Encyclopedia of Chemical Processing

EDITED BY

Sunggyu Lee

Volume 1

Encyclopedia of Chemical Processing

VOLUME I

Encyclopedias from Taylor & Francis Group

Encyclopedia of Biomaterials and Biomedical Engineering (2 Volume Set)
Edited by Gary E. Wnek and Gary Bowlin
ISBN: 0-8247-5562-6

Encyclopedia of Biopharmaceutical Statistics, Second Edition
Edited by Shein-Chung Chow
ISBN: 0-8247-4261-3

Encyclopedia of Chemical Processing (5 Volume Set)
Edited by Sunggyu Lee
ISBN: 0-8247-5563-4

Encyclopedia of Chromatography, Second Edition (2 Volume Set)
Edited by Jack Cazes
ISBN: 0-8247-2785-1

Encyclopedia of Clinical Pharmacy
Edited by Joseph DiPiro
ISBN: 0-8247-0752-4

Encyclopedia of Corrosion Technology, Second Edition
Edited by P.E. Schweitzer
ISBN: 0-8247-4878-6

Encyclopedia of Dietary Supplements
Edited by Paul M. Coates, Marc R. Blackman, Gordon M. Cragg, Mark Levine, Joel Moss, and Jeffrey D. White
ISBN: 0-8247-5504-9

Encyclopedia of Library and Information Science, Second Edition (4 Volume Set)
Edited by Miriam Drake
ISBN: 0-8247-2075-X

Encyclopedia of Medical Genomics and Proteomics (2 Volume Set)
Edited by Jürgen Fuchs and Maurizio Podda
ISBN: 0-8247-5564-2

Dekker Encyclopedia of Nanoscience and Nanotechnology (5 Volume Set)
Edited by James A. Schwarz, Cristian Contescu, and Karol Putyera
ISBN: 0-8247-5055-1

Encyclopedia of Optical Engineering (3 Volume Set)
Edited by Ronald Driggers
ISBN: 0-8247-0940-3

Encyclopedia of Pharmaceutical Technology, Second Edition (3 Volume Set)
Edited by James Swarbrick
ISBN: 0-8247-2825-4

Encyclopedia of Public Administration and Public Policy (2 Volume Set)
Edited by Jack Rabin
ISBN: 0-8247-4748-8

Encyclopedia of Supramolecular Chemistry (2 Volume Set)
Edited by Jerry L. Atwood and Jonathan Steed
ISBN: 0-8247-5056-X

Encyclopedia of Surface and Colloid Science (4 Volume Set)
Edited by Arthur T. Hubbard
ISBN: 0-8247-0633-1

Dekker Agropedia Collection (7 Volume Set)

ISBN: 0-8247-2194-2
(Also available individually)

Encyclopedia of Agricultural, Food, and Biological Engineering
Edited by Dennis R. Heldman
ISBN: 0-8247-0938-1

Encyclopedia of Animal Science
Edited by Wilson G. Pond and Alan Bell
ISBN: 0-8247-5496-4

Encyclopedia of Pest Management
Edited by David Pimentel
ISBN: 0-8247-0632-3

Encyclopedia of Plant and Crop Science
Edited by Robert M. Goodman
ISBN: 0-8247-0944-6

Encyclopedia of Soil Science, Second Edition
Edited by Rattan Lal
ISBN: 0-8493-3830-1

Encyclopedia of Water Science
Edited by B.A. Stewart and Terry Howell
ISBN: 0-8247-0948-9

Coming Soon

Encyclopedia of Biotechnology in Agriculture and Food
Edited by Dennis R. Heldman
ISBN: 0-8493-5027-1

Encyclopedia of Pharmaceutical Technology, Third Edition (6 Volume Set)
Edited by James Swarbrick
ISBN: 0-8493-9399-X

Encyclopedia of Surface and Colloid Science, Second Edition (8 Volume Set)
Edited by Poniss Somasundaran
ISBN: 0-8493-9615-8

Encyclopedia of Energy Engineering (2 Volume Set)
Edited by Barney L. Capehart
ISBN: 0-8493-3653-8

These titles are available both in print and online. To order, visit:

www.crcpress.com

Telephone: 1-800-272-7737 • Fax: 1-800-374-3401

E-Mail: orders@crcpress.com

Encyclopedia of Chemical Processing

VOLUME 1

Edited by:

Sunggyu Lee

Department of Chemical Engineering
University of Missouri - Columbia
Columbia, Missouri
U.S.A.

Taylor & Francis
Taylor & Francis Group
New York London

Published in 2006 by
Taylor & Francis Group
270 Madison Avenue
New York, NY 10016

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 0-8247-5563-4 (Set)
International Standard Book Number-10: 0-8247-5500-6 (Vol 1)
International Standard Book Number-10: 0-8247-5557-X (Vol 2)
International Standard Book Number-10: 0-8247-5558-8 (Vol 3)
International Standard Book Number-10: 0-8247-5559-6 (Vol 4)
International Standard Book Number-10: 0-8247-5560-X (Vol 5)
International Standard Book Number-13: 978-0-8247-5563-8 (Set)
International Standard Book Number-13: 978-0-8247-5500-3 (Vol 1)
International Standard Book Number-13: 978-0-8247-5557-7 (Vol 2)
International Standard Book Number-13: 978-0-8247-5558-4 (Vol 3)
International Standard Book Number-13: 978-0-8247-5559-1 (Vol 4)
International Standard Book Number-13: 978-0-8247-5560-7 (Vol 5)

Library of Congress Cataloging-in-Publication Data

Catalog record is available from the Library of Congress

informa
Taylor & Francis Group is the Academic Division of Informa plc.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

This Encyclopedia is dedicated to
my wife

Kyung Paik Lee

who put up with me,
lovingly supported me,
and helped make this venture
a success

Sunggyu Lee
Editor

Department of Chemical Engineering
University of Missouri–Columbia
Columbia, Missouri, U.S.A.

Editorial Advisory Board

Contributors

Kim Aasberg-Petersen / *Haldor Topsøe A/S, Lyngby, Denmark*

Mohamed O. Abdalla / *Department of Chemistry, Tuskegee University, Tuskegee, Alabama, U.S.A.*

Abdullah M. Aitani / *King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia*

G. Akay / *Process Intensification and Miniaturization Centre, School of Chemical Engineering and Advanced Materials and Institute for Nanoscale Science and Technology, University of Newcastle, Newcastle Upon Tyne, U.K.*

S. Al-Malaika / *Polymer Processing and Performance Research Unit, School of Engineering and Applied Science, Aston University, Birmingham, U.K.*

Lyle F. Albright / *School of Chemical Engineering, Purdue University, West Lafayette, Indiana, U.S.A.*

T. L. Alford / *Department of Chemical and Materials Engineering, Arizona State University, Tempe, Arizona, U.S.A.*

S. W. Allison / *Engineering Science and Technology Division, Oak Ridge National Laboratory (ORNL), Oak Ridge, Tennessee, U.S.A.*

M. Cengiz Altan / *School of Aerospace and Mechanical Engineering, University of Oklahoma, Norman, Oklahoma, U.S.A.*

Ramin Amin-Sanayei / *Arkema Inc., King of Prussia, Pennsylvania, U.S.A.*

Paul Andersen / *Coperion Corporation, Ramsey, New Jersey, U.S.A.*

Ee Lui Ang / *Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

Piero M. Armenante / *Otto H. York Department of Chemical Engineering, New Jersey Institute of Technology, Newark, New Jersey, U.S.A.*

David S. J. Arney / *3M Company, St. Paul, Minnesota, U.S.A.*

David B. Asay / *Department of Chemical Engineering, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.*

Gabriel Ascanio / *URPEI, Department of Chemical Engineering, Ecole Polytechnique, Montreal, Quebec, Canada*

W. R. Ashurst / *University of California–Berkeley, Berkeley, California, U.S.A.*

Naveed Aslam / *University of South Florida, Tampa, Florida, U.S.A.*

Victor Atiemo-Obeng / *Engineering Science and Market Development, The Dow Chemical Company, Midland, Michigan, U.S.A.*

André Bakker / *Fluent Inc., Lebanon, New Hampshire, U.S.A.*

Michael W. Balakos / *R&D, Süd-Chemie Inc., Louisville, Kentucky, U.S.A.*

Shankha K. Banerji / *Department of Civil & Environmental Engineering, University of Missouri–Columbia, Columbia, Missouri, U.S.A.*

Jimmie R. Baran / *3M Company, St. Paul, Minnesota, U.S.A.*

Stanley M. Barnett / *University of Rhode Island, Kingston, Rhode Island, U.S.A.*

Shubhayu Basu / *Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio, U.S.A.*

Roger N. Beers / *The Goodyear Tire & Rubber Company, Akron, Ohio, U.S.A.*

Céline T. Bellehumeur / *Department of Chemical and Petroleum Engineering, University of Calgary, Calgary, Alberta, Canada*

T. J. Bencic / *Optical Instrumentation Technology Branch, NASA John H. Glenn Research Center at Lewis Field, Cleveland, Ohio, U.S.A.*

Jonathan W. Bender / *Department of Chemical Engineering, University of South Carolina, Columbia, South Carolina, U.S.A.*

David A. Benko / *The Goodyear Tire & Rubber Company, Akron, Ohio, U.S.A.*

Sujata K. Bhatia / *Dupont Central Research and Development, Wilmington, Delaware, U.S.A.*

Surita R. Bhatia / *Department of Chemical Engineering, University of Massachusetts–Amherst, Amherst, Massachusetts, U.S.A.*

P. R. Bishnoi / *Department of Chemical and Petroleum Engineering, University of Calgary, Calgary, Alberta, Canada*

T. Reg. Bott / *School of Engineering, Chemical Engineering, University of Birmingham, Birmingham, U.K.*

Andrea Bozzano / *UOP LLC, Des Plaines, Illinois, U.S.A.*

Mike Bradford / *Jacobs Engineering Group Inc., Houston, Texas, U.S.A.*

Ian D. Brindle / *Brock University, St. Catharines, Ontario, Canada*

Edmundo Brito-De La Fuente / *Departamento de Alimentos y Biotecnología, UNAM, México, México*

Michael C. Brooks / *U.S. Environmental Protection Agency, Kerr Research Center, Ada, Oklahoma, U.S.A.*

Nigel D. Browning / *Department of Chemical Engineering and Materials Science, University of California Davis, Lawrence Berkeley National Laboratory, Berkeley, California, U.S.A.*

David A. Bruce / *Department of Chemical and Biomolecular Engineering, Clemson University, Clemson, South Carolina, U.S.A.*

Joel G. Burken / *Department of Civil, Architectural and Environmental Engineering, University of Missouri–Rolla, Rolla, Missouri, U.S.A.*

J. R. Burns / *Protensive Ltd., Bioscience Centre, Centre for Life, Newcastle Upon Tyne, U.K.*

Richard V. Calabrese / *Department of Chemical and Biomolecular Engineering, University of Maryland, College Park, Maryland, U.S.A.*

Gregg Caldwell / *3M Company, St. Paul, Minnesota, U.S.A.*

Gerard T. Caneba / *Department of Chemical Engineering, Michigan Technological University, Houghton, Michigan, U.S.A.*

Giovanni Maria Carlomagno / *Department of Energetics, Thermofluidynamics and Environmental Control (DETEC), University of Naples Federico II, Napoli, Italy*

C. Carraro / *University of California–Berkeley, Berkeley, California, U.S.A.*

Eldon D. Case / *Chemical Engineering and Materials Science Department, Michigan State University, East Lansing, Michigan, U.S.A.*

M. R. Cates / *Engineering Science and Technology Division, Oak Ridge National Laboratory (ORNL), Oak Ridge, Tennessee, U.S.A.*

John C. Chadwick / *Dutch Polymer Institute (DPI), Laboratory of Polymer Chemistry, Eindhoven University of Technology, Eindhoven, The Netherlands*

Louay M. Chamra / *Southeast Cooling, Heating and Power Application Center, Department of Mechanical Engineering, Mississippi State University, Mississippi State, Mississippi, U.S.A.*

Jamal Chaouki / *Ecole Polytechnique, Montreal, Quebec, Canada*

Vicki Chen / *UNESCO Centre for Membrane Science and Technology, School of Chemical Engineering, University of New South Wales, Sydney, New South Wales, Australia*

Zhilei Chen / *Center for Biophysics and Computational Biology, University of Illinois, Urbana, Illinois, U.S.A.*

Hyoung J. Choi / *Department of Polymer Science and Engineering, Inha University, Incheon, Korea*

Kyu Yong Choi / *Department of Chemical Engineering, University of Maryland, College Park, Maryland, U.S.A.*

Pyoungho Choi / *Albany Nanotech, The College of Nanoscale Science and Engineering (CNSE), State University of New York, Albany, New York, U.S.A.*

T. C. Chung / *Department of Materials Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.*

Matthew A. Clarke / *Department of Chemical and Petroleum Engineering, University of Calgary, Calgary, Alberta, Canada*

Gary Combes / *School of Chemical Engineering and Industrial Chemistry, The University of New South Wales, Sydney, New South Wales, Australia*

Richard F. Cope / *Fluid Mechanics and Mixing Group, The Dow Chemical Company, Midland, Michigan, U.S.A.*

Richard Corkish / *ARC Centre of Excellence for Advanced Silicon Photovoltaics and Photonics, University of New South Wales, Sydney, New South Wales, Australia*

R. A. Cottis / *School of Materials, Corrosion and Protection Centre, University of Manchester, Manchester, U.K.*

Sean A. Curran / *Honeywell International Inc., Morristown, New Jersey, U.S.A.*

Wayne R. Curtis / *Department of Chemical Engineering, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.*

Teresa J. Cutright / *Department of Civil Engineering, The University of Akron, Akron, Ohio, U.S.A.*

Qizhou Dai / *Advanced Biomaterials Chemistry, University of British Columbia, Vancouver, British Columbia, Canada*

A. K. Dalai / *Catalysis and Chemical Reactor Engineering Laboratories, Department of Chemical Engineering, University of Saskatchewan, Saskatoon, Canada*

Douglas A. Dale / *Genencor International, Palo Alto, California, U.S.A.*

Rohit P. Datar / *Technical Operations, CPKelco, Okmulgee, Oklahoma, U.S.A.*

Ravindra Datta / *Fuel Cell Center, Worcester Polytechnic Institute, Worcester, Massachusetts, U.S.A.*

Sharad M. Dave / *Bhabha Atomic Research Center, Mumbai, India*

Forrest M. Davidson, III / *The University of Texas, Austin, Texas, U.S.A.*

Frank Davis / *Cranfield University, Silsoe, U.K.*

Mark DeDecker / *Firestone Polymers, Bridgestone/Firestone Research LLC, Akron, Ohio, U.S.A.*

Fariba Dehghani / *School of Chemical Engineering and Industrial Chemistry, The University of New South Wales, Sydney, New South Wales, Australia*

Shuguang Deng / *Chemical Engineering Department, New Mexico State University, Las Cruces, New Mexico, U.S.A.*

Amy S. Determan / *Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa, U.S.A.*

Glenn B. DeWolf / *URS Corporation, Austin, Texas, U.S.A.*

Partha Dey / *P.A. Consulting, Nashville, Tennessee, U.S.A.*

R. Dhib / *Department of Chemical Engineering, Ryerson University, Toronto, Ontario, Canada*

Huu D. Doan / *Department of Chemical Engineering, Ryerson University, Toronto, Ontario, Canada*

Mildred S. Dresselhaus / *Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

Nishith Dwivedi / *Department of Chemical Engineering, Indian Institute of Technology, New Delhi, India*

Vahid Ebadat / *Chilworth Technology, Inc., Princeton, New Jersey, U.S.A.*

Sina Ebnesajjad / *DuPont Fluoroproducts, Chestnut Run Plaza, Wilmington, Delaware, U.S.A.*

Jeremy S. Edwards / *Department of Chemical Engineering, University of Delaware, Newark, Delaware, U.S.A.*

Brian W. Eggiman / *Purdue University, West Lafayette, Indiana, U.S.A.*

J. Richard Elliott, Jr. / *Department of Chemical Engineering, University of Akron, Akron, Ohio, U.S.A.*

Morinobu Endo / *Faculty of Engineering, Shinshu University, Wakasato, Nagano-shi, Japan*

Rolf Erni / *Department of Chemical Engineering and Materials Science, University of California Davis, Lawrence Berkeley National Laboratory, Berkeley, California, U.S.A.*

Arthur W. Etchells / *DuPont Fellow, Philadelphia, Pennsylvania, U.S.A.*

L.-S. Fan / *Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio, U.S.A.*

Rajeev Farwaha / *Celanese Polymers, Bridgewater, New Jersey, U.S.A.*

James J. Feng / *Department of Chemical and Biological Engineering and Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada*

D. Ferdous / *Catalysis and Chemical Reactor Engineering Laboratories, Department of Chemical Engineering, University of Saskatchewan, Saskatoon, Canada*

Jim C. Fitch / *Noria Corporation, Tulsa, Oklahoma, U.S.A.*

J. F. Forbes / *Department of Chemical and Materials Engineering, University of Alberta, Edmonton, Alberta, Canada*

Neil R. Foster / *School of Chemical Engineering and Industrial Chemistry, The University of New South Wales, Sydney, New South Wales, Australia*

Jonathan Francis / *University of Central Lancashire, Preston, U.K.*

Matthew H. Frey / *3M Company, St. Paul, Minnesota, U.S.A.*

Joanna D. Fromstein / *Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, Ontario, Canada*

Mina Gabriel / *Honeywell International Inc., Morristown, New Jersey, U.S.A.*

Alfred Gaertner / *Genencor International, Palo Alto, California, U.S.A.*

Prabhu Ganesan / *Department of Chemical Engineering, University of South Carolina, Columbia, South Carolina, U.S.A.*

Shubhra Gangopadhyay / *University of Missouri–Columbia, Columbia, Missouri, U.S.A.*

Hugo S. Garcia / *UNIDA, Instituto Tecnologico de Veracruz, Veracruz, Mexico*

Dinesh Gera / *Fluent Incorporated, Morgantown, West Virginia, U.S.A.*

Richard Gilbert / *Department of Wood and Paper Science, North Carolina State University, Raleigh, North Carolina, U.S.A.*

Giuseppe Giorleo / *Department of Materials and Production Engineering (DIMP), University of Naples Federico II, Napoli, Italy*

S. M. Goedeke / *Engineering Science and Technology Division, Oak Ridge National Laboratory (ORNL), Oak Ridge, Tennessee, U.S.A.*

Scott Gold / *Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

Vincent G. Gomes / *University of Sydney, Sydney, New South Wales, Australia*

James G. Goodwin, Jr. / *Department of Chemical Engineering, Clemson University, Clemson, South Carolina, U.S.A.*

John R. Grace / *University of British Columbia, Vancouver, British Columbia, Canada*

Brian P. Grady / *School of Chemical, Biological and Materials Engineering, University of Oklahoma, Norman, Oklahoma, U.S.A.*

Dan F. Graves / *Firestone Polymers, Bridgestone/Firestone Research LLC, Akron, Ohio, U.S.A.*

Erich Grotewold / *Department of Plant Cellular and Molecular Biology and Plant Biotechnology Center, The Ohio State University, Columbus, Ohio, U.S.A.*

Rajiv Grover / *Jacobs Engineering Group Inc., Houston, Texas, U.S.A.*

Ronald W. Gumbs / *Gumbs Associates Ltd., East Brunswick, New Jersey, U.S.A.*

L. Jay Guo / *Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan, U.S.A.*

Ram B. Gupta / *Department of Chemical Engineering, Auburn University, Auburn, Alabama, U.S.A.*

Byung Gwon Lee / *Environment and Process Technology Division, Korea Institute of Science and Technology, Cheongryang, Seoul, South Korea*

Don C. Haddox / *Waterbury, Vermont, U.S.A.*

Joel M. Haight / *Pennsylvania State University, University Park, Pennsylvania, U.S.A.*

Pradeep Haldar / *Albany Nanotech, The College of Nanoscale Science and Engineering (CNSE), State University of New York, Albany, New York, U.S.A.*

Ian Hamerton / *University of Surrey, Surrey, U.K.*

Youssef K. Hamidi / *School of Aerospace and Mechanical Engineering, University of Oklahoma, Norman, Oklahoma, U.S.A.*

Tobias Hanrath / *The University of Texas, Austin, Texas, U.S.A.*

Douglas P. Harrison / *Gordon A. and Mary Cain Department of Chemical Engineering, Louisiana State University, Baton Rouge, Louisiana, U.S.A.*

Gareth P. Harrison / *Institute for Energy Systems, School of Engineering and Electronics, University of Edinburgh, Edinburgh, U.K.*

Takuya Hayashi / *Faculty of Engineering, Shinshu University, Wakasato, Nagano-shi, Japan*

Douglas G. Hayes / *Department Biosystems Engineering and Environmental Science, University of Tennessee, Knoxville, Tennessee, U.S.A.*

Donald T. Haynie / *Biomedical Engineering and Physics, Bionanosystems Engineering Laboratory, Center for Applied Physics Studies, Louisiana Tech University, Ruston, Louisiana, U.S.A.*

Roland H. Heck / *Princeton University, Princeton, New Jersey, U.S.A.*

William L. Hergenrother / *Bridgestone Americas Center for Research and Technology, Bridgestone/Firestone Research LLC, Akron, Ohio, U.S.A.*

Andrew M. Herring / *Department of Chemical Engineering, Colorado School of Mines, Golden, Colorado, U.S.A.*

Séamus P. J. Higson / *Cranfield University, Silsoe, U.K.*

Charles G. Hill, Jr. / *Department of Chemical and Biological Engineering, University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.*

John O. Hill / *La Trobe University, Melbourne, Victoria, Australia*

Hugh W. Hillhouse / *Purdue University, West Lafayette, Indiana, U.S.A.*

B. Keith Hodge / *Southeast Cooling, Heating and Power Application Center, Department of Mechanical Engineering, Mississippi State University, Mississippi State, Mississippi, U.S.A.*

J. D. Holladay / *Pacific Northwest National Laboratory, Richland, Washington, U.S.A.*

W. A. Hollerman / *Department of Physics, University of Louisiana at Lafayette, Lafayette, Louisiana, U.S.A.*

W. C. Hsu / *Sino-American Silicon Product Inc., Hsinchu, Taiwan, ROC*

X. D. Hu / *R&D, Süd-Chemie Inc., Louisville, Kentucky, U.S.A.*

Yinlun Huang / *Department of Chemical Engineering and Materials Science, Wayne State University, Detroit, Michigan, U.S.A.*

Anton Huber / *PolySaccharide Initiative, Institut für Chemie, Karl-Franzens, Universität Graz, Graz, Austria*

Kang Moo Huh / *Department of Polymer Science and Engineering, Chungnam National University, Daejeon, South Korea*

Raymond L. Huhnke / *Biosystems and Agricultural Engineering, Oklahoma State Unviersity, Stillwater, Oklahoma, U.S.A.*

Steven C. Hukvari / *Parr Instrument Company, Moline, Illinois, U.S.A.*

Scott M. Husson / *Department of Chemical and Biomolecular Engineering, Clemson University, Clemson, South Carolina, U.S.A.*

Sangchul Hwang / *Department of Civil Engineering and Surveying, University of Puerto Rico, Mayagüez, Puerto Rico*

Oleg Ilinich / *Engelhard Corporation, Iselin, New Jersey, U.S.A.*

R. J. J. Jachuck / *Process Intensification and Clean Technology (PICT) Group, Department of Chemical Engineering, Clarkson University, Potsdam, New York, U.S.A.*

Manish Jain / *3M Company, St. Paul, Minnesota, U.S.A.*

Krishnan Jayaraman / *University of Auckland, Auckland, New Zealand*

Nigil Satish Jeyashekar / *Department of Chemical Engineering, University of Mississippi, University, Mississippi, U.S.A.*

Myung S. Jhon / *Department of Chemical Engineering and Data Storage Systems Center, Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A.*

Tyler Johannes / *Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

Bob Johnson / *School of Physical and Chemical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia*

Joshua Jurs / *Departments of Chemistry, Mechanical Engineering and Materials Science, and Center for Nanoscale Science and Technology, Rice University, Houston, Texas, U.S.A.*

John F. Kadla / *Faculty of Forestry, Biomaterials Chemistry, The University of British Columbia, Vancouver, British Columbia, Canada*

Kishore K. Kar / *Fluid Mechanics and Mixing Group, The Dow Chemical Company, Midland, Michigan, U.S.A.*

Thomas E. Karis / *Hitachi Global Storage Technologies, San Jose Research Center, San Jose, California, U.S.A.*

S. Komar Kawatra / *Department of Chemical Engineering, Michigan Technological University, Houghton, Michigan, U.S.A.*

David O. Kazmer / *Department of Plastics Engineering, University of Massachusetts, Lowell, Massachusetts, U.S.A.*

Jason M. Keith / *Department of Chemical Engineering, Michigan Technological University, Houghton, Michigan, U.S.A.*

Sunil Kesavan / *Akebono Corporation, Farmington Hills, Michigan, U.S.A.*

Daeik Kim / *University of Southern California, Los Angeles, California, U.S.A.*

Ji W. Kim / *Department of Chemical Engineering and Data Storage Systems Center, Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A.*

Jin-Kuk Kim / *Centre for Process Integration, University of Manchester, Manchester, U.K.*

Seong H. Kim / *Department of Chemical Engineering, Materials Research Institute, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.*

Yoong-Ahm Kim / *Faculty of Engineering, Shinshu University, Wakasato, Nagano-shi, Japan*

Brian P. Kirkmeyer / *International Flavors and Fragrances, Union Beach, New Jersey, U.S.A.*

Dehong Kong / *Chilworth Technology, Inc., Monmouth Junction, New Jersey, U.S.A.*

Brian A. Korgel / *The University of Texas, Austin, Texas, U.S.A.*

Milivoje M. Kostic / *Department of Mechanical Engineering, Northern Illinois University, DeKalb, Illinois, U.S.A.*

J. Kouvetakis / *Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona, U.S.A.*

Andrzej Kraslawski / *Lappeenranta University of Technology, Lappeenranta, Finland*

Suzanne M. Kresta / *Department of Chemical and Materials Engineering, University of Alberta, Edmonton, Alberta, Canada*

Satoshi Kubo / *Faculty of Forestry, Biomaterials Chemistry, The University of British Columbia, Vancouver, British Columbia, Canada*

Mark A. Kuehne / *NL Chemical Technology, Inc., Mount Prospect, Illinois, U.S.A.*

Ashok Kumar / *Department of Civil Engineering, University of Toledo, Toledo, Ohio, U.S.A.*

Vimal Kumar / *Department of Chemical Engineering, Indian Institute of Technology, New Delhi, India*

Arunabha Kundu / *Department of Chemical Engineering, Indian Institute of Technology, New Delhi, India*

Charlotte T. M. Kwok / *Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

K. S. Lackner / *Department of Earth and Environmental Engineering, Columbia University, New York, New York, U.S.A.*

Harry J. Lader / *Harry Lader and Associates, Inc., Cleveland, Ohio, U.S.A.*

Yadunandan Lal Dar / *Corporate Research, National Starch and Chemical Company, Bridgewater, New Jersey, U.S.A.*

Joseph M. Lambert, Jr. / *Parr Instrument Company, Moline, Illinois, U.S.A.*

C. W. Lan / *Department of Chemical Engineering, National Taiwan University, Taipei, Taiwan, ROC*

H. Bryan Lanterman / *Department of Chemical Engineering, University of Missouri–Columbia, Columbia, Missouri, U.S.A.*

Pierre Le-Clech / *UNESCO Centre for Membrane Science and Technology, School of Chemical Engineering, University of New South Wales, Sydney, New South Wales, Australia*

James M. Lee / *Department of Chemical Engineering, Washington State University, Pullman, Washington, U.S.A.*

L. James Lee / *Department of Chemical Engineering, The Ohio State University, Columbus, Ohio, U.S.A.*

Sunggyu Lee / *Department of Chemical Engineering, University of Missouri–Columbia, Columbia, Missouri, U.S.A.*

Yoon Seob Lee / *Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio, U.S.A.*

Youn-Woo Lee / *Environment and Process Technology Division, Korea Institute of Science and Technology, Cheongryang, Seoul, South Korea*

Kwok-Wai Lem / *Honeywell International Inc., Morristown, New Jersey, U.S.A.*

Douglas E. Leng / *Leng Associates, Midland, Michigan, U.S.A.*

Alan G. Letki / *Alfa Laval Inc., Warminster, Pennsylvania, U.S.A.*

Christopher Lew / *Department of Chemical and Environmental Engineering, University of California at Riverside, Riverside, California, U.S.A.*

Randy S. Lewis / *School of Chemical Engineering, Oklahoma State University, Stillwater, Oklahoma, U.S.A.*

Bingyun Li / *Biomedical Engineering and Physics, Bionanosystems Engineering Laboratory, Center for Applied Physics Studies, Louisiana Tech University, Ruston, Louisiana, U.S.A.*

Jane C. Li / *NL Chemical Technology, Inc., Mount Prospect, Illinois, U.S.A.*

Jun Li / *Department of Polymer Science, University of Southern Mississippi, Hattiesburg, Mississippi, U.S.A.*

Norman N. Li / *NL Chemical Technology, Inc., Mount Prospect, Illinois, U.S.A.*

Shuang Li / *Department of Chemical and Environmental Engineering, University of California at Riverside, Riverside, California, U.S.A.*

Wenping Li / *Department of Chemical Engineering, University of Houston, Houston, Texas, U.S.A.*

Xiaodong Li / *Department of Mechanical Engineering, University of South Carolina, Columbia, South Carolina, U.S.A.*

Zijian Li / *Department of Chemical and Environmental Engineering, University of California at Riverside, Riverside, California, U.S.A.*

Michael K. Lindell / *Hazard Reduction and Recovery Center, Texas A&M University, College Station, Texas, U.S.A.*

Chung-Chiun Liu / *Electronic Design Center and Chemical Engineering Department, Case Western Reserve University, Cleveland, Ohio, U.S.A.*

Henry Liu / *Freight Pipeline Company, Columbia, Missouri, U.S.A.*

Daniel G. Löffler / *Quarens Technologies, Inc., Bend, Oregon, U.S.A.*

Ali Lohi / *Department of Chemical Engineering, Ryerson University, Toronto, Ontario, Canada*

Stephen J. Lombardo / *Department of Chemical Engineering, University of Missouri–Columbia, Columbia, Missouri, U.S.A.*

Helen H. Lou / *Department of Chemical Engineering, Lamar University, Beaumont, Texas, U.S.A.*

Jorge A. Lubguban / *University of Missouri–Columbia, Columbia, Missouri, U.S.A.*

Douglas K. Ludlow / *Chemical and Biological Engineering, University of Missouri–Rolla, Rolla, Missouri, U.S.A.*

Adriane G. Ludwick / *Department of Chemistry, Tuskegee University, Tuskegee, Alabama, U.S.A.*

Juergen Lueske / *Fluid Mechanics and Mixing Group, The Dow Chemical Company, Niedersachsen, Germany*

Xiaoliang Ma / *Clean Fuels and Catalysis Program, The Energy Institute, and Department of Energy and Geo-Environmental Engineering, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.*

R. Maboudian / *University of California–Berkeley, Berkeley, California, U.S.A.*

Sundararajan V. Madihally / *School of Chemical Engineering, Oklahoma State University, Stillwater, Oklahoma, U.S.A.*

Surya K. Mallapragada / *Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa, U.S.A.*

Raffaella Mammucari / *School of Chemical Engineering and Industrial Chemistry, The University of New South Wales, Sydney, New South Wales, Australia*

Elizabeth Marden Marshall / *Fluent Inc., Lebanon, New Hampshire, U.S.A.*

T. E. Marlin / *Department of Chemical Engineering, McMaster University, Hamilton, Ontario, Canada*

Stanley Marple / *Chemical Engineering Department, University of Houston, Houston, Texas, U.S.A.*

Richard I. Masel / *Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

Takeshi Matsuura / *Department of Chemical Engineering, University of Ottawa, Ottawa, Ontario, Canada*

J. W. Mayer / *Department of Chemical and Materials Engineering, Arizona State University, Tempe, Arizona, U.S.A.*

Kevin P. Menard / *PerkinElmer Thermal Laboratory, Materials Science Department, University of North Texas, Denton, Texas, U.S.A.*

Carosena Meola / *Department of Energetics, Thermofluidynamics and Environmental Control (DETEC), University of Naples Federico II, Napoli, Italy*

John C. Middleton / *BHR Group Ltd., Cranfield, Bedfordshire, U.K.*

Jan D. Miller / *Department of Metallurgical Engineering, University of Utah, Salt Lake City, Utah, U.S.A.*

Patrick L. Mills / *Chemical Science and Engineering Laboratory, DuPont Company, Wilmington, Delaware, U.S.A.*

Kishore K. Mohanty / *Department of Chemical Engineering, University of Houston, Houston, Texas, U.S.A.*

Sanat Mohanty / *3M Company, St. Paul, Minnesota, U.S.A.*

Satish C. Mohapatra / *Advanced Fluid Technologies, Inc., Dynalene Heat Transfer Fluids, Whitehall, Pennsylvania, U.S.A.*

Alexander B. Morgan / *Dow Chemical Company, Core R&D, Midland, Michigan, U.S.A.*

Sarah E. Morgan / *Department of Polymer Science, University of Southern Mississippi, Hattiesburg, Mississippi, U.S.A.*

Maximiliano Luis Munford / *Group of Organic Optoelectronic Devices, Departamento de Física, Universidade Federal do Paraná, Curitiba, Paraná, Brazil*

Balaji Narasimhan / *Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa, U.S.A.*

Amarnath Nareddy / *Department of Chemical and Biomolecular Engineering, Clemson University, Clemson, South Carolina, U.S.A.*

Robert Naumann / *University of Alabama, Huntsville, Alabama, U.S.A.*

Flora T. T. Ng / *Department of Chemical Engineering, University of Waterloo, Waterloo, Ontario, Canada*

Anh V. Nguyen / *Discipline of Chemical Engineering, The University of Newcastle, Callaghan, New South Wales, Australia*

Zheng Ni / *Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

K. D. P. Nigam / *Department of Chemical Engineering, Indian Institute of Technology, New Delhi, India*

Takamasa Nonaka / *Faculty of Engineering, Department of Applied Chemistry and Biochemistry, Kumamoto University, Kurokami, Kumamoto-shi, Japan*

Paul O'Connor / *Akzo Nobel Catalysts, Amersfoort, The Netherlands*

Kimberly Ogden / *Department of Chemical and Environmental Engineering, University of Arizona, Tucson, Arizona, U.S.A.*

Masahiro Ohshima / *Department of Chemical Engineering, Kyoto University, Nishikyo-ku, Kyoto, Japan*

Cristina Otero / *Departamento de Biocatálisis, Instituto de Catalisis y Petroleoquimica, CSIC, Campus Universidad Autonoma, Cantoblanco, Madrid, Spain*

Richard P. Palluzi / *ExxonMobil Research and Engineering Company, Annandale, New Jersey, U.S.A.*

A.-H. Park / *Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio, U.S.A.*

Kinam Park / *Departments of Pharmaceutics and Biomedical Engineering, Purdue University, West Lafayette, Indiana, U.S.A.*

Jae Hyung Park / *Department of Advanced Polymer and Fiber Materials, Kyung Hee University, Gyeonggi-do, South Korea*

Kenneth R. Parker / *Ken Parker Consultant APL, West Midlands, U.K.*

André Avelino Pasa / *Thin Films and Surfaces Group, Departamento de Física, Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, Brazil*

Shwetal Patel / *Department of Chemical Engineering, University of Delaware, Newark, Delaware, U.S.A.*

Rajam Pattabiraman / *Central Electrochemical Research Institute, Karaikudi, India*

Edward L. Paul / *Merck and Co., Inc., Sea Girt, New Jersey, U.S.A.*

Jing Peng / *Department of Applied Chemistry, College of Chemistry, Peking University, Beijing, People's Republic of China*

A. Penlidis / *Institute for Polymer Research, Department of Chemical Engineering, University of Waterloo, Waterloo, Ontario, Canada*

W. Roy Penney / *Department of Chemical Engineering, University of Arkansas, Fayetteville, Arkansas, U.S.A.*

Mario A. Perez / *3M Company, St. Paul, Minnesota, U.S.A.*

Ronald W. Perry / *School of Public Affairs, Arizona State University, Tempe, Arizona, U.S.A.*

Ralph W. Pike / *Department of Chemical Engineering, Louisiana State University, Baton Rouge, Louisiana, U.S.A.*

Branko N. Popov / *Department of Chemical Engineering, University of South Carolina, Columbia, South Carolina, U.S.A.*

Mark A. Prelas / *Nuclear Science and Engineering Institute, University of Missouri–Columbia, Columbia, Missouri, U.S.A.*

Todd Pugsley / *University of Saskatchewan, Saskatoon, Saskatchewan, Canada*

Peter R. Pujadó / *UOP LLC, Des Plaines, Illinois, U.S.A.*

Yunying Qi / *Shell Global Solutions (US) Inc., Westhollow Technology Center, Houston, Texas, U.S.A.*

Mehrdad Rafat / *Department of Chemical Engineering, University of Ottawa, Ottawa, Ontario, Canada*

Theodore W. Randolph / *Department of Chemical Engineering, University of Colorado, Boulder, Colorado, U.S.A.*

Harish G. Rao / *LFR Inc., Elgin, Illinois, U.S.A.*

Sanjeev N. Rao / *University of Auckland, Auckland, New Zealand*

James F. Rathman / *Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio, U.S.A.*

Louis G. Reifschneider / *Department of Technology, Illinois State University, Normal, Illinois, U.S.A.*

David G. Retzloff / *Department of Chemical Engineering, University of Missouri–Columbia, Columbia, Missouri, U.S.A.*

Richard W. Rice / *Department of Chemical Engineering, Clemson University, Clemson, South Carolina, U.S.A.*

J. A. Richardson / *Anticorrosion Consulting, Durham, U.K.*

Peter L. Rinaldi / *Department of Chemistry, The University of Akron, Akron, Ohio, U.S.A.*

Clayt Robinson / *Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio, U.S.A.*

Ken K. Robinson / *Mega-Carbon Company, St. Charles, Illinois, U.S.A.*

Alexander F. Routh / *Department of Chemical and Process Engineering, University of Sheffield, Sheffield, U.K.*

Wolfgang Ruettinger / *Engelhard Corporation, Iselin, New Jersey, U.S.A.*

Gregory T. Rushton / *Department of Chemistry and Biochemistry, Kennesaw State University, Kennesaw, Georgia, U.S.A.*

James M. Ryan / *Ryan Consulting, Inc., Fort Myers, Florida, U.S.A.*

Ajit Sadana / *Department of Chemical Engineering, University of Mississippi, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.*

Krishnendu Saha / *Nuclear Science and Engineering Institute, University of Missouri–Columbia, Columbia, Missouri, U.S.A.*

Sangrama K. Sahoo / *Department of Chemistry, The University of Akron, Akron, Ohio, U.S.A.*

C. Sanchez / *Chimie de la Matière Condensée, Université Pierre et Marie Curie, Jussieu, Paris, France*

Abhay Sardesai / *Department of Chemical Engineering, University of Missouri–Columbia, Columbia, Missouri, U.S.A.*

Robert J. Schmidt / *UOP LLC, Des Plaines, Illinois, U.S.A.*

Karl B. Schnelle, Jr. / *Department of Chemical Engineering, Vanderbilt University, Nashville, Tennessee, U.S.A.*

M. J. Scorah / *Institute for Polymer Research, Department of Chemical Engineering, University of Waterloo, Waterloo, Ontario, Canada*

Edmund G. Seebauer / *Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

Arup K. Sengupta / *Department of Civil and Environmental Engineering, Lehigh University, Bethlehem, Pennsylvania, U.S.A.*

Sukalyan Sengupta / *Civil and Environmental Engineering Department, University of Massachusetts, Dartmouth, Massachusetts, U.S.A.*

Selim M. Senkan / *Department of Chemical Engineering, University of California, Los Angeles, California, U.S.A.*

Xinming Shao / *Akebono Corporation, Farmington Hills, Michigan, U.S.A.*

Zengyi Shao / *Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

J. M. Shaw / *Department of Chemical and Materials Engineering, University of Alberta, Edmonton, Alberta, Canada*

H. H. Sheena / *Polymer Processing and Performance Research Unit, School of Engineering and Applied Science, Aston University, Birmingham, U.K.*

Ken D. Shimizu / *Department of Chemistry and Biochemistry, University of South Carolina, Columbia, South Carolina, U.S.A.*

Allen R. Siedle / *3M Company, St. Paul, Minnesota, U.S.A.*

Michael R. Simurdiak / *Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

Azad Singh / *Department of Chemical Engineering, Indian Institute of Technology, New Delhi, India*

Shivaji Sircar / *Chemical Engineering Department, Lehigh University, Bethlehem, Pennsylvania, U.S.A.*

Andrew W. Sloley / *VECO USA, Inc., Bellingham, Washington, U.S.A.*

Gary S. Smith / *Arkema lnc., King of Prussia, Pennsylvania, U.S.A.*

John M. Smith / *School of Engineering, University of Surrey, Guildford, Surrey, U.K.*

Robin Smith / *Centre for Process Integration, University of Manchester, Manchester, U.K.*

Ryan G. Soderquist / *Department of Chemical Engineering, Washington State University, Pullman, Washington, U.S.A.*

Stephen Sohn / *UOP LLC, Des Plaines, Illinois, U.S.A.*

G. J. A. A. Soler-Illia / *Unidad de Actividad Química, CNEA, Centro Atómico Constituyentes, San Martín, Buenos Aires, Argentina*

Chunshan Song / *Clean Fuels and Catalysis Program, The Energy Institute and Department of Energy and Geo-Environmental Engineering, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.*

Richard Q. Song / *NL Chemical Technology, Inc., Mount Prospect, Illinois, U.S.A.*

James G. Speight / *CD & W Inc., Laramie, Wyoming, U.S.A.*

Vijay R. Srinivas / *Arkema lnc., King of Prussia, Pennsylvania, U.S.A.*

Jason Stephenson / *Departments of Chemistry, Mechanical Engineering and Materials Science, and Center for Nanoscale Science and Technology, Rice University, Houston, Texas, U.S.A.*

Michael R. Stoner / *Department of Chemical Engineering, University of Colorado, Boulder, Colorado, U.S.A.*

Truman S. Storvick / *Chemical Engineering Department, University of Missouri–Columbia, Columbia, Missouri, U.S.A.*

Kenneth Strawhecker / *Department of Chemical Engineering, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.*

Kyung W. Suh / *Midland, Michigan, U.S.A.*

Timothy P. Sullivan / *Randolph Air Force Base, Texas, U.S.A.*

Steve Sund / *Honeywell International Inc., Morristown, New Jersey, U.S.A.*

Aydin K. Sunol / *University of South Florida, Tampa, Florida, U.S.A.*

Sermin G. Sunol / *University of South Florida, Tampa, Florida, U.S.A.*

Kenji Takeuchi / *Faculty of Engineering, Shinshu University, Wakasato, Nagano-shi, Japan*

Philippe A. Tanguy / *URPEI, Department of Chemical Engineering, Ecole Polytechnique, Montreal, Quebec, Canada*

Andreas Taubert / *Department of Chemistry, University of Basel, Basel, Switzerland*

Maruicio Terrones / *Advanced Materials Department, IPICyT, San Luis Potosí, SLP, México*

Francis Thibault / *URPEI, Department of Chemical Engineering, Ecole Polytechnique, Montreal, Quebec, Canada*

Louis J. Thibodeaux / *Louisiana State University, Baton Rouge, Louisiana, U.S.A.*

Cristina U. Thomas / *3M Company, St. Paul, Minnesota, U.S.A.*

Karsten E. Thompson / *Gordon A. and Mary Cain Department of Chemical Engineering, Louisiana State University, Baton Rouge, Louisiana, U.S.A.*

Frank M. Tiller / *Department of Chemical Engineering, University of Houston, Houston, Texas, U.S.A.*

Maria P. Torres / *Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa, U.S.A.*

James M. Tour / *Departments of Chemistry, Mechanical Engineering and Materials Science, and Center for Nanoscale Science, and Technology, Rice University, Houston, Texas, U.S.A.*

Linh T. T. Tran / *Discipline of Chemical Engineering, The University of Newcastle, Callaghan, New South Wales, Australia*

Drew D. Troyer / *Noria Corporation, Tulsa, Oklahoma, U.S.A.*

Maxwell Tsai / *NL Chemical Technology, Inc., Mount Prospect, Illinois, U.S.A.*

Tom Chunghu Tsai / *The Dow Chemical Company, Houston, Texas, U.S.A.*

Uday T. Turaga / *Clean Fuels and Catalysis Program, The Energy Institute and Department of Energy and Geo-Environmental Engineering, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.*

Simant R. Upreti / *Department of Chemical Engineering, Ryerson University, Toronto, Ontario, Canada*

Vikrant N. Urade / *Purdue University, West Lafayette, Indiana, U.S.A.*

Ivo F. J. Vankelecom / *Centre for Surface Chemistry and Catalysis, Department of Interphase Chemistry, Faculty of Agricultural and Applied Biological Sciences, Katholieke Universiteit Leuven, Leuven, Belgium*

Charanya Varadarajan / *Department of Civil Engineering, University of Toledo, Toledo, Ohio, U.S.A.*

Angel Velez / *Nuclear Science and Engineering Institute, University of Missouri–Columbia, Columbia, Missouri, U.S.A.*

Abhilash Vijayan / *Department of Civil Engineering, University of Toledo, Toledo, Ohio, U.S.A.*

V. V. Viswanathan / *Pacific Northwest National Laboratory, Richland, Washington, U.S.A.*

Dionisios G. Vlachos / *Department of Chemical Engineering and Center for Catalytic Science and Technology, University of Delaware, Newark, Delaware, U.S.A.*

Tuan Vo-Dinh / *Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.*

Bipin V. Vora / *UOP LLC, Des Plaines, Illinois, U.S.A.*

Chun Wang / *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

J. R. White / *School of Chemical Engineering and Advanced Materials, University of Newcastle Upon Tyne, Newcastle Upon Tyne, U.K.*

Kimberly A. Woodhouse / *Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, Ontario, Canada*

Guy B. Woodle / *UOP LLC, Des Plaines, Illinois, U.S.A.*

Jiangning Wu / *Department of Chemical Engineering, Ryerson University, Toronto, Ontario, Canada*

Eleanore T. Wurtzel / *Department of Biological Sciences, Lehman College, The City University of New York, Bronx, New York, U.S.A.*

Nicholas Patrick Wynn / *Sulzer Chemtech GmbH, Neunkirchen, Germany*

Xuekun Xing / *NTK Powderdex, Inc., Wixom, Michigan, U.S.A. and Department of Chemical Engineering, Case Western Reserve University, Cleveland, Ohio, U.S.A.*

Yushan Yan / *Department of Chemical and Environmental Engineering, University of California at Riverside, Riverside, California, U.S.A.*

Shang-Tian Yang / *Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio, U.S.A.*

Xiaobo Yang / *Institute of Physical Chemistry and Electrochemistry, University of Hanover, Hanover, Germany*

Hirotsugu K. Yasuda / *University of Missouri–Columbia, Columbia, Missouri, U.S.A.*

Mingli Ye / *Departments of Pharmaceutics and Biomedical Engineering, Purdue University, West Lafayette, Indiana, U.S.A.*

Jeffrey H. Yen / *Arkema Inc., King of Prussia, Pennsylvania, U.S.A.*

Teh Fu Yen / *University of Southern California, Los Angeles, California, U.S.A.*

W. S. Yip / *Suncor Energy Inc., Fort McMurray, Alberta, Canada*

W. C. Yu / *Department of Molecular Science and Engineering, National Taipei University of Technology, Taipei, Taiwan, ROC*

John Zabasajja / *MOS 12, Technology and Manufacturing, Motorola Semiconductor Products Sector (SPS), Chandler, Arizona, U.S.A.*

Jacques L. Zakin / *Department of Chemical Engineering, The Ohio State University, Columbus, Ohio, U.S.A.*

Sergei V. Zelentsov / *Chemical Department, Nizhny Novgorod State University, Nizhny Novgorod, Russia*

Nadezda V. Zelentsova / *Chemical Department, Nizhny Novgorod State University, Nizhny Novgorod, Russia*

Ying Zhang / *Department of Chemical Engineering, The Ohio State University, Columbus, Ohio, U.S.A.*

Huimin Zhao / *Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

Haishan Zheng / *Department of Chemical Engineering, Michigan Technological University, Houghton, Michigan, U.S.A.*

X.-Y. Zou / *Oilphase-DBR, Schlumberger, Edmonton, Alberta, Canada*

Contents

Volume 1

Volume 2

Volume 3

Volume 4

Volume 5

Topical Table of Contents

Advanced Materials

Analytical Methods and Characterization

Batteries and Fuel Cells

Biomaterials, Materials for Biological and Biomedical Applications

Biotechnology and Biological Processing

Catalysis and Catalyst Preparation

Catalysis and Catalyst Preparation (*cont'd*)

Electrochemical Processes and Products

Energy Processing and Renewable Energy

Environmental Technology and Regulations

Mass Transfer and Mixing

Mass Transfer and Mixing (*cont'd*)

Materials Application and Processing

Nanotechnology

Particle Technology

Petrochemicals and Petrochemical Processing

Petroleum and Fuel Processing

Pipeline Technology

Plasma Technology

Polymer Processing

Polymeric Materials and Polymerization

Process Design, Control, Optimization, and Simulation

Process Safety and Loss Prevention

Reactor Engineering and Design

Semiconductor Processing and Microelectronics

Separation Processes

Separation Processes (*cont'd*)

Specialty Chemicals

Supercritical Fluid Technology

Thermodynamics and Process Applications

Transport Phenomena and Applications

Unit Operations and Design

Preface

The *Encyclopedia of Chemical Processing* is an authoritative, dynamic, and most comprehensive multi-volume reference work on the broad subject of chemical processing, which will enable readers to have an enriching experience about general as well as targeted knowledge in this field. The *Encyclopedia* caters to engineers, scientists, researchers, inventors, professors, and students, as well as general readers in academia, industry, research institutions, government, and legal practices. In addition, the *Encyclopedia* has been designed to address the needs of practicing engineers and scientists, businessmen, lawyers, industrial executives, and professionals in the chemical processing and technology arena.

The *Encyclopedia* encompasses the entire realm of chemical processing, offering up-to-date, reliable, and comprehensive coverage of process technologies that have steadily progressed over the years, and at the same time identifying and addressing new breakthroughs and emerging technologies in chemical processing. The *Encyclopedia* contains a large number of entries that are devoted to life science subjects and futuristic materials and technologies, namely, biotechnology, nanomaterials and nanotechnology, and materials and technologies geared for microelectronics. Under the advice of an editorial advisory board comprised of distinguished and renowned scholars from around the world, the *Encyclopedia* will serve as the most respected reference work in the field of chemical processing.

The *Encyclopedia* covers cradle-to-grave information on processing novel materials, emerging process technologies and resultant materials, and manufacturing organic and inorganic chemicals. Specific topics of interest include synthesis reactions, properties and characterization of materials, appropriate choice of catalysts, reactor design, process flowsheets, energy integration practices, pinch design, design of separation equipment and peripherals, environmental aspects of chemical plant operation such as safety and loss prevention, obedience of environmental regulations, waste reduction and management, and much more. The *Encyclopedia* also contains descriptions of different types of reactors and separation systems and their design, unit operations, system integration, process system peripherals such as pumps, valves, and controllers, analytical techniques and equipment, as well as pilot plant design and scale-up criteria.

Fundamental aspects of industrial catalytic processes are detailed including catalyst preparation, characterization, structure-property relationships, deactivation and defouling, and catalyst regeneration methods. Examples of industrial processes that use different types of catalysts for chemical manufacture are also detailed. Identification and utilization of alternative resources for complementing our energy needs are addressed, which include renewable energy resources, oxygenated fuels, biofuels, fuel cells, and batteries.

Polymers are ubiquitous in today's life, and their utilization is limited only by chemists' and chemical engineers' imaginations. The *Encyclopedia* attempts to cover the wide spectrum of polymerization and polymer processing, including metallocene processes, description of structure, properties and end use of different polymers, copolymers, polymer blends and composites, polymer coatings, and rubber compounds. Additional topics of interest that are also covered include but are not limited to polymer characterization, molding technology, and polymer and rubber recycling. Advanced materials being used in myriad applications are also accounted for; examples of these are ceramics, nanomaterials, nanocomposites, carbon nanotubes, hydrophilic polymers, photovoltaic materials, biomaterials, and biomedical materials.

The issues potentially related to global warming cannot be understated in the present world, largely due to the increasing use of fossil fuels by automobiles and industries around the world. The degree to which this environmental issue affects society and the remedial measures needed in order to alleviate these concerns are well-addressed in the *Encyclopedia*, which covers emerging environmental technology, bioremediation, greenhouse gas alleviation, waste minimization, sequestration of carbon dioxide, etc.

Biotechnology is a rapidly growing area of chemical, biological, and life sciences, and as such is also well-covered in the *Encyclopedia*. Enzymes, biomaterials, bioseparation, bioprocessing, bioreactor design, biocatalysis, BioMEMS, protein design, chiral drug separation, and hydrogels are few of the topics of merit that are included.

The *Encyclopedia* also identifies and addresses emerging technologies in great detail including but not limited to nanotechnology, plasma technology, thin film technology, supercritical fluid technology and its applications, as well as microfabrication and micromachining for the microelectronics area.

The authors of this initial printed version of the *Encyclopedia* are recognized experts in their fields, lending credibility and prestige to the *Encyclopedia*. All the authors were invited based on their records of accomplishment in the chosen topical areas. All entries were individually reviewed by peers as well as the Editor. As part of the review and revision processes, every effort was exercised to maintain the consistency, accuracy, readability, and up-to-date nature of the information presented.

The *Encyclopedia* is published in both online and printed formats. The printed version consists of multiple traditional hardbound volumes with articles arranged alphabetically. The online version of the *Encyclopedia* is created by coupling the content of the printed edition with a powerful search engine, user-friendly interface, and customer-focused features. The online database is dynamic and evolving in nature, with additional articles added each quarter.

The Editor feels honored to have been asked to undertake the important and challenging endeavor of developing the *Encyclopedia of Chemical Processing* that will cater to the needs of the rapidly changing world of the 21st century. The Editor is humbled to follow the impeccable work of the previous editor, Professor John J. McKetta, who led the development of the *Encyclopedia of Chemical Processing and Design*, a total of 69 volumes, which has become one of the most authoritative reference sources for scientists, engineers, and practitioners for several decades.

I would like to express my most sincere thanks and appreciation to the authors for their excellent professionalism and dedicated work. Needless to say, an encyclopedia of this nature would never exist if the expert authors had not devoted their valuable time to preparing the authoritative entries on their assigned topics. I wish to thank all my colleagues and friends as well as the editorial board members for all their suggestions, comments, assistance, volunteerism, and patience. In particular, I appreciate the encouragements, guidance, and assistance provided by Mr. Russell Dekker, Dr. Chai-sung Lee, Dr. John C. Angus, Dr. C. C. Liu, Dr. James G. Speight, Dr. Robert Dye, Dr. Sunil Kesavan, Dr. John Zabasajja, Dr. J. Richard Elliott, Jr., Dr. Abhay Sardesai, Dr. Hirotsugu Yasuda, Dr. David G. Retzloff, Dr. Patricia Roberts, Dr. Kelly Clark, Dr. Jeffrey Yen, Dr. Peter Pujado, and Dr. Stephen J. Lombardo. I also would like to thank Mr. Jonathan E. Wenzel, Ms. Leah A. Leavitt, Dr. Teresa J. Cutright, Dr. H. Bryan Lanterman, Dr. Qingsong Yu, and Dr. Patricia A. Darcy for providing various assistance while editing. I am also deeply indebted to the former and current employees of the Publisher for their dedicated work toward successful completion of the project, to name a few, Ms. Alison Cohen, Ms. Oona Schmid, Ms. Marisa Hoheb, Ms. Maria Kelley, Ms. Meaghan Johnson, and Ms. Joanne Jay. The contributions of those mentioned made this *Encyclopedia* possible.

Sunggyu Lee
Editor

Absorption Equipment

Karl B. Schnelle, Jr.
Chemical Engineering Department, Vanderbilt University, Nashville, Tennessee, U.S.A.

Partha Dey
P. A. Consulting, Nashville, Tennessee, U.S.A.

INTRODUCTION

Absorption is a mass transfer operation in which a soluble gaseous component is removed from a gas stream by dissolving in a liquid. Absorption can be used to recover valuable gaseous components such as hydrocarbons or to remove unwanted gaseous components such as hydrogen sulfide from a stream. A valuable solute can be separated from the absorbing liquid and recovered in a pure, concentrated form by distillation or stripping (desorption). The absorbing liquid is then used in a closed circuit and is continuously regenerated and recycled. Examples of regeneration alternatives to distillation or stripping are removal through precipitation and settling; chemical destruction through neutralization, oxidation, or reduction; hydrolysis; solvent extraction; and liquid adsorption. Absorption is one of the main methods of separation used in the chemical processing industry. Accompanied by chemical reaction between the absorbed component and a reagent in the absorbing fluid, absorption can become a very effective means of separation. Absorption can also be used to remove an air pollutant like an acid gas from stream. Then, the system could be a simple absorption in which the absorbing liquid is used in a single pass and then disposed of while containing the absorbed pollutant.

Operations of Absorption Towers

In the past it was the custom to call absorbers operating as cleanup towers to remove undesirable gaseous effluents by the name of scrubber. At that time most of the effluent gases being removed were acid gases being scrubbed with water. The designation of scrubber to scrub the discharge gas and clean it seemed rather natural. Today the same kind of operation is carried out, but with more stringent regulations imposed by the local air pollution control agency. The name scrubber is now applied to those operations in which particulate matter is removed but the scrubbing operation may also include the simultaneous removal of gaseous pollutants. In this chapter the term absorber will refer to the removal of gaseous contaminants.

General Considerations

Filters, heat exchangers, dryers, bubble cap columns, cyclones, etc., are ordinarily designed and built by process equipment manufacturers. However, units of special design for one-of-a-kind operations such as packed or plate towers are quite often designed and built under the supervision of plant engineers. Thus, there is a large variety of this type of equipment, none of it essentially standard.

TYPES OF ABSORPTION EQUIPMENT

Absorption takes place in either staged or plate towers or continuous or packed contactor. However, in both cases the flow is continuous. In the ideal equilibrium stage model, two phases are contacted, well mixed, come to equilibrium, and then are separated with no carryover. Real processes are evaluated by expressing efficiency as a percentage of the change that would occur in the ideal stages. Any liquid carryover is removed by mechanical means.

In the continuous absorber the two immiscible phases are in continuous and tumultuous contact within a vessel that is usually a tall column. A large surface is made available by packing the column with ceramic or metal materials. The packing provides more surface area and a greater degree of turbulence to promote mass transfer. The penalty for using packing is in the increased pressure loss in moving the fluids through the column, which causes an increased demand for energy. In the usual countercurrent flow column, the lighter phase enters the bottom and passes upward. Transfer of material takes place by molecular and eddy diffusion processes across the interface between the immiscible phases. Contact may be also cocurrent or cross-flow. Columns for the removal of

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007642

air contaminants are usually designed for countercurrent or cross-flow operation.

Absorption can take place in a countercurrent, cocurrent, or cross-flow device. Vertical countercurrent towers are either built with a metal, plastic, or ceramic packing or constructed as plate towers with various types of plates. This chapter will discuss the solvents used to carry out absorption and the various types of absorption equipment.

ABSORPTION SOLVENTS

Absorption systems can be divided into those that use water as the primary absorbing liquid and those that use a low-volatility organic liquid. The gas solubility should be high in the absorbing solvent. The gas leaving an absorber is usually saturated with the solvent; therefore, the solvent should have a low vapor pressure. A lower viscosity solvent is advantageous to promote more rapid absorption rates and improve flooding characteristics. The solvent should not be corrosive to the materials of construction of the absorber. It should be nontoxic and nonflammable. Depending on the region where the absorber is to be constructed, the solvent should have a low freezing point.

Nonaqueous Systems

At first glance, an organic liquid appears to be the preferred solvent for absorbing hydrocarbon and organic vapors from a gas stream because of improved solubility and miscibility. The lower heat of vaporization of organic liquids results in energy conservation when solvent regeneration must occur by stripping. Many heavy oils such as No. 2 fuel oil or heavier and other solvents with low vapor pressure can do extremely well in reducing organic vapor concentrations to low levels. Care must be exercised in picking a solvent that will have sufficiently low vapor pressure so that the solvent itself will not become a source of volatile organic pollution. Obviously, the treated gas will be saturated with the absorbing solvent. An absorber–stripper system for recovery of benzene vapors has been described by Crocker.[1] Other aspects of organic solvent absorption requiring consideration are stability of the solvent in the gas solvent system, for example, its resistance to oxidation, and its possible fire and explosion hazard.

Although water is the most common liquid used for absorbing acidic gases, amines (monoethanol-, diethanol-, and triethanolamine; methyldiethanolamine; and dimethylaniline) have been used for absorbing SO_2 and H_2S from hydrocarbon gas streams. Such absorbents are generally limited to solid particulate free systems because solids can produce difficult to handle sludge as well as use up valuable organic absorbents. Furthermore, because of absorbent cost, absorbent regeneration must be practiced in almost all cases.

Aqueous Systems

Absorption is one of the most frequently used methods for removal of water-soluble gases. Acidic gases such as HCl, HF, and SiF_4 can be absorbed in water efficiently and readily, especially if the last contact is made with water that has been made alkaline. Less soluble acidic gases such as SO_2, Cl_2, and H_2S can be absorbed more readily in a dilute caustic solution. The scrubbing liquid may be made alkaline with dissolved soda ash or sodium bicarbonate, or with sodium hydroxide, usually with no higher a concentration in the scrubbing liquid than 5–10%. Lime is a cheaper and more plentiful alkali, but its use directly in the absorber may lead to plugging or coating problems if the calcium salts produced have only limited solubility. A technique often used is the two-step flue gas desulfurization process, where the absorbing solution containing NaOH is used inside the absorption tower, and then the tower effluent is treated with lime externally, precipitating the absorbed component as a slightly soluble calcium salt. The precipitate may be removed by thickening and the regenerated sodium alkali solution is recycled to the absorber. Scrubbing with an ammonium salt solution can also be employed. In such cases, the gas is often first contacted with the more alkaline solution and then with the neutral or slightly acid contact to prevent stripping losses of NH_3 to the atmosphere.

When flue gases containing CO_2 are being scrubbed with an alkaline solution to remove other acidic components, the caustic consumption can be inordinately high if CO_2 is absorbed. However, if the pH of the scrubbing liquid entering the absorber is kept below 9.0, the amount of CO_2 absorbed can be kept low. Conversely, alkaline gases, such as NH_3, can be removed from the main gas stream with acidic water solutions such as dilute H_2SO_4, H_3PO_4, or HNO_3. Single-pass scrubbing solutions so used can often be disposed of as fertilizer ingredients. Alternatives are to remove the absorbed component by concentration and crystallization. The absorbing gas must have adequate solubility in the scrubbing liquid at the resulting temperature of the gas–liquid system.

For pollutant gases with limited water solubility, such as SO_2 or benzene vapors, the large quantities of water that would be required are generally impractical on a single-pass basis, but may be used in unusual circumstances. An early example from the United Kingdom is the removal of SO_2 from flue gas at the Battersea and Bankside electric power stations, which is described by Rees.[2] Here, the normally alkaline

water from the Thames tidal estuary is used in a large quantity on a one-pass basis.

PACKED TOWERS

There are two major types of packing, random dumped pieces and structured modular forms. The structured packing is usually crimped or corrugated sheets. The packing provides a large interfacial area for mass transfer and should have a low-pressure drop. However, it must permit passage of large volumes of fluid without flooding. The pressure drop should be the result of skin friction and not form drag. Thus, flow should be through the packing and not around the packing. The packing should have enough mechanical strength to carry the load and allow easy handling and installation. It should be able to resist thermal shock and possible extreme temperature changes, and it must be chemically resistant to the fluids being processed.

Random or Dumped Packing

Random packings are dumped into the tower during construction and are allowed to fall at random. The tower might be filled with water, first to allow a gentler settling and to prevent breakage, especially with ceramic-type packing. Random dumped tower packing comes in many different shapes. Two of the most popular are rings and saddles. Sizes range from 0.25 in. to 3.5 in., with 1 in. being a very common size. The choice of a packing is mostly dependent on the service in which the tower will be engaged. Packings are made of ceramic, metal, or plastic, depending on the service. Ceramic materials will withstand corrosion and are therefore used where the solutions resulting are aqueous and corrosive. Metals are used where noncorrosive organic liquids are present. Plastic packing may be used in the case of corrosive aqueous solutions and for organic liquids that are not solvents for the plastic of which the packing is made. Metal packing is more expensive, but provides lower pressure drop and higher efficiency. When using plastic materials, care must be taken that the temperature is not too high and that oxidizing agents are not present. Ring-type packings are commonly made of metal or plastic, except for Raschig rings, which are generally ceramic. Ring-type packings lend themselves to distillation because of their good turndown properties and availability in metals of all types that can be press formed. Usually, ring-type packings are used in handling organic solutions when there are no corrosive problems. However, rings do not promote redistribution of liquids, and Raschig rings may even cause maldistribution. Saddles are commonly made from ceramic or plastic and give good corrosion resistance. Saddles are best for redistribution of liquid and, thus, serve as a good packing for absorption towers.

Structured Packing

Early on after the production of random packings had been used extensively, stacked beds of the conventional random packings such as larger-sized Raschig rings were used as ordered packings. Owing to the high cost of installation of this type of packing, it was largely discontinued. At that time multiple layers of corrugated metal lath formed into a honeycomb structure came into use. Later on, a woven wire mesh arranged in rows of vertically corrugated elements came into use. Subsequently, other wire-mesh structures have gained favor. Then, a sheet metal structured packing was developed to reduce the expense of the wire-mesh type. The use of this structured type of packing not only promotes mass transfer owing to increased surface area, but also has less pressure drop in many different services.

Tower Considerations

Materials of construction

Random packing can be made from ceramic materials, plastic, or metal. Most structured packing and plates in staged towers are made from metal although there are simple woven types of plastic materials that can be considered as structured packing. The tower packing, plates, and tower materials must be compatible with the fluids flowing through the towers. Of particular significance would be acid gases that may have a deleterious effect on metal tower internal parts and organic solvents that may have a serious effect on plastic materials. It is also necessary to consider the case where there may be a high heat of absorption emitted. The internal tower may have to be cooled to withstand the temperature that results from the heat of absorption.

Flow arrangements

In diffusional operations such as absorption where mass is to be transferred from one phase to another, it is necessary to bring the two phases into contact to permit the change toward equilibrium to take place. The transfer may take place with both streams flowing in the same direction, in which case the operation is called concurrent or cocurrent flow. When the two streams flow in the opposite direction, the operation is termed countercurrent flow, an operation carried out with the gas entering at the bottom and flowing

upward and the liquid entering at the top and flowing down. This process is illustrated in Fig. 1. A combined operation in which the contaminated gas is first cleaned in a countercurrent operation, as shown in Fig. 2, and then the gas is further treated to remove more of the contaminant as shown in the cocurrent operation that follows.

Countercurrent operation is the most widely used absorption equipment arrangement. As the gas flow increases at constant liquid flow, liquid holdup must increase. The maximum gas flow is limited by the pressure drop and the liquid holdup that will build up to flooding. Contact time is controlled by the bed depth and the gas velocity. In countercurrent flow mass transfer driving force is maximum at the gas entrance and liquid exit. Cocurrent operation can be carried out at high gas velocities because there is no flooding limit. In fact, liquid holdup decreases as velocity increases. However, the mass transfer driving force is smaller than in countercurrent operation.

Some processes for both absorption and the removal of particulates employ a cross-flow spray chamber operation. Here, the water is sprayed down on a bed of packing material. The carrier gas containing pollutant gas or the particulate flows horizontally through the packing, with the spray and packing causing the absorbed gas or particles to be forced down to the bottom of the spray chamber where they can be removed. Fig. 3 illustrates a cross-flow absorber. The design of cross-flow absorption equipment is more difficult than vertical towers because the area for mass transfer is different for the gas and liquid phases.

Continuous and steady-state operation is usually most economical. However, when smaller quantities of material are processed, it is often more advantageous to charge the entire batch at once. In fact, in many cases this is the only way the process can be done. This is called batch operation and is a transient operation from start-up to shut-down. A batch operation presents a more difficult design problem.

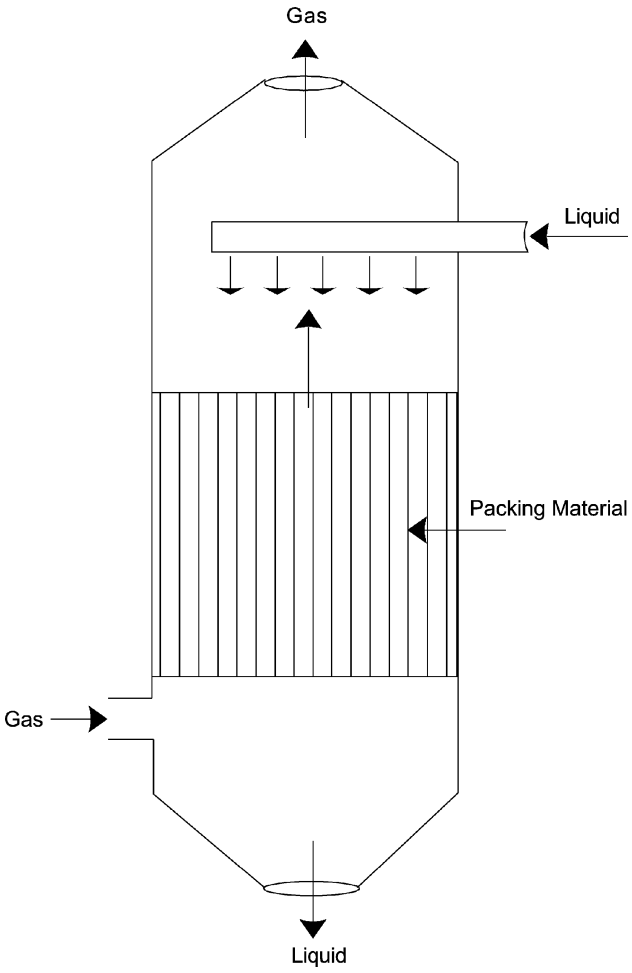


Fig. 1 Countercurrent flow packed tower.

Packed tower internals

In addition to the packing, absorption towers must include internal parts to make a successful piece of operating equipment. Fig. 4 illustrates the placement of the tower internals. These internals begin with a packing support plate at the bottom of the tower. The packing support plate must physically support the weight of the packing. It must incorporate a high percentage of free area to permit relatively unrestricted flow of downcoming liquid. A flat plate has the disadvantage in that both liquids and gases must pass countercurrently through the same holes. Therefore, a substantial hydrostatic head may develop. Furthermore, the bottom layer of packing partially blocks many of the openings reducing the free space. Both of these conditions lower tower capacity. A gas injection plate provides separate passage for gas and liquid and prevents buildup of hydrostatic head.

Liquid distributors are used at all locations where an external liquid stream is introduced. Absorbers and strippers generally require only one distributor, while continuous distillation towers require at least two, at the feed and reflux inlets. The distributors should be 6–12 in. above packing to allow for gas disengagement from the bed. The distributor should provide uniform liquid distribution and a large free area for gas flow.

Liquid redistributors collect downcoming liquid and distribute it uniformly to the bed below. Initially, after entering the tower the liquid tends to flow out to the wall, the redistributor makes that portion of the liquid more available again to the gas flow. It also breaks up the coalescence of the downcoming liquid, and it will eliminate factors that cause a loss of efficiency in the tower and reestablish a uniform pattern of liquid

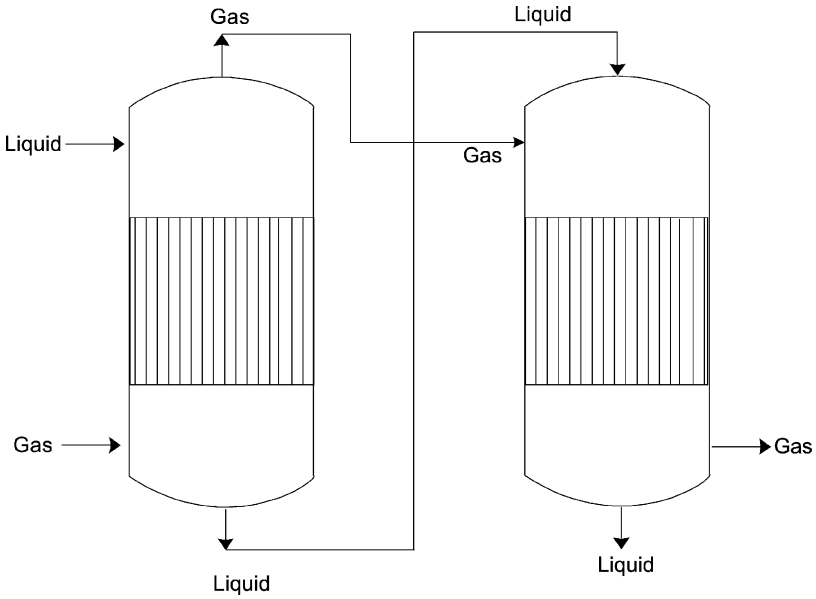


Fig. 2 Combined countercurrent and cocurrent operation.

irrigation. A bed depth of up to 6 m (20 ft) should be alright before redistribution is needed.

Retaining and hold-down plates are used only with ceramic or carbon tower packing. They prevent the upper portion of the packed bed from becoming fluidized and from breaking up during surges in pressure or at high-pressure drop. The plates rest directly on packing and restrict movement by virtue of the weight of the plate. Retainers or bed limiters prevent bed expansion or fluidization. When operating at high-pressure drops, retainers are fastened to the wall. They are designed to prevent individual packing pieces from passing through the plate openings.

PLATE TOWERS

Plate or tray towers are vertical cylinders in which the gas and liquid are contacted on horizontal plates in a stepwise fashion. By the nature of the operation plate towers are countercurrent flow devices. Fig. 5 shows a typical arrangement. In plate columns the gas is introduced at the bottom. Contact between gas and liquid is obtained by forcing the gas to pass upward through small orifices, bubbling through a liquid layer flowing across a plate. The liquid is introduced at the top and passes downward by gravity over the plate and through a downcomer onto the next plate. The

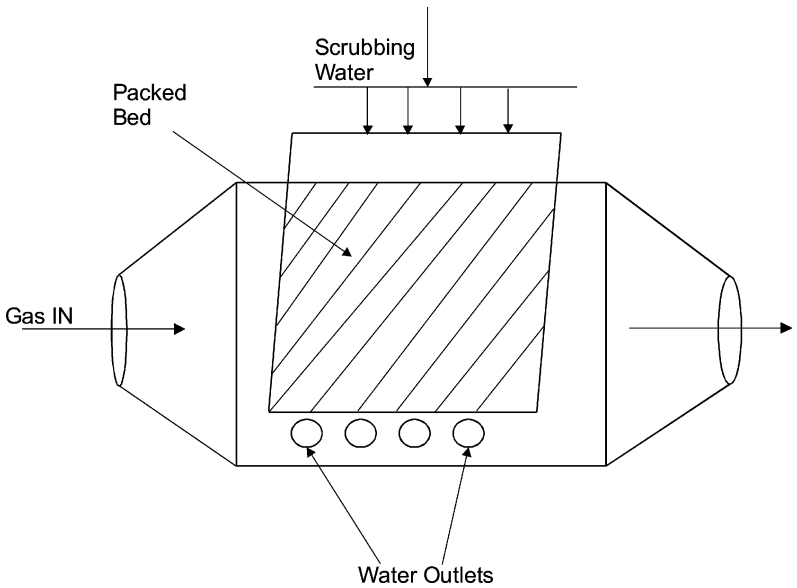


Fig. 3 Cross-flow absorber operation.

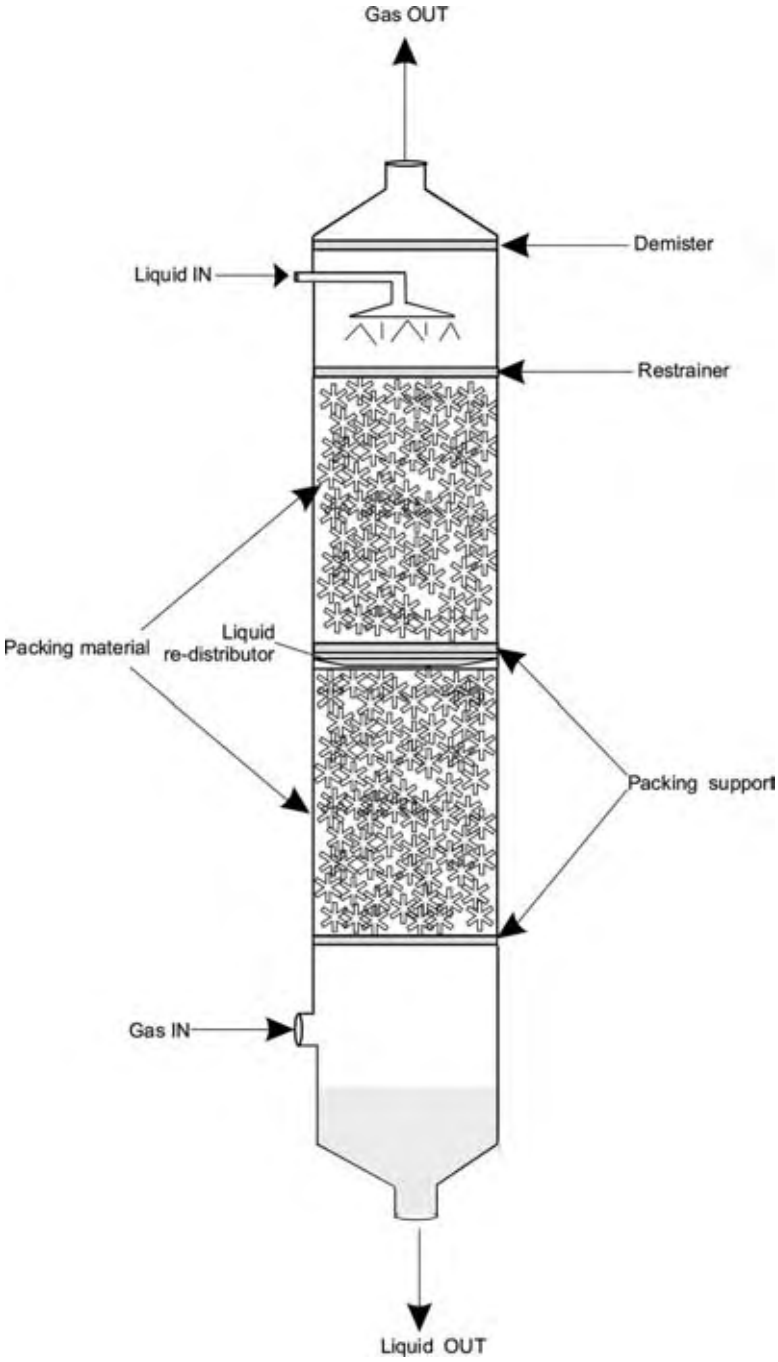


Fig. 4 Packed column with internal hardware indicated. *(View this art in color at www.dekker.com.)*

bubble cap tower is the classical contacting device. Each plate of the tower is a stage in which interphase diffusion occurs and the fluids are separated. Ideally, the vapor and liquid would reach equilibrium at each stage. The number of these ideal stages that are required is determined by the difficulty of separation. The number is calculated from the mass and energy balances around the plate and the tower. The stage or tray efficiency is determined by the mechanical design. Higher contact times result in higher efficiencies. Deeper pools of liquid on the plates promote higher contact times, and higher gas velocities promote better efficiency as well. Unfortunately, these conditions can lead to flooding of the plates and a severe drop in efficiency or foaming of the liquid on the plates. Thus, an inoperative situation might result. Turndown ratio is defined as the ratio of design rate to minimum rate. In many instances of tower design and operation, the performance when operating below the design rate becomes important. Furthermore, plate towers experience the same materials limitations as discussed earlier for packed towers. When selecting

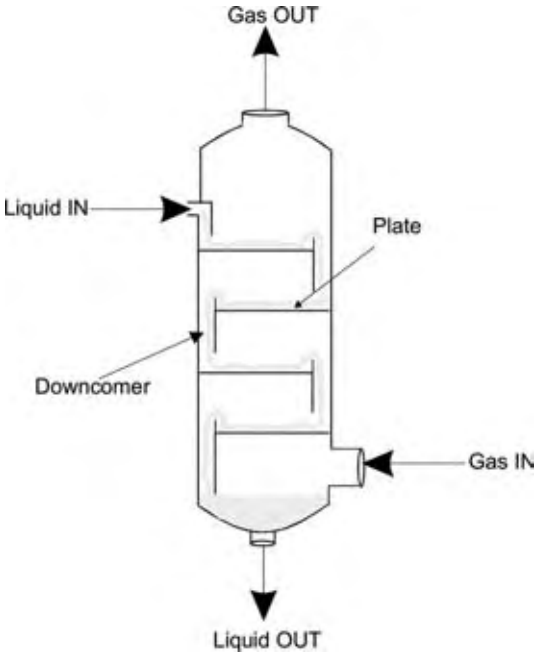


Fig. 5 Plate column. *(View this art in color at www.dekker.com.)*

plate types in addition to materials of construction, turndown ratio, pressure drop, capacity, and efficiency must be considered.

Plate Types

A plate type column may be operated in either a cross-flow or a counterflow method. Cross-flow plates are the most common types. Perry and Green and Wankat discuss the advantages of both type plates.[3,4] Cross-flow plates use a downcomer to transport the liquid from the upper plate to the lower plate. They offer greater mass transfer efficiency and operating range. The downcomer may be located to control the liquid flow pattern. Newer designs of cross-flow plates employ perforations, which may be simple round orifices or may contain movable valve-like vents that act like variable orifices. These type plates will be discussed in the following section.

In counterflow plates there are no downcomers and the liquid and gas use the same openings for flow. The openings are usually round perforations or long slots. The plates may be corrugated to segregate the liquid and gas flow.

Bubble Caps

Bubble cap trays, a cross-flow type of plate, were originally the most common type of tray. On these trays risers lead the gas up through the tray and underneath a cap that is mounted on top of the riser. A series of slots are cut into the cap through which the gas passes into the liquid that is flowing across the plate. They have the advantage of being able to handle wide ranges of liquid and gas flows. However, the new types of trays are much less expensive; therefore, bubble caps are being phased out of use in new tower designs.

Sieve Trays

Sieve plates are simple flat plates perforated with small holes. The advantages are low cost and high plate efficiency but they have narrow gas flow operating ranges. These trays may be subject to flooding because of liquid backing up in the downspouts or excessive entrainment. Fig. 6 is a schematic illustrating bubble cap and sieve trays. Efficiency remains good at design conditions. However, turndown is relatively poor, and therefore, the trays are not flexible in operating conditions. Sieve trays are relatively resistant to clogging and they can have large holes that make them easier to clean. Entrainment is much less than that experienced in bubble cap trays; therefore, plate spacing can be smaller than in bubble cap trays.

Valve Trays

A variation of the bubble cap tray is the valve tray, which permits greater variations in gas flow without dumping the liquid through the gas passages. Valve trays are also cross-flow type and can be described as sieve trays with large variable openings. The openings are covered with movable caps that rise and fall as the gas rate increases and decreases. This keeps the gas velocity through the slots essentially constant. Valve

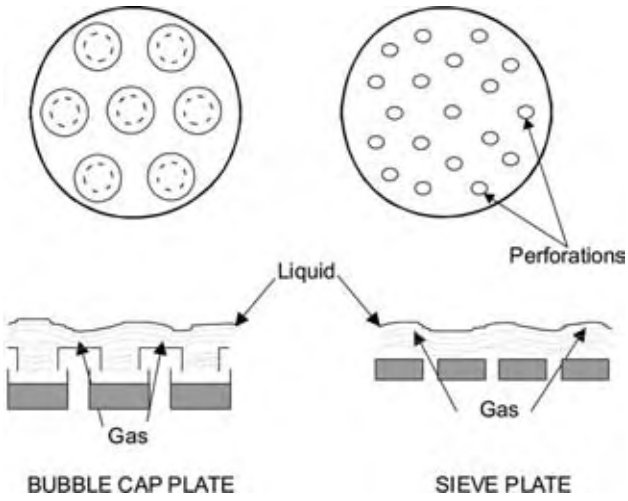


Fig. 6 Bubble cap and sieve tray. *(View this art in color at www.dekker.com.)*

trays are designed to have better turndown ratios than sieve trays and their efficiency remains high as the gas rate drops. Valve trays are more likely to plug if solids are present and are more costly than sieve trays.

Baffle Tower

A counterflow plate-contacting device for absorption is the baffle tower, which has been employed occasionally when plugging and scaling problems are expected to be severe. Fig. 7 illustrates a baffle tower. Gases passing up the tower must pass through sheets of downwardly cascading liquid, providing some degree of contact and liquid atomization. Baffle tower design may use alternating segmental or disk and doughnut plates. Here, the gas alternately flows upward through central orifices and annuli traversing through liquid curtains with each change in direction. Mass transfer is generally poor, and information on design parameters is hard to find.

Spray Chambers

A liquid may be introduced into a tower as fine drops through a nozzle. This device is known as a spray chamber. The flow could be cocurrent or countercurrent. A countercurrent spray chamber is shown in Fig. 8. These towers are considerably more resistant to plugging when solid particulates are present in the inlet gas. However, difficulties with plugging in spray towers and erosion can be troublesome when the spray liquid is recycled. Particle settling followed by fine strainers or even coarse filters is beneficial to eliminate

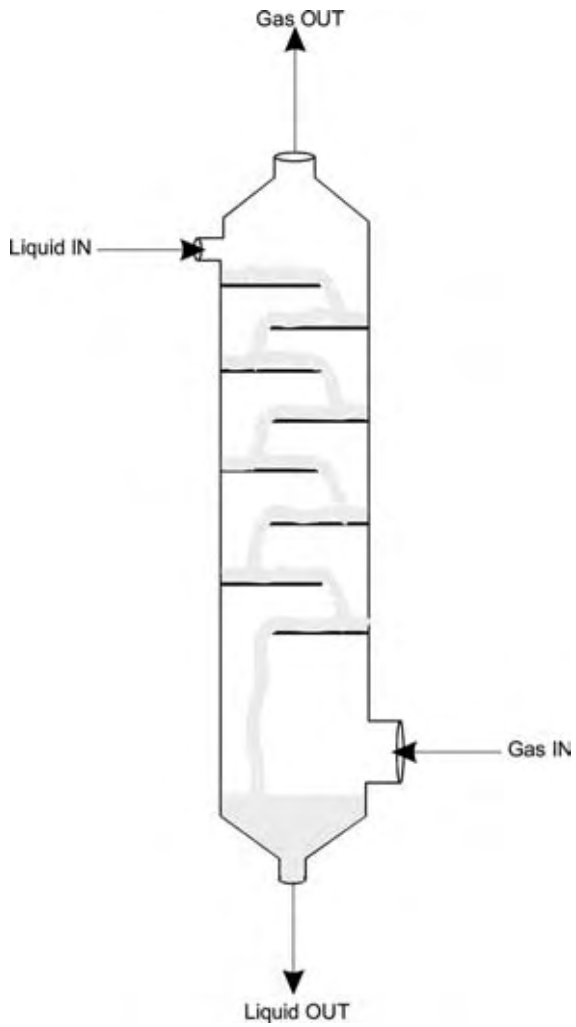


Fig. 7 Baffle tray tower. *(View this art in color at www.dekker.com.)*

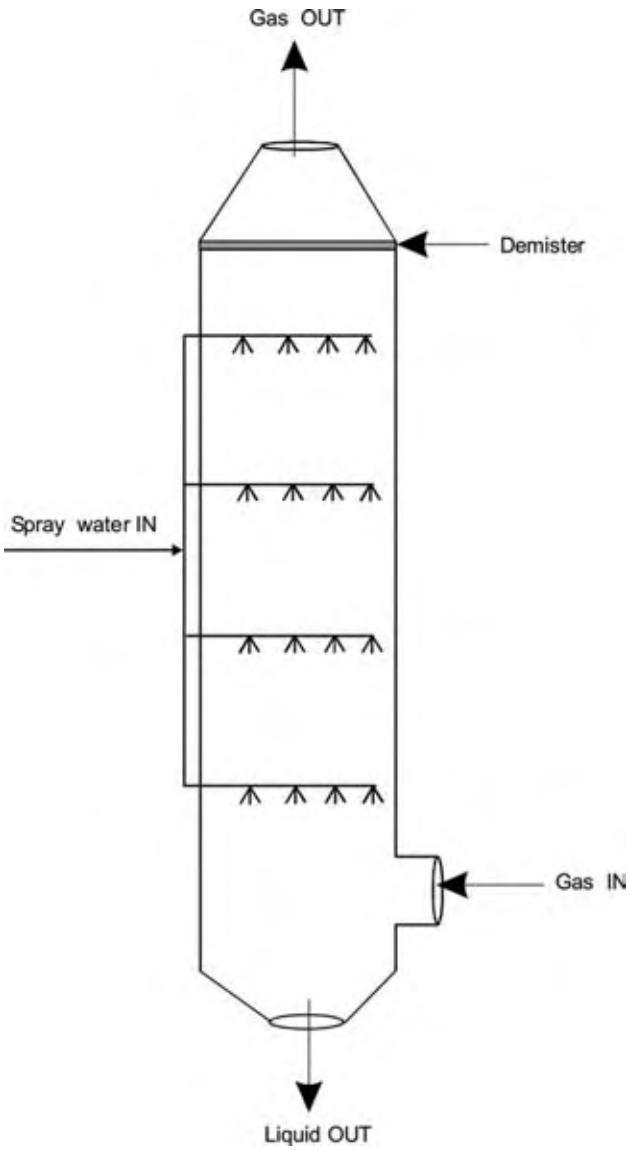


Fig. 8 Spray chamber. *(View this art in color at www.dekker.com.)*

this problem. These devices have the advantage of low-pressure drop but there is a tendency for the liquid to be entrained in the gas leaving the tower. Mist eliminators can help reduce this problem. An additional disadvantage is the cost of pumping the liquid to force it through the nozzles. The efficiency of spray chambers can be improved by introducing the feed into the tower in a cyclonic manner.

ABSORPTION FOR AIR POLLUTION CONTROL

Absorption plays a major role today in air pollution control. In the first part of this chapter it was noted that absorption was referred to as scrubbing especially when associated with cleaning up a stream containing an acid gas before it was emitted to the atmosphere. Today in air pollution control technology the term scrubber continues to be used with reference to cleanup of pollutant gases but usually when the equipment used also removes particulate matter. The combined action of removal of particulates and gases takes place in venturi scrubbers, spray towers, plate towers, and other types of devices. Absorption of sulfur oxides and nitrogen oxides are the most common devices where the combined action of particulate removal and absorption of gases takes place. It must be noted, however, that if an absorption process is going to be used to remove sulfur oxides it should not precede an electrostatic precipitator. Removal of the sulfur molecules before the precipitator will change the electrical properties of the gas and may result in loss of the ability to remove the particulate matter in the precipitator.

There continue to be many absorbers for the removal of water-soluble gases. Acid gases and some volatile organic compounds can be absorbed readily in water by the types of equipment previously discussed. These processes are essentially absorption with chemical reaction. For a discussion of absorption in air pollution control and a description of several absorption systems for sulfur dioxide and nitrogen oxide removal, see Schnelle and Brown.[5] A more detailed discussion of many more processes for flue gas desulfurization employing absorption is given by Lunt and Cunic.[6]

The United Kingdom has a long history of flu gas desulfurization. The world's first such system was installed in a power plant at Battersea, England, in 1936.[7] At Bankside a 228 MW boiler was the first installed. It was fitted with a once-through scrubber system using water from the condensers, which was dosed with alkali. A further unit was commissioned in 1949. Both units continued to operate until the early 1970s. At Fulham power station, a 120 MW boiler system was operated with recycled lime dosed scrubbing liquor from 1936 to 1940.

It should also be noted that absorption has been used to remove contaminants from natural gas streams during processing. In the early 1930s di-ethanolamine was used as an absorbent for both hydrogen sulfide and carbon dioxide.[8] This process became known as the ''Gerbitol Process.'' Other alkanolamines such as mono-ethanolamine and di-isopropanolamine have also found wide application.

Plate Towers

Originally, bubble cap plates had been used for absorption of pollutant gases such as sulfur dioxide. However, the solids in the slurries used as absorbents can more readily plug bubble caps. Typical absorbents used in current processes include, for example, conventional lime slurry; lime-limestone slurries; mixed sodium sulfite/lime slurries; and magnesium sulfite/bisulfite mixed with lime slurries. Conventional lime slurry towers may consist of a multilevel spray tower combined with a venturi scrubber. Venturi scrubbers will be discussed briefly below. Mixed sodium sulfite/lime slurries may be contacted in a plate tower. Sieve plates might be used with larger than normal holes to help prevent plugging due to the solids in the slurries.

Venturi Scrubbers

Venturi scrubbers are designed on the basis of the venturi flow-metering device. The flow channel is narrowed down so that the velocity will greatly increase at the throat. Then, as in the flow-metering device the flow channel widens out. For ease of fabrication venturi scrubbers are designed with a rectangular cross section. The absorption fluid is injected into the Venturi at the throat where the velocity is the greatest. For particulate removal plain water could be used. As in the plate columns discussed above, for the simultaneous removal of particulate matter and sulfur dioxide, soda ash or caustic soda slurries could be used for absorption of the gas. Venturis are frequently used in conjunction with plate towers. They also serve as stand-alone removal devices in some cases.

CONCLUSIONS—A COMPARISON OF PACKED COLUMNS AND PLATE COLUMNS

Absorption is a mass transfer operation that is commonly used to recover valuable gaseous components or to remove undesirable components of a gas stream. It is one of the main methods of separation in the chemical process industries. Absorption can take place in packed towers or in plate towers. Perry and Green (9) compare the advantages of packed and plate towers.[3]

Table 1 Economic factors in packed tower design operating and capital cost factors

Operating costs	Capital costs
Pumping power for gas and liquid	Tower and shell, packing or plates
Labor and maintenance	Packing support
Steam and cooling water	Gas and liquid distributor
Loss of unabsorbed material	Pumps, blowers, and compressors
Disposal of absorbed material	Piping and ducts
Solvent makeup	Heat exchangers
Solvent purification	Solvent recovery system

When column diameters are less than 0.6 m (2.0 ft) packed towers can be considerably cheaper. However, if alloy metals are necessary, plate towers may result in less cost. Using ceramic or other similar resistant materials for packing and materials of construction, packed towers can serve to handle corrosive materials and acids. Because the gas flow in packed towers may offer less degree of agitation, packed tower operation may be better for liquids that tend to foam. When liquids are thermally sensitive, packed columns may offer less holdup and thus prevent changes taking place in the liquids due to thermal reaction.

When solids are contained in liquids or when solids have the chance of condensing out of the gas steam, plate columns offer the advantage of being able to be designed to be more readily cleaned. Plate columns can more readily absorb thermal expansion, which might result in breakage of packing as it is inserted into the column or during the operation. Cooling coils can be more readily installed in plate towers than in packed towers. Flooding may occur with high liquid rates in packed columns, whereas a plate tower may be designed to handle the higher liquid rate. Low liquid flow rates may result in poor wetting of packing, thereby resulting in poor mass transfer. Thus, a plate tower may more readily handle lower liquid flow rates.

Economic Factors Affecting Packed Tower Construction

When a tower is designed for treating a given quantity of gas per hour, the height of the tower, especially in packed towers, is determined from mass transfer considerations. The diameter or cross-sectional area is determined by fluid dynamics from the gas velocity in the empty tower cross section in packed towers and by the velocity through the bubble caps or other openings in plate towers. The smaller the diameter of the tower, the higher the gas velocity that will help the gas to overcome the tower pressure drop. This could result in a lower cost of pushing the gas through the tower. The chief economic factors to be considered in tower design are listed in Table 1.

REFERENCES

1. Crocker, B.B. Capture of hazardous emissions. In *Control of Specific Toxic Pollutants*, Proceedings of the Conference, Air Pollution Control Association, Gainesville, FL, Feb 1979; Air and Waste Management Association: Pittsburgh, PA, 1979; 414–433.
2. Rees, R.L. The removal of oxides of sulfur from flue gases. J. Inst. Fuel **1953**, *25*, 350–357.
3. Perry, R.H.; Green, D.W. *Perry's Chemical Engineer's Handbook*; 7th Ed.; McGraw-Hill: New York, 1997; 14–24, 14–25, 14–39, 14–40.
4. Wankat, P.C. *Equlibrium Staged Separations*; Prentice Hall: Upper Saddle River, NJ, 1988; 369–379.
5. Schnelle, K.B., Jr.; Brown, C.A. *Air Pollution Control Technology Handbook*; CRC Press: Boca Raton, FL, 2002.
6. Lunt, R.R.; Cunic, J.D. *Profiles in Flue Gas Desulfurization*; American Institute of Chemical Engineers: New York, 2000.
7. http://www.dti.gov.uk/energy/coal/cfft/pub/cb013-a.pdf (accessed Dec 2004).
8. http://www.r-t-o-l.com/learning/studentsguide/sru.htm (accessed Dec 2004).

Activated Sludge Process

Shankha K. Banerji
Department of Civil and Environmental Engineering, University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

Wastewater treatment occurs in a treatment plant in several stages depending on the degree of treatment desired. In the first stage, the preliminary treatment processes prepare the influent wastewater for treatment in subsequent processes. Bar screens, grit chamber, and flow equalization tank are some of the processes included in the preliminary treatment. There is no significant removal of biodegradable organic matter expressed in terms of 5-day biochemical oxygen demand (BOD) or suspended solids by these processes. The next stage is the primary treatment process where settleable (and floatable) solids present in the wastewater are removed by gravity sedimentation. In some rare instances, the flotation process can be used instead of gravity sedimentation for the removal of settleable solids. The primary treatment process can remove up to 40% of the incoming BOD and 50–70% of the suspended solids.[1] The subsequent stage is the secondary treatment process, which is needed to remove the remaining soluble and colloidal organic matter from the wastewater that was not removed during the primary treatment processes. The secondary processes invariably use aerobic biological treatment processes to remove the soluble and colloidal organic matter from the wastewater. The biological treatment process converts the soluble and colloidal organic matter into settleable solids and micro-organisms (sludge), which are removed in the secondary settling tank leaving a clearer supernatant effluent for discharge. Thus, the settling tank following the aeration tank is an integral part of the process. In this entry, the secondary tank details are not included. These processes in combination with the primary process can remove 90+% BOD (carbonaceous BOD) and suspended solids. Thus, the secondary wastewater treatment processes can meet the current US Environmental Protection Agency mandated effluent requirements of 30 mg/L of BOD and 30 mg/L of suspended solids for municipal wastewater treatment.[1]

There are two types of secondary aerobic biological treatment processes: suspended growth processes and attached growth processes. In the suspended growth process, the micro-organisms responsible for the biochemical conversion of organic matter are kept in suspension by aeration or agitation in a tank where the wastewater is introduced. The micro-organisms assimilate the organic compounds for synthesis of new cells (biomass) and for respiration, which provides the energy for the synthesis and other cellular processes. Activated sludge process and its modifications are suspended growth processes. In the attached growth process, the micro-organisms are present in an attached form (biofilm) on a medium, either stone, treated wood, or synthetic plastic materials. The wastewater comes in contact with these attached micro-organisms, and the same biochemical processes as in the suspended growth process take place, namely, cell synthesis and respiration. Trickling filters and rotating biological contactors are the two most common attached growth secondary biological treatment processes used.[1]

In this article, activated sludge process and some of its modifications are discussed at some length. Only the details of carbonaceous BOD removal from wastewater are included.

CONVENTIONAL PROCESS

The conventional activated sludge process consists of an aeration tank followed by a settling tank, as shown in Fig. 1. The wastewater from the primary settling tank enters the aeration tank and mixes with the micro-organisms or biomass present. A portion of the settled sludge (biomass) in the secondary settling tank is recycled back to the head of the aeration tank. This recycled sludge is referred to as return activated sludge (RAS). The term sludge in the secondary settling tank refers to solids that have settled in the tank bottom because of gravity forces. The recycling of the sludge maintains a desired amount of biomass concentration in the aeration tank. The solids responsible for the bio-oxidation of organic matter consist of micro-organisms (biomass), biodegradable and nonbiodegradable organic matter, and some inert solids in the aeration tank known as mixed liquor suspended solids (MLSS). The mixture of wastewater and these solids is called mixed liquor. The organic components of the MLSS are known as mixed liquor volatile suspended

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007653

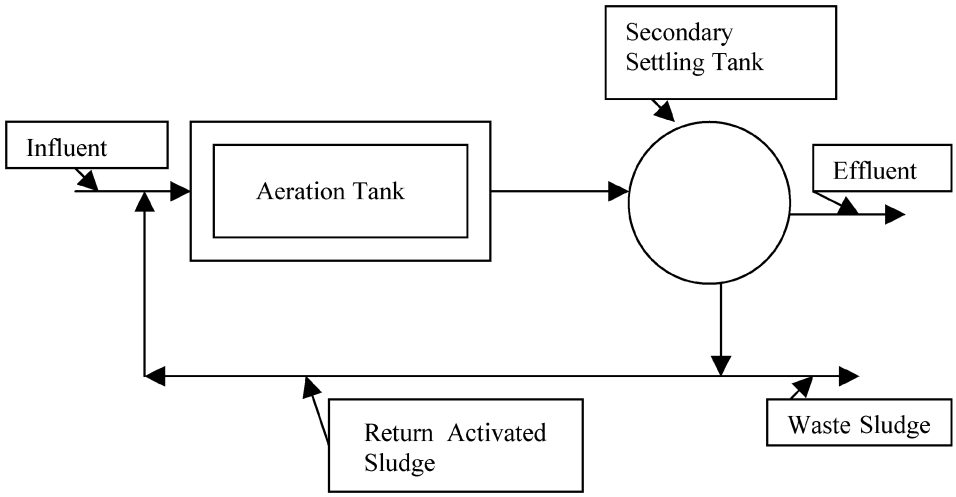


Fig. 1 Conventional activated sludge process.

solids (MLVSS). Mixed liquor volatile suspended solids are often considered to represent the active biomass in the system. The remaining settled sludge from the secondary settling tank is withdrawn as waste activated sludge (WAS) for further processing before disposal.

The biological oxidation process taking place in the aeration tank can be described by the following equation:

$$\text{Organic matter} + \text{N} + \text{P} + \text{O}_2 + \text{micro-organisms} = \text{CO}_2 + \text{H}_2\text{O} + \text{new micro-organisms}$$

In the above equation, N and P are the nitrogen and phosphorus compounds (sometimes called nutrients) that are needed for micro-organism metabolism and growth. In most instances, they are already in excess amounts in domestic wastewater, but for special cases they have to be supplied if not present in adequate quantities. The nutrient requirements in an aerobic biological treatment process are based on the BOD load imposed on the system. For every 100 kg of BOD introduced to the system, 5 kg of N and 1 kg of P should be available.[2] The oxygen needed has to be in a dissolved state to be available to the cells. The organic matter must be a biodegradable type that can be utilized by the micro-organisms present. The transport of the organic molecules inside the cells occurs through a microbial cell membrane having a pore size of the order of 5 Å, which means only small molecules that are soluble can be assimilated by the cell. Larger molecules are broken down to smaller sizes outside the cell through exogenous enzymes secreted by the microbes. The pH of the process should be within the range 6.5–8.5. Under some conditions, the ammonia present in the wastewater can be oxidized by the nitrifying bacteria to nitrite and nitrate molecules. This is known as nitrification process. If necessary, nitrification can be achieved in the process by appropriate sludge recycling and the organic loading rate of the process. The nitrification process is not discussed in this entry, but more information can be obtained in Ref.[1].

The activated sludge process was developed in England by Ardern and Lockett in 1914 based on experiments conducted at the Lawrence Experiment Station in Lawrence, MA, in the early 1900s.[2] Presently, it is one of the most common secondary treatment processes used throughout the world. The conventional process has been modified to improve its performance. These modifications are described later.

MICROBIOLOGY OF THE PROCESS

The micro-organisms present in the aeration tank of the activated sludge process are quite varied and depend on the type of wastewater being introduced and the environment of the aeration tank (e.g., temperature, pH, etc.). The predominant micro-organisms constantly change depending on the environmental conditions. The micro-organisms range from very small virus particles to much larger multicellular worms. But the predominant species are heterotrophic bacteria, with lesser amounts of autotrophic bacteria. Taxonomically, bacteria are prokaryotic protista, having mostly single-cell structure. The heterotrophic bacteria use organic compounds as their source of energy and cell carbon (electron donor), while autotrophic micro-organisms oxidize inorganic compounds for generating energy (electron donor) and use inorganic carbon (bicarbonates or CO_2) as a source of cell carbon. One of the prominent autotrophic bacteria present is the nitrifying bacteria, which transforms ammonia to nitrite and finally to nitrate.[1]

Usually, there are large numbers of bacterial species present, making the system a mixed culture. Because the wastewater contains many different types of organic compounds, the presence of these different micro-organisms with varied metabolic capabilities enhances the possibility of degradation of these compounds. Most of the bacteria present in the activated sludge process are aerobic, meaning that they use oxygen as the ultimate electron acceptor, which produces energy for their growth and other uses. There are some facultative anaerobic bacteria present as well, which can survive in the presence or absence of oxygen. In the absence of oxygen, they can use organic compounds as the ultimate electron acceptor to form reduced organic compounds.

Besides bacteria, other micro-organisms present are protozoa, fungi, and rotifers. Protozoa are eukaryotic protists. Most of the protozoa are unicellular organisms. Wastewater contains many different species of protozoa—flagellates, ciliates, amoebas, and rotifers. These organisms are predators for the bacteria and may help in flocculation and clarification in the secondary settling tank. Fungi are common in the activated sludge process operating at lower pH values. They include organisms such as yeasts and molds. They are basically saprophytic organisms feeding on organic matter. Their numbers are smaller than other species in activated sludge. In addition to these organisms, activated sludge may also contain nematodes (roundworms) and other worms. They play no part in the wastewater treatment process.

The activated sludge occurs in the aeration tank in the form of flocs. These flocs are made up of micro-organisms, inorganic and organic colloidal, and particulate matter. They are bound together in an organic matrix. Their size may vary from 50 to 1000 μm.[3] The shape depends on the materials encased, the organisms present, and the mechanical forces applied inside the tank. The floc containing mixed liquor leaves the aeration tank and normally settles out in a compact form in the secondary settling tank within 2–4 hr. Sometimes the settling of the sludge in the secondary settling tank is disrupted. The sludge is said to be a ''bulking sludge.'' One of the proposed theories on bulking suggests that the preponderance of filamentous bacteria can cause poor settling sludge. It has been suggested that in a good settling sludge floc, the floc forming bacteria and filamentous bacteria are present in balanced numbers to give a compact sludge mass. In the floc, the filamentous bacteria form a backbone that provides its structure and strength, while the floc forming bacteria grow around the filamentous types. The gelatinous matrices of the floc forming bacteria, sometimes known as *Zoogloea* bacteria, entrap other micro-organisms, colloidal and particulate matter to give the floc its shape. The preponderance of filamentous bacteria over floc forming bacteria could cause an unbalanced situation and hence poor settling of the resulting floc. The filamentous bacteria found in activated sludge are *Thiotrix* spp., *Nocardia* spp., *Sphaerotilus natans*, *Beggiatoa*.[4]

KINETICS AND DESIGN EQUATIONS FOR CONVENTIONAL PROCESSES

The design of conventional biological wastewater treatment processes depends on the reaction rates of the metabolism of organic matter by the micro-organisms present. They use a part of the substrate (organic matter) for cell growth and the balance to produce energy to satisfy the cell needs. The reactor (tank) hydraulics is also an important factor in the design of the process. There are two idealized flow patterns that are considered in a suspended growth reactor—completely mixed and plug-flow conditions. In the completely mixed reactor, the influent flow is instantaneously mixed with the reactor content such that the concentration of the organic matter in terms of (BOD or COD) is the same throughout the reactor and in the effluent. In the plug-flow reactor, there is no such mixing in the longitudinal direction, but right angles to the flow there is complete mixing in the reactor section. Thus, there is a substrate concentration gradient longitudinally along the reactor as the wastewater organic matter is metabolized by the micro-organisms. From reactor engineering, it can be shown that for the same influent substrate concentration, reaction rate, and set removal efficiency, plug-flow reactor will have a lower volume than a completely mixed reactor. Thus, for most municipal wastewater treatment applications, where the wastewater does not contain toxic ingredients, plug-flow reactors have been most commonly used.

However, there are instances where a completely mixed reactor may be advantageous. In situations where periodically hazardous or toxic wastes are present in the influent wastewater, the entry of such a waste to the reactor causes an immediate reduction of concentration of the toxic component because of dilution with the entire tank content. This may reduce the adverse impact of the toxic component to the micro-organisms present with no significant impact on their waste treatment ability. If such a waste were introduced to a plug-flow reactor, it would cause an immediate toxic effect on the micro-organisms at the head end of the plant causing progressive process failure. In addition, the oxygen uptake rates throughout the completely mixed reactor are the same, which makes the design of the aeration process simpler. In a plug-flow reactor, the oxygen uptake rate is higher at the head end and decreases as the wastewater proceeds

down the reactor length. This may cause an unbalanced oxygenation system in the tank, especially if the aeration devices are equally spaced throughout the tank length. Higher influent substrate concentration at the head end of the plug-flow reactor also favors the floc forming bacteria over the filamentous type in the reactor, which helps in the settling of these micro-organisms later in the settling tank.[4]

Completely Mixed Reactor with Recycle

Using a mass balance on biomass X and substrate S around the reactor in Fig. 2 under steady state conditions ($dX/dt = 0$ and $dS/dt = 0$), one can arrive at the following equation:

$$\mu = \frac{(Q - Q_w)X_e + Q_w X_r}{VX} \quad (1)$$

where

μ is the specific growth rate of cell mass (time^{-1}) $= \frac{dX/dt}{X}$;

V is the aeration tank (reactor) volume (m^3);

Q is the wastewater influent flow rate (m^3/day);

Q_w is the sludge waste rate from the settling tank bottom (m^3/day);

X is the biomass concentration in the aeration tank (g VSS/m^3);

X_e is the concentration of biomass in the effluent (g VSS/m^3); and

X_r is the concentration of biomass in the return sludge (g VSS/m^3).

It should be noted that substrate S is expressed as BOD or soluble COD (sCOD). The sCOD is obtained by using a wastewater sample that has been filtered through a 0.45 μm membrane filter.

The inverse of the term on the left-hand side of Eq. (1) is known as solid retention time (SRT) or sludge age, θ_c:

$$\theta_c = \frac{VX}{(Q - Q_w)X_e + Q_w X_r} \quad (2)$$

The specific growth rate constant μ is affected by the substrate concentration S. The relationship between S and μ is given by the following equation named after Monod:[1]

$$\mu = \frac{\mu_m S}{K_s + S} \quad (3)$$

where

μ_m is the maximum specific growth rate coefficient (time^{-1});

S is the substrate concentration (BOD or sCOD) in the reactor (g/m^3);

and

K_s is the half-velocity constant (g/m^3).

K_s represents the substrate concentration at half the maximum specific substrate utilization rate, μ_m. K_s is a measure of the affinity of the micro-organism to the substrate. The lower the K_s value, the greater is the affinity of the organism to the substrate. When two organisms are competing for the same substrate in a limiting substrate condition, the organism with lower K_s value will have more success in growing. It should be recognized that with a mixed culture containing mixed substrate as in typical wastewater, the Monod

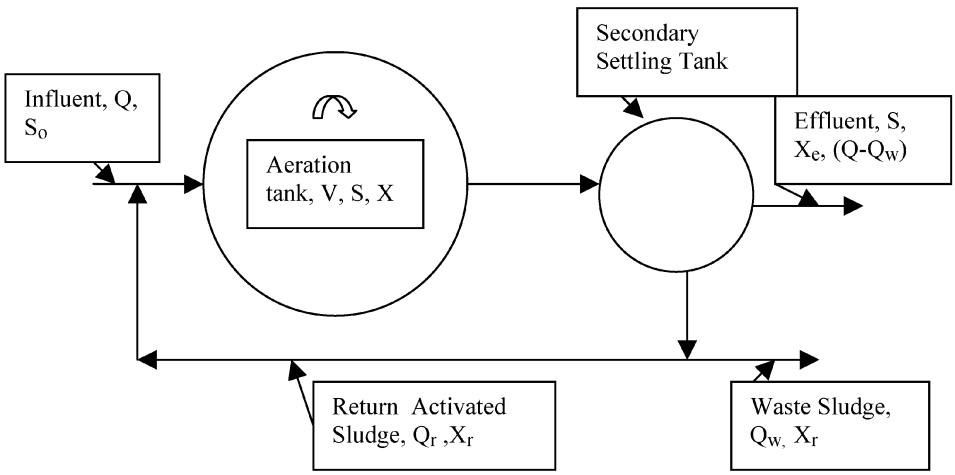


Fig. 2 Completely mixed activated sludge process schematic diagram. (*Note*: The aeration tank is a completely mixed reactor.)

equation described here gives only an approximation of the process kinetics.

Another term used is the food to micro-organism ratio (F/M), which is given by the equation:

$$\frac{F}{M} = \frac{QS_o}{VX} \tag{4}$$

where

F/M is the food to micro-organism (biomass) ratio (g BOD/g MLVSS day); and

S_o is the influent BOD or biodegradable sCOD (g/m^3).

F/M ratio is related to the term U, the specific substrate utilization rate as follows:

$$U = \frac{(F/M)E}{100} \tag{5}$$

E is the % BOD or sCOD removal efficiency in the process

$$E = \left(\frac{S_o - S}{S_o}\right)100 \tag{6}$$

By substituting the expressions for F/M and E, Eq. (5) can be rearranged as:

$$U = \frac{(S_o - S)Q}{VX} \tag{7}$$

U is also given by the equation:

$$U = \left(\frac{dS}{dt}\right)\Big/X = \frac{kS}{K_s + S} \tag{8}$$

because

$$\frac{dS}{dt} = \frac{kSX}{K_s + S}$$

where k is the maximum substrate utilization rate (g substrate (BOD)/g biomass day).

The amount of cell synthesis or growth can be related to the substrate removed by the micro-organisms. Cell yield Y is given by (g biomass produced/g substrate utilized).

$$Y = \frac{dX/dt}{dS/dt} = \frac{\mu_m}{k} \tag{9}$$

where Y is the yield coefficient (g VSS/g BOD or sCOD).

The observed yield coefficient, especially at low growth rates, is less than that given by Eq. (9). This is because of the need for maintenance energy by the cells for nongrowth needs. Eq. (10) provides an expression for observed yield coefficient, Y_{obs}:

$$Y_{obs} = \frac{Y}{1 + k_d\theta_c} \tag{10}$$

where

k_d is the endogenous respiration coefficient (g VSS/g VSS day); and

θ_c is the solid retention time, or sludge age ($time^{-1}$).

The reactor biomass (MLVSS) X is given by the following equation:

$$X = \frac{\theta_c Y(S_o - S)Q}{V(1 + k_d\theta_c)} = \frac{\theta_c Y(S_o - S)}{t(1 + k_d\theta_c)} \tag{11}$$

where $V/Q = t$ is the hydraulic retention time (hr).

The relationship between θ_c and U is shown in Eq. (12):

$$\frac{1}{\theta_c} = YU - k_d \tag{12}$$

If the process θ_c has been selected, then U will have a fixed value as Y and k_d are constants.

Substituting U from Eq. (8) we get:

$$\frac{1}{\theta_c} = \frac{YkS}{K_s + S} - k_d \tag{13}$$

The reactor substrate concentration (also the effluent substrate concentration) S is given by the equation:

$$S = \frac{K_s(1 + k_d\theta_c)}{\theta_c(Yk - k_d) - 1} \tag{14}$$

The production of excess biomass or waste activated sludge per day can be calculated from the following equation:

$$P_x = Q\left[\frac{Y(S_o - S)}{1 + k_d\theta_c} + X_{inv} + X_{ivn}\right] \tag{15}$$

This is the amount of excess sludge that is formed by the conversion of soluble and colloidal organic matter to settleable sludge (biomass) in the aeration tank. This sludge has to be properly treated and disposed of. Typical sludge treatment may consist of thickening,

stabilization (anaerobic digestion), followed by land application.[1]

In Eq. (15) the last two terms, X_{inv} and X_{ivn}, are the nonvolatile (inorganic) suspended solids and influent volatile nonbiodegradable solids, respectively, entering the secondary process. These solids pass through the process unchanged.

The amount of RAS from the secondary settling tank can be obtained from the following equation:

$$\frac{1}{\theta_c} = \frac{Q}{V}\left(1 + R - R\frac{X_r}{X}\right) \tag{16}$$

where

R = sludge recycle ratio = Q_r/Q;

Q_r is the rate of settled sludge recycle (m^3/day);

X_r is the concentration of settled sludge in secondary settling tank (g/m^3) = concentration of recycled sludge (g/m^3).

The poor settling of the sludge in the secondary settling tank causes X_r to be lower, which requires a higher sludge recycle ratio to maintain the same mixed liquor concentration in the aeration tank.

The oxygen required for the biooxidation of organic matter can be estimated from Eq. (17):

$$R_o = Q(S_o - S) - 1.42P_x \tag{17}$$

where R_o is the mass of oxygen required for oxidation of organic matter per day (kg/day).

The factor 1.42 represents the oxygen equivalent of cell mass produced. For only BOD removal, the oxygen requirements can vary from 0.8 to 1.3 kg O_2/kg BOD removed for most conventional activated sludge processes.

Process stability depends on the process solid retention time (θ_c). For a particular influent wastewater quality, reactor configuration, and a given microorganism community in the reactor, Y, K_s, k_d, and k are relatively constant. For a selected process U value, the θ_c becomes fixed. If the operating θ_c is below a specific value, the process becomes unstable and biomass washout may occur. This minimum θ_c value (θ_{c_m}) can be determined from Eq. (13) by putting $S = S_o$, the influent substrate concentration indicating no wastewater treatment:

$$\frac{1}{\theta_{c_m}} = \frac{YkS_o}{K_s + S_o} - k_d \tag{18}$$

In most situations, K_s is much smaller than S_o and can be ignored. Hence, Eq. (18) becomes:

$$\frac{1}{\theta_{c_m}} \approx Yk - k_d = \mu_m - k_d \tag{19}$$

Activated sludge treatment processes should not be designed with θ_c less than θ_{c_m}.

The details about the derivation of the equations in the preceding section can be found in several studies, notably Metcalf & Eddy.[1]

Plug-Flow Reactor with Recycle: The flow schematic of a plug-flow reactor is shown in Fig. 3. The modeling of a plug-flow reactor is mathematically more challenging than the completely mixed activated sludge process. By assuming biomass change to be negligible compared to the total amount present (i.e., MLVSS in the tank is constant), the integration of the substrate mass balance equation yields:[5]

$$(S_i - S) + K_s \ln\left(\frac{S_i}{S}\right) = \frac{\mu_m X_{av} V}{Y(1 + R)} \tag{20}$$

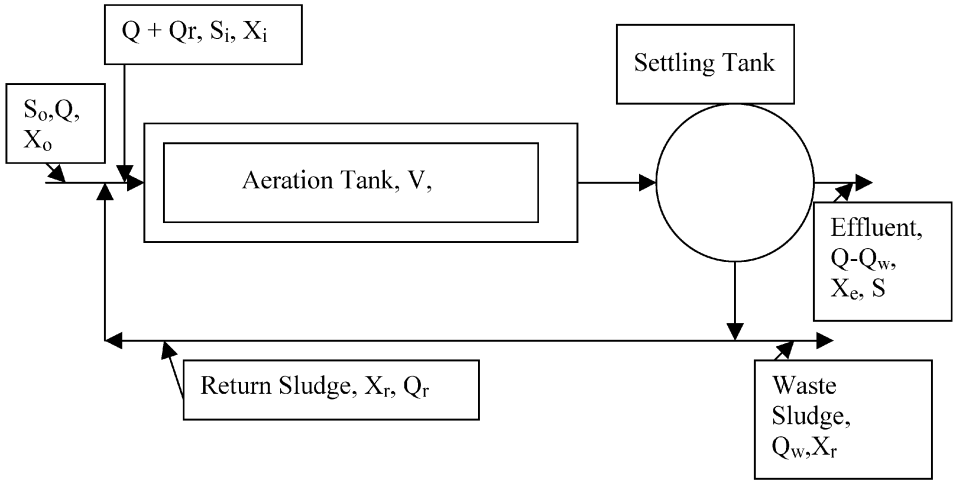


Fig. 3 Plug-flow activated sludge process schematic. (*Note*: The aeration tank is a plug-flow reactor with an average biomass concentration of X_{av} and effluent substrate concentration of S.)

where S_i is the concentration of substrate in the aeration tank after mixing with the recycle sludge flow,

$$S_i = \frac{S_o + RS}{1 + R} \tag{21}$$

where

X_{av} is the average biomass concentration in the tank (g/m^3); and

S is the concentration of substrate in the effluent from the aeration tank (g/m^3).

The small change in biomass is given by:

$$X - X_i = Y(S_i - S) - \frac{k_d X_{av} V}{Q(1 + R)} \tag{22}$$

where X_i is the concentration of biomass in the aeration tank after mixing with the recycle sludge flow,

$$X_i = \frac{X_o + RX_r}{1 + R} \tag{23}$$

where X_o is the concentration of biomass in the influent flow (g MLVSS/m^3) = usually 0. Eq. (23) becomes:

$$X_i = \frac{RX_r}{1 + R} \tag{24}$$

θ_c can be approximated by the following equation:

$$\frac{1}{\theta_c} = \frac{Yk(S_o - S)}{(S_o - S) + (1 + R)K_s \ln(S_i/S)} - k_d \tag{25}$$

In actual practice, a truly plug-flow or completely mixed-flow regime in a reactor is not attained because of longitudinal dispersion and nonideal mixing conditions. The equations reported here approximate the actual conditions in the field.

AERATION AND MIXING REQUIREMENTS

The aeration is a necessary part of the activated sludge treatment process as it supplies the dissolved oxygen (DO) needed for the biooxidation of the organic matter. If a liquid is unsaturated with respect to DO concentration, the natural diffusion process would transport enough oxygen from the air to bring the liquid to saturation value. This transport is based on Fick's law of diffusion.

The oxygen transfer to the liquid phase is best described by the two-film theory. According to this theory, oxygen mass to the liquid is transported through air and water films at the interface because of the concentration gradients. For gases with low solubility such as oxygen, the diffusion through the liquid film is the rate limiting step.[5]

Fig. 4 shows the oxygen concentration gradients at the water interface during mass transport operation. The mass transport process can be expressed by Fick's equation:

$$\frac{dM}{dt} = -D_l A \frac{dC}{dy_f} \tag{26}$$

where

M is the mass of oxygen transported (g);

D_l is the diffusion coefficient for oxygen in water (m^2/time);

C is the dissolved oxygen concentration (g/m^3);

A is the cross-sectional area through which O_2 transport occurs (m^2);

and

y_f is the liquid film thickness (m).

As the liquid film thickness is quite small, the differential quantity dC/dy_f can be replaced by linear approximation of the concentration gradient as:

$$\frac{dC}{dy_f} \approx \frac{C_s - C}{y_f} \tag{27}$$

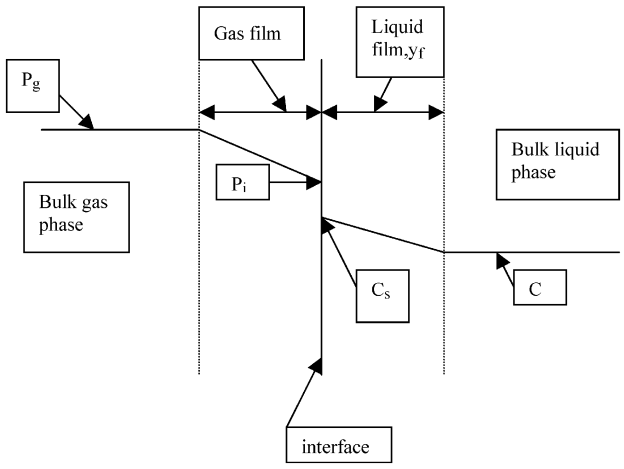


Fig. 4 Two-film oxygen mass transfer theory at liquid interface. (*Note*: P_g = partial pressure of oxygen in bulk gas phase; P_i = partial pressure of oxygen at the interface; C_s = saturation concentration of dissolved oxygen at the interface; C = dissolved oxygen concentration in bulk liquid.)

where

C_s is the saturation dissolved oxygen concentration at the interface layer (g/m^3); and

C is the dissolved oxygen concentration in the bulk liquid (g/m^3).

Dividing both sides of the Eq. (26) by V, the volume of the liquid in the container, and substituting the dC/dy_f from Eq. (27), we get:

$$\frac{dM}{V\,dt} = \frac{dC}{dt} = -D_l A \frac{C_s - C}{V y_f} \tag{28}$$

As y_f cannot be measured easily, it is combined with D_l to form another constant:

$$K_1 = \frac{D_l}{y_f}\ (\text{m/time})$$

Eq. (28) becomes:

$$\frac{dC}{dt} = -K_1 \frac{A}{V}(C_s - C) \tag{29}$$

The area of the bubbles through which the oxygen mass transfer takes place is quite difficult to measure. Hence, the parameters A/V are combined and represented by a, the specific surface area. Eq. (28) is simplified as:

$$\frac{dC}{dt} = K_l a(C_s - C) \tag{30}$$

where $K_l a$ is the overall mass transfer coefficient (time^{-1}).

Under turbulent conditions in the bulk liquid, Eq. (30) can be integrated with boundary conditions: at time $= t_0$, $C = C_o$; at time $= t$, $C = C$:

$$K_1 a = \frac{\ln[(C_s - C_o)/(C_s - C)]}{t - t_0} \tag{31}$$

The coefficient $K_l a$ is a good measure of the efficiency of an aerator. It depends on many factors, such as temperature, wastewater composition, tank geometry, and turbulence.

Measurement of $K_l a$ of aeration devices in clean water can be determined by standard procedures developed by the American Society of Civil Engineers.[6] In the standard procedure, the clean water (usually tap water) in a tank is deoxygenated by adding a reducing agent, sodium sulfite, and a catalyst, cobalt chloride. Thereafter, aeration is resumed with DO being measured periodically at several points in the tank to measure the rate of aeration. The $K_l a$ can be determined from the slope of a plot of $\ln(C_s - C)$ vs. time. The value of C_s, the DO saturation, for the temperature and dissolved solid concentrations of the test water can be obtained from several studies including Ref.[1]. For extrapolation of the clean water $K_l a$ to field conditions with wastewater, some corrections must be applied. Wastewater constituents (especially surface active agents) can reduce the field $K_l a$ value. A factor α can be used to make a correction to the measured clean water $K_l a$:

$$\alpha = \frac{K_1 a_{\text{wastewater}}}{K_1 a_{\text{cleanwater}}} \tag{32}$$

Another factor β corrects for the difference in DO saturation value of clean water and wastewater:

$$\beta = \frac{C_{s_{\text{wastewater}}}}{C_{s_{\text{cleanwater}}}} \tag{33}$$

Eq. (34) shows a temperature correction factor for $K_l a$

$$K_l a_{T^0} = K_l a_{20^0} \theta^{(T-20)} \tag{34}$$

where

θ is the temperature correction factor for reaeration, with a usual value of 1.024.

$K_l a_{T^\circ}$ is the overall mass transfer coefficient at T°C (time^{-1}); and

$K_l a_{20^\circ}$ is the overall mass transfer coefficient at 20°C (time^{-1}).

The standard oxygen transfer rate (SOTR) in clean water can be calculated by knowing the average $K_l a$ in the tank at 20°C and at zero DO concentrations:

$$\text{SOTR} = V(K_1 a_{\text{av}} C_{\text{sav}})(\text{kg } O_2/\text{hr}) \tag{35}$$

where

V is the volume of the aeration tank;

$K_l a_{\text{av}}$ is the average $K_l a$ value in the tank (time^{-1}); and

C_{sav} is the average DO saturation value in the tank (g/m^3);

The field oxygen transfer rate (OTR) in wastewater can be estimated from the SOTR by applying the

appropriate correction factors:

$$\mathrm{OTR} = \alpha\left(\frac{\beta C_s - C}{C_{s,20}}\right)1.024^{(T-20)}\mathrm{SOTR} \quad (36)$$

where

C_s is the DO saturation in clean water in the aeration tank at the prevailing pressure and temperature (g/m^3);

C is the operating DO concentration (g/m^3);

$C_{s,20}$ is the DO saturation in clean water at 20°C (g/m^3).

Details about different types of aeration system for the activated sludge process can be found in Metcalf and Eddy.[1]

Mixing of the mixed liquor is also accomplished by the aeration system. The mixed liquor solids must be kept in suspension for proper operation of the process. The mixing requirement depends on types of aeration equipment, depth, and width of the tank. Typically, a mixed liquor velocity of 0.15 m/sec in the aeration tank provides adequate mixing. For mechanical aerators, power input of 16–30 W/m^3 is often specified for mixing of biomass in the tank.[2]

DESIGN PARAMETERS

The design of an activated sludge system requires the determination of the following items: aeration tank volume V, oxygen requirement, daily sludge wasting rate, and sludge recycle rate R.

The selection of the process θ_c value usually depends on past experience rather than kinetic considerations, because use of kinetic equations gives a much smaller θ_c value.[2] The θ_c is selected to provide mixed liquor that settles well in the settling tank. For a conventional activated sludge process, it varies from 5 to 15 days[7] (see Table 1). Once θ_c has been selected, the tank volume can be calculated from Eq. (11) for a completely mixed reactor or from Eq. (21) for a plug-flow reactor. The constants Y, k_d, k, and K_s must be estimated for calculating the volume V, and X must also be selected. Typically, X in the aeration tank varies from 1500 to 3000 mg MLVSS/L. The selection of the MLVSS in the aeration tank depends on the influent BOD load and desired U or θ_c values. The oxygen requirement for carbonaceous BOD removal can be estimated from Eq. (17). The aeration equipment can be selected to supply the needed oxygen based on the manufacturer's information. The sludge mass wasted per day can be calculated from Eq. (15). The volume of the wasted sludge will depend on the settled sludge concentration and specific gravity of the settled solids. The sludge recycle rate R depends on the settled sludge concentration as shown in Eq. (16). If the settled sludge concentration is low, more sludge has to be recycled, making R greater.

Conventional activated sludge process loading rate in terms of F/M ratio (mass loading; g BOD/g MLVSS day) varies from 0.2 to 0.6. In some instances, the loading rate is based on the tank volume. The volumetric loading rate in terms of BOD applied per unit tank volume varies from 0.3 to 0.6 kg BOD/m^3 day.[7] The hydraulic residence time (V/Q) varies from 4 to 8 hr.

PROCESS MODIFICATIONS

Over the years, the conventional activated sludge process has been modified to improve or to suit a specific operational condition. Some of the more common modifications are extended aeration process, contact

Table 1 Design and operating parameters for activated sludge process and its modifications

	Hydraulic retention time, t (hr)	F/M (kg BOD/kg VSS day)	Solid retention time, θ_c (day)	Volumetric loading rate (kg BOD/m^3 day)	MLSS (mg/L)
Conventional plug-flow	4–8	0.2–0.4	5–15	0.3–0.6	1500–3000
Extended aeration	18–36	0.05–0.15	20–30	0.1–0.4	1500–5000
Contact stabilization	0.5–1.0[a]	0.2–0.6	5–15	1.0–1.2	1000–3000[a]
	2–4[b]	—	—	—	4000–9000[b]
Step aeration	3–5	0.2–0.4	5–15	0.6–1.0	2000–3500
Tapered aeration	4–8	0.2–0.4	5–15	0.3–0.6	1500–3000
High rate	2–4	0.4–1.5	5–10	1.6–16	3000–6000
Pure oxygen	1–3	0.25–1.0	3–10	1.6–3.3	3000–8000

[a]In contact tank.
[b]In stabilization tank.
(From Ref.[7].)

stabilization process, step aeration, tapered aeration, high rate, and pure oxygen process.

Extended Aeration Process: This process is very much like the conventional process but the loading rates are much lower. The sludge age is high and the hydraulic residence time is higher than the conventional process. These conditions result in a better effluent quality, with some nitrification of the wastewater. In addition, the process is quite stable under varying loading rates, producing good settling sludge. The design parameters for this process are given in Table 1. The high hydraulic residence time increases the capital and operating cost for the process. A modified configuration of this process is known as oxidation ditch, where the mixed liquor is moved around a racetrack style reactor. The movement of the wastewater and aeration along the ditch is facilitated by a brush type or vertical rotor aerator, which ensures a fluid velocity of about 0.3 m/sec so that the mixed liquor solids are not settled out in the channels.

Contact Stabilization Process: In this modification of the conventional process, the waste is contacted in the aeration tank with the return sludge (which has been previously aerated) for a relatively short time, about 30–60 min. The mixed liquor is then separated in a settling tank, with the settled sludge being aerated in a separate tank for 3–6 hr before it is returned to the aeration tank (see Fig. 5). The short contact time allows the colloidal and particulate waste constituents to adsorb onto the micro-organism flocs. The stabilization of the adsorbed organic matter occurs when the sludge is reaerated. By following this scheme, the overall tank volume requirement reduces by about 50%.[5] The process is only successful where a large fraction of the influent BOD is in colloidal or particulate form, but for normal domestic wastewater it does not provide equivalent secondary treatment effluent quality.

Step Aeration: In this process, the influent feed is introduced into the plug-flow reactor at two or more points, which distributes the organic load along the length of the tank (Fig. 6). Thus, the oxygen uptake rate along the tank length becomes more uniform rather than high at the start and low at the end as in the conventional system. This gives a better performance during most operating conditions. All other design parameters are the same as in the conventional system as seen in Table 1.

Tapered Aeration: This process corrects the problem of unbalanced aeration supply in a plug-flow conventional activated sludge system by providing more diffusers at the head end, which decrease progressively along the length of the tank as the BOD concentration decreases. The loading rates are the same as in a conventional system (Table 1). This arrangement reduces blower capacity and operating costs, and provides a greater degree of operational flexibility.[8] Fig. 7 shows a schematic of the process.

High Rate: This process is characterized by a shorter hydraulic retention time of the mixed liquor in the aeration tank and a higher loading rate than the conventional process. Consequently, the effluent leaving the process is not as high a quality as in the conventional process, i.e., it has a higher BOD and suspended solids. It often precedes a second-stage nitrification process.

Pure Oxygen Process: This process uses compressed pure oxygen instead of air, resulting in increased DO in the mixed liquor. Advantages include reduced power for oxygen diffusion, faster rate of organic matter stabilization, better settling sludge, and the ability to treat higher BOD wastewater. The process uses covered, completely mixed tanks in three or four stages with oxygen gas and wastewater entering at the head end. Each stage is mixed with a surface aerator. The exiting offgas contains only about 10% oxygen as the rest is used up in the biochemical reactions inside the reactors. In recent practice, the MLSS concentrations in the tanks vary from 1000 to

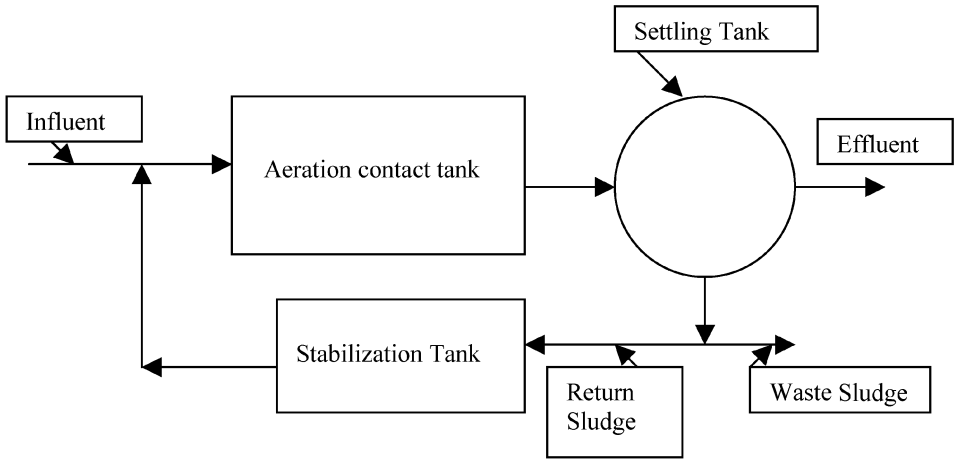


Fig. 5 Contact stabilization process.

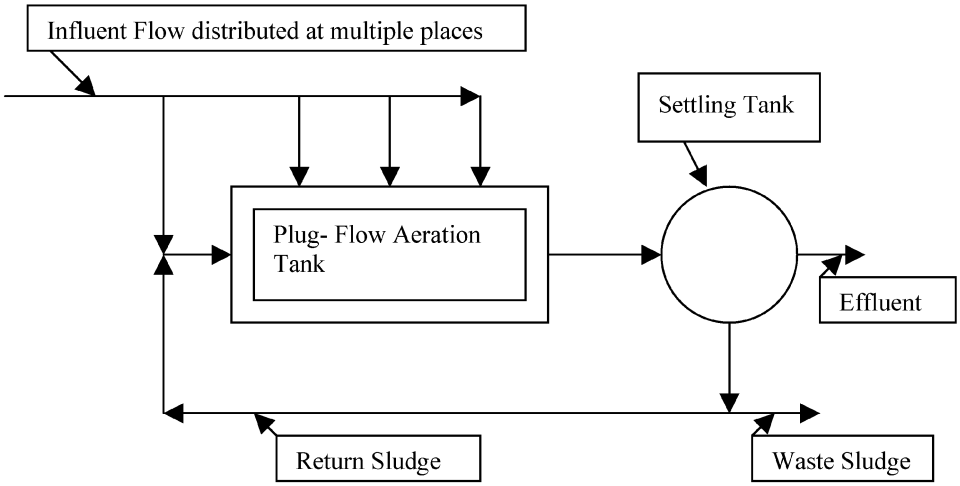


Fig. 6 Step aeration process.

3000 mg/L with DO levels typically in the range of 4–10 mg/L.[9] The hydraulic retention in the aeration tank is usually 1–3 hr. The loading rate for this process is higher than the conventional process, which requires less aeration tank volume than the conventional process for comparable wastewater quality. The oxygen needed for the process has to be produced in situ, which increases the construction cost for the process. The common methods for producing oxygen gas are cryogenic process, pressure swing adsorption, and vacuum swing adsorption. Details of these processes can be found in Ref.[9]. There have been debates whether these processes have significant advantages over the conventional process. The consensus is that the advantages, if any, are marginal at best.[5] Nitrification of the wastewater may occur as a result of higher DO levels. The process does create foaming because of Nocardia accumulation in the aeration tank. It can also depress the pH of the mixed liquor as the CO_2 formed is not stripped by nitrogen gas as with the air diffusion system. The construction materials for the tank and accessories have to be selected properly as the atmosphere inside is more corrosive.[2]

OPERATIONAL PROBLEMS

The activated sludge plant often has operational problems, which could be attributable to unusual characteristics of the influent wastewater or could be because of improper design and operation. These problems can be characterized by two factors: low soluble BOD removal and poor settling of solids in the secondary settling tank.

The low BOD removal could be caused by many reasons such as higher influent organic loading, influx of toxic or inhibitory chemicals, change of pH in the aeration tank beyond the acceptable range 6.5–8.5, insufficient aeration, and insufficient biomass in the aeration tank.

An increased F/M ratio beyond the process design value caused by higher influent BOD concentration

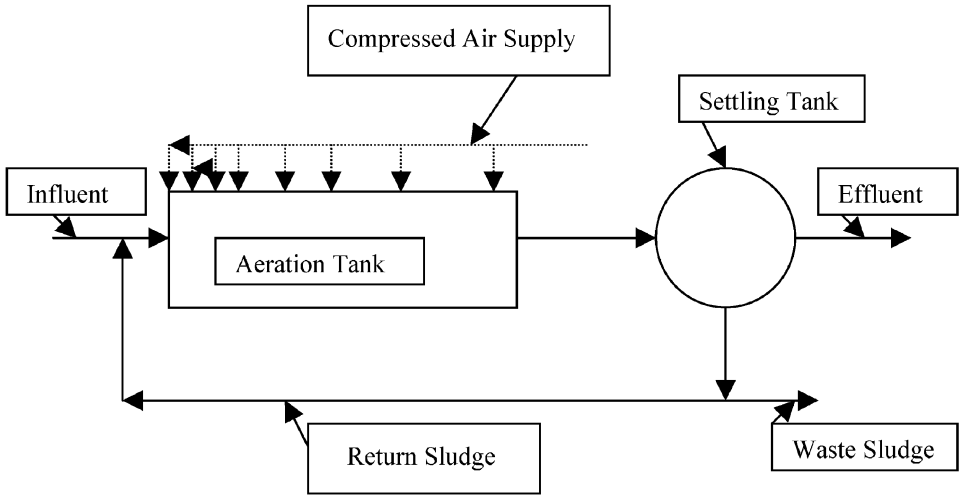


Fig. 7 Tapered aeration process.

could eventually cause low BOD removal. Increased sludge recycle rate will increase process micro-organism concentration and balance the F/M ratio, which will correct the poor performance results. Variable and shock influent organic loading rates could also cause poor BOD removals. An equalizing tank at the head end could minimize the effects of shock or variable BOD influent conditions. Equalizing tank equalizes the diurnal flow variations by providing storage for the excess flow during peak flow hours and a constant withdrawal throughout the day.

Toxic and inhibitory chemicals come as a result of a spill or temporary problems at chemical or other industrial plants discharging effluents into the sewer system. Sewer use ordinance prohibits the discharge of these materials; but in case of accidents these chemicals can inhibit the activities of the micro-organisms in the process. If detected early the wastewater containing the toxic wastes can be diverted to a holding tank, if available, and the unaffected micro-organisms can be nurtured under proper environment till nontoxic wastewater flow is resumed. In some cases where most of the micro-organisms have been destroyed, later external seed sludge may need to be imported to make the process restart. Accidental discharge of acidic or basic wastes from industries may change the wastewater pH beyond the acceptable range of 6.5–8.5. Prolonged operation of the process at these pH values may affect the performance of the plant. Neutralizing chemical may be needed to correct the pH to a safer range.

In a conventional activated sludge plug-flow plant with diffused aeration, the inlet end has a higher oxygen uptake rate than at the end. In cases where the influent BOD load exceeds the oxygen supply, oxygen deficiency may occur and the effluent performance could suffer. The effluent could be high in suspended solids as well as BOD. This type of condition can be remedied by changing the aeration system to the tapered aeration process mentioned earlier or changing the feed introduction by following the step aeration process. In addition, if the air blower capacity is limited, it can be replaced by a larger unit.

Insufficient biomass in the aeration tank could occur if the return sludge pumps do not have enough capacity to supply the increased sludge flow needed at higher influent BOD loading conditions. It could also occur if the settled sludge has a low concentration because of poor settling properties of the sludge, which would require a much higher sludge return flow rate. Addition of larger sludge return pump could help this situation to some extent but may not be able to cope with poor settling sludge conditions.

The problem of poorly settling sludge in the secondary settling tank could be caused by two separate conditions. The first, known as ''bulking sludge,'' is the most common problem of the activated sludge plants. The sludge does not settle well in the tank, giving a low solids concentration to the settled sludge. This requires a much larger amount of the settled sludge recycle rate to maintain a set MLSS level in the aeration tank. In some extreme cases where the sludge is too bulky, no amount of sludge recycling can maintain the process performance. The escape of unsettled sludge through the settling tank weirs can cause higher than the required BOD and suspended solids concentrations. Common factors that cause bulking are nutrient (N and P) deficiency, addition of septage to the influent, insufficient aeration capacity, low pH conditions, and influx of toxic wastes. The nutrient deficiencies of the influent wastewater can be fixed by adding appropriate amounts of nitrogen and phosphorus compounds. A preaeration system included for the septage before it enters the aeration tank could reduce its adverse impacts. Adding additional aeration capacity or modifying the aeration system to tapered aeration could reduce the problems of bulking sludge caused by insufficient aeration capacity. Low pH could occur from the entry of some industrial acidic wastes with the influent wastewater. This can cause the growth of fungi, which are filamentous, and bulking. Proper pH control can overcome this problem. Inadvertent influx of toxic wastes to the plant could also cause a change in the biota of the process and eventually cause bulking. The control of entry of such wastes can correct this problem. In some situations where the bulking is caused by the growth of filamentous organisms, the addition of an oxidizing agent such as chlorine or hydrogen peroxide can selectively reduce their numbers and solve the problem. Chlorine doses from 0.1 to 2.5 g Cl_2/kg of returned sludge dry mass have been successful in controlling sludge bulking.[8]

The other sludge settling problem is called ''rising sludge.'' In this case, settled sludge flocs tend to float up to the top giving it the name. This occurs in situations where the wastewater is nitrified to a great extent with nitrate-N present in the liquid phase. The environment at the bottom of the settling tank is suitable for denitrification of the nitrate molecules, i.e., anoxic conditions with organic carbon available from sludge deposits. Under these conditions, nitrogen gas bubbles are formed from nitrate molecules that attach onto the sludge flocs to float them to the top. By reducing the length of time the settled sludge stays at the settling tank bottom and by increasing the sludge recycle rate to the aeration tank, denitrification can be reduced in the settling tank.[8] In addition, if nitrification is not needed, increased organic loading rate to the aeration tank can reduce nitrification and remove the rising sludge problem.

CONCLUSIONS

Activated sludge process is one of the most common secondary treatment processes available for treating wastewater. It depends on the microbial metabolism of soluble and colloidal organic matter in the presence of dissolved oxygen and nutrients. The resulting biomass from the process is subsequently settled in a settling tank and a portion of the settled biomass (sludge) is recycled back to the aeration tank. Excess biomass from the process is further treated before disposal. The process in combination with the primary treatment can remove up to 90+% of the incoming carbonaceous BOD and suspended solids. Under some conditions, it can also convert influent ammonia to nitrate, i.e., nitrification. The reactors used for the aeration process could be plug-flow or completely mixed type. In most cases, plug-flow configuration is used. The design of the reactor depends on the biokinetic parameters of the mixed micro-organisms developed in the aeration tank for the type of wastewater entering the system. The design of the system requires the determination of the following items: aeration tank volume, oxygen requirement, daily sludge wasting, and recycling rates.

Many modifications for the conventional process have been proposed for improving the system or for treating a specific type of wastewater. These modified processes are extended aeration, contact stabilization, step aeration, tapered aeration, high rate, and pure oxygen process.

Operational problems to activated sludge process can cause low soluble BOD removal and poor settling of the sludge in the secondary settling tank. The reasons for these problems could be because of unusual characteristics of the influent wastewater, or improper design and operation.

REFERENCES

1. Metcalf & Eddy, Inc. *Wastewater Engineering: Treatment and Reuse*, 4th Ed.; McGraw-Hill: New York, 2003.
2. Water Environment Federation. *Wastewater Treatment Plant Design*; Vesilind, P.A., Ed.; Water Environment Federation: Alexandria, VA, 2003.
3. Hänel, K. *Biological Treatment of Sewage by Activated Sludge Process*; Ellis Horwood Ltd.: Chichester, U.K., 1988.
4. Horan, N.J. *Biological Wastewater Treatment Systems*; John Wiley & Sons: Chichester, U.K., 1990.
5. Sundstrom, D.W.; Klei, H.E. *Wastewater Treatment*; Prentice-Hall, Inc.: Englewood Cliffs, NJ, 1979.
6. American Society of Civil Engineers. *Measurement of Oxygen Transfer in Clean Water*; ANSI/ASCE Standard 2-91, 2nd Ed.; American Society of Civil Engineers: Reston, VA, 1992.
7. Droste, R.L. *Theory and Practice of Water and Wastewater Treatment*; John Wiley & Sons: New York, 1997.
8. Gray, N.F. *Activated Sludge—Theory and Practice*; Oxford University Press: Oxford, U.K., 1990.
9. American Society of Civil Engineers (ASCE). *Design of Municipal wastewater Treatment Plants*, 4th Ed., ASCE Manual and Report on Engineering Practice No. 76; American Society of Civil Engineers: Reston, VA, 1998; Vol. 2.

Adsorption

Shivaji Sircar
Chemical Engineering Department, Lehigh University, Bethlehem, Pennsylvania, U.S.A.

INTRODUCTION

The separation and purification of fluid mixtures (gas or liquid) by adsorption is a major unit operation in the chemical, petrochemical, environmental, pharmaceutical, and electronic gas industries. A list of the key commercial applications of this technology is given in Table 1.[1] The phenomenal growth in the development of this technology is demonstrated by Fig. 1, which shows a year-by-year tally of U.S. patents issued between 1980 and 2000 on five different topics of adsorption.[1] The total number of patents is overwhelming.

ADSORPTION AS A SEPARATION PROCESS

Adsorption is a surface phenomenon. When a multicomponent fluid mixture is contacted with a solid adsorbent, certain components of the mixture (adsorbates) are preferentially concentrated (selectively adsorbed) near the solid surface creating an adsorbed phase. This is because of the differences in the fluid–solid molecular forces of attraction between the components of the mixture. The difference in the compositions of the adsorbed and the bulk fluid phases forms the basis of separation by adsorption. It is a thermodynamically spontaneous process, which is exothermic in nature. The reverse process by which the adsorbed molecules are removed from the solid surface to the bulk fluid phase is called desorption. Energy must be supplied to carry out the endothermic desorption process. Both adsorption and desorption form two vital and integral steps of a practical adsorptive separation process where the adsorbent is repeatedly used. This concept of regenerative use of the adsorbent is key to the commercial and economic viability of this technology.

Three generic adsorptive process schemes have been commercialized to serve most of the applications shown in Table 1. They include 1) temperature swing adsorption (TSA); 2) pressure swing adsorption (PSA); and 3) concentration swing adsorption (CSA).[2–9] The fluid mixture (feed) to be separated is passed over a regenerated adsorbent (contained in an adsorber vessel) to produce a stream enriched in the less strongly adsorbed components of the mixture, followed by desorption of the adsorbed components, which produces a stream enriched in the more strongly adsorbed components of the mixture. The TSA processes are generally designed for removal of trace impurities from a mixture (gas or liquid), where the desorption is effected by heating the adsorbent. The PSA processes are designed for separation of the components of a bulk gas mixture or for removal of dilute impurities from a gas stream, where the desorption is effected by lowering the gas phase partial pressure of the adsorbed components within the adsorber. The CSA processes are designed for separation of bulk liquid mixtures, where the desorption is effected by flowing a less selectively adsorbed liquid (eluent or desorbent) over the adsorbent. Numerous variations of these processes have been developed to achieve different separation goals by using 1) different modes and conditions of operation of the adsorption and the desorption steps in conjunction with a multitude of other complementary steps (designed to improve separation efficiency and product quality); 2) different types of adsorbents; 3) different process hardware designs; 4) different process control logic, etc.

Several families of micro- and mesoporous adsorbents offering a spectrum of adsorption characteristics are also available for these separations. Consequently, the technology has been a very versatile and flexible separation tool, which provides many different paths for a given separation need. This availability of multiple design choices is the driving force for innovations.[10] Commercial success, however, calls for a good marriage between the optimum adsorbent and an efficient process scheme. Several emerging concepts in this field can potentially expand its scope and scale of application. These include 1) rapid PSA processes; 2) novel adsorber configurations; 3) use of reversible chemisorbents; 4) adsorbent membranes; 5) simultaneous sorption and reaction, etc.[10]

The design and optimization of adsorptive processes typically require simultaneous numerical solutions of coupled partial differential equations describing the mass, heat, and momentum balances for the process steps. Multicomponent adsorption equilibria, kinetics, and heat for the system of interest form the key fundamental input variables for the design.[11,12] Bench- and pilot-scale process performance data are generally needed to confirm design calculations.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007656

Table 1 Key commercial applications of adsorption technology

Gas separation
Gas drying
Trace impurity removal
Air separation
Carbon dioxide–methane separation
Solvent vapor recovery
Hydrogen and carbon dioxide recovery from steam-methane reformer off-gas
Hydrogen recovery from refinery off-gas
Carbon monoxide–hydrogen separation
Alcohol dehydration
Production of ammonia synthesis gas
Normal–isoparaffin separation
Ozone enrichment
Liquid separation
Liquid drying
Trace impurity removal
Xylene, cresol, cymene isomer separation
Fructose–glucose separation
Fatty chemicals separation
Breaking azeotropes
Carbohydrate separation
Environmental separation
Municipal and industrial waste treatment
Ground and surface water treatment
Air pollution control
VOC removal
Mercury vapor removal
Bioseparation and pharmaceutical separation
Recovery of antibiotics
Purification and recovery of enzymes
Purification of proteins
Recovery of vitamins
Separation of enantiomers of racemic compounds
Removal of micro-organisms
Home medical oxygen production
Electronic gas purification
Production of ultrahigh-purity N_2, Ar. He, H_2, O_2
Purification of fluorinated gases NF_3, CF_4, C_2F_6, SiF_4
Purification of hydrides NH_3, PH_3, ASH_3, SIH_4, Si_2H_6

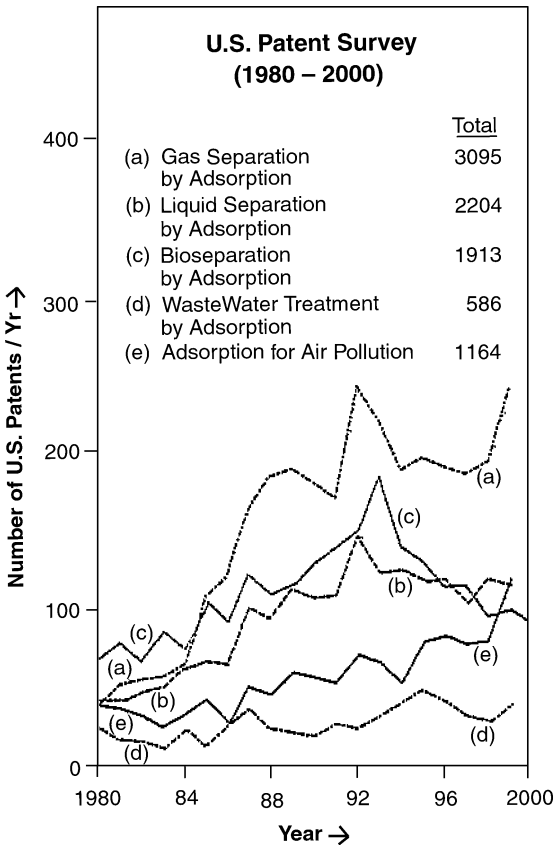


Fig. 1 U.S. patent survey of adsorption topics.

ADSORBENT MATERIALS

A key factor in the development of adsorption technology for the fluid separation has been the availability of appropriate adsorbents. The most frequently used categories include crystalline materials like zeolites, and amorphous materials like activated carbons, silica and alumina gels, polymeric sorbents, and ion-exchange resins. These materials exhibit a large spectrum of pore structures (networks of micro- and mesopores of different shapes and sizes) and surface chemistry (degrees of polarity), which provide a large choice of core adsorptive properties (equilibria, kinetics, and heat) to be utilized in the design of the separation processes. Table 2 lists some of the physical properties of common adsorbents.

The microporous alumino-silicate zeolites (Types A, X, and mordenite are frequently used) provide a variety of pore openings (3–10 Å), cavity and channel sizes, and framework Si/Al ratios. They are also available in various cationic exchanged forms (Na, K, Li, Ag, Ca, Ba, Mg), which govern their pore openings and cationic adsorption site polarities. They are highly hydrophilic materials and must be dehydrated before use. The amorphous adsorbents contain an intricate network of micropores and mesopores of various shapes and sizes. The pore size distribution may vary over a wide range. The activated carbons and the polymeric sorbents are relatively hydrophobic in nature. The silica and alumina gels are more hydrophilic (less than zeolites) and they must also be dehydrated before use.

Commercial adsorbents are generally produced in bound forms (0.5–6.0 mm diameters) in regular particle shapes (beads, pellets, extrudates, granules, etc.). The purpose is to reduce pressure drops in adsorbers. Clay, alumina, polymers, pitch, etc. are used as binders, which typically constitute 10–20% (by weight) of the final product. The binder phase usually contains a network (arteries) of meso- and macropores (0.5–50.0 μm diameters) to facilitate the transport of the adsorbate

Table 2 Physical properties of some adsorbents

	NaX Zeolite (Bayer, Germany)[b]	BPL Carbon (Calgon, U.S.A.)	Molecular Sieve Carbon (Takeda, Japan)	H151 Alumina (Alcoa, U.S.A.)	Silica Gel (Grace, U.S.A.)
BET area (m^2/g)	—	1100	—	350	800
Pore volume (cm^3/g)	0.54	0.70	0.43	0.43	0.45
Bulk density (g/cm^3)	0.65	0.48	0.67	0.85	0.77
Mean pore diameter (Å)	7.4[a]	30	3.5	43	22

[a]Crystal pore aperture size.
[b]Manufacturer given in parentheses.

molecules from the bulk fluid phase to the adsorption sites (within zeolite crystals and micropores of amorphous adsorbents) and vice versa. Adsorption of fluid molecules on the binder material is generally very weak. Fig. 2 shows a schematic drawing of a bound zeolite pellet depicting the pathways for transport of the adsorbate molecules.

The vast majority of fluid separation by adsorption is affected by the thermodynamic selectivity of the adsorbent for certain components of the fluid mixture over others. Physisorption is the dominant mechanism for separation. Thus, it is governed by the surface polarity of the adsorbent and the polarizability and permanent polarity of the adsorbate molecules. All adsorbate molecules, in this case, have access to the adsorption sites. The separation can also be based on a kinetic selectivity by the adsorbent where certain molecules of the fluid mixture diffuse into the adsorbent pores faster than the others because of their relative size differences. Size or steric exclusion of certain components of a fluid mixture from entering the adsorbent pores (typically for zeolites) is also possible. The last case is known as ''molecular sieving.'' Adsorbents may be energetically homogenous, containing adsorption sites of identical adsorption energy (heat of adsorption), or energetically heterogenous, containing a distribution of sites of varying energies. The cause of adsorbent heterogeneity is generally physicochemical in nature. It is created by a distribution of micro- and mesopores of different sizes and shapes within the adsorbent particle as well as by a distribution of adsorption sites of different surface chemistry and polarity within the micropores.

An adsorbent is often tailor-made to suit a separation need or a process can be designed to best fit the properties of an adsorbent. Special adsorbents are also available for specific applications (e.g., removal of mercury vapor, drying of reactive fluids, resistance to acids, etc). More recently, adsorbents have been produced that use reversible chemisorption as the mechanism for gas separation.[13] Creation of new adsorbents and modification of existing adsorbents continue to be an active area of research and development.

KEY ADSORPTIVE PROPERTIES FOR SEPARATION

All practical adsorptive separation processes are carried out using a stationary packed bed (adsorber) of the adsorbent particles. Each particle is subjected to the adsorption, the desorption, and the complementary

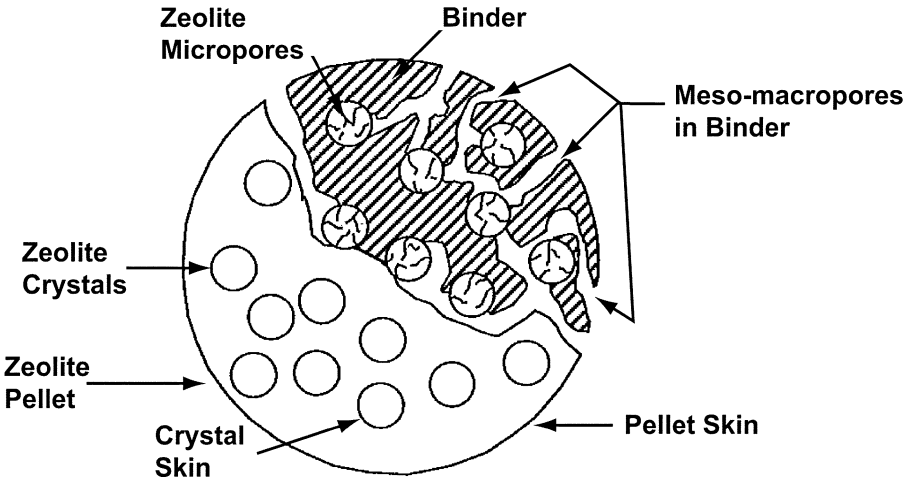


Fig. 2 Schematic drawing of a bound adsorbent particle.

steps of the process in a cyclic fashion. The ad(de)-sorption characteristics exhibited by the particle during different periods of the cycle are governed by the multi-component adsorption equilibria, kinetics, and heat for the fluid mixture of interest under the local conditions (e.g., fluid phase pressure, temperature and composition, adsorbate loadings in the particle, and its temperature) that the particle experiences. As these conditions can vary over a wide range during a process cycle, it is imperative that those adsorptive properties be accurately known over that range for reliable process design.

ADSORPTION EQUILIBRIA

Adsorption equilibria determine the thermodynamic limits of the specific amounts of adsorption (mol/g) of a pure gas or the components of a fluid mixture (gas or liquid) under a given set of conditions [pressure (P), temperature (T), and mole function (y_i or x_i) of component i] of the bulk fluid phase. The simplest way to describe adsorption equilibria of pure gas i is in the form of adsorption isotherms where the amount adsorbed (n_i^0) is plotted as a function of gas pressure (P) at a constant temperature (T). The pure gas adsorption isotherms can have various shapes (Types I–V) by Brunauer classification depending on the porosity of the adsorbent (microporous, mesoporous, or nonporous) and the system temperature (below or above the critical temperature of the adsorbate).[9] However, the most common isotherm shape is Type I, which is depicted by most microporous adsorbents of practical use. These isotherms exhibit a linear section in the very low-pressure region (Henry's law region) where the amount adsorbed is proportional to the gas pressure [$(n_i^0) = K_i P$]. The proportionality constant is called Henry's law constant (K_i), which is a function of temperature only. The amount adsorbed monotonically increases with increasing pressure beyond the Henry's law region with a progressively decreasing isotherm slope and finally the amount adsorbed asymptotically approaches the saturation adsorption capacity (m_i) of the adsorbate. Figs. 3A and 3B show examples of Type I isotherms for adsorption of pure N_2 and pure O_2, respectively, on various zeolites at 25°C.[2] The figures demonstrate that N_2 is more strongly adsorbed than O_2 on all zeolites and their adsorption characteristics are significantly affected by the structure of the zeolite, as well as by the nature of the cation present in them. The LSX zeolites in Fig. 3 represent X zeolite structure with low Si/Al ratio. The amounts adsorbed of a pure gas at any given pressure decrease with increasing temperature because of the exothermic nature of the adsorption process.

The equilibrium amounts adsorbed of component i from a binary gas mixture (n_i) are generally described as functions of gas phase mole fractions (y_i) at a constant system temperature (T) and total gas pressure (P). An example is given in Fig. 4 for adsorption of binary N_2–O_2 mixtures on Na–mordenite at various temperatures where the total gas pressure was 1.0 atm.[2] These binary isotherm shapes are typical for Type I adsorption systems on microporous adsorbents.

The relative adsorption between components i and j of a gas mixture is expressed in terms of the selectivity of adsorption ($S_{ij} = n_i y_j / n_j y_i$). Component i is more selective than component j if $S_{ij} > 1$. The thermodynamic selectivity decreases with increasing T for any given values of n_i. For adsorption on a homogenous adsorbent at constant T, S_{ij} can be constant, increase, or decrease with adsorbate loading depending on the size differences between the molecules of components

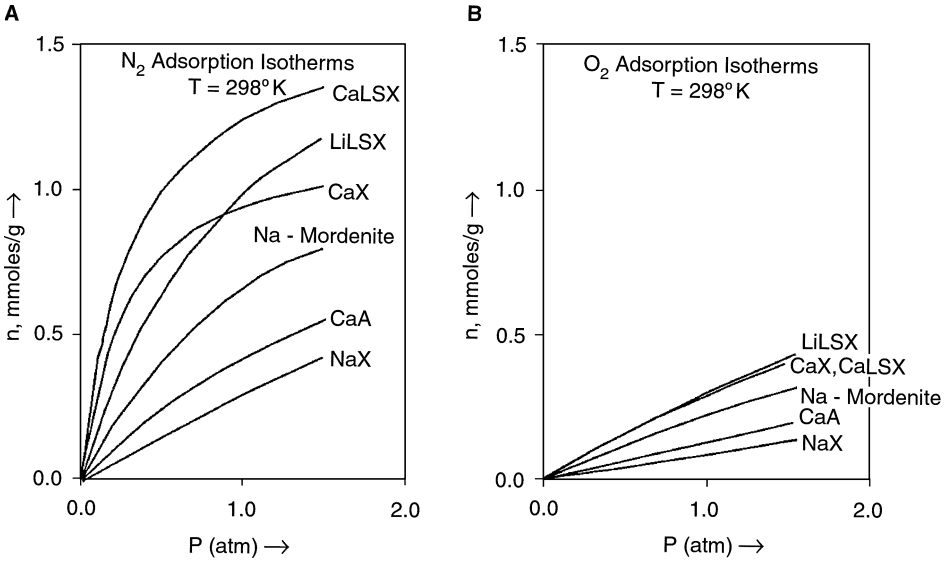


Fig. 3 Pure gas adsorption isotherms for (A) nitrogen and (B) oxygen on various zeolites.

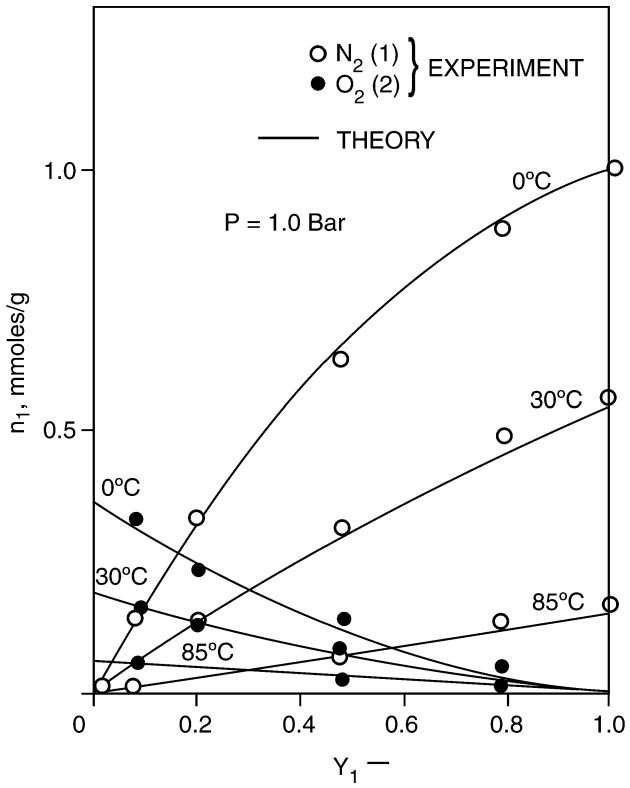


Fig. 4 Binary gas adsorption isotherms for nitrogen (1) and oxygen (2) mixtures on sodium mordenite.

i and j.[14] For adsorption on a heterogenous adsorbent, S_{ij} generally decreases with increasing adsorbate loading.[2] Table 3 gives a list of Henry's law selectivity ($S_{ij}^* = K_i/K_j$) for several binary gas mixtures at 30°C on a zeolite and an activated carbon.[2] The first-mentioned gas of a pair is the more selectively adsorbed component.

Separation of a gas mixture by a time dependent kinetic selectivity [$S_{ij}(t) = n_i(t)y_j/n_j(t)y_i$] has also been used in practice when there is a difference in the rates of adsorption of the components of the gas mixture.

Table 3 Selectivities of binary gas mixtures

Gas mixture	5A Zeolite	BPL Carbon
CO_2–CH_4	195.6	2.5
CO_2–CO	59.1	7.5
CO_2–N_2	330.7	11.1
CO_2–H_2	7400.0	90.8
CO–CH_4	3.3	0.33
CO–N_2	5.6	1.48
CO–H_2	125.0	12.1
CH_4–N_2	1.7	4.5
CH_4–H_2	37.8	36.6
N_2–H_2	22.3	8.2

$n_i(t)$, in this case, is the amount of component i adsorbed at time t.

Numerous models have been developed to describe pure and multicomponent gas adsorption on porous adsorbents. The analytical models are, however, most useful for process design. A few analytical models for Type I adsorption systems, which are thermodynamically consistent, are given below:[2]

$$\text{Langmuir:} \quad b_iPy_i = \theta_i/(1 - \theta_i) \tag{1}$$

$$\text{Multisite Langmuir:} \quad b_iPy_i = \theta_i \Big/ \left(1 - \sum \theta_i\right)^{a_i} \tag{2}$$

$$\text{Martinez–Basmadjian:} \quad b_iPy_i = \left\{\theta_i \Big/ \left(1 - \sum \theta_i\right)^{a_i}\right\} \exp\left(-\sum a_i w_i \theta_i\right) \tag{3}$$

$$\text{Toth:} \quad b_iPy_i = \theta_i \Big/ \left\{1 - \left(\sum \theta_i\right)^{k}\right\}^{1/k} \tag{4}$$

The frequently used Langmuir model describes adsorption of equal-sized adsorbates ($m_i = m_j$) on an energetically homogenous adsorbent. The multisite Langmuir model is an extension to include the effects of dissimilar adsorbate sizes ($m_i \neq m_j$). The Martinez–Basmadjian model is a further extension to include lateral interactions in the adsorbed phase. The Toth model is developed to describe adsorption of equal-sized molecules on an energetically heterogenous adsorbent. The variables of Eqs. (1)–(4) are the fractional coverage of component i of the gas mixture ($\theta_i = n_i/m_i$) at P, T, and y_i, the number of adsorption sites occupied by the adsorbate type i (a_i), the energy of lateral interactions between i molecules in the adsorbed phase (w_i), the gas–solid interaction parameter for component i (b_i), and the adsorbent heterogeneity parameter for all adsorbates ($k < 1$). The temperature coefficient of the parameter b_i is given by:

$$b_i = b_i^0 \exp(q_i^*/RT) \tag{5}$$

where b_i^0 is a constant. q_i^* is the isosteric heat of adsorption of pure gas i in the Henry's law region. R is the gas constant. The pure gas adsorption isotherms (θ_i^0 vs. P) for these models can be obtained by setting $y_i = 1$.

The extent of specific equilibrium adsorption of component i from a liquid mixture having mole fraction x_i for that component is expressed in terms of a variable called the Gibbsian surface excess [$n_i{}^e$ (mol/g)], which is related to the actual amounts adsorbed by[15]

$$n_i^e = n_i - \left(\sum n_i\right)x_i \tag{6}$$

The surface excess of component i is equal to its actual amount adsorbed ($n_i^e \sim n_i$) only when $x_i \ll 1$ and the

component i is very selectively adsorbed ($S_{ij} \gg 1$) over other components of the mixture.[15] The binary liquid phase surface excess adsorption isotherm (n_1^e vs. x_1) at a constant temperature (pressure is not a variable) on a microporous adsorbent is often U-shaped, as shown in Fig. 5A. Component 1 is selectively adsorbed if $n_1^e > 0$. For adsorption of a dilute solute from a liquid mixture, on a microporous solid, the excess isotherm is similar in shape as Type I isotherm for vapor adsorption (as shown in Fig. 5B).

Numerous analytical models have also been developed for binary liquid phase surface excess isotherms. A model equation that accounts for adsorbate size differences, bulk liquid phase nonideality, as well as a simplified description of adsorbent heterogeneity is given below:[16]

$$n_1^e = \frac{m_1}{(S_{0H} - S_{0L})}\left\{\frac{x_2}{a_1}\left[\left(\frac{S_H}{S_{0H}}\right)^{1/(\beta-1)} - \left(\frac{S_L}{S_{0L}}\right)^{1/(\beta-1)}\right] + \frac{a_2}{a_1(\beta-1)}\left(\frac{S_{0H}}{S_H} - \frac{S_{0L}}{S_L}\right)\right\} \quad (7)$$

$$\left(\frac{S_H}{S_{0H}}\right) = (S_H a_1 + a_2)^{(\beta-1)/\beta}; \qquad \left(\frac{S_L}{S_{0L}}\right) = (S_L a_1 + a_2)^{(\beta-1)/\beta} \quad (8)$$

where the variable $\beta = (m_1/m_2)$, a_i is the activity of component i in the bulk liquid phase, and S_{0L} and S_{0H} are the selectivities of adsorption of component 1 over component 2 at the limit of $x_1 \rightarrow 0$ at the lowest and the highest energy sites of the adsorbent.

HEAT OF ADSORPTION

The pertinent thermodynamic variable to quantitatively describe the thermal effects in the exo(endo) thermic gas ad(de)sorption process is called the isosteric heat of adsorption.[17] The isosteric heat of adsorption of component i of an ideal gas mixture (q_i) at adsorbate loading of n_i and temperature T is given by the following thermodynamic relationship:[17]

$$q_i(n_i) = RT^2\left\{\frac{\delta \ln(Py_i)}{\delta T}\right\}_{n_i} \quad (9)$$

Eq. (9) is frequently used to obtain the isosteric heat of adsorption of a pure gas i (q_i^0) as a function of n_i^0 and T from measured isotherms at different temperatures (y_i is equal to unity in that case). Estimation of isosteric heat of adsorption of component i of a gas mixture by using Eq. (9) is, however, not practical and they must be obtained by calorimetric measurements. In the absence of lateral interactions, the isosteric heat of adsorption of a pure gas on an energetically homogenous adsorbent is independent of adsorbate loadings. It remains constant at its value at the Henry's law region (q_i^*) at all coverages. The presence of lateral interactions between adsorbed molecules (pronounced at higher coverages) can increase q_i^0 with increasing n_i^0. The isosteric heat decreases with increasing adsorbate loading when the adsorbent is energetically heterogenous. The isosteric heat is generally a very weak function of T. Fig. 6 shows several calorimetrically measured examples of these behaviors for adsorption of pure SF_6 on various micro- and mesoporous adsorbents.[18] The isosteric heat of adsorption of a component of a gas mixture is equal to that of the pure gas for adsorption on a homogenous adsorbent.

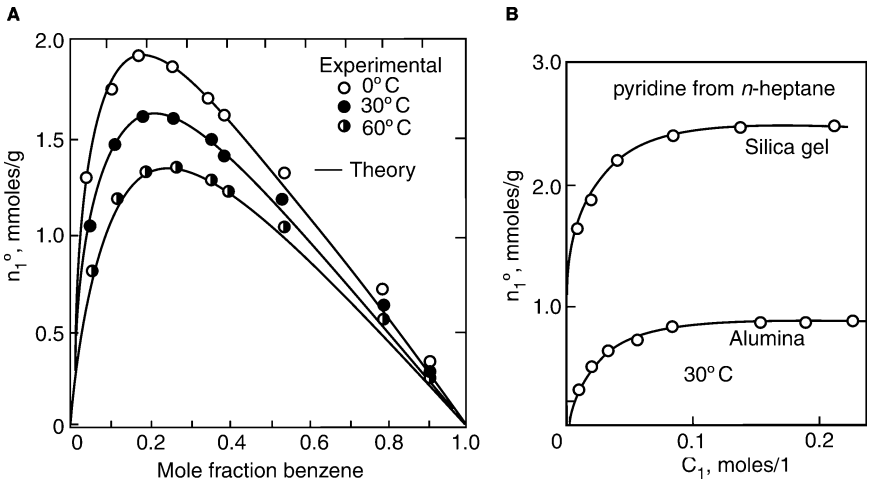


Fig. 5 Binary surface excess isotherms for adsorption of liquid mixtures: (A) benzene (1) + cyclohexane (2) on silica gel and (B) pyridine (1) + n-heptane (2) on silica gel and alumina.

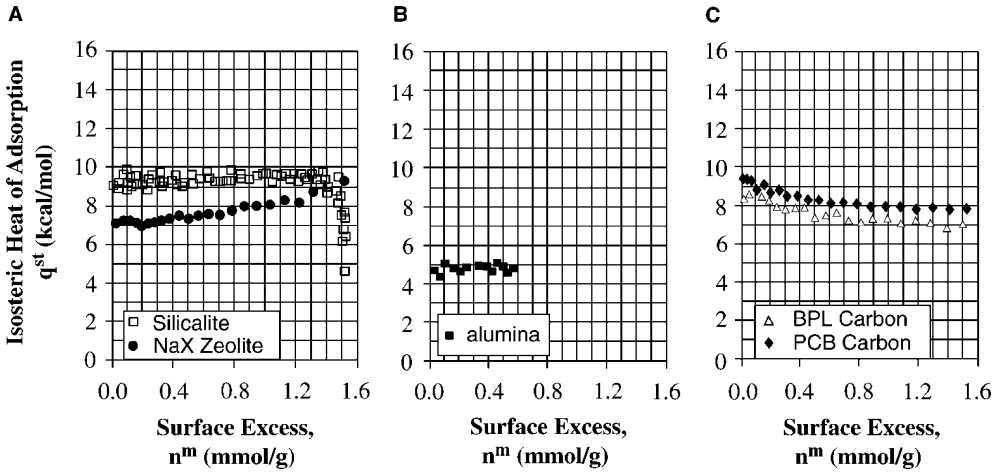


Fig. 6 Isosteric heat of pure SF_6 on various absorbents: (A) zeolites, (B) alumina, and (C) activated carbons.

However, the component isosteric heat of a mixture can be substantially different from that of the pure gas at the same loadings when the adsorbent is energetically heterogenous.[17] Furthermore, the dependence of component isosteric heat on the adsorbate loadings can be very complex in that situation.[17]

HEAT OF IMMERSION

The thermodynamic property to quantify the heat effects for ad(de)sorption of a liquid mixture is the heat of immersion. Fig. 7 shows examples of heat of immersion of binary benzene–cycloxane mixtures on two activated carbons at 30°C.[15] The corresponding surface excess isotherms are U-shaped (e.g., see Fig. 5A). Benzene is the more selectively adsorbed species. The

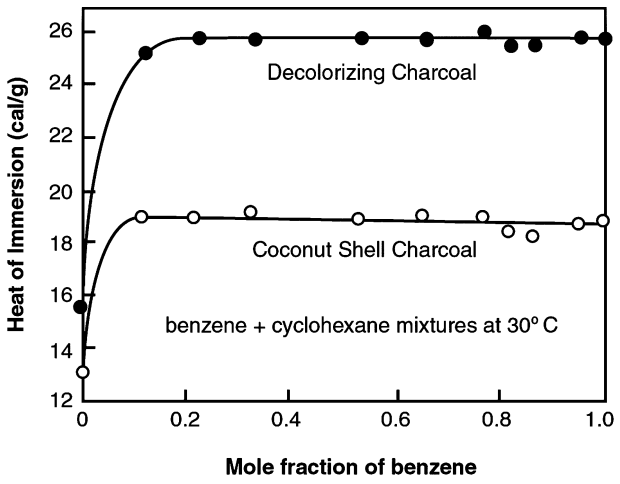


Fig. 7 Heat of immersion of binary liquid mixtures of benzene (1) + cyclohexane (2) on activated carbons.

temperature coefficients of liquid phase adsorption isotherms are generally much smaller than those for gas phase adsorption. The temperature changes in a liquid phase adsorber because of the ad(de)sorption processes are also small owing to the high heat capacity of the liquid.

ADSORPTION KINETICS

The actual kinetics of the ad(de)sorption process is generally very fast (in the order of microseconds). However, a significant resistance may exist for transfer of the adsorbate molecules from the bulk fluid phase to the adsorption sites inside the micropores of the adsorbent (see Fig. 2). For gas adsorption, these resistances may be caused by 1) molecular diffusion through a fluid film outside the adsorbent particles (for mixtures only); 2) Knudsen diffusion through the meso- and macropores of the adsorbent and the binder material in parallel with surface diffusion of adsorbed molecules on the walls of those pores (if any); and 3) activated (hopping) diffusion of adsorbed molecules inside the micropores. The same resistances exist for adsorption of liquid mixtures except that the flow through the meso-macropores of the adsorbent and the binder is controlled by molecular and surface diffusion through liquid filled pores. Additional mass transfer resistance called ''skin resistances'' at the surface of the adsorbent particles or zeolite crystals has also been observed.[2] Some or all of these processes are strongly influenced by the local fluid phase concentration and temperature within the adsorbent particle.

The magnitudes of gas diffusivities by different mechanisms follow the order: meso-macropore gas diffusivity (D_p) $\gg$ surface diffusivity (D_s) in those pores $\geq$ micropore diffusivity (D_m). For example, D_p

and D_s for adsorption of water vapor into the mesopores of an activated alumina sample were ~5–10 × 10^{-2} cm^2/sec and ~3 × 10^{-6} cm^2/sec, respectively.[4] The micropore diffusivity of water vapor into the crystals of NaA zeolite was ~1 × 10^{-6} cm^2/sec.[4]

The diffusivity of a liquid adsorbate through the meso-macropores is much slower than that for a gas. For example, the diffusivity of bulk liquid phase water from ethanol into an alumina sample was ~6 × 10^{-7} cm^2/sec.[4]

The time constant for meso-macropore diffusion into an adsorbent particle of radius R_p is given by (D_p/R_p^2). The time constant for micropore diffusion into a pore (having a characteristic diffusional distance of R_m) is given by (D_m/R_m^2). Because R_m (1–2 μm) is typically much smaller than R_p (1–2 mm), the meso-macropore diffusion generally controls the overall mass transfer into a practical adsorbent particle, even though $D_m \ll D_p$. An exception will be the case where the diameter of the adsorbate molecule is very close to that of the micropore, so that D_m is extraordinarily small, and the micropore diffusion becomes the controlling mechanism. Table 4 shows the micropore diffusivities of various gases in the Henry's law region into the crystals of several zeolites at 300 K.[2]

The table shows the remarkable decrease in the micropore diffusivity of a gas when its molecular diameter approaches that of the zeolite pore. The temperature coefficients of D_s and D_m are given by the Arrhenius relationship [D_s, $D_m = D^0 \exp(-E/RT)$] because these diffusions are activated processes. E is the activation energy for the diffusion process and D^0 is a constant. These diffusivities can also be complex functions of adsorbate loadings and compositions.[19]

The most rigorous formulation to describe adsorbate transport inside the adsorbent particle is the chemical potential driving force model. A special case of this model for an isothermal adsorption system is the Fickian diffusion (FD), model which is frequently used to estimate an effective diffusivity for adsorption of component i (D_i) from experimental uptake data for pure gases.[2,19] The FD model, however, is not generally used for process design because of mathematical complexity. A simpler analytical model called linear driving force (LDF) model is often used.[20] According to this model, the rate of adsorption of component i of a gas mixture into an adsorbent particle of radius R is given by:[11]

$$\frac{d\bar{n}_i}{dt} = k_{ii}(\bar{n}_i^* - \bar{n}_i) + \sum k_{ij}(\bar{n}_j^* - \bar{n}_j) \qquad (10)$$

where $\bar{n}_i$ is the average adsorbate loading of component i in the particle at time t. The variable $\bar{n}_i^*$ represents the adsorbate loading of component i that would be in equilibrium with the superincumbent gas phase conditions at time t. k_{ii} and k_{ij} are straight and cross mass transfer coefficients for component i. Further simplifications are often made by assuming that $k_{ij} = 0$ and k_{ii} is a function of T only. The relation between k_{ii} and D_i is generally given by $K_{ii} = \Omega(D_i/R^2)$, where Ω is a constant.[20] A value of $\Omega = 15$ is often used even though other values are also possible.[20] A parallel formulation called the surface excess linear driving force (SELDF) model for describing adsorption kinetics from liquid mixtures using Gibbsian surface excess as the variable has also been used successfully.[4]

Table 5 shows examples of LDF mass transfer coefficients for adsorption of several binary gas mixtures on BPL activated carbon particles (6–16 mesh) at ~23–30°C.[21] The data show that the mass transfer coefficients are relatively large for these systems. There is a scarcity of multicomponent adsorption equilibria, kinetics, and heat data in the published literature. This often restricts extensive testing of theoretical models for prediction of multicomponent behavior.

DESCRIPTION OF SELECTED ADSORPTIVE SEPARATION PROCESSES

Adsorption has become the state of the art technology for many separation applications as listed in Table 1. The more prolific areas include:

a. Drying of gaseous and liquid mixtures.
b. Production of oxygen and nitrogen enriched gases from air.
c. Production of ultrapure hydrogen from various gas sources.
d. Separation of bulk liquid mixtures where distillation is not convenient.

Table 4 Examples of pure gas diffusivities into zeolites at 300 K[a]

Gas	Kinetic diameter (Å)	NaA Zeolite (4A) (m^2/sec)	Na-CaA Zeolite (5A) (m^2/sec)	Nax Zeolite (13X) (m^2/sec)
N_2	3.70	3.1 × 10^{-14}	1.1 × 10^{-10}	Fast
Kr	3.65	1.2 × 10^{-17}	—	—
CH_4	3.76	1.2 × 10^{-15}	6.0 × 10^{-10}	—
n-C_4H_{10}	4.69	9.6 × 10^{-20}	1.5 × 10^{-13}	2.4 × 10^{-11}

[a]Effective crystal pore openings of 4A, 5A, and 13X zeolites are 4.0, 4.9, and 7.6 Å, respectively.

Table 5 Examples of binary LDF mass transfer coefficients on BPL carbon

Gas mixtures	k_{11} (sec^{-1})	k_{22} (sec^{-1})
CO_2 (1) + He (2)	0.44	—
CH_4 (1) + He (2)	1.42	—
N_2 (1) + He (2)	3.33	—
CO_2 (1) + CH_4 (2)	0.35	0.76
CO_2 (1) + N_2 (2)	0.35	0.66

There has been extensive research and development in all of these areas during the last 30 yr. For example, topics (a)–(c) alone have generated more than 600 U.S. patents during the period 1980–2000.[1] They have been assigned to 160 different corporations around the world. For the sake of brevity, only a selected few commercial processes will be discussed here.

Adsorptive Drying

Both TSA and PSA processes are commercially used for removal of trace or dilute water contamination from a gas. They are commercially designed to handle 1–40,000 ft^3 of feed gas per minute. The basic steps of a conventional TSA process consists of: 1) flowing the contaminated gas over a packed bed of a desiccant (silica gel, alumina, zeolite) at a near-ambient temperature and withdrawing a dry product gas until the moisture concentration in the product gas rises to a preset level; 2) heating the adsorbent to ~150–300°C by flowing a hot dry gas countercurrently through the dbed and rejecting the water laden effluent gas; and 3) cooling the bed to feed gas temperature by countercurrently flowing a dry gas through the bed at feed gas temperature while rejecting the effluent gas. The cycle is then repeated. A part of the dry gas (~10–30%) produced in step 1 is generally used to supply the gas for steps 2 and 3. The effluent gas from step 3 is often heated to supply the gas for step 2. Fig. 8A is a schematic diagram of a three-column TSA gas drying unit. The total cycle time (all steps) for TSA processes generally varies between 2 and 8 hr. A typical dynamic water removal capacity of an alumina dryer is ~5–15% by weight. Product gas dew points of less than −40°C can be easily obtained. Many different process modifications like thermal pulsing, elimination of the cooling step, lower temperature regeneration, etc. are also used for decreasing the costs of drying.[4]

The TSA dryers for liquid mixtures use similar process steps to those used for gas dryers except that the adsorbers are first drained to remove the void liquid before they are heated countercurrently using a hot gas. Heating vaporizes the adsorbed water as well as liquid films adhering to the adsorbent particles. The hot effluent gas is cooled to condense out the components of the feed liquid mixture. The adsorber is then cooled by countercurrently flowing a cold gas and refilled with dry liquid before starting a new cycle.[4]

The basic steps of a conventional PSA gas drying process (Skarstrom cycle) consists of: 1) adsorption of water vapor from the feed gas by flowing the gas over a desiccant bed (silica gel, alumina, zeolite) at an elevated pressure (say 5–15 atm) and withdrawing a dry product gas at feed pressure; 2) countercurrently depressurizing the adsorber to near-ambient pressure and venting the effluent; 3) countercurrently purging the adsorber with a part of the dry gas produced by step 1 at near-ambient pressure while venting the effluent; and 4) countercurrently pressurizing the adsorber with a part of the product gas from step 1. Adsorption at a relatively lower feed gas pressure (1.3–1.7 atm) and desorption under vacuum (both depressurization and purge steps) are also practiced. Fig. 8B is a schematic diagram of a two-bed PSA dryer. The process can be used to obtain a very dry product (say −60°C dew point). A typical process uses ~15–30% of the dry product gas as purge. The total cycle times for PSA processes generally vary between 2 and 6 min.[4]

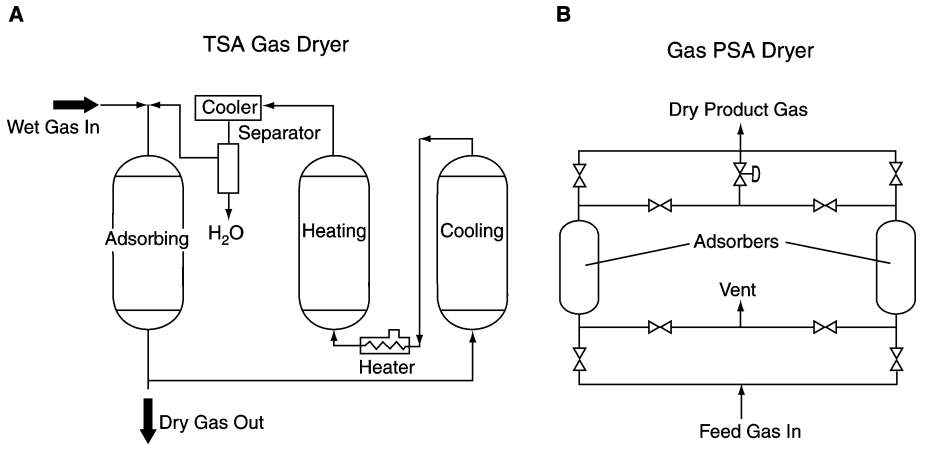


Fig. 8 Schematic drawings of (A) three-column TSA and (B) two-column PSA processes.

Air Fractionation

A large variety of PSA process concepts have been commercialized for: 1) production of 23–95 mol% oxygen using a N_2 selective (thermodynamic) zeolite; 2) production of 99+ mol% nitrogen using an O_2 selective (kinetic) carbon molecular sieve; and 3) simultaneous production of 90+ mol% O_2 and 98+ mol% N_2 using a N_2 selective zeolite from ambient air. Some of these concepts are called vacuum swing adsorption (VSA) because the final desorption pressure is subatmospheric. Commercial units are designed in the size range of 0.012–100 tons of oxygen per day (TPD).[2]

A VSA process for production of 90% O_2 from air using a LiX zeolite uses the cyclic steps of: 1) pressurizing the adsorber from an intermediate pressure level (P_I) to the final adsorption pressure level of P_A ($\sim$1.43 atm) with compressed air; 2) flowing compressed air feed at P_A through the adsorber and producing the 90% O_2 enriched product gas; 3) countercurrently depressurizing the adsorber to near-ambient pressure and venting the effluent gas; 4) countercurrently evacuating the adsorber to a vacuum level of P_D ($\sim$0.34 atm) and countercurrently purging the adsorber with a part of the O_2 product gas from step 2 at that pressure while venting the effluent gas; and finally (5) countercurrently pressurizing the adsorber from P_D to P_I using a part of the product gas. The cycle is then repeated.[22] Using a total cycle time of 70 sec, the O_2 productivity by the process, in terms of bed size factor (BSF), was 830 lb of zeolite per TPD of oxygen. The total power requirement was 11.6 kW/TPD when the product O_2 gas was delivered at a pressure of 6.45 atm by recompression.[22]

The LiX zeolite is currently favored for O_2 production. Table 6 demonstrates its superiority over other zeolites for making 90 mol% O_2 from air using a specific VSA cycle operating between final adsorption and desorption pressures of 1.48 and 0.25 atm, respectively.[23]

The Carbon Molecular Sieves (CMS) contain constricted pore mouths that permit the slightly smaller O_2 molecules to diffuse into the pores of the carbon faster than the N_2 molecules from air. This produces an O_2 enriched adsorbed phase based on kinetic selectively when the CMS is contacted with air for a short period of time. The material has practically no thermodynamic selectivity for either gas.[24] A simple four-step PSA process using a CMS consists of: 1) passing compressed air into the adsorber to pressurize it to P_A, and then withdrawing a N_2 enriched product gas at that pressure; 2) connecting the adsorber with another adsorber that has completed step 3 below, to pressure equalize the adsorbers; 3) countercurrently depressurizing the adsorber to near-ambient pressure and venting the O_2 enriched gas, and finally 4) pressure equalizing the adsorber with the companion adsorber, which has completed step 1. The cycle is then repeated. Using a feed air pressure of $\sim$8.1 atm ($=P_A$), the process could produce a 99.0 mol% N_2 product gas with a N_2 recovery of 49.4% from the feed air. The N_2 productivity was $\sim$92 ft^3/hr per cubic foot of adsorbent.[24] The O_2 enriched waste gas contained 33.8% O_2. The N_2 productivity and recovery were reduced to 73% and 39.2%, respectively, when its purity was raised to 99.5 mol%. Other process cycles have been designed using CMS adsorbents to raise the N_2 product gas purity above 99.9 mol%.[25]

Table 6 Comparative performance of various zeolites for O_2 production by a VSA process

Zeolite	BSF	Energy of separation
NaX	1.00	1.00
CaX	1.28	1.78
LiX	0.51	0.88

Production of Hydrogen

The most common industrial method to make ultrapure hydrogen is by steam-methane reforming (SMR) using a catalyst at the temperature 890–950°C. The reformed gas is then subjected to a high temperature water gas shift (WGS) reaction at 300–400°C. The WGS reactor effluent typically contains 70–80% H_2, 15–25% CO_2, 1–3% CO, 3–6% CH_4, and trace N_2 (dry basis), which is fed to a PSA system at a pressure of 8–28 atm and a temperature of 20–40°C for production of an ultrapure (99.99+ mol%) hydrogen gas at the feed pressure. Various PSA systems have been designed for this purpose to produce 1–120 million cubic feet of H_2 per day.

A popular PSA cycle called polybed process consists of 11 cyclic steps.[26] They include: 1) passing the feed gas at pressure P_F through a packed adsorbent column and withdrawing the high-purity H_2 product gas; 2) cocurrently depressurizing the adsorber from pressure P_F to P_I while producing a stream of essentially pure H_2; 3) further cocurrently depressurizing the adsorber from pressure P_I and P_{II} and withdrawing another stream of H_2 enriched gas; 4) even further depressurizing the adsorber cocurrently from pressure P_{II} and P_{III} and again withdrawing a stream of H_2 enriched gas; 5) countercurrently depressurizing the adsorber from pressure P_{III} to a near-ambient pressure (P_D) and venting the effluent gas; 6) countercurrently purging the adsorber with a stream of essentially pure H_2 produced by a companion adsorber undergoing step 4; 7) countercurrently pressurizing the adsorber from P_D to P_{II} by introducing the effluent gas from

another adsorber carrying out step 3; 8) further pressurizing the adsorber countercurrently from pressure P_{II} and P_I using the gas produced by another adsorber undergoing step 2; and 9) finally pressurizing the adsorber countercurrently from pressure P_I to P_F using a part of the product gas produced by another adsorber carrying out step 1. The cycle is then repeated. Increasing the pressure ratio between the feed gas and the purge gas (P_F/P_D) generally improves the separation performance of the PSA process by: 1) providing higher specific adsorption capacities for the selectively adsorbed components of the feed gas during the adsorption step (thus reducing adsorbent inventory) and 2) lowering the quantity of the purge gas required for adsorbent regeneration (thus increasing product recovery). However, increased feed gas pressure also increases the void gas quantity inside the adsorber at the end of the adsorption step, which: 1) increases the amount of adsorbent needed to contain the impurities during the subsequent cocurrent depressurization steps and 2) increases product loss during the countercurrent depressurization step. Consequently, there is an upper feed gas pressure limit (typically <40 atm) for optimum operation of the PSA cycle.

The most distinguishing feature of the polybed process is that step 1 is terminated when there is a substantial adsorption capacity left for the feed gas impurities inside the column so that these impurities are removed from the expanding void gases in steps 2–4 to produce high-purity H_2 effluents. The adsorbers are generally packed with a layer of an activated carbon at the feed end primarily for selective removal of H_2O and CO_2. A layer of 5A zeolite at the product end removes any remaining CO, CH_4, and N_2. These adsorbents are chosen due to the ease of desorption of the impurities from them. Fig. 9A shows a schematic diagram of a polybed system containing nine parallel adsorbers. Using a total cycle time of 13.3 min, the polybed process could produce a 99.999% pure H_2 product from a WGS reactor effluent gas at 20.7 atm and 21°C with a H_2 recovery of 86.0%. The feed processing capacity was ~35 ft^3 of feed gas per cubic foot of total adsorbent in the system/cycle.[26]

Pressure swing adsorption processes are also designed to produce high-purity (99.95+ %) H_2 products from refinery-off gases containing H_2 (65–90%) and C_1–C_5 hydrocarbon impurities with high H_2 recoveries (~86+ %). Silica gel and activated carbons are used as adsorbents.[27]

Separation of Bulk Liquid Mixtures

A variety of simulated moving bed (SMB) processes has been commercialized for adsorptive separation of bulk liquid mixtures where distillation is not cost-effective. They are designed to process 100–120,000 tons of bulk chemicals per year. Key examples include separation of: 1) *n*-paraffins from branched and cyclic compounds; 2) olefins from paraffins; 3) xylene isomers; 4) glucose–fructose mixture; and 5) enantiomers of racemic compounds. The SMB concept is based on the principles of liquid phase chromatography. It uses a stationary bed of an adsorbent with multiple inlet and outlet ports along its length as shown by Fig. 9B. Continuous countercurrent flow of the solid and the liquid phases inside the adsorber is mimicked by periodically moving the feed introduction and product withdrawal points to and from the adsorber using a rotary valve (RV). A weakly adsorbed desorbent liquid (D) is continuously circulated through the column with a pump. The pump flow rate is periodically changed with the change in the position of the RV. The process splits a feed liquid mixture (A + B) into two easy to separate liquid mixtures (A + D and B + D), which can then be further separated by distillation, if needed. For example, an SMB process could separate a *n*-paraffin + *n*-olefin feed mixture of C_{11}–C_{14} hydrocarbons containing ~9.0 wt% olefins to produce an olefin enriched stream containing ~96.2% olefins with an olefin recovery of ~94.0%. The paraffin enriched stream contained ~98.5 wt%

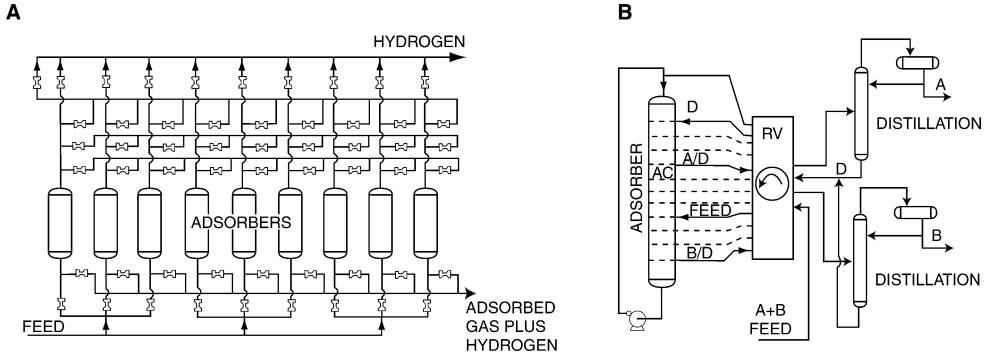


Fig. 9 Schematic drawings of (A) nine column polybed PSA and (B) SMB process with two distillation columns.

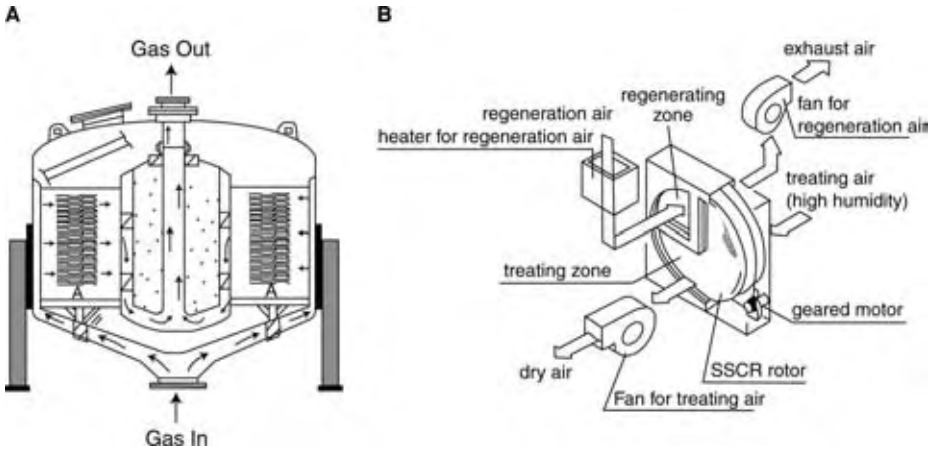


Fig. 10 Schematic drawings of (A) radial bed adsorber and (B) rotary bed adsorber.

paraffins. A hydrocarbon mixture of lower boiling range was used as the desorbent liquid.[6] The preferred adsorbent was NaX zeolite (caustic treated).[3,6]

EMERGING ADSORPTIVE SEPARATION PROCESSES

Design of new adsorptive process cycles using new (or modified) and existing adsorbents continues to grow. The following areas have attracted considerable attention in recent years.

Rapid PSA Cycles

The concept involves the use of faster cycle times (seconds) than those for conventional PSA cycles (minutes) to increase the productivity rate (volume of product gas/volume of adsorbent/hour) by several folds. This is achieved by operating an existing PSA cycle faster using modified hardware or by designing new process cycles using conventional hardware. An example of the latter case is a rapid PSA (RPSA) process for production of ~27.5% O_2 from air.[2] It uses multiple layers of a N_2 selective zeolite stacked in a single adsorber vessel. The layers are cyclically subjected to: 1) simultaneous air pressurization and adsorption and 2) simultaneous depressurization and back purging with O_2 product gas. Using a total cycle time of 10 sec, the O_2 productivity could be raised to ~2300 ft^3/hr per cubic foot of adsorbent with an O_2 recovery of ~64%. This productivity is an order of magnitude higher than conventional PSA processes.[2]

Radial and Rotary Bed Adsorbers

The maximum permissible gas flow rate through a conventional packed bed adsorber is governed by pressure drop, local fluidization, and channeling within the column. Radial bed adsorbers are designed to circumvent some of these problems. The adsorbent is placed in an annular section between two coaxial cylindrical chambers with perforated walls, and the gas flows radially through the arrangement. Fig. 10A is a schematic drawing of a radial bed adsorber.[28] The adsorber allows lower pressure drops, faster cycle times, higher throughputs, and elimination of fluidization, but they are more expensive to build. Radial beds have been used for both PSA and TSA processes.

The rotary bed adsorber (also called adsorption wheel) is designed to provide a truly continuous TSA system.[29] It uses a shallow wheel-shaped adsorption bed that continuously turns about an axis inside a fixed supporting frame. A section of the wheel is used to adsorb impurities from a feed gas while the other section is used to thermally regenerate the adsorbent. Fig. 10B is a schematic drawing of the adsorption wheel arrangement. The adsorbent is made from a

Table 7 Separation performance of MSC membranes

Mixture	Permeance of component 1 (Gas permeation unit)	Selectivity of component 1
O_2 (1) + N_2 (2)	16.0	13.6
H_2 (1) + CH_4 (2)	365.5	500.0
CO_2 (1) + CH_4 (2)	91.0	50.0
C_3H_6 (1) + C_3H_8 (2)	183.0	12–15

Table 8 Hydrocarbon rejections by SSF membrane

Hydrocarbons	Feed gas (mol%)	Rejections (%)
C_3H_8	2–7	94
C_3H_6	2–7	94.5
C_2H_6	7–15	84–91
C_2H_4	5–7	86–91
CH_4	35–50	47–55
H_2	14–30	—

honeycomb-shaped alumina substrate, which can be coated with layers of different adsorbents. The rotary adsorber has been designed for dehumidification of gases, solvent vapor recovery, volatile organic compounds (VOC) removal, gas deodorization, etc.

Hybrid Gas Separation Using Adsorption

Hybrid concepts combining the principles of selective adsorption with those of membrane technology and reaction engineering have been developed to enhance the overall separation or reaction efficiency.[10] Two examples are given below.

Nanoporous Carbon Membrane

The nanoporous carbon membrane consists of a thin layer (<10 μm) of a nanoporous (3–7 Å) carbon film supported on a meso-macroporous solid such as alumina or a carbonized polymeric structure. They are produced by judicious pyrolysis of polymeric films. Two types of membranes can be produced. A molecular sieve carbon (MSC) membrane contains pores (3–5 Å diameters), which permits the smaller molecules of a gas mixture to enter the pores at the high-pressure side. These molecules adsorb on the pore walls and then they diffuse to the low-pressure side of the membrane where they desorb to the gas phase. Thus, separation is primarily based on differences in the size of the feed gas molecules. Table 7 gives a few examples of separation performance of MSC membranes.[30] Component 1 is the smaller component of the feed gas mixture.

The pores of a selective surface flow (SSF) membrane are large enough (6–7 Å) to permit all molecules of a gas mixture to enter the pores. However, the larger and more polar molecules are selectively adsorbed at the high-pressure side. Then, they selectively diffuse on the pore walls to the low-pressure side, where they desorb into the gas phase. The adsorbed molecules also block the passage of smaller molecules through the void space within the pores. Thus, separation is primarily based on selective adsorption and surface diffusion. One key advantage of the SSF membrane is that the smaller molecules of the feed gas mixture (often the desired species) are enriched in the high-pressure side, which reduces subsequent recompression, if needed. The membrane can be used to enrich H_2 from mixtures with C_1–C_4 hydrocarbons, separate H_2 and CH_4 from H_2S, and dehumidify a gas stream at moderate feed gas pressures.[30] Table 8 shows an example of the performance of an SSF membrane in treating a refinery waste gas (pressure ~3.5 atm) where a hydrogen recovery of 60% was achieved.[30]

Simultaneous Adsorption and Reaction

The concept is based on Le Chatelier's principle that the conversion of an equilibrium controlled reaction as well as the rate of the forward reaction can be significantly enhanced by selectively removing one of the reaction products from the reaction zone. A selective adsorbent can be used for this purpose by mixing it with the catalyst in a reactor. The adsorbent, however, must be periodically regenerated. Such concepts are often called ''pressure swing reactors'' when the principles of PSA are used for regeneration.[10] One example is a cyclic process called sorption enhanced reaction process (SERP), which was developed for producing H_2 by SMR ($CH_4 + 2H_2O \leftrightarrow CO_2 + 4H_2$) using an admixture of a conventional SMR catalyst and a novel chemisorbent (K_2CO_3 promoted hydrotalcite) for selective removal of CO_2 from the reaction zone. The process used steam purge under vacuum for regeneration of the chemisorbent. The process could directly produce an essentially CO_x-free H_2 product containing ~95% H_2 and 5% CH_4 at a feed pressure of ~26 psia using a H_2O/CH_4 ratio of 6 : 1 as feed. A CH_4 to H_2

Table 9 Performance of SERP concept for H_2 production[a]

	Product composition (dry)				
	H_2	CH_4	CO_2	CO	Conversion (%)
SERP concept	94.4%	5.6%	40 ppm	<30 ppm	73.0
Conventional SMR	67.2%	15.7%	15.9%	1.2%	52.6

[a]$T = 490°C$, $P = 26.1$ psia, $H_2O/CH_4 = 6.1$.

conversion of ~73% could be achieved at a reaction temperature of only ~500°C, which was much lower than the temperature of a conventional SMR reaction (~850°C). Table 9 describes the pilot test performance of the SERP concept.[13] It also shows the equilibrium limitation of an SMR reactor operated without using the concept. The advantages are obvious.

CONCLUSIONS

The range of applications for gas and liquid separation and purification by adsorption is large and growing. The strong research and development activity in this area is facilitated by the flexibility of practical adsorptive process designs such as pressure and thermal swing adsorption, and SMB adsorption, as well as by the availability of a large spectrum of new and old micro- and mesoporous adsorbents.

The fundamental adsorptive properties governing the performance of the separation processes are the multicomponent equilibria, kinetics, and heat. A large volume of data, as well as models to describe them, exist in the published literature only for adsorption of pure gases and binary liquid mixtures. Binary gas adsorption data are sporadic. Multicomponent data are rare. Existence of adsorbent heterogeneity can introduce severe complexity in the multicomponent adsorption behavior.

Despite these limitations, adsorption has become the state of the art technology for many commercial separations. Examples include: 1) gas and liquid drying; 2) production of oxygen and nitrogen enriched air; 3) hydrogen purification; and 4) several bulk liquid phase separations of close boiling compounds. Emerging topics on adsorption research and development include: 1) rapid pressure swing adsorption; 2) novel adsorber configurations; 3) adsorbent membranes; and 4) simultaneous reaction and adsorption.

REFERENCES

1. Sircar, S. Publications on adsorption science and technology. Adsorption **2000**, *6*, 359–365.
2. Sircar, S.; Myers, A.L. Gas separation by zeolites. In *Handbook of Zeolite Catalysts and Microporous Materials*; Aurbach, S.M., Carrado, K.A., Dutta, P.K., Eds.; Marcel Dekker, Inc.: New York, 2003; Chapter 22, 1063–1105.
3. Kulaprathipanja, S.; Johnson, J.A. Liquid separations. In *Handbook of Porous Solids*; Schüth, F., Sing, K.S.W., Weitkamp, J., Eds.; Wiley-VCH: Weinheim, Germany, 2002; Vol. 5, Chapter 6.4, 2568–2622.
4. Sircar, S. Drying processes. In *Handbook of Porous Solids*; Schüth, F., Sing, K.S.W., Weitkamp, J., Eds.; Wiley-VCH: Weinheim, Germany, 2002; Vol. 5; Chapter 6.3, 2533–2567.
5. Crittenden, B.; Thomas, W.J. *Adsorption Technology and Design*; Butterworth-Heinemann: Oxford, U.K., 1998.
6. Gembicki, S.A.; Oroskar, A.R.; Johnson, J.A. Adsorption, liquid separation. In *Encyclopedia of Separation Technology*; Ruthven, D.M., Ed.; John Wiley: New York, 1997; Vol. 1, 172–199.
7. Ruthven, D.M.; Farooq, S.; Knaebel, K.S. *Pressure Swing Adsorption*; VCH Publishers: Boca Raton, FL, 1994.
8. Sircar, S. Pressure swing adsorption technology. In *Adsorption: Science and Technology;* NATO ASI Series E; Rodrigues, A.E., Levan, M.D., Tondeur, D., Eds.; Kluwer Academic Publishers: London, 1989; Vol. 158, 275–321.
9. Yang, R.T. *Gas Separation by Adsorption Processes*; Butterworths: London, 1987.
10. Sircar, S. Application of gas separation by adsorption for the future. Adsorpt. Sci. Technol. **2001**, *19* (5), 347–366.
11. Sircar, S. Pressure swing adsorption: research needs by industry. Proceedings of the Third International Conference on Fundamentals of Adsorption, Sonthofen, Germany, May 5–9, 1989; Mersmann, A.B., Scholl, S.E., Eds.; Engineering Foundation: New York, 1991; 815–843.
12. Hartzog, D.G.; Sircar, S. Sensitivity of PSA process performance to input variables. Adsorption **1995**, *1*, 133–151.
13. Waldron, W.E.; Hufton, J.R.; Sircar, S. Production of hydrogen by cyclic sorption enhanced reaction process. AIChE J. **2001**, *47* (6), 1477–1479.
14. Sircar, S.; Rao, M.B. Effect of adsorbate size on adsorption of gas mixtures on homogeneous adsorbents. AIChE J. **1999**, *45*, 2657–2661.
15. Sircar, S.; Myers, A.L. Liquid adsorption operations: equilibrium, kinetics, column dynamics and applications. Sep. Sci. Technol. **1986**, *21*, 535–562.
16. Sircar, S. Adsorption from binary liquid mixtures of unequal adsorbate sizes on heterogeneous adsorbents. Surf. Sci. **1984**, *148*, 489–498.
17. Sircar, S.; Rao, M.B. Heat of adsorption of pure gas and multicomponent gas mixtures on microporous adsorbents. In *Surfaces of Nanoparticles and Porous Materials*; Schwarz, J.A., Contescu, C.I., Eds.; Marcel Dekker Inc.: New York, 1999; Chapter 19, 501–528.
18. Cao, D.V.; Sircar, S. Heat of adsorption of SF_6 on various adsorbents. Adsorption **2001**, *7*, 73–80.
19. Karger, J.; Ruthven, D.M. *Diffusion in Zeolites and Other Porous Solids*; John Wiley: New York, 1992.

20. Sircar, S.; Hufton, J.R. Why does linear driving force model for adsorption kinetics work? Adsorption **2000**, *6*, 137–147.
21. Sircar, S.; Kumar, R. Column dynamics for adsorption of bulk binary gas mixtures on activated carbons. Sep. Sci. Technol. **1986**, *21* (9), 919–939.
22. Leavitt, F.W Vacuum Pressure Swing Adsorption ProcessU.S. Patent 5,415,693, May 16, 1995.
23. Leavitt, F.W Vacuum Pressure Swing Adsorption ProcessU.S. Patent 5,074,892, Dec 24, 1991.
24. Sircar, S. Adsorption technology—a versatile separation tool. In *Separation Technology—The Next Ten Years*; Garside, J., Ed.; Institute of Chemical Engineers: Rugby, U.K., 1994; 47–72.
25. Lemcoff, N.C.; Gmelin, R.C Pressure Swing Adsorption Method for Separating Gaseous Mixtures. U.S. Patent 5,176,722, Jan 5, 1993.
26. Fuderer, A.; Rudelsdorfer, E Selective Adsorption Process. U.S. Patent 3,986,849, Oct 19, 1976.
27. Yamaguchi, T.; Kobayashi, Y Gas Separation Process. U.S. Patent 5,250,088, Oct 5, 1993.
28. Smoralek, J.; Leavitt, F.W.; Nowobiliski, J.J.; Bergsten, E.; Fassbaugh, J.H Radial Bed Vacuum/Pressure Swing Adsorber Vessel. U.S. Patent 5,759,242, Jun 2, 1998.
29. Hirose, T.; Kuma, T. Honeycomb rotor continuous adsorber for solvent recovery and dehumidification. 2nd Korea–Japan Symposium on Separation Technology, 1990.
30. Sircar, S.; Rao, M.B. Nanoporous carbon membranes for gas separation. In *Recent Advances on Gas Separation by Microporous Membranes*; Kanellopoulos, N., Ed.; Elsevier: Amsterdam, The Netherlands, 2000; 473–496.

Advanced Oxidation

Sangchul Hwang
Department of Civil Engineering and Surveying, University of Puerto Rico, Mayagüez, Puerto Rico

INTRODUCTION

Chemical oxidation technologies are defined as the processes that use oxidizing agents to degrade or transform complex hazardous chemicals to simpler nontoxic ones. Advanced oxidation processes (AOPs) constitute, in general, the generation and the use of hydroxyl radicals ($^{\bullet}OH$) to oxidize hazardous chemicals, which are otherwise very recalcitrant to conventional oxidation processes.

Advanced oxidation processes have been used for the treatment of drinking water, wastewater, and soil/groundwater contaminated with unwanted and hazardous substances. The processes are, in general, based on the generation and the use of highly reactive hydroxyl radicals ($^{\bullet}OH$) that react indiscriminately with many organic and inorganic substances. This entry provides the readers with an overview of such AOPs in terms of fundamentals of the reaction mechanisms and their application to drinking water, wastewater, and soil/groundwater treatment processes.

REDUCTION AND OXIDATION

This section includes a brief technical discussion on the fundamentals related to the basic reduction and oxidation reactions.

Redox Reaction

A reduction and oxidation (redox, hereafter) reaction is an electron transfer reaction between an oxidizing agent and a reducing agent. Oxidizing agents (oxidants) are substances that cause oxidation, whereas reducing agents (reductants) are those that cause reduction. Losing an electron is oxidation and gaining an electron is reduction. Hence, oxidizing agents gain electrons or are reduced and reducing agents lose electrons or are oxidized. Oxidation state is a measure of the charge on an atom in any chemical substance. More information on the oxidation state of the substance is addressed in the following section.

A redox reaction can be separated into a reduction half-reaction and an oxidation half-reaction. In each of these reactions, the number of electrons lost or gained is equal to the change in oxidation state of the oxidized or reduced substances. Also, both reactions are needed to be properly balanced.[1,2]

Electrode Potential

The power of an oxidant or a reductant is measured by the electrode potential of the substance.[3] Under standard conditions, the electrode potential is defined as the standard electrode potential, E°. Standard conditions are the cases at 25°C, 1 atm pressure, and unit activity of all species. Table 1 lists the values of E° for some chemical oxidants used in water and wastewater treatment processes.

The standard free energy ΔG° is defined as follows:

$$\Delta G^{\circ} = -nFE^{\circ} = -RT \ln K \quad (1)$$

where n is the number of electrons transferred in the redox reaction, F is Faraday's constant (23,062.4 ± 0.3 cal/V/eq), R is the gas constant (1.9872 cal/deg/mol), and T is the absolute temperature (K). For example, when water containing manganese ions is ozonated under acidic conditions at 25°C, a reduction half-reaction and an oxidation half-reaction can be written, respectively, as follows:

$$5[O_3 + 2H^+ + 2e^- \rightarrow O_2 + H_2O] \quad (2)$$

$$2\left[Mn^{2+} + 4H_2O \rightarrow MnO_4^- + 8H^+ + 5e^-\right] \quad (3)$$

Consequently, the net redox reaction for such conditions is as follows:

$$5O_3 + 2Mn^{2+} + 3H_2O \rightarrow 2MnO_4^- + 6H^+5O_2 \quad (4)$$

For the first reduction half-reaction [Eq. (2)] E° is found to be 2.07 V, and for the second oxidation half-reaction [Eq. (3)] the value is obtained to be (−)1.49 V by reversing the sign of the E° value for the reduction half-reaction of permanganate under acidic conditions.[1,2] Therefore, the value of E° for the net redox reaction is the sum of two values [2.07 + (−)1.49 = 0.58 V]. The values of ΔG° and K of the net redox reaction are calculated from Eq. (1) to be (−)133 kcal/mol and 5×10^{97}, respectively, indicating that the reaction is very favored thermodynamically.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008088

Table 1 Standard electrode potentials for selected chemical oxidants

Oxidants	Standard electrode potentials (V)
Chloride	1.36
Hypochlorous acid	1.49
Permanganate	1.68
Hydrogen peroxide	1.78
Ozone	2.07
Hydroxyl radical	2.78

(From Refs.[1,2].)

At conditions where the thermodynamic activities of all substances are not unity (i.e., nonstandard conditions), such concentration effects on redox potential are expressed by the Nernst equation as follows:

$$E = E^\circ - \frac{RT}{nF} \ln \frac{a_{red}}{a_{ox}} \quad (5)$$

where a_{red} and a_{ox} represent the chemical activities of all of the species that appear on the reduced side and the oxidized side, respectively, in the redox reaction.

Analogous to pH, it is convenient to express the activity of electron as pE because electron activities may vary many orders of magnitude. In this sense, the value of pE (pE° if at standard conditions) is defined as follows:

$$pE^\circ = -\log(a_{e^-}) = \frac{E^\circ/2.303RT}{F} = \frac{E^\circ}{0.0591} \quad (6)$$

where a_{e^-} is the activity of electron in solution. Accordingly, if the electron activity were increased by a factor of 10, the pE value would be changed by (−)1.0. For example, if E° (or pE°) for Eqs. (2), (3), or (4) is positive, the reaction will proceed to the right direction as written because ΔG° becomes negative. The greater the positive value of E° or pE°, the greater the tendency of the reaction to proceed.

FUNDAMENTALS OF AOP

The use of chemical reagents with high oxidizing potential is the most effective way of oxidizing substances. The most highly oxidizing reagent available is the hydroxyl radical ($^\bullet$OH). In this regard, most AOPs are performed in conjunction with the generation of $^\bullet$OH to initiate oxidations.

A radical is a compound containing an atom with a single unpaired electron.[4] Structurally, $^\bullet$OH is a highly reactive radical because of its orbital characteristics. The four outermost orbitals of $^\bullet$OH have only seven electrons so that it has a great tendency to gain the eighth one to form the stable state. In the presence of an organic contaminant(s), $^\bullet$OH can abstract a hydrogen atom, thereby provoking contaminant oxidation. Such abstraction reaction is thermodynamically favored strongly, releasing about 119 kcal/mol of energy:

$$^\bullet OH + {}^\bullet H \rightarrow H_2O \quad \Delta H = (-)119\,\text{kcal/mol} \quad (7)$$

where ΔH is the enthalpy of the reaction. If the reaction has a negative ΔH, the reaction releases the heat of the reaction (i.e., exothermic reaction). Conversely, if the reaction has a positive ΔH, it is called an endothermic reaction.

The required dissociation energy (i.e., ΔH for breaking bond) for a C–H bond ranges from 80 to 104 kcal/mol, depending on the C–H location from which H is abstracted. In this regard, ΔH for the chemical oxidation via H abstraction by $^\bullet$OH is negative so that the reaction is energetically permitted.[4]

In this section, brief fundamental reaction mechanisms for each AOP are addressed. Included as AOPs are individual and combinational processes in the use of ultraviolet (UV) irradiation, catalyzed titanium dioxide oxidation, Fenton's reagent oxidation, ozonation, peroxone oxidation, and permanganate oxidation.

Photochemical Oxidation

Ultraviolet and visible lights have sufficient energy to alter the electronic configuration of a molecule that is known to be at its ground state. When a molecule absorbs light of energy $h\nu$, one of the two electrons with opposite spin occupying the same orbital can be promoted to a vacant orbital of higher energy. This phenomenon is called electron transition, allowing the molecule to be in an excited state. In some cases, the promoted electron maintains a spin opposite to that of its former partner, resulting in an excited singlet state. At other times, a spin is reversed, leading to an excited triplet state.[4,5]

The relative energies of the ground and excited states are depicted in Fig. 1. The easiest electronic transition is the excitation of a nonbonded electron (n) into a π antibonding orbital (i.e., $n \rightarrow \pi^*$ transition). The other one is excitation from a π bonding orbital to a π antibonding orbital (i.e., $\pi \rightarrow \pi^*$ transition). Excitation of an electron from a σ bonding molecular orbital to a σ antibonding orbital (i.e., $\sigma \rightarrow \sigma^*$ transition) is impractical. This is, for example, because in dilute aqueous solutions the solvent is present at much greater concentrations and would absorb most of the light.[4,5]

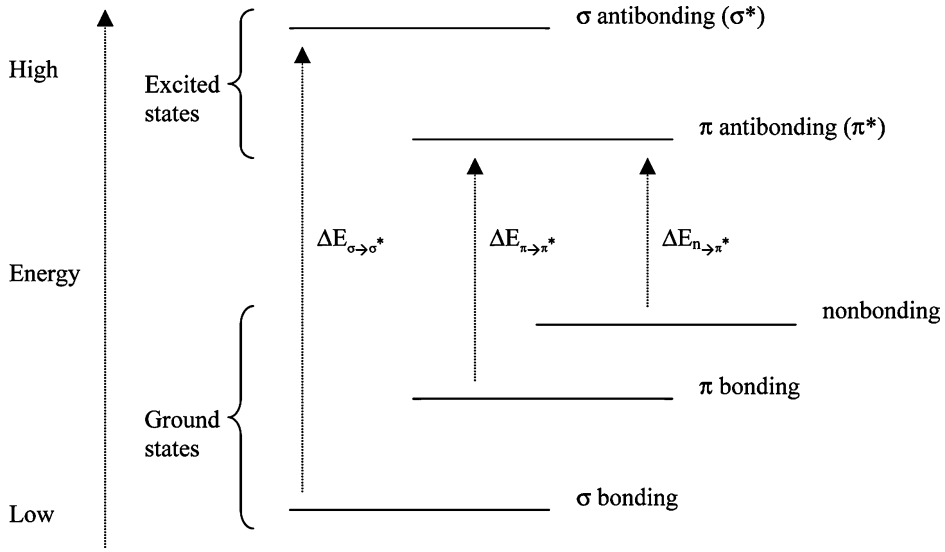


Fig. 1 The relative energies of the ground and excited states.

In general, a molecule in the ground state X absorbs $h\nu$ to produce an excited state Y^* or Z^* as follows:

$$X \xrightarrow{h\nu} Y^* \rightarrow \text{products} \quad (8)$$

$$Y^* \rightarrow Z^* \quad (9)$$

$$Z^* \rightarrow \text{products or X} \quad (10)$$

Typical reactions of excited states include rearrangement, abstraction, addition, and cleavage reactions.

Photocatalyzed Titanium Dioxide Oxidation

Titanium dioxide (TiO_2) has been the most commonly applied photocatalyst because of its chemical stability and low toxicity. The decomposition of organic contaminants in TiO_2 suspensions is initiated by photogenerated electron/hole pairs as follows:[6,7]

$$TiO_2 \xrightarrow{h\nu} TiO_2(e^- + h^+) \quad (11)$$

The holes (h^+) react with electron donors (H_2O and OH^-) or are adsorbed to TiO_2 to produce $^{\bullet}OH$ [Eq. (12)]. At the same time, dioxygen molecules react with the electrons to yield superoxide radical anions ($^{\bullet}O_2^-$) [Eq. (13)], which are in turn protonated to generate hydroperoxy radicals ($HO_2^{\bullet}$) [Eq. (14)].

$$h^+ + H_2O \text{ (or } OH_{surface}^-) \rightarrow {}^{\bullet}OH + H^+ \quad (12)$$

$$O_2 + e^- \rightarrow {}^{\bullet}O_2^- \quad (13)$$

$$^{\bullet}O_2^- + H^+ \rightarrow HO_2^{\bullet} \quad (14)$$

The surface of TiO_2 exhibits the positive charge because of an excess of protons from H_2O or in an acidic solution. In this regard, $^{\bullet}OH$ photogenerated on the TiO_2 surface acts as the major oxidant for the negatively charged contaminants that are attracted to the TiO_2 surface via coulombic forces. Contaminants can also be oxidized at the solution bulk phase as follows:

$$^{\bullet}OH + \text{compounds} \rightarrow \text{products} \quad (15)$$

Oxidation in Use of Hydrogen Peroxide

The simplest way for the $^{\bullet}OH$ generation is the direct photochemical cleavage of H_2O_2 as follows:[8]

$$H_2O_2 \xrightarrow{h\nu} 2{}^{\bullet}OH \quad (16)$$

However, the above theoretical amount of $^{\bullet}OH$ quantum yield can be reduced to about 0.5 because of the radical–radical recombination [Eq. (17)] and other scavenging compounds, such as H_2O_2 itself [Eq. (18)] and bicarbonate ions (HCO_3^-) [Eqs. (19–21)]

$$^{\bullet}OH + {}^{\bullet}OH \rightarrow H_2O_2 \quad (17)$$

$$^{\bullet}OH + H_2O_2 \rightarrow H_2O + HO_2^{\bullet} \quad (18)$$

$$^{\bullet}OH + HCO_3^- \rightarrow CO_3^{-\bullet} + H_2O \quad (19)$$

$$^{\bullet}OH + CO_3^{2-} \rightarrow CO_3^{-\bullet} + OH^- \quad (20)$$

$$CO_3^{-\bullet} + H_2O_2 \rightarrow HCO_3^- + {}^{\bullet}O_2^- + H^+ \quad (21)$$

Another common method of generating $^{\bullet}OH$ for the oxidation of organic compounds is via Fenton's reagent. Fenton's reagent oxidation involves the decomposition of hydrogen peroxide in the presence

of reduced iron salts into $^{\bullet}OH$.[9] Hydrogen peroxide (H_2O_2) reacts with ferrous iron (Fe^{2+}) to yield $^{\bullet}OH$ and ferric iron (Fe^{3+}) [Eq. (22)]; Fe^{3+} is reduced back to Fe^{2+} via reaction with H_2O_2 [Eq. (23)], the superoxide radical ($^{\bullet}O_2^-$) [Eq. (24)], or the perhydroxyl radical ($HO_2^{\bullet}$) [Eq. (25)], which is the protonated form of $^{\bullet}O_2^-$ ($pK_a = 4.8$). The primary reactions are as follows:

$$H_2O_2 + Fe^{2+} \rightarrow Fe^{3+} + OH^- + {}^{\bullet}OH \quad (22)$$

$$H_2O_2 + Fe^{3+} \rightarrow Fe^{2+} + {}^{\bullet}O_2^- + 2H^+ \quad (23)$$

$$^{\bullet}O_2^- + Fe^{3+} \rightarrow Fe^{2+} + O_2 \quad (24)$$

$$HO_2^{\bullet} + Fe^{3+} \rightarrow Fe^{2+} + HO_2^- \quad (25)$$

Hydroxyl radicals are highly reactive and are involved in nonspecific reactions with a wide range of compounds.[10] These reactions involve moderate to moderately high second-order reaction rate constants (10^7–$10^9\,M^{-1}\,sec^{-1}$) between the radical and compounds as shown in Table 2 and represent a promising option for the remediation of hazardous chemicals.[11] Fenton's reagent remediation is discussed in detail later in this entry.

Ozonation

Ozone (O_3) can directly oxidize organic compounds. Typical direct ozonation involves the insertion of the O_3 molecule into unsaturated carbon–carbon bond, resulting in the formation of an ozonide. This direct mechanism is very selective. The other mode of oxidation is through the reaction with $^{\bullet}OH$, which is generated via several O_3-based reactions.[12]

Ozone is very unstable in water with the half-life being in the range of seconds to hours. This is because the bonds that hold the O atoms together in O_3 molecule are very weak. The stability of ozone is largely dependent on the pH of the water. In fact, hydroxide ions play an important role and accelerate the O_3 decomposition as follows:

$$O_3 + OH^- \rightarrow HO_2^- + O_2 \quad (26)$$

$$O_3 + HO_2^- \rightarrow {}^{\bullet}O_2^- + {}^{\bullet}OH + O_2 \quad (27)$$

$$O_3 + {}^{\bullet}O_2^- \rightarrow {}^{\bullet}O_3^- + O_2 \quad (28)$$

Table 2 Rate constants for $^{\bullet}OH$ reactions with selected environmental contaminants

Reactants	Rate constants ($M^{-1}sec^{-1}$)
Benzene	7.8×10^9
2-Chlorophenol	1.2×10^{10}
Naphthalene	5×10^9
Nitrobenzene	3.9×10^9
Phenol	6.6×10^9
Tetrachloroethylene	2.6×10^9
Toluene	3.0×10^9
Trichloroethylene	4.2×10^9
Vinyl chloride	1.2×10^{10}

(From Ref.[10].)

Peroxone Oxidation

Ozone decomposition can be accelerated by the addition of H_2O_2. The reaction between H_2O_2 and O_3 is known to produce the $^{\bullet}OH$. This reaction is called peroxone.[13] The formation of the $^{\bullet}OH$ during peroxone oxidation is as follows:

$$H_2O_2 + H_2O \rightarrow HO_2^- + H_3O^+ \quad (29)$$

$$O_3 + HO_2^- \rightarrow HO_2^{\bullet} + {}^{\bullet}O_3^- \quad (30)$$

$$HO_2^{\bullet} \rightarrow H^+ + {}^{\bullet}O_2^- \quad (31)$$

$$O_2^{\bullet -} + O_3 \rightarrow {}^{\bullet}O_3^- + O_2 \quad (32)$$

$$^{\bullet}O_3^- + H^+ \rightarrow HO_3^{\bullet} \quad (33)$$

$$HO_3^{\bullet} \rightarrow O_2 + {}^{\bullet}OH \quad (34)$$

As such, the peroxone process results in the formation of $^{\bullet}OH$ through the reaction of O_3 with H_2O_2. A difference between the ozonation and peroxone process is that the former relies mainly on the direct oxidation by O_3, whereas the latter depends primarily on the oxidation with $^{\bullet}OH$. The O_3 residual in the peroxone process is short-lived because the O_3 decomposition is accelerated by the addition of H_2O_2, leading to a more reactive and faster oxidation in the peroxone process compared to the ozonation.

Permanganate Oxidation

Oxidation of organic chemicals using permanganate (MnO_4^-, either as potassium permanganate, $KMnO_4$ or sodium permanganate, $NaMnO_4$) involves, in general, direct electron transfer rather than free radical processes. The primary redox reactions for

permanganate are pH-dependent as follows:[14]

$$MnO_4^- + 8H^+ + 5e^- \rightarrow Mn^{2+} + 4H_2O \quad (\text{for pH} < 3.5) \tag{35}$$

$$MnO_4^- + 2H_2O + 3e^- \rightarrow MnO_2(s) + 4OH^- \quad (\text{for } 3.5 < \text{pH} < 12) \tag{36}$$

$$MnO_4^- + e^- \leftrightarrow MnO_4^{2-} \quad (\text{for pH} > 12) \tag{37}$$

The basic reactions involved in $KMnO_4$ oxidation in the presence of water are as follows:

$$RH_2 + 2KMnO_4 + 4H_2O \rightarrow ROH + 2MnO_2 + 2KOH \tag{38}$$

For example, the stoichiometric reaction for the complete destruction of trichloroethylene (C_2HCl_3) is as follows:

$$2KMnO_4 + C_2HCl_3 \rightarrow 2CO_2 + 2MnO_2 + 2K^+ + H^+ + 3Cl^- \tag{39}$$

Similar to the scavenging reactions in Fenton systems, in permanganate systems, there is a background oxidant demand that imposes a demand on the permanganate ion, which in turn reduces process efficiency. This demand, resulting from reaction with a wide range of reduced, naturally occurring chemical species, can often be greater than the demand imposed by the target contaminant to be oxidized. Permanganate is more expensive ($KMnO_4$—\$1.40/lb, \$162/1000 eq; $NaMnO_4$—\$5.95/lb, \$620/1000 eq) than H_2O_2 (\$0.26/lb; \$39/1000 eq). For further information on principles and practices of permanganate oxidation, the readers are highly recommended to refer to Ref.[14].

APPLICATION OF AOP

This section provides brief examples of the application of AOPs to the treatment of drinking water, wastewater, and soil/groundwater contaminated with unwanted and/or hazardous materials.

Drinking Water Treatment

The application of O_3 in drinking water treatment is prevalent because of its capability of disinfection and oxidation.[15] Ozone, as a disinfectant, is unstable in water and undergoes reactions with water components, whereas O_3 decomposes to $^{\bullet}OH$ so that advanced oxidation occurs. Unfortunately, undesired oxidation/disinfection by-products (DBPs) can be formed from the reaction of O_3 and $^{\bullet}OH$ with water matrix components. Such organic and inorganic DBPs can be treated with a subsequent treatment process, such as biological filtration. Care should be taken during ozonation of bromide-containing waters, which forms bromate that is not degraded by biological means. In addition, micro-organisms such as *Cryptosporidium parvum* oocysts are so resistant against disinfection that higher O_3 exposures are required and in turn more DBPs are formed.[11]

In an attempt to remove arsenic from potable municipal water and groundwater, Krishna et al. employed the treatment with Fenton's reagent followed by passing through iron scrap.[16] Their results indicated that the two-stage approach was capable of removing 2.5 mg/L arsenic to lower than the United States Environmental Protection Agency's (USEPA) guideline value of 10 μg/L.

Wastewater Treatment

Arslan et al. have investigated the AOPs for the treatment of reactive dye wastewater.[7] The investigators found that all investigated AOPs (e.g., UV/H_2O_2, photo-Fenton, UV/TiO_2) were capable of completely decolorizing and partially mineralizing the dye wastewater within 1 hr. Among those processes, photo-Fenton oxidation achieved the highest removal efficiency.

Wanpeng et al. have also applied a Fenton reagent oxidation for the treatment of a dye H-acid containing wastewater.[17] These investigators found that the process not only removed chemical oxidant demand effectively, but also improved the biodegradability by combining with the coagulation process.

Using a batch recycle mode, oily wastewater was treated in the photoassisted advanced oxidation systems.[18] Their results indicated that UV/H_2O_2 oxidized oily compounds into organic acids with the efficiency being greater at acidic pH. The oxidation rate was enhanced in the presence of Fe^{3+}.

Ozone, H_2O_2, and UV irradiation were applied separately and in all possible combinations for the treatment of textile wastewater.[19] Among the investigated AOPs, the most effective system was the simultaneous use of all three agents (i.e., $O_3/UV/H_2O_2$). Effective performance was also achieved in combined treatment O_3/H_2O_2. The authors pointed out that it was advisable to assess the costs of different AOPs to make practical comparison among the treatment schemes.[19]

Soil and Groundwater Treatment

One of the commonly used methods of treating soil and groundwater contaminated with hazardous

materials is the use of catalyzed hydrogen peroxide in the presence of iron species (i.e., Fenton's reagent oxidation). Watts et al. applied a higher H_2O_2 concentration than would be stoichiometrically required to oxidize pentachlorophenol (PCP) in soil systems to enhance mineralization.[20] With a pH of 2–3 being an optimum parameter, they achieved a rapid decrease in both PCP and total organic carbon contents within 24 hr.

Photochemical oxidation processes have also been used for the remediation of contaminated soil and groundwater environments. For example, Lewis et al. documented a field evaluation of the UV/oxidation technology ($UV/O_3/H_2O_2$) to treat volatile organic contaminants in groundwater.[21] Greater than 90% removal efficiency was reported for most contaminants.

Another AOP commonly used in in situ oxidation of contaminants in soil and groundwater environments is the permanganate oxidation. Although the process is active at most environmental pHs [Eq. (36)], Eqs. (36), (38), and (39) also indicate that manganese dioxide (MnO_2) is formed as a reaction by-product. Under most environmental conditions (pH 3.5–12), MnO_2 is present as a precipitate, i.e., $MnO_2(s)$. Under some conditions where background oxidant demand is high and large volumes and concentrations of MnO_4^- are needed, accumulation of large quantities of $MnO_2(s)$ in the subsurface can have negative consequences. For example, $MnO_2(s)$ encrustment on dense nonaqueous phase liquids (DNAPLs) may interfere with mass transfer of DNAPL components to the aqueous phase. Permeability reductions in the porous media may result from the formation of $MnO_2(s)$ on the well screen, sand/gravel pack, or aquifer material. Permeability loss can also result from the ion exchange of Na^+ or K^+ for divalent cations in the aquifer matrix. Undesirable reaction by-products may result from the oxidation and mobilization of metals or from the heavy metal content of the MnO_4^- product itself. Finally, there is a secondary drinking water standard established by the USEPA for manganese (0.05 mg/L) based on color, staining, and taste. Manganese imparts a black to brown color to water, can cause black staining on almost everything but glass, and imparts a bitter metallic taste.

Advanced Oxidation Processes as Pretreatment for Biological Treatment

Because of the presence of biorefractory materials, the efficiency of the biological treatments is often hampered. In general, AOP could reduce the toxicity that would otherwise detrimentally affect the biological processes. Additionally, AOP could transform biorefractory materials to more biodegradable compounds.[22,23]

As such, the use of AOP as pretreatment for the biological processes is promising. However, it should be noted that there is a great necessity to improve our understanding of such combined chemical and biological systems. This is because, for example, excess hydrogen peroxide concentration may show toxicity to the micro-organisms, and the by-products produced in the advanced oxidation process may sometimes be toxic to the micro-organisms as well.

FENTON'S REAGENT OXIDATION

This section introduces unconventional reaction mechanisms and the results of Fenton's reagent oxidation.

Hydroxyl-Radical-Independent Fenton Remediation

As mentioned earlier, the classical Fenton reagent oxidation constitutes the generation of $^{\bullet}OH$ as follows:[9]

$$H_2O_2 + Fe^{2+} \rightarrow Fe^{3+} + OH^- + {}^{\bullet}OH \qquad (22)$$

However, because of the small rate constant of the above Fenton reaction ($<10^2\,M^{-1}\,sec^{-1}$), the oxidation of Fe^{2+} by $^{\bullet}O_2^-$ [Eq. (40)] may completely overshadow the Fenton reaction and may have a significant role because of its much higher rate constant ($\sim 10^7\,M^{-1}\,sec^{-1}$).

$$^{\bullet}O_2^- + Fe^{2+} \rightarrow Fe^{3+} + H_2O_2 \qquad (40)$$

In Fe^{2+}-aerobic condition, an iron–oxygen complex, perferryl ion can be produced as an intermediate as follows:

$$Fe^{2+} + O_2 \leftrightarrow \left[\left(Fe^{2+}{-}O_2\right) \leftrightarrow \left(Fe^{3+}{-}{}^{\bullet}O_2^-\right)\right] \leftrightarrow Fe^{3+} + {}^{\bullet}O_2^- \qquad (41)$$

A common representation of perferryl ion is $Fe^{5+}{=}O$, with the formal charge of iron as (+)5. Its high electron affinity may replace $^{\bullet}OH$ as the oxidant.

Another iron–oxygen complex can be ferryl ion as follows:

$$^{\bullet}O_2^- + Fe^{3+} \rightarrow Fe^{2+} + O_2 \qquad (24)$$

$$Fe^{2+} + H_2O_2 \rightarrow Fe^{2+}O + H_2O \qquad (42)$$

$$Fe^{2+} + \left(Fe^{2+}{-}O_2\right) \rightarrow \left(Fe^{2+}{-}O_2{-}Fe^{2+}\right) \rightarrow \left(2Fe^{2+}{-}O\right) \qquad (43)$$

A common representation of ferryl ion is $Fe^{4+}{=}O$ or $Fe^{2+}{-}O$, with the formal charge of iron (+)4. Like perferryl ion, the high electron affinity of ferryl ion may also replace $^{\bullet}OH$ as the oxidant.

A chelating agent diethylenetriaminepentaacetic acid (DTPA) inhibits the reduction of Fe^{3+} by $^{\bullet}O_2^-$ [i.e., prevents the reaction in Eq. (24) from going to the right]. In comparison, ethylenedinitrilotetraacetic acid (EDTA) stimulates the reduction of Fe^{3+} by $^{\bullet}O_2^-$ [Eq. (44)]. Hence, DTPA and EDTA reduce and enhance the Fenton reaction, respectively [Eq. (22)].[24]

$$(Fe^{3+}-EDTA) + {}^{\bullet}O_2^- \rightarrow O_2 + (Fe^{2+}-EDTA) \quad (44)$$

$$H_2O_2 + Fe^{2+} \rightarrow Fe^{3+} + OH^- + {}^{\bullet}OH \quad (22)$$

Effect of pH on Fenton's Reagent Oxidation

Reaction pH has been observed to significantly influence the degradation of organic contaminants. In most cases, the optimum pH has been at 3.[20] At lower pH (i.e., pH < 2.5), the rate and extent of oxidation can decrease because of several reasons: (1) $[Fe(II)(H_2O)]^{2+}$ is formed, which reacts with H_2O_2 more slowly and therefore $^{\bullet}OH$ is produced less; (2) hydrogen ions can scavenge $^{\bullet}OH$; and (3) the reaction of Fe^{3+} with H_2O_2 is inhibited and therefore less Fe^{2+} is available for $^{\bullet}OH$ production.[10,11]

Despite pH 3 being recommended as the operating pH, the percentage of Fenton destruction of 2-methylnaphthalene, *n*-hexadecane, and diesel fuel was higher at pH 7 than at pH 2.[20,25] Beltran et al. also documented enhanced Fenton oxidation rates for polycyclic aromatic hydrocarbons (PAHs) at pH 7 than at pH 2.[26]

At pH higher than 8, Fe^{3+} exists mostly as hydroxy complexes like $FeOH^{2+}$ or $Fe(OH)_2^-$ or as insoluble Fe_2O_3 or $Fe(OH)_3$. In addition, because the oxidation rate of Fe^{2+} with oxygen is proportional to the square of the OH^- concentration [i.e., $(C_{OH^-})^2$], while that with H_2O_2 is only proportional to C_{OH^-}, the formation of Fe^{3+} complexes or hydroxides is favored at high pH and $^{\bullet}OH$ generation through the initiation reaction Eq. (22) or (45) reduces. In this regard, even slower and worse Fenton oxidation of PAHs was achieved at pH 10 or higher than pH 7.[26]

$$H_2O_2 + Fe^{2+} \rightarrow Fe^{3+} + OH^- + {}^{\bullet}OH \quad (22)$$

$$H_2O_2 + Fe(OH)^+ \rightarrow Fe(OH)_2 + 2{}^{\bullet}OH \quad (45)$$

Effect of Fe Source on Fenton's Reagent Oxidation

Grigoropoulou and Philippopoulos reported that ferric salts [$Fe_2(SO_4)_3$ and $FeCl_3$] provided better oxidation rates of phenolic compounds than ferrous ones [$FeSO_4$ and $Fe(NH_4)_2(SO_4)_2$], and the nature of the anions present did not affect the reaction rates.[27] However, the use of iron salts has been limited mainly because of formation of amorphous ferric hydroxide precipitate at neutral and high pH values at the concentrations required for effective Fenton oxidation.

Recent studies have showed that catalytic chemical oxidation can also occur by the interaction of H_2O_2 with iron oxide minerals, such as goethite (α-FeOOH). In model subsurface systems using goethite, Watts et al. reported an enhanced H_2O_2 decomposition and nitrobenzene oxidation at pH 7 than lower pHs.[28] In a heterogeneous Fenton-like reduction, Gurol and Lin reported a lack of the pH effect in the range of 5–9 on chlorobutane oxidation, although the decomposition rate of H_2O_2 was enhanced with increasing pH.[29]

Iron oxide surface can coordinate with protons and hydroxide ions to form different surface groups.[29]

$$\equiv Fe^{III}OH_2^+ \leftrightarrow \equiv Fe^{III}OH + H^+ \quad \left(pK_{a1}^S\right) \quad (46)$$

$$\equiv Fe^{III}OH \leftrightarrow \equiv Fe^{III}O^- + H^+ \quad \left(pK_{a2}^S\right) \quad (47)$$

where $\equiv Fe^{III}OH$ is the surface site of iron oxides, and $\left(pK_{a1}^S\right)$ and $\left(pK_{a2}^S\right)$ are the surface acidity constants for iron oxides. The protonated surface sites of iron oxides (i.e., $\equiv Fe^{III}OH_2^+$) are expected to dominate at pH below $\left(pK_{a1}^S\right)$, whereas the deprotonated surface (i.e., $\equiv Fe^{III}O^-$) would dominate at pH higher than $\left(pK_{a2}^S\right)$. Accordingly, H_2O_2 reacts faster with the negatively charged, deprotonated surface sites than the positively charged, protonated sites. As such, the rate of H_2O_2 decomposition and, consequently, the rate of $^{\bullet}OH$ generation was accelerated at a higher pH at the surface, but this did not necessarily improve compound oxidation because of a lesser number of target compound molecules being held on the surface sites at a higher pH.[29]

From the viewpoint of electron transfer phenomena, the decomposition of H_2O_2 can provide electrons [Eq. (48)]. By gaining these electrons, Fe^{2+} are produced from the reductive dissolution of goethite under acidic conditions [Eq. (49)]. Hydroxyl radicals are therefore produced by the classical Fenton reaction [Eq. (22)].

$$H_2O_2 \rightarrow 2H^+ + O_2 + 2e^- \quad (48)$$

$$\alpha FeOOH + 3H^+ + e^- \rightarrow Fe^{2+} + 3H_2O \quad (49)$$

$$H_2O_2 + Fe^{2+} \rightarrow Fe^{3+} + OH^- + {}^{\bullet}OH \quad (22)$$

The presence of Fe^{2+} can also be attributed to the redox reaction between the intermediate species from the compound oxidation and the surface sites of

the goethite. The following is an example from the 2-chlorophenol (2CP) oxidation.

$$2CP + {}^{\bullet}OH \rightarrow H_2O + {}^{\bullet}R \quad (50)$$

$${}^{\bullet}R + \alpha FeOOH_{(s)} \rightarrow Fe^{2+} + {}^{\bullet}RO + \alpha FeOOH_{(s)} \quad (51)$$

$$H_2O_2 + Fe^{2+} \rightarrow Fe^{3+} + OH^- + {}^{\bullet}OH \quad (22)$$

Therefore, 2CP oxidation reaction is sustained through Eqs. (50), (51), and (22).

Versatile Fenton's Reagent Oxidation

As mentioned previously, $^{\bullet}OH$ is highly reactive and is involved in nonspecific reactions with a wide range of environmental contaminants. These reactions involve moderate to moderately high second-order reaction rate constants between radical and target contaminants and represent a promising option for the remediation of hazardous pollutants.[11]

Superoxide radical, Fe^{2+}, and the hydroperoxide anion (HO_2^-) are potential reducing agents and may facilitate reductive transformations of the contaminants. Therefore, in a treatment system generally perceived to be oxidative, contaminant transformation by reductive pathways may also occur. Indeed, reductive transformation of chloroaliphatic compounds has been reported.[30] However, the specific mechanisms and the environmental conditions in which they are facilitated have not been identified or optimized yet for contaminant transformation.

At many hazardous waste sites, codisposal of contaminants has occurred that require both reductive and oxidative transformations. Under this set of conditions, neither reductive nor oxidative transformations alone are sufficient to achieve the treatment objective. Rather than employing a complex sequence of oxidative and reductive technologies, or vice versa, the Fenton mechanism, if optimized, may be capable of addressing both treatment needs.

CONCLUSIONS

The fundamental mechanisms involved in advanced oxidations and the application of AOPs to the treatment of environmental contaminants have been introduced. As outlined earlier, AOPs are broadly defined as the processes that generate $^{\bullet}OH$ in sufficient quantities to oxidize hazardous contaminants. The generation of $^{\bullet}OH$ is generally accelerated by combining individual (advanced) oxidation processes (e.g., O_3/H_2O_2, UV/H_2O_2, UV/Fenton). Because AOPs hold great promise of oxidizing refractory organic contaminants, extensive research efforts have been conducted in application to the treatment of drinking water, wastewater, and soil and groundwater contaminated with recalcitrant hazardous substances. However, more research is still needed for better understanding and control of the AOPs, which will lead to an improvement of treatment efficiency and costs.

ACKNOWLEDGMENT

This entry was initially written while the author (Dr. S. Hwang) held a National Research Council Research Associateship Award at the National Risk Management Research Laboratory of the U.S. Environmental Protection Agency, Ada, Oklahoma, U.S.A.

REFERENCES

1. AWWA. In *Water Quality and Treatment: A Handbook of Community Water Supplies*, 4th Ed.; McGraw-Hill, Inc.: New York, 1990.
2. Snoeyink, V.L.; Jenkins, D. *Water Chemistry*; John Wiley & Sons: New York, 1980.
3. Manahan, S.E. *Environmental Chemistry*, 6th Ed.; Lewis Publishers: Boca Raton, FL, 1994.
4. Bruice, P.Y. *Organic Chemistry*; Prentice Hall: Englewood Cliffs, NJ, 1994.
5. Oppenländer, T. *Photochemical Purification of Water and Air: Advanced Oxidation Processes (AOPs) Principles, Reaction Mechanisms, Reactor Concepts*; Vch Verlagsgesellschaft Mbh, 2003.
6. Watanabe, N.; Horikoshi, S.; Kawabe, H.; Sugie, Y.; Zhao, J.; Hidaka, H. Photodegradation mechanism for bisphenol A at the TiO_2/H_2O interfaces. Chemosphere **2003**, *52* (5), 851–859.
7. Arslan, I.; Balcioglu, I.A.; Tuhkanen, T.; Bahnemann, D. H_2O_2/UV-C and Fe^{2+}/H_2O_2/UV-C versus TiO_2/UV-A treatment for reactive dye wastewater. J. Environ. Eng. **2000**, *126* (10), 903–911.
8. Carey, J.H. An introduction to advanced oxidation processes (AOP) for destruction of organics in wastewater. Water Pollut. Res. J. Can. **1992**, *27* (1), 1–21.
9. Fenton, H.J.H. Oxidation of tartaric acid in presence of iron. J. Chem. Soc. **1894**, *65*, 899–910.
10. Buxton, G.V.; Greenstock, C.L.; Helman, W.P.; Ross, A.B. Critical review of rate constants for reactions of hydrated electrons, hydrogen atoms and hydroxyl radicals ($^{\bullet}OH/O^-$) in aqueous solution. J. Phys. Chem. Ref. Data **1988**, *17* (2), 513–886.
11. Neyens, E.; Baeyens, J. A review of classic Fenton's peroxidation as an advanced oxidation technique. J. Hazard. Mater. **2003**, *B98*, 33.

12. von Gunton, U. Ozonation of drinking water: part I. Oxidation kinetics and product formation. Water Res. **2003**, *37*, 1443–1467.
13. Kuo, C.-H.; Chen, S.-M. Ozonation and peroxone oxidation of toluene in aqueous solutions. Ind. Eng. Chem. Res. **1996**, *35*, 3973–3983.
14. Siegriest, R.L.; Urynowicz, M.A.; West, O.R.; Crimi, M.L.; Lowe, K.S. *Principles and Practices of in situ Chemical Oxidation Using Permanganate*; Battelle Press: Columbus, OH, 2001.
15. Gottschalk, C.; Libra, J.A.; Saupe, A. *Ozonation of Drinking Water and Wastewater*; Vch Verlagsgesellschaft Mbh, 2000.
16. Krishna, M.V.B.; Chandrasekaran, K.; Karunasagar, D.; Arunachalam, J. A combined treatment approach using Fenton's reagent and zero valent iron for the removal of arsenic from drinking water. J. Hazard. Mater. **2001**, *B84*, 229–240.
17. Wanpeng, Z.; Zhihua, Y.; Li, W. Application of ferrous-hydrogen peroxide for the treatment of H-acid manufacturing process wastewater. Water. Res. **1996**, *30* (12), 2949–2954.
18. Philippopoulos, C.J.; Poulopoulos, S.G. Photo-assisted oxidation of an oily wastewater using hydrogen peroxide. J. Hazard. Mater. **2003**, *B98*, 201–210.
19. Perkowski, J.; Kos, L.; Ledakowicz, S. Advanced oxidation of textile wastewaters. Ozone Sci. Eng. **2000**, *22*, 535–550.
20. Watts, R.J.; Udell, M.; Tauch, P.A.; Leung, S.W. Treatment of pentachlorophenol-contaminated soils using Fenton's reagent. Hazard. Waste Hazard. Mater. **1990**, *7* (4), 335–345.
21. Lewis, N.; Topudurti, K.; Foster, R.A. A field evaluation of the UV/oxidation technology to treat contaminated groundwater. Hazard Mater. **1990**, *3*, 42–55.
22. Beltran, F.J.; Garcia-Araya, J.F.; Alvarez, P. Impact of chemical oxidation on biological treatment of a primary municipal wastewater. 1. Effects on COD and biodegradability. Ozone Sci. Eng. **1997**, *19*, 495–512.
23. Bertanza, G.; Collivignarelli, C.; Pedrazzani, R. The role of chemical oxidation in combined chemical–physical and biological processes: experiences of industrial wastewater treatment. Water Sci. Technol. **2001**, *44* (5), 109–116.
24. Afanas'ev, I.B. *Superoxide Ion: Chemistry and Biological Implications*; CRC Press, Inc.: Boca Raton, FL, 1988.
25. Chen, C.T.; Tafuri, A.N.; Rahman, M.; Foerst, M.B. Chemical oxidation treatment of petroleum contaminated soil using Fenton's reagent. J. Environ. Sci. Health **1998**, *A33* (6), 987–1008.
26. Beltran, F.J.; Gonzalez, M.; Rivas, F.J.; Alvarez, P. Fenton reagent advanced oxidation of polynuclear aromatic hydrocarbons in water. Water, Air, and Soil Pollut. **1998**, *105*, 685–700.
27. Grigoropoulou, H.; Philippopoulos, C. Homogenous oxidation of phenols in aqueous solution with hydrogen peroxide and ferric ions. Water Sci. Technol. **1997**, *36* (2–3), 151–154.
28. Watts, R.J.; Foget, M.K.; Kong, S.-H.; Teel, A.L. Hydrogen peroxide decomposition in model subsurface systems. J. Hazard. Mater. **1999**, *B69*, 229–243.
29. Gurol, M.D.; Lin, S.-S. Hydrogen peroxide/iron oxide-induced catalytic oxidation of organic compounds. Water Sci. Technol. **2001**, *1* (4), 131–138.
30. Watts, R.J.; Bottenberg, B.C.; Hess, T.F.; Jensen, M.D.; Teel, A.L. Role of reductants in the enhanced desorption and transformation of chloroaliphatic compounds by modified Fenton's reactions. Environ. Sci. Technol. **1999**, *33* (19), 3432–3437.

Alkaline Zn–MnO_2 Batteries

Chung-Chiun Liu
Electronic Design Center and Department of Chemical Engineering, Case Western Reserve University, Cleveland, Ohio, U.S.A.

Xuekun Xing
NTK Powerdex Inc., Wixom, Michigan, U.S.A. and Department of Chemical Engineering, Case Western Reserve University, Cleveland, Ohio, U.S.A.

INTRODUCTION

The alkaline Zn/MnO_2 primary battery is one of the important aqueous electrolyte primary battery systems and has been widely and increasingly employed in consumer and industrial markets, since its commercialization in 1960. The total production volume of alkaline Zn-MnO_2 batteries approaches about 12 billion cells in 2001.[1] The alkaline Zn/MnO_2 battery can be regarded as a variant and new development of the traditional Leclanche cell, which consists of Zn anode, MnO_2 cathode and aqueous $NH_4Cl/ZnCl_2$ electrolyte. There are several major changes in the alkaline Zn/MnO_2 battery when compared with the Leclanche cell, which include electrolyte (change from acidic to alkaline, typically, ~30% concentrated KOH solution), the form of the cathode active material [change from chemical manganese dioxide (CMD) to electrolytic manganese dioxide (EMD) with higher purity], the form of the anode active material (change from zinc foil to zinc powder with large surface area), and the cell structure,[2–5] As a result of these changes, the alkaline Zn/MnO_2 battery demonstrates a number of advantages over the Leclanche cell, which include higher energy density, higher rate capability, longer shelf life or lower shelf-discharge rate, better dimensional stability,and others. Although these advantages are gained at the expense of higher initial cost for battery manufacture, they make the alkaline battery an attractive power source to replace at least in part the Leclanche battery. Today, alkaline Zn/MnO_2 batteries are being widely employed for various consumer products and electronic devices, such as toys, camera flash, light flash, radios, portable audio and video devices, and others.

CELL COMPONENTS AND CELL CHEMISTRY

Anode

The anode active material of the alkaline Zn/MnO_2 cell is zinc powder with median particle diameters of about 200 μm. Different from zinc foil used in Leclanche cells, zinc powders provide much higher surface area for the anode, which can carry substantially larger discharge current and, thus, effectively improves the battery rate capability. Actual anode is a compressed mixture of zinc powder, a polymer-based binder (such as polyacrylate or carboxymethyl cellulose) and the gelled electrolyte. The typical anode reaction in the concentrated KOH electrolyte can be described as follows:[6,7]

$$Zn + 4OH^- \rightarrow Zn(OH)_4^{2-} + 2e^- \quad (1)$$

It has been established that the actual anode reaction is more complicated and it takes place through the following mechanism: First, zinc is oxidized and reacts with hydroxide ions OH^- to form a transient intermediate zinc hydroxide $Zn(OH)_2$,[8] and then $Zn(OH)_2$ undergoes a dissolution process in the concentrated alkali solution to form zincate $Zn(OH)_4^{2-}$, namely,

$$Zn + 2OH^- \rightarrow Zn(OH)_2 + 2e^- \quad (2)$$

$$Zn(OH)_2 + 2OH^- \rightarrow Zn(OH)_4^{2-} \quad (3)$$

Although dissolved $Zn(OH)_4^{2-}$ can exist in concentrated KOH solutions at high concentrations, precipitation takes place with continuing discharge once $Zn(OH)_4^{2-}$ in the solution approaches a certain concentration in the alkaline electrolyte. Normally, ZnO is the precipitate in concentrated alkali solutions.[9] It is known that the solubility of ZnO in alkali solutions is only about one half of that of $Zn(OH)_2$,[10] and the overall reaction of the ZnO precipitation is as follows:

$$Zn(OH)_4^{2-} \rightarrow ZnO + H_2O + 2OH^- \quad (4)$$

This reaction takes place through a two-step mechanism: when $Zn(OH)_4^{2-}$ in the solution approaches a critical concentration at zinc surface

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007658

(exceeds saturation with respect to $Zn(OH)_2$), $Zn(OH)_2$ is formed:

$$Zn(OH)_4^{2-} \rightarrow Zn(OH)_2 + 2OH^- \quad (5)$$

Followed by a facile dehydration of $Zn(OH)_2$ in concentrated alkali solutions with ZnO being the precipitate, specially in the case of limited electrolyte:

$$Zn(OH)_2 \rightarrow ZnO + H_2O \quad (6)$$

In less concentrated alkali solutions, the rate of the dehydration decreases, and thus, $Zn(OH)_2$ becomes a main precipitate, especially in the case of excess electrolyte.[11]

Therefore, taking the ZnO precipitation reaction into account, one can describe the overall anode reaction of the alkaline Zn/MnO_2 cell as follows:

$$Zn + 2OH^- \rightarrow ZnO + H_2O + 2e^- \quad (7)$$

Cathode

The cathode active material of the alkaline Zn/MnO_2 battery is the EMD. Among various forms of MnO_2, gamma-MnO_2 is believed to be an active form. The EMD material has higher Mn content in its structure and higher purity, when compared with the CMD employed in Leclanche batteries, which helps reduce gassing and improve battery energy density. The actual cathode of the alkaline Zn/MnO_2 cell consists of EMD, carbon (usually graphite), binder, electrolyte, and selected additives. The cathode reaction at the first stage of the discharge is an one-electron transfer reaction to form manganese oxyhydroxide (MnOOH),[12] which can be described as follows:

$$MnO_2 + H_2O + e^- \rightarrow MnOOH + OH^- \quad (8)$$

At the second stage of the discharge, the produced MnOOH can be further reduced (discharged at a low discharge rate and a lower cut-off voltage) to Mn_3O_4:

$$3MnOOH + e^- \rightarrow Mn_3O_4 + OH^- + H_2O \quad (9)$$

which is equivalent to an additional 0.33 electron reduction per mole MnO_2.

Eq. (8) is a homogeneous proton-transfer reaction because MnOOH forms a solid solution with MnO_2, which results in a characteristic sloping discharge voltage profile for the first stage discharge of the alkaline Zn/MnO_2 cell, as shown in Fig. 1. Eq. (9) is a heterogeneous reaction because Mn_3O_4 has different structure from that of MnOOH, therefore, the voltage profile at this second discharge stage is quite flat. The further reduction of MnOOH is a complex process,

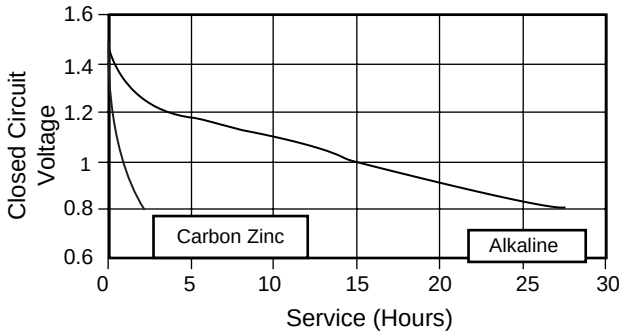


Fig. 1 Typical discharge curves of D size alkaline and carbon-zinc cells at high rate (500 mA). (From Ref.[14].) *(View this art in color at www.dekker.com.)*

and it is believed to place at least in part via the following reaction scheme:

$$MnOOH + H_2O + e^- \rightarrow Mn(OH)_2 + OH^- \quad (10)$$

A dissolution–precipitation mechanism for this reaction was proposed by Kozawa and Yeager.[13]

Cell Reactions

Combining both anode and cathode reactions, the overall cell reaction of the alkaline Zn/MnO_2 battery can be described as follows:

1. At the first stage of the discharge (one electron transfer per mole MnO_2):

$$2MnO_2 + Zn + H_2O = 2MnOOH + ZnO \quad (11)$$

2. The total cell reaction for 1.33 electron transfer per mole MnO_2:

$$3MnO_2 + 2Zn = Mn_3O_4 + 2ZnO \quad (12)$$

The open circuit voltage (OCV) of the alkaline Zn/MnO_2 cell is about 1.55 V at room temperature.

Electrolyte

Concentrated KOH aqueous solution (35–52%) is used for the electrolyte of the alkaline Zn/MnO_2 cell. Certain amount of ZnO is usually added into the electrolyte to suppress gassing process.[14] The electrolyte can be immobilized or gelled with addition of polymer materials, such as carboxymethyl cellulose. The major advantage of this electrolyte is that it has a high ionic conductivity, which helps improve battery rate

capability. In addition, this electrolyte is the source of water that acts as not only the solvent, but also as one of the reactants for the cell reactions, as discussed earlier.

Separator

Most commonly used separator materials for alkaline Zn/MnO_2 batteries are nonwoven polymers, such as cellulose, vinyl polymers, polyolefin, and others. The separator materials must be chemically stable in concentrated KOH solutions and electrochemically stable under both oxidizing and reducing conditions in the cell. In addition to its good electronic insulation, physical strength, and porous structure, good wettability to concentrated KOH solutions is especially crucial to provide a good ionic pathway for the battery operation.

Additives

In order to suppress or inhibit zinc corrosion and hydrogen evolution at zinc anode, a number of additives are selected and used for the anode. Because of its high over potential to the hydrogen evolution reaction, mercury is an effective gassing suppresser and used to be widely employed in alkaline Zn/MnO_2 batteries. In the case of using mercury as anode additive, zinc anode is usually amalgamated to certain level of mercury content. With increasing environmental concern on mercury, manufacturers of alkaline Zn/MnO_2 batteries have reduced mercury usage continuously and have developed mercury-free alkaline Zn/MnO_2 batteries, in which mercury is completely eliminated and the anode is made of zinc alloys with small amount of indium, bismuth, aluminum or calcium, instead. In addition, some organic additives have also been developed for the gassing inhibitor.

CELL CONSTRUCTION AND CELL DESIGN

There are two main cell constructions in consumer and industrial markets for alkaline Zn/MnO_2 batteries. The most common construction is the cylindrical cell, which is most widely being used for various industrial and consumer applications and it is generally categorized as D, C, AA, AAA sizes. Fig. 2 shows the cross-section of the typical construction of cylindrical cells. It is noted that the cylindrical steel can functions as both cathode current collector and the cell container. The inner surface of the can is usually plated with Ni or treated with conductive carbon coatings in order to improve contact with cathode mixture. On the other

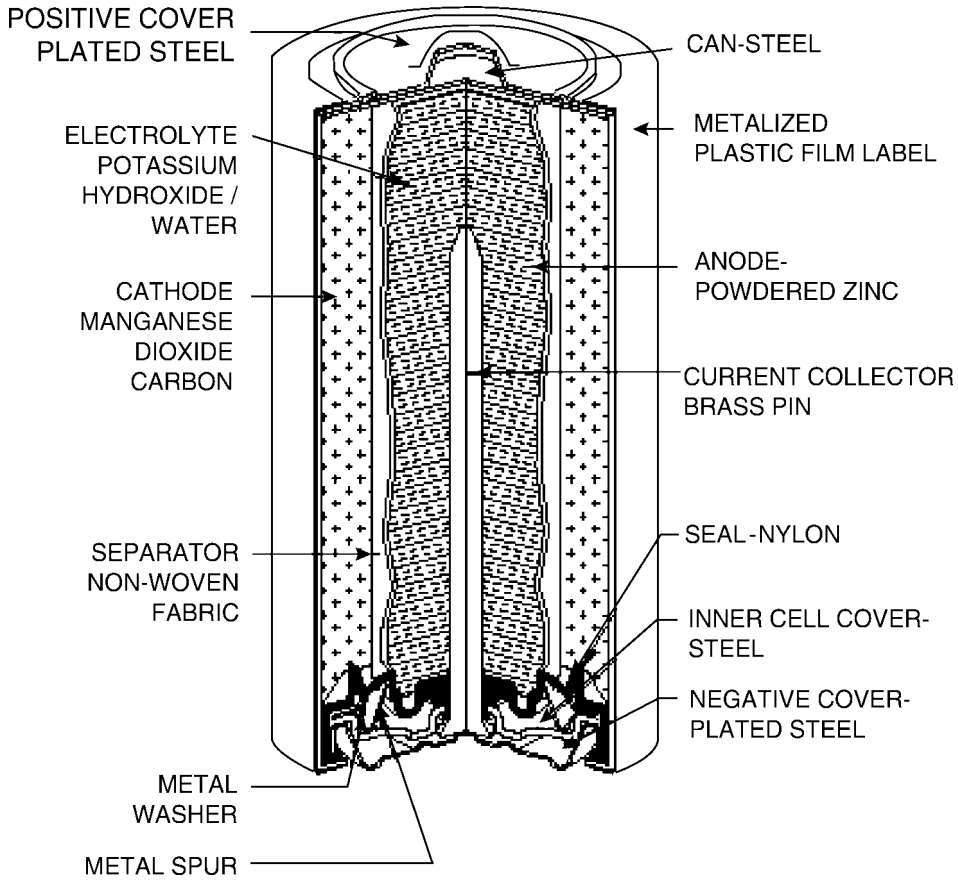


Fig. 2 Typical cell construction of the cylindrical alkaline cell. (From Ref.[14].)

hand, the anode current collector is usually made of brass strip or pins, and it is located in the center of the cylindrical can as shown in Fig. 2. The second major cell construction is the button cell, and its structure is schematically shown in Fig. 3. In this construction, there are two cups (anode cup at the top and cathode cup at the bottom, and both are made of stainless steel), a separator between them, and a thin plastic gasket to provide seal. The button cell structure is usually used for miniature cells that are widely being used as power sources of many low-drain devices such as memory back up, watches, calculators, and others. There are also alkaline Zn/MnO_2 batteries, typically, 9 V batteries, with a rectangular shape.

The alkaline Zn/MnO_2 cell is usually designed as ''anode-limited'' in terms of input capacity, which can prevent hydrogen gassing from the cathode, if the capacity of the cathode is consumed first otherwise.

CELL PERFORMANCE

General Performance Characteristics

The alkaline Zn/MnO_2 battery has a superior performance to the LeClanche battery. Table 1 summarizes a comparison of some characteristics between these two batteries.[15] Compared with Leclanche batteries, alkaline Zn/MnO_2 batteries have significantly higher energy density and power density because of its special chemical and structural characters, as discussed previously. The rate capability of alkaline Zn/MnO_2 batteries is especially superior to LaClanche. A comparison of discharge curves between these two systems is shown in Fig. 1.

Discharge Performance

Depending on applications, alkaline Zn/MnO_2 batteries can be discharged under different conditions, such as constant current, constant resistance, and constant power. In case of the constant current discharge, the capacity utilization can be affected by the applied current, and it decreases with increasing currents, as shown in Fig. 4 for Energizer's alkaline Zn/MnO_2 batteries. From Fig. 4, it is seen that the capacity utilization is relatively stable at low and medium currents, but it decreases rapidly when the current drain rate approaches 100 mA or higher. Besides, the discharge voltage decreases gradually with increasing current drain rates. On the other hand, the discharge capacity is also dependent on the cut-off voltage, with decreasing cut-off voltages, the capacity increases quite significantly, and this feature is also shown in Fig. 4.

Table 1 Typical performance characteristics of the alkaline and the LeClanche batteries

	Alkaline	LeClanche
Typical service capacity	30 mAh to 45 Ah	Several hundred mAh
Energy density, wh/liter	150–440	150
Specific energy, wh/kg	125–225	90
Operating temp. range	−18°C to 55°C	−5°C to 55°C
Storage temp. range	−40°C to 50°C	−40°C to 50°C
Low temp. performance	Good	Poor
Internal resistance	Very low	Moderate
Gassing	Low	Medium
% capacity loss per year at 0°C	1%	3%
% capacity loss per year at 20°C	3%	6%
% capacity loss per year at 40°C	8%	20%

Impact of Temperature on the Discharge

Normal operational temperature range of the alkaline Zn/MnO_2 battery is −18°C − +55°C (0°F–130°F).

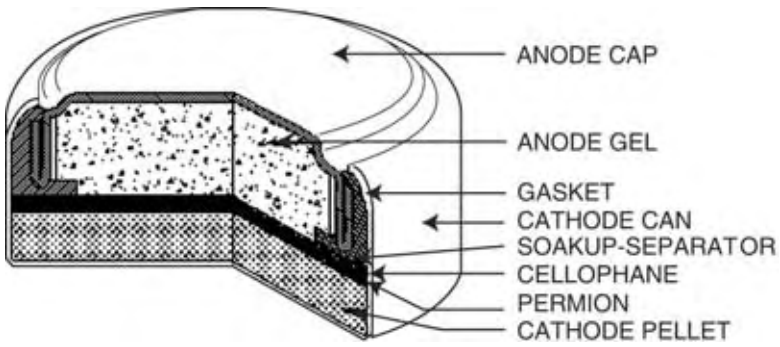


Fig. 3 Typical cell construction of the button alkaline cell. (From Ref.[14].)

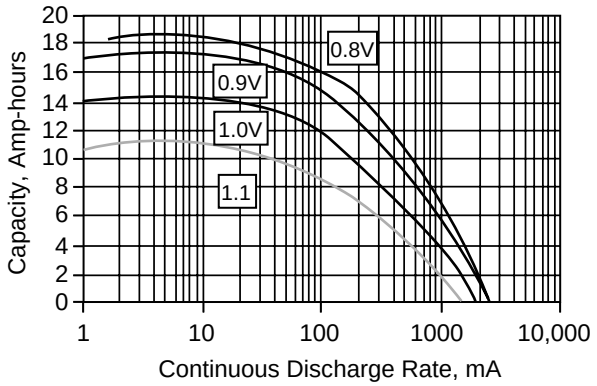


Fig. 4 Effect of drain rate and cut-off voltage on discharge capacity of alkaline cells. (From Ref.[14].) *(View this art in color at www.dekker.com.)*

In general, the capacity utilization decreases with decreasing operation temperatures, as shown in Fig. 5. The discharge voltage also decreases accordingly. Over-high operation temperature may adversely affect the battery sealing and, thus, the battery operational life. On the other hand, over-low operation temperature may lead to a significant increase in the battery internal resistance and, thus, affect discharge performance.

Impact of Temperature on the Self-Discharge and Storage Life

The self-discharge rate of the alkaline Zn/MnO_2 battery increases with increasing storage temperatures. Fig. 6 shows typical results of the impact of storage temperature on the cell capacity retention. The storage temperature of the alkaline Zn/MnO_2 battery is normally ranged in –40°C – +50°C, and it is preferable to store batteries at relatively lower temperatures, which provides a significantly longer storage time period compared with that at higher storage temperatures, while maintaining the same capacity retention, as shown in Fig. 7.

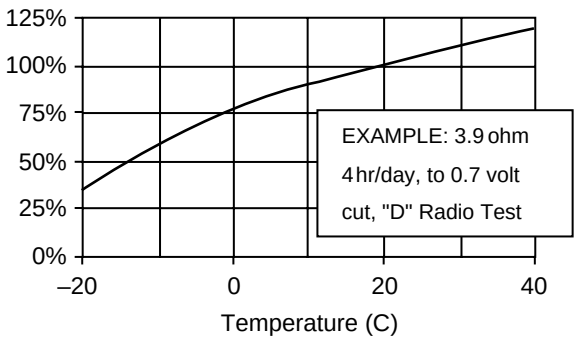


Fig. 5 Effect of temperature on capacity utilization of D size alkaline cells (3.9 ohm load). (From Ref.[14].)

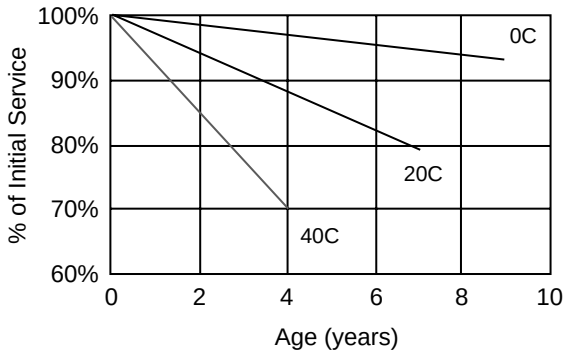


Fig. 6 Effect of storage temperature and time duration on relative service time of alkaline cells. (From Ref.[14].) *(View this art in color at www.dekker.com.)*

RECHARGEABLE ALKALINE Zn/MnO_2 BATTERY

The rechargeable (or reusable) alkaline Zn/MnO_2 battery has been mass-produced by Rayovac, since 1993. The technology was licensed from Kordesh's breakthrough work in late 1970s to 1980s, and even in the earlier work, which can be traced back to the development devoted by Union Carbide in 1970s. The rechargeable alkaline Zn/MnO_2 battery has the same cell reaction as that of the primary alkaline Zn/MnO_2 battery, described by Eq. (11) above, but the reaction must be restricted strictly to the one-electron transfer stage (the discharge cut-off voltage $\geq$0.9 V) to ensure rechargeability of the cathode MnO_2 material. The rechargeable alkaline Zn/MnO_2 battery has similar cell composition as that of the primary alkaline Zn/MnO_2 battery, namely, EMD cathode, zinc powder anode and concentrated KOH electrolyte, but it has a stronger two- or multi-layer separator to prevent internal short circuit caused by zinc dendrites, which may be formed during the charging process of the battery. The rechargeable alkaline Zn/MnO_2 battery has similar cell construction as that of the primary alkaline Zn/MnO_2 battery too, the most common cell construction is the cylindrical AA and AAA type.

The rechargeable alkaline Zn/MnO_2 battery has a normal cycle life of about 25 cycles at 100% DOD (depth of discharge), and the cycle life is much longer if shallow charge-discharge cycles (low DOD) are applied. The rate capability of the rechargeable Zn/MnO_2 battery is lower than its primary counterpart mainly because of its higher internal resistance. In general, this battery can be charged with constant current, constant voltage or pulse current modes, and the overcharge must be avoided. Compared with the rechargeable Ni–Cd batteries, the rechargeable alkaline Zn/MnO_2 battery has a number of advantages, such as lack of toxic and heavy metals, low cost,

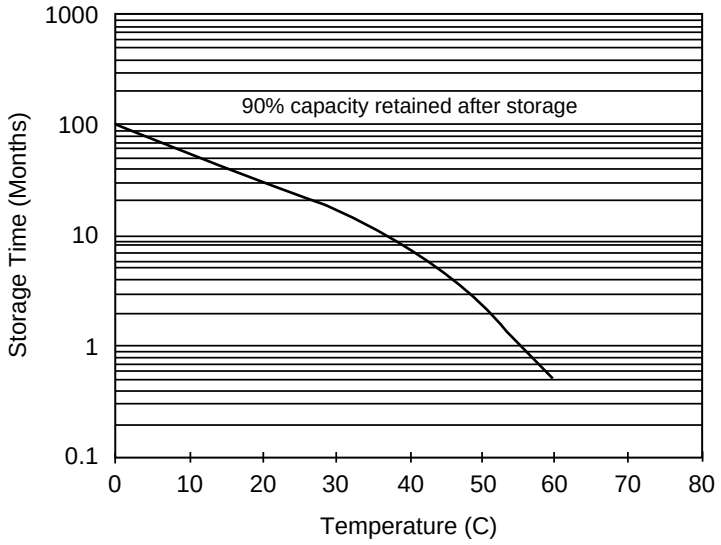


Fig. 7 Relationship between storage temperature and storage time period to keep 90% capacity retention for alkaline cells. (From Ref.[14].)

long shelf life, and lack of the memory effect. The main disadvantages of the rechargeable Zn/MnO_2 battery are its short cycle life and relatively low drain rate.

CONCLUSIONS

The alkaline Zn/MnO_2 battery is one of well-developed battery systems in the industrial and consumer markets, and is still undergoing continuous developments to improve its performance. The developments include: the improvement in active materials and its purity to enhance battery energy capacity; the improvement in the cell structure and cell composition to enhance battery power density; the improvement in sealing materials and the sealing structure to enhance the battery storage and operational life, and to further prevent leakage.

REFERENCES

1. Brodd, R. Private communications, 2002.
2. Falk, S.U.; Salkind, A.J. *Alkaline Storage Batteries*; Electrochemical Society Services; Wiley: New York, 1969.
3. Barak, M. *Electrochemical Power Sources*; Peter Peregrinus: London, 1980.
4. Linden, D. *Handbook of Batteries*, 2nd Ed.; McGraw Hill: New York, 1995.
5. Vincent, C.A.; Scrosati, B. *Modern Batteries*, 2nd Ed.; John Wiley & Sons: New York, 1997.
6. Dirkse, T.P. The nature of the zinc-containing ion in strongly alkaline solutions. J. Electrochem. Soc. **1954**, *101*, 328.
7. Boden, D.P.; Wylie, R.B.; Spera, V.J. The electrode potential of zinc amalgam in alkaline zincate solutions. J. Electrochem. Soc. **1971**, *118*, 1298.
8. Payne, D.A.; Bard, A.J. The mechanism of the zinc(II)-zinc amalgam electrode reaction in alkaline media as studied by chronocoulometric and voltammetric techniques. J. Electrochem. Soc. **1972**, *119*, 1665.
9. Dirkse, T.P. Electrolytic oxidation of zinc in alkaline solutions. J. Electrochem. Soc. **1955**, *102*, 497.
10. Schumacher, E.A. Chapter 3, Primary cells with caustic alkali electrolyte. In *The Primary Battery*; Heise, G.W., Cahoon, N.C., Eds.; Wiley: New York, 1971; Vol. 1, 169.
11. Nikitina, Z. Passivation of a zinc electrode in galvanic elements with alkaline electrolytes. Ya Zh. Prikl. Khim. **1958**, *31*, 218.
12. Kozawa, A.; Powers, R.A. Cathodic reduction of β-manganese dioxide and γ-manganese dioxide in ammonium chloride and potassium hydroxide electrolytes. Electrochem. Technol. **1967**, *5*, 535; The manganese dioxide electrode in alkaline electrolyte; the electron-proton mechanism for the discharge process from MnO_2 to MnO_{xt}. J. Electrochem. Soc., **1966**, *113*, 870.
13. Kozawa, A.; Yeager, J.F. The cathodic reduction mechanism of electrolytic manganese dioxide in alkakline electrolyte. J. Electrochem. Soc. **1965**, *122*, 959; Cathodic reduction mechanism of MnOCH to $MN(OH)_2$ in alkaline electrolyte. **1968**, *115*, 1003.
14. Bockris, J.O'M.; Nagy, Z.; Damjanovic, A. Deposition and dissolution of zinc in alkaline solutions. J. Electrochem. Soc. **1972**, *119*, 285.
15. Adapted from Energizer Alkaline Application Manual (accessed from 2002).

Alkylation Processes to Produce High-Quality Gasolines

A

Lyle F. Albright
School of Chemical Engineering, Purdue University, West Lafayette, Indiana, U.S.A.

James M. Ryan
Ryan Consulting, Inc., Fort Myers, Florida, U.S.A.

INTRODUCTION

Isobutane is alkylated with C_3–C_5 olefins to produce the highest-quality and cleanest-burning gasolines. More than 1.1 million barrels of alkylate/day are produced in the United States (P. Pyror, personal communication). The amounts of alkylate produced in the remainder of the world are considerably less, but they are increasing at a significant rate. In the United States, about 11–13% of the total gasoline pool is alkylate. This percentage depends on the location in the United States and on the season. Alkylate production will likely increase substantially in the future.

Alkylation was first practiced for gasoline production about 60 yr ago. At that time, most of the alkylate was used as fuel for the airplanes used in World War II. Four quite distinct reactors were developed in which isobutane and olefins were introduced as liquids to the reactor. In the reactor, the hydrocarbon liquids are contacted with either liquid sulfuric acid or liquid hydrofluoric acid (HF), which acts as a catalyst. Dispersions of these two relatively immiscible liquids are formed. The alkylate product formed is a mixture of mainly C_5–C_{16} isoparaffins. Alkylate products often have research octane numbers (RONs) varying from 93 to 98 (the motor octane numbers tend to be two to three units lower).

The alkylates and by-products produced using sulfuric acid and HF have compositions that differ in several respects, as discussed later. About 50 yr ago, considerably more alkylate was produced using sulfuric acid. For the next 30–35 yr, the relative importance of the two processes using HF as catalyst increased until essentially equal amounts of alkylate were produced with each acid. In the late 1980s, however, two events occurred that indicated safety concerns with HF needed to be re-examined. First tests were conducted in which 1000 gallons of liquid HF were ejected into the atmosphere.[1] An aerosol cloud formed that contained lethal levels of HF at ground level for 5–6 mi downwind. Alkylation units at the time sometimes had much larger liquid HF inventories. Second, a major release of gaseous HF occurred in an American refinery as the result of an accident. No human fatalities occurred probably because the release was gaseous HF and no aerosol cloud formed. Many residents living within several blocks of the refinery, however, required medical care. Liquid HF often results in serious skin burns and even damage to bones. Also in the late 1980s, two workers in an alkylation plant were killed because of an HF release. Since that time, no grassroot alkylation plants have been built in the United States, but several with sulfuric acid have been.

In current processes that use either sulfuric acid or HF, isobutane in large excess and olefins are introduced as liquids into the reactor. After completion of the reactions, the liquid–liquid dispersions are separated by decanting. The alkylate product is separated by distillation or stripping from the unreacted isobutane, which is recirculated to the reactor. This entry reviews the chemistry, physicochemical phenomena, current processes, and finally suggests methods to improve significantly the alkylation process.

CHEMISTRY OF ALKYLATION

Production of Alkylate: The basic chemistry that explains the wide range of qualities (or octane numbers), by-products and side reactions, and yields has been reported.[2–5] The following two reactions have often been used to represent the overall process:

$$\textit{tert}\text{-butyl cation } (t\text{-}C_4H_9^+) + \text{isobutylene (or a 2-butene)} \rightarrow \text{trimethylpentyl cation } (TMP^+) \quad (1)$$

There are four different TMP^+s.

$$TMP^+ + \text{isobutane} \rightarrow \text{trimethylpentane (TMP)} + t\text{-}C_4H_9^+ \quad (2)$$

The overall chemistry is much more complicated, however, and these two equations have been misinterpreted. For example, they suggest that isobutane and the olefins react by a chain sequence. Olefins and, especially, isobutylene and i-C_5 olefins, react instead prior to and more rapidly than isobutane.[6] Although

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120015543

isobutane contributes the hydride ion (H^-) that converts TMP^+ (and other isoalkyl cations) to TMP (and C_5–C_{16} isoparaffins), the transfer is not by the simple chemistry of Eq. (2).[7] Extensive experimental evidence indicates that the H^- from isobutane is predominately transferred first to conjunct polymers dissolved in sulfuric acid or HF; then the H^- is transferred to TMP^+ or other isoalkyl cations located at or at least near the interface between the acid and hydrocarbon liquid phases in the dispersion. Isobutane and olefins often react on essentially an equimolar basis, but as indicated later more moles of isobutane often react than those of olefins, especially when HF is the catalyst.[8] The above two equations fail to explain the C_8 isoparaffins produced with 1-butene. When 1-butene reacts with a t-$C_4H_9^+$, a dimethylhexyl cation (DMH^+) is produced that is then converted to a dimethylhexane (DMH) isomer. When HF is used as the catalyst, DMHs are indeed major products. With sulfuric acid, TMPs are, however, the major C_8 family produced. The above equations also provide no explanation for the production of lighter and heavier isoparaffins.

Four reaction sequences (mechanisms) actually occur to produce C_5–C_{16} isoparaffins, as discussed next.[2,7,8]

Mechanism 1: This mechanism is the only true alkylation sequence. Only C_7, C_8, or C_9 isoparaffins are produced when isobutane is alkylated with propylene, C_4 olefins (four isomers), or C_5 olefins (six isomers), respectively. Important intermediates formed include isoalkyl sulfates because sulfuric acid reacts with much of either propylene, n-C_4 olefins, or n-C_5 olefins. Hydrofluoric acid also reacts but to a much reduced extent with n-olefins producing isoalkyl fluorides. When n-C_4 olefins are used forming sec-butyl sulfates (or sec-butyl fluoride), the sulfates (or fluorides) react at preferred operating conditions with isobutane to form TMPs in high yields with the regeneration of sulfuric acid (or HF). The intermediate sec-butyl sulfates explain why both 2-butenes and 1-butene produce essentially identical high-quality alkylates when sulfuric acid is the catalyst. For alkylations using HF, considerably lower-quality alkylates containing more DMHs are produced when 1-butene is part of the olefin feed.

Much smaller fractions of isobutylene and i-C_5 olefins react via mechanism 1 as compared to n-olefins. *Iso*-olefins instead react in substantial amounts via mechanisms 2 and 3.

Mechanisms 2 and 3: These two mechanisms are related in that two or more olefins react initially with a single t-$C_4H_9^+$ forming i-C_{10}^+ to i-C_{20}^+s. For mechanism 3, a H^- (generally from conjunct polymers) transfers to the heavy cations to produce C_{10}–C_{20} (and more generally C_{10}–C_{16}) isoparaffins. With mechanism 2, which is always of greater importance than mechanism 3, the heavy cation fragments, because they are relatively unstable, form i-C_4 to i-C_{16} cations and olefins. These latter cations and olefins are converted to i-C_4 to i-C_{16} isoparaffins after suitable H^- transfer, as already discussed, plus H^+ transfer (from the acid). Three points need to be emphasized: first, light isoparaffins such as i-C_5 to i-C_7s are produced (this is frequently the only way such isoparaffins are produced when pure C_4 olefins are used); second, some isobutane is also produced, as has been indicated by tagged carbon research; third, production of DMH is relatively high, which explains the rather high concentrations produced when isobutylene is the olefin feed.[4]

Mechanism 4: This mechanism is always of major importance when HF is the catalyst, but is generally of little or no importance with sulfuric acid. The reaction sequence is complicated, but the overall reaction is often reported in an oversimplified manner as follows:[9]

$$\text{2-isobutane} + n\text{-olefin} \rightarrow n\text{-paraffin} + i\text{-}C_8H_{18} \quad (3)$$

Several points need to be emphasized. First, the i-C_8H_{18} (frequently indicated in the literature as TMP) is really a mixture of C_5–C_{16} isoparaffins, often with RON values in the 93–94 range. This mixture (or alkylate) is formed basically by mechanism 2 reactions. Second, when n-olefins (propylene, n-butenes, and n-pentenes) are used, the light n-paraffins formed are propane, n-butane, or n-pentane, respectively; none are suitable in the gasoline pool. Third, isobutane consumption per production of a given quantity of alkylate is often increased by 6–10% when HF is employed because of the importance of mechanism 4, as compared to little or most likely no importance for alkylations with sulfuric acid.[8] Fourth, C_5 olefins are not usually used in the feedstocks when HF is the catalyst because of the large amounts of isopentane and n-pentane produced; further, isobutane consumption increases.

With HF as catalyst, mechanism 4 can be promoted by higher temperatures and also higher isobutane/olefin ratios.[9] Plant operators using HF often deliberately promote mechanism 4 for alkylations using propylene or 1-butene to produce higher-quality alkylates and more TMPs. Operating costs, however, increase because of increased isobutane consumption and higher costs to recover and recycle the additional unreacted isobutane.

Conjunct Polymer Formation: The polymers (also called acid-soluble oils or red oil) dissolve in the acid phase (sulfuric acid and HF). They have a hydrogen/carbon atomic ratio (H/C) of about 1.75 and a molecular weight of 270–325.[10] They contain numerous –C=C– bonds, which react to a significant extent with sulfuric acid to produce conjunct acid sulfates or with HF, conjunct polymer fluorides. These sulfates (or fluorides)

serve as surfactants and some collect at the liquid–liquid interfaces. As they contain numerous tertiary C–H bonds, they are the main source of H^-s involved in the production of alkylate. Laboratory evidence indicates that these polymers are formed mainly from olefins; pseudoalkylates are also simultaneously produced.[11]

Pseudoalkylate: Using C_4 olefins as an example, *sec*-butyl acid sulfates are first formed; they dissolve in the sulfuric acid phase.[11] When the resulting acid mixture is allowed to warm to room temperature, almost identical amounts of conjunct polymers and pseudoalkylates are produced within several minutes. The following two simplified equations plus carbon and hydrogen balances support this finding:

$$n\text{-}C_4H_8(H/C = 2) + \text{sulfuric acid} \rightarrow \textit{sec}\text{-butyl acid sulfate} \quad (4)$$

$$\textit{sec}\text{-butyl acid sulfate} \rightarrow \text{conjunct polymers sulfate } (H/C = 1.75) + \text{pseudoalkylate (H/C ratio of 2.25)} \quad (5)$$

The RON values of the pseudoalkylate are in the low 80s. Both conjunct polymers and pseudoalkylate are also produced when light olefins are bubbled through sulfuric acid.

Hydrocarbon layers frequently form above used acids in storage tanks. This layer is with high probability partly, if not mainly, a pseudoalkylate produced when isoalkyl acid sulfates dissolved in the used acid decompose. In at least two storage tanks, serious explosions of hydrocarbon–air mixtures have occurred.[12,13]

Decomposition of TMPs and Other Isoparaffins: TMPs and other isoparaffins in alkylates contain tertiary C–H bonds, which permit degradation reactions. First, sulfuric acid acts as an oxidant to start a series of reactions that convert TMP into a mixture of C_4–C_{16} isoparaffins.[14] Second, because of their tertiary C–H bonds, TMPs react (are actually alkylated) with C_3–C_5 olefins; mechanism 2-type reactions are predominant. Both types of reactions lower the quality and the quantity of the alkylate.

Oxidation Reactions in Used Sulfuric Acid: Sulfuric acid reacts with conjunct polymer dissolved in the acid.[15] Sulfur dioxide and water are produced. Such oxidation has occurred on occasion with large releases of sulfur dioxide from storage tanks or railroad tank cars, especially in summer when the acid heats up.

PHYSICOCHEMICAL PHENOMENA RELATIVE TO ALKYLATION

The interfacial area between the acid/hydrocarbon dispersions is the location at which most, if not all, of the C_5–C_{16} isoparaffins are produced.[5,16,17] In industrial units, the preferred dispersions are acid-continuous as compared to hydrocarbon-continuous. Albright and am Ende found at least with sulfuric acid that acid-continuous dispersions tend to have significantly larger interfacial areas.[18] Some important experimental findings related to physicochemical phenomena are discussed next.

Increased levels of agitation that result in increased interfacial areas were found in alkylation experiments using 2-butenes and sulfuric acid to increase the RON value; in one comparison, the RON increased from 86.2 at 1000 rpm to 98.7 at 4000 rpm.[19] Increased agitation also increases the kinetics of alkylation. In one case, the optimum residence time of reactants in the reactor was decreased to about 5 min. In commercial reactors, residence times are often in the 20–30 min range. In commercial reactors using HF as the catalyst, the alkylation reactions are often completed in 10–15 sec or even less; two factors that promote these fast reactions are higher operating temperatures and especially much larger interfacial areas.[3]

In addition to agitation, the interfacial areas are significantly affected by the following: acid/hydrocarbon ratio, acid composition (and especially the amounts of dissolved conjunct polymers), and temperature.[3,18] Conjunct polymers are surfactants that collect in appreciable concentrations at the interface. Here, they act as a reservoir of H^-s in the transfer steps from isobutane or other isoparaffins to the i-C_5 to i-C_{16} cations. Conjunct polymers also have a major effect on the viscosity and other physical properties of the acid phase. In the alkylation reactor, the preferred sulfuric acid and HF phases contain appreciable amounts of conjunct polymers; optimum amounts result in higher RON values, higher yields, less by-products, etc.

At least with sulfuric acid, key intermediates transfer from and across the interfaces. For example, *sec*-butyl acid sulfate transfers into the acid phase, whereas di-*sec*-butyl sulfate transfers into and dissolves mainly in the hydrocarbon phase. No information is available on the fraction of isoalkyl fluorides in the two phases. This fraction likely depends on the amount of conjunct polymers in the HF phase.

In one reactor, the sulfuric acid/hydrocarbon ratio in the dispersion varied substantially with location.[20] Variations occurred depending on the distance from the agitation impeller and the height. Clearly, droplets were coalescing and fragmenting in the reactor; in addition, partial separation of the two phases plus later redistribution was occurring. Obviously, the average residence times of the acid and of the hydrocarbon phase in commercial reactors often differ significantly. In most, if not all, commercial reactions, the average residence time of the acid phase is greater.

Because the main alkylation reactions occur at the interface, both isobutane and olefins in the dispersed droplets are transferred to the interface, and the resulting C_5–C_{16} isoparaffins are transferred from the interface back into the droplet.[5,16,17] Experimental data indicate that such transfer steps are in part at least rate controlling steps. In any case, each droplet acts as a different reaction zone (basically a separate mini-reactor). As droplets of different compositions and sizes occur in all commercial reactors, the alkylation results differ in various droplets, i.e., different alkylates, RONs, yields, amounts of by-products, etc. Improved results would occur if alkylation reactors could be designed and operated so that all the alkylate was produced only at optimal conditions.

FEEDSTREAMS FOR ALKYLATION

Isobutane and light olefins are the desired hydrocarbon feeds. Unfortunately, impurities such as acetylenes, dienes, sulfur- and oxygen-containing hydrocarbons, cyclopentene, and water are also often present. Purification of the feeds is expensive, but is sometimes cost-effective as a means of reducing the buildup in the acids of conjunct polymers. Dry hydrocarbon feeds are preferred, especially with HF. The water transfers to the HF and is a concern relative to metal corrosion. Solid adsorbents are often used for drying of feedstocks.

The feed ratios of isobutane to olefins with sulfuric acid tend to be lower, such as 6 : 1 to 8 : 1. When HF is used, the ratios are often 10 : 1 to 15 : 1; higher ratios then promote both mechanisms 1 and 4. Higher ratios, however, increase operating costs substantially with regard to the recovery and recycling of the unreacted isobutane. Isopentane, which is also readily alkylated, is sometimes deliberately added to supplement isobutane. Much lower-quality alkylates are produced, especially when propylene is a feed olefin, in which case, dimethylhexanes are produced in large amounts.

The compositions (or concentrations) of both the feed acid and the acid present in the reactor are of major importance. The feed sulfuric acids are frequently in the 98.5–99.5% range. The more concentrated feed acid results in the production of more alkylate per unit weight of feed acid.[21] The preferred acid concentration in the reactor varies from about 90% to 95% depending especially on the composition of the olefin feeds. Stronger acids are preferred with propylene-rich olefins, whereas the lower concentrations are preferred with C_5-rich olefins. Less information is available with HF catalysts. Eastman et al. report the HF compositions in a Phillips reactor as 82–85.5% HF, 8–12% acid-soluble oils (or conjunct polymers), and 0.1% water.[22] The unaccounted material in this analysis is probably HF that has reacted to form conjunct polymer fluorides.

In commercial reactors, the acid phase is recirculated numerous times. A small amount of the used acid is removed and sent to the regeneration unit. A similar amount of feed acid is added to the recycle stream. At least with sulfuric acid, major reductions of acid consumption can be obtained when two and preferably more reactors are used in a refinery. Proper arrangements of the acid flows, method of adding different olefins, and adjusting operating conditions in the different reactors can often result in the reduction of acid composition from about 0.5–0.8 to 0.25–0.3 lb/gal.

REACTORS USING SULFURIC ACID

Two types of reactors have been used for over 50 yr. Modifications have occurred with time.

Stratco Reactors: These reactors designed by Stratco, Inc. produce about 34% of the alkylate produced worldwide (P. Pyror, personal communication). In 2003, Stratco was purchased by E.I. du Pont de Nemours, Inc. A reactor, often called a Contactor, is a horizontal cylindrical vessel, as shown in Fig. 1, with the following features.[21,23] An impeller at one end recirculates the acid/hydrocarbon dispersion many times (on average) over a U-tube bundle to regulate the temperature of the dispersion at about 5–10°C. A cylindrical circulation shell located in the reactor provides an annular space so that the dispersion flows from one end of the reactor to the other, where it makes a 180° turn and flows back over the tube bundle.

Injection devices are provided at or near the eye of the impeller for the hydrocarbon feed mixture of isobutane and olefins and for the acid feed (mostly recycled acid). A small fraction (mix to settler or decanter) of the dispersion flowing in the annular region is removed from the reactor through the outlet line positioned on top at the other end (relative to the impeller). The average time that the dispersion remains in the reactor is about 20 min, but obviously there is a wide range of times.

After the dispersion leaving the reactor is separated in the settler (decanter), the liquid hydrocarbon product stream is partially flashed, forming a vapor phase, mainly isobutane. The remaining liquid is hence cooled, and it is used as the coolant in the tube bundle of the reactor. As heat is transferred there, more hydrocarbons vaporize, forming a liquid–gas mixture. Obviously, temperature gradients occur on both sides of the tube bundle as a function of reactor length. Heat transfer (and temperature control) is an important design consideration in contactors.

In addition to temperature variations in the reactor, there are variations throughout the reactor in the composition and size of dispersed droplets, and also different acid/hydrocarbon ratios. Stratco has rather

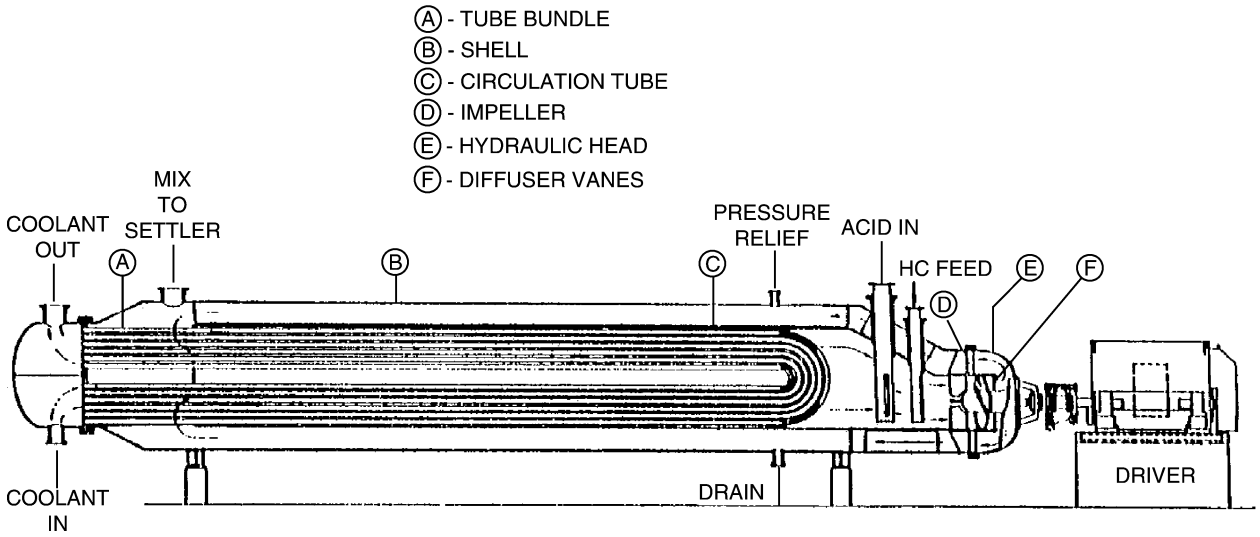


Fig. 1 Contactor designed by Stratco, Inc.: employs sulfuric acid as catalyst.

recently made several modifications including improved and bigger heat transfer systems, improved injectors and impellers, and lowering the impeller to below the axis of the reactor to obtain more uniform dispersions. More modern reactors can provide the following advantages (or some combination): lower operating temperatures, improved quality alkylates, lower acid consumption, larger production capacities, and/or lower compressor duty (or less refrigeration costs).

The emulsion leaving the reactor enters a settler. Residence times there often average up to 60 min to permit separation of the two liquid phases. Most of the acid phase is recycled to the reactor, being injected near the eye of the impeller. The hydrocarbon phase collects at the top of the decanter; it contains unreacted isobutane, alkylate mixture, often some light *n*-paraffins, plus small amounts of di-isoalkyl sulfates. The sulfates must be removed to prevent corrosion problems in the distillation columns. Caustic washes are often employed to separate the sulfates; they result in destruction of the sulfates. Acid washes have the advantage that most of the sulfates eventually react to reform sulfuric acid, which is reused, and to produce additional alkylate product.

Cascade Reactors: At least two types of these reactors and/or operating conditions are currently employed. Units designed by M.W. Kellogg were built until the mid-1960s, and some still produce about 8% of the total alkylate (P. Pyror, personal communication). Exxon-Mobil now designs units, as shown in Fig. 2, which account for about 6% of all alkylate produced worldwide. They refer to their units as ''stirred autorefrigerated alkylation reactors.'' At least one refinery claims their cascade reactors differ from the above two.

All cascade reactors are relatively long cylindrical vessels positioned horizontally. Reactors are partitioned into 4–10 reactor compartments; Fig. 2 shows 10 compartments. Each compartment contains an agitator, which produces and maintains the acid/hydrocarbon dispersion that is circulated repeatedly. Part of the dispersion in a compartment flows over the vertical plate, separating the compartments; this overflow enters the next compartment. The acid to the cascade reactor enters the first compartment, as does part of the isobutane feed (designated in Fig. 2 as recycle refrigerant). The feed isobutane, recycled isobutane from the distillation section, and all of the olefins are premixed, and essentially identical amounts are jetted into each compartment, probably into the eye of the impeller. Details of the design of the impeller, injection device, etc. have not been reported by Exxon-Mobil. The level of agitation is, however, thought to be less than that in Stratco units, based on agitation energy requirements.[24]

In each compartment, the pressure is regulated so that some of the dispersed hydrocarbons vaporize. Light hydrocarbons and especially isobutane vaporize; this vaporization removes most but not all of the following: heats of reaction plus energy provided by agitation. Ackerman et al. have reported that there is a temperature increase from the first to the last compartment; simultaneously, the quality of the alkylate decreases.[24] The vaporization, with high certainty, occurs mainly in those droplets positioned at or near the top surface of the dispersion, where the pressure is lowest. Hence, many of the droplets that pass over the vertical plate from one compartment to the next have experienced vaporization (and have reduced concentrations of isobutane).

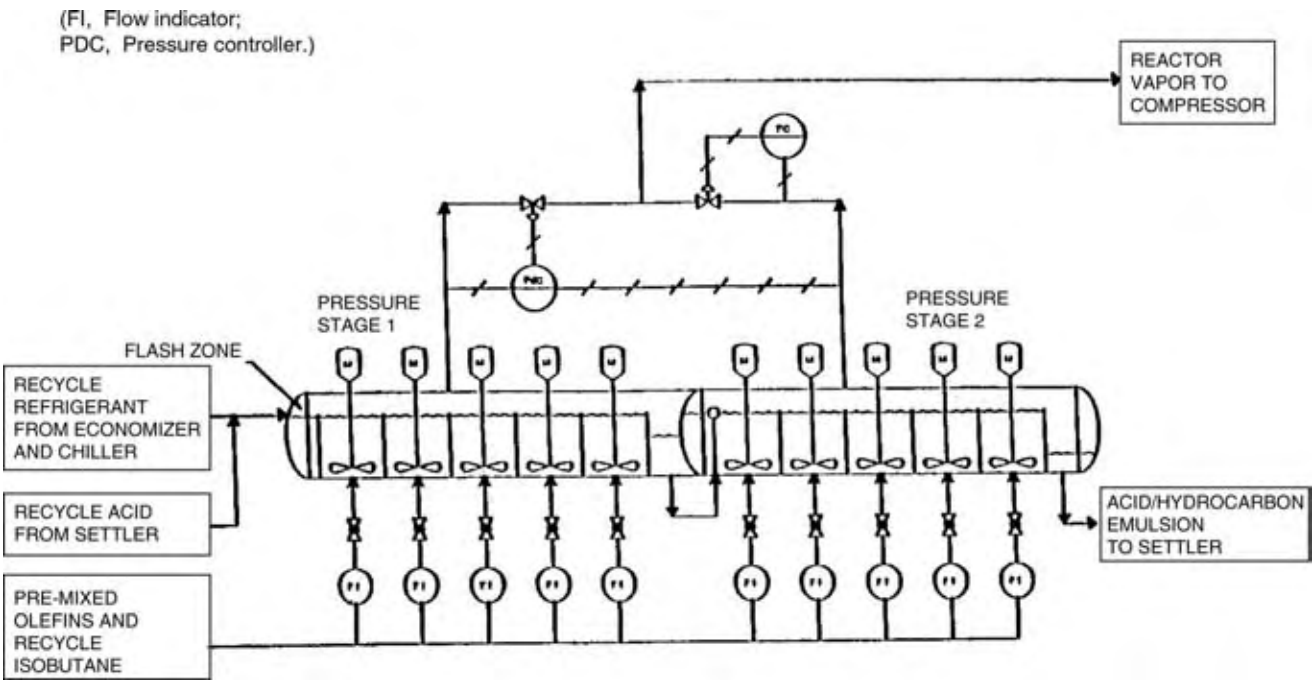


Fig. 2 Reactor designed by Exxon Research and Engineering Co.: employs sulfuric acid as catalyst.

The compositions and sizes of droplets obviously vary greatly in a given compartment. In the first reactor compartment, droplets containing all of the following occur: relatively pure isobutane, feed mixture of isobutane and olefins, hydrocarbon mixture of unreacted isobutane and alkylate, hydrocarbon mixture of alkylate plus the remaining isobutane after vaporization, and combinations of the above. Coalescing and fragmentation of droplets obviously occur as the dispersion flows toward and into the eye of the impeller and as the dispersion flows away from the impeller.[20] There are also significant differences in the droplets from the first to the last compartment. Since residence times in cascade reactors are often as high as 30 min, a small but significant amount of TMPs degrade in cascade reactors, probably a higher fraction than in Stratco reactors.[21]

The gases (mainly isobutane) leaving the top of cascade reactors are compressed, condensed, and then cooled by partial vaporization; the resulting liquid isobutane (refrigerant) is recycled to the reactor. The dispersion leaving the reactor is separated by decanting. Most of the acid is then recycled to the reactor.

A comparison of Stratco and Cascade reactors indicates the following differences:

1. More isobutane needs to be flashed in Stratco reactors to provide sufficient temperature differences for adequate heat transfer. Compressor costs in Stratco units tend to be higher. Less isobutane, however, generally needs to be separated by distillation in Stratco processes; hence, the energy costs for distillation tend to be lower.
2. Stratco reactors are generally smaller, often producing 3500–4500 barrels of alkylate/day. Cascade reactors tend to be larger, often producing 10,000 barrels/day or even greater. Plant operators often think large reactors reduce operating costs per unit amount of alkylate produced. But as already discussed earlier, combinations of two or more small reactors often result in much reduced acid consumption plus improved quality alkylates.[25,26] Such arrangements can result in significant economic benefits.

REACTORS USING HF

The reactors designed by Phillips Petroleum (now Conoco–Phillips) and UOP produce about 32% and 20%, respectively, of all alkylate produced worldwide (P. Pyror, personal communication). The HF/hydrocarbon dispersions in these reactors have much larger interfacial areas (probably by a factor of about 10) as compared to areas in sulfuric acid/hydrocarbon dispersions.[3] This is a major reason why the kinetics of alkylation is much higher with HF.

Conoco–Phillips Reactor: This reactor is basically a vertical tube in which the HF/hydrocarbon dispersion

flows upward, as shown in Fig. 3. Most of the reactions occur in the bottom section of the reactor, within several seconds, as indicated by temperature increases there of several degrees Celsius.

When the feed mixture of isobutane and olefins is introduced in the liquid HF, it provides sufficient energy so that the resulting liquid–liquid dispersion flows upward in the reactor riser and exits into a settler (decanter). The hydrocarbon product stream collects in the top portion of the decanter. The HF phase collects in the bottom of the settler and then flows downward because of gravity through the reactor standpipe to heat exchangers. Here, the HF is cooled using cooling water as the coolant. An advantage of this process is no pump is needed to recirculate the liquid HF. The reactor temperatures probably vary from 30°C to 40°C depending to some extent on the temperature of the available cooling water. The higher temperatures likely occur during summer or in the hotter regions of the country.

UOP Reactors: These reactors are vertical cylindrical vessels with cooling coils inside. The HF flows upward and mixtures of isobutane and olefins are injected at several different heights to form HF-continuous dispersions. Once again the reactions occur within several seconds and at about 30°C to 40°C. A pump is needed in this process to maintain the flow of liquid HF. Such a pump has resulted in HF leaks and must be handled with care.

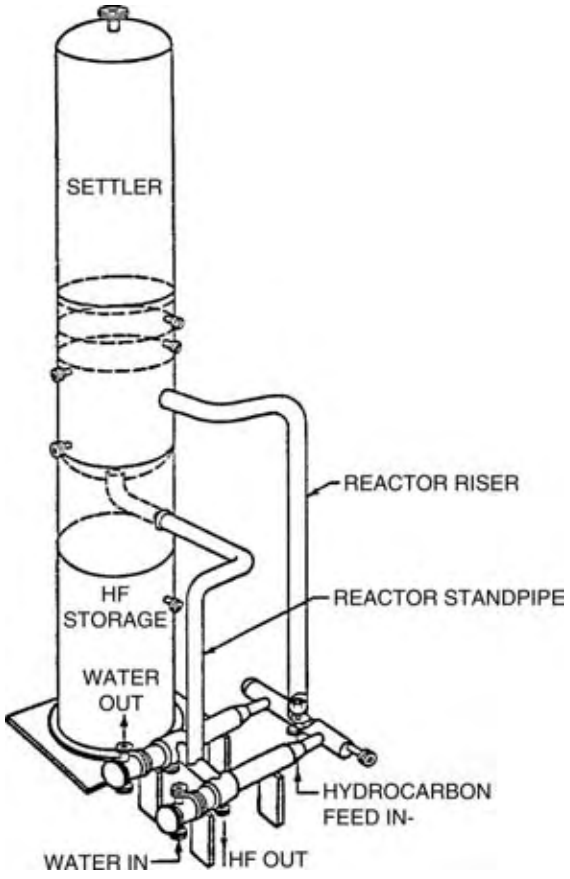


Fig. 3 Reactor designed by Phillips Petroleum Co.; employs HF as catalyst.

Both Phillips and UOP have employed additives dissolved in the liquid HF. The objective is to reduce the tendency to form, as a result of an accident, highly toxic aerosol clouds. Claims have been made that the additives reduce the amount of the HF that enters such a cloud by 65–90%. Although such a reduction is obviously important, no information is available on the hazards that still exist or how far toxic levels of HF might still be transferred. The additives eventually need to be separated, recovered, and recycled. The concentrations of the additives in the HF have apparently not been reported in any detail. Additives increase operating costs significantly. Somewhat high-quality alkylates are, however, produced in at least some cases when additives are used. Conoco–Phillips use sulfolane as their additive while UOP has tested both pyridine and picolene. These additives likely form hydrogen bonds with HF, and also affect the interfacial areas of the dispersions.

Water sprays have also been proposed to reduce HF releases in case of an accident. How to direct the spray at the location of the leak, designing the spray to remain operative after an accident, and providing sufficient water are all concerns. Numerous refineries have also installed ''protected'' storage facilities for feed HF and used HF to minimize potential leaks in or near the refinery.

SEPARATION AND RECOVERY OF ALKYLATE

Two to four distillation columns are usually required to separate the liquid hydrocarbon product stream that contains unreacted isobutane, alkylate mixture, *n*-butane, and propane. The major column is designated as the deisobutanizer (DIB) column. Often this column separates the isobutane as the overhead stream, the alkylate as the bottom stream, and a *n*-butane rich sidestream. In many plants, the feed isobutane is also fed to the DIB to remove most of the *n*-butane. A second column is generally needed to remove propane from the isobutane. Sometimes a third column is provided to purify further the *n*-butane sidestream and to recover more isobutane. In an alternate arrangement, the bottom stream of the DIB column is a mixture of alkylate and *n*-butane. This mixture is then separated in another column.

Alkylation processes using HF require in general larger columns, which also have significantly higher operating costs. First, they operate in general with

substantially higher ratios of isobutane to olefins. Hence, more isobutane is vaporized and separated, which necessitates higher energy costs. Second, some HF is dissolved in the hydrocarbon phase (or is present as isoalkyl fluorides). During distillation, HF is freed and is recovered as an immiscible liquid. The presence of HF necessitated safety provisions to prevent HF releases. Third, as propane and *n*-butane are produced as by-products of mechanism 4, more of both needs to be separated.

In HF-type processes, used HF (containing dissolved conjunct polymers) is separated by distillation to obtain relatively pure liquid HF and conjunct polymers. This column is relatively small and with much lower energy requirements as compared to the DIB column. Safety provisions are important since relatively pure HF is produced.

COMPARISON OF ALKYLATION PROCESSES WITH DIFFERENT CATALYSTS

Up to 10–15 yr ago before HF safety concerns had been determined in detail, it was often thought that HF-type processes had lower operating costs, primarily because of the following two features. First, cooling water is used as a coolant; such cooling is considerably cheaper than using isobutane as a refrigerant. Relatively large compressors are then needed to obtain operating temperatures in the reactor of about 5–15°C. Second, regeneration of used HF is much cheaper than that of used sulfuric acid. Several years ago, sulfuric acid costs accounted for perhaps 25–30% of the total operating costs.

Sulfuric acid-type processes have the following advantages. First, separation (or distillation) costs are substantially cheaper. Second, costs of feedstocks (and especially isobutane) are smaller by several percentages.[8] Third, higher-quality alkylates can be produced from 1-butene. Fourth, it is often feasible to use C_5 olefins as feeds.

Recently, significant improvements have been made for processes using sulfuric acid. First, acid consumption has been reduced by as much as 50% in at least select refineries. Second, simultaneously the quality of alkylates has increased. Even further improvements of these processes seem likely, especially if improved reactors are built.

Although sulfuric acid is potentially hazardous, it seems to be significantly less hazardous even when additives or water sprays are employed in units employing HF. Schechler and Schmidt reported in 1997 that sulfuric acid type processes are the benchmark alkylation processes, which they likely will continue to be in the foreseeable future.[27]

REFERENCES

1. Van Zele, R.L.; Diener, R. On the road to HF mitigation. Hydrocarbon Process. **1990**, Jun 92–98, Jul 77–80.
2. Albright, L.F.; Spalding, M.A.; Faunce, J.; Eckert, R.E. Alkylation of isobutane with C_4 olefins. III. Two-step process using sulfuric acid as catalyst. Ind. Eng. Chem. Res. **1988**, *27*, 391–397.
3. Albright, L.F.; am Ende, D.J. Alkylation-homogeneous. In *Encyclopedia of Catalysis*, Horvath, I.T., Ed.; John Wiley and Sons: Hoboken, NJ, 2003; Vol. 1, 191–210.
4. Hofmann, J.E.; Schriesheim, A. Ionic reactions occurring during sulfuric acid catalyzed alkylation. J. Am. Chem. Soc. **1962**, *84*, 953–961.
5. Kramer, G.M. *Alkylation Studies in Industrial and Laboratory Alkylation*; Albright, L.F., Goldsby, A.R., Eds.; ACS Symposium Series 55; American Chemical Society: Washington, DC, 1977; 1–26.
6. Shlegeris, R.J.; Albright, L.F. Alkylation of isobutane with various olefins in the presence of sulfuric acid. Ind. Eng. Chem. Process Des. Dev. **1969**, *8*, 92.
7. Albright, L.F. Updating alkylate gasoline technology. Chemtech **1998**, *28* (6), 40–47.
8. Albright, L.F. Alkylation of isobutane with C_3–C_5 olefins: feedstock consumption, acid usage, and alkylate quality for different processes. Ind. Eng. Chem. Res. **2002**, *41*, 5627.
9. Hutson, T.; Hays, G.E. Reaction mechanisms for hydrofluoric acid alkylation. In *Industrial and Laboratory Alkylation*; Albright, L.F., Goldsby, A.R., Eds.; ACS Symposium Series 55; American Chemical Society: Washington, DC, 1977.
10. Miron, S.; Lee, R.J. Molecular structure of conjunct polymers. J. Chem. Eng. Data **1963**, *8*, 150–160.
11. Albright, L.F.; Spalding, M.A.; Kopser, C.G.; Eckert, R.E. Alkylation of isobutane with C_4 olefins. II. Production and characterization of conjunct polymers. Ind. Eng. Chem. Res. **1988**, *27*, 386.
12. Kalas, D.; Acid storage tank blows at Martinez CA Refinery. Contra Costa Times and Orlando, FL, Newspaper, Oct 10, 1993.
13. United States Chemical Safety and Hazards Investigation Board. *Findings and Recommendations of Motiva Refinery Sulfuric Acid Tank Farm Disaster*; Aug 28, 2002.
14. am Ende, D.J.; Albright, L.F. Degradation and isomerization of isoparaffins while in contact with sulfuric acid in alkylation units: chemistry and reaction kinetics. Ind. Eng. Chem. Res. **1994**, *33*, 840.
15. Sung, S.; Szechy, G.; Albright, L.F. Decomposition of spent alkylation sulfuric acid to produce

sulfur dioxide and water. Ind. Eng. Chem. Res. **1993**, *32*, 2490.
16. Albright, L.F.; Doshi, B.M.; Ferman, M.A.; Ewo, A. Two-step alkylation of isobutane with C_4 olefins: reaction of isobutane with initial reaction products. In *Industrial and Laboratory Alkylations*; Albright, L.F., Goldsby, A.R., Eds.; ACS Symposium Series 55; American Chemical Society: Washington, DC, 1977; Chapter 7, 109–127.
17. Albright, L.F. Mechanism for alkylation of isobutane with light olefins. In *Industrial and Laboratory Alkylations*; Albright, L.F., Goldsby, A.R., Eds.; ACS Symposium Series 55; American Chemical Society: Washington, DC, 1977; Chapter 8, 128–146.
18. am Ende, D.J.; Eckert, R.E.; Albright, L.F. Interfacial area of dispersions of sulfuric acid and hydrocarbons. Ind. Eng. Chem. Res. **1995**, *34*, 4343.
19. Li, K.W.; Eckert, R.E.; Albright, L.F. Alkylation of isobutane with light olefins using sulfuric acid: operating variables affecting physical phenomena only. Ind. Eng. Chem. Process Des. Dev. **1970**, *9*, 434.
20. McConnell, J.A.; Smuck, W.W. Gamma backscatter technique for level and density detection. Chem. Eng. Progr. **1967**, *63* (8), 79–82.
21. Albright, L.F. Alkylation-industrial. In *Encyclopedia of Catalysis*; Horvath, I.T., Ed.; John Wiley and Sons: Hoboken, NJ, 2003; Vol. 1, 262–281.
22. Eastman, A.D.; Randolph, R.B.; Moore, W.P.; Heald, R.L. Online monitoring of HF. Hydrocarbon Process. **2001**, Aug, 95–100.
23. Albright, L.F. Modern alkylation: H_2SO_4, HF process compared and new technologies revealed. Oil Gas J. **1990**, Nov 26, 70–77.
24. Ackerman, S.; Lerner, H.; Cavanaugh, T.A. Autorefrigerated reactor in sulfuric acid alkylation. In *Pet. Technol. Q.*; Crambeth Allen Publishing: Craven Arms, U.K., 1996; 27–33.
25. Graves, D.C.; Kranz, K.E.; Millard, J.K.; Albright, L.F. Alkylation by Controlling Olefin Ratios US Patent 5,841,014, Nov 24, 1998.
26. Graves, D.C.; Kranz, K.E.; Millard, J.K.; Albright, L.F. Alkylation by Controlling Olefin Ratio US Patent 6,194,625, Feb 27, 2001.
27. Schechler, J.C.; Schmidt, R.J. Motor fuel alkylation advances beyond liquid acid catalysts, Paper at Annual Meeting of National Petroleum Refiners Association, San Antonio, TX, Paper AM-97-47, Mar 16–18, 1997.

Animal Cell Culture

Shang-Tian Yang
Shubhayu Basu
Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio, U.S.A.

INTRODUCTION

Animal cell culture or the ability to continuously grow animal cells in vitro after removing them from animal tissue opens up a plethora of windows in the field of biology and medicine. They provide a platform to investigate the normal physiology and biochemistry of cells, test the effects of drugs and other compounds in vitro, produce artificial tissue for implantation, synthesize valuable products from large-scale cultures, and can even be used as models in studying diseases. Animal cells in culture have been used as end products to provide artificial skin grafts, islet cells, hepatocytes, and bone marrow implants and to produce recombinant and natural proteins like human growth hormone, nerve growth factor, epidermal growth factor (EGF), monoclonal antibodies (MAb), vaccines, interferons, and blood clotting factors. Animal cells are often more advantageous than yeast or bacterial cells for the production of recombinant proteins on a large scale. Proteins that need to be heavily glycosylated for function, or need a proper folding environment because they have a large number of disulfide bonds, are often made in animal cells. Bacterial cells cannot perform glycosylation or phosphorylation, and yeast cells cannot carry out complex glycosylation reactions. Often the glycosylation carried out by yeast cells is not authentic. Animal cells not only secrete the products efficiently but also provide excellent glycosylation and phosphorylation.

Animal cells have very different morphology and characteristics from microbial cells. They secrete an extracellular matrix (ECM), which provides a medium for the cells to interact and migrate. Animal cell lines can be anchorage dependent or independent, and require a suitable medium, temperature, pH, and dissolved oxygen concentration to grow. This entry discusses the structure of the animal cell, the basic procedures associated with animal cell culture, and the important process parameters governing it. The growth environment is highly important for animal cells as changes in the environmental factors can induce growth arrest or apoptosis or even stimulate proliferation or differentiation. Animal cell culture kinetics follows a similar trend as microbial cultures, but, in general, the cell cycle duration for animal cells is much longer than for prokaryotic cells. General animal cell culture kinetics and the cell cycle are reviewed. The intricacies of the animal cell and its shear sensitivity and anchorage dependence make it difficult to scale up bioprocesses involving animal cells. This entry discusses these problems and a variety of bioreactors designed to meet these requirements for large-scale animal cell culture processes.

ANIMAL CELL

Cell Structure

A variety of cell types, including epithelial cells, fibroblasts, muscle cells, nerve cells, cardiac cells, mesenchymal cells, endocrine cells, and embryonic stem cells, have been successfully cultured and maintained in vitro. These animal cells can vary widely in shape, size, and function depending on their sources. Fig. 1 shows a typical animal cell with its internal structure and organelles. In general, animal cells are much larger than microbial cells with a diameter between 10 and 30 μm. Though they lack chloroplasts (and therefore are not photosynthetic), they have many specialized cell organelles, such as Golgi bodies, endoplasmic reticulum, and mitochondria, each surrounded by its own plasma membrane. The nucleus is the most prominent organelle in the animal cell and houses chromosomes and other nuclear proteins. Also present in the nucleus are one or more nucleoli, which are involved in ribosome synthesis. The compartmentalization of different organelles facilitates the separation of specific metabolic functions that are incompatible otherwise. Other membrane bound organelles include lysozomes and peroxisomes, which mainly contain digestive enzymes for specific metabolic processes. It is important to note that, apart from chloroplasts, animal cells also lack vacuoles and a rigid cell wall that are present in plant cells. The lack of a cell wall makes the animal cell highly sensitive to shear forces.

Nonmembranous organelles within the cell include microtubules, microfilaments, and centrioles. The microtubules and microfilaments form a framework

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120017564

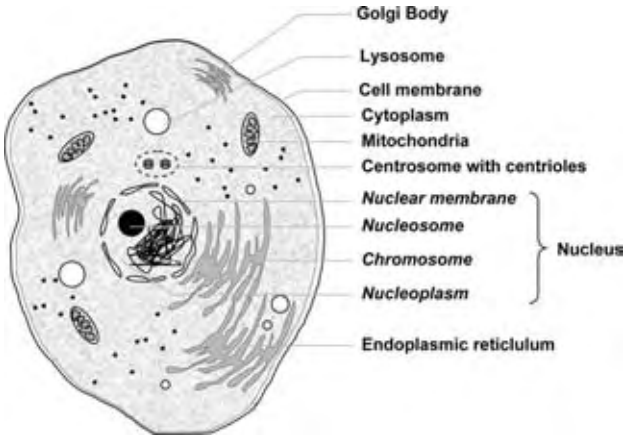


Fig. 1 The structure of a typical animal cell. *(View this art in color at www.dekker.com.)*

called the cytoskeleton that helps the animal cell to maintain its form and keep the organelles in place. The cytoskeleton is also involved in the process of cell division and migration. The larger the cell, the more intricate and elaborate its internal cytoskeletal structure.

Extracellular Matrix

In contrast to prokaryotes, which are unicellular and therefore can live as single entities, animal cells need to socialize—an observation that can be attributed to their origin from higher multicellular organisms. Cells in tissues are usually in contact with a complex network of secreted extracellular proteins referred to as the ECM. This matrix holds the cells together and provides a medium for the cells to interact and migrate.[1] The ECM is secreted mainly by the cells within it and comprises primarily two classes of macromolecules: 1) the glycosaminoglycans, which are linked to proteins in the form of proteoglycans, and 2) the fibrous proteins. Some fibrous proteins are structural proteins such as collagen and elastin, which provide rigidity and elasticity to tissue, respectively, and some are mainly adhesive proteins such as fibronectins and laminins. Fibronectin is a large glycoprotein with binding domains for other ECM proteins and is of primary importance. Mice with their fibronectin gene ''knocked out'' either die in the embryonic stage or grow up to have multiple morphological defects. Of equal importance are the integrins—transmembrane proteins that interact with the cellular cytoskeleton and thereby anchor the cells to the ECM. They are heterodimers, which mediate bidirectional interaction between the cytoskeleton and the ECM. The fibronectins and the integrins together function to attach the cells to their surroundings. It is because of this interdependence of the integrins and the fibronectins that some animal cells can be cultured only when they have a substratum to grow on. This anchorage dependence of some animal cells helps each cell to attach to the substrate and spread along it to interact with other neighboring cells and thus mimics the functions of a tissue in vitro. Fig. 2 illustrates how the ECM molecules help cells attach and spread on the surface of the substratum. Also, cells may have different shapes when attached on the two-dimensional surface or in a three-dimensional space with cell–cell interactions.

Transformed and tumorigenic cells are different from normal cell lines in that they are not usually anchorage dependent. They exhibit a spherical shape, increased life span and lateral diffusivity of membrane proteins, decreased cell receptors and membrane proteins, and a different cytoskeletal structure.[2] The decrease in the concentration of the cell adhesion molecules in the cell membrane of these cells causes the anchorage independence. Transformed cell lines also do not assemble a normal ECM. It is important to note that some cell lines (e.g., lymphocytes) that are normally anchorage dependent can be induced and then adapted to become anchorage independent. This is of tremendous importance to recombinant protein production as discussed later, because the scale-up of suspension cultures is easier than that of anchorage-dependent cell lines.

Cell Lines

The first stage of tissue culture is the primary cell culture. After isolating a desired piece of tissue, it is disaggregated either mechanically or enzymatically. The resulting tissue fragments are then used to inoculate the culture vessel that contains medium. Most normal cells are anchorage dependent. Hence, some of these fragments attach to the vessel wall and migrate out along the surface. Such a culture, before it is first passaged or subcultured, is called a primary culture. As the cells proliferate, they keep spreading out on the culture dish surface until the dish is covered by a single layer of cells. The cells are then said to have reached confluence. Once the cells form this continuous sheet, they stop proliferating because of contact inhibition. Transferring them at low densities (a process termed subculturing or passaging) to new culture vessels that contain fresh medium induces them to resume proliferation. Such a cell population that can continue to grow through many subcultures is called a normal, untransformed cell line.

After about 50 divisions, however, proliferation slows down and the cells show senescence and begin to die. Some of the cells in the culture may undergo some genetic modifications or transformation, which allows them to escape senescence. As long as they are

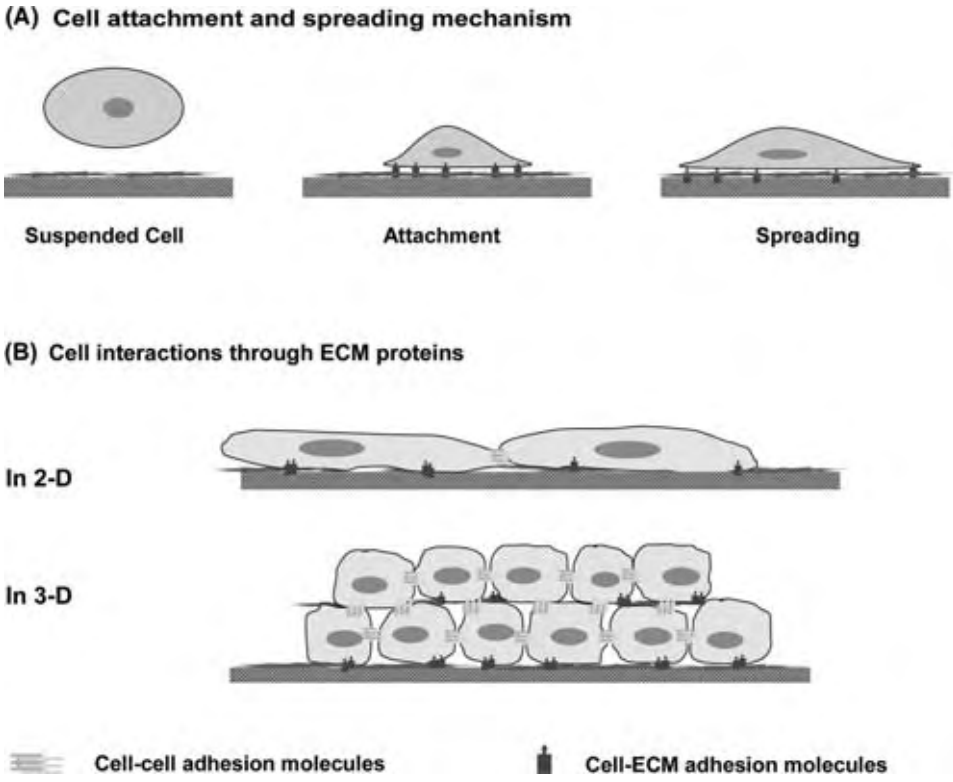


Fig. 2 Cell attachment on surface through cell–ECM interactions. (A) Cell attachment and spreading mechanism. (B) Cell–cell and cell–substratum interactions in two-dimensional and three-dimensional environments. *(View this art in color at www.dekker.com.)*

subcultured periodically, they can grow indefinitely. Such cells with an infinite life span are said to have undergone ''immortalization'' and the cell line is called a transformed cell line. Transformed cells have an enhanced growth rate and may lose anchorage dependence. Tumorigenicity is analogous to transformation but not all transformed cells are tumorigenic or malignant. However, all tumorigenic cells have an infinite life span and enhanced growth rate akin to transformed cells and are mostly anchorage independent.

Table 1 shows some animal cell lines that are commonly used in various applications. Creating a stable, permanent cell line is the first critical step in producing recombinant proteins for therapeutic and diagnostic applications. Hybridoma, commonly used in the production of MAb, are generated by fusing antibody-producing spleen cells, which have a limited life

Table 1 Common cell lines and their application

Cell line	Origin	Cell type	Application
BHK	Baby hamster kidney	Fibroblast	Vaccine production
COS	African green monkey kidney	Fibroblast	Transient expression of recombinant genes
Vero	African green monkey kidney	Fibroblast	Human vaccine production
3T3	Mouse connective tissue	Fibroblast	Development of cell culture technique
CHO	Chinese hamster ovary	Epithelial	Recombinant glycoprotein production
HeLa	Human cervical carcinoma	Epithelial	Animal cell model
MDCK	(Madin Darby) canine kidney	Epithelial	Veterinary vaccine production
Namalwa	Human lymphoma	Lymphoblast	α-Interferon production
NS0	Myeloma	Lymphoblast	MAb production
MPC–11	Mouse myeloma	Lymphoblast	Immunoglobulin production
HKB11	Human somatic hybrid	Hybrid	Recombinant protein production
ES-D3	Mouse embryo	Pluripotent embryonic stem cell	Insulin production

(Adapted from Ref.[3].)

span, with cells derived from an immortal tumor of lymphocytes (myeloma). The resulting hybrid is capable of unlimited growth and producing the antibody. Industrial cell lines used for recombinant protein production usually have been genetically engineered to improve the cell in its growth and ability to produce the protein product at a high expression level.

CULTURING CONDITIONS FOR ANIMAL CELLS

Animal cells can be anchorage dependent or independent but all cell lines need nutrients, a suitable temperature, pH, and dissolved oxygen level to grow. The growth environment for animal cells is highly important because environmental stimuli can trigger different responses from the cell. Changes in growth conditions can induce growth arrest or even apoptosis, or may also stimulate proliferation or differentiation. It is therefore crucial to maintain and closely monitor the growth environment so that the cells grow to high cell densities, have stable genotypic and phenotypic expression, and are able to efficiently express recombinant genes if desired. Some important parameters in animal cell culture are discussed here.

Substrate for Cell Attachment

In earlier days, reusable borosilicate glass bottles were used for animal cell cultures. The hydrophilic glass surface was suitable for cell attachment and growth. With the advent of the plastic age, presterilized polystyrene culture apparatus are readily available for cell growth. The polystyrene surface is usually sulfonated to make it hydrophilic and is sterilized by γ-irradiation. They are meant for a ''one-time'' use and reduce the risk of contamination. Popular forms of culture flasks are T-flasks, Petri dishes, and multiwell plates. Cells grown on the flat or two-dimensional surface such as in a T-flask usually stretch and show a somewhat flattened morphology (Fig. 3A). Stretching allows cells to migrate on the surface and promotes proliferation.

Microcarriers have also provided a good way to increase the available surface area per unit volume in large-scale bioreactors. Various types of microcarriers have routinely been used for the growth of anchorage-dependent cells.[4] Cylindrical cellulose-based microcarriers (DE-53) were among the first used.[5] Since then, different materials like collagen coated-glass beads, gelatin, DEAE dextran, glass-coated plastic, collagen-coated polystyrene, and polyacrylamide have been used.[6] These beads range from 90 to 330 μm in diameter and can be optimized for different cell lines. Calcium alginate gel beads and new surface modified polystyrene have also been used.[7] Cells

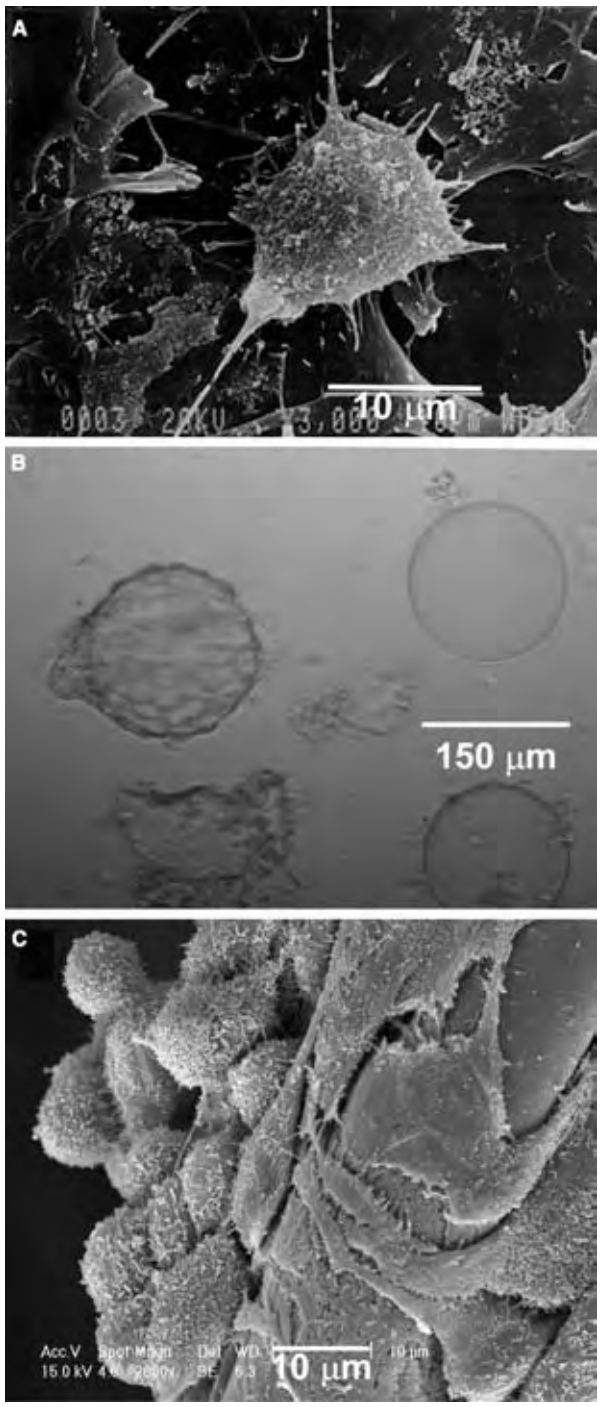


Fig. 3 Different cell morphologies observed for growth on different substrata: (A) cow luteal cells grown on the surface of a T-flask; (B) osteosarcoma cells grown on microcarriers; and (C) osteosarcoma cells grown on a fiber in nonwoven polyester fabrics. *(View this art in color at www.dekker.com.)*

attached and grown on microcarriers usually form a monolayer and reach confluence during the culture (Fig. 3B).

So far, discussions have focused on animal cell growth in monolayers or two dimensions. However,

for tissue engineering applications, cell masses need to aggregate into a three-dimensional tissue construct. To aid this process, tissue scaffolds are used as a matrix to guide animal cell growth in three dimensions. In general, a variety of scaffold types have been used. A broad classification would be foam-like scaffolds (e.g., alginate sponge, chitosan, collagen foam, and PLA foam) and fibrous scaffolds like collagen fibers, nonwoven polyethylene terephthalate, and polypropylene mesh. Cells cultivated in the three-dimensional support environment can not only attach to the surface of the substratum but also grow into the three-dimensional space to form aggregates (Fig. 3C).

Culture Medium

Various types of media have been used to cultivate different cell lines. The choice is mostly empirical, but formulations can be optimized for different cell lines and purposes. Most media, however, have the following essential components: balanced salt solutions (BSS), essential amino acids, glucose, vitamins, buffers, and antibiotics. The BSS provides a concoction of inorganic salts required by the cells and usually has an osmolality between 260 and 320 mOsm/kg, which is similar in range to that experienced by cells in vivo.[8,9] Balanced salt solution often contains sodium bicarbonate and phosphates, which apart from nutrient value, also act in a buffering capacity.

Glucose is the major carbon and energy source in medium formulation, but its concentration varies in different media. Eagle's minimum essential medium contains 1 g/L of D-glucose, whereas Dulbecco's modification incorporates a higher glucose concentration of 4.5 g/L. Amino acids are important nutrients for cell growth and are additional sources for carbon and energy.[10] Essential amino acids are those that cannot be synthesized by the cell metabolic machinery and, therefore, need to be supplemented through the media. Most cell culture media contain 2–4 mM of L-glutamine.

Buffers are inherent constituents of all media formulations and can maintain the medium pH within an acceptable range.[9] The most commonly used buffer is sodium bicarbonate and CO_2, which is usually provided in air at 5%. Dissolved CO_2 reacts with water to form carbonic acid, which dissociates into the bicarbonate ion [Eq. (1)]:

$$CO_2 + H_2O \leftrightarrow H_2CO_3 \leftrightarrow H^+ + HCO_3^- \quad (1)$$

So, increased amounts of dissolved CO_2 increase the acidity of the medium. This action is countered by the presence of sodium bicarbonate [Eq. (2)]:

$$NaHCO_3 \leftrightarrow Na^+ + HCO_3^- \quad (2)$$

The dissociation of sodium bicarbonate to bicarbonate ions in solution shifts the equilibrium of the reaction in Eq. (1) back, thereby effectively maintaining the pH at 7.4. Good et al. came up with a range of zwitterionic buffers.[11] Such buffers, like *N*-2-hydroxyethylpiperazine-*N*′-2-ethanesulfonic acid, have a pK_a of 7.31, which is optimal for cell culture, do not penetrate the cell membrane, and equilibrate with air.[9]

The pH indicator phenol red (phenol sulfonphthalein) is often added to commercially available media. It is pale yellow at pH 6.5, orange at pH 7.0, and red at pH 7.4. It becomes purplish above pH 7.4. Growth media are also often supplemented with antibiotics to promote the growth and propagation of antibiotic resistant strains and also to prevent contamination by micro-organisms. But the presence of antibiotics in the medium does not obviate the use of good aseptic techniques. Nowadays, however, most commercially available media are presterilized as are the polystyrene culture dishes.

Serum

Different types of serum have been used to supplement media with various necessary growth factors and hormones that cells need for their growth. Serum also contains various adhesion factors and antitrypsin activity, which promotes cell attachment. Serum components can act as buffers and as chelators for labile or water insoluble nutrients, bind and neutralize toxins, and provide protease inhibitors. Serum can also reduce oxidative injury to cells caused by ferrous ions.[12] Reduced serum conditions have also been reported to increase the susceptibility of cells to apoptosis.[13]

Various types of serum are fetal bovine serum—the most widely used type, newborn calf serum—which is derived from animals less than 10 days of age, donor calf serum—which is obtained from processed whole blood of calves up to 8 mo old, and horse serum. Human serum and chicken serum have specialized uses for those cells that require a serum derived from similar species, e.g., chicken serum has been repeatedly used for growing various types of avian cells.[9] Serum can be stored safely for over 12 mo at −20°C and longer storage is possible at −70°C. Serum can also be heat inactivated (incubation at 56°C for 30 min) to remove toxic compounds or other agents that can interfere with tissue typing assays.

However, as serum contains a wide range of components whose exact concentration is not known, there is a lot of variability from batch to batch. The presence of serum in culture medium is undesirable in cases where a protein or product has to be purified for commercial use. The cost of serum and the possibility of viral contamination also inhibit its use.

Serum-Free Medium

Many attempts have been made to replace serum and add defined amounts of the essential components of serum to form what is known as serum-free medium (SFM). Serum-free medium, generally, consists of a basal medium and additional supplements. The basal medium provides the essential and nonessential amino acids, vitamins, nucleic acids, lipids, inorganic salts, and a carbon source. The additional supplements are: growth promoters such as insulin, insulin-like growth factors, EGF, platelet-derived growth factor, estradiol, and dexamethasone; attachment factors like collagen, fibronectin, and laminin; and transport proteins and detoxifying agents like albumin and transferrin.[14] Ito et al. reported that insulin immobilized on microcarriers promoted growth of anchorage-dependent cells in a protein-free cell culture system.[15] Some serum-free systems have even used the structural heterogeneity of high-density lipoproteins to influence cell proliferation.[16] Serum-free systems are also very useful in reducing downstream processing steps for recombinant protein production using animal cell culture. Unfortunately, the transition to SFM has not been easy. Different cell lines require several growth factors and the specific growth rates of cells are usually slower in SFM. Moreover, SFM can also be expensive.

Temperature and pH

In contrast to micro-organisms, mammalian cells do not show great adaptation to varying temperature or pH ranges. Most cell lines prefer a pH of 7.4. Eagle reported that normal fibroblasts grow within a pH range of 7.4–7.7, while transformed cells prefer a pH between 7.0 and 7.4.[17] Most cell lines grow best at a temperature of 37°C. Cells can tolerate considerably large drops in temperature in that they can be stored cryogenically in liquid nitrogen at −196°C for several months. At temperatures slightly lower than 37°C, the growth rate decreases, but the cellular metabolic activity does not cease totally. Reduction of temperature to 33.5°C resulted in a lowering of the specific growth rate of Chinese hamster ovary (CHO) cells while having no effect on the cell proliferation.[18] This could be attributed in part to the physical state of the lipid bilayer that makes up the plasma membrane. However, cells in general die at temperatures higher than 42–48°C, where the lipid bilayer exhibits a liquid crystal (fluid) behavior.[19] Temperature can also be used as a tool to control recombinant protein production by engineered cell lines. Hendrick et al. (2001) increased the productivity of tissue plasminogen activator under the control of the SV40 promoter in CHO cell, by a shift in temperature from 37°C to 32°C.[20]

Dissolved Oxygen

Oxygen is required for respiration and is thus a key nutrient for animal cell cultures even though requirements vary between cell lines.[21] Because of the low solubility of oxygen in water, oxygen must be provided continuously, usually by aerating the culture medium. Several authors have reported the importance and effects of oxygen in animal cell cultures. It is difficult to conclude a general trend for oxygen dependence vs. cell metabolism. In the case of antibody production using the AB2-143.2 hybridoma cell line, a pO_2 of 50% air saturation was optimum, while the highest immunoglobulin yield from human lymphoblastoid cells (RPMI #7430) was obtained at a low pO_2.[22,23] The trend is opposite for cell biomass production. While more hybridoma cells were obtained at a low pO_2 of 0.5% air saturation, more lymphoblastoid cells were produced at the highest atmospheric pO_2.[22,23] So, each cell line can be studied and grown at its own optimal pO_2 depending on the desired end product. Ma et al. also reported that a low (2%) oxygen tension promoted proliferation while a high (20%) oxygen tension induced differentiation in human trophoblast cells.[24]

CELL CULTURE KINETICS

The aim of animal cell culture can be either to use the cells as end products or to develop enough biomass (cells) to express a certain target protein of interest in economically viable amounts.

Growth Kinetics

Most animal cell cultures exhibit similar trends in their growth kinetics. While the specific growth rate of each cell line varies, all cell lines do exhibit a characteristic growth curve similar to the one shown in Fig. 4. Cells go through an initial phase of adjustment, the lag phase, after being either subcultured or dissociated from tissue. After this preliminary lag phase, the cells start growing exponentially and this phase is called the logarithmic phase. Following this phase of active growth, the cell growth rate reduces because of nutrient limitation and product accumulation. The total cell number then ceases to increase, a phenomenon triggered off not only by contact inhibition between cells but also because the cell culture reaches a dynamic equilibrium between the rate of cell growth and cell death. The last phase is the death phase, where cell number is reduced because of death caused by either apoptosis or necrosis.

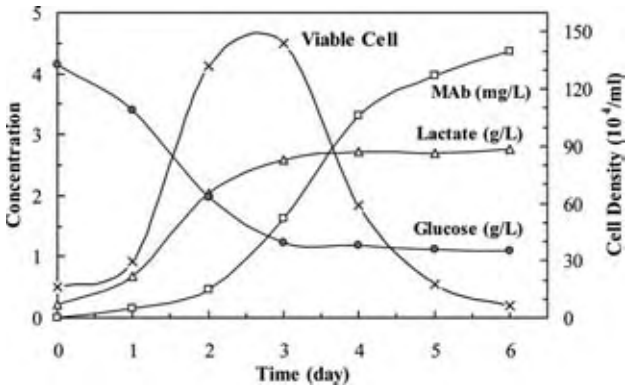


Fig. 4 Typical batch kinetics of a hybridoma culture showing characteristic growth curve and growth-associated production of lactate and nongrowth-associated production of antibodies.

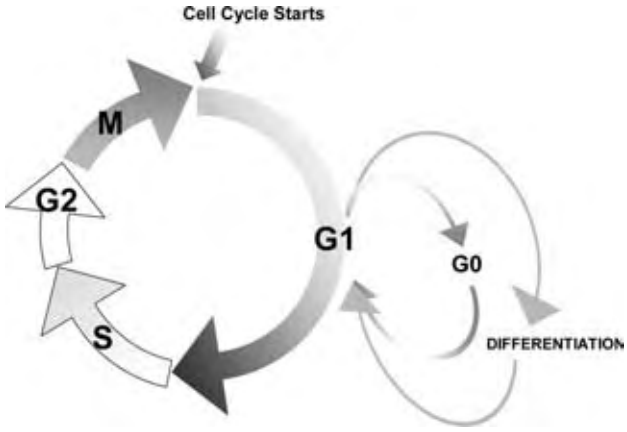


Fig. 5 The cell cycle. A proliferating cell goes through four phases: an initial growth phase (G1), a DNA synthesis phase (S), a second growth phase (G2), and, finally, a mitotic phase (M). During G1 the cell may also choose to differentiate into a new cell type or go into a quiescent state or G0 phase. *(View this art in color at www.dekker.com.)*

Cell Metabolism

Lactic acid is often a product of glucose metabolism via the glycolytic pathway, while ammonia is produced from the catabolism of glutamine and other amino acids. Cell growth can be inhibited by the accumulation of lactate and ammonia in the culture medium. Production of lactic acid can also lower the medium pH below the physiological range. Reducing or selectively removing toxic metabolites in the culture medium is critical to the efficient production of recombinant proteins by animal cells. The production of primary metabolites such as lactate is usually growth associated, whereas protein expression can be either growth associated or nongrowth associated. Fig. 4 shows a typical batch hybridoma culture with the production of MAb mainly in the stationary phase.

Cell Cycle

For a cell to grow and reproduce, it must go through the cell cycle shown in Fig. 5. The two major events in the cell cycle are DNA replication followed by cell division into two new daughter cells. A cell entering the cell cycle first goes through a gap phase (G1). During this phase, it undergoes protein synthesis, which primes the cells for the next phase, the synthesis (S) phase. A second copy of the cellular genome is made during the S phase, thus ensuring that fidelity is maintained when the cell divides. The third phase (G2) is another phase of protein synthesis and it prepares the cell for division, and mitosis finally occurs in the M phase. Essentially, the duration of the cell cycle is an important factor in determining the fraction of dividing cells in a given population. The cell cycle not only controls the rate of cell proliferation and growth, but also may affect protein expression as production of some recombinant proteins by animal cells is cell-cycle dependent.

Cell lines vary in their cell cycle duration, but all of them follow a basic pattern and are tightly regulated at certain checkpoints. If anything goes wrong, the regulatory machinery promptly causes a cell cycle arrest and eventual cell death. Deprivation of growth factors can cause a cell to exit from a proliferating mode to a quiescent mode (G0). During this phase, the whole cell metabolic machinery is suppressed. Addition of growth factors can again stimulate a cell to re-enter the G1 phase. The cell can also exit the cell cycle to differentiate into a new lineage or head toward programmed cell death or apoptosis.

Apoptosis

Apoptosis or programmed cell death contributes to cell death in in vitro cultures under suboptimal conditions. Nutrient deprivation like glucose, glutamine, serum or oxygen limitation, or mild hydrodynamic stress can induce apoptosis. High levels of apoptosis have been reported following deprivation of glucose and essential amino acids.[25] A number of studies have demonstrated that the suppression of this death pathway, by means of overexpression of survival genes such as *bcl*-2, results in improved cellular robustness and antibody productivity during batch culture.[26]

It is important to know the cell cycle and the points where it can be regulated. This is because cancer basically results when the cell cycle control machinery fails and the cell undergoes rapid proliferation and metastasis. On the other hand, apoptosis is also not desirable

in cells growing in bioreactors as it would lower the specific productivity of cells.

There is a plethora of proteins responsible for controlling and regulating the cell cycle and thereby deciding which course the cells should take. They can be broadly divided into the cyclins and cyclin-dependent kinases (Cdks), which interact with each other during regulation. In general, cyclin Ds are associated with the G1 phase, cyclin Es with the transition from G1 to S, cyclin As with the S phase, and both cyclin As and Bs with the transition from G2 to M. The Cdk–cyclin complexes can be inhibited by the Cdk inhibitors, which add an extra level of regulation.

ANIMAL CELL BIOREACTORS

A variety of bioreactors have been developed and used for animal cell cultures, from simple static T-flasks and roller bottles to more complicated multitray and rotating disk reactors. Because most animal cell lines are anchorage dependent, scale-up for animal cell cultures is usually based on providing the maximum surface area for cell attachment. For this reason, microcarriers have been developed for use in culturing animal cells in conventional stirred-tank and airlift reactors. A recent industrial trend is to adapt animal cells to grow in suspension without needing microcarriers for surface attachment. Immobilized cultures in hollow-fiber, packed-bed, and fluidized-bed reactors also have been used to greatly increase cell density and reactor productivity. Improvements in bioreactor design have focused on increasing oxygen transfer, reducing shear and bubble damages, and increasing cell density by cell recycle or immobilization in perfusion cultures. Fig. 6 shows different types of animal cell bioreactors.

Bioreactors for Suspension Cultures

Bioreactors for suspension cultures are similar to the common microbial fermenters. However, the increased shear sensitivity of animal cells necessitates a few changes. The fermenters for this purpose usually lack baffles and other sharp projections that can cause turbulence.[27] The interior of these fermenters is usually lined with glass or finished to a high grade of smoothness to minimize mechanical damage and to enhance cleanliness. The impeller designs are different too.[28,29] Modified marine and pitched-blade impellers usually cause much less cell damage than conventional disk turbine blades. Nevertheless, cell damage and death caused by mechanical agitation and gas sparging is still a major concern in reactor design and scale-up. For cell lines like hybridoma, which are extremely sensitive to shear stress, novel methods have been developed. The Vibro fermenter (Chemap) uses a plate that vibrates (0.1–3 mm) in the vertical plane to achieve mixing. Airlift bioreactors do away with mechanical agitation and achieve mixing by the process of aeration itself. Airlift bioreactors in general consist of two concentric tubes. The inner tube carries a sintered steel ring or other oxygenation apparatus through which air containing 5% CO_2 is bubbled. Air escapes at the top and the liquid comes down the outer tube. Hybridoma cells have been successfully grown in airlift bioreactors.[30] In general, cell densities in suspension cultures are lower than 10^6–$10^7\,ml^{-1}$ because of limited oxygenation at low agitation and aeration rates used to avoid severe cell damages.

Apart from ingenious modifications to the typical bioreactor, rotating wall culture vessels have also been in use (1). But a recent resurgence in their use has been triggered off by the discovery that gravity, or rather the lack of it, plays an important role in the morphology and physiology of tissue constructs.[1,31] The reactor cultures cells in a slowly rotating horizontal cylinder, which produces low shear stress and the continuous rotation keeps the cells always in a state of free fall to simulate microgravity conditions.

Perfusion Cultures

To achieve high cell densities, perfusion cultures have also been used. Perfusion implies the continuous or semicontinuous addition of fresh medium and withdrawal of used medium. This, however, dictates the need for cell separation. This has been achieved by using spin filters, hollow-fiber filters, gravitational settling, or centrifugal separation. Spin filters are usually attached to the stirrer shaft. As they spin, they create a boundary layer effect around them that reduces cell attachment and clogging. Centrifugal methods employ a recycle stream that passes through the centrifuge. Hollow-fiber filters operated under tangential flow allow continuous removal of medium filtrate without significant membrane fouling. Gravitational settlers using inclined tubes allow cells in the outflowing stream to return to the reactor. Bierau et al. used ultrasound to aggregate cells for their fast separation from liquid medium in the settler.[32]

Microcarrier Cultures

As microcarriers provide a good surface area for attachment per unit volume, various types have been routinely used to grow anchorage-dependent cell lines in bioreactors primarily used for suspension cultures. Airlift reactors can also be operated using microcarriers. Wang et al. reported the use of a fluidized-bed

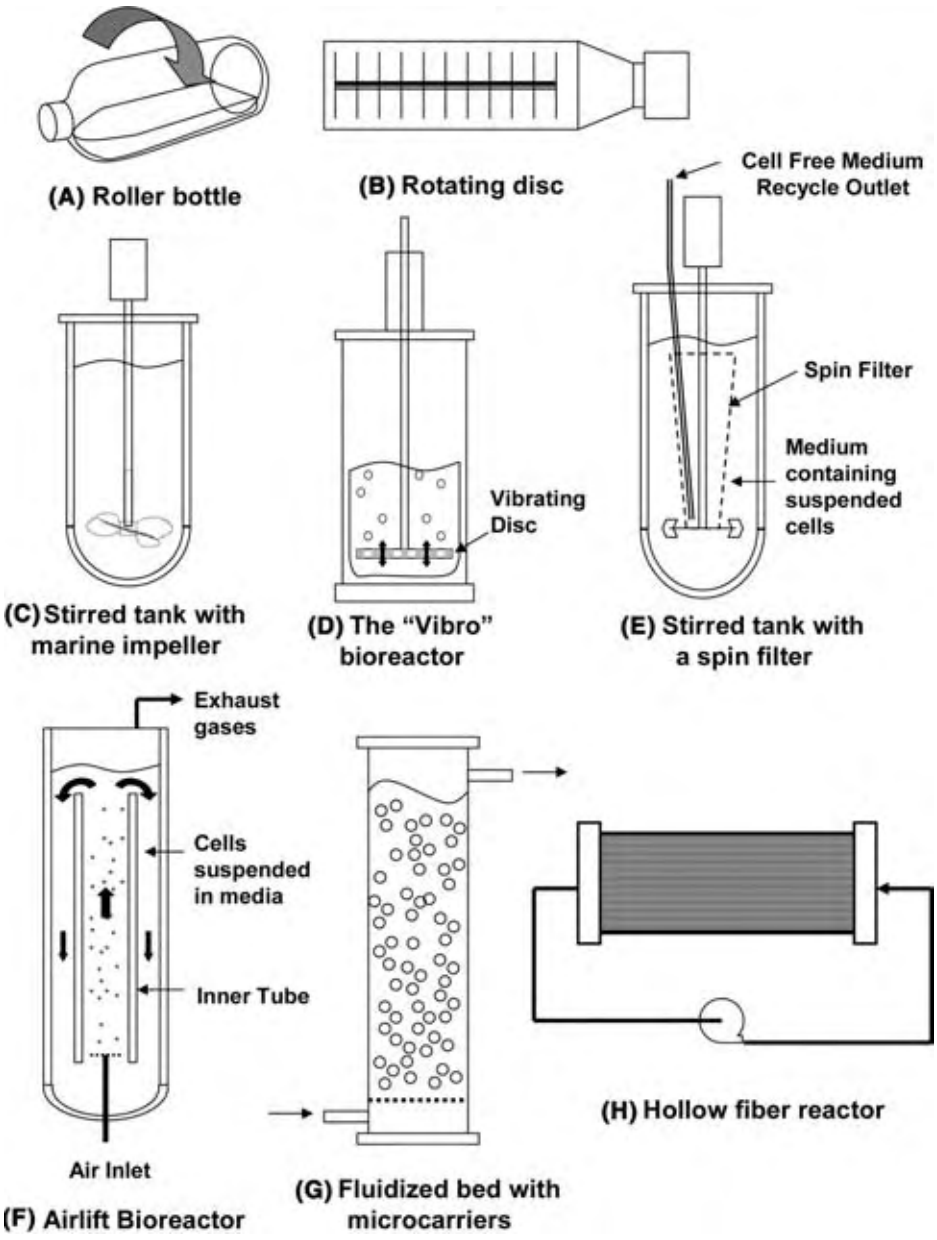


Fig. 6 Various types of animal cell bioreactors: (A) roller bottle; (B) rotating disk; (C) stirred tank with a marine impeller; (D) tank with a pulsating agitator; (E) stirred tank with a spin filter; (F) airlift; (G) fluidized bed; and (H) hollow fiber. *(View this art in color at www.dekker.com.)*

bioreactor in which cells were attached to Cytoline-1 macroporous microcarriers.[33] The upward flow of the medium fluidized the beads or carriers and provided a unique perfused system that gave higher erythropoietin (EPO) production than a conventional stirred-tank bioreactor. Fluidized-bed bioreactors have also been used for suspension cultures of hybridoma.[34] Chong et al. used microcarriers to form a packed-bed reactor to grow CHO cells and report cell density as high as 2×10^7 cells/ml.[35]

However, cells grown on solid microcarriers are often subjected to fluid mechanical damages caused by small turbulent eddies as well as by collision between microcarriers and against the impellers and other bioreactor parts.[36] The development of macroporous microcarriers has largely solved this problem of shear damage to cells. As the name suggests, macroporous microcarriers have a network of pores within them, which not only present a larger surface area for cell attachment but also protect the cells from shear damage. Macroporous microcarriers are mostly made of gelatin, collagen, cellulose, polystyrene, and polyethylene, and the cell distribution in them can be studied by confocal laser scanning microscopy.[37,38]

Hollow-Fiber Reactors

Hollow-fiber reactors are widely used in the production of MAb and can reach cell densities higher than

2×10^8 cells/ml of the fiber volume.[39,40] With continuous perfusion and intermittent harvesting, hollow-fiber reactors can give high reactor productivity and produce antibodies at a high concentration comparable to or even higher than that in mouse ascites.[41] The gradients of metabolites and nutrients created along the axis of the hollow-fiber reactor are undesirable. This problem can be overcome with the radial flow hollow-fiber reactor, which consists of a central flow distributor tube surrounded by an annular bed of hollow fibers.[42] The central flow distributor ensures an axially uniform radial convective flow of nutrients across the fiber bed. In this reactor, the cells grow on the outer side of the fibers. However, in conventional hollow-fiber reactors, the cells could be on either side depending on the inoculation method used and the way the process is carried out. Hollow-fiber reactors have also been used for perfusion cultures. A practical problem with these systems is that the cell density cannot be directly monitored. Hollow-fiber reactors are also expensive, and their uses are limited to small-to-medium scale production of antibodies.

Fibrous-Bed Bioreactors

A promising new technology has been the immobilization of animal cells to fibrous beds rather than microcarriers. The CelliGen Plus® bioreactor uses polyester fabric disks packed in a basket inside a stirred tank for immobilizing animal cells. The reactor productivity was reported to be as high as 12-fold of that in static and stirred suspension culture systems for antibody production.[43] Chen et. al. also developed a fibrous-bed bioreactor to successfully grow osteosarcoma cells to cell density as high as $3 \times 10^8\,ml^{-1}$ with 90% cell viability for as long as 4 mo.[44] The three-dimensional structure provided by the fibrous matrix has been shown to have profound effects on cell growth and protein production.[25,45] This three-dimensional culturing method provides a new technique for the scale-up of animal cell culture.

INDUSTRIAL APPLICATIONS

Animal cell cultures are increasingly being used for the production of recombinant glycoproteins, viral vaccines, and MAb in the biotechnology industry. They are far superior to yeast and bacterial cells in carrying out the complex post-translational modifications that major recombinant protein products require. Not only can animal cells provide better glycosylation and phosphorylation of complex proteins, but they can also carry out authentic proteolytic cleavage, subunit association, and chemical derivatization. However, large-scale animal cell cultures are more difficult to commercialize than microbial cultures because the growth rate of animal cell cultures is much slower, the nutrient requirements are more complex, and the growth conditions more stringent. In addition, there are several parameters that need to be considered. Not only is the required inoculum size large ($\sim 10^5\,ml^{-1}$ or $1–5 \times 10^4\,cm^{-2}$), but the cell proliferation rate is also much slower. The productivity of the target protein is also in the milligram per liter range as compared to the higher production (often in g/L) in microbial cells. The medium is usually more expensive and the cells are highly sensitive to toxic metabolites and shear. In the case of suspension cultures, the scale-up is relatively easy as a range of fermentation equipment developed for microbial cultures can be modified to adapt to animal cell cultures. However, anchorage-dependent cells require a large surface area per unit volume and are therefore more difficult to scale up.

In spite of these hurdles, the last two decades have seen an immense leap in animal cell culture technology both at the laboratory scale as well as the industrial scale. A variety of bioreactors and instrumentation have been ingeniously been devised for the scale up and process control of animal cell cultures. Serum-free media development has considerably reduced the downstream processing costs in the recombinant protein production and purification process. The capability to induce some cell lines to lose anchorage dependence has also been an important breakthrough.

Table 2 lists various types of biopharmaceutical products from animal cell cultures. Viral vaccines are usually produced by first culturing the host cells (e.g., MRC-5 and WI-38) to form a cell layer on the surface of substratum. Seed virus is then added and incubated for about 3 weeks for replication in the host cells without killing them. After washing to remove the medium components, the cells are lysed to release the virions for harvesting and purification. The inactivated viral vaccine is produced by inactivation with formaldehyde and adsorption onto aluminum hydroxide adjuvant.

Roller bottle reactors have been widely used in the past and can generate cell densities upto 5.4×10^6 cells/ml.[46] However, roller bottles are difficult to scale up and cannot meet the growing demand for therapeutic recombinant proteins. Their popularity is on the decline and are largely replaced by microcarriers, and stirred-tank or airlift bioreactors in process scale-up. Initially, industrial production of EPO by CHO cells is carried out in hundreds of roller bottles in incubation rooms equipped with robots for medium changes and product harvesting. The newer production plant for second-generation EPO employs state-of-the-art bioreactors and has three times the production capacity of the old EPO plant.

Table 2 Some important biopharmaceutical products from animal cell cultures

Type	Examples
Vaccines	Polio, hepatitis A, measles, mumps, rubella, yellow fever, rabies, and influenza
Glycoproteins	Interferons
	Blood clotting factors (factors VIII, IX)
	Glycoprotein hormones, EPO
	Plasminogen activators, t-PA
MAb	Diagnostics
	Therapeutics—prevention of respiratory syncytial virus (RSV) infection, treatments of inflammation, breast cancer, non-Hodgkin's lymphoma, and treatment or prevention of transplant rejection
Hormones	Human growth hormone, insulin, calcitonin, and parathyroid hormone
Growth factors	Nerve growth factor and EGF
Proteases	Urokinase

Monoclonal antibodies have been widely used in biomedical research and in diagnostics. Because MAb bind to specific cell surface receptors, they can be used for treatments of transplant rejection, cancer, autoimmune and inflammatory diseases, and infectious diseases. Currently, MAb products comprise about 25% of all biotech drugs in clinical development. Commercial MAb production uses two methods: 1) in vivo cultivation in mouse or rabbit ascites and 2) in vitro cell culture in tissue flasks or bioreactors. For mass production of therapeutic antibodies, the latter method is used. For example, recombinant paliviumab is produced by culturing murine myeloma cells (NS0) in a stirred-tank fed-batch bioreactor. The manufacturing process starts with a vial containing about 10 million frozen cells, which are cultured in T-flask and then in spinner flask to expand the number of cells. These cells are subsequently used in the larger-scale bioreactor process with the culture volume increased incrementally to the final volume of 10,000 L. After inoculation of the production bioreactor, the fermentation takes about 20 days to reach the final titer of the MAb product, which is about 1 g of MAb per liter of the culture medium. The high titer of MAb production in this system was accomplished through extensive research and development works on cell line improvement, medium optimization, and process optimization and control.

CONCLUSIONS

Animal cells in culture have been widely used to study the physiology, metabolism, and biochemistry of cells; test the effect of compounds on different cell types; in tissue engineering applications; and for the production of recombinant glycoproteins, viral vaccines, and MAb. Though several challenges remain to be overcome, the future of developing more products using animal cell cultures is very bright.

REFERENCES

1. Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K.; Watson, J.D. *Molecular Biology of the Cell*, 3rd Ed.; Garland Publishing Inc.: New York, 1994.
2. Prokop, A. Implications of cell biology in animal cell biotechnology. In *Animal Cell Bioreactors*; Ho, C.S., Wang, D.I.C., Eds.; Butterworth-Heinemann: Stoneham, MA, 1991; 21–58.
3. Butler, M.J. *Animal Cell Culture and Technology: The Basics*; Oxford University Press Inc.: New York, 1996.
4. Chu, L.; Robinson, D.K. Industrial choices for protein production by large-scale cell culture. Curr. Opin. Biotechnol. **2001**, *12*, 180–187.
5. Lazar, A.; Silberstein, L.; Reuveny, S.; Mizrahi, A. Microcarriers as a culturing system of insect cells and insect viruses. Dev. Biol. Stand. **1987**, *66*, 315–323.
6. Griffiths, J.B. Scaling up of animal cell cultures. In *Animal Cell Culture: A Practical Approach*; Freshney, R.I., Ed.; IRL Press at Oxford University Press: Oxford, 1992; 47–93.
7. Zuehlke, A.; Roeder, B.; Widdecke, H.; Klein, J. Synthesis and application of new microcarriers for animal cell culture. Part I: design of polystyrene based microcarriers. J. Biomater. Sci. **1993**, *5* (1–2), 65–78.
8. Freshney, R.I. *Culture of Animal Cells—A Manual of Basic Technique*, 4th Ed.; Wiley-Liss Inc.: New York, 2000.

9. Harrison, M.A.; Rae, I.F. *General Techniques of Cell Culture*; Cambridge University Press: Cambridge, 1997.
10. Butler, M.; Christie, A. Adaptation of mammalian cells to non-ammoniagenic media. Cytotechnology **1994**, *15*, 87–94.
11. Good, N.E.; Winget, G.D.; Winter, W.; Connolly, T.N.; Izawa, S.; Singh, R.M.M. Hydrogen ion buffers and biological research. Biochemistry **1966**, *5*, 467–477.
12. Song, J.H.; Harris, M.S.; Shin, S.H. Effects of fetal bovine serum on ferrous ion-induced oxidative stress in pheochromocytoma (PC12) cells. Neurochem. Res. **2001**, *26* (4), 407–413.
13. Geng, Y.; D'Souza, S.; Xin, H.; Walter, S.; Choubey, D. p202 levels are negatively regulated by serum growth factors. Cell Growth Differ. **2000**, *11* (9), 475–483.
14. Keenan, J.; Meleady, P.; Clynes, M. Serum free media. In *Animal Cell Culture Techniques*; Clynes, M., Ed.; Springer: Berlin, 1998; 54–56.
15. Ito, Y.; Uno, T.; Liu, S.Q.; Imanishi, Y. Cell growth on immobilized cell growth factor. Biotechnol. Bioeng. **1992**, *40*, 1271–1276.
16. Chen, J.; LaBrake, S.; McClure, D. Functional roles of high density lipoproteins in serum-free medium. In *Growth and Differentiation of Cells in Defined Environment*, Proceedings of the International Symposium, 1984; Murakami, H., Ed.; Kodansha: Tokyo, Japan, 1985.
17. Eagle, H. The effect of environmental pH on the growth of normal and malignant cells. J. Cell Physiol. **1973**, *82*, 1–8.
18. Ducommun, P.; Rueux, P.-A.; Kodouri, A.; Von Stockar, U.; Marison, I.W. Monitoring of temporary effects on animal cell metabolism in a packed bed process. Biotechnol. Bioeng. **2002**, *77*, 838–842.
19. Mamdouh, Z.; Giocondi, M.-C.; Laprade, R.; Le Grimelle, C. Temporary dependence of endocytosis in renal epithelial cells in culture. Biochim. Biophys. Acta Biomembr. **1996**, *1282*, 171–173.
20. Hendrick, V.; Winnepenninckx, P.; Abdelkafi, C.; Vandeputte, O.; Cherlet, M.; Marique, T.; Renemann, G.; Loa, A.; Kretzmer, G.; Werenne, J. Increased productivity of recombinant tissue plasminogen activator (tPA) by butyrate and shift of temperature: a cell cycle analysis. Cytotechnology **2001**, *36*, 71–83.
21. Spier, R.E.; Griffiths, B. An examination of the data and concepts germane to the oxygenation of cultured animal cells. Proceedings of the 5th Meeting of ESACT, Copenhagen, Denmark, 1982. Karger: Basel, 1982; 81–92.
22. Miller, W.M.; Wilke, C.R.; Blanch, H.W. Effects of dissolved oxygen concentration on hybridoma growth and metabolism in continuous culture. J. Cell Physiol. **1987**, *132*, 524–530.
23. Mizrahi, A. Oxygen in human lymphoblastoid cell line cultures and effect of polymers in agitated and aerated cultures. Proceedings of the 5th Meeting of the ESACT; Copenhagen, Denmark, 1982. Karger: Basel, 1982; 93–102.
24. Ma, T.; Yang, S.-T.; Kniss, D.A. Oxygen tension influences proliferation and differentiation in a tissue engineered model of placental trophoblast-like cells. Tissue Eng. **2001**, *7*, 495–506.
25. Luo, J. Three Dimensional Culturing of Animal Cells in Fibrous Bed Bioreactor. Ph.D. Thesis, The Ohio State University, 2002.
26. Fassnacht, D.; Rössing, S.; Singh, R.P.; Al-Rubeai, M.; Pörtner, R. Influence of bcl-2 on antibody productivity in high cell density perfusion cultures of hybridoma. Cytotechnology **1999**, *30*, 95–106.
27. Griffiths, B.J. Scale up of suspension and anchorage dependent animal cells. In *Methods in Molecular Biology, Animal Cell Culture*; Pollard, J.W., Walker, J.M., Eds.; Humana Press: Clifton, NJ, 1989; Vol. 5.
28. Kamen, A.A.; Tom, R.L.; Caron, A.W.; Chavarie, C.; Massie, B.; Archambault, J. Culture of insect cells in a helical ribbon impeller bioreactor. Biotechnol. Bioeng. **1991**, *38*, 619–628.
29. Shi, Y.; Ryu, D.D.Y.; Park, S.H. Performance of mammalian cell culture bioreactor with a new impeller design. Biotechnol. Bioeng. **1992**, *40*, 260–270.
30. Huelscher, M.; Onken, U. Influence of bovine serum albumin on the growth of hybridoma cells in airlift loop reactors using serum-free medium. Biotechnol. Lett. **1988**, *10*, 689–694.
31. Freed, L.; Langer, R.; Martin, I.; Pellis, N.R.; Vunjak-Novakovic, G. Tissue engineering of cartilage in space. Proc. Natl. Acad. Sci. USA **1997**, *94*, 13885–13890.
32. Bierau, H.; Perani, A.; Al-Rubeai, M.; Emery, A.N. A comparison of intensive cell culture bioreactors operating with hybridomas modified for inhibited apoptotic response. J. Biotechnol. **1998**, *62*, 195–207.
33. Wang, M.-D.; Yang, M.; Huzel, N.; Butler, M. Erythropoietin production from CHO cells grown by continuous culture in a fluidized bed bioreactor. Biotechnol. Bioeng. **2002**, *77*, 194–203.
34. Ray, N.G.; Tung, A.S.; Ozturk, S.S.; Pang, R.H.L. Animal cell culture using fluidized-bed culture technology. Food Bioprod. Process. **1993**, *71* (C2), 124–126.

35. Chong, C.; Chang, Y.; Deng, J.; Xiao, C.; Su, Z. A novel scale up method for mammalian cell culture in packed-bed bioreactor. Biotechnol. Lett. **2001**, *23*, 881–885.
36. Janowski, T.; Grajek, W. Mechanical damage of animal cells in bioreactors. Biotechnologia **1993**, *3*, 166–183.
37. Shiragami, N.; Honda, H.; Unno, H. Anchorage-dependent animal cell culture by using a porous micro-carrier. Bioprocess. Eng. **1993**, *8* (5–6), 295–299.
38. Bancel, S.; Hu, W.-S. Confocal laser scanning microscopic examination of cell distribution in macroporous microcarriers. Biotechnol. Prog. **1996**, *12*, 398–402.
39. Gloeckner, H.; Lemke, H.-D. New miniaturized hollow fiber bioreactor for in vivo like cell culture, cell expansion, and production of cell derived products. Biotechnol. Prog. **2001**, *17*, 828–831.
40. Patankar, D.; Oolman, T. Wall-growth hollow-fiber reactor for tissue culture: I. Preliminary experiments. Biotechnol. Bioeng. **1990**, *36*, 97–103.
41. Gramer, M.J.; Britton, T.L. Antibody production by a hybridoma cell line at high cell density is limited by two independent mechanisms. Biotechnol. Bioeng. **2002**, *79*, 277–283.
42. Tharakan, J.P.; Chau, P.C. A radial flow hollow fiber bioreactor for the large-scale culture of mammalian cells. Biotechnol. Bioeng. **1986**, *28*, 329–342.
43. Wang, G.; Zhang, W.; Jacklin, C.; Freedman, D.; Eppstein, L.; Kadouri, A. Modified Celligen-packed bed bioreactors for hybridoma cell cultures. Cytotechnology **1992**, *9*, 41–49.
44. Chen, C.; Huang, Y.L.; Yang, S.-T. A fibrous bed bioreactor for continuous production of developmental endothelial Locus-1 by osteosarcoma cells. J. Biotechnol. **2002**, *97*, 23–29.
45. Chen, C.; Huang, Y.L.; Yang, S.-T. Effects of three-dimensional culturing on osteosarcoma cells grown in a fibrous matrix: analyses of cell morphology, cell cycle, and apoptosis. Biotechnol. Prog. **2003**, *19*, 1574–1582.
46. Berson, R.E.; Pieczynski, W.J.; Svihla, C.K.; Hanley, T.R. Enhanced mixing and mass transfer in a recirculation loop results in high cell densities in a roller bottle reactor. Biotechnol. Prog. **2002**, *18*, 72–77.

Antioxidants

S. Al-Malaika
H. H. Sheena
Polymer Processing and Performance Research Unit, School of Engineering and Applied Science, Aston University, Birmingham, U.K.

INTRODUCTION

Antioxidants are ubiquitous and essential ingredients to life and living matter, as well as natural and synthetic organic materials. This entry deals with the fundamental and applied aspects of hydrocarbon oxidation and antioxidants with particular emphasis on polymers. Some aspects of polymer oxidation are introduced first before discussing antioxidants and their mechanisms of action in polymer stabilization. Examples of major antioxidant classes are given together with critical evaluation of performance in selected polymers. Recent progress in the areas of biological, reactive, and macromolecular antioxidants are also presented.

OXIDATION AND ANTIOXIDANTS

Hydrocarbon Oxidation

Molecular oxygen is the major cause of irreversible deterioration of hydrocarbon substrates, leading to the loss of useful properties and to the ultimate failure of the substrate. The oxidation process of hydrocarbons is autocatalytic: oxidation starts slowly, sometimes with a short induction period, followed by a gradual increase in the rate, concomitant with the build up of hydroperoxides, which eventually subside, giving rise to a sigmoidal oxidation curve.[1–6] When initiators such as peroxides are present, the length of the induction period is absent, or very short, but it can be prolonged by antioxidants, as shown in Fig. 1. The basic autoxidation theory of hydrocarbons involves a complex set of elementary reaction steps in a free radical-initiated chain reaction mechanism; the basic tenets of this theory apply equally to polymer oxidation.

Polymer Oxidation

The basic autoxidation theory of hydrocarbons, which involves a complex set of elementary reaction steps: initiation, propagation, and termination, is similarly valid for polymer oxidation. Other factors such as heat, mechanical stress, light, and transition metal impurities, contribute to the initiation step, which leads to the formation of the first macro-alkyl radicals (see Scheme 1, Reaction 1).

Propagation reactions involve the very fast reaction of oxygen biradical ($O^{\bullet}$–$O^{\bullet}$) with polymer alkyl radicals leading to the formation of macro-alkyl peroxyl radicals (Scheme 1, Reaction 2). This is followed by abstraction of hydrogen from another macromolecule resulting in hydroperoxide formation, the first molecular product of the chain oxidation process (Scheme 1, Reaction 3). This reaction involves the breaking of a C–H bond and, therefore, requires higher activation energy than Reaction 2. The rate of Reaction 3, which in most polymers at normal oxygen pressures determines the overall oxidation rate, is a function of both the C–H bond dissociation energy (allyl < benzyl < tertiary < secondary < primary) and the stability of the resulting macro-alkyl radical.[5] The macro-hydroperoxides formed can undergo homolysis under the effect of heat, light, or metal ions giving rise to alkoxyl- and hydroxyl-macroradicals (Scheme 1, Reaction 4). Both these radicals can abstract hydrogen from another polymer molecule, leading to new macro-alkyl radicals (Reactions 5 and 6a), which continue the chain reaction. Alkoxyl radicals can undergo further reactions, e.g., β-scission (see Reaction 6b), that would lead to cleavage of the macromolecular backbone and the generation of further radicals.

The oxidative process is terminated through radical combination and disproportionation reactions; the exact nature of the terminating step is determined by the polymer structure and the oxygen concentration. Since Reaction 3 (Scheme 1) is rate determining, alkylperoxyl radicals are the predominant reactive species under normal oxygen pressure (oxygen saturation), i.e., $[ROO^{\bullet}] > [R^{\bullet}]$, and termination occurs primarily through Reaction 7, giving rise to diperoxides, carbonyl compounds, and alcohols.[7] By contrast, under oxygen-deficient conditions, alkyl radicals predominate, i.e., $[R^{\bullet}] > [ROO^{\bullet}]$, and bimolecular termination steps 8–10 dominate, leading to cross-linking (and increased molar mass, Reaction 9) and/or disproportionation (with no

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007666

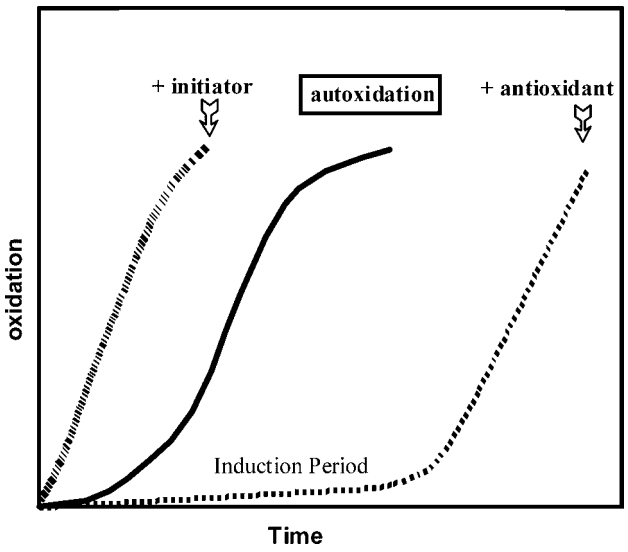
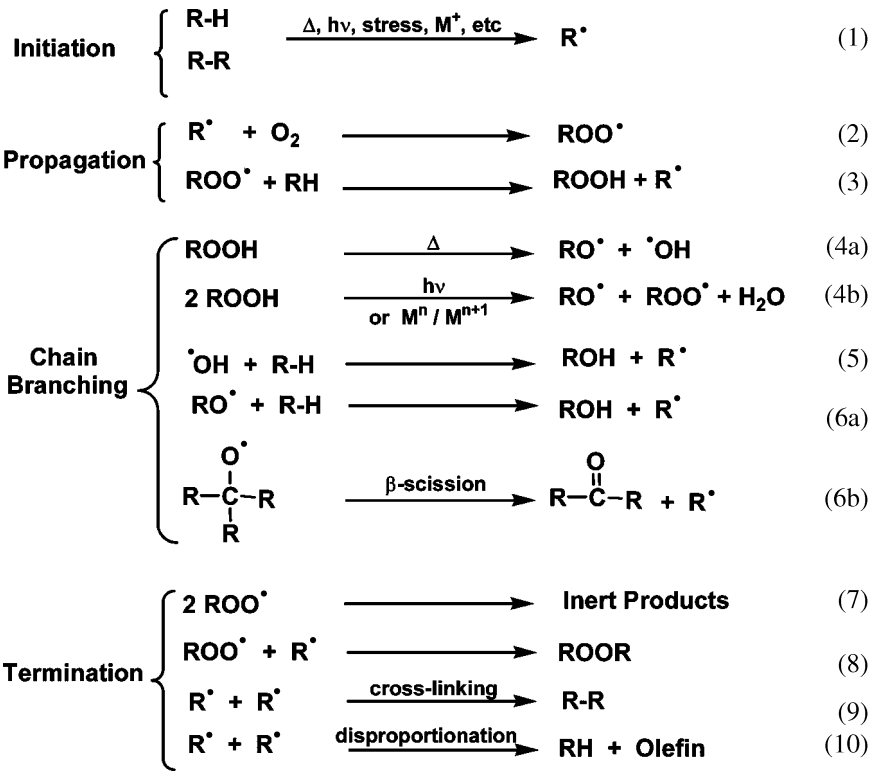


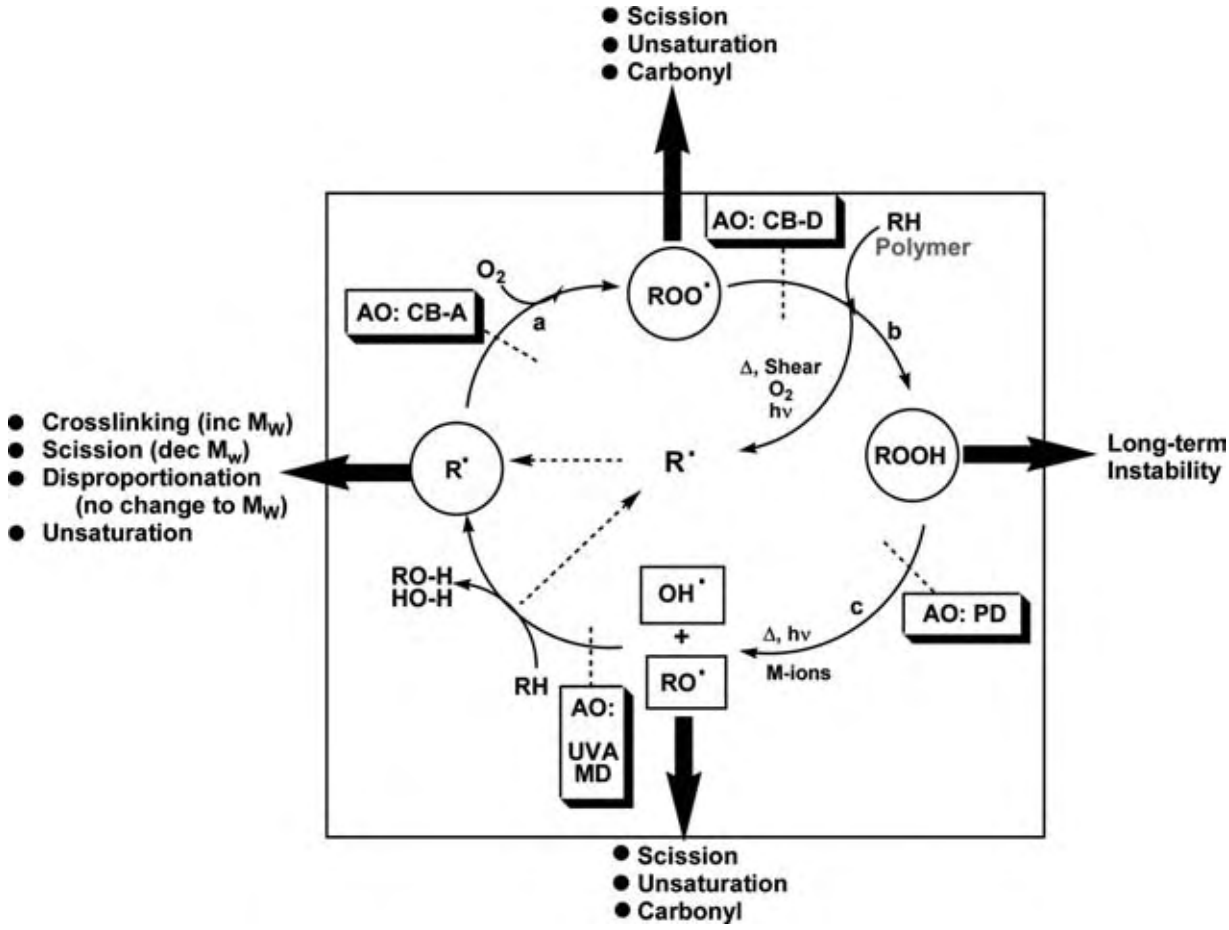
Fig. 1 Generalized shape of autoxidation curve and changes caused by the addition of an initiator and an antioxidant.

change in molar mass Reaction 10) polymer reactions (see Scheme 1). Both "limiting conditions'' are encountered throughout the polymer lifecycle: oxygen-rich environment during end-product use, and oxygen-deficient conditions during polymer processing (e.g., extrusion), as well as in thick cross-sections, when the rate of oxidation is controlled by the oxygen diffusion rate. Other radical-polymer reactions, including fragmentation and addition to double bonds to give rise to further free radicals, can also take place depending on the reaction conditions.[8–10] These macromolecular changes lead ultimately to the loss of mechanical properties (e.g., impact strength, tensile strength, elongation), changes in surface appearance (e.g., yellowing, crack formation, loss of gloss, "chalking''), and premature failure. The deleterious effect of molecular oxygen is accelerated by many other factors: sunlight, heat, ozone, atmospheric pollutants, water, mechanical stress, adventitious metal, and metal ion contaminants. Further, the prior thermal-oxidative history of polymers determines, to a large extent, their photoxidative behavior in service during outdoor weathering.[11] The cyclical oxidation pathways and the nature and effects of oxidation products arising from the different propagating species are highlighted in Scheme 2.

In the presence of oxygen, mechanical, thermal, photochemical, and environmental factors are extremely detrimental to polymer properties: during processing, storage, and first and subsequent lives as end-use products. Inhibition of the oxidative degradative process in polymers can be achieved by the incorporation of low levels (0.03–2%, but for plastics usually at 0.03–0.3%) of antioxidants and stabilizers during the fabrication process. Antioxidants and stabilizers, therefore, occupy a key position in the market of compounding ingredients for polymers, in particular the high volume commodity polymers, e.g., polypropylene, polyethylene, and polyvinyl chloride.



Scheme 1 Basic autoxidation reactions.



Scheme 2 Schematic presentation of the cyclical oxidation process and some of the main reactions/products formed from the propagating radicals. The antioxidant mechanisms interrupting the oxidative cycles are also shown.

Antioxidants and Mechanisms of Action

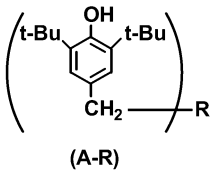
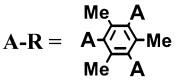
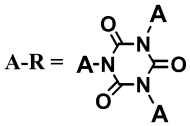
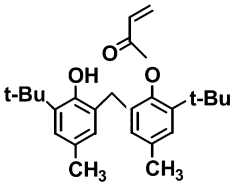
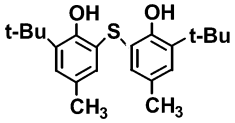
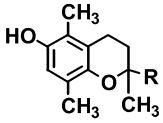
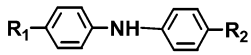
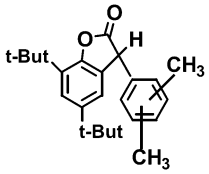
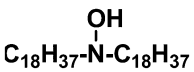
Antioxidants are chemical compounds or species that interfere with the degradative oxidative cycles in polymers, through chemical processes involving inhibition or retardation mechanisms. The two major antioxidant mechanisms, chain breaking and the preventive, are outlined in Scheme 2.[12,13] Examples of commercial antioxidants are shown in Table 1, under their primary mode of action, albeit most antioxidants operate by more than one mechanism.

Chain breaking (CB) antioxidants (sometimes referred to as primary antioxidants) act by removing the propagating radicals (the alkyl peroxyl $ROO^{\bullet}$ and the alkyl $R^{\bullet}$) formed during the primary oxidation cycle.[12,14,15] Chain breaking donor (CB-D) antioxidants operate by reducing the radical $ROO^{\bullet}$ to ROOH (Reaction 1a). Therefore, good CB-D antioxidants (e.g., hindered phenols, AH) must be able to compete effectively with the chain propagating Reaction 1b. The antioxidant radical ($A^{\bullet}$) produced from reaction 1a should also lead to stable molecular products. Hindered phenols and aromatic amines (see Table 1, AOs 1–12) are important examples of commercial CB-D antioxidants. Chain breaking acceptor antioxidants (CB-A) operate by oxidizing $R^{\bullet}$ in a stoichiometric reaction (Reaction 2a) and are effective only under oxygen deficient conditions (i.e., must be able to compete effectively with the chain propagating Reaction 2b). Quinones and stable free radicals, which can act as alkyl radical trapping agents are representative examples of CB-A antioxidants. Hindered amine derivatives (alias hindered amine light stabilizers, HALS, e.g., AOs 25–27, Table 1) operate also by a chain breaking mechanism and, through their transformation products, are able to trap both $R^{\bullet}$ and $ROO^{\bullet}$ through a cyclical regenerative mechanism,[16–22] see Scheme 3.

$$AH + POO^{\bullet} \xrightarrow{CB\text{-}D} POOH + A^{\bullet} \quad \text{(Stabilization Reaction)} \quad \text{Non-radical products} \qquad (1a)$$

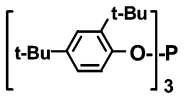
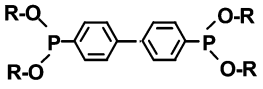
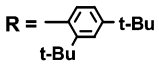
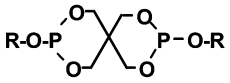
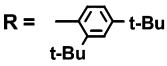
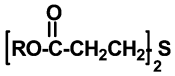
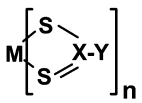
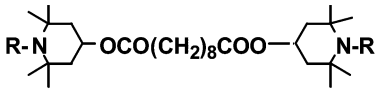
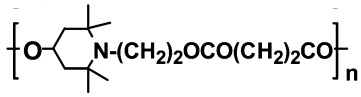
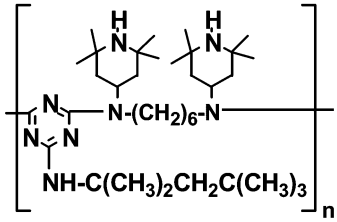
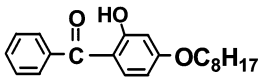
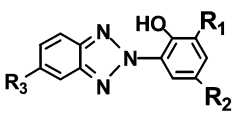
$$P\text{-}H + POO^{\bullet} \longrightarrow POOH + P^{\bullet} \qquad (1b)$$

Table 1 Some examples of commercial antioxidants representing the major antioxidant mechanisms

Antioxidant	Code and trade name		
CHAIN BREAKING ANTIOXIDANTS			
Hindered Phenols			
 (A-R)	A-R = A-H	AO-1	BHT
	A-R = 	AO-2	Irganox 1330
	A-R = 	AO-3	Irganox 3114
	A-R = A-$CH_2CO_2C_{18}H_{37}$	AO-4	Irganox 1076
	A-R = [A-$CH_2CO_2CH_2$]$_4$C	AO-5	Irganox 1010
	A-R = [A-$CH_2CONH(CH_2)_3$]$_2$	AO-6	Irganox 1098
	A-R = [A-$CH_2COO(CH_2)_3$]$_2$	AO-7	Irganox 259
		AO-8	Irganox 3052
		AO-9	Irganox 1081
	R = -$(CH_2)_3$CH(CH_2)$_3$CH(CH_2)$_3$CH	AO-10	Irganox E 201
Aromatic Amines			
	R1 = R2 = t-Oct	AO-11	Nonex OD
	R1 = H, R2 = NH-	AO-12a	Nonex DPPD
	R1 = H, R2 = NHCOC(CH_3)=CH_2	AO-12b	——
Lactones			
		AO-13	HP 136
Hydroxylamines			
		AO-14	Irgastab FS 042

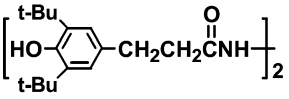
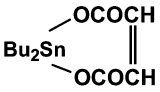
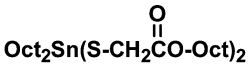
(Continued)

Table 1 Some examples of commercial antioxidants representing the major antioxidant mechanisms *(Continued)*

Antioxidant	Code and trade name		
PREVENTIVE AND PHOTO-ANTIOXIDANTS			
Peroxide Decomposers			
$(RO)_3P$	AO-15a	Ultranox TLP	Phosphite esters
R= $-C_{12}H_{25}$	AO-15b	Irgafos TNPP	
R= $-C_9H_{19}$			
	AO-16	Irgafos 168	
 R = 	AO-17	Irgafos P-EPQ	
 R = 	AO-18	Ultranox U- 626	
 $R = C_{18}H_{37}$ $R = C_{12}H_{25}$	AO-19 AO-20	Irganox PS 802 Irganox PS 800	Sulfur Compounds
 X=C, Y=NR_2, R = C_4H_9, M=Zn, n=2 X=C, Y=NR, M=Fe, n=3 X=P, Y=OR_2, M=Ni, n=2 X=C, Y=OR, M=Ni, n=2	AO-21 AO-22 AO-23 AO-24	Robec Z bud Iron dithiocarbamate Nickel dithiophosphate Nickel xanthate	
Photo-antioxidants, Hindered amine stabilizers			
 R = H R = CH_3	AO-25 AO-26a	Tinuvin 770 Tinuvin 765	HALS
	AO-26b	Tinuvin 626	
	AO-27	Chimassorb944	
UV-Absorbers			
	AO-28	Chimassorb 81	UVA
 $R_1 = R_2 = t\text{-}Bu$, $R_3 = H$ $R_1 = C_{12}H_{25}$, $R_2 = CH_3$, $R_3 = H$ $R_1 = R_3 = H$, $R_2 = CH_3$ $R_1 = t\text{-}Bu$, $R_2 = CH_3$, $R_3 = Cl$	AO-29 AO-30 AO-31 AO-32	Tinuvin 320 Tinuvin 571 Tinuvin P Tinuvin 326	

(Continued)

Table 1 Some examples of commercial antioxidants representing the major antioxidant mechanisms *(Continued)*

Antioxidant	Code and trade name		
Metal Deactivators			
	AO-33	Irganox MD 1024	MD
	AO-34	Eastman OABH	
Hydrogen Chloride Scavengers			
	AO-35	Dibutyl tin maleate	HCl scavenger
	AO-36	Dioctyltin thioglycolate	

$$LA + P^{\bullet} \xrightarrow[(O_2\ \text{deficiency})]{CB\text{-}A} \text{Non-radical products} \quad (2a)$$

(Stabilisation Reaction)

$$O_2 + P^{\bullet} \longrightarrow POO^{\bullet} \quad (2b)$$

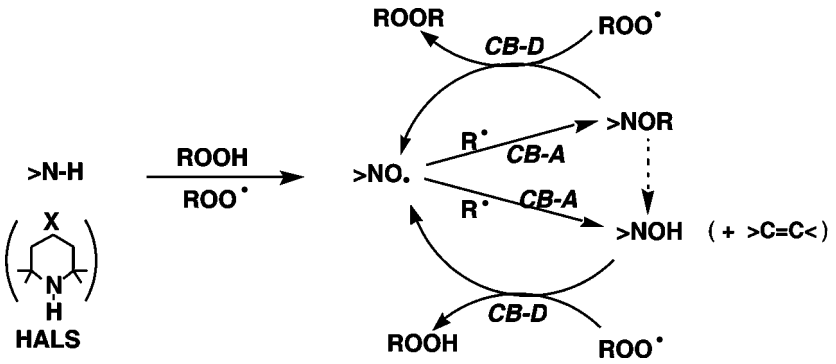
(Propagation Reaction)

where AH is a CB-D antioxidant, $A^{\bullet}$ is the antioxidant radical and LA is a CB-A antioxidant, PH is a polymer.

Preventive antioxidants (sometimes referred to as secondary antioxidants) act by interrupting the secondary oxidation cycle to prevent or inhibit the generation of free radicals.[13] The most important preventive mechanism is the nonradical hydroperoxide decomposition, PD. Phosphite esters and sulfur-containing compounds, e.g., AOs 15–24, Table 1, are the most important groups of peroxide decomposers. Phosphite esters are known as stoichiometric peroxide decomposers (PD-S): they reduce hydroperoxides to alcohols and are themselves oxidized to the corresponding phosphates in 1 : 1 stoichiometric reactions (see Reaction 3). Additionally, some phosphite esters can act as catalytic peroxidolytic agents (PD-C), whereas others (e.g., hindered aryl phosphites, AOs 11 and 12, Table 1) can act by a chain breaking (CB) mechanism; the relative contribution of each of these modes of action to the overall mechanism depends on the structure of the phosphite, the oxidizability of the substrate and the reaction conditions.[23,24]

$$P(OR)_3 + ROOH \xrightarrow{PD\text{-}S} O{=}P(OR)_3 + ROH \quad (3)$$

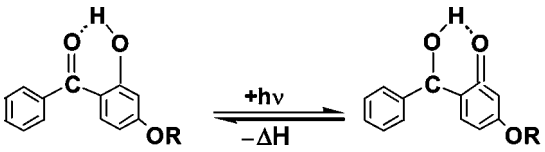
Sulfur compounds are known as catalytic hydroperoxide decomposers (PD-C): one antioxidant molecule destroys several hydroperoxide molecules by the action of intermediate sulfur acid moieties.[25–28] Thioethers and esters of thiodipropionic acid and metal dithiolates are examples of commercial significance (see Table 1, AOs 19–24).



Scheme 3 Reaction mechanism of HALS.

Metal deactivators (MD) act, primarily, by retarding metal-catalyzed oxidation of polymers and are added to polymers used in contact with metals, e.g., wires and power cables. Metal deactivators are normally polyfunctional metal chelating compounds (e.g., Table 1, AOs 33 and 34) that can chelate with metals and decrease their catalytic activity.[29–31]

UV absorbers (UVA) act by absorbing UV light to retard the photolysis of hydroperoxides. Typical examples are based on 2-hydroxybenzophenones (AO 28) and 2-hydroxybenzotriazoles (e.g., Table 1, AOs 29–32); both are photo-stable with high molar absorptions over the region 300–360 nm. Their activity is based essentially on absorption of the harmful UV-radiation and its harmless dissipation as heat. For example, in 2-hydroxybenzophenones, UV-light induces intramolecular hydrogen transfer to yield an enol, which reverts back to the original ketone in a radiationless process,[32] Reaction 4.



(4)

ANTIOXIDANT PERMANENCE: EFFECTS OF CHEMICAL AND PHYSICAL FACTORS

There are many limitations to the use of antioxidants in applications involving the human environment, e.g., food packaging, medical applications, etc. Apart from the risks associated with premature failure of a product, the physical loss of antioxidants into the contact media can also have adverse toxicological effects. Although all antioxidants, which are licensed for use in polymers for food contact and medical applications, have to undergo strict toxicity testing regimes, their approval, however, does not necessarily mean that their oxidation products (derived from the parent antioxidant during processing or as a result of its antioxidant action in the substrate) are also nontoxic.

Low molar mass antioxidants are known to suffer from physical loss and research in recent years has, therefore, focused on finding substantive alternatives. For example, reactive antioxidants, which can be chemically anchored on the polymer backbones to prevent physical loss during processing or in-service, have been explored as nonmigratory antioxidants.[33–39] Clearly, there are many factors that need to be considered before choosing antioxidants in end-use applications. The success of an antioxidant package is critically dependent on the underlying chemical and physical factors that influence antioxidant performance in different substrates and environments.

Chemical Factors

The intrinsic chemical activity of an antioxidant is a function of its molecular structure. This can be determined accurately in a model substrate where the antioxidant is fully soluble and in a test where no physical loss is possible. However, using the chemical activity alone as an indicator, can lead to unreliable predictions of the efficiency of antioxidants in polymers under practical conditions because of the dominating influence of physical factors under certain environments. For example, the hindered phenol BHT (AO 1, Table 1) is amongst the most efficient antioxidants known for liquid hydrocarbons (based on its intrinsic chemical activity determined by oxygen absorption), but is ineffective in protecting thermoplastic polymers (during accelerated air oven ageing) because of physical loss through volatilization.[11] Higher molecular mass antioxidants, based on the same hindered phenol function, e.g., Irganox 1010, Irganox 1330 (Table 1, AOs 5 and 2) outperforms BHT under these conditions.[40–42]

Physical Factors

The physical behavior of an antioxidant is a major factor affecting its permanency, efficiency, and acceptability, especially when the polymer (e.g., fibers, films) artifact is placed directly in contact with aggressive environments, e.g., organic solvents including dry cleaning solvents, hot water and detergents, hot oils, acids, fatty food, hot air, and intense solar radiation. Physical factors, which control the permanence and efficacy of an antioxidant include distribution, solubility, diffusivity, volatility, and leachability.

Distribution of antioxidants and polymer morphology

In order to inhibit the oxidation of polymers, the antioxidant has to be present in sufficient concentration at the various oxidation sites. In this respect, both the distribution of antioxidants and the morphology of the host polymer assume greater significance. Examination of the distribution of photo-antioxidants in typical commercial semi-crystalline polymers, such as polyolefins, has shown[43,44] that they are rejected into the amorphous region on the boundaries of spherulites. Such nonuniform distribution of antioxidants leads to an increase in their concentration in the amorphous region,[43] which is most susceptible to oxidation (the crystalline phase is normally impermeable to oxygen). However, in the case of polymer blends, a nonuniform distribution of antioxidants can undermine the overall stability of the blend, especially when the more oxidizable component of the polymer blend is left unprotected.

Compatibility of antioxidants with polymers

Antioxidants are generally less soluble in polymers than in lower molar mass liquid models. Although antioxidants are usually highly soluble at the elevated processing temperatures (and form with the polymer a homogenous solution), they come out of solution on cooling at room temperature with the solidified polymer supersaturated with the antioxidant. In turn, the antioxidant may precipitate as a separate phase, and exudes to the polymer surface (this is called ''blooming''),[43] leading to a build up of a concentration gradient near the surface and greater tendency of migration of the antioxidant from the bulk. Consequently, an antioxidant with low solubility and high diffusion rate is susceptible to blooming and loss to the surrounding medium by evaporation (air stream as contact media) or through leaching (extraction by a liquid contact media), leaving behind an unprotected polymer surface.

Generally, the compatibility of antioxidants in polymers is improved when the antioxidant and the host polymer have similar characteristics. Compatibility of antioxidants in nonpolar hydrocarbon polymers, therefore, decreases with increasing antioxidant polarity and increases with the number, length, and branching of the inert alkyl substituents attached to the antioxidant function.[45,46] Many commercial antioxidants with higher molecular masses (e.g., Table 1, compare AO 5 with AO 1) have been developed and many have inert long (8 to 18 C-atoms) alkyl chains (e.g., Table 1, AO 4).

Antioxidant diffusion, volatility, and leachability

The permanency of antioxidants is affected by diffusion characteristics of the antioxidant, the nature (gas, liquid, and solid) of the surrounding medium, and the temperature.[43,45,47] Generally, the diffusion coefficient of antioxidants decreases with increasing polar interactions with the polymer, increasing molar mass of antioxidants and branching in their alkyl side chain.

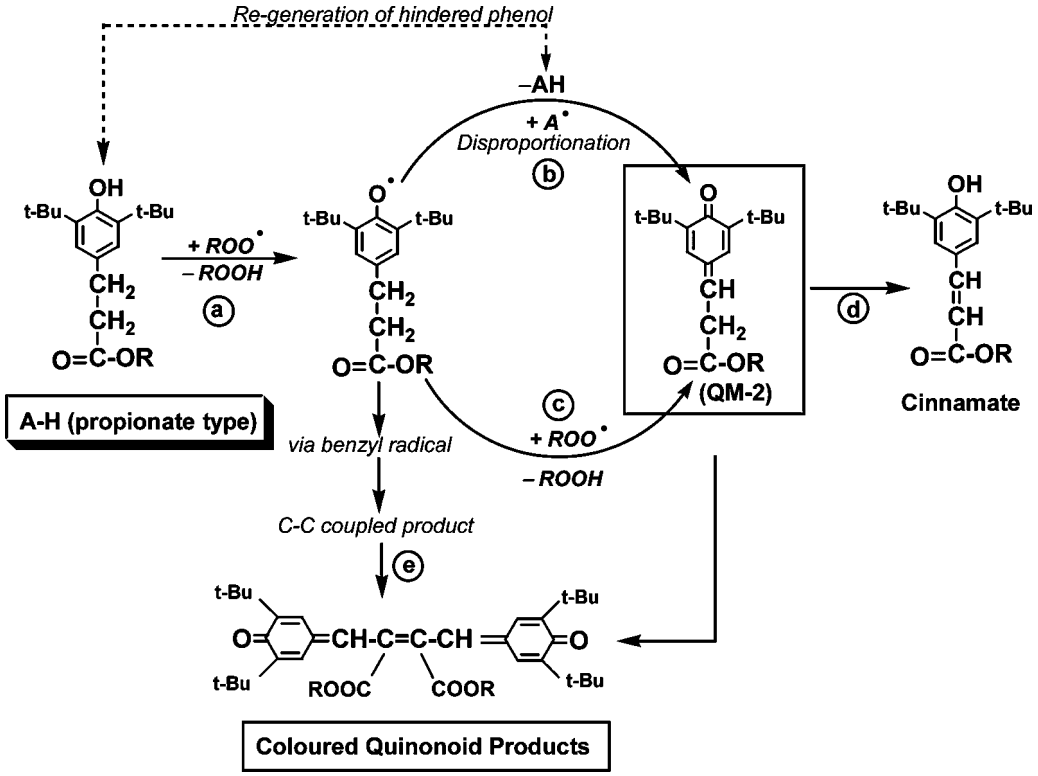
In the presence of a stream of hot air or high temperature and low pressure (e.g., during polymer melt processing) volatility becomes very important: it is governed by the rate of diffusion of antioxidants, which, in turn, determines the rate at which the surface is replenished.[43] The influence of polymer sample shape and the structure and molar mass of antioxidants on volatility has received much attention. The rate of evaporation of antioxidants from rubber and polyethylene, for example, was found[45] to be inversely proportional to the thickness of the sample and directly proportional to its surface area. Furthermore, an increase in molar mass results in an increase in intermolecular dispersion forces, which brings about a decrease in volatility of the antioxidants from the polymers.

In liquid contact media (e.g., foodstuffs, oils, solvents), the rate of loss of antioxidants from the polymer surface depends both on their diffusion coefficient and their partition coefficient between the liquid and the polymer. This is complicated by the varying ability of different solvents (liquid media) to swell polymers, giving rise to an increase in the diffusion coefficient of the antioxidant, which leads to higher loss rates.[47] As in the case of volatilization, the rate of leachability of antioxidants from the surface of polymers into liquid contact media increases with temperature and surface area-to-volume ratio, e.g., in thin polymer samples.

ANTIOXIDANTS: THERMAL STABILIZATION

Stabilization of polymers against the effect of temperature, shear, and other oxidizing agents is required, both to prevent mechano-thermal oxidation at the processing stage (processing antioxidants) and to afford protection during in-service (long-term thermal stabilizers). In addition, the thermal stability of reprocessed polymers (recycled) must be preserved and reinforced by additional stabilization (recycling stabilizers) for further reprocessing and end-use. Different classes of antioxidants are used, and their stabilizing action in polymer melts is outlined below.

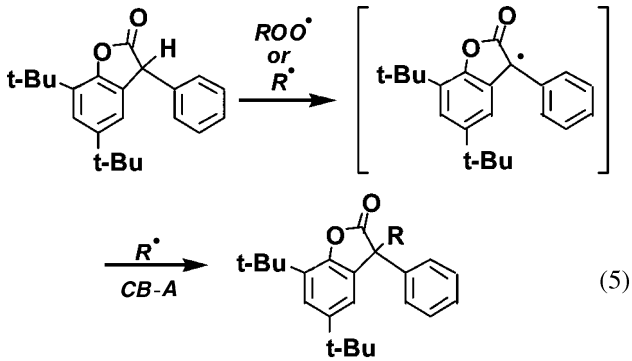
Aromatic amine antioxidants based on *p*-phenylenediamine, e.g., AOs 11 and 12, Table 1, are highly effective peroxyl radical scavengers (more effective than hindered phenols). Their use, however, is limited to elastomers because of their extreme staining power. Sterically hindered phenols (also efficient CB-D antioxidants), on the other hand, have been used extensively as melt processing stabilizers in plastics, as they do not suffer as much from the problem of polymer discoloration during melt processing, but yellowing can still occur because of the formation of colored conjugated quinonoid-type transformation products with visible absorption wavelengths, extending to more than 400 nm.[48] Hindered phenols containing propionate esters, e.g., AOs 4 and 5 are good stabilizers for polyolefins and styrenics, the propionamide AO 6 is used in aliphatic polyamides, and the ester AO 7 is suitable for polyesters. Scheme 4 shows some key reactions of propionate-based hindered phenols used for polymer stabilization, in order to illustrate their antioxidant action and the formation of discoloring products. The primary step in their stabilizing action involves the scavenging of alkylperoxyl ($ROO^{\bullet}$) radicals; a step, which also leads to the

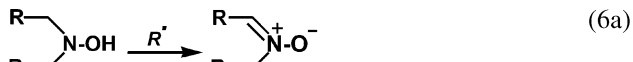


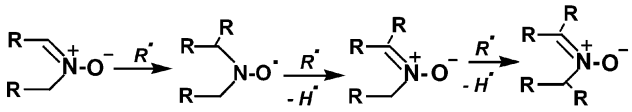
Scheme 4 Key reactions that take place during the antioxidant action of 4-propionate-hindered phenols.

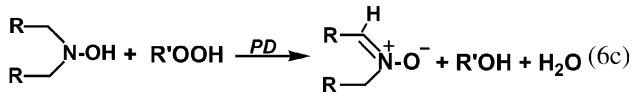
production of harmful hydroperoxides, see Scheme 4a. In practice, therefore, hindered phenols are generally used in combination with peroxide decomposers, typically a hydrolysis-resistant phosphite.[34,41,49] It is clear from Scheme 4 that the major transformations of phenolic antioxidants during melt processing consist of quinone-type products, C–C and C–O coupling product. Transformation products of antioxidants can further influence their overall stabilization effect by exerting either anti- or pro- oxidant effects, thereby synergizing or antagonizing the action of the parent antioxidant.

Antioxidants that act by CB-A mechanism are particularly suitable for melt stabilization. The bifunctional hindered phenol semi-acrylate antioxidant, AO 8 in Table 1, which acts primarily by deactivating alkyl radicals, is an effective processing stabilizer for styrene-butadiene-styrene copolymers.[50] Lactone stabilizers, see AO 13, Table 1, which are used commercially in phenol-free blends (e.g., with phosphites) or in blends containing combinations of hindered phenols and phosphites, have been shown[51,52] to be highly efficient in oxygen deficient-environment particularly during melt processing. They act by trapping macroalkyl radicals, as well as peroxyl radicals, through the intermediacy of the resonance stabilized lactone (benzfuranyl) radical, Reaction 5. Similarly, the hydroxylamines, see AO 14, Table 1, which can also scavenge the macroalkyl radicals via the intermediacy of nitrones, as well as reducing hydroperoxides to alcohols, have been shown[53] to be highly efficient processing stabilizers, Reaction 6.





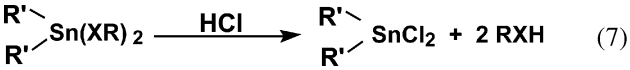




Alkyl and aryl phosphite esters are also effective melt stabilizers. They are often used in combination with hindered phenols to give highly efficient melt stabilizing systems and to reduce discoloration of the polymer because of the oxidation products of the phenols present. Phosphites (particularly those derived from aliphatic alcohols and unhindered phenols) are, however, generally susceptible to hydrolysis. Consequently, moisture-sensitive phosphites affect adversely the handling characteristics (i.e., flow properties) of the additive package and are a source of other problems: corrosion of metal surfaces, formation of dark colored spots, and gel formation. In practice, hydrolysis-resistant phosphites based on sterically hindered phenols are used, e.g., AOs 17 and 18, Table 1.

Sulfur-containing compounds are highly effective PD antioxidants, but they are more complex in their behavior than the phosphites. They react with hydroperoxides in a catalytic process, which leads to the formation of sulfur acids that are responsible for the catalytic nonradical destruction of hydroperoxides. The ''simple'' dialkyl sulfides, e.g., thiodipropionate esters, such as Irganox PS800 and PS 802 (AOs 19 and 20, Table 1), function as weak melt stabilizers as they give rise to a pro-oxidant effect in the polymer because of the generation of free radicals by an initial redox reaction with hydroperoxides, see Scheme 5.[54] In view of the pro-degradant effects of thiodipropionate esters, these sulfide antioxidants are always used in combination with CB-D antioxidants, such as Irganox 1010 (AO 5) in order to deactivate the free radicals formed.

Metal carboxylates (e.g., lead carbonate, calcium, barium, and zinc soaps) and tetravalent derivatives of tin (e.g., dibutyl tin maleate, DBTM, AO 35, Table 1) are frequently used as antioxidants for PVC stabilization. Melt stabilization of PVC is normally aimed at reducing the formation of HCl and hydroperoxides in the polymer, and at removing the developing unsaturation, which is both the source of color and further instability. In general, plasticized PVC is processed at lower temperatures than rigid PVC, and lesser damage is expected during melt processing of the former. The elimination of labile chlorine atoms from the polymer backbone is the most important stabilization mechanism for PVC. Dialkyl tin maleates (e.g., DBTM) and thioglycollates (e.g., DOTG, AO 36, Table 1) function by eliminating HCl, see Reaction 7. In addition, the maleates also act by removing the unsaturation and limiting color development, whereas the thioglycollates have an additional peroxidolytic function.[34]



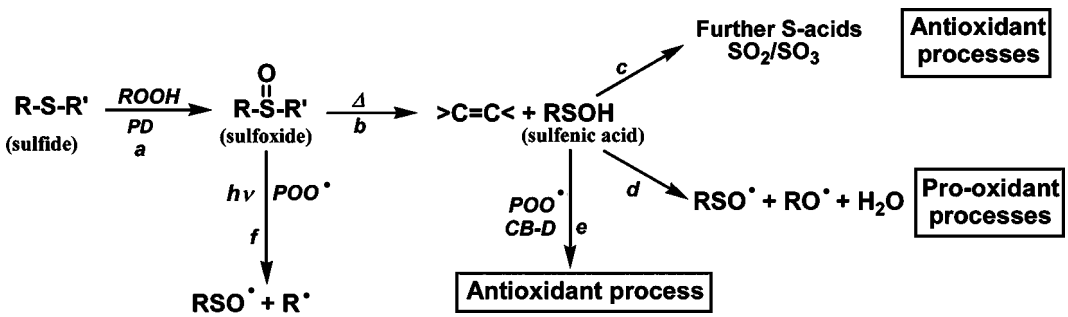
(7)

e.g. DOTG. where:
XR = SCH_2CO_2R and R′ = Oct

ANTIOXIDANTS: PHOTOSTABILIZATION

As with thermal stabilizers, photostabilizers must satisfy basic chemical and physical requirements (see the section titled ''*Antioxidant permanence: effects of chemical and physical factors*''). In addition, they must be photo-stable, i.e., stable to UV-light, to withstand continuous periods of UV-exposure, without being prematurely destroyed or effectively transformed into sensitizing products. There are essentially three classes of compounds that are categorized as photostabilizers/photoantioxidants: UV-absorbers and pigments, peroxide decomposers including nickel complexes, and sterically hindered amine light stabilizers.

Most of the UV-absorbers (UVA) used commercially fall into two main classes of compounds, the 2-hydroxybenzophenones, e.g., AO 28, Table 1, and the 2-hydroxybenzotriazoles, e.g., AOs 29–32, Table 1, with the 2-hydroxy group being essential for their activity.[55] These UVAs operate by additional mechanisms too, for example, by removing initiating radicals (e.g., alkoxyl radicals) in a weak chain breaking-donor (CB-D) mechanism.[56] UV absorbers, such as AO 28, also synergize effectively with peroxide decomposers, e.g., metal dithiolates, see Table 4.



Scheme 5 Peroxidolytic and chain breaking activity of phosphorus antioxidants.

Metal complexes of dithioic acids, e.g., dithiocarbamates, dithiophosphates, xanthates (MDRC, MDRP, MRX, respectively) see AOs 21–24, Table 1, which are highly effective catalytic peroxide decomposers (PD-C) and excellent melt stabilizers, are generally effective photo-antioxidants.[13,25,26] Their effectiveness is attributed mainly to the much higher UV stability of dithiolates compared to simple sulfides. Both the nature of the metal ion in the dithiolate complexes and their concentration play a crucial part in their overall effectiveness as UV stabilizers. Transition metal complexes containing Ni, Co, and Cu are more photostable than Zn- and Fe-containing complexes, hence are more effective, see Table 2. Furthermore, the photo-antioxidant activity of nickel dithiolates is greatly affected by their solubility in the polymer.[46] These peroxide decomposers also synergize effectively when used in combination with UV-absorbers, Table 2.

Hindered piperidines (known as HALS, hindered amine stabilizers), see AOs 25 and 26, Table 1, have been used extensively as photo-stabilizers in commercial polymers. HALS is a unique class of photostabilizers that do not function as UV screens singlet oxygen, triplet carbonyl quenchers, or as peroxide decomposers. Their effectiveness is attributed to their primary transformation product, the corresponding nitroxyl radicals (>NO$^{\bullet}$), which is capable of scavenging alkyl radicals in competition with oxygen, i.e., an effective CB-A antioxidant. Its photostabilizing mechanism also involves the regeneration of the nitroxyl radical from both the corresponding hydroxylamine (>NOH an effective CB-D) and alkylhydroxylamine,[22,40,57,58] see Scheme 3. Overall, the photoantioxidant activity of HALS can be ascribed to a regenerative donor-acceptor (CB-A/CB-D) antioxidant mechanism involving >NO$^{\bullet}$ and >NOH.

Table 2 Effect of different metal dithiolates on the photostability (embrittlement time, EMT) of PP processed at 190°C (For structures, see Table 1)

Antioxidant	Concentration (mol/100 g × 10^4)	UV-EMT (hr)
Unstabilized PP	0	90
ZnDEC	2.5	175
NiDEC	2.5	740
NiDEC	10	840
FeDMC	0.25	85
FeDMC	2.5	150
FeDMC	5	336
CoOct X	3	1600
Chimassorb 81	3	245
NiDBP + Chimassorb 81	6	2650

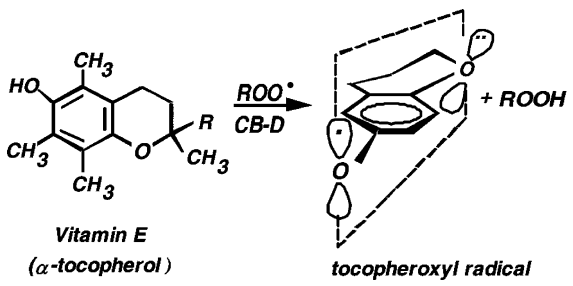
ANTIOXIDANTS: METAL DEACTIVATOR

The main function of metal deactivators (MD) is to retard efficiently metal-catalyzed oxidation of polymers. Polymer contact with metals occur widely, for example, when certain fillers, reinforcements, and pigments are added to polymers, and, more importantly when polymers, such as polyolefins and PVC, are used as insulation materials for copper wires and power cables (copper is a pro-oxidant since it accelerates the decomposition of hydroperoxides to free radicals, which initiate polymer oxidation). The deactivators are normally polyfunctional chelating compounds with ligands containing atoms like N, O, S, and P (e.g., see Table 1, AOs 33 and 34) that can chelate with metals and decrease their catalytic activity. Depending on their chemical structures, many metal deactivators also function by other antioxidant mechanisms, e.g., AO 33 contains the hindered phenol moiety and would also function as CB-D antioxidants.

ANTIOXIDANTS: BIOLOGICAL

The use of antioxidants in human-contact applications, e.g., food-contact, medical, and pharmaceutical, present a challenge in terms of their safety and level of migration into the contact media, e.g., food and body fluids. The biological antioxidant vitamin E, which is a suitable candidate for such areas of application, is a fat-soluble, and sterically hindered phenol antioxidant with the most bioactive form of the vitamin being the α-tocopherol (Table 1, AO 10).

In-vitro rate studies on the antioxidant activity of α-tocopherol has shown that it is one of the most efficient alkylperoxyl radical traps, far better than the commercial hindered phenols, e.g., BHT (2,6-di-*tert*-butyl-4-methylphenol, AO 1).[59] Its efficiency was attributed to stereo electronic effects: the electronic synergy between a fully methylated aromatic ring and the chroman moiety results in a highly stabilized tocopheroxyl radical, formed during the rate limiting step, Reaction 8, because of the interaction between the p-orbitals on the two oxygen atoms.[59]



(8)

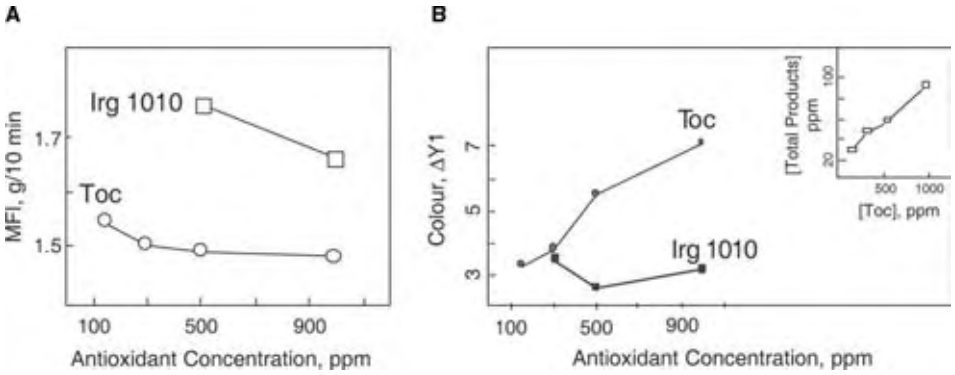


Fig. 2 Melt and color stability of PP processed in the presence of antioxidants. *(View this art in color at www.dekker.com.)*

α-tocopherol has been shown to be a very effective melt processing antioxidant for PP, especially at low concentration.[49,60,61] A comparison of the antioxidant efficiency of α-tocopherol with that of Irganox 1010 (AO 5) during melt extrusion of PP at 260°C shows (see Fig. 2A) clearly the superior performance of the former at all concentrations and, in particular, at very low concentrations. Tocopherol can, therefore, be used cost-effectively at only one-quarter of the concentration typically required for the stabilization of PP by synthetic hindered phenols, such as Irganox 1010. The activity of the former is attributed to the rapid rate of deactivation of radicals responsible for PP chain scission. A further important contribution to its antioxidant activity stems from its oxidation products, which are formed during polymer melt processing; these products were shown to be very effective antioxidants.[61] Tocopherol products formed during PP and PE melt processing consist mainly of direct coupling products, leading to the formation of dihydroxydimer, DHD and quinonoid-type products, trimers, TRI, spirodimers, SPD, quinone methides, QM, together with some aldehydes, ALD,[62] see Fig. 3 for structures of these products. All the oxidation products were shown to be more highly colored than tocopherol itself, with the aldehydes being the most colored, and the trimers the least colored.

In general, sterically hindered phenols contribute to some discoloration (yellowing) of polyolefins during processing. Yellowing of polyolefins containing hindered phenols has been attributed to a number of factors including the formation of colored oxidation products, e.g., quinonoid structures, and interactions between the phenols and transition metal ion catalyst residues from the polymerization stage.[60,63] The extent of discoloration depends on the chemical structure of the parent antioxidant, the oxidation products, and the type and amount of catalyst residues in the polymer. It was shown[49] that at low concentration, both tocopherol and Irganox 1010 cause comparable

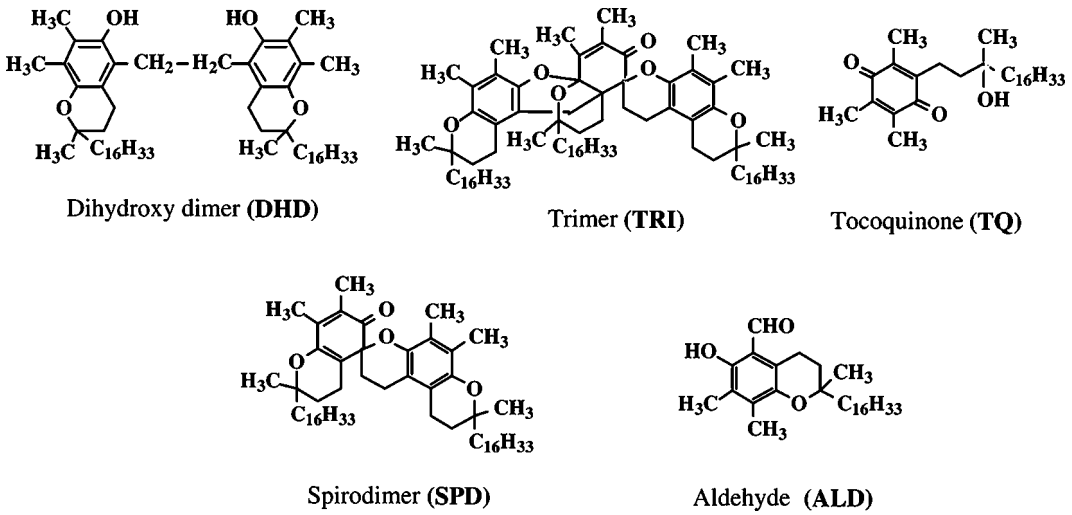


Fig. 3 Structures of the main transformation products of α-tocopherol formed during melt processing of polyolefins.

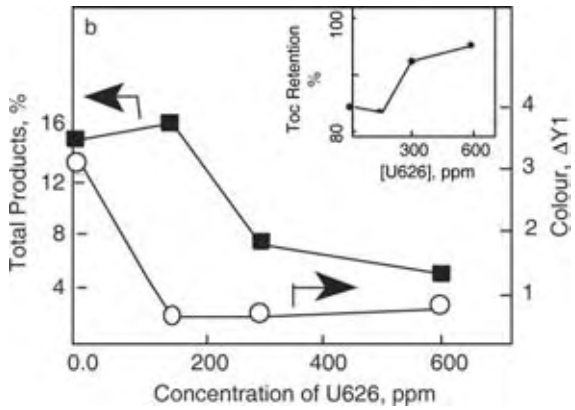


Fig. 4 Color stability of PP processed in presence of 300 ppm tocopherol in combination with the phosphate U-622. Total amount of products formed from tocopherol is also shown.

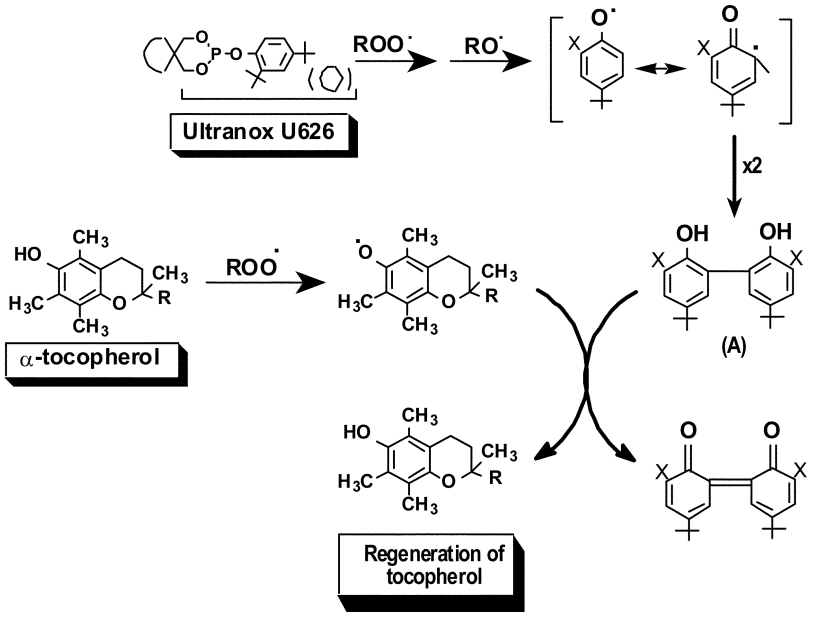
levels of discoloration during melt processing of PP. At increasingly higher concentrations, however, the extent of discoloration affected by tocopherol is higher (Fig. 2B). In order to reduce the extent of discoloration, very small concentrations, e.g., 300 ppm, of a phosphate antioxidant can give rise to a pronounced color suppression, together with higher levels of retention of the tocopherol in the polymer (Fig. 4). The higher retention of the tocopherol antioxidant observed when a small amount of the phosphite U-626 was used has been attributed to[61] first, a reduction in the amount of the more intensely colored transformation products, and second to the regeneration of tocopherol in the presence of the phosphite, via a redox-type reaction, through the formation of a phosphite-phenol C–C coupled product (Scheme 6).

ANTIOXIDANTS: REACTIVE

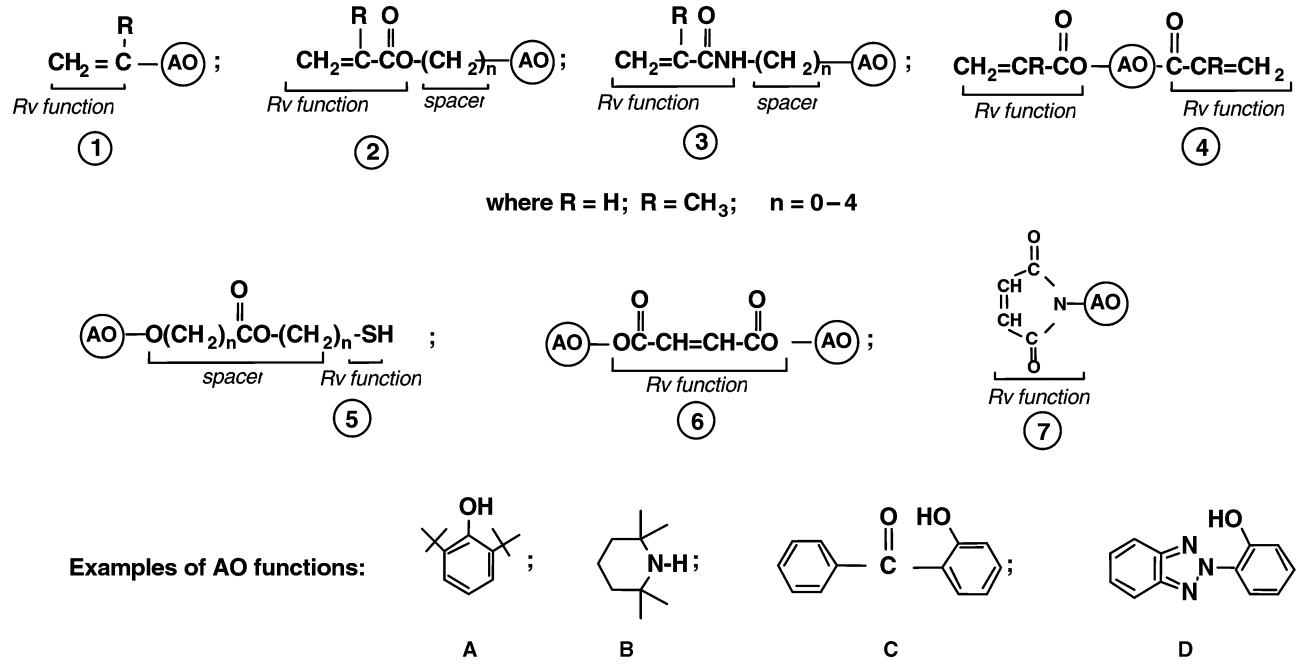
A strategy that is based on the use of reactive antioxidants can also be explored to achieve stabilization of polymers suitable for human-contact applications. Reactive antioxidants that become an integral part of the macromolecular chain can result in nonmigratory stabilizer systems that would be unaffected by extractive hostile contact media.

In general, reactive antioxidants are compounds that contain one or more antioxidant functions (the antioxidant, AO component) and one or more chemical functions that are capable of reacting either with monomers (same or different) or with polymers (the reactive, Rv, component). The AO function is based on any antioxidant moiety (see examples A–D in Scheme 7), whereas the reactive moiety can either be a polymerizable (e.g., Rv functions 1–4) or nonpolymerizable (e.g., Rv functions 5–7) groups, and may or may not contain a spacer (an inert flexible and short chemical link connecting the antioxidant moiety to the reactive function).

Reactive antioxidants may either be copolymerized with monomers during polymer synthesis or grafted on preformed polymers; they are therefore linked to the polymer. Although the copolymerization route has been successfully exploited,[35,64,65] it has not received greater attention because of cost incurred in the synthesis and production of tailor-made "speciality" materials for low-volume specific applications. On the



Scheme 6 Redox reactions resulting in the regeneration of tocopherol.



Scheme 7

other hand, grafting of antioxidants on preformed polymer melts can offer a more flexible and versatile approach where standard compounding and processing machines are used to conduct chemical grafting reactions. Both routes, however, offer tremendous advantages in terms of the physical persistence of antioxidants in the polymer. In both cases, the process of chemical attachment (target reaction) of antioxidants onto the polymer backbone proceeds in competition with other undesirable processes. The main prerequisite here, therefore, is to achieve the target reaction without detriment to the overall polymer properties and the fabrication process.

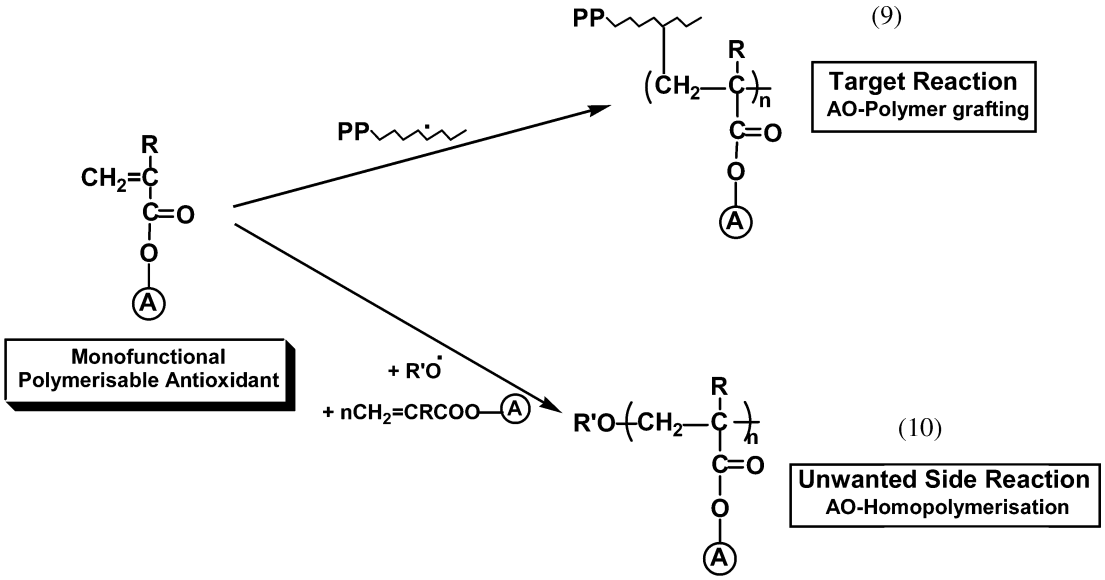
Both thermal- and photo-antioxidant functions have been grafted on polyolefins during melt processing in the presence of a free radical initiators.[66–68] The practical success of in situ melt grafting of antioxidants on polyolefins, however, depends on the correct choice of chemical systems and processing variables that would reduce the interference of side reactions, without altering significantly the polymer characteristics, e.g., molar mass, morphology, and physical properties.

The three different types of reactive antioxidant systems typically used for grafting reactions on polyolefins are briefly as described in the following.

Monofunctional Polymerizable Antioxidants

The use of polymerizable monofunctional antioxidants with one reactive group per antioxidant molecule is considered here. Production of these antioxidants is generally straightforward. Therefore, it can offer, a broad, versatile, and economic route for the production of a range of polymer-grafted antioxidants and antioxidant concentrates. Different monofunctional antioxidants have been reactively grafted on polyolefins, e.g., PP, LDPE, HDPE, poly (4-methyl-1-ene), in the presence of free radical initiators using single- or twin screw extruders, or internal mixers.

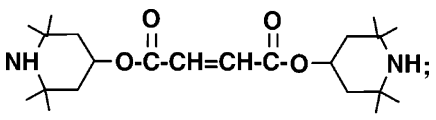
It has been demonstrated, however, that the efficiency of chemical attachment of such monofunctional polymerizable antioxidants on polyolefins (Reaction 9) is always low.[66,69] This is mainly because of the highly competitive homopolymerization reaction of the reactive antioxidant (Reaction 10). For example, studies on the effect of processing variables on the extent of melt grafting on PP of different mono-acryloyl-containing hindered phenol (DBBA) and hindered amine (AOTP) antioxidant functions have shown that grafting efficiency is less than 50%.[69,70] The remaining ungrafted antioxidants were recovered, almost completely, as homopolymers of the parent antioxidants, which were incompatible with the host polymer and were readily removed by extractive solvents. Furthermore, the performance of these homopolymers, when incorporated in the polymer matrix as conventional antioxidants, is very poor. The problems of homopolymer formation and low efficiency of grafting of mono-functional polymerizable antioxidants in polyolefins were subsequently



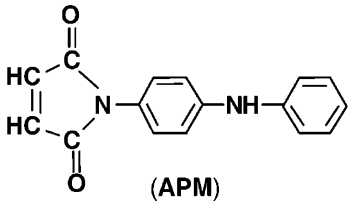
addressed by alternative approaches (see Reactions 9 and 10, above).

Monofunctional Nonpolymerizable Antioxidants

Non-polymerizable monofunctional antioxidants were subsequently used to avoid the problem of homopolymerization of the antioxidant. For example, melt grafting of the two maleated antioxidants, BPM and APM (see below), on PP was shown to lead to high grafting efficiencies (up to 75% in the former and >90% in the latter), which were attributed to the nonpolymerizable nature of the maleate (maleimide) functions.[71] The performance of these antioxidants, especially under extractive organic solvent conditions, was also shown to far exceed that of conventional antioxidants with similar antioxidant function.



(BPM)



(APM)

Bifunctional Polymerizable Antioxidants

The use of reactive antioxidants containing two polymerizable polymer-reactive functions in the same antioxidant molecule is outlined here. Careful choice of the processing parameters, the type, and the amount of free radical initiator can lead to very high levels of antioxidant grafting.[71] For example, melt grafting of concentrates (e.g., 5–20 wt.%) of the di-acrylate hindered piperidine, AATP, on PP in the presence of a peroxide initiator has led to almost 100% grafting. This exceptional grafting efficiency of AATP is in marked contrast with the much lower grafting levels achieved with the mono-functional HALS analogues, e.g., MyATP and AmyTP.[70] Examination of the mechanisms involved in the grafting process of such bifunctional antioxidants has shown that the grafting reaction occurs through the intermediacy of a cross-linked structure, involving the polymer and the reactive antioxidant, lead finally to an antioxidant grafted polymer product, which remains comparable in its general characteristics, e.g., solubility, crystallinity, molar mass, to a conventionally stabilized sample.[70,71]

Monofunctional Polymerizable Antioxidants in the Presence of a Comonomer

The use of a reactive di- or poly-functional comonomer (nonantioxidant), which can co-graft with a monofunctional polymerizable antioxidant on polymers, can improve the grafting efficiency from as low as 10–40% to an excess of 80–90%. This strategy, however,

presents immense challenges because of the presence of more than one polymerizable group in the comonomer, which could lead to additional undesirable (competing) side reactions, complicated by the possibility of comonomer-induced crosslinking reactions of the polymer. The success of this ''one-pot'' synthetic approach lies in the ability to achieve a delicate balance between the composition of the chemical system (antioxidant, comonomer, free radical initiator) and reaction conditions (e.g., temperature, residence time) with the aim of promoting the target grafting reaction at the expense of all competing side reactions.[69,70]

In practice, the success of this method has been clearly illustrated.[69,70] The novelty of this approach lies in the fact that co-grafting of polymerizable polyfunctional agents (traditionally used as crosslinkers, e.g., the trimethylol propane triacrylate, Tris) with mono-vinyl antioxidants (and other additives) in extruders or mixers leads to the production of highly grafted antioxidants in a noncrosslinked polymer. This co-grafting method can be applied to a wide range of antioxidant functions (e.g., HAS, UVA, hindered phenols, aromatic amines) to achieve outstanding levels of antioxidant grafting. Table 3 shows an example, which illustrates the excellent performance, especially under extractive conditions, of a highly bound synergistic antioxidant system (hindered phenols + UV absorber) produced by this method in PP compared to a conventional (unbound) commercial antioxidant system.

ANTIOXIDANT MIXTURES: SYNERGISM AND ANTAGONISM

The interaction between two or more antioxidants (or antioxidant functions) in plastics formulations can lead to enhanced performance by more than the sum of their individual effects; i.e., synergistic effects. Synergism can result from the combined action of two chemically similar antioxidants (homosynergism), e.g., two hindered phenols, or from two different antioxidant functions present in the same stabilizer molecule (autosynergism), e.g., Irganox 1081 (AO 9, Table 1), or when separate stabilizer molecules that carry different antioxidant functions are physically blended in a stabilizer formulation (heterosynergism). Conversely, antioxidant combinations resulting in reduced performance, relative to the sum of their individual contribution, are called antagonistic.

Highly effective UV stabilizing systems can be achieved by the use of synergistic mixtures of compounds acting by different mechanisms. Table 4 illustrates the synergism obtained from combinations of different metal dithiolates with the UV absorber Chimassorb 81 (AO 28, Table 1) in LDPE. Hindered phenol antioxidants combined with sulfur-containing compounds exhibit synergism during thermal stabilization of polyolefins. In contrast, similar combinations of Irganox 1076 (AO 4, Table 1) with different metal dithiolates lead to antagonism during photostabilization, see Table 4.[26,72] This antagonistic behavior has been attributed to sensitization leading to photolytic destruction of the dithiolates by oxidation products of phenols, particularly stilbenequinones. Stabilizers that contain two different antioxidant functions (e.g., PD and CB activities) in the same molecule, such as phenolic sulfides, show much higher molar intrinsic activity as thermal antioxidants (because of autosynergism) than conventional hindered phenols with only CB activity.[11]

HALS exhibit a complex behavior when present in combination with other antioxidants and stabilizers. Effective synergism in both melt and thermal stabilization has been achieved when secondary and tertiary HALS were used in combination with both aromatic and aliphatic phosphites; the synergistic optimum depends on the structure of the phosphite.[73] HALS also synergize the action of UV-absorbers, e.g., benzotriazoles, in different polymers such as polypropylene, polystyrene, and ABS.[74]

Table 3 Comparison of the antioxidant performance (accelerated UV aging) of synergistic mixture (melt grafted in presence of Tris) with a conventional antioxidant mixture based on the same antioxidant functions (at 1 : 1 w/w ratio)

Antioxidant (0.4% in PP films)	UV embrittlement time (hr)	
	Unextracted	Extracted
None	75	70
DBBA[a]	205	80
HAEB[b]	330	70
PP-g(DBBA-HAEB)$_{Tris}$	1160	1130

UV531 is a commercial UV absorber, 2-hydroxy-4octyloxy benzophenone Tris is a triacrylate comonomer.
[a]Unbound, processed alone (no Tris) as a conventional antioxidant.
[b]See Table 2 for structures.

CONCLUSIONS

Antioxidants and stabilizers are generally used in the chemical processing industry for the protection and preservation of properties of materials, including food, with the aim of prolonging and extending their shelf life. They are essential ingredients for the long-term durability of many polymers, such as

Table 4 Synertistic and antagonistic effects on photostability of LDPE processed at 150°C

Stabilizer system	Concentration ($\times 10^{-4}$ M 100 g^{-1})	Photo-embrittlement time (hr)		References
		Observed	Calculated effect[a,b]	
Control PP (no antioxidant)	—	1000	—	[26]
ZnDEC	3	1400	—	[26]
NiDEC	3	1800	—	[26]
Tinuvin 770	3	2400	—	[26]
Irganox 1076	3	1750	—	[26]
NiDBP	2.5	2800	—	[72]
NiBX	2.5	2500	—	[72]
CuDIP	2.5	2300	—	[72]
Chimassorb 81	3	1650	—	
Synergistic systems				
NiDEC + Chimassorb 81	3 + 3	—	—	[26]
ZnDEC + Chimassorb 81	3 + 3	4000	3000	[26]
NiDBP + Chimassorb 81	2.5 + 2.5	5500	4500	[72]
NiBX + Chimassorb 81	2.5 + 2.5	4900	4200	[72]
CuDIP + Chimassorb 81	2.5 + 2.5	5350	4000	[72]
Antagonistic systems				
NiDEC + Irganox 1076	3 + 3	1580	3550	[26]
ZnDEC + Irganox 1076	3 + 3	1250	3150	[26]
NiDEC + Tinuvin 770	3 + 3	1850	4200	[26]
NiDBP + Irganox 1076	2.5 + 5	—	—	[72]

[a]Concentration of HOBP in this case was 5×10^{-4}M 100 g^{-1}.
[b]% Synergism = $\{[(E_s - E_c) - (E_1 - E_c) + (E_2 - E_c)]/(E_1 - E_c + E_2 - E_c)\} \times 100$, where E_s is the embrittlement time of synergist, E_c the embrittlement time of control, E_1 the embrittlement time of antioxidant 1, and E_2 the embrittlement time of antioxidant 2.

polyolefin, and are crucial to the upgrading of their performance, and for achieving the benefits of sustainable development in polymers recycling programs.

Antioxidants and stabilizers are chosen for target applications on the basis of chemical, physical, toxicological, and economic factors. The final selection of an antioxidant package must take into consideration the performance requirements of the end-use polymer article including toxicity, compatibility, appearance, and color. Issues of efficacy and safety have been the driving force behind much of the recent progress made in the areas of biological, reactive, and macromolecular antioxidants.

More stringent regulations and legislations for certain applications of stabilized polymers, such as in food, toys, medicine, and other health-related areas, would promote further interest in the use of biological and naturally occurring antioxidants and reactive antioxidants for chemical processing, and for producing safe and ''permanently'' stabilized polymer compositions. Current emphasis on sustainable development and green chemistry approaches should lead to further exploration of the benefits of renewable resources and environmentally benign synthetic routes in the development and procurement of new antioxidants, or for replacing existing ones.

Compared to conventional antioxidants, reactive antioxidants that are capable of becoming covalently bound to the polymer backbone are not readily lost from polymers during fabrication and in-service. There is a lot of evidence that demonstrates the performance (in terms of polymer protection) of ''immobilized'' antioxidants in practice, especially when polymer products are subjected to harsh environment, e.g., exposure to high temperatures, UV-light and leaching solvents. It is clear from this that high mobility of low molar mass antioxidants is not a necessary prerequisite to achieving stabilization and attachment of antioxidants to polymers can be industrially beneficial.

Reactive antioxidants grafted on polymer melts behave in a similar way to low molar mass conventional antioxidants, but offer many additional advantages. The polymer-linked antioxidants do not suffer from the problem of compatibility, volatility, and migration, i.e., they do not suffer physical loss even under highly aggressive and extractive environments. Such antioxidant systems would be much more risk-free and environmentally friendly. The ability to produce highly grafted antioxidant concentrates (master batches), which can be used in conventional (the same or different) polymers, as ''normal'' additives would extend the use of reactive antioxidants to new areas of application.

REFERENCES

1. Bolland, J.L.; Gee, G. Kinetic studies in the chemistry of rubber and related materials. II. The kinetics of oxidation of unconjugated olefins. Trans. Faraday Soc. **1946**, *42*, 236–243.
2. Bolland, J.L. Kinetic studies in the chemistry of rubber and related materials. VI. Benzyl peroxide-catalyzed oxidation of ethyl linoleate. Trans. Faraday Soc. **1948**, *44*, 669–677.
3. Bolland, J.L. Kinetics of olefin oxidation. Quart. Rev. **1949**, *3*, 1–21.
4. Bateman, L.; Morris, A.L. Initiation efficiencies in olefin autoxidation. Trans. Faraday Soc. **1952**, *48*, 1149–1155.
5. Al-Malaika, S. Autoxidation. In *Atmospheric Oxidation and Antioxidants*; Scott, G., Ed.; Elsevier Applied Science Publishers: London, 1993; 76, Vol. 1, 45–82.
6. Scott, G. Initiators, proxidants and sensitizers. In *Atmospheric Oxidation and Antioxidants*; Scott, G., Ed.; Elsevier Applied Science Publishers: London, 1993; 76, Vol. 1, 83–119.
7. Howard, J.A Absolute rate constants for reactions of oxyl radicals. Adv. Free Radical Chem. **1972**, *4*, 49–173.
8. Iring, M.; Tudos, F. Thermal oxidation of polyethylenes and polypropylene: effects of chemical structure and reaction conditions on the oxidation process. Prog. Polym. Sci. **1990**, *15* (2), 217–262.
9. Hinsken, H.; Moss, S.; Pauquet, J.-R.; Zweifel, H. Degradation of polyolefins during melt processing. Polym. Deg. Stab. **1991**, *34*, 279–293.
10. Johnston, R.T.; Morrison, E.J. Thermal scission and cross-linking during polyethylene melt processing. In *Polymer Durability: Degradation, Stabilisation and Lifetime Prediction*; Clough, R.L., Billingham, N.C., Gillens, K.T., Eds.; Advances in Chemistry Series-249, A.C.S.: Washington, 1996; 651–682.
11. Al-Malaika, S.; Scott, G. Thermal stabilizers of polyolefins. In *Degradation and Stabilisation of Polyolefins*; Allen, N.S., Ed.; Applied Science: London, 1983; 247–281.
12. Scott, G. Antioxidants: chain breaking mechanisms. In *Atmospheric Oxidation and Antioxidants*; Scott, G., Ed.; Elsevier Applied Science: London, 1993; Vol. 1, 121–160.
13. Al-Malaika, S. Antioxidants: preventive mechanisms. In *Atmospheric Oxidation and Antioxidants*; Scott, G., Ed.; Elsevier Applied Science: London, 1993; Vol. 1, 161–224.
14. Pospisil, J. Chain-breaking antioxidants in polymer stabilization. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1979; Vol. 1, 1–37.
15. Pospisil, J. Aromatic amines antidegradants. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1984; Vol. 7, 1–63.
16. Chakraborty, K.B.; Scott, G. Mechanisms of antioxidant action: the behavior of hindered piperidine UV stabilizers during the processing of LDPE. Chem. Ind. **1978**, *7*, 237.
17. Carlsson, D.J.; Garton, A.; Wiles, D.M. The photo-stabilization of polyolefins. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1979; Vol. 1, 219–259.
18. Denisov, E.T. Inhibitor regeneration in oxidation. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1980; Vol. 3, 1–20.
19. Bagheri, R.; Chakraborty, K.B.; Scott, G. Mechanisms of antioxidant action: behavior of a hindered piperidine and related oxidation products during processing and photo-oxidation of polypropylene. Polym. Deg. Stab. **1982**, *4* (1), 1–16.
20. Shlyapintokh, V.Ya.; Ivanov, V.B. Antioxidant action of sterically hindered amines and related compounds. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1982; Vol. 5, 41–70.
21. Berger, H.; Bolsman, T.A.B.M.; Brouwer, D.M. Catalytic inhibition of hydrocarbon autoxidation by secondary amines and nitroxyls. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1983; Vol. 6, 1–27.
22. Al-Malaika, S.; Omikorede, E.O.; Scott, G. Mechanisms of antioxidant action: mechanochemical transformation products of 2,2,6,6-tetramethyl-4-hydroxypiperidinoxyl in polypropylene. J. Appl. Polym. Sci. **1987**, *33*, 703–713.
23. Schwetlick, K. Mechanisms of antioxidant action of phosphite and phosphonite esters. In *Mechanisms of Polymer Degradation and Stabilisation*; Scott, G., Ed.; Elsevier Science: New York, 1990; 23–60.
24. Pobedimskii, D.G.; Mukmeneva, N.A.; Kirpichnikov, P.A. *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1980; Vol. 2, 125–184.
25. Al-Malaika, S. Antioxidant mechanisms of derivatives of dithiophosphoric acid. In *Mechanisms of Polymer Degradation and Stabilisation*; Scott, G., Ed.; Elsevier Science: New York, 1990; 61–107.
26. Al-Malaika, S.; Chakraborty, K.B.; Scott, G. Peroxidolytic antioxidants: metal complexes containing sulfur ligands. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1983; Vol. 6, 73–120.
27. Scott, G. Peroxidolytic antioxidants: sulphur antioxidants and autosynergistic stabilizers based on

alkyl and aryl sulphides. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1983; Vol. 6, 29–71.
28. Shelton, J.R. Organic sulphur compounds as preventive antioxidants. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1981; Vol. 4, 23–69.
29. Osawa, Z. Inhibition of metal-catalyzed degradation of polymers. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1984; Vol. 7, 193–232.
30. Chan, M.G. Metal deactivators. In *Oxidation Inhibition of Organic Materials*; Klemchuk, P., Pospisil, J., Eds.; CRC Press: Boca Raton, 1990; Vol. 1, 225–246.
31. Muller, H. Metal deactivators. In *Plastics Additives Handbook*, 2nd Ed.; Gachter, R., Muller, H., Eds.; Hanser: Munich, 1987; 75–95.
32. Gugumus, G. Light stabilizers. In *Plastics Additives Handbook*, 5th Ed.; Zweifel, H., Ed.; Hanser: Munich, 2001; 141–425.
33. Scott, G. Substantive antioxidants. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1981; Vol. 4, 181–221.
34. Scott, G. Macromolecular and polymer-bound antioxidants. In *Atmospheric Oxidation and Antioxidants*; Scott, G., Ed.; Elsevier Applied Science: London, 1993; Vol. 2, 279–326.
35. Fu, S.; Gupta, A.; Albertsson, A.C.; Vogl, O. New polymerizable 2(2-hydroxyphenyl)2H-benzotriazole ultraviolet absorbers: 2[2,4-dihydroxy-5-vinyl (isopropenyl)phenyl]2,3–2H-dibenzotriazole. In *New Trends in the Photochemistry of Polymers*; Allen, N.S., Rabek, J.F., Eds.; Elsevier Applied Science: London, 1985; 247–264.
36. Pospisil, J. Stabilizers mixtures and polyfunctional stabilizers. In *Oxidation Inhibition of Organic Materials*; Klemchuk, P., Pospisil, J., Eds.; CRC Press: Boca Raton, 1990; Vol. 1, 173–224.
37. Al-Malaika, S. Reactive modifiers for polymers. In *Chemical Reactions on Polymers*; Benham, J.L., Kinstle, J.F., Eds.; ACS Symposium Series-364, ACS: Washington, 1988; 409–425.
38. Al-Malaika, S. Reactive processing and polymer performance. Polym. Plast. Technol. Eng. **1990**, *29* (12), 73–86.
39. Al-Malaika, S. Tying additives down. Chemtech **1990**, *6*, 366–371.
40. Glass, R.D.; Valange, B.M. Antioxidant 'crossover effect' in oven ageing of polypropylene. Polym Deg. Stab. **1988**, *20* (3,4), 355–363.
41. Gugumus, F. Stabilization of plastics against thermal oxidation. In *Oxidation Inhibition of Organic Materials*; Klemchuk, P., Pospisil, J., Eds.; CRC Press: Boca Raton, 1990; Vol. 1, 61–172.
42. Zweifel, H. Effect of stabilization of polypropylene during processing and its influence on long-term behavior under thermal stress. In *Polymer Durability: Degradation, Stabilisation and Lifetime Prediction*; Clough, R.L., Billingham, N.C., Gillens, K.T., Eds.; Advances in Chemistry Series-249, ACS: Washington, 1996; 373–396.
43. Billingham, N.C.; Calvert, P.D. The physical chemistry of oxidation and stabilization of polyolefins. In *Developments in Polymer Stabilization*; Scott, G., Ed.; Applied Science: London, 1980; Vol. 3, 139–190.
44. Billingham, N.C.; Prentice, P.; Walker, T.J. Some effects of morphology on oxidation and stabilization of polyolefins. J. Polym. Sci., Polym. Symp. 1976, Deg. Stab. Polyolefins **1977**, *57*, 287–297.
45. Luston, J. Physical loss of stabilizers from polymers. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1980; Vol. 2, 185–240.
46. Al-Malaika, S.; Desai, P.; Scott, G. Mechanisms of antioxidant action: photostabilization of polypropylene using dithiophosphoryl compounds—effects of alkyl substituents. Plast. Rubber Proces. Appl. **1985**, *5* (1), 15–18.
47. Al-Malaika, S.; Goonetileka, M.D.R.J.; Scott, G. Migration of 4-substituted 2-hydroxyphenones in low density polyethylene: Part I. Diffusion characteristics. Polym. Deg. Stab. **1991**, *32* (2), 231–247.
48. Klemchuk, P.P.; Horng, P.L. Transformation products of hindered phenolic antioxidants and color development in polyolefins. Polym. Deg. Stab. **1991**, *34*, 333–346.
49. Al-Malaika, S.; Goodwin, C.; Issenhuth, S.; Burdick, D. The antioxidant role of α-tocopherol in polymers II- melt stabilizing effect in polypropylene. Polym. Deg. Stab. **1999**, *64* (1), 141–156.
50. Yachigo, S.; Ida, K.; Saski, M.; Inoue, K.; Tanaka, S. Studies on polymer stabilizers. Part V. Influences of structural factors on oxidative discoloration and thermal stability of polymers. Polym. Deg. Stab. **1993**, *39*, 317–328.
51. Nesvadba, P.; Krohnke, C. A new class of highly active phosphorous free processing stabilizers for polymers. Proceedings of the Sixth International Conference Additives 97, ECM: New Orleans, 1997.
52. Zweifel, H. *Stabilisation of Polymeric Materials*; Springer-Verlag: Berlin, Germany, 1998; 53 pp.
53. Horsey, D. Hydroxylamines, a new class of low color stabilizers for polyolefins. Proceedings of the Fifth International Conference Additives 96, ECM: Houston, 1996.

54. Henman, T.J. Melt stabilization of polypropylene. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1979; Vol. 1, 39–99.
55. Rabek, J.F. *Mechanisms of Photophysical Processes and Photochemical Reactions in Polymers: Theory and Applications*; Wiley: New York, USA, 1987; 594 pp.
56. Vink, P. Loss of UV stabilizers from polyolefins during photo-oxidation. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Applied Science: London, 1980; Vol. 3, 117–138.
57. Scott, G. Stable radicals as catalytic antioxidants in polymers. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Elsevier Applied Science: London, 1984; Vol. 7, 65–104.
58. Al-Malaika, S.; Scott, G. Photostabilization of polyolefins. In *Degradation and Stabilisation of Polyolefins*; Allen, N.S., Ed.; Applied Science: London, 1983; 283–335.
59. Burton, G.W.; Le Page, Y.; Gabe, E.J. et al. Antioxidant activity of vitamin E and related phenols: importance of stereoelectronic factors. J. Am. Chem. Soc. **1980**, *102* (26), 7791–7792.
60. Al-Malaika, S.; Ashley, H.; Issenhuth, S. The antioxidant role of α-tocopherol in polymers. I The nature of transformation products of α-tocopherol formed during melt processing of LDPE. J. Polym. Sci. Part A, Polym. Chem. **1994**, *32*, 3099–3113.
61. Al-Malaika, S.; Issenhuth, S. The antioxidant role of α-tocopherol in polymers. III. Nature of transformation products during polyolefins extrusion. Polym. Deg. Stab. **1999**, *65* (1), 143–151.
62. Al-Malaika, S.; Issenhuth, S.; Burdick, D. The antioxidant role of vitamin E in polymers V-separation of stereoisomers and characterization of oxidation products of dl-α-tocopherol formed in polyolefins during melt processing. Polym. Deg. Stab. **2001**, *73* (3), 491–503.
63. Pospisil, J. Antioxidants and related stabilizers. In *Oxidation Inhibition of Organic Materials*; Klemchuk, P., Pospisil, J., Eds.; CRC Press: Boca Raton, 1990; Vol. 1, 33–59.
64. Vogl, O.; Albertsson, A.C.; Janovic, Z. Polymerizable, polymeric, polymer-bound (ultraviolet) stabilizers. In *Polymer Stabilisation and Degradation*; Klemchuk, P., Ed.; ACS Symposium Series-280, American Chemical Society: Washington, 1985; 197–210.
65. Bartus, J.; Goman, P.; Sustic, A. Polym. Prep., Div. Polym. Chem., Am. Chem. Soc. **1993**, *34* (2), 158–159.
66. Munteanu, D. Polyolefin stabilization by grafting. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Elsevier Applied Science: London, 1987; Vol. 8, 179–208.
67. Al-Malaika, S. Reactive antioxidants for polymers. In *Reactive Modifiers in Polymers*; Al-Malaika, S., Ed.; Blackie Academic Professional: London, 1997; 266–302.
68. Scott, G. Mechanochemical modification of polymers by antioxidants and stabilizers. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Elsevier Applied Science: London, 1987; Vol. 8, 209–238.
69. Al-Malaika, S.; Suharty, N. Reactive processing of polymers: mechanisms of grafting reaction of functional antioxidants on polyolefins in the presence of a coagent. Polym. Deg. Stab. **1995**, *49* (1), 77–89.
70. Al-Malaika, S.; Scott, G.; Wirjosentono, B. Mechanisms of antioxidant action: polymer-bound hindered amines by reactive processing. III. Effects of reactive antioxidant structure. Polym. Deg. Stab. **1993**, *40* (2), 233–238.
71. Al-Malaika, S.; Ibrahim, A.Q.; Rao, J.; Scott, G. Mechanisms of antioxidant action: photoantioxidant activity of polymer-bound hindered amines. II. bis acrylates. J. Appl. Polym. Sci. **1992**, *44*, 1287–1296.
72. Bauer, I.; Habicher, W.D.; Rauteneberg, C.; Al-Malaika, S. Effects of antioxidants and stabilizers. In *Comprehensive Polymer Science*; Eastmond, G.C., Ledwith, A., Russo, S., Sigwalt, P., Eds.; Pergamon: New York, 1989; Vol. 6, 539–578.
73. Bauer, I.; Habicher, W.D.; Rauteneberg, C.; et al. Antioxidant interaction between organic phosphites and hindered amine light stabilizers during processing and thermoxidation of polypropylene. Polym. Deg. Stab. **1995**, *48* (3), 427–440.
74. Gugumus, F. Developments in the uv stabilization of polymers. In *Developments in Polymer Stabilisation*; Scott, G., Ed.; Elsevier Applied Science: London, 1987; Vol. 8, 209–238.

Biocatalysis

Tyler Johannes
Michael R. Simurdiak
Huimin Zhao
Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.

INTRODUCTION

Biocatalysis may be broadly defined as the use of enzymes or whole cells as biocatalysts for industrial synthetic chemistry. They have been used for hundreds of years in the production of alcohol via fermentation, and cheese via enzymatic breakdown of milk proteins. Over the past few decades, major advances in our understanding of the protein structure–function relationship have increased the range of available biocatalytic applications. In particular, new developments in protein design tools such as rational design and directed evolution have enabled scientists to rapidly tailor the properties of biocatalysts for particular chemical processes. Rational design involves rational alterations of selected residues in a protein to cause predicted changes in function, whereas directed evolution, sometimes called irrational design, mimics the natural evolution process in the laboratory and involves repeated cycles of generating a library of different protein variants and selecting the variants with the desired functions (see the entry ''Protein Design''). Enzyme properties such as stability, activity, selectivity, and substrate specificity can now be routinely engineered in the laboratory. Presently, approximately 100 different biocatalytic processes are implemented in pharmaceutical, chemical, agricultural, and food industries.[1] The products range from research chemicals to commodity chemicals and the number of applications continue to grow very rapidly. In spite of these successes, however, the vast potential of biocatalysis has yet to be fully realized.

In this entry, we briefly outline the scope of biocatalysis and discuss its advantages and disadvantages as compared to chemical catalysis. We then review such topics as enzyme and whole-cell based biocatalysis, biocatalysts used in nonaqueous media, biocatalyst immobilization, discovery and engineering of novel enzymes, and hybrid approaches combining chemical and biological synthesis. An overview of the six general classifications of enzymes along with their relative use in industry is discussed. Selected industrial applications of whole-cell based biocatalysis including the production of lactic acid and 1,3-propanediol are also studied.

THE SCOPE OF BIOCATALYSIS

Advantages and Disadvantages of Biocatalysis vs. Chemical Catalysis

Similar to other catalysts, biocatalysts increase the speed in which a reaction takes place but do not affect the thermodynamics of the reaction. However, they offer some unique characteristics over conventional catalysts (Table 1). The most important advantage of a biocatalyst is its high selectivity. This selectivity is often chiral (i.e., stereo-selectivity), positional (i.e., regio-selectivity), and functional group specific (i.e., chemo-selectivity). Such high selectivity is very desirable in chemical synthesis as it may offer several benefits such as reduced or no use of protecting groups, minimized side reactions, easier separation, and fewer environmental problems. Other advantages, like high catalytic efficiency and mild operational conditions, are also very attractive in commercial applications.

The characteristics of limited operating regions, substrate or product inhibition, and reactions in aqueous solutions have often been considered as the most serious drawbacks of biocatalysts. Many of these drawbacks, however, turn out to be misconceptions and prejudices.[2,3] For example, many commercially used enzymes show excellent stability with half-lives of months or even years under process conditions. In addition, there is an enzyme-catalyzed reaction equivalent to almost every type of known organic reaction. Many enzymes can accept non-natural substrates and convert them into desired products. More importantly, almost all of the biocatalyst characteristics can be tailored with protein engineering and metabolic engineering methods (refer to the section Biocatalyst Engineering and see also the entry ''Protein Design'') to meet the desired process conditions.

Biocatalytic processes are similar to conventional chemical processes in many ways. However, when considering a biocatalytic process one must account for enzyme reaction kinetics and enzyme stability for single-step reactions, or metabolic pathways for multiple-step reactions.[4] Fig. 1 shows the key steps

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120017565

Table 1 Advantages and disadvantages of biocatalysis in comparison with chemical catalysis

Advantages	Disadvantages
Generally more efficient (lower concentration of enzyme needed)	Susceptible to substrate or product inhibition
Can be modified to increase selectivity, stability, and activity	Solvent usually water (high boiling point and heat of vaporization)
More selective (types of selectivity: chemo-selectivity, regio-selectivity, diastereo-selectivity, and enantio-selectivity)	Enzymes found in nature in only one enantiomeric form
Milder reaction conditions (typically in a pH range of 5–8 and temperature range of 20–40°C)	Limiting operating region (enzymes typically denatured at high temperature and pH)
Environment friendly (completely degraded in the environment)	Enzymes can cause allergic reactions

(From Ref.[2].)

in the development of a biocatalytic process. It usually starts with the identification of a target reaction, followed by biocatalyst discovery, characterization, engineering, and process modeling. In many cases, biocatalyst engineering is the most time-consuming step, often involving two major approaches: rational design and directed evolution. In addition to biocatalyst development, product isolation is an important step. The overall process economics depends on all these factors, which needs to be demonstrated in a pilot-scale plant before scale-up. Biocatalysts can constitute a significant portion of the operating budget; however, their cost can be reduced by reusing them when immobilized (refer to the section Biocatalyst Immobilization).

Enzyme Based Biocatalysis vs. Whole-Cell Biocatalysis

Both isolated enzymes and whole cells can be used as biocatalysts. Compared to whole cells, isolated enzymes offer several benefits, including simpler reaction apparatus, higher productivity owing to higher catalyst concentration, and simpler product purification.[2] Until recently, only enzymes that were abundantly produced by cells could be used in industrial applications. Now it is possible to produce large amounts of an enzyme through the use of recombinant DNA technology. In brief, the DNA sequence encoding a given enzyme is cloned into an expression vector and transferred into a production host such as

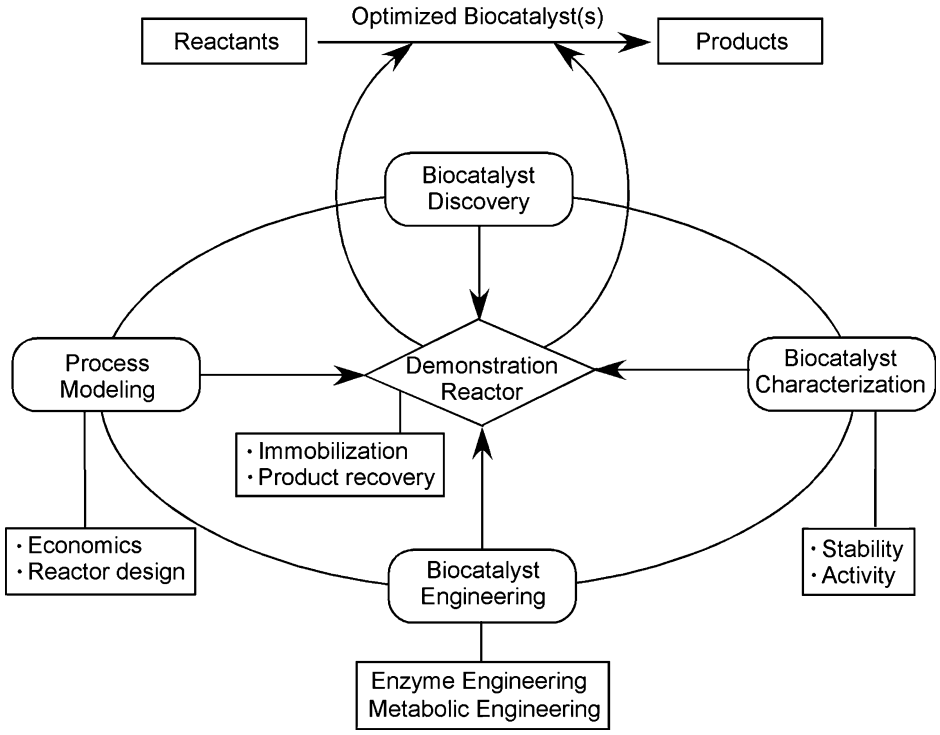


Fig. 1 Flowchart of the development of a biocatalytic process.

Escherichia coli or *Saccharomyces cerevisiae* for gene expression. The overexpressed enzymes are purified from the cell extracts based on their chemical and physical properties. The most commonly used enzyme purification techniques include electrophoresis, centrifugation, and chromatography. Centrifugation separates enzymes based on their differences in mass or shape, whereas electrophoresis separates enzymes based on their differences in charge. Liquid chromatography separates enzymes based on their differences in charge (ion-exchange chromatography), in mass (gel filtration chromatography), or in ligand-binding property (affinity chromatography).

The whole-cell biocatalysis approach is typically used when a specific biotransformation requires multiple enzymes or when it is difficult to isolate the enzyme. A whole-cell system has an advantage over isolated enzymes in that it is not necessary to recycle the cofactors (nonprotein components involved in enzyme catalysis). In addition, it can carry out selective synthesis using cheap and abundant raw materials such as cornstarches. However, whole-cell systems require expensive equipment and tedious work-up because of large volumes, and have low productivity. More importantly, uncontrolled metabolic processes may result in undesirable side reactions during cell growth. The accumulation of these undesirable products as well as desirable products may be toxic to the cell, and these products can be difficult to separate from the rest of the cell culture. Another drawback to whole-cell systems is that the cell membrane may act as a mass transport barrier between the substrates and the enzymes.

Nonaqueous Biocatalysis

Historically, enzymes have been used extensively in aqueous media. Enzymes are well suited to their natural aqueous environment; however, biotransformations in industrial synthesis often involve organic molecules insoluble in water. More importantly, because of its high boiling point and high heat of vaporization, water is usually the least desired solvent of choice for most organic reactions. Thus, shifting enzymatic reactions from an aqueous to an organic medium is highly desired.

Over the past 15 yr, studies have shown that enzymes can work in organic solvents.[5] However, the enzymatic activity is quite low in an organic solvent compared to that in water. Recent advances in protein engineering and directed evolution have aided in the development of enzymes that show improved activity in organic solvents. Progress has also been made in developing simple, scalable, and low-cost techniques to produce highly active biocatalyst preparations for use in organic solvents.[6] One such method improves enzyme activity in organic solvents by lyophilizing (freeze-drying) an aqueous biocatalyst solution in the presence of organic and inorganic molecules called excipients. These excipients include nonbuffer salts, crown ethers, cyclodextrins, and solid-state buffers.[7] Some remarkable results have also been achieved by using ionic liquids as solvents in biocatalytic reactions.[8]

Biocatalyst Immobilization

Immobilization is the process of adhering biocatalysts (isolated enzymes or whole cells) to a solid support. The solid support can be an organic or inorganic material, such as derivatized cellulose or glass, ceramics, metallic oxides, and a membrane. Immobilized biocatalysts offer several potential advantages over soluble biocatalysts, such as easier separation of the biocatalysts from the products, higher stability of the biocatalyst, and more flexible reactor configurations. In addition, there is no need for continuous replacement of the biocatalysts. As a result, immobilized biocatalysts are now employed in many biocatalytic processes.

More than one hundred techniques for immobilizing enzymes have been developed which can be divided into five major groups summarized in Table 2.[9] Adsorption of the enzyme onto a surface is the easiest and the oldest method of immobilization. Entrapment and cross-linking tend to be more laborious enzyme fixation methods, but they do not require altering the enzyme as much as other techniques. The formation of the covalent linkage often requires harsh conditions, which can result in a loss of activity because of conformational changes of the enzyme. It is important to note that most of these techniques can also be used to immobilize whole cells. In addition, although these types of immobilization are considered to be relatively old and well established, the emerging field of nanotube biotechnology has created another possible means of immobilizing biocatalysts.[10]

Biocatalyst Discovery: Sources and Techniques

Traditionally, potentially commercial enzymes are identified by screening micro-organisms, which are frequently isolated from extreme environments, for biocatalytic activity. Commercial enzymes are selected by probing libraries of related enzymes for a range of properties, including activity, substrate specificity, stability over a temperature range, enantio-selectivity, or compatibility under various physical and chemical conditions. Unfortunately, most of the commercially viable enzymes have been isolated in only a few microbial species such as *Bacillus* and *Pseudomonas* because

Table 2 Methods of enzyme immobilization

A. Covalent attachment
Isolated enzymes usually attached through amino or carboxyl groups to a solid support
Variety of supports such as porous glass, cellulose, ceramics, metallic oxides
B. Adsorption
Ion-exchangers frequently used in industry because of simplicity
Industrial applications include anion-exchangers diethylaminoethyl cellulose (DEAE-cellulose) and the cation-exchanger carboxymethyl cellulose (CM-cellulose)
C. Entrapment in polymeric gels
Enzyme becomes trapped in gel volume by changing temperature or adding gel-inducing chemical
Enzymes may be covalently bound to gel (for instance, polyacrylamide cross-linked with *N,N′*-methylenebisacrylamide) or noncovalently linked (calcium alginate)
D. Intermolecular cross-linking
Enzyme cross-linked with bifunctional reagents
Popular cross-linkers are glutaraldehyde, dimethyl adipimidate, dimethyl suberimidate, and aliphatic diamines
E. Encapsulation
Enzymes enveloped in semipermeable membrane, which allows low molecular weight substrates and products to pass through the membrane
Enclosed in a variety of devices: hollow fibers, cloth fibers, microcapsules, film

(From Ref.[9].)

of the limitations in micro-organism cultivation techniques.[11] It has been widely acknowledged that the majority of microbial species (up to 99%) have never been cultivated and thus have never been investigated. To access this vast untapped microbial diversity, several companies such as Diversa Corporation (San Diego, California, U.S.A.) and TerraGen Discovery (Vancouver, British Columbia, Canada) have successfully developed modern bioprospecting techniques such as multiple metagenome cloning to isolate novel industrial enzymes.[12]

New methods for exploring natural biodiversity have been greatly facilitated by high-throughput screening technologies and robust expression in recombinant organisms. Recombinant DNA technology makes it possible to produce enzymes at levels 100-fold greater than native expression and allows expression of genes from organisms that cannot be cultured. Although some problems may be resolved by screening larger libraries of DNA, this may not be the most efficient or expedient method of obtaining a viable biocatalyst. A more efficient means of obtaining a good biocatalyst may involve engineering the catalyst itself using various protein engineering and metabolic engineering techniques (refer to the section Biocatalyst Engineering and see also the entry ''Protein Design'').

Biocatalyst Engineering

Nature has supplied us with a vast array of biocatalysts capable of catalyzing numerous biological reactions. Unfortunately, naturally occurring biocatalysts are often not optimal for many specific industrial applications, such as low stability and activity. Moreover, naturally occurring biocatalysts may not catalyze the reaction with the desired non-natural substrates or produce the desired products. To address these limitations, molecular techniques have been developed to create improved or novel biocatalysts with altered industrial operating parameters. It should be noted that, for enzyme based biocatalysts, many molecular techniques have been developed for engineering enzymes with novel or improved characteristics. Readers are referred to the entry ''Protein Design.'' In this section, we mainly discuss the molecular techniques used for whole-cell based biocatalyst engineering, or metabolic engineering.

Metabolic engineering is a rapidly growing area with great potential to impact biocatalysis.[13] It has been broadly defined as ''the directed improvement of product formation or cellular properties through modifications of specific biochemical reaction(s) or the introduction of new one(s) with the use of recombinant DNA technology.''[14] In an industrial context, the ultimate goal of metabolic engineering is the development of optimal biocatalysts. In the past two decades, metabolic engineering has been successfully used to engineer micro-organisms to produce a wide variety of products, including polymers, aromatics, carbohydrates, organic solvents, proteins, antibiotics, amino acids, and organic acids. According to the approach taken or the aim, these applications can be classified into seven groups: 1) expression of heterologous genes for protein production; 2) extension of the range of substrate for cell growth and product formation; 3) design and construction of pathways for the production of new products; 4) design and

construction of pathways for degradation of xenobiotics; 5) engineering of cellular physiology for process improvement; 6) elimination or reduction of by-product formation; and 7) improvement of yield or productivity.[15] Several of these applications have been implemented at industrial-production scale (refer to the section Industrial Applications of Whole-Cell Based Biocatalysis) and the number of applications should continue to grow. In particular, with the recent advances in genomics, proteomics, and bioinformatics, many new genes and pathways will be discovered and the regulation of metabolic network will also be better understood, all of which will accelerate the development of more commercially viable bioprocesses through metabolic engineering.

Hybrid Approaches Combining Chemical Synthesis and Biocatalysis

Biocatalysts exhibit exquisite catalytic efficiency that is often unmatched by conventional catalysis. Nonetheless, conventional organic synthesis will likely remain the staple of the chemical and pharmaceutical industries. In the future, the integration of these two approaches will probably offer the optimal route for industrial synthesis. An illustration of this principle can be found in the selective deprotection of reactive functional groups. Enzymes are unique deprotecting tools for combinatorial synthesis because of their remarkable selectivity and ability to operate under mild reaction conditions. A recent example is the synthesis of long multiply lipidated peptides containing various side-chain functional groups.[16] In this study, penicillin acylase was used for selective N-deprotection of a highly labile S-palmitoylated oligopeptide. After removal of the protecting group, the S-palmitoylated oligopeptide was used as a building block in further synthetic steps.

INDUSTRIAL APPLICATIONS OF ENZYME BASED BIOCATALYSIS

With the rapid technical developments in gene discovery, optimization, and characterization, enzymes have been increasingly used as biocatalysts. According to the International Union of Biochemistry and Molecular Biology (IUBMB) nomenclature system, all enzymes are classified into six classes on the basis of the general type of reactions that they catalyze (Table 3). Within each class are subclasses and the enzymes themselves. The result is an ordered system of enzymes and the reaction(s) that each catalyzes. It is important to note that, in biological processes, every class of enzyme is utilized in the cell to a large extent. However, this is not the same in industrial processes, where certain classes of enzymes are used more often than others. As shown in Fig. 2, most of the enzymes that have been used as biocatalysts in industry are hydrolases (~65%), even though oxidoreductases are typically much more useful than hydrolases as catalysts. The utility of an enzyme class depends on the relative commercial importance of the products that each enzyme produces, the accessibility of the enzymes, and the specific characteristics of the enzymes (e.g., stability, activity, and selectivity).

Oxidoreductases

Oxidoreductases catalyze oxidation and reduction reactions that occur within the cell. They are very appealing for industrial uses because of the reactions that they are able to catalyze. However, they often need expensive cofactors such as nicotinamide adenine dinucleotides (e.g., NAD^+/NADH) and flavines (e.g., $FAD/FADH_2$) in the reactions. In fact, nicotinamide adenine dinucleotides are required by about 80% of oxidoreductases. Fortunately, several NAD(H)

Table 3 Classification of enzymes

Enzymes	Type of reactions	Representative subclasses
Oxidoreductases	Catalyze the transfer of hydrogen or oxygen atoms or electrons from one substrate to another	Oxidases, oxygenases, peroxidase, dehydrogenases
Transferases	Catalyze the group transfer reactions	Glycosyltransferases, transketolases, methyltransferases, transaldolases, acyltransferases, transaminases
Hydrolases	Catalyze hydrolytic reactions	Esterases, lipases, proteases, glycosidases, phosphatases
Lyases	Catalyze the nonhydrolytic removal of groups	Decarboxylases, aldolases, ketolases, hydratases, dehydratases
Isomerases	Catalyze isomerization reactions	Racemases, epimerases, isomerases
Ligases	Catalyze the synthesis of various types of bonds with the aid of energy-containing molecules	Synthetases, carboxylases

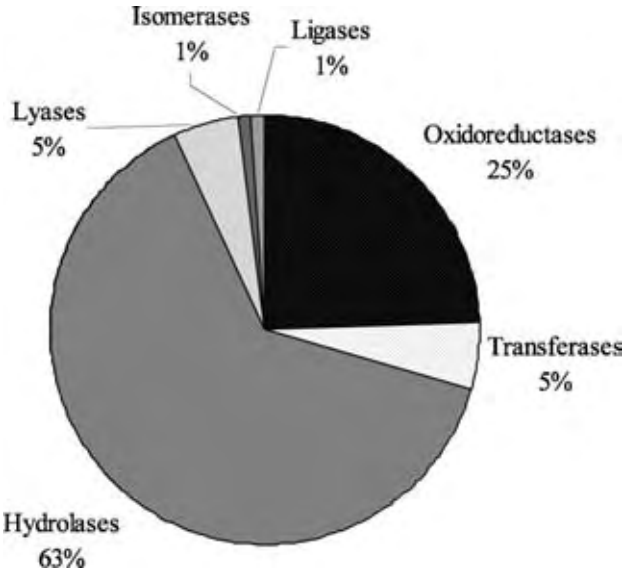


Fig. 2 The relative use of enzyme classes in industry. (From Ref.[2].)

regeneration systems have been developed, the most widely used being the formate/formate dehydrogenase (FDH) system.[17]

An example of a pharmaceutical synthesis reaction involving an oxidoreductase is the synthesis of 3,4-dihydroxylphenyl alanine (DOPA).[2] 3,4-Dihydroxylphenyl alanine is a chemical used in the treatment of Parkinson's disease. The industrial process that synthesizes DOPA utilizes the oxidoreductase polyphenol oxidase. As shown in Fig. 3, the monohydroxy compound is oxidized by the regio-specific addition of a hydroxyl group. It is worth mentioning that epinephrine (adrenaline) can also be synthesized by a similar reaction path using the same enzyme.[2]

Another example is the use of leucine dehydrogenase coupled with FDH for the reductive amination of trimethylpyruvate to L-*tert*-leucine (Fig. 4). The whole

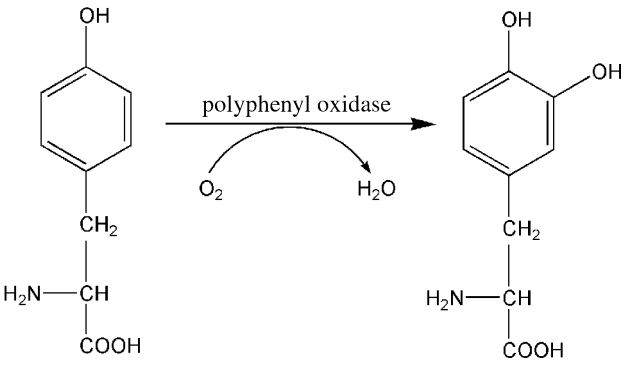


Fig. 3 Enzymatic synthesis of 3,4-dihydroxylphenyl alanine (DOPA).

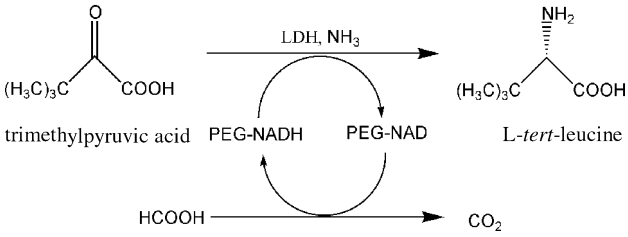


Fig. 4 Enzymatic synthesis of L-*tert*-leucine.

process is carried out in a membrane reactor in which the cofactor NAD^+ is regenerated by FDH. This process has now reached ton-scale production at Degussa (Germany).[18]

Transferases

Transferases catalyze the transfer of functional groups such as methyl, hydroxymethyl, formal, glycosyl, acyl, alkyl, phosphate, and sulfate groups by means of a nucleophilic substitution reaction. They are not widely used in industrial processes; however, there are a few examples of industrial processes that utilize transferases.

A classical example of industrial application of transferases is the use of various glycosyltransferases for the synthesis of oligosaccharides. Oligosaccharides and polysaccharides are important classes of naturally occurring compounds, which play vital roles in cellular recognition and communication processes.[19] Because of the required use of many protection and deprotection groups, chemical synthesis of complex oligosaccharides represents a daunting challenge in synthetic organic chemistry. By contrast, enzymatic synthesis of oligosaccharides by glycosyltransferases requires very few protection and deprotection steps because of the high regio- and stereoselectivity of glycosyltransferases, thus offering an attractive alternative.[2] Another example is the use of a glucokinase (a transferase) in combination with an acetate kinase for the production of glucose-6-phosphate (Fig. 5). This process is carried out in multikilogram scale by the Japanese company Unitika.[1]

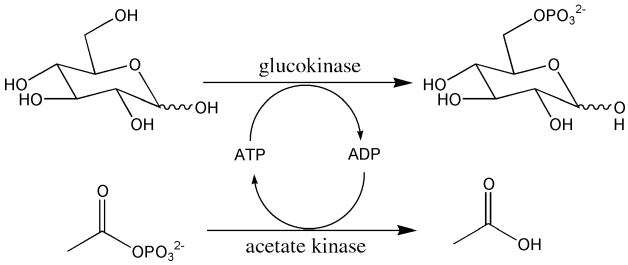


Fig. 5 Enzymatic synthesis of glucose-6-phosphate.

Hydrolases

Hydrolases catalyze the addition of water to a substrate by means of a nucleophilic substitution reaction. Hydrolases (hydrolytic enzymes) are the biocatalysts most commonly used in organic synthesis. They have been used to produce intermediates for pharmaceuticals and pesticides, and chiral synthons for asymmetric synthesis. Of particular interest among hydrolases are amidases, proteases, esterases, and lipases. These enzymes catalyze the hydrolysis and formation of ester and amide bonds.

Lipases can hydrolyze triglycerides into fatty acids and glycerol. They have been used extensively to produce optically active alcohols, acids, esters, and lactones by kinetic resolution. Lipases are unique, in that they are usually used in two-phase systems. A classic example is the use of a lipase for the production of (*S,R*)-2,3-*p*-methoxyphenylglycyclic acid, an intermediate for diltiazem. In this process, methyl-*p*-methoxyphenylglycidate is stereospecifically hydrolyzed by a lipase immobilized in a hollow fiber membrane reactor. The enzyme is located at the interfacial layer between an organic and an aqueous phase.[1]

Proteases such as α-chymotrypsin, papain, and subtilisin are also useful biocatalysts for regio-selective or stereoselective hydrolytic biotransformations. For example, dibenzyl esters of aspartic and glutamic acid can be selectively deprotected at the 1-position by subtilisin-catalyzed hydrolysis (Fig. 6).[2] In addition, α-chymotrypsin is used in the kinetic resolution of α-nitro-α-methyl carboxylates, which results in L-configured enantiomers of the unhydrolyzed esters with high optical purity (>95% e.e.).[2]

Lyases

Lyases are the enzymes responsible for catalyzing addition and elimination reactions. Lyase-catalyzed reactions involve the breaking of a bond between a carbon atom and another atom such as oxygen, sulfur, or another carbon atom. They are found in cellular processes, such as the citric acid cycle, and in organic synthesis, such as in the production of cyanohydrins.[2]

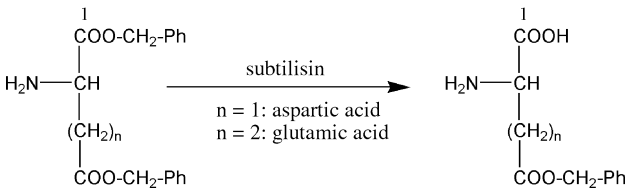


Fig. 6 Regio-selective ester-hydrolysis catalyzed by subtilisin.

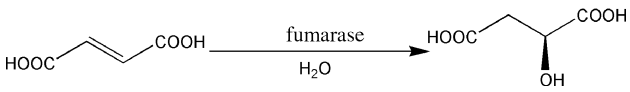


Fig. 7 Enzymatic synthesis of (*S*)-malic acid.

Several industrial processes using lyases as catalysts have been reported. Perhaps the most prominent lyase-catalyzed process is the production of acrylamide from acrylnitrile. This process is carried out by the Nitto Chemical Company of Japan at a scale of more than 40,000 tons per year.[20] Another example is the use of a fumarase for the production of (*S*)-malic acid from fumaric acid. As shown in Fig. 7, a water molecule is added to the double bond in fumarate by means of an addition reaction. The result is a cleavage of the carbon–carbon double bond, and a formation of a new carbon–oxygen bond. A third example is biocatalytic production of a cyanohydrin from a ketone. This reaction is catalyzed by a lyase called oxynitrilase. It consists of the cleavage of one carbon–oxygen bond, and the addition of a HCN molecule. The chirality of the product is based on the form of the enzyme used (*R*-oxynitrilase or *S*-oxynitrilase).[2]

Isomerases

Isomerases catalyze isomerization reactions such as racemization and epimerization. They have not been used in many industrial applications. However, one of the most successful enzyme based biocatalytic processes involves an isomerase: the use of glucose isomerase for the production of high-fructose corn syrup (HFCS) (Fig. 8). High-fructose corn syrup is used as an alternative sweetener to sucrose in the food and beverage industry. The isomerization of glucose to HFCS on an industrial scale is carried out in continuous fixed-bed reactors using immobilized glucose isomerases. The total amount of HFCS produced by glucose isomerase exceeds a million tons per year.[21]

Ligases

Ligases catalyze reactions that involve the creation of chemical bonds with nucleotide triphosphates. They

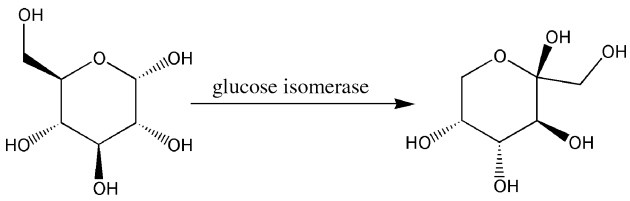


Fig. 8 Enzymatic synthesis of high-fructose corn syrup.

are important in certain cellular processes, such as connecting nucleotides in DNA replication. However, similar to isomerases, ligases have very few industrial applications.[2] It is important to note that DNA ligases are essential tools in recombinant DNA technology and are used almost in every biology-related laboratory.

INDUSTRIAL APPLICATIONS OF WHOLE-CELL BASED BIOCATALYSIS

Whole-cell based biocatalysis utilizes an entire micro-organism for the production of the desired product. One of the oldest examples for industrial applications of whole-cell biocatalysis is the production of acetic acid from ethanol with an immobilized *Acetobacter* strain, which was developed nearly 200 yr ago.[1] The key advantage of whole-cell biocatalysis is the ability to use cheap and abundant raw materials and catalyze multistep reactions. Recent advances in metabolic engineering have brought a renaissance to whole-cell biocatalysis. In the following sections, two novel industrial processes that utilize whole-cell biocatalysis are discussed with emphasis on the important role played by metabolic engineering.

Lactic Acid

L-lactic acid has long been used as a food additive and has recently received great attention because it can be used as an important feedstock for the production of other chemicals such as polylactic acid (PLA), acetaldehyde, polypropylene glycol, acrylic acid, and pentadione.[22] Among them, PLA is the most important product as it can be used to manufacture thermoformed containers, packaging, nonwovens, paper-coated articles, and film products.[23] Lactic acid can be produced from sucrose, whey (lactose), and maltose or dextrose from hydrolyzed starch using *Lactobacillus* strains.

Compared to other polymeric materials such as polyethylenes, polylactic acid has several advantages including an increased hydrophilicity, resistance to ultraviolet light, ability to be dyed with dispersion dyes, a range of melting temperatures between 120°C and 170°C, and low flammability and smoke generation. Most importantly, polylactic acid is biodegradable and is derived from renewable resources, utilizing energy from the sun and lowering the fossil fuel dependence for production.[23]

In recognition of the superior properties and the huge potential market of polylactic acid, Cargill Inc. and The Dow Chemical Company started a joint venture Cargill Dow LLC to produce lactic acid using fermentation (whole-cell biocatalysis) several years ago. A production plant was built in Blair, NE, in 2001 and it now produces 140,000 metric tons of polylactic acid per year. It is predicted that the eventual cost of polylactic acid will be between $0.50 and $0.75 per pound (http://www.cargilldow.com).

One of the drawbacks in the current commercial fermentation process is that the predominant form of the product is the deprotonated lactate rather than lactic acid, requiring more expensive and wasteful product purification steps. This is because the *Lactobacillus* fermentation operates at a minimum pH of 5.0–5.5 which is above the pK_a of lactic acid (3.87). To overcome this limitation, a powerful strain improvement method, genome shuffling, was used to improve the acid tolerance of a poorly characterized industrial strain of *Lactobacillus*.[24] A population of strains with subtle improvement in pH tolerance was isolated using classical strain improvement methods such as chemostats, and were then shuffled by recursive pool-wise protoplast fusion to create mutant strains that grow at substantially lower pH than does the wild-type strain.

1,3-Propanediol

1,3-Propanediol is an intermediate that is widely used in the synthesis of polyesters and polyurethanes. Polymers based on 1,3-propanediol are very useful in the carpet and textile industry because of their good light stability and biodegradability.[25] The conventional methods for producing 1,3-propanediol rely on petroleum derivatives and are quite capital intensive and/or generate waste streams containing environmental pollutants. Thus, the use of micro-organisms to produce 1,3-propanediol from glucose represents an attractive alternative.

Both the biological production of 1,3-propanediol from glycerol and that of glycerol from glucose have been known for many years.[13] However, there is no single micro-organism that could convert basic carbon sources such as glucose to the desired 1,3-propanediol end-product. Such a micro-organism is highly desired in the process as it requires less energy input and uses an inexpensive starting material.

A team of researchers from DuPont and Genencor has successfully used metabolic engineering techniques to engineer such a micro-organism. The conversion of glucose to 1,3-propanediol requires the combination of two natural pathways: glucose to glycerol and glycerol to 1,3-propanediol. The best natural pathway for the production of glycerol from glucose was found in the yeast *Saccharomyces cerevisiae*, which consists of two enzymes: dihydroxyacetone-3-phosphate dehydrogenase and glycerol-3-phosphate phosphatase. The best natural

pathway for production of 1,3-propanediol from glycerol was found in *Klebsiella pneumoniae*, which consists of glycerol dehydratase and 1,3-propanediol dehydrogenase. The genes encoding these two natural pathways were cloned and expressed in *E. coli*. *E. coli* was chosen as the production strain because it has been used in large-scale production on an industrial level, it has many genetic tools, and its metabolism and physiology are well characterized. This engineered *E. coli* was found to produce over 120 g/L of 1,3-propanediol in 40 hr fed-batch fermentation.[13]

CONCLUSIONS

Biocatalysis has become an important tool for industrial chemical synthesis and is on the verge of significant growth. In the past several decades, many biocatalytic processes have been implemented to produce a wide variety of products in various industries. Most of them use naturally occurring enzymes or micro-organisms as catalysts. With the help of innovative biocatalyst discovery methods and advances in protein engineering and metabolic engineering, the time and cost of developing new biocatalysts can be reduced significantly. Most importantly, the biocatalysts can be readily tailored to their specific applications and process conditions through protein engineering and metabolic engineering. It is possible that in the future they can be rationally designed to act specifically in any chemical reaction of interest, fulfilling the holy grail of catalysis: catalysis by design. In addition, the use of biocatalysts in organic solvents in combination with the integration of biocatalysis and chemical catalysis will continue to broaden the scope of the applications of biocatalysts. New advances in genomics, proteomics, and bioinformatics will fuel the development of biocatalysis as an integral part of industrial catalysis.

REFERENCES

1. Wandrey, C.; Liese, A.; Kihumbu, D. Industrial biocatalysis: past, present, and future. Organic Process Res. Dev. **2000**, *4* (4), 286–290.
2. Faber, K. Biotransformations. In *Organic Chemistry: A Textbook*, 3rd Ed.; Springer-Verlag: Berlin, Germany, 1997.
3. Rozzell, J.D. Commercial scale biocatalysis: myths and realities. Bioorg. Med. Chem. **1999**, *7* (10), 2253–2261.
4. Schmid, A.; Dordick, J.S.; Hauer, B.; Kiener, A.; Wubbolts, M.; Witholt, B. Industrial biocatalysis today and tomorrow. Nature **2001**, *409* (6817), 258–268.
5. Klibanov, A.M. Improving enzymes by using them in organic solvents. Nature **2001**, *409* (6817), 241–246.
6. Schoemaker, H.E.; Mink, D.; Wubbolts, M.G. Dispelling the myths—biocatalysis in industrial synthesis. Science **2003**, *299* (5613), 1694–1697.
7. Lee, M.Y.; Dordick, J.S. Enzyme activation for nonaqueous media. Curr. Opin. Biotechnol. **2002**, *13* (4), 376–384.
8. Kragl, U.; Eckstein, M.; Kaftzik, N. Enzyme catalysis in ionic liquids. Curr. Opin. Biotechnol. **2002**, *13* (6), 565–571.
9. Klibanov, A.M. Immobilized enzymes and cells as practical catalysts. Science **1983**, *219* (4585), 722–727.
10. Martin, C.R.; Kohli, P. The emerging field of nanotube biotechnology. Nat. Rev. Drug Discovery **2003**, *2* (1), 29–37.
11. Dalboge, H.; Lange, L. Using molecular techniques to identify new microbial biocatalysts. Trends Biotechnol. **1998**, *16* (6), 265–272.
12. Burton, S.G.; Cowan, D.A.; Woodley, J.M. The search for the ideal biocatalyst. Nat. Biotechnol. **2002**, *20* (1), 37–45.
13. Chotani, G.; Dodge, T.; Hsu, A.; Kumar, M.; LaDuca, R.; Trimbur, D.; Weyler, W.; Sanford, K. The commercial production of chemicals using pathway engineering. Biochim. Biophys. Acta **2000**, *1543* (2), 434–455.
14. Stephanopoulos, G.N.; Aristidou, A.A.; Nielsen, J. *Metabolic Engineering: Principles and Methodologies*; Academic Press: London, U.K., 1998.
15. Nielsen, J. Metabolic engineering. Appl. Microbiol. Biotechnol. **2001**, *55* (3), 263–283.
16. Machauer, R.; Waldmann, H. Synthesis of lipidated eNOS peptides by combining enzymatic, noble metal- and acid-mediated protecting group techniques with solid phase peptide synthesis and fragment condensation in solution. Chemistry **2001**, *7* (13), 2940–2956.
17. Chenault, H.K.; Whitesides, G.M. Regeneration of nicotinamide cofactors for use in organic synthesis. Appl. Biochem. Biotechnol. **1987**, *14* (2), 147–97.
18. Kragl, U.; VasicRacki, D.; Wandrey, C. Continuous production of L-tert-leucine in series of two enzyme membrane reactors—modelling and computer simulation. Bioprocess Eng. **1996**, *14* (6), 291–297.
19. Ginsburg, V., Robbins, P.W., Eds. *Biology of Carbohydrates*; Wiley: New York, 1984.
20. Zaks, A. Industrial biocatalysis. Curr. Opin. Chem. Biol. **2001**, *5* (2), 130–136.
21. Gerhartz, W., Ed. *Enzymes in Industry: Production and Applications*; VCH: New York, 1990.

22. Varadarajan, S.; Miller, D.J. Catalytic upgrading of fermentation-derived organic acids. Biotechnol. Prog. **1999**, *15* (5), 845–854.
23. Hunt, J. Large-scale production, properties and commercial applications of polylactic acid polymers. Polym. Degrad. Stab. **1998**, *59* (1–3), 145–152.
24. Patnaik, R.; Louie, S.; Gavrilovic, V.; Perry, K.; Stemmer, W.P.; Ryan, C.M.; Del Cardayre, S. Genome shuffling of Lactobacillus for improved acid tolerance. Nat. Biotechnol. **2002**, *20* (7), 707–712.
25. Zhu, M.M.; Lawman, P.D.; Cameron, D.C. Improving 1,3-propanediol production from glycerol in a metabolically engineered Escherichia coli by reducing accumulation of sn-glycerol-3-phosphate. Biotechnol. Prog. **2002**, *18* (4), 694–699.

Biofilms

T. Reg. Bott
School of Engineering, Chemical Engineering, University of Birmingham, Birmingham, U.K.

INTRODUCTION

The term ''biofilm'' refers to a colony of living matter, usually microorganisms including algae, fungi, or bacteria, together with extracellular material attached to a solid surface. In natural conditions it will almost certainly contain several different microbial species, and also depending on the prevailing conditions, the biofilm may also contain adventitious particulate matter. In processing operations, biofilms can be an asset or a hindrance. In some cases, the biofilm is essential for achieving the objective of the process, e.g., the use of trickle filters for the removal of contamination to improve the quality of potable water.[1] On the other hand, the effectiveness of a cooling water system can be reduced by the presence of biofilms on heat transfer surfaces.[2]

Biofilm formation is a natural phenomenon, so that wherever suitable conditions exist, it would be anticipated that a biofilm would develop. Essential, of course, are the presence of nutrients to sustain life and a suitable temperature. Biofilms are formed on surfaces; so the availability of a suitable surface that may be colonized by microorganisms is vital. The structure and the composition of the biofilm are greatly dependent on the environment in which it grows. The composition of the fluid (usually water) in contact with the biofilm, including organic matter, oxygen, and trace elements, its pH, and its flow conditions across solid surfaces, all affect the rate of growth and the robustness of the biofilm structure that develops. The removal of waste products from the vicinity of the biofilm will also depend on the structure of the biofilm, in terms of porosity and thickness, and the flow conditions internal and external to the biofilm. The structure of the biofilm will also influence the availability of nutrients within the biofilm, because of its effect on the penetration of the fluid carrying the nutrients within the biofilm. The biofilm activity may affect the surface on which it resides. For example, the pH near the base of the biofilm may be different from that of the bulk fluid external to the biofilm, because of the metabolism of the microorganisms. This may give rise to corrosion or deterioration of the surface. Clearly, this is an important consideration in relation to the integrity of processing equipment.

The initiation and continued growth of a biofilm is an extremely complex process, subject to many internal and external influences. A discussion of these factors forms the basis of this entry.

MICROBIAL ACTIVITY

Bacteria may be regarded as more versatile than algae or fungi because they are not limited by the need for light or a consumable substrate. In addition, there is considerable variation among bacteria, which is due, in part at least, to the differences in the properties of cell surface polymers.[3] Fletcher[4] suggests that the bacteria attached to a surface appear to be metabolically different from their planktonic or ''free swimming'' counterparts.

Algae utilize carbon dioxide and inorganic chemicals as their primary source of nutrients, and light (usually sunlight) to photosynthesize sugars. The associated biofilms are generally composed of single cells or filamentous organisms, but other forms can exist depending on the conditions. In natural conditions, algal biofilms are found on rocks or stones, in rivers, or on the seashore. Colonies can also reside on man-made structures such as bridge supports standing in water or offshore oil rigs.

Fungi require a fixed organic source of carbon. The rigid cell wall limits them to being saprophytic on organic substrates or as parasites on animals, plants, algae, or even other fungi. In fact, fungi may be found on any solid that provides an organic substrate, provided that the local conditions are satisfactory.

Microbial cells are surrounded by a cell wall, which retains the cell contents, and is the primary barrier between the cell surface and the environment in which it exists. The quality of the cell wall, in terms of selective permeability, maintains the necessary levels of nutrients, trace elements, and cell internal pH. The cell membrane is the site of transfer processes: water is able to pass through this membrane, in or out of the cell, depending on the thrust of the osmotic pressure. The chemistry of the cell wall affects its properties in terms of surface electric charge and the availability of binding ions.

In many organisms, the cell wall is rigid, giving a characteristic shape to individual cells such as rod, filament, or sphere. The rigidity of the cell wall allows the development of structures that may be beneficial for the maintenance of a coherent biofilm. Some cells, however, do not have a rigid cell wall and, therefore, require an intrinsic mechanism to control osmotic

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030621

pressure for the prevention of damage by excess water intake. A few freshwater organisms can dispose of water or imbibe it through contractile vacuoles. Certain algae can control the condition of soluble metabolites to counter the effects of osmotic pressure.

Microorganisms can produce extracellular materials, such as slimes of polysaccharides and mucilages, which may help to maintain attachment to the solid substrate, provide a source of nutrients if the nutrient availability declines for any reason, or enhance protection of the cells.

Clearly, the availability of nutrients will determine whether or not a biofilm can form and develop. In common with all living matter, the elements that constitute microorganisms are associated with organic chemistry in the widest sense, including carbon, hydrogen, oxygen, nitrogen, sulfur, phosphorus, and other inorganic molecules. The requirements are a source of energy, carbon, and reducing chemicals. Although CO_2 is an increasing component of the atmosphere, it is used as a nutrient source of carbon only by algae, via photosynthesis. The usual source of carbon for microorganisms is carbohydrates, especially polysaccharides because of their abundance. Nevertheless, other sources used include fats, hydrocarbons such as methane, and proteins. Cellulose is preferred by some fungi.

The inertness of atmospheric nitrogen precludes its availability for microorganisms. The usual source is nitrogen-containing compounds such as amino acids, ammonia, nucleotides, uric acid, and urea.

Sulfur is plentiful in naturally occurring compounds. Inorganic sulfates can be reduced via the sulfide and incorporated into amino acids. Hydrogen sulfide is used as a source of sulfur by some microorganisms. Organic sulfur may represent an alternative sulfur-containing nutrient.

Trace elements, including potassium, magnesium, and iron, are required by microorganisms, while calcium, sodium, and silica may be necessary for the growth of some species. There may also be a requirement for traces of zinc, copper, cobalt, manganese, and molybdenum.

The metabolism of microorganisms gives a basis for classification:[2]

1. Phototrophs obtain their energy directly from the sun.
2. Chemotrophs acquire their energy from the oxidation of organic as well as inorganic compounds.
3. Autotrophs are capable of synthesizing their cell carbon from simple compounds such as CO_2.
4. Heterotrophs require fixed sources of carbon.
5. Lithotrophs produce the reducing equivalents required for cell synthesis from inorganic substances like H_2S and ferrous iron.
6. Organotrophs obtain chemical reduction from the oxidation of organic molecules.

Movement to a favorable location for nutrient assimilation is achieved by the use of flagella, as with some bacteria, or by the use of fibrils, as with some algae cells to glide across surfaces.

During metabolism, a single cell will take in nutrients allowing the synthesis of macromolecules that provide the basis for growth. As the cell increases in size, it will eventually divide. Cell division in bacteria is well ordered; the daughter cells that result from the division are uniform in dimension and genetic character. The time to reach full size can be very short (as low as a few hours) depending on the prevailing conditions, particularly temperature and nutrient availability.

Cell division in some fungi starts with the production of a ''bud'' on the cell surface. Of course, the bud is much smaller than the original cell, but it grows till it reaches adult dimensions, and the reproductive process is repeated. There is a limit to the number of buds that can be produced by a single parent cell, as only one can be grown at a particular point on the surface of the cell. Fungi grow almost exclusively through what is known as apical growth, though they can also reproduce by means of sexual and asexual cycles.

For the development of algal biofilms, as with bacteria, few cells are needed to attach to a surface, as cell division rapidly gives rise to colonies that eventually coalesce to form a compact biofilm.[5]

THE IMPORTANCE OF A SURFACE

It would appear that many microorganisms tend to reside on a surface, in discrete colonies, or in a film. It has been suggested[3] that under natural conditions, i.e., in rivers or lakes, 90% of the microorganisms in the biosphere exist in biofilms. Of fundamental importance is that the microorganisms contained in a developing biofilm are exposed to a continuous supply of nutrients. The disturbance of the surface of the water source (e.g., a lake) in contact with the atmosphere ensures there is aeration, which would be advantageous for aerobic organisms. Furthermore, the movement of the water would assist the removal of waste products of the metabolic processes in the cells. The advantages experienced by the sessile cells (those attached to a surface) may be contrasted with the depravation of the planktonic cells (those suspended in water) where nutrient availability at the cell surface is largely because of Brownian diffusion, the cells moving with the water flow along the water streamlines.

Other advantages that accrue for biofilms on surfaces[4] include protection from short-term fluctuations in pH, salt and biocide concentrations, and dehydration. Because of the mass-transfer resistance of the biofilm to the transport of nutrient, conditions will change throughout the biofilm depth. Where several

different microorganisms are present, it is possible for symbiosis to occur. Symbiosis may be defined as a mutually beneficial partnership between organisms of different kinds. Examples include the utilization and breakdown of less readily degradable substances by specialized organisms, which provides a source of nutrients for other microorganisms, e.g., the degradation of cellulose by fungi. The creation of ecological niches, such as anaerobic zones under aerobic biofilms, allows the growth of anaerobic cells in otherwise aerobic environments.

COLONIZATION OF SURFACES

It is generally accepted that the colonization of a surface by microorganisms is a multistage process, as suggested by Characklis.[6] Five steps are involved.

Formation of an Organic Conditioning Layer

A conditioning layer on the virgin solid surface, newly immersed in water, is created by the adsorption of macromolecules. Chamberlain[7] has studied the importance of these adsorbed macromolecules in the attachment of microbes to solid surfaces. The adsorbed macromolecules on surfaces are generally organic in character, although there may be other compounds included in the adsorbed layer, such as metallic hydroxides and mineral particles. Because macromolecules frequently possess multiple attachment points, either functional groups or segments of more hydrophobic character, their attachment is likely to be irreversible. In freshwater systems, such as cooling water circuits, the macromolecules involved, including humic acids and polysaccharides, are likely to be derived from formally living matter. The humic components originate primarily from lignin-type materials and may contain carboxylic and phenolic residues that impart high reactivity. These components are generally highly aromatic. In seawater used for cooling purposes, humic compounds are somewhat different in character and generally comprise complex condensation molecules with polysaccharide, peptide, and lipid components. In contrast to fresh water systems, the marine humic compounds are generally aliphatic in character. In food-processing streams, the macromolecules are often proteins or glycoproteins, which may bring about the attachment of casein and other milk proteins to stainless steel, the usual material of construction in food processing equipment. The rate of adsorption of macromolecules is, in part at least, controlled by their concentration in the water, their relative affinities for the material with which the surface is made, and the hydrodynamics of the flow across the surface.

Transport of Microorganisms to a Surface

The transport of microbial cells dispersed in the liquid stream toward the surface is strongly influenced by the regime of flow across the surface, i.e., laminar or turbulent (see item elsewhere in this encyclopedia). The liquid immediately in contact with the solid surface may be regarded as being stationary, because of the viscous drag exerted by the solid surface. The velocity of the liquid layers gradually increases at distances at right angles to the surface. When close to the surface, these layers are slow moving and constitute what is generally known as the viscous or laminar sublayer. In simple terms, if the bulk flow is turbulent, a boundary layer exists between the slow-moving laminar sublayer and the turbulent bulk flow. For microorganisms to approach the surface and become attached to the conditioning layer, they have to pass from the turbulent bulk liquid, through the boundary layer and into the viscous sublayer. Initially, the cells are carried by eddy diffusion (i.e., resulting from the turbulence in the bulk liquid), but as they approach and enter the viscous sublayer, the eddy diffusion is damped out. Within the viscous sublayer itself, the transport mechanism is caused by molecular diffusion or Brownian motion. The thickness of the viscous sublayer is strongly dependent on the velocity in the bulk liquid. The greater the bulk flow, the thinner the viscous sublayer; i.e., the resistance to the transport of the cells to the surface is lower, and hence a more rapid build-up of cells on the surface is possible. Frictional drag forces become significant in the viscous sublayer, which slows the cells as they move towards the surface. Complex interactions between the cells and the effects of the flow conditions are also likely to affect the approach velocity; for instance, the disturbance of the laminar sublayer by the turbulence in the bulk flow.

Thermophoresis, where particles move in response to a temperature gradient, may also affect the movement of cells in relation to a surface. Because cells have a large water content, gravity will have little effect on settlement. As already stated, cells may respond to chemical stimuli that influence movement in relation to surfaces, and it is possible that the conditioning layer, because of its composition, attracts cells by chemotaxis, thereby facilitating colonization.

Attachment to a Surface

The attachment of particles in a fluid suspension to a solid surface is extremely complex and involves long-range attraction forces to bring the particles to the surface and provide a basis for further interaction. The forces involved may include van der Waals forces

and electrostatic forces. Because of the complexity, simplifying assumptions are made in an attempt to provide a reasonable explanation of the phenomena involved.

The majority of solid surfaces, when immersed in an aqueous solution, acquire an electrical surface charge. As a result, an electrical potential between the surface and the bulk fluid is created. There is an attraction between oppositely charged ions (counterions) in the fluid and the surface. At the same time, similarly charged ions (coions) are repelled away from the surface. The electric discontinuity created at the surface is generally known as the ''electrostatic double layer.'' Fig. 1 is a simplified illustration of the situation that applies to a spherical colloid particle immersed in water. The diffuse electrostatic double layer plays an important role in the interaction between the solid surface and microorganisms and colloids in suspension. Indeed, microorganisms in suspension have been called ''living colloids,'' the ''bridging'' may take the form of chemical bonding with the adsorbed macromolecules on the surface. The basic concepts of particle adhesion theory may be modified where microorganisms are the adhering particles. Mozes,[8] applying adhesion theory to the formation of biofilms, makes the following observations in relation to microbial cells in suspension:

1. It is not a homogeneous, rigid, smooth, and spherical ideal colloid particle.
2. It is not inert and in equilibrium with its environment.
3. It may be capable of independent movement.
4. It may respond physiologically to contact with a surface.

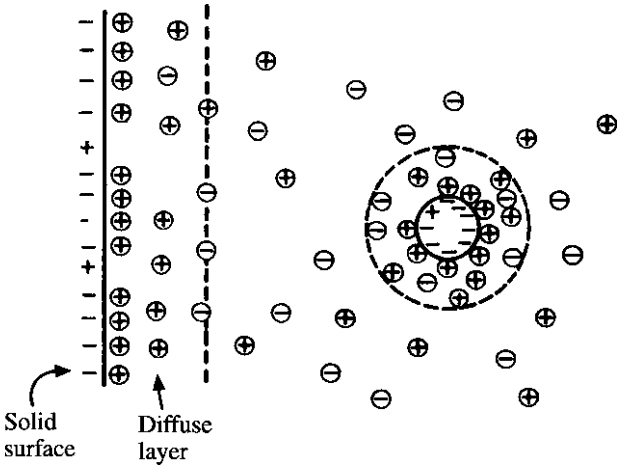


Fig. 1 An idealized representation of an electrostatic double layer and a solid colloidal particle. (From Ref.[2].)

Both reversible and irreversible attachments are possible. Microbial cells can be held in close proximity to a surface by the long-range forces, but still be capable of Brownian motion. Under the flow conditions, mild shear forces, brought about by flow and disturbances in the viscous sublayer, may also be present. These effects may restrict or prevent attachment. It is anticipated that this could be the situation if no conditioning molecules were residing on the surface and would explain the reason for the observed delay in biofilm formation under flowing conditions. Any adhesion of cells under these conditions could be said to be reversible.

Once there is a bonding between the conditioning molecules and the microbial cells, removal by shear forces becomes more difficult and the attachment can be regarded as irreversible. As the biofilm develops, however, shear forces caused by the fluid flow may be capable of removing cells from the surface. The extent of this removal is dependent on the quality of the biofilm and the bulk flow velocity. Furthermore, the establishment of irreversible attachment is dependent on the quality of the surface involved. Rough surfaces are known to facilitate the formation of biofilms. If the crevices on the surface are relatively large, a bacterium cell that is only 1 or 2 μm in size, could easily ''hide'' in a crevice and be unaffected by the shear forces. Under these circumstances, it is possible that there is a greater opportunity to form a bond with the conditioning molecules. Once these ''survivors'' are attached, the biofilm can begin to develop and irreversible attachment established. In the food industry, great care is taken to make sure that the surfaces of equipment in contact with food products are exceptionally smooth to reduce the risk of the formation of biofilms that could adversely affect food quality.

Nutrient Transport and Assimilation

After colonization of the surface, the microorganisms utilize the available nutrients to grow, multiply, and synthesize both intracellular products and extracellular polymeric substances that constitute the substance of the biofilm. Significant amounts of biofilm can be produced under ideal conditions, and even if planktonic cells are no longer present in the flowing water, the sessile cells already on the surface can provide the basis for biofilm development.[9]

The transport of nutrients from the bulk water to the developing biofilm depends on the concentration difference of nutrients between the bulk water and the surface of the biofilm. A concentration driving force is established by the removal of nutrients by the growing biofilm, assuming that the nutrient concentration in the bulk is maintained. As with the initial mass

transfer of cells to the surface, the boundary layers adjacent to the biofilm constitute a resistance to the transport of nutrients to the biofilm; the rate of growth may be limited by this resistance. The extent of the resistance will depend on the flow conditions outside the boundary layers. In broad terms, the higher the bulk velocity, the lower the resistance to mass transfer of nutrients.

As the biofilm develops, the nutrient availability to the bulk biofilm may become affected. The biofilm, despite its voids and channels, offers a further resistance to mass transfer. The cells within the biofilm consume nutrients that diffuse through the biofilm in response to the difference in concentration between nutrients at the biofilm surface and the cells attached to the conditioning layer. As a consequence, it is entirely possible that cells in the region of the solid surface are likely to become starved of nutrients. The properties of the biofilm may be different, therefore, in the layers where nutrient is available compared with the regions where there is little or no nutrient. For instance, the lack of oxygen may encourage anaerobic species to develop (some bacteria can exist as aerobes or anaerobes), with attendant changes to the quality of the biofilm.

Biofilm Removal

As is apparent from the foregoing discussion, the growth of the biofilm is a complex interaction between the flowing fluid and the physiology of the biofilm structure. Furthermore, the properties of the biofilm may change with time as it grows, resulting in the removal of part of the biofilm by the shear forces acting on the outer layers of the biofilm. The process is generally referred to as ''sloughing.''

THE STRUCTURE OF BIOFILMS

Concepts on the detailed structure of biofilms prior to the invention of the confocal scanning laser microscope were largely the results of intuition and speculation, because, in general, the biofilm had to be removed from the environment in which it was formed. Such a procedure could lead to the damage of the biofilm and loss of moisture that would change its appearance.

With the invention of newer methods of investigation, detailed examinations of fully hydrated biofilms, where water flows across the surface of the biofilm, could be made. The results of the observations revealed a heterogeneous structure consisting of cell clusters separated by interstitial voids and channels.[10] Real-time video imaging of biofilms growing on surfaces subject to fast flowing water revealed the presence of

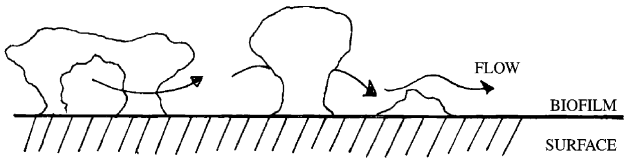


Fig. 2 Schematic concept of flow within a biofilm. (From Ref.[10].)

''streamers'' oscillating rapidly in the flow.[11] These streamers, which oscillate in response to the effects of turbulence in the flowing water, may assist the assimilation of nutrients by the destruction of the laminar sublayer around the colony. In other words, the resistance to the mass transfer of the nutrients is reduced.

The cell clusters can be composed of a single species that is likely to be the result of the initial colonization of the surface. It is possible to have groups of cells containing a mixture of various species, and this might be the result of synergy, i.e., an interactive association between two populations of microorganisms, not necessarily for survival, but for each group's benefit. Extracellular polysaccharide may be present around the clusters or less densely distributed in the spaces between the microcolonies. The result is a random array of channels through which water can flow. Fig. 2 illustrates the possible flow pattern between clusters of cells. Clearly, this water flow has implications for the delivery of nutrients to the cells in the lower regions of the biofilm.

Under high flow rates of fluid, when shear forces are strong, it is possible that the surviving cells lie so that the removal forces they experience are reduced to a minimum,[12] and the biofilm is less prone to sloughing.

THE DEVELOPMENT OF BIOFILMS

The discussion so far has dealt with scientific background to the formation of biofilms. It is of interest therefore to examine, as far as possible, the pattern of development in industrial process operations. The work described here is related to fresh water fouling by biofilms and carried out in a pilot scale laboratory apparatus, but it gives an indication of the effects of the principal operating variables on biofilm development with time. Biofilms in cooling water systems reduce the cooling efficiency because they are prone to occur on heat transfer surfaces, where, in general, conditions are conducive to development. The organisms involved are usually bacteria. Unless corrective action is taken, the biofilms so produced can cause serious operating problems with attendant higher operating costs.

The change in biofilm accumulation on a surface with time, under flowing water conditions, would be expected to follow the idealized curve in Fig. 3. Three regions can be seen in the diagram:

1. Initiation and growth, which is related to the laying down of the conditioning macromolecules on the surface.
2. After the colonization, there is a period of rapid growth, where the conditions, in terms of temperature, availability of nutrients, and their transport to the biofilm, are conducive to sustained growth.
3. The rate of growth gradually falls off, till a stable thickness of biofilm is reached, which might vary from a few micrometers to several millimeters, depending on the prevailing conditions.
4. The leveling off of the biofilm growth is attributed to a balance between growth and removal, depending on nutrient availability and temperature on the one hand, and shear forces caused by water flow on the other.

Under practical operating conditions, it will be seen that this ideal representation of biofilm development with time will be modified. Fig. 4 demonstrates that practical data roughly follow the idealized curve, but with considerable fluctuation of. The reason for this "saw tooth" appearance is attributed to the growth and partial sloughing of parts of the biofilm because of the shearing action of the flowing water at the surface of the biofilm. It is interesting to note that there is little effect on the development of the biofilm by the elimination of the planktonic bacteria in the flowing water. Once the surface has been colonized, development continues provided that a nutrient supply is maintained. An understanding of the effect of different variables on biofilm growth is essential for devising

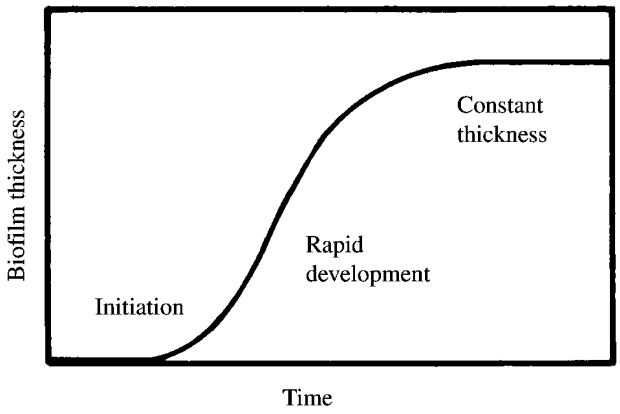


Fig. 3 Idealized concept of biofilm growth on a surface with time. (From Ref.[2].)

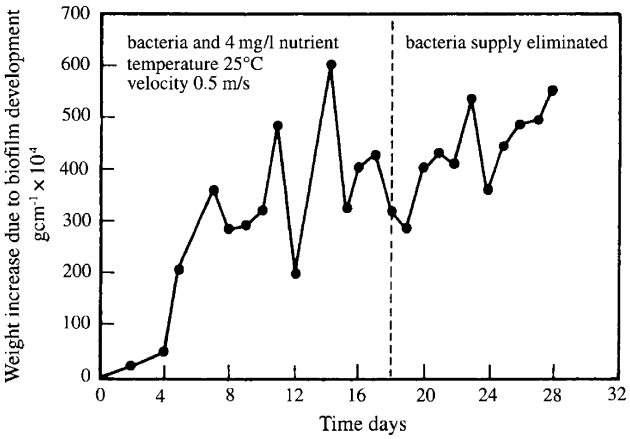


Fig. 4 Practical biofilm growth curve on the inside of an aluminum tube. (From Ref.[9].)

methods for biofilm control where biofilms pose an operational problem, in cooling water circuits, for instance, or in enhancing growth where a biofilm is the key to the effectiveness of the process, such as water purification.

MEASUREMENT OF BIOFILM ACCUMULATION

Very relevant to the representation of biofilm growth data is how to measure biofilm accumulation. There are various methods available, but a detailed discussion is outside the scope of this encyclopedia. For many years, direct weighing was employed, i.e., to determine the biofilm accumulation by weighing a tube or insert plate before and after contamination with microorganisms. The technique is very unsatisfactory for a number of rather obvious reasons. The removal of the test surface from the rest of the system for weighing may damage the biofilm, and the adventitious moisture associated with the biofilm is likely to give misleading information. Furthermore, the method does not lend itself to continuous tracking of the change in biofilm accumulation with time. Another method that has been used is to carry out a cell count on a unit area of test surface, but, apart from being labor intensive, this method also has the same problems of direct weighing. A number of methods have been developed to overcome the problem.[13] Some of the data reproduced here involve the use of infrared absorbance as a measure of biofilm thickness.[14]

THE AVAILABILITY OF NUTRIENTS

It is to be expected that the availability of nutrients is crucial to the development of biofilms. The effects of

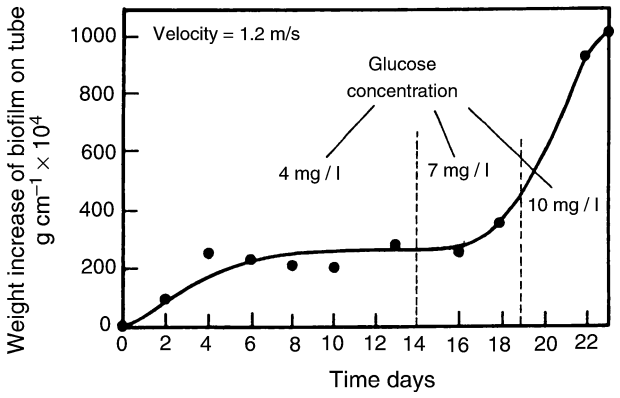


Fig. 5 The change in biofilm accumulation on the inside of a tube with three different nutrient concentrations expressed in terms of glucose concentration. (From Ref.[9].)

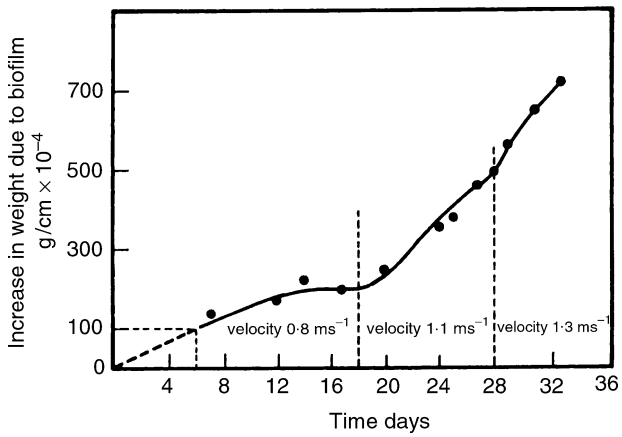


Fig. 7 The effect of velocity on biofilm growth at constant concentration. (From Ref.[9].)

the changes in nutrient concentration on biofilm growth are shown in Fig. 5. Fig. 6 demonstrates the importance of trace elements on biofilm development. Until trace elements were added to the system, little or no growth was evident.

THE EFFECT OF BULK WATER VELOCITY

The effects of changes in bulk water velocity are basically twofold. Higher velocities represent greater turbulence in the bulk flow and a reduction in the thickness of the boundary layers adjacent to the biofilm residing on the solid surface. As the velocity increases, the availability of nutrients, for a given concentration, to the biofilm increases because of the lower resistance to mass transfer of nutrients to the biofilm. It would be expected that this would be evident in the higher rate of biofilm growth. As the velocity increases, however, the attendant shear forces acting on the biofilm also increase. It would be expected that this would be manifest in a reduction of biofilm accumulation, till a new plateau is reached, i.e., the competing effects of increased growth because of greater nutrient availability and the removal of biofilm by the increased velocity. The result is apparent in Fig. 7.

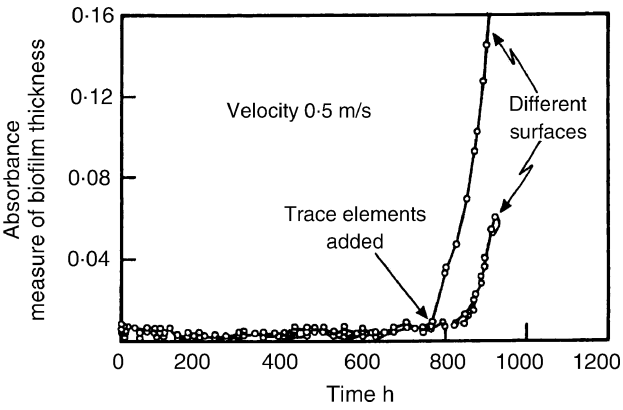


Fig. 6 The effect of adding trace elements to the nutrient on the growth of a biofilm on the inside of a glass tube, using infrared monitoring. (From Ref.[15].)

THE EFFECT OF TEMPERATURE

Microorganisms have an optimum growth temperature, when, provided that there are sufficient nutrients available, the growth is maximum. The optimum temperature is different for different species, on account of various metabolic characteristics. It is usually in the range of 20–50°C, with many in the range of 35–40°C. Fig. 8 shows the very pronounced effect of a relatively small temperature change on the development of a biofilm of *Escherichia coli*.[16]

THE EFFECT OF SURFACE

Mott and Bott[17] illustrated the effect of different materials on the accumulation of *Pseudomonas fluorescens* biofilms on the inside of tubes under identical operating conditions (see Fig. 9). The differences between the effects of the materials occur for two reasons: roughness and surface electrical properties. The quality of the surface, in terms of roughness, on which microorganisms attach, can affect the biofilm accumulation as discussed earlier. The effect of roughness is illustrated in Fig. 9 by the difference of biofilm accumulation between electropolished and "as received" 316 stainless steel. The rougher stainless steel is seen to be more hospitable to biofilm growth.

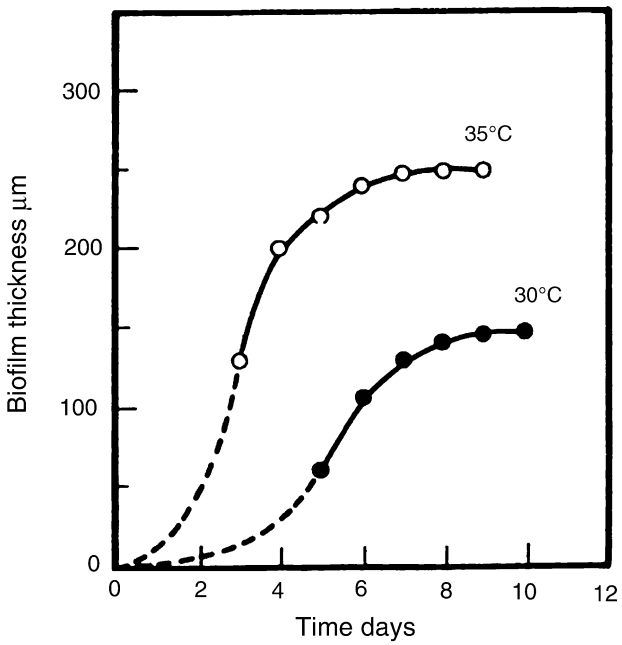


Fig. 8 The effect of temperature on the growth of a biofilm of *E. coli*. (From Ref.[16].)

In the light of these results, modification of the surface is a possible way of changing biofilm accumulation, i.e., to reduce roughness where the biofilm represents a nuisance or to use a rough surface where the biofilm performs a useful function.

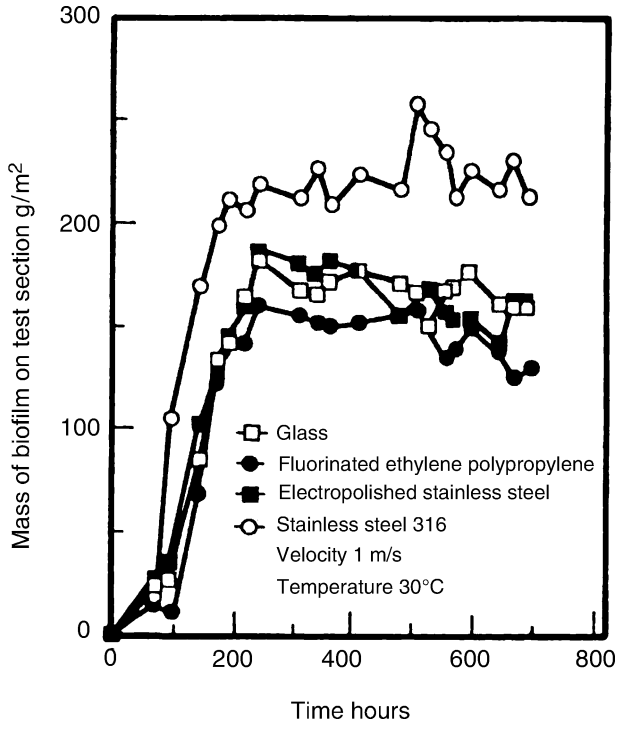


Fig. 9 Development of a biofilm of *P. fluorescens* on different materials. (From Ref.[7].)

CONTROL OF BIOFILMS

Reference in the discussion so far has been made to the need of controlling biofilm growth, a requirement that arises often. Where it is required to enhance growth, such as the treatment of wastewater, it is simply a question of ensuring that the operating conditions are optimum for the species present. There may be, of course, limited opportunity for this approach by virtue of the existing conditions of the process. It is unlikely that there would be much opportunity, for instance, to raise the temperature of the incoming stream because of the difficulties this would present and the costs involved.

Of greater significance is the need to control the growth of biofilms, as encountered in the use of naturally occurring water for cooling purposes. Some mention has been made in the preceding discussion, but it is useful to summarize the opportunities available. At the onset, it is important to stress that the tactic adopted will very much depend on the operating conditions and particularly on the quality of the water utilized. The obvious choice of method for controlling biofilm growth is to "dose" the cooling water with a biocide that will kill the microorganisms present in the system. For many years, this was the technique employed and the preferred biocide was chlorine, as it was very effective, available, and relatively cheap. As the cooling water is usually discharged back after use to the source from which it was obtained, e.g., a lake, river, or canal for fresh water systems, or the sea where this was more convenient, it has led to concern for the environment. Although chlorine is still used in many cooling systems, it is coming under increasingly tighter control to reduce the threat to the environment. As a consequence, there has been considerable investment in the search for a reliable alternative.

Perhaps the solution lies in the use of physical method, i.e., to avoid the use of chemicals altogether. The earlier discussion had indicated that velocity plays an important part in the development of biofilms, so that higher velocities might be used for biofilm control, as this would increase the removal forces. Although a realistic possibility, the technique is likely to be costly in terms of energy usage, because of the serious increase in pressure drop through the system that this would entail. In general terms, it has to be remembered that pressure drop increases as the square of velocity increases. A common technique in the power industry is to circulate sponge rubber balls with the cooling water through the steam condensers. It has to be said, however, that the opportunities for physical control have not been fully explored. Some physical methods for control that were described in a recent article are now being investigated.[18] They include the use of ultrasound and inserts in heat exchanger tubes, and the circulation of polymer fibers. The latter technique

would appear to show promise because the fibers reach all parts of the system in the same way as dissolved chemical biocides do, thereby keeping all parts of the system clean, provided that the fluid velocity is maintained. The other two techniques tend to give more localized control, in particular in heat exchangers, where the biofilm growth is likely to be most prevalent owing to the favorable temperature. An alternative physical strategy is to use modified surfaces to reduce the adherence of the microorganisms to the surface to facilitate removal by the shear forces. The extra cost of treatment of the surface such as electropolishing or ion implantation could be high. The alternative of coating the surfaces with, say, a polymer could invoke questions of its integrity over long periods of time.

In response to the environmental issues, the so-called ''environment friendly'' biocides have been or are being developed. The concept is that after a relatively short time, they decompose to innocuous breakdown products, some of which could be nutrients for biological activity that gives safe disposal. Although this suggests a contradiction in terms, there need not be any difficulties, provided the problem is recognized and handled properly. The cost of these biocides is relatively high, partly because the dose required is, in general, much higher than the traditional chemicals such as chlorine. For this reason, the dosing strategy requires careful attention.[19] The advantages of this approach include optimizing the chemical use and minimizing the maintenance costs. Some biocides that have been used for a number of years are naturally environment friendly and include the oxidizing agents, ozone and hydrogen peroxide, which breakdown to oxygen and water, respectively. A major difficulty with ozone is that it has to be generated as required, because it cannot be stored. Hydrogen peroxide is sometimes a constituent of proprietary biocide formulations.

CONCLUSIONS

Biofilms are common in natural environments and may include bacteria, algae, and fungi. Microorganisms prefer to reside on surfaces because the surface offers a certain degree of protection. But at the same time, they have a ready access to nutrients contained in the fluid passing over the surface. The important variables that affect the stability of the biofilm are temperature and the velocity of the fluid in contact with the biofilm. There is an optimum temperature and an optimum velocity to sustain the maximum biofilm thickness, provided that there is an adequate supply of nutrients. In processing, biocides can either be an aid or a disadvantage, for instance, as seen in wastewater treatment and fouling of heat exchangers in cooling water circuits, respectively. Where biofilms are an impediment to processing, control is implemented by chemical or physical means.

ARTICLE OF FURTHER INTEREST

Fluid Flow, p. 975.

REFERENCES

1. Hope, C.K.; Bott, T.R. Laboratory modelling of manganese biofiltration using biofilms of *Leptothrix discophora*. Water Res. **2004**, *38*, 1853–1861.
2. Bott, T.R. Fouling of heat exchangers. *26 Chemical Engineering Monographs*; Elsevier: Amsterdam, Holland, 1995.
3. Costerton, J.W.; Geesey, G.G.; Cheng, K.J. How bacteria stick. Sci. Am. **1978**, *238* (1), 86–96.
4. Fletcher, M. Bacterial metobolism in biofilms. In *Biofilms—Science and Technology*; Melo, L.F., Bott, T.R., Fletcher, M., Capdeville, B., Eds.; Kluwer Academic Publishers: Dordrecht, Holland, 1992; 113–124.
5. Leadbeater, B.S.C.; Callow, M.E. Formation, composition and physiology of algal biofilms. In *Biofilms—Science and Technology*; Melo, L.F., Bott, T.R., Fletcher, M., Capdeville, B., Eds.; Kluwer Academic Publishers: Dordrecht, Holland, 1992; 149–162.
6. Characklis, W.G. Microbial fouling – a process analysis. In *Fouling of Heat Transfer Equipment*; Somerscales, E.F.C., Knudsen, J.C., Eds.; Hemisphere Publishing: New York, U.S.A., 1981; 251–291.
7. Chamberlain, A.H.L. The role of adsorbed layers in bacterial adhesion. In *Biofilms—Science and Technology*; Melo, L.F., Bott, T.R., Fletcher, M., Capdevilee, B., Eds.; Kluwer Academic Publishers: Dordrecht, Holland, 1992; 59–67.
8. Mozes, N. The ways we study interfacial phenomena of living cells. In *Adhesion des Microorganisms aux Surfaces*; Bellon-Fontaine, M.W., Fourniat, J., Eds.; Chatenay-Malaby, France, June 2729, 1994; Lavosier TEC and DOC: Paris, 1995; 3–13.
9. Bott, T.R.; Miller, P.C. Mechanisms of biofilm formation on aluminium tubes. J. Chem. Technol. Biotechnol. **1983**, *33B*, 177–184.
10. Stoodley, P.; deBeer, D.; Lewandowski, Z. Liquid flow in biofilm systems. Appl. Environ. Microbiol. **1994**, *60*, 2711–2716.
11. Lewandowski, Z.; Stoodely, P. Flow induced vibrations, drag force and pressure drop in conduits covered with biofilm. Wat. Sci. Technol. **1995**, *32*, 19–26.
12. Santos, R.C.; Callow, M.E.; Bott, T.R. The structure of *Pseudomonas fluorescens* biofilms in

contact with flowing systems. Biofouling **1991**, *4*, 319–336.
13. Lens, P.; Moran, A.P.; Mahony, T.; Stoodley, P.; O'Flaherty, V. *Biofilms in Medicine, Industry and Environmental Biotechnology*; Part 3, Section 5; IWA Publishing: London, U.K., 2003; 441–470.
14. Bott, T.R. Monitoring biofouling using infrared absorbance. In *Biofilms in Medicine, Industry and Environmental Biotechnology*; Lens, P., Moran, A.P., Mahony, T., Stoodely, P., O'Flaherty, V., Eds.; IWA Publishing: London, U.K., 2003; 461–470.
15. Santos, R.C. Polymer coatings in relation to single and mixed population biofilms. Ph.D. Thesis, University of Birmingham, 1993.
16. Bott, T.R.; Pinheiro, M.M.V.P.S. Biological fouling—velocity and temperature effects. Can. J. Chem. Eng. **1977**, *55*, 473–474.
17. Mott, I.E.C.; Bott, T.R. The adhesion of biofilms to selected materials of construction for heat exchangers. Proceedings of the Ninth International Heat Transfer Conference, Jerusalem, Israel, 1991; 5, 21–26.
18. Bott, T.R. Potential physical methods for the control of biofouling in water systems. Trans. I. Chem. A: Chem. Eng. Res. Design **2002**, *79*, 484–490.
19. Grant, D.M.; Bott, T.R. Biocide dosing strategies for biofouling control. Heat Trans. Engineering **2005**, *24* (1), 44–50.

Biofuels and Bioenergy

Dinesh Gera
Fluent Incorporated, Morgantown, West Virginia, U.S.A.

INTRODUCTION

Biofuel is any gas, liquid, or solid fuel derived either from recently living organisms or from their metabolic by-products, including dedicated energy crops and trees, agricultural food and feed crop residues, aquatic plants, wood and wood residues, animal wastes, and other waste materials. It is a renewable energy, unlike other fossil fuel sources such as coal, petroleum, or nuclear energy. Biomass (any plant-derived organic matter) is a subset of biofuels, and it is used repeatedly in this entry. The term ''bioenergy'' refers to the production of energy (liquid, solid, and gaseous fuels) and heat using the biofuels or biomass.

Biomass is a very heterogeneous and chemically complex renewable resource. Understanding this natural variability and range of chemical compositions is essential for scientists and engineers conducting research and developing energy technologies using biomass resources. Biomass that is available for energy has a potential to produce an array of energy-related products, including liquid, solid, and gaseous fuels; electricity; heat; chemicals; and other materials.

Most scientists now agree that temperatures around the world are rising, and global warming may be occurring. There will be high societal costs if unchecked growth in atmospheric concentrations of greenhouse gases continues. Bioenergy and bio-based industrial feedstocks offer sound, economically friendly, and environmentally beneficial ways to reduce the pace at which CO_2 and other global warming gases accumulate in the atmosphere. The utility of bioenergy and its coproducts can play a major role in mitigation strategies for reducing greenhouse gases. Potential environmental benefits from biomass include offsetting these greenhouse gas emissions and sequestering carbon, improving water quality, and reducing soil erosion through the use of perennial cropping systems on marginal lands and by recovering wastes and capturing methane emissions.

The relative performance of biomass energy systems in reducing net greenhouse gas emissions depends on the sustainability of the sources of biomass feedstock, the energy requirements of the conversion systems, and the overall conversion efficiencies. Unlike fossil fuels, biomass production systems recapture emissions of carbon dioxide. It has been well documented that biomass energy and product systems have the potential to substantially reduce net greenhouse gas emissions. These vary greatly across biomass systems. A 1999 life cycle analysis indicated a 95% reduction in carbon dioxide emissions from a woody crop fired integrated gasification combined-cycle system relative to the average coal-fired power system. Direct-fired biomass systems, using residues that otherwise would have gone to landfills, generated even greater reductions in greenhouse gas emissions by avoiding methane production. The United States Department of Agriculture (USDA) and the Department of Energy (DOE) studies have shown that compared to gasoline, greenhouse gas emissions can be reduced on a per-gallon basis by 20–30% with the use of corn ethanol and 85–140% with that of cellulosic ethanol.[1]

This entry is organized into three major parts. The first identifies the biomass resources in the form of conventional forestry, agricultural crops and residue, and oil-bearing plants, among others. The second describes the conversion processes of bioresources into biofuels, and it is followed by the end product usage of biofuels in producing electricity in power plants.

BIORESOURCES

A variety of fuels can be made from biomass resources, including liquid fuels such as ethanol, methanol, biodiesel, Fischer–Tropsch diesel; gaseous fuels such as hydrogen and methane, and solid fuels such as switch grass, walnut shells, and sawdust, among others. Examples of biofuel include alcohol (from fermented sugar), black liquor from the paper manufacturing process, and soybean oil. Biofuel contains no petroleum, but it can be blended at any level with petroleum fuel to create a biofuel blend. It can be used in conventional heating equipment or diesel engines with no major modifications. Biofuel is simple to use, biodegradable, nontoxic, and essentially free of sulfur and aromatics. A typical blend of 20% biofuel with 80% of convention petroleum or fossil fuel has demonstrated significant environmental benefits. Biofuel is the only alternative fuel to have fully completed the health effects testing of the Clean Air Act. The use of biofuel in a conventional heating system or diesel engine results in a substantial reduction of unburned

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120018027

hydrocarbons, carbon monoxide, and particulate matter compared to the emissions from heating oil or diesel fuel. In addition, the exhaust emissions of sulfur oxides and sulfates (major components in acid rain) from biofuel are essentially eliminated compared to heating oil or diesel fuel.

Bioresources include any organic matter available on a renewable basis, including dedicated energy crops and trees, agricultural food and feed crops, agricultural crop wastes and residues, wood wastes and residues, aquatic plants, animal wastes, municipal wastes, and other waste materials.

Herbaceous energy crops: These are perennials that are harvested annually after taking 2–3 yr to reach full productivity. They include such grasses as switchgrass, miscanthus (also known as elephant grass or e-grass), bamboo, sweet sorghum, tall fescue, kochia, wheatgrass, and others.

Woody energy crops: Short-rotation woody crops are fast-growing hardwood trees harvested within 5–8 yr after planting. These include hybrid poplar, hybrid willow, silver maple, eastern cottonwood, green ash, black walnut, sweetgum, and sycamore.

Industrial crops: Industrial crops are developed and grown to produce specific industrial chemicals or materials. Examples include kenaf and straws for fiber, and castor for ricinoleic acid. New transgenic crops are being developed that produce the desired chemicals as part of the plant composition, requiring only extraction and purification of the product.

Agricultural crops: These feedstocks include the currently available commodity products such as cornstarch and corn oil, soybean oil and meal, wheat starch, other vegetable oils, and any newly developed component of future commodity crops. They generally yield sugars, oils, and extractives, although they also can be used to produce plastics and other chemicals and products.

Aquatic crops: A wide variety of aquatic biomass resources exist, such as algae, giant kelp, other seaweed, and marine microflora. Commercial examples include giant kelp extracts for thickeners and food additives, algal dyes, and novel biocatalysts for use in bioprocessing under extreme environments.

Agriculture crop residues: Agriculture crop residues include biomass (primarily stalks and leaves) that are not harvested or removed from the fields in commercial use. Examples include corn stover (stalks, leaves, husks, and cobs), wheat straw, and rice straw. With approximately 80 million acres of corn planted annually, corn stover is expected to become a major biomass resource for bioenergy applications.

Forestry residues: Forestry residues include biomass that is not harvested or removed from logging sites in commercial hardwood and softwood stands, as well as material resulting from forest management operations, such as precommercial thinnings and removal of dead and dying trees.

Municipal waste: Residential, commercial, and institutional postconsumer wastes contain a significant proportion of plant-derived organic material that constitutes a renewable energy resource. Waste paper, cardboard, wood waste, and yard wastes are examples of biomass resources in municipal wastes.

Biomass processing residues: All processing of biomass yields by-products and waste streams collectively called residues, which have significant energy potential. Residues are simple to use because they have already been collected. For example, processing of wood for products or pulp produces sawdust and a collection of bark, branches, and leaves/needles.

Animal wastes: Farms and animal processing operations create animal wastes that constitute a complex source of organic materials with environmental consequences. These wastes can be used to make many products, including energy.

Biofuels can be converted into all three natural forms, i.e., liquid, gas, and solid, from bioresources, using the biochemical, photobiological, and thermochemical processes described in the following section.

BIOMASS CONVERSION PROCESSES

Biochemical conversion processes: Enzymes and micro-organisms are frequently used as biocatalysts to convert biomass or biomass-derived compounds into desirable products. Cellulase and hemicellulase enzymes break down the carbohydrate fractions of biomass into five- and six-carbon sugars, a process known as hydrolysis. Yeast and bacteria ferment the sugars into products such as ethanol. Biotechnology advances are expected to lead to dramatic biochemical conversion improvements.

Photobiological conversion processes: Photobiological processes use the natural photosynthetic activity of organisms to produce biofuels directly from sunlight. For example, the photosynthetic activities of bacteria and green algae have been used to produce hydrogen from water and sunlight.

Thermochemical conversion processes: Heat energy and chemical catalysts are used to break down biomass into intermediate compounds or products. In gasification, biomass is heated in an oxygen-starved environment to produce a gas composed primarily of hydrogen and carbon monoxide. In pyrolysis, biomass is exposed to high temperatures in the absence of air, causing it to decompose. Solvents, acids, and bases can be used to fractionate biomass into an array of products including sugars, cellulosic fibers, and lignin.

Biofuels exist in all three natural forms, that is, gas, liquid, and solid:

1. Gas fuel
    a. Methane—Methane can be produced by the natural decay of garbage dumps over time. Additionally, biomass can be gasified to produce a synthesis gas composed primarily of hydrogen and carbon monoxide, also called syngas or biosyngas. Hydrogen can be recovered from this syngas, or syngas can be catalytically converted to methanol. It can also be converted using the Fischer–Tropsch catalyst into a liquid stream with properties similar to diesel fuel, called Fischer–Tropsch diesel. However, all of these fuels also can be produced from natural gas or coal using a similar process.
2. Liquid biofuels
    a. Bioalcohols are ethanol and methanol, when they are not produced from petroleum. A significant amount of ethanol is produced from sugar beets and corn using the fermentation process. Ethanol is the most widely used biofuel today, with a current capacity of 1.8 billion gallons per year based on starch crops such as corn. Ethanol produced from cellulosic biomass is currently the subject of extensive research and development, and demonstration efforts. It is currently being used as an automotive fuel and a gasoline additive.
    b. Straight vegetable oil (SVO) is a waste product of the food service industry, also called fryer grease. It can be used to run a conventional diesel engine if the oil is clean and heated to the appropriate temperature before being injected into the engine. Using SVO as a fuel greatly lowers health and environmental hazards caused by using petroleum fuels while boosting energy security.
    c. Biodiesel is produced through a process in which organically derived oils are combined with alcohol (ethanol or methanol) in the presence of a catalyst to form ethyl or methyl ester using transesterification. The biomass-derived ethyl or methyl esters can be blended with conventional diesel fuel or used as a neat fuel (100% biodiesel). Biodiesel can be made from soybean or canola (rapeseed) oils, animal fats, waste vegetable oils, or microalgae oils. It can be used in unaltered diesel engines. Biodiesel can be mixed with petroleum diesel. Vehicle performance on biodiesel is almost similar to vehicle performance on conventional petroleum diesel.
3. Solid fuel
    a. Biomass, such as switch grass, walnut shells, and sawdust, is currently used as the cofiring fuel in industrial boilers to reduce the nitrogen oxide (NO_x) emissions.

GLOBAL BENEFITS AND IMPACTS

A 1999 biofuel life cycle study, jointly sponsored by the USDOE and the USDA concluded that biofuel reduces net CO_2 emissions by 78% compared to petroleum fuels from biofuel's closed carbon cycle.[1] The CO_2 released into the atmosphere when biofuel is burned is recycled by growing plants, which are later processed into fuel.

Scientific research confirms that biofuel has a less harmful effect on human health than petroleum fuel. Biofuel emissions have decreased levels of polycyclic aromatic hydrocarbons (PAHs) and nitrited PAH compounds (nPAH), which have been identified as cancer causing compounds. Test results indicate that PAH compounds were reduced by 50–85%. Targeted nPAH compounds were reduced by 90%. Biofuel is nontoxic and biodegradable. In addition, the flash point (the lowest temperature at which it can form an ignitable mix with air) is 300°F, well above petroleum fuel's flashpoint of 125°F.

Several developed countries have introduced policies encouraging use of biofuels made from grain, vegetable oil, or biomass to replace part of their fossil fuel use in transport. These initiatives generally have at

least three goals: 1) to prevent environmental degradation by using cleaner fuel; 2) to reduce dependence on imported, finite fossil fuel supplies by partially replacing them with renewable, possibly domestic, sources; and 3) to provide a new demand for crops to support producer incomes and rural economies.

The use of biomass for power could have impacts on local air pollution. As an example, direct combustion and cofiring of biomass offers improvements in conventional air pollutants, NO_x, sulfur dioxide (SO_2), particulate matter (PM_{10} and $PM_{2.5}$), carbon monoxide (CO), and volatile organic chemicals (VOC) relative to direct-firing of coal. It follows, then, that certain emissions from advanced coal energy systems, which are now under development by the USDOE, also may be moderated by biomass cofiring or cogasification. The USDOE is conducting research on biomass gasification systems that most likely would reduce the overall emissions relative to most conventional fossil systems.

Because the impact of conventional pollutants is primarily local or regional, facility location is also a significant environmental consideration. Utilization of bio-based chemicals may help to reduce risks to human health from environmental releases and workplace exposure to toxic chemicals, particularly those derived from petroleum-based feedstocks. Substitution of bio-based products for petroleum-based end products has the potential to reduce pollution from virtually all stages of production, from extraction of the raw material to final product manufacturing and product disposal. Substituting bio-based products for inorganic-based products, e.g., wood for steel, biocement for cement, and cotton insulation (formaldehyde-free) for fiberglass, can have even more substantial effects on reducing greenhouse gas emissions, as the inorganic-based products are extremely energy intensive to produce. Substituting a biochemical for a petrochemical that has different properties can reduce the pollution generated by the production of the petrochemical, and also may reduce the environmental impact associated with using the chemical in manufacturing final consumer products. Finally, even substituting a biochemical for an identical petrochemical can reduce the upstream impacts associated with the extraction of the material.

Using the advanced bioethanol technology available, it is possible to produce ethanol from any cellulose/hemicellulose material, which means any plant or plant-derived material. Many of these materials are not just underutilized and inexpensive, but also create disposal problems. For example, rice straw and wheat straw are often burned in the field, a practice that is becoming limited by air pollution concerns. Also, much of the material now going into landfills is cellulose/hemicellulose material and could be used for bioethanol production. Technology is now available to convert the municipal solid waste to ethanol. Wastes from many paper mills, food processing, and other industries also may be converted to bioethanol.

Another possible bioethanol feedstock is corn stover (corn stalks and leaves). Because of the stover's large volume and proximity to current ethanol production facilities and because it is already there for the taking, biofuels program analysts expect stover to be one of the primary feedstocks for advanced bioethanol production. Currently, stover is left in the fields in some cases and plowed under in others. Appropriate levels of stover harvesting will need to be carefully determined to avoid losing stover's value for erosion control and soil enrichment. In many areas though—particularly in northern areas where springtime soil warming is of concern—stover is routinely plowed under—mostly to get rid of it. Harvesting part of that stover might allow farmers to switch to a no-till operation, which soil scientists recognize to be far better from a soil erosion and fertilizer use perspective.

The total bioenergy potential of the base year, 1990, was estimated at 225 EJ (1 EJ = 1 exajoule = 10^{18} J) or 5.4 billion (10^9) tons of oil equivalent (Gtoe).[2] For comparison, the actual use of bioenergy in 1990 was 46 EJ or 1.1 Gtoe. By the year 2050, this potential was estimated to have grown to between 370 and 450 EJ (8.8 and 10.8 Gtoe). The slowest growth is expected to occur in the ''crop residues'' category (because of increasing the harvesting index). To put the estimated totals into perspective, the estimated global energy value from photosynthesis is 4000 EJ.[2]

SOME ISSUES IN BIOMASS HANDLING

Material handling, collection logistics, and infrastructure are important aspects of the biomass resource supply chain.

Biomass material handling: Materials handling systems for biomass constitute a significant portion of the capital investment and operating costs of a bioenergy conversion facility. Requirements depend on the type of biomass to be processed as well as the feedstock preparation requirements of the conversion technology. Biomass storage, handling, conveying, size reduction, cleaning, drying, and feeding equipment, and systems require special attention. Biofuels can be stored in existing fuel tanks. Biofuel is completely compatible with petroleum fuels.

Biomass collection logistics and infrastructure: Harvesting biomass crops, collecting biomass residues, and storing and transporting

biomass resources are critical elements in the biomass resource supply chain.

Biomass feedstock production shares many of the potential environmental effects associated with other agricultural systems, including soil erosion, fertilizer, and pesticide runoff. However, biomass production systems using perennial crops, such as trees and grasses, that are being developed by the Agricultural Research Service, the Forest Service, the Oak Ridge National Laboratory, land grant universities, and others will lower overall chemical use and reduce soil erodable from rates associated with conventional monoculture crop production. Positive environmental effects have been documented where biomass (perennial crop) production replaces conventional crop production on marginal, highly erodable lands. Additional research is needed to optimize the management systems for a wide variety of soil and climate conditions. Utilization of biomass products such as compost also can serve as a component of an overall strategy to reduce, reuse, and recycle solid waste. For example, promoting composting is entirely consistent with efforts to enhance recycling. In providing incentives to promote biomass energy from waste, it will be important to understand and mitigate potential instances where diverting waste for energy might have impacts on programs and incentives to promote waste reduction and recycling. It is important to keep in mind that the established, national hierarchy for materials utilization places reuse and recycling above the use of materials for energy recovery. Improvements in handling and management of solid waste, such as the capture and use of landfill gas, can reduce emissions of methane, a powerful greenhouse gas, as well as other pollutants.

THERMOCHEMICAL PROPERTIES OF BIOMASS FEEDSTOCKS AND FUEL

Biomass feedstocks and fuels exhibit a wide range of physical, chemical, and agricultural/process engineering properties. Despite their wide range of possible sources, biomass feedstocks are remarkably uniform in many of their fuel properties (see Table 1), compared with competing feedstocks such as coal or petroleum. For example, there are many kinds of coals whose gross heating value ranges from 20 to 30 MJ/kg (million joules per kilogram; 8600–12,900 Btu/lb). However, nearly all kinds of biomass feedstocks destined for combustion fall in the range 15–19 MJ/Kg (6450–8200 Btu/lb). For most agricultural residues, the heating values are even more uniform—about 15–17 MJ/Kg (6450–7300 Btu/lb); the values for most woody materials are 18–19 MJ/Kg (7750–8200 Btu/lb). Moisture content is probably the most important determinant of heating value. Air-dried biomass typically has about 15–20% moisture, whereas the moisture content for oven-dried biomass is around 0%. Moisture content is also an important characteristic of coals, varying in the range of 2–30%. However, the bulk density (and hence energy density) of most biomass feedstocks is generally low, even after densification—between about 10% and 40% of the bulk density of most fossil fuels—although liquid biofuels have comparable bulk densities.

Table 1 Thermochemical properties of various bioenergy feedstocks and fuels

		Heating value (MJ/kg)	Ash (%)	Sulfur (%)	Potassium (%)	Ash melting temperature (°C)
Bioenergy feedstocks	Corn stover	17.6	5.6			
	Sugarcane bagasse	18.1	3.2–5.5	0.1–0.15	0.73–0.97	
	Hardwood	20.5	0.45	0.009	0.04	900
	Softwood	19.6	0.3	0.01		
	Hybrid poplar	19.0	0.5–1.5	0.03	0.3	1350
	Bamboo	18.5–19.4	0.8–2.5	0.03–0.05	0.15–0.5	
	Switch grass	18.3	4.5–5.8	0.12		1016
	Miscanthus	17.1–19.4	1.5–4.5	0.1	0.37–1.12	1090
Liquid biofuels	Bioethanol	28		<0.01		
	Biodiesel	40	<0.02	<0.05	<0.0001	
	Black liquor	11–13	40–45	3–8	0.1–8	
Fossil fuels	Coal (low rank; lignite/sub-bituminous	15–19	5–20	1.0–3.0	0.02–0.3	~1300
	Coal (high rank; anthracite/bituminous	27–30	1–10	0.5–1.5	0.06–0.15	~1300
	Oil	42–45	0.5–1.5	0.2–1.2		

(From Refs.[3,4].)

Most biomass materials are easier to gasify than coal, because they are more reactive, with higher ignition stability. This characteristic also makes them easier to process thermochemically into higher-value fuels such as methanol or hydrogen. Ash content is typically lower than for most coals, and sulfur content is much lower than for many fossil fuels. Unlike coal ash, which may contain toxic metals and other trace contaminants, biomass ash may be used as a soil amendment to help replenish the nutrients removed by harvest. A few biomass feedstocks stand out for their peculiar properties, such as high silicon or alkali metal contents. These feedstocks may require special precautions for harvesting, processing, and combustion equipment. Note also that mineral content can vary as a function of soil type and the timing of feedstock harvest. In contrast to their fairly uniform physical properties, biomass fuels are rather heterogeneous with respect to their chemical elemental composition.

Among the liquid biomass fuels, biodiesel (vegetable oil ester) is noteworthy for its similarity to petroleum-derived diesel fuel, apart from its negligible sulfur and ash content. Bioethanol has only about 70% of the heating value of petroleum distillates such as gasoline, but its sulfur and ash contents are also very low. Both of these liquid fuels have lower vapor pressure and flammability than their petroleum-based competitors—an advantage in some cases (e.g., use in confined spaces such as mines) but a disadvantage in others (e.g., engine starting at cold temperatures).

It should be mentioned here that almost all fuels suitable for traditional transport vehicles are compounds containing predominantly carbon and hydrogen. In the case of fossil derived fuels, the only other constituent elements, such as nitrogen and sulfur, are generally regarded as undesirable contaminants. Compounds such as tetraethyl lead are typically added to modify fuel properties so as to reduce the tendency of the fuel to ''knock'', while others, such as nitromethane, are added to specialty fuels to improve power output. However, biofuels differ chemically from fossil-based fuels in that they usually contain oxygen in addition to carbon and hydrogen. Like fossil fuels, they also may contain other elements, notably nitrogen, which is an undesirable impurity. Table 2 compares the approximate compositions and properties for selected transport fossil fuels and biofuels.

Biofuels are highly volatile and have a relatively high hydrogen (H) content. The portion of the volatiles in biofuels is around 70–80%. The carbon content and heat values are low compared to fossil fuels. Biofuels, including black liquor, are renewable fuels, which do not increase the carbon dioxide (CO_2) burden globally. The age of biofuels ranges from a few months to several thousands of years. Different energy plants and annual crops, which are grown on fields, are the most short-lived biofuels, with their age varying from a few months to some years. A characteristic feature of solid biofuels is their low sulfur (S) and sodium (Na) contents, but relatively high chlorine (Cl) and potassium (K) contents. The net heating value (LHV) of most biofuels is in the range of 6–10 MJ/kg in as fired conditions. This range is only about one-third of the heating value of coal.

The relatively low sulfur content in most solid biofuels may introduce corrosion problems in the superheaters. The inhibiting tendency of sulfur to limit superheater corrosion has already been recognized in the 1970s. It is generally known that when the sulfur to chlorine molar ratio (S/Cl) in the fuel is higher than 2.0, corrosion is diminished significantly. The best way to prevent molten phase corrosion is to keep the superheater metal temperature below the first melting temperature of deposits, in practice below 500°C when firing biofuels.

Table 2 Approximate compositions and properties for selected transport fossil fuels and biofuels

Fuel	Approximate average formula	Average molecular weight	Energy density (MJ/L)	CO_2 emissions (G/MJ)
Natural gas	$\sim CH_{3.85}$	18.2	38.2 MJ/m^3	51.3
LNG	$\sim CH_{3.85}$	18.2	25.0	51.3
CNG	$\sim CH_{3.85}$	18.2	38.2 MJ/m^3	51.3
LPG	$\sim C_3H_{7.8}$	49	25.7	60.2
Gasoline	$\sim C_{5.4}H_{10.7}$	80	35.2	65.8
Automotive diesel	$\sim C_{15.2}H_{22.2}$	212	38.6	65.8
Methanol	CH_3OH	32.04	15.8	60.8
Ethanol	C_2H_5OH	46.07	23.4	64.3
RME biodiesel	$\sim C_{13}H_{29}O$	201	33.3	85.0

(From Ref.[5].)

Biofuels can be combusted with various technologies, but only a few of them have gained a well-established position. The moisture content of biofuels is relatively high (compared to fossil fuels), which places certain requirements on the combustion systems designated for these fuels. In this context, the following technologies are available: 1) grate-firing technology, which is the oldest but still widely used technology for bark and for straw combustion; 2) bubbling fluidized bed technology, which is the most suitable for high moisture fuels; and 3) pulverized fuel injection technology in tangentially fired (t-fired) boilers. The first two technologies are getting obsolete; hence, a combustion example with the t-fired industrial boiler is demonstrated in the following section.

END PRODUCT USAGE OF BIOFUEL

Cofiring Biofuel with Coal in an Industrial Boiler

Because coal-fired boilers are a significant source of power generation in the United States and abroad, cofiring biomass at levels of 2–15% (heat basis) could provide a significant increase in bioenergy utilization while taking advantage of the existing utility infrastructure. Cofiring biomass with coal results in NO_x reduction and helps in reducing the unburned carbon in ash. Some utility companies have had success when cofiring 6 mm (0.25 in.) topsize biomass at full load.[6] In such situations, the computational fluid dynamics (CFD) modeling can provide insight into the behavior of the 6 mm particles as opposed to the smaller sizes. For example, Table 3 illustrates the complex behavior when cofiring a dry switchgrass in a 150 MWe t-fired pulverized coal utility boiler with four burner levels at 13.5 m, 12.0 m, 10.5 mm, and 9.0 m from the ground (A, B, C, D, where D is the lowest level near the bottom ash hopper, and A is the uppermost level) as shown in Fig. 1. Using an eastern bituminous low-sulfur coal in this CFD simulation, switchgrass is evenly distributed among the burners at a total 10% energy basis cofiring level, using a broad size distribution of about 5% plus 6 mm, 14% plus 4 mm, 37% plus 2 mm, 60% plus 1 mm, and 23% minus 0.5 mm with a mean particle size of 2 mm. Table 1 presents the fate of biomass particles—whether particles distribute to the fly ash or bottom ash, their effective biomass particle residence time (from burner injection level to either the convective section entrance or the bottom ash hopper), and combustion efficiency (CE)—as viewed from one corner of the boiler.

In this CFD simulation, pulverized coal and switchgrass achieve an average residence time of over 3 and 4 sec, respectively, with average CEs of 99.9% and 99.3%, respectively. Although switchgrass particles are an order of magnitude larger than pulverized coal, they achieve high CEs because of their high volatile content along with residence time enhancements. Table 1 shows that the smallest switchgrass particles behave in an expected fashion, with longer residence times observed for particles injected in the lower furnace, with essentially complete burnout as particles enter

Table 3 Computational fluid dynamics biomass particle size impacts—residence time, CE, and fly ash/bottom ash partitioning for 10% switchgrass cofiring at full load in a four-level burner 150 MWe t-fired boiler

	Switchgrass					
Burner level	**0.5 mm**	**1 mm**	**3 mm**	**4 mm**	**5 mm**	**6 mm**
A Upper	2.1 sec No sparklers 99.9% CE All fly ash	1.5 sec Few sparklers 99.9% CE All fly ash	6.4 sec Sparklers 97% CE All fly ash	8.1 sec Sparklers 96.3% CE 90% fly ash	3.2 sec Sparklers 95.3% CE Bottom ash	2.5 sec Sparklers 95.4% CE Bottom ash
B	2.6 sec No sparklers 99.9% CE All fly ash	3.4 sec Few sparklers 99.9% CE All fly ash	3.4 sec Sparklers 97.2% CE 6% fly ash	2.8 sec Sparklers 95.5% CE Bottom ash	2.3 sec Sparklers 95.2% CE Bottom ash	1.9 sec Sparklers 95.2% CE Bottom ash
C	3.1 sec No sparklers 99.9% CE All fly ash	2.1 sec Few sparklers 99.9% CE All fly ash	2.8 sec Sparklers 95.7% CE Bottom ash	2.2 sec Sparklers 95.3% CE Bottom ash	1.9 sec Sparklers 95.1% CE Bottom ash	1.7 sec Sparklers 95.1% CE Bottom ash
D Lower	7.2 sec No sparklers 99.9% CE 39% fly ash	7.6 sec Few sparklers 99.2% CE 28% fly ash	1.4 sec Sparklers 95.2% CE Bottom ash	1.2 sec Sparklers 95.1% CE Bottom ash	1.2 sec Sparklers 95% CE Bottom ash	1.2 sec Sparklers 95.1% CE Bottom ash

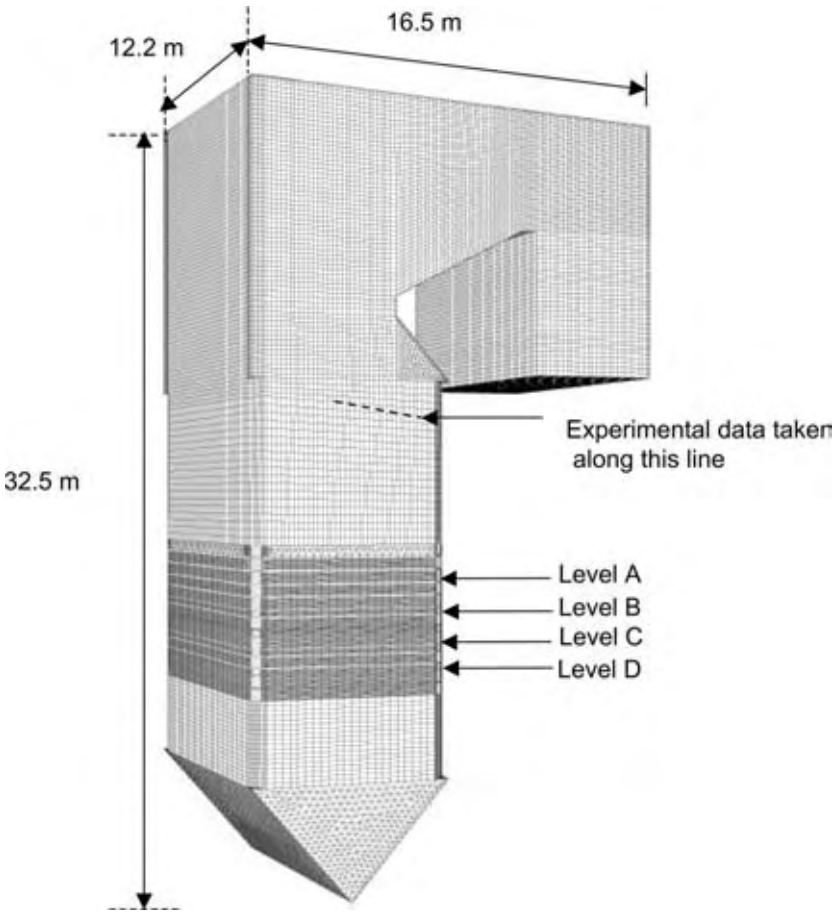


Fig. 1 Schematic of an industrial t-fired (150 MW_t) boiler.[6]

the convective pass and report to fly ash. However, the larger switchgrass particles behave quite differently as the relative contributions of gravity, buoyancy, and drag forces alter particle trajectories and effective residence times inside the turbulent flow field of the t-fired boiler. This behavior can be seen by the presence of still-burning sparklers entering the convective pass for intermediate particle sizes, and at larger sizes, still-burning sparklers that simply drop into the bottom ash. While the overall CE of 99.3% for switchgrass is very good, the presence of still-burning sparklers could be an issue from the standpoint of boiler operations and bottom ash handling.

CONCLUSIONS

In response to now incontrovertible evidence of human-induced global warming caused by the emission of atmospheric warming gases, renewable bioenergy generation is the potential vehicle to offset the greenhouse gases in the future. It is anticipated that the new electrical capacity from future power plants may come from energy crops and forestry residue combustion. Interestingly, combustion of biofuels releases CO_2, but because biofuels are made from plants that absorb CO_2 from the atmosphere, release is recaptured when new biomass is grown to produce more biofuels. Depending on how much fossil energy is used to grow and process the biomass feedstock, substantially reduced net greenhouse gas emissions result. Making biodiesel from soybeans reduces net emissions nearly 80%. The combination of reducing both gasoline use and fossil electrical production can mean a greater than 100% net greenhouse gas emission reduction and less dependence on foreign oil. As with any fossil-fuel combustion, no matter how much is spent on improved technology and emission controls, gasoline and diesel use in motor vehicles will, by definition, contribute to global climate change. To whatever extent biofuels are used instead, global warming will be reduced. Biofuels are highly volatile and hence can be used in big particle sizes as a reburning fuel. Cofiring biomass with coal in the existing power plants can significantly reduce the NO_x emissions, and hence there will be less fear of acid rain.

REFERENCES

1. Wang, M.; Saricks, C.; Santini, D. Effects of fuel ethanol use on fuel-cycle energy and greenhouse gas emissions. Argonne National Laboratory, ANL/ESD-38, 1999; http://www.transportation.anl.gov/pdfs/TA/58.pdf (accessed Nov 16, 2003).
2. Guenther, F.; Schrattenholzer, L. Global bioenergy potentials through 2050. Biomass Bioenergy **2001**, *20*, 151–159.
3. Scurlock, J. Bioenergy feedstock characteristics. Oak Ridge National Laboratory, Bioenergy Feedstock Development Programs, P.O. Box 2008, Oak Ridge, TN 37831; http://bioenergy.ornl.gov/papers/misc/biochar_factsheet.html (Netherlands accessed Nov 16, 2003).
4. PHYLLIS—database on composition of biomass and waste. Energy Research Centre of the Netherlands; http://www.ecn.nl/phyllis/ (accessed Nov 16, 2003).
5. Calais, P.; Sims, R.S. A comparison of life-cycle emissions of liquid biofuels and liquid and gaseous fossil fuels in the transport sector, Solar 2000 Conference, Dec 4, 2000; Australian, New Zealand Solar Energy Society: Brisbane (available on CD ROM).
6. Gera, D.; Freeman, M.; Mathur, M.P.; Robinson, A. Effect of large aspect ratio of biomass particles on carbon burnout in a utility boiler. Energy Fuels **2002**, *16*, 1523–1532.

Bioinformatics and Modeling Biological Systems

B

Shwetal Patel
Jeremy S. Edwards
Department of Chemical Engineering, University of Delaware, Newark, Delaware, U.S.A.

INTRODUCTION

Bioinformatics is the science of accumulating, analyzing, storing, assimilating, and transmitting large quantities of biological data. The completion of genome sequencing projects for a large number of organisms has kindled a quest to understand the complex programs encoded in the sequences. Because these programs are executed through coded proteins, a tremendous amount of effort is now directed toward identifying and characterizing all the proteins, their interactions, and the genetic structure defining them. Consequently, there is a great demand for computational tools to efficiently analyze and assimilate the large amount of biological data into knowledge. Particularly important are computational tools that utilize genomic, transcriptomic, proteomic, and metabolomic data to reconstruct biological pathways and networks so as to facilitate the understanding of their function and interaction.

The availability of complete genomes of microbes has revealed their remarkable industrial potential. Microbial genomic data are paving the way for novel and improved vaccines and therapeutics, better-tasting and healthier foods and beverages, stronger biodefense, efficient bioprocess alternatives, and a cleaner environment (Table 1). Bioinformatics tools will greatly facilitate a holistic understanding of the functioning of these microbes and enable their manipulation to perform desired tasks with unprecedented precision, flexibility, and efficiency. Therefore, the recent advances in bioinformatics have created a stir in the chemical process industry. The purpose of this entry is to provide an overview of bioinformatics tools and their applications toward understanding metabolic networks, which are likely to have a major impact on the chemical process industry.

Genome research is yielding enormous amounts of information at a rate that continues to increase. The nature of this information varies from static (i.e., sequences/structures of DNA and protein molecules) to dynamic (i.e., gene expression patterns and proteomic data). A large number of public databases now house these data. In the following section, the nature of biological information and some public databases for storing biological data are described. A major challenge is the computational integration of highly diverse data types, ranging from qualitative information such as gene function and protein modification sites to more quantitative information such as gene expression patterns and flux through metabolic networks.[8]

Pathway modeling, the computational representation of genetic and metabolic pathways, effectively builds on existing tools that analyze raw sequence data. It has been extensively utilized to understand the mechanism underlying the complex behavior of biological networks and even to design synthetic networks with desired properties. One such technique, namely, metabolic flux balance analysis, has been successfully applied to define the capabilities of biological systems on a genome scale and to develop an understanding of physiological function in microbes. Because bioinformatics applied to metabolic flux balance analysis shows great promise in revolutionizing the chemical process industry, an entire section is devoted to the basic principles underlying this technique and the methods for its implementation.

BIOLOGICAL INFORMATION

The sequencing of the genomes of numerous organisms has created a large amount of primary nucleotide sequence data that are stored in numerous public databases. Recent efforts to understand the function of gene products (proteins) and their complex interactions have resulted in a steadily increasing amount of biological information. As a result, there are now databases for almost all biological functional elements—from genes and gene expression to protein structure for many common organisms (Table 2). The management of this biological information has been possible because of the tremendous advances in computing capabilities in recent times. In conjunction with the development of modern database technology, the World Wide Web has become the natural medium for managing and distributing genomic data. The database at the National Center for Biotechnology Information (NCBI: http://www.ncbi.nih.gov) is the largest repository for molecular biology information. In addition to biological data, the web site has a large software collection for analyzing nucleotide and protein sequences.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016148

Table 1 Examples of industries that will benefit significantly from advances in bioinformatics

Industry	Microbes	Reference
Health		
Genome sequence helps understand how the species achieves virulence, and identify genes and potential vaccines	*Neisseria meningitis* *Streptococcus agalactiae*	[1] [2]
Food		
Genome sequence enables their engineering to produce better-tasting and safer foods such as cheddar cheese and cabernet sauvignon	Lactic acid bacteria	[3]
Biodefense		
Knowledge of the complete genome sequence can help develop rapid detection techniques of potential bioterror agents causing deadly diseases	*Brucella suis*	[4]
Chemicals		
Genomic sequence helps in metabolic engineering to produce expensive chemicals by altering natural pathways	*Escherichia coli*	[5]
Environment		
Analysis of genomic sequence provides an understanding of physiological function and the ability to engineer bacteria to clean up hazardous waste and the environment	*Deinococcus radiodurans* *Rhadocyclus tenius*	[6] [7]

A major task now is to develop computational tools to analyze these data in the databases so as to create knowledge. There are several ongoing programs to understand the molecular machines and their molecular level controls for both individual microbes and communities of microbes sufficiently so that we can engineer them to address national needs. The computational tools required to achieve this end range from sequence analysis algorithms (comparative genomics) to statistical mechanical methods for describing macromolecular motions, structures (structural bioinformatics), and interactions (docking). A final goal is to integrate the output from these diverse tools into comprehensive computational models to perform mechanistic-based cellular simulations. Toward this goal, new technologies are required to integrate the resulting data and build multidimensional models at the cell and cell community levels that reflect the results of functional genomics studies (Fig. 1). In what follows, a relatively well-developed technique for modeling and analyzing biological systems, namely, pathway modeling, is described.

BIOLOGICAL PATHWAY MODELING

Biological systems function by transmitting and processing information in genetic, protein, and metabolic networks. Therefore, a large chunk of postgenomic research is devoted to understanding the complex interactions that are responsible for the functioning of biological networks.[9] There are two principal approaches to modeling biological systems. Kinetic modeling attempts to derive the properties of enzymatic pathways from knowledge of the kinetic parameters associated with the activity of each component and the order in which those components act.[10] Steady-state or static modeling searches for global properties of networks based on generic properties of the system such as stoichiometry and Boolean logic.[11]

Kinetic Modeling

In the first approach, a set of rate expressions are derived to describe the temporal variation of the species as a function of the activities of the binding and catalytic events (for example, using detailed mass action laws or Michaelis–Menten kinetics). This approach has been extensively used in areas such as metabolic control analysis.[12] Early studies addressing the connection between genotype and phenotype were undertaken using this technique. In particular, the phenomenon of dominance because of near-neutrality of gene activity was explained; a dominant phenotype is not affected by minor perturbations in the genotype.[12] According to the analysis, if a microbe has evolved to optimize its growth rate, then metabolic flux should be as high as possible, and the activity of each enzyme (which is a function of its level of expression as well as catalytic rate constant set by the protein structure) will take on a value that places it close to the point at which the maximal flux reaches a plateau.[12] Under these conditions, halving the dosage of the enzyme will

Table 2 Some biological databases

Database	Description and URL
Gene	
NCBI	A general repository for molecular biology information http://www.ncbi.nlm.nih.gov
GenBank	Largest public sequence database at NCBI
EMBL	European Molecular Biology Laboratory http://www.embl-heidelberg.de/
MGDB	Mouse Genome Database http://www.informatics.jax.org
Protein	
Swiss–Prot	Annotated structure and function of proteins http://expasy.ch
PIR	Annotated structure and function of proteins http://pir.georgetown.edu
PDB	3-D biological macromolecular structural data http://www.rcsb.org/pdb/
Gene expression	
GEO	Repository for expression data at NCBI http://www.ncbi.nlm.nih.gov/geo/
Protein–protein interactions	
BIND	Biomolecular Interaction Network Database http://www.blueprint.org/index.phtml?page = databases
DIP	Catalogs experimentally determined interactions between proteins http://dip.doe-mbi.ucla.edu
Biochemical pathways	
BRENDA	Comprehensive enzyme information system http://www.brenda.uni-koeln.de
EcoCyc	Encyclopedia of *E. coli* genes and metabolism http://ecocyc.pangeasystems.com/ecocyc/ecocyc.html
EPD	Eukaryotic promoter databases http://www.epd.isb-sib.ch
ExPASy	Thumbnail sketches of metabolism and cellular biochemistry http://www.expasy.ch/chi-bin/search-biochem-index
KEGG	Kyoto encyclopedia of genes and genomes http://www.genome.ad.jp/kegg/kegg2.html
LIGAND	Database for enzyme, compounds, and reactions http://www.genome.ad.jp/dbget/ligand.html
MIRAGE	Molecular informatics resource for the analysis of gene expression http://www.ifti.org

have little effect on the total flux, and hence the gene tends to be dominant.[12] In addition, slight changes in activity because of new mutations and polymorphisms will not greatly affect flux, so the genetic variants should have little effect on fitness, and their dynamics will be in accordance with the expectations of neutral theory. A considerable body of experimental data from *Escherichia coli* in particular corroborates this conclusion.[13]

A complete metabolic network for the human red blood cell (hRBC, see Fig. 2) was analyzed using kinetic information available in the literature.[10] The mass balance equations describing the temporal variation of the metabolite concentrations, x, are represented as

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \mathbf{S}\mathbf{v}(x) \tag{1}$$

where $\mathbf{v}$ is the vector of enzymatic reaction rates and $\mathbf{S}$ is the stoichiometric matrix for the enzymatic reactions. The dynamic model was analyzed using phase plane analysis, temporal decomposition, and statistical analysis. The results revealed that the formation of pseudoequilibrium concentration states was a characteristic feature of hRBC metabolism. The analysis

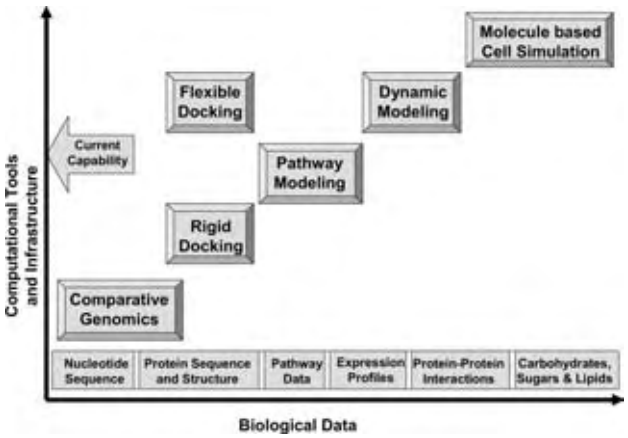


Fig. 1 The amount of biological information is growing steadily. From nucleotide sequence to protein–protein interactions, a large amount of data are being generated for each biologically functional element. Computational tools are now available to perform analysis of the data for these elements and convert them into knowledge. To address the complex interactions between various functional elements, there is a need to develop new computational analysis, modeling, and simulation capabilities. Research is currently under way to create the mathematics, algorithms, and computer architectures required to understand each level of biological complexity.

enabled the definition of physiologically meaningful pools of metabolites.[10] Analysis of this nature elucidates the nuances underlying the functioning of biological networks and the results can greatly improve understanding and facilitate manipulation.

Models based on mass action kinetics have proven useful in describing cellular processes such as receptor–ligand binding, signal transduction, and cellular behavior.[14] Biochemical networks involved in cellular signaling pathways possess rich dynamical characteristics. Experimental elucidation of the dynamic behavior of these pathways is a daunting task. Therefore, there is great interest in developing theoretical models of biochemical networks involved in signaling pathways. In conjunction with experimental data, these models reveal that signaling pathways can display ultrasensitivity, flexible bistability, and oscillatory behavior.[15]

The processes occurring during cell division in prokaryotes and eukaryotes represent a classic example of a tightly controlled oscillatory biochemical network. Models of biochemical networks involved in cell division have proven extremely useful in testing various hypotheses regarding the underlying mechanism. Specifically, it was found that toggle-like switching of the dynamics of the network arising from positive feedback is the driving force for cell division transitions into and out of mitosis.[16]

Networks of signaling pathways are also implicated in performing temporal decoding functions involved

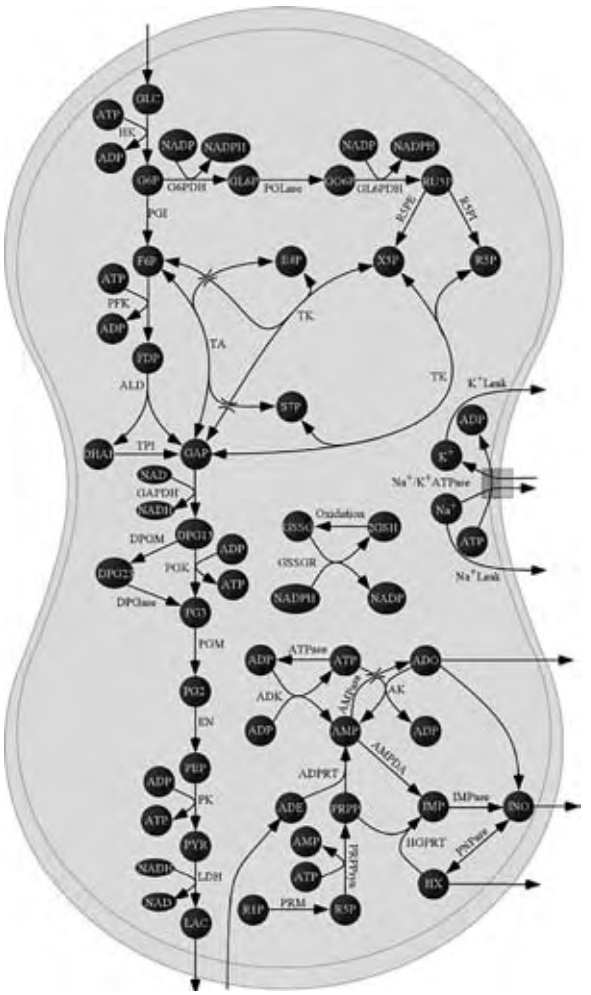


Fig. 2 The red blood cell has played a special role in the development of mathematical models of metabolism given its relative simplicity and the detailed knowledge about its molecular components. The model comprises 44 enzymatic reactions and membrane transport systems and 34 metabolites and ions. The model includes glycolysis, the Rapaport–Leubering shunt, the pentose phosphate pathway, nucleotide metabolism reactions, the sodium/potassium pump, and other membrane transport processes. Analysis of the dynamic model using phase planes, temporal decomposition, and statistical analysis shows that hRBC metabolism is characterized by the formation of pseudoequilibrium concentration states: pools or aggregates of concentration variables. (From Ref.[10].)

in bacterial chemotaxis. Using mass action kinetic models of the networks, it is possible to understand information processing (temporal tuning) by the network. Such models predict two primary mechanisms for temporal tuning of pathways: a weighted summation of responses of pathways with different timing, and the presence of biochemical feedback loops with emergent dynamics.[17]

Kinetic models of biological systems show great potential in guiding the manipulation biological

systems to precisely perform the desired tasks. Examples of such applications are the engineering of synthetic gene circuits and metabolic engineering.[10,18] Hasty et al. demonstrated that gene circuits engineered using computational modeling and experimental molecular biology can lead to insights into some of the basic modules that comprise complex, naturally occurring gene networks.[18] Drawing an analogy with established techniques in electrical engineering, the gene circuit approach uses mathematical and computational tools in the analysis of a proposed circuit diagram, while novel experimental techniques are used to construct the networks to perform the desired task. Using this technique, it is possible to design basic building blocks of gene circuits, namely, autoregulatory toggle switches (Fig. 3), logic gates, and oscillators.

The lack of kinetic information makes it difficult to develop detailed models of biological systems. Therefore, most studies attempt to model subnetworks to gain an insight into biological function. In most cases, reasonable values are assigned to biochemical parameters to develop a first-generation model. The results of this model are compared with available experimental data to test the underlying hypothesis. The model may then be used to design experiments, the results of which can help fill the knowledge gaps in the model.

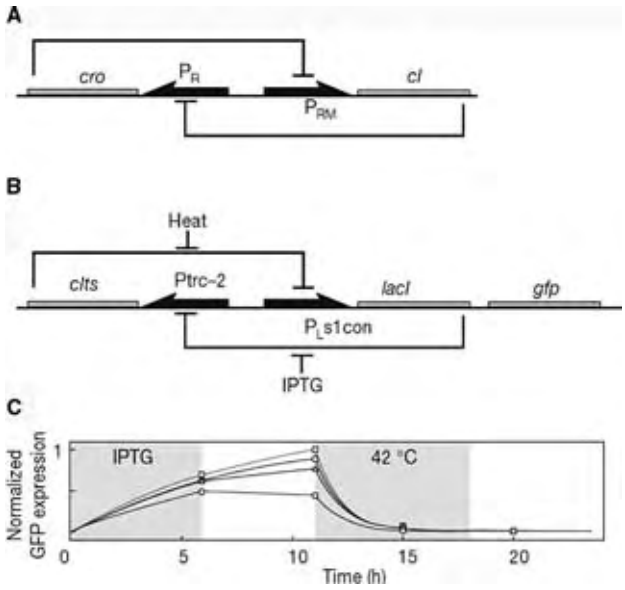


Fig. 3 Genetic toggle switch: (A) The P_R and P_{RM} system in λ phage. Cro controlled by P_R represses P_{RM} whereas CI controlled by P_{RM} represses P_R (B) The synthetic genetic toggle switch uses the promoter Ptrc-2 to control the production of a temperature-sensitive version of the CI protein (expressed by *cIts*); CI acts to repress the promoter P_Ls1con. Conversely, P_Ls1con controls transcription of the gene *lacI*, whose product LacI (lactose repressor) represses Ptrc-2. (C) Experimental results showing bistability of a genetic toggle switch in *E. coli*. The response of green fluorescent protein (GFP) is shown, which corresponds to expression of the *cI* gene. Shaded regions indicate periods of induced switching. (From Ref.[18].)

Steady-State or Static Modeling

The second broad class of genomic modeling eschews any explicit reference to the kinetic properties of defined pathways, and instead aims to find generic systems properties that emerge from the logic and connectivity of interacting networks. The approach has been applied to study genetic regulatory networks as well as metabolic networks.

The universal properties of gene regulatory networks may be described using a combination of Boolean switches.[11] In such a model, each gene interacts with k other genes. Using such a model, Kauffman illustrated the underlying mechanisms responsible for order, adaptation, and coevolution.[11] The particular biological behavior that was predicted to occur was found to be a function of the degree of interaction within the network, embodied in k. A notable observation was that, for very large values of k, a phase transition from order to chaos occurs. Therefore, there is a limit to the degree of interconnectivity that can be supported by biological networks. The application of this modeling technique to the genome-wide analysis of metabolic networks has received lot of attention in recent times. The basic principles underlying this technique and the methods for its implementation are described in the forthcoming section, followed by a description of applications in the field of metabolic engineering.

METABOLIC FLUX BALANCE ANALYSIS

Metabolic networks can be quantitatively and qualitatively studied without enzyme kinetic parameters by using a constraints-based approach.[19] Metabolic networks must obey the fundamental physicochemical laws, such as mass, energy, redox balances, diffusion, and thermodynamics. Therefore, when kinetic constants are unavailable, cellular function can still be mathematically constrained based on the mass and energy balance. Flux balance analysis (FBA) is a mathematical modeling framework that can be used to study the steady-state metabolic capabilities of cell-based physicochemical constraints.[19]

The FBA approach is attractive as it does not require kinetic parameters. Additionally, the stoichiometric parameters that are required for FBA are constant and unambiguously unknown. The stoichiometric coefficients are invariant and can be identified

from genome sequence information using bioinformatics. Therefore, like all static modeling attempts, this approach greatly benefits from comparative genomics: the presence of homologs of known enzymes can be used to reconstruct the biochemical pathway.[20] Functional mapping based on homology is possible because of the evolutionary conservation of proteins, and about 70% of gene products from each of the sequenced genomes having homologs in distant genomes.[20] Thus, the functions of many newly sequenced genes can be predicted simply by comparing different genomes and by transferring functional annotation of proteins from better-studied organisms to their homologs from lesser-studied organisms. Once these functions are known, a putative metabolic network can be constructed and bioinformatics tools can be applied to analyze the network.

FBA is attractive as a predictive tool. That is, once the model is constructed, one can predict the metabolic behavior under a number of different conditions.[21] The basic principle underlying FBA is the steady-state conservation of mass, energy, and redox potential. A dynamic mass balance can be written around each metabolite (X_i) within a metabolic network. Fig. 4 shows a hypothetical network with the fluxes (V) affecting a metabolite (X_i). The dynamic mass balance for X_i is:

$$\frac{dX_i}{dt} = V_{syn} - V_{deg} - V_{use} \pm V_{trans} \quad (2)$$

where the subscripts syn and deg refer to the synthesis and degradation reactions. The V_{trans} and V_{use} metabolic fluxes correspond to the uptake/secretion and utilization reactions, respectively. Eq. (2) can be rewritten as:

$$\frac{dX_i}{dt} = V_{syn} - V_{deg} - V_{use} + b_i \quad (3)$$

where b_i is the net transport of X_i into the defined metabolic system. The growth and the maintenance requirements (V_{use}) can be represented as fluxes in the metabolic network, which are mathematically indistinguishable from other metabolic reactions. The stoichiometry and magnitude of the growth and maintenance fluxes can be estimated from the literature or measured.[22,23] For a metabolic network that contains m metabolites and n metabolic fluxes, all the transient material balances can be represented by a single matrix equation:

$$\frac{d\mathbf{X}}{dt} = \mathbf{S} \cdot \mathbf{v} + \mathbf{b} \quad (4)$$

where $\boldsymbol{X}$ is an m dimensional vector defining the metabolite concentrations, $\mathbf{v}$ is the vector of n metabolic fluxes, $\mathbf{S}$ is the $m \times n$ stoichiometric matrix, and $\mathbf{b}$ is the vector of metabolic transport processes.

The time constants characterizing metabolic transients are typically very rapid compared to those of cell growth and process dynamics; therefore, the transient mass balances can be simplified to consider only the steady-state behavior.[24] Eliminating the time derivative in Eq. (4) and rearranging the equation yields:

$$\mathbf{S} \cdot \mathbf{v} + \mathbf{I} \cdot \mathbf{b} = 0 \quad (5)$$

Eq. (5) states that at steady-state (or pseudo-steady-state), all the formation fluxes of a metabolite must be balanced by the degradation fluxes.

Not all metabolites are capable of being transported into or out of the cell; therefore, the $\mathbf{I} \cdot \mathbf{b}$ term can be simplified by removing the rows in the $\mathbf{b}$ vector that correspond to metabolites that are not transported, thus forming a vector $\mathbf{b}_r$. Additionally, the corresponding columns in $\mathbf{I}$ are eliminated, thus forming the matrix $\mathbf{U}$. Furthermore, the stoichiometric matrix can be partitioned such that the metabolic reactions within the system boundary are defined by $\mathbf{S}_{reactions}$, $\mathbf{S}_{use}$, representing the internal fluxes and the biomass/maintenance requirement fluxes, respectively. The $\mathbf{U}$ matrix allows certain metabolites to be transported into and out of the system. Eq. (5) can therefore be rewritten as:

$$\begin{aligned} &\mathbf{S} \cdot \mathbf{v} + \mathbf{U} \cdot \mathbf{b}_r = 0 \\ &[\mathbf{S}_{reactions} | \mathbf{S}_{use} | \mathbf{U}] \begin{bmatrix} \mathbf{v}_{reactions} \\ \mathbf{v}_{use} / \mathbf{b}_r \end{bmatrix} = 0 \end{aligned} \quad (6)$$

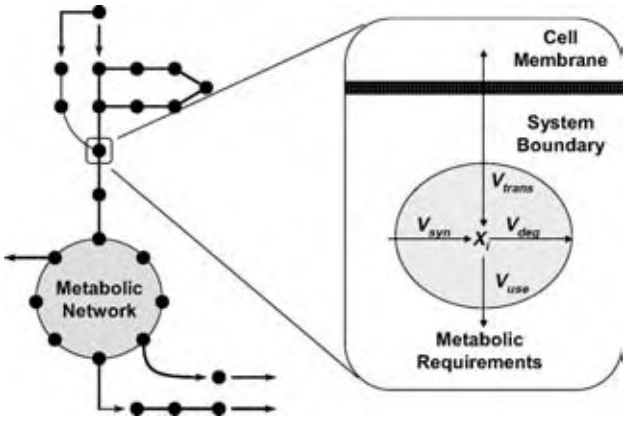


Fig. 4 Flux balance models use material balances around each metabolite in a metabolic network. The concentration of each metabolite, X_i, is affected by various fluxes, V_j. V_{trans} is the uptake or secretion flux, while V_{use} is the flux required for growth and maintenance. V_{syn} and V_{deg} refer to the fluxes resulting from the metabolic synthesis and degradation of the metabolite.

where

$$\mathbf{S}' = [\mathbf{S}_{\text{reactions}}|\mathbf{S}_{\text{use}}|\mathbf{U}]$$

$$\mathbf{v}' = \left[\frac{\mathbf{v}_{\text{reactions}}}{\mathbf{v}_{\text{use}}/\mathbf{b}_r}\right]$$

Therefore, we generate the following equation:[21]

$$\mathbf{S}' \cdot \mathbf{v}' = 0 \tag{7}$$

Every metabolite inside the system boundary corresponds to a row in the stoichiometric matrix; however, some of these metabolites are intracellular, while the rest are extracellular (see Fig. 4). The stoichiometric matrix is arranged such that the m_{i} internal metabolites are entered first, and the m_{e} external metabolites second ($m = m_{\text{i}} + m_{\text{e}}$). When the stoichiometric matrix is arranged in this manner, the $\mathbf{S}_{\text{use}}$ and the $\mathbf{U}$ matrices take on the following form:

$$\mathbf{U} = \left[\frac{0}{I}\right]$$

$$\mathbf{S}_{\text{use}} = \left[\frac{\mathbf{S}^u_{\text{use}}}{0}\right] \tag{8}$$

where $\mathbf{U}$ is an $m \times m_{\text{e}}$ matrix, $\mathbf{I}$ is the $m_{\text{e}} \times m_{\text{e}}$ identity matrix, and $\mathbf{S}^u_{\text{use}}$ is a matrix with m_i rows.

Eq. (7) defines the mass, energy, and redox potential constraints on the metabolic network, thus effectively defining the capabilities of the metabolic network. Flux distributions (the flux vector, $\mathbf{v}$) that satisfy Eq. (7) are possible metabolic phenotypes. However, this estimation of the metabolic capabilities is conservative and can be further reduced by adding additional constraints.

Additional Constraints

Eq. (7) defines the mass, energy, and redox balance constraints on the metabolic system. However, additional physicochemical constraints are typically placed on the metabolic network. For example, the value of the flux through each of the metabolic reactions can be constrained based on reversibility, transport limitations, and/or maximum allowable fluxes. Irreversibility constraints are enforced by placing a lower limit of zero on the respective irreversible metabolic reactions:

$$0 \leq v_j \leq \beta_j \tag{9}$$

In the absence of additional information, β_j (metabolic flux upper limit) is set to infinity. The constraints in Eq. (9) can be representative of a maximum allowable flux through an irreversible reaction by setting β_j to a finite value. The flux limitations may be known from enzyme expression levels and $V_{\max}$ data. The following equation can be used if information regarding the metabolic flux range is known, rather than simply specifying the maximal flux levels:

$$\alpha_j \leq v_j \leq \beta_j \tag{10}$$

The restriction of the flux through certain reactions can be used to model the regulatory events occurring within the cell. Further, any experimental information can be simulated by appropriately assigning values to α and β. It is also possible to incorporate regulatory information using this methodology. For instance, setting fluxes to zero or some low level can represent the repression of certain enzymes, such as *ppc* and *pps* during growth on glucose. Thus, the use of additional constraints provides increased flexibility in analyzing the metabolic network by incorporating additional knowledge about a particular cell.

Commonly, restrictions are placed on the transport fluxes (corresponding to the various uptake processes). For example, when specific information regarding an uptake rate is not known, the maximal uptake rate is set by using the constraints defined by Eq. (10) (β is defined as the maximal uptake rate or $\mathbf{b}_r$ element). However, often in experimental systems, the uptake rates have been measured, and the flux value can be fixed by using the formalism described by Eq. (10). Transport reactions for metabolites that are not present in a simulated culture condition are constrained to zero ($0 \leq v_j \leq 0$).

Solution Method

Typically, the number of metabolic fluxes is greater than that of mass balances (i.e., $n > m$) resulting in a plurality of feasible flux distributions or flux vectors that satisfy Eq. (7). This range of solutions is indicative of the flexibility in the metabolic network. Although infinite in number, the solutions to Eq. (7) are in the null space of the matrix $\mathbf{S}$.[25] However, in the analysis of metabolic systems, we are interested in the feasible set.[26] The feasible set is the region of intersection between the null-space of $\mathbf{S}'$ and the linear inequality constraints Eqs. (9) and (10). The feasible set defines the capabilities of the metabolic genotype of a given organism [more accurately than just considering the solutions to Eq. (8), as discussed above], because it defines all the metabolic flux distributions that can be achieved with a particular set of genes.[19] Each flux distribution is a specific metabolic phenotype that is expressed under particular conditions.

There are two fundamentally different methods to define a specific metabolic phenotype or flux vector.

First, one can experimentally measure as many fluxes (or more) as the dimension of the null-space, so as to uniquely calculate the remaining fluxes.[24] This approach is called metabolic flux analysis. Alternatively, an objective of the metabolic network can be chosen to computationally explore the best use of the metabolic network by a given metabolic genotype. Herein, we pursue the second option. The solution to Eq. (7) subject to the linear inequality constraints can be formulated as a linear programming (LP) problem, in which one finds the flux distribution that minimizes a particular objective. Mathematically, the LP problem is stated as:

Minimize Z, where

$$Z = \sum c_i \cdot v_i = \langle \mathbf{c} \cdot \mathbf{v} \rangle \quad (11)$$

where Z is the objective that is represented as a linear combination of metabolic fluxes v_i. Herein, the vector $\mathbf{c}$ is defined as the unit vector in the direction of the growth flux, v_{growth}. The optimization can also be stated as the equivalent maximization problem, i.e., by changing the sign on Z.

Objective Function

This general representation of Z enables the formulation of a number of diverse objectives. These can be design objectives for a strain, exploitation of the metabolic capabilities of a genotype, or physiologically meaningful objective functions, such as maximization of cellular growth rate. For instance, growth can be defined in terms of biosynthetic requirements based on cellular biomass composition, as defined in the literature.[22] Thus, biomass generation is defined as a reaction flux draining the intermediate metabolites in the appropriate ratios, and the negative of this flux is defined as the objective function to be minimized, Z. The diversity of metabolic objectives that can be represented by Eq. (11) provides the necessary flexibility to answer a number of important questions.

APPLICATIONS

Flux balance analysis has been used for over 15 yr to study the metabolic flux distribution.[19] Initially, the applications of FBA were primarily educational; however, recently, the utility has grown. The FBA has been applied to study the effect of gene deletions, the design of bacterial metabolism for industrial and environmental applications, or for the computational exploration of cellular physiology.[19] Metabolic engineering has been successfully applied to engineer micro-organisms to produce valuable biochemicals for the pharmaceutical industry. For example, Stafford et al. successfully used *Rhodococcus* sp. to convert indene into precursors of the HIV protease inhibitor, and Chang et al. engineered *E. coli* to produce optically pure isomers of lactate.[27,28] The FBA will likely find additional applications in the near future.

Traditionally, to identify the function of a protein the first step was to obtain a mutant and study its properties. This same procedure can also be applied to whole-cell models. A detailed understanding of the metabolic network is possible by understanding the behavior of the system without a key component. The FBA has been used to predict the whole-cell metabolic flux distribution for a number of mutants under different conditions.[19] The results of the computational gene deletion analysis indicate that the model accurately predicts the qualitative growth behavior in over 85% of the cases. Genome enabled studies, such as the analysis of large sets of mutants in parallel and whole-cell transcript profiling, can be further aided by the interpretation of the data from a metabolic model.[19]

The FBA can provide insights for engineering metabolic networks. For example, FBA can be used to identify the important fluxes for the production of a metabolite or protein. Once the key points are identified, the manipulations can be carried out using molecular genetics. In the past two decades, the ability to make precise manipulations of the cellular genetic content has greatly outpaced our ability to predict the effect of these changes. Through techniques such as FBA, metabolic engineers have identified tools and techniques that can overcome this limitation. Thus, in the near future, metabolic engineers and bioinformatists will be able to harness the power of having completely sequenced genomes and a descriptive mathematical model for their particular system of interest.

Finally, mathematical modeling in biology can provide valuable insight into the integrative behavior of cellular networks. For example, FBA can be used to predict the metabolic by-products in *E. coli* that are formed during aerobic and anaerobic metabolism.[19] Furthermore, FBA has been utilized to generate several novel hypotheses regarding the function of the *E. coli* metabolic network. These were demonstrated to be consistent with available experimental data.[21]

HIGH-THROUGHPUT EXPERIMENTS

An area that is receiving considerable attention is the development of computational tools to meaningfully analyze data obtained using high-throughput experiments. These techniques have the potential to reveal important information otherwise buried within a sea

of data. Gene expression, defined by the levels of cellular mRNA, is the first aspect of gene function amenable to genome-scale measurements with readily-available technology. It is now possible to carry out massively parallel analysis of gene expression on tens of thousands of genes from a given sample.[29] A number of techniques have been developed to perform statistical analysis of the measurements, including public databases and proprietary software. The most popular method for analyzing gene expression profiles is hierarchical clustering. This approach is extremely effective in identifying groups of coregulated genes, and is incorporated into the Cluster/TreeView software (http://rana.lbl.gov/EisenSoftware.htm). A particularly impressive application of this strategy was the compendium of expression profiles approach in which over 300 different strains and conditions were contrasted.[29] This technique can be used to study the coregulation of groups of genes under different conditions and upon treatment with drugs. Compared to the method of assessing gene function by coregulation, the compendium approach has a significant advantage in that it does not rely on the regulatory characteristics of the gene of interest. Furthermore, the same compendium used to characterize mutants can also be used for other perturbations, including treatments with pharmaceutical compounds, and potential disease states as well. For instance, the compendium mentioned here was used to identify a novel target for the commonly used drug dyclonine.[29]

The results of incisive analysis of data obtained from high-throughput experiments have the potential to provide important information regarding the underlying structure of the biological networks. Information about the architecture of biological networks can provide tremendous insights into their structure and function. Moreover, such data are extremely useful in the development of dynamic models of these pathways. For example, using a network clustering method to analyze high-throughput protein–protein interaction data obtained using two-hybrid screens, it is possible to unravel the signaling-protein modules (Fig. 5).[30] In Fig. 5, the data are seen to decompose into well-defined clusters. For example, Ras-pathway proteins form a single cluster. The organization of a pathway into separate protein clusters reflects the existence of more than one module within the pathway. The results indicate that pathway models can then be developed in a modular manner.

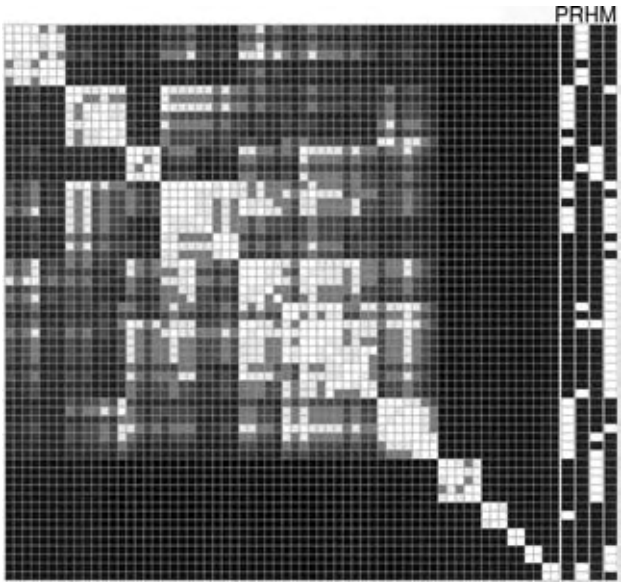


Fig. 5 Diagram showing the clustering of proteins involved in the yeast signaling network. From a global two-hybrid screen, 64 proteins that have at least one interaction with another signaling protein were selected. A symmetrical matrix of these 64 proteins was clustered identically in both dimensions. The cluster tree is not shown. Each row or column represents a protein. Each feature is the intersection of two proteins and is a grayscale representation of pairwise protein association (see text). Direct interactions are white. Indirect interactions of increasing distance (weaker association) are progressively darker. All features on the diagonal (self-associations) are white. Columns to the right of the clustered network represent Munich Information Center for Protein Sequences (MIPS)-defined signaling pathways [P, polarity-PKC; R, Ras; H, HOG; M, mating/filamentation MAPK (mfMAPK)]. White bars in the MIPS pathway columns indicate protein members of the pathway. (From Ref.[30].)

CONCLUSIONS

Recent advances in molecular biology have created a wealth of information from which it is possible to understand biological function at the molecular level. Toward this end, computational tools are being rapidly developed to process this information and convert it into knowledge. The availability of biological information about microbes and the ability to effectively process that information to define the initial capabilities in both normal and perturbed states have tremendous potential in the chemical process industry. Industries that will greatly benefit include the pharmaceutical, biotechnology, health, food, defense, biochemical, and environmental.

A significant effort in the information processing direction has been devoted to reconstructing, analyzing, and understanding the biological pathways involved in gene regulation, signaling, and metabolism. Techniques are now available to model biological subnetworks and, in some cases, even complete networks. Their applications have revealed that biological networks exhibit complex dynamical behavior,

including ultrasensitivity, flexibility, bistability, oscillatory behavior, and temporal tuning. A good understanding of these functions will be extremely useful in engineering biological systems to perform the desired tasks and, therefore, will have tremendous applications in process design, development, and optimization.

The flux balance approach is an elegant technique for analyzing biological pathways in that the approach allows metabolic networks to be quantitatively and qualitatively studied without knowledge of specific enzyme kinetic parameters. A combination of FBA and bioinformatics provides a rapid, economical, and predictive in silico screening method for potential strategies to produce a desired compound.

In the near future, there will be tremendous progress in the development of sophisticated computational tools to assimilate biological information into quantitative models. Such models will facilitate a deeper understanding of biological function, thereby allowing their precise engineering to perform the desired tasks. A major challenge in modeling and simulating systems biology is integrating high- and low-level models so that a more accurate picture of the entire biological process can be obtained. Integrating models of protein structure and function with those of biochemical pathways promises to provide insights into the complexities of biological function.

High-throughput experiments will also significantly benefit from powerful computational tools to extract biologically meaningful information from a large amount of data. The results of these experiments will be extremely useful in providing information about the underlying structure of biochemical pathways and in refining mathematical and computational models. With the development of sophisticated bioinformatic tools for understanding and manipulating biological function with amazing precision, there is considerable excitement in the chemical process industry.

ACKNOWLEDGMENT

This work was supported by the Microbial Genome Program grant from the U.S. Department of Energy.

REFERENCES

1. Pizza, M.; Scarlato, V.; Masignani, V.; Guiliani, M.M.; Arico, B.; Comanducci, M.; Jennings, G.T.; Baldi, L.; Bartolini, E.; Capecchi, B.; Galeotti, C.L.; Luzzi, E.; Manetti, R.; Marchetti, E.; Mora, M.; Nuti, S.; Ratti, G.; Santini, L.; Savino, S.; Scarselli, M.; Storni, E.; Zuo, P.; Broeker, M.; Hundt, E.; Knapp, B.; Blair, E.; Mason, T.; Tettelin, H.; Hood, D.W.; Jeffries, A.C.; Saunders, N.J.; Granoff, D.M.; Venter, J.C.; Moxon, E.R.; Grandi, G.; Rappuoli, R. Identification of vaccine candidates against serogroup *B meningococcus* by whole-genome sequencing. Science **2000**, *287*, 1816–1820.
2. Tettelin, H.; Masignani, V.; Cieslewicz, M.J.; Eisen, J.A.; Peterson, S.; Wessels, M.R.; Paulsen, I.T.; Nelson, K.E.; Margarit, I.; Read, T.D.; Madoff, L.C.; Wolf, A.M.; Beanan, M.J.; Brinkac, L.M.; Daugherty, S.C.; DeBoy, R.T.; Durkin, A.S.; Kolonay, J.F.; Madupu, R.; Lewis, M.R.; Radune, D.; Fedorova, N.B.; Scanlan, D.; Khouri, H.; Mulligan, S.; Carty, H.A.; Cline, R.T.; Van Aken, S.E.; Gill, J.; Scarselli, M.; Mora, M.; Iacobini, E.T.; Brettoni, C.; Galli, G.; Mariani, M.; Vegni, F.; Maione, D.; Rinaudo, D.; Rappouli, R.; Telford, J.L.; Kasper, D.L.; Grandi, G.; Fraser, C.M. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc. Natl. Acad. Sci. USA **2002**, *99* (19), 12391–12396.
3. Klaenhammer, T.; Altermann, E.; Arigoni, F.; Bolotin, A.; Breidt, F.; Broadbent, J.; Cano, R.; Chaillou, S.; Deutscher, J.; Gasson, M.; van de Guchte, M.; Guzzo, J.; Hartke, A.; Hawkins, T.; Hols, P.; Hutkins, R.; Kleerebezem, M.; Kok, J.; Kuipers, O.; Lubbers, M.; Maguin, E.; McKay, L.; Mills, D.; Mauta, A.; Overbeek, R.; Pel, H.; Pridmore, D.; Saier, M.; van Sinderen, D.; Sorokin, A.; Steele, J.; O'Sullivan, D.; de Vos, W.; Weimer, B.; Zagorec, M.; Siezen, R. Discovering lactic acid bacteria by genomics. Antonie Van Leeuwenhoek **2002**, *82* (1–4), 29–58.
4. Paulsen, I.T.; Seshadri, R.; Nelson, K.E.; Eisen, J.A.; Heidelberg, J.F.; Read, T.D.; Dodson, R.J.; Umayam, L.; Brinkac, L.M.; Beanan, M.J.; Daugherty, S.C.; Deboy, R.T.; Durkin, A.S.; Kolonay, J.F.; Madupu, R.; Nelson, W.C.; Ayodeji, B.; Kraul, M.; Shetty, J.; Malek, J.; Van Aken, S.E.; Riedmuller, S.; Tettelin, H.; Gill, S.R.; White, O.; Salzberg, S.L.; Hoover, D.L.; Lindler, L.E.; Halling, S.M.; Boyle, S.M.; Fraser, C.M. The *brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. Proc. Natl. Acad. Sci. USA **2002**, *99* (20), 13148–13153.
5. Dayem, L.C.; Carney, J.R.; Santi, D.V.; Pfeifer, B.A.; Khosla, C.; Kealey, J.T. Metabolic engineering of a methylmalonyl-CoA mutase-epierase pathway for complex polyketide biosynthesis in *Escherichia coli*. Biochemistry **2002**, *41* (16), 5193–5201.
6. White, O.; Eisen, J.A.; Heidelberg, J.F.; Hickey, E.K.; Peterson, J.D.; Dodson, R.J.; Haft, D.H.; Gwinn, M.L.; Nelson, W.C.; Richardson, D.L.;

Moffat, K.S.; Qin, H.; Jiang, L.; Pamphile, W.; Crosby, M.; Shen, M.; Vamathevan, J.J.; Lam, P.; McDonald, L.; Utterback, T.; Zalewski, C.; Makarova, K.S.; Aravind, L.; Daly, M.J.; Minton, K.W.; Fleischmann, R.D.; Ketchum, K.A.; Nelson, K.E.; Salzberg, S.; Smith, H.O.; Venter, J.C.; Fraser, C.M. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science **1999**, *286*, 1571–1517.

7. McMahon, K.D.; Jenkins, D.; Keasling, J.D. Polyphosphate kinase genes from activated sludge carrying out enhanced biological phosphorous removal. Water Sci. Technol. **2002**, *46* (1–2), 155–162.
8. Sensen, C.W. From model organisms to organismal models: visualizing complex genomic datasets. Biosilico **2003**, *1* (1), 23–26.
9. Frazier, M.E.; Johnson, G.M.; Thomassen, D.G.; Oliver, C.E.; Patrinos, A. Realizing the potential of the genome revolution: the genomes to life program. Science **2003**, *300*, 290–293.
10. Kauffman, K.J.; Pajerowski, J.D.; Jamshidi, N.; Palsson, B.O.; Edwards, J.S. Description and analysis of metabolic connectivity and dynamics in the human red blood cell. Biophys. J. **2002**, *83* (2), 646–662.
11. Kauffman, S.A. *The Origins of Order*; Oxford University Press: Oxford, 1993.
12. Kacser, H; Burns, J. The molecular basis of dominance. Genetics **1981**, *97* (3–4), 639–666.
13. Dykhuizen, D.; Dean, A.; Hartl, D. Metabolic flux and fitness. Genetics **1987**, *115* (1), 25–31.
14. Lauffenburger, D.A.; Linderman, J.J. *Receptors: Models for Binding, Trafficking, and Signaling*; Oxford University Press: New York, 1993.
15. Neves, S.R.; Iyengar, R. Modeling of signaling networks. Bioessays **2002**, *24* (12), 1110–1117.
16. Sha, W.; Moore, J.; Chen, K.; Lassaletta, A.D.; Yi, C.; Tyson, J.J.; Sible, J.C. Hysteresis driven cell-cycle transitions *in Xenopus laevis* egg extract. Proc. Natl. Acad. Sci. USA **2003**, *100* (3), 975–980.
17. Bhalla, U.S. Mechanism for temporal tuning and filtering by postsynaptic signaling pathways. Biophys. J. **2002**, *83* (2), 740–752.
18. Hasty, J.; McMillen, D.; Collins, J.J. Engineered gene circuits. Nature **2002**, *420*, 224–230.
19. Edwards, J.S.; Covert, M.; Palsson, B.O. Metabolic modeling of microbes: the flux-balance approach. Environ. Microbiol. **2002**, *4* (3), 133–140.
20. Galperin, M.Y.; Koonin, E.V. Comparative genome analysis. In *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 2nd Ed.; Baxevanis, A.D., Ouellette, B.F.F., Eds.; John Wiley: New York, 2001; 359–392.
21. Edwards, J.S.; Ibarra, R.U.; Palsson, B.O. In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. Nat. Biotechnol. **2001**, *19*, 125–130.
22. Neidhardt, F.C.; Ingraham, J.L.; Schaechter, M. *Physiology of the Bacterial Cell*; Sinauer Associates, Inc.: Sunderland, MA, 1990.
23. Varma, A.; Palsson, B.O. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. Appl. Environ. Microbiol. **1994**, *60*, 3724–3731.
24. Vallino, J.; Stephanopoulos, G. Metabolic flux distributions in *Corynebacterium glutmicum* during growth and lysine overproduction. Biotechnol. Bioeng. **1993**, *41*, 633–646.
25. Strang, G. *Linear Algebra and Its Applications*; 3rd Ed.; Saunders College Publishing: Forth Worth, 1988.
26. Chvatal, V. *Linear Programming*; W.H. Freeman and Company: New York, 1983.
27. Stafford, D.E.; Yanagimachi, K.S.; Stephonopoulos, G. Metabolic engineering of indene bioconversion in *Rhodococcus* sp. Adv. Biochem. Eng. Biotech. **2001**, *73*, 85–101.
28. Chang, D.; Jung, H.; Rhee, J.; Pan, J. Homofermentative production of D- or L-lactate in metabolically engineered *Escherichia coli* RR1. Appl. Environ. Microbiol. **1999**, *65* (4), 1384–1389.
29. Hughes, T.R.; Marton, M.J.; Jones, A.R.; Roberts, C.J.; Stroughton, R.; Armour, C.D.; Benett, H.A.; Coffey, E.; Dai, H.; He, Y.D.; Kidd, M.J.; King, A.M.; Meyer, M.R.; Slade, D.; Lum, P.Y.; Stepaniants, S.B.; Shoemaker, D.G.; Chakraburtty, K.; Simon, J.; Bard, M.; Friend, S.H. Functional discovery via a compendium of expression profiles. Cell **2000**, *102*, 109–126.
30. Rives, A.W.; Galitski, T. Modular organization of cellular networks. Proc. Natl. Acad. Sci. USA **2003**, *100* (3), 1128–1133.

Biomass to Ethanol

Randy S. Lewis
School of Chemical Engineering, Oklahoma State University, Stillwater, Oklahoma, U.S.A.

Rohit P. Datar
Technical Operations, CPKelco, Okmulgee, Oklahoma, U.S.A.

Raymond L. Huhnke
Biosystems and Agricultural Engineering, Oklahoma State Unviersity, Stillwater, Oklahoma, U.S.A.

INTRODUCTION

Since the energy crisis of the 1970s, the development of low-cost, sustainable, and renewable energy sources has been a major focus of research. Fuel-grade ethanol is a transportation energy source that can be produced from biomass. Biomass resources for ethanol include sugar-based crops such as beets and cane, starch-based crops such as corn and potatoes, and lignocellulosic feedstocks such as wood, corn stover, and grasses.

This entry provides an overview on the processes and limitations for making ethanol from biomass using either a fermentable sugar platform or a fermentable producer gas platform. Fig. 1 shows a simplified schematic of the platforms that are described later. Two of the platforms convert biomass (sugar, starch, or lignocellulose) into fermentable sugars, while the other platform converts lignocellulose into fermentable producer gas. Since an azeotrope occurs in the distillation process (3–5% water remains), further dehydration of ethanol can be accomplished through azeotropic distillation or molecular sieve dehydration.

The advantages and disadvantages of using ethanol as a transportation fuel are also described along with the issues that must be resolved in providing a consistent supply of high-quality and low-cost biomass for the production of ethanol. Several technological barriers remain in producing ethanol, including high enzyme costs, the inability of microbial strains to simultaneously and efficiently utilize a variety of sugar substrates, and the low ethanol yield in producer gas fermentation. However, with the development of inexpensive routes for production, low-cost ethanol could soon become a reality.

ETHANOL AS A BIOFUEL

In 2001, 80% of worldwide energy was supplied by fossil fuels, with only 1.4% being supplied by sustainable biomass used for electricity, heat production, and transportation. With regards to transportation energy, 97% was supplied by fossil fuels in industrialized nations.[1] Due to concerns over the depletion of fossil fuels, utilizing renewable resources would be beneficial for supplementing energy needs currently obtained from fossil fuels.

Fuel-grade ethanol (bioethanol) and other valuable commercial products can be produced by utilizing renewable resources (i.e., biomass) as a starting raw material. Over the last few years, there has been a tremendous growth in the U.S. ethanol industry fueled by a growing demand for renewable fuels. In 2002, the total ethanol produced in the United States was 2 billion gallons, representing a 20% increase over 2001 and a 45% increase over 1999.[2] The potential world ethanol production from biomass is 130 billion gallons per year, with lignocellulosic biomass contributing to 90% of this amount. This production capacity is 16 times higher than the current ethanol production in the world and could replace 32% of the global gasoline consumption.[3] As a supplement to gasoline, ranging from 10% (E10) to 95% (E95), ethanol is beneficial for environmental and economic/national security reasons.

Environmental Benefits

The Clean Air Acts Amendments of 1990 mandated the addition of oxygenates to gasoline to reduce the formation of harmful ground-level ozone (photochemical smog) in urban areas. In addition, the reduction of toxic gaseous pollutants and volatile organic compounds from emissions of motor vehicles was mandated. Methyl tertiary butyl ether (MTBE) has traditionally been used as an oxygenate in gasoline due to its high octane number, low sulfur content, and relatively low production cost as compared to other high-octane components.[4] However, in recent years, MTBE has been shown to be a potential carcinogen in humans and animals and a source of ground- and surface-water contamination. MTBE is currently being eliminated through political

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039776

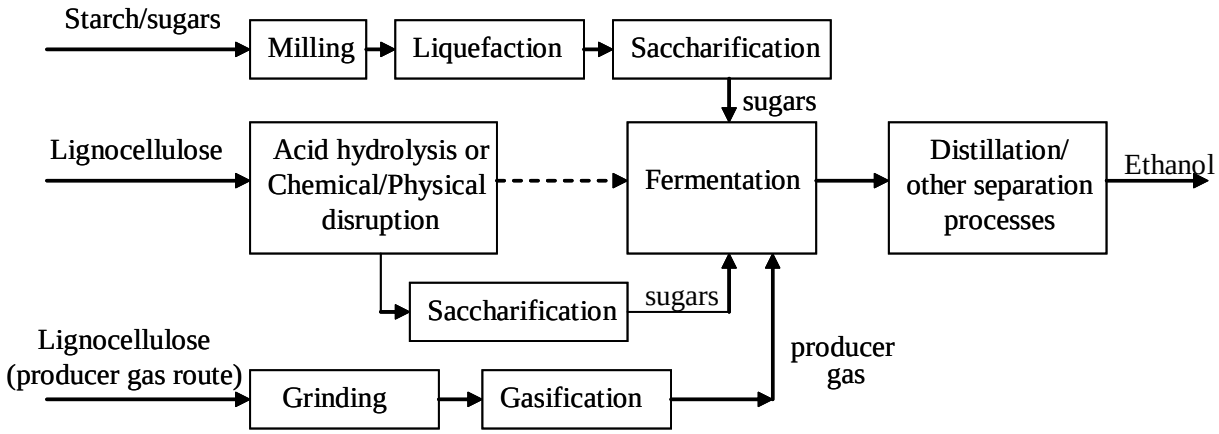


Fig. 1 Processes for biomass conversion to ethanol.

means, although efforts are being made to prevent the ban.[5] Ethanol is a very strong candidate to replace MTBE despite higher production costs, since it is environmentally friendly and can be produced domestically from renewable sources.

The addition of ethanol to gasoline reduces harmful vehicle emissions and facilitates the reduction or removal of toxic air components, which are commonly found additives in gasoline.[2] Emissions of carbon dioxide (CO_2), the primary greenhouse gas contributing to global warming, have increased steadily, and the transportation sector alone accounts for 33% of total emissions.[6] It is estimated that an 85% ethanol fuel would reduce greenhouse gas emissions between 15% and 30%.[7] Ethanol-blended fuels would reduce particulate emissions, primarily a concern in diesel engines, by as much as 40%.[2]

Economic/National Security Benefits

Currently, the United States imports about 57% of its oil, creating a US$ 66 billion trade deficit; the deficit is expected to rise to US$ 170 billion by 2020.[8] Projections show that oil imports could grow to 68% in 2025, thus increasing the dependence on crude oil from potentially unstable regions around the world.[2] It is estimated that a 10% ethanol blend would reduce oil requirements by 3%, while a 95% ethanol blend could reduce the oil requirement by as much as 44%.[9] Thus, it is beneficial to increase local energy production via fuels from renewable sources to prevent sudden energy disruptions, which could have severe economic ramifications.

Drawbacks

There have been ongoing arguments over the use of ethanol as a fuel. Ethanol can be corrosive to some metals, gaskets, and seals in car engines. The heating value of ethanol is 76,000 Btu/gal compared to 115,000 Btu/gal for gasoline; hence, more ethanol is required to travel the same distance. There have been conflicting reports on the preceding statement, since some argue that there is statistically no mileage difference due to the higher combustion efficiency of ethanol compared to gasoline.[10] Some argue that the cost of ethanol is too high for it to be used as an additive or a replacement to gasoline and that the ethanol industry will not survive without tax incentives.[11] Arguably, ethanol as a fuel will not be able to compete with gasoline unless less expensive routes for ethanol production are investigated and developed.

BIOMASS RESOURCES

The major sources of biomass feedstock are low-cost residues, wastes, and byproducts such as corn stover, wheat and rice straw, municipal solid waste, sawdust, dedicated energy crops such as switchgrass, hybrid poplars, and hybrid willows, and corn starch. The total estimated availability of usable biomass in the world is about 2 billion dry tons per year.[3] Thus, considerable effort has been devoted towards developing energy crops to meet the estimated increase in energy usage. An estimated 549 million acres of land will be available to cultivate such energy crops (based on estimates by the Oak Ridge National Laboratory—Biofuels Feedstock Development Program). The use of herbaceous energy crops, such as switchgrass, for the production of ethanol also offers a significant energy gain compared to corn. The ratio of total available energy to the total energy input for harvesting and handling is 4.43 and 1.21 for switchgrass and corn, respectively, translating to energy gains of 343% and 21% respectively.[12] In addition to its energy gain, switchgrass

has been chosen as the model crop for biomass utilization because of its high productivity over a wide geographic range, ability to grow on marginal lands, low water and nutrient requirements, and environmental benefits such as improved soil conservation.[13]

One of the primary reasons for the slow acceptance of biomass to ethanol conversion is the lack of a reliable and sustainable lignocellulosic supply infrastructure to maintain a commercial conversion facility. There are many issues that must be resolved in providing a consistent supply of high-quality, low-cost biomass.[14]

- *Feedstock quality*. To operate round the year, most conversion facilities need to utilize a variety of biomass feedstocks. Therefore, changes in feedstock quality, including physical and chemical characteristics, must be taken into account in the design, operation, and performance of a commercial plant.
- *Harvesting and collection*. Crop residues must be collected soon after the principal crop has been removed to reduce field losses and avoid contamination. Other feedstocks such as perennial grasses are harvested only during selected months of the year. The timing of harvest and collection of crop residues and energy crops often necessitates the use of storage facilities.
- *Transportation*. Movement of biomass from the original source to a processing facility can be very costly because of its relatively low bulk density. For example, the bulk density of coal is over three times that of most standard biomass packages.
- *Storage*. The quantities required for a commercial conversion facility demand the use of storage facilities even if only used as a supply buffer at the facility itself. For dry biomass, protection from the elements is often required to maintain acceptable moisture levels and to minimize storage losses. In addition, fire protection is an important consideration when storing large quantities of biomass at one site.

BIOMASS TO ETHANOL: FERMENTABLE SUGAR PLATFORM

A majority of ethanol produced from biomass utilizes the fermentation of sugars that are obtained from polysaccharides prevalent in the biomass. Feedstocks include sugar-based crops such as beet and cane, starch-based crops such as corn and potato, and lignocellulosic plants such as wood, corn stover, and grasses. Fermentation increases in difficulty from sugar crops to starchy crops to lignocellulosic plants as a result of the increasing complexity and heterogeneity of sugar components of these biomass resources. Sugar crops contain the simplest sources of sugars in the form of sucrose (Fig. 2), a disaccharide consisting of

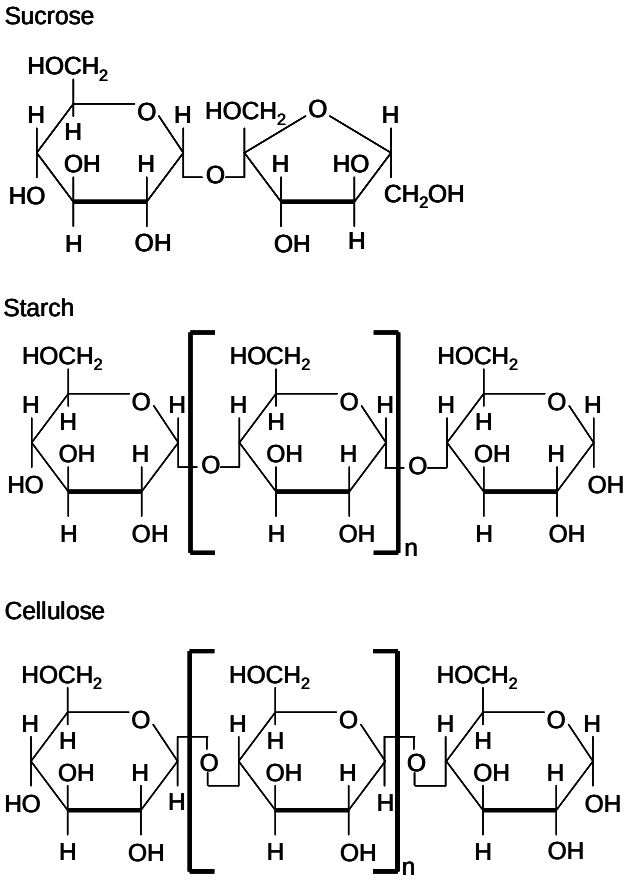


Fig. 2 Sugar units of biomass.

glucose and fructose (both are six-carbon sugars). Starch is a polysaccharide that serves as the nutritional reservoir in plants and is composed of glucose residues in α-1,4 linkages that must be converted to single glucose units for fermentation. Starch consists of either an unbranched form (Fig. 2) or a branched form containing additional α-1,6 linkages of glucose every 20–30 glucose units.

Lignocellulosic materials are characterized by varying amounts of cellulose, hemicellulose, lignin, and small quantities of other extractives. Typically, the composition by weight is 40–50% cellulose, 20–40% hemicellulose, and 10–30% lignin.[15] Cellulose (Fig. 2) is an unbranched polysaccharide that serves a structural role in plants and is composed of glucose residues joined by β-1,4 linkages, which are stronger and more difficult to break than the starch linkages. Hemicellulose is a branched polysaccharide, consisting primarily of five-carbon sugars such as xylose and arabinose in addition to six-carbon sugars such as glucose and mannose. These complex sugars must be converted to single sugar units for fermentation. Lignin is a group of amorphous, high molecular weight compounds that cannot be fermented.

Fermentation of Sugar and Starch Crops

Most current commercial ethanol production is from the fermentation of sugar and starch crops. Yeast can rapidly convert sucrose to ethanol with a theoretical carbon conversion of 67%. The production of ethanol from corn grew to about 1.9 billion gallons in 2001.[16] This accounted for 90% of the total ethanol production and an estimated 615 million bushels of corn (6.2% of total corn produced) were consumed. The remaining 10% of ethanol production was by fermentation of grain sorghum, barley, wheat, cheese whey, and potatoes.

The basic steps for the conversion of sugar and starch crops to ethanol are: 1) The biomass either undergoes a grinding process (dry milling) or is chemically treated (wet milling) to reduce the size of the feedstock; 2) if starch crops are used, the starch is converted to sugars by enzymatic treatment—called saccharification; 3) the sugars are fermented to ethanol by yeast; and 4) the ethanol is purified from the fermentation broth by distillation (Fig. 1). The remaining solid residue is used for cattle feed or, in some cases, as a fuel for boilers. The process for the starch/sugar feedstock can either be a dry mill process, in which the entire feedstock is fed to the fermentation unit, or a wet mill process, in which feedstock components are separated and the starch is fed to the fermentation unit. For the dry milling process, a liquefaction step occurs, in which a caustic solution and enzymes are added in a heated environment to obtain a slurry solution.

Such fermentation processes are often used in geographic locations in which the crops are grown in abundance, due to the low transportation costs of the feedstock supply. Low volumetric productivities and long fermentation times are disadvantages of the processes, which often have a high cost and require federal subsidies.[17] Therefore, to expand ethanol production throughout the United States, alternative raw materials, such as underutilized biomass and low-cost cellulosic feedstock, are being investigated.

Fermentation of Sugars from Lignocellulosic Biomass

Due to abundant quantity and competitive prices, lignocellulosic feedstocks have a greater potential for ethanol production than starch and sugar crops. Importantly, lignocellulosic feedstocks do not interfere with food security[3] and can be cultivated on marginal lands. The most common lignocellulosic materials are corn stover, grasses, wood chips, paper wastes, and agricultural residues. Since cellulose and hemicellulose are complex sugars not directly utilized by yeast or bacteria, they need to be reduced to fermentable sugars. Several processes have been utilized to obtain the fermentable sugars, including: 1) acid hydrolysis followed by fermentation; 2) acid and enzymatic hydrolysis followed by fermentation; 3) physical/chemical disruption and enzymatic hydrolysis followed by fermentation; and 4) pretreatment followed by simultaneous saccharification and fermentation.

Acid hydrolysis followed by fermentation

In the dilute acid process, biomass is treated with dilute sulfuric acid (2–5%) at about 160°C under a pressure of about 10 atm[18] to break down the cellulose and hemicellulose components to fermentable sugars. The lignin fraction is separated from the hydrolysate. The five- and six-carbon sugars are then fermented to ethanol by using genetically engineered organisms such as *Escherichia coli*,[19] *Saccharomyces cerevisiae*,[19] and *Zymomonas mobilis*.[20] A drawback is that this process results in low glucose yields (50–60%) from cellulose and hence gives low ethanol yields.[21] A concentrated acid solution (10–30%) can also be utilized at 100°C and atmospheric pressure to hydrolyze the hemicellulose and cellulose.[22] The benefits of the concentrated acid hydrolysis process over the dilute acid hydrolysis process are the lower operating temperatures and pressures, as well as higher glucose yields. A challenging issue in both the dilute acid hydrolysis and concentrated acid hydrolysis processes is the disposal of lignin, although it can be used as a fuel. Also, a major disadvantage is the formation of toxic by-products, such as furfural and hydroxymethyl furfural, that can affect the conversion rates.

Acid and enzymatic hydrolysis followed by fermentation

For the enzymatic process, the feedstock is first pretreated with a dilute acid to break down the lignocellulose into lignin, hemicellulose, and cellulose. Enzymes called cellulases and xylanases are then used to break down the cellulose and hemicellulose fractions into six- and five- carbon sugars, respectively. Without the dilute acid treatment, the enzymes would not be able to come in efficient contact with the cellulose and hemicellulose. The sugars are then fermented to ethanol using organisms such as *E. coli*,[19] *S. cerevisiae*,[23] and *Z. mobilis*.[20] The pretreatment of lignocellulose with dilute acid followed by enzymatic breakdown of cellulose and hemicellulose has shown to maximize the overall process yields as compared to processing with dilute acid alone. The disadvantage of this method is its high cost and requirement of enzymes,

although the costs have been greatly reduced over the last several years.[24] Also, this process is complicated by the accumulation of soluble products such as cellobiose and cellotriose, which act as competitive inhibitors of hydrolysis.[19]

Physical/chemical disruption and enzymatic hydrolysis followed by fermentation

Several processes, other than dilute acid hydrolysis, have been used to make biomass accessible to enzymatic breakdown for the generation of fermentable sugars. For ammonia disruption, the lignocellulose is exposed to ammonia at a high pressure and a temperature ranging from 25°C to 90°C.[25] The elevated pressure and temperature causes swelling and decrystallization of the biomass. The pressure is then suddenly lowered, causing the biomass to explode. This makes the biomass accessible to enzymes that hydrolyze the cellulose and hemicellulose to fermentable sugars, as noted above. For steam disruption, the biomass is fed into a high-pressure cylinder and treated with steam. The biomass is then passed into a flash tank, causing an explosion due to the sudden pressure change. The explosion causes autohydrolysis of the hemicellulose to xylose. The cellulose is treated with enzymes to form glucose. A disadvantage of the process is that volatile organics (e.g., furfural) are formed, which are toxic to the microbial catalysts. Additional processes that break down the biomass to allow enzymatic treatment include mild alkaline extraction with $Ca(OH)_2$ or NaOH.[26] In all cases, the sugars obtained from the process are fermented to ethanol using the organisms identified in the previous enzymatic method.

Pretreatment followed by simultaneous saccharification and fermentation

To eliminate the separate steps of formation of fermentable sugars followed by fermentation, the simultaneous saccharification and fermentation process has been developed following pretreatment with acid or physical/chemical disruption methods.[27] Since glucose is an inhibitor of cellulase activity, this process effectively removes the glucose and provides higher yields.

Major disadvantages of the processes mentioned include the current need for economic subsidies, the high cost of enzymes, and the formation of waste streams (such as acid pretreatment materials and toxic compounds found in acidic hydrolysates of biomass), although utmost efforts are being made to eliminate these drawbacks.[24,28] When biomass is utilized, lignin (approximately 10–40 wt% of biomass) cannot be broken down into fermentable components. Thus, 10–40% of the biomass is not incorporated into products. However, utilization of the lignin waste stream is currently being explored. Developments in conversion technologies over the last two decades have reduced the cost of bioethanol production from US$ 5.00/gal to approximately US$ 1.20/gal.[29]

BIOMASS TO ETHANOL: FERMENTABLE PRODUCER GAS PLATFORM

A large variety of biomass substrates, including solid municipal waste and waste paper, can be converted to producer gas via gasification. Producer gas can then be converted to ethanol by chemical catalysts, but microbial catalysts offer several advantages, since they require significantly lower temperature and pressure conditions (usually atmospheric conditions) and are less susceptible to varying feed gas compositions. Chemical catalysts are more susceptible to poisoning, and their specificity is lower compared to microbial processes, although faster conversion times are often possible.

A large number of bacterial strains have been isolated that have the ability to ferment producer gas (the CO, CO_2, and H_2 components) to ethanol, acetic acid, and other useful liquid products. *Clostridium ljungdahlii* was the first recognized organism to form ethanol from components of producer gas.[30] The organism favors the production of acetate at a higher pH (5–7), but ethanol is the dominant product at pH between 4 and 4.5. Recently, an additional clostridial acetogen was isolated and was shown to produce ethanol from producer gas generated from biomass.[31,32] Other organisms that can produce ethanol from producer gas, although not as the major product, include *Butyribacterium methylotrophicum*[33] and *Clostridium autoethanogenum*.[34]

Formation of Producer Gas via Gasification

Gasification is a heterogeneous thermochemical process in which a carbon-based feedstock, usually fed in a ground or pelletized form, is converted to a gaseous product by means of a gasification agent. The gaseous product, termed producer gas, consists primarily of CO, CO_2, N_2, methane (CH_4), H_2, and water. Trace amounts of higher hydrocarbons, ash, tar, and other contaminants are also present. Gasification can be classified into indirect and direct gasification. Direct gasification involves the use of an oxidizing agent, either air or oxygen, to partially oxidize the feedstock, while indirect gasification occurs in the absence of oxygen. Steam is a common gasifying agent for indirect gasification. The type of gasifier used depends on many

factors, such as the type and size of biomass, the desired composition of producer gas, and the scale of operation.

Common gasifier configurations include fixed-bed gasifiers, moving-bed gasifiers, and fluidized-bed gasifiers. Though simple and inexpensive to operate, fixed-bed gasifiers have disadvantages such as channeling and limited scale-up capacity. Fluidized-bed gasifiers offer excellent heat and mass transfer rates between biomass and the gas stream and have been used for a variety of feedstocks such as sawdust, sugarcane bagasse, rice husk, woodchips, and solid waste. They offer a reliable alternative for small-particle feedstocks such as switchgrass, which have a tendency to agglomerate in other types of gasifiers. Fluidized-bed gasifiers can also be scaled up easily for industrial operations.[35] Unlike other processes described previously for conversion of biomass to ethanol, in which the lignin fraction of lignocellulose remains unutilized, it is possible to convert all the carbon present in lignocellulosic substrate to a gaseous mixture using gasifiers.

Pathways of Producer Gas Fermentation

For acetogens, such as the ones identified for producer gas fermentation, ethanol synthesis is reported to follow the Wood–Ljungdahl pathway.[36] The overall stoichiometry for the formation of ethanol (C_2H_5OH) from CO, CO_2, and H_2 is

$$6CO + 3H_2O \rightarrow C_2H_5OH + 4CO_2 \quad (1)$$

$$6H_2 + 2CO_2 \rightarrow C_2H_5OH + 3H_2O \quad (2)$$

With CO as the sole substrate carbon source, one-third of the carbon from CO can theoretically be converted to ethanol [Eq. (1)]. Since reducing equivalents are necessary to fuel the reductive metabolic pathway, some of the CO is oxidized to produce reducing equivalents (via carbon monoxide dehydrogenase) and CO_2; the carbon is not available for ethanol formation. Although ethanol can be theoretically produced from CO alone, previous studies have shown that an external supply of CO_2 is required in addition to CO to sustain cell growth and ethanol formation, although CO_2 is generated during the fermentation process.[37]

For a 3:1 H_2/CO_2 ratio, all of the carbon can theoretically be converted to ethanol [Eq. (2)], although such as ratio does not occur with biomass gasification. However, it is important to note that CO_2 can be used by acetogens if H_2 is present. For an equimolar mixture of H_2 and CO, two-thirds of the carbon from CO can theoretically be converted to ethanol as shown in Eq. (3) [a combination of Eqs. (1) and (2)].

$$6CO + 6H_2 \rightarrow 2C_2H_5OH + 2CO_2 \quad (3)$$

In both these latter cases, reducing equivalents are produced from H_2 via the hydrogenase enzyme, resulting in greater carbon conversion to ethanol. Thus, the theoretical yield of carbon in the producer gas towards ethanol production depends upon the composition of the producer gas.

Since biomass is generally of the form $(CH_2O)_n$, the potential exists for a significant amount of H_2 to be formed in addition to the CO and CO_2. According to the biomass structure, the amount of H_2 cannot be in excess of the CO and CO_2 such that the theoretical carbon conversion is not greater than 67% [see Eq. (3)]. Process inefficiencies and reactions in the gasifier, such as the water gas shift reaction and formation of tars, results in lower carbon conversion. However, unlike the fermentation of sugars obtained from biomass, in which the theoretical carbon conversion is 67% of the non-lignin component of biomass (based on glucose to ethanol), the lignin in the gasification process can contribute to the carbon source. It is important to note that optimization issues must still be addressed before comparisons can be made between the gasification–fermentation process and the fermentation process involving biomass-generated sugars.

Fermentation of Producer Gas

Commercialization of producer gas fermentations is currently hindered by low productivity in the bioreactor. Several factors, such as low cell density, lack of regulation of metabolic pathways to yield only the desired product, inhibition of the biological catalysts by products and substrates, and low gas–liquid mass transfer, need to be addressed to establish the economic feasibility of producer gas fermentations.[33]

The transport of gases to the bulk liquid through the liquid film around gas bubbles is the rate-limiting step in most fermentation processes that involve sparingly soluble gaseous substrates. Producer gas fermentations are primary examples of such mass transfer limited fermentations. At mild temperatures, CO and H_2 have aqueous solubilities of 60% and 4%, respectively, compared to oxygen on a mass basis. These low solubilities result in low concentration gradients and, hence, low mass transfer rates.[33] Higher mass transfer rates can be achieved by the use of an agitator system. Increasing the operating pressure can also increase the gas mass transfer rate in producer gas fermentations. Microbubble dispersions, bubbles with

diameters of about 50–100 μm, have been used to provide a large gas transport area at low power consumption.[33]

Continuous stirred-tank reactors (CSTRs) have been routinely employed for producer gas fermentations. A two-stage reactor system has also been used to maximize ethanol production and minimize the formation of byproducts. Carbon monoxide and hydrogen conversions of 90% and 70%, respectively, were observed in the first reactor, while they were about 70% and 10% in the second reactor. High ethanol-to-acetate ratios were achieved by the use of such a dual reactor system.[38] Bubble columns are also commonly used for industrial fermentations. A comparative study was performed between a CSTR and a bubble column reactor for CO fermentation using *Peptostreptococcus productus*. Higher conversion rates of CO were observed with the bubble column without the use of any additional agitation.[39] Producer gas fermentation with packed bubble columns and trickle bed reactors has also been studied. The trickle bed reactor has a low pressure drop and liquid hold-up, and the conversion rates were the highest compared to CSTRs and bubble columns.[40]

The alteration of fermentation conditions, such as pH, drastically affects product concentrations. Research with *C. ljungdahlii* has shown that at high pH values (5.5–6), acetate was the dominant product, while at a lower pH (4–4.5), there was a drastic shift towards the production of ethanol.[41] Inhibition by end products or intermediates is the principal factor that limits metabolic rates and final product concentrations in many fermentation processes. Product inhibition can greatly affect the economics of commercialization. With regards to ethanol inhibition, growth of *B. methylotrophicum* was inhibited at alcohol concentrations of 5 g/L.[42] However, a recently isolated clostridial strain was shown to tolerate ethanol concentrations up to 78 g/L.[31] Efforts have been made to eliminate the drawbacks of inhibition by improvement of bacterial strains to tolerate higher product concentrations and/or by use of novel separation coupled fermentation processes such as pervaporation, extraction, and membrane separation.

CONCLUSIONS

With regards to lignocellulosic processes, producer gas fermentation is still in the early research stages compared to fermentation of sugars from lignocellulose. Both processes have advantages and disadvantages with reference to the process barriers described previously. Unlike the commercialized sugar/starch ethanol process, the lignocellulosic processes are in the early stages of commercialization.[43] It is important to realize that a single process may not be the only solution, and that there is a strong potential for multiple processes to coexist depending upon the biomass feedstock.

The development of new processes for the conversion of low-cost renewable biomass to ethanol is important as a possible alternative to the fast-depleting petroleum resources. Despite the current low utilization of ethanol compared to gasoline, ethanol is fast becoming a choice as an additive to gasoline, considering its positive environmental impact due to lower toxic and greenhouse gas emissions. Although a federal subsidy presently contributes to the success of ethanol as a fuel, its future certainly looks promising. The cost of ethanol could be reduced dramatically if the current research efforts to utilize low-cost biomass are successful. Several technological barriers remain, such as the high cost of enzymes, the inability of microbial strains to simultaneously and efficiently utilize a variety of sugar substrates, and the low ethanol yield in producer gas fermentation. However, with the development of inexpensive routes for production, low-cost ethanol could soon become a reality.

REFERENCES

1. Goldenberg, J.; Johansson, T.B. *World Energy Assessment Overview: 2004 Update*; United Nations Development Programme: New York, 2004; 30.
2. *Ethanol Industry Outlook 2003: Building a Secure Energy Future*; Renewable Fuels Association: Washington, DC, 2003; 1–14.
3. Kim, S.; Dale, B.E. Global potential bioethanol production from wasted crops and crop residues. Biomass Bioenergy **2004**, *26* (4), 361–375.
4. Ahmed, F.E. Toxicology and human health effects following exposure to oxygenated or reformulated gasoline. Toxicol. Lett. **2001**, *123*, 89–113.
5. Dinneen, R. Market for ethanol to replace MTBE is uncertain. Ind. Bioprocess. **2004**, *26* (4), 8–9.
6. Greene, D.L.; Schafer, A. *Reducing Greenhouse Gas Emissions from U.S. Transportation*; Pew Center on Global Climate Change: Arlington, VA, 2003; 1–80.
7. Wang, M.; Saricks, C.; Santini, D. Greenhouse gas emissions of fuel ethanol produced from corn and cellulosic biomass. EM: Air Waste Manage. Assoc. Mag. Env. Managers **1999**, *October*, 17–25.
8. *NREL Public Affairs Online Fact Sheet*, ES-130-24866; National Renewable Energy Laboratory: Golden, CO, 1998.
9. Yacobucci, B.D.; Womach, J. *Fuel Ethanol: Background and Public Policy Issues*, RL 30369;

Congressional Research Service: Washington, DC, 2000.
10. *Ethanol: Separating Fact from Fiction*, DOE/GO-10099-736; U.S. Department of Energy: Washington, DC, 1999.
11. Bungay, H.R. Confessions of a bioenergy advocate. Trends Biotechnol. **2004**, *22* (2), 67–71.
12. McLaughlin, S.B.; Walsh, M.E. Evaluating environmental consequences of producing herbaceous crops for bioenergy. Biomass Bioenergy **1998**, *14* (4), 317–324.
13. Sanderson, M.A.; Reed, R.L.; McLauglin, S.B.; Wullschleger, S.D.; Conger, B.V.; Parrish, D.J.; Wolf, D.D.; Taliaferro, C.; Hopkins, A.A.; Ocumpaugh, W.R.; Hussey, M.A.; Read, J.C.; Tischler, C.R. Switchgrass as a sustainable bioenergy crop. Bioresource Technol. **1996**, *56* (1), 83–93.
14. *Roadmap for Agriculture Biomass Feedstock Supply in the United States*, DOE/NE-ID-11129; U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, Biomass Program: Washington, DC, 2003.
15. McKendry, P. Energy production from biomass (Part 1): Overview of biomass. Bioresource Technol. **2002**, *83* (1), 37–46.
16. Ugarte, D.D.L.; Walsh, M. Synergism between agricultural and energy policy: the case of dedicated bioenergy crops. In Southern Agricultural Economics Association Annual Meeting, Orlando, FL, Feb 2–5, 2002; 1–10.
17. Chum, H.L.; Overend, R.P. Biomass and renewable fuels. Fuel Process. Technol. **2001**, *73* (1–3), 187–195.
18. Iranmahboob, J.; Nadim, F.; Monemi, S. Optimizing acid-hydrolysis: a critical step for production of ethanol from mixed wood chips. Biomass Bioenergy **2002**, *22*, 401–404.
19. Ingram, L.O.; Lai, X.; Moniruzzaman, M.; Wood, B.E.; York, S.W. Fuel ethanol production from lignocellulose using genetically engineered bacteria. In *Fuels and Chemicals from Biomass*, American Chemical Society Symposium Series; Saha, B.C., Woodward, J.E., Eds.; American Chemical Society Press: Washington, DC, 1997; Vol. 666, 57–73.
20. Picataggio, S.K.; Zhang, M.; Eddy, C.; Deanda, K.A.; Finkelstein, M. Recombinant Zymomonas for Pentose Fermentation. U.S. Patent 5,514,583, 1997.
21. Himmel, M.E.; Adney, S.A.; Baker, J.O.; Elander, R.; McMillan, J.D.; Nieves, R.A.; Sheehan, J.J.; Thomas, S.R.; Vinzant, T.B.; Zhang, M. Advanced bioethanol production technologies: a perspective. In *Fuels and Chemicals from Biomass*, American Chemical Society Symposium Series; Woodward, J., Saha, B., Eds.; American Chemical Society Press: Washington, DC, 1997; Vol. 666, 1–45.
22. Clausen, E.C.; Gaddy, J.L. Concentrated sulfuric acid process for converting lignocellulosic materials to sugars. U.S. Patent 5,188,673, 1993.
23. Ho, N.W.Y.; Chen, Z.; Brainard, A.P.; Sedlak, M. Genetically engineered *Saccharomyces* yeasts for conversion of cellulosic biomass to environmentally friendly transportation fuel ethanol. In *Green Chemical Syntheses and Processes*; American Chemical Society Symposium Series; Anastas, P.T., Heine, L.G., Williamson, T.C., Eds.; American Chemical Society Press: Washington, DC, 2000; Vol. 767, 143–159 (Chapter 12).
24. Novozymes Press Release, Novozymes Wins Technology Leadership Award for Biofuels Achievement, 25 July 2005. Novozymes: Bagsvaerd, Denmark.
25. Shleser, R. *Ethanol Production in Hawaii: Processes, Feedstocks and Current Economic Feasibility of Fuel Grade Ethanol Production in Hawaii*, HD9502.5.B54.S4; Hawaii State Department of Business, Economic Development and Tourism: Honolulu, HI, 1994.
26. Reith, J.H.; den Uil, H.; van Veen, H.; de Laat, W.T.A.M.; Niessen, J.J.; de Jong, E.; Elbersen, H.W.; Weusthuis, R.; van Dijken, J.P.; Raamsdonk, L. Co-production of bio-ethanol, electricity and heat from biomass residues. In *12th European Conference and Technology Exhibition on Biomass for Energy, Industry and Climate Protection*; Amsterdam, Netherlands, June 17–21, 2002; ETA-Florence: Amsterdam, Netherlands, 2002 # ECN-RX-02-030.
27. Kadar, Z.; Szengyel, Z.; Reczey, K. Simultaneous saccharification and fermentation (SSF) of industrial wastes for the production of ethanol. Ind. Crops Prod. **2004**, *20* (1), 103–110.
28. Weil, J.R.; Dien, B.; Bothast, R.; Hendrickson, R.; Mosier, N.S.; Ladisch, M.R. Removal of fermentation inhibitors formed during pretreatment of biomass by polymeric adsorbants. Ind. Eng. Chem. Res. **2002**, *41* (24), 6132–6138.
29. Wyman, C.E. Potential synergies and challenges in refining cellulosic biomass to fuels, chemicals, and power. Biotechnol. Prog. **2003**, *19* (2), 254–262.
30. Phillips, J.R.; Clausen, E.E.; Gaddy, J.L. Synthesis gas as substrate for the biological production of fuels and chemicals. Appl. Biochem. Biotechnol. **1994**, *45–46*, 145–157.
31. Liou, S.; McGuire, J.E.; Tanner, R.S. Conversion of synthesis gas to ethanol by clostridial strain P7. In *American Society of Microbiology*. Proceedings, Salt Lake City, UT, May 20; Salt Lake City, UT, 2002; O 10–11.

32. Datar, R.P.; Shenkman, R.M.; Cateni, B.G.; Huhnke, R.L.; Lewis, R.S. Fermentation of biomass-generated producer gas to ethanol. Biotechnol. Bioeng. **2004**, *86* (5), 587–594.
33. Bredwell, M.D.; Srivastava, P.; Worden, R.M. Reactor design issues for synthesis-gas fermentations. Biotechnol. Prog. **1999**, *15* (5), 834–844.
34. Abrini, H.; Naveau, H.; Nyns, E.J. Clostridium autoethanogenum, sp. nov., an anaerobic bacterium that produces ethanol from carbon monoxide. Arch. Microbiol. **1994**, *161*, 345–351.
35. Bridgewater, A.V. Renewable fuels and chemicals by thermal processing of biomass. Chem. Eng. J. **2003**, *97*, 87–102.
36. Ljungdahl, L.G. The autotrophic pathway of acetate synthesis in acetogenic bacteria. Ann. Rev. Microbiol. **1986**, *40*, 415–440.
37. Rajagopalan, S.; Datar, R.P.; Lewis, R.S. Formation of ethanol from carbon monoxide via a new microbial catalyst. Biomass Bioenergy **2002**, *23*, 487–493.
38. *Bench-Scale Demonstration of Biological Production of Ethanol from Coal Synthesis Gas*, DE-AC22-92PC92118; U.S. Department of Energy: Washington, DC, 1994.
39. Vega, J.L.; Clausen, E.C.; Gaddy, J.L. Design of bioreactors for coal synthesis gas fermentations. Resources Conserv. Recycl. **1990**, *3*, 149–160.
40. Klasson, K.T.; Ackerson, M.D.; Clausen, E.C.; Gaddy, J.L. Bioconversion of synthesis gas into liquid or gaseous fuels. Enz. Microb. Technol. **1992**, *14* (8), 602–608.
41. Klasson, K.T.; Ackerson, M.D.; Clausen, E.C.; Gaddy, J.L. Biological conversion of coal and coal-derived synthesis gas. Fuel **1993**, *72* (12), 1673–1678.
42. Worden, R.M.; Grethlein, A.J.; Zeikus, J.G.; Datta, R. Butyrate production from carbon monoxide by *Butyribacterium methylotrophicum*. Appl. Biochem. Biotechnol. **1989**, *20–21*, 687–698.
43. Badger, P.C. Ethanol from cellulose: a general review. In *Trends in New Crops and New Uses*; Janick, J., Whipkey, A., Eds.; ASHS Press: Alexandria, VA, 2002; 17–21.

Biomaterials

Sujata K. Bhatia
Dupont Central Research and Development, Wilmington, Delaware, U.S.A.

Surita R. Bhatia
Department of Chemical Engineering, University of Massachusetts–Amherst, Amherst, Massachusetts, U.S.A.

INTRODUCTION

Biomaterials science is a multidisciplinary endeavor incorporating chemical engineering, medicine, biology, chemistry, materials science, bioengineering, and biomechanics. The past few years have witnessed an explosion in the field of biomaterials, with an expansion of both the compositions and the applications of medical implant materials. As the prevalence of chronic diseases such as diabetes, cardiovascular disease, and neurodegenerative disease increases, there will be an even greater need for innovative biomaterials. This entry reviews the current status of the field of biomaterials, and highlights new developments in biomaterials. The entry will provide an overview of medical applications of biomaterials, and will describe current classes of biomaterials, including metals, ceramics and glasses, and polymeric materials. The entry will then discuss the next generation of biomaterials, including surface-modified biomaterials, smart biomaterials, bioactive materials, biomimetic materials, patterned biomaterials, and tissue engineering and regenerative medicine.

DEFINITIONS

A commonly used definition of a biomaterial, endorsed by a consensus of biomaterials experts, is "a nonviable material used in a medical device, intended to interact with biological systems."[1] An essential characteristic of biomaterials is biocompatibility, defined as "the ability of a material to perform with an appropriate host response in a specific application."[1] The goal of biomaterials science is to create medical implant materials with optimal mechanical performance and stability, as well as optimal biocompatibility.

OVERVIEW OF BIOMATERIALS APPLICATIONS

Biomaterials are used in diverse clinical applications. Table 1 lists several examples of applications of biomaterials in medicine.[2] Note that metals, ceramics, polymers, glasses, carbons, and composite materials are listed.

TYPES OF BIOMATERIALS

Metals

Metals and alloys have long been used in surgical and dental applications where materials with high strength are required. Metals are excellent for providing specific mechanical properties, including strength and ductility; however, corrosion of metallic implants in biological environments remains a concern. Corrosion not only limits device lifetime but also causes release of toxic metal ions that are often carcinogenic or mutagenic.[3,4] Thus, much of the current research focuses on minimizing and reducing corrosion of metallic biomaterials. Blackwood has recently reviewed common types of corrosion encountered in metal implants in vivo, as well as physiological parameters relevant to corrosion.[3] For surgical implants, relevant parameters are chloride content, pH, and dissolved oxygen levels in blood.[3] In vitro tests of corrosion resistance are typically performed in aqueous solutions containing 0.9% NaCl, with a pitting resistance number greater than 26 desirable for implanted materials.[3] Differences between in vitro and in vivo response are often attributed to dissolved oxygen content, sulfur-containing amino acids present in blood, and pathological changes associated with implantation such as generation of hydrogen peroxide and lowered pH (as low as 4) at the implant site.[3] Corrosion is an even greater concern in dental applications, because of the high acidity and chloride ion levels in many foods. Additionally, the corrosiveness of saliva is highly dependent on oral hygiene.[3] Thus, while reasonable in vitro models for saliva are available, it is difficult to predict in vivo corrosive resistance of dental materials.

Most metallic biomaterials fall into one of four categories: stainless steels, titanium and titanium-based alloys, cobalt–chromium alloys, and amalgams.[3] Additionally, research is under way on a number of

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120040342

Table 1. Examples of clinical applications of biomaterials

Application	Types of materials
Orthopedic	
Joint replacements (hip, knee)	Titanium, Ti–Al–V alloy, stainless steel, polyethylene
Bone plate for fracture fixation	Stainless steel, cobalt–chromium alloy
Bone cement	Poly(methyl methacrylate)
Bony defect repair	Hydroxyapatite
Artificial tendon and ligament	Teflon™, Dacron™
Cardiovascular	
Blood vessel prosthesis	Dacron™, Teflon™, polyurethane
Heart valve	Reprocessed tissue, stainless steel, carbon
Catheter	Silicone rubber, Teflon, polyurethane
Pacemaker	Polyurethane, silicone rubber, platinum electrodes
Ophthalmologic	
Intraocular lens	Poly(methyl methacrylate)
Contact lens	Silicone-acrylate, hydrogel
Corneal bandage	Collagen, hydrogel
Dental	
Dental implant for tooth fixation	Titanium, alumina, calcium phosphate
Neurologic	
Cochlear implant	Platinum electrode
General surgery	
Skin repair template	Silicone–collagen composite
Sutures	Silk, nylon, poly(glycolide-*co*-lactide)
Adhesives and sealants	Cyanoacrylate, fibrin
Organ replacement	
Heart–lung machine	Silicone rubber
Artificial kidney (hemodialyzer)	Cellulose, polyacrylonitrile
Artificial heart	Polyurethane

(Adapted from Ref.[2].)

next-generation metallic biomaterials, including rare earth materials and shape-memory alloys. Of the stainless steels, type 304 had been used previously in medical applications, but problems with localized corrosion and tumor formation were sometimes reported.[3] Type 316L SS is currently the most widely used in biomedical applications.[5] This is an iron–chromium–nickel alloy with a low carbon content, where chromium provides corrosion resistance. The resistance to pitting corrosion can be improved if nitrogen additions are made.[3] More recently, additional corrosion-resistant stainless steels have been developed, including 316LVM grade with a typical composition of 18Cr14Ni3Mo. For dental materials, which must be extremely corrosion resistant, ultraclean high nitrogen austenitic stainless steels are recommended, such as 21Cr10Ni3Mo0.3Nb0.4N.[3] Finally, mixing different grades of stainless steels is not recommended, as this can lead to galvanic corrosion and failure. In general, stainless steels display better mechanical and formability properties but worse corrosion resistance than titanium-based alloys. Release of chromium presents a concern, although the levels of chromium are lower in stainless steels than in cobalt–chromium alloys.[3]

Titanium displays excellent biocompatibility and corrosion resistance; however, it does not have the high strength necessary for several biomedical applications.[3,5] The most popular material for load-bearing orthopedic applications is Ti6Al4V, a dual-phase alloy comprising an Al-stabilized η-phase and a V-stabilized β-phase.[5] Other alloys in use for medical applications include Ti2.5Al2.5Fe, Ti6Al7Nb, and Ti50Ta.[3,5] While all these alloys exhibit a higher tensile strength than titanium, their corrosion resistance is not as high.[3] They also display poor shear strength and thus should not be used for applications such as screws.[3]

Common cobalt–chromium alloys include CoCrMo, used in dentistry and artificial joint applications, and CoCrNiMo, used as a part of replacements for heavily loaded joints because of its very high tensile strength.[3] Other cobalt–chrome biomaterials include CoCrWNi, MP35N, and ASTM F1058 (40Co12Cr15Ni7Mo).[3,5] In all these materials, chromium is present at fairly high levels. Thus, the release of chromium upon corrosion is a concern, as chromium is a carcinogen.[3]

Amalgams are typically used in dental applications and are multiphase alloys. Silver–tin amalgam carries a risk of mercury release through corrosion of the

Sn_7Hg γ_2 phase.[3] Newer high copper amalgams reduce the risk of mercury release, as preferential corrosion of the η' phase, Cu_6Sn_5, typically occurs.[3] However, release of mercury can still occur even in these materials. Older silver–tin amalgams are based on a silver–tin alloy, while high copper amalgams are based on either a silver–copper–tin alloy or a mixture of silver–tin and silver–copper alloys.[3]

Next-generation metallic biomaterials include porous titanium alloys and porous CoCrMo with elastic moduli that more closely mimic that of human bone; nickel–titanium alloys with shape-memory properties for dental braces and medical staples; rare earth magnets such as the NdFeB family for dental fixatives; and titanium alloys or stainless steel coated with hydroxyapatite for improved bioactivity for bone replacement.[3,5,6] The corrosion resistance, biocompatibility, and mechanical properties of many of these materials still must be optimized; for example, the toxicity and carcinogenic nature of nickel released from NiTi alloys is a concern.[3]

Ceramics and Glasses

Many ceramic materials possess improved biocompatibility as compared to metals, and corrosion is typically not an issue. Ceramics often have high strength but display brittleness, poor crack resistance, and low ductility.[7] Several ceramic materials are bioinert, bioactive (forming bonds with the surrounding tissue such as bone), or bioresorbable (as in the case of some porous ceramics).[7]

Arguably, the most important ceramic biomaterial is hydroxyapatite (HAP), $Ca_{10}(PO_4)_6(OH)_2$, a synthetic analog of bone mineral.[7] Natural bone is a composite comprising small crystalline HAP platelets bonded to collagen.[7] Synthetic hydroxyapatite is known to be bioactive, forming a strong bond with adjacent bone tissue and inducing bone growth along the interface (termed osteoconduction).[6] Hydroxyapatite and its derivatives are used most often in orthopedic and dental applications, for applications such as repair of bone defects and tooth root implants.[7] As mentioned above, HAP powders and coatings are also sometimes used in conjunction with metallic implants to induce adhesion with the surrounding tissue and to promote bioactivity.[6] Solid HAP is very robust in physiological environments and can remain in the body for 5–7 yr, while porous HAP can be resorbed by the body after approximately 1 yr.[7] Over 30 yr of research and clinical practice suggest that HAP and its related compounds are generally nontoxic and produce little or no inflammation or foreign body response. There have been some unfavorable biological reactions reported with porous HAP; these have been attributed to irritation from sharp edges of the implant and micromovement of the implant.[7]

Materials derived from HAP include Mg-HAP, which has been investigated for bone restoration in osteochondrosis; carbonate-HAP, with potential applications in the deposition of calcium and in some root canal fillings; silver-HAP, of interest for infected bone defects; and fluorine-HAP, of interest in the treatment of tooth defects and as coatings for metallic biomaterials.[7] Closely related materials include tricalcium phosphate, $Ca_3(PO_4)_2$, and certain calcium-phosphate glasses (described below), which form microcrystals of HAP on their surfaces in vivo, and thus exhibit similar degrees of bonding to bone tissue as HAP and its derivatives.[7] Additionally, tricalcium phosphate and HAP can both be combined with glasses to optimize the resorption time in vivo or to tune the bioactivity. Many of these so-called ''glass-ceramics'' typically contain SiO_2 in concentrations from 12% to 50% by mass, and may also contain CaO, Na_2O, or P_2O_5; or less commonly K_2O, MgO, CaF_2, Al_2O_3, and B_2O_3. Glass-ceramic implants with a lower calcium concentration may be resorbed in the body in as little as 10–30 days.[7] Taken together, HAP, tricalcium phosphate, and related materials have been investigated for nearly every conceivable type of bone reconstruction, including skull restoration after surgery or trauma, tooth-root implants, and tooth fillings, maxillofacial reconstruction, joint reconstruction, repair of load-bearing skeletal elements, grafting and stabilizing skull bone, repair of alveolar clefts and augmentation of alveolar ridges, cervical spine fusion, and reconstructive surgery of the middle ear.[7] Additionally, because glass-ceramics containing approximately 45% SiO_2 display strong bonding with soft tissue, these materials are of interest for the restoration of tendons, ligaments, small blood vessels, and nerve fibers; as well as catheters for infected wounds and as microsurgical joints.[7] Finally, HAP is used as a component of composite materials for bone replacement applications, as reviewed recently.[8,9] Examples include collagen–HAP composites and polyethylene–HAP composites.

Another important class of ceramic biomaterials is based on Al_2O_3 (aluminum oxide or corundum). These materials are bioinert and are widely used for the knob of replacement hip joints and other joint endoprostheses.[7] Other uses include tooth-root implants, parts for osteosynthesis, maxillofacial implants, and skull surgery.[7] Aluminum oxide ceramics prepared for use as biomaterials must have a low concentration of impurities such as silicon oxide and alkaline earth oxides, which can interfere with the bioinertness of the implant.[7]

A third class of bioceramics are based on ZrO_2, stabilized by Y_2O_3 or CeO_2. These materials are close to aluminum oxide materials in terms of biocompatibility but exhibit a higher bend strength and crack resistance, though with lower compressive strength.[7] Zirconium dioxide ceramics can be used for many of

the same types of biomedical applications as aluminum oxide and is in fact replacing aluminum oxide in many applications.

Glass-based biomaterials have been reviewed recently by Hench, Xynos, and Polak; Stroganova, Mikhailenko, and Moroz; and Knowles.[6,10,11] As alluded to above, biocompatible glasses are often used in conjunction with ceramics or alone for applications involving bone and joint reconstruction. Bioactive glasses are more quickly dissolved and resorbed into the body than bioceramics; furthermore, the dissolution time can be tuned using a variety of techniques.[11] Much of the research in this area has involved the ternary Na_2O–CaO–P_2O_5 system, which is of interest for both tissue engineering and antibacterial applications.[11] Calcium phosphate glasses with additional components, such as SiO_2, K_2O, MgO, and Al_2O_3, have also been investigated for bone restoration applications.[10]

Polymers

The literature on polymeric biomaterials is vast. Polymeric materials typically are easier to process than metals or ceramics, and often have lower strength and moduli. Their microstructure and transport properties can often be tuned by varying processing conditions, polymer molecular weight, cross-link density, and environmental conditions (temperature, pH, ionic strength, etc.). For these reasons, they are used more often for soft tissue engineering and drug delivery applications. Polymer-based materials with biomedical applications include hydrogels, temperature, and pH-responsive gels, conventional thermoplastics, block copolymers, polysaccharides, and artificial proteins, to name a few. Several of these are described later in this entry (see sections Smart Biomaterials and Bioactive Biomaterials below) as well as in other entries in this Encyclopedia on hydrogels, functional biomaterials, and so on. To avoid repetition with these entries, here we summarize the major biomedical applications of different classes of polymers, and refer the reader who wishes to have more information to recent reviews as well as related entries in this Encyclopedia. Table 2 lists some important classes of polymeric materials along with their main biomaterials applications.[12]

THE NEXT GENERATION OF BIOMATERIALS

Surface-Modified Biomaterials

While traditional biomaterials have saved millions of lives, these materials still suffer from problems with infection, thrombosis, inflammation, and poor healing with resultant fibrous encapsulation of biomaterials. Implanted biomaterials induce an inflammatory foreign body reaction that prevents normal wound healing; the foreign body response may result from nonspecific protein adsorption to biomaterial surfaces. One approach to overcoming these problems is to engineer the tissue-biomaterial interface by modifying the biomaterial surface.[13] Biomaterial surfaces may be engineered to create nonfouling, or stealth, surfaces that inhibit protein adsorption. Poly(ethylene glycol) (PEG) has been attached to biomaterials to create nonfouling materials that repel protein adsorption and cell attachment in vitro; PEG-modified surfaces have been found to resist cell adhesion for up to 2 weeks in vitro. While PEG is effective, its nonfouling properties are dependent on surface chain density, and it is easily damaged by oxidants. Poly(ethylene glycol) oligomers in self-assembled monolayers have been applied to create precision surfaces that are highly protein resistant.[14] Additionally, PEG-like surfaces have been prepared by plasma deposition of tetraethylene glycol dimethylether (tetraglyme) to form a highly nonfouling cross-linked structure; these surfaces are resistant to protein adsorption, as well as adhesion of platelets, monocytes, endothelial cells, and bacteria in vitro. In addition to PEG and PEG-like surfaces, nonfouling surfaces have been constructed using phospholipids, including phosphatidyl choline, and saccharides. In general, a number of different strategies have been successful in reducing nonspecific protein adsorption to biomaterials in vitro, but this has not yet translated into success in vivo. For example, tetraglyme-coated implants still induce fibrous capsule formation when implanted subcutaneously in mice, and tetraglyme-treated implants exhibit significantly higher macrophage adhesion than untreated implants in vivo.[13] Better model systems that accurately capture the in vivo environment will be required for the development of nonfouling surfaces for implanted biomaterials.

Smart Biomaterials

In numerous medical device applications, it may be desirable to have biomaterials that can respond to changes in the surrounding environment. Environmental triggers can be used in implanted biomaterials to activate or deactivate drug delivery, cell attachment, or a change in mechanical properties. Smart biomaterials have been designed that are sensitive to changes in pH, temperature, and other physical and chemical stimuli.[15] Hydrogels that exhibit pH-dependent swelling behavior can be created from ionic networks. For example, hydrogels prepared from poly(methacrylic acid) grafted with poly (ethylene glycol) (PMAA-g-PEG) shrink at pH 2 because of formation of interpolymer complexes, but swell 3–25

Table 2 Important classes of polymeric biomaterials

Polymer	Main biomaterials applications
Proteins and protein-based polymers	
Collagen	Soft tissue engineering and implants, absorbable sutures, wound dressings, drug delivery
Albumin	Cell encapsulation and drug delivery
Poly(amino acids)	Oligomeric drug carriers, polyelectrolyte complexes for cell encapsulation
Polysaccharides and derivatives	
Carboxymethyl cellulose	Drug delivery, dialysis membranes, polyelectrolyte complexes for cell encapsulation, and cell immobilization
Cellulose sulfate	Complexes for cell encapsulation
Agarose	Supporting material for clinical analysis, cell immobilization
Alginate	Cell encapsulation and immobilization, immobilization of enzymes, controlled release, injectable microcapsules
Carrageenan	Microencapsulation, thermoreversible gelation
Hyaluronic acid	Lubrication applications
Heparin and heparin-like glycosaminoglycanes	Antithrombotic and anticoagulant properties; used in surgery
Dextran and derivatives	Drug delivery
Chitosan and derivatives	Gels, membranes, and microspheres for drug delivery; polyelectrolyte complexes for encapsulation
Aliphatic polymers	
Poly(lactic acid), poly(glycolic acid), and their copolymers	Biodegradable sutures, drug delivery systems, and tissue engineering scaffolds
Poly(hydroxy butyrate), poly(caprolactone), and their copolymers	Biodegradable matrices for controlled release and cell encapsulation
Polyamides	Sutures, dressings, hemofiltration membranes
Polyanhydrides	Biodegradable tissue engineering scaffolds and devices
Poly(ortho esters)	Surface-eroding materials for sustained release and ophthalmology
Poly(cyano acrylates)	Biodegradable surgical adhesives and glues
Polyphosphazenes	Drug delivery, hydrogels, and thin films
Thermoplastic polyurethanes	Permanent implants, prostheses, vascular grafts, catheters, and drug delivery devices
Polyethylene	Sutures, catheters, membranes
Poly(vinyl alcohol)	Gels and membranes for drug delivery and cell encapsulation
Poly(ethylene oxide) and copolymers	Used to render surfaces biocompatible and resistant to protein adhesion; copolymers with poly(propylene oxide) form thermoreversible gels for drug delivery
Poly(hydroxyethyl methacrylate)	Hydrogels for soft contact lenses, drug delivery, skin coatings
Poly(methyl methacrylate)	Dental implants, bone replacement
Polytetrafluoroethylene	Vascular grafts, clips, and sutures
Polydimethylsiloxanes	Implants in plastic surgery and orthopedics blood bags and pacemakers
Poly(vinyl methyl ether)	Temperature-sensitive materials; shape-memory materials
Poly(*N*-alkylacrylamides)	Temperature-sensitive gels

(Adapted from Ref.[12].)

times in size at a physiological pH 7.[16] Ionic strength-dependent and pH-dependent swelling has also been observed in gels of PMAA or poly(acrylic acid) with poly(hydroxyethyl methacrylate).[15] These gels can be loaded with drug and will trap the drug at low pH, then swell to release the drug at a high pH. Such gels may be promising for oral delivery of biopharamaceuticals, as the gel will protect the drug from the acidic pH of the stomach for subsequent successful delivery in the small intestine. Insulin-loaded PMAA-g-PEG gels have been

administered orally to diabetic mice, and the glucose levels of the mice decreased following gel administration, suggesting that the gel successfully protected and delivered the protein therapeutic.[17]

In addition to pH-sensitive gels, thermally sensitive biopolymers have been investigated; such materials can respond to changes in temperature from room temperature to body temperature. These hydrogels exhibit a lower critical solution temperature (LCST); the gels expand when cooled below the LCST and contract when heated above the LCST. Hydrogels of poly(N-isopropylacrylamide) (pNIPAAm) have an LCST of 32°C and have been studied extensively for drug release applications.[15] Temperature-modulated drug release from pNIPAAm hydrogels can be achieved via bulk squeezing: drug that is distributed inside the matrix is squeezed out when the hydrogel contracts with heating above the LCST. Another application of pNIPAAm hydrogels is the creation of surfaces for cell culture systems. Coatings of pNIPAAm are hydrophilic below the LCST and hydrophobic above the LCST. Because cells attach on hydrophobic surfaces and detach from hydrophilic surfaces, cells cultured on pNIPAAm surfaces can be nontraumatically lifted as intact sheets from these surfaces, simply by lowering the temperature.

Polymers of elastin-like peptide (ELP) also exhibit temperature-sensitive behavior.[13] The LCST of ELP can be controlled by varying the length of the ELP molecule or its amino acid composition to obtain transition temperatures ranging from 30°C to 90°C. The ELPs have been used for chondrocyte encapsulation and as drug carriers for cancer therapy. An additional thermoresponsive biomaterial is the biodegradable triblock copolymer poly(ethylene glycol)–poly(lactic acid-co-glycolic acid)–poly(ethylene glycol) (PEG–PLGA–PEG). This copolymer exhibits a sol-to-gel transition with increases in temperature.[15] This property may be important for drug delivery applications, as PEG–PLGA–PEG can be injected as a free-flowing solution at room temperature and then becomes a gel upon reaching body temperature. Hydrogels with certain compositions may demonstrate both pH- and temperature sensitivity. For example, hydrogel copolymers of MAA and NIPAAm sense small changes in pH and temperature. Controlled release of antithrombotic agents, including streptokinase and heparin, has been demonstrated with these dual-sensitivity hydrogels.[15]

Other stimuli in addition to pH and temperature have been investigated to create smart biomaterials. These include external physical stimuli such as light, magnetic fields, electric current, and ultrasound. Specific biochemical stimuli also may be used: calcium-responsive hydrogels, antigen-responsive hydrogels, and microbial infection-responsive hydrogels have all been designed for drug delivery applications.[14]

Bioactive Biomaterials

Several advances have been made in rendering materials biologically active.[18] Bioactivity may be used to impart pharmacological activity to a biomaterial, to modify the biocompatibility of a material, or to tune the lifetime and degradation of a biomaterial. One method for creating bioactive biomaterials is to incorporate pharmacological agents into materials. Polymeric implants that release therapeutic drugs are already in clinical use; examples include progesterone-releasing implants for fertility regulation, and lupron-releasing implants for prostate cancer treatment. Polymer-coated arterial stents that elute sirolimus and tacrolimus have revolutionized the treatment of coronary artery disease, by preventing restenosis following stent placement. Site-specific delivery of chemotherapeutics for brain cancer treatment has been achieved using polyanhydride disks loaded with carmustine (BCNU).[19] Biological growth factors may also be incorporated into materials, either through surface display or through controlled release systems. Polypeptide growth factors may regulate a variety of cellular responses, including cell migration, proliferation, survival, and differentiation. Growth factors have been exploited to create biomaterials that deliver angiogenic growth factors to induce vascular repair; neuronal survival and differentiation factors to treat neurodegenerative disease; transforming growth factor β to induce bone repair; bone morphogenetic protein-4 to enhance bone formation; and tissue growth factors to heal chronic ulcers.

Another approach to creating biological activity in biomaterials is to incorporate adhesion factors, including adhesion-promoting oligopeptides or oligosaccharides.[18] While traditional biomaterials promote cell adhesion via nonspecific adsorption of proteins, a greater degree over cell migration, cell adhesion, and cell-type selectivity can be achieved by incorporating adhesion-promoting factors directly into biomaterials. For example, the RGD tripeptide from fibronectin may be either immobilized on biomaterial surfaces or included directly into the backbone of polymer chains to induce cell adhesion, spreading, focal contact formation, and cytoskeletal organization. The YIGSR domain and the SIKVAV domain from laminin are migration-promoting peptides, and have been incorporated into gels to promote neuronal cell infiltration and nerve regeneration. The tetrapeptide REDV from fibronectin may be employed in vascular grafts to specifically support adhesion and migration of vascular endothelial cells, while also preventing the adhesion of clot-forming platelets. The protein osteopontin has been immobilized on poly(2-hydroxyethyl methacrylate) surfaces to promote endothelial cell adhesion.[13] These approaches to biomaterial modification are clinically relevant, as it has been shown in clinical studies that enhancement of cell

adhesion strength improves the performance of endothelial cell-seeded vascular grafts in high-flow regions.[20]

Still another method for engineering bioactive biomaterials is to design materials containing enzymatic recognition sites.[18] Incorporation of enzymatic cleavage sites into a biomaterial allows the degradation rate of the material to be tuned, and also allows the biomaterial to be proteolytically remodeled. For example, gels containing PEG chains with central oligopeptide sites that are substrates for collagenase or plasmin are degradable by cell-associated enzymatic activity.

Tissue Engineering and Regenerative Medicine

The goal of tissue engineering is to replace lost tissues or organs with polymer constructs that contain specific populations of living cells. Traditional strategies for replacing lost tissue include organ transplantation and mechanical device implantation; however, organ transplantation is limited by donor organ shortages and transplant rejection, and mechanical devices have limited durability and limited bioactivity. Tissue engineering has the potential to overcome these limitations by creating functional tissues that have the capacity for growth, remodeling, and self-repair. The general approach to creating engineered tissue is to harvest specific cell populations from the tissue of interest, seed the cells into a biodegradable polymer scaffold, and cultivate the cell/polymer construct in a bioreactor prior to implantation.[21] Upon implantation, the biodegradable polymer will degrade and gradually be replaced by regenerated tissue.

Successful tissue engineering requires appropriate selection of cells, polymers, and bioreactor conditions for the application of interest. Cells used in tissue engineering may be drawn from either primary tissues or cell lines; the bulk of tissue engineering experiments have utilized primary autogenous cells. Scaffold polymers for tissue engineering must be biodegradable, biocompatible, and readily processible into appropriate anatomical shapes. The most commonly used scaffold polymers are poly(glycolic acid) (PGA), poly(lactic acid) (PLA), and their copolymers (PLGA). Polyanhydrides, polycarbonates, and polyurethanes have also been investigated, and hydrogels, particularly algal polysaccharides, have been utilized as cell delivery matrices. Bioreactors for tissue engineering are designed to allow optimal conditioning of the cell/polymer construct to initiate tissue formation prior to implantation. It is desirable to achieve a high and uniform cell density throughout the polymer scaffold. The bioreactor is a dynamic tissue culture environment where gas and nutrient exchange are augmented by constant turnover of tissue culture medium, and where tissue-specific mechanical forces (stretch, pressure, shear forces) are recapitulated.[22]

The tissue engineering approach has been applied to many tissues including skin, cartilage, bone, liver, intestine, urologic tissue, cardiovascular tissue, and neural tissue.[19,22] Tissue engineered skin regeneration systems are already in clinical use to repair burns, wounds, and chronic ulcers. In one approach, neonatal dermal fibroblasts are placed on PLGA scaffolds and grown into sheets to create skin.[19] In another approach, a bilayer system is used, in which a lower dermal layer of human fibroblasts and an upper epidermal layer of human keratinocytes are seeded onto a bovine collagen matrix. Cartilage regeneration has been achieved by delivering chondrocyte suspensions to focal articular cartilage defects; an autologous chondrocyte product has been FDA approved for clinical application.[22] Cartilage tissue has also been engineered in the configuration of the ear, nasal septum, and trachea using PGA and PLGA constructs seeded with chondrocytes; such constructs have also been shown to close full-thickness cranial defects in animals. Clinical trials are under way to use chondrocyte/polymer constructs for cartilage replacement in humans. In the area of bone replacement, PGA meshes seeded with periosteal cells have been shown to generate new bone in animal models. For replacement of liver function, hepatocytes are seeded on PGA sheets; implantation of these sheets into liver enzyme-deficient animals results in partial correction of the enzyme deficiency. Copolymers of PGA and PLA scaffolds fabricated to contain vascular-like channels and seeded with cocultures of hepatocytes and endothelial cells demonstrate remodeling and formation of vascular channels in vitro, suggesting that complex tissue architecture can be achieved with an appropriate cell/polymer construct. For intestinal replacement, mixed enterocyte/stromal cell populations have been seeded onto PGA meshes; such constructs demonstrate formation of structures resembling intestinal villi and crypts when implanted in animals. In the urologic system, functional ureters have been created using tubular PGA constructs seeded with urothelial cells and smooth muscle cells, and functional bladder tissue has been created from PGA sheets seeded with urothelium and smooth muscle cells. For cardiovascular tissue engineering, endothelial cells have been seeded on both tubular polymer constructs to create blood vessels and leaflet-like constructs for heart valve reconstruction; tissue engineered heart valves are functional in large animals.[22,23] For neural regeneration, an electrically conducting polymer, polypyrrole, has been shown to provide a substrate for nerve regrowth in animal models.

CONCLUSIONS

Biomaterials have improved the lives of millions by providing material solutions to biomedical problems.

Biomaterials have been applied in a variety of clinical disciplines, including cardiovascular medicine, orthopedics, and ophthalmology, and new materials are in use or under development for virtually every organ system in the body. Traditional biomaterials have been designed from polymers, ceramics, and metals. The next generation of biomaterials will incorporate biomolecules, therapeutic drugs, and living cells. Innovative new biomaterials, including surface-modified biomaterials, smart biomaterials, bioactive biomaterials, and tissue-engineered materials, will have improved properties of biocompatibility, tunability, and biological functionality. Successful development of new biomaterials will require an increased understanding of cell–material interactions, as well as better model systems for the biological environment.

ARTICLES OF FURTHER INTEREST

Biomolecular Engineering, p. 171.
Functional Biomaterials, p. 1099.
Hydrogels, p. 1307.
Tissue Engineering, p. 3115.

REFERENCES

1. Williams, D.F. Definitions in biomaterials. Proceedings of a Consensus Conference of the European Society for Biomaterials, Elsevier: New York, 1987.
2. Ratner, B.D.; Hoffman, A.S.; Schoen, F.J.; Lemons, J.E. Biomaterials science: an introduction to materials in medicine. Academic Press: New York, 1996.
3. Blackwood, D.J. Biomaterials: past successes and future problems. Corros. Rev. **2003**, *21* (2–3), 97–124.
4. Krug, H.F. Metals in clinical medicine: the induction of apoptosis by metal compounds. Mat. Wiss. Werkstofftech. **2002**, *33*, 770–774.
5. Kannan, S.; Balamurugan, A.; Rajeswari, S.; Subbaiyan, M. Metallic implants—an approach for long term applications in bone related defects. Corros. Rev. **2002**, *20* (4–5), 339–358.
6. Hench, L.L.; Xynos, I.D.; Polak, J.M. Bioactive glasses for in situ tissue regeneration. J. Biomater. Sci. Polym. Ed. **2004**, *15* (4), 543–562.
7. Dubok, V.A. Bioceramics—yesterday, today, tomorrow. Powder Metall. Met. Ceram. **2000**, *39* (7–8), 381–394.
8. Mano, J.F.; Sousa, R.A.; Boesel, L.F.; Neves, N.M.; Reis, R.L. Bioinert, biodegradable and injectable polymeric matrix composites for hard tissue replacement: state of the art and recent developments. Compos. Sci. Technol. **2004**, *64*, 789–817.
9. Kikuchi, M.; Ikoma, T.; Itoh, S.; Matsumoto, H.N.; Koyama, Y.; Takakuda, K.; Shinomiya, K.; Tanaka, J. Biomimetic synthesis of bone-like nanocomposites using the self-organization mechanism of hydroxyapatite and collagen. Composites Sciences and Technology **2004**, *64*, 819–825.
10. Stroganova, E.E.; Mikhailenko, N.Y.; Moroz, O.A. Glass-based biomaterials: present and future (a review). Glass Ceram. **2003**, *60* (9–10), 315–319.
11. Knowles, J.C. Phosphate based glasses for biomedical applications. J. Mater. Chem. **2003**, *13*, 2395–2401.
12. Angelova, N.; Hunkeler, D. Rationalizing the design of polymeric biomaterials. Trends Biotechnol. **1999**, *17*, 409–421.
13. Ratner, B.D.; Bryant, S.J. Biomaterials: where we have been and where we are going. Annu. Rev. Biomed. Eng. **2004**, *6*, 41–75.
14. Ratner, B.D. Reducing capsular thickness and enhancing angiogenesis around implant drug release systems. J. Controlled Release **2002**, *78* (1–3), 211–218.
15. Peppas, N.A.; Bures, P.; Leobandung, W.; Ichikawa, H. Hydrogels in pharmaceutical formulations. Eur. J. Pharm. Biopharm. **2000**, *50* (1), 27–46.
16. Kim, B.; La Flamme, K.; Peppas, N.A. Dynamic swelling behavior of pH-sensitive anionic hydrogels used for protein delivery. J. Appl. Polym. Sci. **2003**, *89* (6), 1606–1613.
17. Lowman, A.M.; Morishita, M.; Kajita, M.; Nagai, T.; Peppas, N.A. Oral delivery of insulin using pH-responsive complexation gels. J. Pharm. Sci. **1999**, *88* (9), 933–937.
18. Hubbell, J.A. Bioactive biomaterials. Curr. Opin. Biotechnol. **1999**, *10* (2), 123–129.
19. Langer, R. Biomaterials in drug delivery and tissue engineering: one laboratory's experience. Acc. Chem. Res. **2000**, *33* (2), 94–101.
20. Meinhart, J.; Deutsch, M.; Zilla, P. Eight years of clinical endothelial cell transplantation. Closing the gap between prosthetic grafts and vein grafts. ASAIO J. **1997**, *43* (5), M515–M521.
21. Langer, R. Selected advances in drug delivery and tissue engineering. J. Controlled Release **1999**, *62* (1–2), 7–11.
22. Marler, J.J.; Upton, J.; Langer, R.; Vacanti, J.P. Transplantation of cells in matrices for tissue regeneration. Adv. Drug Deliv. Rev. **1998**, *33* (1–2), 165–182.
23. Nugent, H.M.; Edelman, E.R. Tissue engineering therapy for cardiovascular disease. Circ. Res. **2003**, *92* (10), 1068–1078.

BioMEMS

L. James Lee
Department of Chemical Engineering, The Ohio State University, Columbus, Ohio, U.S.A.

INTRODUCTION

Miniaturization methods and materials are well developed in the integrated circuit industry. They have been used in other industries to produce microdevices, such as camera and watch components, printer heads, automotive sensors, micro-heat exchangers, micropumps, microreactors, etc., in the last 15 years.[1,2] These new processes are known as microelectromechanical systems (MEMSs), with a combined international market of over US$ 15 billion in 1998.[3] In recent years, MEMS applications have also been extended to the optical communication and biomedical fields. The former are called micro-optic electromechanical systems (MOEMSs), while the latter are known as biomicroelectromechanical systems (bioMEMSs). Potential MOEMS structures include optical switches, connectors, grids, diffraction gratings, and miniature lenses and mirrors. Major potential and existing bioMEMS products are biochips/sensors, drug delivery systems, advanced tissue scaffolds, and miniature bioreactors.

Future markets for biomedical microdevices for human genome studies, drug discovery and delivery in the pharmaceutical industry, clinical diagnostics, and analytical chemistry are enormous (tens of billions of U.S. dollars).[4] In the following sections, major bioMEMS applications and microfluidics relevant to bioMEMS applications are briefly introduced. Because of the very large volume of publications on this subject, only selected papers or review articles are referenced in this entry.

BIOMEMS APPLICATIONS

Biochips/Biosensors

Chip-based microsystems for genomic and proteomic analysis are the first bioMEMS products to have been commercialized. A large number of articles have been published in this field in recent years. Here, a brief introduction is given based on several recent review articles.[5–9] Biosensors are not necessarily microsystems.[10,11] MEMS techniques, however, may greatly enhance the performance of biosensors and reduce their cost. Microfabricated biosensors can be considered a division of biochips.[10–12]

Most molecular and biological assays and tests are very tedious, as shown in Fig. 1. They include the following steps: 1) obtaining a cellular sample (e.g., blood or tissue); 2) separating the cellular material of interest; 3) lysing the cells to release the crude DNA, RNA, and protein; 4) purifying the crude lysate; 5) performing necessary enzymatic reactions, such as denaturing, cleaving, and amplifying of the lysate by polymerase chain reaction (PCR); 6) sequencing DNA/genes using gel or capillary electrophoresis; and finally, 7) detecting and analyzing data. This process requires skilled technicians working in well-equipped biomedical laboratories, for periods of time ranging from many hours to several days to analyze a single sample. Much of today's diagnostic equipment is costly and bulky. It has limited use in medical diagnostics and is unsuited for emergency response at sites of care. To improve public health services, there is a great need to develop efficient and affordable methods and devices that can simplify the diagnostic process and be used as portable units. In recent years, the concept of integrating many analysis systems into one microdevice has attracted a great deal of interest in industry and academia. Such devices are called ''laboratories-on-chips.'' They combine a number of biological functions (such as enzymatic reactions, antigen–antibody conjugation, and DNA/gene probing) with proper microfluidic techniques (such as sample dilution, pumping, mixing, metering, incubation, separation, and detection in micrometer-sized channels and reservoirs) in a miniaturized device. The integration and automation involved can improve the reproducibility of results and eliminate the labor, time, and sample preparation errors that occur in the intermediate stages of an analytical procedure. The miniaturized devices also allow realization of low-energy and ''point-of-care,'' parallel detection from a very small sample size, and easy data storage and transfer through computers and the Internet.

Biochips used for genomic analysis range from those used for separations for DNA sequencing, to those used in microvolume PCR, to complete analysis systems. Sequencing separations of single-stranded DNA fragments on a microchip follow the same principles as in conventional capillary electrophoresis.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030622

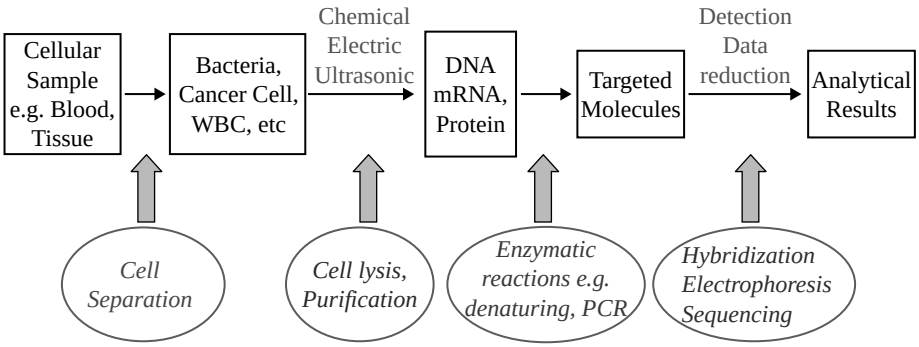


Fig. 1 Schematic of molecular diagnostics. *(View this art in color at www.dekker.com.)*

The process, however, is much faster because a higher electric field can be applied to the micrometer-sized separation channels without Joule heating problems. Automatic injection of a very small sample volume and parallel processing of a large number of separations can also be easily achieved. PCR allows amplification of a specific region of a DNA chain. PCR carried out on microchips is much more efficient than that on commercial PCR thermocyclers. In PCR, a sample solution is mixed, containing DNA, primers (synthesized short oligomers whose sequences flank the region of DNA to be amplified), a thermostable DNA polymerase enzyme, and the individual deoxyribonucleotides. Melting (or denaturization) of the double-stranded DNA molecules to single-stranded ones is done by heating the sample to ~95°C. The system is then quickly cooled to ~60°C for annealing; during this process, the added primers adhere to the single-stranded DNA. Finally, the sample is heated to ~72°C, at which temperature the polymerase is most active. During this extension period, complementary dinucleotide triphosphates are added to the growing strand using the target DNA as template. Each PCR cycle may double the amount of DNA of the required length. In the ideal case, 1 mol of targeted DNA fragments can be produced after 79 cycles. Practically, 20–50 cycles are needed to obtain a measurable quantity. Due to the high surface area/volume ratio associated with microdevices (it is important that PCR-friendly surfaces are produced in these devices), heat transfer is more efficient and temperature control is much easier than in large systems. PCR time can be easily reduced from hours to minutes, particularly in continuous-flow PCR chips.[6]

Miniaturized proteomic analysis devices include enzyme assays and immunoassays. The enzyme assay chip is mainly a sophisticated incubator and flow-through system. It can perform multiple functions typically required by the biochemist, namely, diluting substrate and buffer, mixing enzyme and substrate, incubating during conversion, and allowing for detection in a flow channel. Immunoassay chips are similar to enzyme assay chips, except that the main focus is antigen–antibody interaction for clinical diagnostics and drug discovery.

There are a small number of commercially available biochips in the market today. Most are microarray-based systems, with biomolecules such as DNA probes, enzymes, and antigens being immobilized on the chip surface (e.g., GenChip® from Affymetrix, NanoChip™ from Nanogen, Inc., and GeneXpert® from Cepheid), or simple microfluidic systems capable of DNA sequencing by either electrophoresis (e.g., LabChip® from Caliper, Inc., and LabCard™ from ACLARA BioSciences, Inc.) or PCR. DNA microarray chips and DNA sequencing microfluidic chips have contributed to the Human Genome Project.

In addition to commercially production, a great deal of research and development work on biochips has been going on both in industry and in academia. Genomic analysis of DNA and RNA continues to be the focus of interest, but more and more effort is being spent on proteomic analysis of proteins and peptides.[5–9] Several enzyme assays and immunoassays designed based on microarray-based systems with simple microfluidic control are close to commercialization. They can be a vital tool in clinic diagnostics, drug discovery, and biomedical research.

Completely integrated micro-total analysis systems (μ-TAS) that can perform all the functions mentioned in Fig. 1 would be very valuable for high-throughput drug screening and personalized healthcare. However, only model systems have been proposed by research groups at present.[9] The mass production of such complicated systems at low cost is a challenging issue. Silicon and glass have been the most popular materials for fabricating microchips, but polymers are increasingly being used because of the availability of flexible, low-cost, high-throughput manufacturing methods for the micro-/nanoscale features needed for these types of applications.

Sensitive detection in microfluidic analytical devices is a challenge because of the extremely small detection volumes available. In conventional capillary

electrophoresis, the most commonly employed detection method is UV absorption. In microscale biochips, laser-induced fluorescence (LIF) in conjunction with optical microscopy is currently the dominant detection technique because of its high sensitivity and noncontact nature. LIF microscopy, however, is costly, and the equipment size is quite large. To ensure wide application of the miniaturized biomedical devices, simple, portable, and low-cost detection methods are essential. Considerable efforts have been made lately to explore electrochemical methods, because the use of electrodes for detection leads to smaller instruments and cost reduction. Amperometry, conductimetry, and electrochemiluminescence are also likely methods to complement fluorescence detection for on-chip analysis.[13]

Drug Delivery

Self-regulated and controllable drug delivery systems

Most conventional drug delivery systems are based on polymers or lipid vesicles. Drug safety and efficacy can be greatly improved by encapsulating the drug inside or attaching it to a polymer or lipid. The three general mechanisms by which drugs are delivered from polymer or lipid systems are: 1) diffusion of the drug species through a polymer membrane; 2) a chemical or enzymatic reaction leading to cleavage of the drug from the system; and 3) solvent activation through swelling or osmosis of the system.[14] A major limitation of currently available delivery devices is that they release drugs at a predetermined rate. Certain disease states, such as diabetes, heart disease, hormonal disorders, and cancer, require drug administration either at a life-threatening moment or repeatedly at a certain critical time of day. Drug delivery technology can be taken to the next level by the fabrication of ''smart'' polymers or devices that are ''responsive'' to the individual patient's therapeutic requirements and deliver a certain amount of drug in response to a biological state. Given the miniature size of implantable devices, micromachining techniques will be essential for their manufacture. Currently, there are no commercial products based on the micromachined responsive drug delivery approach, and only some early research activity is seen in this direction.

The controlled release of drugs has been explored by adapting intelligent polymers, such as functional hydrogels, which respond to stimuli such as magnetic fields, ultrasound, electric current, temperature, and pH change.[15–17] These chemically synthesized materials are biocompatible and have good functionality. However, they often lack well-defined properties because of their inherent size and structure distribution resulting from chemical synthesis.[18] On the other hand, microfabrication technology developed for microelectronic applications is capable of mechanically creating devices with more precisely defined features, in a size range similar to that of polymeric and lipid materials.[19] Using hydrogels as switches or gates for controlled drug delivery and microfluidics has been explored recently by several researchers.[20,21] In a recent paper, Cao, Lai, and Lee[22] describe the design of a self-regulated drug delivery device based on the integration of both mechanical and chemical methods. A pH-sensitive hydrogel switch is used to regulate the drug release, while a constant release rate is achieved by carefully designing the shape of the drug reservoir.

Biocapsules, membranes, and engineered particles for drug delivery

Immunoisolation is the protection of implanted cells from the host's immune system by the complete prevention of contact of immune molecules with the implanted cells, generally by the use of a semipermeable membrane. To achieve this without preventing nutrients from reaching the cells or waste from being removed, it is necessary to have an absolute pore size just below the minimum size needed to block out the smallest immune molecule, immunoglobulin G (IgG). The polymer and ceramic membranes used currently in biomedical devices possess nanopores with nonuniform size distributions, which makes it difficult to control the passage of drugs and immunoglobulins (~30–50 nm in size) through these membranes.[23,24] Nonuniform porosity also requires the use of long, torturous flow paths, necessitating the use of thick membranes. Nanoscale resistance to flow in such thick membranes is high, so that high applied pressures (~1–4 MPa) are needed, which further complicates use.[25] These membranes also show incomplete virus retention.

Ferrari and co-workers[19] examined the feasibility of using microfabricated silicon nanochannels for immunoisolation. A suspension of cells was placed between two microfabricated structures with nanoporous membranes to fabricate an immunoisolation biocapsule. Characterization of diffusion through the nanoporous membranes demonstrated that 18 nm channels did not completely block IgG but did provide adequate immunoprotection (immunoprotected cells remained functional in vitro in a medium containing immune factors for more than two weeks, while unprotected cells ceased to function within two days). A major application of biocapsules containing nanochannels is immunoisolation of transplanted cells for the treatment of hormonal and biochemical deficiency diseases, such as diabetes.

Polymer microparticles have attracted much attention for drug delivery applications. Traditional microparticle

fabrication protocols, such as phase separation, emulsification, and spray drying, have been successfully used for the production of drug delivery microspheres.[26,27] However, due to the surface-driven manufacturing process of these methods, the structural complexity of the resulting particles is limited. These methods are also difficult to apply for producing a monodispersed particle size distribution. Size control of microparticles is an important factor, since there are many routes of drug administration. According to DeLuca et al.,[28] very large microparticles (>100 μm) with a broad particle size distribution are acceptable for embolization and drug delivery by implantation. Microparticles in the size range of 10–100 μm can be used for subcutaneous and intramuscular administration. Here, the particle size distribution is not a critical factor. Intravenous administration results in localization in the capillary vasculature and uptake by macrophages and phagocytes. Microparticles larger than 8 μm lodge predominately in the lung capillaries, whereas those smaller than 8 μm may clear the lung and be localized in the liver and spleen. Therefore, it is most important to control the size of the largest particle.

During inhalation administration to the lung, filtering of particles occurs in the upper airways by inertial impaction, with large particles (aerodynamic diameter $d_a > 5\,\mu m$) being deposited in the mouth and the first few generations of airways. Very small particles ($d_a < 1\,\mu m$) are dispersed by diffusion, and a large fraction of these particles remain suspended in the airflow and are exhaled. Microparticles with the optimal size range of 1–5 μm are deposited in the central and peripheral airways and in the alveolar lung region by a combination of inertial impaction and sedimentation.[29] Therefore, the size and distribution of microparticles for inhalation therapies must be closely controlled to achieve high efficiency. Inhalation is a noninvasive drug delivery route and has been used widely for the treatment of diseases such as asthma, cystic fibrosis, and chronic obstructive pulmonary disease. Potential applications of new inhalation products in the near future include the treatment of diseases such as diabetes, pain, and growth deficiency, where proteins and lipids-based drugs will be used. For these biomolecule-based drugs, processing conditions such as high temperature and long solvent contact time may result in drug denaturization, so particle formation methods must avoid such conditions. For certain envisioned functional features of drug delivery vehicles, such as targeted and controlled vector release on cancer tumors, ''highly engineered'' microparticles (i.e., each particle is essentially a microdevice) may be required. This is another limiting factor for the traditional microparticle fabrication methods.

Compared to conventional polymer microparticle fabrication methods, microfabrication offers greater control of particle features and geometries. The shape and size of the particles can be controlled tightly. Perhaps more importantly, the components and surface properties can be designed to achieve particular functions. Using soft lithography (this fabrication method is explained in a later section), Guan and Hansford[30] recently developed a simple method to fabricate nonspherical polymer microparticles of precise shape and size, which can serve as either drug delivery vessels or substrates for further processing to produce functional drug delivery devices. Fig. 2 shows a micrograph of thin, platelike microparticles fabricated using this method. Combining the surface micropatterning and surface-tension self-assembly of autofolding,[31,32] well-defined 3D micropolyhedra, e.g., cubes and pyramids, can be fabricated from metals.[33] In our laboratory, similar micropolyhedra are currently being developed using functional polymers (e.g., biodegradable polymers and functional hydrogels). Large protein and gene molecules may be wrapped in such well-defined microstructures and delivered to targeted sites by either pulmonary delivery or intravenous administration.

Tissue Engineering

Tissue engineering is the regeneration, replacement, or restoration of human tissue function by combining synthetic and living molecules in appropriate configurations and environments.[34] The scaffold, the cells, and the cell–scaffold interactions are the three major components of any tissue-engineered construct. Although many tissue scaffold materials, such as foams and nonwoven fabrics, have been developed and used,[35,36] many challenges must be overcome for the promise of tissue engineering to become a reality. These include: 1) low-cost fabrication of well-defined 3D scaffold configurations at both micro- and nanoscale; 2) incorporation of appropriate biocompatibility, bioactivity, and biodegradability in the scaffolding construct to manipulate cellular and subcellular functions; and 3) active control of transport phenomena and cell growth kinetics to mimic microvasculature functions. Micro-/nanofabrication technology of polymers has tremendous potential in this field because it can achieve topographical, spatial, chemical, and immunological control over cells and thus create more functional tissue engineering constructs.[37]

An ideal tissue scaffolding process should be able to produce well-controlled pore sizes and porosity, provide high reproducibility, and use no toxic solvents. This is because these physical factors are associated with nutrient supply and vascularization of the cells in the implant as well as the development of a fibrous tissue layer that may impede nutrient access to the cells. Current processing methods used for polymer

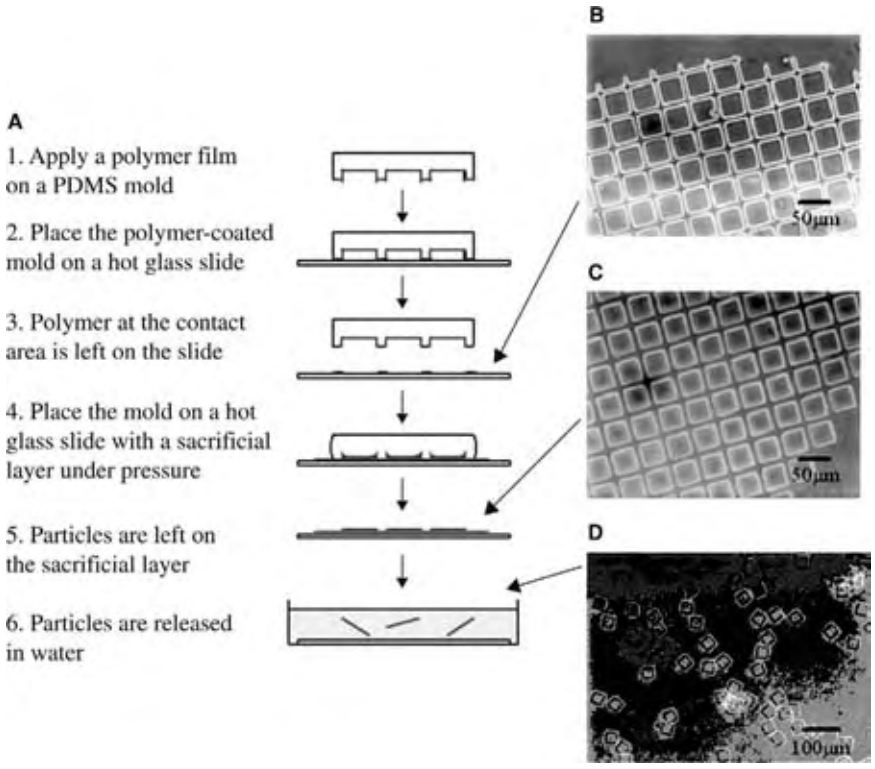


Fig. 2 Thin platelike microparticle fabricated by hot stamping. (A) The schematic of the procedure; (B) polymer at the contact area left on the glass slide; (C) polymeric microparticles left on the sacrificial layer after hot stamping; (D) and in water after release.

scaffolds include solvent casting, plastic foaming, fiber bonding, and membrane lamination.[35] However, precise, reproducible features in the micrometer and nanometer range are difficult to attain using these methods. By combining living cells and microfabricated 2D and 3D scaffolds with carefully controlled surface chemistry, investigators have begun to address fundamental issues such as cell migration, growth, differentiation, apoptosis, orientation, and adhesion, as well as tissue integration and vascularization.

The functioning of tissues such as retinal, cardiac, and vascular tissue is dependent on the controlled orientation of multiple cell types. A key issue in the engineering of these tissues is control of the spatial distribution of cells in vitro to recreate a lifelike environment. The current approach to seeding cells is to allow cells to be randomly distributed in the scaffold. Microfabrication techniques, on the other hand, can produce short- and long-range surface patterns to mediate cell distribution and adhesion, biological interaction, and immune responses. Porous scaffolds without integrated blood supply rely solely on diffusion for mass transfer. They are limited to millimeters in size, while normal tissues leverage convection from blood vessels to enable oxygenation of large tissues.[38] Incorporating microfluidic networks in 3D tissue scaffolds for cell culture and implantation can be achieved by microfabrication techniques.[47] This new approach can provide the functional equivalents of microvasculature and enable scale-up of tissue engineering. Many cell-based bioreactors can be designed in a similar manner.[39]

Scientific and commercial work to date in tissue engineering has been largely devoted to clinical needs and focused on physiological aspects. There is a lack of low-cost, solventless, and mass-producible processing methods to fabricate scaffolds with well-defined micro- and nanostructures. In our laboratory, a manufacturing protocol is currently being developed for 3D tissue scaffolds of various shapes. The scaffold can be easily fabricated by combining micropatterned biodegradable polymers and supercritical CO_2 foaming technology.

Depending on the type of bioMEMS application, the polymers used can range from low-cost commodity plastics for disposable biochips to biodegradable and biofunctional polymers for drug delivery and tissue engineering. The feature size in these microsystems can be in either the micrometer or the nanometer scale. For instance, 10–100 μm is the desired microchannel size in microfluidic biochips. Below that, detection is too difficult, and, above that, mixing, heat transfer, and mass transfer are too slow. Particles used in drug delivery and the cell size in tissue scaffolds are also in the micrometer range. On the other hand, nanosized features are essential for immunoisolation in cell-based gene therapy and cell culture in tissue engineering. The enabling processing methods need to cover a broad size range, be mass producible and affordable, and be

compatible with the polymers and biomolecules used in the process. Fluid transport in bioMEMS devices is crucial in many applications such as fast DNA sequencing, protein separation, drug delivery, and tissue generation.

Microfluidics

Microfluidics is the manipulation of fluids in channels, with at least two dimensions at the micrometer or submicrometer scale. This is a core technology in a number of miniaturized systems developed for chemical, biological, and medical applications. Both gases and liquids are used in micro-/nanofluidic applications,[40,41] and generally, low-Reynolds-number hydrodynamics is relevant to bioMEMS applications. Typical Reynolds numbers for biofluids flowing in microchannels with linear velocity up to 10 cm/s are less than 30.[42] Therefore, viscous forces dominate the response and the flow remains laminar.

Fluid motion in these small-scale systems can be driven by applied pressure difference, electric fields associated with charged Debye double layers (or electrical double layer—EDL)—common when ionic solutions are present, or capillary driving forces owing to wetting of surfaces by the fluid.[42] Pressure-driven flow is similar to the classic Poiseuille flow. Electrokinetic effects can result in either electro-osmotic flow (EOF) or electrophoretic responses. Electro-osmotic flow is a bulk flow driven by stresses induced in the thin EDL (i.e., 1–10 nm) near the channel walls, caused by an electric field imposed across the channel length. The velocity profile in the core of the channel is plug-like, even for a channel height as small as 24 nm.[43] Higher electrical permittivity of the fluid, imposed electric field strength, and zeta potential on the wall surface may all increase the flow rate.

Electrophoretic response, on the other hand, is the motion of charged molecules in a fluid caused by an electric field imposed across the channel length. Positively charged molecules move to the negative electrode, while negatively charged molecules move towards the positive electrode, leading to molecule separation. Electrophoresis is the most widely used separation method in the biotechnology field today. Typically, a buffer solution is chosen such that all biomolecules in the fluids, e.g., DNA/RNA fragments and proteins, are negatively charged. They all migrate from the sampling point to the detection point. Since DNA molecules have the same charge/mass ratio, separation is usually achieved by placing an immobilized gel or a mobile ''gel'' solution in the separation channel. Electro-osmotic flow may cause unwanted washout of the gel solution during electrophoresis, so some sort of channel coating may be necessary for EOF suppression if the channel wall has a high zeta potential (e.g., glass). On the other hand, undesirable electrophoretic separation may occur in EOF if the sample solution contains components with different charges. A high-ionic-strength-plugs method has been developed to facilitate sample transport. The use of solutions at different ionic strengths and therefore different electroosmotic mobility, however, creates a quite complex situation in microfluidics. Electrokinetic flows work very well in microchannels because of the large surface-to-volume ratio, which minimizes the Joule heating problem in this type of flow. Very high electric field strength (i.e., hundreds to thousands of volts per centimeter channel length) can be easily applied in microdevices to speed up the processing time from hours to minutes or even seconds. For nanosized channels, it has been found that very low electric power (e.g., several volts per micrometer channel length) can generate a volume flow rate that is practical for controlled drug delivery.[43]

Capillary separation is also highly favored in microfluidics. This method is simple and low-cost, but a gas–liquid interface must exist. It is mainly used for reagent loading and release in portable biochips and drug delivery systems. The velocity profile is similar to that in pressure-driven flow, but the flow is very sensitive to the surface tension of the fluid, solid surface energy and roughness, and channel shape.[44] Active control of surface tension forces to manipulate flows in microchannels can be achieved by forming gradients in interfacial tension on the channel surface[45] or by electrowetting.[46]

For most cases involving the flow of small-molecule liquids, such as buffer solutions, the standard continuum description of transport processes works very well, except that surface forces (surface tension, electrical effects, van der Waals interactions, and, in some cases, steric effects) play a more important role than usual. Although some discrepancies have been reported between pressure-driven flow measurements made in microchannels and calculations based on the Navier–Stokes equations, most have been found to be experimental errors.[42] This is because the pressure drop, as a function of flow rate, varies as the inverse fourth power of channel radius (or inverse third power of channel height), and a small change in the radius (or channel height) due to manufacturing imperfections or channel-wall contamination produces large changes in the flow. Since the volumetric flow rate varies linearly with channel radius (or height) for electrically driven flow,[43] EOF is a more reliable way than pressure-driven flow to verify microfluidic experiments with calculations. A recent study[43] shows that calculated flow rates from classical EOF analysis agree well with experimental data for channel heights in the range of 10–20 nm.

Retardation of flow of ionic liquids and solutions in microchannels, however, can be significant when the

channel walls have either the same static charge[47] (e.g., glass surface is negatively charged) in pressure-driven flow or opposite charges in EOF.[43] In the former case, the flow causes charges inside the EDL to accumulate downstream, while charges on the solid channel wall remain immobile. Such excess charge creates a potential drop in the channel direction, causing a ''backflow.'' For channel height in the range of 100 μm, this electroviscous effect (flow retardation is often counted as an increase in fluid viscosity) is small. But a retardation of 70% is observed when the glass channel diameter is in the range of several micrometers.[47] In the latter case, the backflow can be manipulated by surface micropatterning of opposite charges on the walls of the microchannel to achieve laminar chaotic mixing[48] or controllable membrane permeation.

In many bioMEMS applications, the sample fluid contains molecules and particles of various sizes. Small organic molecules are a few angstroms in size, typical protein molecules are about 2–5 nm, and large DNA molecules and cells are in the range of 1–10 μm. In some cases, the radii of near-spherical fluid droplets or gas bubbles are comparable to that of the channel, but in others, their lengths may be larger.[42] Non-Newtonian fluid and multiphase flow mechanics must be applied. In microchannels, the shear rate can be very high, e.g., $10^7\,\mathrm{sec}^{-1}$, even though the Reynolds number is low. Rheological characterization of polymeric fluids and biofluids in such a flow field has recently been studied in our laboratory.[48] It was found that the standard rheological analysis used at the macroscale also works at the microscale. The high-shear Newtonian plateau can be easily observed. For solutions containing large polymer (or DNA) molecules, polymer degradation and wall slip are substantial when the flow rate is high. Rheology in microchannels needs to be studied further because many biofluids are highly non-Newtonian. One advantage of microfluidics is that a single biomolecule such as DNA can be isolated and analyzed on a biochip containing small channels or wells.[49,50] Since the molecule size is comparable to the channel (well) dimension, understanding and manipulating both the macroscopic and microscopic transport phenomena of the confined molecule undergoing flow is an active area of research.[42]

CONCLUSIONS

The miniaturization of biomedical and biochemical devices for bioMEMSs has gained a great deal of attention in recent years. Products include biochips/biosensors, drug delivery devices, tissue scaffolds, and bioreactors. In the past, MEMS devices have been fabricated almost exclusively in silicon, glass, or quartz because of the comparable technology available in the microelectronics industry. For applications in the biochemistry and biomedical field, polymeric materials are desirable because of their lower cost, good processability, and biocompatibility. Polymer microfabrication techniques, however, are still not well developed.

REFERENCES

1. Madou, M.J. *Fundamentals of Microfabrication: The Science of Miniaturization*, 2nd Ed.; CRC Press: Boca Raton, FL, 2002.
2. Jensen, K.F. Microchemical systems: status, challenges, and opportunities. AIChE J. **1999**, *45*, 2051.
3. Freemantle, M. Downsizing chemistry: chemical analysis and synthesis on microchips promise a variety of potential benefits. Chem. Eng. News **1999**, *77*, 27–36.
4. Snyder, M.R. Micromolding technology extends sub-gram part fabrication capability. Mod. Plast. **1999**, *76* (1), 85.
5. Bousse, L.; Cohen, C.; Nikiforov, T.; Chow, A.; Kopf-Sill, A.R.; Dubrow, R.; Parce, J.W. Electrokinetically controlled microfluidic analysis systems. Annu. Rev. Biophys. Biomol. Struct. **2000**, *29*, 155.
6. Sanders, G.H.W.; Manz, A. Chip-based microsystems for genomic and proteomic analysis. Trends Anal. Chem. **2000**, *19* (6), 364.
7. Carrilho, E. DNA sequencing by capillary array electrophoresis and microfabricated array systems. Electrophoresis **2000**, *21*, 55.
8. Kricka, L.J. Microchips, microarrays, biochips and nanochips: personal laboratories for the 21st century. Clin. Chim. **2001**, *307*, 219.
9. Krishnan, M.; Namasivayam, V.; Lin, R.; Pal, R.; Burns, M.A. Microfabricated reaction and separation systems. Curr. Opin. Biotechnol. **2001**, *12*, 92.
10. Vo-Dinh, T.; Cullum, B. Biosensors and biochips: advances in biological and medical diagnostics. Fresenius J. Anal. Chem. **2000**, *366*, 540.
11. Wang, J. Glucose biosensors: 40 years of advances and challenges. Electroanalysis **2001**, *13* (12), 983.
12. Lauks, I.R. Microfabricated biosensors and microanalytical systems for blood analysis. Acc. Chem. Res. **1998**, *31*, 317.
13. Schwarz, M.A.; Hauser, P.C. Recent developments in detection methods for microfabricated analytical devices. Lab on a Chip Miniaturis. Chem. Biol. **2001**, *1* (1), 1.
14. Langer, R. Drug delivery and targeting. Nature (Suppl.) **1998**, *392*, 5.

15. Lowman, A.M.; Peppas, N.A. Analysis of the complexation/decomplexation phenomena in graft copolymer networks. Macromolecules **1997**, *30*, 4959.
16. Torres-Lugo, M.; Peppas, N.A. Molecular design and in vitro studies of novel pH-sensitive hydrogels for the oral delivery of calcitonin. Macromolecules **1999**, *32*, 6646.
17. Traitel, T.; Cohen, Y.; Kost, J. Characterization of glucose-sensitive insulin release systems in simulated in vivo conditions. Biomaterials **2000**, *21*, 1679.
18. Lanza, R.P.; Chick, W. Encapsulated cell therapy. Sci. Am. Sci. Med. **1995**, *2* (4), 16.
19. Desai, T.A.; Hansford, D.; Ferrari, M. Characterization of micromachined silicon membranes for immunoisolation and bioseparation applications. J. Membrane Sci. **1999**, *159*, 221.
20. Kaetsu, I.; Uchida, K.; Shindo, H.; Gomi, S.; Sutani, K. Intelligent type controlled release systems by radiation techniques. Radiat. Phys. Chem. **1999**, *55*, 193.
21. Liu, R.H.; Yu, Q.; Bauer, J.M.; Jo, B.-H.; Moore, J.S.; Beebe, D.J. In-channel processing to create autonomous hydrogel microvalues. In *Micro Total Analysis systems 2000*, Proceedings of the 4th μ_{TAS} symposium, Enschede, Netherlands, May 14–18 2000; 45–48.
22. Cao, X.; Lai, S.; Lee, L.J. Design of a self-regulated drug delivery device. Biomed. Microdev. **2001**, *3* (2), 109.
23. Colton, C.K. Implantable biohybrid artificial organs. Cell Transplant. **1995**, *4* (4), 415.
24. Desai, T.A.; Hansford, D.J.; Kulinsky, L.; Nashat, A.H.; Rasi, G.; Tu, J.; Wang, Y.; Zhang, M.; Ferrari, M. Nanopore technology for biomedical applications. Biomed. Microdev. **2000**, *2* (1), 11.
25. Kim, K.J.; Stevens, P.V. Hydraulic and surface characteristics of membranes with parallel cylindrical pores. J. Membrane Sci. **1997**, *123*, 303.
26. Jain, R.A. The manufacturing techniques of various drug loaded biodegradable poly(lactide-co-glycolide) (PLGA) devices. Biomaterials **2000**, *21*, 2475.
27. Langer, R. Biomaterials in drug delivery and tissue engineering: one laboratory's experience. Acc. Chem. Res. **2000**, *33*, 94.
28. DeLuca, P.P.; Mehta, R.C.; Hausberger, A.G.; Thanoo, B.C. Biodegradable polyesters for drug and polypeptide delivery. In *Polymer Delivery Systems, Properties and Applications*; El-Nokaly, M.A., Piatt, D.M., Charpentier, B.A., Eds.; ACS Symposium Series 520; Amercian Chemical Society: Washington, DC, 1993; 53–79 (Chapter 4).
29. Edwards, D.A. Delivery of biological agents by aerosols. AIChE J. **2002**, *48* (1), 2.
30. Guan, J.; Lee, L.J.; Hansford, D.J. Layered thin-film polymer microparticles fabricated by soft lithography. To be submitted to chemistry of marterials.
31. Green, P.W.; Syms, R.R.A.; Yeatman, E.M. Demonstration of three-dimensional microstructure self-assembly. J. Microelectromech. Syst. **1995**, *4* (4), 170.
32. Harsh, K.F.; Bright, V.M.; Lee, Y.C. Solder self-assembly for three-dimensional microelectromechanical systems. Sensors Actuators **1999**, *77*, 237.
33. Gracias, D.H.; Kavthekar, V.; Love, J.C.; Paul, K.E.; Whitesides, G.M. Fabrication of micrometer-scale, patterned polyhedra by self-assembly. Adv. Mater. **2002**, *14* (3), 235.
34. Langer, R.; Vacanti, J.P. Tissue engineering: the design and fabrication of living replacement devices for surgical reconstruction and transplantation. Lancet **1999**, *354*, 23.
35. Mikos, A.G.; Sarakinos, G.; Leite, S.M.; Vacanti, J.P.; Langer, R. Laminated three-dimensional biodegradable foams for use in tissue engineering. Biomaterials **1993**, *14* (5), 323.
36. Li, Y.; Yang, S.-T. Effects of three-dimensional scaffolds on cell organization. Biotechnol. Bioprocess Eng. **2001**, *6*, 311.
37. Desai, T.A. Micro- and nanoscale structures for tissue engineering constructs. Med. Eng. Phys. **2000**, *22*, 595.
38. Griffith, L.G.; Noughton, G. Tissue engineering—current challenges and expanding opportunities. Science **2002**, *295* (5557), 1009.
39. King, K.R.; Terai, H.; Wang, C.C.; Vacanti, J.P.; Borenstein, J.T. Microfluidics for tissue engineering microvasculatuer: endothelial cell culture. In *Micro-total Analysis Systems*; 2001; Proceedings of the 5th μ_{TAS} 2001 Symposium, Monterey, CA, USA, October 21–25, 2001; 247–249.
40. Gad-el-Hak, M. The fluid mechanics of microdevices. J. Fluids Eng. **1999**, *121*, 5.
41. Giordano, N.; Cheng, J.-T. Microfluid mechanics: progress and opportunities. J. Phys. Condens. Matter **2001**, *13*, R271.
42. Stone, H.A.; Kim, S. Microfluidics: basic issues, applications, and challenges. AIChE J. **2001**, *47* (6), 1250.
43. Conlisk, A.T.; McFerran, J.; Zheng, Z.; Hansford, D. Mass transfer and flow in electrically charged micro- and nanochannels. Anal. Chem. **2002**, *74*, 2139.
44. Kang, K.; Lee, L.J.; Koelling, K.W. High shear microfluidics and its application in rheological measurements. Experiments in Fluids **2005**, *38*, 222–232.

45. Gallardo, B.; Gupta, V.K.; Eagerton, F.D.; Jong, L.I.; Craig, V.S.; Shah, R.R.; Abbott, N.L. Electrochemical principles for active control of liquids on submillimeter scales. Science **1999**, *283*, 57.
46. Pollack, M.G.; Fair, R.B.; Shenderov, A.D. Electrowetting-based actuation of liquid droplets for microfluidic applications. Appl. Phys. Lett. **2000**, *77* (11), 1725.
47. Kulinsky, L.; Wang, Y.; Ferrari, M. Electroviscous effects in microchannels. SPIE Proc. **1999**, *3606*, 158.
48. Stroock, A.D.; Weck, M.; Chiu, D.T.; Huck, W.T.S.; Kenis, P.J.A.; Ismagilov, R.F.; Whitesides, G.M. Patterning electro-osmotic flow with patterned surface charge. Phys. Rev. Lett. **2000**, *84* (15), 3314.
49. Smith, D.E.; Babcock, H.P.; Chu, S. Single-polymer dynamics in steady shear flow. Science **1999**, *283*, 1724.
50. Shrewsbury, P.J.; Muller, S.J.; Liepmann, D. Effect of flow on complex biological macromolecules in microfluidic devices. Biomed. Microdev. **2001**, *3* (3), 225.

Biomolecular Engineering

Zengyi Shao
Ee Lui Ang
Huimin Zhao
Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.

INTRODUCTION

Biomolecular engineering is an emerging field of academic research and industrial practice having the goal of engineering value-added biomolecules and biomolecular processes for applications in medical, chemical, agricultural, and food industries.[1] This new subject is very broad and highly interdisciplinary, including, but is not limited to, protein engineering, metabolic engineering, bioinformatics, bioprocessing, gene therapy, drug design, discovery and delivery, biomaterials, and nano-biotechnology. In the past decades, biomolecules such as protein pharmaceuticals, DNA vaccines, monoclonal antibodies, and industrial enzymes have been increasingly commercialized. In this entry, the key concepts and technologies in biomolecular engineering and their applications in engineering proteins, pathways, and nucleic acids are introduced.

KEY CONCEPTS AND TECHNOLOGIES EMPLOYED IN BIOMOLECULAR ENGINEERING

Protein Engineering

Protein engineering refers to the ability to alter protein structure to achieve a desired protein function. Two main protein engineering approaches, rational design and directed evolution, have been developed in the past two decades (Fig. 1). The former involves alterations of selected residues in a protein via site-specific mutagenesis to achieve predicted changes in function. In comparison, directed evolution mimics the process of natural evolution in the test tube, involving repeated cycles of creating molecular diversity by random mutagenesis/gene recombination, followed by screening/selecting the functionally improved variants.

Metabolic Pathway Engineering

Metabolic pathway engineering involves the directed improvement of product formation or cellular properties through the modification of specific biochemical reaction(s) or through the introduction of new one(s) by recombinant DNA technology.[2] Specifically, metabolic engineering includes identification of metabolic pathways, elucidation of regulatory mechanisms, metabolic flux analysis, metabolic control analysis (MCA), identification of inter- or intra-cellular transport mechanisms, and discovery and manipulation of biosynthetic pathways. This area is particularly important to biotechnology because it offers ways for improving existing bioprocesses, designing new bioprocesses, as well as producing novel chemicals and pharmaceuticals.

Many molecular biology and analytical chemistry tools have been developed for metabolic engineering.[3] For example, various vectors have been designed for the optimal expression of heterologous genes in industrial hosts. Several gene-cloning techniques have been developed to isolate novel genes from cultivable or uncultivable organisms, whereas RNA-antisense techniques have been used to silence gene expressions. In addition, directed evolution methods have been used to construct genes, pathways, or whole genomes with altered functions. As important as these genetic engineering tools, a number of powerful analytical techniques have also been developed for metabolic pathway analysis and analyses of cellular functions, such as gas chromatography and mass spectrometry (GC–MS), nuclear magnetic resonance (NMR), two-dimensional gel electrophoresis, DNA chips, and protein chips.

Because of the complexity of metabolic networks, rationally designed metabolic pathways often have undesired metabolic consequences on unrelated cellular properties. To address this limitation, a new metabolic engineering strategy of particular interest, ''inverse metabolic engineering'' (IME), was developed[4] (Fig. 2), which integrates directed evolution principles with the ''direct'' classical metabolic engineering. The strategy begins with the construction and identification of a desired phenotype; then the genetic basis or environmental factor for the desired phenotypic characteristic is determined; finally, this phenotype is endowed on another strain or organism by genetic manipulation. The essential and challenging step of IME is to identify the genetic basis of the desired phenotype,[5] which is illustrated later.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120019077

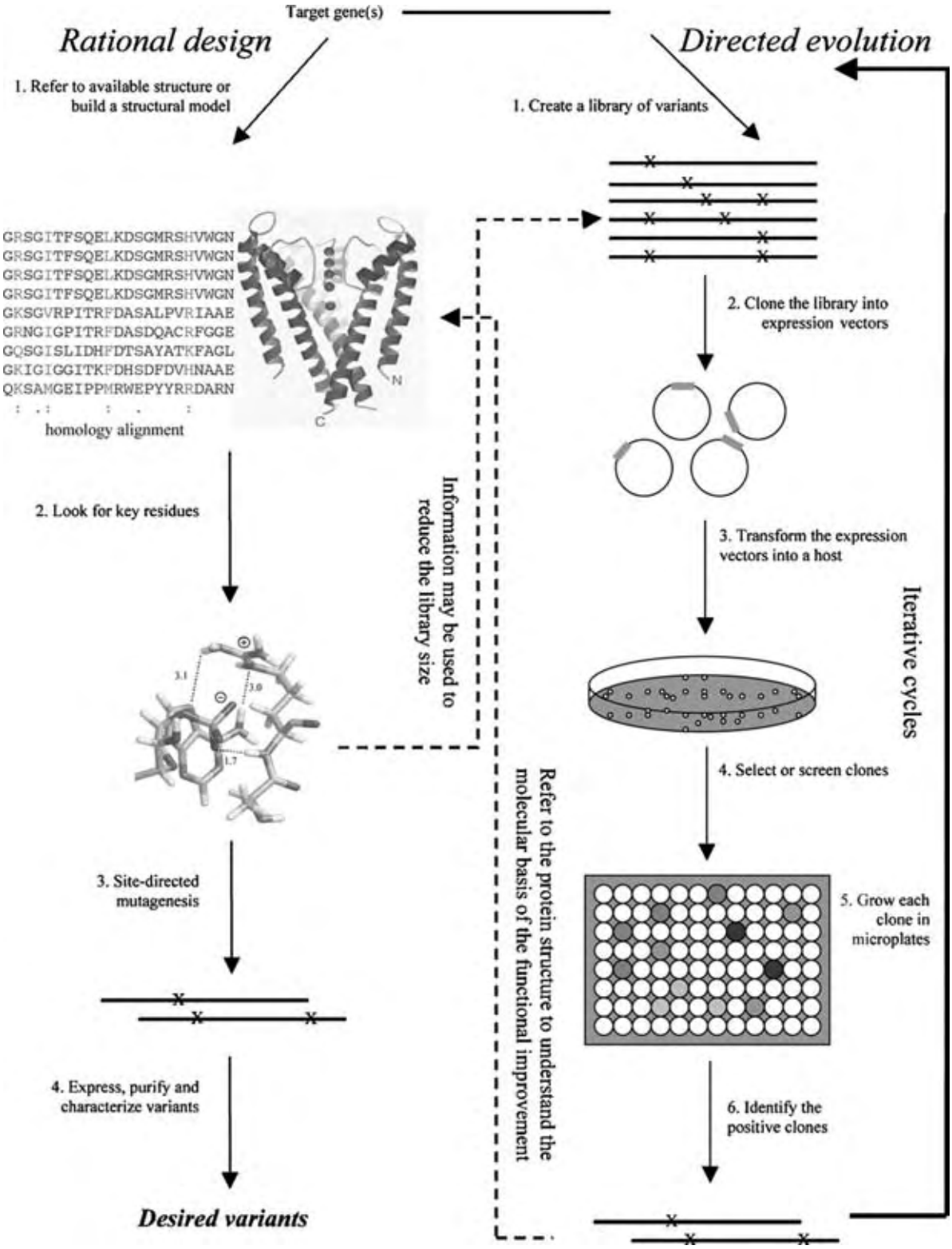


Fig. 1 Comparison between rational design and directed evolution. *(View this art in color at www.dekker.com.)*

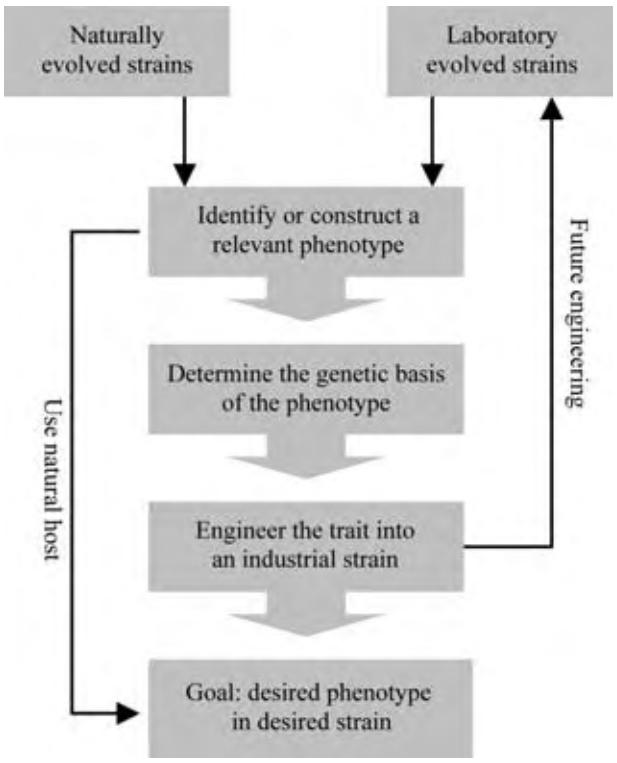


Fig. 2 The scheme of the inverse metabolic engineering (IME) approach. (From Ref.[5].)

Another two key concepts in metabolic engineering are metabolic pathway analysis and metabolic pathway modeling. The former is used for assessing inherent network properties in the complete biochemical reaction networks. It involves identification of the metabolic network structure (or pathway topology), quantification of the fluxes through the branches of the metabolic network, and identification of the control structures within the metabolic network.[3]

To identify metabolic network structure of common organisms, it is often helpful to do an extensive literature review. Also, a few pathway databases, such as BioCyc Knowledge Library (www.biocyc.org), may be used to identify specific metabolic pathways. However, in some cases, it is not easy to identify the complete metabolic network. This will cause a serious problem as some parts of the pathways that carry significant metabolic fluxes will possibly be ignored, resulting in only partial or even false estimation and analysis of the network structure. In these cases, enzyme assays and isotope-labeling techniques combined with either NMR or GC–MS are valuable tools for confirming the presence of specific enzymes and pathways. Once the whole metabolic network structure has been identified, quantification of the flux distribution through branches can be determined by solving a series of algebraic mass balance equations in terms of substrates, products, biomass constituents, and intracellular metabolites. For example, Fig. 3 shows how the carbon fluxes are distributed through the individual branches of metabolic pathways of *Saccharomyces cerevisiae* under anaerobic growth. It offers insights into other important cell physiological characteristics that can be used for controlling and directing metabolic fluxes, identifying alternative pathways, and calculating the maximum theoretical yields of the products.[6] Once the fluxes through the different branches are quantified, controlling the fluxes can be attempted. It is necessary and important to understand how the enzymes at the branch points are regulated. Various types of regulatory mechanisms have been identified, such as feedback inhibition, cooperativity, modification of covalent enzymes, and control of enzyme synthesis.[6] Determining the metabolite concentrations is also important for understanding how metabolic fluxes are controlled. In this case, GC–MS and LC–MS–MS are good choices for measuring a large number of metabolites.[3] Furthermore, to quantify the flux control, the concept of metabolic control analysis is commonly used with flux control coefficients (FCC), which is defined as the relative change in the steady-state flux, resulting from an infinitesimal change in the activity of an enzyme of the pathway, divided by the relative change of enzymatic activity.[6]

For pathway modeling, the typical aims include exploration of the possible behavior of a system, interpretation and evaluation of measured data, systematic analysis of a metabolic network structure and its qualitative behavior, and designing/predicting/optimizing the outcome of future experiments. The different modeling approaches currently used, including structure model, stoichiometric model, carbon flux model, stationary and nonstationary mechanistic model, and gene regulation model, are discussed and evaluated by Wiechert.[7]

Nucleic Acid Engineering

Previously, nucleic acids were considered as the carriers of genetic information in cells, with very few other functions. However, with the recent discovery of ribozymes (catalytic RNA), aptamers (binding RNA/DNA), and DNAzymes (catalytic DNA), nucleic acids have also been increasingly explored for diagnostic and therapeutic applications. As nucleic acids carry both structural (genotype) and functional (phenotype) information in a single molecule, they are particularly amenable to powerful in vitro selection methods such as SELEX (Systematic Evolution of Ligands by Exponential Enrichments) (Fig. 4).[8] The principle behind these selection methods is the same as that of directed protein evolution, where the DNA or RNA molecules are put through iterative rounds of diversification and selection.

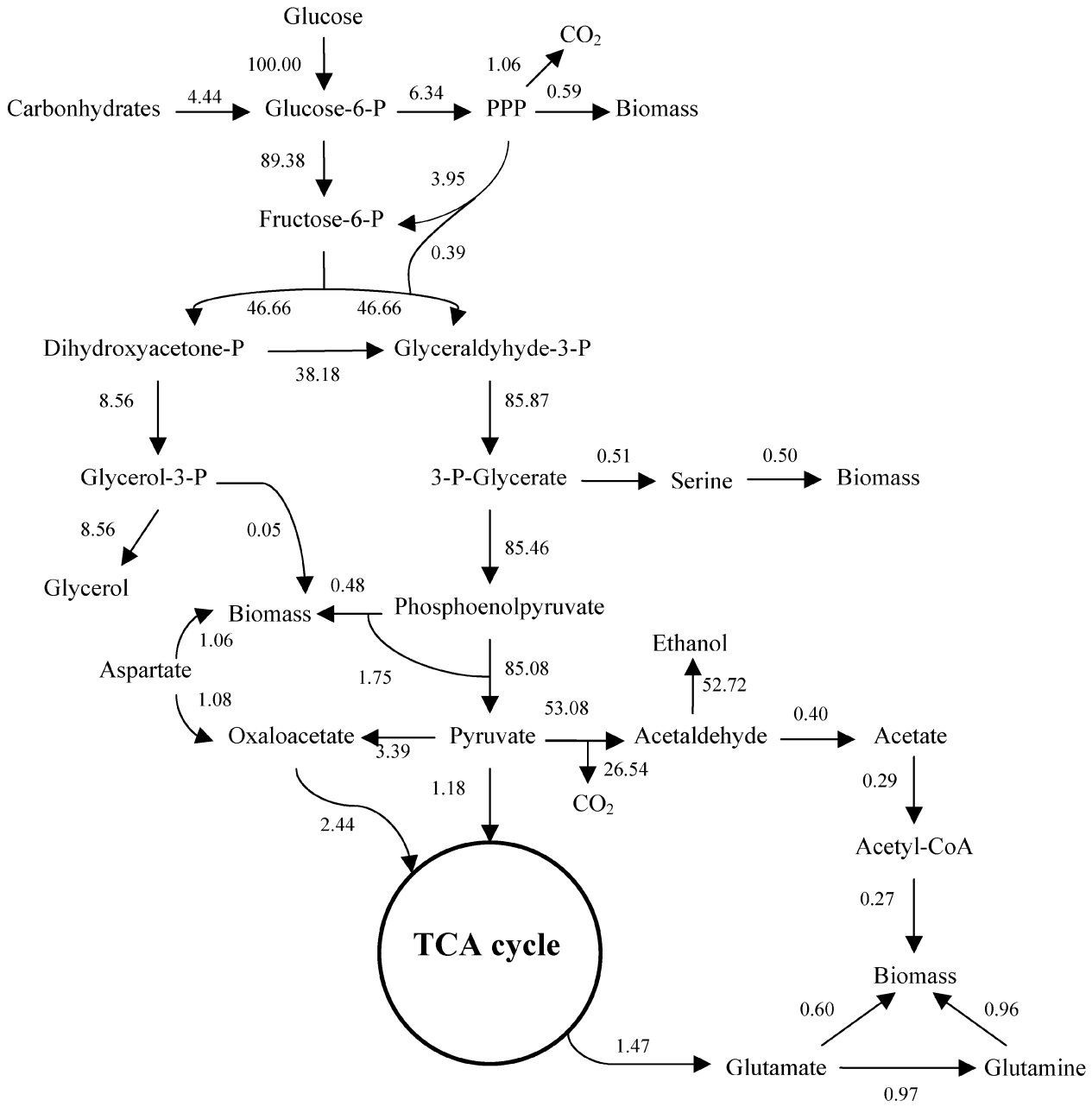


Fig. 3 An example of metabolic flux analysis showing fluxes through the different metabolic pathways of *S. cerevisiae* under anaerobic growth. (Adapted from Ref.[6].)

Although the different classes of therapeutic nucleic acids work via different mechanisms to treat diseases, the development of all forms of nucleic acid molecules into practical therapeutic agents faces a common obstacle—the omnipresence of nucleases in biological fluids. Nucleases degrade the DNA or RNA agents in the body, shortening the life span and consequently the efficiency of the nucleic acid therapeutic agents.[9] As a result, larger doses are required to achieve desired results, making the treatment very expensive. There are several methods to overcome this limitation, first of which is modification of bases. By replacing the 2′-OH-group in RNA with a host of different bases, such as 2′-NH_2– and 2′-F-pyrimidines, the stability of aptamers that were used as RNA-based therapeutic agents can be greatly increased.[10] However, 2′-modified aptamers must be compatible with the overall SELEX protocol. Therefore, any modified nucleotides used must be recognizable by the various polymerases used in the amplification process. Alternatively, the use of "Transcription Free SELEX" bypasses the need for polymerase compatibility by allowing random RNA

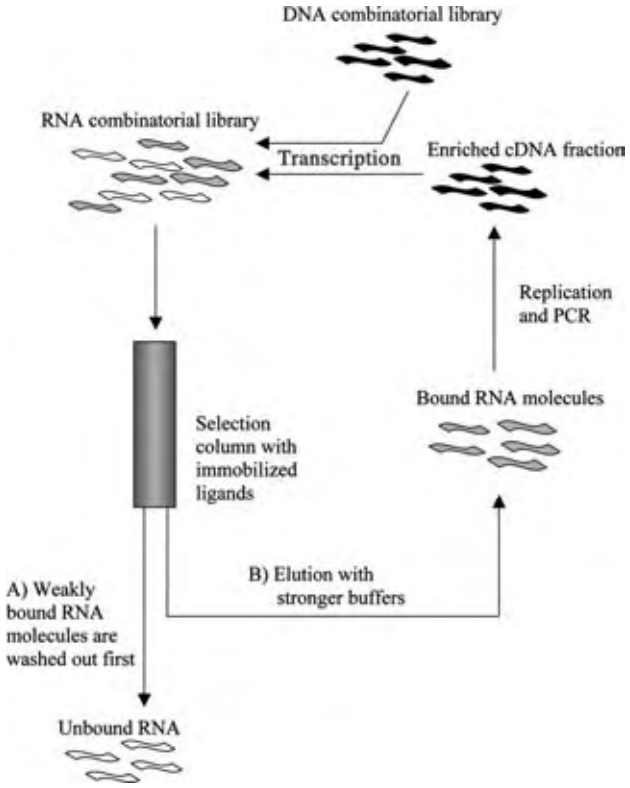


Fig. 4 The scheme of the SELEX method.

nucleotide fragments to bind to random DNA templates. Standard chemical reactions such as carbodiimide[11] or cyanogen bromide condensations[12] can then be used to ligate these fragments, and the RNA aptamers can then be recovered by melting the duplex.[13]

The modification of bases has also been applied to another type of RNA therapeutic agent, the antisense oligonucleotides. Currently, the modifications can be classified into various generations. In the first generation, modifications were done on the backbone of the oligonucleotides, a common example is the replacement of one nonbridging oxygen atom with sulfur.[14,15] These were known as phosphorothioate (PS) oligonucleotides. One successful example is Vitravene^{TM}, which is currently the only FDA-approved antisense drug in the market. The second generation contains sugar modifications, with the substitution of the 2′-OH groups being the most common. Oligonucleotides with 2′-O-alkyl derivatives were found to be more resistant to DNA- or RNA-cleaving enzymes. Currently, 2′-O-methyl and 2′-O-alkyl have already been incorporated into a number of oligonucleotides in clinical trials.[16] The third generation of antisense RNA consists of zwitterionic oligonucleotides.[17] It was found that a modification of 2′-O-aminopropyl (AP)-RNA displayed a much higher resistance to snake venom phosphodiester (SVPD) compared with a 2′-O-butyl-modified RNA.[18] AP-RNAs were able to competitively inhibit the degradation of single-stranded DNA by *Escherichia coli* Klenow fragment (KF) 3′-5′ exonuclease and SVPD. Crystal structure studies of AP-RNA revealed that the positively charged 2′-O-substituent is able to interfere with the metal-ion binding site B of the KF exonuclease, slowing down the degradation process.

Another method of improving nucleic acid stability is the use of mirror image molecules. Substrate recognition by natural nucleases is inherently stereospecific because these enzymes consist of only L-amino acids. Thus, only (D-)oligonucleotides can be recognized and degraded by the nucleases. To increase the stability of aptamers in biological fluids, nuclease resistant aptamers can be generated by using non-natural L-nucleotide aptamers, which are also known as spiegelmers (''spiegel'' meaning mirror in German).[19] However, because of the very reason that spiegelmers can escape nuclease recognition, no natural enzymes are able to recognize and amplify such nucleic acids. This greatly limits the screening of spiegelmers for potential drug agents as they cannot be directly screened using SELEX methods. To overcome this problem, unmodified D-RNA or D-DNA libraries were screened against the mirror image of the natural drug target instead.[19] Following the rules of symmetry, the mirror image of the selected D-aptamers will in turn bind to the natural drug targets.

Bioinformatics

Bioinformatics is a rapidly expanding field, involving the application of computer technology to the management of biological information. Here, only two of the key bioinformatics components related to biomolecular engineering, database and computer modeling, are discussed.

The successful sequencing of the genomes from more than 100 organisms, including humans, has led to the increasing use of genomic databases such as GenBank, European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL), and DNA Data Bank of Japan (DDBJ). Each of these three international databases collects and archives the sequencing data reported worldwide on a daily basis.[20] One extremely powerful tool for searching these nucleotide and protein databases is the Basic Local Alignment Search Tool (BLAST),[21] which compares a target DNA or protein sequence to all the existing sequences in the databases to find known sequences with high sequence similarities, and thereby obtain insights into the structure and function of the DNA or protein of interest. This tool is freely available to the public on the

National Center for Biotechnology Information (NCBI) website (www.ncbi.nlm.nih.gov/BLAST). It should be noted that many powerful tools for web-based database searching and sequence analysis have been developed, such as the San Diego Supercomputer Center Biology Workbench (SDSC Biology Workbench, workbench.sdsc.edu). In the SDSC Biology Workbench, database searching is combined with access to a wide variety of sequence analysis and modeling tools, which makes it very convenient for researchers to analyze nucleic acids and proteins.

The availability of a protein crystal structure is important for understanding the molecular basis of protein functions as well as for engineering new or improved protein functions. However, the protein structures are typically solved by x-ray crystallography and nuclear magnetic resonance (NMR), which are very time-consuming and laborious processes. As a result, the number of solved protein structures is far fewer than that of known proteins. Within this context, the use of computer-modeling algorithms and programs to rapidly create a structural model of the target protein in silico is a very attractive alternative to experimental structure-determination methods. In general, there are three basic methods for structure prediction: ab initio, homology (comparative) modeling, and fold recognition (threading).

Ab initio methods rely on the fact that the folded protein is in a state of lowest free energy; hence, the predicted structures are created through energy minimization based solely on possible interactions between the residues. As ab initio methods do not refer to existing protein structures, they are typically computation intensive. However, theoretically these methods could be superior, as, unlike homology modeling and fold recognition, they are not biased by existing structural information. Many different types of empirical force fields and levels of structure description are used in ab initio methods. Homology modeling attempts to predict protein structure based on the similarity between the sequence of a protein and the sequences of other proteins of known structures, whereas fold recognition defines a database of known structures and then fits (threads) the sequence to that database, evaluating the score that assesses the suitability of each possible fit to a given fold. For high sequence identities ($>30\%$), homology modeling usually gives a reasonably accurate model. In cases of low sequence identities ($<30\%$), threading methods often yield more accurate models than does homology modeling,[22] especially considering that there are some cases where folds are the same, yet sequences are very different (low identity).

As mentioned earlier, many bioinformatics tools were developed by individual research groups all around the world and are freely available on the Internet. In addition, there are also a few commercial software programs, such as Insight II (Accelrys Inc., San Diego, California) and SYBYL®/Base (Tripos, Inc., St. Louis, Missouri). These programs integrate force fields, energy minimization, homology modeling, molecular dynamics simulation, and molecular visualization into a single platform, making them very powerful tools for protein analysis, modeling, and design.

EXAMPLES

Protein Engineering

There are many examples of protein engineering that use rational design and directed evolution.

Metabolic Pathway Engineering

Metabolic pathway engineering has numerous applications in food, agriculture, chemical, and pharmaceutical industries. Examples include, but are not limited to, increasing the yield of antibiotics, biosynthetic precursors, or polymers, expanding the metabolic capacity to degrade harmful compounds, or producing novel compounds that cannot be found in nature.

Biosynthesis

To establish metabolic pathways in a production host that are able to channel carbon flux to a desired product at a high yield requires careful consideration of the entire metabolic microenvironment in the host. These efforts normally include four key elements: 1) direct and optimize the primary metabolic pathway flux to the target product, including removal of rate-limiting steps, and transcriptional and allosteric regulation; 2) genetically block competing branch pathways; 3) modify secondary metabolic pathways to enhance energy metabolism and availability of required enzymatic cofactors; and 4) remove detrimental side-products.[23] For example, the central glucose metabolic pathway of *E. coli* was engineered to achieve high recombinant protein production. The acetate accumulated at the end of the central glycolysis pathway affected both the recombinant protein production and the cell density detrimentally.[24] Strategies investigated to reduce acetate accumulation included modification of glucose uptake rate by a glucose analog methyl α-glucoside,[25] redirection of carbon flux toward a less inhibitory byproduct, acetoin, by introducing acetolactate synthase gene of *Bacillus subtilis* into *E. coli* to convert pyruvate to acetoin instead of acetate,[26] and elimination of critical enzymes, including acetate kinase and phosphotransacetylase that are involved directly in

the formation of acetate.[27] Another example is synthesis of hydrocortisone, an important starting material for steroidal drug synthesis, from glucose in yeast.[28] The whole process involved the natural yeast biosynthetic pathways, a yeast pathway rerouted by one plant enzyme and other five enzyme steps catalyzed by eight mammalian proteins. Experimentally, recombinant *S. cerevisiae* was engineered to express 13 assembled genes, while several unwanted side reactions brought about by endogenous yeast genes were disrupted, efficiently directing endogenous carbon flux toward the target product, hydrocortisone.

Biodegradation

Some natural organisms can degrade harmful organic compounds, referred to as xenobiotics, such as aromatics, halogenated aliphatics, and pesticides. However, these naturally occurring degradation processes are extremely slow and inefficient. Thus, speeding up these processes by metabolic engineering has become an attractive strategy. Most work in this field has focused on biodegradation of aromatic hydrocarbons, among which the TOL pathway of *Pseudomonas putida* was studied extensively.[29] The TOL pathway is a plasmid-encoded pathway and can utilize toluene, *m*-, and *p*-xylene as sole sources of carbon and energy. However, it cannot use benzene, which is often present with toluene and xylene as a mixture. Heterologous expression of a toluene dioxygenase in a *P. putida* strain carrying the TOL pathway resulted in a recombinant micro-organism that can degrade all these four aromatic compounds at a high rate.[30]

Inverse metabolic engineering

The strategy of IME starts from variant construction, which can be classified into two categories: ''exogenous'' mutagenesis and ''endogenous'' mutagenesis. The former, which is usually plasmid based, involves directed evolution of proteins, whereas the latter mainly involves evolutionary engineering of host chromosome(s). An approach to create more robust strains by random mutagenesis of the *E. coli* chromosome was successfully carried out by transforming an exogenous plasmid pmut containing the *mutD* gene into the host.[31] The *mutD* gene encodes the ε-subunit of DNA polymerase III, which is responsible for proofreading of DNA replication. *mutD5*, which carries two amino acid substitutions, is non-functional, but still competes with the functional MutD protein produced from the chromosomal copy *mutD* gene, resulting in an increase of the mutation frequency of *E. coli*. In this way, a number of *E. coli* cells with a broad spectrum of base substitutions and even frameshift mutations were obtained. Among these, it was found that three bacterial strains can tolerate dimethylformamide up to 10–20 g/L. Another approach to create novel traits in microbes is genome shuffling. In this strategy, classical strain-improvement method is used to create a population (pool) of microbes with small functional improvements that are then shuffled by recursive pool-wise protoplast fusion.[32] For example, it was reported that the acid tolerance of *Lactobacillus* was improved by genome shuffling.[33] The lower pH is desirable because, at low pH, the fermentation product of *Lactobacillus*, lactic acid, is mostly in the free-acid form and can be purified much more easily than its lactate form (which is the predominant form at higher pH). The initial genetic diversity was first created in a fermenter by slowly decreasing the fermentation pH from 6.0 to 4.1 over a period of 1200 hr, resulting in a stable population of *Lactobacilli* variants growing at pH 4.1, a pH that severely inhibits growth of wild type strain. Then genome shuffling was carried out within the selected population by means of pool-wise recursive protoplast fusion, and finally a population of microbes that can grow at pH 3.8 was obtained. The yield of lactic acid of these improved microbes was threefold higher than that of the wild type microbes at pH 4.0.

Identifying the genetic basis for the desirable phenotype remains one of the central challenges in IME. Traditional approaches include DNA sequencing of the inserts after identifying desirable clones from the plasmid-based library and evaluation of gene disruption libraries to identify the genes essential for a desired phenotype.[5] However, these approaches are quite time-consuming and expensive, because it is not easy to ensure that all the relevant genes have been analyzed or the same insert will not be repeatedly sequenced. Thus, new approaches such as DNA microarray combined with molecular bar codes[34] and insertional mutagenesis[35] were developed to address these limitations. For both approaches, a pool of cell variants was generated, with its specific gene interrupted by DNA fragments (molecular bar codes or insertional elements) that can facilitate later identification and quantification of each interrupted gene by DNA microarray. Cell variants, growing in different selective conditions competitively, were combined. Under a selective pressure, the genes involved in the biosynthesis of nutrients not provided in the media would be significantly enriched,[35] while the sequence tags of the interrupted genes would diminish in the culture if the deleted genes were important for the growth.[34] Compared with the wild type or unselected library profile, the relevant genes for a specific phenotype can be identified and quantified through DNA microarray.

Nucleic Acid Engineering

Ribozymes

Ribozymes are RNA molecules with catalytic activities. These molecules were discovered by Cech et al. in 1981,[36] and have been found to catalyze a variety of reactions in the cell such as RNA splicing, RNA processing, the replication of RNA genomes, and peptide bond formation during translation.[37] The main function of naturally occurring ribozymes is the sequence-specific cleavage of RNA molecules. Ribozymes can function either in the *cis* manner, in which they catalyze the splicing of their own RNA sequence, or in the *trans* manner, in which they catalyze the cleavage of other RNA molecules.

One of the most studied *trans*-acting ribozymes is the hammerhead ribozyme. It consists of three base-paired helices surrounding a ''core'' sequence (Fig. 5). Stems I and III bind to complementary sequences on the target RNA, and the central region catalyzes the cleavage of the RNA at the 3′ end of a UH sequence (where H = U, A, or C). Thus the ribozyme can theoretically be designed to target any RNA molecule containing the UH sequence. Hammerhead ribozymes possess great therapeutic potential and have been targeted at numerous genes, ranging from viral disease genes, such as hepatitis B and HIV-1, to cancer related genes, such as multidrug resistance (MDR-1).[37] The hammerhead ribozyme can be improved in a variety of ways such as in vitro selection to isolate ribozymes with higher activity and chemical modification of the nucleotide bases to improve nuclease resistance. These engineering methods have been covered in the review by Usman et al.[38] In a recent study, a hammerhead ribozyme (RzA) that cleaves the hepatitis B virus (HBV) poly A signal was connected to the viral encoded RNA involved in the packaging of bacterial virus Φ23 DNA into procapsids (pRNA).[39] The chimeric ribozyme (pRNA-RzA) was able to cleave the substrate HBV-polyA with nearly 100% efficiency. Furthermore, *e*-antigen assays and northern blot analysis showed that the chimeric ribozyme displayed better performance in inhibiting the HBV-polyA than the ribozyme alone in cell cultures. These findings suggest that pRNA can chaperone and escort the hammerhead ribozyme to function in the cell, enhancing the cleavage efficiency and inhibition effect of the ribozyme on HBV.

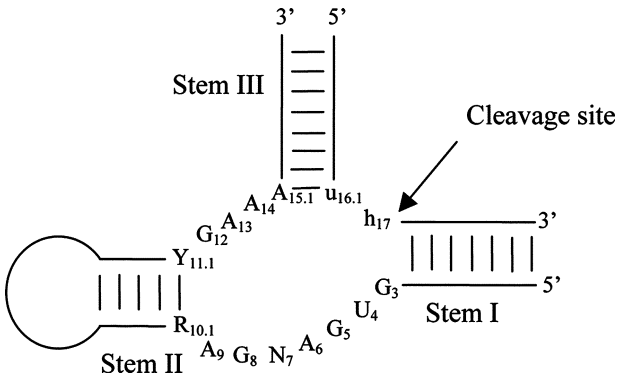


Fig. 5 The hammerhead ribozyme, showing the three non-conserved stems and the cleavage site. Ribozyme nucleotides are in uppercase letters, while substrate nucleotides are in lowercase. Y represents C or U; R represents A or G; N represents A, C, U, or G; h represents A, C, or U. (From Ref.[48].)

Another type of ribozyme is the Group I intron, which carries out a self-splicing reaction via a two-step *trans*-esterification reaction. As such, *trans*-splicing Group I ribozymes can be used as a form of treatment for genetic disorders as they can repair RNA transcripts from mutated genes. The Group I ribozyme recognizes the target mRNA by base pairing to an accessible region of the transcript upstream of a non-sense or mis-sense mutation via an internal guide sequence (IGS). After base pairing, the ribozyme cleaves off the downstream region and splices in a corrected transcript to restore the correct genetic information (Fig. 6).[40] The Group I ribozyme has been adopted to mediate the repair transcripts of mutated p53 gene, a tumor suppressor gene that is mutated in many cancers.[41] Using in vitro selection, two ribozymes, Rib41 and Rib65, which were able to cleave the majority of p53 transcripts and yield products of the correct size, were isolated from a ribozyme library containing randomized IGS, 5′-GNNNNN-3′ (where N represents any of the four nucleotides). These ribozymes were able to repair the defective p53 RNA transcripts with high fidelity and specificity. In addition, the corrected transcripts were found to be functionally translated, resulting in a 23-fold induction of a p53 responsive promoter and a threefold reduction in the MDR-1 gene promoter.

DNA enzymes

DNA has long been regarded as a passive molecule that is ideal for carrying genetic information but is structurally monotonous and thus functionally impoverished. However, this notion was changed with the discovery of catalytic RNA.[42] Since then, DNA enzymes, which are cation-dependent enzymatic molecules composed entirely of DNA, have been developed.[43,44]

In the seminal work by Breaker and Joyce,[43] a mixture of random N_{50} DNA oligomers were tethered to a matrix by a RNA nucleotide-containing linker.

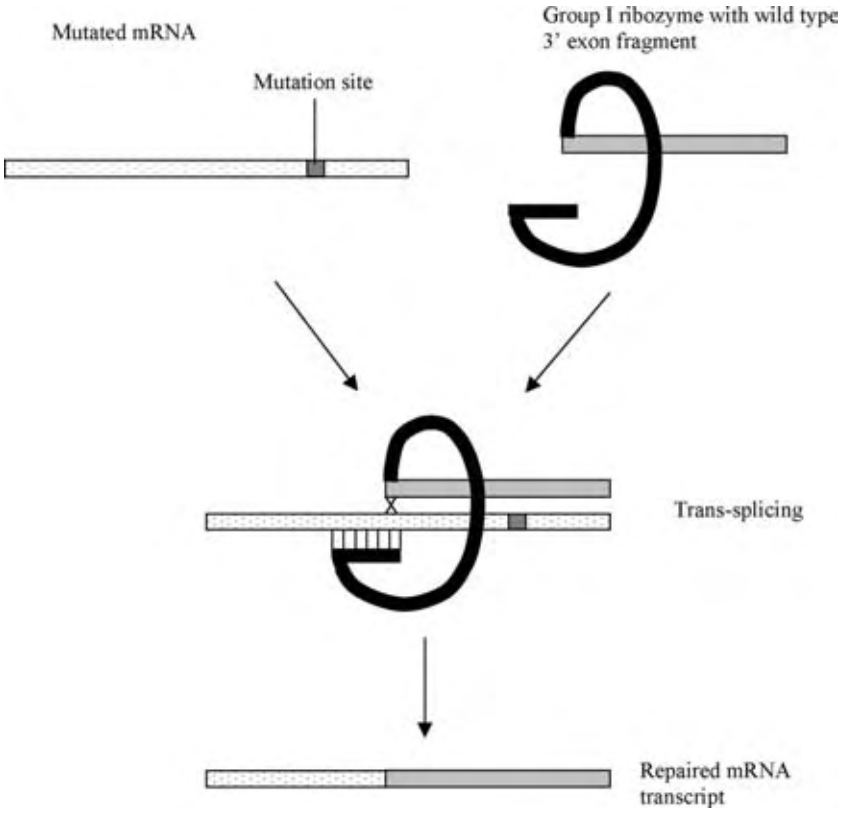


Fig. 6 The scheme of mRNA repair by Group I ribozyme.

When washed with lead solution, oligomers capable of utilizing lead ions to cleave the tethering RNA were eluted together with the solution. Hence the DNA enzymes were able to ''select'' themselves. It was also found that these enzymes can be engineered to cleave RNA substrates specifically and efficiently. DNA enzymes with 10–23 motifs or 8–17 motifs are prominent examples of such biomolecules. Both the DNA enzymes consist of a nucleotide catalytic core flanked by two substrate-binding arms of seven to eight bases each.[44] By changing the substrate-binding domain sequence, the DNA enzyme can be designed to target the AU nucleotides of different mRNA substrates with high specificity.

One potential target for DNA enzyme therapy is angiogenesis of solid tumors. Although angiogenesis is an important step in tumor development, current understanding of key transcription factors regulating it is limited.[45] However, it is known that one gene involved in angiogenesis is the early growth response (Egr-1) gene. A 10–23 motif DNA enzyme designed to target a specific motif in the 5′-untranslated region of Egr-1 mRNA was able to inhibit Egr-1 protein expression, microvascular endothelial cell replication and migration, and microtubule network formation on basement membrane matrices.[45] In addition, the Egr-1 DNA enzymes suppressed tumor growth by fivefold without influencing body weight, wound healing, blood coagulation, or other hematogical parameters of mice. This study not only shed light on the understanding of key transcription factors regulating angiogenesis but also showed the huge potential of DNA enzyme as therapeutic drug in cancer treatments.

Aptamers

Aptamers are ligands of double-stranded DNA or single-stranded RNA that can be generated against amino acids, drugs, proteins, and other molecules. They have potential applications in analytical devices, such as biosensors, and as therapeutic agents. When combined with in vitro selection techniques,[8,46] aptamers can become powerful screening tools for rapidly identifying targets for new pharmaceuticals.

A spiegelmer, NOX 1255, with a high affinity (K_D = 20 nM) for the peptide hormone, gonadotropin-releasing hormone (GnRH), was isolated by using the mirror image selection method.[47] The stability of this molecule was further increased by adding a 40 kDa polyethylene glycol moiety onto its 5′ end (NOX 1257). The spiegelmer was also found to be highly specific as it did not show any inhibition by buserelin, a peptide analog of GnRH. Furthermore, the anti-GnRH spiegelmer

exhibited a very low immunogenic potential in in vivo experiments with rabbits, suggesting that spiegelmers are not recognized by the immune system and can be administered repeatedly over long periods without adverse effects. This study demonstrates the potential of spiegelmers as a new pharmaceutical approach against GnRH and other targets.

Selected spiegelmers must be synthesized chemically, as no enzyme is able to synthesize L-DNA or L-RNA. However, with the current technology the efficient chemical synthesis of oligoribonucleotides is limited to 60 nt. Typical spiegelmers identified through the SELEX method are 60–90 nt long consisting of a 30–40 nt long randomized region and fixed primer sites of about 15–25 nt on each end. Truncation of these flanking primer regions from selected spiegelmers is required before further testing in biological systems. However, the flanking primers may affect the binding ability of the spiegelmer to its target and spiegelmers may not work well if they are truncated. To reduce this uncertainty, Vater et al. developed a strategy known as Tailored-SELEX to minimize the number of fixed primer regions in a spiegelmer.[9] Using this strategy, a rat α-CGRP binding spiegelmer with a calculated affinity of $K_D = 2.5\,nM$ at 37°C, which is close to the binding constant of the neuropeptide to its receptor ($EC_{50} = 1\,nM$), was isolated.

CONCLUSIONS

With the advances in high-throughput screening method for discovering target biomolecules and the accumulation of data in functional genomics and proteomics, the rate of designing and discovering valuable biomolecules will rapidly grow. Many enzymes and proteins have been engineered for process improvement and generation of high-value products at a low cost. A few biomolecular techniques have been utilized in disease diagnosis, and some engineered biomolecules have entered clinical trials for therapeutic uses. When the enabling technologies such as protein engineering, metabolic engineering, microarray, mass spectrometry, and bioinformatics grow more mature, and when our understanding of cells becomes more thorough, biomolecular engineering will no doubt become one of the most important research areas in both academia and industry.

ACKNOWLEDGMENT

We thank the Office of Naval Research and National Science Foundation for supporting our work in biomolecular engineering.

ARTICLES OF FURTHER INTEREST

Biocatalysis, p. 101.
Protein Design, p. 2467.

REFERENCES

1. Ryu, D.D.; Nam, D.H. Recent progress in biomolecular engineering. Biotechnol. Prog. **2000**, *16* (1), 2–16.
2. Stephanopoulos, G. Metabolic fluxes and metabolic engineering. Metab. Eng. **1999**, *1* (1), 1–11.
3. Nielsen, J. Metabolic engineering. Appl. Microbiol. Biotechnol. **2001**, *55* (3), 263–283.
4. Bailey, J.E.; Sburlati, A.; Hatzimanikatis, V.; Lee, K.; Renner, W.A.; Tsai, P.S. Inverse metabolic engineering: a strategy for directed genetic engineering of useful phenotypes. Biotechnol. Bioeng. **2002**, *79* (5), 568–579.
5. Gill, R.T. Enabling inverse metabolic engineering through genomics. Curr. Opin. Biotechnol. **2003**, *14* (5), 484–490.
6. Stephanopoulos, G.N.; Aristidou, A.A.; Nielsen, J. *Metabolic Engineering: Principles and Methodologies*; Academic Press: London, U.K., 1998.
7. Wiechert, W. Modeling and simulation: tools for metabolic engineering. J. Biotechnol. **2002**, *94* (1), 37–63.
8. Tuerk, C.; Gold, L. Systematic evolution of ligands by exponential enrichment—RNA ligands to bacteriophage-T4 DNA-polymerase. Science **1990**, *249* (4968), 505–510.
9. Vater, A.; Jarosch, F.; Buchner, K.; Klussmann, S. Short bioactive spiegelmers to migraine-associated calcitonin gene-related peptide rapidly identified by a novel approach: Tailored-SELEX. Nucl. Acids Res. **2003**, *31* (21), e130.
10. Eulberg, D.; Klussmann, S. Spiegelmers: Biostable aptamers. Chem. Biol. Chem. **2003**, *4* (10), 979–983.
11. Dolinnaya, N.G.; Sokolova, N.I.; Gryaznova, O.I.; Shabarova, Z.A. Site-directed modification of DNA duplexes by chemical ligation. Nucl. Acids Res. **1988**, *16* (9), 3721–3738.
12. Dolinnaya, N.G.; Sokolova, N.I.; Ashirbekova, D.T.; Shabarova, Z.A. The use of Brcn for assembling modified DNA duplexes and DNA–RNA hybrids—Comparison with water-soluble carbodiimide. Nucl. Acids Res. **1991**, *19* (11), 3067–3072.
13. Smith, J.D.; Gold, L. Transcription-Free SELEX US Patent 6,387,620, 2002.
14. Agrawal, S.; Goodchild, J.; Civeira, M.P.; Thornton, A.H.; Sarin, P.S.; Zamecnik, P.C. Oligodeoxynucleoside phosphoramidates and

phosphorothioates as inhibitors of human immunodeficiency virus. Proc. Natl. Acad. Sci. U.S.A. **1988**, *85* (19), 7079–7083.
15. Matsukura, M.; Shinozuka, K.; Zon, G.; Mitsuya, H.; Reitz, M.; Cohen, J.S.; Broder, S. Phosphorothioate analogs of oligodeoxynucleotides: Inhibitors of replication and cytopathic effects of human immunodeficiency virus. Proc. Natl. Acad. Sci. U.S.A. **1987**, *84* (21), 7706–7710.
16. Corey, D.R. Telomerase inhibition, oligonucleotides, and clinical trials. Oncogene **2002**, *21* (4), 631–637.
17. Urban, E.; Noe, C.R. Structural modifications of antisense oligonucleotides. Farmaco **2003**, *58* (3), 243–258.
18. Teplova, M.; Wallace, S.T.; Tereshko, V.; Minasov, G.; Symons, A.M.; Cook, P.D.; Manoharan, M.; Egli, M. Structural origins of the exonuclease resistance of a zwitterionic RNA. Proc. Natl. Acad. Sci. U.S.A. **1999**, *96* (25), 14240–14245.
19. Klussmann, S.; Nolte, A.; Bald, R.; Erdmann, V.A.; Furste, J.P. Mirror-image RNA that binds D-adenosine. Nat. Biotechnol. **1996**, *14* (9), 1112–1115.
20. Stoesser, G.; Sterk, P.; Tuli, M.A.; Stoehr, P.J.; Cameron, G.N. The EMBL nucleotide sequence database. Nucl. Acids Res. **1997**, *25* (1), 7–14.
21. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. J. Mol. Biol. **1990**, *215* (3), 403–410.
22. Schoonman, M.J.; Knegtel, R.M.; Grootenhuis, P.D. Practical evaluation of comparative modelling and threading methods. Comput. Chem. **1998**, *22* (5), 369–375.
23. Chotani, G.; Dodge, T.; Hsu, A.; Kumar, M.; LaDuca, R.; Trimbur, D.; Weyler, W.; Sanford, K. The commercial production of chemicals using pathway engineering. Biochim. Biophys. Acta **2000**, *1543* (2), 434–455.
24. Yang, Y.; Bennett, G.N.; San, K.Y. Genetic and metabolic engineering. J. Biotechnol. **1998**, *1* (3), 134–141.
25. Chou, C.H.; Bennett, G.N.; San, K.Y. Effect of modulated glucose uptake on high-level recombinant protein production in a dense *Escherichia coli* culture. Biotechnol. Prog. **1994**, *10* (6), 644–647.
26. Aristidou, A.A.; San, K.Y.; Bennett, G.N. Modification of central metabolic pathway in *Escherichia coli* to reduce acetate accumulation by heterologous expression of the *Bacillus subtilis* acetolactate synthase gene. Biotechnol. Bioeng. **1994**, *44* (8), 944–951.
27. Yang, Y.T.; Aristidou, A.A.; San, K.Y.; Bennett, G.N. Metabolic flux analysis of *Escherichia coli* deficient in the acetate production pathway and expressing the *Bacillus subtilis* acetolactate synthase. Metab. Eng. **1999**, *1* (1), 26–34.
28. Szczebara, F.M.; Chandelier, C.; Villeret, C.; Masurel, A.; Bourot, S.; Duport, C.; Blanchard, S.; Groisillier, A.; Testet, E.; Costaglioli, P.; Cauet, G.; Degryse, E.; Balbuena, D.; Winter, J.; Achstetter, T.; Spagnoli, R.; Pompon, D.; Dumas, B. Total biosynthesis of hydrocortisone from a simple carbon source in yeast. Nat. Biotechnol. **2003**, *21* (2), 143–149.
29. Assinder, S.J.; Williams, P.A. The TOL plasmids: determinants of the catabolism of toluene and the xylenes. Adv. Microb. Physiol. **1990**, *31*, 1–69.
30. Lee, J.Y.; Jung, K.H.; Kim, H.S. Amplication of toluene dioxygenase genes in a hybrid pseudomonas strain to enhance the biodegradation of benzene, toluene, and *p*-xylene mixture. Biotechnol. Bioeng. **1995**, *45* (6), 488–494.
31. Selifonova, O.; Valle, F.; Schellenberger, V. Rapid evolution of novel traits in microorganisms. Appl. Environ. Microbiol. **2001**, *67* (8), 3645–3649.
32. Zhang, Y.X.; Perry, K.; Vinci, V.A.; Powell, K.; Stemmer, W.P.; del Cardayre, S.B. Genome shuffling leads to rapid phenotypic improvement in bacteria. Nature **2002**, *415* (6872), 644–646.
33. Patnaik, R.; Louie, S.; Gavrilovic, V.; Perry, K.; Stemmer, W.P.; Ryan, C.M.; del Cardayre, S. Genome shuffling of *Lactobacillus* for improved acid tolerance. Nat. Biotechnol. **2002**, *20* (7), 707–712.
34. Giaever, G.; Chu, A.M.; Ni, L.; Connelly, C.; Riles, L.; Veronneau, S.; Dow, S.; Lucau-Danila, A.; Anderson, K.; Andre, B.; Arkin, A.P.; Astromoff, A.; El-Bakkoury, M.; Bangham, R.; Benito, R.; Brachat, S.; Campanaro, S.; Curtiss, M.; Davis, K.; Deutschbauer, A.; Entian, K.D.; Flaherty, P.; Foury, F.; Garfinkel, D.J.; Gerstein, M.; Gotte, D.; Guldener, U.; Hegemann, J.H.; Hempel, S.; Herman, Z.; Jaramillo, D.F.; Kelly, D.E.; Kelly, S.L.; Kotter, P.; LaBonte, D.; Lamb, D.C.; Lan, N.; Liang, H.; Liao, H.; Liu, L.; Luo, C.; Lussier, M.; Mao, R.; Menard, P.; Ooi, S.L.; Revuelta, J.L.; Roberts, C.J.; Rose, M.; Ross-Macdonald, P.; Scherens, B.; Schimmack, G.; Shafer, B.; Shoemaker, D.D.; Sookhai-Mahadeo, S.; Storms, R.K.; Strathern, J.N.; Valle, G.; Voet, M.; Volckaert, G.; Wang, C.Y.; Ward, T.R.; Wilhelmy, J.; Winzeler, E.A.; Yang, Y.; Yen, G.; Youngman, E.; Yu, K.; Bussey, H.; Boeke, J.D.; Snyder, M.; Philippsen, P.; Davis, R.W.; Johnston, M. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature **2002**, *418* (6896), 387–391.
35. Badarinarayana, V.; Estep, P.W., III; Shendure, J.; Edwards, J.; Tavazoie, S.; Lam, F.; Church, G.M.

Selection analyses of insertional mutants using subgenic-resolution arrays. Nat. Biotechnol. **2001**, *19* (11), 1060–1065.
36. Cech, T.R.; Zaug, A.J.; Grabowski, P.J. In vitro splicing of the ribosomal RNA precursor of Tetrahymena: Involvement of a guanosine nucleotide in the excision of the intervening sequence. Cell **1981**, *27* (3 Pt 2), 487–496.
37. Puerta-Fernandez, E.; Romero-Lopez, C.; Barroso-delJesus, A.; Berzal-Herranz, A. Ribozymes: Recent advances in the development of RNA tools. FEMS Microbiol. Rev. **2003**, *27* (1), 75–97.
38. Usman, N.; Beigelman, L.; McSwiggen, J.A. Hammerhead ribozyme engineering. Curr. Opin. Struct. Biol. **1996**, *6* (4), 527–533.
39. Hoeprich, S.; Zhou, Q.; Guo, S.; Shu, D.; Qi, G.; Wang, Y.; Guo, P. Bacterial virus phi29 pRNA as a hammerhead ribozyme escort to destroy hepatitis B virus. Gene Therapy **2003**, *10* (15), 1258–1267.
40. Sullenger, B.A.; Cech, T.R. Ribozyme-mediated repair of defective mRNA by targeted, trans-splicing. Nature **1994**, *371* (6498), 619–622.
41. Watanabe, T.; Sullenger, B.A. Induction of wild-type p53 activity in human cancer cells by ribozymes that repair mutant p53 transcripts. Proc. Natl. Acad. Sci. U.S.A. **2000**, *97* (15), 8490–8494.
42. Kruger, K.; Grabowski, P.J.; Zaug, A.J.; Sands, J.; Gottschling, D.E.; Cech, T.R. Self-splicing RNA—Auto-excision and auto-cyclization of the ribosomal-RNA intervening sequence of *Tetrahymena*. Cell **1982**, *31* (1), 147–157.
43. Breaker, R.R.; Joyce, G.F. A DNA enzyme with Mg^{2+}-dependent RNA phosphoesterase activity. Chem. Biol. **1995**, *2* (10), 655–660.
44. Santoro, S.W.; Joyce, G.F. A general purpose RNA-cleaving DNA enzyme. Proc. Natl. Acad. Sci. U.S.A. **1997**, *94* (9), 4262–4266.
45. Fahmy, R.G.; Dass, C.R.; Sun, L.Q.; Chesterman, C.N.; Khachigian, L.M. Transcription factor Egr-1 supports FGF-dependent angiogenesis during neovascularization and tumor growth. Nat. Med. **2003**, *9* (8), 1026–1032.
46. Ellington, A.D.; Szostak, J.W. In vitro selection of RNA molecules that bind specific ligands. Nature **1990**, *346* (6287), 818–822.
47. Wlotzka, B.; Leva, S.; Eschgfaller, B.; Burmeister, J.; Kleinjung, F.; Kaduk, C.; Muhn, P.; Hess-Stumpp, H.; Klussmann, S. In vivo properties of an anti-GnRH Spiegelmer: an example of an oligonucleotide-based therapeutic substance class. Proc. Natl. Acad. Sci. U.S.A. **2002**, *99* (13), 8898–8902.
48. Hertel, K.J.; Pardi, A.; Uhlenbeck, O.C.; Koizumi, M.; Ohtsuka, E.; Uesugi, S.; Cedergren, R.; Eckstein, F.; Gerlach, W.L.; Hodgson, R. Numbering system for the hammerhead. Nucl. Acids Res. **1992**, *20* (12), 3252.

Bioprocess and Chemical Process Intensification

G. Akay
Process Intensification and Miniaturization Centre, School of Chemical Engineering and Advanced Materials and Institute for Nanoscale Science and Technology, University of Newcastle, Newcastle Upon Tyne, U.K.

INTRODUCTION

Process intensification (PI), motivated by plant cost reduction, emerged as a design strategy in which the processing volume is reduced (and ultimately miniaturized) while enhancing the processing field strength. Process intensification can be based on purely physical considerations. However, when there is a synergy between miniaturized processing volume and intensified processing field, PI can be considered to be inherently intensive or phenomenon based. It is therefore often necessary to consider PI with process miniaturization. Literature on PI and PI miniaturization (PIM) is reviewed within a unifying framework that allows the identification of intensification field(s) and the underlying phenomenon responsible for PI. The scope and limitations of intensified processes are examined, and through the survey of patent literature, PIM-based technology transfers are revealed. Although some of the technology transfer involves the replacement of existing unit operations with more efficient intensified processes, there are examples where PI, miniaturization, and integration are combined to address the ultimate aim of these design objectives. Despite the possibility of achieving bioprocess intensification (BI) that approaches 100-fold or more, compared with the existing technology, such intensifications are not based on that of processing field but on miniaturization in which the biocatalyst–support interactions are important. Therefore, microbioreactors coupled with highly selective biocatalysts can achieve intensification over 100-fold, making biotransformations more cost-effective compared with chemical PI. In addition to microreactor fabrication, PIM also requires the development of catalysts with not only enhanced surface area and selectivity but also accessibility of the active sites. Active site accessibility can be achieved through the development of nanostructured microporous catalysts and catalyst support with arterial channels. Such systems will enhance the scope and acceptance of PIM as they provide a clear path for sustainable technology.

When first introduced during the 1970s, PI represented a novel design strategy that aimed at reduction in the processing volume by at least an order of magnitude, compared with the existing technology, but without any reduction in process output. This restricted view of PI was relativistic and represented a design objective driven primarily by cost savings, because a significant part of the plant cost is associated with piping, support structures, and civil engineering items.[1–3] On the other hand, process miniaturization (PM) in the chemical industry has existed and continues to make excellent progress in the form of analytical equipment and sensors. Therefore, for a given production objective (or processing rate), PI and PM represent top-down and bottom-up approaches in process design, respectively. These two design approaches can be integrated with the sole aim of plant size reduction to provide major savings on capital and operating costs.

However, the integration of PIM also creates synergy in the development of intensified processes, novel product forms, and size dependent phenomena, which in turn provides novel intensified processes. Process intensification-miniaturization is seen as an important element of sustainable development because it can deliver: 1) at least a 10-fold decrease in process equipment volume; 2) elimination of parasitic steps and unwanted by-products, thus eliminating some downstream processing operations; 3) inherent safety because of reduced reactor volume; 4) novel product forms; 5) energy, capital, and operating cost reduction, and an environment friendly process; 6) plant mobility, responsiveness, and security; and 7) a platform for other technologies.

DRIVING FORCES FOR PI

Process intensification is achieved by the superimposition of two or more processing fields (such as various types of flow, centrifugal, sonic, and electric fields), by operating at ultrahigh processing conditions (such as deformation rate and pressure), a combination of the two, or by providing selectivity or extended interfacial area or a capacity for transfer processes. In heat and mass transfer operations, drastic reduction in diffusion/conduction path results in equally impressive transfer rates. As the processing volume (such as reactor

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120022532

size) is reduced by several orders of magnitude, the processing equipment ultimately becomes miniaturized.

The summary of PI driving forces, type of intensification/associated phenomenon, and their applications are shown in Table 1.

TYPES OF PI: PHYSICAL AND PHENOMENON-BASED INTENSIFICATIONS

Operations at high/ultrahigh process field strengths (driving force) such as pressure, temperature, or reduction in conduction/diffusion path often increase transport processes, and such processes can then be deemed to be intensified. Process integration can also have the same effect, often providing energy efficiency and reduced reactor volume. We can call this physical PI, which is limited by the availability of manufacturing technology to carry out such processes.

Miniaturized systems where the transport processes occur across a length scale of 100–0.1 μm (or less) not only provide a reduced diffusion/conduction path length, but can also offer process selectivity. When the enhanced transfer rates are solely caused by reduction in length scale or enhanced surface area, the intensification is still physical. However, when there is an underlying phenomenon associated with a small processing volume, then this phenomenon can be used to obtain PI. In a large number of physical, chemical, and biological processes stage-wise operations are common. When such processes are repeated several thousand times in microscopic volumes, process selectivity at each stage results in PI. There are four main reasons for PI in miniaturized systems: 1) reduced diffusion/conduction path, which can be viewed as a physical means of PI; 2) interaction between the walls of the microscopic processing volume and the fluid microstructure; 3) the ability to control the transport (momentum, heat, and mass) conditions within a small processing volume, which in turn can enhance the selectivity/efficiency of the target reaction/process; and 4) behavior of matter and microorganisms within the microenvironment.

Clearly, the above argument indicates the multidisciplinary nature of PI, which involves materials science and engineering and a detailed understanding of the molecular mechanisms of chemical processes and bioprocesses. Chemical PI and BI based on selectivity are therefore associated with an underlying phenomenon, the macroscopic manifestations of which result in

Table 1 Summary of driving forces in process intensification

Driving force	Type of intensification	Applications
Body forces	Heat and mass transfer	Rotating packed beds Spinning disks
Flow field/fluid microstructure interactions (TSVs $\Leftrightarrow$ DSVs)[a]	Flow induced phase inversion (FIPI) Flow induced phase change	Agglomeration, Microencapsulation Emulsification Structuring
High/ultrahigh pressure	Mass transfer and chemical reaction	Food processing Chemical reaction
Electric field	Mass transfer/selectivity	Cross-flow filtration Phase separation Chemical synthesis Powder dispersion
Ultrasonic and sonoelectric field	Mass transfer and chemical reaction	Cross-flow filtration Phase separation Chemical synthesis Powder dispersion
Diffusion/conduction path reduction	Heat and mass transfer	Thin film heat exchangers Membrane processes Miniaturization Catalysis
Carrier mediation	Mass transfer	Surfactant-based separations Pore accessibility
Size dependent phenomena	Selective biochemical reactions	In vitro organs In vitro biochemical synthesis Bioprocess intensification

[a]Thermodynamic state variables (TSVs) and deformation state variables (DSVs) are interchangeable in FIPI/phase change.
(From Ref.[1].)

intensification. Such an intensification technique is termed as inherently intensive PI or phenomenon-based PI in which the small processing volume is a prerequisite to an intensified process viability and efficiency. The interaction between the process driving force, processing volume, and type of intensification is shown in Fig. 1.

BIOPROCESS INTENSIFICATION

Process intensification in biotechnology has certain inherent restrictions in terms of "intensification fields," or PI-driving forces, such as temperature, pressure, concentration of reactants/products, mechanical stresses and deformation rates, and the electric field.[4–8] The above-mentioned PI fields/driving forces are commonly used in chemical PI, often in combination. Chemical PI increases with increasing field strength and therefore the PI is only limited by the limits of the reactor engineering. However, in most cases, the above-referred intensification fields cannot be used in BI. Because of these restrictions on the type of PI-driving forces, BI can therefore be achieved, in the first instance, through the reduction of the diffusion path for the reactants and products, and through the creation of the most suitable environment for the biocatalysts and micro-organisms, which can enhance selectivity, resulting in phenomenon-based intensification. It is likely that the optimization of the strength and type of the intensification field will be required in BI.

PHYSICAL PI

Physical PI based on the intensification of the processing field and subsequent reduction in processing volume as a result of enhanced momentum, heat, and mass transfer is mostly encountered in multiphase systems. In such systems, the primary aim is to create a large surface area between the phases and to achieve a high-interface renewal rate and product removal rate. Some of these processes are commonly provided using stirred-tank reactors in which the volume is very large and the process driving force (shear or extension rate) is low, nonselective with a broad distribution.

It is well known that mixing can be conducted using flow instabilities, especially when the instability is time dependent, such as those encountered in capillary entrance flows of viscoelastic fluids, so that the instability in the form of vortices causes mixing and the primary flow discharges the mixed "batch" from the mixing unit. Flow instabilities in thin liquid films also have the same effect, and in all cases, a small mixing volume is utilized at a very high rate, because such instabilities require a high flow rate (which can be quantified by the Reynolds number) for a given flow geometry. There are a number of "intensified" reactors available that can be further modified to include other processing fields (such as ultrasound and electric field) superimposed on the flow field. Process intensifications based on the generation of flow instabilities are summarized below.

Spinning Disk Reactors

Spinning disk reactors (SDRs) have been developed for the intensification of processes involving low-viscosity liquid-phase reactions, including polymerization, crystallization, and catalytic organic reactions.[9–12] Thin films with intense surface ripples are generated under the effect of high acceleration fields created by rotation of the disk surface. Film thicknesses of the order of 50–200 μm are typical. Vigorous mixing characteristics and enhanced heat and mass transfer rates, obtained in

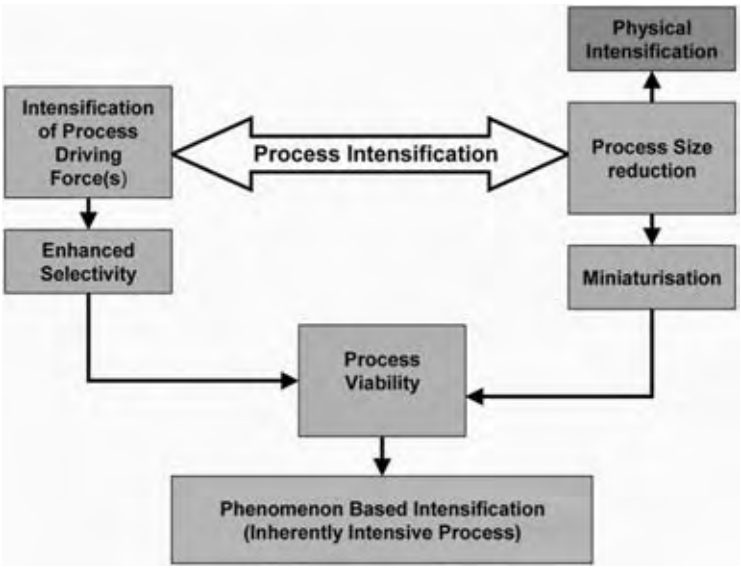


Fig. 1 Relationship between PI fields, physical PI, process miniaturization, selectivity, process viability, and phenomenon-based intensification. *(View this art in color at www.dekker.com.)*

the thin films, make the SDR an ideal reactor for operations involving fast reactions, which require rapid mixing/mass transfer and highly exothermic reactions that require an efficient heat removal system. The thin films also offer the potential of uniform penetration of ultraviolet (UV) radiation for the efficient processing of UV-initiated polymerizations.

The thin film characteristics of the SDR have been exploited for polymerization reactions proceeding by three mechanisms: free-radical polymerizations initiated by a chemical initiator and by UV-irradiation and condensation polymerizations.[9,10,11] Generally, faster rates of polymerization were observed in the SDR than in a batch reactor or a thin static film system with the SDR also producing a polymer with controlled molecular weight. For instance, the photopolymerization of *n*-butyl acrylate in the SDR showed that a high conversion (>90%), high molecular weight polymer with narrow MWD (M_n up to 35,000, M_w up to 70,000, and polydispersity index of 1.8–2.4) was obtained from the SDR under exposure to modest UV light intensities of 25 mW/cm^2, at residence times of about 2 sec.[10] In contrast, only 30% conversion could be achieved in a thin static film, after a longer exposure time of 10 sec at the same UV intensity, with a lower molecular weight polymer being formed. These results demonstrate that the high-intensity mixing taking place with the thin SDR film allows the polymerization to proceed faster than under static thin film conditions and that a better-quality polymer can also be obtained from processing in the SDR.

The vigorous mixing regime in the thin films of the SDR has also been exploited for crystallization processes.[12] Homogeneous nucleation in the thin films results in a narrow crystal size distribution for crystals formed in the SDR. Uniformly shaped barium sulfate crystals of 0.5–1 μm have been produced with a reasonably low power dissipation of 115 W/kg.[12]

Oscillatory Flow Reactors

Oscillatory flows are known to increase the transfer processes.[13] In oscillatory flow reactors (OFRs), there is a controlled stage-wise mixing present, which is achieved using tubes fitted with low-constriction orifice plate baffles and oscillatory motion (range: 0.5–10 Hz) superimposed upon the net flow of the process fluid.[14,15] The combination of the baffles and the oscillatory motion creates a flow pattern conducive to efficient heat and mass transfer while maintaining plug flow. Unlike conventional tubular reactors where a minimum Reynolds number must be maintained, the degree of mixing is independent of the net flow, allowing long residence times to be achieved in a reactor of greatly reduced length-to-diameter ratio.

An OFR can therefore be much more compact than conventional plug flow reactor designs, allowing reactor designs for longer residence times to be of practical dimensions. The main niche application envisaged is to allow the conversion of long residence time batch processes to continuous processing, when other continuous reactor designs are too expensive or impractical. This is not the only application of this technology: it can be used in a range of unit operations, including heat and mass transfer, multiphase mixing, particle suspension, bioreactions, and fermentations. Some selected examples are given below.

Mass transfer rates have been shown to be increased in yeast cultures because of increased breakup of bubbles and increased gas holdup resulting in enhanced aeration. Three-phase heterogeneous photocatalytic OFR that suspends catalyst-coated particles while they are contacted with air and exposed to UV radiation from an axially located lamp represents a good example of the utilization of several processing fields to enhance reaction rates. In this reactor, particle disengagement occurs in a baffle-free zone at the top of the column. Conversion of a batch saponification reaction to continuous processing in an OFR has been shown to result in a 100-fold reduction in reactor size, as well as greater operational control and flexibility. The mixing in OFRs is very uniform, as the radial and axial velocities are of the same magnitude. This has been shown to be an advantage when handling shear-sensitive materials, such as certain pharmaceutical crystals and in flocculators. The scale-up and design of OFRs are well understood. Indeed, the understanding of the OFR's scale-up is believed to be one of its advantages. A full review of the history and further details of the theory and applications of OFR technology are available.[15]

Controlled Deformation Static or Dynamic Mixers/Reactors

The above-described mixers are essentially low-viscosity devices. In many operations where the viscosity is high, when dealing with concentrated multiphase gas–liquid–solid binary or tertiary systems, or when liquid-to-solid phase transformation occurs during mixing, novel equipment designs are needed to intensify the heat/mass transfer processes. The multiphase fluids also represent an important class of materials that have microstructure developed during processing and subsequently ''frozen-in,'' ready for use as a product. To deliver certain desired functions, the control of microstructure in the product is important. This microstructure is developed in most cases by the interaction between the fluid flow and the fluid microstructure; hence, uniformity of the flow field is important.

Typical products with microstructure are detergents, emulsions, suspensions, paints, food emulsions, margarines, low-fat spreads, ice cream, cosmetics, personal care products, microcapsules, and agglomerated powders. The classical processing of these materials aims for low-viscosity operations, which usually result in either dilute systems or high operation temperature. Intensified processes involving microstructured products (also known as intensive structuring) deliberately use low temperature/high concentration (which results in high viscosity—non-Newtonian flow behavior), small processing volume (provides uniform deformation fields), and high/ultrahigh deformation rates under laminar flow conditions so that high/ultrahigh stresses can be generated and sustained to achieve a small microstructure at a fast production rate. These requirements therefore qualify this processing strategy as intensive processing.

All flows can be decomposed into shear and extensional components. The effectiveness of the flow field is dependent on the deformation rate, the relative values of shear and extension, and the microstructure of the fluid.[16,17] Extensional flows are more effective in microstructure development, such as droplet breakup and mechanochemical reactions. However, such flows are difficult to generate and to maintain and therefore, in practical applications, capillary entrance/exit flows provide a suitable means of achieving extensional flows where the shear component of the flow field changes with the capillary entry–exit angle.[16,17] Indeed, OFRs also generate extensional flows, which result in efficient droplet breakup in emulsification. Mixers in which the deformation rate and the relative values of shear/extension rates can be controlled are known as controlled deformation mixers.[18–20] The multiple expansion–contraction static mixer (MECS mixer), which is just a series of short capillaries separated by flow dividers, is an example of a controlled deformation static mixer and provides all of the above requirements because ultrahigh shear/extension rates approaching $10^6\,\mathrm{sec}^{-1}$ can be generated and maximum shear/extension rate ratio can be controlled.[6,17,21] Although MECS mixers are very useful in the processing of liquid/liquid systems (emulsions), they are not suitable for solid/liquid systems, in which case the controlled deformation dynamic mixer (CDDM) can be used.[19,20,22–25] Such mixers are suitable for highly viscous fluids and can also have pumping action and heat transfer facilities.

Controlled deformation dynamic mixers can be regarded as stage-wise reactors in which each stage is connected to the neighboring stages in parallel and in series. The fluid motion is three-dimensional; when the fluid leaves a stage in series, it undergoes extensional deformation while parallel-stage transfer provides mixing. Such devices have two elements, a rotor and a stator, both of which have cavities machined on them. The rotor and stator cavities can match, thus providing a suitable geometry for extensional flow at the exit/entrance to a cavity. Shear deformation is dominant within the cavities, and shear and extensional flows are superimposed in the regions where there is minimum rotor/stator clearance.

There are two such devices available. They are either in the form of two concentric cylinders (inner cylinder is the rotor) or as two (or four) disks as shown in Fig. 2.[19,20,22–25] The cavity arrangement in a disk CDDM is shown in Fig. 3. The rotor (Fig. 3A) and stator (Fig. 3B) disks can be further modified to have microporous disks for simultaneous reaction and separation or an electric field can be applied across the disk clearance.[22] Furthermore, the disks can have

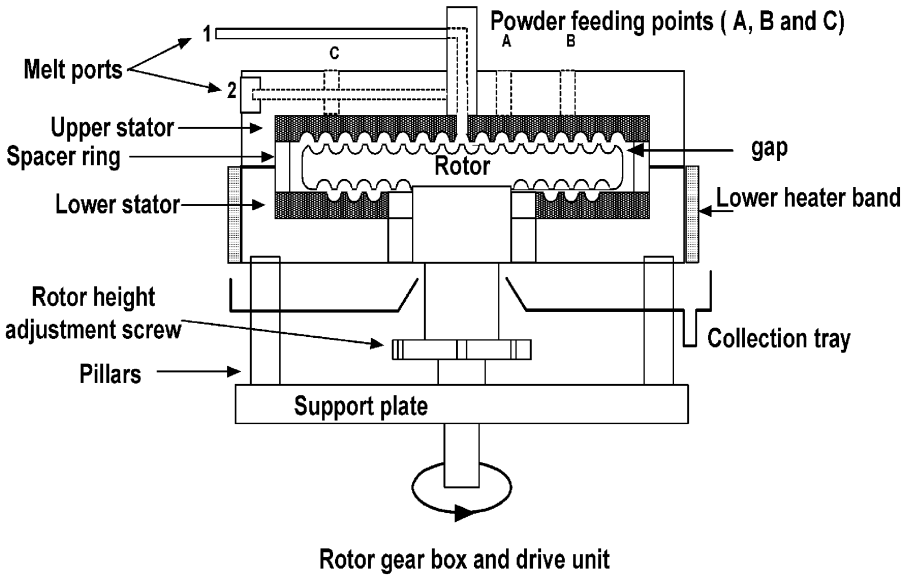


Fig. 2 Diagrammatic view of a CDDM in rotor–stator disk configuration. (From Ref.[22].)

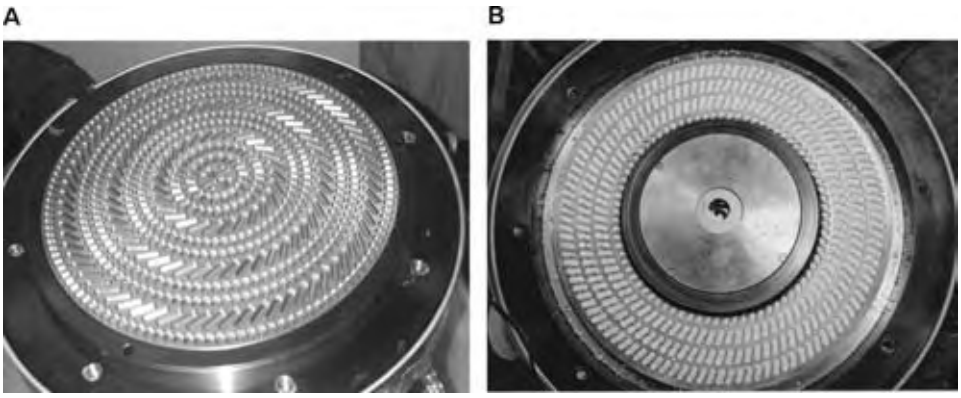


Fig. 3 Cavity arrangement in the upper rotor and lower stator disks of a CDDM in rotor–stator disk configuration. Elongated cavities are more efficient in pumping/conveying and smaller-aspect ratio cavities are for mixing. (A) Upper rotor. (B) Lower stator. (From Ref.[23].) *(View this art in color at www.dekker.com.)*

a temperature profile imposed on them. Controlled deformation (static and dynamic) mixers have been used in the intensive structuring of concentrated detergent formulations, agglomeration and microencapsulation of fine powders, and intensive separation of stable water-in-crude-oil emulsions.[19,20,22,26–31]

PHENOMENON-BASED PI

Miniaturization or reduced processing volume can result in physical PI by virtue of increasing surface area per unit volume. Miniaturized heat exchangers/reactors can be fabricated using microchannels, but such fabrication techniques are expensive and are limited by the availability of the fabrication technique for industrial scale applications. On the other hand, if the small processing volume results in enhanced selectivity, then the subsequent process has a greater degree of intensification. Therefore, in phenomenon-based PI, for a given reaction, both the reactor size/material and the nature/strength of the processing field must be considered carefully to achieve an even more impressive degree of intensification. Here, we review some of the phenomena that have been used in PI and already utilized in industry.

Flow Induced Phase Inversion Phenomenon and PI in Particle Technology

Flow induced phase inversion (FIPI) has been observed by the author and applied to intensive materials structuring such as agglomeration, microencapsulation, detergent processing, emulsification, and latex production from polymer melt emulsification.[16–18,21,29,32–35] A diagrammatic illustration of FIPI is shown in Fig. 4. When material A is mixed with material B, in the absence of any significant deformation, the type of dispersion obtained [(A-in-B) or (B-in-A)] is dictated by the thermodynamic state variables (TSVs) (concentration, viscosity of components, surface activity, temperature, and pressure). If the prevailing TSVs favor the formation of (A-in-B) dispersion, phase inversion to (B-in-A) dispersion can be achieved by changing the TSVs (thermodynamically driven process). Alternatively, the dispersion can be subjected to a well-prescribed deformation, characterized by its rate and type of deformation state variables (DSVs) to invert the dispersion under constant thermodynamic conditions; this phenomenon is known as FIPI. It is found that FIPI is not catastrophic and the dispersion goes through an unstable, cocontinuous state denoted as [AB], followed by a relatively stable

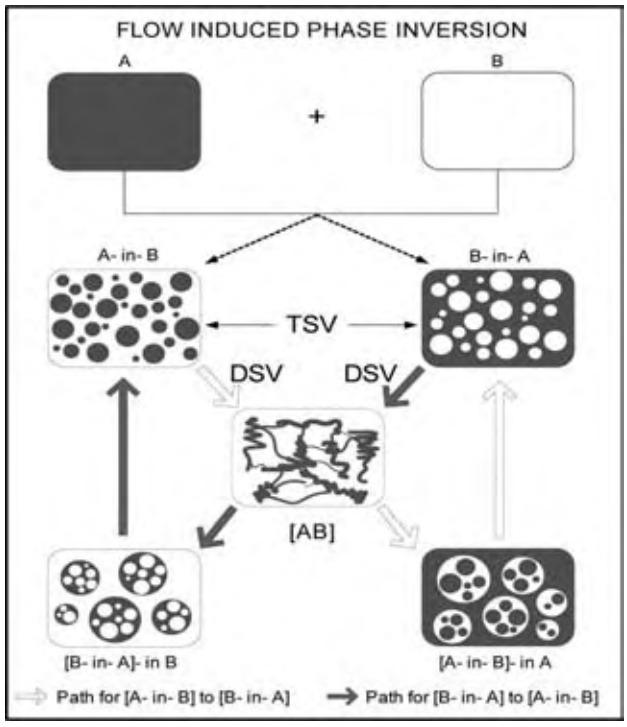


Fig. 4 Isothermal FIPI paths for the inversion of (A-in-B) or (B-in-A) emulsions through a cocontinuous unstable emulsion phase (AB). TSV, thermodynamic state variable; DSV, deformation state variable. *(View this art in color at www.dekker.com.)*

multidispersion state, denoted as [(A-in-B)-in-A], before complete phase inversion to (B-in-A). Therefore, the interchangeability of TSVs with DSVs forms the basis of FIPI processes.

The importance of FIPI is twofold. It can be used to promote phase inversion without changing the thermodynamics of the system to obtain a higher entropy state, or it is possible to delay phase inversion while reducing the system entropy.[36] The characteristics of the microstructure formed (such as emulsion droplet size) are dependent on the type of microstructure and deformation (shear, extension, or combined), as well as the deformation rate. To maximize the fluid microstructure/flow field interactions, the flow field must be uniform, which requires the application of the flow field over a small processing volume, which can be achieved by using MECS mixers or CDDMs.

Isothermal FIPI Emulsification

Isothermal emulsification of viscous oils to obtain concentrated oil-in-water emulsions is carried out by inverting a water-in-oil emulsion under flow conditions using an MECS mixer.[16,17,21] Initial water-in-oil emulsion is prepared using a batch mixer (typical shear rate of $100\,\text{sec}^{-1}$), which is subsequently inverted to oil-in-water emulsion at extension rates approaching $10^6\,\text{sec}^{-1}$. It is found that the FIPI starts at a critical extension rate and the extension rate has to be increased to achieve full inversion resulting in submicrometer size emulsion droplets with a very narrow size distribution (size span < 1).[17]

In a novel process, FIPI was also applied to the emulsification of polymer melts in water, thus providing an alternative method to emulsion polymerization for the production of latexes.[18,21,32–34] In fact, some thermoplastic melts (such as polyethylene) cannot be obtained through the emulsion polymerization route; hence, the present technique is an example of PI providing a novel product form. To achieve the emulsification of thermoplastics, it is necessary to operate near or above 100°C and at elevated pressures, which necessitates the use of polymer processing equipment fitted with a MECS mixer at the outlet. It was found that molecular surfactants could not be used to obtain the initial (water-in-polymer melt) emulsion. Instead, hydrophobically modified water-soluble polymers were used as the surface active material. After the phase inversion in the MECS mixer, the resulting emulsion was diluted to the level required. This also freezes the molten latexes. The important attributes of FIPI emulsification include a low level of surfactant use, low temperature processing, production of submicrometer particles with a narrow size distribution, and production of novel products.

Nonisothermal FIPI and Intensive Granulation Technology

Although the underlying manifestation of FIPI is the same, detailed molecular or microscopic mechanisms are different from one multiphase system to another. When FIPI is applied to granulation (i.e., agglomeration and microencapsulation of fine particles in the absence or presence of liquids) using polymers as binders, the dominant mechanism for phase inversion is based on the behavior of macromolecules in flow fields, which includes stress-induced crystallization, diffusion, macromolecular chain scission, macromolecular confirmation, deposition, and reaction. Therefore, high/ultrahigh stresses must be generated and sustained to achieve these phenomena. A combination of these phenomena results in FIPI in polymer melt/solid particle dispersions, which is further enhanced by operations at low temperatures as a result of high viscosity (i.e., high stress generation) and differential crystallization, because of the distribution of the molecular weight of polymeric binders.

Intensive granulation technology

Agglomeration and encapsulation of fine particles are encountered in several industrial sectors such as pharmaceuticals, detergents, fertilizers, and animal feedstocks. The volume throughput in these processes can vary enormously. Batch processing is suitable for pharmaceutical applications where the volume of production is quite low and product values are very high, while a continuous process is essential in other sectors where the volume throughput is very high.

Currently available agglomeration techniques are either based on the provision of a suitable processing environment for primary particle collisions (growth agglomeration) or on the compression of particles to form densified particles (pressure agglomeration) in the presence or absence of a binder. These techniques are well known and periodic reviews are available. The growth agglomeration technique is often conducted in batch mixers or fluidized beds in which particle breakup and agglomeration occur simultaneously. The residence times are relatively long and the variety of possible primary particles for agglomeration and binders is restricted. In pressure agglomeration, pressurization is confined to a small volume and there are restrictions on the raw materials.

The so-called multinucleus-type capsules/microcapsules are essentially agglomerated particles with prescribed released characteristics, which can be controlled by the concentration and solubility of the binder. In agglomeration, the amount of binder is kept to a minimum, as long as the agglomerates are

sufficiently strong. Therefore, in principle, multinucleus microcapsules can be manufactured through an agglomeration process, where the binder concentration should be controlled. However, currently available agglomeration processes do not have this flexibility and therefore such capsules/microcapsules are produced through specific techniques.

A novel agglomeration/encapsulation (granulation) technique based on isothermal FIPI has been disclosed recently and applied to several industrial processes. In an isothermal FIPI granulation, a concentrated suspension of powder in liquid is caused to crumble (phase invert) through the further addition of a powder (a ''crumbling agent''), i.e., a phase inversion from a solid phase suspended in liquid (a paste) to a liquid phase suspended in solid (an agglomerated granular material). The process requires a degree of mechanical energy input to disperse the two phases and, therefore, it must be carried out in some kind of mixing environment. Also, the critical deformation rate at which FIPI starts is dependent on the thermodynamic state of the structured fluid as well as the type of flow field, i.e., shear, extensional, or combined flows. This mixing flexibility allows FIPI in either a batch or a continuous mode of operation. The FIPI granulation technique combines both growth and pressure agglomeration processes. For a typical isothermal FIPI agglomeration/microencapsulation process, phase inversion takes place at a critical dispersed phase concentration (crumbling concentration), when the primary particle size and mixing conditions are kept constant.

The use of temperature and deformation as processing parameters can therefore promote phase inversion in solid particle/polymer systems and can be utilized in PI in particle technology. The reduction in temperature in polymer melt/particle mixture (paste) results in a reduction of the continuous phase volume because of the crystallization and stress induced deposition of the higher molecular weight fractions of the polymer. Furthermore, at lower temperatures, because of the reduced polymer chain mobility and interactions with the large numbers of small primary particles, which make a three-dimensional cage, the effective mobile-phase volume decreases further with increasing primary particle concentration. Consequently, the continuous phase can no longer sustain the increased phase volume of the dispersed phase and phase inversion takes place in the form of the melt fracture. This phenomenon now forms the basis of the present granulation technique.

Intensive granulator

The complete granulation process is carried out in a system that consists of an extruder and a CDDM in the form of rotor–stator disks with heat transfer facilities. A diagrammatic illustration of the granulator is shown in Fig. 2. It consists of two back-to-back rotors (referred to as upper and lower rotors) made from a single circular block and two stators (also referred to as upper and lower stators) facing the rotors. As shown in Fig. 3, the rotors and stators contain cavities on their surfaces, which are designed to pump/convey and/or achieve mixing. The cavities on the rotor and stator are offset by half a cavity length so that the rotor and stator cavities never exactly match with each other. The separation (gap) between the upper or lower rotor and the corresponding stators can be varied independently.

The filled polymer melt from the extruder output flows into the gap between the upper stator and rotor. The rotor block (incorporating the upper and lower rotors) spins sandwiched between the upper stator and lower stator of the granulator. The upper rotor temperature is equal to the temperature when solidification of the binder starts, while the upper stator temperature is maintained at the temperature when the solidification is complete. Granulation takes place inside the gap between the upper stator and rotor. Once the granules are formed, they travel through the gap between the lower stator and the rotor and finally emerge and drop onto the collection tray. The upper stator also contains three auger powder feeders located symmetrically at various distances from the center. Additional powder can be added into the filled polymer melt after phase inversion. The addition of the extra powder is useful to increase the particle content of the agglomerates and to form a shell-and-core-type microcapsule.

Samples from various regions were recovered and analyzed. It was found that granulation takes place over four steps. These steps are associated with the following zones: 1) melt flow zone; 2) granule nucleation zone; 3) crumbling (melt fracture and fragmentation) zone; and 4) granule transport zone. The granulation occurs over a short distance, within a distance of 2–5 mm. The location of the crumbling (phase inversion) is important as it dictates the size of the granulator. This location is dependent on the effectiveness of heat transfer, the temperature of the polymer melt from the extruder, the polymer concentration, and the clearance between the rotor and the stator.

Fig. 5 illustrates the granule particle size distribution as a function of gap size between rotor and stator.[25] As can be seen, the average particle size $D(50)$ is approximately equal to the gap size, and the particle size span (ca. 0.5) is very low and independent of gap size. Typically, the concentration of the binder is 25 wt%. In all of the known agglomeration and microencapsulation techniques, the concentration of

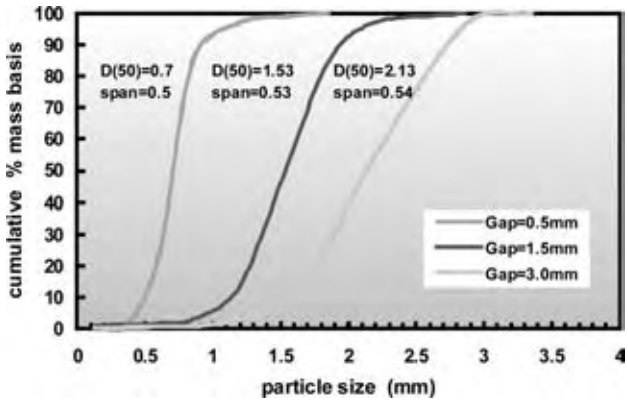


Fig. 5 Granule size distribution as a function of gap (clearance) between the rotor and stator disks. Average granule size $D(50)$ in mm and particle size span are also indicated. (From Ref.[24].) *(View this art in color at www.dekker.com.)*

the active material (filler) is dependent on the granule size, larger granules having a higher binder concentration. Because of the fact that the granulation is induced from a molten state where the filler concentration is uniform, it is therefore not surprising that the granules have a constant active (filler) concentration.

CHEMICAL PI USING NANOSTRUCTURED MICROPOROUS MATERIALS AND CATALYSTS

B

Preparation of Novel Porous Materials for PI

Functionalized microporous polymers (with a superimposed nanostructure in the form of nanopores within the walls of the micropores) have been prepared through a high internal phase emulsion (HIPE) polymerization route.[35,37–40] These polymers (known as PolyHIPE Polymers, PHP) were subsequently metallized by solution deposition followed by heat treatment.[39,41]

Microporous polymers have been used in the intensification of several processes encountered in nuclear, petroleum, bioconversion, tissue engineering, chemical, and environmental technologies.[22,31,39–45] The advantages of PHP and its metallic form are associated with the accessibility of their pores, controllability of the pore and interconnect structures, versatility of fabrication, and chemical modification of their walls. The structure of these materials is shown in Figs. 6A–C. These polymers can be manufactured over a wide range of pore size (D) ($0.5\,\mu m < D < 5000\,\mu m$) and interconnect size (D) ($0 < d/D < 0.5$).

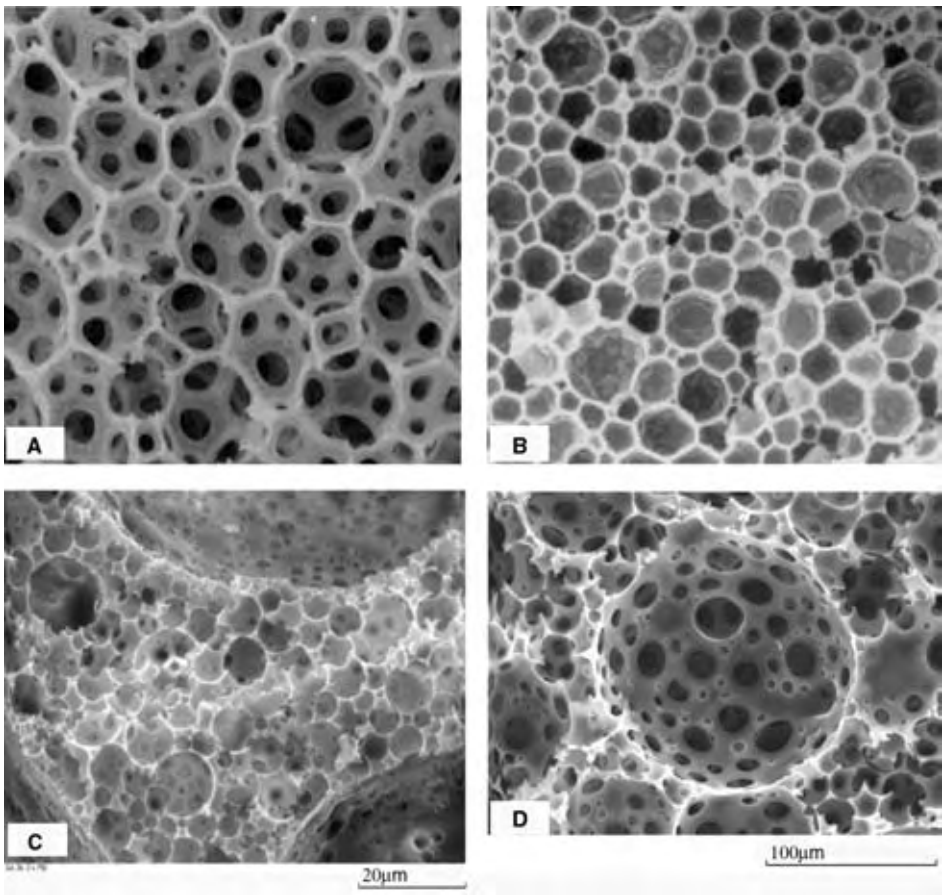


Fig. 6 Basic PHP structures: (A) primary pores with large interconnecting holes; (B) primary pores with nanosized interconnecting holes; (C) large coalescence pores (three such pores are partially shown) dispersed into the primary pores in the process of coalescence; and (D) detail of the coalescence pores. Note that these pore structures can be prepared over a wide size range.

Pores with size above ~200 μm are obtained through a coalescence polymerization route.[35,39]

Nanostructured microporous catalysts or catalyst supports offer intensified catalysis as they provide enhanced surface area accessible to the reactants and products. In nonstructured catalysts, although the surface area may be large, they are often inaccessible as a result of surface fouling and diffusion resistance can slow down the rate of reaction. In a recent development, microporous materials were used as templates for the solution deposition of metals, which were subsequently heat treated to obtain porous metallic structures, where the size of the pores ranged from 10 μm to 10 nm.[39,41] The relative phase volume of these two regions can be controlled and the overall porosity can be in excess of 50%. Fig. 7 illustrates the size scale of structures ranging from 10 μm to 10 nm.

These attributes are used in devising intensified processes as well as in the discovery of several size dependent phenomena, especially in biology, which are subsequently utilized in BI. Currently, PHPs (both hydrophobic and hydrophilic versions) are used in the intensification of stable water-in-oil emulsion separation (demulsification), gas–liquid separation, as applied to tar and water removal from biogas produced through the gasification of biomass, BI, tissue engineering, and metal ion/toxin removal from contaminated water.[8,22,31,35,39,40,42–45] The accessibility of the pores increases both the rate and the capacity of metal ion removal. In currently available ion-exchange resins, only the outermost part of the resin beads is utilized, because as soon as the resin surface is saturated or fouled, metal ion removal rate and capacity are drastically reduced.

Intensification of Stable Water-in-Oil Emulsion Separation

The breakdown of the stable emulsions and subsequent separation to oil and water (demulsification) are important in nuclear, petroleum, and environmental technologies. The emulsion stability is primarily induced by the use of surfactants and is enhanced by reduced size and narrow size distribution of the emulsion droplets. Disruption to low interfacial activity (hence instability) can be achieved by using demulsification agents, which are, however, costly and environmentally undesirable, as they are irrecoverable. Demulsification can also be achieved by electric and/or centrifugal fields, or by chemical treatment of the emulsion.

During the reprocessing of depleted uranium, a highly stable and viscous water-in-oil emulsion is

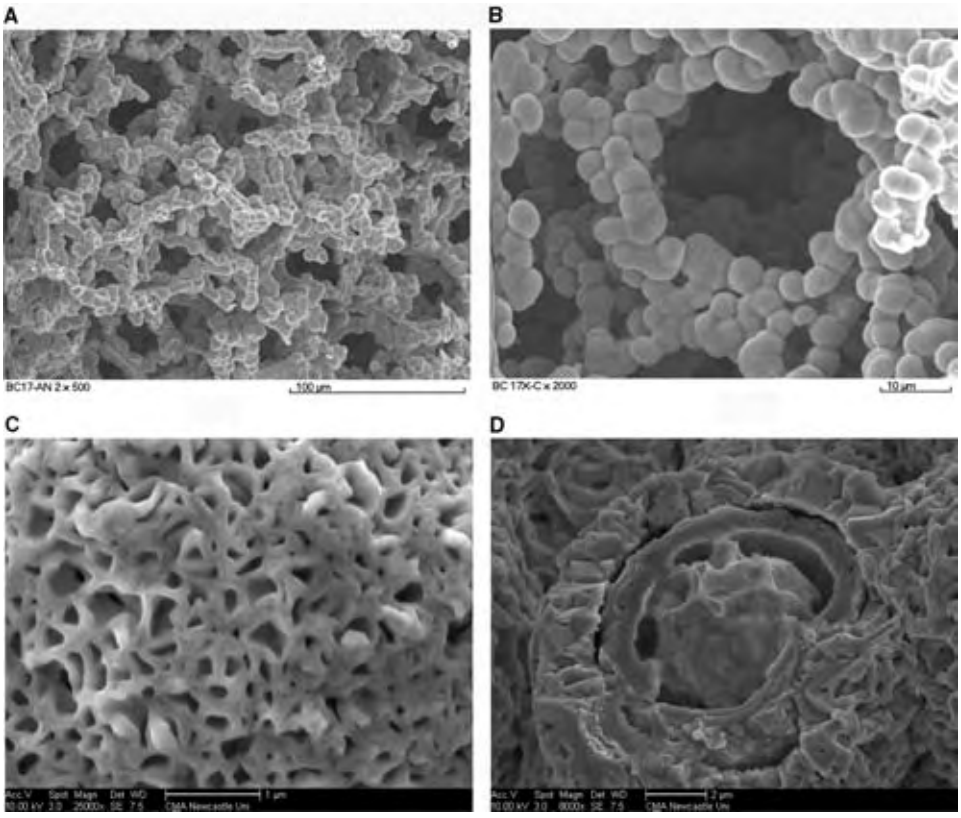


Fig. 7 Nanostructured microporous nickel with accessible pores showing the hierarchy of the pore sizes: (A) micropores created by fused metal grains; (B) structure of the metal grains before heat treatment; (C) surface pore structure of the grains; and (D) inner structure of metal grains showing the existence of a porous inner core.

formed called interfacial crud (or crud). This emulsion deposits in the process lines, thus reducing the heat/mass transfer rates, making it necessary to periodically remove crud by solvent washing.

The removal of water from crude oil and the subsequent treatment of the water produced is a very important process as the crude oil may contain up to 90% water, although it is still economically viable to separate crude from water. In offshore applications, if the separation is achieved at the seabed, massive savings in pumping and heating costs can be achieved. Such a separation process must be intensive, because of the volume of crude oil produced and the hostile environment, or limited space such as the seabed or offshore oil platform. However, such a separation technology at the source would be sustainable utilizing the potential energy of the crude oil from the well, i.e., high temperature and pressure, both of which increase the demulsification rate. Furthermore, the produced water must also be treated, if it is to be discharged into the sea.

Recently, we have shown that sulfonated PHP can act as a demulsifier for highly stable emulsions like crud and water-in-crude oil emulsions.[31,44] The mechanism of demulsification is that sulfonated PHP removes selectively surface active species in the emulsion, causing destabilization. At the same time, it also adsorbs metal ions, thus achieving two functions at the same time. As a result, these materials are called demulsifier adsorbers. In highly stable emulsions where neither electric field nor demulsifier adsorbers are effective, the combination of these two methods appears to create synergy for separation.

The intensification of crude oil-in-water emulsions was tested using BP Amoco crude (Harding Field, North Sea) with specific gravity of $0.8\,g/cm^3$ and viscosity of 150 mPa sec at 25°C. The aqueous phase is a model seawater containing 28.1 g/L NaCl, 0.6 g/L $CaCl_2$, 5 g/L $MgCl_2$ in double-distilled water. The emulsion viscosity at 25°C and shear rate of $1000\,sec^{-1}$ is 1030 mPa sec. These emulsions show no sign of separation after 4 weeks of standing. The tests were carried out using a flow-through electric field separator described in Refs.[44,46]. This cell contains a circular channel of 1 cm diameter with two electrodes separated by 10 cm in which the anode is electrically insulated and the counter electrode is earthed. Freshly prepared emulsion containing 0.5 g demulsifier per kilogram of 50/50 emulsion was fed from the center of the circular channel and oil-rich and water-rich phases were collected from the top (anode) and the bottom (cathode) and allowed to separate. The overall degree of separation was measured within 10 min of collection of the sample.

The effect of PHP demulsifier adsorber is illustrated in Figs. 8 and 9. Fig. 8 shows the variation of percentage separation as a function of flow rate under a constant electric field, while Fig. 9 shows the variation of separation as a function of the electric field at a constant flow rate. Both figures clearly show the synergy present between the two types of intensification in oil–water emulsion separation. A continuous intensified separator based on the rotating disk in rotor–stator arrangement with electric field has been disclosed recently.[22]

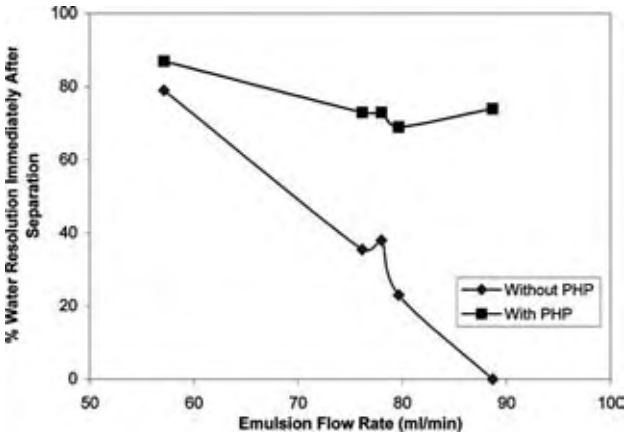


Fig. 8 Process intensification in water-in-crude oil separation under electric field with or without PHP: variation of percentage separation immediately after passing through the electric field in the absence or presence of the sulfonated PHP. Electric field strength of 2.5 kV applied over a distance of 10 cm. (From Ref.[43].) *(View this art in color at www.dekker.com.)*

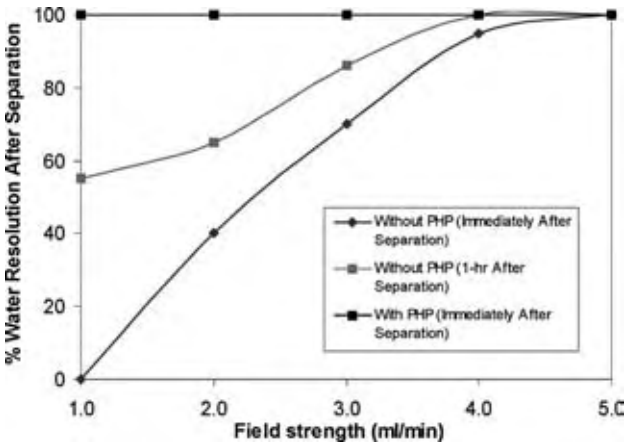


Fig. 9 Process intensification in water-in-crude oil separation under electric field with or without PHP: variation of percentage of separation with electric field strength applied over a distance of 10 cm when the emulsion flow rate is kept constant at 60 ml/min. Percentage separation into oil–water layers is carried out either immediately (within 10 min) or after 1 hr of emulsion passing through the electric field. (From Ref.[43].) *(View this art in color at www.dekker.com.)*

CARRIER-MEDIATED INTENSIFICATION OF SEPARATION PROCESSES

Toxins in the form of heavy metal ions or organic molecules in water (and indeed from gases) can be removed by using carriers with suitable surface chemistry, which can bind the toxic chemicals in their structure and are subsequently removed by filtration. These carriers include small microscopic particles or surfactants.[47–50] Because of their unique ability to form well-ordered molecular aggregates and structures, ranging from a few nanometers to several hundred micrometers, surfactants can be used in the intensification of separation processes, such as the removal of organic and heavy metal ion contaminants from water. Surfactants offer a very large surface area for mass transfer and surfactant assemblies can be modified to be selective.[49] Surfactant-mediated separations can be further enhanced by the orthogonal superimposition of cross-flow separation field and an electric field.[51,52] In such separation processes, both the permeate flux and the surfactant rejection can be enhanced by over 100-fold each.[51,52] However, this enhancement is by no means common, as it depends on the surfactant concentration, as well as surfactant head-group charge density and surfactant phase behavior. Furthermore, surfactants can degrade electrochemically, therefore the selection of surfactant becomes important in electric field enhanced mediation processes.[52]

During the cross-flow microfiltration of dilute surfactant solutions, surfactants form highly stable gel phases within micropores. These gel phases do not dissolve even though they are in contact with surfactant-free water flowing at shear rates in excess of $5 \times 10^4\,\text{sec}^{-1}$. Such stable-strong surfactant phases are expected to form at concentrations over 30 wt%. This phenomenon, first observed by the author, is utilized in cross-flow filtration and forms the bases of the demulsification of highly stable water-in-oil mulsions as well as the removal of tar/water from gases.[22] These observations indicate that microporous materials (with a varying degree of hydrophilicity) preferentially adsorb surface-active species, thus forming the basis of oil–water separation using PHPs.

In the intensification of chromate removal from water, a double-chain cationic surfactant, dioctadecyldimethylammonium chloride (DODDMAC), was used as a carrier and a cross-flow electrofiltration was used, in which both the transient and the steady-state fluxes and the rejection of metal ions and surfactant were measured.[51,52] Dioctadecyldimethylammonium chloride in water forms multilamellar droplets, even at very low concentrations. This structure is shown in Fig. 10. Metal ions are entrapped within the water layers and organic toxins can be immobilized within the surfactant bilayers. Under an electric field, DDOMAC multilamellar droplets with entrapped anions migrate to the negative electrode, forming a stable gel.

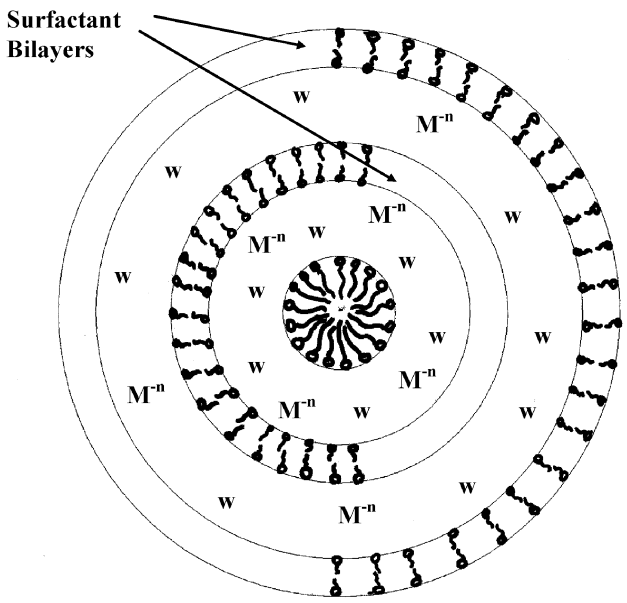


Fig. 10 Schematic representation of surfactant in multilamellar droplet phase with entrapped metal ions (M^{-n}) in the aqueous phase (W) layers, which are separated by surfactant bilayers. The number of layers (hence the size of the surfactant droplet) is dependent on temperature and concentration. When the surfactant head group is positively charged thus encapsulating oppositely charged metal ions, under an electric field, surfactant lamellar droplets migrate to the anode and form a highly stable viscous gel in which the positively charged metal ions are concentrated at the anode only separated by the surfactant bilayer. (From Ref.[52].)

The electric field intensification in the surfactant-mediated separation processes is shown in Table 2.[52] In the feed, surfactant and metal ion concentrations are denoted by C_{SF} and C_{MF}, respectively, while the corresponding concentrations in the permeate under steady-state conditions are denoted by C^*_{SP} and C^*_{MP}. Surfactant and metal ion rejections at a steady state are defined as $R^*_S = (1-C^*_{SP}/C_{SF})$ and $R^*_M = (1-C^*_{MP}/C_{MF})$. The molar ratio of metal ion to surfactant is denoted by Y_{MS}. The separation of the electrodes is 3 mm. In Table 2, the initial current, pH, and solution conductivity are also given. It shows that both metal and surfactant are separated effectively under an electric field in which the permeate flux, surfactant, and metal ion rejections are enhanced. Economic analysis of the process indicates that some 20- to 50-fold efficiency increase is achieved compared with the no electric field case. For the process to be economical, low-solubility surfactants that can form multilamellar droplets should be used as carriers.

Table 2 Process intensification in the removal of heavy metal ions from water using surfactants as carriers under electric field[a]

E (V)	C_{SF} (mM)	C_{MF} (mM)	Y_{MS}	J^* (L/hr/m^2)	R^*_M (%)	R^*_S (%)	I_0 (A)	pH_0	K_0 (μsec/cm)
0	3	0.2	0.067	254	99.4	—	—	6.70	65
0	3	0.5	0.17	190	65.3	88.7	—	6.59	70
10	3	0.2	0.067	1500	—	—	0.1	6.75	61
30	3	0.2	0.067	2390	99.5	—	0.1	6.68	68
30	1	0.5	0.5	1060	60.0	76.5	0.3	7.55	120
30	3	0.5	0.17	2760	96.1	97.0	0.4	7.24	119
30	5	0.5	0.1	2490	97.9	—	0.4	6.25	—
30	10	0.5	0.05	2450	98.0	97.4	0.5	6.45	138
50	3	0.5	0.17	4450	99.1	98.5	0.7	7.24	93

[a] J^*, steady state flux; R^*_M, metal rejection; R^*_S, surfactant rejection; pH_0, initial feed pH; K_0, conductivity; I_0, initial cell currents; E, electric field; C_{SF}, feed surfactant concentration; C_{MF}, feed metal concentration; Y_{MS}, metal to surfactant molar ratio.
(From Ref.[52].)

PHENOMENON-BASED BIOPROCESS INTENSIFICATION

The most suitable driving force in BI is the reduction of the diffusion path that already operates in transport processes across biological bilayers. Consequently, biocatalyst membranes and specially designed bioreactors, such as jet loop and membrane reactors, are available to intensify biochemical reactions.[4–8,53] Supported biocatalysts are often employed to enhance catalytic activity and stability and to protect enzymes/microorganisms from mechanical degradation and deactivation.[5,6,8,54] Immobilization of the cells is one of the techniques employed to improve the productivity of bioreactors.

Immobilized cells are defined as cells that are entrapped within or associated with an insoluble matrix. Various methods that are used for immobilization include covalent coupling, adsorption, entrapment in a three-dimensional scaffold, confinement in a liquid–liquid emulsion, and entrapment within a semipermeable membrane. Bioreactors with immobilized cells have several advantages over those operating with free cells or immobilized enzymes. Immobilized cell systems permit the operation of bioreactors at flow rates that are independent of the growth rate of the micro-organisms employed. Catalytic stability can be greater for immobilized cells than for free cells. Some immobilized micro-organisms tolerate higher concentrations of toxic compounds compared with free cells, when the cell support media act as a temporary sink for the excess toxin. However, in the current biocatalyst support technology, the presence of the support itself introduces mass transfer restrictions for the substrate/product/nutrient diffusion to and from the biocatalyst. These disadvantages are also valid when supports are used to grow animal cells in vitro.[35,40] In this case, well-designed cell support systems are necessary for cell proliferation and viability. The efficiency of cell function (for example, production of collagen II during the in vitro growth of chondrocytes) is dependent on the microarchitecture, i.e., pore and interconnect size, and surface chemistry of the support.[35,40,54,55]

Microporous polymers with a well-prescribed internal microstructure were prepared in monolithic form to form a flow-through microbioreactor in which phenol degrading bacteria, *Pseudomonas syringae*, were immobilized. Initially, bacteria were force-seeded within the pores and subsequently allowed to proliferate, followed by acclimatization and phenol degradation at various initial substrate concentrations and flow rates. Two types of microporous polymer were used as the monolithic support. These polymers differ with respect to their pore and interconnect sizes, macroscopic surface area for bacterial support, and phase volume. Poly-HIPE polymer with a nominal pore size of 100 μm with phase volume of 90% (with a highly open pore structure) yielded reduced bacterial proliferation while the polymer with nominal pore size of 25 μm with phase volume of 85% (with small interconnect size and large pore area for bacterial adhesion) yielded monolayer bacterial proliferation. Bacteria within the 25 μm polymer support remained monolayered without any apparent production of extracellular matrix during the 30-day continuous experimental period as shown in Fig. 11(A). The microbioreactor performance was characterized in terms of volumetric utilization rate and compared with the published data, including the case where the same bacteria were immobilized on the surface of microporous polymer beads and used in a packed bed during the continuous degradation of phenol.[42] It has been shown that at a similar initial substrate concentration, the volumetric utilization in the microreactor is at least 20-fold more efficient than the packed bed, depending on the

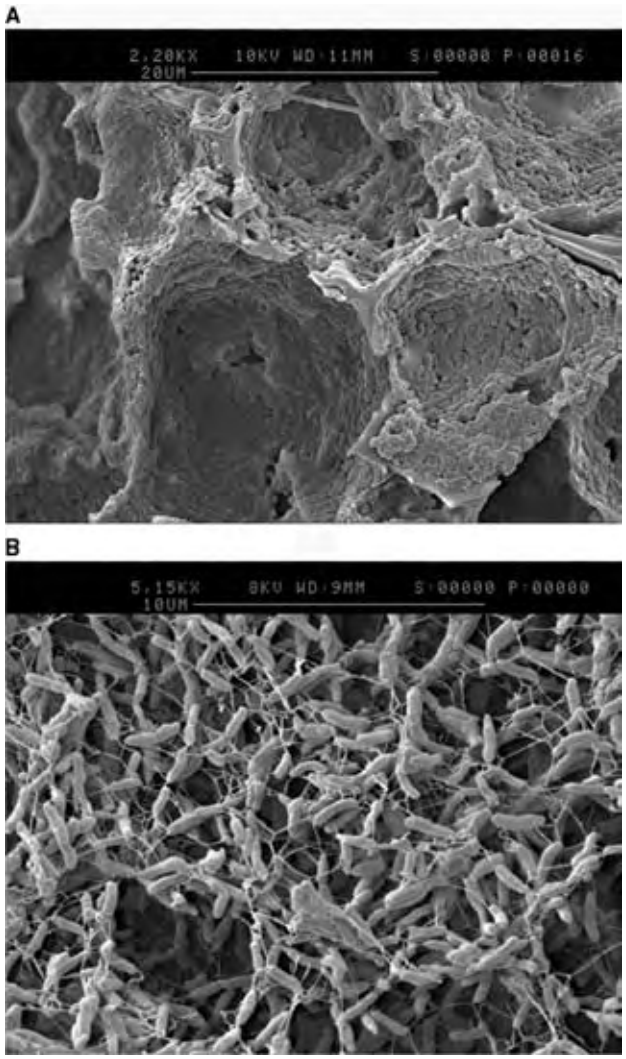


Fig. 11 Pore size dependence of bacterial behavior in microbioreactors. (A) Monolayer bacteria coverage in a small pore size (25 μm) microbioreactor. (B) Biofilm formation by the same bacteria in a large pore size (100 μm) microbioreactor. (From Ref.[8].)

flow rate of the substrate solution. The concentration of the bacteria within the pores of the microreactor decreases from 2.25 cells/μm^2 on the top surface to about 0.4 cells/μm^2 within 3 mm reactor depth. If the bacteria-depleted part of the microreactor is disregarded, the volumetric utilization increases by a factor of 30-fold compared with the packed bed. This efficiency increase is attributed to the reduction of the diffusion path for the substrate and nutrients and enhanced availability of the bacteria for bioconversion in the absence of biofilm formation, as well as the presence of flow over the surface of the monolayer bacteria. When 100 μm pore size PHP is used as a support, or when the flow through the support is absent (i.e., beads used in packed beds), biofilm formation occurs, as shown in Fig. 11(B).[42]

CONCLUSIONS

The combination of process intensification and process miniaturization is an important element of sustainable technology and a platform for advancement in multidisciplinary science and engineering. There are now several examples of intensified processes being utilized in industry, as apparent from patent activity. Some of these intensified processes are confined to a single unit operation, while others integrate several intensified unit operations in the more efficient production of existing/improved/novel products.

Physical PI is limited in the scope and degree of intensification that it can achieve, while phenomenon-based intensifications can result in more impressive levels of intensification. They also provide new opportunities in understanding the behavior of structured matter, macromolecules, and microorganisms within a confined environment with superimposed transport processes. The understanding of the transport processes in human body and organ structures is very useful in PIM, while the application of the techniques and principles of PIM to biology and medicine can result in advances in tissue engineering and biology.

The ultimate aim of PIM is to establish intensified plants by integrating intensified unit operations that are already multifunctional. An example of such a plant is a ''biomass power plant'' in which an intensified gasifier is integrated with an intensified catalyst system and intensified environmental remediation (process water and gas cleaning) to produce electricity through several routes, including internal combustion engine and high-temperature fuel cells.[45,56]

ACKNOWLEDGMENTS

I wish to thank Drs K. Boodhoo, M. Dogru, and A. Harvey for helpful discussions. Bioprocess intensification and PI developed by the author have been supported by the United Kingdom Engineering and Physical Sciences Research Councils (EPSRC) and the Department of Trade and Industry (United Kingdom), as well as by the industry, notably, AstraZeneca, BLC Research, British Nuclear Fuels Ltd (BNFL), BP Amoco, Cytec, Exxon Mobil, ICI, Intensified Technologies Incorporated (ITI), Morecroft Engineers, Norsk-Hydro, Safety-Kleen Europe, Triton Chemicals, Unilever, and Willacy Oil Services. I am grateful for their support.

REFERENCES

1. Akay, G.; Mackley, M.R.; Ramshaw, C. Process intensification: opportunities for process and

product innovation. In *IChemE Research Event*; Chameleon Press: London, 1997; Vol. 1, 695–703.
2. Stankiewicz, A.; Moulijn, J.A. Process intensification. Ind. Eng. Chem. Res. **2002**, *41*, 1920–1924.
3. Akay, G. Upping the ante in the process stakes. Chem. Eng. **2004**, *752*, 37–39.
4. Chisti, Y.; Moo-Young, M. Bioprocess intensification through bioreactor engineering. Trans. IChemE. **1996**, *74*, 575–583.
5. Giorno, L.; Drioli, E. Biocatalyst membrane reactors: applications and perspectives. TIBTECH **2000**, *18*, 339–349.
6. Akay, G.; Erhan, E.; Keskinler, B.; Algur, O.F. Removal of phenol from wastewater by using membrane-immobilized enzymes: part 2. Cross-flow filtration. J. Membr. Sci. **2002**, *206*, 61–68.
7. Pekdemir, T.; Keskinler, B.; Yildiz, E.; Akay, G. Process intensification in wastewater treatment: ferrous iron removal by a sustainable membrane bioreactor system. J. Chem. Technol. Biotechnol. **2003**, *78*, 773–780.
8. Akay, G.; Erhan, E.; Keskinler, B. Bioprocess intensification in flow through micro-reactors with immobilized bacteria. Bioeng. Biotech. **2005**, *90*, 180–190.
9. Boodhoo, K.V.K.; Jachuck, R.J.J. Process intensification: spinning disc reactor for styrene polymerisation. Appl. Therm. Eng. **2000**, *20*, 1127–1146.
10. Boodhoo, K.V.K.; Dunk, W.A.E.; Jachuck, R.J.J. Continuous photopolymerisation in a novel thin film spinning disc reactor. In *Photoinitiated Polymerization*; Crivello, J.V., Belfield, K.D., Eds.; ACS Symposium Series No. 847, Washington, DC, 2003; 437–450.
11. Boodhoo, K.V.K.; Jachuck, R.J.J. Process intensification: spinning disc reactor for condensation polymerisation. Green Chem. **2000**, *2*, 235–244.
12. Cafiero, L.M.; Baffi, G.; Chianese, A.; Jachuck, R.J. Process intensification: precipitation of barium sulphate using a spinning disc reactor. Ind. Eng. Chem. Res. **2002**, *41*, 5240–5246.
13. Erkoc, E.; Yapici, S.; Keskinler, B.; Cakici, A.; Akay, G. Effect of pulsed flow on the performance of carbon felt electrode. Chem. Eng. J. **2002**, *85*, 153–160.
14. Mackley, M.R.; Ni, X. Mixing and dispersion in a baffled tube for steady laminar and pulsatile flow. Chem. Eng. Sci. **1991**, *46*, 3139–3151.
15. Baird, M.H.I.; Harvey, A.P.; Mackley, M.R.; Ni, X.; Rama Rao, N.V.; Stonestreet, P. Mixing through oscillations and pulsations—a guide to achieving process enhancements in the chemicals and process industries. Chem. Eng. Res. Des. **2003**, *81*, 373–383.
16. Akay, G. Flow induced phase inversion in powder structuring by polymers. In *Polymer Powder Technology*; Narkis, M., Rosenzweig, N., Eds.; Wiley: New York, 1995; 542–587.
17. Akay, G. Flow induced phase inversion in the intensive processing of concentrated emulsions. Chem. Eng. Sci. **1998**, *53*, 203–223.
18. Akay, G Stable Oil in Water Emulsions and a Process for Preparing Same. European Patent 649,867, Oct 17, 2001.
19. Akay, G.; Irving, G.N.; Kowalski, A.J.; Machin, D Process for the Production of Liquid Compositions. European Patent 799,303, Oct 31, 2001.
20. Akay, G.; Irving, G.N.; Kowalski, A.J.; Machin, D Dynamic Mixing Apparatus for the Production of Liquid Compositions. US Patent 6,345, 907, Feb 12, 2002.
21. Tong, L.; Akay, G. Process intensification in particle technology: flow induced phase inversion in the intensive emulsification of polymer melts in water. J. Mater. Sci. **2002**, *37*, 4985–4992.
22. Akay, G Method and Apparatus for Processing Flowable Materials and Microporous Polymers. International Patent Application; PCT Publication WO 2004/004880, Jan 15, 2004.
23. Akay, G.; Tong, L.; Addleman, R. Process intensification in particle technology: intensive granulation of powders by thermo-mechanically induced melt fracture. Ind. Eng. Chem. Res. **2002**, *41*, 5436–5446.
24. Akay, G.; Tong, L. Process intensification in polymer particle technology: granulation mechanism and granule characteristics. J. Mater. Sci. **2003**, *38*, 3169–3181.
25. Akay, G.; Tong, L. Process intensification in particle technology: intensive agglomeration and microencapsulation of powders by non-isothermal flow induced phase inversion process. Int. J. Transport Phenom. **2003**, *5*, 227–245.
26. Akay, G Agglomerated Abrasive Material Compositions Comprising Same, and Process for Its Manufacture. US Patent 4,988,369, Jan 29, 1991.
27. Akay, G Coating Process. European Patent 0,382,464, Jan 21, 1993.
28. Akay, G Process for the Manufacture of an Agglomerated Abrasive Material. European Patent 0,303,416, Feb 10, 1993.
29. Akay, G. Flow induced phase inversion agglomeration. Polym. Eng. Sci. **1994**, *34*, 865–880.
30. Akay, G Detergent Powders and Process for Preparing Them. European Patent 534,525, Dec 27, 1996.
31. Akay, G.; Vickers, J Method for Separating Oil in Water Emulsions. European Patent 1,307,402, May 7, 2003.

32. Akay, G.; Tong, L. Preparation of low-density polyethylene latexes by flow-induced phase inversion emulsification of polymer melt in water. J. Colloid Interface Sci. **2001**, *239*, 342–357.
33. Akay, G.; Tong, L.; Hounslow, M.J.; Burbidge, A.S. Intensive agglomeration and microencapsulation of powders. Colloid Polym. Sci. **2001**, *279*, 1118–1125.
34. Akay, G.; Tong, L.; Bakr, H.; Choudhery, R.A.; Murray, K.; Watkins, J. Preparation of ethylene vinyl acetate copolymer latex by flow induced phase inversion emulsification. J. Mater. Sci. **2002**, *37*, 4811–4818.
35. Akay, G.; Dawnes, S.; Price, V.J Microcellular Polymers as Cell Growth Media and Novel Polymers European Patent 1,183,328, Jun 3, 2002.
36. Akay, G. Flow induced phase inversion: mechanism and applications. In *Recent Advances in Transport Phenomena*; Dincer, I., Yardim, F., Eds.; Elsevier: Paris, 2000; 11–17.
37. Williams, J.; Wrobleski, D.A. Spatial distribution of the phases in water-in-oil emulsions: open and closed microcellular foams from cross-linked polystyrene. Langmuir, **1988**, *4*, 656–662.
38. Akay, G.; Bhumgara, Z.; Wakeman, R.J. Self-supported porous channel filtration modules: preparation, properties, and performance. Chem. Eng. Res. Des. **1995**, *73*, 782–797.
39. Akay, G Microporous Polymers. International Patent Application; PCT Publication WO 2004/005355, Jan 15, 2004.
40. Akay, G.; Birch, M.A.; Bokhari, M.A. Microcellular polyhipe polymer (PHP) supports osteoblastic growth and bone formation in vitro. Biomaterials **2004**, *25*, 3991–4000.
41. Akay, G.; Dogru, M.; Calkan, B.; Calkan, O.F. Flow induced phase inversion phenomenon in process intensification and micro-reactor technology. Process intensification in water-in-crude oil emulsion separation by simultaneous application of electric field and polymeric demulsifiers. In *Microreact Technology and Process Intensification*; Wang, Y., Halladay, J., Eds.; Oxford University Press: Oxford, 2005; Chapter, 18.
42. Erhan, E.; Yer, E.; Akay, G.; Keskinler, B.; Keskinler, D. Phenol degradation in a fixed bed bioreactor using micro-cellular polymer-immobilized Pseudomonas syringae. J. Chem. Technol. Biotech. **2004**, *79*, 195–206.
43. Erhan, E.; Keskinler, B.; Akay, G.; Algur, O.F. Removal of phenol from wastewater by using membrane immobilized enzymes: part 1. Dead end filtration. J. Membr. Sci. **2002**, *206*, 361–373.
44. Akay, G.; Noor, Z.Z.; Dogru, M. Process intensification in water-in-crude oil emulsion separation by simultaneous application of electric field and polymeric demulsifiers. In *Microreactor Technology and Process Intensification*; Wang, Y., Halladay, J., Eds.; Oxford University Press: Oxford, 2005; Chapter 23.
45. Dogru, M.; Akay, G. Gasification International Patent Application; PCT Publication WO 2005/047435, 26 May, 2005.
46. Pekdemir, T.; Akay, G.; Dogru, M.; Merrells, R.E.; Schleicher, B. Demulsification of highly stable water-in-oil emulsions. Sep. Sci. Technol. **2003**, *38*, 1161–1184.
47. Akay, G.; Keskinler, B.; Cakici, A.; Danis, U. Phosphate removal from water by red mud using crossflow microfiltration. Water Res. **1998**, *32*, 717–726.
48. Keskinler, B.; Akay, G.; Bayhan, Y.K.; Erhan, E. The effect of ionic environment on the crossflow microfiltration behaviour of yeast cell suspensions. J. Membr. Sci. **2002**, *206*, 351–360.
49. Scamehorn, J.F.; Christian, S.D.; El-Sayed, D.A.; Uchiyama, H.; Younis, S.S. Removal of divalent metal cations and their mixtures from aqueous streams using micellar enhanced ultrafiltration. Sep. Sci. Technol. **1994**, *26*, 809–830.
50. Akay, G.; Odirile, P.T.; Keskinler, B.; Wakeman, R.J. Cross microfiltration characteristics of surfactants. In *Surfactant Based Separations: Science and Technology*; Scamehorn, J.F., Harwell, J.H., Eds.; ACS Symposium Series: Washington, DC, 2000; Vol. 740, 175–200.
51. Akay, G.; Wakeman, R.J. Electric field intensification of surfactant mediated separation processes. Chem. Eng. Res. Des. **1996**, *74*, 517–525.
52. Akay, G.; Odirile, P.T. Metal ion removal using electric field intensified mediated crossflow filtration process. J. Membr. Sci. *in press*.
53. Keskinler, B.; Akay, G.; Pekdemir, T.; Yildiz, E.; Nuhoglu, A. Process intensification in wastewater treatment: oxygen transfer characteristics of a jet loop reactor for aerobic biological wastewater treatment. Int. J. Environ. Technol. Manag. **2004**, *4*, 220–235.
54. De Bartolo, L.; Morelli, S.; Bader, A.; Drioli, E. The influence of polymeric membrane surface free energy on cell metabolic functions. J. Mater. Sci. Mater. Med. **2001**, *12*, 959–963.
55. Bokhari, M.; Birch, M.; Akay, G. Polyhipe polymer: a novel scaffold for in vitro bone tissue engineering. Adv. Exp. Med. Biol. **2003**, *534*, 247–254.
56. Akay, G.; Dogru, M.; Calkan, O.F.; Calkan, B. Biomass processing in biofuel applications. In *Biofuels for Fuel Cells*; Lens, P., Westerman, P., Haberbauer, M., Menero, A., Eds.; IWA Publishing: London, 2005; Chapter 4.

Bioprocessing

Ryan G. Soderquist
James M. Lee
Department of Chemical Engineering, Washington State University, Pullman, Washington, U.S.A.

INTRODUCTION

Bioprocessing utilizes biological processes to convert raw materials into useful substances. For centuries mankind has converted raw materials into desirable products by harnessing biological processes. In fact the use of baker's yeast, or leaven, to enhance the quality and texture of bread, as well as the fermentation of grape or cereal extracts to generate ethanol, are common bioprocesses that have been utilized since ancient times. Very little was known about the mechanisms responsible for these important conversions, however, until 1856 when Pasteur characterized yeast as a living organism converting sugar to ethanol.[1]

Scientific progress in more recent times has not only revealed the mechanisms of traditional bioprocesses, but has also improved the ability of bioprocessing to generate useful products. Advances in pure culture and other classical bacteriological methods initiated the first commercial production of lactic acid by fermentation in 1881.[1] The subsequent development of the first cholera, diphtheria, and tetanus vaccines from 1885 to 1890 greatly improved human health.[1] The production of penicillin and the subsequent emergence of recombinant DNA technology in the early 1970s, however, significantly enhanced the potential of bioprocessing.

Bioprocessing is now a diverse technology, and the potential of bioprocessing seems to have very few limits. New applications of bioprocessing overshadow its traditional role as the fundamental technology for fermented food and beverage production. Pharmaceutical production is one of the fastest growing fields of bioprocessing. Biological processes facilitate the production of a wide array of pharmaceutical compounds, which are typically proteins that act as enzymes or antibodies. Biological processes are also critical in toxic waste degradation, enhanced oil recovery, and other environmental efforts.[2] Modern bioprocessing has several common components and fundamentals, which are the focus of this entry.

BIOPROCESSING TECHNOLOGY

Biological processes use living cells (microbial, animal, or plant), or components extracted from living cells such as enzymes, to create desired chemical or physical changes.[2] The major products of bioprocessing include bulk organics, organic acids, amino acids, antibiotics, extracellular polysaccharides, nucleotides, enzymes, vitamins, alkaloids, pigments, vaccines, therapeutic proteins, monoclonal antibodies, and insecticides.[3] Biological processes are also used in wastewater treatment, sludge processing, composting, and bioremediation (i.e., the degradation of crude oils and toxic organic chemicals).[1] These more recent applications of bioprocessing are in addition to the manufacture of traditional foods and beverages such as yogurt, bread, vinegar, soy sauce, beer, and wine.

The use of bioprocessing to induce the desired physical or chemical changes has advantages and disadvantages over traditional chemical processes that do not employ living cells and their components. The advantages of biological processes include mild reaction conditions, specificity, and the use of renewable resources.[2] Most biological processes are conducted at temperatures in the range of 25–37°C, at a pH range of 5–8, and at atmospheric pressure. Therefore, the operation is less hazardous and the manufacturing facilities are less complex compared to a traditional chemical process. The use of biomass is also a noteworthy production advantage, as biomass is easily regenerated without deleterious effects on the environment.

The disadvantages of biological processes for industrial-scale processes, however, include complex product mixtures, dilute aqueous environments, contamination, and variability.[2] When living cells are used to produce a desired product, multiple enzyme reactions typically occur, and the final product mixture contains a complex mixture of cell mass and its components, metabolic by-products, and remnants of the original nutrients. The desired product must be separated from this complex product mixture. Biological processes secrete small amounts of a product into an inherently dilute aqueous environment, which further complicates separation techniques. Contamination from unwanted environmental bacteria and molds is a paramount concern with biological processes, particularly when slow-growing plant and animal cells are used. Living cells tend to mutate over time and can lose characteristics vital for the success of a process.[2]

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120025834

Some cellular components are also unstable at room temperatures for long periods of time, which can limit a continuous process. Thus, the advantages and disadvantages of biological processes should be compared against traditional chemical methods on a case-by-case basis.

ENZYME-MEDIATED PROCESSING

To convert raw materials into a desired product, it is advantageous to utilize cellular components instead of complete cells in some instances. Enzymes are the most common cellular components used to catalyze biochemical reactions. They are a class of proteins that increase the rate of a reaction without undergoing permanent chemical changes themselves. Enzymes can be obtained either from a pure culture of micro-organisms or directly from plants and animals.[2] A small amount of enzyme can produce large amounts of a product because enzymes are highly specific and catalyze specific chemical reactions. They also increase the rate of a reaction much faster than nonbiological catalysts.

A historical use of enzymes is the use of cow stomachs, containing the enzyme rennin, to curdle milk and initiate the cheese-making process. Enzymes are now applied commercially on both large and small scales. The three major categories for commercial production are industrial enzymes, analytical enzymes, and medical enzymes.[2] Industrial enzymes are applied in tons, while analytical and medical enzymes are applied in the range of milligrams to grams.

There are several bulk industrial enzymes that are applied on a large scale, as outlined by Ward.[1] Carbohydrases are used to process starches. Cellulase can convert cellulose, the largest component of biomass, into sugars that can be further converted into fuel ethanol via fermentation. Proteases have several industrial applications, particularly in the food processing industry. Lipases and lactases are used in several food processing applications as well.

Ward lists several key enzyme properties that are important for applications in bioprocessing.[1] The primary considerations are the specificity of an enzyme with respect to the type of reaction being carried out, the substrates used, and the nature of the product. Another vital aspect for efficient processing is the enzyme solubility and stability at different pH and temperature conditions. The structure, charge, and shape of an enzyme must also be known when purification and recovery of the enzyme are desirable.

It is necessary to develop rate equations from kinetic studies that quantify how the rate of an enzymatic reaction is affected by various chemical and physical conditions to calculate reaction times, yields, and optimum economic conditions. These studies are typically done by measuring the concentrations of product and substrate with respect to time at different enzyme concentrations. Lee describes the three most common approaches to derive the rate equation for an enzymatic reaction.[2] The Michaelis–Menton method assumes that there is an enzyme–substrate complex that forms much faster than the product-releasing step. In this approach, the rate equation is simplified by assuming that the slow step determines the rate, and that the other step is at equilibrium. The Briggs–Haldane method assumes that the change in the enzyme–substrate concentration with respect to time is negligible. This is also known as the pseudo–steady-state assumption. Finally, the numerical approach determines the overall rate expression for an enzymatic reaction by numerically solving the differential equations simultaneously for substrate and product concentrations.

The reuse of an enzyme can be economically favorable when a high-cost enzyme is used. It can be difficult to separate and reuse an enzyme because enzymes are typically globular proteins that are highly soluble in water. A common technique to facilitate the reuse of a high-value enzyme is to immobilize the enzyme onto a surface, inside of an insoluble matrix or within a semipermeable membrane. Both chemical and physical means can be employed to immobilize enzymes. The former method involves the covalent attachment of enzymes to water-insoluble supports and is the most widely used method for enzyme immobilization.[2]

Lee outlines three different physical methods that are commonly utilized for enzyme immobilization.[2] Enzymes can be adsorbed physically onto a surface-active adsorbent, and adsorption is the simplest and easiest method. They can also be entrapped within a cross-linked polymer matrix. Even though the enzyme is not chemically modified during such entrapment, the enzyme can become deactivated during gel formation and enzyme leakage can be problematic. The microencapsulation technique immobilizes the enzyme within semipermeable membrane microcapsules by interfacial polymerization. All of these methods for immobilization facilitate the reuse of high-value enzymes, but they can also introduce external and internal mass-transfer resistances that must be accounted for in design and economic considerations.

CELL-MEDIATED PROCESSING

Bioprocessing also uses living cells as an alternative to enzymes and other cellular components. Each different cell species contains a variety of cellular components that uniquely transform raw materials into desirable products. The use of living cells thus eliminates the need for production and purification of a specific enzyme.

Living cells, however, can also produce by-products that are potentially harmful to certain products.

Fig. 1 shows a typical biological process that utilizes living cells.[2] First, a culture of the desired cell line is used to inoculate a small flask, referred to as a shake flask. This culture is then used to inoculate a small bioreactor. The culture contents of this small bioreactor are then used to inoculate a larger bioreactor. Inoculums are subsequently transferred to reactors of increasing size until the culture volume is large enough to inoculate a production-scale bioreactor. The cells are then separated from the liquid medium when the culture growth in the production-scale bioreactor is complete. The subsequent recovery steps depend on whether the product of interest was secreted to the supernatant or retained within the cells during production. A series of purification steps are then typically required before a product is ready for packaging.

Ward discusses several considerations in the selection of organisms utilized in a bioprocess.[1] The gaseous requirements of an organism, particularly oxygen and carbon dioxide requirements, can influence reactor design parameters. The nutritional requirements of different organisms are also a consideration. Bacteria, fungi, and yeasts can require nitrogen, minerals, and specific nutrients in addition to the carbohydrates used as a carbon source. Plant cell cultures require only a well-defined mixture of salts and carbohydrates. Mammalian cells, on the other hand, need a variety of complex nutritional requirements that can be very expensive. The strength of an organism's cell wall is a key consideration, because the ability of an organism to withstand shear forces generated from agitation plays an important role in reactor design. The growth rate of a cell line also influences processing considerations. As a general rule, the average growth rate of a cell population decreases with increasing cell complexity. Hence, mammalian cells will grow slower than plant cells, which in turn will grow much slower than yeasts and bacteria.

Small-scale trials are conducted to determine the growth and production characteristics of a cell line as a function of the culture environment once a suitable host organism is selected. A small-scale culture typically requires shake flasks with capacities in the range of 125 ml to 1 L. The primary environmental conditions that are optimized at this stage of the process include the medium composition, pH, and temperature.[3] Other parameters that can be calculated from small-scale experiments include cell growth rates, specific productivities, and product yields.

Understanding the growth kinetics of a cell system is particularly important in bioreactor design. Lee describes a typical growth cycle for the cultivation of cells in a batch process.[2] The lag phase of growth typically occurs immediately following inoculation, and is a period where cellular growth and division is minimal. This lag phase usually occurs because the cells must adjust to the new medium before growth can begin. The cells may increase in size during this period even though the cell density does not increase. The growth of a culture then accelerates and the cell density starts to increase until the cellular division rate approaches a maximum value. During the exponential growth phase, the cellular division rate remains at a maximum and constant value, and the cell density increases exponentially as the cells rapidly divide. At the close of the exponential growth phase, the growth decelerates because the cell density begins to approach a maximum value, and there is a consequent decrease in the cellular division rate. During the stationary phase of growth, the cell density remains at its maximum value and does not increase any further. Finally, during the death phase of growth toxic metabolic by-products accumulate and the cells deplete nutrients, which results in cellular lysis.

BIOREACTOR UTILIZATION

A bioreactor is a device that transforms raw materials into products by means of biological processes conducted by enzymes or living cells.[2] Bioreactors are frequently called fermenters when they generate a product by means of microbial cells, because the growth of microbes in a vessel is commonly called fermentation. There are several different bioreactor designs that can be customized to a desired process.[3] The processing conditions in a reactor, such as temperature, pH, and aeration, depend on the aspects of a particular

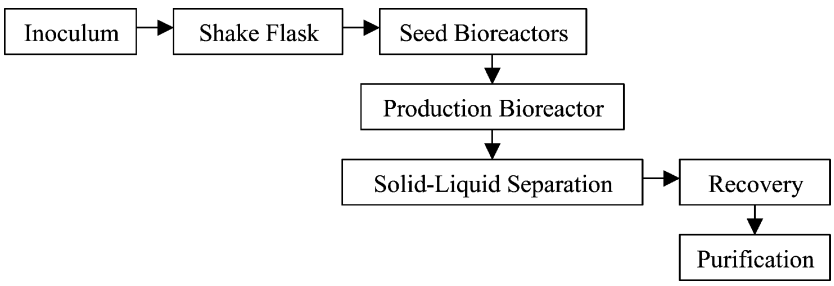


Fig. 1 Steps in a typical bioprocess.

reactor. The mode of reactor operation, such as batchwise or continuous production, is also configured for a specific biological process.

The main challenge in bioreactor design for aerobic cell lines is to provide adequate mixing and aeration. A stirred tank reactor is most commonly employed for large-scale aerobic bioprocesses.[2] In a stirred tank reactor, mechanical agitation by an impeller accomplishes the mixing and bubble dispersion, and there are a variety of impeller configurations available. A conventional stirred tank reactor also contains a number of baffles at the vessel perimeter.[1] When air is introduced at the base of the tank through the use of a ring sparger, the rotating impeller and baffle system facilitate the dispersion of air bubbles throughout the medium.[1] The agitator can, however, damage shear-sensitive systems such as mammalian and plant cell cultures.[2] A common design assumption is that a stirred tank reactor is well mixed at each point in the vessel during all of the production phases.

A bioreactor is referred to as a batch reactor when all of the substrates are added to the vessel before inoculation, and the products are removed only at the end of the process. This definition of batch operation refers to the addition and removal of liquid and solid components from the system, but does not include the continuous addition of oxygen that must be supplied to aerobic systems. Most commercial bioreactors are operated in the batch mode.[3] To comply with manufacturing regulations, a batch process is typically advantageous because contamination or genetic changes in a cell line can occur during a long-term process.

Fed-batch operation incorporates intermittent or continuous feeding of nutrients during production. This type of operation can mitigate the effects of catabolite repression, reduce the viscosity of the medium, lessen the effects of toxic medium components, or extend the production phase of a process for as long as possible.[1] Examples of processes where fed-batch operation is particularly useful include the production of baker's yeast and penicillin.[3]

Stirred tank reactor systems can also operate in a continuous mode. In this configuration, fresh medium is continually supplied to the reactor and the desired products are continuously removed in the course of production. A continuous system is referred to as a chemostat when the flow rate is set to a constant value. It is further known as a turbidostat when the flow rate is set to maintain a constant turbidity or cellular concentration.[2] Continuous reactor systems are commonly abbreviated as CSTR or CSTF and they refer to continuous stirred tank reactor and continuous stirred tank fermenter, respectively.

An important parameter to minimize in the design and operation of a continuous reactor is the residence time, which is the ratio of the reactor volume over the reactor flow rate. When a product is formed during the exponential growth phase, continuous operation is advantageous over batch operation because the residence time is lower in the first scenario. Batch operation is preferred when most of the production occurs during the stationary phase of growth, because the residence time for a batch operation in this scenario is much smaller.[2] Concerns with continuous operation arise from foaming and plugging of the system with large cell aggregates, along with contamination and genetic instability of a cell line. Continuous reactors are particularly useful when a process employs immobilized cells and enzymes, or when cell recycling can be incorporated.

The tubular flow reactor is another common reactor configuration. In this reactor, cells and medium are continuously added at one end of a cylindrical tube and biological processes occur as the reaction mixture flows along the tube. The fluid in a tubular reactor can flow either vertically or horizontally.[3] There is no fluid mixing in this method of operation, and thus the stream composition varies in both the radial and the longitudinal directions. Variations in the radial direction, however, are typically small compared to variations in the longitudinal direction. When radial variations are neglected, the reactor is known as a plug flow reactor, frequently abbreviated as PFR. Plug flow operation can be a simulation of batch operation in a continuous flow system because substrate and product conversions vary as the reaction proceeds along the tube in a manner analogous to batch operation.

Doran lists several considerations between batch, continuous stirred tank, and plug flow reactor configurations.[3] A number of continuous stirred tank reactors configured in series will theoretically approach the same reaction characteristics of those in a plug flow or a batch reactor. For single reactors not in series, the reaction kinetics determine the benefits between reactor configurations. If the reaction is zero order, there is no difference between batch, continuous stirred tank, and plug flow reactors, as far as the reaction considerations are concerned. For first-order reactors, however, batch and plug flow reactor operations will lead to higher conversions than continuous stirred tank reactors with the same volume. This is because high reactant concentrations are optimal for first-order reaction kinetics and the reactants in a CSTR are instantly diluted as they enter the vessel. When a reaction is autocatalytic, which is the case with biological reactions where reaction rates increase as the amount of biomass increases, the rate of conversion in a continuous stirred tank reactor can be significantly greater than in a batch or plug flow reactor.

In addition to the standard modes of reactor operation, alternative reactor configurations also facilitate

complicated bioprocesses. These alternative configurations are typically designed to improve on the disadvantages of traditional stirred tank reactors such as the high-power requirements of impeller operation and shear damage from impeller tips. Additionally, alternative reactor configurations can more efficiently meet the specific requirements of a system, such as better aeration, more effective heat removal, increased cell separation and cell retention, and cell immobilization.[2]

Airlift reactors are an alternative bioreactor in which the agitation is accomplished by a clear cyclic pattern of air flow through both a riser and a downflow reactor compartment. Stirred tank and airlift reactors are most often used for aerobic cell cultures. Airlift reactors are often utilized to culture plant and animal cells because they typically have lower shear levels.[3]

Another alternative bioreactor is the bubble column, which is composed of a long cylindrical vessel with a sparging device at the bottom. In a bubble column, the mixing action of rising bubbles fulfills the oxygen needs of the cells.[2] Bubble columns are used industrially for the production of baker's yeast, beer, and vinegar, and for wastewater treatment.[3] While column reactors are the simplest type of reactor, they are inflexible and limited to a relatively narrow range of operating conditions.[2] For example, if the fluid has a high viscosity, a column reactor cannot provide adequate mixing and mass-transfer characteristics.[3]

Apart from the aeration, mixing, and conversion parameters associated with a particular reactor configuration, there are also several practical considerations that must be accounted for in the design of an industrial bioreactor. Bioreactors must be able to withstand up to 3 atm of positive pressure and temperatures of up to 150–180°C, which is why most large vessels are typically made of steel.[3] The aseptic operation of a vessel is a key concern as 3–5% of industrial fermentations are lost because of a failure in sterilization procedures.[3] Bioreactor control is also an important consideration as deviations in reactor conditions can impact cellular growth and metabolism. Most bioreactors contain on-line measurement devices that determine the temperature, pressure, aeration rate, agitation rate, pH, and other key parameters for a designated biological process.

Doran outlines several of the key parameters involved in the scale-up of a biological process from laboratory-scale shake flasks to production-scale bioreactors.[3] In the first stage of studies a bench-top bioreactor, typically 1–2 L, is used to determine the oxygen requirements of the cells, their shear sensitivity, foaming characteristics, and any limitations that the reactor imposes on the organism. The results of these early studies enable decisions regarding operation in the batch, fed-batch, or continuous mode. A pilot-scale bioreactor, which typically has a capacity of 100–1000 L, is used in the next phase of process development. The pilot-scale reactor maintains an equivalent geometry, method of aeration and mixing, impeller design, and other important features of the bench-top bioreactor. A loss or variation in the performance of a cell line often occurs in the conversion from a bench-top to a pilot-scale bioreactor, even though the described reactor aspects remain constant. Thus, the results of pilot-scale findings often necessitate the reassessment of economic projections and design considerations. The industrial-scale phase of process development commences after the successful completion of the bench-top and pilot-scale phases. During this phase of development, the production-scale bioreactor is tested, along with all of the auxiliary services, such as the air supply, sterilization process, steam and cooling water supply lines, and reactor control devices.

DOWNSTREAM PROCESSING

After biological reactions have generated a product of interest, it is necessary to recover this product from a liquid mixture that typically contains several undesirable components. The treatment of any culture broth after bioreactor cultivation is known as downstream processing. Downstream processing can account for 60–80% of the total production cost, particularly in the production of modern recombinant proteins and monoclonal antibodies.[2,3] A typical downstream process requires several steps in the areas of solid–liquid separation, cell rupture, product recovery, and product purification. It is important to minimize the number of downstream processing steps required because significant product losses inevitably occur during each step.[3]

The product from a desired fermentation is either secreted from the cells into the fermentation broth or is entrapped within the cells themselves. In either case, the first step in downstream processing is to separate the cells from the fermentation broth. This step is typically accomplished by filtration or centrifugation and is known as solid–liquid separation.

Filtration separates cells from a fluid by forcing the fluid through a porous filter medium, which deposits solids as liquids pass through. Vacuum or positive-pressure equipment is used to create the driving force for filtration. The main advantages of filtration include high rates of separation, low cost, mechanical simplicity, and relative ease of maintenance. However, it can have a low retention or poor containment, and can require the addition of a filter aid to ensure good filtration when solids accumulate on the membrane.[4]

Centrifugation is an alternative to filtration. In bioprocessing, centrifugation removes cells from the fermentation broth, eliminates cell debris, and collects precipitates.[3] The equipment used for this purpose is usually more expensive than filtration equipment. It, however, is more effective than filtration when particles are small and difficult to filter. There are two basic types of centrifuges that are employed on a large scale. A tubular centrifuge is a hollow cylindrical tube in which a suspension is fed through the bottom and the clarified liquid is removed from the top.[2] Solids are deposited on the wall of a tubular centrifuge during this process. The solids are removed manually from the wall, which hinders continuous operation. A disk centrifuge is more commonly utilized when continuous operation of the centrifuge is required. This kind of centrifuge is a short and wide bowl that spins on a vertical axis and contains closely spaced cone-shaped disks that increase the capture of particles on the surface. The clarified liquid flows toward the center of the bowl and is expelled through the annulus. The solids are thrown outward and can be discharged intermittently or continuously.[2]

If the desired product is entrapped within the cells themselves, the cells are disrupted by mechanical or chemical means. The former include grinding with abrasives, high-speed agitation, high-pressure pumping, and ultrasonic techniques.[3] The latter include the treatment of cells with surfactants, alkalis, organic solvents, or osmotic shock.[2] In general, the mechanical methods of disruption have the widest application, while the chemical methods are generally cell and product specific and thus not applicable to all systems.[5] While cell rupture techniques have to be very powerful, care must taken not to damage the desired components. For example, low temperatures are often required during cellular disruption procedures to prevent product loss.

The recovery phase of a bioprocess begins once the products are freely suspended in a solution, either in the supernatant from the solid–liquid separation or in the lysate from the cellular disruption. Extraction is the process of separating the desired components in a liquid solution by contact with another insoluble liquid into which the desired product will be selectively extracted. This process is also commonly known as liquid–liquid extraction. The solvent-rich phase is called the extract and the residual liquid from which the desired product has been removed is called the raffinate.

Adsorption serves the same purpose as extraction in the recovery, or isolation, of the desired product from a dilute solution. In adsorption, a specific substance is adsorbed selectively by certain solids as a result of a specific physical or chemical interaction.[2] Three common adsorption methods include conventional adsorption, ion-exchange adsorption, and affinity adsorption. The adsorption of organic chemicals onto charcoal or porous polymeric adsorbents is a broad application of conventional adsorption.[3] Conventional adsorption is a reversible process caused by intermolecular forces of attraction.[2] Ion-exchange adsorption occurs when charged ions reversibly interact with charged surfaces on the molecule of interest, and it is an established practice for the recovery of amino acids, proteins, antibiotics, and vitamins.[3] Affinity adsorption is a very selective process that depends on specific chemical interactions between the molecule of interest and a solid resin.[1]

The product is often purified further after the recovery phase, particularly when the desired product is a protein of pharmaceutical interest. Precipitation is a widely used method that can be induced by the addition of salts, organic solvents, or heats.[2] It can both purify and concentrate a particular protein fraction, and is frequently accomplished by the addition of salts for a ''salting out'' effect.[1] The addition of a salt precipitates proteins because increasing salt concentrations reduce the solubility of a protein in a solution.[2] Even though precipitation is an effective and relatively inexpensive method, it is also a fairly crude step and is often followed by chromatographic separations.[6]

Chromatography is a purification method similar to adsorption as both methods involve an interaction between a solute and a solid matrix. Many principles of adsorption also apply to chromatography. A solution composed of several different solutes is injected at one end of a chromatography column, and each solute moves at a rate dependent on its relative affinity for a solid and immobile resin, which is known as the stationary phase. Solutes with a higher affinity for the stationary phase will move through the column more slowly than those with a lower affinity for the stationary phase. Chromatography differs from adsorption because it is based on the different rates of movement for solutes in a column, rather than the capture of one solute by an adsorbent.[2] The stationary phase in chromatography is commonly packed in a cylindrical column and can be an adsorbent, an ion-exchange resin, an affinity resin, a porous solid, or a gel.[2] A chromatography process is typically scaled up by maintaining a constant column height, fluid velocity, and sample concentration, while increasing the column diameter and the volumetric flow rate in proportion to the column cross-sectional area.[7]

In electrophoresis, the net surface charge and the size of molecules separate them from one another in an electric field.[1] Electrophoresis is one of the most effective methods of protein separation and purification. It has a very high resolving power to clearly separate charged protein molecules. This method of

separation, however, is not amenable to all mixtures as the components must have an ionic form and each component must migrate differently in an electric field.[2]

Membrane separation is a purification step that uses polymeric membranes to separate components primarily on the basis of size. Pressure is the main driving force in membrane separations. Microporous membrane filters are frequently used to sterilize biological solutions and have pores ranging from 0.025 μm to several micrometers in diameter.[1] Reverse osmosis and ultrafiltration membranes can separate on the basis of molecular size with ranges of 500–1000 Da and 1000–1,000,000 Da, respectively.[1] A cross-flow configuration, in which recirculation of the liquid through thin channels provides a constant flow parallel to the membrane surface, frequently reduces the accumulation of molecules on the membrane surface.[2]

After purification, product packaging and marketing complete a bioprocess. Lyophilization is a common packaging technique for long-term storage. Lyophilization utilizes the principle of sublimation to dry a product in a manner that facilitates rehydration.[1] Medical and clinical trials are required before a product can be released as a health-care application if the product is a new pharmaceutical. Most regulatory agencies also require a set of well-documented manufacturing protocols and standards during all of the production phases.[3]

CONCLUSIONS

As the potential applications of biotechnology are vast, bioprocessing will be a key technology in the future. Scientific breakthroughs made in gene expression and protein engineering will rely on bioprocessing techniques to incorporate these breakthroughs into new products and services. Developments in biological processing will not only continue to impact food and beverage production, but will also significantly advance health-care and environmental conservation efforts. A sound understanding of bioprocessing principles will be increasingly important to the scientists and engineers of the future to utilize new biological processes on a large scale.

REFERENCES

1. Ward, O.P. *Bioprocessing*; Bryant, J.A., Kennedy, J.F., Eds.; Open University Press Biotechnology Series; Open University Press: Buckingham, U.K., 1991.
2. Lee, J.M. Biochemical engineering. In *Prentice Hall International Series in the Physical and Chemical Engineering Sciences*; Amundson, N.R., Ed.; Prentice Hall: Englewood Cliffs, NJ, 1992.
3. Doran, P.M. *Bioprocess Engineering Principles*; Academic Press Limited: London, 1995.
4. Reed, I.; Mackay, D. Clarification techniques. In *Protein Purification Techniques*, 2nd Ed.; Roe, S., Ed.; Oxford University Press: Oxford, 2001; 51–82.
5. Cumming, R.H.; Iceton, G. Cell disintegration and extraction techniques. In *Protein Purification Techniques*, 2nd Ed.; Roe, S., Ed.; Oxford University Press: Oxford, 2001; 83–108.
6. Harris, E.L.V. Concentration of the extract. In *Protein Purification Techniques*, 2nd Ed.; Roe, S., Ed.; Oxford University Press: Oxford, 2001; 111–154.
7. Hagel, L. Separation on the basis of chemistry. In *Protein Purification Techniques*, 2nd Ed.; Roe, S., Ed.; Oxford University Press: Oxford, 2001; 155–183.

Bioremediation

Teresa J. Cutright
Department of Civil Engineering, The University of Akron, Akron, Ohio, U.S.A.

INTRODUCTION

Biologists agree that the natural biodegradation of decaying plants, animals, and other organics by organisms has occurred since the beginning of time. The natural degradation of organic matter is the direct result of the microorganism's metabolism to protect itself from the environment, obtain nutrients, produce energy in a usable form, convert nutrients into cellular form, discard unnecessary products, and replicate genetic information. Mankind has been able to manipulate microbial metabolism for his own benefit. The first documented use of this dates back to the use of yeast for bread in ancient Egypt and fermentation of grapes in ancient Greece.[1] The implementation of microbiological processes evolved with man to encompass other food and pharmaceutical applications. By the beginning of the 20th century, bacteria applications had expanded to include the treatment of wastewater. However, it was not until the early 1970s that scientists investigated the viability of the bioremediation of contaminated soil.[2] Although the mechanisms behind bioremediation have existed since the dawn of time, its application for the treatment of hazardous contaminants is still considered a new technology.

Bioremediation as a remedial technology has several advantages. It can:

- Simultaneously treat subsurface soil and water.
- Be completed in less time than traditional pump-and-treat operations.
- Eliminate long-term liability.
- Result in complete mineralization of the contaminants.
- Be conducted under mild operating conditions.
- Easily be combined in a treatment train.
- Be completed with little or minimal stress on environmental and ecological systems.

Because bioremediation is still considered a relatively new technology, it can be scientifically intensive at the beginning of the project. For instance, a thorough site characterization can be required to ascertain the extent of contamination. In addition, extensive monitoring can be required to gain an understanding of the chemical and biological process responsible, and to track degradation by-products. Overcoming the mass transfer limitations during in situ applications and long treatment times are two other common disadvantages. However, the primary limitation to successful bioremediation is that it is often initiated without knowing the full picture. For instance, other unidentified contaminants, or even the degradation by-products may be toxic to the microbes. This can result in increasing the time required for the treatment or stopping remediation altogether. Bioremediation is still an extremely desirable treatment approach as the advantages still far outweigh the disadvantages. Bioremediation is desirable because it is very cost-effective and the most environment-friendly remediation technology. This entry contains an overview of bioremediation including the common terms associated with bioremediation, the requirements for successful applications, the key steps for treatability studies, and an introduction to degradation pathways.

TERMINOLOGY AND CATEGORIES ASSOCIATED WITH BIOREMEDIATION

The most basic definition of bioremediation is the ''use of any living organism to convert contaminants into less harmful compounds.'' Previously this definition was restricted to microorganisms or microbial processes (e.g., microbial mediated enzyme reactions), where microorganisms referred to eubacteria, archebacteria, and unicellular organisms such as bacteria, yeasts, fungi, and protozoa. Although ''living organisms'' encompass multicellular organisms, it was not until the early 1990s that bioremediation applications were extended to include plants (i.e., phytoremediation). As shown in Fig. 1, the bioremediation umbrella contains several classifications depending on the approach implemented, additive(s) introduced, and location of the treatment. The subdivision of organisms based on their metabolic characteristics (i.e., chemotrophs, phototrophs, mesophiles, etc.) are not presented here but can be found in detail in several texts.[3–5]

The top of Fig. 1 contains the most common descriptors applied to bioremediation. Mineralization is the complete conversion of an organic compound into biomass, carbon dioxide, water, and salts. The difference between the degradation of natural organic

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007679

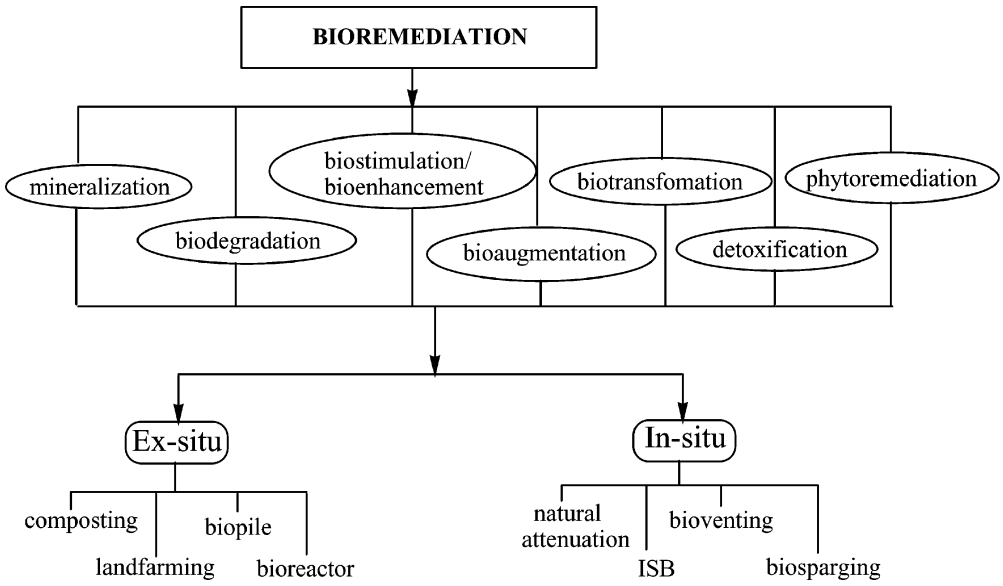
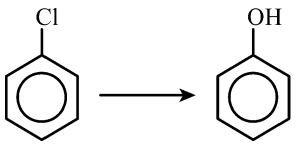
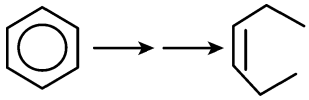


Fig. 1 Descriptive terms for bioremediation applications.

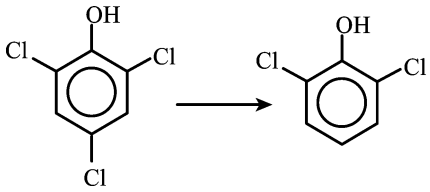
matter and chemical contaminants is the ease of degradation, complexity of the structure to be degraded, and time required for achieving complete mineralization. Biodegradation is also used to describe any of the microbial mediated degradation steps prior to mineralization. It includes substitution or transformation reactions that cause a simplification in the compound structure. For instance, a substitution process would be the introduction of a hydroxyl group in place of chloride [Eq. (1)], while transformation would facilitate the breaking of intermolecular bonds [Eq. (2)]. The generic term of transformation does not differentiate between whether the primary structure of the parent compound is still evident or not. As shown in Eq. (2), the structure of the parent compound is no longer easily discernable. Conversely, biotransformation is the alteration of the molecular structure via microbial mediated enzymatic catalysis where the structure of the parent compound is still evident [Eq. (3)]. The goal of the intermediate steps in both biodegradation and biotransformation is to yield a compound that is more susceptible to microbial attack (i.e., less recalcitrant). Detoxification refers to both abiotic and biotic processes that render the compound less toxic. The biotransformation of 2,4,6-chlorophenol to 2,6-chlorophenol as shown in Eq. (3) is an example of detoxification because toxicity generally increases with the increasing number of chloride substitutions.



(1)



(2)



(3)

The bottom of Fig. 1 contains the subcategories used to describe the application, approach, and location of the bioremediation process. In defining the different approaches, it is important to note that indigenous species are those microorganisms inherent to the contaminated media or site. The terms biostimulation and bioenhancement are often interchangable. Both refer to the addition of a terminal electron acceptor (TEA), supplemental nutrients, auxiliary carbon sources, or substrate analogs to stimulate the activity of the indigenous species. The subtle distinction occurs with regard to the location of the additives. Biostimulation is the delivery of the nutrients and/or TEA to the indigenous microbes. Occasionally, bioenhancement is used to distinguish when the stimulation of the ''natural microbes'' occurs in aboveground containers. There are three instances where the indigenous species cannot be stimulated, viz., the appropriate microbial population is not present, bacterial numbers are too low and cannot be increased in an acceptable time frame, and contaminant concentrations or type

are toxic to the bacteria.[4] When the indigenous species cannot be stimulated, bioaugmentation can be tried. Bioaugmentation is the addition of foreign microorganisms for the remediation of a contaminated site. Foreign does not imply the use of genetically engineered species, but refers to any microbe not inherent to the contaminated site. In selecting the foreign species, it is important to ascertain whether the microbes will survive in the new environment, are effective for the target compounds, and can easily be distributed in the subsurface.[6] Often the addition of supplemental nutrients and TEAs (i.e., biostimulation) occurs in conjunction with bioaugmentation to assist with biomass generation. As previously mentioned, phytoremediation is the use of plants for the direct (phytoextraction) or indirect (stimulation of the rhizosphere) treatment of contaminated media.

The physical location of the bioremediation process is divided into two primary categories, i.e., in situ and ex situ, which can also be used in conjunction with the descriptors defined above. In situ applications are those processes that take place in the contaminated media without excavation. ISB is the acronym commonly used for in situ bioremediation. Natural attenuation is also referred to as intrinsic bioremediation. It is the result of several natural processes (abiotic transformation, biodegradation, sorption, etc.) that reduce contaminant concentrations in the environment without human intervention. Bioventing and biosparging are used to distinguish the location of delivered nutrients/TEAs. Bioventing introduces the TEA to stimulate activity in the vadose zone (i.e., the top 2 ft of the unsaturated soil) at approximately 1/10 the delivery rate of soil vapor extraction. Biosparging introduces the material below the groundwater level. Although biosparging is predominantly associated with TEA delivery, it can also be used to describe the introduction of nutrients and/or microbes.

Ex situ applications are processes that occur after the contaminated media have been excavated. Once excavated the media can be treated on-site or taken off-site for subsequent treatment. Composting is the oldest ex situ treatment that utilizes a mixture of mesophilic and thermophilic organisms. Windrows of soil are constructed to a height of 3–4 ft and length of 50–200 ft. Water is added weekly to provide the necessary moisture and regulate the internal temperature of the compost pile. Landfarming occurs on lined containers of various dimensions. The soil is applied at a maximum of 1 ft and is tilled one or two times a month to enhance aeration and nutrient delivery. Biopiles are 6 ft mounds of excavated soil that were premixed to provide a ''uniform'' dispersion of the contaminant. An array of air ducts is placed within the bottom of the pile to introduce the required TEA. Periodically, the piles are sprinkled with water and nutrients to maintain the required moisture and nutrient levels. As with composting and landfarming applications, biopiles are constructed on engineered liners to contain any contaminant releases. Bioreactors are used to describe any contained ex situ biological treatment that does not fall into the categories mentioned above. These systems can be operated in batch, semibatch, or continuous modes with either attached growth or free suspensions of bacteria.

CONTAMINANT CLASSIFICATIONS APPLICABLE TO BIOREMEDIATION

Bioremediation has been successfully demonstrated for a variety of contaminant classifications. The majority of the studies have focused on petroleum compounds (BTEX, gasoline, diesel, jet fuel, etc.) because of their widespread occurrence as a contaminant. The other major waste classifications where bioremediation has been successful are solvents (toluene, trichloroethylene, etc.), creosote, pulp and paper, pesticides, textiles, polycyclic aromatic hydrocarbons (PAHs), and polychlorinated biphenyls (PCBs). Table 1 contains a partial list of the microbial genus successfully implemented for these contaminants. For aqueous petroleum contaminants, bacteria and yeasts are the most prevalent degraders. In contaminated soil systems, bacteria and fungi are the microorganisms responsible for degradation.[7]

Petroleum compounds have also been widely studied as a result of their susceptibility to microbial attack. There are several rules of thumb associated with the ease of degradation in terms of compound structure. In general, as the compound molecular weight increases or the compound's structure becomes more complex, degradation becomes more difficult. The relative biodegradability of contaminants under aerobic conditions, in order of decreasing ease (i.e., the easiest first, the more recalcitrant last) is:[3,4,8]

- Linear alkanes (C_{10}–C_{19})
- Gases (C_2–C_4)
- Branched alkanes (to C_{12})
- Alkanes without substitutions
- Alkenes (C_3–C_{11})
- Branched alkenes
- Aromatics
- Cycloalkanes
- PAHs
- PCBs.

Polychlorinated biphenyls and other halogens can often be reductively dehalogenated much faster anaerobically (i.e., in the absence of oxygen) as compared to their aerobic counterparts.

Table 1 Select microbial genera with successful demonstration of treating different contaminant classifications

Contaminant	Microbial genus		
Aliphatics	*Achromobacter*	*Bacillus*	*Flavobacterium*
	Acinetobacter	*Micrococcus*	*Pseudomonas*
	Arthrobacter	*Mycobacterium*	*Vibrio*
Creosote	*Arthrobacter*	*Cunninghmella*[a]	*Penicillium*[a]
	Aspergillus[a]	*Fusarium*[a]	*Pleurotus*[a]
	Cladosporium[a]	PAH degraders	*Thauera*
Pesticides, herbicides	*Achromobacter*	*Athiorocaceae*	*Penicillium*[a]
	Alcaligenes	*Corneybacterium*[a]	*Methylomonas*
	Arthrobacter	*Flavobacterium*	*Zylerion*[a]
	Pseudomonas		
Dyes	*Aeromonas*	*Phaneorchaete*[a]	*Trametes*[a]
	Micrococcus	*Pseudomonas*	
	Klebsiella	*Shigella*	
Food, dairy, and meat	*Acinetobacter*	*Brevibacterium*	*Rhodococcus*
	Arthrobacter	*Nitrosomonas*	*Vibrio*
	Bacillus	*Pseudomonas*	
PAHs	*Aeromonas*	*Brevibacterium*	*Phanerochaete*[a]
	Achromobacter	*Cunninghmella*[a]	*Pichia*[a]
	Acinetobacter	*Corneybacterium*	*Candida*[a]
	Anthrobacter	*Flavobacterium*	*Pseudomonas*
	Bacillus	*Penicillium*[a]	*Sporobolmyces*[a]
PCBs	*Acinetobacter*	*Corneybacterium*	*Ralstonia*
	Arthrobacter	*Pseudomonas*	*Vibrio*
	Alcaligenes[a]		
Pulp and paper	*Arthrobacter*	*Sporotrichum*[a]	*Trichoderma*[a]
	Eisenia	*Talaromyces*[a]	*Xanthomonas*
	Chromobacter		
Solvents	*Alcaligenes*	*Morganeela*	*Pseudomonas*
	Citrobacter	*Mycobacterium*	*Rhodococcus*
	Desulfomonite	*Nitrosomonas xanthobacter*	
	Enterobacter	*Nocardia*	

[a]Fungi.

It is important to note that there are specific bacteria-contaminant combinations that will be the exceptions to the trend. In addition, this list does not include the impact of bioavailability or concentration. For instance, normal alkanes are considered the most biodegradable of all petroleum hydrocarbons. However, at elevated concentrations the C_5–C_{10} compounds inhibit the activity of several hydrocarbon degraders.[9] Furthermore, the number, type, and position of substitutions will also influence the ease of degradation in branched chains, aromatics, and multiple ringed compounds. As the complexity of the compound's structure and the number of compounds present at a site increase, one microbial strain will not be able to facilitate the complete bioremediation of target compounds and degradation by-products. In these situations, the use of a consortium, or mixed population of bacteria, should enable complete mineralization via cometabolism. Cometabolism is the indirect metabolism of a nongrowth substrate or recalcitrant compound during the transformation of the required growth substrate.

ELEMENTS ESSENTIAL TO BIOREMEDIATION

There are several basic microbial requirements that must be in place if bioremediation is to be successful.

They are essential for supporting any microbial ecosystems and include the appropriate microbes, a substrate for carbon and energy source, TEA, an inducer to facilitate enzyme production, nutrients for microbial maintenance and growth, mechanisms to degrade metabolic by-products, the absence of toxicity and competing organisms, and the appropriate environmental conditions.

Adequate Microbial Population

The primary requirement for any successful biological treatment is the presence of an adequate microbial population. As shown earlier in Table 1, there are a wide variety of microbial species with demonstrated success at remediating an array of contaminants. Soils that have favorable growth conditions will exhibit viable microbial numbers of greater than 10^4 colony forming units (CFU)/g of dry soil. Normal activities in ''clean'' groundwater are in the range of 10^6–10^8 CFU/L. Soils and groundwaters depicting activities $<10^3$ CFU/g soil and $<10^6$ CFU/L are considered nonviable for growth. Such low numbers indicate the lack of an essential nutrient/TEA, the presence of toxic material, an extremely high contaminant concentration, or a combination of all of the above. The cause of nonviable populations can be determined, and possibly corrected for, by a thorough site characterization and treatability study.

Terminal Electron Acceptor and Redox Potential

Aerobic bacteria complete most of the petroleum bioremediation applications, particularly those above the groundwater table. Aerobes are those bacteria that require an oxygen source as their TEA. Conversely, anaerobic species require the absence of oxygen (anoxic conditions) for their respiration. In situ anaerobic bioremediation is typically only conducted in the saturated zone because of the difficulty in maintaining a strict anaerobic environment. In some instances, facultative anaerobes are utilized because they can alter the respiration to be metabolically active under both anaerobic and aerobic conditions. As such, the type of TEA available will dictate the metabolism and subsequent degradation mode. The most common TEAs used for bioremediation are listed in Table 2. Careful selection of microbe–TEA combinations can enable a specific degradation pathway to facilitate cometabolism and prevent undesired degradation by-products.

Appropriate TEA selection is also essential for establishing redox conditions near the microbes.[2] During normal cell metabolism, energy for growth is obtained from the transfer of electrons from the donor to the acceptor. The redox potential is used to describe the availability of electron donors (preferably the contaminants). Reported redox potentials (in millivolts) for each respiration mode vary; however for ISB, redox potentials decrease and become more anaerobic as the depth below the surface increases. For instance, redox potentials >400 mV are representative of aerobic conditions in the vadose zone. Values from 100 to 50 mV are facultative, while those $\leq$300 mV are associated with obligate anaerobic environments.

Table 2 Common TEAs used for aerobic, anaerobic, and fermentation respiration

Respiration mode	Terminal electron acceptor
Aerobic	Oxygen
	H_2O_2
	Various organics
Anaerobic	NO_3^-
	SO_4^{2-}
	Various organics
Methane fermentation	CO_2
	Various organics

Nutrient Requirements

Nutrient amendments are often listed as the most effective approach for increasing biodegradation efficiencies with nutritional requirements being approximately the same as the bacteria's cell composition.[10,11] Using *Escherichia coli* as the reference bacterium, the generic formula for biomass (i.e., bacteria cells) is $C_5H_7NO_2$.[2–4,12] Thus, carbon and nitrogen represent 53% and 12% of the cell, respectively, and are the two nutrients needed most for cellular maintenance and growth. Heterotrophic bacteria can utilize organic contaminants as their primary carbon source. In some instances, the microbe may need greater levels of carbon than is readily available from the contaminant. In these situations, an auxiliary carbon source may be required to either induce degradation or compensate for when contaminant concentrations are too low (not bioavailable). During soil remediation, if the soil organic content is too high it may serve as the preferred carbon source, thereby limiting degradation of the target compound.

Nitrogen is the most commonly added nutrient for bioremediation processes. It is added primarily, in the form of ammonium or nitrate ions, for cellular growth. Nitrate can also be used as an alternative TEA. However, it is often difficult to manage subsurface nitrogen levels because of the microbial conversion of ammonia

to nitrate, nitrate to nitrite, etc. Phosphorus is the second most common nutrient added for cellular growth and maintenance. Potassium phosphate can also serve as a buffering agent. Currently, there is no exact method to predetermine the exact nutrient sources to use at a site. The general rule of thumb for groundwater amendments is to employ a supplement that will result in a carbon : nitrogen : phosphorous ratio of 100 : 10 : 1. During the remediation of contaminated soils, the nutrients may be present, but not bioavailable or in a preferred form. In these situations, the nitrogen : phosphorous ratio is changed to 10 : 5 or 10 : 10, depending on the soil characterization results.[3] For both groundwater and soil remediation, the introduction of supplementary nutrients is limited to $3\,lb/yd^3$ soil to prevent osmotic shock.[4] In other words, if too many salts are added at the same time, the change in osmotic pressure can rupture the cell wall. Terminal electron acceptor and nutrient amendment deliveries are staggered to prevent fouling of the injection well and the occurrence of localized population growth.

Moisture and Other Environmental Parameters

Moisture is also required for microbial activity as the primary component in bacteria protoplasm is water. If the bacteria do not receive enough water, they will die. However, oversaturation conditions can be detrimental. Too much water can inhibit gas exchanges, thereby depleting the airborne oxygen and rendering an anaerobic zone.[9] Moisture levels are typically maintained by using water as the delivery medium for nutrient amendments and removal of degradation by-products.

The pH level of the surrounding environment can affect microbial growth, cellular functions (fluidity, membrane transport, etc.), equilibrium of enzymatic reactions, and availability of nutrients. Most bacteria can exist in pH ranges between 5 and 9, with an optimal growth around 7. Maintaining the soil pH near 6.5 has the added advantage of preventing the solubilization of heavy metals that are often toxic to bacteria. However, natural bacterial adaptations have enabled successful bioremediation for soils outside the optimal pH range.[2]

The surrounding temperature will also have a great impact on microbial activity. Most of the microorganisms identified for the bioremediation of petroleum contaminants are mesophilic (i.e., optimal growth temperature of 30°C). Although microbes have an optimal growth temperature, they may be able to thrive over a 10°C or a 40°C range, depending on whether they are stenothermal or eurythermal, respectively. The cation exchange capacity and soil type (organic matter, clay content, etc.) will also affect achievable bioremediation efficiencies. The impact of each of these parameters, optimal levels, and examples of potential methods to alter subsurface conditions can be found in several standard textbooks.[2–5,9]

Compound Specific Factors

Compound dependent factors that impact bioremediation efficiencies are the contaminant concentration, adaptation of microbial community, co-oxidation, and cross-acclimation because of substrate analogs.[12] The contaminant concentration is a direct function of the available concentration and solubility, often reported in terms of bioavailability. Depending on the contaminants' inherent physicochemical characteristics, it may be toxic at parts per billion concentrations.[13] In other situations, a contaminant may be considered toxic only when present at high concentrations (>10,000 ppm). The exact impact of concentration is dependent on the micro-organism(s). Typically, biodegradation rates are modeled as first order but will change with time. Rate changes occurring over a long period of time (months and years) are often the result of the time required by the bacterium to alter its metabolism for inducing enzymes essential for that catabolic pathway.[12] Conversely, short time frame changes are often associated with those systems that can be represented as Michaelis–Menton kinetics.[9]

The number of active species as well as the dominant genus in a microbial population can change (i.e., shift) significantly over time. For example, hydrocarbon degraders are typically <1% of the total organisms in clean areas but comprise almost 100% of the active species in a contaminated area.[14] Population changes can be because of nutrient amendments, TEA availability, competitive inhibition within the community, or metabolic stresses posed by the contaminant or degradation by-products.[15] Co-oxidation and cross-acclimation are prevalent at sites with multiple contaminants. Both processes can contribute to changes in the microbial community and are a direct example of why consortiums are advantageous over pure cultures.

Once all of the ''controllable'' factors are in place, it is critical to select the proper bacterial combination(s) that enables the appropriate rate of biomass growth and enzyme production. For instance, the primary step in trichloroethylene bioremediation is the production of oxygenases, such as toluene dioxygenase (tod). However, oxidation of trichloroethylene generates intermediates that inactivate the tod enzyme and can even be toxic to the cells themselves, i.e., catabolite repression.[16,17] Thus, the biomass growth rate and

enzyme regeneration must be high enough to exceed enzyme inactivation and cellular toxicity if the bioremediation process is to be successful.

METABOLIC PATHWAYS

Understanding the metabolic, or degradation, pathways involved in the bioremediation of contaminants can be invaluable. For certain contaminants, only after the responsible pathway(s) is known can mineralization be considered a possibility.[18] Metabolic pathways can provide critical information regarding the:

- Intermediate metabolites' form, concentration, toxicity, and rate limiting steps that may occur.
- Blockage in a reaction sequence (i.e., dead end pathway).
- Potential health concerns.
- Degree of remediation that can be/has been achieved.

Aliphatic and Alicyclic Metabolic Pathways

The aerobic degradation of both nonhalogenated and halogenated alkanes, alkenes, and cycloalkanes has been widely studied.[2,4,7,14,18–21] As such, only an example of the general pathway and summary of key findings is presented. The reaction(s) is initiated by the production of mono- and dioxygenase enzymes. Aliphatic compounds are susceptible to enzymatic attack at either end and will follow the generic pathway depicted in Eq. (4). Further degradation at one end (β-oxidation) or at both ends (ω-oxidation) of the compound results in smaller fatty acids that are easily degradable.[12]

$$\text{aliphatic} \xrightarrow{O_2} \text{alcohol} \rightarrow \text{aldehydes} \rightarrow \text{carboxylic acid} \quad (4)$$

Degradation of the parent compounds and intermediates is predominantly an aerobic process. Most of the contaminated sites typically contain adequate indigenous populations and may, therefore, require only supplemental nutrients or TEA. Other generalizations pertaining to alkane, alkene, and alicyclic degradation include the following:

- Long-chain compounds are easier to degrade.
- Compounds with less than nine carbon atoms typically have to be degraded via cometabolism.
- Degradation rate of saturated compounds is much greater than that of unsaturated compounds.
- Rate of straight chain degradation is much greater than that of branched chain compounds.

Aromatic Metabolic Pathways

Biodegradation of an aromatic compound can be divided into two primary steps: activation of the ring and subsequent ring cleavage. Ring activation incorporates molecular oxygen into the ring structure by either mono– or dioxygenase enzymes. The first primary intermediate from mono-oxygenase reactions is an epoxide followed by the formation of transdihydrodiols.[9] Aromatics degraded by the catalysis of aromatic ring dioxygenases, such as tod, incorporate two oxygen molecules to form dihydroxy intermediates.[12,22] Ring activation can be completed under both aerobic and anaerobic conditions. In general, degradation of halogenated aromatics occurs at a faster rate under anaerobic conditions. However, there are a few dehalogenase-mediated mechanisms that can remove a halogen prior to ring cleavage, thereby making it more susceptible to aerobic degradation.[18,23]

A key step in the anaerobic degradation of nonoxygenated aromatics is the incorporation of oxygen into its structure. This rate limiting step typically occurs by the introduction of a hydroxyl group via hydrolysis reactions. Anaerobic degradation of nonoxygenates can occur; however, the primary intermediate is more thermodynamically stable and therefore more resistant to subsequent degradation. For both oxygenated and nonoxygenated compounds, degradation rates are higher under denitrifying than under methanogenic conditions.

An essential criterion in any metabolic pathway is that the metabolism of the contaminant will eventually result in an intermediate that can enter the tricarboxylic acid (TCA) cycle.[13,24] For the aerobic or anaerobic degradation of aromatic compounds, the pivotal point in ring activation to ensure that this will happen is the formation of either catechol or protocatechuate.[4,24–26] Examples of some of the aromatics that can be aerobically degraded to catechol and protocatechuate are contained in Figs. 2 and 3, respectively. After the formation of the pivotal compound, ring cleavage will occur at either the ortho or the meta position.[24] Typical pathways associated with the ortho cleavage of catechol and protocatechuate are shown in Fig. 4, with muconic acids and mucolactones as the primary end products. The meta-cleavage pathway (Fig. 5) yields pyruvate and acetaldehydes. The intermediates formed by ortho and meta breakage then enter the 3-ketoadipic acid cycle (i.e., Krebs cycle). As shown in Fig. 6, the Krebs cycle results in the formation of either succinate or coenzyme A compounds (acetyl-coenzyme A or succinyl-coenzyme A), which can easily enter the TCA cycle. As a result, most practitioners will track the biodegradation byproducts until either catechol or protocatechuate is formed. Once these intermediates are formed,

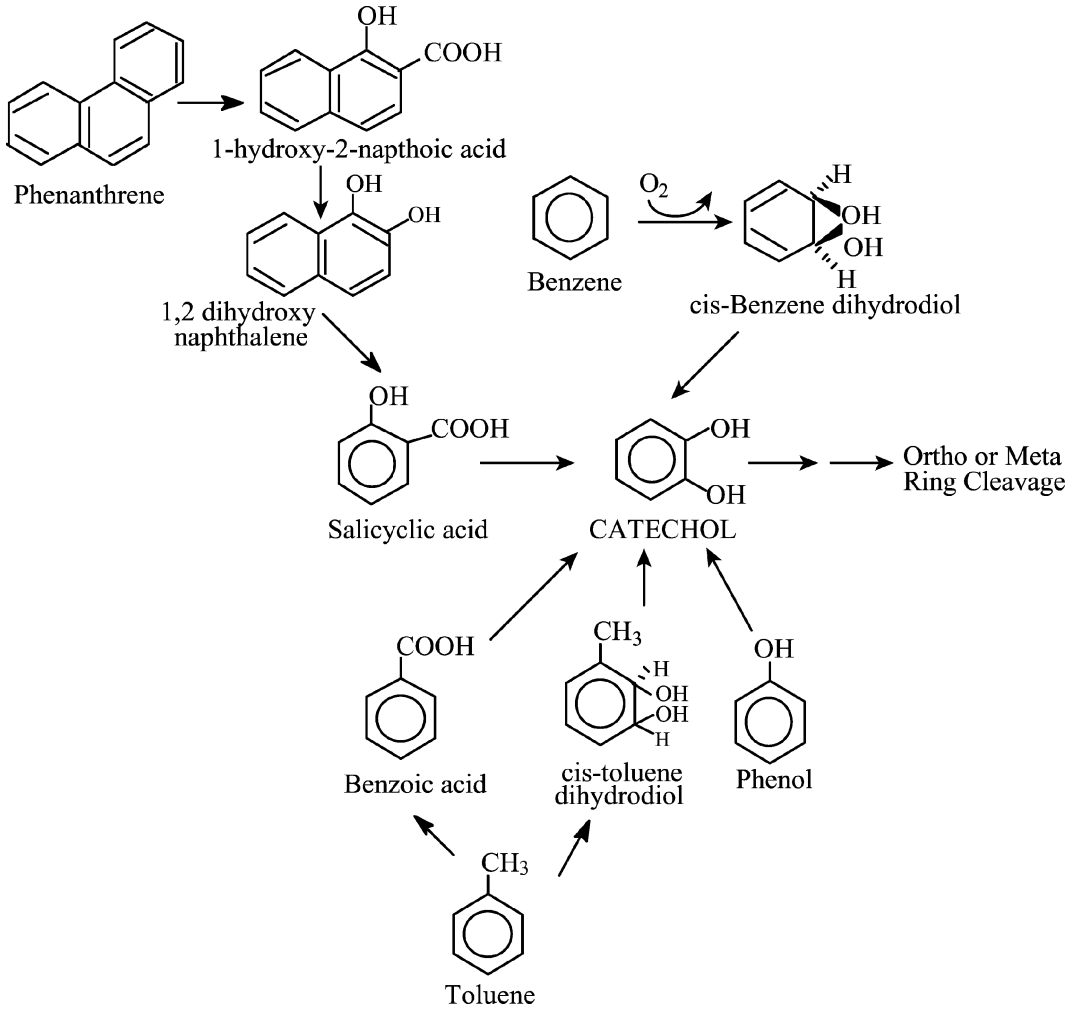


Fig. 2 Representative schematic aerobic degradation of aromatics to catechol.

mineralization is almost a certainty as long as the bacterial activity is maintained by adequate nutrients and TEA.

The degradation of aromatic compounds has demonstrated success under both aerobic and anaerobic conditions. For simple aromatics, such as benzene and toluene, all modes of anaerobic degradation (sulfate reducing, methanogenic, and nitrate respiration) have been successful.[4] However, this is not the case for more complex aromatics such as PAHs and PCBs. For all aromatics, the number, type, and location of substitution will influence the ease of degradation. A few of the other generalizations for multiringed aromatics include the following:

- Polycyclic aromatic hydrocarbons degrade one ring at a time by mechanisms similar to that of monoaromatic compounds.[12,13]
- As the number of rings increases, ease of degradation decreases.[27]
- Cometabolism and analog substrates are required for compounds with 4+ rings.[2,27]
- Initial ring oxidation is the rate limiting step.[4,24,28]
- Degradation decreases with increasing saturation.[2,28,29]
- The effectiveness of aromatic or anaerobic degradation of halogenated compounds is compound bacteria specific (i.e., no universal applications).

Additional information on microbial mediated reactions for a host of xenobiotics can be found at http://umbbd.ahc.umn.edu. The database contains information on some of the key degradation pathways associated with bioremediation.

TREATABILITY STUDIES

The primary objective of a treatability study is to evaluate the overall feasibility of bioremediation. The

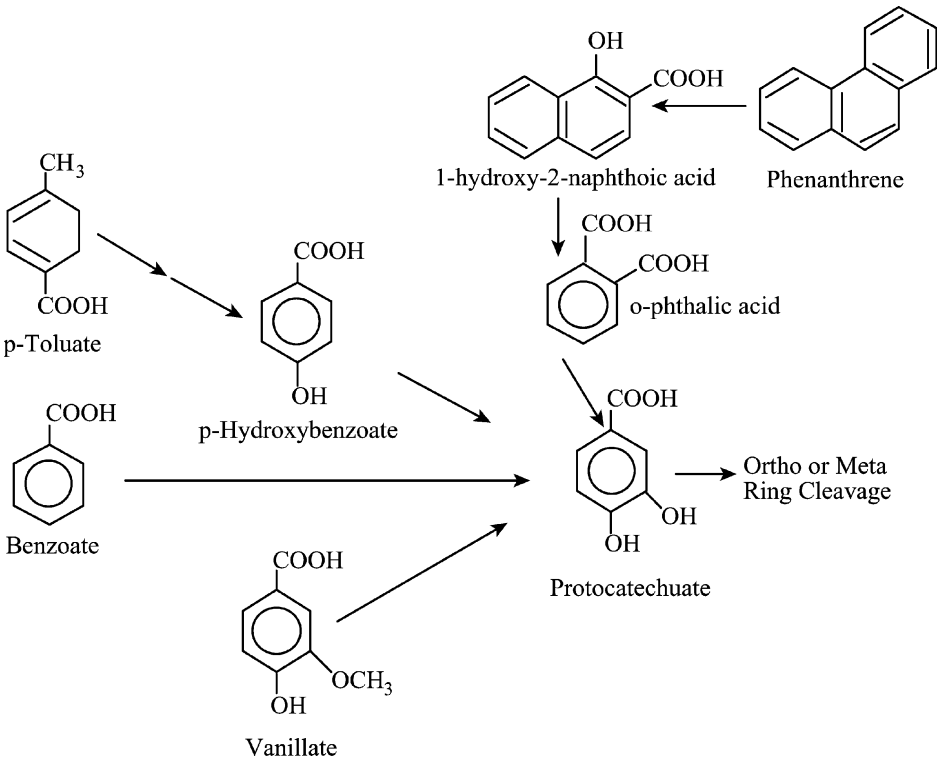


Fig. 3 Representative aerobic degradation of aromatics to protocatechuate.

studies can be conducted either in laboratory setting or in field tests. One of the goals of a treatability study is to characterize the microbial population that can ascertain the critical metabolic pathways. It can also determine if a change in respiration mode (TEA) is needed for complex, highly substituted compounds (e.g., PCBs). The second goal is to ascertain biodegradation potential. The potential can be as simple as whether bioremediation will or will not occur. Studies to evaluate potential can also

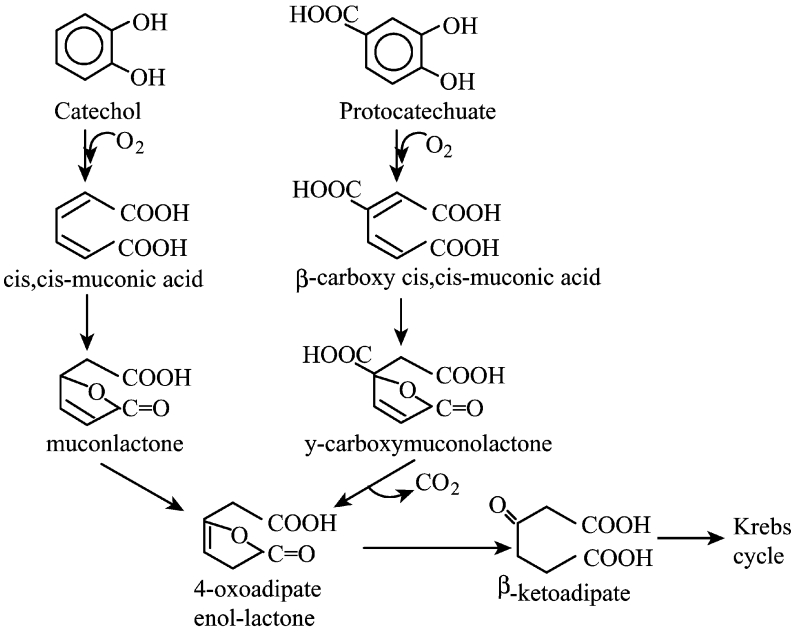


Fig. 4 Ortho ring cleavage of catechol and protocatechuate.

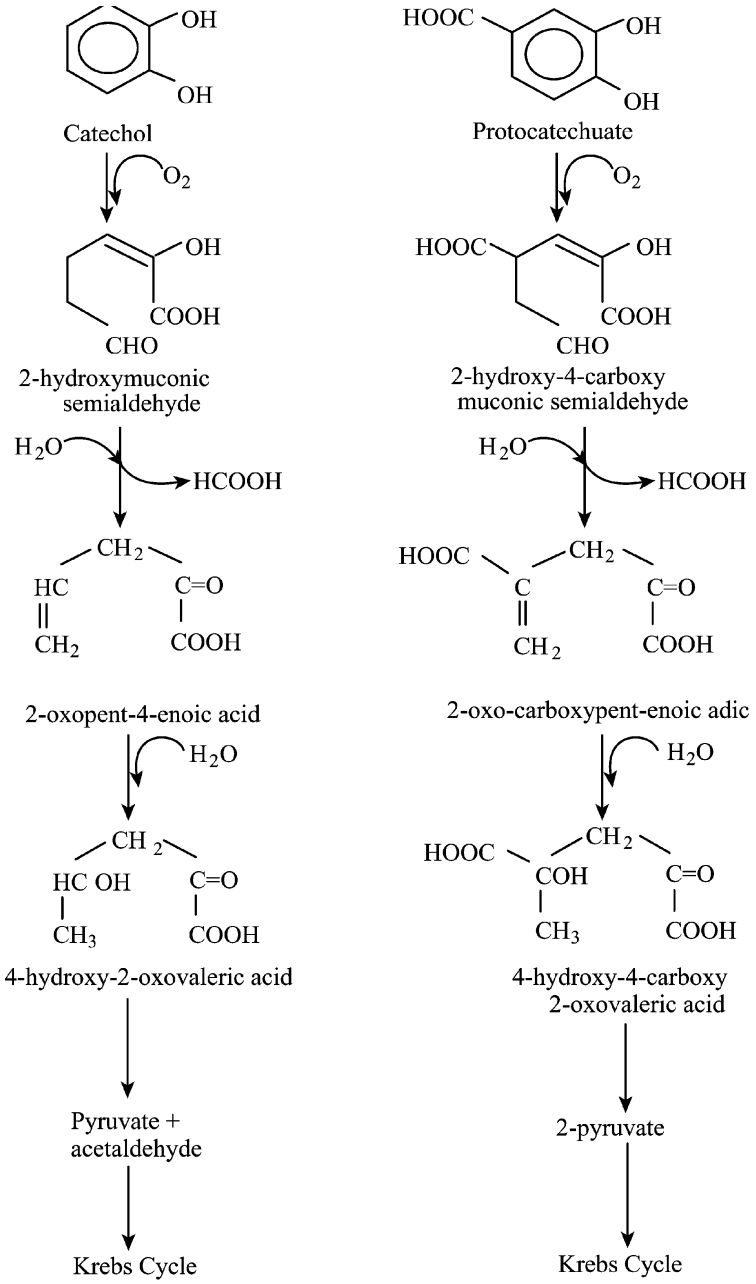


Fig. 5 Meta ring cleavage of catechol and protocatechuate.

be more rigorous to determine the actual biodegradation rates, achievable threshold levels, required time, etc. The third key application of treatability studies is for process optimization. These tests can determine the source, amount, and dosage frequency of the supplemental nutrients and/or TEA. It may also assess if the site will require bioaugmentation or if biostimulation suffices. In evaluating bioaugmentation approaches, a treatability study can test novel designs (i.e., designer microbes for specific enzyme induction) at cost-effective, safe environments. Last, a treatability study can assist with the monitoring of ongoing projects. For instance, the results can identify what key degradation by-products need to be tracked to prove that the treatment is proceeding as expected.

LONG-TERM CHALLENGES

Although bioremediation has been very successful over the past decades, it is still faced with several long-term challenges. The first one facing researchers and practitioners is the accurate, quantitative determination of in situ degradation rates. Almost all of the published degradation rates have been based on laboratory

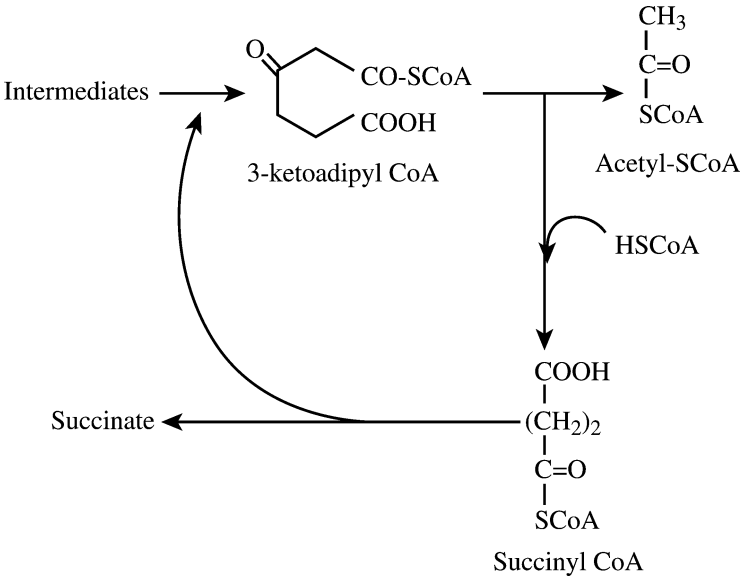


Fig. 6 Representative Krebs cycle.

studies that do not reflect subsurface environments. In fact, a 15–20% loss in efficiency is expected to occur when going from the laboratory to the field scale.[4] The lack of information pertaining to ISB rates is confounded by the fact that efficiencies of ISB treatments are typically site specific, even when the same bacteria consortium–contaminant combination was in place. This is largely because of the nonheterogeneity of soils leading to mass transfer limitations, which interferes with remedial and monitoring activities. Current rate assessment techniques implement either ex situ methods (i.e., laboratory scale methods) or in situ methods with their own inherent limitations (in situ respiration, tracer studies, etc.).[30]

Assessing in situ rates is directly related to the issues surrounding bioavailability (sorption to soils). Even at highly contaminated sites, pollutant transfer and availability may be limited, thereby decreasing achievable degradation rates. This has led to in situ bioremediation being incorrectly classified as ''ineffective'' because of the required time for treatment. One approach to decrease treatment time is to increase contaminant bioavailability. The introduction of commercial surfactants, such as Tween 80®, or utilization of biosurfactants is one method of enhancing solubility/bioavailability. For example, *Pseudomonas auerigenosa* is prevalent in subsurface soils and will produce rhamnolipid surfactants when in the stationary phase. However, increasing a contaminant's bioavailability too much can also be detrimental. For instance, the commercial surfactant may bind to the soil, thereby becoming a contaminant itself. Conversely, the new elevated concentration may be toxic to the degrading organisms. Thus, the challenge with bioavailability pertains to the appropriate selection (synthetic vs. biosurfactant) and application/production of a surfactant to enhance contaminant solubility in a cost-effective, environment-friendly manner without exceeding substrate utilization rates and toxic levels.

Another long-term challenge for ISB is determining how to address sites with multiple contaminants. A large fraction of National Priority List sites still await cleanup because of the problems associated with one of the multiple contaminant classifications inhibiting one or more of the remedial options.[4,5] For instance, most sites contain a mixture of solvents, petroleum products, and heavy metals.[4] At moderate to high concentrations, heavy metals are toxic to microbial activity. One approach to overcome the toxicity would be to stimulate or introduce microbes with natural metal resistance mechanisms. Several strains of the *Sphaerotilus*, *Aspergillus*, and *Pseudomonas* species can oxidize specific metals to less toxic forms (e.g., $Fe^{6+} \rightarrow Fe^{3+}$) or even precipitate the metals by altering the surrounding pH.

In addition to potential metal toxicity, sites with multiple contaminants have difficulty meeting regulatory cleanup levels within acceptable time frames. It is important to note that at these highly contaminated sites, one remedial approach will not be able to address all of the contaminants and media. The key will be to employ remedial treatments that can be easily adapted as a part of the treatment train.

CONCLUSIONS

As shown by the brief examples presented here, bioremediation is extremely versatile. It has demonstrated

success with almost every contaminant classification across all media—soil, water, sediment, and even air. Given adequate time, microbes can adapt to almost any environment. As such, bioremediation is the only other treatment method besides incineration that can result in the complete mineralization of a contaminant. This is particularly advantageous in that mineralization is the end treatment. As scientists and engineers learn more about micro-organisms and their metabolic pathways, the long-term challenges will be overcome and bioremediation will move to the forefront of remediation technologies. For instance, advances in catalysis and microbiology may enable the implementation of bacteria specifically designed to produce key enzymes for the degradation of highly recalcitrant compounds.

REFERENCES

1. Hale, W.H. *Ancient Greece*; American Heritage Inc.: New York, 2001.
2. Norris, R.D.; Hinchee, R.E.; Brown, R.; McCarty, P.L.; Wilson, J.T.; Reinhard, M.; Bouwer, E.J.; Borden, R.C.; Vogel, T.M.; Thomas, J.M.; Ward, C.H. *Handbook of Bioremediation*; CRC Press: Ann Arbor, MI, 1994.
3. Baker, K.H.; Herson, D.S. *Bioremediation*; McGraw Hill: New York, 1994.
4. Cookson, J.T., Jr. *Bioremediation Engineering: Design and Application*; McGraw-Hill: New York, 1995.
5. Rittman, B.E.; McCarty, P.L. *Environmental Biotechnology: Principles and Applications*; McGraw-Hill: New York, 2001.
6. Major, D.W.; Fitchko, J. *Emerging On-Site and In-situ Hazardous Waste Treatment Technologies*; Cahners Publishing: Des Plaines, IL, 1990; 2.19–2.29.
7. Balba, M.T.; Al-Awadhi, N.; Al-Daher, R. Bioremediation of oil-contaminated soil: microbiological methods for feasibility assessment and field evaluation. J. Microbiol. Methods **1998**, *32* (2), 155–164.
8. Riser-Roberts, E. *Remediation of Petroleum Contaminated Soils: Biological, Physical, and Chemical Processes*; Lewis Publishers: New York, 1998; 120–127.
9. Eweis, J.B.; Ergas, S.J.; Chang, D.P.Y.; Schroeder, E.D. *Bioremediation Principles*; McGraw-Hill: New York, 1998.
10. Bachoon, D.S.; Araujo, R.; Molina, M.; Hodson, R. Microbial community dynamics and evaluation of bioremediation strategies in oil-impacted salt marsh sediment microcosms. J. Ind. Microbiol. Biotechnol. **2001**, *27* (2), 72–79.
11. Joo, C.S.; Oh, Y.S.; Chung, W.J. Evaluation of bioremediation effectiveness by resolving rate-limiting parameters in diesel-contaminated soil. J. Microbiol. Biotechnol. **2001**, *11* (4), 607–613.
12. Mitchell, R. *Environmental Microbiology*; John-Wiley and Sons: New York, 1992.
13. Landis, W.G.; Yu, M.H. *Introduction to Environmental Toxicology: Impacts of Chemicals on Ecological Systems*; Lewis Publishers: Boca Raton, 1995.
14. Atlas, RM. Biodegradation of hydrocarbons in the environment. Basic Life Sci. **1986**, *45*, 211–22.
15. Macnaughton, S.J.; Stephen, J.R.; Venosa, A.D.; Davis, G.A.; Chang, Y.J.; White, D.C. Microbial population changes during bioremediation of an oil spill. Appl. Environ. Microbiol. **1999**, *65* (8), 3566–3574.
16. Cox, C.D.; Woo, H.J.; Robinson, K.G. Cometabolic biodegradation of trichloroethylene (TCE) in the gas phase. Water Sci. Technol. **1998**, *37* (8), 97–104.
17. Sun, A.K.; Wood, T.K. Trichloroethylene degradation and mineralization by pseudomonads and *Methylosinus trichosporium* OB3b. Appl. Microbiol. Biotechnol. **1996**, *45*, 248–256.
18. Webb, M.D.; McCarty, A.J. Microbial transformation of chlorinated phenols. Recent Res. Dev. Microbiol. **2001**, *5* (1), 261–281.
19. Bartha, R. Biotechnology of petroleum pollutant degradation. Microb. Ecol. **1986**, *12*, 155–172.
20. Pitter, P.; Chudoba, J. *Biodegradability of Organic Substances in the Aquatic Environment*; CRC Press: Boca Raton, 1990.
21. Vogel, T.M.; McCarty, P.L.; Criddle, C.S. Transformations of halogenated aliphatic compounds. Environ. Sci. Technol. **1987**, *21* (8), 722–736.
22. Power, M.; Van der Meer, J.R.; Tchelet, R.; Egli, T.; Eggen, R. Molecular-based methods can contribute to assessments of toxicological risks and bioremediation strategies. J. Microbiol. Methods **1998**, *32*, 107–119.
23. Schlomann, M.; Kaschabek, S.R.; Reineke, W. Pathways of bacterial degradation of chloroaromatic compounds—special aspects of chlorocatechol pathways of *Rhodococcus*. Biol. Abwasserreiningung **1999**, *12*, 29–61.
24. Dagley, S. Catabolism of aromatic compounds by microorganisms. Adv. Microb. Physiol. **1971**, *6*, 1–46.
25. Greenberg, A. Exploration of selected pathways for metabolic oxidative ring opening of benzene

based on estimates of molecular energetics. Struct. Energ. React. Chem. Ser. **1995**, *3*, 401–432.
26. Harwood, C.S.; Burchhardt, G.; Herrmann, H.; Fuchs, G. Anaerobic metabolism of aromatic compounds via the benzoyl-CoA pathway. FEMS Microbiol. Rev. **1998**, *22* (5), 439–458.
27. Cutright, T.; Lee, S. Bioremediation of PAH contaminated soil: microorganisms and metabolic pathways. Fresenius Environ. Bull. **1994**, *3* (7), 413–421.
28. Lee, J.Y.; Jung, K.H.; Choi, S.H.; Kim, H.S. Combination of the tod and tol pathways in redesigning a metabolic route of *Pseudomonas putida* for the mineralization of benzene, toluene, and p-xylene mixture. Appl. Environ. Microbiol. **1995**, *61* (6), 2211–2217.
29. Renganathan, V.; Johnston, J.B. Catechols of novel substrates producing the toluene ring oxidation pathway of *Pseudomonas sp.* strain T-12. Appl. Microbiol. Biotechnol. **1989**, *31* (4), 419–424.
30. Korda, A.; Santas, P.; Tenente, A.; Santas, R. Petroleum hydrocarbon bioremediation: sampling and analytical techniques, in-situ treatments, and commercial microorganisms currently used. Appl. Microbiol. Biotechnol. **1997**, *48* (3), 677–686.

Bioseparations

Shubhayu Basu
Shang-Tian Yang
Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio, U.S.A.

INTRODUCTION

A series of downstream processing and separation steps have long been used in food, chemical, pharmaceutical, and biotechnology industries in the recovery and production of biologically derived products, including enzymes, antibiotics, polysaccharides, organic acids, amino acids, vitamins, and alcohols. This sequence of recovery steps constitutes the field of bioseparation, which applies fundamental engineering principles and biochemistry to the design of new, or modification of, conventional chemical engineering unit operation processes. Since the invention of recombinant DNA technology and the birth of modern biotechnology industry in the late 1970s, bioseparation has faced a much greater challenge in dealing with more complicated recombinant protein products that require an extremely high purity to satisfy the stringent Food and Drug Administration (FDA) regulation for therapeutic and other clinical applications. Today, bioseparation is a significant and separate discipline in itself.

Although there are a wide variety of bioproducts that need to be separated and purified after fermentation, this entry focuses on the downstream processing of recombinant protein products, which are taking the pharmaceutical world by storm. Currently, a plethora of recombinant human proteins with immense therapeutic value are being produced in fermentation by various host organisms, including bacteria, fungi, and animal cells. Depending on the protein products and the host organisms, different bioseparation processes need to be designed to suit their specific needs. Here, a general bioseparation process design flowsheet is discussed first, followed by two case studies to help the reader understand better how each method can be integrated to obtain a product. These studies outline the design of the downstream processing units for obtaining recombinant human glial cell line derived neurotrophic factor (rhGDNF), which is produced commercially as an inclusion body (IB) in *Escherichia coli*, and erythropoeitin (EPO), a secreted protein from an engineered Chinese hamster ovary (CHO) cell line. Companies producing these recombinant protein products do not release their actual patented processes. However, the two flowsheets presented here are simulated based on the general process design principles to closely resemble the actual processes. The two case studies should help the reader in understanding the difference in strategy in purifying intracellular and extracellular protein products.

CLASSIFICATION OF BIOPRODUCTS

Bioproducts generally can be classified on the basis of their physical, chemical, and biological properties, which together with their concentrations present in the fermentation broth usually dictate how they can be separated and purified. Table 1 shows some industrial bioproducts with their concentrations produced from fermentation and typical downstream processing methods used for their separation and purification. Small molecules such as sugars, amino acids, vitamins, alcohols, and organic acids have sizes less than 2 nm and molecular weights (MW) less than 1000 Da. Large ones such as proteins and polysaccharides can be as big as 3–20 nm and have MW ranging between 10^3 and 10^7 Da. These small and large molecules are soluble in water and are usually present in the liquid phase during the fermentation. On the other hand, ribosomes, viruses, bacterial cells, yeast cells, and animal cells are much larger (18 nm to 10 μm) and are usually found as particulates or suspended solids in the fermentation broth.

The choice of bioseparation methods depends on the type of bioproduct, but the sequence of unit operations can often be dictated by the location of the products with respect to the cells. Whole cells like single-cell proteins, bacterial, yeast, and animal cells are relatively easy to separate and purify from the fermentation broth, and generally require only a few simple downstream processing steps. Extracellular products like some secreted enzymes, antibiotics, organic acids, and alcohols are slightly more complex to isolate and purify from the fermentation broth, but their relatively high concentrations and good stability render their recovery and purification easier and amenable for large-scale industrial operation. However, recombinant proteins for therapeutic use and other

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016151

Table 1 Some bioproducts produced by fermentation and downstream processing steps used in their recovery and purification

Product	Concentration (g/L)	Steps used in downstream processing	Reference
Ethanol	70–120	Stripping, distillation	[1]
Citric acid	50–100	Precipitation/solvent extraction, crystallization	[2]
Antibiotics (e.g., penicillin G)	10–30	Filtration, solvent extraction, crystallization, centrifugation, drying	[1,2]
Amino acids	1–100	Filtration, precipitation, crystallization, membrane filtration	[3]
Single-cell protein	30–50	Flocculation, filtration, centrifugation, drying	[1]
Xanthan gum	25–50	Alcohol precipitation, centrifugation, drying	
Extracellular enzymes	2–5	Precipitation, adsorption, chromatography sequence	[1]
Monoclonal antibody	0.01–1	Precipitation/centrifugation, chromatography, diafiltration, ultrafiltration	[4]
Tumor plasminogen activator	0.045	Mainly a sequence of chromatography steps	[5]
Vitamin B_{12}	0.02–0.06	Flocculation, filtration, drying, adsorption, elution, two-phase separation, crystallization	[1]

clinical applications usually are produced in a relatively low concentration but must be purified to a near-100% purity. Therefore, they require the most cumbersome purification steps and are usually difficult and costly to produce. The degree of difficulty in recovering the recombinant proteins also depends on whether they are present in IBs, or as extracellular or intracellular proteins.

BASIC BIOSEPARATION STEPS

The steps in purifying any bioproduct are governed by its properties, cellular location, initial broth characteristics (viscosity, concentration, impurities) and final product concentration, and the physical form desired. The unit operations also differ from their counterparts in the chemical industry because most biological molecules are temperature and pH sensitive, and are produced in low concentrations.

Some bioproducts are derived from fermentation by living organisms but immobilized enzymes can also catalyze a biotransformation. If the product is extracellular, then the initial steps of recovery include removal of the cells and other particulate matters from the broth. If the product is intracellular, it would be necessary to lyse the cells to release the product into the broth. The cell debris is then separated before the product is recovered from the broth. In certain cases, proteins produced as IBs need to be solubilized and the proteins renatured before further recovery steps. In the case of biotransformed products, the immobilized cells or enzymes and their support need to be removed initially.

Fig. 1 illustrates the various key steps required in downstream processing and the characteristic pathways used for recovery of cellular, extracellular, and intracellular products. Their function in downstream processing can be summarized as follows:

1. *Separation of cells and other particulates*: Filtration, sedimentation, centrifugation, and decanting are common methods.
2. *Cell disintegration*: In the case of intracellular products, cells need to be lysed by high-pressure homogenization, by wet milling, or by enzymes, detergents, and other chemicals that disintegrate the cell wall and membrane.
3. *Clarification*: With or without cell lysis, the fermentation broth consists of a mixture of components from which the final product needs to be clarified. Primary steps in clarification usually involve centrifugation and filtration to remove cell debris and other particulates. The product purity in the outlet stream is usually still low, between 1% and 10% (w/v), at the end of these steps.
4. *Enrichment*: More sophisticated separation techniques such as solvent extraction, precipitation, adsorption, and ultrafiltration (UF) are commonly used to concentrate and partially purify the product. The purity of the product usually reaches 10–80% (w/v) at the end of these steps.
5. *Protein refolding*: For recombinant proteins present as IBs in bacterial cells, they need to be solubilized first and then properly refolded to renature the proteins to regain their biological activities.
6. *Purification*: These operations are usually high-resolution techniques that are often difficult to scale up beyond the laboratory or pilot scale.

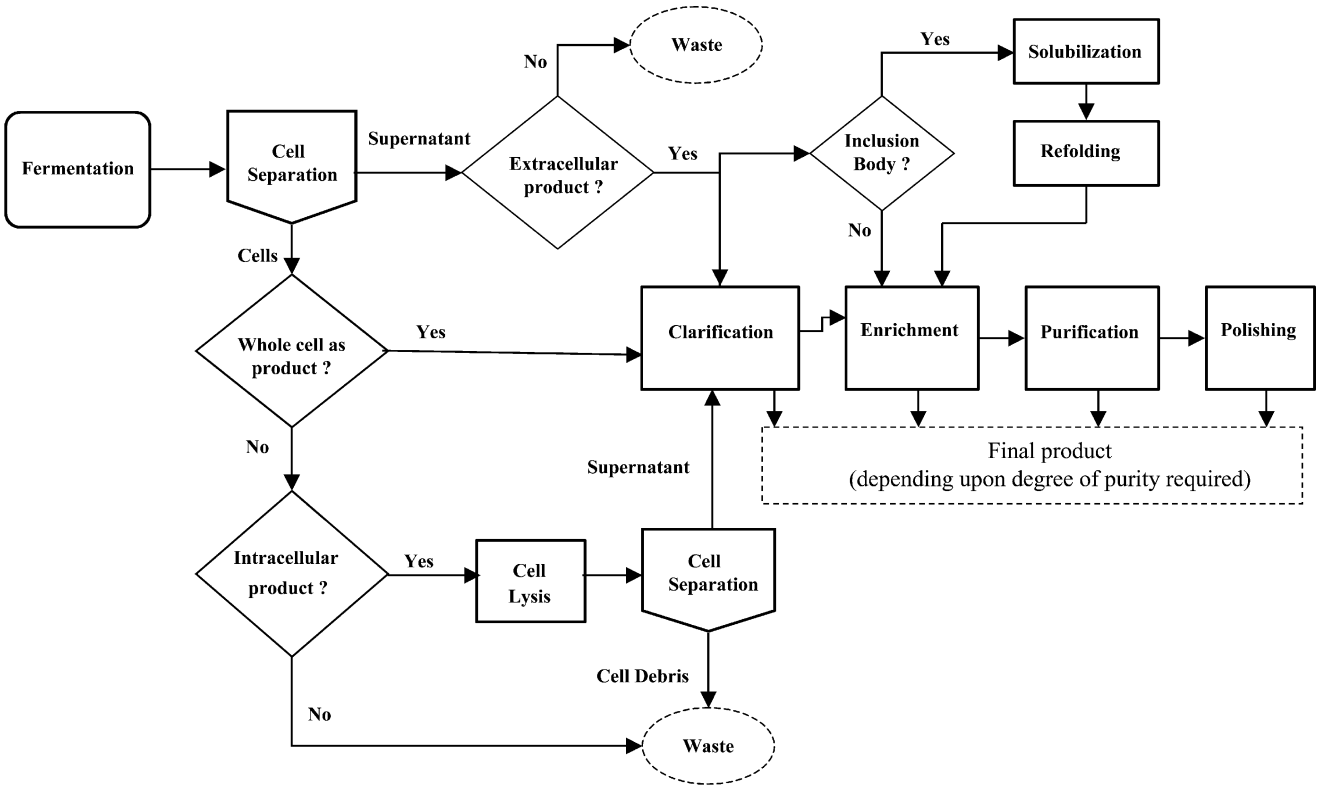


Fig. 1 A general process flowsheet for bioseparations.

However, various kinds of chromatography techniques utilize various physicochemical properties of the desired product to purify it. Product purity is often greater than 90–99% (w/v) at the end of a series of chromatography steps.

7. *Polishing*: The last steps provide the product at the desired concentration, purity, and physical form. These steps may consist of sterile filtration, diafiltration, UF, crystallization, freeze drying, lyophilization, and spray drying.

It is important to note that the design of a downstream process depends on the intended use of the product, the final concentration and purity needed, its marketable price, and the speed of recovery required because of the labile nature of most biological substances.

CELL SEPARATION

Common methods to separate cells and other particulates from the fermentation broth are sedimentation, centrifugation, and filtration. Sedimentation and centrifugation utilize the difference in density of the particle and its surrounding fluid to separate particulates from the fermentation broth. Sedimentation is settling in a gravitational field, whereas centrifugation uses centrifugal forces to amplify the settling velocity.

Sedimentation

Sometimes cells have a normal tendency to flocculate or aggregate based on charge differences. This allows for a simple and cost-effective cell recovery system. Although numerous studies have been done on the principles that govern sedimentation, it is worth noting that when settling occurs the particle has no net acceleration. The opposing drag and buoyancy forces negate gravity so that the particle settles at a velocity known as the terminal velocity, v_g, which is a function of the particle diameter, the density difference between the particle and the fluid, and the viscosity of the fluid. Sedimentation is often used in large-scale wastewater treatment processes as well as in traditional biosynthetic pathways involving fermentation. However, settling gives rise to a sludge that still contains ample amount of the fluid. Settled sludge biomass might need to be further dewatered using centrifugation.

Centrifugation

In the process of centrifugation, the applied centrifugal force accelerates the settling velocity of a particle. The settling velocity, v_ω, is additionally dependent on the angular velocity and radius besides the parameters governing v_g. The empirical equations governing the

settling velocity vary at different Reynolds numbers (Re) and are given elsewhere.[6] In brief, an Re below 0.4 is defined as the Stokes' regime, whereas that greater than 500 is defined as the Newtonian regime. Allen regime applies for all Re in between these values.

Although most protein settling occurs in the Stokes' regime, where the centrifugation efficiency is purely dependent on the particle size, shape, and density, things become a little more complicated in the Allen and Newtonian regimes, where high particle velocities, wall effects, and hindered settling occur. Hindered settling velocities are correlated to v_ω and are usually dependent on the particle volume fraction, shape, and size.[7]

Various types of centrifuges exist and choosing the correct one to achieve a certain separation efficiency is important. Table 2 summarizes the types, applications and advantages.

The choice of the correct centrifuge is always difficult, but keeping the parameters outlined in Table 2 in mind, one can design the downstream operation in an efficient way. It is important to remember that a decanter or a scroll-type centrifuge is less efficient for separation of micro-organisms, but it can handle large solid particles. Hence, this type can be used in series before another centrifuge to prevent clogging in the latter. In general, tubular bowl centrifuges are better for protein precipitates, disk centrifuges for cells and cell debris, and decanter centrifuges for noncellular, large particles such as antibiotic crystals.[8] It is important to sterilize and contain centrifuges as contamination of a bioproduct at any stage of recovery can be disastrous for the process.

Filtration

Filtration can be used to separate cells or cell debris, concentrate cells or protein solution, and remove or exchange salts. Two broad categories of filtration are conventional dead-end filtration, where fluid flow is normal to the plane of the membrane, and cross-flow filtration (or tangential flow filtration), where the fluid flows parallel to the membrane surface.

It is beyond the scope of this entry to review the basic principles governing filtration. However, it is interesting to note that filtration produces a more concentrated and dewatered cell sludge (20–35% w/v) or cell solids ($\geq$40% w/v) than settling. A variety of filter media, membranes, and equipment are commercially available. In the case where the deposited cake is compressible with low permeability and thereby adds more resistance to filtration, filter aids or precoats often alleviate the problem. Two of the most widely used filter aids are diatomaceous earth and perlite.

Various types of filtration equipment are available commercially and can be operated in batch, semicontinuous, or continuous modes. Among the commonly used types are the plate and frame filter, rotary drum filter, leaf filter, plate filter, and tray filter. Apart from the plate and tray filters, all other are enclosed and therefore are easy to work with when sterility of the solids is an important issue. Moreover, all these filters are examples of dead-end filters. Cross-flow filtration is mostly used in the purification stage through membranes with very low pore sizes and is discussed later.

CELL LYSIS

To obtain intracellular products, the cells need to be disintegrated and the cell debris discarded. The product then has to be purified from the supernatant. Cell lysis can be achieved by different methods, and the level of harshness required depends on the rigidity of the cell wall and hence the type of cell. Sometimes products are stored in the periplasmic space of the cells and complete cell lysis is not necessary. Prokaryotes, fungi, and plant cells have rigid cell walls, while animal cells do not. This makes disruption of the former category more difficult than in animal cells. Ironically though, most animal cells are efficient in secreting the product they synthesize.

Cell lysis methods can be broadly classified into mechanical methods using homogenizers, bead mills, and sonicators, and chemical methods involving osmotic shock, enzymes, and detergents. As the name suggests, mechanical methods rupture the cell membrane by shear forces or compression forces generated in homogenizer valves or by the grinding between glass spheres in bead mills, while sonicators rupture cell membranes using ultrasound waves. Nonmechanical methods can use osmotic shock, which can burst cells without a cell wall. A rapid decrease in the extracellular solute concentration can lyse cells. Enzymes like lysozyme can digest the cell wall of prokaryotes and plant cells, and make the resulting protoplasts susceptible to rupture by osmotic shock. Detergents and solvents can also be used to dissolve cell membranes. However, it is important to remember that addition of any chemical to lyse cells also warrants its separation from the broth to purify the desired product. In such a case where the presence of detergents or other chemicals can destabilize proteins, mechanical methods are more advantageous.

CLARIFICATION

Primary clarification steps usually employ centrifugation or filtration to partially purify the products from

Table 2 Comparsion of various types of centrifuges

Type	Mode of operation	Centrifugal force (*g*)	Particle size (μm)	Maximum throughput (m^3/hr)	Advantages	Disadvantages	Applications
Tubular bowl	Batch or semicontinuous	20,000	0.1–100	0.01–5	1. High centrifugal force 2. Good dewatering 3. Easy cleaning 4. Simple operation	1. Limited solid capacity 2. Foaming 3. Solid recovery is cumbersome	Pilot-scale separation of animal, plant, and microbial cells
Multiple-chamber bowl	Batch	5,000–9,000	0.5–20,000	150	1. Large solid boiling capacity 2. Good dewatering 3. Bowl cooling possible	1. Cleaning more difficult than for tubular bowl 2. Solid recovery difficult	Separation of human blood plasma components
Basic disk type	Batch or semicontinuous	5,000–11,000	0.5–500	30–60	1. Big discharge under pressure eliminates framing 2. Bowl cooling possible	1. Poor dewatering 2. Difficult to clean	Production of glucose isomerase from *Bacillus coagulans*
Disk with nozzle discharge	Continuous	6,000–9,000	0.5–500	300	1. Column solid discharge 2. Bowl cooling possible	1. Poor dewatering 2. Difficult to clean	Production of baker's yeast
Basket	Batch	400–1,300	10–1,000	—	1. Good dewatering 2. Large solid holding capacity	1. Low centrifugal force 2. Recovery difficult 3. Cannot separate fine particles	Removal of adsorbents such as cellulose and agarose
Scroll/decanter	Continuous	—	5–50,000	200	1. Continuous solid discharge 2. High-feed solid concentration	1. Turbulence created by scroll 2. Low centrifugal force	Antibiotic production

the fermentation broth. The operating principles for both unit operations are exactly the same as discussed earlier. However, filtration processes involved in this step are more refined and usually perform separation on a much smaller scale. As discussed before, there are two modes of operating a filtration unit: dead-end filtration and cross-flow filtration. While dead-end filtration is typically used for removal of macrosized particles (50–1000 μm), cross-flow filtration membranes can have different pore sizes, which dictate the type of filtration being performed.

Microfiltration (MF) is usually used to separate micrometer-sized particles (between 0.1 and 20 μm) from the fluid. Ultrafiltration is used to separate macromolecules or polymers such as proteins and polysaccharides with MW much greater than 1000, nanofiltration (NF) for oligomers with MW between 100 and 1000, and reverse osmosis (RO) for small solutes such as salts from water. The UF, NF, and RO membranes are semipermeable and are the finest forms of filtration presently known. Their separation mechanism is mainly based on the pore size of the membrane as characterized by their MW cutoff. Reverse osmosis membranes have the smallest pores and usually allow only the passage of water through the membrane. This process requires a high pressure to overcome the osmotic pressure, which could be several hundred atmospheric pressures under a high salt concentration. Ultrafiltration membranes are comparatively the largest in pore size and require a relatively low pressure as the driving force because there is negligible osmotic pressure and the pressure difference across the membrane is low. This process is useful for separating proteins from smaller molecules and is an important primary clarification technique as it helps pretreat and concentrate the broth before other forms of purification such as chromatography. It is also a vital procedure in desalinating the protein and, if used, is done as a purification step. The dilution mode filtration or diafiltration is usually placed after UF units to remove salts from the concentrated protein solution. Diafiltration is done by keeping a constant volume upstream of the membrane. Water is added at the same rate as the solution is being removed through the membranes. This decreases the salt concentration in the protein solution. Fig. 2 illustrates the operations of UF and diafiltration and how they change salt and protein concentrations in the feed solution.

A simple centrifugation or MF step in the primary clarification may directly be succeeded by a secondary clarification or an enrichment step, such as extraction, precipitation, or adsorption. The concentrated product may then be subjected to membrane filtration processes. Ultrafiltration might be done earlier and then followed by extraction or precipitation with salts. Diafiltration units can subsequently be used to remove the salts. This flexibility in rearranging various unit operation steps is common and varies between products.

ENRICHMENT

Extraction

Extraction is a process that utilizes the difference in solubility of the product of interest between two immiscible phases. Liquid–liquid extraction is commonly used in classical biotechnological downstream processing for production of antibiotics and other small molecules. The second liquid phase is usually a water immiscible organic solvent. Reciprocating plate, Podbielniak centrifugal extractors, Delaval contactors, and Westfalia extraction-decanters are examples of some of the equipment available for carrying out extraction.

For the extraction of proteins, aqueous two-phase systems (ATPS) are preferred over organic solvents, which usually denature the proteins and render them biologically inactive. They consist of polyethylene glycol (PEG), and a salt (e.g., potassium phosphate) or dextran in water. At concentrations above a critical value, the mixture separates into two phases—one rich in PEG and the other in dextran or salt. In industrial systems, salts are more commonly used because they are relatively inexpensive as compared to dextran. The MW, charge and surface properties of the protein decide how the protein partitions in the system. The nature of the phase components, the MW of the polymer, and the concentration and type of salt used also affect the distribution.[9]

The partition coefficient of proteins can be increased selectively by chemically attaching affinity ligands to the hydroxyl groups of the PEG. The protein can then be back-extracted from this phase by adding a fresh salt phase under conditions where the protein prefers the salt phase or by immobilizing the ligand complex in Sepharose beads, which separate into the top phase.[10,11] Liquid–liquid extraction is commonly used in the industry because of its low cost and easy scale-up potential, but it is rarely used for protein separation even though some proteins can be effectively partitioned in ATPS.

Precipitation

Precipitation is a simple yet large-scale method of purifying proteins and polysaccharide bioproducts. Except for therapeutic recombinant proteins, most proteins such as industrial enzymes can usually be efficiently recovered and concentrated using precipitation techniques. Precipitation utilizes the physical interaction

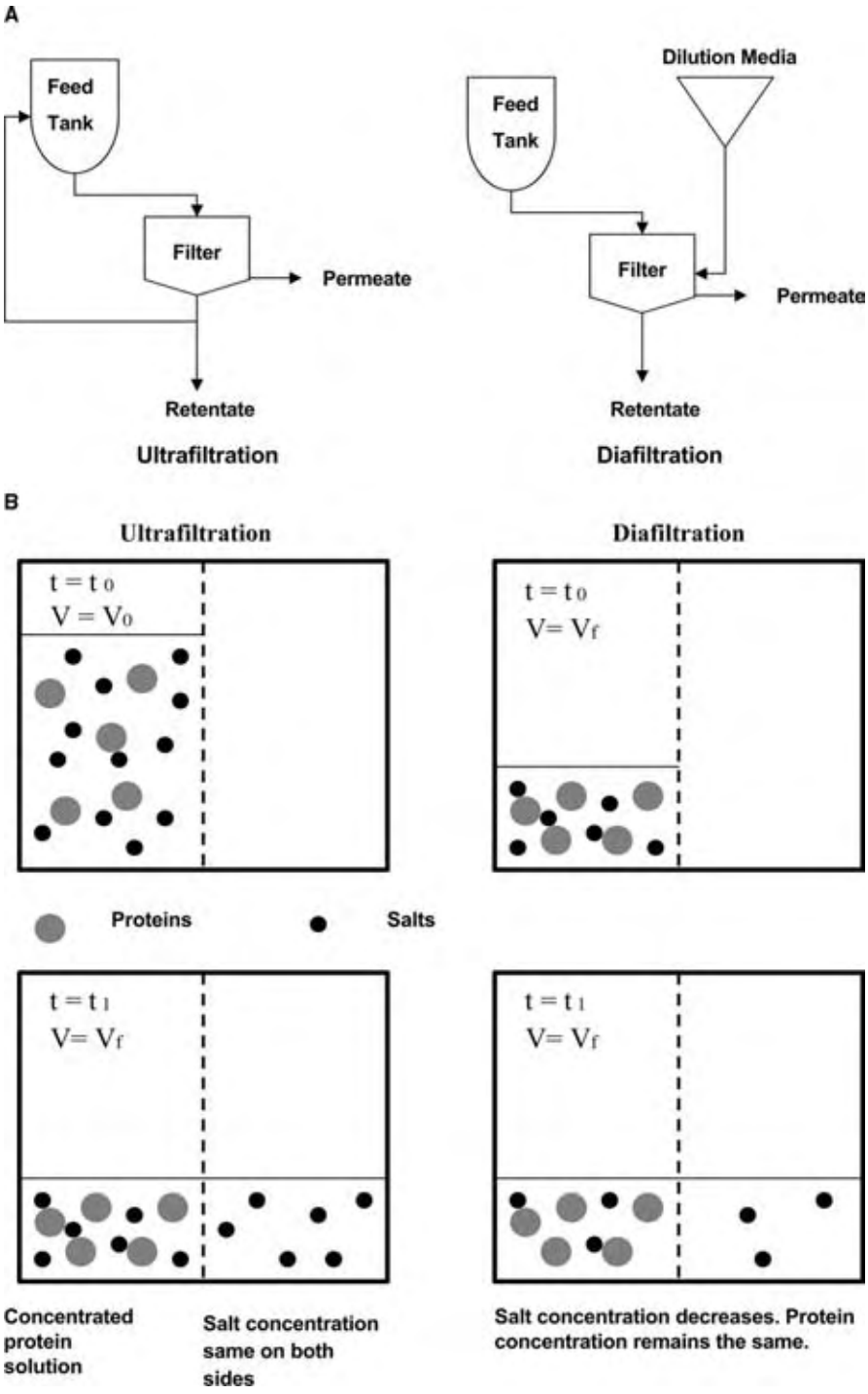


Fig. 2 (A) Schematic of the flow streams involved in UF and diafiltration. (B) Schematic of the process governing UF and diafiltration.

forces like hydrophobic and electrostatic forces and solvation effects to bring solutes out of solution. Table 3 outlines the different ways to precipitate proteins.

Protein precipitation using a change in pH is usually successful for concentrated systems. It may not be efficient for low concentrations however. In general, protein solubility rises with increasing salt concentration up to 0.2 M. This process is known as salting in. Under such a condition, a protein in solution can be precipitated out just by diluting the solution. A low temperature increases precipitate yield in such a case and avoids denaturation. For proteins that are thermolabile and yet can be easily renatured, selective precipitation by thermal methods is a very suitable option. This is a classic example of a ''kinetic'' as opposed to an ''equilibrium'' based separation method. As temperature increases, denatured protein selectively

Table 3 Methods of protein precipitation

Method	Example	Theory or principle
Change in pH	pH = p*I*	At pH other than p*I* polar interactions with the solvent or nonionic interactions with the salt keeping the protein in solution; at pH = p*I*, the protein has no net charge and therefore minimum solubility
Heat denaturation	70°C	As one increases the temperature, denatured protein selectively precipitates out
Salting out	Citrate salts, $(NH_4)_2SO_4$	At higher salt concentrations the water of hydration near the hydrophobic patches on proteins is used for solvating the ions; this exposes the hydrophobic regions of two or more protein molecules, which spontaneously aggregate together and precipitate out of solution
Addition of organic solvent	Acetone, ethanol	Hydrophobic patches are solvated by the solvent but the charged regions interact via electrostatic forces, causing precipitation
Addition of nonionic polymer	PEG	Very efficient for low-MW proteins like globulins; it precipitates protein by effectively reducing the amount of water required for protein solvation
Addition of ionic polymer	Polyacrylic acid, polyethyleneimine	Opposite charge on polymer and proteins cause complex formation

precipitates out at a rate proportional to the concentration of the denatured protein.

Salting in is dependent on surface charges and the polar interactions between the proteins and the solvent. In contrast, salting out is greatly dependent on hydrophobic forces. The addition of salt needs water molecules to solvate the ions. While the free water is used at first, higher salt concentrations use the ordered water molecules near the hydrophobic patches on proteins. This exposes the hydrophobic regions, which spontaneously interact and cause the proteins to form aggregates and precipitate. Most proteins precipitate out of solution over a narrow range of salt concentrations. Ammonium sulfate is widely used for precipitating proteins because it has a high efficacy of precipitation, high solubility in water ($\sim$533 g/L at 20°C), low cost, and a low density as compared to protein aggregates, which helps in protein recovery. Potassium salts have a lower solubility and higher density than ammonium salts in general, and are therefore less desirable than ammonium salts. Citrate salts are very effective in precipitation only above a pH of 7.0. Below that, they cannot be used because of their buffering action.

Organic solvents reduce the dielectric constant of the medium and solvate the hydrophobic regions of the protein molecule. Hence, the charged regions are exposed and interact to precipitate the protein. The mechanism is in a way similar to that of salting out. The solvent should neither interact directly with the protein nor denature it. Also, solvent requirement increases considerably at high ionic strengths.

Polyethylene glycol is available over a large range of MW (6000–20,000) and is widely used in precipitating proteins. Most water-soluble polymers lead to large increases in viscosity at concentrations >20% (w/v). Polyethylene glycol is an exception and is useful for precipitation of low-solubility proteins like globulins. With higher-solubility proteins like albumins, the difference in density between the aggregate and solution is negligible. This makes centrifugation impossible. Recovery of protein from the PEG–protein complex after precipitation is, however, very slow using dialysis.

Crystallization differs from precipitation in that it produces a regular lattice of molecules of the product. As biological products can be obtained at very high purity by crystallization, it is also often used as a polishing step. Sometimes, crystallization as a purification step can obviate the need for expensive chromatographic steps. Crystallization can be induced by pH adjustment, lowering ionic strength, or adding seed crystals of the desired protein to the solution.[8]

Adsorption

Adsorption is based on the principle of the differential affinity of dissolved solutes in a mobile liquid phase to a stationary solid phase. Adsorption columns usually have a smaller capacity but higher specificity than extraction. Various adsorbents have been in use, including activated charcoal (usually from vegetable sources), and more recently, customized resins that can have either positive or negative charges on them, or

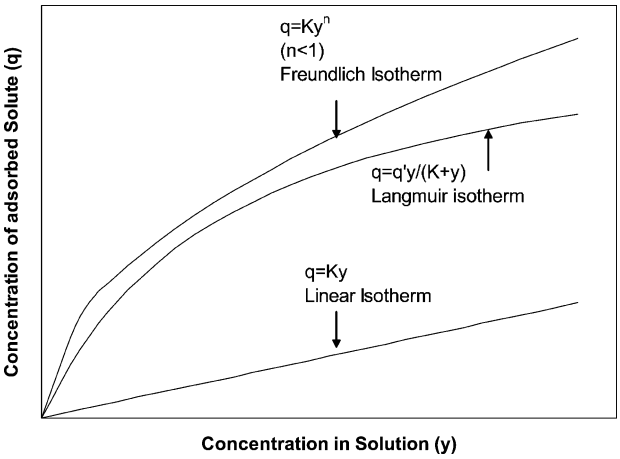


Fig. 3 Three different types of adsorption isotherms.

are hydrophobic in nature. Polystyrene, methacrylate, and acrylate based adsorbents are used for low-MW compounds, cellulose based adsorbents for proteins, poly(styrene-divinylbenzene) resins for nonpolar solutes, and acrylic esters for hydrophobic solutes.[2,12]

Mathematical analysis of adsorption is beyond the scope of this entry, but it is important to point out that the equilibria used are often based on adsorption isotherms. The three isotherms shown in Fig. 3 occur in practice, although linear isotherms are very rare. The nonlinearity of the Langmuir and Freundlich isotherms makes scale-up and analysis tedious. The Freundlich isotherm is based on empirical data analysis and occurs more commonly in practice than the Langmuir isotherm, which has a theoretical basis.

Adsorption is usually done in a batch mode in fixed, agitated, or expanded bed adsorption units. The process usually consists of cyclic operations of adsorption and desorption in two steps. Desorption or recovery of the adsorbed solutes from the adsorbents can be done by either changing the temperature or pH, or using a solvent, which also regenerates the adsorbents for use in the next operation cycle.

PROTEIN REFOLDING

Formation of IBs during protein expression poses a bottleneck in the efficient downstream processing of therapeutic proteins. The reasons for IB formation are not fully known. Because translation is a slower process than protein folding, it is likely that the misfolding of translation intermediates plays some role. Further, since post-translational modifications, such as glycosylation and lyposylation, which are known to affect the secondary structure of proteins, are absent in bacteria, the non-modified protein structure may cause misfolding. The recovery of soluble active protein from purified IBs requires the denaturation of the polypeptide by a denaturing agent such as 8 mol/L urea or 6 mol/L guanidine HCl and its subsequent refolding back to an active form. Fig. 4 depicts the protein configuration during the refolding step. There is no universal method to refold any protein. The conditions needed to properly refold the protein of interest remain an empirical science and need to be optimized for each protein.[13] In general, a solubilized protein solution is usually added into a large volume of a refolding buffer to reduce the concentration of the denaturing agent and also to avoid aggregate formation of protein molecules in the course of renaturation.[14] Refolding can also be done by dialysis against a refolding buffer and buffer exchange by gel chromatography.[15,16] The dilution method is usually the easiest but greatly decreases the concentration of the target protein. Also, the requirement of a large volume and the consequent inefficient mixing in large tanks lead to heterogeneity in folding conditions.[17] Often, the refolding conditions optimized for a protein in the laboratory scale might not work on a large scale.

PURIFICATION

Chromatographic techniques are the sole method for high-resolution purification of proteins and many other bioproducts. As opposed to adsorption, where the feed is continuously passed through the column until breakthrough of the strongly adsorbed components occurs, chromatography separates components in samples that are injected in a pulsatile manner. Elution of the components is achieved either in an isocratic mode, where the same mobile phase is maintained throughout, or by the gradient elution mode. In the former, the separation of components depends on the differences in the retention time of each component in the column. In the latter, the sample is introduced as a pulse and is followed by a different mobile phase containing a displacer.

Conventional batch chromatography is relatively simple and offers operating flexibility, but suffers from several drawbacks as well. To obtain high purity and yield (>99%), a large amount of solvent is needed and column utilization is inefficient.[18,19] To get around these disadvantages, continuous countercurrent chromatographic methods were developed. This mode of operation maximizes the mass transfer, but it is difficult to maintain a stable solid phase velocity.[19,20] An alternative to continuous countercurrent chromatography is to simulate the movement of the solid phase by periodically moving the inlet and outlet ports, while keeping the bed stationary. This mode of operation is called simulated moving bed chromatography and has the advantage of significantly higher

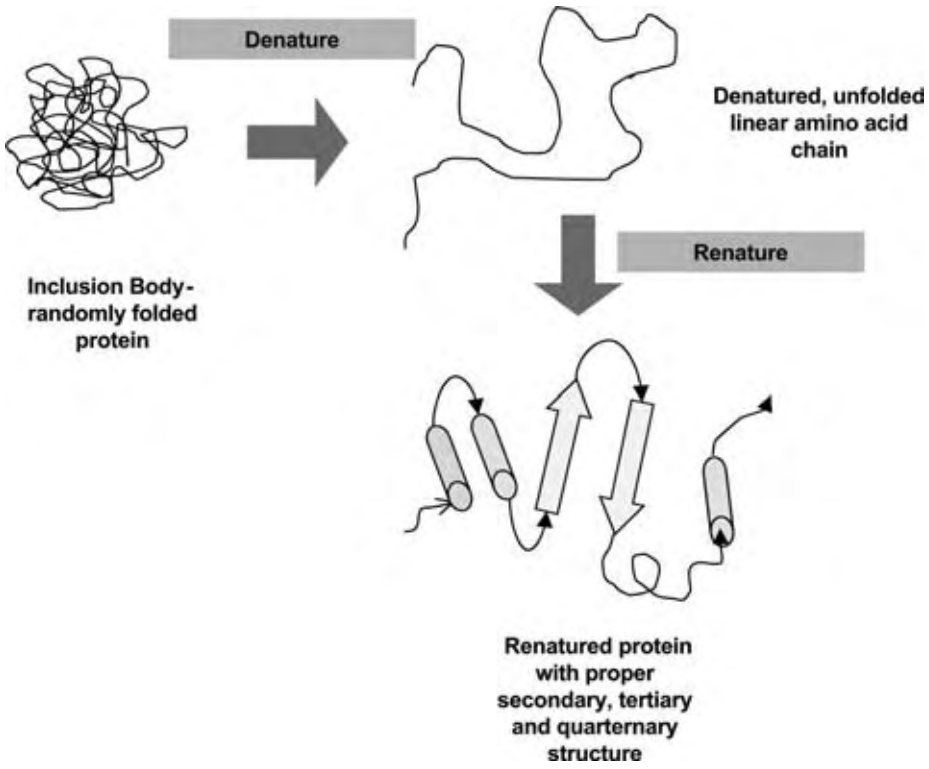


Fig. 4 Schematic depiction of the protein refolding process. *(View this art in color at www.dekker.com.)*

throughput per bed volume and low solvent use.[19] Column switching, chromatography with recycle, and reciprocating size exclusion chromatography in a semi-continuous frontal mode are refinements to conventional methods and can further reduce solvent use, increase peak resolution, and facilitate faster recovery of large molecules.[21–23]

The type of matrix in the column dictates the various kinds of chromatography techniques. All matrix materials need to be inert, chemically and physically stable to withstand the harsh cleaning procedures, be rigid to withstand the high flow rates, and have the porosity to provide a high surface area for adsorption. Matrices can be inorganic like porous silica, controlled pore glass, and hydroxyapatite, synthetic organic polymers like polyacrylamide, polymethyl methacrylate, or natural polysaccharides like cellulose, dextran, and agarose. The type of matrix often defines the basis of chromatography, be it size exclusion, ion exchange, hydrophobic interaction, or affinity. Table 4 summarizes the different types of chromatography and their basis of separation.

The pattern of separated proteins obtained on the detector is termed a chromatogram. The quality of a chromatographic separation can be quantitatively appreciated by calculating the hypothetical number of plates in the peak of a chromatogram (N). The theoretical plate height (H) is also often used as a parameter to quantify a chromatogram. Fig. 5 shows a typical peak in a chromatogram. N for such a peak and is given as:

$$N = 16\left(\frac{t_r}{t_w}\right)^2 \quad (1)$$

Table 4 Various types of liquid chromatography

Types	Basis	Application
Ion exchange	Ionic charge	Proteins
Gel filtration/size exclusion	Size	Desalting, large molecules
Affinity	Specific binding	Antibodies, antigens
Reverse phase	Hydrophobicity	Peptides
Hydrophobic interaction	Hydrophobicity	Proteins
Chromatofocusing	Isoelectric point	Proteins
Ligand exchange	Adsorption	Sugars

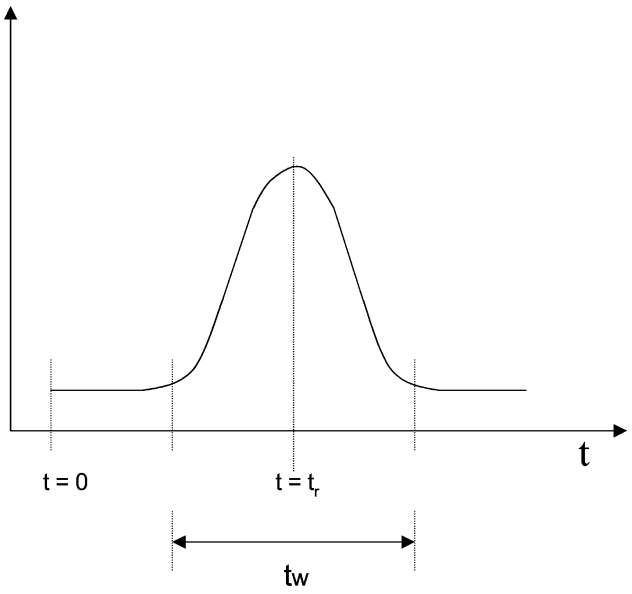


Fig. 5 Typical chromatogram peak characteristics.

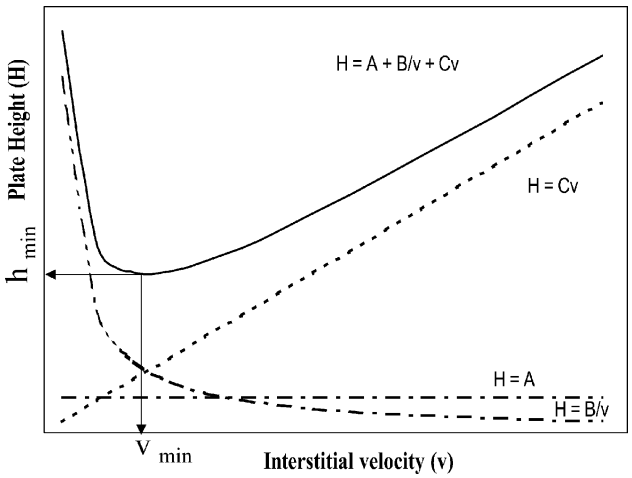


Fig. 6 Graphical representation of the van Deemter equation showing the effect of eluent velocity on the plate height of a chromatography column.

A greater number of plates indicate higher efficiency of separation. This implies that a smaller base width (t_w) will indicate a better resolution. Further, if each plate has a height H, then,

$$H = L/N \tag{2}$$

where L is the length of the column. Obviously, a lower plate height indicates a better resolution. H also depends on the interstitial velocity v of the mobile phase. The relation between H and v is given by the Van Deemter equation:

$$H = A + \frac{B}{v} + Cv \tag{3}$$

where A, B, and C are constants. Knowing the Van Deemter constants for a system can give an idea about the operating velocity at which resolution of the eluting peaks will be the best. As shown in the plot of H vs. v (Fig. 6), there is an optimum v at which H is minimum. Numerically, H is minimum when $v = (B/C)^{1/2}$.

POLISHING

After a series of chromatography steps yield a highly pure product in the liquid state, the product needs to undergo one or more steps to convert it to the form in which it will be sold. Drugs need to be formulated and mixed with binders, colors, and other excipients if they are sold in a solid form. They can also be sold as liquids or aerosols, suspensions or emulsions. In such a case, they need to be diluted and mixed at the correct concentration.

Bulk biochemicals like organic acids and solvents are often formulated as concentrated solutions after removal of water from the stream exiting from the final separation step. Antibiotics are usually crystallized from solution to a very high purity. Recombinant proteins are usually sensitive and can lose bioactivity if not stored in a proper manner. Proteins are formulated mostly as dry powders, and a variety of stabilizing agents are added to preserve the bioactivity of the protein.[12] Enzymes are usually freeze dried or lyophilized to obtain a solid form. Lyophilization basically sublimes the water to vapor under vacuum. It is a nondenaturing process and is used for drying pharmaceutical products, therapeutic proteins, food items, viruses, and bacteria. If the product is not very sensitive to heat, it can be dried by spray dryers, vacuum shelf dryers, or rotary dryers. These are well-known and widely used unit operations in the chemical industry. Details about these unit operations can be found in many classic chemical engineering handbooks.

CASE STUDIES

In this section, we try to develop downstream processing strategies for two different products: rhGDNF as an IB in *E. coli* and EPO as a protein secreted from CHO cells. The genomes of the two different organisms used have been genetically engineered to produce the proteins. The purpose of such an exercise is to learn to integrate the basic steps discussed so far in simulating an actual process. The cases will combine the spectrum of both intracellular and extracellular proteins as well as prokaryotic and eukaryotic expression.

Production of rhGDNF by *E. coli*

GDNF is an approximately 20 kDa, glycosylated polypeptide that exists in its native form as a homodimer.[24] The gene for GDNF has been mapped to human chromosome 5 at p12–p13.1 and gives rise to two alternatively-spliced forms that code for prepropeptides of 211 and 185 amino acid (aa) residues, respectively.[25] Both the long (alpha) and the short (beta) forms yield equivalent 134-aa residue mature forms after proteolytic cleavage.[24,26] Although the native molecule contains two potential glycosylation sites, nonglycosylated rhGDNF has full biological activity.[24]

GDNF has the ability to promote the survival and differentiation of dopaminergic neurons in primary cultures of the embryonic ventral midbrain.[24] It has also been found to have trophic effects on dopaminergic neurons (increasing the number surviving in culture, dopamine uptake, cell size, and extension of neurites). It also protects mecencephalic dopaminergic cells against the deleterious effects of the neurotoxin 1-methyl-4-phenylpyridinium (MPP^+).[27] This finding is central to the development of GDNF for the treatment of Parkinson's disease because the primate model of Parkinson's disease relies on the use of 1-methyl-4-phenyl,1,2,3,6-etrahydropyridine, a precursor to the neurotoxin MPP^+. By virtue of its effects on cell culture and efficacy in animal studies, GDNF constitutes a powerful therapeutic candidate for several neurodegenerative diseases, especially Parkinson's.[25,28]

For large-scale production of GDNF, *E. coli* is a very good choice as a host. It has several advantages in this case. First, nonglycosylated GDNF, though highly potent as a therapeutic agent, has a relatively small size (134 aa, MW ~15 kDa) and simple structure. Second, the recombinant nonglycosylated form is as functionally active as the glycosylated native form found in the human brain. In general, bacterial host systems cannot reproduce the authentic glycosylation of eukaryotic proteins, and in such cases where glycosylation is important for the protein to function, a more complex host system has to be chosen. For this case, therefore, *E. coli* is very suitable as a host for the production of rhGDNF. However, the recombinant protein often forms an IB in such a bacterial system. This causes problems in the extraction and downstream processing of the protein.

A flowsheet of the steps in downstream processing is given in Fig. 7. At the end of the fermentation, the cells are transferred into a blending tank, which isolates the upstream from the downstream section of the plant. The broth is then fed into a disk stack centrifuge and centrifuged to harvest the cells. The supernatant is centrifuged again and the cell sludges from both streams are resuspended in ice-cold buffer. The buffer facilitates the separation of the cell debris from the IBs. The suspension is maintained at 4°C to prevent heating and subsequent proteolytic degradation of the target protein during the cell lysis steps, which is done in a high-pressure homogenizer. The cell lysate from the homogenizer is a viscous extract. To reduce the viscosity, the lysate is treated with lysozyme (added to a final concentration of 100 μg/ml) and then homogenized again.

The second homogenized stream is then treated with half the volume of protease inhibitor cocktail (without ethylenediaminetetraacetic acid) to inhibit any proteolytic degradation of the target protein. It is then centrifuged and the IBs are recovered in the heavy phase while most of the cell debris remains in the supernatant. This is possible because the density (1.3 g/L) and the size (φ: 1 μm) of the IBs are significantly larger than the cell debris. The IB sludge is then washed with wash buffer containing Triton X100 to remove any loosely associated contaminants. The mixture is centrifuged, and the IB sludge is then resuspended in phosphate buffer solution (PBS) in a blending tank. The PBS is drained in a microfilter, and the IBs are transferred to a glass-lined blending tank and mixed with a solubilizing buffer. The solubilizing buffer contains urea solution, a chaotropic agent that dissolves the denatured protein in the IBs.

At the end of the solubilization, the solution is concentrated in a diafiltration unit. All the remaining fine particles (biomass debris and IBs) are removed using a polishing dead-end filter. This polishing filter protects the chromatographic units that are used next. The filtered stream is purified using anion exchange chromatography. Two quaternary amine Sepharose columns (Q-Sepharose, Pharmacia) work in parallel. The feed is applied with 10 mM Tris-HCl (pH 8.0) containing 4 M urea and the column is developed with a linear gradient of 0–0.5 M NaCl in this buffer.[24] The next step involves the removal of the SO_3 moieties from the cystine side chains in the rhGDNF to allow proper disulfide bonding and correct refolding of the protein to its native form. Pooled fractions from the Q-Sepharose column are mixed with 19 volumes of refolding buffer in a blending tank.[24] Dilution is necessary both in the solubilization and in the refolding steps to minimize the intermolecular interactions that can lead to protein inactivation.

At this stage, the stream contains the dissolved rhGDNF, which is then concentrated using a diafilter followed by a dead-end polishing filter to remove any non-native protein aggregates, and then subsequently fed with 50 mM Tris–HCl (pH 8.0) into two S-Sepharose (sulfonate based cation exchanger) columns working in parallel. The column is developed with a linear gradient of 0–1 M NaCl in this buffer.[24] Fractions containing rhGDNF are applied to a hydroxyapatite column in 25 mM sodium phosphate (pH 6.8). The purified protein

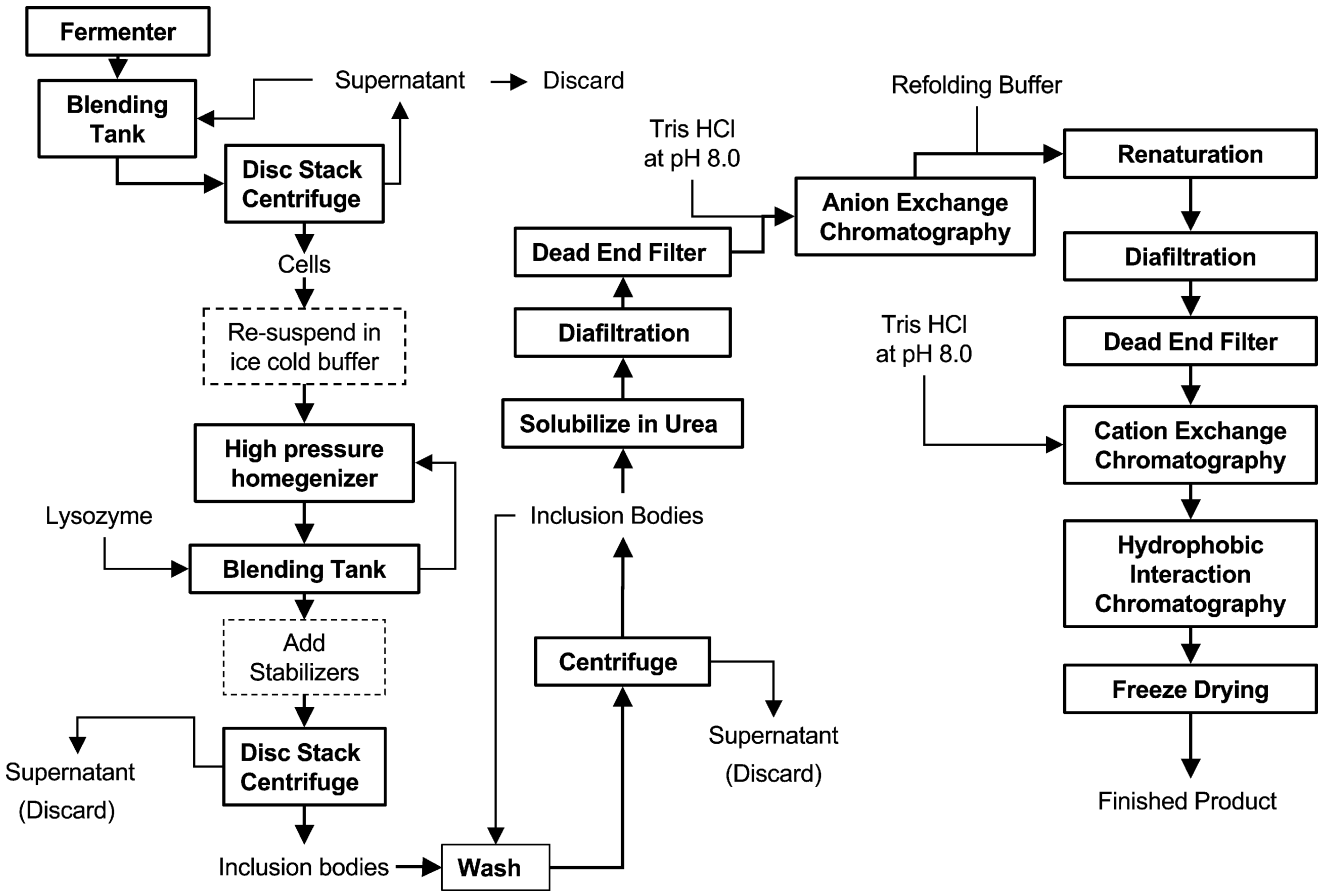


Fig. 7 A process flowsheet for rhGDNF production.

is lyophilized to yield a powdered form. It is 99.99% pure and contains <0.01% bacterial proteins and <5 pg/mg bacterial endotoxins.

Liew et al. reported that GDNF was a recalcitrant protein with a strong tendency to form IBs in *E. coli*.[29] However, insolubility has its advantages too:

a. Inclusion bodies can represent the highest yielding protein fraction of the target protein.
b. They are easy to isolate as an efficient first step in a purification scheme. Nuclease treated washed IBs are usually 75–95% of pure target protein.
c. Target proteins in IBs are generally protected from proteolytic breakdown.
d. The target protein is inactive and cannot harm the host if it is a potential toxin.

The steps followed in downstream processing of rhGDNF expressed in *E. coli* are very similar to any conventional method used to purify target proteins expressed as IBs in bacteria. However, the specific reagents used at different stages are decided after optimization at a lower scale. Alternative methods exist to substitute various steps in the process, but their applicability depends on the specific protein under consideration. For example, if we produce rhGDNF as a fusion protein with markers such as thioredoxin or glutathione *S*-transferase, then the protein is more soluble and, therefore, the purification strategy would be different and less cumbersome. Also, depending on the chemical properties of these markers, the chromatography steps would also change.

Production of EPO by CHO Cells

Epotein alpha or Epogen is an aqueous formulation of recombinant human EPO expressed in CHO cells. Erythropoietin is a 166-aa glycoprotein with an MW of 34 kDa.[30] It is heavily glycosylated with carbohydrates accounting for 38% of the MW.[31] Human EPO has three *N*-glycosylation sites and one *O*-glycosylation site.[32] Epogen is the trade name for recombinant EPO manufactured by Amgen and is formulated as a sterile colorless liquid in an isotonic sodium chloride/sodium citrate buffered solution for intravenous or subcutaneous administration in both preservative-free

single-dose and preservative (benzyl alcohol) containing multiple-dose vials. Johnson and Johnson, Ortho Biologics, and Kirin Pharmaceuticals (Japan) are some of the other companies manufacturing EPO.

The 2.7 billion dollars earned by EPO sales accounted for ~48% of Amgen's total revenue in 2002. However, with the market growing at ~23% by the last quarter of 2002, the sale of EPO products made nearly $7.8 billion in the U.S. market alone in 2004, making it by far the largest selling biopharmaceutical product ever.[33] In fact EPO sales in the United States is projected to reach a whopping $15 billion in 2010.[33] Erythropoiesis, the formation of red blood cells in the body, is regulated by EPO. Recombinant EPO expressed by CHO cells has a structural conformation similar to that of the native human protein and the market for recombinant EPO has been created because of its efficacy in relieving anemia in patients suffering from chronic renal failure.[30,31] Epogen increases the RBC levels in such patients, thereby alleviating the need for blood transfusions. Besides this well-established activity, EPO also has therapeutic potential in the treatment of stroke, head trauma, and epilepsy.[30]

Erythropoeitin is heavily glycosylated and its carbohydrate moiety has been shown to be important for its biological activity.[34] It is also sialylated, and this is crucial for in vivo biological activity.[35] Although almost all the EPO obtained earlier was just purified from human urine, the advent of recombinant DNA technology had many groups trying different methods to produce recombinant EPO.[36] However, the complex glycosylation pattern of the protein and the need for high sialic acid content make mammalian cell culture the best available option to produce EPO.

In large-scale production, EPO is produced by a CHO cell line under serum-free conditions by fed-batch fermentation.[36,37] The use of serum-free media makes the subsequent recovery of EPO much easier. CHOpDSVL-gHuEPO cells are grown as a suspension culture in bioreactors to produce EPO at a concentration of 200–300 mg/L.[32,36]

A flowsheet of the downstream processing steps is given in Fig. 8.[36] Postfermentation, the broth is cooled and clarified using a disk stack centrifuge followed by filtration. Care should be taken when choosing cell separation equipment because cell lysis at this point can contaminate the supernatant and make purification of the EPO difficult. Cell separation equipment specially designed for separating fragile eukaryotic cells, e.g., CSC6 [6000 m^2 ECA (which stands for effective calrification

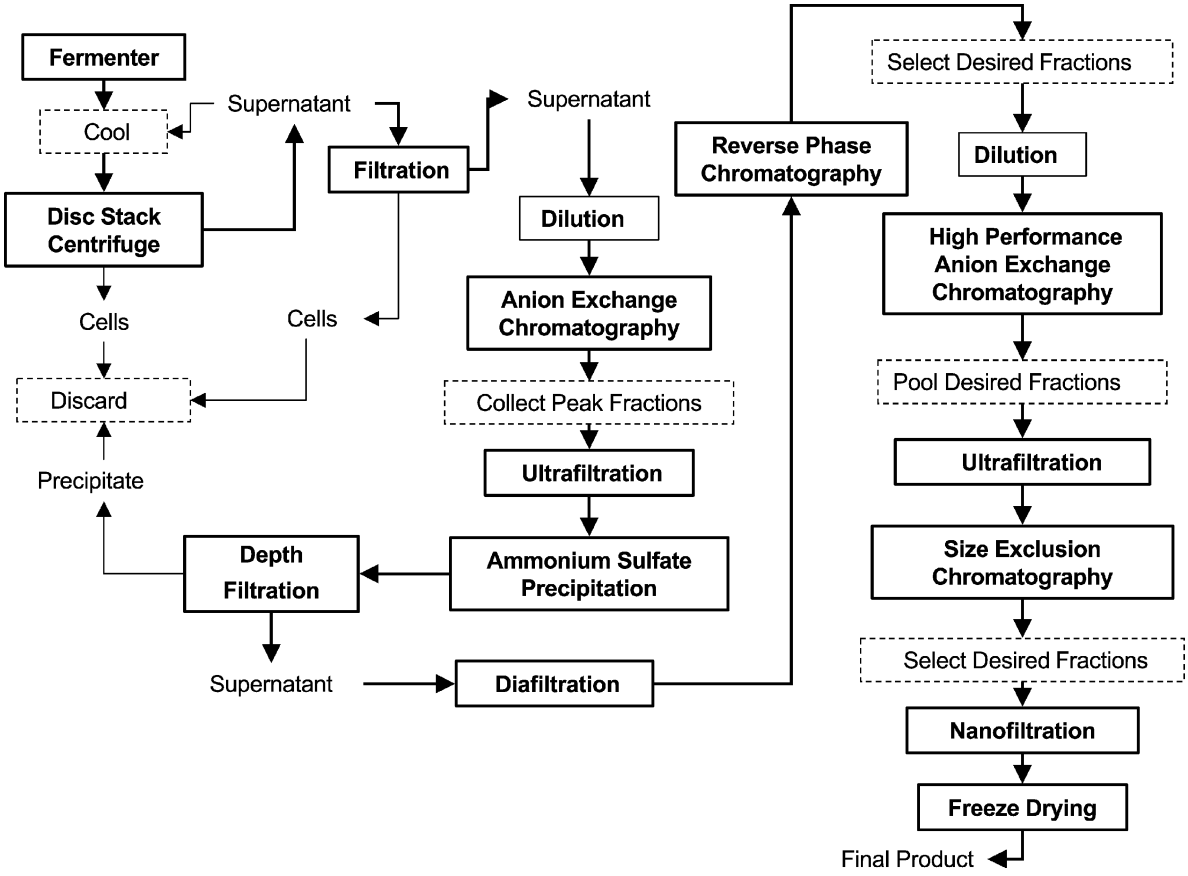


Fig. 8 A process flowsheet for rhEPO production.

area), 15,500× *g*, 200 L/hr, WestFalia], with hydrohermetic feed inlet, or BTPX 250 (11,000 m^2 ECA, 13,000× *g*, 300 L/hr, Alfa Laval), with gentle disk inlet, should be used to minimize shear stress induced lysis of the cells. Filtration is followed by another centrifugation step to further clarify the broth.

The crude extract after clarification is then diluted with water before being applied to an anion exchange chromatography column equilibrated with 20 mM Tris (pH 7.5) and 50 mM NaCl. The product is eluted by step elution and pooling the peak fractions gives a yield of ~50–60%. It is important to note that the capture step works at a neutral pH because a pH less than 6 would activate the endogenous proteases. The pooled fractions obtained from the anion exchange column are concentrated by UF and then further purified by using 2.4 M ammonium sulfate to precipitate the contaminating host cell proteins. At this concentration of ammonium sulfate, almost none of the EPO precipitates out. The salt concentration is then reduced using diafiltration to less than 240 mM.

The reversed phase chromatography step is then performed to remove different isoforms of the protein according to their sugar moiety and degree of sialylation, as well as viruses and the remaining host cell proteins. Acetonitrile is used as a solvent to elute out the protein, and because the heavily glycosylated and sialylated forms elute first, the first half of the eluting fractions is pooled together and diluted. The high-performance anion exchange chromatography then separates the different isoforms again, but this time according to their isoelectric point, and not the degree of hydrophobicity. As the degree of sialylation affects the isoelectric point, the column is an efficient way to separate the desired isoforms. The correct fractions are pooled together, concentrated by UF, and then subjected to size exclusion chromatography to remove any potential dimers or higher aggregates. The final step of dead-end NF removes any potential viral contaminants from the purified protein, which is then freeze-dried and sent for formulation.

CONCLUSIONS

To engineer living organisms to produce bioproducts of our choice is a relatively easy task compared to the process of recovering them from the fermentation broth. In most cases, the bioproduct is produced in very small quantities. Every single step in a downstream processing sequence not only increases the cost, but also decreases the product yield. Although novel chromatographic methods are increasing the specificity and yield for the products of interest, they are costly and difficult to scale up. However, it is to be realized that not all bioproducts need to go through every step outlined in this entry. Commercial enzymes are usually prepared as crude extracts, while polysaccharides and biogum recovery do not require as sophisticated purification steps as recombinant proteins.[7] Economics drives the industry, and therefore, innovation needs to be made both in the upstream part to increase productivity as well as in the downstream processing section to purify the product in as few steps as possible and yet to the highest degree achievable. In the foreseeable future, bioseparations will remain a challenging engineering field that must continue to develop to economically produce a myriad of recombinant therapeutic protein products currently under development and clinical trials.

REFERENCES

1. Atkinson, B.; Mavituna, F. *Biochemical Engineering and Biotechnology Handbook*; The Nature Press: New York, 1983; 890–931.
2. Belter, P.A.; Cussler, E.L.; Hu, W-S. *Bioseparations: Downstream Processing for Biotechnology*; Wiley Interscience: New York, 1998.
3. Yamada, K. Recent advances in industrial fermentation in Japan. Biotechnol. Bioeng. **1977**, *19*, 1563.
4. Matejtschuk, P.; Baker, R.M.; Chapman, G.E. Purification of characteristics of monoclonal antibodies. In *Bioseparation and Bioprocessing: A Handbook*; Subramaniam, G., Ed.; Wiley-vett: Wennihem, Germany, 1998; Vol. II, 223–252.
5. Bykowska, K.; Rijken, D.C.; Collen, D. Purification and characterization of the plasminogen activator secreted by a rat brain tumor cell line in culture. Thromb. Haemost. **1981**, *46*, 642–644.
6. Garcia, A.A.; Bonen, M.R.; Ramirez-Vick, J.; Sadaka, M.; Vuppu, A. *Bioseparation Process Science*; Blackwell Science: Malden, MA, 1999.
7. Bailley, J.E.; Ollis, D.E. Product recovery operations. In *Biochemical Engineering Fundamentals*; 2nd Ed.; McGraw-Hill Inc.: New York, 1986; 726–797.
8. Harrison, R.G.; Todd, P.; Rudge, S.R.; Petrides, D.P. *Bioseparations Science and Engineering*; Oxford University Press: New York, 2003.
9. Kelley, B.D.; Halton, T.A. Protein purification by liquid-liquid extraction. In *Biotechnology, Vol. 3, Bioprocessing*; Stephanopolous, G., Ed.; VCH Publishers: Weinheim, 1993; 593.
10. Mattiasson, B.; Kaul, R. Use of aqueous two phase systems for recovery and purification in biotechnology. In *Separation, Recovery and Purification in Biotechnology, Recent Advances and Mathematical Modelling*; Asenjo, J.A., Hong, J., Eds.; American Chemical Society: Washington, DC, 1986; 78–92.

11. Mattiasson, B.; Ling, T.G.I. Efforts to integrate affinity interactions with conventional separation technologies: affinity partition using biospecific chromatographic particles in aqueous two-phase systems. J. Chromatogr. **1986**, *376*, 235–243.
12. Kaul, R-H.; Mattiasson, B. Downstream processing in biotechnology. In *Basic Biotechnology*; 2nd Ed.; Ratledge, C., Kristiansen, B., Eds.; Cambridge University Press: Cambridge, U.K., 2001; 187–212.
13. Fischer, B.; Summer, I.; Goodenough, P. Isolation, renaturation and formation of disulfide bonds of eukaryotic proteins expressed in *Escherichia coli* as inclusion bodies. Biotechnol. Bioeng. **1993**, *41*, 3–13.
14. Buchner, J; Rudolph, R. Renaturation, purification and characterization of recombinant Fab-fragments produced in *Escherichia coli*. Bio/Technology **1991**, *9*, 157–162.
15. Saito, Y.; Ishii, Y.; Niwa, M.; Ueda, I. Direct expression of a synthetic somatomedine C gene in *Escherichia coli* by use of a two cistron system. J. Biochem. **1987**, *101*, 1281–87.
16. Smith, A.T.; Santama, N.; Dacey, S.; Edwards, M.; Bray, R.C.; Thorneley, R.N.F.; Burke, J.F. Expression of a synthetic gene for horseradish peroxidase C in *Escherichia coli* and activation of the recombinant enzyme with Ca^{2+} and heme. J. Biol. Chem. **1990**, *265*, 13335–13343.
17. Masaaki, T.; Keiko, S.; Shigeo, K. Effective refolding of fully reduced lysozyme with a flow-type reactor. Process. Biochem. **1996**, *31* (4), 341–345.
18. Navarro, A.; Carmel, H.; Rigal, L.; Phemius, P. Continuous chromatographic separation process: simulated moving bed allowing simultaneous withdrawal of three fractions. J. Chromatogr. A, **1996**, *770*, 39–50.
19. Xie, Y.; Koo, Y-M.; Wang, N.-H.L. Preparative chromatographic separation: simulated moving bed and modified chromatography methods. Biotechnol. Bioprocess. Eng. **2001**, *6*, 363–375.
20. Berg, C. Hypersorption process for separation of light gases. Trans. A.I.Ch.E. **1946**, *42*, 665–680.
21. Wankat, P.C. *Rate Controlled Separation*; Elsevier Applied Science: New York, 1990.
22. Wankat, P.C. *Large Scale Adsorption and Chromatography*; CRC Press: Boca Raton, FL, 1986; Vol. II.
23. Chang, W.-J.; Koo, Y.M. On line recovery of large molecules from mixtures using reciprocating size exclusion chromatography. Biotechnol. Tech. **1999**, *13*, 211–214.
24. Lin, L.-F.H.; Doherty, D.H.; Lile, J.D.; Bektesh, S.; Collins, F. GDNF: a glial cell line derived neurotrophic factor for midbrain dopaminergic neurons. Science **1993**, *260*, 1130–1132.
25. Lin, L-F.H. Glial cell line derived neurotrophic factor (GDNF): a comprehensive review. Neural Notes II **1996**, (3), 3–7.
26. Schindelhauer, D.; Schuffenhauer, S.; Gasser, T.; Steinkasserer, A.; Meituyer, T. The gene coding for glial cell line derived neurotrophic factor (GDNF) maps to chromosome 5p12-p13.1. Genomics **1995**, *28*, 605–607.
27. Hou, J.G.G.; Lin, L-F.H.; Mytilineou, C. Glial cell line derived neurotrophic factor exerts neurotrophic effects on dopaminergic neurons *in vitro* and promotes their survival and regrowth by 1-methyl 4-pyridinium. J. Neurochem. **1996**, *66*, 74–82.
28. Lapchak, P.A.; Gash, D.M.; Shoushu, J.; Miller, P.J.; Hilt, D. Glial cell line derived neurotrophic factor—a novel therapeutic approach to treat motor dysfunction in Parkinson's disease. Exp. Neurol. **1997**, *144*, 29–34.
29. Liew, O-W.; Choo, A.B.H.; Too, H.P. Parameters influencing the expression of mature glial cell line–derived neurotrophic factor in *E. coli*. Biotechnol. Appl. Biochem. **1997**, *25* (3), 223–233.
30. Zanette, D.; Soffientini, A.; Sottani, C.; Sarubbi, E. Evaluation of phenylboronate agarose for industrial scale purification of erythropoietin from mammalian cell cultures. J. Biotechnol. **2003**, *101*, 275–287.
31. Rader, R.A. *BioPharma: Biopharmaceutical Products in the US market*, 2nd Ed.; Biotechnology Information Institute: Rockville, MD, 2003; 42–45.
32. Jacobs, K.; Shoemaker, C.; Rudersdorf, R.; Neill, S.D.; Kaufman, R.J.; Mufson, A.; Seehra, J.; Jones, S.S.; Hewick, R.; Fritsch, E.F.; Kawakita, M.; Shimizu, T.; Miyake, T. Isolation and characterization of genomic and cDNA clones of human erythropoietin. Nature **1985**, *313*, 806–810.
33. U.S. Blood Growth Factors Markets. *Erythropoietins Market*. Market Analysis and Forecasts; Frost and Sullivan Research Report, Sep 21, 2004. www.frost.com.
34. Dube, S.; Fisher, J.W.; Powell, J.S. Glycosylation at specific sites of erythropoietin is essential for biosynthesis, secretion and biological function. J. Biol. Chem. **1988**, *25*, 17516–17521.
35. Yoon, S.K.; Song, J.Y.; Lee, G.M. Effect of low culture temperature on specific productivity, transcription level and heterogeneity of erythropoietin in Chinese hamster ovary cells. Biotechnol. Bioeng. **2003**, *82* (3), 289–298.
36. Alliger, P.; Palma, N. Chromatographic Purification of Recombinant Human Erythropoietin. WO Patent 03045996, 2003.
37. Lin, F.K. Production of Erythropoietin US Patent 5955422, 1993.

Blowing Agent

Kyung W. Suh
Midland, Michigan, U.S.A.

INTRODUCTION

Cellular plastics, also known as plastic foams, or foamed plastics can be produced primarily by extrusion, injection molding, rotational molding, blow molding, thermal expansion, calendaring, mechanical frothing, or spray-on conveyors, using physical blowing agents (PBAs) or chemical blowing agents (CBAs). Most plastic foams are blown with inert gases (CO_2, N_2, H_2O), CBAs that release inert gases, hydrocarbons (HCs) containing three to six carbon atoms, chlorinated HCs, chlorofluorocarbons (CFCs) such as CFC-11 (R-11), CFC-12 (R-12), CFC-113 (R-113), and CFC-114 (R-114), hydrochlorofluorocarbons (HCFCs) such as HCFC-141b (R-141b) and HCFC-142b (R-142b), and hydrofluorocarbons (HFCs) such as HFC-152a (R-152a) and HFC-134a (R-134a). Since the theory of ozone depletion by chlorofluorocarbons was first published in 1974, extensive research has been conducted by government, academia, and industry to define the science and develop alternatives.[1–4] Emission of certain volatile organic compounds (VOC) generates smog photochemically. Because CFCs have low reactivity in the lower atmosphere, they will migrate to the stratosphere in about 10 yr, depleting ozone in the upper atmosphere and possibly affecting global climate change. Substitution of more photochemically reactive compounds such as HCs, HCFCs, and HFCs for CFCs may reduce the ozone depletion potential (ODP), as well as global warming potential (GWP) in the stratosphere, but may adversely affect indoor air quality in the lower atmosphere. Therefore, interaction with the total environment must be considered in developing nontoxic, environmentally acceptable blowing agents that are sustainable and economically viable.

NOMENCLATURE

A cellular plastic is defined as a two-phase gas–solid plastic system whose apparent density is decreased substantially by the presence of numerous cells or voids disposed throughout its mass. These cells or voids are chiefly formed by using CBAs, PBAs, microballoons, or fillers containing gases or liquids. If these cells are discrete and the gas phase of each is independent of that of the other cells, the material is termed closed cell. If these cells are interconnected, the material is termed open cell. The nomenclature of cellular polymers is not standardized; classifications have been made according to the properties of the base polymer, the methods of manufacture, the cellular structure, or some combination of these. According to an ASTM test, foamed plastics are classified as rigid or flexible.[5] A flexible foam is one that does not rupture when a 20 cm × 2.5 cm × 2.5 cm piece is wrapped around a 2.5 cm mandrel at a uniform rate of one lap per 5 sec, at 15–25°C. Rigid foams rupture under this test. This classification is used here. The term structural foam has not been exactly defined, but is used here to refer to rigid foams produced with densities greater than 320 kg/m^3. In the case of cellular rubber, the ASTM uses several classifications based on the method of manufacture.[6,7] Cellular rubber is a general term covering all cellular materials that have an elastomer as the polymer phase. Sponge rubber and expanded rubber are cellular rubbers produced by expanding bulk rubber stocks, and are open cell and closed cell, respectively. Latex foam rubber, which is also a cellular rubber, is produced by frothing a rubber latex or liquid rubber, gelling the frothed latex, and then vulcanizing it in the expanded state. Various blowing agents used to prepare cellular polymers may be classified according to the properties of the resultant foamed plastics. High-density structural foams are usually produced with CBAs. On the other hand, low-density foams with density less than about 320 kg/m^3 are typically produced with PBAs or a combination of both CBAs and PBAs. Cross-linked polyolefin foams can be produced with CBAs by heating rather than decompression.

PHYSICAL BLOWING AGENT

Selection of suitable blowing agents for each polymer is very important for producing dimensionally stable foams with the desired cellular structures and physical properties in various end-use applications. Some of the desirable properties of suitable blowing agents are as follows.

1. High solubility of blowing agents at high processing temperatures and low solubility at the

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016152

ambient temperature are desirable for improving processability due to lower system pressure, better blowing agent mixing, better cell size control, and also for improving physical properties due to less plasticization of cellular polymers at the ambient temperature.

2. The permeability of the blowing agent or blowing agent mixtures through cellular plastics should be similar to or higher than that of air through the same cellular plastics. Nakamura showed in his patent that to produce a dimensionally stable foam, the permeability of a mixed blowing agent system should be between 0.75 and 6 times that of air through the same styrenic polymers.[8] However, this range of blowing agent permeability may be different depending on the type of polymer foams.
3. The heat of vaporization of a suitable blowing agent should be high at the foaming temperature to reduce the time to stabilize or solidify the foam after expansion by decreasing the temperature of the polymer phase as the blowing agent vaporizes and expands to produce cells or voids.
4. The rate of temperature drop with respect to the corresponding rate of pressure drop in an adiabatic expansion, known as Joule–Thompson co-efficient, should be large enough to make the expanding foam cool faster, and thereby the foam solidifies or sets up faster to produce low-density, dimensionally stable foam. The total internal cooling of the expanding foam consists of evaporative cooling by the heat of vaporization of the blowing agent and the adiabatic expansion cooling by the Joule–Thompson effect.
5. A suitable blowing agent or blowing agent mixture for the thermal insulation foam containing at least one or more insulating blowing agents should have low vapor thermal conductivity (preferably lower than about 0.015 W/m/K) and low gas permeability to produce a thermal insulation foam with low long-term K-factor. The long-term K-factor is defined as the aged K-factor when the foam is saturated with one atmosphere of air inside the cells. Suh showed that an insulating blowing agent as defined should have low permeability, less than about 1.0 cm^3mil/day per 100 in^2. atm [6.0×10^{-13} cm^3 (STP) cm/cm^2/sec/cmHg].[9,10] This will insure that a dimensionally stable low-density insulation foam with low long-term K-factor can be produced by using about 30/70 to 70/30 mixture by weight of one or more noninsulating gases [fugitive blowing agent = permeability greater than about 1.0 cm^3 mil/day per 100 in^2 atm or $6.0 \times 10^{-13}$$cm^3$ (STP) cm/cm^2/sec/cmHg] and one or more insulating gases.
6. A suitable blowing agent should have low toxicity, low odor, low flammability, low corrosiveness, low molecular weight, high specific gravity, non-VOC (photochemical reactivity not greater than that of ethane), low ODP, low GWP, and low cost.
7. A suitable blowing agent should have low vapor pressure at room temperature, which will facilitate easy handling under ambient condition.

Environmental Issues

A successful selection of blowing agent depends on many factors as described earlier and also on the ability to produce foams with the desired cellular structures, cross-sectional size and shape, and properties required in end-use applications. Therefore, foam developments often have been motivated by issues associated with the blowing agents. Currently, there are three environmental issues facing the blowing agents: plant emission of VOC, ozone depleting potential, and GWP. Table 1 shows the physical properties of potential blowing agents. In response to the growing scientific consensus that CFCs would deplete the ozone layer in the stratosphere, the U.S. government banned the use of CFCs in aerosols in 1979 and the United Nations Environmental Programme (UNEP) made a concerted effort to obtain international agreements aimed at protecting the ozone layer from 1981. In March 1985, the Vienna Convention for the protection of the ozone layer was convened and developed a framework for international cooperation in research, environmental monitoring, and information exchange. In September 1987, the Montreal Protocol on ozone depleting substances was signed by 24 nations and took effect on January 1, 1989. This treaty called for limiting production of specified hard CFCs, including CFC-11, -12, -113, -114, and -115, to 50 % of the 1986 levels by 1998. Dow was the first polystyrene foam producer to develop a formulation that replaced ozone depleting substances like CFC-12 with HCFC-142b, which has substantially reduced the ODP of the insulating blowing agent used to make insulation foams.[9,10,15,16] The Montreal Protocol was followed by further agreements to eliminate the use of HCFCs and also to reduce the release of greenhouse gases by reducing energy consumption (Kyoto Protocol). Many countries have since taken a more aggressive schedule for eliminating blowing agents that have ODP values greater than zero. Since the ban of CFCs in foam applications, polystyrene foam producers started to switch to HCFC-142b in 1989, and the polyurethane and phenolic foam industry made the transition to HCFC-141b. Of the common HCFCs, 141b has the highest ODP. As a result, it was targeted for phaseout

Table 1 Physical properties of potential physical blowing agents

Blowing agent	Chemical formula	MW	BP°C	Density (g/cm^3)	Heat of vaporization (cal/g)	Thermal conductivity (w/m/K)	Flamm. limits (Vol%) in air	ODP	GWP	VOC	Toxicity TLV (ppm)
Inert gas											
Nitrogen	N_2	28	−196	1.146	—	0.0261	NF	0	0	No	No
Air	$79N_2/21O_2$	29	—	1.184	—	0.0260	NF	0	0	No	No
Carbon dioxide	CO_2	44	−78.5	1.811	137	0.0166	NF	0	0.00025	No	No
Water	H_2O	18	100	1.000	262	0.0178	NF	0	0	No	No
HC											
Propane	C_3H_8	44.1	−42.1	0.585	101.8	0.0178	2.1–9.5	0	—	Yes	1000
n-Butane	C_4H_{10}	58.1	−0.5	0.56	92.1	0.0159	1.8–8.4	0	—	Yes	600
i-Butane	$(CH_3)_2CHCH_3$	58.1	11.7	—	—	0.0163	1.8–8.4	0	—	Yes	No
n-Pentane	C_5H_{12}	72.2	36.1	0.616	86.0	0.0135	1.5–7.8	0	—	Yes	1000
i-Pentane	$(CH_3)_2CHC_2H_5$	72.2	27.9	0.615	31.5	—	1.4–7.6	0	—	Yes	No
Cyclopentane	C_5H_{10}	70.1	49.2	0.740	99.7	—	F	0	—	Yes	No
n-Hexane	C_6H_{14}	86.2	68.7	0.655	80.3	—	1.2–7.4	0	—	Yes	No
Cyclohexane	C_6H_{12}	84.2	80.8	0.774	94.9	—	F	0	—	Yes	No
CFC											
CFC-11	CCl_3F	137.4	23.8	1.467	43.1	0.0074	NF	1.0	1.0	No	No
CFC-12	CCl_2F_2	120.9	−29.8	1.311	39.5	0.0095	NF	0.95	3.1	No	No
CFC-113	$C_2Cl_3F_3$	187.4	47.6	1.565	35.1	0.0078	NF	0.6	—	No	No
CFC-114	$C_2Cl_2F_4$	170.9	3.8	1.456	32.5	0.0112	NF	0.7	4.0	No	No
CHC											
Methyl chloride	CH_3Cl	50.5	−24.2	0.997	121.9	0.0105	10.7–17.4	0.02	—	No	Yes 50
Methylene chloride	CH_2Cl_2	84.9	39.8	1.320	78.7	0.0122	NF	—	—	Yes	Yes
Ethyl chloride	C_2H_5Cl	64.5	12.4	0.90	92.3	0.0095	3.6–15.4	0.003	—	Yes	Yes 150
Oxygenated HC											
Ethanol	C_2H_5OH	46.1	78.3	0.785	204.1	0.0157	3.3–19.0	0	—	Yes	No
2-Propanol	$(CH_3)_2CHOH$	60.1	82.3	0.780	175.7	0.0141	2.2–13.5	0	—	Yes	No

(Continued)

Table 1 Physical properties of potential physical blowing agents *(Continued)*

Blowing agent	Chemical formula	MW	BP°C	Density (g/cm^3)	Heat of vaporization (cal/g)	Thermal conductivity (w/m/K)	Flamm. limits (Vol%) in air	ODP	GWP	VOC	Toxicity TLV (ppm)
HCFC											
HCFC-22	$CHClF_2$	86.5	−40.8	1.194	55.8	0.0118	NF	0.055	0.35	No	500
HCFC-141b	CH_3CCl_2F	117.0	32.1	1.233	49.8	0.0100	7.6–17.7	0.11	0.09	No	500
HCFC-142b	CH_3CClF_2	100.5	−9.4	1.12	51.4	0.0094	6.7–14.9	0.055	0.38	No	1000
HCFC-123	$CHCl_2CF_3$	152.9	27.9	1.463	44.8	0.0104	NF	0.01	0.019	No	10
HCFC-124	$CHClFCF_3$	136.5	−12.2	1.364	—	—	NF	0.02	0.097	No	500
HFC											
HFC-152a	CH_3CHF_2	66.1	−24.7	0.950	53.3	0.0094	3.7–14.9	0	0.03	No	1000
HFC-134a	CH_2FCF_3	102.0	−26.5	1.21	50.8	0.0141	NF	0	0.27	No	1000
HFC-143a	CH_3CF_3	84.0	−47.4	1.176	—	0.0094	6.0–	0	0.74	No	1000
HFC-245fa	$CF_3CH_2CHF_2$	134	—	—	—	—	—	0	—	No	No
HFC-365mfc	$CF_3CH_2CF_2CH_3$	148	40.2	1.27	42.3	0.0106	3.5–9.0	0	—	No	No

MW, molecular weight; BP, boiling point (or sublimation temperature in the case of CO_2); ODP, ozone depletion potential; GWP, global warming potential; VOC, volatile organic compounds; TLV, threshhold limit value; ppm, parts per million.
(From Refs.[11–14].)

by January 1, 2003, while other products had until 2005. Recently, attention has been focused on the GWP of the blowing agents. This is an area that is unsettled at the present time, but is expected to become a significant consideration in the use of certain blowing agents. An especially important point to consider is that the best alternatives for desirable insulating gases are HFCs.[9,10,17] The HFC alternatives have zero ODP and, furthermore, GWP values much lower than those of hard CFCs or HCFCs. Dow and BASF have independently developed polystyrene foam formulations that do not contain HCFC or HFC blowing agents.[18–22] Because these foams do not contain captive blowing agents that have desirable insulating properties, infrared attenuators have been employed to improve the thermal conductivity of the foam products.[23,24] Certain organic compounds generate smog photochemically (VOC). Because halogenated HCs have low reactivity in the lower atmosphere, substitution of photochemically reactive compounds for the current blowing agents may reduce the ozone depletion in the stratosphere, but could adversely affect the indoor air quality. Therefore, interaction with the total environment must be considered in developing environmentally acceptable blowing agents.

Solubility and Permeability

The solubility of a blowing agent in a polymer at any temperature and pressure can be approximately predicted by using the Flory–Huggins equation as a first approximation:[25–29]

$$\ln a = \ln v_o + (1 - 1/m_n)v_p + xv_p^2 \quad (1)$$

where

$$a = f/f^o = p/p^o \quad (2)$$

and a = activity, f = fugacity, f^o = fugacity of pure blowing agent, p = vapor pressure above the polymer solution, p° = vapor pressure of pure blowing agent, v_o and v_p = volume fraction of blowing agent and polymer, respectively, m_n = number average chain length of the polymer, and x = Flory–Huggins interaction parameter, which can be expressed as

$$x = 0.34 + V_o(d_o - d_p)^2/RT \quad (3)$$

where $V_o$is the molar volume of the blowing agent, and d_o and d_p are the solubility parameter of blowing agent and polymer, respectively. For high molecular weight polymers, $m_n$becomes very large and $(1 - 1/m_n)$ is reduced to 1. For blowing agent mixtures, the above equations may be expressed as follows:[26–28]

$$\ln a_m = \ln v_m + (1 - 1/m_n)v_p + xv_p^2 \quad (4)$$

where

$$a_m = f_m/f_m^\circ = p_m/p_m^\circ \quad (5)$$

$$p_m = \sum_{i=1}^{n} p_i y_i \quad i = 1, 2, 3, \ldots, n \quad (6)$$

$$p_m^\circ = \sum_{i=1}^{n} p_i^\circ y_i \quad i = 1, 2, 3, \ldots, n \quad (7)$$

$$x = 0.34 + V_m(d_m - d_p)^2/RT \quad (8)$$

$$v_m = \sum_{i=1}^{n} v_i \quad i = 1, 2, 3, \ldots, n \quad (9)$$

$$v_o + v_p = 1 \quad (10)$$

$$v_m + v_p = 1 \quad (11)$$

$$1/V_m = v_1/V_1 + v_2/V_2 + \cdots \quad (12)$$

$$d_m = d_1v_1 + d_2v_2 + \cdots \quad (13)$$

Here, Scott's ''single-liquid'' approximation was used in evaluating d_m and V_m values. This is a good approximation if the interaction of blowing agents with the polymer is very small as compared to the interaction between the two blowing agents.[30] For more accurate approximation, the equation of state proposed by Sanchez and Lacombe or by Panayiotou and Vera may be used but may require experimentally determined interaction constants for specific compounds.[31–33] For most engineering purposes, however, the Flory–Huggins equation may be sufficient for predicting the solubility of potential blowing agents in a given polymer. Accurate prediction of solubility of blowing agent in a molten polymer flowing through the extrusion process is very difficult because of the fact that the blowing agent and molten polymer mixture are not in an equilibrium state but in a dynamic state. Therefore, dynamic solubility rather than equilibrium solubility is required to predict the solubility of potential blowing agents in the extrusion process. This may be one of the reasons why it is so difficult to predict the rheology, pressure drop in the process equipments, foam nucleation, and foam expansion rate. In the mid-1970s, Suh developed a concept and coined the term ''critical dynamic system pressure'' over the blowing agent/polymer mixture in the extrusion process in Dow and used the dynamic solubility for the process control in producing low-density extruded

foam products of good quality.[34,35] The critical dynamic system pressure is defined as the minimum system vapor pressure required to keep the equilibrium amount of blowing agent dissolved in the molten polymer, which is considerably higher than the equilibrium vapor pressure above the solution. Details of this predictive method are not within the scope of this chapter. Permeability of blowing agents through plastic foams depends on the diffusivity of the blowing agent through the foam, its solubility in the polymer, open cell content, foam density, and cellular structure.[36–40] The permeability of the gaseous blowing agents through the plastic membrane can be expressed as a product of diffusion constant and solubility in the polymer as follows:

$$P = DS \tag{14}$$

Permeation of the blowing agent through a plastic membrane can usually be treated as an activated process with an activation energy E_p and permeability at the standard or room temperature

$$P(T) = P_o \exp(-E_p/RT) \tag{15}$$

There are many empirical correlations and theoretical treatments for predicting solubility, diffusivity, and permeability constants.[41] Table 2 shows the permeability and solubility data for gases in polystyrene and low-density polyethylene. In the extrusion process, the polymer is plasticated in the extruder and the blowing agent is mixed with molten polymer, which is cooled down to the foaming temperature and expanded against atmospheric or subatmospheric pressure to produce a low-density extruded foam. The pressure inside the cells decreases from the system pressure at the beginning of the foam expansion to slightly higher than ambient pressure as the foam expands to the final density. As the foam temperature decreases, the foam begins to solidify and reaches ambient temperature. The cell pressure experiences a partial vacuum as a result of PVT relationship. If a blowing agent or any major component of the blowing agent mixture has a permeability higher than that of air, the cell pressure decreases further before increasing as the air starts to permeate into the foam. On the other hand, if a blowing agent or any major component of the blowing agent mixture has a permeability much lower that that of air, the cell pressure starts to increase continually and reaches a maximum pressure considerably higher than the atmospheric pressure when the foam becomes saturated with one atmosphere of air and then the cell pressure decreases to atmospheric pressure as the low-permeability blowing agent diffuses out during the aging period. Therefore, one has to consider dimensional stability during aging at these cell pressure extremes depending on the permeability of the blowing agent used and also on the strength of the polymer. If the cellular polymer does not have enough strength to withstand these pressure extremes, the foam

Table 2 Permeability and solubility data for gases in polystyrene and low-density polyethylene (density = 920 kg/m^3)

Gas	Polystyrene permeability[a,b] [$\times 10^{-13}$ cm^3 (STP) cm/cm^2/sec/cmHg]	Polystyrene solubility[b] (pph/atm)	Low-density polyethylene permeability[c] (nmol/m/sec/GPa)
N_2	282	0	287
O_2	2055	0	963
Air	611	0	413
CO_2	8569	0.4	3430
CFC-11	3.77	13.6	2640
CFC-12	2.64	3.3	946
CFC-114	NA	NA	473
HCFC-22	167	3.0	1310
HCFC-141b	NA	NA	NA
HCFC-142b	0.42	5.5	NA
HCFC-124	1.43	2.3	NA
HFC-152a	NA	NA	NA
HFC-134a	5.03	1.8	NA

[a](From Refs.[8–10].)
[b]Allied-Signal data.
[c](From Ref.[13].)

tends to collapse at the minimum cell pressure and also tends to expand at the maximum cell pressure. The permeability of a blowing agent through polyethylene can be modified to improve the dimensional stability of the foam by combining the polymer with low-permeability polymers or by adding a small amount of a partial ester of a fatty acid to the polymer.[42–45] These formulation and process innovations have allowed Dow to make blowing agent changes from CFC-114 to HCFC-142b in 1989, which in turn was replaced with hydrocarbon blowing agent in 1994.[46–48] Hydrocarbons have several drawbacks including flammability, high permeability, and VOC concerns.

Organic PBA

The original concept of cellular polystyrene is credited to the Swedish inventors Munters and Tandberg, who filed a patent on ''Foamed Polystyrene'' in 1931, and subsequently obtained U.S. patent 2,023,204 in 1935 using methyl chloride, CO_2, and others. In 1941, The Dow Chemical Company began to develop a commercial process known as the Tower Process for the production of cellular polystyrene using a low-boiling compound such as butylene or methyl chloride.[49] In the late 1940s through 1950s, the first continuous extrusion process for producing a polystyrene foam was developed, and it is still the basis for the manufacture of polystyrene foams and polyolefin foams developed by Rubens using CFC-114 as well as other thermoplastic foams.[50,51] In 1963, Dow converted the process to a mixed blowing agent containing methyl chloride and CFC-12 to improve the dimensional stability of polystyrene foam board by attempting to match the permeability of blowing agent mixture with that of air through the foam. Polyurethane foams were developed by Otto Bayer and associates in Germany in the early 1940s using water-generated carbon dioxide as a blowing agent.[52] This blowing agent was subsequently replaced with CFC-11 to improve the cell uniformity and the K-factor of polyurethane foam or polyisocyanurate foam. The first cellular synthetic plastic was phenol–formaldehyde resin produced unintentionally by Baekeland and others in this field.[53] The elimination of cell formation produced by steam generated during the reaction of these resins, as described by Baekeland in his 1909 patent, is generally considered the beginning of the plastic industry. Foams produced in this way have nonuniform and coarse cell structures with large voids or blowholes. Therefore, volatile liquids or CBAs are incorporated into phenolic foam formulations to control cell size, cell morphology, and foam density. Volatile liquids such as CFC-113 and other polyhalogenated saturated fluorocarbons with a boiling point between 30°C and 100°C were preferred blowing agents for producing phenolic foams with more uniform, finer cell sizes having improved flame retardancy and thermal insulation value.[54] The ideal CFC replacement must be non-VOC, nontoxic, and must have zero ODP and a short atmospheric lifetime to ensure a low GWP. The general strategy of The Dow Chemical Company was to develop a two-stage replacement program in co-operation with Pennwalt Corp. (now AtoFina), Du Pont, and Allied Chemical Corp. (now Honeywell). The first stage of the program was to incorporate at least one hydrogen in the proposed CFC substitute's structure, which provides a destruction mechanism via hydrogen atom abstraction by troposphere hydroxyl radicals.[9,10] The HCFCs, especially HCFC-141b, HCFC-142b, HCFC-22, were selected as the optimum intermediate replacements for CFC-11 and CFC-12, respectively.[10,15,16] The second stage of the program was to replace all chlorine atoms in the structure with hydrogen (HFC-134a, -143a, -152a) to obtain zero ODP.[17] The HFCs such as HFC-245fa, -365mfc, and HFC-134a, -143a, -152a have been considered as the optimum replacements for HCFC-141b and HCFC-142b, respectively.[55]

Hydrocarbons with C_1–C_6 and chlorocarbons such as methyl chloride and ethyl chloride have been also used for producing plastic foams as a single blowing agent or as a coblowing agent to facilitate the processability and to provide the performance and structural properties required in the end-use applications. Some of these compounds have concerns with flammability, VOC, toxicity, or cellular structures. There are many techniques to introduce fluorine into organic compounds, but hydrogen fluoride (HF) is considered to be the most economical source of fluorine for many industrial applications. Although HF is a relatively ineffective exchange agent when used alone, it readily replaces only active halogens as shown below for the commercial production of CFCs and HCFCs by the successive replacement of chlorine with fluorine using HF:[56]

$$CHCl_3 \xrightarrow{HF} CHCl_2F + CHClF_2 + CHF_3 \quad (16)$$

$$CCl_4 \xrightarrow{HF} CCl_3F + CCl_2F_2 + CClF_3 \quad (17)$$

$$CCl_3CCl_3 \xrightarrow{HF} CCl_2FCCl_3 + CCl_2FCCl_2F + CClF_2CCl_2F + CClF_2CClF_2 \quad (18)$$

$$CH_3CCl_3 \xrightarrow{HF} CH_3CCl_2F + CH_3CClF_2 + CH_3CF_3 \quad (19)$$

Hydrofluorocarbons are also produced from acetylene or olefin and HF. A commercial production of 1,1-difluoroethane, HFC-152a, is conducted in the

vapor phase over an aluminum fluoride catalyst:

$$C_2H_2 + 2HF \xrightarrow{\text{catalyst}} CH_3CHF_2 \quad (20)$$

Inorganic PBA

The most widely used inert inorganic PBAs are CO_2,N_2, and water. Although these compounds are environmentally acceptable, each has some drawbacks as a single blowing agent. CO_2 and N_2 have very low solubility in most polymers and zero heat of vaporization at their foaming temperatures. Therefore, they require very high system pressure to dissolve enough blowing agent to produce a low-density foam. They also tend to produce a very small cell size and ''hot foams'' at the center of the foam board, which makes it very difficult to produce an extruded low-density foam with a large cross section without possible thermal collapse at the center of foam boards of very low density. The CO_2 blowing agent was first used commercially by The Dow Chemical Company in 1982 as a coblowing agent to produce the Styrofoam brand polystyrene insulation foam products.[15] Since then, Dow has led the development of blowing agent technology and the process innovations necessary for successful implementation of HCFC-142b globally to reduce the ODP and subsequently HFC-134a or 100% CO_2 blowing agent systems in some countries of Europe to entirely eliminate ozone-depletion concerns.[17,19,20,34] The use of cell size enlarging agents may be necessary with HFC-134a or 100% CO_2 blowing agents to control the cell size for different applications.[57,58] Water blowing agent is corrosive, especially when used with a halogenated flame retardant. It also lacks the blowing power to produce low-density extruded thermoplastic foams. However, it should be considered as an alternative blowing agent because of its environmental acceptability and low cost.[23,59–62] Water can also be used for producing polyurethane foams by reacting with isocyanates to generate urea and CO_2, which acts as a foaming agent. This water-generated CO_2 may increase the density and flammability of the resultant polyurethane foams.

Mixtures of organic and inorganic PBAs

There are advantages and disadvantages in using one or more organic PBAs or one or more inorganic PBAs. Organic PBAs provide processability and performance properties, whereas inorganic PBAs provide better environmental acceptability at the expense of processability and performance properties. However, mixtures of organic and inorganic PBAs may provide the best possible options to minimize the environmental issues while maintaining the performance properties required in end-use applications. Mixtures of water or CO_2 with HCs, chlorocarbons, HCFCs, or HFCs may be such examples.[15,21,22]

CHEMICAL BLOWING AGENT

Chemical blowing agents are normally used in producing high-density structural foams such as polyolefin insulation materials for coaxial wires and cables, polystyrene extrusion profiles, sheets, pipes, and other flexible or semiflexible cellular polymers such as cellular rubbers, latex foams, PVC foams, etc. In chemical foaming process, the expanding gas is generated in situ as a by-product of a chain extension or cross-linking of the polymer, as in the decomposition of CBA by adding CBAs to the polymer under heat, or as in the formation of polyurethane foams or polyisocyanurate foams by adding water reacted with isocyanate groups generating CO_2, or as in the condensation of phenol and formaldehyde producing water as a by-product in the form of steam, which is capable of foaming the resulting phenolic resin. The process for producing low-density polyolefin foams normally has three separate stages: sheet formation, cross-linking, and foaming. In real situations, there may not be three distinct separate stages depending on the process and the type of polymer, but there may be some overlaps and sometimes concurrent operations may take place at different rates to control the process and product properties. The selection of a suitable CBA depends on the type of polymer and process used, the type of cross-linking, radiation or chemical cross-linking, and the shape and form of the final foam products desired. The key desired properties of suitable blowing agents are given below:

1. The decomposition temperature of the CBA should be higher than the melting point (T_m) or glass transition temperature (T_g) of the polymer. The blending operation of CBA and the polymer should be conducted at a temperature sufficiently above T_m (close to the order–disorder temperature $T_{od} = T_m$ or $T_g + 55°C$) but preferably below the decomposition temperature of the CBA.
2. The amount of gas yield should be large.
3. The gas should be released within a narrow temperature range.
4. The rate of gas release should be rapid, and controllable by temperature, pressure, and the type and concentration of decomposition activators.
5. The decomposition products of CBA should be noncorrosive, nontoxic, nonflammable, colorless, and environmentally acceptable.

6. The CBA and the gas released should be readily dispersed in the polymer.
7. The decomposition products of CBA should not affect the rates of nucleation, polymerization, cross-linking, and decomposition of CBA.

Various CBAs are used in producing cellular polymers. They are either organic or inorganic compounds that decompose under the influence of heat to yield at least one gaseous decomposition product. They are either exothermic or endothermic but the effect of heat release or heat intake should not be large enough to cause polymer degradation or polymer temperature control. Table 3 shows the physical properties of potential CBA.

Organic CBA

The most widely used organic CBA is azodicarbonamide (ADC), also known as azobisformamide (ABFA). The decomposition of this compound is first order according to the following reaction:

$$H_2N–CO–N=N–CO–NH_2 \longrightarrow NH_3 + CO + N_2 + HNCO \quad (21)$$

For all practical purposes, ADC and its residues are odorless, nondiscoloring, and have low flammability, good storage stability, and good dispersability in all elastomers and plastics. Although the ADC decomposition temperature is relatively high, it can be modified by adding numerous activators such as heavy metal salts to reduce the decomposition temperature range to as low as 130–220°C from 190–220°C. This decomposition range makes it particularly suitable for producing a wide range of cellular polymers including styrenic and olefinic polymers, vinyls, rubbers, and elastomers. Because the gas yield of ADC is the highest among commercially available organic blowing agents, it is very economical to use for many applications. However, caution should be used in the generation of NH_3, which may cause rusting and staining of molds as well as polymer degradation. Various grades of modified ADC are available from the manufacturers for improving flow characteristics, dispersability, mold plate-out, and storage stability. Different particle sizes are also available from the manufacturers for improving dusting, cell size control, activator efficiency, and compatibility with polymer matrix. Both activators and particle size affect the decomposition rate. The smaller the particle size, the more effect activators will have on the decomposition rate. Common examples of strong activators are zinc oxide, zinc stearate, dibasic lead phthalate, and dibasic lead phosphite, while dibutyl tin dilaurate and barium/cadmium stearate combinations are mild activators. Table 4 shows the effect of activators on the decomposition temperature of ADC. Among other aliphatic azocompounds, azobisisobutyronitrile and dinitrosopenta-methylene tetramine may deserve some consideration. The former has limited applications owing to toxicity of the residue, and the latter is a widely used blowing agent in the rubber industry but of limited use in the plastics industry owing to its high decomposition exotherm and the unpleasant odor of its residue.[11] However, *N*,*N*′-dimethyl-*NN*′-dinitrosoterephthalamide has a low-decomposition exotherm, which is very suitable for producing thick cellular vinyl polymers without concern for the polymer degradation inside the thick sections of the foam caused by the high-decomposition exotherm of most organic CBAs. Of the many commercially available sulfonyl hydrazides, only 4,4′-oxybis(benzenesulfonyl hydrazide) (OBSH) is of commercial importance in the extrusion of cellular polyethylene for wire insulations, in casting PVC foams, and in rubber-resin blends.[11] It is also used to simultaneously expand and cross-link the polymer.

Inorganic CBA

The most important inorganic CBAs are sodium bicarbonate and sodium borohydride, which decompose by endothermic reaction, generating water/CO_2 and hydrogen, respectively. The former takes advantage of the chemical reactivity of inorganic bicarbonates with acidic compound to generate carbon dioxide, which is much faster than the thermal decomposition of sodium bicarbonate in the absence of acidic compound. The latter is based on the reaction with water or steam to generate copious amounts of hydrogen gas at very high rates. The most common acid and bicarbonate systems are based on sodium bicarbonate and citric acid, which are available as 100% powders or as concentrates in low molecular weight polystyrene or polyethylene carriers. The concentrate products are available in up to 70% of active concentrates and may contain various additives to improve the dispersability, cell size uniformity, screw slippage problem, and so on. The acid/bicarbonate systems are typically used as nucleators and also as blowing agents in the extrusion process as well as in the injection molding process or structural foam molding process.

The other bicarbonate system of some importance is ammonium bicarbonate, which decomposes in the range of 40–60°C. This material is difficult to process owing to the low decomposition temperature and high rate of decomposition with liberation of strong ammonia odor in the extrusion or injection molding processes. However, ammonium bicarbonate can be

Table 3 Physical properties of potential chemical blowing agents

Blowing agent	Trade symbol	Decomposition temperature (°C)	Gas yield (cm^3/g)	Decomposition products		Trade name (supplier)
				Gas type	Residue type	
Azodicarbonamide (1,1′-azobisformamide)	AZ (ABFA)	200–215	220	36 wt%	64 wt% Biuret, cyanuric acid, cyamelide, urea, triuret	Kempore (Uniroyal)
				N_2 65%		Celogen (Uniroyal)
				CO 32%		Porofor (Mobay/Bayer)
				CO_2 3%		Azofoam (Biddle Sawyer)
						Unicel (Dong Jin)
						Azo (Dong Jin)
						Ficel (Schering Berin)
N,*N*′-Dimethyl-*N*, *N*′-dinitrosoterephthalamide	DTA	90–105	126	N_2	Esters of terephthalic acid	Nitrosan (Du Pont)
Dinitrosopenta-methylenetetramine	DNPT	200–205	240	N_2, NO	Water, formaldehyde, amines	Opex (Uniroyal)
						DNPT (Dong Jin)
						Unicell (Dong Jin)
						Prespersion (Synthetic Products)
						Warecel (Santech)
4,4′-Oxybis(benzene-sulfonyl hydrazide)	OBSH	155–160	125	N_2	Water, aromatic sulfur polymer	Azocel (Fairmount)
						OBSH (Dong Jin)
						Celogen OT (Uniroyal)
						Unicel OT (Dong Jin)
						Santechem (Santech)
Sodium borohydride	$NaBH_4$	—	2370	H_2	$NaBO_2$	
Sodium bicarbonate/citric acid	$NaHCO_3$	100–210	100–160	$CO_2 \cdot H_2O$	Sodium citrate	Hydrocerol (Henley)
						Safoam (Reedy)

(From Refs.[12,13].)

Table 4 Activators for ADC

Activator system	Decomposition temperature (°C)	Polymers
None	200–220	PP, PS, HDPE, EVA, ABS, SMA, polyolefin copolymers, blends, cross-linked polymers
Silica/ZnO	190	LDPE, EVA, LLDPE
Dinitrosopentamethylene tetramine	155	EVA, LDPE, ionomer
Zinc sulfide	170	Polyolefin, LLDPE, HDPE, ionomer

PP, polypropylene; PS, polystyrene; HDPE, high-density poylethylene; EVA, ethylene vinyl acetate; ABS, acrylonitrile–butadiene–styrene; SMA, styrene maleic anhydride; LDPE, low-density polyethylene; LLDPE, linear low-density polyethylene. (From Ref.[12].)

used as a blowing agent and cross-linking agent that consumes ammonia, eliminating the odor problems associated with ammonia liberation. Caution should be used in handling ammonium bicarbonate for the possible liberation of ammonia during storage and its effect on health and corrosive properties.

SOLID ADSORBENT CONTAINING VOLATILE LIQUID

Many solid adsorbents liberate gas as a result of desorption of volatile liquids under the influence of heat. Typical adsorbents with microporous structures such as activated carbons, or precipitated silicas and renewable resources have been used as a coblowing agent in producing low-density extruded polystyrene foam boards.[63–65] Incorporation of corn cobs or other renewable vegetable matter containing about 10% water together with a primary PBA into polystyrene in the extrusion process produced a low-density polystyrene foam board with bimodal cellular structures. This type of foam with bimodal cell structures has about 10–15% lower K-factor than similar foams without bimodal cellular structures.[65] Similar results were obtained with a precipitated silica for producing a low-density extruded polystyrene foam with bimodal cellular structures.[63,64]

SAFETY AND ENVIRONMENTAL ASPECTS

Flammability

Among the blowing agents, HCs and some of the HCFCs and HFCs are flammable and pose a fire hazard in handling at the manufacturing plants. Plastic foams are organic in nature and, therefore, are combustible. All plastic foams containing flammable blowing agents should be handled, transported, and used according to the manufacturers' recommendations and local and national regulations.

Atmospheric Emissions

Certain organic compounds are found to be smog generating substances (VOC) because of their high photochemical reactivity at ambient conditions, Substitution of photochemically reactive compounds for the current blowing agents may reduce the ozone depletion in the stratosphere, but may have unacceptable GWP and air quality in the lower atmosphere. Therefore, the interaction of the blowing agent with the total environment needs to be considered in developing environmentally acceptable alternative blowing agents.[66]

CONCLUSIONS

In the past decade, significant advances have been made in the selection of environmentally acceptable alternative PBAs for a given polymer to comply with government regulations and local codes. However, there has been very little change in the use of CBAs except that much improvement has been made in the areas of product quality and formulation flexibility.

Some of the important factors to consider in the development of alternative physical blowing agents are solubility, permeability, boiling point, vapor pressure, flammability, toxicity, VOC, ODP, GWP, availability, and cost. Similarly, the key important factors to be considered in the selection of CBAs for a given polymer are decomposition temperature, gas yield, rate of gas release, environmentally acceptable decomposition products, toxicity, and dispersability in the polymer matrix.

Finally, the processability, performance properties, and cellular morphology of the products play key roles

in the successful commercialization of alternative blowing agents that are environmentally sustainable.

REFERENCES

1. Molina, M.J.; Rowland, F.S. Stratosphere sink for chlorofluoromethanes—chlorine atomic-catalysed destruction of ozone. Nature (Lond.) **1974**, *249*, 810.
2. Stolarski, R.S.; Cicerone, R.J. Stratospheric chlorine: a possible sink for ozone. Can J. Chem. **1974**, *52*, 1610.
3. Crutzen, P. Estimates of possible future ozone reductions from continued use of fluorochloromethanes (CF_2Cl_2, $CFCl_3$). Geophys. Res. Lett. **1974**, *1*, 205.
4. Palmer, A.R. *Economic Implications of Regulating Chlorofluorocarbon Emissions from Nonaerosol Applications*, R-2524-EPA, a report prepared for the U.S. EPA; Rand Corp.: Santa Monica, CA, 1980.
5. Junger, H. Kunstst. Rundsch. **1962**, *9*, 437.
6. Goggin, W.C.; McIntire, O.R. Br. Plast. **1947**, *19* (223), 528.
7. Randolph, A.F., Ed. *Plastics Engineering Handbook*, 3rd Ed.; Reinhold Publishing Corp.: New York, 1960; 137–138.
8. Nakamura, M. The Dow Chemical Co.; U.S. Patent 3,960,792, 1976.
9. Suh, K.W.; Killingbeck, G.W. The Dow Chemical Co.; Canadian Patent 1,086,450, 1980.
10. Suh, K.W.; Killingbeck, G.W. The Dow Chemical Co.; British Patent 1,537,421, 1978.
11. Lasman, H.R. *Modern Plastics Encyclopedia*; McGraw-Hill Book Co.: New York, 1965; Vol. 66, 394.
12. Throne, J. Foaming agents. In *Thermoplastic Foams*; Sherwood Publishers, 1996; 125.
13. Park, C.P. Polyolefin foam. In *The Handbook of Polymeric Foams and Foam Technology*; Hanser Publishers, 1991; 187–242.
14. Smart, B.E.; Fernandez, R.E. Fluorinated aliphatic compounds. In *Kirk-Othmer Encyclopedia of Chemical Technology*; 1996; Vol. 11, 506.
15. Suh, K.W.; Kennedy, J.M. The Dow Chemical Co.; U.S. Patent 4,636,527, 1987.
16. Suh, K.W.; Severson, J.L. The Dow Chemical Co.; U.S. Patent 4,916,166, 1990.
17. Suh, K.W. The Dow Chemical Co.; U.S. Patent 5,011,866, 1991.
18. Suh, K.W.; Yamada, M.; Shmidt, C.D.; Imeokparia, D.D. The Dow Chemical Co.; U.S. Patent 5,679,718, 1997.
19. Paquet, A.N.; Priddy, D.B.; Vo, C.V.; Pike, W.C.; Hahnfeld, J.L. The Dow Chemical Co.; U.S. Patent 5,650,106, 1997.
20. Vo, C.V.; Paquet, A.N. The Dow Chemical Co.; U.S. Patent 5,389,694, 1995.
21. Voelker, H.; Gerhard, A.; Schuch, H.; Weilbacher, M.; Weber, R. BASF; U.S. Patent 5,182,308, 1993.
22. Voelker, H.; Gerhard, A.; Weilbacher, M.; Weilbacher, F.; Weber, R.; Schuch, H. BASF; U.S. Patent 5,334,337, 1994.
23. Paquet, A.N.; Suh, K.W. The Dow Chemical Co.; U.S. Patent 5,210,105, 1993.
24. Bartz, A.M.; Hitchcock, M.K. The Dow Chemical Co.; U.S. Patent 5,373,026, 1994.
25. Flory, P.J. Statistcal thermodynamics of polymer solutions. In *Principles of Polymer Chemistry*; Cornell University Press: Ithaca, NY, 1953, ch. 12, 495 pp.
26. Suh, K.W.; Clark, D.H. Cohesive energy densities of polymers from turbidimetric titrations. J. Polymer Sci. A-1 **1967**, *5*, 1671.
27. Suh, K.W.; Corbett, J.M. Solubility parameters of polymers from turbidimetric titrations. J. Appl. Polymer Sci. **1968**, *12*, 2359.
28. Suh, K.W.; Liou, D.W. Phase euqlibria in polymer-liquid 1-liquid 2 systems. J. Polymer Sci. A-2, **1968**, 813.
29. Hildebrand, J.H.; Scott, R.L. High polymer solutions. In *The Solubility of Nonelectrolytes*; Rheinhol Publishing Corp.: New York, 1950, ch. 20, 346 pp.
30. Scott, R.L. J. Chem. Phys. **1949**, *17*, 268.
31. Sanchez, I.C.; Lacombe, R.H. An elementary molecular theory of classical fluids. J. Phys. Chem. **1976**, *80*, 2352.
32. Sanchez, I.C.; Lacombe, R.H. Statistical thermodynamics of polymer solutions. Macromolecules **1978**, *11*, 1145.
33. Panayiotou, C.; Vera, J.H. Polym. J. **1982**, *14*, 681.
34. Suh, K.W. Method for the Preparation of Styrene Polymer Foam, Blowing Agents for the Preparation of Such Foam and Styrene Polymer Foams Prepared with Such Blowing Agents. The Dow Chemical Co.; Canadian Patent 1,309,221, 1992.
35. Suh, K.W.; Hood, L.S. *Thermoplastic foam processing with supercritical blowing agents*, United Engineering Foundation Conference on Supercritical Fluids in Materials and Synthesis, Davos, Switzerland, Sep 26 to Oct 1, 1999.
36. Norton, F.J. J. Cell. Plast. **1967**, *3* (1), 23.
37. Cuddihy, E.F.; Moacanin, J. J. Cell. Plast. **1967**, *3* (2), 73.
38. Shankland, I.R. Adv. Foam Aging **1986**, *1*, 60.
39. Sheffield, C.F. *Description and application of a diffusion model for rigid closed cell foams*, ORNL Symposium on Mathematical Modeling of Roofs, Oak Ridge National Laboratory, Oak Ridge, TN, Sep 15, 1988.

40. Shankland, I.R. *The effect of cell structure on the rate of foam aging*, International Workshop on Long Term Thermal Performance of Cellular Plastics, SPI, Canada, Oct 1989.
41. Van Krevelen, D.W. *Properties of Polymers*, 2nd Ed.; Elsevier: Amsterdam, 1976.
42. Ehrenfreund, H.A. Polytech Systems, Inc.; U.S. Patent 3,810,964, 1974.
43. Cronin, E.W. Haskon, Inc.; U.S. Patent 3,644,230, 1972.
44. Watanabe, S.; Matsuki, Y. The Dow Chemical Co. U.S. Patent 4,343,913, 1982.
45. Park, C.P. The Dow Chemical Co.; U.S. Patent 4,714,716, 1987.
46. Kolosowski, P.A. The Dow Chemical Co.; WO Patent 92/19439, 1992.
47. Park, C.P The Dow Chemical Co.; U.S. Patent 4,528,300, 1985.
48. Park, C.P. The Dow Chemical Co.; U.S. Patent 4,640,933, 1987.
49. McIntire, O.R. The Dow Chemical Co.; U.S. Patent 2,450,436, 1948; 2,515,250, 1950.
50. McCurdy, J.L.; DeLong, C.E. The Dow Chemical Co.; U.S. Patent 2,669,751, 1954.
51. Rubens, L.C.; Griffin, J.D.; Urchick, D. The Dow Chemical Co.; U.S. Patent 3,067,147, 1962.
52. Backus, J.K. Rigid polyurethane foams. In *Handbook of Polymeric Foams and Foam Technology*; Klempner, D., Frisch, K., Eds.; Hanser Publishers: Munich, 1991.
53. Baekeland, L.H. U.S. Patent 942,699, Dec 7, 1909.
54. D'Alessandro, W.J. Union Carbide Corp.; U.S. Patent 3,389,094, 1968.
55. McCoy, M.C & EN. October 30, 2000, 22.
56. Hudlicky, M. Chemistry of organic fluorine compounds, 2nd Ed.; Ellis Horwood Ltd.: Chichester, U.K., 1976.
57. Suh, K.W.; Amos, C.R. The Dow Chemical Co.; U.S. Patent 4,229,396, 1980.
58. Suh, K.W.; Wakabayashi, M.; Vo, C.V.; Paquet, A.N. The Dow Chemical Co.; U.S. Patent 5,489,407, 1996.
59. Paquet, A.N.; Suh, K.W. The Dow Chemical Co.; U.S. Patent 5,240,968, 1993.
60. Paquet, A.N.; Suh, K.W. The Dow Chemical Co.; U.S. Patent 5,332,761, 1994.
61. Paquet, A.N.; Suh, K.W. The Dow Chemical Co.; U.S. Patent 5,369,137, 1994.
62. Paquet, A.N.; Suh, K.W. The Dow Chemical Co.; U.S. Patent 5,380,767, 1995.
63. Motani, S.; Saito, T.; Ito, T. Kaneka. U.S. Patent 4,990,542, Feb 1991.
64. Motani, S.; Saito, T.; Ito, T. Kaneka. U.S. Patent 5,064,874, Nov 1991.
65. Hurps, R.L.; Suh, K.W. The Dow Chemical Co.; U.S. Patent 4,559,367, Dec 17, 1985.
66. *Handbook for the Montreal Protocol on Substances That Deplete the Ozone Layer*, 2nd Ed.; Ozone Secretariat, UNEP: New York, Oct 1991.

Branching Level Detection in Polymers

M. J. Scorah
Institute for Polymer Research, Department of Chemical Engineering, University of Waterloo, Waterloo, Ontario, Canada

R. Dhib
Department of Chemical Engineering, Ryerson University, Toronto, Ontario, Canada

A. Penlidis
Institute for Polymer Research, Department of Chemical Engineering, University of Waterloo, Waterloo, Ontario, Canada

INTRODUCTION

Long-chain branching (LCB) affects a variety of polymer properties from dilute to melt state. Because many commercial polymers often contain a fraction of branched material, understanding the relationship between polymer properties and molecular architecture is of tremendous importance. This entry deals with the detection of branching in polymers in two parts. It begins with a description of the most important and the most well established solution and rheological properties influenced by branching. Solution studies have shown that a branched molecule is more compact than a linear analog and that various contraction factors can be used to quantify the reduction of a molecule's size because of branching. Rheological investigations have revealed two opposing effects attributed to LCB. Branching leads to a reduction in the size of a polymer molecule and, in turn, fewer entanglements. However, when the branches reach a critical length, the overall number of entanglements is increased. With an understanding of the impact of branching on polymer properties, methods to detect or estimate LCB can be devised. Methods based on spectroscopic, chromatographic, and rheological techniques are the focus of the second section of this entry. Although both spectroscopic and chromatographic methods are extremely useful in the characterization of polymers, they are unable to detect the very low levels of branching that can still alter a polymer's flow behavior. The drawback of rheological methods is that the rheology of such polymers is still not fully understood. As a result, the analysis by such methods is frequently qualitative.

BACKGROUND

The occurrence of branching was postulated, roughly 70 years ago, by Staudinger and Schulz to explain certain unexpected observations with polystyrene.[1] Flory later showed that transfer reactions during the free radical polymerization of styrene could produce long-chain branches.[2] Already by the early 1950s, a noteworthy amount of work dealt with the mechanism and kinetics of branching reactions, the effect of branching on polymer properties, and the possible determination of LCB. In 1953, Stockmayer and Fixman reviewed the state of dilute solution properties of branched polymers.[3] Since that era, numerous reviews of branching have been compiled. In 1968, Graessley summarized the methods for the detection of branching based on dilute solution methods.[4] In the same year, Dexheimer et al. compiled a list of studies dealing not only with the effects of branching on polymer properties but also the kinetics and mechanism of branching.[5] Small wrote a notable review of long-chain branching, examining its influence on various polymer properties, and its estimation.[6] More recent reviews include works by Burchard[7,8] and Mays and Hadjichristidis,[9] who examined the solution properties of branched macromolecules, and Roovers,[10] who provided a comprehensive survey of the literature for branched polymers in general. Reviews on the effect of branching in the melt state include those of Graessley[11] and Vega et al.[12]

Branched polymers can have various structures depending upon their synthesis. A great deal of research has involved the synthesis of model branched polymers[13,14] with narrow molecular weight distributions (MWD) to specifically observe the effect of branching. These results help to construct fundamental theories, which can be used in the investigation of randomly branched polymers.

BRANCHED POLYMER PROPERTIES

Dilute Solution Properties

Mean-square radius

The size of a macromolecule is one of its most fundamental properties. Although there are several ways to

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120027021

represent the dimensions of a polymer chain, the mean-square radius is a typical measure of size, given by the following:

$$\langle s^2 \rangle = \left\langle \sum_{i=1}^{N} \frac{r_i^2}{N} \right\rangle \quad (1)$$

where the polymer molecule is considered to be comprised of N small elements of identical mass and r_i is the distance of the ith unit from the polymer molecule's center of gravity. The use of angled brackets denotes that the summation is averaged over all possible conformations that the polymer chain can assume. The term radius of gyration is widely used when referring to a polymer molecule's size and is simply the square root of the mean-square radius:

$$R_g = \langle s^2 \rangle^{\frac{1}{2}} \quad (2)$$

Theoretical calculations for the mean-square radius of gyration usually assume that a polymer molecule can be represented by a random flight chain made up of N freely jointed units. Discrepancies between the model and actual chains arise for two reasons, known as short- and long-range effects. The short-range effects are because of units not being completely free to rotate but having bond restrictions, while long-range effects occur because intersections of units are impossible. Short-range effects are addressed by dividing the polymer chain into longer segments of several bonds so that each unit can be considered to be freely jointed. If long-range effects, also known as volume exclusion, are absent, then chains obey the random flight model and take an ''unperturbed'' state. In this case, the mean-square radius is represented by $\langle s_0^2 \rangle$ (the subscript 0 denotes an unperturbed value). From the work of Flory,[15] it can be seen that long-range effects are nonexistent at the θ point (particular temperature, T_θ, for a specific solvent) and

$$\langle s_0^2 \rangle = \langle s_\theta^2 \rangle \quad (3)$$

Compared to a linear chain of the same number of units, a branched chain is more compact. As a result, the impact of branching on the size of a polymer chain is to decrease the mean-square radius as branching increases. To assess the decrease in size because of branching, the mean-square radius of a branched polymer is compared to the size of a linear analog of identical molecular weight. Quantitatively, this was defined by Zimm and Stockmayer[16] with the following branching or contraction factor:

$$g = \left. \frac{\langle s^2 \rangle_{br}}{\langle s^2 \rangle_l} \right|_M = \left. \frac{R_{g\,br}^2}{R_{gl}^2} \right|_M \quad (4)$$

The subscript M indicates that both the branched (br) and the linear (l) chains have identical molecular weights. Because branched polymers are more compact and have smaller dimensions, g will always be less than unity with smaller values being an indication of a higher amount of branching. Theoretical equations for the calculation of contraction factors for various types of branching have been developed and are given in Table 1. Although it is not a complete list of the results in the literature, Table 1 does summarize the earlier work which provided expressions for more common types of branching. The equations for g given in Table 1 are based on the random flight model where chains are assumed to be in an unperturbed state. If Eq. (3) is assumed to hold, these equations can be compared to experimental values of g_θ, where the mean-square radius for both linear and branched chains is measured at the θ point.

Values of $\langle s^2 \rangle$ can be obtained experimentally from radiation or neutron scattering experiments. By measuring the angular dependence of the intensity of scattered radiation between the particles and the probing radiation, the particle size can be determined. Well developed techniques for determining a particle's size include light scattering (LS), small-angle x-ray scattering (SAXS), and small-angle neutron scattering (SANS).[17] The theory of light scattering from macromolecular solutions is provided in the works of Debye[18] and Zimm[19,20] and is reviewed and applied to branched polymers by Burchard.[8] Light scattering experiments provide a z-average estimate of the radius of gyration ($R_{g(z)}$). For samples with a narrow molecular weight distribution, this does not provide a problem and $R_g = R_{g(z)}$. However, when samples with a broad molecular weight distribution are analyzed, the increase in polydispersity (M_w/M_n) has a substantial effect and $R_{g(z)}$ will increase. This increase in the z-average radius of gyration is significant enough that it will counterbalance the decrease in size because of branching. As a result, $R_{g(z)}$ for a polydisperse branched polymer may seem identical to that of a polydisperse linear sample. Therefore, some fractionation method must be used to obtain monodisperse fractions where the contraction factor can be calculated. Further discussion concerning the determination of g for polydisperse samples is provided in the section for detecting LCB.

The radius of gyration can be related to molecular weight by an equation of the following form:

$$R_g = K_{R_g} M^v \quad (19)$$

where K_{Rg} and ν are constants. The exponent can vary between 0.33 for hard spheres and 1 for a rigid rod. In the case of linear chains, an exponent of 0.5 refers to an unperturbed state, while in good solvents, ν is closer to 0.6.

Table 1 Theoretical equations for mean-square radius contraction factor for several branched structures

Branching type	Theoretical branching factor	References
Regular stars		
Monodisperse arms	$g = \frac{3f - 2}{f^2}$ (5)	[16]
Polydisperse arms	$g = \frac{6f}{(f + 1)(f + 2)}$ (6)	[16]
	$\langle g \rangle_z = \frac{3f}{(f + 1)^2}$ (7)	[8]
Stars with unreacted functional groups		
Monodisperse arms	$g = \frac{1 + 3(f - 1)\alpha}{(2 + (f - 1)\alpha)^2}$ (8)	[21]
Polydisperse arms	$g = 4\frac{1 + (f - 1)\alpha}{(2 + (f - 1)\alpha)^2}$ (9)	[21]
Symmetrical combs	$g = \lambda - \frac{1}{n + 1}\lambda^2(1 - \lambda) + \frac{2}{n}\lambda(1 - \lambda)^2 + \left(\frac{3n - 2}{n^2}\right)(1 - \lambda)^3$ (10)	[22]
Random combs		
n is constant	$g = \lambda + \frac{2}{n}\lambda(1 - \lambda)^2 + \left(\frac{3n - 2}{n^2}\right)(1 - \lambda)^3$ (11)	[22]
n varies	$g = \lambda + \frac{3(1 - \lambda)^2}{\bar{n}} + \frac{(1 - \lambda)^3}{\bar{n}^2}$ (12)	[22]
Random branching		
n is constant	$g = 3\left(\frac{\pi}{2(f - 1)(f - 2)_n}\right)^{\frac{1}{2}} - \frac{2(6 - f)}{(f - 1)(f - 2)n}$ (13)	[16]
n varies	$g_3 = \left[\left(1 + \frac{\bar{n}}{7}\right)^{\frac{1}{2}} + \frac{4\bar{n}}{9\pi}\right]^{-\frac{1}{2}}$ (14)	[16]
	$g_4 = \left[\left(1 + \frac{\bar{n}}{6}\right)^{\frac{1}{2}} + \frac{4\bar{n}}{3\pi}\right]^{-\frac{1}{2}}$ (15)	
n varies polydisperse	$\langle g_3 \rangle_w = \frac{6}{n_w}\left[\frac{1}{2}\left(\frac{2 + n_w}{n_w}\right)^{\frac{1}{2}} \times \ln\left(\frac{(2 + n_w)^{\frac{1}{2}} + n_w^{\frac{1}{2}}}{(2 + n_w)^{\frac{1}{2}} - n_w^{\frac{1}{2}}}\right) - 1\right]$ (16)	
	$\langle g_4 \rangle_w = \frac{1}{n_w}\ln(1 + n_w)$ (17)	[16]
	$\langle g \rangle_z = \frac{1}{1 + \frac{(f - 1)(f - 2)n_w}{6}}$ (18)	

Notes: f is the functionality of the branch point (i.e., the number of chains attached to branch point); α is the number of unreacted functional groups; λ is the fraction of polymer in the backbone chain; n is the number of branch points per chain; $\bar{n}$ is the average number of branch points per chain; n_w is the weight-averaged number of branch points per chain; π is 3.14159 . . . ; g_3 and g_4 denote trifunctional and tetrafunctional branch points, respectively; subscripts n, w, or z represent the number, weight, and z-averages, respectively.

An excellent summary of $R_g - M$ data by Fetters et al. provides estimates for these parameters in both θ and good solvents for a number of linear polymers.[25] The parameters in Eq. (19) are not only influenced by the experimental conditions (solvent, temperature) but are also affected by the polymer's structure. For monodisperse stars, it has been found that increasing the number of arms decreases K_{Rg} while ν remains identical to that of the linear polymer (see Fig. 1). However, for randomly branched polymers, it has been found that ν is closer to 0.5 and in some cases much lower. Such low values of this exponent might be considered as an indication of the branched polymer being in an unperturbed state; however, this is not the case. An explanation as to why ν is so small for randomly branched polymers has been found using fractal behavior, and an overview is given by Burchard.[8]

Intrinsic viscosity

The increase in viscosity with the addition of a polymer into a solvent is an important property. By measuring the solution viscosity as a function of polymer concentration, useful information about the polymer's molecular properties can be determined. From the solution data, the intrinsic viscosity (also known as the limiting viscosity number or Staudinger index) can be calculated:

$$[\eta] = \lim_{c \to o} \frac{1}{c} \left(\frac{\eta}{\eta_0} - 1 \right) \tag{20}$$

where c is the concentration of polymer in solution, η the solution viscosity, and η_0 the viscosity of the pure solvent. The presence of branching leads to smaller intrinsic viscosity values. Comparable to Eq. (4), a branching factor can be also defined using the intrinsic viscosity:

$$g' = \left. \frac{[\eta]_{\mathrm{br}}}{[\eta]_l} \right|_{\mathrm{M}} \tag{21}$$

Because of the ease in measuring intrinsic viscosity relative to the radius of gyration, considerably more experimental works have reported intrinsic viscosity data for branched molecules. As a result, attempts have been made to relate the two contraction factors. However, the efforts to find an encompassing relationship have not been completely successful. Thurmond and Zimm[26] proposed the following equation:

$$g' = g^{\varepsilon} \tag{22}$$

with a value of 1.5 for ε, but found results supporting and opposing the use of such a value for their polystyrene star samples. Zimm and Kilb[27] later came to the numerical conclusion that $\varepsilon = 0.5$ for certain star polymers. Burchard proposed that the relationship between the two branching factors could not be adequately described by a simple power law and proposed the following equation for star polymers:[8]

$$g' = (a + (1 - a)g^p)g^{\varepsilon} \tag{23}$$

with $a = 1.104$, $p = 7$, and $\varepsilon = 0.906$.

Overall, studies have found that ε varies with experimental conditions and the type of branching.

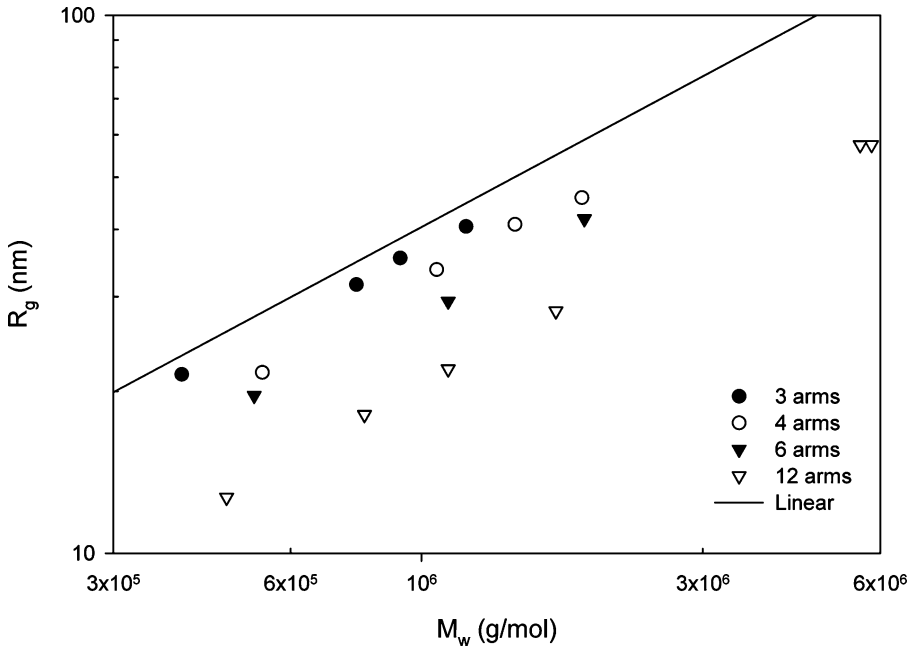


Fig. 1 Radius of gyration as a function of molecular weight for monodisperse polystyrene stars. (Data from Refs.[23,24]. Parameters for linear polystyrene from Ref.[25].)

Typically, stars have been found to have the lowest values ($\varepsilon = 0.5$), while combs are in the upper limit ($\varepsilon = 1.5$). Results for randomly branched polymers tend to fall somewhere in between these limits. However, these are only general trends and upon examination of the results for polyethylene, a variety of values for ε can be shown (see Table 2). For randomly branched polystyrene with tetra-branch points, values ranging from 0.5 in cyclohexane to 0.72 in toluene have been found.[28] Further work found that styrene copolymerized with divinylbenzene produced polymer with an exponent varying from 0.65 at low conversion to 1.41 at higher conversions.[29] Berry and Orofino determined that for their poly(vinyl acetate) combs in benzene, ε varied from 0.5 to 1.0, while randomly branched poly(vinyl acetate) in benzene exhibited a range of 0.7–0.9.[30] Foster et al. also found $\varepsilon = 1$ for randomly branched poly(vinyl acetate) in tetrahydrofuran.[31] Values of ε greater than 1.5 have been observed, including the work with styrene–butadiene graft copolymer where ε was estimated to be 2.[32] An extensive review on the dilute solution properties of star polymers has found that over a large range of functionalities, no constant exponent could be found.[33]

The intrinsic viscosity in turn can be related to a molecular weight with a power law expression known as the Kuhn–Mark–Houwink–Sakurada (KMHS) equation:

$$[\eta] = K_{[\eta]} M^{\alpha} \tag{24}$$

Similar to the behavior of $R_g - M$ for stars, the exponent α does not vary with star functionality (see Fig. 2). However, for randomly branched polymers, it is found that α is smaller when compared to the exponent for a linear polymer. In fact, α decreases with increasing molecular weight for randomly branched polymers. Because of the nature of branching reactions in randomly branched polymers (e.g., transfer to polymer), the number of branches per macromolecule increases with molecular weight. It is for this reason that α decreases with increasing molecular weight.

Rheological Properties

Rheological properties of polymers are highly sensitive to macromolecular structure. In some cases, the incorporation of minute amounts of branching can alter the flow properties of a polymer even though the dilute solution properties are unchanged. The rheological behavior of branched polymers is influenced by a variety of factors including the number, location, architecture, and length of branches.[11] Similar to dilute solution properties, the differences between the viscoelastic behaviors of linear and branched polymers can be masked by large polydispersities.

There has been a considerable amount of research on the rheological properties of polymers with uniform model structures such as stars and combs with narrow molecular weight distributions. These studies provide insight into the effects of branching. The presence of branching has two opposing effects. First, branching is known to decrease the size of polymer molecules compared to a linear chain of identical molecular weight. In such a case, a smaller size will result in fewer chain entanglements. Second, at a critical molecular weight, the branches become long enough to get entangled and the overall number of entanglements increases. It is because of these two opposing effects that the impact of branching on rheological properties is highly dependent upon the nature of branching present in the polymer.

Table 2 Values of ε for polyethylene[a]

Group	ε	Conditions	Method	References
Foster et al.	0.75	TCB at 140°C	SEC with calibration curve and Eq. (14) to calculate average number of branches. This is compared with average number of branches obtained from NMR data and best agreement gives ε	[31]
Hert and Strazielle	0.7–0.9; 0.9–1	Autoclave reactor; tubular reactor; TCB at 135°C	SEC-Visc to determine g' and fractions collected from SEC analyzed with MALLS to obtain g	[34]
Grinshpun et al.	0.68–0.88	TCB at 145°C	SEC-LALLS and NMR data to calculate ε	[35]
Tackx and Tacx	1–1.5; 1.2–1.8	Autoclave reactor; tubular reactor; TCB at 140°C	SEC-MALLS to calculate g and KMHS equation with universal calibration to get g'	[36]
Wang et al.	0.2–1.8	TCB at 135°C	SEC-MALLS-Visc to obtain g and g'	[37]

TCB, 1,2,4-trichlorobenzene; SEC, size exclusion chromatography; NMR, nuclear magnetic resonance; LALLS, low-angle laser light scattering; MALLS, multiangle laser light scattering; Visc, viscometer; KHMS, Kuhn–Mark–Houwink–Sakurada equation.
[a]Selected references after 1980.

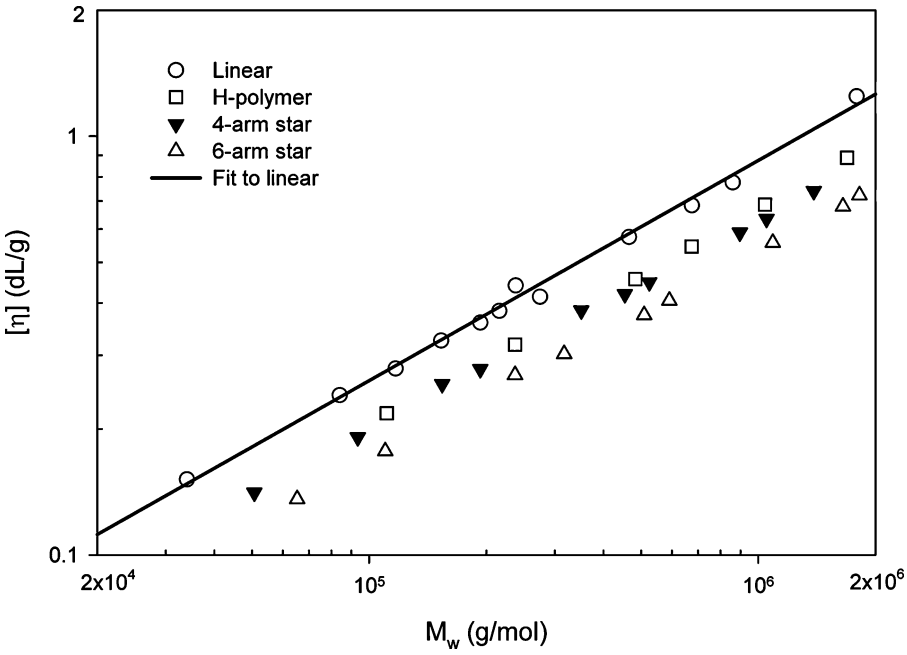


Fig. 2 Intrinsic viscosity as a function of weight-average molecular weight for branched polystyrenes. (From Refs.[38,39].)

Zero-shear properties

The zero-shear viscoelastic properties of concentrated polymer solutions or polymer melts are typically defined by two parameters: the zero-shear viscosity (η_0) and the zero-shear recovery compliance (J_e^o). The former is a measure of the dissipation of energy, while the latter is a measure of energy storage. For model polymers, the influence of branching is best established for the zero-shear viscosity. When the branch length is short or the concentration of polymer is low (i.e., for solution rheology), it is found that the zero-shear viscosity of the branched polymer is lower than that of the linear. This has been attributed to the smaller mean-square radius of the branched chains and has led to the following relation:[40,41]

$$\eta_{0\mathrm{br}} = g^a \eta_{0l} \begin{cases} a = 1 & \text{when } M < M_c \\ a = 3.4 & \text{when } M > M_c \end{cases} \qquad (25)$$

where g is the ratio of mean-square radii of the branched and linear polymers of identical molecular weight in the unperturbed state, and M_c is a critical molecular weight where the exponent in Eq. (25) changes from 1 (unentangled linear polymers) to 3.4 (entangled linear polymers). Experimental evidence of a smaller zero-shear viscosity for branched polymers has been shown for a variety of polymer architectures such as stars,[38,42] combs,[43,44] and randomly branched polymers.[45] In some cases, a reduction in zero-shear viscosity has been observed at higher polymer concentrations[42] and even in the melt state.[11]

In certain cases, the behavior of branched polymers cannot be described by Eq. (25). These situations can arise for higher molecular weights, higher polymer concentrations, longer branch lengths, or longer spacing between branch points (i.e., situations where branching can increase the number of chain entanglements). For these conditions, it has been found that the zero-shear viscosity of the branched polymer can be considerably higher than that of the linear. This viscosity enhancement has been detected for several polymers of varying branching structure such as randomly branched polybutadiene,[46] randomly branched poly(vinyl acetate),[47] star poly(α-methylstyrene),[48] comb polystyrene,[48,49] H-shaped polystyrene,[39] star polyethylene,[50] comb, regular and irregular stars, H-shaped and pom-pom polyethylene,[51] and star polyisoprene.[42,52] In Fig. 3, the behavior of zero-shear viscosity vs. molecular weight for a series of H-shaped polystyrenes is shown. At low molecular weights and hence smaller branch lengths, there is a viscosity reduction, while at higher molecular weights, there is a viscosity enhancement.

To understand this viscosity enhancement, it is easier to start with the theory for linear polymers. The behavior of linear polymers can be described by the reptation model.[53,54] For a linear polymer of high molecular weight in the melt, chains can be modeled as a confined tube where the diffusion of the chain is restricted along the tube contour. Entanglements are formed between chains where the reptation of a chain along its contour becomes the dominant mode of movement. The addition of a branch point prevents reptation and other forms of movement must occur for the chain to change its configuration. In the case

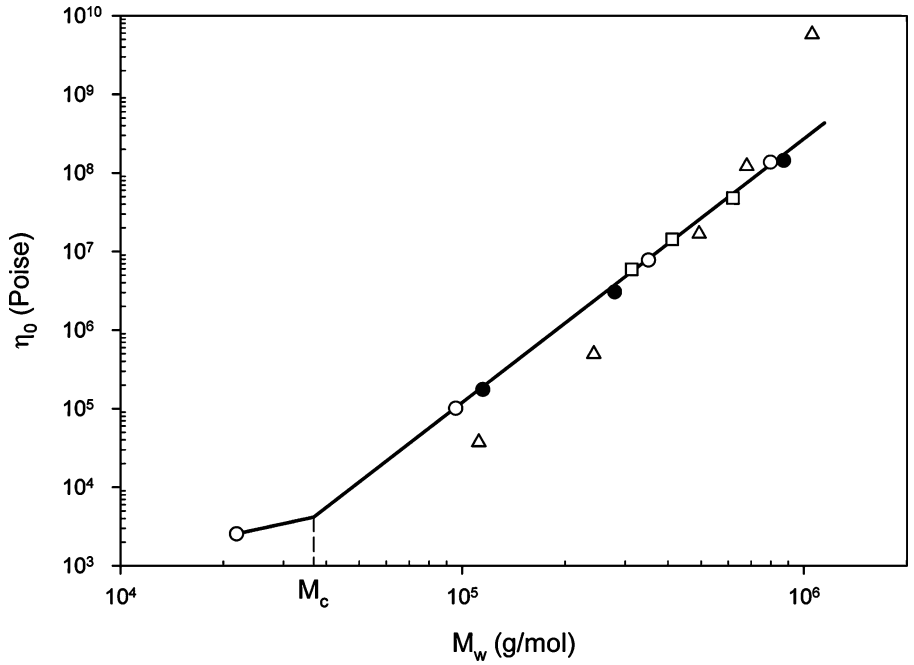


Fig. 3 Zero-shear viscosity at 169.5°C as a function of weight-averaged molecular weight for polystyrene melts. Open and closed circles and squares, linear; triangles, H-polymers. (From Ref.[39].)

of a star polymer, the arms retract partway down its contour tube and then move outward along a different trajectory.[54] This process is slow relative to that of linear polymers and is influenced by the branching length. For longer branches, there are a higher number of entanglements per branch, leading to a higher zero-shear viscosity. In the case of star polymers, it has been found that the viscosity increases exponentially with arm molecular weight and that the effect of functionality is no longer significant for $f > 4$ (i.e., total star molecular weight does not influence viscosity; only arm molecular weight does).[52] For star polymers, the exponential dependence of zero-shear viscosity on the number of entanglements per arm is given by:

$$\eta_o \propto \left(\frac{M_a}{M_e}\right)^{1/2} \exp\left(v' \frac{M_a}{M_e}\right) \tag{26}$$

where M_a is the arm molecular weight, M_e the molecular weight between entanglements, and ν' a constant. To properly describe the dynamics of branched melts, other effects aside from arm retraction that were not included in the original treatment of linear polymers, such as contour length fluctuations and dynamic tube dilation are important.[55] In the case of comb and H-shaped polymers, the length of branches is not the only factor as it has been found that the length between branch points is a significant variable.[43]

The zero-shear recovery compliance is another viscoelastic property of polymers that is noticeably influenced by the presence of branching, and similar to viscosity, the effect of branching is dependent upon several factors including polymer concentration and molecular weight. In dilute solutions, J_e^o is expected to decrease with branching according to:[40]

$$J_e^o = \frac{2g_2M}{5cRT} \tag{27}$$

where c is the concentration of polymer and g_2 is defined as follows:

$$g_2 = \left.\frac{\left[\frac{\langle s^4\rangle - \langle s^2\rangle^2}{\langle s^2\rangle^2}\right]_{br}}{\left[\frac{\langle s^4\rangle - \langle s^2\rangle^2}{\langle s^2\rangle^2}\right]_{l}}\right|_M \tag{28}$$

Similar to other branching factors, g_2 is unity for linear polymers and less than 1 for branched polymers. Equations for the determination of g_2 for certain branching structures have been determined including those for uniform stars and combs.[56] This reduction in J_e^o has been observed for low molecular weight star polymers at low and high concentrations and for high molecular weight stars at low concentrations.[42] However, in most instances, J_e^o is much larger for branched polymers compared to a linear polymer of the same molecular weight[39,42,44–47,57] and Eq. (27) does not accurately predict the behavior of a branched polymer. Based on the theory used to derive Eq. (26), a similar relationship for the zero-shear compliance and molecular

weight of star polymers is expressed as follows:

$$J_e^o = v' \frac{M_a}{M_e} \frac{1}{G_N^o} \quad (29)$$

where G_N^o is the plateau modulus which has been found to be independent of branching.[57] As well, the enhancement of η_0 and J_e^o does not necessarily coincide at the same polymer concentration and some polymers (e.g., polystyrene) are less prone to enhancement than others.[11]

DETECTION OF LONG-CHAIN BRANCHING

For the most part, detection of long-chain branching relies upon indirect methods that compare the properties of a branched polymer to the corresponding linear polymer. However, spectroscopic or chemical methods can quantitatively determine the number of branches or end groups without needing a linear reference. The results from different methods typically do not agree as each technique has a different criterion as to how long a branch must be so as to be considered ''long.'' Long-chain branches are usually defined as being comparable in length to the main chain. Using polyethylene as an example, it has been found that size exclusion chromatography (SEC) can detect a minimum length for a straight chain alkane branch greater than six and less than twelve carbons.[35] It is generally accepted that six carbons or longer can be detected by ^{13}C-nuclear magnetic resonance (NMR) spectroscopy.[58] In the case of rheology, a definite length is not completely established. It has been proposed that the length of an LCB corresponds to the critical entanglement molecular weight.[59] For polyethylene, this equates to 270 carbon atoms or greater (reported values of M_c vary from 2100 to 5200[12]). Other sources have found that branches longer than twice the molecular weight between entanglements are considered long in rheological terms.[60] In the case of polyethylene, this corresponds to 180 carbon atoms in length. Because of this variability in the methods of detecting LCB, agreement between techniques can be considered fortuitous.

Spectroscopic Methods

Spectroscopic methods rely on the chemical difference between end groups or branch points in a polymer. In the case of polyethylene, the most common spectroscopic technique for determining long-chain branching has been nuclear magnetic resonance spectroscopy. Using ^{13}C-NMR, branches of one to five carbon atoms can be distinguished, while a six-carbon-atom branch produces the same spectral pattern as any subsequent branch of greater length.[58] Peak assignments for various chemical shifts in polyethylene have been well documented.[61] Fig. 4 is a typical ^{13}C-NMR spectrum of polyethylene showing the various peak assignments).[62] Recent studies employing NMR have reported long-chain branching as low as 0.2 branches per 10,000 carbon atoms in polyethylene.[62] Although NMR is seen as an absolute technique for determining LCB in polyethylene, studies have found limitations when compared to methods utilizing rheological measurements. A series of commercial high-density polyethylenes (HDPE) with similar molecular weight and

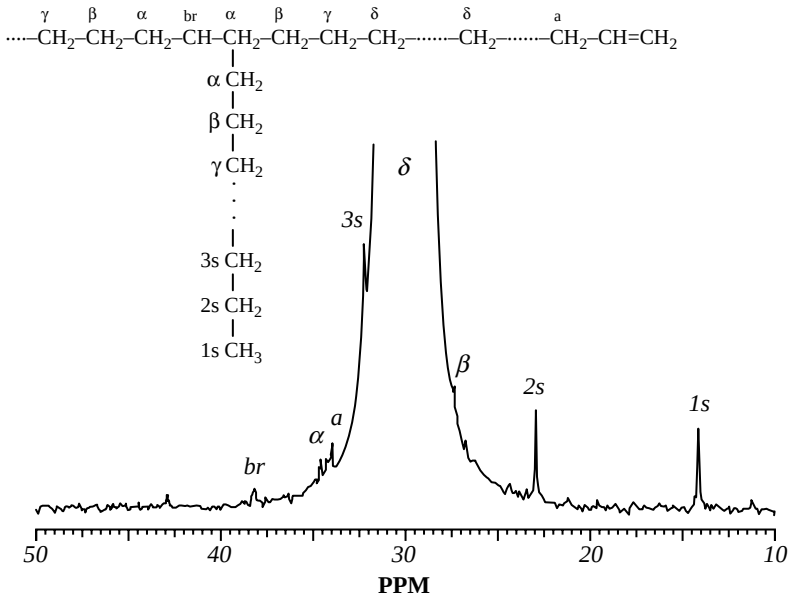


Fig. 4 ^{13}C-NMR spectrum of polyethylene sample measured at 120 °C using deuterium *o*-dichlorobenzene and 1,2,4-trichlorobenzene as solvents. (From Ref.[62].)

molecular weight distributions were found to have large and systematic increases in zero-shear viscosity, which were attributed to long-chain branching.[59] However, the number of branches determined with NMR showed random scatter. In the same study, it was found that NMR could not detect the presence of long-chain branching in peroxide modified high-density polyethylene, nor could it identify LCB in multiple pass extruded HDPE. In both cases, samples showed large changes in their rheological behavior relative to the unmodified samples.

Chromatographic Methods

Grubisic et al. proposed that SEC separates polymer molecules, regardless of chemical composition and large-scale structure, but according to a hydrodynamic volume, V_H.[63] They found this hydrodynamic volume to be proportional to the product of intrinsic viscosity and molecular weight:

$$V_H \propto [\eta]M \tag{30}$$

The excellent correlation between hydrodynamic volume and ($[\eta]M$) has become the basis of the universal calibration curve (UCC) for polymers. Although other size parameters, namely those dependent on R_g, have been suggested, the applicability of UCC has recently been verified for a variety of polymers including those with a high chemical and molecular weight asymmetry (e.g., miktoarm stars where arms are of different composition).[64]

Methods for determining long-chain branching in polymers through the use of SEC depend upon the type of detectors employed, and thus the type of information collected. For conventional SEC equipped with only a concentration detector, a calibration curve is constructed where molecular weight is a function of elution volume. For a linear and branched macromolecule of the same molecular weight, it is known that the branched macromolecule will have a smaller intrinsic viscosity ($[\eta]_{br} < [\eta]_l$) and as a result, it will elute later than the linear macromolecule. Because of this, SEC with only a concentration detector will underpredict the molecular weight of branched polymers. Using the KMHS relation for the linear polymer, two branching ratios can be defined for SEC:[65]

$$g'_{[\eta]} = \left(\frac{[\eta]_{br}}{[\eta]^*}\right)^{\alpha+1} \tag{31}$$

$$g'_M = \left(\frac{M^*}{M_{br}}\right)^{\alpha+1} \tag{32}$$

where $[\eta]^*$ and M^* are the intrinsic viscosities and molecular weight of linear molecules leaving the column at the same elution volume as the branched molecules having intrinsic viscosity and molecular weight of $[\eta]_{br}$ and M_{br}. Several workers have applied an iterative method to derive the branching density (λ) based on SEC chromatograms.[66–68] Assuming a value of λ, the number of branches is calculated as a function of molecular weight ($n = \lambda M$) and using the appropriate equation, g can be determined for each molecular weight (see Table 1). Then assuming a value for ε, g' and then $[\eta]_{br}$ can be calculated for each elution volume. The total intrinsic viscosity can be determined as a weighted sum of the individual elution volumes. This total intrinsic viscosity calculated from SEC data can be compared to the measured value. The process is repeated by changing λ until the two values agree. Problems with this method arise from the various assumptions made, including constant values of λ and ε. Several studies have found that the branching density varies with molecular weight for a particular sample and a constant value of ε cannot be assumed.[29,33,37] Another problem lies in the fact that theoretical equations for g are for chains in an unperturbed state, best represented by a θ solvent and temperature. Typically, SEC measurements are made in good solvents. For these reasons, values of the branching density based on this method are only valid for relative comparisons.

Most SEC studies of branching employ at least one other detector aside from a concentration detector. In the past, the most common additional detector for measuring long-chain branching was an online viscometer. By using UCC and a viscometer, both molecular weight and intrinsic viscosity distributions can be determined for a polymer sample. Low-angle light scattering detectors with UCC have also been employed for the determination of molecular weight and intrinsic viscosity distributions, and in turn, the detection of long-chain branching. However, the use of a viscometer has been found to be more appropriate for studying branched polymers. If a branched sample has a distribution in the extent or type of branching, the fractionation with SEC will not be complete. It is possible for two polymer molecules, one more branched than the other, to have the same hydrodynamic volume ($[\eta]M$) and coelute. The more highly branched chain will have a smaller intrinsic viscosity, but it may also have a higher molecular weight. In this case, the detector cells do not contain monodisperse fractions, and the results from the detectors at each elution volume are average results. Light scattering is known to provide a weight-average molecular weight estimate and from the work of Hamielec and Ouano,[69] it was determined that UCC provides the number-average molecular weight for

polymer molecules of the same hydrodynamic volume but differing molecular weight.

With the intrinsic viscosity distributions obtained from either a viscometer or light scattering detector, previous studies have attempted to quantitatively determine the amount of long-chain branching. First, g' was related to g using an assumed value of ε, and the appropriate equation from Table 1 was used to calculate the number of branches per molecule. However, because of the assumptions involved, it is more appropriate to determine the level of branching in a more qualitative manner. From the molecular weight and intrinsic viscosity data collected, a KMHS plot can be constructed for each polymer sample and comparisons can be made to the linear polymer. In Fig. 5, an example of a KMHS plot for branched polystyrene compared to a linear sample obtained using SEC is shown. As the level of branching increases, the $[\eta] - M$ relationship can no longer be expressed as a linear relation in a log–log plot. Therefore, deviations

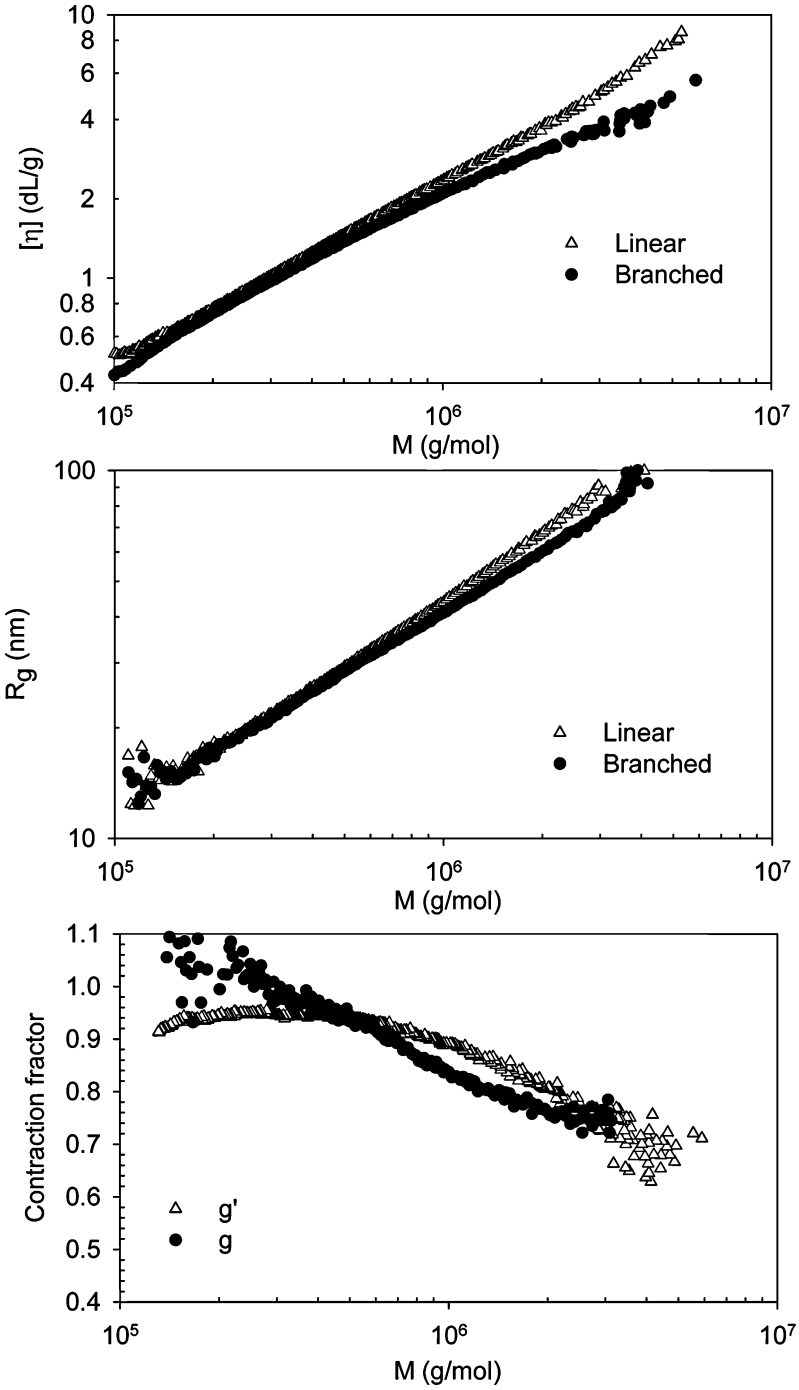


Fig. 5 Intrinsic viscosity, radius of gyration, and contraction factors as a function of molecular weight for polystyrene (SEC at 30°C with 1.0 ml/min of tetrahydrofuran). (From Ref.[74].)

from the KMHS plot for a linear sample are an indication of branching, with more curvature being an indication of higher levels of branching.

The introduction of commercial low angle laser light scattering (LALLS) detectors for SEC in the 1970s was followed by multiangle laser light scattering (MALLS) detectors in the 1980s. Coupled with SEC, a MALLS detector allows for the determination of the radius of gyration as a function of molecular weight for a polymer sample. In plots of $R_g - M$, the presence of branching is detected by lower trends for branched polymers compared to a linear sample (see Fig. 5). The use of MALLS also allows for the determination of g directly and the branching density can be determined from a suitable equation from Table 1. Most studies now involve the combination of three detectors in the analysis of branched polymers (concentration, light scattering, and viscosity detectors). Mourey et al. have reported on a unique method to analyze branched polymers with triple detector SEC.[70–73] Because light scattering reports a weight-average molecular weight, and viscometry with UCC provides an estimate of the number-average molecular weight, it is possible to determine the local polydispersity for each elution volume. However, significant molecular weight heterogeneity is needed between the types of branched macromolecules to detect any local polydispersity.[73]

Rheological Methods

The most common rheological method is to compare the zero-shear viscosity of the branched polymer to that of the linear. It is well established that for linear polymers above the critical molecular weight for entanglement, zero-shear viscosity scales with molecular weight to a power of 3.4. For the most part, η_0 is independent of the molecular weight distribution as long as there is no significant fraction of low molecular weight material below the entanglement threshold.[75,76] Thus, deviation from the power law for linear samples implies the presence of branching. Whether there is a reduction or enhancement of zero-shear viscosity depends upon polymer molecular weight and branching characteristics (number, length, and architecture). Gabriel and Münstedt completed a study on the long-chain branching in polyethylene where reduction or enhancement was observed in various samples.[76] For highly branched low-density polyethylene (LDPE), a reduction in zero-shear viscosity was observed as high branching densities (i.e., a high number of branches of relatively short length) led to a lower radius of gyration and fewer entanglements. However, for metallocene-catalyzed polyethylenes with very low levels of long-chain branching, a significant increase in η_0 was observed and attributed to a higher number of entanglements from longer branches.

Recently, a model has been proposed for the relationship between zero-shear viscosity and weight-average molecular weight to also account for the effect of branching.[77] Although the model has limitations and uncertainties in some parameters, it was found to predict the reduction and enhancement in the $\eta_0 - M$ behavior of branched polyesters and branched polyethylenes. The authors attribute an increase in η_0 to cases where the molecular weight between branch points is much larger than the entanglement molecular weight, while decreases in η_0 are found in polymers with high branching densities where the molecular weight between branch points is small. The main limitation of the model is to account for varying types of branching within a polymer sample. In Fig. 6, it is illustrated how the number of branches per chain can influence the zero-shear viscosity based upon this model.

Unlike the zero-shear viscosity, the zero-shear compliance is significantly influenced by the molecular weight distribution, particularly the high molecular weight fractions. For linear polymers, increases in polydispersity correspond to an increase in elasticity (increase in J_e^o). In the case of branched polymers, the zero-shear compliance can experience a reduction or enhancement depending on a number of factors including molecular weight, number, length and type of branching, and molecular weight between branch points. Fewer studies rely on the detection of long-chain branching from zero-shear compliance estimates.[76] This is most likely because of the difficulties in separating the effects of polydispersity and LCB.

Shroff and Mavridis have completed a number of studies on the determination of branching in polyethylene.[59,78,79] The authors have reviewed and proposed several indices relying on rheological data to quantify the amount of long-chain branching including what they have termed as the long-chain branching index (LCBI):[78]

$$\text{LCBI} = \frac{\eta_0^{1/a_3}}{[\eta]_{\text{br}}} \frac{1}{K_3^{1/a_3}} - 1 \tag{33}$$

where k_3 and a_3 are the constants for a linear polymer in a $\eta_0 - [\eta]$ power law equation as follows:

$$\eta_0 = K_3[\eta]^{a_3} \tag{34}$$

The LCBI is zero for linear polymers with larger numbers denoting higher levels of long-chain branching. A significant advantage of this index is that it is independent of molecular weight and molecular weight distribution. The LCBI has been developed for essentially linear polymers, and its derivation relies upon

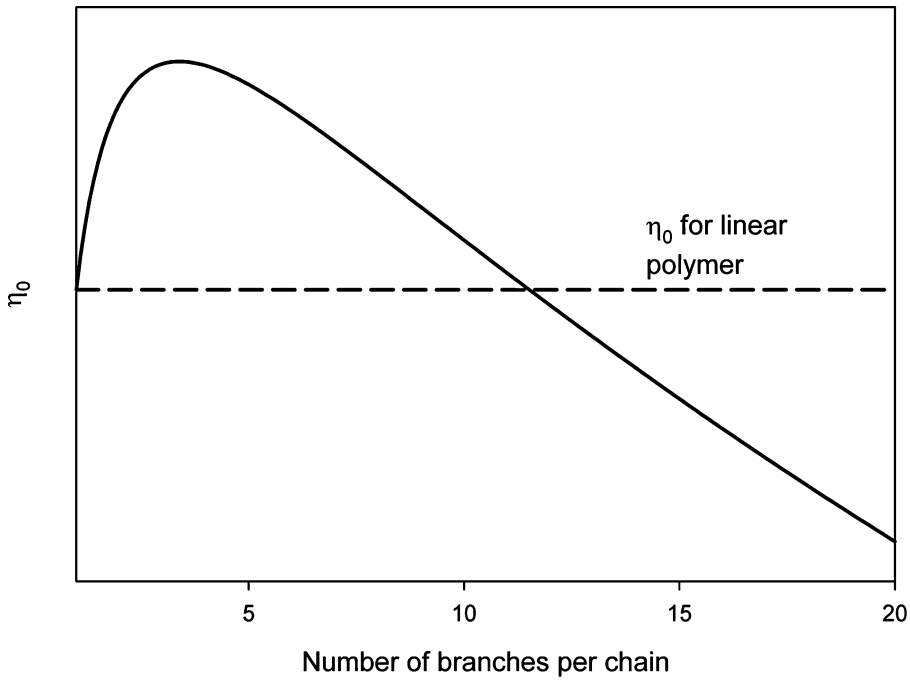


Fig. 6 Zero-shear viscosity as a function of the number of branches per chain for polyethylene. (Based on Ref.[77].)

the assumption that $g \approx g' \approx 1$. In other words, the LCBI is only valid for low levels of long-chain branching not detected by solution methods. A modification of Eq. (33) has been proposed to accommodate those polymers where $g' < 1$. However, this method relies on assuming a value for ε, and as previously noted, this parameter is found to vary with long-chain branching and molecular weight. As Shroff and Mavridis show a series of low-density polyethylenes where branching was detected with solution methods, the expected trend of samples with increasing LCB is only found at $\varepsilon = 0.5$ and not for other values of ε reported for polyethylene in the literature.[78] Shroff and Mavridis have also reviewed and extended a list of rheological polydispersity indices.[79] It was found that the presence of even small amounts of long-chain branching significantly alters most of the measures of rheological polydispersity, while no changes were observed in the MWD obtained from SEC. One such measure proposed by the authors is E_R, which is independent of molecular weight but influenced by molecular weight distribution and long-chain branching:[79]

$$E_R = C_1 G'|_{\text{at } G''_{\text{ref}}} \tag{35}$$

where G''_{ref} is a reference low modulus value corresponding to low frequencies and C_1 is a constant. For polyolefin melts, the authors have suggested $G''_{\text{ref}} = 5000\,\text{dyn/cm}^2$ and $C_1 = 1.781 \times 10^{-3}\,\text{cm}^2/\text{dyn}$. In Fig. 7, an example of a $\eta_0 - M_w$ plot for various polyethylene samples where very small amounts of LCB produced an increase in the zero-shear viscosity but were not detected with SEC is provided. The samples categorized as having lower E_R values do not display as much viscosity enhancement as the samples with much higher values of E_R.

Another index developed to detect long-chain branching is the Dow rheology index (DRI).[80] It was specifically derived for polyethylenes with similar, narrow molecular weight distributions ($M_w/M_n \sim 2$) and is not able to distinguish between the effects of polydispersity and branching.

$$\text{DRI} = 3.65 \times 10^5 \frac{\tau_0}{\eta_0} - 0.1 \tag{36}$$

where τ_0 is a characteristic time determined from the Cross equation.

Another technique to calculate the amount of LCB in polymers was developed by Wood-Adams and Dealy.[81] The procedure involves transforming the complex viscosity data as a function of frequency into a molecular weight distribution. This distribution is then compared to that obtained from SEC and the difference in the location of peaks is correlated to a level of branching.

Although providing only a qualitative determination of the level of branching, comparing the linear viscoelastic data of samples from dynamic rheology measurements is a common technique. Plots of the loss angle, δ, as a function of frequency are altered because of branching:

$$\delta = \tan^{-1}\left(\frac{G''(\omega)}{G'(\omega)}\right) \tag{37}$$

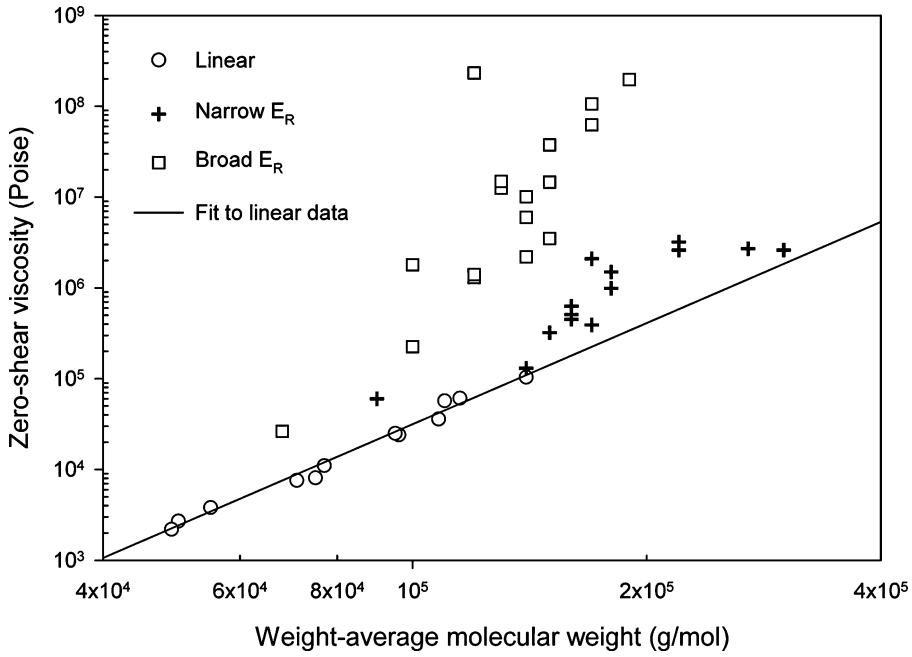


Fig. 7 Zero-shear viscosity as a function of weight-averaged molecular weight for various polyethylene samples at 190°C. (Circles) Fairly narrow MWD ($M_w/M_n \sim 2$–6) linear samples; (crosses) broad to very broad MWD ($M_w/M_n \sim 6.6$–30) branched samples but with narrow rheological polydispersity ($E_R \sim 1.7$–3.7); (squares) broad MWD ($M_w/M_n \sim 6.0$–14) branched samples with large rheological polydispersity ($E_R \sim 6.0$–14). (From Ref.[78].)

where $G''(\omega)$ and $G'(\omega)$ are the loss and storage moduli and ω is the angular frequency. For a perfectly elastic solid, the loss modulus is zero and so also is the loss angle. For a perfectly viscous fluid, the storage modulus is zero and the loss angle is 90°. In Fig. 8, a plot of the loss angle as a function of frequency for high-density polyethylenes from metallocene catalysts is provided. In general, the presence of branching causes a plateau in the loss angle with the magnitude and breadth of the plateau depending on the degree of LCB.[82] A variation of this technique is known as the reduced van Gurp–Palmen plot where the loss angle is plotted as a function of the ratio of the complex modulus and plateau modulus ($G_{red} = |G^*|/G^o_N$).[83,84] The plot is found to be temperature invariant and because a reduced modulus is used, the curves are not influenced by polymer composition. In Fig. 9, an example of a reduced van Gurp–Palmen plot is provided. For a linear polymer, the loss angle starts as a plateau at 90° for low G_{red}. As the reduced modulus increases, the loss angle decreases past an inflection point toward a minimum at $G_{red} = 1$ and then increases

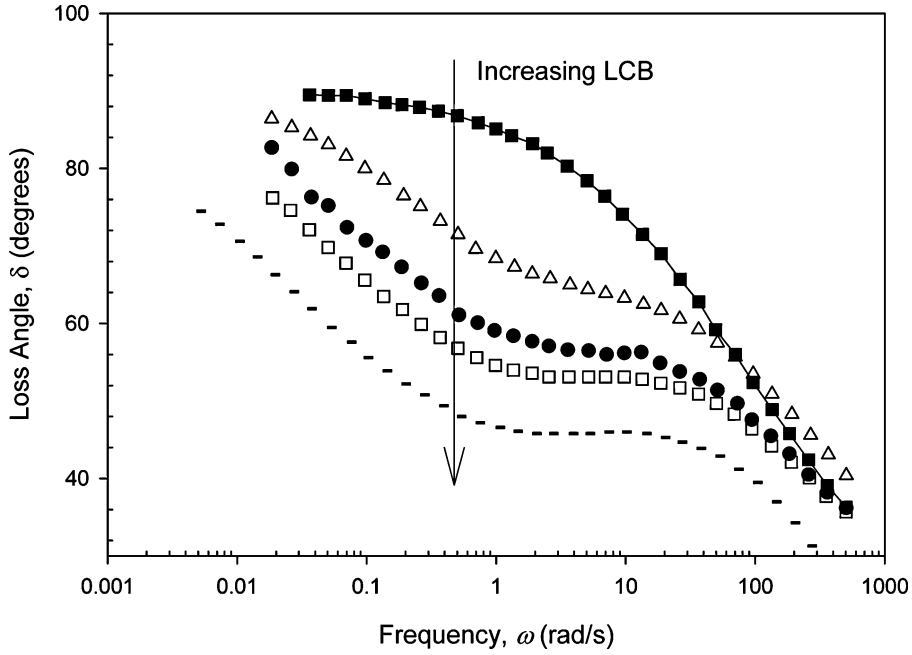


Fig. 8 Loss angle as a function of frequency at 150°C obtained for high-density, metallocene catalyzed polyethylene samples of increasing long-chain branching. (From Ref.[82].)

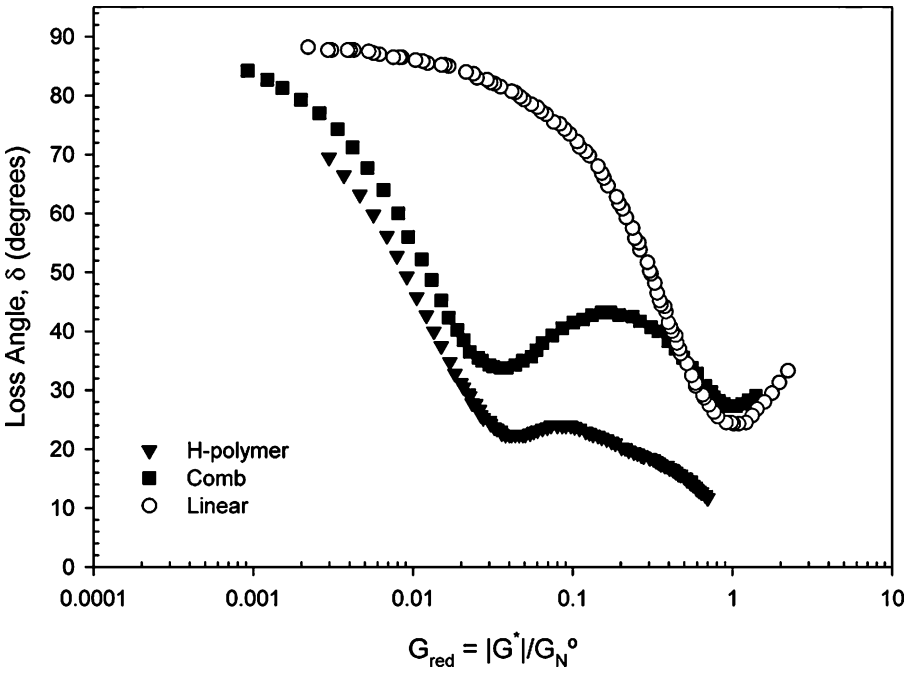


Fig. 9 Loss angle as a function of the reduced modulus (ratio of complex modulus and plateau modulus) for polystyrene samples. (From Ref.[84].)

again. The effect of molecular weight is to lower the value of the loss angle at the minimum, while a broader molecular weight distribution tends to cause a less steep drop in δ toward the minimum (i.e., δ begins to decrease from 90° at much lower values of G_{red}). The impact of branching is to produce a plateau or even a second minimum at the inflection point of $G_{red} < 1$. This presence of a second minimum has also been found for blends of linear high and low molecular weight materials.[83]

CONCLUSIONS

The task of quantifying branching in polymers is complicated by their very nature. Typical branched macromolecules are extremely heterogeneous in terms of the number, type, and length of branches, molecular weight variations, and in most instances, the presence of linear or cross-linked fractions. Spectroscopic methods such as NMR are some of the best methods for quantitatively determining the level of branching. Chromatographic techniques can provide more information such as molecular weight and branching distributions depending upon the detectors employed. Yet both spectroscopic and chromatographic methods are not sensitive enough to be able to detect the very low levels of branching that can still significantly alter a polymer's flow behavior. In such instances, rheological characterization becomes the only solution. However, the rheology of branched polymers is still not fully understood and the techniques typically provide a qualitative view on branching.

REFERENCES

1. Staudinger, H.; Schulz, G.V. Highly polymerized compounds. CXXVI. Comparison of osmotic and viscosimetric molecular-weight determinations of polymer-homologous series. Ber. **1935**, *68B*, 2320–2335.
2. Flory, P.J. Mechanism of vinyl polymerizations. J. Am. Chem. Soc. **1937**, *59*, 241–253.
3. Stockmayer, W.H.; Fixman, M. Dilute solutions of branched polymers. Ann. N.Y. Acad. Sci. **1953**, *57* (4), 334–352.
4. Graessley, W.W. Detection and measurement of branching in polymers. In *Characterization of Macromolecular Structure*; National Academy of Sciences: Washington, DC, 1968; 371–388.
5. Dexheimer, H.; Fuchs, O.; Suhr, H. Branching of macromolecules and its effect on properties. Deut. Farben-Z. **1968**, *22* (11), 481–494.
6. Small, P.A. Long-chain branching in polymers. Adv. Polym. Sci. **1975**, *18*, 1–64.
7. Burchard, W. Static and dynamic light scattering from branched polymers and biopolymers. Adv. Polym. Sci. **1983**, *48*, 1–124.
8. Burchard, W. Solution properties of branched macromolecules. Adv. Polym. Sci. **1999**, *143*, 113–194.

9. Mays, J.W.; Hadjichristidis, N. Dilute solution properties of branched macromolecules. J. Appl. Polym. Sci.: Appl. Polym. Symp. **1992**, *51*, 55–72.
10. Roovers, J. Branched polymers. In *Encyclopedia of Polymer Science and Engineering*; Wiley: New York, 1983; 478–499.
11. Graessley, W.W. Effect of long branches on the flow properties of polymers. Acc. Chem. Res. **1977**, *10*, 332–339.
12. Vega, J.; Aguilar, M.; Peon, J.; Pastor, D.; Martinez-Salazar, J. Effect of long chain branching on linear-viscoelastic melt properties of polyolefins. E-Polymers **2002**, *46*, 1–35.
13. Barner, L.; Barner-Kowollik, C.; Davis, T.P.; Stenzel, M.H. Complex molecular architecture polymers via RAFT. Aust. J. Chem. **2004**, *57* (1), 19–24.
14. Hadjichristidis, N.; Pitsikalis, M.; Pispas, S.; Iatrou, H. Polymers with complex architecture by living anionic polymerization. Chem. Revs. **2001**, *101* (12), 3747–3792.
15. Flory, P.J. *Principles of Polymer Chemistry*; Cornell University Press: Ithaca, NY, 1953.
16. Zimm, B.H.; Stockmayer, W.H. The dimensions of chain molecules containing branches and rings. J. Chem. Phys. **1949**, *17* (12), 1301–1314.
17. Chu, B.; Lui, T. Characterization of nanoparticles by scattering techniques. J. Nanopart. Res. **2000**, *2*, 29–41.
18. Debye, P. Light scattering in solutions. J. Appl. Phys. **1944**, *15*, 338–342.
19. Zimm, B.H. Molecular theory of the scattering of light in fluids. J. Chem. Phys. **1945**, *13*, 141–145.
20. Zimm, B.H. The dependence of the scattering of light on angle and concentration in linear polymer solutions. J. Chem. Phys. **1948**, *52*, 260–267.
21. Burchard, W. Statistics of star-shaped molecules. II. Stars with homodisperse side chains. Macromolecules **1974**, *7* (6), 841–846.
22. Casassa, E.F.; Berry, G.C. Angular distribution of intensity of Rayleigh scattering from comblike branched molecules. J. Polym. Sci. A: Polym. Chem. **1966**, *4*, 881–897.
23. Khasat, N.; Pennisi, R.W.; Hadjichristidis, N.; Fetters, L.J. Dilute solution behavior of asymmetric three-arm and regular three- and twelve-arm polystyrene stars. Macromolecules **1988**, *21*, 1100–1106.
24. Douglas, J.F.; Roovers, J.; Freed, K.F. Characterization of branching architecture through ''Universal'' ratios of polymer solution properties. Macromolecules **1990**, *23*, 4168–4180.
25. Fetters, L.J.; Hadjichristidis, N.; Lindner, J.S.; Mays, J.W. Molecular weight dependence of hydrodynamic and thermodynamic properties for well-defined linear polymers in solution. J. Phys. Chem. Ref. Data **1994**, *23* (4), 619–640.
26. Thurmond, C.D.; Zimm, B.H. Size and shape of the molecules in artificially branched polystyrene. J. Polym. Sci. **1952**, *8* (5), 477–494.
27. Zimm, B.H.; Kilb, R.W. Dynamics of branched molecules in dilute solution. J. Polym. Sci. **1959**, *37*, 19–42.
28. Kurata, M.; Abe, M.; Masamichi, I.; Matsushima, M. Randomly branched polymers. I. Hydrodynamic properties. Polym. J. **1972**, *3* (6), 729–738.
29. Ambler, M.R.; McIntyre, D. Randomly branched styrene/divinylbenzene copolymers. II. Solution properties and structure. J. Appl. Polym. Sci. **1977**, *21*, 2268–2282.
30. Berry, G.C.; Orofino, T.A. Branched polymers. III. Dimensions of chains with small excluded volume. J. Chem. Phys. **1964**, *40* (6), 1614–1621.
31. Foster, G.N.; Hamielec, A.E.; MacRury, T.B. The molecular weight and branching distribution method. ACS Symp. Ser. **1980**, *138*, 131–148.
32. Vega, J.R.; Estenoz, D.A.; Oliva, H.M.; Meira, G.R. Analysis of a styrene-butadiene graft copolymer with the help of a polymerization model. Int. J. Polym. Anal. Charact. **2001**, *6*, 81–87.
33. Roovers, J. Dilute solution properties of regular star polymers. Plast. Eng. **1999**, *53*, 285–341.
34. Hert, M.; Strazielle, C. Study of the branching structure of low-density polyethylene by viscosity and light scattering measurements on fractions from exclusion GPC. Makromol. Chem. **1983**, *184* (1), 135–145.
35. Grinshpun, V.; Rudin, A.; Russell, K.E.; Scammell, M.V. Long-chain branching indexes from size-exclusion chromatography of polyethylenes. J. Polym. Sci. B: Polym. Phys. **1986**, *24* (5), 1171–1176.
36. Tackx, P.; Tacx, J.C.J.F. Chain architecture of LDPE as a function of molar mass using size exclusion chromatography and multi-angle laser light scattering (SEC-MALLS). Polymer **1998**, *39* (14), 3109–3113.
37. Wang, W.-J.; Kharchenkob, S.; Miglerb, K.; Zhu, S. Triple-detector GPC characterization and processing behavior of long-chain-branched polyethylene prepared by solution polymerization with constrained geometry catalyst. Polymer **2004**, *45* (19), 6495–6505.
38. Utracki, L.A.; Roovers, J.E.L. Viscosity and normal stresses of linear and star branched polystyrene solution. I. Application of corresponding states principle to zero-shear viscosities. Macromolecules **1973**, *6* (3), 366–372.
39. Roovers, J. Melt rheology of H-shaped polystyrenes. Macromolecules **1984**, *17*, 1196–1200.
40. Ham, J.S. Viscoelastic theory of branched and cross-linked polymers. J. Chem. Phys. **1957**, *26* (3), 625–633.

41. Bueche, F. Viscosity of molten branched polymers and their concentrated solutions. J. Chem. Phys. **1964**, *40* (2), 484–487.
42. Graessley, W.W.; Masuda, T.; Roovers, J.E.L.; Hadjichristidis, N. Rheological properties of linear and branched polyisoprene. Macromolecules **1976**, *9* (1), 127–141.
43. Roovers, J.; Graessley, W.W. Melt rheology of some model comb polystyrenes. Macromolecules **1981**, *14* (3), 766–773.
44. Nagasawa, M.; Fujimoto, T. Preparation, characterization and viscoelastic properties of branched polymers. Prog. Polym. Sci. Jpn. **1972**, *3*, 263–314.
45. Masuda, T.; Nakagawa, Y.; Ohta, Y.; Onogi, S. Viscoelastic properties of concentrated solutions of randomly branched polystyrenes. Polym. J. **1972**, *3* (1), 92–99.
46. Valentine, R.H.; Ferry, J.D.; Homma, T.; Ninomiya, K. Viscoelastic properties of polybutadienes—linear and lightly crosslinked near the gel point. J. Polym. Sci., Part A-2 **1968**, *6*, 479–492.
47. Graessley, W.W.; Shinbach, E.S. Flow properties of branched polydisperse polymers. J. Polym. Sci.: Polym. Phys. Ed. **1974**, *12* (10), 2047–2063.
48. Takahashi, Y.; Suzuki, F.; Miyachi, M.; Noda, I.; Nagasawa, M. Zero-shear viscosity of branched polymer solutions. Polymer J. **1986**, *1968* (18), 1–89.
49. Isono, Y.; Fujimoto, T.; Kajiura, H.; Nagasawa, M. Viscoelastic properties of branched polymers. II. In concentrated solutions. Polym. J. **1980**, *12* (6), 369–378.
50. Raju, V.R.; Rachapudy, H.; Graessley, W.W. Properties of amorphous and crystallizable hydrocarbon polymers. IV. Melt rheology of linear and star-branched hydrogenated polybutadiene. J. Polym. Sci.: Polym. Phys. Ed. **1979**, *17* (7), 1223–1235.
51. Lohse, D.J.; Milner, S.T.; Fetters, L.J.; Xenidou, M.; Hadjichristidis, N.; Mendelson, R.A.; Garcia-Franco, C.A.; Lyon, M.K. Well-defined, model long chain branched polyethylene. 2. Melt rheological behavior. Macromolecules **2002**, *35* (8), 3066–3075.
52. Fetters, L.J.; Kiss, A.D.; Pearson, D.S.; Quack, G.F.; Vitus, F.J. Rheological behavior of star-shaped polymers. Macromolecules **1993**, *26* (4), 647–654.
53. de Gennes, P.G. *Scaling Concepts in Polymer Physics*; Cornell University Press: Ithaca, NY, 1979.
54. Doi, M.; Edwards, S.F. *The Theory of Polymer Dynamics*; Oxford University Press: Oxford, England, 1986.
55. Watanabe, H. Viscoelasticity and dynamics of entangled polymers. Prog. Polym. Sci. **1999**, *24*, 1253–1403.
56. Pearson, D. S.; Raju, V. R. Configurational and viscoelastic properties of branched polymers. Macromolecules **1982**, *15* (2), 294–298.
57. Graessley, W. W.; Roovers, J. Melt rheology of four-arm and six-arm polystyrenes. Macromolecules **1979**, *12* (5), 959–965.
58. Randall, J. C. Characterization of long-chain branching in polyethylenes using high-field. ^{13}C-NMR. ACS Symp. Ser. **1980**, *142*, 93–118.
59. Shroff, R.N.; Mavridis, H. Assessment of NMR and rheology for the characterization of LCB in essentially linear polyethylenes. Macromolecules **2001**, *34* (21), 7362–7367.
60. Jordan, E.A.; Donald, A.M.; Fetters, L.J.; Klein, J. Transition from linear to star-B branched diffusion in entangled polymer melts. Polym. Prep. **1989**, *30* (1), 63–64.
61. De Pooter, M.; Smith, P.B.; Dohrer, K.K.; Bennett, K.F.; Meadows, M.D.; Smith, C.G.; Schouwenaars, H.P.; Geerards, R.A. Determination of the composition of common linear low density polyethylene copolymers by ^{13}C-NMR spectroscopy. J. Appl. Polym. Sci. **1991**, *42* (2), 399–408.
62. Wang, W.-J.; Yan, D.; Zhu, S.; Hamielec, A.E. Kinetics of long chain branching in continuous solution polymerization of ethylene using constrained geometry metallocene. Macromolecules **1998**, *31* (25), 8677–8683.
63. Grubisic, Z.; Rempp, P.; Benoit, H. Universal calibration for gel permeation chromatography. J. Polym. Sci. **1967**, *5* (9), 753–759.
64. Stogiou, M.; Kapetanaki, C.; Iatrou, H. Examination of the universality of the calibration curve of size exclusion chromatography by using polymers having complex macromolecular architectures. Int. J. Polym. Anal. Charact. **2002**, *7*, 273–283.
65. Scholte, Th.G.; Meijerink, N.L.J. Gel permeation chromatography on branched polymers. Polym. J. **1977**, *9* (2), 133–139.
66. Drott, E.E.; Mendelson, R.A. Determination of polymer branching with gel-permeation chromatography. I. Theory. J. Polym. Sci.: Polym. Phys. Ed. **1970**, *8*, 1361–1371.
67. Ram, A.; Miltz, J. New method for molecular weight distribution (MWD) determination in branched polymers. J. Appl. Polym. Sci. **1971**, *15* (11), 2639–2644.
68. Kurata, M.; Okamoto, H.; Iwama, M.; Abe, M.; Homma, T. Randomly branched polymers. II. Computer analysis of the gel-permeation chromatogram. Polym. J. **1972**, *3* (6), 739–748.

69. Hamielec, A.E.; Ouano, A.C. Generalized universal molecular weight calibration parameter in GPC. J. Liq. Chromatogr. **1978**, *1* (1), 111–120.
70. Mourey, T.H.; Vu, K.A.; Balke, S.T. Use of multidetector SEC for determining local polydispersity. ACS Symp.Ser. **1999**, *731*, 20–34.
71. Thitiratsakul, R.; Balke, S.T.; Mourey, T.H.; Schunk, T.C. Detecting ''perfect resolution'' local polydispersity in size exclusion chromatography. Int. J. Polym. Anal. Charact. **1998**, *4* (4), 357–377.
72. Balke, S.T.; Mourey, T.H.; Karami, A. Evaluating size exclusion chromatography fractionation. Int. J. Polym. Anal. Charact. **2000**, *6* (1), 13–33.
73. Balke, S.T.; Mourey, T.H. Local polydispersity detection in size exclusion chromatography: method assessment. J. Appl. Polym. Sci. **2001**, *81* (2), 370–383.
74. Scorah, M.J. Experimental and Modelling Investigation of a Novel Tetrafunctional Initiator in Free Radical Polymerization. Ph.D. Thesis, Department of Chemical Engineering, University of Waterloo, Waterloo, Canada, 2005.
75. Dealy, J.; Wissbrun, K.F. *Melt Rheology and Its Role in Plastics Processing*; Van Nostrand Reinhold: New York, 1989.
76. Gabriel, C.; Münstedt, H. Influence of long-chain branches in polyethylenes on linear viscoelastic flow properties in shear. Rheol. Acta **2002**, *41* (3), 232–244.
77. Janzen, J.; Colby, R.H. Diagnosing long-chain branching in polyethylenes. J. Molecul. Struct. **1999**, *485-486*, 569–584.
78. Shroff, R.N.; Mavridis, H. Long-chain branching index for essentially linear polyethylenes. Macromolecules **1999**, *32*, 8454–8464.
79. Shroff, R.; Mavridis, H. New measures of polydispersity from rheological data on polymer melts. J. Appl. Polym. Sci. **1995**, *57* (13), 1605–1626.
80. Lai, S.; Plumley, T.A.; Butler, T.I.; Knight, G.W.; Kao, C.I. Dow rheology index (DRI) for insite technology polyolefins (ITP): unique structure-processing relationships. ANTEC Soc. Plast. Eng. **1994**, *52* (2), 1814–1815.
81. Wood-Adams, P.M.; Dealy, J.M. Using rheological data to determine the branching level in metallocene polyethylenes. Macromolecules **2000**, *33* (20), 7481–7488.
82. Wood-Adams, P.; Dealy, J.M.; deGroot, A.W.; Redwine, O.D. Effect of molecular structure on the linear viscoelastic behavior of polyethylene. Macromolecules **2000**, *33* (20), 7489–7499.
83. Trinkle, S.; Friedrich, C. Van Gurp-Palmen plot: a way to characterize polydispersity of linear polymers. Rheol. Acta **2001**, *40* (4), 322–328.
84. Trinkle, S.; Walter, P.; Friedrich, C. Van Gurp-Palmen plot. II. Classification of long chain branched polymers by their topology. Rheol. Acta **2002**, *41* (1–2), 103–113.

Bubble Cap Tray

Stanley Marple
Chemical Engineering Department, University of Houston, Houston, Texas, U.S.A.

INTRODUCTION

Bubble cap trays are commonly used to provide mass or heat transfer between liquid and vapor streams. Its advantages include minimum liquid leakage, wide range of operating rates, reasonable cost, and usefulness at very low liquid rates.

Sieve and valve trays have replaced bubble cap trays in many new construction projects because of lower cost. Yet bubble cap trays are still often used for the superior characteristics mentioned earlier.

This entry summarizes the technology of bubble cap trays. Industrial-scale data contributed by Fractionation Research, Inc. (FRI) are included. More detailed technical material is available in Ludwig[1] and in Refs.[2–5]

GENERAL DESCRIPTION

Bubble cap trays are commonly used to provide mass or heat transfer between liquid and vapor streams. The most common installation is the multiplate column containing two to several hundred trays at vertical spacing of 6–36 in. (Fig. 1).

Fig. 2 shows an assembled tray with 3 in. caps. A closer view of the caps is seen in Fig. 3. Fig. 4 displays the dimensions of 3 in. caps used by FRI in their experimental studies (Fractionation Research, Tulsa, Oklahoma, U.S.A., private communication).

Caps may be 2–8 in. in diameter or larger, but there is a performance advantage for the smaller 3 or 4 in. caps. The vapor slots are typically 0.25–0.50 in. wide, with total slot area perhaps 10% of the tower cross-sectional area. The vapor risers (Fig. 4) prevent liquid from flowing down through the openings in the tray.

Caps are sometimes tack-welded into place for long-term durability, or they may be fastened with double lock-nuts or hold-down bars installed parallel to the liquid flow.

Liquid flow may be directed onto the tray across a low entrance weir to even out the flow, or the downcomer skirt clearance alone may be relied upon for this function. The liquid is agitated into a frothy mix as it flows across to the outlet weir. The outlet weir is normally 1–4 in. high to provide a liquid seal and leads to a downcomer that occupies about 8–20% of the tower cross-sectional area.

The overflow weir should be at least 0.50 in. higher than the downcomer skirt clearance to prevent vapor from passing up the downcomer, thus escaping vapor–liquid contact on the tray above.

Sometimes for large liquid rates, the downcomer skirt clearance is increased by recessing the bottom end of the downcomer into a sump built into the tray floor. Such a sump increases tray cost and may trap debris.

Weep Holes and Tray Leakage

During shutdown, it is necessary to drain the trays and seal pans of unevaporated liquid and steam condensate. At least two drain holes are needed for each tray or seal pan, at 0.50 in. minimum diameter and preferably 0.75 in. to avoid plugging. As a rule of thumb, a 0.75 in. hole will deliver about 2 gpm of liquid during tower drainage.

If tray leakage during operation is found to be harmful, gasketing or seal welding may be employed.

Sequential Start-up

It may be necessary to establish liquid flow before introducing vapor flow through the tray, especially with small downcomers. However, large ones usually do not require this measure.

Direction of Liquid Flow

The most common flow arrangement (Fig. 1) has liquid alternating direction as it passes down the tower. Lewis studied three different flow arrangements[6] and found that the most efficient method is to have the liquid flowing in the same direction in plug flow on each tray. The reversing arrangement is, however, mechanically expedient.

Hanging Downcomers

It is a common practice, as shown in Fig. 1, to slope the downcomer wall inward so that the bottom of the

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008098

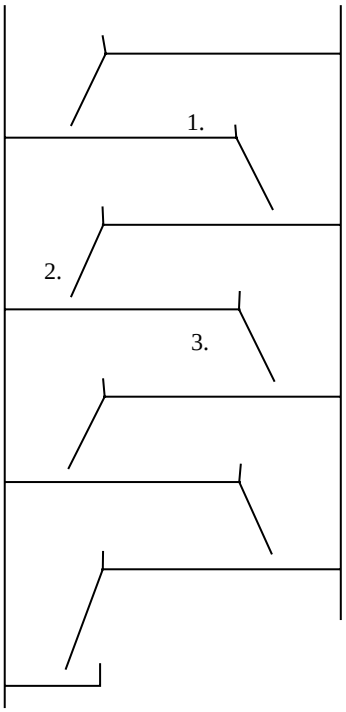


Fig. 1 Schematic arrangement of bubble cap trays: 1) capped trays; 2) liquid downcomers; and 3) disengagement zones.

downcomer occupies less tray area than the top. Active area is thus increased, and the downcomer is equally effective because froth settling gives lower downward velocity in the lower part of the downcomer.

In recent years, downcomer area encroachment on the active area has been further reduced by a concept called "hanging downcomer." The resulting lower slot velocities allow more vapor-handling capacity. One such arrangement is the UOP (originally the Union Carbide) MD Tray (Fig. 5).[7] A restricting slot in the bottom of the downcomer box causes liquid to back up, providing a seal to prevent vapor passage up the downcomer.

The capacity advantage provided by the hanging downcomer depends on how much additional area is provided. For example, if the inflow downcomer for a conventional tray covers 15% of the tray area, the capacity improvement for a hanging downcomer might be 10% or 12%.

Trays with Centrifugal Action

At least two manufacturers offer trays that depend on centrifugal action to separate vapor and liquid streams. Each tray is an array of centrifugal contacting and separating devices, which brings the streams together

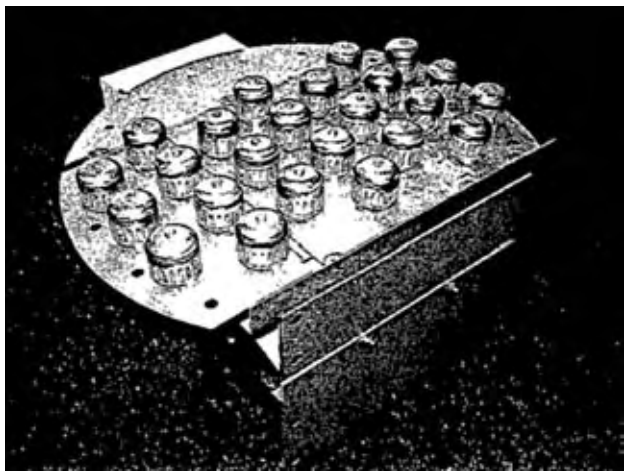


Fig. 2 Bubble cap tray assembly. (Photograph courtesy of ACS Industries, Inc.)

and then separates them in a swirl within each device. Because the separating force can be much greater than gravity, the towers can operate with very high vapor velocities. The capital cost of these centrifugal trays is high at present.

Manning at Shell Research was one of the first to experiment with centrifugal separation on an industrial scale.[8]

PREDICTING SYSTEM PERFORMANCE WITH BUBBLE CAP TRAYS

Correlations of data for either capacity or separating power cover a wide range of hardware and distilling

Fig. 3 Side view of 3 in. bubble caps as installed. (Photograph courtesy of ACS Industries, Inc.)

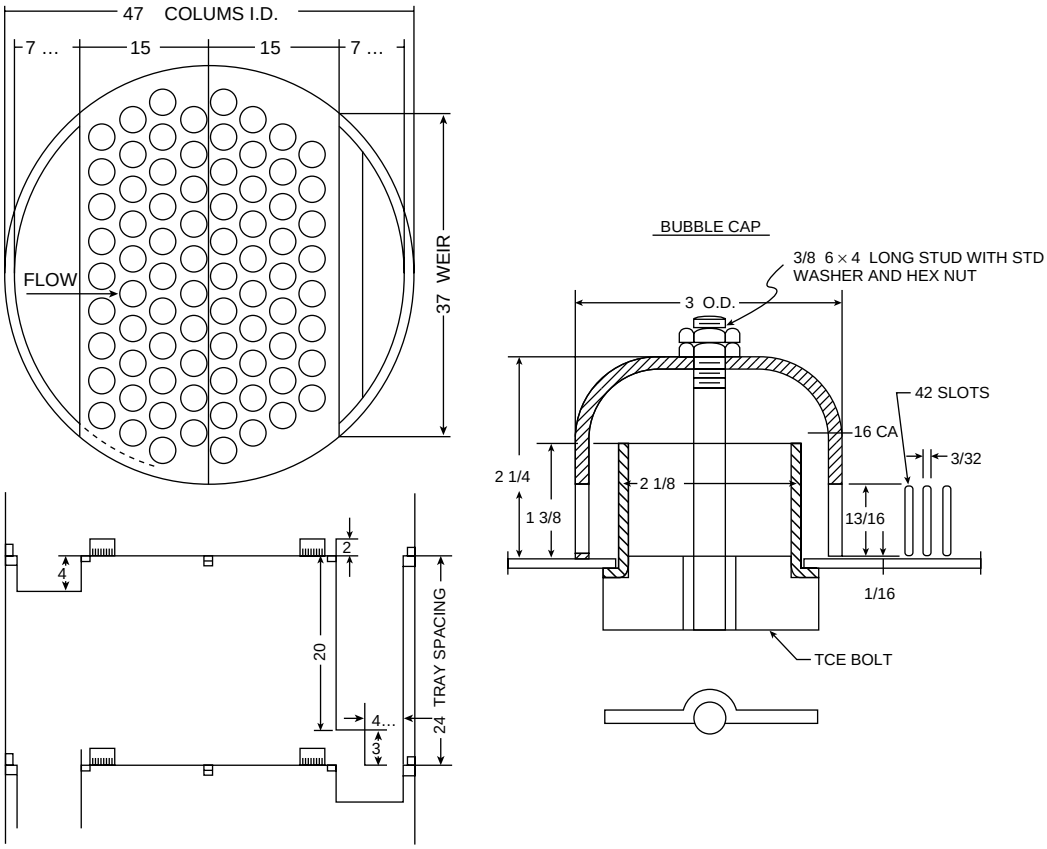


Fig. 4 Mechanical details of 3 in. cap installation at FRI. (Drawing courtesy of Fractionation Research, Inc.)

systems. The most dependable predictions for a new system are therefore based on operating data for similar equipment and distilling systems. To this end, pertinent data obtained by FRI are given in the following material (Fractionation Research, Tulsa, Oklahoma, U.S.A., private communication, 2003).

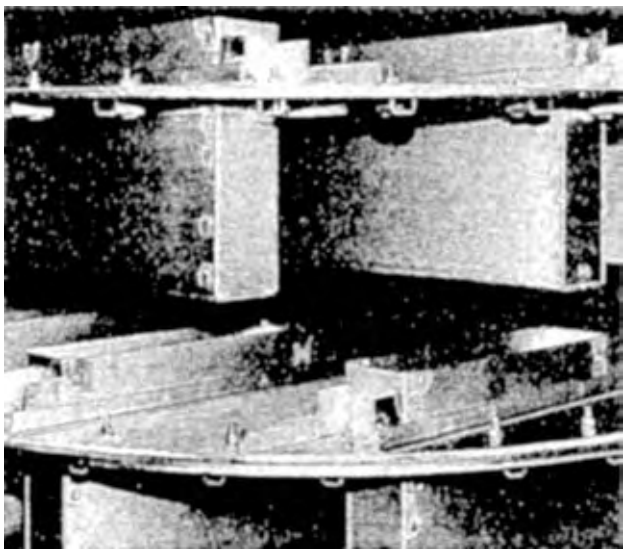


Fig. 5 Side view of MD trays. (Courtesy of UOP.)

Vapor-Limited Capacity

Operation of bubble cap trays at high vapor rate in effect may be limited by massive entrainment of a liquid into the vapor stream. This causes high-pressure drop for vapor flow, which in turn prevents liquid from flowing down the tower. Thus, while the term ''vapor flood'' is often applied to conditions of high entrainment, it is also true that flooding refers to a failure to handle liquid flow.

Fig. 6 includes full-scale capacity data for 3, 4, and 6 in. caps at low pressure obtained by FRI and by Shell for 6 in. caps only (Fractionation Research, Tulsa, Oklahoma, U.S.A., private communication).[9] Both sets of data for 6 in. caps show a capacity deficit compared to the smaller caps; however, the Shell data show approximately 80% of the capacity found by FRI. The difference may be caused by system properties or by the more conservative definition of flood point used by Shell. Data at higher pressure with the butane–isobutane system at the 4 ft diameter were also obtained by FRI (Fig. 8).

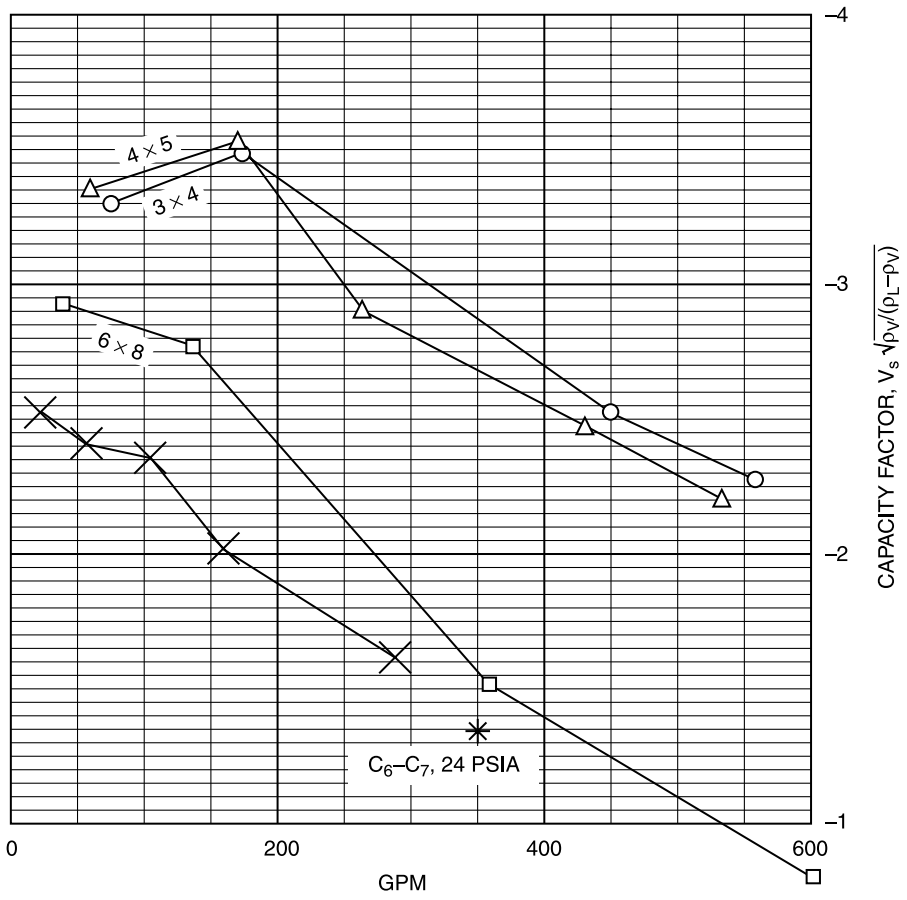


Fig. 6 Capacity of bubble cap trays, low-pressure systems. V_s, superficial vapor velocity (ft/sec); ρ_v, vapor density (any units); ρ_l, liquid density (same units as vapor); ○, data supplied by FRI for 3 in. caps and on 4 in. triangular pitch; Δ, FRI data, 4 in. caps on 5 in. triangular pitch; □, FRI data, 6 in. caps on 8 in. triangular pitch; ×, Shell data, 6 in. caps on 8.5 in. triangular pitch, 2,2,4-trimethyl pentane/toluene system, 19.7 psia; ∗, except Shell data. (From Ref.[9].)

Effect of Foaming Tendency

Systems of low-foaming tendency tend to conform to a standard capacity prediction as defined by liquid and vapor densities. Systems of higher molecular weight also have higher boiling points. Thus, at a given operating pressure, higher distilling temperatures for heavier materials cause foaming tendencies to be similar to those for distilling lighter materials. There is, however, no guarantee that a new system may not have unusual foaming tendency.

The degree of foaming tendency may be checked by laboratory foaming tests, but it is difficult to predict what effect a given degree of foaming may have on tray capacity.

At least four types of foaming systems have been recognized in the distilling industry. They are:

1. The ''soapy'' foam in which bubbles are stabilized by surface-adsorbed components. These foams cover a wide range of system factors. Sometimes chemical foam-breaking additives are used.
2. A second type of foam is found in which the liquids have high viscosity so that the liquid film forming the bubbles drains slowly. An example of this type is encountered when stripping small amounts of volatile solvents from heavy hydrocarbon lubricating oils.
3. Another type of foam has been observed when a multicomponent liquid has a composition close to a separation into two liquid phases. Surface tension drops to a very low value, allowing easy formation of bubbles. An example is found in solvent recovery from lubricating-oil extraction processes. This type of foam has been referred to as ''Ross'' foam,[10] named after an investigator of these foams.
4. The fourth type, Maragoni foam,[11] so called after Maragoni's observation of liquid film stabilization on evaporation, has been described by Zuiderweg and Harmens.[12] This foam occurs in a complex system when surface tension increases as evaporation proceeds, thus strengthening the bubbles. It is usually weak, thus often felt to have little industrial significance. On the other hand, there appears to be no other way to

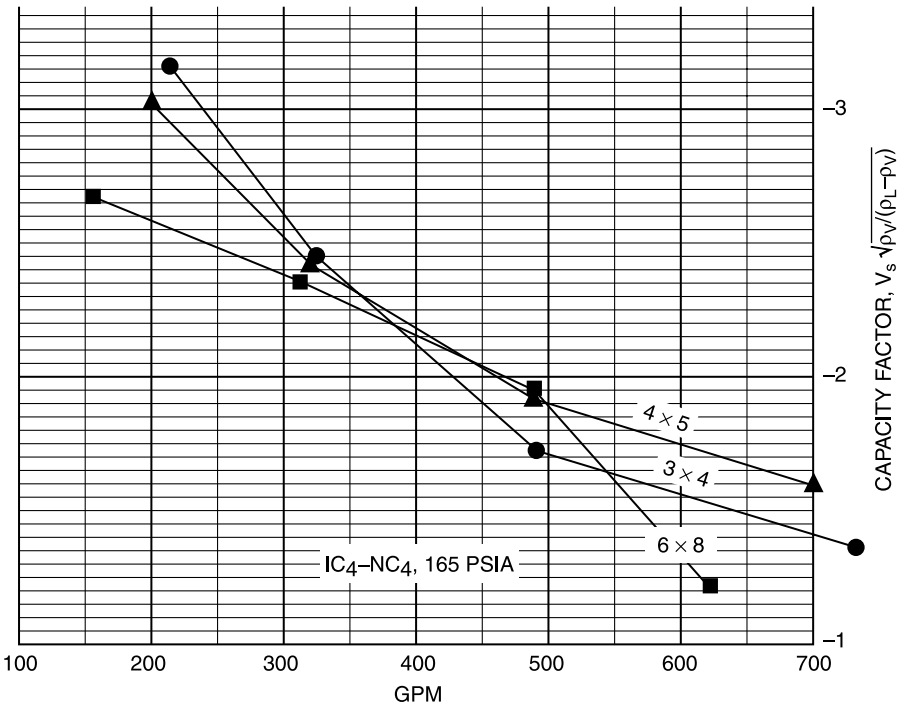


Fig. 7 Capacity of bubble cap trays, butane–isobutane system (FRI 4 ft. I.D. Tower). •, 3 in. caps on 4 in. triangular pitch; ▲, 4 in. caps on 5 in. centers, triangular pitch; and ■, 6 in. caps on 8 in. centers, triangular pitch. (Data courtesy of Fractionation Research, Inc.)

explain the strong capacity decrease observed in the upper trays of absorption plant stripping towers in the petroleum refining industry. Of course, slurries of fine particles are a special case of foaming.

The observed system capacity factors have been reported by Glitsch in Table 1.

Predicting Vapor-Limited Capacity

The most conservative definition of vapor flooding capacity is the load for which tower pressure drop exhibits a sharp increase, signifying liquid buildup at some tray. It is, however, possible to operate the tower at somewhat higher load, perhaps 10% more. As the load is increased, reboiler pressure rises, allowing a semistable operation, albeit with reduced separating efficiency. Fractionation research with others is understood to use a higher limiting load, i.e., the load at which the tower becomes inoperable.

When flooded zones expand, they move upward in the tower. The limiting tray can often be identified from pressure measurements.

Souders and Brown introduced the relationship $U_{\max} = KV\sqrt{D_v/(D_l - D_v)}$, a modified kinetic energy ratio to correlate limiting entrainment rates for various systems.[13] In this relationship V is the vapor velocity; D_v is the vapor density; and D_l is the liquid density.

Fair has published a refinement of the Souders–Brown correlation for bubble cap tray capacity,[3] while Glitsch has provided a vapor flooding relationship for valve trays, which also seems to represent bubble cap tray capacity for 3 or 4 in. bubble caps at low pressure.[14] In the latter, originally developed for valve trays and called ''equation (13)'' by Glitsch, the effect of hydraulic liquid head gradient is expressed as an equivalent added vapor load:

$$\frac{\%\ \text{Flood}}{100} = \frac{V_{\text{load}} + \text{gpm} \times \text{FPL} \div 13{,}000}{\text{AA} \times \text{CAF}}$$

In this relationship, V_{load} is the vapor rate, cfs $\times \sqrt{D_v/(D_l - D_v)}$; FPL is the flow path length across the tray (in.); gpm is the liquid flow rate (gal/min); AA is the active tray area (ft^2); and CAF is the Glitsch capacity factor (Fig. 9).

Table 1 System factors for capacity estimation

Nonfoaming regular systems	1.0
Fluorine systems, e.g., BF_3, Freon	0.9
Moderate foaming, e.g., oil absorbers, amine, and glycol regenerators	0.85
Heavy foaming, e.g., amine and glycol absorbers	0.73
Severe foaming, e.g., MEK units	0.60
Foam-stable systems, e.g., caustic regenerators	0.30

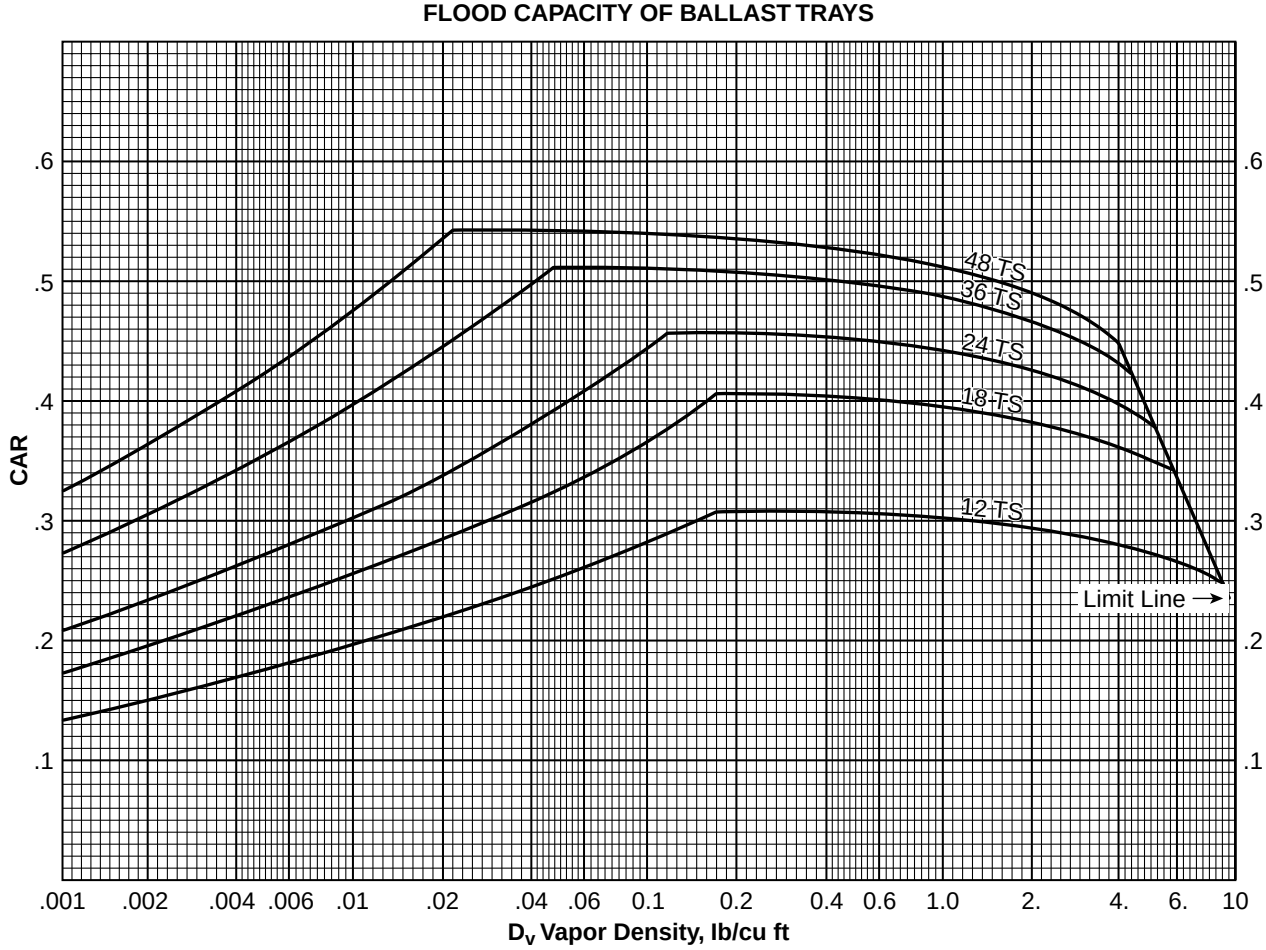


Fig. 8 Glitsch capacity factors for valve trays or bubble cap trays with small caps. TS, tray spacing. (Courtesy of Koch–Glitsch, Inc.)

A conservative design would be to take 90% of the Glitsch value as the flooding load with 3 in. bubble caps, and 70% of the valve-tray flooding load for a tray with 6 in. caps (Fractionation Research, Tulsa, Oklahoma, U.S.A., private communication).

Liquid Rate Flooding

Flooding is much more sensitive to vapor rate than to liquid rate. Trays can, however, definitely be flooded by high liquid rates. Prediction of maximum liquid capacity is based on hydraulic calculation of liquid backup in the downcomers. For low-foaming systems, it is often assumed that liquid backup cannot be more than half of the downcomer height, calculated as clear liquid. For high vapor density, e.g., 3.0 lb/ft^3, the limiting backup is even less.

Above the limiting liquid load, the vapor is entrained downward in the downcomer, the tray liquid level builds up, and the liquid becomes entrained into the vapor stream. The tower floods with the liquid, separation degrades, and eventually the liquid carries overhead just as in vapor-initiated flooding.

Predicting Liquid-Limited Capacity

The backup of the liquid in the downcomer is the liquid depth on the tray plus the equivalent tray pressure drop and the head loss under the downcomer skirt. The latter can be somewhat decreased by using a shaped downcomer skirt edge, but this measure is often found to be inconvenient.

The dry tray head loss for vapor passing through the caps has been measured by a number of investigators.[2,15] In general, we add the head loss through the riser, the slots, and the reversal area to give very roughly three times the velocity head ($U_{\text{riser}} \times D_{\text{vapor}}$) (in ft) of the fluid handled. To this dry tray drop, we add the submergence at the bottom of the slot opening, calculated from the weir height and the weir crest, plus half of the liquid-head gradient on the tray, to get the wet tray pressure drop. Bolles and Dauphine, as

summarized by Ludwig, give a detailed exposition that is too lengthy to include here.[1,2,15]

The downcomer liquid backup, then, is:

$$H_d = h_w + h_{ow} + \Delta + h_d + h_t$$

where H_d is the downcomer backup (in. of liquid); h_w is the weir height (in.); h_{ow} is the weir crest (in. of liquid); Δ is the hydraulic gradient across the tray (in.); h_d is head loss under the downcomer skirt (in. of liquid), and h_t is the total wet tray pressure drop (in. of liquid).

The wet tray pressure drop is obtained by calculating the dry tray drop, then adding the slot submergence, including the weir crest and half of the average hydraulic gradient.

The weir crest is obtained from the Francis weir formula as $h_{ow} = 0.4(\text{gpm}/L_{wi})^{2/3}$, in which L_{wi} is the effective weir length in inches.

The slot opening is easily obtained from the volumetric vapor flow at the tray pressure and temperature, and the slot geometry.

It is also required to size the downcomer for adequate phase separation. One method often employed for setting downcomer design velocity is laid out in the Glitsch design manual for valve trays.[13] Here, the downcomer design velocity is correlated from the difference between liquid and vapor densities and the system factor as described above. For many low-pressure liquids, downcomer area at the top might allow 200–250 gpm/ft^2 at 24 in. tray spacing.

Hydraulic Gradient

Large liquid head gradients reduce tray capacity and may damage separation by causing maldistribution of vapor flow across the tray. In fact, an extremely large head gradient can cause some caps to dump the liquid to the tray below.

Bolles gives helpful guides to predict the hydraulic gradient and are also reproduced in Ludwig.[1,2]

On the other hand, Shell research staff found only a small measured hydraulic gradient on a well-designed low-pressure tower at 5 ft diameter with a liquid load of 121 gpm/ft of weir length.[9]

Trays for High Liquid Loads

Liquid-handling capacity can be improved by splitting the flow between two paths, usually called ''passes'' (Fig. 9). It is frequently possible to use four or more passes per tray, but it is difficult to distribute liquid evenly in a cylindrical tower if more than two passes are used. Another method of handling large liquid flow is to split each flow path into two or more levels with the liquid flowing over a weir at each level.

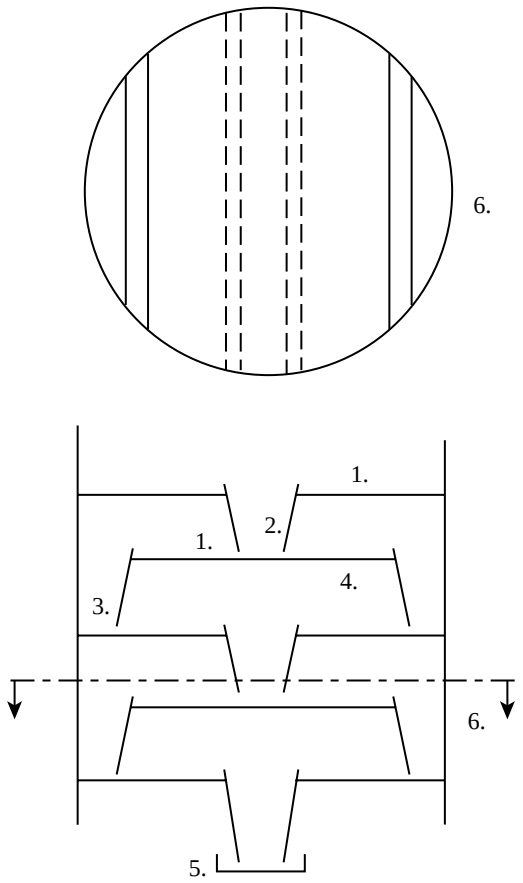


Fig. 9 Two-pass trays: 1) capped areas; 2) center liquid downcomers; 3) side liquid downcomers; 4) disengagement zones; 5) seal pan; and 6) cross-sectional view.

A swept-back weir is often used to increase weir length for high liquid rates. Caps with directional vapor flow may also be used to push liquid across the tray.

''Hanging downcomers'' as in the UOP MD tray (Fig. 5) are useful for high liquid loads, but there is efficiency loss because of the short flow path lengths. This problem may be countered by using more trays at reduced spacing.

Tray Separating Efficiency

Quantification of separating efficiency is useful in: 1) design and 2) rating performance of operating or experimental equipment.

The most commonly used definition of efficiency is the overall efficiency, defined as the number of theoretical trays required for a separation divided by the number of actual trays. A theoretical tray is defined as a contacting stage for which the vapor and the liquid streams leaving would be in thermodynamic equilibrium. The overall efficiency is inexact because separation

within one tower will vary with species properties and tray hydraulics. Practically speaking, it is usually constant in one section of the tower within 10% and is often a basis for design.

For an individual tray, the concept of efficiency often used is the Murphree vapor efficiency, defined as the ratio of the composition change in the vapor stream to the composition change if the vapor leaving the tray were in thermodynamic equilibrium with the liquid leaving. Thus,

$$e_{\mathrm{mv}} = \frac{y_n - y_{n-1}}{y_n^* - y_{n-1}}$$

where y_n is the vapor mole fraction of the light key. The subscript identifies the tray number. y_n^* is the vapor composition in equilibrium with the liquid leaving the tray.

Tray efficiencies are usually assumed to have the same value for each component in the separation. However, actual separating efficiency for several components may vary because of differing molecular properties, such as diffusivity.

In research and development work, a concept called point efficiency is useful. Considering a local element of tray contacting area, the point vapor efficiency is defined as the composition change in the vapor stream passing through the element, divided by the composition change that would be obtained if the vapor leaving the element were in equilibrium with the liquid leaving it.

$$e_{\mathrm{l}} = \frac{y_{\mathrm{l},n} - y_{\mathrm{l},n-1}}{y_{\mathrm{l},n}^* - y_{\mathrm{l},n-1}}$$

where e_{l} is the point vapor efficiency; $y_{\mathrm{l},n}$ is the composition of the vapor leaving the tray element; $y_{\mathrm{l},n-1}$ is the composition of the vapor entering the tray element from the tray below; and $y_{\mathrm{l},n}^*$ is the vapor composition that would be in equilibrium with the liquid leaving the tray element.

Point efficiency is difficult to measure directly because it is hard to get a vapor sample experimentally that does not contain entrained liquid droplets.

It would be possible to do design work integrating local efficiencies for each tray through the entire tower, though calculating the detailed work required would be a lengthy procedure.

A related design concept would be to use a film diffusion model to predict composition changes for each element of tray area, integrating over the entire tower. This ''rate based'' concept has attraction in special cases and is of course the normal method of design for packed distilling towers, as contrasted to plate towers.

For design purposes, it would be accurate to use the Murphree efficiency that is applied to each tray and integrated for the entire tower. Some day it may be feasible to do that. The amount of calculation required for a complete Murphree efficiency approach would be large, and often the overall efficiency is close to the Murphree tray efficiency.

Prediction of Separating Efficiency

An important reference often used for design is O'Connell's correlation of tray efficiencies.[16] This work is an improvement of an earlier publication by Drickamer and Bradford.[17] In O'Connell's correlation (Fig. 10), it is assumed that mass and heat transfer depend on diffusivities and liquid viscosity, and that both effects can be correlated by using the liquid viscosity and the component relative volatility for the separation.

The next step in predicting tray efficiency for design would be to take into account the effects of 1) tray and tower configuration and 2) hydraulic conditions on the tray. For the effects of the former, one can refer to a comprehensive treatise by W. K. Lewis, Jr.[6] In this somewhat idealized study, the advantage of plug flow of liquid across the tray and its flow direction for the liquid on each tray is shown. Such an arrangement can produce Murphree tray efficiencies as high as 150% from a point efficiency of 80%.

A few towers with very low liquid flow rates have been set up with pipe downcomers arranged for liquid flow in the same direction on successive trays. On the other hand, it is much simpler mechanically to provide reversed flow of liquid on adjacent trays.

If the liquid flow is reversed on successive trays, the cross-flow enhancement of efficiency is greatly reduced, but still present. Thus, tray efficiencies of 110% or more are sometimes observed for low-viscosity systems, such as depropanizers or butane–isobutane splitters.

The effects of tray leakage and liquid entrainment in the vapor stream may need to be taken into account. For bubble cap trays, tray leakage is normally of no significance. Entrainment, on the other hand, can decrease separation markedly. Quantitative prediction of entrainment is sometimes possible; however, the effect of entrainment is usually combined with the effect of vapor velocity on local efficiency, using the results of experimental full-scale studies, such as those performed by FRI.

Optimization of Tray Efficiency

It is tempting to reduce capital cost by designing distilling towers that operate at a high percentage of maximum load, say 85% or more, though the optimum design may be at lower loads; for example, see Fig. 11.

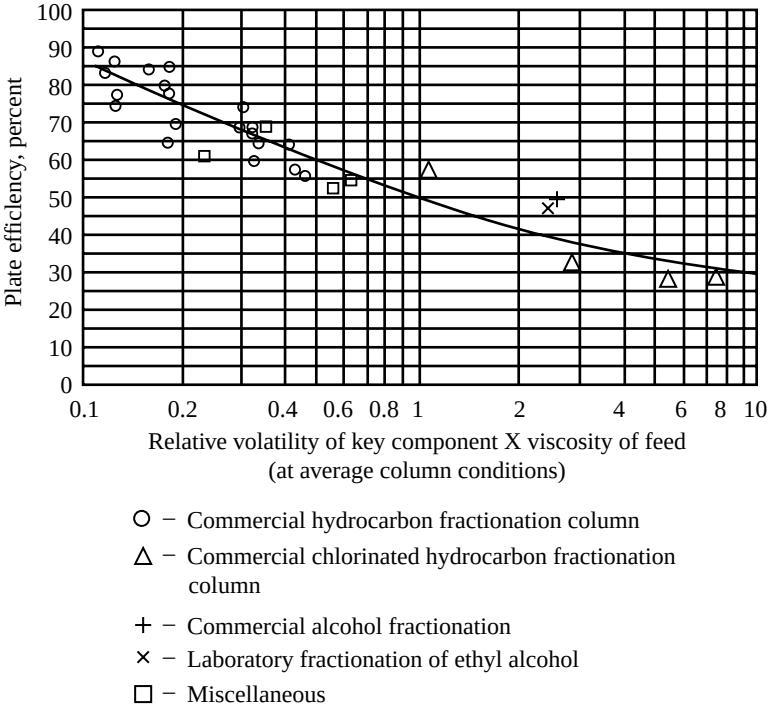


Fig. 10 O'Connell correlation of tray efficiencies. (From Ref.[16].)

The maximum efficiency is frequently found at approximately 60% of the maximum vapor load. At higher loads, vapor contact time is decreased and entrainment may significantly reduce tray efficiency.

It has been confirmed that the effect of vapor rate on entrainment has a higher exponent for dense vapors as in high-pressure distillation.[18] Thus, the effect of entrainment on efficiency is less marked for high-pressure distillation, because as soon as entrainment sets in with increasing vapor rate, the column is overcome by the massive entrainment rate. This effect may mean that it is reasonable to design for a somewhat higher fraction of flooding vapor load in a high-pressure system than in a low-pressure system. See Fig. 12 for FRI data on the effect of vapor rate on efficiency at high pressures.

It is rare to encounter a satisfactory long-term operation at more than about 92% of vapor capacity. Above about 95% of maximum load, small surges in feed rate or reboil rate may cause temporary local flooding, which interferes with good separation.

The effect of entrainment on separating efficiency depends on each case. In general, entrainment is more harmful in rectifying sections than in stripping sections, because it carries less volatile components toward the draw point for the volatile product. It is, in fact, possible to estimate the reduction in tray efficiency caused by a given amount of entrainment. The amount can be estimated roughly by the method of Simkin, Strand, and Olney.[18]

Tray Spacing, Hydraulic Regime, and Entrainment

Direct observations of operating distilling towers have established that the liquid/vapor mixed phase on a bubble cap tray usually exists in three zones. On the bottom there is a thin layer, a few millimeters thick, of dense liquid. Above this is a somewhat foamy layer containing a lot of vapor, but essentially liquid-continuous. This layer extends roughly to the top of the caps or, perhaps, a few centimeters higher. Then the upper part of the tray space contains a spray zone, varying in density so that relatively little liquid is carried to the tray above.

As pointed out by Bennett and Novak,[19] at low liquid rates the liquid-continuous foamy layer may practically disappear and mass transfer takes place in a vapor-continuous spray regime.[19] Operation in the spray regime is less effective for mass transfer and should be avoided when possible.

When the vapor rate is increased as much as possible, the region of higher spray density expands upward and an exponential increase in liquid tray-to-tray entrainment takes place.

Larger tray spacings reduce entrainment markedly. The optimum spacing will vary with the system. For example, systems that leave corrosion products behind must be maintained more frequently. It may then be necessary to have 2 ft. tray spacings to allow convenient access. The optimum spacing for clean systems will be less, perhaps 15 or 18 in.

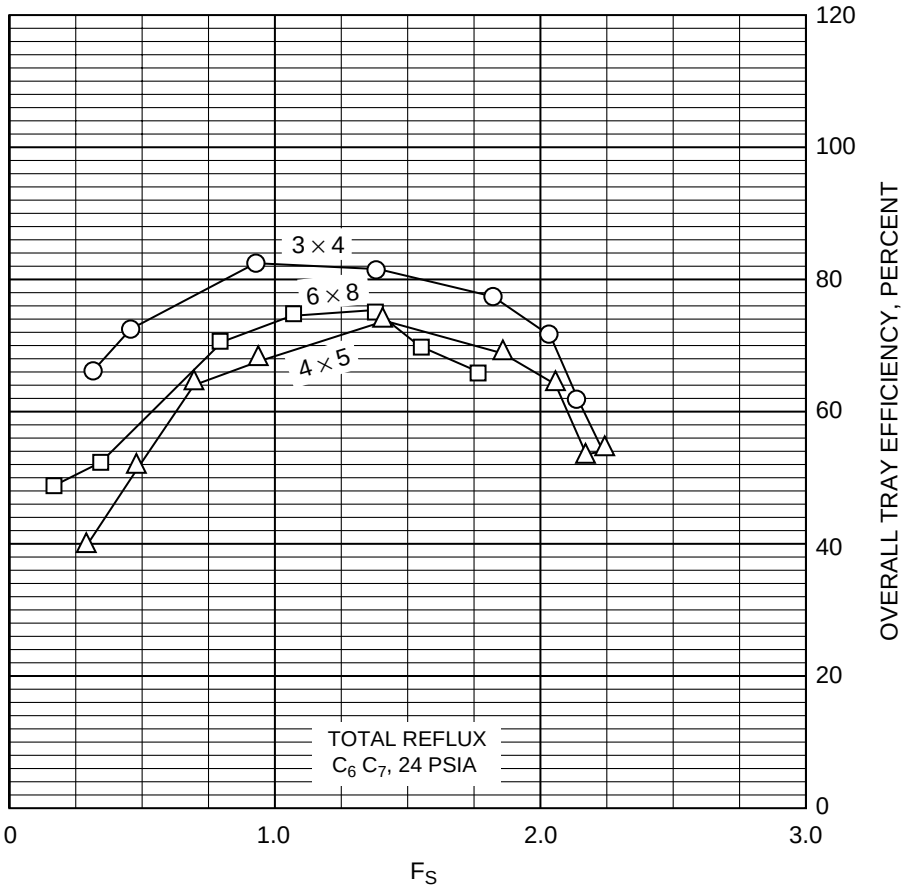


Fig. 11 Data on efficiency of bubble cap trays at low pressure. $F_s = V_s P_v^{1/2}$ (for legend see Fig. 6). (Data courtesy of Fractionation Research, Inc.)

Large-diameter towers require more spacing to provide heavy tray supports.

Effect of Liquid Leakage

The effect of liquid leakage from the tray is more important for stripping trays than for rectifying trays. For stripping trays, liquid leakage carries more volatile components toward the draw point for the less volatile product, thus reducing the effectiveness of the stripper. Fortunately, for most bubble cap trays, there is no tray leakage of consequence.

VACUUM TOWERS

Bubble cap trays are often used in vacuum towers with absolute top pressures of 50–120 mm mercury (Torr.). For lower top pressures or large numbers of theoretical trays, structured packing may be preferred because of its lower pressure drop per theoretical stage. If practicable, the bubble cap trays have the advantage of lower installed cost and easier cleanup in fouling service.

The importance of tray pressure drop in vacuum columns might be illustrated by the following example: Take a limiting case with 50 mm mercury (Torr) top tower pressure. If tray pressure drop is at best 2 mm mercury/tray and 20 actual trays are needed (assume 10 theoretical stages), then the bottom pressure would be at least 90 mm mercury. The reboil pressure would need to be higher, perhaps 120 mm mercury absolute. In a typical situation, the bottom oil product might tolerate 800°F for a short time. The maximum equivalent atmospheric initial boiling point for the bottom product oil would be about 750°F. This boiling point would be too low for most lubricating oils; but if a large flow of steam is introduced into the tower, the reduction in oil partial pressure might allow an equivalent atmospheric boiling point of the bottom product as high as 950°F, or 510°C. This initial boiling point is in the range of industrial and automotive lubricating oils.

There are many cases in which deeper distillation cuts are needed. For such cases, the packed towers can go somewhat deeper. Furthermore, heat-sensitive materials may require lower bottom tower temperatures. A more expensive device, such as the spinning band column, can go to very low pressures because

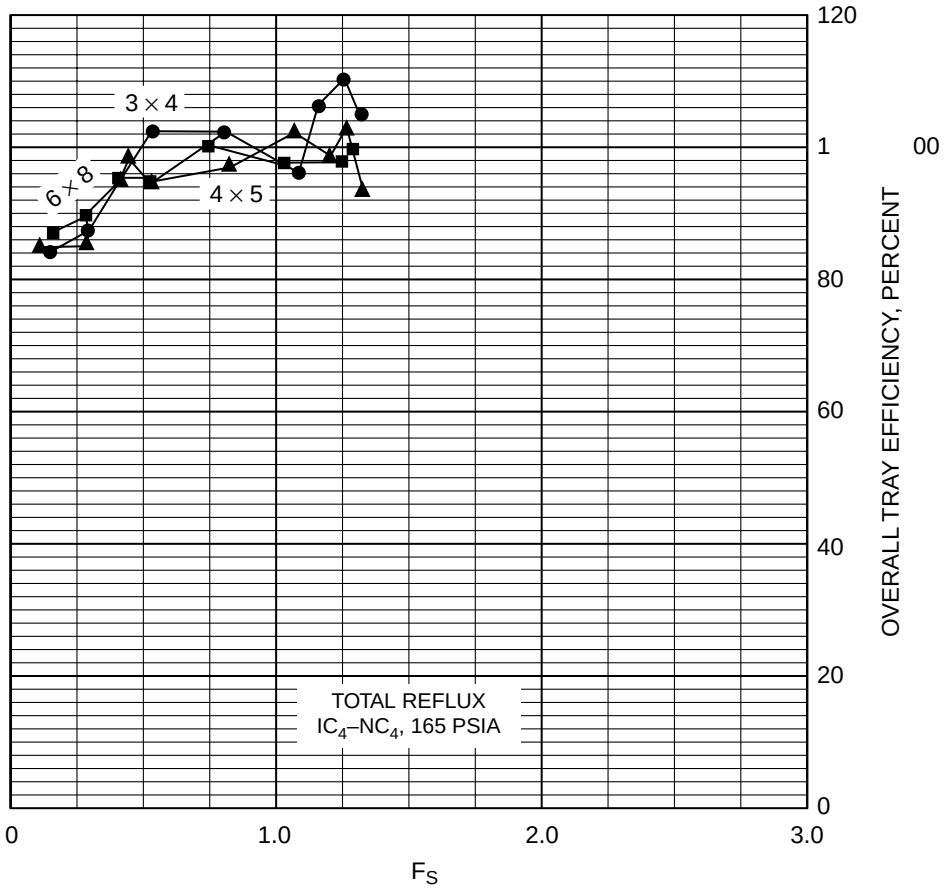


Fig. 12 Tray efficiency of bubble cap trays at high pressure. In FRI 4 ft tower. $F_s = V_s P_v^{1/2}$ (for legend see Fig. 7). (Data courtesy of Fractionation Research, Inc.)

the contacting dispersion is supplied by an external source of energy.

Another advantage of bubble cap trays for vacuum distillation is that there is no tray leakage. At the low liquid rates used in vacuum distillation, tray leakage, if present, would have a large effect on separation.

Bubble cap trays in vacuum service normally have a large number of small caps, say 3 in. diameter. Slot submergence is reduced by using a small exit weir height. The slot submergence is often only 0.25 in. or so to reduce tray pressure drop. Maintaining liquid seal on the tray can be made more certain by recessing the bottom end of the downcomers below the tray deck level. Recessed seal pans have an advantage by allowing larger skirt clearance for any kind of service. As pointed out earlier, there are also disadvantages in the use of recessed downcomer bottoms.

Blowing at Low Liquid Loads

Vacuum towers have high vapor velocities because of low vapor density; they also have correspondingly low liquid loads. The liquid/vapor flow parameter, $F_{lv} = L/G\sqrt{D_v/D_l}$ dimensionless, in which L is the liquid rate (weight units); G is the vapor rate (weight units); D_v is the vapor density; D_l is the liquid density (same units as vapor density); if below about 0.005, is considered to be on the low side. It is possible under such conditions for the high cap-slot velocities to pick up a major fraction of the liquid on the tray and blow it into the exit downcomer, thus bypassing good contact in the active zone of the tray.

Two methods are used to prevent tray liquid blowing. The simplest is probably the picket-fence exit weir (see Fig. 13). In this arrangement, a high exit weir interrupts the trajectory of flying liquid droplets. The slots control liquid depth on the tray.

The second antiblowing technique is to install a vertical baffle upstream of the exit weir. Substantial clearance above and below the baffle allows normal tray function (Fig. 14).

FEEDING ARRANGEMENTS

Feed streams are often introduced by means of a perforated pipe distributor (Fig. 15), which directs the

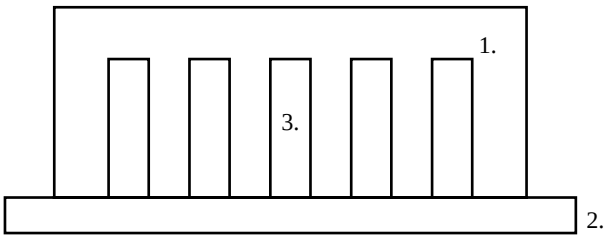


Fig. 13 Picket-fence exit weir: 1) spray-catching baffle, 12 in. or higher; 2) nominal exit weir height; 3) and open slots to allow liquid passage.

flow against the wall of the downcomer from the tray above the feed tray. Often the liquid introduced may not reach equilibrium with the tower internal liquid stream before it reaches the exit weir. To prevent capacity limitation by boiling the liquid in the downcomer, it is a good practice to provide extra spacing for the tray below the feed tray as well as on itself.

Similarly, cold condensate reflux may not be completely heated to top tray temperature by passage over one tray. Providing extra tray spacing below the top tray may prevent hydraulic disturbance at the top tray.

Superheated feeds from chemical reactors or petroleum cracking units are a special case, often requiring a baffle-tray desuperheating section in the tower.

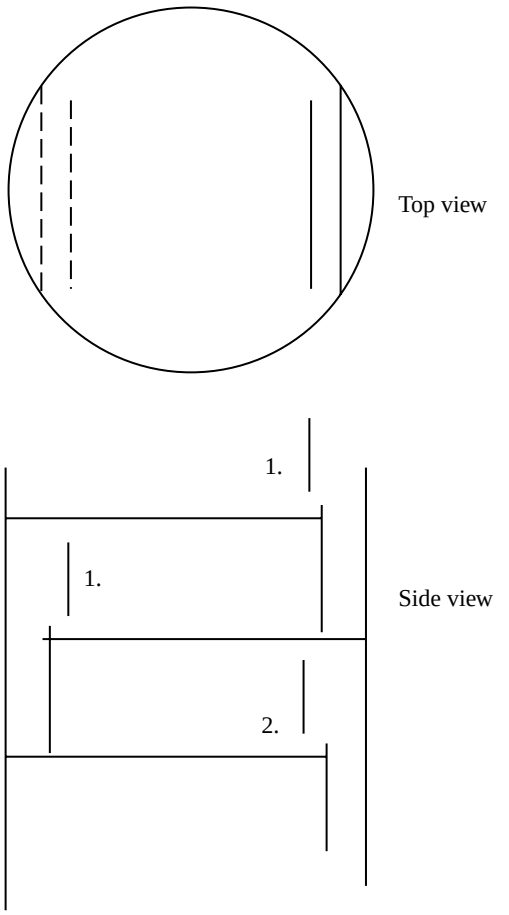


Fig. 14 Antiblowing baffle arrangement: 1) antiblowing baffle and 2) clearance for liquid passage.

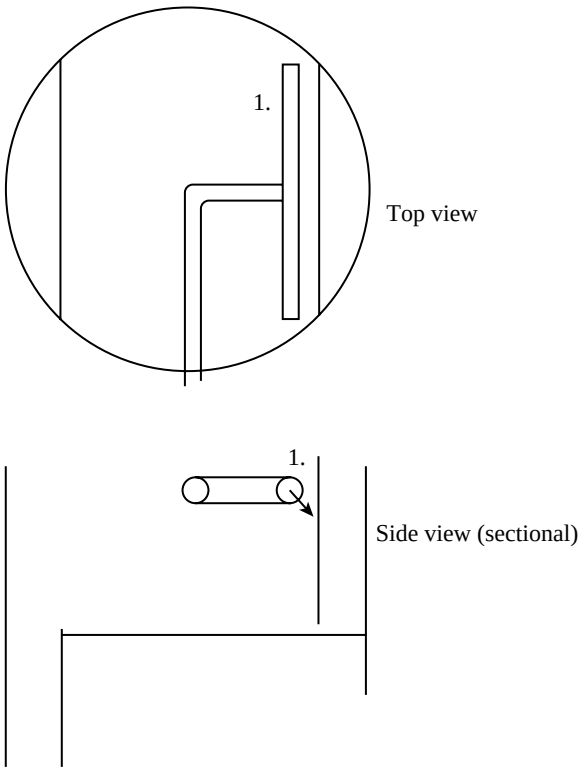


Fig. 15 Slotted feed manifold, showing stream entry. 1) Slotted feed delivery pipe.

SIDE-DRAW PRODUCTS

Partial Draws

If the draw is markedly less in rate than the internal liquid flow, it can be drawn under flow control from an internal or external sump, and arranged so that the tower internal reflux flows through the sump. It is a good idea, nonetheless, to have a seal baffle or a seal pan set up so that if the side draw rate is set higher than the internal reflux flow, the downcomer will not become unsealed.

Total Draw Streams

If the draw stream rate is to be at or near the total internal liquid flow rate, it is best to provide an internal or external sump, with the draw rate controlled by the liquid level. The residual internal reflux can be a controlled flow to the tray below the draw.

Vapor Side Draws

Vapor side draws are rarely used. Usually, extra tray space is provided with a demister mat placed to exclude liquid droplets.

CONCLUSIONS

Bubble cap trays are used in thousands of distilling, absorption, and stripping towers to provide mass and heat transfer between liquid and vapor streams. In new construction projects, valve trays and sieve trays are used more often because of lower capital cost. However, bubble cap trays continue to be preferred when tray leakage must be prevented as, for example, if liquid flow rate is unusually low. They are also superior in effective range of throughputs. It has been found that the best capacity and efficiency are obtained with small (3 or 4 in. diameter) round caps. Those with such caps have a capacity advantage of 10–25% over trays with larger caps, such as those with 6 in. diameter. Apparently, the larger caps have more resistance to liquid flow, causing the liquid to accumulate on the tray, thus leading to more entrainment and downcomer backup.

For pressures near atmospheric, bubble cap trays with the smaller caps are approximately equivalent in hydraulic capacity and separating efficiency to valve or sieve trays.

Modern tray designs often increase hydraulic capacity by modifying the liquid downcomers to increase bubbling area. For example, some trays use hanging downcomers, in which no active tray area is sacrificed for liquid delivery.

REFERENCES

1. Ludwig, E.E. *Applied Process Design for Chemical and Petrochemical Plants*, 2nd Ed.; Gulf Publishing Company: Houston, 1979; Vol. 2, 61–128.
2. Bolles, W.L. Optimum bubble-cap tray design. Pet. Process. **1956**, February through May.
3. Fair, J.R. *Perry's Chemical Engineer's Handbook*, 7th Ed.; McGraw-Hill: New York, 1997; 14–5 through 14–61.
4. Kister, H.Z. *Distillation Design*; McGraw-Hill: New York, 1992.
5. Kister, H.Z. *Distillation Operation*; McGraw-Hill: New York, 1989.
6. Lewis, W.K., Jr. Ind. Eng. Chem. **1936**, *28*, 399.
7. UOP. *Trays for Distillation, Absorption, Stripping & Extraction*; UOP Process Equipment: Tonawanda, NY.
8. Manning, E., Jr. High capacity distillation trays. Ind. Eng. Chem. **1964**, *56*, 14–19.
9. Manning, E., Jr.; Marple, S.; Hinds, G.P. A plant-scale unit for distillation-tray research. Ind. Eng. Chem. **1957**, *49*, 2051.
10. Ross, S.; Nishioka, G. Foaminess of binary and ternary solutions. J. Phys. Chem. **1975**, *79*, 1561.
11. Maragoni, M. Nuovo Cimento **1871**, *2* (5–6), 239.
12. Zuiderweg, F.J.; Harmens, A. Chem. Eng. Sci. **1958**, *9*, 89.
13. Souders, M.; Brown, G.G. Ind. Eng. Chem. **1934**, *26*, 98.
14. Glitsch, Inc. (now Koch-Glitsch). *Ballast Tray Design Manual*, Bulletin No. 4900, 3rd Ed.; p.13.
15. Dauphine, T.C. Pressure Drops in Bubble Trays. Sc. D. Thesis, Massachusetts Institute of Technology, 1939.
16. O'Connell, H.E. Plate efficiency of fractionating columns and absorbers. Trans. AIChE **1946**, *42*, 741.
17. Drickamer, H.G.; Bradford, J.B. Overall plate efficiency of commercial hydrocarbon fractionating towers. Trans. AIChE **1943**, *39*, 319.
18. Simkin, D.J.; Strand, C.P.; Olney, R.B. Entrainment from bubble caps. Chem. Eng. Prog. **1954**, *50*, 565.
19. Bennett, D.L.; Novak, K.W. Optimize distillation columns—part I. Trayed columns. Chem. Eng. Prog. **2000**, *96*, 19–26.

Bulk Molding and Sheet Molding Compounds

Sanjeev N. Rao
Krishnan Jayaraman
University of Auckland, Auckland, New Zealand

INTRODUCTION

Both bulk molding compound (BMC) and sheet molding compound (SMC) are precompounded reinforcements intended for fast processing by compression, transfer, and injection molding. As its name implies, SMC is supplied in sheet form, while BMC, though similar in constitution, is supplied in bulk form, usually as a thick rope or a log. The compounds consist of reinforcing fibers (almost always random discontinuous E-glass), fillers, and resin [which is usually either unsaturated polyesters (UPE) or vinyl ester]. The SMC has a leathery consistency and may be cut with shears or blanked into shapes, whereas BMC is in the form of dough generally extruded as a log or a rope that is weighed and consolidated by hand first. The charge is then placed in the mold. This entry discusses some general aspects of BMC and SMC, their variants, molding technology, reuse, and applications.

BACKGROUND

Polymer composites are widely used in the automobile, electrical, and building industries because of their properties. Earlier, conventional methods (where the fibers and resin were combined and wetted out within the process) were used in the manufacturing of composite parts, but as technology advanced, there have been innovations in the field of formulation, compounding methods, and molding techniques.

These improvements have made it possible to develop compounds with excellent and varied performance characteristics at low compound cost by the choice and regulation of the proportions of the compound ingredients.

BMCs and SMCs are based on unsaturated polyesters and short glass fiber reinforcements. They have found wide applications, particularly in automotive and electrical industries, in parts such as automobile bumpers, light housings, and reflectors because of their low price and easy processing. The earliest BMC was manufactured as early as the 1950s, employing a process that impregnated the glass fibers with resins, fillers, and other additives. Today BMC and SMC are mostly accepted as high-performance thermoset molding compounds used extensively in the manufacture of electrical and automobile components. This entry deals with an overview of BMC and SMC, their manufacturing techniques, compositions, properties, and applications.

BULK MOLDING COMPOUNDS AND SHEET MOLDING COMPOUNDS

BMCs and SMCs are polyester-based thermosetting composite materials that are a premix of fiber reinforcements, fillers, and additives that cure under heat and pressure. Unlike conventional manufacture of composites, BMCs and SMCs are fully compounded products that are readily available to mold. The constituents of both compounds are similar but differ in proportions; BMCs generally have high filler and low fiber content, making it flow easily and improving its moldability. The properties vary according to their area of application;[1] however, they are typically materials having high rigidity and heat resistance, good electrical insulation, shrinkage, and good dimensional stability. BMC is also termed dough molding compound because of its appearance, and it is usually extruded as a long rope or a log, while SMC is usually in the form of sheets that appear in the form of continuous rolls. Typical SMC sheets are 1–2 m wide and around 5 mm thick.

FORMULATIONS

The main constituents of BMC and SMC are thermosetting resin, fiber reinforcements, and some fillers. Table 1 gives the formulations of BMC, SMC, and their variants.

Resins

The polymer component, which is a thermosetting resin, forms cross-linked bonds between chains of molecules, which enhances the characteristic of the product. Unsaturated polyesters and vinyl esters are primarily used as resin systems. Epoxies are also used in some cases, but the cure cycle is longer. Phenolic resins have gained importance these days because of their inherent properties, especially in applications that require lower flammability, reduced smoke generation, and higher thermal stability. New resin systems such as

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007684

Table 1 Formulation of BMC/SMC and their variants

	BMC (PHR[a])	SMC (PHR)	ZMC[b] (PHR)	LPMC[c] (PHR)
Polyester resin	60.0	55.0	65.5	65.0
Low profile additive	40.0	40.0	40.0	
Styrene	5.0	5.0	5.0	5.0
Initiator	1.5	1.5	1.5	1.2
Inhibitor	Trace amount	250 ppm	100 ppm	
Mold release	4.0	4.0	4.0	5.0
Pigment	0.25	1.0		1.2
Thickening agent	1.0	2.0		
Filler	50–200	150–250	220	220
Glass fiber[d]	10–25%	25–30%	15–25%	25–35%
Paste	75–90%	70–75%	75–85%	70–75%

[a]Per hundred resin.
[b]Z Molding Compound.
[c]Low Pressure Molding Compound.
[d]25.4 mm used for SMC.
(From Ref.[2].)

a hybrid of unsaturated polyester and urethane are also becoming popular.[2]

Fiber Reinforcement

Glass fibers used for reinforcement have the obvious effect of strengthening the material. The length of the glass fibers varies between 3 and 25 mm depending on the type of mold and the molding conditions. E-glass is most commonly used as reinforcement in the composites; other fibers used include carbon, amarid (Kevlar), and S-glass. The glass fiber loading is normally around 30% by weight in compression-molded composites, but can vary from 18% to 65%.[2] Table 2 shows typical fibers used in polymer composites.

Natural fibers have gained importance because of their biodegradable nature. While natural fibers offer properties similar to synthetic ones, their moisture absorption characteristic undermines their efficient use in composite manufacturing.[3] The use of natural fibers goes way back to 1950s when sisal fibers were used as reinforcements for producing automobile heater housings.[2]

Fillers

Both BMC and SMC can accept relatively large filler loadings; most of the fillers used are inorganic and are of mineral type.[4] The fillers used usually produce the following effects:

1. Increase in viscosity of the liquid resin system.
2. Reduction of shrinkage during curing.
3. Act as a heat sink by decreasing the peaking exothermic reaction on cure.
4. Increase in specific gravity of the resin (by using denser fillers).
5. Acceleration or retardation of gel time during cure.
6. Reduction in cost of the composite as compared to pure raw material.

Table 2 Common fibers used in BMC and SMC

Fiber	Specific gravity	Tensile strength (GPa)	Tensile modulus (GPa)	Tensile failure strain (%)	Coefficient of thermal expansion ($\times 10^{-6}$/°C)
E-glass	2.54	3.45	72.2	4.8	5
S-glass	2.48	4.30	86.9	5.0	2.9
Carbon (graphite)	1.76–2.15	1.5–5.6	220–690	0.3–1.2	10.1–1.2 (longitudinal), 7–12 (radial)
Kevlar 49	1.45	3.62	131	2.8	−2 (longitudinal), 59 (radial)

(From Ref.[2].)

The fillers are normally added after thorough dispersion of catalyst in the resin and other additives such as internal mold release, pigments, wetting agents. But the addition may precede the dispersion of catalyst if it is a shrink control filler. Table 3 gives some examples of the fillers used.

Calcium Carbonate Fillers

Calcium carbonate fillers are either of ground limestone type or precipitated chalk type. They are inexpensive and have low oil–absorption, which allows considerable freedom in filler levels while keeping the viscosity under control. Relatively coarse particle size fillers ranging from 30 to 100 mesh are used to incorporate high loadings.

Clay Type Fillers

Clay fillers are used to improve electrical properties of the compound and enhance flow in molding. They are also chemically inert when compared to carbonate fillers, which enhances their chemical resistance.

Alumina Trihydrate Fillers

Alumina trihydrate fillers (ATH) are largely used for imparting flame retardation property to the compound. This property conferred is because of the energy absorption that occurs as three water molecules are liberated from each molecule of ATH at combustion temperatures. Both SMC and BMC employ filler loading levels from 150 to over 200 parts of ATH filler per 100 parts of resin.

Pigments and Colorants

Inorganic pigments such as titanium oxides, and organic pigments such as the phthalocyanine pigments are used as dyes. Care must be taken when using colorants in polyester pigments so that they do not interfere with the proper cure of the resin. With peroxide catalyzed systems, the colorants must also be able to resist the oxidizing action of peroxide.

Thickeners

Thickening agents are integral parts of BMC and SMC. They are used to thicken the molding compounds and make them easy to handle. The most common thickening agents are metal oxides and hydroxides of magnesium and calcium.

Internal Mold Release Agent

Internal mold release agents are added to facilitate part ejection from the mold. The most common types of agents include metal stearates, fatty acids, fatty acid amides, and hydrocarbon waxes. Zinc stearate and calcium stearate are the widely used mold release agents in SMC and BMC.

Low Profile Additives

Low profile plastics are added to reduce shrinkage during cure. They are normally thermoplastics that include polyvinyl acetates, polymethyl methacrylate, and copolymers with other acrylate, vinyl chloride–vinyl acetate copolymers, polyurethane, polystyrene, polycaprolactone, cellulose acetate butyrate, saturated polyester, and styrene butadiene copolymers. More details about the low profile additive (LPA) mechanism are published in the literature.[5]

Table 3 Additive families developed for shrinkage control properties

Category	Examples
First generation. Nonpolar polymer additives that are noncompatible with the resin system	Polystyrene Polyethylene
Second generation. Systems with reduced or no molding shrinkage developed for automotive and other commercial applications	Polyvinyl acetates Polymethylmethacrylate and copolymers Polyvinyl chloride–vinyl acetate copolymers Styrene–butadine copolymers Cellulose acetate butyrate Modified polyvinyl acetates
Third generation. Highest levels of shrinkage control. Accuracy of mold reproduction and surface smoothness	Dual thickening systems containing isocyanate prepolymers Saturated polyesters Saturated polyester blends with polyvinyl chloride Polyurethanes

(From Ref.[1].)

Initiators and Inhibitors

Initiators are used to initiate the curing reaction at elevated temperatures. Cross-linking or polymerization occurs by a free radical mechanism in which the double bond of the polyester chain reacts with the vinyl monomer that is usually styrene, and this reaction provides a three-dimensional network that converts the viscous resin to a hard thermoset solid. The initiators added decompose at elevated temperatures thus providing free radicals to initiate the cross-linking. Peroxyesters and peroxyketals are the most common classes of peroxides used as initiators.

Inhibitors are added in small quantities to increase the shelf life of the compounds. They also help in modifying the cure rate and the magnitude of the exotherm to prevent cracking in thick sections of the molded components. Substituted phenolic derivatives and the quaternary ammonium salts, like hydroquinone and *p*-benzoquinone, are two general classes of inhibitors usually used. More information about initiator chemistry is available in the literature.[5]

MANUFACTURE OF BMC/SMC

Standard SMC consists of chopped strand E-glass blended with resin and fillers supplied in continuous rolls. The schematic of the manufacturing process is shown in Fig. 1. Two sheets of polyethylene films are brought together from different levels and are coated with a resin that has been mixed with fillers, initiators, inhibitors, and thickening agents. The glass fibers are deposited over the coated film and are consolidated to homogeneous sheets. On exit from the machine, the sheets are coiled into rolls and stored in a maturation room.

BMCs are manufactured in a large hopper where all the ingredients are mixed with Z-blades for a set time.

VARIANTS OF BMC/SMC

Z Molding Compounds

Z molding compound (ZMC) was first prepared in France in 1979.[2] This compound needs a special type of injection machine—a combination of plunger and screw—and was indigenously developed by *Billion* in France. The machine uses a screw to homogenize and measure the shot. The injection is made like a plunger by the displacement of the screw and inner barrel inside the main barrel. Compared to SMC, ZMC parts have lower mechanical properties, but higher performance when compared to conventional injection molded BMC.

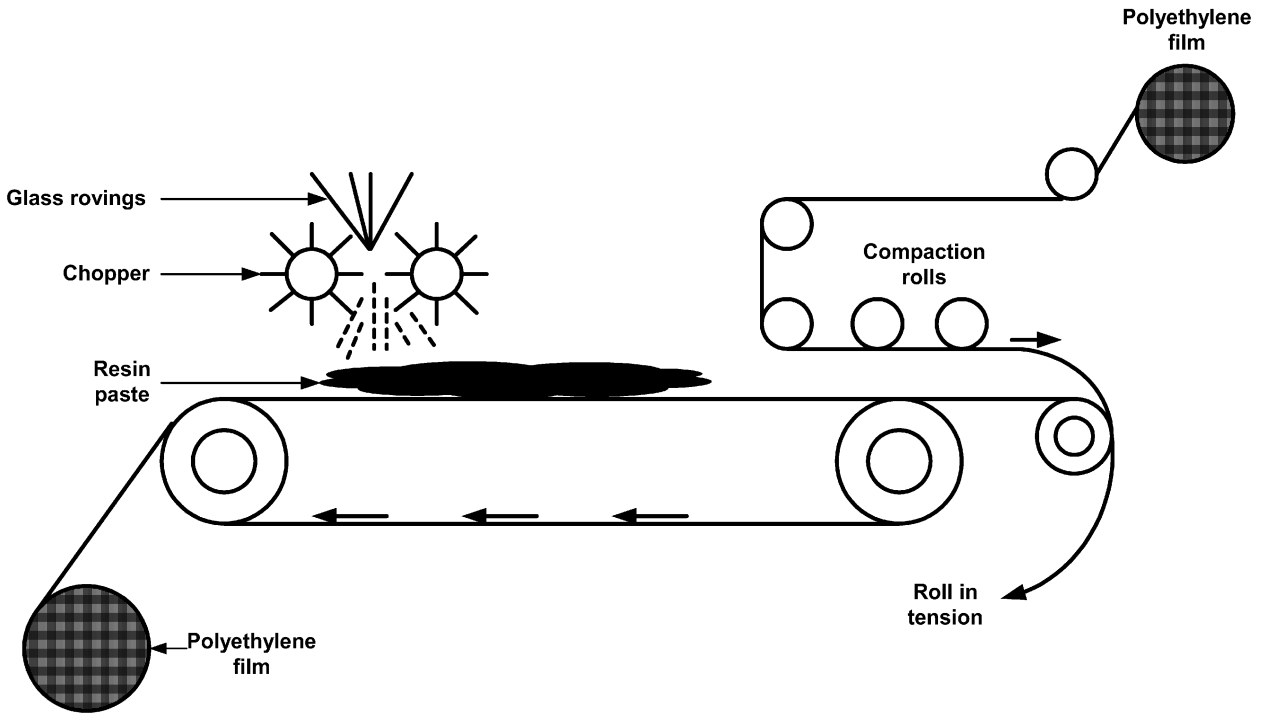


Fig. 1 Manufacture of SMC.

Thick Molding Compounds and Continuous Impregnated Compounds

Thick molding compounds (TMCs) (Fig. 2) and continuous impregnated compounds (CICs) (Fig. 3) are rather recent developments in the manufacture of polyester compounds.[1] They are manufactured by a continuous mixing process that enables high filler and glass loadings. The compounding and equipment used are similar for both the compounds. The resin premix is distributed on two large mixing rolls and then the chopped glass fibers are distributed onto the rolls, and efficient mixing and impregnation takes place as the resin–glass compound passes between the rollers. In TMC, the compound is sandwiched between polyethylene films and passed through impregnation rolls. But in CIC, the compound is removed from the mixing rolls by doctor blades and transported by a screw or plunger into boxes or drums to give the material the physical characteristic of BMC. Modified CIC is also known as kneaded molding compound (KMC).[2]

Low Pressure Molding Compound

Low pressure molding compound (LPMC) is an SMC kind of material. LMPC is made by replacing the chemical thickening mechanism with a physical thickening mechanism using crystic polyester. The material is first heated to melt crystic polyester, and then the other additives are added and run on a modified SMC machine at elevated temperatures. By using cooling rolls, the cooling process can be paced and thus the material could be molded right off the SMC machine without having to wait for 24–48 hr, which is the standard thickening time for SMCs.[2] The shelf life of LMPC is much longer than that of SMC, and the physical properties comparable. LPMC allows the use of lower tonnage presses to mold larger parts, thus reducing the tooling costs.

Glass-Mat Reinforced Thermoplastic

Glass-mat reinforced thermoplastic (GMT) is similar to SMC and is available in sheet form, reinforced with random oriented fibers that may be continuous or discontinuous. It is manufactured on a double belt press where the glass fiber mats are sandwiched between layers of extruded molten polymer before it enters the press. The thermoplastic is usually about a few millimeters thick and thus stiff, and hence stored as flat sheets. Another method involves deposition of a molten mixture of resin, chopped fibers, and additives over a moving belt where the water is driven off. The fiber volume fraction in this compound is usually 0.1–0.3 and fiber lengths are in the range of 10–30 mm, unless the reinforcement is continuous.

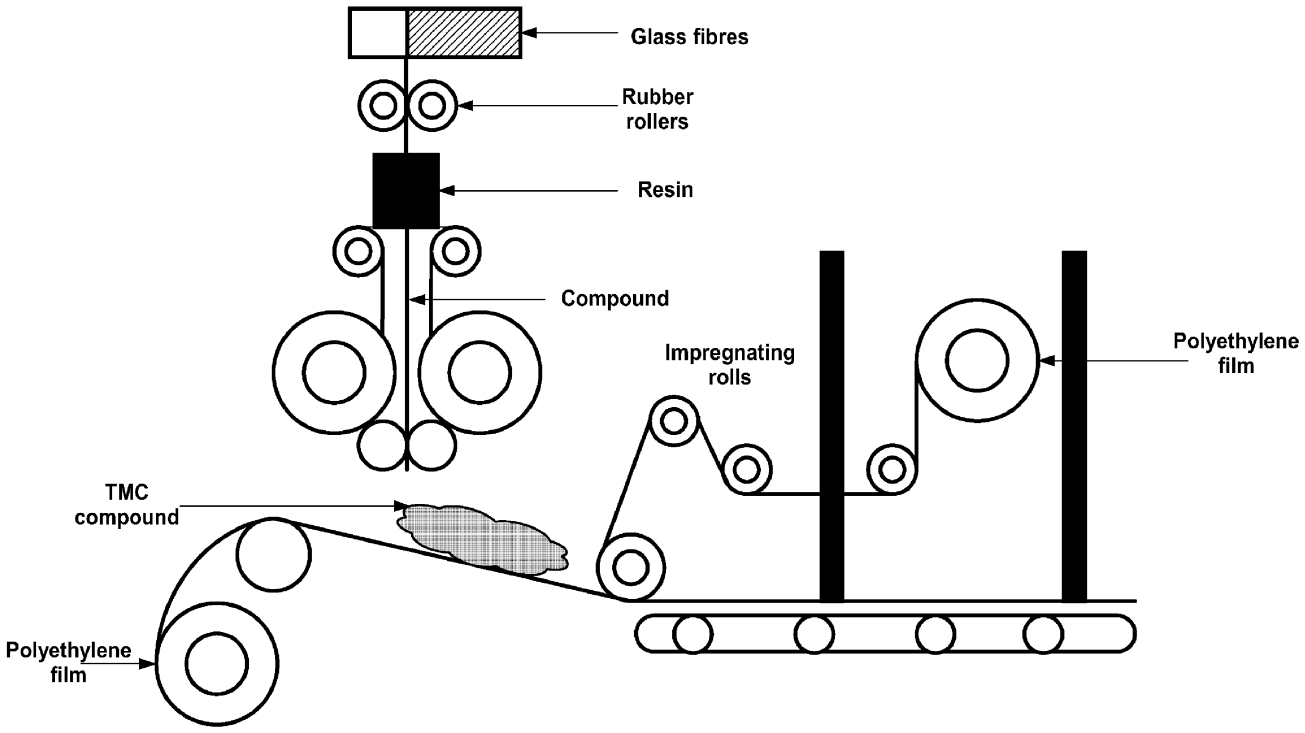


Fig. 2 Manufacture of TMC.

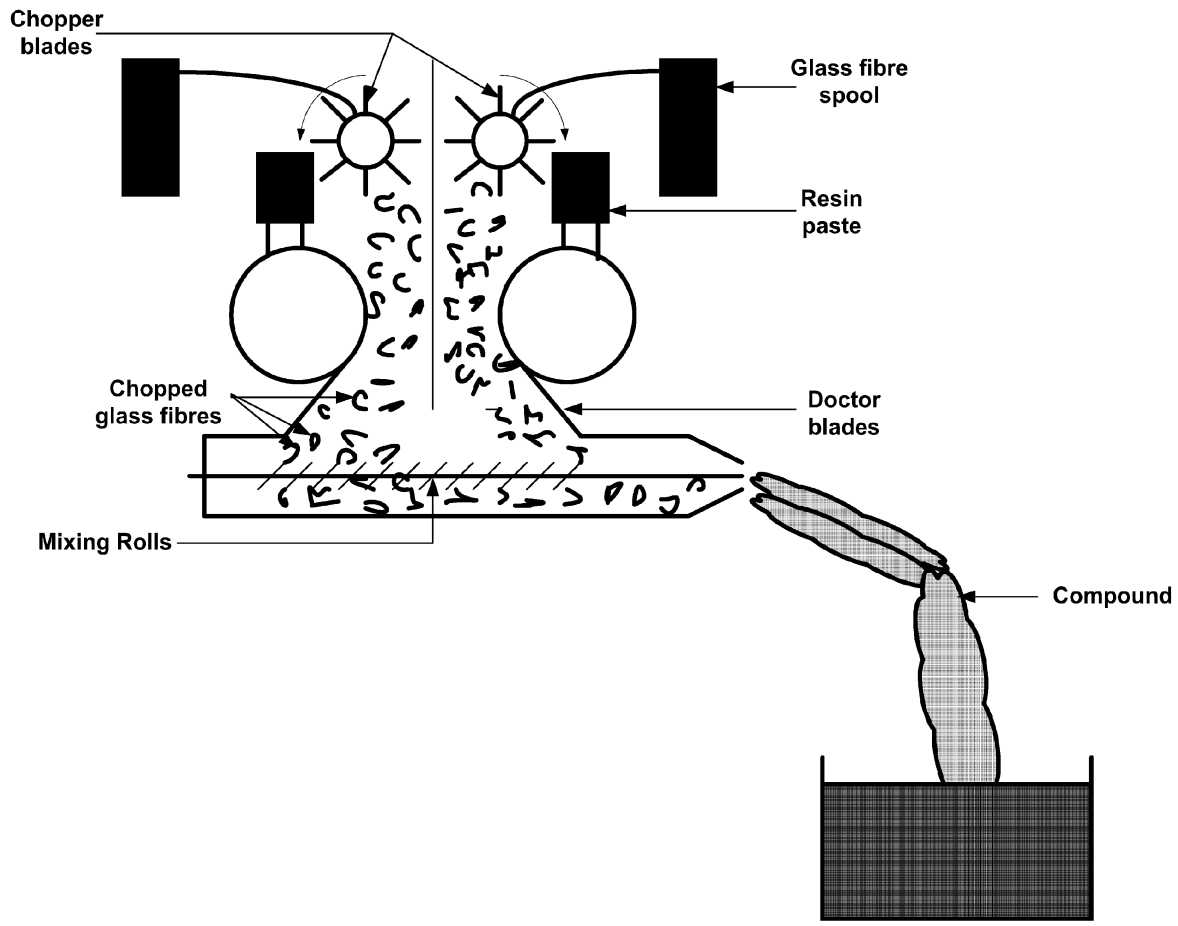


Fig. 3 Manufacture of CIC.

MOLDING OF BMC/SMC

Compression Molding

Matched metal compression molding is one of the oldest manufacturing techniques in the plastics/composites industry. Compression molding (Fig. 4) involves placing a charge into a heated mold and using the closing action of the mold to form the required shape, with cure taking place within the mold. The general technique of the molding process is the same for both SMC and BMC, but the parameters such as pressure and temperature may vary. The pressure applied is low for BMC (35 bar is sufficient for most moldings) and considerably high (70 bar) for SMC to make them flow in the mold. Operating temperature is around 135°C for SMC while it is 120°C for BMC, owing to the fact that BMC flows easily at room temperature. Compression molding is a complex process, and it may result in some surface imperfections that can be controlled by proper material selection, part design, and molding techniques. Compression molding involves four distinct steps.[2]

Charge preparation and placement

The sheets of SMC are cut into a desirable shape (charge is weighed in case of BMC) and stacked on the lower half of the mold, and the movable upper half is brought down to close the mold. BMC material is normally consolidated by hand before they are placed in the mold; in most cases the charge is placed centrally in the mold. The placement determines the quality of the part as it influences the length of the flow in the mold, fiber orientation, flow line, and other surface defects.

Mold closing and filling

Once the charge is properly placed in the mold, it is closed quickly to contact the top surface of the charge. Mold closure is a critical factor: delay in closure time may cause some precure owing to the gelation of the top surface, occurring before the part is molded, and faster closure rates trap air, which induces surface imperfections, which is undesirable. A 10-sec closure time has been suggested as optimum to avoid any precure.[1]

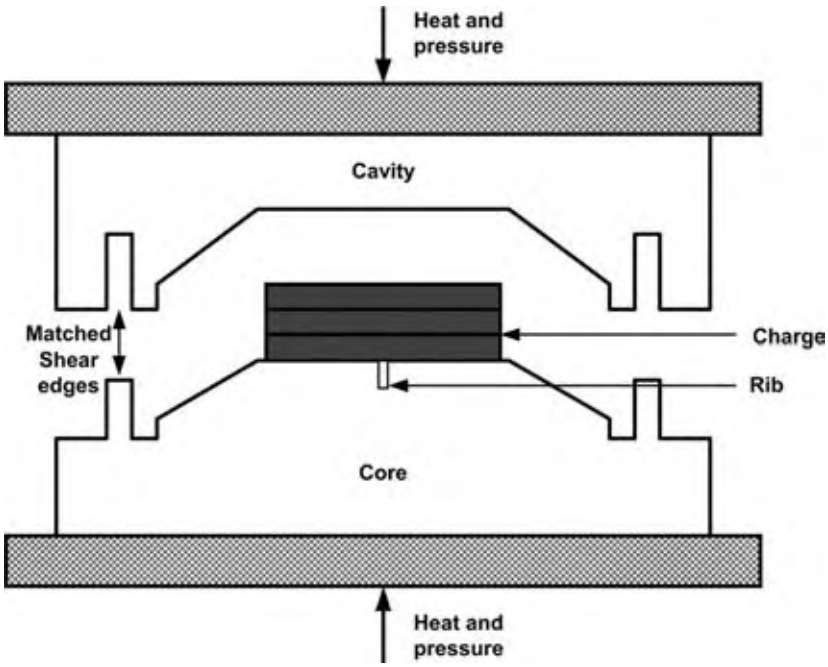


Fig. 4 Compression molding technique.

Curing

After filling, the charge remains in the hot mold for the cross-linking reaction to be completed and solidify. The curing time depends on several factors such as resin–initiator–inhibitor reactivity, part thickness, and mold temperature. More about curing is explained in the cure section.

Part ejection and post cure

At the end of the molding cycle, the mold is opened and the part is ejected from the core. The part is allowed to cure under ambient temperature, during which shrinkage may occur because of the residual stresses developed owing to differential cooling at variable sections of the part.

Transfer Molding

Transfer molding process (Fig. 5) is another alternative process for compression molding so as to accommodate multi-impression molding of the same material. This process is mainly used in molding parts that involve a lot of insertions into the part. Integral type transfer molding is commonly used, but where auxiliary ram type is used, the pot is maintained at lower temperature than that of the mold, thus maintaining the BMC at higher viscosity and rendering the flow control easier. It is often seen that the strengths of the transfer molded components are lower compared to compression molded ones, which is because of some damage caused to the glass fibers when the compound passes through the runner and the gate system. However, this could be minimized by designing larger gate and runner sizes. An important factor to be considered while designing a mold is the fiber orientation, which depends on the size and shape of the mold cavity and the point where the material enters first. The operating temperatures and curing rates are more or less the same as in compression molding. The main advantage of this process is high production volumes because of the facility of feeding multiple cavities from one charge of BMC in the transfer pot.

Injection Molding

Injection molding is a completely automated process, providing economic benefits over other processes. It consists of an injection screw and a barrel, with the screw having constant pitch to accommodate constant feed rate. Fig. 6 shows a schematic of a simple force feed hopper system consisting of a pressurized ram and a cylinder that forces the BMC to the start of the screw. The jacket provided is used to heat the barrel and the charge passing through it. The flight of the rotating screw causes the material to move through a heated ''extruder'' barrel, where it softens (being made fluid) and so it can be fed into the shot chamber (front of screw). This motion generates a pressure that causes the screw to retract, and when the preset limit is reached and the shot size met, the screw stops rotating. At a preset time, the screw acts as a ram to push the melt into the mold.

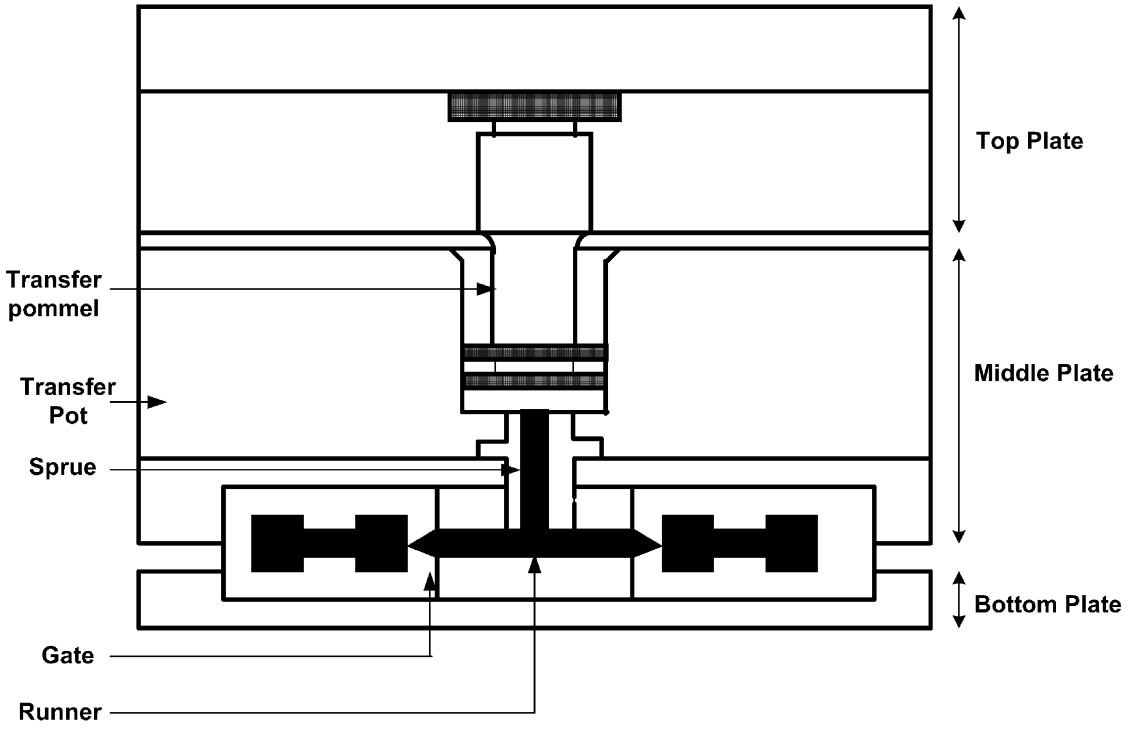


Fig. 5 Transfer molding technique.

Factors influencing injection molding

Time, temperature, and pressure are the primary factors influencing the surface finish of the molded part. Time factors include rate of injection, duration of ram pressure, time of cooling, cross-linking time, and rotation of the screw (RPM). Pressure factors are applied pressure (high or low), back pressure on the extruder screw, and pressure loss before the charge enters the cavity, which can be caused by a variety of restrictions in the mold cavity. Temperature factors are mold (cavity and core), barrel, and nozzle temperatures, as well as the melt temperature because of back pressure, screw speed, and frictional heat.

GENERAL MOLD DESIGN CONSIDERATIONS

All mold components contacted by the molding compound—runners, gates, and cavities—should be made of hardened steel, hardened to 65–68 on Rockwell scale, highly polished, and hard chrome plated.

The ejector pins should have an adequate cross-sectional area to minimize the possibility of distorting or puncturing the molded plastic at the time of ejection from cavities, as most of the thermoset compounds are slightly soft during that time.

In automatic molds, it is vital to ensure that the undercuts, hold down pins, and the molded part remain in the desired half of the mold, so that when the parts of the runner system are ejected, they do not get dislocated with that movement.

Flash removal at the end of each cycle is important for successful automatic molding.

The molds should have the ability to maintain constant temperature despite continual heat removal by the relatively cooler molding compounds; this can be monitored by placing temperature sensors and heat cartridges to provide uniform temperature. Insulating blankets may prove beneficial in minimizing mold heat losses and variations caused by local air currents around the molds and presses. A cut out should be provided, which will cut off the power to the heat cartridges as and when it senses a few degree rise than the desired temperature.

COMPARISON OF MOLDING PROCESSES

BMC can be molded by compression-, transfer-, or injection molding processes. Compression molding is the oldest and most commonly used process. It consists of forming the material between the heated metal punch and die. This process is the most economic one; however there are a few components that cannot be compression molded owing to their intricate shapes. Another drawback of compression molding process involves weighing of the charge before it is laid out

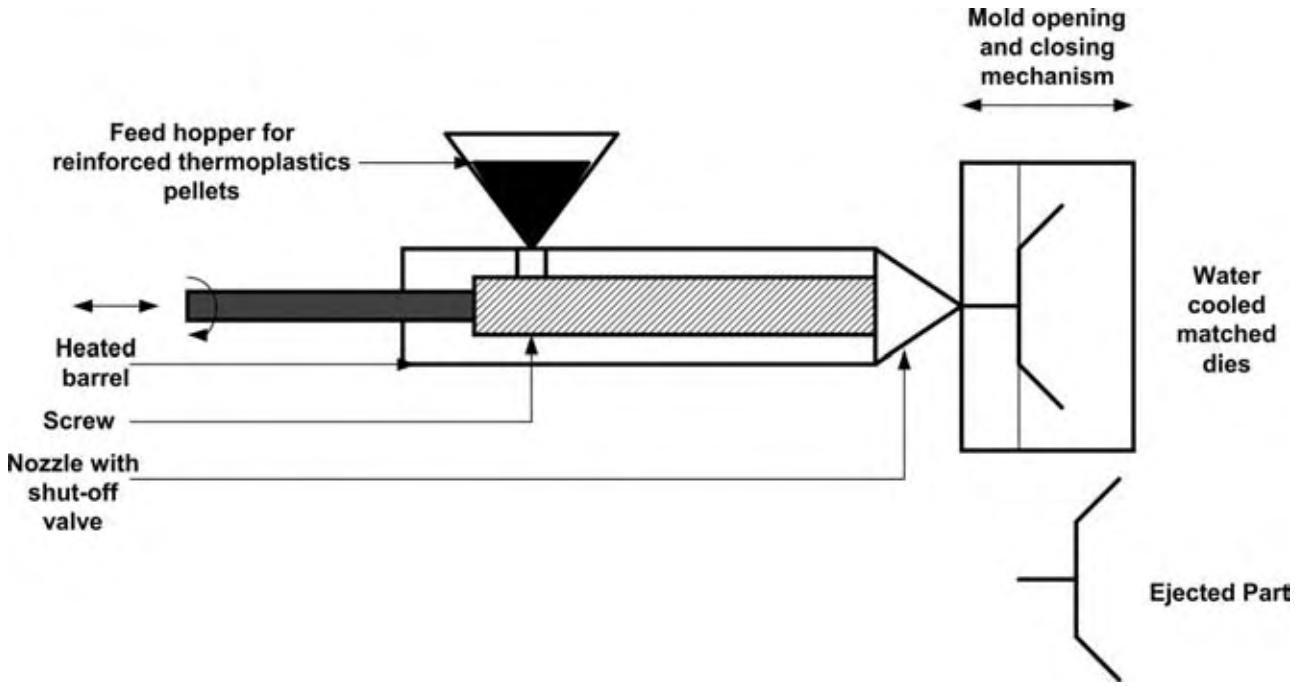


Fig. 6 Injection molding process.

in the mold cavity, which would make it tedious for the molder. Transfer molding or injection molding processes are preferred to avoid this difficulty and reduce production cycle time and labor costs. Transfer molding again involves weighing out and dispensing of one quantity of charge per production cycle, which is still a manual operation, but injection molding process provides its own means of weight calculation, no matter whether it is a single or a multi-impression mold.

The production cycle is delayed in compression- and transfer molding as it takes time to bring the charge to the mold temperature where the cross-linking reaction initiates, thereby increasing cure time. This is avoided in injection molding as the charge is preheated in the barrel of the injection unit, and moreover because of the frictional heating as the material passes through the runner system and the gate. As a consequence, the charge enters the mold at a higher temperature and the cross-linking reaction starts as soon as the charge enters the mold, thus reducing the cure time.

Cure time for compression molding is influenced by the molding thickness, and very thick moldings may require more time than thin ones. Compared to this, injection molding is much less influenced by the molding thickness, and very thick moldings may require a little more cure time that thin ones.

Another difference between the processes is the flash formation and its removal. As it is reinforced with fibers, it becomes difficult to remove the flash formed. As compression molding uses semipositive molds, there would be some material loss in the form of flash, and deflashing is normally done by hand. This process would make it more expensive and less satisfactory, because the finish line is visible and this may be objectionable. Injection molding would prove to be advantageous as the flash formed is less, because the mold is already closed and the two halves fit close to each other (Table 4).

As BMC and SMC are fiber reinforced, fiber orientation is very important. During compression molding, the central loading of the mold may lead to radial orientation of the fiber as the material flows out to the extremities of the cavity. Transfer- and injection molding are likely to give stronger orientational effects because the filling of the cavity is from one fixed point, which is the gate. The location of the gate depends on the mold designer and many other parameters that are beyond the scope of this entry.

Comparison of the molded parts yields a conclusion that injection- and transfer molding would be advantageous in mass production as the production cycle time is much shorter compared to compression molding. However, compression molded parts are much stronger compared to injection- and transfer molded parts, which is because of the damage done to the fibers when the material flows through the runner system and the gate of the mold. Another aspect of quality of the product is surface finish, which is not consistent and attractive in the compression molded parts, as the handling of material may transfer some dirt from the hand to the working surfaces. As BMCs are fast curing, some amount of precure is inevitable before the

Table 4 Comparison of typical production rates for different processes

	Small molding[a]			Large molding[b]		
Cycle element	**Compression**	**Transfer**	**Injection**	**Compression**	**Transfer**	**Injection**
Cure time (sec)	30	30	25	45	40	30
Press open time (sec)	50	15	10	15	15	10
Total cycle time (sec)	80	45	35	60	55	40
Saving as a % of compression cycle time	—	44	56	—	8	33

[a]Thin section/10 g weight, 12-impression mold.
[b]Thick section/1000 g weight, single impression mold.
(From Ref.[1].)

mold is closed. Injection molding differs significantly as the material is transferred very rapidly from the barrel of the machine to the cavity, and thus any precure can be avoided; being an automated process, it involves less material handling.

CURE OF BMC/SMC

Thermosets are widely used for their cost effectiveness and outstanding mechanical properties, especially at high temperatures. Unfortunately, they run the risk of being over- or undercured, which plays a substantial role in determining the product quality. The term ''cure'' refers to the monomeric liquid resin becoming a cross-linked rigid solid, which produces a tightly bound three-dimensional network.

As the quality of the product predominantly depends on the curing reaction, monitoring the cure reaction is most important. Infrared spectroscopy, nuclear magnetic resonance spectroscopy, and chromatography are typical methods in research and development. However, for practical reasons, these methods are difficult to incorporate into the production process. The macroscopic attributes that change during cure are mechanical and electrical material parameters. Typical mechanical parameters are complex elastic or shear modulus, and viscosity. These quantities can be measured by dynamic mechanical analysis (DMA), ultrasound propagation, and viscosimetry.[6] Table 5 shows the sensors incorporated in the molds for monitoring and their relative costs.

Computer simulation software are commonly used as cure monitoring tools basically in visualizing the flow front of the charge inside the mold. They reduce time and cost of the cure monitoring. The basis for the models developed for simulation is the analysis of the flow behavior taking into account the thermoset cure. The flow analysis would involve calculation of elementary viscosity, pressure distribution as the flow front expands, correcting at all times for elementary cure level. Temperature calculation is based on the equation that accounts for convection, conduction, viscous heating, and cure reaction. The elemental cure time can be calculated from the material database developed, which stores data files on curing behavior of the compound such as viscosity, gel time, and heat of reaction. The development of these simulation programs depends on the need to standardize material tests necessary to provide reliable data for computation.

PROPERTIES OF BMC/SMC

Properties of a thermoset can be varied by using different formulations such as fiber content, filler content, etc.; this uniqueness makes it difficult to generalize the definition of properties of the material. Its characteristics are used as guidelines in part design and material selection to understand the effect of the change in formulation on mechanical properties. Table 6 shows static and impact properties of SMC, BMC, LPMC, and ZMC. Tensile and flexural properties are routinely

Table 5 Sensors incorporated in the molds for monitoring and their relative costs

	Configuration of use	**Monitored value**	**Cost**	**Use in operation**
Fiber-optic sensors	Embed	Chemical constitution refractive index, strain, temperature	High	Good
Dielectric sensors	Embed	Electrical permittivity	Intermediate	Poor
Piezoelectric sensors	Embed	Impedance	Intermediate	Good
Ultrasonic sensors	Attach to die	Sound velocity, attenuation	Low	No

(From Ref.[6].)

Table 6 Static and impact properties of SMC, BMC, LPMC, and ZMC

	Tensile strength (MPa)	Tensile modulus (GPa)	Flexural strength (MPa)	Flexural modulus (GPa)	IZOD impact (un-notched)	Specific gravity	Coefficient of thermal expansion ($\times 10^{-6}/^{\circ}C$)
SMC	65–100	9.5–14	130–200	8–14	600–1200	1.3–2.0	8–14
BMC	30–70	8–12	50–150	9–17	100–700	1.7–2.1	15–20
ZMC	30–70	8.5–12.5	50–150	7–12	200–500	1.8–2.0	11–27
LPMC	65–100	9.5–14	120–200	8–14	600–1200	1.8–2.0	7–10

(From Ref.[2].)

measured in general; compression and shear for special purposes. The static properties are highly dependent on the fiber length, loading, and its orientation. Tensile strength increases rapidly with the increase in fiber content and fiber length, but the modulus is affected moderately. The type of fiber being used as reinforcement plays an important part (E-glass, S-glass) in both tensile strength and modulus during low fiber loading.

Several other tests are used to correlate properties with the working conditions. The dynamic mechanical analyzer (DMA) is used to measure modulus and viscoelastic properties related to ductility. By using the DMA test method, the degradation in modulus could be understood and applied to the end use application. The effect of the environmental conditions is also tested; the esthetic durability in heated environments is not restricted to warp and relaxation attributed to creep. Creep and stress relaxation tests are also done on SMC/BMC for their structural applications. The retention of color is critical for a heated appliance application. BMC/SMC can be formulated to offer excellent color stability when exposed to high temperatures.

RECYCLING OF BMC/SMC

Thermoset polymers are considered to be nonrecyclable plastic materials. Once they are formed into a shape and cross-linked, they cannot be remolded or melted. When they are used as matrix materials for composites, the recycling problem becomes more difficult because of the multiphase nature of the composite. Some methods of recycling[7] are explained in the following sections. Table 7 gives the property of BMC reinforced with recycled glass reinforced SMC.

Mechanical Degradation

Mechanical recycling methods are based on the use of SMC scrap directly without altering its chemical properties. The SMC scrap is shredded or ground using mechanical hammers into a form that can be used as fillers. The entire SMC could be ground into fine powder (including glass fiber). The other method is to mill the SMC in such a way that some glass fibers could be recovered, and later separated and reused as reinforcement. Recovery of some glass is very important from the cost effectiveness point of view as the fibers have more value as reinforcement. Moreover, it requires more energy to grind all the fiber into powder.

Pyrolysis

Pyrolysis can be considered as a nonconventional method of recycling plastic materials and is especially

Table 7 Property of BMC reinforced with recycled glass reinforced SMC

		Tensile			Flexural		Izod impact	
Sample	SMC added (%)	Strength (MPa)	Modulus (MPa)	Elongation (%)	Strength (MPa)	Modulus (MPa)	Notched (J/m)	Un-notched (J/m)
BMC control	0	27.9	13,099	0.44	96.7	10,548	270	361
Standard BMC + SMC	6	16.1	9,858	0.68	68.6	9,789	270	278
Standard BMC + SMC	12	17.3	12,685	0.22	71.4	10,203	209	274
BMC resin + SMC	70	25.8	7,859	0.36	55.4	6,411	89	114

(From Ref.[7].)

appropriate for thermoset composites that cannot be remolded. In the pyrolysis process (heating without oxygen), the organic part of the material is decomposed to gases and liquids, which can be useful as fuels or sources of chemicals. Inorganic components such as glass fibers and calcium carbonate filler ($CaCo_3$), which are left relatively unmodified, could be reused as a whole or individual components in other composite materials.

Chemical Degradation

This process involves reversing chemical reaction where the organic polymer is converted back to more basic chemical building blocks. Among these processes, hydrolysis, glycolysis, methanolysis, and aminolysis are well known. This depolymerization process will also free the glass fiber reinforcement, which is an added value.

Heat Cleaning

This process involves completely volatizing the resin and freeing the glass fibers. The SMC is passed through a furnace maintained at a set temperature where the compound becomes completely volatile and frees the glass fibers that can be recovered. The gases liberated during the process are also potentially useful by-products.

GENERAL APPLICATIONS

BMC provides cost advantages particularly for high load and temperature bearing products, thus finding application in electrical and mechanical industries. The BMC/SMC are tested for short circuiting where heaviest current loads are applied, which produces an electric arc inside the switch and raises the temperature to several thousands of degrees in a short period of time. BMC is used for household appliances, circuit breaker housings, heat shields for electric irons and headlamp reflectors. BMC/SMC provide economic solutions for electrical equipment because of its electrical insulation, thermal resistance, and freedom of design of properties. BMC has become a choice for low and high voltage applications such as switches and arc barriers. They provide efficient power distribution and protect the end user from dangerous effects of short circuiting.

As SMC is transparent to electromagnetic waves and offers additional benefits such as low cost compared to metals, dimensional stability, stiffness, low weight, and high temperature resistance, DaimlerChrysler is considering using class A SMC to manufacture body parts in one of their coupe models.[8]

CONCLUSIONS

The compounds, both SMC and BMC are pre-compounded reinforcements designed for fast processing by compression-, transfer-, and injection molding. Unlike conventional methods of composite manufacturing where the fibers and resin were combined and wetted out within the process (which was also time consuming), BMC/SMC offer the advantage of being ready to mold mix, which saves a lot of time and hence production volumes. Injection molding and transfer molding processes are used for mass production, and by optimizing mold design, high volume products with superior surface finish can be obtained. As BMC/SMC can be tailored accordingly, products with high mechanical strengths, and electrical and thermal properties can be obtained. Recyclability and reuse of SMC/BMC parts add to its advantage. Hence, the compounds find a wide range of applications in mechanical and electrical industries.

REFERENCES

1. Monk, J.F. *Thermosetting Plastics*, 2nd Ed.; Addison Wesley Longman: Dorchester, 1997; 238 pp.
2. Haque, E.; Burr; Leach, L. Matched metal compression moulding of polymer composites. In *Handbook of Composites: Matched Metal Compression Moulding of Polymer Composites*; Peters, S.T., Ed.; Chapman & Hall, 1998; 378–396.
3. Bledzki, A.; Gassan, J. Composites reinforced with cellulose based fibers. Prog. Polym. Sci. **1999**, *24*, 221–274.
4. Dominick, V.; Rosato, P.E. *Plastics Processing Data Handbook*, 2nd Ed.; Chapman & Hall, 1997.
5. Kia, H. In *Sheet Molding Compound: Science and Technology*; Kia, H., Ed.; Hanser/Gardner Publications, Inc.: Cincinnati, 1993.
6. Kosaka, T.; Fukuda, T. Cure and Health Monitoring. In *Encyclopedia of Smart Materials*; Schwartz, M., Ed.; John Wiley-Interscience: New York, USA, 2002; Vols. 1 and 2, 291–318.
7. Pebly, H.E. In *Reuse and Disposal*; Peters, S.T., Ed.; Chapman & Hall, 1998; 883–904.
8. Anonymous. DaimlerChrysler fits coupes with SMC. Reinf. Plast. **2003**, *4*.

Capsule Pipeline

Henry Liu
Freight Pipeline Company, Columbia, Missouri, U.S.A.

INTRODUCTION

The use of various types of capsule pipelines for materials transport or conveying in industry is described. They include small pneumatic capsule pipeline (PCP) systems with nonwheeled capsules for dispatch of materials and documents over short distances, often within a building complex or between neighboring buildings; large PCPs with wheeled capsules for bulk materials transport for longer distances in lieu of using trucks or conveyor belt; short as well as long hydraulic capsule pipeline (HCP) systems for transporting grain, minerals, construction materials, and solid wastes; and product capsule pipeline (PRCP) systems for transporting coal, petroleum coke, and certain minerals. Also included is a brief discussion of the economic, safety, and environmental benefits of capsule pipelines.

This entry focuses on the application of capsule pipelines, including: 1) selection of the appropriate types of capsule pipeline for any given application; 2) understanding of the underlying theory and basic equations needed for design; 3) special design considerations; and (d) key operational properties of capsule pipelines for specific applications. It provides important information to engineers for planning and preliminary design of an appropriate capsule pipeline system for any given application. Reference materials are also provided to enable the reader to find more detailed information on the subject.

TYPES OF CAPSULE PIPELINE

Capsule pipeline is the transport of solids by using capsules (cargo-carrying vessels or vehicles) moving through pipelines. The capsules are generally cylindrical in shape, having a diameter slightly (about 10%) smaller than the pipe diameter, and a length two to five times the capsule diameter. They are suspended and/or propelled by the fluid moving through the pipe. The fluid can be any liquid or gas. When a gas is used as the fluid to propel capsules, the system is called PCP, whereas when a liquid is used as the fluid, the system is called HCP. Air is usually the most practical gas for use in PCP, whereas water is the most practical for HCP. Both PCP and HCP are discussed separately.

Pneumatic Capsule Pipeline

Because the density of the gas used in PCP is 100–1000 times smaller than that of the liquid used in HCP, both the buoyancy and the lift forces on capsules are much smaller for PCP than for HCP. They are insufficient to overcome the weight of the capsules and cannot suspend the capsules in the pipe. Thus, PCP capsules slide along the bottom of the pipe and generate large contact friction. This causes large energy loss and wear of capsules and pipe. To avoid such undesirable effects, PCP capsules, especially those that are large and/or carry heavy cargoes, have wheels. Fig. 1 shows two types of wheeled PCP capsules: those for use in circular pipes and those for use in rectangular or square pipes. Both types have been utilized successfully in Japan.[1] Note that while the square and rectangular PCPs use bottom wheels mounted on the capsule bottom frame, the round (circular) PCPs use gimbals-type wheel assemblies mounted on the centerline of both ends of each capsule—compare (A) with (B) in Fig. 1. The gimbals-type wheel assemblies in circular pipes allow the wheels to rotate freely around the pipe's inside circumference, thereby keeping capsule bodies stable in the pipe. Without the gimbals arrangement, the capsule body would be unstable and would rotate in the pipe, causing spillage of cargoes and other problems. An example of a round PCP used in Japan has a one-way length of 3.2 km using a 1 m-diameter pipe.[2] The pipeline transports 2 million tons of limestone per year from a mine to a cement plant. This pipeline was built in 1983, and it has operated very successfully ever since, having compiled an availability record exceeding 98%. A square-type PCP was also used successfully in Japan.[3] This was a temporary pipeline used in the construction of a long and large tunnel for bullet trains to cross a mountain. The capsules in this pipeline carried ready-mixed concrete and other construction materials into the tunnel, and on their return trip carried the excavated soil and rocks. As the tunnel construction progressed, the square-section

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120017571

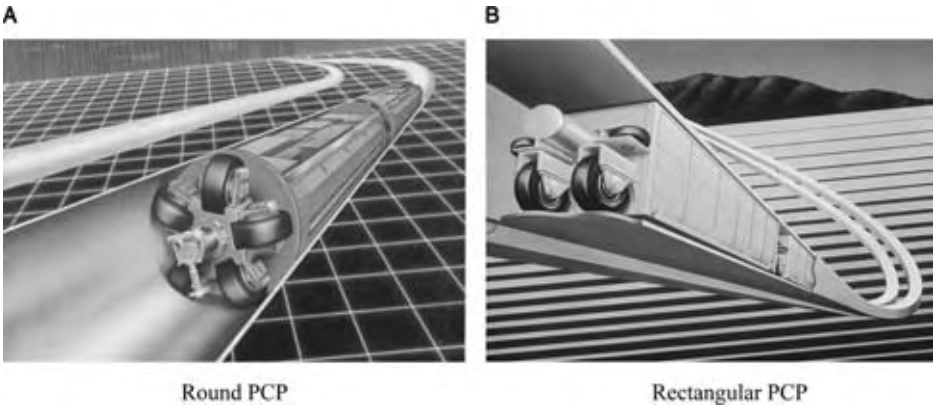


Fig. 1 Pneumatic capsule pipeline capsules in circular and rectangular conduits. (Courtesy of Sumitomo Metal Industries, Ltd.) *(View this art in color at www.dekker.com.)*

pipeline was extended into the tunnel by adding new sections to its front end. The pipe was assembled from panels made of lightweight concrete. The pipeline was dismantled and discarded (used for landfill) upon completion of the tunnel. In addition, the former Soviet Union also used 1 m-diameter round PCPs for transporting rocks.[4]

During the first half of the 20th century, non-wheeled PCPs of 300 mm or less were used extensively around the world for transporting light cargoes such as mail, telegrams, cash, and receipts.[5] Six major cities in the United States, New York, Washington, DC, Boston, Philadelphia, Chicago, and St. Louis, used them for transporting mail (including small parcels) between these cities' central post office and branch offices. By 1950, they were all replaced by automobiles.[6] Today, continued use of such small PCPs, called "tube transport," exists in the United States at large hospitals, airport terminals, drive-in banks, large factories, etc. Fig. 2 shows a system used at a large hospital. All current PCP systems, large or small, are highly automated and computerized. It is a far cry from the old tube transport systems of the past.

Fig. 2 Use of a computerized small PCP ("tube transport") system at the Johns Hopkins Hospital in Baltimore, MD, for transporting medical supplies. (Courtesy of PEVCO.)

Hydraulic Capsule Pipeline

The liquid used in HCP has a large density, which enables the development of a large-buoyancy force on the capsules in HCP, and a significant lift force even at relatively low velocity. For this reason, HCP capsules need no wheels, and they can be operated at relatively low velocity to conserve energy. The energy intensiveness (EI) of HCP, which is the energy consumed by HCP for transporting a unit weight of freight over a unit distance, is much smaller for HCP than for PCP. The conventional unit of EI used in the United States is Btu/TM (British thermal unit of energy consumed per ton of cargo transported per mile of distance). According to Liu, the values of EI for HCP and PCP are approximately 700 and 1800 Btu/TM, respectively, whereas for train and truck they are approximately 680 and 2300 Btu/TM, respectively.[7] This shows the advantage of HCP from the energy conservation standpoint, making the system attractive for long-distance transportation of bulk materials. Proper operation of HCP requires that the capsules be lifted off the pipe floor (by both buoyancy and lift), which minimizes not only the contact friction but also the wear and tear. Because of the lack of contact friction in HCP, there is no need for capsules to have wheels. This in turn reduces the cost of capsules and maintenance needed. How HCP capsules are suspended by the flow is an interesting topic to be discussed in "Theory and Equations."

The capsules in HCP may use either rigid containers or flexible containers. Fig. 3 shows a 7 in.-diameter steel capsule tested for grain transport in an 8 in. steel pipe at the University of Missouri–Columbia (UMC). With rigid containers, a second pipeline is needed to return the empty capsules, as is the case with PCP. If flexible containers made of plastic film are used to hold

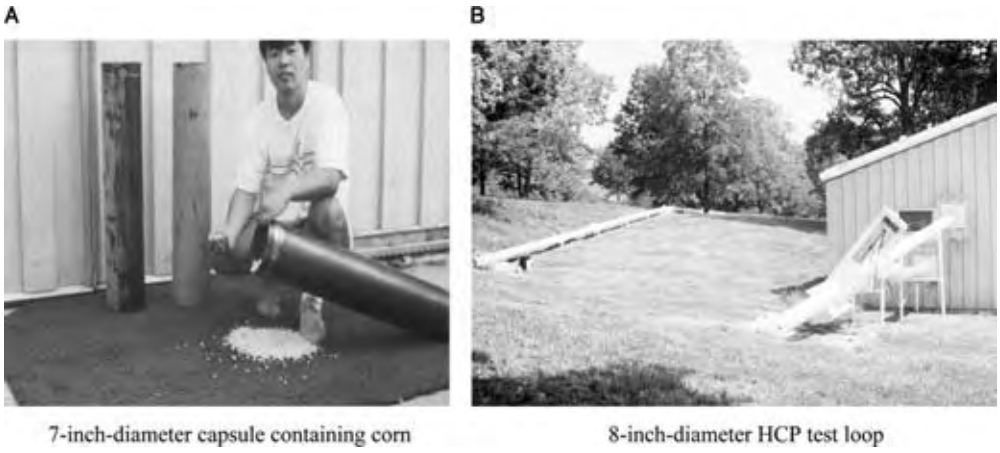


Fig. 3 Grain capsules tested at University of Missouri–Columbia. *(View this art in color at www.dekker.com.)*

the cargo, the plastic film can be recycled for other uses at the pipeline outlet, and the second pipeline is avoided, which reduces the overall cost, especially if the transportation distance is long. In the early 1980s, the W.R. Grace Company proposed to build such a pipeline from western Colorado to the coast of California for transporting powdered coal for export. However, the project was unable to attract financiers because of the lack of research to prove that the plastic film of the coal capsules can endure the long-distance transport without breakage.

A third type of HCP is PRCP, in which the capsule is made of the product (goods) to be transported, and no capsule shell or container is needed at all. This alleviates the need for both containers and the return pipeline. This type of HCP has been developed for transporting coal—the coal log pipeline (CLP). The CLP technology was studied extensively for 10 yr (1991–2000) at the Capsule Pipeline Research Center of the UMC. The Center was a State/Industry University Cooperative Research Center of the National Science Foundation. Both binder and binderless methods were developed to produce strong and wear-resistant coal logs for hydrotransport through pipe.[8,9] Fig. 4 shows the 5.4 in.-diameter coal logs produced by a hydraulic press tested in a 6 in.-diameter pipe test loop.[10] The 5.4 in. logs were also tested in a commercial pipeline in Kansas over a distance of 5 mi.[11] Coal logs of 1.9 in. diameter were tested extensively in a 2 in.-diameter pipe recirculating test loop at UMC, and the best logs tested were able to travel over 200 mi in the pipe with only 5% weight loss because of wear.[12] In addition to coal, petroleum coke (pet-coke) has been tested for compaction into logs for hydrotransport, and the result is promising. At about 50°C, pet-coke can be compacted without binder

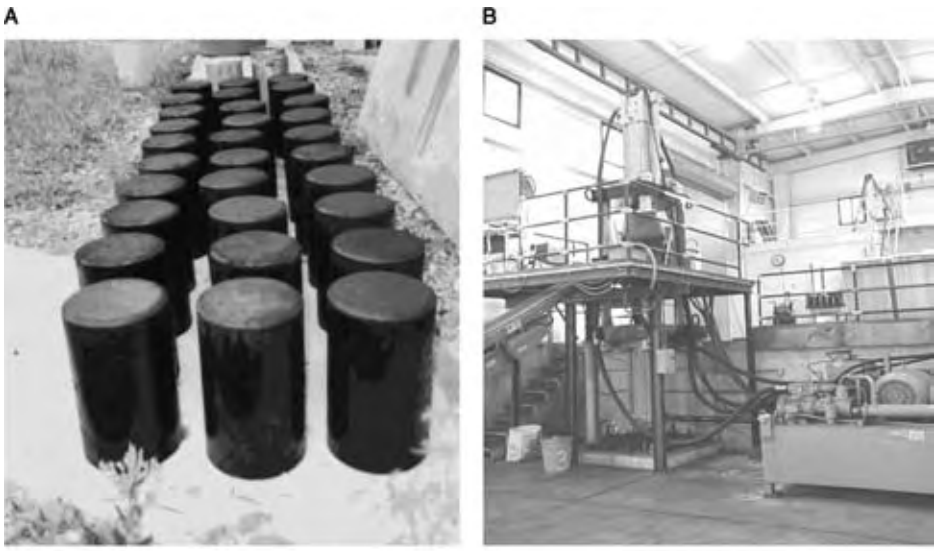


Fig. 4 Coal logs of 5.4 in. diameter and the press to produce such logs at University of Missouri–Columbia. *(View this art in color at www.dekker.com.)*

to produce strong and water-resistant logs for hydrotransport by pipe. At much higher temperatures, such as those experienced when the pet-coke is first produced before quenching with water, it can be cast into logs for hydrotransport. Another type of PRCP involves cans; the cans themselves are the capsules. It would be relatively simple to transport cans for short distances in a pipe of a diameter approximately 90% the can diameter. Whenever applicable, PRCP appears to be the most economically attractive type of HCP.

A fourth type of HCP is slurry capsule pipeline (SCP), which uses slurry instead of pure liquid (water) for transporting capsules through pipe. This is especially advantageous for long-distance transport of minerals. The mineral in the slurry of SCP may be either the same or different from that transported by the capsules in the pipe, depending on the situation and need. Using slurry instead of water for transporting capsules has three advantages: 1) more solids (minerals) can be transported by the pipe; 2) less water is used for the transport; and 3) the density of the slurry is significantly higher than that of water, and this increases the buoyancy and the lift on capsules, which in turn reduces contact friction between capsules and pipe, and hence reduces energy loss and wear. The last advantage is especially important when heavy minerals such as iron ore are to be transported by HCP. A dense slurry can be used in such a case, which greatly facilitates capsule transports. The concept of SCP has been proven in a laboratory study.[13]

Comparison Between PCP and HCP

A comparison of HCP with PCP is provided in Table 1. It is important to keep the different factors listed in Table 1 in mind when determining whether HCP or PCP is more suitable for a given application. For instance, for transporting cargoes that require speed, PCP is likely to be a better choice. On the other hand, when energy consumption is substantial and when cargo transport speed is immaterial, HCP is likely to be a better choice. All other key factors must be considered as well to make a wise selection between HCP and PCP. In addition, a comparison of the costs of the two systems for any given application should also be conducted using the life-cycle cost analysis to be discussed later.

Table 1 Comparison between HCP and PCP

Items for comparison	HCP	PCP
Fluid used	Liquid (water)	Gas (air)
Buoyancy and lift	Substantial	Negligible
Contact friction encountered	Low	High
Need for wheels	No	Yes
Capsule speed	Low (<3 m/sec)	High (>8 m/sec)
EI	Low (about 700 Btu/TM)	High (about 1800 Btu/TM)

ADVANTAGES OF CAPSULE PIPELINES

Use of capsule pipelines instead of trucks for freight transport reduces the number of trucks on highways, streets, and in and around industrial plants, thereby reducing traffic jams, accidents, air and noise pollution, and security problems caused by trucks. Also, trucks being mobile can be used by terrorists as bombs to attack selected targets. Capsule pipelines being fixed to the ground have much fewer security problems. From the security standpoint, it is especially important to use capsules pipelines instead of trucks to transport hazardous materials and explosives. When compared with conveyor belt systems, capsule pipelines have the advantage of being enclosed and are hence much safer to people, and less likely to pollute the air. Furthermore, because capsule pipeline is mostly underground, it is better from the land-use standpoint, and is unaffected by the weather.

THEORY AND EQUATIONS

Pneumatic Capsule Pipeline

The motion of capsules in a PCP is driven by the pressure drop across each capsule, Δp_c, generated by the air movement through the pipe. The faster the air flows through the pipe, the greater the pressure drop that develops across each capsule, and the faster the capsule moves. However, even at high air velocity, V, the capsule velocity V_c still trails the air velocity. The velocity difference, $V - V_c$, is related to the pressure drop Δp_c as follows:

$$\Delta p_c = C_D \frac{\rho(V - V_c)^2}{2} \quad (1)$$

where C_D is the drag coefficient of the capsule, and ρ is the air density at the capsule location in the pipe.

As shown in Fig. 5, most PCP capsules have two plates attached to the two flat ends of each capsule, called ''seal plates,'' or two rings around the capsule body near the two ends, called ''seal rings.'' Their purposes are the same: to seal the ends of the capsule so that little air passes through the seal and around the capsule. This generates a large drag coefficient

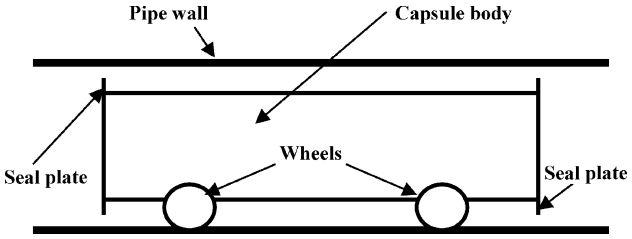


Fig. 5 Basic configuration of a capsule in a PCP of rectangular cross section.

C_D, which from Eq. (1) causes the slip velocity, $V_d = V - V_c$, to diminish for a given capsule pressure difference Δp_c, which means little slip between the capsule and the air. The value of C_D for a capsule with two seal plates or end rings can be calculated from the following equation derived by Kosugi:[14]

$$C_D = \frac{4k_s^4}{(1 - k_s^2)^2} = \frac{4k_A^2}{(1 - k_A)^2} \tag{2}$$

In the above equation, k_s is the seal plate diameter ratio, which is the seal plate diameter, D_s, divided by the inner diameter of the pipe, D. For PCP of rectangular or square cross-section, the seal plate area ratio k_A should be used, namely, $k_A = A_s/A = k_s^2$, where A_s is the seal plate area and A is the pipe cross-sectional area. For a capsule with only one seal pate, the factor 4 in Eq. (2) should be changed to 2. Note that in commercial PCPs, while the seal plate diameter D_s is usually greater than 97% of the pipe diameter D, the capsule body diameter is usually much smaller—about 90% of D. The smaller body diameter is needed to allow capsules to pass through PCPs having bends of reasonable curvature. With $k_s = 0.97$, Eq. (2) yields a drag coefficient of 1014. Most commercial PCPs have a drag coefficient greater than 1000.

From Eq. (1), the drag force F_D on the capsule is

$$F_D = A\Delta p_c = AC_D \frac{\rho(V - V_c)^2}{2} \tag{3}$$

For air moving at a constant speed V through a PCP having an upward slope of angle θ, the capsule velocity can be obtained from the Newton's second law, which yields

$$V_c = V - V_d$$
$$\text{where } V_d = \sqrt{\frac{2W_c(\eta \cos\theta + \sin\theta)}{C_D A\rho}} \tag{4}$$

In Eq. (4), V_d is the slip velocity, which is the difference between the air velocity V and the capsule velocity V_c; W_c is the capsule weight; and η is the contact friction coefficient between the capsule and the pipe (i.e., the rolling friction or wheel friction). When the pipe is horizontal, the slip velocity in Eq. (4) becomes $V_d = \sqrt{2\eta W_c/(C_D A\rho)}$.

Although Eq. (4) was derived for a PCP with an upward slope, it is also applicable to the downward slope if θ is taken to be negative, as long as the slope angle θ is less than $\tan^{-1}\eta$. When θ is greater than $\tan^{-1}\eta$ on a downslope, the capsule velocity now exceeds the air velocity V, and the capsule speed becomes

$$V_c = V + V_d$$
$$\text{where } V_d = \sqrt{\frac{2W_c(\sin\theta - \eta\cos\theta)}{C_D A\rho}} \tag{5}$$

The pressure drop along a PCP can be calculated by separately calculating the pressure drop across each capsule in the pipe, that across each capsule-free space between neighboring capsules, that between the blower outlet and the first capsule in the pipe, and that between the last capsule in the pipe and the pipe outlet. Summing all these pressure drops yields the total pressure drop along the PCP. The pressure drop across each capsule can be determined from Eq. (1) using the local values of ρ, V, and V_c. To calculate the pressure drop along the pipe in the capsule-free regions, equations in fluid mechanics can be used. The Darcy–Weisbach or Fanning equation can be used if the flow is assumed incompressible. The incompressible flow assumption is good if the total pressure drop along the entire length of the pipe is no more than a few pounds per square inch. For a higher total pressure drop along the pipe, the flow must be treated as compressible, and the following formula for compressible isothermal flow can be used:

$$\frac{fL}{D} = \frac{1}{\gamma M_1^2}\left[1 - \frac{p_2^2}{p_1^2}\right] + 2\ln\frac{p_2}{p_1} \tag{6}$$

In the above equation, L is the length over which the pressure drop $p_1 - p_2$ is to be calculated; p_1 is the absolute pressure of the flow at an upstream point 1, and p_2 is the absolute pressure of the flow at a downstream point 2; f is the Darcy–Weisbach friction factor that can be determined from the Moody Diagram; γ is the adiabatic exponent (equal to 1.4 for air); and M_1 is the Mach number of the flow at the upstream point 1. In addition to the pressure equations, the following equation of state of ideal gas is also needed:

$$\frac{p_2}{p_1} = \left(\frac{\rho_2}{\rho_1}\right)^{\gamma} \tag{7}$$

Using Eqs. (1), (2), (4) [or (5)], (6), and (7) for various reaches along the pipe, and using the boundary

condition $p = p_a$ (atmospheric pressure) at both ends of the pipe, and the pressure across the blower, p_b, the equations can be solved simultaneously to determine the variation of p, ρ, V, and V_c along the PCP. This in a nutshell explains how the pressure, air density, air velocity, and capsule velocity along a PCP can be calculated by assuming compressible flow. More details of such computations with illustrative examples are given in Liu and in York.[15,16]

Hydraulic Capsule Pipeline

Because of the fact that the capsules in HCP have no wheels and are strongly affected by the buoyancy and hydrodynamic lift, the fluid mechanics of HCP is far more complicated than that of PCP. However, extensive research has been conducted in HCP since 1960 by more than 30 researchers in about 10 nations to understand the basic nature of HCP flow and to derive the necessary equations for the design of HCP. In what follows, only the most pertinent theories and equations needed for the design of commercial HCPs are discussed. Readers interested in knowing more about the theory of HCP should read the vast body of papers and reports published; they total well over 100.

It is permissible to assume that all steady flows in HCP are incompressible. By using the one-dimensional continuity equation for incompressible flow, the following holds:

$$VA = V_c A_c + V_a(A - A_c) \quad \text{or} \quad V_a = \frac{V - k^2 V_c}{1 - k^2} \tag{8}$$

In Eq. (8), A, A_c, and $(A - A_c)$ are the cross-sectional areas of the pipe, the capsule, and the capsule–pipe annulus, respectively; V_a is the mean velocity of the fluid in the annulus; and k is the diameter ratio D_c/D. From Eq. (8), one can conclude the following:

- If $V > V_c$, then $V_a > V$ or $V_a > V > V_c$.
- If $V < V_c$, then $V_a < V$ or $V_c > V > V_a$.
- If $V = V_c$, then $V_a = V$ or $V_a = V = V_c$.

Perhaps the most important progress made in the last decade in understanding HCP flow is the four-regime theory. The theory was developed by Henry Liu and verified by experiments conducted by several of his graduate students at the Capsule Pipeline Research Center, UMC. The theory is briefly described as follows.

Consider a cylindrical capsule immersed in water in a pipe, with the capsule being denser than the water. Assume that initially the water is at rest. In this case, the capsule sinks to rest on the floor of the pipe. Later, if the water is allowed to flow slowly through the pipe at low velocity, there is insufficient drag generated on the capsule to cause it to move, and the capsule remains at rest in the pipe despite the fluid motion. This is Regime 1, in which the capsule velocity is 0 and $V > V_c = 0$. Fig. 6A depicts Regime 1. If the water velocity in the pipe is increased sufficiently, a point will be reached when the capsule will start to slide. The water velocity at this point is called the ''incipient velocity,'' designated by V_i. The capsule starts to slide at V_i because the drag on the capsule by the flow has exceeded the resistance caused by contact friction. If the water velocity V is increased to beyond V_i, the flow is in Regime 2, in which the capsule slides on the pipe floor, and the water velocity is higher than the capsule velocity ($V > V_c$). The situation is depicted in Fig. 6B. As the velocity of the fluid increases in Regime 2, more and more lift is generated on the capsule to reduce contact friction and to increase capsule speed. A point is reached where the capsule velocity

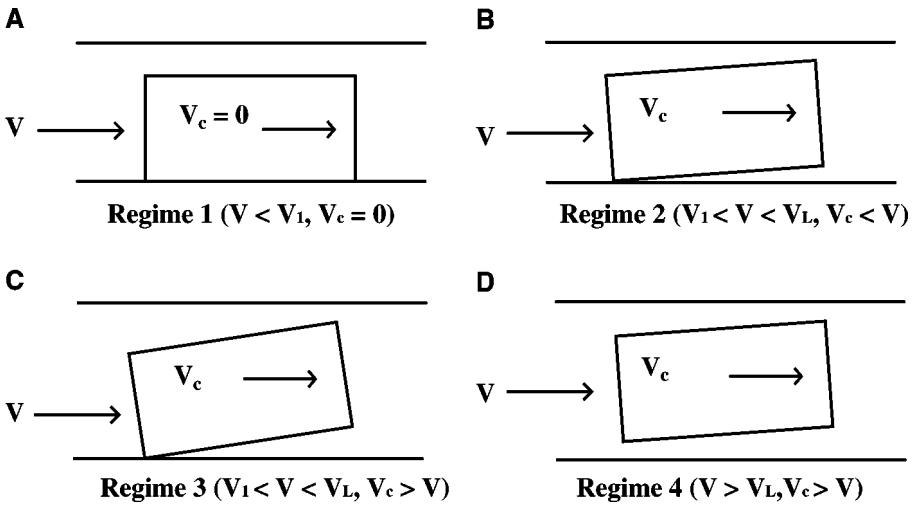


Fig. 6 The four regimes of capsule flow in pipe.

becomes equal to the fluid velocity. This velocity, marking the end of Regime 2 and the beginning of Regime 3, is called the critical velocity, V_0.

If the water velocity is increased to beyond the critical velocity, the drag (owing to pressure gradient across the two ends of the capsule) and the lift (because of pressure differential between the bottom and the top of the capsule) continue to increase, whereas the contact friction continues to decrease. This makes the capsule move faster than the fluid ($V_c > V > V_0$), and the flow is said to be in Regime 3 (see Fig. 6C). As the water velocity continues to increase in Regime 3, the drag and lift on the capsule further increase, and the contact friction continues to decrease. There comes a point when the combined upward force on the capsule because of buoyancy and lift equals the downward force on the capsule owing to capsule weight. When this happens, the capsule is completely lifted off the pipe floor, in a nose-up position as shown in Fig. 6D. The fluid velocity at the end of Regime 3 is called the ''lift-off velocity,'' designated as V_L. If the fluid velocity continues to increase beyond the lift-off, the capsule is completely suspended by the flow, and the flow is in Regime 4. In this regime, $V_c > V > V_L$, and the capsule velocity is about 15% greater than the fluid velocity (namely, $V_c \cong 1.15V$). This is not a desirable situation because a turbulent wake is generated behind the capsule, causing violent vibration of the capsule in the pipe, which in turn causes the capsule to impact the pipe wall. Such random impacts result in high energy loss and severe abrasion of the capsules and the pipe. The most desirable operational condition for HCP is when the fluid velocity is slightly (say 10%) less than the lift-off velocity (namely, $V \cong 0.9V_L$). Under such a condition, the capsules in the pipe start to lift off but are not yet completely lifted off. They move in a stable nose-up position with the front being lifted up and the tail touching the pipe gently. Both headloss and abrasion are minimized under such a condition.

To determine the flow regime of an HCP flow, the three characteristic velocities that mark the boundaries between regimes (i.e., V_i, V_0, and V_L) must be known. They can be calculated as described in Ref.[15]. Once these three characteristic velocities are found for any given HCP flow, the flow regime becomes known. With the flow regime known, one can then use the equations derived for this regime to calculate the capsule speed and pressure gradient along the pipe, as discussed in Gao.[17] More about the fluid mechanics of capsule pipeline together with illustrative examples is given in Ref.[15]. Note that the four-regime theory, discussed above conveniently in terms of water, is actually applicable to any fluid—both liquid and gas. Thus, the theory is applicable to both HCP and PCP involving any fluid. However, in the case of PCP, the buoyancy and lift caused by air (or another gas) are usually so small that they have little effect on capsule motion. Consequently, all horizontal PCPs operate only in Regimes 1 and 2, with capsules either resting on the pipe floor or rolling in the pipe at a velocity less than the fluid velocity. Only in rare cases does a PCP operate in Regime 3.

SYSTEM COMPONENTS AND OPERATION

System Description

All capsule pipeline systems include a pipe, many capsules, an injection system to inject capsules into the pipe, a pumping system to pump the fluid and capsules through the pipe, an ejection system to eject the capsules from the pipe, an instrumentation system to monitor the flow rate, pressure, temperature, and the passage of capsules at key locations, and an automatic control system that uses programmable logic controllers and an SCADA (Supervisory Control and Data Acquisition).

Capsule Design

Different types of capsule pipelines require different capsule designs. Information on designing capsules for PCP can be found in Refs.[1–3] and from the vendors of PCP; for HCP it can be found in Ref.[18]. Usually, several PCP capsules are linked together to form a train, and there are usually more than one train moving through the pipe at any time. In contrast, HCP capsules are not linked—they operate better without being linked because collision between capsules is not a problem for HCP.

Pipe Types

The types of pipe required differ with the types and applications of capsule pipelines. Generally, for PCP of circular cross section, steel pipes of sufficient rigidity (thickness) are used. Pressure concrete cylinder pipes may also be used if the pipe joints are smooth, and if the capsules use rubber wheels, which do not abrade the concrete. For PCPs of rectangular or square cross section, the pipe (rectangular conduit) may be made of steel or concrete. The square PCP used in Japan for constructing the long tunnel for bullet trains was made of precast concrete plates linked together to form the rectangular conduit.[7] Other arrangements are also possible. Note that the pressure in PCP is not high (normally not more than 1 or 2 atm), and the pipe

can tolerate a small amount of leakage without serious consequence.

Loading/Unloading at Inlet/Outlet

The PCP in Japan for transporting limestone uses a rotary loader, which automatically loads each capsule of a train successively.[6] Upon reaching the destination, each capsule dumps its load by opening the bottom door without having to stop. An alternate unloading system has also been developed, which rotates the capsules 180° (upside down) so that the cargo can be dumped through the same opening from which it enters the capsules. The foregoing system used in Japan for loading/unloading has proved rather reliable. It is an in-line system because the capsules are loaded and unloaded when they are moving in-line. Capsules never leave the line except for occasional cleanup and maintenance. The capsule trains in the pipe are recirculated through the dual pipe. A review of different types of injection and ejection systems for HCP is given in Liu.[19] The most practical type of capsule injector for HCP is the multilock launcher described in Ref.[20]. The system can achieve over 90% of linefill (% total length of pipe occupied by the capsules), which is essential for cost-effective use of HCP. Loading/unloading cargo for the system is done off-line.

Pumps/Blowers

All current commercial systems of PCPs use blowers to blow air through the pipe, and the air in turn drives the capsules through the pipe. Because capsules cannot pass through the blowers, a swinging pipe segment is used to cause each capsule train to bypass the blower. This greatly impedes capsule traffic, causing the PCP to have a very low linefill rate, usually less than 5%. It also makes the use of booster pumps difficult, thereby limiting the distance of PCP to only a few miles. Research has been done in recent years to solve this problem by using electromagnetic capsule pumps, which allow the capsules to pass through the pumps unimpeded.[21,22] The electromagnetic capsule pump can increase the linefill of PCP to 20%, and it makes the use of booster pumps a simple matter. For HCP, the multilock injection system causes capsules to enter the pipe without having to go through the pump. It can pump high (90%) linefill through the pipe until the capsules arrive at the first booster station. Then, a special pump-bypass is used for adding energy to the flow and causing the capsules to move to the next booster station and so forth. Information on the design of such booster stations can be found in Refs.[20,23]. Electromagnetic capsule pumps have also been developed for use as booster pumps in HCP.[24]

Operational Considerations

An important consideration in the operation of HCP is the weight of the capsules. Whenever possible, each capsule should be filled to a weight corresponding to a specific gravity between 1.00 and 1.05, which will optimize operation—minimize energy loss and abrasion (wear). While this may not be a problem for capsules loaded with cargo, it becomes a problem for empty capsules that must be returned via a return pipeline. Those returning capsules must be filled either with another cargo or water, so that they can have a specific gravity in the 1.00–1.05 range. With the returning capsules carrying water, there will be an excess of water at the pipeline intake reservoir, which is the same as the outlet reservoir of the return pipeline. This problem can be solved easily by running the cargo delivery pipeline at a velocity faster than the return pipeline, so that the water flow rate will be the same in the two pipelines. To keep the number of capsules delivered per second the same for the two pipelines of different velocities, the delivery line should have a linefill rate of capsules lower than that of the return line, which has little consequence other than a small reduction of capacity. The linefill rate can be controlled by the rate of capsule injection into the pipe at the inlet.

COST ANALYSIS

The cost-effectiveness of any capsule pipeline system can be determined by comparing the unit freight transportation cost (i.e.,"unit cost" in short), in dollars per ton of the cargo transported through the pipeline, to the unit cost or tariff (also in $/ton) for the same distance charged by competing modes of freight transport including truck, railroad (if there is a rail link), conveyor belts, etc. Rigorous determination of the unit cost of capsule pipeline requires the use of the life-cycle cost analysis. The life-cycle cost includes not only capital cost but also operation and maintenance costs, which in turn include energy, fuel, salaries, corporate income tax, insurance, etc. A reasonable annual profit (return on investment), such as 15%, should also be included so that the unit "cost" obtained for the pipeline can be readily compared to the tariffs of other modes. From the life-cycle cost and from the annual cargo throughput of the pipeline, one can calculate the unit cost as explained in detail in Ref.[25]. All future projects of capsule pipelines should conduct such an analysis before the project is approved or authorized.

CONCLUSIONS

Extensive R&D has been conducted on various types of capsule pipelines in the last 50 years to warrant their commercial use for transporting solids, at least for short and intermediate distances. In considering a capsule pipeline for a given project, the first matter to be determined is which type of capsule pipeline is best suited for the specific application. The information provided in this entry, together with other practical considerations, will lead to the selection of an appropriate type. Once the type of capsule pipeline is determined, a preliminary system can be designed using the information provided in this entry, including details provided in the listed references, and using standard knowledge of fluid mechanics and pipeline design. The design parameters obtained will include the pipe diameter, fluid speed, capsule speed, linefill, total pressure drop along the pipe, total power needed, distance between booster stations, pump head for each booster station, horsepower of each pump (or blower), etc. Using such a preliminary design, one can determine the cost of each individual item, and calculate the life-cycle cost and the unit cost. By comparing the profit-included unit cost of the selected capsule pipeline system with the tariffs charged by competing modes of freight transport, the designer can determine whether the capsule pipeline considered is the most cost-effective way to transport the solids for the given project. Other factors, such as convenience, environment, safety, and security, should also be compared with other modes to determine whether the capsule pipeline system considered should be the choice.

In passing, it may be of interest to point out that in 2004, the New York state government sponsored a study to determine the technical and economic feasibility of using large PCP systems for underground freight transport in New York City.[26] The study found that all of the six applications investigated are technically feasible, and five of them are also cost-effective.

REFERENCES

1. Kosugi, S. Pneumatic capsule pipelines in Japan and future development. Proceedings of the 1st International Symposium on Underground Freight Transport, Columbia, MO, Sep 2–3, 1999; Henry Liu, Ed.; University of Missouri–Columbia: Columbia, MO, 2000.
2. Kosugi, S. A capsule pipeline system for limestone transportation. Proceedings of the 7th International Symposium on Freight Pipelines, Wollongong, Australia, Jul 6–8, 1992; Institution of Engineers: Barton Act, Australia, 1992.
3. Kosugi, S.; Uchida, M. Pneumatic capsule pipelines for tunnel construction. Proceedings of the 8th International Symposium on Freight Pipelines, Pittsburgh, PA, 1995; Round, G., Ed.
4. Jvarsheishvili, A.G. Pneumatic capsule pipelines in USSR. J. Pipelines **1981**, *1* (1), 109–110.
5. Zandi, I.; Gimm, K.K. *Transportation of Solid Commodities via Freight Pipeline, Volume II, Freight Pipeline Technology*; Report DOT-TST-78T-36; U.S. Department of Transportation: Washington, DC, 1976.
6. Cohen, R.A. The pneumatic mail tubes: New York's hidden highway and its development. Proceedings of the 1st International Symposium on Underground Freight Transport, Columbia, MO, Sep 2–3, 1999; Henry, Liu, Ed.; University of Missouri–Columbia: Columbia, MO, 2000.
7. Liu, H. Introduction. In *Pipeline Engineering*; CRC Press: Boca Raton, FL, 2003.
8. Tao, B.N. Binder Concentration and Lubricant Effect on Coal Log Quality; M.S. Thesis, Department of Mechanical & Aeronautical Engineering. University of Missouri–Columbia: Columbia, MO, 1997.
9. Gunnink, B.W.; Kananur, J.; Chen, F. Binderless compaction of hot water formed coal logs for pipeline transport. Fuel Process. Technol. **1995**, *43* (3), 191–212.
10. Liu, H. Freight transport by underground pipelines: state-of-the-art assessment. Proceedings of the 3rd International Symposium on Underground Freight Transport, Bochum, Germany, Sep 19–20, 2002; Stein, D., Ed.; Ruhr University of Bochum: Bochum, Germany, 2002.
11. Liu, H.; Cheng, C.C. Coal log abrasion in pipelines. Proceedings of the 20th Technical Conference on Coal Utilization and Fuel Systems, Clearwater, FL, Mar 1995; Coal Technology Association: Washington, DC, 1995.
12. Cheng, C.C. Wear and Damage of Coal Logs in Pipeline; Ph.D. Dissertation, Department of Civil Engineering. University of Missouri–Columbia: Columbia, MO, 1994.
13. Seaba, J.P.; Xu, G. Capsule transport in slurry medium. J. Fluids Eng. **1995**, *117*, 691–695.
14. Kosugi, S. A Basic Study on the Design of the Pneumatic Capsule Pipeline System; Ph.D. Dissertation, Department of Mechanical Engineering. Osaka University: Osaka, Japan, 1985.
15. Liu, H. Capsule pipelines. In *Pipeline Engineering*; CRC Press: Boca Raton, FL, 2003.
16. York, K. Predicting the Performance of a PCP System Using a Linear Induction Motor for Capsule Propulsion; M.S. Thesis, Department of Civil Engineering. University of Missouri–Columbia: Columbia, MO, 1999.

17. Gao, X. Hydrodynamics of HCP with Slopes and Bends; Ph.D. Dissertation, Department of Civil Engineering. University of Missouri–Columbia: Columbia, MO, 1999.
18. Assadollabaik, M.; Liu, H.; Westrich, P. Design and test of hydraulic capsules to transport grain. In *Freight Pipelines*; Liu, H., Round, G.F., Eds.; Hemisphere Publishing Corporation: New York, 1990.
19. Liu, H. Hydraulic capsule pipeline. J. Pipelines **1981**, *1* (1), 11–23.
20. Liu, H. Design and operational considerations of hydraulic capsule pipelines. Proceedings of the Workshop on Capsule Pipelines, Tokyo, Japan, Jun 1992. Japanese Society of Multiphase Flow: Tokyo, Japan, 1992; 26–50.
21. Liu, H.; O'Connell, R.; Plodpradista, W.; York, K. Use of linear induction motors for pumping capsules in pneumatic capsule pipeline. Proceedings of the 1st International Symposium on Underground Freight Transport, Columbia, MO, 1999. University of Missouri–Columbia: Columbia, MO, 2000; 84–94.
22. Plodpradista, W. Study of Tubular Linear Induction Motor for Pneumatic Capsule Pipeline System; Ph.D. Dissertation, Department of Electrical Engineering. University of Missouri–Columbia: Columbia, MO, 2002.
23. Wu, J.P. Dynamic Modeling of an HCP System and Its Control; Ph. D. Dissertation, Department of Civil Engineering. University of Missouri–Columbia: Columbia, MO, 1994.
24. Assadollabaik, M.; Liu, H. Optimal design of electromagnetic capsule pump for capsule pipeline. J. Pipelines **1986**, *5*, 157–169.
25. Liu, H.; Noble, J.S.; Wu, J.P.; Zuniga, R. Economics of coal log pipeline for transporting coal. Transport. Res. (A) **1998**, *32* (4), 377–391.
26. Liu, H. *Feasibility of Underground Pneumatic Freight Transport in New York City*; Columbia, MO, 2002, Final Report for project sponsored by the New York State Energy Research and Development Authority (NYSERDA), contract no. 7643, 95 pp. (Report published on the following website: www.freightpipelinecompany.com).

Carbon Dioxide Capture and Disposal: Carbon Sequestration

C

K. S. Lackner
Department of Earth and Environmental Engineering, Columbia University, New York, New York, U.S.A.

A.-H. Park
L.-S. Fan
Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio, U.S.A.

INTRODUCTION

Fossil fuels provide a large, affordable source of energy that is limited by environmental impacts rather than resource constraints. A major concern in using fossil fuels is the emission of CO_2 to the atmosphere. CO_2 is a potent greenhouse gas and the dominant contributor to anthropogenic climate change. It is also an acid gas that changes the chemistry of the surface ocean, which is in equilibrium with the atmosphere. Since CO_2 is physiologically active, in plants as well as animals, a change in CO_2 concentrations is likely to have widespread ecological effects even without climate change. In order to stabilize the level of CO_2 in the air, emissions have to be reduced by a factor of three or more. In what follows, the options for the capture of CO_2 and its subsequent disposal are outlined. Together, these technologies are known as carbon sequestration.

BACKGROUND

The atmospheric concentration of CO_2 rose from 280 ppm in 1800 to 370 ppm in 2000, mainly due to the consumption of fossil fuels.[1] This increase in CO_2 concentration is expected to have various environmental and ecological effects. For example, doubling of the CO_2 concentration in the atmosphere reduces the rate of calcium carbonate deposition in coral reefs by 30–40%.[2] Most of this rise has occurred over the last few decades, and, unless action is taken, the projected growth over the 21st century could lead to a doubling or tripling of the preindustrial level of CO_2.[3] Unlike for SO_2 emission, the total accumulation of CO_2 matters rather than the rate of CO_2 emission. Oceanic uptake of CO_2 can compensate for some emissions,[4] but this uptake will collapse once CO_2 concentrations in the air are held constant. Oceanic uptake depends on a concentration gradient in the surface ocean maintained by the currently rising CO_2 concentration in air. At an increased CO_2 level in air, the ocean uptake rate will decrease and, with it, the world's annual emission allowance. In order to stabilize the atmospheric CO_2 level, emissions have to be reduced by several factors. As a result, the world's per capita emission would have to be a factor of 10–30 smaller than current per capita levels in industrial countries.[5] Stabilization of CO_2 concentrations in the atmosphere either requires abandoning fossil energy or capturing and storing most, if not all, CO_2 produced.

The term ''sequestration'' refers to ''putting aside and locking up'' carbon that is freed in the extraction of fossil fuels. Among some researchers, the term sequestration has become synonymous with biological sequestration, i.e., the cultivation of biomass to capture and store CO_2.[6] Here, the term is used more broadly: Sequestration refers to any storage scheme that can keep CO_2 out of the atmosphere. Carbon sequestration schemes can be ranked by capacity and permanence of storage, by environmental consequences, and by cost of implementation.

Carbon sequestration involves three steps—capture, transport, and disposal. The current estimated cost of capture and compression of CO_2 to pipeline pressure (110 bar) is approximately US$ 30–50 per ton of CO_2. Transport of CO_2 adds US$ 1–3 per ton per 100 km.[7] Disposal costs vary, but in some cases, they are quite low, e.g., US$ 4 per ton of CO_2 for some forms of underground injection.

Since CO_2 sequestration raises the cost of energy, it is unlikely that sequestration will be introduced without regulatory pressure. The trade-off between ignoring the risk of climate change and paying for a carbon-neutral energy infrastructure is easier if the cost of the carbon management can be held low.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007689

Sequestration will only be considered if it is cheaper than the alternatives for reducing CO_2 emissions. Alternative forms of energy, including wind and nuclear energy, limit the acceptable price of CO_2 sequestration.

The enabling technologies for sequestration fall into four categories:

- CO_2 capture from large industrial sources, such as power plants, refineries, cement plants, steel mills, or hydrogen production facilities.
- Capture of carbon from natural and short-lived carbon reservoirs, including the capture of CO_2 directly from the air.
- CO_2 transport from the point of capture to the point of disposal.
- CO_2 disposal in locations that allow for long-term storage of large quantities of CO_2.

CO_2 CAPTURE AT CONCENTRATED SOURCES

CO_2 sequestration requires economically competitive capture of CO_2 at industrial sources. This involves its separation from other gaseous constituents in plants' effluents and its preparation for transport and disposal. Fossil fuel burning power plants are among the largest sources of CO_2 emission worldwide (30%); other large sources include cement plants (6%), steel mills (6%), and hydrogen production facilities (1%). Capture strategies need to be tailored to each specific case. Most research has concentrated on power plants, for which a number of strategies have been developed. They differ in the location in the process where CO_2 is removed from the material flow stream.

Upstream Capture or Decarbonizing of Fuel

Carbon capture occurs upstream of the power plant while creating an intermediate carbon-free fuel from a carbonaceous energy resource such as coal. Capture is combined with fuel preparation and refining, and thus, the energy conversion itself does not involve emission of CO_2. In this case, the carbon capture becomes more akin to operations in a refinery or bulk chemical production plant.

Fuel decarbonizing applies to other sectors of the energy economy as well. Conversion from natural gas to electric power in commercial or domestic applications would move the carbon problem upstream and thus represent a strategy of decarbonizing of the consumption sector. Another example would be the transition to hydrogen in the transportation sector.[8] As long as coal and natural gas provide hydrogen at a fraction of the cost of the electricity that would enter into electrolysis, hydrogen is likely to be produced from fossil energy resources. Centralizing hydrogen production decarbonizes the transportation sector.[9]

Upstream capture also involves various gasification schemes that could be used to create a carbon-free chemical fuel from carbonaceous input. Typically, this fuel would be hydrogen, although other energetic compounds, such as ammonia,[10] metals (e.g., magnesium[11]), and metals hydrides (LiH[12]), have been considered in the past. Any carbon-free fuel could be consumed without concern over greenhouse gas emissions. Like any upstream technology change, the major cost of upstream CO_2 capture lies in the need to replace a large existing infrastructure.

As an example, consider methane steam reforming, water–gas shift, and partial oxidation producing hydrogen and CO_2, where the CO_2 is removed from the pressurized gas stream. Since hydrogen production is an established technology, this option is already available. Improvements would involve better gas separation techniques. Typical cleanup processes involve physical and chemical adsorption of CO_2. Glycol-based compounds (i.e., dimethylether of polyethylene glycol) and amine solutions are examples of the commercial solvents used for physical absorption that require a sufficiently high partial pressure of CO_2 ($>500\,kPa$). Other solvents include Selexol and Rectisol[13] and KS-2.[14] Other adsorption systems capture CO_2 on a bed of adsorbent materials such as molecular sieves or activated carbon.[15] CO_2 can also be separated from other gases by condensation at cryogenic temperatures.[16] Polymers, metals such as palladium, and molecular sieves are also being evaluated for membrane-based separation processes.[13] Although operating conditions require high pressures, most of these separation methods require low operating temperatures. Thus, the flue gas stream has to be cooled before the separation process, which leads to a rather large energy penalty. The details of these processes can be found in the following section.

On the other hand, hydrogen membranes and the reaction-based processes using solid sorbents can separate CO_2 while operating at the elevated temperatures of the gasifier. Such methods utilize the carbonation/calcination cycle of a solid metal oxide (MO).[17–19]

$$MO + CO_2 \leftrightarrow MCO_3 + \text{heat} \quad (1)$$

At low temperatures or high partial pressure of CO_2, the reaction proceeds to the right. At high temperatures or low partial pressure of CO_2, the reaction proceeds to the left. Gaseous CO_2 is bound into the solid matrix during the carbonation process, yielding metal carbonate. The metal oxide is then regenerated via either temperature or pressure swings. At temperatures beyond the calcination temperature that is specific for

the given partial pressure of CO_2, a pure CO_2 stream is produced as the metal oxide is regenerated. For the $CaO/CaCO_3$ system, the transition at 1 atm of CO_2 is at 890°C (calcination temperature), and, as the partial pressure of CO_2 increases, the carbonation/calcination temperatures of the system also rise. The advantages of the reaction-based separation process are: 1) CO_2 capture under A wide range of flue gas conditions (i.e., 150–700°C and <1 atm); 2) high equilibrium capacity; 3) low equilibrium concentration; and 4) generation of a pure CO_2 stream. $CaCO_3$, $MgCO_3$, $ZnCO_3$, $PbCO_3$, $CuCO_3$, and $MnCO_3$ are the potential candidates based on the range of carbonation and calcination temperatures, and reactivity.[17]

Downstream Flue Gas Scrubbing

The opposite approach to carbon capture is downstream flue gas scrubbing. The power plant remains unchanged until just prior to the release of flue gases. At this point, CO_2 sorbents or membranes are used to remove the CO_2 from the exhaust stream. Unlike the operating conditions in upstream designs, the temperature of the exhaust gas and the CO_2 partial pressure tend to be low for downstream capture. In pulverized coal combustion, CO_2 concentrations in flue gas are around 10–15%; in gas turbine exhausts, they are between 3% and 5%. Thus, in some cases, the flue gas has to be pressurized slightly to enhance the separation process. In addition to the chemical solvents listed earlier, alkanolamines [i.e., monoethanolamine (MEA) and diethanolamine (DEA)] can be used to separate CO_2. Particularly, the MEA process is currently considered to be the most attractive choice for separating CO_2 from flue gases, since this system can absorb CO_2 at low partial pressures.

Adsorption on solids is another well-known method of capturing gases from a gas stream. A column of adsorbent materials such as molecular sieves, activated carbon, and zeolite capture CO_2 on materials with high surface area. At high pressures, CO_2 is removed from the gas stream, and at low pressure, the adsorbed CO_2 is recovered. However, this method is usually low in capacity and often not very selective in the adsorption process. Cryogenic methods, while possible, are usually too expensive and are limited to purifying CO_2 for specialty applications.

Flue gas scrubbing is a very simple change in the design of a power plant, but since the CO_2 stream is large, the overall energy and cost penalty is large. Typical penalties are around 30–40% of the power output of the power plant. Thus, the cost per ton of CO_2 avoided includes the revenue loss due to the downrating of the power plant. Design improvements seem to have some potential to lower this cost.

Integrated Zero-Emission Power Plants

Not all power plant designs fit into an upstream or downstream category. Integrated systems let carbon move through the entire process, but they prevent normal dilution of the output flue gas, so that the effluent is concentrated CO_2. While most of the plants in this category are still in an early development phase, they promise to combine high efficiency, virtually zero atmospheric pollution, and complete capture of all CO_2. All avoid the intake of air.

If CO_2 is collected from the N_2-rich flue gas stream, the effluent from the power plant only contains condensable materials, making it possible to cap the flue stack and create a power plant with zero emissions to the air. Such plants integrate CO_2 capture with other pollutant control mechanisms and eliminate the expense of even more stringent flue gas scrubbing.

Oxyfuel combustion with flue gas recirculation represents a method of CO_2 capture that could be implemented today.[20] In this technique, input air is replaced by a mixture of recycled nitrogen-free flue gas (mainly CO_2) and pure oxygen. The recirculated CO_2 maintains an optimal oxygen concentration in intake gas, which, for a retrofit, would be near that of ambient air. The major cost and energy penalty in such a plant results from the operation of the oxygen plant, which consumes approximately one-third of the electricity output of the power station. At the current price of oxygen (US$ 4 per 100 m^3),[21] the oxygen would add US$ 0.018/kWh to the net electricity produced. However, once the oxygen demand and the economic scale of the oxygen production are increased, this cost is likely to be reduced. Noncryogenic approaches, such as high-temperature mixed solid oxide membranes, may be used to separate oxygen from air. The high-temperature units allow for much tighter integration into the power plant. For example, they could be integrated into the combustion unit of a gas turbine, which would then keep the compressed combustion products out of the turbine stream.[22]

Solid oxide fuel cells would integrate electricity production with air separation. The fuel need not be hydrogen; it could be any mixture of CO and H_2, or might even include methane, which undergoes steam reforming within the fuel cell chamber.[23] In the case of coal or biomass, the fuel gas is produced in an upstream gasification step. The fuel cell separates the oxygen from the air while generating electricity and heat. By transferring the heat to the upstream gasification process, it is possible to achieve extremely high rates of efficiency. The ZECA plant design aims at 70% conversion efficiency while delivering CO_2 in a concentrated stream.[24]

Another approach to bringing oxygen together with fuel is through a chemical looping process.

Chemical looping involves a sorbent, typically a metal or, more likely, a low-oxidation-state metal oxide that can be oxidized in air.[25] The oxide is reduced by fuel gases in a subsequent step. A variation of this approach oxidizes the metal not in air but in a chemical reaction with water, producing a pure stream of hydrogen. The advantage of this approach to hydrogen production is that it eliminates separation of hydrogen from the fuel gas, the use of a separate water–gas shift reactor, and the cleanup of fuel gas prior to hydrogen production.[26] The oxidized fuel gas is disposed of completely without any emissions to the atmosphere. From chemical looping, a high-pressure sequestration-ready CO_2 stream can also be produced.

Conventional gasification requires separation of hydrogen and CO_2 from the fuel gas mixture. This can be accomplished through chemical or physical sorbents or through membrane separation. As mentioned earlier, at low temperatures, physical and chemical adsorption is commonly used to separate hydrogen from the gas stream. Many of the zero-emission plants, however, could improve efficiency dramatically if CO_2 capture could be achieved at high temperatures. Solid chemical sorbents, such as CaO, have been considered potential candidates.

Membrane-based separation processes to capture either H_2 or CO_2 from the gasifier are new and less studied methods of CO_2 separation and capture. Membranes separate the desired gas component without requiring phase changes or chemical or physical sorption. The cost of membrane separation is generally dictated by the overall pressure drop. Membranes made of various types of materials such as polymers, metals, and rubber composites have been investigated.[10,27] Palladium and molecular sieves are currently under study.[10]

CAPTURE FROM DIFFUSE SOURCES

Capture of CO_2 at concentrated sources is easier than capture from the environment, but it requires significant changes in the existing infrastructure. Furthermore, approximately half of all emissions arise from small, distributed sources. Many of these emitters are vehicles, for which capture on board is not practical. For small, mobile emitters, the subsequent capture of carbon dioxide from the air provides another alternative.

The atmosphere represents a large and well-mixed buffer of carbon dioxide. Emissions over the course of a year are about 1% of the total CO_2 in the atmosphere. As a result, it is not necessary to capture a specific CO_2 emission; it suffices to capture an amount of CO_2 that cancels an emission that has happened or is about to happen. In general, CO_2 capture from air occurs at the disposal site and thus avoids the CO_2 transportation infrastructure necessary for capture at centralized sources.

Biomass sequestration, which has been embraced by coal companies as the least expensive way to start carbon management, relies on capture from air. Cofiring with biomass followed by sequestration also leads to a net CO_2 reduction in the atmosphere.[28] On a per-ton basis, this option is quite cost effective, but the land area that is required is too large to be practical. Consequently, it is worthwhile to consider other options. Currently, studies are underway that consider capture of CO_2 directly from the air by chemical means. These processes require good chemical sorbents that do not pose environmental concerns in their own right and that can be regenerated cost-effectively.

The scale of an air capture system is surprisingly small. At a wind speed of 6 m/s, the U.S. per capita emissions of 22 t/year flow through an opening of $0.2\,m^2$, the size of a television screen. Consequently, an air contacting apparatus, whether it just stands in the wind or utilizes wind flows through tower-like structures, can be very compact. The major challenge being addressed by current research is the design of a sorbent recycle system. Strong alkali solutions including NaOH and $Ca(OH)_2$ readily pull CO_2 out of the air, but the recovery is energy intensive. $Ca(OH)_2$ in contact with air turns into $CaCO_3$; NaOH turns into Na_2CO_3. The latter is reconverted to sodium hydroxide in a process used in the pulp and paper industry known as the Kraft process.[29] The effective equation is

$$Na_2CO_3 + Ca(OH)_2 \rightarrow 2NaOH + CaCO_3 \qquad (2)$$

In the end, both approaches require the calcination of limestone to lime. The energy penalty in this process is about 4.5–5 GJ per ton of CO_2, which amounts to a 30–40% energy penalty for a transportation sector that uses air extraction for managing its own CO_2 emissions. This is comparable to the energy penalty incurred in the conversion of fossil fuels into hydrogen as a transportation fuel.

TRANSPORT OF CO_2

A CO_2 transportation infrastructure would be a rather conventional approach. On land, pipelines for long-distance CO_2 transport already exist. For example, a pipeline system more than 800 km long connects CO_2 fields in Southern Colorado to oilfields in West Texas. The CO_2 is purchased at about US$ 15/t for tertiary oil recovery.

Ocean transport might involve pipelines or ships. Small CO_2 tankers already exist.[30] However, there is currently no use of larger vessels. Nevertheless, the

technology for large vessels seems less daunting than that for liquefied natural gas tankers, which is a well-established technology.

DISPOSAL OF CO_2

Disposal or long-term storage of CO_2 is the final step in the carbon management chain. It is made difficult by the large scale of storage. At current rates, the total CO_2 production for a century would be 650 Gt of carbon (GtC), or 2500 Gt of CO_2. For comparison, Lake Michigan contains 5000 Gt of water.

While the sequestration technology is still too new to settle on one specific approach, any storage option, to be useful beyond small niche markets, must have large capacity, be safe, be environmentally benign, and guarantee the long-term stability of storage. Storage lifetime constraints will become more stringent over time. The total leakage of all storage must be less than environmentally tolerable emissions. Leakage emissions, $L(t)$, are a function of time, t. They are related to total stored carbon, $S(t)$, and the representative lifetime of all storage, τ, through

$$L(t) = S(t)/\tau \tag{3}$$

Leakage rates start small, because S is initially small. As S increases, lifetime constraints grow more stringent. Until there is a comprehensive understanding of the global carbon cycle, the maximum allowable leakage rate will remain uncertain. It is, however, unlikely to surpass 1 GtC/yr. This suggests that τ will have to be in excess of 1000 years even before $S(t)$ reaches 1000 GtC. Unless fossil fuels are phased out, such amounts will be stored long before the end of the 21st century. This is a serious constraint: Biomass turnover is measured in years to decades, and ocean turnover times are measured in centuries.

In the following section, the various options in order of their likely capacity are discussed.

Biomass Sequestration

One of the largest natural carbon flows through the environment is driven by photosynthesis. Plants take up CO_2 and water and turn them into reduced carbon compounds such as starch and cellulose. Photosynthesis fixes about 100 GtC per year. Most of this is returned via respiration and decomposition, but it seems that over the last decade, mid-latitude forests have been net sinks for CO_2.[31] Trees and other plants sequester CO_2 during periods of growth. Thus, forestation and agricultural fixation of carbon, in either biomass or soil carbon, can play a role in carbon sequestration. A mid-latitude forest stores about 60 tC/ha aboveground and another 100 tC/ha belowground. Uptake rates can be as much as 3–10 tC/yr, but the capacity is limited. Worldwide biomass amounts to 600 GtC and soil carbon to 1400 GtC.[32] This defines the yardstick against which biological sequestration has to be measured. To match the CO_2 emission from a single GW coal plant, the forested area should be increased at a rate of 1 ha (2.5 acres) per hour. Keeping up with a GW plant requires a growing forest of 370,000 ha (950,000 acres). Biomass sequestration could make up for past deforestation, which is responsible for a small fraction of the increased atmospheric CO_2. Biomass sequestration is most useful when it is the by-product of other societal goals. Environmental concerns arise where monocultures for the purpose of biomass fixation replace less carbon-efficient natural systems.

Another biomass approach is ocean fertilization. This method spurs biomass growth in areas that are low in productivity due to lack of critical nutrients. It, too, is limited in capacity, and has raised environmental concerns over changes in the natural food chain. The cost of biomass conversion is often quite low. Paustian et al.[33] estimate that planting forests in developing countries could cost as low as US$ 3–10 per ton of CO_2 captured.

Ocean Sequestration

A large fraction of the emitted CO_2 is naturally absorbed by the ocean. Therefore, nonaction equates to partial ocean sequestration. In equilibrium, the partitioning of CO_2 between ocean and atmosphere is roughly 4 : 1.[34] As the partial pressure of CO_2 in the air rises, the ratio gradually moves toward a higher fraction of CO_2 remaining in the atmosphere. The CO_2 carrying capacity of the ocean is determined by its carbonate alkalinity, which is the charge-weighted sum of ion concentrations, excluding protons, hydroxyl ions, and bicarbonate and bicarbonate ions. Charge neutrality demands that the alkalinity A is equal to the sum of the ionic charge in these remaining ions:

$$\mathrm{A} = [\mathrm{OH}^-] + [\mathrm{HCO_3}^-] + 2[\mathrm{CO_3}^{2-}] - [\mathrm{H}^+] \tag{4}$$

For ocean water with a pH near 8, one can neglect protons and hydroxyl ions; thus,

$$\mathrm{A} \approx [\mathrm{HCO_3}^-] + 2[\mathrm{CO_3}^{2-}] \tag{5}$$

At constant alkalinity, the ocean can absorb more CO_2 by replacing one CO_3^{2-} with two HCO_3^-. Since 90% of all the inorganic CO_2 in the ocean is in the form of HCO_3^-, the uptake capacity of the ocean is far less

than the 39,000 GtC present in the ocean. Eliminating all carbonate ions would reduce the pH to about 4 and store no more than 3500 GtC. In equilibrium with 550 ppm of CO_2 in the air, the pH of the ocean would drop by 0.3, and if the CO_2 were stirred into the entire ocean, it could pick up about 1000–1500 GtC. Without active interference, the approach to this equilibrium would take millennia.[35]

In spite of these limitations, estimates for ocean storage capacity in excess of 100,000 billion GtC have been cited.[36] Such numbers would only be feasible if alkalinity (i.e., NaOH) is added to the ocean to neutralize the corresponding amounts of the carbonic acid. Over thousands of years, the dissolution of calcareous oozes at the bottom of the ocean could provide such alkalinity.[4,37]

Ocean disposal options fall into several categories, as illustrated in Fig. 1. One is to dilute the dissolved CO_2 at a depth below the mixed layer. Carbon dioxide can be stored using this method for decades to centuries, but capacities are limited. Another approach is to form lakes of CO_2 at the bottom of the ocean. Below 2700 m, compressed CO_2 is denser than seawater and sinks to the bottom. In addition, CO_2 reacts with seawater to form a clathrate, a cage structure with approximately six water molecules per CO_2.[38] Clathrates are solids that can form at temperatures slightly higher than the melting point of water and thus form spontaneously in the presence of liquid CO_2 near the bottom of the ocean. However, they are not stable, and as a result, they dissolve into ocean water once the CO_2 concentration in the water drops back to the normal levels. Since 1977, when Marchetti[39] first introduced the concept of direct ocean sequestration of CO_2, significant research effort has taken place, and now the cost for the direct disposal of CO_2 in the oceans is estimated to range from as low as US$ 1–6 per ton of CO_2,[40] to a more realistic estimate of US$ 5–15 per ton of CO_2.[41]

The environmental impact may be the most significant factor determining the acceptability of ocean storage, since the strategy is predicated on the notion that the impacts on the ocean will be less than the avoided impact of these emissions on the atmosphere. Environmental concerns in ocean sequestration focus on long-term chronic issues of changing the ocean chemistry as well as on the local effects of low pH (as low as 4) and its effect on marine organisms (e.g., stunting coral growth). While it is true that the oceans will get acidified as well, if the problem is ignored, the need for dealing with elevated levels of CO_2 in the atmosphere may not be limited to climate change concerns but also to changes in the ocean water chemistry. It does not help to replace one environmental problem with another. In addition, there is the possibility of sudden release of the sequestered CO_2 due to thermal plumes and volcanic action in the ocean. It has been reported that when a lake in Africa erupted, an enormous amount of CO_2 was discharged from the bottom of the lake, killing thousands of animals and humans by asphyxiation.[42]

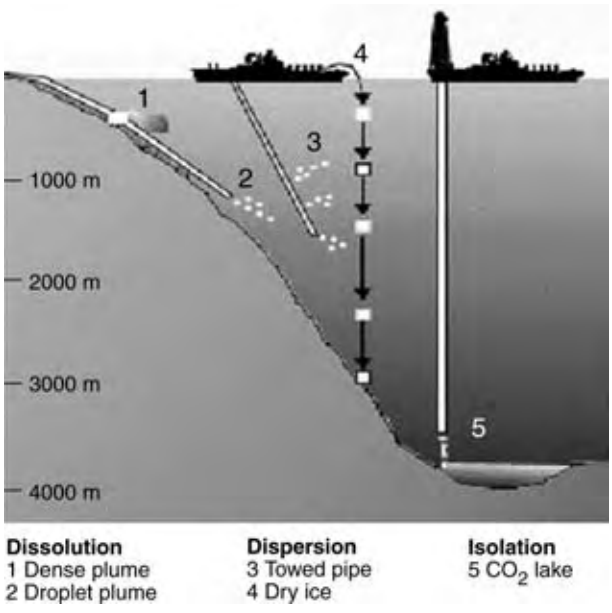


Fig. 1 Ocean disposal of CO_2: five injection scenarios. (From Ref.[34].) *(View this art in color at www.dekker.com.)*

Geological Sequestration

Underground storage in geological formations is a major option for disposing of CO_2 and is already practiced. Potential sites for geological storage are active oil fields, coal beds, depleted oil and gas reservoirs, deep saline aquifers, and mined salt domes and rock caverns.[43] In West Texas, approximately 20 million tons of CO_2 are consumed in tertiary oil recovery.[44] These injections do not qualify as sequestration as the CO_2 has been extracted from underground wells about 500 miles away. However, these projects and similar projects in Canada demonstrate that the disposal of CO_2 is practically feasible. Nevertheless, the capacity of tertiary oil and gas recovery is quite limited. Estimates range from 20 to 60 GtC.[45] Displacing methane tied up in deep, unmineable coal adds another small carbon sink to the portfolio of options. This method is attractive in the sense that most of the injected CO_2 will be immobilized by either physical or chemical absorption on the coal surface. On the other hand, in other reservoirs such as aquifers, the injected CO_2 will likely exist as a supercritical phase without being fully dissolved for thousands of years,

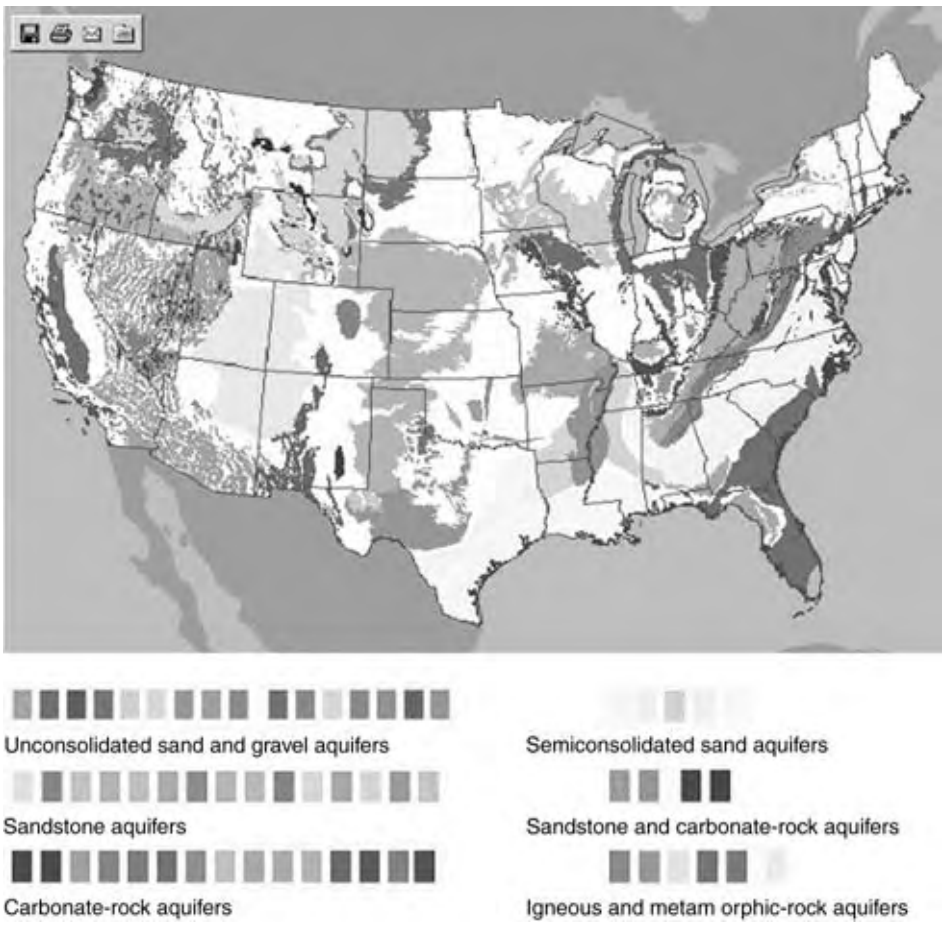


Fig. 2 Saline aquifers in the United States based on the U.S. Geological Survey. (From Ref.[45].) *(View this art in color at www.dekker.com.)*

and that creates a potential problem of a higher leakage rate.[46]

Beyond those examples where CO_2 can be utilized in the recovery of large underground reservoirs of oil and gas, deep aquifers represent the best long-term underground geological storage option. Such aquifers are generally saline and are separated from shallower aquifers and surface water supplies. The estimated storage capacities of these aquifers in the United States are 5–500 billion tons of CO_2.[47] This technology has been demonstrated by Statoil at the Sleipner project in the North Sea. It shows that injection into deep aquifers is economically feasible.[48] This project is ultimately driven by a tax of US$ 50 per ton of CO_2. Fig. 2 shows the location of deep saline aquifers in the United States.

There are a number of additional projects underway or being brainstormed that will further outline the viability and long-term capacity of geological sequestration. The central question is long-term safety, long-term leakage rate, and, ultimately, the capacity of these projects. It is likely that there will a large number of sites where CO_2 can be stored safely, for all practical purposes permanently. While some proponents of the scheme have suggested storage capacities of the order of 300 GtC, the real number at this point is still very uncertain and may prove to be much larger. For the application of geological sequestration, both near-surface and at-depth monitoring systems should also be developed to minimize CO_2 leakage from underground.

Mineral Sequestration

As the amount of CO_2 stored increases, it becomes progressively more difficult to guarantee a physical barrier that prevents CO_2 from returning to the atmosphere. Chemical conversion to a thermodynamically lower state would thus be desirable and is indeed possible. CO_2 is the anhydrous form of carbonic acid and therefore can be used to displace weaker acids such as silicic acid. The formation of carbonates from silicates is well known as geological weathering. Thermodynamically, CO_2 can be bound as a carbonate. In many instances, these carbonates dissolve in water, but some, such as magnesium and calcium carbonates, are remarkably stable as solids. Thus, mineral sequestration would provide a means of storing CO_2.

Fig. 3 Worldwide distribution of magnesium-rich ultramafic rocks. (From Ref.[49].) *(View this art in color at www.dekker.com.)*

Mineral carbonation is a new and, consequently, less studied method of sequestration. Mineral resources are plentiful for storing all the carbon that could ever be released in the consumption of fossil fuels. This sequestration process offers a safe and permanent method of CO_2 disposal, since there is almost no possibility of accidental release of CO_2 from the disposal site, as CO_2 is chemically incorporated into the mineral and immobilized. Furthermore, the reactions that bind CO_2 to the mineral are exothermic in nature, leading to the formation of thermodynamically stable carbonate forms.[49]

Ultramafic igneous rocks (primarily peridotites and serpentinites) contain large amounts of MgO bound into a silicate structure. Dunite is a rock made up entirely of olivine (Mg_2SiO_4). A more abundant ultramafic rock is serpentine [$Mg_3Si_2O_5(OH)_4$], which is hydrolyzed olivine. Both olivine and serpentine also contain some iron in place of magnesium. Peridotite and serpentinite rocks containing olivine and serpentine exist in amounts far exceeding fossil carbon resources, and Fig. 3 shows the map of the worldwide locations of those deposits. There are large deposits in the United States and Puerto Rico.[50] There are also large deposits of these minerals in Canada and parts of Europe.

Mineral sequestration can proceed in industrial processes or underground. Industrial processes could be based on gas–solid reactions, aqueous processes, or combinations of the two. While the cost of mining, mineral preparation, and tailing disposal is well known in industry and relatively inexpensive, the chemical reaction process is still too expensive for practical implementation.[51] The bottleneck is the slow kinetics of the process and the high cost of accelerating the process by various means. Heat treatment of serpentine, external attrition grinding, and acid dissolution have all been shown to work, but the cost of the process will still have to be reduced from current estimates of US$ 69 per ton of CO_2[52] to more acceptable numbers like US$ 30/t.[49] Since carbonic acid has proven to be too weak to dissolve serpentine or olivine with sufficient speed, some of the more promising approaches are based on the use of mixtures of weak acids that enhance the dissolution of serpentine and olivine while still allowing for their recovery at minimal energy cost.[53,54]

In the second process, CO_2 could be injected into selected underground mineral deposits for carbonation (in situ carbonation). This process envisions pumping CO_2 directly into an underground deposit of porous magnesium- or calcium-bearing rock. In contrast to permanent underground storage of CO_2 gas (as in aquifers and depleted oil/gas reservoirs), this process would result in chemically stable carbonates; it therefore poses substantially lower long-term risk. The reaction would be aided by naturally high pressures (overburden) and could proceed more rapidly than mineral weathering on the Earth's surface.

CONCLUSIONS

While, in the short term, the costs of CO_2 capture and transport are likely to dominate the cost of carbon sequestration, in the end, the cost of disposal will become more important. There are two major reasons. First, as technologies improve and power plants are designed around the concept of carbon capture, the incremental cost of capturing carbon dioxide will become smaller. Indeed, in some of the zero-emission designs, the capture processes separate CO_2 while capturing other pollutants, and thus the net costs of carbon capture may be close to zero.

Pipeline costs will get smaller as the pipeline network gets larger. At the same time, however, the demands on carbon dioxide storage will gradually increase. As the reservoirs get larger, concerns over environmental consequences will increase and the concerns over leakage will become greater. Meanwhile, the easily accessible sites will be used up, and the cost of monitoring will increase.

CO_2 sequestration technology is a viable contender among a set of options to stabilize the atmospheric CO_2 level over the next few decades.[55] The technical ability to start such a process exists today, but the institutional structures required to reduce CO_2 emissions are still missing. Carbon sequestration at the very least will buy time for alternatives to fossil fuels to become competitive. It is, however, equally possible that carbon sequestration may remove the major environmental obstacles to the use of fossil fuels, in which case they may prove competitive with alternatives for a long time to come.

REFERENCES

1. Fankhauser, S.; Leemans, R.; Erda, L.; Ogallo, L.; Pittock, B.; Richels, R.; Rosenzweig, C.; Safriel, U.; Tol, R.S.J.; Weyant, J.; Yohe, G. Vulnerability to climate change and reasons for concern: a synthesis. In *Climate Change 2001: Impacts, Adaptation, and Vulnerability—A Report of Working Group II of the Intergovernmental Panel on Climate Change*; McCarthy, J.J., Canziani, O.F., Leary, N.A., Dokken, D.J., White, K.S., Eds.; Cambridge University Press: Cambridge, 2001; 941–967.
2. Langdon, C.; Takahashi, T.; Sweeney, C.; Chipman, D.; Goddard, J.; Marubini, F.; Aceves, H.; Barnett, H.; Atkinson, M.J. Effect of calcium carbonate saturation state on the calcification rate of an experimental coral reef. Global Biochem. Cycles **2000**, *14* (2), 639–654.
3. Lackner, K. Climate change: a guide to CO_2 sequestration. Science **2003**, *300* (5626), 1677–1678.
4. Archer, D.; Kheshgi, H.; Maier-Reimer, E. Multiple timescales for neutralization of fossil fuel CO_2. Geophys. Res. Lett. **1997**, *24* (4), 405–408.
5. Schimel, D.; Enting, I.; Heimann, M.; Wigley, T.; Raynaud, D.; Alves, D.; Siegenthaler, U. CO_2 and the carbon cycle. In *IPCC Special Report on Climate Change 1994: Radiative Forcing of Climate Change and An Evaluation of the IPCC IS92 Emission Scenarios*; Houghton, J.T., Meira-Filho, L.G., Bruce, J., Lee, H., Callander, B.A., Haites, E., Harris, N., Maskell, K., Eds.; Cambridge University Press: Cambridge, 1995; 35–71.
6. Harvey, L.D.D. Declining temporal effectiveness of carbon sequestration: implications for compliance with the United National Framework Convention on Climate Change. Climate Change **2004**, *63* (3), 259–290.
7. Wallace, D. Capture and storage of CO_2—what needs to be done. The 6th Conference of the Parties, COP 6, to the United Nations Framework Convention on Climate Change, The Hague, The Netherlands, Nov. 13–24, 2000. www.iea.org/envissu/index.htm, 2000.
8. Ogden, J.M. Prospects for building a hydrogen energy infrastructure. Annu. Rev. Energy Environ. **1999**, *24* (1), 227–279.
9. Johansson, T.B.; Williams, R.H.; Ishitani, H.; Edmonds, J.A. Options for reducing CO_2 emissions from the energy supply sector. Energy Policy **1996**, *24* (10–11), 985–1003.
10. McFarlan, A.; Pelletier, L.; Maffei, N. An intermediate-temperature ammonia fuel cell using Gd-doped barium cerate electrolyte. J. Electrochem. Soc. **2004**, *151* (6), A930–A932.
11. Hasvold, O.; Lian, T.; Haakaas, E.; Storkersen, N.; Perelman, O.; Cordier, S. CLIPPER: a long-range, autonomous underwater vehicle using magnesium fuel and oxygen from the sea. J. Power Sources. **2004**, *136* (2), 232–239.
12. Kong, V.C.Y.; Kirk, D.W.; Foulkes, F.R.; Hinatsu, J.T. Development of hydrogen storage for fuel cell generators II: utilization of calcium hydride and lithium hydride. Int. J. Hydrogen Energy **2003**, *28* (2), 205–214.
13. Reimer, P.; Audus, H.; Smith, A. Carbon dioxide capture from power station. Proceeding of IEA Greenhouse R&D Programme, Cheltenham, Gloucester, U.K., 2001; www.ieagreen.org.uk.
14. Mimura, T.; Simayoshi, H.; Suda, T.; Iijima, M.; Mitsuake, S. Development of energy saving technology for flue gas carbon dioxide recovery by chemical absorption method and steam system in power plant. Energy Convers. Manage. **1997**, *38* (Suppl.), S57–S62.
15. Hoffert, M.I.; Wey, Y.C.; Callegari, A.S.; Broeker, W.S. Atmospheric response to deep sea injections of fossil fuel CO_2. Climate Change **1979**, *2*, 53–68.
16. Halman, M.M.; Steinberg, M. *Greenhouse Gas Carbon Dioxide Mitigation: Science and Technology*; Lewis Publishers: Boca Raton, FL, 1999.
17. Iyer, M.; Gupta, H.; Sakadjian, B.B.; Fan, L.-S. Multicyclic study on the simultaneous carbonation

and sulfation of high-reactivity CaO. Ind. & Eng. Chem. Res. **2004**, *43*, 3939–3947.

18. Balasubramanian, B.; Lopez Ortiz, A.; Kaytakoglu, S.; Harrison, D.P. Hydrogen from methane in a single-step process. Chem. Eng. Sci. **1999**, *54*, 3543–3552.
19. Fink, C.; Curran, G.; Sudbury, J.D. CO_2 acceptor process pilot plant—1974, Rapid City, South Dakota. Proceedings of the Sixth Synthetic Pipeline Gas Symposium, Chicago, IL, Oct. 28–30, 1974.
20. Andersson, K.; Johnsson, R.; Stroemberg, L. Large scale CO_2 capture—applying the concept of O_2/CO_2 combustion to commercial process data. VGB Powertech. **2003**, *10*, 1–5.
21. Kirschner, M.J. Oxygen. In *Ullman's Encyclopedia of Industrial Chemistry*, 7th Ed.; John Wiley & Sons, Inc.: New York; 2003.
22. Griffin, T.; Sundkvist, S.G.; Asen, K.; Bruun, T. Advanced zero emissions gas turbine power plant. Proceedings of ASME Turbo Expo: Power for Land, Sea & Air, Atlanta, GA Jun. 16–19, 2003; 836–842.
23. George, R.A. Status of tubular SOFC field unit demonstrations. J. Power Sources **2000**, *86* (1–2), 134–139.
24. Lackner, K.S.; Ziock, H.-J. The US zero emission coal alliance. VGB Powertech. **2001**, *12*, 57–61.
25. Leithner, R. Energy conversion processes with intrinsic CO_2 separation. Proceedings of SME, Denver, CO, Feb. 2004.
26. Mattisson, T.; Johansson, M.; Lyngfelt, A. Multicycle reduction and oxidation of different types of iron oxide particles—application to chemical-looping combustion. Energy Fuels **2004**, *18* (3), 628–637.
27. Bugante, E.C.; Shimomura, Y.; Tanaka, T.; Taniguchi, M.; Oi, S. Methane production from hydrogen and carbon dioxide and monoxide in a column bioreactor of thermophilic methanogens by gas recirculation. J. Ferment. Bioeng. **1989**, *67* (6), 419–421.
28. Keith, D.W. Sinks, energy crops and land use: coherent climate policy demands an integrated analysis of biomass. Climatic Change **2001**, *49* (1–2), 1–10.
29. Bujanovic, B.; Cameron, J.H.; Yilgor, N. Comparative studies of Kraft and Kraft–borate pulping of black spruce. J. Pulp Paper Sci. **2003**, *29* (6), 190–196.
30. Veder, A. Coral Carbonic. http://www.anthonyveder.nl/fleet/ccarbonic.html (accessed June 2004).
31. Houghton, R.A. Why are estimates of the terrestrial carbon balance so different? Global Change Biol. **2003**, *9* (4), 500–509.
32. Griffin, K.L.; Seemann, J.R. Plants, CO_2 and photosynthesis in the 21st century. Chem. Biol. **1996**, *3* (4), 245–254.
33. Paustian, K.; Cole, C.V.; Sauerbeck, D.; Sampson, N. CO_2 mitigation by agriculture: an overview. Climate Change **1998**, *40* (1), 135–162.
34. Butler, J.N. *Carbon Dioxide Equilibria and Their Applications*; Lewis Publisher: Chelsea, MI, 1991.
35. Lackner, K.S. Carbonate chemistry for sequestering fossil carbon. Annu. Rev. Energy Environ. **2002**, *27* (1), 193–232.
36. Herzog, H.; Drake, E.; Adams, E. *CO_2 Capture, Reuse, and Storage Technologies for Mitigating Global Climate Change—A White Paper*, Final Report, DE-AF22-96PC01257; MIT Energy Laboratory: Cambridge, MA, Jan. 1997.
37. Broecker, W.S.; Takahashi, T. Neutralization of fossil fuel CO_2 by marine calcium carbonate. In *The Fate of Fossil Fuel CO_2 in the Oceans*; Andersen, N.R., Malahoff, A., Eds.; Plenum Publishing Corp.: New York, 1978.
38. Dendy, E.; Sloan, J. Fundamental principles and applications of natural gas hydrates. Nature **2003**, *426*, 353–359.
39. Marchetti, C. On geoengineering and the CO_2 problem. Climatic Change **1977**, *1*, 59–68.
40. Freund, P.; Ormerod, W.G. Progress toward storage of carbon dioxide. Energy Convers. Manage. **1997**, *38* (Suppl.), S199–S204.
41. Herzog, H.; Adams, E.; Auerbach, D.; Caulfield, J. *Technology Assessment of CO_2 Ocean Disposal*, Report 95-001; MIT Energy Laboratory: Cambridge, MA, 1995.
42. Stager, C. Silent deaths from Cameroon's killer lake. Natl. Geogr. **1987**, *172* (3), 404–420.
43. Herzog, H.; Drake, E.; Tester, J.; Rosenthal, R. *A Research Needs Assessment for Capture Utilization, and Disposal of Carbon Dioxide from Fossil Fuel-Fired Power Plan*, DOE/ER-30194; U.S. DOE: Washington, DC, 1993.
44. Ruether, J.; Dahowski, R.; Ramezan, M.; Schmidt, C. *Prospects for Early Deployment of Power Plants Employing Carbon Capture*; National Energy Technology Laboratory: Pittsburgh, PA, 2002.
45. Holloway, S. Storage of fossil fuel-derived carbon dioxide beneath the surface of the Earth. Annu. Rev. Energy Environ. **2001**, *26*, 145–166.
46. Gale, J. Geological storage of CO_2: What do we know, where are the gaps and what more needs to be done? Energy **2004**, *29*, 1329–1338.

47. Bergman, P.D.; Winter, E.M. Disposal of CO_2 in aquifers in the U.S. Energy Convers. Manage. **1995**, *36*, 523–526.
48. Herzog, H.; Eliasson, B.; Kaarstad, O. Capturing greenhouse gases. Sci. Am. **2000**, *2*, 72–79.
49. Lackner, K.S.; Wendt, C.H.; Butt, D.P.; Joyce, E.L.; Sharp, D.H. Carbon dioxide disposal in carbonate minerals. Energy **1995**, *20* (11), 1153–1170.
50. Goff, F.; Lackner, K.S. Carbon dioxide sequestering using ultramafic rocks. Environ. Geosci. **1998**, *5* (3), 89–101.
51. Lackner, K.S.; Butt, D.P.; Wendt, C.H.; Goff, F.; Guthrie, G. *Carbon Dioxide Disposal in Mineral Form: Keeping Coal Competitive*, LA-UR-97-20941997, Los Alamos National Laboratory Report; Los Alamos National Laboratory: Los Alamos, NM, 1997.
52. Penner, L.R.; O'Connor, W.K.; Gerdemann, S.J.; Dahlin, D.C. Mineralization strategies for carbon dioxide sequestration. Proceedings of the 20th Pittsburgh Coal Conference, Pittsburgh, PA, Sep. 15–19, 2003.
53. Park, A.-H.; Jadhav, R.; Fan, L.-S. CO_2 mineral sequestration: chemical enhanced aqueous carbonation of serpentine. Can. J. Chem. Eng. **2003**, *81* (3–4), 885–890.
54. Park, A.-H.; Fan, L.-S. CO_2 mineral sequestration: physically activated dissolution of serpentine and pH swing process. Chem. Eng. Sci. **2004**, *59* (22-23), 5241–5247.
55. Lackner, K.S. Can carbon fuel the 21st century? Int. Geol. Rev. **2002**, *44* (12), 1122–1133.

Carbon Fibers from Lignin-Recyclable Plastic Blends

C

Satoshi Kubo
John F. Kadla
Faculty of Forestry, Biomaterials Chemistry, The University of British Columbia, Vancouver, British Columbia, Canada

INTRODUCTION

Recently, environmental concerns have sparked interest in utilizing biodegradable and bio-derived materials in various industrial fields. Particularly, the utilization of plant-derived materials such as agricultural residues and forest products has received increasing attention. Lignin is one of the most abundant biomacromolecules existing in the plant kingdom. An enormous amount of lignin is produced as a by-product of the pulp and paper industry. As a result, a number of systems have been proposed for the utilization of lignin as a renewable polymeric material. However, lignin utilization is still limited, with less than 2% of the lignin produced being used in commercial applications; lignin is used as dispersants, adhesives, and surfactants. This has been attributed, at least partially to the fact that most of the systems being developed require chemical modification of the lignin or the use of noncommercial lignins such as organosolv lignins. Polymer blending is a convenient and inexpensive method to develop novel polymeric materials. In previous studies, we had shown that lignin forms a miscible polymer blend with poly (ethylene oxide) (PEO). Blending with PEO improved some of the unfavorable properties of lignin, such as moldability and flexibility. However, these blends are still unattractive as a commercial material because of their dark color and unsuitable mechanical properties for structural materials.

One potential application of lignin-based fibers is carbon fibers. Carbon fibers are lightweight materials with prominent mechanical properties. The conversion of lignin fibers into carbon fibers resulted in improved mechanical properties. In our research, it was found that the porosity of lignin–synthetic plastic-based carbon fibers could be controlled by careful selection of the blending synthetic plastic. Immiscible blend fibers resulted in hollow and porous fibers without any further treatment after carbonization. Due to the difference in thermal stability of the synthetic polymer and the lignin, careful control of thermal treatment conditions enables the removal of the synthetic polymer from the lignin fiber. In the manufacturing process of carbon fibers, three fundamental steps are involved: fiber spinning, thermostabilization, and carbonization. In the case of the lignin, the thermostabilization process is sometimes omitted because of the auto-oxidative reactions that occur when lignin is heated to the high temperatures during carbonization. However, the softening point of the lignin fiber is low, and this can make it difficult to maintain the fibrous form during carbonization. The addition of a synthetic polymer will also affect this process; the use of a low T_g material will decrease the softening point of the blend relative to lignin, whereas a high T_g material will increase it. Therefore, depending on the synthetic polymer being blended with the lignin, the necessity of the thermostabilization stage will vary.

In this entry, the effect of blending recyclable poly(propylene) (PP) and poly(ethylene terephthalate) (PET) with lignin on carbon fiber production is presented. We discuss the effects of lignin structure and specific intermolecular interactions on lignin thermal properties as well as the effect of blend composition on surface morphology, mechanical properties, and the manufacturing process of lignin/recyclable plastic-based carbon fibers.

UTILIZATION OF LIGNIN

Extensive cross-linking and strong intramolecular interactions of polymeric lignins constrain the utilization of polymeric lignin in solid material systems;[1,2] hardwood lignins are more easily processed than softwood lignins. Due to the inherent chemical and molecular weight heterogeneity, lignin and lignin derivatives have limited utility in applications demanding a constant well-defined feedstock. In fact, less than 2% of the total available lignin is reportedly used in higher value products.[3] In 1996, total sales in lignin-based specialty products, such as animal bypass protein, agrochemicals, dispersants, adhesives, and surfactants, were reported at $600 million.

One of the most important applications of lignin and its derivatives is in the area of plastic products. Lignin displays unique thermal behavior, showing not only thermoplastic but also thermosetting properties. Thermal properties, such as glass transition, thermal fluidity, and others, are also dependent on

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039775

the isolation method and the lignin source (discussed earlier). Since the 1930s, lignin has been used in plastic materials.[4] As a result, a considerable amount of information exists regarding the modification of lignins toward the engineering of plastics. Unfortunately, in most cases, the incorporation of various monomers or polymers into the lignin structure results in properties unsuitable for structural materials.[1] Glasser and coworkers[5,6] have shown that through the manipulation of the network structure and substituents in lignin, the physical properties of lignin-based materials can be manipulated. They found that the noted brittleness of lignin, caused by the globular structure of lignin fragments, could be abolished by incorporating a variety of polyether components in the network structure. That is a decrease in glass transition temperatures and brittleness could be achieved through the introduction of ''soft'' molecular segments capable of a plastic response to mechanical deformation.

Lignin-based polyblends or polymer alloys have also been developed to enhance the thermoplastic behavior of lignin.[7] However, the amount of lignin incorporation into the polymeric materials is generally limited, because of the inherent brittleness of the lignin phase or phase separation due to immiscibility. Recently, Li, Mlynar, and Sarkanen[2] have produced homogeneous blends containing 85% (w/w) underivatized industrial kraft lignin with poly(vinyl acetate) and two plasticizers. Likewise, we have shown that incorporation of small amounts of PEO into blends with kraft[8] and organosolv[9] lignin can dramatically improve the thermal and physical properties of lignin-based materials. The various lignin-based thermoplastics exhibited promising mechanical properties, with the tensile behavior of these polymeric materials directly dependent upon the degree of association between the intrinsic kraft lignin components.

LIGNIN-BASED CARBON FIBERS

One area of lignin utilization that is receiving increasing attention is in the development of advanced composite fibers. Carbon fibers are one of the most important engineering materials in advanced composites. They are lightweight, fatigue resistant materials that possess high strength and high stiffness. These unique properties result from their flawless structure and the development of highly anisotropic graphic crystallites orientated along the fiber axis during the production process.[10] Carbon fibers are manufactured by thermally treating fibers at 1000–2000°C in an inert atmosphere while maintaining the fibrous structure. This is aided by a stabilization stage in which the precursor fibers are heated under tension at 200–300°C in the presence of air. This causes cross-linking on the fiber surfaces, among other reactions, and prevents shrinking, melting, and fusing.

The first commercially produced carbon filament was made from a cellulosic precursor for its application as an incandescent lamp filament in 1879.[11,12] Its use as a reinforcing material in composite products began in the 1950s as the aircraft and aerospace industries led the search for new materials with properties superior to the then available metals. Cost was seldom a deciding factor. As the years passed, the advantages of advanced composite materials in other applications became apparent. Today carbon fiber composite products are routinely used in sports equipment, marine products, construction and automotive industries to name a few. However, unlike in its infancy, the cost of carbon fiber production has limited its widespread use. The fundamental difficulty is reducing the cost of manufacturing the precursor, the so-called ''white fiber.'' There are primarily three types of precursor materials of commercial significance: pitch (petroleum or coal), viscose rayon, and polyacrylonitrile (PAN).[13] Of these three, PAN is most important for structural applications.[a] It is an excellent precursor material that has been widely researched and is in wide commercial production. Almost 80% of commercially available carbon fibers are derived from PAN. However, current PAN technology is expensive, thereby limiting its utilization in lower cost general performance (GP) applications, and research continues toward decreasing the PAN precursor costs, e.g., Amlon (BP Chemicals).[14]

The overwhelming success of PAN-based carbon fibers over rayon and pitch can be attributed to several key aspects.[13] Structurally, PAN has a faster rate of pyrolysis without much disturbance to its basic structure and to the preferred orientation of the molecular chains along the fiber axis present in the original fiber. By contrast, carbon fibers from rayon suffer from extremely low carbon yield (20–25%) due to chain fragmentation, which eliminates the orientation of the precursor fiber. While improved properties can be achieved by stretch graphitization, this process is expensive and does not compensate for the low yields.

Pitch-based carbon fibers, on the other hand, provide properties not readily obtainable with PAN-based fibers. PAN-based fibers have excellent tensile strength at a modulus of 200 GPa, but the strength decreases as the modulus is increased.[14] Pitch-based fibers have lower tensile strengths, but are capable of modulus levels up to the theoretical modulus of graphite, 1000 GPa, and have much better thermal and electrical conductivity properties than PAN-based fibers.

[a]Developed in the 1940s by DuPont and Union Carbide, PAN began to be used as a carbon fiber precursor material in the 1960s through work done in Japan (Toray) aand Great Britain (Courtaulds).

Thus, pitch and PAN-based fibers are really complementary to one another, each filling a different set of commercial needs.

In addition to rayon, pitch, and PAN, many polymeric materials have been used to make carbon fibers.[15–24] However, only the big three, rayon, pitch, and PAN, have endured the high-performance markets. Their price has dropped over the years, but remains high, accounting for over one-half of the production costs, too high for the GP markets. The literature supports that recycled petrochemical polymers and fibers and renewable cellulosics and lignins, which are inexpensive and widely available, may be potential feedstocks for GP carbon fibers.[15–24]

Carbon fibers have been manufactured from lignin. One of the first reports of lignin-based carbon fibers was from Otani et al.[25] Utilizing thiolignin, alkali lignin, and lignosulfonates, carbon fibers were produced by both thermal and solvent spinning followed by thermal processing to 600–1000°C. Lignin fibers were produced by thermal spinning of softwood and softwood/hardwood kraft lignin blends. A purification process was needed to remove impurities and facilitate fiber spinning. The fiber quality was poor, requiring slow winding rates (5–10 m/min) and producing low tensile strength (300 MPa) carbon fibers. Improvements in spinnability and tensile strength were obtained by using dry spinning. Satisfactory lignin fibers were formed from kraft lignin; alkali lignin; and lignosulfonate spun from water, alkaline water, and glycerol. Blending with poly(vinyl alcohol) (PVA) or PAN enhanced the spinnability and the tensile strength of the fibers. Winding rates of 600 m/min were achieved when softwood kraft lignin (SKL) was blended with PVA (3/4, w/w). Fine fibers with excellent mechanical properties were produced. In the production of carbon fibers, an oxidation process, referred to as stabilization or ''thermostabilization,'' is generally conducted to induce cross-linking to maintain fiber integrity during the subsequent carbonization process. It was expected that lignin-based carbon fibers, due to the high oxygen content in lignin, could be produced with shorter stabilization periods at lower temperature than other feedstocks. The various hydroxyl and ether groups, via oxygen-based radicals, were expected to facilitate cross-linking within the lignin. In fact carbon fibers were prepared with very little thermostabilization, and tensile strengths as high as 800 MPa were obtained.[25]

Around the same time, Mansmann and coworkers reported the production of carbon fibers from a variety of dry-spun materials, including lignin (lignosulfonates) by the simple addition of small amounts of PEO or acrylic acid–acrylamide copolymers.[26] Although similar to the procedure of Ohtani, Mansmann employed acidic rather than neutral and/or alkaline spinning conditions. It was reported that the spinning under acidic conditions enhanced the mechanical properties of the lignin and carbon fibers. Furthermore, the amount of synthetic polymer required to enhance spinnability was much less in comparison to the method of Otani. The concentration of the polymer required for favorable spinning was dependent on its molecular mass. Only 0.3–0.6% (wt%) PEO, DP = 100 K, was needed for good spinning. In addition, dry spinning from acidic conditions reduced the thermostabilization temperature to between 100°C and 150°C. Tensile strengths of ~100 MPa and ~600–800 MPa, and Young's modulus of 2–8 GPa and 33.3 GPa were reported for the lignin and carbon fibers, respectively.

Based on these encouraging results, pilot-scale production of lignosulfonate-based carbon fibers, Kayacarbon, was developed and made commercially available by Nippon Kayaku Co.[27] Carbon fiber was prepared by dry spinning a lignosulfonate plasticized with PVA. Pilot trials revealed that lignin-based carbon fibers could be produced at lower temperatures with shorter stabilization periods than other feedstocks, such as pitch, because cross-linking reactions that produce the graphitic structure are facilitated by oxygen-based radicals derived from the various hydroxyl and ether groups within lignin. Close examination of the carbon fibers produced showed poorly developed fibrillar structures and lack of homogeneity between the crystalline planes. A large number of inclusions were found, arising from catalytic graphitization by sodium impurities, which contributed to a weakening of the fibers.[28,29] Microvoids of complicated size and shape were also observed.[30,31] The sizes of microvoids, as well as of the solid phase around them, increased with increasing heat treatment temperature; stretching of the fiber promoted further increases. These flaws, combined with the lack of orientation, resulted in low modulus and low tensile strength carbon fibers. As a result, these carbon fibers were abandoned because they did not meet the specifications of the high-performance aerospace industry, which dominated the carbon fiber market.

Sudo and coworkers demonstrated that steam-exploded lignin could be converted into a pitchlike molten material using hydrogenolysis followed by solvent extraction and polycondensation under high temperatures.[32,33] The lignin pitch had H/C ratios between 1.03 and 1.06, a softening point of 110°C, and melted at over 145°C to form a viscous liquid with suitable properties for thermal spinning and thermostabilization. A continuous fine filament could be produced at a speed of over 100 m/min. The fibers were heated in air from 1–2°C/min to 210°C, converting the lignin fibers to an infusible material, enabling them to be carbonized at a heating rate of 5°C/min up to 1000°C in a stream of nitrogen. The typical properties

of the lignin-based carbon fiber were: fiber diameter = 7.6 ± 2.7 μm; elongation = 1.63 ± 0.29%; tensile strength = 660 ± 230 MPa; modulus of elasticity = 40.7 ± 6.3 GPa. Unfortunately, the overall yield of the carbon fiber was only 20.7%, well below that of other lignin-based carbon fibers. To enhance the overall yield, hydrogenolysis was replaced by phenolation using creosote.[34,35] This resulted in yields of 40.0–49.2%.

In addition to steam-exploded lignin, Alcell and acetic acid lignins were studied. The spinnability of the phenolated Alcell was better than that of the other phenolated lignins. The phenolated Alcell lignin had a softening point of 150–170°C and could be continuously spun for 500 sec at a winding rate of 1000 m/min. min. To improve the spinnability of the phenolated acetic acid lignin, it was blended with pitch in the ratio of 1/1 (w/w). The overall yield of the carbon fibers made by this method was substantially higher than that of the hydrogenolysis lignin. Overall yields of 40.0–49.2% and 49.6–57.3% were reported for the phenolated and pitch-added phenolated lignin-based carbon fibers, respectively. Superior properties to that of the Kayacarbon were observed, with a tensile strength of 614 MPa, Young's modulus of 31.8 GPa, and 1.94% elongation being reported for the phenolated acetic acid lignin carbon fiber.

More recently, Sano et al. produced lignin fibers suitable for carbon fibers by thermal spinning organosolv lignin obtained from aqueous acetic acid pulping of birch.[36–38] Unlike the previously reported acetic acid lignin fibers, this system did not require any chemical modification (e.g., phenolysis). The spinnability of the organosolv lignin was attributed to the polydispersity and partial acetylation of the lignin due to the pulping process. The raw acetic acid lignin showed poor spinnability due to the formation of pyrolyzed gaseous materials in the extruder nozzle. Thermal treatment at 160°C under vacuum for 30 min removed this pyrolyzable material permitting excellent fiber spinning at more than 400 m/min. Tensile strength of 30.8 ± 3.5 MPa, Young's modulus of 3.59 ± 0.43 GPa, and 0.81 ± 0.08% elongation were reported for this lignin fiber. This fiber was easily carbonized after a thermostabilization process (0.5°C min^{-1} to 250°C). Tensile strength of 355 ± 53 MPa, Young's modulus of 39.1 ± 13.3 GPa, and 0.98 ± 0.25% elongation was reported for this lignin fiber. The tensile strength increased with decreasing fiber diameter and was comparable to previously reported lignin-based carbon fibers if fine carbon fiber was made from this lignin. Overall carbon fiber yield of 32.7% was obtained, comparable to that of isotropic petroleum pitch, ~30–35%, but lower than that reported for the phenolated acetic acid lignin of Sudo and coworkers.[34] Shortly thereafter, Sano et al. reported on the production of carbon fiber from softwood acetic acid lignin. The infusible softwood lignin was made fusible by removing the infusible high molecular mass fraction (~30% of the lignin) by successive extraction with various concentrations of acetic acid. Good fiber spinning was produced after thermal treatment of this lignin fraction. Moreover, this fiber was converted to carbon fiber without a thermostabilization process. Tensile strength of 26.4 ± 3.1 MPa, 147 ± 51 MPa, Young's modulus of 3.59 ± 0.43 GPa, 19.5 ± 5.5 GPa, and 0.71 ± 0.14%, 0.75 ± 0.27% of elongation were reported for this lignin fiber and carbon fiber, respectively.

At about the same time, Itoh reported that lignin prepared by the organosolv pulping using phenols and water could be thermally spun into lignin fibers for the production of carbon fiber.[39] Fiber spinning was achieved at a spinning temperature of between 150°C and 250°C without any lignin modification required. Tensile strength of 3–60 MPa was reported for the thermostabilized fiber; however, no information about the properties of the carbon fiber has been reported.

Although both petrochemical polymers and lignin were successfully used as carbon fiber precursors with good carbon fiber properties, these materials were gradually abandoned as the high-performance aerospace market, which relied increasingly on PAN and pitch, came to dominate carbon fiber production. Today however, carbon fibers are routinely being used in marine products, construction, the transportation industry, and the like. Cost, not performance, is now the driving force in their utilization. Thus, any technology that can produce a low cost fiber precursor, suitable for carbon fiber, will have a great impact on these markets. As discussed, recycled petrochemical polymers and industrially produced lignins can be used to produce carbon fibers. However, in each system discussed here, problems concerning production costs exist. Either plasticization or lignin modification was required, or noncommercially produced lignins (e.g., steam-exploded, acetic acid) were used. Therefore, to overcome this we set out to utilize an industrially produced technical lignin for the production of carbon fibers with properties suitable for the general purpose carbon fiber market.

Recently, we reported the first carbon fibers produced from an industrially produced kraft lignin, without any chemical modification. The process involved thermal spinning followed by carbonization. A fusible lignin with excellent spinnability to form a fine filament was produced with a thermal pretreatment under vacuum. Blending the lignin with PEO further facilitated fiber spinning, but at PEO levels greater than 5–10%, the blends could not be stabilized without the individual fibers fusing together. Carbon fibers produced had an overall yield of 45%. The tensile strength and modulus increased with decreasing fiber

diameter and are comparable to those of much smaller diameter carbon fibers produced from phenolated-exploded lignins. In view of the mechanical properties, tensile strength 400–550 MPa and modulus 30–60 GPa, kraft lignin should be further investigated as a precursor for general grade carbon fibers.

Lignin is a renewable, nontoxic, commercially available, and low cost natural resource. Lignin has enormous potential as a raw material for the polymer and composites industries. In spite of many years of development effort, the full potential of lignin has not been fully utilized. One area of potential utilization is in polyblend fibers for composite applications. In our previous lignin/synthetic polymer blend carbon fiber studies, pure PEO has been used to improve spinning ability. More recently, we have developed the lignin-based carbon fibers using recyclable plastics, such as PET and PP. In this entry, we introduce the preparation method, processability, and properties of these hardwood kraft lignin (HKL)/recyclable plastics-based carbon fibers.

THERMAL AND CHEMICAL PROPERTIES OF LIGNIN

Lignin is a natural polymer found in wood. Lignin is readily available and relatively inexpensive. Lignin structure is dependent on wood species and processing conditions used to isolate it. Lignin does not have a well-ordered structure like most of the other natural and synthetic polymers. It is a polyaromatic polyol made up of phenyl propane monomer units linked by carbon–carbon and ether linkages. The monomer units can be classified into three different types: guaiacyl, syringyl, and 4-hydroxyl-phenyl propane structures (Fig. 1). These monomer units undergo random polymerization during the lignification process. The macromolecular structure of lignin, type and ratio of typical interlinkages of the monomer units differ between genera, tissues, and even cell wall layers within the same plant.

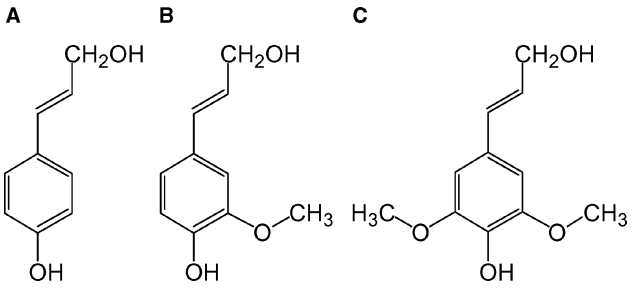


Fig. 1 Precursors of lignin biosynthesis. (A) *p*-Coumaryl alcohol (4-hydroxyl-phenyl propane); (B) coniferyl alcohol (guaiacyl); and (C) sinapyl alcohol (syringyl).

Currently, there is no method to isolate native lignin from plants in its native form. Chemical and physical modifications are unavoidable during the lignin isolation process. Commercial lignins, such as those isolated from alkaline, acidic, or organic solvent-based processes, have undergone extensive fragmentation and degradation. As a result, a wide variety of lignins, in terms of chemical properties and structures, can be obtained. Specifically, depending on the type and length of chemical processing, the lignins will vary in molecular weight, functional groups present, degree of condensation, types of inter-monomeric linkages, and types and ratios of monomeric units. However, the chemical structure of lignin is typically drawn as a statistical representation of the fundamental interunit linkages (Fig. 2).

There are several reports in the literature describing the effect of chemical structure on the observed thermal properties of lignin.[1,3,40–42] In the thermal processing of lignin for structural applications, the relationship between chemical structure and thermal properties is of extreme importance. Differential scanning calorimetric (DSC) analysis of HKL and SKL is shown in Fig. 3. The glass transition temperature (T_g) of the HKL appears at 93°C, much lower than that of the SKL, 119°C.

The observed difference in thermal transition temperatures between the softwood and hardwood lignins can be explained based on differences in the chemical structure between the two lignins. Hardwood lignins are made up of both guaiacyl (G) and syringyl (S) structures, whereas softwood lignins are primarily composed of guaiacyl units (Fig. 2). As a result, softwood lignins contain more condensed interunit linkages, such as 5-5′ and β-5. These ''cross-links'' or ''branch-points'' produce a more complex network-like structure, and hinder the rotational and translational motions within the lignin macromolecule increasing the T_g. Although a minor component of native lignins, accounting for approximately 20% and 10% of the interunit linkages in softwood and hardwood lignins, respectively, these condensed units are less reactive toward chemical processing. Thus, the relative amount of these condensed linkages increases, becoming more significant in residual and technical lignins.[43] In addition, processing conditions can also affect the functional groups and molecular weight of the respective lignins.[44]

To better understand the differences in thermal properties between the various lignins, further structural investigations were performed using nuclear magnetic resonance (NMR) spectroscopy. Fig. 4 shows the oxygenated aliphatic region of the HMQC NMR spectra for an HKL and an SKL. Included in the spectra are the assignments for the pinoresinol (β-β), phenylcoumaran (β-5), and β-*O*-4 structure along with α- and γ-hydroxyl group assignments. As can be clearly seen, β-*O*-4 structures survive chemical processing and remain in the industrial lignins. Further, qualitatively

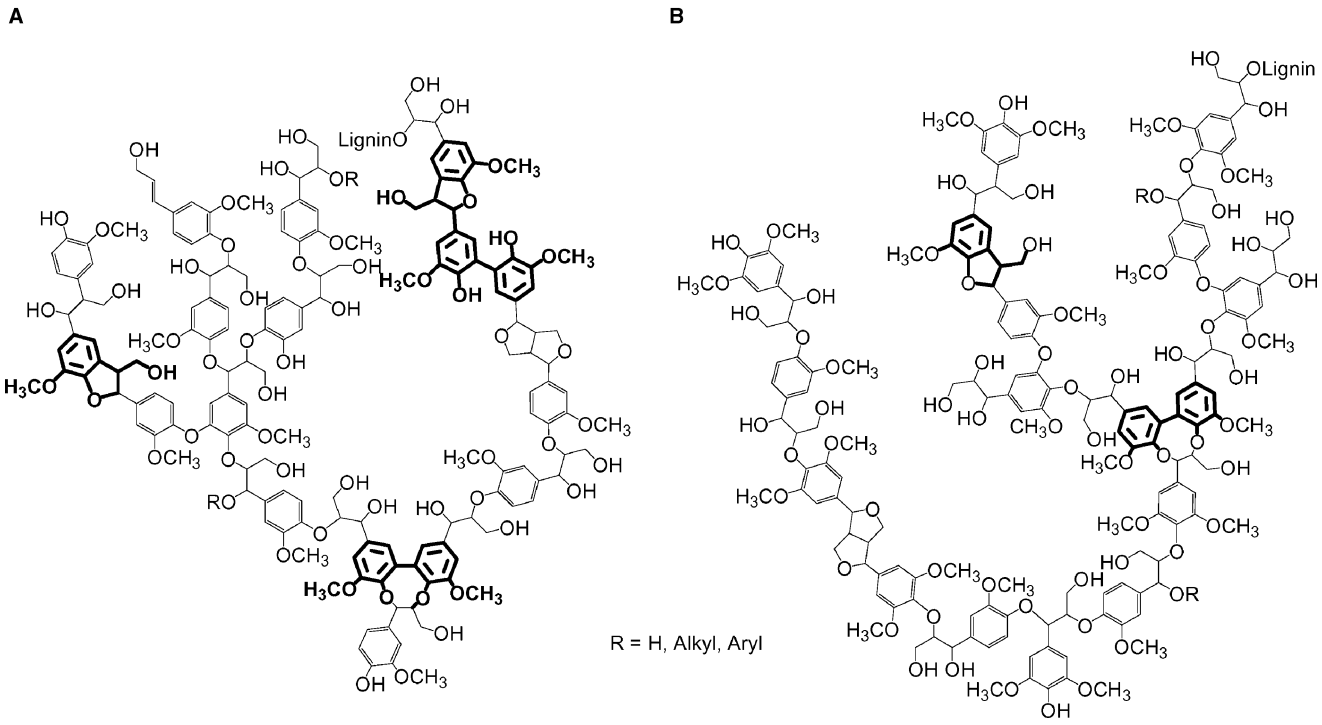


Fig. 2 Structural representation of softwood (A) and hardwood (B) lignin.

the two lignins contain many of the same interunit linkages and appear structurally quite similar. Therefore to better understand the differences in thermal behavior between the HKL and the SKL, other detailed chemical analyses were performed and the results are summarized in Table 1.

POLYMER BLENDING AND FIBER SPINNING

Generally, in thermoforming and/or molding of polymer blends, the processibility relates to the mixing property, e.g., miscibility. To determine miscibility of polymer blends thermal analysis, DSC in particular is widely used. In DSC, a single compositional-dependent glass transition temperature (T_g) is an indication of full miscibility at a dimensional scale between 5 and 15 nm.[45] Recently, we have investigated various lignin/synthetic polymer blends, including poly(ethylene) (PE), PET, PEO, and PVA.[8,9,46,47] Most polymer blends are immiscible due to the low entropy of mixing. In fact most of the lignin polymer blend systems we have studied are immiscible. However, we have found that lignin forms miscible blends with PEO and PET. DSC curves of HKL/PEO and HKL/PET are shown in Fig. 5. Only one T_g is observed in these blends appearing between the T_g values of the individual blend components. Further, a decrease in the melting temperature of the synthetic polymer component is also observed. The DSC results confirm the formation of miscible blends in the amorphous phase. By contrast, two T_gs are observed for HKL/PP blend fibers, as was the melting peak of the PP fraction in all blend fibers (Fig. 5B).

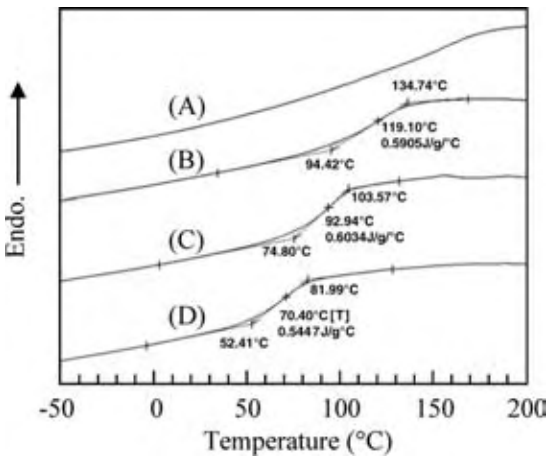


Fig. 3 DSC curves of isolated lignin. (A) Sodium lignosulfonate; (B) SKL; (C) HKL; and (D) organosolv lignin.

In all these systems, the lignin was prepared according to our thermal extrusion method for carbon fiber applications. It is known that lignin has unique thermal properties, wherein during typically thermal processing it undergoes polycondensation reactions. Although classified as a thermoplastic material, exposure to excess thermal energy can transform lignin into a thermosetting material. In our method, fine HKL fibers

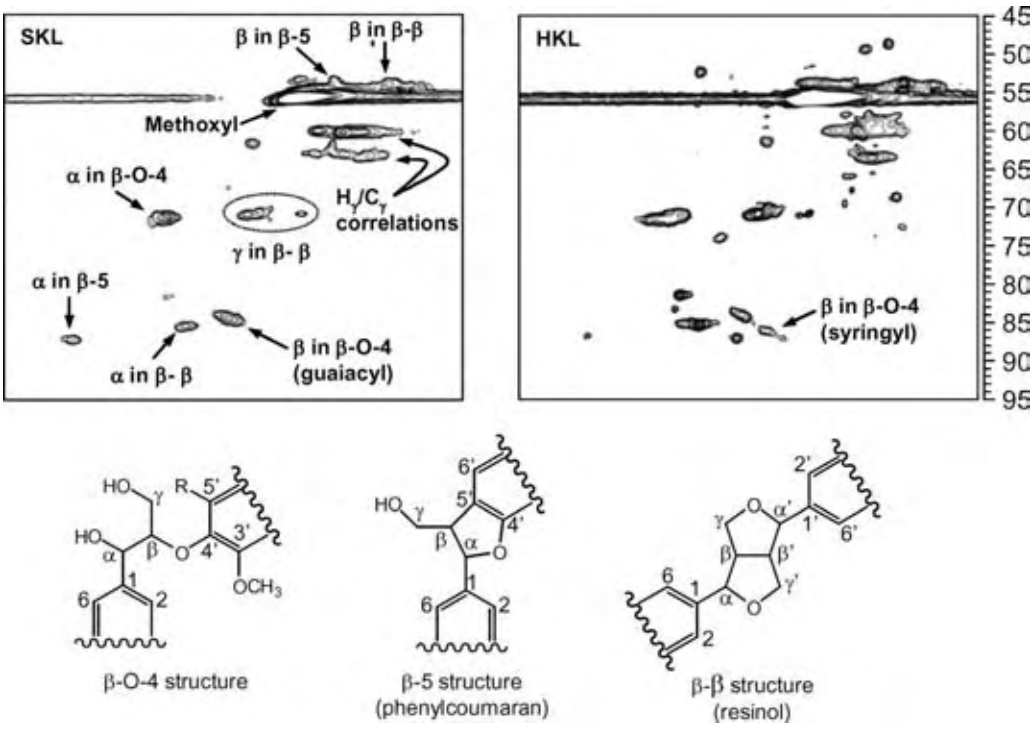


Fig. 4 Oxygenated aliphatic region of the HMQC NMR spectra of SKL and HKL.

can be spun by thermal extrusion without significant chemical modification. However, careful temperature control is required for successful fiber spinning. The thermal fiber spinning of lignin is improved through polymer blending. Our initial studies focused on using PEO kraft lignin blends. As discussed earlier, kraft lignin forms miscible blends with PEO. The blending with PEO dramatically improved the fiber spinning of the lignin-based fibers. However, incorporation of high levels of PEO had a negative impact on the subsequent carbon fibers. Due to the low glass transition temperature of PEO, during the thermostabilization process (discussed later), the lignin/PEO fibers fused together as a result of the heat treatment. Therefore, PEO levels of less than 5–10 wt% were necessary to maintain fiber integrity.

The thermal blending and the fiber spinning properties of HKL/synthetic polymer blends are not only affected by the T_g of the blending polymer, but also by the specific interactions formed between the blend components. As described earlier, PET is believed to form miscible blends with HKL. HKL/PET fibers can be prepared by thermal extrusion.[46] However, the thermal fiber spinning of HKL/PET blends is problematic. The thermal viscosity dramatically decreases at the spinning temperature, reducing the processibility of the HKL/PET blends. Such thermal viscosity behavior is not observed for either homopolymer. This suggests that specific interactions form between blend components during the spinning process. This is currently being further investigated.

By contrast, other polymers such as polyolefins and PVA form immiscible blends with HKL. In fact, phase separation is observed in fibers produced from the thermal spinning of HKL and PP. The fiber spinning properties of several PP samples having various melt viscosities and tacticities were examined. Most of the HKL/PP blends show poor or bad fiber spinnability. However, excellent fiber spinning was achieved with PP samples having a thermal viscosity comparable to the HKL.

Table 1 Chemical properties of SKL and HKL

	Functional groups (mmol g^{-1})					
	Hydroxyl groups				Molecular mass	
Lignin	Aliphatic	Phenolic	Methoxyl	S G^{-1}	M_w	Dispersity
SKL	5.6	3.8	4.2		2,800	2.0
HKL	4.1	4.3	5.9	1.2	2,400	1.8

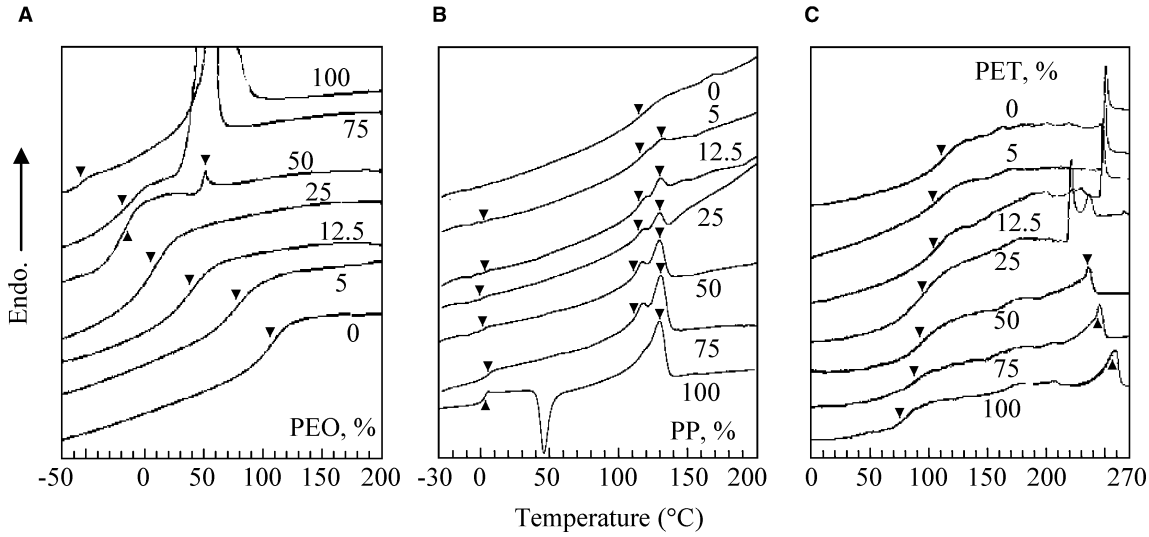


Fig. 5 DSC curves of HKL/synthetic polymer blends. (A) HKL/PEO; (B) HKL/PP; (C) HKL/PET.

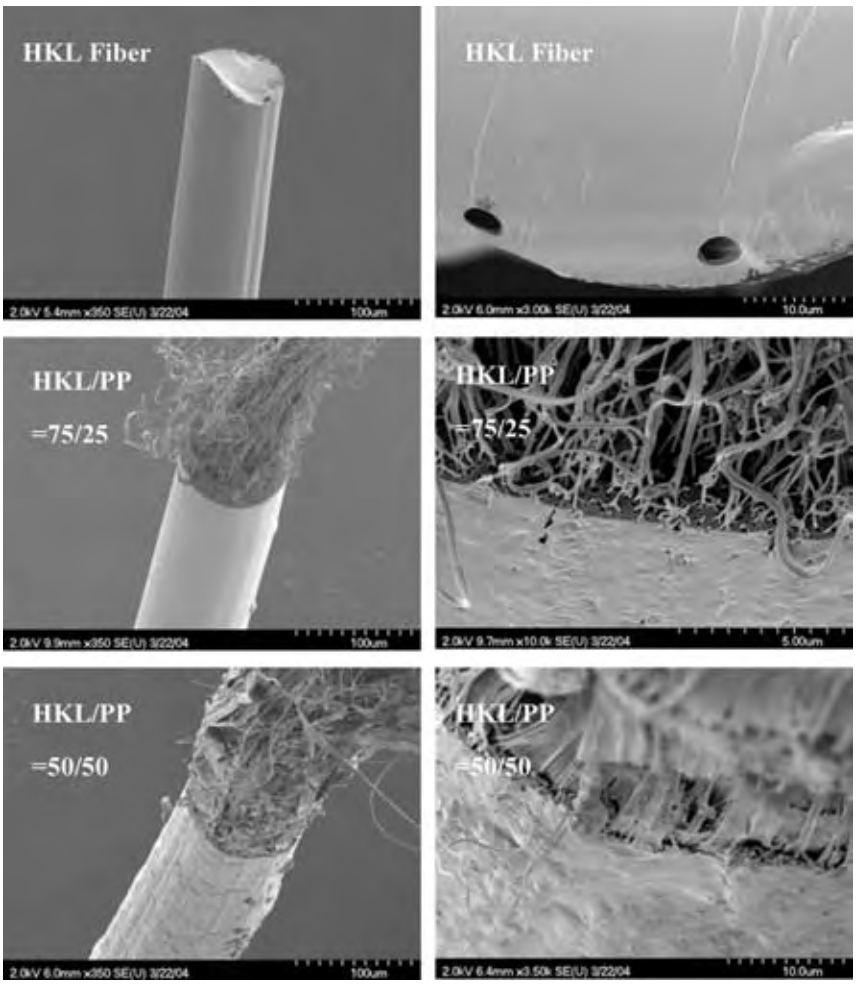


Fig. 6 SEM image of HKL and HKL/PP blend fibers.

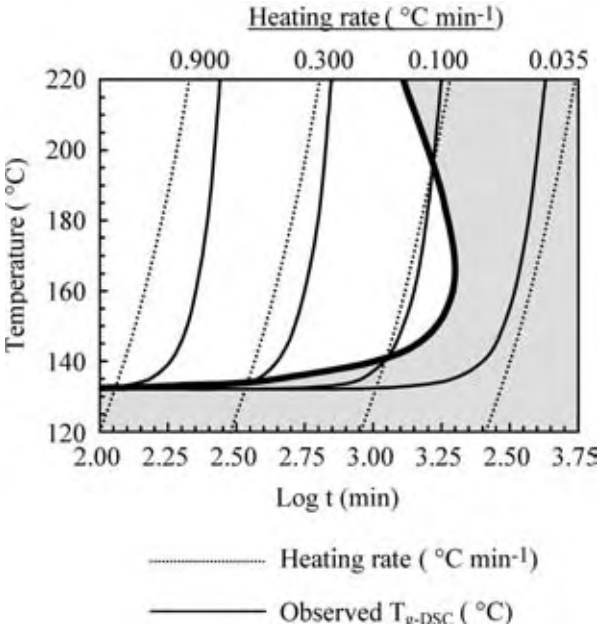


Fig. 7 CHT diagram for lignin oxidation.

FIBER MORPHOLOGY

Lignin/synthetic polymer blend fibers have varying surface morphology depending on the miscibility of blend components. Fig. 6 shows scanning electron microgram (SEM) images of HKL and HKL/PP blend fibers. The HKL fibers have a relatively smooth surface with the appearance of some large voids in the cross section. During thermal processing, some thermal decomposition may occur. These pores may be the result of entrained decomposed gases. However, specific fine pores are not observed on the fiber surface. The miscible lignin/synthetic polymer blend fibers also have smooth surfaces, although a fine regular pattern is observation under high magnification (160,000X).

By contrast, the immiscible lignin/PP blend fibers form unique core/shell structure due to the difference in molten viscosity between HKL and PP.[48,49] The PP component is located in the core of the fiber. Depending on the amount of PP incorporated, various morphologies are observed. In high HKL content blend fibers, the PP appears as nanoscale fibers bundled within the lignin shell. As the PP content is increased, the PP nanoscale fibers merge forming ribbonlike structures within the core (HKL/PP = 50/50 blend fiber).

THERMOSTABILIZATION PROPERTIES OF LIGNIN AND LIGNIN/SYNTHETIC POLYMER BLENDS

To convert the fibers produced from thermal spinning into carbon fibers, a thermostabilization stage is necessary to prevent fiber fusing. In the case of petroleum pitch, infused oxygen bridging and cross-linking mainly in the alkyl components is introduced by thermal oxidation during this process. However, in the case of lignin, the alkyl components are already partially oxidized (Fig. 2). Therefore, thermostabilization will also involve polycondensation and dehydration in addition to some oxidation of the lignin macromolecule. As a result, increases in molecular weight and in T_g will occur. As lignin contains various functional groups with the potential to undergo thermally induced chemical reactions during heating, it has been speculated that a thermostabilization step may not be necessary in the manufacturing process of lignin-based carbon fibers. However, our results indicate that these reactions do not produce a sufficient increase in the T_g to prevent softening during carbonization.

Ideally, the heating rate during thermostabilization must be slow enough to allow the T_g to increase faster than the thermostabilization temperature. To predict the behavior of lignin and its polymer blends during thermostabilization, we studied the effect of heating rate on changes in the thermal properties of lignin. A continuous heating transformation (CHT) diagram,

Table 2 Thermostabilization properties of HKL and HKL/synthetic polymer (75/25 w/w) fibers under air atmosphere

Fiber	Heating rate ($^\circ C\,hr^{-1}$)					
	12	30	60	120	180	300
HKL 100%	•	○	○	○	⊗	⊗
HKL/PEO[a]	⊗	⊗	⊗	⊗	⊗	⊗
HKL/PET	•	•	•	•	•	•
HKL/PP	○	○	○	⊗	⊗	⊗

•Excellent fiber integrity.
○Good fiber formation, fibers stick together but could be separated.
⊗Poor fiber integrity, fibers fuse together.
[a]HKL/PEO = 87.5/12.5.

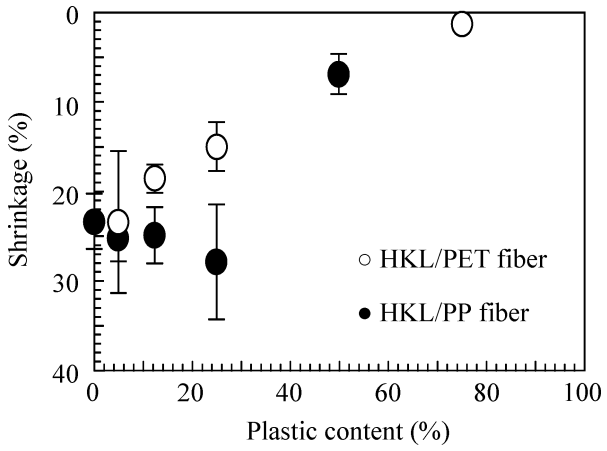


Fig. 8 Thermal dimensional stability of lignin/synthetic plastic in the thermostabilization process.

which is similar to the ones used by Wisanrakkit and Gillham,[50] was constructed from the DSC analysis using various heat treated lignin under different heating rate. The lignin CHT diagram is shown in Fig. 7.

The CHT diagram is a phase diagram consisting of plots of reaction temperature and T_g vs. time for various constant heating rates. The dotted lines show the temperature of each heating rate in time. The solid black lines are the T_g contours. The bold curve represents the points for which reaction temperature equals the T_g. The region to the left of the bold curve represents conditions that lead to softening of the sample. In order to avoid softening, the sample must be heated at a rate that falls outside of this region (gray area).

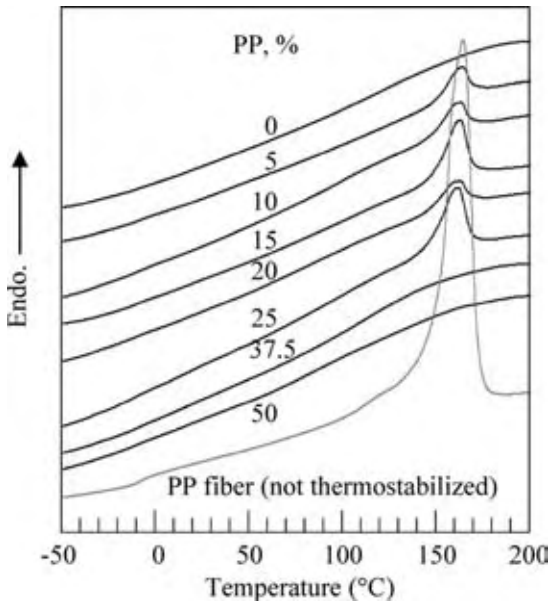


Fig. 9 DSC curves of thermostabilized HKL and all PP/HKL blend fibers.

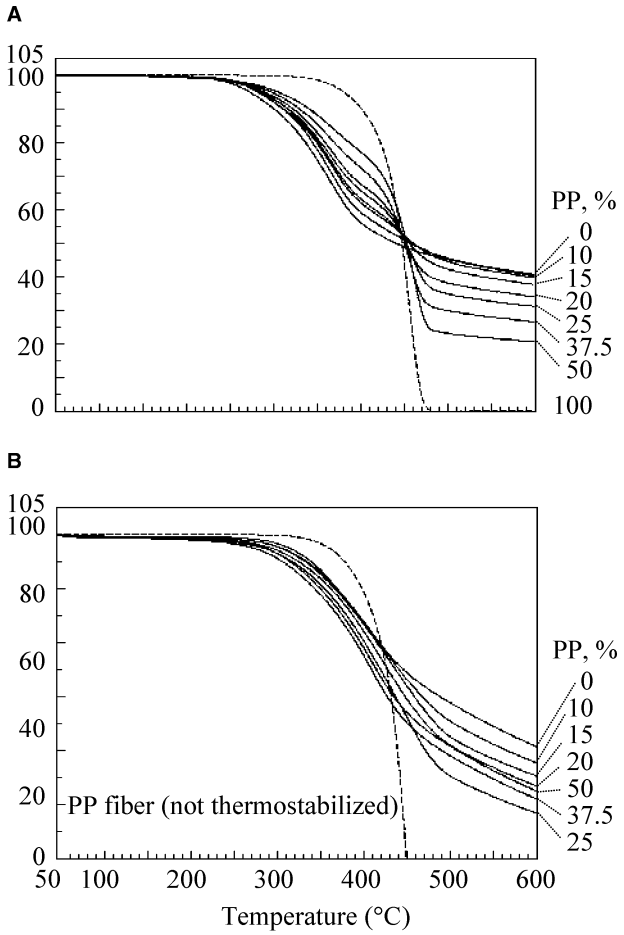


Fig. 10 TGA curves of thermostabilized HKL and all PP/HKL blend fibers. (A) HKL/PP blend fibers and (B) HKL/PP thermostabilized fibers.

As seen in Fig. 7, the fastest allowable heating rate for ground lignin is about $0.06^{\circ}C\,min^{-1}$. In the case of lignin fibers, diffusion becomes an issue in terms of overall conversion and T_g. Lignin fibers can be thermostabilized at a heating rate of $0.2^{\circ}C\,min^{-1}$ (Table 2), substantially higher than that required to maintain an increase in T_g. Not surprisingly T_g is not the only factor affecting fiber integrity. Other thermal properties, such as viscosity, will be very important. However, this result shows that thermostabilization is an essential process in lignin carbon fiber production.

The thermostabilization properties of the various lignin/synthetic polymers are included in Table 2. It can be seen that polymer blending with PET and PP affects the thermostabilization process differently. Significantly higher heating rates can be used in the thermostabilization of lignin/PET blend fibers than for lignin or lignin/PP fibers. (Both PET and PP homofibers cannot be thermostabilized.) This is attributed to the physical properties of the respective polymers. The melting temperature of PET (ca. 270°C) is

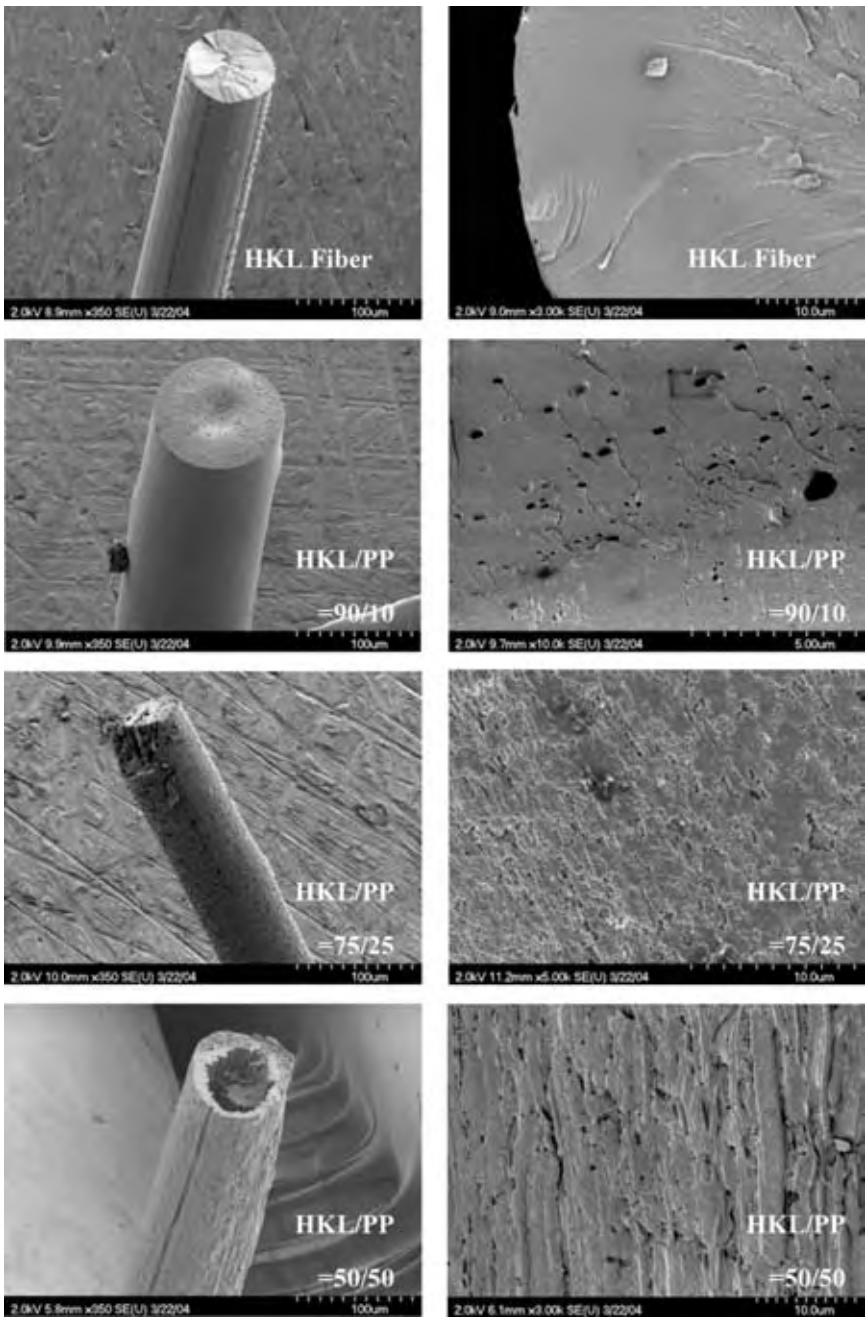


Fig. 11 SEM images of thermostabilized HKL and HKL/PP blend fibers.

higher than the softening point of lignin, even though the glass transition temperatures of both polymers are almost the same. Although the melting temperature of PET was decreased upon blending with lignin, it is still higher than the oxidation temperature. Moreover, the fiber spinning temperature of the HKL/PET blends is much higher than that of HKL. As discussed earlier, the higher spinning temperature is indicative of the higher thermal viscosity of the polyblend, the result of specific interactions between the components. By contrast, the HKL/PP blends show almost the same thermostabilization properties as the HKL fiber. The T_g and melting temperature of PP are lower than that of the HKL. Further, the fiber spinning temperature for PP is approximately the same as that of HKL. This, combined with the fact that the blends are immiscible and do not exhibit any specific interactions, makes it difficult to increase the heating rate of thermostabilization beyond that of the HKL fibers.

Polymer blending also affects the dimensional stability of the HKL-based fibers during the thermostabilization process. The HKL homofibers undergo more than 20% shrinkage during thermostabilization. This fiber deformation affects the utility and the

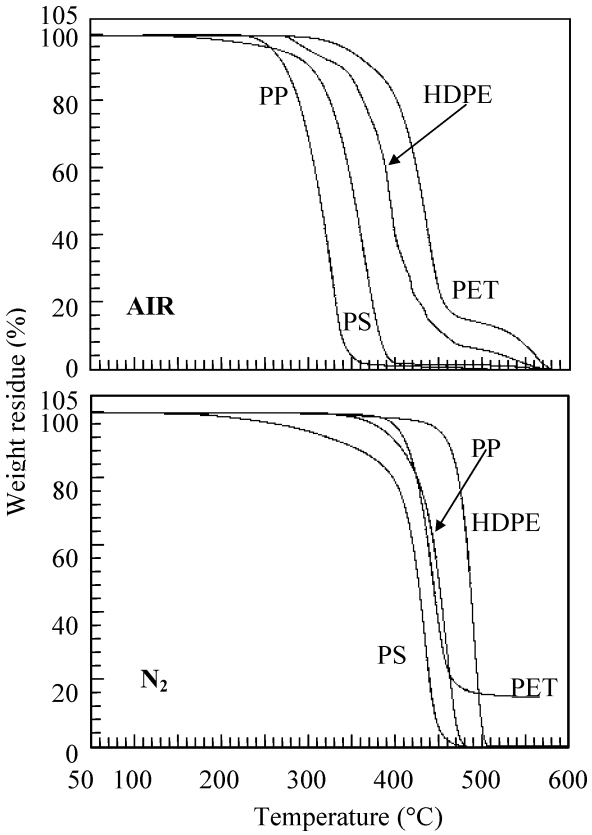


Fig. 12 TGA curves of various recyclable plastic measured under nitrogen and air conditions.

processibility of lignin for carbon fiber production; fiber shrinkage can destroy the fiber integrity. However, the dimensional stability of the lignin fibers can be improved through polymer blending with PET and PP. Fig. 8 shows the effect of blend components on fiber shrinkage. Increasing incorporation of both PP and, especially, PET effectively improves fiber stability, with fiber shrinkage decreasing with increasing polymer blending.

MORPHOLOGICAL CHANGES OF HKL/PP BLEND FIBERS IN THE THERMOSTABILIZATION PROCESS

Of the various lignin-based fibers, the HKL/PP fibers displayed the most dramatic change in morphology. DSC curves of the thermostabilized HKL and various PP/HKL blend fibers are shown in Fig. 9. The T_g of HKL was not observed, indicating consistent thermostabilization of the various fibers under these conditions. The melting peak of the PP fractions was also observed with the exception of HKL/PP = 50/50 and 62.5/37.5 blends, even though the original PP content of these fibers is higher than the other blend fibers. This suggests that there are no crystalline PP structures in these HKL/PP blend fibers. However, the T_g of the amorphous fraction of the PP component was also not observed. Thermogravimetric analysis (TGA) of the thermostabilized HKL/PP blends is shown in Fig. 10B. The TGA profile is different from the original HKL/PP blend fiber (Fig. 10A). The residual weight at 600°C for the thermostabilized HKL/PP = 50/50 and 62.5/37.5 blends is higher than that of the HKL/PP = 75/25 fiber. It appears that in the 25 wt% PP fibers sufficient lignin is lost in conjunction with the PP during thermal treatment. This is not the case in the fibers with high PP content, 50 wt%, and to a lesser extent, 37.5 wt%. It is likely that in the 25 wt% PP fibers, a sufficient amount of lignin is entrapped within the PP phase, which with increasing PP incorporation decreases due to an increase in the PP domain size. Further support of this can be seen from microscopy analysis of the fibers.

Table 3 Yields of HKL/synthetic polymer blend carbon fibers in each step of carbon fiber production

Precursor	Production step			
	HT	Thermostabilization	Carbonization	Overall
HKL 100%	99.0	94.4	51.6	34.3
HKL/PP				
95/5	99.1[a]	95.4	53.2	47.7[b]
87.5/12.5	99.1[a]	94.3	50.5	44.8[b]
75/25	99.3[a]	92.7	43.0	37.6[b]
50/50	99.5[a]	80.6	35.0	26.6[b]
HKL/PET				
95/5	99.1[a]	94.4	51.6	45.8[b]
87.5/12.5	99.1[a]	93.0	48.7	42.7[b]
75/25	99.3[a]	92.5	44.1	38.4[b]
Tar-pitch	40.0	110	80	33.4

[a]Calculated from yield for 100% HKL fiber.
[b]Yield of spinning process for blend sample is assumed to be 95%.

Scanning electron micrograms of the thermostabilized fibers are shown in Fig. 11. In contrast to the precursor fibers, there are small pores observed on the surface of the HKL thermostabilized fibers, but not in the cross section. By contrast, the thermostabilized blend fibers have pores of various sizes throughout the fiber. The thermostabilized HKL/PP = 75/25 fiber has fine pores on both the surface and the cross section. Interestingly, the thermostabilization of HKL/PP = 50/50 fibers has a porous surface, hollow core morphology. Comparison to the lignin fibers reveals that the pores are created during low temperature thermal processing. TGA analysis of various recyclable plastics measured under nitrogen and air atmospheres is shown in Fig. 12. The weight loss of PP in air is much faster than that under a nitrogen atmosphere. Thus, the pores observed in the HKL/PP thermostabilized fiber are the result of the oxidative degradation of the PP component during heat treatment in air.

CARBONIZATION

In addition to physical properties, an important aspect in carbon fiber production is yield. Higher weight losses or lower yield result in higher production costs. The carbon content of lignin is lower than that in petroleum and coal tar pitch. However, the total production yield of our HKL-based carbon fiber is higher (Table 3). In the case of pitch-based materials, several treatments, such as thermal pretreatments for polycondensation reactions and hydrogenation, are performed to convert the pitch-based material into suitable materials for fiber spinning. As a result of these processing steps, a critical loss in mass occurs. Due to its structure, lignin requires only a simple thermal treatment prior to the spinning process. The accompanying weight loss is very low ~1%. Unlike pitch, a small loss in yield is observed during the thermostabilization (ca. 3–7%) of lignin fibers. As lignin has a highly oxidized/functionalized structure, as described earlier, a net loss in weight/yield is observed during thermostabilization. In pitch, due to the unoxidized structure, a weight/yield increase occurs because of the introduction of oxidized functional groups into the pitch molecules. The majority of weight loss in lignin-based fibers occurs during the carbonization step. This weight loss is substantially higher than that observed for pitch (20%) due to the lower carbon content of lignin (~60%). However, the yields obtained from our kraft lignin-based carbon fibers are approximately 10% higher than those previously reported for the carbonization of organosolv lignins and comparable to those from phenolated steam-exploded lignins. The blending of PP and PET, due to the pyrolyzable properties of

Table 4 Typical mechanical properties of HKL and HKL/recyclable plastic carbon fiber

Precursor fiber	Tensile strength (MPa)	Young's modulus (GPa)
HKL	605	61
HKL/PP		
95/5	332	57
87.5/12.5	437	54
75/25	155	29
50/50	167	28
HKL/PET		
95/5	669	84
87.5/12.5	682	84
75/25	703	84
Isotropic pitch (GP grade carbon fiber)	720	32
Activated carbon fiber (pitch)	100–250	—

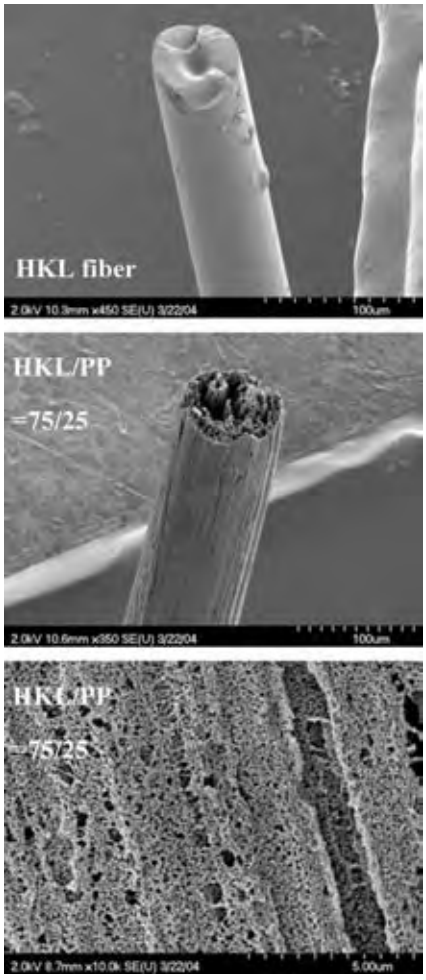


Fig. 13 SEM images of HKL and HKL/PP blend carbon fibers.

these polymers, decreased carbon fiber yield relative to lignin alone.

The mechanical properties of the carbon fibers are listed in Table 4. Tensile strength of the HKL/PET-based carbon fibers increased with PET content up to a maximum at 25% PET. Increasing the PET content beyond 25% leads to a decrease in mechanical properties, which is likely the result of off gases generated due to the pyrolysis of the PET component. The tensile strength and the Young's modulus of the HKL/PET = 75/25 carbon fibers are comparable and three times greater, respectively, than that of GP grade pitch-based carbon fibers. By contrast, the mechanical properties of the HKL/PP blend carbon fibers decreased with increasing PP content. As expected, various sized pores are observed on the carbon fiber surfaces, with pore size increasing with PP content.

Scanning electron micrograms of the HKL/PP carbon fibers are shown in Fig. 13. The pores observed on the HKL thermostabilized fiber decrease in size and/or disappear after carbonization. This indicates that a reorganization of the lignin molecules occurs during the carbonization process. The HKL/PP carbon fibers also show some differences as compared to the thermostabilized fibers. There are a lot of fine pores observed on the fiber surface and cross sections as compared to the corresponding thermostabilization fiber. From the DSC (Fig. 9) and TGA (Fig. 10) analysis of the thermostabilized HKL/PP fibers, it was observed that some PP component remains in the fiber after the oxidation process. Therefore, the small pores in the HKL/PP blend carbon fibers are likely due to the pyrolysis of the residual PP from the carbonized lignin framework. All of the various thermal treatments, pretreatment, thermostabilization, and carbonization, produce pores in the HKL/PP blend fibers. However, the mechanism(s) of pore growth are different between these processes. Pores are created by the oxidative degradation of PP during the thermostabilization process, while the gasification of pyrolyzable PP is the main factor for pore growths under carbonization.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Richard Gilbert of NC State University for his comments and contributions to this work. We would also like to thank the Microscope Lab of University of British Columbia for their help with SEM observations.

REFERENCES

1. Glasser, W.G.; Sarkanen, S. *Lignin: Properties and Materials*; American Chemical Society: Washington DC; 1989.
2. Li, Y.; Mlynar, J.; Sarkanen, S. The first 85% kraft lignin-based thermoplastics. J. Polym. Sci. Part B—Polym. Phys. **1997**, *35* (12), 1899–1910.
3. Glasser, W.G.; Northey, R.A.; Schultz, T.P. *Lignin: Historical, Biological, and Materials Perspectives*; American Chemical Society: Washington, DC, 2000; 559 pp.
4. Esselen, G.J.; Bacon, F.S. Raw materials of the plastics industry. Ind. Eng. Chem. **1938**, *30* (2), 125.
5. Glasser, W.G.; Barnett, C.A.; Rials, T.G.; Saraf, V.P. Engineering plastics from lignin, II. Characterization of hydroxyalkyl lignin derivatives. J. Appl. Polym. Sci. **1984**, *29* (5), 1815–1830.
6. Glasser, W.G.; Jain, R.K. Lignin derivatives, I. Alkanoates. Holzforschung **1993**, *47* (3), 225–233.
7. Feldman, D.; Banu, D.; Lacasse, M.; Wang, J.; Luchian, C. Lignin and its polyblends. J. Macromol. Sci.—Pure Appl. Chem. **1995**, *A32* (8–9), 1613–1619.
8. Kadla, J.F.; Kubo, S. Miscibility and hydrogen bonding in blends of poly(ethylene oxide) and kraft lignin. Macromolecules **2003**, *36* (20), 7803–7811.
9. Kubo, S.; Kadla, J.F. Poly(ethylene oxide)/organosolv lignin blends: relationship between thermal properties, chemical structure and blend behavior. Macromolecules **2004**, *37* (18), 6904–6911.
10. Riggs, J. *Encyclopedia of Polymer Science and Engineering*, 2nd Ed.; Wiley: New York, 1985; Vol. 2, 640–685.
11. Swan, J.W. British Patent 4933, 1880.
12. Edison, T.A. US Patent 223,398, 1880.
13. Donnet, J.B.; Bahl, O.P. *Encyclopedia of Physical Science and Technology*; Academic Press, 1987; Vol. 2, 515–529.
14. Donnet, J.-B. Carbon Fibers. Marcel Dekker, 1998.
15. Hatori, H.; Yamada, Y.; Shiraishi, M. Inplane orientation and graphitizability of polyimide films, II. Film thickness dependence. Carbon **1993**, *31* (8), 1307–1312.
16. Sliva, D.E.; Selley, W.G. Continuous Method for Making Spinnable Polyacetylene Solutions Convertable to High Strength Carbon Fibers. US Patent 3928516, 1975.
17. Krutchen, C.M. Melt Extrudable Polyacetylene Copolymer Blends. US Patent 3852235, 1974.
18. Kobayashi, R.W.; Zaldivar, R.D. Carborane Catalyzed Graphitization of Polyarylacetylene (PAA). US Patent 52,88438, 1994.
19. Nagasaki, A.; Ashitaka, H.; Kusuki, Y.; Oda, D.; Yoshinaga, T. Process for Producing Carbon Fiber. US Patent 4131644, 1978.
20. Horikiri, S.; Iseki, J.; Minobe, M. Process for Producing Carbon Fiber. US Patent 40,70446, 1978.
21. Araki, T.; Takita, H.; Asano, K. Verfahren zum herstellen von kolenstoffasern. German Patent 20,24,063, 1970.

22. Seo, I.; Oono, T.; Murakami, Y. Process for Producing Raw Material Pitch for Carbon Materials. European Patent 38,1493, 1990.
23. Shiokawa, M.; Matsumoto, T. Production of Pitch-based Carbon Fiber. Japanese Patent 12,82,349, 1989.
24. Ashitaka, H.; Kusuki, Y.; Yamamoto, S.; Ogata, Y.; Nagasaka, A. Preparation of carbon-fibers from syndiotactic 1,2-polybutadiene. J. Appl. Polym. Sci. **1984**, *29* (9), 2763–2776.
25. Otani, S.; Fukuoka, Y.; Igarashi, B.; Sasaki, K. Method for Producing Carbonized Lignin Fiber. US Patent 3,461,082, 1969.
26. Mansmann, M.; Winter, G.; Pampus, P.; Schnoring, H.; Schon, N. Stable Lignin Fibers. US Patent 3,723,609, 1973.
27. ''Kayacarbon'' *Manufacturer's Brochure*; Nippon Kayaku Co. Ltd..
28. Johnson, D.J.; Tomizuka, I.; Watanabe, O. Fine-structure of lignin-based carbon-fibers. Carbon **1975**, *13* (4), 321–325.
29. Johnson, D.J.; Tomizuka, I.; Watanabe, O. Fine-structure of pitch-based carbon-fibers. Carbon **1975**, *13* (6), 529–534.
30. Tomizuka, I.; Johnson, D.J. Microvoids in pitch-based and lignin-based carbon fibres as observed by x-ray small angle scattering. Yogyo-Kyokai-Shi **1978**, *86* (4), 186.
31. Tomizuka, I.; Kurita, T.; Tanaka, Y.; Watanabe, O. Voids in the carbon fibers produced from lignin and PVA. Yogyo-Kyokai-Shi **1971**, *79* (12), 460.
32. Sudo, K.; Shimizu, K. A new carbon-fiber from lignin. J. Appl. Polym. Sci. **1992**, *44* (1), 127–134.
33. Sudo, K.; Okoshi, M.; Shimizu, K. Carbon-fiber from lignin—improvement of conversion process of lignin. Abstr. Pap. Am. Chem. Soc. **1988**, *195*107–CELL.
34. Sudo, K.; Shimizu, K.; Nakashima, N.; Yokoyama, A. A new modification method of exploded lignin for the preparation of a carbon-fiber precursor. J. Appl. Polym. Sci. **1993**, *48* (8), 1485–1491.
35. Kubo, S.; Uraki, Y.; Sano, Y. Thermomechanical analysis of isolated lignins. Holzforschung **1996**, *50* (2), 150–156.
36. Kubo, S.; Ishikawa, N.; Uraki, Y.; Sano, Y. Preparation of lignin fibers from softwood acetic acid lignin—relationship between fusibility and the chemical structure of lignin. Mokuzai Gakkaishi **1997**, *43* (8), 655–662.
37. Kubo, S.; Uraki, Y.; Sano, Y. Preparation of carbon fibers from softwood lignin by atmospheric acetic acid pulping. Carbon **1998**, *36* (7–8), 1119–1124.
38. Uraki, Y.; Kubo, S.; Nigo, N.; Sano, Y.; Sasaya, T. Preparation of carbon-fibers from organosolv lignin obtained by aqueous acetic-acid pulping. Holzforschung **1995**, *49* (4), 343–350.
39. Itoh, K. Preparation of Lignin Fiber. Japan Patent H1239114, 1989.
40. Rouilly, A.; Rigal, L. Agro-materials: a bibliographic review. J. Macromol. Sci.—Polym. Rev. **2002**, *C42* (4), 441–479.
41. Wrzesniewska-Tosik, K.; Tomaszewski, W.; Struszczyk, H. Manufacturing and thermal properties of lignin-based resins. Fibres & Textiles in Eastern Europe **2001**, *9* (2), 50–53.
42. Yoshida, H.; Morck, R.; Kringstad, K.P.; Hatakeyama, H. Fractionation of kraft lignin by successive extraction with organic-solvents, II. Thermal-properties of kraft lignin fractions. Holzforschung **1987**, *41* (3), 171–176.
43. Kondo, R.; McCarthy, J.L. Condensation of lignins with coniferyl alcohol in alkaline aqueous solutions. J. Wood Chem. Technol. **1985**, *5* (1), 37–52.
44. Braun, J.; Holtman, K.; Kadla, J.F. Lignin-based carbon fibers: oxidative thermostabilization of kraft lignin. Carbon **2005**, *43*, 385–394.
45. Paul, D.R.; Bucknall, C.B. *Polymer Blends*; Wiley: New York, 2000.
46. Kadla, J.F.; Kubo, S. Lignin-based polymer blends: analysis of intermolecular interactions in lignin–synthetic polymer blends. Composites, Part A: Appl. Sci. Manuf. **2004**, *35* (3), 395–400.
47. Kubo, S.; Kadla, J.F. The formation of strong intermolecular interactions in immiscible blends of poly(vinyl alcohol) (PVA) and lignin. Biomacromolecules **2003**, *4* (3), 561–567.
48. Kadla, J.F.; Gilbert, R.D.; Venditti, R.A.; Kubo, S. Porous Fiber Manufacture from Natural/Synthetic Polymer Blends. US Patent 2003212157, 2003.
49. Kadla, J.F.; Kubo, S.; Venditti, R.A.; Gilbert, R.D. Novel hollow core fibers prepared from lignin polypropylene blends. J. App. Polym. Sci. **2002**, *85* (6), 1353–1355.
50. Wisanrakkit, G.; Gillham, J.K. Continuous heating transformation (CHT) cure diagram of an aromatic amine epoxy system at constant heating rates. J. Appl. Polym. Sci. **1991**, *42* (9), 2453–2463.

Carbon Nanotubes

Morinobu Endo
Yoong-Ahm Kim
Takuya Hayashi
Kenji Takeuchi
Faculty of Engineering, Shinshu University, Wakasato, Nagano-shi, Japan

Maruicio Terrones
Advanced Materials Department, IPICyT, San Luis Potosí, SLP, México

Mildred S. Dresselhaus
Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.

INTRODUCTION

Carbon nanotubes (CNTs) are tubular sp^2-like carbon molecules with nanoscale dimensions that exhibit fascinating physicochemical properties that make them potentially useful in the fabrication of novel nanotechnology devices. Carbon nanotubes consist of rolled graphene (a hexagonal sp^2 carbon layer) cylinders of nanometer size in diameter; they could be either single- or multiwalled (nested cylinders).[1–3] The diameters of such tubules could be in the Fullerene size range (around 1 nm), and the minimum diameter that has been observed is 0.4 nm.[4,5] Multiwalled carbon nanotubes (MWNTs) are strongly related to three-dimensional graphite or turbostratic hexagonal carbon (see Fig. 1). The nanotube curvature introduces some sp^3-like bonding into the sp^2 planar bonding of the graphene sheet. Unusual properties of CNTs arise from the curved sp^2 graphene layers by imposing additional quantum confinement and topological constraints in the circumferential direction of the cylinders. In the following sections, their structure, growth, and applications of CNTs are presented. We will also discuss the synthesis, properties, and applications of another family of tubular carbon, known as vapor-grown carbon fibers (VGCFs).

BONDING NATURE OF CARBON ATOMS

Carbon is the sixth element in the periodic table, and has the lowest atomic number of any of the elements in column IV. Each carbon atom has six electrons, which occupy $1s^2$, $2s^2$, and $2p^2$ atomic orbitals. The $1s^2$ orbital contains two strongly bound core electrons. Four more weakly bound electrons occupy the $2s^22p^2$ valence orbitals. In the crystalline phase, the valence electrons give rise to 2s, $2p_x$, $2p_y$, and $2p_z$ orbitals, which are important in forming covalent bonds within carbon materials. Since the energy difference between the upper 2p energy levels and the lower 2s level in carbon is small compared to the binding energy of the chemical bonds, the electronic wave functions for these four electrons can readily mix with each other, thereby changing the occupation of the 2s and three 2p atomic orbitals, so as to enhance the binding energy of a C atom with its neighboring atoms. The general mixing of 2s and 2p atomic orbitals is called hybridization, whereas the mixing of a single 2s electron with one, two, or three 2p electrons is called sp^n hybridization with $n = 1, 2, 3$, respectively.

Thus, three possible hybridizations could occur in carbon: sp, sp^2, and sp^3, while other group IV elements, such as Si and Ge, exhibit primarily sp^3 hybridization. Carbon differs from Si and Ge because it has no inner atomic orbitals, except for the spherical 1s orbitals, and the absence of nearby inner orbitals facilitates hybridizations involving only valence s and p orbitals for carbon. The various bonding states are connected with certain structural arrangements, so that sp bonding gives rise to a chain structure such as carbynes, sp^2 bonding to planar structures such as graphene, and sp^3 bonding to a tetrahedral structure such as diamond.

A CNT is a graphene sheet appropriately rolled into a cylinder of nanometer size diameter (see Fig. 1).[1–3] The curvature of the nanotubes admixes a small amount of sp^3 bonding so that the force constant (bonding) in the circumferential direction is slightly weaker than along the nanotube axis. Since the single-walled CNT (SWNT) is only one atom thick and has a small number of atoms around its circumference, only a few wave vectors are needed to describe the periodicity of the nanotubes. These constraints lead to quantum confinement of the wavefunctions in the radial and circumferential directions, with plane wave motion occurring only along the nanotube axis, corresponding to a large number of closely spaced allowed wave vectors. Thus, although CNTs are related to a two-dimensional (2D) graphene

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120025012

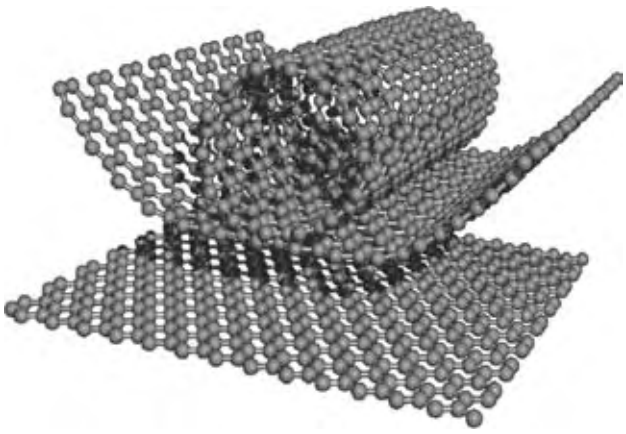


Fig. 1 Schematic diagram of an individual carbon layer in the honeycomb graphite lattice, called a graphene layer, and how it can be rolled to form a CNT. *(View this art in color at www.dekker.com.)*

sheet, the tube curvature and the quantum confinement in the circumferential direction lead to 1D properties that are different from those of a graphene sheet.

STRUCTURE OF CARBON NANOTUBES

The structure of CNTs has been explored early on by high-resolution transmission electron microscopy (HRTEM) and scanning tunneling microscopy (STM) techniques, yielding direct confirmation that the nanotubes are seamless cylinders derived from the honeycomb lattice representing a single atomic layer of crystalline graphite, called a graphene sheet (Fig. 2A).[1,6–8] The structure of a single-wall CNT is conveniently explained in terms of its 1D unit cell, defined by the vectors $\boldsymbol{C_h}$ and $\boldsymbol{T}$ in Fig. 2A.

The circumference of any CNT is expressed in terms of the chiral vector $\boldsymbol{C}_h = n\boldsymbol{a}_1 + m\boldsymbol{a}_2$, which connects two crystallographically equivalent sites on a 2D graphene sheet (Fig. 2A).[9] The construction in Fig. 2A depends uniquely on the pair of integers (n, m) that specify the chiral vector. Fig. 2A shows the chiral angle θ between the chiral vector $\boldsymbol{C_h}$ and the "zigzag" direction ($\theta = 0$) and the unit vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ of the hexagonal honeycomb lattice of the graphene sheet. The zigzag and armchair nanotubes, respectively, correspond to chiral angle $\theta = 0$ and 30°, and chiral nanotubes correspond to $0 < \theta < 30°$. The intersection of the vector $\overrightarrow{OB}$ (which is normal to $\boldsymbol{C_h}$) with the first lattice point determines the fundamental 1D translation vector $\boldsymbol{T}$. The unit cell of the 1D lattice is the rectangle defined by the vectors $\boldsymbol{C_h}$ and $\boldsymbol{T}$ (Fig. 2A).

The cylinder connecting the two hemispherical caps of the CNT (Fig. 3) is formed by superimposing the

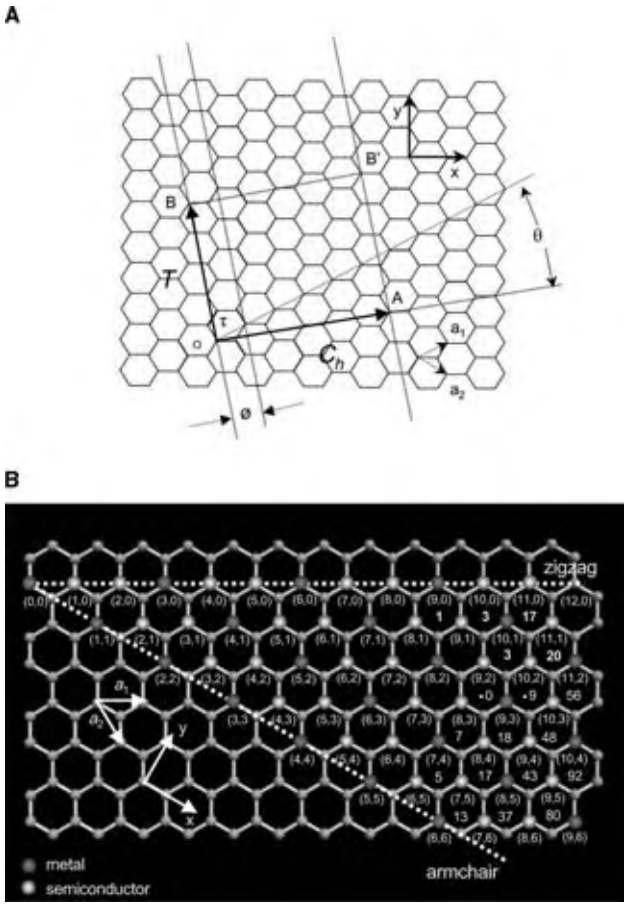


Fig. 2 (A) The chiral vector $\overrightarrow{OA}$ or $\boldsymbol{C_h} = n\mathbf{a}_1 + m\mathbf{a}_2$ is defined on the honeycomb lattice of carbon atoms by unit vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ and the chiral angle θ with respect to the zigzag axis. Along the zigzag axis $\theta = 0°$. Also shown are the lattice vector $\overrightarrow{OB} = \boldsymbol{T}$ of the 1D nanotube unit cell and the rotation angle ψ and the translation τ, which constitute the basic symmetry operation $R = (\psi/\tau)$ for the CNT. The diagram is constructed for $(n, m) = (4, 2)$. (From Ref.[9].) (B) Possible vectors specified by the pairs of integers (n, m) for general CNTs, including zigzag, armchair, and chiral nanotubes. Below each pair of integers (n, m) is listed the number of distinct caps that can be joined continuously to the CNT denoted by (n, m). (From Ref.[9].) The encircled dots denote metallic nanotubes while the small dots are for semiconducting nanotubes. (From Ref.[3].). *(View this art in color at www.dekker.com.)*

two ends of the vector $\overrightarrow{\boldsymbol{C_h}}$ and the cylinder joint is made along the two lines $\overrightarrow{OB}$ and $\overrightarrow{AB'}$ in Fig. 2A. The lines $\overrightarrow{OB}$ and $\overrightarrow{AB'}$ are both perpendicular to the vector $\boldsymbol{C_h}$ at each end of $\boldsymbol{C_h}$.[9] In the (n, m) notation for $\boldsymbol{C_h} = n\mathbf{a}_1 + m\mathbf{a}_2$, the vectors $(n, 0)$ or $(0, m)$ denote zigzag nanotubes and the vectors (n, n) denote armchair nanotubes. All other vectors (n, m) correspond to chiral nanotubes.[10] The nanotube diameter d_t is given by

$$d_t = \sqrt{3}a_{C-C}(m^2 + mn + n^2)^{1/2}/\pi = \boldsymbol{C_h}/\pi \quad (1)$$

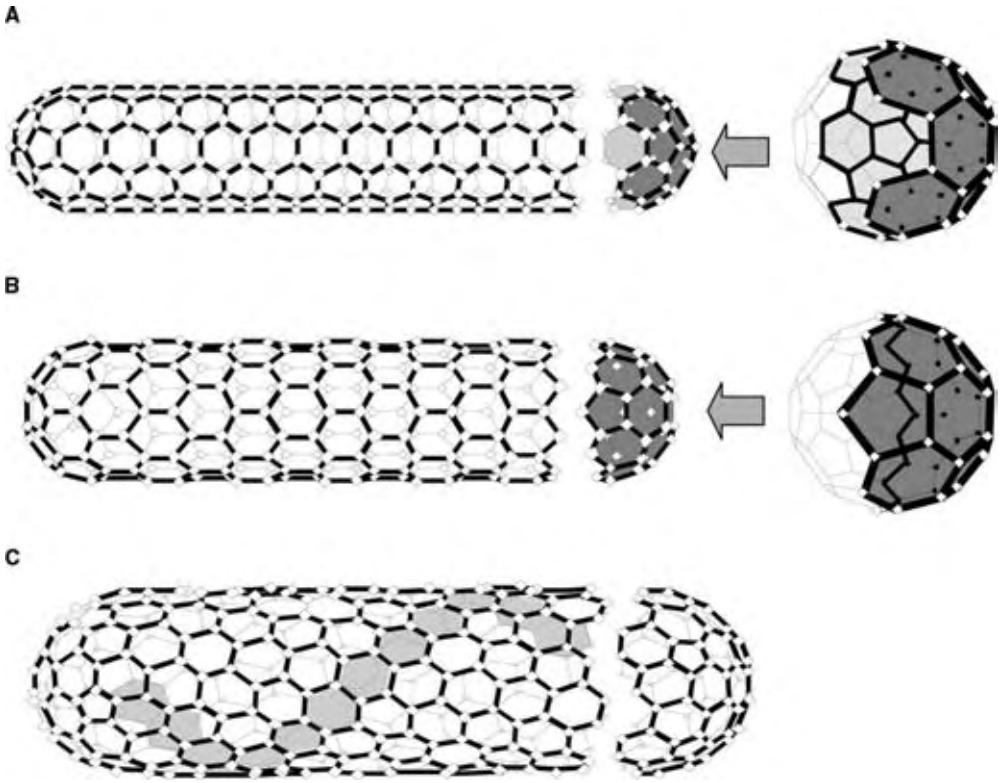


Fig. 3 Schematic model for single-wall CNTs with the nanotube axis normal to the chiral vector, which, in turn, is along (A) the $\theta = 30^\circ$ direction [an "armchair" (n, n) nanotube], (B) the $\theta = 0^\circ$ direction [a "zigzag" $(n, 0)$ nanotube], and (C) a general θ direction, such as $\overrightarrow{OB}$ (see Fig. 2), with $0 < \theta < 30^\circ$ [a "chiral" (n, m) nanotube]. The actual nanotubes shown here correspond to (n, m) values of: (A) (5, 5), (B) (9, 0), and (C) (10, 5). The nanotube axis for the (5, 5) nanotube has fivefold rotation symmetry, while that for the (9, 0) nanotube has threefold rotation symmetry. *(View this art in color at www.dekker.com.)*

where $\boldsymbol{C_h}$ is the length of $\boldsymbol{C_h}$, a_{C-C} is the C–C bond length (1.42 Å). The chiral angle θ is given by

$$\theta = \tan^{-1}\left[\sqrt{3n/(2m + n)}\right] \tag{2}$$

From Eq. (2) it follows that $\theta = 30^\circ$ for the (n, n) armchair nanotube and that the $(n, 0)$ zigzag nanotube would have $\theta = 60^\circ$. From Fig. 2A it follows that if we limit θ to between 0° and 30°, then by symmetry $\theta = 0^\circ$ for a zigzag nanotube. Both armchair and zigzag nanotubes have a mirror plane and thus are considered as chiral. Differences in the nanotube diameter d_t and chiral angle θ give rise to differences in the properties of the various CNTs. The symmetry vector $R = (\psi/\tau)$ of the symmetry group for the nanotubes is indicated in Fig. 2A, where both the translation unit τ (or pitch) and the rotational angle ψ are shown. The number of hexagons, N, per unit cell of a chiral nanotube, specified by integers (n, m), is given by

$$N = 2\left(m^2 + n^2 + nm\right)/d_R \tag{3}$$

where $d_R = d$ if $n - m$ is not a multiple of $3d$ or $d_R = 3d$ if $n - m$ is a multiple of $3d$ and d is defined as the largest common divisor of (n, m). Each hexagon in the honeycomb lattice (Fig. 2A) contains two carbon atoms. The unit cell area of the CNT is N times larger than that for a graphene layer and consequently the unit cell area for the nanotube in reciprocal space is correspondingly $1/N$ times smaller. Fig. 2B indicates the nanotubes that are semiconducting and those that are metallic, and shows the number of distinct fullerene caps that can be used to close the ends of an (n, m) nanotube, such that the fullerene cap satisfies the isolated pentagon rule.

On the other hand, in MWNTs, generally no graphitic 3D AB or ABC stacking is established, even though an individual shell of the multilayers consists of perfect graphene sheets. Also, each individual tubule forming the MWNT has different and independent chirality, which might contribute to a larger intershell spacing (e.g., 0.342 nm) than that found in graphite (e.g., 0.335 nm). Multiwalled carbon nanotubes relate closely to the layer stacking of graphite because the interlayer spacing is found to vary from 3.4 to 3.6 Å between adjacent cylinders, possibly due to glide defects and

tube curvature.[11] Magnetoresistance measurements provide a highly sensitive measure of the degree of 3D crystalline order using the following equation:

$$\Delta\rho/\rho_0 = [\rho(B) - \rho(0)]/\rho(0) \quad (4)$$

where $\rho(B)$ and $\rho(0)$ are the electrical resistivity with and without a magnetic field B, respectively. As shown in Fig. 4, highly oriented pyrolytic graphite with the highest degree of structural 3D perfection shows large positive transverse magnetoresistance, with smaller values of positive $\Delta\rho/\rho_0$ [mesophase-pitch carbon fibers (MPCFs) at 3000 °C] corresponding to more disordered graphite.[12] As more disorder is introduced, the 3D ordering disappears and the materials assume a 2D or turbostratic ordering, where there is no interlayer site correlation between adjacent graphene planes. The magnetoresistance for the turbostratic carbon is characteristically negative, with the magnitude of the negative magnetoresistance decreasing as the disorder increases. For the case of VGCFs heat treated at 2000 °C and MWNTs, the negative magnetoresistance strongly suggests that the stacking of the shells is quite similar to that of turbostratic graphite regarding their electronic properties along the tube axis (see the dotted curve in Fig. 4).[12,13] The curvature of the coaxial layers, however, excludes the possibility of the AB or ABC stacking of the graphene layers, and in this sense, the stacking of layers in an MWNT is equivalent to turbostratic stacking. Also, for small-diameter CNTs, there are a limited number of possible nanotube diameters, and these depend also on the nanotube chirality, so that it is not, in general, possible to find a nested set of nanotubes that all have the same chirality and also show a close packing of cylindrical shells. Thus, the electronic properties of MWNTs have some differences and some similarities when compared with turbostratic carbons. These characteristic structures of single- and multiwalled CNTs indicate that they are unique 1D materials with fascinating electronic and mechanical properties.

CARBON NANOTUBES DERIVED FROM VGCFs

Carbon fibers were developed in the early 1960s, and have provided many breakthroughs to carbon science due to their unique fibrous morphology. Therefore, it is necessary to compare the structure and properties of CNTs with conventional carbon fibers exhibiting diameters around 10 μm. Furthermore, there are also very important materials that can be related to both CNTs and carbon fibers. These are called the VGCFs synthesized directly by the catalytic decomposition of hydrocarbons.[6,14–16] Vapor-grown carbon fibers have diameters ranging from 1 nm to 10 μm, the thinner ones corresponding to CNTs and the thicker ones similar to carbon fibers. It has been reported that the

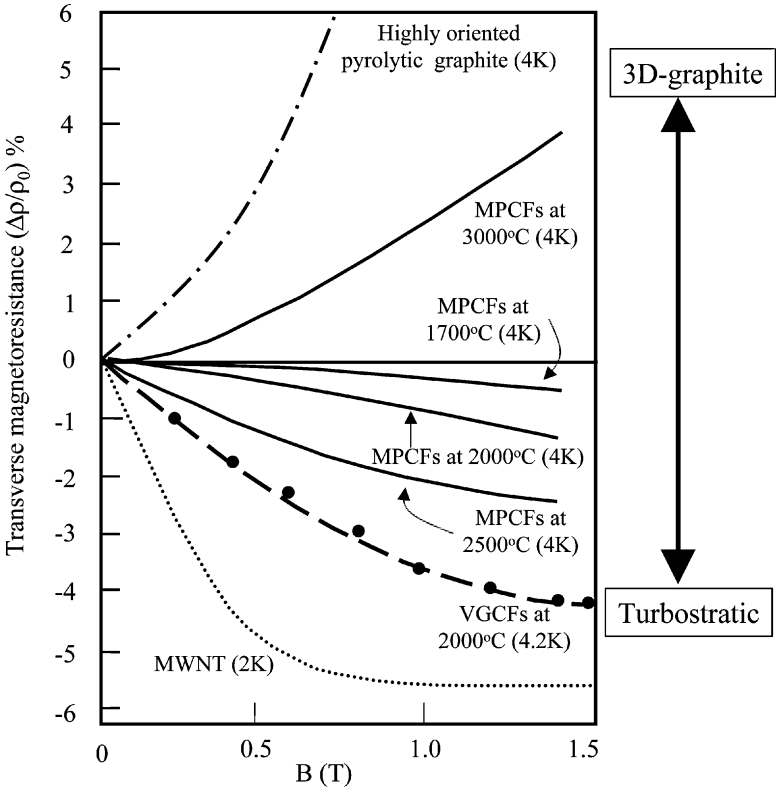


Fig. 4 Comparative study of transverse magnetoresistance at low temperature (as indicated in parentheses) for highly oriented pyrolytic graphite (the dot–dash curve), typical MPCFs heat treated at different temperatures (the solid curves), VGCFs (the dashed curve), and MWNTs (the dotted curve). Vapor-grown carbon fibers exhibit a similar magnetoresistance behavior to that of MWNTs.

central core of a VGCF consists of a CNT.[17] Through precise control of the synthesis conditions, it is possible to produce CNTs and carbon fibers on a large scale. Both materials have been industrialized for different practical applications. Therefore, the structures of CNTs and carbon fibers as well as VGCFs, in which the latter connects both typical materials, are strongly emphasized in this account. Vapor-grown carbon fibers have a very special structure like annular-rings (see Fig. 5A) and are synthesized by a somewhat different formation process than that used to produce polyacryonitrile-based carbon fibers and MPCFs. In particular, VGCFs are not prepared from a fibrous precursor, but rather from hydrocarbon gas, using a catalytic growth process outlined in Fig. 5B.[6,14–17] Ultrafine transition metal particles, such as iron particles with diameter less than 10 nm, are dispersed on a ceramic substrate, and a hydrocarbon, such as benzene diluted with hydrogen gas, is introduced at temperatures of about 1100 °C. Hydrocarbon decomposition takes place on the catalytic particle, leading to a continuous carbon uptake by the catalytic particle and a continuous output by the particle of well-organized tubular filaments of hexagonal sp^2-carbon. The rapid growth rate of several tens of micrometers per minute, which is 10^6 times faster than that observed for the growth of common metal whiskers, allows the production of commercially viable quantities of VGCFs. Evidence in support of this growth model is the presence of catalytic particles at the tips of the resulting VGCFs (Fig. 5C).[18] The primary thin hollow tube is first formed by the catalytic process (with a diameter of less than several nanometers), and the tube is then thickened by a successive CVD (chemical vapor deposition) process, corresponding to the deposition of pyrolytic carbon layers on the primary tubular core (Fig. 5D).

Such a high growth rate of the primary core tube provides the possibility of employing a 3D growth process involving a floating catalyst in a reaction chamber without the presence of a solid substrate (see Fig. 6A). In this process, the nanometer-sized catalytic particles of metal can float (or be suspended) for a specified time in the hot zone of the reaction chamber in order to produce thin (80 nm homogenous diameter) nanotubes of several hundred micrometers in length.[18,19] The resulting nanotubes shown in Fig. 6B consist of straight fibers with high purity, implying that the individual tubes experienced uniform reaction conditions as they passed through the reactor.[19] By proper choice of growth conditions, large quantities of high-purity nanotubes can be obtained continuously. Furthermore,

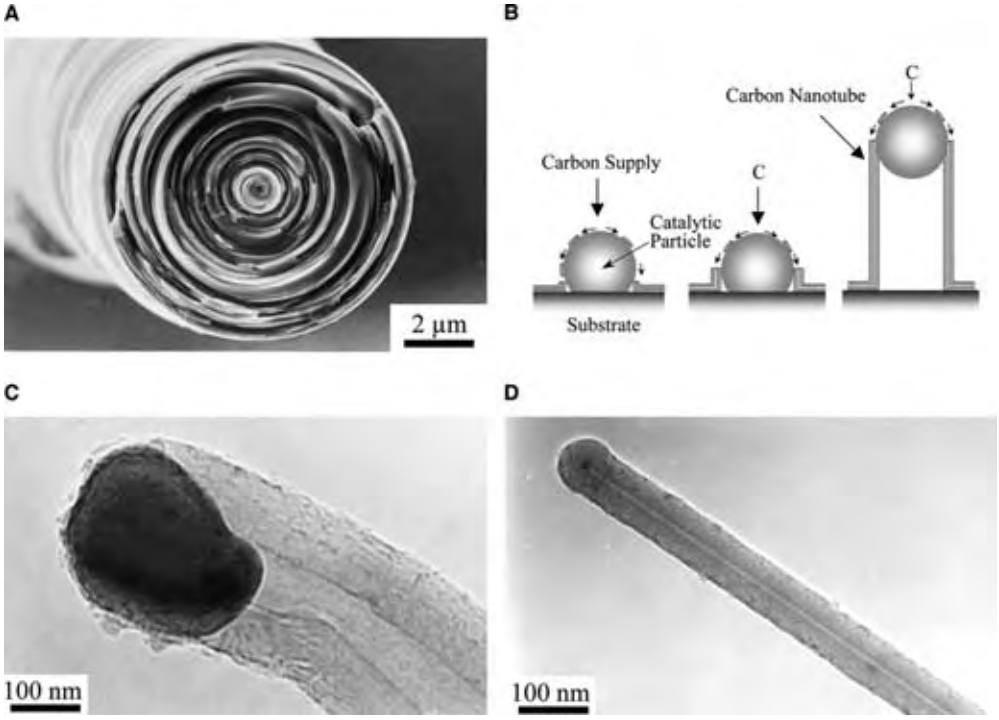


Fig. 5 (A) Scanning electron microscopic image of VGCFs. (From Ref.[19].) (B) Suggested growth mechanism of VGCFs using ultrafine catalytic metal particles. (From Ref.[6].) (C) Very early stage of fiber growth in which the catalytic-particle is still active for promoting elongation growth. The primary fiber thus formed acts as a core for VGCFs. (From Ref.[19].) (D) The fiber is obtained through a thickening process, such as the pyrolytic deposition of carbon layers on the primary thin fibers. (From Ref.[20].) The encapsulated catalytic particle can be seen at the tip of the hollow core. *(View this art in color at www.dekker.com.)*

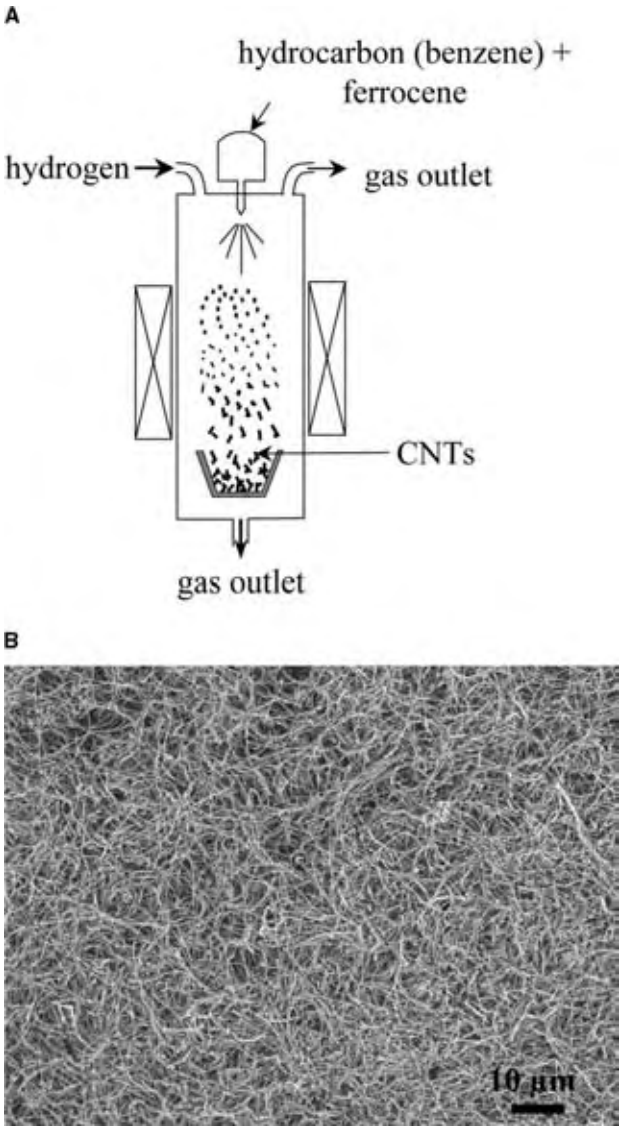


Fig. 6 (A) A vertical-type fiber synthesis system for the floating reactant method in which benzene vapor containing an organic–metallic compound such as ferrocene is introduced into a vertical-type reactor by hydrogen gas. The fibers grow while the catalytic particles are floating in the growth region of the reactor. (B) Vapor-grown carbon fibers, obtained under controlled growth conditions, exhibit relatively homogenous diameter along the fiber length and also high purity. (From Ref.[18].)

the tube diameter can be varied through control of the residence time of the tube in the reaction zone and of the pressure of the hydrocarbon feedstock. This basic process could be used for the large-scale production of CNTs by a CVD process based on the catalytic growth.[16]

In Fig. 7A, we observe the broken edge of a thick VGCF. At the center of the annular ring structure, we can clearly observe an extruded CNT (with a diameter of ~5 nm), which serves as a template for growing the

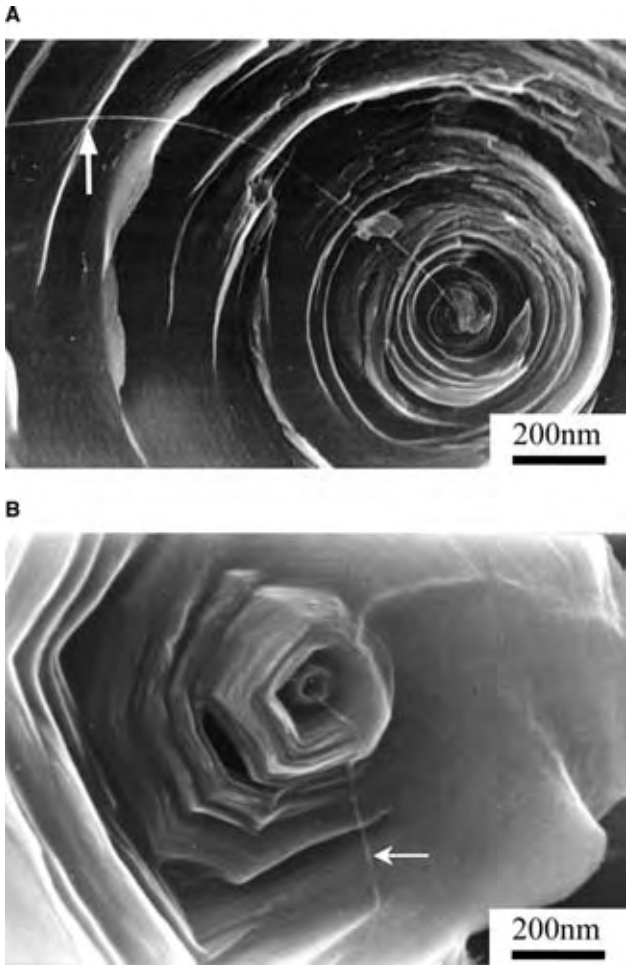


Fig. 7 (A) Carbon nanotube exposed at the breakage edge of a VGCF as (A) grown and (B) heat-treated at 3000 °C. The sample is fractured by pulverization and the core diameter is ~50 Å. (B) These photos suggest a structural discontinuity between the nanotube core of the fiber and the CVD deposited carbon layers. The images show the strong mechanical properties of the nanotube core, which maintains its form after breakage of the periphery. (From Ref.[20].)

thicker carbon fiber.[17] This tubular core implies a discontinuity in the structure between the core and the thickened periphery of the carbon fiber. Such a discontinuity in structure is retained even after heat treatment at temperatures as high as 3000 °C. This graphitization process introduces a fully developed graphite structure in the peripheral region of a VGCF exhibiting a polygonal shape (see Fig. 7B). Fig. 8A shows the early stage of fiber growth of a thin tubule corresponding to an SWNT, which has both a bare region and a region partially coated with pyrolytically deposited carbon.[20] Sometimes it is possible to observe an isolated double-walled carbon nanotube (DWNT) without a pyrolytic carbon coat at the early stage of formation (Fig. 8B) [6] This type of nanotube can be produced by the same process as VGCFs by carefully

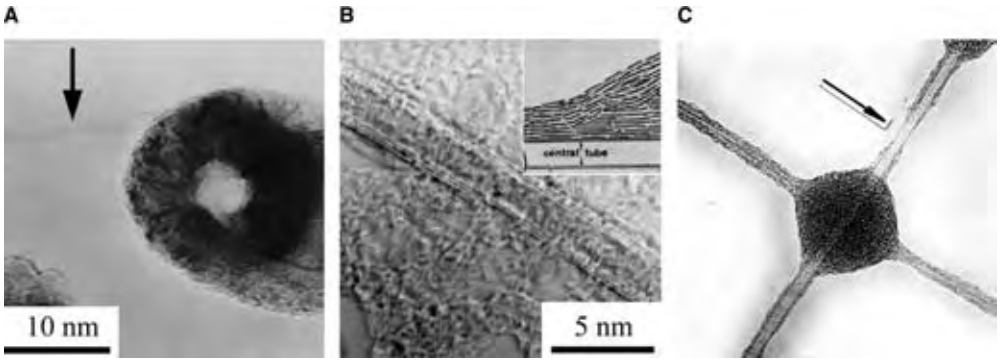


Fig. 8 (A) Coexistence of a VGCF and an SWNT (with a diameter of about 20 nm) obtained by the catalytic decomposition of benzene. (From Ref.[20].) The deposition of a partial carbon layer on a carbon nanotube during the thickening process is observed. (B) Double-walled carbon nanotube (obtained by benzene decomposition) and subsequently heat treated at 2800 °C, yielding the same structure as nanotubes prepared by the arc method. (From Ref.[6].) Insert is a schematic diagram of DWNTs. (From Ref.[6].) (C) High-resolution transmission electron microscope image of two crossing SWNTs coated with amorphous carbons indicates that the structure consists of an individual graphene cylinder in projection. (From Ref.[6].)

controlling the vapor pressure of the hydrocarbon and using a much smaller size of catalytic particle.[6] Fig. 8(C) exhibits an HRTEM image of crossing single-wall carbon nanotube coated with amorphous carbons.To my knowledge, this report in 1976 imaged the first ever observed SWNTs and MWNTs, which were tubular graphite of nanometer scale.[6]

SYNTHESIS OF CARBON NANOTUBES

Up to now, various synthetic methods for producing carbon nanotubes have been reported. Among them, we consider the three main synthetic techniques: 1) arc discharge of graphite electrodes; 2) laser vaporization of graphite–metal targets; and 3) thermolysis of hydrocarbons over metal catalysts. The synthesis of MWNTs via arc discharge involves the passage of a direct current through two high-purity graphite electrodes in a He atmosphere. Under this process, some carbon atoms, which are evaporated from the anode, condense on the cathode (negative electrode) forming a deposit. In addition to MWNTs, SWNTs can be produced using the arc-discharge method if a transition metal catalyst is added in the process. The production of SWNTs using this technique was reported almost simultaneously by two research groups.[21,22] Nowadays, SWNTs can be produced using the carbon arc method in conjunction with different metals such as Gd, Co–Pt, Co–Ru, Co, Ni–Y, Rh–Pt, and Co–Ni–Fe–Ce. For Ni–Y–graphite mixtures, Journet et al. found that high yields (<90%) of SWNTs (average diameter of 1.4 nm) can be produced.[23] This Ni–Y mixture is now used worldwide for production of SWNTs in high yield. The arc method usually involves high-purity graphite electrodes, metal powders, and high-purity He and Ar gases: thus, the costs associated with the production of SWNTs and MWNTs are high. Although the crystallinity of the material is also high, there is no control over the length and diameter of nanotubes. Also, by-products such as polyhedral graphite particles, encapsulated metal particles, and amorphous carbon are also formed.

In 1996, the discovery by the Rice group of a synthetic method involving laser vaporization of a graphite target that leads to high-quality single-wall nanotubes had a great impact on nanotube science.[24] The technique involves the laser evaporation of a target composed of Co–Ni/graphite at 1200 °C. The x-ray diffraction and TEM showed that the bundles (ropes) mainly contained (10, 10) nanotubes (13.8 Å diameter) packed in a crystalline form, and also consist of 100–500 SWNTs packed in a 2D triangular constant lattice of 17 Å. This technique has the advantage of producing high (above 70%) yields of SWNTs, but not economical advantage due to the presence of high-purity graphite rods, high-power lasers, and a relatively low production yield.

Synthesis of CNTs by a catalytic CVD method involves the catalytic deposition of hydrocarbons (e.g., benzene, acetylene, ethylene, etc.) over the surface of metal particles (e.g., Co, Ni, Fe, and Pd on substrate).[6,16] By simultaneously feeding hydrocarbons and catalytic particles in the gas phase into the reaction chamber (microsprayer or evaporation), MWNTs were synthesized on a large scale. On top of that, this type of tube was commercialized as the conductive filler in anode material of lithium-ion secondary battery.[25] Recently, the trend is to synthesize CNTs using a catalytic CVD method because this technique is very promising for the large-scale production of both SWNTs and MWNTs, especially using a floating reactant technique.[16] For SWNTs, the development of high-pressure carbon monoxide (HiPCO) process

(9 kg/day) gave impetus on the scientific study and applications of SWNTs.[26] In terms of the manufacturing process, this process has been shown to be more controllable and cost efficient as compared to either the arc discharge or the laser vaporization methods. Through exact control of synthetic conditions, it is possible to synthesize DWNTs selectively by using this technique. For the fabrication of a CNT transistor, this synthesis technique appears to be highly efficient due to the reproducibility and low cost when compared to other routes.

Recently, DWNTs, which consist of two concentric cylindrical shells, have attracted the attention of numerous scientists because it is believed that these tubes are more thermally and chemically stable and mechanically strong when compared to single-walled carbon nanotubes (SWNTs) (see Figs. 9A and 9C). In addition, as with SWNTs, these double-layered tubes should behave as quantum wires due to its narrow diameters (e.g., <2 nm). Through the right combination of judiciously selected catalyst metals using the seeding method (called a catalytic CVD method), which has been considered as a powerful tool for mass production of nanotubes, and the optimizied purfication method, highly purified DWNTs with a narrow diameter distribution, exhibiting a hexagonally packed array, were obtained as shown in Fig. 9B.[6,16] It is well known that the radial breathing mode frequency is inversely related to tube diameter. For as-grown DWNTs (Fig. 9D), Raman peaks appear above 250 cm^{-1} (corresponding to the inner shells of DWNTs), and below 250 cm^{-1} (usually associated with the outer shells of DWNTs). We envisage this material to be useful in the fabrication of novel sensors, nanocomposites, field emission sources, nanobearings, nanotube bi-cables, and electronic devices.

THE SMALLEST FREESTANDING SWNTs

The demand for small-diameter CNTs is important for taking advantage of the size effect that is predicted to lead to properties different from those of larger nanotubes. The SWNTs were produced by means of a nanozeolite floating reactant method. By combining improved versions of both the floating reactant method

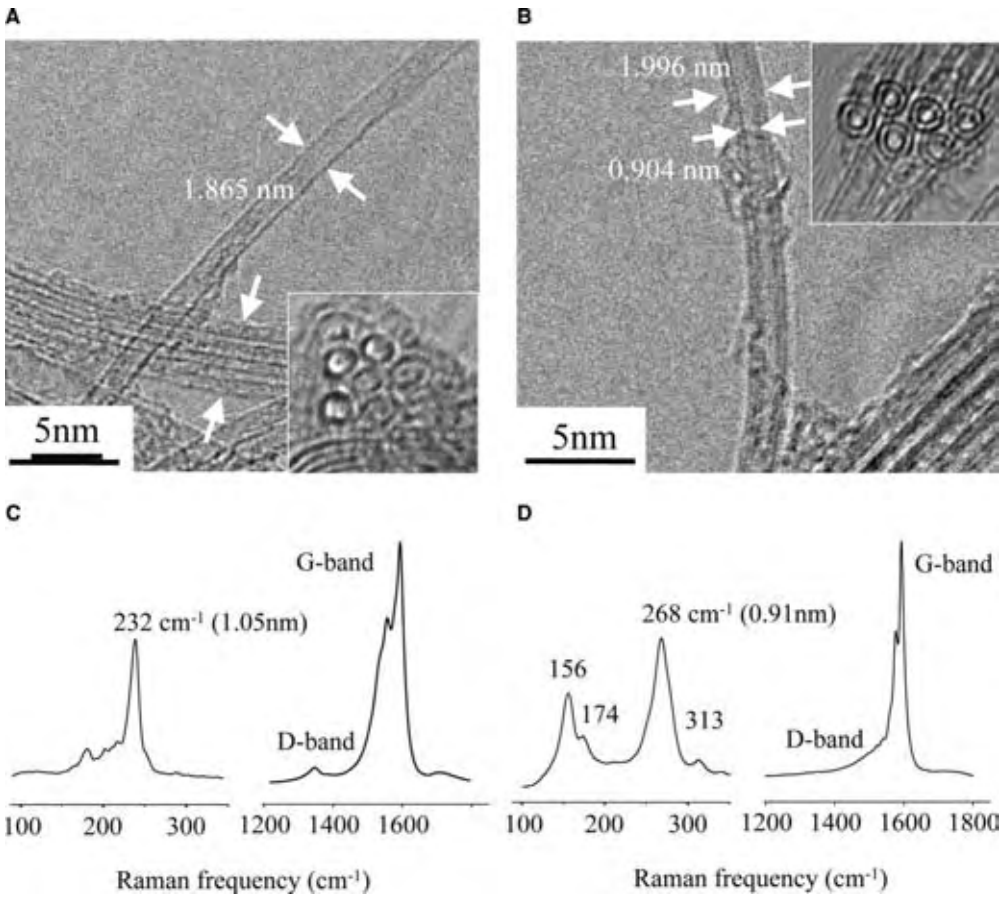


Fig. 9 (A) Typical HRTEM image of SWNT (insert is a cross-sectional TEM image). (B) Typical HRTEM image of DWNTs (insert is a cross-sectional TEM image). (C) Raman spectra of SWNT. (D) Raman spectra of DWNTs. Note that the bent nanotube is passed through the image [see inserts in (A) and (B)].

and the zeolite template method, a very attractive method for the mass production of thin SWNTs was suggested.

Transmission electron microscopic images of these SWNTs are very different from those shown in previous publications because in the latter SWNTs are not embedded on an amorphous film that disturbs the observation, nor is the tube located inside the central core part of an MWNT. Another important point is that the small-diameter SWNTs were produced in an unconfined system unlike those reported previously, using either a zeolite or an MWNT as templates. These distinctions are very important for the advancement of nanotube field and applications because the production of isolated and suspended SWNTs of small diameter is very important. In this context, a small-diameter SWNT (Fig. 10A) was found to be isolated from other bundles and could be clearly observed through the use of an energy filter attached to the transmission electron microscope.[5] The diameter of the tube was measured to be ca. 0.426 nm, which is the smallest-diameter nanotube yet to be produced by the CVD method, and observed as a freestanding (isolated) structure. The tube was unstable under electron beam irradiation during TEM observation in comparison with other thicker SWNTs, and deformed within a minute. From the observed diameter, and also based on the comparison with the observed and simulated TEM images of the SWNT model, we conclude that this particular nanotube can be assigned to the (5, 1) indices with a diameter of ca. 0.436 nm. We could not find the end of this particular tube due to its extreme length and because it was mostly overlapped with either bundles of SWNTs or carbon particles. It is noteworthy that the diameter of this small tube is highly consistent with the diameter of a C_{20} molecule, which could be related to the end of a (5, 1) nanotube (see inset of Fig. 10A, which is capped with six adjacent pentagonal rings). There are two major mechanisms for the catalytic growth of CNTs depending on the location of the catalyst during growth. As shown in Fig. 10(B), one is the tip-growth mechanism whereby the catalyst particle is at the tip of the tube, and the other is the root-growth mechanism, whereby the catalytic particle is placed at the root of the growing tube.

PHYSICOCHEMICAL PROPERTIES OF CARBON NANOTUBES

Exceptional mechanical properties of CNTs are expected due to their strong carbon–carbon bond, if few structural defects are present. The TEM image (Fig. 11A) and a theoretical simulated image (Fig. 11B) demonstrate that CNTs have nearly ideal mechanical properties because nanotubes do not break even after twisting, bending, or flattening. The first report on Young's modulus of isolated MWNTs carried out inside the TEM indicated values of 1–1.8 TPa, which are much higher than those of commercially available carbon fibers (ca. 800 GPa).[27] From the theoretical calculation, Young's modulus of SWNTs, of around 1 nm, is in the

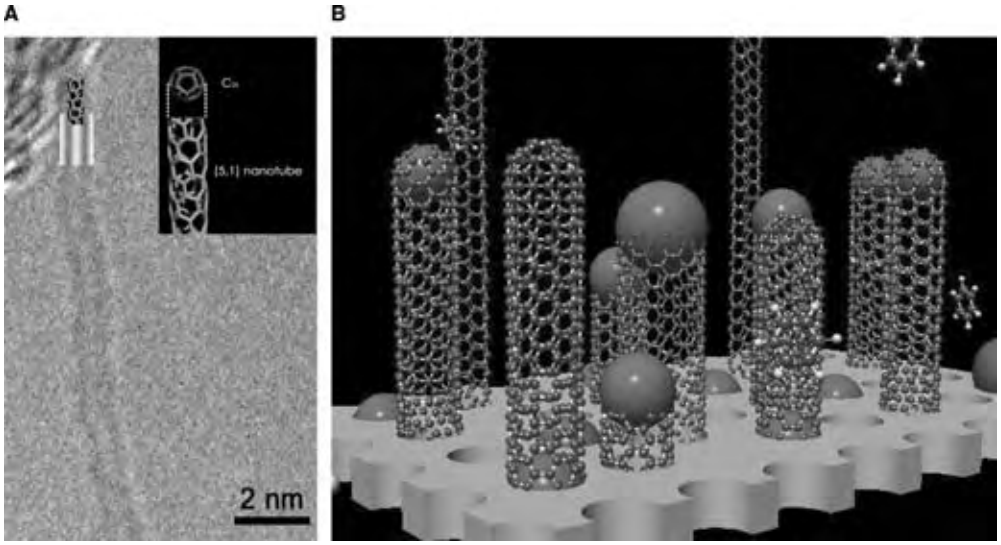


Fig. 10 (A) High-resolution transmission electron microscope image of a small-diameter SWNT. Inserted images are the model of (5, 1) tube and the TEM simulated image, which is in good agreement with the observation. (From Ref.[5].) [Inserted image is simulated model of the cap of a (5, 1) nanotube. The cap consists of six pentagonal rings, just like part of a C_{20} molecule, but slightly deformed.] (B) Schematic image of the tip-growth and root-growth models, in which the catalytic particles (red balls) are located at the tip and at the root, respectively. The blue balls are the carbon atoms and the brown base is the simplified model of a zeolite. (From Ref.[5].) *(View this art in color at www.dekker.com.)*

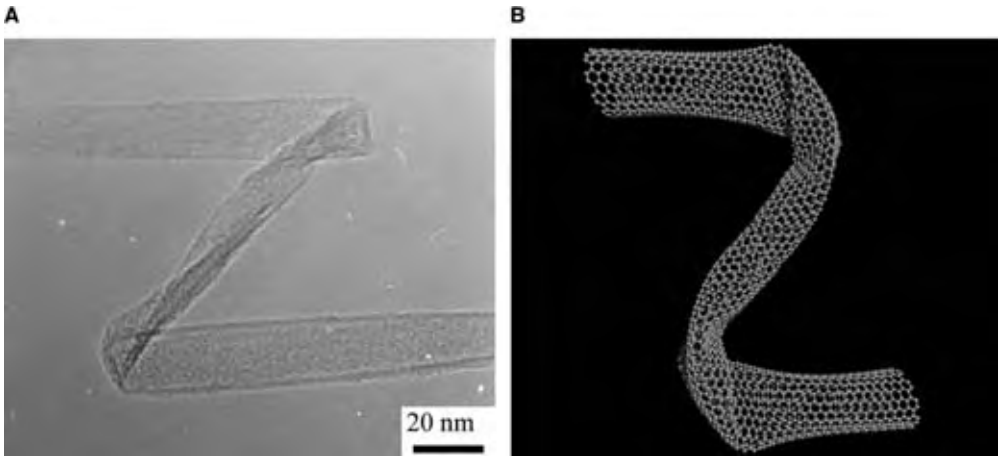


Fig. 11 (A) High-resolution transmission electron microscope image of a distorted SWNT and (B) a computer simulated model. This image shows the high flexibility of carbon nanotube. *(View this art in color at www.dekker.com.)*

range of 0.6–0.7 TPa. It is remarkable that the density-normalized modulus and strength of SWNTs are ca. 19 times higher than those of a steel wire.

Transport conductivity measurement on a single MWNT revealed that MWNTs exhibited both metallic and semiconducting behavior (resistivities at 300 K of $\sim 1.2 \times 10^{-4} - 5.1 \times 010^{-6}\,\Omega\,\text{cm}$; activation energies <300 meV for semiconducting tubes).[28,29] This experimental result indicates that the geometric difference (e.g., defects, chirality, diameter, etc.) and degree of crystallinity (hexagonal lattice perfection) strongly affect the electronic structure of MWNTs. In 1997, the first transport measurement on individual SWNTs (1 nm in diameter) demonstrated that SWNTs could behave as quantum wires, in which electrical conduction occurs via well-separated, discrete electron states that are quantum-mechanically coherent over long distances.[30] By utilizing scanning tunneling spectroscopy (STS) on individual SWNTs, it was found that they could be either metallic or semiconductors depending on small variations in the chiral angle or diameter.[31] Carbon nanotubes have the highest thermal conductivity because the phonon contribution dominates at all temperatures. The measured thermal conductivity of individual MWNTs (3000 W/m/K) is higher than that of diamond and that of graphite (ca. 2000 W/m/K).

CNT APPLICATIONS

The potential applications of CNTs due to their small dimension and excellent physicochemical properties make them useful in wide ranges from multifunctional composites, electrochemical electrode, and/or additives, field emitters, to nanosized semiconductor devices. Up to now, commercialized fields of CNTs are the filler in anode materials of lithium-ion secondary battery.[25] The desirable characteristics of MWNTs as composite fillers, especially in the electrodes of Li-ion batteries are summarized as follows: 1) The small diameter of the tubes makes it possible to distribute the fibers homogenously in the thin electrode material, thus introducing a larger surface area that could react efficiently with the electrolyte. 2) The improved electrical conductivity of the electrode is related to the high electrical conductivity of the tubes, and the network formation of the tubes in the anode forms a tube-mat. 3) As compared with that of conventional whiskers, the relatively high intercalation ability of MWNTs did not lower the capacity of anode materials upon cycling. 4) High flexibility of the electrode is achieved due to the network formation of MWNTs in the fiber-mat structure. 5) The high endurance of the electrode is because MWNTs absorb the stress caused by intercalation of Li-ions. 6) There is improved penetration of the electrolyte due to the homogenous distribution of the tubes surrounding the anode material. 7) As compared with that of carbon black, the cyclic efficiency of the Li-ion battery was improved for a relatively long cycle time. In the near future, the application of MWNTs as a filler material in various electrochemical systems is expected to become widespread due to their outstanding properties, especially for applications where improved conductivity is needed.

By utilizing MWNTs as scanning probe microscope tips, high-resolution images are obtained. On top of that, chemically functionalized tubes gave a high sensing ability for chemical and biological groups interacting with a different surface. Flat-panel display using SWNTs and MWNTs as a field emission electron source will be commercialized in the near future because CNT has a small diameter, high structural integrity, high electrical conductivity, and chemical stability. Also, utilization of CNTs as platinum support material in the fuel cell system will be realized.

In addition, the combination of high-aspect ratio, small diameter, excellent strength, stiffness, low density, and high conductivity makes nanotubes ideal candidates as fillers in polymer composites. Improved mechanical properties of nanotube/semicrystalline polymer composites are thought to enhance the crystallization of semicrystalline polymer because CNTs could act as nucleation sites that result in enhanced stress transfer from the matrix to the fillers. Further progress has to be done in order to fully utilize these nanotube/polymer composites, such as the optimization of surface properties, the homogenous dispersion without physical damages, the development of the alignment method (also evaluation method) and processing.

The fabrication of supercapacitors and electrochemical actuators that could be used as artificial muscles is another alternative for nanotube applications. Supercapacitors already were built on hybrid vehicles because this could provide rapid acceleration and store braking energy electrically. When using sheet electrodes with SWNT and MWNTs, it is possible to obtain specific capacitance of 180 and 102 F/g and power densities of 20 and 8 kW/kg, respectively.[32] On the other hand, nanotube actuators can work at low voltages and temperatures as high as 350 °C. For example, maximum stress observed in SWNTs is ~26 MPa, a value that is 100 times larger than that observed in natural muscles.

The possibility of using carbon nanotubes as nanowires is envisaged due to their observed ballistic transport. Unfortunately, the large contact resistance between the nanotube and the external circuitry has to be solved for this application. For the fabrication of nanotube field effect transistors, SWNTs were connected to metal nanoelectrodes (see Fig. 12). The performance is excellent in terms of switching speed owing to low capacitance.[33] An inherent problem associated with CNT lies in the difficulty in manipulating them. From a commercial viewpoint, further technical progress is required, such as selective growth of nanotubes using self-assembly techniques.

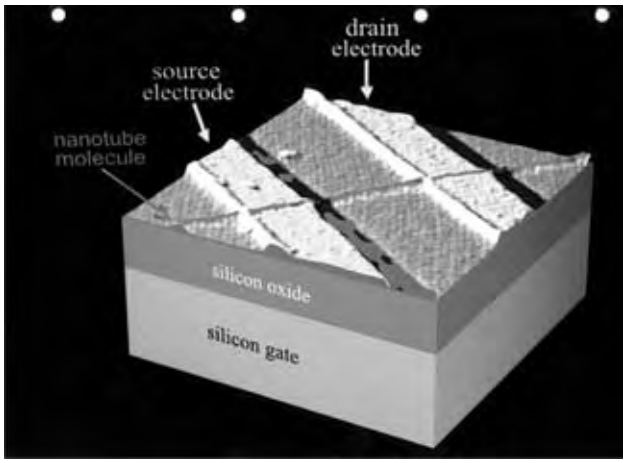


Fig. 12 An electronic device based on a single rolled-up sheet of carbon atoms. (From Ref.[33].) In the figure, a CNT (red; about 1 nm in diameter) bridges two closely spaced (400 nm apart) platinum electrodes (labeled source and drain) atop a silicon surface coated with an insulating silicon oxide layer. Applying an electric field to the silicon (via a gate electrode, not shown) turns on and off the flow of current across the nanotube, by controlling the movement of charge carriers onto the nanotube. *(View this art in color at www.dekker.com.)*

CONCLUSIONS

The 19th century can be remembered as an iron age while the 20th century is recognized to be the century founded on silicon technology. Now much attention is paid to tubular nanocarbon, ''carbon nanotube,'' as one of the promising candidates that could revolutionize technology in the 21st century. Carbon nanotubes are blooming as among other nanomaterials, and they could be visualized as connecting bridges between molecular and macrodevices. However, some challenges have to be solved in order to realize the potential uses of CNTs. The first one is the large-scale synthesis producing defect-free CNTs at low costs. Second, it is important to control the size and chirality of CNTs. The final and most important theme is how to manipulate these tiny molecules in order to fabricate novel nanodevices and materials. We envisage that in less than 10 yr various nanotube-based devices will be taking the place of emerging technologies.

ACKNOWLEDGMENTS

This work was supported by the CLUSTER of Ministry of Education, Culture, Sports, Science, and Technology. We also acknowledge CONACYT-México grants: W-8001-millennium initiative (MT), G-25851-E (MT), 37589-U (MT), and 41464-Inter American Collaboration (MT).

REFERENCES

1. Iijima, S. Helical microtubules of graphitic carbon. Nature **1991**, *354* (7), 56–58.
2. Saito, R.; Dresselhaus, M.S.; Dresselhaus, G. *Physical Properties of Carbon Nanotubes*; Imperial College Press: London, 1998.
3. Dresselhaus, M.S.; Dresselhaus, G.; Eklund, P.C. *Science of Fullerenes and Carbon Nanotubes*; Academic Press: San Diego, CA, 1996.
4. Qin, L.C.; Zhao, X.; Hirahara, K.; Miyamoto, Y.; Ando, Y.; Iijima, S. The smallest carbon nanotube. Nature **2000**, *408* (2), 50

5. Hayashi, T.; Kim, Y.A.; Matoba, T.; Ezaka, M.; Nishimura, K.; Tsukada, T.; Endo, M.; Dresselhaus, M.S. The smallest freestanding single-wall carbon nanotube, Nano Lett. **2003**, *3* (7), 887–889.
6. Oberlin, A.; Endo, M.; Koyama, T. Filamentous growth of carbon through benzene decomposition. J. Cryst. Growth **1976**, *32*, 335–349.
7. Wilder, J.W.G.; Venema, L.C.; Rinzler, A.G.; Smalley, R.E.; Dekker, C. Electronic structure of atomically resolved carbon nanotubes. Nature **1998**, *39*, 59–62.
8. Odom, T.W.; Huang, J.L.; Kim, P.; Lieber, C.M. Atomic structure and electronic properties of single-walled carbon nanotubes. Nature **1998**, *391*, 62–64.
9. Dresselhaus, M.S.; Dresselhaus, G.; Saito, R. Carbon fibers based on C_{60} and their symmetry. Phys. Rev. B **1992**, *45*, 6234–6242.
10. Saito, R.; Fujita, M.; Dresselhaus, G.; Dresselhaus, M.S. Electronic structure of chiral graphene tubules. Appl. Phys. Lett. **1992**, *60* (18), 2204–2206.
11. Sun, X.; Kiang, C.H.; Endo, M.; Takeuchi, K.; Furuta, T.; Dresselhaus, M.S. Stacking characteristics of graphene shells in carbon nanotubes. Phys. Rev. B **1996**, *54* (18), 1–4.
12. Issi, J.P.; Charlier, J.C. Electrical transport properties in carbon nanotubes. In *The Science and Technology of Carbon Nanotubes*; Tanaka, K., Yamabe, T., Fukui, T., Eds.; Elsevier: New York, 1999; Chapter 10, 107–127.
13. Endo, M.; Hishiyama, Y.; Koyama, T. Magnetoresistance effect in graphitizing carbon fibres prepared by benzene decomposition. J. Phys. D: Appl. Phys. **1982**, *15*, 353–363.
14. Baker, R.T.K. Catalytic growth of carbon filaments. Carbon **1989**, *27* (3), 315–323.
15. Tibbetts, G.G. Why are carbon filaments tubular? J. Cryst. Growth **1984**, *66*, 632–637.
16. Endo, M. Grow carbon fibers in the vapor phase. Chem. Tech. **1988**, *18*, 568–576.
17. Endo, M.; Takeuchi, K.; Kobori, K.; Takahashi, K.; Kroto, H.W.; Sarkar, A. Pyrolytic carbon nanotubes from vapor-grown carbon fibers. Carbon **1995**, *33* (7), 873–881.
18. Endo, M.; Saito, R.; Dresselhaus, M.S.; Dresselhaus, G. From carbon fibers to carbon nanotubes. In *Carbon Nanotubes*; Ebbesen, T.W., Ed.; CRC: New York, 1997; 35–105.
19. Dresselhaus, M.S.; Dresselhaus, G.; Sugihara, K.; Spain, I.L.; Goldberg, H.A. *Graphite Fibers and Filaments*; Springer Series in Materials Science; Springer-Verlag: Berlin, 1988; Vol. 5.
20. Endo, M.; Takeuchi, K.; Igarshi, S.; Kobori, K.; Shiraishi, M.; Kroto, H.W. The production and structure of pyrolytic carbon nanotubes (PCNTs). J. Phys. Chem. Solids **1994**, *54* (12), 1841–1848.
21. Iijima, S.; Ichihashi, T. Single-shell carbon nanotubes of 1-nm diameter. Nature **1993**, *363*, 603–605.
22. Bethune, D.S.; Kiang, C.H.; de Vries, M.S.; Gorman, G.; Savoy, R.; Vazquez, J.; Beyers, R. Cobalt-catalyzed growth of carbon nanotubes with single-atomic-layer walls. Nature **1993**, *363*, 605–606.
23. Journet, C.; Maser, W.K.; Bernier, P.; Loiseau, A.; Lamy de la Chapelle, M.; Lefrant, S.; Deniard, P.; Lee, R.; Fischer, J.E. Large-scale production of single-walled carbon nanotubes by the electric-arc technique. Nature **1997**, *388*, 756–758.
24. Thess, A.; Lee, R.; Nikolaev, P.; Dai, H.; Petit, P.; Robert, J.; Xu, C.; Lee, Y.H.; Kim, S.G.; Rinzler, A.G.; Colbert, D.T.; Scuseria, G.E.; Tomanek, D.; Fisher, J.E.; Smalley, R.E. Crystalline ropes of metallic carbon nanotubes. Science **1996**, *273*, 483–487.
25. Endo, M.; Kim, Y.A.; Hayashi, T.; Nishimura, K.; Matushita, T.; Miyashita, K.; Dresselhaus, M.S. Vapor-grown carbon fibers (VGCFs) basic properties and battery application. Carbon **2001**, *39* (9), 1287–1297.
26. Nikolaev, P.; Bronikowski, M.J.; Bradley, R.K.; Rohmund, F.; Colbert, D.T.; Smith, K.A.; Smalley, R.E. Gas-phase catalytic growth of single-walled carbon nanotubes from carbon monoxide. Chem. Phys. Lett. **1999**, *313*, 91–97.
27. Treacy, M.; Ebbesen, T.W.; Gibson, J.M. Exceptional high Young's modulus observed for individual carbon nanotubes. Nature **1996**, *381*, 678–689.
28. Dai, H.; Wong, E.W.; Lieber, C.M. Probing electrical transport in nanomaterials: conductivity of individual carbon nanotubes. Science **1996**, *272*, 523–556.
29. Ebbesen, T.W.; Lezec, H.J.; Hiura, H.; Bennett, J.W.; Ghaemi, H.F.; Thio, T. Electrical conductivity of individual carbon nanotubes. Nature **1996**, *382*, 54–56.
30. Tans, S.J.; Devoret, M.H.; Dai, H.; Thess, A.; Smalley, R.E.; Geerlings, L.J.; Dekker, C. Individual single-wall carbon nanotubes as quantum wires. Nature **1997**, *386*, 474–477.
31. Wildoer, J.W.G.; Venema, L.C.; Rinzler, A.G.; Smalley, R.E.; Dekker, C. Electronic structure of atomically resolved carbon nanotubes. Nature **1998**, *391*, 59–62.
32. Baughman, R.H.; Zakhidov, A.A.; de Heer, Walt A. Carbon nanotubes—the route toward applications. Science **2002**, *297*, 787–792.
33. Tans, S.J.; Verschueren, R.M.; Dekker, C. Room-temperature transistor based on a single carbon nanotube. Nature **1998**, *393*, 49–51.

Catalyst Preparation

X. D. Hu
Michael W. Balakos
R&D, Süd-Chemie Inc., Louisville, Kentucky, U.S.A.

INTRODUCTION

Heterogeneous catalysts are defined as solids or mixture of solids that are used to accelerate a chemical reaction without undergoing change themselves. The types of solids used in industry as heterogeneous catalysts include simple oxides, mixed oxides, metal salts, solid acids and bases, metals, and dispersed metals. Catalysts are used in a wide variety of chemical and environmental processes worldwide. The global value of fuels and chemicals produced by catalytic routes is about US$ 2.4–3 trillion per year.[1,2] About 20% of all products produced in the United States are derived from a catalytic process of some form.[1] As important as catalysis is to the world economy, the number of various chemicals used as a catalyst as well as the form and shape of the material vary as much as the number of processes that use catalysts. Fig. 1 is a picture of a number of various types of catalysts and illustrates the numerous possibilities of shapes and sizes. Naturally, the preparation processes of such a wide variety of products is also numerous.

The preparation of heterogeneous catalysts is a series of unit operations, particularly involving the processing of solids, to produce a catalyst with specific chemical and physical properties that are important in the performance of the material in the chemical process that it is being used.[3–6] The type and order of unit operations involved in the preparation vary greatly from catalyst to catalyst. There are numerous possibilities of manufacturing processes for particular catalyst formulations with each affecting the chemical and physical properties of the final product.

HETEROGENEOUS CATALYSIS

The choice of catalyst for a specific process depends on the performance of the material. For a catalyst to be economically effective, it must possess good activity, selectivity, and stability.[7] Activity is the rate at which the reaction proceeds with a specific volume of catalyst and flow rate of reactants. It is desirable to maximize the activity per volume of a catalyst, thereby minimizing the volume of the reactor. Selectivity of a catalyst is the ratio of the rate at which the desired reaction takes place over a catalyst to the rate of undesired side reactions. Since the side reactions convert feed to an unwanted product, a highly selective catalyst is economically desirable. The stability of a catalyst is a measure of the deactivation rate. For a catalyst to be viable, the deactivation rate should be low so that the catalyst change-out frequency is low, minimizing process down-time, catalyst costs, and turnaround costs.

The components of catalysts vary widely from process to process, but the common goal is to maximize the active material on the surface. In the case of a metal oxide, the oxide should have a large surface area, thereby exposing a large amount of active sites to the reactants. When the active species is a reduced metal or an expensive material, it is common practice to deposit the metal on a high surface area oxide support to maximize the active sites. The percentage of active metal that is exposed to the surface is the dispersion of the metal. If the metal atom is considered the active site, then a high dispersion is desirable to obtain a high activity per unit metal.

To eliminate intraparticle transport limitations, the particle size and average pore size must be carefully controlled during manufacture. Other physical properties that become important in industry have to do with the physical integrity of the catalyst particles. These properties include bulk density, crush strength, resistance to abrasion, and attrition. These properties are very important when working with reactors that contain a large amount of a particular catalyst. Fig. 2 lists a number of chemical and physical properties that affect catalyst performance.

The physical and chemical properties discussed above are influenced by every step of the preparation process as well as the choice of raw materials. In addition, several preparation routes may be available to obtain a catalyst with specific properties. On developing a catalyst for industrial applications, the influence of the preparation procedure on these properties must be taken into consideration as well as economical production of a material. Examples of industrial catalytic processes and the method of manufacture of the catalysts are shown in Table 1. This table illustrates the many different types of catalysts as well as preparation methods used in industry.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007693

Fig. 1 Heterogeneous catalysts of various shapes and sizes. (Courtesy of Süd-Chemie Inc.) *(View this art in color at www.dekker.com.)*

UNIT OPERATIONS OF CATALYST PREPARATION

The preparation of a heterogeneous catalyst consists of a series of unit operations that mostly deal with solids processing. A partial list of unit operations that are used in catalyst manufacture is given in Table 2. The table is segregated into four basic categories needed for the manufacture of heterogeneous catalysts: 1) precursor formation, 2) purification, 3) posttreatment,

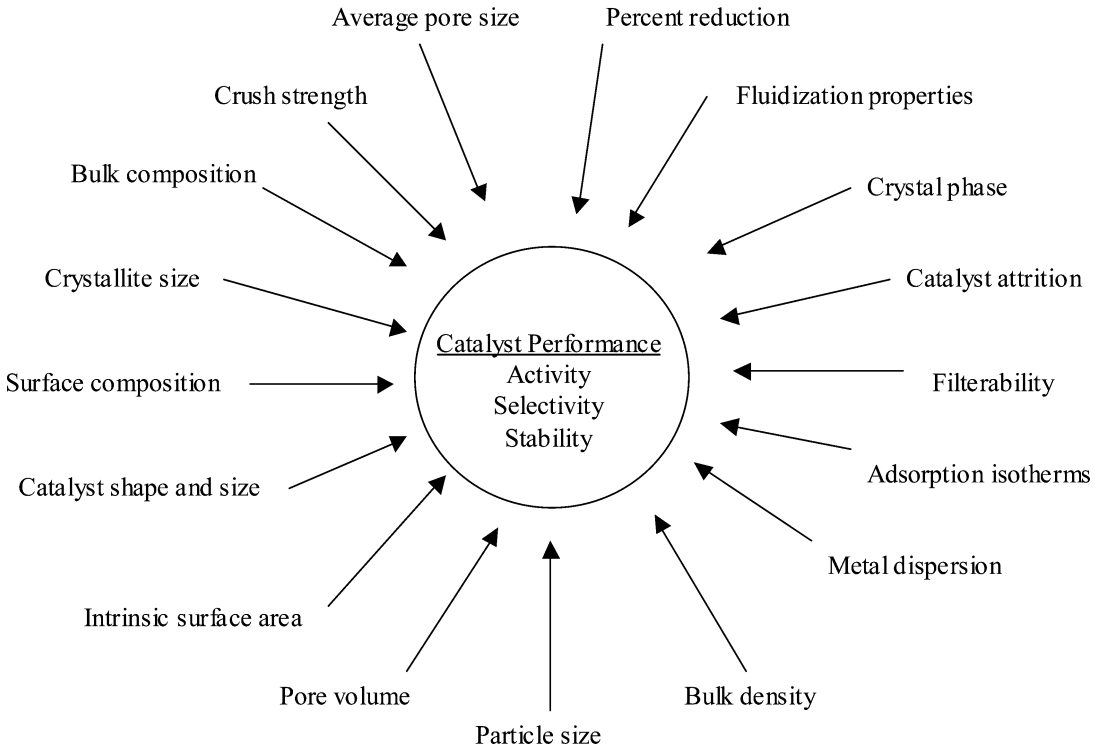


Fig. 2 Properties of heterogeneous catalysts that may be important in performance under industrial conditions.

Table 1 Examples of industrial catalytic reactions, the type of catalyst used and the typical preparation method

Industrial reaction	Catalyst	Catalyst preparation method
Ethylene oxychlorination	$CuCl_2/Al_2O_3$	Impregnation/thermal spreading
Ethylene hydration	WO_3/SiO_2	Precipitation/deposition
Polygas oligomerization	H_3PO_4/SiO_2	Compounding
C_5/C_6 isomerization	Pt/mordenite	Precipitation/ion-exchange/dipping
CO hydrogenation to liquid fuels	Cu-Fe-K/SiO_2, Co/support	Co-precipitation/impregnation
CO hydrogenation to methanol	$CuO/ZnO/Al_2O_3$	Co-precipitation
N_2 and H_2 to ammonia	Fe_2O_3	Fusing/smelting
Selective hydrogenation	Pd/Al_2O_3	Impregnation
Benzene hydrogenation	Ni/Al_2O_3, Pt/Al_2O_3	Precipitation/ionic-adsorption
Fats and oils hydrogenation	Ni/SiO_2-Al_2O_3	Precipitation
Ethylbenzene dehydrogenation	Fe_2O_3	Compounding
Ethylene oxydehydrogenation	Bi-Mo/SiO_2	Fluid bed impregnation
Ethylene epoxidation	Ag/Al_2O_3	Vacuum impregnation
Propylene oxidation	MMO	Precipitation
Butane oxidation	VPO	Precipitation

and 4) particle size enlargement. The preparation of catalysts may include all or only one of the four main steps. In addition, the steps may be combined and repeated in many different ways to achieve the desired properties as economically as possible. How the unit operations are carried out and the types of precursors greatly affect the properties of the catalyst can determine the activity, selectivity, and lifetime of the catalyst in an industrial process.

Precursor Formation

Precursor formation is the initial formation of the active solid material. This can be done by precipitation or coprecipitation of the required chemical species, decomposition, hydrothermal synthesis, adsorption or impregnation of active species onto a support, and other methods of synthesis.

Precipitation

The starting materials of the active components of catalysts are often in the form of soluble metal salts that are precipitated to form a solid precursor to a catalyst. Among the very limited soluble salts, metal nitrate is the preferred salt over fluoride, chloride, phosphate, or sulfate because of less corrosiveness and low residual anion on finished catalyst. Sometimes more expensive metal carboxylates such as formate, acetate, glycolate, or oxalate are used in the situation where disposal of nitrate is prohibited by environmental regulations or residual nitrogen is detrimental to the reaction.

The precipitation of solid catalyst precursor is caused by acid–base neutralization. The solution of metal salts is normally acidic while the base typically uses carbonate, bicarbonate, ammine, or hydroxide. The coprecipitated precursor is typically composed of mixed basic carbonates, hydroxycarbonates, and hydroxides. These compounds have the general formula[8]

$$Me^{2+}{}_x Me^{3+}{}_y (OH^-)_{2x+3y-az} (A^{a-})_z \cdot nH_2O \quad (1)$$

where Me^{2+} = metal cations with 2+ valence (Mg, Mn, Fe, Co, Ni, Cu, or Zn, etc.); Me^{3+} = metal cations with 3+ valence (Al, Fe, or Cr, etc.); A^{a-} = compensating anions (CO_3^{2-}, SO_4^{2-}, NO_3^-, or Cl^-, etc.); x, y, z, n = stoichiometric constants.

The conditions and procedures of precipitation play a significant role in catalyst morphology, texture, pore structure, physical strength, and consequently the performance (activity, selectivity, and stability) of the catalyst. By merely changing the sequence of solution addition, catalyst components can be precipitated simultaneously or sequentially. The methods of precipitation most often used are constant pH coprecipitation, sequential precipitation, acid-to-base precipitation, and base-to-acid precipitation.

Constant pH coprecipitation takes place by feeding acidic and basic solutions simultaneously into the precipitation tank. Usually, acidic solution is fed at a constant rate and the feed rate of basic solution is varied to maintain constant pH, typically at the acid–base neutralization point. The acid feed rate is predetermined based on consideration of residence time, batch time, or aging time. The coprecipitation reaction is typically controlled isothermally at 40–70°C with a constant stir

Table 2 Unit operations involved in catalyst manufacture

Precursor formation	Precipitation
	Co-precipitation
	Decomposition
	Hydrothermal synthesis
	Adsorption/impregnation
Solid-solution separation and purification	Impurity removal from solid Washing Ion exchange Decomposition
	Water or solvent removal Filtration Drying
	Volatiles removal
	Waste recovery
Posttreatment	Calcincation
	Reduction
	Stabilization
Catalyst formation	Preparation of coarse solid Mixing Milling
	Spray agglomeration
	Granulation Tabletting Pelletizing Extruding Spheridizing
	Washcoating

at 300–1500 rpm.[9] The precipitation can be batch or continuous operation. In the case of continuous operation, the slurry is continuously withdrawn from the precipitation tank to either an aging tank or an in-line filter press.

A uniform distribution of catalyst components in the precipitated agglomerates is expected by constant pH coprecipitation; however, there may still exist a concentration gradient from the surface to the core of the particle due to solid-state diffusion. As an example, copper is surface deficient, as evidenced by x-ray photoelectron spectroscopy (XPS) analysis on an industrial $CuO/ZnO/Al_2O_3$ catalyst for methanol synthesis reaction. Table 3 illustrates how the components of the catalyst can vary from the surface to the core of the particle.[10]

Sequential precipitation is used when a nonuniform distribution of catalyst components throughout a catalyst particle is desirable. In this method, the pH of the precipitation is controlled at specific levels at different

Table 3 A $CuO/ZnO/Al_2O_3$ catalyst shown to be Cu deficient on the surface by XPS analysis

Sample no.	Surface Cu/(Cu + Zn)	Bulk Cu/(Cu + Zn)	Surface/bulk	Relative rate const.
A	0.30	0.41	0.63	1.00
B	0.33	0.43	0.69	1.23
C	0.33	0.42	0.67	1.33
D	0.38	0.48	0.78	1.48
E	0.36	0.45	0.77	1.42

(From Ref.[10].)

times so that the individual component of the catalyst is precipitated sequentially. For example, in preparing a zinc–aluminum binary oxide catalyst, the precipitation of aluminum hydroxide takes place at a pH of 5.5 and the precipitation of zinc hydroxycarbonate takes place at a pH of 7.0–7.5.[11] Likewise, in preparing a copper–zinc–alumina ternary oxide catalyst, precipitation of aluminum can be as early as a pH of 2.9 and precipitation of copper hydroxycarbonate occurs at 4.4.[11] To create a multicomponent catalyst with desired heterogeneity by design, the components can be fed at different times even though they may precipitate at a similar pH.

Acid-to-base precipitation is used for batch operation, in which the mixed metal salt solution is fed into the basic solution (precipitation agent) in the reactor. During the acid-to-base operation, the pH continuously decreases to a predetermined lower value. The end pH can be further adjusted by adding an additional amount of acid or base. Base-to-acid precipitation is also an alternative, with the pH continuously increasing during the precipitation.

The physical and chemical properties of precipitated particles can be tailored by the precipitation conditions such as solution concentration, precipitation temperature, pH, method of addition, and mixing intensity.[4] The adjustment of these conditions can be achieved by using equipment with a setup similar to that depicted in Fig. 3.[9] The parameters most often controlled are the solution concentration, precipitation temperature, and the pH of solution.

The precipitation process, the reverse of solid dissolution, is dictated by solution thermodynamics. The solution reaches saturation state when the dissolution rate equals the precipitation rate. Nucleation starts when the solution concentration exceeds the saturation concentration. The solution concentration affects the crystallite size of the precipitated catalyst. For example, diluted solution is beneficial to crystal growth due to slow nucleation and the presence of a few nuclei. In contrast, submicrometer- or even nano-sized amorphous gel or sol can be formed starting with a more concentrated solution.

In general, nucleation and crystal growth rates reach a respective maximum with increasing temperature. In many systems, the temperature of the maximum nuclei formation rate is lower than that of the maximum crystal growth rate. Therefore, an intermediate precipitation temperature may be chosen for obtaining fine particles.

To determine precipitation temperature on a commercial scale, consideration is also given to such factors as production time, material handling, ease of subsequent processes, and production efficiency. Many catalyst manufacturing processes choose a precipitation temperature between 70°C and 80°C.

The pH of precipitation can be calculated from thermodynamic data such as the solubility product

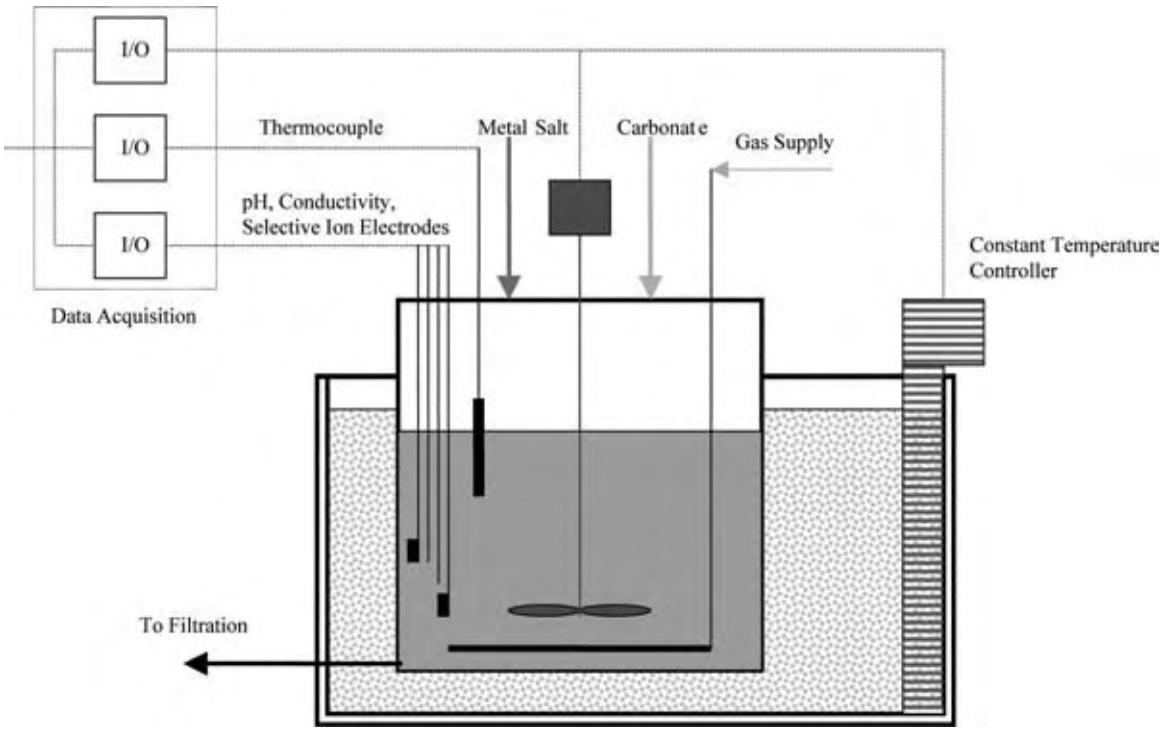


Fig. 3 Schematic of the precipitation process. *(View this art in color at www.dekker.com.)*

Table 4 Calculated solution pH for forming simple hydroxides of various compounds

Hydroxide	pH	Hydroxide	pH
$Mg(OH)_2$	10.5	$Fe(OH)_2$	5.5
Ag(OH)	9.5	$Cu(OH)_2$	5.3
$Mn(OH)_2$	8.5–8.8	$Cr(OH)_3$	5.3
$La(OH)_3$	8.4	$Zn(OH)_2$	5.2
$Ce(OH)_3$	7.4	$Al(OH)_3$	4.1
$Co(OH)_2$	6.8	$Zr(OH)_4$	2.0
$Ni(OH)_2$	6.7	$Fe(OH)_3$	2.0

constant for simple compounds or determined from a titration (neutralization) curve for solid mixture. The pH of forming a simple hydroxide solid is listed in Table 4. A pH of precipitation higher than the theoretical neutralization point is usually chosen to ensure completion.

Decomposition Deposition

In contrast to precipitation, deposition of a solid can occur by decomposition of a premixed solution upon changing temperature and pH. For example, by heating a mixture of aluminum salt and urea solution to 90–100°C, aluminum hydroxide will be precipitated due to an increase of the solution OH^- group from urea hydrolysis:

$$(NH_2)_2CO + 3H_2O \longrightarrow 2NH_4^+ + 2OH^- + CO_2 \quad (2)$$

The aim is to form a less soluble or nonsoluble solid deposited onto a support by decomposition, disproportionation, or hydrolysis of complexes or other organic compounds.

An example is the thermal decomposition of a metal onto a support involving an ammine complex having the formula $[Me(NH_3)_6]^{2+}$ CO_3^{2-} where Me is the metal. In this complex, six ammonia molecules coordinate symmetrically from the central metal ion and form an octahedral structure. It is therefore possible to form small metal crystallite-sized catalyst by using ammine complex as a precursor. The unique features of a catalyst made from ammine decomposition include uniform distribution of metal throughout the catalyst particles, smooth and homogenous surface morphology, and high thermal stability.[11,12]

A common procedure in decomposition deposition involves mixing ammine solution with a highly dispersed carrier material, and then slowly raising the reaction temperature to the decomposition temperature.[12] During the process, ammonia is released from the complex and the metal hydroxycarbonate species are deposited onto the carrier material. As shown in a typical process flow sheet for this method in Fig. 4,

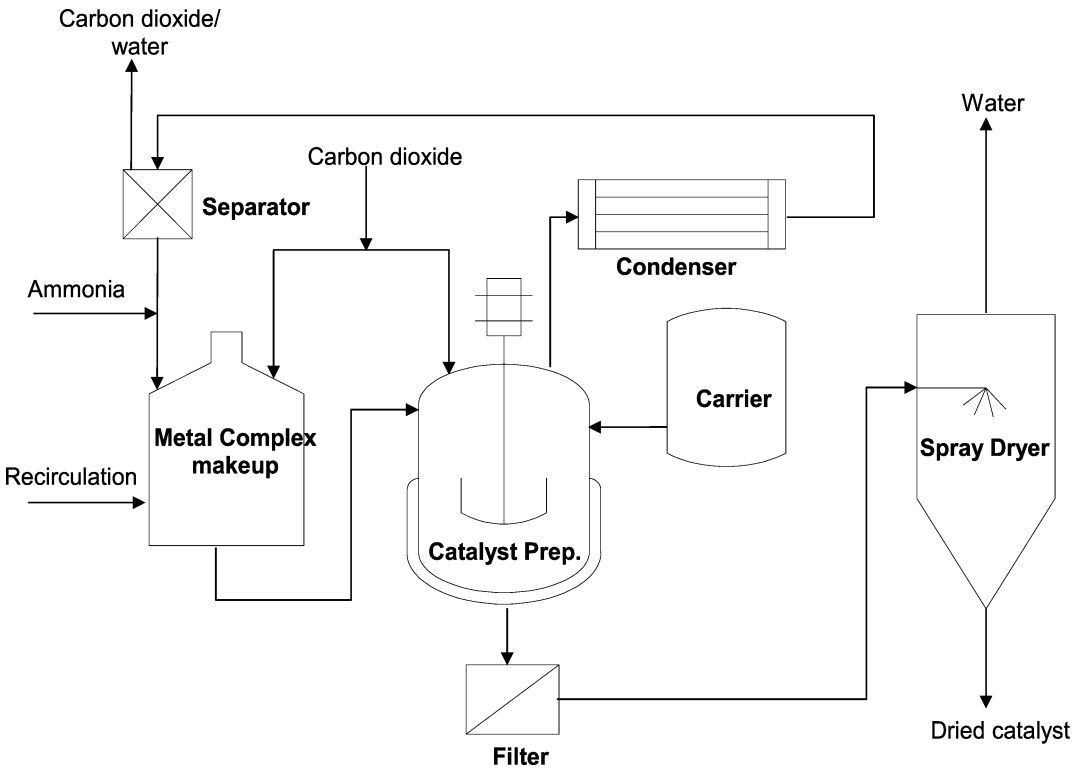


Fig. 4 Schematic of ammine decomposition process.

ammonia and carbon dioxide are recovered. Water is separated from the catalyst precursor and can be recovered as well. Therefore, the water consumption in the decomposition deposition process is much lower than in acid–base coprecipitation process.

Compounding

Physical mixing of different components of a catalyst is one of the simplest methods of catalyst preparation. The catalyst components can be added in a mixer in one step or in several steps. Good mixing is a key to ensuring a homogenous dispersion of all the components. The advantage of the compounding method is that it usually does not require filtration and washing, so no wastewater is generated. The disadvantage of the method is that interdispersion of the components is poor in some cases and dust pollution is problematic.

An example of using the compounding method is an iron catalyst for dehydrogenation of ethyl benzene to styrene.[13] For making a catalyst containing 11.2% potassium oxide and 88.8% iron oxide, a mixture of the required amounts of potassium carbonate and unhydrated iron oxide is dry-blended with a small amount of organic lubricant. Water is then added into the oxide mixture to form an extrudable paste that is then formed in cylindrical pellets, dried, and calcined at 600°C. It should be noted that not only can the compounding method be used to make a mixture of oxides, it can also be used to make nano-phased mixed oxide through solid-state reaction using solvent-aided intensive mixing and appropriate posttreatment. In many cases, compounding is therefore not only physical mixing but also chemical activation and physicochemical modification of the solid under the influence of mechanical energy. An illustration of such activation is the preparation of a supported copper catalyst.[14] First, $CuCl_2$ is compounded homogenously with γ-Al_2O_3. The solid mixture is then subjected to a heat treatment at 300°C. During mixing and heat treatment, $CuCl_2$ is spontaneously dispersed onto alumina to form a monolayer. The highly dispersed chloride compound cannot be detected by x-ray diffraction.[14]

Wet compounding is another way to hasten the formation of a solid solution. In this method, a homogenous solution or a colloidal solution containing active components of the catalyst is mixed with other components of the catalyst in solid form. The reaction between solution and solid is activated by vigorous mixing and subsequent calcination. Solid phosphoric acid catalyst for hydrocarbon polymerization is prepared by this method.[15]

Compounding can be carried out by several different types of mixers/blenders. Based on the degree of mixing and time required, mixers are selected in accordance with their capabilities to generate diffusive, convective, and/or shear mixing. A tumbler gives diffusive mixing where movement of individual particles in random motion leads to a high degree of homogeneity in a long period of time.[16] A mix muller, shown in Fig. 5, is a type in which pan and muller turret move in opposite directions, and gives mostly convective mixing. For rapid compounding, a rotating pan mixer equipped with baffle and agitator blades is very often used. As shown in Fig. 6, the pan mixer generates actions of both shearing and convective mixing.

Impregnation/Coating

This method is used to incorporate active components (as solution) onto solid support. The support can be in the form of tablet, beads, extrudates, granules, or even powder. The processes of the preparation include incipient wetness, excess solution adsorption, and surface coating. The driving force of the solute moving from solution to solid in a porous support is either van der Waals (physical adsorption), chemical binding (chemical adsorption), or capillary. The pore radius in most porous supports ranges from several to several hundred angstroms. The capillary force calculated from such small pores can be equivalent to a value of tens to thousands of bars. By considering the viscosity of impregnation solution, the time required to penetrate into the internal pore structure can be estimated by the following equation:

$$t = \frac{2\eta}{\delta \cos\theta} \frac{r^2}{\overline{R}} \tag{3}$$

where η is the viscosity coefficient of impregnation solution, r is the penetration distance (depth) at time

Fig. 5 Mix muller-type mixer. (Courtesy of Süd-Chemie Inc.) *(View this art in color at www.dekker.com.)*

Fig. 6 Rotating pan mixer. (Courtesy of Eirich Machines, Inc.) *(View this art in color at www.dekker.com.)*

t, δ is the surface tension of solution, $\overline{R}$ is the mean pore radius, θ is the contact angle between solution and support surface.

The equation can also be expressed as in Ref.[17]:

$$\xi = \Omega t^{0.5} \tag{4}$$

where ξ is the dimensionless penetration depth, Ω is the penetration coefficient, t is the impregnation time.

In incipient wetness, the volume of solution used is equal to the pore volume of support to ensure total adsorption of solution into the mesopores. This process may be repeated several times in order to reach the desirable concentration level of the active components. If there is an excess amount of solution, it will be coated on the surface of the carrier to form an active component eggshell.

In excess solution adsorption, the support material is submerged in excess amount of impregnation solution (the volume of impregnation solution is much higher than the pore volume of the support). The excess solution is filtered after adsorption equilibrium is reached. In many cases, competitive adsorption between solvent and solutes and/or between different solutes leads to a nonuniform distribution of active components throughout the support particles. This phenomenon can be utilized to enhance performance (normally selectivity) of certain types of catalysts. The distribution of the active components can also be tailored by the manipulation of the pore structure of the support, pH and viscosity of the solution.[17]

One of the most important parameters of the solid surface is the isoelectric point (IEP), or point of zero charge (PZC). As a rule, an oxide particle dipped in a solution at a pH lower than IEP tends to polarize positively to adsorb compensating anions. The same particle dipped in a solution at a pH higher than the IEP has a negative surface charge that is compensated by adsorbed cations. Thus, by choosing the isoelectric point of the oxide, the pH value of the impregnating solution, or the type of impregnating compound, the adsorption strength and, consequently, the behavior of thermal stability of the catalyst can be altered as shown in Table 5.[18]

Novel Preparation Methods

The activity and selectivity of a solid catalyst toward a desirable product are often limited by the randomness of the molecular arrangements of the active components in the catalysts prepared by traditional methods. It is therefore most challenging to devise new solid catalyst preparation methods to control metal dispersion, metal–support interaction, and pore structure on the molecular or nanometer level. It is equally challenging to form a heterogeneous catalyst without the need for further steps (catalytic or noncatalytic) to treat the effluents in the entire preparation process. Emerging new techniques in catalyst preparation are summarized in the following areas.

Nano-scale synthesis

The methods that are commonly used to produce semiconductors and electronic materials are now being used in catalyst preparation. As recently reported, lithography has been used to deposit metal in predetermined patterns.[19] Deposition of small clusters using nano-scale scanning tunneling microscopic tip has also been demonstrated.

Table 5 Effect of point of zero charge (PZC) on adsorption strength

Adsorption strength	Strong	Weak
Sintering upon heating	Difficult	Easy
pH < PZC		
Positive surface polarity		
a) Anionic metal compound	X	
b) Cationic metal compound		X
pH > PZC		
Negative surface polarity		
a) Anionic metal compound		X
b) Cationic metal compound	X	
With secondary competitive ions		X
Form surface complex	X	

(From Ref.[18].)

High-throughput synthesis

Combinatorial synthesis of intermediates in the pharmaceutical industry is being widely used for efficient preparation of solid catalysts. The high-throughput combinatorial method is particularly useful in making a large array of samples for combinatorial testing. By this method, it is possible to search for ''out-of-box'' formulations as well as systematically investigate new reaction and preparation conditions. For example, a new robotic workstation has been reported to allow wet or dry impregnation of solution containing Mo/V/Nb/W on 48 alumina samples in parallel.[20]

Preparation using alternate energy source

In contrast to traditional energy sources, microwave, laser, and sonic energy can be delivered to specific solid sites or small areas by fine-tuning frequencies and energy levels. It is especially suitable for preparation of nano-structured catalyst. For example, inherent ''eggshell'' catalysts can be produced with sonochemical preparation by which nano-dispersed metal particles are formed with instantaneous decomposition of metal solution by high-intensity irradiation of ultrasound in local area. In the preparation of Mo_2C on ZSM-5 with ultrasound at 20 kHz, narrowly distributed particles of about 2 nm in diameter are uniformly dispersed on the outer surface of the ZSM-5 support.[21]

Green chemistry synthesis

New methods are focused on the preparation of solid catalysts without release of gaseous or liquid waste streams.[22] In zeolite synthesis, the chemical compounds generated during combustion of templates (structure-directing agent) are often hazardous and harmful to the environment. A new method teaches disassembling the template within zeolite pore space into small fragments without combustion.[23]

Solid–Solution Separation and Purification

Before forming into a shape, the precursor in slurry needs to be purified and then concentrated from the mother liquor. Because most industrial catalysts cannot tolerate many trace impurities (Na, K, Cl, S, P, etc.), the purification process is often exhaustive with single or even multistep unit operations such as decantation, ion exchange, and filtration.

Ionic impurities are adsorbed on the surface of the catalyst precursor, especially on gelatinous material with high surface area. Water washing with up to several hundred times more than the amount of solid may not be sufficient to remove the impurities to the desired low level. In those cases, ion exchange with ''benign'' ionic compounds, which are not detrimental to the reaction, removed by heat treatment are frequently used. For example, dilute NH_4OH solution is used to replace Na in purifying alumina gel. However, washing and ion exchange are often a continuation of the aging process. One often observes color change during extensive washing, indicating transformation of crystallite phases of the precursor. Therefore, the aging effects of washing on properties of the catalyst precursor must be considered when choosing a proper washing method, equipment, conditions, and ion exchange reagent.

Posttreatment

The purpose of heat treating a solid precursor is to remove volatiles (typically water) and to convert the solid to a desirable amorphous or crystallite phase. It is during heat treatment that the precursor converts to a physically robust and chemically active catalyst. It affects such properties of the catalyst as surface acidity, number of active sites, surface area, pore structure, and crush strength. Several operations that involve heat treatment include drying, calcination, reduction, and stabilization and they are frequently employed in multiple steps before and after forming catalyst pellets.

Drying is defined as the process of eliminating the solvent from the solid precursor. Drying is accomplished by diffusion of the solvent from the solid interior to the surface and by evaporation of the solvent with heat and/or vacuum. The pore structure may start to form during this process.

Solvent can be associated with the solid in different chemical and physical states. It can be part of the crystalline structure, chemically adsorbed in solid interior or physically penetrated/adsorbed inside the interstices of the solid. The interstices of the solid are subjected to a large capillary tension upon drying. The force exerted to solid can be estimated by Laplace's law:[21]

$$\Delta P = \frac{2\delta}{r} \tag{5}$$

where δ is the surface tension of the solvent (70 dyn/cm for water) and r is the radius of the capillary.

The large capillary force tends to collapse the interstices among the solid particles and bring the particles together.[21] For example, a pressure of about 800 bar is created within 8 nm particle interstices.

Drying under supercritical conditions proceeds in the absence of liquid/gas interfaces and, therefore, is free of the effects of surface tension. Supercritical drying avoids collapse of the solid matrix and forms an aerogel with extremely high surface area and high

Table 6 Effect of drying conditions on physical properties of the alumina catalyst after drying

Drying conditions	Density (g/cm^3)	BET surface area (m^2/g)	Radical shrinkage (%)
Ambient evaporation	1.214	171	74
Supercritical CO_2	0.058	718	8
Supercritical acetone	0.146	716	34

(From Ref.[24].)

pore volume with low density.[24,25] However, the disadvantage of aerogel as catalyst material is that it cannot withstand high temperature calcinations. A comparison of conventional and supercritical drying is given in Table 6.[24]

Several unit operations can be adapted for drying purposes according to the location of the solvent in the solid and requirement of pore volume and surface area of the final catalyst. Besides oven drying, the common drying methods include spray drying, vacuum drying, and fluid bed drying. A drier of 7–30 ft in diameter equipped with an atomizer is commonly used for spray drying. A double cone mixer with vacuum capability is used for vacuum drying. A fluid bed coater can also be used for fluid bed drying. Fluid bed drying can be used in conjunction with impregnation in one step or in separate steps. The effects of drying methods on catalyst surface area and pore volume are demonstrated in drying of silica material, as shown in Table 7.[26]

Calcination is a process of heat treatment at temperatures higher than the drying temperature. The phenomena of thermal decomposition, crystallization and recrystallization, and sintering are observed during calcination.

Thermal decomposition is the process wherein the structure of the catalyst is formed by the heat treatment of the precursor after volatile components are decomposed or chemical water associated with the lattice structure of the solid is removed. Examples of such a phenomenon are the decomposition of metal nitrate, hydroxide, carbonate, chloride, sulfate, phosphate, hydroxy salts, or oxy salts to corresponding oxides. The following equation shows the decomposition of cobalt nitrate coupling with partial oxidation of Co^{2+} to Co^{3+} and formation of spinel Co_3O_4.

$$3Co(NO_3)_2 \rightarrow Co_3O_4 + 6NO_x + (7 - 3x)O_2 \ (x < 2.333) \quad (6)$$

As seen from the equation, two moles of NO_x are generated for every mole of $Co(NO_3)_2$ decomposed. Caustic scrubbing or catalytic reduction is most commonly used in commercial catalyst production for NO_x removal.

During calcination, the precursor undergoes a continuous crystallization and recrystallization process during which thermodynamically unstable or metastable phases are converted to more stable ones. X-ray diffraction patterns in Fig. 7 show phase transformation of hydrous titanium oxide from 100°C to 550°C.[27]

Sintering is a phenomenon that occurs at a calcination temperature higher than the Tamman temperature or roughly above two-thirds the melting temperature (in K) of the solid. Sintering results in agglomeration of solid particles. During calcination, the surface area of the solid may initially increase due to recrystallization and then decreases monotonically due to sintering. As shown in Fig. 8, addition of rare earth metals to the oxide (hydrous titanium oxide in this case) may delay sintering and, consequently, stabilize the surface area of the catalyst.[27]

It should be noted that pore volume and average pore size of the catalyst are changing with the decrease of surface area during calcination according to the relationship

$$\bar{r} \propto \frac{2V}{S} \quad (7)$$

The average pore radius will be increased if pore volume (V) remains constant upon heating. The average pore radius will decrease or remain constant if pore volume is also decreased. As shown in Table 8, surface area of zirconia is gradually decreased from 165 to $\sim$2 m^2/g when being calcined to higher temperatures, whereas mesopores less than 100 Å initially become more populous and are eventually eliminated when the material is severely sintered.

Table 7 Effect of drying methods on physical properties of a silica catalyst

Drying methods	BET surface area (m^2/g)	Pore volume (cm^3/g)
Spray	105	0.15
Vacuum	132	0.21
Fluid bed	145	0.26

(From Ref.[26].)

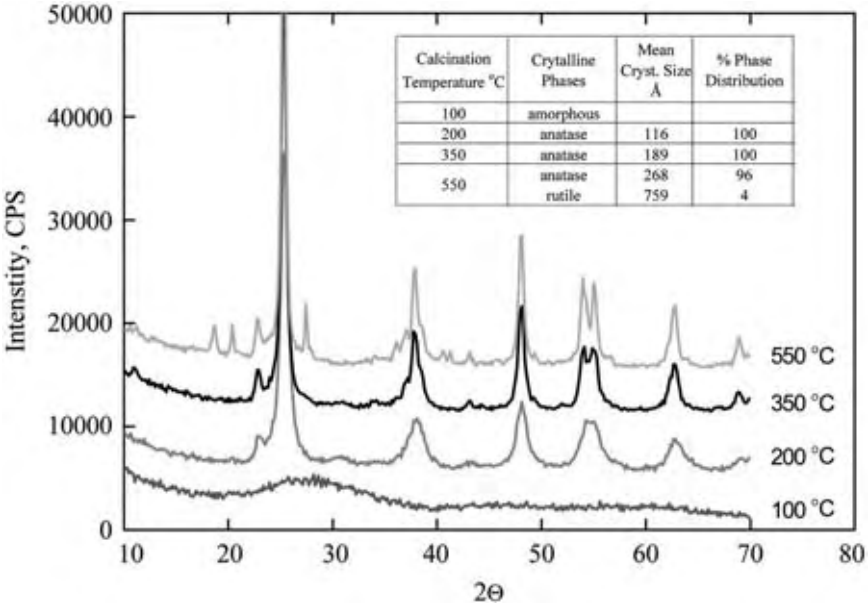


Fig. 7 XRD pattern of TiO_2 showing phase transformation upon calcination. (From Ref.[27].) *(View this art in color at www.dekker.com.)*

Calcination on an industrial scale usually uses a box kiln, belt calciner, or rotary calciner. A rotary calciner is the most widely used calcination equipment owing to its good solid–solid mixing, enhanced solid–gas energy exchange, large contact area with heating surface, short residence time, and economic operation. It is also used in the reduction of precious metal on support and catalyst regeneration.

The time of passage (TOP) or mean residence time of catalyst particles is stated to be the primary variable affecting rotary calciner operation.[28] It can be expressed as a function of feed rate or rotation speed:

$$\text{TOP} = \text{Holdup/Feed rate} = \text{W/Fw} \tag{8}$$

$$= \alpha \frac{L}{\tan\theta N^b D} \tag{9}$$

where α is a constant between 0.2 and 0.3, b is a constant between 0.9 and 1.0, θ is an angle of inclination on holdup, N is the rotation speed of the rotary, D is the diameter of the calciner, and L is the length of the calciner.

The TOP of a commercial scale rotary calciner is typically 30–60 min and depends primarily on feed rate and tube length. The TOP in a rotary calciner is much shorter than that in a belt calciner. A continuous multipass operation is sometimes employed for materials requiring a long calcination time.

The majority of base metal and precious metal catalysts used in industry are active in the reduced state.[7,29] Flowing hydrogen diluted in nitrogen over the oxide at elevated temperatures is the most common procedure for reducing the oxide to its most active state.[7] Reduction can be done in situ in the industrial reactor before the catalyst is put into service or at the catalyst manufacturing plant. In a catalyst manufacturing plant, the reduction is done in a fixed bed or fluid bed furnace. Since water is the by-product of reduction, the hydrogen concentration must be high enough to reduce the oxide in a reasonable time frame while low enough to prevent the product steam from sintering the metals. In addition, reduction of metal oxides is typically an exothermic process; therefore, a lower H_2 concentration prevents thermal runaway.

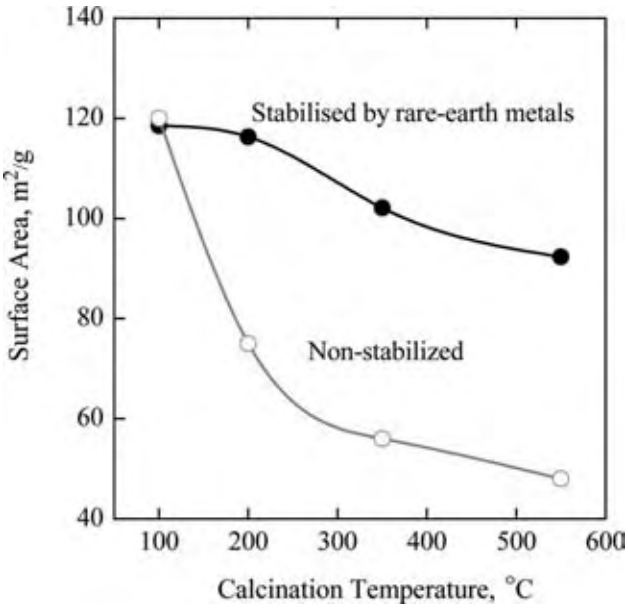


Fig. 8 Stabilization of hydrous titanium oxide surface by the addition of rare-earth oxide to the support. (From Ref.[27].) *(View this art in color at www.dekker.com.)*

Table 8 Effect of calcination conditions on a ZrO_2 support

Material	ZrO_2			
Pretreatment conditions	300°C	550°C	700°C	1200°C
Surface area (m^2/g)	165	81	47	2.4
Pore volume (cm^3/g)	0.22	0.33	0.29	0.14
Particle size distribution (Å)				
<100	14.98	34.38	34.07	1.05
100–1000	5.52	5.93	4.53	9.08
1000–5 μm	79.51	59.59	61.40	79.73
5–10 μm				10.13

Some metal compounds can be reduced by a chemical reagent such as formaldehyde, formic acid, or hydrazine. These methods are generally more expensive and are used only when conventional hydrogen methods are not feasible.

Once reduced, the metals may be pyrophoric or self-heating upon exposure to air. This is problematic if the reduction is done at the catalyst manufacturing plant and must be shipped to the end-user's plants. In this case, a stabilization procedure is employed at the catalyst manufacturing plant so the material is safe for transport as per U.S. Department of Transportation standards.

Typically, stabilization of the catalyst consists of forming an oxide layer on the outer surface of the metal, while keeping the inner surface reduced. This is the result of controlled oxidation of the catalyst at low temperature. The effect is that 5–20% of the metal is oxidized and shields the inner reduced material from air. The outer oxide layer readily reduces in situ with hydrogen or with the reactant gases at a temperature often lower than the bulk reduction temperature when the plant is started.

Another common stabilization procedure is to submerge the reduced catalyst into a solvent under N_2 atmosphere. Since air cannot get to the catalyst surface, the catalyst can be shipped, submerged in the solvent. The catalyst is carefully loaded into the reactor without exposure to air and the solvent washed away. Normally, a solvent that is similar to the reaction product is used for such stabilizations.

An example of immersion stabilization is nickel catalysts for the hydrogenation of edible oils.[30] The nickel catalyst is a powder that is used in batch slurry reactors. Since nickel is pyrophoric, the material must be stabilized for shipment to end user plants. After reduction in a fluid bed furnace, the catalyst is dropped under a N_2 atmosphere into melted, fully hydrogenated edible oil. The thick slurry is then pastillated into droplets and solidified at room temperature. The result is small droplet-shaped solids of reduced nickel catalyst embedded in hardened edible oil. When used at the edible oil facility, the fully hydrogenated oil readily dissolves in the reaction mixture, exposing the catalyst to the reactants.

Catalyst Particle Forming

In order to be uniformly filled in a commercial reactor without causing flow distribution problems, an industrial catalyst must be formed in a certain shape and size. The specifications for catalyst size, shape, and mechanical strength are decided in accordance with the energy balance, pressure drop, and flow dynamics of the reactor design.

Forming (or granulation) is a process of particle enlargement by agglomerating small particles together into large pellets, tablets, extrusions, spheres, and microspheres.[31] In fluid bed reactors, microspherical catalyst powders with medium particle size of 60–80 μm give the best fluidization properties. Shaped extrudates, tablets, or pellets are generally used in fixed bed reactors. The particles that provide the largest external geometric surface areas (EGSA) are usually desirable. Relative pressure drops of the catalyst with different shapes and sizes are shown in Table 9.

The methods of forming can be put into four categories: pressure forming, rotation forming, liquid-bound agglomeration, and oil-drop forming. The applicability of the particles from different forming methods is summarized in Table 10.

Pressure forming by tabletting, pelletizing, or extruding is used most frequently in industrial catalyst production. A tabletting machine is employed for compacting dry powder into cylindrical tablets. The momentum of the compacting force is delivered by two moving punches to a die. During the tabletting process, powder is continuously fed into the die from a feeder and tablets are continuously ejected from the die. Fluidity of the catalyst powder is the key to forming a strong tablet. If the powder lacks the required fluidity and lubricity, two measures are usually taken—powder densification by an additional kneading step and/or addition of lubricating material such as graphite, paraffin oil, clay, talc, stearic acid, or metal stearate.

Table 9 Relative pressure drop across a catalyst bed of various geometry and size

Shape	Height (inch)	Diameter (inch)	Hole diameter (inch)	EGSA (sf/cf)	Void	Relative ΔP. (%)
Cylinder	>0.24	0.12	n/a	231.6	0.49	1.00
One-hole ring	0.16	0.16	0.079	368.8	0.54	1.04
Three-hole ring	0.20	0.22	0.067	311.4	0.57	0.89
Trilobe	0.20	0.22	0.067	393.2	0.53	0.78

Pressure forming can also use a pellet press (or pellet mill or pelletizer) operating at a relatively lower pressure. Pellets are agglomerated by forcing them through orifices of the die by a rotating (gear) or rolling (roller) element. Similar to the tabletting machine operation, the material is densified, deformed, and pressed into large particles. For material sensitive to heat generated by compression, the feed hopper and rolls are usually cooled to control the temperature of the material. Lubrication is critical for adequate friction between the surfaces of the catalyst and die. Polymers are often used as a lubricant. The melting point of the polymer chosen should be slightly higher than the maximum temperature in the chamber of the press.

Compaction conditions affect the total pore volume and shape of pores for some materials. In general, pore volume is smaller for a catalyst granulated by a tabletting machine than that by a pellet press. Partial pore closure is possible for tablets having a very high crush strength.

A screw extruder is another type of pressure forming equipment used. The material is fed into the rear of the extruder, conveyed along the screw inside the barrel, and forced through an orifice die plate. The barrel can be heated or cooled to mediate the flow properties of the material. The screw is driven by a motor with variable speeds. The orifice of the die plate can be shaped into tri-lobe, quad-lobe, or other geometric shapes. The material flow within the barrel is primarily due to the drag force of the screw and can be estimated using the equation (for a single-screw machine):

$$Q = 0.5\pi^2 D^2 NL \sin\theta \cos\theta \quad (10)$$

where: Q = volume flow rate, D = barrel diameter, L = chamber length, N = screw speed, and θ = screw flight angle.

Binder is needed for forming extrudates. Besides water, colloidal alumina, silica, or cellulose are frequently used as binding material. The crush strength of the extrusion is normally weaker than the products granulated by compaction or compression. However, the advantages of extruding are high capacity and low cost.

Catalysts on structured supports are becoming more common in industry. The main application for such materials has been for environmental processes such as NO_x abatement and autocatalytic converters, but the concept has been gaining ground on other industrial applications.[32] The most common form of structured catalyst is the monolith, which is made of many channels where the active catalytic component is on the walls of the channels. Such structures can be made to have up to 1200 cpsi (channels per square inch) or more. The advantages of monolithic-type catalysts are lower pressure drop, greater geometric surface area, and no attrition due to vibrations. The disadvantages are that laminar flow through the small channels may cause mass transport limitations and no interconnectivity between channels causes poor radial heat conductivity.

The monolithic substrate can be made of either ceramic or steel. Ceramic monoliths are extruded in the form with cordierite being the main material used. Metallic monoliths are made by corrugation, followed by rolling or folding the metal into monoliths of the desired shape and size.

The washcoating procedure may be such that the catalyst powder is put onto the surface of the substrate

Table 10 Characteristics of catalyst forming methods

Forming method	Used for reactor type	Common shapes	Normal size range (mm)
Compacting	Tubular bed	Tablet, ring,	3–55
Compressing	Fixed bed	Pellet, ring	1.5–10
Extruding	Fixed bed	Extrudate, ring	1.2–6
Rotating	Fixed bed	Spheres	2–5
Oil-drop	Moving bed	Small spheres	1–3
Spraying	Fluid bed, bubble column	Micro-spheres	0.2–1.0

or the oxide support is put onto the substrate and then the active metal component is put on by impregnation, adsorption, or ion exchange.[33] Washcoating is the operation of dipping the substrate into a slurry of the material to be washcoated, followed by drying and calcination. The calcination fixes the powder to the surface of the substrate so that the walls of the channels are covered to the desired thickness. The operation is much more complicated in that process parameters such as density, viscosity, pH, particle size, and others affect the thickness and adhesion of the catalytic layer.

CONCLUSIONS

The preparation of heterogeneous catalysts is a complex operation that can incorporate many unit operations. The operation involves the formation of the catalyst precursor, separation, and purification, posttreatment, and forming. These operations can be carried out in many ways. All of these operations affect the physical and chemical properties of the final catalysts, which affect the performance in industrial use. Last, the economics of the process that uses the catalyst will be affected and is the ultimate driving force.

REFERENCES

1. Heinemann, H. Development of industrial catalysis. In *Handbook of Heterogeneous Catalysis*, Ertl, G., Knozinger, H., Weitkamp, J., Eds.; Wiley-VCH: Weinheim, Germany, 1997; Vol. 1, 35–48.
2. Schmidt, F. New catalyst preparation technologies—observed from an industrial viewpoint. Appl. Catal. **2001**, *221*, 15–21.
3. Campanati, M.; Fonasari, G.; Vaccari, A. Fundamentals in the preparation of heterogeneous catalysts. Catal. Today **2003**, *77*, 299–314.
4. Perego, C.; Villa, P. Catalyst preparation methods. Catal. Today **1997**, *34*, 281–305.
5. Pernicone, N. Scale-up of catalyst production. Catal. Today **1997**, *34*, 535–547.
6. Le Page, J.F.; Cosyns, J.; Courty, P.; Freund, E.; Fanck, J.-P.; Jacquin, Y.; Juguin, B.; Macilly, C.; Martino, J.; Miquel, R.; Montarnal, R.; Sugier, A.; van Landeghem, H. *Applied Heterogeneous Catalysis: Design, Manufacture, Use of Solid Catalysts*; Technip: Paris, 1987.
7. Satterfield, C.N. *Heterogeneous Catalysis in Industrial Practice*; McGraw-Hill: New York, 1991.
8. Puxley, D.C.; Kitchener, I.J.; Komodromos, C.; Parkyns, N.D. The effect of preparation method upon the structure, stability and metal/support interactions in nickel/alumina catalysts. In *Proceeding of the Third International Symposium: Scientific Bases for the Preparation of Heterogeneous Catalysts*; Poncelet, Grange, G.P., Eds.; Elsevier: Amsterdam, 1983; 237–271.
9. Hu, X.D. Scientific Basis for Design of Heterogeneous Catalysts—A Study of Higher Alcohol Synthesis Catalysts, Ph.D dissertation, University of Delaware, 1989.
10. Hu, X.D.; Peng, X.D.; Toseland, B. Challenges in catalyst design and preparation for alternative energy sources. Tri-State Catalysis Society Spring Symposium. Lexington, KY, Apr. 24, 1999.
11. Courty, P.; Marcilly, C. A Scientific Approach to the Preparation of Bulk Mixed Oxide Catalyst. In *Proceedings of the Third International Symposium: Scientific Bases for the Preparation of Heterogeneous Catalysts*; Poncelet, G., Grange, P., Jacobs, P.A., Eds.; Elsevier: Amsterdam, 1983; 485–519.
12. Hu, X.D.; Wagner, J. Promoted and Stabilized Copper Oxide and Zinc Oxide Catalyst and Preparation. U.S. Patent 5,990,040, Nov. 23, 1999.
13. Williams, D.; Mishima, Y.; Rokicki, A. Dehydrogenation Catalysts. US Patent 6,465,704, Oct. 15, 2002.
14. Xie, Y.; Tang, Y. Spontaneous monolayer dispersion of oxides and salts onto surfaces of supports: applications to heterogeneous catalysis. Adv. Catal. **1990**, *37*, 1–43.
15. Chao, T.H.; Wilcher, F.P.; Ford, M.R.; Ringwelski, A.Z. Solid Phosphoric Acid Catalyst. US Patent 4,946,815, Aug. 7, 1990.
16. Goldberger, W.M.; Robbins, L.A.; Fielder, R.A.; Jepsen, T.L.B.; Knoll, F.S.; Maloney, J.O.; Mitchell, D.W.; Parekh, B.K.; Sorenson, T.C.; Stavenger, P.L.; Thelen, R.L.; Treybal, R.E.; Wechsler, I. Solid-solid and liquid-liquid systems. In *Perry's Chemical Engineer's Handbook*, 6th Ed.; Perry, R.H., Green, D.W., Maloney, J.O., Eds.; McGraw Hill: New York, 1984, section 21, 1–83.
17. Iglesia, E.; Soled, S.; Baumgartner, J.E.; Reyes, S.C. Synthesis and catalytic properties of eggshell cobalt catalysts for the Fischer-Tropsch synthesis. J. Catal. **1995**, *153*, 108–122.
18. Hu, X.D. Modification of surface morphology of cobalt based Fischer-Tropsch catalysts. Presented in the 17th North American Catalysis Society Meeting, Toronto, Canada, June 3, 2001.
19. Kung, H.H.; Kung, M.C. Heterogeneous catalysis: what lies ahead in nanotechnology. Appl. Catal. A, Gen. **2003**, *246*, 193–196.
20. Grasso, G.; Harji, B.; Belochapkine, S.; Ross, J.R.H. A new robotic system for catalyst preparation by wet or dry impregnation. Presented in EuroCombicat, Ischia, Italy, Jun 4, 2002.

21. Danstin, G.; Suslick, K. Sonochemical preparation of a nanostructured bifunctional catalyst. J. Am. Chem. Soc. **2000**, *122*, 5214–5215.
22. Hu, X.D.; Ladebeck, J. An overview of environmentally responsible technologies for solid catalyst preparation. Presented in 18th North American Catalysis Society Meeting, Cancun, Mexico, June 4, 2003.
23. Lee, H.; Zones, S.; Davis, M.E. A combustion-free methodology for synthesizing zeolite and zeolite-like materials. Nature **2003**, *425*, 385–388.
24. Mizushima, Y.; Hori, M. Alumina aerogel catalysts prepared by two supercritical drying method used in methane combustion. J. Mater. Res. **1996**, *10* (6), 1424–1428.
25. Liang, C.; Sha, G.; Guo, S. Resorcinol-formaldehyde aerogels prepared by supercritical acetone drying. J. Non-Cryst. Sol. **2002**, *271*, 167–170.
26. Takaho, S., Harako, F., Eds.; *Catalyst Preparation*; Koudansha Publisher: Tokyo, 1974.
27. Hu, X.D.; Davies, S. Preparation of high surface area hydrous titanium oxide. Presented in 214th ACS National Meeting, Las Vegas, NV, Sep. 19, 1997.
28. Sudah, O.S.; Chester, A.W.; Kowalski, J.A.; Beeckman, J.W; Muzzio, F.J. Quantitative characterization of mixing processes in rotary calciners. Powder Technol. **2002**, *126* (2), 166–173.
29. Delmon, B. Formation of final catalyst. In *Handbook of Heterogeneous Catalysis*; Ertl, G., Knozinger, H., Weitkamp, J., Eds.; Wiley-VCH: Weinheim, Germany, 1997; Vol. 1, 264–286.
30. Patterson, H.B.W. *Hydrogenation of Fats and Oils: Theory and Practice*; AOCS Press: Champaign, IL, 1994.
31. Capes, C.E. *Particle Size Enlargement*; Elsevier: Amsterdam, 1980.
32. Cybulski, A., Moulijn, J.A., Eds. *Structure Catalysts and Reactors*; Marcel Dekker: New York, 1998.
33. Nijhuis, T.A.; Beers, A.E.W.; Vergunst, T.; Hoek, I.; Kapteijn, F.; Moulijn, J.A. Preparation of monolithic catalysts. Catal. Rev. **2001**, *43* (4), 345–380.

Catalytic Combustion for Thermal Energy Generation

Daniel G. Löffler
Quarens Technologies, Inc., Bend, Oregon, U.S.A.

INTRODUCTION

Catalytic combustion became probably the most ubiquitous catalytic reaction after the implementation of the automotive converter 30 years ago. Other uses include space heaters, flammable gas detectors, and miniaturized reactors for endothermic reactions. More recently, it has been applied to reduce NO_x production in gas turbines for power generation. This article describes the engineering principles of catalytic combustion and discusses in some detail their application to gas turbines. The breakthrough inventions that allowed catalytic combustion to move forward in the field of power generation are illustrated using a simple computer model. The discussion is limited to catalytic combustion applications involving thermal energy production; exhaust gas cleanup applications, such as unburned hydrocarbon or volatile organic compound removal are not discussed. Therefore, citations are limited mostly to contributions that resulted in significant advances to the development of catalytic combustion for heat generation.

CATALYTIC COMBUSTION IN POWER GENERATION

Conventional gas turbines use diffusion flame combustors which operate at temperatures high enough for reactions of nitrogen and oxygen in air to form substantial amounts of nitrogen oxides (NO_x), predominantly nitric oxide (NO) and smaller amounts of nitrogen dioxide (NO_2) and nitrous oxide (N_2O). Low levels of nitrogen oxides in the air cause eye and respiratory tract irritation and contribute to the formation of acid rain that damages vegetation. Other ill consequences of nitrogen oxide emissions include the formation of atmospheric particles that cause visibility impairment, and the formation of toxic chemicals such as nitrate radicals, nitroarenes, and nitrosamines, that could cause biological mutations. In addition, nitrous oxide is a greenhouse gas contributing to global warming. For these reasons, current regulations mandate NO_x levels of <10 ppm in combustion effluent streams.

Fig. 1A shows a schematic representation of a gas turbine fitted with a diffusion flame combustor. The turbine drive and the air compressor are connected by the power shaft, which can also be connected to any system using the power of the turbine, such as an electrical generator. A portion of the compressed air stream is mixed with the fuel and burned in the combustor. The effluent from the combustor is mixed with the remaining air and directed to the drive, where the gases are expanded to impart rotating power to the turbine shaft. The adiabatic flame temperature of the fuel/air mixture must be high, typically around 1800°C, for the diffusion flame to be stable. The combustion flue gas is then mixed with the remaining air to lower the temperature of the mixture to 1300°C. Gases at this temperature can be safely injected into the turbine expander without damage to the turbine blades.

Diffusion flame combustors are inexpensive to build and simple to operate. Their main drawback is that they produce substantial levels of nitrogen oxides. Fig. 2 shows plots of NO_x concentration vs. temperature in a well-mixed reactor under conditions typical of turbine combustors. The level of NO_x increases exponentially with temperature, reaching over 300 ppm at the operating temperature of the diffusion flame combustor. It is clear from the figure that essentially no NO_x would be formed if the combustion could be sustained at the temperature at which the gases enter the turbine drive. This suggests that adjusting the amount of fuel to keep the adiabatic flame temperature at around 1300°C upstream of the burner could reduce the formation of NO_x. Unfortunately, the lean combustion mixtures required to reduce NO_x to single-digit ppm levels lead to unstable flames. Frequent flameout and reignition events generate vibrations that impair the operation and even the mechanical stability of the turbine. In contrast, fuel and air mixtures sustain stable combustion on a catalyst at much lower temperatures and without the requirement for a flame.

Catalytic Combustor

The fundamental features of a catalytic combustor for a gas turbine were disclosed in a patent granted in 1975, but it took three decades and several additional inventions to turn that concept into a commercial product.[2] In late 2003, the first commercial gas turbine equipped with a catalytic combustor was put into operation, providing power to 120 buildings in a health care facility in California.[3] It should be noted that the only reason to use a catalytic combustor instead of a

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120025013

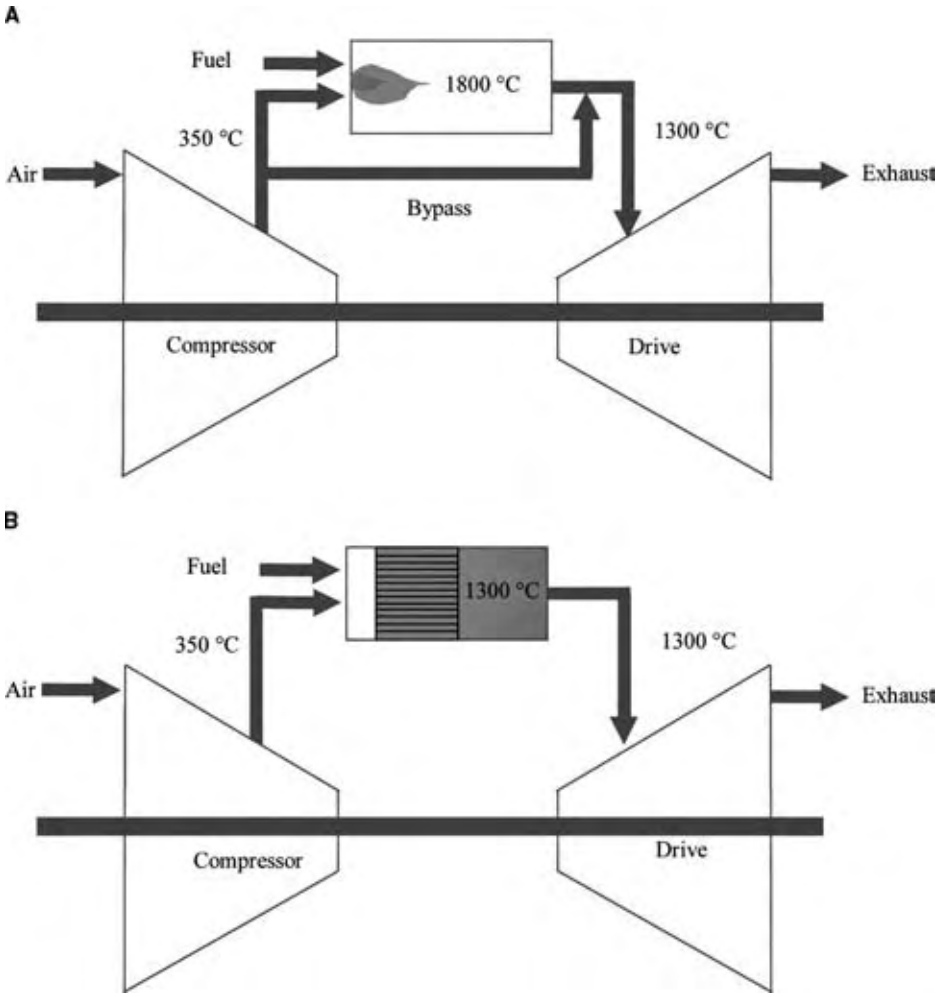


Fig. 1 (A) Schematic representation of a gas turbine fitted with a diffusion flame combustor and (B) fitted with a catalytic combustor. The bypass air stream is used to lower the temperature of the diffusion flame outlet to a level not damaging to the turbine blades. The catalytic combustor requires no bypass stream because the catalyst sustains stable combustion at 1300°C. *(View this art in color at www.dekker.com.)*

conventional diffusion flame combustor to generate the hot gas stream needed to drive a gas turbine is to reduce the level of nitrogen oxides produced in the combustion process. The use of catalytic combustion technology is then regulation-driven because there is no reduction in capital cost or gain in fuel efficiency. But without some type of nitrogen oxide abatement system, gas turbines would have been legislated to extinction in most of the industrialized world.

Fig. 1(B) shows a schematic representation of a gas turbine fitted with a catalytic combustor.[2] The entire air stream from the compressor is mixed with fuel and fed to the combustor. The adiabatic temperature of this mixture is 1300°C, therefore, the combustor exhaust gases are fed to the power turbine directly and without further dilution. Fuel is uniformly distributed over the cross-section of the combustor and mixed with air upstream of the catalyst. The catalyst is most frequently dispersed within a washcoat coated onto the surface of a monolith to minimize pressure losses and intraparticle transport restrictions.

Monolithic catalysts can be described as ''continuous unitary structures that contain many small, mostly parallel passages,'' with catalytic materials usually washcoated on the walls of those passages or channels.[4] Thus, the monolith can be visualized as a bundle of wall-coated parallel plug flow reactors. Little or no heat transfer occurs in the radial direction; hence, to ensure uniform temperatures over the cross-section of the monolith, the fuel/air ratio and the flow rates at the channel inlet must be the same for all channels. In those channels, the interplay of heat and mass transfer phenomena associated with fast and highly exothermic combustion reactions generate conditions conducive to ''hot spots'' and multiple steady states with the consequent appearance of ignition–extinction behavior.

The catalysts used in catalytic combustors must be active enough for the combustor to ignite (light off) at relatively low inlet temperatures, preferable below the compressor discharge temperature, to avoid using flame preheaters that would add complexity and possibly generate NO_x. Precious metals typically palladium and platinum show high activity for methane oxidation. Those metals are generally dispersed on porous supports to maximize the exposed surface area, and consequently, the catalytic activity per unit volume.

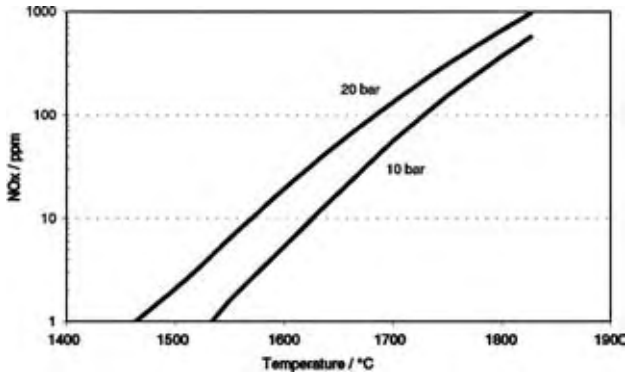


Fig. 2 Calculated NO_x levels vs. temperature for the pressures indicated. Curves were calculated using the GRI-Mech combustion kinetics database for an isothermal combustor with 20 msec residence time and an inlet composition showed in Table 1. (From Ref.[1].) *(View this art in color at www.dekker.com.)*

Early catalytic combustors built according to the teachings of reference[23] deactivated rapidly because high wall temperatures led to sintering of support materials and to evaporation of precious metals.[2] A staged combustor was suggested to solve this problem.[5] In this approach, an ignition stage, having a metal catalyst, was used to start the combustion. The subsequent stages, operating at high temperatures, carried the combustion to completion. The temperature in the ignition stage was to be controlled by simply reducing the length of the ignition catalyst so that not all the fuel was consumed. This approach failed because although exit gas temperatures were lower in the shorter ignition stage, wall temperatures were unchanged by these changes in residence time. A detailed understanding of heat and mass transfer in catalytic combustion became essential to recognize why shorter monoliths did not result in lower wall temperatures and to devise methods to control local temperatures at reactor walls.

The key interplay of reaction kinetics and transport phenomena in a catalytic combustor must be treated using rigorous reactor models. In the next section, we use a simple model to describe the behavior of a catalytic combustor and to interpret the technology breakthroughs that led to the successful implementation of catalytic combustion to reduce NO_x in power generation. The model will be kept simple, even though its additional complexities are readily incorporated, because our purpose is to show the main characteristics of a catalytic combustor rather than to provide accurate simulations of expected performance.

A mathematical model for a catalytic combustor

A number of mathematical models with various degrees of detail have been developed to simulate the behavior of monolith catalytic combustors. The simplest models reduced the simulation of the entire monolith to a description of a single channel by assuming adiabatic reactor operation, no thermal interaction between adjacent passages (thus, no net radial heat transfer), and equal flow rates in all monolith passages. The complete set of assumptions required for these models is described in reference.[4]

We shall develop next a single-channel model that captures the key features of a catalytic combustor. The catalytic materials are deposited on the walls of a monolithic structure comprising a bundle of identical parallel tubes. The combustor includes a fuel distributor providing a uniform fuel/air composition and temperature over the cross section of the combustor. Natural gas, typically >98% methane, is the fuel of choice for gas turbines. Therefore, we will neglect reactions of minor components and treat the system as a methane combustion reactor. The fuel/air mixture is lean, typically 1/25 molar, which corresponds to an adiabatic temperature rise of about 950°C and to a maximum outlet temperature of 1300°C for typical compressor discharge temperatures (~350°C). Oxygen is present in large stoichiometric excess and thus only methane mass balances are needed to solve this problem.

Assuming identical passages and uniform temperature and fuel distribution over the cross section of the monolith, we obtain the axial composition and temperature distributions by using mass and thermal energy balances over a plug-flow volume element in one channel. If we ignore homogenous gas phase reactions, all composition and temperature changes in the gas phase are due to mass and energy transfer through the washcoat interface. We assume further that there are no radiation losses, that the gas is well mixed radially across the channel, and that all resistances for mass and heat transfer between the gas phase and the washcoat can be lumped into the mass and heat transfer coefficients. The steady state mass and heat balances for the gas phase are described by two coupled ordinary differential equations similar to those for plug flow tubular reactors, except that the chemical reaction and heat generation terms are now located on the wall rather than in the bulk fluid.

$$\frac{d\xi}{dz} = \frac{S\,km(y_g - y_w)}{G} \tag{1}$$

$$\frac{d(T_g)}{dz} = \frac{ha_L(T_g - T_w)}{G\,cp_g} \tag{2}$$

where ξ is a dimensionless extent of reaction, z is the passage length, S is the cross-sectional area, km is a mass transfer coefficient, y_g and y_w are the molar fractions of methane in the bulk gas and at the washcoat

interface, respectively, T_g and T_w are the bulk gas and washcoat temperatures, respectively, G is the mass flow rate, h is a heat transfer coefficient, a_L is the area of the tube wall, and cp_g is the heat capacity of the gas. For methane combustion rates that are first order in methane and zero order in oxygen concentration, the rates of heat removal and heat generation are given by Eqs. (3) and (4), respectively.[6]

The rate of heat removal from the washcoat is proportional to the heat transfer coefficient and to the temperature difference between the gas and the washcoat, while the rate of heat generation is the product of the rate of mass transfer times the heat of reaction ($-\Delta H$), as indicated by Eqs. (3) and (4) below, where k_r is a first-order reaction rate constant. At the steady state the rates of heat generation and removal are equal.

$$Q_r = h(T_w - T_g) \tag{3}$$

$$Q_g = (-\Delta H) y_g \Big/ \left(\frac{1}{k_r} + \frac{1}{k_m}\right) \tag{4}$$

$$Q_g = Q_r \tag{5}$$

Eqs. (1), (2), and (5) must be solved simultaneously to obtain the methane conversion and temperature profiles for the monolith. Gas phase composition y_g and temperature T_g are calculated for each volume element from the mass and heat balances, Eqs. (1) and (2), and then substituted in the washcoat heat balance, Eq. (5). At the inlet of the monolith, where both the composition and the temperature of the gas phase are known, Eq. (5) can be solved independently.

Monolith with all channels coated with catalysts

Fig. 3 shows plots of Q_g and Q_r vs. T_w at the inlet of the monolith for three different gas inlet temperatures. The rate of heat generation has a sigmoidal shape, while the rate of heat removal is represented by straight lines. At low temperatures, Q_g presents an Arrhenius temperature dependence because k_r is the dominant term in Eq. (4). As the washcoat temperature increases, the process becomes mass transfer controlled, k_m dominates and the rate of heat generation becomes almost independent of temperature because of the weak temperature dependence of k_m. Eq. (5) is satisfied at the points of intersection between curves Q_g with the straight lines Q_r, which can evidently lead to more than one solution. For example, when the inlet gas temperature is 280°C, Eq. (5) is satisfied for three values of T_w. As the temperature of the inlet gas is increased, the two lower intersection points approach each other and eventually both points merge. A further

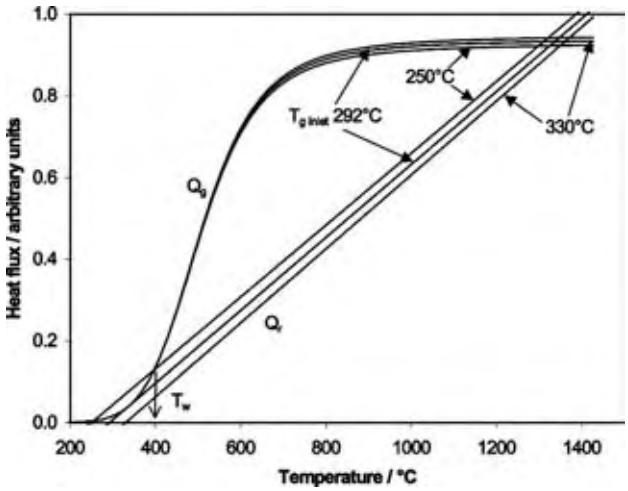


Fig. 3 Heat generated and removed at the inlet of a monolith combustor vs. temperature, calculated from Eqs. (3) and (4) for the conditions presented in Table 1. The straight lines represent the heat transfer curves in the absence of radiation losses. When the inlet gas temperature is 280°C, Eq. (5) is satisfied for three values of T_w: 297°C, 371°C, and 1326°C. As the temperature of the inlet gas is increased, the two lower intersection points approach each other and eventually both points merge at $T_w = 335$°C when the inlet gas temperature is 292°C. This is referred to as the catalytic ignition or light-off temperature. A further increase in the inlet gas temperature results in a situation where there is only one intersection point. *(View this art in color at www.dekker.com.)*

increase in the inlet gas temperature results in a situation where there is only one intersection point. The system operates then at a much higher T_w with a methane conversion nearly 100%. In practice, the inlet gas temperature should be higher than this ignition temperature to ensure that the combustor will operate with maximum conversion.

The upper branch of the sigmoidal curve corresponds to mass transfer conditions; hence, the wall temperature can be calculated by substituting Eqs. (3) and (4) in Eq. (5) and letting $k_r \to \infty$. Assuming that the resistances to heat and mass transfer can be represented by the film thickness δ_h and δ_m, respectively, we obtain, after some algebra, Eq. (6), where the Lewis number Le represents the ratio of the heat transferring capability of the gas to the rate of diffusion mass transfer. For mixtures of methane and air, $Le \sim 1$. Since $\delta_h \sim \delta_m$, the temperature of the washcoat at any point within the ignited monolith is close to the adiabatic temperature of the fuel/air mixture. For the inlet conditions in Table 1, the group $(\delta_h/\delta_m)/Le = 1.12$, and the inlet wall temperature estimated from Eq. (6) is 1371°C, as shown in Fig. 4.

$$T_w = T_g + \frac{\delta_h}{\delta_m}\frac{\Delta T_{ad}}{Le} \tag{6}$$

Table 1 Parameters used to simulate a catalytic combustor

Monolith	200 cpsi
Inlet conditions	
Pressure	12 bar
Temperature	330°C
Feed fuel/air ratio	0.0421
Adiabatic temperature rise ΔT_{ad}	933
Methane oxidation kinetics	
Activity	1740 mol/(sec cm^2 bar)
Activation energy	72 kJ/mole

Temperature and methane conversion profiles along the monolith calculated solving Eqs. (1), (2), and (5) for the conditions listed in Table 1 are also shown in Fig. 4. The washcoat temperature is highest at the inlet and then decreases slightly thereafter, asymptotically reaching the adiabatic temperature for the reactant mixture. The temperature is highest at the inlet because the concentration of methane, and hence the adiabatic temperature rise, is highest at that point. As the methane fuel is consumed, ΔT_{ad} decreases while the temperature of the gas T_g increases to become the dominant term in Eq. (8). Finally, washcoat and gas temperatures converge, reaching the adiabatic flame temperature of the fuel/air mixture after all fuel is combusted.

Fig. 4 shows that a short combustor would not lead to lower washcoat temperatures, because the washcoat reaches its maximum temperature immediately after ignition near the inlet of the channels. A short combustor would lead to incomplete methane conversion and low gas phase temperatures, both undesirable effects, without achieving the desirable effect lowering the washcoat temperature and preventing structural degradation of the catalyst. Surface temperatures reach the adiabatic temperature (~1300°C) even before detectable amounts of methane are combusted. Reducing the inlet temperature moves the ignition point farther away from the inlet, without significantly lowering washcoat temperatures (Fig. 5). A similar effect is expected from an increase in reactant flow rates.

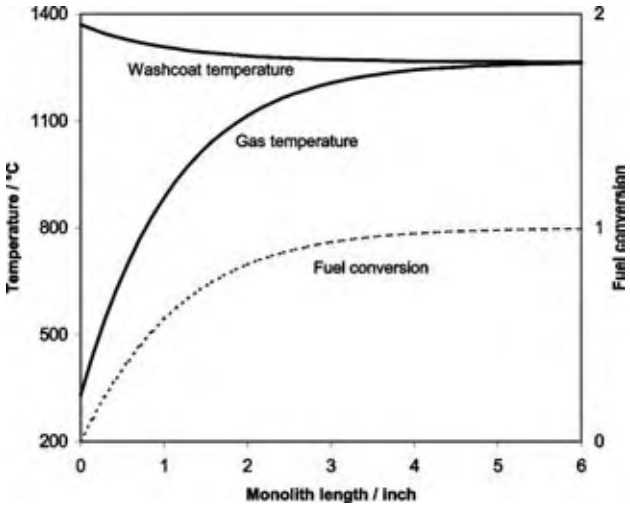


Fig. 4 Axial temperature and methane conversion distributions in a catalytic combustor calculated by solving Eqs. (1), (2), and (5) for the conditions presented in Table 1. The maximum washcoat temperature is obtained at the monolith inlet, where both gas temperature and fuel conversion are very low. Shortening the monolith would result in lower gas outlet temperatures and incomplete fuel conversion, but washcoat temperatures will remain unchanged. *(View this art in color at www.dekker.com.)*

Coupling of catalytic and noncatalytic channels by selective washcoating of a fraction of the monolith channels

An incremental improvement in the path to practical applications of catalytic combustion was disclosed by Yasuyoshi et al., who recognized the important applications of manipulating the heat balance in Eq. (5).[7] These authors coated the walls of alternate channels to have only half of the reactant flow passing through catalytic channels. For similar heat transfer coefficients in coated and uncoated channels, and assuming that there is no temperature gradient in the channel wall, the heat removal term in the heat balance is given by Eq. (7).

$$Q_r = h(T_w - T_{gcat}) + h(T_w - T_{gnoncat}) \quad (7)$$

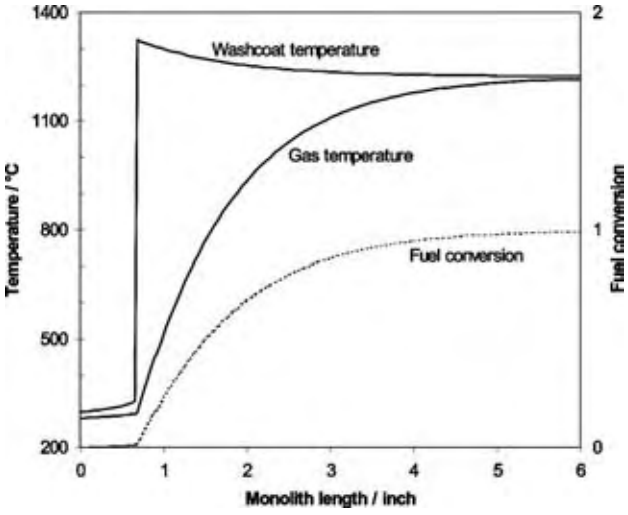


Fig. 5 Axial temperature and methane conversion distributions in a catalytic combustor. Conditions are as in Fig. 4, except that the temperature of the feed is now 280°C. The lower inlet temperature moves the ignition point downstream from the inlet of the monolith, but washcoat temperatures are not significantly reduced. *(View this art in color at www.dekker.com.)*

Because the gas flow rate is the same in both catalytic and noncatalytic channels, the gas temperatures T_{gcat} and $T_{gnoncat}$ are the same; hence, the heat removal rate is twice as fast as in the fully coated monolith. Heat and mass balances over the whole monolith show that only half the fuel fed can be consumed, even assuming full conversion in the catalytic channels. In consequence, the maximum outlet temperature of the gas phase will be the temperature for adiabatic combustion of the feed gas at 50% conversion.

Gas and wall temperature profiles and methane conversions obtained by integrating Eqs. (1), (2), and (5) after substituting Eq. (7) in the heat balance Eq. (5) are shown in Fig. 6. As for fully coated monoliths, the washcoat temperature is the highest at the inlet and decreases slightly thereafter, asymptotically merging with the gas temperature. Washcoat temperatures can be calculated from Eq. (8), which is derived by substituting Eqs. (4) and (7) in Eq. (5) and letting $k_r \rightarrow \infty$. The half-factor reflects the fact that the gas temperatures in the catalytic and noncatalytic channels are the same.

$$T_w = T_g + \frac{1}{2}\frac{\delta_h}{\delta_m}\frac{\Delta T_{ad}}{Le} \tag{8}$$

The inlet wall temperature estimated from Eq. (8) is 810°C for the conditions in Table 1, well below the value of 1371°C calculated in fully coated monoliths. The light-off temperature is 327°C, greater than the value of 292°C for the fully coated monolith, because heat transfer area is twice as large; hence, it takes a faster rate of heat generation to reach the ''runaway'' ignition conditions.

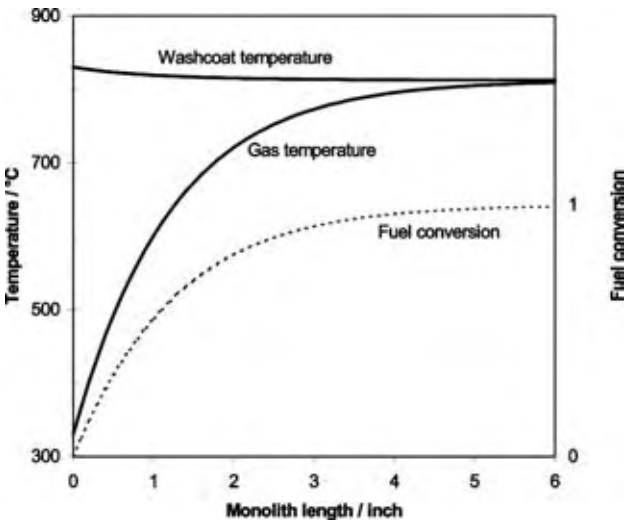


Fig. 6 Axial temperature and methane conversion distributions in a catalytic combustor with alternate channels coated. One-half of the total gas flow passes through catalytic channels. Curves are calculated solving Eqs. (1), (2), and (5) for the conditions presented in Table 1, after substituting Eq. (7) in the heat balance Eq. (5). Temperatures are much lower than in the fully coated monolith of Fig. (4). *(View this art in color at www.dekker.com.)*

Catalyst life was significantly improved by implementing the teachings of reference[7] but two problems remained.[7] First, the washcoat temperature, although much lower than in the fully coated monoliths, remained high enough to cause sintering of highly dispersed precious metals. Also, early catalytic combustors contained ceramic monoliths, which limited channel size and were easily destroyed by rapid temperature cycling. These problems were addressed in a patent application by Retallick and Alcorn, which disclosed corrugated metal foil monoliths and procedures to assemble those monoliths by coating one side of the metal foil strip and then rolling it to form a channel structure with alternating catalytic and noncatalytic channels.[5] Metal foils with chevrons or herringbone corrugations are coated with catalyst on one side, laid one on top of the other, and then rolled into a cylinder to form a monolith structure. This monolith contains alternating coated and uncoated channels with the walls of the coated channels cooled by the uncombusted fuel–air mixture flowing in the uncoated channels. The metal strips are able to expand thermally in both length and width and these metal monoliths are far more tolerant to thermal shock than ceramic structures.

Retallick and Alcorn also showed that similar structures could be built up by alternating catalyst-coated flat strips with uncoated strips having straight or other corrugations. By varying the height of the corrugations or the number of noncatalytic channels between catalytic channels it is possible to design structures where more gas would flow through noncatalytic channels than through catalytic channels. The temperature of the gas in the noncatalytic channels will then be lower than in the catalytic channels. The mixing-cup temperature of the gas leaving the monolith can be calculated from a heat balance over the volume of the monolith. For example, if the gas flow in the coated channels is 30% of the total gas flow rate, assuming full conversion in the catalytic channels the mixing-cup temperature of the outlet gas will be equal to the temperature for adiabatic combustion of the feed gas at 30% conversion.

Gas and wall temperature profiles and methane conversions, calculated integrating Eqs. (1) and (2) after substituting Eq. (7) in the heat balance Eq. (5), are shown in Fig. 7 for a monolith in which 30% of reactant inlet flow occurs in catalytic channels. As for fully coated monoliths, the washcoat temperature is highest at the channel inlet and decreases slightly thereafter, asymptotically reaching the gas temperature. Washcoat temperatures become much lower in

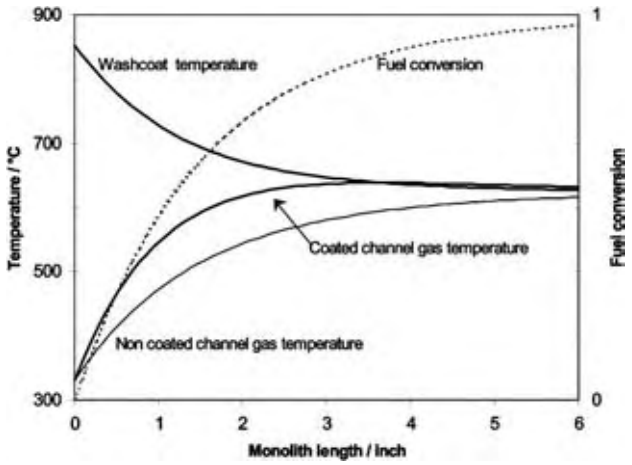


Fig. 7 Axial temperature and methane conversion distributions in a catalytic combustor with alternate channels coated. Thirty percent of the total gas flow passes through catalytic channels. Curves are calculated solving Eqs. (1), (2), and (5) for the conditions presented in Table 1, after substituting Eq. (7) in the heat balance Eq. (5). Gas and wall temperatures are considerably lower than in the case of Fig. 6 except for the inlet wall temperature, which is independent of the fraction of flow passing through catalytic channels when only one wall of the channels is coated. *(View this art in color at www.dekker.com.)*

monoliths with half of the channel coated at all points in the combustor, except at the inlet. At this point, the temperature of the gas is the same in both catalytic and noncoated channels and the washcoat temperature is governed by Eq. (8). It follows that the temperature of the washcoat at the inlet of the partially coated monolith is largely independent of the fraction of coated channels. For similar reasons, the light-off temperature at the inlet of a partially coated monolith does not change with the fraction of coated channels.

As indicated above, the inlet temperature obtained using partially coated monoliths, although substantially lower than those observed in fully coated systems, may not be low enough to prevent sintering of the highly dispersed metal catalysts. A new invention was needed to reduce the temperature of the leading edge of the monolith to allow low-temperature operation of highly active catalysts.

Partially coated monolith with diffusion barrier layer

The temperature of the leading edge of the monolith is governed by Eq. (8); hence, to lower the wall temperature (T_w) the terms in the right-hand side of the equation need to be reduced. Reducing the inlet gas temperature would require cooling down the inlet air stream, an operation that would impact on the energy efficiency of the gas turbine. Also, if the inlet gas temperature (T_g) were reduced, more active catalysts would be needed to maintain light-off at reasonable inlet temperatures. Those catalysts must be even more dispersed and, consequently, more sensitive to sintering than less active catalysts used at higher inlet temperatures. The adiabatic temperature rise ΔT_{ad} cannot be reduced, because that would decrease turbine efficiency, and the Lewis number cannot be varied because it reflects intrinsic transport properties of the fuel–air mixture. Hence, the only parameter left to change is the ratio of transport resistances δ_h/δ_m. This ratio can be manipulated by coating an inert porous layer on top of the washcoat, adding heat and mass transfer resistances in series with those of the boundary layer. In the porous layer, the resistance to mass transfer is larger and the resistance to heat transfer is smaller than in the boundary layer because mass transfers through the void fraction only, while heat transfers through gas and solid and the heat conductivity of the solid is much larger than that of the gas. The net effect of adding an inert porous layer is then to reduce the washcoat temperature by decreasing the mass flow to the catalytic washcoat without significantly changing the rate of heat transfer.

Dalla Betta et al. first proposed an inert porous layer, or diffusion barrier, to prevent temperature runaway, and loosely interpreted the effect in terms of a reduction in the rate of combustion.[8] A more rigorous interpretation of the effect of an inert porous layer on catalyst temperature was provided by McCarty et al, who also described the desired properties for diffusion layer materials, including a high thermal conductivity and low specific combustion activity.[9] These authors stated that the high washcoat temperatures found in catalytic combustion of natural gas were due to the high diffusivity of methane in air, which causes the diffusion rate to the catalyst surface to match the rate of heat dissipation by conduction to the gas phase. The diffusion barrier decreases the rate of diffusion of methane to the catalyst surface, thus reducing the catalyst temperature. Modeling work by Hayes et al. confirmed those concepts.[10]

It should be noted that the porous layer has no effect on the light-off temperature at the leading edge of the monolith. The onset of light-off occurs at low temperatures, when the reaction rate constant k_r dominates the denominator of Eq. (4). By coating a small fraction of the active channels and providing a porous layer, the temperature of the washcoat can be kept low enough to prevent from catalyst degradation while allowing light-off at compressor discharge temperature.

The effect of a diffusion barrier layer on combustor performance is evident in the model calculations shown in Fig. 8. Temperatures at the monolith inlet are dramatically lower than those presented for the

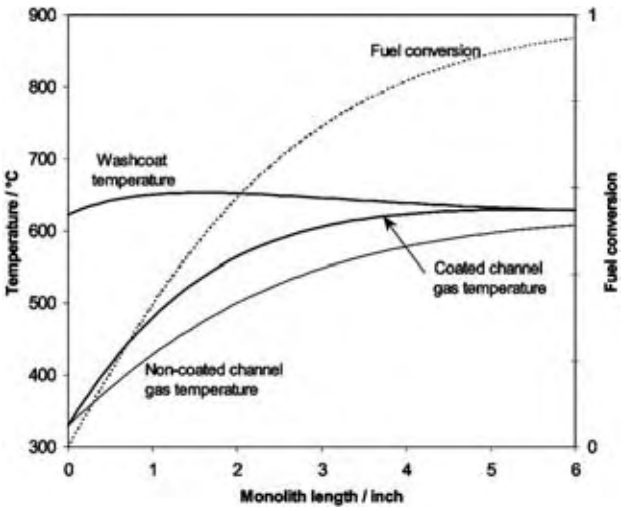


Fig. 8 Axial temperature and methane conversion distributions in a catalytic combustor with alternate channels coated and a diffusion barrier layer deposited on top of the washcoat. The thickness of the diffusion barrier layer decreases exponentially downstream from the inlet. Thirty percent of the total gas flow passes through catalytic channels. Curves are calculated solving Eqs. (1), (2), and (5) for the conditions presented in Table 1, after substituting Eq. (7) in the heat balance Eq. (5). Gas and wall temperatures are close to those in Figure 7 except for the inlet wall temperature, which is much lower now because the diffusion barrier layer limits the diffusion rate of methane to the washcoat. *(View this art in color at www.dekker.com.)*

uncoated monolith in Fig. 7, while fuel conversion and outlet gas temperatures do not change significantly.

Homogenous combustion after partially coated monoliths

Methane reactants within noncatalytic channels remain unburned and they must be combusted in a fully catalytic stage or in a homogenous flame. A fully catalytic monolith would cause the catalyst to reach adiabatic temperatures and to deactivate. Thus, the only practical option is to complete the combustion in a homogenous combustion process.

The partially combusted gases exiting the monolith will ignite after a period of time that depends on their temperature and composition at the outlet of the monolith. The temperature of the fuel–air mixture at that point must be high enough that the homogenous combustion is completed within residence times typical of turbine combustors (~20 msec). Longer residence times might result in combustors too big to fit within the space limitations of gas turbine enclosures. Homogenous kinetic simulations indicate that >300 msec residence time is required to combust the outlet gases from the monolith (Fig. 8) to CO levels below 10 ppm. Clearly, higher monolith exit temperatures are required to ensure complete homogenous combustion. One approach would be to use multiple catalytic combustor stages in which the channels in the first stage are coated with highly dispersed metal catalysts for low-temperature light-off. To protect these highly dispersed metals, the first-stage monolith should have a small fraction of coated catalysts and an inert diffusion layer, such as in the case of Fig. 8. The outlet gases from the coated and noncoated channels are mixed at the outlet of the first stage and then introduced into a second stage, in which more methane is converted and the stream temperature rises to levels required for fast homogenous ignition. The catalyst in this second stage need not be as active as in the first stage because the higher inlet temperatures allow much faster light-off along the channel. Less active catalysts are generally poorly dispersed and more tolerant of high temperatures.

A two-stage combustor operating under the inlet conditions as shown in Table 1 and using monoliths with 30% and 40% of the flow though catalytic channels in the first and second stages, respectively, will have second stage outlet gas temperature of 888°C. The time evolution of CH_4 and CO concentration and temperature in the homogenous combustion process after the second catalytic stage, calculated using the GRE-Mech kinetics, is shown in Fig. 9.[1] The sharp drop in methane concentration accompanied

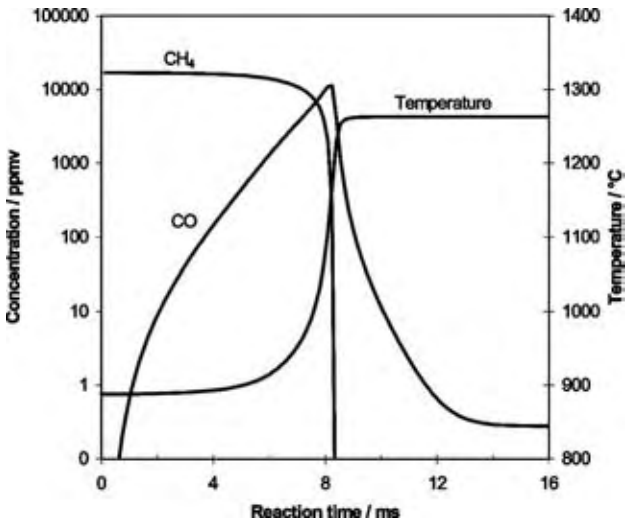


Fig. 9 The time evolution of CH_4 and CO concentration and temperature in an adiabatic flame calculated using the GRE-Mech kinetic mechanism. The flame temperature in this example is 1263°C, but this design allows for much higher flame temperatures, if needed. (From Ref.[1].) *(View this art in color at www.dekker.com.)*

by a steep rise in temperature at 8 msec denotes flame ignition. Combustion of CO to a level below 1 ppm occurs at 12 msec. Only a very small amount of CO is produced in the catalytic stages, most of the CO forms before ignition in the homogenous combustion stage.[11]

Catalytic combustion catalyst

Palladium oxide supported on metal oxides such as zirconia with various additives is the catalyst of choice for the combustion of natural gas in gas turbines because of its high activity for methane oxidation, which leads to light-off at low temperatures, and its low volatility, which minimizes sintering rates and evaporative metal losses. This catalyst presents an unusual situation in that the thermodynamically stable phase can be either Pd metal or PdO, depending on the oxygen partial pressure and the temperature. Low temperatures and high oxygen concentrations favor the oxide phase, which shows higher combustion activity than metal surfaces. Because the metal-oxide phase transition occurs under conditions of oxygen concentration and temperature found in turbine combustors, the mechanism of PdO formation and decomposition in a reaction stream containing methane and oxygen, and the kinetics of methane oxidation on both oxide and metal phases, have received considerable attention.[12] Thermodynamic calculations indicate that PdO is stable in the first-stage monolith of Fig. 6, while Pd metal is favored in the second-stage monolith. The activity of Pd metal is considerably lower than that of the oxide, but it is still high enough to allow light-off and to maintain the combustion reaction under mass-transfer control. For this reason, the peculiar behavior of the rate of methane oxidation versus temperature is largely invisible to the users of catalytic combustors.

OTHER APPLICATIONS OF CATALYTIC COMBUSTION

Catalytic combustion is used in radiant heaters for appliances such as stoves for outdoor use, that take advantage of the insensitivity of catalytic burners to windy conditions, and in industrial applications such as paint drying and plastic thermoforming, where the heat source operates in potentially explosive atmospheres. An interesting application of catalytic combustion in a consumer product is the cordless curling iron that works with a small cartridge of butane fuel. In this device, heat production and dissipation are matched to yield the desired temperature of 55°C with a fuel that could have an adiabatic flame temperature higher than 2000°C.[13] Catalytic combustion devices are also used to detect flammable gases in air. The concentration of the gas is measured through the heat released by combustion, either by placing a thermocouple on a dispersed metal catalyst or by measuring the electrical resistance of a catalytic metal wire.

Endothermic reactions, such as steam reforming, are usually carried out in long narrow tubes filled with catalysts and externally heated by flames. The heat could be provided more uniformly and more accurately at the necessary level by a combustion catalyst coated on the outside of the tubes, and heat transfer rates could be further improved by coating the endothermic reaction catalyst on the inner wall of the tube. In this way, the heat of combustion is transferred to the heat sink (the endothermic reaction) through the solid wall, avoiding solid–gas heat transfer resistances. However, the tubular geometry is not most efficient for this application because of the difficulty to coat the inside of the tubes and the need to include static mixers to facilitate mass transfer to the catalytic surfaces.

A preferred embodiment of this concept is the catalytic plate reactor, which consists of catalytically coated metal plates so that exothermic and endothermic reactions take place in alternate channels. In addition to minimizing the heat transfer resistances, this reactor facilitates mass transfer to the catalytic surface by reducing the diffusion length. Catalytic plate reactors can find applications in steam reforming, dehydrogenation, and hydrocarbon cracking which are strongly endothermic processes. In recent years those reactors have received considerable attention as steam reformers for fuel cell applications.[14]

ISSUES WITH CATALYTIC COMBUSTION

There are a few drawbacks associated with catalytic combustion. First, as noted by Pfefferle, the power density is low.[15] The volumetric heat release rates of catalytic combustors (without a homogenous flame downstream) are much lower than those found in conventional flame combustors because catalytic combustors are mass transfer limited, and mass transfer coefficients are relatively low.

Catalyst durability remains a concern, even after early problems related to thermal stability have been solved allowing turbine operation for >8000 hr.[3] Airborne impurities, such as silicones (organo-silicon compounds containing the functional group $Si(OH)_x$) can lead to stoichiometric deactivation. When a silicone molecule burns on an active site, the carbon and hydrogen atoms react to form carbon oxides and water leaving a silicon residue to deactivate the active site. Metal-containing organic molecules may cause the same effect. Rigorous filtering of the intake air is absolutely needed to ensure stable operation.

A major consideration in power generation applications is the sensitivity of the catalytic combustor to homogenous combustion in noncatalytic channels. If for some reason (for example, a momentary reduction in gas flow rate) the homogenous combustion backs up to the exit of the second-stage monolith, the gas in the uncoated channels may ignite. The adiabatic temperature of the gas at the outlet of the noncatalytic channels is the temperature of the gases at the outlet of the second stage plus the adiabatic temperature rise of the gases at the inlet of that stage. For the two-stage combustor discussed above, operating with the inlet conditions of Table 1 and using the monolith of Fig. 8 in the first stage, the adiabatic temperature rise of the feed to the second stage is 643°C. Since gas at the outlet of the second stage monolith is at 888°C, the gas in the uncoated channels will burn with an adiabatic flame temperature of 1531°C, effectively melting the metal monolith. Thus, catalytic combustion is very unforgiving of flow rate fluctuations; it takes only one accident to destroy the combustor.

This problem can be circumvented in a fuel-rich approach to catalytic combustion for gas turbines recently proposed.[12] In this method fuel is mixed with air to form a fuel-rich mixture that is reacted over the catalyst to produce both partial and total oxidation products. The reaction products are then mixed with excess air and burned in a homogenous flame. Because the gases exiting the catalyst are fuel-rich, they cannot sustain combustion in the event of a homogenous flame backup. The promise of this method needs to be confirmed in full-scale turbine tests.

CONCLUSIONS

Advances in chemical reaction engineering and catalytic materials have allowed catalytic combustion for thermal energy generation to be commercialized in consumer and industrial applications. The development of catalysts coated on one side of a metal substrate, coupled with the use of diffusion barriers, has allowed controlling the combustion temperature to suit diverse applications. Catalytic materials have been developed to remain active for thousands of hours under conditions deemed too severe just a few years ago. We may expect that the need for clean distributed power increases the demand for gas turbines fitted with catalytic combustors and promotes the development of catalytic burners to be used in fuel processors for fuel cell power systems.

REFERENCES

1. Smith, P.R.; Golden, D.M.; Frenklach, M.; Moriarty, N.W.; Eiteneer, B.; Goldenberg, M.; Bowman, C.T.; Hanson, R.K.; Song, S.; William, C.; Gardiner, W.C., Jr.; Lissianski, V.V.; Qin, Z; http://www.me.berkeley.edu/gri_mech/ (accessed Oct 2003).
2. Pfefferle, W.C. Catalytically-supported thermal combustion US Patent 3,928,961, Dec 10, 1975.
3. http://www.catalyticaenergy.com/xonon/commercialization.html (accessed Oct 2003).
4. Cybulski, A.; Moulijn, J. Monoliths in heterogeneous catalysis. Catalysis Rev. Sci. Eng. **1994**, *36*, 179–270.
5. Retallick, W.B.; Alcora, W.R. Combustion apparatus for high-temperature environment US Patent 5,202,303, Apr 13, 1993.
6. Au-Yeung, J.; Chen, K.; Bell, A.T.; Iglesia, E. Isotopic studies of methane oxidation pathways on PdO catalysts. J. Catalysis **1999**, *188*, 132–139.
7. Yasuyoshi, K.; Kunihiko, K.; Masao, O. Catalyst body for combustion Japan Patent 59136140, Aug 4, 1984.
8. Dalla-Betta, R.A.; Tsurumi, K.; Shoji, T. Graded palladium-containing partial combustion catalyst US Patent 5,258,349, Nov 2, 1993.
9. McCarty, J.G.; Wong, V.L.; Wood, B.J. Catalytic combustion process U.S. Patent 6,015,285, Jan 18, 2000.
10. Hayes, R.E.; Kolaczkowski, S.T.; Li, P.K.C.; Awdry, S. The palladium catalysed oxidation of methane: reaction kinetics and the effect of diffusion barriers. Chem. Eng. Sci. **2001**, *56*, 4815–4835.
11. Dalla Betta, R.A.; Löffler, D.G. Selectivity considerations in methane catalytic combustion. In *Heterogeneous Hydrocarbon Oxidation*; Warren, B.K., Oyama, S.T., Eds.; ACS Symposium Series 638; American Chemical Society: Washington, DC, 1996; 36–47.
12. Forzatti, P. Status and perspectives of catalytic combustion for gas turbines. Catalysis Today **2003**, *83*, 3–18.
13. Saint-Just, J.; der Kinderen, J. Catalytic combustion: from reaction mechanism to commercial applications. Catalysis Today **1996**, *29*, 387–395.
14. Zanfir, M.; Gavriilidis, A. Catalytic combustion assisted methane steam reforming in a catalytic plate reactor. Chem. Eng. Sci. **2003**, *58* (17), 3947–3960.
15. Pfefferle, L.D.; Pfefferle, W.C. Catalysis in Combustion. Catalysis Rev. Sci. Eng. **1987**, *29* (2–3), 219–267.

Catalytic Cracking

Paul O'Connor
Akzo Nobel Catalysts, Amersfoort, The Netherlands

INTRODUCTION

Catalytic cracking is a very flexible process used to reduce the molecular weight of hydrocarbons. Today, fluid catalytic cracking (FCC) remains the dominant conversion process in petroleum refineries. Although FCC is sometimes considered to be a fully matured process, new challenges and opportunities in its application and a continuing stream of innovations in the process and catalyst field ensure that it will remain an important and dynamic process in the future of refining.

HISTORICAL PERSPECTIVE

Prior to 1925, the higher-boiling heavy crude oil molecules were chemically changed to smaller naphtha (gasoline) molecules by thermal decomposition using a process called thermal cracking. In the late 1920s, Eugene Jules Houdry demonstrated that the catalytic cracking process yields more gasoline with higher octane content. The first full-scale commercial fixed-bed catalytic cracking unit began production in 1937.

PROCESS DEVELOPMENT

During catalytic cracking, the catalysts are covered after a short time by a deactivating layer of coke. This coke can be removed and regenerated by burning, but the regeneration time is relatively long compared to the reaction time. An efficient way to solve this problem is to move the catalyst from one reactor (for hydrocarbon cracking) to another reactor (for catalyst regeneration).

The first continuous circulating catalyst process using bucket elevator thermofor catalytic cracking (TCC) was started up in Paulsboro, New Jersey, in 1941. The moving bed solved the problem of moving the catalyst between efficient contact zones. However, the catalyst beads used were still too large, limiting the regenerator temperatures to avoid intraparticle heat exchange and therefore requiring a large regenerator and catalyst holdup.[1]

Moving the solid catalyst in this way remained a challenge, which was solved by making use of the theory that it is possible to make a powder flow in a manner similar to a liquid if enough gas flows through it. This phenomenon is called fluidization, and the FCC process was introduced, which uses fine powdered catalysts that can be fluidized. The first commercial circulating fluid bed process went into production in 1942 in Baton Rouge, Louisiana.[2] By the 1970s, FCC units replaced most fixed- and moving-bed crackers.

Generally, FCC units operate in a heat-balanced mode whereby the heat generated by the burning of coke is equal to the heat needed for the vaporization of the feed plus the heat of cracking. Also, the pressure balance of an FCC unit is very important to ensure proper catalyst circulation and to prevent contact between the hydrocarbons (reactor) and air (regenerator). Overall, this makes the optimal operation of a unit a very interesting challenge.

The FCC process equipment and operations have continued to coevolve with the catalyst and the changing economical and environmental requirements. Key developments in FCC processes and machinery are:[1,3]

- Short-contact-time riser reactors.
- Feed distribution and atomization.
- Feed prevaporization or "supercritical" injection.
- Multiple feed injection.
- Quick product disengagement and separation from catalyst.
- Quick product quenching.
- More efficient stripping.
- Downer (downflow) reactors.
- Improved regenerator efficiency and lower inventory.
- Improved control of combustion (reduction in CO, CO_2, SO_x, NO_x).
- Improved air-grid designs.
- Catalyst coolers (internal and external heat removal).
- Power recovery from flue gas.
- Improved high-flux standpipes.
- High-efficiency cyclone separators.
- Cyclones without diplegs.
- Third-, fourth-, and fifth-stage particulate capture systems.
- Erosion- and high-temperature-resistant metallurgy.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007694

These developments have led to dramatic reductions in the size (elevation, volume, catalyst inventory) and, hence, costs of an FCC unit per barrel of feed charged and converted.

CHEMISTRY AND KINETICS

The catalytic cracking of hydrocarbons is a chain reaction that is believed to follow the carbonium ion theory, involving three steps: initiation, propagation, and termination. The initiation step is represented by the attack of an active site on a reactant molecule to produce the active complex that corresponds to the formation of a carbo-cation. Chain propagation is represented by the transfer of a hydride ion from a reactant molecule to an adsorbed carbonium ion. Finally, the termination step corresponds to the desorption of the adsorbed carbenium ion to give an olefin while restoring the initial active site.[4]

The carbenium-ion cracking mechanism produces a higher yield of a much more desirable gasoline than thermal cracking. While thermally cracked gasoline is quite olefinic, catalytically cracked gasoline contains a large amount of aromatics and branched compounds, which is beneficial for the gasoline's octane numbers (research octane number—RON, and motor octane number—MON). Table 1 illustrates the differences in the kinetics of thermal and catalytic cracking at about equal conversion. The data indicate that even for catalytic cracking, wide differences are possible. Only carbenium-ion cracking involving a tertiary carbenium ion produces branched compounds. A second type of cracking, protolytic[5] cracking can take place, yielding more ''radical cracking''-like products such as methane and ethane (fuel gas).

CATALYSTS

Table 2 lists the types and forms of cracking catalysts developed and used over the years. The way in which catalysts are built up from the separate components (catalyst assembly) and the form of the catalysts have a significant impact.[6]

The first FCC catalysts, super filtrols, were produced by activating clays with acid. These materials were originally used to bleach edible oils and decolorize hydrocarbons. Synthetic mixed oxide catalysts followed, some of which were two to three times more active than the activated-clay-based types. The Al_2O_3 of $SiO_2{\cdot}Al_2O_3$ was optimized by impregnating dry SiO_2 gels in the 10–25% Al_2O_3 range. The high-Al_2O_3 catalysts (25% Al_2O_3) exceeded the low-Al_2O_3 catalysts and super filtrol steamed activity level. Later on, clay was added in the preparation of synthetic $SiO_2{\cdot}Al_2O_3$ catalysts to provide additional macroporosity.

Cocurrently, microspheroidal (MS) catalysts were developed, recognizing the advantages of a catalyst with higher alumina content and improvements in impregnation efficiency with small particles compared to the traditional lumps of silica hydrogel. The original process was quite burdensome and involved an emulsion process, which was soon replaced by spray drying of the impregnated gel. Spray drying is still the way all FCC manufacturers compound and form their MS catalysts.

In the 1950s, zeolites were invented, and their potential application in catalysis received great deal of attention.[7] In the 1960s, Mobil[8] introduced zeolites into FCC catalysts, leading to very substantial increases in conversion and gasoline production, as shown in Table 3.

The first-generation zeolite catalysts were based predominantly on $SiO_2{\cdot}Al_2O_3$ gels (''matrix'') in which the zeolite was added at some point prior to spray drying. In the ''in situ'' crystallization method as applied by Engelhard, kaolin-based microspheres are prepared and calcined, after which zeolites are crystallized in the microspheres, leaving zeolite in an ''Al_2O_3-enriched matrix.''

In the early 1970s, Grace Davison introduced the use of a silica hydrosol-based binder for the incorporation of zeolites. Silica hydrosol is polymerized silica dispersed in water to form a clear continuous phase; it is a binder giving dramatic improvements in attrition resistance and density. The result is the creation of particles that are encased in a hard, resilient shell of a vitreous material. The low activity of the silica hydrosol relative to the $SiO_2{\cdot}Al_2O_3$-based systems enhanced the selectivities of nearly pure zeolite cracking, resulting in further improvements in gasoline and coke yields. Nevertheless, alternate routes for using Al_2O_3

Table 1 Catalytic vs. thermal cracking

Conversion	*i*C5/*n*C5 ratio	(C1 + C2)/*i*C4 ratio	Fuel gas (mol/mol cracked)
Thermal	0	66	2
Activated carbon	0.06	27	0.54
Alumina	0.2	14	2.72
Silica–alumina	3.8	0.6	0.41

Table 2 Early days of cracking catalysts

Year	Process	Reactor system	Catalyst type	Catalyst form
1920	McAfee	Batch	$AlCl_3$	Granulated
1939	Houdry	Fixed bed	Clay	Acid treated Granulated
1940	Suspensoid	Liquid phase	Clay	Ex-lube oil decolorizing Powdered
1942	FCC	Fluid bed	Clay	Super filtrol Acid treated Powdered
1945	TCC, Houdry flow	Moving bed	Clay	Acid treated Pellets
1942	FCC	Fluid bed	Synthetic $SiO_2{\cdot}Al_2O_3$	Ground
1946	FCC	Fluid bed	Synthetic $SiO_2{\cdot}Al_2O_3$	Microspheres
1965	FCC	Fluid bed	X, Y zeolites	Microspheres

gel and/or $SiO_2{\cdot}Al_2O_3$ gels as binders and as catalytic functional materials are still being pursued because of the significance of the porosity and permeability of the microspheres.

The importance of diffusion restrictions in FCC catalysis is often questioned. Short-contact-time pilot riser experiments confirm that combining zeolites with ''diffusion-enhancing'' matrices can result in significant product selectivity and product property improvements. The industrial benefits of a nonzeolite ''matrix'' on bottoms cracking in heavy vacuum gas oil (VGO) and resid FCC (RFCC) have also been confirmed in practice. Recently, some new methods have been used[6,11] to measure the accessibility of FCC catalysts.

Besides faujasite (Y) zeolites, today's catalysts contain several additional functional materials,[9] such as metal traps, nickel-resistant matrices, bottoms-cracking matrices, and small pore zeolites, such as, for instance, ZSM-5. These zeolites are often added as separate (additive) particles with the intention of boosting the gasoline octanes and/or the production of light olefins (propylene).

Also, separate FCC additives are produced, usually consisting of metals (Pt, Pd, Ce, V, Cu, Co, Zn) on alumina and/or alumina–magnesia supports for CO_2, CO, SO_x, and NO_x control of the regenerator off-gas[10] and for sulfur reduction of gasoline.

Table 3 Improvements in zeolite cracking catalysts

Catalyst	Conversion (vol.%)	Gasoline (vol.%)
Silica–alumina gel	56	40
REHX	68	52
REHY	75	58

Catalyst Aging and Deactivation

FCC catalysts are deactivated via several mechanisms,[11] all of which result in loss of activity and change in yield selectivity. Catalysts age, meaning they undergo change in chemical and physical structure due to the (hydro-) thermal conditions during the 10,000–50,000 reaction and regeneration cycles they endure. Catalysts can also be poisoned, whereby their active sites are covered by coke and/or polar molecules (nitrogen), which neutralize catalytic activity, or by metals (V, Ni, Na), which can destroy or alter the activity. Catalysts can also be deactivated by fouling, whereby coke and/or metals are deposited and block the catalyst pores, thereby limiting mass transfer.

Reversible deposits (nitrogen, coke) are removed during regeneration. If we assume that the poisoning effect increases with the concentration of the deposits, then the poisoning effect is inversely proportional to the catalyst-to-oil ratio, and therefore is dependent on the coke selectivity of the catalyst. Irreversible catalyst poisons (metals), on the other hand, build up and continue to interact with the catalyst.

The very detrimental effects of contaminants such as Fe and Ca on the accessibility and performance of catalysts has been reported.[6,12,13] Apparently, these contaminants can result in (liquid) eutectic melts on the surface of the catalyst particles, which can block their entrance pores and even cover the whole catalyst surface.

FEEDSTOCKS, PRODUCTS, AND THE ENVIRONMENT

To strike a balance between product demand and refinery feed composition, more residue with high

metal content and Conradson carbon residue (CCR) is being included in FCC feedstocks. Improvements in process and catalyst technology have resulted in a further opening of the feedstock processability window.

Environmental regulations are becoming a key driving force for reducing FCC process air-pollutant emissions[10] and for changing the composition of fuel products.[9] This is affecting the design and operation of FCC and is providing new opportunities for the development of catalyst and additive technology.

FUTURE ROLE OF FCC IN REFINING: FEEDSTOCK CHALLENGES

Fluid catalytic cracking is sometimes said to be an evolved process. However, new challenges and opportunities are ever-present in its application, and constant innovations ensure that it will be crucial in the future of refining.

From the beginning, FCC was employed to upgrade atmospheric (or long) residue. By distillation under vacuum, FCC feedstock (also called VGO) was separated from vacuum (or short) residue. Distillation techniques have been extended to increase the yield of FCC feedstock, and this development (revamps of vacuum units) is still taking place at various refineries. This deeper distillation does sufficiently restrict metals (Ni, V, etc.) and nonreactive and coke-producing components (CCR) in the distillate FCC feedstock. In some cases, deasphalting was and is employed to further increase the yield of acceptable FCC feedstock.

With the advent of better catalysts and effective additives, FCC feed preparation by vacuum distillation has become redundant for certain low-metal/CCR crude oils, so that the atmospheric residue can be processed in an FCC unit directly. The range of crude oils from which the atmospheric residue can be processed ''directly'' can be extended by revamping existing FCC units or building special new FCC units, RFCC units, and/or heavy-oil crackers.

The main feature of such an RFCC revamp is increasing the coke-burning capacity of the regenerator while maintaining the FCC unit in heat balance (often by incorporating catalyst coolers in the regenerator). Also, the employment of improved catalysts that reduce coke formation and enhance tolerance for metals (notably V, Fe, and Na) has enabled this development in RFCC.

In the refining world, the FCC process as a residue upgrader has to compete with other catalytic processes (such as hydrocracking) and thermal processes (such as thermal cracking, visbreaking, and coking). During the period 1983–2003, RFCC emerged from a very small base to deployment of nearly 4 million barrels a day, while several other new processes failed and passed from the scene.[14] Among the trends over the last few years are:

- Continued large increase in delayed coking capacity, particularly for processing of the lowest-quality vacuum residues.
- Significant increases in RFCC capacity in both grassroots facilities (China, India, Middle East) and FCC revamps.
- Continued growth and extension of the use of hydrodesulfurization and hydrodemetallization (HDS/HDM) processes to treat poor-quality residues for the preparation of RFCC feedstocks.
- Broader commercial application of residue hydrocracking technologies to vacuum residues, especially with ebullating bed processes.
- Continued rapid expansion of hydrogen manufacture to meet the needs of increased residue conversion capacity and to meet environmental needs for transportation fuels.
- The emergence of coproduction, via residue gasification, of hydrogen and power in refinery settings.
- A dramatic expansion of ''field'' upgrading operations in the oil sand bitumen deposits in Canada and the extra-heavy-oil Orinoco deposits in Venezuela.

An overview of the worldwide residue processing capacity is show in Table 4, taken from a recent study[14] considering the units in operation by March 2003.

An RFCC in this study is defined as an FCC that can handle feedstocks containing more than 2% CCR. For RFCC to continue to compete in this arena, the range of crude oils/residues that can be processed needs to be increased; only the crude oil/residues with the highest CCR feeds should be left for cokers.

FUTURE ROLE OF FCC IN REFINING: ADVANCED FUELS

In recent decades, product quality development has been considerable, starting with the phase-out of lead in gasoline in the last decades of the 20th century and the start of the phase-down of the sulfur level in diesel and gasoline from the 1990s. In this century, the further phase-out of sulfur in both gasoline and diesel continues. For gasoline, the main source of sulfur is FCC gasoline, and desulfurization has to be executed with care so as not to reduce octane quality. Various licensors have found commercially viable solutions. Another option that is available to refiners is the

Table 4 Worldwide residue processing capacity (million barrels per day)

Conversion technology	United States	Canada/Mexico/ Venezuela	Japan	Europe	Rest of the world	World total	% Distribution
Thermal							
Cracking/visbreaking	44	331	24	2,260	1,635	4,293	26
Coking	2,245	951	66	673	1,169	5,104	31
Deasphalting	283	39	16	46	75	458	3
Hydroprocessing							
Fixed bed	499	30	591	149	1,042	2,312	14
Ebullating bed	102	244	23	79	49	497	3
Slurry phase	—	4	—	—	—	4	0
RFCC	831	281	318	681	1,832	3,942	24
Total	4,002	1,879	1,037	3,889	5,801	16,609	100
% of crude capacity	24	30	22	15	19	20	

use of a sulfur reduction FCC catalyst and/or additive. This option is not expected to bring down the sulfur level sufficiently, but it may be considered in a transient period and/or can serve to enhance feedstock flexibility.

A point of great concern remains the fate of light cycle oils (LCOs) in FCC refineries. These molecules are in the gasoil range, but it is increasingly difficult to blend them with the traditional, more demanding, gasoil products. Likewise, it will become more difficult to employ LCO as a diluent in the diminishing residual fuel oil pool. The necessity for inventiveness in this area should provide a fertile ground for the conception of new ideas in the catalyst field as well.

Fig. 1 illustrates the generic routes by which transportation fuels can be produced in refineries. The pure hydroconversion route 1 is usually cost prohibitive; on the other hand, the much cheaper FCC and hydrotreating route 3 is not appropriate to produce the desired fuels. Consequently, usually the thermal conversion route (e.g., delayed coking) combined with hydrotreating (route 2) is selected to improve the transportation fuels slate of a refinery.

As it stands now, FCC is not the preferred route to producing low-aromatics transportation fuels. A definite advantage of FCC is that it can produce high yields of valuable propylene and butylene, which are used as feedstock in the chemical industry or as building blocks for a very clean gasoline, i.e., alkylate, but this does not compensate enough for the poor middle distillate quality. The basic problem is that the FCC process produces a significant amount of extra aromatics (which is good for high octane gasoline), which need to be removed and/or saturated with hydrogen to produce acceptable low-aromatics diesel, advanced diesel, and/or advanced gasoline.

The challenge for FCC will be to present a ''fourth route,'' reinventing the FCC process and catalyst into a refinery solution that will give good conversion of heavy hydrocarbons to low-aromatics transportation fuels, with the option of also making light olefins from heavier residual feedstocks. The objective of this

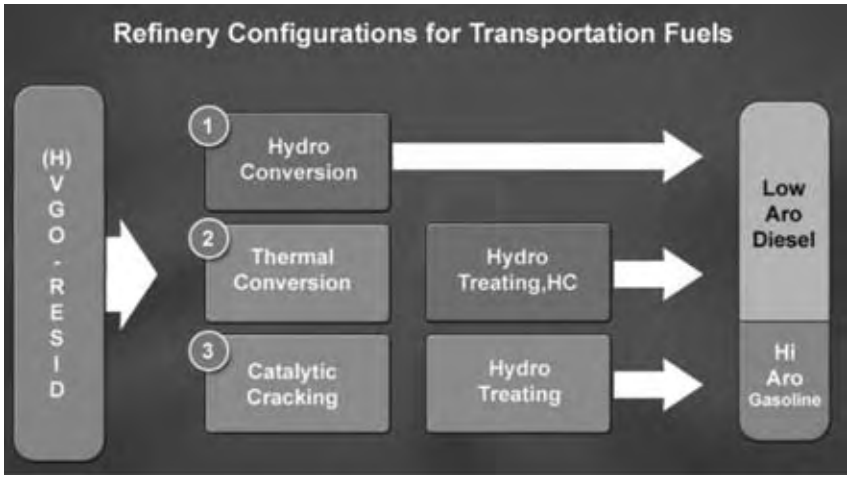


Fig. 1 Generic and simplified refinery configurations. *(View this art in color at www.dekker.com.)*

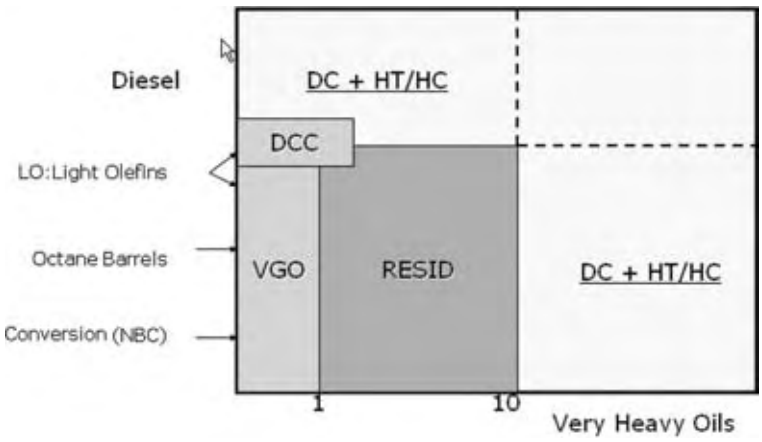


Fig. 2 Constraints on FCC technology. *(View this art in color at www.dekker.com.)*

"fourth way" will be to offer an option to the refinery industry based on the utilization of already existing FCC capacity (low investment), which will then enable the production of the advanced fuels desired in the future at low energy costs, minimum hydrogen consumption, and, hence, low extra CO_2 emissions. Preliminary scouting experiments indicate that a radical change in the FCC catalyst design strategy is possible and aromatics production can be reduced in FCC. Combining these new catalyst developments with innovations on the process side[15] opens the way to revitalizing the role of FCC in the conversion of undesirable heavy oils to clean and advanced fuels.

Also, in terms of flexibility for processing heavier oils, FCC faces tough competition from other processes. Fig. 2 illustrates how FCC technology is constrained within a box as regards product quality as well as feedstock quality. The general industry trend is towards heavier oils, which the FCC process cannot handle economically, and toward lower-aromatics transportation fuels such as diesel, which the existing FCC process cannot produce. It is very obvious that unless radical changes are made, the FCC process is approaching the end of its career as the heavy oil conversion workhorse in the refinery world.

Will the FCC process be able to compete and flourish in this difficult arena?

In a comparison of strengths and weaknesses, the FCC process competes very well with hydroconversion processes because of its simple, robust, low-pressure process configuration. On the other hand, compared to thermal conversion routes such as thermal cracking (visbreaking, deep thermal cracking) and coking, FCC runs into tough competition, especially when heavier high-metal feeds need to be processed, resulting in high catalyst consumption and hence catalyst costs. In many new refinery or refinery expansion projects, the thermal cracking option is therefore often preferred to RFCC. Here too, new breakthroughs in FCC catalyst and process technology to process the very heavy feeds will be able to change the foregoing picture.

The overall assessment, therefore, is that FCC still has a bright future ahead, albeit a future with many changes to come.

REFERENCES

1. Avidan, A.A. Origin, development and scope of FCC. In *Fluid Catalytic Cracking Science and Technology*; Magee, J.S., Mitchell, M.M., Eds.; Studies in Surface Science and Catalysis; Elsevier, 1993; Vol. 76.
2. *The Fluid Bed Reactor, A National Historic Chemical Landmark*. American Chemical Society, 1998 (published to commemorate the designation of the fluid reactor as a National Historic Chemical Landmark).
3. Schlosser, C.R.; Pinho, A.; O'Connor, P.; Baptista, R.D.; Sandes, E.F.; Torem, M.A. Residue catalytic cracking technology: the state of the art and future developments. *Akzo* Nobel Catalysts ECO-MAGIC Symposium, Noordwijk, The Netherlands, June 2001.
4. Corma, A.; Orchilles, A.V. Current views on the mechanism of catalytic cracking. Micropor. Mesopor. Mater. **2000**, *35–36*, 21–30.
5. Wojciechowski, B.V.; Corma, A. *Catalytic Cracking: Catalysis, Chemistry, Kinetics*; Marcel Dekker Inc.: New York, NY, 1986.
6. O'Connor, P.; Imhof, P.; Yanik, S.J. Catalyst assembly technology in FCC. In *Fluid Catalytic Cracking V. Materials and Technology Innovations*; Occelli, M.L., O'Connor, P., Eds.; Studies in Surface Science and Catalysis; Elsevier, 2001; Vol. 134.
7. Boyle, J.; Pickert, J.; Rabo, J.; Stamires, D. 2nd International Congress on Catalysis; Editions Technol.: Paris, July 1960; 2055 pp.

8. Plank, C.J. The invention of zeolite cracking catalysts. Chem. Tech. **1984**, April, 243.
9. Biswas, J.; Maxwell, I.E. Recent process and catalyst related developments in FCC. Appl. Catal. **1990**, *63*, 197–258.
10. Cheng, W.C.; Kim, G.; Peters, A.W.; Zhao, X.; Rajagopalan, K. Environmental FCC technology. Catal. Rev. Sci. Eng. **1998**, *40* (1&2), 39–79.
11. O'Connor, P.; Pouwels, A.C. FCC catalyst deactivation. In *Catalyst Deactivation 1994*; Studies in Surface Science and Catalysis; Elsevier, 1994; Vol. 88.
12. Hodgson, M.C.J.; Looi, C.K.; Yanik, S.J. Avoid excessive RFCCU catalyst deactivation: improve catalyst accessibility. Akzo Nobel Catalysts Symposium, Noordwijk, Netherland, June 1998.
13. Yaluris, G.; Cheng, W.; Hunt, J.L.; Boock, L.T. The effects of Fe poisoning on FCC catalysts. In NPRA Annual Meeting 2001, AM-01-59.
14. *Upgrading Heavy Crude Oils and Residues to Transportation Fuels*: *Technology*, *Economics and Outlook*, *Phase 7*; SFA Pacific, Inc., 2004.
15. Henz, H.; Avezevedo, F.; Chamberlain, O.; O'Connor, P. Re-inventing fluid catalytic cracking. Hydrocarbon Processing, September 2005.

Catalytic Dehydrogenation

Bipin V. Vora
Peter R. Pujadó
UOP LLC, Des Plaines, Illinois, U.S.A.

INTRODUCTION

Dehydrogenation, in general, concerns the abstraction of hydrogen from a compound to produce a less saturated analog. Although dehydrogenation can be used on compounds that contain heteroatoms (e.g., alcohols to aldehydes), by and large, the most important reactions pertain to hydrocarbons, principally paraffins to olefins and diolefins, or alkyl aromatics to alkenyl aromatics (e.g., ethylbenzene to styrene). Dehydrogenation can be effected thermally or catalytically. Thermal dehydrogenation is best exemplified by the pyrolysis of hydrocarbons to produce olefins, usually in the presence of steam, in pyrolysis furnaces or steam crackers at elevated temperatures. By controlling temperatures and residence times, thermal pyrolysis can be surprisingly selective, as is the case, for example, in the conversion of ethane to ethylene (~80–90 wt% yields) or in the production of α-olefins by wax cracking. In general, however, thermal pyrolysis has limited selectivity patterns. Thus, for example, the thermal pyrolysis of a naphtha will typically yield only about 29–33 wt% ethylene and 14–17 wt% propylene. But these yields are a strong function of the feedstock composition, and significantly higher yields of ethylene can be obtained if the feed naphtha is rich in *n*-paraffins.

Higher selectivities can be obtained when the dehydrogenation reactions proceed catalytically, but always subject to per-pass conversion limitations imposed by thermodynamic equilibria, so that recycle of unconverted hydrocarbons is usually required. This entry concerns mostly the application of catalytic dehydrogenation. Catalytic paraffin dehydrogenation for the production of olefins has been in commercial use since the late 1930s, while catalytic paraffin oxydehydrogenation for olefin production has not yet been commercialized. However, there are some interesting recent developments worthy of further research and development. During World War II, catalytic dehydrogenation of butanes over a chromia–alumina catalyst was performed for the production of butenes, which were then dimerized to octenes and hydrogenated to octanes to yield high-octane aviation fuel. More recently, platinum or modified platinum catalysts have been used instead of chromia–alumina catalysts. Important aspects in dehydrogenation entail the approach to equilibrium or near-equilibrium conversions while minimizing side reactions and coke formation.

Commercial processes for the catalytic dehydrogenation of propane and butanes attain per-pass conversions in the range of 30–60%, while the catalytic dehydrogenation of C_{10}–C_{14} paraffins typically operates at conversion levels of 10–20%. In 2000, nearly 7 million metric tons of C_3–C_4 olefins and 2 million metric tons of C_{10}–C_{14} range olefins were produced via catalytic dehydrogenation. Oxydehydrogenation employs catalysts containing vanadium and, more recently, platinum. Oxydehydrogenation at ~1000°C and very short residence time over Pt and Pt–Sn catalysts can produce ethylene in higher yields than in steam cracking. However, there are a number of issues related to safety and process upsets that need to be addressed. Important objectives in oxydehydrogenation are attaining high selectivity to olefins with high conversion of paraffins and minimizing potentially dangerous mixtures of paraffin and oxidant. More recently, the use of carbon dioxide as an oxidant for ethane conversion to ethylene has been investigated as a potential way to reduce the negative impact of dangerous oxidant–paraffin mixtures and to achieve higher selectivity.

While catalytic dehydrogenation reflects a relatively mature and well-established technology, in many respects oxydehydrogenation can be characterized as still being in its infancy. Oxydehydrogenation, however, offers substantial thermodynamic advantages and is an area of active research on many fronts.

CATALYTIC DEHYDROGENATION OF PARAFFINS TO OLEFINS

Dehydrogenation reactions are difficult. High temperatures and low pressures are required to achieve reasonable per-pass conversions. Hydrocarbon dehydrogenation reactions are endothermic to the tune of about 30–35 kcal/mol and thus have large heat input requirements. The equilibrium constant, K_P, of a typical dehydrogenation reaction, $A \rightarrow B + aH_2$, starting with 1 mol of A, m mol of B, and n mol of an

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016154

inert diluent (e.g., steam), can be expressed as follows:

$$K_P = P^a \frac{x}{1 - x}\left[\frac{m + ax}{1 + m + ax + n}\right]^a \quad (1)$$

where x is the equilibrium conversion of component A. This expression is needed in many dehydrogenation reactions that are conducted in the presence of hydrogen in the feed. Although hydrogen limits the extent of conversion, its use is often necessary to prevent excessive coking of the catalyst. If the reaction takes the simpler form A $\rightarrow$ B + H_2, starting with 1 mol of A and n mol of inert diluent, the expression for K_P becomes:

$$K_P = Px^2/[(1 - x)(1 + x + n)] \quad (2)$$

In these expressions, P is the total absolute pressure and $\ln(K_P) = -\Delta G/RT$, where ΔG is the change in Gibbs' free energy. Use of diluents has the effect of lowering the partial pressure and thus can lead to higher levels of conversion. Often, however, the expression can be more complex because other dehydrogenated by-products are also produced (e.g., olefins and diolefins). In the absence of any diluent, we have

$$K_P = Px^2/(1 - x^2) \quad (3)$$

$$x^2 = K_P/(K_P + P) \quad (4)$$

which allows us to quickly estimate the potential maximum conversion under a given set of conditions, T and P.

Selection of operating conditions for dehydrogenation requires a compromise. The temperature must be high enough for a favorable equilibrium and for a good reaction rate, but not so high as to cause excessive cracking or catalyst deactivation. The rate of dehydrogenation reactions diminishes as conversion increases. The ideal temperature profile in a reactor would show an increase with distance, but because dehydrogenation reactions are strongly endothermic, attainable profiles normally show a steep temperature decline or, at best, are isothermal when sufficient heat can be provided across the walls of the catalyst bed.

The reactor pressure should be as low as possible without excessive recycle costs or equipment size. Usually, it is near atmospheric, though reduced pressures (under vacuum) have been used in the Houdry butane dehydrogenation process. In any case, the catalyst bed must be designed for a low pressure drop.

Rapid preheating of the feed is desirable to minimize cracking. Rapid cooling or quenching at the exit of the reactor is sometimes necessary to prevent condensation reactions of the olefinic products. Construction materials must be resistant to attack by hydrogen, capable of prolonged operation at high temperature, and not unduly active for conversion of hydrocarbons to carbon. Alloy steels containing chromium are usually favored. Steels containing nickel are also used, but they can cause trouble from carbon formation. If steam is not present, traces of sulfur compounds may be needed to avoid carbonization. Both steam and sulfur compounds act to keep the metal walls in a passive condition.

HISTORICAL OVERVIEW AND CHROMIA–ALUMINA CATALYSTS

Paraffin dehydrogenation for the production of olefins has been in use since the late 1930s. During World War II, catalytic dehydrogenation of butanes over a chromia–alumina catalyst was done for the production of butenes that were then dimerized to octenes and hydrogenated to octanes to yield high-octane aviation fuel.

Dehydrogenation of butanes over a chromia–alumina catalyst was first developed and commercialized at Leuna in Germany and was also independently developed by UOP (then Universal Oil Products) in the United States, together with ICI in England. The first UOP-designed plant came onstream in Billingham, England, in 1940 and was soon followed by two more units in Heysham, England, in 1941.[1] The primary purpose of these butane dehydrogenation units was to produce butenes that were then dimerized to octenes using solid phosphoric acid catalysts discovered by R.E. Schaad and V.N. Ipatieff.[2]

Other companies soon followed these pioneering efforts. For example, Phillips Petroleum built a multitubular dehydrogenation reactor near Borger, TX, in 1943.[1] However, the most significant development was made by Houdry using dehydrogenation at less than atmospheric pressure for higher per-pass conversions. This process, which came onstream toward the end of World War II, was also used for the production of butenes. After the war, Houdry further developed and commercialized the chromia–alumina dehydrogenation system and extended it to the production of butadiene, in what came to be known as the Catadiene™ process.[3]

In the dehydrogenation process using chromia–alumina catalysts, the catalyst is contained in a fixed shallow bed located inside a reactor that may be either a sphere, a squat vertical cylinder, or a horizontal cylinder. The actual design reflects a compromise between gas flow distribution across a large cross-sectional area and the need to maintain a low pressure drop. A significant amount of coke is deposited on the catalyst during the dehydrogenation step; therefore, a number of reactors are used in parallel—some for dehydrogenation while the rest are being purged or regenerated.

The dehydrogenation reactions are strongly endothermic, and the heat is provided, at least in part, by the sensible heat stored in the catalyst bed during regeneration (carbon burn); additional heat is provided by direct fuel combustion and also by heat released in the chromium redox cycle. The length of the total reactor cycle is limited by the amount of heat available and can be as short as 10–20 min.

The Houdry Catadiene process was used extensively for the production of butadiene, either by itself (*n*-butane to butadiene) or in conjunction with catalytic oxydehydrogenation of *n*-butene to butadiene. The latter was commercialized by the Petro-Tex Chemical Corp. and was called the Oxo-D™ process.[3] A similar oxydehydrogenation approach for the production of butadiene was also adopted by Phillips Petroleum in their O-X-D™ process.[3]

Other companies, including Exxon (then Esso), Shell, Gulf, and Dow, also followed similar dehydrogenation technologies for the conversion of *n*-butenes to 1,3-butadiene, often in the presence of steam. Steam serves as a heat carrier, to lower the hydrocarbon partial pressure, and to minimize coke deposits on the catalyst. Catalyst compositions varied considerably. Thus, Exxon initially used a catalyst based on MgO with hourly regenerations but later changed it to a catalyst based on iron oxide with only very infrequent regenerations. Similarly, Shell's 205 catalyst (62.5% iron oxide, 2.3% chromia, and 35.3% potassium carbonate) required about 12 mol of steam per mole of feed butenes and had a butadiene selectivity of about 75–80 mol% at about 30% per-pass conversion. Dow's catalyst Type B contained Ca and Ni phosphates with a small amount of chromia and had a selectivity to butadiene of about 90 mol%, but at a higher steam ratio of about 20 mol per mole of feed butenes; hourly regenerations were still required.

Large quantities of butadiene have become available over the past 30 yr, mostly as a by-product from the thermal cracking of naphtha and other heavy hydrocarbons. This marked shift has resulted in the shutdown of all on-purpose catalytic dehydrogenation units for butadiene production in North America, Western Europe, and the Far East.

In the late 1980s, the application of chromia–alumina catalysts was extended by Houdry to the dehydrogenation of propane to propylene and isobutane to isobutylene. The new process application called Catofin™ operates on the same cyclic principle as in the former Catadiene process.[4,5] As of late 2000, a total of eight Catofin units existed for the production of isobutylene (including two converted older Catadiene units) with an aggregate capacity of about 2.8 million metric tons per annum (MTA) of isobutylene. In addition, two Catofin units were built for the production of propylene, but it is understood that only one is operational with a nameplate dehydrogenation capacity of about 250,000 MTA propylene, but usually operating only on a seasonal basis. Plans for another 450,000-MTA Catofin propane dehydrogenation unit in Saudi Arabia have also been announced. The Catofin process technology is currently owned by Süd-Chemie and is offered for license by ABB Lummus.

Around 1959, an alternative chromia–alumina catalytic dehydrogenation process was developed in the former Soviet Union. This method avoided the use of the cyclic operation by using a fluidized-bed reactor configuration similar to the fluid catalytic cracking (FCC) process used in refineries.[6] However, back mixing common to dense fluidized-bed operations results in poor selectivity and increases the formation of heavies, sometimes called ''green oils.'' A circulating regenerated catalyst is used to provide the heat of reaction in the riser, and the spent catalyst is reheated by carbon burn in the regenerator. During the 1990s, a large-scale fluid-bed isobutane dehydrogenation unit for about 450,000 MTA isobutylene was commercialized by Snamprogetti (an Italian company) in Saudi Arabia based on technology from Yarsintez in Russia.[6] Recent literature reports further improvements by Snamprogetti.[7,8]

NOBLE METAL DEHYDROGENATION CATALYSTS

A different approach to catalytic dehydrogenation was first introduced in the mid-1960s for the supply of long-chain linear olefins for the production of biodegradable detergents.

Synthetic detergents, based on the use of branched alkylbenzene sulfonates derived from propylene tetramer and benzene, had been introduced in the 1940s. By the early 1960s, however, it became apparent that branched dodecylbenzene-based detergents, though very active and offering excellent performance characteristics, did not biodegrade readily and were accumulating in the environment. The need for biodegradable detergents prompted the development of catalytic dehydrogenation of long-chain linear paraffins to linear olefins.

The work on catalytic reforming with noble metal (Pt) catalysts done in the 1940s by V. Haensel clearly demonstrated that Pt-based catalysts had high activity for the dehydrogenation of paraffins to the corresponding olefins.[9] In the 1960s, Herman Bloch further extended this thinking by developing Pt-based catalysts that could selectively dehydrogenate long-chain linear paraffins to the corresponding internal mono-olefins with high activity and stability and with minimum cracking.[10] This was the basis for the UOP Pacol™ process for the production of linear olefins for the

manufacture of biodegradable detergents.[11] In 1999, there were more than 30 commercial Pt-catalyzed dehydrogenation units in operation for the manufacture of detergent alkylate.

Long-chain paraffins are both valuable and highly prone to cracking. Therefore, to maintain high selectivity and yield, it is necessary to operate at relatively mild conditions, typically below 500°C, and at relatively low per-pass conversions. While this is economical for the production of heavy linear olefins, it is not for the production of light olefins.

Paraffin dehydrogenation is an endothermic reaction that is limited by chemical equilibrium and, according to Le Chatelier's principle, higher conversion will require either higher temperatures or lower pressures. As noted earlier, the maximum equilibrium conversion in a simplified form can be approximated as follows:

$$X_e^2 = \frac{K_P}{K_P + P} \tag{5}$$

where X_e is the equilibrium conversion, P is the total absolute pressure, and K_P is the equilibrium constant for the dehydrogenation reaction. The equilibrium constant can be easily calculated from Gibbs' free energies as tabulated in the API 44 report or in similar sources of thermodynamic data. Figs. 1 and 2 illustrate the equilibrium conversion levels that can be obtained for propane at 1 atm. abs. and at 0.23 atm. abs. (175 torr), respectively (read the respective component mole fractions as the vertical intervals between lines at a given temperature).

The equilibrium constant for paraffin dehydrogenation increases significantly as the carbon number increases. Fig. 3 shows the equilibrium constant for the dehydrogenation of *n*-paraffins ranging from ethane to pentadecane.[12]

Fig. 4 shows the temperatures required to achieve 10% and 40% equilibrium conversion based on these equilibrium constants. It indicates that the temperature required for the dehydrogenation of light paraffins is much higher than that for heavy paraffins. For example, for 40% conversion, the dehydrogenation of propane requires a temperature of at least about 580°C, while dodecane can be theoretically dehydrogenated to the same extent at only 450°C. The equilibrium conversion increases at higher temperatures, but side reactions, coke formation, and catalyst deactivation are also accelerated. Thus, extrapolation directly from heavy olefins to light olefins cannot be done without taking other factors into consideration.

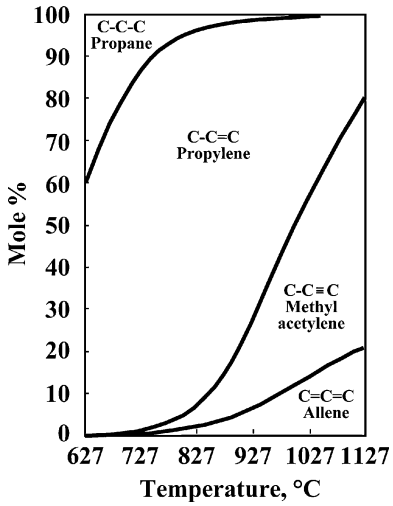


Fig. 1 Propane dehydrogenation equilibrium at 1 atm. abs. pressure.

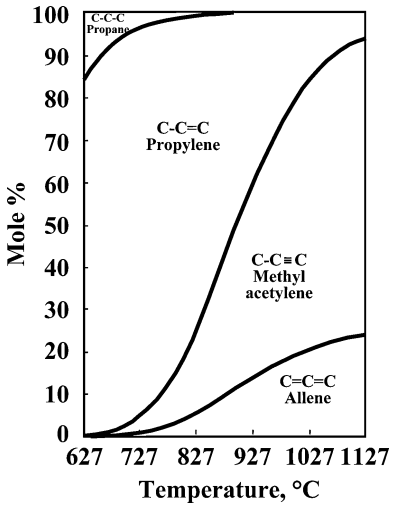


Fig. 2 Propane dehydrogenation equilibrium at 0.23 atm. abs. pressure.

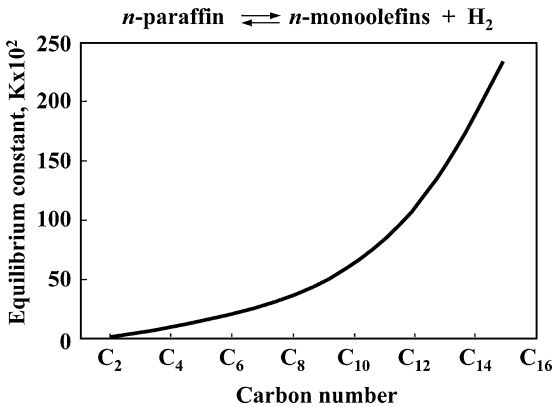


Fig. 3 Equilibrium constants for *n*-paraffin dehydrogenation at 500°C.

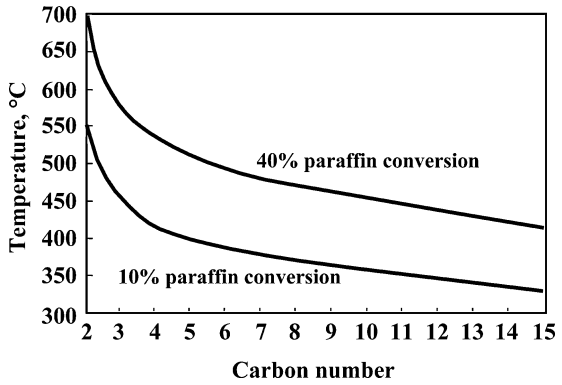


Fig. 4 Temperature required to achieve 10% and 40% conversion of C_2–C_{15} *n*-paraffins at 1 atm. abs. pressure.

Production of light olefins by the catalytic dehydrogenation of light paraffins must be able to maintain reasonable per-pass conversion levels and high olefin selectivity. Very importantly, it must be able to produce olefins in high yields over long periods of time without shutdowns.

In the early 1970s, UOP introduced continuous catalyst regeneration (CCR) technology that enabled noble metal catalysts to remain at their most desirable stable activity for several years without having to shutdown the reactor for catalyst regeneration. The combination of noble metal catalysts operating at high severity in conjunction with CCR technology made it possible to design, build, and economically operate large catalytic dehydrogenation units that can produce light olefins, in particular propylene and isobutylene, at high selectivities while still operating at superatmospheric pressures. This technology is known as the UOP Oleflex™ process. As of late 2003, there were five propane dehydrogenation units, five isobutane dehydrogenation units, and one combined propane/isobutane dehydrogenation unit of this type in commercial operation with an aggregate operating capacity of 1.3 million MTA polymer grade propylene and 2.3 million MTA isobutylene.

The world propylene production capacity based on the use of catalytic dehydrogenation of propane has increased steadily over the past 10 yr and is expected to grow even further under the right economic conditions relative to the availability and pricing of propane.[13] On the other hand, environmental concerns on the use of methyl-*tert*-butyl ether (MTBE), an oxygenated gasoline additive, are expected to adversely impact the future expansion of isobutane dehydrogenation applications.

Although production of ethylene via catalytic dehydrogenation over Pt catalysts is very selective (about 95%), extension of this dehydrogenation technology to ethane has not taken place because of the need for even more severe operating conditions: higher temperatures and lower pressures. Such conditions cause excessive coking of the catalyst or require costlier operation under vacuum.

Practically all existing catalytic dehydrogenation capacity based on Pt catalysts is based on the Oleflex process with CCR; however, there are also two smaller units for isobutane dehydrogenation for 118,000 and 13,000 MTA isobutylene, respectively, both based on the steam active reforming technology (STAR) developed by Phillips Petroleum and derived from their earlier multitubular reactor design experience. This reactor design resembles a typical steam reformer that is operated until the catalyst deactivates as a result of coke deposition; banks of tubes are sequentially taken out of service for catalyst regeneration. The STAR technology is currently owned and licensed by Krupp-Uhde.

PROCESS CHEMISTRY

The main reaction in catalytic dehydrogenation is the formation of mono-olefins from the corresponding feed paraffin. Others include consecutive and side reactions. The reaction pathways involved in heavy paraffin dehydrogenation (e.g., detergent range C_{10}–C_{14} *n*-paraffins) are more complicated than those in light paraffin dehydrogenation (e.g., propane and isobutane). The main difference in reaction pathways is that a significant amount of cyclic compounds can form via dehydrocyclization from heavy paraffins; this is not the case for light paraffins. Figs. 5 and 6 illustrate the possible reactions that take place on platinum (Pt) and acid (A) sites, respectively, in the dehydrogenation of light and heavy paraffins when the catalyst is not selective, e.g., unmodified platinum catalysts supported on alumina.

The consecutive reactions, the dehydrogenation of mono-olefins to diolefins and triolefins, are catalyzed on the same active sites as the dehydrogenation of paraffins to mono-olefins. The consecutive reactions that

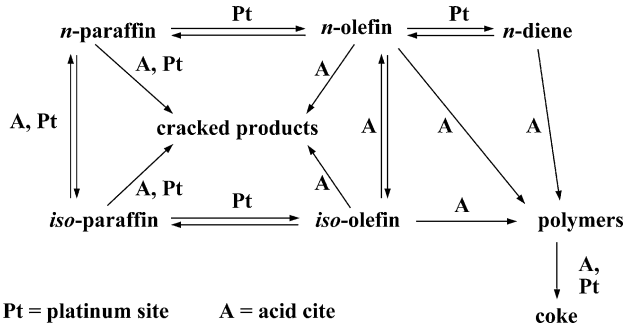


Fig. 5 Reactions in platinum and acid sites in light paraffin dehydrogenation with unmodified catalyst.

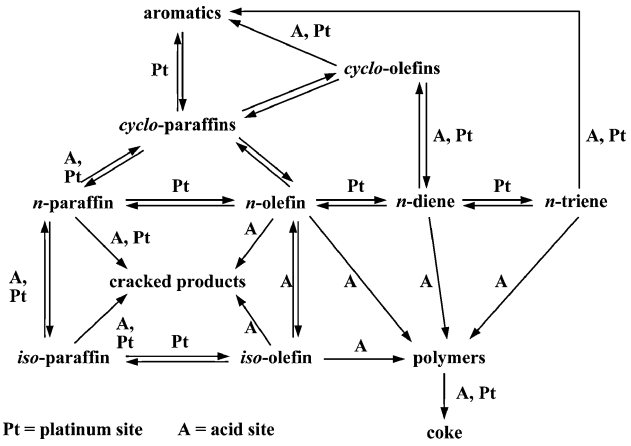


Fig. 6 Reactions in platinum and acid sites in heavy paraffin dehydrogenation with unmodified catalyst.

form triolefins, aromatics, dimers, and polymers must be suppressed kinetically or by catalyst modifications.

ROLE OF CATALYSTS AND SUPPORTS

The discussion in this section pertains to alumina-supported platinum catalysts. The work by C.P. Poole and D.S. MacIver provides an extensive review of chromia–alumina catalysts.[14,15]

The key role of dehydrogenation catalysts is to accelerate the main reaction while controlling other reactions. Unmodified alumina-supported platinum catalysts are highly active but are not selective to dehydrogenation. Various by-products, as indicated in Figs. 5 and 6, can also form. In addition, the catalyst rapidly deactivates because of fouling by heavy carbonaceous materials. Therefore, the properties of platinum and the alumina support need to be modified to suppress the formation of by-products and to increase catalytic stability.

The reaction of olefins on platinum is faster than that of paraffins because olefins interact with platinum more strongly than do paraffins. The role of platinum modifiers is to weaken the platinum–olefin interaction selectively without affecting the platinum–paraffin interaction. Among metals, arsenic, tin, germanium, lead, and bismuth are reported as platinum activity modifiers. The consecutive dehydrogenation rate of mono-olefins and diolefins is decreased by this modification without lowering the rate of paraffin dehydrogenation significantly. The modifier also improves the stability against fouling by heavy carbonaceous materials.

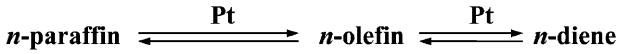
Platinum is a highly active catalytic element and is not required in large quantities to catalyze the reaction when it is dispersed on a high surface area support. The high dispersion is also necessary to achieve high selectivity to dehydrogenation relative to undesirable side reactions, such as cracking.

The typical high surface area alumina supports employed have acidic sites that accelerate skeletal isomerization, cracking, oligomerization, and polymerization of olefinic materials, and enhance coke formation. Alkalis or alkaline earth metals assist in the control of the acidity. Also, α-alumina supports that have essentially no acidity can be utilized; however, the challenge is to obtain high dispersion of platinum on such very low surface area supports. Therefore, acidity must be eliminated by using suitable modifiers.

Modified catalysts possess high activity and selectivity to mono-olefins. The major by-products are diolefins that can be controlled kinetically. Coke formation is also suppressed, and therefore, stability is greatly improved. Over modified catalysts, the major reaction pathways for both light and heavy paraffin dehydrogenation systems are simpler (Fig. 7).

Alumina has excellent thermal stability and mechanical strength under processing, transport, and catalyst regeneration conditions. However, the most important reason alumina is used as a support material is its superior capability to maintain a high degree of platinum dispersion, which is essential for achieving high dehydrogenation activity and selectivity.

Light paraffin dehydrogenation



Heavy paraffin dehydrogenation

Fig. 7 Paraffin dehydrogenation on modified Pt catalysts.

The catalytic reaction rate is limited by the intraparticle mass transfer rate. If the rate is relatively slow, both activity and selectivity are lowered. As a result, the support must have a low pore diffusional resistance (high effectiveness factor). For a given pore volume, the surface area and the strength of the support increase as the pore diameter decreases, and the pore diffusional resistance decreases as the pore diameter increases. Thus, an appropriate pore structure must be determined for the support to achieve optimal catalytic performance.

Selectivity decreases as the conversion increases because *n*-mono-olefins are consecutively converted into by-products. Selectivity decreases sharply as conversion approaches equilibrium because the main dehydrogenation process is limited by equilibrium, but other reactions continue to occur. Therefore, if side reactions are controlled, the selectivity is improved as the equilibrium conversion becomes higher by increasing the temperature and by decreasing the pressure and the feed ratio of hydrogen to paraffin.

The relationship between selectivity and conversion can be simulated if rate functions, relative rate constants, and equilibrium constants are known. Fig. 8 shows simulated selectivities to *n*-heptene and *n*-heptadiene for the dehydrogenation of *n*-heptane. In this simulation, the relative rate constants used are unity, which represents that the catalyst possesses perfect selectivity regarding consecutive dehydrogenation; the dehydrogenation rate of paraffin is equal to that of mono-olefin and diolefin. Experimental selectivities obtained over a UOP dehydrogenation catalyst show that the catalyst has >90% selectivity for consecutive dehydrogenation steps, as seen in Fig. 8.

The rate of light paraffin conversion over a Pt catalyst (Oleflex type process) can be expressed as a modified first-order equation according to a Langmuir–Hinshelwood mechanism. Other equations may be derived accordingly.

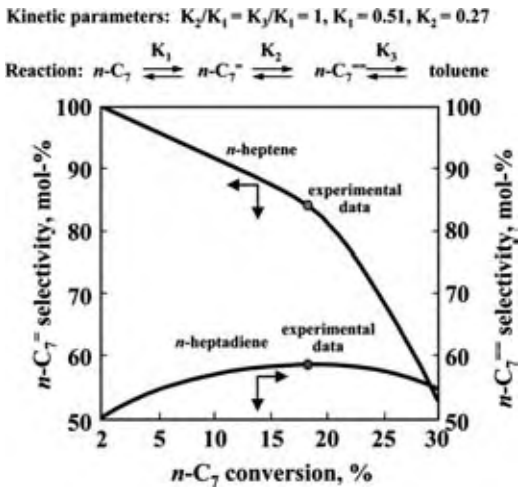


Fig. 8 Simulation of selectivity for dehydrogenation of *n*-heptane. *(View this art in color at www.dekker.com.)*

CATALYST STABILITY AND REGENERATION

The dehydrogenation of long-chain paraffins is performed under relatively mild temperature conditions of 400–500°C. Thus, the catalyst can maintain a long life even at high space velocity and catalyst productivity. Therefore, it is not economical to build facilities for catalyst regeneration.

Because of equilibrium limitations, the dehydrogenation of light paraffins requires significantly higher temperatures above 600°C to achieve economically attractive conversions. The catalyst deactivation is accelerated under high-temperature conditions, and frequent catalyst regeneration is necessary for light paraffin dehydrogenation. For the dehydrogenation of light paraffins, a number of different types of reactor-regeneration systems are commercially utilized.

- Houdry's Catofin and similar processes employ a cyclic sequence of steps—process, purge, air regeneration, purge, hydrogen reduction, and back to process.
- The Phillips STAR process also regenerates the catalyst on a cyclic basis, but while the Houdry regeneration is actually a mechanism to provide the heat for the reaction even when coke buildup is still very low, the catalyst in the isothermal STAR process is only regenerated after coke has accumulated to appreciable levels that result in low catalyst activity.
- UOP's Oleflex process uses multistage adiabatic reactors with CCR.
- Snamprogetti's dehydrogenation process consists of a fluidized-bed reactor and regeneration system. Here too, the coke buildup is very low and the ''regeneration'' loop is actually a means of supplying heat to the reactor.

HEAT OF REACTION

The heat of reaction for paraffin dehydrogenation is about 30 kcal/mol (125 kJ/mol). In a cyclic adiabatic operation (e.g., Houdry), it is provided by reheating the catalyst and the added inert materials to a high temperature during the regeneration step, so that the catalyst cools down and conversion decreases during the reaction step. Because several reactors are used in parallel, average conversion is obtained. In an isothermal process (e.g., STAR), the catalyst is loaded inside vertical tubes inside a furnace and the heat is

introduced through the tube walls. In a fluidized reactor, the temperature profile can be maintained uniformly in the back mixed zone of the bed, while heat is provided by introducing a hot regenerated catalyst. In Oleflex adiabatic reactors, a significant temperature drop occurs across the catalyst bed, which lowers the equilibrium conversion level; a multistage reactor system with interstage reheating is used for higher paraffin conversions.

Fig. 9 illustrates conversion, equilibrium conversion, and temperature along the catalyst bed in a three-stage adiabatic reactor system for the dehydrogenation of isobutane. For propane dehydrogenation, a four-stage reactor system becomes more economical because higher average temperatures are needed. A multistage reactor system also affords lower inlet temperatures, relative to a single-stage reactor system (Fig. 10). Thus, thermal cracking and catalyst deactivation, which are accelerated at higher temperatures, can be controlled to low levels.

PROCESS FLOW AND REACTOR CHARACTERISTICS

Cyclical Processes

As described earlier, the Houdry Catadiene process, the Houdry Catofin process, and other similar cyclical processes make use of parallel reactors that contain a shallow bed of chromia–alumina catalyst. Fig. 11 illustrates a schematic of such a process.

This technology has been used extensively for the production of butadiene and, in more recent years, for that of isobutylene and propylene.[16,17] The feed is preheated through a fired heater before being passed over the catalyst in the reactors. The hot reactor effluent is cooled, compressed, and sent to the product fractionation and recovery section. The dehydrogenation reactors are refractory-lined carbon steel vessels (i.e., cold wall design). To accommodate continuous flow of the main streams—hydrocarbons and regeneration air—the reactors are operated on a timing cycle that satisfies the following requirement:[1]

$$\text{Onstream time} + \text{Regeneration time} + \text{Purge time} = \text{Total cycle time} \quad (6)$$

The number of reactors in each cycle is the prorated time fraction of the total cycle time. Thus, with five reactors, two reactors can be onstream simultaneously, two on regeneration, and one on purge, evacuation, and valve changes. Fig. 12 provides a typical timing cycle for a five-reactor unit, but as many as eight reactors in parallel have been provided in some units. The total cycle time is usually in the range of 15–30 min.

The onstream period at subatmospheric pressure is followed by a purge. Next comes regeneration at essentially atmospheric pressure, followed by purge, hydrogen reduction, and evacuation to reaction pressure, after which the reactor is ready for another onstream period. Process streams enter and leave the reactors through fast-acting gate valves. The gate valves can range up to 40 in. (1 m) in diameter, are designed for high-temperature service, and are equipped with a pressurized inert seal in the bonnet to prevent leakage of air into the process gas when the valve is closed. Overall, this mechanical design has proven to be very reliable over many years of operation.

The regeneration is done with air that has been preheated through a direct fired burner or, alternatively, with the exhaust of a gas turbine. The regeneration step is intended to preheat the catalyst to the onstream temperature necessary to initiate the next process cycle and to remove coke deposits on the catalyst. Flue gas sensible heat may be recovered in a waste heat boiler. The hydrogenation step prepares the catalyst for the dehydrogenation phase and also contributes additional heat from the reduction of Cr^{6+} to Cr^{3+}.

Another cyclical process is the Phillips STAR process.[18] It uses a fixed-bed fired-tube reactor operating at a positive superatmospheric pressure. In many respects, it is similar in design to a steam reforming furnace with the heat of reaction provided by firing outside the tubes, thus operating at near-isothermal conditions. Steam is used as a diluent to lower the partial pressure of the reactants and, thus, to achieve reasonable conversion levels of about 30–40% for propane and 45–55% for butanes. It also helps slow down the deposition of carbon (coke) on the catalyst, thereby extending cycle time from minutes to hours.

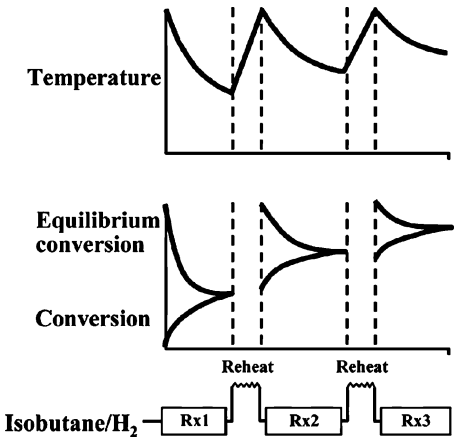
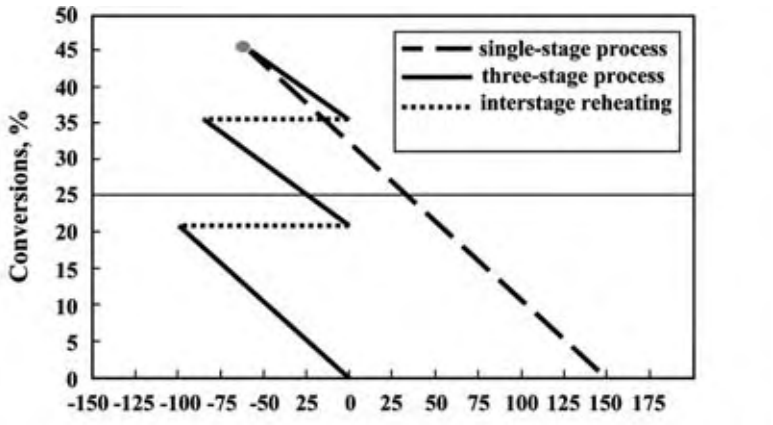


Fig. 9 Temperature profile and conversions of three-stage isobutene dehydrogenation process.



Typical temperature endotherms per stage, in °C, required to achieve about 45% isobutane conversion. In a single-stage approach the feed would have to be preheated about 150°C higher than in a three-stage process

Fig. 10 Isobutane dehydrogenation. *(View this art in color at www.dekker.com.)*

Periodic catalyst regeneration or carbon burn off is required to maintain the activity of the catalyst. Typical cycle time is reported to be at least 8 hr, with 7 hr of process time and 1 hr of regeneration time. For continuous operation, various furnace modules can be operated such that, for example, seven operate in the process mode while one is in the regeneration mode. Fig. 13 shows a schematic diagram of a STAR process unit.[18]

Continuous Processes

Snamprogetti has commercialized fluidized-bed dehydrogenation (FBD) for the catalytic dehydrogenation of light paraffins using a chromia–alumina catalyst with an alkaline promoter, which is used primarily for the dehydrogenation of isobutane to isobutylene in the manufacture of MTBE.[6–8] The catalyst is microspheroidal with an average diameter of $<100\,\mu m$ and

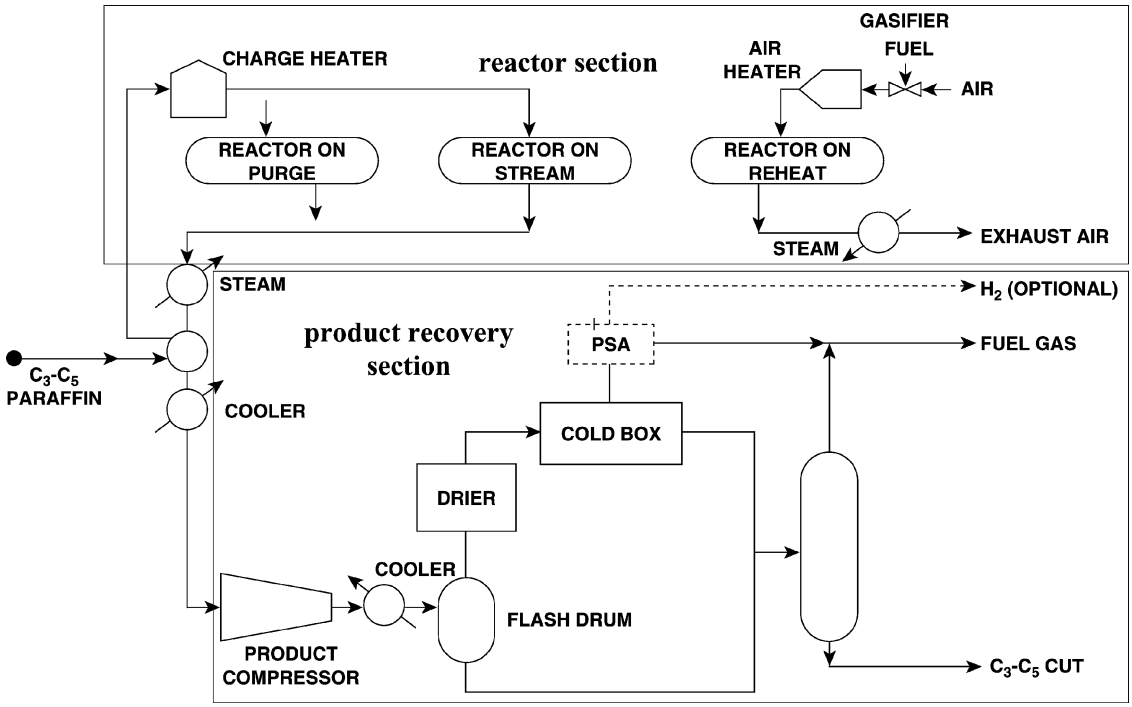


Fig. 11 Catofin process flow diagram.

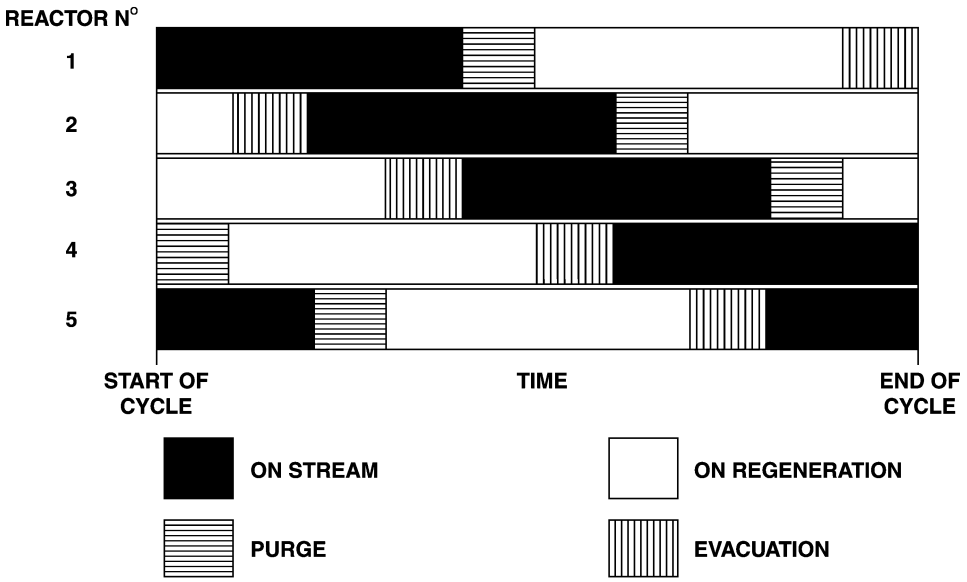


Fig. 12 Olefins from paraffins.

an apparent bulk density of $<2000\,kg/m^3$.[19] The heat of the reaction is provided by the circulating hot regenerated catalyst back to the reactor. In all concepts, the FBD process is very similar to the FCC process units commonly used in petroleum refineries. However, because back mixing has a negative effect on the yields, horizontal baffles with suitable openings are inserted within the fluidized bed to limit the backflow of solids, such that the fluidized bed is split into a series of stages, each comparable to a continuous stirred tank reactor (CSTR).[19] A typical process scheme is shown in Fig. 14.

Fresh feed is vaporized, mixed with the recycle of unconverted paraffins, and fed to the fluidized reactor through a distributor for optimal even distribution. Entrained catalyst is removed from the product off-gas by means of cyclones. The catalyst circulates continuously from the reactor to the regenerator and vice versa by means of transfer lines. Coke deposited on the catalyst is burnt off in the regenerator; however, because the amount of coke is relatively small, additional fuel must be burnt in the regenerator to satisfy the thermal requirements of the endothermic dehydrogenation reaction. However, while this approach is similar to that in the Houdry process, FBD does not have a catalyst reduction step with hydrogen before proceeding to the dehydrogenation cycle; lack of this step is believed to be somewhat detrimental to the overall performance of the process.

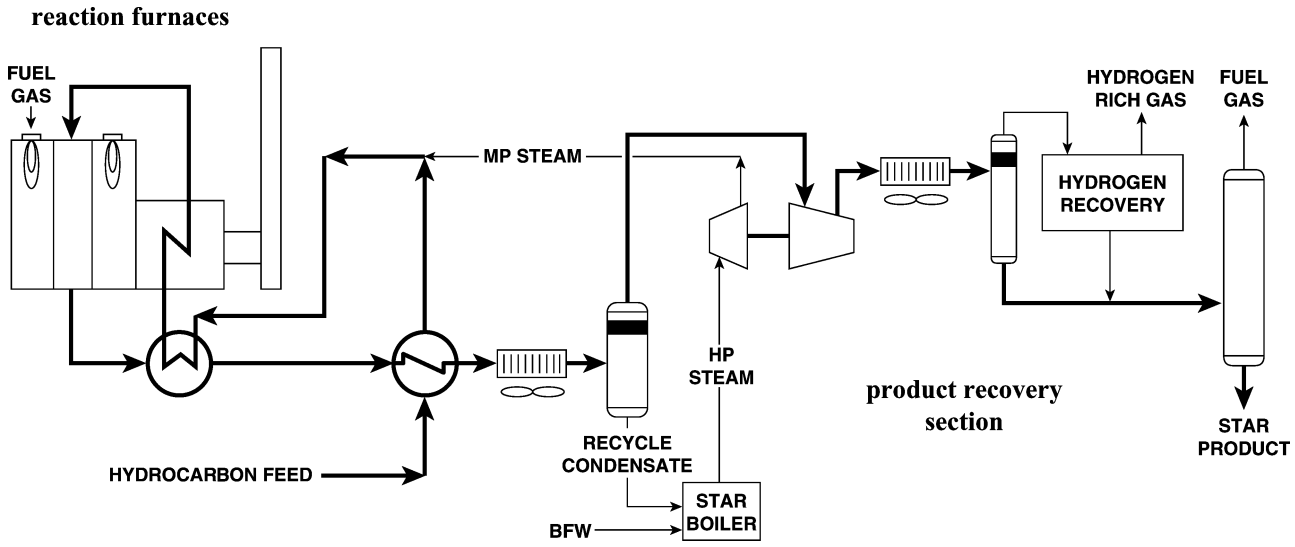


Fig. 13 Phillips Petroleum Company—STAR process.

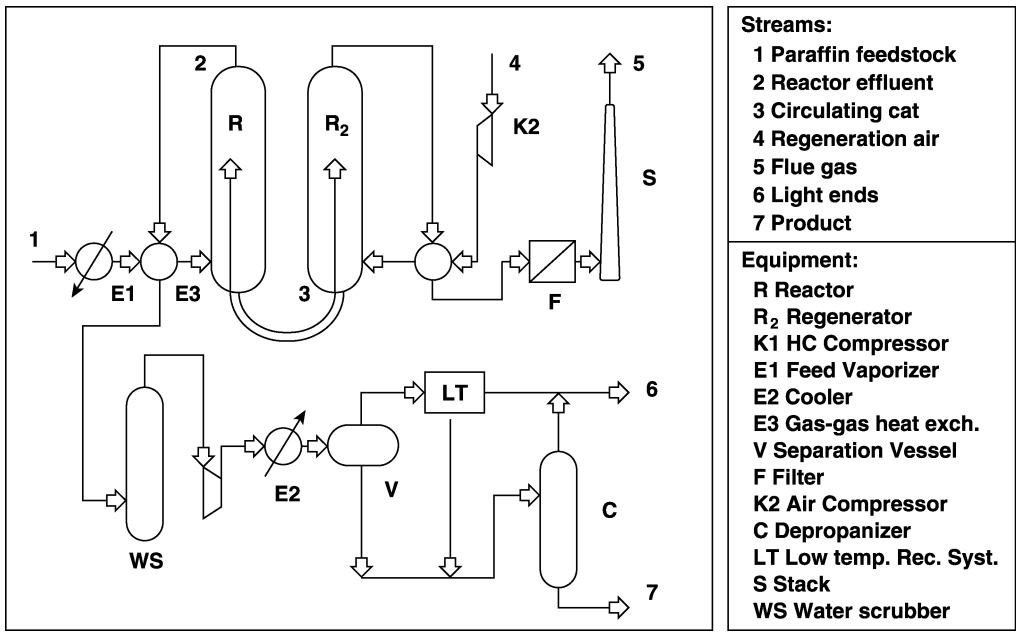


Fig. 14 Snamprogetti's FBD process scheme.

UOP's catalytic dehydrogenation processes typically make use of radial-flow adiabatic fixed-bed (or slowly moving bed) reactors with modified Pt–alumina catalysts.

The UOP Pacol process for selective long-chain paraffin dehydrogenation to produce linear mono-olefins is shown in Fig. 15 in combination with the UOP detergent alkylation process. The Pacol process consists of a radial-flow reactor and a product recovery section. Worldwide, more than 2 million metric tons per year of linear alkyl benzene is produced employing this process.[20]

The flow diagram of the UOP Oleflex process is seen in Fig. 16. The process consists of a reactor section and a product recovery section. The former consists of three or four stages of radial-flow reactors, charge and interstage heaters, reactor feed–effluent exchangers, and the CCR section (Fig. 17). As noted earlier, more than 1 million metric tons of propylene and 2 million metric tons of isobutylene are now produced via this route.[13] The performances of the Pacol and Oleflex processes are summarized in Table 1.

Use of the Oleflex process for the dehydrogenation of ethane to ethylene has also been investigated but, to date, the economics do not appear to be favorable because of the low equilibrium conversion and the need to operate at a pressure lower than atmospheric if a reasonable ethane conversion is to be expected;

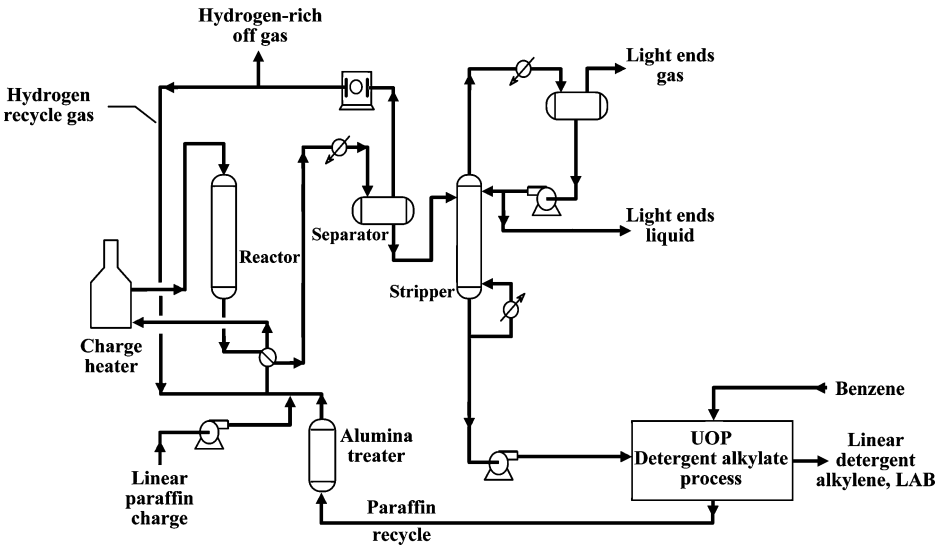


Fig. 15 UOP Pacol dehydrogenation process.

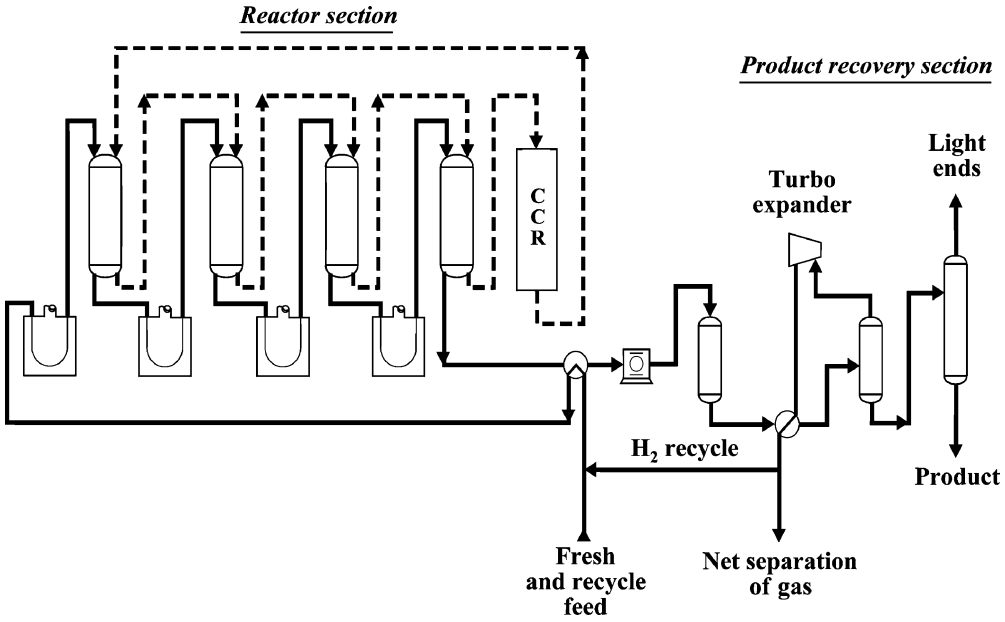


Fig. 16 UOP Oleflex process.

the cost of fractionating ethylene in an ethane–ethylene splitter is otherwise too high. Dow Chemical has recently been awarded a patent for the dehydrogenation of ethane over a metal–mordenite catalyst complex at relatively low conversions in which the product ethylene is selectively recovered from the dilute ethylene–ethane stream by alkylating it with benzene.[21]

REACTOR DESIGN OPTIONS

The choice of reactor design plays a very important role in the success of catalytic processes. The following types are commercial reactor designs for endothermic catalytic dehydrogenation processes:

- Downflow adiabatic fixed bed.
- Radial-flow fixed bed or moving bed adiabatic.
- Tubular isothermal.
- Fluidized bed.

Table 2 summarizes the main characteristics of the four reactor systems. The dot represents beneficial characteristics of each reactor type.

The choice of the right reactor depends on the catalyst and the selection of operating conditions.

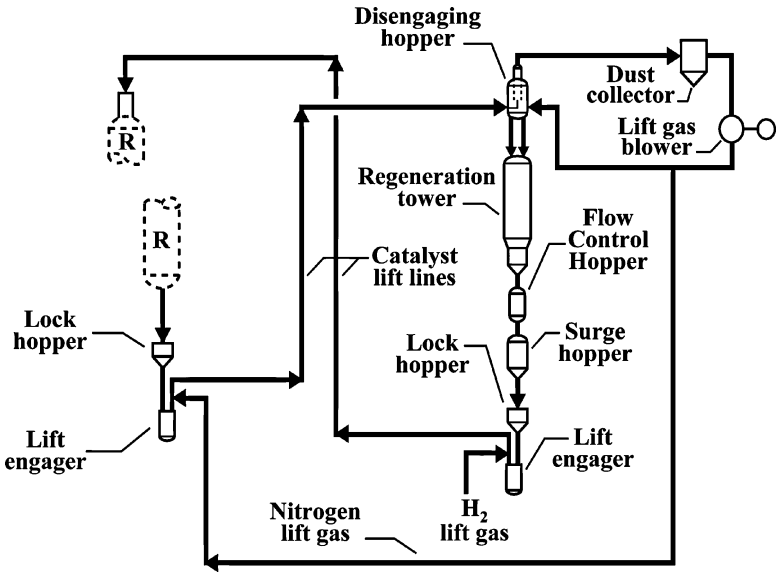


Fig. 17 Oleflex catalyst regeneration section.

Table 1 Performance of Pacol and Oleflex processes

Process	Feed	Conversion (%)	Selectivity (%)
Oleflex	Propane	40	90
	n-Butane	50	85
	Isobutane	50	92
Pacol	*n*-Heptane	20	90
	n-C_{10}–C_{13}	13	90
	n-C_{11}–C_{14}	13	90

ETHYLBENZENE DEHYDROGENATION

Although the previous discussion had centered on the catalytic dehydrogenation of paraffins, a study on the subject would not be complete without analysis of the dehydrogenation of ethylbenzene to styrene.

Styrene is an important monomer or comonomer in the manufacture of a number of polymers: polystyrene, acrylonitrile/butadiene/styrene, styrene/acrylonitrile, etc. There are two main processes for the manufacture of styrene. In one process styrene is made a coproduct with propylene oxide:

Ethylbenzene + air → ethylbenzene hydroperoxide

Ethylbenzene hydroperoxide + propylene → methyl phenyl carbynol + propylene oxide

Methyl phenyl carbynol → styrene + water

This process was originally developed and commercialized by Oxirane (a joint venture company between ARCO Chemical, now Lyondell, and Halcon) and independently by Shell Petrochemical Company. At present, this is one of the main processes for the commercial manufacture of propylene oxide (the other is a variant of the same that starts with isobutane instead of ethylbenzene, and produces propylene oxide together with *tert*-butyl alcohol, isobutylene, and ultimately MTBE, as the main coproducts); so significant amounts of styrene are obtained therefrom.

The demand for styrene, however, outstrips the supply available from its coproduction with propylene oxide; so the other major process for the production of styrene is the catalytic dehydrogenation of ethylbenzene:

Ethylbenzene → styrene + H_2

The dehydrogenation reaction proceeds over an iron or an iron–chromium catalyst that usually also contains potassium in the form of potassium carbonate, so that at elevated temperatures various complex mixed carbonates and oxides are formed, e.g., $KFeO_2$. Temperatures are elevated, in the order of 630°C, and pressures are usually subatmospheric for improved per-pass conversions. Steam dilution is performed to further lower the partial pressure of the reactants. Because the reaction is strongly endothermic, various reaction stages with interheat (and interstage addition of superheated steam) are normally employed. Fig. 18 illustrates a typical process scheme for the dehydrogenation of ethylbenzene to styrene.

Usually, two adiabatic reactors in series with interheat are used, but a third reactor may be present or added for increased capacity as shown in Fig. 19.

In an interesting variant of the conventional process, part of the hydrogen that is produced in the first stage of conversion is selectively reacted with oxygen over a separate bed of catalyst such that significant amounts of heat are released internally within the reactor system. The hydrogen oxidation catalyst is selected to ensure that there is practically no conversion or degradation of either ethylbenzene or styrene. While this process could be thought of as an oxydehydrogenation, in reality it is just a conventional dehydrogenation coupled with an oxidative reheat step. The alternative ethylbenzene dehydrogenation process, known commercially as Styrene Monomer by Advanced Reheat Technology (SMART™), was invented by Dr. Dennis J. Ward of UOP LLC in an original version called Styro-Plus demonstrated at Mitsubishi Chemicals, Kashima, Japan, and is now licensed commercially jointly by UOP LLC and ABB Lummus Global Inc. In addition to supplying the heat of reaction internally, the SMART process benefits from the equilibrium displacement that results from the selective removal of one of the reaction products, hydrogen, as illustrated in Fig. 20.

It has been shown[22] that SMART process technology is best suited for revamps of existing ethylbenzene dehydrogenation units, which in the case of the UOP/Lummus technology are called the Classic™ process. Figs. 19, 21, and 22 illustrate various modes

Table 2 Characteristics of four reactor systems

	Downflow	Radial flow	Tubular	Fluidized bed
Low pressure drop		•		•
Plug flow	•	•	•	
Catalyst addition or removal		•		•
High heat transfer, near isothermal			•	•
Variable heat transfer coefficient				

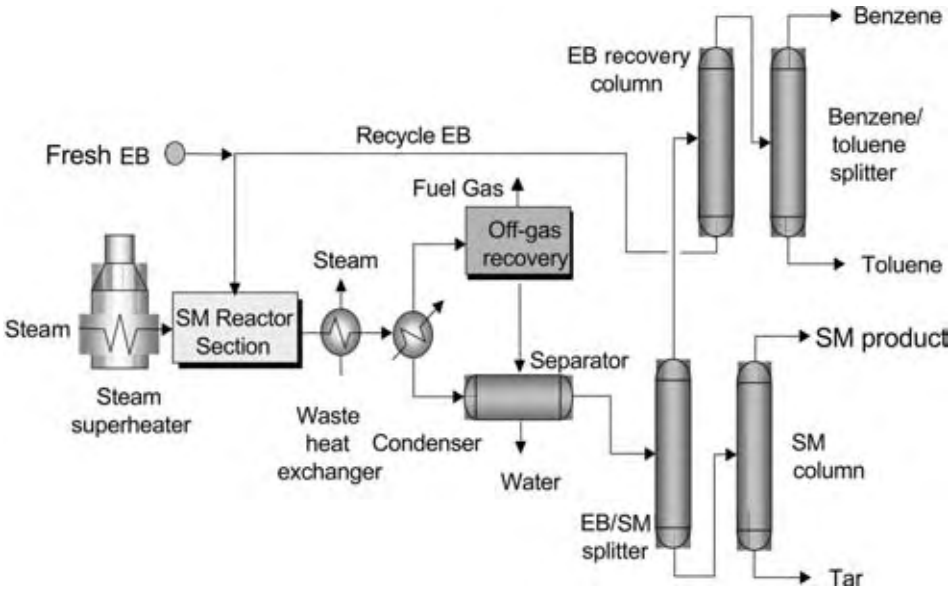


Fig. 18 Typical ethylbenzene dehydrogenation unit for the production of styrene monomer. *(View this art in color at www.dekker.com.)*

of revamping an existing ethylbenzene dehydrogenation process unit.

- In a classic revamp (Fig. 19), a new dehydrogenation reactor with its own feed reheat exchanger is inserted between the two existing dehydrogenation reactors.
- In a partial SMART revamp (Fig. 21), one SMART reactor that contains both oxidation catalyst and dehydrogenation catalyst is placed between the two existing dehydrogenation reactors. No reheat exchanger is needed for the SMART reactor.
- In a full SMART revamp (Fig. 22), an additional oxidation reactor is added after the intermediate SMART reactor and before the last conventional dehydrogenation reactor, so that no feed reheat exchanger is needed for the final stage either.

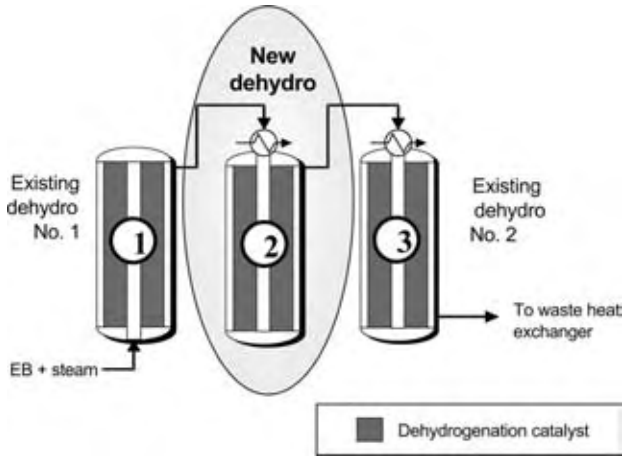


Fig. 19 Possible conventional revamp of a classic styrene monomer unit. *(View this art in color at www.dekker.com.)*

Table 3 compares the results of the various revamp options.

OTHER DEHYDROGENATION TECHNOLOGIES

The processes discussed above are for the direct catalytic dehydrogenation of paraffins to the corresponding olefins or of olefins to diolefins. Other methods have also been considered, although none has reached the level of commercialization. Some of the most notable are:

- Halogen-assisted dehydrogenation.
- Oxydehydrogenation.

Use of halogens for the dehydrogenation of paraffins has been proposed in different ways. For example, heavy paraffins were first chlorinated and then dehydrochlorinated to heavy olefins commercially in

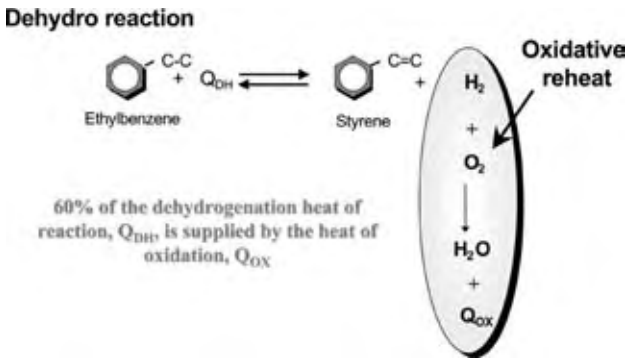


Fig. 20 Oxidative reheat mechanism. *(View this art in color at www.dekker.com.)*

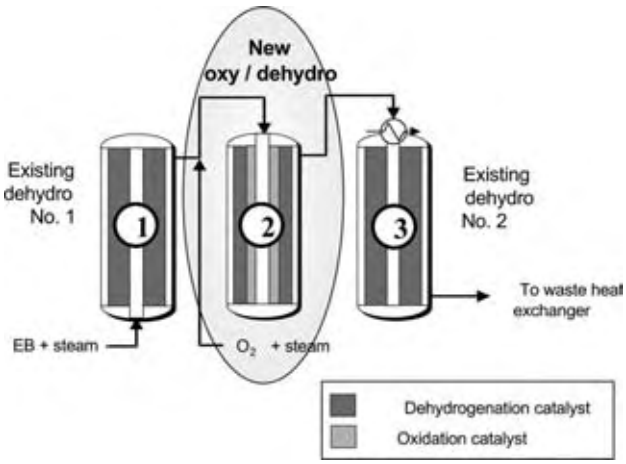


Fig. 21 Partial SMART revamp. *(View this art in color at www.dekker.com.)*

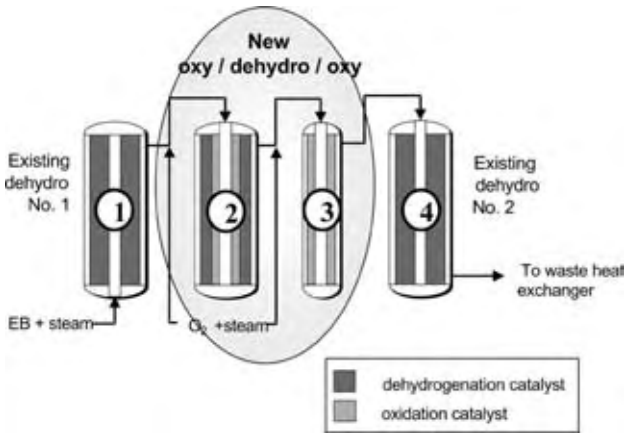


Fig. 22 Full SMART revamp. *(View this art in color at www.dekker.com.)*

the past both by Shell and Hüls, among others. Pyrolysis of methane in the presence of chlorine has been proposed by Prof. Sidney Benson for the production of acetylene and ethylene.[23] Other chlorination/dehydrochlorination cycles have been proposed for the production of ethylene from ethane. Propane dehydrogenation in the presence of iodine via a propyl iodide intermediate has also been proposed.[4,24] Apart from the apparent corrosion problems associated with the use of halogens, other difficulties readily come to mind owing to the relatively high cost of chlorine, and even more so of iodine, and the need to either dispose of or recycle vast quantities of halogens.

Oxydehydrogenation or oxidative dehydrogenation can be considered in at least two different ways.

- Use of oxygen to oxidize the hydrogen coproduct from dehydrogenation, and thus to displace the dehydrogenation equilibrium to higher conversions as discussed earlier for the dehydrogenation of ethylbenzene to styrene. Although a similar approach has been proposed for the dehydrogenation of paraffins, it has not been commercialized.[25–28]
- Direct use of oxygen as a means of dehydrogenating, for example, ethane to ethylene. Oxydehydrogenation has successful commercial applications in the conversion of *n*-butenes to butadiene (e.g., as in the Oxo-D process referred to earlier), but not yet for the production of ethylene or propylene.

Use of oxydehydrogenation relative to straight catalytic dehydrogenation must be viewed both in terms of safety issues and in an economic context.

Table 3 Results of various revamp options

	Classic revamp	Partial SMART	Full SMART
Reactor section	Add conventional dehydrogenation reactor	Add smart oxy/dehydrogenation reactor	Add smart oxy/dehydrogenation reactor and another oxydehydrogenation reactor
EB conversion (%)	70–72	74–76	80–82
Capacity increase (%)	25–30	25–40	40–50
Superheater	New heater	No change—minor modification	Minor modification
Off-gas compressor	New compressor	No change	No change
Vent-gas recovery	Bottleneck	No change	No change
Fractionation	Modification required (no major retraying required for front-end columns if there is <20% capacity increase)	Modification required (no major retraying required for front-end columns if there is <25% capacity increase)	Modification required (no major retraying required for front-end columns if there is <35% capacity increase)

On the latter, even though oxydehydrogenation offers advantages as a means of overcoming thermodynamic equilibrium limitations, it also leads to the total or partial loss of by-product hydrogen, which in some instances can have a very significant economic impact.

Although less apparent, oxydehydrogenation also plays a role in the work done by BP Amoco, Asahi, and others to extend ammoxidation to the direct conversion of propane to acrylonitrile. It is believed that the ammoxidation of propane proceeds through a transient propylene intermediate from which acrylonitrile is derived through a conventional ammoxidation pathway.[29] Similarly, the conversion of *n*-butane to maleic anhydride could also be regarded as a form of oxydehydrogenation, except that in this case it leads to the formation of a new compound that incorporates oxygen in the molecule.

CONCLUSIONS

Catalytic dehydrogenation of paraffins and of ethylbenzene is a commercial reality in numerous applications, from the production of light olefins, heavy olefins, to that of alkenylaromatics. Oxydehydrogenation, on the other hand, is still in the developmental stage, but, if successful, holds great promise on account of its potential energy savings.

REFERENCES

1. Hornaday, G.F.; Ferrell, F.M.; Mills, G.A. Manufacture of mono- and diolefins from paraffins by catalytic dehydrogenation. In *Advances in Petroleum Chemistry and Refining*; Interscience Publishers, 1961; Vol. 4.
2. Egloff, G.; Davis, R.F. Polymerization of monoolefins with solid phosphoric acid, XII International Congress of Pure and Applied Chemistry, New York, 1951; 10–13.
3. Waddams, A.L. *Chemicals from Petroleum*, 4th Ed.; Gulf Publishing Company, 1978 (reprinted 1980).
4. Weiss, A.H. The manufacture of propylene. (158th Meeting of the American Chemical Society, Sep 10–12, 1969.) In *Refining Petroleum for Chemicals*; Advances in Chemistry Series; American Chemical Society: Washington, DC, 1970; Vol. 97.
5. Craig, R.G.; Spence, D.C. Catalytic dehydrogenation of liquefied petroleum gas by the Houdry Catofin and Catadiene processes. In *Handbook of Petroleum Refining Processes*; Robert, A.M., Ed.; McGraw-Hill, 1986; Section 4.1.
6. Sanfilippo, D.; Buonomo, F.; Fusco, G.; Miracca, I. Paraffins activation through fluidized bed dehydrogenation: the answer to light olefins demand increase. In *Studies in Surface Science and Catalysis*; Elsevier, 1998; Vol. 119, 919–924.
7. Iezzi, R.; Bartolini, A. US Patent 5,633,421, May 27, 1997.
8. Luckenbach, E.C.; Zenz, F.A.; Papa, G.; Bertolini, A. US Patent 5,656,243, Aug 12, 1997.
9. Haensel, V. US Patent 2,602,772, Jul 8, 1952 (and other patents on catalytic reforming with Pt catalysts assigned to UOP).
10. Bloch, H.S. UOP discloses new way to make linear alkylbenzene. Oil Gas J. **1967**, 79–81. US Patent 3,448,165, Jun 3, 1969 (and other patents on the catalytic dehydrogenation of paraffinic hydrocarbons assigned to Herman Bloch and UOP).
11. Berg, R.C.; Vora, B.V. Detergent alkylate. In *Encyclopedia of Chemical Processing and Design*; Marcel Dekker, Inc., 1982; Vol. 15, 266–284.
12. Imai, T.; Vora, B.V.; Bricker, J.C. Development of Dehydrogenation Catalysts and Processes. Petroleum and Petrochemical College, Chulalongkorn University, Bangkok, Dec 1989.
13. Gregor, J.H.; Antonelli, R.; Foley, T.D.; Arnold, E.C. Increased opportunities for propane dehydrogenation. DeWitt World Petrochem. Rev. Houston, Tex. **1999,** *Mar*, 23–25.
14. Poole, C.P.; MacIver, D.S. The physical–chemical properties of chromia–alumina catalysts. In *Advances in Catalysis*; Academic Press, 1967; Vol. 17, 223–314.
15. Poole, C.P.; Kehl, W.L.; MacIver, D.S. Physical properties of coprecipitated chromia–alumina catalysts. J. Catal. **1962**, *1*, 407–415.
16. Tucci, E.L.; Dufallo, J.M.; Feldman, R.J. Commercial performance of the Houdry Catofin process for isobutylene production for MTBE, Workshop on Catalysts and Catalytic processes, Research Institute, King Fahd University of Petroleum and Minerals: Dhaharan, Saudi Arabia, Nov 6, 1991.
17. Gussow, S.; Whitehead, R. Isobutane dehydrogenation by Catofin as feed for motor fuel ether, NPRA Annual Meeting, National Petroleum Refiners' Association (NPRA): San Antonio, TX, Mar 17–19, 1991.
18. Dunn, R.O.; Brinkmeyer, F.M.; Schuette, G.F. The Phillips STAR process for the dehydrogenation of C_3, C_4, and C_5 paraffins, NPRA Annual Meeting, National Petroleum Refiners' Association (NPRA): New Orleans, LA, Mar 22–24, 1992.
19. Buonomo, F.; Fusco, G.; Miracca, I.; Papa, G.; Sanfilippo, D. *Fluid Bed Dehydrogenation of Light Paraffins: The FBD Technology*; Petrotech-96, Bahrain, Jun 10–12, 1996.

20. Lawson, R.J.; Fritsch, T.R. *Developing a Clean, Efficient Process*; Asia-Pacific Chemicals, 1999; Vol. 99, 8–9.
21. Campbell, A.Q.; Garcés, J.M.; May, T.M.; Pogue, R.F. Preparation of Ethylbenzene and Substituted Derivatives by Alkylation Using Unpurified Reaction Products of Ethylene Prepared by Dehydrogenation of Ethane World Patent WO 96/34843, Nov 7, 1996 (assigned to Dow Chemical Co.).
22. Jeanneret, J.J.; Tagamolila, C.P.; Woodle, G.B.; Gandhi, D. Get SMART—start planning your styrene capacity expansion now, Styrene Conference 1000, Süd-Chemie AG and ABB Lummus: Louisville, KY, Aug 13–17, 2000.
23. Benson, S Conversion of methane using chlorine.. US Patent 4,199,533, Apr 22, 1996 (assigned to the University of Southern California).
24. Weiss, A.H. Which propylene process is best? Hydrocarbon Process. **1968**, *47*, 123–127.
25. Laegreid, T.; Rönnekleiv, M.; Sobbaken, A. Proceedings of the DGMK Conference, 11–12, Nov 11–17, 1993; 147.
26. Lodeng, R.; Soaker, P. Reactor for Catalytic Dehydrogenation of Hydrocarbons with Selective Oxidation of Hydrogen US Patent 5,997,826, Dec 7, 1999 [assigned to Den Norske Stats Oljeselskap A.S. (Statoil)].
27. Tsikoyiannis, J.G.; Stern, D.L.; Grasselli, R.K. Metal oxides as selective hydrogen combustion (SHC) catalysts and their potential in light paraffin dehydrogenation. J. Catalysis. **1999**, *184*, 77.
28. Agaskar, P.A.; Grasselli, R.K.; Michaels, J.N.; Reischman, P.T.; Stern, D.L.; Tsikoyiannis, J.G Process for the Catalytic Dehydrogenation of Alkanes to Alkenes with Simultaneous Combustion of Hydrogen. US Patent 5,430,209, Jul 4, 1995 (assigned to Mobil Oil Corp.).
29. Brazdil, J.F.; Cavalcanti, A.P.; Padolewski, J.P. Method for Preparing Vanadium Antimony Oxide Based Oxidation and Ammoxidation Catalysts US Patent 5,693,587, Dec 2, 1997 [assigned to The Standard Oil Company of Ohio (Sohio/BP)].

Catalytic Naphtha Reforming

Abdullah M. Aitani
King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

INTRODUCTION

The use of catalytic naphtha reforming as a process to produce high-octane gasoline continues to be important as it has been over the 55 yr of its commercial use. The catalytic reformer occupies a key position in a refinery providing high value-added reformate for the gasoline pool, hydrogen for hydroprocessing operations, and, frequently, benzene, toluene, and xylene aromatics for petrochemical uses. The main objective of catalytic reforming is to transform paraffins and naphthenes in naphtha to aromatics-rich products with as little ring opening or cracking as possible. Aromatics have very high octane numbers (>100) and can be tolerated in gasoline up to almost 50 vol%. However, because of stringent environmental regulations, the refining industry has taken significant steps to reduce the level of aromatic compounds in gasoline by adapting its formulation. Many countries have reduced the total aromatics content in gasoline from 42–35 vol%.

Naphtha feed to the reformer contains a mixture of C_6–C_{11} paraffins, naphthenes, and aromatic hydrocarbons in the temperature range of 85–200°C. Most reformers process straight-run naphthas with qualities that vary significantly depending on the crude oil origin and boiling range.[1–3] Typical straight-run naphtha contains 65 wt% paraffins, 20 wt% naphthenes, and 15 wt% aromatics. Cracked naphthas, such as visbreaker, coker, and heavy hydrocracked, are another source of reformer feeds. In motor gasoline applications, naphtha contains the full range of C_6–C_{11} components to maximize gasoline production. In petrochemical applications, naphtha feed may be adjusted to contain a more select range of hydrocarbons to maximize the production of aromatics.

Currently, there are more than 700 commercial installations of catalytic reforming units worldwide with a total capacity of about 11.4 million barrels per day (b/d). Table 1 presents a worldwide distribution of catalytic reforming capacity.[4] About 40% of this capacity is located in North America followed by 20% each in Western Europe and Asia-Pacific region.

REFORMING CHEMISTRY

Several catalytic reactions take place during the naphtha reforming that involve the rearrangement of hydrocarbon molecules. Because naphtha contains many paraffin and naphthene isomers, multiple reforming reactions take place simultaneously in the reforming reactors. Naphthenes react rapidly and efficiently to aromatics, while paraffins react slowly and with poorer specificity. Aromatic compounds pass through the reforming unit relatively unchanged.[5] These reactions comprise the following:

- Dehydrogenation and dehydroisomerization of naphthenes to aromatics.
- Aromatization of paraffins.
- Isomerization of normal paraffins and naphthenes.
- Hydrocracking of paraffins and naphthenes to lighter saturated paraffins.
- Dealkylation of aromatics.

The above reactions take place concurrently. The first two are desirable for the production of aromatics, while the last two are generally undesirable because of the production of lower-value light products. The dehydrogenation of naphthenes proceeds rapidly and produces hydrogen as well as aromatics. The reactions take place on the surface of the catalyst and are very much dependent, among other factors, on the right combination of interactions between catalyst ingredients. Most of them are promoted by the dual metal–acid functions of the catalyst. The majority of these reactions involve the conversion of paraffins and naphthenes and result in an increase in octane number and a net production of hydrogen. The overall reactions are highly endothermic and require continuous supply of heat to maintain reaction temperature.

PROCESS CLASSIFICATION

The reforming process is generally classified into three types:

1. Semiregenerative
2. Cyclic (fully regenerative)
3. Continuous regenerative (moving bed).

This classification is based on the frequency and mode of regeneration. Table 2 presents a region-wise distribution of catalytic reforming capacity by process

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039766

Table 1 Worldwide distribution of naphtha reforming by capacity and process type

Region	No. of refineries	Crude capacity (1000 b/d)	Reforming capacity (1000 b/d)	Reforming as % crude capacity
N. America	159	20,476	4,193	20.5
W. Europe	104	14,727	2,147	14.6
Asia-Pacific	161	20,695	1,917	9.3
E. Europe	93	10,236	1,520	14.9
Middle East	45	6,472	661	10.2
S. America	67	6,573	429	6.5
Africa	46	3,230	478	14.8
Total	675	82,409	11,347	13.8

(From Ref.[4].)

type.[4] The semiregenerative scheme dominates the reforming capacity at about 57% of the total capacity followed by continuous regenerative at 27% and cyclic at 11%. Most grassroots reformers are designed for continuous catalyst regeneration (CCR). In addition, many units that were originally built as semiregenerative units have been revamped to continuous regeneration units. The semiregenerative type requires unit shutdown for catalyst regeneration, whereas the cyclic process utilizes a swing reactor for regeneration in addition to regular reactors. The continuous process permits catalyst replacement during normal operation.

Semiregenerative Process

The semiregenerative process is characterized by continuous operation over long periods, with decreasing catalyst activity. Eventually, the reformers are shut down periodically as a result of coke deposition to regenerate the catalyst in situ. Regeneration is carried out at low pressure (approximately 8 bar) with air as the source of oxygen. The development of bimetallic and multimetallic reforming catalysts with the ability to tolerate high coke levels has allowed the semiregenerative units to operate at 14–17 bar with similar cycle lengths obtained at higher pressures. It is believed that all reforming licensors have semiregenerative process design options.[6]

The semiregenerative process is a conventional reforming process that operates continuously over a period of up to 1 year. As the catalytic activity decreases, the yield of aromatics and the purity of the by-product hydrogen drop because of increased hydrocracking. Semiregenerative reformers are generally built with three to four catalyst beds in series. The fourth reactor is usually added to some units to allow an increase in either severity or throughput while maintaining the same cycle length. The longer the required cycle length, the greater the required amount of catalyst. Conversion is maintained more or less constant by raising the reactor temperatures as catalyst activity declines. Sometimes, when the capacity of a semiregenerative reformer is expanded, two existing reactors are placed in parallel, and a new, usually smaller, reactor is added. Frequently, the parallel reactors are placed in the terminal position. When evaluating unit performance, these reactors are treated as though they are a single reactor of equivalent volume.[1,6]

Table 2 Distribution of naphtha reforming by process type

Region	Reforming capacity (1000 b/d)	Share of reforming capacity by process type (%)			
		Semiregenerative	Continuous	Cyclic	Other
N. America	4,193	46.4	26.8	22.2	4.6
W. Europe	2,147	54.0	31.5	11.0	3.5
Asia-Pacific	1,917	42.4	44.8	1.6	11.2
E. Europe	1,520	86.4	11.0	1.1	1.5
Middle East	661	63.0	23.1	7.2	6.7
S. America	429	80.4	9.3	0.6	3.5
Africa	478	81.9	0.0	1.8	16.3
Total	11,347	56.8	26.9	11.1	5.2

(From Ref.[4].)

The research octane number (RON) that can be achieved in this process is usually in the range of 85–100, depending on an optimization between feedstock quality, gasoline qualities and quantities required, as well as the operating conditions required to achieve a certain planned cycle length (6 mo to 1 yr). The catalyst can be regenerated in situ at the end of an operating cycle. Often the catalyst inventory can be regenerated 5–10 times before its activity falls below the economic minimum, whereupon it is removed and replaced.

Cyclic (Fully Regenerative) Process

The cyclic process typically uses five or six fixed catalyst beds, similar to the semiregenerative process, with one additional swing reactor, which is a spare reactor. It can substitute any of the regular reactors in a train while the regular reactor is being regenerated. In this way, only one reactor at a time has to be taken out of operation for regeneration, while the process continues in operation. Usually, all the reactors are of the same size. In this case, the catalyst in the early stages is less utilized; therefore, it will be regenerated at much longer intervals than in the later stages. The cyclic process may be operated at a low pressure, a wide boiling range naphtha feed, and a low hydrogen-to-naphtha ratio.[6] Coke lay-down rates at these low pressures and high octane severity (RON of 100–104 range) are so high that the catalyst in individual reactors becomes exhausted in time intervals of less than 1 week to 1 mo.

The process design of the cyclic process takes advantage of low unit pressures to gain a higher C_{5+} reformate yield and hydrogen production. The overall catalyst activity, conversions, and hydrogen purity vary much less with time than in the semiregenerative process. However, a drawback of this process is that all reactors alternate frequently between a reducing atmosphere during normal operation and an oxidizing atmosphere during regeneration. This switching policy needs a complex process layout with high safety precautions and requires that all the reactors be of the same maximum size to make switches between them possible.

Continuous Regenerative Process

The CCR process is characterized by high catalyst activity with reduced catalyst requirements, more uniform reformate of higher aromatic content, and high hydrogen purity. The process can achieve and surpass reforming severities as applied in the cyclic process but avoids the drawbacks of the cyclic process. The continuous process represents a step change in reforming technology compared to the semiregenerative and cyclic processes. Since its introduction in the early 1970s, it has gained wide acceptance in refining and petrochemical industries worldwide.

In this process, small quantities of catalyst are continuously withdrawn from an operating reactor, transported to a regeneration unit, regenerated, and returned to the reactor system. In the most common moving-bed design, all the reactors are stacked on top of one another. The fourth (last) reactor may be set beside the other stacked reactors. The reactor system has a common catalyst bed that moves as a column of particles from top to bottom of the reactor section. Coked catalyst is withdrawn from the last reactor and sent to the regeneration reactor, where the catalyst is regenerated on a continuous basis. However, the final step of the regeneration, i.e., reduction of the oxidized platinum and second metal, takes place in the top of the first reactor or at the bottom of the regeneration train.

Fresh or regenerated catalyst is added to the top of the first reactor to maintain a constant quantity of catalyst. Catalyst transport through the reactors and the regenerator is by gravity flow, whereas that from the last reactor to the top of the regenerator and back to the first reactor is by the gas-lift method. Catalyst circulation rate is controlled to prevent any decline in reformate yield or hydrogen production over time onstream.

In another design, the individual reactors are placed separately, as in the semiregenerative process, with modifications for moving the catalyst from the bottom of one reactor to the top of the next reactor in line. The regenerated catalyst is added to the first reactor, and the spent catalyst is withdrawn from the last reactor and transported back to the regenerator. The continuous reforming process is capable of operation at low pressures and high severity by managing the rapid coke deposition on the catalyst at an acceptable level. Additional benefits include elimination of downtime for catalyst regeneration and steady production of hydrogen of constant purity (93% compared to 80% in the semiregenerative process). Operating pressures are in the 3.5–17 bar range and design research octane number is in the 95–108 range.

COMMERCIAL REFORMING PROCESSES

Several commercial reforming processes are available for license worldwide. A list of reforming processes with a summary of key process features is presented in Table 3. Several commercial processes are available, dominated by UOP and Axens Technologies for semiregenerative and continuous reforming. Other licensors included Houdry Division, Chevron, Engelhard, Exxon-Mobil, and Amoco, but none of them is currently

Table 3 Licensors of naphtha reforming processes

Licensor	Process name	Process type and key features	Remarks
UOP	Platforming	Semiregenerative; CCR; CycleMax regenerator; product recovery system	Over 800 units with 8 million b/d
Axens	Octanizing; Aromizing	Semiregenerative; CCR; dual forming for conventional process revamp	Over 100 licensed units
Houdry Div. of Air Products	Houdriforming	Semiregenerative; high-octane gasoline and aromatics	0.3 million b/d
Engelhard	Magnaforming	Semiregenerative or semicyclic	1.8 million b/d
ExxonMobil	Powerforming	Semiregenerative or cyclic	1.4 million b/d
Chevron	Rheniforming	Semiregenerative; low-pressure operation	1 million b/d
Amoco	Ultraforming	Semiregenerative or cyclic	0.5 million b/d

(From Ref.[6].)

actively licensing.[6] The processes differ in the type of operation (semiregenerative or CCR), catalyst type, catalyst regeneration procedure, and process engineering design.

All licensors agree on the necessity of hydrotreating the feed to lower the level of poisons for the platinum-based reforming catalyst.[2] Temporary poisons are sulfur and nitrogen, while As, Pb, and other metals are permanent poisons. Proper conditions of hydrogen, pressure, temperature, and space velocities are able to reduce these poisons to the acceptably low levels of modern catalysts. Numerous process design modifications and catalyst improvements have been made in recent years.

Platforming Process

Platforming was the first process to use platinum-on-alumina catalysts. The first UOP semiregenerative platforming unit went onstream in 1949. UOP technology is used in over 50% of all reforming installations with more than 750 units in service worldwide.[7] The platforming process has been adapted to bimetallic catalyst and to both semiregenerative and continuous operations. UOP offers semiregenerative units that use staged loading of catalyst for increased production. New catalysts are continuously aimed at maximizing hydrogen yield, increasing operating flexibility, and maximizing C_{5+} product yield.

The total capacity of the semiregenerative units exceeded 5.0 million b/d while CCR units reached 3.8 million b/d. The simultaneous use of CCR technology and bimetallic catalysts has given UOP a unique position in the field of catalytic reformer process licensing. Recent catalyst formulations have improved both aromatic and reformate yields. UOP has improved the performance of the conventional platforming process by incorporating a CCR system. The process uses stacked radial-flow reactors and a CCR section to maintain a steady-state reforming operation at optimum process conditions: fresh catalyst performance, low reactor pressure, and minimum recycled gas circulation. The flow pattern through the platforming unit with CCR is essentially the same as with conventional fixed-bed units. The effluent from the last reactor is heat exchanged against combined feed, cooled, and phase split into vapor and liquid products in a separator. A schematic flow diagram of the CCR platforming process is presented in Fig. 1.

Catalyst flows vertically by gravity down the stack, while the feed flows radially across the annular catalyst bed. The predominant reactions are endothermic; hence, an interheater is used between each reactor to reheat the charge to reaction temperature. The catalyst is continuously withdrawn from the last reactor and transferred to the regenerator. The withdrawn catalyst flows down through the regenerator where the accumulated carbon is burned off. The regenerated catalyst is purged and then hydrogen is lifted to the top of the reactor stack, maintaining nearly fresh catalyst quality. Because the reactor and regenerator sections are separate, each operates at its own optimum conditions.

Typical operating conditions for the current design of the UOP CCR process are: reactor pressure 6.8 bar; liquid hourly space velocity (LHSV) 1.6 per hour; H_2/hydrocarbon (HC) molar ratio 2 : 3; and RON-clear 100–107. The CCR unit operates at higher severity and lower reactor catalyst inventory. In addition, the CCR unit runs continuously compared to 12 mo semiregenerative cycle lengths. Typical product yields for the CCR and semiregenerative platforming units operating at the same conditions are presented in Table 4.[8] Many of the benefits of CCR include higher hydrogen yield and purity as well as higher octane barrels.

One recent development in CCR technology is second-generation CCR platforming with several modifications in the reactor and regenerator sections.

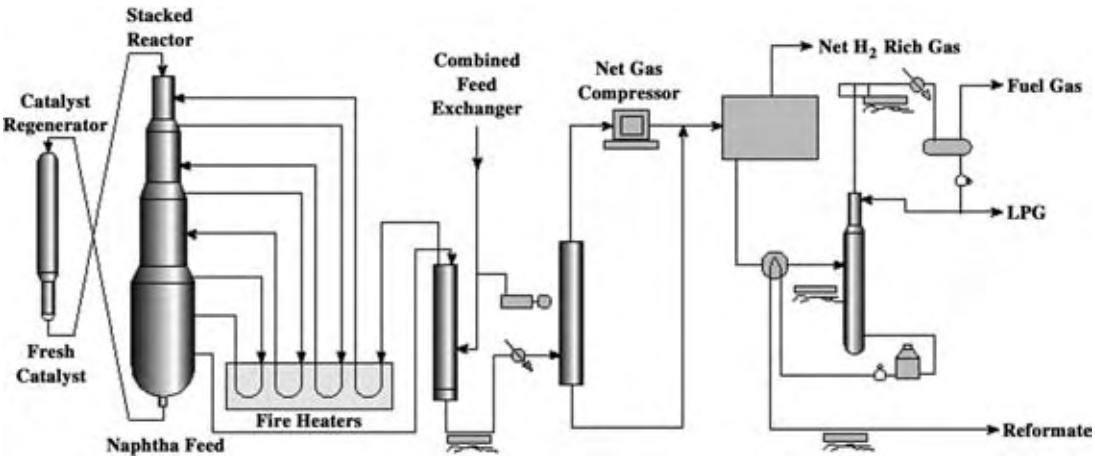


Fig. 1 Schematic of UOP CCR platforming process.

The high-efficiency regenerator design resulted in an increased coke-burning capacity with reduced regeneration severity and complexity. CycleMax regenerator provided easier operation and enhanced performance over other regenerator designs. The operation at ultra-low pressure (3.4 bar) and the use of low-platinum catalyst ensured the highest yield of the reformate and aromatic product with more cost-effective process operation. The net gas recovery schemes maximized the yields of reformate and hydrogen. Moreover, the new regenerator allowed higher regeneration rates to support the coke generation of the low-pressure operation and high conversion levels.

UOP has developed the RZ platforming process for high aromatics selectivity, in the range of 80% or higher. The selectivity of conventional reforming for benzene and toluene is significantly lower than for the C_8 aromatics. Although UOP CCR platforming is the most efficient means for producing xylenes from heavy naphtha fractions, its conversion of C_6 and C_7 paraffins to aromatics is normally below 50%, even at low pressure. In general, RZ platforming configuration is consistent with other UOP platforming systems. The process employs adiabatic, radial-flow reactors that are arranged in a conventional side-by-side pattern. Interheater is used between each reactor to reheat the charge to reaction temperature. Treated naphtha feed is combined with recycled hydrogen and sent to the reactor section. The effluent from the last reactor is heat exchanged against combined feed, cooled, and phase split into vapor and liquid products in a separator. The liquid from the recovery section is sent to a stabilizer where light saturates are removed from the C_{6+} aromatic product. Typical cycle lengths are 8–10 mo and the units are designed for efficient in situ catalyst regeneration.

Table 4 Typical yields of UOP semiregenerative and CCR units for a Middle East naphtha feed

Parameter	Semiregenerative	CCR
Catalyst type	R-72	R-274
Stream factor (day/yr)	330	360
Pressure (kPa)	1380	345
Yield		
Hydrogen (scfd)	1270	1690
C_{5+} (wt%)	85.3	91.6
RON-clear	100	100

(From Ref.[8].)

Octanizing Process

The Axens octanizing reforming technology is based on Institut Français du Pétrole (IFP) and Procatalyse's reforming expertise for the upgrade of various types of naphtha to produce high-octane reformate, BTX, and LPG. The reforming process can be supplied in either semiregenerative or continuous operation.[9–11] The Axens semiregenerative version is a conventional reforming process in which the catalyst is regenerated in situ at the end of each cycle. The operating pressure of this process is in the range of 12–25 bar with low-pressure drop in the hydrogen loop. The product RON-clear is in the range 90–100. Trimetallic catalyst formulations for semiregenerative applications offer higher selectivity and stability.

The decrease in the reformate yield during the run of the semiregenerative version (in spite of improved catalyst stability) has led Axens to develop a catalyst moving-bed system that allows continuous regeneration of the catalyst. The octanizing process is an advanced design that reflects the results of several decades of research and development efforts. The heart of the technology lies in its original catalyst circulation and CCR systems. A schematic flow diagram of the

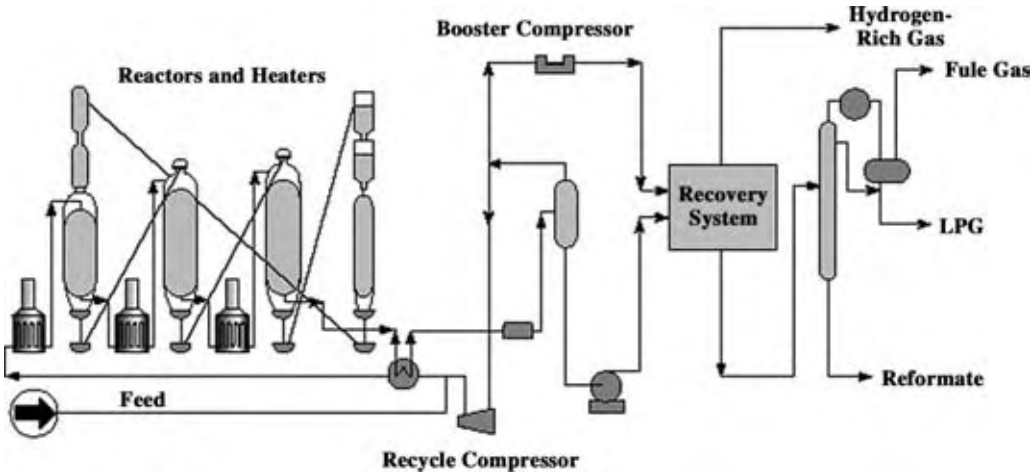


Fig. 2 Schematic of Axens octanizing process.

octanizing process is presented in Fig. 2.[12] The overall process comprises the following:

- A conventional reaction system consisting of a series of four radial-flow reactors using a highly stable and selective catalyst suitable for continuous regeneration.
- A catalyst transfer system using gas lift to carry the catalyst from one reactor to the next and finally to the regenerator.
- A catalyst regeneration section, which includes a purge to remove combustible gases, followed by catalyst regeneration.

The octanizing process features high onstream efficiency, flexibility, and reliability. Major improvements, compared with previous designs, are the development of catalysts of increased activity, selectivity, and hydrothermal stability along with substantial increases in the yields of C_{5+} reformate and hydrogen. Table 5 presents typical yields of the Axens conventional and regenerative process.

Aromizing is Axens CCR reforming process for the selective production of aromatics. It is the petrochemical complement to the octanizing process. The technology employs an advanced catalyst formulation to achieve high BTX aromatics yield and offers high aromatics yields, low investment and operating costs, and high onstream factor. In the octanizing or aromizing process, treated naphtha is mixed with recycled hydrogen, preheated, and passed through a series of adiabatic reactors and heaters, where it is converted to aromatics-rich stream and hydrogen. The effluent is cooled by heat exchange, and the liquid product is separated from recycled and hydrogen gases. Axens regenerative technology has been improved to allow faster circulation of the catalyst and, as a consequence, increased regeneration frequency as required by the more severe operating conditions (low pressure, low H_2/HC ratio).

Table 5 Typical yields of Axens semiregenerative and octanizing processes for a 90–165°C cut light Arabian feedstock

Parameter	Semiregenerative	Octanizing process
Pressure (kg/cm^2)	10–15	<5
Yield (wt%)		
Hydrogen	2.7	3.8
C_{5+}	83	88
RON-clear	100	102
MON-clear	89	90.5

(From Ref.[9].)

Houdriforming Process

The houdriforming process was licensed by the Houdry Division of Air Products and Chemicals, Inc. It can be used to upgrade various naphtha to aviation blending stocks, aromatics, and high-octane gasoline in the range of 80–100 RON-clear. The process operates in a conventional semiregenerative mode with four reactors in series for BTX production, compared with three for gasoline. The catalyst used is usually Pt/Al_2O_3 or a bimetallic. A small ''guard case'' hydrogenation pretreater can be used to prevent catalyst poisons in the naphtha feedstock from reaching the catalyst in the reforming reactors. The guard case is filled with the usual reforming catalyst but is operated at a lower temperature. It is constructed as an integral stage of the houdriforming operation when required for the feedstock.

At moderate severity, the process may be operated continuously for either high-octane gasoline or

aromatics, without provision for catalyst regeneration. However, operation at high severity requires frequent in situ catalyst regeneration. Typical operating conditions are: temperature 755–810 K, pressure 10–27 bar, LHSV 1–4 per hour, and H_2/HC ratio 3 : 6. The total capacity of houdriforming units is about 250,000 b/d.

Magnaforming Process

The magnaforming process, which was licensed by Engelhard Corporation, can be used to upgrade low-octane naphtha to high-octane reformate. The product is a premium blending component or aromatic hydrocarbon source. The feature that most distinguishes the magnaforming process from other reforming processes is its use of a split hydrogen recycle stream to increase liquid yields and improve operating performance. About half of the recycled gas is compressed and recycled to the first two reactors, which are operated under mild conditions. The greater portion of the recycled gas is returned to the terminal reactors that operate under severe conditions. It is believed that substantial compressor power savings can be achieved by splitting the recycle. It is also claimed[3] that the lower H_2/HC ratio in the first two reactors benefits naphthene dehydrogenation, while the higher ratio in the latter reactors provides improved catalyst stability.

Engelhard does not offer a continuous design, but it did provide a semiregenerative design and a combination of semiregenerative and cyclic regeneration. The combination design is made by supplementing the terminal reactors by a swing reactor that can alternate with the terminal reactors. It can also operate in parallel with the terminal reactors, permitting these reactors to be regenerated without unit shutdown. Smaller magnaforming units use a conventional three-reactor system, compared with four reactors in large units.

The process was initially designed to operate with established monometallic platinum catalysts but was adapted to include the newer platinum–rhenium-based catalyst of the E600 and E800 series. The catalysts provide greater activity and stability, enabling use in units where high-severity operation and long cycle lengths are required. A wide range of catalysts have been used in the magnaforming process to optimize operating performance and to produce the desired product specifications. Many units incorporate sulfur guard technology to reduce sulfur in reformer feed to ultralow levels. There are approximately 150 units totaling 1.8 million b/d using the Engelhard reforming technology.

Powerforming Process

Powerforming is offered by ExxonMobil to produce gasoline blending stocks from low-octane naphthas. Alternatively, the process may be operated to give high yields of benzene or other aromatics, or to produce aviation blending stocks. It also produces large quantities of hydrogen that can be used to hydrotreat or improve other products. The process features a semiregenerative or cyclic configuration and proprietary catalyst system tailored to the client's specific needs. Advantages include cost-competitive installations for units less than 12,000 b/d, 98 RON products achievable, mild severity operations, and long cycle lengths.

The staged catalyst system has high stability, good selectivity and selectivity maintenance. It uses the high-activity catalyst with very high stability and benzene/toluene yields. KX-130 is an excellent debottlenecking catalyst or BTX-producing catalyst. The dual catalyst system offers high-activity catalyst, high benzene/toluene yields, and a C_{5+} yield advantage. ExxonMobil's catalyst management techniques for hot flue gas regeneration in cyclic units, onstream chlorination, and analytical tools for monitoring the catalyst help enhance the unit's performance.

Cyclic powerformers are designed to operate at low pressure, with a wide range of feed boiling point and low hydrogen-to-feed ratio. The unit has four reactors in series plus a swing reactor. The use of the swing reactor allows any of the onstream reactors to be taken out of service for regeneration while maintaining continuous operation of the unit. The frequency of regeneration can be varied to meet changing process objectives as well as to operate under high-severity, low-pressure conditions, as coke deposition is maintained at low levels. Regeneration is generally performed on a predetermined schedule to avoid having two or more reactors regenerated at the same time. Usually, the terminal reactors are scheduled for more frequent regenerations than the early-stage reactors. Run lengths of up to 6 yr between shutdowns could be achieved. Powerforming has a wide range of potential refining applications and has been commercially applied in nearly 50 semiregenerative units and over 30 cyclic units ranging from 1000 b/d to 65,000 b/d.

Rheniforming Process

The rheniforming process can be used to convert naphthas to high-octane gasoline blendstock or aromatics plant feedstock. The process gained wide acceptance when Chevron patented the bimetallic Pt/Re catalyst in 1968. Rheniforming is basically a semiregenerative process comprising a sulfur sorber, three radial-flow reactors in series, a separator, and a stabilizer. The process is characterized by the sulfur control sorber, which reduces sulfur to 0.2 ppm in the reformer feed. A new rheniforming catalyst system has been used, which permits low-pressure operation. The high

resistance to fouling of the catalyst system increases the yields of aromatic naphtha product and hydrogen due to the long cycle lengths, which reach 6 mo or more. Optimized operating techniques permit maintenance of high catalyst activity throughout each cycle and return to fresh activity after each regeneration.

The increased resistance to fouling also provides for expansion of existing plants by using higher space velocities, lower recycle ratios, or increased product octane. Converted units operate with H_2/HC ratios of 2.5 : 3.5 and long cycles between regeneration. It is believed that a total of 73 rheniformers are onstream with a total capacity of more than 1 million b/d.

Ultraforming Process

The Amoco's ultraforming process (now BP's) can be used to upgrade low-octane naphthas to high-octane blending stocks and aromatics. The process is a fixed-bed cyclic system with a swing reactor incorporated in the reaction section that is usually specified for aromatic (BTX) production. The system can be adapted to semiregenerative operation with the conventional three radial-flow reactors in series. The process uses rugged, proprietary catalysts, permitting frequent regeneration and high-severity operations at low pressures. The catalyst system has a relatively low precious metals' content, and its estimated life is perhaps 4 yr for cyclic operation vs. 8 yr for semiregenerative operation. The swing reactor in cyclic operation replaces any reactor while the catalyst bed in this reactor is being regenerated. Normally, the reactors are of the same size; however, the first reactor is loaded with half the usual amount of catalyst.

Ultraformers may be designed to produce high-purity xylenes and toluene, which can be separated by straight distillation before the extraction step. The benzene fraction can be recovered by extractive distillation. High yields of C_{5+} reformate and hydrogen have been reported for the ultraforming process. The total capacity of ultraforming is over 530,000 b/d for 39 commercial units worldwide. However, no new ultraformers have been licensed in recent years.

Zeoforming Process

The zeoforming process is a new reforming technology for high-octane gasoline production from hydrocarbon raw material of various origins (straight-run naphtha of gas condensates and crude oils, condensates of accompanying gases, olefin-containing gases, secondary hydrocarbon fractions of refinery and petrochemical plants), boiling up to 200–250°C. The process has been developed by the SEC Zeosit Center of the Siberian Branch of the Russian Academy of Sciences. A special highly stable and selective IC-30-type catalyst based on modified pentasil zeolite is used in the process.[13]

Because of the acid–base shape-selective mechanism of the hydrocarbon conversion in the zeoforming process, the gasoline product has a lower aromatics content compared to the reformate with the same octane number. Reactions proceeding under zeoforming process conditions allow benzene concentration to decrease below 1 vol% and to produce gasoline with low sulfur content without naphtha hydrotreating owing to insensibility of the catalyst to sulfuric compounds in the feedstocks. The process is based on the catalytic conversion of linear paraffins and naphthenes into iso-paraffins and aromatics over zeolite-containing catalysts, which allows the motor octane number (MON) of naphtha to increase from 45–60 up to 80–85.

Stepanov et al.[13] reported that a small-scale plant of 120 b/d of naphtha was operated successfully in the north of Siberia, Russia in 1992. Catalyst life reached more than 1.5 yr. A new industrial plant based on zeoforming process of 1000 b/d capacity has been in operation at Glimar Refinery in Poland since 1997.

COMMERCIAL REFORMING CATALYSTS

In the 1950s, naphtha-reforming catalysts were essentially heterogeneous and monometallic and composed of a base support material (usually chlorided alumina) on which platinum metal was placed. These catalysts were capable of producing high-octane products; however, because of quick deactivation as a result of coke formation, they required high-pressure, low-octane operations. Typically, the time between regenerations is a year or more.

In the early 1970s, bimetallic catalysts were introduced to meet the increasing severity requirements. Platinum and another metal (often rhenium, tin, or iridium) account for most commercial bimetallic reforming catalysts. The Pt–Re catalyst has proven to be reliable and easily regenerable, making it the primary catalyst in use today in semiregenerative units.[14] The catalyst is most often presented as 1/16, 1/8, or 1/4 in. Al_2O_3 cylindrical extrudates or beads into which Pt and other metals have been deposited.

In commercial catalysts, platinum concentration ranges between 0.3% and 0.7% and chloride is added (0.1–1.0%) to the alumina support (eta or gamma) to provide acidity. Those of CCR are mainly bimetallic Pt–Sn on chlorided alumina in a spherical form. The advantages include increased activity and selectivity during paraffin aromatization and decreased rate of catalyst deactivation.

At present, there are six international vendors of reforming catalysts producing more than 80 different types of catalysts suitable for different applications

and for a variety of feedstocks. The current demand for reforming catalysts is mainly a replacement market with about 75–80% of it as bimetallics. Reforming catalyst vendors continue to develop new catalyst formulations designed to meet a wide array of challenges. Many of these challenges involve environmental regulations that refiners have been, and will be, required to meet during the coming few years. Table 6 presents a compilation of commercial reforming catalysts that are available by sale or license to refiners.[15] The list provides information on catalyst supplier, catalyst type, and other selected catalyst properties.

Table 6 List of commercial naphtha reforming catalysts

Catalyst designation	Type	Application	Active agents on alumina
Axens			
AR-405, -501	Bimetallic spherical	Aromatic production	Pt–Sn
CR-201, -301, -401	Bimetallic	Continuous	Pt–Sn
CR-502	Monometallic	Cyclic, semiregenerative	Pt
CR-701, -702	High stability bimetallic	Continuous	Pt–Sn
RG-412	Monometallic	Semiregenerative	Pt
RG-534	Bimetallic	Cyclic	Pt–Re
RG-492	Skewed bimetallic	Semiregenerative	Pt–Re
RG-534	Monometallic	Cyclic	Pt
RG-582, -682	Bimetallic	Semiregenerative	Pt–Re
Criterion Catalyst Co.			
P-15	Mono high activity	Semiregenerative	Pt–Cl
P-93, -96	Monometallic	Semiregenerative	Pt–Cl
PHF-43, -46	Monometallic	Semiregenerative	Pt–Cl
PR-9, -11	Multimetallic	Continuous	Pt–Sn–Cl
PR-28, -291	Multimetallic	Continuous	Pt–Sn–Cl
PS-7, -10, -20, -30, -40	Multimetallic	Continuous	Pt–Sn–Cl
ExxonMobil Research & Engineering			
KX-120; KX-130	Multimetallic	Semiregenerative, cyclic	Pt–Re–Cl
KX-130	Multimetallic	Semiregenerative, cyclic	Pt–Ir–Cl
KX-160, 170	Multimetallic	Semiregenerative	Pt–Re–Cl
KX-190	Multimetallic	Cyclic	Pt–Sn
Indian Petrochemicals Corp.			
IRC-1001	Monometallic	Semiregenerative	Pt–Cl
IRC-1002	Monometallic low Pt	Semiregenerative	Pt–Cl
IPR-2001	Bimetallic	Continuous	Pt–Re–Cl
IPR-3001	Multimetallic	Continuous	Pt–Re
Instituto Mexicano del Petroleo (IMP)			
RNA-1	Bimetallic	Semiregenerative	Pt–Re
RNA-1(M)	Bimetallic wider range	Semiregenerative	Pt–Re
RNA-2	Bimetallic trilobe	Aromatics	Pt–Re
RNA-4	Bimetallic	CCR	Pt–Sn
UOP			
R-50	Bimetallic	Semiregenerative	Pt–Re
R-55	Monometallic	Semiregenerative	Pt
R-56	Bimetallic	Semiregenerative	Pt–Re
R-62	Bimetallic spherical	Semiregenerative	Pt–Re
R-72	Proprietary spherical	Semiregenerative	P
R-85	Monometallic	Semiregenerative	Pt
R-86	Bimetallic	Semiregenerative	Pt–Re
R-132, -134	Bimetallic	Continuous	Pt–Re
R-162, -164	High density; bimetallic	Continuous	Pt–Re
R-232, -234	Low coke; bimetallic	Continuous	Pt–Re
R-272, -274	High yield; bimetallic	Continuous	Pt–Re
RZ-100	Monometallic	Aromatics	Pt

(From Ref.[15].)

CONCLUSIONS

Catalytic reforming remains the primary process in the refinery for producing high-octane gasoline to be blended into the gasoline pool. As the needs for gasoline have risen, the demands on the reformer have also increased. The catalytic reformer is one of the few units to provide hydrogen in the refinery to meet clean fuel requirements. Reforming is also a major source of aromatics for petrochemical production. Nowadays, the catalytic reformer produces more hydrogen by increasing the severity of operation or by improving selectivity to aromatics. New reforming technology, in the form of new catalysts or minimal revamp process improvements, has improved the operation of the catalytic reformer.

ACKNOWLEDGMENT

The author acknowledges the support of the King Fahd University of Petroleum & Minerals, Dhahran, during the preparation of the manuscript.

REFERENCES

1. Edgar, M. Catalytic reforming of naphtha in petroleum refineries. In *Applied Industrial Catalysis*; Leach, B., Ed.; Academic Press: New York, 1983; Vol. 1, 123.
2. Lovink, H. Naphtha hydrotreatment. In *Catalytic Naphtha Reforming*; Antos, G. et al., Eds.; Marcel Dekker: New York, 1995; 257.
3. Little, D. *Catalytic Reforming*; PennWell: Tulsa, OK, 1985; 25.
4. Nakamura, D. Worldwide refining capacity creeps ahead in 2004. Oil Gas J. **2004**, Dec 20, 40.
5. Parera, J.; Figoli, N. Reactions in the commercial reformer. In *Catalytic Naphtha Reforming*; Antos, G. et al., Eds.; Marcel Dekker: New York, 1995; 79.
6. Aitani, A. Reforming processes. In *Catalytic Naphtha Reforming*, Antos, G. et al., Eds.; Marcel Dekker: New York, 1995; 409.
7. Godwin, G.; Moser, M.; Marr, G.; Gautam, R. Latest developments in CCR Platforming catalyst technology, Presented at the 40th International Petroleum Conference, Slovak Research Institute of Pertoleum and Hydrocarbon Gases: Bratislava, Sep 2001.
8. Dachos, N.; Kelly, A.; Felch, D.; Reis, E. UOP Platforming process. In *Handbook of Petroleum Refining Processes,* 2nd Ed.; Meyers, R., Ed.; McGraw-Hill: New York, 1997; 4.3.
9. Refining Handbook 2000. In *Hydrocarbon Processing*; Gulf Publishing Company: Houston; Nov 2000; 97.
10. Clause, O.; Dupraz, C.; Frank, J. Continuing innovation in catalytic reforming, Presented at the NPRA Annual Meeting, San Antonio, Mar 1998.
11. Le Goff, P.; Pike, M. Increasing semi-regenerative reformer performance through catalytic solutions, Presented at the 3rd European Catalyst Technology Conference, Euro Petroleum Consultants (EPC): Amsterdam, Feb 2002.
12. *Axens Process Brochures for Octanizing and Aromizing Processes*; Axens IFP Group Technologies, Rueil-Malmaison, Paris, France, 2002.
13. Stepanov, V.; Snytnikova, G.; Ione, K. A new effective process for motor gasoline production over zeolite catalysts, Presented at the 5th European Congress on Catalysis, European Federation of Catalysis society (EFCATS): Limerick, Sep 2001.
14. Antos, G.; Moser, M.; Lapinski, M. The new generation of commercial catalytic naphtha-reforming catalysts. In *Catalytic Naphtha Reforming*; Antos, G., Aitani, A., Eds.; Marcel Dekker: New York, 2004; 335.
15. Stell, J. Catalyst developments driven by clean fuel strategies. Oil Gas J. **2003**, Oct 6, 42.

Centrifuges

Alan G. Letki
Alfa Laval Inc., Warminster, Pennsylvania, U.S.A.

INTRODUCTION

Centrifuges are a means of mechanical separation. They can be used to separate insoluble solids and liquids mechanically based on particle size and/or density, or immiscible liquids based on density. Most commonly, the solids have a higher density than the liquid(s) from which they are separated. Centrifuges occupy two separate branches of the mechanical separation tree (Table 1), as they are of value in enhancing both sedimentation-based and filtration-based separations.[1] Their design has evolved since the time when Swedish engineer Gustaf de Laval in 1878 patented the first successful industrial centrifuge—a cream separator. Operating in batch, batch automatic, or continuous modes, centrifuges are now used by process industries to perform many separation duties. These include clarification, classification, dewatering, extraction, purification/concentration, rinsing, washing, and thickening. The aim of this entry is to describe general operating characteristics. More theoretical issues are discussed elsewhere.[2]

SEPARATION DUTIES

Each separation duty and its function is described below.

Clarification is the separation of insoluble particles from a liquid. The goal is to maximize the clarity of a single liquid phase via solid removal.

Classification splits a slurry into two streams. The aim is to achieve the separation of solids based on particle size and/or density differences.

Dewatering is the removal of the vast majority of solids from a slurry to produce a cake that has body and integrity. The objective is to produce a conveyable/truckable cake, increase the fuel value, or reduce the processing cost of the feed solids by removing the liquid, which is usually water.

Extraction involves the mixing of a liquid, typically an immiscible organic liquid with an aqueous slurry. The purpose is to transfer the solute from the aqueous to the organic phase and recover the organic phase.

Purification and concentration describe the separation of two immiscible liquids, with or without the presence of insoluble solids. The former process aims to produce a very clean light-phase liquid, whereas the latter seeks to yield a very clean heavy-phase liquid.

Rinsing is the removal of mother liquor from suspended solids after separation but before they are removed from the centrifuge. The intent is to produce a less contaminated solid.

Washing is the countercurrent or cross-current cleaning performed by dissolving impurities in suspensions or crystal/amorphous solid slurries. The objective is to use a series of centrifuges to enhance the purity of the solids or to minimize the amount of wash liquid.

Thickening is the removal of the vast majority of suspended solids from a liquid to produce a fluid solid stream of acceptable viscosity. The usual aim is to reduce the liquid content and maintain a pumpable/mixable solid slurry.

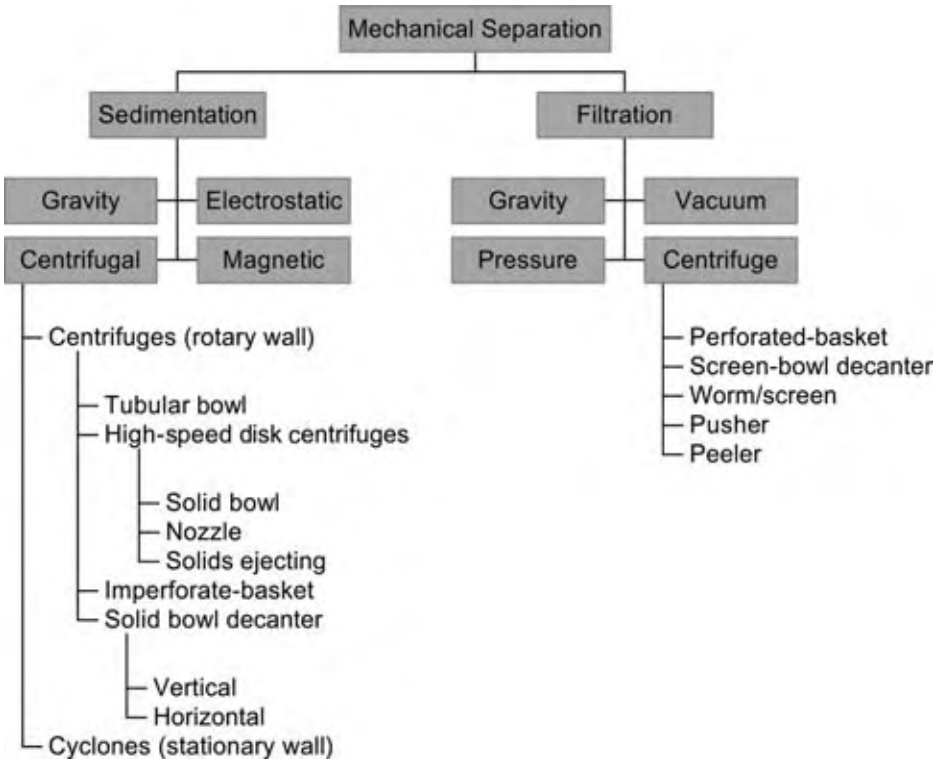
SEDIMENTATION VS. FILTRATION

The basic difference between filtration and sedimentation is media dependence. Filtration removes solids from a liquid by directing the stream at the media. It succeeds by collecting the solids on, or within the thickness of, the media, while allowing the liquid to pass through. The process is usually used for materials that form incompressible (rigid) cakes, especially if rinsing is required. Premixing of the slurry with additives (filter aid) is done sometimes to provide the necessary cake consistency. At other times, a precoat layer, normally in conjunction with cloth, may be used to enhance filtration. Sedimentation is employed on materials that form compressible cakes, especially soft, slimy, gelatinous materials, or on those materials with particle sizes too small to be efficiently collected by filter media. Whether a cake acts compressibly or incompressibly may vary with particle size distribution and the shapes of the mixture of particles, and may be affected by the G-level applied.

While most solids have a higher gravity than the liquor from which they are separated, floating solids that are usually not captured in ordinary sedimenting centrifuge may well be collected in filtering centrifuges if the particles are large enough.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120006800

Table 1 Mechanical separation tree



MECHANICAL COMMONALITY

Centrifuges consist of the following common mechanical parts: a bowl that generally rotates on either a horizontal or a vertical axis in which the separation takes place, a casing that collects the solids or liquids discharged from the bowl and diverts them in different streams away from the bowl, a frame that supports the casing and bowl, and a drive mechanism to supply power for turning the bowl. Some method of feed input/distribution into the bowl as well as a means of discharging solids and/or liquids from the bowl is also required. Sedimenting centrifuges have mostly solid or imperforate bowls, while perforate ones are used to support the media, usually screens or cloths of filtering centrifuges.

THE ROTATING BOWL

The fundamental mechanical component of a centrifuge is a rotating bowl. When properly designed and utilized, it augments separation. The magnitude of the enhancement is based both on the radius of the bowl and on the speed of rotation, and is sometimes incorrectly described as G-force. The relative centrifugal force or G-level is not a force—it is the ratio of the acceleration of the centrifugal field to the acceleration caused by the earth's gravity, which is a dimensionless number:

$$G_s = \omega^2 r/g \text{ dimensionally } (1/\text{time}^2 \times \text{distance})/(\text{distance}/\text{time}^2) \text{ dimensionless} \quad (1)$$

This ratio or level may reach 60,000 on small laboratory units, 20,000 on small industrial scale units, or may be limited to a few hundreds on large industrial centrifuges. Although the level tends to decrease as the size of the rotor increases, it is normally large enough so that a rotor rotating horizontally is considered to have the same separating capacity as it would have if it rotated vertically, i.e., the influence of the earth's gravitational field is negligible while the rotor is spinning. The choice of a vertical or horizontal centrifuge is therefore based on other factors than the effect of the earth's gravitational pull on the material being centrifuged.

The rotation of the bowl results in power being consumed and stresses being placed on the rotating bowl shell.[3]

SEDIMENTATION

Starting with Newton, when a force is applied to a particle, it is accelerated:

$$F = ma \quad (2)$$

In a static settling tank under the influence of the earth's gravity, the particle settles along the radius of the earth. When g is the gravitational constant,

$$F = mg \quad (3)$$

In a centrifugal field, the acceleration, $\omega^2 r$, results in a force that acts normal to the axis of rotation:

$$F = m\omega^2 r \quad (4)$$

In a sedimenting centrifuge, a continuous liquid phase moves through the rotor. To accomplish a useful separation, the discontinuous phase, either the insoluble solids or the immiscible liquid drops (or both), must move in a direction different from the flow of the continuous liquid. Stokes' law is usually applied to describe this relationship. The effective force accelerating the particle in a centrifugal field is then given as:

$$F_p = (m_p - m_l)\omega^2 r \quad (5)$$

where m_p is the mass of the particle and m_l is the mass of the liquid displaced by the particle. If we define $\Delta\rho = (\rho_p - \rho_l)$, the difference in the density between the particle and the continuous liquid phase, then for a spherical particle of diameter D:

$$F_p = (\Pi/6)\Delta\rho D^3\omega^2 r \quad (6)$$

If the diameter is small or the viscosity is high, the particle moves at a velocity below the turbulent range and Stokes' law defines the force of the liquid phase resisting the particle as:

$$F_l = 3\Pi\eta D v_s \quad (7)$$

If the particle settles long enough (reaches equilibrium), then

$$F_l = F_p \quad (8)$$

Then, in a centrifugal field:

$$v_s = (\Delta\rho D^2\omega^2 r)/18\eta \quad (9)$$

In the earth's gravitational field:

$$v_g = (\Delta\rho D^2 g)/18\eta \quad (10)$$

The difference between the velocity in the centrifugal field (v_s) and in the earth's gravitational field (v_g) is an essential reason for utilizing sedimenting centrifuges. In a centrifuge, as in a gravity thickening device, the diameter of the particle, the difference in density between the particle and the continuous-phase liquid, as well as the viscosity of the liquid, are important process parameters.

From an operational standpoint, it may be useful to consider that in the earth's gravitational field, if the center of the earth were viewed as "downhill"—the direction in which movement usually occurs—the direction in a centrifuge corresponding to this downhill movement would be away from the center of rotation, i.e., from smaller to larger radius.

SIGMA THEORY

A method of sizing/scaling centrifuges called the sigma theory had been developed by Ambler over half a century ago.[4] Ambler recognized that there are two sets of factors influencing separation performance: those that are characteristic of the feed and others that are typical of the centrifuge. The former include particle size, densities, and viscosity and the latter include bowl speed and bowl dimensions. The theory calculates the equivalent area of a centrifuge so that the feed rate to the centrifuge divided by the sigma factor yields a clarification capacity in units of mass/area/time. Other slightly different factors have been advanced over the years.[5]

INLETS AND OUTLETS

Considerable engineering effort is typically expended in the design of inlets and outlets. In selecting the correct bowl type and operating speed, it is important to ensure that the advantages achieved not be diminished by either the inlet or the outlet chosen for the desired application.

If, as discussed, the square of particle diameter is related to settling velocity, it is obvious that separation may be enhanced by not reducing the diameter of the particle in the inlet to the centrifuge. It is perhaps less obvious that the type of discharge/outlet may reduce particle size and be detrimental to further downstream processing.

The inlet, including the feed pipe, serves several purposes and should therefore exhibit various properties. One among them is that it should be free from plugging. To achieve this, grinders may be used (especially on waste applications). When feed pipe nozzles are used, screens may be employed ahead of the centrifuge to remove oversized materials and prevent plugging. In addition to carrying feed slurry into the centrifuge, concentrically designed tubes may also be used to transport a second or even a third fluid into

the centrifuge. In horizontal decanters, the inlet may convey a flocculent that is added internally to flocculate shear-sensitive materials. In screen bowls, separate rinse liquor for distribution as screen rinse may also be transported. A similar function is also observed with cleaning solutions, especially on hygienic applications. The inlet also helps determine the characteristics of the feed system. It may be a single open pipe discharge as in the case of horizontal decanters or disk centrifuges that would have a fairly low pressure drop associated with them. It could be a nozzle to jet the feed into the bowl as is typical of the tubular bowls and vertical decanters that would require higher pressure. In the case of some hermetic disk centrifuges, the feed system is used to cause flow through the entire centrifuge. With high solid slurries or particles that tend to settle quickly, the feed pipe may have several discharge points to evenly distribute solids, especially in large basket centrifuges, and may be fed from a circulating loop to prevent settling of the feed slurry.

Outlets of both solids and liquids, if not properly chosen, may adversely affect separation to the extent that it may even be impossible to effect separation in the bowl. Because choosing the correct outlet size can be critical, most sedimenting centrifuges are designed so that at least the liquid outlet radius can be changed relatively easily. Selecting the appropriate outlet radius is particularly important in separating immiscible liquids because two must be chosen.[6]

Open discharges, i.e., the simple flow of liquid phase over the bowl weir of sedimenting centrifuges, if properly designed, tend to consume less energy but require more space in the casing. The absence of adequate casing space may promote remixing of discharged components in, or while being discharged from, the casing. Paring disks or skimmer tubes, while consuming power, may be used not only to carry isolated liquid away from the unit, but may also be critical in minimizing air entrainment or foam production on some materials. Although an overwhelming number of materials are acceptably processed, when cake discharge causes unacceptable degradation of some of them, they can be removed from batch operations by completely stopping the centrifuge and hand scooping the material or by manually removing bags or liners designed to fit in the centrifuge rotor.

SEDIMENTING CENTRIFUGES

Sedimenting centrifuges use a solid/imperforate bowl to enhance sedimentation. The most common sedimenting centrifuges listed in the order of decreasing G-level, are as follows: tubular bowl, disk stack separator, decanter, and imperforate baskets.

Several tables are provided to show the 40-fold difference in diameter and the 300-fold difference in attainable G-level. Table 2 lists the separation duties normally associated with the centrifuge type, while Table 3 shows the type of separation and the mode of operation when solids are present in the feed. The typical feed solids and particle sizes associated with the centrifuges are shown in Table 4. Table 5 provides the common feed rate ranges and solid concentrations achieved. Solids in Tables 4 and 5 are given in typical %vol/vol and in typical %wt/wt (insoluble or suspended solids) because the preference for volume or weight measurements varies by application. Tubular bowls (Fig. 1) use a small diameter to allow operation at an extremely high G-level and a relatively high length to diameter ratio. The small diameter limits the stress on the bowl shell allowing relatively common materials such as stainless steel to be utilized in bowl construction. To allow reliable operation, the rotor is vertically mounted and acts as a pendulum. The small diameter and high-speed rotation preclude using a solid-discharge mechanism. Hence, solid–liquid separations require manual solid removal.

Disk stack separators are vertical bowl centrifuges that are characterized by the presence of a nested conical stack of disks. The centrifuges are usually categorized by the method of solid removal from the bowl and include solids-retaining, solids-ejecting, and nozzle-discharge units. The feed enters the bowl at

Table 2 Separation duties

	Type of separation			Separation mode (with solids in feed)		
	Liquid–solid	**Liquid–liquid–solid**	**Liquid–liquid**	**Batch**	**Batch automatic**	**Continuous**
Tubular bowl	X	X	X	X		
High speed solids-retaining type	X	X	X	X		
High-speed nozzle solids-discharging type	X					X
High-speed solids-ejecting type	X	X	X			X
Decanter type	X	X				X
Basket	X			X	X	

Table 3 Separation type and mode

	Basic separation duties							
	Clarification	**Classification**	**Dewatering**	**Extraction**	**Purification**	**Rinsing**	**Washing**	**Thickening**
Tubular bowl	X		X	X	X			
High-speed solids-retaining type	X				X			
High-speed nozzle solids-discharging type	X	X		X	X	X	X	X
High speed solids-ejecting type	X		X	X	X		X	
Decanter type	X	X	X			X	X	X
Basket			X			X		X

the center of rotation, is distributed across the bottom of the bowl, and then flows up through a disk stack. The solids that are more dense than the liquid travel radially outward. Those that enter the disk stack are forced to the underside of the conical-shaped disk, and then along the disk surface to the periphery of the bowl. The lighter liquid moves radially inward and up the stack. In the case of two immiscible liquids, the heavier (more dense) one forms an intermediate position between the light phase and the heavier solids. The presence of the heavy liquid requires the presence of a special top disk and a provision for keeping the liquids separate while exiting the bowl and casing. The settling distance is limited by the spacing between the disks. Because the distance that the particles or immiscible droplets have to settle is considerably shortened, settling capacity is significantly increased. This makes these units the most efficient of clarifiers, purifiers, or concentrators per unit of floor space.

Solids-retaining separators (Fig. 2) are batch centrifuges. They are generally larger in diameter than tubular bowls. Therefore, they are large enough to contain a disk stack but are dependent on at least partial disassembly for manual removal of solids. The larger diameter means that even though they operate at high G-levels, the G-level is somewhat lower than that of a tubular bowl. It also indicates that they have higher solids-holding capacity than the tubular bowl. The bowl shown in this figure is equipped for a liquid–liquid separation. The outlet dams would be adjusted to devote the majority of the disk stacks to the phase requiring greater purity. In the purification process, which aims at producing a clean light phase, the interface (e-line) would be moved outward. On the other hand, in the concentration process, which seeks to yield a clean heavy phase, the e-line would be moved inward. The feed would normally be introduced near the e-line.

Solids-ejecting separators (Fig. 3) are designed for continuous processing with intermittent discharge of solids. The bowl consists of a sliding bottom and a top or hood. At its widest part, the body has a series of discharge openings. In normal operation, the bottom is held against the hood, closing off these openings. When the bottom drops to its lower position, the space between the it and the hood exposes the discharge ports. Accumulated solids are ejected or shot through the ports and into the collecting cover. In the

Table 4 Typical feed concentration and particle size

	Feed rate (gal/min)			Solid dryness							
				vol%				wt%			
	0–5	**5–150**	**>150**	**0–10**	**10–35**	**35–60**	**60–95**	**4–10**	**10–15**	**15–25**	**>25**
Tubular bowl	X					X	X	X	X	X	X
High speed solids-retaining type	X	X		X	X	X	X	X	X	X	X
High-speed nozzle solids-discharging type	X	X	X	X	X	X		X			
High-speed solids-ejecting type	X	X		X	X	X		X	X		
Decanter type	X	X	X		X	X	X	X	X	X	X
Basket	X	X	X				X			X	X

Table 5 Typical feed rate and cake dryness

	Feed solid concentration											Solid particle size (mm)			
	vol%							wt%							
	0–0.1	0.1–1	1–5	5–8	8–20	20–40	>40	0–1.5	1.5–5	5–10	>10	0–1	1–5	5–500	500–10,000
Tubular bowl	X	X	X	X	X			X	X			X	X	X	
High speed solids-retaining type	X	X						X				X	X	X	
High-speed nozzle solids-discharging type					X	X		X	X	X		X	X	X	
High speed solids-ejecting type	X	X	X	X	X			X	X			X	X	X	
Decanter type			X	X	X	X	X	X	X	X	X		X	X	X
Basket				X	X	X	X	X	X	X	X			X	X

bowl shown here, an operating fluid and springs control the opening and closing of the bowl. Numerous variations in discharge methods have been employed. It is generally considered that the faster the shot, the greater the potential for a drier cake.

Nozzle-discharge separators (Fig. 4) consist of a solid bowl that has a double cone shape. In place of discharge ports that are intermittently open, it contains nozzles that are usually along the periphery. As solids flow toward the nozzle, they are concentrated before being continuously discharged. The bowl shown here has been modified and can introduce rinse water or recycled nozzle flow between the disk stack and the nozzle. With the introduction of rinse, the solids are forced through the layer of rinse before being discharged through the nozzle. At least some portion of the rinse liquor tends to travel inward and be discharged through the stack. It cannot be discharged separately from the centrate unless it is immiscible and provision is made for liquid–liquid discharge. The recycling of nozzle flow would allow the use of larger nozzles to produce the same nozzle concentration,

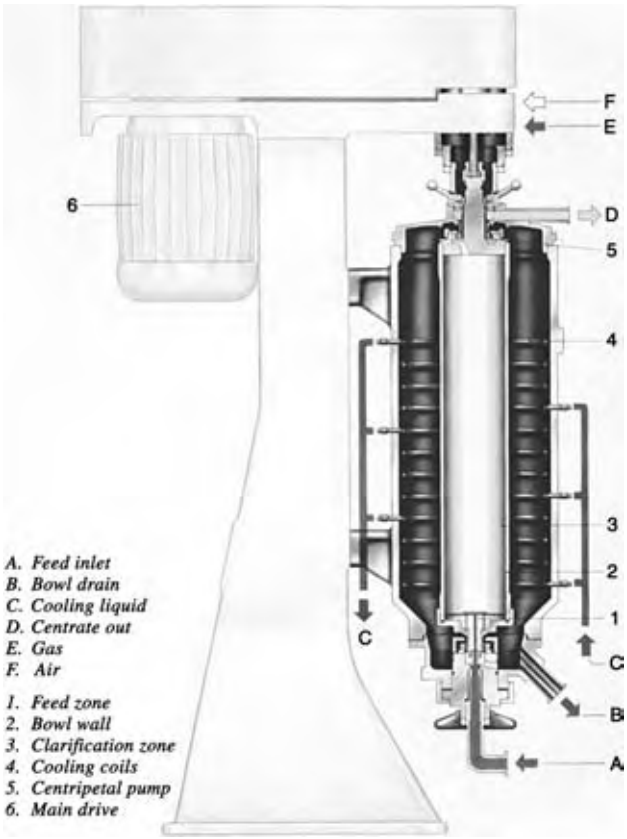


Fig. 1 Tubular bowl centrifuge. *(View this art in color at www.dekker.com.)*

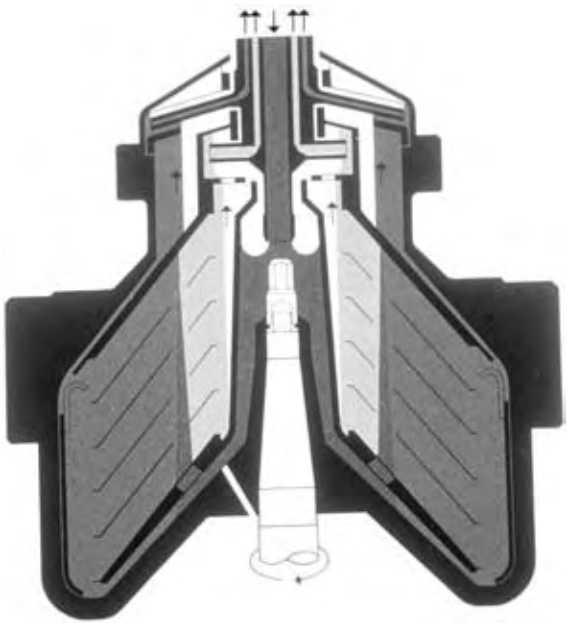


Fig. 2 Solids-retaining disk stack separator. *(View this art in color at www.dekker.com.)*

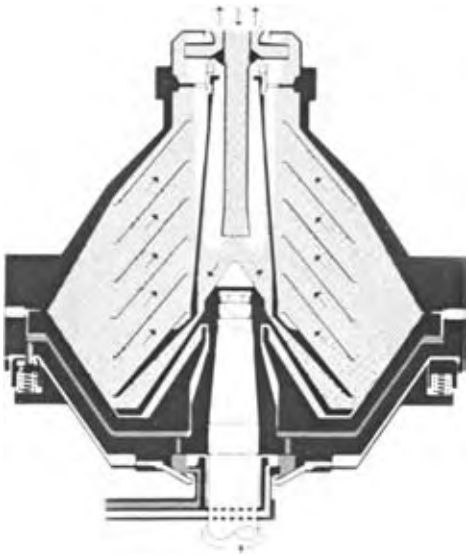


Fig. 3 Solids-ejecting disk stack separator. *(View this art in color at www.dekker.com.)*

or may permit the concentration to be controlled externally via the recycle rate.

Solid bowl decanters (Fig. 5) are characterized by the presence of a screw concentrically located within the rotating bowl, a bowl that is cylindrical with a conical beach extension, and the use of a gear reducer to establish a differential speed between the conveyor and the bowl. The feed material is introduced through a feed tube at the center of rotation. It is discharged inside the conveyor hub and accelerated through feed ports into the bowl. This means that the energy to accelerate the feed is provided by the centrifuge driver—usually an electric motor—and not by the feed pump. In the bowl, the solids are continuously separated from the liquid. Being denser than the liquid, they settle against the bowl wall. The liquid or the

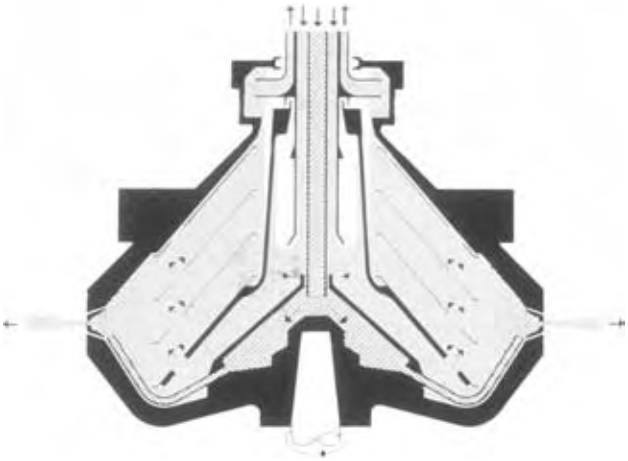


Fig. 4 Nozzle-discharge disk stack separator. *(View this art in color at www.dekker.com.)*

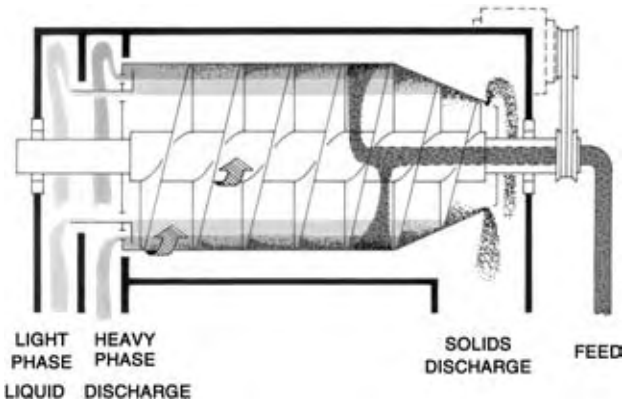


Fig. 5 Solid bowl decanter centrifuge. *(View this art in color at www.dekker.com.)*

two immiscible liquids are decanted over adjustable weirs, usually on the large (cylindrical end) side of the bowl. The conveyor moves at a speed slightly different from the bowl because of the gear reducer, which is often augmented by a variable speed driver—usually electric or hydraulic. This differential speed conveys the solids along the cylindrical portion of the bowl and up the conical beach, discharging them at the opposite end of the bowl from the liquid. If the electric or hydraulic drive is infinitely variable, then the bowl conveyor differential is also infinitely variable over its operating range. The unit can be used on solid–liquid or solid–liquid–liquid separations. By installing nozzles above the beach and using a concentric feed tube, with the inner tube transporting feed slurry and the outer tube carrying rinse liquor, it is possible to do some rinsing in this unit. If the rinse is added on the beach, it must ultimately find its way to the centrate end and be commingled with the centrate before discharge from the bowl.

Solid bowl scroll decanter centrifuges have also been built vertically. The pendulum-type construction allows the bowl to expand more easily when operating at higher temperatures and makes mechanical sealing for elevated pressures less cumbersome. Although these units have been designed for operation up to 10 bars, they often operate at 5 bars or less.

Imperforate basket centrifuges are vertical baskets that make up a relatively small portion of those sold. They are most often used to collect fine or amorphous solids, especially slimy or sticky ones. The solids are fed into the basket and those that are heavier than the liquid accumulate on the wall, much as they do in the tubular bowl. Because the diameter of the basket is at least an order of magnitude larger than that of the tubular bowl, there is usually space for a discharge mechanism inside the basket. Imperforate bowls often have two instruments: a skimmer for liquids and a plow for solids. The clear, less dense liquid supernates

and is discharged over a top ring or is removed through a skimmer tube. After the liquid is removed, the plow is inserted and the solids usually drop out through the bottom. As the diameter becomes larger than 4–5 ft, the G-level drops to a few hundred times *g*. As the unit is batch operated, the spin times can be increased, which sometimes compensates for the decreased G-level.

THE BATCH CYCLE

Centrifuges may be batch, batch automatic, or continuous. Handling of the separated solids is the primary reason for the difference in operating modes. Therefore, a centrifuge type may be a batch operation when separating solids from liquids, but a continuous one when separating two immiscible liquids.

Although from an economic and production standpoint the batch cycle analysis certainly applies to the tubular bowl and the solids-retaining disk stack centrifuge, the batch cycle (Fig. 6) is most often considered when operating basket centrifuges. The cycle as shown is most often associated with a filtering centrifuge. The feeding, spinning, and rinsing can be individually varied to optimize process performance. Likewise, the bowl speed for spinning and discharge is sometimes altered for optimal performance, usually to avoid compression or to minimize particle breakage. Vertical baskets or pendulum-type units are usually operated at variable basket speeds to allow higher G-levels for dewatering and lower levels to minimize particle breakage on discharge. The peeler is typically a horizontal basket that plows the cake for discharge at high speeds to minimize the cycle time. For vertical baskets, the cycle and its component times are usually measured in minutes, and in extreme cases, in hours, while the corresponding cycle component times in peelers may be measured in seconds. In lieu of accelerating and decelerating times, the peeler may have an additional rinsing time after the solids are discharged. The residual solids on the screen may be dissolved on an intermittent basis. This conditioning of the screen can be programmed into the batch cycle to maintain drain rates and cycle times.

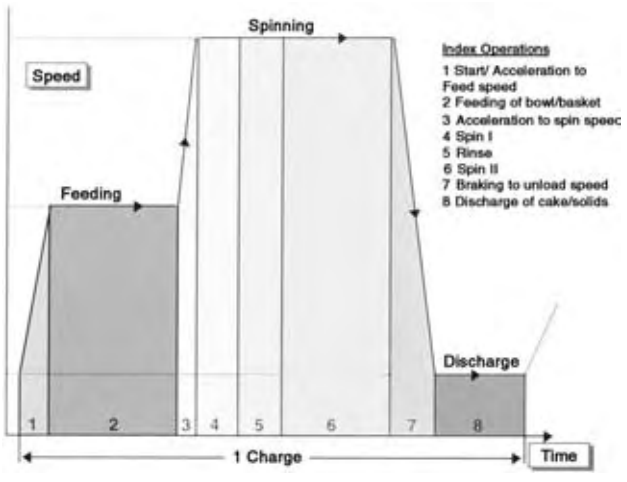


Fig. 6 Batch cycle diagram. *(View this art in color at www.dekker.com.)*

The batch cycle with respect to imperforate bowls would most often substitute skimming for rinsing. As the solids settle against the bowl wall, there may be a clean liquid layer in the basket above the settled solids. To maximize cake dryness, the clear layer is taken off through a skimmer tube. There may be one or more feed and skim cycles before the cake solids are plowed out for discharge.

FILTERING CENTRIFUGE

While the tendency is to consider hydraulic rates on sedimenting centrifuges, filtering units are commonly considered in terms of tonnages. High tonnages are often achieved, not by processing high liquid rates, but by processing high to extremely high slurry solid concentrations. Mechanical comparisons are often made on filtering area or volume per batch/tonnage per cycle. Cake thickness is an important consideration as is rinse efficiency.

The screen bowl decanter (Fig. 7) can be thought of as a horizontal solid bowl decanter with a cylindrical screen inserted between the top of the conical beach and the bowl discharge. The screen opening is of the order of 150 mesh. In addition to having the ability to segregate the rinse liquid from other centrifuge discharges, the decanter allows cake rinsing to be much more comparable to other filtering centrifuges than to the solid bowl. Alternately, it can provide a clean centrate and a more uniform crystal size at the cake discharge by allowing the fine particles to pass through the screen and be isolated for processing separately from the cake and centrate discharges.

The vertical or pendulum perforate basket (Fig. 8) is perhaps the most commonly used centrifuge. The perforated basket in its simplest form is lined with a cloth, wire mesh, or a bag that accumulates solids as the liquid passes through. When the solid space is filled, feeding is discontinued and the relevant portion of the batch cycle is continued. In smaller units, the bowl may simply spin down to a complete stop for manual removal of the cake or replacement of the bag. At higher G-levels or in larger diameter units, wire mesh or screen is added to support the filter media. Slowing the unit and plowing the solids out at low speed complete the batch cycle. Rinse can be isolated

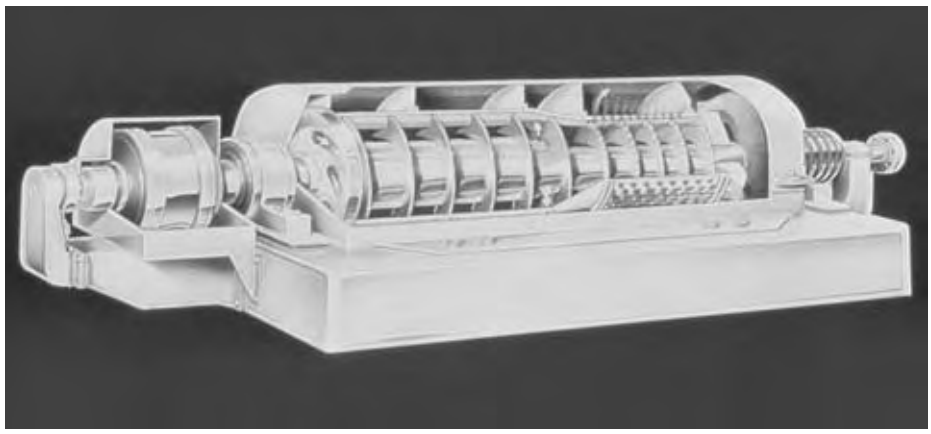


Fig. 7 Screen bowl decanter centrifuge. *(View this art in color at www.dekker.com.)*

from centrate by diverting the liquid discharge at the appropriate time.

Major types of horizontal perforate basket centrifuges are peelers, pushers, and worm/screen centrifuges. The fundamental purpose of the horizontal operation is to keep the operating mechanisms, including drives, isolated from the process streams. The variation in size, G-level, and application is not nearly as high as it is with the sedimenting centrifuges. The fundamental concern is the processing of crystals, with emphasis on throughput, dryness, particle integrity, and purity.

Peelers are characterized by conical baskets (Fig. 9) that are fitted with a heavy-duty plow mechanism that ''peels'' the cake away from the bowl for discharging at operating speed. This minimizes the batch cycle time and is normally used to process more rugged crystals. Peelers may provide cake rinse and may be equipped to dissolve the residual heel cake on line, promote drainage, and maintain short cycle times.

Cylindrical screen sections over which solids are ''pushed'' by the reciprocating action of a piston characterize pushers. They may consist of one, two, or multiple stages. The cake may receive one or more rinses, if required.

A conical screen with an internal screw or ''worm'' characterizes the worm/screen. Because the crystals are

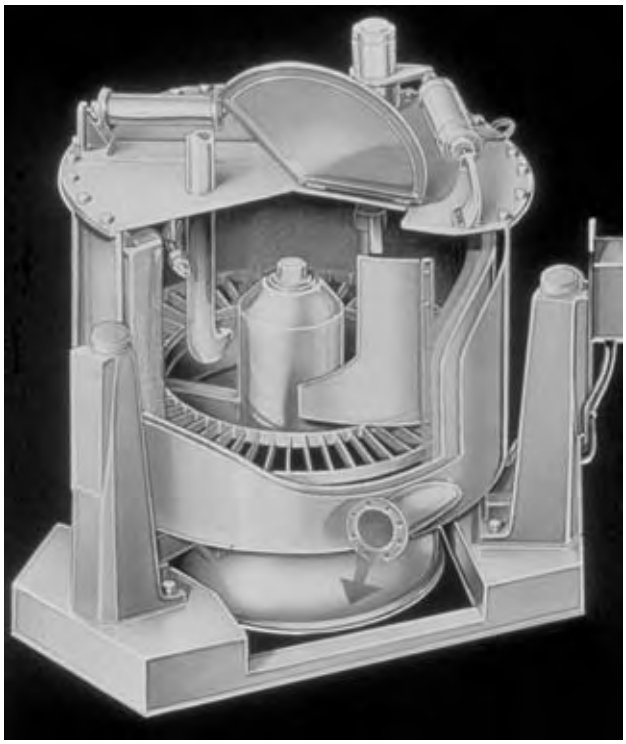


Fig. 8 Vertical perforate basket centrifuge. *(View this art in color at www.dekker.com.)*

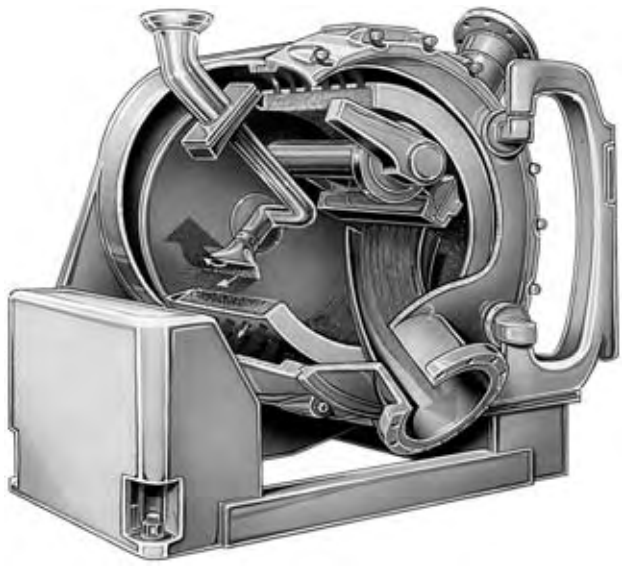


Fig. 9 Peeler centrifuge. *(View this art in color at www.dekker.com.)*

fed at the small end of the cone, the internal screw helps control the rate at which the crystal travels down the screen to discharge. The differential speed between the screw and the screen is sometimes externally adjustable. Multiple rinses may be employed and are sometimes kept separate from each other.

CONCLUSIONS

This entry has presented a brief description of sedimenting and filtering centrifuges. The types of separation, separation duties, and modes of operation have been discussed. The basic common and distinct mechanical elements have also been dealt with. References are included for those readers seeking a more in-depth historical, theoretical, as well as practical background on the subject.

REFERENCES

1. Letki, A.G. Know when to turn to centrifugal separation. Chem. Eng. Progr. **1998**, *Sep*, 29–24.
2. Ambler, C.M. Theory. Ind. Eng. Chem. **1961**, *53*, 430–433.
3. Letki, A.; Moll, R.T.; Shapiro, L. Separation, centrifugal. In *Encyclopedia of Chemical Technology*, 4th Ed.; Kirk-Othmer, Ed.; John Wiley and Sons Inc.: New York, 1997; Vol. 21, 826–875.
4. Ambler, C.M. The evaluation of centrifuge performance. Chem. Eng. Progr. **1952**, *Mar*, 150–158.
5. Delogio, T.; Letki, A.G. New directions in centrifuging. Chem. Eng. **1994**, *Jan*, 70–76.
6. Axelsson, H.; Madsen, B. Sedimenting centrifuges. In *Ulmann's Encyclopedia of Industrial Chemistry*, 6th Ed., Wiley–VCH Verlag GmbH & Co: Weinhein, 2000.

Ceramics

Stephen J. Lombardo
Department of Chemical Engineering, University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

This entry will be divided into sections covering the classification of ceramics, a brief history of ceramics, raw materials for ceramics, properties and applications of ceramics, processing of ceramics, and a brief commentary on future trends. The field of ceramics is very broad and encompasses not only well-known, conventionally used materials and technology but also much newer compositions, processing methods, and applications. More detailed information on all of these sections is available in printed format[1–9] and on the web.[10–14]

CLASSIFICATION

Ceramics, which come from the Greek words *keramikos* and *keramos* for pottery or potter's clay, are defined as inorganic nonmetallic solids and are a separate class of materials as compared to metals or organic solids such as polymers and plastics. This simple definition, which encompasses many materials, can be divided into broad application categories as whiteware, structural clay, refractory, glass, abrasive, advanced, and other (see Fig. 1). These groups can be further divided into more specific market segments. The last heading in Fig. 1 of "Other" contains cement, diamond, graphite, and silicon. Although these materials do not completely fit the definition of ceramic materials, they also do not fully meet the criteria to be classified solely as organic or metallic solids either. They are thus often included with the more conventional ceramics in Fig. 1 because of similarity in properties, application, or processing.

In addition to grouping ceramics by general application as in Fig. 1, alternative taxonomic schemes are possible. Table 1 indicates ways that ceramics can be classified according to application, composition, crystal structure, forming method, microstructure, and synthesis method. The subheadings in each category are often combined in conversational usage so that one may refer to "ferroelectric electronic ceramics," "chemical vapor deposited oxide films," "single crystal sensors," or "monolithic ceramics for wear applications."

BRIEF HISTORY

In the social, cultural, and technological evolution of mankind, ceramic materials have played a significant role. The earliest man-made ceramics, which were figurines, were produced around 24,000 B.C. By 10,000 B.C., vessels fabricated from ceramic materials were in use. By 5000 B.C., a number of cultures worldwide used ceramics in basic functions such as storage vessels for food and drink, and as cookware; ceramics also served as decorative, ornamental, and possibly symbolic functions in different cultures as well. By 1500 B.C., glass products were produced in many areas of the world. These early developments of ceramic technology, which postdate the Stone Age and predate metal processing, are exemplary of mankind's creativity whereby common raw materials are transformed by sequential processing steps into value-added, often highly functional products. In fact, the basis for many aspects of modern ceramic processing can be found in technologies developed hundreds or thousands of years ago.

Today, ceramic materials find widespread applications in both the consumer and industrial market segments. In the home, ceramics are commonly used in kitchens and bathrooms and for decoration and ornamentation. As construction materials, ceramics are widely used as bricks, tiles, and concrete sidewalks.

Although many of the common uses of ceramics listed above are well known, the foundation role that ceramic materials have played in technology development and in modern society is often less widely appreciated. Ceramic materials are technology enablers for the processing of a variety of commodity and advanced materials. For example, the inertness of ceramic crucibles and refractory bricks enables the high-temperature processing of metal alloys and glass. Ceramic and glass materials support silicon wafers in many of the processing steps used to manufacture integrated circuits. Ceramic coatings on gas turbine blades allow for operation at higher temperature, which leads to more efficient energy generation.

In addition to their role as technology enablers, ceramic components find widespread use today in electronics, wireless communications, fiber optic cables, lasers, digital data storage, and capacitors. Emerging technologies in which ceramics play a key role include

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007696

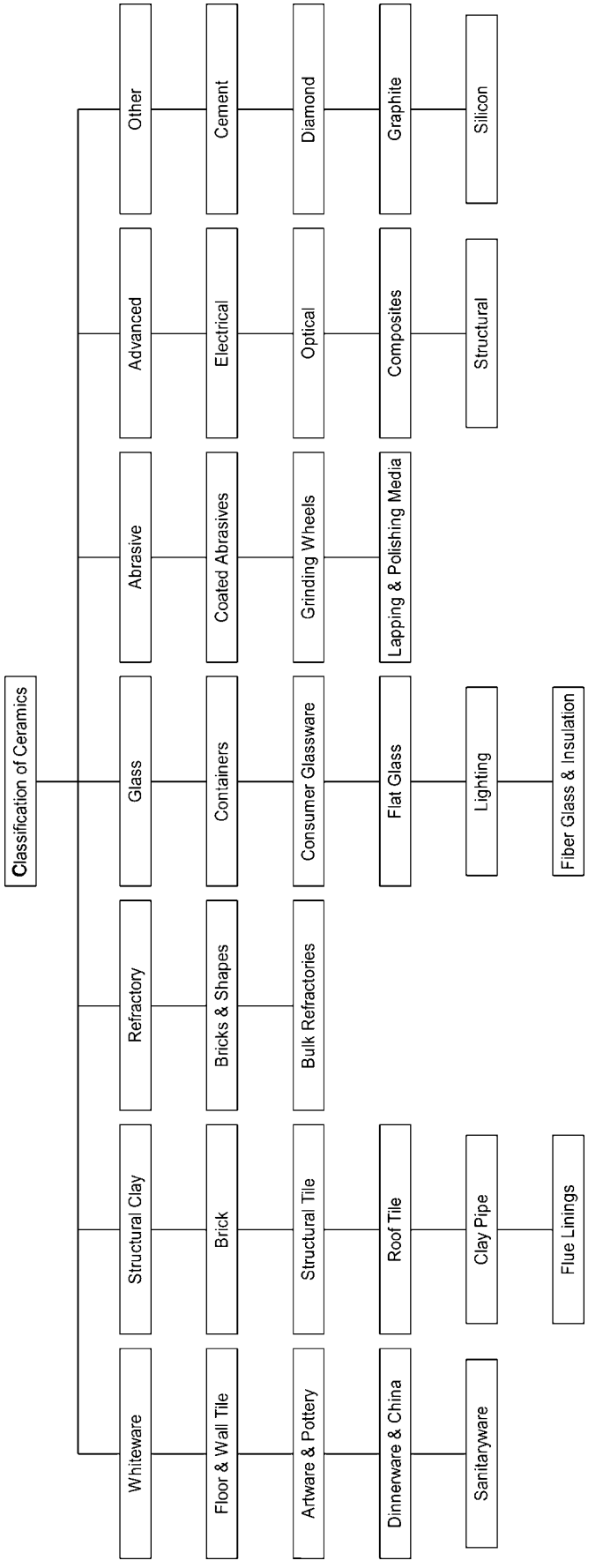


Fig. 1 Classification of ceramics by application.

Table 1 Different classification schemes for ceramics

Application
Whiteware
Structural clay
Refractory
Glass
Abrasive
Advanced
Wear
Membrane
Electronic packaging
Composition
Oxide
Nonoxide
Silicate
Ferrite
Carbide, nitride, boride
Crystal structure
Clay
Perovskite
Spinel
Amorphous
Graphitic
Forming method
Tape
Hot pressed
Slip cast
Extrudate
Coating
Film
Laminate
Microstructure
Single crystal
Polycrystalline
Monolithic
Composite
Foam
Glassy
Synthesis method
Solid-state
Sol–gel
Reaction bonded
Chemical vapor deposition

superconductors, sensors, microelectromechanical devices (MEMS), nanomaterials, photonics, electro-optics, and fuel cells.

RAW MATERIALS FOR CERAMICS

Ceramic materials of the clay family occur widely in nature, and the many different forms of clay differ in both their composition and crystal structure. In general, the structure of clay is noted for its layered arrangement of aluminosilicate sheets. Upon addition of water, the clay–water mixture responds plastically, which aids in the ability to form complex shapes. Two examples of clay compositions are kaolinite [$Al_2(Si_2O_5)(OH)_4$] and montmorillonite [$(Al_{2-y}X_y)(Si_2O_5)_2(OH)_2$, where X is Na or Mg and y is 0.33].

Nonclay minerals also widely used as ceramic raw materials are silica (SiO_2, obtained from quartzite, sandstone, and sand), lime (CaO, derived from limestone, $CaCO_3$), talc ($Mg_3Si_4O_{10}(OH)_2$), and feldspar (aluminosilicates containing oxides of sodium, potassium, and calcium). Other ceramic materials are derived from naturally occurring minerals. Alumina (Al_2O_3) is obtained from the mineral bauxite, which is a mixture of gibbsite [$Al(OH)_3$] and diaspore ($HAlO_2$). Zirconia (ZrO_2) is found as the mineral baddeleyite or is derived from zircon ($ZrSiO_4$). Titania (TiO_2) is found as the minerals rutile and anatase. Many of the ceramic raw materials found in nature may need to undergo powder beneficiation prior to being further processed for specific applications.

Some ceramic materials are not found widely or at all in nature, and thus are synthesized for use. To prepare more complex ceramic compositions such as perovskites of general structural formula ABO_3, and ferrites, of formula MFe_2O_4, the individual oxides or salts of the cations A, B, and M are often combined as powders and then reacted at high temperature by a solid-state diffusion mechanism. Silicon nitride (Si_3N_4) can be manufactured from either the nitridation of silicon metal or from the reaction of silicon tetrachloride with ammonia. Silicon carbide (SiC) is obtained from the reduction of silica with a carbon containing source.

Other synthesis methods have also been developed but may not be practiced on a large commercial scale as compared to those indicated above. Examples of syntheses that begin from liquid phase precursors include the sol–gel, hydrothermal, and Pechini methods. In the course of these reaction schemes, polycondensation or precipitation occurs, and the volatile components are removed, often by thermal methods. Other synthesis routes to ceramic materials include the nitridation of metals to form metal nitrides and the carbothermal reduction of oxides to form carbides and borides.

PROPERTIES AND APPLICATIONS

Ceramic materials have a number of properties that make them useful in a wide variety of applications. Although from a fundamental viewpoint material properties can be classified as equilibrium (e.g., stress–strain), transport (e.g., thermal conductivity), hysteretic (e.g., ferromagnetization), and irreversible (e.g., hardness), a more conventional grouping is used here. Some of the main physical, chemical, mechanical,

thermal, electrical, magnetic, and optical properties of ceramics are listed in Table 2.

The properties of ceramics are related to the type of bonding in the material and whether the material is amorphous or crystalline. Glasses are a common example of amorphous materials in which the atoms comprising the material do not exhibit a highly regular long-range arrangement of atoms in space.

The crystal structure of ceramics, i.e., the periodic arrangement of atoms in space, ranges from simple to complex. The complexity arises from the occurrence of both covalent and ionic bonding and from the diverse compositions of ceramics, especially when multiple cations or anions are present. In order of decreasing symmetry, seven crystal systems exist (cubic, tetragonal, orthorhombic, monoclinic, rhombohedral, hexagonal, and triclinic); ceramics are found in all these crystalline forms. The crystal structure of a ceramic material can be the determining factor whether certain mechanical, electrical, and magnetic properties exist.

Table 2 Selected properties of interest for ceramic materials and applications

Mechanical properties
Density
Elastic modulus
Bulk modulus
Poisson's ratio
Tensile strength
Compressive strength
Modulus of rupture
Fracture toughness
Hardness
Fatigue
Creep
Electrical properties
Resistivity
Conductivity
Dielectric constant
Loss factor
Breakdown strength
Electromechanical coupling constant
Magnetic properties
Permeability or susceptibility
Optical properties
Refractive index
Dispersion
Reflection
Refraction
Absorption
Transmission
Color
Thermal properties
Thermal expansion
Conductivity
Heat capacity
Emissivity
Thermal shock resistance
Chemical properties
Corrosion resistance
Adsorption
Catalytic

Physical Properties

One primary physical property is density, and as a class of materials, the density of ceramics (2–7 g/cm^3) is intermediate between the densities of polymers (0.4–2 g/cm^3) and metals (>7 g/cm^3, with the exception of alloys of aluminum, magnesium, and titanium). Thus, ceramics are attractive materials when compared to many metals, when a reduction in weight per unit volume is desired.

Another physical property is melting point. Many ceramics have melting points above 500°C; many oxide, carbide, and nitride ceramics, however, have melting points (or decompose) near or well above 1500°C. These values far exceed those of polymers, which generally melt or begin to decompose below 400°C. Metals, on the other hand, often possess melting points between 500 and 1500°C; exceptions to this are refractory metals such as tungsten and molybdenum, which melt above 2500°C, but these two materials are restricted to use at elevated temperature in high vacuum or reducing environments. As a consequence of their high melting points, ceramics are often used in high-temperature applications. Some ceramics, especially the nonoxides, do not exhibit a true melting point but rather decompose or dissociate at elevated temperature.

Chemical Properties

One important chemical property of ceramics is resistance to chemical corrosion at both room and elevated temperatures. In fact, the inertness of some ceramic materials, combined with high melting point and the retention of at least modest mechanical strength, allows them to be widely used in the high-temperature processing of other materials such as molten metals and glasses. Quartz, for example, is the crucible material in the Czochralski process for growing single crystal silicon. Refractory bricks, which are used in furnace construction, glass melting, and steel making, are made from compositions mainly consisting of alumina, silica, and magnesia (MgO). Clay-based filters are employed in the filtering of molten metals before they are cast into molds. Silicon carbide, quartz, and other ceramic materials are holders for silicon wafers in the fabrication of integrated circuits.

Some ceramics exhibit biocompatibility in the human body. Alumina and zirconia are employed as the ball for hip replacements. Hydroxyapatite ($Ca_{10}(PO_4)_6(OH)_2$) is used as bone replacements, as ocular implants, and as a coating for metallic implants. Ceramics also find application in dentistry for restorative work.

The adsorption properties of ceramics for liquids and gases also make them widely used for separations based on chemical affinity; silica, alumina, and zeolites (aluminosilicates with highly regular microporous intracrystalline pores) are widely used as adsorbents in the chemical process industries. Inorganic membranes based on silica, alumina, titania, and zeolites are also emerging; in these applications, their surface properties and highly tailored pore size are used to effect gas and liquid separations. Ceramic membranes, as compared to polymeric membranes, can also operate in aggressive (highly basic or acidic) chemical environments and at high temperature.

Mechanical Properties

The mechanical properties of interest for ceramics include tensile strength, strain at failure, compressive strength, modulus of elasticity, hardness, and fracture toughness. Of all materials, ceramics have some of the highest values of compressive strength, modulus of elasticity, and hardness. These attributes, alone or in combination, often make ceramics suitable for applications wherever wear resistance is important. Silicon nitride, for example, is used to make ceramic ball bearings. Alumina is used for femoral hip replacements. Silicon carbide is used as water pump seals in automobile engines and alumina is used as seals in water faucets. Abrasive grains of alumina, silicon carbide, boron carbide (B_4C), silica, cubic boron nitride (BN), and diamond (C) are used in grinding, lapping, and polishing operations. Silicon nitride is used in high performance cutting tools and is being considered for use as engine valves in diesel and internal combustion engines. Concrete, which contains cement, is used as a durable and inexpensive road and sidewalk material.

For tensile strength and fracture toughness, lower values, as compared to advanced metal alloys are often realized for ceramics. This arises, to a certain extent, from flaws that arise in the material during processing. To extend the range of application of ceramic materials, much current process development, compositional development, and microstructure engineering are aimed at improving the tensile strength and fracture toughness.

Many applications of ceramics rely on their mechanical properties at high temperature. This advantage of ceramics is realized because of their high melting points and because of the inertness of the materials to chemical reaction, which can include resistance to oxidation, reduction, and chemical corrosion. In combination, these properties often lead to superior creep and fatigue resistance of ceramic materials. Thus, many ceramics find their application as high-temperature barrier materials in the form of tiles, bricks, crucibles, and supports for silicon wafers in the semiconductor industry. For silicon nitride, the retention of high-temperature mechanical properties in extreme environments has led to the use of this material in gas turbines and automobile turborotors.

Thermal Properties

The thermal properties of interest for ceramics are thermal expansion, thermal conductivity, specific heat, and emissivity. The thermal expansion of ceramics tends to be lower than that of metals and this has both positive and negative consequences. Because of the low thermal expansion coefficient of some ceramics, they tend to withstand thermal shock, and thus can be subjected to temperature cycling. This same low thermal expansion, however, leads to strain mismatch when ceramic components, such as turborotors, are joined to metallic parts, such as the turborotor shaft.

The thermal conductivity of ceramics can span a range of values and, for a given material, can also be adjusted by addition of second phases, including air (pores), dopants, and by grain boundary engineering. To lower the thermal conductivity of ceramics, porosity is often designed into the microstructure of the materials. This can be accomplished by incomplete sintering, by preparing ceramic foams, and by preparing fibrous materials. The low thermal conductivity of these materials, along with high melting points, makes them extremely useful for thermal insulation as tiles, heat shields, and as fibrous insulation. Fiberglass insulation is one such example; a second is the tile for the space shuttle, which is mainly fibrous silica containing approximately 90% porosity. In addition to dissipating heat, the shuttle tiles can also withstand the extreme thermal shock associated with exposure to shaded and sunlit regions of outer space.

Electrical Properties

The electrical properties of interest for ceramics include conductivity, resistivity, dielectric breakdown strength, dielectric constant, loss factor, and electromechanical coupling. Most ceramics do not have high electrical conductivity, and thus ceramics have found application as electrical insulators for many years. The electrical insulating capability of some ceramics is also retained under high electric field; this is referred to as high dielectric breakdown strength

and is important in electricity transmission and in other high voltage applications.

Although the vast majority of ceramic materials are electrical insulators because of their high band gap, some ceramics, such as silicon carbide, strontium titanate, and silicon, are semiconducting. The semiconducting behavior of these materials is often modified by the incorporation of impurities or dopants into the material: n-type dopants lead to semiconduction whereby electrons (negative particles) are the predominant charge carriers. Conversely, p-type dopants lead to semiconduction whereby holes (positive particles) are the predominant charge carries. Doping is widely used to tailor the semiconducting properties of silicon in integrated circuits.

Another type of electrical conductivity observed in ceramics is ionic conductivity, which often occurs appreciably at elevated temperature; a widely used material exhibiting this behavior is zirconia doped with other oxides such as calcia (CaO) or yttria (Y_2O_3). For this material, atomic oxygen is the mobile ionic species. Doped zirconia finds widespread use as oxygen sensors, especially as part of automobile emission control systems, where the oxygen content of the exhaust gas is monitored to control the air-to-fuel ratio. Other applications of ionic conducting ceramics are as the electrolyte phases in solid-oxide fuel cells and in sodium–sulfur batteries.

Ceramic compositions also exhibit superconductivity, and what is notable is that they become superconducting at much higher temperatures as compared to other metallic and intermetallic compositions. The oxide composition $YBa_2Cu_3O_7$ is superconducting near 90 K, which is above the temperature of liquid nitrogen. Other ceramic oxide compositions containing bismuth, strontium, barium, calcium, and copper are superconducting at temperatures between 100 and 150 K.

Ceramics can possess a wide range of dielectric constants from 2 to 10,000, and thus find use as capacitors for charge storage and electronic filtering. Ceramic dielectrics include porcelain materials and many titania- or titanate-based compositions. In electrical applications involving alternating current, the dielectric loss factor is important for determining how efficiently electrical energy is utilized.

Selected classes of asymmetric crystal structures exhibit the property of piezoelectricity. With the application of a mechanical strain, piezoelectric materials develop an electrical potential difference across them; conversely, when a potential difference is applied to these materials, a displacement occurs. The efficiency of the conversion between mechanical energy and electrical energy is described by the electromechanical coupling constant, which practically ranges to values as high as 0.7; a value of 1 would imply complete conversion between mechanical and electrical energy. The piezoelectric behavior of ceramics is used widely. The conversion of electrical energy to displacement underlies the applications of ultrasonic mixing, sound wave generation from ultrasonic transducers, and precision-controlled positioning devices. The conversion of strain to an electrical signal is the basis of operation for ultrasonic sensors, accelerometers, strain gauges, and sound generation from phonograph recordings. Ferroelectric materials that find widespread use in these applications are barium titanate ($BaTiO_3$), lead titanate ($PbTiO_3$), and lead zirconate ($PbZrO_3$). Because the response of piezoelectric devices is very sensitive to aspects of ceramic processing and microstructure, additives are often used with the primary compositions to fine-tune the performance. Such fine-tuning may lead to improved coupling constants and better mechanical, temperature, and frequency performance.

Magnetic Properties

Ceramic materials also exhibit magnetic behavior, which is described in terms of the magnetic permeability or the magnetic susceptibility. Magnetic ceramics find application as permanent magnets, as circuit elements in electronic devices, and as digital information storage media in the form of coatings on recording tape and on disk substrates. Many magnetic ceramics are of the ferrite composition and have the general chemical formula of MFe_2O_4, where M is a divalent cation such as iron, nickel, copper, manganese, or magnesium. More than one of the divalent cations is often included in ferrite compositions to fine-tune the magnetic properties. Other ceramics that exhibit magnetic behavior include the rare earth garnets, orthoferrites, ilmenites, and hexagonal ferrites.

Optical Properties

The optical properties of ceramics are useful in the ultraviolet, visible, and infrared ranges of the electromagnetic spectrum, and one key quantity used to describe the optical property of a material is the refractive index, which is a function of the frequency of the electromagnetic radiation. Other quantities used to characterize optical performance are absorption, transmission, and reflection; these three properties sum to unity and are also frequency dependent. The last three properties govern many aspects of how light interacts with materials in windows, lenses, mirrors, and filters. In many consumer, decorative, and ornamental applications, the esthetic qualities of the ceramic, such as color, surface texture, gloss, opacity, and translucency, depend critically on how light interacts with the material.

In addition to the well-known optical uses listed above, the behavior of light with ceramic materials

is used to advantage in a number of advanced applications. Fiber optic glass cables exhibit nearly 100% total internal reflectance and are used to transmit both light images and digital information. Glass beads can exhibit directional reflectance and are used as additives in paint to improve the light reflectivity of signage and pavement lines. Solid-state lasers from ruby and yttrium aluminum garnet (YAG) are also widely fabricated from ceramic materials. Windows for gas lasers are often fabricated from ceramics materials; the ceramic used depends on its ability to transmit light at the frequency of interest.

PROCESSING

The majority of ceramic components are derived from powder processing routes, which are in contrast to the casting or molding of metals, polymers, and glass. The difficulty in casting ceramic shapes arises because of the high melting points of ceramic materials. In addition, because of their brittleness, it is not possible to stamp or forge ceramic shapes either. For applications requiring thin cross-sectional area, ceramics are prepared as coatings and as thin films.

Powder Processing

A generic processing route to fabricate a ceramic component from a powder raw material is contained in Fig. 2, and the individual processing steps are described in more detail below. Myriad variations of this processing scheme are possible, but in general, the manufacture of any ceramic component by a powder processing route will contain a number of steps of the type and in the order indicated below.

Powder beneficiation

As received, ceramic powders often need to be modified before they can be used to manufacture a component. Such powder beneficiation includes changing the particle size, shape, surface area, particle size distribution, and purity. Reduction of the particle size is generally accomplished by milling (particle size comminution). Although ball milling has been widely used to modify the particle size, shape, and size distribution, higher energy methods, such as jet milling, attrition milling, and vibratory milling, are also available and can lead to finer particle sizes. Jet milling is accomplished by impinging a high velocity stream of particles onto itself; the energy of the impacts causes particle fracture and hence size reduction. Attrition and vibratory milling often use relatively small milling media, as compared to ball milling, but rely on increased frequency of impacts between the milling media and the powder particles for size reduction.

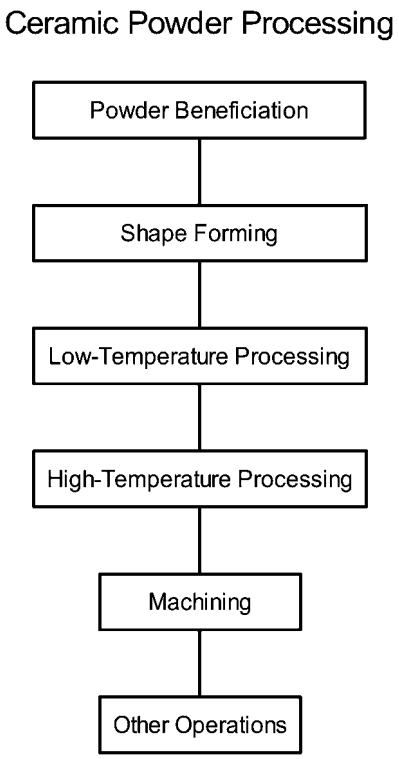


Fig. 2 Example of a process flow diagram for preparing a ceramic component by a powder processing route.

As part of powder beneficiation, compositional modification may also be performed by the addition of second phases, dopants, etc. These changes may be required to influence the sintering behavior of the ceramic powders or to tailor the structural, thermal, electrical, and optical properties of the ceramic products. The source of these additives may be from powders such as oxides, glasses, or salts. To remove unwanted impurities, second phases, surface contaminants, or other by-products, powders may beneficiated by washing, sedimentation, filtration, and magnetic separation.

In the processing of some ceramic powders, enlargement of the size of the solid particulates is required. Granulation and spray drying are two process operations that convert small particles that do not readily flow, into flowable solids; this facilitates the forming of shapes. Granulation is accomplished by agitating solid particulates against each other in drums or by forcing particles between plates, blades, narrow gaps, or rollers. Spray drying is performed by atomizing a slurry that contains ceramic particles, a carrier fluid—often water or a solvent—and organic phases such as dispersants and binders. As the slurry is atomized from a nozzle into a heated chamber, droplets form and dry into spherically shaped solid particulates.

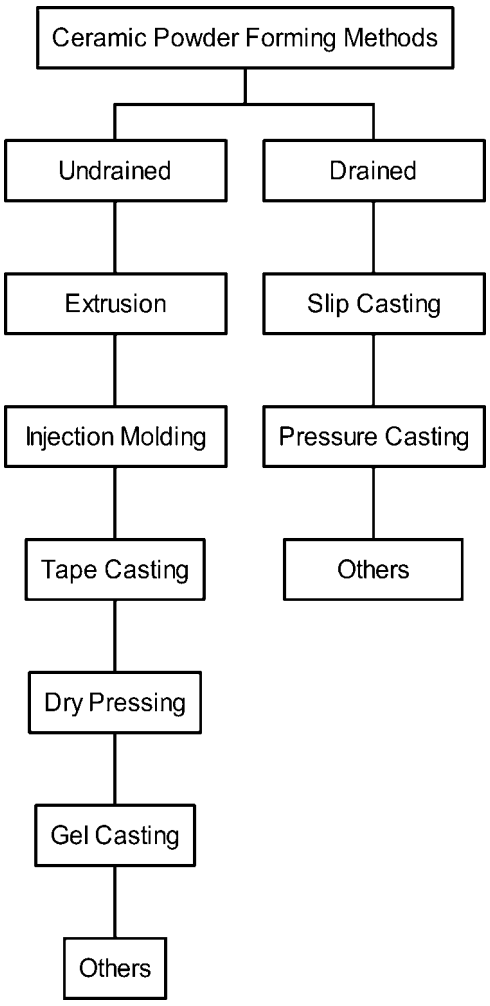


Fig. 3 Classification of powder-based forming methods used to make ceramic shapes.

Shape forming

Many shape forming techniques based on powder processing exist for ceramics, and some of the main types are listed in Fig. 3. These fabrication techniques can be grouped into drained and undrained methods; these techniques differ in the degree of dimensional tolerance that can be achieved. In general, net-shape forming methods are preferred, but in the fabrication of ceramics, only near-net shape is often realized.

In drained forming, the carrier fluid that is used to facilitate the motion of the ceramic particles is removed as the shape is formed. Slip casting (see Fig. 4) and pressure casting are examples of drained methods. Because in drained forming the carrier fluid is removed as the particles pack into a shape, the green density of the formed components is higher than the density of particles in solution, and in fact, drained forming can lead to very high green densities, as compared to undrained methods.

Undrained forming methods are those in which the carrier fluid used to facilitate the motion of particles is not removed as the shape is made. The carrier fluid thus remains in the formed shape and must be removed prior to high-temperature processing of the component. Extrusion, injection molding, tape casting, and gel casting are all undrained methods. For extrusion and injection molding (see Fig. 5), the carrier fluids are often mixtures of organic species, which collectively is referred to as binder. For tape casting (see Fig. 6), the carrier fluid is predominantly either solvents or water; binders are again used to impart plasticity to the tapes. For gel casting, the carrier fluids contain monomeric species and initiators that cause polymerization directly in the mold. Because in undrained forming methods the carrier fluid remains in the mold, the green density, i.e., the bulk density of the particulate body, is often close to the density of the particles in the carrier fluid.

Another widely used method for forming simple ceramic shapes is dry pressing (see Fig. 7). If for dry pressing, the carrier fluid is viewed as the particle phase itself which flows under the influence of gravity, then dry pressing can be included as an undrained forming method. The other fluid phase present during dry pressing is air, a compressible fluid, which neither imparts much flow to the particles, nor does it impede particle packing in the mold.

Low-temperature processing

After the powder is formed into a shape, the resulting component, termed the green body, is subjected to processing operations at low temperature to remove the carrier fluid, which may consist of one or more of solvents, water, dispersants, and binders, depending on which forming method is used. Solvents and water removal may be accomplished initially at temperatures below 100°C whereas binders are removed at temperatures of 100–800°C. Although most binders are mainly decomposed below 500°C, higher temperatures are sometimes used to remove any trace carbon from the material. For bodies moderately to highly loaded with organic phases, such as those obtained by injection molding, extrusion, and tape casting, the removal of binder can be a very slow process and can take days to avoid forming defects in the body.

After drying and binder removal, the powder particles are often not strongly adherent to each other, and because of this, the body is sometimes presintered, or bisque fired, to improve the mechanical strength for handling. Presintering may be performed at temperatures of 800–1500°C, depending on the ceramic. Because the low-temperature furnace operations span a range of temperatures, each step may be performed in separate ovens or furnaces or the steps may be combined into one or more heating units, where the

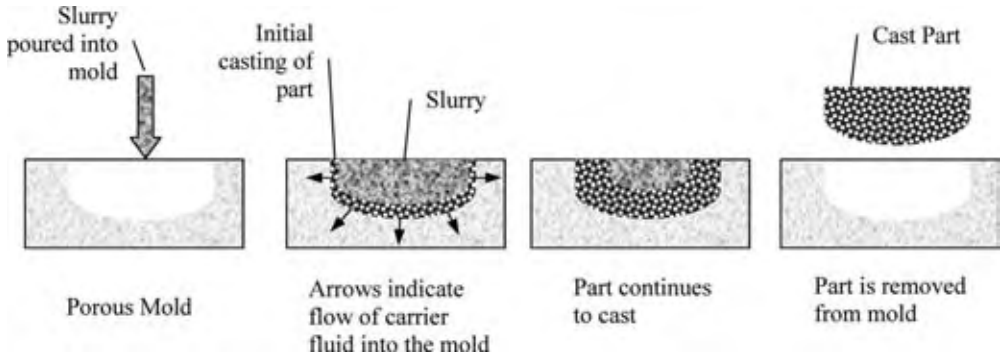


Fig. 4 Schematic of the slip casting process for a ceramic shape. A slurry of ceramic particles and a carrier fluid is poured into a porous mold. The carrier fluid flows into the pores of the mold, and the ceramic particles remain behind and build up a solid cake in the mold. After the fluid has drained from the part, the cast shape can be removed from the mold.

time–temperature profile is tailored to meet the different processing objectives.

High-temperature processing

After the porous green body has been dried and the organic phases have been removed, the component is sintered at elevated temperature below the melting point of the material. This process step is generally one of the most important, whereby the microstructure of the material is tailored to meet specific, end-use properties.

During sintering, the amount of porosity and the surface area decrease, and the pore size, grain size, and density all increase. The sintering process is generally divided into three stages: initial, intermediate, and final. During the initial stage of sintering, necks or bridges form between particles, and a small amount of particle rearrangement takes place. Generally, the amount of shrinkage is small during this stage of sintering. The most dramatic changes in the microstructure occur during the intermediate stage of sintering. A substantial reduction in the total porosity, surface area, and number of pores takes place; the pore size, however, increases. In the final stage of sintering, the most important processes are the final elimination of porosity and the growth of grains. A scanning electron micrograph of a sintered, nearly dense microstructure is shown in Fig. 8. The elimination of porosity that occurs during sintering may be performed by different transport mechanisms including vapor phase transport, solid-state diffusion, and liquid phase dissolution and reprecipitation.

After sintering, a final process step at elevated temperature may be performed to impart specific properties to the component. This may include fine-tuning of the crystallinity, porosity, grain size, stress, and electrical and magnetic properties of the material.

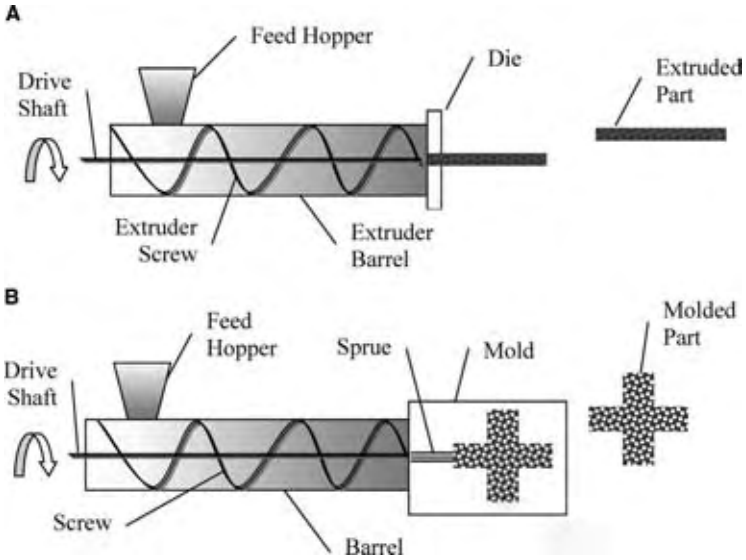


Fig. 5 (A) Schematic of screw extrusion of a ceramic shape. The ceramic feed material dispersed in a binder flows from the hopper into the barrel where the rotating screw transports the material through the die opening. (B) Schematic of injection molding of a ceramic shape. The ceramic feed material dispersed in a binder flows from the hopper into the barrel where the rotating screw transports the material through the sprue into a closed mold. After the mold is opened, the part is ejected.

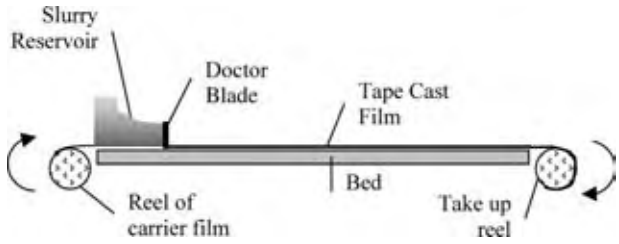


Fig. 6 Schematic of a tape casting process used to form thin sheets of ceramic material. The slurry of ceramic particles and carrier fluid flows beneath the doctor blade onto a moving carrier film. As the tape moves down the bed, it dries and is then spooled on a take up reel.

Machining

After the processing operations at high temperature, machining of the component may be performed to meet the final product specifications for shape, dimensions, and surface finish. For ceramics, this requires expensive diamond tooling such as slicing and grinding wheels and lapping and polishing media. Alternative methods to machining with diamond tooling are electric discharge, ultrasonic, laser, and abrasive flow machining.

Machining may also be performed at other points in the process flow schematic in Fig. 2, such as after the drying step or after presintering. These are the likely process points because the green bodies may then have sufficient strength to withstand the contact stresses of the tooling with the component.

Other operations

Before the manufacturing process for a ceramic is complete, other processing operations may be performed which are specific to an application. These may include applying a coating, metallization, characterization, proof testing, inspection, and packaging.

Glass Processing

Because glasses can be melted at relatively low temperatures, alternative methods to powder-based forming are primarily used to shape glass components. One technique is pressing, in which force is applied to a viscous gob of glass so that it adopts the shape of a mold. Pressing is used to form glass components with thick cross sections such as dishes, lenses, and plates and is also used to fabricate preforms for later processing operations such as blowing. Glass blowing is used to manufacture thin-walled shapes such as light bulbs, glass containers, and art objects. Drawing is used to form long highly symmetric shapes such as sheets, rods, and tubes. Fiberglass reinforcing filaments, fiber optic cables, and thermal insulation are all made by drawing operations. For very thin fibers used in mat or wool form, as is found in thermal insulation, fibers are produced in large quantities by a spinning operation, in which molten glass flows through a multiholed device called a spinneret. It can be mentioned here that there is a special class of ceramics referred to as glass ceramics. Components from these materials are formed from the melt obtained by glass processing methods. They then undergo crystallization. The advantage of this processing route is that the large shrinkage that occurs during sintering is avoided.

Coatings and Thin Film Processing

Coatings or thin films of ceramics can be prepared on a wide variety of substrates for a wide range of applications. The benefits derived from coatings can be

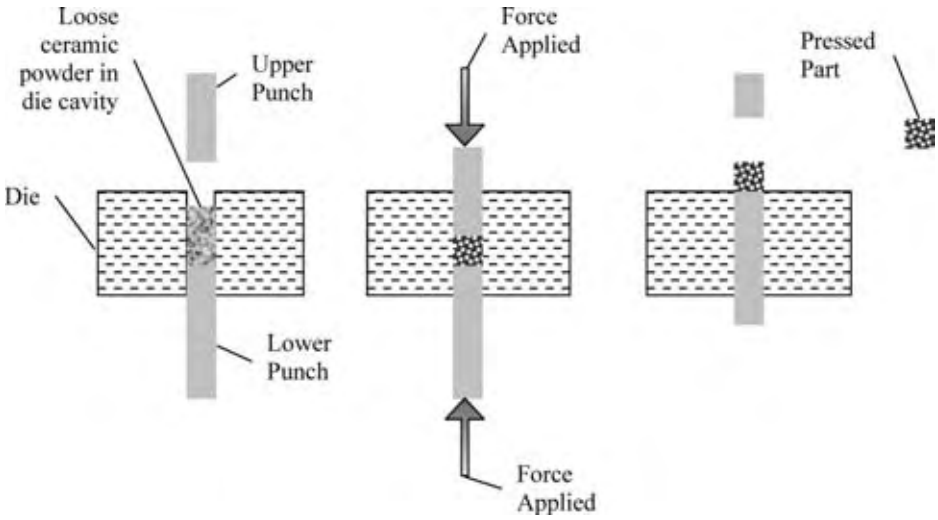


Fig. 7 Schematic of dry pressing of a ceramic shape. The ceramic particle feed flows into the die cavity. Force is applied to one or both punches and the powder is pressed into shape. The part is then ejected from the die cavity.

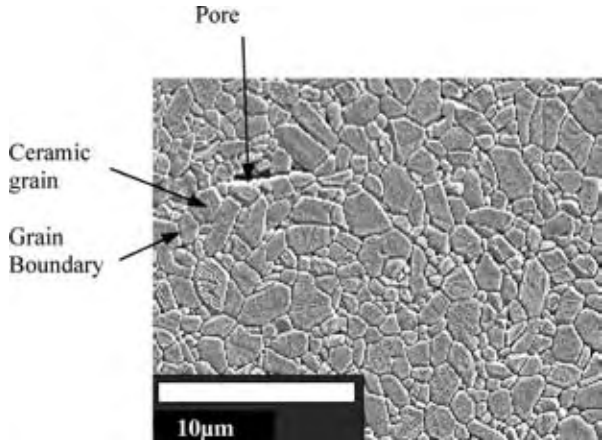


Fig. 8 The microstructure of alumina sintered at 1450°C for 3 hr. The micrograph was taken by a scanning electron microscope. Grains, grain boundaries, and pores are evident in the nearly dense microstructure. The magnification bar is 10 μm, or about one-tenth the width of a human hair.

improved wear-, thermal-, and corrosion resistance. Thin films are also widely used in the fabrication of integrated circuits in which different regions of the substrate have different compositions with highly tailored electrical properties.

Coating and thin films can be applied by a number of methods. In thermal or plasma spraying, a ceramic feedstock, either a powder or a rod, is fed to a gun from which it is sprayed onto a substrate. For the process of physical vapor deposition (PVD), which is conducted inside an enclosed chamber, a condensed phase is introduced into the gas phase by either evaporation or by sputtering. It then deposits by condensation or reaction onto a substrate. A plasma environment is sometimes used in conjunction with PVD to accelerate the deposition process or to improve the properties of the film. For coatings or films made by chemical vapor deposition (CVD), gas phase chemicals in an appropriate ratio inside a chamber are exposed to a solid surface at high temperature; when the gaseous species strike the hot surface, they react to form the desired ceramic material. CVD-type reactions are also used to infiltrate porous substrates [chemical vapor infiltration (CVI)]. For some applications, the CVD reactions take place in a plasma environment to improve the deposition rate or the film properties.

CONCLUSIONS

Ceramic materials have played a key role in the economic, societal, cultural, and technological development of mankind and will continue to do so. For many commodity and advanced applications, ceramic materials, as compared to metals and polymers, are the only suitable options either as technology enablers or in end-use products. For many of the most demanding advanced applications relying on electrical, magnetic, optical, wear, hardness, strength, and thermal properties, new or improved ceramic materials are the likely candidates. This will require continued investment in the research and development of new ceramic materials, new processing methods, and new characterization tools. In addition to the key role ceramics play in the world today, they will continue to do so in the future in science, medicine, communication, transportation, core industrial infrastructure, environmental remediation, energy efficiency, and in the chemical process industries.

REFERENCES

1. Kingery, W.D.; Bowen, H.K.; Uhlmann, D.R. *Introduction to Ceramics*, 2nd Ed.; Wiley Interscience: New York, 1976.
2. Richerson, D.W. *Modern Ceramic Engineering*, 2nd Ed.; Marcel Dekker: New York, 1992.
3. Rahaman, M.N. *Ceramic Processing and Sintering*; Marcel Dekker: New York, 1995.
4. Chiang, Y.-M.; Birnie III, D.P.; Kingery, W.D. *Physical Ceramics: Principles for Ceramic Science and Engineering*; John Wiley & Sons: New York, 1997.
5. Schneider, S.J. Technical Chairman. *Engineered Materials Handbook, Vol. 4, Ceramics and Glasses*; ASM International: Materials Park, OH, 1991.
6. Reed, J.S. *Introduction to the Principles of Ceramic Processing*; John Wiley & Sons: New York, 1988.
7. Van Vlack, L.H. *Physical Ceramics for Engineers*; Addison-Wesley: Reading, MA, 1964.
8. Ring, T.A. *Fundamentals of Ceramic Powder Processing and Synthesis*; Academic Press: San Diego, 1996.
9. Callister, W.D. *Material Science and Engineering, An Introduction*, 6th Ed.; John Wiley & Sons: New York, 2003.
10. http://www.ceramics.org (accessed July 2005).
11. http://www.ceramics.com (accessed July 2005).
12. http://www.ceramics.nist.gov/ (accessed July 2005).
13. http://www.ikts.fraunhofer.de (accessed July 2005).
14. http://www.advancedceramics.org/ (accessed July 2005).

Chemical Mechanical Planarization in Integrated Circuit Manufacturing

C

John Zabasajja
MOS 12, Technology and Manufacturing, Motorola Semiconductor Products Sector (SPS), Chandler, Arizona, U.S.A.

INTRODUCTION

The use of semiconductors has become a common occurrence in everyday life. Desktop and laptop computers, automobiles, medical devices, cellular phones, and all entertainment electronic gadgetry including toys, game boys, and television sets, etc., use semiconductors. For all these applications, miniaturization and fast speed of the gadgetry have become a desirable marketing tool that device and electronic manufacturers utilize to lure customers. The manufacture of integrated circuits, which are the essential building blocks for semiconductors, has continued to grow in complexity. Miniaturization requires that millions of these integrated circuits be built on a single chip. The ability to deliver multiple functions to the customer with this single chip has resulted in increased sophistication in their production.

In integrated circuit manufacturing, there are multiple processing steps, from the fabrication of the silicon wafer starting with a cylindrical silicon ingot to packaging of the chip as an end product to deliver to device and electronic manufacturers. Chemical mechanical planarization (CMP) is one of the back-end processing steps used in multibillion dollar manufacturing complexes known as fabs to fabricate these integrated chips. The history of CMP has been well documented elsewhere.[1,2] Ever since the early 1980s with its discovery and application by International Business Machines (IBM) and other companies, such as Motorola, CMP has become an essential step in the manufacture of all integrated circuits in the semiconductor industry. As microprocessor, logic, and wireless device dimensions have continued to scale down to meet high-performance requirements, CMP has become the enabling technology for planarizing metal and dielectric layers. Planarization is needed to lay down multiple layers of metal to build the circuitry, with millions of chips being packed into smaller areas of the wafer. Table 1 shows the National Technology road map for semiconductors, indicating the increased interconnectivity and metal levels that are being incorporated into current and future semiconductor devices.[2]

As technology continues to evolve and as integrated circuits become more sophisticated, integrated device manufacturers such as Intel, Motorola, IBM, Micron, Texas Instruments, and others in Europe and Asia have continued to work in instituting improvements to reduce the cost of ownership of integrated process manufacturing technologies, while continuing to seek an edge in device fabricating methods.[3] Chemical mechanical planarization, as one of the critical process technologies in this fabrication, is considered to be an enabling technology in the manufacture of leading edge devices. In this article, fundamental aspects of CMP will be discussed. Applications of CMP in front-end manufacturing such as in building shallow trench isolation (STI) layers that are used in the front end to isolate devices on a silicon chip, and back-end manufacturing methods such as copper (Cu) CMP that are used to build the electrical interconnectivity of the chip to the outside world will be discussed.

THE CMP SYSTEM

Fig. 1 shows a schematic representation of a CMP system with a rotating polishing table, a wafer carrier where the pressure or down force is applied, a pad conditioner, a stacked polish pad, and a slurry delivery tube, which delivers slurry to the polishing table. The pad typically consists of a hard pad (made of a material such as polyurethane). Fig. 2 shows a scanning electron microscopy (SEM) of the surface of a used polyurethane (IC1000 pad from Rodel) polishing pad. The open voids in the subsurface are clearly shown in the SEM. The pad is a closed cell, void filled polyurethane foam. As can be seen, the top of the pad surface is rough (with voids in the subsurface in the range of 20–100 μm), compared with the scale of typical local wafer features (smallest features for current and future technologies are <0.1 μm).

Polishing pad structure and material pad properties are important determinants in the polish rate and planarization in a CMP process. The impact of pad properties on rate, uniformity, and planarization is still not well understood. However, pad properties that are

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008022

Table 1 National Technology road map for semiconductors

	Year				
Technology characteristics	**1998**	**2001**	**2004**	**2007**	**2010**
Minimum feature size (μm)	0.25	0.18	0.13	0.1	0.07
Maximum wafer diameter (mm)	200	300	300	400	400
Number of metal levels					
DRAM	2–3	3	3	3	3
Microprocessor	5	5–6	6	6–7	7–8
Interconnection metal	Al, Cu	Al, Cu	Cu, Al	Cu, Al	Cu, Al
Interlevel dielectric constant	<3	<2.5	<2	1–2	1–2

considered to be important determinants of polishing performance include specific gravity (a measure of porosity), hardness, and compressibility. It has been shown that harder pads with lower compressibility provide better planarization.[2]

As the pad is used, or as it ages, the pad starts losing its roughness and porosity, resulting in lower removal rates. More pad ''glazing'' occurs with a subsequent loss in surface roughness or porosity. To maintain the porosity, a process known as pad conditioning, which involves utilizing diamond conditioner grit moving across the diameter of the pad in a programmed sequence to provide maximum conditioning, is used. Fig. 3 shows the surface height probability density function (PDF) for an IC1000 pad.[4] It shows that above a certain height (mean surface) on the pad, the polishing behavior of the pad is influenced by conditioning and abrasive wear. With optimized pad conditioning, pad life can be extended by as much as 100 times in terms of the numbers of wafers polished on that pad.

As shown in Fig. 1, down-force or pressure is applied on the wafer through a wafer carrier. The wafer carrier plays an important role in polishing nonuniformity. Tool manufacturers have designed wafer carriers with flexible membranes in the center and retaining rings on the edges of the wafer.[5] Utilizing backside pressures to the membranes, differential pressures can be applied across the wafer starting from the edge to the center of the wafer. If faster edge removal is desired, for instance, to account for the higher deposition profile on the edge of the wafer, a higher pressure through multizone control will be applied at the edge of the wafer (on the retaining ring) vis-à-vis the center of the wafer (on the membrane) to provide the faster edge removal rate.[6] Also, it has been shown that matching the carrier and platen angular velocities while rotating is essential to achieving uniform removal rate across the wafer.

Fig. 4 shows a multistep chemical mechanical polishing (CMP) tool utilized in current production. The tool consists of three polishing platens with each platen having all the features that have been discussed previously. As processing is completed on the first polishing platen, the wafer can be transferred to the next polishing platen for the next stage of processing.

For example, for a typical metal stack such as 6 kA Cu/400 kA tantalum (Ta)/2.5 kA TEOS on top of a

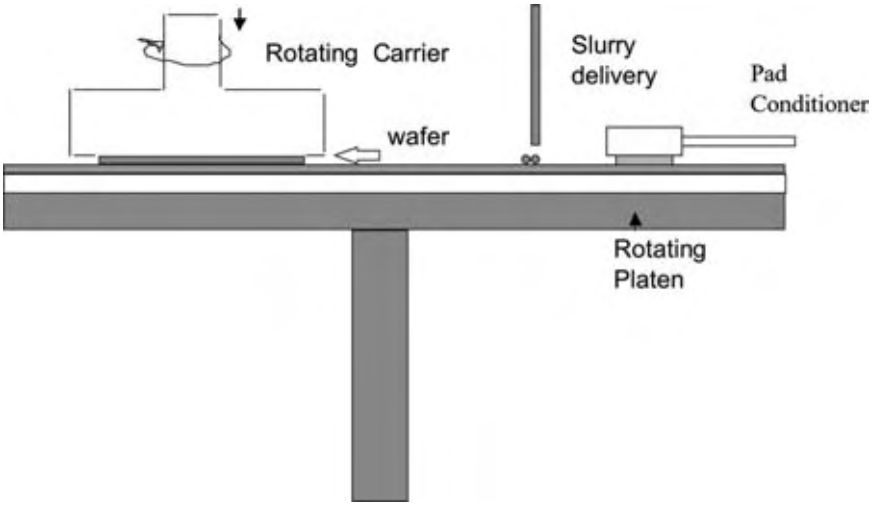


Fig. 1 Schematic of the CMP system. *(View this art in color at www.dekker.com.)*

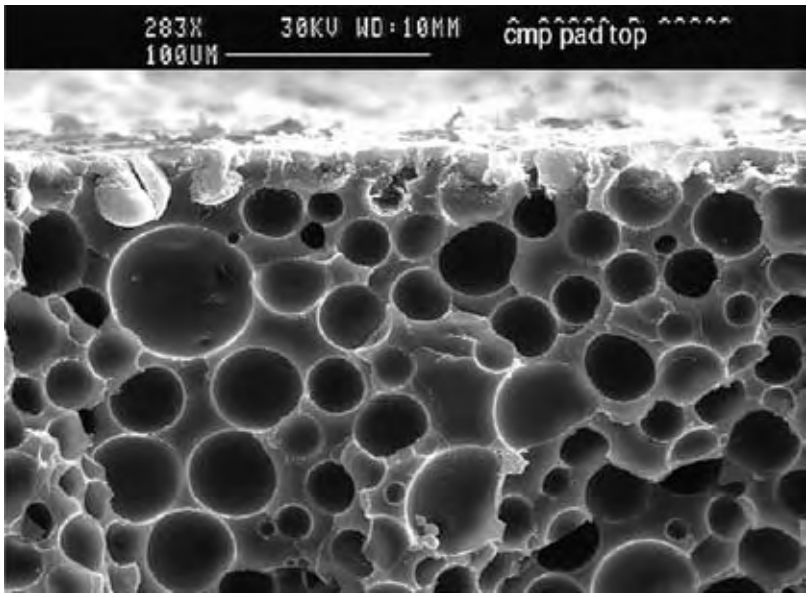


Fig. 2 Cross-section of a Rodel IC1000 polishing pad.

blanket silicon wafer, the bulk Cu will be polished on the first platen to remove most of the 6 kA thickness of Cu. The wafer will then be transferred to the second platen to polish the remaining Cu film, which has been reduced to <6 kA thickness and the process on this platen will then transit to an overpolished step after the end point has been triggered by the polishing tool with the detection of the 400 kA Ta barrier film. The wafer is transferred to a third platen to remove the remaining Ta film and also polish the oxide to buff the oxide layer. For selectivity purposes, the first two platens could use different slurry chemistries to polish the Cu, while the third platen will also use an entirely different slurry, also to polish the oxide film. The multistep tool provides the versatility to handle multiple slurries for different polishing steps and films. After the wafers are polished, the wafers are transferred to a cleaner, which serves the purpose of removing any remaining slurry particles. In the cleaner, removal of slurry postpolish and wafer drying is done. Slurry and other particulate defects are removed from the wafer surface through a combination of chemical and mechanical action.[7] For most scrubbers, a polyvinyl alcohol brush is used to provide mechanical action. Chemical is added through the brushes. The cleaning chemical is selected based on pH to give like charges to the particles, wafer surface, and brush. This selection prevents brush loading and defect redeposition

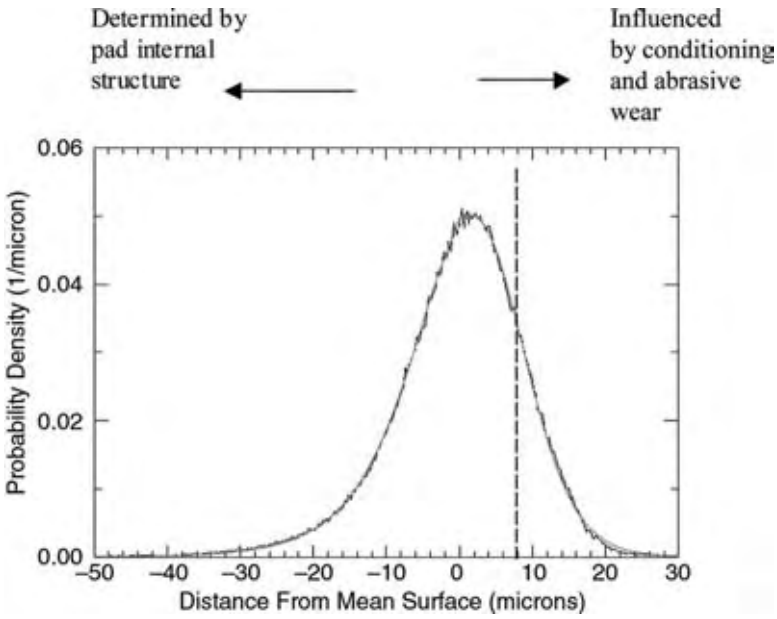


Fig. 3 Measured PDF for an IC1000 polishing pad. *(View this art in color at www.dekker.com.)*

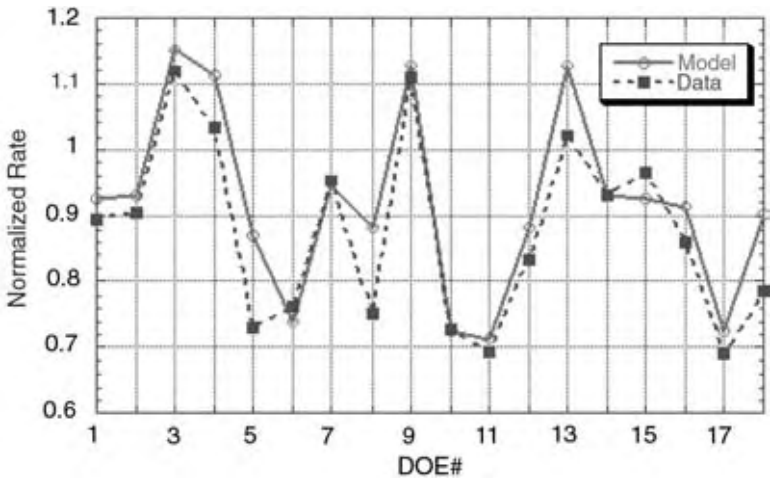


Fig. 4 Fit of Preston's model to experimental data for W CMP design of experiment. *(View this art in color at www.dekker.com.)*

on the wafer. A fundamental understanding of zeta potential effects is needed to enable cleaning chemistry to remove slurry particles from the wafer surface. For most applications, oxide CMP does not require any chemical to be applied in the scrubber to remove the slurry particles, while for metal applications, specifically tungsten (W) and Cu CMP, high- and low-pH chemistries such as NH_4OH and citric acid, respectively, are needed at different stages of cleaning, for effective removal of abrasive slurry particles.[2,8]

Most typical slurries consist of at least two components. The first component is the slurry abrasive, which typically provides the mechanical shear force that is needed to remove the oxide or metal films from the wafer. Most abrasive particles are in the range of 40–150 nm in size. The slurry will also have an oxidizer component such as H_2O_2 to provide the right chemical environment or chemical forces for the oxidation of the metal layer. For oxide polishing, mechanical forces predominate. Typical oxide slurries such as SC112 (KOH based slurry with silica abrasive from Cabot) do not have any oxidizer component. For metal polishing, it is theorized that the metal film has to go through a transformation of forming the oxide, which in most cases is fairly soluble and can easily be removed by mechanical action on the wafer surface.

PRESTON'S EQUATION

Chemical mechanical planarization is the selective planarization of a wafer by mechanical or chemical removal of material. The removal of material is governed by the Preston equation:[2,9]

$$\text{Removal rate} = \Delta H/\Delta t = KP^m V^n \tag{1}$$

where ΔH is the change in the thickness of the material that is being removed from the wafer, Δt is the elapsed time, K is Preston's coefficient, P is the pressure, and V is the relative velocity of the wafer carrier with respect to the polish pad. Both P and V could be a function of the position of the wafer to explain the variation of the polish rate across the wafer. K is dependent on process and consumable variables such as slurry concentration and pad properties, and m and n are constants. Typically, a value of 1 has been used for both m and n for oxide polishing; however, values ranging from $-4/3$ to $4/3$ have been reported in the literature based on slurry particle and pad properties.[2,9] Preston's equation has been found to be reasonably accurate for oxide, W and Cu CMP processes, although the dependence of K on slurry and pad properties is still not well understood and is still being investigated.[2] Both mechanical and chemical forces are important parameters and are expressed through the constant K. For example, in oxide polish, mechanical forces dominate and K can be expressed in terms of the mechanical properties of the film being polished.[2] K can be expressed in terms of Young's modulus:

$$K = (1/2)E \tag{2}$$

where E is Young's modulus of the film being polished. Thus, in cases where polishing is dominated by mechanical forces, Preston's equation is:

$$\text{Removal rate} = \Delta H/\Delta t = ((1/2)E)P^m V \tag{3}$$

Further attempts to elucidate the factors incorporated in this equation are given elsewhere.[2]

For metal polish, chemical effects can be explicitly delineated from Preston's coefficient Eq. (1):[9]

$$\text{Removal rate} = k^1 C_{ox} P^m V^n \tag{4}$$

where $K = k^1 C_{ox}$, and C_{ox} is the concentration of the oxidant in the slurry. All the other material properties are expressed through k^1. All pad and slurry properties

are expressed through the constant k^1. The chemical Preston's model [Eq. (4)] predicts that the polish rate will be directly proportional to the concentration of the slurry, and the oxidizer concentration (C_{ox}) in particular. Using experimental data at different pressure and velocity settings, best fit values for the parameters in Eq. (4) can be generated from a design of experiments (DOE).[4]

Table 2 shows a summary of DOE data for normalized values of P [represented by the down-force variable, and relative velocity V between the platen and carrier (represented by the variables, the platen speed, and the wafer carrier speed)], with the corresponding normalized removal rate for the W CMP process. Fig. 5 shows a good fit between Preston's model for the process and empirical data with normalized polish rate for each of the values shown in Table 2.

Also, the pressure P can be written as:

$$P = F/A \tag{5}$$

where F = force and A = area. Planarization is the process of removing up-areas at a faster rate than down-areas on a wafer.[2] From Eq. (5), since the pressure P is dependent on the force applied and the surface area of the features being polished on the wafer, dense areas and large features will polish slower than isolated or smaller features. Thus, removal rates and planarization are pattern dependent, as predicted by Preston's equation. We will now discuss the individual CMP processing steps that are typically used to build a device.

SHALLOW TRENCH ISOLATION CMP

For sub-0.5 μm devices, shallow trenches are used to isolate active areas in CMOS circuits. Fig. 6 is a simplified schematic showing a three-level metal device on top of a silicon wafer with the shallow trench in the front end near the devices, and the dielectric and metal layers in the back end to provide the interconnectivity. The trench elements are used to isolate the device elements. Due to the fact that these isolation areas are closest to the device areas and to the silicon substrate, care must be taken to ensure that all processes to form these elements have low defectivity and good selectivity to enable the process to stop at an oxide or nitride layer without processing through to the silicon substrate. Fig. 7 shows the processing steps in STI CMP process starting from photopatterning with photoresist to forming the device. Prior to STI CMP, etching through the nitride and oxide forms a trench using oxidizing gases after photopatterning. The trench is filled with SiO_2 using chemical vapor deposition. The interlevel dielectric (ILD), which in this case is SiO_2, is then planarized. It is critical to completely clear the SiO_2 and stop at the nitride layer during CMP. Thus, having a good consumable set of pads and slurries that will enable high polish rate of uniformity with good stop features at nitride, while exhibiting low defectivity, is desirable.

Typical slurries for polishing SiO_2 are KOH or NH_4OH based at high pH (10–12) with equal or slightly harder material such as silica. Careful particle size control of the slurry using filtration is required to minimize defects such as micro scratches and gouging.[10]

INTERLEVEL DIELECTRIC CMP

Dielectric CMP is used to planarize intermetal dielectrics (IMD) (represented by ILD1, ILD2, or ILD3) and premetal dielectrics (represented by ILD0). The slurry typically used in this process is at high pH (10–12), KOH or NH_4OH based and silica abrasive. The hydroxide ion is needed to hydrolyze the substrate, which in this case is SiO_2.

The planarization and uniformity resulting from ILD CMP improve the depth of focus for photolithography by reducing the variation at the top of the ILD or oxide layer and also enable reduction of the metal step coverage at that step and future steps. Fig. 8 shows the topography associated with an underlying structure, which contains a metal feature, prior to undergoing CMP processing. Fig. 9 shows the processing steps needed to planarize the feature and build the device and the interconnecting structure. Several ILD or IMD steps have to be conducted to provide the planarity to build the devices, the metal contacts, and interconnecting vias. As previously discussed, removal rate and planarization in ILD CMP are pattern-dependent and could be determined by consumable factors such the compressibility of the pad, how the pad is conditioned, and slurry properties such as the size of the abrasive particles. Further, these slurry properties could also play another important role in ILD CMP in determining the level of defectivity.

Both the defectivity and the removal rate can be determined during ILD polish by making measurements using stand-alone or integrated metrology tools.[11,12] For rate determination, thickness measurements can be determined by stand-alone metrology spectrophotometric tools or by tools that are fully integrated into the polisher for in situ measurements, which can provide a closed loop feedback system for measuring thickness while polishing.[13]

TUNGSTEN CMP

The first metal CMP application to be used in the industry was W CMP. Tungsten CMP slurries typically

Table 2 Normalized rate and uniformity data for W CMP design of experiment

Pattern	DOE	P Spd	C Spd	DF	BSA	OX	Abr	Norm rate	W uniformity	P/P Spd (RPM)	P/P C Spd (RPM)	P/P DF (PSI)	P/P BSA (PSI)	FE400	WA400	W rate
++−+−−	1	1	1	−1	1	−1	−1	0.8939	13	1	1	−1	1	−1	−1	3575.6
0	2	0	0	0	0	0	0	0.90333	12.9	0	0	0	0	0	0	3613.3
+−+−+−	3	1	−1	1	−1	1	−1	1.11843	4.12	1	−1	1	−1	1	−1	4473.7
+−++−+	4	1	−1	1	1	−1	1	1.03333	3.33	1	−1	1	1	−1	1	4133.3
−++−+	5	−1	1	1	−1	−1	1	0.7294	10	−1	1	1	−1	−1	1	2917.6
−+−−+−	6	−1	1	−1	−1	1	−1	0.76128	7.16	−1	1	−1	−1	1	−1	3045.1
+−−++−	7	1	−1	−1	1	1	−1	0.952	13	1	−1	−1	1	1	−1	3808
−−++−−	8	−1	−1	1	1	−1	−1	0.75305	9.49	−1	−1	1	1	−1	−1	3012.2
+++++	9	1	1	1	1	1	1	1.10995	10.2	1	1	1	1	1	1	4439.8
−−−+++	10	−1	−1	−1	1	1	1	0.72848	7.42	−1	−1	−1	1	1	1	2913.9
−+−+−+	11	−1	1	−1	1	−1	1	0.69365	12.6	−1	1	−1	1	−1	1	2774.6
−−+−++	12	−1	−1	1	−1	1	1	0.83233	7.51	−1	−1	1	−1	1	1	3329.3
+++−−−	13	1	1	1	−1	−1	−1	1.02025	8.99	1	1	1	−1	−1	−1	4081
0	14	0	0	0	0	0	0	0.93063	8.01	0	0	0	0	0	0	3722.5
++−−++	15	1	1	−1	−1	1	1	0.9647	8.29	1	1	−1	−1	1	1	3858.8
+−−−+	16	1	−1	−1	−1	−1	1	0.85998	4.83	1	−1	−1	−1	−1	1	3439.9
−−−−−	17	−1	−1	−1	−1	−1	−1	0.6902	8.47	−1	−1	−1	−1	−1	−1	2760.8
−++++−	18	−1	1	1	1	1	−1	0.78535	7.44	−1	1	1	1	1	−1	3141.4

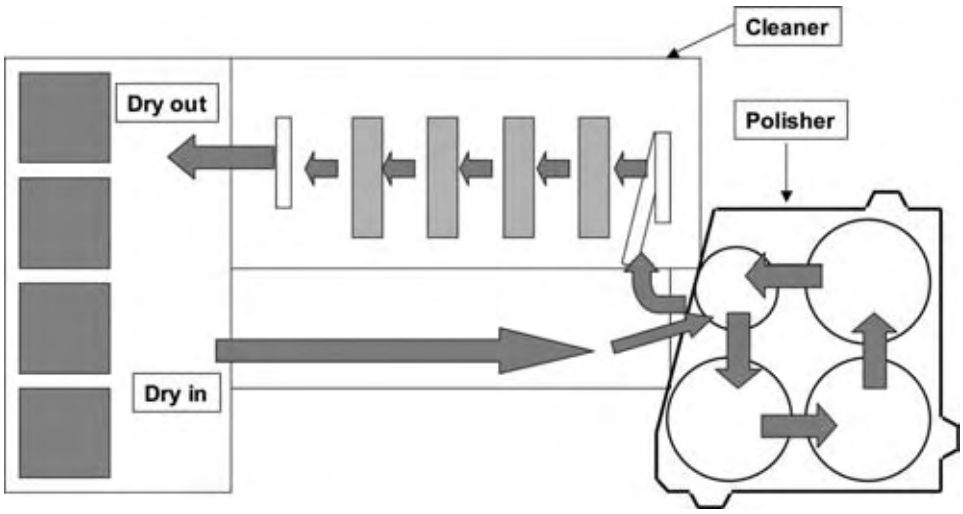


Fig. 5 Typical configuration of a multihead, multiplaten CMP system. *(View this art in color at www.dekker.com.)*

consist of an oxidizer to provide the chemical and abrasive material to provide the mechanical component of the polish. Typical oxidizers include low-pH strong oxidizers such as H_2O_2, $FeNO_3$, and KIO_4. The abrasives are typically composed of Al_2O_3 or SiO_2.

It has been discussed widely in the literature that the mechanism for W CMP consists at the beginning of the polish, the formation and removal of a surface layer of WO_3 that forms naturally on the W surface that is being polished:[2]

$$W + (OX)^{n-} + nH_2O \rightarrow WO_3 + (OX)^{(n+1)-} + 2nH^+ \quad (6)$$

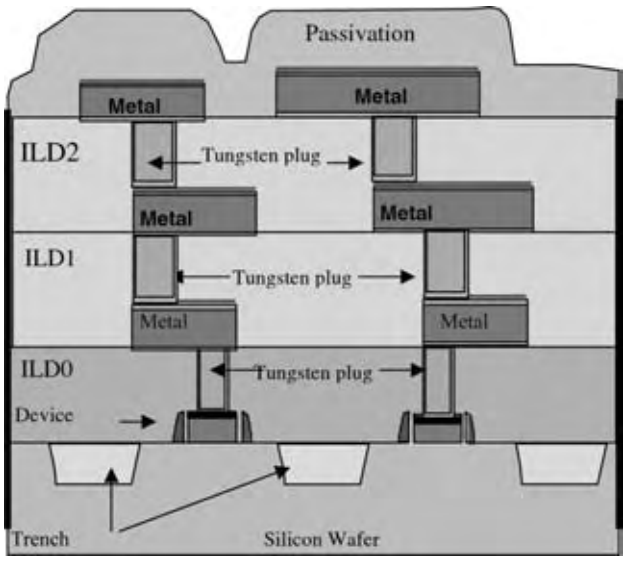


Fig. 6 Cross-section of a three-level metal device showing the trench isolation, W plugs, metal, and interlevel dielectric layers. *(View this art in color at www.dekker.com.)*

WO_3 is a surface oxide that prevents further oxidation and dissolution of the underlying W by the slurry once it is formed. In the high surface areas where the pad contacts the surface, mechanical abrasion of the abrasive in the slurry removes the surface layer and exposes W, while in the low area where the pad does not contact the surface (and abrasion does not exist with no pad to wafer contact), the W film remains intact due to the presence of the oxide layer. Once both the high- and the low-surface areas of the wafer are planarized and are fully in contact with the pad, mechanical abrasion of the underlying film takes place.

Fig. 10 shows the W deposition, W CMP, aluminum (Al) physical vapor deposition, and patterning processes that are used to form the contact plug. These processes are repeated at each metal level as shown earlier in Fig. 6. In order to maintain the planarity of the W plugs with the surrounding oxide layers, dishing of the metal plugs and erosion of the oxide must be minimized. Schematic definitions of dishing and erosion are shown in Fig. 11. Dishing occurs primarily because of pad bending into the plugs to remove metal. Erosion is the thinning of the ILD layer. Interlevel dielectric erosion is defined as the difference in the ILD thickness before and after the polish step. Both dishing and erosion occur during the overpolish step and are highly pattern dependent.[2] The wider the metal lines, the more likely it is that a pad will exert more pressure to remove the material within the plug or recess, increasing the likelihood of dishing. As metal pattern density increases, oxide erosion increases. Also, both pad and slurry properties (such as W: oxide selectivity of the slurry), in addition to pattern density, will affect dishing and erosion.[2] Further detailed discussions of W CMP are given elsewhere.[2]

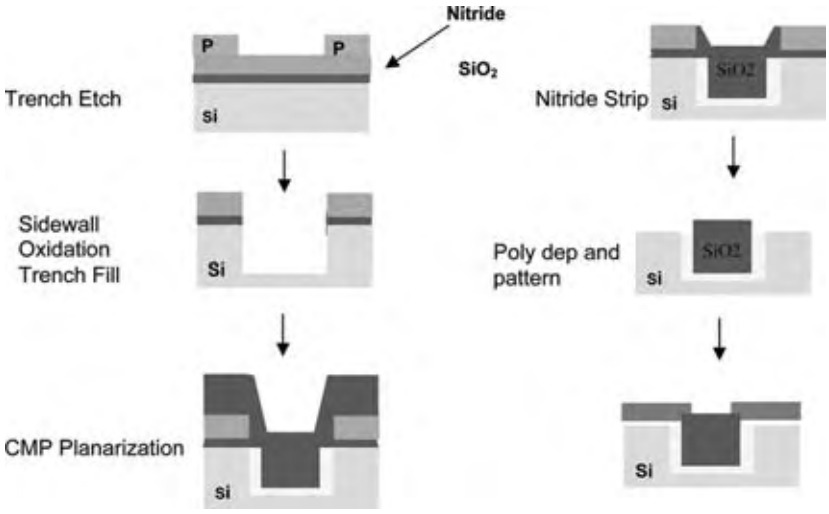


Fig. 7 Shallow trench isolation processing. *(View this art in color at www.dekker.com.)*

COPPER CMP

As device features diminish in size, Cu CMP has become the dominant CMP process for processing this device in the back end. Cu has several attributes that make it attractive as an interconnect material for such devices.[14] Electrochemically, Cu is noble compared to W and Al. Its hardness falls between W and Al and will not scratch as easily as Al. It has higher electromigration resistance than W or Al/Cu, which has been conventionally used as the interconnect metal in current devices.

One of the parameters for measuring device performance is interconnect delay or the time constant.[2]

The interconnect delay is due to the capacitance C and resistance R associated with metal lines. The resistance of a line is given by:

$$R = \rho(l/A) \tag{7}$$

where ρ is the metal resistivity, l is the length, and A is the cross-sectional area of the line. For devices, as feature sizes reduce (i.e., as A decreases), the metal resistivity ρ must also be significantly reduced to maintain the same RC, time constant. The speed of current devices is limited by metal layers (W and Al and Al/Cu alloys and Cu) and the capacitance C is governed by the intrinsic nature of the dielectric insulating material such as SiO_2. Table 3 shows the resistivity and dielectric constant of various materials. It is clear from the table that utilizing pure Cu in devices is advantageous. For transistors with >1 GHz speeds or Giga bit DRAM or >1 billion transistors, switching to Cu and low dielectric constant (κ) materials is absolutely necessary.

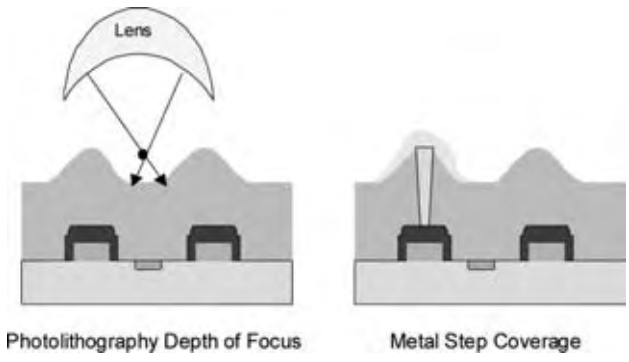


Fig. 8 Topography and its effects prior to CMP. *(View this art in color at www.dekker.com.)*

For Cu CMP, there are two distinct mechanisms for processing that have been proposed in the literature: mechanical abrasion of the Cu surface and subsequent transportation of the abraded material from the vicinity of the wafer surface. In order to fully dissolve all of the abraded material, the solubility of the Cu and the dissolution rates into the slurry must be high. Cu CMP slurries like W CMP slurries employ either a combination of complexing agent and oxidizing agent or an oxidizing acid. Also, like W CMP, surface film formation is also important. Depending on additives in the slurry, nonnative surface films can be formed to protect the film from further dissolution or corrosion during or after polish. Additives like benzonitrile (BTA) can be incorporated in the slurry to form Cu–BTA surface complex to protect the surface from corrosion after CMP. It has been shown elsewhere that the nature of the dishing of the Cu surface will be affected by the presence of the additive.[2]

For Cu CMP, the slurry will typically contain an oxidizer such as H_2O_2 with Al_2O_3 or SiO_2 based abrasives and could be at low or high pH. The most efficient Cu polish is conducted on a multihead and multiplaten tool like the one shown in Fig. 4. Bulk

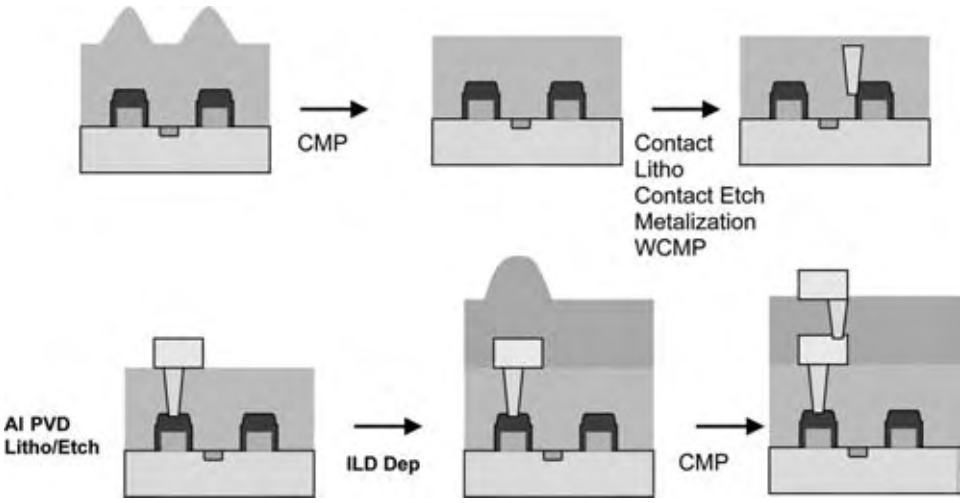


Fig. 9 Back end device construction with CMP. *(View this art in color at www.dekker.com.)*

removal of the polish is first conducted with high polish down-force on a hard pad with an oxidizer and an abrasive on platen 1. The second part of the polish is conducted with the same slurry at low down-force to clear the remaining Cu and the polish is stopped using an end point once the barrier metal is detected. The third part of the polish is used to clear the barrier metal and buff the underlying oxide layer. The slurry on the third platen will be similar to the slurry used for oxide polish.

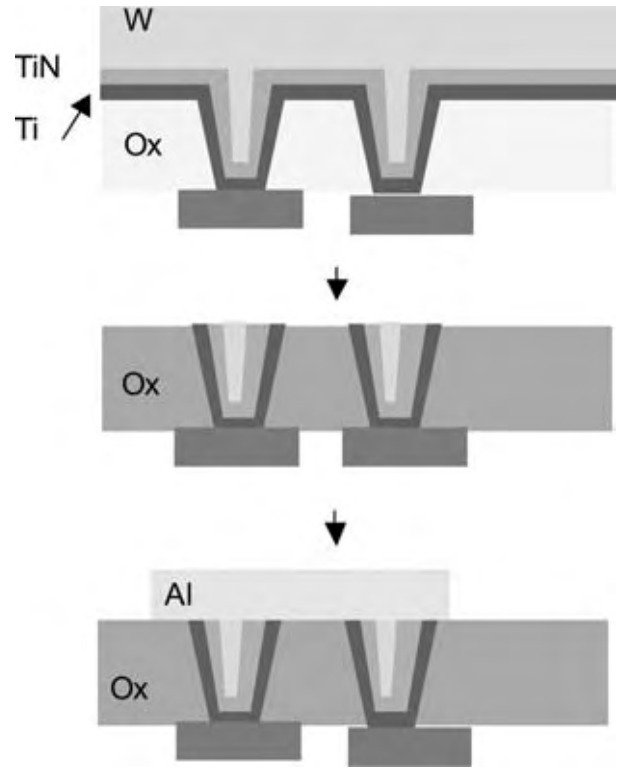


Fig. 10 Formation of Al line on top of W contact plug and Ti/TiN barrier after WCMP and deposition. *(View this art in color at www.dekker.com.)*

A process that is typically used for Cu processing is the damascene process. In the damascene process, the Cu is deposited over the entire wafer surface, including holes (vias or contacts) and grooves (metal lines). All metal that is above the surface is removed through polish.[15] In damascene processing, metal lines are not etched; it is only the dielectric that is etched to form plugs and vias. There are two types of damascene processes, dual damascene and single damascene. The difference between single and dual damascene is that in dual damascene, both the metal plugs and the metal lines are filled and polished at the same time; while in single damascene processes, the plugs and the metal lines are deposited, filled, and polished separately, increasing the number of polishing steps.[16] In single damascene processes with single via polish, maintaining the uniformity of the polish and lowering the defectivity are much more critical due to the higher level of topography associated with having vias as single features sticking out above the surrounding topography.

In addition to reducing the resistance of devices by migrating to higher-conductivity metallic materials such as Cu, a need has also risen to lower the capacitance of the devices to reduce the interconnect delay or time constant. This is done by utilizing low κ materials to replace SiO_2. Examples of low κ materials are shown in Table 3. Utilization of these low κ (dielectric) materials has created more challenges in Cu CMP. One example of a low κ material that has emerged to replace the SiO_2 is black diamond.[17] The dielectric constant of low κ films is on the order of 3.0–2.8. Ultralow κ films with <2.3 dielectric constants are also being considered as device dimensions diminish. Due to the softness of these films, superior planarization

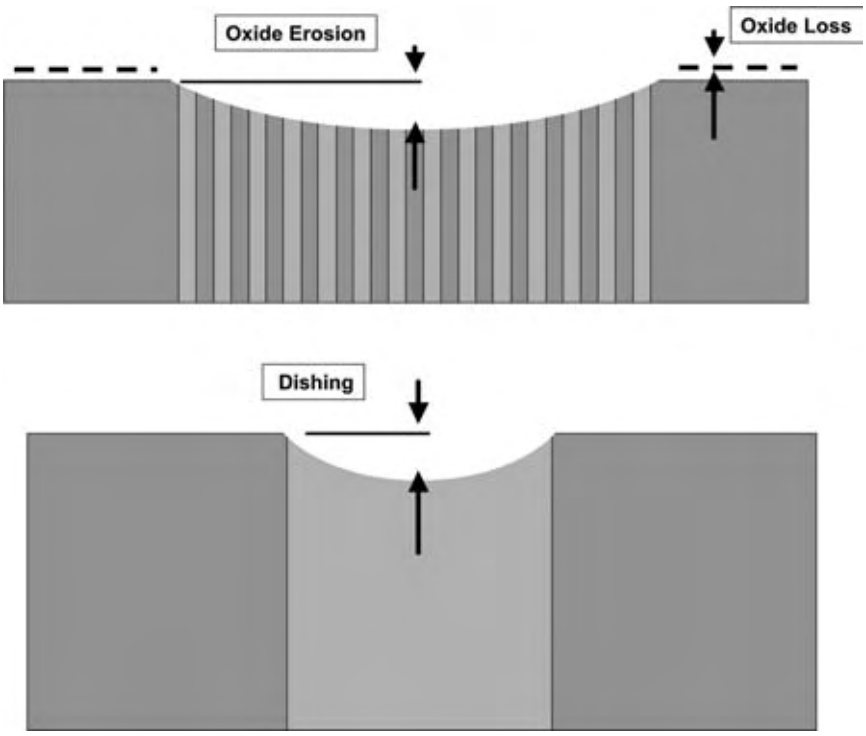


Fig. 11 Schematic definition of dishing and erosion. *(View this art in color at www.dekker.com.)*

and polishing techniques are needed to ensure low dishing and erosion and lower defectivity.

Defectivity in Cu CMP along with Cu-line dishing and oxide erosion is recognized to be an important problem. In-line probe tests immediately after polish are instituted to detect residual metal. Defects due to Cu corrosion are generated using electrochemical and photoinduced mechanisms because of the presence of water and ionized components in the slurry. Galvanic corrosion is also driven by device layout on the wafer formed as a result of galvanic cells generated between adjacent dies. During processing, exposure of the wafer to light must be minimized to limit this corrosion. Also, queue times are instituted immediately after Cu polish and scrub prior to the next processing steps to minimize Cu exposure to ambient conditions, which will cause corrosion. Also, due to the complexity of the slurry, particles and organic defects such as carbon particles generated from slurry residue are generally found and must be removed during the later stages of the polish or at the scrub step.

CONCLUSIONS

In summary, CMP is an enabling process technology that has become critical for manufacturing transistor devices. It has evolved from being an art in the late 20th century to a complex process technology that is an integral part of advanced semiconductor manufacturing for the 21st century. As technologies transitioned to the below 0.1 μm node, complex integration requirements have demanded more robust CMP processes. Due to the increasing sophistication and miniaturization of devices, CMP has become an even more critical technology for device manufacturing. The utilization of more advanced and complex materials such as Cu and low κ materials to meet device requirements will result in demands to understand and optimize process technologies such as CMP, to improve the robustness and flexibility of IC devices. Also, with the advent of more complex and integrated manufacturing fabs with larger-diameter wafers, the challenges of designing CMP processes to optimize the manufacturing

Table 3 Properties of common metals and dielectric materials

	Resistivity	Dielectric material	Dielectric constant, κ
Tungsten/aluminum	4.0	Silicon dioxide (SiO_2)	Thermal = 3.9, PECVD = 4.2
Aluminum	2.4	Fluorinated SiO_2	3.5
Copper	1.7	Inorganic polymers	2.7–3.5

processes in terms of cost will not diminish. With the increasing popularity and convergence of computing, entertainment, home, transportation, and mobile devices in the electronic space, the use of CMP in integrated circuit manufacturing will continue to grow.

ACKNOWLEDGMENTS

The author would like to thank the Technology and Manufacturing Division of Motorola Semiconductor Products Sector (SPS) for granting the permission, opportunity, and resources to publish this article. Materials included in this article are published with the sole permission and acknowledgment of Motorola SPS.

REFERENCES

1. Beyer, K. The inception of chemical-mechanical polishing for device. Applications at IBM; http://www-3.ibm.com/chips/micronews/vol. 5 no 4/beyer.html (accessed Jan 2002).
2. Steigerwald, J.; Murarka, S.; Gutmann, R. *Chemical Mechanical Planarization of Microelectronic Materials*; John Wiley & Sons, Inc.: New York, 1997; Vol. 5, 298–303.
3. Braun, A. Chemical mechanical polishing becomes more gentler, more efficient. Semiconductor International; http://www.einsite.net/semiconductor/index.asp?layout=articlePrint&articleID=CA180078 (accessed Jun 2001).
4. Lawing, A. Improving the results of post-CMP wafer-scale thickness measurements. Microbiology **2002**, *20* (1), 31–37.
5. Chen, K.W.; Wang, Y.L.; Chang, L.; Liu, C.W.; Lin, Y.K.; Wang, T.C.; Chang, S.T.; Lo, K. Novel strategies of FSG-CMP for within-wafer uniformity improvement and wafer edge yield enhancement beyond 0.18 micro technologies. IEEE International Symposium Conference Proceedings, (Cat. No.01CH37203) San Jose, CA, Oct 8–10, 2001; Vol. 1, 259 pp.
6. Dejule, R. CMP grows in sophistication, semiconductor international; http://www.einsite.net/semiconductor/index.asp?layout=articlePrint&articleID=CA170658 (accessed Jan 2003).
7. Post-CMP Wafer Cleaning; http://sst.pennnet.com/Articles/Article_Display.cfm?Section=Archives&Subsection=Display&ARTICLE_ID=4784&KEYWORD=CMP%20scrubbers (accessed Jan 2003).
8. Peters, L. A host of challenges for copper CMP cleaning. Semiconductor International; http://www.einsite.net/semiconductor/index.asp?layout=articlePrint&articleID=CA164130 (accessed Jan 2003).
9. Zabasajja, J.; Merchant, T.; Ng, B.L.; Banerjee, S.; Green, D.; Lawing, S.; Kura, H. Modeling and characterization of tungsten chemical and mechanical polishing process. J. Electrochem. Soc. **2001**, *148* (2), G73.
10. Geanne, V.; Lin, Z.; Budge, J.; Somit, J.; Basab, C. *Copper CMP Defect Reduction Using POU Filtration*; Semicon West Technical Program, Jul 2000.
11. Ge, L.M.; Dawson, D.J.; Cunningham, T.; Andricacos, P.C.; Searson, P.C.; Reidsema-Simpson, C.; Allongue, P.; Stickney, J.L.; Oleszek, G.M. Copper CMP characterization by atomic force profilometry. In *Electrochemical Processing in ULSI Fabrication and Semiconductor/Metal Deposition II*, Proceedings of the International Symposium (Electrochemical Society Proceedings, Vol. 99), 1999; 238 pp.
12. Steckenrider, J.S.; Guha, S.; Sethuraman, A.; Ra, Y.; Kim, H. Classifying defects for copper CMP process modules. Microbiology **2001**, *19* (8), 33–36, 38, 40, 42–45.
13. Israel's Nova expands metrology line. Solid state technology; http://sst.pennnet.com/Articles/Article_Display.cfm?Section=Archives&Subsection=Display&ARTICLE_ID=4215&KEYWORD=Nova (accessed Jan 2003).
14. Peters, L. Exploring advanced interconnect reliability. Semiconductor Int. **2002**, *25* (6), 63–64, 70–72.
15. Grief, M.; Schlueter, J.; Chadda, S. Interaction of copper CMP with interconnect integration; http://solidstate.archives.printthis.clickability.com/pt/cpt?action=cpt&expire=&urlID=5079 (accessed Jan 2003).
16. Zhang, Z.; Zheng, G.; Huang, R.; Yang, X.; Shao, B.; Zong, X. Application of copper interconnect and damascene technology in deep submicronic. Res. Progr. SSE. **2001**, *21* (4), 407–414.
17. Bremmer, J. A new class of insulating materials. Emergence of ultra low-k; http://solidstate.archives.printthis.clickability.com/pt/cpt?action=cpt&expire=&urlID=5079 (accessed Jan 2003).

Chemical Vapor Deposition

David G. Retzloff
Department of Chemical Engineering, University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

Chemical vapor deposition (CVD) is a relatively old process technology. Its early application can be traced back to the production of incandescent lamps in the 1880 time frame, where it was used to enhance the strength of carbon filaments via metal deposition. However, it was the use of CVD in semiconductor fabrication in the 1960s that ushered in the widespread application of this technology to a variety of applications. Chemical vapor deposition is defined as a process in which chemical reactions occur on or near the surface of a solid material called the substrate that result in gaseous molecules called precursors being deposited on the substrate surface as a thin film of solid material. It is the chemistry component that distinguishes this process from physical vapor deposition, which is primarily an evaporation/condensation process that does not involve chemical reactions. The CVD process occurs in a vessel termed the reactor and the residual gaseous material is pumped out of the reactor. The process pressure spans a range from 1 Torr to atmospheric. The temperatures involved range from 300°C to 2000°C. The CVD process is characterized by the method used to activate the reaction(s) that produce the deposited coating or the type of precursor that is used. Thus, thermal CVD, plasma enhanced CVD (plasma assisted CVD, plasma enhanced CVD), thermal laser CVD (LCVD), photochemical vapor CVD (PCVD or photo-CVD), and chemical beam epitaxy are classifications of the CVD process based on the method of activation of the reaction. A subclassification identifies the pressure used in the CVD process. Examples are low-pressure CVD and atmospheric pressure CVD. The only existing classification based precursor is metallo-organic CVD (MOCVD). It is not unusual for CVD occurring under atmospheric conditions to be identified solely by the method for activating the CVD reaction. The CVD reactors can be either vertical or horizontal depending on the application. Epitaxial pancake-type CVD and barrel reactors are vertical. Most other types of reactors are horizontal. These reactors are further classified as either hot-wall or cold-wall reactors. The rationale for choosing a specific CVD process and reactor is to produce a coating (film) on the substrate that strongly adheres to the substrate, is uniform in both composition and thickness, and has specific physical properties. These criteria are the reasons for the large variety of CVD processes and reactors designs.

Chemical vapor deposition is a mature technology and as such has a large literature covering both the theoretical aspects of the process and the applications. This literature is growing due to new applications of CVD as well as new understandings of the fundamental processes that occur in this technology. A review of CVD technology is presented in what follows. The CVD process and general CVD reactors will be discussed first, followed by a discussion of the theoretical aspects of CVD processing. Next, an overview of the equipment used in CVD processing will be considered. Finally, a list of CVD applications is presented to give the reader a feel for the diversity of applications using CVD.

THE CVD PROCESS AND CVD REACTIONS

An essential element of the CVD process is the chemical reaction that occurs to produce the coating/film. The precursors that constitute the chemical reactants must, of course, contain the chemical elements that will ultimately constitute the coating. In addition, these precursors must be stable at room temperature, react cleanly in the reactor without side reactions, not condense in the transfer lines, and be easily produced. The number of CVD reactions that are currently used is quite extensive and a complete listing of every reaction is beyond the scope of this review. However, a classification according to the method to activate the reactions for generic reaction type with specific examples is given in what follows. This will provide the reader with an overview of the diversity of applications of CVD.

Thermal Decomposition Reactions

In this method thermal energy from a heater is used to activate the reaction process. Some typical examples are:

Hydrocarbon deposition: $CH_4 \rightarrow C(s) + 2H_2$

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008100

This is used to produce graphite and diamond.

Halide decomposition: $WF_6 \rightarrow W + 3F_2$

Hydride decomposition: $GeH_4 \rightarrow Ge(s) + 2H_2$

Hydrogen reduction:
$SiCl_4 + 2H_2 \rightarrow Si(s) + 4HCl$

This is used to produce epitaxial silicon on semiconductor silicon wafers.

Plasma CVD Reactions

In this process either radiofrequency (RF) or microwave sources are used to activate the reactions:

Epitaxial silicon deposition:
$SiHCl_3 + H_2 \rightarrow Si(s) + 3HCL$

Ceramic materials (silicon carbide):
$3SiH_4 + C_3H_8 \rightarrow 3SiC + 10H_2$

Insulators (silicon nitride):
$3SiH_4 + 2N_2 \rightarrow Si_3N_4 + 6H_2$

Microwave CVD to produce diamond:
$CH_4 \rightarrow C(diamond) + 2H_2$

Thermal Laser CVD

In this process a laser at thermal equilibrium is used to initiate the reactions:

Metal deposition: $Cd(CH_3)_2 \rightarrow Cd + C_2H_6$

Photo-CVD

In this process the photons from a light source are used to initiate the reactions:

Silicon oxide: $SiH_4 + 2O_2 \rightarrow SiO_2 + 2H_2O$

Metallo-organic CVD

Gallium arsenide:
$(CH_3)_3Ga + AsH_3 \rightarrow GaAs + 3CH_4$

As can be seen from this short classification the applications of CVD are both numerous and widespread. The particular reactions shown are not unique to the CVD process listed. In fact, most of the reactions can be initiated by all of the methods listed. The rationale for choosing a particular reaction initiating method is discussed next.

The thermal CVD process in which thermal energy in the form of heat is used to activate the reaction process is the most common one used. Thermal activation typically occurs at temperatures above 900°C but can be substantially lower when metallo-organic precursors are used. Thus, the coating/film produced by this technique must be thermally stable at such temperatures and the thermal activation process should not lead to the incorporation of unwanted species.

Plasma activation of the reactions occurs at temperatures in the 300°C to 500°C range. Either an RF generator typically operating at 13.56 MHz to conform to FCC regulations or a microwave generator can be used to produce the plasma. The microwave-generated plasma has the advantage that the electrons that form the plasma have higher energy and can initiate reactions requiring more energy. Radiofrequency plasmas are more common because of their extensive use in plasma processing in the microelectronics industry. However, one cautionary point must be emphasized. Highly energetic plasma electrons can initiate unwanted side reactions when chemically possible. Thus, consideration should be given to this possibility when considering plasma CVD. The lower temperature deposition is the principal advantage of plasma CVD. A secondary benefit of this type of CVD process is the sputtering of the substrate that inevitably occurs. This sputtering changes the chemical character of the substrate surface. This is especially important for substrates that normally have an oxide formed on their surface, which is removed by the sputtering action. In all cases the surface becomes ''chemically active'' and the coating/film that results has better adhesion properties. A less commonly used process for plasma CVD is the plasma arc, which requires a large amount of energy and operates at extremely high temperatures. The plasma arc has been most notably used in the deposition of diamond.

Thermal laser CVD involves the same chemical deposition processes that occur in thermal CVD. It, therefore, is applicable to depositing the same types of materials that the thermal CVD process does. Its major application is the direct writing of thin films in semiconductor processing.

Photo-CVD involves the activation of the reaction by photons. As a result, no thermal energy is required for the deposition. This is the feature that distinguishes it from LCVD. Because the intensity of the photons from current sources is low, the deposition process is slow. Thus, this process currently has limited applicability.

Metallo-organic CVD involves the use of metallo-organic compounds as precursors. Its principal

advantage is the high quality of films produced by this process. Most MOCVD reactions occur at temperatures between 500°C and 1100°C and pressures between 1 Torr and atmospheric. Both the equipment and the chemicals are relatively expensive, resulting in high production costs. Metallo-organic CVD is used in microwave applications, field-effect transistors, solar cells, infrared detectors, and optical cavity lasers. However, the number of applications using this method is growing due to the refinements made in the process.

Chemical vapor deposition is also being used to manufacture complex parts both at the microminiature and the macro scale. The part is represented in a computer as a series of two-dimensional slices. The physical component is built one layer at a time using a laser with computer control to literally write the two-dimensional layer represented in the computer. For macro scale parts the process is slow due to the time it takes to produce a single two-dimensional layer. However, it can be applied to parts that are very difficult to make by any other method. This application is currently in the experimental stage but represents a future direction for the use of CVD.

A variation of CVD termed chemical vapor infiltration is being used to make fiber reinforced ceramic and metal composites as well as ceramic fibers. For reinforced fibers and ceramic composites the gaseous precursors diffuse into the porous structure of the substrate, which can be a fibrous weave or mat or inorganic porous foam. Upon reaction the substrate is densified to form a composite. Because the precursors must diffuse into the porous substrate, the process is operated in the kinetically controlled regime to avoid plugging the pores near the surface and stopping the densification process prematurely. This is inherently a slow process. Ceramic fibers can be produced using LCVD by focusing the laser beam on the end of a growing fiber in a CVD reactor. Either the fiber or the laser beam can be moved as the fiber grows to continue the process. This process is currently in the experimental stage but may be another future direction for the application of CVD.

A variant of CVD, termed plasma polymerization, has been used to produce polymer coatings. So far, it has been used in a limited number of applications most notably to produce biocompatible coatings for electronic sensors that are implanted in humans. This process is also being investigated for polymer coating applications. As the name implies, an RF generator is used to induce a plasma environment into which the monomer components are introduced. The electrons from the plasma induce the polymerization reaction at the surface of the substrate. The power input to the plasma must be kept low to prevent secondary reactions from destroying the polymer structure.

THEORETICAL CONSIDERATIONS IN CVD PROCESSING

A key consideration in CVD is the feasibility of the chemical reactions involved in producing the coating/film. The first step in assessing the viability of the proposed chemical reactions is to establish the thermodynamic feasibility of the reaction process. This is accomplished by determining the change in the Gibbs free energy for the reaction process under the conditions existing in the CVD reactor. The change in Gibbs free energy can be calculated from the free energies of formation of the reactants and products from the formula

$$\Delta G_r = \Sigma \Delta G_f \text{ (products)} - \Sigma \Delta G_f \text{ (reactants)}$$

where ΔG_f is the Gibbs free energy of formation and ΔG_r is the change in Gibbs free energy for the reaction. In most cases the Gibbs free energies of formation of the chemical species involved are readily available in tabulated form. When the Gibbs free energy of formation data are not available for a particular chemical species the well-known group contribution method may be used to estimate both the heat of formation and the entropy of formation. The Gibbs free energy of formation may then be calculated using the thermodynamic relation between the Gibbs free energy, the enthalpy, and the entropy. The thermodynamic feasibility of the CVD reaction(s) then follows from the equilibrium thermodynamic principle that a process will evolve in a direction that will minimize the Gibbs free energy. Thus, a chemical reaction is thermodynamically feasible if ΔG_r is negative. Furthermore, if chemical reaction equilibrium prevails in the CVD reactor, then the Gibbs free energy change for the reaction at constant temperature and pressure is 0. At chemical equilibrium $\Delta G_r = 0$ and it follows from the relationships

$$\Delta G_r = \Delta G_r^0 + RT \ln K, \quad K = \Pi (a_i)_i^v$$

that

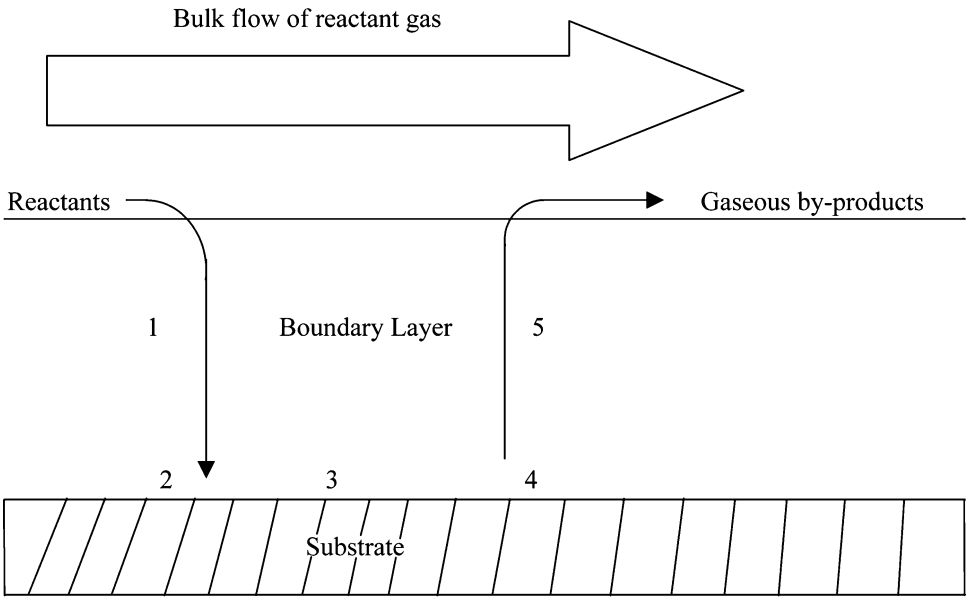
$$RT \ln K = -\Delta G_r^0$$

where a_i is the activity of species i, K is the equilibrium constant, and ΔG_r^0 is the standard Gibbs free energy change for the reaction. In this event the concentration of the chemical species present may be calculated from the equilibrium constant and the relationship between the activity and the concentration of species i. At a given temperature, pressure, and inlet concentration of the reactants, the composition and amount of material deposited can be determined using thermodynamic

principles. In addition, the possibility of multiple reactions, the existence of multiple solid phases and their composition can be determined using equilibrium thermodynamics. However, thermodynamic equilibrium occurs rarely in a CVD reactor and the values of the quantities calculated from thermodynamic considerations should be viewed as theoretical limiting values.

In general, the reactions that occur in the CVD process are kinetically controlled. Accurate theoretical predictions of reaction rates and kinetic parameters are currently not possible from fundamental principles. Thus, to obtain reliable data on the properties of the coating/film that will be obtained as well as the rate at which the coating/film will be formed, laboratory experiments are usually required. Fortunately, such experiments are easy to design and perform at a reasonable cost.

The CVD process is described in much the same manner as a chemical reactor in which a gas–solid reaction is taking place. The reactant gases flow past the substrate, forming a hydrodynamic boundary layer. The reactants diffuse through the boundary layer to the interface, where adsorption of the reactants on the substrate takes place. A chemical reaction occurs on the surface forming the coating/film. The by-products of the reaction are desorbed and diffuse through the boundary layer to the reactor outlet. The process is depicted in Fig. 1. Two important considerations emerge from this description with respect to operation of the CVD reactor. First, the concentration of reactants at the surface of the substrate decreases along the direction of flow. This has implications with respect to the uniformity of the thickness of the coating/film and must be taken into account in the design of the CVD reactor. Second, the rate-limiting step in CVD reactors can be the gas-phase kinetics (this is uncommon), the mass transport across the boundary layer, or the surface reaction kinetics. When the CVD process is limited by mass transport, either the diffusion of the reactants through the boundary layer to the substrate or the diffusion of the reaction by-products from the substrate to the bulk flow is the limiting step. This frequently is the result of a thick boundary and/or a high substrate temperature. For the case when the surface reaction kinetics is the rate-limiting step the reaction proceeds slowly relative to diffusion. This is usually associated with a thin boundary layer due to high gas velocity and low



The processes that occur during CVD are:

1. Diffusion of reactants through the boundary layer
2. Adsorption of the reactants on the substrate surface
3. Chemical reaction on the substrate surface forming the coating
4. Desorption of by-products from substrate surface
5. Diffusion of by-products to bulk flow stream

Fig. 1 Basic phenomena that occur in the CVD process: 1) diffusion of reactants through the boundary layer; 2) adsorption of the reactants on the substrate surface; 3) chemical reaction on the substrate surface forming the coating; 4) desorption of by-products from substrate surface; and 5) diffusion of by-products to bulk flow stream.

substrate temperature. The diffusion coefficients are large and the gas pressure is low. It is possible to change from kinetics as the rate-limiting step to mass transport by changing the temperature. This is because surface kinetics is the rate-limiting step at a lower temperature while mass transport is the rate-limiting step at a higher temperature. Similarly, pressure can be used to control the rate-limiting step since the diffusivity of a gas is inversely proportional to the pressure. Reducing the pressure to the Torr range can change the diffusivity by several orders of magnitude. This type of analysis is used to manipulate the deposition process to obtain the desired coatings/films.

There have been numerous theoretical studies of the CVD process, which include numerical simulations. Generally, the objective of these simulations is to predict the time required to achieve a specified coating thickness and the uniformity of the coverage of the substrate. The investigations are tailored to the specific geometry of the CVD reactor being studied. It is not the objective of this review to enumerate all or even a representative number of the studies that have been done and the reactor geometries that have been investigated. However, two investigations will be mentioned because they represent some of the early studies of the CVD reactor performance and established the basic models that form the basis of current theoretical research on this process. One of the early theoretical studies on the horizontal CVD reactor configuration commonly used in the semiconductor industry to deposit epitaxial silicon on a silicon wafer was undertaken by Jensen and Graves.[1] They applied conservation laws and a kinetic model to predict the growth rate of the epitaxial layer and study the effects of recycling on uniformity and deposition rate. Subsequently, Takoukis and coworkers[2] undertook a similar investigation of a pancake-type CVD reactor, which is a vertical reactor configuration with a rotating susceptor that holds the silicon wafers. This reactor configuration is also used in the semiconductor industry for the formation of epitaxial silicon on a semiconductor wafer. They used conservation laws and a kinetic model for the di-chlorosilane reaction to obtain the equations modeling the process. In their study both the vector velocity field and the growth rate of the epitaxial layer were studied. These numerical investigations illustrate the principles used to understand what is occurring in a CVD reactor as well as the depth of understanding of the physics and chemistry of the process. What is notably missing is a detailed fundamental understanding of what variables control the structure and morphology of the coating/film. The films may be crystalline or amorphous. The crystalline films can be columnar with or without facets or composed of fine grains. What is known from the experiment is that the microstructure of the CVD deposit can be manipulated by changing the temperature, pressure, and/or degree of supersaturation of the vapor. In this regard, at low pressure the boundary layer thickness is small and surface kinetics is the controlling step. Under these conditions the coating/film has a tendency to be fine grained. This is also the case for low temperature and high supersaturation. However, as coating thickness increases a columnar-like grain structure develops and becomes more pronounced as the thickness increases further.

The mechanisms by which a film is formed on a substrate by CVD are still not clearly understood. Several theories have been developed using thermodynamics and/or nucleation theory. However, which view is correct is still a matter of ongoing controversy. It is unlikely that this will be resolved in the near future.

CHEMICAL VAPOR DEPOSITION EQUIPMENT

There are a large variety of designs for CVD equipment. Commercial manufacturers generally offer a limited number of designs that are targeted to specific application needs. However, the total number of applications is large. Chemical vapor deposition equipment is frequently custom built by the user for a specific application. Thus, the discussion of CVD equipment presented here will focus on those features of CVD equipment that are generic to all CVD equipment of a particular class. Because most precursors used in CVD are in the gaseous state, CVD equipment frequently has an inlet train consisting of gas purifiers and flowmeters to regulate the flow of reactants to the CVD reactor. In some cases when the precursor is in the liquid state a vaporizer will also be present in the inlet train.

At the basic level CVD reactors fall into two classifications—open and closed reactor systems. In the closed CVD system the precursors are loaded into the CVD chamber together with the specimens to be treated. The system is then closed and the temperature increased to initiate the chemical process. The process continues for a time sufficient to produce the required effect. The temperature is then reduced to ambient so that the reactor may be opened and the specimens removed. This reactor design is frequently used for the purification of metals and chromizing parts. The dominant type of CVD reactor is the open type. Flowing precursors continuously enter this reactor and the gaseous by-products of the chemical process are continuously removed from the reactor (usually with a vacuum pump) and appropriately treated for discharge into the environment.

The thermal CVD reactors mentioned in the introduction are divided into two classes—hot wall reactors and cold wall reactors. Hot wall reactors consist of

constant temperature furnaces heated by resistance heating elements. The substrates being coated are placed in the reactor, the reactor temperature is raised to the desired level, and the precursors are fed to the reactor. For this type of reactor the temperature control is very tight. However, deposition occurs everywhere inside the reactor including the reactor walls. This requires periodic cleaning of the reactor walls to prevent deposit buildup and cross-contamination. In cold wall reactors the substrate is heated either by induction or by radiant heating. This causes the coating to deposit predominately on the substrate with only minimal coating of the reactor walls. Thermal CVD reactors can be either open or closed reactors.

Chemical vapor deposition reactors that employ a plasma environment to initiate the chemical reactions have the advantage of being able to form coatings at significantly lower temperatures than those required in thermal CVD reactors. In fact, if these temperatures were used for the thermal CVD reactors, no chemical reactions would occur and no coatings would result. These reactors use an RF or microwave generator to produce an electromagnetic field that interacts with the reaction process. The reason why the plasma CVD reactors can operate successfully at these low temperatures is that the electromagnetic field produced by the generator can efficiently couple to the electrons present to produce very energetic electrons (with a characteristic temperature of 24,000 K) that participate as reactants in the chemical reaction process. Essentially, the radiation field supplies the energy required by the chemical reactions. This drives the relevant CVD chemical reaction process at significantly lower thermal temperatures than would be required in a thermal CVD reactor. For thermally sensitive substrates or substrates that are adversely affected by high temperatures, plasma CVD is the method of choice. Plasma CVD reactors are normally open reactors that operate at pressures in the Torr range.

Thermal laser CVD reactors use a laser to provide thermal energy at the adsorbing substrate to initiate the chemical reactions leading to deposition. Due to the laser wavelength used there is essentially no energy adsorption by the gas molecules. This reactor design allows ''direct writing'' of the film on the substrate because the thermal energy required for the CVD reactions to occur is localized at the focal point on the surface of the substrate. This feature is very desirable for applications in the semiconductor industry.

Photo-CVD reactors employ an optical source to supply radiation at a wavelength sufficient to initiate the reaction process. This wavelength is typically in the ultraviolet (UV) part of the electromagnetic spectrum because it is at these wavelengths that the photons have sufficient energy to break chemical bonds. Ultraviolet lasers are frequently used because they are capable of providing the required intensity to achieve a reasonable deposition rate. The energy for the reaction process is supplied by the photon field. As a result no heat is required for this type of reactor. Also, in contrast to LCVD reactors there is no constraint on the type of substrate that can be coated, as there is no requirement for adsorption of the photons by the substrate. A drawback of this type of CVD reactor is the slow rate of deposition that is normally achieved with current UV-lasers. Both laser and photo CVD reactors are more expensive to use when compared to thermal CVD reactors because of the expense of the source used to initiate the reaction process. As a result they are used in more specialized applications where alternatives are limited. Photo-CVD reactors can be either open or closed reactors and operate at subatmospheric pressures.

The MOCVD process does not involve a new type of CVD reactor. The reactor used is basically a thermal CVD rector. Metallo-organic compounds are used as precursors usually in conjunction with other reactants to reduce the operating temperature required in a thermal CVD reactor. The precursors and equipment costs are high for MOCDV reactors. As a result MOCVD is used when high-quality coating is required. These reactors are typically open-type reactors that operate in the Torr to atmospheric pressure range.

Chemical vapor deposition reactors are either horizontal- or vertical-type reactors. The orientation assigned to the reactor (horizontal or vertical) refers to the orientation of the flow and the substrates within the reactor. The effects of buoyancy and inertial forces on the flow within the reactor are the principal distinction for these two types of reactors. A generic depiction of these two types of reactors is given in Fig. 2. A complete list of examples of CVD reactors and the design criteria used for each specific reactor is beyond both the scope and page limitations of this review as the variations in reactor configuration within each classification are truly endless, being limited only by the inventiveness and objectives of the designer.

CVD APPLICATIONS

The number of uses of CVD in providing coatings/films is enormous and growing. In this section a reasonably complete list of applications of CVD is presented to provide the reader an appreciation of the extent to which CVD is used in commercial applications. To make this task manageable and the presentation useful for the reader, the presentation is organized by the coating material produced. The applications of the specific coating are included for each material. The first materials considered are metal coatings followed by nonmetallic materials, semiconductor materials, and ceramic materials.

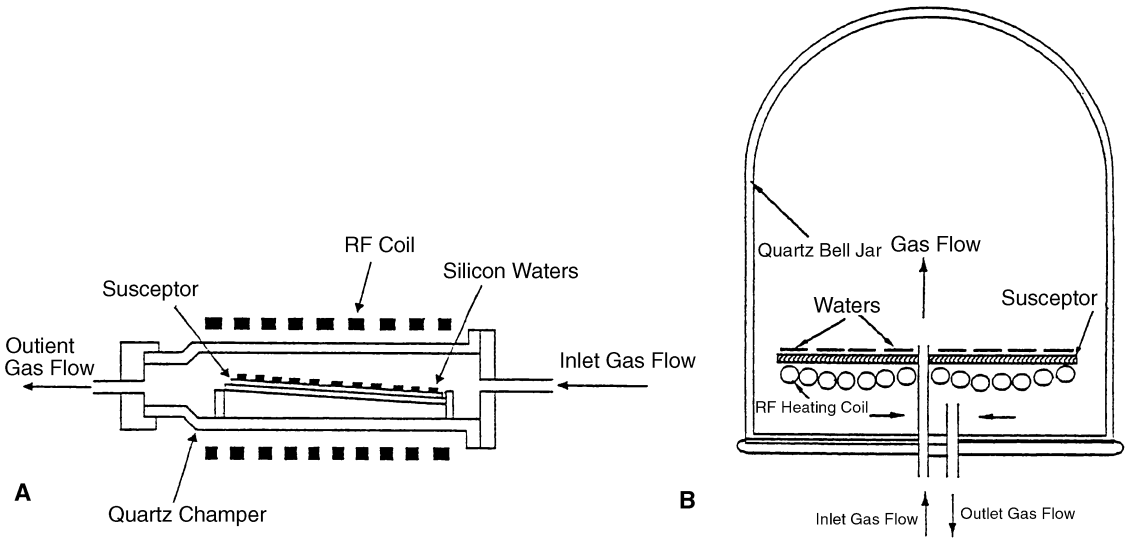


Fig. 2 (A) Schematic of a horizontal CVD reactor. (B) Schematic of a vertical CVD reactor.

Metals

- Beryllium: Wall coatings for nuclear fusion reactors.
- Chromium: Corrosion and oxidation protection of metal parts.
- Copper: Conductive films semiconductor devices.
- Gold: Metallization of contacts in semiconductor applications.
- Iridium: Corrosion resistant coatings for rocket engines.
- Molybdenum: Semiconductor gate metallization, high-power laser mirrors, and coatings for solar converters.
- Nickel: Forming tools for metal and plastic processing, electrical contacts.
- Niobium: Nuclear fuel particle coating.
- Platinum: Catalysts in automobile emissions, ohmic contacts, coatings for high-temperature crucibles.
- Rhenium: Crucibles, high-temperature furnace heaters.
- Tantalum: Corrosion resistant coatings, thin film capacitors.
- Titanium: Corrosion resistant coatings.
- Tungsten: Coatings for targets in x-ray cathode tubes, absorber coatings for solar collectors.
- Zinc oxide: Photoconductor device coatings, piezoelectric devices.

Nonmetallic Material

- Boron: Wall coating for nuclear reactors.
- Graphite: Coating for nuclear reactors, heating electrodes.
- Diamond: Coating for cutting tools, coatings for wear resistance, optical coatings, semiconductor devices.

Semiconductor Applications

- Aluminum nitride: Packaging material for electronic devices, dielectric layers.
- Bismuth titanate: Dielectric material.
- Boron nitride: X-ray lithography masks, high-temperature crucibles.
- Germanium: Photovoltaic devices, photodetectors, tailor bandgap on silicon.
- Lead titanate: Ferroelectric material.
- Silicon dioxide: Electrical insulator.
- Silicon nitride: Diffusion barrier and electrical passivation.
- Epitaxial silicon: To improve the performance of semiconductor devices by reducing defects and impurities.
- Silicides ($AlSi_2$, $CrSi_2$, Cr_3Si, $CoSi_2$, $MoSi_2$, $TiSi_2$, WSi_2, $TaSi_2$, $CoSi_2$): Gate metallization and interconnects.
- III–V and II–VI compounds (GaAs, GaN, GaP, InAs, InP, AlAs, BP, InGaAs, AlInAs, AlGa, As, GaPAs, InGaN, ZnSe, ZnS, ZnTe, CdS, HgTe, CdMnTe): Microwave devices, photovoltaic devices, light emitting diodes, field effect transistors.
- Strontium titanate: Dielectric material.
- Tantalum oxide: Gate insulators in metal–oxide–semiconductor devices, dielectrics for capacitors, optical coatings.
- Titanium oxide: Dielectric layer in thin film capacitors.

Ceramic Materials

- Aluminum oxide: Thin film applications in field effect transistors, coatings for carbide tools.
- Borides (HfB_2, Mo_2B_5, NbB_2, TaB_2, TiB_2, W_2B_5, ZrB_2): Wear and corrosion resistance applications.
- Boron carbide (B_4C): Coating for jet engine nozzles, coatings for shielding in nuclear reactors.
- Chromium carbide: Coating for wear and corrosion resistance.
- Hafnium carbide: Coatings for superalloys.
- Iron oxide: Beam splitters in optical devices.
- Silicon carbide: Coatings for ceramic heat exchangers, radiation resistant semiconductor applications, susceptor coatings, heteroepitaxial film on silicon, coatings for nuclear waste containers.
- Tin oxide: Coatings for solar cells, antistatic coatings, coatings for plate glass and light bulbs.
- Titanium carbide: Coatings for cutting and milling tools, coatings for ball bearings, coatings for extrusion nozzles.
- Tungsten carbide: Porous carbon coating for catalytic applications.
- Zinc selenide: Optical and infrared windows.
- Zinc sulfide: Optical and infrared windows.
- Zirconium oxide: Piezoelectric devices, oxygen sensors.

This list of materials deposited by CVD techniques and their applications is by no means exhaustive. Their number continues to increase as people explore new materials and applications for CVD coatings. What is apparent is the broad range of materials and applications for CVD coatings/films. The CVD process is extremely versatile and simplistic. All that is required are gas phase precursors, an energy source to initiate the chemical reactions involved, and a chamber in which to conduct the CVD process. Experimentation with physical parameters will yield high-quality coatings/films with excellent adhesion properties. What is missing is a detailed understanding at the fundamental level of the cause and effect relationships between the velocity field, reactor variables (temperature, pressure, flow rate) and the structural properties of the coating, adhesion to the substrate, and uniformity of the coverage. The numerical simulations of the flow fields that exist in the CVD reactor show that such flows are quite complex. The details of coating/film growth from the molecular point of view is evolving. It is premature to attempt a prediction of the resulting coating/film and its properties from the operating conditions in the CVD reactor. In the foreseeable future one must be content to have general guidelines coupled with experience to choose a CVD reactor design suitable for the application being considered, and it is equally true that experience and testing will remain the main tools for adjusting reactor variables to achieve the required coating/film properties.

CONCLUSIONS

Chemical vapor deposition has developed into a mature process technology. It has numerous applications in the areas of coatings, semiconductor devices, metal purification, and fiber reinforced ceramic material. It has also been used in conjunction with a closed reactor system to purify metals. It is currently being investigated for applications in producing microminiaturized parts and parts that are too complex to be readily made by standard manufacturing techniques. Plasma CVD techniques have been used with monomer precursors to produce polymer coatings. It is expected that new CVD applications will continue to be developed in the foreseeable future. This is also true for the design of CVD reactors, which is frequently dictated by the application. However, advances in the fundamental understanding of the physics and chemistry that occur in the CVD process are likely to proceed at a much slower pace. The process by which a coating assembles itself on the surface of the substrate is extremely complex. It will take major new insights to identify and describe in mathematical terms the details of this process. The interactions between the transport process, the chemistry, and the mechanism(s) by which the film grows on the substrate surface are not well understood and this is not likely to change in the near future. Fortunately, advances in the applications of CVD can and do proceed without this detailed fundamental knowledge.

REFERENCES

1. Jensen, K.F.; Graves, D.B. Modeling and analysis of low pressure CVD reactors. J. Electrochem. Soc. **1983**, *130*, 1950–1957.
2. Oh, I.; Takoukis, C.G.; Neudeck, G.W. Mathematical modeling of epitaxial silicon growth in pancake chemical vapor deposition reactors. J. Electrochem. Soc. **1991**, *138*, 554–567.

Chiral Drug Separation

Bingyun Li
Donald T. Haynie
Biomedical Engineering and Physics, Bionanosystems Engineering Laboratory, Center for Applied Physics Studies, Louisiana Tech University, Ruston, Louisiana, U.S.A.

INTRODUCTION

Consideration of chirality is now an integral part of drug research and development and the regulatory process. There is no choice! Enantiomeric forms of a drug can differ in potency, toxicity, and behavior in biological systems. Enantiomers of all chiral bioactive molecules have to be separated and tested. The Food and Drug Administration (FDA, U.S.A.), and regulatory authorities in Europe, China, and Japan have provided guidelines indicating that preferably only the active enantiomer of a chiral drug should be brought to market.

This entry discusses the main chiral drug separation methods, viz. gas chromatography (GC), high-performance liquid chromatography (HPLC), capillary electrophoresis (CE), and some new techniques. The first part presents the concept of molecular chirality and some examples of chiral drugs. The importance of chiral drug separation is briefly discussed. The second part describes the main chiral drug separation techniques and related chiral recognition mechanisms, as well as available chiral selectors. Some typical examples are given, and the pros and cons of chromatographic and capillary electrophoretic techniques are compared. Readers are also referred to specialized review articles for details. The last part summarizes and briefly discusses future development trends in chiral drug separation techniques.

BACKGROUND

What are Chiral Molecules?

A molecule with an asymmetric carbon center has a unique three-dimensional shape and is called chiral, from Greek *cheir* meaning ''hand.'' A chiral molecule and its mirror image are not completely identical: they differ in their ''handedness'' and are not superimposable. Such chiral molecules are called ''enantiomers,'' from Greek *enantios*, meaning ''opposite.'' A human right hand and left hand are enantiomers: they are mirror images, but a ''left'' glove cannot be worn on a right hand. The arrangement of thumb and fingers in three dimensions makes a right hand and a left hand distinctly different from each other. The two different forms of a chiral pair comprise the same number and types of atoms, and they are commonly described as D- and L-isomers with reference to their ability to rotate polarized light. An equimolar mixture of enantiomers is called a ''racemate.''

Fig. 1 depicts 2-butanol. The top two structures are mirror images. Below, ''right-handed'' 2-butanol has been rotated so that the OH group points to the left for comparison with ''left-handed'' 2-butanol. Although the carbon frameworks of the two molecules align, the position of the hydroxyl group is different. In the rotated version of right-handed 2-butanol, in the lower left panel, the hydroxyl group points into the page; in left-handed 2-butanol, in the lower right panel, the hydroxyl group points out from the page. It is impossible to align the two molecules completely without breaking bonds. Left-handed and right-handed forms of a molecule can have profoundly different properties in a biological context. In the food industry, for instance, left-handed limonene smells like lemons, while the right-handed molecule smells like oranges. Different enantiomers can differ widely in their biological properties because chirality is related to the three-dimensional structure, and one form may be more suitable for specific interaction with a biological molecule, such as a receptor, enzyme, etc.

Chiral Drugs

Most synthetic drugs developed in the past were not chiral, though some were. Drugs developed from natural products are largely chiral. Currently, about 40% of the drugs in use are known to be chiral. A report from Technology Catalysts International, Falls Church, VA, states that worldwide sales of chiral drugs in single-enantiomer dosage forms grew at an annual rate of more than 13% to $133 billion in 2000, and that sales could hit $200 billion by 2008.[1] About one-third of all dosage-form drug sales in 2000 were single enantiomers. By geography, the United States is the largest consumer of enantiomeric fine chemicals, contributing 60% of the worldwide total. Drug companies continue

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039232

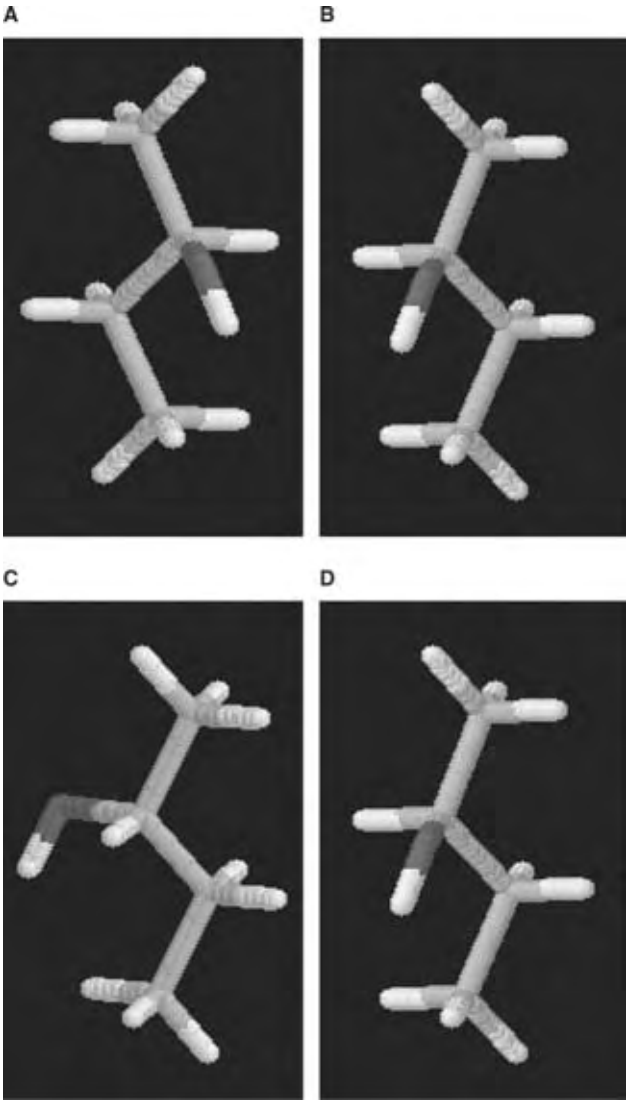


Fig. 1 2-Butanol chiral molecules: (A) right-handed (*R*) 2-butanol, (B) left-handed (*S*) 2-butanol, (C) rotated (*R*) 2-butanol, and (D) (*S*) 2-butanol. *(View this art in color at www.dekker.com.)*

to develop chiral drugs as single enantiomers and use chirality to manage drug life cycles.

Most commercially available drugs are both synthetic and chiral. A large number, however, are still marketed as racemic mixtures. Only about one-third of drugs are administered as pure enantiomers. The respiratory drug montelukast (Merck), the antirheumatoidal infliximab (Centocor), the ophthalmic drug for the treatment of glaucoma latanaprost (Pharmacia), and the prostate hyperplasia agent tamsulosin (Boehringer Ingelheim Pharmacy) are among the best-selling single-enantiomer drugs. The chiral drug (*S*)-(+)-ibuprofen is marketed as fast-acting, and it reaches therapeutic concentrations in blood in 12 vs. 30 min for racemic mixtures. Allegra, an isomer from a metabolite of seldane, is used as an allergy medication. Table 1 lists further examples of chiral drugs and their bioactivities. The enantiomeric forms of a drug can differ markedly in potency, toxicity, and behavior in biological systems.

Importance of Chiral Drug Separation

In the early 1980s analytical chiral separation was a rather difficult task, and preparative synthetic and separation methods were not as advanced as today. Nevertheless, it was clear that chiral drugs should be enantioseparated and that each enantiomer should be used separately. Nowadays, enantiomers are considered distinctly different compounds, as enantiomers of drug substances may have distinct biological interactions and, consequently, profoundly different pharmacological, pharmacokinetic, or toxicological activities.[2] The body is highly chiral selective; it will interact with each racemic drug differently and metabolize each enantiomer by a separate pathway to produce different pharmacological activity. One isomer may thus produce the desired therapeutic activities, while the other may be inactive or produce unwanted side effects (see Table 1). Even when side effects are not serious, the inactive enantiomer must be metabolized and thus represents an unnecessary burden for the organism.

One chiral form of the drug naproxen has 28 times the anti-inflammatory activity of the chiral relative. One isomer of dopamine, used to treat Parkinson's disease, acts on nerve cells to control tremor, while the other is toxic to nerve cells. Racemic mixtures of the drug thalidomide were marketed to pregnant women in the 1960s to counter morning sickness. Therapeutic activity, however, comes exclusively from the (*R*)-(+)-enantiomer. It was discovered only after several hundred birth defects had resulted from administration of thalidomide that the (*S*)-(+)-enantiomer is teratogenic. By then, consideration of chirality has become an integral part of drug research and development and of the regulatory process.

In general, use of the more active isomer of a drug has the following advantages:

- Fewer or diminished side effects, which may result from the unwanted isomeric form.
- Automatically halved dosage for a patient.
- Decreased waste due to decrease in manufacturing of unwanted isomer.
- New commercial opportunities for "racemic switching:" a drug previously marketed as a racemate can be redeveloped and introduced as an enantiomerically pure form, possibly useful for extending patent protection of a key product.

Table 1 Examples of chiral drugs and functions

Chiral drugs	Bioactivity
Albuterol	D-isomer may provoke airway constriction; L-isomer avoids side effects
Ethambutol	The (*S*,*S*)-form of ethambutol is a tuberculostatic agent; the (*R*,*R*)-form causes optical neuritis that can lead to blindness
Levodopa	The levodopa (L-dopa) is a Parkinson's disease agent; the D-form causes serious side effects, such as granulocytopenia
Penicillamine	The (*S*)-enantiomer has antiarthritic activity; the (*R*)-form is extremely toxic
Propanolol	Racemic compound is used as drug; however, only the (*S*)-(−)-isomer has the desired β-adrenergic blocking activity
Propoxyphene	α-L-isomer is antitussive (cough); α-D-isomer is analgesic (pain)
Thalidomide	The (*S*)-isomer has the desired antinausea effects; the (*R*)-form is teratogenic and causes fetal abnormalities, such as severely underdeveloped limbs

There are obvious benefits to studying the properties of the enantiomers of a chiral drug molecule with respect to therapeutic efficacy and safety. In view of this, since 1992 the FDA and the European Committee for Proprietary Medicinal Products have required that the properties of each enantiomer be studied separately before decisions are taken to market the drug as one of the enantiomers or as a racemate.[3] This requires powerful means of chiral drug detection and separation. In addition, there is increasing awareness of the need to reevaluate the properties of individual enantiomers of currently marketed racemic drug molecules. The effect has been a significant increase in demand for sensitive chiral analytical and separation methods. At the same time, the number of new chemicals entering development within the pharmaceutical industry has increased significantly. By now most drug companies have clear guidelines recommending that the enantiomers of all chiral bioactive molecules be separated and tested.

The ideal way to obtain pure drug enantiomers would be enantioselective synthesis. This, however, is rarely practical, usually complicated, and almost always expensive. There is little control over which chiral form of a chemical compound will be formed during a typical production process. This lack of control generally results in production of equal amounts of the various possible chiral forms of a compound. Often, therefore, separation of intermediates or final products from a racemic mixture is required. Increasing interest is being paid to development of methods of efficient, high throughput, and sensitive chiral separations, control of chemical synthesis, assessment of enantiomeric purity, and determination of pharmacodynamics.

The various examples of different therapeutic, toxicological, and pharmacokinetic properties of the enantiomers of chiral drugs provide a strong impulse for the development of techniques for chiral drug separation. Enantiomers can differ in absorption, distribution, protein binding, and affinity to the receptor. Such properties have been exploited for the development of powerful techniques for achieving analytical-scale chiral separations, quantifying minor enantiomeric impurities in chiral drugs, and preparing gram to multi-ton amounts of enantiomerically pure chiral drugs. Chromatographic techniques, such as HPLC, GC, thin layer chromatography, and supercritical fluid chromatography have been developed for chiral separations. Capillary zone electrophoresis, capillary gel electrophoresis, electrokinetic chromatography, and capillary electrochromatography have proved powerful alternatives to chromatographic techniques.

CHIRAL DRUG SEPARATION PRINCIPLES AND TECHNIQUES

Principles of Chiral Separation and Chiral Selectors

Principles of chiral separation

Separation of enantiomers has been achieved using GC, HPLC, and CE. Chiral recognition generally

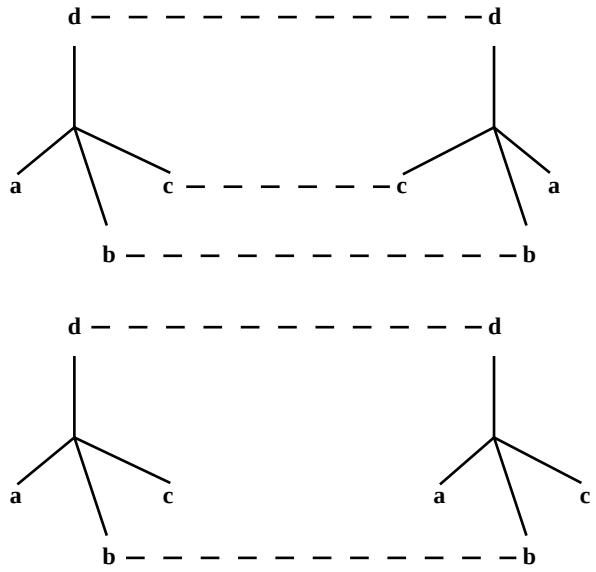


Fig. 2 Three-point interaction model.

depends on a minimum of three simultaneous interactions between the selector and selected—the so-called three-point-interaction rule of Dalgliesh (Fig. 2).[4,5] At least one of these interactions must be stereoselective to form diastereomeric complexes, and thereby enable chiral separation.

The principle task of chiral separation is to create the selectivity essential for separation of stereoselectively different forms of compounds, which may be recognized as such only during the interaction with a chiral selector. This is the separation principle for chromatographic techniques and also for chiral CE. In chromatographic chiral separation, there is a distribution of analyte between two immiscible phases, and these should exhibit different mobilities. Commonly one phase is mobile and the other is stationary. In chiral CE there are not two immiscible phases present but pseudophases at best, or even only one monophasic, homogenous system. Chiral recognition, however, occurs at the molecular level, not on the macroscopic level of the phases. A separation technique therefore must allow the transfer of the molecular level event (in this case chiral recognition) to macroscopic phenomena: different retention times of enantiomers in chromatography and different effective mobilities of enantiomers in electrophoretic techniques.

Immiscibility of phases is a prerequisite in chromatographic separation because pressure as a driving force cannot select a given component from several species in the same phase. Under certain circumstances, however, electrically driven mobility can be selective for one or several species residing in the same phase. So, the immiscibility requirement of the two phases does not apply to chiral CE. In other words, the principal difference between chromatographic techniques and CE is that pressure cannot distinguish between different molecular components in a monophasic system, whereas electrically driven mobility can do this under certain circumstances. The phenomenon responsible for chiral separation is the same in chromatographic and electrophoretic techniques: enantioselective interaction between the analyte enantiomers and a chiral selector.

Chiral selectors

Enantiomers are distinguished on the basis of their interaction with a chiral selector. Development of chiral selectors or chiral stationary phases (CSPs) for GC, HPLC, and CE has rapidly opened a new dimension in the area of chiral drug separation techniques. There are different chiral selectors available for enantiomeric separation of drugs and pharmaceuticals. Finding a suitable chiral selector, whether immobilized on a solid support (GC, HPLC) or added to a running buffer (HPLC, CE), is still often based on trial and error. A few predictions can be made, however, if common structural elements are present. After a selector has been chosen, variables, such as the nature, ionic strength, and pH of buffer, can be varied, as can presence of organic modifiers, temperature, and so on.

Among available selectors, native and derivatized cyclodextrins (CDs) are the most widely used ones at this time. A majority of drug and pharmaceutical applications have been achieved with CDs. Use of CDs as chiral selectors is the subject of a number of reviews (e.g., Refs.[6,7].). Native CDs are cyclic oligosaccharides consisting of six α-CD-, seven β-CD- or eight γ-CD-glucopyranose units. A truncated cone provides a hydrophobic cavity; the exterior surface, surrounded by hydroxyl group, is hydrophilic. Low UV absorbance, low cost, and water solubility are attractive properties of CDs for use as chiral selectors. In addition to CDs, other chiral selectors, such as natural and synthetic chiral surfactants, crown ethers, proteins, oligo- and polysaccharides, macrocyclic antibiotics, and chiral ligands have been applied to chiral separations. Important selectors and some chiral recognition mechanisms are given in Table 2. Some chiral selectors developed thus far can efficiently resolve enantiomers of various important drugs.

Chiral Drug Separation Techniques

The main methods used for chiral drug separation are GC, HPLC, and CE.[2,4,6–25] Other techniques, such as chiral crystallization and enzyme-based kinetic separation, have also attracted attention.[26]

Table 2 A variety of chiral selectors employed for chiral separation and related chiral recognition mechanisms

Chiral selector	Chiral recognition mechanism	Examples	Separated enantiomers
CDs	Chiral recognition is based on inclusion of the bulky hydrophobic group of the analyte into the hydrophobic cavity of the CD and on lateral interactions of the hydroxyl groups, such as hydrogen bonds and dipole–dipole interactions, with the analyte.	Carboxymethylated β-CD, heptakis-*O*-methyl-β-CD, hydroxyethyl-β-CD, mono(6-β-aminoethylamino-6-deoxy)-β-CD, and mono(6-amino-6-deoxy)-β-CD.	Acebutolol, acenocoumarol, carnitine, cathinone, ephedrine, epinephrine, glutethimide, ketotifen, thioridazine, etc.
Macrocyclic antibiotics	The multiple chiral atoms and several functional groups allow multiple interactions with the analytes to enable chiral recognition. The primary interaction is ionic; secondary interactions include hydrogen bonding, dipole–dipole, π–π, hydrophobic interactions, and steric repulsion.	Rifamycin B, rifamycin SV, ristocetin A, teicoplanin, fradiomycin, kanamycin, ansamycins, avoparcin, and vancomycin.	Amephoterin, α-aminoadipic acid, flurbiprofen, fenoprofen, methionine, methotrexate, ketoprofen, and suprofen, etc.
Proteins/ polypeptides	Proteins and peptides are naturally chiral and they often have different qualitative interactions (e.g., different binding sites) and/or quantitative interactions (i.e., different affinity or binding capacity) with the stereoisomers of chiral molecules. Very specific high-affinity binding often occurs, and any sort of intermolecular interactions (e.g., hydrophobic interactions, electrostatic interactions, etc.) may play a role in the binding, which is often reversible.	Bovine serum albumin, human serum albumin, pepsin, lysozyme, avidin, ovomucoid, casein, conalbumin, streptavidin, trypsin, ovoglycoprotein, and β-lactoglobulin.	Benzoin, bunitrolol, epinastine, flurbiprofen, ibuprofen, ketoprofen, leucovorin, pindolol, promethazine, propranolol, trimeputine, warfarin, etc.
Polysaccharides	Hydrogen bonds and dipole–dipole interactions with hydroxyl groups of the sugar molecules are assumed to be the main interactions. In some cases, the helical structure of dextrins might be responsible for chiral recognition.	Heparin, dextran sulfate, dermatan sulfate, streptomycin sulfate, amylose, chondroitin sulfate C, laminaran, dextrin sulfopropyl, and kanamycin sulfate.	Doxylamine, laudanosine, naproxen, oxamniquine, pheniramine, primaquine, timepidium, trimetoquinol, etc.
Chiral surfactants	The chiral separation of analytes is based on their partition coefficients between the chiral micelle phase and the electrolyte bulk phase.	Alkylglucosides, alkylmaltoside, sodium cholate, saponines, sodium dodecyl sulfate, sodium taurocholate.	Ephedrine, fenoldopam, hexobarbital, ketamine, pindolol, timolol, etc.
Chiral crown ethers	Two different diastereomeric inclusion complexes are formed. The primary interactions for complexation are hydrogen bonds between the three amine hydrogens and the oxygens of the macrocyclic ether in a tripod arrangement. Ionic, dipole–dipole interactions, or hydrogen bonds between the carboxylic groups and polar groups of the analytes may act as additional supporting interactions.	18-Crown-6-tetracarboxylic acid, dicyclohexyl-18-crown-6, 18-crown-6-2,3,11,12-tetracarboxylic acid, benzo-monoaza-15-crown-5, and (*R*,*R*)-2,12-bis(hydroxymethyl)-2, 12-dimethyl-18-crown-6.	Aminoglutethimide, aminophenol, baclofen, dopa, ephedrine, mexiletine, noradrenaline, norephedrine, octopamine, primaquine, etc.

(From Refs.[6–8].)

Applications of GC to chiral separation

The first separation of enantiomers was achieved by Gil-Av, Feibush, and Charles-Sigler[9] using capillary GC. Separation of enantiomers using CSPs involves hydrogen bonding, coordination, and inclusion. Typical chiral selectors include modified CDs, derivatized amino acids, and terpene-derived metal coordination compounds. The scope and limitations, applications, and mechanistic considerations of chiral separation by GC have been reviewed by Schurig and Francotte.[10,11]

The main applications of enantiomeric separation by GC concern precise determination of enantiomeric composition of chiral research chemicals, drugs, intermediates, metabolites, pesticides, flavors and fragrances, etc. CHIRBASE, a database of chiral compounds, provides comprehensive structural, experimental, and bibliographic information on successful and unsuccessful chiral separations, and rule sets for each CSP and information about the processes of chiral separations.[27] According to CHIRBASE, an appropriate CSP is available for almost every racemic mixture of compounds ranging form apolar to polar. Some 22,000 separations of enantiomers, involving 5,500 basic chiral compounds and documented in 2,200 publications, have been achieved by GC. This method is particularly suitable for volatile compounds such as inhalation anesthetic agents, e.g., enflurane, isoflurane, desflurane, and racemic α-ionone.

A particularly attractive feature of GC, one that distinguishes it from liquid chromatography methods, is the lack of a sensitive dependence on solvents, modifiers, and gradient elution systems. Prerequisites for the use of GC, however, are volatility, thermostability, and resolvability of the chiral analyte. Obviously, this restricts the utility of the method.

Applications of HPLC to chiral separation

Chromatographic methods have dominated separation of enantiomers during the past several decades, especially HPLC.[4,15–17] Numerous book chapters and review articles deal with the separation of chiral drugs by this method (e.g., Refs.[2,6,12–14].). Chiral HPLC is more versatile than chiral GC for enantiomeric separation because it can separate a wide variety of nonvolatile compounds. It can be used to develop fast and accurate methods of chiral drug separation, and it allows on-line detection and quantitation of both mass and optical rotation of enantiomers when appropriate detection devices are used.

Current chiral separation methods using liquid chromatographic techniques can be divided into two categories: a direct method based on diastereomer formation on CSPs or in mobile phases, and an indirect method based on diastereomer formation by reaction with a homochiral reagent. Direct chiral separation using CSPs is more widely used and predictable in mechanistic terms than methods involving chiral additives in the mobile phase. To date, more than a hundred HPLC CSPs are commercially available. No single CSP can be considered universal; none has the ability to separate all classes of racemic compounds.

Three components should be considered in developing a chiral separation method: analyte, CSP, and mobile phase. The key to successful chiral separation of a particular class of racemates on a given CSP is awareness of possible chiral recognition mechanisms. The enormous increase in recent years in number of groups working on chiral chromatography has led to a rapid and impressive accumulation of data in CHIRBASE.[27]

Some examples of chiral HPLC separations of racemic drugs are the following. Typical chromatograms of the simultaneous determination of isopyramide and its active metabolite, mono-*N*-dealkyldisopyramide, in drug-free human plasma, human plasma spiked with disopyramide and mono-*N*-dealkyldisopyramide, and treated subject plasma are presented in Fig. 3.[4] Chiral HPLC has been used to separate chlorpheniramine and its main monodesmethyl metabolite, verapamil and its metabolite, and tramadol and its metabolite.[15–17]

Applications of CE to chiral separation

CE has become widely popular for enantiomer separation over the past decade as a very powerful complementary or alternative technique to HPLC in pharmaceutical science and industry. Several chiral separation principles successfully applied in HPLC have been transferred to CE. The first chiral separation by CE was by Gassmann, Kuo, and Zare in 1985.[18] This approach offers key advantages such as high efficiency, feasibility of incorporating a large number of chiral selectors that greatly facilitates method development, small amounts of chiral selector and solvents, speed of analysis, low overall cost, and minimal environmental impact. The use of low-UV wavelength (e.g., 200 nm) in CE allows detection of impurities with poor chromophores, which may be difficult or impossible to detect by other methods. CE is suitable for charged and polar compounds for which chromatographic methods are not very strong.

Several comprehensive review articles have appeared in recent years dealing with general aspects and applications of chiral CE separation.[8,19–21] A comprehensive list of more than 280 drugs separate by chiral CE, including the respective chiral selectors and background electrolytes, appears in a review article by Gübitz and Schmid.[21] Chiral CE has also

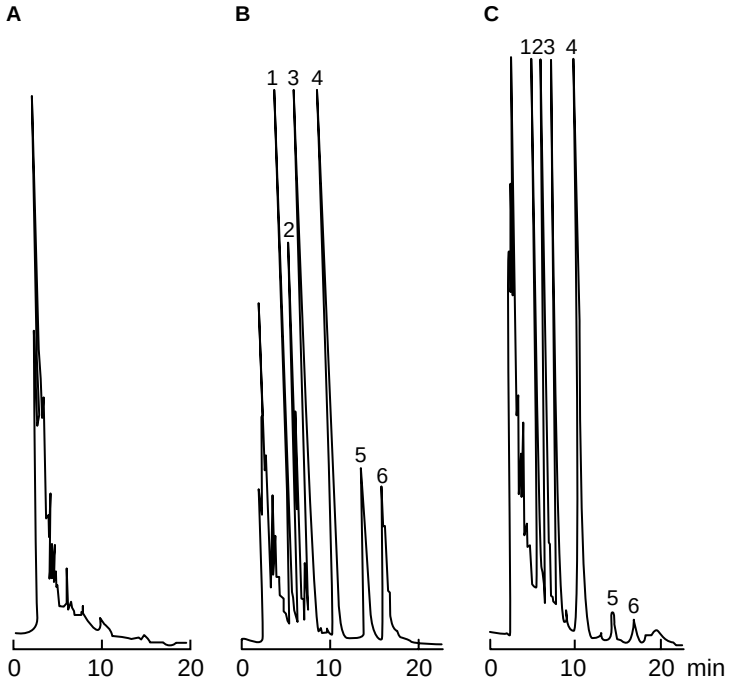


Fig. 3 Chromatograms showing analysis of disopyramide and mono-*N*-dealkyldisopyramide enantiomers in plasma: (A) blank plasma, (B) plasma spiked with 625 ng/ml of disopyramide and mono-*N*-dealkyldisopyramide enantiomers, (C) plasma sample from a healthy volunteer collected 6 hr after administration of 100 mg of Dicorantil. Peak assignments: 1, (*S*)-(+)-disopyramide; 2, (*R*)-(−)-disopyramide; 3, 4, metoprolol; 5, (*S*)-(+)-mono-*N*-dealkyldisopyramide; and 6, (*R*)-(−)-mono-*N*-dealkyldisopyramide. Chromatographic conditions: Chiralpak AD column (250 × 4.6 mm I.D., 10 μm particle size); hexane–ethanol (91 : 9, v/v) mobile phase plus 0.1% diethylamine; 1.2 ml/min flow rate; and detection at 270 nm. (From Ref.[4].)

been the theme of several special issues of *Electrophoresis* and *Journal of Chromatography A*.

Chiral separation in all CE techniques, including capillary zone electrophoresis (CZE), capillary gel electrophoresis (CGE), and capillary isoelectric focusing (CIEF), relies on enatioselective noncovalent/intermolecular interactions between the analyte and a chiral selector, which may be expressed as effective mobility difference (CZE and CGE), stereoselective shift of the acid–base equilibrium (CIEF), etc. Although the fundamental enantioselection mechanisms in CE are the same as in chromatography, migration is driven by electrophoresis. Enantiomers of a chiral drug compound will have the same charge density, so chiral separation in CE is not commonly based on electrophoretic separation, in which different migration velocities are an effect of different charge densities of analyte components. For enantiomers, both the electroosmotic flow and the electrophoretic mobility of the analyte are equally nonstereoselective. What distinguishes enantiomers in chiral CE is their interaction with a chiral selector.

Chiral separation by CE can be achieved by a direct or an indirect method. Direct separation is the more common approach. The chiral selector is dissolved in the running buffer, where it interacts selectively with the enantiomers to form reversible and transient diastereoisomeric or inclusion complexes of differing effective mobility. In indirect chiral CE separation, the enantiomers form covalent diastereomeric derivatives with a chiral reagent. A chiral selector is unnecessary in this approach because of the different electrophoretic mobilities of the diastereomeric derivatives.

Examples of chiral CE separations of racemic drugs are the following. (*R*)-(−)-ketoprofen has successfully been separated from ketoprofen and detected (Fig. 4).[25] (*S*)-(+)-ketoprofen is the active component. Also, simultaneous chiral separation of a basic drug compound, 2(*R*)-*N*-[1-(6-amino-pyridin-2-ylmethyl) piperidin-4-yl]-2-[(1*R*)-3,3-difluorocyclopentyl]-2-hydroxy-2-phenyl-acetamide, and its chiral acidic intermediate, (*R*,*R*)1-(2,2-difluorocyclopentyl)-phenylacetic acid, has been achieved by CE using a single-isomer CD, octakis (2,3-diacetyl-6-sulfo)-γ-CD (ODAS-γ-CD).[22] Carnitine has been separated using 50 mM DM-β-CD in 20 mM phosphate buffer (pH 4.3) as chiral selector.[23] The separations are done at 30°C in a fused-silica capillary, dynamically coated with triethanolamine present in the background electrolytes.

Similarities and differences in chiral separation by chromatographic and capillary electrophoretic techniques

HPLC remains the dominant technique for chiral separation in industry. CE has become well accepted in academic laboratories. Current GC, HPLC, and CE instruments are automated. Chiral separation in CE relies on a chromatographic separation principle. Nevertheless, there are significant differences, as shown in Table 3, between these techniques. The property of

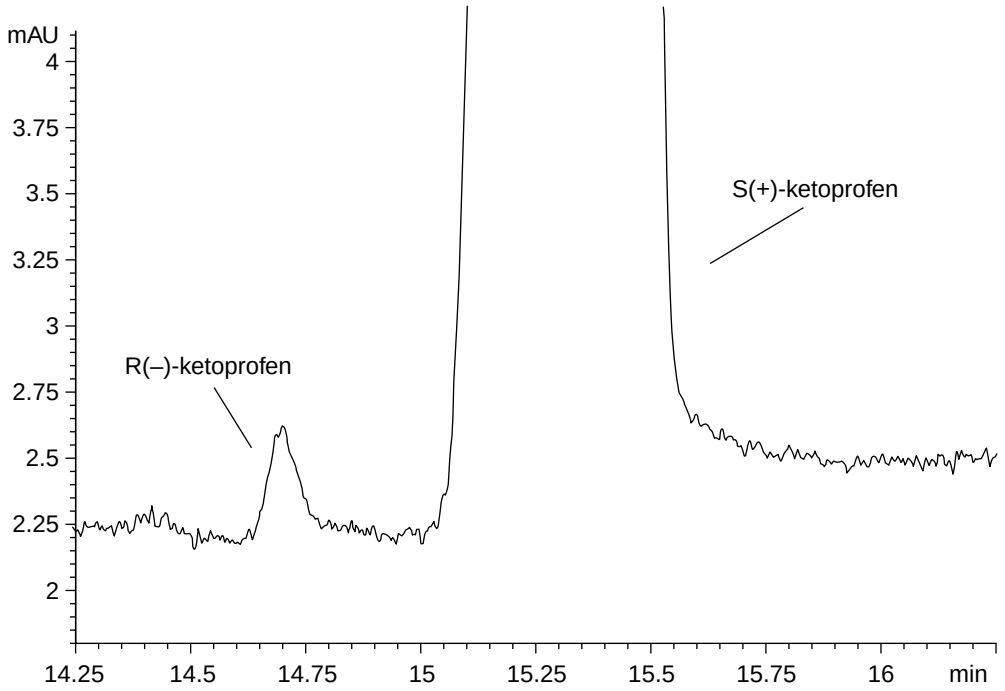


Fig. 4 Detection of the (*R*)-(−)-ketoprofen impurity contained in Enantyum tablets using chiral CE. Conditions: a 20 mM phosphate–20 mM triethanolamine background electrolyte at pH 5.0, with 50 mM of tri-*O*-methyl-β-CD; temperature, 35°C; applied voltage, 20 kV, 8.1 mA. (From Ref.[25].)

electrophoretic mobility in chiral CE in particular, and its ability to be selective for the analytes residing in the same phase, is responsible for all the differences. Another important point is that in chromatographic techniques, except with a chiral mobile phase additive, the analyte is virtually immobile when associated with a chiral selector. By contrast, in CE the analyte selector complex is commonly mobile.

In studies of a chiral drug candidate and its possible metabolites in preclinical and clinical Phase I to Phase III, most of the biological sample matrices, such as urine, plasma, serum, saliva, cerebrospinal fluid, and tissue homogenates, are more compatible with CE than chromatographic techniques. Moreover, it is not possible in GC and difficult in chiral HPLC to achieve the chiral separation of a drug and its phase I and phase II metabolites in a single run. By contrast, this is possible with chiral CE, as in the simultaneous chiral separation of Phase I and Phase II metabolites of chiral antihistaminic drug dimethindene.[24]

In summary, there are significant differences between chiral separations in pressure-driven HPLC and electrically driven CE systems. These differences are advantageous in that they make the techniques complementary. The rules and dependencies observed in one technique, however, are not necessarily applicable to the other.

Recently developed techniques for chiral separation: chiral membranes

HPLC is useful because of its preparative-scale capability. The method is generally slow and labor-intensive, however, requiring specialized engineering approaches for acceptable throughout. In comparison, chiral membrane separation offers significant advantages in simplicity, cost, and throughput.[28]

Polypeptides or modified polypeptides have been tested for use in chiral membrane separation. The partition behavior of optical isomers will largely be influenced by the structure and number of ''recognition sites'' in the polypeptide membrane. Polypeptides could be designed on demand and used to build polypeptide films/membranes.[29] Polypeptide membranes have shown very high enantiomer permeation rates with encouraging selectivity for chiral drug separation.[30] Peptide hydrophobicity, molecular weight, and secondary-structure formation propensity of specially designed peptides will be considered in future work on chiral separation membranes, as polypeptides could be designed to behave like certain proteins in human body and thus to achieve high biomimetic enantioselectivity. This approach would appear to promise very high enantioselectivity and high permeation rates.

Table 3 Differences between HPLC and CE as chiral separation techniques

	HPLC	CE
Instrument, cost, and safety	Expensive columns, consumption of a relatively high amount of buffer solutions, and hazardous organic modifiers.	Simple instrument: no pump, injector valves, and detector cells are required; a minute amount of solvent and an extremely low amount of chiral selector and buffer are required; environmentally friendly and inexpensive.
Chiral selector	Immobilized; great number of commercially available CSPs; a combination of chiral selectors is difficult or at least time-consuming.	Commonly mobile, commercially available chiral selectors are inexpensive, chiral selectors can be mixed in any desired ratio (only limited by the solubility).
Selectivity and efficiency	Chiral separation selectivity may in the best case approach the thermodynamic selectivity of the chiral recognition but will never exceed it; the separation efficiency sometimes is poor.	Separation selectivity may easily exceed the thermodynamic selectivity of the recognition; a chiral separation even in the absence of chiral recognition is, in principle, feasible; high peak efficiency.
Manipulation of mobility terms	Impossible to adjust the selectivity of chiral separation without changing the affinity pattern of the enantiomers toward chiral selector.	Possible to adjust the enantiomer migration order without reverting the affinity pattern between enantiomers of the analyte and a chiral selector.
Separation scale	Semipreparative and preparative scale.	Analytical scale, very small sample volumes.
Reproducibility	Good; high success rate.	Relatively poor and low success rate.
Method development	Relatively slow; changing and conditioning a column is time-consuming.	Rapid; changing a capillary and/or chiral selector takes only few minutes.

CONCLUSIONS

It is necessary to consider the chiral nature of a compound in drug research and development and the drug regulatory process. Enantiomers of all chiral bioactive molecules must be separated and tested. The FDA and regulatory authorities in Europe, China, and Japan have provided guidelines indicating that only the active enantiomer of a chiral drug should be brought to market. Chiral techniques have been developed for the separation and analysis of chiral drugs. Among these, HPLC based on CSPs is widely employed for the assays of drug enantiomers in pharmaceutical preparations and biological fluids. More recently, CE plus chiral selector additives to the running buffer have been used for the same purpose. In both chromatographic and electrophoretic methods, different types of chiral selectors, including CDs, crown ethers, quinine, chiral surfactants, polysaccharides, proteins, and macrocyclic antibiotics, have successfully been used for chiral separation of drugs and drug candidates. In many cases, HPLC and CE are complementary with respect to enantioselectivity. The development of novel separation techniques is an active area of research. Ones showing high throughput and high enantioselectivity are likely to be important to commerce, in view of the increasing trend in marketing single-enantiomer drugs.

ARTICLE OF FURTHER INTEREST

Chromatographic Separation, p. 481.

REFERENCES

1. Stinson, S.C. Chiral pharmaceuticals. Chem. Eng. News **2001**, *79* (40), 79–97.
2. Vermeulen, N.P.E.; Koppele, J.M. te. Stereoselective biotransformation: toxicological consequences and implications. In *Drug Stereochemistry: Analytical*

Methods and Pharmacology, 2nd Ed.; Wainer, W.I., Ed.; Marcel Dekker, Inc.: New York, 1993; 245–280.
3. Announcement: FDA's policy statement for the development of new stereoisomeric drugs. Chirality **1992**, *4* (5), 338–340.
4. Bortocan, R.; Lanchote, V.L.; Cesarino, E.J.; Bonato, P.S. Enantioselective analysis of disopyramide and mono-*N*-dealkyldisopyramide in plasma and urine by high-performance liquid chromatography on an amylose-derived chiral stationary phase. J. Chromatogr. B **2000**, *744*, 299–306.
5. Dalgliesh, C.E. The optical resolution of aromatic amino acids on paper chromatograms. J. Chem. Soc. **1952**, *137*, 3940–3942.
6. Gübitz, G.; Schmid, M.G. Chiral separation by chromatographic and electromigration techniques. A review. Biopharm. Drug Dispos. **2001**, *22*, 291–336.
7. Boer, T.; Zeeuw, R.A.; Jong, G.J.; Ensing, K. Recent innovations in the use of charged cyclodextrins in capillary electrophoresis for chiral separations in pharmaceutical analysis. Electrophoresis **2000**, *21*, 3220–3239.
8. Amini, A. Review: recent developments in chiral capillary electrophoresis and applications of this technique to pharmaceutical and biomedical analysis. Electrophoresis **2001**, *22*, 3107–3130.
9. Gil-Av, E.; Feibush, B.; Charles-Sigler, R. Separation of enantiomers by gas liquid chromatography with an optically active stationary phase. Tetrahedron Lett. **1966**, *7* (10), 1009–1015.
10. Schurig, V. Separation of enantiomers by gas chromatography. J. Chromatogr. A **2001**, *906*, 275–299.
11. Francotte, E. Chromatography as a separation tool for the preparative resolution of racemic compounds. In *Chiral Separations: Applications and Technology*; Ahuja, S., Ed.; American Chemical Society: New York, 1996; 271–308.
12. Allenmark, S.G. *Chromatographic Enantioseparation: Methods and Applications*, 2nd Ed.; Ellis Horwood Ltd: New York, 1991.
13. Taylor, D.R.; Maher, K. Chiral separations by high-performance liquid-chromatography. J. Chromatogr. Sci. **1992**, *30* (3), 67–85.
14. Haginaka, J. Review: pharmaceutical and biomedical applications of enantioseparations using liquid chromatographic techniques. J. Pharm. Biomed. Anal. **2002**, *27*, 357–372.
15. Hiep, B.T.; Khanh, V.; Hung, N.K.; Thuillier, A.; Gimenez, F. Determination of the enantiomers of chlorpheniramine and its main monodesmethyl metabolite in urine using achiral–chiral liquid chromatography. J. Chromatogr. B **1998**, *707*, 235–240.
16. Brandsteterová, E.; Wainer, I.W. Achiral and chiral high-performance liquid chromatography of verapamil and its metabolites in serum samples. J. Chromatogr. B **1999**, *732*, 395–404.
17. Campanero, M.A.; Calahorra, B.; Valle, M.; Troconiz, I.F.; Honorato, J. Enantiomeric separation of tramadol and its active metabolite in human plasma by chiral high-performance liquid chromatography: application to pharmacokinetic studies. Chirality **1999**, *11*, 272–279.
18. Gassmann, E.; Kuo, J.E.; Zare, R.N. Electrokinetic separation of chiral compounds. Science **1985**, *230*, 813–815.
19. Gübitz, G.; Schmid, M.G. Recent progress in chiral separation principles in capillary electrophoresis. Electrophoresis **2000**, *21*, 4112–4135.
20. Vespalec, R.; Boček, P. Chiral separations in capillary electrophoresis. Chem. Rev. **2000**, *100*, 3715–3753.
21. Gübitz, G.; Schmid, M.G. Review: chiral separation principles in capillary electrophoresis. J. Chromatogr. A **1997**, *792*, 179–225.
22. Zhou, L.; Thompson, R.; French, M.; Ellison, D.; Wyvratt, J. Simultaneous enantioseparation of a basic drug compound and its acidic intermediate by capillary electrophoresis. J. Sep. Sci. **2002**, *25*, 1183–1189.
23. Mardones, C.; Ríos, A.; Valcárcel, M.; Cicciarelli, R. Enantiomeric separation of D- and L-carnitine by integrating on-line derivatization with capillary zone electrophoresis. J. Chromatogr. A **1999**, *849*, 609–616.
24. Rudolf, J.; Blaschke, G. Stereoselectivity of the Phase-II-metabolism of the H-1 antihistaminic drug dimethindene investigated by capillary electrophoresis. Enantiomer **1999**, *4* (3–4), 317–323.
25. Blanco, M.; Coello, J.; Iturriaga, H.; Maspoch, S.; Pérez-Maseda, C. Chiral and nonchiral determination of ketoprofen in pharmaceuticals by capillary zone electrophoresis. J. Chromatogr. A **1998**, *799*, 301–307.
26. Shahani, S. *Opportunities in Chiral Technology*; Business Communications Co.: Norwalk, CT, 2004.
27. Piras, P.; Roussel, C.; Pierrot-Sanders, J. Reviewing mobile phases used on Chiralcel OD through an application of data mining tools to CHIRBASE database. J. Chromatogr. A **2001**, *906*, 443–458.
28. Warner, T.N.; Nochumson, S. A work in process: membrane-based chromatography is paving the way for throughput biopharmaceutical processing. Mod. Drug Discov. **2003**, *6*, 45–49.
29. Li, B.; Haynie, D.T. Multilayer biomimetics: reversible covalent stabilization of a nanostructured biofilm. Biomacromolecules **2004**, *5* (5), 1667–1670.
30. Rmaile, H.H.; Schlenoff, J.B. Optically active polyelectrolyte multilayers as membranes for chiral separations. J. Am. Chem. Soc. **2003**, *125*, 6602–6603.

Chlorofluorocarbons

Byung Gwon Lee
Youn-Woo Lee
Environment and Process Technology Division, Korea Institute of Science and Technology, Cheongryang, Seoul, South Korea

INTRODUCTION

The chlorofluorocarbons (CFCs) have been the most important organic fluorine chemicals during the past 70 years or so. They have unique properties: chemically stable, low in toxicity, nonflammable, noncorrosive, and compatible with other materials. In addition, the CFCs have thermodynamic and physical properties that make them ideal for a variety of uses. They have been used as refrigerants; as blowing agents in the manufacture of insulations, packaging, and cushioning foams; as cleaning agents for metal and electronic components; and in many other applications.[1–3]

Selected for development by Midgley and his coworkers in the late 1920s, CFC-12 (CF_2Cl_2) was considered to be the ideal refrigerant to replace sulfur dioxide and ammonia.[4] The outstanding properties of CFC-12 quickly led to the search for other CFCs. Kinetic Chemicals, a joint venture between Du Pont and General Motors, was formed to manufacture the new materials, and later it became part of the Freon Division of Du Pont.[3]

The introduction of CFCs marked the inception of the fluorocarbon industry. In terms of both tonnage produced and product value, CFCs have dominated the fluorocarbon industry and provided feedstocks for the development of other products, such as fluoropolymers. By the early 1970s, when they were first linked to the destruction of the ozone layer, CFCs were being produced in many countries around the world. Some major producers and their trade names are shown in Table 1.

The possible implication of CFCs in the depletion of stratospheric ozone, postulated in 1974 and reinforced in the late 1980s by the discovery of a possible link to the thinning of the ozone layer over Antarctica during springtime, had profound effects on the fluorocarbon industry.[5,6] The discovery of the ozone-depleting properties of CFCs led to the Montreal Protocol and the London and Copenhagen amendments (1990, 1992), which scheduled the end of production of CFCs by the end of 1995.

PROPERTIES

The CFCs usually have high densities and low boiling points, viscosities, and surface tensions. The irregularity in the boiling points of the fluorinated methanes and ethanes, however, does not appear in the chlorofluorocarbons. Their boiling points consistently increase with the number of chlorines present. The physical and environmental data of some CFCs are summarized in Table 2.

Although the CFCs are not as stable as the fluorocarbons (FCs), they are unusually stable compounds. Dichlorodifluoromethane, CCl_2F_2 (CFC-12), is stable at 500°C in quartz. Trichloromonofluoromethane, CCl_3F (CFC-11), begins to decompose at 450°C. The CFCs react with molten alkali metals and CCl_2F_2 reacts vigorously with molten aluminum, but with most metals they do not react below 200°C. An exception is the dechlorination of chlorofluorocarbons with two or more carbon atoms in the presence of Zn, Mg, or Al in polar solvents.[1,2]

Most CFCs are hydrolytically stable, CCl_2F_2 being considerably more stable than either CCl_3F or $CHCl_2F$. Chlorofluoromethanes and chlorofluoroethanes disproportionate in the presence of aluminum chloride. For example, CCl_3F and CCl_2F_2 give $CClF_3$ and CCl_4; $CHClF_2$ disproportionates to CHF_3 and $CHCl_3$.

The carbon–chlorine bond in most chlorofluorocarbons can be homolytically cleaved under photolytic conditions (185–225 nm) to give chlorine radicals. This photochemical decomposition is the basis of the prediction that chlorofluorocarbons that reach the upper atmosphere deplete the earth's ozone shield.

PRODUCTION

Industrial production of CFCs is achieved by successive replacement of chlorine in chlorocarbons by fluorine, using anhydrous hydrogen fluoride and a catalyst.[3,7] This use consumes about 50% of the hydrogen fluoride manufactured in the world. The extent of chlorine exchange can be controlled by

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007703

Table 1 Trademarks and manufacturers of CFCs

Country	Trademark	Manufacturer
Australia	Isceon	Australian Fluorine Chemicals
Czechoslovakia	Ledon	Slovek Pro Chemickov
France	Flugene Forane	Pechiney-Saint Gobain S.A. Ugine
Germany	Frigedohn Frigen Fluogen Fridohna	Fluorwerke Dohna Farbwerke Hoechst Chemische Fabrik von Heyden VEB Chemiewerk Nunchritz
Italy	Algofrene	Montecatini
Japan	Asahiflon Daiflon Flon	Ashai Glass Daikin Kogyo Mitsui Fluorochemicals
Korea	Korfron	Ulsan Chemicals
Netherlands	Fcc	Akzo
United Kingdom	Arcton Isecon	Imperial Chemical Industries Imperial Smelting
United States of America	Freon Genetron Isotron Ucon	E. I. Du Pont de Nemours & Co., Inc Allied Chemical Corporation Pennwalt Corporation Union Carbide Corporation

varying the hydrogen fluoride concentration, the contact time, or the reaction temperature. Carbon tetrachloride and hexachloroethane (or tetrachloroethylene plus chlorine) are commonly used starting materials for one- and two-carbon CFCs. These feedstocks are prepared by direct chlorination of hydrocarbons.

Historically, most of the CFCs have been made by a liquid-phase process, employing antimony pentachloride as a catalyst. The antimony may be added either as the trichloride and converted to the pentachloride in situ with chlorine or, more conveniently, as the liquid pentachloride itself. Antimony pentachloride reacts with hydrogen fluoride to give mixed chlorofluorides, which are the actual catalysts, and hydrogen chloride [reaction (1)]. The exchange reaction then takes place between the chlorocarbon feedstock and the mixed chlorofluoride, producing the desired CFC and regenerating antimony pentachloride [reaction (2)]. Carbon tetrachloride serves as the feedstock for CFC-11, -12, and -13 [reaction (3)]. The chlorines are exchanged stepwise, the reaction temperature determining the number of fluorines introduced. In commercial reactors, temperatures usually range from 100°C to 150°C and pressures from 10 to 30 atmospheres (atm). The hydrogen chloride is conveniently removed as a liquefied gas for other use or absorbed in water for sale as hydrochloric acid. The crude CFC is washed with base to remove traces of hydrogen chloride and any hydrogen fluoride, dried, and purified by distillation.

$$SbCl_5 + nHF \longrightarrow SbF_nCl_{5-n} + nHCl \quad (1)$$

$$nRCl + SbF_nCl_{5-n} \longrightarrow nRF + SbCl_5 \quad (2)$$

$$CCl_4 \xrightarrow{HF} CCl_3F + CCl_2F_2 + CClF_3 \quad (3)$$

A logical feedstock for the two-carbon CFCs is hexachloroethane. Since it is a high-melting solid, it is generated in situ from tetrachloroethylene and chlorine in the fluorination process. The antimony pentahalides catalyze the addition of chlorine in the fluorination process and may be reduced to the trihalide before subsequent reoxidation with more chlorine. If the feed ratios are not in balance, the excess of hydrogen fluoride may add across the double bond leading to hydrogen-containing compounds. The reaction for the synthesis of CFC-111, -112, -113, and -114 is shown in reaction (4). Again, the temperature of the reaction usually determines the number of chlorines exchanged. While isomers are possible, the carbon atom that is fluorinated is usually the one bearing the most chlorine atoms.

$$CCl_3CCl_3 \xrightarrow{HF} CCl_2FCCl_3 + CCl_2FCCl_2F + CClF_2CCl_2F + CClF_2CClF_2 \quad (4)$$

Table 2 Physical and enivironmental data for some CFCs

Property	CCl_3F	CCl_2F_2	$CClF_3$	CCl_2FCCl_2F	CCl_2FCClF_2	$CClF_2CClF_2$	$CClF_2CF_3$
CFC number	11	12	13	112	113	114	115
Molecular weight	137.4	120.9	104.5	203.8	187.4	170.9	154.5
Boiling point (°C)	23.8	−29.8	−81.4	92.8	47.6	3.8	−39.1
Melting point (°C)	−111	−158	−181	26	−35	−94	−106
Crit. temp (°C)	198.0	112.0	28.9	278.0	214.1	145.7	80.0
Crit. pressure (MPa)	4.4	4.1	3.9	3.4	3.4	3.3	3.1
Liquid density at 25°C (g/ml)	1.476	1.311	1.298_{-30}	1.634_{30}	1.565	1.456	1.291
Heat capacity (J/(kg · K))							
Liq. (25°C)	0.870	0.971	1.033_{-30}	—	0.912	1.016	1.192
Vap. (25°C, 101.3 kPa,)	0.565	0.607	0.661	—	0.674_{60}	0.711	0.686
Latent heat of vaporization (bp) (kJ/kg)	180.3	165.1	148.4	—	146.7	136.0	126.0
Viscosity (cP)							
Liq. (25°C)	0.430	0.214	0.195_{-45}	—	0.680	0.360	0.193
Vap. (25°C, 101.3 kPa)	0.011	0.023	0.015	—	$0.010_{10\ kPa}$	0.012	0.013
Relative dielectric strength ($N_2 = 1$)	3.71	2.46	1.65	—	$3.90_{44\ kPa}$	3.34	2.54
Dielectric constant							
Liq. (25°C)	2.28_{29}	2.13_{29}		2.54	2.41	2.26	—
Vap. (50.65 kPa)	1.0036_{24}	1.0032_{26}	1.0024_{29}	—	—	$1.0043_{26.8}$	1.0035_{27}
Surface tension at 25°C, (dyn/cm)	18	9	14_{-73}	23_{30}	17.3	12	5
Solubility in water at 25°C and 101.3 kPa (wt.%)	0.110	0.028	0.009	—	—	0.013	0.006
Ozone depletion potential (CFC-11 = 1.0)	1.000	0.900	1.000	—	0.900	0.850	0.400
Global warming potential ($CO_2 = 1.0$, 100 yr)	4000	8500	11700	—	5000	9200	9300

Antimony-based catalysts may become deactivated with use. Simple reduction to the trihalide is easily remedied by oxidation with chlorine. Over time, however, tar formation limits the activity of the catalyst and it needs either purification or removal as waste. Owing to environmental concerns over the disposal of toxic compounds, there is considerable current interest in the detoxification and recovery of antimony waste.[8,9]

This process is being replaced later by a continuous vapor-phase process that employs gaseous hydrogen fluoride in the presence of catalysts. Vapor-phase exchange processes have been developed that avoid the antimony disposal problem. In these processes, vaporized mixtures of the chlorocarbon feedstock and an excess of anhydrous hydrogen fluoride vapors are passed over catalysts at temperatures ranging from 250°C to 400°C. Once again the exchange reaction occurs stepwise, increasing catalyst bed temperatures yielding increased fluorination. The most common catalyst is trivalent chromium, either as the fluoride or as the oxide.[10] The catalyst may be used on a support such as alumina. Other catalysts include iron and fluorinated alumina. Research activity related to the search for CFC substitutes has produced claims of improvements in chromia catalysts, as well as those based on other metals such as cobalt, nickel, manganese, and lanthanum.[11–14]

Chromia (Cr_2O_3) catalysts and, to various extents, the other vapor-phase catalysts, tend to produce some rearrangement of the halogens. This scrambling may be troublesome in the two-carbon series when, for example, some CF_3CCl_3 is produced along with $CF_2ClCFCl_2$ One particular process achieves the halorination of the in situ generated chlorocarbon simultaneously. The catalyst is fluorinated alumina impregnated with thorium fluoride.

Just as with the liquid-phase process, the catalyst requires regeneration from time to time. The usual problem is ''coking,'' which requires treatment of the catalyst bed with oxygen or air at high temperatures. There have been some successful attempts to prolong

catalyst life by adding oxygen along with the feedstocks to remove the coke continually. Reactivation of alumina-based catalysts with high-temperature steam has also been reported.

Vapor-phase fluorinations are very attractive industrially. Reactions can be conducted continuously at modest pressures (10 atm), with appropriate recycle streams as necessary. However, since an excess of hydrogen fluoride is usually employed, downstream purification involving CFC–HF azeotropes may cause difficulties.

While high-molecular-weight CFCs can be synthesized via a variety of methods, from a practical standpoint such materials are best prepared by free-radical telomerization of chlorotrifluoroethylene (CTFE). Either the conversion is carried out in the presence of a chain-transfer agent or higher-molecular-weight polymers are cracked to lighter species. Carbon tetrachloride [reaction (5)] and CFC-113 [reaction (6)] have both been used as telogens. The telomer chains are formed predominantly, but not exclusively, via radical attack on the CF_2 carbon. Trichloromethyl (CCl_3) end groups must be fluorinated, at least partially, to make useful oils or grease since these groups are too reactive for an inert oil. Whichever way the oils are made, they must be free from unsaturation, since oxygen tends to react with C=C sites leading to the formation of acid fluorides and hence unusable material. Saturation and end-group fluorination are generally accomplished using a combination of chlorine and chlorine trifluoride or cobalt trifluoride. Once stabilized, the telomers are fractionated into various standard viscosities for sale as oils. Higher-molecular-weight material forms the basis of the waxes, while grease is made by the addition of suitable thickening agents.

$$CCl_4 + nCF_2{=}CFCl \longrightarrow CCl_3(CF_2CFCl)_nCl \qquad (5)$$

$$CF_2ClCFCl_2 + nCF_2{=}CFCl \longrightarrow CF_2ClCFCl(CF_2CFCl)_nCl \qquad (6)$$

Usually there is an overabundance of lower-molecular-weight telomers, which are not as useful as the heavier products; thus, coupling processes have been developed in which two units are dimerized through the removal of end-group chlorine. One such coupling reaction is achieved by heating telomers with a mixture of copper and zinc powders at 220°C [reaction (7)]. While coupling of CCl_3 end groups is the easiest, it produces CCl_2–CCl_2 units in the middle of the chains, which are difficult to fluorinate. If any remains in the product, stability is adversely affected.

$$2CF_2ClCFClCF_2CFCl_2 \longrightarrow (CF_2ClCFClCF_2CFCl)_2 \qquad (7)$$

APPLICATIONS

It is impossible to overstate the importance of the role that CFCs have played in making refrigeration, air-conditioning, and, to some extent, aerosols such essential facets of modern existence. Life, as we know it, would be impossible in some parts of the world without the means for cooling food and houses. Of course, the very property that made the CFCs so useful, their stability, allows them to reach the stratosphere and there release chlorine atoms, which interact with the ozone layer. There will be a drastic decline in the use of CFCs in most of their high-volume applications of the heavier CFCs and the volumes will be inconsequential compared to the 2.5 billion pounds of volatile CFCs produced worldwide in 1986.

Refrigeration

This has been the area where CFCs have excelled in their application. The air-conditioning industry owes much of its tremendous growth to the chlorofluorocarbons. The CFCs were judged to be safe alternatives to materials such as sulfur dioxide and ammonia over 70 yr ago and allowed refrigeration to become available safely in houses and automobiles.

CFC-12 was commonly used in domestic and mobile air-conditioning. It normally achieved temperatures down to −20°C, and was suitable for use with centrifugal, rotary, and reciprocating compressors. CFC-11 was used in industrial chillers and other units when a temperature of 5–10°C was adequate. For applications where colder temperatures are required, azeotropic blends involving CFC-115 (CF_3CF_2Cl) and HCFC-22 ($CHClF_2$) are employed.

Aerosols and Propellants

Another important commercial application of CFCs has been in the aerosol field as a propellant for a wide variety of products. Personal products such as antiperspirants and hair sprays account for ca. 85% of this use.[1] The characteristic stability, nonflammability, low toxicity, and lack of odor provide a unique combination of properties desirable for propellants.

The most common CFC propellants are CFC-12 (CCl_2F_2), CFC-11 (CCl_3F), and CFC-114 ($CClF_2CClF_2$), used either by themselves or in combination. The principal propellant for nonfood aerosols is $CC1_2F_2$. A combination of $CC1_3F$ and $CC1_2F_2$ is used to reduce the propellant pressure and to increase the compatibility of other compounds in the aerosol formulation. Approximately 60% of the total manufactured CCl_2F_2 and CCl_3F is used in aerosols.

1,2-Dichlorotetrafluoroethane is used primarily in the cosmetic field, often in combination with CCl_2F_2. As CFCs have begun to be replaced, there has been an increase in the use of hydrocarbons, sometimes mixed with HCFC-22 (CHF_2Cl) to lower the flammability.

Foams

Polymers in a foamed or cellular state have many useful properties. Energy-efficient buildings have sheets (or sprayed foams) of polystyrene foam insulation in their construction, sheets made by compounding CFC-12 under pressure with liquid polymer followed by release to atmospheric pressure via extrusion. Polyurethane foams were produced using CFC-11, which was vaporized by the heat of reaction of the isocyanate with the polyol. The foaming mass was usually heat-treated to complete the process.

When employed as a foam to fill cavities, however, CFC-12 was used to expand the polymer mass. Although some CFC, especially CFC-11, remained behind in the foam, almost all of the material was lost to the atmosphere during foaming operations. This was one application, therefore, where recycling was not feasible. Approximately 30% of the CFC-11 was produced as a blowing agent for rigid and flexible polyurethane foams. The CFC trapped in the rigid foam significantly increased its insulation properties.

Minor quantities of rigid polyurethane foam are manufactured by a frothing process using CCl_2FCClF_2 and $CClF_2CClF_2$. Extruded polystyrene foams have used CCl_2F_2 as the principal blowing agent. Both $CClF_2CClF_2$ and $CClF_2CClF_2$ are used in low-density polyethylene and polystyrene foams.

Chlorofluorocarbons were completely removed from plastic foam products for food packaging, such as egg cartons and meat trays in the early 1990s; currently, HCFC-22 mixed with carbon dioxide, HFCs, supercritical fluids, pentane, or isopentane is being used as an alternative. Dow Chemical has reported that it uses HCFC-142b in its polystyrene and polyethylene fabricated foams.[1]

Solvents

CFC-113 ($CF_2ClCFCl_2$) was an ideal solvent and cleaning agent for the electronics and aerospace industries.[15] It had excellent wetting properties, was noncorrosive, and was easily removed, leaving no residue. Oil and greases were dissolved without affecting most plastic, elastomeric, or metal components. For applications where greater solvent power was needed, it might be mixed with alcohols and ketones, usually at an azeotropic composition. Solvents based on CFC-113 were used to remove flux from printed wiring boards, and as degreasers for precision and polished surfaces where no residue could be tolerated. CFC-113 was also used in the dry-cleaning of expensive, heat-sensitive fabrics. CFC-11 has also been used in dry-cleaning because of its better solvency. Other minor applications of chlorofluorocarbons included their use as dielectric fluids and in wind tunnels and bubble chambers.

Lubricants

The chlorotrifluoroethylene (CTFE) telomer oils were originally developed as lubricants for the mechanical equipment needed in the separation of uranium isotopes using the extremely reactive uranium hexafluoride. These inert oils, greases, and waxes now find many applications in aggressive environments where a hydrocarbon oil will not survive. Like hydrocarbon oils, they are sold as blend according to certain viscosities. Greases are made by mixing an appropriate oil with a thickener such as silica or higher-molecular-weight CTFE polymer.

Chlorotrifluoroethylene telomer lubricants are used primarily where chemical inertness and nonflammability are required. The chemical industry and the cryogenic gas industry (primarily oxygen) are the major users of these materials. They are used to lubricate all types of process equipment, such as dryers, conveyers, pumps, valves, and compressor seals. Extreme pressure tests using the four-ball method show that they are good lubricants, without any seizure even at an applied load of 800 kg.

In the aerospace industry, as well as finding use in the oxygen-delivery system to the space shuttle oxidizer tanks, a light CTFE oil is under development as a nonflammable hydraulic fluid for future aircraft. Since the CTFE oil is much heavier than presently used hydraulic oils, retrofitting of existing aircraft is not feasible. The CTFE oil operates in smaller lines at much higher (8000 psi) pressure. New elastomeric seals, which are compatible with the oil, have been developed. In modern aircraft, serious fires can occur during landing if a hydraulic hose ruptures and sprays fluid onto the hot braking system. Tests done with CTFE hydraulic fluid show the material is completely nonflammable during such an event.

Chlorotrifluoroethylene telomers find use as vacuum pump oils in the electronics industry, where reactive gases and aluminum chloride would be harmful to hydrocarbon oils. They are also used as instrument fill fluids where they substitute for glycerice or silicone fluids in pressure gauge sensors and diaphragm seals in aggressive chemical service. Other uses include application as a cutting or drawing oil in tantalum, molybdenum, and niobium processing, as

a lubricant for life-support systems in oxygen-rich atmospheres, and as a cold-temperature batch fluid in laboratory apparatus.

Chlorotrifluoroethylene oils are used as flotation fluids in gyroscopes. Specific density and viscosity values are achieved by blending together oils of different weights. Navigational devices containing these oils are found in many commercial and military aircraft, as well as missiles. Newer technologies such as fiber optics are now increasingly used, but the floated gyroscopes have a well-deserved reputation for accuracy and a long field life.

Chemical Intermediates

Chlorofluorocarbons function as chemical intermediates, provided they are consumed rather than released into the atmosphere. CFC-113 serves as the starting material for the production of CTFE monomer. Despite the many years of research into a catalytic vapor-phase process for this conversion, the preferred current method still involves zinc dechlorination in methanol [reaction (8)].[16,17]

$$CF_2ClCFCl_2 + Zn \longrightarrow CF_2{=}CFCl + ZnCl_2 \quad (8)$$

Chlorotrifluoroethylene monomer serves as a building block for the CTFE telomer oils as well as the solid higher polymer and various copolymers. CFC-113 also may be used in the production of trifluoroethylene monomer by vapor-phase reduction using hydrogen and a precious metal catalyst, usually palladium [reaction (9)]. Copolymers of trifluoroethylene and vinylidene fluoride show interesting piezoelectric properties.

$$CF_2ClCFCl_2 + H_2 \longrightarrow CF_2{=}CHF + 3HCl \quad (9)$$

CFC-113 may also serve as the starting material for some of the CFC alternatives. Conversion to CFC-113a [reaction (10)] may be accomplished using aluminum chloride as a catalyst, followed by reduction with sodium sulfite under basic conditions to give HCFC-123 [reaction (11)]. Alternatively, CFC-113a may be reduced with isopropanol in the presence of UV light [reaction (12)]. Both these processes produce HCFC-123 cleanly, with no over-reduction, and essentially free from HCFC-123a ($CF_2ClCHFCl$), which is usually found in several percent yield when HCFC-123 is made by the direct fluorination of tetrachloroethylene. The presence of the less stable HCFC-123a isomer may be problematic in some projected uses of HCFC-123.

$$CF_2ClCFCl_2 \longrightarrow CF_3CCl_3 \quad (10)$$

$$CF_3CCl_3 + Na_2SO_3 + NaOH \longrightarrow CF_3CHCl_2 + NaCl + Na_2SO_4 \quad (11)$$

$$CF_3CCl_3 + (CH_3)_2CHOH \longrightarrow CF_3CHCl_2 + HCl + (CH_3)_2C{=}O \quad (12)$$

CFC-113a can be fluorinated to CFC-114a and CFC-115 using vapor-phase techniques [reaction (13)]. Vapor-phase reduction of CFC-114a [reaction (14)] gives HCFC-124 and HFC-134a, both of which are under development as CFC alternatives. A similar reduction of CFC-115 gives its possible replacement HFC-125 [reaction (15)].

$$CF_3CCl_3 + HF \longrightarrow CF_3CFCl_2 + CF_3CF_2Cl \quad (13)$$

$$CF_3CFCl_2 + H_2 \longrightarrow CF_3CHClF + CF_3CH_2F \quad (14)$$

$$CF_3CF_2Cl + H_2 \longrightarrow CF_3CHF_2 + HCl \quad (15)$$

REGULATION OF CFCs PRODUCTION

Depletion of Ozone Layer

In the 1970s Lovelock and coworkers, using his newly developed electron capture detector, demonstrated that CFCs 11 and 12 were trace constituents in the atmosphere.[18] In 1972, at a meeting initiated by Du Pont, representatives of the leading CFC manufacturers discussed the environmental fate of their products. Ray McCarthy succinctly summarized their opinion in the following way: ''Fluorocarbons are intentionally or accidentally vented to the atmosphere world-wide at a rate approaching one billion pounds per year. These compounds may be either accumulating in the atmosphere or returning to the surface, land or sea, in pure form or as decomposition products. Under any of these alternatives it is prudent that we investigate any effects which the compounds may produce on plants or animals now or in the future.''

As a result, they initiated the industry research program to investigate the environmental impact of CFCs through an organization called the ''Fluorocarbon Program Panel'' (FPP) in 1972, 2 yr before publication of the Rowland–Molina hypothesis implicating chlorofluorocarbons in stratospheric ozone depletion.[5] This organization and its variously named successors sponsored much of the fundamental research into

atmospheric chemistry through the 1970s. Although the chemical industry is often depicted as not caring about environmental concerns, in the case of the CFCs the fluorocarbon producers adopted a demonstrably responsible attitude. Through FPP they jointly funded an extensive and high-quality atmospheric research program to test the validity of the Rowland and Molina hypothesis using independent experts who were free to publish their results as peer reviewed papers in reputable journals. Simultaneously, each company initiated its own research program to identify CFC replacements that retained the advantages of CFCs, notably low toxicity and nonflammability, while avoiding their implied threat to the ozone layer.

The atmospheric research program confirmed that CFCs were likely to deplete stratospheric ozone, as predicted by Rowland and Molina. The high stability of the chlorofluorocarbons, which made them so attractive as refrigerants, etc., was thought to enable themselves to pass unchanged through the troposphere to the stratosphere; once there, it was pointed out that the intense UV radiation from the sun could cause the carbon–chlorine bonds to break homolytically, giving chlorine atoms (e.g., $CF_2C1_2 + UV \rightarrow Cl^{\bullet} + {}^{\bullet}CF_2Cl$), which were known from laboratory experiments to catalyze the destruction of ozone: [$Cl^{\bullet} + O_3 \rightarrow ClO^{\bullet} + O_2$; $ClO^{\bullet} + O \rightarrow Cl^{\bullet} + O_2$ (net: $O_3 + O \rightarrow 2O_2$)]. The trace gas ozone, which is present throughout Earth's atmosphere, reaches its peak concentration within the stratosphere. The ozone layer plays an important role for life on Earth by absorbing biologically harmful UV radiation in the shorter-wavelength regions below 320 nm, i.e., in the UV-C (<290 nm) and UV-B (290–320 nm) regions. The stratospheric chlorine catalyzes the reversal of the photochemical generation of ozone (trioxygen, O_3) from the more familiar respirable form of oxygen (dioxygen, O_2): $O_2 + UV$ (<240 nm) $\rightarrow O_2$, $2O$; $O + O_2 \rightarrow O_3$; $O_3 + UV$ (210–320 nm) $\rightarrow O + O_2$. One million ozone molecules may be decomposed by just 10 chlorine atoms before the latter are eliminated, mainly through encounters with natural methane ($Cl^{\bullet} + CH_4 \rightarrow HCI + {}^{\bullet}CH_3$). The hydrogen chloride produced ultimately precipitates as hydrochloric acid in rain (''rains out'') in quantities negligible compared with natural background concentrations.

The observational techniques in the late 1970s were thought to be incapable of detecting any depletion that had already occurred. It was concluded that CFCs should be phased out, but that this could occur over a sufficiently long period to minimize the economic impact of the change to CFC users. Nevertheless, several developed countries including the United States pursued this changeover more proactively by banning the use of CFCs in most aerosols and by 1980 this had significantly reduced world CFC production.

Replacements for CFCs

Initial evidence of the effect of chlorofluorocarbons on the ozone layer was very limited and based mainly on computer models of the atmosphere. In the 1970s there was no observational support for ozone depletion, but the refrigeration manufacturers responded positively to the challenge. By 1976 most companies had initiated substantial research programs to identify environmentally acceptable replacements for CFCs. Alternatives to the CFCs need to retain the attractive properties of CFCs, but avoid any adverse environmental impact. Low toxicity, nonflammability, good thermodynamic properties, and accessibility via economically viable manufacturing routes were key factors in the selection. The inclusion of one or more hydrogen atoms in their molecules results in products being largely destructible in the lower atmosphere by naturally occurring hydroxyl radicals, ensuring that relatively little of the material survives to enter the stratosphere.

In the first stage of the research, producers were primarily seeking alternatives for the major CFC refrigerants and foam blowing agents CFC-12 and CFC-11, which were also the main aerosol propellants. This search was crucially influenced by the major customer: the refrigeration and air-conditioning industry. Between 1931 and 1974 the Industry owed its rapid development to the availability of the CFCs, designing its equipment around their specific thermodynamic properties. By careful, evolutionary development the industry had acquired an enviable reputation for reliability. Not surprisingly, the refrigerant manufacturers were expected to develop substitutes whose physical properties were similar to those of the CFCs fluids that were being replaced. In particular, similar vapor pressures were needed, since pressure was a key factor in refrigerator and air-conditioning design. Essentially, this meant searching for fluids with similar boiling points to those of CFC-l2 and CFC-11, which also met the other criteria outlined above.

By 1976, HFC-l34a had emerged from the research program as the fluid closest in physical properties to CFC-l2 and therefore the most acceptable to the refrigeration and air-conditioning industry. Its isomer HFC-134 was not favored, however, because of its significantly lower vapor pressure. HFC-l52a was rejected, primarily because of its flammability. The search for a CFC-11 replacement proved more difficult; but by the end of the 1970s, HCFC-123 and HCFC-141b had been identified as potential substitutes. Other candidates, such as HCFC-l33a, HCFC-31, and HCFC-21, had failed on toxicity grounds. To hasten the development program, manufacturers shared data about the toxicity of candidate compounds to avoid needless expense and time-consuming duplication.

In parallel with their own research programs, the manufacturers, through the FPP also jointly funded research to study the atmospheric chemistry of CFCs in order to assess the extent of any risk they might pose. Independent research workers of universities and research institutes worldwide were contracted to measure the rates of reactions, which were essential input data for the complex computer models needed to predict the rate of ozone depletion. This value could not be measured directly in the 1970s because the large daily and seasonal fluctuations in stratospheric ozone concentrations swamped the modest depletion expected from CFCs.

By 1980, the consensus of informed scientific opinion based on the best available evidence was that, although CFCs could cause depletion of stratospheric ozone, the overall effect at the production levels would be less then 3%. This suggested that action to replace CFCs would be required, but this could be taken over a sufficiently long period to minimize the cost to the refrigeration industry and ultimately to the customers. As a follow-up to the 1978 ban in the United States on the use of CFCs as aerosol propellants, the U.S. Environmental Protection Agency undertook in 1980 to promote legislation restricting the manufacture and other uses of CFCs. However, the perceived low rate of ozone depletion coupled with changing U.S. political priorities removed the urgency to enact the appropriate legislation.

Discovery of Antarctic Ozone Depletion

In 1984, a remarkable and totally unpredicted phenomenon was discovered by the British Antarctic Survey, the so-called ''ozone hole.''[2] The discovery of ozone depletion over Antarctica during the spring period provided the first observational support for the possible effect of CFCs on the stratospheric ozone. However, the observation of ozone loss did not indicate its cause. From 1984 to 1988, several theories were postulated from CFC chemistry to atmospheric dynamics or even to cosmic electron fluxes. It was not until 1988, with the results of the 1987 Airborne Antarctic Ozone Expedition, that a probable link with CFCs was established. This prompted the Natural Resources Defense Council, an American pressure group, to sue the EPA to fulfill its 1980 promise to seek legislation to further control the manufacture and use of CFCs in the United States of America.

Industrial research programs to develop CFC substitutes were renewed with vigor, and the refrigerant manufacturers set up in 1988 the jointly funded ''Program for Alternative Fluorocarbon Toxicity Testing'' (PAFT) to initiate and manage the toxicity testing of the preferred candidates. The refrigerant manufacturing industry also renewed its research into the atmospheric aspects of fluorocarbon fluids by forming in 1989 and funding an organization called ''Alternative Fluorocarbons Environmental Acceptability Study'' (AFEAS). Continuing the work started in the 1970s, this organization contracted independent research scientists to develop the understanding of the complex interactions of chemical species in the atmosphere.

An important outcome of this work has been the formulation of the ozone depletion potentials (ODPs) for chlorine-containing fluids. Arbitrarily. CFC-11 is assigned an ODP of unity; the ODPs of other fluids are normalized to that of CFC-11 on a mass-for-mass basis. Important factors in determining the ODP of a fluid include its atmospheric lifetime and the quantity of chlorine it contains. Hydrochlorofluorocarbons (HCFCs) have much lower ODPs than CFCs—a consequence mainly of the former's lower atmospheric lifetimes. Hydrofluorocarbons (HFCs), which by definition contain no chlorine, have zero ODPs.

The Montreal Protocol

Worldwide concern over the depletion of the ozone layer was discussed at the Vienna Convention in 1985 under the auspices of the United Nations Environment Program (UNEP). This culminated in the Montreal Protocol, the first international agreement made to protect the global environment. Signed in September 1987 and revised in June 1990, the Protocol now controls, throughout much of the world, the use and production of CFCs, halons (firefighting agents), carbon tetrachloride, and 1,1,1-trichloroethane (base solvent for degreasing formulations used in the electronic industry).

This agreement originally mandated a 50% reduction in CFC production and consumption by July 1, 1999, but, importantly, allowed for future revision in light of new scientific evidence. Subsequent atmospheric studies suggested that ozone was likely to be depleted at a faster rate than had been previously thought. Following revisions of the Protocol the complete phase-out of CFCs and HCFCs were also mandated. Although the European Union did not convince other signatories to the Protocol to accept proposals for earlier CFC and HCFC phase-out, it decided to include them in its own regulations. The European Union unilaterally banned the use of all recycled CFCs from January 2000. Chlorofluorocarbon recycling is still allowed in the United States and it has even become attractive to recover CFC-12 from its azeotrope with CHF_2CH_3 (HFC-152a) known as R500.

CONCLUSIONS

Since the early 1930s, the CFCs have been widely used as refrigerants, foam blowing agents, propellants, and

so on due to their outstanding physical and thermodynamic properties. However, the CFCs were unfortunately revealed to be responsible for depleting the ozone layer of the stratosphere on Earth. Production and use of CFCs began to be regulated by the international agreement, Montreal Protocol. The HCFCs and HFCs were recommended as the promising alternatives of the CFCs.[19,20] Outstanding physical properties of CFCs, in particular CFC-12, are very difficult to match or substitute by alternative materials in a variety of industrial applications. This fact has rendered a challenge to the research on CFC alternatives.

REFERENCES

1. Resnick, P.R. Fluorine compounds, organic. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 3rd Ed.; Grayson, M., Ed.; John Wiley & Sons: New York, 1980; Vol. 10, 860–870.
2. Howe-Grant, M. *Fluorine Chemistry*; John Wiley & Sons: New York, 1995; 264–281.
3. Elliott, A.J. Chlorofluorocarbons. In *Organofluorine Chemistry*; Banks, R.E., Smart, B.E., Tatlow, J.C., Eds.; Plenum Press: New York, 1994; 145–157.
4. Midgley, T., Jr.; Henne, A.L.; McNary, R.R. Heat Transfer US Patent 1,833,847, Nov. 24, 1931.
5. Molina, M.J.; Rowland, F.S. Stratospheric sink for chlorofluoromethanes: chlorine atom catalyzed destruction of ozone. Nature **1974**, *249*, 810–812.
6. Farman, J.C.; Gardiner, B.G.; Shanklin, J.D. Large losses of total ozone in Antarctica reveal seasonal ClO_X/NO_X interaction. Nature **1985**, *315*, 207–210.
7. Hudlicky, M. *Chemistry of Organic Fluorine Compounds*, 2nd Ed.; Ellis Horwood, Ltd.: Chichester, U.K., 1976.
8. Kalcevic, V.; McGahan, J.F. Process for Treating Spent Catalyst Including Antimony Halides from Chlorofluorocarbon Production US Patent 4,751,063, Jun. 14, 1988.
9. Lee, S.K. Detoxification of Spent Antimony Halide Catalyst and Recovery of Antimony Values US Patent 4,411,874, Oct. 25, 1984.
10. Knaak, J.F. Use of Hexagonal Chromium (111) Oxide Hydroxide Catalyst in Fluorination Process US Patent 3,978,145, Aug. 31, 1976.
11. Gumprecht, W.H.; Manzer, L.E.; Rao, V.N.M. Improved Process for the Manufacture of 1,1,1-Trifluorodichloroethane and 1,1,1,2-Tetrafluorochloroethane EP Patent 313,061, Apr. 26, 1989.
12. Carlson, E.J.; Armor, J.N.; Cunningham, W.J.; Smith, A.M. Chromium Aerogel Method of Producing Same and Fluorinating Process Utilizing Same. US Patent 4,828,818, May 9, 1989.
13. Manzer, L.E. Gas-Phase Fluorination Process EP Patent 331,991, Sep. 13, 1989.
14. Manzer, L.E.; Rao, V.N.M. Gas-Phase Fluorination Process US Patent 4,766,260, Aug. 23, 1988.
15. Murphy, K.P. CFC 113: a unique member of the halogenated solvent family. J. Test. Eval. **1989**, *17*, 87–89.
16. Gardner, L.E. Hydrodehalogenation Process and Catalyst US Patent 3,636,173, Jan. 18, 1972.
17. Nychka, H.R.; Eibeck, R.E. Dechlorination of Haloethanes Using Ethylene US Patent 4,155,941, May 22, 1979.
18. Lovelock, J.E. Atmospheric fluorine compounds as indicators of air movements. Nature **1971**, *230*, 379.
19. McCulloch, A. CFC and halon replacements in the environment. J. Fluorine Chem. **1999**, *100*, 163–173.
20. Powell, R.L. CFC phase-out: have we met the challenge. J. Fluorine Chem. **2002**, *114*, 237–250.

CHP Technology/Systems

Louay M. Chamra
B. K. Hodge
Southeast Cooling Heating and Power Application Center, Department of Mechanical Engineering, Mississippi State University, Mississippi State, Mississippi, U.S.A.

INTRODUCTION

The traditional model of electric power generation and delivery is based on the construction of large, centrally located power plants. ''Central'' means that a power plant is located on a hub surrounded by major electrical load centers. For instance, a power plant may be located close to a city to serve the electrical loads in the city and its suburbs, or a plant may be located in the midpoint of a triangle formed by three cities. Power must be transferred from a centrally located plant to the users. This transfer is accomplished through an electricity grid that consists of high-voltage transmission systems and low-voltage distribution systems. The traditional structures of the electrical utility market, green engineering issues, and environmental acceptability have resulted in a relatively small number of public electric utilities. However, today's technology permits development of smaller, less expensive power plants, bringing in new, independent producers. Competition from these independent producers, along with the re-thinking of existing regulations, is affecting the conventional structure of electric utilities. Concerns about the environment and the availability of ''green'' power have also become more important factors in the willingness to consider alternatives to the traditional centrally located electrical power plants. The restructuring of the electric utility industry and the development of new ''onsite and near-site'' power generation technologies have opened up new possibilities for buildings, building complexes, and communities to generate and sell power. Competitive forces have created new challenges as well as opportunities for companies that can anticipate technological needs and emerging market trends.

Historically, research, development, and commercialization efforts have been focused on individual system components (cooling, thermal storage, heating, ventilation air, and power). A new category consisting of power generation equipment coupled with thermally activated components has evolved as cooling, heating, and power (CHP). A successful CHP system requires a need for both generated electric/shaft power and thermal energy. An operation that does not have a need for both electric/shaft power and thermal energy will not likely benefit from CHP. Cooling, heating, and power is especially beneficial to the buildings that typically require electric power and can utilize thermally activated HVAC system components. The application of CHP system focuses on onsite fuel conversion, combining power generation and HVAC system optimization, and integration with other innovative technologies. Cooling, heating, and power holds some of the answers to the efficiency, pollution, environmental, and deregulation issues that the utility industry currently faces. Economic considerations are an important factor in the specification or retrofit of CHP systems. Not only are first costs and operating and maintenance (O and M) costs important, but also O and M costs uncertainties, especially for new and emerging technologies, can present barriers to CHP acceptance.

A review of CHP technology and systems is presented in this entry. Specific types of distributed power generation (DPG) and thermally activated technologies will be introduced and briefly previewed. Then, an overview of the equipment used in CHP systems is discussed. Finally, a bibliography of CHP systems and related equipment is included at the end of the entry.

THE CHP SYSTEM

Inefficiencies are associated with the traditional method of electric power generation and delivery. Figs. 1 and 2 illustrate the losses inherent in the generation and delivery of electric power in traditional power plants and in combined-cycle power plants. Traditional power plants convert about 30% of the fuel's available energy into electric power, and highly efficient, combined-cycle power plants convert over 50% of the available energy into electric power. The majority of the energy content of the fuel is lost at the power plant through the discharge of waste heat. Further, energy losses occur in the transmission and distribution of electric power to the individual user. Inefficiencies, environmental, and pollution issues associated with conventional power plants provide the impetus for new developments in ''onsite and near-site'' power generation.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120041074

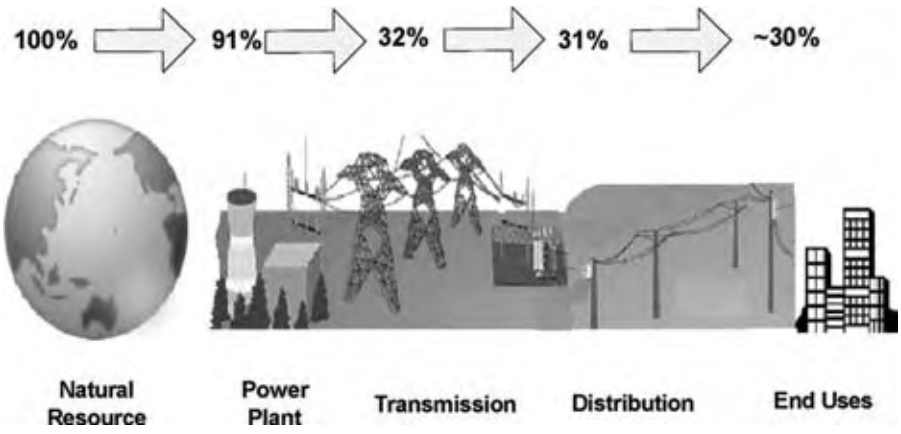


Fig. 1 Efficiency of central power generation. *(View this art in color at www.dekker.com.)*

Cooling, heating, and power combines DPG with thermally activated components to meet the CHP needs of buildings. Cooling, heating, and power systems consist of integrated power generation equipment (gas turbines, microturbines, internal combustion (IC) engines, and fuel cells), thermal systems (water chillers, absorption chillers, air conditioners and refrigeration—electric or engine driven, boilers, and thermal storage), ventilation/IAQ systems (desiccant, enthalpy, and other energy recovery devices), and building control and system integration technologies. Technological advances in both power generation and thermally activated systems have contributed to the development of diverse CHP applications.

Cooling, heating, and power has the potential to reduce carbon dioxide and air pollutant emissions and to increase resource energy efficiency dramatically. It produces both electric or shaft power and useable thermal energy onsite or near-site, converting as much as 80% of the fuel into useable energy. A higher efficiency in energy conversion means less fuel is necessary to meet energy demands. Also, onsite power generation reduces the load on the existing electricity grid and infers better power quality and reliability. Fig. 3 illustrates the increase in efficiency of CHP systems over the power plant efficiencies seen in Figs. 1 and 2. The economic viability of CHP systems is dependent on a number of factors: 1) the cost of CHP-generated electricity and recovered thermal energy vis-à-vis the cost of electrical and thermal (gas) power purchased from the grid; 2) the cost of grid connect charges (and penalties) for back-up and outage power requirements; 3) additional costs incurred by CHP component maintenance (e.g., overhauls and oil changes); and 4) costs associated with additional personnel time for CHP operations.

DISTRIBUTED POWER GENERATION

A number of technologies are commercially available for generating electric power or mechanical shaft power onsite or near-site where the power is used. The three major categories for distributed generation are combustion turbines, IC engines, and fuel cells.

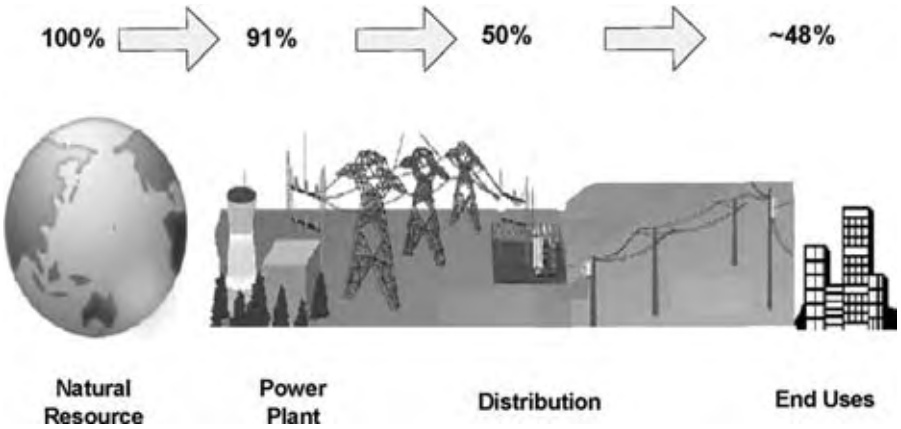


Fig. 2 Efficiency of combined-cycle power generation. *(View this art in color at www.dekker.com.)*

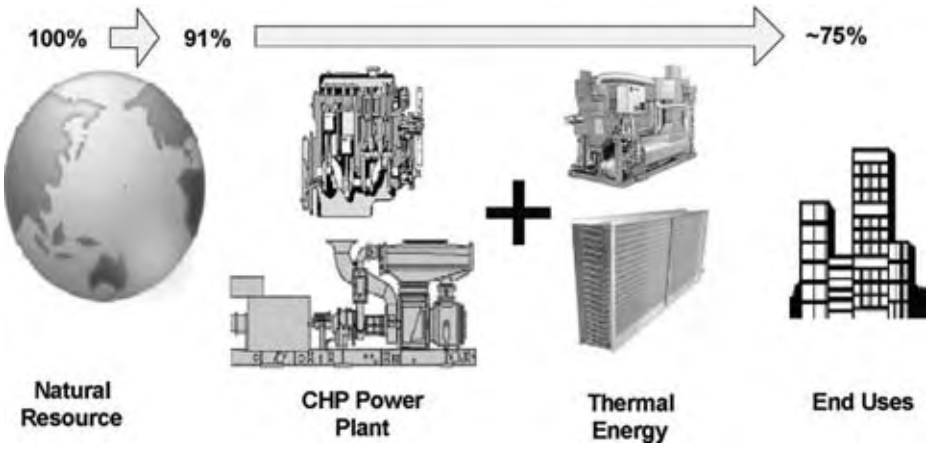


Fig. 3 Efficiency of CHP systems. *(View this art in color at www.dekker.com.)*

Combustion Turbines

Combustion turbines are based on gas turbines and use a variety of fuels, including natural gas, fuel oil, or bio-derived fuels. They can also use recuperators to recover thermal energy in the turbine exhaust streams for pre-heating the air/fuel mixtures for the combustor sections.

The efficiency of electric power generation for combustion turbine systems, operating in a simple-cycle mode (i.e., without external use of heat recovery in the turbine exhaust), ranges from 21% to 40%. Combustion turbines produce high-quality heat that can be used to generate steam or hot water for thermally activated applications, including heating and cooling.

Utilization of thermal energy in a combustion turbine exhaust stream significantly enhances fuel efficiency. Maintenance costs per unit of power output for combustion turbines are among the lowest of all power-generating technologies. As the output capacity of combustion turbines decreases with the increases in ambient air temperature, in hot weather climates or on hot days, cooling of inlet air has been found to be cost effective for many power plants.

Industrial turbines and microturbines are the two types of combustion turbines that are commercially available. Industrial turbines represent a well-established technology for power generation and also represent the ''high'' end of power generating capacity equipment. These can provide from 1 MW to more than 100 MW of electric power. Most CHP systems need capacities below 20 MW, sufficient for large office buildings, hospitals, or small campuses of offices and commercial buildings. The thermal efficiency of industrial gas turbines for power generation ranges from 25% to 40%. A picture of the rotating spool of an industrial turbine is shown in Fig. 4.

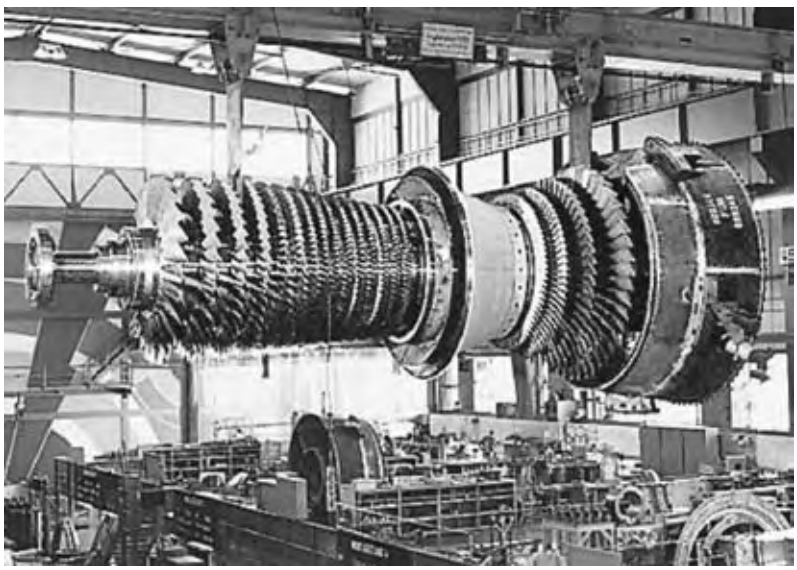


Fig. 4 Industrial turbine spool by Siemens Westinghouse. *(View this art in color at www.dekker.com.)*

Fig. 5 Thirty-kilowatt microturbine by Capstone. *(View this art in color at www.dekker.com.)*

Microturbines are a new generation of smaller turbines. The capacities of microturbines range from 25 to 500 kW. Fig. 5 pictures a 30-kW microturbine by Capstone Turbine, Inc.

Microturbines can use natural gas, propane, and bio-derived gases produced from landfills, sewage treatment facilities, and animal waste processing plants as a primary fuel. The fuel source versatility of microturbines allows their application in rural as well as urban areas. Microturbines evolved from automotive and truck turbochargers, auxiliary power units for airplanes, and small jet engines used on remotely piloted military aircraft. Because microturbines have fewer moving parts than conventional generating equipment of similar capacity, they have the potential to significantly reduce maintenance and operating costs as compared to traditional distributed-energy prime movers. By using recuperators, microturbine systems are capable of energy efficiencies for power generation in the 25–30% range. These turbines have a significant potential for onsite power generation for CHP systems.

Internal Combustion Engines

A reciprocating engine, either four-cycle IC or diesel, is used for producing mechanical shaft power. Shaft power can be used to drive a generator to produce electric power or to operate other equipment including air compressors for process or vapor compression systems for space conditioning. These applications of reciprocating engines are very well established and widespread. Engines can use natural gas, propane, diesel fuel, or bio-derived fuels and are available in capacities ranging from 5 kW to 10 MW. A diesel fuel engine generator is pictured in Fig. 6.

Fig. 6 Engine-generator set by Caterpillar. *(View this art in color at www.dekker.com.)*

Reciprocating engines used for power generation have low capital cost, easy startup, proven reliability, good load-following characteristics, and significant heat recovery potential. They are the most widespread distributed-generation technology in the world today. Existing engines achieve electric generation efficiencies in the range from 25% to over 40%. The incorporation of exhaust catalysts and better combustion design and control has significantly reduced pollutant emissions during the past few years.

Thermal energy in the engine exhaust gases and from the engine cooling system can be employed to provide space heating, hot water, or to power thermally activated equipment. Emissions of engines tend to be higher than those of microturbines and fuel cells. In some locations, depending on local air quality standards, engine emissions may limit reciprocating engine applications for CHP systems.

In a gas engine-driven chiller, the engine produces mechanical shaft power that is used for operating a chiller compressor. Such chillers are very similar to conventional electric-driven chillers. The only difference is that the electric motor that drives the compressor in an electric chiller is replaced with a reciprocating engine.

Fuel Cells

Fuel cells produce electric power by electrochemical reactions, generally between hydrogen and oxygen without the combustion processes. Unlike turbine- and engine-generator sets, fuel cells have no moving parts and, thus, no mechanical inefficiencies.

Phosphoric acid fuel cells (PAFCs) are commercially available. Several hundred PAFC units, most of the order of 200 kW, are operating worldwide. PAFCs are realizing efficiencies of up to 40%. The only byproducts of PAFC operation are water and heat. However, enriched hydrogen fuel must be produced by subjecting hydrocarbon resources (natural gas or

Fig. 7 PureCell™ fuel cell by United Technologies. *(View this art in color at www.dekker.com.)*

methanol) to a reforming or gasification process. This process results in chemical reactions that produce carbon dioxide and other environmental emissions.

Like a battery, a fuel cell produces direct current (DC). However, fuel cells come in a complete package in which the fuel cell stack is integrated with an inverter to convert DC to alternating current and a reformer to provide the hydrogen-rich fuel. Thus, a complete fuel cell system includes a fuel reformer, a fuel cell stack, and a power conditioner. A 200-kW PAFC unit by United Technologies Company is illustrated in Fig. 7.

There are other types of fuel cells of which proton exchange membranes (PEMFC), molten carbonate (MCFC), and solid oxide (SOFC) are the most promising. These fuel cells are at various stages of technology demonstration and are not commercially available. Each type of fuel cell has its own ''preferred'' range of capacities and waste heat temperatures that determine where they can be used to best advantage in CHP systems. Table 1 gives an overview of fuel cell characteristics for the PAFC, SOFC, MCFC, and PEMFC.

Fuel cells are expected to deliver electrical conversion efficiencies in the range of 40–60%. These are used in cogeneration applications that could realize thermal efficiencies of 80–90% for the overall system. Cooling, heating, and power applications can benefit from the high electrical conversion efficiencies of fuel cells and the even higher overall energy conversion efficiencies of fuel cells used in combined cycles and fuel cells coupled with thermally activated components. Fuel cells produce electricity without combustion and with very low NO_x emissions, therefore limiting adverse environmental effects in populated CHP installation

Table 1 Overview of fuel cell characteristics

Electrolyte	PAFC	SOFC	MCFC	PEMFC
Size range	100–200 kW	1 kW–10 MW	250 kW–10 MW	3–250 kW
Operating temperature	200°C (400°F)	1000°C (1800°F)	650°C (1200°F)	90°C (200°F)
Fuel	Natural gas, landfill gas, digester gas, propane	Natural gas, hydrogen, landfill gas, fuel oil	Natural gas, hydrogen	Natural gas, hydrogen, propane, diesel
Efficiency (%)	40–45	50–60	50–60	40–50
Thermal energy applications	Hot water	Hot water, LP steam, HP steam	Hot water, LP steam, HP steam	Hot water (80°C)
Emissions	Nearly zero	Nearly zero	Nearly zero	Nearly zero
Commercial Status	Some available	Nearing availability	Some available	Some available

(From Ref.[1].)

Table 2 Strengths and weaknesses of fuel cells

Electrolyte	PAFC	SOFC	MCFC	PEMFC
Strengths				
Quiet	☐	☐	☐	☐
Low emissions	☐	☐	☐	☐
High efficiency	☐	☐	☐	☐
Proven reliability	☐			
Self reforming		☐	☐	
High energy density		☐		
Automotive application				☐
Weaknesses				
High costs	☐	☐	☐	☐
Low energy density	☐			
R&D stage		☐	☐	☐
Low CHP potential				☐

(From Ref.[1].)

locations. Residential and commercial buildings will also benefit from the quiet and reliable operation of fuel cells. Strengths and weaknesses of each type of fuel cell are listed in Table 2.

Distributed power generation is a required component of a CHP system. Internal combustion engines, combustion turbines, and fuel cells are the current prime movers that have the most potential for DPG. Characteristics of these DPG technologies such as electric efficiency, power output, and cost are presented in Table 3.

HEAT RECOVERY

In most CHP applications, the exhaust gas from a prime mover is ducted to a heat exchanger to recover the thermal energy in the gas stream. Generally, these heat exchangers are air-to-water heat exchangers, where the exhaust gas flows over some form of tube-and-fin heat exchanger surface and the heat from the exhaust gas is used to make hot water or steam. The hot water or steam is then used to provide process energy and/or to operate thermally activated equipment, such as absorption chillers or desiccant dehumidifiers. Many of the thermal-recovery technologies used in building CHP systems require hot water, some at moderate pressures of 15–150 psig. In the cases where additional steam or pressurized hot water is needed, supplemental heat can be added to the exhaust gas with a duct burner.

In some applications, air-to-air heat exchangers can be used. If the emissions from the generation equipment are low enough, the hot exhaust gases can be mixed with make-up air and vented directly into the heating system for building space heating.

In the majority of installations, a flapper damper or ''diverter'' valve is employed to vary the flow across the heat exchanger to maintain a specific design temperature of the hot water or a specific steam generation rate. In some CHP designs, the exhaust gases are used to activate a thermal enthalpy wheel or a desiccant dehumidifier. A thermal wheel uses the exhaust gas to heat a rotating wheel with a medium that absorbs the heat and then transfers the heat into the incoming airflow into which the wheel is rotated.

THERMALLY ACTIVATED DEVICES

Thermally activated devices are based on technologies that use thermal energy, preferably heat from the exhaust gases of power generation equipment, instead of electric energy for providing heating, cooling, or humidity control for buildings. The two primary components for thermally activated devices for application in CHP systems are absorption chillers and desiccant dehumidifiers.

Absorption Chillers

Absorption chillers use heat as the primary source of energy for driving an absorption refrigeration cycle. These chillers require very little electric power (0.02 kW/ton) compared to electric chillers that need 0.47–0.88 kW/ton, depending on the type of electric chiller. Absorption chillers have fewer and smaller moving parts and are quieter during operation than electric chillers. These chillers are also environment-friendly in that they use non-CFC refrigerants.

Commercially available absorption chillers can utilize the following sources of heat:

- Steam
- Hot water
- Exhaust gases
- Direct combustion

Absorption chillers, except those that use direct combustion, are excellent candidates for providing some or the entire cooling load in a CHP system for a building. Modern absorption chillers can also provide heat during winter and feature electronic controls that provide quick startup, automatic purge, and greater turndown capability than many electric chillers. Maintenance contracts and extended warranties are also available for absorption chillers at costs similar to those for electric chillers. Many facilities across the United States are already benefiting from the use of absorption chillers, such as the one pictured in Fig. 8.

Table 3 Comparison of DPG technologies

	Diesel engine	Natural gas engine	Gas turbine	Microturbine	Fuel cell
Electric efficiency	30–50%	25–45%	25–40%	20–30%	40–70%
Power output (MW)	0.05–5	0.05–5	3–200	0.025–0.25	0.2–2
CHP installed cost ($/kW)	800–1500	800–1500	700–900	500–1300	>3000
Footprint (ft^2/kW)	0.22	0.22–0.31	0.02–0.61	0.15–1.5	0.6–4
O and M costs ($/kW hr)	0.005–0.010	0.007–0.015	0.002–0.008	0.002–0.01	0.003–0.015
Availability (% online)	90–95	92–97	90–98	90–98	>95
Time between overhaul	25,000–30,000 hr	24,000–60,000 hr	30,000–50,000 hr	5,000–40,000 hr	10,000–40,000 hr
Start-up time	10 sec	10 sec	10 min–1 hr	60 sec	3 hr–2 days
Fuels	Diesel and residual oil	Natural gas, biogas, propane	Natural gas, biogas, propane, distillate oil	Natural gas, biogas, propane, distillate oil	Hydrogen, natural gas, propane
Fuel pressure (psi)	<5	1–45	125–500 (may require compressor)	40–100 (may require compressor)	0.5–45
NO_x emissions (lb/MW hr)	3–33	2.2–28	0.3–4	0.4–2.2	<0.02
CHP output (Btu/kW hr)	3,400	1,000–5,000	3,400–12,000	4,000–15,000	500–3,700
CHP temperature (°F)	180–900	300–500	500–1,100	400–650	140–700
Recovered heat uses	Hot water, LP steam, district heat	Hot water, LP steam, district heat	Direct heat, hot water, L/HP steam, district heat	Direct heat, hot water, LP steam	Hot water, L/HP steam
Noise	Moderate to high	Moderate to high	Moderate	Moderate	Low

(From Ref.[2].)

Fig. 8 Absorption chiller by broad air conditioning. *(View this art in color at www.dekker.com.)*

Two types of absorption chillers are commercially available: single effect and multiple effect. Compared to single-effect chillers, multiple-effect absorption chillers cost more (higher capital cost) but are more energy efficient and are, thus, less expensive to operate (lower energy cost). The overall economic attractiveness of each chiller depends on many factors, including the cost of capital and the cost of energy.

Desiccant Dehumidifiers

There are two separate aspects of space conditioning for comfort cooling:

- Lowering the temperature of the air (sensible cooling), and
- Reducing humidity in the air (latent cooling)

Fig. 9 Desiccant dehumidifier. *(View this art in color at www.dekker.com.)*

The humidity level should remain below 60% relative humidity to prevent growth of mold, bacteria, and other harmful microorganisms in buildings and to prevent adverse health effects.

Traditionally, temperature and humidity control have been accomplished using a single piece of equipment that reduces the air temperature below the dew point temperature. Moisture in the incoming air condenses on the outside of a cooling coil over which the air passes, and cooler air, containing less moisture, is sent to the space being conditioned. Reducing humidity in the air by cooling often requires lowering the air temperature below a comfortable level and may necessitate reheating of the dehumidified air to achieve comfort.

Desiccant dehumidifiers reduce humidity in the air by using solid desiccant materials or liquid desiccant materials to attract and hold moisture. They can operate independent of chillers and can also be operated in series or parallel with chillers. Recoverable heat from the exhaust gases of turbines and engines for power generation or engine-driven chillers is used for regenerating desiccant material in these dehumidifiers.

In some CHP systems, the moisture content of the air is reduced using a desiccant dehumidifier and the dehumidified air is then cooled using conventional cooling equipment. By reducing the moisture content of the air, desiccant dehumidifiers satisfy the latent cooling load and, thus, reduce the load of the chillers to the sensible cooling (reducing the temperature). Alternatively, a desiccant dehumidifier can be used to further dehumidify and partially reheat cool, saturated air leaving a conventional cooling coil. By positioning the desiccant dehumidifier after the cooling coil, the dehumidification performance of the desiccant is enhanced. This allows the use of moderate or lower temperatures, typical of CHP systems, for regenerating the desiccant. A typical, commercially available desiccant system is shown in Fig. 9.

CASE STUDY: MISSISSIPPI BAPTIST MEDICAL CENTER

The Mississippi Baptist Medical Center (MBMC) in Jackson, Mississippi, represents an excellent example of a CHP system with a long and economically successful record. The MBMC CHP system was brought online in 1991, so it has a long operational history. This case study examines the details of the system as well as the operating experiences.

Fig. 10 is a site photograph of the MBMC. The full-service hospital has 694 beds, a 24-hr emergency room, a medical staff of 500, and more than 3000 total employees.

The history of CHP at MBMC started in 1990 when hospital officials were interested in exploring options to reduce energy costs. In 1990, the hospital had a large electricity requirement, a large steam requirement, a significant price differential per Btu between electricity and gas, a centralized physical plant, and small daily variations in energy requirements. Taken together, the 1990 energy profile indicated CHP, or cogeneration as it was then called, to be a viable economic option, and a CHP system was proposed. The system was expected to provide 70% of the electricity requirement, 95% of the steam required, and 75% of the cooling load. The system contains the following components:

1. Gas turbine generator set: Solar Centaur H Model
2. Diverter valve

Fig. 10 Site view of the MBMC. *(View this art in color at www.dekker.com.)*

3. ABCO waste heat recovery boiler
4. Economizer
5. Steam absorption chiller
6. Primary switch gear

The performance specifications as well as the details of each of the system components are presented in the following paragraphs. System operation can be best understood by examining Fig. 11, which presents a schematic of the system.

The Solar Centaur Turbine is fired by natural gas; shaft power is extracted from the turbine, via a gearbox, and used to drive the generator. The Solar Centaur H Turbine is rated at 5600 hp with an electrical output of 4.0 MW on an ISO standard day. The turbine is controlled by an Allen Bradley PLC 5/20 microprocessor with starting and synchronizing controls, a fire detection/protection system, and vibration analyzer capability. The nominal generator output is 13, 800 V. Fig. 12 shows the turbine installation arrangement.

The diverter valve operation is controlled by the waste heat recovery boiler (WHRB) and directs the exhaust gas to the WHRB boiler or out of the bypass stack to maintain the required steam pressure. The ABCO WHRB is rated at 30,000 lb/hr and has two firing modes. In the turbine-firing mode, a 5.8 MMBtu duct burner is available to supplement the turbine exhaust stream. In the direct-fire mode, used when the turbine is offline, the direct fresh-air fire at 41.5 MMBtu is available.

The economizer utilizes the remaining waste heat to preheat boiler feed water; water treatment chemicals

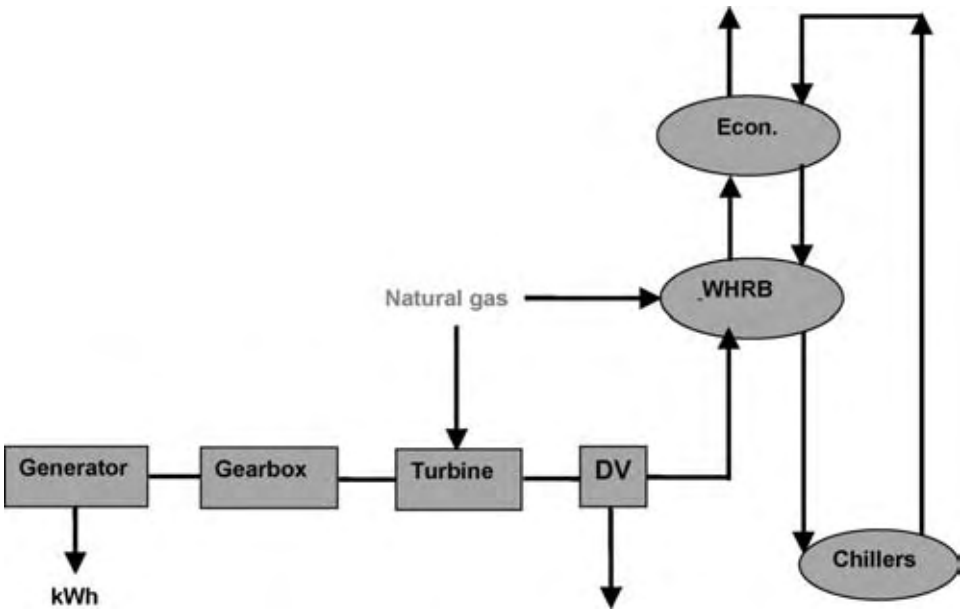


Fig. 11 MBMC CHP system schematic. *(View this art in color at www.dekker.com.)*

Fig. 12 Centaur H turbine installation. *(View this art in color at www.dekker.com.)*

are added to the feed water prior to the economizer. Two absorption chillers are used by the system: a 1250-ton York Paraflow double effect chiller utilizes 115 psi steam at 11.8 lb/hr-ton to produce chilled water at 42°F and a 750-ton Trane double effect chiller utilizes 115 psi steam at 9.6 lb/hr-ton to produce chilled water at 42°F. The chillers supply approximately 60% of the MBMC's 2002 total chilled water requirements. Additions to the facility since 1991 have added cooling load and resulted in 60% rather than 70% of the chilled water load being supplied by the absorption chillers. Fig. 13 shows the two chillers and their installation arrangements.

The primary switchgear is Powell metal-clad 4-bay switchgear that uses vacuum breakers for generator output and primary utility feed and contains generator and utility protective relays as well as synchronizing controls for generator/utility grid interconnect. The turbine-generator does not supply 100% of the electrical load, so the MBMC remains connected to the grid in normal operation. If a CHP problem is sensed by the switchgear, the full hospital load is instantaneously shifted to the grid. If the primary utility grid connection fails, the switchgear will shift the electrical load to a secondary utility feed in 2 sec. The combination of turbine-generator, primary grid, and secondary grid provides the MBMC with triple redundancy for emergency situations.

Schmidt and Hodge[3] examined in detail the economics of the MBMC CHP system and concluded that the actual yearly cost avoidance was close to the original estimate of $800,000/yr. Their economic study results are presented in Table 4.

The actual average yearly cost avoidance for 1994–1996 was $701,000, which compares favorably with the original estimate. The actual payback period

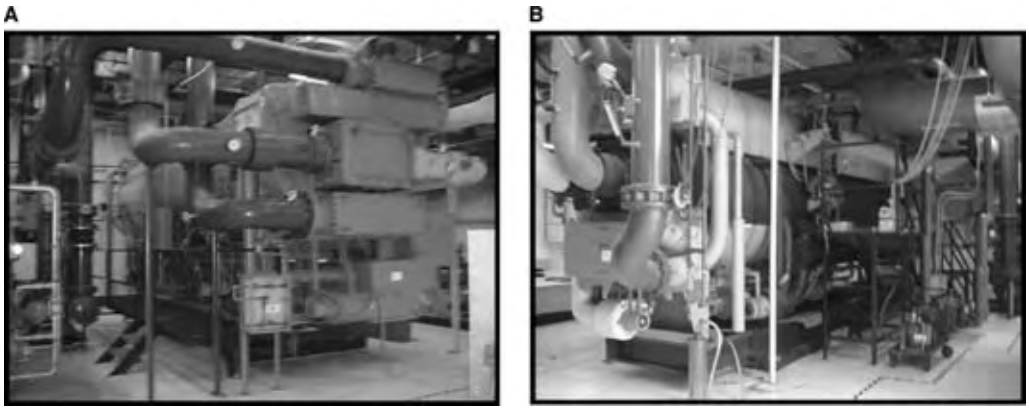


Fig. 13 The York and Trane absorption chillers: (A) York chiller; (B) Trane chiller. *(View this art in color at www.dekker.com.)*

Table 4 Actual cost avoidance

Year	Electricity savings ($)	Natural gas ($)	Maintenance ($)	Savings ($)
1994	1,250,000	402,000	159,000	686,000
1995	1,240,000	432,000	159,000	648,000
1996	1,400,000	468,000	163,000	770,000
Average				701,000

was about 6 yr, which again compares favorably with the estimated payback period of the original economic study of 1990.

One indication of the overall system performance is the percentage of time the CHP system was online. Table 5 presents the online percentage for the years for which the data were available as well as the total electrical generation as a percentage of the electrical requirement for the entire hospital.

As illustrated in Table 5, except for 1998 and 2001, the system provided in excess of 70% of the electricity for the hospital and online percentages mirrored the percentage generation. In 1998, generator problems caused a multi-day outage, and in 2001 the high spike in natural gas costs caused the MBMC to rely solely on grid electricity during the peak-cost period. Taken over the 11 yr of operation, the system was online and producing electricity for all but 2 out of 132 months!

Performance data for a recent calendar year of operation are provided in Table 6. These data illustrate the effectiveness of the system in the MBMC.

The average steam requirements for the absorption chillers and the ancillary hospital usage were virtually satisfied by the turbine exhaust as the duct burner used only 5% of the total gas usage for the system (turbine plus duct burner). Moreover, 2001 was the year in which gas costs caused a curtailment in the turbine usage; nonetheless, the turbine generator accounted for more than 60% of the required electricity.

Many factors have contributed to the successful long-term operation of the MBMC CHP system. The most important of these factors are delineated as follows:

1. A high reliability for all system components.
2. A comprehensive maintenance program.
3. An enthusiastic and competent power-house staff.
4. An accurate and consistent monitoring procedure. A no-penalty switchover-to-grid electrical rate structure. A continuous assessment and improvement process.

Based on the initial desire to avoid electricity costs and the initial estimate of cost avoidance of about $800,000/yr, the CHP systems at MBMC must be rated a success. In the United States, situations where CHP make sense exist by the thousands. This case study illustrates how effective CHP can be as a solution to energy costs.

CONCLUSIONS

For myriad reasons including increasing concerns about energy security, escalating energy costs, power quality requirements, environmental issues, and overall thermal efficiency, CHP is emerging as a viable alternative to conventional grid-based electricity. The driving potential behind CHP systems is the thermal efficiency that can be achieved. Projected system efficiencies of 75% are well above the overall thermal efficiencies experienced by conventional energy systems. Reliability, enhanced power quality, and lower emissions are additional benefits that make CHP systems attractive. In most instances (except where power quality and

Table 5 MBMC online and generation percentages

Year	Generation (%)	Online (%)
1994	74	94
1995	74	98
1996	78	98
1998	64	84
1999	73	98
2001	61	82
2002	70	93

Table 6 Calendar-year 2001 performance data

Total electricity used	34,870,334 kW hr
Electricity generated	21,181,009 kW hr
Electricity purchased	13,689,325 kW hr
Turbine gas	334,221 MMBtu
Duct burner gas	18,286 MMBtu
Average steam production	19,130 lb/hr

Table 7 Rankings for DPG technologies

Category	Positive ⇒ Negative
Efficiency	Fuel cell ⇒ IC engine ⇒ Gas turbine ⇒ Microturbine
Cost/technology status	IC engine ⇒ Gas turbine ⇒ Microturbine ⇒ Fuel cell
Emissions	Fuel cell ⇒ Microturbine ⇒ Gas turbine ⇒ IC engine
Noise	Fuel cell ⇒ Microturbine ⇒ Gas turbine ⇒ IC engine
Small scale capacity	Microturbine ⇒ Fuel cell ⇒ IC engine ⇒ Gas turbine

dependability are primary concerns), the deciding factor in specifying or retrofitting CHP systems will be driven by economic considerations. The case study demonstrated long-term favorable economics for a CHP system in a large hospital. Many such situations exist; indeed, examples abound where long-term favorable economics have been demonstrated. However, the energy mix in the United States is very complicated and especially in times of escalating and uncertain energy costs, a detailed economic analysis is needed if an informed decision is to be made concerning a CHP installation. As more electricity is being generated by natural gas and as natural gas is a likely first choice for many CHP installations, in the long-term, electricity generated by centrally located power plant and that by CHP installations are likely to share common economics. That being the case, the increased thermal efficiency of CHP systems will continue to favorably impact the economics of CHP even in the face of increasing natural gas prices.

A variety of DPG technologies may be selected for a CHP system. Table 7 illustrates how these technologies compare in efficiency, cost/technology status, emissions, noise, and small-scale capacity. The technologies are ranked from those having the most positive characteristics to those having the most negative characteristics. For example, fuel cells have the lowest emissions and, therefore, the best characteristics in this category, and IC engines have the most emissions and, therefore, the most negative characteristics in this category. The rankings in Table 7 are based on the technologies as a whole and may vary in some cases for specific components.

REFERENCES

1. http://www.energy.ca.gov/distgen/. California Distributed Energy Resources Guide.
2. http://www.eere.energy.gov/de/pdfs/chp/. Distributed Energy Resources Combined Heat and Power Program, DOE.
3. Schmidt, D.; Hodge, B.K. Cogeneration at the Mississippi Baptist Medical Center. Eng. Syst. **1998**, *15* (7), 74–78.

Chromatographic Separations

C

Scott M. Husson
Department of Chemical and Biomolecular Engineering, Clemson University, Clemson, South Carolina, U.S.A.

INTRODUCTION

This contribution provides an introduction to various types of chromatographic separations, operating configurations, and methods to characterize chromatographic separation data. After a brief introduction and background, the three chromatography process techniques are described. Gas and liquid chromatography (LC) concepts are presented along with brief introductions to the common stationary phase materials and geometries, the various mechanisms of retaining solutes on the stationary phase, and the common methods of detection. Next, modes of operation from batch to continuous, simulated moving-bed (SMB) chromatography are described with references given to commercialized systems. Finally, chromatographic separation theory is introduced, leading to the definitions of chromatographic figures of merit, including the number of theoretical plates and peak resolution.

The basis of chromatographic separation is the partitioning of chemical solutes between a moving fluid phase, called the mobile phase, and a stationary phase over which the mobile phase flows. The mobile phase can be either a gas or a liquid, while the stationary phase is either a solid or a liquid applied to a solid substrate. Separation of the chemical solutes in the mobile phase occurs based on differences in their partitioning between the two phases. Those that partition to the stationary phase to a greater extent will take longer to flow past it.

Categorization of chromatographic separations begins with process techniques: elution development, displacement development, and frontal analysis chromatography. Next, the types of chromatographic separations can be categorized based on mobile phase: gas or liquid. Further categorization is based on the mechanism of retaining solutes on the stationary phase, e.g., adsorption, ion exchange, etc. Still further categorization is based on the stationary phase geometry: thin layer, packed bed, capillary column, membrane module, etc. Finally, the mode of operation also has categories: batch, continuous, moving-bed operation; continuous, SMB; continuous annular chromatography, etc. For detailed information on these topics, see the Bibliography section.

BACKGROUND

The Russian botanist Mikhail Tswett is generally recognized as the inventor of chromatography for his work 100 yr ago. His work involved separating plant pigments by eluting a mixture of the pigments on a column of calcium carbonate. The pigments separated into colored bands on the column, hence the name chromatography.[1] However, the observations of Tswett attracted little attention until about the 1950s, when Martin and Synge demonstrated its usefulness for biological separations. In 1952, Martin and Synge received the Nobel Prize for their work, which revolutionized analytical chemistry. In addition to its wide use in analytical chemistry and laboratory-scale separations, chromatography is used on a preparative scale ($<$1 ton/day) in the production of pharmaceuticals and fine chemicals, and on a larger, bulk scale ($>$50 tons/day) in separations of close-boiling isomers and mixed sugars.[2,3] To understand these and other aspects of chromatography, limited definitions and nomenclature are presented as needed. For a more comprehensive listing of chromatography nomenclature, see Ref.[4].

PROCESS TECHNIQUES IN CHROMATOGRAPHIC SEPARATIONS

The term development is used to describe the process by which chemical solutes are transported through the stationary phase. For batch modes of operation, solute development is done by elution or displacement processes. For continuous modes of operation, solute development is accomplished by the process of frontal analysis.

Elution Development

Elution development is the most common process technique for chromatographic separations. A pulse of sample mixture is introduced into the mobile phase at a point near the inlet of the stationary phase, and all of the chemical solutes migrate through the

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007704

stationary phase at rates that depend on their degrees of interaction with the stationary phase material. If the migration rates of the solutes differ, the solutes will gradually form distinct zones separated by pure mobile phase solvent. Fractional collection of the fluid zones as they exit the column results in pure solute products diluted in the mobile phase solvent.

Displacement Development

In displacement development chromatography, a pulse of the sample is introduced into the system at the inlet of the stationary phase. The difference from elution development is the use of one or more mobile phase solvents, or displacing agents that interact strongly with the stationary phase material. If the chromatographic conditions and the mobile phase are kept constant throughout the entire experiment, then the mode of operation is termed isocratic elution. In the case where multiple mobile phase solvents are used, a first solvent is chosen, which interacts with the stationary phase more strongly than does the least-interacting solute. Once this solute has been displaced far enough along the stationary phase, a second mobile phase solvent is introduced, which is able to displace the next most weakly interacting solute, and so forth. Gradient elution is a type of displacement development in which the mobile phase continuously changes composition, gradually becoming stronger in its ability to displace bound solutes.

Frontal Analysis

Frontal analysis chromatography allows greater flexibility for continuous scale-up than elution or displacement chromatography. The primary difference is that the sample to be separated is introduced as a step change in frontal analysis. As the sample mixture is fed continuously into one end of the stationary phase bed, the least retained solute begins to emerge at the outlet of the bed as a pure product. This product can be recovered until, eventually, the stationary phase material becomes saturated with the next most strongly retained solute, which then begins to ''break through'' into the product stream. Operating as such, frontal analysis allows recovery of only one pure product. Advances in stationary phase–mobile phase contactor design overcome this apparent limitation; these advances are described under the section ''Modes of Operation.''

GAS CHROMATOGRAPHY

In gas chromatography (GC), a sample is vaporized into a flowing gas stream called the carrier gas, and the components of the sample are separated based on differences in the extent to which they are retained by a stationary phase. Fig. 1 illustrates the components of a typical GC instrument. A standard setup comprises a sample introduction system, the column and oven heater, and the detector. Microliter samples are injected through a rubber septum into a heated sample port at the inlet of the column that holds the stationary phase. For applications that use a capillary column, a stream-splitter is employed that directs only a fraction of the sample through the column. A carrier gas (mobile phase) moves the sample through the system. An oven is used to control the temperature of the column, detector, and injection port. The sample port is maintained at a temperature above the normal boiling point of the least volatile component. A limitation for GC separations, then, is that the sample components must be stable thermally at the temperatures used in the sample port and column oven. The oven may operate at a constant temperature, or it may be programmed in time to ramp to higher temperatures. Most GC systems now have computer software packages that control all aspects of operation and data collection, including operation of automated injection systems.

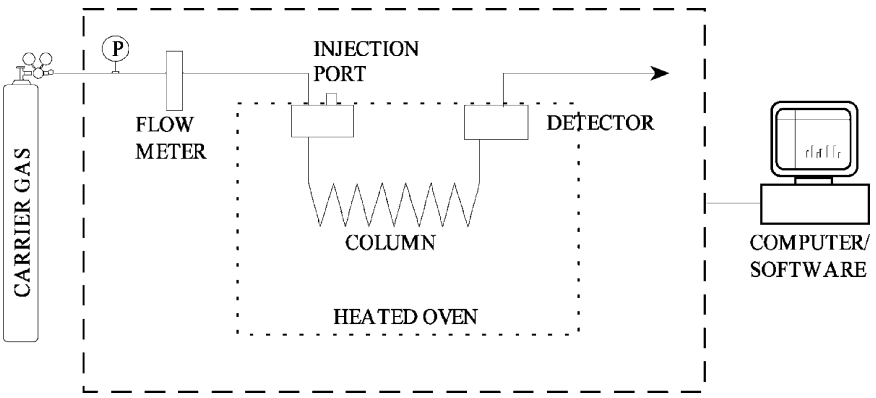


Fig. 1 Schematic illustration of a basic GC system.

Carrier Gases for GC

The most common carrier gases are helium, nitrogen, argon, and hydrogen. Research-grade gases (preferably 99.995%) are used for GC separations to avoid low-level contaminants such as water, hydrocarbons, and carbon dioxide. Even at low concentrations, these impurities can affect detector stability seriously.[5] Carrier gases are supplied from a generator or a cylinder with a regulator that allows pressure regulation between 10 and 250 psig at its outlet. Selection of appropriate carrier gases depends primarily on the type of detector, as described in the following sections.

Detectors for GC

Dozens of detectors have been developed for use in GC. The two most widely used, low-cost detectors are the thermal conductivity detector (TCD) and the flame ionization detector (FID). For descriptions of other detectors see Grob (1995) in Bibliography.

Thermal conductivity detector

A TCD uses heating filaments to measure differences in thermal conductivity between the sample stream exiting from the column and pure carrier gas. As the concentration of sample in the gas phase affects the ability of the gas to conduct heat, differences in the thermal conductivity are proportional to the molar concentration of sample in the gas. Maximum signal response occurs with hydrogen as the carrier gas, because its high thermal conductivity maximizes the difference in thermal conductivity between sample and carrier. Helium is used commonly and offers a safer option than hydrogen. Good control of carrier flow rate is needed for TCD, because it detects changes in sample concentration, not mass. The minimum detectable level (MDL) for TCD is $\sim 5 \times 10^{-10}$ g/ml.[6]

Flame ionization detector

For FID detection, hydrogen is mixed with the carrier gas, and the column effluent stream is burned in air. The FID measures the ionization of the sample stream, which is proportional to the mass of the sample in the gas. Flame ionization detector has high sensitivity (MDL is $\sim 10^{-12}$ g/sec), but it has two main disadvantages: FID cannot detect compounds that do not produce ions in a hydrogen–oxygen flame (e.g., H_2O, CO_2, CO, CS_2, N_2, H_2S, formic acid, and nitrogen oxides). Further, FID destroys the sample material in the detection process.[6] Therefore, if the gas effluent is to be examined further, an effluent splitter is required prior to the FID detector.

Gas Chromatography Columns

Both packed and capillary columns can be used to house the stationary phase for GC. Packed columns offer high capacity and low cost. Packed GC columns are typically used in process laboratories for separating or determining components in a sample, or for gas analysis applications (e.g., for the determination of BTU content in natural gas). Capillary columns offer higher efficiencies, but are generally more expensive. These are used commonly in analytical laboratories when resolution, rather than capacity, is most important (e.g., chiral separations). In capillary chromatography, the stationary phase material is coated on the inner surface of a capillary tube. Hybrids of a sort are the support-coated open tubular and porous layer open tubular (PLOT) columns that use a deposited layer of liquid-coated particles or porous particles on the inner wall of a length of tubing. Column lengths vary from 2 to 3 m for typical packed columns to 25–50 m for typical capillary columns.

Gas Chromatography Stationary Phases

Gas chromatography is often divided into categories based on the type of stationary phase used. Gas–liquid chromatography (GLC) implements a porous, inert solid support that is coated with a viscous, nonvolatile liquid phase. On the other hand, gas–solid chromatography (GSC) uses a solid adsorbent as the stationary phase. Klee offers these general rules-of-thumb for selection of stationary phase materials: use solid adsorbents to separate room-temperature gases, liquid stationary phases to separate room-temperature liquid and solid mixtures, polar phases for polar solutes, and nonpolar phases for nonpolar solutes.[7] Table 1 lists common liquid- and solid-stationary phase materials available for use in capillary columns. Barry cross-refers numerous column materials from nine different manufacturers.[8]

Gas–liquid chromatography

In GLC, separation occurs based on differences in partitioning of the sample components between the carrier gas and the liquid phase. A wide selection of liquid phases makes GLC a versatile separation technique. Further, the liquid phase can be a polymer or a chemically bonded phase. In all cases, the liquid phase is film coated or chemically bonded onto a solid support surface or a column wall. Liquid-bonded phases overcome the problem of leakage of the stationary phase material into the carrier. They are used commonly in LC also, and the process to fabricate

Table 1 Common capillary GC stationary phase materials and applications

Stationary phase type[a]	Stationary phase material	Maximum operating temperature (°C)	Separation applications
NP liquid; bonded	Poly(dimethylsiloxane)	400	General purpose
NP liquid; bonded	Poly(phenylmethylsiloxane)	320	Pesticides, herbicides, PCBs, and PAHs
Liquid; nonbonded	Derivatized cyclodextrans	260	Chiral separation of small molecules
P liquid; bonded	Poly(alkylene glycol)	240	Fatty acid methyl esters, phenols, fragrance oils, and aldehydes/ketones
P liquid; nonbonded	Base-modified poly(ethylene glycol)	200	Amines and other volatile basic compounds
P liquid; bonded	Acid-modified poly(ethylene glycol)	250	Volatile organic acids without derivatization and phenols
NP solid; PLOT	Carbon	250	Permanent gases and C1–C5 hydrocarbons
NP solid; PLOT	Molecular sieve 5 A	300	Permanent and noble gases
P solid; PLOT	Alumina	200	General use. C1–C8 hydrocarbon isomers
NP solid; PLOT	Poly(styrene-divinylbenzene)	290	CO_2, C1–C4 hydrocarbons, sulfur gases, and polar compounds
P solid; PLOT	Poly(divinylbenzene-ethylene glycol dimethacrylate)	190	C1–C7 hydrocarbons, amines, and organic solvents

[a]NP, nonpolar; P, polar.

''bonded-phase'' materials is described under the section ''Partition Liquid Chromatography.''

Gas–solid chromatography

In GSC, separation occurs based on differences in the adsorption of the various components in the sample onto the solid adsorbent. While GSC may not offer as much flexibility in stationary phase functionality as GLC, it has its own advantages. For separation applications, advantages include higher available operating temperatures, higher column efficiencies, and no stationary phase leakage.[5] Typical solid phases for GSC include zeolites, silica gel, activated alumina, carbon, carbon molecular sieves, diatomites, and porous polymers.

Separation Applications of GC

Gas chromatography finds routine use as an analytical tool in numerous industries, universities, and government and private laboratories. Gas chromatography is used to separate and identify components of gaseous, liquid, or volatile solid mixtures. For nonvolatile compounds, special techniques can be used to derivatize the sample to a form amenable to GC. Alternatively, nonvolatile samples can be pyrolyzed and their pyrolysis products are separated and analyzed.[5]

Separation applications are far too numerous to list; a few applications that demonstrate the breadth of use are listed. In the petroleum industry, for example, GC is used to separate complex mixtures during exploration, production, and refining of crude oil, and to simulate distillation of petroleum fractions. In clinical medicine and forensic science, GC is used to separate and identify drug compounds in biological fluids and tissues. In environmental laboratories, GC is used to separate and identify chemical compounds in air, soil, and water samples. In food science, GC is used to separate and identify antioxidants and preservatives, among other applications.

LIQUID CHROMATOGRAPHY

In LC, the components of a flowing liquid sample are separated based on differences in the extent to which they are retained by a solid stationary phase. The stationary phase may be packed into a glass, plastic, or metal column or cartridge (column chromatography); alternatively, it can be spread out on a flat glass, plastic, or metal surface (layer chromatography). Fig. 2 illustrates a typical LC instrument used for analytical or laboratory-scale applications; LC systems for continuous, large-scale separations are described under the section ''Modes of Operation.'' A standard setup has a solvent delivery system, a column that performs the separation, a detection system, and a fraction collection system. For low-pressure LC, the solvent delivery system can be a reservoir of mobile phase that drains into the column by gravity flow. For high-pressure (or high-performance) LC (HPLC), the solvent delivery system comprises a solvent reservoir(s), a mechanism for solvent degassing, and a solvent pump. The column is commonly held in a thermostatted chamber to control temperature. Sample components emerge from the end of the column and are

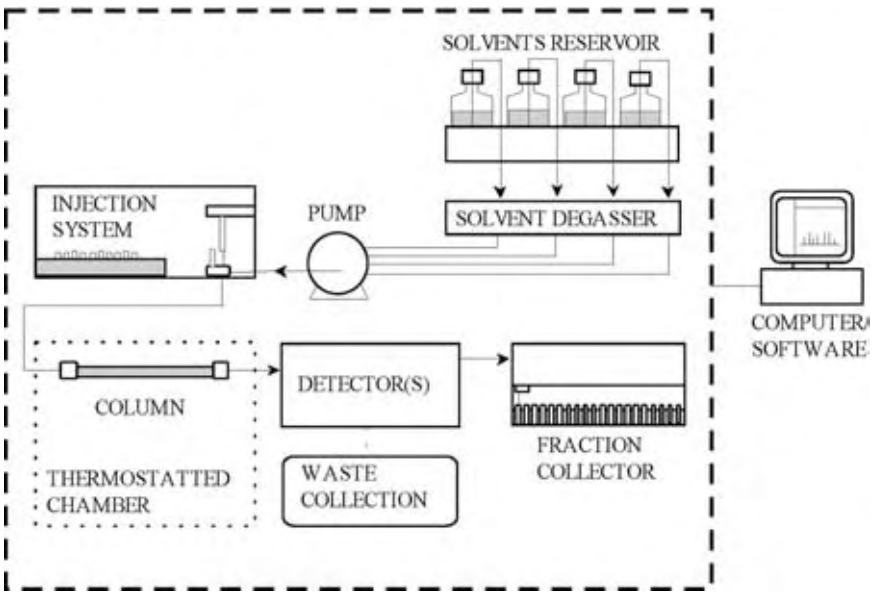


Fig. 2 Schematic illustration of a basic LC system.

monitored by a detector and/or collected in fractions. As with GC, most LC systems now have computer software that controls all aspects of operation and data collection, including operation of automated injection systems and time-dependent programming of temperature and solvent composition.

Detectors for LC

As with GC, numerous detectors have been developed for LC. Each can be categorized as one of two types of detectors: solute property and bulk property. The two most commonly used detectors for LC are the ultraviolet-photometric (UV) detector and the refractive index (RI) detector. Wheals describes these and other common LC detectors.[9]

Ultraviolet-photometric detector

The UV detector is the most widely used detector for LC. It is a solute property detector that is suitable for those solute compounds that absorb radiation in the UV range (~190–400 nm). Ultraviolet-photometric detectors are relatively insensitive to temperature and flow rate fluctuations. The sensitivity to solute detection is high (noise equivalent concentration $\sim 10^{-10}$ g/ml).[9] Ultraviolet-photometric detectors are also well suited to applications that use gradient elution, given that many common LC solvents have low UV absorptivities.

Refractive index detector

Refractive index detectors continuously monitor the difference in bulk RI of the mobile phase and that of a reference mobile phase containing no solute. As such, RI is a bulk property detector. Unlike UV detectors, RI detectors are highly sensitive to temperature fluctuations and somewhat sensitive to flow rate fluctuations.[10] Temperature should be maintained within $\pm 0.001^{\circ}$C for high-sensitivity measurements.[10] Noise equivalent concentration for RI is $\sim 10^{-7}$ g/ml.[9]

Liquid Chromatography Columns

For analytical LC, standard packed columns (4–8 mm I.D.), capillary packed columns (50–100 μm I.D.), and microbore packed columns (0.5–1.5 mm I.D.) are used widely. Column sizes depend on the application, e.g., analytical, preparative, or commercial separations. Special configurations also exist, including membrane chromatography modules (stacks and hollow fibers) that offer lower pressure drops and easier scale-up than packed beds.

Liquid Chromatography Stationary Phases

Gas chromatography achieves separation based on solute partitioning into a thin liquid coating or adsorption onto a solid stationary phase material, whereas LC uses these and other mechanisms to achieve separation. The type of stationary phase and, in some cases, mobile phase determines the mechanism to achieve separation. Descriptions that follow are of the mechanisms and materials used most often.

Absorption (partition) chromatography

Nowadays, liquid stationary phases in LC comprise a liquid material bonded covalently to a solid support material. These bonded-phase materials allow liquid phase partition chromatography without the concern of leakage of the stationary phase into the flowing mobile liquid phase. Partition chromatography offers a wide range of selectivity effects; separations depend on differences among the functional groups in the solutes and also the two liquid phases. The ability to tune the solute selectivity by changing both liquid phases makes partition chromatography particularly well suited for separations that involve mixtures of chemicals from a homologous series.[11]

A common method to prepare bonded-phase materials is to use silica gel as the solid support and to react functionalized organosilanes onto the surface hydroxyl groups of the silica gel. Of course, other solid support materials can be used provided that they present surface functional groups (e.g., OH, COOH, and NH_2) that can participate in modification reactions. The flexibility offered by this approach leads to stationary phase materials with myriad functionalities. Fig. 3 illustrates commonly used organosilane reactants and two example reaction paths that lead to bonded-phase silica. For more discussion on bonded-phase materials, see Ref.[12].

Adsorption chromatography

Adsorption chromatography exploits differences in the relative affinity of solutes for a solid adsorbent used as the stationary phase. Common stationary phase materials for adsorption chromatography are porous silica gel, activated alumina, activated carbon, magnesium oxide, carbonates, and highly cross-linked polymers such as styrene-divinylbenzene and methacrylates.[11] The chemical natures of these adsorbent stationary phase materials make them well suited for separations of solute mixtures that differ in polarity and chemical functionality. For example, silica is an acidic adsorbent that retains basic compounds to a greater extent than nonbasic ones. In contrast, alumina

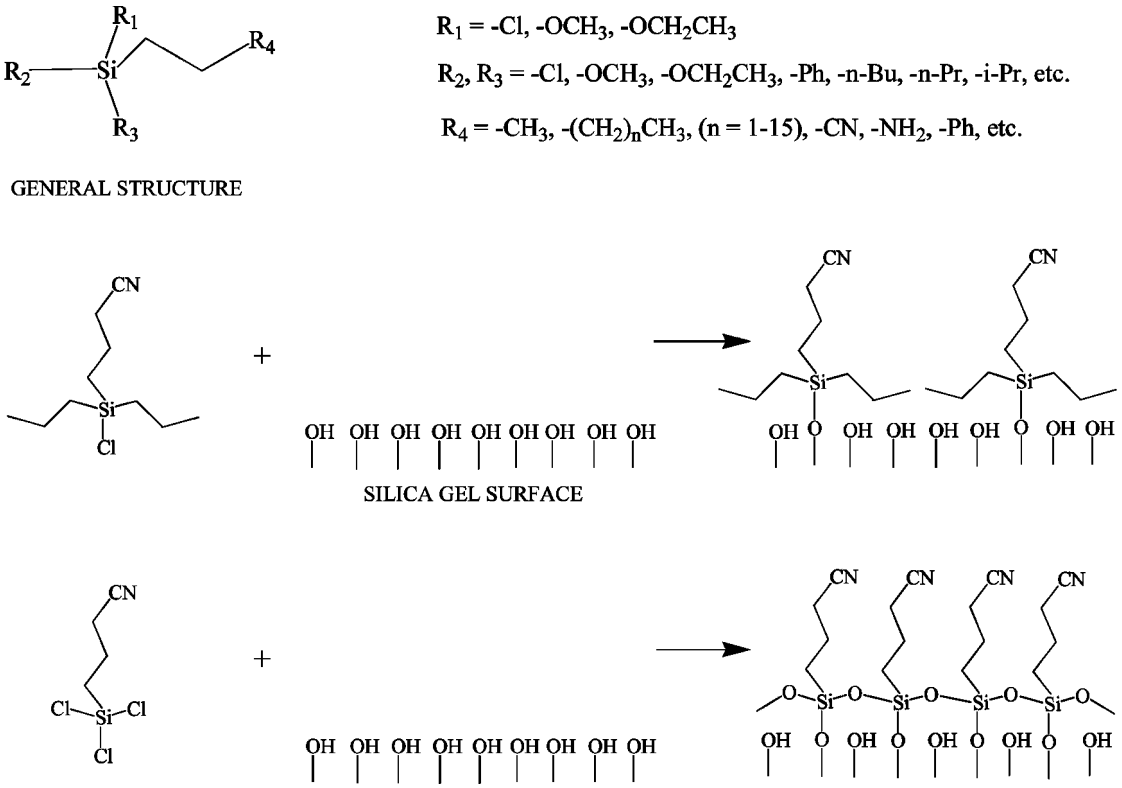


Fig. 3 Chemistries involved in the preparation of bonded phases from organosilane modifiers.

is a basic adsorbent that retains acidic compounds preferentially.

Ion exchange chromatography

Ion exchange chromatography utilizes a reversible chemical reaction as the mechanism to achieve separation. The stationary phase for ion exchange chromatography comprises a charged support, with freely exchangeable counterions. Charged solutes ($M^{\pm}$) in solution become fixed to the stationary phase according to the following ion exchange reaction:

$$M^{\pm} + \left(C^{\pm}R^{\mp}\right)_{solid} \leftrightarrow \left(M^{\pm}R^{\mp}\right)_{solid} + C^{\pm} \qquad (1)$$

In effect, ion exchange chromatography exploits differences in the degree of charge and affinity (measured by the reaction equilibrium constant or molar selectivity coefficient) of ions in solution for sites of opposite charge in the stationary phase to accomplish separation.

Depending on the ionic charge(s) of the solutes to be separated, anion or cation exchange stationary phases must be used. Common ion exchange support materials include silica, poly(styrene-divinylbenzene), and cellulose. Common anion exchange groups anchored to these supports include 1°, 2°, 3°, and 4° amines. The latter results in a strong anion exchange phase. Common cation exchange groups are carboxylic and sulfonic acids. The latter produces a strong cation exchange phase.

Ion exclusion and ion pair chromatography also exploit differences in solute charge to accomplish separation. In ion exclusion, fully dissociated compounds are restricted from accessing the stationary phase pores caused by Donnan exclusion, thereby eluting at the column void volume, whereas nondissociated compounds have full access to pore volume. Ion pair chromatography uses an ion exchange compound in the mobile phase, along with a nonfunctionalized poly(styrene-divinylbenzene) stationary phase to accomplish separation of solute ions of opposite charge. (For details on ion exchange, ion exclusion, and ion pair chromatography, see Weiss in ''Bibliography.'')

Size exclusion chromatography

Size exclusion chromatography (SEC) separates molecules based on differences in their hydrodynamic volume (or size). For this reason, SEC has become a standard technique for determining molecular weight distributions of polymer samples. Common separation applications include polymer, protein, and nucleic

acid separations; plasmid purification; and polysaccharide fractionation. When applied to organosoluble polymers (e.g., polystyrenes, polyurethanes, and polysiloxanes), SEC is often called gel permeation chromatography; when applied to water-soluble polymers and biopolymers (e.g., polyacrylamide, proteins, peptides, and oligonucleotides), it is sometimes called gel filtration chromatography (GFC). Size exclusion chromatography can be performed at high temperatures (>200°C) for the analysis of compounds that are only sparingly soluble at room temperature, including high molecular weight polymers.

The column packing in SEC comprises porous, spherical gel beads with a defined pore size distribution. Most often, these beads are made from poly(styrene-divinylbenzene). (For GFC, cross-linked dextran and agarose gels are often used.[13]) The sample is dissolved in a suitable solvent that is often used as the mobile phase as well. Separation occurs as a result of differences in accessibility of pore volume. Small molecules can freely access the whole pore volume; as a result, the column retards these molecules the greatest. As molecular volume increases, less and less pore volume is accessible for molecules to sample, and elution times decrease. For all molecules with hydrodynamic volumes that are too large to penetrate into the pores of the packing, elution occurs at the (interstitial) void volume of the column. The retention volume for each solute can be described mathematically as:

$$V_R = V_0 + KV_P \quad (2)$$

where V_R is the retention volume, V_0 is the void volume, V_P is the pore volume, and K is a distribution constant. In SEC, the distribution constant is defined as the ratio of the average concentration of the solute in the pores to that in the interstitial volume. For systems where separation is based on hydrodynamic volume only, i.e., no enthalpic interactions between solute and packing, values for K are bound between 0 and 1. Small molecules have $K = 1$, whereas large molecules that cannot penetrate into the pore volume have $K = 0$ (see Fig. 4). (For details on SEC see Mori and Barth in ''Bibliography.'')

Affinity chromatography

Affinity chromatography is a form of adsorption chromatography, wherein the solid support surface has been functionalized with an immobilized compound (ligand) that complements the solute (ligate) to be separated from a complex mixture.[13] It is an important technique for difficult biological separations. Here, it exploits the unique ability of proteins to specifically bind molecules noncovalently.

A common solid support is functionalized agarose beads, which provide high capacity, low nonspecific solute binding, and high stability.[13] The ligand (e.g., a protein, nucleic acid, biopolymer, etc.) is anchored covalently to the support matrix using well-developed coupling chemistries. A solution containing the protein of interest is flowed past this affinity stationary phase, whereby the protein binds to the tethered ligand. After this step, the protein is recovered from the stationary phase by displacement with a wash solution. Displacement can be done by adding free ligand to the wash solution, or by changing its pH,

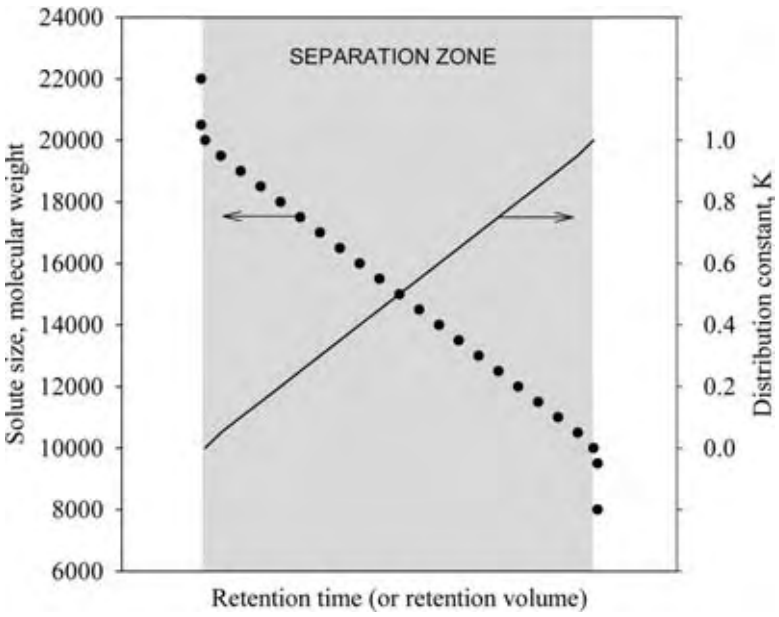


Fig. 4 Illustration of an SEC calibration curve. High-MW molecules that have $K = 0$ and low-MW molecules that have $K = 1$ pass through the column with no separation. For molecules that have $0 < K < 1$, size-based separation occurs.

ionic strength, or temperature relative to the initial protein solution.

Mobile Phase Selection for LC

Selection of a mobile phase for LC is more difficult than for GC. In LC, solute interactions with the mobile phase affect solute retention dramatically. The goal is to optimize the interactions between the solutes and the mobile phase and the solutes and the stationary phase to accomplish separation. If the solutes are to be separated based on differences in polarity, then a common approach is to use a polar stationary phase with a less polar mobile phase; this approach is termed normal-phase chromatography. If the solutes are to be separated based on differences in polarizabilities, then a common approach is to use a nonpolar stationary phase and a polar mobile phase; this approach is termed reverse-phase chromatography.

Practically speaking, the mobile phase should not degrade the stationary phase, should have adequate solvency for the solute mixture, and should not interfere with product recovery in preparatory or commercial separations.[14] For guidelines on solvent selection for partition and adsorption LC, see Refs.[12,14,15]. (For guidelines in ion chromatography see Weiss in ''Bibliography.'')

Gradient Elution

Gradient elution refers to the use of a mobile phase that changes composition during separation. As such, gradient elution is actually a displacement development technique whereby the solvent strength is increased gradually from the start to the end of separation. Solutes that are held weakly to the stationary phase elute from the column in the early stages when the mobile phase solvent strength is low. More strongly held solutes are eluted by mobile phase with high solvent strength. After separation, the stationary phase must be washed before it is reused to flush away the last part of the mobile phase gradient. For ion exchange chromatography, solvent strength can be altered by solvent composition, temperature, ionic strength, and pH.

MODES OF OPERATION

Batch Operation

The commonly used elution and displacement chromatography process techniques use batch operation, wherein a pulse of the mixture to be separated is loaded onto the inlet of the chromatography column. For this reason, batch operation is suitable for small-scale separations. For scale-up to high-volume separations, continuous modes of operation are used typically.

Continuous Operation

Continuous modes of LC are based on the frontal analysis process technique, whereby the mixture to be separated is introduced to the chromatography column as a step change, rather than a pulse. The inherent difficulty for continuous operation is that the stationary phase has a finite capacity that becomes saturated after relatively few bed volumes of liquid have passed.[3] Also, in nearly all cases, the stationary phase must be regenerated for reuse. To achieve continuous operation, the stationary phase must pass through a loading zone, then through a regeneration zone, subsequently back through the loading zone, and so forth. Two options exist for moving the solid stationary phase through these loading and regeneration zones: physically moving the solids (moving bed) and simulating the movement of the solids by changing the locations of the feed and regenerant entry points (simulated moving bed).

Moving-bed operation

Humphrey and Keller note that few successful moving-bed systems have been in operation for liquid systems.[3] Most moving-bed systems fail owing to stationary phase attrition, equipment erosion, and difficulty in achieving plug flow of the solid and liquid phases.[3] Nevertheless, two continuous moving-bed contactors have been used successfully for ion exchange separations: the Higgins contactor and the Himsley fluidized-bed process. The reader is directed to Ref.[11] for more discussion on these contacting devices.

Simulated moving-bed operation

To overcome problems associated with moving solid stationary phase materials, SMB operation fixes the location of the stationary phase and changes the feed and regenerant entry points and the raffinate and product removal points with time. Single- and multicolumn SMB systems are in use. Humphrey and Keller[3] instruct that the single-column system is favored economically when processing feed flows over $\sim$50 million lb/yr, while the multicolumn system is favored for processing lesser feed amounts.

A commercial example of the single-column SMB system is the Sorbex technology of UOP. Table 2 lists

Table 2 Commercial technologies based on UOP Sorbex Technology

Process name[a]	Separation
Parex™	*p*-Xylene from mixed C8 aromatics
Ebex™	Ethylbenzene from mixed C8 aromatics
MX Sorbex™	*m*-Xylene from mixed C8 aromatics
Molex™	Linear paraffins from branched/cyclic hydrocarbons
Olex™	Olefins from paraffins
Cresex™	*p*-Cresol or *m*-cresol from other cresols
Cymex™	*p*-Cymene or *m*-cymene from other cymenes
Sarex™	Fructose from mixed sugars

[a]Trademarks of UOP LLC.

examples of commercial processes licensed by UOP that utilize the Sorbex technology.

In a multicolumn SMB system, the stationary phase is distributed among a number of separate columns that are connected to one another and to lines that introduce feed and regenerant and that recover products. A commercial example of a multicolumn SMB system is the ISEP Continuous Contactor from Advanced Separation Technologies, Inc., in which a series of stationary phase columns rotate on a carousel that delivers them to loading and regeneration zones. For more discussion and illustrations of single- and multicolumn SMB systems, see Ref.[3]. (For further discussion on preparative and production-scale chromatography see Ganetsos and Barker in ''Bibliography.'')

INTRODUCTION TO CHROMATOGRAPHIC SEPARATION THEORY

Consider a closed system comprising two phases and two solutes as illustrated in Fig. 5. When the number of moles of solute in each phase becomes ''statistically'' constant, then the system is in equilibrium. The term ''statistically'' was introduced to emphasize that equilibrium is a dynamic state in which molecules move back and forth between the two phases. At equilibrium, a distribution constant (or partition coefficient) can be defined as

$$K = \frac{c_S}{c_M} = \frac{\text{Molar concentration in stationary phase}}{\text{Molar concentration in mobile phase}} \quad (3)$$

The units of concentration in Eq. (3) are typically moles per volume. However, in some cases (e.g., with capillary columns), it is more convenient to express concentration of solute in the stationary phase in units of moles per mass of stationary phase, rather than per volume.[16]

A second way of defining the distribution constant results from considering a single solute molecule. Under the conditions of dynamic equilibrium, this single molecule spends some of its time in each phase. The time spent in the stationary phase relative to the time spent in the mobile phase is also given by the distribution constant. This definition forms the basis of the chromatography theory.

Fig. 5 illustrates a physical model of the chromatography process. Initially, there is a dynamic equilibrium of molecules between the phases. Then, one phase is moved relative to the other with an average velocity, v. In the stationary phase, molecules do not move; while in the mobile phase, molecules move with a velocity equal to v. Provided that the interphase mass transfer rate is fast relative to the flow rate of the mobile phase, the time-average distribution of a molecule between the phases is statistically equal to the equilibrium distribution as determined by the distribution constant.

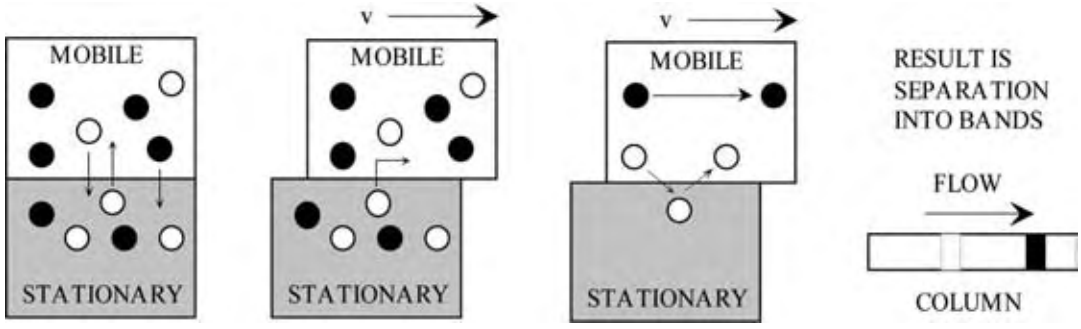


Fig. 5 Model of the physical process used to separate mixtures by chromatography.

With the time-average distribution equal to the equilibrium distribution, the fraction of time spent by a particular molecule in the mobile phase, τ_i, can be evaluated mathematically as

$$\tau_i = \frac{V_M}{V_M + K_i V_S} \quad (4)$$

where V_M is the volume of the mobile phase, K_i is the distribution constant of component i (based on a molar concentration per volume of stationary phase), and V_S is the volume of the stationary phase.

As the molecule is immobile during its time in the stationary phase, it will move down the chromatographic column with an effective velocity given as

$$u_i = \tau_i v \quad (5)$$

where u_i is the effective velocity of component i and v is the average velocity of the mobile phase. Therefore, if two solutes have different distribution constants, they will move down the length of the column with different effective velocities. The solutes will elute from the column in discrete bands. In a real system, many molecules are present, and these will disperse about an average value. For components with linear adsorption isotherms, the result for each component approximates a Gaussian distribution whose width depends on the molecular dispersion of that component.[17]

Calculating the Distribution Constant from the Chromatogram

With a suitable detector placed at the exit of the column, the concentration of solute in the mobile phase can be monitored continuously. A plot of the concentration vs. time is called a chromatogram. Fig. 6 shows a typical chromatogram.

The time at which a peak appears in the detector response is called the retention time of the solute; it is the time it takes a given solute to travel the length of the column. For a Gaussian peak distribution, the retention time corresponds to the peak center. Mathematically, the retention time of component i is simply

$$t_{Ri} = \frac{L}{u_i} \quad (6)$$

where L is the column length. If a compound does not partition into the stationary phase, then it will travel down the length of the column with a velocity equal to the mobile phase. The time that it takes this unretained compound is called the holdup time, and is defined as

$$t_0 = \frac{L}{v} \quad (7)$$

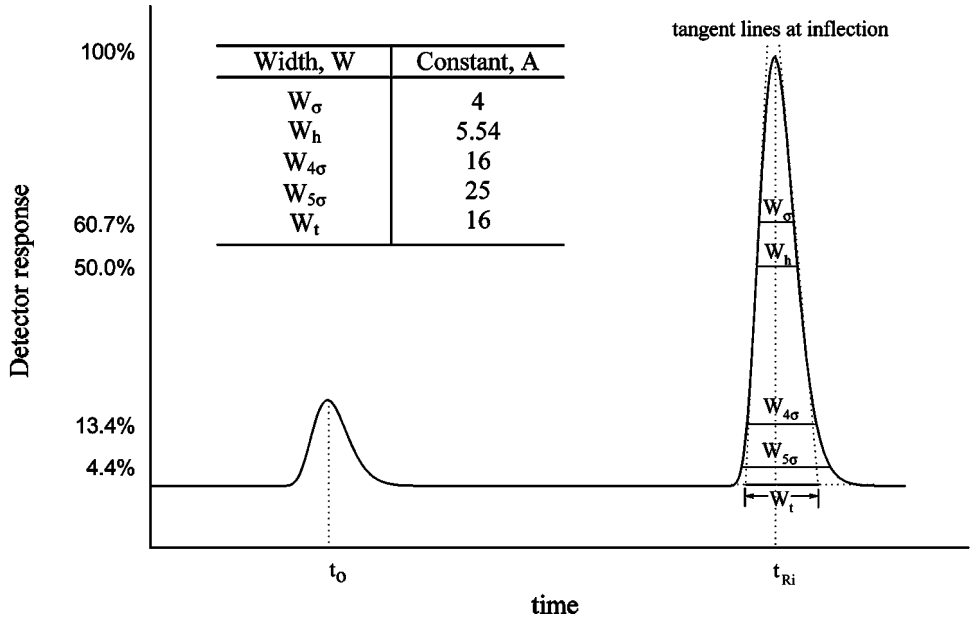


Fig. 6 Chromatogram for a two-component mixture with Gaussian responses. The small peak represents an unretained compound. The large peak is used to illustrate common definitions of peak width.

Combining Eqs. (4)–(7), the distribution constant for a given solute i is given as

$$K_i = \frac{V_M}{V_S}\left(\frac{t_{Ri} - t_0}{t_0}\right) \quad (8)$$

The ratio in brackets is called the capacity factor, k; it is used commonly as a measure of the ability of the stationary phase to retain a solute. Using Eq. (8), the distribution constant of component i can be calculated from readily measurable quantities.

To standardize reported retention times for a given column, manufacturers often report adjusted or relative retention factors rather than distribution constants. The adjusted retention time of a given compound is reported relative to the retention time of an unretained component. The relative distribution coefficient is reported relative to the distribution coefficient of a reference compound:

$$t'_{Ri} = t_{Ri} - t_0 \quad (9)$$

$$K'_i = \frac{K_i}{K_{i,REF}} = \frac{t'_i}{t'_{i,REF}} \quad (10)$$

Column Efficiency

Column efficiency refers to the amount of band broadening that occurs when a compound passes through the column. The typical Gaussian shape of the chromatographic band results from the random motions of the molecules making up the band as it moves down the column. The result of these random individual motions is a symmetric spread of velocities around the mean value. This spreading is referred to as molecular diffusion.

Chromatodiffusion also contributes to this band broadening. At the center of a component band, the chemical potential of that component is high. At the edges of the band, it is low. There is a net diffusion of molecules from the center of the band to the edges because of this chemical potential difference.

Nonuniform column packing also affects band broadening. Different molecules travel along various paths through the column. Depending on the length of their respective path, two molecules of the same component will have varying retention times.

The plate height concept applied to chromatographic separations

A good chromatographic method will usually produce peaks with adequate separation and symmetry to allow the user to treat all peaks as Gaussian distributions. This treatment leads to a simple definition of the number of theoretical plates for a given separation.[5]

$$N_i = A\left(\frac{t_{Ri}}{W_A}\right)^2 \quad (11)$$

where N_i is the number of theoretical plates for the separation of species i, W is the width of the peak in time units, and A is a constant based on the definition of W. Fig. 6 summarizes the common definitions of W, along with the corresponding value of the constant A.

From the usual definition of equilibrium-staged separations, the plate height is calculated as $H = L/N$. The efficiency of chromatographic columns increases as the number of plates increases and as the plate height decreases. Both of these values are used widely in the literature as measures of column performance.

Effect of velocity on plate height

Some of the variables that affect column efficiency can be controlled to improve separations. The extent of band broadening, and thus column efficiency, depends on the amount of contact time between the mobile and stationary phases. Therefore, column efficiency depends on the flow velocity of the mobile phase. Plots of plate height vs. velocity are called van Deemter curves. The plots are fit to the van Deemter equation

$$H = A + \frac{B}{v} + Cv \quad (12)$$

where the constants A, B, and C are dependent on uniformity of packing, molecular diffusion, and chromatodiffusion, respectively. (For capillary columns, $A = 0$.) Fig. 7 shows typical van Deemter curves for GC and LC.

For LC, the second term in Eq. (12) is often negligible because liquid diffusion coefficients are low ($\sim 10^5$ times lower than for gases). Furthermore, for LC, the parameter A depends on v, and A and C are coupled.[18] For these reasons, LC plate heights increase with velocity for all reasonable operating values of v. One might be tempted to use low velocities in LC to achieve low plate heights; however, lower velocities translate into longer retention times.

For GC, the optimum velocity corresponds to the minimum plate height, which occurs at $v = (B/C)^{1/2}$. The first term in Eq. (12) is independent of velocity and depends only on how efficiently the column is packed. To obtain a small H and, thus, efficient separation, uniform packing is essential.

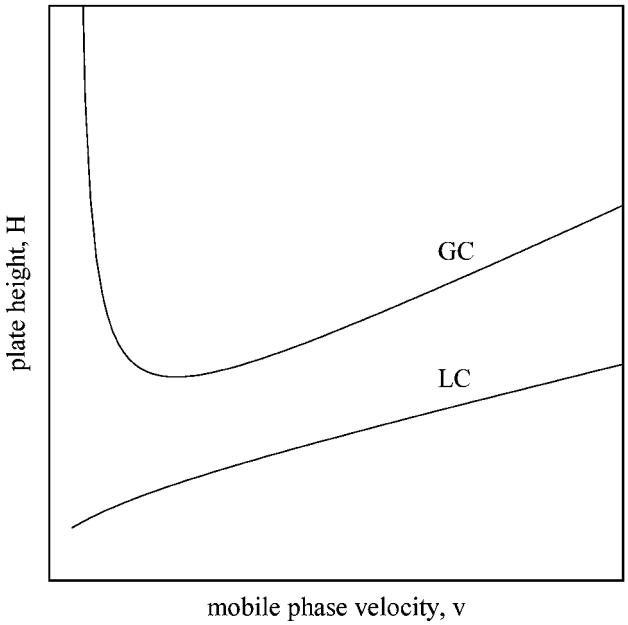


Fig. 7 Typical van Deemter plots for GC and LC.

Column Resolution

The resolution R_s of a column describes the ability to separate two compounds, 1 and 2; assuming that both peaks can be treated as Gaussian distributions, a resolution can be calculated:

$$R_S = \frac{2(t_{R1} - t_{R2})}{W_1 + W_2} \quad (13)$$

In Eq. (13), W_1 and W_2 are measured at the base of each peak. For Gaussian peak distributions, these widths equal 4σ, and a column resolution of 1.0 means that there is 4.6% overlap of peak 1 in peak 2 and vice versa; a resolution of 1.5 means there is 0.3% overlap between the two peaks. Generally, the resolution for a given separation can be increased by lengthening the column and, thus, increasing the number of theoretical plates. Here again, the trade-off for this increased resolution is a longer retention time.

Effect of the distribution constant on resolution

Given information about the distribution constants for two solutes, the column length necessary to provide a given resolution between them can be calculated, with reasonable assumptions, from the definition of resolution.[17]

$$L = 16R_S^2 H\left(\frac{K_2/K_1}{K_2/K_1 - 1}\right)^2\left(\frac{1 + k_2'}{k_2'}\right)^2 \quad (14)$$

The result demonstrates that the distribution constants of 1 and 2 must have different values for separation to occur. If they were exactly the same, then resolution would have to be zero.

Asymmetric Peak Distributions

Real chromatographic peaks generally have some degree of asymmetry. For asymmetric peaks, it is a mistake to use equations based on the Gaussian distribution function to calculate peak parameters. More accurate functions are available for use with asymmetric peaks, including the commonly used exponentially modified Gaussian function (EMG). Di Marco and Bombi present tabulated information on empirical functions for characterizing chromatographic peaks.[19] Foley and Dorsey give relations between the measurable retention time, peak width, asymmetry factor, and chromatographic figures of merit, including the number of theoretical plates, for the EMG model.[20]

CONCLUSIONS

Chromatography is used both as an analytical method and an industrial unit operation to separate and/or identify the chemical components in complex mixtures. Its basis is the partitioning of solutes between a moving fluid phase and a stationary phase. The fluid phase can be either a gas or a liquid, while the stationary phase is either a solid or a liquid applied to a solid substrate.

REFERENCES

1. Johnson, E.L.; Stevenson, R. *Basic Liquid Chromatography*; Varian Associates: Palo Alto, 1978.
2. Skoog, D.A.; West, M.W.; Holler, F.J. *Analytical Chemistry, An Introduction*, 6th Ed.; Saunders College Publishing: Philadelphia, 1994.
3. Humphrey, J.L.; Keller, G.E., II. *Separation Process Technology*; McGraw-Hill: New York, 1997.
4. Ettre, L.S. New, unified nomenclature for chromatography. Chromatographia **1994**, *38* (7–8), 521–526.
5. Grob, R.L. Theory of gas chromatography. In *Modern Practice of Gas Chromatography*, 3rd Ed.; Grob, R.L., Ed.; Wiley: New York, 1995; 51–121.
6. Henrich, L.H. Detectors and data handling. In *Modern Practice of Gas Chromatography*, 3rd Ed.; Grob, R.L., Ed.; Wiley: New York, 1995; 265–321.

7. Klee, M.S. Optimizing separations in gas chromatography. In *Modern Practice of Gas Chromatography*, 3rd Ed.; Grob, R.L., Ed.; Wiley: New York, 1995; 225–261.
8. Barry, E.F. Columns: packed and capillary/column selection in gas chromatography. In *Modern Practice of Gas Chromatography*, 3rd Ed.; Grob, R.L., Ed.; Wiley: New York, 1995; 123–224.
9. Wheals, B.B. Detectors for HPLC. In *Techniques in Liquid Chromatography*; Simpson, C.F., Ed.; Wiley: New York, 1982; 121–140.
10. Byrne, S.H., Jr. Detectors in liquid chromatography. In *Modern Practice of Liquid Chromatography*; Kirkland, J.J., Ed.; Wiley: New York, 1971; 95–124.
11. Seader, J.D.; Henley, E.J. *Separation Process Principles*; Wiley: New York, 1998.
12. Horvath, C. Bonded phase chromatography. In *Techniques in Liquid Chromatography*; Simpson, C.F., Ed.; Wiley: New York, 1982; 229–301.
13. Hill, E.A. Design of media for high performance chromatography of large biomolecules. In *Techniques in Liquid Chromatography*; Simpson, C.F., Ed.; Wiley: New York, 1982; 367–377.
14. Synder, L.R. The role of the mobile phase in liquid chromatography. In *Modern Practice of Liquid Chromatography*; Kirkland, J.J., Ed.; Wiley: New York, 1971; 125–157.
15. Scott, R.P.W. Selectivity in chromatographic separations. In *Techniques in Liquid Chromatography*; Simpson, C.F., Ed.; Wiley: New York, 1982; 141–183.
16. Lindsey, S. *High-Performance Liquid Chromatography*; Wiley: New York, 1987.
17. Simpson, C.F. An introduction to liquid chromatography and some fundamental relations. In *Techniques in Liquid Chromatography*; Simpson, C.F., Ed.; Wiley: New York, 1982; 1–29.
18. Karger, B.L. The relationship of theory to practice in high-speed liquid chromatography. In *Modern Practice of Liquid Chromatography*; Kirkland, J.J., Ed.; Wiley: New York, 1971; 3–53.
19. Di Marco, V.B.; Bombi, G.G. Mathematical functions for the representation of chromatographic peaks. J. Chromatogr. **2001**, *931*, 1–30.
20. Foley, J.P.; Dorsey, J.G. Equations for calculation of chromatographic figures of merit for ideal and skewed peaks. Anal. Chem. **1983**, *55*, 730–737.

BIBLIOGRAPHY

Ganetsos, G., Barker, P.E., Eds. *Preparative and Production Scale Chromatography*; Dekker: New York, 1993.

Grob, R.L., Ed. *Modern Practice of Gas Chromatography*, 3rd Ed.; Wiley: New York, 1995.

Mori, S.; Barth, H.G. *Size Exclusion Chromatography*; Springer-Verlag: Berlin, 1999.

Simpson, C.F., Ed. *Techniques in Liquid Chromatography*; Wiley: New York, 1982.

Weiss, J. *Ion Chromatography*, 2nd Ed.; VCH Verlagsgessellschaft mbH: Weinheim, Germany, 1995.

Coal–Water Slurries

S. Komar Kawatra
Department of Chemical Engineering, Michigan Technological University, Houghton, Michigan, U.S.A.

INTRODUCTION

A slurry is defined as a fluid mixture of a pulverized solid with a liquid, usually water.

The term ''coal slurry,'' therefore, includes all mixtures of coal and water that can flow as a fluid. There are four main types of high-percent-solid coal slurries discussed in the literature, with different characteristics depending on the application:

1. *Coal–water slurry fuels*: High-percent-solid slurries that are stable against settling, so they can be pumped, stored, handled, and burned much like heavy fuel oil. Coal–water fuels have about half the heating value of fuel oil, and because of their water content they burn cooler than pulverized coal, leading to lower production of nitrogen oxides.

2. *Coal slurry pipelines*: Slurry is used for transporting the coal through pipelines, but the coal is dewatered at the end before utilization. The slurry is unstable, and the particles are kept in suspension by the turbulence of the flow as they travel through the pipeline. The particles rapidly settle out of suspension once the slurry stops moving.

3. *Gasifier feedstock*: Integrated gasification/combined cycle (IGCC) plants use coal slurry as their feedstock. This simplifies transport of the coal and provides the water needed along with the coal in the gasification reaction. The slurries used must flow and atomize well in the reactor, but do not need to have long-term stability against settling.

4. *Coal slurry waste*: This is fine coal processing tailings that contain a significant amount of coal and enough water so that they can flow. These slurries can be processed to recover fine coal particulates, or disposed of in tailings impoundments. These slurries can easily flow over large areas and cause considerable damage if the impoundment dam breaks. Coal slurry wastes are very high-percent solids and are described as having ''about the consistency of molasses.''

The properties of coal slurries that are of most interest are: 1) the viscosity/rheology, which controls the ease of pumping and atomization; 2) the stability against settling, which controls the ability of the slurry to be stored in tanks without the solids forming a hard mass that clogs the system; and 3) the solid concentration and coal type/composition, which determines the heating value per unit volume of slurry.[1]

Of the four applications for coal slurries given here, coal slurry fuels have the most stringent requirements. Hence, the following discussion of slurry properties is with reference to coal slurry fuels.

RHEOLOGY OF COAL SLURRIES

There are two competing considerations for the flow properties (rheology) of coal-water slurry fuels. The first is that the slurry must have a low viscosity (preferably less than 2000 centipoise) at high shear rates so that it can be pumped and atomized readily.[1] The second is that it should be sufficiently viscous so that the particles will not settle out during storage, as otherwise the settled solids will clog the pipes leading from the storage tank. There are two factors that contribute to prevention of slurry settling: development of a yield stress and thixotropy.

In a slurry that develops a yield stress, the slurry does not begin to flow until a particular level of shear stress is applied to it. These types of slurries are classified as either pseudoplastic with yield, or Bingham plastic, as shown in Fig. 1. Pseudoplastic and Bingham plastic slurries have a high apparent viscosity (ratio of the shear rate to the shear stress) at low shear rates, but at higher shear rates the apparent viscosity decreases. Coal slurry rheology can generally be modeled using Herschel and Buckley's yield-power law:[2]

$$\tau = \tau_0 + K\gamma^n$$

where τ is the shear stress (in Pa), τ_0 is the yield stress required to initiate slurry flow (in Pa), γ is the shear rate (sec^{-1}), n is the flow behavior index (equals 1 for Newtonian slurries, <1 for pseudoplastic slurries, and >1 for dilatant slurries), and K is the consistency number (equal to the apparent viscosity for Newtonian slurries).

Slurries that exhibit a yield stress will resist settling when stationary, which is a low shear condition, but

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008028

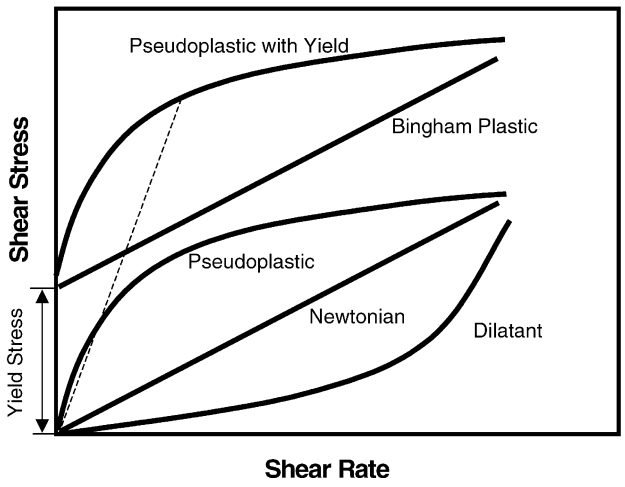


Fig. 1 Classes of rheological behavior that can be shown by coal slurries, as they appear when plotted on a shear rate/shear stress graph. It is desirable for coal slurries to be Bingham plastic or pseudoplastic with yield, as such slurries flow readily at high shear rates (such as during pumping or atomization), while remaining stable against settling at low shear rates because of their yield stress. Dilatant slurries are completely unsuitable for coal slurry applications because they are extremely difficult to pump.

will flow readily when a higher shear force is applied, such as by a pump or an atomizer.

In a thixotropic slurry, the slurry will, when left undisturbed, develop a three-dimensional network structure where particles attach loosely to each other in an extremely weak gel. This holds the solid particles suspended indefinitely. The shear stress required to cause the thixotropic slurry to flow can then be broken into two parts:[3]

$$\tau = \tau_e + \tau_s$$

where τ is the total shear stress (in Pa) required to produce slurry flow, τ_e is the equilibrium shear stress (in Pa) reached after agitating the slurry for an extended time, and τ_s is the structural shear stress (in Pa) resulting from the formation of the structural network.

The structure can then be broken up by stirring, so that the slurry can be refluidized to a free-flowing state. The structural shear stress of the thixotropic slurry decreases with time during agitation according to the following expression:[3]

$$d\tau_s/dt = [\tau_{s\infty} - \tau_s]/\lambda_s - K\eta^{1/2}\gamma\tau_s$$

where $d\tau_s/dt$ is the rate of change of structural shear stress with respect to time, τ_s is the structural shear stress (in Pa), η is the apparent viscosity (in Pa sec), γ is the shear rate (sec^{-1}), $\tau_{s\infty}$ is the final structural shear stress achieved under shear-free conditions, λ_s is the time constant for buildup of structural stress, and K is the proportional constant for the rate of structure breakdown.

Solid Loading

The rheological properties of coal slurries depend on the quantity of solids suspended in the liquid and on the physical and surface chemical properties of the solid particles. The viscosity of the slurry will typically increase gradually with increasing solid loading until a critical point is reached where interparticle friction becomes important. Beyond this point, the viscosity will rapidly increase until the slurry ceases to flow, as shown in Fig. 2. The value of the critical solid loading will vary considerably between coal slurries, depending on the physical and chemical properties of the particles. For most coal slurries, and particularly for coal–water slurry fuels, it is important to be able to make the solid loading of the slurry as high as possible so that the energy content per liter of fuel is maximized, while keeping the viscosity low enough so that the slurry can be pumped economically.

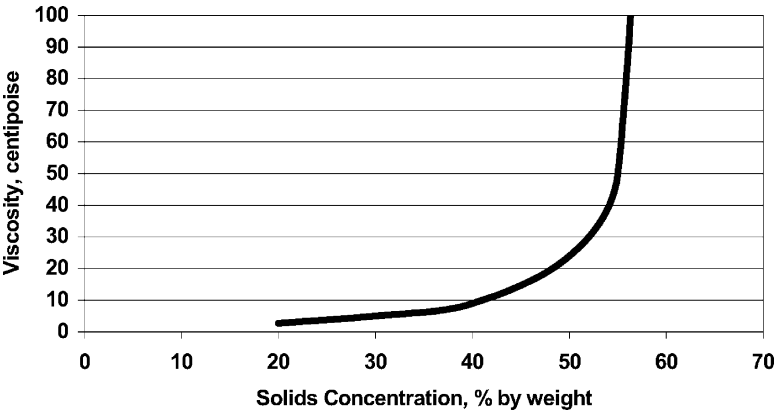


Fig. 2 Effects of solid concentration on the viscosity of a particular coal slurry. Beyond a certain limiting concentration, the viscosity rapidly increases to extremely high values. (From Ref.[4].)

Particle Size Distribution

The size distribution of particles will control the amount of liquid needed to fluidize a given quantity of coal. In general, a fine size distribution will produce a more viscous slurry than a coarse size distribution at the same wt% solids, and the fine particles will produce a more non-Newtonian rheological curve. This can be seen in the laboratory results shown in Fig. 3, which compares a coarse coal slurry to a fine coal slurry. It is clearly seen that the fine slurry is much more viscous, its pseudoplastic character is very pronounced, and its yield value is high, while the coarse coal slurry is clearly a Bingham plastic.[5]

A graded size distribution, where fine particles fill the interstices between coarse particles as shown schematically in Fig. 4, will minimize the amount of void space that must be filled by fluid, and so will reduce the quantity of liquid needed to produce a flowable slurry. The best size distributions for this purpose have proven to be multimodal distributions, made up of several fairly narrow size fractions. An example of such a size distribution is shown in Table 1.[1] To achieve such a multimodal size distribution, coal slurry production facilities are designed to generate several coal streams, which are each ground and sized to the desired particle sizes and then combined to give the proper size distribution.

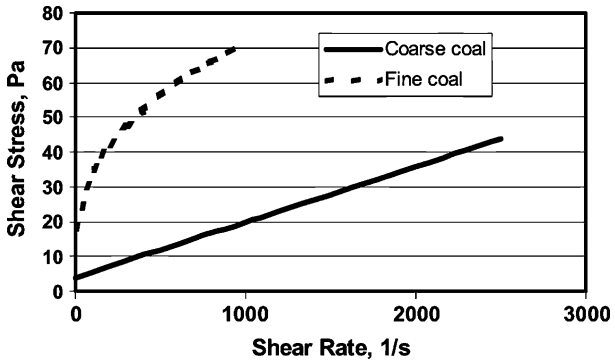


Fig. 3 Comparison of the rheological curves for a fine coal slurry (80% passing 34 μm, top size 100 μm, 52 wt% solids) and for a coarse coal slurry (58 wt% solids). Neither slurry used any additives. Because it is extremely difficult to measure the rheology of unstable slurries with conventional rheometers, these results were obtained using a continuous-pressure-vessel rheometer, which was specially designed for this purpose. The fine coal curve is the average of 10 measurements and the coarse coal curve is the average of 5 measurements, and the standard error of the shear rate measurements was approximately 1.0 Pa for these slurries. The fine coal slurry is clearly pseudoplastic with a yield value of approximately 18 Pa, while the coarse coal slurry is Bingham plastic with an estimated yield value of 4 Pa.

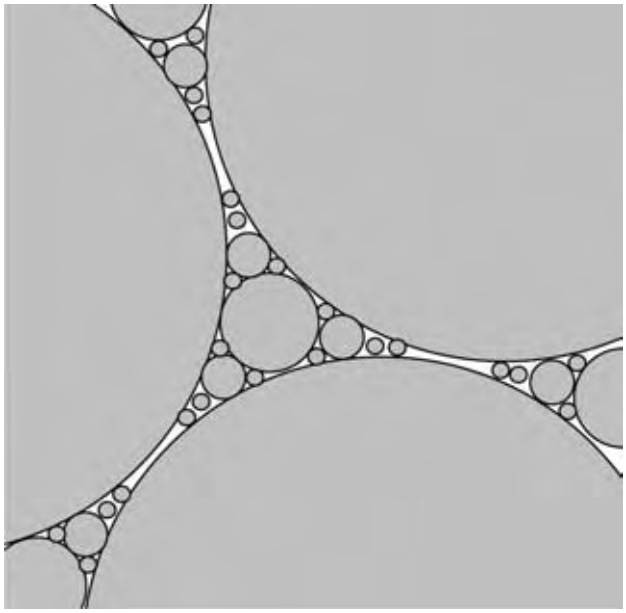


Fig. 4 Graded size distribution to maximize the packing fraction of ideal spherical particles, and thereby minimize the amount of water needed to fluidize the slurry. In this illustration, there are four different sizes of particles present. A similar principle will maximize the packing fraction of real, irregularly shaped particles.

An example of how slurry rheology is affected by combining a ''coarse'' size distribution (208–279 μm) with a ''fine'' size distribution (smaller than 45 μm) can be seen in Figs. 5 and 6. Adding increasing amounts of coarse particles made the suspensions examined less viscous until approximately 40% of the solids were coarse particles. Increasing the proportion of coarse particles beyond that level produced no further viscosity decrease and tended to destabilize the slurry so that the particles would rapidly settle out.[6]

Surfactants and Dispersants

The interparticle friction in the slurry can be reduced through the use of surface-active reagents (surfactants). The surfactants, therefore, alter the viscosity and overall rheology of the slurry. It is important to

Table 1 Example of a multimodal size distribution to produce a very high solid loading in a coal slurry (This results in a powder with only 24% pore volume)

Size fraction (μm)	Total solids by weight (%)
701–833	60
88–104	30
38–43	8
<25	2

(From Ref.[1].)

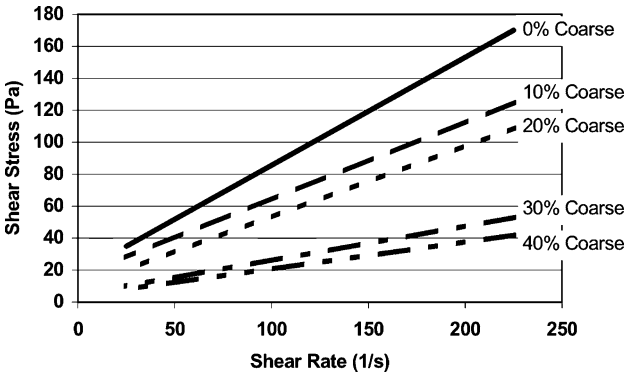


Fig. 5 Changes in the rheology of 40 wt% coal suspensions as the proportion of coarse particles (208–279 μm) to fine particles (<45 μm) is increased. As the fraction of coarse particles increases, the viscosity drops and the yield stress is decreased. (From Ref.[6].)

remember that coal is not a single tightly defined substance. It covers a wide range of fossilized high-carbon solids, which vary a great deal in composition, heating value, and chemistry. High-rank, high-volatile coals tend to be more hydrophobic than low-rank or low-volatile coals, which contain larger quantities of moisture and oxygen in their structure. It is therefore critical that the surfactants used be selected to match the properties of the coal being used for the production of the slurry, as surfactants that are highly successful in one application may be largely ineffective in another.

Because of their structure and chemistry, surfactants are chemicals that tend to concentrate at surfaces in preference to dissolving in the volume of the liquid. They can, therefore, be used to selectively alter the surfaces of particles, changing the way they interact with each other.

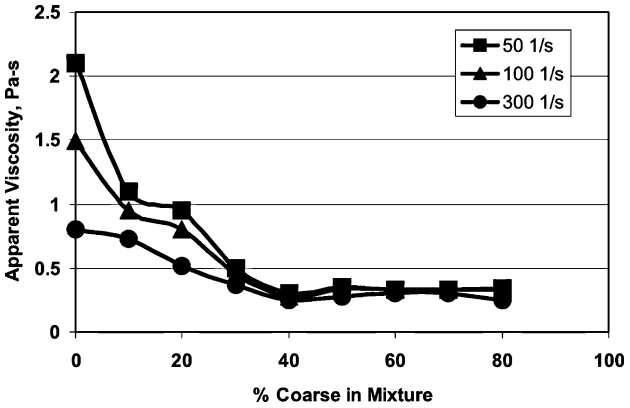


Fig. 6 Apparent viscosity of a 50% solid coal slurry as the proportion of coarse particles (208–279 μm) to fine particles (<45 μm) is increased, at three different shear rates. The viscosity was minimized at 40% coarse particles. (From Ref.[6].)

In the case of coal particles suspended in water, there are two effects that control interparticle interactions: hydrophobic properties and electrical charge effects. Strongly hydrophobic coal particles will tend to flocculate because of the reduced surface energy of a coal–coal interface compared to a coal–water interface. Surfactants that increase hydrophobicity will therefore tend to increase viscosity owing to increased interparticle interaction. The electrical charge on the particle surfaces, which arises by ion exchange with the fluid, will also affect the viscosity. A low magnitude of charge, within approximately ±5 mV of zero net charge, will allow particles to interact and flocculate readily, resulting in a high viscosity and significantly non-Newtonian behavior. A high magnitude of charge, exceeding –20 mV, will cause particles to repel each other, minimizing flocculation and reducing viscosity.

There are hundreds of different surfactants that have been considered for control of coal slurry rheology. The majority of these reagents fall broadly into the three classes described in Table 2: anionic surfactants, nonionic surfactants, and polysaccharides. Anionic surfactants are those that have a negative charge in solution. When they adsorb on the coal surface, they provide it a net negative charge, which increases interparticle repulsion. Nonionic surfactants and polysaccharides have no net charge in solution and are believed to form a more physical barrier that prevents particles from interacting with each other.

Other Factors Affecting Rheology

The rheology of a coal-water slurry is also affected by the composition of the coal. It has been determined that increases in coal ash content lead to increasing viscosity and that slurries are more viscous at pH 6 than at pH 8.[7] Many coals, particularly low-rank or partially oxidized coals, contain humic acids, which act as surfactants and can affect the properties of the slurry.[8] The presence of these humic acids helps to decrease slurry viscosity and increase the maximum particle loading. Additionally, the viscosity is affected by the temperature of the slurry, with increasing temperature tending to decrease the slurry viscosity. It has also been reported that high shear agitation can reduce the viscosity of the coal slurry, apparently by improving the dispersion of surfactants.[9]

APPLICATIONS OF COAL SLURRIES

Slurry Fuels

As petroleum-based fuels are being depleted, it is expected that coal-derived fuels such as coal slurry

Table 2 Types of surfactants that have been used for preparing highly concentrated coal-water slurries: Many of these surfactants are used in combination with formulations that depend on the characteristics of the coal

Anionic surfactants	Nonionic surfactants	Polysaccharides
Ethylene diamine tetra-acetic acid (EDTA)	Polyethylene glycol ethers	Carboxymethyl cellulose salts
Calcium lignosulfonate	Polypropylene glycol	Cellulose
Ammonium lignosulfonate	Ethoxylated/propoxylated alcohols	Wood particles
Sodium lignosulfonate	Alkylphenyl decaethylene glycol ether	Uronic acid
Sodium tripolyphosphate	Polyethyleneimine	Hydroxypropyl cellulose
Polyalkylene oxide	Propylene oxide polymers	Polysaccharose (produced by lipomyces micro-organisms)
Polysodium styrene sulfonate	Ethylene oxide polymers	
Polyacrylic acid	Phenol butyl naphthalene polymers	Cellulose ether
Sodium allyl sulfonate	Methacrylic acid polymers	
Polyacrylamide	Butyric acid lauryl ethylene diamine salt	
Polyoxyethylene nonyl phenol ether		
Polyvinyl sulfonic acid		
Sulfonated benzene sodium salt		
Naphthalene sulfonic acid		
Tallow soap		
Isobutylene		
Maleic anhydride		
Sodium humate (extracted from brown coal)		

(From Ref.[1].)

fuels will become more important. These fuels are attractive mainly because they are made from coal, which is of low cost compared to petroleum products, and can be pumped, stored, and burned much like fuel oil. The demonstrated coal reserve base in the United States is 508 billion tons, with predicted reserves as high as 1.7 trillion tons.[10] At the current rate of consumption (approximately 1 billion tons/year) this would last for 500–1700 yr, which means that coal, and therefore coal slurry fuels, will be available long after petroleum deposits are depleted. Most of the activity in coal slurry fuel development has used highly cleaned high-rank coals, which can be produced using advanced physical, chemical, and biological techniques to remove ash-forming minerals and sulfur.[11] In addition, processes have been developed to upgrade low-rank coals so that they can be used to produce coal slurries with a high heating value.[12]

Coal slurry fuel production plants have been constructed for producing coal-water slurry fuels. These plants consist of facilities for grinding highly cleaned coal to the desired fineness, adding stabilizing reagents, and increasing the solid content of the slurry to approximately 65–70% solids by weight. A schematic of a typical plant is shown in Fig. 7. The key design features of such a plant are:

1. *Particle size control*: The plant pictured produces a multimodal size distribution, consisting of relatively large particles from the first stage of grinding, an intermediate size of particles from the second stage of grinding, and particles from the hydrocyclone oversize product that are reground to a very fine size when they are returned to the second grinding stage.

2. *Water content control*: Grinding and size classification of the coal require a slurry with considerably more water than is desired in the final fuel. Hence, filters must be used to remove this excess water once the particles reach their target size.

3. *Reagent addition*: Before the reagents are added, the material coming from the drum filter will be extremely viscous. Pug mills, which can handle a very high solid content, are, therefore, used to combine the reagents with the coal slurry both to combine the desired particle size fractions and to uniformly disperse the reagents through the coal.

Coal slurry fuels are of interest for the following markets:

Replacement of Heavy Fuel Oil: The original concept for coal-water slurry fuels was as a direct replacement for heavy #5 and #6 fuel oils. The intent was that power plants and industrial heating operations that were using these fuel oils could switch directly to a coal-water slurry with a minimum of modification to their combustors. This requires a very

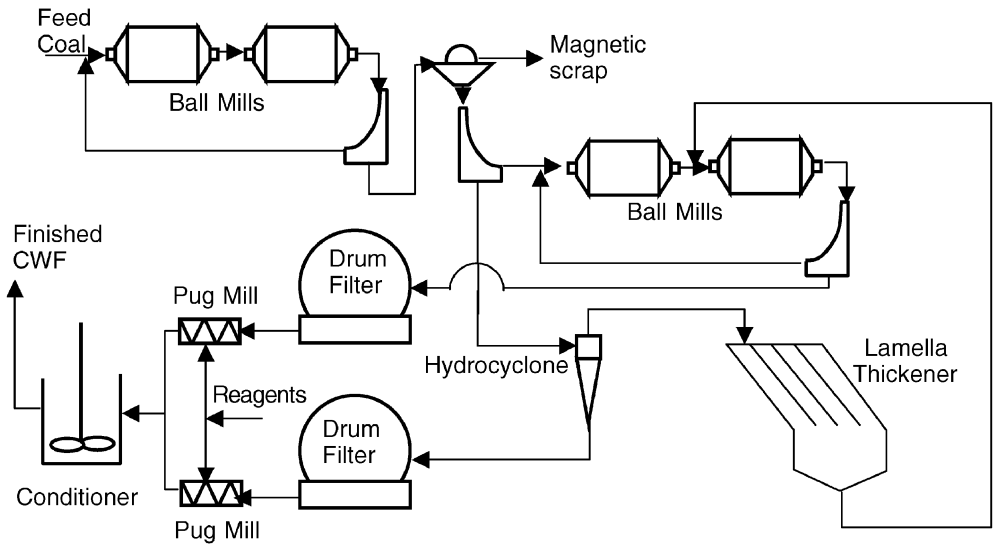


Fig. 7 Simplified schematic of the Beijing Coal Water Fuel Plant, China. Clean coal is first ground in multiple stages to produce a graded size distribution that will fluidize with the minimum amount of water. Drum filters are used to remove excess water from the slurry, and surfactant reagents are combined with the slurry using pug mills. This plant processes 35 metric tons of coal per hour and produces a slurry that is 65% coal by weight. (From Ref.[22].)

low-ash coal to prevent fouling of the burners and is most attractive when fuel oil prices abruptly increase. The disadvantage of coal-water slurry fuels in this application is that they only provide about half the energy per unit volume as the fuel oil does. Currently, there do not appear to be any industries that see these fuels as an attractive alternative, and they are not being sold commercially for this purpose.

Supplement for Pulverized Coal Combustors: The water content of coal-water slurry fuels has made them attractive for control of nitrogen oxide emissions from pulverized coal combustors at power plants. If the slurry fuel is injected as a supplement for the main fuel (dry coal), the water lowers the combustion temperature and prevents nitrogen oxides from forming, while still allowing complete combustion of the coal. As nitrogen oxides are troublesome to remove or destroy once they are formed, it is generally preferable to prevent their formation at the source. This technology is currently in use for low-NO_x coal combustors. It has also been determined that, when coal-water slurry is added in pulverized coal combustors, a greater fraction of the inorganic hazardous air pollutants (including mercury) became trapped in the bottom ash rather than escaping with the flue gases.[13]

Injection into Fluid-Bed Combustors: Similarly, fluid-bed combustors can be operated with part of the fuel supplied as a coal-water slurry. In Italy, coal-water slurries are currently considered to be a standard fuel for this application.[14]

Diesel Engines: Coal slurry fuels have been proposed as diesel engine fuel as a direct replacement for fuel oil, and diesel engines designed for this purpose have been developed and successfully operated.[15] The main issues with using coal slurry in this application are: first, to prevent clogging of the injectors while still adequately atomizing the slurry; second, to minimize wear of the piston rings caused by abrasive particles in the coal; and third, to control the higher level of particulate emissions compared to conventionally-fueled diesel engines. This application requires particularly low-ash coal for production of the slurry to minimize the quantity of ash formed in the engine.

Pipeline Transport

A coal slurry pipeline is a system for transporting solid coal particles in a liquid carrier. Long-distance coal slurry pipelines are an alternative to railroad transport, and their practicality and economics are largely dependent on land ownership, terrain, water availability and water contamination concerns, political considerations, and coal demand.[16] An example of pipeline transport is the 440 km Black Mesa pipeline, which was completed in 1970 and is currently the only long-distance coal slurry pipeline operating in the United States.[17] Slurry pipelines are also used over shorter distances to transport material within a processing facility.

Unlike coal slurry fuels, which are utilized in a fluid form, coal transported by pipeline is used as a solid after dewatering. As a result, it does not need to be stable when stored in tanks, and it is not so critical that its solid loading be maximized. The following variables are important for a pipeline system:

- Solid concentration range
- Flow velocity range
- Pipe diameter
- Friction losses

These are, in turn, controlled by the size distribution and density of the coal particles. To maintain homogenous flow, which is necessary to minimize wear of the pipeline, a coal slurry should have the following properties, assuming a coal specific gravity of 1.4:[4]

- Maximum particle size: 2400 μm
- Percentage finer than 44 μm: 20%
- Slurry concentration: 50 wt%.

To keep the particles in suspension, the flow should be at least 0.15 m/sec faster than either: 1) the critical deposition velocity of the coarsest particles, or 2) the laminar/turbulent flow transition velocity. The flow rate should also be kept below approximately 3 m/sec to minimize pipe wear. The critical deposition velocity is the fluid flow rate that will just keep the coarsest particles suspended, and is dependent on the particle diameter, the effective slurry density, and the slurry viscosity. It is best determined experimentally by slurry loop testing, and for typical slurries it will lie in the range from 1 m/s to 4.5 m/sec. Many empirical models exist for estimating the value of the deposition velocity, such as the following relations, which are valid over the ranges of slurry characteristics typical for coal slurries:[18]

$$F_L = V_c/[2gD(\rho_s - \rho z_f)/\rho_f]^{1/2}$$

where

$$F_L = \exp[0.165 - 0.073C_D - 12.5K_2]$$
$$K_2 = [K_1 - 0.14]^2$$
$$K_1 = [\mu_L/\rho_L]^{2/3}/g^{1/3}d_{50}$$

C_D = particle drag coefficient
d_{50} = mass median diameter of $+74\,\mu m$ particles (m)
μ_L = viscosity of liquid (Pa sec)
V_c = deposition velocity (m/sec)
g = gravitational acceleration (m/sec^2)
D = pipe diameter (m)
ρ_s = density of solid particles coarser than $74\,\mu m$ (kg/m^3)
ρ_L = density of fluid (kg/m^3)
$\rho_f = [\rho_s C_f + (1 - C_t)\rho_L]/[1 - C_t + C_f]$
$C_t = C_f + C_r$
C_f = volume fraction of $-74\,\mu m$ particles in slurry
C_r = volume fraction of $+74\,\mu m$ particles in slurry

Once the coal slurry has been transported to its destination by pipeline, it is dewatered for utilization.[19] Chemical additives are not normally used in slurry pipelines, as they would be an unnecessary added cost and a possible source of water contamination. It is important to monitor the quality of the water removed from the slurry, as it could contaminate waterways if it dissolves harmful materials before being released.

Coal Gasification

For integrated gasification combined cycle (IGCC) plants, coal slurries are pumped under pressure into the reactor, where they react with water to produce combustible gases, particularly methane. There are two reasons for using coal slurries for this application rather than dry coal: 1) a fluid slurry can be easily pumped into a pressurized reactor, whereas it is much more difficult to convey dry solids against a pressure gradient, and 2) the gasification reaction requires water in any case; hence, it is simplest to introduce water along with the coal particles.

The requirements for coal slurry intended to be used in IGCC are similar to those for pipeline transport, with the added requirement that, as for coal-water fuels, the slurry needs to atomize easily in the reactor. Atomization has been found to be easier (producing smaller-diameter droplets at lower jet pressures) as the size of the coal particles in the slurry increases, because of weaker capillary forces holding the slurry together and decreased interparticle friction.[20] As the capillary forces and interparticle friction are also responsible for increasing the viscosity in coal slurries, it is clear that low-viscosity slurries will atomize more readily than high-viscosity slurries.

Coal Slurry Wastes

The bulk of the fine coal that makes up coal slurry wastes is produced in the mine by the miner cutting heads. These fine particles are typically mixed with

large quantities of mineral fine substances, particularly clays, to the extent that the fine coal has a very high ash content. Additional fine particles are produced by rotary breakers and other size reduction equipment that is in use in most coal-washing plants. Processing methods exist that can be used to separate coal from the mineral matter in these fine coal slurries, specifically, froth flotation techniques. However, not all plants have the facilities to recover this fine coal, either because their coal is not suited to froth flotation processing, or because they have difficulty meeting ash and moisture specifications with fine coal produced by this means. As a result, many plants must dispose of their fine coal slurries as waste.

Large volumes of coal waste slurry are disposed of in surface impoundments. As the slurry remains easily liquefied long after being impounded, it represents a serious hazard if the impoundment fails. An example of this occurred in Martin County, KY, in 2000, when a breakthrough in a coal slurry impoundment released 245 million gallons of slurry into 75 mi of streams, and caused in excess of $37 million in damages.[21]

Unlike other coal slurry applications where the slurry must be kept fluid and pumpable, coal slurry in impoundments needs to be stabilized to prevent flow. Treatments of coal waste slurry, therefore, concentrate on promoting complete settling, strong interparticle interactions and high viscosities so that impoundment breaks will not release the material as a damaging flow.

CONCLUSIONS

Coal slurries are of great importance in a wide variety of coal-handling situations, as the slurry form makes the coal easy to handle by pumping with minimal dust. Highly concentrated slurries that can be burned directly as fuel make it possible to eliminate the need for handling dry coal in the combustion process. The primary considerations in producing and handling coal-water slurries are managing the slurry rheology through manipulation of particle size and surface chemistry and minimizing the water content needed to maintain fluidity. Currently, the most common use of coal-water slurries is as a means for transporting coal, either within a plant or through pipelines. While coal-water slurry fuels have not yet become a widespread replacement for liquid fossil fuels, they are a viable alternative to heavy fuel oils provided that they are made from coal with sufficiently low ash content.

REFERENCES

1. Mishra, S.K.; Kanungo, S.B. Factors affecting the preparation of highly concentrated coal-water slurry (HCCWS). J. Sci. Ind. Res. **2000**, *59*, 765–790.
2. Pulido, J.E.; Rojas, C.P.; Acero, G.; Duran, M.; Orozco, M. Rheology of Columbian coal-water slurry fuels: effect of particle-size distribution. Coal Sci. Technol. **1995**, *24*, 1585–1588.
3. Usui, H.; Saeki, T.; Hayashi, K.; Tamura, T. Sedimentation stability and rheology of coal water slurries. Coal Prep. **1997**, *18*, 201–214.
4. Kawatra, S.K.; Bakshi, A.K. The on-line pressure vessel rheometer for concentrated coal slurries. Coal Prep. **2002**, *22* (1), 41–56.
5. Nguyen, Q.D.; Logos, C.; Semmler, T. Rheological properties of South Australian coal-water slurries. Coal Prep. **1997**, *18*, 185–199.
6. Mishra, S.K.; Senapati, P.K.; Panda, D. Rheological behavior of coal water slurry. Energy Sour. **2002**, *24*, 159–167.
7. Pawlik, M.; Laskowski, J.S.; Liu, H. Effect of humic acids and coal surface properties on rheology of coal-water slurries. Coal Prep. **1997**, *18*, 129–149.
8. Takao, S.; Ozaki, H. The effects of agitation on the rheology of a coal water mixture. Coal Prep. **1997**, *18*, 215–225.
9. EIA. U.S. Coal Reserves. 1997 Update. Energy Information Administration; http://www.eia.doe.gov/cneaf/coal/reserves/front-1.html.
10. Kawatra, S.K.; Eisele, T.C. *Coal Desulfurization—High Efficiency Preparation Methods*; Taylor & Francis: New York, 2001.
11. Usui, H.; Tasukawa, T.; Saeki, T.; Katagiri, K. Rheology of low rank coal slurries prepared by an upgrading process. Coal Prep. **1997**, *18*, 119–128.
12. Nodelman, I.G.; Pisupati, S.V.; Miller, S.F.; Scaroni, A.W. Partitioning behavior of trace elements during pilot-scale combustion of pulverized coal and coal-water slurry fuel. J. Hazard. Mater. **2000**, *74*, 47–59.
13. Okasha, F.; Miccio, M. Prediction of coal-water slurry dispersion in a fluidized bed combustor, Twenty-Sixth International Symposium on Combustion; The Combustion Institute, 1996; 3277–3285.
14. Hsu, B.D. Coal fueled diesel engine development update at GE transportation systems. In *Coal-Fueled Diesel Engines*; Caton J.A., Webb, H.A., Eds.; ASME, 16-ICE, 1992; 1–9.

15. Cox, C.B. *Comparing the Studies of a Coal Slurry Pipeline*; Virginia Water Resources Research Center, Special Report No. 17; Virginia Polytechnic Institute and State University: Blacksburg, VA, 1983.
16. Anonymous. Black Mesa Pipeline. Black Mesa Pipeline Co.; http://www.blackmesapipeline.com/index.htm, 2003.
17. Gandhi, R.L.; Snoek, P.E. Design aspects of slurry pipelines. In *Handbook of Fluids in Motion*; Cheremisinoff, N.P., Gupta, R., Eds.; Ann Arbor Science Publishers: Ann Arbor, MI, 1983; Chapter 36, 945–968.
18. Gilles, R.G.; Shook, C.A. A deposition velocity correlation for water slurries. Can. J. Chem. Eng. **1991**, *69*, 1225–1227.
19. Guo, J.; Hodges, S.; Uhlherr, P.H.T. Dewatering of HTD coal slurry by mechanical expression. Coal Prep. **1997**, *18*, 227–239.
20. Son, S.Y.; Kihm, K.D. Effect of coal particle size on coal-water slurry (CWS) atomization. Atomization Sprays **1998**, *8*, 503–519.
21. Office of Surface Mining. *Report on October 2000 Breakthrough at the Big Branch Slurry Impoundment*; U.S. Department of the Interior, 2002; http://www.osmre.gov/pdf/martincounty030402.pdf.
22. Metso Minerals. Coal water fuel (CWF) plants: raw coal transformed into liquid fuel. http://www.metsominerals.com/inetMinerals/mmcontent2.nsf/WebWID/WTB-030508-2256B-5BA7F/$File/FLUID_CARBON.pdf.

Computational Fluid Dynamics

André Bakker
Elizabeth Marden Marshall
Fluent Inc., Lebanon, New Hampshire, U.S.A.

INTRODUCTION

Computational fluid dynamics (CFD) is an engineering field in which flow fields and fields of related scalar variables, such as temperature and chemical species concentrations, are calculated in great detail for the domain of interest. The flow field calculations are performed by solving discretized forms of the equations for the conservation of mass, momentum, energy, and other relevant variables. The final results are detailed maps of the flow field showing fluid velocities, temperatures, and species concentrations throughout the system. These results can be used to gain a better understanding of the system of interest. Results of computational fluid dynamics can illustrate how a piece of equipment operates, how to troubleshoot problems, how to optimize performance, and how to design new equipment. During the 1950s and 1960s, CFD was used mainly in the aerospace and defense industries. This was followed by the power generation industry (especially the nuclear industry) and the automotive industry during the 1970s. The process industries did not start to use CFD seriously until the 1980s and 1990s. One of the important reasons for the relatively late adoption was that chemical reactors are characterized by so many complex phenomena. These include heat transfer, multiphase flow, and homogenous and heterogenous reactions. The development of methods to model these complex physical phenomena has occurred in recent years. Much progress has been made, and CFD is now used to model accurately many pieces of plant equipment, such as fluidized beds, packed-bed reactors, cyclone separators, spray dryers, bubble columns, stirred vessels, static mixers, and more.

One way the CFD results differ from traditional analysis methods is in the amount of detail they provide. Whereas traditional methods, such as design based on correlations of experimental data or overall thermodynamic balances, provide general design guidelines, they do not usually provide insight into the internal working of the equipment. For example, in a chemical reactor where several competing reactions take place, the final product composition may depend on nonuniformities in the concentration, flow, or temperature fields inside the reactor. Traditional methods that make use of average values inside the reactor are not able to take into account the effect of such nonuniformities. Computational fluid dynamics methods, which typically calculate flow field variables at hundreds of thousands of points inside the reactor to come up with overall reaction rates, are far better suited for the analysis of such systems. Another difference between CFD and traditional design methods is the minimal reliance of CFD on experimental data and extrapolation of that data to different scales, a process known as scale-up. Computational fluid dynamics relies on solving the fundamental equations of motion and conservation. These equations are scale independent and can be solved directly for the full-scale equipment.

While CFD provides a number of advantages over other analysis methods, high-fidelity CFD can come at a cost. It typically requires the use of specialized CFD software that offers extensive modeling capabilities, the availability of a skilled CFD analyst to set up the model and interpret the results, and high-end computing resources. In recent years, however, the cost of high-end computing has dropped substantially as the speed and power of processors have increased at a similar rate. Clusters of inexpensive computers can now do jobs that previously required supercomputers. In addition, new CFD software products have been introduced with limited capabilities and easy-to-use front ends. These new tools are designed to make CFD available to nonspecialists who want quick decision support for product design or troubleshooting. In this entry, the conservation equations that describe fluid motion are reviewed, and the general methodology behind CFD is summarized. The steps that are typically followed during an analysis are described and illustrated by an example of a reacting fluidized bed.

CONSERVATION EQUATIONS

Computational fluid dynamics is based on the principle of solving conservation equations for all relevant variables. The conservation equations include the transport of the variable throughout the domain, as well as either its creation or its destruction. Conserved variables include:

- Mass
- Momentum

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016187

- Enthalpy
- Turbulent kinetic energy
- Turbulent energy dissipation rate
- Chemical species concentrations
- Local reaction rates
- Local volume fractions for multiphase problems.

The equation for conservation of mass, also termed the continuity equation, has the form:

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_i}(\rho U_i) = 0 \tag{1}$$

Here, ρ is the fluid density and U_i is the fluid velocity in the x_i direction. When an index, such as i, is repeated in the same term, it means that the term is a summation over all possible values of the index. The first term on the left-hand side describes the change in fluid density over time, and the second term describes the transport of the fluid. For incompressible fluids, which have a constant density, the continuity equation reduces to the following, simpler form:

$$\frac{\partial U_i}{\partial x_i} = 0 \tag{2}$$

The statement for the conservation of momentum, also known as the Navier–Stokes equation, is as follows:

$$\begin{aligned}\frac{\partial(\rho U_i)}{\partial t} + \frac{\partial}{\partial x_j}(\rho U_i U_j) &= -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j}\left[\mu\left(\frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i} - \frac{2}{3}\frac{\partial U_k}{\partial x_k}\delta_{ij}\right)\right] \\ &\quad + \rho g_i + F_i\end{aligned} \tag{3}$$

The first term on the left-hand side describes the variation of the fluid momentum in time and the second term describes the transport of the momentum in the flow (convective transport). The first term on the right-hand side describes the effect of gradients in the pressure p; the second term, the transport of momentum due to the molecular viscosity μ (diffusive transport); the third term, the effect of gravity g; and in the last term, F_i lumps together all the other forces acting on the fluid. Techniques for solving the set of four equations (one continuity and three momentum equations) are discussed in a later section of this entry. When the flow is compressible, it is usually necessary to close the system of equations listed above using a thermodynamic equation of state (such as the ideal gas law) that calculates the density as a function of temperature and pressure.

For steady-state flows, the first term on the left-hand side is zero. For turbulent flows, which are unsteady, it saves time and computer resources to calculate a single, time-averaged flow field only, instead of solving for the full time-dependent flow field. To do this, the velocity is written as the sum of a constant and a fluctuating component $(U_i + u_i')$, and the momentum equation is averaged over time, resulting in the following modified conservation equation for momentum:

$$\begin{aligned}\frac{\partial}{\partial x_j}(\rho U_i U_j) &= -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j} \\ &\quad \times \left[\mu\left(\frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i} - \frac{2}{3}\frac{\partial U_k}{\partial x_k}\delta_{ij}\right)\right] \\ &\quad + \frac{\partial}{\partial x_j}(-\rho\overline{u_i' u_j'}) + \rho g_i + F_i\end{aligned} \tag{4}$$

A new term is introduced, the so-called Reynolds stresses $\overline{u_i' u_j'}$. The overbar denotes a time average. This term is the correlation between the turbulent velocity fluctuations u_i' and u_j', and it describes the transport of momentum in the mean flow due to turbulence. This term is difficult to model, and over the years a variety of turbulence models have been developed. Turbulence models are necessary for calculating time-averaged flow fields directly, without first having to calculate a fully time-dependent flow field and then doing time averaging. The use of turbulence models is therefore much more computationally efficient. A detailed discussion is beyond the scope of this entry, but it is important to note that not all turbulence models are equally suited for all types of flow. Table 1 summarizes the most common turbulence models and their properties.

For the other conserved quantities, conservation equations of a form similar to the momentum conservation equation are solved. These usually include a time-dependent term, a convective transport term (describing the transport of the variable due to the mean flow), a diffusive transport term, and a generalized source term describing all other physical effects.

COMPUTATIONAL FLUID DYNAMICS METHODOLOGY

Overview

The general methodology followed during a CFD analysis comprises the following steps:

1. Geometry creation
2. Grid generation
3. Defining the model by specifying physical models, boundary conditions, and initial conditions
4. Specification of the numerical methods to be used for solving the model equations
5. Performing the calculation
6. Analysis and interpretation of the results.

Table 1 Summary of turbulence models

Model	Description, advantages, and disadvantages
Standard k–ε	The most widely used model, it is robust, economical, and time tested. The Reynolds stresses are not calculated directly, but are modeled in a simplified way by adding a so-called turbulent viscosity to the molecular viscosity. Its main advantages are a rapid, stable calculation, and reasonable results for many flows, especially those with a high Reynolds number. It is not recommended for highly swirling flows, round jets, or flows with strong flow separation
RNG k–ε	A modified version of the k–ε model, this model yields improved results for swirling flows and flow separation. It is not well suited for round jets and is not as stable as the standard k–ε model
Realizable k–ε	Another modified version of the k–ε model, the realizable k–ε model correctly predicts the flow in round jets and is also well suited for swirling flows and flows involving separation
Spalart–Allmaras	The Spalart–Allmaras model was developed for external flows in aerospace applications. It provides good answers for attached flows and flows with mild flow separation. It is not commonly used for process industry applications
RSM	The full Reynolds stress model (RSM) provides good predictions for all types of flows, including swirl, separation, and round and planar jets. Because it solves transport equations for the Reynolds stresses directly, longer calculation times are required than for the k–ε models
LES	Large eddy simulation (LES) is a transient formulation that provides excellent results for all flow systems. It solves the Navier–Stokes equations for large-scale turbulent fluctuations and models only the small-scale fluctuations (smaller than a computational cell). Because it is a transient formulation, the required computational resources are considerably larger than those required for the RSM and k–ε family of models. In addition, a finer grid is needed to gain the maximum benefit from the model and to accurately capture the turbulence in the smallest, subgrid-scale eddies. Analysis of LES data usually requires some degree of planning in advance of building the model.

Preprocessing

The first two steps, geometry and grid creation, are also referred to as preprocessing. In the first step, geometry creation, a computer model is created of all flow passages and of those solid components that need to be included in the heat transfer analysis. The geometry is created in a way that is similar to three-dimensional solid model creation in computer aided design (CAD) software, and in software used for structural analysis. In fact, it is becoming more and more common to import CAD geometries into the CFD preprocessor instead of creating the whole geometry from scratch using the preprocessor. An important difference between the computer models constructed for structural analysis applications and those constructed for CFD is that the former requires details of the solid parts, whereas the latter requires details of the flow passages or of the space enclosed by the solid parts. Hence, when CAD models are imported into CFD preprocessing software, new volumetric entities for the flow passages usually have to be created from the imported solid surfaces. Furthermore, certain geometric details that can be present in the CAD model (e.g., boltheads, weldlines, and flange details) but not relevant to the CFD analysis can (and should) be removed.

Although it is common to use three-dimensional geometries, some problems can be solved suitably in two-dimensions. For example, the flow in a straight circular pipe in the absence of gravity or asymmetric heating is independent of the angular position. In such cases, it is sufficient to create a two-dimensional geometry, taking into account the axial and radial directions, but ignoring the circumferential (''tangential'') coordinate.

Once the geometry has been created, a computational grid, or mesh, has to be generated. During grid generation, the flow domain defined in the geometric model is subdivided into a large number of computational cells. The CFD solver will later solve conservation equations inside each computational cell, keeping track of what enters and leaves through each of the cell faces. Generating a suitable grid is therefore of utmost importance for the success of the model calculations. Different cell types with different shapes can be used to this effect. For two-dimensional boundary layer flow, for example, it is better to use up to 10 layers of quadrilateral cells than triangular cells in

the near-wall region. To accurately capture a two-dimensional jet entering a plenum, at least 10 cells should be used for the jet cross section to avoid excessive spreading. These guidelines can be extended to three-dimensional cases, and depend on the overall dimensions and requirements of the case at hand. In most software, the initial grid can be refined or coarsened during the course of the solution, to better resolve gradients or capture other effects, if needed. Fig. 1 shows some of the most commonly used types. The corner points of the cells are referred to as nodes. For two-dimensional models it is common to use triangular or quadrilateral cells. For three-dimensional models it is common to use hexahedral, tetrahedral, prismatic, or pyramid cells. Of course, these are not all of the possible cell types. For example, clipping corners of hexahedral cells can create other cell shapes. To accommodate such shapes, some CFD software allows the use of arbitrary polygonal cells, with any number of faces.

There are different ways in which the cells can be distributed throughout the domain. When the mesh consists solely of quadrilateral or hexahedral cells distributed throughout the domain in a regular, stacked pattern, such as that shown in Fig. 2, the mesh is referred to as a structured mesh. The use of such structured meshes, in which all cells are referenced by an i, j, k index in the computational domain, was common in the past because of the advantages it offered in the form of simplified programming and straightforward data interpolation routines. The main disadvantage of using structured meshes lies in the difficulties they pose in representing complex shapes. Now it is more common to use unstructured meshes, such as that shown in Fig. 3. In unstructured meshes, cells can be arranged in an arbitrary fashion. The user can choose the cell arrangement that is most suitable for reproducing the details of the geometry. Unstructured meshes can use only one cell type (e.g., all hexahedral or all tetrahedral), or mix different cell types as shown in the mesh in Fig. 3. Meshes that combine different cell types are usually referred to as hybrid meshes.

For very complex geometries, meshing can be simplified by subdividing the flow domain into different volumetric zones that can be meshed independently, choosing the best mesh type for that particular zone. The meshes for the different zones then can be combined into one mesh, usually called a multiblock mesh. When combining the meshes for the different zones, the cells and the nodes at the shared faces do not necessarily have to be lined up exactly. Multiblock meshes are called conformal meshes when the nodes at the shared faces do line up, and when they do not line up they are called nonconformal meshes.

Model Definition

Once the mesh has been created, the scope of the problem to be solved has to be specified. This step includes specifying the material properties, defining the physical models, setting the boundary conditions, and specifying the initial conditions.

A wide range of physical models is available in most commercial CFD software. At a minimum, the flow field will be calculated by solving the conservation equations for mass and momentum. In addition to flow, many of the problems encountered in the process industry involve heat transfer also. For such applications, the temperature field can also be calculated, which is commonly done by solving a conservation equation for enthalpy. For problems involving chemical reaction, the transport equations for the chemical species involved in the reaction(s) will be solved. The creation and destruction of the species due to the reaction are modeled by means of source terms in these equations. The reaction rates determining these source terms are calculated locally, based on the values of species concentrations and temperature at each

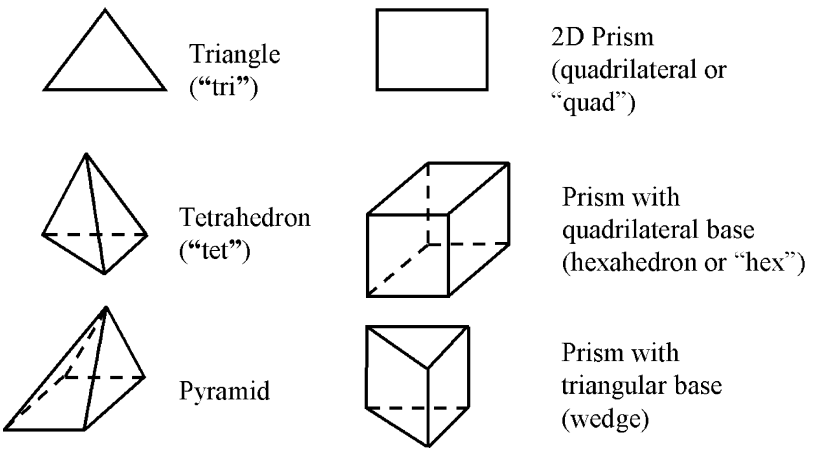


Fig. 1 Examples of three-dimensional element shapes that can be used in CFD simulations.

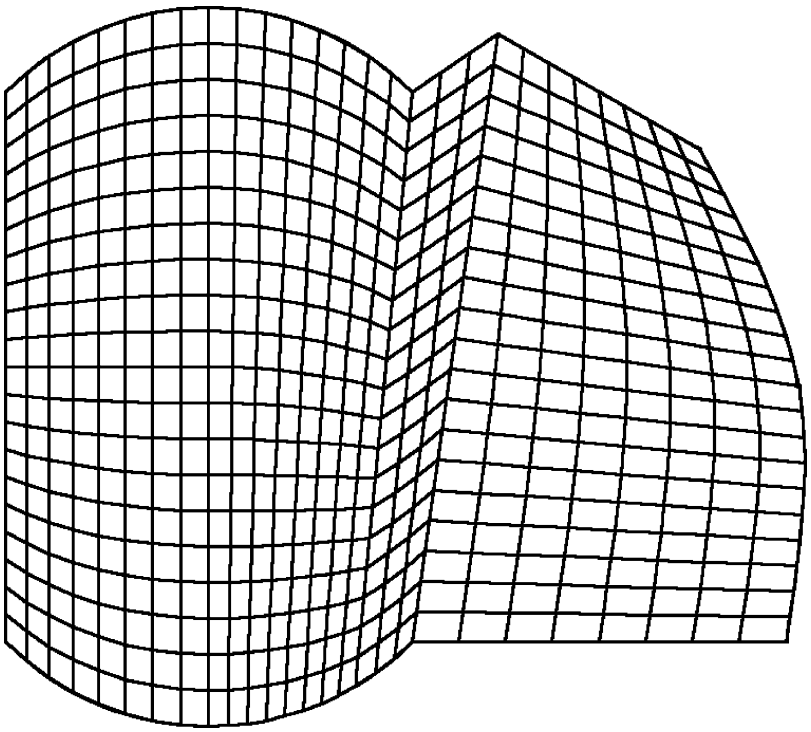


Fig. 2 An example of a two-dimensional structured grid.

variable storage site (a point at the center of each computational cell). The quality of the CFD predictions for reacting systems depends largely on the availability of accurate reaction rate constants. Many real-life systems involve a number of main reactions and also a number of side reactions that progress at much slower rates. When the total number of reactions is very large, or when the reaction constants for the less important side reactions are not known, the reaction model is often simplified to include only the most important reactions.

Multiphase flows are common in many types of process equipment. In such equipment there is usually one main continuous phase (a primary phase), and one or more dispersed (or secondary) phases. Multiphase flows can be modeled in a number of different ways. The most straightforward method is to calculate a single-phase flow field and then to calculate the trajectories of a finite number of particle streams for the dispersed phase that include the effects of drag, buoyancy, and turbulence. (Here, the term particle is used in the most general sense; it includes drops and bubbles also.) This so-called Lagrangian particle tracking calculation shows how the particles are dispersed throughout the domain. Physical phenomena such as particle evaporation, boiling, or surface reaction can be included in the tracking algorithm as well. The effect that the particles have on the flow field can be modeled by including particle trajectory-dependent source terms in the momentum (or other) equations for the fluid. Particle tracking models are best suited for systems with low volume fractions ($<10\%$) of the dispersed phase. Examples of systems that are commonly modeled using this method are spray dryers and liquid fuel or coal combustors.

Eulerian multiphase flow models work differently. Rather than compute individual particle trajectories, the primary and secondary phases are modeled as interpenetrating continua using separate sets of conservation equations. In addition to a set of momentum and continuity equations for each phase, a volume fraction equation for each secondary phase is also solved. Momentum exchange terms are included, which depend on the buoyancy, drag, and lift of the secondary phase particles. Computational fluid dynamics predictions are for the flow fields for both phases, which are usually different, depending on the system characteristics. Eulerian multiphase flow models can be formulated for any number of phases. Breakup and coalescence can be included for droplets or bubbles, along with mass transfer between the phases. When one of the secondary phases consists of a granular material, a maximum packing density can be prescribed for that phase, and a special treatment is used for the granular viscosity and granular pressure. Eulerian multiphase flow models are commonly used for fluidized beds, stirred vessels, and bubble columns.

Other types of multiphase flows are free surface flows, where there is a well-defined interface between two continuous phases. Examples of such flows can be found in liquid separators, unbaffled mixing vessels where surface deformation occurs when a central vortex forms, mold filling applications, and blow

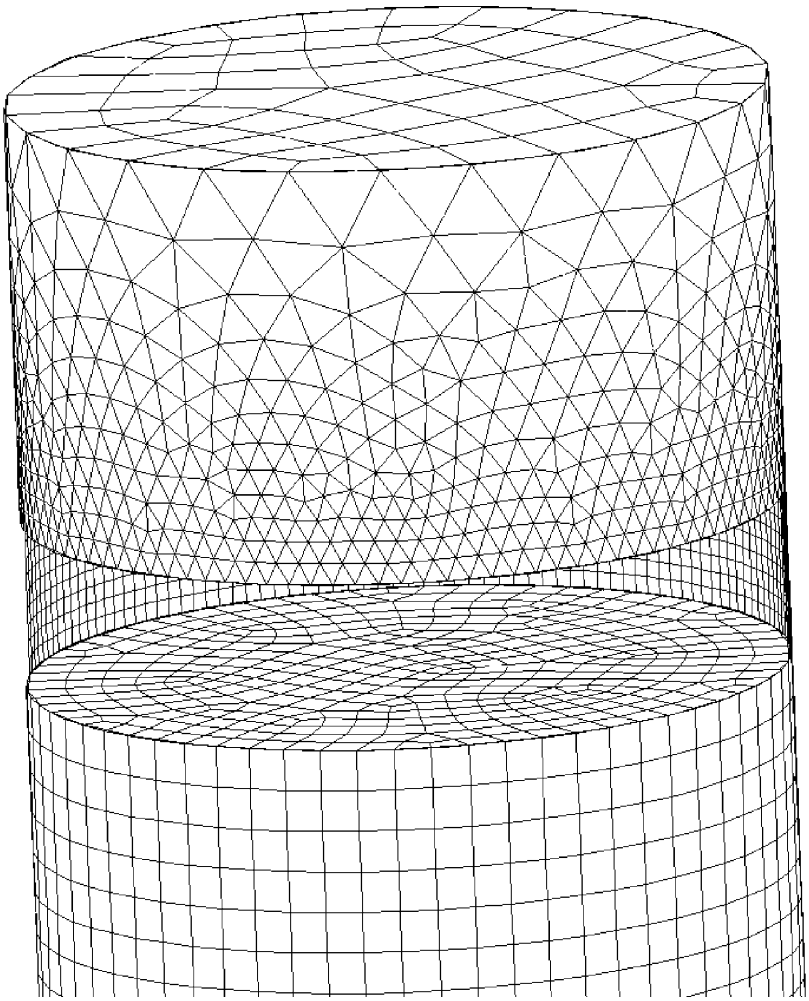


Fig. 3 An example of a three-dimensional unstructured grid.

molding applications. With free surface flows, mass and momentum conservation equations are solved, along with an equation that describes the movement of the interface.

The material properties that have to be specified depend on the physical models chosen. For flow field calculations only, it is usually sufficient to specify the density and the viscosity. The fluid can be incompressible (constant density) or compressible (a compressible liquid or an ideal or nonideal gas). The viscosity can be either Newtonian or non-Newtonian (independent or dependent on shear rate, respectively). Any type of non-Newtonian fluid can be modeled, including shear-thinning fluids, yield-stress fluids, and viscoelastic fluids, assuming that the viscosity behavior is known. Unfortunately, this is not always the case. For fluids with complex rheological behavior such as paper pulp, dough, and other viscoelastic fluids, it is often difficult to determine the rheology. In such cases, simplified rheologies often have to be used, and it is not uncommon to model the fluid as Newtonian using an effective average viscosity that is known to give good results for such quantities as pressure drop or torque predictions in rotating equipment.

When heat transfer is involved, the density and viscosity can be functions of temperature, and the thermal properties of the fluid have to be prescribed as well. These properties include heat capacity, thermal conductivity, molecular weight (for gases), and thermal expansion coefficient. For problems involving chemical species, all physical properties have to be specified for all of the species along with a method to calculate the average property for mixtures in each cell using the properties of the component species.

To complete the model specification, boundary conditions have to be specified. These describe the flow conditions at the domain boundaries. At flow inlets one can usually specify the fluid velocity, a mass flow rate, or an inlet pressure. Depending on the problem definition, the inlet temperature, species concentrations, turbulence properties, and volume fractions of any secondary phase must also be supplied. At flow exits, one usually specifies an outlet pressure, and if entrainment through a flow exit is anticipated, the exit

should be moved downstream or have boundary conditions carefully specified for the entrained flow. For problems with only a single flow exit—far from recirculation zones—no such boundary conditions are required. One can also set guidelines regarding how much of the mass entering the domain from the inlet(s) exits through each of the outlets. At the domain walls, the wall can be prescribed to move at exactly the same velocity as the adjacent fluid with no imposed frictional drag (a so-called slip condition), or the wall can be assigned the more common no-slip condition, where a constant wall velocity (tangent to the fluid flow), usually zero, is set. The shear stress at the wall can also be prescribed. When the problem involves heat transfer, boundary condition options include a specified heat flux (which can be zero to model well-insulated, or adiabatic, walls), a specified temperature, or an external heat transfer boundary condition that may include heat losses due to radiation.

For problems involving repeating geometry and flow conditions, it is often sufficient to model the flow in only one section of the domain, which is identical to that in other sections of the same size. At the appropriate boundaries of that section, either periodic or symmetric boundary conditions can be prescribed.

For problems where the fluid behavior is steady and the final solution is independent of the initial conditions, calculations can usually be carried out in a steady-state fashion. For problems that describe transient behavior or that require a time-marching solution technique, initial conditions must be supplied in addition to boundary conditions. These include the state of the flow field, concentration fields, and temperature fields at time zero at all locations in the domain.

As part of the specification of the physical models and boundary conditions, the user has to specify the body forces on the fluid. Body forces are those forces that act on the entire volume of fluid throughout the domain. These forces include gravity, electromagnetic forces (if relevant), and the Coriolis force for rotating domains.

Numerical Methods

There are a variety of numerical methods available to solve the conservation equations. The most commonly used method in commercially available CFD software today is the finite volume method. Excellent descriptions of this method can be found in Refs.[1,2]. With the finite volume method, an integral form of the conservation equations is solved by performing a mass and momentum balance over all faces of each computational cell. There are, however, many other methods available, such as the finite element method (where the equations are solved in differential form instead of in integral form), vorticity-based methods (where conservation equations for vorticity are solved instead of those for momentum), and spectral element methods (where the equations are transformed into Fourier space before being solved). The main advantage of the finite volume method over other methods is that it conserves mass on coarse as well as fine meshes, including those with large numbers (tens of millions) of computational cells, and performs well for turbulent flows.

One disadvantage of the finite volume method, however, is that because velocities and other variables have to be interpolated from the cell centers (the storage sites) to the cell faces so that face fluxes can be computed, a certain amount of numerical error is introduced. These errors tend to have the effect of decreasing the sharpness of gradients in the flow when compared to real life, similar to how diffusion smoothens out gradients in all scalar fields. This error is therefore termed numerical diffusion, and it is present in all CFD simulations performed using the finite volume method. It can be minimized by using meshes that are finer in regions where steep gradients in the flow occur, and by choosing accurate interpolation methods (also called numerical schemes) that result in a small error (usually at the expense of increased computation time). However, for flows where gradients tend to be relatively small, such as in time-averaged turbulent flows, the numerical diffusion may be insignificant. The finite element method is not as prone to numerical diffusion, and therefore tends to give solutions that are more accurate for a given number of cells for flows that have sharp gradients. Examples of such flows are laminar flows at very low Reynolds numbers. The finite element method is not as computationally efficient as the finite volume method for turbulent flows or for meshes with large numbers of cells. Most applications in the process industries involve turbulent flows in complex geometries that require fine meshes. The finite volume method is used, therefore, for most applications. The finite element method tends to be the better choice for applications involving very viscous fluids, such as those encountered in blow molding and extrusion processes.

Solution

Once the solution method has been chosen, the actual calculations are performed using an iterative procedure. The solution procedure is as follows. Starting with an initial solution provided by the user (which may be a value of zero for all variables throughout the domain), the solver determines the error in the mass and momentum balances for each computational cell. This error is also known as the residual. Based on

that error, the solver determines what adjustments need to be made to various variables to improve the balances for each cell. These adjustments are applied to all cells in the domain, and the procedure is repeated until the error in the solution is smaller than a preset tolerance specified by the user. At that time, the solution is considered converged and the calculation is halted. There are many different ways in which this concept can be implemented, and many mathematical techniques are available to enhance the rate of convergence and to improve the solver's ability to converge to a stable and accurate solution. A discussion of these techniques is beyond the scope of this entry.

Postprocessing

The final step in a CFD simulation is to analyze and interpret the results, a step also referred to as postprocessing. Typical CFD results include values for the pressure, the three velocity components, turbulence properties, temperature, and species concentrations at all grid cells in the domain (which is typically on the order of several hundred thousand in today's applications). The total amount of available data is thus extremely large and detailed. To be useful to an engineer the data have to be presented in an understandable way.

The objective of the simulation might be to obtain quantitative estimates for certain global variables, such as an overall reaction rate, the temperature at a certain location, or the pressure drop across the domain. These can be calculated from the CFD solution in a relatively straightforward manner. The results also can be presented graphically, and there are many options available for doing so.

Iso-surfaces can be constructed by connecting all points in the flow domain that have exactly the same value of a given variable. These surfaces can be colored by the contours of any other variable. Iso-surfaces are useful to visualize certain flow-field features, such as the central vortex in the cyclone separator shown in Fig. 4. Contour plots can be used to show the local values of scalar variables. A cross section of the flow domain is created and colored by the local value of the variable of interest, e.g., temperature. A color scale is used indicating how each color corresponds to a certain value of the variable. Such plots can be used to quickly see the variation of important variables throughout the domain. An example of a contour plot is found in Fig. 5, which is described in the example in the following section.

A commonly used graphical visualization of the flow field is through plots of velocity vectors. An example, also taken from the simulation described in the following section, is shown in Fig. 6. The velocity vectors point in the direction of the fluid flow where the vector originates. The length of the vector can be made proportional to the local velocity magnitude, such that longer vectors indicate higher velocities. The vectors can also be colored by the velocity magnitude, or by any other variable, as desired.

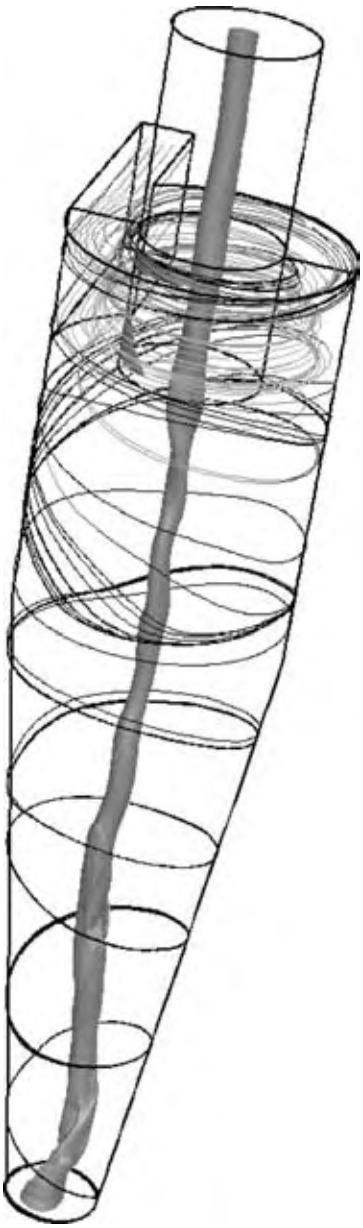


Fig. 4 An iso-surface of axial vorticity in a cyclone, colored by velocity magnitude, is used to show the central vortex; pathlines are used to show the swirling flow. *(View this art in color at www.dekker.com.)*

Other methods to visualize the flow field make use of flow lines, of which different types are available. Streamlines are useful in two-dimensional simulations. They are curves that are tangent to the flow field and whose spacing is such that a fixed mass of fluid passes between any pair of lines. For three-dimensional

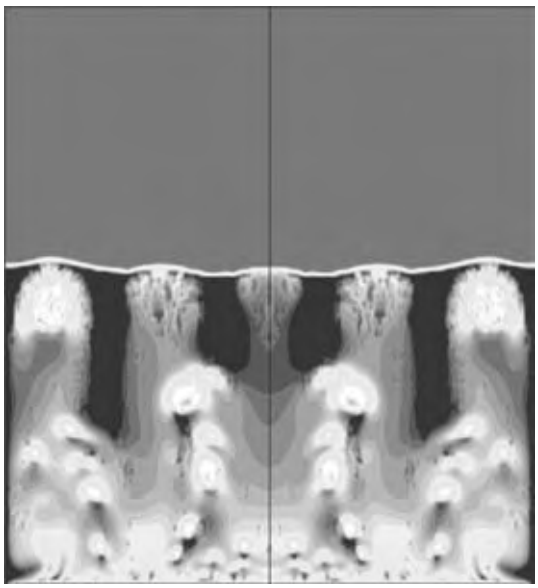


Fig. 5 Gas volume fraction in a reacting fluidized bed 0.6 sec after the operation begins. *(View this art in color at www.dekker.com.)*

geometries, pathlines can be constructed by calculating the trajectories of massless particles that follow the mean flow. Streaklines are formed by a series of particles that are continuously injected into the flow from a number of fixed locations. For steady two-dimensional flows, streamlines, pathlines, and streaklines are the same. However, for unsteady (time varying) flows, they can be very different.

Other types of flow lines also can be used. Oil-flow lines are pathlines that are constrained to a given boundary surface. When calculating the pathline, velocity components that are tangent to the given boundary surface are included and normal velocity components are ignored. This is useful for visualization of the flow field over an object. Timelines are formed by placing a series of particles along a line in the flow field at time zero. The particles are tracked for a given amount of time, and the curve connecting the particles is redrawn. Timelines also are called material lines, and they are commonly used to study the amount of stretching and deformation that occurs in mixing applications.

REACTING FLUIDIZED BED

One example that illustrates many of the concepts described in this entry is the conversion of ozone to oxygen in a fluidized bed. The particles in the bed serve as a catalyst where surface reactions take place. To model this complex process, the commercial software FLUENT is used. The fluidized bed is simulated using the Eulerian granular multiphase model, and a reaction is specified in the bed to account for the decomposition of ozone and subsequent production of oxygen.

A two-dimensional axisymmetric model of a column 0.23 m in diameter and 0.25 m high is used.[3] At rest, the bed is 0.12 m high, and contains sand particles impregnated with iron oxide, 117 μm in diameter with a density of 2.65 g/cm^3. Two gas phase species (ozone and oxygen) are specified. Ozone enters from inlets at the base of the bed at speeds that range from 4 to 14 cm/sec. The gas lifts and separates the particles in the bed, allowing for better interaction between the ozone and iron oxide at the particle surfaces. The decomposition rate is expressed as

$$K = 1.57\alpha[O_3] \tag{5}$$

where α is the volume fraction of the catalyst and $[O_3]$ is the concentration of ozone. After the reaction has

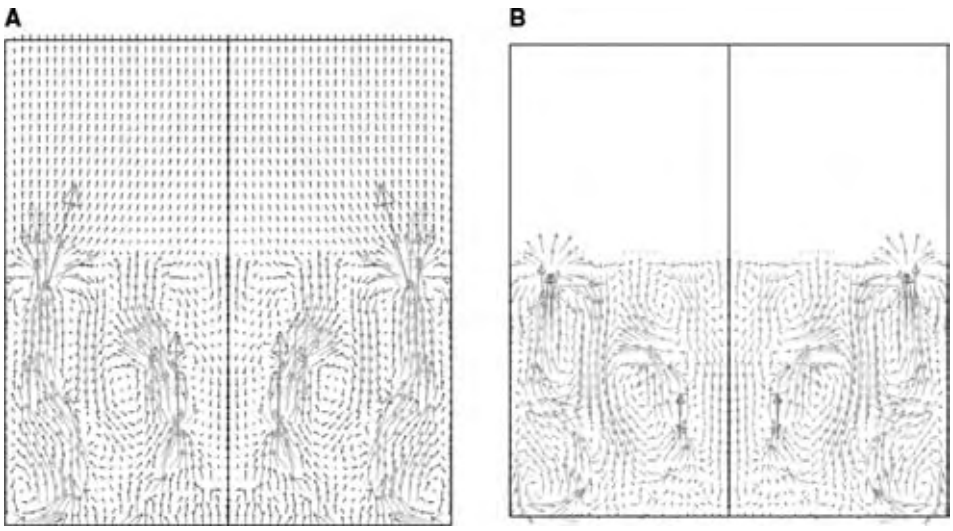


Fig. 6 Velocity vectors of (A) the gas phase and (B) the granular phase (sand) in a fluidized bed after 0.6 sec of operation; the sand motion is different from that of the gas, and is restricted to the bed region. *(View this art in color at www.dekker.com.)*

taken place, a mixture of ozone and oxygen exits through the top of the column. Of particular interest are the conversion characteristics of the bed, where conversion is defined as

$$1 - \frac{C_{\text{out}}}{C_{\text{in}}} \tag{6}$$

where C_{out} and C_{in} are the outgoing and incoming concentrations of ozone, respectively. If $C_{\text{out}}/C_{\text{in}}$ is small, conversion is high and the bed is operating efficiently.

Fig. 5, discussed earlier, shows contours of the gas volume fraction 0.6 sec after the flow is initiated. Light gray (or red) indicates regions of pure gas, and dark gray (or blue) indicates regions where the sand volume fraction is maximum. The flow field is the same whether the reaction in the gas phase takes place or not. Bubbles are formed near the bottom of the bed and migrate upward. The simulated bubble shape and size are grid dependent. If a coarse mesh is used, the bubbles are fewer in number and rounder. On finer meshes, the bubbles are denser and more irregular. The number and size of the bubbles have a significant impact on the conversion. The more bubbles in the domain, the greater the interaction between the ozone and sand particles, so the higher the conversion.

Fig. 6 illustrates the velocity fields for the gas (left) and granular (right) phases after 0.6 sec of operation. The vectors are plotted on every fifth grid cell to make the display easier to read. The velocity fields are different in several ways. First, there is no granular motion on the top of the bed, where only the oxygen and ozone gases are escaping. Second, while some of the larger flow features are the same for both phases, some of the smaller ones are not. For example, a rather large recirculation pattern dominates the gas flow on either side of the bed. In the sand phase, perhaps owing to the increased inertia of the material, the recirculation patterns are not as strong, and have slightly different shapes.

In Fig. 7, curves for $C_{\text{out}}/C_{\text{in}}$ (pink) and gas holdup (red) are plotted as functions of the gas velocity. The gas holdup is defined as the ratio of the gas volume in the bed to the total volume of the bed. The figure shows that the gas holdup increases with gas velocity up to a point, after which saturation occurs. This is because at the low end, the increasing gas velocity forces the bed to lift more. When saturation occurs, the bed can no longer rise and hold additional gas.

The curve of $C_{\text{out}}/C_{\text{in}}$ is in good agreement with data from Ref.[3]. The results follow the trends in the gas holdup. At low gas speeds, the residence time is long, so there is more time for the ozone to come into contact with the particles. $C_{\text{out}}/C_{\text{in}}$ is small, so conversion is high. At high speeds, the residence time reduces, and the rate of conversion tends toward a constant value.

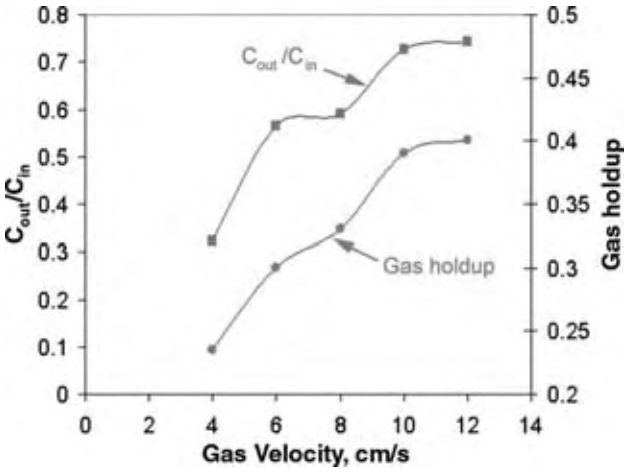


Fig. 7 Ozone conversion (pink) and gas holdup (red) as functions of gas velocity for the fluidized bed. *(View this art in color at www.dekker.com.)*

CONCLUSIONS

Before embarking on any CFD simulation, it is important to realize that the care that goes into the setup of the model is directly related to the quality of results that are obtained. The grid should be fine enough and of optimal quality to capture the flow details of interest. The choice of physical models, for turbulence or multiphase, for example, should be based on the flow regime and expected flow features. Boundary conditions and physical properties should be chosen carefully so that they represent the full range of anticipated behavior. Approximations—what to keep in the model and what to neglect—can also play a significant role in the type of effects that the simulation can capture. Finally, once the setup is complete and the simulation is running, it is important to make sure that proper convergence has been obtained before using the results for any type of careful analysis.

If these issues are heeded, the potential benefits that can be extracted from a simulation are numerous. With the computational resources available today, both in terms of speed and power, engineers are more able to use high-density grids and multiple, detailed physical models than at any time in the past. In the years to come, computing power will only continue to increase, so CFD results will become more and more accurate as turnaround times shorten. Interoperability between CFD and other computer-aided engineering packages is on the rise. In many industries, this has led to a recent expansion of CFD from its use in the analysis and design departments to being a major contributor to automated plant operation and process optimization.

ACKNOWLEDGMENTS

The authors wish to thank their colleagues Dr. Bin Yang, Dr. Wei Zhou, Dr. Ahmad Haidari, Dr. Ravi Prasad, Dr. Mike Slack, and Mr. John Murray who have helped to make this work possible.

NOTATION

α	Volume fraction of gas in fluidized bed
C_{in}	Inlet concentration (kg/m^3)
C_{out}	Outlet concentration (kg/m^3)
δ_{ij}	Kronecker delta function
ε	Turbulent kinetic energy dissipation rate (m^2/sec^3)
F_i	Net force in the i direction (N)
g	Gravitational acceleration (m/sec^2)
k	Turbulent kinetic energy (m^2/sec^2)
K	Rate of ozone decomposition ($kg\,mol/m^3\,sec$)
μ	Molecular viscosity ($kg/m\,sec$)
p	Pressure (Pa)
ρ	Density (kg/m^3)
t	Time (sec)
u'_i	Fluctuating component of the velocity in the direction i (m/sec)
U_i	Velocity in the direction i (m/sec)
x_i	Spatial coordinate in the direction i (m)

REFERENCES

1. Patankar, S.V. *Numerical Heat Transfer and Fluid Flow*; Hemisphere: Washington, DC, 1980.
2. Versteeg, H.K.; Malalasekera, W. *An Introduction to Computational Fluid Dynamics: The Finite Volume Method*; Longman Scientific & Technical: Essex, 1995.
3. Fryer, C.; Potter, O.E. Experimental investigation of models for fluidized bed catalytic reactors. AIChE J. **1976**, *22*, 38–47.

Computer-Aided Process Engineering

Andrzej Kraslawski
Lappeenranta University of Technology, Lappeenranta, Finland

INTRODUCTION

This entry presents the structure of computer-aided process engineering (CAPE) from the perspective of activities, problems, and methods. More detailed information is presented on process synthesis and chemical product development. Future prospects of process and product design are mentioned briefly.

BACKGROUND

CAPE develops methodologies and their implementations for designing activities and processes that enable the manufacturing of products from raw materials. The manufacturing process is realized by a chain of physical, biological, and chemical processes. Other functions of CAPE stress on process modeling or decision making in chemical supply chain. The term ''process systems engineering'' (PSE) is sometimes used instead of CAPE.

CAPE concentrates on finding a representation of problems, generation of several alternative solutions, and selection of an appropriate solution. Majority of the problems handled by CAPE is NP-hard. This means that the computation time increases exponentially with the problem size. It is an illustration of the fact that the design is usually ill defined (no clear information about its goals) and open ended (usually a few equally good designs exist). NP-hard problems are the main reason for the generation of several alternatives instead of providing straightaway optimal solution.

CAPE uses methods and tools developed in chemistry, physics, biology, mathematics, economics, and systems engineering, and applies them to products and production methods in process industries. Its original contribution is in the development of new representations for the description of the problems in process industries as well as that in computationally efficient solution methods and tools.

This entry gives an introduction to the history of CAPE, general characteristics, and, structure. It describes in more detail the most significant and characteristic branch of CAPE—process synthesis, and introduces the most important new area of CAPE applications—product design. A few conclusions and ideas about the future development of CAPE are also provided.

HISTORY

The development of chemical industries is characterized by the following main phases: first, extraction and purification of useful materials (metals from ores, dyes from plants and insects, etc.); next, production of naturally existing chemicals from various raw materials (acetic acid, furfural, and vanillin, to name a few); and finally, synthesis of chemicals that are not available naturally (nylon and other polymers). The degree of complication of the production processes increased in every phase. It was manifested by the fact that more and more equipment had to be used in the production to perform the so-called unit operations. The first design issue related to process synthesis was how to combine those unit operations to get products of good quality, in a short time, and at a reasonable cost. The advent of computers and scientific insights have resulted in treating the problem of connecting the equipment for realization of various unit operations as a mathematical programming problem,[1] and as such to be tractable by computers. Historically, the first applications of CAPE were in process design for petrochemical industry in the late 1950s.

The first conference on computer-aided design was organized in 1972.[2] Now there are cyclic conferences for the presentation of new results, namely, ESCAPE, PSE, FOCAPD, and FOCAPO. CAPE is also very educative for chemical engineers. Important journals on CAPE are *Computers and Chemical Engineering*, *AICHE Journal*, *Industrial & Engineering Chemistry Research*, *Chemical Engineering Science*, *Chemical Engineering*, and *Processing and Chemical Engineering Research and Design*, while the most popular textbooks are Refs.[3–6].

GENERAL CHARACTERISTICS

The activities and processes of CAPE are interdisciplinary, multiobjective, complex, uncertain, as well as dynamic.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030505

Typical activities are analysis of the consumer needs; its translation into technical specifications of the product; generation of ideas on how to produce such a material; synthesis of the production scheme, e.g., flowsheet; design of the process, its operation, and control; waste treatment; planning and controlling of the flow of raw materials, utilities, and final products; and analyses of the market response leading to the changes of the actual products. The interdisciplinary nature of CAPE results from the need of applying technical, scientific, economic, and management data, information, and knowledge to deal with the above-mentioned activities.

The multiobjective nature of the activities covered by CAPE results from the fact that production processes are dependent on technical, economic, environmental, and social systems. Each has its own priorities and objectives (e.g., technical—maximizes yield and purity; economic—maximizes profit; environmental—minimizes the amount of wastes and use of natural resources; and social—full employment for the workforce). Moreover, there are often several objectives inside the systems themselves. Usually there are conflicts between these objectives with their role being twofold. On one hand, they introduce serious constraints and limitations leading to the search of trade-offs between the objectives and finally Pareto optimal solutions. On the other, attempts to remove these conflicts could lead to the generation of highly innovative technical, economic, and organizational solutions.

The complexity of the activities and processes where CAPE is involved is a consequence of new economic, technical, scientific, and social opportunities. The first of these is manifested by such phenomena as mergers and acquisitions of companies as well as emergence of new markets. This has led to new problems related to the complicated supply chain and knowledge management in heterogeneous organizations. It resulted in the need for analysis and synthesis of multiscale systems. The maturity of the chemical industry has resulted in its shift from being process based toward product based. New challenges emerged in the development and operation of processes as well as in the design of products. On one hand, a lot of efforts have to be taken to improve the existing processes (usually running in nearly optimal state), and on the other to develop new products characterized by complex structural properties and unconventional formulations.

The complexity in CAPE applications also results from new opportunities offered by proteomics, materials science (self-assembled materials, nanomaterials, biomimetic materials), and pharmaceutical science (combinatorial design of molecules, drug delivery systems).

Social changes like new beauty standards and food habits resulted in the development of new industries that produce highly complex products (e.g., cosmetics, food products with long shelf life, etc.).

The dynamic aspects related to CAPE applications result from the earlier needs of process control to run the processes in unsteady state and control them, as well as from new requirements like the shortening of the life cycle of products or management of abnormal situations.

STRUCTURE

The structure of CAPE is complex because of the broad spectrum of the problems that are solved and the variety of methods applied. Structuring can be done from the various classifications presented in Table 1.

From the viewpoint of scientific methodology there are three main tasks in CAPE: representation of the problem, generation of several alternative solutions, and selection of the best one. These tasks correspond to the activities realized in four phases of any scientific method: analysis (description of the problem and identification of the objectives), hypothesis (generation of solutions), synthesis (comparing the solutions), and validation (formulation of conclusions). The activities realized in the last two phases correspond to the selection task in CAPE.

The CAPE activities specified from the point of view of realization of any industrial project are product design, process synthesis, process design, process operation, and process decommissioning. Looking from the historical perspective, the economic interest and R&D activities in particular phases of an industrial project development and realization have been changing depending on the economic factors in the world. The social, political, and technological changes in the last 20 years have been manifested by the growing demand for new products. This trend has resulted in the mergers and acquisitions of many businesses at a scale not observed earlier. Changes in the way businesses are organized have very deeply influenced the manufacturing and processing industries. The main consequences have been low profit margins of commodities, growing pressure on environmental aspects of production and conservation of raw materials, very high costs of R&D, significance of the low-tonnage and high value-added products, growing importance of the customer-oriented products, reduction of time from the development stage to the market, shortening of the life cycle of the products, and issues related to the intellectual property rights.

All these factors have caused the shift from process-oriented to product-centered businesses in processing industries. Consequently, the most interesting economic and scientific activities are now realized in

Table 1 The structure of CAPE

Classification					
Task-based		**Representation**	**Generation of Alternatives**		**Selection**
Activity-based	**Product design**	**Process synthesis**	**Process design**	**Process operation**	**Process decommissioning**
Problem-based	**Conceptualization** Identification of chemical path Flowsheet synthesis Solvent and catalyst selection Recycling decisions Location of sensors	**Change of Scale** Property-structure design Life-cycle analysis Properties prediction	**Intensification** Retrofit analysis Debottlenecking Batch processes planning and scheduling Data reconciliation Fault diagnosis Real time optimisation Heat and mass and momentum Modeling Supply chain management High throughput product development Multifunctional units modeling		**Integration** Heat and mass integration Sustainability analysis Safety analysis Linking of business decisions and process design

(Continued)

Table 1 The structure of CAPE *(Continued)*

Classification					
Method - based	**Optimization**		**Modelling and Simulation**	**Search**	**Selection**
	Continuous variables		Neural networks	Branch and bound algorithms	Rule-based Reasoning
	Optimization		CFD	Best-first search algorithms	Case-based reasoning
	MINLP & MINLP		Steady and dynamic state simulation	Branch-and-cut algorithms Means-ends analysis	Analytical hierarchy process
	Genetic algorithms		Parametric programming		
	Annealing techniques		Constraint programming		
	Dynamic programming		Multi-scale computations		
	Collocation methods				
	Stochastic optimisation				
	Agent-based computations				
	Disjunctive programming				
	Tabu search				
Tool - based	**Solvers**	**Simulators**	**Databases**	**Equipment designers**	**Decision support systems**
	GAMS	ASPENS	DETHERM	FLUENT	PROSYN
	ASCEND	gPROMS	CHEMSAFE	Engineers Aide Toolbox (Chempute Software)	KRUST
		ROMEO	Physprops		
			NISTData		

product- and process design, with particular stress on the development of new products, intensification of the existing processes and their integration into the operation of bigger systems, as in a whole division of industry, or the natural and social environment.

The classes of problems to be identified in the contemporary CAPE concern the problem of conceptualizing change of scale, intensification, and integration. There are two main subjects of conceptualization. The first one is generation of alternatives and selection of appropriate one for the identification of the chemical path that transforms the given raw materials into the required product. The second is identification of the structure of the flowsheet, i.e., what units are to be used to produce the required material and how they should be connected. More detailed information about the solvents and the control structure is also decided in the conceptualization phase.

The problems of change of scale result from extending the applications and research in CAPE to micro- and macroscales. As shown in Fig. 1, the activities of CAPE are no longer limited to mesoscale (unit operations and processing units). The research in microscale, in terms of dimensions of the object as well as duration of the phenomena, is characteristic of the activities in the field of product design. The CAPE activities in macroscale correspond to applications at the company, industry, or even global scale.

The intensification in CAPE concerns the changes and improvements of equipment and flowsheets, modifications of control procedures, as well as new approaches to organization and management of large-scale activities in process industries.

The integration activities have been mainly related to heat and mass exchangers. However, there are many more popular applications related to integration of flowsheet creation with simultaneous consideration of safety aspects. The design of control system is also beginning to be integrated with process synthesis and process design. The large-scale integration is aimed at combining technical and business decisions to optimize the value chain in the company.

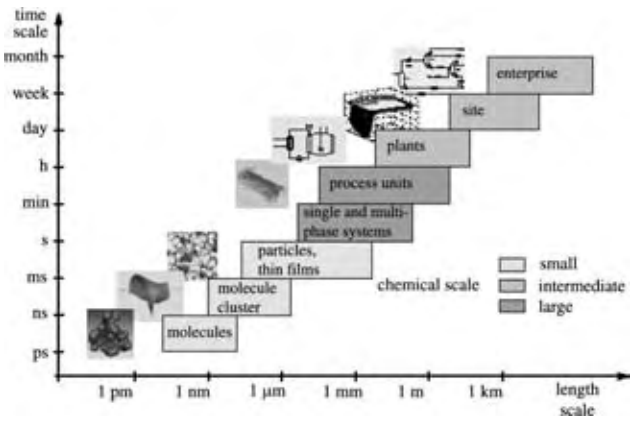


Fig. 1 Scales in process engineering. (From Ref.[43].)

The method-based classification of CAPE identifies the following main groups: optimization, modeling and simulation, search, and selection. Most R&D activities involve the first two. The methods of search and decision support relating to artificial intelligence seem to get less attention than they had in the 1980s and 1990s.

The tools used in CAPE could be specified as solvers, simulators, databases, equipment designers, and decision support systems. Several tools, such as simulators, databases, and decision supporting systems, are often embedded as one computing environment, e.g., ASPEN or ICAS. A few examples of the existing programs and environments are given in Table 1.

Process Synthesis

The most characteristic task of CAPE and historically the most popular is process synthesis.[7,8] The development of process synthesis is confirmed by the use of a very broad spectrum of methods for building a process flowsheet.

A quite common approach to solving the synthesis problem is application of optimization methods.[9] Optimization-based methods use not only traditional algorithmic approaches, such as mixed-integer nonlinear programming (MINLP), but also stochastic ones, e.g., simulated annealing and evolutionary algorithms.[10] Two common features of these methods are the formal, mathematical representation of the problem, and the subsequent use of optimization. The advantage of this approach is the provision of a systematic framework for handling a variety of process synthesis problems and the more rigorous analysis of features such as structure interactions and capital costs. An important drawback of optimization-based methods is the lack of ability to automatically generate a flowsheet superstructure. Another disadvantage is the huge computational effort required, and the fact that the optimality of the solution can be guaranteed only with respect to the alternatives that have been considered a priori.

Heuristic methods were developed by well-experienced engineers and researchers. The first attempt to develop a systematic heuristic approach for the synthesis of multicomponent separation sequences was made by Siirola and Rudd.[11] Common example is hierarchical heuristic approach.[12] Heuristic rules are applied at five design levels to generate and evaluate the alternatives using economic criteria. The hierarchical heuristic method emphasizes the strategy of decomposition and screening. It allows for a quick location of flowsheet structures that are often ''near''

optimal solutions. The major limitation of this method, because of its sequential nature, is the impossibility to manage interactions between different design levels. The same reason causes problems in the systematic handling of multiobjective issues within the hierarchical design. The hierarchical heuristic method offers no guarantee of finding the best possible design. The heuristic approach has been used in many applications, e.g., synthesis of separation systems,[13] waste minimization,[14] and metallurgical process design.[15]

A different approach to process synthesis is offered by means–ends method.[16] It is based on the observation that the purpose of material processing is to apply various operations in such a sequence that the differences in properties between the raw materials and the products are systematically eliminated. As a result, the raw materials are transformed into the desired products. The means–ends method starts with an initial state and successively applies transformation operators to produce intermediate states with fewer differences until the goal state is reached. The hierarchy for the reduction of property differences is as follows: identity, amount, concentration, temperature, pressure, and finally form. This property changing method has its limitations, as it ignores the influences and the impacts on other properties. Moreover, the search method takes an opportunistic approach, which cannot guarantee the generation of a feasible flowsheet. The means–ends analysis approach has been used as a systematic process synthesis method for overall process flowsheet synthesis, as well as for the more detailed case of a separation system to resolve the concentration differences in nonideal systems that include azeotropes.

The phenomena-driven method for the process synthesis analyzes not the processing units, the so-called building blocks, but the phenomena that occur in those blocks. This method is based on opportunistic task identification and integration. It was applied to separation process synthesis, based on thermodynamic phenomena.[17] It explored the relationship between the physicochemical properties, separation techniques, and conditions of operation. The number of alternatives for each separation task is reduced by systematically analyzing these relationships. Then, possible flowsheets are produced with a list of alternatives for the separation tasks.

The conflict-based method for process synthesis applies TRIZ approach.[18] TRIZ is a method for the identification of the system's conflicts and contradictions for the generation of innovative designs. The conflict-based approach decomposes a design problem into subproblems instead of applying hierarchical design. It is an efficient method for modifying the solution space and screening alternatives at an early stage. The design problem is represented by the conflicts between the inter-related design objectives or the characteristics of the process flowsheet. It overcomes the difficulties of considering the interconnection of the hierarchical design levels as well as the limitations of insufficient problem representation. CBA has been developed and used for the preliminary design of distillation and reactor/separator systems, as well as for waste management.[18,19]

The next method for process synthesis is case-based reasoning (CBR). CBR reuses solutions that were applied in the past to similar problems. It is composed of several steps combined into a cyclic procedure. In the first step, retrieval, a new problem is matched against problems of previous cases by calculating the value of the similarity functions to find the most similar problem and its solution. If the proposed solution does not meet the necessary requirements of the new problem, the adaptation phase is applied. The returned solution and a new problem together form a new case that is incorporated into the case base during the learning step. The main disadvantage of CBR is the very strong influence of old design cases and the lack of sufficient adaptation methods to support innovative design. CBR has been used for the synthesis of separation systems.[20]

The emerging approach to process synthesis is based on the principle of equipartition of the driving forces.[21] According to this principle, design could be considerably improved by the equal distribution of the driving forces throughout the process by assuming that the rates of entropy production are proportional to the square of the driving forces. Although the fundamentals of this principle are the subject of discussion, the potential importance of the method is hard to overestimate. The example of application of this method has been presented for the design chemical reactor.[22]

Chemical Product Design

A new and very active field of research is chemical product design.[23–25] The activity of CAPE in product design concentrates on the analysis of the relations of material function, its properties, and composition.

There are three main classes of methods applicable to product design: experimental design, knowledge-based methods, and computer-aided molecular design.

There are two formulations of the problem of product design (Fig. 2).

The first is forward problem and occurs when the design starts from the given structure of a molecule or composition of the material, and aims at the determination of the properties of this material. The second approach, reverse problem, starts with the given properties of material and finds the molecular composition fulfilling the requirements. The experimental methods

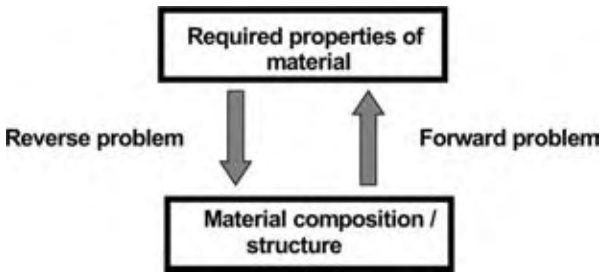


Fig. 2 Forward and reverse formulation of chemical product design problem.

are an example of the forward problem formulation. The knowledge-based as well as computer-aided molecular design methods could be used in forward or reverse problem formulation.

Experimental methods of product design

Historically, the oldest method is experimental product design, which evolved into sophisticated statistical methods, capable of handling very complicated mixing and processing problems. The CAPE activities concentrate on the statistical analysis of the processes that can lead to the identification of new products.

There are two major types of experimental design: planned experiments and high-throughput experiments (HTE). Recently, the second type has been attracting the interest of CAPE practitioners.

HTE is used to tackle problems where the parameter space is too large to be handled efficiently using conventional approaches. HTE consists in the use of miniaturized laboratory equipment, robots, screening apparatus, and computers. The main field of HTE application is in the determination of drug composition, multifunctional materials, coatings, and catalysts, as well as that of their formulation. The fundamental problems related to the application of CAPE to HTE consist in handling the complexity related to the amount of data and highly complicated phenomena under consideration. As a consequence, research is now on the design of databases, data mining, integration, and representation; conversion of data to knowledge; experimental control systems; and decision supporting system to facilitate ''hit-to-lead'' process that aims at maximizing the number of successful designs.

The following are examples of CAPE applications in HTE.

The catalyst has been designed as given in Ref.[26] The proposed process of ''knowledge extraction'' consists in planning HTE experiments in a way allowing for the discrimination of the models of catalytic reactions, determination of the kinetic constants, and relating them to the catalyst microstructure. The proposed forward modeling is realized by the application of the rules capturing human expertise, neural networks (NN), and genetic algorithms (GA).

An example of CAPE tool application to HTE in pharmaceutical design is the use of CBR for identifying the required conditions for protein crystallization.[27] The application of HTE and CAPE to drug development in the context of knowledge discovery is presented in Ref.[28]

Knowledge-based methods of product design

The use of CAPE knowledge-based methods is aimed mainly at limiting the cost and time of new product development. They are especially useful because of the existence of a huge amount of data and information stored by the companies. The common feature of these methods is the use of historical data, and information and generation, on this basis, of a new knowledge applicable to the actual product design problems. The proposed methods are rule-based reasoning, CBR, NN, GA, and data mining.

The CBR has been used in various applications, e.g., formulation of tablets, plastics, rubber, and agrochemicals. There are several works dealing with tablet formulation, e.g., Ref.[29] The CBR–rule-based hybrid system is aimed at retrieval as well as adaptation. The potentially interesting formulation is adapted by the application of the rule-based system. It allows to determine the most appropriate ingredients by the elimination of various conflicts and constraints related to the simultaneous presence of some of these in a tablet. The adaptation is realized using the voting mechanism, which selects the most frequently used ingredients among the retrieved cases. Another new application of CBR in chemical product design is presented in Ref.[30] The problem consists in the determination of the recipe for tire tread manufacturing (elastomers, silica, carbon black, accelerants, etc.) as well as determination of the conditions for compound mixing and vulcanization. The design of lubricating oils applying CBR has been introduced in Ref.[31] The main problem tackled is design of the additive formula that should be combined with the oil to create a lubricating agent.

The application of NN in product design concentrates on a few classes of materials. The fuel additives[32] have been used as combustion modifiers, antioxidants, inhibitors of corrosion, deposit controllers, etc. The authors have proposed a hybrid approach combining modeling and NN. The results obtained from modeling have been introduced into the net and compared with the experimental results. Such an approach enabled the tuning of the model allowing for the prediction of the build-up of the deposit as a function of the composition of the additives.

In Ref.[33] the author has proposed application of NN for the formulation of a rubber mixture. The net was used for the purpose of direct (what are the properties of the mixture when the composition is given) and indirect modeling (what should be the composition to ensure the required properties of the mixture). There have been adaptive usage of NN in learning how to map the relations between the inputs and outputs.

In Ref.[34] the authors have applied NN to predict the molecular surface of the acid, reactive and direct dyes. The mapping of the three-dimensional molecular surfaces into Kohonen network enabling the prediction of substantivity is an example of the direct formulation of the product design problem.

There are several applications of NN in drug design, especially in tablet formulation. They concentrate on finding the relation between the composition of the tablet and the required release time, prediction of the physicochemical properties of the substances based on their molecular composition, and formulation of the special-purpose tablets (e.g., fast disintegrating). An example of application of NN to the direct and reverse problems of chemical product design is given in Ref.[35] In the first case, the objective was to determine the optimal composition ensuring the required release of the active substance from the tablet. The second dealt with the determination of the optimal composition of the mixture of ketoprofen hydrogels composed of two gel bases, two penetration enhancers, and two solvents. The objective function was the required rate of penetration of active substance and skin irritation.

NN have also been used in the design of refrigerants. In Ref.[36] the authors have studied the prediction of the specific volume of refrigerant–absorbent couple in the function of system composition, pressure, and temperature using NN. The objective was to determine the properties of the couple such as ensuring zero ozone depletion, thermal resistance, high evaporation heat at low pressure, low specific volume of vapor, low solidification, and high critical temperatures.

The application of NN to the formulation of the polymer composites has been reviewed in Ref.[37] The authors presented a comprehensive overview of NN application in the estimation of fatigue life, design of unidirectional and laminate composites, assessment of wear of composites, and optimization of the processing.

GA are gaining a lot of attention in CAPE. This very promising tool of artificial intelligence has been used to design polymers[38] and catalysts.[39] Very often GA are combined with first-principles modeling and other artificial intelligence methods.

Nowadays the ''classical'' rule-based systems are relatively seldom encountered as a tool for the support of the product design. However, there are several examples of hybrid systems in the design of cosmetics (rule-based and first principles)[40] and pharmaceuticals (rule-based and NN).[41]

Data mining is a term used for a group of techniques used in finding useful patterns in the datasets. This approach consists in the analysis of huge data depositories in search of useful correlations between material properties and their composition. A simple example of data mining application in rubber design is given in Ref.[42]

CONCLUSIONS

CAPE is a multidisciplinary branch of process engineering spreading very rapidly into new fields outside the processing industries. The emerging fields for CAPE applications are proteomics, financial engineering, and material sciences, to name a few.[43–45]

In the near future, the main concerns of CAPE will probably focus on issues related to product design (molecular modeling for solving function–property–composition problems), supply chain management (cost reduction of raw materials, effective use of energy and its new sources), and life cycle assessment (mitigation of climate change, process sustainability).

The problems likely to be handled by CAPE methods and tools that to respond to the above-mentioned challenges are multiscale modeling, designing of metabolic processes, solving of quantitative and qualitative models, reuse of the designs, development of creativity-supporting methods for promotion of innovative design, uncertainty handling, effective methods for solving large-scale dynamic problems, and management of information overload.

REFERENCES

1. Sargent, R.W.H. Integrated design and optimisation of processes. CEP **1967**, *63*, 9.
2. John, W.R. Process synthesis: poised for a wider role. CEP **2001**, *97*, 59–65.
3. Douglas, J.M. *Conceptual Design of Chemical Processes*; McGraw-Hill Science: New York, 1988.
4. Biegler, L.T.; Grossmann, I.E.; Westerberg, A.W. *Systematic Methods of Chemical Process Design*; Prentice Hall: Upper Saddle River, 1997.
5. Turton, R.; Richard, R.C.; Whiting, W.B.; Shaeiwitz, J.A. *Analysis, Synthesis, and Design of Chemical Processes*; 2nd Ed.; Prentice Hall: Upper Saddle River, 2002.
6. Seider, W.D.; Seader, J.D.; Lewin, D.R. *Product and Process Design Principles*; 2nd Ed.; Wiley, 2003.

7. Westerberg, A.W. A retrospective on design and process synthesis. Comput. Chem. Engng **2004**, *28*, 447–458.
8. Barnicki, S.D.; Siirola, J.J. Process synthesis prospective. Comput. Chem. Engng **2004**, *28*, 441–446.
9. Biegler, L.T; Grossmann, I.E. Retrospective on optimization. Comput. Chem. Engng **2004**, *28*, 1169–1192.
10. Sahinidis, N.V. Optimization under uncertainty: state-of-the-art and opportunities. Comput. Chem. Engng **2004**, *28*, 971–983.
11. Siirola, J.J.; Rudd, D.F. Computer-aided synthesis of chemical process designs. Ind. Eng. Chem. Fundam. **1971**, *10*, 353.
12. Douglas, J.M. A hierarchical decision procedure for process synthesis. AICHE J. **1985**, *31*, 353.
13. Seader, J.D.; Westerberg, A.W. A combined heuristic and evolutionary strategy for synthesis of simple separation sequences. AICHE J. **1977**, *23*, 951.
14. Douglas, J.M. Process synthesis for waste minimization. Ind. Eng. Chem. Res. **1992**, *31*, 238.
15. Linninger, A.A. Metallurgical process design—a tribute to Douglas's conceptual design approach. Ind. Eng. Chem. Res. **2002**, *41*, 3797.
16. Siirola, J.J. Strategic process synthesis: advances in the hierarchical approach. Comput. Chem. Engng **1996**, *20*, S1637–SI643.
17. Jaksland, C.A.; Gani, R.; Lien, K.M. Separation process design and synthesis based on thermodynamic insights. Comput. Chem. Engng **1995**, *50*, 511.
18. Li, X.N.; Rong, B.G.; Kraslawski, A. Synthesis of reactor/separator networks by the conflict-based analysis approach. In *Proceedings of European Symposium on Computer Aided Process Engineering—12*; Grievink, J., van Schijndel, J., Eds.; Elsevier, 2002: 241–246.
19. Li, X.N.; Rong, B.G.; Kraslawski, A.A. Conflict-based approach for process synthesis with wastes minimization. In *Proceedings of European Symposium on Computer Aided Process Engineering—13*; Kraslawski, A., Turunen, I., Eds.; Elsevier, 2003: 209–214.
20. Seuranen, T.; Hurme, M.; Pajula, E. Synthesis of separation processes by case-based reasoning. Comput. Chem. Engng *in press.*
21. Sauar, E.; Ratkje, K.S.; Lien, M.K. Equipartition of forces: a new principle for process design and optimization. Ind. Eng. Chem. Res. **1996**, *35*, 4147–4153.
22. Kjelstrup, S.; Sauar, E.; Bedeaux, D.; van der Kooi, H. The driving force distribution for minimum lost work in chemical reactors close to and far from equilibrium. 1. Theory. Ind. Eng. Chem. Res. **1999**, *38*, 3046–3050.
23. Gani, R. Computer-aided methods and tools for chemical product design. Chem. Eng. Res. Des. **2004**, *82*, 1494–1504.
24. Cussler, E.L.; Wei, J. Chemical product engineering. AICHE J. **2003**, *49*, 1072–1075.
25. Hill, M. Product and process design for structured products. AICHE J. **2004**, *50*, 1656–1660.
26. Caruthers, J.M.; Lauterbach, J.A.; Thomson, K.T.; Venkatasubramanian, V.; Snively, C.M.; Bhan, A.; Katare, S.; Oskarsdottir, G. Catalyst design: knowledge extraction from high-throughput experimentation. J. Catal. **2003**, *216*, 98–109.
27. Jurisica, I.; Rogers, P.; Glasgow, J.I.; Fortier, S.; Luft, J.R.; Wolfley, J.R.; Bianca, M.A; Weeks, D.R.; DeTitta, G.T. Intelligent decision support for protein crystal growth. IBM Syst. J., **2001**, *40*, 394–409.
28. Gardner, C.R.; Almarsson, O.; Chen, H.; Morissette, S.; Peterson, M.; Zhang, Z.; Wang, S.; Lemmo, A.; Gonzalez-Zugasti, J.; Monagle, J.; Marchionna, J.; Ellis, S.; McNulty, C.; Johnson, A.; Levinson, D.; Cima, M. Application of high throughput technologies to drug substance and drug product development. Comput. Chem. Engng **2004**, *28*, 943–953.
29. Craw, S.; Wiratunga, N.; Rowe, R. Case-based design for tablet formulation. Proceedings of the 4th European Workshop on Case-Based Reasoning, 1998; Springer-Verlag, 358–369.
30. Bandini, S.; Colombo, E.; Colombo, G.; Sartori, F.; Simone, C. The role of knowledge artifacts in innovation management: the case of a chemical compound designer CoP; http://www.disco.unimib.it/simone/gestcon/c&T03-Bandini-et.al.doc (accessed 29 March 2004).
31. Shi, Z.; Zhou, H.; Wang, J. Applying case-based reasoning to engine oil design. Artif. Intell. Engng **1997**, *11*, 167–172.
32. Sundaram, A.; Ghosh, P.; Caruthers, J.; Venkatasubramanian, V. Design of fuel additives using neural networks and evolutionary algorithms. AIChE J. **2001**, *47*, 1387–1406.
33. Borosy, A. Quantitative composition–property modelling of rubber mixtures by utilising artificial neural networks. Chemomet. Intell. Lab. Syst. **1999**, *47*, 227–238.
34. Greaves, A.J.; Gasteiger, J. The use of self-organising neural networks in dye design. Dyes Pigments **2001**, *49*, 51–63.
35. Takayama, K.; Fujikawa, M.; Obata, Y.; Morishita, M. Neural network based optimization of drug formulations. Adv. Drug Deliv. Rev. **2003**, *55*, 1217–1231.

36. Sözen, A.; Özalp, M.; Arcaklyoglu, E. Investigation of thermodynamic properties of refrigerant/absorbent couples using artificial neural networks. Chem. Eng. Process. **2004**, *43*, 1253–1264.
37. Zhang, Z.; Friedrich, K. Artificial neural networks applied to polymer composites: a review. Compos. Sci. Technol. **2003**, *63*, 2029–2044.
38. Venkatasubramanian, V.; Chan, K.; Caruthers, J.M. Evolutionary design of molecules with desired properties using genetic algorithms. J. Chem. Inf. Comp. Sci. **1995**, *35*, 188–195.
39. Corma, A.; Serra, J.M.; Chica, A. Discovery of new paraffin isomerization catalysts based on SO_4^{2-}/ZrO_2 and WO_x/ZrO_2 applying combinatorial techniques. Catal. Today **2003**, *81*, 495–506.
40. Wibowo, Ch.; Ng, K.M. Product-oriented process synthesis and development: creams and pastes. AIChE J. **2001**, *47*, 2746–2767.
41. Guo, M.; Kalra, G.; Wilson, W.; Peng, Y.; Augsburger, L.L. A prototype intelligent hybrid system for hard gelatine capsule formulation development pharmaceutical technology. Pharmaceutical Research **2002**, *26*, 44–60.
42. Chen, N.; Zhu, D.D.; Wang, W. Intelligent materials processing by hyperspace data mining. Engng. Applic. Artif. Intell. **2000**, *13*, 527–532.
43. Schneider, R.; Marquardt, W. Information technology support in the chemical process design life cycle. Chem. Eng. Sci. **2002**, *57*, 1763–1792.
44. Gani, R. CAMD: computer aided molecular design—examples of applications; http://www.capec.kt.dtu.dk/documents/overview/camd-overview-2004.pdf (accessed 24 April 2005).
45. Grossmann, I.E. Challenges in the new millennium: product discovery and design, enterprise and supply chain optimization, global life cycle assessment. Comput. Chem. Engng **2004**, *29*, 29–39.

Conductive Polymers

Ronald W. Gumbs
Gumbs Associates Ltd., East Brunswick, New Jersey, U.S.A.

INTRODUCTION

Organic polymers that possess the electronic, magnetic, and optical properties of metals are known as conductive polymers (CPs). Because of their conjugated π electron backbones, they can be oxidized or reduced more easily and more reversibly than conventional polymers with charge-transfer agents, also commonly called dopants, a term borrowed from condensed matter physics. While retaining some of the mechanical properties of polymers, they do not melt or dissolve in common organic solvents, a major impediment to their widespread commercialization in the same manner as traditional plastics. The same electronic structure that confers electrical conductivity to these polymers also contributes to their intractability and instability.

Conductive Polymers is a new field in materials science utilizing the unique electronic properties of a class of easily synthesized polymers with the predominant property of high and controllable conductivity. Written from the perspective of an industrial polymer chemist, this entry reviews the search for processible CPs and covers the synthesis, properties emanating from their conductivity, and applications with emphasis on the processing of these polymers for various commercial applications.

OVERVIEW

Ranby[1] has appropriately classified polymers into four generations that are available as commercial products and summarized in Table 1.

According to Ranby,[1] the fourth generation of polymers started with the discovery[2] that polyacetylene (PA) could be rendered electrically conductive by doping; the chemistry of its doping is described elsewhere.[3] Curiously enough, some of the major CPs were first prepared in the 19th century before the general concept of polymers was well understood. According to Naarmann,[4] the term PA was used in 1866 and a high molecular weight copper acetylene was formed when acetylene was passed over copper powder below 250°C. Thiophene and pyrrole were polymerized in 1883 and 1888, respectively. Polyaniline (PANi) was synthesized in 1891 and initial attempts to prepare polyphenylene (PP) were made as far back as 1842.

The Concept of Doping

Conjugated polymers are either electrical insulators or semiconductors, and those that can have their conductivity increased by several orders of magnitude above that of semiconductors are commonly referred to as electronic polymers. Indeed, the conductivity attainable by an electronic polymer has been increased infinitely by the discovery[5] of superconductivity in regio-regular poly(3-hexylthiophene) at 2°K.

The concept of doping[3,6] is what distinguishes CPs from most organic polymers. During doping, an organic polymer, either an insulator or semiconductor possessing a low conductivity, typically in the range from 10^{-10} to 10^{-5} S/cm, is transformed into a synthetic metal (ca. 1–10^4 S/cm). In essence, the controlled addition of small concentrations of chemical species results in dramatic changes in the properties of the polymer. Doping can be reversed to produce the original pristine polymer with minor or no degradation of the polymer backbone.

Synthetic metals are prepared by oxidation or reduction in which some of the π electrons are removed (*p*-doping) or external electrons added (*n*-doping). Simultaneously, there is an insertion of counterions near the charged sites, resulting from the reduction or oxidation of the dopants. These counterions stabilize the doped state, and both doping and undoping can be carried out chemically or electrochemically.

The backbone of the CP in its doped state primarily consists of a delocalized π-system, but in its undoped state, the polymer may have a conjugated backbone such as in *trans*-PA or it may have a nonconjugated backbone as in the leucoemeraldine and emeraldine base forms of PANi.

There are many reviews of the vast literature on CPs, including: 15 books;[6–21] compilations in various encyclopedias;[4,22–28] and an excellent overview by Kanatzidis.[29] They cover all aspects of the theory, energy band structure, chain structure, morphology, comparison of conductivities with metals, semiconductors and insulators, doping, synthesis and characterization, electrochemistry, processing, and potential

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007714

Table 1 Evolutionary development of synthetic polymers

Generation	Period introduced	Examples	Distinguishing features
First	1930–1950	Vinyls, polystyrenes, acrylics, epoxies, phenolics	Low density and ease of processing for applications in packaging, paints, coatings, laminates, and adhesives
Second	1950–1965	High-density polyethylene, isotactic polypropylene, polycarbonate, polysulfones, linear polyesters, synthetic rubbers	Improved mechanical strength
Third	1965	PPS, polyimides, aromatic polyesters, aromatic polyamides, fluoropolymers, thermoplastic elastomers	High mechanical strength, very high melting point, superior chemical resistance
Fourth	1977	Intrinsically CPs	Electrical conductivity

commercial applications of the fourth generation of polymers.

CPs AS SEMICONDUCTORS

The electrical property of a material is determined by its electronic structure, and the relevant theory that explains the electronic structure in the solid state is band theory. This theory, however, does not fully explain conductivity in polymers. It is noteworthy that the energy spacing between the highest occupied and lowest unoccupied bands is called the bandgap; the highest occupied band is called the valence band and the lowest unoccupied band is called the conduction band. The bandgaps of insulators and semiconductors are wide and narrow, respectively; there are no bandgaps in metals.

According to band theory, the electrical properties of a solid depend on how the bands are filled. There is no conduction when the bands are filled or empty. If the material has a narrow bandgap, thermal excitations of electrons from the valence band to the conduction band occur, giving rise to conductivity in classical semiconductors.

Because CPs conduct current without having a partially empty or partially filled band, concepts from solid state physics are used to explain the electronic phenomenon in these polymers. Thus, chemists refer to solitons, polarons, and bipolarons when they discuss the fundamentals of CPs. And, Fig. 1 shows the energy level diagram for an undoped, slightly doped, and heavily doped polymer to further illustrate the concept of doping.

The semiconductor band structure of CPs enables electronic excitations or electron transfer, e.g., from the valence to the conduction band, leading to most of the properties that are of interest. Excitation of electrons from the valence band to the conduction, e.g., by photons, typically yields excited state properties such as photoluminescence and nonlinear optical (NLO) properties.

Oxidation of the CP, i.e., essentially a removal of electrons from the valence band leads to the presence of charges on the CP. These charges are strongly delocalized over several monomeric units in the CP. They also cause a relaxation or structural distortion of the geometry of the CP to a more energetically favored conformation. In addition, a charge may also be donated to the conduction band of the CP, causing reduction of the CP.

SYNTHESIS OF CPs

Many of the procedures used to prepare neutral CP precursors are commonly employed in the polymer industry. Hence, the polymerization methods of Ziegler-Natta, Friedel-Crafts, and nucleophilic displacements yield PA, PP, and polyphenylene sulfide (PPS), respectively. Other methods include: Dields-Alder elimination, Wittig, and electrochemical coupling. The procedures used to prepare CPs in this vast arsenal are generally divided into two main categories: chemical and electrochemical.

Chemical Methods

There are two general routes to the synthesis of conjugated polymers: addition polymerization of unsaturated monomers and condensation polymerization or stepwise coupling of monomers with difunctional groups.

Addition polymerization

An alkyne, such as acetylene or any of its derivatives, undergoes addition polymerization to yield PAs. A typical procedure involves adding acetylene under 80 kPa (600 mm Hg) of acetylene to a solution of $Al(C_2H_5)_3 : Ti(OC_4H_9)_4$ (4 : 1) in toluene (0.1–0.2 molar

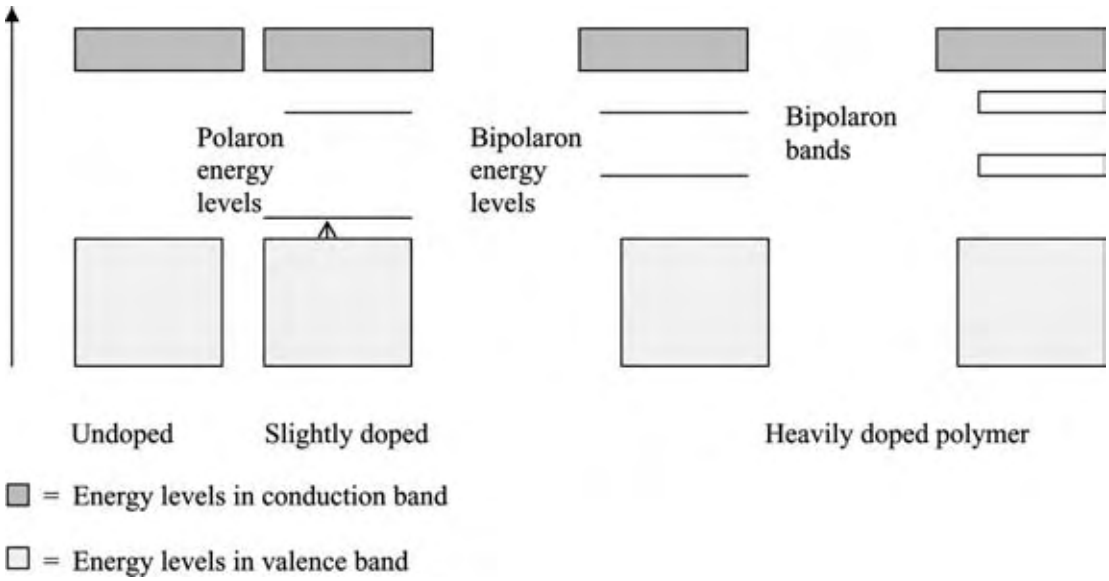


Fig. 1 Band structure of a CP. *(View this art in color at www.dekker.com.)*

in Ti) at –78°C and 80 kPa (600 mm Hg) for 20 min. Polyacetylene is also prepared by ring-opening metathesis polymerization of cyclo-octatetraene. A summary of the methods used to prepare PA and its derivatives is included in reviews.[13,21]

$$n\mathrm{HC}{\equiv}\mathrm{CH} \longrightarrow -(-\mathrm{CH}{=}\mathrm{CH}-)_n-$$

Condensation polymerization

This method has been used successfully to prepare polyarylenes and poly(arylene vinylenes) but not in the synthesis of PA and its derivatives. It is a straightforward extension of well-established aryl coupling reactions, which have been comprehensively reviewed.[13,21] Essentially, stepwise coupling of monomers with difunctional groups is carried out in a more controlled fashion than during addition polymerization, but conversion rates are low and purification of monomers is necessary. It includes: Wurtz–Wittig coupling of dichlorides with sodium or potassium; Ullman coupling of the di-iodide with copper; Grignard coupling of the dibromide with magnesium in ether and a complex of nickel chloride; diazonium coupling; oxidative coupling; and elimination reactions.[4]

Oxidative coupling of monomers, such as pyrrole, thiophene, aniline, aromatic, and other heteroaromatic systems is the most widely used method. In brief, it involves the use of an oxidant that reacts with the monomer to generate a radical cation and a radical anion. The former combines with another radical cation to form a dimer, which undergoes further oxidation to propagate the polymerization.

There are, of course, numerous examples of elimination reactions involving the loss of water, hydrogen halides, oxygen, nitrogen, and other small molecules with the formation of a double bond. Both direct and indirect methods are employed, including processes with soluble precursor polymers. It is interesting to note that polymerization of phenyl vinyl sulfoxide and subsequent elimination of benzenesulfonic acid lead to the formation of PA.

Electrochemical Methods

Electrochemical oxidation of resonance stabilized aromatic molecules is a common and singular method used to prepare CPs. It involves the oxidative coupling of monomers and is an ideal method to prepare conductive films that show reversible redox reaction. The polymerization can be monitored by in situ combination of the electrolysis system and appropriate spectroscopic or electro-analytical techniques; the electrochemistry of CPs is the subject of a recent review[13] with 748 references.

Polymerization of monomer is initiated by species, typically radical cations, formed in an electrode reaction during electrolysis. For each electron transferred through solution, a corresponding chemical reaction must occur on the potential of the electrode together with the discharge potential of the various species present. The system consists of an electrolysis cell with electrodes, electrolytes, monomer, and solvent. Important steps in the process are: initiation takes place at the surface of the electrode and propagation is either homogeneous or heterogeneous; the growth of the polymer may occur on the electrode surface or in the

bulk of the electrolysis solutions, which is similar to reaction in chemically initiated polymerization; chain transfer and termination are also generally similar to the corresponding steps in the chemical method.

SOLUBILITY OF CPs

The discovery[22] in 1986 that AsF_5-doped CPs are soluble in AsF_3 created much surprise and skepticism among workers in the field. It generated a search for less exotic solvents and, as a result, the literature contains many reports of soluble CPs. On closer examination, however, many of these materials in their doped states do not form genuine solutions. Evidently, the authors of these reports were unaware that light scattering or ultracentrifugation of mixtures of solvents and polymers can in fact determine whether the mixture is a solution or dispersion.

For purposes of clarity, it is perhaps instructive to describe the distinction between a solution and dispersion. The term ''solution'' refers to a complete solvent shell around an atom, molecule, or a single polymeric chain on the molecular scale. It refers to a single phase and a homogeneous solution. The term ''dispersion'' means that particles larger than 1 nm are surrounded by and embedded in a medium. Thus, dispersions are not homogeneous, but consist of at least two phases. This heterogeneous system contains a pure solvent phase with no dissolved molecules and the pure dispersed phase. Both phases approach each other at the interfaces.

Solution in Corrosive Solvents

There are reports[22] that doped CPs are soluble in unusual solvents, which are difficult to handle for many reasons. Polyaniline dissolves in concentrated acids such as sulfuric acid, methane, and trifluoromethane sulfonic acids; concentrations as high as 20% by weight are possible. After precipitation of the polymer in methyl alcohol, the polymer retains its conductivity and exhibits some crystallinity. And recently, MacDiarmid and others have spun fibers from concentrated H_2SO_4 solutions.

PPS doped with AsF_5 dissolves readily in AsF_3,[22] but cast films are no longer soluble. Frommer has suggested that the solubilization mechanism involves solvation of both reactive radical intermediates and dopant counterions.[22] In addition, dialkyl esters of phosphoric acid dope PANi and render it soluble in certain solvents such as decalin. The resulting solutions can be mixed with conventional polymers and used to prepare films and fibers.

On the other hand, Wessling insists that these mixtures ''look like solutions (clear green), but they are dispersions.'' In comprehensive fashion, he shows that there is a decrease in viscosity and absorbance over time, which is accompanied with flocculation. He also presents convincing evidence to support his dispersion theory. He does concede, however, that the AsF_5/AsF_3 system is a genuine solvent for CPs based on the evidence.

Solution in Water

The sodium salts and the acids of poly-3(2-ethanesulfonate) thiophene and poly-3-(4-butanesulfonate) thiophene are the first known examples of water soluble CPs;[21] many other examples of water-soluble, self-doped CPs are given in recent reviews.[20] Indeed, it is well known that the incorporation of highly soluble sulfonate groups on the CP backbone is a common method of achieving water solubility in a doped CP. The two most popular routes are the use of sulfonated polymeric dopants and the polymerization of sulfonated monomers.

Substituted Derivatives

Polymers formed from the electropolymerization of 3-thiophenecarboxylic acid, 3-cyanothiophene, and 3-nitrothiophene are soluble in acetonitrile. In addition, substitution of alkyl groups with more than six carbon atoms at the 3-position of the thiophene molecule and subsequent polymerization renders the reduced form of the polymer soluble. Indeed, solubility increases with chain length of the alkyl group and the de-doped or reduced forms are truly soluble in solvents such as toluene, THF, and methylene chloride.[17]

PROCESSING OF CPs

In view of the chemical nature of CPs, many clever schemes have been devised to achieve processibility while maintaining electrical conductivity. Most reports of solubility and processibility involve many schemes, such as the use of corrosive solvents, e.g., concentrated H_2SO_4 and AsF_3, processing of the de-doped polymers followed by re-doping, dispersion of the CPs as colloids, thermoforming of blends with thermoplastics, incorporation of solubilizing groups, the use of soluble precursor polymers, substitution of long-chain alkyl groups, copolymerization, and in situ polymerization.

Solution Processing

The processibility of CPs is of course related to their solubility, and many applications involve processing of solutions.

Processing of Neutral Form Followed by Doping

The most popular method of processing is based on limited evidence that the reduced or de-doped form is soluble. It is widely applied to PANi based on extensive experimental studies by MacDiarmid and his group. In brief, a solution of the de-doped and neutral form of the CP (the emeraldine base) in an organic solvent is used to coat a substrate and the cast film is chemically doped after drying.

There are, however, some problems with this method. First, high molecular weight fractions are not soluble in the strongest solvent, *N*-methyl pyrrolidone, for emeraldine base. Second, Wessling has shown by membrane filtration and photon correlation spectroscopy that these transparent, clear blue liquids are in fact dispersions and not solutions.

Based on the PANi model, it is generally assumed that this method is applicable to all CPs as a means of processing. It is known that some CPs exhibit some solubility in organic solvents in their de-doped states, but the fractions with the highest molecular weight do not dissolve and cast films frequently do not re-dissolve.

Counterion-Induced Processibility

Cao and coworkers reported that problems in processing of doped PANi could be solved by the use of functionalized protonic acids. Indeed, solutions of camphor–sulfonic acid doped with PANi in *m*-cresol show liquid crystallinity and high electrical conductivity, according to Smith and his coworkers. This prompted additional work with this and other polymers. Unfortunately, the conductivity of cast films of solutions of polypyrrole (PPy) doped with dodecyl benzene sulfonic acid in *m*-cresol is low (ca. 10^{-3} S/cm).

Copolymerization

Since conventional polymers are very miscible with organic solvents and melt-processible, it was thought that a true copolymer with conjugated and nonconjugated segments might lead to processibility. Consequently, many attempts have been made during the past 20 yr to prepare electrically conductive vinyl copolymers that are processible. And while conductive random, alternating, block, and graft copolymers have been prepared, the processibility of conventional polymers has not been achieved.

In the most recent attempt,[30] an inverted emulsion process was used to synthesize conducting copolymers of aniline and acrylonitrile using benzoyl peroxide as the oxidizing agent. The solubility of the copolymers in dimethyl sulfoxide was confirmed by electronic absorption spectroscopy, and the conductivity of the copolymers ranged from 0.01 to 0.04 S/cm. It is interesting to observe that MacDiarmid and others are skeptical of any finding that aniline is capable of copolymerizing with vinyl monomers.

Processing via In Situ Polymerization

Many commercial applications of CPs involve coatings on various substrates and the most important method for processing CPs as coatings is the in situ polymerization of monomers on the surfaces of substrates. It is commonly used to coat fabrics and to fabricate Langmuir–Blodgett films.[27]

Dispersibility of CPs

Spinks and coworkers[26] have reviewed the preparation of sterically stabilized PA, PP, and PANi colloids. There is a large body of significant work on the CP coated particles, CP-silica nanocomposites, and the other CP colloids.

Colloidal Dispersions

Techniques using colloidal dispersions of CPs are widely used to process CPs as solutions, i.e., sols and gels.[26] The initial challenge, however, is to prepare dispersions that are stable and to this end additives such as surfactants and soluble polymers are employed. The most commonly used method of preparing stable dispersions is to polymerize the monomer, e.g., pyrrole in an aqueous solution of a water soluble polymer such as methylcellulose and subsequently washing the CP with water at 70°C where gelation occurs in order to remove the unreacted monomer, oxidant ($FeCl_3$), and $FeCl_2$. Other water soluble polymers are used, including the following: poly(vinyl alcohol), poly(acrylamide), and poly(vinyl pyrrolidone). The problem is that the washing in hot water is not as effective as with the cellulose ether.

Alloying and Blending of CPs with Conventional Thermoplastics

In order to disperse low concentrations of CPs in thermoplastics[28] and elastomers,[25] it is necessary to prepare highly structured fine aggregates to permit percolation in the polymer blend. With better polymerization techniques and new blending methods, the percolation threshold decreases. Indeed, a critical concentration as low as 1.5 vol% and a saturation

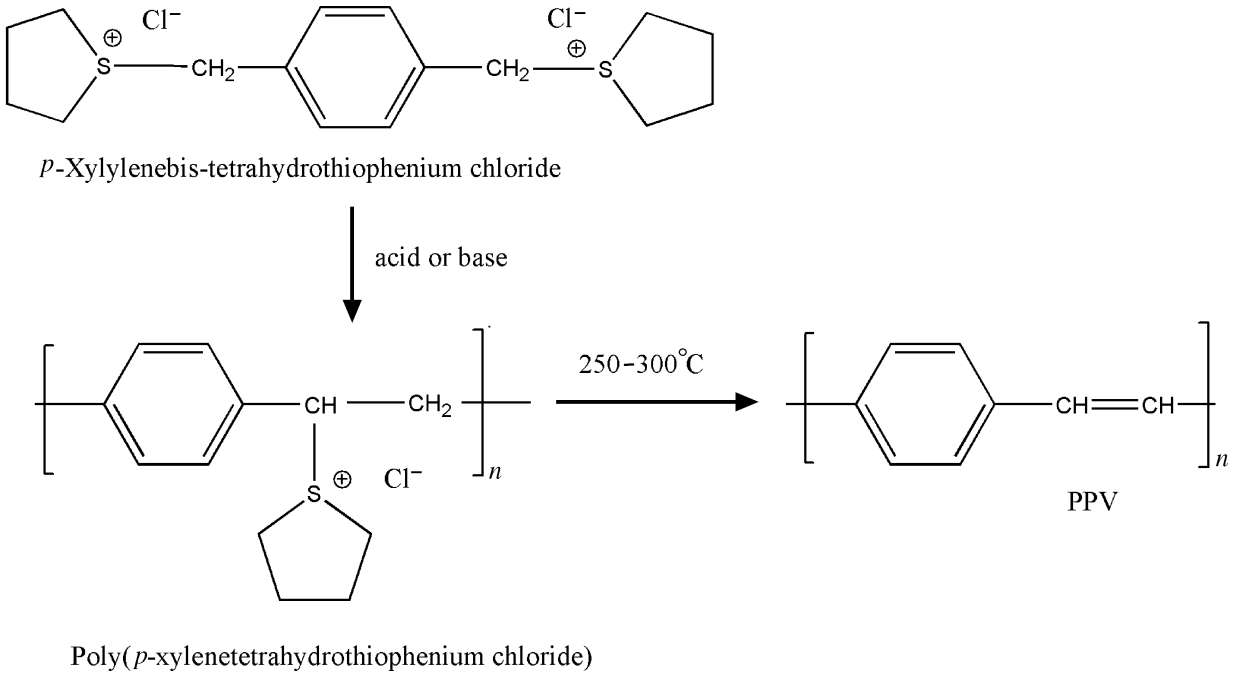


Fig. 2 Preparation of PPV via a precursor route.

concentration of 25 vol% are possible with PANi in PMMA.

Alloying and Blending of CPs with Conventional Thermosets

It is a much more formidable problem to disperse a CP, such as PANi, in a thermosetting polymer system, such as a polyurethane or an epoxy resin. As thermosets are reactive, they react with active hydrogens of the CP, resulting in de-doping and loss of conductivity. The di-isocyanate groups react with secondary amine groups and with residual water in PANi. Encapsulation of the CP leads to a serious decrease in conductivity of the fully cured blend. But some clever ruses have been devised to prepare conductive epoxy resins by using di-anhydride curing agents with emeraldine base. Apparently, the di-anhydride is an effective doping agent, whereas amine curing agents are de-doping agents.

Precursor Routes

Processing through a precursor route involves the use of an intermediate, processible polymer that can be later converted into the fully conjugated material. It is most often used with poly(*para*-phenylene vinylene (PPV), a relatively stable and insoluble polymer that can be manipulated as soluble precursor polymers to form films and fibers. There are many different precursor routes to PPV, including the sulfonium and alkoxy derivatives as described in a recent review.[13] Fig. 2 shows a typical chemical reaction commonly used to prepare films of PPV. In fact, a 2.5% solution of the precursor in water is commercially available. In addition, nine monomeric precursors for PPV derivatives are available to produce red, orange, and green emitters for light-emitting diodes (LED) devices.

GENERAL CLASSES AND APPLICATIONS OF CPs

Table 2 contains idealized structures of some CPs with typical dopants and values for the conductivities of thin films. The exact structures of PPy and polythiophene (PT) are unknown. Polyacetylene is the most crystalline and PANi can exist in several oxidation states with electrical conductivities varying from 10^{-1} S/cm to the values reported in Table 2. In its undoped state, PPS is an engineering thermoplastic with a conductivity of less than 10^{-16} S/cm. Upon doping with AsF_5, conductivities as high as 200 S/cm have been obtained after casting a film from a solution of AsF_3.[22]

Classes of CPs

Among the many conjugated polymers, PA, PANi, PPy, PT, PPS, and PPV have received the most

Table 2 Structure and conductivity of some common CPs

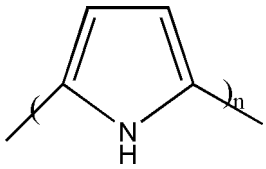
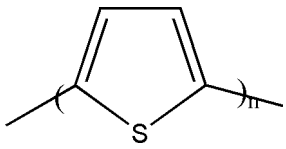
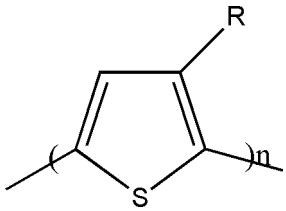
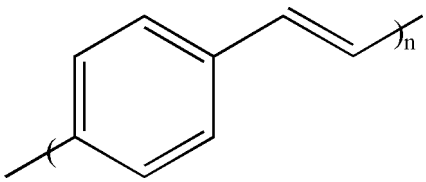
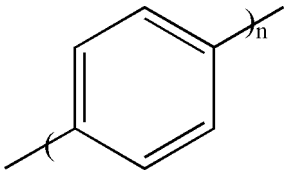
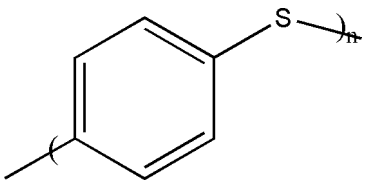
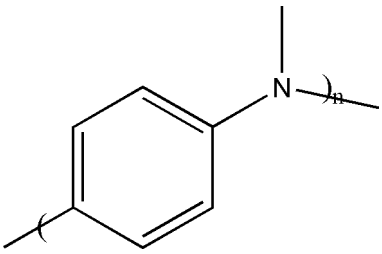
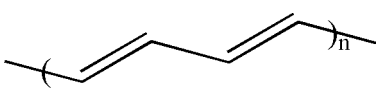
Polymer	Structure	Dopants	Approximate conductivity (S/cm)
PPy		BF_4^-, $C1^-$	100–500
PT		BF_4^-, $C1O_4^-$, $FeC1_4^-$	1,000
Poly(3-alkylthiophene)		BF_4^-, $C1O_4^-$, $FeC1_4^-$	1,000
PPV		AsF_5	1,000
PP		AsF_5, Li, K	1,000
PPS		AsF_5	200
PANi		HCl	5–100
PA		I_2, Br_2, Li, Na, AsF_5	1,000

Table 3 Types of CPs

Polymer class	References
PAs	[20]
PANis	[1,20]
PPys	[5]
PTs	[9,13,20]
Silicon-containing PTs	[20]
Poly(thienylene vinylene)	[13]
Poly(*p*-phenylenes)	[13,20]
Poly(l,4-phenylene sulfide)	[22]
PPV	[13]
Poly(azines)	[20]
Poly(azulenes)	[9]
Cyclopolymers	[13,24]
Poly(quinolines)	[9]
Poly(toluidines)	[20]
Poly(diacetylenes)	[21]
Poly(diphenyl amine)	[21]
Poly(*N*,*N*′-diphenylbenzidine)	[21]
Metallophthalocyanine-based conducting polymers	[13,20]
Organometallic conducting polymers	[20]
Self-doped conducting polymers	[17,20]
Low bandgap conducting polymers	[13]
Conductive copolymers	[15,20]
Ionically conducting polymers	[15]
Nonconjugated conducting polymers	[23]

attention. A list of the general classes of CPs with references is shown in Table 3.

Applications of CPs

All books, reviews, and entries on CPs describe the potential applications. Chandrasekhar[21] and others[11–20] have reviewed in comprehensive fashion the applications of CPs, including: batteries; sensors; electro-optic and optical devices; microwave- and conductivity-based technologies; electrochromic devices; electrochemomechanical and chemomechanical devices; corrosion protection; semiconductor, lithography, and electrically related applications—photovoltaics, heterojunction, and photoelectrochemical cells; capacitors; electrolytic and electroless metal plating; CP-based ''molecular electronic devices;'' catalysis and delivery of drugs and chemicals; membranes; and LEDs.

Conductive polymers have been studied for a variety of applications in batteries, primarily for secondary or rechargeable batteries. They have been investigated as the anode as well as cathode materials, with cathode materials in Li secondary batteries being the major focus. Suggested uses have included CP (anode/cathode) batteries, lead-acid batteries, and Zn batteries. Polyaniline and polypyrrole have been the primary focus, but other common CPs have been investigated. To be sure, the advantages of CPs as battery electrodes are light weight, low cost, processibility into odd shapes and sizes, less corrosive nature, and compatibility with both organic liquid and solid electrolytes.

Conductive polymers are attractive materials for application as LEDs because they require low voltage, of the order of 2–5 V, and they are processible in their undoped state to yield films and odd shapes/sizes. In addition, they are inexpensive, and exhibit increasingly higher quantum efficiencies. Their emission colors are tunable by tailoring the CP backbone, i.e., copolymerization, layering or combining multiple CPs, or several other techniques, from the visible region to the near IR. Whether they can displace liquid crystals in displays is an open question as of this writing.

Conductive polymers have been investigated for applications as sensors, with sensing accomplished through the following modes:

Conductometric, using change in conductivity.

Potentiometric, monitoring the change of open circuit potential at a CP electrode.

Amperometric, monitoring the current at a CP electrode.

Voltammetric, monitoring the change in current while varying the applied potential at a CP electrode.

Gravimetric, entailing the effect of a weight change in a CP.

Optical, typically involving the effect of doping or analyte binding on the optical properties of a CP.

pH-Based, using the effect of change in pH in the sensor environment.

The electro-optic and optical devices based on CPs can be classified into three categories: those based on second order NLO properties; those based on third order NLO properties; those based on properties other than NLO; CP-based lasers. The driving forces behind the development of these devices are discussed elsewhere.[12]

The military applications of CPs have been extensively investigated during the past 25 yr with dramatic results in the following areas: electromagnetic impulse (EMI) shielding; conductive coatings and composites for electrostatic discharge dissipation (ESD) or

antistatic applications; passive absorbers for microwave-region radiation for radar cross section (RCS) reduction; dynamic (switchable) microwave absorption and RCS control; microwave smoke camouflage; and several other applications that are classified.

Another application of CPs with equal significance to the military is in the area of electrochromic devices, which may be of two types:

Mode of function: Reflectance-, transmittance-, and cumulative-mode devices.

Spectral region of response: Near-UV-visible, near-UV-visible-NIR, IR, visible-NIR-IR, micro/millimeter-wave, and wide spectral.

Because of the demonstrated success achieved with electrochromism in CPs, there is an ongoing search for a CP-based device capable of modulating the entire spectrum from the visible through the IR to the microwave region.

The reversible reaction between the doped and undoped CP is significant in terms of military applications. When combined with the physical, chemical, optical, and electronic properties of both forms, it is easy to comprehend why CPs are truly the fourth generation of polymers in modem warfare. In the area of stealth, PANi is used in the smart skins of radar-evading aircraft. But, as the microwave absorption is converted to thermal energy, these aircraft can be observed by their heat signature. And, because there is a significant difference between the IR specular reflectance of doped and undoped CPs, devices based on CPs can be used to track, camouflage, and evade incoming heat seeking missiles. In the visible and near IR regions of the spectrum, electrochromic devices with exceptional contrast are available.

Polyaniline has been successfully exploited because of its relatively low cost, acceptable environmental stability, and ease of processing. And even though PPy is inherently more difficult to process, it has been used as a coating on fabrics for several applications—EMI shielding and camouflage applications—because of its higher conductivity.[13]

The electrochromic properties of PANi from the visible to the microwave region can be improved by copolymerization of aniline with other aromatic amines such as diphenyl amine and *N*,*N*′-diphenylbenzidine. Devices with solid electrolytes, rapid switching, high contrast, and acceptable cycling stability have been discovered and successfully exploited by Chandrasekhar.[21]

One of the more fascinating applications is the electromechanical actuation with CPs, leading to artificial muscles, wings, and propellers. The basis behind such actuation is that CPs swell on doping, with an increase in volume of about 35%. Thus, if a bilayer of a CP and a flexible material that does not swell or contract is created, electrochemical or chemical redox of the CP will result in bending of the bilayer.

Corrosion of metals is a serious problem which has been successfully addressed with CPs, and consequently anticorrosion coatings represent the largest market for CPs. The driving forces behind this development are: the issue of cracks and pinholes in conventional protective coatings, as well as slow diffusion of the corroding species to the metal surface; environmental issues, such as the unacceptability of chromates; cost; and ease of application. The principles underlying the use of CPs for this application are based on the fact that metal corrosion is primarily an oxidative process, involving the loss of electrons from the metal to an oxidizing species in the environment. Hence, the fundamental nature of corrosion is electrochemical, and a CP is capable of serving as an in situ oxidant or anodic protective material. In other words, it oxidizes the surface of the metal with which it is in contact, thereby being itself reduced.

COMMERCIAL PRODUCTION

Very few CPs are produced in bulk quantities. Polyphenylene sulfide, a member of the third generation of polymers, was produced in bulk quantities many years before CPs were established and its dopability was elucidated. Polyethylenedioxythiophene is commercially available as a water-based colloidal dispersion (Baytron P water dispersion), and presumably as dispersible powders. The powders with a conductivity of 5–10 S/cm can be dispersed in thermoplastic polymers and in organic solvents such as xylene. Polyaniline doped with dodecylbenzene sulfonic acid and complexed with zinc dodecylbenzene sulfonate is commercially available as a powder, which can be dispersed in polyolefins. The same polymer doped with *p*-toluenesulfonic acid is also available as a dispersible powder, Ormecon, and in a predispersed form for solution processing in polar and nonpolar media. Based on Ormecon PANi, there are many commercial products marketed for many different applications.

Other CPs are produced in large quantities inhouse by many companies for marketing as products for specific applications. Milliken Research Corporation, for example, supplies PPy-coated textiles for microwave absorption, ESD, and other uses.[13] Ashwin-Ushas Corporation markets several different electrochromic devices for the visible, IR, and microwave regions. Sigma-Aldrich, Inc. offers a wide range of CPs from the main polymer classes and a complete complement of monomer precursors.

For the bulk production of CPs, the chemical method is used. It is essentially a scaleup of the procedures used in the laboratory. In the case of PANi, benzidine is a byproduct formed during oxidation of aniline and is a carcinogen. Consequently, it must be removed by extraction or copolymerization during oxidation. In addition, the aqueous waste must be either treated before release into the environment or recycled.

Electrochemical methods involving the continuous production in a flow-through cell and the partially immersed cylindrical horizontally revolving electrode are being used to produce CPs inhouse. For instance, flexible films of PPy with excellent mechanical properties and chemical stability have been produced continuously by electrochemical polymerization. The films with an average thickness of 80 μm are commercially available.

CONCLUSIONS

The main characteristic of a CP is a conjugated backbone that can be subjected to oxidation or reduction, resulting in what is frequently termed *p*-type of *n*-type doping, respectively. Doping leads to a dramatic increase in electrical conductivity, but also to a decrease in stability, solubility, and fusibility. A vast arsenal of synthetic methods is available to prepare CPs, which can be dispersed in conventional polymers to form processible blends or alloys. To be sure, polymers have numerous technological applications because of their low cost and density combined with flexibility. But the ultimate goal of preparing a pure polymer—a synthetic metal—with the physical properties of a metal remains as elusive as that of the alchemist's dream of converting lead into gold.

REFERENCES

1. Ranby, B. Background—polymer science before 1977. In *Conjugated Polymers and Related Materials—The Interconnection of Chemical and Electronic Structure*, Proceedings of the Eighty-first Nobel Symposium, Lulea, Sweden, Jun 13–18, 1991; Salanek, W.R., Lundstrom, I., Ranby, B., Eds.; Oxford University Press: New York, 1993; 24–25.
2. Chiang, C.K.; Fincher, C.R., Jr.; Park, Y.W.; Heeger, A.I.; Shirakawa, H.; Louis, E.J.; Gau, S.C.; MacDiarmid, A.G. Electrical conductivity in doped polyacetylene. Phys. Rev. Lett. **1997**, *59*, 1098.
3. Pekker, S.; Janossy, A. Chemistry of doping and distribution of dopants in polyacetylene. In *Handbook of Conducting Polymers*; Skotheim, T.A., Ed.; Marcel Dekker, Inc: New York, 1986; Vol. 1, 45–79.
4. Naarmann, H. Electrically conducting polymers. In *Industrial Polymers Handbook—Products, Processes, Applications*; Wilkes, E.S., Ed.; Wiley-VCH: New York, 2001; Vol. 3, 1209–1241.
5. Schon, J.H.; Dodabalapur, A.; Bao, Z.; Kloc, C.; Schenker, O.; Batlogg, B. Gate-induced superconductivity in a solution-processed organic polymer film. Nature **2001**, *410*, 189.
6. Heeger, A.J. Semiconducting and metallic polymers: the fourth generation of polymeric materials. Synth. Met. **2002**, *125*, 23–42.
7. Seymour R.B., Ed. *Conductive Polymers*; Plenum Press: New York, 1981.
8. Cotts, D.B.; Reyes, Z. *Electrically Conductive Organic Polymers for Advanced Applications*; Noyes Data Corporation: Park Ridge, NJ, 1986.
9. Skotheim T.A., Ed. *Handbook of Conducting Polymers*; Marcel Dekker, Inc.: New York, 1986; Vol. 1.
10. Skotheim, T.A., Ed. *Handbook of Conducting Polymers*; Marcel Dekker, Inc.: New York, 1986; Vol. 2.
11. Alcacer L., Ed. *Conducting polymers—special applications*. Proceedings of the Workshop, Sintra, Portugal, Jul 28–31, 1986; D. Reidel Publishing Company: Dordrecht, Holland, 1987.
12. Prasad, P.N., Ulrich, D.R., Eds. Nonlinear optical and electroactive polymers. Proceedings of an American Chemical Society Symposium on Electroactive Polymers, Denver, Colorado, Apr 6–10, 1987; Plenum Press, 1988.
13. Skotheim, T.A., Elsenbaumer, R.L., Reynolds, J.R., Eds. *Handbook of Conducting Polymers*; 2nd Ed.; Marcel Dekker, Inc.: New York, 1998.
14. Skotheim, T.A., Ed. *Electroresponsive Molecular and Polymeric Systems*; Marcel Dekker, Inc.: New York, 1988; Vol. 1.
15. Margolis J.M., Ed. *Conductive Polymers and Plastics*; Chapman and Hall: New York, 1989.
16. Aldissi, M. *Inherently Conducting Polymers—Processing, Fabrication, Applications, Limitations*; Noyes Data Corporation: Park Ridge, NJ, 1989.
17. Bredas, J.L., Silbey, R., Eds. *Conjugated Polymers—The Novel Science and Technology of Highly Conducting and Nonlinear Optically Active Materials*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1991.
18. Salaneck, W.R., Lundstrom, I., Ranby, B., Eds. Conjugated polymers and related materials—the interconnection of chemical and electronic structure. Proceedings of the Eighty-first Nobel Symposium; Oxford University Press: New York, 1993.

19. Aldissi, M., Ed. *Intrinsically Conducting Polymers: An Emerging Technology*, Proceedings of the NATO Advanced Research Workshop on Applications of Intrinsically Conducting Polymers, Oct 12–15, 1992; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1993.
20. Nalwa, H.S., Ed. *Handbook of Organic Conductive Molecules and Polymers: Conductive Polymers: Synthesis and Electrical Properties*; John Wiley & Sons Ltd.: New York, 1997; Vol. 2.
21. Chandrasekhar, P. *Conducting Polymers, Fundamentals and Applications: A Practical Approach*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1999.
22. Frommer, J.E.; Chance, R.R. Electrically conductive polymers. In *Encyclopedia of Polymer Science and Engineering*, 2nd Ed.; Kroschwitz, J.I., Ed.; John Wiley & Sons: New York, 1986; Vol. 5, 462–507.
23. Glatzhofer, D.T.; Deshpande, S. Conducting polymers (nonconjugated). In *Encyclopedia of Materials: Science and Technology*; Buschow, K.H.J., Cahn, R.W., Flemings, M.C., Ilschner, B., Kramer, E.J., Mahajan, S., Eds.; Elsevier: Amsterdam, 2001; Vol. 2, 1439–1446.
24. Wang, C.-S.; Dotrong, M. Conjugated ladder polymers. In *Encyclopedia of Materials: Science and Technology*; Buschow, K.H.J., Cahn, R.W., Flemings, M.C., Ilschner, B., Kramer, E.J., Mahajan, S., Eds.; Elsevier: Amsterdam, 2001; Vol. 2, 1461–1468.
25. De Paoli, M.-A.; Zoppi, R.A.; Felisberti, M.I. Conductive elastomeric blends. In *Encyclopedia of Materials: Science and Technology*; Buschow, K.H.J., Cahn, R.W., Flemings, M.C., Ilschner, B., Kramer, E.J., Mahajan, S., Eds.; Elsevier: Amsterdam, 2001; Vol. 2, 1455–1461.
26. Spinks, G.M.; Eisazadeh, H.; Wallace, G.G. Conducting polymer colloids. In *Encyclopedia of Materials: Science and Technology*; Buschow, K.H.J., Cahn, R.W., Flemings, M.C., Ilschner, B., Kramer, E.J., Mahajan, S., Eds.; Elsevier: Amsterdam, 2001; Vol. 2, 1421–1432.
27. Mattoso, L.H.C.; Oliveira, O.N., Jr.; Ferreira, M. Conducting polymers (for Langmuir–Blodgett film fabrication). In *Encyclopedia of Materials: Science and Technology*; Buschow, K.H.J., Cahn, R.W., Flemings, M.C., Ilschner, B., Kramer, E.J., Mahajan, S., Eds.; Elsevier: Amsterdam, 2001; Vol. 2, 1432–1439.
28. Krohnke, C. Electrically conducting composites. In *Encyclopedia of Materials: Science and Technology*; Buschow, K.H.I., Cahn, R.W., Flemings, M.C., Ilschner, B., Kramer, E.I., Mahajan, S., Eds.; Elsevier: Amsterdam, 2001; Vol. 3, 1953–1981.
29. Kanatzidis, M.G. Conductive polymers. Chem. Eng. News **1990**, *36*, 36–54.
30. Jeevananda, T.; Siddaramaiah; Seetharamu, S.; Saravanan, S.; D'Souza, L. Synthesis and characterization of poly(aniline-co-acrylonitrile) using organic benzoyl peroxide by inverted emulsion method. Synth. Met. **2004**, *140* (2–3), 247–260.

Contact Angles, Surface Tension, and Capillarity

C

Peter R. Pujadó
UOP LLC, Des Plaines, Illinois, U.S.A.

INTRODUCTION

Contact angles are an external manifestation of surface tensions or interfacial tensions and, as such, can be used to determine the value of such surface tensions.

Surface tensions arise from the unbalance of molecular attractive (and repulsive) forces that result at an interface from the different nature of the materials that come in contact. Surface tensions are more readily observed in liquids—between immiscible liquids or between a liquid and a gas. Surface tensions are also present on the surface of solids, but the rigidity of the solid structure prevents their manifestation in an observable way, except in combination with the surface of a liquid in contact with the solid surface. Thus, while the surface tension and the surface free energy in a liquid are the same, this is not the case on solid surfaces.

The balance of forces between surface tensions at the contact line results either in the Neumann triangle for a liquid/liquid/liquid or liquid/liquid/gas system or in the Young–Dupré equation on a liquid/liquid/solid or a liquid/gas/solid system (Fig. 1). While the Neumann triangle represents a true balance of forces, the Young–Dupré equation is little more than a definition of the $(\sigma_{AS} - \sigma_{BS})$ term, a difference between the respective solid/fluid surface free energies and not truly solid/fluid interfacial tensions.

BACKGROUND INFORMATION

As a rule, any system in equilibrium will tend to adopt a configuration that minimizes its energy state within the constraints or degrees of freedom that the system faces. When liquid interfaces are present, in addition to gravitational, electrostatic, and other forms of energy, we must include the energy of the surfaces. In the absence of other energy manifestations, liquids will adopt configurations that minimize their surface energy—in effect, their surface area. Thus, in the absence of gravity, a liquid droplet will be spherical because a sphere encompasses the smallest possible surface area around a fixed given volume. In the presence of gravity, a droplet sitting on a flat solid surface will adopt a ''sessile'' configuration; if suspended from the bottom of a flat surface, it will adopt a ''pendent'' configuration. Axisymmetric sessile drops are particularly useful in determining the contact angle at the contact line between the liquid and the solid. Similar measurements can be taken at the contact line between immiscible liquids, particularly at the contact line of an axisymmetric lens (a sessile lens) of an immiscible liquid floating on the surface of another liquid.

The equations that govern the shape of sessile and pendent configurations are very simple but nonlinear and cannot be solved analytically except for some very simple configurations or in limiting cases. Fortunately, numerical solutions are readily available such that, given the physical dimensions of a particular, say, sessile drop configuration, the interfacial forces can be readily calculated.

The effects of surface tension on sessile and pendent drops or lenses are but a simple manifestation of capillary hydrostatics. The field of capillarity can be far more extensive, principally when coupled with electromagnetic forces and also for liquid interfaces in motion, or in the motion in liquid interfaces that may result from local variations in surface tension as may be caused, for example, by local variations in temperature, or by the localized introduction of surfactants (interfacial tension modifiers), or by localized space-delimited chemical reactions. Wicking flows (as in ''heat pipes'') and flows in porous media (as in petroleum reservoir displacement) are a few of many other examples in which interfacial forces play a predominant role.[1]

Although often disregarded and not strikingly apparent in everyday life, capillary phenomena are everywhere and frequently can be found playing an important role in physical systems. Capillary phenomena occur wherever there is an interface between gas and liquid or between two liquids. The importance of capillarity depends on the ratio between the area of the region in which surface effects are dominant and the volume of the bulk phases.

In classical natural philosophy capillarity was regarded as a branch of hydrostatics and its study was focused on two fundamental areas: 1) nature of capillary forces and interpretation of their origin and 2) external manifestations of capillary forces. It is not known who first became aware of the existence of surface phenomena but the discovery of capillarity has been attributed to Leonardo da Vinci. Maxwell in his famous contribution to the ninth edition of

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007713

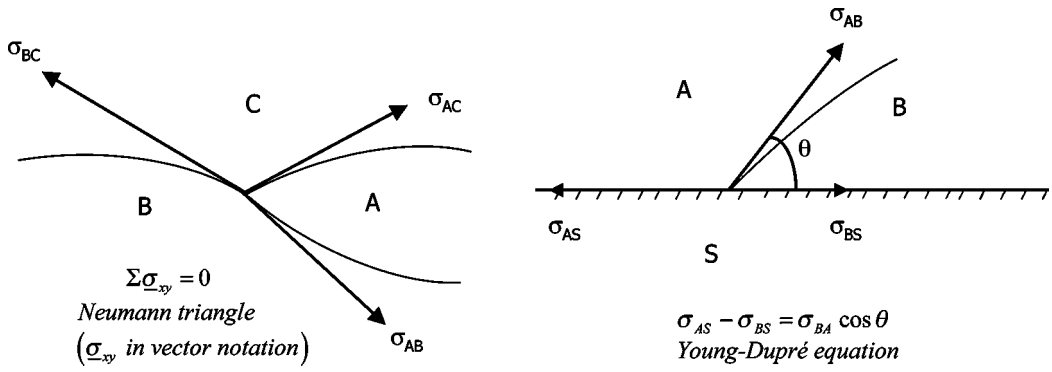


Fig. 1 Balance of forces between surface tensions at the contact line: Neumann triangle and Young–Dupré equation. *(View this art in color at www.dekker.com.)*

The Encyclopaedia Britannica outlined the history of the subject. Additions were made by Lord Rayleigh when he revised Maxwell's article for the 11th edition.[2] An annotated historical bibliography can be found in Young's *Lectures*.[3] Capillary forces as such were first studied by Laplace.[4] Gauss subsequently inaugurated an alternative mathematical approach.[5] The subject was later taken up by Lord Rayleigh, van der Waals, von Neumann, Bakker, and many others.[6–10] In recent years statistical mechanics has provided new insights into the theory of capillary forces.[11,12] First appearing in 1954, Lifshitz's theory of collective intermolecular forces in condensed media seems to represent a breakthrough that has led to a better understanding of the nature of capillary forces.[13–15]

THE EQUATION OF CAPILLARITY

The configuration of a capillary meniscus is governed by a balance of fluid pressures and surface tension and is correctly described by the Young–Laplace equation:[1,16,17]

$$\Delta p = 2H\sigma_{AB} \tag{1}$$

subject to the appropriate boundary conditions. In this equation, H is the local mean curvature of the meniscus while σ_{AB} represents the value of the surface tension between two fluids A and B, assumed to be uniform throughout each individual liquid/liquid or liquid/gas interface. If the principal radii of curvature at any given point of a surface are r_1 and r_2, $2H = 1/r_1 + 1/r_2$. At every ordinary point of a surface there is a direction for which the radius of normal curvature is a maximum and a direction for which it is a minimum, and they are at right angles to each other.[18] These maximum and minimum radii, r_1 and r_2, are the principal radii of curvature at that point. Except for a plane or a sphere, $r_1 \neq r_2$. In the case of a sphere with uniform radius, r, $2H = 2/r$.

More sophisticated analyses add higher-order terms to the Young–Laplace equation, but such refinements are neither apparent nor necessary because the basic equation accurately represents even the most complex systems.

An easily verifiable manifestation of capillarity and the contact angle can be observed by dipping a thin glass tube in a liquid such that the radius r of the tube is small enough for the shape of the liquid meniscus to be approximately spherical so that $2H = 2\cos\theta/r$, where θ is the contact angle between the liquid and the glass and $r/\cos\Delta z$ is the radius of curvature of the meniscus. If the liquid wets the glass, the liquid will climb a certain height, Δz, inside the tube.[19] We can then write the equation of capillarity as:

$$\frac{2\sigma\cos\theta}{r} = g\Delta z(\rho_B - \rho_A) \tag{2}$$

If the liquid wets the glass, θ is either zero or close to zero and not easily observable. Contact angles are more easily observable if the liquid does not wet the glass, $\theta > 90°$, so that instead of the liquid climbing inside the tube, the liquid is depressed to a lower level relative to the free liquid surface. This is the case, for example, if the glass tube is immersed in mercury or if the glass tube is coated with paraffin wax internally and is submerged in water.

Another way to see the contact angle is by depositing a droplet of liquid on a flat horizontal surface. By using appropriate combinations of liquids and solid surfaces we can obtain contact angles that may range from $\theta = 0°$ (totally wetting and spreading liquid) to $\theta = 180°$ (totally nonwetting liquid), as may be obtained, for example, by depositing a droplet of water on a smooth horizontal surface coated with a perfluoroethylene polymer (e.g., Teflon^{TM}).

Provided that the droplet is very small the effects of gravity can be neglected and the droplet may be assumed to be spherical with radius r. If so, the Young–Laplace equation becomes

$$\Delta p = 2H\sigma = 2\sigma/r \tag{3}$$

It follows from this simple relationship that the vapor pressure over a convex surface is greater than over a plane.[20] Various interesting consequences arise from this fact. One, for example, is that water vapor will not condense in a totally dust-free environment for lack of nucleating centers. Another is that when mercury is spilled and the drops break down into very small microdroplets, the increased vapor pressure greatly augments the risks of toxic exposure to mercury.

In the general form of the Young–Laplace equation, $\Delta p = p_A - p_B$ (A above and B below), and the local mean curvature is positive for a meniscus that is concave upward. The Young–Laplace equation is more often formulated in terms of the hydrostatic pressures developed in a uniform body force field, such as that of gravity near the earth's surface:

$$2(H - H^o)\sigma = g(\rho_B - \rho_A)(z - z^o) \tag{4}$$

where g is the local acceleration of gravity, ρ_B and ρ_A the densities of the fluids below and above the interface, respectively, and H^o the mean curvature of the interface at a datum elevation z^o. Owing to the way in which mean curvature depends on the first and second derivatives of elevation on the interface, this equation is strongly nonlinear. Solutions in closed form are available only for well-known cases of interfaces with translational symmetry, which have the shape of elasticas or lintearіae, and a few limiting cases of interfaces with rotational symmetry when $g \to 0$ or $\Delta p \to 0$, such as spheres, nodoids, cylinders, unduloids, and catenoids.[21] In general, the Young–Laplace equation has to be solved numerically and, even so, practically all the numerical solutions are limited to rotationally symmetric (axisymmetric) systems, except for a few ''three-dimensional'' interfaces studied by Petrov and Chernous'ko, Concus and Finn, Larkin, and others.[22–27] While, in principle, the nonaxisymmetric problem is not too difficult to tackle numerically, fitting the boundary conditions often becomes an impossible task.

The first solutions of the differential equation for rotationally symmetric menisci were computed geometrically, or graphically, by Thomas Young himself.[28] Young's method was employed by John Perry beginning in 1874, on instructions from Sir William Thomson, who later was to become Lord Kelvin and who wrote that between 1854 and 1859 he had worked out the method with what may be a slightly more general starting procedure than Young's.[29] The first numerical solutions of note were those for sessile and pendent drops developed by J. C. Adams, probably around 1855, and applied by Bashforth soon thereafter.[30] Subsequent developments have been reviewed by Bakker, from the van der Waals's school, and Padday, a British surface chemist and industrial scientist.[31–33] Padday's own computations are by far the most comprehensive that have appeared to date. A different and more efficient method was developed by Chun Huh and further expanded by Pujadó.[34,35] This method can be adapted to determine configurations of interfaces acted on by normal viscous stress as well as pressure. In fact, this is the boundary condition that should be used for the solution of the Navier–Stokes flow equation in a domain bounded by a free surface, the location and configuration of which is unknown a priori.

PARAMETRIZATION OF THE EQUATIONS

The explicit gravity-dependent Young–Laplace equation can be rewritten in dimensionless form as follows:[35]

$$2Ha = \frac{1}{x}\frac{d}{dx}\left[\frac{x\,y'}{\left(1 + (y')^2\right)^{1/2}}\right] = \varepsilon y \tag{5}$$

where $x = r/a$ [a is a capillary constant defined by $a^2 = \varepsilon\sigma/(g(\rho_B - \rho_A))$, with $\varepsilon = +1$ when $\rho_B - \rho_A > 0$, a sessile profile, and $\varepsilon = -1$ when $\rho_B - \rho_A < 0$, a pendent profile] and $y = (z - z^o)/a + y^o$, where y^o is the dimensionless ordinate of the datum point (usually the apical point) and is given by $y^o = \varepsilon 2H^o a$.

Then, by introducing $y' = dy/dx = \tan\varphi$, the above dimensionless equation can be decomposed into a system of first-order nonlinear differential equations:[6,24,34]

$$\begin{aligned}\frac{dx}{d\phi} &= \frac{x\cos\phi}{\varepsilon xy - \sin\phi}\\ \frac{dy}{d\phi} &= \frac{x\sin\phi}{\varepsilon xy - \sin\phi}\end{aligned} \tag{6}$$

which is a convenient form for numerical integration in terms of the independent variable φ and can be used with no difficulty for the entire range $0 \le \varphi < \infty$ in the sessile case ($\varepsilon = +1$) and up to the first inflexion point $0 \le \phi < \phi_{in}^1$ in pendent configurations ($\varepsilon = -1$).

To overcome the difficulty encountered at the inflexion points, we need a system of equations in terms of an independent variable that increases or decreases monotonically with arc length regardless of the profile of the meniscus. For this the natural choice is the dimensionless arc length, s, itself. It is defined as

$$ds = [(dx)^2 + (dy)^2]^{1/2} \tag{7}$$

The Young–Laplace equation can now be written as a system of three first-order nonlinear differential equations:

$$\begin{aligned} \frac{dx}{ds} &= \cos\phi \\ \frac{dy}{ds} &= \sin\phi \\ \frac{d\phi}{ds} &= \varepsilon y - \frac{\sin\phi}{x} \end{aligned} \tag{8}$$

Both sessile and pendent profiles form continuous families either of which can be generated by using these equations regardless of the value of ε. The choice of ε, given a convention for the changes of s and φ, affects only the sessile or pendent character of the starting point and its neighborhood. The initial value chosen for y for any given abscissa, x_0, sets the curvature at that point and fully determines the profile. The initial value chosen for s is irrelevant; the integration depends only on the step size, Δs, and its sign.

The actual configuration of a meniscus is given by that member of the family of solutions of the Young–Laplace equation that satisfies a given set of boundary conditions. By using this method we effectively change a tedious two-point boundary-value problem into a much simpler initial-value problem and the final solution is obtained by interpolating or iterating over a sequence of profiles. Larkin used a similar initial-value artifice to integrate the two-dimensional partial differential equation of Young and Laplace over a domain lacking rotational symmetry.[25] It should be noted though that the Young–Laplace equation is elliptic and its solution over a domain subject to a given set of boundary conditions has to be carried out by solving a system of algebraic equations defined over the mesh points of a finite-difference lattice.

BOUNDARY CONDITIONS AND INTEGRATION START-UP FOR POINTS OF ZERO SLOPE

Starting the integration of the Young–Laplace equation at a point of zero slope is exceedingly convenient whenever possible, the reason being that the solution in the vicinity of such a point behaves as a Bessel function or a modified Bessel function of zero order. Several types of profiles, real or hypothetical, can be considered. For the purposes of this article, only sessile drops and bubbles will be analyzed.

Sessile capillary menisci in the vicinity of a point with zero slope are described by the modified Bessel equation

$$y'' + (1/x)y' - y = 0 \tag{9}$$

which can be obtained by expanding $(1 + (y')^2)^n$, $n = 1/2$, $3/2$, etc. about $y' = 0$, and neglecting all terms $(y')^n$ with $n \geq 2$.

When the point of zero slope lies on the axis of rotational symmetry, as is the case with axisymmetric drops and bubbles, the solution for points in the immediate vicinity of the axis is

$$y(x) = \tan\phi^+ \frac{I_0(x)}{I_1(x^+)} \quad 0 \leq x \leq x^+ \tag{10}$$

where x^+ is an abscissa for which the above approximation still holds within a given tolerance. Normally, this is the abscissa corresponding to a φ^+ of up to 0.5°; values of φ^+equal to 0.1° or smaller can be chosen for safety.

This method automatically gives us the coordinates of the starting point of the curve at (x^+, y^+)

$$y^+ \equiv y(x^+) = \tan\phi^+ \frac{I_0(x^+)}{I_1(x^+)} \tag{11}$$

and the dimensionless apical curvature, i.e., the curvature at the point $x = 0$

$$y_0 \equiv y(0) = 2H^o a = \tan\phi^+ \frac{1}{I_1(x^+)} \tag{12}$$

The choice of a sign for φ^+ determines whether we obtain a sessile bowl ($\varphi^+ > 0$) or a sessile dome ($\varphi^+ < 0$) according to our convention for curvature signs.

Using the Bessel approximation as a start-up artifice always gives us a two-parameter family of solutions (or a one-parameter family for each initial abscissa x_0) as is always the case with a second-order differential equation. The parameter x^+ is directly related to the curvature y_0 or y^+ of a given profile. However, to singularize a profile that passes through a point (x^o, y^o) three parameters are necessary (x^o, y^o, φ^o), although solutions may not exist for some combination of parametric values (for example, if $x^o = 0$, no profiles with finite nonzero slope at x^o exist). In all cases, once x^+ and y^+ have been determined, we may proceed

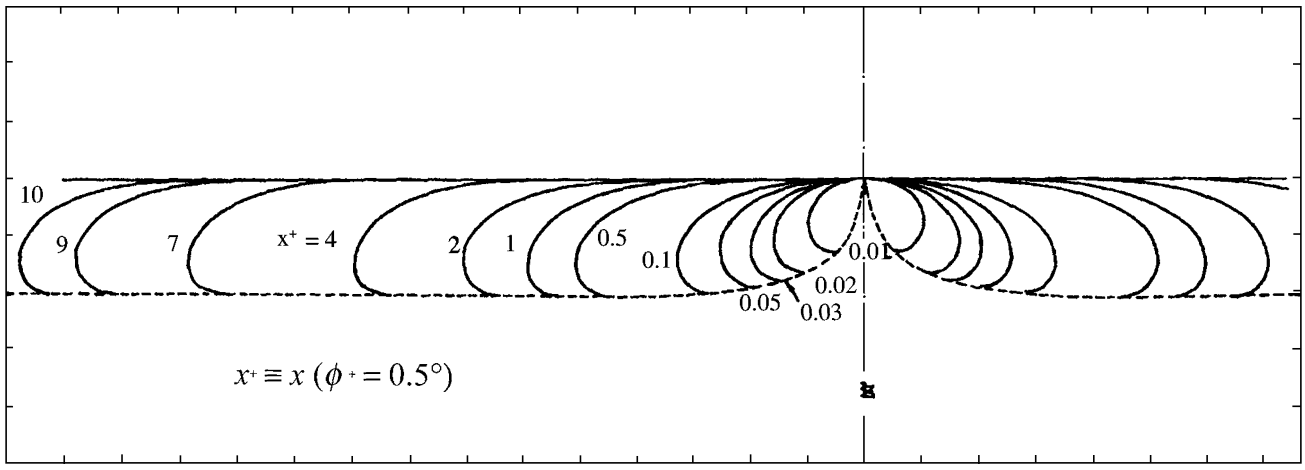


Fig. 2 Axisymmetric sessile drops.

with the integration of the equations by making use of some standard Runge–Kutta subroutine or an equivalent method.

REPRESENTATIVE ROTATIONALLY SYMMETRIC MENISCI

Figs. 2 and 3 show a family of bounded axisymmetric sessile configurations integrated all the way from $\varphi^+ = 0°$ (or $\varphi = \varphi^+$) to $\varphi = 180°$.

For any given angle, $\varphi > 0$, the thickness of the meniscus increases rapidly with increasing radius, passes through a maximum, and decreases slightly as it tends to an asymptotic limit as the radius tends to infinity. The limit is given by the two-dimensional solution of a translationally symmetric configuration—the separating elastica. If the integration is continued beyond $\varphi = \pi$, the profile is found to coil over itself indefinitely, as indicated in Fig. 4, where the first few loops are shown. If the integration is extended it will be noticed that the "centers" of the loops fall close to a straight line—the generatrix of a conical surface—which passes through the origin ($x = 0$, $y = 0$) (Fig. 5).

This is characteristic of all solutions of the Young–Laplace equation that have a nodoidal (i.e., coil-like) character and appears also in many other representative cases. The curves in the figures have been drawn for sessile drops with the denser liquid—shaded area—below. Sessile bubbles have identical configurations except that they are turned upside down, yet have the heavier fluid still on the lower side. All of these curves are parametrized with respect to their curvature at the origin, y_0, which is equivalent to parametrization by the value of their starting x^+.

The family of so-called "bounded" sessile configurations, that is, sessile drops and bubbles centered at the axis, belongs to the one-parameter subfamily

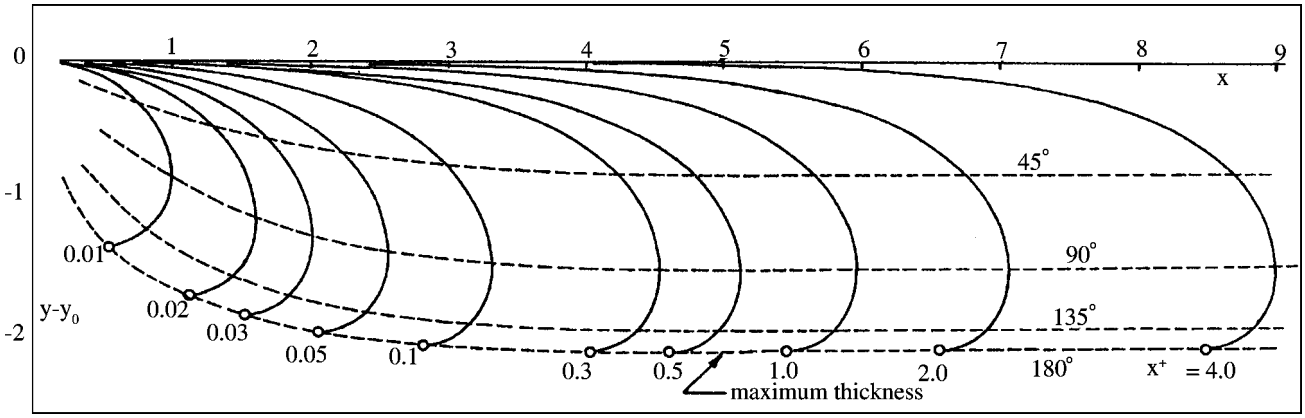


Fig. 3 Axisymmetric sessile drops. Locus of various contact angles: 45°, 90°, 135°, 180°.

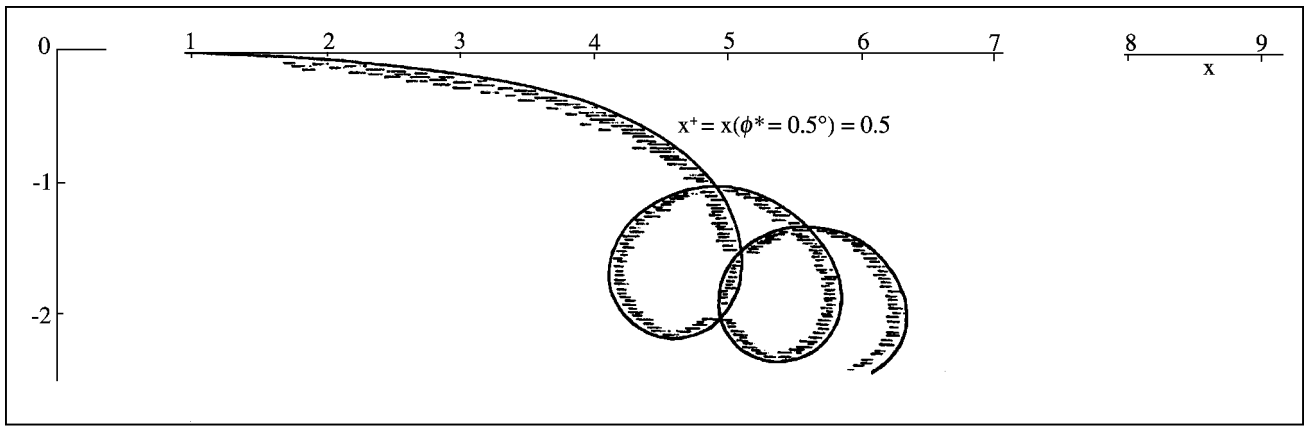


Fig. 4 Axisymmetric sessile profile integrated beyond 180°.

of solutions mentioned above. So does the family of bounded pendent configurations, that is, pendent drops and bubbles centered at the axis. Figs. 6 and 7 show equilibrium configurations of pendent drops with increasing curvature (decreasing x^+).

Similar calculations can be done for families of unbounded sessile profiles. An unbounded sessile profile can be visualized physically as the configuration of an axisymmetric dry patch, or that of a meniscus around a cylindrical rod with its axis perpendicular

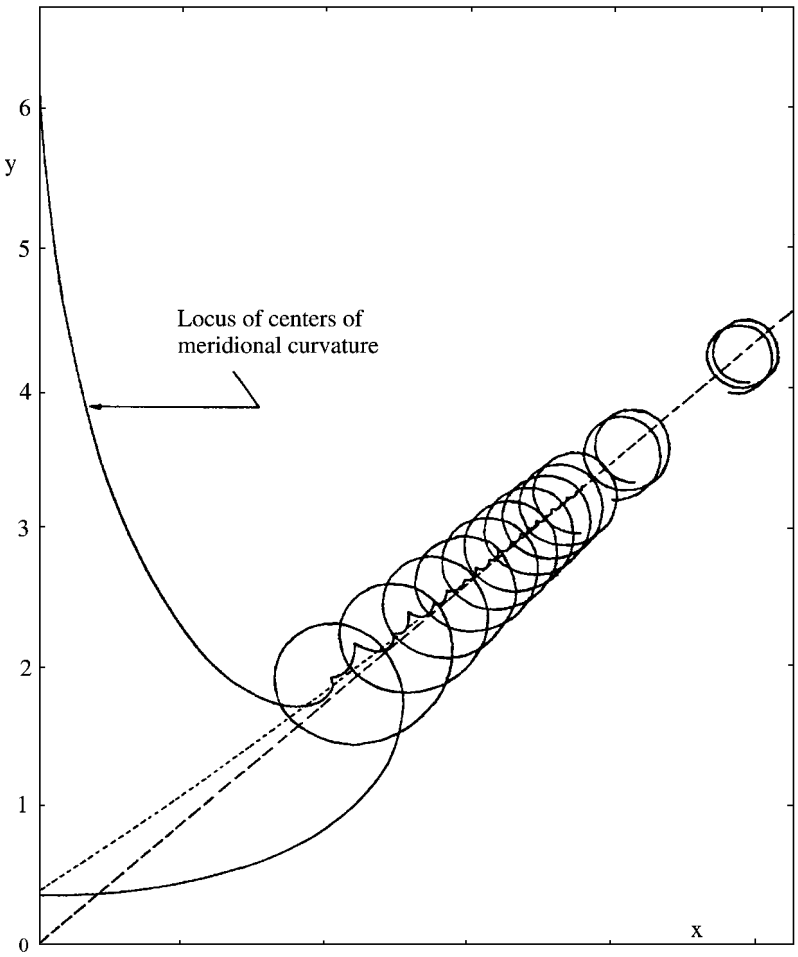


Fig. 5 Axisymmetric sessile profile with extended integration.

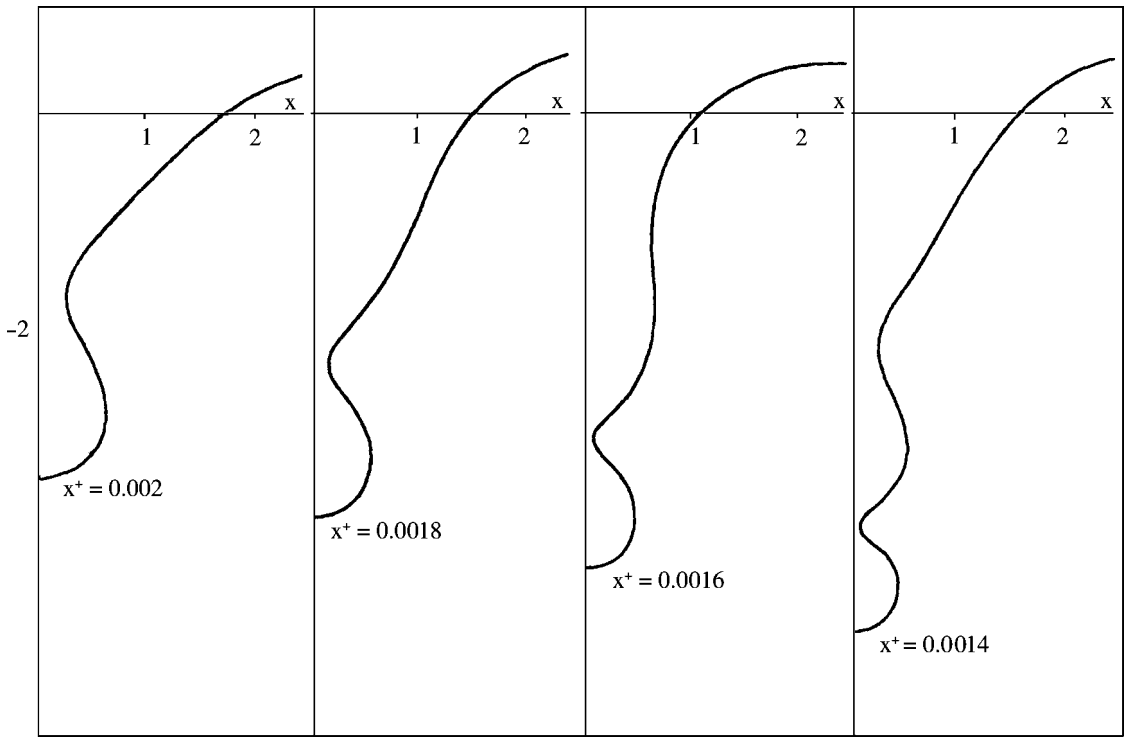


Fig. 6 Axisymmetric pendent drop assumed to be in static equilibrium (for $x^+ = 0.002,\ 0.0018,\ 0.0016$, and 0.0014).

to the free liquid interface, or the external meniscus around the ring of a de Noüy tensiometer. In this case, the boundary condition is

$$y'(x) \rightarrow 0 \quad \text{as } x \rightarrow \infty \tag{13}$$

and the solution of the Bessel approximation becomes:

$$y(x) = \tan \phi^+ \frac{K_0(x)}{K_1(x^+)} \quad x^+ \leq x \tag{14}$$

Likewise, the same treatment can be applied to the calculation of the shapes of pendent drops and bubbles. The behavior of pendent profiles about the apical point of zero slope, $y'(x_0) = 0$ at $x_0 = 0$, is given by the Bessel equation:

$$y'' + (1/x)y' + y = 0 \tag{15}$$

of which, the solution is

$$y(x) = \tan \phi^+ \frac{J_0(x)}{J_1(x^+)} \quad 0 \leq x \leq x^+ \tag{16}$$

where again the starting point for numerical integration is located at

$$y^+ \equiv y(x^+) = \tan \phi^+ \frac{J_0(x^+)}{J_1(x^+)} \tag{17}$$

such that the apical curvature is $y^o \equiv y(0) = a2H^o = \tan \varphi^+(1/J_1(x^+))$.

Not discussed here, but also equally feasible is the calculation of the equilibrium profiles of sessile and pendent lenticular configurations—for example, drops of an immiscible liquid floating on an interface (like oil on water) form a sessile lens as illustrated in Fig. 8.[36]

AXISYMMETRIC MENISCI AND THE DETERMINATION OF CONTACT ANGLES

It may be observed in Figs. 2 and 3 that the dimensionless profiles provide a unique correlation between the diameter of the meniscus (or the diameter of the meniscus at the contact line) and the elevation of the apical point with respect to the plane on which the meniscus rests. Therefore, once we have established the apical curvature of the meniscus that corresponds to that profile, it is a fairly straightforward task

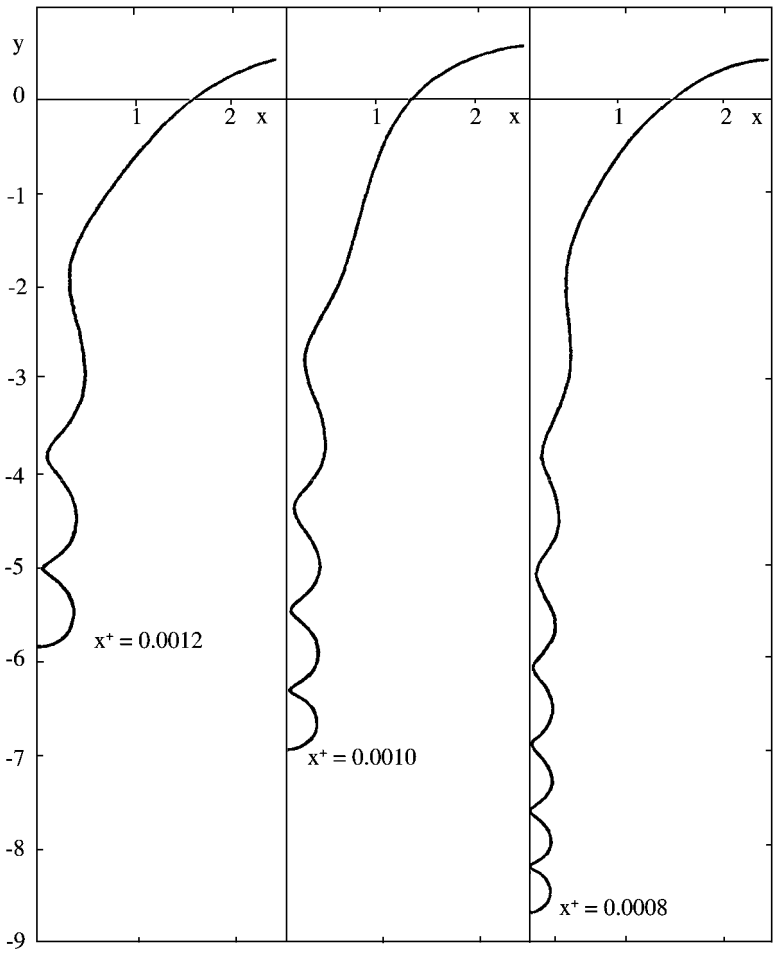


Fig. 7 Axisymmetric pendent drop assumed to be in static equilibrium (for $x^+ = 0.0012$, 0.0010, and 0.0008).

to determine the values of the capillary constant, a^2, and the surface tension, σ. The dimensions of the meniscus should be established by taking precise microscopic measurements.[37–44]

Alternatively, equivalent correlations can be derived for other menisci or capillary phenomena that will lead to similar determinations of the surface tension σ. These could depend on the relative elevations of the menisci inside thin capillary tubes or in between plates, or in more accurate measurements using ring tensiometers or other commercially available devices.

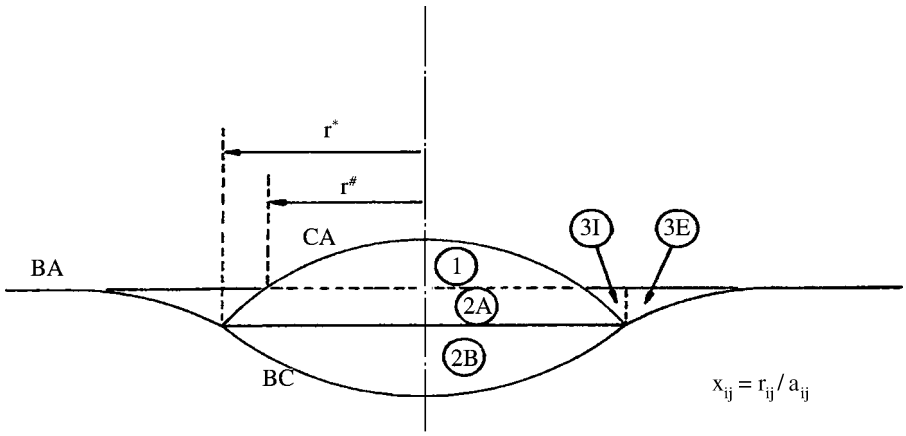


Fig. 8 Sessile lenticular configuration.

CONCLUSIONS

Capillarity phenomena are everyday occurrences that result from the existence of surface tension or interfacial tensions. In addition to the static phenomena discussed herein, surface tension and capillarity are also responsible for numerous dynamic phenomena that may result from localized gradients in temperatures or in compositions; the study of dynamic capillary phenomena (e.g., Marangoni flows, Bénard cells) is the subject of much literature coverage and is beyond the scope of this survey.

Static capillary phenomena lead to precisely determined geometrical shapes like sessile menisci, pendent menisci, minimal surfaces, which can be used for the physical determination and measurement of the surface tension or the interfacial tensions between fluids. In addition to the simple forms considered herein, more complex forms (e.g., sessile lenticular drops) can be studied.[35,36] Mathematical resolution of these shapes is a combination of the (numerical) solution of the highly nonlinear Young–Laplace equation together with an appropriate set of boundary conditions. For practical purposes, only axisymmetric forms are readily amenable to mathematical analysis.

REFERENCES

1. Udell, K.S. Heat transfer in porous media considering phase change and capillarity . . . the heat pipe effect. J. Heat Mass Transfer **1985**, *28* (2), 485–495.
2. Maxwell, J.C. Capillary attraction. In *Encyclopaedia Britannica*, 9th Ed.; 1878; Vol. 5; Rayleigh's revisions, 11th Ed.; 1911.
3. Young, T. A course of lectures on natural philosophy and the mechanical arts, 2 volumes; J. Johnson: London, 1807.
4. Laplace, P.S. (Marquis de). *Theory of Capillary Attraction;* Supplement to the 10th book of Celestial Mechanics, 1805; Bowditch, N., Transl., 1839.
5. Gauss, C.F. Principia generalia theoriae figurae fluidorum. Comment. Soc. Regiae Scient. Gottingensis Rec. **1830**, *7*.
6. Strutt, J. (Lord Rayleigh). On the theory of surface forces. Phil. Mag. **1890**, *30* (5), 285–298, 456–475.
7. Strutt, J. (Lord Rayleigh). on the theory of the capillay tube. Proc. R. Soc. **1915**, *A92*, 184–195.
8. van der Waals, J.D. Thermodynamische Theorie der Kapillarität unter Voraussetzung stetiger Dichteänderung. Ver. Kon. Ak. Wet. (Amsterdam) **1893**, *1* (8). Zitschr. Phys. Chem. **1894**, *13*, 657.
9. Neumann, F. Von. *Vorlesungen über die Theorie der Kapillarität*; B. G. Teubner: Leipzig, 1894.
10. Bakker, G. Théorie de la couche capillaire plane des corps purs. Scientia **1911**, *31*.
11. Kirkwood, J.G. *Collected Works—Theory of Liquids*; Gordon and Breach: London, 1968.
12. Buff, F.P. The theory of capillarity. In *Handbuch der Physik;* Springer Verlag, 1960; Vol. 10.
13. Dzyaloshinskii, I.E.; Lifshitz, E.M.; Pitaevskii, L.P. The general theory of van der Waals forces. Adv. Phys. **1961**, *10*, 165–209.
14. Ninham, B.W.; Parsegian, V.A. van der Waals forces—special characteristics in lipid-water systems and a general method of calculation based on the Lifshitz theory. Biophys. J. **1969**, *10*, 646–663.
15. Ninham, B.W.; Parsegian, V.A. van der Waals forces across triple-layer films. J. Chem. Phys. **1970**, *52*, 4578–4587.
16. Adam, N.K. *The Physics and Chemistry of Surfaces*; Oxford University Press, 1941; Dover Publications, 1968.
17. Adamson, A.W. *Physical Chemistry of Surfaces*; Wiley-Interscience, 1967.
18. Eisenhart, L.P. *A Treatise on the Differential Geometry of Curves and Surfaces*; Ginn and Co., 1909; Dover Publications, 1960.
19. Zhmud, B.V.; Tiberg, F.; Hallstensson, K. Dynamics of capillary rise. J. Colloid Interface Sci. **2000**, *228* (2), 263–269.
20. Wayner, P.C., Jr.; Wang, Y.X.; Zheng, L.; Plawsky, J.L. Optical evaluation of the effect of curvature and apparent contact angle in droplet condensate removal. J. Heat Transfer **2002**, *124* (4), 729–738.
21. Greenhill, A.G. *The Application of Elliptic Functions*; McMillan & Co., 1892; Dover Publications, 1959.
22. Petrov, V.M.; Chernous'ko, F.L. Determining the equilibrium form of a liquid subject to gravity forces and surface tension. Fluid Dyn. **1966**, *1* (5), 109–112.
23. Concus, P.; Finn, R. On the behavior of a capillary surface in a wedge. Proc. Natl. Acad. Sci. **1969**, *63* (2), 292–299.
24. Concus, P.; Finn, R. On a class of capillary surfaces. J. Analyse Math. **1970**, *23*, 65–70.
25. Larkin, B.K. Numerical solutions of the equation of capillarity. J. Colloid Interface Sci. **1967**, *23*, 305–312.
26. Kim, H.Y.; Lee, H.J.; Kang, B.H. Sliding of liquid drops down an inclined solid surface. J. Colloid Interface Sci. **2002**, *247* (2), 372–380.
27. Zhou, D.; Blunt, M.; Orr, F.M. Hydrocarbon drainage along corners of noncircular capillaries. J. Colloid Interface Sci. **1997**, *187* (1), 11–21.

28. Young, T. Essay on the cohesion of fluids. Philos. Trans. R. Soc. **1805**, *95*, 68–86.
29. Thomson, W. (Lord Kelvin). Capillary attraction. Proc. R. Inst. **1886**, *11* (3), 483–507; Nature **1886**, *34*, 270–272, 290–294, 366–369; Popular Lectures Addresses **1981**, *1*, 1–79.
30. Bashforth, F.; Adams, J.C. *An Attempt to Test the Theory of Capillary Action*; Cambridge University Press, 1883.
31. Bakker, G. Kapillarität und Oberflächenspannung. In *Handbuch der Experimental Physik*; Akademie Verlag: Leipzig, 1928; Vol. 6.
32. Padday, J.F. The profiles of axially symmetric menisci. Philos. Trans. R. Soc. **1971**, *A269*, 265–293.
33. Padday, J.F.; Pitt, A. Axisymmetric meniscus profiles. J. Colloid Interface Sci. **1972**, *38* (2), 323–334.
34. Huh, C.; Scriven, L.E. Shapes of axisymmetric fluid interfaces of unbounded extent. J. Colloid Interface Sci. **1969**, *30*, 323–337.
35. Pujadó, P.R. Capillary Hydromechanics; Dissertation, University of Minnesota, 1972.
36. Seeto, Y.; Puig, J.E.; Scriven, L.E.; Davis, H.T. Interfacial tensions in systems of three liquid phases. J. Colloid Interface Sci. **1983**, *96* (2), 360–372.
37. Jennings, W.J.; Pallas, N.R. An efficient (calculation) method for the determination of interfacial tensions from drop profiles (for pendent or sessile drops oriented up or down). 195th ACS National Meeting, 3rd Chemical Congress of North America, Toronto, Jun 5–10, 1988; Abstr. Coll. *37*, V1.
38. Briant, J.; Theibot, B. Measuring very low interfacial tensions and contact angles at high pressures and temperatures. Institut Français du Pétrole, World Surface—Active Agents Congress, Munich, May 6–10, 1984; Rev. Inst. Fr. Pet. **1985**, *40* (2), 241–250.
39. Allain, C.; Ausserre, D.; Rondelez, F. A new method for contact angle measurements of sessile drops. J. Colloid Interface Sci. **1985**, *107* (1), 5–13.
40. Rotenberg, Y. The determination of the shape of nonaxisymmetric drops and the calculation of surface tension, contact angle, surface area, and volume of axysymmetric drops. Diss. Abstr. Int. B **1984**, *44* (9).
41. Malcolm, J.D.; Elliott, C.D. Interfacial tension from the height and diameter from a single sessile drop or captive bubble. Can. J. Chem. Eng. **1980**, *58* (2), 151–153.
42. Carroll, B.J. The accurate measurement of contact angle, phase contact areas, drop volume, and Laplace excess pressure in drop-on-fiber systems. J. Colloid Interface Sci. **1976**, *57* (3), 488–495.
43. Ortiz-Arroyo, A.; Larachi, F.; Iliuta, I. Method for inferring contact angle and for correlating static liquid hold-up in packed beds. Chem. Eng. Sci. **2003**, *58* (13), 2835–2855.
44. Flock, D.L.; Le, T.H.; Gibeau, J.P. The effect of temperature on the interfacial tension of heavy crude oils using the pendent drop apparatus. J. Can. Pet. Technol. **1986**, *25* (2), 72–77.

Corrosion in the Process Industries

J. A. Richardson
Anticorrosion Consulting, Durham, U.K.

R. A. Cottis
School of Materials, Corrosion and Protection Centre, University of Manchester, Manchester, U.K.

INTRODUCTION

Corrosion makes a large impact on the economics of chemical processing. A recent U.K. government study estimated that corrosion expenditure amounts to ~4.5% of the turnover of the U.K. chemical/petrochemical industry. Much of this total expenditure represents investment in materials and protection practices to manage corrosion in new equipment. However, a significant proportion arises from the failure to identify and mitigate well-known corrosion risks at the design stage, and the U.K. study estimated that 15% of the total estimated cost may be saved by the application of existing corrosion prevention technology.

This entry provides an introduction to the management of corrosion in the process industries. It is necessarily rather general in character, and it should not be relied on for decision making; rather it provides a background for nonspecialists to help them to appreciate the decisions and advice of specialists.

MATERIALS FOR PROCESS EQUIPMENT

Process equipment has to operate over wide ranges of temperature, pressure, and fluid composition. Volatile hydrocarbons are stored at temperatures well below –100°C, and furnace tubes may be required to operate at temperatures above 1000°C. Crude oil distillation equipment operates commonly under vacuum, whereas supercritical processes operate at pressures of several hundred atmospheres. Aqueous solutions of mineral acids, alkalis, and salts can be extremely corrosive toward metallic materials, whereas plastic materials are much more vulnerable to organic solvents. The wide diversity of commercial chemical process conditions dictates that all classes of engineering materials find use in chemical process equipment.

The safe design and operation of pressurized equipment demands levels of strength and toughness that are generally achievable only in metallic materials. Consequently, the great majority of process equipment is of metallic construction. Economics dictates carbon and low alloy steels as preferred first option materials, but they lack useful strength above ~600°C, have relatively poor toughness below 0°C, and have poor corrosion resistance. Stainless steels and nickel alloys have strength and toughness capabilities across much wider temperature ranges, and wide ranging corrosion resistance, depending on the alloy. Copper, aluminum, and lead alloys have limited temperature capabilities, but find niche applications based on their specific corrosion resistances. Titanium, zirconium, and tantalum alloys are amongst the most corrosion-resistant metals available, though at a price, and are commonly used as lining materials on cheaper, steel substrates. Cobalt alloys are used principally for their wear resistances in the handling of solids, slurries, etc.

Plastic materials have limited strength and temperature capabilities, but provide cost-effective corrosion resistances toward many fluids that are highly corrosive to metallic materials. Thermoplastic materials such as polyethylene, polypropylene, polyvinyl chloride) and the fluoropolymers (PVDF, FEP, PTFE), and thermosetting polyester, epoxy and furane resins reinforced with glass or carbon fibers (FRPs), find uses in relatively small, low-pressure vessels and piping systems, and in valves and pumps. Elastomeric materials, ranging from natural rubbers to relatively expensive, synthetic fluoroelastomers, are widely used as joint sealing materials. All the polymeric materials find uses as corrosion-resistant lining materials for steels.

Ceramic materials have excellent temperature capabilities and good corrosion resistances, but are brittle, and thus unsuitable for pressure containment. In volume terms, their major use is as refractories for the thermal protection of steels. At lower temperatures, chemical stoneware; acid-resistant bricks; and specific oxide, carbide, and nitride ceramics have niche applications as corrosion-resistant linings and surfacings, in valves and pumps, and in seals and bearings. Silicon carbide and impervious graphite find use as heat exchange materials for particularly corrosive duties. Solid borosilicate glass is used for relatively small vessels and piping systems operating at low pressures, and

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007716

proprietary glass formulations are used as corrosion-resistant linings for steels.

CORROSION CHEMISTRY

Corrosion is fundamentally a chemical reaction between a metal and its environment.[a] As such it is a heterogeneous reaction between a fluid and a solid. At higher temperatures (when the environment is a gas rather than a liquid), the reaction is typically a direct reaction between oxygen and the metal to form the metal oxide. The oxide will form as a solid on the metal surface,[b] and oxidation will be controlled by the transport of oxygen and metal ions through the corrosion product.

In fluids that allow conduction of ionic charge, such as solutions of salts in water or molten salts, the corrosion reactions may split into two parts (which may occur in different places): the anodic reaction that results in the conversion of metal to metal cations and produces surplus electrons in the metal, and the cathodic reaction that consumes electrons and hence balances the anodic reaction:

Anodic reaction :

$$Fe \rightarrow Fe^{2+} + 2e^{-} \quad \text{iron corrosion}$$

Cathodic reaction :

$$2H^{+} + 2e^{-} \rightarrow H_2 \quad \text{hydrogen evolution}$$

Cathodic reaction :

$$O_2 + 2H_2O + 4e^{-} \rightarrow 4OH^{-} \quad \text{oxygen reduction}$$

Note that the two cathodic reactions are given—the first tends to be more important in acids and for reactive metals such as aluminum, while the second tends to be more important in neutral solutions and for less-reactive metals such as steel and copper alloys.

An important aspect of corrosion in liquids is the nature of the corrosion product. If this takes the form of a soluble salt, then the metal will remain in contact with the liquid, and rapid corrosion can be expected (this is known as active corrosion). If the corrosion product is solid (often an oxide or hydroxide), then it will tend to form a barrier between the metal and the solution, and this will tend to stifle the corrosion process (in this situation the metal is described as being passive, and the protective corrosion product is called a passive film).

CORROSION ENVIRONMENTS

There are an essentially infinite number of corrosion environments, but there are a number of key classes in the process industries:

Atmospheric: Atmospheric corrosion due to the combined effects of rain and the deposition of salt and other pollutants will affect most equipment. Corrosion occurs while the metal surface is wet, and is strongly influenced by the composition of deposits (such as sulfates from industrial atmospheres and chlorides from marine atmospheres). External corrosion of steel and stainless steel process equipment beneath thermal insulation and fireproofing is of particular concern.

Gases: While most corrosion is electrochemical, corrosion in noncondensing gases is chemical, because there is no possibility for the transport of ionic charge in the environment, or of dissolution of the corrosion product. This leads to a rather different character for corrosion in gases, which will consist of the growth of scales of corrosion product. This requires the transport of oxygen and metal ions through the scale, and this generally requires a high temperature to occur at a significant rate.

Waters: Water plays an important part in corrosion at lower temperatures. It acts as a solvent for salts that modify the corrosion process; it ionizes these salts, and thereby allows the passage of ionic charge; it may dissolve the corrosion products; it can dissolve a moderate amount of oxygen; and it can act as an oxidant itself with the evolution of hydrogen. The content of soluble salts, and hence the ionic conductivity, is important. Fresh water (e.g., unpolluted river water and tap water) has a low salt content, and in consequence, it is relatively noncorrosive toward carbon steel. This is especially true at moderate temperatures when oxygen is excluded from the water; many heating and cooling systems (including most domestic central heating systems) rely on deaeration of tap water (often by reaction with the steel in a closed system) to control corrosion. Similarly, the major water treatment for lower pressure boiler waters is concerned with the maintenance of a low salt content and the removal of oxygen. Even in the presence of oxygen, clean river water gives an acceptable corrosion rate of carbon steel (many steel-hulled boats for use only in freshwater are used without any corrosion protection). As the salt content increases, the water becomes more corrosive, especially in the presence of oxygen, and seawater is very corrosive to many metals.

[a]The term ''corrosion'' may also be applied to nonmetals, and these are touched on in this article, but the chemistry in this case will be different, and the wide variety of nonmetallic materials precludes detailed coverage here.

[b]At higher temperatures and in the presence of salt deposits, the corrosion product may be liquid, in which case the behavior may be more like that of aqueous corrosion. At very high temperatures, the oxide may be volatile, in which case very rapid corrosion can occur.

Acids: Acids tend to produce high corrosion rates, both because protective oxides become soluble in acids and the reduction of hydrogen ions becomes an effective cathodic reaction. There is a major distinction between oxidizing acids (notably nitric acid), which provide an alternative cathodic (oxidation) reaction and can corrode even noble metals such as copper and silver, and so-called reducing acids (such as hydrochloric and sulfuric, though the term ''nonoxidizing acids'' might be more accurate), which require an alternative reactant (such as dissolved oxygen) to corrode the noble metals, but which can corrode steels and more reactive alloys with hydrogen evolution. Note that oxidizing acids can cause passivation of iron and steels; so iron will corrode rapidly in concentrated hydrochloric acid, but it will passivate in concentrated nitric acid.

Alkalis: Most metals are protected by a passive oxide in mildly alkaline solutions, but the protective oxide will redissolve in strong alkali to form oxy-anions of the metal, allowing corrosion to occur. For carbon steels, the region of corrosion in alkali is very limited, but it can lead to the serious problem of caustic stress corrosion cracking (SCC).

Salts: In general, the nature of the anion present in dissolved salts is important for corrosion, largely as a result of the interaction of the anion with protective passive films. An important property of the anion present in salts is the pKa of the corresponding acid, as many localized corrosion processes involve the production of local acidic conditions at anodic regions; anions of weak acids will act as buffers and limit the drop in pH. Halides tend to be particularly aggressive, both because they have a low pKa and support the formation of strong acids, and because metal halides tend to be soluble. In contrast, carbonates and phosphates have a high pKa, and therefore restrict any drop in pH and tend to inhibit corrosion. The cation present in the salt can also affect corrosion, most commonly because of the precipitation of hydroxides at local cathodic regions (which tend to become more alkaline than the bulk solution). Thus, calcium and zinc cations tend to act as cathodic inhibitors for iron and steels in neutral solutions.

TYPES OF CORROSION

Corrosion can produce a range of morphologies, some of which can be particularly damaging for a given amount of metal loss. Thus uniform corrosion, in which the metal loss is distributed over the entire metal surface, usually causes relatively slow, manageable loss in section thickness of components, whereas some localized forms of corrosion can lead to failure of components in a matter of months, weeks, or even days.

Uniform or General Corrosion

Uniform corrosion usually occurs in fairly aggressive environments that attack the whole surface. Examples include carbon steel in seawater or acids, or aluminum alloys in strong alkali. The rate of metal loss is usually rather high, but, because it is distributed over the whole surface, the performance can usually be predicted, and managed with corrosion allowances, in most situations. Thus, sheet steel piling is often used in seawater without any corrosion protection, the corrosion rate of around 0.1 mm/yr,[c] coupled with the relatively thick steel sections, giving an acceptable life.

Galvanic Corrosion

When two dissimilar metals are electrically connected together, such that electrons can flow from one metal to the other, it is probable that the anodic, metal dissolution reaction will concentrate on one metal, while the cathodic reaction occurs on both. This accelerates the corrosion of the anodic component. The acceleration will be particularly severe if the area of the cathodic metal is much larger than that of the anodic metal (Fig. 1). While the anodic metal will corrode more, the cathodic metal will normally corrode less, and one form of galvanic corrosion provides a method of corrosion control (know as cathodic protection), in which the anodic, corroding metal is provided deliberately in order to restrict the corrosion of the cathodic metal.

Dealloying or Selective Attack

Dealloying occurs when one component of an alloy is lost preferentially. Thus, brass is an alloy of zinc (a rather active metal) and copper (a rather noble metal). Consequently, the zinc tends to be lost in preference to the copper. Often the copper will form a ''seal'' over the surface, preventing further corrosion, but if conditions do not allow this, then the corrosion can penetrate into the component, removing most of the zinc. The result is a porous copper component, which has little mechanical strength, and the problem is often discovered when the component fractures. Similarly, one component of a two-phase alloy can

[c]The corrosion rate quoted is a typical average value. The rate varies according to depth, and a number of other factors, with the region around high tide normally having the highest corrosion rate. Recently, the phenomenon of accelerated low water corrosion has been identified, with corrosion rates of 1 mm/yr or more, and corrosion protection should be considered for all new structures.

Fig. 1 Galvanic corrosion of a steel blank flange around a threaded brass coupling from a liquid fertilizer storage tank. *(View this art in color at www.dekker.com.)*

corrode preferentially, when the process is better called selective attack (Fig. 2).

Differential Aeration Corrosion

When there is a difference in oxygen content between two parts of a component, the part with the higher concentration will tend to become more alkaline (and may passivate), while the part with the lower concentration will tend to become more acidic (and will corrode more readily). The acceleration in corrosion rate due to differential aeration is relatively limited, because there must be a relatively large anodic area compared to the cathodic area in order to produce enough alkalinity to passivate the steel.

Crevice Corrosion and Pitting Corrosion

These forms of corrosion are similar to differential aeration corrosion, in that an oxygen-free region becomes acidic by virtue of the net anodic reaction and consequently corrodes rapidly when coupled to a region in aerated solution. However, a key difference is that these forms of corrosion occur on alloys that are initially passive;[d] so there is no limit to the area of the passive, cathodic region. Thus, the severity of the attack may be much greater (Fig. 3). Crevice corrosion occurs when there is a narrow gap between two pieces of metal or a piece of metal and an insulator. The oxygen is consumed by the slow passive corrosion in the crevice, causing the crevice to become sufficiently acidic that the passivity breaks down and active corrosion starts. This then produces stronger acidification and reinforces the process. Pitting corrosion occurs on a free surface, often initiated by a local defect in the passive oxide, such as a surface sulfide inclusion (though in service, it is often initiated by a local crevice under a particle of dirt or similar). Once the pit has initiated, it is stabilized in essentially the same way as crevice corrosion. While they also occur in other passive alloys, crevice corrosion and pitting corrosion are particular problems with stainless steels, particularly in chloride-containing solutions. The main approach to control of these problems is in alloy composition, where increasing chromium, molybdenum, and nitrogen contents improve the resistance to pitting and crevice corrosion.

Flow Effects

Solution flow typically enhances corrosion rates, by increasing the transport of dissolved oxygen to the metal surface, by increasing the rate of removal of protective corrosion products, and, in extreme cases, by physically removing the corrosion products or even metal (in the case of erosion by suspended particles or cavitation) (Fig. 4). In a few situations flow can be beneficial; thus for stainless steels in chloride solutions, flow can prevent the development of the acidification that is necessary for pitting and crevice corrosion.

Stress Corrosion Cracking

The term SCC can be defined as ''the initiation and propagation of cracks in a metallic component as a result of the combined influence of a mechanical stress and a specific corrosive environment''. The stresses involved may be externally applied working stresses or internal residual stresses (often produced by deformation

[d]It can be argued that any form of corrosion that occurs in a crevice is crevice corrosion, and any form of corrosion that leads to small local penetration is pitting corrosion.

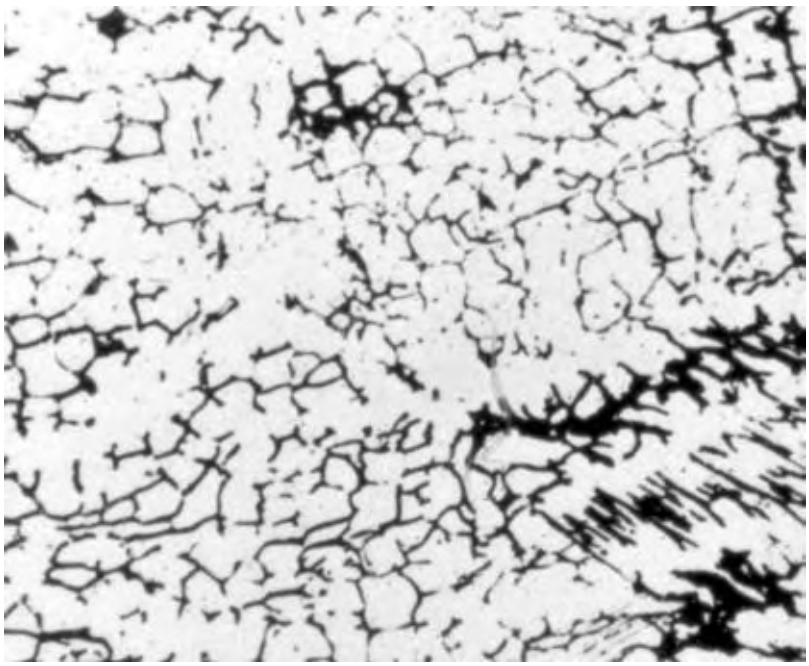


Fig. 2 Selective attack of the black ferrite phase in a stainless steel weld bead. *(View this art in color at www.dekker.com.)*

or welding during fabrication). Most alloys will suffer from SCC in some environments, usually ones that give a relatively low rate of general corrosion. Crack growth rates vary over a wide range, with typical times to failure ranging from hours to years (Fig. 5).

Corrosion Fatigue

Metal fatigue is the process of crack initiation and growth due to the action of a fluctuating mechanical stress. Corrosion fatigue is simply metal fatigue that

Fig. 3 Crevice and pitting corrosion of a stainless steel autoclave head. Note the crevice corrosion underneath the bolts (now removed) and in the gap between the two parts that are still assembled, and the pitting corrosion on the free surface. This corrosion was probably caused by chloride derived from thermal insulation. *(View this art in color at www.dekker.com.)*

Fig. 4 Flow-induced corrosion of a pump impeller. *(View this art in color at www.dekker.com.)*

is accelerated by the action of a corrosive environment (Fig. 6). The typical effect of corrosion is to reduce the time to failure compared to testing in air, especially at lower stresses and cyclic load frequencies, and to reduce or eliminate the fatigue limit.

Hydrogen Effects

Hydrogen embrittlement

Hydrogen will dissolve easily in metals, and, once in solution, it can cause brittle fracture. In general, body-centered cubic metals, such as ferritic steels, are most susceptible to hydrogen embrittlement, while face-centered cubic metals, such as austenitic stainless steels, are much less susceptible. Additionally, the susceptibility increases as the strength of the material increases (Fig. 7).

Hydrogen-induced cracking (HIC) and stress-oriented HIC

These processes involve crack formation by the precipitation of dissolved hydrogen onto lamellar nonmetallic inclusions. They are a particular problem

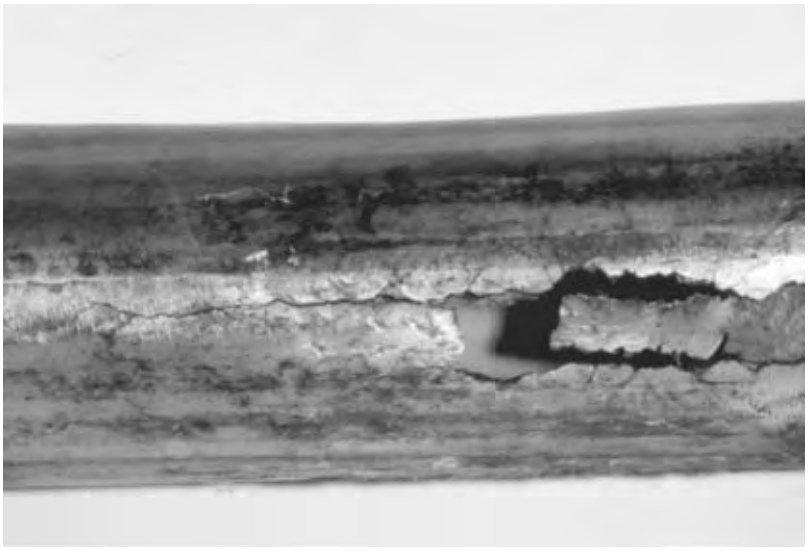


Fig. 5 External SCC of a stainless steel pipe after contact with wet, chloride-containing, thermal insulation.

Fig. 6 Environmentally assisted cracking of a weld in a deaerator, caused by a combination of static residual and cyclic operating stresses, and revealed by magnetic particle inspection.

in equipment handling 'sour' hydrocarbon fluids that contain hydrogen sulfide (Fig. 8).

Hydrogen attack

This is a high temperature process relevant to equipment in petroleum refineries and petrochemical plants containing hydrogen. It is associated with the formation of methane by the reaction of hydrogen and carbon inside the steel. This results in decarburization, and the methane precipitates at the nonmetallic inclusions to produce cracking in a similar way to HIC.

Fretting Corrosion

Fretting corrosion occurs when two metal surfaces are rubbed together, usually in a relatively noncorrosive environment, such as moist air. The rubbing removes the protective oxide, allowing further oxidation to occur, and the oxide produced acts as an abrasive to accelerate this process.

Microbial Corrosion

Corrosion of steels and stainless steels may be induced by the metabolic products of microbial growth and

Fig. 7 Cracks in a high strength steel relief valve spring. The cracks have been revealed using a dye that penetrates into the crack then seeps back out into a porous white coating. The cracks were caused by hydrogen embrittlement of the excessively hard steel, coupled with tensile residual stresses and the storage environment on a chemical plant (with an occasional trace of H_2S). *(View this art in color at www.dekker.com.)*

Fig. 8 Hydrogen-induced cracking—note the internal split produced by the precipitation of hydrogen gas onto flattened sulfide inclusions.

reproduction. The classic case is the corrosion of steel by sulfate reducing bacteria under anaerobic conditions in soils, or under deposits and tubercles in waters. Stainless steels are also vulnerable to pitting in waters containing iron and iron manganese bacteria, in the event that hydrotest or cleaning waters are retained in equipment for long periods prior to recommissioning.

MANAGEMENT OF CORROSION IN NEW EQUIPMENT

Management of Corrosion by Design

The corrosion of an alloy in an environment is influenced by chemistry, temperature, stress, geometry, and galvanic effects, which are exacerbated in process equipment by heat transfer and fluid flow. All these factors can be varied by design, and it follows that the propensity of process equipment to corrosion can be influenced strongly by design detail.

Chemistry definition

The commonest cause of unpredicted corrosion problems is the failure to define, accurately, the chemistry of process streams, including startup, shutdown, and transient conditions, or to anticipate changes in chemistry at specific locations in equipment. The corrosivity of a process stream is often determined by its minor components, e.g., the presence of 10s–100s ppm chlorides can promote localized corrosion of stainless steels and other passive alloys. It is important that minor components are defined, quantified, and evaluated at the design stage, including their possible local concentration such as in distillation and separation equipment.

Heat transfer is a particular promoter of the development of local chemistries that are very different from the bulk fluid chemistry, in particular, where phase changes occur. Early condensates from acid gas/vapor streams can be very concentrated and corrosive relative to bulk condensates as in the cases of carbonic acid from steam, sulfuric acid from flue gases, and hydrochloric acid from refinery overhead streams. Initially, benign condensates or cooling fluids can concentrate due to intermittent contact with surfaces hot enough to promote concentration or dryout, as in the chloride-induced, external SCC of austenitic stainless steels under thermal insulation. Extremely high concentration factors are possible under boiling heat transfer in crevices or under deposits, or under film boiling conditions at high heat flux regions, as in water side, on-load corrosion in process and utility boilers. There are practical design measures, which mitigate the risks of developing corrosive microenvironments in heat transfer equipment, e.g., preferring tubeside water and horizontal orientation, or venting of top tubesheets in vertical, stainless steel water coolers, and the attenuation of heat fluxes in highly rated boilers by tube inlet ferrules.

Temperature definition

Although normally straightforward for bulk process streams, it is important to select materials for skin rather than bulk temperatures in heat transfer equipment, and to evaluate the effects of mixing exotherms on local bulk fluid temperatures, such as that may occur in acid addition/dilution equipment.

Control of stress and stress concentration

Design codes for process equipment specify maximum allowable stresses for materials, the basis for which varies, depending on the type of equipment and the class of material. Corrosion allowances are specified commonly in design to ensure that pressurized components remain thick enough throughout their service lives to maintain membrane stresses below the design maximum allowable levels.

In practice, process equipment commonly contains regions with substantially higher stresses. The codes sometimes allow higher stresses in local regions of the equipment, design stresses can be concentrated by defects in materials, and fabrication processes such as cold working and welding introduce ''residual'' stresses of the order of the yield stress.

Environmentally assisted cracking (EAC) problems such as SCC and corrosion fatigue commonly initiate at such high stress regions. Clearly, locally high stresses permitted by the design code are ''given'', and if they present a risk of EAC to a specific material, then an alternative must be specified. However, stress concentration effects arising from defects can be mitigated in design

by specifying detail such as transitional radii at changes in cross section, machining/grinding of weld toes, quality of surface finish, etc. Residual stresses can be reduced in design by specifying processes such as thermal stress relief, temper bead welding, peening, etc.

Control of geometry

The major effects of geometry relate to vulnerability of process equipment to crevice corrosion and erosion–corrosion.

Crevice corrosion in process equipment is associated most commonly with flanged joints with gaskets, heat exchanger joints, and weld defects. Screwed and socket welding flanges present crevices to the fluid, whereas slip-on welding and welding-neck flanges avoid crevices. Care is needed in the specification and sizing of gaskets to avoid crevices. The ubiquitous tube/tubeplate joint in tubular heat exchangers is inevitably at potential risk of crevice corrosion, particularly where high heat transfer rates into the crevice can superheat the crevice relative to the bulk fluid.

Consideration might be given in design to locating fluids that might promote crevice corrosion on the tubeside of the exchanger, and/or to procedures for closing off crevices such as tubeplate back face bore welding. Plate and other compact heat exchangers, with very large areas of compression or welded joints, can be particularly vulnerable to crevice corrosion, and consideration might be given to upgrading the material to mitigate risks. Crevice corrosion can occur at welds where lack of sidewall fusion or root penetration creates a crevice, and appropriate fabrication and inspection procedures must be implemented to prevent such defects.

Erosion–corrosion is avoided in design through the use of limiting velocities for fluids that have been established by testing and/or experience. Beyond that, it is important to design and fabricate fluid systems to minimize flow disturbance, by using where appropriate long radius bends/elbows and gradual changes in cross section, ensuring that flanges are aligned correctly, avoiding excessive weld root protrusion in fabrication, etc. It is also important to select, size, and locate components that disturb flows, such as probes, valves, and orifice plates, to minimize local and downstream turbulence. Thus, it is a good design practice to locate control valves and orifice plates away from bends, etc.

Control of galvanic effects

Inevitably, process systems are constructed in a variety of materials, and the potential for galvanic interaction needs to be addressed in the design. Although the risk of galvanic corrosion is confined mostly to mixed metal systems, carbon/graphite (widely used in heat exchange and jointing systems) and silicon carbide (mostly a heat exchanger material) are sufficiently electrically conducting for their potential contributions to galvanic corrosion to require evaluation.

Available galvanic series of materials have been obtained in specific environments, mostly seawater, and are of limited use in galvanic corrosion risks in chemical process fluids that may have very different conductivities, tendencies to promote passivity on specific materials, etc. However, as a general rule, it is a good design practice to avoid material combinations that are widely separated in such series, and in particular, to avoid unfavorable area ratios involving small areas of ''active,'' anodic metal coupled to large areas of ''noble,'' cathodic material, e.g., more active fasteners and weld filler metals. Galvanic effects may be mitigated through coatings. Preferably, both couple members are coated, but if only one can be coated, it should be the noble, cathodic member. Equipment items judged to be vulnerable can, in principle, be isolated electrically using insulated flange joints, but in practice this may be frustrated by remote earthing via steel supports, pipe hangers, etc. Galvanic effects at material breaks in piping systems can be mitigated by physical separation of couple members using insulated spool pieces made from FRP, or coated, more noble couple member. Galvanic effects can be overridden by applying cathodic protection to the couple, as in the protection of the water boxes of seawater coolers by sacrificial anodes.

Management of Corrosion by Materials Selection

Materials selection for corrosion resistance is as reliable as the information upon which it is based. Corrosion information derives from two sources—experience or test results.

Experience is sometimes codified in the form of industry guides that address generic, industry-wide corrosion problems. More commonly, experience takes the form of corporate experience of materials performance in specific applications, obtained from the operation and inspection of existing equipment. This information may be accessible across corporate boundaries through informal industry networks, depending on its commercial significance, and electronic networks, such as CORROS-L (see www.jiscmail.ac.uk) and the NACE Corrosion Network (see www.nace.org), which are increasingly important.

In the absence of relevant experience, materials selection decisions have to be based on the results of corrosion tests, which range from relatively simple,

laboratory procedures, through semitechnical/pilot plant tests, to operating plant tests and trials. Corrosion test results are widely available in the public domain, usually as tabulated corrosion rate data for specific environments, from materials suppliers and professional corrosion organisations. Alternatively, new corrosion tests can be undertaken to inform materials selection decisions for applications where no relevant experience or test data is available.

Corrosion tests have inevitable limitations in their capacities to mimic actual service conditions of equipment. Standard, ambient pressure, immersion test procedures, with intermittent fluid refreshment, are available for both metallic and nonmetallic materials, but are limited to the ambient pressure boiling point of the fluid, and provide limited scope to simulate the effects of stress, geometry, heat transfer, and fluid flow. Such test procedures can be conducted at plant pressures and temperatures in autoclaves, and can be upgraded to focus on specific factors such as fluid flow and heat transfer. Even so a laboratory test, however elaborate, is a poor substitute for a test in the plant itself.

The key advantage of plant testing is direct exposure to the service fluid, thereby avoiding the risks of missing undefined, significant fluid constituents, or time-dependent chemistry changes, in unrefreshed laboratory liquors. Test coupons can be installed in plants in racks, or bolted directly onto trays, baffles, flanges, slip rings, etc, or onto retractable coupon holders. Care is needed to locate test coupons in the right places, particularly where phase and/or chemistry changes occur such as in separation and distillation equipment. Test plant components such as spool pieces, impellers, heat exchanger tubes, agitator paddles, column internals, etc., provide perfect simulation of all potential contributory factors to corrosion, as long as the attendant risks of premature failure can be accommodated.

It is not uncommon for new processes to encounter significant corrosion problems during the early stages of commercial operation, reflecting the poor reliability of the information used to select materials. Public domain corrosion tables are useful guides to corrosion performance but, not least because of the common absence of test details, are relatively unreliable information sources. Laboratory and/or plant coupon testing are more reliable, but the most reliable information derives from field performance data from operating plant inspections and/or long-term plant component tests. It is also important that materials selection is an integral part of process design from the earliest, chemistry definition stages through process design iterations, and that any laboratory, semitechnical, pilot plant, and sidestream activities are used to evaluate candidate materials as an integral part of process development.

Management of Corrosion by Inhibition

Although its scientific basis is beyond the scope of this contribution, corrosion control by the addition of corrosion inhibiting chemicals is important in several areas of chemical processing. The major application is in the management of corrosion, principally of steel, in cooling water and steam raising systems. Well-established practices are available commercially, and the great majority of corrosion problems in process industry water systems arise from the failure to specify at the design stage, and/or manage in service, appropriate water treatment practices. Inhibition is practised in several other areas of chemical processing, including the control of overheads corrosion of steel and other alloy heat exchangers by condensing hydrochloric acid in oil refining, and the control of wet carbon dioxide corrosion of steel vessels and pipelines in oil/gas recovery and acid gas stripping.

Management of Corrosion by Coatings and Linings

A wide variety of coatings and lings are available for the protection of steel in chemical process equipment, and they often provide cost-effective alternatives to more expensive, corrosion-resistant alloys.

Chemical plant atmospheres can promote particularly rapid atmospheric corrosion. For structural steelwork, galvanized zinc and/or paint coatings are usually specified. As alternatives to galvanizing, a variety of zinc- or zinc phosphate-containing primers are specified routinely, overcoated with compatible paint systems, including micaceous iron oxide epoxy, acrylic polyurethanes, silicones, etc. Paint coatings are also used for the prevention of atmospheric corrosion of process equipment, although care is needed in the use of metallic zinc, because of concerns about polarity reversal on warm/hot surfaces. A particular concern is the control of external corrosion beneath thermal insulation and fireproofing, for which relatively high-quality paint systems are required for the protection of steel, while thin aluminum foil is an alternative to paint coatings for reducing the risk of external SCC of stainless steel surfaces.

High performance paint coatings can be used for the internal corrosion protection of steel process equipment such as storage tanks, pipelines, vessels, and even heat exchangers, although a detailed account of the options is beyond the scope of this contribution. In environments where the corrosion rate of steel would be $>\sim 0.5$ mm/yr if exposed, thicker linings are preferred to relatively thin paint coatings. A wide range of thermoplastic, FRP, rubber, and glass linings are available, depending on the application.

Corrosion-resistant metals and alloys, including stainless steels, nickel alloys, titanium, zirconium, and tantalum can be applied as linings or claddings to cheaper steel substrates. Most (>90%) are applied by roll bonding, but weld overlaying and explosive bonding are also used, as appropriate. Particularly expensive metals, such as tantalum, can be used as very thin (~0.5 mm) loose linings.

Management of Corrosion by Electrochemical Intervention

Wet corrosion of metals being an electrochemical process, it can be controlled, in principle, by electrochemical intervention, and commercial practices are available for specific applications. The most widely used technique is the cathodic protection of steel in soils and waters, which involves lowering the potential of steel to levels at which the rate of corrosion is negligible using distributed sacrificial anodes of zinc, magnesium or aluminum alloys, or inert impressed current anodes driven by an external power supply. Cathodic protection is used commonly in association with coatings, to confine current demand to defect sites in an otherwise protective coating. It is used in the process industries for the external protection of buried/immersed pipelines and infrastructure, and occasionally the internal control of corrosion, including galvanic corrosion, in tubular water coolers.

Anodic protection involves the promotion of passivity on steel or stainless steel equipment by raising and controlling the potential into the passive range in the process fluid, using auxiliary cathodes, or coupling to more noble materials, or by the addition of oxidizing species to the fluid (a special case of inhibition). Anodic protection is used mostly to extend the useful range of stainless steels in reducing sulfuric and phosphoric acids, where it finds practical usage in heat exchangers and storage tanks. It is used, not merely to control corrosion at acceptable rates, but also to prevent unacceptable contamination of process fluids with corrosion products.

MANAGEMENT OF CORROSION IN OPERATING EQUIPMENT

The corrosion performance of equipment in service is determined largely by decisions taken at the design stage. However, design decisions often anticipate appropriate management and maintenance processes in service, and if these are not delivered, then problems may ensue. For example, if the plant is operated, knowingly or otherwise, outside the design envelope, the risk of corrosion will increase, and it can only be mitigated by appropriate review of the original materials of construction under the revised process conditions, and upgrading if required. If corrosion is to be controlled by corrosion inhibition, or electrochemical intervention, then corrosion risks in service will be determined by the extent to which these processes are managed and maintained according to the design intent.

In modern practice, corrosion risks in operating chemical process equipment are managed within the broader context of loss-of-containment risk mitigation, based mainly on risk-based inspection (RBI). The principles of RBI are beyond the scope of this contribution, but are based essentially on risk mitigation by optimizing inspection frequency and coverage, using appropriate tools, techniques, procedures, and practices. The results of these activities define whether the design intent is being realized in practice, or whether further risk mitigation is required in the forms of equipment replacement or repair, or equipment modification and redesign (increased corrosion allowances, changes in materials, addition of coatings/linings, addition of corrosion inhibitors, etc). They also define the need for additional monitoring of corrosion rates through appropriate inspection and/or monitoring techniques.

The inspection of process equipment to detect and size corrosion damage is beyond the scope of this contribution, and the following summary is confined to techniques that can be used for on-line corrosion monitoring. Techniques such as radiography and ultrasonics can be used externally to track corrosion distribution, wall thickness, and defect size through periodic measurements, but there are limitations relating to sensitivity and conditions of measurement. Periodic or continuous fluid analysis for the presence/concentrations of corrosion products can provide useful information, but the locations and distributions of the corrosion from which the products arise can only be inferred. Otherwise, the available techniques for corrosion monitoring fall into three broad categories.

Firstly, there are techniques that provide information on accumulated metal loss, from which average corrosion rates can be obtained by periodic measurements. The simplest approach involves the use of coupons, which can be retrieved at intervals from critical locations on the plant to provide average corrosion rates from weight or dimensional change. Alternatively, probes with elements comprising strips, wires, or thin-walled tubes of the test material can be inserted at critical locations, and the changes in their dimensions monitored by changes in their electrical resistance, which can be measured readily and intermittently online, the latest commercial technology being sensitive to nanometer changes in dimension. In a third technique, hydrogen that has been generated during

corrosion of the internal surface permeates the wall of a pipe or vessel and is collected in a chamber on the external surface, where the change of pressure can be related to metal loss on the internal surface.

A second category of techniques based on electrochemistry, provide instantaneous measurements of corrosion rate/state. Linear polarisation resistance and electrochemical noise measurements require the introduction of probes with varying numbers of elements of test material. Measurements are made across individual pairs of elements, involving the detection of small changes in potential or current, which are generated spontaneously, or in response to the application of small potential or current perturbations. Corrosion rates, and information on the propensity to localized corrosion, can be generated from the measurements, the principles behind which are beyond the scope of this contribution. The corrosion potentials of plant components can be measured online, and although they contain no information as to corrosion rate, they can provide useful information as to whether components are passive or not, or close to potential ranges where there may be a vulnerability to localized corrosion. Finally, in an electrochemical variation to the technique referred to earlier, hydrogen that permeates through a pipe or vessel wall can be oxidized electrochemically in a cell strapped onto the pipe exterior, thereby providing an indirect measurement of corrosion rate on the interior surface.

The final two techniques allow remote, direct monitoring of metal loss from plant components. Thin layer activation (TLA) uses isotopes created in a surface layer of the component by proton irradiation. Corrosion is monitored through periodic measurements of the loss in activity as the surface corrodes, which can be made outside the pressure envelope. The electric field signature method (FSM) measures changes in the electric field pattern in a metallic structure induced by corrosion, from voltage measurements amongst pairs of an array of pins distributed on the external surface. TLA and FSM are sensitive to around 1% of layer depth and wall thickness, respectively. They are relatively expensive techniques, but find use in the diagnosis and control of critical and expensive corrosion problems on operating plant.

CONCLUSIONS

Corrosion will always remain a significant challenge to the process industries. Process technologies will continue to develop into areas where there is no relevant experience of materials performance. There are inevitable risks in the use of corrosion testing to predict materials performance. With the exception of a relatively few generic problems in the nuclear and oil/gas processing fields, corrosion models do not allow reliable prediction of materials performance. However, for the majority of process equipment, corrosion prevention is a relatively mature technology, based on major progress in the understanding of the mechanisms and control of corrosion phenomena over the last 40–50 yr. Much of this progress has been made through industry initiatives, and the accumulated knowledge and experience can be sourced in the activities and publications of bodies such as the National Association of Corrosion Engineers (NACE, http://www.nace.org), American Petroleum Institute (API, http://www.api.org), and the Materials Technology Institute (MTI, http://www.mti-link.org) in the USA, and the European Federation of Corrosion (EFC, http://www.efcweb.org) and DECHEMA (http://www.dechema.de) in Europe. Of particular relevance are the NACE Special Technical Groups (STGs) on Water Treatment (STG11), Petroleum Refining and Gas Processing (STG34), Process Industry (STG36–39) and Energy Generation (STG41), and the EFC Working Parties (WPs) on Corrosion and Scale Inhibition (WP1), Corrosion by Hot Gases and Corrosion Products (WP3), Nuclear Corrosion (WP4) and Corrosion in the Refinery Industry (WP15). Most process/chemical chemical engineers receive little or no training in corrosion and its management in their graduate education and early formative years. When corrosion issues arise in the design or the operation of process equipment, it is advisable to consult corrosion specialists on all but very routine matters, and the earlier the involvement of the specialist, the better.

BIBLIOGRAPHY

Agree, M. *Betz Handbook of Industrial Water Conditioning*, 9th Ed.; Betz Laboratories Inc.: Trevose, 1991.

Behrens, D. *DECHEMA Corrosion Handbook*, 12 Volumes, 4th Ed.; John Wiley and Sons, 1991. (Also available on CD-ROM: Elsevier Science; 2001 onwards.).

Craig, D.B.; Anderson, D.S. *Handbook of Corrosion Data*, 2nd Ed.; ASM International: Ohio, 1995.

Dillon, C.P. *Corrosion Control in the Chemical Process Industries*, 2nd Ed.; Materials Technology Institute: St Louis, 1994.

Francis, R. *Galvanic Corrosion: A Practical Guide for Engineers*; NACE International: Houston, 2000.

Gaverick, L. *Corrosion in the Petrochemical Industry*; ASM International: Ohio, 1994.

Gravener, D.L. *Corrosion Data Survey—Metals Edition*, 6th Ed.; NACE International: Houston, 1985.

Guides to Corrosion Control, Dept. of Industry: London. Available in print or online from the U.K. National Corrosion Service. (http://www.npl.co.uk/ncs/).

Hamner, N.E. *Corrosion Data Survey—Nonmetals Edition*, 5th Ed.; NACE International: Houston, 1975.

Holman, R. Corrosion and Associated Costs in the U.K. Chemicals and Petrochemicals Sector; Paint Research Association: London, 2002 (http://www.pra.org.uk/projects/costofcorrosion.htm).

Kemmer, F.N. *The Nalco Water Handbook*, 2nd Ed.; McGraw Hill Inc.: New York, 1989.

Lai, G.Y. *High Temperature Corrosion of Engineering Alloys*; ASM International: Ohio, 1990.

Moniz, B.J.; Pollack, W.I. *Process Industries Corrosion—The Theory and Practice*; NACE International: Houston, 1986.

Revie, R.W. *Uhlig's Corrosion Handbook*, 2nd Ed.; John Wiley and Sons Inc.: New York, 2000.

Schweitzer, P.A. *Corrosion Resistance Tables*, 5th Ed.; Marcel Dekker Inc.: New York, 2004.

White, R.A. *Materials Selection for Petroleum Refineries and Gathering Systems*; NACE International: Houston, 1998.

Critical Phase Behavior

J. Richard Elliott, Jr.
Department of Chemical Engineering, University of Akron, Akron, Ohio, U.S.A.

INTRODUCTION

The critical pressure is defined as the pressure above which no vapor can exist. The critical temperature is defined as the temperature above which no liquid can exist. The supercritical fluid (SCF) region is, therefore, the region above both the critical temperature and the critical pressure. The subject of critical phase behavior can be perplexing to the uninitiated and painfully obvious to the fully indoctrinated. This entry is directed primarily at the former, especially those who have an appreciation for general thermodynamics and chemical processing but lack exposure to the nuances of the critical region. We also include a broad survey of patents and technology that impact chemical processing and relate to critical phase behavior.

BACKGROUND

For pure fluids, the critical region is complicated by behavior that cannot be easily correlated. The most obvious instance is the curvature of the temperature–density coexistence curve. Classical equations tend to overestimate the critical point when fit to data outside the critical region. In other words, the true curvature is flatter than the curvature of a classical equation. Mathematically, $(\rho^L - \rho^V) \sim (T_c - T)^{0.325}$ instead of the classical curvature given by $(\rho^L - \rho^V) \sim (T_c - T)^{0.5}$. Remedies for this problem generally take the form of either scaling crossover equations[1] or multiparameter equations.[2] Extensions to mixtures further complicate the issue. Critical scaling tends to be a specialized subject with no single approach being recognized as a conventional standard. Fortunately, the qualitative behavior from classical equations is generally correct once allowance is made for deviations very near the critical point. Because most applications of critical phase behavior are related to mixtures, the focus is on mixtures in this entry.

The critical region becomes very interesting when one considers the behavior of mixtures. At its most basic level, the critical point is the point at which the vapor and the liquid become indistinguishable at fixed overall composition. The path to the critical point follows an indirect route when the composition is rich in a supercritical component like CO_2 and leans in a heavy component like hexadecane. Above the critical temperature, one may easily find a dew point at low pressure, and an increasing fraction of liquid as one raises the pressure from that point. But increasing the pressure at fixed temperature may result in the fraction of liquid decreasing before one reachs the bubble point. How can this be possible? Raising the pressure at fixed temperature makes the vapor phase denser, making it a better ''solvent.'' The vapor phase then dissolves more of the heavy component. Ultimately, one reachs a point where all but one drop is vapor.

With this preparation, the reader can appreciate the key elements of a subject that can be challenging. The critical region is a space where phases are barely distinguishable; hence subtle changes make large differences. The focus in this entry is on critical phase behavior as it pertains to extraction and separation. We omit coverage of supercritical reactions because the phase behavior is qualitatively similar with or without reaction. Reactions merely alter the compositions. The impact on reactions of high compressibility in the critical region is of special interest, however, and initial study can be facilitated through the compilation of Noyori.[3]

BASICS OF CRITICAL PHASE BEHAVIOR

Fig. 1 provides the most logical starting point for any introduction to critical phase behavior. The solid curve represents the distinct boundary between the vapor and the liquid phases along the vapor pressure curve. The dashed lines, on the other hand, represent regions of the phase diagram that have been defined by convention for purposes of general reference. Physical properties like density, viscosity, and diffusivity vary continuously across these imaginary boundaries, but are generally similar within the designated regions. We focus especially on the properties near the critical point and in the SCF region because the properties in that region are sensitive to variables within the engineer's control, not just temperature and pressure, but also composition. This sensitivity can provide engineers with the ability to tune the phase behavior. If one includes liquid–liquid and micellar critical behavior within the scope of this subject, advanced topics like nanoscale morphology can be contemplated as well as more conventional topics like extraction.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007719

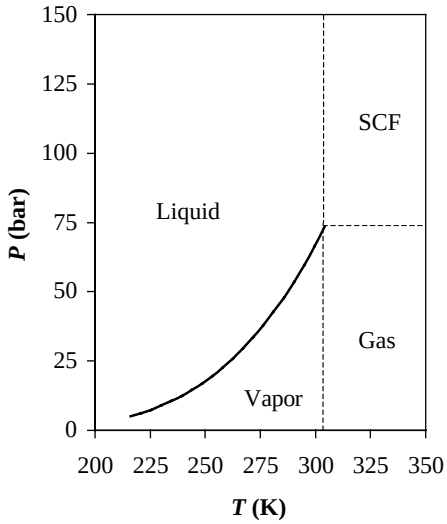


Fig. 1 Phase diagram for pure CO_2.

The key is to recognize the engineering possibilities when the thermodynamics are indefinite. Thermodynamics can be difficult and critical phase thermodynamics excruciating, but opportunities abound and much can be accomplished with a simple, general understanding of what is occurring.

Phase Diagrams and Classifications of Phase Behavior

Despite the many opportunities in the critical region, there are also constraints. Some of the most striking are the discontinuities that occur in the critical loci of mixtures. At first glance, one might assume that the critical points of mixtures should vary continuously from one pure component to the other as the composition varies. That does happen occasionally, but rarely in applications of interest.

In mixtures like CO_2 and polystyrene, the CO_2 phase can hold only a limited amount of polymer before polymer-rich phase begins to precipitate. One might suggest increasing the pressure to increase the density and the carrying power of the CO_2. But then a second liquid phase is more likely to form than a truly SCF phase. Or, at temperatures around 50–80°C, vapor–solid equilibrium may predominate. At the polymer-rich end of the critical locus, the critical temperature is extremely high. A small amount of CO_2 can be dissolved under these conditions, but the upper critical locus diverges toward very high pressures instead of approaching the critical point of CO_2 as the temperature is reduced. Such problems with polymers may seem relatively obvious in retrospect, but similar problems occur with lower molecular weight components like vegetable oil. Hydrogenation of vegetable oil is a common application, but one should consider the difference in critical temperatures between hydrogen (33 K) and vegetable oil (>600 K).

Clearly, we must carefully consider the constraints that arise when we combine components in the critical region. Fortunately, van Konynenburg and Scott[4] have analyzed these constraints in a very general manner and organized the analysis into a classification scheme like Aristotle's genus and species. Their approach was remarkably simple but effective. They studied the van der Waals equation of state for mixtures with a range of choices for the parameters a and b appearing in the equation. Qualitatively, these parameters are capable of describing the critical phase behavior of nearly all binary mixtures. Illustrations of the phase diagram types are given in Fig. 2.[5] Note that type VI behavior is not mentioned by van Konynenburg and Scott. However, type VI behavior can be exhibited by the van der Waals equation.[6] This behavior appears to be exclusive to aqueous systems.[7] Other type variations have been suggested over the years,[8,9] but the current consensus holds that these six types are sufficient.

Types II and III are the most commonly encountered in SCF applications. Classes I and V represent mixtures for which $H^E < 0$, where H^E is the excess heat of mixing. Generally, $H^E < 0$ when solvation interactions occur in the absence of association, as in the acetone + chloroform system. Strongly solvating mixtures like these are relatively uncommon.

These diagrams illustrate the trends of the critical loci as the composition changes. Critical loci roughly correspond to the upper limit of temperature and pressure for a phase envelope at fixed composition. Fig. 3 illustrates in three dimensions the compositions leading to each critical point. Above the critical locus, we would expect to see only a single phase. The slashes along the critical locus indicate the side on which the phase envelopes open up into two phases. In types III–V, the critical locus is interrupted by a vapor–liquid critical end point (VLCEP). At the VLCEP, the liquid phase splits into two liquid phases. An upper critical end point (UCEP) is similar, but the vapor merges with one of the liquid phases. In types IV and V, the lower critical end point (LCEP) also plays a significant role. The LCEP is a condition at which a liquid phase merges with a vapor phase. The resulting liquid–liquid (LL) condition gradually evolves along the critical locus into a vapor–liquid (VL) condition as the two tangent points on the Gibbs energy curve shift toward the VL critical point of pure component 2.

The work of van Konynenburg and Scott stands out as a classic throughout all of thermodynamics. A basic understanding can be achieved by noting how phase behavior depends on the strength of the molecular

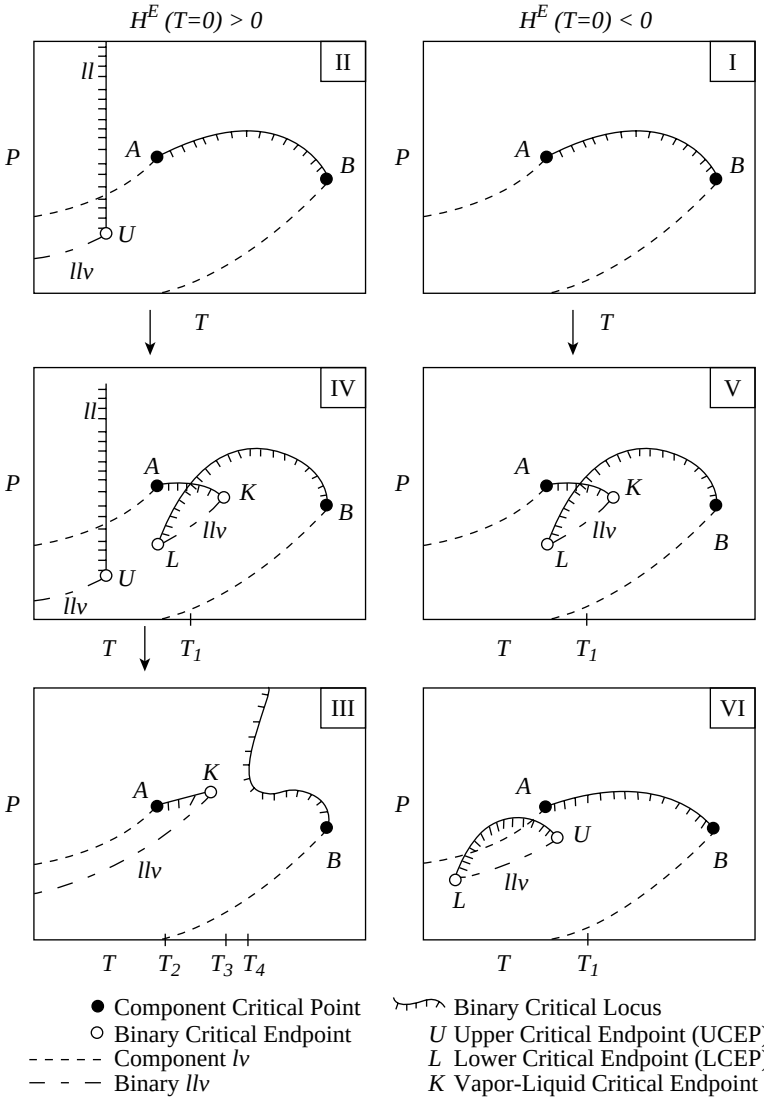


Fig. 2 Progression of binary phase behavior with increasing molecular asymmetry according to van Konynenburg and Scott. Arrows denote progressions of phase behavior expected by theory. Experimental progressions frequently differ. (From Ref.[77].)

interactions. Considering the quadratic mixing rule: $a = x_1^2 a_{11} + 2x_1 x_2 a_{12} + x_2^2 a_{22}$, with $a_{12} = (1 - k_{ij}) \times (a_{11} a_{22})^{1/2}$ we may observe that molecular attractions are weaker as k_{ij} becomes more positive. Strong molecular attractions are indicated by negative values for k_{ij}. van Konynenberg and Scott generalized their results by defining two dimensionless quantities describing the inherent difference between the two pure components, $\zeta \equiv (a_{11}/b_{11}^2 - a_{22}/b_{22}^2)/(a_{11}/b_{11}^2 + a_{22}/b_{22}^2)$, and the strength of their binary interaction by $\Lambda \equiv (a_{11}/b_{11}^2 - 2(a_{12}/b_1 b_2) + a_{22}/b_{22}^2)/(a_{11}/b_{11}^2 + a_{22}/b_{22}^2)$. By convention, the compound with the lower value of T_c is taken as component 1. Hence, a large value of ζ indicates that the components are very different in volatility. Note that $\Lambda = 0$ when $a_{12} = (a_{11} + a_{22})/2$, i.e., the arithmetic mean. A large positive value of Λ generally indicates weak binary attractions, but a better characterization of binary interaction is indicated by the curve for $k_{ij} = 0$. In the absence of any other guideline, the best approach is to follow the $k_{ij} = 0$ line when assessing the critical phase behavior for any given mixture.

The transition from I to II to III occurs as the binary interactions become weaker. This is especially clear when the pure components are similar in attractive interactions ($\zeta \sim 0$–0.4). Basically, the weakening binary attractions lead to less stable solutions. At high values for ζ, the differences between the pure components are so large that slight weakness in the binary attraction destabilizes the liquid phase. Indeed, types III–V all exhibit two liquid phases. This is surprising for the type V case, because the binary attractions are quite strong.

It is tempting to ascribe the trend in ζ to a difference in sizes of the molecules, because the larger molecules tend to have larger values for a. On the other hand, Fig. 4 assumes that the molecules have equal size; hence, there must be some other explanation. To clarify,

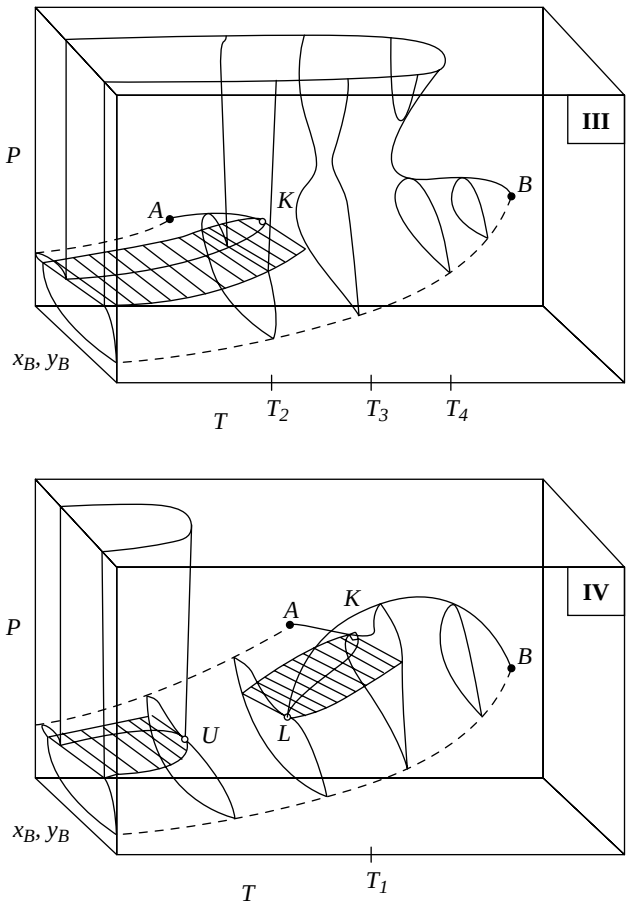


Fig. 3 Type III and IV phase behavior illustrated in three-dimensional diagrams. Symbols and labels are the same as Fig. 2. (From Ref.[77].)

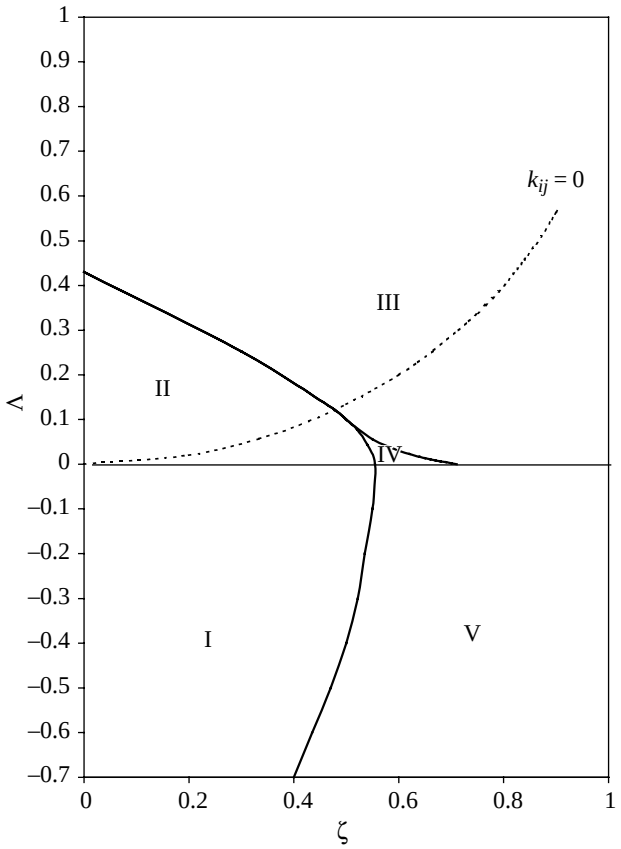


Fig. 4 van der Waals master diagram for equal-sized molecules. (From Ref.[4].)

Elliott and Lira[5] show that $a/b^2 \sim \delta^2$, where δ is the solubility parameter. For large molecules, the solubility parameter varies little with respect to molecular weight. Thus, increases in ζ correspond primarily to increases in cohesive energy density, not molecular size. Taking pentane as an example of component 1, we can estimate ζ from solubility parameters to be 0.16, 0.27, 0.33, 0.40, 0.52, 0.62, 0.84 for cyclohexane, benzene, acetone, n-hexanol, ethanol, methanol, and water. This provides an idea of the range of chemical functionalities addressed in Fig. 4.

The van der Waals equation was an excellent choice for van Konynenburg and Scott, but it does have its limitations. By choosing the van der Waals equation, the fundamental model was greatly simplified. Simplifying as much as possible is highly advisable given such a complex problem to begin with. Unfortunately, this choice results in a certain degree of ambiguity when it comes to classifying the phase behavior for a given mixture of interest. Because the van der Waals equation gives only qualitative accuracy, reliable values of the binary interaction parameters are rarely known for mixtures of interest. With these limitations, the accepted practice is to classify the phase behavior according to the experimental observations rather than relying on a strict interpretation of the van der Waals equation. For example, cyclohexane and benzene are best represented by $k_{ij} > 0$, but they are not believed to exhibit LL behavior. Hence, they are classified as type I.

One might assume that reproducing the van Konynenburg and Scott analysis for any new equation should be straightforward given the advances in computers since the work was originally performed in 1968. Figs. 5 and 6 shed light on why this is not the case. These figures show how subtle changes in the Gibbs energy lead to substantial changes in the phase behavior. In some cases, the important values of Gibbs energy look like little more than slips of the pen. In Fig. 5, a bend in the Gibbs energy at low x_2 indicates the existence of a vapor root, but the tangent line connecting this root to the liquid lies above the tangent line connecting the two liquid roots. As Gibbs energy must be minimized, the tangent line between the two liquid roots is favored if a sufficient amount of component 2 is present. Otherwise, the VL region is the most relevant. In Fig. 6, all three roots lie on a

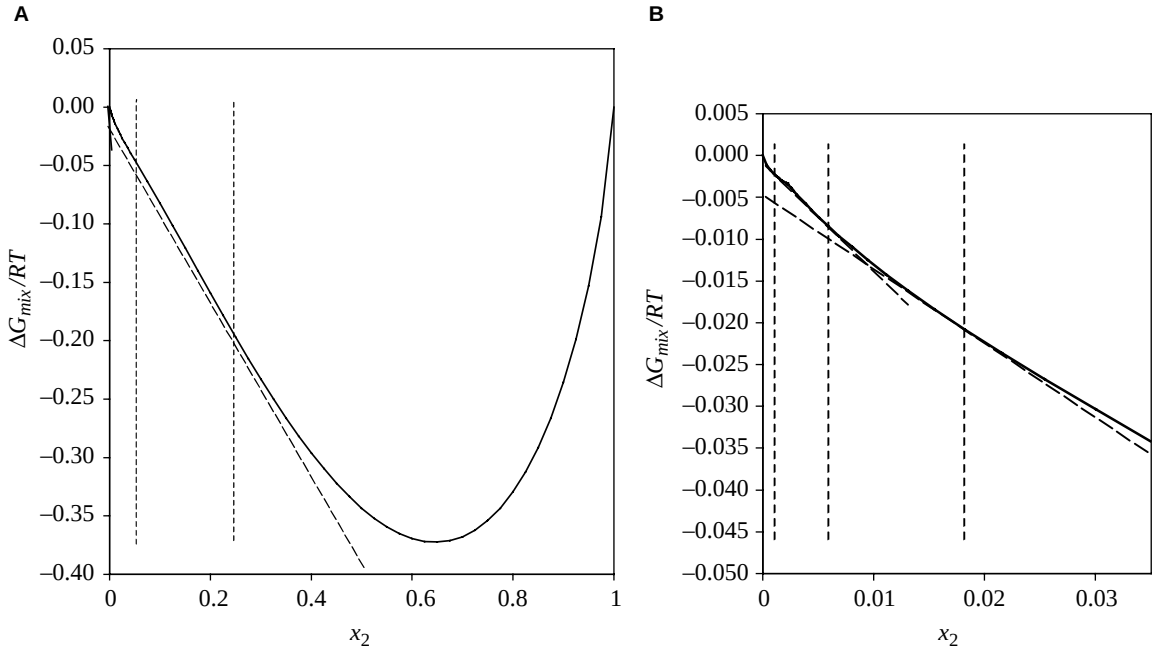


Fig. 5 Gibbs energy of a system inside the VLL region of a type IV or V system showing the LL tangent (A) and the VL tangent (B) at the same conditions of $T_r = 1.005$, $P_r = 1.0$. *(View this art in color at www.dekker.com.)*

single tangent line, indicating that this condition lies exactly on the liquid–liquid–vapor (LLV) line. Noting that Figs. 5 and 6 represent the same pressure at increasing temperatures, we may wonder what happens when the temperature is further increased at constant pressure. Basically, the middle root recedes from tangent line and we are left with equilibrium between the vapor and lower liquid. Note that van Konynenburg and Scott show this same transition analysis over a wider range of conditions, but their figures are highly idealized, masking the subtlety of the changes in Gibbs energy.

Figs. 5 and 6 illustrate instances of type IV or V behavior near the high-pressure LLV line. This LLV line is a short segment, in contrast to the low-pressure

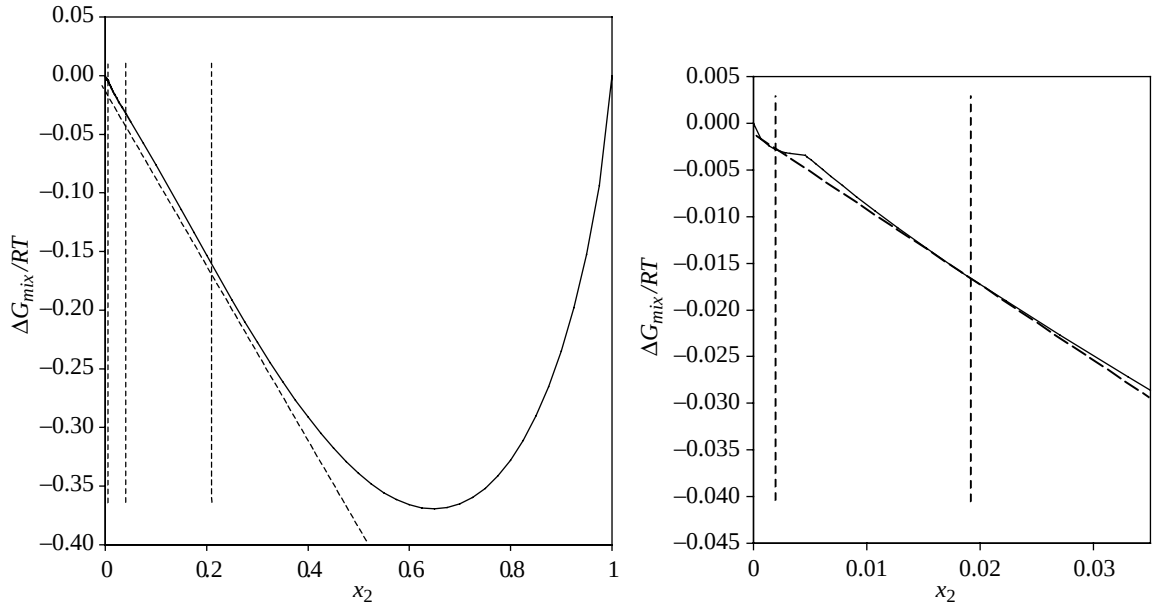


Fig. 6 Gibbs energy of a type IV or V system showing the LLV three-point tangent at $T_r = 1.010$, $P_r = 1.0$. *(View this art in color at www.dekker.com.)*

LLV line in types II–IV. The low-pressure LLV line is relatively easy to identify computationally, because the liquid binodal region increases in size as one lowers the temperature. Therefore, it is simple to find the LLV region at a low temperature, then steadily raise the temperature using the previous result as the initial guess for the next. Initially, locating the high-pressure LLV line is more difficult because it only exists over a narrow range of conditions. Kolafa et al.[10] have developed a modern program for mapping global phase diagrams. By incorporating hydrogen bonding into their equation of state, they were able to easily reproduce the qualitative features of closed loop (Type VI) diagrams.

Finally, we should address the trend as the size ratio varies. The most substantial alteration is the expansion of type IV behavior. We can estimate the result at $k_{ij} = 0$ by considering the critical loci (SRK) as computed by the Soave–Redlich–Kwong[11] equation of state. Soave's SRK equation was shown to accurately correlate the critical loci of a large number of systems.[12] Considering only the behavior along $k_{ij} = 0$, we obtain a plane with ζ on the abscissa and ξ on the ordinate, where $\xi \equiv (b_2 - b_1)/(b_2 + b_1)$. In this instance, we compute ζ and ξ from a and b parameters of the SRK equation, which are different from a and b parameters of the van der Waals equation. Fig. 7 is the result of this procedure.

Several caveats should be noted while referring to Fig. 7. First, van Konynenburg and Scott observe a narrow range of type IV behavior from $\zeta = 0.26$–0.30 that is qualitatively consistent with Fig. 7, but their value of ξ was 0.333. Fig. 7 indicates a value of $\xi = 0.1$ to match this ζ range. We may explain this by noting that the values of a and b parameters are different for the two equations of state. Second, types I and V are omitted from Fig. 7 although many phase diagrams have been classified as types I and V experimentally. Similar to van Konynenburg and Scott, we attribute this to interference from solid phase boundaries in the experimental systems, hypothesizing that an LL region must exist at $T \to 0$ when the geometric mean is applied. Hence, type II systems will be classified experimentally as type I if the LL region lies below the solid phase boundary, and similarly for type IV and V systems. Third, it should be noted that for the SRK equation, the $\xi - \zeta$ relation increases up and to the left along a homologous series; hence, the number of components in a series that overlap the type IV region may be greater than initially anticipated. On a related note, solute molecular weights greater than ~500 all lie in a tiny portion at the upper left corner of Fig. 7. Thus, a vast amount of polymer solution experience lies in a region of Fig. 7 where types II–IV are barely distinguishable based on the current analysis, and small variations in k_{ij} would drastically alter the phase behavior. Finally, one should note that negative values of ξ are omitted in Fig. 7.

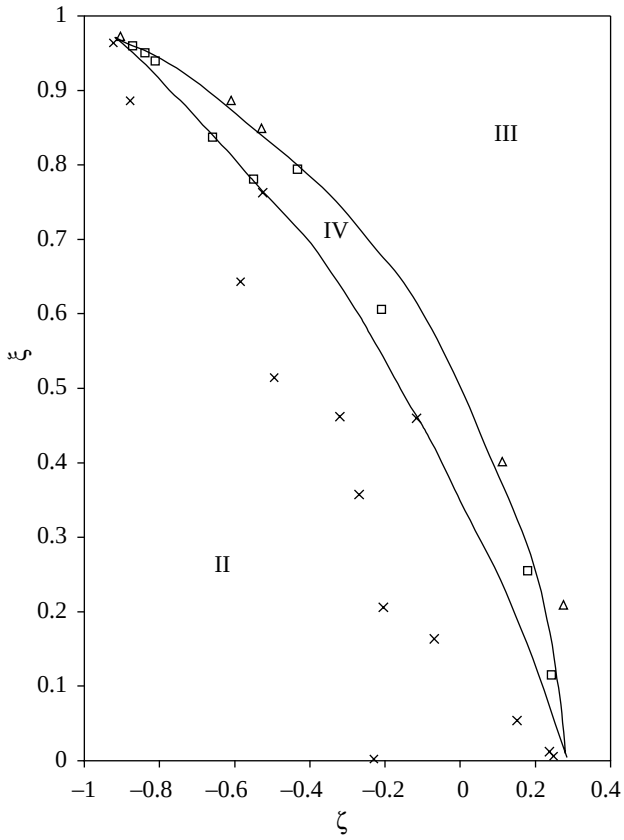


Fig. 7 SRK master diagram for phase behavior at $k_{ij} = 0$. X, Δ, □ are the computed type for II, III, and IV regions. Computations are based on n-alkanes with N_2, CH_4, C_2H_6, CO_2, CH_3OH, and H_2O as solvents.

A small number of the sampled binary combinations resulted in negative ξ values and no type IV behavior was observed, but the sampling was so sporadic as to be inconclusive.

With this background, the reader should be ready to consider the trends in phase behavior for a number of experimental systems. Several compilations are available for experimental systems and their type classifications.[7,13–16] These are typically in terms of homologous series, similar to the experimental programs themselves. We prefer to organize the list according to the type behavior. The type behavior plays a larger role the chemical processing.

Type I systems

A large number of these have been compiled by Rainwater[17] and Luks and Miller.[15] For the most part, these are mixtures of components with similar molecular weights and chemical structures like $N_2 + CH_4$, $N_2 + O_2$, or $\text{Refrigerant}_{22} + \text{Refrigerant}_{114}$. The components may differ structurally to an extent that the solution behavior would be classified as nonideal,

however, and azeotropes are possible. Some of the more unusual examples include: HCl + dimethyl ether, methane + pentane (and lower paraffins), ethane + n-heptadecane, propane + nonacosane, CO_2 + n-hexane, and N_2O + n-heptadecane[18] and ethyl ether + n-butanol. Simple systems like these can play important roles in refrigeration, especially if they form azeotropes. Applications of azeotropes in refrigeration are complicated by small variations in the azeotropic composition with respect to pressure. The result in the presence of a small leak is that the composition of the refrigerant varies depending on whether the leak is at high pressure or low pressure. This necessitates a protocol of replacing all refrigerant of an azeotropic system if any leak is suspected. Hence, even the thermodynamically simplest systems can induce frustration in a particular application.

Type II systems

The primary examples include CO_2 + n-heptane through CO_2 + n-dodecane. The CO_2 + n-hexane system is a prime example of a system that would very likely be type II if not for the interference of the solid phase boundary. The SRK equation of Fig. 7 overestimates the range of type II behavior for CO_2 systems, terminating with pentadecane instead of dodecane. Small positive values for k_{ij} would improve the likelihood of quantitative agreement in this regard, but distract from the utility of Fig. 7 as a general, qualitative guideline. CO_2 mixtures with 2-hexanol and 2-octanol also exhibit type II behavior. NH_3 + n-butane provides another interesting example of type II behavior. Other mixtures of NH_3 + higher hydrocarbons are likely to exhibit this type of phase behavior as well.

Type III systems

Type III behavior indicates the most extreme asymmetry between the components of a binary mixture. Nearly all H_2 systems supply striking examples of type III behavior. CO_2 mixtures with 2,5-hexanediol and 1-dodecanol are also classified as type III. The system CO_2 + n-tridecane is peculiar because it was classified by van Konynenburg and Scott as type III, whereas Enick et al.[19] have classified it as type IV, owing to experimental identification of a three-phase region. The system CO_2 + n-tetradecane is a variation on type III, where the solute-rich locus terminates in a solid(wax)–liquid–liquid boundary. Several important systems fall into a similar category. For example, CO_2 + naphthalene is commonly used as a model system for supercritical extraction. The naphthalene system differs from the n-tetradecane system in that the solute-rich locus terminates at a higher temperature than the solvent-rich locus, leaving an open gap. For CO_2 + tetradecane, on the other hand, the two loci overlap in temperature. Wax precipitation from natural gas (mostly CH_4) probably has a similar gap. In the temperature range of the gap, there is a large region of vapor + solid equilibrium, uninterrupted by liquid phases. The *x-P* plots commonly seen in the supercritical literature appear as *P-x* projections at constant temperature when considered in this light. These distinctions are clarified in Fig. 13.6 of Ref.[5]. Branching in the solute favors lower melting temperatures resulting in normal type III behavior, as exemplified by CO_2 + squalane.[20]

Type IV and V systems

We consider these two types simultaneously because they share their distinctive feature. That feature is an interruption in the critical locus where two liquid phases appear over a short range of compositions before the critical locus reappears as a liquid–liquid critical point. Unlike low-temperature LL behavior, varying the pressure has a strong impact on type IV or V liquid–liquid–equilibria, (LLE) making it appear or entirely disappear over a remarkably narrow range of pressures. Systems that exhibit type IV behavior include methane + 1-hexene and benzene + polyisobutylene, the only polymer solution mentioned by van Konynenburg and Scott. Peters has also speculated that methane and ethane mixed with alkylbenzenes will form type II–IV solutions, in contrast to the I, III, V solutions of the n-alkanes.[21]

The alkane mixtures provide the prototypical examples of type I → type V behavior. Methane + hexane (and higher alkanes), ethane + octadecane, and propane + pentatriacontane are all type V. The upper LL regions of these systems are noteworthy in that the temperature difference between the UCEP and the LCEP seems to monotonically increase with increasing carbon number.[21] Ultimately, this trend must reverse as type III behavior sets in, but no indication of this reversal has been observed experimentally. Mixtures of methane with hexane isomers provide unusual examples of type V phase behavior. Type V behavior is exhibited for all isomers except 2,2-dimethyl butane. Ternary mixtures of methane with the 2,2 and 2,3-isomers provide a rare example of tricritical behavior. Turning to another example, the type V LLV locus becomes extremely short as the asymmetry of the mixture increases to the point where transition to type III behavior is approached. Ethane + *p*-dichlorobenzene provides an example of this phenomenon, with an LLV locus extending over a mere 0.6 K.[22] Such an odd effect may seem to have little practical significance, unless one considers the impact of an unexpected precipitation on a critical pipeline.

A prominent chemical processing application of type V behavior is exemplified in the Selexol process. The Selexol process was originally based on type V behavior in propane + triglyceride systems and has also been applied to propane + fatty acids. Dimerization of the acids makes them perform effectively like ''diglycerides'' thermodynamically. Hixson et al. have published extensively on this process.[23,24] They have also developed the concept of using temperature gradients within an extraction column to enhance separations, a concept recently discussed in the context of SCF CO_2 extractions.[25] The glycerides partition favorably into the propane-rich phase. Downstream manipulations of the pressure and temperature lead to delicate fractionation of the glyceride constituents. As an example of a nutraceutical application, vitamin A has been produced in this way. Applications based on CO_2 or dimethyl ether remain as possibilities.

Summary of Phase Diagram Basics

Summing up, we reiterate that critical phase diagrams describe an array of subtle, complex behaviors. Some of these create opportunities for chemical processing like the Selexol process. Some characterize potential problems like wax precipitation. Either way, the phase diagrams provide concise insight into a broad scope of possibilities that researchers in high pressure chemical processing need to be aware of. It should be noted that the discussion here is actually quite circumscribed in that only binary mixtures have been explicitly considered. Recent work shows a quantum leap of higher order complexity when it comes to treating ternary mixtures and cosolvent effects.[26]

RESEARCH TOPICS

Beyond the basics, there are a number of efforts underway to advance the role of critical phase behavior in chemical processing. A few of these are already making an impact, but most remain as concepts under development. We briefly summarize key elements of these and provide links that should facilitate a broader understanding on subjects of interest. This list does not aspire to be comprehensive. Rather, it reflects the author's biased impressions of subjects that might resonate with developers of chemical processes. It is hoped that this brief survey will improve with feedback from readers and the passage of time.

Theory

Much of the traditional theory behind critical phase behavior has been discussed in the context of phase diagrams. This section is devoted to a brief review of fundamental studies that have been less fully developed when it comes to chemical processing in the critical region.

Among the early theoretical developments in SCF studies was the finding that solvent molecules tend to cluster around solute molecules,[27] giving rise to local density functionals that may deviate markedly from the bulk density. Similarly, local compositions differ markedly from bulk compositions. These composition enhancements have been demonstrated experimentally by fluorescence spectroscopy and theoretically by molecular simulation[28] and by integral equation theories.[29] This enhances the solubility of heavy solutes to a great extent. Another impact of the local composition is on reaction kinetics. One might suspect that reaction kinetics should follow the local composition rather than the bulk composition. In fact, this has been demonstrated experimentally and described theoretically.[30,31] There appears to be little recognition of these phenomena in chemical processing, but that may be because of lack of awareness on the part of reaction engineers. Alternatively, critical phase specialists may not be aware of high-pressure reactive processes that could benefit from this knowledge. Either way, this peculiar property of the critical region remains to be exploited in chemical processing.

One theoretical subject that has received much attention is the subject of equations of state. Equations of state based on conventional mixing rules, like the Peng–Robinson[32] or SRK equations, tend to underestimate clustering effects. Clustering is essentially a local composition effect. Another strong contributor to local composition anomalies is hydrogen bonding. The local composition mixing rules are typified by the Wong–Sandler modification of the Peng–Robinson equation (PRWS).[33] Methods like this are capable of correlating complex phenomena, but they show signs of detachment from physical reality. For example, the k_{ij} parameters tend to be extremely large and erratic in value. Nevertheless, this flexibility can be applied to clustering that arises from critical fluctuations or from hydrogen bonding, obviating the need to draw distinctions. At present, the PRWS approach is probably the least likely to fail when it comes to correlating complex phase equilibria. This makes it the next logical model to consider after trying a simple equation like the Peng–Robinson model. Hydrogen bonding equations of state are typified by the statistically associating fluid theory (SAFT) equation.[34,35] These models apply Wertheim's[36] perturbation theory to obtain an accurate characterization of the association and solvation interactions that are primarily responsible for the difficulties experienced in representing phase equilibria involving polar substances. These theories are very appealing from a chemical perspective and are promising for

predictions. The solvation parameters follow trends that would be expected from correlations of spectroscopic measurements like the Kamlet–Taft parameters.[37] One of the more important advantages of the SAFT approach is that extension to polymer solutions is entirely straightforward, which cannot be said of the PRWS approach. Furthermore, the formalism lends itself to treatment of copolymers just as straightforwardly. For example, Feelly et al.[38] were able to accurately predict copolymer solubilities in CO_2 over the range of 350–550 K and 1100–2100 bars using k_{ij} values calibrated for homopolymers. Optimization on the molecular scale is greatly facilitated by the SAFT approach, showing great promise for future development in chemical processing. A recent review of modifications and applications of the SAFT approach has been written by Muller and Gubbins.[39]

Petroleum Applications

Oil and gas recoveries provide examples of chemical processing at high pressure that arise quite naturally. A short list of related research topics would include CO_2 flooding for improved oil recovery, and wax and hydrate remediation in gas recovery.

The temperature and the pressure at which oil recovery takes place are steadily increasing as reserves are sought at progressively greater depths. The added expense of operating at these depths also intensifies motivation to enhance the productivity of each well as much as possible. One strategy for this enhancement has been to recover or generate CO_2 at the wellhead and inject it into the well to displace residual oil. The CO_2 is generally miscible with the oil, reducing its density and viscosity and facilitating recovery. The problem with this approach is that some of the CO_2 bypasses the oil and goes directly to the reservoir outlet. The method becomes ineffective when these ''fingers'' of bypass flow become numerous. To overcome the fingering problem, dissolving polymers into the CO_2 can enhance the viscosity. The problem then becomes improving polymer solubility. This has been a longstanding problem. Early work showed that additives like toluene were effective, but the amount required (>10 wt%) was too large to be economical.[40] An alternative approach is to customize the molecular structure to enhance solubility. Beckman et al. have found that fluorination promotes solubility to a sufficient extent that viscosity can be increased by a factor of 3 with 4 wt% polymer.[41] Unfortunately, fluorinated polymers are expensive. When considering the immense volumes of materials being pumped around in an oil field, a small percentage is multiplied by a very large number. To be economical, the viscosity enhancing agents need to be nearly as inexpensive as the oil itself. The subject of polymer solubility in CO_2 has stimulated research with biomedical applications. These are discussed in a later section.

Wax precipitation from natural gas is another longstanding problem being exacerbated by progressively deeper drilling. Assuming that natural gas is predominantly methane and the relevant phase diagram is similar to type III, it may be possible to surpass the critical locus at the elevated temperatures and pressures of reservoir conditions. As a related phenomenon, high-pressure synthesis of polyethylene from ethylene is known to operate in this upper region of the type III diagram.[6] All temperature and pressure drops would then result in precipitation in the recovery stream. Pedersen described one method of predicting solid–vapor precipitation based on the assumption of an ideal solution in the solid phase.[42] Recently, Coutinho et al.[43] have reported improved results when a predictive activity model is applied to the solid phase. The fluid phases have been accurately modeled with conventional equations and mixing rules in these instances. The availability of reasonable models for these systems provides opportunities to brainstorm through a number of prospective remedies. Promising strategies focus on precipitating at controlled locations and seeding heavy components that lower the viscosity of any wax that does form. With knowledge of the phase diagrams, it should be possible to accommodate wax precipitation problems with reasonable success.

Gas hydrates form when small molecules like nitrogen, methane, CO_2, or propane come in contact with water at pressures above 30 bars and temperatures of roughly 275 ± 30 K. The composition of the gas species in the hydrate is in the order of 5 wt%. Technically, hydrate description requires an entirely new family of phase diagrams. We defer to existing literature for a detailed discussion.[44] Like waxes, hydrates are a nuisance to gas recovery. A commonly suggested treatment is to dope the gas stream with methanol. Methanol disrupts the formation of hydrates, necessitating higher pressures before the hydrates can form. Recently, Peters et al. have reported that additives like cyclohexane or cyclobutanone can have the opposite effect, reducing the pressure or elevating the temperature of hydrate formation.[45] At first, this may seem like an added nuisance, but it adds a remediation strategy. If precipitation can be effected at controlled locations, then the problem may be remedied.

Bio/Medical Applications

The studies most commonly identified with supercritical fluids and critical phase behavior are those concerned with extraction of natural products. No review of this subject would be complete without

reference to the decaffeination of coffee.[46] A commercial process for many years, it still acts as a model system for treatment of natural products.[47] In this application, purification of the caffeine enhances the value of both the caffeine and the residue. More recent applications have focused on simply extracting the valuable component. Typical extracts have been alkaloids, triglycerides, fatty acids, and antioxidants. Any given issue of the *Journal of Supercritical Fluids* is likely to contain 3–4 articles on similar extractions. Fatty alcohols can be obtained from fatty acids by hydrogenolysis in near critical propane. The phase behavior of this system shows signs of type IV behavior with a pressure-dependent LL region near room temperature and 30 bars.[48] Apparently, these efforts are still largely academic. In a brief search, 72 patents since 1980 mentioned supercritical extraction in their titles. Many of the early patents were focused on upgrading fuels. Thirteen patents target specific natural products, with the earliest dating from 1989. Fatty acids and taxol related species are mentioned twice each. Three patents focused on flavors. The rest of the patents targeted esoteric species, with one recent patent focusing on a specific protein. None of the patents for natural products were from the United States. Three were from Europe, four from China, and six from Japan. This trend suggests that much of the chemical processing interest in natural products derives from overseas.

Greater interest in the United States has been focused on the formation of microparticles for encapsulation and time release applications. DeBenedetti et al. have described two processes for microparticle production: one by rapid expansion of supercritical solutions (RESS)[49,50] and one based on a supercritical antisolvent (SAS), for which they hold a patent.[51] In the RESS process, a homogeneous polymer + CO_2 solution is expanded through a nozzle, dropping the pressure and precipitating polymeric particles whose size and shape depend on the precise conditions of the expansion. Nineteen patents identified the concept of expansion from supercritical solution since 1988, the substantial majority of which targeted production of microparticles. In the SAS process, a homogeneous solution is exposed to a supercritical phase that does not dissolve the polymer, inducing precipitation. Eight patents identified the concept of SAS, including the original in 1993. Four of these were international patents and three were US patents. Five of these patents targeted drugs or proteins. Recent references to these applications provide links to further study.[52,53]

The attraction of CO_2 in these applications is its biocompatibility. Any residual CO_2 solvent will be readily metabolized. Use of CO_2 as a biocompatible solvent has required several adaptations to enhance the solubility of proteins and polymers in the CO_2 phase. Most of these have focused on fluorination,[54] although recent efforts have turned toward cheaper esters and ethers.[55] One such adaptation of CO_2 won a Green Chemistry Award for 2002.[56,57] Graft polymerization in CO_2 is another proven means of altering the compatibility and solubility.[58] Johnston et al. have demonstrated how fluorinated surfactants can be used to dissolve proteins into supercritical micelles.[59,60] Johnston is named in seven patents on drug and microparticle preparations. Taken together, the bio/medical applications appear to be some of the most promising for chemical processing in the near term.

Green Chemistry

Similar to CO_2's attractiveness from a biological perspective, it is also environmentally friendly relative to traditional solvents. This property suggests its extended use as a substitute in many applications where traditional solvents have been applied. Two examples can be cited from the paint and dry cleaning industries.[61]

Donohue and collaborators have developed a process for spray painting with CO_2 substituted in place of the usual solvents.[62,63] This is a fully developed, viable process, but it must compete with continually advancing technologies in the paint industry. The competing technologies are already reducing the proportions of solvent by lowering the molecular weight of the paint and making it reactive after coating.[64]

Dry cleaning with CO_2 is rapidly becoming a more competitive process. Two similar technologies were developed at Los Alamos and by DeSimone et al. These processes effectively use near-critical CO_2 for washing and a thermal cycle for regenerating the CO_2. This chemical process has received a surge of interest from ICI and Linde.[65] With this investment, it is likely that this process will increase market share over the near term. A related technology is dedicated to the use of CO_2 as a cleaning agent in the electronics industry. Los Alamos and DeSimone, among others, are now competing in the development of that processing technology.

Carbon dioxide is not the only nonflammable, biocompatible, and widely available solvent. Let us not forget about water. In particular, supercritical water oxidation and related reactive processes have shown a tremendous capacity for reducing toxic chemicals to innocuous constituents. Akiya and Savage have authored a recent review.[66] Of particular interest is the section on hydrolysis in SCF water. Numerous references are tabulated according to chemical family. With regard to the more toxic compounds, Klein et al. have been especially active for many years[67,68] and Tester et al. have focused on alkyl halides.[69]

Modell et al. have patented this technology for nonincineratory destruction of toxic chemicals.[70,71] Early work in this area was related to the supercritical extraction/depolymerization of coal. Lee et al. have found it possible to depolymerize scrap polymer such that clean monomers can be recovered, turning a liability into an asset.[72,73] This approach is especially interesting in the case of polyvinyl chloride (PVC). Whereas PVC normally degrades into polychlorinated hydrocarbons, the presence of oxygen and water promotes decomposition into the monomer and hydrochloric acid. Potential problems in this developing technology include clogging and corrosion.[74]

Summary of Research Topics

Widespread applications of small-scale critical phase technology, like dry cleaning and paint spraying, are likely to have a subtle, autocatalytic effect on general implementation. High-pressure processing equipment is generally perceived as being inherently high in capital cost. In actuality, doubling the mass of steel in one or two key components of an overall process makes a small difference in the overall cost. Major costs are involved, however, in custom developed equipment that requires repeated testing and redesign and cannot be mass produced. One way of representing this is by the size exponent of a cost estimation chart. An exponent of 0.24 indicates the importance of economies of scale for SCF equipment.[75] On the other hand, global demand for a specialty flavor or a nutraceutical is not the sort of thing that can be scaled to large values. Hence, one key to economical critical phase chemical processing is mass production of interchangeable components for small-scale operation. The economics and the prevalence of critical phase behavior in chemical processing could change substantially if such a development were to take place.

CONCLUSIONS

We have summarized briefly the key concepts and applications of a broad and complex subdiscipline of thermodynamics. Necessarily, selective judgments have been made as to what aspects of critical phase behavior are most important to chemical processing. Alternative perspectives on the role of critical phase behavior in chemical processing are readily available. For example, the role of reactive synthetic processes at critical conditions has received little attention here.[76] As one prospective source for further links, the article by Perrut provides an especially relevant survey.[75] Readers are encouraged to survey this literature for themselves and keep abreast of the latest developments. It is a subject of current research and its status in chemical processing is constantly evolving.

REFERENCES

1. Kiselev, S.B.; Ely, J.F.; Lue, L.; Elliott, J.R. Computer simulations and crossover equation-of-state of square-well fluids. Fluid Phase Eq. **2002**, *200*, 121–145.
2. Younglove, B.A.; Ely, J.F. Thermophysical properties of fluids. II. Methane, ethane, propane, isobutane and normal butane. J. Phys. Chem. Ref. Data **1987**, *16*, 577–798.
3. Noyori, R. Supercritical fluids: Introduction. Chem. Rev. **1999**, *99*, 353–634.
4. Van Konynenburg, P.H.; Scott, R.L. Critical lines and phase behavior in binary van der Waals mixtures. Phil. Trans. Roy. Soc. London Ser. **1980**, *A298* (1442), 495–540.
5. Elliott, J.R.; Lira, C.T. *Introductory Chemical Engineering Thermodynamics*; Prentice-Hall: Englewood Cliffs, NH, 1999.
6. De Loos, T. *Supercritical Fluids*; Kiran, E., Levelt Sengers, J.M.H., Eds.; Kluwer: Amsterdam, 1994; 65–89.
7. Streett, W.B. *Chemical Engineering at Supercritical Fluid Conditions*; Paulaitis, M.E., Penninger, J.M.L., Gray, R.D., Davidson, P., Eds.; Ann Arbor Science: Ann Arbor, MI, 1983; 3–30.
8. Ochel, H.; Becker, H.; Maag, K.; Schneider, G.M. J. Chem. Thermo. **1993**, *25*, 667.
9. Brunner, E. Fluid mixtures at high-pressures.9. phase-separation and critical phenomena in 23 (normal-alkane + water) mixtures. J. Chem. Thermo. **1990**, *22*, 220.
10. Nezbeda, I.; Pavlicek, J.; Kolafa, J.; Galindo, A.; Jackson, G. Global phase behavior of model mixtures of water and n-alkanols. Fluid Phase Eq. **1999**, *158*, 193–199.
11. Soave, G. Equilibrium constants from a modified Redlich-Kwong equation of state. Chem. Eng. Sci. **1972**, *27*, 1197.
12. Elliott, J.R.; Daubert, T.E. Evaluation of the equation of state method for calculation of the critical properties of mixtures. Ind. Eng. Chem. **1987**, *26*, 1686–1691.
13. Rowlinson, J.S.; Swinton, F.L. *Liquids and Liquid Mixtures*, 3rd Ed.; Butterworths: Boston, 1982.
14. Schneider, G.M. High-pressure investigations on fluid systems — a challenge to experiment, theory, and application. J. Chem. Thermo. **1991**, *23*, 301–326.
15. Miller, M.M.; Luks, K.D. Observations on the multiphase equilibria behavior of CO_2-rich and

ethane-rich mixtures. Fluid Phase Eq. **1989**, *44*, 295.
16. Brunner, E. Fluid mixtures at High Pressures. VII. Phase separations and critical phenomena in 12 binary mixtures containing ammonia. J. Chem. Thermo. **1988**, *20*, 1397.
17. Rainwater, J.C. *Supercritical Fluid Technology: Reviews in Modern Theory and Applications*; Bruno, T.J., Ely, J.F., Eds.; CRC Press: Boca Raton, FL, 1991; 57–148.
18. Jangkamolkulchai, A.; Lam, D.H.; Luks, K.D. Fluid Phase Eq. **1989**, *50*, 175–187.
19. Enick, R.M.; Holder, G.I.; Morsi, B.I. Critical and 3 phase-behavior in the carbon-dioxide tridecane system. Fluid Phase Eq. **1985**, *22*, 209–224.
20. Schneider, G.M. Angew. Chem. Int. Ed. Engl. **1978**, *17*, 716–727.
21. Peters, C.J. *Supercritical Fluids*; Kiran, E., Levelt Sengers, J.M.H., Eds.; Kluwer: Amsterdam, 1994; 117–145.
22. Diepen, G.A.M.; Scheffer, F.E.C. On critical phenomena of saturated solutions in binary systems. J. Am. Chem. Soc. **1948**, *70*, 4081.
23. Hixson, A.N.; Miller, R. US, 2,219,652, 1940.
24. Bogash, R.; Hixson, A.N. Chem. Eng. Progress **1949**, *45*, 597–601.
25. Stahl, E.; Quirin, K.-W.; Gerard, D. *Dense Gases for Extraction and Refining*; Springer-Verlag: New York, 1988.
26. Scheidgen, A.L.; Schneider, G.M. Fluid phase equilibria of (carbon dioxide plus a 1-alkanol plus an alkane) up to 100 MPa and T = 393 K: cosolvency effect, miscibility windows, and holes in the critical surface. J. Chem. Thermo. **2000**, *32*, 1183–1201.
27. Kim, S.; Johnston, K.P. Clustering in supercritical mixtures. AIChE J. **1987**, *33*, 1603.
28. Knutson, B.L.; Tomasko, D.L.; Eckert, C.A.; Debenedetti, P.G.; Chialvo, A.A. *Supercritical Fluid Technology*: *Theoretical and Applied Approaches in Analytical Chemistry*; Bright, F.V., McNally, M.E.P., Eds.; American Chemical Society: Washington, DC, 1992; Vol. 488, pp. 60.
29. Wu, R.S.; Lee, L.L.; Cochran, H.D. Structure of dilute supercritical solutions — clustering of solvent and solute molecules and the thermodynamic effects. Ind. Eng. Chem. Res. **1990**, *29*.
30. Brennecke, J.F.; Chateauneuf, J.E. Homogeneous organic reactions as mechanistic probes in supercritical fluids. Chem. Rev. **1999**, *99*, 433.
31. Roek, D.P.; Chateauneuf, J.E.; Brennecke, J.F. A fluorescence lifetime and integral equation study of the quenching of naphthalene fluorescenceby bromoethane in super-and subcritical ethane. Ind. Eng. Chem. Res. **2000**, *39*, 3090.
32. Peng, D.Y.; Robinson, D.B. A new two-constant equation of state. Ind. Eng. Chem. Fundam. **1976**, *15*, 59.
33. Wong, D.S.H.; Sandler, S.I. A theoretically correct mixing rule for cubic equations of state. AIChE J. **1992**, *38*, 671.
34. Chapman, W.G.; Gubbins, K.E.; Jackson, G.; Radosz, M. New reference equation of state for associating liquids. Ind. Chem. Eng. Res. **1990**, *29*, 1709.
35. Huang, S.; Radosz, M. Equation of state for small, large, polydisperse, and associating molecules: Extension to fluid mixtures. Ind. Eng. Chem. Res. **1991**, *30*, 1994.
36. Wertheim, M.S. Fluids with highly directional attractive forces. III. Multiple attraction sites. J. Stat. Phys. **1986**, *42*, 459.
37. Abraham, M.H.; Andonian-Haftvan, J.; Kamlet, M.J.; Whiting, G.S.; Leo, A.; Taft, R.W. J. Chem. Soc. Perkin Trans. **1994**, *2*, 1777.
38. Feelly, T.; Becker, F.; Latz, H.; Sadowski, G.; Buback, M. In Paper 82b, AIChE Fall National Mtg; AIChE: New York, Indianapolis, 2002.
39. Müller, E.A.; Gubbins, K.E. Molecular-based equations of state for associating fluids: a review of SAFT and related approaches. Ind. Eng. Chem. Res. **2001**, *40*, 2193.
40. Irani, C.; Zajac, J. Handling of high pour point West-African crude oils. J. Pet. Tech. **1982**, *34*, 289–298.
41. Shi, C.M.; Huang, Z.H.; Beckman, E.J.; Enick, R.M.; Kim, S.Y.; Curran, D.P. Semi-fluorinated trialkyltin fluorides and fluorinated telechelic ionomers as viscosity-enhancing agents for carbon dioxide. Ind. Eng. Chem. Res. **2001**, *40*, 908–913.
42. Pedersen, K.S. Prediction of cloud point temperatures and amount of wax precipitation. SPE Prod. and Fac. **1995**, 46.
43. Coutinho, J.A.P.; Pauly, J.; Daridon, J.L. A thermodynamic model to predict wax formation in petroleum fluids. Braz. J. Chem. Eng. **2001**, *18*, 411–423.
44. Sloan, E.D.J. *Clathrate Hydrates of Natural Gases*; 2nd Ed.; Marcel Dekker: New York, 1998.
45. Mooijer-van den Heuvel, M.M.; Peters, C.J.; de Swaan Arons, J. Gas hydrate phase equilibria for propane in the presence of additive components. Fluid Phase Eq. **2002**, *193*, 245.
46. Prasad, R.; Gottesman, M.; Scarella, R.A. Decaffeination of aqueous extracts of roasted coffee. General Foods Corp. USA, US, 4,246,291, 1981.
47. Mohamed, R.S.; Saldana, M.D.A.; Mazzafera, P.; Zetzl, C.; Brunner, G. Extraction of caffeine, theobromine, and cocoa butter from Brazilian cocoa beans using supercritical CO_2 and ethane. Ind. Eng. Chem. Res. **2002**, *41*, 6751–6758.

48. Rovetto, L.J.; Bottini, S.B.; Peters, C.J. Phase equilibrium data on binary and ternary mixtures of methyl palmitate, hydrogen and propane. *J. Supercrit. Fluids* **2004**, 31, 111–121.
49. Debenedetti, P.G.; Tom, J.W.; Kwauk, X.; Yeo, S.D. Rapid expansion of supercritical solutions (RESS). Fluid. Phase Eq. **1993**, *82*, 311–321.
50. Smith, R.D. Supercritical fluid molecular spray thin films and fine powders. Battelle Memorial Institute, USA. US, 4,734,451, 1988.
51. Debenedetti, P.G.; Lim, G.B.; Prud'homme, R.K. Formation of protein microparticles by antisolvent precipitation. Princeton University. Eur, 542314, 1993.
52. Weber, M.; Russell, L.M.; Debenedetti, P.G. Mathematical modeling of nucleation and growth of particles formed by the rapid expansion of a supercritical solution under subsonic conditions. J. Supercrit. Fluids **2002**, *23*, 65–80.
53. Werling, J.O.; Debenedetti, P.G. Numerical modeling of mass transfer in the supercritical antisolvent process: miscible conditions. J. Supercrit. Fluids **2000**, *18*, 11–24.
54. DeSimone, J.M.; Guan, Z.; Elsbernd, C.S. Science **1992**, *257*, 945.
55. Conway, S.E.; Byun, H.S.; McHugh, M.A.; Wang, J.D.; Mandel, F.S. Poly(lactide-co-glycolide) solution behavior in supercritical CO_2, CHF_3, and $CHClF_2$. J. App. Polym. Sci. 80, 1155–1161.
56. Sarbu, T.; Styranec, T.J.; Beckman, E.J. Design and synthesis of low cost, sustainable CO_2-philes. Ind. Eng. Chem. Res. **2000**, *39*, 4678–4683.
57. Ritter, S.K. Green Challenge. C&E News **2002**, *80*, 26–30.
58. Trivedi, A.H.; Kwak, S.; Lee, S. Grafting of poly (vinyl chloride) and polypropylene with styrene in a supercritical CO_2 solvent medium: synthesis and characterization. Polym. Eng. Sci. **2001**, *41*, 1923–1937.
59. Johnston, K.P.; Harrison, K.L.; Clarke, M.J.; Howdle, S.; Heitz, M.P.; Bright, F.V.; Carlier, C.; Randolph, T.W. Science **1996**, *271*, 624.
60. Jacobson, G.B.; Tumas, W.; Johnston, K.P. Biphasic catalysis in water/carbon dioxide micellar systems. US, 6,479,708, 2002.
61. Tullo, A.H. Dry cleaning and paint spraying are CO_2 'pioneers'. C&E News **1999**, *77*, 13.
62. Lee, C.; Hoy, K.L.; Donohue, M.D. Supercritical fluids as diluents in liquid spray application of coatings and apparatus therefor. Union Carbide Corp., USA, EP, 1989.
63. Argyropoulos, J.N.; Lear, J.J.; Hoy, K.L.; Donohue, M.D. Pressurized fluid composition and process for making same. Union Carbide and Plastics Technology Corp., USA, EP, 481431, 1992.
64. Tullo, A.H. Automotive coatings. C&E News **2002**, *80*, 27–30.
65. Tullo, A.H. ICI enters CO_2 dry cleaning. C&E News **2002**, *80*, 12.
66. Akiya, N.; Savage, P.E. Roles of water for chemical reactions in high-temperature water. Chem. Rev. **2002**, *102*, 2725–2750.
67. Abraham, M.A.; Klein, M.T. Pyrolysis of benzylphenylamine neat and with tetralin, methanol, and water solvents. Ind. Eng. Chem. Res. **1985**, *24*, 300–306.
68. Izzo, B.; Klein, M.T.; LaMarca, C.; Scrivner, N.C. Hydrothermal reaction of saturated and unsaturated nitriles: Reactivity and reaction pathway analysis. Ind. Eng. Chem. Res. **1999**, *38*, 1183–1191.
69. Salvatierra, D.; Taylor, J.D.; Marrone, P.A.; Tester, J.W. Kinetic study of hydrolysis of methylene chloride from 100 to 500°C. Ind. Eng. Chem. Res. **1999**, *38*, 4169–4174.
70. Modell, M. Treatment of organic material in supercritical water. Modar, Inc., USA. US, 4,338,199, 1982.
71. Modell, M.H.; Edward, G.; Gairns, Stuart, A. Method and apparatus for treating paper-mill effluents. Modell Environmental Corp., USA. WO, 9,510,486, 1995.
72. Lilac, W.D.; Lee, S. Kinetics and mechanisms of styrene monomer recovery from waste polystyrene by supercritical water partial oxidation. Adv. Envi. Res. **2001**, *6*, 9–16.
73. Lee, S.; Gencer, M.A.; Fullerton, K.L.; Azzam, F.O. Oxidation-depolymerization process for waste polymers. University of Akron, USA. US, 5,386,055, 1995.
74. Mitton, D.B.; Yoon, J.-H.; Cline, J.A.; Kim, H.-S.; Eliaz, N.; Latanision, R.M. Corrosion behavior of nickel-based alloys insupercritical water oxidation systems. Ind. Eng. Chem. Res. **2000**, *39*, 4689–4696.
75. Perrut, M. Supercritical fluid applications: Industrial developments and economic issues. Ind. Eng. Chem. Res. **2000**, *39*, 4531–4535.
76. McCoy, M. 'Green' processes based on supercritical carbon dioxide are moving out of the lab. C&E News **1999**, *77*, 11–13.
77. Lira, C.T. In *Supercritical Fluid Technology in Oil and Lipid Chemistry*; King, J.W., List, G.R., Eds.; AOCS Press: Champaign, IL, 1996. Reproduced with permission.

BIBLIOGRAPHY

Chemical Reviews **1999**, *99*, 565–590.
Ind. Eng. Chem. Res. **2000**, *39*, 4441–5048.

Cross-Linked Polyethylene

Carosena Meola
Giovanni Maria Carlomagno
Department of Energetics, Thermofluidynamics and Environmental Control (DETEC), University of Naples Federico II, Napoli, Italy

Giuseppe Giorleo
Department of Materials and Production Engineering (DIMP), University of Naples Federico II, Napoli, Italy

INTRODUCTION

In people's opinion, plastic is a light and weak substance that easily melts when warmed. And yet, cross-linking the carbon atoms suffices to transform such material into a superior material that may be resistant to temperature, pressure, corrosion, and that can be used in a variety of applications. In fact, polyethylene (PE) once cross-linked is advantageously employed in the fabrication of blanket insulation for electrical and telephone wires, pipes for the transport of cold and hot liquids, prostheses for the human body, and so forth. Since the late 1960s, when the European scientist Engel first succeeded in cross-linking PE, there has been a proliferation of cross-linking methods with the intention of fabricating a type of PE suitable for a specific need. There are many cross-linking methods; each method has advantages and disadvantages and no one method works well for every product. In fact, a cross-linked polymer obtained with one method may be excellent for one application, but it may be very inadequate for another application. It is important to choose the most effective method for the specific need and to comply with quality standards. Otherwise, a fixed sequence of operations alone does not ensure the product quality, but it is also important to characterize the products through appropriate testing and nondestructive evaluation. Another important point is with regard to the hazards and risks related to the substances, or devices, employed; manufacturers must comply with safety regulations and ecosustainability. The intended purpose of this article is to expose the reader to an overview of the existing methods and to give indications and suggestions about the most appropriate method for a specific application with the existing legislation involving both quality and safety aspects.

CROSS-LINKING METHODS

Polyethylene is certainly the most used kind of plastic. In fact, there are many things made of PE, from shop bags to envelopes for alimentary packaging, cases to hold and protect electronic devices, medical prostheses, and so forth. Indeed, the list could go on without end!

A long chain of carbon atoms (CH_2) represented as:

$$(-CH_2–CH_2–) \tag{1}$$

is also called linear, or high-density polyethylene (HDPE). However, the ethylene molecules do not always add on in a regular fashion, but sometimes, under high-pressure polymerization, ethylene molecules attach as short branches leading to low-density polyethylene (LDPE). Low-density polyethylene is much more economical to produce and can be used in many applications, which do not require specific material characteristics.

Since its accidental discovery in the early 1930s in Great Britain from the failure of a chemical reaction under pressure, researchers' efforts have been driven toward obtaining a PE with specific chemical, mechanical, and thermal characteristics for the fabrication of complex-shaped tools, or for use in adverse environmental conditions.[1] The fundamental way to improve material properties such as impact strength, chemical resistance, and thermal characteristics is via cross-linking. Indeed, the introduction of cross-linked polyethylene (PEX) in the early 1970s was another milestone in the plastic era. From that date, PEX has captured a giant share of the market because of its superior characteristics with respect to other plastics.

Modifications in the polymeric structure can be brought about by several methods, which differ from each other in two main factors. One factor is the state (i.e., molten or solid) of the polymer during cross-linking. The other factor is the type of activator used to promote cross-linking. In this article the different methods are grouped into two main categories:

- Chemical processes, which require a chemical initiator (peroxide or silane) to induce links in the polymer chain.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007720

- Radiation processes, which involve exposure to ionizing radiation from either radioactive sources or highly accelerated electrons, to liberate free radicals for cross-linking.

Chemical Processes

Cross-linking is activated by a chemical substance, which could be:

- A peroxide—the method is called peroxide initiated cross-linking. The resulting PE is also called PEX-A in the European standards.
- A silane—the method is called cross-linking via silane or moisture-based vinyl silane cross-linking. The resulting PE is also called PEX-C in the European standards.

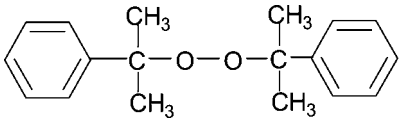
Peroxide processes

The basic process consists of the decomposition of a peroxide at high temperature and the creation of carbon bonds along the PE chain; cross-linking occurs in the molten state. This process was developed in the late 1960s in central Europe by the German scientist Thomas Engel and it is called the Engel process. Indeed, there are several different peroxide-based processes. In the Engel process, a granulated blend of PE, peroxide, and stabilizers is sintered together under high pressure; cross-linking occurs during extrusion through a long heated die. This process is generally used for the production of HDPE. In the Pont a'Mousson process, PE is mixed with peroxide, extruded, and cross-linked in a salt bath at high temperature; the resulting product is a low, or medium, density PE. In the Daoplas process instead, the peroxide is incorporated after extrusion and activated in equipment downstream (the extruder) at high temperature and pressure.

An organic peroxide is a carbon-based chemical that includes a minimum of two oxygen atoms bonded together (–O–O–).[2–4] The general formula is:

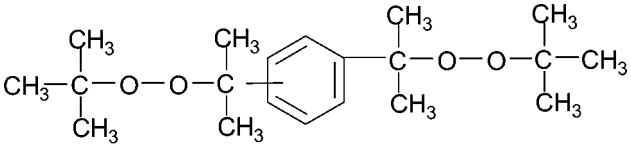
$$R^1\text{–O–O–}R^2 \qquad (2)$$

where R^1 and R^2 can be aryl, alkyl, or acyl groups. There are also peroxides with a second –O–O– bond and three R-groups. Owing to the chemical structure of such R-groups several different families such as alkyl, aryl, and acyl peroxides and peroxyketals are identified. The alkyl peroxides produce the most reactive free radicals and they are the most used for cross-linking.[4] Such peroxides may contain one –O–O– group, as in dicumyl peroxide:



(3)

or two –O–O– groups as in 1,3-1,4bis(*tert*-butylperoxyisopropyl) benzene:



(4)

The peroxyketals contain two peroxy groups bonded on the same carbon atom; this peculiarity causes instability and high reactivity and thereby complicates peroxide transportation, handling, and storage operations.

The peroxide cross-linking reaction occurs in three steps:

Step I: The addition of heat causes peroxide thermal decomposition (i.e., the oxygen bonds break via homolysis). One unpaired electron remains in each oxygen atom and promotes the formation of peroxide radicals.

Step II: Each peroxide radical reacts with the PE molecule, i.e., abstracts a hydrogen atom from the polymer chain, becoming a stable ROH species. The abstraction of hydrogen causes the formation of polymer radicals.

Step III: Two polymer radicals react with each other forming stable PEX.

Peroxides are mainly used for the production of HDPE pipes. The basic process consists of three steps.

- *Mixing*: The peroxide, in liquid or molten phase, is sprayed on PE granules.
- *Extrusion*: The compound (soaked granules plus eventual additives) is poured into the extruder where it is melted and shaped. The cross-linking reaction takes place either in the extruder (Engel) or in the downstream equipment.
- *Curing*: The extruded product enters the equipment where cross-linking is completed under controlled temperature and pressure. The use of high-pressure tanks is necessary for degassing of volatiles and avoiding the formation of voids inside the PEX.

Organic peroxides belong to the aromatic hydrocarbon/alkylbenzene/dialkylbenzene chemical family and are designated by the International Occupational Safety and Health Information Centre (CIS) with the symbols O (oxidizing), X_i (irritating), and N (environmentally

dangerous) and are collocated in the Hazard class 5.2. UN No. 3110. They are generally shock, heat and friction sensitive, and are incompatible with strong oxidizing agents, can react violently with reducing agents, heavy metals, concentrated acids, and concentrated bases, and may ignite organic materials on contact. Such substances can cause local irritations and burns of the skin and mucous membranes of the eyes and the respiratory tract. The risk and safety phrases are: R7-36-38-51-53 and S3-7-17-36-37-39. Thorough care and special technical regulations are required for transportation, handling, and storage.

Vinyl silane moisture

In this process, the cross-linking is activated by silane coupling agents. The organic moieties of silanes can react with many chemicals, including polymers, via typical organic chemistry reactions.[4–7] The organosilane molecule includes a central silicon atom (Si) bounded to two different categories of groups (vinyl and alkoxy), which exhibit different reactivity. The vinyl groups allow silane grafting to the PE backbone. The hydrolyzable alkoxy groups in the presence of water or moisture react (via condensation and hydrolysis) and generate a three-dimensional network of siloxane linkages.

The PE is cross-linked in the crystalline state. The use of silanes causes siloxane (Si–O–Si) bridges to form, which are less rigid than carbon-to-carbon (C–C) bonds induced by peroxides. On the other hand, because cross-linking is accomplished on the shaped product in the presence of water (70–95°C), there is a certain flexibility in the decision making about the desired degree of cross-linking. In fact, one may interrupt the process at a certain time to control the material rigidity. There are two main processes for cross-linking PE with moisture-cured vinyl silanes:

1. One-step (Monosil) process, which involves a continuous feeding of liquid silanes during extrusion. The extruder is equipped with a long barrier screw and an injection system. This speeds up production but poses safety problems owing to the volatile and flammable nature of silanes.
2. Two-step (Sioplas) process, which involves a step to prepare cross-linkable PE and another step to create the final product.

Step 1: A silane blend containing peroxide is melt processed with PE and formed into pellets, which can be stored for several months. Such copolymers can be obtained by two methods:

a. Reactor copolymers involving polymerization of ethylene with vinyl silane. Such technology produces stable copolymers with long shelf life. Conversely, polymerization in a high-pressure reactor results in highly branched products.
b. Silane-grafted PEs are fabricated in compounding equipment involving a single- or twin-screw extruder. Silane may be grafted to any ethylene-polymer, and so PEX of specific characteristics can be obtained. The resulting silane-grafted PE also contains hydrolyzable alkoxy groups, which, in the presence of water, react to join adjacent grafted PE molecules to form stable three-dimensional cross-linked networks of siloxane linkages.[6]

Step 2: The grafted PE pellets, combined with a catalyst masterbatch, is extruded and cured in a high-temperature water bath, or in a steam sauna, to reach the desired cross-linking level.

Polyethylene can also be cross-linked in a single step by using dry silane masterbatches (a powder composed of porous polymer carriers) with the dry silane process.[4]

Silane coupling agents possess the propensity to decompose in contact with atmospheric moisture and to produce hydrogen chloride, methyl alcohol, alkyl ethers of ethylene glycol, and other hydrolysis products. Under certain conditions autoignition can cause an explosion.[8] Thus, carefulness is essential in transportation, handling, and storage; personnel must wear protective equipment to avoid eye and skin contact. It is advised that silane coupling agents be kept away from fire and moisture.

Radiation Processes

These processes do not require addition of any chemicals to the original PE compound. The main effect of radiation is triggering of the chemical ionization process, which results in irreparable damage to the life sustaining chemistry of living organisms and the initiation of cross-linking in polymeric materials. Thus, radiation, apart from improvement of polymer characteristics, is used also for sterilization purposes and is particularly convenient for the cross-linking of PE for medical implants. Ionizing radiation results in energy transfer to the electrons orbiting the polymer nuclei, which causes excitation of the PE molecules. Two effects may be observed:

- Chain scission (radiolysis) with loss of hydrogen atoms and formation of free radicals; extensive chain scission entails reduction of molecular weight with material degradation.
- Cross-linking with merging of two hydrogen atoms to form a hydrogen molecule and merging of two

polymer radicals (via formation of carbon-to-carbon bonds). Cross-linking transforms a linear network into a three-dimensional one with an increase of the molecular weight.

Indeed, during radiation, both chain scission and cross-linking occur simultaneously and competitively.[9,10] The predominance of one, or the other, depends on many factors involving the sensitivity of the polymer to radiation effects (chemical yield), the irradiation dose, the dose rate, the presence of stabilizers or radical scavengers, steric hindrance effects, and the polymer radiation environment.[9] In the presence of oxygen, scission predominates over cross-linking, while in an inert environment, such as nitrogen, cross-linking predominates over scission.

There are two basic radiation methods: gamma radiation (γ) and electron beam radiation (EBR). The first involves exposure of products to a radioactive isotope as in Cobalt 60 (Co^{60}) or Cesium 137 (Cs^{137}). Electron beam radiation consists of bombarding the PE with high-energy electrons, which are accelerated to near light speed. Several kinds of accelerators are commercially available and classified as electrostatic direct current, electrodynamic direct current, radio-frequency linear accelerators, magnetic induction, and continuous wave machines. The target material is passed under the accelerator window by conveyors, reel-to-reel equipment, or other handling means. The governing parameters are electron energy, pulse current, dimensions of the irradiation window, conveyor speed, dose rangeability, beam distribution, properties of the material such as density and hydrogen content, and the number of transits under the window.[11] The dose, D, or energy per mass unit, is calculated by the relationship:

$$D = K\frac{I}{V} \tag{5}$$

with I being current intensity, V feed ratio, and K equipment factor.

Radiation is potentially hazardous and requires ad hoc facilities and procedures. Both γ-rays and electrons are highly penetrating; thus, a shielded environment with thick walls (i.e., 6 ft) constructed of high-density concrete and according to standard legislation is necessary. The product to be treated is loaded into carriers on a conveyor that passes into the cell through a labyrinth. Personnel in these facilities must comply with standard Health and Safety legislation.

Cross-linking of thin-film PE can also be induced by excimer UV lamp irradiation, for coating purposes, to change a hydrophilic surface into a hydrophobic one. With this method a PE film can be bonded to another material without adhesive.

APPLICATIONS

Cross-linked polyethylene is currently employed in civil engineering, electric/electronic fields, medicine, and the packaging industry, and as technology evolves it may be used in many other fields. However, each field has specific requirements and so a specific material must be fabricated with compliance to existing standards (i.e., ASTM F648 for Surgical Implants, ASTM F876 for PEX Tubing, ASTM F877 for PEX Hot and Cold Water Distribution Systems, ASTM D3555 for Wire and Cable Insulation, and so forth). Potentially, the thermomechanical characteristics of PEX can be tailored by appropriate selection of compound ingredients, cross-linking level, and sequence of manufacturing operations. Herein, the attention is focused on three main applications, pipes, electrical cable insulation, and medical implants.

Pipes

Since the introduction of PEX in the early 1970s, plastic pipes became a viable alternative to traditional materials such as copper and clay and cement. Cross-linked polyethylene was immediately considered as a good compromise between safety and economy. On the one hand such material has high resistance to chemical and electrochemical corrosion (due to a bad electric conductor), low encrustation tendency and low head losses, long-term pressure resistance, noise dampening-elastic properties. On the other hand it is light, flexible, and easy to transport and install because it can be delivered in coils; this also minimizes the number of joints and, in turn, the leakage rates.[6,12]

Polyethylene pipes are grouped in classes according to their lifetime rating. The currently used classes are PE80 and PE100, which should withstand a hoop stress of 8 and 10 Mpa, respectively, for 50 yr at room temperature and include a safety factor that accounts for the variability of the working conditions. Now, other classes are under development.[12] Cross-linked polyethylene pipes can fail in a ductile way under high stress, or in a brittle way at low stress. The second failure is caused by slow crack growth (SCG), which represents the weakness of PEX pipes and is most probably induced by defects, or solid particles, such as residual catalyst, attached to the inner pipe surface.[13] To improve stress cracking resistance, a third-generation PE, also called bimodal PE, has been developed.[13] A desired molar weight distribution is obtained with a combination of fractions of branched high- and linear low-molecular-weight PEs. The higher-molecular-weight fraction enhances strength while the linear lower-molecular-weight fraction facilitates processing.

Pipes may be susceptible to gas permeability, which can lead to two adverse effects. One concerns permeability to oxygen, which, in closed circuits such as underflow heating, or radiation heating, causes the steel components to corrode. The other drawback is with regard to the fouling of potable water by carbon dioxide, nitrogen, and other gases from contaminated soil. A possible solution may be the adoption of multilayer structures involving a copolymer of ethylene and vinyl alcohol between two PEX layers. Thus, gas permeation is avoided by the hydrophilic barrier while permeation to liquids is prevented by the hydrophobic PEX.

Electrical Cable Insulation

Application of PE in the insulation of electrical cables requires giving the material specific characteristics such as thermal stability under load, heat resistance, and heat shrinkability. This is attained through four main steps: compounding, forming, cross-linking, and expansion. In particular, fabrication of heat shrinkable cables is based on the shape-memory phenomenon. The technological process consists of conveying elongational stresses into the preform by expanding and then cooling it under stress.[14] The frozen stresses are released when the cable is reheated (during application) causing shrinkage. Generally, both LDPEs and HDPEs are used for cable insulation.

When PEX was first introduced in the electric cable insulation field (almost 40 yr ago) it appeared as the solution to many problems because of its excellent dielectric strength, low dielectric permittivity, low loss factor, high resistance to chemicals, and good mechanical properties. Indeed, substitution of impregnated paper and mineral oil filled cables with PEX for insulation of underground high-voltage electric wires offers many advantages in manufacturing, transportation, installation, and environmental benefits because of the elimination of oil leakage. Unfortunately, many premature failures occurred before the expected duration of 40 yr. The failures were caused by electrical treeing initiated by water treeing. This problem raised a question of great concern among power utilities worldwide.

Many studies have been developed to understand the complex mechanism of tree inception and propagation with insulation breakdown, and to search for any solution without changing material (i.e., without discarding PEX). It is supposed that, under normal operating voltage stress, breakdown channels originate at the tip of defects present in the form of microvoids, gas cavities, conducting inclusions, or intrusions in the insulation structure.[15,16] Water trees are then supposed to consist of tracks of oxidized (hydrophilic) polymer within the hydrophobic PEX. Such tracks do not cause insulation failure but under favorable conditions (lightning surges), evolve into an electrical tree with deleterious consequences.[16] The growth of water trees may be contrasted by hydrophilic clusters, but additives, while improving water tree resistance also reduce hydrophobic and dielectric characteristics. Thus, technology is still being sought to increase tree resistance without affecting valuable PEX characteristics.

Medical Implants

Ultra-high-molecular-weight polyethylene (UHMWPE) was introduced in hip arthroplasty in the early 1960s by the English surgeon Sir John Charnley as a solution to severe hip arthritis and is still the current material of choice for the bearing surface in total joint replacement prostheses. The majority of total hip prostheses implanted in the following three decades have included an acetabular UHMWPE cup articulating against a femoral ball of cobalt–chromium alloy. Wearing of UHMWPE components was observed with undesirable biological response and in 1994 the National Institutes of Health (NIH) officially assessed the tissue inflammation and bone resorption (osteolysis) as caused by the PE wear debris.[17,18]

In 1998, the development of a highly cross-linked polyethylene covered by a U.S. patent was a milestone in arthroplasty.[19] The benefits in wear resistance have led to a proliferation of cross-linking technology into hip and knee replacements. Chemical cross-linking typically involves the generation of noxious fumes and sensitizing by-products of peroxide degradation. Thus, almost all manufacturers use ionizing radiation. Trade names like Crossfire™ (75 kGy gamma radiation and 25 kGy gamma radiation sterilization under nitrogen environment), Durasul™ (95 kGy EBR and ethylene oxide sterilization), Longevity™ (100 kGy EBR and gas plasma sterilization), and Marathon™ (50 kGy gamma radiation and gas plasma sterilization) are commercially available.

It has to be noted that in addition to potentially improving the wear resistance, cross-linking can adversely affect other physical properties such as tensile strength and elongation with susceptibility to crack formation at the articulating surface.[20,21] Another great problem is oxidation induced by the free radicals, which remain entrapped within the crystalline phase of the UHMWPE during radiation. To minimize such undesirable effects it was suggested to remelt the material after gamma irradiation to extinguish the residual free radicals and then to sterilize with ethylene oxide or gas plasma to avoid the reintroduction of free radicals that would occur during radiation sterilization.[20] Another method was warm irradiation with adiabatic melting to make cross-linked UHMWPE with high-energy (10 MeV, 40 kW) EBR without sacrificing the mechanical properties.[21]

The method suitability was assessed through simulator studies that otherwise seemed incapable of predicting the in vivo performance.[22] In fact, there are many factors, apart from the material properties and manufacturing processes, such as the surgical technique and the patient response, which complicate the determination of a cause–effect relationship. Such considerations prompted in June 1999 a Safety Notice from the Medical Devices Agency (MDA SN1999) cautioning careful monitoring of postimplant patients with highly cross-linked UHMWPE components.[23] Despite many studies concerning material improvements in total joint replacement, failures continue to occur and some questions remain about the UHMWPE cross-linking level and procedure to prevent wear. Of course, the patient's response is a primary factor, and ascertaining the benefits of a kind of prosthesis will need several years of clinical data.

TESTING

A product of given characteristics can be obtained by mixing pellets of different polymers, additives such as colorants, stabilizers, antioxidants, flame-retardants, and so forth. Thermomechanical modifications can be brought about either by silanes or peroxides or by exposure to ionizing radiation. A product of good quality is certainly the result of experience, ability, and good knowledge of the interactions between the different substances, and also characterization of final products with the most effective techniques is of vital importance. Many techniques, collected in Table 1, are available for the characterization of PEX.[23–29]

Chemical Analysis

This analysis consists basically of the evaluation of chemical modifications induced in PE by cross-linking methods. The most important parameter to measure is the gel content (fraction or gel%), which indicates the cross-linking degree. It is generally measured, following the ASTM D 2765 standard, as the percentage of the original weight of a sample after extraction for 24 hr in boiling toluene (or xylene) and successive drying in a vacuum oven at 90°C. The variation of gel% with curing time (silane cross-linking) and irradiation dose (EBR) for LDPE is shown in Fig. 1. Cross-linking of 70% can be obtained with exposure to a radiation dose of 90 kGy, or 50 hr curing in hot water with silane. However, the main difference between the two methods is that in EBR the cross-linking degree can be easily increased by increasing the dose. Instead, in the silane method, lengthening curing has no significant effect; an increase in the cross-linking degree requires the addition of an increased percentage of silane to the compound.

Another important parameter is the swelling ratio. A sample is placed in a solvent and the variation of its height is monitored as a function of time through a contact probe. A relationship between sample and solvent allows for the evaluation of the cross-link density and the chain length between cross-links. The technique is named swell ratio testing (SRT) in the American ASTM F2214-02, or hot set test (OST) in the German DIN 57472 Standard.[23,24]

Of vital importance, especially for medical implants, is quantifying the presence of free radicals, which can react with oxygen forming carbonyls and cause material

Table 1 A collection of techniques useful for the characterization of PEX

Chemical analysis	Mechanical analysis	Physical analysis	Thermal analysis	Nondestructive evaluation
Gel content measurements ASTM D2765	Compression tests ASTM D2990	Specific gravity determination	Dynamic mechanical analysis (DMA)	Ultrasonics
Fourier transform infrared spectroscopy (FTIR) ASTM E1421	Elongation tests	Scanning electronic microscopy (SEM)	Differential scanning calorimetry (DSC) ASTM D3417	Scanning acoustic microscopy (SAM)
Resonance spectroscopy (RS)	Fatigue tests ASTM E647	Transmission electronic microscopy (TEM)	Thermogravimetric analysis (TGA)	Photothermal radiometry
Swell ratio testing (SRT) ASTM F2214-02 Hot set test (OST) DIN 57472	Small punch analysis		Thermomechanical analysis (TMA)	Dielectric spectroscopy
Trace element analysis (TEA) ASTM F648	Shore hardness testing (SHT) ASTM D2240			Partial discharge (PD)
	Tensile tests ASTM D2990, D638			

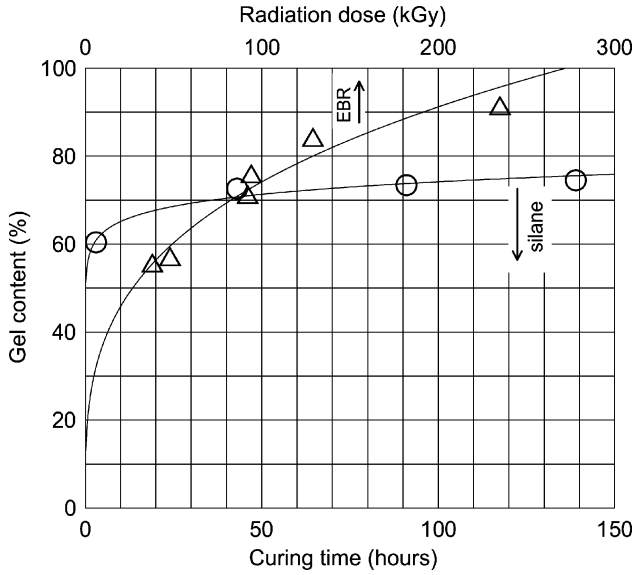


Fig. 1 Variation of gel content with curing time and irradiation dose.

embrittlement. This is done with resonance spectroscopy. The chemical structure of a molecule can also be analyzed by Fourier transform infrared spectroscopy. Many radiolytic products are visible in the infrared spectrum and can be detected by an infrared detector like the mercury–cadmium–telluride detector. The absorbance is proportional to the concentration of the chemical species active at the selected frequency. Again, for medical implants a trace element analysis is performed to ensure absence, or allowed percentages, of some substances like titanium, calcium, chlorine, etc.

Mechanical Analysis

The mechanical properties, such as tensile, compression, shear, and fatigue strengths are fundamental for the product's lifetime. Each product must be tested by taking into account the effective operative stresses. A hip cup should allow a patient to take as many as a million steps a year. Pipes and cable insulation for underground conveyance must perform properly in severe soil conditions and seismic zones; pipes should resist scratches and stresses due to high pressure and temperature. Tensile, compression, elongation, and fatigue tests are commonly performed according to well-traced guidelines to evaluate the mechanical characteristics of materials and their suitability for different applications. The hardness is generally evaluated with the Profiler-P10 and with imprints obtained statically and dynamically. The hardness coupled with mechanical properties is helpful in predicting the material's response to the interaction with other objects, or substances.

A novel method is the small punch analysis, or pin on disk test, which is used to evaluate the weight loss of friction material.[10] It is performed with a metallic pin in friction contact on a sample (small disk); the pin, moving cyclically, yields a stress–strain curve related to the specimen wear. This test is generally performed on retrieved, or aged, UHMWPE components.

Physical Analysis

The density, ρ_p, or specific gravity, can be evaluated by applying the Archimedean principle in water:[24]

$$\rho_p = \rho_w \frac{W_{pa}}{W_{pa} - W_{pw}} \tag{6}$$

where ρ_w is the density of water and W_{pa} and W_{pw} are the weight of the polymer in air and water, respectively.

Morphological modifications arising through cross-linking can be observed by scanning electronic microscopy (SEM). Three images are shown in Fig. 2 for pure LDPE (A), silane grafted and cured in hot water for 80 hr (B), and electron beam irradiated at a dose of 80 kGy (C). A fragment of each material was gold-coated and viewed by SEM. Changes in crystallinity can be observed by transmission electronic microscopy.

Thermal Analysis

Several methods may be used to analyze the material behavior under controlled temperature variations. Differential scanning calorimetry (DSC) is the most utilized thermal analysis technique. It consists of observing and recording (thermogram) exothermic and endothermic phenomena that occur in a sample sealed in an aluminum sample chamber under temperature variations. It is possible to evaluate glass transition temperature (T_g), melting point (T_m), crystallinity degree from the fusion enthalpy, and crystallization temperature (T_c). Typical outputs are shown in Fig. 3 for thermo shrinking insulation of low-voltage cables cross-linked via silane grafting and curing for 91 hr in hot water (A) and electron beam irradiation at a dose of 129 kGy (B). The different heat flow levels are linked to the different cross-linking mechanisms. In fact, irradiation was performed at ambient temperature while the material was in the solid state; instead, curing in hot water, when the material was not completely solid and had poor structural stability, probably led to the formation of molecular defects. Indeed, the response to DSC measurements is strictly related to molecular structure; a low-density material behaves differently from a high-density one.[25]

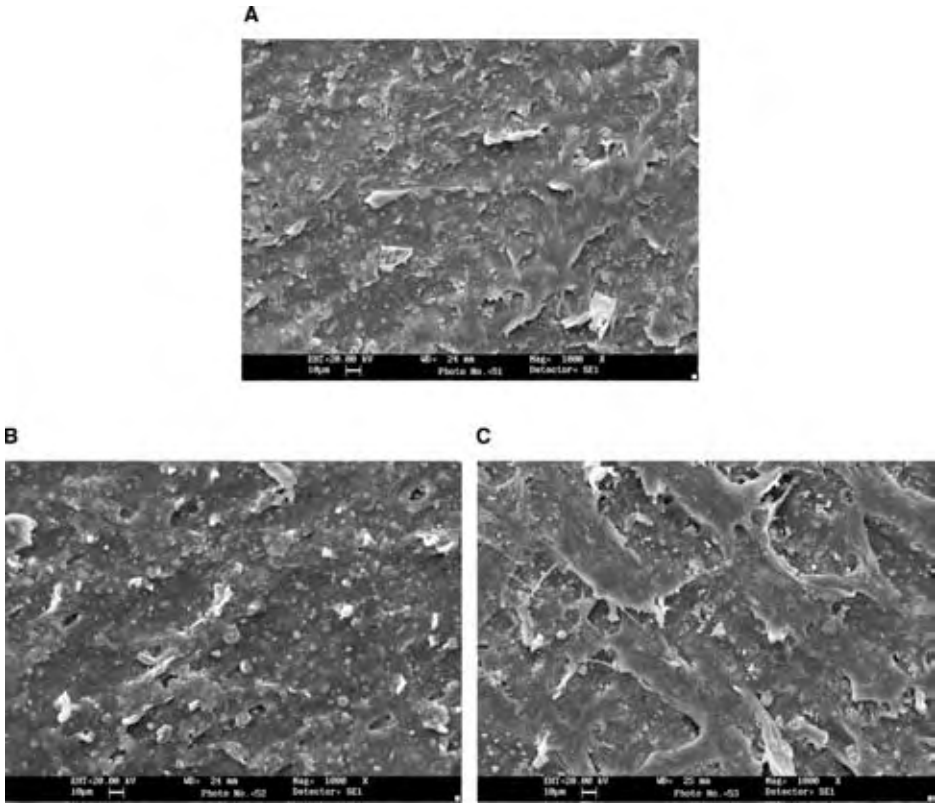


Fig. 2 Scanning electron microscopy images: (A) pure LDPE; (B) silane grafted and cured in hot water; and (C) electron beam irradiated.

The dynamic mechanical analysis gives detailed information about the viscoelastic properties of a sample when heated, cooled, or held under isothermal conditions. The three α, β, and γ peaks displayed by the material before melting can be used to evaluate the effects on the PE molecular structure of additives and cross-linking methods. With the thermogravimetric analysis, information about weight loss, degradation onset temperature, and degradation rate of a sample when heated, or held isothermally, in nitrogen (or argon) atmosphere is attained. The dimensional variations of a sample when it is heated, cooled,

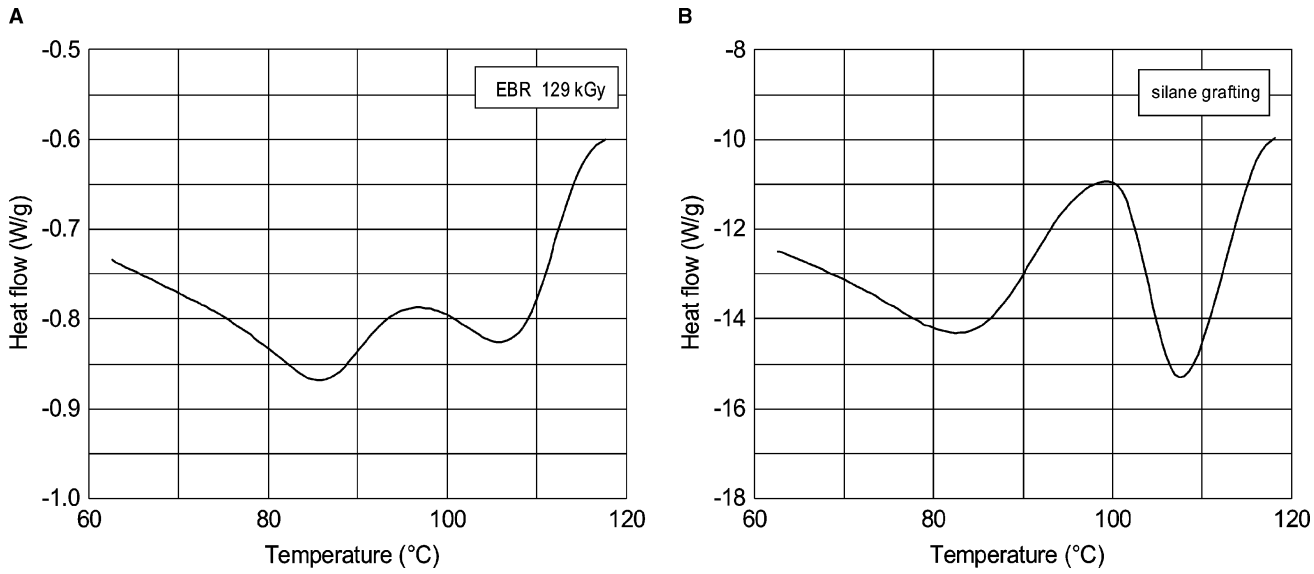


Fig. 3 Differential scanning calorimetry profiles: (A) electron beam irradiation and (B) silane grafting and curing in hot water.

or held under isothermal conditions are observed by the thermomechanical analysis technique.

Nondestructive Evaluation

Cross-linked polyethylene may include impurities and voids from which the major causes of premature failures (SCG in pipes, treeing in electric cables, and cracking in medical implants) can originate. Thus, nondestructive evaluation with effective techniques that can discover defects at the incipient stage, before the component is put in operation, is of vital importance. Conventional ultrasound (up to 10 MHz), which is the current technique for detection of flaws in metal piping and vessels, is limited by the attenuated nature of polymers. A scanning acoustic microscope with operating frequencies up to 100–150 MHz has been found to be more effective.[27]

Another technique that has proved to be practical for the evaluation of PEX is photothermal radiometry, commonly known as lockin thermography, which is the multiplexed version. Studies present in the literature demonstrate the success of such a technique when investigating the effect of molecular orientation in stretched PE, the loss of mechanical properties, local material inhomogeneities due to extrusion and cross-linking processes, as well as material differences linked to the different compounds.[28,29] The test procedure consists of acquiring phase (or amplitude) images while the specimen surface is thermally stimulated with a sinusoidal heat flux. Two phase images, taken at a heating frequency $f = 0.12$ Hz, are shown in Fig. 4 for LDPE cross-linked with EBR. The first image (A) shows a piece of commercial (Megarad, Italy) low-voltage electric cable insulation after expansion to three times its original diameter; it is possible to see yield tracks that are probably due to local material inhomogeneities or nonuniform distribution of forces during enlargement. The second image (B) refers to a sheet irradiated at increasing dose from right to left; it is possible to see dark zones on the left due to material degradation under overdose.

Lockin thermography can be exploited for characterization of PEX through the variation of the phase angle. In fact, the local phase angle variation is well correlated to the local variation of the elastic modulus under variation of irradiation dose, as shown in Fig. 5; more specifically, a differential phase angle (with respect to the phase angle of untreated material) is considered. The modulus of elasticity is evaluated at 150°C above the melting point to relate it directly to the cross-linking degree.[26] Through the variation of the phase angle the variation of the material density can also be evaluated as demonstrated in Ref.[28] for a retrieved UHMWPE hip cup after 9 yr of implantation.

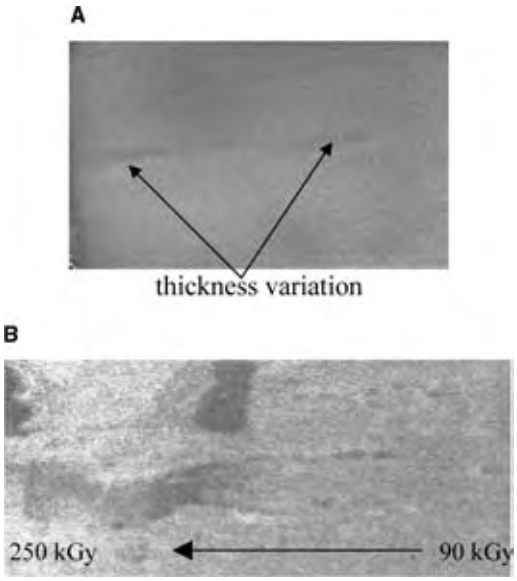


Fig. 4 Phase images for $f = 0.12$ Hz. (A) post expansion cable insulation; and (B) sheet irradiated at increasing dose from 90 kGy (right) to overdose (left). *(View this art in color at www.dekker.com.)*

The electric utility industry is today forced by the Market Directive to adopt condition-based maintenance to ensure steady supply. The method universally adopted to ensure proper installation of cable system accessories and to periodically determine the state of the insulation for decision making about repairing or replacing before failure occurs is partial discharge (PD). This method is based on the time-of-flight principle of induced PD signals; a time variation is attributed to local faults. Calibration is needed to establish a correct relationship between the magnitude

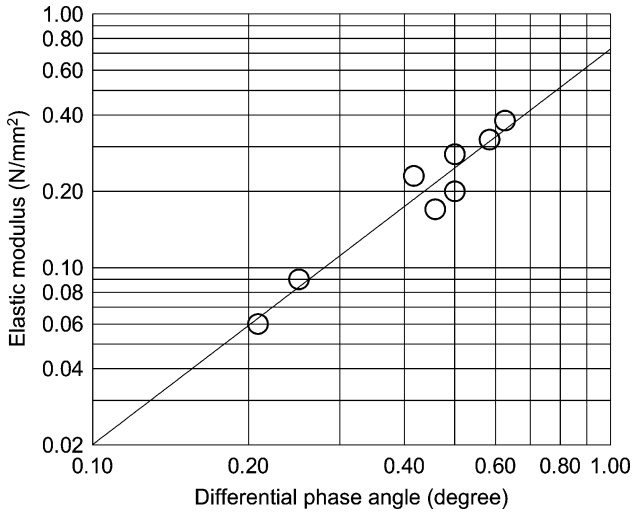


Fig. 5 Differential phase angle against modulus of elasticity for LDPE for varying the electron beam dose.

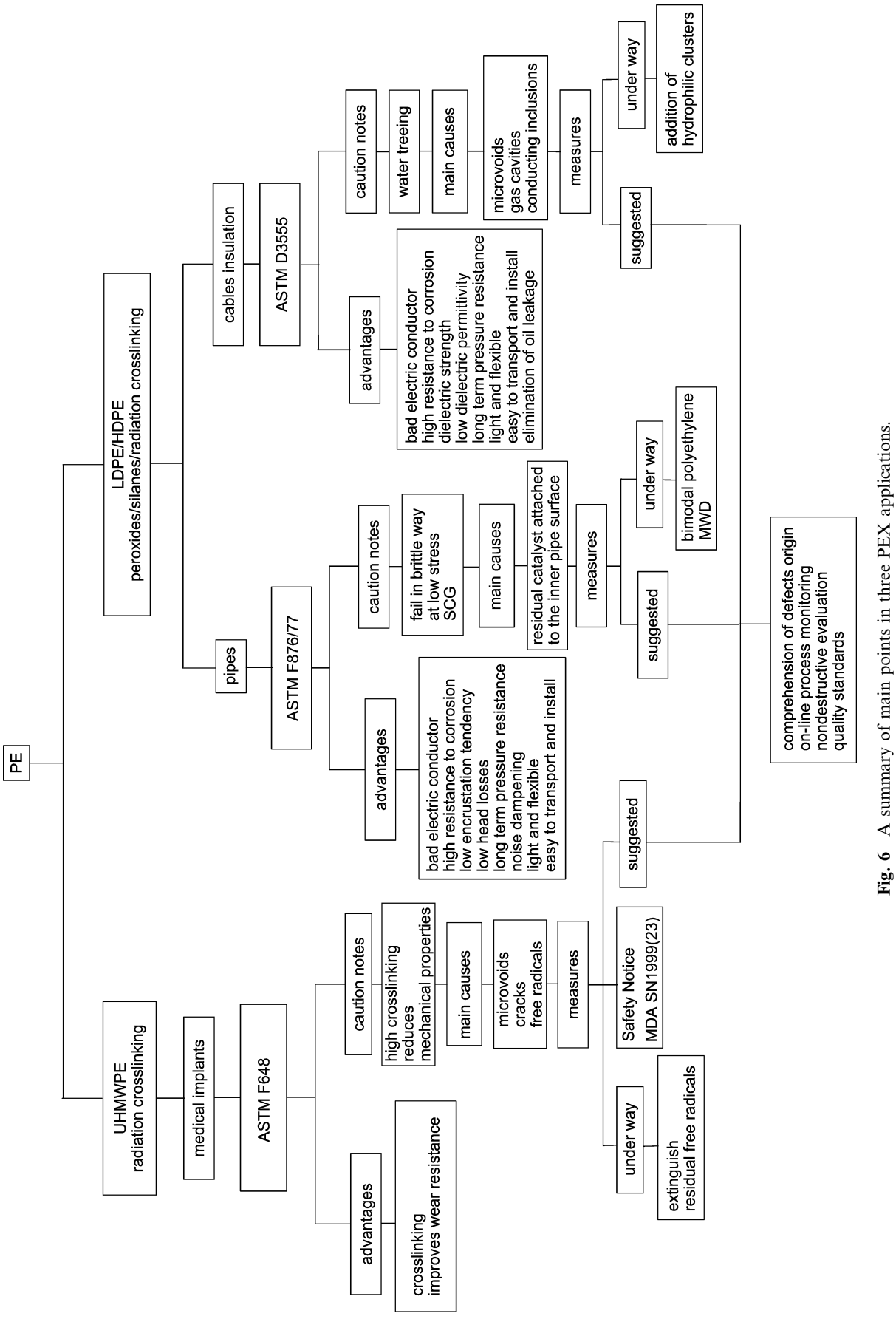


Fig. 6 A summary of main points in three PEX applications.

of the discharge in the cable sample and the signal received; quantitative measurements are complicated by interference and noisy environments. Recently, a new on-site and on-line calibration method, which allows overcoming of the problems of previous methods, has been developed.[30] The insulation deterioration for water treeing is associated with a change in dielectric properties such as loss factor ($\tan\delta$) and capacitance to be measured with dielectric spectroscopy.

CONCLUSIONS

The benefits of PEX over other plastics has led to a proliferation of cross-linking technologies and, of course, each manufacturer claims the superior quality of its product. Indeed, PEX can be advantageously used in many applications owing to caution in:

- Selection of the compound ingredients; fractions of LDPEs and HDPEs; and specific additives such as stabilizers, antioxidants, flame-retardants, and so forth.
- Selection of the most suitable cross-linking method (peroxide, silane, or radiation) that gives the material the thermomechanical properties adequate for the specific application.
- Complying with local and international standards (DIN, UNI, UNE, BS, ASTM) regarding mechanical and physical properties to ensure a safe life without unexpected failure and meet hygienic requirements (ANSI, NSF) to avoid health effects.

Application of PEX in three main fields: piping, cable insulation, and medical implants has been examined. Many premature failures of PEX products in each field were reported in the literature; the origin was recognized in the presence of defects forming during fabrication. A summary with advantages and caution notes is sketched in Fig. 6. It seems that production standards are established mainly on the basis of destructive tests and statistical inference with little attention to in-process monitoring and nondestructive evaluation. In this context, infrared thermography, as a remote imaging system of temperature mapping and nondestructive evaluation, may be advantageously exploited.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to: Megarad s.r.l. (Italy) for supplying specimens and related information, Dr. Paolo Suriano for assistance in performing DSC and mechanical tests, and Dr. Benedetto De Vito for the SEM images.

REFERENCES

1. Gibson, R.O. The discovery of polythene. R. Inst. Chem. Lect. Ser. **1964**, *1*, 1–30.
2. Smedberg, A.; Hjertberg, T.; Gustafsson, B. Effect of molecular structure and topology on network formation in peroxide cross-linked polyethylene. Polymer **2003**, *44*, 3395–3405.
3. Anbarasan, R.; Babot, O.; Maillard, B. Cross-linking of high-density polyethylene in the presence of organic peroxides. J. Appl. Polym. Sci. **2004**, *93*, 75–81.
4. www.specialchem4polymers.com (accessed Jun 2004).
5. Shah, G.B.; Fuzail, M.; Answar, J. Aspects of cross-linking of polyethylene with vinyl silane. J. Appl. Polym. Sci. **2004**, *92*, 3796–3803.
6. Unidelta S.p.A. Crosslinked polyethylene pipes. In *Technical Handbook T4001*; 2000.
7. PolyOne^{TM} Corporation. Technical Report N. 66; 2002.
8. Tamanini, F.; Chafee, J.L.; Jambar, R.L. Reactivity and ignition characteristics of silane/air mixtures. Process Saf. Prog. **1998**, *17* (6), 243–258.
9. Ivanov, V.S. *Radiation Chemistry of Polymers*; VSP: The Netherlands, 1992.
10. Valenza, A.; Visco, A.M.; Torrisi, L.; Campo, N. Characterization of ultra-high-molecular-weight polyethylene (UHMWPE) modified by ion implantation. Polymer **2004**, *45*, 1707–1715.
11. Becker, R.C.; Bly, J.H.; Cleland, M.R.; Farrell, J.P. Accelerator requirements for electron beam processing. Rad. Phys. Chem. **1979**, *14*, 353–375.
12. Scheelen, A. *Recent Developments in PE Pipe Materials with Outlook to the Future*, Technical Report; Solvay Polyolefins Europe: Belgium.
13. Hubert, L.; David, L.; Séguéla, R.; Vigier, G.; Degoulet, C.; Germain, Y. Physical and mechanical properties of polyethylene for pipes in relation to molecular architecture. I. Microstructure and crystallisation kinetics. Polymer **2001**, *42*, 8425–8434.
14. Morshedian, J.; Khonakdar, H.A.; Mehrabzadeh, M.; Eslami, H. Preparation and properties of heat-shrinkable cross-linked low-density polyethylene. Adv. Polym. Technol. **2003**, *22* (2), 112–119.
15. Boggs, S.; Xu, J. Water treeing-filled versus unfilled cable insulation. IEEE Electrical Insulation Mag. **2001**, *17* (1), 23–29.
16. Sarathi, R.; Das, S.; Kumar, C.R.A.; Velmurugan, R. Analysis of failure of crosslinked polyethylene cables because of electrical treeing: a physicochemical approach. J. Appl. Polym. Sci. **2004**, *92*, 2169–2178.

17. Hovie, D.W. Tissue response in relation to type of wear particles around failed hip arthroplasties. J. Arthropl. **1990**, *5*, 337–348.
18. Total Hip Replacement NIH Consensus Statement, Sep 12–14, 1994; Vol. 12 (5), 1–31.
19. Non-Oxidizing Polymeric Medical Implant U.S. Patent 5,414,049, May 9, 1998.
20. McKellop, H.; Shen, F.W.; Lu, B.; Salovey, R.; Campbell, P. The effect of sterilization method and other modifications on the wear resistance of acetabular cups of ultra-high molecular weight polyethylene. A hip simulator study. J. Bone and Joint Surgery **2000**, *82* (12), 1708–1725.
21. Muratoglu, O.K.; Bragdon, C.R.; O'Connor, A.S.; Jasty, M.; Harris, W.H. A novel method of cross-linking ultra-high molecular-weight polyethylene to improve wear reduce oxidation, and retain mechanical properties. J. Arthoplasty **2001**, *16*, 149–160.
22. Greenwald, A.S.; Bauer, T.W.; Ries, M.D. New polys for old: contribution or caveat? J. Bone Joint Surg. **2001**, *83* (2), 27–31.
23. Spiegelberg, S. Analytical techniques for assessing the effects of radiation on UHMWPE. Society for Biomaterials Annual Conference, St. Paul, MN, 2001.
24. Khonakdar, H.A.; Morshedian, J.; Wagenknecht, U.; Jafari, S.H. An investigation of chemical crosslinking effect on properties of high-density polyethylene. Polymer **2003**, *44*, 4301–4309.
25. Vallés-Lluch, A.; Contat-Rodrigo, L.; Ribes-Greus, A. Differential scanning calorimetry studies on high- and low-density annealed and irradiated polyethylenes: influence of aging. J. Appl. Polym. Sci. **2003**, *89*, 3260–3271.
26. Meola, C.; Nele, L.; Giuliani, M.; Suriano, P. Chemical and irradiation crosslinking of polyethylene. Technological performance over costs. Polym. Plast. Technol. Eng. **2004**, *43* (3), 629–646.
27. Avila, S.M.; Horvath, D.A. Microscopic void detection as a prelude to predicting remaining life in electric cable insulation. International Topical Meeting on Nuclear Plant Instrumentation, Controls and Human-Machine Interface Technologies (NPIC&HMIT 2000), Washington, DC, Nov 2000.
28. Busse, G.; Eyerer, P. Thermal wave remote and nondestructive inspection of polymers. Appl. Phys. Lett. **1983**, *43*, 355–357.
29. Meola, C.; Carlomagno, G.M.; Prisco, U.; Vitiello, A. Non-destructive control of polyethylene blanket insulation. Res. Nondestr. Eval. **2004**, *15* (2), 55–63.
30. Zhong, L.; Chen, G.; Xu, Y. A novel calibration method for PD measurements in power cables and joints using capacitive couplers. Meas. Sci. Technol. **2004**, *15*, 1892–1896.

Crystal Growth

C. W. Lan
Department of Chemical Engineering, National Taiwan University, Taipei, Taiwan, ROC

W. C. Yu
Department of Molecular Science and Engineering, National Taipei University of Technology, Taipei, Taiwan, ROC

W. C. Hsu
Sino-American Silicon Product Inc., Hsinchu, Taiwan, ROC

INTRODUCTION

Crystal growth is not only a physical process that is interesting to science, but also a technologically important subject in the chemical and material industry. In this entry, we give an introduction to crystal growth physics including the thermodynamics and kinetics for nucleation and growth. Growth mechanisms, surface roughing, and growth inhibitions are also reviewed. Finally, the technologies of crystal growth and the current issues are also discussed.

BACKGROUND

Crystals are solids in which molecules are arranged in a regular, repetitive three-dimensional (3-D) pattern, and crystal growth is a process of building up this structure. In nature, the building units (unit cells) of the crystals are divided into only seven kinds, the so-called seven crystal systems. The molecules can be placed at cell corner, face, or center. Hence, a total of 14 so-called Bravais lattices can be produced. Because of the repetitive nature of the unit cells, crystals can take on characteristic and interesting forms, for example, the colorful and beautiful shine of various gem stones like diamond, ruby, and sapphire. Although the basic structure unit determines the intrinsic properties of a material, the ultimate form and properties of the crystal are determined by the growth process. For example, the formation of beautiful snowflakes can have all kinds of morphologies, and this is controlled by growth conditions. Even in biology, the formation of structures of oyster shells, corals, ivory, and bones is determined by crystal growth, which has evolved to a well-known field called biomineralization.[1,2]

Crystal growth is not only a physical process that is interesting to science, but also a technologically important subject in the chemical and materials industry. Crystals, especially semiconductor single crystals, are important materials in modern technology since the invention of transistors at Bell Laboratories.[3–8] Most of the modern electronic, optical, and optoelectronic devices, as well as integrated circuits and optics, are built upon on the substrate sliced from a single crystal. The miniaturization of the devices and circuits requires large and perfect single crystals, and this has promoted the fast development of crystal growth technology in the past two decades. Crystal production globally is estimated at more than 200,000 tons/yr, of which the largest fraction of about 60% is the semiconductors silicon, GaAs, InP, GaP, and its alloys. In fact, for the solar cells alone, more than 10,000 tons of silicon per year has been used globally. Quartz is the second largest commodity crystal and its production rate is more than 300 tons/yr in the world. Table 1 gives a brief view of the typical single crystals available in the market including their applications and growth methods, which will be discussed later.[3–6] Besides quartz, other oxide crystals such as $LiNbO_3$ and $LiTaO_3$ play an important role in wireless communication, particularly in the surface acoustic wave (SAW) filters.[9] Even for sapphire crystals, their use as substrates for the growth of GaN film used in blue and white light emitting diodes (LEDs) has increased dramatically in recent years.[10,11] Besides the bulk single-crystal growth, it is one of the myths of the thin-film technologies that the crystals are grown layer by layer on an oriented substrate, the so-called epitaxy. The epitaxial growth, especially from the vapor phase, has been routinely used for making LEDs, diode lasers, and quantum-well devices.[12]

Crystal growth starts with the first-order phase transition from an either homogenous or heterogenous nucleation, followed by the construction (growth) of surfaces and morphologies. The growth can be carried out by solidification from a melt, growth from a supersaturated solution, condensation from a vapor phase, or grain growth from solid phases. In terms of thermodynamics, as illustrated in Fig. 1, the crystallized phase

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030516

Table 1 Typical commercial crystals: Their market share, applications, and growth methods

Crystals (% market share)	Examples	Applications	Growth methods
Semiconductor (60%)	Si, SiGe, GaAs, InGaAs, InP	ICs, transistors, solar cells, diodes, etc.	Czochralsiki (Cz), Bridgman one melting
Scintillation crystals (12%)	$Bi_4Ge_3O_{12}$, Lu_2SiO_5, BaF_2, CaF_2, $PbWO_4$, Tl:CsI, Na:CsI, NaI	Radiation detectors, positron emission tomography (PET)	Cz, Bridgman
Optical crystals (10%)	Sapphire, CaF_2, BaF_2, MgF_2, MgO, quartz, Si, Ge, ZnSe, ZnS, $CaCO_3$, YVO_4, $LiNbO_3$, α-BaB_2O_4 (BBO), $Y_3Al_5O_{12}$ (YAG)	Substrates, lenses, mirror, prisms, wave plate, polarizer, etc.	Bridgman, Cz, Kyropoulos
Acoustic crystals (10%)	$LiNbO_3$, $LiTaO_3$, quartz	SAW filter, sensor, etc.	Cz, Bridgman
Nonlinear optics and laser crystals (5%)	YAG, KH_2PO_4 (KDP), $Bi_{12}SiO_{20}$ (BSO), $KTiOAsO_4$ (KTA), β-BBO, LiB_3O_5 (LBO), $KTiOPO_4$ (KTP), $AgGaS_2$, $ZnGeP_2$, Mg:$LiNbO_3$, Nd:YVO_4, Nd:YAG, Nd:$GdVO_4$, Ti:sapphire	Laser applications, optical communication, etc.	Cz, solution growth
Jewelry (3%)	Sapphire, ruby, amber, $MgAl_2O_4$, $CaCO_3$	Decoration, watch window	Cz, Vernuil

should be more stable. In other words, the solution phase has a higher Gibbs free energy because of supersaturation or undercooling. The starting phase can be different for the same crystal. For industrial applications, growth from the melt remains the major method for semiconductors and oxides. If the molten state of the material is not thermodynamically stable or an undesirable phase is expected to be encountered during cooling down, growth from a solution or a vapor phase at a lower temperature is preferred. For a material having high vapor pressure, vapor growth can also be a convenient approach. For the growth of a single crystal, a seed is often required to avoid parasitic nucleation.

Crystal growth is also a convenient way for purification, which is known as recrystallization, because of the different solubility of a solute in the solution and the solid phases. In metallurgy, purification of metals by melting and solidification is a common process. Organic chemists prefer the recrystallization from solution. By repeating the solidification or recrystallization process several times, the purity of the material could be greatly improved.[13] The best example is the zone refining process, a technique of fractional crystallization developed by Pfann.[14] This economical and continuous process has been widely used in industry. Materials with purity up to 6 or 7 N have been routinely produced. Without these high-purity materials, the modern semiconductor industry would not be possible. Furthermore, because of its purification (or concentration) nature, crystal growth or crystallization is essentially a chemical process for the production of a wide range of specialty chemicals and pharmaceuticals. This kind of application is usually referred to as industrial or mass crystallization, which concerns the

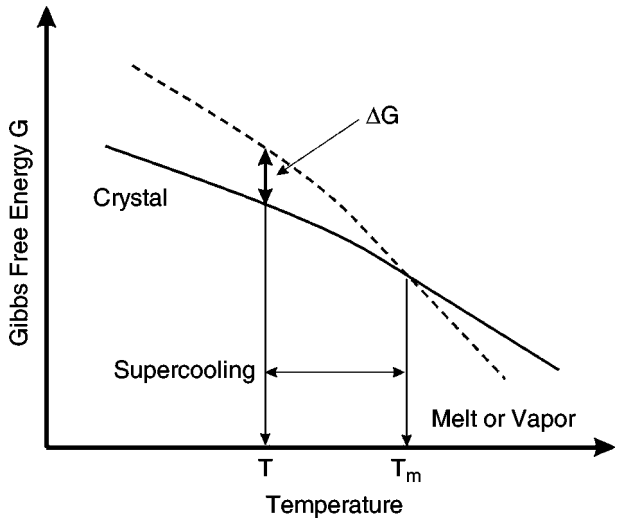


Fig. 1 The Gibbs free energy as a function of temperature for solid and fluid phases.

nucleation and growth of a huge number of small crystals with sizes around 200–1000 micron.[15,16]

Crystal growth is a science of great depth and breadth, and the technology for crystal growth has been developed for a century. Many extensive texts have been written, e.g., Refs.[3,5–8]. Several journals, as well as handbooks, have been published on the subject.[17] Therefore, this entry attempts to give a brief overall view of the subject for beginners. We shall start with some fundamental concepts for crystal growth. Based on the point of view of thermodynamics and kinetics, we will discuss when and how the growth can proceed and what form the crystal can take. The effect of impurities is discussed briefly. Segregation for melt growth is also introduced, followed by the role of dopant mixing. Last, some typical growth methods used in laboratories and industries for growing single crystals are introduced. However, vapor growth, industrial crystallization, and biomineralization are not discussed here.

THERMODYNAMICS AND KINETICS

Driving Force for Nucleation and Crystal Growth

As mentioned for Fig. 1, the first-order transition from a fluid phase to a solid phase requires an excess Gibbs free energy ΔG, in other words, the supersaturation σ or supercooling ΔT, for nucleation and the following growth. The driving force per unit area, f, is proportional to the Gibbs free energy difference per molecule as shown below:

$$f = -\Delta G/V = kT \ln(1 + \sigma)/V \sim (kT/V)\sigma \quad (1)$$

where V is the volume per molecule, k the Boltzmann constant, T the temperature, and σ the dimensionless supersaturation defined as $\sigma = P/P^{\text{sat}} - 1$ for the gas phase or $C/C^{\text{sat}} - 1$ for the solution phase. P is the partial pressure and C the solute concentration. For the growth of a single crystal in applications, the supersaturation is usually kept small to avoid parasitic nucleation. Therefore, $\ln(1 + \sigma)$ approaches σ and the driving force is proportional to supersaturation as shown in Eq. (1). For melt growth,

$$f \sim (\Delta H/V)\sigma$$

where ΔH is the latent heat per molecule and $\sigma = \Delta T/T_{\text{m}}$, the dimensionless supercooling.[18] When a driving force is large enough to overcome the interfacial energy for forming the crystal, nucleation occurs and crystal growth can then continue.

Nucleation

Nucleation is the first step of the first-order phase transition, but a barrier exists in the formation of embryos for the new phase. The interfacial energy is always a positive term to destabilize the nuclei. But once the nuclei grow large enough, the free energy drops rapidly with the new phase formation. The critical size enabling the nucleus to dissolve or to grow is at the local maximum of the Gibbs free energy. A schematic of the Gibbs free energy as a function of the nucleus size is shown in Fig. 2. It should be pointed out that the interfacial energy is closely related to the geometry of the nucleus. With foreign particles or substrates, the interfacial energy between the nucleus and the foreign media can be significantly reduced.[19] This can be interpreted by the larger bonding energy between the nucleus and the foreign molecules than the salvation energy. On the other hand, if a lattice-matched substrate is used, the barrier for forming nucleus is significantly reduced, and this is the reason that the seeded growth can be carried out at a small supercooling or supersaturation. The classic nucleation theories give the critical radius with the following form for a spherical nucleus:[19]

$$r_{\text{c}} = 2V\gamma/\Delta G \quad (2)$$

where γ is the interfacial energy; again, ΔG measures the free energy for phase change. The corresponding nucleation barrier is given as the following:

$$\Delta G_{\text{n}} = (16/3)\pi\gamma^3(V/kT\sigma)^2 \sim \gamma^3/\sigma^2 \quad (3)$$

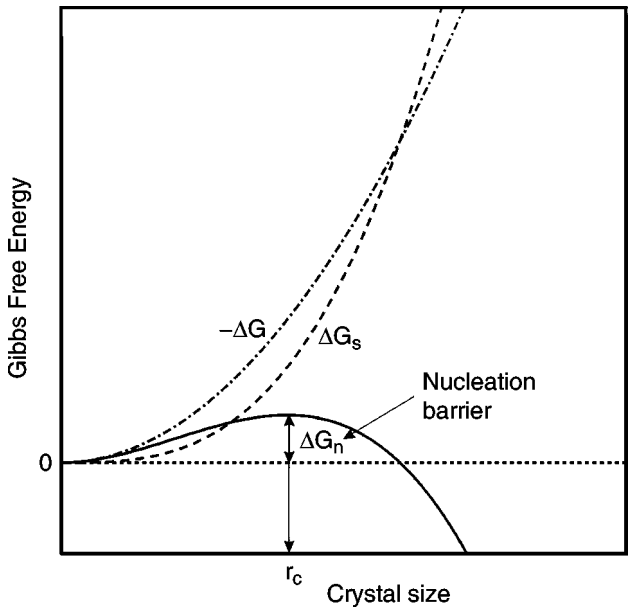


Fig. 2 Gibbs free energy as a function of crystal size during nucleation.

Or, in general $\Delta G_n = B\gamma^3/\sigma^2$, where B is a factor depending on the nucleus geometries. With this energy barrier, the steady-state nucleation rate can be obtained with the form of

$$J_n = A\exp(-\Delta G_n/kT) = A\exp[-B\gamma^3/(kT\sigma^2)] \quad (4)$$

where A is a constant.[19] For 2-D nucleation, Liu et al. have shown that the nucleation barrier is proportional to γ^2/σ.[20] Direct validation of the nucleation theories is difficult because of lack of direct observation and measurements. However, in an experimental model system of a colloidal monolayer, the 2-D theory has been found to be in good agreement with the nucleation data.[21] Further discussion can be found elsewhere.[2,22] The induction period for nucleation that takes the reciprocal of nucleation rate is often used.[15,16] For single-crystal growth, a longer induction period indicates a more stable solution which makes the growth easier.

Growth Kinetics

After nucleation, crystal growth continues from the nucleus surface. When the crystal surface is rough, the growth kinetics is usually simple because the growth sites are randomly distributed. The growth rate is in general proportional to the local driving force. However, in reality, the crystal surface consists of steps, with terraces and kinks, as well as adatoms and vacancies, as illustrated in Fig. 3. The preferred growth sites are the kinks due to more bonding to grasp the adatoms. The growth requires adsorption and then surface diffusion of atoms to the growth sites. The simple step model with surface diffusion can lead to a simple linear law for the growth, i.e., the growth rate being linearly proportional to the supersaturation σ, similar to the one in the rough surface formed by kinks.[18] However, when the surface is smooth at the end of step growth, the kinetics depends on the growth mechanisms. A new 2-D nucleation is a way to start the next layer, as illustrated by Fig. 4A. However, this often requires a substantial supersaturation (10–30%), which is less common. An outlook of an as-grown 4-*N*,*N*-dimethylamino-4′-*N*-methyl-stilbazolium tosylate (DAST) crystal from methanol shows several nucleated crystals on the surface under large supersaturation (10%). Very often the growth can still continue even with only 1% or less supersaturation, and this is possible through the spiral growth of a screw dislocation.[23]

Based on the spiral growth, Burton, Cabrera, and Frank proposed a model (referred to as the BCF theory) considering the transport of solute molecules from the bulk solution to the surface, and then surface diffusion to the kinks of the spiral generated steps from a screw dislocation.[23] A schematic of the spiral generated step is shown in Fig. 4B. Such a mechanism requires a much smaller driving force than the 2-D nucleation and is a favorable growth mode in nature. The BCF theory also predicts a quadratic kinetics at low supersaturation. The BCF theory has been quite successful in comparison with experimental observation and measurements (detailed discussions can be found in, e.g., Refs.[24,25].). For melt growth, the growth

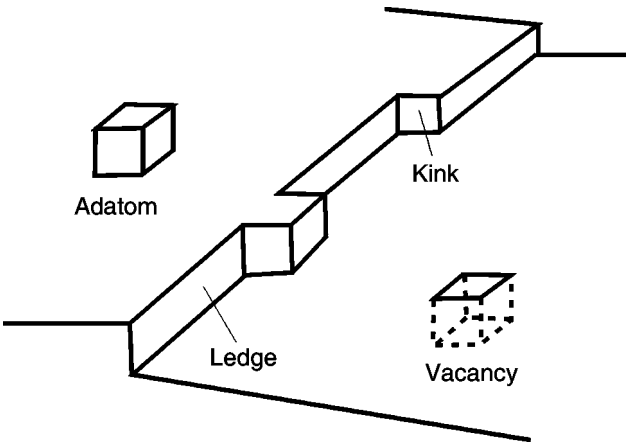


Fig. 3 A simple surface model on the crystal surface.

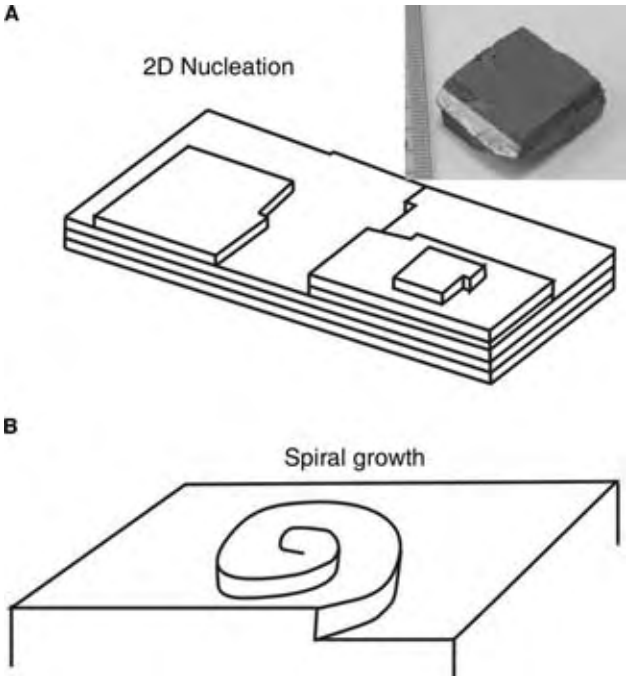


Fig. 4 Growth mechanisms: (A) surface nucleation and (B) BCF spiral growth; an as-grown DAST crystal showing the 2-D nucleation is also illustrated.

rate based on screw dislocations is also quadratic, and confirmed by many examples. The typical growth curves based on different mechanisms are illustrated in Fig. 5, where the growth on the kinks or a rough surface is referred to as the adhesive growth. The consideration of multiple screws can lead to a more complicated kinetic behavior. Nevertheless, the BCF theory lays the foundation for crystal growth theory, and has a strong impact in metallurgy, materials science, and the semiconductor industry.

Moreover, the steps and dislocations appear randomly on the crystal surface, their growth rates (flow) in a particular direction do not occur at the same speed. In some cases, the step growth is inhibited by impurities. Then, step bunching can occur and lead to macrosteps, as illustrated in Fig. 6, which can be visible up to several millimeters. An SEM image of the macrosteps on the surface of a DAST crystal is also illustrated in Fig. 6. Similarly, growth hillocks can also be generated from the spiral growth because the growth speed near the screw center is faster. Furthermore, the step growth can be affected further by fluid flow leading to the formation or dissipation of the kinematic waves, thus roughening or smoothening the crystal surface.[26,27] An extensive and thorough review of the growth mechanisms has been given by van der Eerden.[28] Recently, the use of atomic force microscopy in the observation of the growth steps has led to significant progress in understanding the growth mechanisms.[29,30] Some computer modeling studies have also been presented based on the various models, and the simulated results give a deeper insight of the growth dynamics, e.g., Refs.[31,32].

In the melt growth, the driving force is usually high so that the equilibrium growth habit is suppressed, and in most cases the interface is rough. Nevertheless, if the growth front is curved, and the thermal gradient is low,

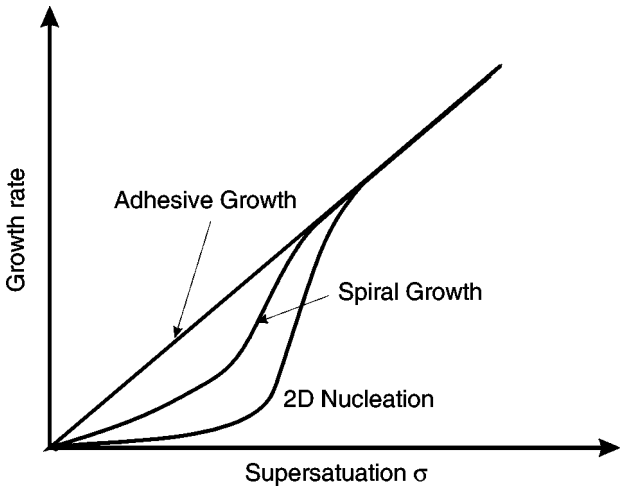


Fig. 5 Growth rate dependence on the supersaturation for various growth mechanisms.

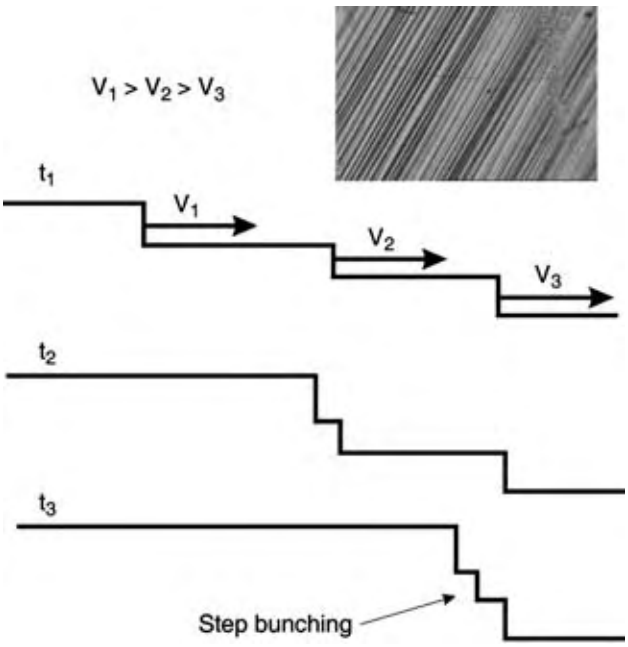


Fig. 6 Step bunching mechanism forming macrosteps; the photograph shows the typical macrosteps (several microns) on the surface of a DAST single crystal.

facets, as the result of step or spiral growth, may appear in some parts of the interface. Because the segregation in the rough and faceted surface is different, faceting can affect significantly the crystal quality and dopant segregation.

Growth Inhibition

The step growth can be inhibited by many factors, especially by impurities and foreign particles. Several mechanisms exist, such as step pinning, incorporation, kink blocking, and surfactant effects.[33] Step pinning, as sketched in Fig. 7 (top), and its effect on the growth rate is also shown schematically. As shown, once the impurities are adsorbed on the surface, the step needs to grow around the pinning points. If the impurity level is low, the average pinning distance is much greater than the radius of curvature, and the step can advance freely. However, with a high impurity level, if the supersaturation is low, the growth can be stopped almost completely forming the so-called dead zone.[34] The dead zone size increases with the impurity concentration. The step spinning can change the face speed, and the impurity pinning can be quite selective to faces depending on the bonding nature. Therefore, the final crystal shape, as a result of face competition, can be changed as well.

Growth inhibition is often observed in biomineralization.[2] Because the incorporation of extraneous ions

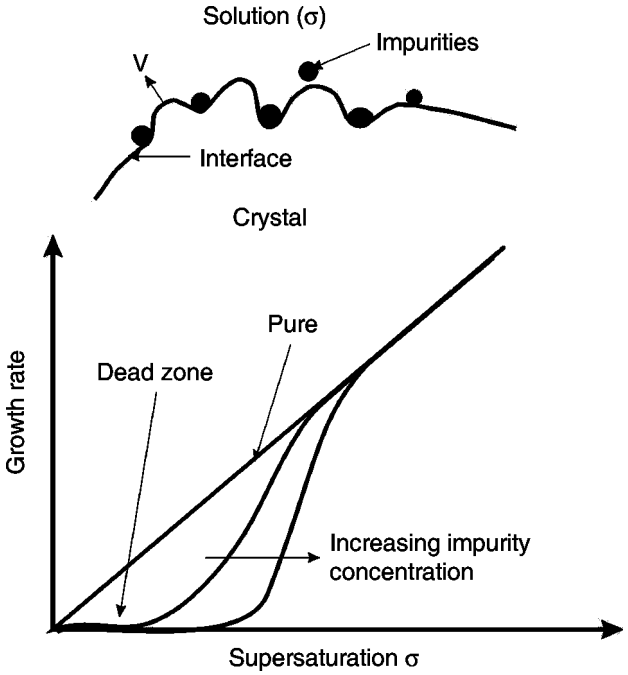


Fig. 7 Step-pinning inhibition of the growth rate due to impurities.

is very selective to the nature and surface binding sites, the growth of some crystallographic planes is inhibited, leading to an extraordinary outlook of the mineral structure. Furthermore, block incorporation of the organic macromolecules can also lead to intercalated structures.

Surface Smoothness

When the growth takes place at low temperature, such as the solution growth of organic crystals, the resultant crystal surface is often smooth. However, at high temperature, this is not always true. In fact, at absolute zero temperature (0 K), a surface at equilibrium should contain no growth steps. At higher temperature, there are adatoms due to thermal energy, and the adatom density ρ_a can be described by the Gibbs formula as:[18,35]

$$\rho_a = \exp(-\Delta G_a/kT) \tag{5}$$

where ΔG_a is the Gibbs free energy needed to extract an atom from a step to make it an adatom. Detailed calculations of the energy involved are required to estimate the roughening transition temperature, which provides a thermodynamic criterion for a smooth surface formation.[18,35] Jackson took a rather different approach to estimate the transition.[36] He formulated an α factor that takes into account the contact nature of the solution and the surface. The α factor is defined as:

$$\alpha = (\Delta H/kT)f_k \tag{6}$$

where ΔH is the latent heat of the phase change and f_k (<1) is a crystallographic factor representing the fraction of all first neighbors lying in a plane parallel to the crystal surface. With a large α factor ($\alpha \gg 2$), the surface is almost smooth because of the strong bonding energy as compared to the thermal fluctuation. This is typical for organic crystals. For most of the metallic materials, α factors are low, so that their surfaces are rough and less prone to faceting. There are also other advanced theories and computational approaches to predict the surface roughness.[28,35] Nevertheless, the Jackson's α factor provides a first estimate of the surface quality.

Growth Morphology

The growth kinetics mentioned above can be different for different crystallographic planes. When the surface is not rough, the facets, or singular surfaces, corresponding to a minimum surface energy determine the final form of the grown crystal. According to the Gibbs's criterion, for a crystal containing facets, the criterion of the crystal shape is given by:

$$\min \sum_i \gamma(n_i)A(n_i) \tag{7}$$

where $A(n_i)$ is the area of a facet of orientation n_i. The geometry features of the crystal shape that minimizes surface free energy were developed by Wulff in 1901 (also see the detailed description by Bennema).[37] His theorem states that, as illustrated in Fig. 8, under equilibrium, there is a point, i.e., the Wulff point, such that the perpendicular distance l_i from any surface

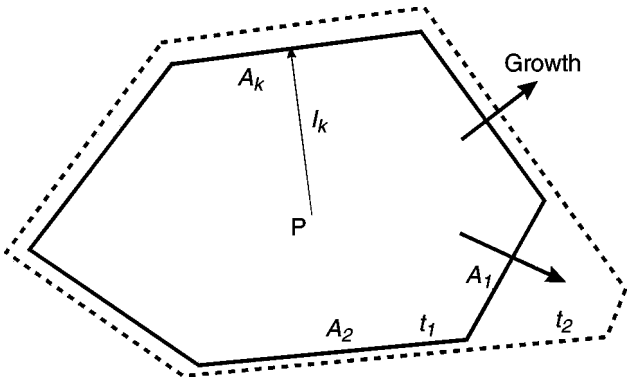


Fig. 8 Equilibrium shape of a crystal and competition of the faces.

tangent plane of orientation n_i to the Wulff point is proportional to γn, such that

$$\frac{\gamma(n_1)}{l_1} = \frac{\gamma(n_2)}{l_2} = \cdots = \frac{\gamma(n_i)}{l_i} \quad (8)$$

Then, the smallest polyhedral bounded by the tangent planes having the smallest volume at $\sum_i l_i A(n)_i/3$; is the equilibrium shape of the crystal. Unfortunately, the nature is not so simple, because crystal growth is not at an equilibrium condition. Because the driving force is inversely proportional to l_i, when the crystal is small, the driving (restoring) force back to the equilibrium is high. Therefore, the equilibrium shape of a small crystal can be maintained. Once the crystal size gets bigger, the restoring force becomes smaller, and the crystal shape is often a result of the competition of the plane growth. As illustrated in Fig. 8 at t_2 ($t_2 > t_1$), the lower growing planes will survive at the final shape. Again, when the growth of some planes is inhibited by impurities or foreign macromolecules, the crystal morphology can be affected significantly. A recent review by Winn and Doherty, though focused on organic solution growth, is a good reference to learn more about the prediction of crystal morphology.[38]

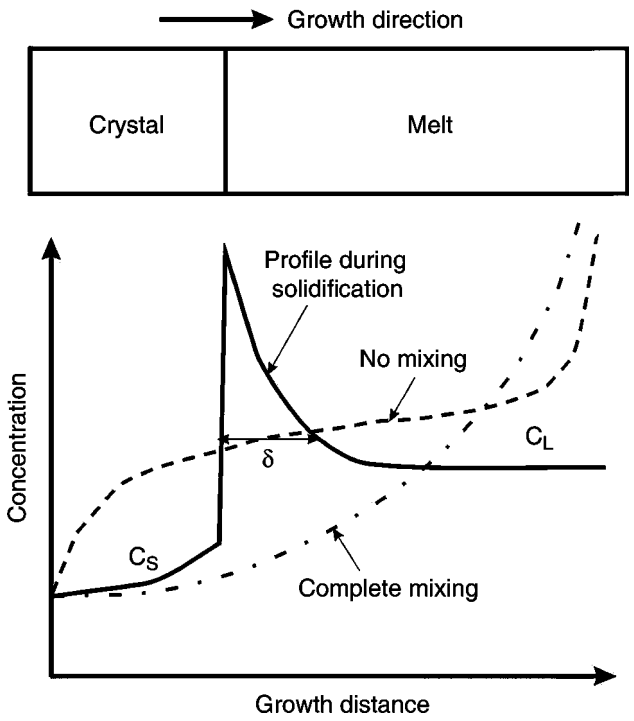


Fig. 9 A schematic sketch of distribution and segregation of impurities during normal freezing. The dashed lines show the final concentration distribution for the cases of complete and no mixings in the melt, respectively.

Segregation and Morphological Instability

In most of the industrial applications, dopants are added into the crystals for tailoring their electrical or optoelectronic properties. However, because of the different dopant solubility in the crystal and in the liquid, dopant segregation (inhomogeneity) during growth is inevitable. This is a problem particularly for a batch growth process. Let us take the normal freezing, as illustrated in Fig. 9, as an example. The crystal growth starts from one end to the other by solidification. With an initial dopant concentration in the melt at C_0 and no solid-state diffusion, the axial dopant segregation with complete dopant mixing (without solid-state diffusion) can be described by the Schiel equation:[39,40]

$$C_s = C_0(1 - f_s)^{k-1} \quad (9)$$

where f_s is the fraction of solidification and k the segregation coefficient. The segregation coefficient k is the ratio of the dopant solubility in the solid and that in the melt. The schematic dopant distribution for $k < 1$ having complete dopant mixing in the solid after solidification is illustrated by the lower dashed line in Fig. 9. However, in practice, the dopant is not completely mixed in the melt. A dopant boundary layer is established in front of the interface during solidification, as illustrated by the solid line in Fig. 9. The dopant concentration profile for the case of no mixing is illustrated by the upper dashed line. The Schiel equation is often used to fit the dopant profile in the crystal by using the segregation coefficient as a parameter, even though, the dopant is not well mixed. The fitting value is often referred to as the effective segregation coefficient k_{eff}. Burton et al. further proposed a simple model (the so-called BPS theory) to correlate the effective segregation coefficient with convection, which is characterized by a boundary layer thickness δ, as:[41]

$$k_{\text{eff}} = \frac{k}{k + (1 - k)e^{-V\delta/D}} \quad (10)$$

where V is the solidification speed and D the dopant diffusivity in the melt. Although the BPS theory is not quite correct for the closed system, it still gives a good physical insight for the dopant segregation under convection. Apparently, from Eq. (10), to improve axial dopant uniformity, one can increase the solidification speed or reduce dopant mixing (having a thicker boundary thickness), so that k_{eff} can be closer to unity. In other words, if the solidification distance is long enough, the axial dopant distribution can be rather uniform. Hence, the reduction of mixing, such as the growth in space, has been an active research topic in crystal growth. Suppressing the flow by magnetic

fields, and thus reducing axial segregation, has also been extensively used in practice.[42]

Furthermore, because of the dopant segregation, the dopant boundary layer affects the solidification temperature (T_L) in front of the growth interface. If the thermal gradient G_T is smaller than the gradient of the solidification temperature at the growth front, constitutional supercooling can occur.[43] This can be described by:

$$G_T \leq m\,dC/dz \tag{11}$$

where m is the slope of the liquidus temperature in the phase diagram. Once the supercooling is large enough to overcome the interfacial energy, the microscopically planar interface can break down into cellular cells or dendrites. Mullins and Sekerka gave an excellent analysis on the morphological instability, which predicts the onset wavelength that forms a steady cellular array.[44]

Despite the segregation in the axial direction, segregation can also occur in the lateral (or radial) direction because of the nonuniform dopant boundary layer. The control of lateral segregation is extremely important in industry to ensure that the wafers cut from the ingot have a good uniformity, especially for large wafers. The interface shape and the convection in the bulk phase are the key factors affecting the lateral segregation. Extensive research efforts through computer modeling have been made to control the transport processes and the interface shape for improving crystal uniformity.[45,46]

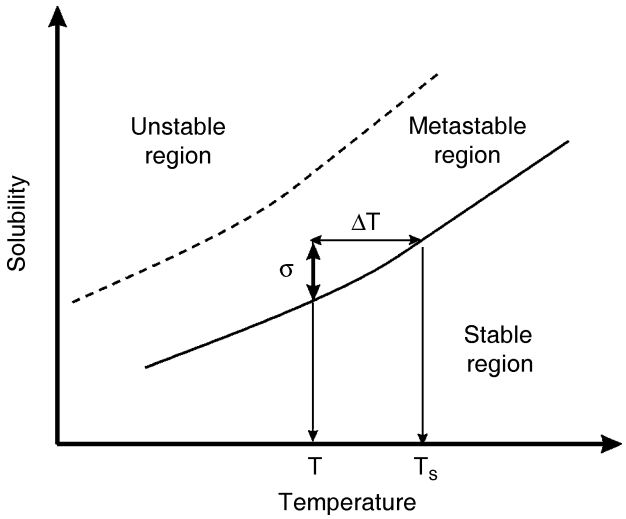


Fig. 10 Stability diagram for solution growth.

CRYSTAL GROWTH METHODS

There are many ways to grow crystals. However, the growth of a bulk single crystal remains a special interest to industry because of the need of single-crystal substrates for device applications. As mentioned previously, a good growth method provides uniform supersaturation or supercooling around the seed crystal, while the other place remains undersaturated or superheated to avoid parasitic nucleation. Some growth techniques have been widely adopted in industry. They can be divided into three categories, namely, the solution growth, melt growth, and vapor growth. Here, we give

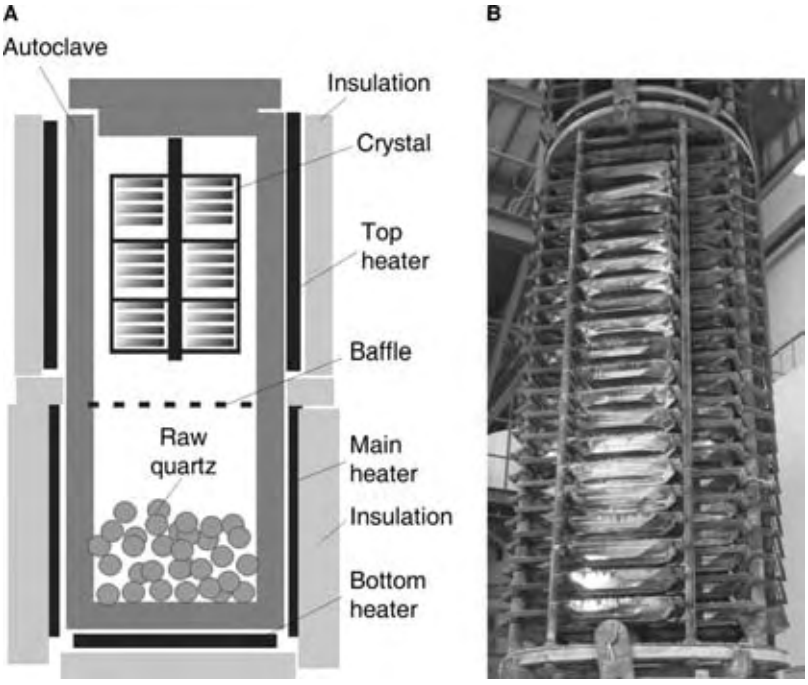


Fig. 11 (A) Schematic sketch of hydrothermal growth of quartz crystals. (B) The crystal basket after growth. (Courtesy of Hantek Inc., Taiwan.) *(View this art in color at www.dekker.com.)*

a brief introduction to the solution growth and melt growth methods through their major applications. The discussion of the growth from the vapor phase can be found elsewhere.[5–7]

Solution Growth

Growth of a bulk crystal from the solution is straightforward. A seed can be selected from the crystallization of a supersaturated solution by slow cooling or slow evaporation of solvent. Then, the growth of a bulk crystal can be carried out from the seed crystal by controlling the solution at the metastable region, as shown in Fig. 10; outside the metastable zone, the growth of crystals from the seeds only is not possible. To maximize the growth speed (production rate), the solubility needs to be as large as possible. This could be achieved by using either a better solvent or a higher temperature or pressure. The growth of quartz in an autoclave is a typical example. Because of its piezoelectric properties and low cost, quartz, next to silicon, is the second most widely used electronic material. More than 300 tons of single crystals have been used each year. Even at high temperature and pressure in an autoclave, to have enough solubility, the natural quartz nutrient, lascas, needs to be dissolved by forming a complex in alkali metal hydroxide or carbonate aqueous solution. A schematic of the hydrothermal growth is illustrated in Fig. 11A. In practice, hundreds of thin plate seed crystals are placed in the growth zone for growth, and the growth takes several months. A crystal basket taken out from the autoclave after growth is shown in Fig. 11B. The inner volume of the autoclave is up to $3.5\,m^3$. The temperature difference between the two zones is the chief process control for the growth rate, and the baffle in between is to reduce the thermal mixing of both zones. Two similar processes, the low and

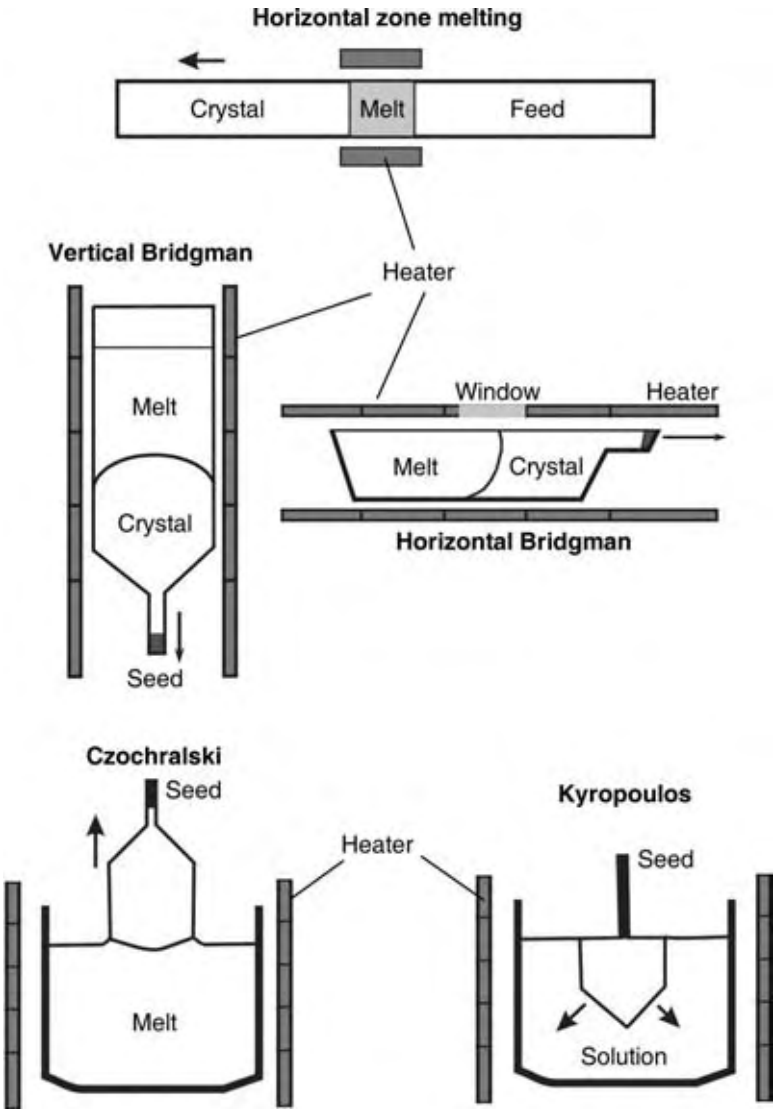


Fig. 12 Major melt growth methods. *(View this art in color at www.dekker.com.)*

high pressure, are widely used in industry.[47] The low-pressure process operated at about 70–100 MPa uses up to 1.0 M sodium carbonate solution, and the temperature at the growth chamber is around 345°C with a temperature difference of 10°C from the nutrient chamber. On the other hand, the high-pressure process operated from 100 to 150 MPa uses up about 1.0 M sodium hydroxide solution and grows the crystal at 360°C. The growth rate is about 1 mm/day.

High-temperature solution growth has also been adopted to grow many kinds of crystals, but often at normal pressure. An extensive introduction is given by Elwell and Scheel.[48] Besides the growth at high temperature, there are a number of important crystals that can be grown from low-temperature solutions. The most well-known one is the growth of potassium dihydrogen phosphate (KDP), an important nonlinear optical crystal for laser applications. Significant progress has been made because of the development of the fast growth process using a careful control of solution purity and particles.[49,50] By removing particles and clusters through reheating and filtration of the circulated solution, the growth rate up to 60 mm/day is possible for KDP.[51]

Melt Growth

The growth of a bulk crystal from the melt is the fastest approach because of the higher growth temperature and the larger driving force. Typical melt growth methods are illustrated in Fig. 12. The zone melting method at the top of the figure is more popular for purification instead of growing crystals. For purification, multiple zones are usually adopted.[14] The Bridgman method, both vertical and horizontal, has several unique advantages as compared with other methods. This method is simple and does not require skillful crystal growers. Hence, a fully automatic operation can be easily achieved. As illustrated in the figure, this technique melts the material in an ampoule and solidifies it directionally from one end to the other by moving the ampoule (or heating profile). Because a very low thermal gradient can be used, this technique is particularly useful for the crystals vulnerable to thermal stresses, such as III–V or II–VI compound semiconductors. Large-size (up to 8 in. diameter) GaAs single crystals with low dislocation density have been grown.[52] Lately, the etched-pit density (EPD) has been reduced to $10^4/cm^2$ for 8 in.-diameter crystals.[53] The horizontal configuration is less popular in applications, because the D-shaped crystals obtained from the horizontal boat are more difficult in postprocessing.

The most popular technique in melt growth, as shown in Fig. 12 (bottom), is the Czochralski (Cz) method; the Kyropulos method is quite similar but the growth is proceeded by slow cooling. Fig. 13A shows a typical industrial puller for the growth of 8 in.-diameter silicon; a photograph of a silicon crystal after growth is also shown in Fig. 13B. The Cz

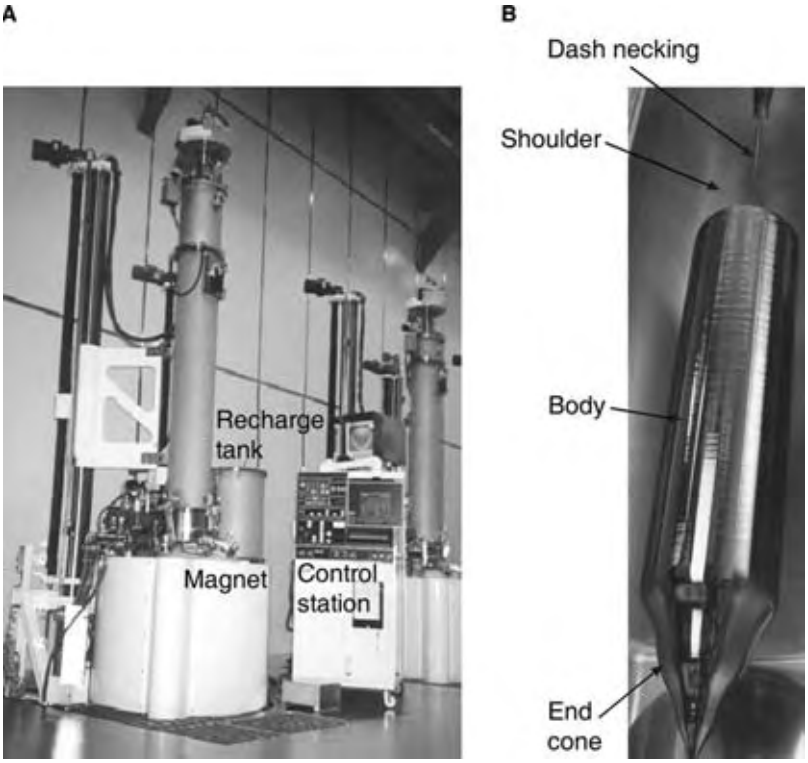


Fig. 13 (A) An outlook of a commercial puller having magnetic fields and recharge tank. (B) An as-grown 8 in.-diameter silicon single crystal. (Courtesy of Taisil Electronic Materials Inc., Taiwan.) *(View this art in color at www.dekker.com.)*

process was first introduced by Czochralski for the kinetic study of crystallization.[54] The development of the technology for semiconductors was credited to Teal and Little.[55] They started using a seed crystal to define the crystal orientation, and introduced the concepts of diameter control and dopant distribution by controlling the temperature and the crystal/crucible rotation. However, a dislocation-free single crystal was not possible until a necking procedure was proposed by Dash in 1958.[56] Before the growth, when the seed crystal is dipped into the melt, the thermal shock can generate defects, especially in the form of dislocations. The Dash necking is to let the seed grow with a small diameter, so that the dislocation lines can extend and end at the surface quickly. After some distance of necking, a dislocation-free crystal is obtained. Then, the shouldering procedure is to reach the full diameter. The body, the part that can be sliced into wafers, is pulled up to a desired length. To finish the growth, the end coning is required before detaching the crystal from the melt surface. Again, this is to minimize the thermal shock during detachment. Over three decades of development, dislocation-free silicon single crystals up to 16 in. in diameter have been grown,[57] while 8–12 in.-diameter silicon crystals have been produced routinely for integrated-circuit applications.

Besides silicon, the Cz method has also been used extensively for compound semiconductors, such as GaAs and InP. Up to 8 in.-diameter GaAs and 6 in.-diameter InP ingots with an excellent quality have been grown, and are now available in the market.[58] Furthermore, many oxide crystals used in solid-state lasers, SAW devices, and nonlinear optics are often grown by this technique as well. Other examples are summarized in Table 1. It should be pointed out that the Czochralski method is a meniscus control process in which the crystal shape is determined by the meniscus. Therefore, the on-line control of the crystal diameter is possible by either monitoring the optical ring (mainly for silicon) of the meniscus or weighting the growing crystal. The control variable can be either the pulling speed or the melt temperature.

CONCLUSIONS

In this article, we give a brief overall view of the crystal growth. The fundamental and application aspects are discussed briefly. In general, crystal growth starts from the nucleation of a thermodynamically stable or metastable phase. Once the driving force overcomes the interfacial energy for forming a new crystal, crystal growth can proceed, but is still limited by the supply of the nutrient through mass transport and the growth kinetics on the crystal surface. The impurities and foreign particles play a crucial role on the growth kinetics and thus the final morphologies. Indeed, how the foreign molecules attach to the crystal surface is a chemistry issue. Therefore, pH, temperature, and solvent are effective parameters to modify the crystal morphologies. By choosing a proper template the metastable form of the crystals could be obtained. In industrial applications, crystal growth is also an important process to purify the materials. Several economical continuous processes have been adopted, such as zone refining, fractional recrystallization, sublimation, etc. Furthermore, the growth of a high-quality bulk crystal is important for many modern technologies. In addition to a proper growth method, high-purity raw materials and a clean growth environment, skillful techniques, and process control are necessary. Because the growth takes time, it always needs to compromise between the quality and the yield. The beautiful appearance of a crystal also requires a perfect structure for applications. The requirement of the crystal quality increases dramatically as the device size diminishes. The stringent specifications for the substrates continue to challenge crystal growers to optimize the process and develop new processes in the future.

REFERENCES

1. Mann, S. *Biomineralization Principle and Concepts in Bioorganic Materials Chemistry*; Oxford University Press: Oxford, 2001.
2. Dove, P.M.; DeYoreo, J.J.; Weiner, S. *Biomineralization*; Mineralogical Society of America: Washington, DC, 2003.
3. Scheel, H.J. Historical introduction. In *Handbook of Crystal Growth 1a: Thermodynamics and Kinetics*; Hurle, D.T.J., Ed.; North-Holland: Amsterdam, 1994; Chapter 1, 1–42.
4. Scheel, H.J. The development of crystal growth technology. In *Crystal Growth Technology*; Scheel, H.J., Fukuda, T., Eds.; John Wiley & Sons: New York, 2003, Chapter 1.
5. Laudise, R.A. *The Growth of Single Crystals*; Prentice-Hall: Englewood Cliffs, NJ, 1970.
6. Brice, J.C. *Crystal Growth Processes*; John Wiley and Sons: New York, 1986.
7. Pamplin, B.A., Ed.; *Crystal Growth*, 2nd Ed.; Pergamon Press Inc.: Elmsford, New York, 1980.
8. Hurle, D.T.J. *Crystal Pulling from the Melt*; Springer-Verlag: Berlin, 1993.
9. Campell, C.K. *Surface Acoustic Wave Devices for Mobile and Wireless Communications*; Academic Press: San Diego, 1998.
10. Talbot, D. LEDs vs. the light bulb. Technol. Rev. **2003**, *106*, 30–36.

11. Zorpette, G. Blue chips. Sci. Am. **2000**, *282*, 30–21.
12. Stringfellow, G.B. *Organometallic Vapor-Phase Epitaxy*, 2nd Ed.; Academic Press: San Diego, 1999.
13. Zief, M.; Wilcox , W.R., Eds.; *Fractional Solidification*; Marcel Dekker Inc.: New York, 1967.
14. Pfann, W.G. *Zone Melting*; John Wiley & Sons Inc.: New York, 1958.
15. Myerson, A.S. *Handbook of Industrial Crystallization*; Butterworth-Heinemann: Boston, 1993.
16. van der Heijden, A.E.D.M.; van Rosmalen, G.M. Industrial mass crystallization. In *Handbook of Crystal Growth 2a: Basic Techniques*; Hurle, D.T.J., Ed.; North-Holland: Amsterdam, 1994; Chapter 7, 315–415.
17. Hurle, D.T.J. *Handbooks of Crystal Growth*; North-Holland: Amsterdam three volumes, 1993.
18. Mutaftschiev, B. Nucleation theory. In *Handbook of Crystal Growth 1a: Thermodynamics and Kinetics*; Hurle, D.T.J., Ed.; North-Holland: Amsterdam, 1993; Chapter 4, 307–475.
19. Markov, I.V. Crystal growth for beginners: fundamentals of nucleation. In *Crystal Growth, and Epitaxy*; World Scientific: Singapore, 1995.
20. Liu, X.Y.; Maiwa, K.; Tsukamoto, K. Heterogeneous two-dimensional nucleation and growth kinetics. J. Chem. Phys. **1997**, *106*, 1870–1879.
21. Zhang, K.-Q.; Liu, X.Y. In situ observation of colloidal monolayer nucleation driven by an alternating electric field. Nature **2004**, *429*, 740–743.
22. Liu, X.Y. Generic mechanism of heterogeneous nucleation and molecular interfacial effects. In *Advances in Crystal Growth Research*; Sato, K., Nakajama, K., Furukawa, Y., Eds.; Elsevier Science: Amsterdam, 2001; 42–61.
23. Burton, W.K.; Cabrera, N.; Frank, F.C. The growth of crystals and the equilibrium structures of their surfaces. Philos. Trans. R. Soc. Lond. A **1951**, *243*, 243–299.
24. Bennema, P; Glimer, G.H. Kinetics of crystal growth. In *Crystal Growth*; Hartmann, P., Ed.; North-Holland: Amsterdam, 1973; 263–327.
25. Chernov, A.A. Formation of crystals in solutions. Contemp. Phys. **1989**, *30*, 251–276.
26. Frank, F.C. *Growth and Perfection of Crystals*; Roberts, B.W., Turnbull, D., Eds.; John Wiley: New York, 1958; 411.
27. Bennema, P.; Glimer, G.H. Kinetics of crystal growth. In *Crystal Growth*; Hartmann, P., Ed.; North-Holland: Amsterdam, 1973; 263–327.
28. van der Eerden, J.P. Crystal growth mechanisms. In *Handbook of Crystal Growth 1a: Thermodynamics and Kinetics*; Hurle, D.T.J., Ed.; North-Holland: Amsterdam, 1993; Chapter 6, 307–475.
29. DeYoreo, J.J.; Orme, C.A.; Land, T.A. Using atomic force microscopy to investigate solution crystal growth. In *Advances in Crystal Growth Research*; Sato, K., Nakajama, K., Furukawa, Y., Eds.; Elsevier Science: New York, 2001; 361–380.
30. DeYoreo, J.J. Eight years of AFM: what has it taught us about solution crystal growth. 13th International Conference on Crystal Growth, Hibiya, T., Mullin, J.B., Uwaha, M., Eds.; Elsevier: Kyoto, Japan, 2001.
31. Vekilov, P.G.; Lin, H.; Rosenberger, F. Unsteady crystal growth due to step-bunching cascading. Phys. Rev. E **1997**, *55*, 3202–3214.
32. Kwon, Y.I.; Derby, J.J. Modeling of the coupled effects of interfacial and bulk phenomena during solution crystal growth. J. Cryst. Growth **2001**, *230*, 328–335.
33. DeYoreo, J.J.; Vekilov, P.G. Principles of crystal nucleation and growth. In *Biomineralization*; Dove, P.M., DeYoreo, J.J., Weiner, S., Eds.; Mineralogical Society of America, 2003; 57–114.
34. Voronkov, V.V.; Rashkovich, L.N. Step kinetics in the presence of mobile adsorbed impurities. J. Cryst. Growth **1994**, *144*, 107–115.
35. Pimpinelli, A.; Villain, J. *Physics of Crystal Growth*; Cambridge University Press: Cambridge, 1998.
36. Jackson, J.A. Mechanism of growth. In *Liquid Metals and Solidification*; American Society for Metals: Cleveland, OH, 1958; 174.
37. Bennema, P. Growth and morphology of crystals: integration of theories of roughening and Hartman–Perdok theory. In *Growth Mechanisms: Handbook of Crystal Growth 1a: Thermodynamics and Kinetics*, Hurle, D.T.J., Ed.; North-Holland: Amsterdam, 1993; 477–581.
38. Winn, D.; Doherty, M.F. Modeling crystal shapes of organic materials grown from solution. AIChE J. **2000**, *46*, 1348–1367.
39. Tiller, W.A.; Jackson, K.A.; Rutter, J.W.; Chalmers, B. The redistribution of solute atoms during the solidification of metals. Acta Metall. **1953**, *1*, 428–437.
40. Flemings, M.C. *Solidification Processing*; McGraw-Hill: New York, 1974.
41. Burton, J.A.; Prim, R.C.; Slichter, W.P. The distribution of solute in crystals grown from the melt. Part I. Theoretical. J. Chem. Phys. **1953**, *21*, 1987–1991.
42. Hurle, D.T.J.; Series, R.W. Use of a magnetic field in melt growth. In *Handbook of Crystal Growth 2a: Basic Techniques*; Hurle, D.T.J., Ed.; North-Holland: Amsterdam, 1994; Chapter 5, 259–285.
43. Rutter, J.W.; Chalmers, B. A prismatic substructure formed during solidification of metals. Can. J. Phys. **1953**, *31*, 15.

44. Mullins, W.W.; Sekerka, R.F. Stability of a planar interface during solidification of a dilute binary alloy. J. Appl. Phys. **1964**, *35*, 444–451.
45. Brown, R.A. Theory of transport processes in single crystal growth from the melt. AIChE J. **1988**, *34*, 881–911.
46. Lan, C.W. Recent progress in crystal growth modeling and growth control. Chem. Eng. Sci. **2004**, *59*, 1437–1457.
47. Hervey, P.R.; Foise, J.W. Synthetic quartz crystal—a review. Miner. Metall. Process. **2001**, *18* (1), 1–4.
48. Elwell, D.; Scheel, H.J. *Crystal Growth from High Temperature Solutions*; Academic Press: London, 1975.
49. Zaitseva, Z.P.; Rashkovich, L.N.; Bogatyreva, S.V. Stability of KH_2PO_4 and $K(H,D)_2PO_4$ in fast crystal growth rates. J. Cryst. Growth **1995**, *148*, 276–292.
50. Zaitseva, N.; Atherton, J.; Rozsa, R.; Carman, L.; Smolsky, I.; Runkel, M.; Ryon, R.; James, L. Design and benefits of continuous filtration in rapid growth of large KDP and DKDP crystals. J. Cryst. Growth **1999**, *197*, 911–920.
51. Yang, S.; Su, G.; Li, Z.; Jiang, R. Rapid growth of KH_2PO_4 crystals in aqueous solution with additives. J. Cryst. Growth **1999**, *197*, 383–397.
52. Rudolph, P.; Jurisch, M. Bulk growth of GaAs: an overview. J. Cryst. Growth **1999**, *198/199*, 325–335.
53. Borner, F.; Bynger, Th.; Eichler, St.; Flade, T.; Hammer, R.; Jurisch, M.; Kretzer, U. 2nd Asian Conference on Crystal Growth and Crystal Technology, 2002; Paper A.20.
54. Czochralski, J. A new method for the measurement of crystallization rate of metals. Z. Phys. Chem. **1918**, *92*, 219–221.
55. Teal, G.K.; Little, J.B. Growth of germanium single crystals. Phys. Rev. **1950**, *78*, 647.
56. Dash, W.C. The growth of silicon crystals free from dislocations. In *Growth and Perfection of Crystals*; Doremus, R.H., Roberts, B.W., David Turnbull, Eds.; John Wiley and Sons, Inc.: New York, 1958; 361–385.
57. Shiraishi, Y.; Takano, K.; Matsubara, J.; Iida, T.; Takase, N.; Machida, N.; Kuramoto, M.; Yamagishi, H. Growth of silicon crystal with a diameter of 400 mm and weight of 400 kg. J. Cryst. Growth **2001**, *229*, 17–21.
58. Fujita, K. Past, present and future of the growth of compound semiconductor crystals. 2nd Asian Conference on Crystal Growth and Crystal Technology, 2002; Paper A.18.

Cumene Production

Robert J. Schmidt
UOP LLC, Des Plaines, Illinois, U.S.A.

INTRODUCTION

The history of the cumene market has been examined in great detail with much discussion regarding product usage, emerging markets, and process economics over the past 10–20 yr.[1] With more than 90% of the world's phenol production technology currently based on the cumene hydroperoxide route, it is the focus of this entry to review the latest technology improvements in cumene processing made over the past 10 yr. Current state-of-the-art processes for the production of cumene as a feedstock for phenol involve zeolitic catalyst technology offerings from UOP, Badger Licensing (formerly ExxonMobil and the Washington Group), and CDTech based on zeolitic catalysis. Much of the improvements in these technologies relate to yield, stability, and operating costs.

CUMENE PRODUCTION

Cumene is produced commercially through the alkylation of benzene with propylene over an acid catalyst. Over the years, many different catalysts have been used for this alkylation reaction, including boron trifluoride, hydrogen fluoride, aluminum chloride, and phosphoric acid. Cumene processes were originally developed between 1939 and 1945 to meet the demand for high-octane aviation gasoline during World War II.[2,3] By 1989, about 95% of cumene demand was used as an intermediate for the production of phenol and acetone. Today, nearly all cumene is used for production of phenol and acetone with only a small percentage being used for the production of α-methylstyrene. The demand for cumene has risen at an average rate of 2–4% per year from 1970 to 2003.[4,5] This trend is expected to continue through at least 2010.

Currently, over 80% of all cumene is produced by using zeolite-based processes. Early processes using zeolite-based catalyst systems were developed in the late 1980s and included Unocal's technology based on a conventional fixed-bed system and CR&L's catalytic distillation system based on an extension of the CR&L MTBE technology.[6–9] At present, the Q-Max™ process offered by UOP and the Badger Cumene Technology developed by ExxonMobil and offered by Badger Licensing represent the state-of-the-art zeolite-based catalyst technologies. A limited number of cumene units remain using the fixed-bed, kieselguhr-supported solid phosphoric acid (SPA) catalyst process developed by UOP and the homogenous $AlCl_3$ and hydrogen chloride catalyst system developed by Monsanto.

Solid Phosphoric Acid Catalyst

Although SPA remains a viable catalyst for cumene synthesis, it has several important limitations: 1) cumene yield is limited to about 95% because of the oligomerization of propylene and the formation of heavy alkylate by-products; 2) the process requires a relatively high benzene/propylene (B/P) molar feed ratio on the order of 7/1 to maintain such a cumene yield; and 3) the catalyst is not regenerable and must be disposed of at the end of each short catalyst cycle. Also, in recent years, producers have been given increasing incentives for better cumene product quality to improve the quality of the phenol, acetone, and especially α-methylstyrene (e.g., cumene requires a low butylbenzene content) produced from the downstream phenol units.

For the UOP SPA catalyst process, propylene feed, fresh benzene feed, and recycle benzene are charged upflow to a fixed-bed reactor, which operates at 3–4 MPa (400–600 psig) and at 200–260°C. The SPA catalyst provides an essentially complete conversion of propylene on a one-pass basis. A typical reactor effluent yield contains 94.8 wt% cumene and 3.1 wt% diisopropylbenzene (DIPB). The remaining 2.1% is primarily heavy aromatics. This high yield of cumene is achieved without transalkylation of DIPB and is a key advantage to the SPA catalyst process. The cumene product is 99.9 wt% pure. The heavy aromatics, which have a research octane number (RON) of about 109, can be either used as high-octane gasoline-blending components or combined with additional benzene and sent to a transalkylation section of the plant where DIPB is converted to cumene. The overall yield of cumene for this process based on benzene and propylene is typically 97–98 wt% if transalkylation is included or 94–96 wt% without transalkylation. Generally, it has been difficult to justify the addition of a transalkylation section to the SPA process based on the relatively low incremental yield improvement that it provides.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120026490

$ALCL_3$ AND HYDROGEN CHLORIDE CATALYST

Historically, $AlCl_3$ processes have been used more extensively for the production of ethylbenzene (EB) than for the production of cumene. In 1976, Monsanto developed an improved cumene process that uses an $AlCl_3$ catalyst, and by the mid-1980s, the technology had been successfully commercialized. The overall yield of cumene for this process can be as high as 99 wt% based on benzene and 98 wt% based on propylene.[10] Detailed process flow information is widely published in the literature for this technology.[11] Dry benzene, both fresh and recycled, and propylene are mixed in an alkylation reaction zone with an $AlCl_3$ and hydrogen chloride catalyst at a temperature of less than 135°C and a pressure of less than 0.4 MPa (50 psig).[11] The effluent from the alkylation zone is combined with recycled polyisopropylbenzene and fed to a transalkylation zone also using $AlCl_3$ catalyst, where polyisopropylbenzenes are transalkylated to cumene. The strongly acidic catalyst is separated from the organic phase by washing the reactor effluent with water and caustic. The distillation section is designed to recover a high-purity cumene product. The unconverted benzene and polyisopropylbenzenes are separated and recycled to the reaction system. Propane in the propylene feed is recovered as liquid petroleum gas (LPG).

ZEOLITE CATALYSTS

In the past decade, beta zeolite (given the universal BEA) has rapidly become the catalyst of choice for commercial production of EB and cumene. Mobil invented the basic beta zeolite composition of matter in 1967.[12] Since that time, catalysts utilizing beta have undergone a series of evolutionary steps leading to the development of the state-of-the-art catalysts such as QZ-2000™ catalyst and QZ-2001™ catalyst for cumene alkylation.

Much of the effort between 1967 and the early 1980s involved characterization of the perplexing structure of beta zeolite. It was quickly recognized that the BEA zeolite structure has a large-pore, three-dimensional structure, and a high acidity capable of catalyzing many reactions. But it was not until early 1988 that scientists at Exxon finally determined the chiral nature of the BEA structure, which is shown in Fig. 1.

While the structure of beta was being investigated, new uses for this zeolite were being discovered. A major breakthrough came in late 1988 when workers at Chevron invented a liquid phase alkylation process using beta zeolite catalyst. Chevron patented the process in 1990.[13] While Chevron had significant commercial experience with the use of Y (FAU) zeolite in liquid phase aromatic alkylation, Chevron quickly recognized the benefits of beta over Y as well and other

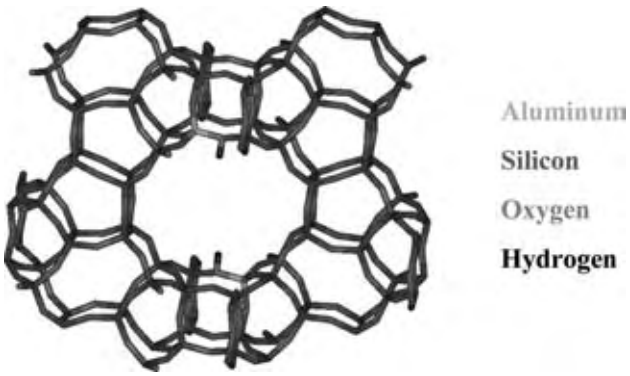


Fig. 1 Beta zeolite. *(View this art in color at www.dekker.com.)*

acidic zeolites used at that time, such as mordenite (MOR) or ZSM-5 (MFI).

Fig. 2 shows a comparison of the main channels of these zeolites. Chevron discovered that the open 12-membered ring structure characteristic of beta zeolite coupled with its high acidity made it an excellent catalyst for aromatic alkylation. These properties were key in the production of alkyl aromatics such as EB and cumene in extremely high yields and with product purities approaching 100%. Moreover, Chevron discovered that the combination of high activity and porous structure imparted a high degree of tolerance to many typical feed contaminants.

From a technical perspective, the process developed by Chevron was a breakthrough technology in that the high cumene yields and purities were not attainable by the other vapor phase or liquid phase processes of the day. Nevertheless, the manufacturing cost of beta zeolite was still too high for catalyst producers to make a commercially viable catalyst. UOP, however, developed new manufacturing technology to make a beta zeolite based catalyst a commercial reality. In 1991 a new cost-effective synthesis route was invented by Cannan and Hinchey at UOP.[14] The new synthesis route patented the substitution of diethanolamine, a much less expensive templating agent, for a substantial fraction of tetraethylammonium hydroxide, which had been used in the synthesis previously. Moreover, the route further enables the use of tetraethylammonium bromide (instead of the hydroxide) as an additional cost saving approach. Finally, the new synthesis route allows the practical synthesis of beta over a wider range of silica to alumina ratios, a factor that has a profound effect on the catalyst's performance.

Subsequently, UOP sought to develop zeolitic catalysts that would overcome the limitations of SPA including a catalyst that is regenerable, produces higher cumene yield, and decreases the cumene cost of production. More than 100 different catalyst materials were screened, including mordenites, MFIs, Y-zeolites,

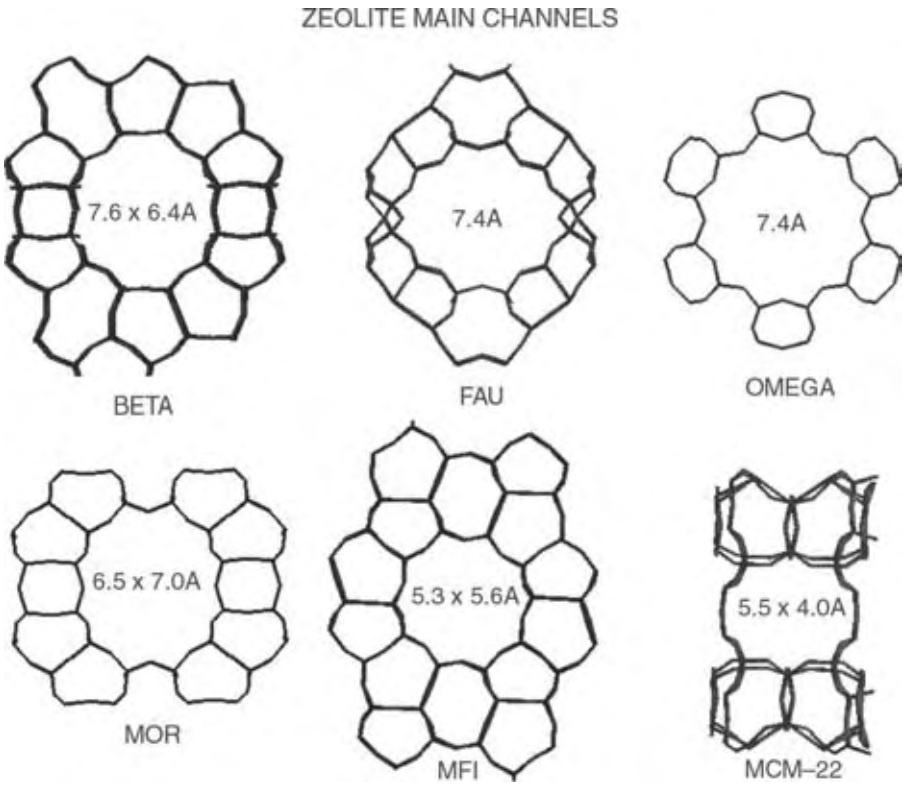


Fig. 2 Comparison of various zeolites. *(View this art in color at www.dekker.com.)*

amorphous silica–aluminas, and beta zeolite. The most promising materials were modified to improve their selectivity and then subjected to more rigorous testing.

On the process side, Unocal developed an early liquid phase fixed-bed reactor system based on a Y-type zeolite catalyst in the 1980s.[15] The higher yields associated with the liquid phase based process were quickly recognized and adapted by the industry with the selectivity to cumene generally falling between 70 and 90 wt% based on converted benzene and propylene dependent on operating conditions. Major side products in the process are primarily polyisopropylbenzenes, which are transalkylated to cumene in a separate reaction zone to give an overall yield of cumene of about 99 wt%. The distillation requirements involve the separation of propane for LPG use, the recycle of excess benzene to the reaction zones, the separation of polyisopropylbenzene for transalkylation to cumene, and the production of a purified cumene product.

By 1992, UOP had selected the most promising catalyst, based on beta zeolite, for cumene production and then began to optimize a liquid phase based process design around this new catalyst. The result of this work led to the commercialization of the UOP Q-Max process and the QZ-2000 catalyst in 1996. More recently in 2001, UOP commercialized a new alkylation catalyst, QZ-2001, which offers improved stability and operation as low as 2 B/P molar feed ratio. The low B/P feed ratio (2) represents the lowest in the industry and affords cumene producers the option to expand capacity and/or revamp existing fractionation equipment with significant cost savings.

The Q-Max process flow scheme based on a liquid phase process is shown in Fig. 3. The alkylation reactor is divided into four catalyst beds contained in a single reactor vessel. Fresh benzene feed is routed through the upper-mid section of the depropanizer column to remove excess water that may be present in the fresh benzene feed. Relatively dry benzene is withdrawn from the depropanizer for routing to the alkylation reactor. Recycle benzene to the alkylation and transalkylation reactors is recovered as a sidedraw from the benzene column. A mixture of fresh and recycle benzene is charged downflow into the alkylation reactor. Propylene feed is divided into portions and injected into the alkylation reactor between the catalyst beds and each portion is essentially completely consumed in each bed. An excess of benzene is used in the alkylation reactor to avoid polyalkylation and to help minimize olefin oligomerization. The alkylation reaction is highly exothermic and the temperature rise in the reactor is controlled by recycling a portion of the alkylation reactor effluent to the reactor inlet to act as a heat sink. In addition, the inlet temperature of each downstream bed is reduced to the same temperature as the first bed inlet by injecting a portion of cooled reactor effluent between beds.

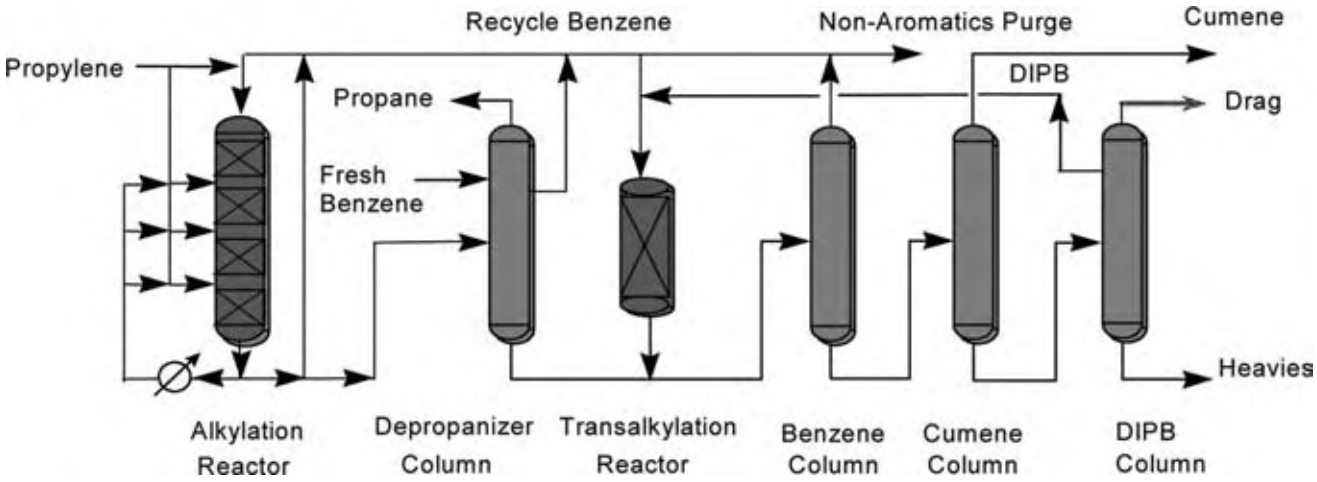


Fig. 3 Q-Max process. *(View this art in color at www.dekker.com.)*

Effluent from the alkylation reactor flows to the depropanizer column, which removes any propane that may have entered with the propylene feed along with excess water that may have entered with the fresh benzene feed. The bottom stream of the depropanizer column goes to the benzene column where excess benzene is collected overhead and recycled. Any trace nonaromatics that may have been in the fresh benzene feed can be purged from the benzene column to avoid an unacceptable accumulation of nonaromatic species in the benzene recycle stream. The benzene column bottom stream goes to the cumene column where the cumene product is recovered overhead. The cumene column bottoms stream, which contains predominantly DIPB, goes to the DIPB column. If the propylene feed contains an excessive amount of butylenes, or if the benzene feed contains an excessive amount of toluene, higher levels of butylbenzenes and/or cymenes can be formed in the alkylation reactor. These compounds are distilled out and purged from the overhead section of the DIPB column. The DIPB stream leaves the DIPB column by a sidedraw and is passed to the transalkylation reactor. The DIPB column bottom stream consists of heavy aromatic by-products, which are normally blended into fuel oil. Steam or hot oil typically provides the heat requirements of the product fractionation section.

The sidedraw from the DIPB column containing mainly DIPB combines with a portion of the recycle benzene and is charged downflow into the transalkylation reactor. In the transalkylation reactor, DIPB and benzene are converted to additional cumene. The effluent from the transalkylation reactor is then sent to the benzene column.

QZ-2000 or the newer QZ-2001 catalyst can be utilized in the alkylation reactor while QZ-2000 catalyst remains the catalyst of choice for the transalkylation reactor. The expected catalyst cycle length is 2–4 yr, and the catalyst should not need replacement for at least three cycles with proper care. At the end of each cycle, the catalyst is typically regenerated ex situ via a simple carbon burn by a certified regeneration contractor. However, the unit can also be designed for in situ regeneration. Mild operating conditions and a corrosion-free process environment permit the use of carbon–steel construction and conventional process equipment.

An alternative zeolite process was developed by CR&L and licensed by CDTech and is based on the concept of catalytic distillation, which is a combination of catalytic reaction and distillation in a single column.[6–9] Catalytic distillation uses the heat of reaction directly to supply heat for distillation of the reaction mixture. This concept has been applied commercially for producing not only cumene but also EB and methyl *tert*-butyl ether (MTBE). The use of a single column that performs both the reaction and the distillation functions has the potential of realizing substantial savings in capital cost by essentially eliminating the need for a separate reactor section. Unfortunately, many available zeolite catalysts that are ordinarily very effective in promoting alkylation in a fixed-bed environment are much less effective when used in the environment of the catalytic distillation column. Also, a separate fixed-bed finishing reactor may be required to ensure that complete conversion (100%) of the olefin occurs to avoid yield losses of propylene to the LPG product stream. As such, the amount of catalyst and physical size of the distillation column may be substantially larger than the benzene column used in the conventional fixed-bed process. Thus, the savings realized by the elimination of the reactor section may be more than offset by the increased catalytic distillation column size, catalyst cost, and addition of a finishing reactor. Depending on the relative values

and trade-off of these considerations including utility values, catalytic distillation may still be an appropriate option for producers in certain circumstances, although market interest in this process option for cumene production has been low in recent years.

Zeolitic Alkylation Chemistry

The production of cumene proceeds by a modified Friedel–Crafts alkylation of benzene with propylene. This reaction can be promoted with varying degrees of effectiveness using many different acid catalysts. The basic alkylation chemistry and reaction mechanism are shown in Fig. 4. The olefin forms a carbonium ion intermediate, which attacks the benzene ring in an electrophilic substitution. The addition of propylene to the benzene ring is at the middle carbon of C3 olefin double bond, in accordance with Markovnikov's rule. The presence of the isopropyl group on the benzene ring weakly activates the ring toward further alkylation, producing DIPB and heavier polyalkylate by-products.

Because new high-activity beta zeolite catalysts such as QZ-2000 catalyst are such strong acids, they can be used at lower reaction temperatures than SPA catalyst or other relatively lower-activity zeolites such as MCM-22 catalyst.[16,17] The lower reaction temperature in turn reduces the olefin oligomerization reaction rate, which is relatively high for SPA catalyst. The result is that beta zeolite catalysts tend to have higher selectivity to cumene and lower selectivity to both nonaromatics that distill with cumene (such as olefins, which are analyzed as Bromine Index, and saturates) and heavy by-products. For example, although butylbenzene is typically produced from traces of butylene in the propylene feed, there is always the potential also for butylbenzene to form through the oligomerization of propylene to nonene, followed by cracking and alkylation to produce butylbenzenes and amylbenzenes. As a result of having relatively high activity and operating at relatively low temperature, beta zeolite catalyst systems tend to eliminate oligomerization. This results in essentially no butylbenzene formation other than that formed from the butylenes in the propylene feed. The cumene product from a beta zeolite based process such as the Q-Max process unit fed with a butylene-free propylene feedstock typically contains less than 15 wt ppm butylbenzenes.

The Q-Max process typically produces near-equilibrium levels of cumene (between 85 and 95 mol%) and DIPB (between 5 and 15 mol%). The DIPB is fractionated from the cumene and reacted with recycle benzene at transalkylation conditions to produce additional cumene. The transalkylation reaction is believed to occur by the acid catalyzed transfer of one isopropyl group from DIPB to a benzene molecule to form two molecules of cumene, as shown in Fig. 5

Beta zeolite catalyst is also an extremely effective catalyst for the transalkylation of DIPB to produce cumene. Because of the high activity of beta zeolite, transalkylation promoted by beta zeolite can take place at very low temperature to achieve high conversion and minimum side products such as heavy aromatics and additional *n*-propylbenzene as highlighted in Fig. 6. Virtually no tri-isopropyl benzene is produced in the beta system owing to the shape selectivity of the three-dimensional beta zeolite structure, which inhibits compounds heavier than DIPB from forming.

As a result of the high activity and selectivity properties of beta zeolite, a beta zeolite based catalyst

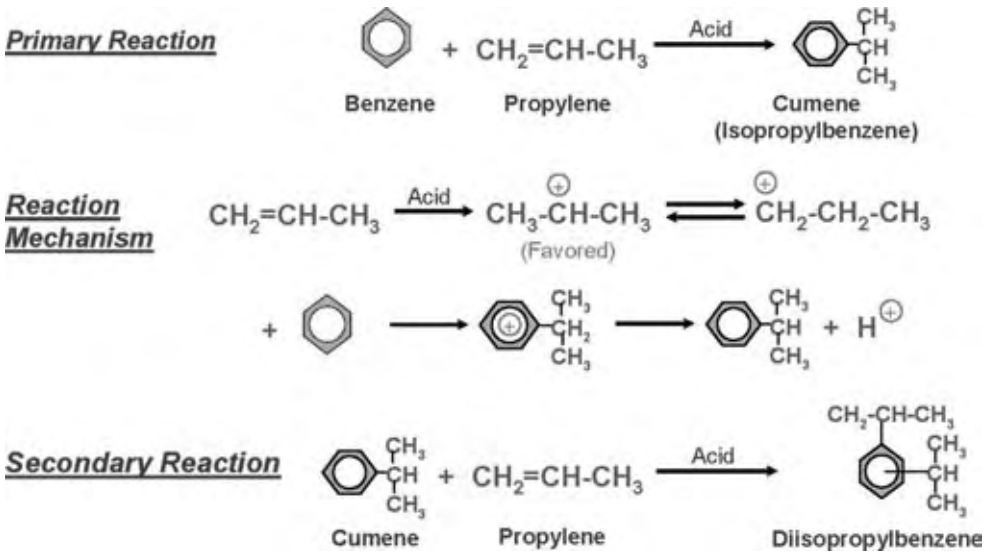


Fig. 4 Alkylation chemistry. *(View this art in color at www.dekker.com.)*

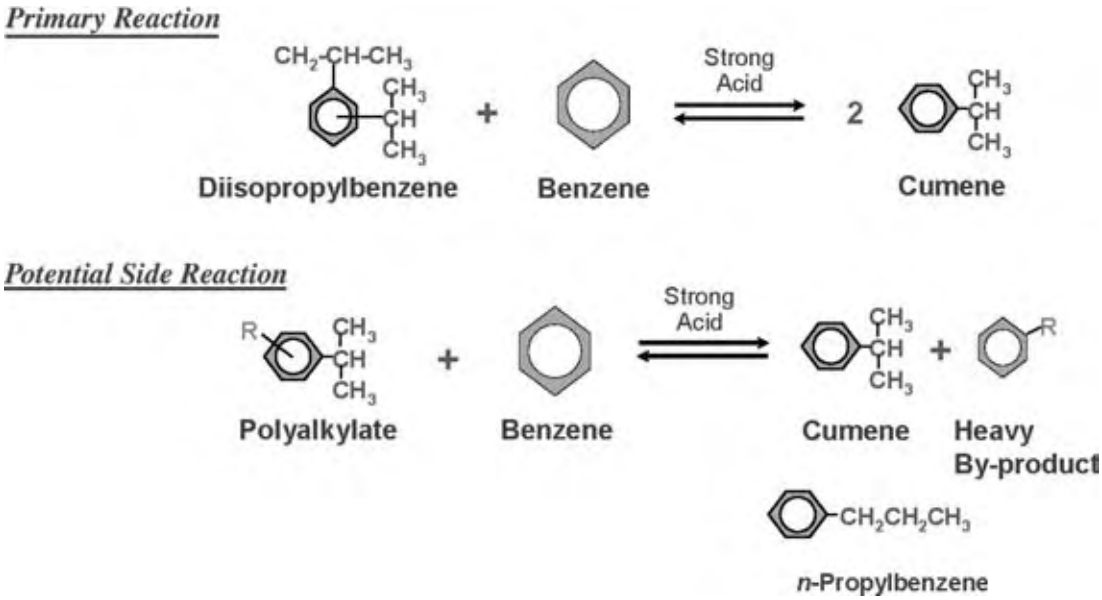


Fig. 5 Transalkylation chemistry. *(View this art in color at www.dekker.com.)*

(e.g., QZ-2001 catalyst or QZ-2000 catalyst) is specified for both the alkylation and the transalkylation sections of the Q-Max process.

With both alkylation and transalkylation reactors working together to take full advantage of the QZ-2001/QZ-2000 catalyst system, the overall yield of cumene based on benzene and propylene feed in the Q-Max process can be at least 99.7 wt% or higher. Because the Q-Max process uses small, fixed-bed reactors and carbon–steel construction, the erected cost is relatively low. Also, because the QZ-2001/QZ-2000 catalyst system is more tolerant of feedstock impurities (such as water, *p*-dioxane, sulfur, etc.) compared to other catalysts available, the Q-Max process requires

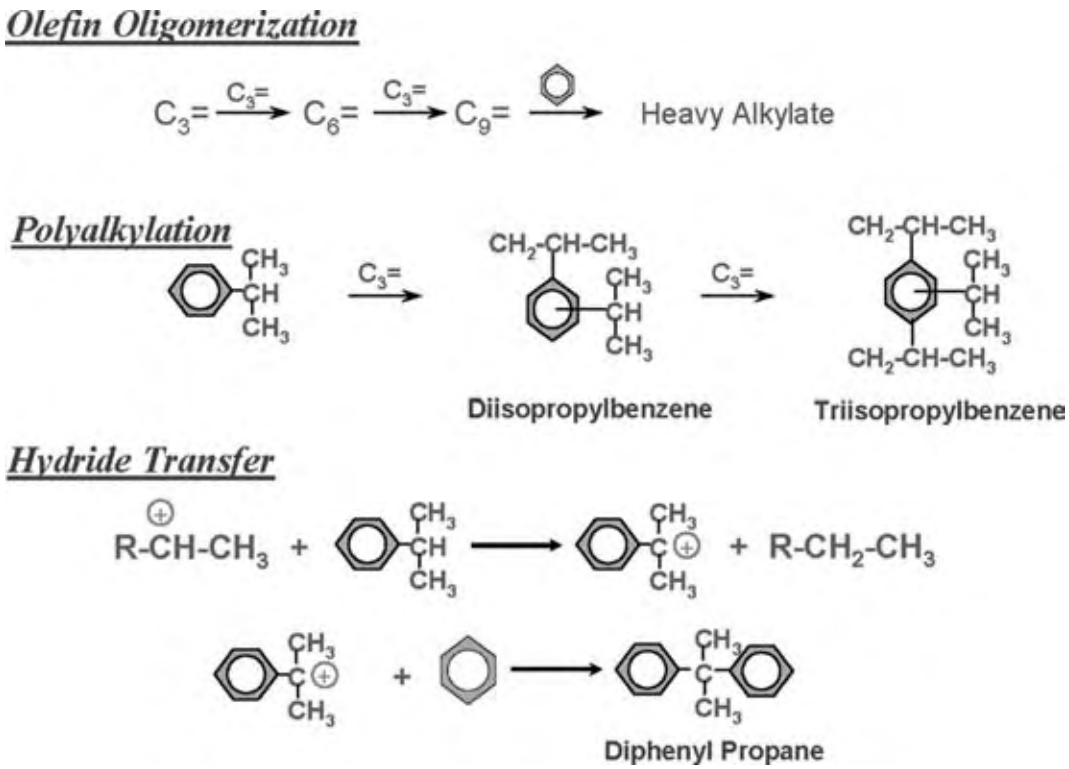


Fig. 6 Possible alkylation side reactions. *(View this art in color at www.dekker.com.)*

minimal pretreatment of the feeds, which further minimizes the capital costs.

This is in distinct contrast to other technologies based on zeolites other than beta where extensive feed contaminant guard beds are required to protect the catalyst from rapid and precipitous deactivation and loss of conversion when exposed to trace amounts of sulfur, water, oxygenates, and nitrogen.

Cumene Product Impurities

Beta zeolite catalyst can be optimized to nearly eliminate all undesirable side reactions in the production of cumene. The improvement in beta zeolite catalyst quality has occurred to the point that any significant impurities in the cumene product are governed largely by trace impurities in the feeds. The selectivity of the catalyst typically reduces by-products to a level resulting in production of ultrahigh cumene product purities up to 99.97 wt%. At this level, the only significant by-product is *n*-propylbenzene with the catalyst producing essentially no EB, butylbenzene, or cymene beyond precursors in the feed. Fig. 7 shows the reactions of some common feedstock impurities that produce these cumene impurities.

Beta Catalyst Resistance to Feed Contaminants

Beta zeolite is a large pore zeolite with optimized acid site densities that exhibits:

- Good mass transfer properties.
- Significant reduction in undesirable polymerization and by-product side reactions.
- High tolerance to feedstock impurities and poisons.

The resistance of new zeolitic catalysts to temporary and permanent catalyst poisons is essential to the economic and commercial success of a zeolitic based cumene process. The following commercial data obtained using beta zeolite as a catalyst illustrates the outstanding ability of beta zeolite to cope with a wide range of feedstock contaminants:

Cyclopropane

n-Propylbenzene (nPB) is formed by alkylation of benzene with cyclopropane or *n*-propanol, and by anti-Markovnikov alkylation of benzene with propylene. Cyclopropane is a common impurity in propylene feed and approximately half of this species is converted to nPB in the alkylation reactor. Essentially, all alkylation catalysts produce some nPB by anti-Markovnikov alkylation of propylene. The tendency to form nPB rather than cumene decreases as the reaction temperature is lowered, making it possible to compensate for cyclopropane in the feed to some extent. As the operating temperature of zeolitic alkylation catalyst is decreased, the deactivation rate increases. However, because of the exceptional stability of the beta zeolite catalyst system, a unit operating with beta zeolite catalyst can be operated for extended cycle lengths and still maintain an acceptable level of nPB in the cumene product. For example, with FCC-grade propylene feed containing typical amounts of cyclopropane,

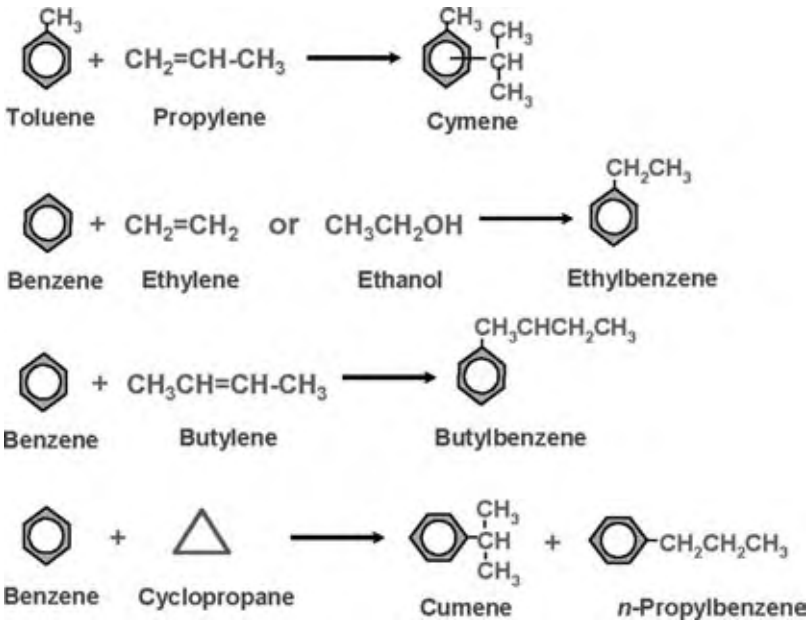


Fig. 7 Reactions of feed impurities. *(View this art in color at www.dekker.com.)*

a beta zeolite based process can produce an overall cumene product containing 250–300 wt ppm nPB while achieving an acceptable catalyst cycle length.

Water

Water can act in this environment as a Bronsted base to neutralize some of the weaker zeolite acid sites. This effect is not harmful to any appreciable extent to the beta zeolite catalyst at typical feed stock moisture levels and under normal alkylation and transalkylation conditions. This includes processing of feedstocks up to the normal water saturation condition (typically 500–1000 ppm) resulting in 10–150 ppm water in the feed to the alkylation reactor dependent on feed and/or recycle stream fractionation efficiency.

Oxygenates

Small quantities of methanol and ethanol are sometimes added to the C_3s in pipelines to protect against freezing because of hydrate formation. Although the beta zeolite catalyst is tolerant of these alcohols, removing them from the feed by a water wash may still be desirable to achieve the lowest possible levels of EB or cymene in the cumene product. Cymene is formed by the alkylation of toluene with propylene. The toluene may already be present as an impurity in the benzene feed, or it may be formed in the alkylation reactor from methanol and benzene. Ethylbenzene is primarily formed from ethylene impurities in the propylene feed. However, similar to cymene, EB can also be formed from ethanol.

p-Dioxane is sometimes present in benzene from extraction units that use ethylene glycol based solvents. It is reported to cause deactivation in some zeolitic alkylation catalysts even at very low ppm levels. However, beta zeolite catalyst appears to be tolerant to *p*-dioxane at levels typically found in benzene extraction processes and does not require costly removal of this impurity.

Sulfur

Sulfur has no significant effect on beta zeolite catalyst at the levels normally present in olefin and benzene feeds considered for cumene production. However, even though the beta zeolite catalyst is sulfur tolerant, trace sulfur that makes its way into the finished cumene unit product may be a feed quality concern for downstream phenol processors where the typical sulfur specification is <1 ppm. The majority of sulfur compounds associated with propylene (mercaptans) and those associated with benzene (thiophenes) are converted to by-products outside the boiling range of cumene. Because some sulfur compounds form by-products that boil within the boiling range of cumene, the sulfur content of the cumene product depends on the sulfur content of the propylene and especially benzene feeds to a certain extent. Sulfur at the levels usually present in propylene and benzene feeds considered for cumene production will normally result in cumene product sulfur content that is within specifications.

Unsaturates

Use of beta zeolite catalyst does not require the benzene feed to be clay treated prior to use in alkylation service. Some of the unsaturated material in the benzene can lead to the formation in the alkylation reactors of polycyclic-aromatic material which will get preferentially trapped in some zeolites having relatively small-sized pores. This can lead to increased deactivation rates in such small-pore zeolites. Beta zeolite's large pore structure makes it possible to more easily handle this polycyclic-aromatic material and as a result does not require further treatment of the benzene feed to remove unsaturated material. In addition, alpha-methylstyrene (AMS) is produced by alkylation of benzene with methylacetylene or propadiene. Some of the AMS alkylates with benzene, forming diphenylpropane, a heavy aromatic that leaves the unit with the DIPB column bottoms.

Nitrogen, Metal Cations, and Arsine

The presence of trace amounts of organic nitrogen compounds and metal cations in the benzene feed or arsine in the olefin feed has been known to neutralize the acid sites of any zeolite catalyst. Good feedstock treating practice or proven guard-bed technology easily handles these potential poisons. For example, basic nitrogen, which can sometimes be present in the benzene fresh feed, is easily removed using very low-cost UOP guard-bed technology. To facilitate monitoring of feeds for potential nitrogen based contaminants, UOP has developed improved analytical techniques to help in the evaluation. These methods include UOP 971 (''Trace Nitrogen in Light Aromatic Hydrocarbons'' by Chemiluminescence) used to detect total nitrogen down to 30 ppb and UOP 974 (''Nitrogen Compounds in Light Aromatic Hydrocarbons'' by GC) used to detect individual nitrogen species down to about 100 ppb.

CURRENT STATE-OF-THE-ART CUMENE TECHNOLOGY

Currently, the major processes for cumene production are liquid phase technologies offered by UOP and

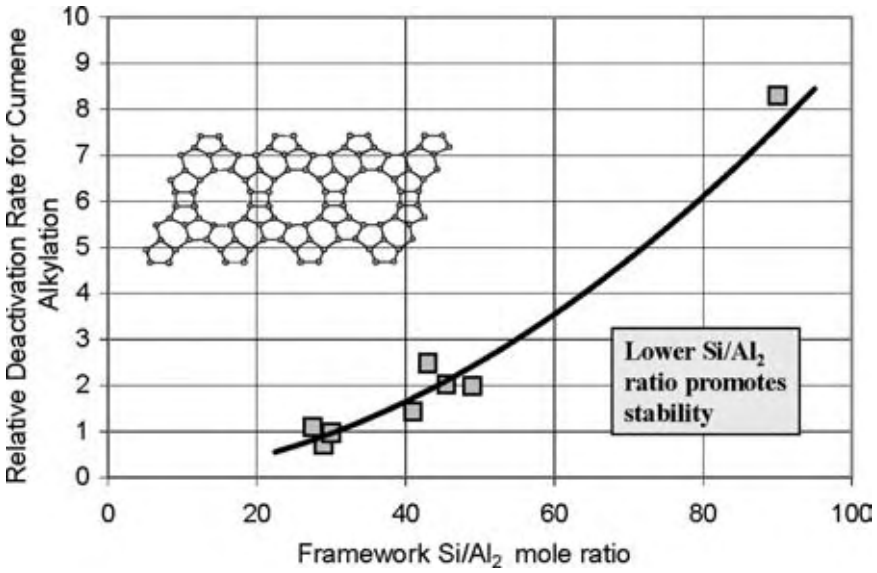


Fig. 8 Effect of framework Al on beta catalyst stability. *(View this art in color at www.dekker.com.)*

Badger Licensing (ExxonMobil technology) based on beta zeolite (such as QZ-2001 and QZ-2000 catalysts) and MCM-22 catalyst, respectively. Over the past decade, great progress has been made in improving and optimizing catalyst formulations for use in applications to produce cumene from benzene alkylation. For example, the ability to synthesize beta zeolite in a wide range of Si/Al_2 ratios has given catalyst designers the ability to tailor the zeolite into a form that optimizes activity and selectivity. A parametric study on the effects of Si/Al_2 ratio on activity and selectivity was published by Bellusi.[18] In this work, it was found that as the silica to alumina ratio was increased from 28 to 70, there was a decrease in both activity and selectivity toward isopropylbenzene-type compounds. Additionally, the less active catalysts had a greater tendency toward oligomerization and were more prone toward coking.

This study parallels work performed at UOP, where, through the use of nonconventional synthesis techniques, samples have also been prepared with Si/Al_2 ratios as low as 10. Through this work it has been found that with a Si/Al_2 ratio of 25, the catalyst maintains sufficient activity to achieve polyalkylate equilibrium (e.g., DIPB equilibrium) and, at the same time, minimizes the formation of heavier diphenyl compounds (and hence maximizes yield) in cumene service.

Perhaps the most critical understanding was developed with regard to the need to minimize the Lewis acidity of the catalyst and at the same time maintain high Brønsted acidity. Studies at UOP demonstrated that olefin oligomerization was directly related to the

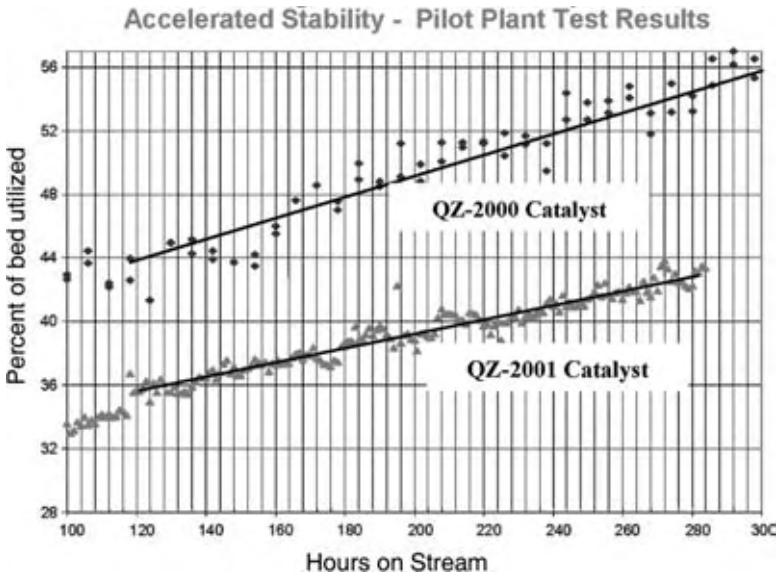


Fig. 9 Stability of QZ-2001 catalyst vs. QZ-2000 catalyst. *(View this art in color at www.dekker.com.)*

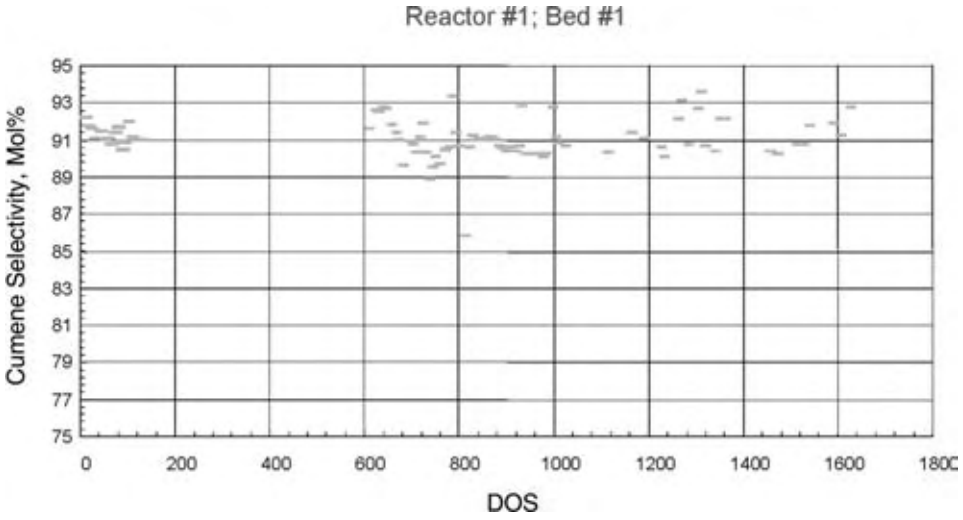


Fig. 10 JLM commercial data on catalyst stability. *(View this art in color at www.dekker.com.)*

Lewis acid function of the catalyst. Olefin oligomerization reactions can lead to the formation of heavy compounds (coke type precursors), which have a negative effect on catalyst stability. Thus, minimization of the Lewis character of the beta leads to a catalyst with high stability. Generally, Lewis acidity in beta zeolite has been attributed to the existence of nonframework aluminum atoms. The most common mechanism for the formation of nonframework alumina is through steam dealumination during the catalyst calcination step of the manufacturing process. By careful control of the temperature, time, and steam levels during the manufacturing process, it is possible to produce a catalyst that is extremely stable at typical alkylation conditions.

From a commercial standpoint this knowledge has had the additional benefit of developing a regeneration protocol that is extremely robust. It has been demonstrated in commercial in situ and ex situ procedures that the beta zeolite catalyst can be regenerated with excellent results providing complete restoration of fresh catalyst performance. The feature of complete regenerability is another attribute that distinguishes beta zeolite catalysts from other commercially available zeolite catalysts such as MCM-22 catalyst, where significant activity and selectivity can be lost upon regeneration.[19] The ability to regenerate catalyst is essential in a commercial environment to provide additional flexibility to cope with a wide range of feedstock sources, feedstock contaminants, and potential operational upsets.

The historical development of beta zeolite showed that early versions of beta catalyst demonstrated less than optimum performance when compared to today's state-of-the-art formulation. Fig. 8 is a plot of the relative stability of beta zeolite as a function of the Si/Al_2 ratio of the beta zeolite structure in which the dominating influence of this parameter is evident. Uop has learned to stabilize the zeolite structure through careful process and chemical means. This has resulted in a catalyst system that is extremely robust, highly regenerable, and tolerant of most common feedstock impurities.

Additional studies of beta zeolite have come to similar conclusions. For example, Enichem has found that beta zeolite is the most effective catalyst for cumene alkylation

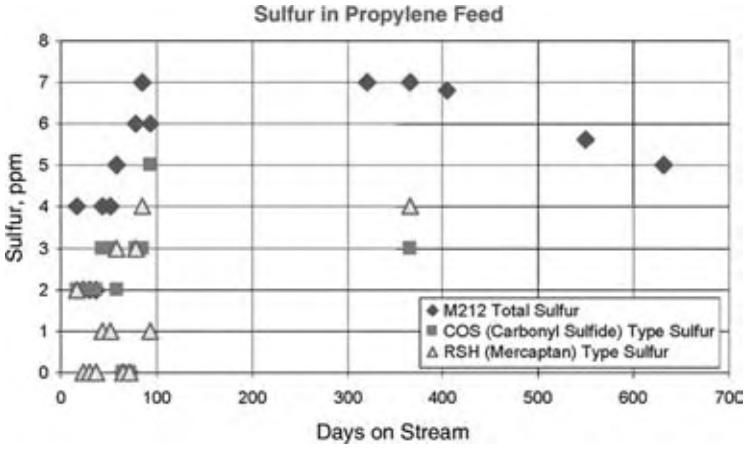


Fig. 11 JLM commercial data on sulfur in propylene feed. *(View this art in color at www.dekker.com.)*

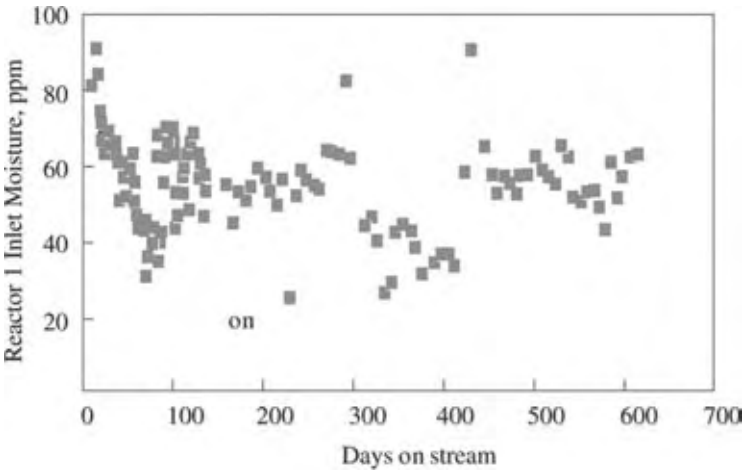


Fig. 12 JLM commercial data on alkylation reactor moisture content. *(View this art in color at www.dekker.com.)*

among other zeolites tested including Y, mordenite, and an isostructural synthesis of MCM-22 catalyst.[20]

The principles described above also led to the development of the new-generation QZ-2001 alkylation catalyst.[21] In Fig. 9, results from accelerated stability testing of QZ-2000 and QZ-2001 catalysts demonstrate the superior stability of the latest catalyst system. QZ-2001 catalyst exhibits as much as twice the stability when compared to QZ-2000 catalyst. The benefit afforded by QZ-2001 catalyst can be utilized by cumene producers in several ways. It can be taken directly through reduction of the catalyst loading for a specific run length, or alternatively as a convenient way to increase run length, or to increase throughput through the cumene unit, by allowing operation at a lower B/P feed ratio.

Long-Term Stability of Beta Zeolite Catalyst

Commercial operation with a wide variety of olefin feedstocks from different sources has demonstrated the flexibility of beta zeolite in the Q-Max process. Refinery grade, chemical grade, or polymer grade propylene feedstocks have been successfully used to make high-quality cumene product in the Q-Max process.

A good example of the ruggedness of the beta zeolite catalyst can be found in the case of JLM's Blue Island (Illinois) Q-Max operation. The operation started in August 1996 as the first Q-Max process operation with UOP beta zeolite catalyst. Initial operating results were reported in 1997.[22] The unit has continued to operate with stable performance for more than 7 yr without catalyst regeneration in spite of the presence of significant levels of feed contaminants.

Excellent monoalkylation selectivity has also been observed over many years of service in the JLM operation as shown in Fig. 10. Under the normal operating conditions of the unit, an equilibrium cumene selectivity of about 91 mol% is predicted. Thus, results clearly show that the beta zeolite catalyst is active enough to achieve near-equilibrium selectivity. This is an important feature of the catalyst as the amount of dialkylate that must be processed in the transalkylator and the subsequent cost of processing this material are minimized.

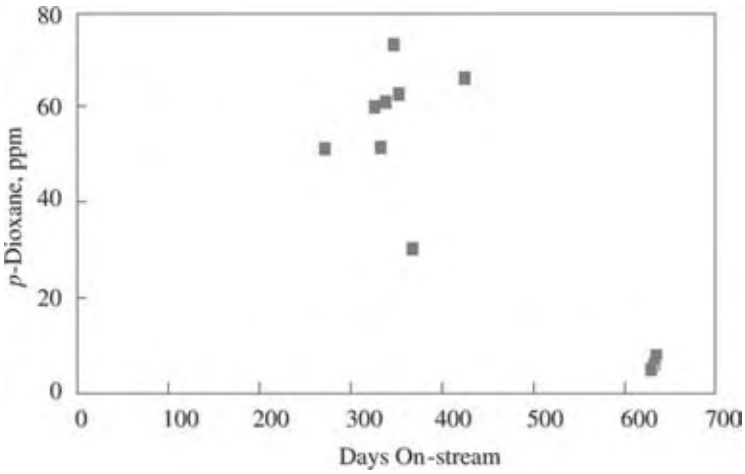


Fig. 13 JLM commercial data on benzene feed *p*-dioxane content. *(View this art in color at www.dekker.com.)*

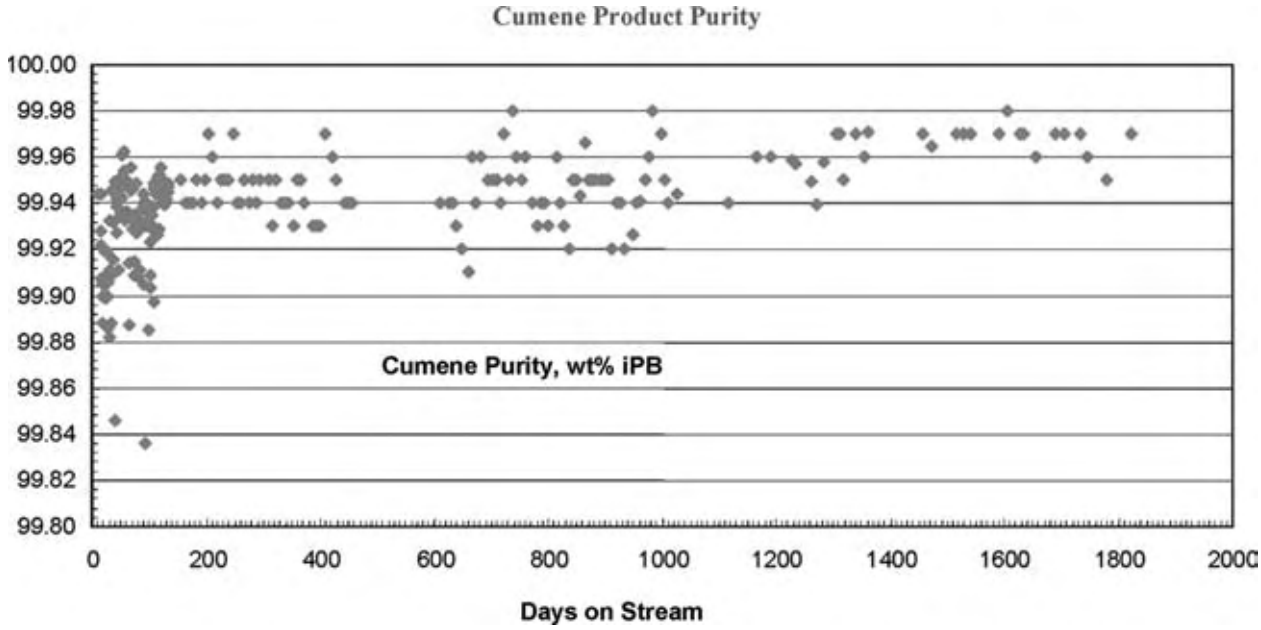


Fig. 14 JLM commercial data on cumene product purity. *(View this art in color at www.dekker.com.)*

Feed contaminants such as sulfur, water, and *p*-dioxane were monitored very closely during the first 2 yr onstream as shown in Figs. 11–13.

The impact of these impurities on cumene product quality was negligible during this period. The customer was able to maintain extremely high-quality cumene throughout the life of the beta catalyst, as shown in Fig. 14.

Another important observation is that sulfur levels from 3 to 7 ppm, moisture levels of 30–70 ppm, and even *p*-dioxane excursions up to 70 ppm had virtually no impact on catalyst stability or performance, as seen in Figs. 10 and 15. Note that the alkylation catalyst selectivity and catalyst bed inlet temperature and weight average bed temperature remain virtually unchanged after years of operation.

Beta Zeolite Catalyst Regeneration

As a result of beta's high activity and robustness, catalyst requirements are minimized. At the end of

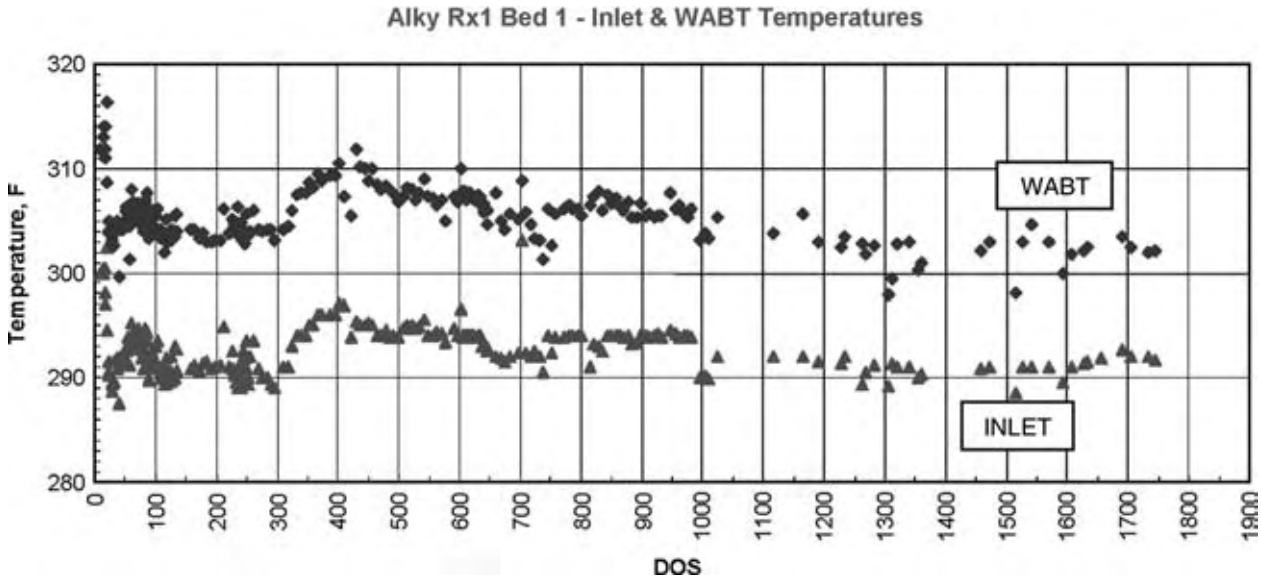


Fig. 15 JLM commercial data. *(View this art in color at www.dekker.com.)*

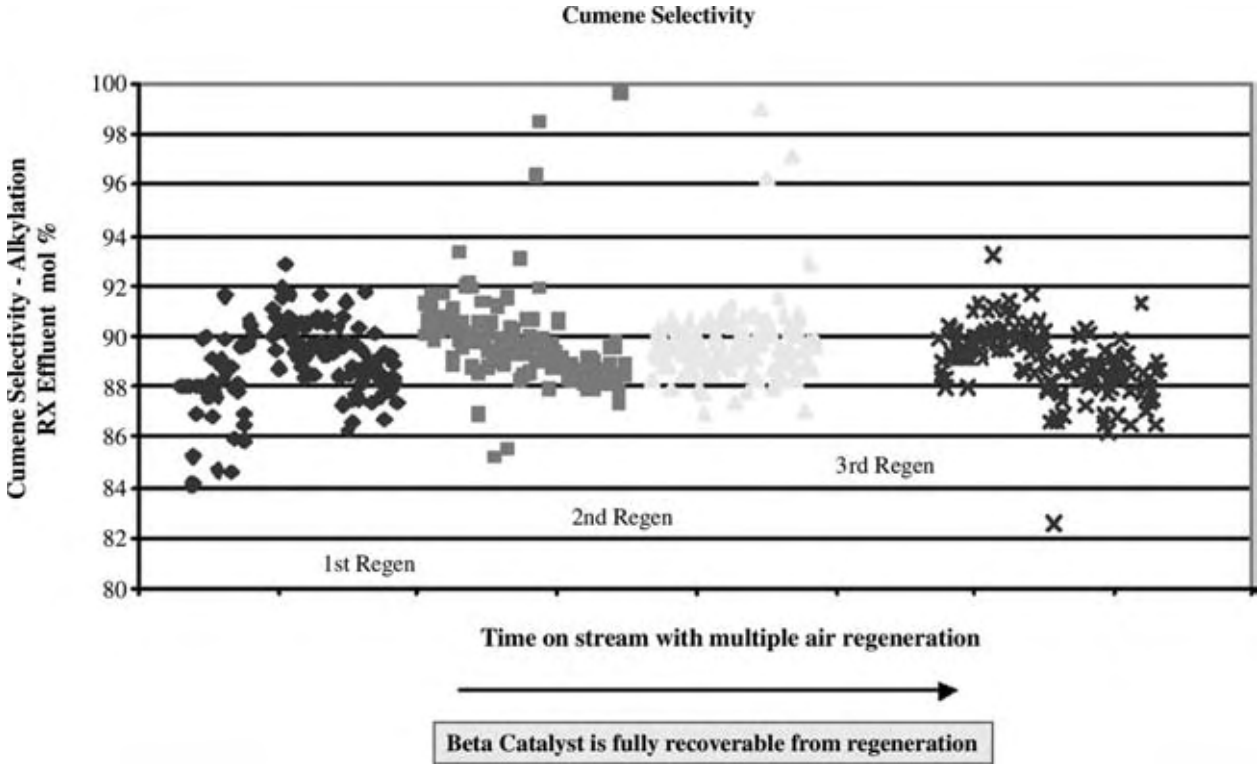


Fig. 16 Commercial Q-Max process data. Effect of in situ QZ-2000 catalyst regeneration on cumene selectivity. *(View this art in color at www.dekker.com.)*

each cycle, the catalyst can be regenerated ex situ by a certified regeneration contractor. The regenerability of beta zeolite catalyst provides an ultimate life of three cycles or more if appropriate processing and regeneration guidelines are followed. If desired, the Q-Max process unit can be designed to accommodate in situ catalyst regeneration. Both options have been successfully demonstrated in commercial operation.

Figs. 16 and 17 show an example of the performance of QZ-2000 catalyst after multiple in situ

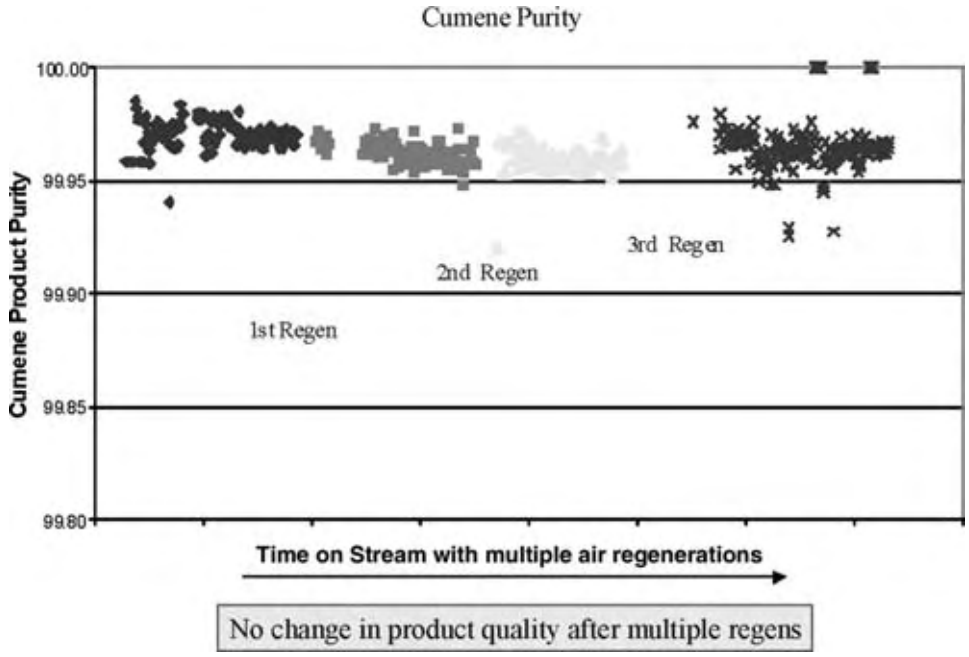


Fig. 17 Commercial Q-Max process data. Effect of in situ QZ-2000 catalyst regeneration on cumene purity. *(View this art in color at www.dekker.com.)*

regenerations in a commercial cumene unit. The cumene unit experienced premature deactivation because of excessive basic nitrogen levels in the feed as well as unsatisfactory plant operation.

Taking advantage of the in situ regeneration capability, the customer opted to regenerate the catalyst three times during this period. The results show the remarkable resilience of the beta zeolite catalyst to the stresses of regeneration with virtually no loss in monoalkylate selectivity or cumene product quality as a result of repeated regenerations.

CONCLUSIONS

Fixed-bed zeolitic cumene technology is the process of choice for the cumene/phenol industry. Most of the previously installed cumene capacity based on older $AlCl_3$ and SPA technologies has now been replaced with the newer zeolitic technology over the past 10 yr. The result is greatly improved yields and reduced operating and capital costs. UOP's Q-Max process based on beta zeolite has emerged as the leader for cumene technology. This is primarily due to the high activity, robustness, and lower operating costs associated with operating a beta zeolite based catalyst system.

REFERENCES

1. Bentham, M., et al. Process improvements for a changing phenol market. In DeWitt Petrochemical Review Conference, Houston, TX, Mar 19–21, 1991.
2. Stefanidakis, G.; Gwyn, J.E. In *Encyclopedia of Chemical Processing and Design*; McKetta, J.J., Cunningham, W.A., Eds.; Marcel Dekker: New York, 1977; Vol. 2, 357.
3. Keim, W.; Roper, M. In *Ullmann's Encyclopedia of Industrial Chemistry*; Gerhartz, W., Ed.; VCH Verlagsgesellschaft: Weinheim, 1985; Vol. A1, 185.
4. Sedaglat-Pour, Z. *Cumene*; CEH Data Summary; SRI International: Menlo Park, CA, Mar 1989, 638.5000A.
5. Chem. Mark. Rep. **1987**, *232* (10), 54.
6. Shoemaker, J.D.; Jones, E.M., Jr. Cumene by catalytic distillation. In 1987 NPRA Annual Meeting, Mar 29–31, 1987.
7. Jones, E.M., Jr.; Mawer, J. Cumene by catalytic distillation, In AICHE Meeting, New Orleans, LA, Apr 6–10, 1986.
8. Stadig, W.P. Cumene. Chem. Process. **1987**, *50* (2), 27.
9. Shoemaker, J.D.; Jones, E.M., Jr. Cumene by catalytic distillation. Hydrocarbon Process **1987**, *66* (6), 57.
10. Canfield, R.C.; Unruh, T.L. Improving cumene yields via selective catalysis. Chem. Eng. **1983**, *90* (6), 32.
11. Canfield, R.C.; Cox, R.C.; McCarthy, D.M. Monsanto/Lummus crest process produces lowest cost cumene. In *AICHE* 1988 Spring Meeting, New Orleans, LA, Apr 6–10, 1986.
12. Wadlinger, R.L.; Kerr, G.T. U.S. Patent 3,308,069, Mar 6, 1967. Mobil Oil Corporation.
13. Innes, R.A.; Zones, S.I.; Nacamuli, G.J. Liquid Phase Alkylation or Transalkylation Process Using Zeolite Beta. U.S. Patent 4,891,458, Jan 2, 1990. Chevron Research and Technology Company: San Francisco, CA.
14. Cannan, T.R.; Hinchey, R.J. Synthesis of Zeolite Beta. U.S. Patent 5,139,759, Aug 18, 1992. UOP: Des Plaines, IL.
15. Inwood, T.V.; Wight, C.G.; Ward, J.W. Liquid-phase Alkylation and Transalkylation Process. U.S. Patent 4,459,426, Jul 10, 1984, Union Oil.
16. Cheng, J.C., et al. A comparison of zeolites MCM-22, beta, and USY for liquid phase alkylation of benzene with ethylene. Sci. Technol. Catal. **1998**, *6*, 52–60.
17. Cheng, J.C., Smith, C.M.; Venkat, C.R.; Walsh, D.E. Continuous Process for Preparing Ethylbenzene Using Liquid Phase Alkylation and Vapor Phase Transalkylation. U.S. Patent 5,600,048, Feb 4, 1997, Mobil Oil Corporation,
18. Bellussi, G., et al. Liquid phase alkylation of benzene with light olefins catalyzed by beta zeolites. J. Catal. **1995**, *157*, 227–234.
19. Dandekar, A.B., et al. Regeneration of Aromatic Alkylation Catalysts Using Hydrocarbon Stripping. WO 01/83408, Nov 2001. Mobil Oil Corp, Baytown, TX.
20. Perego, C., et al. Experimental and computational study of beta, ZSM-12, Y, modernite and ERB-1 in cumene synthesis. Microporous Mater **1996**, *6*, 395–404.
21. Schmidt, R.; Zarchy, A.; Petersen, G. New developments in cumene and ethylbenzene alkylation—Paper 124b. In 1st Annual Aromatic Producers Conference—AIChE Spring Meeting, New Orleans, LA, Mar 10–14, 2002.
22. Jeanneret, J.; Greer, D.; Ho, P.; McGehee, J.; Shakir, H. The UOP Q-Max process: setting the pace for cumene production, Presented at the 22nd Annual De Witt Petrochemical Review, Houston, TX, Mar 18–20, 1997.

Dehumidification

Louay M. Chamra
B. Keith Hodge
Southeast Cooling, Heating and Power Application Center, Department of Mechanical Engineering, Mississippi State University, Mississippi State, Mississippi, U.S.A.

INTRODUCTION

People who experience hot, muggy weather or cold, clammy weather readily understand the discomfort associated with high humidity conditions. Humidity affects our comfort both directly and indirectly in various ways. Controlling humidity in a conditioned space can be important for a wide variety of comfort- and health-related reasons. In addition, humidity control is important in transport and manufacturing processes. Traditional concerns include moisture damage during marine transport, humidity damage for moisture-sensitive artifacts, process improvement for certain manufacturing sectors (confectioneries, electronics, and pharmaceuticals), and product protection from degradation (some foods, grains, and seeds) and from corrosion.

Studies conducted over the last few decades have shown that active humidity control is required in modern, sealed buildings. These studies investigated humidity control using cooling coils [direct expansion (DX) or chilled water], desiccant dehumidification units, mechanical ventilation devices, and humidification units. Additionally, the use of hygroscopically active materials (such as plaster, brick, and ceramic) to control the relative humidity of the surfaces has been considered. This topic is divided into three sections—cooling coil units, desiccant dehumidification devices, and hygroscopically active materials.

COOLING COIL UNITS

A popular humidity-reducing method is to use a cooling coil to sub-cool the air and followed by reheat. In traditional cooling systems, dehumidification is achieved by cooling a moist air stream below its dew point, so that liquid water condenses out of the air. The approximate process path is illustrated in a psychrometric chart in Fig. 1. The process shown is for air being cooled and dehumidified from conditions of 95°F dry bulb (db), 75°F wet bulb (wb) to about 77°F db, 58 grains/lbm_{da}. Initially, the dry bulb temperature of the moist air decreases, while the moisture content remains constant, as shown in Fig. 1. The dry bulb temperature continues to decrease as moisture begins to condense out of the air onto the cooling coil. To deliver air at 77°F db, 58 grains/lbm_{da}, reheat must be used. The path of reheat process is also illustrated in Fig. 1. The resulting air state lies at the center of the ASHRAE summer comfort zone in Fig. 2. The American Society of Heating, Refrigerating, and Air-conditioning Engineers (ASHRAE) defines thermal comfort as ''that condition of mind in which satisfaction is expressed with the thermal environment.'' ASHRAE Standard 55 specifies conditions for which 80% of sedentary (or slightly active) persons find the environment thermally acceptable. In this example, the total net cooling load is 10.8 BTU/lbm_{da}, and, of this, 6.4 BTU/lbm_{da} or about 59% is latent load.

The dehumidification performance of a cooling coil system can be characterized in several ways. The most obvious indicator is the latent capacity, defined as the difference between the total and sensible capacities. The latent capacity increases with reduced airflow and higher entering wet-bulb temperature. However, the sensible capacity changes with these variables as well. As the operation of a DX system is almost always controlled by a dry bulb temperature thermostat, greater latent capacity alone does not necessarily mean lower humidity in a building.

In assessing the ability of an HVAC system to maintain acceptable humidity levels in the building, the more important performance parameter is the sensible heat ratio (SHR) of the equipment. The SHR is defined as the ratio of the sensible cooling load to the total cooling load. Improved latent performance is characterized by a lower SHR. Defining the relationships between SHR and other dehumidification indices is also convenient. In particular, the latent heat ratio (LHR) can be defined as the latent cooling load fraction, and the latent to sensible ratio (LSR) can be defined as the ratio of latent load to sensible load. These ratios appear as

$$\mathrm{SHR} = \frac{Q_{\mathrm{sensible}}}{Q_{\mathrm{sensible}} + Q_{\mathrm{latent}}} \tag{1}$$

$$\mathrm{LHR} = \frac{Q_{\mathrm{latent}}}{Q_{\mathrm{sensible}} + Q_{\mathrm{latent}}} = 1 - \mathrm{SHR} \tag{2}$$

$$\mathrm{LSR} = \frac{Q_{\mathrm{latent}}}{Q_{\mathrm{sensible}}} = \frac{1 - \mathrm{SHR}}{\mathrm{SHR}} \tag{3}$$

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007728

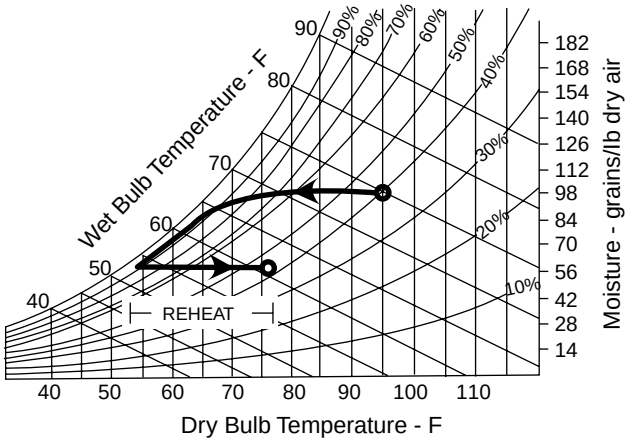


Fig. 1 Sub-cooling dehumidification process. (From Ref.[1].)

The SHR can also be related to the psychrometric properties of the entering and leaving air states.

$$\mathrm{SHR} = \frac{Q_{\mathrm{sensible}}}{Q_{\mathrm{sensible}} + Q_{\mathrm{latent}}} = \frac{c_p(T_{\mathrm{a,ent}} - T_{\mathrm{a,lvg}})}{h_{\mathrm{a,ent}} - h_{\mathrm{a,lvg}}} \quad (4)$$

where Q_{sensible} is the sensible cooling load, Q_{latent} the latent cooling load, c_p the specific heat of air, $T_{\mathrm{a,ent}}$ the entering air temperature, $T_{\mathrm{a,lvg}}$ the leaving air temperature, $h_{\mathrm{a,ent}}$ the entering air specific enthalpy, and $h_{\mathrm{a,lvg}}$ the leaving air specific enthalpy.

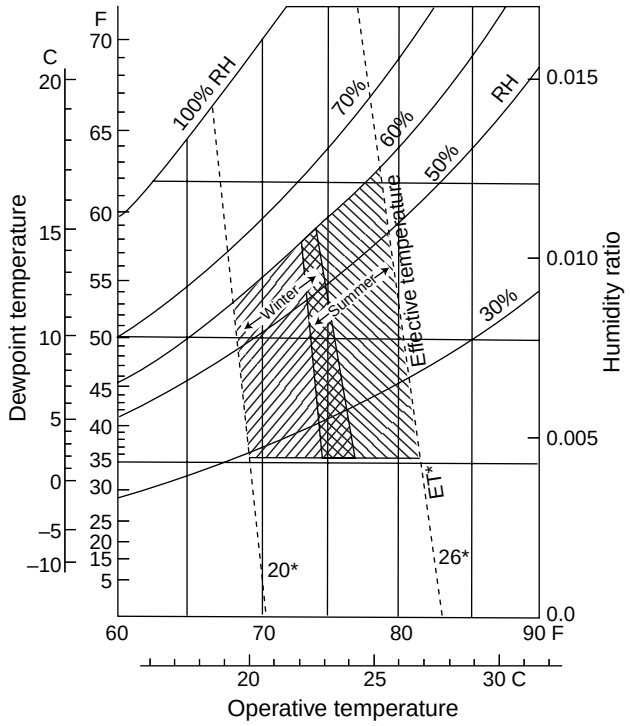


Fig. 2 ASHRAE comfort zones. (From Ref.[2].)

The relationships among the psychrometric states, loads, and sensible heat ratio are shown graphically in Fig. 3. For a given pair of entering and leaving air states, sensible, latent, and total loads are proportional to the differences in temperature, humidity ratio, and enthalpy, respectively, as shown in the figure. The SHR is defined by the slope of the line connecting the two points.

Brandemuehl[3] evaluated the ability of unitary equipment to maintain adequate space humidity levels. The evaluation involved the dehumidification requirements in different buildings and at different climates. He analyzed the building load characteristics to identify and describe key factors that affect humidity control in commercial building applications. The results indicated that ventilation loads dominated the dehumidification characteristics of most commercial buildings and dramatically affect the SHR of the system loads. He concluded that, in most cases, the dehumidification requirements were beyond the capability of conventional unitary HVAC equipment.[3]

In the conventional system, the same equipment is used for both sensible cooling and dehumidification. If both humidity control and temperature control are required, a provision for reheat of the sub-cooled air must be included. In the example in Fig. 1, the net cooling load is 10.8 $\mathrm{BTU/lbm_{da}}$, but the total load on the cooling coil is 16.1 $\mathrm{BTU/lbm_{da}}$ with the difference (5.3 $\mathrm{BTU/lbm_{da}}$) being added back during the reheat process. Thus, energy is used both for the excess sub-cooling and for the reheat.

Another disadvantage of the conventional approach is that the air leaving the evaporator coil is nearly saturated, with a relative humidity typically above 90%. This moist air travels through duct work until the air either is mixed with dryer air or reaches the reheat unit. The damp ducts, along with wet evaporator coils and standing water in a condensate collection pan (Fig. 4), foster problems with microbial growth and the associated problems of health and odor.

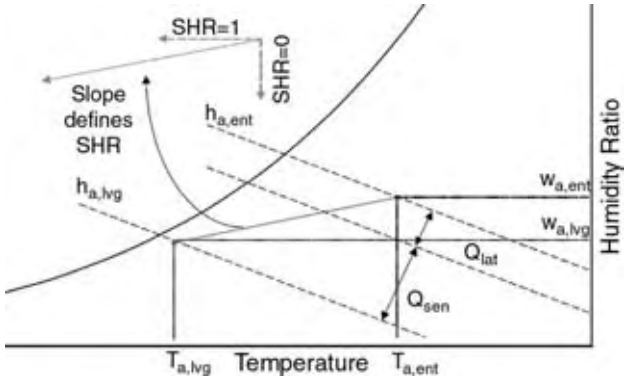


Fig. 3 Coil loads and sensible heat ratio on psychrometric chart.

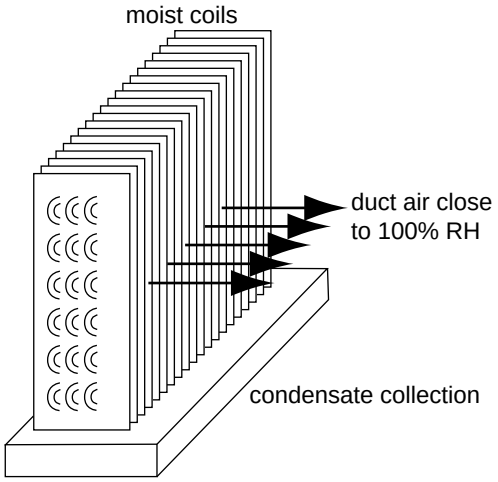


Fig. 4 Damp duct symptoms. (From Ref.[1].)

DESICCANT DEHUMIDIFICATION UNITS

Desiccant dehumidification systems remove moisture from the air by forcing the water vapor directly into a desiccant material. The moisture from the air is attracted to the desiccant because an area of low vapor pressure is created at the surface of the desiccant. The pressure exerted by the water in the air is higher, hence the water molecules move from the air to the desiccant and the air is dehumidified.

The functioning of a desiccant material might be compared to the action of a sponge in collecting a liquid. When the sponge is dry, it soaks up the liquid effectively. Once it becomes saturated, the sponge is taken to a different spot, the liquid is expelled by squeezing the sponge, and the dry sponge is ready to absorb more liquid. In a desiccant system, if the desiccant material is cool and dry, its surface vapor pressure is low and moisture is attracted and absorbed from the air, which has a high vapor pressure. After the desiccant material becomes wet and hot, it is moved to another air stream and the water vapor is expelled by raising the temperature (this step is called ''regeneration''). After regeneration, the desiccant material is ready to be brought back to absorb more water vapor. The entire process involves only water vapor—no liquid is ever condensed.

Desiccant dehumidification units are available with both solid and liquid desiccants. The difference between solid and liquid desiccants is their reaction to moisture. Some simply collect moisture like a sponge collects water. These desiccants are called adsorbents and are mostly solid materials. Silica gel is an example of a solid adsorbent. Other desiccants undergo a chemical or physical change as they collect moisture. These are called absorbents and are usually liquids or solids that become liquid as they absorb moisture. Lithium chloride collects water vapor by absorption. Sodium chloride, common table salt, is another example of an absorbent.

Solid desiccant dehumidification units contain desiccant-impregnated wheels. The solid wheel units come in two common configurations—single-wheel units and dual wheel units. Single-wheel units (sometimes called TWERS—total energy recovery system) are commonly used in conjunction with evaporative cooling systems (such as the one used in ice rinks). Dual wheel (sometimes referred to as DWERS—dual wheel energy recovery system) and liquid units are usually paired with cooling and heating coils for temperature control. Regardless of the type, desiccant units can provide increased energy savings for ventilation.

A variety of factors determine whether an adsorbent will be useful as desiccants. These factors include cost, long-term stability, moisture removal characteristics (rate, capacity, saturation conditions, and suitable temperatures), regeneration requirements (rate of moisture surrender as a function of temperature and humidity), availability, and manufacturing considerations.

Solid Adsorbents

Silica gels and zeolites are used in commercial desiccant equipment. Other solid desiccant materials include activated aluminas and activated bauxites. The choice of desiccant material for a particular application depends on factors such as the regeneration temperature, the level of dehumidification, and the operating temperature.

Solid desiccant materials are arranged in a variety of ways in desiccant dehumidification systems. A large desiccant surface area in contact with the air stream is desirable, and a way to bring regeneration air to the desiccant material is necessary.

The most common configuration for commercial space conditioning is the desiccant wheel shown in Fig. 5A. The desiccant wheel rotates continuously between the process and the regeneration air streams. The wheel is constructed by placing a thin layer of desiccant material on a plastic or metal support structure. The support structure, or core, is formed, so that the wheel consists of many small parallel channels coated with desiccant. Both ''corrugated'' and hexagonal (Fig. 5B) channel shapes are currently in use. The channels are small enough to ensure laminar flow through the wheel. Some kind of sliding seal must be used on the face of the wheel to separate the two streams. Typical rotation speeds are between 6 and 20 rotations per hour. Wheel diameters vary from 1 ft to over 12 ft. Because dust or other contaminants can interfere with the adsorption of water vapor and quickly degrade the system performance, air filters

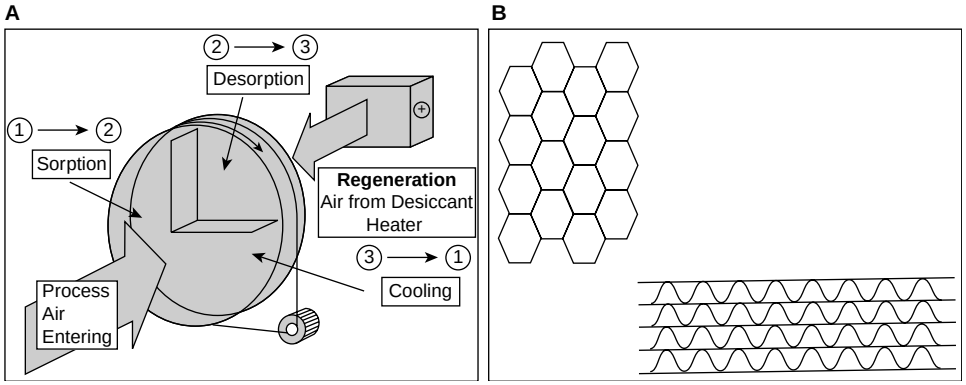


Fig. 5 (A) Desiccant wheel. (B) Corrugated and hexagonal channel shapes. (From Refs.[1,4].)

are an important component of solid desiccant systems. All commercial systems include filter and maintenance protocols for keeping the filters functioning properly.

Liquid Absorbents

Materials that function as liquid absorbents include ethylene glycols, sulfuric acid, and solutions of the halogen group such as lithium chloride, calcium chloride, and lithium bromide (ASHRAE[5]). A generic configuration for a liquid desiccant system is illustrated in Fig. 6. The process air is exposed, usually by spraying the solution through the air stream, to a concentrated desiccant solution in an absorber. As the solution absorbs water from the air stream, the concentration drops, and the weak solution is taken to a regenerator, where heat is used to drive off the water (which is carried away by a regeneration air stream) and the concentrated solution is returned to the absorber. Liquid desiccant systems provide the added advantage of removing many particulates from the air stream. Some liquid desiccants kill bacteria as well. Furthermore, liquid desiccant systems can be configured to operate with very low regeneration temperatures.

Regeneration

For solid or liquid systems, regeneration energy can be drawn from a variety of sources. Because of the relatively low temperature requirements of regeneration (<250°F), waste heat provided by combustion

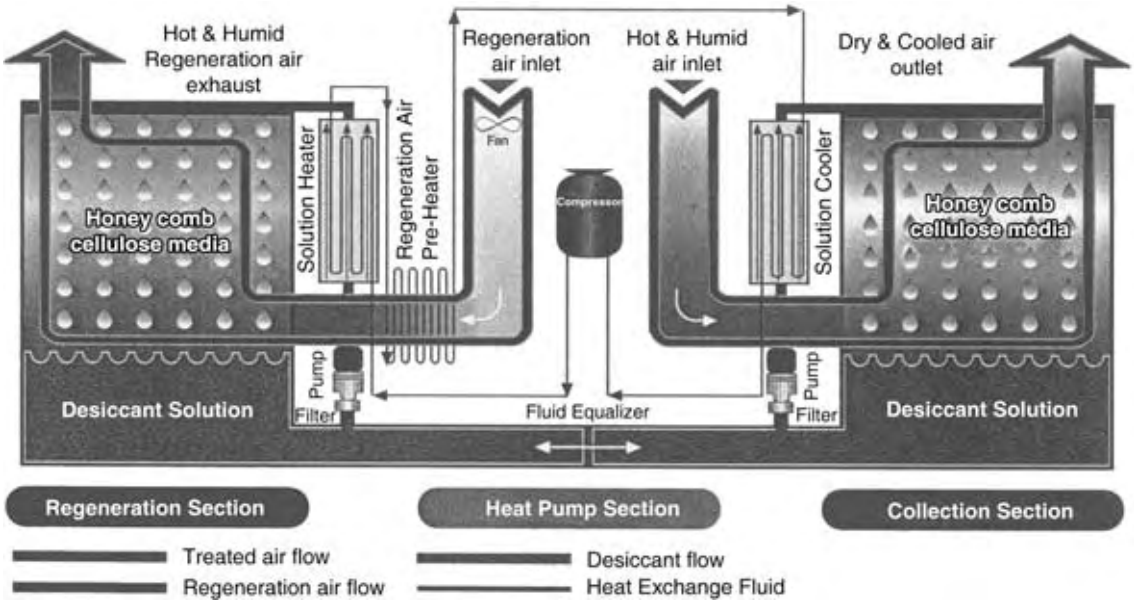


Fig. 6 Liquid desiccant unit schematic. (From Ref.[6].)

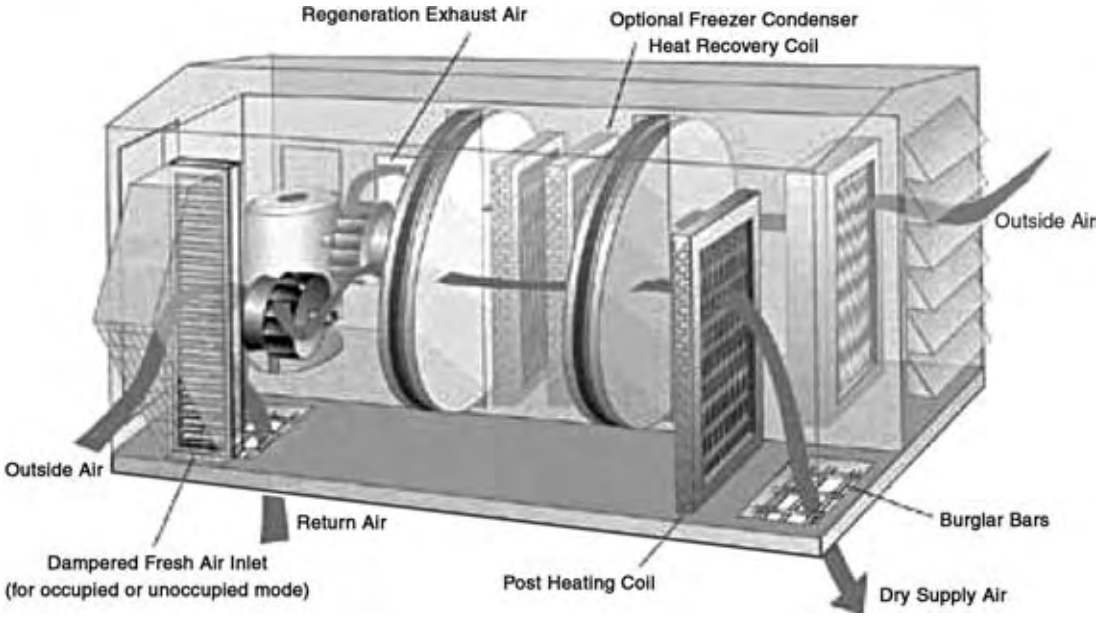


Fig. 7 Dual wheel desiccant schematic. (From Ref.[7].) *(View this art in color at www.dekker.com.)*

turbines, IC engines, and any of the fuel cell technologies is capable of supplying heat at regeneration temperatures.

Solid Desiccant System

The components of a generic solid desiccant dehumidification system are illustrated in Fig. 7. At a minimum, the system will include separated air streams of process and regeneration for the desiccant device and some kind of heat source to raise the temperature of the regeneration air.

The approximate path of the process air through a desiccant device is shown in Fig. 8 for the same inlet and outlet conditions as were shown for the sub-cooling system (Fig. 1). Note that, as implied by the path from point 1 to point 2 in Fig. 8, the desiccant process increases the dry bulb temperature of the process air. For solid desiccant materials, this increase is a result of the ''heat of adsorption'', which consists of the latent heat of vaporization of the adsorbed liquid plus an additional ''heat of wetting.'' The heat of wetting is the energy released during dehumidification, in excess of the latent heat of vaporization. The path from point 1 to point 2 is close to a line of constant enthalpy. After the dehumidification process, the process air must undergo a sensible cooling process to reach the end point.

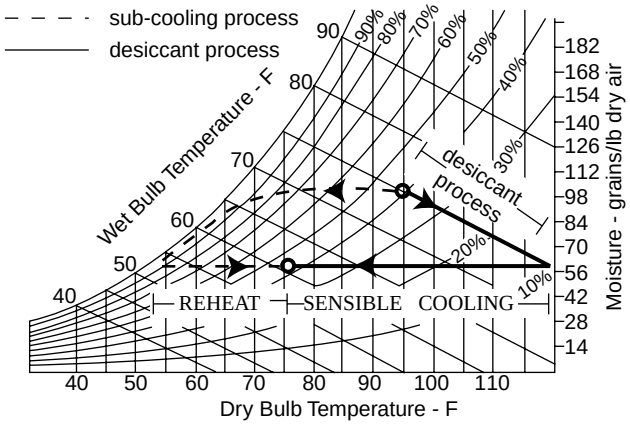


Fig. 8 Dry desiccant dehumidification process. (From Ref.[1].)

A wide variety of configurations of desiccant dehumidification system are available. The air inlet conditions and outlet requirements of process and regeneration call for different configurations depending the individual situations. For illustration purposes, two examples are presented in Figs. 9 and 10.

In Fig. 9, a configuration for a desiccant system that is designed for operation in a ventilation situation is illustrated; that is, 100% outside air is used to supply the conditioned space. The air for regeneration is taken from within the conditioned space and exhausted outside. The outside air starts at state 1 at approximately 95°F db, 75°F wb. From state 1 to state 2, the sensible temperature increases and the moisture content decreases as the outside air passes through a desiccant wheel. From state 2 to state 3, the hot air rejects some heat to the regeneration air stream via a heat exchanger. Finally, the air stream is cooled to the design supply condition by passing it through a conventional cooling coil (state 3 to state 4).

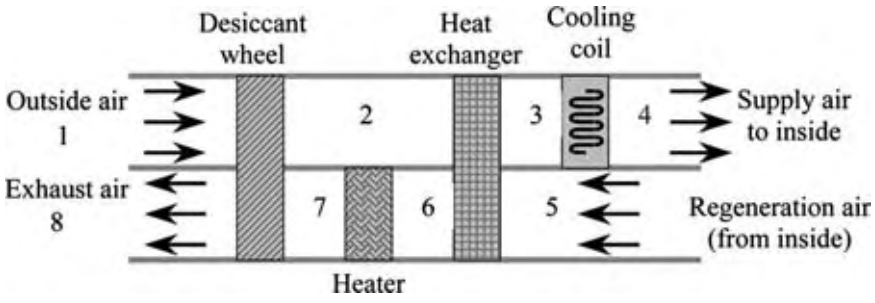


Fig. 9 Ventilated desiccant dehumidification system configuration. *(View this art in color at www.dekker.com.)*

The regeneration air stream starts at approximately 75°F db, 40% rh. From state 5 to state 6, this air stream acquires heat from the process air stream through a heat exchanger, and from state 6 to state 7, the regeneration stream is further heated to an appropriate temperature for desiccant regeneration. Finally, the regeneration air is cooled and humidified as it extracts moisture from the desiccant wheel (state 7 to state 8). At state 8, the regeneration stream is exhausted to the outside.

In Fig. 10, a re-circulating configuration is illustrated; 100% of the process air is drawn from the conditioned space (and all is returned to the conditioned space). The regeneration air is 100% from outside. The equipment arrangement is identical to that of the ventilation illustration in Fig. 9; only the air stream conditions are different.

Benefits and Cost Considerations

To effectively discuss the benefits and disadvantages of desiccant systems, the two common methods of installation must be understood. A ventilation installation is one in which the desiccant unit is completely independent of the air conditioning (A/C) coils. (In this discussion, both the cooling and the heating coils of conventional equipment are referred to as air conditioning coils). Such an installation allows outdoor air to be continuously supplied to the indoor space without requiring the fan of the main air conditioning system to operate simultaneously with the desiccant unit. Another advantage of this installation is that it allows for humidity control across several control zones in a building (or a group of buildings). Common applications for ventilation installations are in large office buildings and schools. A preconditioning installation is one in which the desiccant unit is connected to pretreat the outdoor air before it is treated by the main air conditioning unit. This type of installation requires the fan on the air conditioner to operate when the desiccant unit is operating. One advantage to a preconditioning installation is that it can eliminate the seasonal variations experienced by the A/C coil, which improves the energy efficiency of the coils (Fischer[8]). A typical range of outdoor air conditions throughout a year is shown in Fig. 11 as the teardrop shape filled with diagonal lines. The oval in the center of the seasonal variations is the typical range of air conditions that a desiccant unit supplies year around, when installed as a preconditioning system. Common applications for a preconditioning installation are in homes and small commercial buildings.

Regardless of the variation, desiccant units operate similarly. During the winter mode, the heat recovery (sensible) wheel of a dual wheel system does not rotate. Fig. 12 is an operational diagram, with the states indicated, of a desiccant system operating in ventilation mode; while both summer and winter modes of operation are shown in the figure, only the winter mode will be discussed (for a discussion of the summer ventilation mode of operation refer to Fischer[8]). On the process side, cold, dry air enters the desiccant wheel and is heated and humidified. The air then passes unchanged through the inactive cooling coil and the stationary sensible wheel. On the regeneration side,

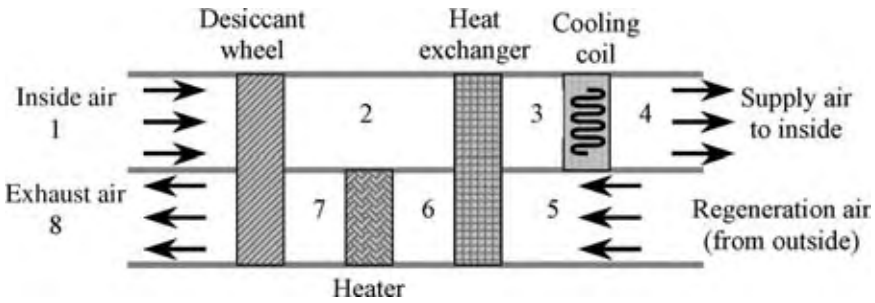


Fig. 10 Re-circulated desiccant dehumidification system configuration. *(View this art in color at www.dekker.com.)*

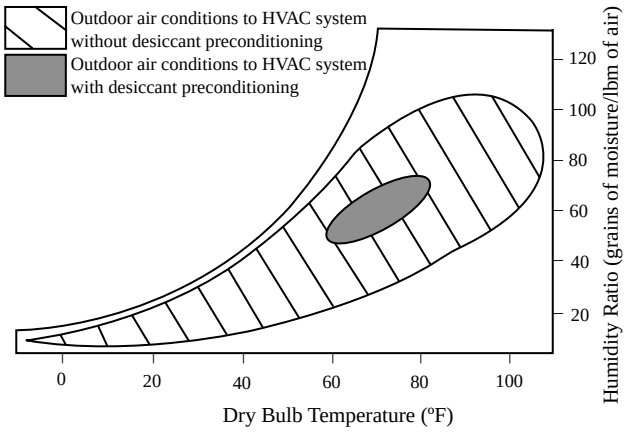


Fig. 11 Desiccant preconditioning target. (From Ref.[8].)

warm moist return air passes unchanged through the stationary sensible wheel and then passes through the desiccant wheel. The return air then exits the unit cooler and drier. In the example shown in Fig. 12, the supply air of 61.5°F and 31 grains is both cooler and less humid than the return air of 70°F and 38 grains. While the outdoor air has been dehumidified, the net effect on the indoor environment is that it has been dehumidified less than in conventional ventilation systems. Green,[9–11] Sale,[12] and Gelperin[13] recommend that the indoor environment be humidified during the winter. While the use of a desiccant unit to condition the ventilated fresh air reduces the humidity in the indoor space, it does not reduce the humidity nearly as much as ventilation alone.

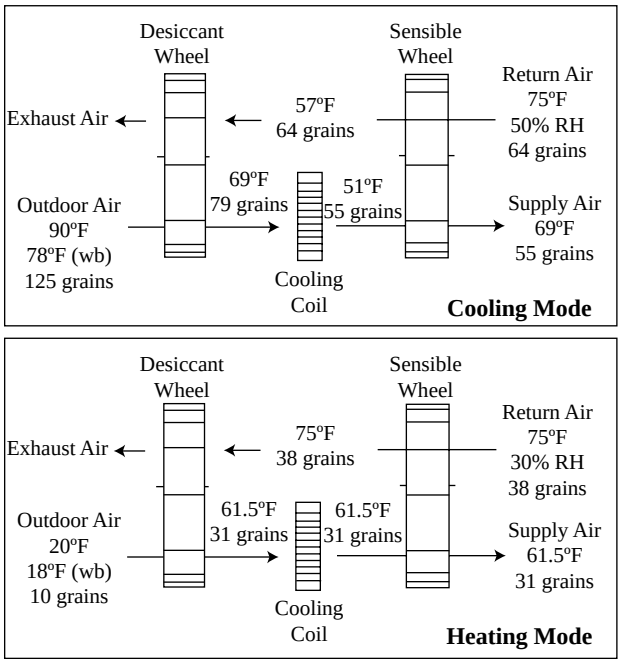


Fig. 12 Operating characteristics of a dual wheel desiccant unit. (From Ref.[8].)

Desiccant Dehumidification vs. Cooling Coil Dehumidification

Desiccant units have several advantages over cooling coil units. Downing and Bayer[14] conducted an investigation into a Georgia school with indoor air quality problems; the indoor air quality problems at the school were so bad that students walked out of class in protest and lawsuits were filed against the school system. As part of this investigation, data were collected for direct expansion cooling coils supplying 5 cfm/person continuously, direct expansion cooling coils supplying 5 cfm/person intermittently, and a dual wheel desiccant unit supplying 15 cfm/person; these data are shown in Fig. 13. As demonstrated in this figure, the direct expansion cooling coils could not maintain the indoor relative humidity low enough to prevent microbial problems. However, the dual wheel desiccant unit was able to maintain the humidity level low enough to help prevent microbial problems. Notice that at 5 cfm/student intermittently, the direct expansion unit operates below the maximum recommended relative humidity during the occupied hours; however, when the units are not run during the unoccupied hours, the relative humidity rises above the threshold. At 5 cfm/student continuously, the direct expansion unit cannot maintain an acceptable relative humidity; again there was a sharp increase in the relative humidity during the unoccupied hours when the direct expansion unit was shut down. The increase in the relative humidity when the direct expansion units were not operating was exacerbated by the evaporation of the condensed water in the cooling coil drain pan. Data for 15 cfm/student intermittently and continuously were taken; however, Downing and Bayer[14] did not present these data as the relative humidity for those cases was significantly higher than that for the 5 cfm/student continuously. When a dual wheel desiccant unit was installed and operated at 15 cfm/student continuously during both occupied and unoccupied hours, the relative humidity remained relatively constant. Fischer[8] does a complete economic analysis comparing desiccant units to direct expansion units providing 15 cfm/person continuously and concluded, as mentioned by Baughman and Arens[15] and Downing and Bayer,[14] that desiccant systems (unlike direct expansion cooling coils) do not require condensate drain pans, which can become a significant source of microbial growth.

The humidity-control capability of desiccant technology offers many potential cost savings when compared with conventional sub-cooling systems. The capital cost of a desiccant system is often more than that of an equivalent sub-cooling system because of extra equipment costs. Installed capital cost for active, solid desiccant systems range from $4 to $9

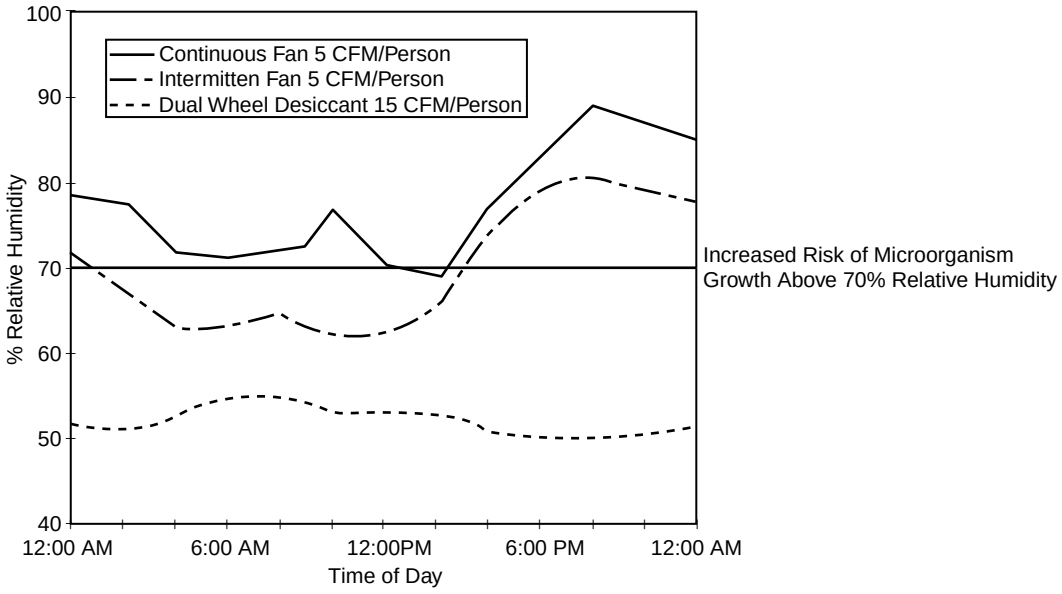


Fig. 13 Daily humidity level comparisons. (From Ref.[14].)

per cfm capacity for air handling, depending upon the total capacity and equipment enclosure requirement. The higher-end of the cost range applies to systems with <5000 cfm. Some desiccant systems, depending on the specific installation, may result in lower capital cost; however, the reduction of latent heating load will usually result in lower operating costs for desiccant systems. Harriman[16] discusses the installation of a desiccant dehumidification system in a medical research building where both capital costs and estimated operating costs were lower for the desiccant system.

Because the humidity and temperature can be controlled independently with a desiccant dehumidification and cooling system, the system performance is often more effective than that obtainable with conventional systems. Analysis of the SHR suggests the potential for saving energy cost that a desiccant system may have. An SHR close to unity implies that very little moisture is removed from the air, while a sensible heat ratio close to zero indicates that most of the load is latent cooling. Air-conditioned environments often have SHR values well below unity, which results in greater energy consumption for a sub-cooling system than that for a desiccant dehumidification and cooling system that meets the same zone temperature requirements.

Some potential cost savings vary with installation and are recognized based on the individual application. These cost savings are related to dehumidification and are implied by the process/product benefits outlined in Table 1. Other potential costs associated with poor humidity control should not be overlooked. Examples include retail establishments where customer discomfort or dissatisfaction because of unpleasant odors and merchandize appearance would be detrimental to business. For grocery stores, frost buildup in refrigerated display cases is unsightly, and its elimination requires expensive defrost cycles. Hospitals and nursing homes require careful attention to minimizing conditions favorable to microbial growth and propagation. Additionally, microbial growth can be hazardous to the health of children in school buildings as well as to the occupants of office and other commercial buildings. Ice rinks can reduce "fogging," condensation in the building, and can improve ice quality with lower humidity levels.

HYGROSCOPICALLY ACTIVE MATERIALS

Building materials can be divided into three categories—non-porous media, hygroscopic-porous media, and capillary-porous media. Arens and Baughman[17] define non-porous materials, such as plastic, glass, and sheet metal, as those that condense moisture on the surface of the material; hygroscopic-porous materials, such as wood, clay, and natural fibers, as those that have microscopic pores that reduce the vapor pressure at the surface and volume changes with the material moisture content; capillary-porous materials, such as brick, concrete, and gypsum board, as those that have visible pores and whose volumes are not affected by moisture levels. Initially, hygroscopic materials may appear to promote biological growth (especially natural fibers); however, wood-based products will adsorb/desorb the moisture in cellulose molecules which removes the water, and thus, not allowing it to be used for microorganism growth.[17]

Table 1 Process/product benefits because of dehumidification

Process	Product benefits
Lithium battery production	Prevent corrosion and improve production
Computer and electronic equipment production	Prevent condensation and corrosion on metal surfaces
Plastic molding	Improve product finish by preventing condensation on metal surfaces
Archives and museums	Increase longevity of books, artwork, and artifacts
Seeds and grain storage	Optimize seed moisture level and minimize microbial deterioration
Confectionary and pharmaceutical packaging	Keep products from deteriorating
Confectionary manufacturing	Improve product and appearance production

(From Ref.[1].)

Straube and deGraauw[18] provide an extensive review of a popular hygroscopically active material. Their research shows that cement-bonded wood fiber (CBWF) has ''an interesting and unique mix of vapor permeability and vapor storage capacity.'' Though CBWF had been known to provide a healthier indoor environment, it was Straube and deGraauw[18] who first attempted to learn why CBWF provides a healthier indoor environment. Their research also concludes that hygroscopically active materials may be able to help prevent mold and fungi problems during conditions of peak relative humidity by adsorbing some of the moisture from the ambient air. One study (West and Hansen[19]) found that hygroscopically active materials can influence the indoor relative humidity by as much as 15–20%. More research in this area is necessary before the widespread use of hygroscopically active materials can be recommended as a means to alleviate problems of indoor air quality. Baughman and Arens[15] report that wood paneling can ''adsorb/desorb large amounts of humidity on a diurnal cycle [over the course of a day] without surface condensation occurring.''

The use of hygroscopically active materials may help in controlling the space relative humidity during the peak humidity periods, which can be significantly high when using natural ventilation, mechanical ventilation, or direct expansion dehumidification units. Effects of hygroscopically active material could be even more significant in high-humidity rooms such as kitchens and bathrooms. Therefore, when no humidity control is present and in rooms that are usually humid, the use of hygroscopically active materials is recommended. However, there is little information, such as sorption rate and moisture capacity, available on most hygroscopic materials; therefore, this recommendation is made with the condition that only materials with known hygroscopic properties are used.

HUMIDITY CONTROL AND INDOOR AIR QUALITY

Humidity control may be accomplished by using cooling coil units, desiccant dehumidification units, and hygroscopically active materials. Effects of most indoor air pollutants are strongly influenced by the

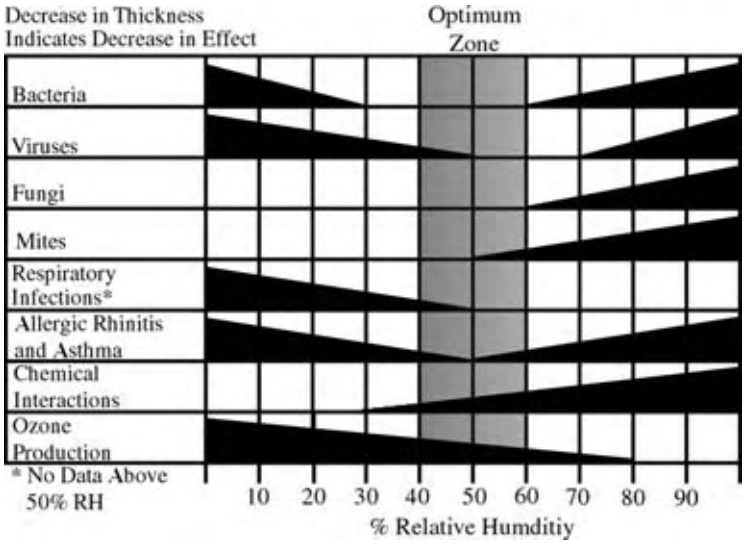


Fig. 14 Relative humidity effects on indoor air pollutants and human illnesses.

relative humidity. In Fig. 14, the effect of relative humidity on common indoor air pollutants and human ailments is illustrated. Here, the optimum relative humidity range is 40–60%. For this reason, common practice is to design humidification systems to maintain a 50% relative humidity; however, the needs of the space will determine the designed humidity set-point. The two most significant pollutants in the 40–50% relative humidity range are chemical interactions and ozone production.

CONCLUSIONS

Desiccant systems are the most convenient means to control the indoor relative humidity. As Downing and Bayer[14] have shown, desiccant systems can supply larger quantities of fresh air without introducing an unacceptably high relative humidity to the indoor environment. As desiccant systems remove the moisture from the process air stream through a mass transfer process, there is little chance for the system to promote microorganism growth. A major limiting factor in desiccant systems is the initial cost; however, the added benefits of the reduced health care costs because of better humidity (and microorganism) control (Fischer[8]). Furthermore, the initial cost of a desiccant unit to handle high ventilation rates is roughly equivalent to purchasing a direct expansion unit that can handle the same ventilation rates.[8]

REFERENCES

1. Chamra, L.; Parsons, J.A.; James, C.; Hodge, B.K.; Steele, W.G. Desiccant Dehumidification Curriculum Module for Engineering/Technology echnology HVAC Courses, Mississippi State University, 2000.
2. W.R. Grace & Co., Davison Silica Gels, Davison Chemical Division, Baltimore, MD, 1996.
3. Brandemuehl, M.J. Humidity Control Options, ASHRAE 1121-RP, Final Report, 2000.
4. Meckler, M.; Heimann, R.; Fisher, J.; McGahey, K. Desiccant Technology Transfer Workshop Manual, American Gas Cooling Center, Arlington, VA, 1995.
5. ASHRAE. *ASHRAE Handbook: Fundamentals*. American Society of Heating, Refrigeration, and Air Conditioning Engineers, Inc.: Atlanta, GA, 2001.
6. DryKor Corporation. Sales Brochure, 2002.
7. McGahey, K. New commercial applications for desiccant-based cooling. ASHRAE J. **1998**, *40* (7), 41–45.
8. Fischer, J.C. Optimizing IAQ, humidity control, and energy efficiency in school environments through the application of desiccant-based total energy recovery systems. *IAQ '96: Paths to Better Building Environments*. ASHRAE: Atlanta, GA, 1996; 179–194.
9. Green, G.H. The effect of indoor relative humidity on absenteeism and colds in school. ASHRAE Trans. **1974**, *80* (Part 2).
10. Green, G.H. The positive and negative effects of building humidification. ASHRAE Trans. **1982**, *88* (Part 1), 1049–1061.
11. Green, G.H. Indoor relative humidities in winter and the related absenteeism. ASHRAE Trans. **1985**, *91* (Part 1B).
12. Sale, C.S. Humidification to reduce respiratory illnesses in nursery school children. Southern Med. J. **1972**, *65* (7), 882–885.
13. Gelperin, A. 1973. Humidification and upper respiratory infection incidence. Heating/Piping/Air Conditioning **1973** March, 77–78.
14. Downing, C.C.; Bayer, C.W. Classroom indoor air quality vs. ventilation rate. ASHRAE Trans. **1993**, *99* (Part 2).
15. Baughman, A.V.; Arens, E.A. Indoor humidity and human health—part I: Literature review of health effects of humidity-influenced indoor pollutants. ASHRAE Trans. **1996**, *102* (Part 1), 193–211.
16. Harriman III, L.G. Improving humidity control for commercial buildings. ASHRAE J. **2000** November, 24–32.
17. Arens, E.A.; Baughman, A.V. Indoor humidity and human health—Part II: Buildings and their systems. ASHRAE Trans. **1996**, *102* (Part 1), 193–211.
18. Straube, J.F.; deGraauw, J.P. Indoor air quality and hygroscopically active materials. ASHRAE Trans. **2001**, *107* (Part 1), 444–450.
19. West, M.K.; Hansen, E.C. Effect of hygroscopic materials on indoor humidity and air quality. Proceedings of IAQ: Desiccant Dehumidification and Cooling, ASHRAE: Atlanta, GA, 1992, 178–182.

Denitrogenation

Daeik Kim
Teh Fu Yen
University of Southern California, Los Angeles, California, U.S.A.

INTRODUCTION

Petroleum (rock-oil, derived from Latin *petra*, meaning rock or stone, and Latin *oleum*, meaning oil) occurs widely in the earth as a gas and a liquid.[1] Petroleum is a mixture of gas, liquid, and solid hydrocarbon-type chemical compounds that occur in sedimentary rock deposits throughout the world. In the crude state, petroleum has minimal value, but when refined, it provides high-value liquid fuels, solvents, lubricants, and many other products.

The fuels derived from petroleum contribute between one-third and one-half of the world's total supply of energy. They are used not only for transportation fuels (e.g., gasoline, diesel fuel, aviation fuel, and jet fuel), but also to heat buildings. Petroleum products are used to lubricate machines and a once-maligned byproduct, asphalt, is used to provide highway surface and roofing materials.

Petroleum is by far the most commonly used liquid fuel source. Nevertheless, dwindling supplies of this source make it essential that other sources of liquid fuels be found. Such sources are heavy oils, oil sand bitumens, coal, and oil shale. In fact, the occurrence of petroleum and its uses as sources of much-needed liquid fuels may be likened to the ''tip of the iceberg'' when compared to the resources of other fossil fuels. For instance, in the case of oil sand, the Alberta oil sand deposits contain approximately 1.6 billion barrels of oil,[2] which is the principle oil source in America. Thus, development of other fossil fuels will be expected to be the major energy recovery as petroleum stocks are depleted. Petroleum and other fuel sources contain nitrogen, sulfur, and metals in small amounts, as well as oil and gas. Those small amounts of compounds tend to degrade the quality of fuels as the final product. In this sense, desulfurization and demetallization technologies have been well developed and much investment was allowed as part of military and commercial projects, such as fuel cell and battle ship desulfurizer. Sulfur and metal removal have drawn more attention since these compounds are contained more in fuel sources than nitrogen. At the same time, sulfur in fuel leads directly to emission of SO_2 and sulfate particular matter (PM) which endanger public health and welfare. The regulation situation has become more stringent, requiring the use of low-sulfur fuel.[3] Nitrogen compounds contained in energy sources are equally problematic, even though its content is generally smaller than sulfur in fuel. Nitrogen compounds directly help produce NO_x and related air pollutions that are worse in urban areas, reacting with ozone in automobile emission and sunlight. Denitrogenation technology is required to decrease the nitrogen content in fuel, which will increase the opportunities for exploiting the energy sources that are presumed to be less valuable. This technology can even upgrade the value of current high-quality fuel products which still contain small amounts of nitrogen compounds, eventually resulting in reduction of air pollution.

NITROGEN AND DENITROGENATION

Nitrogen Compounds in the Environment and Petroleum

Nitrogen is a major element of nutrients for living organisms. It forms a very important component of the building blocks of protein in the living cell, which means that deoxyribonucleic acid (DNA) contains nitrogen. Microorganisms circulate nitrogen in the process of nitrification and denitrification. Finally, nitrogen fixation takes place in plants with the help of microorganisms which helps in maintaining a constant 79% nitrogen composition of air.

Natural gas, petroleum, tarsands, oil shale, and coal are called fossil fuels because of their presumed biogenic origin in buried decaying remains deposited in layers and geochemically transformed through heat and pressure over the course of time into their present form. For this reason, fossil fuels contain nitrogen compounds.

Many fuels and crude oils contain a certain amount of nitrogen in their compounds. For example, the nitrogen content of petroleum is generally within the range of 0.1–0.9%, although crude oils with no detectable nitrogen or even trace amounts are not uncommon.[4] In general, the ''heavier'' the oil, the higher is their nitrogen content. When comparing the amount of nitrogen in petroleum to all other fossil fuels of different locations, oil shale (Colorado) has 0.41%,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120038486

oil shale (Alaska) has 0.30%, kerogene (Green River) has 2.13%, bitumen (Lower Freeport) has 1.42%, and asphaltene (Baxterville) has 0.80%.[5]

Different fossil fuels have different compositions. As mentioned, many nitrogen-containing organic species of biological origin are present in the precursors of fossil fuel deposits that have undergone very different geological transformations through biostratinomy (change after death) and taphonomy (fossilization). This difference not only shows the distinctive features in geochemical genesis and deposition histories of fossil sources, but also provides a new insight into the properties of pollutants and their physiological activities.

Nitrogen, oxygen, and sulfur are called heteroatoms (impurities) in fossil fuels. This trio comes in small amounts in fossil fuels, but they control the quality and treating the impurities improves the quality of fuels. Different fossil fuels with different geological features can have different major heteroatoms (impurities). For instance, the major heteroatom in crude petroleum is sulfur. Nitrogen is the major heteroatom in oil shale, as is oxygen in coal. Generally, oil shale shows higher nitrogen contents than does crude oil.

Why Do We Need Denitrogenation?

Chemically speaking, petroleum is an extremely complex mixture of hydrocarbon compounds, with a small amount of oxygen, nitrogen, sulfur, and metal impurities. These impurities result in the degradation of the quality of petroleum and lessen the production ability of available oil resources. Therefore, it is essential to lower the nitrogen content before any refinery processes are performed.

Nitrogen compounds in oil are converted by combustion or incineration to nitrogen oxides (NO_x), and hence to one of the main sources of acid rain and air pollution. The nitrogen compounds are also undesirable in the various refining processes because they deposit on to the catalyst, causing a loss in catalytic activity. Pyrrolic compounds, also nitrogen-containing (with concentrations as low as 0.01%), can cause both storage instability and the promotion of gum with sediment formation in fuels. The use of heavy oils as catalytic cracking feedstocks has accentuated the harmful effects of the nitrogen compounds, which are more prevalent in the higher boiling crude oils.[6]

Currently, denitrogenation technology for light oils is carried out industrially via the catalytic hydrodenitrogenation (HDN) process and simultaneous hydrodesulfurization (HDS).[7] The HDN of petroleum products is more difficult than the HDS. The production of light oil, which contains both very low nitrogen and sulfur, has a tendency to require inevitably severe operating conditions (high temperature and high pressure) and the use of particular active-catalysts, which induce high costs. Chemically assisted ultrasound process (CAUP) is relatively cost- and energy-effective since the CAUP method can produce localized high temperature and high pressure instantly at low energy consumption to break down the pollutants or target compounds.

Nitrogen Containing Model Compounds (Aniline; C_6H_7N, Indole; C_8H_7N, Carbazole; $C_{12}H_9N$)

Aniline, indole, and carbazole contain nitrogen bonding in their chemical structure. Tables 1–3 show where to these model compounds belong.

In the case of indole, alkaline hydrolysis and putrefaction of proteins result in its formation. This formation in the putrefaction of proteins is presumed to be result of decomposition of tryptophan. The formation of indole from albumin may be stopped by the addition of lactose, while other sugars have varying effects on its production. Indole frequently accompanies pus formation and is found in the human liver, pancreas, brain, and bile. Indole, accompanied by its β-methyl homolog, skatole, is found in the feces of humans and animals and in the contents of the intestines.[8]

Indole and homologs of indole have been found in coal and molasses tar. It is also present in ''practical'' α-methylnaphthalene. Its presence occurs by reaction with oxalyl chloride to give the acid chloride of 3-indoleglyoxylic acid. Indole can be prepared by the reduction of indoxyl by all of the following: sodium amalgam, zinc dust and alkali, catalytically, and dehydrogenation. In the preparation of indoxyl or indoxylic acid in the synthesis of indigo, a small amount of indole is obtained when the melt is overheated. Indole has been prepared in fair yields by adding sodium amalgam or zinc dust to the alkaline melt of indoxylic acid.[8]

Indole derivatives are found in many natural products. Indole has unpleasant (fecal) odors. It has

Table 1 Polycyclic aromatic hydrocarbon (PAH)

PAH: Naphthalene, fluorene, phenanthrene, and their alkylated homologs	
PASH (polycyclic aromatic sulfur hydrocarbon): benzothiophene and dibenzothiophene	PANH (polycyclic aromatic nitrogen hydrocarbon): indole, carbazole, quinoline and their alkylated derivatives

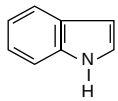
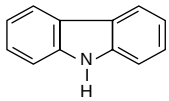
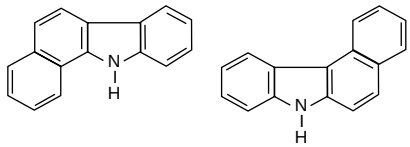
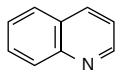
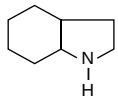
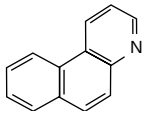
Table 2 Polycyclic aromatic hydrocarbons (PAHs)

Carbonaceous PAHs	Heteroatomic PAHs
Naphthalene, anthracene, pyrene, coronene	Hensofuran, benzothiophene, indole, carbazole benzoxazole, quinoline, isoquinoline, dibenzofuran

been found in *Robinia pseudacacia*, the jasmines, certain citrus plants, the perfume of the *Hevea braziliensis*, and in orange blossom.[8] Indole is also found in the wood of *Celtis reticulosa*. It is even found in Mexican mushroom, Epena snuff, Indian snuff, and the Isatis herb. The indole structures contribute hallucinatory and neurohumorous properties. Indole is usually obtained by repeated extraction of the blossoms with a suitable solvent and subsequent removal of the solvent by distillation.

Many different types of nitrogen-containing compounds have been shown in oil shale. Poulson and co-workers[9] reported that pyridines and pyrroles were major types of nitrogen compounds in shale oil. Cyclic amides and anilides were proposed as possible additional types of nitrogen compounds in shale oil. Van Meter and co-workers found benzonitriles in shale oil naphtha. Brown and co-workers identified alkylpyridines, cycloalkanopyridines, alkylanilines, quinolines, tetrahydroquinolines, and tetrahydroisoquinolines in

Table 3 Basic and nonbasic nitrogen containing species in crude oils

Nonbasic	Pyrrole	C_4H_5N	
	Indole	C_8H_7N	
	Carbazole	$C_{12}H_9N$	
	Benzo(a)carbazole	$C_{16}H_{11}N$	
Basic	Pyridine	C_5H_5N	
	Quinoline	C_9H_7N	
	Indoline	C_8H_9N	
	Benzo(f)quinoline	$C_{13}H_9N$	

(From Ref.[15].)

hydrocracked shale oil naphata. However, major nitrogen compound types in petroleum crude oil have been characterized mainly into four classes—pyridines, diaza compounds, carbazoles, and amides. Some light oils contain considerable amounts of indole and aniline as well.

Petroleum and fuels can be categorized into two kinds, hydrocarbons and nonhydrocarbons. Nitrogen, sulfur, and oxygen belong to the nonhydrocarbon category, which is very recalcitrant and degrades the quality of fuels.

Polycyclic aromatic compounds (PAC) are a significant environmental chemical group, with an associated health effect. PACs includes PAH (polycyclic aromatic hydrocarbon), PASH (polycyclic aromatic sulfur hydrocarbon), and PANH (polycyclic aromatic nitrogen hydrocarbon). The main PAHs are naphthalene, fluorene, phenanthrene, and their alkylated homologs; PASHs are benzothiophene and dibenzothiophene; PANHs are indole, carbazole, quinoline, and their alkylated derivatives[10] (see Table 1).

Polycyclic aromatic hydrocarbons (PAHs) are categorized into carbonaceous PAHs(naphthalene, anthracene, pyrene, and coronene) and heteroatomic PAHs (benzofuran, benzothiophene, indole, benzoxazole, quinoline, isoquinoline, dibenzofuran, and carbazole)[10] (see Table 2).

Nitrogen species are also divided through acidic extraction into basic species (such as aniline, quinoline, benzoquinoline, and their derivatives) and nonbasic species (such as indole, carbazole, and their derivatives)[8] (see Table 3).

Denitrogenation Technology

Microbial degradation[11]

Microbial degradation of nitrogen compounds from fossil fuels was one type of method to remove nitrogen. An example of this is the biodegradation of the nitrogen-containing compound, quinoline, in crude oil. For example, 2 mg/L of carbazole (which accounts for 80% of nitrogen compound), and 500 mg/L of carbazole (which accounts for 24% of nitrogen compound), were removed after 15 days. Microbial degradation however has some drawbacks, such as (1) it is a very slow method; (2) there are difficulties in manipulating the bacterial growth; (3) it needs the skill of a professional operator; and (4) it cannot handle high concentrations of pollutants.

Pyrolysis[12]

Pyrolysis (coal) is another method that is old but still widely used. Pyrolysis oil derived from oil shale was hydrotreated in a stirred reactor at 400 C for 12 hr. The catalysts are nickel- and cobalt-molybdenum types. The results showed that the nitrogen and sulfur contents of the oils were significantly decreased after hydrotreatment and were further decreased with increasing pressure of hydrogen. However, this method requires very high temperature and pressure, which means high costs.

Ferric chloride-clay complexation method[13]

Raw shale oil is allowed to process through $FeCl_3$-clay chromatography. $FeCl_3$-clay is prepared by contacting the attapulgus (Engelhard Minerals and Chemicals) with a methanolic ferric chloride solution (saturated $FeCl_3 \cdot 6H_2O$ in methanol) for 1 hr. $FeCl_3$-clay is filtered, washed, and extracted with pentane for 48 hr in a Soxhlet extractor to remove retained, nonadsorbed, iron salts and then dried under nitrogen. This method is similar to the ion exchange. The data showed that in shale oil 65% of total nitrogen, 56% of total oxygen, and 27% of total sulfur were removed by the $FeCl_3$-clay technique. This method is nondestructive because pollutants stay in another form of chemicals.

Hydrotreating reactions[14]

Denitrogenation reactivities of nitrogen species in gas oils were followed in the hydrotreating reactions at 340°C under 5 MPa of H_2 to quantify their respective reactivities by gas chromatography-atomic emission detection (GC-AED).

The reactivity orders are found as: indole > methylated aniline > methylated indole > quinoline > benzoquinoline > methylated benzoquinoline > carbazole > methylated carbazoles.

Solid acid denitrogenation[15]

A new method to remove the nitrogen compounds of lubricating base oils was suggested. A solid acid was employed as an effective reagent to remove the basic and nonbasic nitrogen compounds. However, the sulfur compound is relatively difficult to remove. Clay treating after the solid acid denitrogenation process can apparently improve the oxidation stability of denitrogenated base oils.

Ultraviolet irradiation and liquid–liquid extraction[7]

Denitrogenation for light oils is based on a combination of ultraviolet (UV) irradiation and liquid–liquid extraction. Two extraction systems, one oil/water and the other oil/acetonitrile, were used for the denitrogenation of three separate light oils of differing

nitrogen and hydrocarbon composition. Photodecomposition of carbazole was found to be suppressed by the presence of double-ring aromatic hydrocarbons. This adverse effect can be reduced by the addition of hydrogen peroxide to the water phase. In the presence of 30% H_2O_2 and 36 hr of photoirradiation, the nitrogen content of light oils was decreased to less than 20% of feed value. In the oil/acetonitrile system, the nitrogen-containing compounds in the light oils were extracted into the acetonitrile phase and were photodecomposed effectively, even in the presence of double-ring aromatics. Then, with 10 hr of photoirradiation, the nitrogen content of the light oils was decreased successfully to less than 3% of initial concentration. This method is time-consuming and is not cost-effective and therefore cannot treat high concentrations of pollutants.

Alkylation and a subsequent precipitation method using alkylating agents (CH_3I and $AgBF_4$)[16]

Nitrogen and sulfur are methylated by the addition of the alkylating agents under moderate conditions and are removed successfully as the precipitates of the corresponding *S*-methylsufonium and *N,N*-dimethlycarbasolium tetrafluoroborates. By these means, the sulfur and nitrogen concentration of vacuum gas oil were reduced simultaneously to less than 0.1 and 7.0% of initial concentration with a 20-fold molar excess of CH_3I and a 10-fold molar excess of $AgBF_4$. Analytic methods are by means of field ionization-mass spectrometry (FI-MS) and gas chromatography-atomic emission detection (GC-AED). This is a nondestruction method.

Chemically assisted ultrasound process

CAUP is an ultrasonic reaction that produces a considerable amount of energy and pressure, which causes large numbers of bubbles or cavitations. With this energy, the bounding of a pollutant can be broken. Most of conventional methods necessarily accompany high energy and pressure all the time, which is expensive. Ultrasound is an amazing source of energy and pressure with relatively low electricity. The mechanism of ultrasound can be summarized in three phenomena:[17–19] (1) solvent compression and rarefaction (rapid movement of fluids); (2) cavitation [high temperature: 20,000°F, high pressure: 75,000 psig, collision time: 1.25 (10^{-5} sec)]; and (3) microstream with little heating (great volume of vibrational energy confined to a small volume of reaction).

In summary, the target compound is dissolved in a solvent. Surfactant entails one phase of emulsion by changing surface properties of target compound and solvent while stirring, then ultrasonic reactions attack the loose bonding of target compound through a high degree of energy and pressure. The bond scission reaction occurs between N–H and C–H bonding. The concentration of the target compound is then noticeably decreased. Fig. 1 shows the mechanism of CAUP.

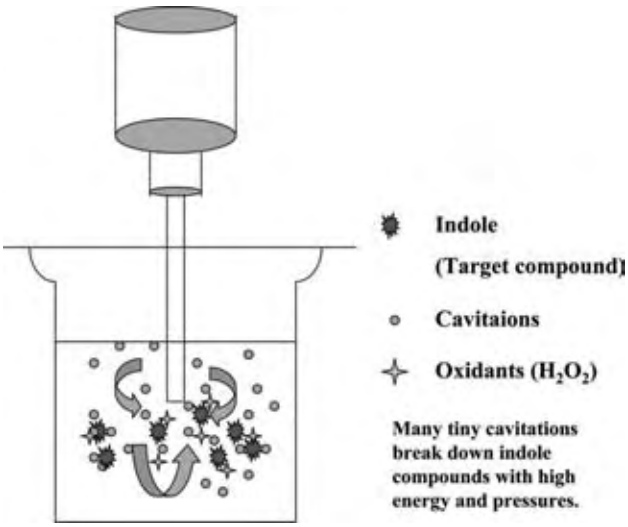


Fig. 1 Diagram showing ultrasonic reaction. *(View this art in color at www.dekker.com.)*

CONCLUSIONS

The trend in fossil fuels shows that heavy bitumen and oil shale consumption will increase over the next century because high quality oil resources are limited. This also means there is no need to expect high quality fuel production only from petroleum or natural gas. Low value energy sources will be commercially viable only if denitrogenation technology is allowed to reduce the nitrogen content. Many profitable resources such as heavy bitumen and oil shale have commanded less attention due to their low commercial value, although the exploration of oil shale for production has become an important energy program. The U.S. reserves the largest amount of oil shale in the world, but can only produce high quality oil as long as the technology is available.

Technology tends to control the industry. Denitrogenation technology is equally important as desulfurization and demetallization when it comes to upgrading low value energy resources. The theory and technology of denitrogenation will have a great impact on commercial applications in industry and scientific areas.

REFERENCES

1. Yen, T.F. *Environmental Chemistry*; Prentice Hall, Inc.: New Jersey, 1999; Vol. 4A, 157.

2. Pakdel, H.; Roy, C. Recovery of bitumen by vacuum pyrolysis of Alberta Tar Sands. Energy Fuels **2003**, *17*, 1145–1152.
3. Mei, H; Mei, B.W.; Yen, T.F. A new method for obtaining ultra-low sulfur diesel fuel via ultrasound assisted oxidative desulfurization. Fuel **2003**, *82*, 405–414.
4. Speight, J.G. *Fuel Science and Technology Handbook*; Marcel Dekker, Inc.: New York, 1990; Vol. 41, 81.
5. Yen, T.F. *Environmental Chemistry*; Prentice Hall, Inc.: New Jersey, 1999; Vol. 4A, 162.
6. Hsu-Chou, R.S.Y.; Mobashery, S.; Yen, T.F. Denitrogenation of shale oil by oxime formation from pyrroles. Energy Sources **1998**, *20*, 857–866.
7. Shiraishi, Y.; Hirai, T.; Komasawa, I. Photochemical denitrogenation processes for light oils effected by a combination of UV irradiation and liquid–liquid extraction. Ind. Eng. Chem. Res. **2000**, *39* (8), 2826–2836.
8. Sumpter, W.C.; Miller, F.M. *Heterocyclic Compounds with Indole and Carbazole Systems*; 1954; 1–2, N.Y. 2.
9. Shue, F.-F.; Yen, T.F. Concentration and selective identification of nitrogen- and oxygen-containing compounds in shale oil. Anal. Chem. **1981**, *53* (13), 2081–2084.
10. Williams, P.T.; Nazzal, J.M. Polycyclic aromatic compounds in shale oils: Influence of process conditions. Environ. Technol. **1998**, *19* (8), 775–787.
11. Sugaya, K.; Nakayama, O.; Hinata, N.; Kamekura, K.; Ito, A.; Yamagiwa, K.; Ohkawa, A. Biodegradation of quinoline in crude oil. J. Chem. Technol. Biotechnol. **2001**, *76* (6), 603–611.
12. Williams, P.T.; Nazzal, J.M. Pyrolysis of oil shales: influence of particle grain size on polycyclic aromatic compounds in the derived shale oils. J. Inst. Energy **1999**, *72* (491), 48–55.
13. Yen, T.F.; Shue, F.-F.; Wu, W.-H.; Tzeng, D. Ferric chloride-clay complexation method: removal of nitrogen-containing compounds from shale oil and related fossil fuels. Anal Chem. **1983**, 457–466.
14. Shin, S.H.; Sakanishi, K.; Mochida, I.; Grudoski, D.A.; Shinn, J.H. Identification and reactivity of nitrogen molecular species in gas oils. Energy Fuels **2000**, *18* (7–8), 539–544.
15. Wang, Y.Z.; Li, R.L. Denitrogenation of lubricating base oils by solid acid. Petrol. Sci. Technol. **2000**, *18* (7–8), 965–973.
16. Shiraishi, Y; Hirai, T; Komasawa, I. A novel desulfurization for fuel oils based on the formation and subsequent precipitation of S-alkylsulfonium salts. 4. Desulfurization and simultaneous denitrogenation of vacuum gas oil. Ind. Eng. Chem. Res. **2001**, *40* (15), 3398–3405.
17. Tu, S.P.; Kim, D.; Yen, T.F. Decolorization and destruction of metallophthalocyanines in aqueous medium by ultrasound; A feasibility study. J. Environ. Eng. Sci. **2002**, *1*, 237–246.
18. Kim, D.; Tu, S.P.; Yen, T.F. Evaluation of versatile ultrasonic effects on degradation of organometallics from petroleum. Korean J. Environ. Eng. Res. **2003**, *8* (2), 11–23.
19. Kim, D.; Shiu, F.J.Y.; Yen, T.F. Devulcanization of scarp tire through matrix modification and ultrasonication. Energy Sources **2003**, *25* (11), 1099–1112.

Design of Extrusion Dies

Milivoje M. Kostic
Department of Mechanical Engineering, Northern Illinois University, DeKalb, Illinois, U.S.A.

Louis G. Reifschneider
Department of Technology, Illinois State University, Normal, Illinois, U.S.A.

INTRODUCTION

The goal of this chapter is to introduce the reader to the importance of extrusion die design as well as the complexities inherent in the task. Extrusion is of vital importance to all plastics processing. In addition to providing raw stock such as sheet for thermoforming and pellets for injection molding and other extrusion processing, numerous end-use products are made with extrusion such as film, tubing, and a variety of profiles. Although the types of extruded products made can differ dramatically in shape, there are a set of common rules that govern basic die design. For example, it is important to streamline the flow from the inlet to the exit, and as a practical measure, to fine-tune the flow balance and product dimensions, flow adjustment devices could be included in the die design.

Several unique products are made by extrusion and the dies needed to make these products are classified as: 1) sheet dies; 2) flat-film and blown-film dies; 3) pipe and tubing dies; 4) profile extrusion dies; and 5) co-extrusion dies. Furthermore, each product type has unique hardware downstream of the die to shape and cool the extruded melt. To aid the reader, detailed illustrations of the various die designs and the complementary downstream cooling and shaping hardware are shown.

Predicting the required die profile to achieve the desired product dimensions is a very complex task and requires detailed knowledge of material characteristics and flow and heat transfer phenomena, and extensive experience with extrusion processing. Extrusion die design is still more an art than a science, even though the latter is becoming more and more relevant for design optimization because of recent advancement in the powerful computation and modeling of complex flow and heat transfer processes, before, through, and after the die.

DESIGN FUNDAMENTALS

Extrusion is a continuous process where solid polymeric materials, either pellets or powders, are sheared and heated as they are conveyed through either a single- or a twin-screw extruder (as described elsewhere) to become a pressurized melt. The pressurized melt flows through a properly shaped orifice, or extrusion die, and then is pulled (with a little pressure) as it is cooled and shaped to a final product called the extrudate. The proper design of an extrusion die is extremely important to achieve the desired shape and accurate dimensions of the extruded product. The function of an extrusion die is to shape the molten plastic exiting an extruder into the desired cross section depending on the product being made. The die provides a passage between the circular exit of the extrusion barrel and the more complex and often much thinner and wider die exit. A schematic of a common die, called a sheet die, is shown in Fig. 1A to illustrate this point. The extrusion process creates products of uniform cross section in a continuous fashion. An ideal passage will:[1,2]

- Balance the melt flow by providing a more uniform exit velocity across the entire die exit.
- Achieve this flow balance with a minimal pressure drop.
- Streamline the flow to avoid abrupt changes in the flow passage that may cause stagnation areas. Stagnated flow may lead to thermal degradation of the plastic melt as the melt is exposed to high heats for long periods.

As a practical measure, flow control devices should be incorporated into the die design to permit fine-tuning of the die passage shape to ensure a proper flow balance. In addition, the design of extrusion dies is complicated by two unique material properties of molten plastics:[3]

- Melts exhibit shear thinning behavior (become less viscous) as they are sheared.
- Melts exhibit viscoelastic behavior, which influences the ''die swell'' on exiting the die.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039324

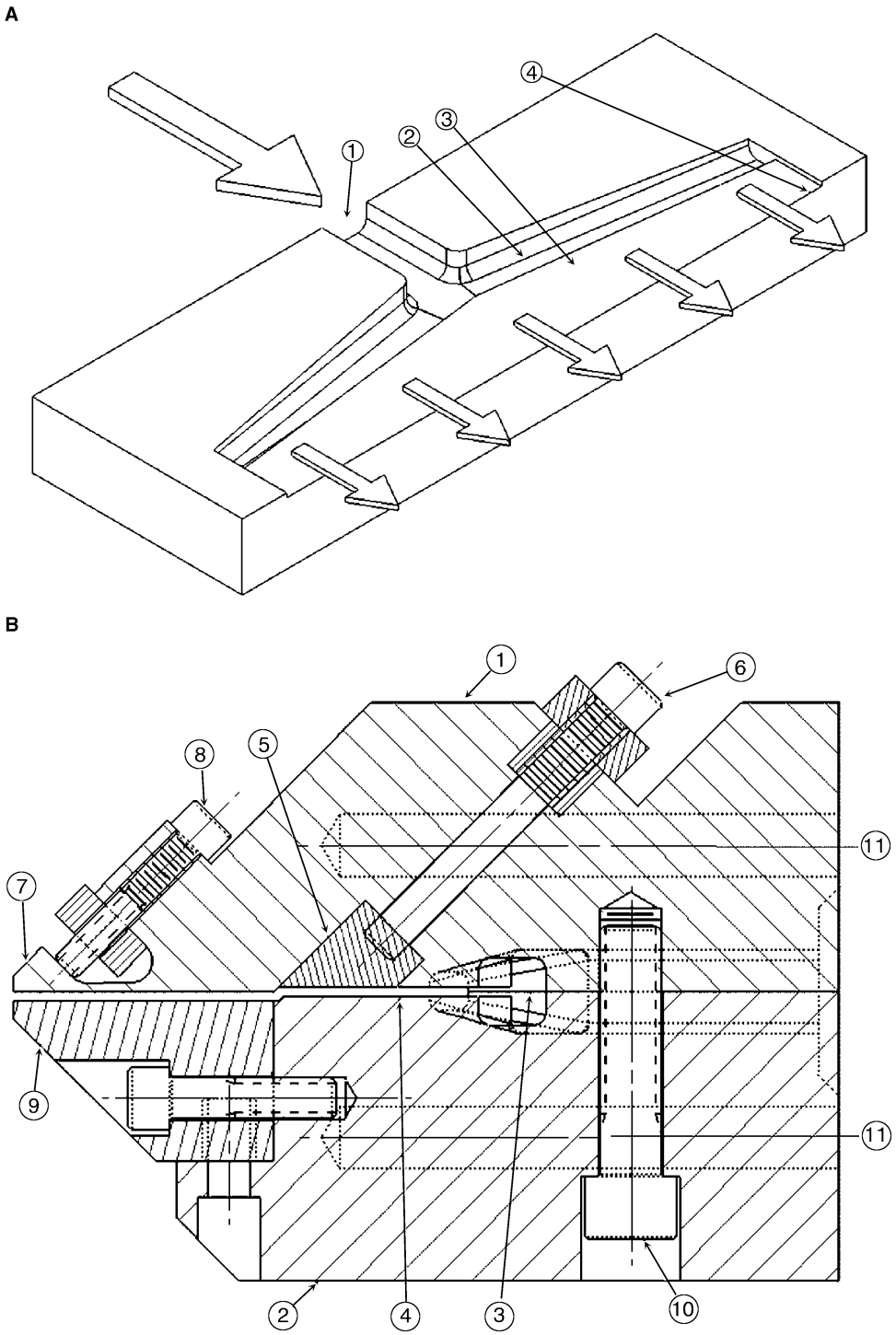


Fig. 1 Coat hanger-type sheet die concept (A): (1) central inlet port; (2) manifold (distributes melt); (3) island (along with manifold, provides uniform pressure drop from inlet to die lip; (4) die lip (die exit forms a wide slit); and schematic of sheet die (B): (1) upper die plate; (2) lower die plate; (3) manifold; (4) island; (5) choker bar; (6) choker bar adjustment bolt; (7) flex die lip; (8) flex lip adjustment bolt; (9) lower lip; (10) die bolt; (11) heater cartridge.

The shear thinning causes the volumetric flow to be very sensitive to slight changes in die geometry. For example, the flow for a typical polymer melt through a slit will vary with the cubic thickness of the gap. Thus, a small change in the die gap along the contour of the die exit may cause considerable change of the melt flow. The term ''die swell'' refers to the enlargement in the direction orthogonal to the flow direction.

Swelling after exiting the die lip is due to two distinct phenomena:[4,5]

- Velocity relaxation (unification) of the melt flow.
- Viscoelastic relaxation of the strained polymer molecules.

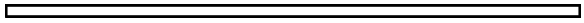
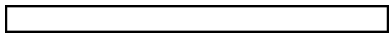
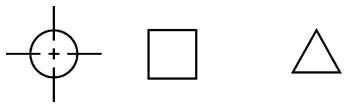
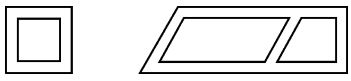
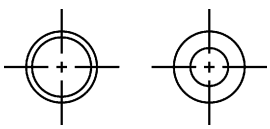
Velocity relaxation occurs because the melt is no longer under shear from the no-slip walls of the extrusion die. The melt assumes a uniform bulk velocity that causes the high-speed areas to slow down and the areas previously retarded from the wall to increase their speed. The net result is enlarging (swelling) of the melt cross-section bulk as it exits the die, while the stagnant outside region and, especially, the corners are stretched and shrunk. Newtonian and non-Newtonian fluids exhibit die swell owing to velocity relaxation. The swelling due to viscoelastic memory is a characteristic of polymeric fluids and occurs because the polymer molecules are stretched in the flow direction while passing through the high-shear area before the die exit. On exiting, the molecules recoil and shorten in the flow direction. The result is an expansion in the direction orthogonal to the flow, a swelling of the diameter of a round strand on exiting the die, for example. The amount of viscoelastic swelling is a combination of the material properties of the polymer as well as process conditions such as melt temperature, shear rate, and residence time under high shear, especially near the die exit.[6]

The design of extrusion dies today is facilitated by computer-based simulation tools. The flow of non-Newtonian fluids through complex passages is routinely performed by computational fluid dynamics (CFD) programs.[7–10] Factors such as shear thinning are readily accounted in die design. The viscoelastic behavior can also be modeled today, although this simulation requires extensive material testing to obtain the required material parameters for accurate simulation results. This is discussed in more detail later in this entry.

Extrusion dies vary in shape and complexity to meet the demands of the product being manufactured. There are five basic shapes of products made with extrusion dies, as illustrated in Table 1.[11,12] Film and sheet dies are called slit dies as the basic shape of the die exit is a slit. Film is also made with annular dies as in the case of blown film. Strand dies make simple geometric shapes, such as circles, squares, or triangles. Pipe and tubing dies are called annular dies as the melt exits the die in the shape of an annulus. The inner wall of the annulus is supported with slight air pressure during extrusion. Open profile dies make irregular geometric shapes, such as ''L'' profiles or ''U'' profiles, and combinations of these. Hollow profile dies make irregular profiles that have at least one area that is completely surrounded by material. Examples of hollow profiles can be simple, such as concentric squares to make a box beam, or a very complex window profile.

Each extruded product relies on a die to shape the moving melt followed by shaping and cooling devices downstream to form the extrudate into the final desired shape and size. A more complete treatment of these devices, which are typically water-cooled metallic

Table 1 Typical extruded product shapes

Films	$t < 0.01$ in.	
Sheets	$t > 0.01$ in.	
Profiles	Strand	
	Open	
	Hollow chamber	
Tubes	$d < 1.0$ in.	
Pipes	$d > 1.0$ in.	

devices that contact the extrudate melt, is presented later in the section ''Extrudate Cooling and Sizing Hardware.''

SHEET DIES

The most common extrusion die for sheet products is the coat hanger-type manifold die as shown schematically in Fig. 1A and as a section view in Fig. 1B. The key elements of Fig. 1A are:

- Central inlet port: connects to the extruder barrel.
- Manifold: provides a streamlined channel to evenly distribute the melt to the island.
- Island: with the manifold serves to create an equal pressure drop from the die inlet to all points across the die exit.
- Die lip: wide slit across the die that provides the final sizing of the melt.

Commercial sheet dies typically employ four features to control the flow to the die lip. These are the combined shape of the manifold and the island as well as the following three features shown in Fig. 1B:

- Choker bar: adjustable along with the width of the die and serves to tune the flow balance across the width of the die.
- Lower lip: sets the nominal sheet thickness.
- Flex-lip: adjustable along the width of the die and provides the final tuning to create uniform flow across the die.

In addition, sheet dies have die bolts that hold the upper and lower die plates together. The die plates are normally heated with cartridge heaters spaced along the width of the die. This type of die is typically made for a specific type of polymer to account for the shear thinning behavior of that polymer.[11] Consequently, the flow distribution in the die will change with melt viscosity, i.e., when the power law viscosity index of the resin changes. As the polymer grades change, flow adjustments can be made at the choker bar and at the flex lip, which both span the width of the die and can be adjusted at numerous points along the width. Clam shelling, or die deflection, is another cause of nonuniform flow across the die.[13,14] The higher pressures along the centerline of the die coupled with the lack of bolting to keep the die plates together cause the centerline of the die gap to widen. Clam shelling can increase as the throughput of the die increases because of higher die pressures. Thus, the flow balance across the die will be sensitive to production rates. Innovations in automatic flow adjustments have been made with designs like the Auto-Flex sheet die where the die lip gap at the flex lip is automatically adjusted by changing the length of the flex lip adjustment bolts.[15,16] The temperature-controlled bolts change length in response to cross-machine scanning of the sheet thickness.

FLAT FILM AND BLOWN FILM DIES

Dies used to make film less than 0.01 in. thick include flat, slit-shaped dies called T-dies and annular dies for blown film (Figs. 2 and 3). The design of the T-die is similar to the coat hanger-type die with the exception that the manifold and the land length are constant along with the width of the die. Consequently, the use of T-dies is often limited to coating applications with low-viscosity resins that resist thermal degradation, as the ends of the manifold in the T-die create stagnation pockets.[13] A common application for a film die is to coat a substrate like paper.

Blown film dies are the most common way of making commercial films. Because the blown film is so thin, weld lines are not tolerated, and the melt is typically introduced at the bottom of a spiral mandrel through a ring-shaped distribution system, as shown in Fig. 3. A series of spiral channels, cut into the mandrel-like multiple threads, smear the melt as it flows toward the die exit. This mixing action ensures that the melt is homogenous on exiting the die. Unlike other extrusion processes, blown film is sized and quenched from melt to solid film without contacting metallic cooling elements. The interior of the melt tube is pressurized with approximately 2 in. of water pressure. This pressure causes the tube to suddenly expand into a bubble as it exits the die. The tube forms a bubble because it is pinched overhead with nip rolls, which retain the air pressure. During the process of expanding, the melt tube undergoes an order of magnitude reduction in thickness and thus cools rapidly. This quenching moment occurs at the frost line of the bubble. The melt quenching occurs with a combination of external cooling air and internal bubble cooling air, as shown in Fig. 3. After the film passes through the nip rolls, it passes through a series of guide rollers to be wound up on to a roll.

PIPE AND TUBING DIES

Both pipe and tubing are made in dies with an annular die exit. A pipe product is defined as being greater than 1 in. in outer diameter and a tube less than 1 in. Dies for these products are made in two styles: 1) in-line dies (also called spider dies) shown in Fig. 4A and

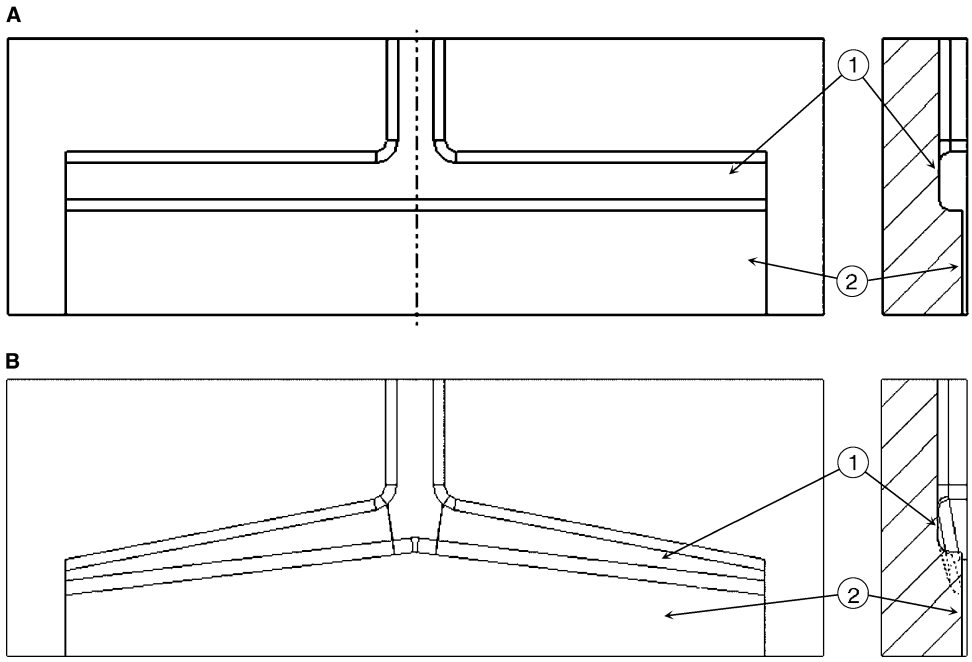


Fig. 2 Comparing designs of T-type die (A) [(1) constant cross-section manifold; (2) constant land length] to coat hanger-type die (B) [(1) manifold cross-section decreases as distance from centerline increases; (2) land length becomes shorter farther from the centerline of the die].

2) cross-head dies shown in Fig. 4B. The key elements of an in-line die are:

- Housing: mounts onto the end of the extruder, provides a circular passage through which the melt flows; it supports the mandrel and retaining ring.
- Mandrel (Torpedo): suspended in the center of the circular passage in the die body with metal bridges called spiders (typically three are used). One spider allows for passage of air into the center of the torpedo, is streamlined to avoid flow stagnation, and supports the die pin.
- Die pin: mates with the torpedo to provide streamlined sizing to the final inner diameter of the melt tube leaving the die; it has an air hole running through it to allow air to pass through the die body to the interior of the melt tube. A slight positive air pressure may be used to keep the inner diameter of the tubular extrudate from collapsing on exiting the die.
- Die land: forms the outer diameter of the tubular extrudate, held in place with a retaining ring and position adjusted with centering bolts. The die land can be changed to create a tube of a different diameter or wall thickness while keeping the original die pin.
- Retaining plate: secures the alignment of the die land with the die pin, bolted to the die body.
- Heater band: closely fitted to the housing (and for larger pipes to the exposed portion of the die pin) to ensure that the die is held at a temperature close to the required temperature of the melt.
- Flange for extruder attachment: tapered flange to permit alignment and attachment to extruder with split locking collars.

The in-line die is the least costly for manufacture of the two designs but can create defects called weld lines in the product. Weld lines occur because the melt is split and rejoined as it passes over the spider legs. A cross-head die can overcome this problem by eliminating the spiders. The melt enters the side of the die and turns 90° as it flows through a coat hanger-type passage that is wrapped around the mandrel. Key elements of a cross-head die that are different from an in-line tubing die are (see Fig. 4B):

- Core tube: mandrel with coat hanger-type passage that splits the flow and uniformly distributes melt along the annulus between the die pin and the die land.
- Side feed: melt enters from the side of the die and flows around the mandrel.
- Air supply: in-line with the die pin support.

Another application for the cross-head-style tubing die is wire coating. The following adjustments are made to a cross-head tubing die to perform wire coating:

- First, instead of passing air through the core tube, a bare conductor wire is pulled through the die entering the core tube inlet and exiting the die pin.

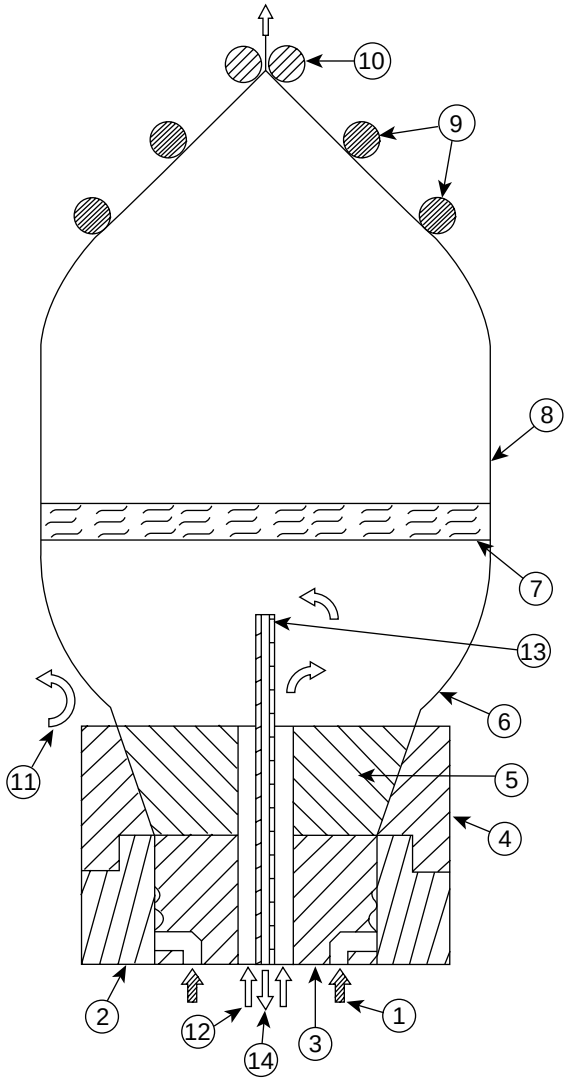


Fig. 3 Schematic of spiral mandrel blown film die operation: (1) ring-shaped melt distribution; (2) die body; (3) spiral flow mandrel; (4) sizing ring; (5) spreader; (6) film bubble; (7) frost line; (8) solidified film; (9) bubble collapsing rollers; (10) nip rollers; (11) external bubble cooling air; (12) internal bubble cooling air inlet; (13) internal bubble cooling pipe; and (14) heated internal bubble air return.

- Second, the length of the die pin is shortened to cause the wire to contact the melt tube before it exits the die land.

PROFILE EXTRUSION DIES

Profile extrusions are the most difficult to make because changes in take-up speed or screw rotational speed alone are not enough to compensate for deficient product dimensions. In the case of sheet and film, if the edges of the sheet are not at the target thickness, they can be trimmed off and sent back to the extruder to be reprocessed. Profile extrudates are significantly affected by nonuniform die swell unlike sheet and tube products. In the case of profiles with corners and other irregularities, like a square profile, the die exit needed to achieve a square profile is not square owing to the influence of die swell. Fig. 5 illustrates a die exit required to achieve a square extrudate. Note that the corners have an acute cusp shape and the side walls are not flat. A melt exiting this required but nonorthogonal shape will swell into a desired, orthogonal square shape. The design of nonorthogonal die exit required to achieve orthogonal profiles is addressed later in this entry in the section Modern Design Simulation and Computational Tools.

Open profile dies are typically shapes, such as ''U-shaped'' or ''L-shaped'' channels, that are not axisymmetric, unlike tube shapes. Consequently, open profiles are more prone to cooling unevenly and thus may generate residual stresses in the solidified (frozen) extrudate that cause the product to bow. A critical design rule for open profiles is to maintain a uniform wall thickness throughout the product cross section. Examples of poor and better profile designs are shown in Fig. 6 with the poorly designed sections shown on the left-hand side and the improved designs on the right-hand side. The difficulty with the original designs of both profiles A and B is the nonuniform wall thickness. The thinner sections will solidify first with the thicker sections still remaining molten in some core area. The result will be additional thermal shrinkage in the thicker regions and thus warpage of the final product. Product A will warp downward and product B will warp toward the right. These warping problems can be alleviated by making the entire cross section with more uniform thickness. Then, the entire cross section will solidify more uniformly and a little residual stress will be trapped in the solid extrudate. Design A illustrates a case where a hollow profile is used to solve the warpage problem whereas design B illustrates the use of an open profile to replace the thick region. The revised product designs of A and B will require more expense to fabricate the dies for these products as a set of mandrels must be made to form the hollow chambers. However, there are key benefits derived by making these changes:

- Better-quality products due to more uniform cooling and shrinkage: straighter products.
- Less material use by removing thick, unnecessary regions: savings of material costs.
- Faster cooling rates due to less hot plastic to cool: higher production rates.

Profile dies are commonly made with a series of plates that are stacked together to form a complex

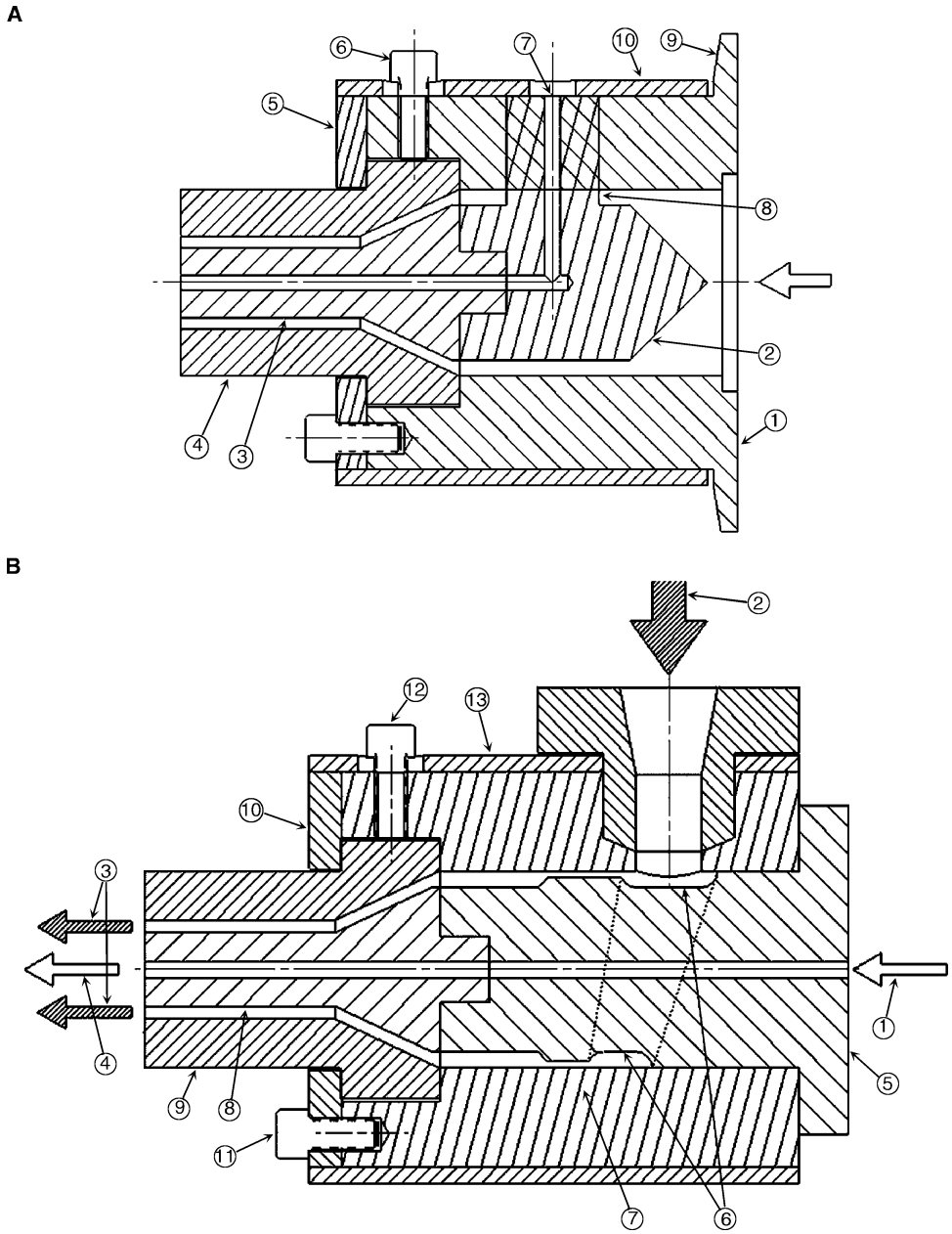


Fig. 4 Schematic of in-line tubing die (A): (1) housing; (2) mandrel (torpedo); (3) die pin (interchangeable); (4) die land (interchangeable); (5) retaining plate; (6) die centering bolt; (7) air hole; (8) mandrel support (spider leg); (9) die flange (mount to extruder with split clamp); (10) heater band; and cross-head tubing (or wire coating) die (B): (1) air or wire conductor inlet; (2) melt inflow (side inlet); (3) melt exit (annulus); (4) air or wire conductor exit; (5) core tube; (6) flow splitter; (7) housing; (8) die pin; (9) die land; (10) retaining plate; (11) retaining ring bolt; (12) die centering bolt; (13) heater band.

passage from the circular exit of the extruder to the required profile die exit. A stacked plate design makes for easier manufacture and permits adjustments to parts of the die assembly as needed during extrusion trials to fine-tune the die flow. An example of a stack plate die that makes a U-shaped profile is shown in Fig. 7. This figure illustrates an exploded view of a stack plate die, a cross-sectional view of the assembled die, and a detail of the die exit compared to the target profile. Stacked plate profile dies typically have these elements:

- Adapter plate: forms transition from circular extruder exit to approximate profile shape.
- Transition plate: forms streamlined transition from adapter plate exit to preland plate inlet.

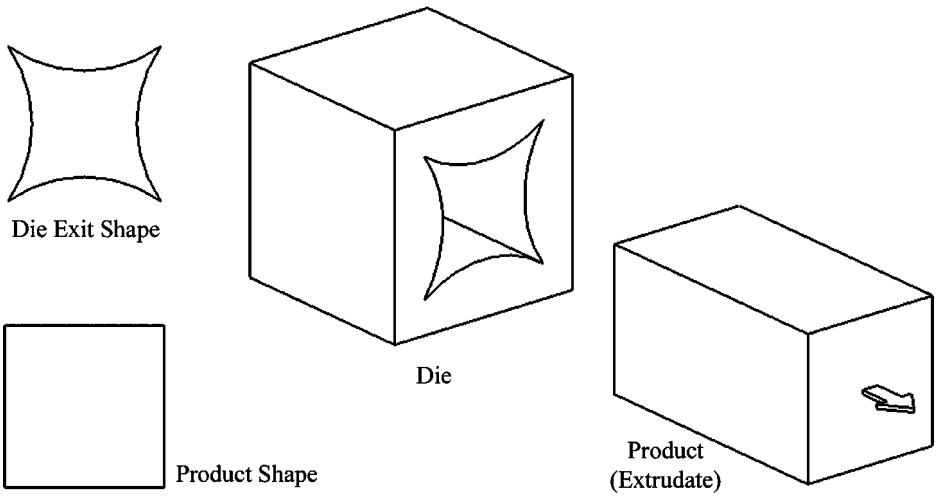


Fig. 5 Irregular die shapes required for regular extrudates.

- Preland plate: imparts significant flow adjustment by reducing thickness in high-flow areas and increasing thickness in low-flow areas anticipated downstream in the die land to make flow more uniform.
- Die land plate: provides a uniform cross-section passage that is typically 10 times longer than the thickness of the extrudate to relax the viscoelastic stresses in the melt before leaving the die (reduces die swell) and forms the shape of the extrudate leaving the die. The die land profile has the required shape to compensate for extrudate deformation after the die (die swell and drawdown) and yield the desired shape downstream.

The die exit profile shown in Fig. 7 creates an extrudate that is a U-shaped profile with three sides of uniform thickness and perpendicular walls to the bottom surface. The irregular shape of the die exit was generated with the aid of CFD as outlined in the section Modern Design Simulation and Computational Tools later in this entry.

COEXTRUSION DIES

Another important product made with extrusion dies is the creation of multilayered materials. Multilayered sheet and film materials have two applications:

- Making more economical material by sandwiching a less costly core material between two more expensive materials.

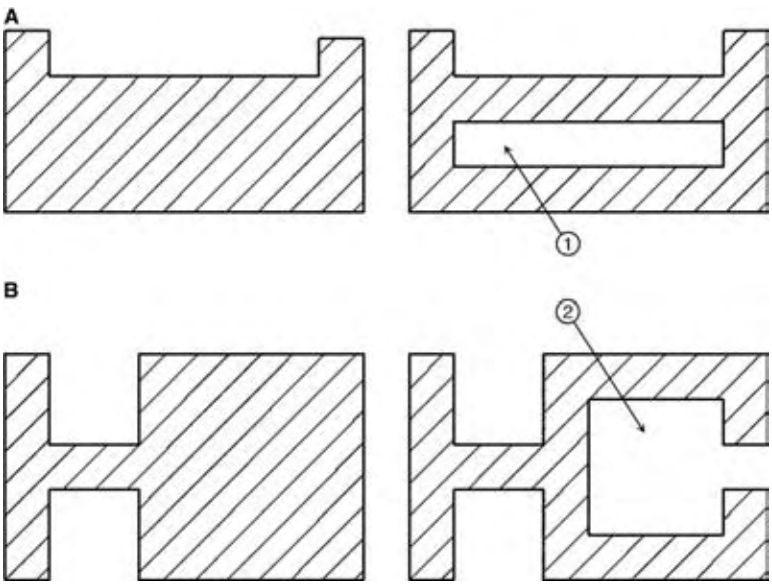


Fig. 6 Examples of poor (on left) and improved extrusion product designs (on right) to achieve uniform product thickness: (1) profile (A) made into a hollow profile and (2) profile (B) made into an open profile.

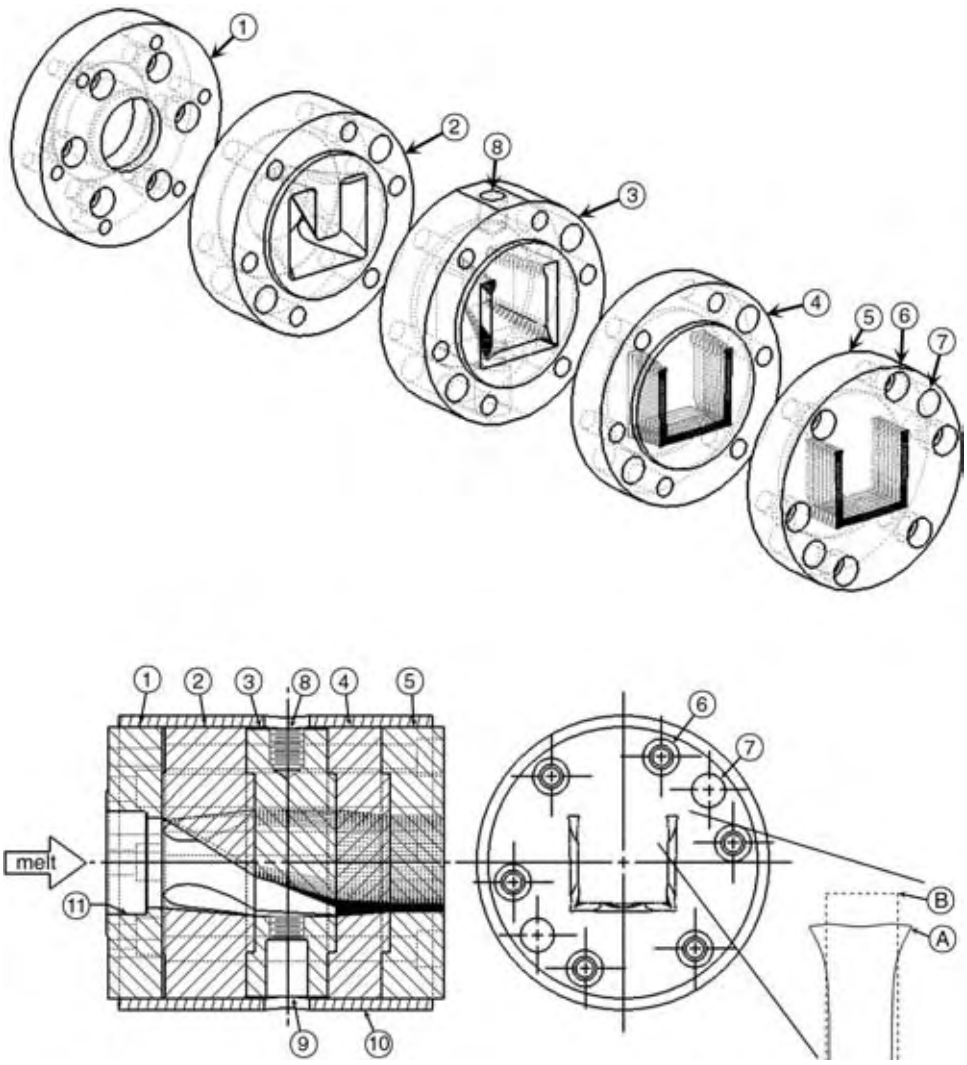


Fig. 7 U-Profile stack die: exploded view (top); section view (lower left); and end view (lower right): (1) extruder mounting plate; (2) die adapter plate; (3) transition plate; (4) preland plate; (5) die land plate; (6) die bolt hole; (7) alignment dowel pin hole; (8) thermocouple well; (9) pressure transducer port; (10) heater band; (11) breaker plate recess. Detail (lower right): (A) die exit profile and (B) product profile.

- Creating a composite material with improved properties by combining two or more materials that each posses a desirable property.

Applications of coextruded material include:

- Sheet stock made with an acrylic topcoat over acrylonitrile–butadiene–styrene (ABS). The acrylic provides UV resistance and gloss while the ABS provides impact resistance.
- Blown film with special barrier properties: one layer limits oxygen migration through the film and another provides protection from UV radiation.

There are two common methods of achieving co-extruded materials: feed block manifolds and multimanifolds within dies.[17] Two, three, or more melt streams may be combined with co-extrusion dies. The key elements of a feed block manifold as shown in Fig. 8A are:

- Inlet ports for the upper layer, middle layer, and lower layer.
- Streamlined melt lamination area that channels separate flow streams into one laminated melt stream inside feed block.
- Adapter plate between the feed block and the sheet die.
- Sheet die, which is identical to a monolayer die. The laminated melt stream enters the center of the die and spreads out along the manifold flowing out of the die exit as a distinct multilayer extrudate.

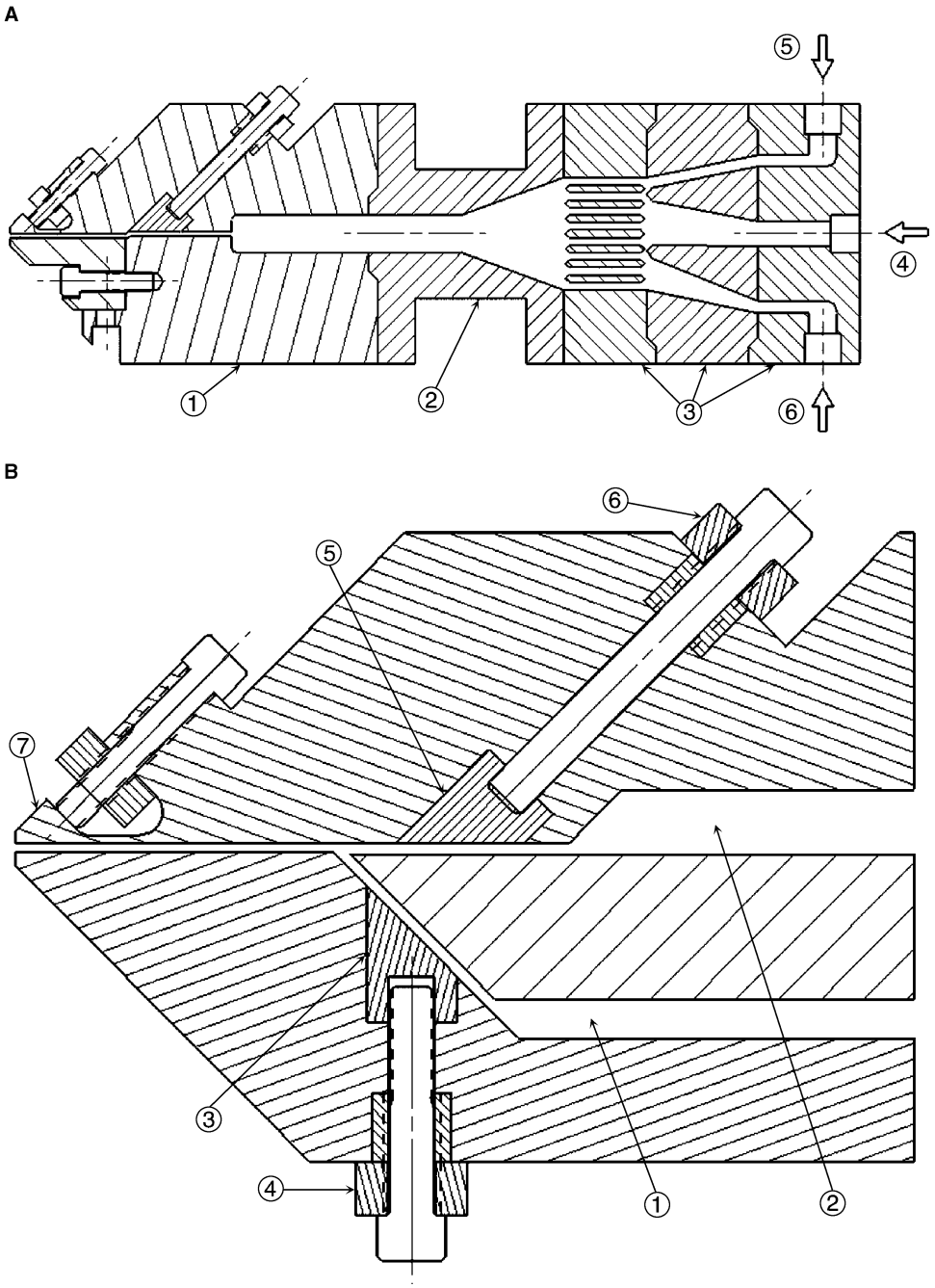


Fig. 8 Coextrusion feed block manifold and sheet die (A): (1) sheet die with flow restriction; (2) adapter plate; (3) feed block asembly; (4) core material layer inlet; (5) upper material layer inlet; (6) lower material layer inlet; and coextrusion multimanifold sheet die (B): (1) lower melt channel; (2) upper melt channel; (3) lower choker bar; (4) lower choker bar adjustment bolt; (5) upper choker bar; (6) upper choker bar adjustment bolt; (7) flex lip.

An alternative to the feed block design is a multimanifold die as depicted in Fig. 8B. The key elements of this design are:

- It is similar to a monolayer extrusion die, except that there is more than one feed channel.
- Each melt channel has its own choker bar for flow control.
- Melt streams converge inside the die near the die exit and emerge as a distinct multilayer extrudate.

The feed block technique is cheaper to implement than the multimanifold approach, but because the melt streams travel some distance before reaching the die exit, irregular flow patterns can develop at the interface of the different melt streams.[18–20] This is especially

true when attempting to coextrude melts of significantly differing viscosities. Lower-viscosity melt tends to encapsulate the more viscous melt. The alternative method is to keep the melt streams separated until just before the die exit as is done with the multimanifold design. Multimanifold designs permit coextrusion of plastic melts having significantly differing viscosities. With any coextrusion process, however, there must exist basic chemical compatibility between the neighboring melt streams to ensure good cohesion between the layers.

EXTRUDATE COOLING AND SIZING HARDWARE

With the exception of blown film and strand profiles, all extrudates require cooling and/or sizing by some metallic element. Table 2 summarizes the type of cooling and sizing hardware used for the various extrusion products.[21] In the case of sheet extrusion, the cooling is achieved with a chill roll stack, schematically illustrated in Fig. 9. The chill rolls are typically highly polished chrome-plated rollers that impart the surface gloss on sheet products and cool the extrudate while pulling the melt away from the die with a constant take-up speed. The average sheet thickness is achieved by the combination of extrusion screw rotational speeds and take-up speed adjustments. If the line speed taking the extrudate melt away from the die is greater than the average die exit speed, the thickness of the sheet decreases. This is called drawdown.

In the case of pipe and tubing products, the nominal outer and inner diameters of the extrudate are made by selecting the appropriate size of the die pin and die land. The final outer dimension of tubing products is achieved with sizing rings that are typically placed in the vacuum water bath, shown schematically in Fig. 10. The outer diameter of the tube is set with the sizing ring as the vacuum, which combined with a slight positive pressure inside the tube, forces the extrudate against the inner race of the ring. The desired inner diameter of the tube is achieved by adjusting the take-up speed of the extrudate relative to the average die exit speed. If the take-up speed is greater than the average die exit speed, the cross-sectional area decreases. Because the outer diameter of the tube is set with a sizing ring, the inner diameter will increase. Thus, the wall thickness of a tube is controlled this way.

The sizing and cooling of profiles have special requirements because of their complex shape.[21,22] These devices are called calibrators and are often as complicated as the die. To maintain the shape of a profile, vacuum is applied while simultaneously cooling the extrudate. Some calibrators, called wet vacuum calibrators, alternately inject water between the extrudate and the calibrator to lubricate and augment the cooling. A schematic of the dry vacuum calibrator setup used to size and cool the U-shaped profile is shown in Fig. 11. Fig. 11B illustrates a partially disassembled vacuum calibrator to reveal the following details: the vacuum channels and the cooling lines that simultaneously hold the moving extrudate in the desired shape while cooling it. Vacuum calibrators for profiles are typically made of stainless steel to withstand the abrasive action of extrudates while in contact with the calibrator during the cooling process. For example, the U-profile shown in Fig. 11 will tend to shrink onto the core feature of the lower calibrator and pull away from the upper calibrator surfaces. This complicates the design of the calibrator as the extrudate will conform to the ideal calibrator shape as it deforms. The cooling of the extrudate also complicates the simulation and design of the calibrator as discussed in the next section.

Table 2 Cooling and sizing devices for various extruded products

Product type	Cooling and sizing device
Film and sheet	Chill roll
Blown film	External and internal bubble air
Profiles—strand	Water tank
Profiles—open and hollow chamber	Vacuum calibrators and water tank
Tubing and pipes	Sizing rings and vacuum water tank

MODERN DESIGN SIMULATION AND COMPUTATIONAL TOOLS

The development of powerful computing hardware and proficient numerical techniques makes it possible now to simulate, analyze, and optimize three-dimensional extrusion processes with complex geometries, including nonlinear and viscoelastic polymer behavior. Numerical simulation has the potential to uncover important interior details of the extrusion process, such as velocity, shear stress, pressure, and temperature fields in the region of interest, which is not possible to do experimentally. A critical challenge for simulation methods is the ability to accurately represent the complex viscoelastic polymer material behavior that is dependent on process parameters, like shearing flow rate and temperature. Another challenge is to accurately model the complex geometry and the boundary conditions of extrusion dies and calibrators, especially for profile extrusions.

Experts in the polymer processing field have cited that the increasing complexity of product designs, coupled

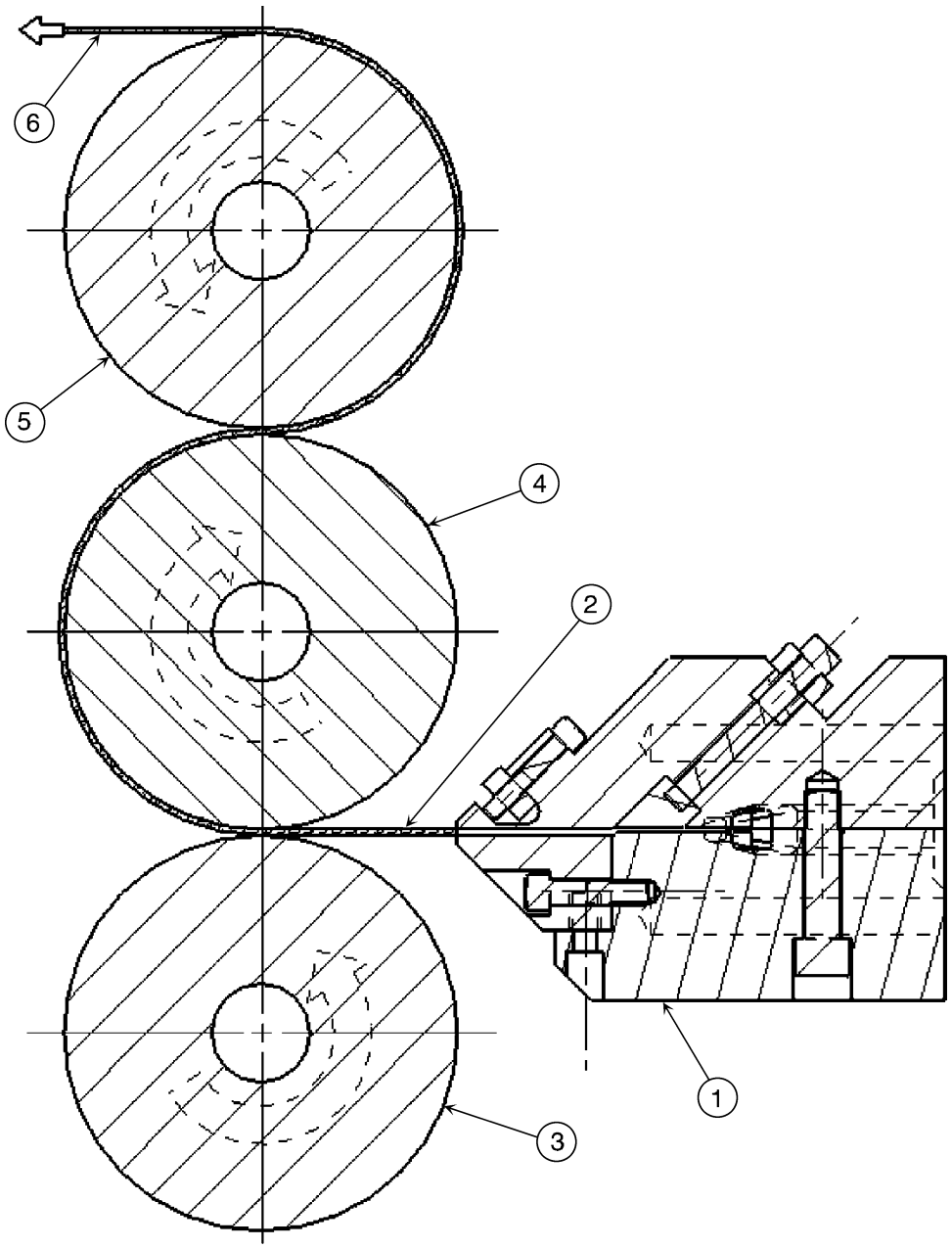


Fig. 9 Chill roll stack for sheet extrusion: (1) sheet die with flow restrictor; (2) molten sheet extrudate; (3) lower chill roll (all chill rolls temperature controlled); (4) middle chill roll (imparts gloss/surface texture to sheet); (5) upper chill roll; and (6) solidified sheet.

with the shorter development times, and a shortage of qualified engineers fuel the need for more process simulation in industry.[23] As already stated, several commercial polymer flow simulation programs are used for profile die design.[7–10] However, because the cooling rate of the extruded product determines the speed of the extrusion line, optimal design of a calibrator is also critical to productive operations. In addition, the design of the calibrator has an influence on the straightness of the final product because of uneven cooling results in unfavorable thermal deformations and warped products.[24]

Analytical solutions have been developed to aid the design of calibrators for simple shapes, such as sheets and pipes.[22] More complex shapes, such as window profiles, require the use of numerical finite element methods that can model arbitrary shapes.[25]

Computational Fluid Dynamics Simulation of Polymer Flow for Die Design

The design process begins with a target product shape. The objective of the simulation is to determine

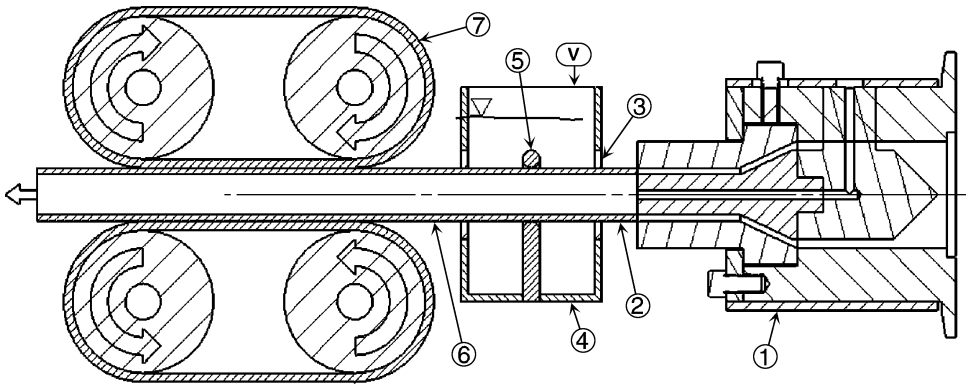


Fig. 10 Tubing vacuum water-bath calibration and take-off: (1) tubing/pipe die; (2) molten tube extrudate; (3) baffle; (4) vacuum water tank; (5) sizing ring; (6) solidified tube; and (7) puller.

the required die passage that results in a balanced mass flow exiting the die and an extrudate shape downstream of the die that matches the target profile.

A commercial polymer flow simulation program was used to simulate the three-dimensional die flow and heat transfer through the U-profile die, shown in Fig. 12.[7,26] Because the last two die plates have the greatest influence on the extrudate profile shape, only these two plates are designed with flow simulation. Simulation requires the following inputs:[7]

1. Geometric model of the die passage
   a. Two-dimensional profile of the inlet plane of the passage
   b. Two-dimensional profile of the target extrudate shape.

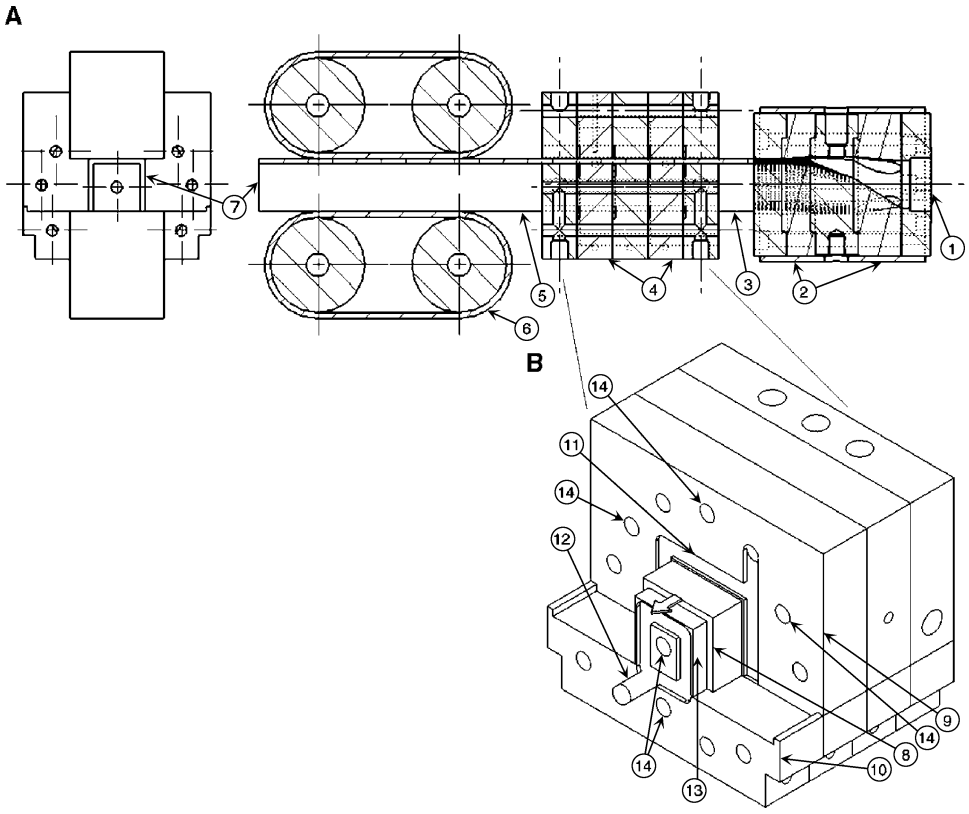


Fig. 11 Profile vacuum calibration and take-off. (A) Section view of calibration process: (1) melt enters profile die; (2) profile die stack; (3) molten profile extrudate; (4) calibrator (cools, shapes, and sizes extrudate); (5) solidified plastic; (6) puller; (7) orientation of profile. (B) Partially disassembled calibrator: (8) profile passing through calibrator; (9) upper calibrator stack; (10) lower calibrator stack; (11) upper vacuum channel; (12) lower vacuum channel; (13) core feature of lower calibrator; (14) cooling line.

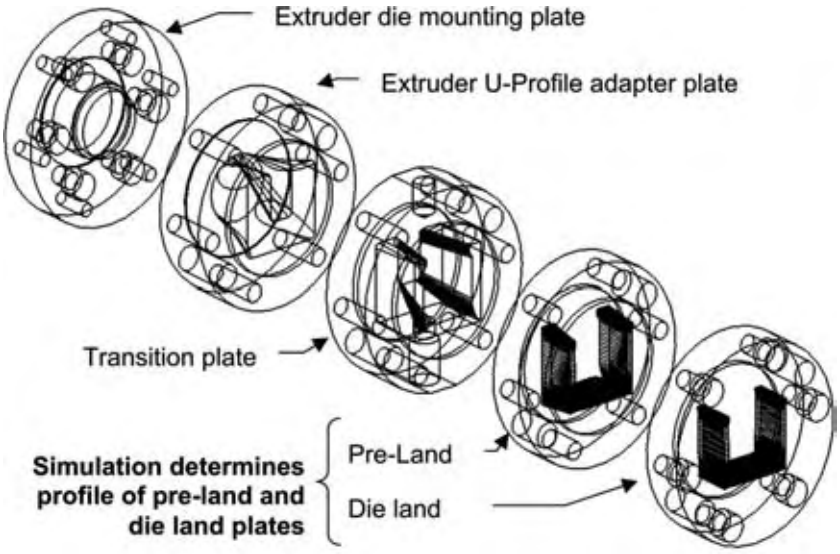


Fig. 12 U-Profile die plates designed with CFD simulation.

c. Specification of the preland length, the die land length (note that die land has a constant cross-sectional profile), and the free surface flow length after the die exit.
2. Thermomechanical properties of the polymer melt: density, heat capacity, and thermal conductivity.
3. Rheological properties of the polymer melt: non-Newtonian viscosity as a function of shear rate and temperature and/or viscoelastic material characteristics.
4. Process conditions: inlet melt temperature, mass flow rate into the die passage (or pressure at the inlet), die wall temperature, and take-up speed of the extrudate downstream from the die.

The computational domain resembles the real three-dimensional die geometry and a free surface flow after the die, where velocity redistribution (equalization) and stress relaxation take place in a short distance downstream from the die exit (Fig. 13A). Because of the symmetry of the die design, only half of the die passage is modeled (Fig. 13B). A finite element model of the die passage and the free surface region consists of 16,592 hexahedral elements and 19,530 nodes, as detailed elsewhere.[26] The computational domain must have appropriate boundary conditions to represent the realistic conditions present as the melt passes through the die and exits into a free surface flow (Fig. 13C). The used commercial CFD program implements an ''inverse extrusion'' solution algorithm, which computes the shape of the die exit (die land profile) required to achieve the target profile dimensions at the exit of the free surface domain.[7] The program solves for the shape of the die land that will achieve the target profile after die swell occurs.[7,26]

Cooling Simulation and Calibrator Design

For optimal design of the profile extrusion calibrators, cooling bath, and other cooling accessories, a comprehensive knowledge about the extrudate heat transfer process (cooling) is necessary. The biggest challenge in modeling of profile extrudate cooling is to specify properly the boundary conditions in every local part of the cooling equipment. It is possible to approximate heat transfer coefficients or determine the values experimentally.[27,28] It may be very difficult to estimate the heat transfer coefficient in a vacuum calibrator because it is not possible to predict, without experimental verification, where the polymer has a good contact with the cooling wall and what is the influence of a thin layer of cooling water being sucked in from the cooling bath. However, even estimated values can be used to get a good overall picture of the process, as the polymer materials have a fairly low thermal conductivity. This means that the obtained results are not exact, but they can be very useful for design. Therefore, the modeling and simulation of extrudate cooling is a useful tool for studying the profile extrusion cooling process, as well as for design improvement of the calibrators and other cooling equipment. Other researchers also indicated that calibration design can be done to estimate the cooling performance.[29,30] An illustration of the type of information that can be obtained by simulating the cooling of the extrudate is shown by cooling simulation of a U-profile extrudate using a commercial simulation program and

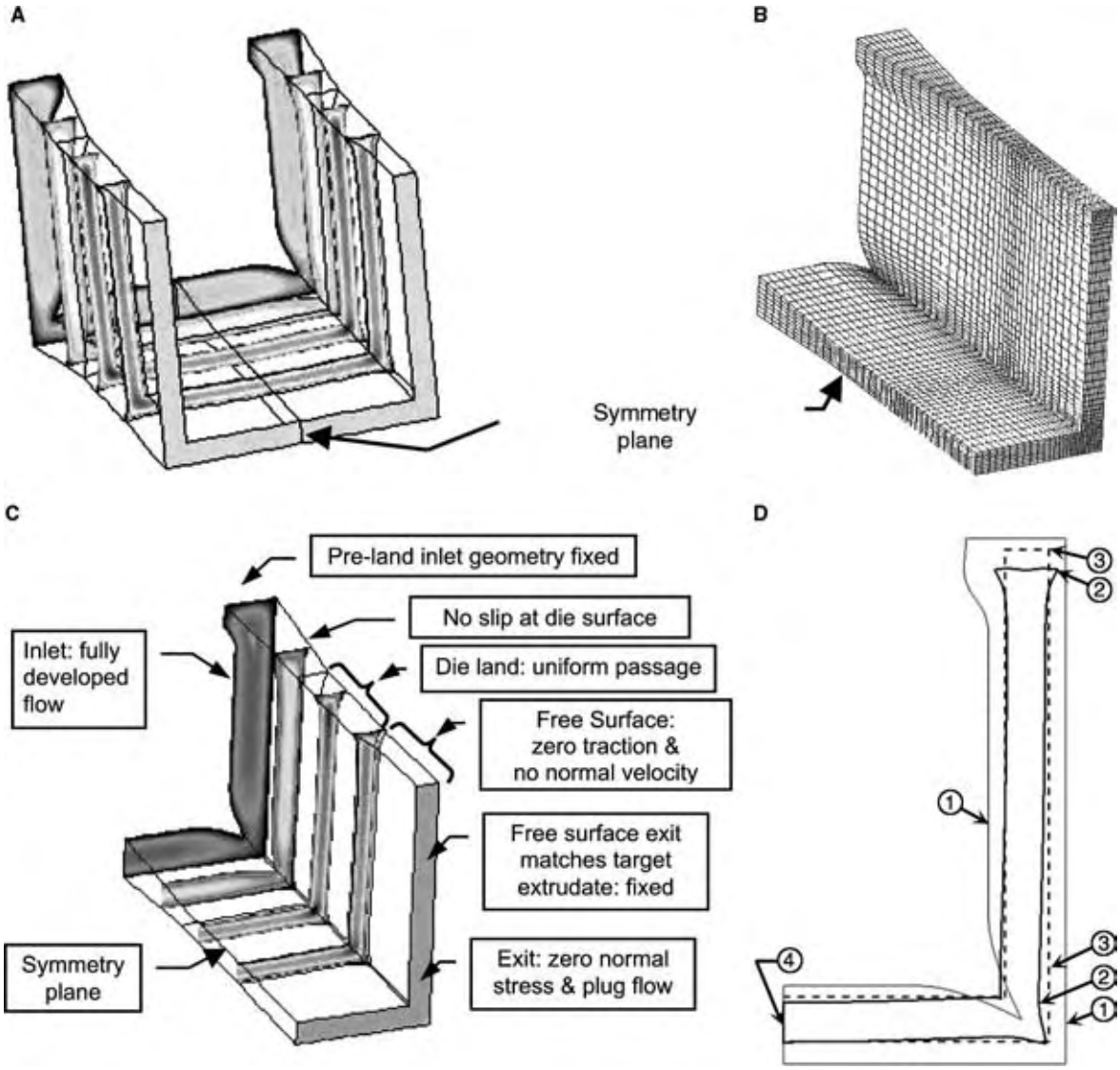


Fig. 13 Computational model for U-Profile die design: (A) preland, die land, and free surface as computational domain; (B) finite element mesh (symmetry exploited to reduce computational requirements); (C) boundary conditions for simulation of polymer flow through die and extrudate free surface; and (D) relevant profiles: (1) preland inlet; (2) die land (uniform along flow length); (3) final free surface (target extrudate profile); and (4) symmetry plane.

experimentally determined heat transfer coefficients in a vacuum calibrator.[26]

CONCLUSIONS AND DESIGN RECOMMENDATIONS

As stated above, extrusion die design is a complex task because the extrudate product dimensions depend not only on the die design (die shape), but also on the polymer properties and extrusion process parameters. The following are general recommendations for extrusion die design:

- Achieve a balanced melt flow exiting the die.
- Minimize the pressure drop required to achieve a balanced flow to permit the maximum mass flow rate with the smallest-sized extruder required.
- Provide flow control devices in the die to optimize the flow distribution.
- Streamline the die flow passage to avoid flow stagnation areas. Such areas facilitate degradation of the polymer melt due to prolonged exposure at elevated temperatures.
- Use modular design with stacked plates for manufacturability, convenient assembly, and disassembly, as well as convenient modifications and cleaning.
- Die land length should be at least 10 times the product thickness (or gap) to facilitate the polymer melt stress relaxation within the die.

- Avoid thick and nonuniform extrudate wall thickness to achieve better flow balance control in the die, minimize material use, reduce cooling times, and minimize postextrusion warping of the product.
- Avoid or minimize hollow profiles as they increase die fabrication costs and complicate the cooling process of the extrudate.

Except for circular dies, it is virtually impossible to design a die geometry to achieve a quality extrudate product for a wide range of polymers and extrusion process conditions. That is why a good die design must incorporate appropriate adjustment features to be set (or tuned) during the extrusion process to compensate for the deficiency of the final product, i.e., the cooled extrudate. For a fixed die geometry, adjustment of the deficient profile may be achieved by changing extrusion process parameters, like temperature, flow rate, cooling rate, and/or take-up speed. However, it is important to optimize the die design to make the necessary adjustments practically possible. This is why polymer extrusion die design has most often relied on experience, empirical data, and expensive trial and error adjustments to design and optimize a die and complementary process parameters. However, by integrating computational simulation with empirical data and by improving the extrusion monitoring instrumentation the die design process can be improved. A better die design method yields improved product quality and a reduction in the time to design and optimize the extrusion process, resulting in lower costs. It is important to state again that computational simulation and empirical extrusion engineering are synergistic in nature. They have their exclusive strengths and weaknesses that cannot replace each other, but, if properly integrated, may significantly improve extrusion die design.

REFERENCES

1. Tadmor, Z.; Gogos, C.G. Die forming. In *Principles of Polymer Processing*; John Wiley & Sons, 1979; 521–524.
2. Rosato, D.V. Die design and performance. In *Extruding Plastics*; Chapman & Hall, 1998; 228–282.
3. Rauwendaal, C. Die forming. In *Understanding Extrusion*; Hanser, 1998; 107–109.
4. Tadmor, Z.; Gogos, C.G. Polymer melt rheology. In *Principles of Polymer Processing*; John Wiley & Sons, 1979; 148–172.
5. Michaeli, W. Monoextrusion dies for thermoplastics. In *Extrusion Dies for Plastics and Rubber*, 2nd Ed.; Hanser, 1992; 195–198.
6. Woei-Shyong, L.; Hsueh-Yu, H. Experimental study on extrudate swell and die geometry of profile extrusion. J. Polym. Eng. Sci. **2000**, *40* (5), 1085–1094.
7. Polyflow (application software); Fluent Inc.: Lebanon, NH; http://www.fluent.com/software/polyflow (accessed Mar 2005).
8. Flow 2000; Compuplast International: Zlin, Czech Republic; http://www.compuplast.com/FLOW_2000.htm (accessed Mar 2005).
9. Dieflow: Chippewa Falls WI; http://www.dieflow.com (accesses Mar 2005).
10. HyperXtrude, Altair Engineering, Inc.: Troy, MI; http://www.altair.com/software/hw_hx.htm (accessed Mar 2005).
11. Michaeli, W. Monoextrusion dies for thermoplastics. In *Extrusion Dies for Plastics and Rubber*, 2nd Ed.; Hanser, 1992; 128–194.
12. Levy, S.; Carley, J.F. Extrusion dies for specific product lines. In *Plastics Extrusion Technology Handbook*, 2nd Ed.; Industrial Press, Inc., 1989; 96–139.
13. Extrusion Dies Industries, LLC: Chippewa Falls, WI; http://www.extrusiondies.com (accessed Mar 2005).
14. Michaeli, W. Monoextrusion dies for thermoplastics. In *Extrusion Dies for Plastics and Rubber*, 2nd Ed.; Hanser, 1992; 147–148.
15. Levy, S.; Carley, J.F. On-line and computer control of the extrusion process. In *Plastics Extrusion Technology Handbook*, 2nd Ed.; Industrial Press, Inc., 1989; 302–304.
16. Tadmor, Z.; Gogos, C.G. Die forming. In *Principles of Polymer Processing*; John Wiley & Sons, 1979; 533–537.
17. Michaeli, W. Coextrusion dies for thermoplastics. In *Extrusion Dies for Plastics and Rubber*, 2nd Ed.; Hanser, 1992; 234–238.
18. Levy, S.; Carley, J.F. Extrusion dies for specific product lines. In *Plastics Extrusion Technology Handbook*, 2nd Ed.; Industrial Press, Inc., 1989; 228–233.
19. Gifford, W.A. A three-dimensional analysis of coextrusion in a single manifold flat die. J. Polym. Eng. Sci. **2000**, *40* (9), 2095–2100.
20. Michaeli, W. Coextrusion dies for thermoplastics. In *Extrusion Dies for Plastics and Rubber*, 2nd Ed.; Hanser, 1992; 215–221.
21. Levy, S.; Carley, J.F. Extrusion dies for specific product lines. In *Plastics Extrusion Technology Handbook*, 2nd Ed.; Industrial Press, Inc., 1989; 188–213.
22. Michaeli, W. Calibration of pipes and profiles. In *Extrusion Dies for Plastics and Rubber*, 2nd Ed.; Hanser, 1992; 311–326.

23. Michaeli, W.; Pfannschmidt, O.; Franz, A.; Vogt, N. Pre-computing developments—progress report on process simulation in industry. Kunststoffe **2001**, *91* (7), 32–39.
24. Brown, R.J. *Predicting How the Cooling and Resulting Shrinkage of Plastics Affect the Shape and Straightness of Extruded Products*. Proceedings of the Annual Technical Conference of the Society of Plastics Engineers, May 4–8, 2000.
25. Sheehy, P.; Tanguy, P.A.; Blouin, D. A finite element model for complex profile calibration. J. Polym. Eng. Sci. **1994**, *34* (8), 650–656.
26. Reifschneider, L.G.; Kostic, M.K.; Vaddiraju, S.R. *Computational Design of a U-Profile Die and Calibrator*. Proceedings of the Annual Technical Conference of the Society of Plastics Engineers, Chicago, May 16–20, 2004.
27. Michaeli, W. Calibration of pipes and profiles. In *Extrusion Dies for Plastics and Rubber*, 2nd Ed.; Hanser, 1992; 324–329.
28. Fredette, L.; Tanguy, P.A.; Hurez, P.; Blouin, D. On the determination of heat transfer coefficient between PVC and steel in vacuum extrusion calibrators. Int. J. Num. Methods Heat Fluid Flow **1996**, *6*, 3–12.
29. Placek, L.; Svabik, J.; Vlcek, J. *Cooling of Extruded Plastic Profiles*. Proceedings of the Annual Technical Conference of the Society of Plastics Engineers, May 4–8, 2000.
30. Carneiro, O.S.; Nobrega, J.M.; Covas, J.A.; Oliveria, P.J.; Pinho, F.T. *A Study of the Thermal Performance of Calibrators*. Proceedings of the Annual Technical Conference of the Society of Plastics Engineers, Chicago, May 16–20, 2004.

Desulfurization

Chunshan Song
Uday T. Turaga
Xiaoliang Ma
Clean Fuels and Catalysis Program, The Energy Institute, and Department of Energy and Geo-Environmental Engineering, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.

INTRODUCTION

Desulfurization of hydrocarbon streams is an important process used in a petroleum refinery to reduce the sulfur concentration in fuels such as gasoline, jet fuel, kerosene, diesel, and heating oil so that the resulting fuels meet environmental protection standards.[1–7] Hydrotreating is one of the most popular and widely practiced desulfurization processes and refers to the catalytic removal of sulfur [hydrodesulfurization (HDS)], nitrogen [hydrodenitrogenation (HDN)], oxygen [hydrodeoxygenation (HDO)], and metals [hydrodemetallation (HDM)] from petroleum distillates in the presence of hydrogen. Hydrotreating first appeared in petroleum refineries as a finishing process in the 1930s.[2] Hydrotreating operates at conditions milder than those typically used in fluidized catalytic cracking (FCC) or hydrocracking. Typical hydrotreating process conditions vary with feedstock and are summarized in Table 1.

In the past two decades petroleum refining has changed extensively and the fortunes of hydrotreating, in particular, have witnessed a sea change. Hydrotreaters now occupy a central role in modern refineries and more than 50% of all refinery streams now pass through hydrotreaters for conversion, finishing, and pretreatment purposes.[2] Hydrodesulfurization is the largest application of catalytic technology in terms of the volume of material processed.[8] On the basis of usage volume, HDS catalysts are ranked third behind catalysts used for automobile emission control and FCC.[8]

Commercial hydrotreating catalysts are, typically, sulfides of Mo or tungsten (W) supported on γ-Al_2O_3 and promoted by either Co or Ni. Nickel, known for its high hydrogenation activities, is preferred as a promoter when feedstocks containing high amounts of nitrogen and aromatics need to be processed. Table 2 provides the compositional range and physical properties of typical hydrotreating catalysts in the oxidic phase.

GROWING DEMAND FOR DESULFURIZATION

The challenge of fulfilling the world's growing transportation energy needs is no longer a simple issue of producing enough liquid hydrocarbon fuels. This challenge is instead accentuated by a complex interplay of environmental and operational issues. Environmental issues include societal demands that liquid hydrocarbon fuels be clean and less polluting. The emergence of new refining processes and the increasing use of new forms of energy production, e.g., fuel cells, exemplify operational issues. Together, these trends are driving the need for deep desulfurization of diesel and jet fuels.

As an example of the kind of regulations that are being specified and contemplated for hydrocarbon fuels, Table 3 lists the compositional and performance properties for diesel. The United States Environmental Protection Agency (USEPA) has mandated that diesel—whose automotive use is now growing at a pace faster than that of gasoline—have no more than 15 parts per million by weight (ppmw) of sulfur by 2006.[9,10] This represents a 97% reduction in the allowable sulfur concentration in diesel from 500 to 15 wppm. The United States now allows up to 3000 wppm of sulfur in jet fuel.[11] With the European Union now demanding no more than 1000 wppm of sulfur in jet fuel, a reduction in the permissible sulfur content of U.S. jet fuel can be expected.[12] New gasoline sulfur regulations will require most refiners to meet a 30 ppmw sulfur average with an 80 ppmw cap for both conventional and reformulated gasoline by January 1, 2006.

Initially, fuel sulfur was regulated to reduce emissions of the oxides of sulfur, which contribute to acid rain, ozone, and smog. The recent and stricter round of sulfur specifications, however, are an effort to reduce automobile emissions of the oxides of nitrogen (NO_x) and particulate matter (PM). For example, the 15 ppmw diesel sulfur limit follows from the USEPA's parallel program of rule making that seeks to reduce automobile NO_x and PM emissions by 95% and 90%, respectively, by 2007. Automobile manufacturers are demanding ultra-low-sulfur fuels because only then would their advanced, sulfur-sensitive after-treatment technologies achieve such drastic reductions in NO_x and PM emissions.[13,14]

Sulfur specifications are the more visible drivers for desulfurization research. Fig. 1 presents a qualitative

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007732

Table 1 Typical hydrotreating process conditions for various feedstocks

Feedstock	Temperature (°C)	Hydrogen pressure (atm)	LHSV[a] (hr^{-1})
Naphtha	320	15–30	3–8
Kerosene	330	30–45	2–5
Atmospheric gas oil	340	38–60	1.5–4
Vacuum gas oil	360	75–135	1–2
Atmospheric residue	370–410	120–195	0.2–0.5
Vacuum residue	400–440	150–225	0.2–0.5

[a]LHSV, liquid hourly space velocity.
(From Ref.[2].)

relationship between the type and size of sulfur molecules in various distillate fuel fractions and their relative reactivities.[5] Various refinery streams are used to produce three major types of transportation fuels, gasoline, jet fuels, and diesel fuels, that differ in composition and properties. Other fuel specifications are equally important, albeit less visible, reasons for continued research and development in desulfurization. For example, European countries typically require diesel fuel with sulfur less than 50 wppm, cetane number of 48–51, and density less than 0.825 g/cm^3. These fuel specifications are more stringent than those enforced in the United States (see Table 3). To be active in the European diesel fuels market—larger than that in the United States—oil refiners will have to produce ultraclean, high-quality, premium diesel fuel through sophisticated refining processes such as the selective ring opening (SRO) of naphthenes.[15]

Ring opening is commonly observed in the hydrocracking process where one or more carbon–carbon (C–C) bonds are broken. Selective ring opening, in contrast, breaks only one C–C bond open, thus preventing extensive reduction in molecular weight.[16] This process, shown for a model diaromatic molecule in Fig. 2, leads to extensive improvement in the quality of diesel fuel by increasing cetane number and decreasing fuel density simultaneously.[16]

Hydrodesulfurization is important for the SRO process because the most effective SRO catalysts are based on noble metals such as iridium, which are highly sensitive to sulfur. Therefore, all SRO feedstocks will have to be extensively desulfurized and hydrogenated before being sent to an SRO reactor. Selective ring opening, while fulfilling important nonsulfur fuel specifications, is an emerging refining process that symbolizes a compelling operational issue that could require more active HDS catalysts.

In recent times, there has been tremendous interest in fuel cells. This interest is set to intensify with the U.S. government's new Freedom Cooperative Automotive Research (Freedom CAR) program that seeks to develop cars based on hydrogen-powered fuel cells.[17] Enough hydrogen to satisfy a fuel cell-based transportation system will have to be produced by processing hydrocarbon fuels through reforming and related processes all of which are sensitive to sulfur. While most fuel cells are also sulfur sensitive, some are intolerant to as little as 0.1 wppm sulfur, e.g., polymer electrolyte membrane fuel cell (PEMFC).[7] As the sulfur compounds in liquid hydrocarbon fuels and the H_2S produced from these sulfur compounds in the hydrocarbon reforming process are poisonous to both the catalysts in hydrocarbon fuel processor and the electrode catalysts in fuel cell stack, the sulfur content in the liquid hydrocarbon fuels has to be reduced to a very low level [<10 ppmw for solid oxide fuel cell and <0.1 ppmw for PEMFC].[7] Consequently, desulfurization will continue to be one of the most important processes in oil refineries.

Table 2 Composition and properties of typical hydrotreating catalysts

Composition and properties[a]	Range	Typical values
Active phase precursors (wt%)		
MoO_3	13–20	15
CoO	2.5–3.5	3.0
NiO	2.5–3.5	3.0
Promoters (wt%)		
SiO_2	0.5–1	0.5
B, P	0.5–1	0.5
Physical properties		
Surface area (m^2/g)	150–500	180–300
Pore volume (cm^3/g)	0.25–0.8	0.5–0.6
Pore diameter (nm)		
Mesopores	3–50	7–20
Macropores	100–5000	600–1000
Extrudate diameter (mm)	0.8–4	3
Extrudate length/diameter	2–4	3
Bulk density (g/ml)	0.5–1.0	0.75
Average crush strength/ length (kg/mm)	1.0–2.5	1.9

[a]Active phase precursors and promoters are supported on a γ-Al_2O_3 carrier.

Table 3 Specifications for diesel fuel

Parameter	U.S.A.		Europe		World Fuels Charter
	EPA 2006	CARB[a] average	Current	2005	
Sulfur (ppmw)	15	15	350	50[b]	30
Density (g/cm^3)			<0.845	(<0.825)	<0.84
T 90% (°C)	338	321			
T 95% (°C)		349	360	(<340)	340
Cetane number		48	51	(>56)	55
Cetane index	40				
Polyaromatics (%)[c]		1.4	11	(<1)	2
Total aromatics (%)	35	10			15

Values in parentheses have been proposed, and are not mandated yet.
[a]California Air Resources Board.
[b]Sulfur content in some European countries is much lower, e.g., Germany requires diesel to have sulfur no more than 10 ppmw.
[c]In this article, unless indicated otherwise, percentage compositions are weight based.

SULFUR COMPOSITION OF KEY HYDROCARBON FUEL

The major organic sulfur compounds in petroleum fractions are thiols (mercaptans), sulfides, disulfides, thiophenes, benzothiophenes (BTs), dibenzothiophenes (DBTs), naphthothiophenes (NTs), benzonaphthothiophenes (BNTs), phenanthro[4,5-b,c,d]thiophenes (PTs), and their alkyl-substituted derivates, as shown in Table 4. There are three major types of transportation fuels, gasoline, diesel, and jet fuels, which differ in composition and properties (Fig. 1, Table 4). The major sulfur compounds existing in commercial gasoline are thiophene, 2-methylthiophene, 3-methylthiophene, dimethylthiophenes, and benzothiophenes, indicating that most sulfur compounds with higher HDS reactivity, including thiols, disulfides, and sulfides, have been removed from the commercial gasoline in conventional petroleum refining processes such as FCC. The major sulfur compounds in the jet fuels such as, JP-8 are dimethylbenzothiophenes and trimethylbenzothiophenes. Most of them have two methyl groups at the 2- and 3-positions, respectively, implying that these alkyl benzothiophenes are more difficult to be removed than others. The sulfur compounds in commercial diesel fuel include alkyl benzothiophenes and alkyl dibenzothiophenes. Of these compounds, alkyl dibenzothiophenes with alkyl groups at the 4- and/or 6-positions are the most difficult to remove by conventional HDS processes.

REACTIVITY AND REACTION MECHANISM OF VARIOUS SULFUR COMPOUNDS

The major fundamental approaches of the numerous theoretical and experimental studies conducted on desulfurization since the 1980s can be grouped into four general areas: 1) determining reaction kinetics and inhibition, over existing and new catalysts; 2) catalyst characterization using spectroscopy; 3) studying coordination chemistry of organometallic complexes; and 4) employing molecular simulation and computational analysis.

The kinetics and poisoning investigations indicate that the sulfur compounds present in hydrocarbon fuels usually show different reactivities and mechanisms for undergoing desulfurization. The HDS reactivity of various sulfur compounds decreases in the order of disulfides > sulfides, thiols > thiophenes > BTs, NTs > BNTs, DBTs without any alkyl group at the 4- and/or 6-position > DBTs with one or two alkyl group(s) at the 4- and/or 6-position(s). For the sulfur compounds without a conjugation structure between the lone pairs on S atom and the π-electrons on the aromatic ring, including disulfides, sulfides, thiols, and tetrahydrothiophene, HDS occurs directly through the hydrogenolysis pathway:

$$R\text{–}S\text{–}H + H_2 \longrightarrow R\text{–}H + H_2S$$
$$R\text{–}S\text{–}R' + 2H_2 \longrightarrow R\text{–}H + H_2S + R'\text{–}H$$
$$R\text{–}S\text{–}S\text{–}R' + 3H_2 \longrightarrow R\text{–}H + 2H_2S + R'\text{–}H$$

These sulfur compounds exhibit higher HDS reactivity than that of thiophenic compounds, because they have higher electron density on the S atom and a weaker C–S bond. For the thiophenic compounds, in which the lone pairs on the S atom conjugate with the π-electrons on the ring, including thiophenes, BTs, DBTs, NTs, PTs, and BNTs, HDS over the commercial catalysts usually proceeds through two pathways, the hydrogenation pathway (hydrogenation followed by hydrogenolysis) and the direct hydrogenolysis pathway (direct elimination of S atom via C–S bond cleavage), as shown below.

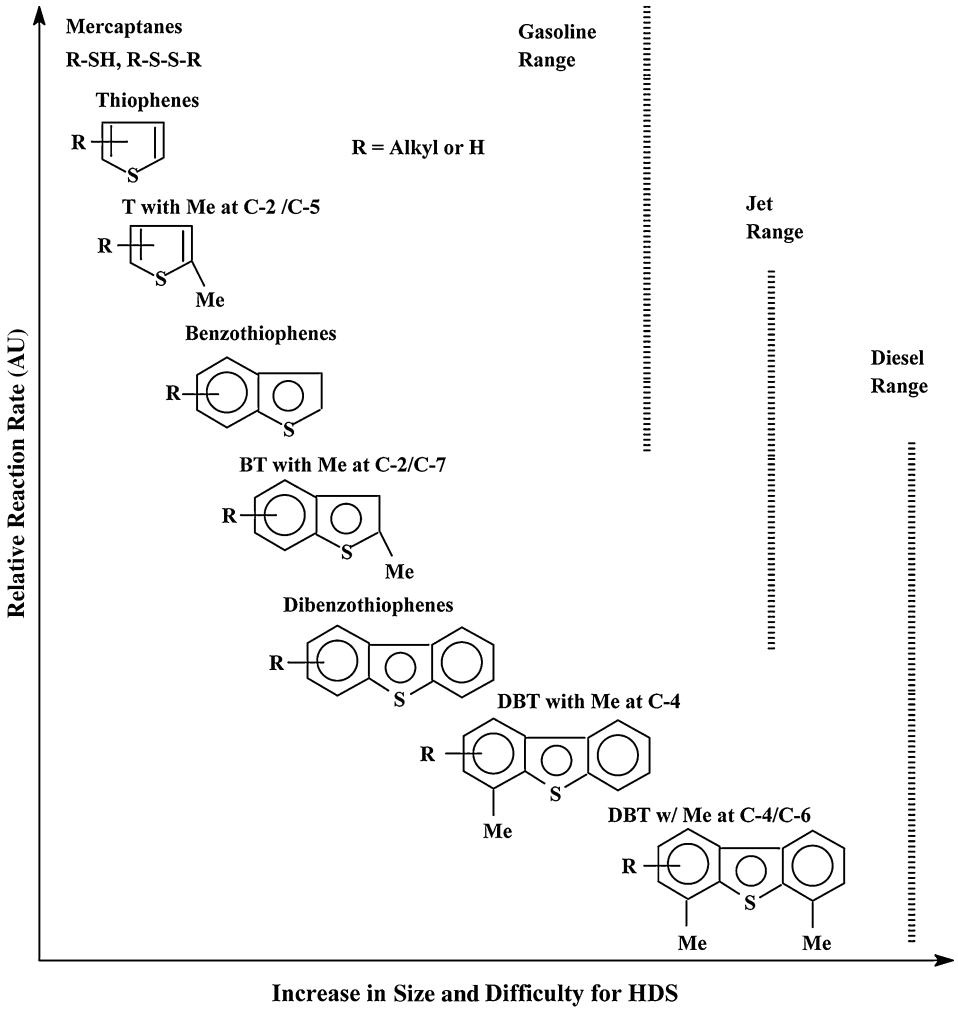


Fig. 1 Reactivity of various organic sulfur compounds in HDS vs. their ring sizes and positions of alkyl substitutions on the ring.

The HDS reactivity of the thiophenic compounds is dominantly dependent on both the electron structure and the steric hindrance of alkyl groups. For thiophene and BT, the total HDS reactivity is greater than that of DBT by about an order of magnitude because there is higher π-electron density at the C(2)–C(3) and C(4)–C(5) bonds in thiophene and at the C(2)–C(3) bond in BT, leading to their greater HDS reactivity through the hydrogenation pathway.[18] The π-electron distribution on dibenzothiophene is more uniform, as on a benzene ring, resulting in its lower hydrogenation reactivity. Thus, HDS of DBT over commercial Co–Mo catalysts proceeds dominantly through the hydrogenolysis pathway (see Scheme 1). However, if one or two alkyl groups are attached at the 4- or/and 6-positions of DBT, the hydrogenolysis pathway will be hindered strongly and the hydrogenation pathway becomes dominant.[3] In this case, the HDS reactivity will

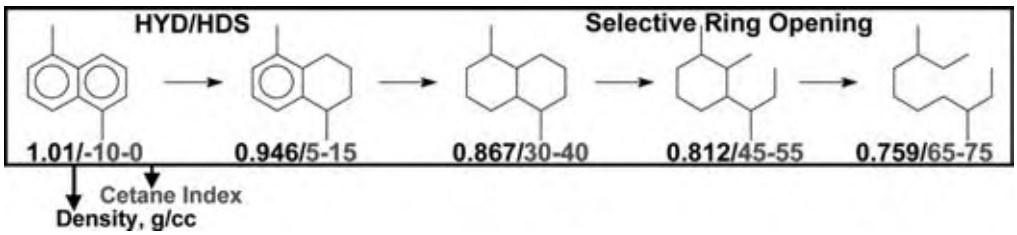
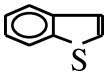
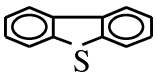
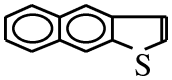
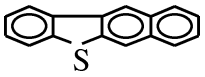
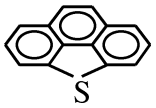


Fig. 2 Selective ring opening of a model diaromatic molecule. (From Ref.[16].) *(View this art in color at www.dekker.com.)*

Table 4 Some typical sulfur compounds in petroleum fractions

Thiols (mercaptans)	R–S–H
Sulfides	R–S–R
Disulfides	R–S–S–R
Thiophenes	
Benzothiophenes (BTs)	
Dibenzothiophenes (DBTs)	
Naphthothiophenes (NTs)	
Benzonaphthothiophenes (BNTs)	
Phenanthro[4,5-b,c,d]thiophenes (PTs)	

decrease significantly. These sulfur compounds are called as the refractory sulfur compounds.

Some coexisting compounds in petroleum fractions and in the products during the HDS process exhibit a strong effect on HDS. H_2S was found to be one of the main inhibitors of the hydrogenolysis pathway, while polyaromatic compounds were found to be the main inhibitors of the hydrogenation pathway.[3] Basic nitrogen compounds affect both the hydrogenolysis pathway and the hydrogenation pathway. Some excellent reviews of this aspect of HDS research are available in the literature.[1–6]

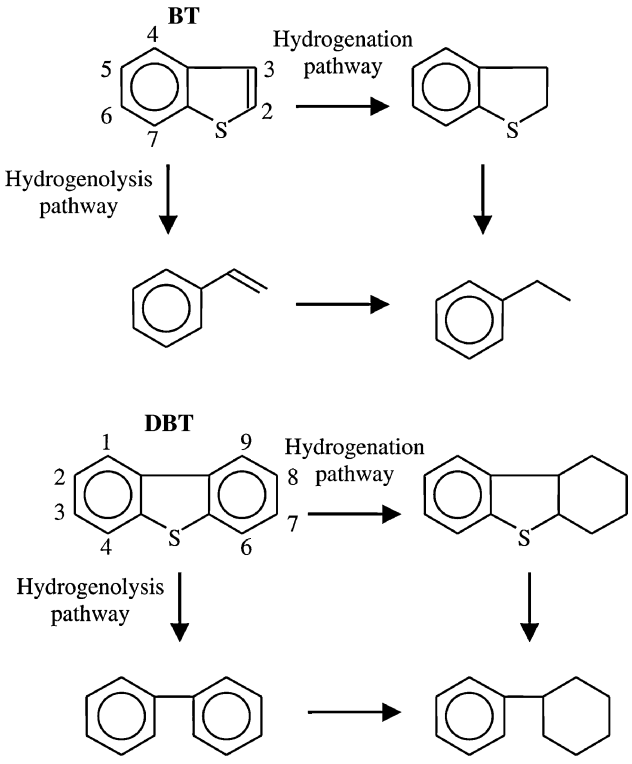
CHALLENGES IN ULTRADEEP DESULFURIZATION PROCESSES

Reactivity of Sulfur Compounds

Recently, investigations demonstrated that sulfur compounds remaining in diesel fuels at sulfur level lower than 500 ppm are dominantly the DBTs with alkyl substituents at the 4- and/or 6-position, and are lower in HDS reactivity.[3–6] These species are termed refractory sulfur compounds. Both steric hindrance and electronic factor are responsible for the observed low reactivity of 4- and 6-substituted DBTs.[19]

Based on HDS reactivity of sulfur compounds in a gas oil, the sulfur compounds can be classified into four groups according to their HDS reactivities that were described by the pseudo-first-order rate constants.[20] The first group is dominantly alkyl BTs; the second, DBT and alkyl DBTs without alkyl substituents at the 4- and 6-positions; the third, alkyl DBTs with only one alkyl substituent at either the 4- or the 6-position; the fourth, alkyl substituents at the 4- and 6-positions. The sulfur distribution in the four groups in the gas oil is 39, 20, 26, and 15 wt%, respectively, and the relative rate constant of HDS for each of the four groups is 36, 8, 3, and 1, respectively.[20] Fig. 3 shows the relative reactor volume requirements for various degrees of sulfur removal by conventional single-stage HDS of diesel fuel.[6] According to this result, when the total sulfur content is reduced to 500 ppmw, the sulfur compounds remaining in the hydrotreated oil are the third and fourth group sulfur compounds. When the total sulfur content is reduced to 30 ppmw, the sulfur compounds remaining in the hydrotreated oil are only the fourth group sulfur compounds, indicating that the lower is the sulfur concentration the lower is the HDS reactivity of the remainder sulfur compounds.

For desulfurization of sulfur species in gasoline, it is not difficult to remove the sulfur compounds in the



Scheme 1 Mechanistic pathways (hydrogenolysis and hydrogenation) for the desulfurization of refractory polyaromatic sulfur compounds.

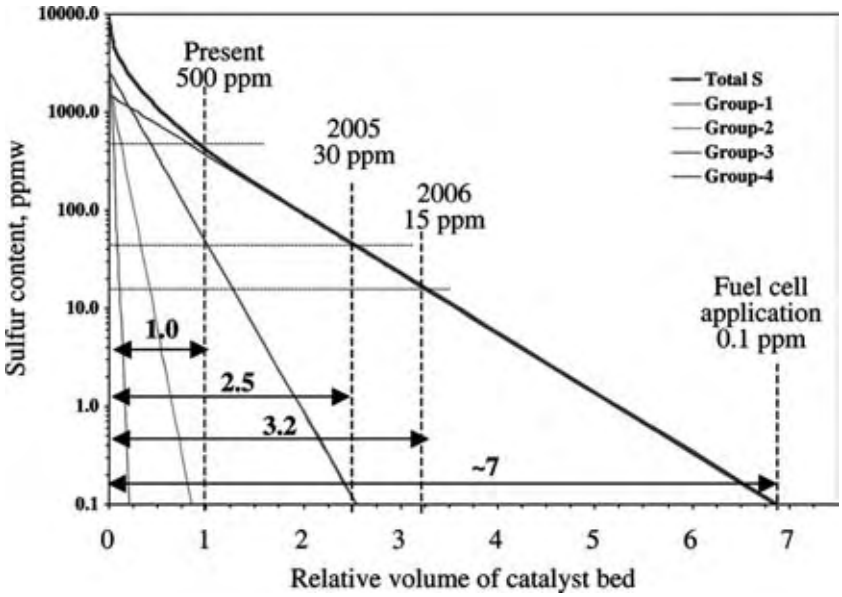


Fig. 3 Simulated HDS of diesel to meet 15 and 0.1 ppm levels on the basis of a conventional single-stage reactor, assuming 1.0 wt% S in the feed. HDS kinetic model: $C_{S,total} = C_{S1o}\,e^{-k1t} + C_{S2o}\,e^{-k2t} + C_{S3o}\,e^{-k3t} + C_{S4o}\,e^{-k4t}$. *(View this art in color at www.dekker.com.)*

naphtha range by current catalytic HDS processes. For the U.S. refineries, most of the sulfur in the gasoline pool is found in FCC naphtha. The challenge in deep desulfurization of FCC naphtha is selective conversion of sulfur compounds without saturation of olefinic compounds, which account for about 15–25 wt% in FCC naphtha and contribute to octane number enhancement. High hydrogen consumption in ultradeep HDS of FCC naphtha is another issue that needs to be considered. For straight-run kerosene, which is used for making jet fuels, the sulfur removal by HDS is more difficult than that from naphtha but less difficult compared to that from gas oil.

Inhibiting Effects of Fuel Components

A major challenge in the ultradeep desulfurization of hydrocarbon fuels is the inhibiting effects of fuel components (e.g., nitrogen compounds) and reaction products (e.g., hydrogen sulfide). While this area has been studied in catalysis literature, it is expected to gain increased attention because of the severity of the new sulfur specifications.[1]

The difference in the catalytic activity between model compounds and real feedstocks is the most visible evidence for the influence of feedstock components on HDS catalysts. Some researchers have shown that an increase of 30–50°C in temperature is required for a commercial catalyst to achieve the same conversions of DBT, 4-MDBT, and 4,6-DMDBT in light gas oil as compared to the model compounds when dissolved in a hydrocarbon solvent.[1,8] Inhibition of HDS activity occurs predominantly through one of the following compounds:

1. Aromatic hydrocarbons
2. Nitrogen-containing compounds
3. Hydrogen sulfide
4. Ammonia, saturated hydrocarbons (solvents), and water.

The influence of these inhibitors is known to increase in the following order:[21] saturated hydrocarbons, monoaromatics < condensed aromatics, oxygen compounds, hydrogen sulfide < organic sulfur compounds < basic nitrogen compounds. Three of these inhibitor classes are discussed here.

Aromatic compounds

Naphthalene and its derivatives are one of the more dominant aromatics present in various diesel and jet fuel feedstocks. Therefore, several investigators have reported the influence of naphthalene on HDS of model compounds. One of the first reports was by Lo who found naphthalene to weakly inhibit the conversion and selectivity of the HDS of DBT.[22] Similarly, LaVopa and Satterfield found little effect of naphthalene and phenanthrene on the HDS of thiophene.[23] Other researchers have, however, found naphthalene to be a stronger inhibitor of HDS activity.[4] Nagai and Kabe, in fact, found naphthalene to significantly reduce catalyst selectivity for the hydrogenation pathway.[24] Isoda et al., on the basis of similar selectivity inhibition, concluded that naphthalene severely inhibits the hydrogenation active sites in a

conventional HDS catalyst.[25] While the inhibition of HDS by naphthalene and other aromatic hydrocarbons is significant, Girgis and Gates believe that it is weaker than inhibition caused by nitrogen-containing heterocyclic compounds.[1]

Nitrogen-containing compounds

Among nonsulfur diesel fuel components, nitrogen-containing organic compounds have received significant attention because of their lower reactivity as compared to that of polyaromatic sulfur compounds (PASCs).[8] Therefore, at very low sulfur levels, refractory nitrogen compounds could significantly influence the deep HDS of PASCs. Atmospheric gas oil, which is frequently used as a diesel feedstock, typically contains nitrogen compounds 70% of which are nonbasic (e.g., carbazole type) while the rest are basic (e.g., quinoline type). Light cycle oil (LCO), a feedstock used for diesel and thermally stable naphthenic jet fuels, contains much higher nitrogen, predominantly nonbasic in nature.[8,11] Diesel and some jet fuel feedstocks, therefore, have enough nitrogen species to significantly influence deep HDS.

Several researchers have clearly established the inhibiting effect of basic nitrogen compounds on the HDS of thiophene and DBT.[4] A much smaller body of work has dealt with the effect of nonbasic nitrogen on HDS and the results seem to be inconclusive. For example, La Vopa and Satterfield and, more recently, Furimsky and Massoth have reported on the inhibiting effects of carbazole and indole, respectively, for the HDS of DBT.[23,26] Similarly, others found the inhibiting effects of carbazole and indole comparable to that of basic nitrogen compounds such as quinoline.[4] However, a very recent study found carbazoles having little negative effect on the HDS of DBT and substituted DBTs found in their blend of gas oil and LCO.[4] They, however, determined carbazoles and substituted carbazoles as the least reactive nitrogen compounds.

In addition to ambiguity on the effect of basic and nonbasic nitrogen, little is known about the exact mechanism of nitrogen poisoning. Kwak et al. found interesting influences of carbazole and quinoline on the HDS of DBT, 4-MDBT, and 4,6-DMDBT over a Co–Mo/γ-Al_2O_3 catalyst.[4] While both nitrogen compounds inhibited conversion of all three PASCs, the HDS of DBT was only mildly poisoned till at least 500 ppm of the nitrogen compound was present in the feedstock. More profoundly, the nitrogen compounds inhibited different types of sites on the catalyst. For example, the HDS of DBT occurred less through the hydrogenation route when poisoned with nitrogen compounds. Nagai and Kabe also obtained similar results with the HDS of DBT on a sulfided Mo/γ-Al_2O_3 catalyst.[4] In the case of the substituted DBTs, however, Kwak et al. noticed that conversion was achieved mainly through the hydrogenation pathway, i.e., the hydrogenolysis route was inhibited.[4]

Hydrogen sulfide

Several researchers have experimentally demonstrated the inhibiting influence of hydrogen sulfide (H_2S) on HDS.[1,4] This inhibiting influence is expected from simple and kinetic and equilibrium considerations. Refiners take great care to keep H_2S in commercial hydrotreaters at an optimum level. For example, hydrogen—used in excess in a hydrotreater—is recirculated after scrubbing out the H_2S by-product carefully. The recycle stream needs to contain an optimum level of H_2S to keep the catalyst as a sulfide and thus maintain its activity and selectivity. Sie has described other process options to minimize inhibition effects by H_2S, e.g., countercurrent flow reactors and monolithic catalyst systems.[27]

Hydrogen sulfide also inhibits HDS activity by modifying the catalyst surface. For example, high concentrations of H_2S are known to increase the density of Bronsted acid sites on a commercial catalyst.[2] Several researchers have found H_2S to poison mainly hydrogenolysis sites on a sulfided Co–Mo/γ-Al_2O_3 catalyst.[2]

Therefore, the petroleum refining industry faces a major challenge to meet the new stricter sulfur specifications and the need for fuel cell applications in the early 21st century when the quality of the crude oils continue to decline in terms of increased sulfur content and decreased API gravity.

APPROACHES TO ULTRADEEP DESULFURIZATION PROCESS

Ultradeep desulfurization approaches include: 1) improving catalytic activity by new catalyst formulation for HDS of 4,6-DMDBT; 2) tailoring reaction and process conditions; 3) designing new reactor configurations; and 4) developing new processes. One or more approaches may be employed by a refinery to meet the challenges of producing ultraclean fuels at an affordable cost.

Improving Catalytic Activity by New Catalyst Formulations

Design approaches for improving catalytic activity for ultradeep HDS focus on how to remove 4,6-DMDBT

more effectively, by modifying catalyst formulations to 1) enhance hydrogenation of the aromatic ring in 4,6-DMDBT by increasing the hydrogenating ability of the catalyst; 2) incorporate acidic feature in the catalyst to induce isomerization of methyl groups away from the 4- and 6-positions; and 3) remove inhibiting substances (such as H_2S). The catalytic material formulations may be improved for better activity by using different supports (carbon, TiO_2, TiO_2–Al_2O_3, HY, MCM-41, etc.) for conventional alumina-supported Co–Mo, Ni–Mo, and Ni–W catalysts; increasing the loading level of active metal (Mo, W, etc.); adding one more base metal (e.g., Ni to Co–Mo or Co to Ni–Mo); and incorporating a noble metal (Pt, Pd, Ru, etc.). Catalyst companies such as Akzo Nobel, Criterion, and Haldor Topsoe have developed more active catalysts for deep HDS.

One catalyst development approach, the inclusion of acidic functionality in HDS catalysts, is discussed briefly here as a case study. Turaga and Song hypothesized that molybdenum sulfide (MoS_2) supported on mesoporous molecular sieve MCM-41 and promoted by cobalt would have superior activity for deep HDS because of MCM-41's 1) high surface area and uniform mesopores and 2) superior acidity as compared to conventional supports such as γ-alumina (γ-Al_2O_3).[28] They synthesized and screened a series of mesoporous aluminosilicate MCM-41 molecular sieves with different silica/alumina (SiO_2/Al_2O_3) ratios (the framework SiO_2/Al_2O_3 ratio determines support acidity) as catalyst supports for varying loadings of cobalt oxide (CoO) and molybdenum oxide (MoO_3).

The activity of the MCM-41-supported catalysts was found to depend on the CoO–MoO_3 loading. At CoO–MoO_3 loadings typical of commercially available HDS catalysts, MCM-41-supported catalysts were only slightly better. At higher loadings—27.0% (by weight) MoO_3 and 5.8% CoO—MCM-41-supported catalysts were twice more active than the commercial catalyst. This difference in activities is related to the degree of MoS_2 stacking as suggested by temperature-programmed reduction and X-ray photoelectron spectroscopy. Additional characterization studies, furthermore, suggest the presence of a different and more active catalytic phase on MCM-41-supported catalysts.

A remarkable increase in the conversion of 4,6-DMDBT was observed over MCM-41-supported catalysts with decreasing SiO_2/Al_2O_3 ratio (or increasing acidity). More significantly, the SiO_2/Al_2O_3 ratio of MCM-41 has a profound effect on product distribution and catalyst selectivity. Irrespective of CoO–MoO_3 loading, catalysts using MCM-41 with a SiO_2/Al_2O_3 ratio of 50 convert more of 4,6-DMDBT through the highly desirable hydrogenolysis pathway. In comparison, the γ-Al_2O_3-supported commercial catalyst relies on the hydrogenation pathway for catalytic conversion. The acidity of these catalysts was measured using temperature-programmed desorption of *n*-butylamine and aluminum magic angle spinning nuclear magnetic resonance and correlated to their selectivities for hydrogenolysis and hydrocracking.

Tailoring Reaction and Process Conditions

Tailoring process conditions aims at achieving deeper HDS with a given catalyst in an existing reactor without changing the processing scheme, with no or minimum capital investment. The parameters include those that can be tuned without any new capital investment (space velocity, temperature, pressure) and those that may involve some relatively minor change in processing scheme or some capital investment (expansion in catalyst volume or density, H_2S scrubber from recycle gas, improved vapor–liquid distributor). First, space velocity can be decreased to increase the reactant-catalyst contact time. More refractory sulfur compounds would require lower space velocity for achieving deeper HDS. Second, temperature can be increased, which increases the rate of HDS. Higher temperature facilitates more of the high activation energy reactions. Third, hydrogen pressure can be increased. Fourth, improvements can be made in vapor–liquid–solid contact, which effectively increases the surface area of the catalyst. Fifth, the concentration of hydrogen sulfide in the recycle stream can be removed by scrubbing. Because H_2S is an inhibitor to HDS, its buildup in high-pressure reactions through continuous recycling can become significant. Recent work on decreasing the concentration of hydrogen sulfide in gas phase has been discussed by Sie.[27] Finally, more volume of catalyst can be used, through either catalyst bed volume expansion or more dense packing.

Designing New Reactor Configurations

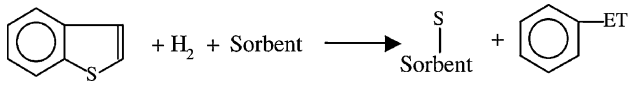
Industrial reactor configuration for deep HDS of gas oils in terms of reaction order, and the effects of the H_2S produced have been discussed by Sie.[27] The reactor design and configuration involve both single-stage and two-stage desulfurization processes. Desulfurization processes in use today in the United States generally have only one reactor, because of the need to only desulfurize diesel fuel to 500 ppmw or lower. Hydrogen sulfide strongly suppresses the activity of the catalyst for converting the refractory sulfur compounds, which should occur in the major downstream part of a cocurrent trickle-bed reactor during deep desulfurization.

The normally applied cocurrent trickle-bed single reactor is, therefore, not the optimal technology for deep desulfurization. A second reactor can be used, particularly to meet the lower sulfur levels. Both desulfurization and hydrogenation in the second reactor can be improved by removing H_2S and NH_3 from the exit gas of the first reactor before entering the second reactor. This last technical change is to install a complete second stage to the existing one-stage hydrotreater. This second stage would consist of a second reactor, and a high-pressure hydrogen sulfide scrubber between the first and the second reactor. Assuming use of the most active catalysts available in both reactors, UOP projects that converting from a one-stage to a two-stage hydrotreater could produce 5 ppm sulfur relative to the current level of 500 ppm today.[29]

A new way of reactor design is to have two or more catalyst beds, which are normally placed in separate reactors, within a single reactor shell and have both cocurrent and countercurrent flows. This new design was pioneered by ABB Lummus and Criterion, as represented by their SynSat process.[30] Traditional reactors are cocurrent in nature. The hydrogen is mixed with the distillate at the entrance to the reactor and the mixture flows through the reactor. The advantage of cocurrent design is practical; it eases the control of gas–liquid mixing and contact with the catalyst. The disadvantage is that the concentration of H_2 is the highest in the front of the reactor and lowest at the outlet. The opposite is true for the concentration of H_2S. The novel solution to this problem is to design a countercurrent reactor, where the fresh H_2 is introduced from the bottom of the reactor and the liquid distillate from the top. Here, the hydrogen concentration is highest (and the hydrogen sulfide concentration is lowest) where the reactor is trying to desulfurize the most difficult (sterically hindered) compounds. The difficulty of countercurrent designs in the case of distillate hydrotreating is vapor–liquid contact and the prevention of liquid flooding and hot spots within the reactor. The SynAlliance (consisting of ABB Lummus, Criterion Catalyst Corp., and Shell Oil Co.) patented a countercurrent reactor design called SynTechnology. With this technology, in a single reactor design, the initial portion of the reactor will follow a cocurrent design, while the last portion of the reactor will be countercurrent.[30]

Developing New Processes

Among the new process concepts, design approaches for ultradeep desulfurization focus on several different areas.[1] Some researchers are looking at adsorption and sulfur atom extraction–removing sulfur by using reduced metals to react with sulfur to form metal sulfides at elevated temperatures under H_2 atmosphere without hydrogenation of aromatics. For instance, Phillips Petroleum (now Conoco Phillips) studied its refineries and concluded the use of hydrotreating technologies to reach ultralow sulfur levels in gasoline to be a cost-prohibitive option. It developed a new process, S Zorb Sulfur Removal Technology for gasoline and diesel, in which the sulfur atom in the sulfur-containing compounds adsorbs onto a sorbent and reacts with the sorbent (see Scheme 2) at elevated temperatures under low H_2 pressure.[31] On another front, researchers at Pennsylvania State University are investigating selective adsorption for removal of sulfur compounds (PSU-SARS) in which sulfur is removed under ambient or mild conditions without using hydrogen by selective interaction with sulfur compounds in the presence of aromatic hydrocarbons.[6,7,32] Researchers at University of Michigan are using pi-complexation for adsorption of sulfur compounds.[33] In other studies, sulfur compounds are oxidized by liquid-phase oxidation reactions with or without ultrasonic radiation, followed by separation of the oxidized sulfur compounds.[34] Finally, biodesulfurization is a process that removes sulfur from fossil fuels using bacteria via microbial desulfurization. Several recent reviews outline the progress in the study of microbial desulfurization from the basic and practical point of view.[35,36]



Scheme 2 Basic principle of the S Zorb Sulfur Removal Technology process developed by Phillips Petroleum for sulfur removal from liquid fuel at elevated temperatures under low H_2 pressure.

CONCLUSIONS

Heightened concerns for cleaner air and increasingly more stringent regulations on sulfur contents in transportation fuels will make desulfurization more and more important. The sulfur problem is becoming more serious in general, particularly for diesel fuels, as the regulated sulfur content is getting an order of magnitude lower, while the sulfur content of crude oils refined in the United States is becoming higher. The chemistries of gasoline and diesel fuel processing have evolved significantly around the central issue of how to produce cleaner fuels in a more efficient, environment friendly, and affordable fashion. New design approaches are necessary for making affordable ultraclean fuels.

REFERENCES

1. Girgis, M.J.; Gates, B.C. Reactivities, reaction networks, and kinetics in high-pressure catalytic hydroprocessing. Ind. Eng. Chem. **1991**, *30*, 2021.
2. Topsoe, H.; Clausen, B.S.; Massoth, F.E. Hydrotreating catalysis. In *Science and Technology*; Springer-Verlag: Berlin, 1996; 310 pp.
3. Whitehurst, D.D.; Isoda, T.; Mochida, I. Present state of the art and future challenges in the hydrodesulfurization of polyaromatic sulfur compounds. Adv. Catal. **1998**, *42*, 345.
4. Babich, I.V.; Moulijn, J.A. Science and technology of novel processes for deep desulfurization of oil refinery streams: a review. Fuel **2003**, *82* (6), 607–631.
5. Song, C. An overview of new approaches to deep desulfurization for ultra-clean gasoline, diesel fuel and jet fuel. Catal. Today **2003**, *86* (1–4), 211–263.
6. Song, C.; Ma, X. New design approaches to ultra-clean diesel fuels by deep desulfurization and deep dearomatization. Appl. Catal. B Environ. **2003**, *41* (1–2), 207–238.
7. Song, C. Fuel processing for low-temperature and high-temperature fuel cells. Challenges, and opportunities for sustainable development in the 21st century. Catal. Today **2002**, *77* (1), 17–50.
8. Prins, R. Catalytic hydrodenitrogenation. Adv. Catal. **2001**, *46*, 399–464.
9. DeCicco, J.; Mark, J. Meeting the energy and climate challenge for transportation in the United States. Energy Policy **1998**, *26*, 395–412.
10. EPA. Control of air pollution from new motor vehicles: heavy-duty engine and vehicle standards and highway diesel fuel sulfur control requirements. Fed. Reg. **2001**, *66*, 5101–5150.
11. Song, C. Introduction to chemistry of diesel fuels. In *Chemistry of Diesel Fuels*; Song, C., Hsu, C.S., Mochida, I., Eds.; Taylor & Francis: New York, 2000; 1–60.
12. National Petroleum Refiners Association. *Hydroprocessing—Fuel Quality, 1998 Q&A Minutes*; National Petroleum Refiners Association: Washington, DC, 1998.
13. EPA. Control of diesel fuel quality. Fed. Reg. **1999**, *64*, 26142–26158.
14. Manufacturers of Emission Controls Association. *The Impact of Sulfur in Diesel Fuel on Catalyst Emission Control Technology*; Manufacturers of Emissions Controls Association: Washington, DC.
15. BP. *BP Statistical Review of World Energy 2001*; BP p.l.c.: London, U.K., 2001.
16. Touvelle, M.S.; McVicker, G.; Daage, M.; Hudson, C.W.; Klein, D.P.; Cook, B.R.; Chen, J.G.; Hantzer, S.; Vaughan, D.E.W.; Ellis, E.S. Selective ring opening of naphthenic molecules. In 17th North American Catalysis Society Meeting, Toronto, Canada, 2001**.**
17. Banerjee, N.; Hakim, D. U.S. ends car plan on gas efficiency; looks to fuel cells. The New York Times **2002,** Jan 9, A1.
18. Nag, N.K.; Spare, A.V.; Broderick, D.H.; Gates, B.C. Hydrodesulfurization of polycyclic aromatics catalyzed by sulfided $CoO–MoO_3/\gamma$-Al2O_3: the relative reactivities. J. Catal. **1979**, *57*, 509–512.
19. Ma, X.; Sakanishi, K.; Isoda, T.; Mochida, I. Quantum chemical calculation on the desulfurization reactivities of heterocyclic sulfur compounds. Energy Fuels **1995**, *9*, 33–37.
20. Ma, X.; Sakanishi, K.; Mochida, I. Hydrodesulfurization reactivities of various sulfur compounds in diesel fuel. Ind. Eng. Chem. Res. **1994**, *33*, 218–222.
21. Schulz, H.; Bohringer, W.; Waller, P.; Ousmanov, F. Gas oil deep hydrodesulfurization: refractory compounds and retarded kinetics. Catal. Today **1991**, *49*, 87–97.
22. Lo, H.S. Modeling of Hydrotreating. Ph.D. Thesis, Department of Chemical Engineering, University of Delaware, Newark, DE, **1981**.
23. LaVopa, V.; Satterfield, C.N. Poisoning of thiophene hydrodesulfurization by nitrogen compounds. J. Catal. **1998**, *110*, 375–387.
24. Nagai, M.; Kabe, T. Selectivity of molybdenum catalyst in hydrodesulfurization, hydrodenitrogenation, and hydrodeoxygenation: effect of additives on dibenzothiophene hydrodesulfurization. J. Catal. **1983**, *81*, 440–449.
25. Isoda, T.; Nagao, S.; Ma, X.; Korai, Y.; Mochida, I. Hydrodesulfurization of refractory sulfur species. 1. Selective hydrodesulfurization of 4,6-dimethyldibenzothiophene in the major presence of naphthalene over $CoMo/Al_2O_3$ and Ru/Al_2O_3 blend catalysts. Energy Fuels **1996**, *10*, 482–486.
26. Furimsky, E.; Massoth, F.E. Deactivation of hydroprocessing catalysts. Catal. Today **1999**, *52*, 381–495.
27. Sie, S.T. Reaction order and role of hydrogen sulfide in deep hydrodesulfurization of gas oil: consequences for industrial reactor configuration. Fuel Proc. Tech. **1999**, *61*, 149–171.
28. Turaga, U.T.; Song, C.; Turaga, T.; Song, C. MCM-41-supported Co–Mo Catalysts for deep hydrodesulfurization of light cycle oil. Catal. Today **2003**, *86* (1–4), 129–140.
29. EPA-Diesel RIA. *Regulatory Impact Analysis: Heavy-Duty Engine and Vehicle Standards and Highway Diesel Fuel Sulfur Control Requirements*, Air and Radiation EPA420-R-00-026;

United States Environmental Protection Agency, Dec. 2000.
30. Dautzenberg, F. A call for accelerating innovation. Cattech **1999**, *3* (1), 54–63.
31. Gislason, J. Phillips sulfur-removal process nears commercialization. Oil Gas J. **2002**, *99* (47), 74–76.
32. Ma, X.; Sun, L.; Song, C. A new approach to deep desulfurization of gasoline, diesel fuel and jet fuel by selective adsorption for ultra-clean fuels and for fuel cell applications. Catal. Today **2002**, *77* (1–2), 107–116.
33. Hernandez-Maldonado, A.J.; Yang, R.T. Desulfurization of liquid fuels by adsorption via complexation with Cu(I)-Y and Ag-Y zeolites. Ind. Eng. Chem. Res. **2003**, *42* (1), 123–129.
34. Aida, T.; Yamamoto, D.; Iwata, M.; Sakata, K. Development of oxidative desulfurization process for diesel fuel. Rev. Heteroatom Chem. **2000**, *22*, 241–256.
35. MacFarland, B.L.; Boron, D.J.; Deever, W.; Meyer, J.A.; Johnson, A.R.; Atlas, R.M. Biocatalytic sulfur removal from fuels: applicability for producing low sulfur gasoline. Crit. Rev. Microbiol. **1998**, *24* (2), 99–147.
36. Ohshiro, T.; Izumi, Y. Microbial desulfurization of organic sulfur compounds in petroleum. Biosci. Biotechnol. Biochem. **1999**, *63* (1), 1–9.

Detergent Alkylate

Bipin Vora
Andrea Bozzano
Stephen Sohn
UOP LLC, Des Plaines, Illinois, U.S.A.

INTRODUCTION

The subject of this entry relates to the manufacture of linear alkylbenzene (LAB), commercially known as detergent alkylate. It is one of the primary surfactant raw materials for the production of household, laundry, dishwashing, and other surfactants.

The alkylation is the substitution of a hydrogen atom bonded to the carbon atom of an aromatic ring by an alkyl group. The alkylations of paraffins, nitrogen, oxygen, and sulfur are described in separate entries. Most of the industrially important alkyl aromatics used for petrochemical intermediates are produced by alkylating benzene with mono-olefins, e.g., ethylene, propylene, and olefins with 10–18 carbons. The liquid anhydrous hydrofluoric acid (HF) was the most commonly used catalyst until the mid-1990s. Environmental concerns associated with mineral acid catalysts have encouraged process changes and the development of solid-bed alkylation processes. This entry summarizes the process steps used produce LAB via both HF and solid-bed alkylation technologies, including economic comparison of both the processes.

HISTORICAL BACKGROUND

The synthetic detergents industry originated in the 1940s, when it was found that a new anionic surfactant type—alkylbenzene sulfonate—had detergent characteristics superior to those of natural soaps. The first surfactant of this kind was sodium dodecylbenzene sulfonate (SDBS). This material was produced by the Friedel–Crafts alkylation reaction of benzene with propylene tetramer (a mixture of C_{12} olefin isomers), followed by sulfonation with oleum or sulfur trioxide and then neutralization, usually with sodium hydroxide. The alkylation was typically performed using homogenous acid catalysts, such as HF or sulfuric acid.

Because of its low, stable price and its effectiveness in a wide range of detergent formulations, SDBS rapidly displaced natural soaps in household clothes washing and dishwashing applications. It became the standard synthetic surfactant for the detergent industry.

In the early 1960s, it became obvious that dodecylbenzene-based detergents were contributing to the pollution of lakes and streams. The reason for this was the poor biodegradability of the highly branched propylene tetramer aliphatic chain. This caused detergents discharged from water treatment plants to maintain their surfactant properties and form relatively stable foams.[1–3]

In the mid-1960s, commercial processes were developed for the production of linear paraffins. These could be used as feedstock to produce linear alkylbenzene sulfonates (LAS), a biodegradable alternative to SDBS.[4,5] The majority of synthetic detergent plants built after the 1970s were for the production of LABs rather than DDB. Reformulation with LAS started the phaseout of DDB, and by the end of the 1990s, most DDB plants were no longer in operation or had been converted to production of LAB.

In the 1980s and 1990s concern over environmental and safety issues associated with the use of HF in alkylation plants increased. This pressure resulted in the development of solid-bed alkylation technology for the production of LAB. The majority of LAB plants built since 1995 utilize solid catalyst technology.

The worldwide annual production of LAB increased from 1.1 million tons in 1980 to 1.8 million metric tons in 1990 and 2.4 million tons in 2000.[6] Linear alkylbenzene sulfonates continue to be one of the most efficient and cost-effective surfactants in the detergent industry.

PRODUCTION OF BRANCHED DODECYLBENZENE

Branched dodecylbenzene is produced by alkylation of polymer tetramer and benzene. The polymer tetramer feedstock is produced in a propylene oligomerization plant using a process such as the UOP catalytic condensation process. Benzene feedstock is typically produced by solvent extraction of reformate or pygas by means of an aromatics extraction process. There is no detailed discussion of DDB production in this entry, as the production of SDBS has been almost entirely phased out. A complete discussion is presented in Ref.[7].

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016170

PRODUCTION OF LAB

Manufacturing Routes

There are several routes for the production of LAB. In most cases, a linear paraffin feedstock is used to produce olefins for alkylation. The paraffin feedstock is typically a mixture of linear paraffins in the range of C_{10}–C_{14}. The paraffins are derived from kerosene by means of adsorptive separation.

Other sources of olefins, such as α-olefins, can also be used to produce LAB. In most cases, these are only used as a supplemental feedstock to alkylation units, as a means of boosting LAB production. Typically, economics preclude the use of purchased olefins as the sole feedstock for the production of LAB. Benzene is another feedstock required for the alkylation of olefins to produce alkylbenzenes.

The most common route to LAB involves HF alkylation to produce LAB from linear mono-olefins and benzene. In this route, linear mono-olefins are produced by catalytic dehydrogenation of linear paraffins. The mono-olefins may then be enriched, by selective hydrogenation of diolefins (formed because of the side reactions in the dehydrogenation step). In the final step, olefins and benzene are fed to the alkylation unit to produce LAB. There are over 30 operating plants worldwide producing LAB by this manufacturing route.

The current state-of-the-art technology for LAB production uses a solid acid catalyst instead of HF. As with HF alkylation, this route uses catalytic paraffin dehydrogenation to produce mono-olefins, and selective hydrogenation to revert diolefins back to mono-olefins. An added step of aromatics removal is used to increase LAB yields in the alkylation step. As of 2003, there are three operating plants producing LAB using this manufacturing route, with several more in the design stage.

A less common route uses an aluminum chloride ($AlCl_3$) alkylation to produce LAB. This route typically uses a linear olefin feedstock produced by either catalytic dehydrogenation or paraffin chlorination/dehydrochlorination. A variation of the aluminum chloride route uses a chlorination step to convert paraffins to monochloroparaffins as an intermediate, which is then fed directly to an $AlCl_3$ unit to form LAB.[8,9] Considerable amounts of indane and tetralin derivatives are formed as by-products from this technology.[10] Aluminum chloride routes were in common use during 1960s and 1970s. However, because of corrosion, higher maintenance, and inferior product quality, the aluminum chloride plants have been mostly phased out and only a few industrial plants continue to use this technology to produce LAB.

Paraffin dehydrogenation followed by alkylation accounts for about 90% of the current world production. Most of the units built prior to the 1990s employ a homogenous HF acid catalyst system and the units built post-1990 employ heterogenous solid acid catalyst systems. As such, detailed discussion in this entry will be focused on these two manufacturing routes with particular emphasis on the heterogenous catalyst system.

Yields and Economics

Typical yields for complexes using HF and solid-bed alkylation routes are shown in Table 1. This table illustrates that the yields for the two routes are similar. For constant production of LAB, paraffin use is approximately equal for both the routes. The ''HAB'' by-product stream consists of heavy alkylate (discussed in more detail in later sections). The HAB by-product is formed in both routes and depending on the properties, may be used in applications, such as heat transfer fluids, or as enhanced oil recovery surfactants in a sulfonated form. Both routes also produce some light products in the form of off-gas and cracked product from the dehydrogenation unit. The solid-bed alkylation route also produces an aromatic by-product stream (PEP™ Extract in Table 1), which consists of aromatics produced in the dehydrogenation unit. While aromatics removal is possible for the HF route, it is typically not practiced. Instead, the HF route has an acid regenerator bottoms stream, which consists of by-products extracted from purification of the HF acid. Both of these by-products are typically recovered for fuel value. In the table Case-1 represents an LAB complex that includes the Pacol™, DeFine™, PEP, and Detal™ processes all licensed by UOP LLC and hereafter referred to as ''Pacol/DeFine/PEP/Detal complex.'' Case-2 represents the Pacol, DeFine, and UOP HF detergent alkylation processes, all licensed by UOP LLC and hereafter referred to as ''Pacol/DeFine/HF Alky complex.''[a]

Typical economics for a modern LAB complex are shown in Table 2.

Linear Alkylbenzene Product Properties

Table 3 compares LAB product properties for the HF, $AlCl_3$, and Detal catalyst systems. Bromine index and sulfonatability are the key measures of product quality, as they affect final product value. High-bromine index LAB also will produce a more colored sulfonate. Tetralin content can be important, as it relates to the ultimate biodegradation of the product and sulfonability.

[a]UOP™, Pacol™, DeFine™, and Detal™ are trademarks and/or service marks of UOP LLC.

Table 1 Yields for LAB complex. Case 1: Solid-bed alkylation; Case 2: HF alkylation

	Units	Case 1 Pacol/DeFine/ PEP/Detal	Case 2 Pacol/DeFine/ HF Alky
Raw materials			
Normal paraffins	MT	78.2	77.8
Benzene	MT	33.2	33.0
Total raw materials	MT	111.4	110.8
Products			
LAB	MT	100.0	100.0
HAB	MT	4.4	4.5
Benzene drag	MT	—	—
PEP extract	MT	2.4	—
Acid Regen Btms (polymer)	MT	—	1.6
Net separator off-gas	MT	1.5	1.6
Stripper Ovhd (Liq + Vap)	MT	3.1	3.1
Total products	MT	111.4	110.8

A significant reduction in tetralin content was brought about by the conversion of diolefins to mono-olefins prior to alkylation. Product linearity is a parameter related to the rate of biodegradation of the ultimate LAS product. The 2-phenyl content of the LAB has an impact on product solubility with a maximum solubility at approximately 30%. As can be seen in Table 3, the Detal LAB has the same or better linearity, improved sulfonate color, and lower tetralin content compared to the LAB produced in either the $AlCl_3$ or the HF alkylation process. It also has higher 2-phenylalkane content, which gives improved solubility in many detergent formulations. The cloud point of a liquid-detergent formulation prepared with LAS derived from Detal LAB is lower than that of the same formulation produced from HF LAB over a wide range of surfactant concentrations.[11] All of these properties demonstrate that the current Detal technology produces a superior product to the HF technology or the older $AlCl_3$ technology.

Table 2 Typical economics for an modern LAB complex

LAB capacity	80,000 MTA
Location	USGC, 2002
Capital costs	ISBL—$52 MM, OSBL—$16 MM
Total EEC	$101 MM
Feedstock	$/MT LAB
Benzene	103
N-Paraffin	357
Total	459
By-product credits	(22)
Consumables (utilities, catalyst and chemicals)	77
Direct fixed costs (labor, maintenance, interest on working capital)	45
Indirect fixed costs	41
Total cash cost of production	600
Simple return on investment	28%

ISBL, inside battery limits; OSBL, outside battery limits; EEC, estimated erected cost.

DISCUSSION OF MANUFACTURING ROUTE

Fig. 1 shows a block flow diagram for the processing flow scheme of a LAB complex. The first step in the complex is the dehydrogenation of paraffins to produce mono-olefins. This is typically done in a Pacol process. Owing to equilibrium limitations, conversion levels in Pacol units are typically 10–15%. This necessitates a large recycle paraffin stream from the alkylation unit. The net products from the Pacol unit are hydrogen, a small amount of light ends (from cracking side reaction), and the paraffin/olefin mixture, which becomes the feed to the DeFine unit. In this unit, diolefins produced in the Pacol process are hydrogenated to mono-olefins or paraffins. The DeFine unit requires a small hydrogen stream to give appropriate conversion/selectivity. The mono-olefin-containing product from the DeFine unit then goes to the PEP unit for removal of aromatics by-products. Aromatics are produced in small quantities in the Pacol unit, and their removal increases product yields in the alkylation unit. As such, one of the side products from the PEP unit is an aromatic by-product stream. Last, the treated product from the PEP unit is routed to the alkylation unit. This can be done with a UOP HF detergent alkylate unit or a Detal solid-bed alkylation unit. Here, the olefins are

Table 3 Typical LAB properties

Properties	$AlCl_3$ detergent alkylate	HF detergent alkylate	Detal detergent alkylate
Specific gravity	0.85–0.87	0.85–0.87	0.85–0.87
Bromine index	<15	<15	<15
Saybolt color	+30	+30	+30
Water (ppm)	<100	<100	<100
Tetralins (wt%)	5–15	<1.0	<0.5
2-Phenylalkanes (wt%)	30	15–18	>25
n-Alkylbenzene (wt%)	90	92–94	92–94
Klett color of 5% solution	20–40	20–40	10–30
Molecular weight	235–260	235–260	235–260

alkylated with benzene to make LAB. A side product from both units is a heavy alkylate stream. An additional side product only from HF units is the acid regenerator column bottoms stream. This stream contains mostly aromatics and heavy alkylate, and is a result of purification of the HF acid phase from the alkylation reaction.

Paraffin Dehydrogenation—Chemistry and Reaction Conditions of the Pacol Unit

The paraffin dehydrogenation reaction scheme is shown in Fig. 2. Paraffins are dehydrogenated to form mono-olefins with the double bond distributed according to thermodynamics (less than 10% in the α position). The extent of the reaction is largely controlled by thermodynamic equilibrium, and typical paraffin conversion levels are limited to 10–15%. The reaction is typically carried out at low pressure to enhance the equilibrium in favor of olefin production.

Dehydrogenation is carried out over a noble metal catalyst in the vapor phase and in the presence of hydrogen at temperatures of greater than 400°C. To minimize coking on the catalyst and improve the catalyst stability some hydrogen is introduced with the feed. In addition to the olefins, hydrogen is the second main product that is recovered and a portion of it is recycled. Because of the high activity relative to equilibrium conversion, a high space velocity is used.

Some undesirable side reactions can also occur. Mono-olefins can be further dehydrogenated to give diolefins, which in turn can be dehydrogenated further to, ultimately, give aromatics. The aromatic compounds are typically disubstituted cyclics with molecular weight similar to the starting paraffins.

Paraffins can also be isomerized to isoparaffins, which in turn are dehydrogenated to *iso*-olefins, branched diolefins, and aromatics (the last two are not shown in Fig. 2). Some cracking of both paraffins and olefins occurs, to give light ends (C_9 and lighter). It is desirable to maximize the yield to mono-olefins while reducing the yield to diolefins, aromatics, and light ends.

Selective Hydrogenation—Chemistry and Reaction Conditions of the DeFine Unit

The reaction of diolefins in the alkylation unit results in the formation of undesirable products: diphenyl alkanes, polymers, and indane/tetralins (Fig. 3). The first two result in yield loss, while the third lowers product quality. Selective hydrogenation can be used to convert diolefins to mono-olefins. The UOP DeFine process was commercialized for this purpose in 1986.[12,13] Use of the DeFine process in LAB production results in approximately 50% reduction in the formation of heavy alkylate and approximately 5% increase in LAB yield. The reaction takes place over

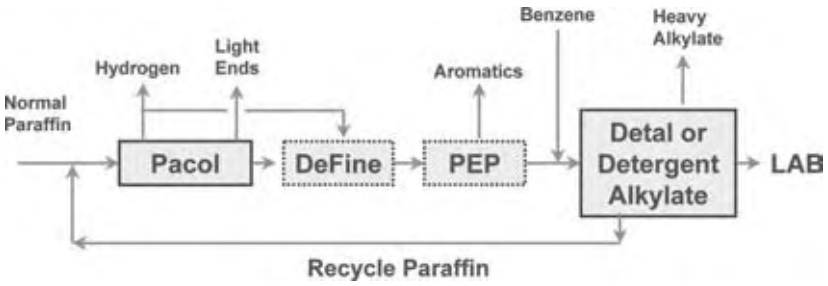


Fig. 1 Linear alkylbenzene complex flow scheme. *(View this art in color at www.dekker.com.)*

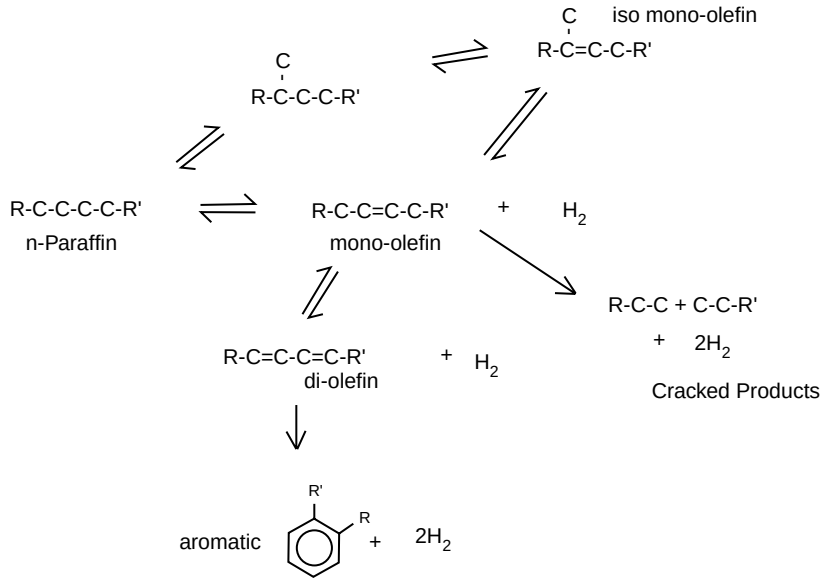


Fig. 2 Pacol reaction scheme. *(View this art in color at www.dekker.com.)*

a catalyst, in the liquid phase, at temperatures from 150°C to 250°C.

Aromatics Removal—PEP Chemistry and Conditions

In the Pacol unit, aromatics produced via dehydrocyclization reaction are of the same carbon range (C_{10}–C_{13}) as the small quantity of alkylaromatics that are present in the feed paraffin. Aromatics are undesirable for several reasons:

- Olefins will alkylate with the aromatic by-product in the alkylation unit forming a heavy alkylate (C_{10}–C_{13} plus C_{10}–C_{13} making C_{20}–C_{26} range heavy alkylate, see Fig. 4).
- Conversion of alkylaromatics in the alkylation unit is low—only 15%.
- This results in a buildup of alkylaromatics in the recycle paraffin stream. This lowers the production capacity of the LAB unit.

The UOP PEP unit is present in all Detal-based complexes to remove alkylaromatics. Aromatic removal is accomplished by molecular sieve adsorption, and exploits differences in adsorption affinity between aromatics and other molecules in the Pacolate. The PEP units consist of several adsorbent chambers

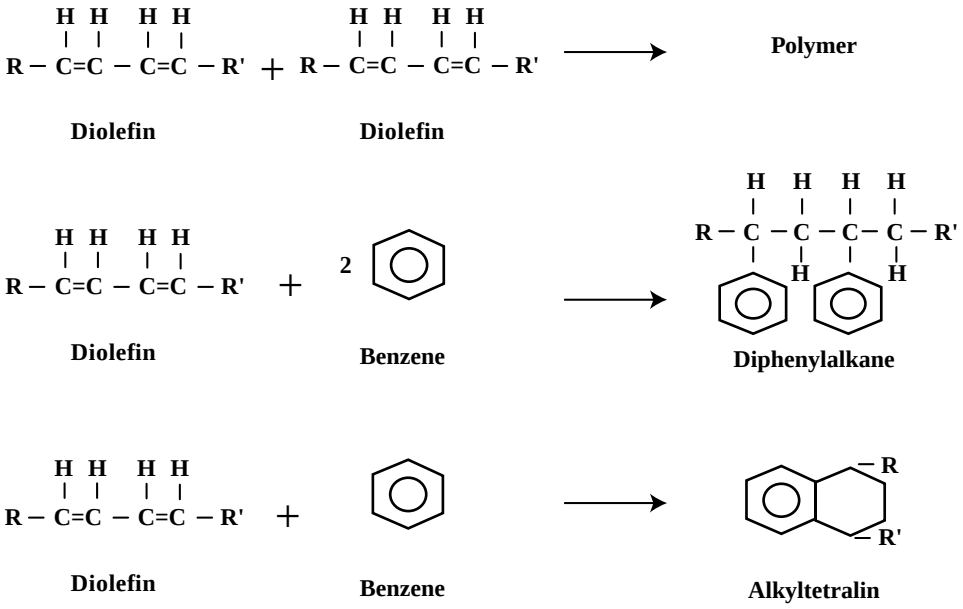


Fig. 3 Undesirable diolefin reactions in alkylation unit. *(View this art in color at www.dekker.com.)*

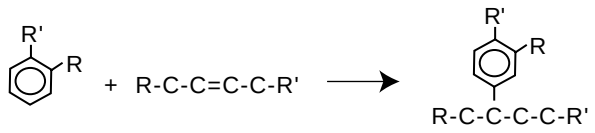


Fig. 4 Undesirable aromatics reactions in alkylation unit. *(View this art in color at www.dekker.com.)*

as shown in Fig. 5. A desorbent displaces the adsorbed aromatics from the molecular sieve. These aromatics are then separated from the desorbent and purge material by fractionation.

Alkylation—Chemistry

As shown in Fig. 6, the primary detergent alkylation reaction is the alkylation of benzene with the straight chain mono-olefins to yield LAB.

This reaction is acid catalyzed and can be generalized as occurring in two steps. The first step is the formation of a carbonium ion or polarized complex. The second step is the electrophilic attachment of the aromatic to the carbonium ion resulting in the LAB molecule.[14] This is accomplished using traditional protonic acid catalysts (H_2SO_4, HF, and phosphoric acid), Friedel–Crafts-type catalysts ($AlCl_3$ and boron fluoride), or any solid acid catalyst having a comparable acid strength. In either case (whether heterogenous or homogenous catalyst), the catalyst must demonstrate sustained activity, selectivity, and stability over long periods of operation.

During alkylation, some undesirable side reactions occur. A couple of these side reactions are outlined in Fig. 7. These secondary reactions include alkylation of olefin with LAB to form dialkylbenzenes, and the reaction of two olefins to form a dimer.

Other undesired reaction can occur between diolefins and benzene (see DeFine section) and Pacol aromatics and benzene (see PEP section).

In addition to the alkylation activity, all acid catalysts possess the ability to shift the olefinic double bond position along the chain. Researchers have shown that the position of the phenyl group in the final product is specific to the catalyst system used.[15,16] Table 4 summarizes some of this work for the reaction of 1-dodecene (112-41-4) with benzene.

HF ALKYLATION—UOP DETERGENT ALKYLATE PROCESS

Processing Conditions

The HF alkylation process is the dominant route for LAB production utilized by the plants built before 1995.[17] The olefin feed to a conventional HF alkylation unit is typically derived from the Pacol process. Olefin feed is combined with an excess of benzene before the addition of the HF acid. Diluting the olefin in benzene followed by intense mixing helps minimize the formation of by-products, such as dialkylbenzene. The typical operating temperature for the HF alkylation reaction is 30–40°C. Operating pressure is sufficient

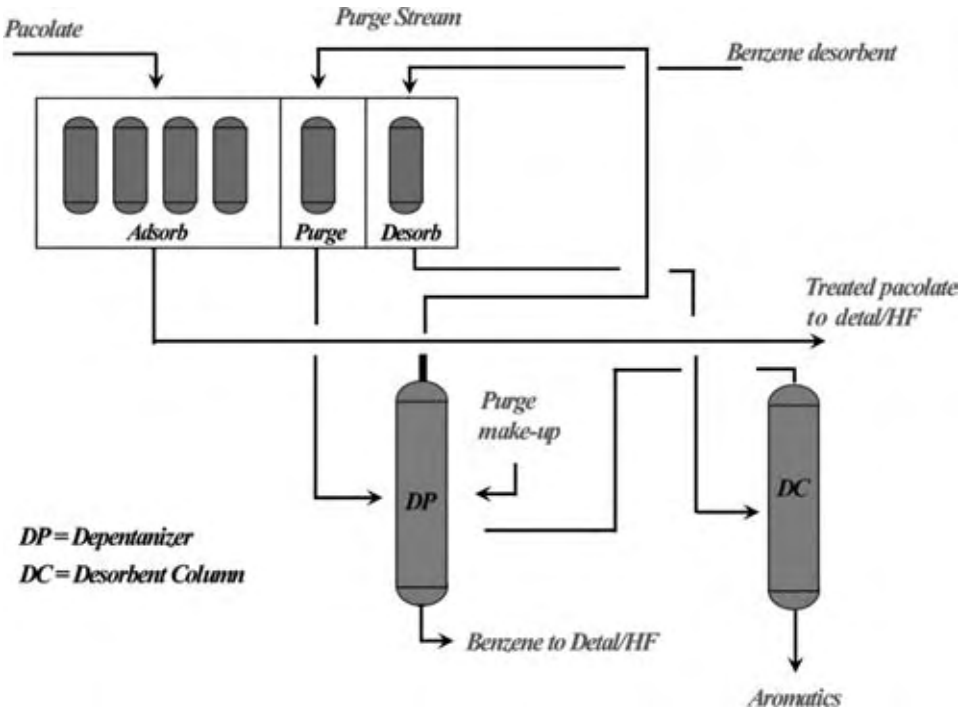


Fig. 5 PEP unit flow scheme. *(View this art in color at www.dekker.com.)*

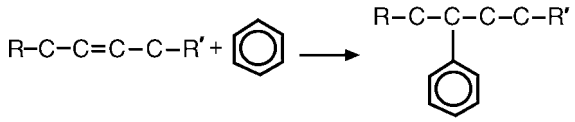


Fig. 6 Formation of LAB from mono-olefins and benzene. *(View this art in color at www.dekker.com.)*

to ensure that the HF acid remains in the liquid phase. A simplified flow diagram for a typical UOP detergent alkylate process is shown in Fig. 8. The olefin feed is combined with makeup and recycled benzene and cooled prior to mixing with HF acid. The reaction section consists of either one or two mixer reactors and acid settlers. A portion of the HF phase from the settler is sent to the HF regenerator, where heavy by-products are removed to maintain the required purity of the HF acid. The balance of the HF phase or circulating acid is returned to the inlet of the HF reactor to mix with the feed prior to entering the reactor. The hydrocarbon phase from the acid settler proceeds to the fractionation section, where the remaining HF catalyst, excess benzene, unreacted *n*-paraffins, heavy alkylate, and the LAB product are separated by fractionation. In these sequential fractionation steps, the recovered HF acid and benzene are recycled to the alkylation reactor. The unreacted *n*-paraffins are passed through an alumina treater to remove combined fluorides and are then recycled back to the dehydrogenation unit. Not shown in Fig. 8 is the HF acid handling and neutralization section. The HF acid from the HF regenerator–HF stripper overhead is recycled to the alkylation reactor, and then, after separation in the settler it is pumped to the alkylation reactor. The HF acid from the first-stage settler is pumped as feed to the HF regenerator.

Olefins react to a small extent with HF to form alkylfluorides that are separated in the paraffin column with the recycle paraffins (Fig. 9). Alumina treaters help to eliminate both combined fluorides and free HF acid from the recycle paraffins.

Metallurgy

Since its introduction during the early 1940s, the HF alkylation process has gone through many changes.

Table 4 Isomer distribution, wt%, of dodecylbenzene from 1-dodecene and benzene

	HF	SiO_2–Al_2O_3	HY–Zeolite
2-Phenyl	16.7	33.4	29.5
3-Phenyl	16.4	21.9	20.2
4-Phenyl	17.5	14.8	17.1
5-Phenyl	24.1	15.6	16.9
6-Phenyl	25.3	14.2	16.3

Its first application was in the production of aviation gasoline, followed by motor fuel blending components, and then extension to the detergent alkylation field.

At low water contents, HF has a very low corrosion rate on mild carbon steel, but as the water content rises above 5%, the corrosion rate rapidly increases. Corrosion is anticipated wherever there is a possibility of acid or acid–hydrocarbon mixtures in contact with water. Copper, silver, and platinum are resistant to corrosion from all concentrations of HF, with the exception that copper and silver are attacked in the presence of sulfur compounds and oxygen. Lead is useful only if the water content is above 35%. Finally, stainless steel has poorer corrosion resistance to HF than carbon steel.

There are four rules for long and successful operation of HF alkylation process units: good engineering design, appropriate materials of construction, diligence in keeping equipment and feedstocks dry, and careful adherence to routine prescribed maintenance practices.

Use of carbon steel may be acceptable where water concentration is assured to be below 5% with varying corrosion allowances for major vessels and piping. However, the HF acid regeneration and the neutralization sections require monel or monel cladded vessels and all monel piping. All instruments in HF or hydrocarbons that contain any amount of HF or may have a possibility of getting some HF acid contamination must have monel parts. Monel is generally specified for the moving parts, for the trim of pumps, in thermowells, in select instrument parts, and as trim on most valves. Monel cladding is necessary for the acid regenerator, a small column, and the piping around the acid regenerator. Kel-F™ and Teflon™ are used largely as

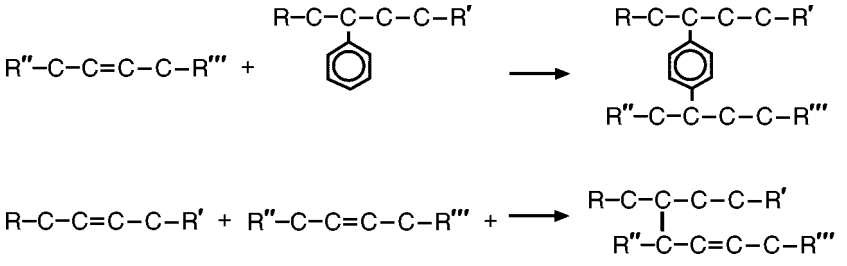


Fig. 7 Alkylation section side reactions. *(View this art in color at www.dekker.com.)*

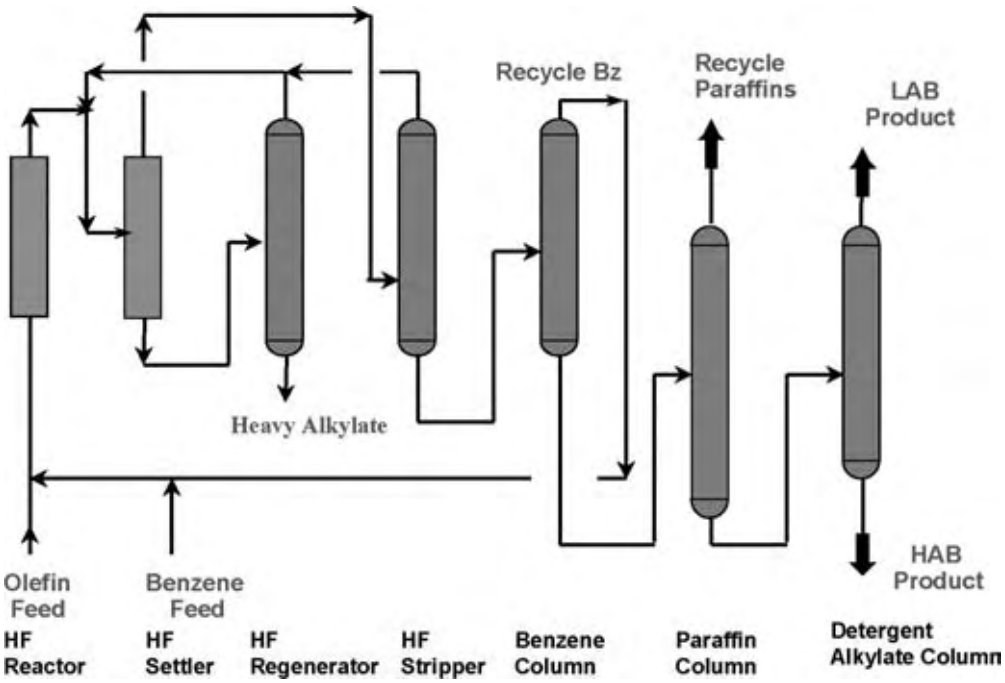


Fig. 8 UOP detergent alkylate process flow scheme. *(View this art in color at www.dekker.com.)*

sealing materials and for certain instrument parts, as well as for packing and trim in some valves.[b]

The hazards of HF acid in HF alkylation have been identified and managed for many years. Plant design, operating practices, and safety equipment are combined to cope with potential risks. Several references deal with safe maintenance and operation.[18–20] Although most of these references are directed toward the HF alkylation process for motor fuel production, general safety rules of handling HF and HF-containing equipment also apply to the detergent alkylation process.

Waste Management

In the UOP detergent HF alkylation process the engineering and design standards have been developed and improved over the years to obtain a process that will operate efficiently and economically. In normal operation, the waste streams from these process units consist of a small water stream from the Pacol stripper overhead and the HF regenerator bottoms.

There may also be certain other undesirable materials in addition to the normal process flow that could be discharged from the unit. These undesirable materials could potentially cause concern if steps were not taken in the waste management process and product treating areas to transform them into either inert or harmless substances.

One common step is to divide the gaseous and liquid waste streams into two separate waste streams, acidic and nonacidic, handled separately. Nonacidic gaseous waste streams from units, such as the dehydrogenation unit, or the acid-free section of the detergent alkylation unit, typically are combined and sent directly to a nonacid flare system where they are disposed of by burning.

Acidic gaseous streams from the detergent alkylation units usually are combined and sent to a relief gas scrubber where HF acid is removed by countercurrent gas–liquid contacting using circulating limewater solution:

$$Ca(OH)_2 + 2HF \rightarrow CaF_2 + 2H_2O$$

Calcium fluoride, being insoluble in water, deposits as sludge in the bottom of circulating limewater tanks and is removed periodically. In some units, KOH solution is used as a neutralizing medium, and spent KOH ($KOH + HF \rightarrow KF + H_2O$) is regenerated by lime: $2KF + Ca(OH)_2 \rightarrow CaF_2 + 2KOH$.

Thus, calcium fluoride becomes the end waste product. Acid-free process wastewater that may include rainwater from the acid-free area of the units is directed to wastewater treating systems. As a general rule there are no nonacid liquid hydrocarbon waste streams.

An acidic liquid waste stream originates from the HF acid regenerator bottoms. This material is neutralized, after which the hydrocarbon portion (heavy

[b]Teflon is a trademark of E.I. DuPont de Nemars and Co. Kel-F is a trademark of Minnesota Mining and Manufacturing Co.

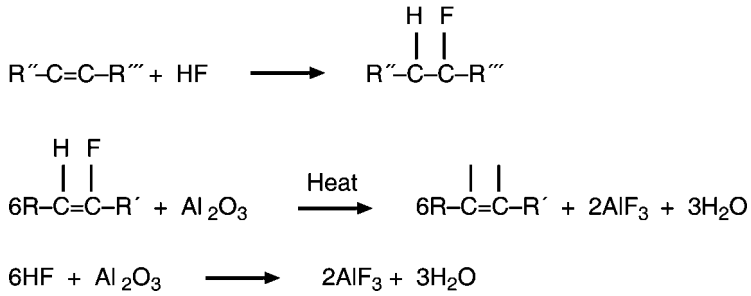


Fig. 9 Alkylfluoride formation and alumina treating in HF units. *(View this art in color at www.dekker.com.)*

alkylate) may be burned in a process heater to recover its heating value. A neutralization basin collects acidic water streams from acid regenerator bottoms (water that enters the alkylation unit usually exits through the regenerator bottoms); rainwater from the acid sections of the unit; and water used for cleaning acid-section vessels or other equipment during shutdowns. In this basin, excess lime is commonly used to neutralize the HF acid portion of these streams. Neutralized water flows from the neutralization basin into the nonacidic wastewater stream and is sent to wastewater treating facilities. Calcium fluoride sludge, which is removed periodically from the neutralization basin, is the end waste product.

SOLID-BED ALKYLATION—UOP/CEPSA DETAL PROCESS

Processing Conditions

The most recent advance in detergent alkylation is the development of a solid catalyst system. UOP and CEPSA (Compania Espanola de Petroleos SA) jointly developed the Detal process, which uses a fixed-bed heterogenous catalyst system for the production of LAB. Detal technology was commercialized by Petresa, a subsidiary of CEPSA, in Quebec, Canada, in 1995.[21] By then, two additional Detal units are in operation. Unlike HF and $AlCl_3$ catalysts, the Detal catalyst is noncorrosive and eliminates problems associated with the handling and disposal of HF (see HF alkylation section Waste Mangement and Metallurgy). Typical solid acid catalysts are amorphous and crystalline alumino-silicates, clays, ion-exchange resins, mixed oxides, and supported acids.[22] Among these solid acid catalysts, ZSM-5, Y-type zeolites, mordenite, and, more recently, MCM-22 and β-zeolite have been identified as the new commercial catalysts for various aromatic alkylation reactions.[23,24] A variety of other potential solid acid catalysts for detergent alkylation have been described in the literature.[25–28]

Flow Scheme

The flow scheme of the Detal process is presented in Fig. 10. The olefin feed and recycle benzene are

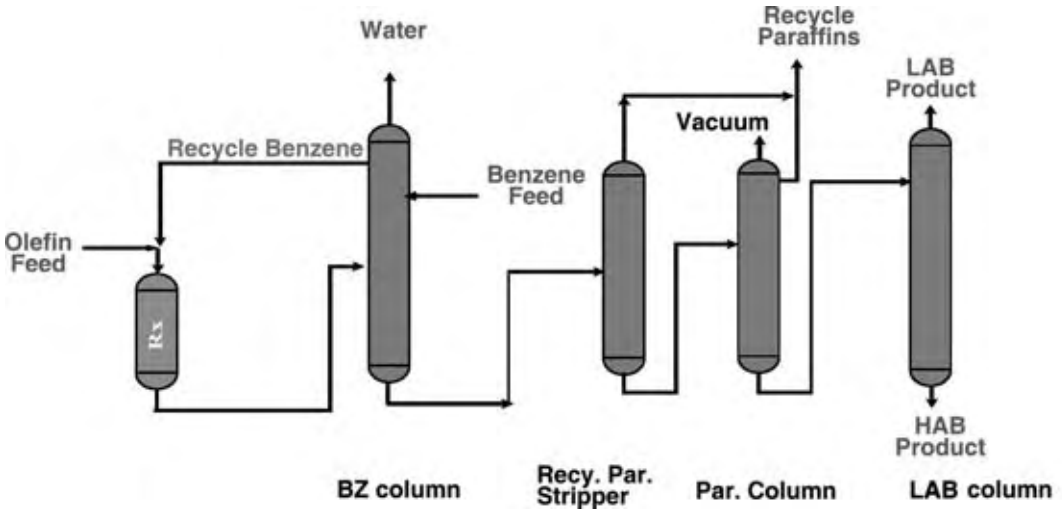


Fig. 10 UOP/CEPSA Detal process for the production of LAB. *(View this art in color at www.dekker.com.)*

combined with makeup benzene before introduction to the fixed-bed reactor containing the solid acid catalyst. As with HF alkylation, an excess of benzene is required. The reaction occurs in the liquid phase and under mild conditions to achieve optimal product quality. The reactor effluent flows directly to the fractionation system, which is similar to that for the HF alkylation process. Because the Detal process eliminates the need for HF, carbon steel metallurgy can be used in vessel and piping construction. Also, the special equipment required for HF handling is no longer needed.

UOP DeFine and PEP units must always be included in Detal-based LAB complexes for optimal performance. All LAB complexes built since 1995 employ solid-bed heterogenous alkylation catalyst system.

CONCLUSIONS

Linear alkylbenzene enjoys worldwide acceptance as a cost-effective and environmentally friendly intermediate used in the production of household and industrial cleaning formulations. Both HF acid and solid-bed alkylation processes are cost-competitive routes for the manufacture of LAB. Solid-bed alkylation process produces fewer environmentally hazardous by-products and employs a less complex facility when compared to HF route. With the breakthrough development of commercially demonstrated and expected future refinement, the solid-bed alkylation technology is expected to meet the continued growth in LAB.

REFERENCES

1. Swisher, R.D. *Surfactant Biodegration*, 2nd Ed.; Marcel Dekker: New York, 1970.
2. Huber, L. Soap cosmet. Chem. Spec. **1989**, *65* (5), 44–46.
3. Davidson, A.S.; Milwidsky, B.M. *Synthetic Detergents*, 7th Ed.; John Wiley & Sons, Inc: New York, 1987.
4. Berg, R.C.; Illingworth, G.E. Linear internal olefins. In *A Commercial Intermediate for Detergents*; Detergent Conference, Barcelona, Spain, Mar 4, 1976.
5. Broughton, D.B.; Berg, R.C. Production of linear mono-olefins by Pacol dehydrogenation and Olex extraction. NPRA Annual Meeting, San Antonio, TX, Mar 23–26, 1969.
6. Imai, T.; Kocal, J.A.; Vora, B. Sci. Technol. Catal. **1994**, 339.
7. Berg, R.C.; Vora, B.V. *Encyclopedia of Chemical Processing and Design*; Marcel Dekker: Vol. 15, 266–284.
8. ARCO Technology, Inc. Hydrocarbon Process. **1985**, *64* (11), 127.
9. Euteco Impianti SPA. Hydrocarbon Process. **1981**, *60* (11), 175.
10. Cavalli, L.; Landone, A.; Pellizzan, T. Linear alkylation for detergency—characterization of secondary components. XIX Jornadas Del Comite Espanola De la Detergeneia, Barcelona, Spain, 1988; 41–52.
11. Berna, J.L.; Moreno, A. European Patent EP, 353,813, 1990
12. Vora, B.V. US. Patent 4,523,048, 1985.
13. Vora, B.V.; Ellig, D.L. US. Patent 4,761,509, 1988.
14. Allinger, N.L.; et al. *Organic Chemistry*, 2nd Ed.; Worth Publishers: New York, 1976; 339.
15. Linfield, W.M., Ed.; *Anionic Surfactants*, Part 1; Marcel Dekker: New York, 1976; Vol. 7, 258.
16. Keim, W.; Roper, M. *Ullmann's Encyclopedia of Industrial Chemistry*; Gerhartz, W., Ed.; VCH Verlagsgesellschaft: Weinheim, 1985; Vol. A1, 195.
17. Vora, B.V.; Pujado, P.R.; Allawala, M.A.; Fritsch, T.R. Production of biodegradable detergent intermediates. Second World Surfactants Congress, Paris, France, May 24–27, 1988.
18. Thorton, D.P. Corrosion free HF alkylation. Chem. Eng. **1970**, *77* (13), 108–112.
19. Forrey, K.; Schrage, C. Trouble shooting HF alkylation. Hydrocarbon Process. **1996**, *45* (1), 107–114.
20. Crowe, W.H.; Marysiuk, A.M. How to work safely with HF alkylation. Hydrocarbon Process. **1965**, *44* (5), 192–194.
21. Vora, B.V.; Pujado, P.R.; Imai, T.; Fritsch, T.R. *Recent Advances in the Production of Detergent Olefins and Linear Alkylbenzenes*; Society of Chemical Industry: University of Cambridge, U.K., Mar 26–28, 1990.
22. Tanabe, K. *Solid Acids and Bases*; Kodansha: Tokyo, 1976; Vol. 1.
23. Cheng, J.C., et al. A comparison of zeolites MCM-22, beta, and USY for liquid phase alkylation of benzene with ethylene. Sci. Technol. Catal. **1988**, *6*, 52–60.
24. Schmidt, R.J.; Zarchy, A.S.; Peterson, G. Paper 124b. AIChE Spring Meeting, Mar 10–14, 2002.
25. Eur. Chem News **1990**, *54* (1428), 26.
26. Venuto, P.B.; Hamilton, A.L.; Landis, P.S.; Wise, J.J. Organic reactions catalyzed by crystalline aluminosilicates. J. Catal. **1996**, *5*, 272.
27. Cao, Y.; Kessas, R.; Naccache, C.; Taarit, Y. Alkylation of benzene with dodecene. The activity and selectivity of zeolite type catalysts as a function of the porous structure. Appl. Catal. A Gen. **1999**, *184*, 231–238.
28. Price, P.M.; Clark, J.H.; Martin, K.; Macquarrie, T.W.; Bastock, T.W. Enhanced selectivity in the preparation of linear alkylbenzenes using hexagonal mesoporous silica supported aluminum chloride. Org. Process Res. Dev. **1998**, *2*, 221–225.

Detergent Enzymes

Michael R. Stoner
Department of Chemical Engineering, University of Colorado, Boulder, Colorado, U.S.A.

Douglas A. Dale
Alfred Gaertner
Genencor International, Palo Alto, California, U.S.A.

Theodore W. Randolph
Department of Chemical Engineering, University of Colorado, Boulder, Colorado, U.S.A.

INTRODUCTION

Modern laundry and dishwashing detergents contain enzymes to improve cleaning efficiency. Proteases are most commonly used, but other enzymes such as amylases, cellulases, and lipases are becoming increasingly popular. This entry describes the important properties of each enzyme class, introduces some of the obstacles involved in formulating them into detergent products, and provides information about how detergent enzymes are commercially produced.

BACKGROUND

The remarkable catalytic properties of enzymes have long been harnessed for human purposes. Ancient people unknowingly used enzymes for such processes as tanning, linen making, and indigo production.[1] Modern biotechnology and recombinant DNA technology have made possible not only to produce and purify large quantities of enzymes, but also to improve upon their naturally occurring properties through protein engineering. This has spawned a $1.5 billion annual market for ''industrial enzymes'' with over 500 commercial products.[2] Applications involve food production (dairy, brewing, baking, wine making, etc.), animal feed (largely as digestive aids), production of fuel ethanol, and various technical uses in the starch, pulp and paper, textile, and leather industries. However, the most important application of industrial enzymes, in terms of both value and volume, is in detergent products.[1]

More than 80% of all laundry detergent products sold in the United States, Europe, and Japan contain enzymes.[3] Several important market drivers support this high level of penetration. First, the last several decades have been characterized by profoundly increased worldwide awareness of the need to lower energy costs and reduce environmental pollution. This has pushed the detergent industry toward enzyme-containing products as biodegradable, low-temperature alternatives to formulations that rely on phosphate builders and high wash water temperatures. Additionally, the stain-removal performance of enzymes, even at very low concentrations, has allowed detergent manufacturers to continue their trend toward increasingly concentrated products, reducing production and shipping costs.

But perhaps the most important business driver for the use of enzymes in detergent is the need for product differentiation and branding. This problem is well illustrated in the U.S. domestic market for detergent products, which generated $3.2 billion in sales during 2003.[4] As a mature market, sales growth is essentially limited by the rate of population growth. Thus, gaining and protecting market share is crucial for the success of detergent manufacturers. With the rise of discount mass-market retailers and warehouse clubs, private label detergent products have become widespread, creating price pressure for established market leaders. This has increased the need for technical innovation to differentiate branded products, leading to the development and commercialization of several new types of detergent enzymes.

TYPES OF ENZYMES USED IN DETERGENTS

To be useful as a detergent additive, an enzyme must be compatible with existing laundry practices while simultaneously boosting soil-removal performance. The criterion of enhancing performance is often easily met, as most stains and soils are biological in nature, and enzymes have evolved to break down all kinds of biological materials, including the proteins, starches, and oils that are common laundry soils. There are some problems, however, in that many stains are

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007733

composed of denatured, cooked, baked, or gelatinized food products, which can be difficult for detergent enzymes to digest. For instance, bleach-containing detergents can cross-link the protein molecules that comprise some stains, thus rendering the soil less accessible to enzymatic removal.

Compatibility with harsh cleaning environments is often more difficult to achieve. Detergent enzymes must perform well in the soft, alkaline wash water that is best for solubilizing common laundry soils. These optimized wash water conditions were serendipitously identified in ancient times, when people cleaned their clothes by trampling them on muddy banks of rivers and then rinsed them with clean water. The process was much more effective in soft water, leading people to congregate in areas with soft water to do their laundry.[5] The beneficial aspects of washing clothes in alkaline solution were first identified around 3,500 BC, when people started adding such materials as soapwort, ashes, and soda to their wash water. These materials boost alkalinity, and accordingly, detergency.[5]

The most notable aspect of modern detergent formulations is the presence of synthetic surfactants. Laundry products contain a mixture of anionic and nonionic surfactants at weight percentages as high as 50%.[6] Modern formulations usually also contain bleaches and water-softening builders. Detergent enzymes must be compatible with all of these materials, representing a significant formulation challenge that is discussed later.

The duration and temperature conditions of automatic washing machine cycles must also be taken into account during the selection of detergent enzymes. These can vary widely from region to region. For example, in Europe, laundry cycles can be more than an hour at temperatures reaching as high as 90°C.[7] In contrast, United States wash cycles are much shorter, lasting between 10 and 12 minutes, at the more moderate temperature of ''hot'' domestic tap water.[7] In Japan, clothes are washed for 10 minutes after a lengthy presoak period in cold or lukewarm water.[8] In general, however, wash temperatures are falling worldwide owing to the desire to conserve energy. Clearly, a versatile detergent enzyme for the worldwide market must be stable for long periods at high temperatures, yet also provide significant activity at lower temperatures.

Despite these obstacles, several classes of enzymes have been successfully incorporated into detergent products. These include proteases, amylases, lipases, cellulases, and mannanases. Market data are shown in Table 1, and each class of enzyme is discussed in further detail in what follows.

Protease

Proteases are enzymes that catalytically degrade proteins, making them useful in the removal of all manner of proteinaceous and protein-containing stains such as blood, egg yolk, milk, grass, and collar soil. Because they perform well against such a wide variety of stains, proteases are the most commonly used type of detergent enzyme (see Table 1).

Proteases were first introduced as a detergent product in 1931 by the German scientist Otto Röhm.[3] This product was essentially a mixture of animal pancreatic enzymes including trypsin and chymotrypsin. It was not very effective, as these enzymes lack broad substrate specificity and do not exhibit high activity or stability at alkaline pH.

World War II spurred much advancement in the production of penicillin, particularly the advent of submerged fermentation processes. This new technique was soon adapted to the large-scale production of bacterial proteases. The first modern detergent protease, isolated from *Bacillus licheniformis*, was introduced in 1962.[3] This enzyme, a particular type of protease called a subtilase, was stable at higher temperatures, had broad substrate specificity and worked well in alkaline conditions. The appearance of this enzyme and others similar to it (i.e., subtilisins) opened up the detergent enzymes market, and by 1969, 50% of the laundry detergent products sold in the United States and Europe contained enzymes.[3]

The subtilisins used in today's detergent products are usually from *Bacillus lentus* and other *Bacillus* species. The survival and growth of these soil-dwelling organisms involves secreting large amounts of

Table 1 1995 Worldwide market for detergent enzymes in millions of U.S. dollars

	Protease	Lipase	Cellulase	Amylase	Total
United States	80	10	40	10	140
Western Europe	100	20	45	20	185
Japan	30	15	20	–	65
Others	110	10	5	–	125
Total	320	55	110	30	515

(From Ref.[34].)

hydrolytic enzymes (such as subtilisin), then absorbing the resulting extracellular degradation products as building blocks and energy sources. Biologically, this is an interesting feat as the organism is able to produce a highly active and nonspecific protease without disrupting or destroying its own internal enzymatic processes. The subtilisin genome indicates how this is possible: the gene contains a signal sequence that targets it for export through the Sec translocase complex. Once outside the cell, the signal sequence, or ''prodomain,'' is proteolytically clipped from the molecule. The final folding of subtilisin is then catalyzed by foldases, bringing it into its final, active conformation.[9] This extracellular secretion mechanism of subtilisin by *Bacillus* species has enabled the production and purification of large amounts of the enzyme in a cost-effective manner.

Another beneficial property of the commonly used *Bacillus* subtilisins is their broad substrate specificity. In contrast to other serine endopeptidases like trypsin and chymotrypsin, the subtilisins have elongated substrate binding sites that allow them to accommodate a wide variety of different peptide sequences.

There are two important disadvantages of subtilisin as a detergent protease. The first is that it is susceptible to methionine oxidation, which results in sharply decreased thermal stability and concomitantly reduced proteolytic activity at high wash temperatures. A further disadvantage is its dependence on bound Ca^{2+} ions for thermodynamic stability. Builders and other chelating detergent ingredients can remove the metal ions, destabilizing the active, properly folded conformation of the enzyme.

A tremendous amount of protein engineering effort has been directed at improving subtilisin as a detergent enzyme. Indeed, patents exist for single amino acid substitutions at nearly every site along the 275 amino acid sequence of subtilisin.[2] The much larger space defined by multisite mutations has not yet been thoroughly explored;[10] however, more sophisticated techniques such as DNA shuffling have already been successfully applied in the development of improved proteases.[11]

Amylase

After protease, the next most common detergent enzyme is α-amylase (or simply amylase). This class of enzymes hydrolyzes the α-1,4-linkages that join glucose subunits into starch polymers. Because starch is often present in food, amylase is particularly effective against food-based stains. Also, as more and more processed food products include starch as a thickening agent, amylase is becoming increasingly useful as a detergent additive. The trend toward lower wash water temperatures has added to the popularity of amylase in detergent; starch solubility is limited in cold water, and by producing the more soluble oligosaccharide and dextrin degradation products, amylase contributes significantly to stain-removing power.

Typically, the amylases used in detergent are derived from bacterial sources. *B. licheniformis* amylase is particularly desirable for detergent applications as it is highly thermostable and maintains its activity even at high pH. The primary drawback of *Bacillus*-derived amylases as detergent enzymes is that their stability (and in many cases also their activity) relies on bound Ca^{2+} ions.[12] This is problematic in detergent formulations that must also contain detergent builders to limit water hardness. Oxidation of a methionine residue close to the active site of *B. licheniformis* amylase is another common stability problem.[12]

Lipase

The use of lipase in detergent has also been driven by the global trend toward lower wash water temperatures. This is because lipase decomposes fats and oils, compounds that are generally only sparingly soluble in cold detergent mixtures, into smaller, more readily solubilized subunits. Lipase is particularly effective against triglyceride-containing stains such as butter, lipstick, and sebum (collar stain).

Lipase is unique amongst the detergent enzymes as its catalytic activity primarily occurs in oil–water interfaces. This has the important ramification that most of its stain-removal effectiveness comes not during the wash cycle, but during the subsequent drying period, when total water content is as low as 25%.[3] Thus, the benefit of lipase in a detergent formulation is not apparent until the second wash cycle. Perhaps not surprisingly, the interfacial activation phenomenon of detergent lipases is easily disrupted by the presence of surfactants.[3] This poses a difficult formulation problem that has limited the widespread usage of lipase in detergent products. Moreover, the release of medium chain fatty acids by lipases has been implicated in the creation of unpleasant odors.

Cellulase

The need for product differentiation has been the primary motivating factor for the addition of cellulase to detergent formulations. This type of enzyme degrades the cellulose that comprises cotton fibers—a fact that is counterintuitive because clothes, not stains, are made of cotton. However, cellulase is beneficial in that the enzyme preferentially acts on the cotton microfibrils that are worn or damaged, removing them from the surface of the fabric and making the garment appear softer and brighter. Removal of the fuzzy outer layer

of the cotton fibers has the added effect of making the fabric more resistant to encrustation of other soils, such as hard water deposits.[3] Detergent cellulases are generally derived from fungal sources such as *Humicola insolens* or *Trichoderma reesei*.

Mannanase

The newest type of detergent enzyme is mannanase. These enzymes catalytically degrade the β-1,4-linked galactomannans commonly found in the seeds of leguminous plants and beans of carob trees and in galactoglucomannans as components of softwood hemicelluloses. These compounds are often referred to as gums, and they are increasingly used as thickeners and gel texture agents in food and consumer products. For example, ice cream, barbecue sauce, processed foods, salad dressing, hair styling aids, and cosmetic makeup all typically contain gums that are susceptible to digestion by mannanase. The gums act like glue for particulate soils, making the clothing fabric appear dull. Full scale washing tests followed with stain-removal evaluation by expert panels have shown that mannanases are highly effective against a broad range of common fabric stains.

FORMULATION CONCERNS

Enzymes are large molecules comprising tens to hundreds of amino acid subunits. Their unique catalytic properties are associated with specific conformations of the amino acid chain. However, these active conformations are thermodynamically quite fragile, typically only 5–15 kcal/mol more stable than unfolded conformations. Additionally, the amino acid subunits that define the protein contain a variety of reactive groups that are subject to chemical degradation pathways resulting in loss of structure and/or activity. Thus, ensuring the physical and chemical integrity of an enzyme's ''native'' structure is a primary concern when adding enzymes to detergent products.

Powdered Formulations

Because of the long-term instability of proteins in aqueous solution, enzyme producers and formulators have attempted to produce stable solid enzyme formulations since enzymes were first used in detergents. Initially, commercially produced protease-containing detergents contained spray dried enzymes. As discussed earlier, proteases have the ability to digest themselves via autolysis and are often incompatible with surfactants. These problems are easily overcome by storing the protease in the solid state. Simple spray-drying provides a fast and cost-effective way of compartmentalizing two incompatible ingredients and maintaining protease activity over a typical detergent product's shelf life. However, the technologies used to produce modern powdered enzymes have evolved far beyond simple spray drying.

The main reason for this is the possible immunogenic nature of enzymes, which, similar to most other proteins, can cause allergic reactions if not handled in a safe manner. Thus, modern enzyme granules need to provide a tough barrier to prevent release of enzyme dust, whilst providing quick release once in the wash application. Enzyme granules contain coatings that are designed to withstand the physical impact and the shear forces typically encountered during powder processing. Many ways of producing enzyme granules have evolved, with just three methods being currently in common use for detergents. These include drum granulation, marumerization/spheronization, and fluid bed coating (Fig. 1). An overview is provided in Table 2. In general, fluid bed technology is the most flexible approach, giving granules a uniform appearance and smooth coating (Fig. 2), while other technologies have an edge in either cost or throughput, as they are more amenable to continuous operation.

Powdered detergents, even though appearing dry, contain between 10 and 30% moisture, allowing slow diffusion of water and other molecules in and out of enzyme granules. Bleach, and in particular its active ingredient, hydrogen peroxide, can travel together with water through thin coating layers over long storage periods. This can cause oxidation of various chemical groups on the enzyme molecule. Thus, enzyme granules for bleach-containing detergents comprise additional protective measures, either in physical form, such as through barrier layers or via chemical additives.

Liquid Formulations

The successful application of enzymes in liquid detergent formulations presents several technical obstacles that are not encountered in detergent powders. These problems stem from the fact that liquid detergent products are complex aqueous solutions, and physical separation of enzymes from potentially harmful detergent components is impractical. Table 3 describes a typical heavy-duty liquid detergent (HDLD) formulation. Surfactants, bleaches, and water-softening builders—all necessary ingredients for a viable detergent product—can affect the physical and chemical stability of enzymes. Additionally, as newer types of detergent enzymes are added to formulations that already contain proteases, proteolytic degradation of enzymes is a concern.

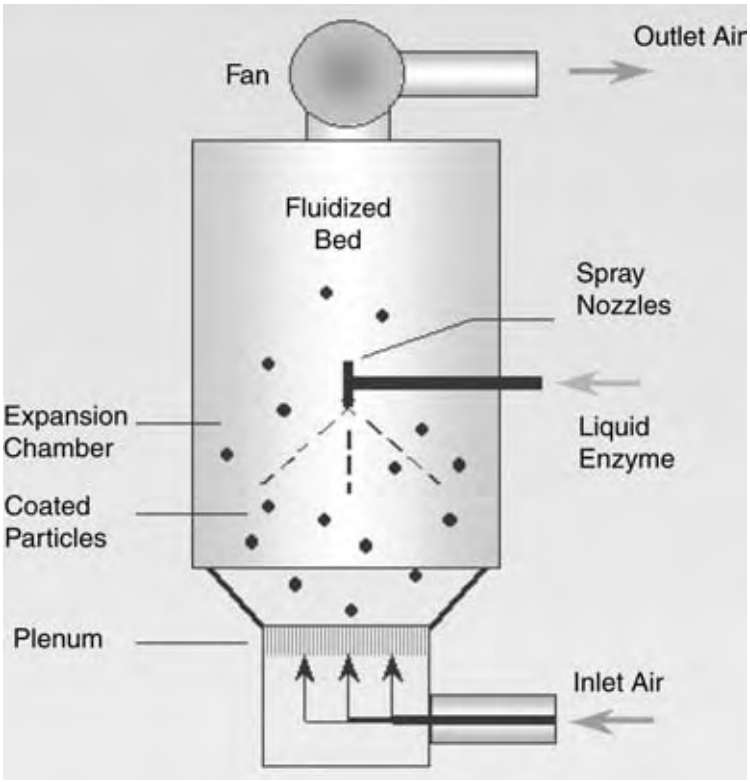


Fig. 1 Schematic view of a fluid bed granulator. *(View this art in color at www.dekker.com.)*

Market trends highlight the importance of overcoming these technical obstacles. In the United States, consumers increasingly choose liquid detergent products over powdered formulations. In 2003, liquid products outsold powders by roughly 2 : 1 margin.[4] This trend is expected to continue, and considerable research resources have been directed toward the challenge of improving enzyme stability in liquid detergent products. Liquid detergents provide convenience to the consumer and eliminate possible solubility concerns, particularly in regions where low wash water temperatures are prevalent.

The patent literature contains numerous examples of enzyme stabilization schemes based on the use of various chemical additives. Common themes are to either lower the amount of water in the formulation while retaining solubility of other components, or to add specific enzyme stabilizer/inhibitors. Some examples are highlighted in Table 4, and many others can be found in the review by Crutzen and Douglass.[3] Few published articles have addressed the mechanisms of enzyme stabilization by additives. One study examines protease stabilization by carboxylic acid salts[13] and another discusses the use of borate in conjunction with propylene glycol to inhibit protease activity.[14]

The primary disadvantages of using chemical additives as enzyme stabilizers are their cost and that valuable formulation space is lost without any direct benefit to detergent performance. Thus, an alternative research strategy involves understanding the detrimental effects of standard detergent ingredients on enzymes. This research has been primarily focused on the adverse effects of surfactants and surfactant blends. Kravetz and Guin demonstrated that anionic surfactants (e.g., linear alkylbenzene sulfonate, α-olefin sulfonate) are more harmful to protease and amylase than nonionic surfactants (e.g., alcohol ethoxylates).[15] Lalonde et al. explored the influence of total surfactant content and surfactant type on protease stability, concluding that the detrimental effects of anionic surfactants could be reduced by decreasing the water content of the formulation.[16] More recently, it has been shown that enzymes could be stabilized against a relatively harsh anionic surfactant, linear alkylbenzene sulfonate (LAS), by including highly ethoxylated cosurfactants in the formulation.[17] A common conclusion from this research is that anionic surfactants, particularly those that entirely lack ethoxylation, destabilize proteins.

Surfactant-induced unfolding of enzymes is an important issue, and it has been addressed in some more fundamental research studies. In the early 1980s, Jones et al. conducted dialysis experiments to examine binding stoichiometry between surfactants and various protein molecules.[18,19] Their work showed that anionic surfactants bind to proteins in a cooperative way, with hundreds of surfactant monomers binding to a single protein molecule. There is experimental evidence that the resulting unfolded protein–surfactant complex

Table 2 Comparison of production methods for powdered enzymes

Technology	Equipment	Advantages	Disadvantages
Prilling (hot melt)	Prilling tower	Continuous process High capacity Easy rework	High product dust Low melting point Requires predried enzyme
Extrusion/spheronization (marum)	Screw extruder, Marumerizer, Fluid bed coater	Inexpensive raw materials Continuous process	Complex multistep process High dust during process and in product Requires highly concentrated enzyme
Drum granulation (high shear)	Lödige mixer, Littleford mixer, Fluid bed or drum coater	High capacity Continuous process Tough granules	Dusty, multistep process Requires highly concentrated enzyme Wide particle size distribution
Fluid bed (layering/coating)	Fluid bed coater	Single, contained reactor Flexible formulation Tough, uniform granules	Batch process Difficult rework

consists of the protein chain wrapped around micelle-like surfactant aggregates. Taken together, these studies indicate that surfactant-induced unfolding is generally a thermodynamically favorable event.

However, even if thermodynamics dictate that unfolding of enzymes in the presence of surfactants will eventually occur, the kinetics of the unfolding process is a much more relevant issue for the detergent formulator. Otzen et al. conducted an extensive protein engineering study to compare the unfolding of cellulase in detergent and guanidine hydrochloride.[20] They reached several important conclusions: amino acid substitutions that increased stability against guanidine generally did not bring about a concomitant increase in stability against surfactant, and electrostatic interactions between cationic amino acids and anionic surfactant molecules significantly influenced the overall unfolding rate. Further studies suggest that the size and shape of surfactant aggregates may also affect the unfolding rate of proteins in detergent.[21]

PRODUCTION OF DETERGENT ENZYMES

Identification of Target Enzymes

Numerous methods have been used to successfully identify new enzymes for detergent applications. Historically, the method of choice has been to screen for new enzymes from naturally occurring microbes. The most commonly used protease, subtilisin, was initially isolated in this manner. More recently, researchers have resorted to exploring extreme environments, such as hot springs, to more closely match

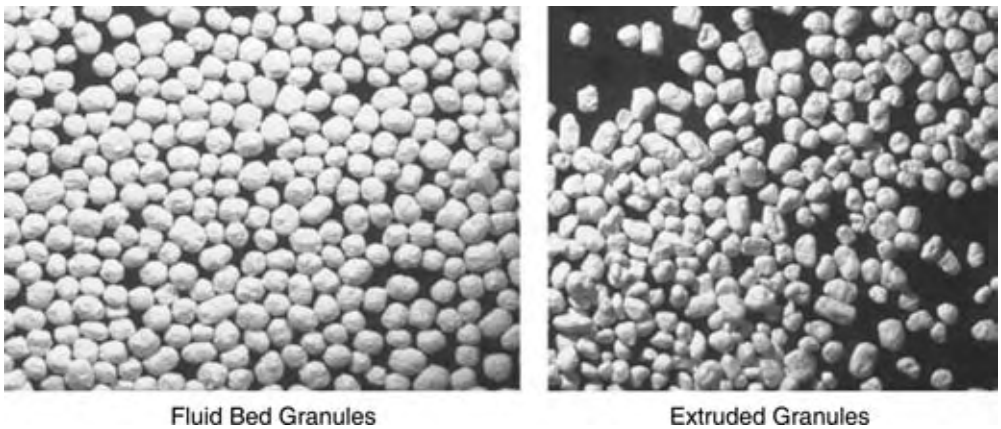


Fig. 2 Comparison of enzyme granules produced by fluidized bed and extrusion methods. *(View this art in color at www.dekker.com.)*

Table 3 Example of heavy-duty liquid detergent (HDLD) formulation

Ingredient	Weight percent
Anionic surfactant (typically alkylbenzene sulfonate, alkyl ether sulfate)	0–40
Nonionic surfactant (alcohol ethoxylate)	0–10
Builder (sodium carbonate, zeolite)	0–35
Bleach (sodium perborate)	0–10
Polymer (product stabilizer)	0–1
Enzyme(s)	0–2
Enzyme stabilizer(s)	0–5
Preservative	0.05–0.2
Fragrance	0–0.6
Colorant	0–0.2
Fluorescent brightener	0–5

(From Ref.[6].)

the enzymes' natural environment with that found in a washing machine. This is often easier said than done in that the so-called ''extremophiles'' living in these relatively harsh environments can be difficult or impossible to culture in a laboratory setting. One way to overcome this problem is to extract the target microbe's DNA, screen it for useful protein sequences, and, if necessary, express the sequences of interest in different host organisms.[22] These methods have been enhanced by the application of robotics and other high throughput techniques, making ''genome-based screens'' a very attractive method of identifying new detergent enzymes.

A variation of this technique is the so-called ''homolog screen,'' whereby sequences of related families of proteins are aligned and probed for the best or most stable mutant.[23] Together with computational methods to identify and characterize the proteins, these technologies have enabled identification and targeted commercialization of new enzymes on an unprecedented short timescale.

Table 4 Patented additives for enzyme stabilization in liquid detergents

Patent	Stabilization scheme
US Patent 4,462,922	Boric acid, polyol, sodium sulfite
US Patent 4,532,064	Boron compounds, dicarboxylic acids
US Patent 4,404,115	Sodium pentaborate
US Patent 4,537,707	Boric acid and formate

Optimization of Enzymes for Detergent Applications

Before the advent of protein engineering, naturally occurring enzymes were commercialized without any form of modification or optimization. An initial technology dislocation occurred in the 1980s and 1990s, when the advent of protein engineering facilitated the improvement of enzyme performance to levels previously thought to be impossible.[6,24] Other methods for optimizing newly identified detergent enzymes include directed evolution or gene shuffling.[25,26]

In either case, whether the enzyme was initially identified using classical or postgenomic screening methods, the proven screener's mantra also applies to optimization: ''You get what you screen for!'' This phrase implies that if the optimization criteria are not specifically designed to mimic the conditions the enzyme encounters during its application, the results of protein selection and improvement will not necessarily meet the appropriate performance criteria. Indeed, heavy reliance on these modern high-throughput technologies without keeping an eye on real-world applications has, somewhat unjustifiably, led to a backlash against applying these techniques in both therapeutic and industrial biotechnology applications.[27]

Fermentation Processes

As evidenced by centuries-long experience with the arts of brewing, yoghurt and cheese making, humans long ago mastered the process of microbial fermentation. Nevertheless, early detergent enzyme production processes relied on extraction from plants or animals. In contrast, modern detergent enzymes are mostly produced using micro-organisms in large-scale fermentations. While there are variations on the theme, such as the traditional Koji process, a moist tray fermentation used in the soy sauce and sake industry and to produce specialty enzymes, today's detergent enzymes are mostly grown in submerged culture in stainless steel fermentation tanks. Both prokaryotic and eukaryotic organisms are used for production. In the case of prokaryotes, *Bacillus* species are common for the production of proteases and amylases, and eukaryotes are often used to manufacture cellulases. A typical industrial fermentor is shown in Fig. 3. Because microbial fermentations require the presence of a sufficient amount of biomass to yield economically viable levels of enzyme, fermentor volumes can be quite large, ranging up to 1,000 m^3 in size. Each large-tank inoculum begins with a seed vial, proceeding through a series of tanks of intermediate size until sufficient cell mass is obtained to inoculate the final fermentor. Many attempts have been made to improve this process, with

Fig. 3 A large-scale fermentor for the production of detergent enzymes. *(View this art in color at www.dekker.com.)*

the most notable improvement being the development of fed batch fermentations, where a slow feed of a limiting nutrient creates controlled growth and enzyme production conditions. Another principal achievement was to engineer bacterial strains that are not capable of sporulation and produce high amounts of enzyme.[28]

Much effort has also been directed toward the development of fully continuous chemostat-type fermentations in which a tank produces a steady stream of enzymes. In this type of operation, product-containing fermentation broth is drawn off while new media is fed to the tank to maintain a constant volume. Reasonable successes with this method have been reported in the biopharmaceutical industry. However, similar examples of successful application of this technique have not yet appeared in the industrial biotechnology literature.

Microbial fermentations must be tightly controlled to ensure optimal growth of micro-organisms and efficient production of enzymes. As shown in Fig. 4, a modern fermentor will allow operators to control the temperature, pH, redox potential, and dissolved oxygen of the growth medium. Large-scale fermentation operations have been optimized and standardized to the point that most room for improvement is associated with the physiology of the microbial production strains themselves. Thus, much research has been done to explore and understand both the metabolic and the catabolic pathways of enzyme-producing bacterial strains, facilitating the design of even more highly optimized fermentation processes.[29] This knowledge has also allowed enzyme manufacturers to implement effective environmental control systems aimed at protecting workers, products, and the environment from the release of enzymes or micro-organisms.

Purification and Formulation

In contrast to fermentation processes where reasonably standardized equipment is used, a number of varying techniques are necessary to accomplish purification and formulation of enzymatic products from the

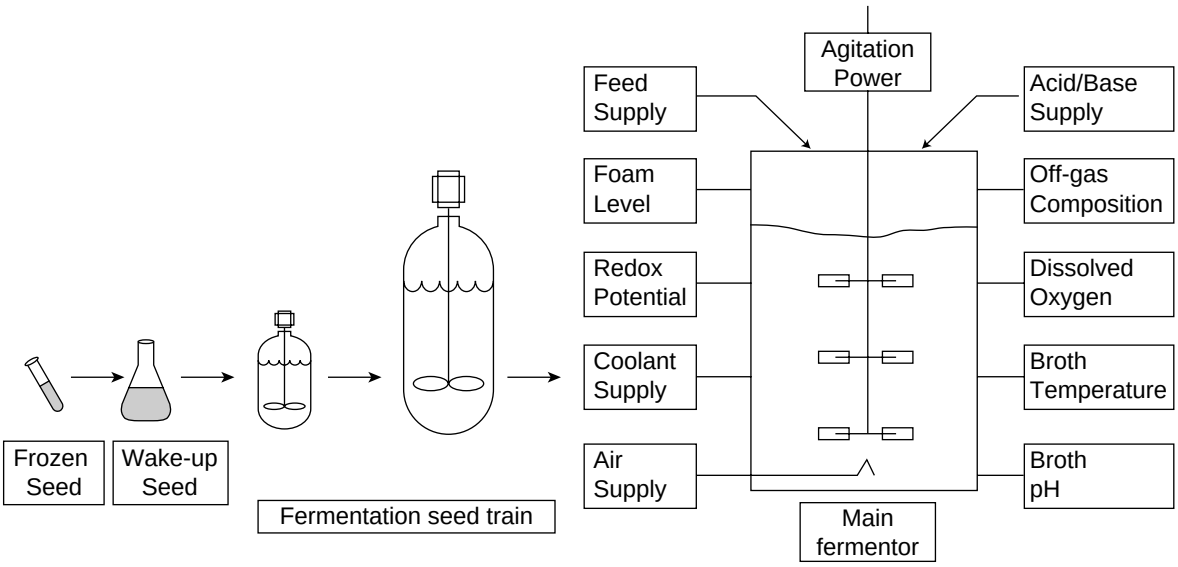


Fig. 4 Industrial fermentation scheme.

fermentation broth. The collective term ''downstream processing'' encompasses methods that can vary widely between enzyme products. A schematic overview of a typical downstream process is given in Fig. 5. The early part of the enzyme separation is usually referred to as recovery, while the elements further downstream are called purification.

To begin, enzyme products can be inside the cell (intracellular), loosely associated with the cell, or secreted (extracellular). Each of these products requires a different approach on how to purify the enzyme. The majority of presently commercialized detergent enzymes is of the extracellular variety and can be recovered directly from the fermentation broth. Thus, the primary recovery challenge involves removing the cells from the broth, aptly called cell separation. Three different techniques are commonly used to achieve this goal: filtration, microfiltration, and centrifugation.

Filtration can include filter presses, rotary vacuum drum filters, and variations on synthetic membrane filtration equipment, such as filter cartridges, pancake filters, or cellulosic filter pads (Fig. 6). These processes typically operate in a batch mode: When the filter press is filled up or the vacuum drum cake is exhausted, a new batch must be started. This type of filtration is also called dead-end filtration because the only fluid flow is through the membrane itself.

Microfiltration is in principle a more technically advanced version of classic dead-end filtration processes. Microfiltration, along with related technologies such as ultrafiltration, nanofiltration, and reverse osmosis, relies on synthetic or ceramic membranes with a defined pore size. These operations use ''cross flow'' or ''tangential flow'' in which the filtrate is pumped in a direction parallel to the membrane surface.[30] The microbes are retained on one side of the membrane, while the aqueous enzyme solution passes through the membrane. The primary advantage of tangential flow is that membrane fouling is reduced to fouling of the membrane pores themselves while particulate

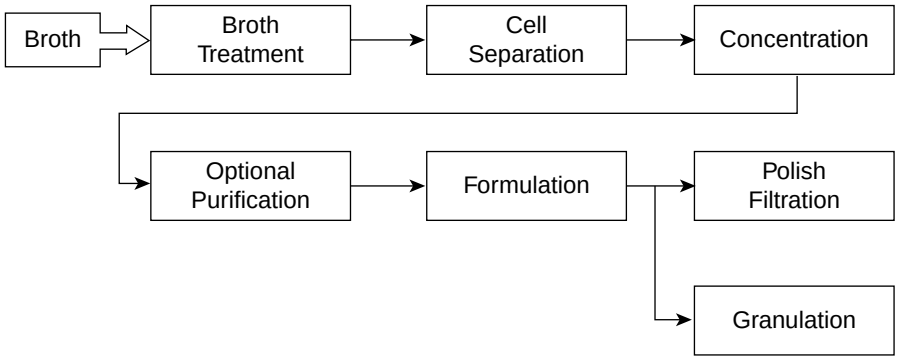


Fig. 5 Recovery of industrial enzymes.

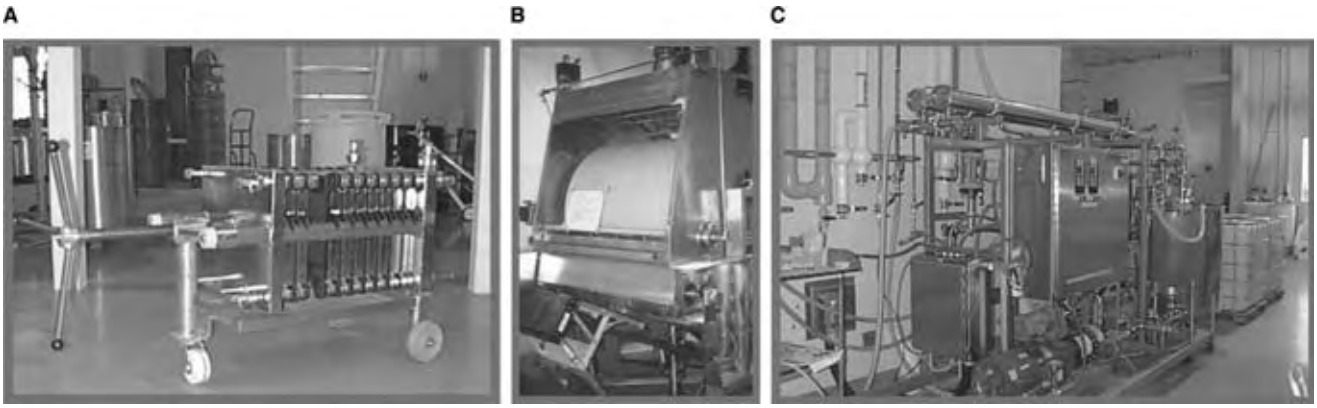


Fig. 6 Examples of downstream process equipment: (A) filter press; (B) rotary vaccum drum; (C) microfiltration unit. *(View this art in color at www.dekker.com.)*

matter on top of the membrane is continuously swept away.

Centrifugation is another viable method of removing microbes from the fermentation broth. Two types of centrifuge are available: disk stack centrifuges (clarifiers) and decanter centrifuges, which contain solid removing scrolls. Both devices can be operated in a continuous mode, speeding processing.

Once the cells have been removed, the enzyme broth is concentrated by evaporation or ultrafiltration. In ultrafiltration, membrane pore sizes are much smaller than in microfiltration, allowing only water and small molecules to travel through the membrane. Thus, enzymes can be concentrated under mild process conditions. Typical molecular weight cut off pore sizes for ultrafiltration membranes are 5000, 10,000, or 30,000 Daltons.

After recovery, purification of the enzyme is the next step. This can be achieved in many ways, through precipitation with salts, crystallization, chromatography, and aqueous two-phase extraction. Many of these methods are associated with substantial capital cost, low throughput, or low yields and are not commonly used for detergent enzymes. However, authors have reported the use of crystallization and aqueous two-phase extraction for large-scale preparations.[31,32] These methods are also effective in concentrating enzyme broths.

The final formulation of liquid detergent enzymes includes addition of stabilizers, antimicrobial substances, and glycols. This is followed by a final polish filtration to provide a clear liquid. If the enzyme broth is slated for dry product, the concentrate can either be stored for later use or applied directly for granulates. As with fermentation and downstream processing, formulations are often kept as closely guarded trade secrets in the biotechnology industry, and few publications on new approaches or new technologies are available in the literature.

ENVIRONMENTAL HEALTH AND SAFETY ASPECTS OF ENZYMES

Enzymes are important factors in the worldwide push to use sustainable and environmentally sound technologies. Enzymes have allowed detergent manufacturers to respond to the consumer desire for lower wash temperatures without compromising performance. Additionally, as completely biodegradable catalysts, enzymes do not persist in the environment.

However, there are some risks associated with the large-scale production and use of enzymes. These are: a) release of micro-organisms into the environment; b) contact by people with enzymes at the enzyme and detergent plants; and c) contact with enzymes by the consumer.

Release of enzymes is not considered a problem because detergent enzymes occur naturally in the environment. In fact *Bacillus* species, common hosts for the production of enzymes, are found in the soil and are part of the microbial population used in composting.[5] Nevertheless, the bacterial fermentations carried out in enzyme plants are carefully contained, and no live organisms should be detectable in the final enzyme product. As an additional safety measure, engineered bacterial strains are designed so that they cannot compete effectively with natural strains in the environment.

As is true for most proteins, detergent enzymes can cause allergic reactions in susceptible individuals. Inhalation of enzyme dust poses the greatest threat. These allergies typically manifest themselves with hay fever-like symptoms. The health effects are completely reversible once the allergen has been removed. Another health risk with proteases can be skin irritation; however, this has been described as a mild side effect.[33] Because of these health risks, enzyme and detergent manufacturers have employed advanced granulation technology to protect plant workers and consumers

from exposure to enzyme dust. In addition to engineering the products themselves, engineering controls, work-practice controls, and personal protective equipment have been implemented in plants handling enzymes to minimize exposure. Thus, enzyme technology provides the contemporary detergent manufacturer with a safe alternative to include more harmful chemical catalysts or surface-active agents in their formulations.

CONCLUSIONS

Enzymes are an increasingly important component in detergent formulations, both in terms of effectiveness and as a means of product differentiation. The primary types of detergent enzymes are protease, amylase, lipase, and cellulase; all are derived from bacterial and fungal sources. Enzyme activity is subject to various forms of chemical and physical degradation, problems that are particularly acute in liquid formulations, where the enzymes cannot be physically isolated from the harmful effects of surfactants and bleaches.

REFERENCES

1. Kirk, O.; Borchert, T.V.; Fuglsang, C.C. Industrial enzyme applications. Current Opinion in Biotechnology **2002**, *13* (4), 345–351.
2. Cherry, J.R.; Fidantsef, A.L. Directed evolution of industrial enzymes: An update. Current Opinion in Biotechnology **2003**, *14* (4), 438–443.
3. Crutzen, A.; Douglass, M.L. Detergent enzymes: A challenge! In *Handbook of Detergents Part A: Properties*; Broze, G., Ed.; Marcel Dekker: New York, 1999; 639–690.
4. Branna, T. Is dial a good call for Henkel? Happi **2004**, *41* (1), 78–90.
5. van Ee, J.H. Historical overview. In *Enzymes in Detergency*; van Ee, J.H., Misset, O., Baas, E.J., Eds.; Marcel Dekker: New York, 1997; 1–10.
6. Sachdev, A.; Krishnan, S. Heavy-duty liquid detergents. In *Liquid Detergents*; Lai, K.-Y., Ed.; Marcel Dekker: New York, 1997; 261–324.
7. Bott, R. Development of new proteases for detergents. In *Enzymes in Detergency*; van Ee, J.H., Misset, O., Baas, E.J., Eds.; Marcel Dekker: New York, 1997; 1–10.
8. Aehle, W. Introduction. In *Enzymes in Industry: Production and Applications*; Aehle, W., Ed.; Weinheim, F.R.D.: New York, 2004; 1–12.
9. Wang, L.; Ruvinov, S.; Strausberg, S.; Gallagher, D.T.; Gilliland, G.; Bryan, P.N. Prodomain mutations at the subtilisin interface – correlation of binding-energy and the rate of catalyzed folding. Biochemistry **1995**, *34* (47), 15415–15420.
10. Bryan, P.N. Protein engineering of subtilisin. Biochimica et Biophysica Acta—Protein Structure and Molecular Enzymology **2000**, *1543* (2), 203–222.
11. Ness, J.E.; Kim, S.; Gottman, A.; Pak, R.; Krebber, A.; Borchert, T.V.; Govindarajan, S.; Mundorff, E.C.; Minshull, J. Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. Nature Biotechnology **2002**, *20* (12), 1251–1255.
12. Nielsen, J.E.; Borchert, T.V. Protein engineering of bacterial alpha-amylases. Biochimica et Biophysica Acta—Protein Structure and Molecular Enzymology **2000**, *1543* (2), 253–274.
13. Crossin, M.C. Protease Stabilization by carboxylic-acid salts – relative efficiencies and mechanisms. Journal of the American Oil Chemists Society **1989**, *66* (7), 1010–1014.
14. Stoner, M.R.; Dale, D.A.; Gualfetti, P.J.; Becker, T.; Manning, M.C.; Carpenter, J.F.; Randolph, T.W. Protease autolysis in heavy-duty liquid detergent formulations: Effects of thermodynamic stabilizers and protease inhibitors. Enzyme and Microbial Technology **2004**, *34*, 114–125.
15. Kravetz, L.; Guin, K.F. Effect of surfactant structure on stability of enzymes formulated into laundry liquids. Journal of the American Oil Chemists Society **1985**, *62* (5), 943–949.
16. Lalonde, J.; Witte, E.J.; Oconnell, M.L.; Holliday, L. Protease stabilization by highly concentrated anionic surfactant mixtures. Journal of the American Oil Chemists Society **1995**, *72* (1), 53–59.
17. Russell, G.L.; Britton, L.N. Use of certain alcohol ethoxylates to maintain protease stability in the presence of anionic surfactants. Journal of Surfactants and Detergents **2002**, *5* (1), 5–10.
18. Jones, M.N.; Manley, P. Binding of normal-alkyl sulfates to lysozyme in aqueous-solution. Journal of the Chemical Society—Faraday Transactions I **1979**, *75*, 1736–1744.
19. Jones, M.N.; Manley, P.; Midgley, P.J.W.; Wilkinson, A.E. Dissociation of bovine and bacterial catalases by sodium normal-dodecyl sulfate. Biopolymers **1982**, *21* (7), 1435–1450.
20. Otzen, D.E.; Christiansen, L.; Schulein, M. A comparative study of the unfolding of the endoglucanase Ce145 from Humicola insolens in denaturant and surfactant. Protein Science **1999**, *8* (9), 1878–1887.
21. Otzen, D.E. Protein unfolding in detergents: Effect of micelle structure, ionic strength, pH and temperature. Biophysical Journal **2002**, *83* (10), 2219–2230.

22. Short, J.M. Recombinant approaches for accessing biodiversity. Nature Biotechnology **1997**, *15*, 1322–1323.
23. Lecompte, O.; Thompson, J.D.; Plewniak, F.; Thierry, J.; Poch, O. Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene **2001**, *270*, 17–30.
24. Wells, J.A.; Cunningham, B.C.; Graycar, T.P.; Estell, D.A. Recruitment of substrate-specificity properties from one enzyme into a related one by protein engineering. Proceedings of the National Academy of Sciences **1987**, *93*, 5167–5176.
25. Stemmer, W.P. Rapid evolution of a protein in vitro by DNA shuffling. Nature **1994**, *370*, 389–391.
26. Amin, N.S.; Wong, S.; Schellenberger, V. Direct transformation of site-saturation libraries in *Bacillus subtilis*. Biotechniques **2003**, *35*, 1134–1140.
27. Mullin, R. As high throughput screening draws fire, researchers leverage science to put automation into perspective. Chemical and Engineering News **2004**, *82*, 23–32.
28. Ferrari, E.; Henner, D.J.; Perego, M.; Hoch, J.A. Transcription of *Bacillus subtilis* subtilisin and expression of subtilisin in sporulation mutants. Journal of Bacteriology **1988**, *70*, 289–295.
29. Sanford, K.; Soucaille, P.; Whited, G.; Chotani, G. Genomics to fluxomics and physiomics – pathway engineering. Current Opinion in Microbiology **2002**, *5*, 318–322.
30. Okamoto, Y.; Ohmori, K.; Glatz, C.E. Harvest time effects on membrane cake resistance of *Escherichia coli* broth. Journal of Membrane Science **2001**, *190*, 93–106.
31. Becker, N.T.; Lawlis, B.V. *Subtilisin crystallization process*; Genencor International: United States, 1991.
32. Kula, M.R. Trends and future prospects of aqueous two-phase extraction. Bioseparations **1990**, *1*, 181–189.
33. Caddow, A.J.; Concoby, B. Enzyme safety and regulatory considerations. In *Enzymes in Industry: Production and Applications*; Aehle, W., Ed.; Weinheim, F.R.D.: New York, 2004; 399–411.
34. Houston, J.H. Detergent enzymes' market. In *Enzymes in Detergency*; van Ee, J.H., Misset, O., Baas, E.J., Eds.; Marcel Dekker: New York, 1997; 1–10.

Diamond Films

Angel Velez
Mark A. Prelas
Nuclear Science and Engineering Institute, University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

Advances in technology are oriented not only toward the optimum performance of components, but also in the selection of the most reliable materials that allow technology to push the envelope. With improved material growing techniques, devices are designed to meet the desired characteristics for a specific application. Diamond materials have met this demand with good performance. With the increased use of diamond film in technology, research is constantly improving the materials and opening new doors. For example, diamond is being incorporated in complex multicomponent matrices. This has been possible because of the development of innovative deposition methods to deposit good-quality diamond films that show good adhesion properties and longer lifetimes. Widespread applications of diamond materials have been found in mining, tooling, electronics, optics, medicine, biology, and related fields.

DIAMOND FILMS

What Is a Diamond Film?

The two most important forms of carbon are graphite and diamond. Graphite consists of six-membered aromatic rings of sp^2 hybridized carbon atoms arranged in layers (Fig. 1). The layers are held together by a weak van der Waals force while the atoms in the ring are tightly bound by covalent bonds. The layers can slip, thus giving the graphite excellent lubrication properties. Diamond is the densest phase of carbon. The carbon atoms are covalently bonded through sp^3 bonds forming tetrahedral cells (Fig. 2). A rare form of diamond, hexagonal diamond, called lonsdaleite is also possible (Fig. 3). Essentially, the difference in the structures is the type of hybridization, sp^2 or sp^3, or the ratio of sp^2 and sp^3 bonds and the structure type.

Diamond-like film or amorphous diamond is made up of sp^2 and sp^3 bonded carbon. The amount of sp^2 bonds depends on the method of deposition and can be as high as 60%. The higher the number of sp^3 bonds the harder the material, the higher the band-gap, and the better the overall physical properties of the material.

HISTORY

Natural Diamond

When humankind first discovered diamond is not known, but the fourth century B.C. Sanskrit manuscript, the Artha-Sastra of Kautilya, tells about the diamond trade in India. The diamonds were mined from riverbeds (alluvial deposits). Industrial uses of diamond were described in a second-century Chinese text suggesting diamond trade with India. India was the only source of diamond until the discovery of diamonds in the Diamantina region of Brazil in 1725. Like India, the deposits were alluvial. In 1866, South Africa's alluvial diamond deposits were discovered on the Vaal River. Other alluvial deposits were discovered on The Orange River. In 1870, a series of diggings in present-day Kimberley identified the nonalluvial source of diamond, a volcanic rock that was named kimberlite. Five kimberlite pipes were discovered in this region. In 1888, DeBeers Consolidated Mines Ltd consolidated the diamond mining operations in Kimberley and became the world's leading diamond producer. A new diamond bearing kimberlite pipe was discovered in Pretoria in 1902. This mine, called the Premier Mine, proved to be very productive and was the site of the largest diamond gem discovered, the 3106-carat Cullinan.

A worldwide search for other diamond bearing deposits followed. In 1908, diamonds were found in the German South West African (present-day Namibia) desert sands. The significance of this discovery is that after World War II, this field came under the control of a mining company called Anglo-American. This mining company was formed in 1917 by Ernest Oppenheimer, who was backed by a mining engineer named Herbert Hoover (future president of the United States), the U.S. Newmont Mining Corporation, and J. P. Morgan Bank. Eventually, Oppenheimer's control of Anglo American Corporation was used as leverage for his chairmanship of DeBeers.

A diamond bearing kimberlite pipe was discovered in Siberia in 1954. There are at least 300 kimberlite pipes in Siberia. The most productive of these is the Udachnaya mine near the town of Udachnaya.

In the early 1970s, the Ellendale pipes were discovered in the Kimberley region of Western Australia.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120025845

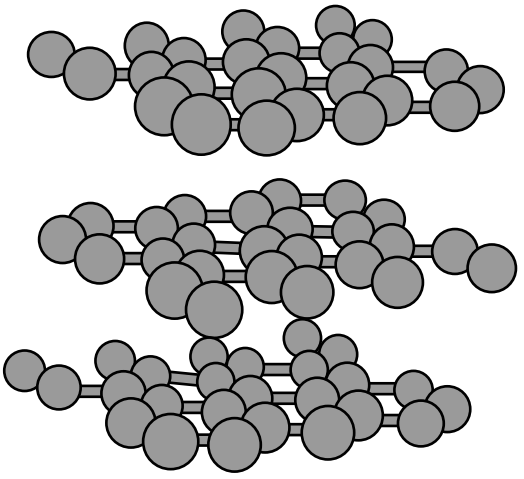


Fig. 1 Structure of graphite.

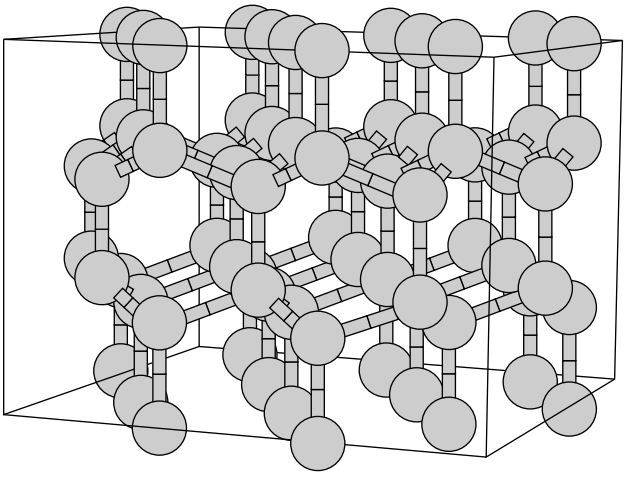


Fig. 3 Structure of lonsdaleite.

This discovery was significant because the pipes were not kimberlite, but a different type of volcanic rock called lamproite. One of the most productive diamond mines in the world, the Argyle mine with an ore that produces 400 carats of diamond per 100 tons, is a lamproite pipe. Albeit, the diamond is small and 90% is industrial grade.

In 1991, Canadian geologist Chuck Fipke's 1991 discovery of diamond bearing kimberlite in the Northwest Territory of Canada led to the construction of Canada's first diamond mine, the Ekati Mine. Canada is poised to become the leading producer of diamonds.

Synthetic Diamond, High Pressure and High Temperature

In 1797, Smithson Tenant discovered that diamond was made of pure carbon.[1] The quest to make synthetic diamond dates back to the early 1800s when C. Cagniard de la Tour claimed to have synthesized diamond from a solution. His crystals turned out to be aluminum and magnesium oxide. Many other claims and attempts were made but until the discovery of diamond in kimberlite, the role of high pressure and high temperature (HPHT) in the formation of diamond was not known. Diamond can be formed at great depths in the earth because of HPHT; it is the dominant phase of carbon; while at low pressures like those attained in a typical laboratory experiment, graphite is the dominant phase (Fig. 4).

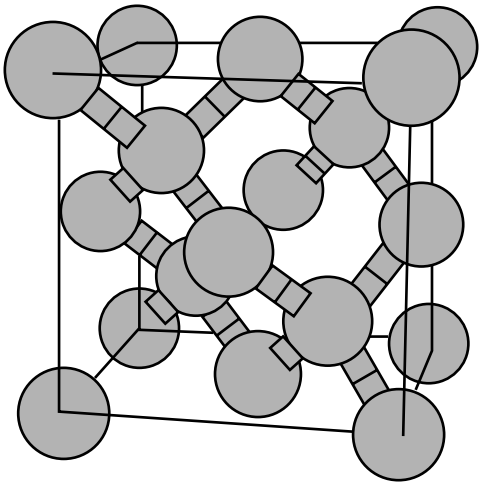


Fig. 2 Structure of diamond.

High-pressure research was revolutionized by Percy Bridgman in 1905 with his high-pressure press. Bridgman in that year solved the problem of the high-pressure seal; the same method used today. He used a very soft solid seal material (Sioux Indian pipe stone), but had the insight to keep the seal at a slightly higher pressure than the sample. He achieved this by using a two-piece piston and the fact that pressure is force acting on an area. Making the area of his seal less than that of his sample, and by exerting the same force on both the sample and the seal with his two-piece piston, he ensured that the seal was at a slightly higher pressure than the sample. Bridgman was also the first to use tungsten carbide, a material with much greater stiffness and hardness than steel, as a high-pressure anvil. Although Bridgman tried to make diamond, pressure alone was not sufficient.

The key to making diamond is the combination of pressure and temperature. The partnership of inventor Baltzar Van Platen and the company Allmanna Svenska Elektriska Aktiebolaget, or ASEA, was the first to grow diamond on February 16, 1953, but strangely failed to announce their discovery to the world. To achieve high pressures, ASEA used von Platen's split sphere device, which incorporated six wedge-shaped anvils, each directed at a cube-shaped sample chamber. The sample was carbon-rich at the center and

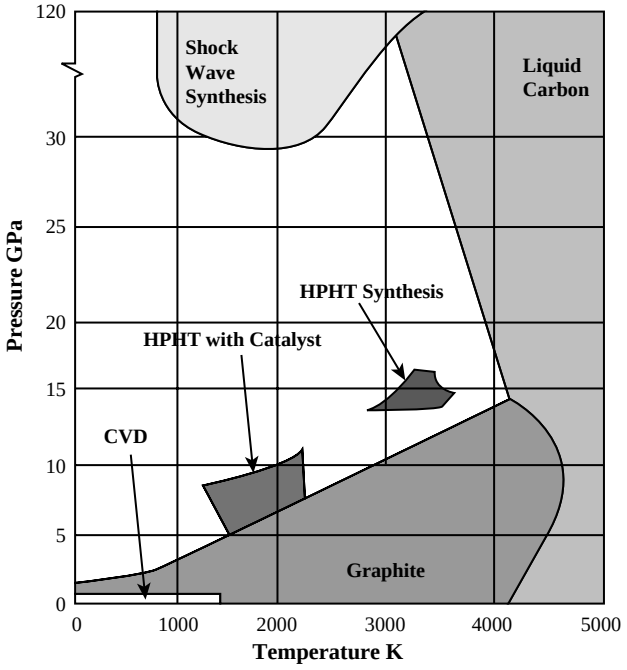


Fig. 4 Phase diagram of diamond.

surrounded by thermite (which ignited and burned to create a high temperature) and a copper jacket. The sample was placed inside the split sphere and this assembly was jacketed with copper. The split sphere was then placed in a high-pressure water tank. The high-pressure water tank was capable of achieving a pressure of 6000 atm. This 6000 atm when applied to the spit sphere would focus more than 80,000 atm of sustained pressure on the sample. In the first experiment that formed diamond, the sustained pressure was 83,000 atm and the carbon-rich material was a mixture of iron carbide and graphite. The iron was a critical component because it acted as a catalyst in the HPHT catalytic regime (see Fig. 4).

A team from GE grew diamond on December 8, 1954. The team had implemented many innovations in their high-pressure apparatus. The first challenge was to secure a suitable sealing material. High-quality Sioux Indian pipe stone was extremely difficult to get and Percy Bridgman kept his source secret. The GE team found an alternate sealing material, the mineral pyrophyllite or wonderstone, which was commercially available. In addition, high temperature was achieved by passing current through a graphite heating element. GE also had two high-pressure press designs, both of which eventually grew diamond, Herb Strong's cone apparatus and Tracy Hall's belt apparatus. The team had the high-pressure and high-temperature capability required to grow diamond but did not understand the need for a catalytic material (Fig. 4). Their breakthrough came when, in November 22, 1954, Herb Strong noted that diamond was discovered in some iron meteorites. He wrote in his notebook: ''The (iron–nickel) alloys should be investigated at high pressure for (melting point) and as a solvent for carbon in the liquid state.'' Indeed, iron, nickel, and cobalt are catalysts for diamond growth in the HPHT catalytic regime. On December 8, 1954, Strong wrapped a commercial carbonizing powder called Steco, with two seed diamonds, in iron foil. He used an estimated pressure of 50,000 atm with a temperature of 1250°C. His experiment was run for 16 hr. On December 15, x-ray analysis proved that two shard-like crystals, 1/16 of an inch in the long direction, were diamond. The results were not immediately repeatable and there is considerable evidence from a 1993 analysis of the larger of the two crystals that the diamonds were probably natural. On December 16, however, Hall used iron sulfide and two diamond seed crystals with his belt press and did produce synthetic diamond. This experiment was repeated at will.

In the last 40 yr, the development of synthetic diamond in various forms has fueled a revolution in the use of diamond as an engineering material. The process of HPHT diamond synthesis was responsible for stunning growth in the abrasives market. During that time, the world's consumption of diamond abrasive materials increased from 5 to over 100 tons/yr.

SYNTHETIC DIAMOND, CHEMICAL VAPOR DEPOSITION

Between November 26, 1952, and January 7, 1953, W. Eversole reported the synthesis of diamond by chemical vapor deposition (CVD). Going strictly by the phase diagram of carbon, it was very difficult for many scientists to accept the proposition that diamond could be obtained by CVD. In 1956, Soviet scientists B. V. Spitsyn and B. V. Deryagin filed a patent disclosure covering the growth of a diamond from carbon tetraiodide, which set the stage for substantial CVD work in the USSR. In the 1960s J. Angus in the United States duplicated the earlier work of Eversole. Meanwhile, progress was also made in Deryagin's laboratory in the USSR. The major breakthrough in the CVD process was the use of atomic hydrogen during the growth phase. Understanding the role of atomic hydrogen began in the late 1960s. Boris Spitsyn noted that when hydrogen gas was added to his thermal decomposition experiment, diamond crystals grew at rates thousands of times faster than in previous experiments. The mechanism of this process was not understood fully until several years later when it was explained independently by Angus' group and by Deryagin's group. In 1971, Derjaguin and Dimitri Fedoseev published a summary of their group's work, which demonstrated

a high-growth-rate CVD process.[2] The Western world became familiar with this work when Derjaguin and Fedoseev published an article in Scientific American.[3] This announcement was greeted with skepticism by scientists worldwide, primarily because Derjaguin's earlier claim of discovering a new form of water called polywater was discredited. In the late 1970s and early 1980s, however, N. Setaka's group in Japan also reported high-growth-rate CVD diamond films. These reports motivated the U.S. Department of Defense to increase research funding for CVD diamond in the mid-1980s.

Chemical vapor deposition diamond films have been expanded into the markets of electronics, cutting tools, and wear-resistant coatings and have demonstrated other applications in the areas of thermal management, optics, and acoustics. The CVD diamond market is growing broader as new products continue to be developed, such as chemical process electrodes and radiation detectors. Additional applications such as cold cathode emitters, high-temperature sensors, and, ultimately, semiconductors are also being developed. The impact of diamond on people's lives, while not always immediately obvious, is definitely becoming increasingly significant.

SYNTHETIC DIAMOND, EXPLOSION

One of the by-products of a large meteor collision with the earth is the formation of fine diamond crystals due to the explosion. The presence of fine diamond crystals in a layer 65 million years old helped to formulate the theory that a large meteor impact was responsible for the extinction of the dinosaurs. One of the distinctive aspects of diamond created by explosion is that cubic as well as hexagonal (lonsdaleite) crystals are formed. An explosion is capable of creating millions of atmospheres of pressure in microseconds. In 1959, Paul DeCarlie and John Jamieson used a simple apparatus where explosives were packed behind an aluminum metal plate and detonated. The plate accelerated into a graphite sample generating a pressure approaching 300,000 atm. The experiment produced fine diamond crystals. In the mid-1960s Dupont used their expertise in explosive technology to develop a commercial diamond producing explosive process. The finely grained diamond product, trade name Mypolex, has sold millions of carats per year because it is suitable for fine polishing applications.

Owing to the short time the pressure is created in an explosion, the diamond growth time is short, thus leading to small crystal size. In the laboratories of the former Soviet Union, an explosive process was developed to create ultra-dispersed detonation nanophase materials (UDDNM), which are composed of nanograin-sized crystals of various materials. This process has been used to synthesize nanophase diamond, zirconium dioxide, and aluminum oxide. This process, having no natural analogs, forms the material by direct synthesis in a detonation wave. Synthesis occurs under HPHT conditions. There are several former defense centers in the former Soviet Union that are capable of producing UDDNM.

The former Soviet Union used materials produced by this method as abrasives for polishing, production of composites, and lubrication. There are other important applications that need to be explored as well, which take advantage of the optical, electronic, and magnetic properties of these unique materials. The ultradispersed detonation process uses an explosive substance to create a synthesis wave. Synthesis occurs at high pressures (20–30 GPa) and temperatures (3000–5000 K) and over short time periods (0.1–0.5 μsec).[4,5] In the fabrication of diamond, the carbon from the explosive substance itself is used in the synthesis process. The resulting particles have specific surface areas of 270–390 m^2/g with shapes close to spherical. The size distribution is between 4 and 6 nm (see Table 1). In addition, the particles have high porosity, a high degree of chemical inertness, and high sorptive capacity.[6]

METHODS OF SYNTHESIS

High Pressure and High Temperature

High pressure and high temperature is a method used to grow diamond under an HPHT environment. The process is performed in the diamond stable region at pressures ranging from 7 to 10 GPa and temperatures ranging from 1700 to 2000 K. Graphite or other carbons are dissolved in molten transition metals and then precipitated in diamond form.

Chemical Vapor Deposition

Chemical vapor deposition is a growing technique in which a solid material is deposited from its gas phase

Table 1 Properties of nanophase diamond

Property	Value
Percentage of diamond phase (%)	85–91
Original particles size (nm)	4–6
Aggregates size (nm)	50–1000
Density (g/cm^2)	3.10
Specific surface area (m^2/g)	300 ± 60
Carbon percentage (%)	85–91
Chemical impurities	O, N, and H

on to a substrate. Diamond crystals grown by this technique require a carbon-containing gas, usually CH_4, and H_2. Hydrogen plays a double role in the active chemistry. In its atomic form it provides the path to carbon matrix formation when it reacts with a previously formed hydrocarbon surface by extracting a hydrogen atom and leaving a carbon surface.

$$H(g) + CH(s) \longleftrightarrow H_2(g) + C^*(s)$$

$$H(g) + C^* \longleftrightarrow CH(s)$$

This mechanism occurs in a continuous process allowing the growth of diamond crystals at low pressure (usually below atmospheric pressure) and high temperature (approximately 2000°C). Hydrogen also prevents the formation of aromatic species that lead to graphite formation through a selective etching process.

Hot Filament

Diamond deposition by the hot filament method consists of a carbon containing gas and hydrogen, which undergo dissociation by passing through a hot filament usually made of tungsten wire. The dissociated molecules then deposit on a substrate (at approximately 900°C) where a carbon matrix grows in the form of diamond. Deposition dynamics is described by the CVD process.

Thermal Decomposition

During diamond growth, carbon atoms are extracted from organic species, usually methane (CH_4), methyl (CH_3), and acetylene (C_2H_2). Atomic hydrogen is necessary to prevent graphitic formations at the diamond surface. To break apart molecular hydrogen, certain conditions of temperature and pressure must be met. At equilibrium, the atomic hydrogen concentration is given by Eq. (1):

$$[H]eq. = \frac{\{[(P^0)^2 + 4K^{-1}P^0P_T]^{1/2} - P^0\}}{2K^{-1}RT} \quad (1)$$

where P^0 is the standard pressure, P_T is the total pressure, R is the gas constant, T is the temperature, and K is the equilibrium constant for hydrogen dissociation.[7]

Because the dissociation reaction is endothermic, diamond synthesis processes like combustion synthesis, plasma torch, and hot filament provide enough energy to maintain a high rate of hydrogen dissociation. Flame nozzles achieve higher deposition rates compared to other CVD processes. The use of multiple flame burners increases the deposition area with high growth rates.[8]

Plasma Phase

Diamond CVD using a DC-arc plasma jet is one of the methods that are used for high growth rate diamond deposition. A gas mixture of H_2/Ar and CH_4 is necessary to enable the deposition of carbon for growing crystals on the substrate. A high-energy thermal plasma produces the high dissociation rates of chemical species that allows a high deposition rate. Thermal plasmas have high energy densities (up to $10^8\,J/m^3$), and temperatures that may exceed 10,000 K. Many parameters have been studied including gas mixtures, temperatures, and environmental conditions in the chamber to maximize the efficiency of the process. Studies of the deposited material demonstrate that a high-quality film can be grown by this technique. The deposition areas still remain relatively small and work continues to enhance the deposition areas. Large deposition areas become especially important for coating processes where uniformity is required. Besides high gas and power consumption, this method is economically reliable for the industrial scale. Linear growth rates of 1 mm/hr have been reported and this rate is three orders of magnitude larger than those achieved with hot filament CVD (HFCVD) and microwave CVD (MWCVD).

Microwave CVD uses microwave energies ranging from several hundreds of watts to tens of kilowatts at a frequency of 915 MHz or 2.45 GHz to dissociate active species. The most common MWCVD systems use 2.45 GHz.

Thermal plasma chemical vapor deposition (TPCVD) has many advantages over other methods, as deposition rate is the determining factor in the cost of diamond-coated materials.

INDUSTRIES

Tool and Die

Tool performance depends greatly on its interface with other materials under circumstances that are less than ideal. The properties of diamond abrasives make it suitable for tool design. Specifically, when operational conditions of heat, corrosion, vibrations, torsion, stress, strain, and force are considered, diamond is the material of choice. Increasing demand in the use of lightweight and other advanced materials has had positive effects on using diamond deposition as an alternative to design tools that hold their properties during cutting, drilling, and polishing work. Although

diamond is considered an ideal material for coating because of its hardness, not every diamond morphology necessarily meets the required standards for a given application. Other aspects, such as adhesion, film thickness, uniformity, and edge sharpness, also play an important role for each application. Diamond coatings on cutting tools can be classified according to their method of synthesis: polycrystalline diamond (PCD), CVD diamond, and single-crystal diamond. Polycrystalline diamond refers to a diamond that is made by sintering diamond particles together in a metal matrix under HPHT conditions.[9] Its structure exhibits a high abrasion resistance and hardness. It is mainly used for cutting nonferrous metals, nonmetallic (e.g., plastics, wood, chipboard, and rubber) substances, and wear parts. Major applications are for automotive and aerospace components. The mechanical and thermal properties of diamond are shown in Table 2.

In the production of wire dies, single-crystal diamond type Ib is used because of its thermal stability at higher temperatures (over 1000°C) and a metal-inclusion-free matrix. The wear region is allocated in the crystallographic {111} plane to get the highest abrasion and wear resistance. Although diamond exhibits more wear resistance characteristics than tungsten carbide, crystal morphologies may cause catastrophic failures during the drawing process owing to high pressures.

Grinding and Polishing

The grinding process is characterized by material removal. It is achieved by applying a force and movement during the process. To increase the removal rate, one of the two parameters should be increased. A diamond tool design is dependent on both diamond toughness and hardness properties and plays a role in the G ratio, which is defined as the volume of material removed over volume of tool worn. Toughness describes the way diamond fractures during use. It is also known as friability. Hardness describes diamond's resistance to abrasion and is also known as scratch and indentation. Its value ranges from 57 to 104 GPa and depends on crystal orientation, e.g., (100) and (111) surfaces, respectively, for single crystals. Polycrystalline diamonds have no orientation dependence for hardness parameters because of the statistical nature of surface orientation. A bulk value for hardness is used instead.

A crystal that exhibits large smooth faces tends to polish because it has a higher rub than cutting action. This creates excessive heat and low material removal. This kind of morphology works better in metals, glass, concrete, and stone. Irregular texture in small crystals gives a friable characteristic to diamond crystals with a higher tendency to fracture. Crystal fracture is a desirable effect in tool performance because it establishes the effective concentration of working crystals on the tool surface and generates new and sharp cutting surfaces. Resin-bonded tools have a longer life because of high crystal retention. Mechanical bonding is strengthened by rough crystal surfaces. Mesh size in the resin bond is an important parameter as smaller sizes can be pulled out with less force. Crystal retention can increase even for small mesh sizes by using spiked metal coatings that increase the surface contact area. Nickel, copper, and titanium are used for coating diamonds. The first two provide an enhanced mechanical bonding in wet and dry grinding processes, respectively. The third provides both mechanical and chemical enhanced bonding. The metal cladding can be 60% of particle weight

Table 2 Mechanical and thermal properties of synthesized diamond

Property	CVD diamond	Single-crystal diamond	PCD
Young's modulus (GPa)	1000–1100	1000–1100	776
Hardness (Gpa)	85–100	50–100	50
Tensile strength (MPa)	450–1100	1050–3000	1260
Compressive strength (GPa)	9.0	9.0	7.60
Fracture toughness [MPa(m)$^{1/2}$]	5.5–8.5	3.4	8.81
Transverse rupture strength (GPa)	1.3	2.9	1.2
Poisson's ratio	0.07	0.07	0.07
Thermal conductivity at 20°C (W/m/K)	500–2200	600–2200	560
Thermal conductivity at 200°C (W/m/K)	500–1100	600–1100	200
Thermal diffusivity (cm^2/sec)	2.8–11.6	5.5–11.6	2.7
Thermal expansion coefficient at 300 K	1.21	1.21	4.2
Thermal expansion coefficient at 500 K	3.84	3.84	
Thermal expansion coefficient at 1000 K	4.45	4.45	6.3

and also can act as a heat sink, improving tool performance in resin-bonded tools like grinding wheels.[10]

High removal rates have been achieved for abrasive tools where cutting surfaces are coated with rough diamond. Roughness depths of around 30 μm in microtool grinding surfaces have performed well in machining brittle materials achieving not only a high removal rate but also high quality.[11] This feature is very important for machining low-dimension structures like printed circuit boards, micromechanical systems (MEMS), and dental tools. In dental burs, alternative techniques like HFCVD for growing crystals and WC–Co substrate pretreatment have shown that adhesion toughness can be enhanced thereby improving tool performance and life.[12,13] In CVD, diamond abrasive roughness value, *Rz*, increases with crystal width ranging from 2 to 6 μm for crystal widths of 17 and 50 μm, respectively.[14] In optical applications, the grinding energy, that is, energy expended for material removal during glass grinding, is independent of the tool speed and in-feed rate; however, for slow tool speed and high in-feed rate an increased bond wear and enhanced self-sharpening effect have been identified.[15] In woodworking industries increased feeding rates up to 150 m/min are needed for milling tools with higher cutting edges. Increasing the number of PCD edges on the tool increases the efficiency removal capacity for this kind of process.[16]

Material removal by electrical discharge diamond grinding is a grinding method in which spark discharges between the metal bond and workpiece thermally soften the material while simultaneously having the diamond grains perform the abrasive action. This method is attractive for grinding conductive materials with a hard profile. Although normal force increases as wheel speed increases, this method reduces the normal force and grinding power when compared with other methods.[17]

Polishing processes are characterized by material removal by abrasive action in which grit size determines material finishing. Diamond grits for this process may be bonded in resin, vitrified, or metal matrix, although most polishing and fine finishing work is performed using an abrasive of loose powder. Major applications are for lens polishing, polishing of diamond tools, glass, ceramics, and semiconductors. Chemical vapor deposition diamond is used as an abrasive because it is found to be efficient in polishing other CVD diamond surfaces by achieving a higher polishing rate and reducing roughness values from 3.32 to 0.55 μm.[18] This has a major impact in applications that require high-quality crystals for electronic, electrical, thermal, and optical applications. Polishing action is direction oriented, having soft and hard directions during the process. Selection of hard direction ⟨100⟩ could end in tool damage. At the microscopic level, polishing action strongly depends on the deformation energy and mechanical characteristics in the diamond matrix that contribute to atom displacements.[19]

Drilling

A major limitation in the use of PCD is that it is not suitable for ferrous materials. Diamond reacts with iron in the presence of oxygen at high temperatures. This limitation forces the use of other hard materials, such as cubic boron nitride.

Substrates help to prolong tool life but can introduce problems of differential thermal expansion because of the inherent characteristics of both materials. Because drilling activity generates high temperatures, it is necessary to use a secondary phase with high thermal stability. Microcrystalline diamond in ceramic material accomplishes this task. Matching coefficient of thermal expansion and diamond cube orientation allows drill bits to remain sharp for drilling medium and hard rock at very high penetration rates.

Increasing the use of microdrills in electronics makes it necessary to enhance the cutting efficiency of tools that are designed to operate at very high cutting speeds. An enhanced technique for HFCVD is used to ensure better thermal distribution and uniform coating of WC–Co microdrills with PCD.[20] Tool design for drilling surfaces depends strongly on crystal formation, nucleation, growth, and adhesion. Any imperfection in these parameters, such as protusions or supercrystal formations, can precipitate a catastrophic tool failure.[21]

Mesh size in drilling tools has a dramatic impact in drilling speeds, which can be increased from 1000 to 6000 rpm along with a force reduction on the diamond core bit by a factor of nearly 6.8.[22]

Mining

Mining work requires specialized tools to perform work under the harshest and most aggressive conditions. This is when diamond abrasive properties combined with the hardness properties of other materials become crucial to drilling or cutting action that is subjected to extreme loading during prolonged periods.

ELECTRONICS

Heat Sinks

The state-of-the-art technology in electronics is characterized by continuous reduction of circuit etch. This represents an enormous electronic activity packed in a reduced volume, giving rise to higher temperatures.

This fact could not be overlooked making it necessary to implement effective heat sinks to ensure the desired device performance.

Diamond has the highest thermal conductivity when compared to other substances. For type Ia natural diamond the thermal conductivity is about 2000–4000 W/m/K. For type IIa it is up to 17500 W/m/K. A major problem in heat transfer is an effective heat sink that depends on the effective contact area, forces between both the materials, and the gap interface. Diamond exhibits the best characteristics for a heat spreader, which is the interface material that transfers heat between heat source and heat sink.

At a microscopic level the contact surface is restricted by peaks and valleys and even highly polished surfaces may exhibit a high peak to valley ratio. This makes it necessary to use an extremely flat contact surface between the heat source and thermal spreader to guarantee an efficient transfer of heat. Chemical vapor deposition diamond with a thickness of 1000 μm has been used for Multi Chip Modules (MCM) for this purpose. Heat spreaders are used in the electronic industry for IC packaging and solid-state lasers.

Insulators

The electrical properties of undoped diamond crystals make it suitable to be used as an insulator. It can support electric field strengths as high as 10 MV/cm before conduction and has a resistivity value as high as $10^{14}\,\Omega$. Its higher band-gap of 5.46 eV makes it difficult to promote an electron to the conduction band. This makes diamond a first-class insulator. Diamond insulators are being used in electronics as an intrinsic interface with doped diamond.

Semiconductors

Diamond's natural properties make it an ideal material for semiconductor applications. Although it has a wide band-gap of 5.5 eV, its intrinsic properties of carrier velocity of 2.7×10^7 cm/sec and mobility of 2200 $cm^2/V/sec$ for electrons, and dielectric constant of 5.5 give this material a broad advantage over other wide-band-gap semiconductors in high-power and high-frequency applications. Its low intrinsic carrier concentration of 10^5 at 1000 K makes it a good candidate for high-temperature electronics. The properties of diamond are attractive for the development of microsystem devices as an interface mainly due to diamond's high heat conductivity.[23]

When diamond is doped with either boron or nitrogen it becomes a semiconductor but it is not activated at room temperature because of the dopants' activation energies of 0.37 and 1.7 eV, respectively. Activation and conductivity at room temperature are problems that can be addressed by the incorporation of other electronic structures that increase carrier transport. Crystal morphology is an important parameter in the boron doping process to determine uncompensated acceptors (N_A–N_D) for different crystal facets as a function of doping concentration. The temperature coefficient of resistance for a CVD diamond film can be changed by boron doping. As conductivity depends on the crystal phase, the combined electromechanical properties can be exploited in sensor applications both for resistive temperature detectors and for pressure transducers.[24] As electrical conductivity is related linearly with boron concentration, a better-controlled process may allow for the development of better semiconductor devices improving crystal quality and operating limits.[25]

Diamond-like carbon (DLC) devices are also being developed in the electronic field. Although potential applications range from heat sinks to solar cells, its low-mobility characteristics, ranging from 10^{-4} to 10^{-5} $cm^2/V/sec$, and crystal structure defects still have to be improved for practical applications.[26] Most recent applications are for masking integrated circuits with satisfactory results for p–n junction yield rate through DLC mask, being better than polymer mask materials.[27]

A larger band-gap enables the fabrication of diodes with higher breakdown voltages. Diamond has a higher band-gap than SiC, GaN, and Si. However, its breakdown voltage is dependent on impurities in the crystal lattice. To achieve the highest breakdown voltage, high-purity diamond is required. Hydrogen and boron deteriorate electrical characteristics. Hydrogen diffusion depends on boron concentration. It has been observed that hydrogen diffuses as a positive ion for a homoepitaxial diamond layer with a high boron concentration and as a negatively charged species for a CVD layer with a low boron concentration.[28] Diamond Schottky diodes have large forward resistances of 300 Ω/cm^2 that are 10 times the value for Si diodes owing to compensated impurities, device geometry, and lack of a good ohmic contact. Under uncompensated conditions the forward resistance value for a Schottky diode could range between 0.01 and 0.1 Ω/cm^2 with a breakdown voltage of 10 kV.[29] A diamond Schottky diode should be a superior device over a Si diode for these parameters.

Chemical vapor deposition films have been grown on an $Ir/SrTiO_3$ substrate of around 0.6 cm^2 for a field effect transistor application achieving for the first time an RF output power for a device operating at frequencies of the order 10^9 Hz. Larger-size substrates will eventually allow the development of power diamond transistors at wafer scale.[30]

With the fast pace development of MEMS, diamond is gaining more applicability because of its flexible

properties of being an insulator or a conductive material when necessary. A major area is microswitching. In air, the frequency of operation of a diamond microswitch is about eight times that of its Si counterpart.[31]

Tribology

Diamond exhibits low-friction properties. This makes the diamond suitable for tribological applications where high friction and wear is present. Tribological properties depend on many factors including crystal orientation and particle size. High material removal rate is achieved during polishing with diamond particles because of their high deformation resistance.[32] Diamond-like carbon films are used as solid lubricants. Laser treatment of the surfaces in tribological DLC films is performed to reduce wear by lithography processes and prolong surface lifetime.[33] Multilayer DLC coatings have a better performance than a single-layer DLC coating; however, the latter exhibits less stress characteristics compared to intrinsic stress and applied stress with higher loads in multilayer DLC. Single-layer friction partners exhibit low friction and wear.[34] Other techniques like sputter ion plating have been used to deposit hydrogenated DLC, a-C : H coatings. These coatings exhibit low friction and wear behavior under high contact pressures ranging from 1 to 3.5 GPa.[35]

CONCLUSIONS

Diamond's properties make it the desirable material in thermal, optical, electrical, electronics, and mechanical applications. It has not been exploited to its maximum potential. Future trends are toward applications in medicine, biology, and the nuclear field. These fields already have applications that use diamond but continuous improvements are being made through research. Diamond is a strong candidate as a substitute for materials currently being used in a variety of applications. Although implementation may be deterred by cost factors or technical issues, the development of new deposition techniques may overcome this limitation. Deposition techniques and a higher control of processes surely will help to launch more sophisticated electronic applications that eventually will realize diamond's superior performance over other materials.

REFERENCES

1. Hazen, R.M. *The Diamond Makers*; Cambridge University Press, 1999.
2. Derjaguin, B.; Fedoseev, D.; Bakul, V.; Ryabov, V.; Spitsyn, B.; Nikitin, Y.; Bochko, A.; Varnin, V.; Laurent'ev, A.; Primatchuk, V. *Physicochemical Synthesis of Diamond from Gas*; Kiev: Technika, 1971.
3. Derjaguin, B.V.; Fedoseev, D.B. The synthesis of diamond at low pressures. Sci. Am **1975**, *233* (5), 102–109.
4. Komarov, W.F.; Petrov, E.A.; Sakovitch, G.V.; Klimov, A.V. Ultra dispersed detonation diamond properties and applications. In *Advances in New Diamond Science and Technology*; Sato, K., Fujimori, N., Fukunaga, O., Kamo, M., Okushi, H., Yoshikawa, M., Eds.; MYU: Tokyo, 1994.
5. Amisichkin, V.; Dolgushin, D. Mechanisms of diamond synthesis from high explosives. In *Advances in New Diamond Science and Technology* ; Sato, K., Fujimori, N., Fukunaga, O., Kamo, M., Okushi, H., Yoshikawa, M., Eds.; MYU: Tokyo, 1994.
6. Vereschagln, A.L.; Sakovich, G.V.; Komarov, V.F.; Petrov, E.A. Properties of ultrafine diamond clusters from detonation synthesis. Diamond Relat. Mater **1993**, *3*, 160–162.
7. Argoitia, A.; Kovach, C.S.; Angus, J.C. Hot filaments CVD methods. In *Handbook of Industrial Diamonds and Diamond Films*; Prelas, M.A., Popovici, G., Bigelow, L.K., Eds.; Marcel Dekker, Inc.: New York, 1998; 797.
8. Ravi, K.V.; Koch, C.A.; Olson, D. Large area diamond synthesis by the combustion flame process. 2nd International Conference on the Applications of Diamond Films and Related Materials, Japan, 1993; Yoshikawa, M., Murakawa, M., Tzeng, Y., Yarbrough, A., Eds.; 491.
9. http://www.e6.com. What is PCD? (accessed Nov 2003).
10. Bryant, J. Which diamond grinding wheel? Manuf. Eng. **1998**, *120* (3), 74–79.
11. Jan, G.; Lothar, S.; Heinz, W. Chemical vapour deposition diamond coated microtools for grinding, milling, and drilling. Diamond Relat. Mater. **2000**, *9*, 921–924.
12. Sein, H.; Ahmed, W.; Rego, C. Application of diamond coatings onto small dental tools. Diamond Relat. Mater. **2002**, *11* (3–6), 731–735.
13. Ahmed, W.; Sein, H.; Ali, N.; Gracio, J.; Woodwards, R. Diamond films grown on cemented WC-Co dental burs using an improved CVD method. Diamond Relat. Mater. **2003**, *12*, 1300–1306.
14. Jan, G.; Lothar, S.; Bernd, M.; Hans-Werner, H. Micro abrasive pencils with CVD diamond coating. Diamond Relat. Mater. **2003**, *12*, 707–710.
15. Toshio, T.; Paul, D.F. Micromechanics of diamond composite tools during grinding of glass. Mater. Sci. Eng. **2000**, *A285*, 69–79.

16. Fath, J.; Wagner, K.W. All-round machine for EDM and grinding. Ind. Diamond Rev. **2003**, *3*, 42–43.
17. Koshy, P.; Jain, V.K.; Lal, G.K. Mechanism of material removal in electrical discharge diamond grinding. Int. J. Machine Tools Manuf. **1996**, *36* (10), 1173–1185.
18. Tang, C.J.; Neves, A.J.; Fernandes, A.J.S.; Grácio, J.; Ali, N. A new elegant technique for polishing CVD diamond films. Diamond Relat. Mater. **2003**, *12*, 1411–1416.
19. Van Bouwelen, F.M. Diamond polishing from different angles. Diamond Relat. Mater. **2000**, *9*, 925–928.
20. Sein, H.; Ahmed, W.; Hassan, I.U.; Ali, N.; Gracio, J.J.; Jackson, M.J. Chemical vapour deposition of microdrill cutting edges for micro- and nanotechnology applications. J. Mater. Sci. **2002**, *37*, 5057–5063.
21. Chaterjee, S.; Edwards, A.G.; Feigerle, C.S. Machining and morphological evaluation of diamond coated tungsten carbide drills. J. Mater. Sci. **2001**, *35*, 5707–5717.
22. Jennings, M. High speed drilling for high speed trains. Ind. Diamond Rev. **2003**, *3*, 31–32.
23. Kohn, E.; Adamschik, M.; Schmid, P.; Denisenko, A.; Aleksov, A.; Ebert, W. Prospects of diamond devices. J. Phys. D Appl. Phys. 2001, *34*, R77–R85.
24. Chalker, P.R.; Johnston, C.; Werner, M. Physical properties of diamond for thermistors and pressure transducers. Semicond. Sci. Technol. **2003**, *18*, S113–S116.
25. Thonke, K. The boron acceptor in diamond. Semicond. Sci. Technol. **2003**, *18*, S20–S26.
26. Milne, W.I. Electronic devices from diamond-like carbon. Semicond. Sci. Technol. **2003**, *18*, S81–S85.
27. Hirakuri, K.K.; Yoshimura, M.; Friedbacher, G. Application of DLC films as masks for integrated circuit fabrication. Diamond Relat. Mater. **2003**, *12*, 1013–1017.
28. Saguy, C.; Cytermann, C.; Fizgeer, B.; Richter, V.; Avigal, Y.; Moriya, N.; Kalish, R.; Mathieu, B.; Deneuville, A. Diffusion of hydrogen in undoped, p-type and n-type doped diamonds. Diamond Relat. Mater. **2003**, *12*, 623–631.
29. Butler, J.E.; Geis, M.W.; Krohn, K.E.; Lawless, J., Jr.; Deneault, S.; Lyszczarz, T.M.; Flechtner, D.; Wright, R. Exceptionally high voltage Schottky diamond diodes and low boron doping. Semicond. Sci. Technol. **2003**, *18*, S67–S71.
30. Kubovic, M.; Aleksov, A.; Schreck, M.; Bauer, Th.; Stritzker, B.; Kohn, E. Field effect transistor fabricated on hydrogen-terminated diamond grown on $SrTiO_3$ substrate and iridium buffer layer. Diamond Relat. Mater. **2003**, *12*, 403–407.
31. Schmid, P.; Adamschik, M.; Kohn, E. Design of high-speed diamond microswitch. Semicond. Sci. Technol. **2003**, *18*, S72–S76.
32. Xie, Y.; Bhushan, B. Effect of particle size, polishing pad, and contact pressure in free abrasive polishing. Wear **1996**, *200*, 281–295.
33. Dumitru, G.; Romano, V.; Weber, H.P.; Pimenov, S.; Kononenko, T.; Hermann, J.; Bruneau, S.; Gerbig, Y.; Shupegin, M. Laser treatment of tribological DLC films. Diamond Relat. Mater. **2003**, *12*, 1034–1040.
34. Sheeja, D.; Tay, B.K.; Krishnan, S.M.; Nung, L.N. Tribological characterization of diamond-like carbon (DLC) coatings sliding against DLC coatings. Diamond Relat. Mater. **2003**, *12*, 1389–1395.
35. Jarrat, M.; Stallard, J.; Renevier, N.M.; Teer, D.G. An improved diamond-like coating with exceptional wear properties. Diamond Relat Mater. **2003**, *12*, 1003–1007.

Diamond-Like Carbon Films

Angel Velez
Mark A. Prelas
Nuclear Science and Engineering Institute, University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

Diamond-like carbon (DLC) films have been known since 1962. The applications of DLC gained serious attention in 1971.[1] A wide range of applications have been found for DLC in the industrial and biological fields with promising growth also in the electronics field. Because DLC is composed of sp^3, sp^2, and possibly sp^1 coordinated atoms, the physical properties vary with sp^3 to sp^2 ratios and deposition conditions. This allows a heterogenous mixture of crystalline and amorphous phases in different film structures as required by specific applications. The sp^3/sp^2 ratio also determines the density of the film and is affected by the presence of hydrogen. Graphite-like films are characterized by a high content of sp^2 bonding while diamond-like films have a high percentage of sp^3 bonding. When the structure content has more than 80% of sp^3 it is classified as tetrahedral amorphous carbon (ta-C) and this type of film is essentially hydrogen free.

APPLICATIONS

Diamond-like carbon has found a number of applications in the areas of optics, semiconductors, electronics, medicine, and nuclear instrumentation. Diamond-like carbon is predominantly made using chemical vapor deposition. As the technology matures new applications will evolve.

Optics

The optical properties of DLC make it suitable for scratch-resistant antireflection coatings. Although it absorbs in the visible region, it has a good transmission in the IR region. It is used in germanium optics and provides protection against harsh environments (e.g., rain, sea) and abrasion. The use of these films sometimes is limited by the allowed thickness because of the mechanical stress and optical absorption limitations. This makes necessary the use of other primary coatings before DLC is deposited.[2] The use of DLC on aluminum mirror surfaces improves the optical performance of the mirror because there is no destructive interference in the coating boundaries, which affects mirror reflectivity. The exhibited high reflectivity, however, limits the use of DLC coatings in photothermal conversion of solar energy where high absorption in the IR region is required. It has been observed that the absorption properties of DLC films within the visible region and near-IR region are strongly affected by the presence of oxygen and silicon impurities.[3]

Semiconductors

The semiconducting properties of DLC are being studied for photovoltaic applications. A boron doped hydrogenated amorphous carbon (a-C : H) on an n-type silicon substrate heterojunction exhibits rectifying behavior when exposed to light. Although a photocurrent in the a-C : H is generated for wavelengths ranging from 300 to 700 nm, the conversion efficiency is still low because of the recombination process at the interface with the Si substrate.[4]

Electronics

Major applications of DLC films in electronics mainly depend on their mechanical and chemical properties. For data storage, both surface roughness and chemical composition are important parameters when hard disks are coated. To increase the data storage it is necessary to reduce the head-disk interfacial spacing, that is, the height between the flying recording head and the disk. Diamond-like carbon coatings of 10 nm have been deposited on hard disk surfaces to achieve high-density recording. Diamond-like carbon films also provide a smoother surface with a roughness value of 2.3 Å and corrosive protection to these devices.[5] Specifically, DLC deposition by ultrathin ion beam is performed to overcoat the magnetic recording medium. Films of 20 Å have been deposited showing good wear and corrosion protection on this medium.[6]

The use of DLC in field emission displays is growing because of its properties as an electron emitter for field emission array applications. By controlling the $sp^3/(sp^3 + sp^2)$ ratio the activation energy can be

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120037648

reduced to increase the release of electrons. The use of DLC can increase emitter current or decrease the necessary voltage. The voltage reduction is achieved through the adjustment of asperity height. Varying the current values during cathodic arc deposition can control these characteristics. It has been found that a maximum ratio of 50% is achieved at 80 A and the maximum asperity roughness at 100 A.[7] The asperity height is the main factor to be controlled because electron emission is highly dependent on it. Another approach in the development of field emission surfaces consists of irradiating DLC films of 70–80% sp^3 content with U or Au ions of approximately 1 GeV or 340 MeV, respectively.[8] With this technique sp^2 cylindrical ion tracks with the conductive properties of graphite-like carbon allow electron transport from the substrate to the surface through the DLC material. Continuous conductance along the track is only achieved by irradiating with heavy ions. Low hydrogen concentration is necessary to prevent the inhibition of sp^3 to sp^2 transformation.

Modifications of surface structures are still under analysis to reduce the onset voltage to operating conditions without damage. Novel structures based on the mechanical characteristics of ta-C combined with SiN are being developed to produce microcantilever structures that enhance the anode field emission.[9]

Medicine

Applications in coating biomedical devices are growing quickly because the DLC material is chemically inert and offers corrosion and wear resistance. Diamond-like carbon or amorphous carbon has the necessary smoothness required by many applications in biomedical devices, being preferable to polycrystalline structures. The development of medical implants requires a biocompatible material that could be exposed to the body fluids, some being highly corrosive, with minimum deterioration while at the same time preventing adverse reactions. Those implants used for joints also need to tolerate sliding motion that causes wear of contacting surfaces. In applications for body implants, wear is a critical parameter because it limits the effective lifetime of the implant. Simulations of the effects of body fluids on the implant have been performed. A lower coefficient of friction is obtained for DLC coatings with a considerably reduced wear rate. This has been observed on titanium surfaces used in orthopedic applications.[10] It has been demonstrated that hydrogen-free coatings like ta-C and a-C perform well during sliding wear tests with coefficients of friction ranging from 0.04 to 0.05. Hydrogenated structures exhibit higher wear rates under water-lubricated conditions and are not suitable for these applications.[11] Diamond-like carbon material should still perform well even if it is delaminated. It has been observed that bone marrow cells do not react to the presence of DLC particulates, and hence, neither inflammatory nor toxic reaction is expected in the body from the use of DLC.[12] The absence of a toxic response, besides the wear and tribological properties, makes DLC coatings attractive for long-term use where biocompatibility is necessary because of the biological system's natural response.

In vivo behavior the tissue–implant interaction and bioreactions related with cell attachment, proliferation, and differentiation are influenced by the surface texture of a film. Surface chemistry also plays an important role in bioreactions. When DLC is alloyed with other materials, the bioproperties of both materials interact to control biological responses like inflammatory reactions, protein adsorption, and cell differentiation. Metal alloy coatings prevent metal ion release, which causes adverse biological reactions like bone resorption, and platelet activation, which triggers thrombosis. This is a critical parameter in intracoronary heart stents. Platelet deposition is highly reduced with DLC coatings. This characteristic makes DLC films hemocompatible. This meets the requirement for implants that are continuously exposed to blood where a high ratio of albumin/fibrinogen is required to lower platelet adhesion to reduce thrombus formation.[13] Diamond-like carbon surfaces also offer other advantages like antibacterial activity, good growth of neurons, and white blood cell interaction.[14]

Diamond-like carbon can be alloyed with toxic materials. This is necessary when no cell adhesion should be allowed specifically when equipments or implants are of temporary use. When this alloy makes contact with the biological environment, the cytotoxic materials, like copper, vanadium, and silver, inhibit cell growth on the material surface because of the toxic action of the alloy. Changing alloy components can control the bioreactions. When DLC is mixed with silicon oxide a reduction of the inflammatory reactions is observed.

Nuclear Instrumentation

In the nuclear field, DLC coatings are used to improve instrumentation performance. In microstrip gas counters the signal generation is due to the avalanche of positive ions migrating away from the anode surface and it is proportional to the energy of the incoming radiation. The cathode and anode strips require a stable substrate with low resistivity, which is necessary to prevent positive ion collections that cause voltage instabilities. To reduce surface charging, DLC films are used for coating ionic conducting glasses to

increase microstrip gas counterstability. This allows an improvement in the rear signal measurement from the anode up to 60% of its value and the construction of two-dimensional detectors with thinner glasses.[15] Diamond-like carbon's surface properties make it suitable for use as a material for storing ultracold neutrons. Suppression factors for neutron transmission of 2.7 and a suppression factor for neutron storage of 4 have been independently observed. The suppression factor for transmission is measured by examining the neutrons being transmitted from a source when surrounded by a quartz wall vs. the neutrons being transmitted from a source when surrounded by a quartz wall coated with deuterated diamond-like films. The suppression factor for storage is measured by examining the neutrons stored from a source when surrounded by a quartz wall vs. the neutrons stored from a source when surrounded by a quartz wall coated with deuterated DLC films. This collection method has been successfully used in neutron–electric dipole moment experiments.[16]

CONCLUSIONS

Diamond-like carbon since its inception in 1962 has found applications in some very important areas. These applications include coatings used in scratch-resistant optics, razor blades, prosthesis in medical applications; electron emission surfaces in electronics; as an insulator material for copper heat sinks; in semiconductors such as solar cells and sensors for visible to infrared radiations; and as structural materials such as deuterated DLC film used for neutron storage in advanced research instrumentation. As technology matures the unique properties of DLC will find new and important applications.

ARTICLE OF FURTHER INTEREST

Diamond Films, p. 685.

REFERENCES

1. Aisenberg, S.; Chabot, R. Ion-beam deposition of thin films of diamond-like carbon. J. Appl. Phys. **1971**, *42* (7), 2953–2958.
2. Lettington, A.H. Applications of diamond-like carbon thin films. Carbon **1998**, *36* (5-6), 555–560.
3. Franta, D.; Ohlidal, I.; Bursikova, V.; Zajickova, L. Optical properties of diamond-like carbon films containing SiO_x. Diamond Relat. Mater. **2003**, *12*, 1532–1538.
4. Soga, T.; Hyashi, Y. Solar photovoltaic application of diamond-like a-C. In *Properties of Amorphous Carbon*; Silva, S.R.P., Ed.; EMIS Data Review Series No. 29; INSPEC, The Institution of Electrical Engineers: U.K., 2003; 355 pp.
5. Fung, M.K.; Lai, K.H.; Lai, H.L.; Chan, C.Y.; Wong, N.B.; Bello, I.; Lee, C.S.; Lee, S.T. Diamond-like carbon coatings applied to hard disks. Diamond Relat. Mater. **2000**, *9*, 815–818.
6. Yongjian, S.; Chu, X.; Yang, M.M. Ultrathin ion beamcarbon as an overcoat for the magnetic recording medium. IEEE Trans. Magnetics **2003**, *39* (1), 594–598.
7. Kan, M.C.; Huang, J.L. Field emission characteristics of amorphous diamond. J. Am. Ceramic Soc. **2003**, *86* (9), 1513–1517.
8. Zollondz, J.-H.; Krauser, J.; Weidinger, A.; Trautmann, C.; Schwen, D.; Ronning, C.; Hofsaess, H.; Schultrich, B. Conductivity of ion tracks in diamond-like carbon films. Diamond Relat. Mater. **2003**, *12*, 938–941.
9. Milne, W.I.; Tsai, J.T.H.; Teo, K.B.K. Novel field emission structure based on tetrahedrally bonded amorphous carbon. Diamond Relat. Mater. **2003**, *12*, 195–200.
10. Xu, T.; Pruitt, L. Diamond-like carbon coatings for orthopedic applications: an evaluation of tribological performance. J. Mater. Sci. Mater. Med. **1999**, *10*, 83–90.
11. Ronkainen, H.; Varjus, S.; Holmberg, K. Tribological performance of different DLC coatings in water-lubricated conditions. Wear **2001**, *249*, 267–271.
12. Hauert, R.; Müller, U. An overview on tailored tribological and biological behavior of diamond-like carbon. Diamond Relat. Mater. **2003**, *12*, 171–177.
13. Hauert, R. A review of modified DLC coatings for biological applications. Diamond Relat. Mater. **2003**, *12*, 583–589.
14. Grill, A. Diamond-like carbon coatings as biocompatible materials—an overview. Diamond Relat. Mater. **2003**, *12*, 166–170.
15. Cicognani, G.; Guerard, B.; Oed, A. Performance of MSGC on diamond-coated glass. Nuclear Instrum. Methods Phys. Res. A **1997**, *392*, 115–119.
16. van der Grinten, M.G.D.; Pendlebury, J.M.; Shiers, D.; Baker, C.A.; Green, K.; Harris, P.G.; Iaydjiev, P.S.; Ivanov, S.N.; Geltenbort, P. Characterization and development of diamond-like carbon coatings for storing ultracold neutrons. Nuclear Instrum. Methods Phys. Res. A **1999**, *423*, 421–427.

Differential Scanning Calorimetry

John O. Hill
La Trobe University, Melbourne, Australia

INTRODUCTION

It has always been recognized that there is a subtle difference between the titles ''thermal analysis'' and ''calorimetry'' and the associated groups of experimental techniques which is not simply one of ''semantics.'' There have been generations of interest in ''thermal events'' as a means of articulating changes in the thermodynamic behavior of materials. The analysis of thermal events may be approached in two different ways which intrinsically overlap. Either the analysis is designed to measure specific properties of a material, such as heat capacity, enthalpy, entropy, or free energy with high precision and accuracy at predefined temperatures and conditions, or thermal properties may be measured over a temperature range using a controlled temperature program. Thus, calorimetry is the measurement of heat changes which occur during a process whereas thermal analysis measures a property of a material as a function of temperature. It is therefore evident that ''thermal analysis'' and ''calorimetry'' are not synonymous terms. However, even with these simple definitions of both types of techniques, it is difficult to classify a technique such as differential scanning calorimetry (DSC) either as a calorimetric or a thermal analysis technique since in DSC, heat changes are measured as a function of temperature and hence DSC clearly epitomizes the synergy between calorimetry and thermal analysis. This review concentrates on the historical development of DSC and on the more recent development of temperature modulated DSC (tmDSC), together with selected applications and possible future enhancements. DSC is recognized as the primary thermal analysis technique which has revolutionized the characterization of the thermal properties of synthetic polymers and the chemical processing thereof. It is used as a routine analytical technique in materials science.

DIFFERENTIAL SCANNING CALORIMETRY

Historically, DSC is a development of differential thermal analysis (DTA) and both techniques measure the heat flux in a sample as a function of temperature. In DTA, the heat flux is measured as a temperature difference between a sample and a reference material as a function of temperature whereas in DSC, the heat flux in the sample is measured directly as a function of temperature. Thus, DSC is essentially ''quantitative DTA,'' or more precisely, DSC is a combination of DTA and adiabatic calorimetry. DSC is thus a ''differential calorimeter'' that achieves a continuous ''power compensation'' between sample and reference.

The practical distinction between DTA and DSC is simply related to the nature of the signal output from the equipment. For DTA, this is proportional to the temperature difference between the sample and a thermally inert reference material, usually alumina, when both are subjected to the same temperature program. For DSC, this is proportional to the difference in thermal power between the sample and the inert reference. For both DTA and DSC, the usual temperature program is a linear temperature change with respect to time. In classical DSC, the sample and the reference material are maintained at the same temperature throughout the controlled temperature program by a differential power input to the sample and reference material. This differential power input is recorded against the programmed (reference) temperature. Thermal events in the sample thus appear as deviations from the baseline in either an endothermic or exothermic direction, depending upon whether more or less energy has to be supplied to the sample relative to the reference material. The operating temperature range for DSC is typically subambient to 700°C. At subambient temperatures, the DSC sample container unit has to be protected from moisture condensation. In simple terms, the area of a DSC endotherm or exotherm is proportional to the associated enthalpy change of the identified thermal event. Thus, DSC instruments require calibration both in terms of ''temperature'' and ''energy'' and certified reference materials (CRMs) have been proposed for this purpose.[1] In simple terms, temperature calibration of a DSC is effected using metals such as indium, tin, or aluminium, which have well-defined melting points and energy calibration is effected using metals such as indium or tin or compounds such as potassium perchlorate or silver sulfate, which have well-defined fusion or transition enthalpies.[1]

A major application of DSC has been the measurement of heat capacity of a material, particularly at subambient temperatures. In simple terms, in the absence

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007736

of a sample and with both sample and reference crucibles empty, the DSC baseline is a horizontal line. When a sample is introduced, the DSC baseline is displaced in the endothermic direction and the displacement is proportional to the total heat capacity of the sample. The associated proportionality constant is determined by using a standard material such as sapphire,[1] scanned under similar conditions as the sample. DSC may also be used to measure the heat capacities of a material before and after a physical transition such as melting and thus a complete monitoring of the heat capacity of a sample is possible.

Another general application of DSC is for the determination of purity as described by Laye.[2] This application depends on an analysis of the shape of the fusion peak of the material in question. The DSC method mirrors the classical approach which depends on the depression of the freezing point by the presence of an impurity, as defined by Raoult's law. There is an increase in fusion peak width with increasing proportion of impurity and the subsequent analysis depends on using fractional areas of the fusion peak to calculate the extent of melting. It is assumed that the components of the impure mixture form an ideal solution on melting. Clearly, the DSC method is not applicable to samples which decompose at or near the melting temperature or to those which interact chemically with the impurity. Despite these restrictions, the DSC method has been widely used to determine the percentage purity of the active ingredient in pharmaceutical preparations and to determine the purity of metals such as silicon in electronic applications.

DSC has a wide and diverse application range. It is routinely applied to study the thermal behavior of polymers, glasses, ceramics, resins, oils, fats, waxes, clays and minerals generally, coal, lignite, wood, liquid crystals, explosives, pyrotechnics, propellants, pharmaceuticals, biological materials, metals and alloys, natural products, and synthetic catalysts over a wide temperature range. It is also used in quality control and chemical processing procedures, particularly as applied to synthetic polymers. DSC curves have also been used in ''fingerprinting'' and identifying components of mixtures. The identification of polymorphs in the context of pharmaceuticals is particularly relevant since different components may have different physiological activities. The investigation of potential reactivity between components of a drug preparation, as revealed by DSC, represents another significant application of this versatile analytical technique.

DSC is a valuable analytical technique in materials science and in particular, in the development of phase diagrams. This application is illustrated with reference to the binary system, triphenylmethane(A)/transstilbene(B).[3] The melting points of the pure components are obtained from the corresponding DSC curves and the melting behavior of mixtures of the components depends on the ''history'' of mixture preparation and for this general discussion, it is assumed that all mixtures have been prepared in completely molten form initially without any volatilization or decomposition. There is no deviation from the baseline until solid B crystallizes out and this corresponds to the first endotherm. The area of this endotherm depends upon the amount of B present in the mixture. As the temperature of the mixture decreases further, the eutectic temperature is reached and subsequently, solid A crystallizes out as a sharp exotherm. By studying a range of mixture compositions by DSC, the complete phase diagram for the binary system is obtained. This procedure has been particularly useful for study of metal mixtures and alloy systems. The phase diagrams of the ternary Ge-Sb-Bi system[4] and the very complicated Ag-Te-S-Se system with 17 phase regions[5] serve to illustrate the power of DSC in rationalizing the phase diagrams of these systems.

The dehydration of inorganic salt hydrates is usually stepwise but not always predictable. For copper sulfate pentahydrate, three DSC peaks are apparent corresponding to the loss of 2, 2, and 1 waters of hydration. Further, there is an increase in dehydration enthalpy of this salt with increasing numbers of hydration waters removed. This type of study is of importance in the estimation of the purity of the sample since the dehydration enthalpy is directly proportional to the water content of the sample. Similarly, DSC can be used to determine the moisture content of materials generally, since the measured endothermic dehydration DSC peaks are directly proportional to the water content of the material.

The application of DSC to study the thermal behavior of synthetic polymers has been most extensive and intensive and has been comprehensively reviewed by Turi.[6] Most solid polymers are formed by rapid cooling to low temperatures from the melt by a process known as quenching and are thus initially in the glassy state. The transition from a glass to a rubber, termed the ''glass transition'' is accompanied by a change in heat capacity but no change in enthalpy. The transition thus appears on the DSC curve as a discontinuity of the baseline at the glass transition temperature. This is a characteristic property of a polymer and of its processing. As the temperature is subsequently increased, the polymer may crystallize giving rise to an exotherm before melting occurs. At higher temperatures, the polymer may decompose (degrade) or oxidize depending on the nature of the surrounding atmosphere. Thus, a single DSC scan of a polymer is able to define its thermal characteristics and, most importantly, its thermal stability. As an example, the DSC analysis of poly (ethyleneterephthalate) (PET) heated in nitrogen reveals the glass transition, exothermic crystallization, and endothermic melting in a single scan.[7]

Recycling of plastic waste has huge commercial significance but the process is slowed by the problems of component identification, sorting, and collection. DSC can be used to identify the components of a plastic waste.[8] The components are unambiguously identified by their respective exothermic crystallization peaks. Other peaks can be used for this purpose but glass transitions are generally not used in fingerprinting operations due to their poorly defined temperature ranges.

DSC has been used to test the completeness of curing of epoxy resins. The completeness of curing is revealed by a second DSC scan of the sample, which shows no residual exotherm but indicates the glass transition for the cured resin. This application again shows the uniqueness of DSC in revealing a complex chemical processing phenomenon.

One of the more celebrated applications of DSC in materials science which is of very significant commercial significance relates to a DSC analysis of high alumina cement.[9] Some three decades ago, alumina diluted cement led to the weakening and in some cases, the collapse of concrete structures, particularly in third world countries. The weakening process is due to a conversion of the initial product of setting into gibbsite or hydrated alumina. DSC can be applied to determine the extent of such conversion by measuring the relative peak heights of the alumina cement DSC scan. In practice, a calibration is carried out using a cement sample of known composition, but the analysis is rapid and reliably quantitative.

DSC has also been used as an analytical technique in food science. Structural changes in fats and waxes characterize the product such that it is possible, for example, to easily differentiate between butter and margarine on the basis of their different DSC profiles.[10] Although it is not usually possible to unambiguously characterize each of the peaks and shoulders, the overall profile is a fingerprint of a particular fat or fat blend.

Finally, DSC has been applied with some success to the study of kinetics of chemical reactions. This study has as its foundation the identification of reaction rate with DSC signal output and the extent of reaction with fractional DSC peak area. This application has many difficulties and shortcomings which have been reviewed by Laye.[11]

These examples are sufficient to clearly reveal the pre-eminence of DSC as the thermoanalytical technique of choice in materials science.

TEMPERATURE MODULATED DIFFERENTIAL SCANNING CALORIMETRY

In classical DSC, the imposed temperature program is a linear change with respect to time. However, more complex programs can be invoked by overlaying the conventional linear heating with a regular modulation, thereby creating temperature modulated DSC. In the original form of the technique described by Reading et al.[12,13] in 1993, the modulation was sinusoidal. Other forms of modulation have since been introduced such as square wave and saw-tooth. Three developments of calorimetry have thus combined over the past decade to dramatically enhance the capabilities of DSC. These are the high precision of adiabatic calorimetry, the speed of operation, and small sample size associated with DSC and the superimposition of measurement of frequency dependence of thermal events thereby leading to the evolution of tmDSC. ''Modulation'' is perhaps the most significant development with respect to thermal analysis techniques, paralleling in impact the ''Fourier transform'' development of infrared spectroscopy.

The essential comparative features between conventional DSC and tmDSC are shown schematically in Figs. 1A and B. In conventional DSC, the multiple temperature sensors average the calorimeter temperatures. In tmDSC, the linear temperature program of conventional DSC is modulated by superimposing a periodic wave form of small amplitude on the linear temperature program. The most commonly applied wave form is a sine wave, as shown by the equations for sample temperature in Fig. 1B. Essentially, tmDSC provides periodic heating and cooling within each cycle but the overall effect is the same as in conventional DSC namely, a linear change in average temperature with time. The resultant heat flow signal is a composite of the response to a thermal event occurring in the sample and the response of the underlying heating program. Five component signals thus derive from a tmDSC experiment, which are described as: i) the average or underlying signal, equivalent to DSC; ii) the in-phase cyclic component; iii) the out-of-phase component; iv) the reversing component to the underlying heat flow; and v) the nonreversing heat flow. Two of these components are inter-related: the reversing component (iv) is derived from multiplying the heat capacity of the sample by the heating rate, and subtraction of (iv) from the underlying signal (i) gives the nonreversing heat flow (v).

Hence, in the simplest terms, tmDSC is a description of the heat flow into the sample resulting from the sinusoidal modulation of the temperature program. Two properties of the sample can be investigated by tmDSC, the heat capacity which is directly related to the ''reversing component'' and a kinetically hindered thermal event which is related to the ''nonreversing'' component. Conventional DSC provides only a measure of the total heat flux into a sample as a function of temperature whereas tmDSC allows the heat capacity and kinetic components to be separated. However,

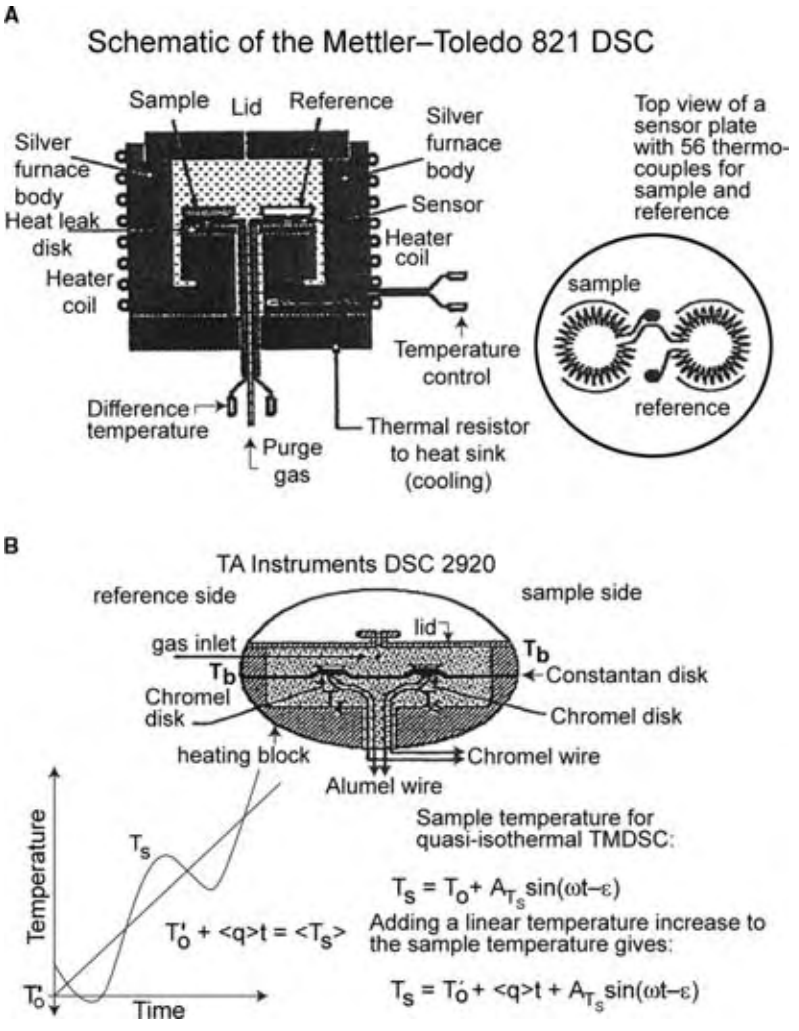


Fig. 1 Schematic diagram of: (A) classical heat flux DSC; (B) tmDSC.

the conditions of a tmDSC experiment are more critical than in conventional DSC since the selection of the period and amplitude of the modulation need to be selected in addition to the underlying heating rate. Sample size and period of modulation influence the ability of the sample to follow the temperature modulation and hence the apparent value of the heat capacity. A more detailed theoretical treatment of the relationship between heat capacity and the reversing component of tmDSC shows that a ''reversing time correction factor'' has to be included which depends on the nature of the sample and its mass. By using multiple frequencies for modulation, the required correction factor can be determined in a single tmDSC run, which further indicates the supremacy of tmDSC for heat capacity measurements. A further variation in tmDSC is that the underlying heating rate can be set to zero, in which case, measurements are carried out under quasi-isothermal conditions. As for DSC, calibration of tmDSC is essential and this has presented a challenge. At present, the nematic to smetic-A transition in a cyanobiphenyl liquid crystal (80CB) has been suggested[12] for calibration of tmDSC.

Alternative modulation functions and data analysis procedures have been applied in tmDSC. Different instrument manufacturers have applied square wave or saw-tooth modulation, coupled with Fourier transform analysis methodologies. Alternative modulation functions provide some additional advantages. For example, a square wave function ensures that a steady state is achieved over an isothermal plateau since the signal during such a period is the nonreversing contribution. The amplitude provides a measure of the reversing signal. However, there are some disadvantages of these alternative modulation functions in that accuracy and resolution may be compromised since not all the necessary data are made available. The Wunderlich ATHAS Laboratory has investigated other types of modulation in tmDSC[14–16] and has shown that such modulations are likely to be incorporated into future generations of tmDSC systems for enhanced precision of measurement of heat capacity and glass transitions.

Wunderlich[17] has also comprehensively reviewed the development of temperature modulated calorimetry into the 21st century in conjuction with its predicted scope for further application in the physical, chemical, and materials sciences.

There are many advantages of tmDSC over conventional DSC. Use of the former technique with appropriate frequencies and amplitudes allows separation of reversing processes, such as glass transitions, from nonreversing processes, such as relaxation endotherms or cure processes. With tmDSC, baseline curvature of the cyclic signal is usually insignificant, thereby making it easier to distinguish between baseline effects and real thermal transitions. The signal-to-noise ratio of the cyclic measurement of heat capacity is generally greater, since all drift or noise at frequencies other than that of modulation is ignored by the Fourier transform analysis. In general, in tmDSC, resolution of thermal events is improved because very low underlying heating rates are used.

Applications of tmDSC in general rely on the enhanced precision of the technique compared to conventional DSC as attained by the maintenance of a steady state and a negligible temperature gradient within the sample throughout the run and hence, it is possible to measure heat capacity quasi-isothermally. In order to achieve the highest precision in tmDSC, the theory of the technique indicates that a high heating rate should be applied together with a large sample mass.[12] These conditions lead to rapid analysis of large samples and provide a further advantageous feature of DSC over other thermal analysis techniques.

It is generally acknowledged that DSC is the pre-eminent thermal analysis technique and that it has progressively become the established technique for the study of the thermal behavior of polymeric materials. Conventional DSC correlates thermal power with heat capacity and the integral thereof to energy and entropy. Thus, DSC has been applied to determine heat capacities of a wide range of materials. Conventional DSC is able to determine heat capacity to an uncertainty of 1–2%: tmDSC is able to measure this parameter to an uncertainty of less than 1% with reproducible reliability. It is the ''temperature modulation'' feature of tmDSC which has confirmed this technique as the most versatile and most reliable of the thermal analysis techniques. Its versatility is further qualified by its ability to characterize the thermal behavior of materials without the need to have a detailed knowledge of the fundamental theoretical principles which underscore the basis of the technique.

Apart from the determination of heat capacities of a wide variety of materials, the other most common use of tmDSC is to determine the thermal behavior of synthetic polymers. A tmDSC analysis of poly(ethyleneterephthalate) has been described by Haines et al.[18]. The five components of the tmDSC signal are revealed and the difference in the heat capacity profiles corresponding to ''reversing Cp,'' ''kinetic Cp,'' and ''nonreversing Cp'' are particularly striking and collectively reveal the ability of tmDSC to provide complete heat capacity characterization of a polymer sample.

tmDSC has major advantages over conventional DSC in terms of differentiating overlapping thermal events. A tmDSC analysis of a poly(ethyleneterephthalate)–acrylonitrile/butadiene/styrene (PET–ABS) polymer blend has been reported by Laye.[19] The total heat flow profile for the blend and its separation into the reversing and nonreversing components is apparent. The glass transition temperature for PET is around 65°C and that for ABS is 105°C, which would mask the glass transition of ABS in conventional DSC. Since the glass transition is a major identification characteristic of a polymer, it is important that this is unambiguously defined in the thermal analysis of polymer blends and hence in the chemical processing of polymers.

tmDSC also offers a quantitative investigation of polymer glass transitions as a function of ''thermal history'' of the sample, since glass transition kinetics vary with the enthalpy of the glassy state. Further, the ability of tmDSC to separate reversing and nonreversing processes during phase transitions offers quantitative investigation of cold crystallization and melting with superheating transitions. tmDSC therefore offers unparalleled opportunities for the quantitative analysis of the thermal properties and thermal behavior of materials, particularly synthetic polymeric materials.

Reactions involving partial diffusion can be studied by tmDSC such as epoxy resin curing.[13] The glass transition temperature is measured as a function of the conversion fraction. tmDSC is able to separate the glass transition and the melting transition in interpenetrating polymer networks prepared by in situ copolymerization.[13] Additionally, an estimate of the degree of crosslinking in the copolymer is possible from the composite thermal data, together with an estimate of the glass transition temperature of the copolymer phase.

Numerous tmDSC applications were presented at the Eighth European Symposium on Thermal Analysis & Calorimetry (ESTAC 8) in Barcelona in 2002[20] and a selection of these is briefly summarized here to illustrate the present scope of applications of this technique. tmDSC has been applied to study the morphology and recrystallization kinetics of nickel sulfide. Such information is of significance in understanding the failure of toughened glass panels when subjected to external stresses. tmDSC is shown to be ideal for rationalizing the melting behavior of polytetrahydrafurans and their blends. The PTHF oligomer shows two endothermic peaks. The lower peak shows temperature dependence on molar mass whereas the

higher peak is independent of molecular mass of the polymer. The order–disorder transition temperature of triblock copolymer gels, swollen with parafinic oils has been studied by tmDSC. The order–disorder transition temperature of these gels is detected as an abrupt change in heat capacity of the sample and is dependent on the polymer content of the gel and on the molar mass of the oil. Thus, tmDSC assists an understanding of the mechanism of order–disorder transitions in triblock copolymers swollen with solvents. tmDSC has been applied to study ''intelligent'' polymer systems which show major changes in properties with small physical or chemical stimuli. The kinetics of ''demixing'' (shrinking) and ''remixing'' (swelling) are important considerations in possible applications of these ultra-new hi-tech systems. In this context, it is shown that tmDSC is ideal for the characterization of the kinetics of phase separation in partially miscible polymer blends and polymer solutions. ''Mixing/demixing'' phenomena are manifested as large excess contributions to the apparent heat capacity signal of the system on the time scale of temperature modulation which enable the real-time monitoring of demixing, remixing, and vitrification processes occurring. tmDSC thus allows investigation of nano-scale miscibility phenomena in water-soluble polymers and related hydrogels. tmDSC has been shown to be valuable for studying the deconvolution of ''nonreversing'' and ''reversing'' heat flow effects in reacting polymer systems. tmDSC has been further extended to study temperature-induced phase separation in poly(ethylene) /poly(ethersulfone) (PEO–PES) blends. tmDSC has also been applied to model the effect of additives on the cure kinetics of epoxy resins and for in situ detection of reaction-induced phase separation. The cure temperature is crucial in determining the final phase-separated morphology and impact properties of such resins. tmDSC, in conjunction with TGA, has been applied to characterize elastomers. From the component pyrolysis temperatures, tmDSC can be used to quantitatively determine the individual polymer compositions. tmDSC has also been applied to study the vitrification, devitrification, and dielectric relaxations associated with the nonisothermal curing of an epoxy-amine system. Such systems are characterized by continuous heating transformation cure diagrams, which are readily obtained by tmDSC. tmDSC has also been applied to study high temperature polymeric materials, such as bis-maleimide resins which find extensive use as lightweight fiber-reinforced structural composites and electrical insulators. The cold crystallization and melting of PET and poly(ethylene-2,6-naphthalene dicarboxylate) (PEN) have been studied by tmDSC with a net zero heating rate.[21]

Finally, maximum precision in tmDSC is obtained with very slow underlying heating rates which allow an adequate number of temperature modulations over a region of changing heat capacity such as a transition region. This condition has recently been achieved by the new DSC cell and ''Tzero'' technology, developed by TA Instruments (U.S.A.).[20] The thermal capacitance and resistance of the DSC measuring system are effectively eliminated and this feature, coupled with the overall temperature control capabilities, has made much shorter modulation periods practical. Hence, ''fast tmDSC'' is now a reality.

SIMULTANEOUS THERMOGRAVIMETRIC ANALYSIS–DIFFERENTIAL SCANNING CALORIMETRY

Half a century ago, the Pauliks[22] launched the ''derivatograph'' which was a combination of a thermogravimetric analysis system and a differential thermal analysis system and constituted the first simultaneous thermal analysis system. STA systems involve the mounting of a DTA head onto the thermobalance suspension or rise rod, which is connected directly to a microbalance. Modern STA instruments are capable of a resolution of the order of 1 microgram with sample masses ranging from 5 to 100 mg. Essentially, STA involves measurements of two thermal properties on the same sample subjected to the same temperature program. There are advantages of STA over individual TGA and DTA systems. Fundamentally, it is difficult to correlate results obtained on different portions of a sample with different thermal analysis techniques due to ''sample'' and ''instrument'' factors whereas, in general, results obtained on the same sample by two thermal analysis techniques simultaneously are reproducible. In general, STA may be viewed as yielding data which are more significant than the sum of the two sets of data obtained individually and thus, a synergistic effect operates in STA. This effect is readily apparent. TGA can only detect thermal events associated with a mass change, so a phase change such as melting is not detected by this technique. However, melting is readily detected by DTA as a sharp endotherm. A combination of TGA and DTA shows no sample mass change associated with the melting DTA peak. Further, melting with decomposition is readily identified by STA since the endothermic DTA peak shows shoulders which correlate with a sample mass decrease. Other obvious advantages of STA are ''time saving''—only one experiment is necessary instead of two—and ''sample saving.'' The latter is particularly advantageous for samples available in short supply or for expensive samples such as the active ingredient of drugs and biological materials. However, there is a perceived disadvantage of STA in that the sensitivity of individual thermal analysis

techniques is decreased on combination due to essential compromises in instrumental design. It is this feature which has limited the development of TGA–DSC systems. It is not possible to incorporate power compensated sensors into a STA system but heat flux sensors have replaced DTA sensors in STA systems and the latter are available commercially. The essential difference between ''heat flux DSC'' and DTA lies in the conversion of the temperature difference between sample and reference into differential power. The algorithm for this conversion is contained in the instrument software and the design of the DSC cell is critical if this algorithm is to be transferable from one experiment to another and independent of the sample. This is the most critical limitation in the design of a combined TGA–DSC cell. The DSC sensor must not change position with different sample masses so as to avoid changes in heat transfer characteristics and subsequent instrumental base-line shifts with concomitant loss of overall sensitivity. However, if such instrumental challenges can be overcome, STA in the form of TGA–DSC offers all the advantages of TGA–DTA with the added advantage of quantification of the identified thermal events both in terms of mass and energy changes.

Applications of TGA–DSC are not as numerous as for TGA–DTA and largely involve studies of solid-state phase changes. Simultaneous TGA–DSC has been applied by Warrington[23] to study the thermal characteristics of ammonium nitrate. The combined curves reveal that partial sublimation, then vaporization of the sample occurs before and after the melting point. More accurate estimates of the transition and fusion enthalpies are thus available based on the associated sample mass, as derived from the TGA curve. Further, it is known that traces of moisture in the sample can affect the temperature ranges of the transitions both on heating and cooling. Using TGA–DSC, the moisture content is known precisely at each stage of the thermal analysis.

TGA has been used with some success to determine the extent of sublimation of volatile solid materials.[23] TGA–DSC can be applied to determine the corresponding sublimation enthalpies, which are usually difficult to determine by conventional methods. The determination of sublimation enthalpies is notoriously difficult since with most volatile solids, the effusion process is rarely smooth and continuous and thus the use of STA for this purpose is a most welcome adjunct to many less reliable analytical techniques which rely on an ''ideal'' sublimation process. For sublimation studies, the sample is sealed in a container with a small pin-hole in the lid. The TGA curve shows a progressive mass loss until the sample is exhausted. The corresponding DSC curve shows a deviation from the base-line equivalent to the sublimation enthalpy which can be correlated with the rate of mass loss. Calibration of the system with a substance of known sublimation enthalpy, such as benzoic acid or ferrocene, allows the derivation of absolute sublimation enthalpies. With certain restrictions, this type of STA system can be applied to determine vaporization enthalpies of liquids. Overall, TGA–DSC is a valuable adjunct to the family of thermal analysis techniques particularly with respect to the specialist applications described herein.

CONCLUSIONS

Differential scanning calorimetry is the most sophisticated member of the family of thermal analysis techniques and bridges the nexus between ''thermal analysis'' and ''calorimetry.'' DSC essentially combines the principles of adiabatic calorimetry with those of differential thermal analysis and is essentially a ''quantitative DTA.'' DSC is of paramount importance in the characterization of the thermal behavior of synthetic polymers and polymer blends. It is also of major significance in the study of resin curing phenomena. DSC is recognized as the primary technique for the determination of heat capacity, particularly at subambient temperatures. DSC has also been applied to determine the percentage purity of a wide range of materials and in this context it is a primary analytical technique in pharmaceutical and medicinal chemistry. DSC is also of major significance in metallurgy in terms of the derivation of alloy phase diagrams and for the purity determination of metals and metalloids associated with the electronics industry. The recently derived technique known as temperature modulated DSC further enhances the precision and accuracy of DSC as a quantitative thermoanalytical technique and further refines the application range of DSC. The particular advantage of tmDSC is that samples are essentially maintained in an environment of thermal equilibrium during the modulated heating program and thereby, the thermodynamic data obtained are reliable, reproducible and relate to defined processes. DSC has been combined with thermogravimetric analysis TGA to form simultaneous thermal analysis TGA–DSC. Although the instrumental challenges are significant, this type of STA system is invaluable for the determination of the thermodynamic parameters associated with solid-state phase changes and the processes of sublimation and vaporization. Since the sublimation process of most volatile solids is complex and rarely complete, the TGA profile for a volatile solid indicates the extent of sublimation and the corresponding DSC profile indicates the sublimation enthalpy over the same temperature range. Overall, DSC, in all of its configurations is the most versatile and most significant thermal analysis technique and its application range in materials science is limitless particularly with respect to materials of technological importance.

REFERENCES

1. Richardson, M.E.; Charsley, E.L. *Handbook of Thermal Analysis and Calorimetry*; Elsevier: Amsterdam, 1998; Chapter 13.
2. Laye, P.G. Differential thermal analysis and differential scanning calorimetry. In *Principles of Thermal Analysis and Calorimetry*; Haines, P.J., Ed.; Royal Society of Chemistry: Cambridge, U.K., 2002; Chapter 3, 86–89.
3. McNaughton, J.L.; Mortimer, C.T. Differential Scanning Calorimetry, IRS-Physical Chemistry, Series 2; Butterworths: U.K., 1975; Vol. 10, 28.
4. Surinach, S.; Baro, M.D.; Tejerina, F. Proceedings of the Sixth ICTA; Wiederman, H.G., Ed.; Birkhauser-Verlag: Basel, 1980; Vol. 1, 155.
5. Boncheva-Mladenova, Z.; Vassilev, V. Proceedings of the Sixth ICTA; Wiederman, H.G., Ed.; Birkhauser-Verlag: Basel, 1980; Vol. 2, 99.
6. Turi, E.A., Ed. *Thermal Characterisation of Polymeric Materials*, 2nd Ed.; Academic Press: San Diego, 1997; Vols. 1 and 2.
7. Laye, P.G. Differential thermal analysis and differential scanning calorimetry. In *Principles of Thermal Analysis and Calorimetry*; Haines, P.J., Ed.; Royal Society of Chemistry: Cambridge, U.K., 2002; Chapter 3, 61.
8. Brown, M.E. *Introduction to Thermal Analysis—Techniques and Applications*; Kluwer: Amsterdam, 2001; Chapter 4, 83–84.
9. Laye, P.G. *Differential Thermal Analysis and Differential Scanning Calorimetry*; Royal Society of Chemistry: Cambridge, U.K., 2002; Chapter 3, 61–62.
10. Brown, M.E. *Introduction to Thermal Analysis—Techniques and Applications*; Kluwer: Amsterdam, 2001, Chapter 4, 87.
11. Laye, P.G. *Differential Thermal Analysis and Differential Scanning Calorimetry*; Royal Society of Chemistry: Cambridge, U.K., 2002; Chapter 3, 83–84.
12. Reading, M. Modulated differential scanning calorimetry—A new way forward in materials characterization. Trends Polym. Sci. **1993**, *8*, 248–253.
13. Reading, M.; Elliott, D.; Hill, V.L. A new approach to the calorimetric investigation of physical and chemical transitions. J. Therm. Anal. **1993**, *40*, 949–955.
14. Wunderlich, B.; Jin, Y.; Boller, A. Mathematical description of differential scanning calorimetry based on periodic temperature modulation. Thermochim. Acta **1994**, *238*, 277–293.
15. Kwon, Y.K.; Andorsch, R.; Pyda, M.; Wunderlich, B. Multifrequency saw-tooth modulation of a power-compensated differential scanning calorimeter. Thermochim. Acta **2001**, *367* (8), 203–215.
16. Pyda, M.; Kwon, Y.K.; Wunderlich, B. Heat capacity measurements by saw-tooth modulated standard heat-flux differential scanning calorimetry with sample temperature control. Thermochim. Acta **2001**, *367* (8), 217–227.
17. Wunderlich, B. Temperature modulated calorimetry in the 21st century. Thermochim. Acta **2000**, *355*, 43–57.
18. Haines, P.J.; Reading, M.; Wilburn, F.W. *Handbook of Thermal Analysis and Calorimetry*; Brown, M.E., Ed.; Elsevier: Amsterdam, 1998; Vol. 1, Chapter 5.
19. Laye, P.G. Differential thermal analysis and differential scanning calorimetry. In *Principles of Thermal Analysis and Calorimetry*; Haines, P.J., Ed.; Royal Society of Chemistry: Cambridge, U.K., 2002; Chapter 3, 68–69.
20. Nomen, R. Symposium Chair. Eighth European Symposium on Thermal Analysis & Calorimetry (ESTAC 8), Barcelona, August 2002; Abstracts cited: 6-02, 6-03, 6-P1, 6-P2, 6-P3, 6-P14, 6-P16, 6-P23, 6-P26, P&F-06; Institute Quimie De Sarria, Universitat Ramon Llull, Barcelona, Spain.
21. Montserrat, S.; Roman, F.; Colomer, P. Study of the crystallization and melting region of PET and PEN and there blends by tmDSC. J. Therm. Anal. Calorimetr. **2003**, *72*, 657–666.
22. Paulik, F.; Paulik, J. Simultaneous thermoanalytical examinations by means of the Derivatograph. In *Comprehensive Analytical Chemistry*; Svehla, G., Ed.; Elsevier: Amsterdam, 1981; Vol. X11(A).
23. Warrington, S.B. Simultaneous thermal analysis techniques. In *Principles of Thermal Analysis and Calorimetry*; Haines, P.J., Ed.; Royal Society of Chemistry: Cambridge, U.K., 2002; Chapter 6, 172–174.

Dimethyl Ether

Abhay Sardesai
Department of Chemical Engineering, University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

The driving force for developing alternative fuels is no longer solely to reduce oil dependence on select oil exporting countries. Noxious exhaust emissions generated by combustion of fossil fuels as well as increasing levels of carbon dioxide, a major contributor toward global warming, have necessitated the need to develop alternative fuels that address clean-burning as an important criterion. Dimethyl ether (DME) is identified as a multisource, multipurpose fuel that can effectively allay both of these concerns by establishing stable indigenous fuel supply and alleviating environmental concerns. Dimethyl ether can be readily produced from natural gas and coal, renewable resources, such as biomass and wood, as well as waste matter. It can be used as a clean-burning fuel in diesel engines, a household fuel [liquefied petroleum gas (LPG) alternative] for heating and cooking, a fuel for gas turbines in power generation, a fuel for fuel cells, a feedstock for producing chemicals and oxygenates, and a propellant in the aerosols market. Applications of DME as a fuel in various industries have been successfully tested by a number of leading companies. Feasibility studies on producing DME from syngas have been carried out by NKK Corporation at their 5 ton/day facility and by Haldor Topsoe at their 50 kg/day facility. In addition, commercial plants have been planned and/or are under way, such as India DME project of 5000 MTPD, DME International Corporation, Japan, of 2500–4500 MTPD, and Japan DME Inc. of 5000 MTPD. Although studies indicate that DME will become economical when oil prices are at or above a threshold price per barrel, it is clear that recent strides in DME synthesis technology show that it holds tremendous promise with a realistic aspiration to develop into an alternative transportation and domestic fuel. It is apparent that DME production costs will get even lower when its production capacities increase as more and more countries start embracing the idea of utilizing this fuel as a viable alternative.

BACKGROUND

Dimethyl ether is gaining worldwide recognition as a multisource, multipurpose clean fuel and chemical feedstock for the 21st century. It is a technically mature, environmentally friendly, and market acceptable alternative fuel. As shown in Fig. 1, DME can be produced from a variety of sources, and its end use includes a number of important applications. Dimethyl ether can be manufactured in large quantities from coal, natural gas, biomass, and municipal solid waste. Current technologies for producing DME on a large scale include those of NKK Corporation, Haldor Topsoe, Air Products, and Toyo Engineering Corporation (TEC).

The simplest ether compound, DME has been shown to be both nontoxic and environmentally benign. Dimethyl ether has a variety of uses in the fuel and the chemical industries. Currently, the major use of DME is as a propellant in the aerosols industry. Its cetane number (a quantitative indicator of the ignition quality of diesel fuel) is high, ranging from 55 to 60, so that it can be used in diesel engines. Its flame is a visible blue flame similar to that of natural gas, and it can be used just as it is in an LPG cooking stove, and it does not produce any aldehydes. The toxicity of DME is low, about the same as that of LPG, and even lower than that of methanol.[1] Because of DME's restrictive use, current world capacity is only 150,000 metric tons/yr.[2] Future mega-plant technology providers include NKK, Haldor Topsoe, Lurgi, Toyo Engineering, Mitsubishi Gas Chemicals, Kvaerner, Synetix, etc.

There have been a number of joint efforts between companies at a multinational level to gather and disseminate know-how about DME properties, DME synthesis processes, and DME applications. These associations have furthered DME's cause by jointly conducting testing studies on DME applications as an automotive fuel, an LPG substitute, and a fuel for power generation. The International DME Association (IDA) plays a significant role in enhancing knowledge about DME to the general public and is a central source of comprehensive DME related information (www.aboutdme.org). Institutional members of this association include BP, Carbinol Technologies, AVL Powertrain Engineering, Inc., Air Products, Renault, Snamprogetti, Lurgi, Akzo Nobel, Shell Global Solutions, Haldor Topsoe, Volvo, Atrax Energi AB, Mitsui, Japan DME Forum (JDF), etc. It is an excellent source of comprehensive information about DME synthesis,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007739

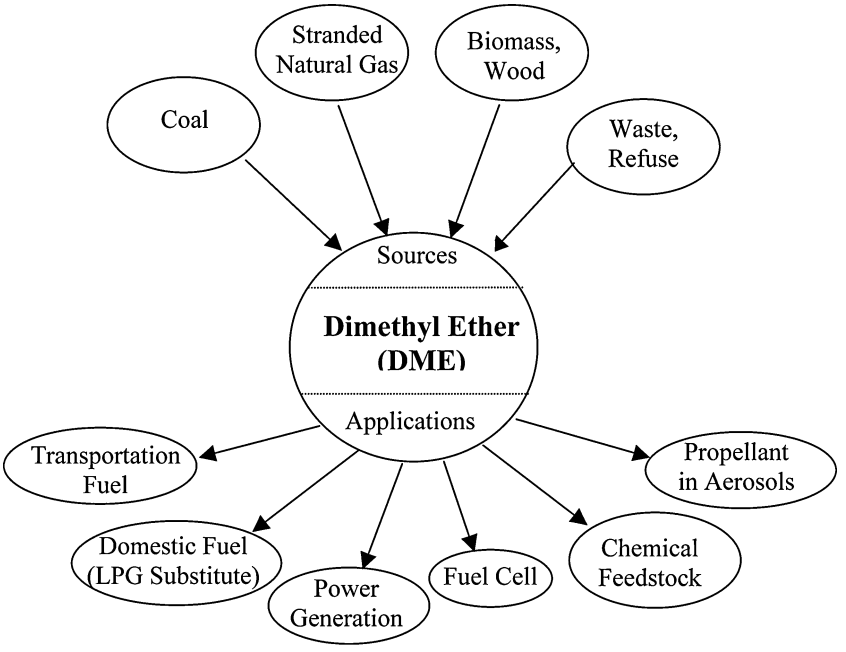


Fig. 1 Dimethyl ether—a multisource, multipurpose chemical.

applications of DME as a fuel in various sectors, as well as the economic, environmental, and commercial aspects of DME utilization. DME International Corporation was established in Tokyo and Japan by a consortium of eight companies to commercialize and market DME as a new source of alternative energy. NKK Corporation, Toyota Tsusho Corporation, and Hitachi, Ltd. are the leading participants. Other participants include a trading company, Marubeni Corporation; two energy-related firms, Idemitsu Kosan Co., Ltd. and INPEX Corporation; an industrial gas supplier, Nippon Sanso Corporation; LNG Japan Corporation; and one non-Japanese company, TotalFinaElf, which is the world's fourth largest petroleum company. The objective of this corporation is to produce DME by a low-cost direct synthesis route by taking advantage of stranded reserves of low-grade coal and smaller gas fields present in Asia. Japan DME Forum was formed by an alliance of 37 Japanese companies, two universities, and several trade organizations tasked with the development of large-scale DME technology. It is also a member of the IDA.

PROPERTIES OF DME

Dimethyl ether, the simplest form of ether, is a colorless gas at room temperature and pressure. It is noncarcinogenic, virtually nontoxic, and also noncorrosive in nature. Dimethyl ether is a liquid at the moderate pressure of approximately 496 kPa. Its handling characteristics are very similar to LPG. It is stored in conventional atmospheric pressure tanks as a refrigerated liquid (−25°C) or in pressurized tanks (5 bar) at 20°C. Dimethyl ether can be ocean transported in LPG tankers. While its net calorific value of 28.88 kJ/kg is lower than that of propane, butane, and methane, it is higher than that of methanol. In gaseous state, its net calorific value is 59,413 kJ/Nm^3, which is higher than that of methane. With regard to the combustion properties, its explosion limit is wider than that of propane and butane, but almost identical to that of methane and narrower than that of methanol. Table 1 compares the properties of DME with those of methane and propane.

USES/APPLICATIONS

High-purity DME is currently used as an aerosol propellant owing to its environmentally benign characteristics. In addition, it can be widely used as an

Table 1 Comparison of DME properties with methane and propane

Property	DME	Methane	Propane
Normal boiling point (°C)	−24.9	−161.5	−42.1
Vapor pressure at 25°C (atm)	6.1	246	9.3
Liquid density (g/cm^3)	0.668	–	0.501
Net heating value (kcal/kg)	6,900	12,000	11,100
Flammability limits in air (vol%)	3.4–17	5–15	2.1–9.4

ultraclean fuel for compression ignition and diesel engines (buses, taxis, trucks, and off-road vehicles such as construction equipment), a household fuel for heating and cooking, a power generation fuel in gas turbines, a hydrogen source in fuel cell cars, and a chemical feedstock for fuel additives and chemicals.

Dimethyl Ether as a Transportation Fuel

The need to develop alternative fuels is motivated by two important issues. Self-reliance in meeting fuel demands is the first factor, as fossil fuel supply is controlled by a small number of oil exporting countries. Second, the deleterious effect of fossil fuel combustion on the environment has prompted urgent use of clean-burning alternative fuels. Harmful exhaust emissions of internal combustion engines, such as nitrogen oxides (NO_x), carbon monoxide, volatile organic compounds, and carbon particulates, as well as high levels of carbon dioxide, a greenhouse gas contributing toward global warming, are causing irreparable harm to the environment. Use of DME, a cleaner burning oxygenate, has a very positive impact on these problems. Properties of DME are compared with those of diesel in Table 2.

Dimethyl ether has attractive advantages as an ultraclean transportation fuel alternative, most notably as follows:

- Dimethyl ether molecules have no carbon-to-carbon bonds, which diminish the tendency to form solid carbon particulates during combustion.
- Dimethyl ether has a low autoignition temperature.
- Dimethyl ether has a high cetane number of 55–60, compared to about 45 for diesel. Molecular bonds of DME break up to form radicals at reasonable activation energy, which leads to high cetane numbers.
- The normal boiling point of DME is −25°C. This provides fast fuel/air mixture formation, reduces ignition delay, and imparts excellent cold start properties.

Table 2 Comparison of DME properties with diesel

Property	DME	Diesel
Normal boiling point (°C)	−25.1	180–370
Liquid density (g/cm^3)	0.67	0.84
Ignition temperature (°C)	235	250
Explosion limit (%)	3.4–17	0.6–0.65
Cetane number	55–60	40–55
Net heating value (kcal/kg)	6,900	10,000

- Oxygen content of DME is 35 wt%, which suppresses the formation of soot and facilitates smokeless combustion.
- Dimethyl ether combustion produces low engine noise or quiet combustion and ultralow exhaust emissions, especially those of NO_x compounds.
- Dimethyl ether qualifies as a renewable fuel, as it can be produced from biomass and wood.

Dimethyl ether economics are comparable to those of diesel fuel, especially at low NO_x regulations, in terms of thermal efficiency. This is possible without major engine modifications, and as such end-user economics could be quite favorable for DME as compared to the costs of converting diesel engines to run on spark-ignited alternative fuels.[3] NKK Corporation, Japan, has been working on a novel process for mass-production and application technology for DME. The company aims to encourage development of DME-fueled vehicles, and is also researching applied technology for DME-fired gas turbine and diesel engine generator systems as well as DME fuel cells. They have tested DME on diesel engines with only minor modifications to the fuel injection system, with regard to engine performance and exhaust gas composition. No black soot or smoke was generated, ignition was quiet, generation of NO_x was reduced by 20–30% as compared to that of diesel fuel at the same cylinder pressure, and combustion time was shorter, translating into higher thermal efficiency.

The Combustion Laboratory at Pennsylvania State University has successfully developed a project involving conversion of a faculty–staff campus shuttle bus to operate on DME. The conversion of the Penn State shuttle bus was a collaborative effort between the Pennsylvania Department of Environmental Protection, Air Products, the Federal Energy Technology Center, Navistar International, Champion Motor Coach, Penn State's Combustion Lab, Fleet Services, and Pennsylvania Transportation Institute. The Combustion Laboratory will eventually convert the shuttle bus engine to operate on blends of DME, diesel fuel, and a lubricity agent.

Dimethyl Ether as a Household Fuel

Dimethyl ether has the capability to substitute LPG as a household fuel for heating and cooking purposes, especially in Asian countries such as Japan, China, and India where the demand for LPG is growing. Utilization of DME as a household fuel can also reduce the burden on developing countries that use solid fuels like coal and firewood, which in turn have harmful effects on the environment. Dimethyl ether can be produced indigenously in developing countries by taking

advantage of stranded natural gas fields in their vicinity. Dimethyl ether is a clean fuel and the total investment cost would be small because existing LPG infrastructures could be used with minor modifications.[4]

Dimethyl ether flame is a visible blue flame similar to that of natural gas, and it can be used in an LPG cooking stove without producing any aldehydes. Its lower explosion limit is higher than that of propane, indicating higher safety than propane in case of leakage. Combustion tests on DME have been conducted with mass-produced household cooking stoves in Japan, and have passed the Japan Industrial Standard combustion test.[4] This study showed that stoves designed to burn city gas (85% methane, 15% propane) could burn DME well with only an adjustment of the variable air dampers.

Dimethyl Ether for Power Generation

Dimethyl ether is a promising new gas turbine fuel. Roughly speaking, 1 bcf/day of gas, converted into about 20,000 tpd of DME, can generate about 10,000 MW of power. Utilization of DME for power generation offers tremendous environmental benefits, in terms of CO_x, SO_x and NO_x emissions.[5] It burns in conventional gas turbines without modifications to the turbine or the combustors. Emissions produced by combustion of conventional fuels in gas turbines include nitrogen oxides, carbon monoxide, unburned hydrocarbons, and sulfur oxides. Dimethyl ether produces no sulfur oxide emission, as the fuel is sulfur free. It generates the least amount of NO_x, CO, and unburned hydrocarbons as compared with natural gas and distillate, and lower CO_2 emissions than the distillates.

BP has introduced a turbine-grade DME fuel that contains 88.0–89.8 wt% DME, 2.9–3.5% water, 7.0–8.0 wt% methanol, and 0.3–0.5 wt% of other oxygenates (with no metals or nitrogen, and with only a trace of sulfur). BP and GE power systems are currently implementing a project in India to use DME as a fuel for power generation.[6] Test results of the BP-introduced fuel at General Electric show that its emission properties and other key combustor operating parameters, including dynamic pressures and metal temperatures, are comparable to those of natural gas. The power generation efficiency (D) is expressed in terms of a heat rate number that corresponds to the amount of thermal energy needed to generate one unit of electrical energy. A lower heat rate number reflects higher power generation efficiency. The estimated performance of a nominal 700 MW combined cycle power plant based on the GE 9E machine indicates that the heat rate using refrigerated DME would be about 1.6% lower than that using natural gas, and about 6.3% lower than that using liquid naphtha. GE is prepared to pursue commercial offers, including standard commercial terms and guarantees, of DME-fired B/E class (diffusion and DLN burners) and F class (diffusion burners) heavy-duty gas turbines.

NKK Corporation has developed a process for DME production, which will greatly enhance the use of DME in power plants in Asia, especially in Japan, which typically use liquefied natural gas (LNG), LPG, fuel oil, and coal. Fuel oil and coal are subject to environmental concerns, most notably the emission of carbon dioxide and treatment or disposal of ash. A possible solution to circumvent this problem would be to generate DME at the mine, which can be shipped and used in a more energy-efficient and environmentally conscious manner. Liquefied natural gas and LPG are cleaner burning fuels, albeit at a much higher investment cost. Japan imports LNG, as fuel for electric power generation. Again, it would be economical to import DME produced from natural gas at medium- and small-scale gas fields considered too small for LNG development for thermal power generation. The supply source for LPG is limited mainly to the Middle East for countries in Southeast Asia and the Far East. Dimethyl ether can substitute LPG in power generation applications, just like in household applications.

Dimethyl Ether for Fuel Cells

Fuel cells can be powered by DME. Daimler Chrysler A. G. has studied the feasibility of using DME as a fuel in a polymer electrolyte membrane (PEM) fuel cell, in collaboration with Ballard Power Systems and University of Technology RWTH Aachen, Germany. Direct oxidation fuel cells, including the ones using methanol as fuel, are hindered by efficiency losses. The effect of methanol fuel crossover oxidation reaction at the cathode is the most significant efficiency loss, the others being use of parasitic fluid pumps, and mild toxicity of methanol vapor. An advantageous finding of the study is that DME is typically not oxidized at the cathode of the fuel cell. This minimizes the unwanted effects of fuel crossover, leading to improved fuel cell efficiencies when compared to direct methanol fuel cells, especially at low to medium current densities.

Researchers at the Electrochemical Engine Center at the Pennsylvania State University have identified additional advantages of using DME as a fuel in PEM fuel cells. It must be noted that DME molecules do not have a carbon–carbon double bond, enabling nearly complete oxidation in low-temperature PEM fuel cells. Also, DME can be stored in high-density liquid phase at modest pressures of around 5 atm, and delivered as a gas-phase fuel in a pumpless operation. Therefore, use of DME can potentially combine

the advantages of easy fuel delivery of pressurized hydrogen and the high energy density storage of liquid fuel.[7]

Hydrolyzing DME with steam can produce hydrogen, carbon oxides being the by-products. This method utilizes an essentially alkali-metal free catalytic composition of copper or nickel in elemental form, which catalyzes the hydrolysis reaction. A water–gas shift reaction converts carbon monoxide, which is usually present in the hydrolysis product, to relatively inert carbon dioxide. The hydrolysis and shift reactions take place in a single reaction zone, or alternatively, in separate reaction zones where reaction conditions can be individually optimized. When separate zones are provided, the heat is essentially transferred from the water–gas shift reaction zone to the hydrolysis reaction zone. The hydroshifted product stream can be utilized as fuel to power a turbine, and an integrated heat transfer scheme can be used to recover heat from the turbine exhaust stream. The reactions are carried out at temperatures between 150 °C and 800 °C, with better results obtained when operating in the range of 350–400 °C. Dimethyl ether conversion of 88.5% and hydrogen to carbon oxide ratio of 2.44 in the product stream were obtained by operating at 350°C and 2°/hr.[8]

Catalytic reaction of steam reforming of DME to hydrogen-rich gas was studied in a fixed-bed continuous-flow reactor at a temperature of 200–360°C under atmospheric pressure over a mechanical mixture of catalysts for DME hydration and for methanol steam reforming. Two catalysts, namely, 12-tungstosilicoheteropolyacid deposited on γ-Al_2O_3 and copper deposited on SiO_2, were used in the study to perform the reactions of DME hydration to methanol and methanol steam reforming, respectively. Mechanical mixtures of these two catalysts were used to perform the steam reforming of DME to produce hydrogen-rich syngas. Dimethyl ether conversion increased with increasing temperature, and so did the concentrations of H_2, CO, and CO_2. Dimethyl ether conversion of 100% and hydrogen outlet concentration of ~71°vol% were obtained at 290°C and GHSV of ~1200/hr.[9]

Dimethyl Ether as a Propellant

Dimethyl ether has been increasingly used as a propellant in aerosol formulations to replace chlorofluorocarbons, which are found to destroy the ozone layer of the atmosphere. Dimethyl ether is nontoxic and easily degrades in the troposphere. Although about 90% of the major current U.S. aerosol industry uses hydrocarbon-based propellants (mostly isobutane and propane), DME could become a more widely used propellant in the coming years. Several aerosol-based household products include colognes, hair sprays and dyes, personal care mousses, antiperspirants, and room air fresheners.

Current suppliers for the DME propellant market include DuPont, Akzo Nobel, and Mitsubishi Gas Chemicals. Demeon D, a DME-based product from Akzo Nobel, is used as a propellant in cosmetic formulations, foam blowing, and paint or other aerosol sprays.

Dymel® A, a product based on DME manufactured by DuPont, is a medium- to high-pressure propellant for general aerosol use, including personal products. It has extremely low toxicity, its lower explosive limit in air is higher than that of propane and isobutane, and it is a chemically stable compound. In aqueous solutions, the propellant is hydrolytically stable over a wide pH range. Dymel A is unique among aerosol propellants in that it has high solubility in both polar and nonpolar solvents. It is completely miscible with most organic solvents, and is by itself a very good solvent for many types of polymers, e.g., hair spray and paint resins. Dymel A has 35 wt% solubility in water, which facilitates formulation of single-phase products with large amounts of water, and is the only liquefied gas aerosol propellant to do so.

Dimethyl Ether as a Chemical Building Block

Dimethyl ether is an essential intermediate in the synthesis of hydrocarbons from coal or natural gas derived syngas. Dimethyl ether is a building block for the preparation of many important chemicals, including methyl sulfate.[10] Dimethyl sulfate is an important commercial commodity as a solvent and also as an electrolyte in high-energy-density batteries. Lower olefins like ethylene and propylene or downstream products, such as gasoline and range boiling hydrocarbons, are produced from syngas using DME as an intermediate.[11,12] A variety of specialty industrial chemicals such as oxygenates, acetaldehyde, acetic acid, ethylene glycol precursor like 1,2-dimethoxyethane, etc. can be formed using DME as a feedstock. Air Products has programs under way to use DME as a chemical building block for higher-molecular-weight oxygenated hydrocarbons.[13] Some of the chemicals that can be synthesized using DME as a building block are shown in Fig. 2.

DIMETHYL ETHER SYNTHESIS

Dimethyl ether can be produced from natural gas, biomass, or other carbon containing materials. Using existing supplies of natural gas combined with current technology, DME can be economically produced on a large scale via synthesis gas. Syngas, or synthesis gas, is

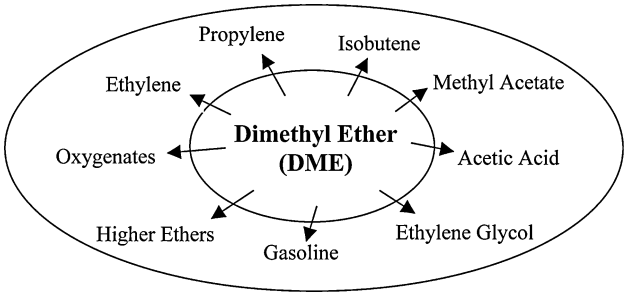


Fig. 2 Dimethyl ether—a building block for chemicals and oxygenates.

produced by coal gasification, biowaste gasification or by natural gas or hydrocarbon reforming. Dimethyl ether synthesis in the vapor phase suffers from low per-pass conversions, mandated in part by the debilitating effects of high temperature on the catalysts. Synthesis of DME in the gas phase has been studied in detail at Mobil Oil.[14,15] Gas-phase DME synthesis processes, in general, suffer from the drawbacks of low hydrogen and CO conversions per pass, along with low yield and selectivity of DME coupled with a high yield of carbon dioxide. These processes are typically expensive due to high capital costs for reactors and heat exchangers, and high operating costs due to inefficient CO utilization and high recycle rates.[13,16] Using an inert liquid as a heat sink for highly exothermic reactions offers a number of opportunities in syngas processing. Heat generated by the exothermic reactions is readily accommodated by the inert liquid medium. This enables the reaction to be run isothermally, minimizing catalyst deactivation, which is commonly associated with the more adiabatic gas-phase technologies.

The single-stage, liquid-phase DME synthesis process incorporates the sequential reaction of methanol synthesis and methanol dehydration in a slurry phase reactor system. Combining these reversible reactions in a single step drives each reaction thermodynamically by utilizing its inhibiting products as reactants in the subsequent reaction. Apart from the superior heat management allowed by the liquid-phase operation, the synergistic effect of these reactions occurring together yields higher quantities of DME than could be obtained from sequential processing.[13,16,17]

Commercial Processes

A number of processes have been developed to convert coal- or natural gas-based syngas into DME. The prominent ones include those by Haldor Topsoe, NKK Corporation, Air Products, Toyo Engineering Company, and the University of Akron, Electric Power Research Institute. A brief description of these processes is given below.

Haldor Topsoe Process

The chemical reactions involved in synthesis of DME from natural gas are as follows:

- Reforming:

$$CH_4 + 3/2O_2 = CO + 2H_2O$$
$$CH_4 + H_2O = 3H_2 + CO$$
$$CO + H_2O = CO_2 + H_2$$

- Dimethyl ether synthesis:

$$3H_2 + CO_2 = CH_3OH + H_2O$$
$$H_2O + CO = H_2 + CO_2$$
$$2CH_3OH = CH_3OCH_3 + H_2O$$

Haldor Topsoe has conducted a considerable amount of research for the purpose of developing DME as a diesel fuel from natural gas. They developed a new process, which is an integrated process for production of methanol from synthesis gas generated from various feedstocks ranging from natural gas to coal and biomass, followed by its subsequent conversion into DME in one single plant. The Haldor Topsoe process can produce neat DME or a raw fuel grade DME that can be specific blends of DME, methanol, and water. As shown in Fig. 3, the process consists of the following main steps: synthesis gas preparation, methanol and DME synthesis, and final purification unit. The synthesis gas preparation process uses autothermal reforming (ATR), consisting of a specifically designed burner (CTS burner). In this process, oxygen is added to desulfurize natural gas and steam. The steam to carbon ratio in the ATR is low (as low as 0.6) and the exit temperature is high. This ensures a favorable synthesis gas composition and low methane content. Synthesis of methanol and conversion of methanol into DME takes place in two separate reactors, which allows both parts of this sequential reaction to be carried out at optimal conditions. Methanol synthesis, which is a more exothermic reaction than DME synthesis, is carried out in a cooled reactor where reaction exotherm is continuously removed. Dimethyl ether synthesis takes place over a proprietary multiple-function methanol/DME catalyst in a loop comprising three adiabatic fixed-bed reactors in series. These have interstage cooling to achieve a high conversion of CO to CO_2. The product mixture of DME/methanol/water is then condensed and separated. The unconverted synthesis gas is split into a recycle stream and a purge stream, which is used as fuel and as hydrogen recycle. Dimethyl ether is purified by distillation in the final purification unit.[18,19]

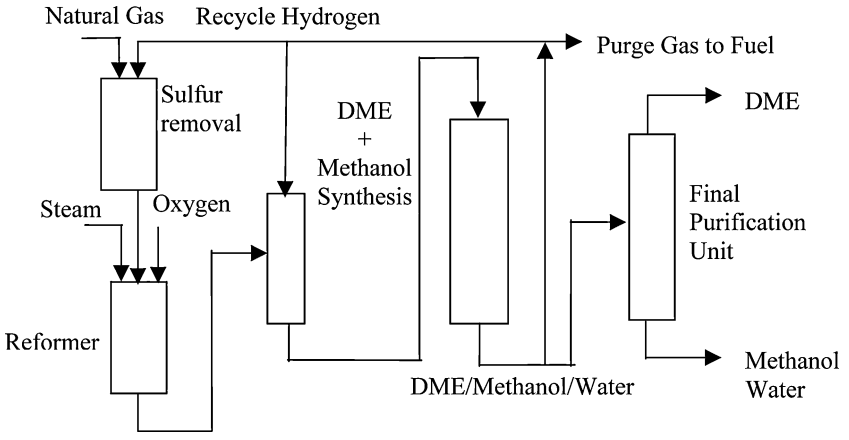


Fig. 3 Haldor Topsoe technology for DME synthesis.

NKK Process

Researchers at NKK Corporation have developed a process for DME synthesis from coal-bed methane using a slurry-bed reactor technology, utilizing a proprietary highly active catalyst for producing DME directly from the syngas at high yields. The success of this technology has been demonstrated in a 5 ton/day pilot plant, which is being operated since 1999.[20]

Chemical reactions involved in the synthesis of DME by NKK process are shown below

$$2CH_4 + O_2 + CO_2 = 3CO + 3H_2 + H_2O \quad (1)$$

Reforming reaction

$$2CO + 4H_2 = 2CH_3OH \quad (2)$$

Methanol synthesis

$$2CH_3OH = CH_3OCH_3 + H_2O \quad (3)$$

Methanol dehydration

$$3CO + 3H_2 = CH_3OCH_3 + CO_2 \quad (4)$$

DME synthesis

$$CO + H_2O = CO_2 + H_2 \quad (5)$$

Water–gas shift reaction

$$2CH_4 + O_2 = CH_3OCH_3 + H_2O \quad (6)$$

Overall process

Synthesis of DME from coal-bed methane consists of three reactions, namely, methane reforming, methanol synthesis, and methanol dehydration. Water produced by the methanol dehydration reaction participates in the water–gas shift reaction, which in turn produces hydrogen that can be utilized for methanol synthesis. In case the CO shift conversion reaction is slow, DME is synthesized by the methanol synthesis reaction and the methanol dehydration reaction.

The technology developed by NKK Corporation in Japan to produce DME from coal-bed methane is shown in Fig. 4. The process scheme consists of four sections: syngas reformer, carbon dioxide removal, DME synthesis, and DME separation/purification. Because the H_2/CO ratios of synthesis gas obtained by the coal gasification range from 0.5 to 1.0, the gas composition is adjusted by the shift reaction so that

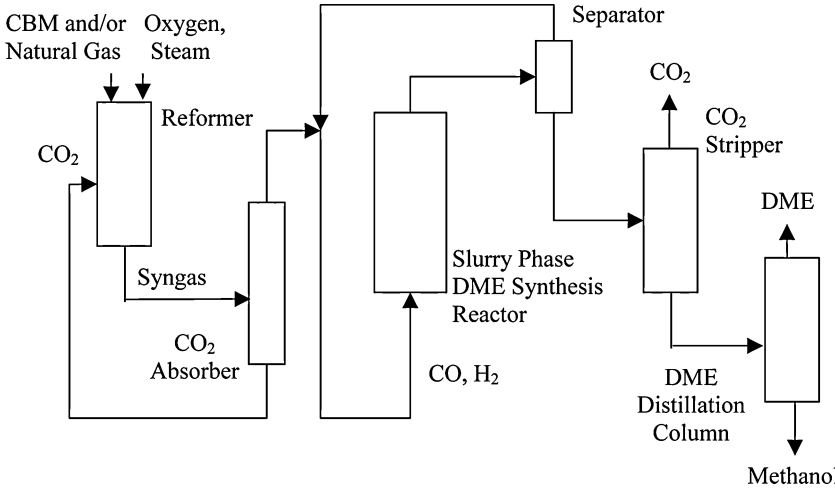


Fig. 4 NKK slurry phase DME synthesis.

$H_2/CO = 1$, and it is then supplied for DME synthesis. The effluent from the slurry reactor is cooled and chilled to separate the liquid phase containing DME, CO_2, and small amounts of methanol and water from the gaseous phase containing unreacted H_2 and CO. Most of the separated gas is recycled to the reactor. After CO_2 removal, the product DME of required purity is obtained by removing the impure water and methanol.

Air Products Process

Air Products has developed a single-step process for the direct synthesis of DME from synthesis gas with or without coproduct methanol in the liquid phase. This process can handle syngas with higher than 50 vol% CO, as is the syngas composition from advanced coal gasification. This process gives higher syngas conversion per pass than can be achieved by methanol synthesis alone. Selectivity of DME and methanol is optimized by varying reaction conditions, catalyst compositions, and/or ratios to suit specific process applications. Conversion of syngas to DME involves three separate reactions. All three reactions are equilibrium limited and exothermic in nature. Based on what they claim, the reaction chemistry is

$$xCO + xH_2O = xCO_2 + xH_2$$

Water–gas shift reaction

$$yCO + 2yH_2 = yCH_3OH$$

Methanol synthesis

$$2CH_3OH = CH_3OCH_3 + H_2O$$

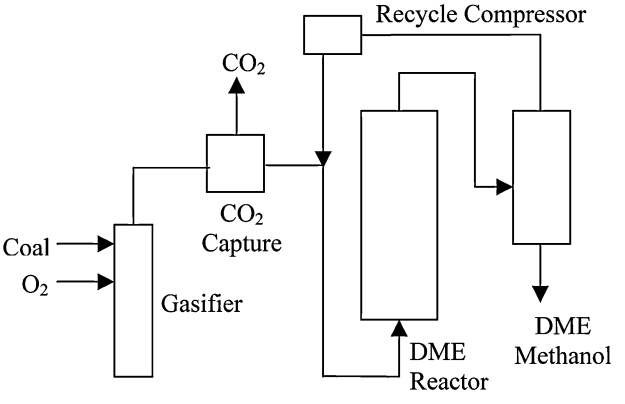
Methanol dehydration

$$(x + y)CO + (2y - x)H_2 = CH_3OCH_3 + CO_2 + (y - 2)CH_3OH + (1 - x)H_2O$$

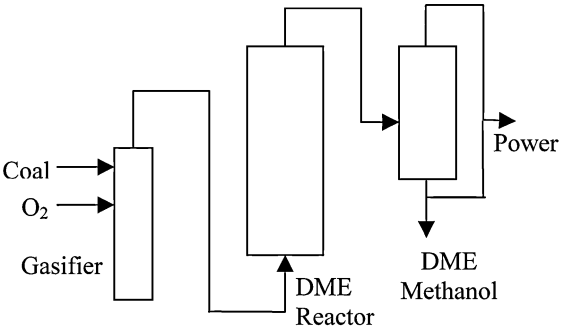
Net reaction

The single-stage synthesis of DME is carried out in a slurry-phase reactor, equipped with six-bladed draft-tube impellers. The catalysts used are powdered commercial catalysts BASF S3-85 and BASF S3-86 for methanol synthesis, γ-alumina catalysts for methanol dehydration, and copper-based BASF K3-110 as a shift catalyst. These catalysts are slurried in either degassed Witco-70 or DRAKEOL-10 mineral oils with slurry concentrations in the range of 15–30 wt% catalyst.

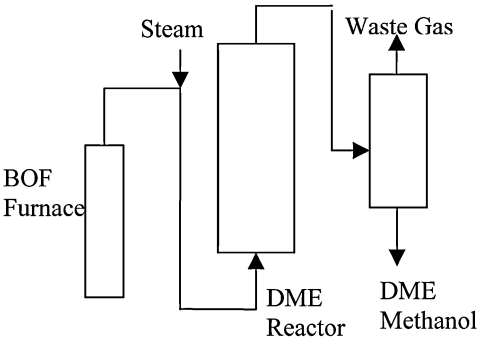
As shown in Fig. 5, three potential commercial modes of operation to produce DME have been investigated by Air Products. The first operating mode uses an oxygen-blown coal gasifier with recycle of the synthesis gas. The coproducts of this process mode are methanol and DME, which have applications in the fuel and petrochemical industry. Per-pass CO conversion was 70% at 1200 L/kg of catalyst per hour. The second operating mode uses the once-through approach of coal-derived syngas. Dimethyl ether and methanol are recovered, and unreacted gas is fired to a turbine. The third operating mode also uses the once-through process, but uses basic oxygen furnace (BOF) off-gas as the feed stream. The hydrogen to carbon monoxide ratio of BOF off-gas is essentially zero, thereby making it imperative for steam to be cofed to the reactor. This process enables the use of off-gases with unfavorable compositions, which are currently recovered only for their fuel value, and upgrading them into more valuable liquid products.[13,16]



Air Products DME Process: Coal Gasification Mode



Air Products DME Process: Once Through CGCC



Air Products DME Process: BOF Offgas Feed

Fig. 5 Air products LPDME process—commercial modes of operation.

Toyo Engineering Corporation Process

Toyo Engineering Corporation has developed a jumbo DME plant from natural gas using existing proven technologies.[21] Toyo Engineering Corporation has developed a number of large-scale methanol synthesis plants in the world. Chemical reactions involved in DME synthesis by TEC's process, as claimed, are shown below:

$$CO + 2H_2 = CH_3OH$$
Methanol synthesis

$$CO_2 + 3H_2 = CH_3OH + H_2O$$
Methanol synthesis

$$2CH_3OH = CH_3OCH_3 + H_2O$$
Methanol dehydration

$$CO + CO_2 + 5H_2 = CH_3OCH_3 + 2H_2O$$
Overall reaction

The configuration of a 7000 MTPD DME plant is based on the combination of methanol synthesis and methanol dehydration. Attractive features of this process include lesser total investment cost and lesser oxygen consumption when compared with the methanol/DME coproduction route or direct DME synthesis route. Also, carbon dioxide is not produced in the DME synthesis step of this process. As shown in Fig. 6, this process utilizes a steam reformer, TEC's TAF-X reactor, oxygen reformer, TEC's MRF-Z® methanol reactor, and TEC's DME reactor.

University of Akron, Electric Power Research Institute (UA-EPRI) Process

Researchers at the University of Akron, in conjunction with the Electric Power Research Institute, have developed a novel liquid-phase process that produces DME in a single stage from CO-rich syngas.[22,23] A process schematic of the DME minipilot plant is shown in Fig. 7. The process chemistry for the novel one-step

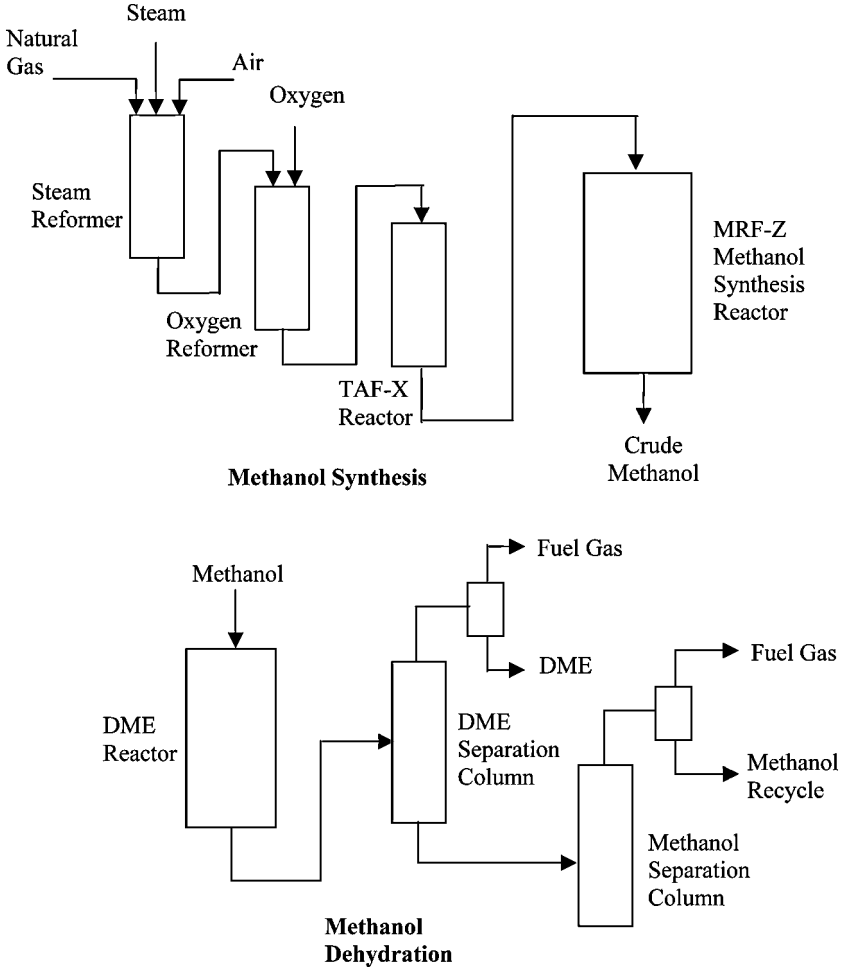


Fig. 6 Toyo Engineering Corporation DME process.

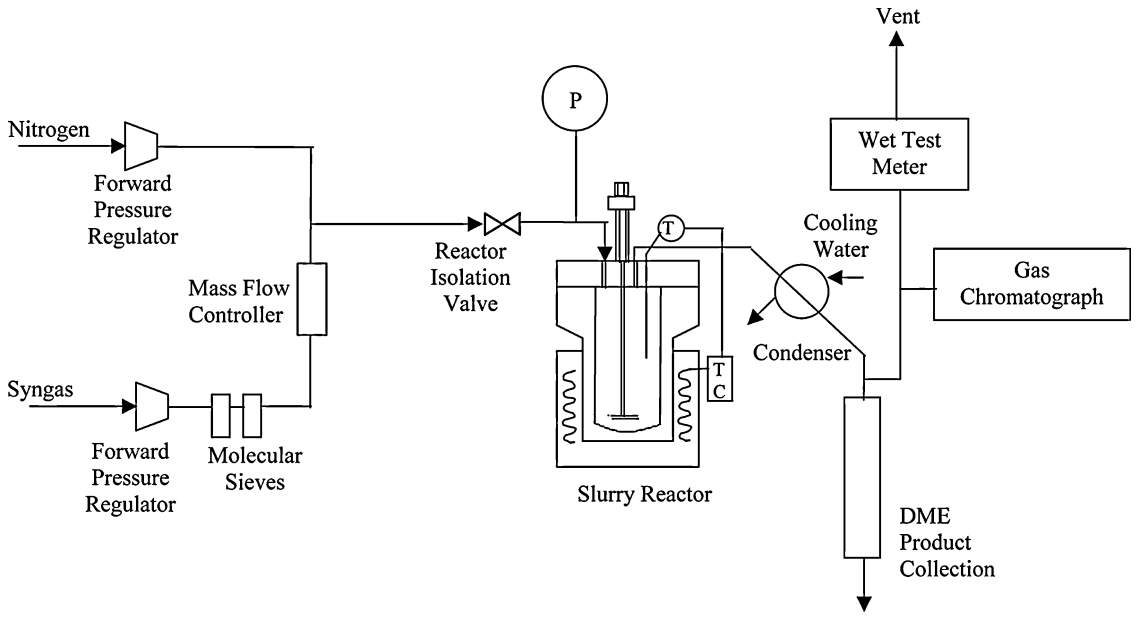


Fig. 7 Process schematic of the UA-EPRI liquid-phase DME experimental process unit.

synthesis of DME in the liquid phase is as follows

$$CO_2 + 3H_2 = CH_3OH + H_2O$$
Methanol synthesis

$$CO + H_2O = CO_2 + H_2$$
Water–gas shift reaction

$$2CH_3OH = CH_3OCH_3 + H_2O$$
Dimethyl ether synthesis

The methanol synthesis reaction and water–gas shift reaction take place over the coprecipitated $Cu/ZnO/Al_2O_3$ catalyst, and the methanol dehydration reaction takes place over a γ-alumina catalyst. The reactions are carried out at 250°C and 70 atm in a liquid phase involving inert oil, such as Witco-40, Witco-70, or Freezene-100 oil.

The single-stage reactor productivity was increased by as much as 60% when the coproduction of methanol and DME was used, as compared to methanol synthesis alone. Per-pass syngas conversion when using the coproduction of methanol and DME approach was increased by as much as 50% over that of methanol synthesis only. At high slurry ratios, when methanol concentration in the liquid phase is very high, the percent increase in single-stage reactor productivity could be as high as 80%. The above fact is very significant from a commercial perspective as almost all the commercial reactors in the liquid-phase synthesis are operated in the gas-to-liquid mass transfer limited region. The process is capable of coproducing methanol and DME in any fixed proportion, at significant synthesis rates of both methanol and DME. The process feasibility has been well demonstrated on their minipilot plant system.[17,22]

CONCLUSIONS

Dimethyl ether is a versatile multisource, clean-burning fuel. It can be readily produced from natural gas and coal, from renewable resources, such as biomass and wood, as well as from waste matter. Production of DME by combining methanol synthesis, methanol dehydration, and water–gas shift reactions in a single stage results in a synergistic effect with favorable thermodynamics, thereby creating a driving force for the overall reaction with very high syngas conversion in a single pass. Dimethyl ether can be used as a clean-burning fuel in diesel engines and fuel cells, as a household fuel (LPG alternative) for heating and cooking, as a fuel for gas turbines in power generation, as a feedstock for producing chemicals and oxygenates, and as a propellant in the aerosols market. Use of DME as a transportation fuel results in lower NO_x, CO_x, and hydrocarbon emissions, and also eliminates soot particulates and evaporative emissions. Acceptance of DME as a clean fuel in terms of public perception is facilitated by the fact that it is nontoxic and noncorrosive and it can be transported, stored, and dispensed as a liquid fuel in the same manner as LPG or propane. In the short term, DME can be easily manufactured from methanol. Meanwhile, commercial production plants are being designed and set up in different parts of Asia, Europe, South America, and the

United States to produce DME directly from natural gas and coal-based syngas in the near future.

REFERENCES

1. Ohno, Y.; Shikada, T.; Ogawa, T.; Ono, M.; Mizuguchi, M. New clean fuel from coal—dimethyl ether. Clean Fuels Symposium, American Chemical Society (ACS) Meeting, San Francisco, CA, Apr 1997.
2. Fleisch, T. Move on dimethyl ether: case is building for DME as clean diesel fuel. Diesel Prog. Engines Drives **1995**, *1*, 42–45.
3. Rouhi, A.M. Amoco Haldor Topsoe develop dimethyl ether as alternative diesel fuel. Chem. Eng. News **1995**, *26*, 37–39.
4. Kato, Y. Trends in energy and environment in the Asian region and feasibility of new clean fuels. Presentation to Asian Economy Joint Study Association, Nov 5, 1999.
5. Fleisch, T. GTL perspectives. In 21st World Gas Conference, Nice, France, 2000.
6. Basu, A.; Wainwright, J. DME as a power generation fuel: performance in gas turbines. Petrotech-2001 Conference, New Delhi, India, Jan 2001.
7. Muller, J.; Urban, P.; Holderich, W.; Colbow, K.; Zhang, J.; Wilkinson, D. Electro-oxidation of dimethyl ether in a polymer-electrolyte-membrane fuel cell. J. Electrochem. Soc. **2000**, *147* (11), 4058–4060.
8. Bhattacharyya, A.; Basu, A. Process for Hydroshifting Dimethyl Ether U.S. Patent 5,626,794, 1997.
9. Galvita, V. Production of hydrogen from dimethyl ether. Appl. Catal. A Gen. **2001**, 85–90.
10. Mills, G. Conversion of synthesis gas to liquid fuels. Fuel **1994**.
11. Sardesai, A.; Tartamella, T.; Lee, S. Synthesis of hydrocarbons from dimethyl ether: selectivities toward light hydrocarbons. Fuel Sci. Technol. Int. **1996**, *14* (5), 703–712.
12. Lee, S. US. Patent, 5,459,166, 1995.
13. Brown, D.M.; Bhatt, B.L.; Hsuing, T.H.; Lewnard, J.J.; Waller, F.J. Novel technology for the synthesis of dimethyl ether from syngas. Catal. Today **1991**, *8*, 279–304.
14. Bell, W.; Chang, C.; Shinnar, R. Method for Generating Power upon Demand U.S. Patent 4,341,069, 1982.
15. Zahner, J. Conversion of Modified Synthesis Gas to Oxygenated Organic Chemicals U.S. Patent, 4,011,275, 1977.
16. Lewnard, J.; Hsuing, T.; White, J.; Brown, D. Single-step synthesis of dimethyl ether in a slurry reactor. Chem. Eng. Sci. **1990**, *45* (8), ISCRE 11.
17. Lee, S. *Dimethyl Ether Synthesis Process*; Final Report, Project 317-6, EPRI TR-100246; Electric Power Research Institute: Palo Alto, CA, Feb 1992.
18. Wilson, R. DME shows promise as drop-in replacement for diesel fuel. Diesel Prog. Engines Drives **1995**, *1*, 108–109.
19. Jones, G., Jr.; Holm-Larsen, H.; Romani, D.; Sills, R. How is DME manufactured in large-scale plants. PETROTECH-2001, New Delhi, India, Jan 2001.
20. Ohno, Y.; Inoue, N.; Ogawa, T.; Ono, M.; Shikada, T.; Hayashi, H. Slurry phase synthesis and utilization of dimethyl ether. NKK Tech. Rev. **2001**, *85*.
21. Mii, T.; Hirotani, K. Economic evaluation of a jumbo DME plant. WPC Asia Regional Meeting, Shanghai, China, Sep 17–20, 2001.
22. Lee, S.; Gogate, M.; Kulik, C. A novel single-step dimethyl ether (DME) synthesis in a three-phase slurry reactor from CO-rich syngas. Chem. Eng. Sci. **1992**, *47* (13/14), 3769–3776.
23. Keil, F. Methanol-to-hydrocarbons: process technology. Microporous Mesoporous Mater. **1999**, *29*, 49–66.
24. Phillips, J.; Reader, G. The use of dimethyl ether as a transportation fuel—a canadian perspective. ASME Fall Technical Conference, Paper No. 98-ICE-152, ICE-31-3, 1998; 65–71.

Dimethylcarbonate

Byung Gwon Lee
Environment and Process Technology Division, Korea Institute of Science and Technology, Cheongryang, Seoul, Korea

INTRODUCTION

Dimethylcarbonate [DMC, $(CH_3O)_2CO$] is a versatile chemical substance, which has been used mainly as a carbonylating, methylating, or methoxycarbonylating agent. It can be effectively used as an environmentally benign substitute for phosgene in polycarbonates and isocyanates synthesis, and for dimethyl sulfate in various methylation reactions.[1,2]

Dimethylcarbonate has two carbon centers at which a nucleophile may react: the carbonyl and the methyl group. When a nucleophile attacks the carbonyl carbon of DMC, the cleavage of acyl–oxygen bond results in a methoxycarbonyl product. Therefore, DMC can replace phosgene as a safe source for carbonic acid derivatives. On the other hand, when a nucleophile attacks at the methyl carbon of DMC, the methylation product is produced by the alkyl–oxygen bond cleavage. As a methylating agent, DMC is a safe substitute for dimethyl sulfate.[2]

Phosgene and dimethyl sulfate are highly reactive chemicals, used since the beginning of the chemical industry. The cost of their utilization, however, is dramatically increasing owing to the growing safety concerns and measures during their production, transportation, storage, and use. Their use in carbonylation and methylation reactions brings about the formation of stoichiometric amounts of by-products such as inorganic chlorides of sulfates, generally in the form of organically contaminated aqueous streams. Dimethylcarbonate does not have this kind of problem at all in its various applications. In addition, DMC is being considered as a component of reformulated fuels owing to its high oxygen content and good blending properties.[3] Dimethylcarbonate has about three times higher oxygen content than methyl *tert*-butyl ether (MTBE) and its synthesis is not dependent on isobutylene feed like MTBE.

PROPERTIES

Dimethylcarbonate is a transparent liquid with a melting and boiling point of 277 and 363 K, respectively, and the specific gravity is close to unity. Thus, in a sense, the physical properties are close to those of water. Dimethylcarbonate is sparingly soluble in water and miscible with the most organic solvent, particularly polar solvents such as esters, ketones, ethers, and alcohols. Dimethylcarbonate is extremely stable to pyrolysis up to at least 623 K. Dimethylcarbonate undergoes base- or acid-catalyzed hydrolysis to produce methanol and carbon dioxide.[1] One of the most important properties of DMC is low toxicity compared with conventional methylating agents or carbonylation agents. Table 1 summarizes the physical and toxicological properties of DMC.

Some data are available in the open literature on the hazards and toxicity of DMC.[3,4] These toxicology studies performed by IITRI (Illinois Institute of The Technology Research Institute) involved eye and skin irritation studies and dermal sensitization studies, along with dermal, oral, and inhalation studies for acute toxicity. The results indicate that DMC is a nonirritant, nontoxic chemical. Dimethylcarbonate is listed in the "Hazard Rating 3"(HR3) category. Hazard rating (HR) is a number between 1 and 3. Ratings are assigned on the basis of low [1], medium [2], or high [3] toxic, fire, explosive, or reactivity hazard. HR3 means the substance has an LD_{50} (lethal dose, 50% of population) below 400 mg/kg and an LC_{50} (lethal concentration, 50% of population) below 100 ppm, or that the material is explosive, spontaneously flammable, or highly reactive. Hazard Rating 2 (HR2) means the substance has an LD_{50} between 400 and 4000 mg/kg, and an LC_{50} between 100 and 500 ppm, or that the material is either highly flammable or reactive. Hazard Rating 1 (HR1) means the substance has an LD_{50} between 4000 and 40,000 mg/kg and an LC_{50} between 500 and 4000 ppm, or that the material is combustible. Hexane and cyclohexane fall into the HR3 category, the same as DMC.[3] In addition, recently it has been found that the photochemical ozone creation potential (POCP) of DMC is the lowest among common volatile organic compounds (2.5, ethylene = 100).[4]

MANUFACTURE

The traditional synthesis of DMC used to require toxic, hazardous phosgene as a reagent. Nonphosgene alternative routes for DMC production basically have

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120006802

Table 1 Physical and toxicological properties of dimethylcarbonate

Physical properties			
Chemical formula	$(CH_3O)_2CO$		
Name	Dimethyl carbonate, carbonic diester, or carbonic acid ester		
Color, odor	No color, no odor		
Molecular weight	90.08		
Boiling point (°C)	90.3		
Melting point (°C)	4.0		
Density (d_4^{20})	1.07		
Viscosity (cP at 20°C)	0.625		
Solubility	13.9 g DMC/100 g H_2O, 4.2 g H_2O/100 g DMC		
Toxicological properties			
	DMC	Phosgene	Dimethyl sulfate
LD_{50} (oral)	13.8 g/kg (rat)	—	0.20 mg/kg (rat)
LD_{50} (inhalation)	140 mg/L (4 hr, rat)	0.02 mg/L (30 min, rat)	0.05 mg/L (10 min, human)

relied on the reaction of methanol with carbon monoxide (oxidative carbonylation) or with carbon dioxide (direct carboxylation with CO_2, or indirect carboxylation, using urea or alkene carbonates as CO_2 carriers), as depicted in Fig. 1.

Phosgene Alcoholysis

The conventional method for synthesizing DMC employs phosgene.[3] Phosgene will react with methanol to form methyl chloroformate.

$$CH_3OH + COCl_2 \rightarrow CH_3COCl + HCl \quad (1a)$$

Methyl chloroformate reacts with an additional mole of methanol to form DMC.

$$CH_3COCl + CH_3OH \rightarrow CH_3OCOOCH_3 + HCl \quad (1b)$$

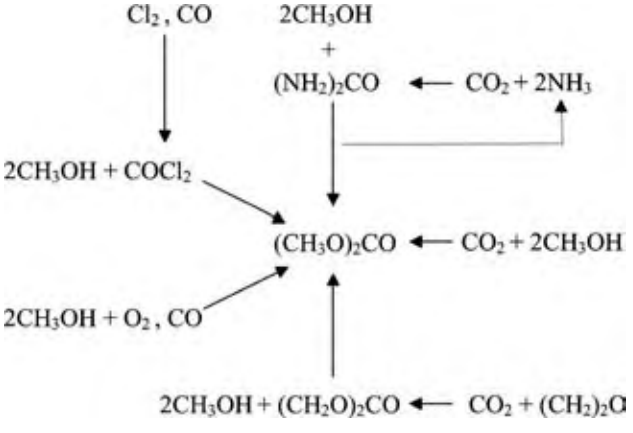


Fig. 1 Dimethylcarbonate synthesis routes.

Reaction (1a) occurs quite easily, whereas reaction (1b) is slower and can be accelerated using an acid scavenger such as a tertiary amine or an inorganic base, e.g., NaOH.[5] While making the carbonate, the amount of alcohol and the reactor residence time are increased. The continuous production of methyl carbonate has been reported in a 1973 French patent.[3] Here, methyl chloroformate and methanol are fed continuously into a packed-bed reactor operating with a bottom to top temperature gradient of 72–127°C. HCl is drawn off the head and dimethylcarbonate (99%) is withdrawn from the base.

Oxidative Carbonylation of Methanol

Liquid-phase synthesis

The methods of DMC preparation that are based on the catalytic reaction of methanol with carbon monoxide and oxygen, according to reaction (2), have been the subject of intensive studies.

$$2CH_3OH + CO + 1/2O_2 \rightarrow (CH_3O)_2CO + H_2O \quad (2)$$

EniChem set up a project aimed at the development of a nonphosgene synthesis of DMC for large volume usage; as a result, a new industrial process was established, based on a liquid-phase methanol oxidative carbonylation in the presence of copper chlorides as catalysts.[6] This catalytic system was highly effective in DMC production: the reaction was carried out by feeding at the same time methanol, carbon monoxide, and oxygen to the suspension of the catalyst in a mixture of water, methanol, and DMC and recovering DMC by distillation after the catalyst separation. Besides, the process

was not influenced by the carbon monoxide purity: diluents such as hydrogen did not interfere with the catalytic system. The first industrial plant, based on the developed technology, went onstream since 1983.

Modification of the copper chloride catalyst has been sought by different research groups in the effort to improve the catalyst performance; examples are: 1) the use of a CuCl/KCl eutectic mixture as catalyst combined with the pressure swing technique; and 2) the introduction of high-boiling cosolvents/ligands in the reaction mixture.[6]

Methanol oxy-carbonylation is a redox two-step reaction. According to a simplified scheme, the reaction proceeds through CuCl oxidation by oxygen to cupric methoxychloride [reaction (3)].[6]

$$2CuCl + 2CH_3OH + 1/2O_2 \rightarrow 2Cu(OCH_3)Cl + H_2O \quad (3)$$

The latter is reduced by carbon monoxide to DMC [reaction (4)], restoring CuCl and allowing the setup of the catalytic cycle.

$$2Cu(OCH_3)Cl + CO \rightarrow (CH_3O)_2CO + 2CuCl \quad (4)$$

This scheme does not appear fully adequate because it was observed that the reduction of cupric methoxychloride under carbon monoxide hardly proceeds under strictly anhydrous conditions, but it is promoted by the addition of even small amounts of soluble cupric species, such as $CuCl_2$. The same effect is obtained by the addition of some water, bringing Cu(II) in solution, by hydrolysis. Moreover, diethylcarbonate is produced when $Cu(OCH_3)Cl$ is reacted in ethanol under CO. Therefore, polymeric, insoluble cupric methoxychloride cannot be considered the true reaction intermediate. As a matter of fact, a methanol dispersion of $Cu(OCH_3)Cl$ synthesized by CuCl oxidation with oxygen in methanol presents an electron spin resonance spectrum with a resolved signal typical of a cupric ion in solution over the enlarged signal particular to the insoluble, polymeric species.

In systems containing high water concentrations, as generated by direct oxygen reoxidation under technical conditions, exhaustive hydrolysis of copper methoxychloride leads to a number of soluble and insoluble cupric species containing chloride and hydroxide anions such as $CuCl_2$ and $Cu_4Cl_2(OH)_6$ (atacamite). This reaction system brings about, by reduction under CO, DMC and cuprous chloride formation, provided that copper and chloride are balanced. Halide anions appear to be very important to establish the catalytic cycle.

In the presence of amine ligands, such as pyridine, dipyridyl, and tetramethylethylene diamine, the catalytic effectiveness increases, but the practical use of this system is hampered because it fails to produce DMC under technical conditions, mainly owing to water sensitivity. In the amine–copper system, catalysis of DMC formation by Cu(I), through the carbonyl complex CuCl(CO), was observed, this feature being reminiscent of the mechanism of CO to CO_2 oxidation by Cu(II) in aqueous solutions.[6]

Other redox metallic systems are able to perform the alcohol oxy-carbonylation to carbonates. Palladium(II) chloride complexes stoichiometrically give carbonates in the presence of a base such as Na_2CO_3 or $N(Et_3)$ under very mild conditions.[7] In the methanol oxy-carbonylation with palladium(II) acetate complexes, dimethyl oxalate is produced along with DMC.[8] The selectivity toward DMC could be driven by the choice of the phosphine ligand, by increasing the carbon monoxide pressure and by the addition of a base. A lot of effort has been put in with the aim of making the oxy-carbonylation catalytic with respect to palladium, by addition of some cocatalyst for palladium(0) reoxidation. When using air or oxygen as oxidants, water is coproduced. It is known that the system Pd(II)–CO in the presence of water easily brings about carbon dioxide formation. It was reported that the addition of phosphonium salts or nitriles as cosolvents/ligands to the reaction mixture improves carbon monoxide selectivity to DMC, ranging from 50% to 80%, according to the reaction conditions.[6] Besides, by addition of alkali earth metal salts and 2-hydroxypyridine to the palladium–copper salts system, the CO selectivity toward DMC could be increased up to 90%. To prevent the formation of water in the reaction medium, use of dialkyl peroxides as oxidants, in the presence of a copper–pyridine complex as cocatalyst, or an electrochemical oxidation has been proposed.[9]

Cobalt(II) complexes bearing ligands with nitrogen and oxygen donors, such as Schiff bases, acetylacetonate, or picolinate, produce DMC with high selectivity in methanol under carbon monoxide and oxygen, especially in the presence of weakly coordinating solvents such as nitriles, amides, or ureas. These cobalt-based systems appear very attractive because they are halogen-free and noncorrosive compared with systems containing chloride, especially copper chlorides. In contrast to palladium and copper, simple cobalt salts, such as chlorides, are inactive as catalysts.[6]

Gas-phase synthesis

A lot of effort has been put in to develop a gas-phase oxy-carbonylation process. First, EniChem described gas-phase DMC production using CuCl dispersed on alumina.[6] Dow Chemical studied $CuCl_2$, alone or in combination with alkali and alkali earth halides, and the pyridine complex of $Cu(OCH_3)Cl$, both supported

on active carbon, as catalysts. The catalyst activity was not high and decreased during the tests in a relatively short time (20–40 hr), depending on the catalyst loading; it was restored by in situ regeneration of the catalyst by addition of HCl. It was reported that both high selectivity and DMC productivity are obtained when using a $CuCl_2$ catalyst supported on active carbon and modified by treatment with alkaline hydroxides; a stable catalytic activity was claimed after 100 hr. Recently, a study of the structural change in a $CuCl_2$ active carbon supported catalyst used in the DMC synthesis was reported.[10]

Cuprous chloride supported on a Y-type zeolite by high-temperature anhydrous reaction gave a DMC productivity as good as the active carbon supported catalyst.[11] Addition of tetraalkyl ammonium chlorides to copper chlorides on zeolites or alumina increased the DMC productivity and the catalyst life. In any case, active carbons are the most preferred supports for this reaction, compared to zeolites or alumina, because of higher DMC productivity when working with copper(II) chlorides.

The DMC productivity increases by addition of palladium salts to copper salts supported on active carbon. It is worth pointing out that, although copper salts without halides are practically inactive in the reaction, the use of halogen-free palladium/copper salts supported on active carbon reportedly gives high DMC productivity and CO selectivity.[6]

When palladium salts are used for methanol oxy-carbonylation to DMC, reaction conditions are milder than using copper only; however, methanol and CO selectivities are lower owing to the formation of DMO as a by-product and to the higher ratio between CO_2 and DMC production rates. Despite the large amount of work on the catalytic systems, no process based on gas-phase direct methanol oxy-carbonylation to DMC has been established.

Carbonylation of Methylnitrite

Recently, two reviews on the synthesis of DMC via methylnitrite (MN) appeared.[12] Starting from methanol, carbon monoxide, and oxygen as raw materials, the process takes place in two steps in the gas phase: in the first, methanol reacts at about 50°C with nitrogen oxides and oxygen to give MN and water, without any catalyst, according to reactions (5) and (6), which involve N_2O_3 as an intermediate species.

$$2NO + 1/2O_2 \rightarrow N_2O_3 \quad (5)$$

$$2CH_3OH + N_2O_3 \rightarrow 2CH_3ONO + H_2O \quad (6)$$

In the second, MN reacts with carbon monoxide to produce DMC at 100–120°C and 0.5–1 MPa, in the presence of a palladium supported catalyst, in a fixed-bed reactor according to reaction (7).

$$2CH_3ONO + CO \rightarrow (CH_3O)_2CO + 2NO \quad (7)$$

In the catalytic process, the NO produced in the latter reaction is converted back to MN according to reactions (5) and (6).

The above DMC synthesis using MN was discovered at UBE, and a plant based on this technology has recently gone onstream with a capacity of about 6000 ton/yr.[13] Reaction (7) is catalyzed by supported palladium(II) halide complexes, which allow high (90–95%) CO selectivity. The addition of a cocatalyst such as copper chloride is required to prevent the reduction of Pd(II) to Pd(0) because Pd(0) tends to accelerate the formation of DMO.

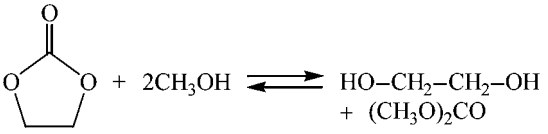
Many different catalytic systems based on palladium with various cocatalysts and supports were studied.[6] The nature of the support influences the catalyst performance; it is important to avoid the presence of both acid sites, to avoid MN decomposition to methylformate and methylal, and basic sites that increase the DMO production. The catalyst must contain chloride anions to be effective and to keep the oxidation state of palladium(II). Therefore, the loss of chlorides in the outlet stream of the reactor as organic or inorganic compounds must be counteracted by addition of chlorinated organic compounds or HCl in the feed, to maintain the catalyst performances.

Surprisingly, palladium(II) salts supported on NaY zeolite produce DMC, even without halogens.[6] The preferred support seems to be active carbon compared to zeolites because of higher DMC selectivities based on both MN and carbon monoxide, >95% in the case of active carbon and 80–90% in the case of zeolites. Palladium chloride/copper chloride on active carbon is likely used as a catalytic system in the industrial process. Because the carbonylation of MN to DMC occurs without water coproduction, the use of palladium salts as catalysts does not adversely affect selectivity. In the carbonylation reactor outlet some amount of methylchloroformate is present, as expected because it is known that palladium(II) chloride supported on alumina or silica catalyzes the reaction between MN, CO, and HCl to give methylchloroformate.[14] The presence of halide ions in the catalytic system and the methylchloroformate generation likely raise some corrosiveness issues.

The methyl nitrite process, studied both at UBE and at Bayer, is a new and valuable process for the DMC synthesis; nevertheless, there are some concerns about the toxicity and handling of MN.[6]

Transesterification of Ethylene Carbonate or Urea

An alternative to the oxy-carbonylation processes is the transesterification of ethylenecarbonate (EC) with methanol. In this process, DMC is cogenerated with ethylene glycol (EG), according to reaction (8).

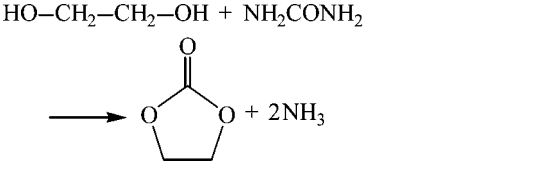


(8)

Ethylene carbonate can be prepared by a well-known process, from ethylene oxide and carbon dioxide with a catalyst such as a quaternary ammonium halide at 150–175°C. Addition of a cocatalyst such as zinc chloride to a quaternary ammonium iodide allows milder reaction conditions (50–100°C).[6]

Reaction (8) takes place in the presence of a catalyst at about 100–150°C at moderate pressure, for example, by working in a homogenous phase in the presence of tin, zirconium, or titanium complexes. Both homogenous and heterogenous basic or acid catalysts can also be used for the reaction; however, the base catalyzed reaction appears to be the most effective for the synthesis of DMC.[15] Recently, a new heterogenous catalyst based on potassium carbonate treated titanium silicate (TS-1) molecular sieve was reported.[16] Unfortunately, the transesterification reaction of EC with methanol is an equilibrium reaction and the formation of DMC is thermodynamically not favored. Several methods have been suggested to improve the low yield of DMC, for example, by removal of the reaction product as DMC/CH_3OH azeotrope from the reaction mixture by distillation or by selective solvent extraction of the DMC produced.[6] In this reaction, EC and methanol conversion are typically in the range 50–60% and 15–20%, respectively, whereas the DMC selectivity, based on both converted EC and methanol, is about 98%.

This process also suffers from the complications associated with the cogeneration of two compounds. This problem could be overcome by recycling ethylene glycol to produce EC, for example, via urea transesterification according to reaction (9).



(9)

The driving force for reaction (9) is ammonia removal. By using an adequate catalyst such as zinc oxide, 95–98% EC selectivities based on ethylene glycol and urea have been reported, at total urea conversion, when using a small excess of ethylene glycol.[6]

Although the direct synthesis of DMC starting from urea and methanol would be very attractive, its thermodynamics is not favorable, the calculated ΔG° for the direct transesterification reaction being about $\cong$3 kcal at 100°C.[3] The reaction involves two steps: first, urea is converted to methylcarbamate by reaction with methanol at relatively low temperatures (ca. 100°C in the presence of a catalyst or ca. 150°C without catalyst); then, the carbamate is further reacted with methanol at 180–190°C in the presence of a catalyst to produce DMC, according to the reactions (10a) and (10b).

$$NH_2CONH_2 + CH_3OH \rightarrow NH_2COOCH_3 + NH_3 \quad (10a)$$

$$NH_2COOCH_3 + CH_3OH \rightarrow (CH_3O)_2CO + NH_3 \quad (10b)$$

As above, the driving force for both reactions is ammonia removal. Both reactions (10a) and (10b) are catalyzed by a combination of a weak Lewis acid and a weak Lewis base, such as equimolar amounts of Al(*i*-Bu)$_2$H and PPh_3. These bifunctional catalysts reduce the formation of by-products from carbamate decomposition. The above reactions could be carried out in two separate steps or in one pot when in the presence of tin(IV) compounds as catalysts for the second reaction.[6]

In a recent development of this route, the reactions take place at the same time in a reactor with a distillation column at about 180°C and 0.5 MPa, in the presence of a tin(IV) alkoxide in a high-boiling solvent such as triglyme, feeding methanol and urea in the reactor containing the catalyst and the solvent and removing DMC and methanol overhead.[17] Dimethylcarbonate selectivity based on methanol is excellent (97–98%), and the DMC yield is higher than previously reported. However, the urea selectivity and conversion are more difficult to estimate. Some products of urea alkylation are produced in 1–3% yields, whereas the formation of urea decomposition products (such as biuret $NH_2CONHCONH_2$ and other high molecular weight by-products) cannot be ruled out because their presence was not checked.

Synthesis of DMC from urea is attractive. In fact, by this process the synthesis of a carbonate starting from an alcohol and carbon dioxide would be achieved, because in principle the evolved ammonia can be recycled to the synthesis of urea.

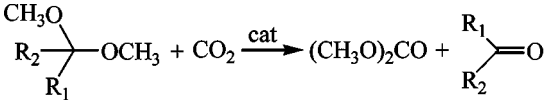
Direct Synthesis of DMC Starting from CO_2

Conversion of carbon dioxide to useful industrial compounds has recently raised much interest in view

of the so-called ''Sustainable Society'' and ''Green Chemistry.'' The direct synthesis of carbonates starting from alcohols and carbon dioxide, studied since the 1980s, fulfils this approach.[6] In particular, the catalytic DMC synthesis starting from carbon dioxide and methanol is currently studied intensively. The reaction occurs at 140–190°C and 5 MPa in the presence of zirconia (ZrO_2) with a very high (ca. 100%) methanol selectivity to DMC.[10] Unfortunately, the methanol conversion is very low because the equilibrium of the reaction (11) is largely shifted to the left as shown:

$$CO_2 + 2CH_3OH \rightarrow (CH_3O)_2CO + H_2O \qquad (11)$$

To avoid the formation of water as by-product, the reaction could be carried out using dehydrated derivatives of methanol, such as acetals [reaction (12)]. The carbonyl compound, formed along with DMC, can be recycled to prepare the corresponding acetal:



(12)

This reaction occurs at 190°C under supercritical CO_2 (30 MPa) via the carbomethoxy methoxy bridged dialkyl tin dimer intermediate. Yields, based on this tin intermediate, are ca. 60%. Dimethylcarbonate appeared to be formed by an intramolecular pathway. Further efforts in this very promising area are necessary to discover more efficient catalytic systems.

Other Routes to DMC

A few other routes to DMC have been studied. Alkylation of metal carbonates with organic halides could be a method to prepare complex, functionalized carbonates under mild conditions, but it is unattractive for the synthesis of DMC from many points of view, both economic and ecological.[6]

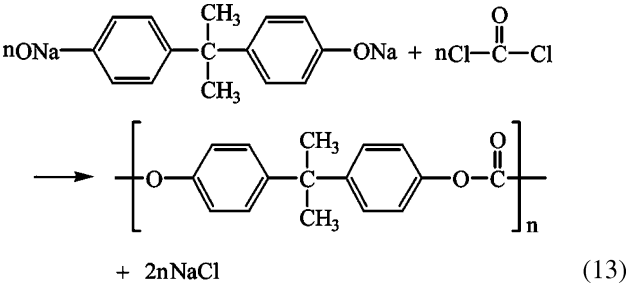
Dimethylcarbonate was prepared by decarbonylation of DMO in the presence of sodium methoxide or tetraphenyl phosphonium chloride as catalyst at 100°C without solvent. Under these conditions, DMO is totally converted and the selectivity to DMC ranges between 85% and 95%.[18] The method could be also used to prepare DMC starting from the corresponding oxalate.

APPLICATIONS

Dimethylcarbonate characteristics, coupled to its nontoxicity, are very attractive for the chemical industry. Dimethylcarbonate applications are sorted according to its use as a chemical intermediate (carbonylating or methylating reagent), solvent, and fuel additive.

Polycarbonate Production

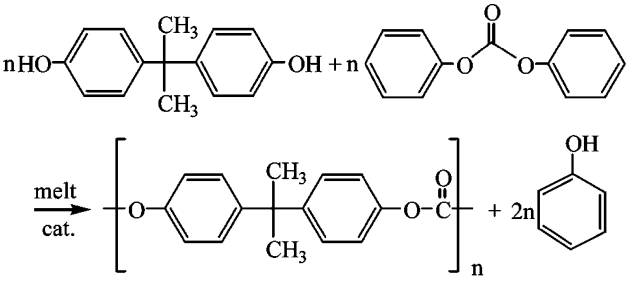
Polycarbonate resins are important engineering thermoplastics with good mechanical and optical properties as well as electrical and heat resistance, useful for many engineering applications. Polycarbonates have been commercially produced by the interfacial polycondensation between a bisphenol-A salt in an aqueous caustic solution and phosgene in an organic solution as follows [reaction (13)]:



(13)

The main disadvantages of this phosgene process are: 1) The high toxicity and corrosiveness of phosgene; 2) The use of copious amounts of methylene chloride solvent (10 times the weight of the product); This solvent is water soluble, so it contaminates the wash water; and 3) The complex cleanup to remove ionic materials.

The most practical nonphosgene process for manufacturing polycarbonates is the transesterification of diphenylcarbonate (DPC) with bisphenol-A. A nonphosgene process for the melt polymerization production of aromatic polycarbonates [reaction (14)], making use of EniChem technologies for the production of DMC and DPC as intermediates, has been commercially established and will account in a short time for about 300,000 ton/yr polycarbonates.



(14)

This process avoids at the same time the use of phosgene as a reactant, and methylene chloride as a solvent, and thus eliminates the coproduction of NaCl salt. Avoiding these drawbacks is a common feature

of DMC-based processes, matching the guidelines for a clean chemistry: besides exploiting the use of a harmless reagent, they avoid the use of solvents and the coproduction of salts or HCl, and generate only easily disposable or recyclable by-products.

The DMC-based route to aromatic polycarbonates takes place via production of DPC as intermediate and successive melt polymerization between DPC and bisphenol-A, overcoming the previous technology, based on interfacial polymerization with phosgene.[6]

The transesterification reaction between DMC and phenol is carried out in the liquid phase in the presence of a variety of homogenous catalysts.[19] According to the process developed by EniChem, DPC is obtained from DMC in two steps [reactions (15a) and (15b)]: in the first, DMC and phenol are reacted in the presence of a titanium alkoxide catalyst to give methylphenylcarbonate (MPC); in the second, MPC is disproportionated to DPC and DMC.[20]

$$PhOH + (CH_3O)_2CO \rightarrow PhO(CO)OCH_3 + CH_3OH \quad (15a)$$

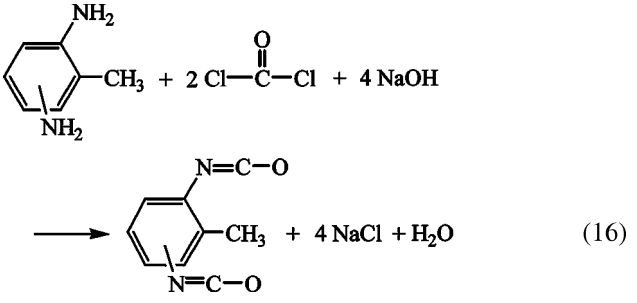
$$2PhO(CO)OCH_3 \rightarrow (PhO)_2CO + (CH_3O)_2CO \quad (15b)$$

There were two critical aspects on the road from DMC to DPC. The first was related to obtaining DPC with high selectivity, because it was known that DMC is a powerful methylating agent toward phenol with the formation of anisole. EniChem first disclosed the DMC to DPC route through the use, as catalysts, of a number of Lewis acids and metal alkoxides, especially titanium tetraphenoxide, allowing selective DPC formation. By the use of titanium compounds, selectivity well over 99.5% to DPC is obtained. Afterward, many other catalysts were found suitable, like Sn, Pb, Mo derivatives, and others.[21] The second critical aspect of DPC production was related to the highly unfavorable equilibrium of the reaction, the equilibrium constant being as low as 7.9×10^{-10} at 25°C. Performing the transesterification reaction only to produce mainly the intermediate MPC and, instead of the second transesterification step from MPC to DPC, carrying out a disproportionation of MPC to DPC and DMC partially overcomes the equilibrium constraints. Still, a very unfavorable equilibrium has to be dealt with in MPC formation. Even if a positive effect of raising the temperature on the equilibrium is observed, because of the slight endothermicity of the reaction, and the use of titanium tetraphenoxide as catalyst allows to carry out the reaction with high selectivity up to at least 200°C, the gain on the equilibrium constant value is not so pronounced as to relieve the problems raised by the reaction thermodynamics: at 200°C, an equilibrium constant of 1.7×10^{-3} still holds for the formation of MPC. Therefore, good process engineering design was required to successfully carry out the reaction to MPC on a large industrial scale, with a reasonable energy consumption. As there are many equilibrium reactions, the process is best carried out in a reactive column with countercurrent phenol/DMC feeds, using optimized operation conditions, to guarantee an isothermal temperature profile in the column, and adopting an excess of feed DMC over phenol. Under the selected conditions, phenol conversion per pass in the range of 25–35% is obtained.

Another well-established application of DMC in the field of polycarbonates relates to the production of polydiethyleneglycol bis(allylcarbonate), a thermosetting resin used in the production of optical glasses and lenses. The nonphosgene process involves the intermediate formation of diallylcarbonate from DMC, whereas the traditional process was based on the use of diethyleneglycol bis(chloroformate), obtained from phosgene, and allows high flexibility in terms of customer tailored products.

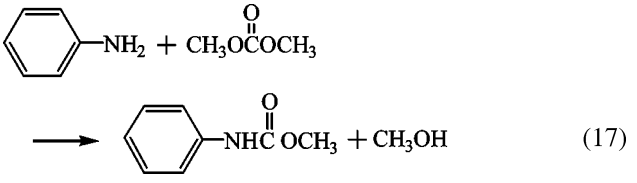
Carbamate and Isocyanate Production

One of the most important uses of phosgene is the production of isocyanate for polyurethane. For example, toluene diisocyanate is produced by the reaction of diaminotoluene with phosgene [reaction (16)].



(16)

The obvious drawbacks of this method are high toxicity of phosgene and handling of a large amount of hydrogen chloride. Therefore, several methods for synthesizing isocyanates (or carbamate ester as a precursor) without using phosgene have been sought.



(17)

The reaction between DMC and primary or secondary amines yields carbamates [reaction (17)]. Suitable catalysts are needed to achieve good reaction rates

and high selectivities. When aliphatic amines are reacted, strong bases, such as alkali alkoxides, can be used to afford quantitative yields under mild conditions.[22] Interestingly, carbon dioxide has also been proved to be an effective catalyst in the reaction. Lewis acids, such as $AlCl_3$ or AlI_3 and mercuric salts, zinc acetate, or lead oxides, and salts are required in the case of anilines.[23] By these methods it is possible to obtain both mono- and dicarbamates.

The nonphosgene production of isocyanates takes place through the thermolysis of the corresponding carbamate. The carbamate synthesis may involve a number of possible alternative ways, such as the reaction of a nitrocompound with CO, or the reaction of an amine with CO and O_2, with urea and alcohol, or with a carbonic ester. Among these routes, the reaction of DMC or DPC with aliphatic amines is a very efficient way to produce carbamates.

A nonphosgene process for the production of methyl isocyanate, starting from methylamine and DPC as raw materials, has been established by EniChem, resulting in the commercialization of two production units in the United States (1988) and China (1994).[24] Selective formation of *N*-methyl *O*-phenyl urethane (MPU) is achieved by the reaction, in nearly stoichiometric amounts, of methylamine and melt DPC. This reaction is quickly and easily performed continuously at 50°C and atmospheric pressure. The resulting liquid product, consisting of an equimolar mixture of MPU and phenol, is thermally treated at 210°C to decompose MPU, then the isocyanate and phenol produced are separately recovered. The latter, after distillation, can be eventually recycled to DPC production. Not only is the use of phosgene avoided by this process, but also CH_3NCO is safely generated as a controlled vapor stream by MPU decomposition in the pyrolysis step, according to the heat provided to the endothermic reaction. The stream of isocyanate can be directly reacted to the finally desired product, which is normally a carbamate pesticide or pharmaceutical, avoiding the need for any risky accumulation or storage of CH_3NCO. By using suitable catalysts, the stationary concentration of CH_3NCO in the reacting solution is kept very low during the entire addition of the isocyanate and reaches negligible concentration values at the end of the addition in short intervals, allowing efficient production under highly safe conditions. 1,8-Diazabicyclo[5.4.0]undec-7-ene (DBU) behaves as a particularly useful catalyst.[25] Also in this process, it is worth noting the absence of coproducts and solvents.

The efficient production of *O*-methyl carbamates from aliphatic amines and DMC requires, as a rule, the use of catalysts such as basic compounds. With such catalysts the reaction is selectively performed under very mild conditions and an excellent hourly productivity is obtained, in the production of both monocarbamates and dicarbamates. The following isocyanate forming decomposition step can be performed in several ways: through purely thermal processes or through the aid of catalysts to improve the performance; in liquid phase, by diluting the carbamate in an inert high-boiling solvent, or in the gas phase; and in continuous or in batch processes. The required temperatures are, as a rule, in excess of 250°C because *O*-alkyl,*N*-alkyl carbamates are more resistant to decomposition than the aryl substituted counterparts. A relevant example is the process for hexamethylene diisocyanate (HDI) production studied and patented by EniChem.[26] Hexamethylenediamine (HDA) is transformed into the corresponding hexamethylene diurethane (HDU) in excess DMC, acting as a reactant and solvent at the same time 99% yield is obtained after a 0.5 hr reaction at 65°C in the presence of $NaOCH_3$ as catalyst. In a second step HDU undergoes a noncatalytic vapor phase thermolysis in a tubular reactor at 420–460°C, 30 torr residual pressure, and 1–3 sec contact time, to afford HDI with a 95% overall process yield, based on the amine.

A process for the nonphosgene production of aromatic isocyanates, like the largely produced commodity isocyanates TDI and MDI, would be of outstanding industrial interest. This goal will require solving formidable chemical and technological problems, to stand the competition with the existing phosgene processes. EniChem is currently actively pursuing the development of a DMC based production process for TDI and, following promising results obtained on the laboratory scale, further activity advancement involving design and operation of a pilot plant has been scheduled.

Dimethylcarbonate as Methylating Reagent

Dimethylcarbonate is a versatile reagent for methylation reactions at C, N, O, and S centers, behaving as a good substitute for dimethylsulfate or methylhalides that are toxic and corrosive.[1] For the mono-methylation of activated methylene groups in substrates such as arylacetonitriles, useful intermediates for antiinflammatory drugs, DMC is better than other methylating agents for selectivity to mono-methylated derivatives. Methylation of amines, amides, imide dyes, phenols, and thiols can be achieved by the reaction with DMC in liquid phase, in the presence of basic catalysts (e.g., carbonates of cesium or potassium) and phase transfer agents, or in gas phase, by gas–liquid phase transfer catalysis in the presence of polyethylene glycol (PEG 6000) or zeolites supported carbonates. Finally, gas-phase methylation of substituted phenols on alumina supported alkali metal salts was recently reported.[6]

Quaternary ammonium salts are obtained by exhaustive methylation of aliphatic tertiary amines.

An important application is the preparation, from trimethylamine and DMC, of electronic-grade tetramethylammonium hydroxide, free from chloride ion, by electrolysis of the hydrogen carbonate.

Dimethylcarbonate as Solvent

Dimethylcarbonate, represents a viable alternative to acetate esters and ketones in most applications, from paints to adhesives, taking advantage of its good solvency power. It should also be reminded that DMC is the file leader of many derived carbonic esters, available by transesterification reactions, whose properties can be tailored according to the target applications, like in the field of lubricating oils. Dimethylcarbonate, as a nonaqueous electrolyte component, is finding increasing application in the field of lithium rechargeable batteries, as witnessed by the number of patents in the area (more than 200 in the last 3 yr). A further example of DMC application is as a blowing agent in polyurethane foam after the ban on CFC.[27]

Dimethylcarbonate as Fuel Additives

In recent years, DMC has been taken into consideration as oxygenate to reduce vehicle emissions associated to environmental and health risks. The reasons are the outstanding oxygen content in the DMC molecule (53.3 wt%) combined with its good blending properties.[2]

In addition to its blending properties, there is one physical characteristic of DMC that is of paramount significance. The distribution of DMC in a gasoline/water two-phase system is much more favorable than for the C_1–C_3 light alcohols. Another characteristic of DMC that is a concern is its freeze point. Pure DMC freezes at 1°C (34°F). However, at 3–4 wt% in gasoline, DMC remains in solution without becoming cloudy or hazy below −40°C (−40°F).

Another characteristic of DMC that is important is hydrolysis. Dimethylcarbonate can react with water to produce methanol and CO_2. Conditions that favor this reaction are elevated temperature, excess free water, and alkali metal carbonates. Under normal circumstances, in gasoline without any free water, the reaction does not proceed at all. With a small amount of free water and adequate time and temperature, some of the DMC will enter aqueous phase and hydrolyze. In a completely aqueous system with an alkali carbonate catalyst present and elevated temperature, DMC hydrolyzes slowly. For example, DMC in water held at 70°C (158°F) for 1 week with ample K_2CO_3 present was hydrolyzed to methanol and CO_2 with a conversion of 50%.[2]

CONCLUSIONS

One of the most important goals of the chemical industry with respect to environmental issues is a reduction in the use of toxic materials and in the amount of waste generated. Industrial development of nonphosgene route to DMC production has resulted in DMC playing a significant role as a nontoxic versatile chemical and as a component of reformulated fuels. It is expected that the future will show a growing expansion of the current and new DMC applications.

REFERENCES

1. Ono, Y. Catalysis in the production and reactions of dimethylcarbonate, an environmentally benign building block. Appl. Catal. **1997**, *155*, 133–166.
2. Ono, Y. Dimethylcarbonate for environmentally benign reactions. Catal. Today **1997**, *35*, 15–25.
3. Pacheco, M.A.; Marshall, C.L. Review of diemthyl carbonate manufacture and its characteristics as a fuel additive. Energy Fuels **1997**, *11*, 2–19.
4. Ravetti, F. The role of dimethylcarbonate in the replacement of hazardous chemicals. Surf. Chem. Catal. **2000**, *3*, 497–503.
5. Buysch, H.-J. Carbonic acid esters. In *Ullmann's Encyclopaedia of Industrial Chemistry*, 5th Ed.; Gerhartz, W., Yamamoto, Y.S., Campbell, F.T., Pfefferkom, R., Rounsaville, J.F., Eds.; VCH Publishers: Weinheim, 1986; Vol. A5, 197–202.
6. Delledonne, D.; Rivetti, F.; Romano, U. Developments in the production and application of dimethylcarbonate. Appl. Catal. A. **2001**, *221*, 241–251.
7. Cavinato, G.; Toniolo, L.J. New aspects of the synthesis of dimethyl carbonate via carbonylation of methyl alcohol promoted by methoxycarbonyl complexes of palladium(II). J. Organomet. Chem. **1993**, *444*, C65–C66.
8. Rivetti, F.; Romano, U. Alcohol carbonylation with palladium(II) complexes, effects of ligands, carbon monoxide, pressure and added bases. J. Organomet. Chem. **1979**, *174*, 221–226.
9. Filardo, G.; Galia, A.; Rivetti, F.; Scialdone, O.; Silvestri, G. Catalytic systems based on transition metals for the carbonylation of methanol to dimethylcarbonate. Electrochim. Acta **1997**, *42*, 1961–1965.
10. Tomishige, K.; Sakaihori, T.; Sakai, S.-I.; Fujimoto, K. Dimethyl carbonate synthesis by oxidative carbonylation on activated carbon supported $CuCl_2$ catalysts: catalytic properties and structural change. Appl. Catal. A. **1999**, *181*, 95–102.

11. King, S.T. Reaction mechanism of oxidative carbonylation of methanol to dimethyl carbonate in Cu-Y zelolite. J. Catal. **1996**, *161*, 530–538.
12. Uchiumi, S.; Ateka, K.; Matsuraki, T. Oxidative reactions by a palladium-alkyl nitrite system. J. Organomet. Chem. **1999**, *576*, 279–289.
13. Miyazaki, H.; Shiomi, Y.; Fujitsu, S.; Masunaga, K.; Yanagisawa Process for the Preparation of Oxalic Acid Diesters. H. U.S. Patent 4,384,133, May 17, 1983.
14. Nishihira, K.; Tanaka, S.; Nishida, Y.; Manada, N.; Kurafuji, T.; Murakami, M. Method of Producing an Ester Compound U.S. Patent 5,869,729, Feb 9, 1999.
15. Knifton, J.F.; Duranlean, R.G. Ethylene glycol-dimethyl carbonate cogeneration. J. Mol. Catal. **1991**, *67*, 389–399.
16. Tatsumi, T.; Watanabe, Y.; Koyano, K.A. Synthesis of dimethyl carbonate from ethylene carbonate and methanol using TS-1 as solid base catalyst. Chem. Commun. **1996**, 2281–2282.
17. Ryu, J.Y. Process for Making Dialkyl Carbonates . U.S. Patent 5,902,894, May 11, 1999.
18. Harada, K.; Imbe, Y.; Nishihira, K.; Tanaka, S.; Fujitsu, S.; Sugise, R.; Kashiwagi, K.; Sumida, T.; Doi, T.; Nishio, M. Catalyst for Decarbonylation Reaction U.S. Patent 5,892,091, Apr 6, 1999.
19. Shaikh, A.G.; Sivaram, S. Organic carbonates. Chem. Rev. **1996**, *96*, 951–976.
20. Rivetti, F.; Paludetto, R.; Romano, U. Continuous Process for the Preparation of Phenyl Methyl Carbonate U.S. Patent 5,705,673, Jan 6, 1998.
21. Shaikh, A.G.; Sivaram, S. Organic carbonates. Chem. Rev. **1996**, *96*, 951–976.
22. Mizia, F.; Rivetti, F.; Romano, U. Procedure for the Preparation of Alkyl Isocyanates U.S. Patent 5,315,034, May 24, 1994.
23. Fu, Z.H.; Ono, Y. Synthesis of methyl *N*-phenyl carbamate by methoxycarbonylation of aniline with dimethyl carbonate using Pb compounds as catalysts. J. Mol. Catal. **1994**, *91*, 399–405.
24. Rivetti, F.; Mizia, F.; Garone, G.; Romano, U. Process for the Production of *N*-Methylcarbamates U.S. Patent 4,659,845, Apr 21, 1987.
25. Mizia, F.; Rivetti, F.; Romano, U. Process for Producing *N*-Alkyl-Carbamates U.S. Patent 5,021,590, Jun 4, 1991.
26. Mizia, F.; Rivetti, F.; Romano, U. Procedure for the Preparation of Alkyl Isocyanates U.S. Patent 5,315,034, May 24, 1994.
27. Stefani, D.; Sam, F.O.; Lunardon, G. Process for Preparing Low-Density Flexible Polyurethane Foams, and the Polyurethane Foams Obtained U.S. Patent 5,340,845, Aug 23, 1993.

Distillation Column Design: Packing

Andrew W. Sloley
VECO USA, Inc., Bellingham, Washington, U.S.A.

INTRODUCTION

Fractional distillation is the most widely used operation in the process industries. Applications cover the range from petroleum refining, gas processing, chemicals, to pharmaceuticals and foods. Process and mechanical design requirements must merge for successful application. Distillation column internal design meshes both mechanical and process requirements in a careful selection of packing and supporting internals.

The objective of mechanical layout and design of a distillation column is to construct and install equipment that creates the conditions that will meet the process design objectives of a distillation service. However, mechanical design considerations and process design objectives interact. Balancing mechanical convenience vs. process needs minimizes investment for any given process. Process design must take equipment limitations into account. It cannot require physically impossible equipment performance. Mechanical design must meet the process objectives. Mechanical convenience cannot override process necessities.

The general design guidelines here apply to liquid–vapor packing for heat-transfer and mass-transfer services. In addition to fractional distillation this includes many stripping, absorption, and other services. This discussion covers a range of common equipment situations. The objective of the discussion is to allow for a design sizing (overall height and diameter) of a straightforward, new distillation column. Complex columns and modified columns present many challenges requiring specialized application of design criteria. The references included direct the engineer to in-depth discussions on special design requirements.

WHEN TO USE PACKING

The main distillation devices are the tray and packing. Both trays and packing come in different types. Device selection requires balancing many competing factors. Table 1 lists the factors that favor each type of equipment. Final equipment selection is usually service specific.

For new tower design, packing is the favored choice in 1) low-pressure operation; 2) low-pressure drop operation; 3) high liquid-to-vapor ratios; 4) low liquid-to-vapor ratios; 5) ceramic and polymer materials of construction; and 6) weight and size critical installations.

PACKING SYSTEM DESIGN

Introduction

A wide variety of random packing and structured packing are available. For new construction, except for especially corrosive services, pressure drop critical applications, and special cases, economics drives fractionation equipment selection toward trays. Special cases include applications of very low liquid rates, and where equipment size is critical. Examples of critical equipment size include units on offshore platforms, towers in severe earthquake zones, and towers to be housed inside buildings.

Severely corrosive services that require ceramic or special polymer materials favor packing design as tray fabrication in anything but metal is extremely difficult and expensive. Packing, with its low pressure drop, is also the equipment of choice in most systems operating under vacuum.

The majority of packing installations go into existing tray towers for new operating conditions. Some combination of capacity increase, product quality improvement, or energy reduction may be the target. Revamp packing installations require critical attention to integration of the packing system with the existing equipment. Many unit failures have occurred from reusing distributors, feed devices, and draw pans designed for tray towers with packed towers.

The complexities of packing refits into an existing tower are well beyond the current scope of discussion. Here, we will concentrate on the simplest element of packed tower design, selection and verification of packing hydraulic performance, along with an introductory discussion of packing efficiency.

When faced with a packed tower, always keep in mind that packing involves an entire system, not just some mass-transfer surface. Packed towers include packing supports, packing, packing hold-down, feed distributors, internal redistributors, and other equipment. Fig. 1 shows a generic packed tower system.

Similar to trays, packing flooding has no clearly defined and commonly accepted definition. Different

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008102

Table 1 Device selection criteria

Device	Capacity	Flexibility	Pressure drop	Notes[a]	Cost
Packing					
First-generation random	Low	Moderate	Medium	1	Inexpensive
Second-generation random	Medium	Moderate	Medium-low	1	Inexpensive
Third-generation random	High	Moderate	Low	1, 5	Moderate
Fourth-generation random	Very high	Moderate	Low	1, 5	Moderate
Structured	Very high	Moderate	Very low	1, 5	High
Grid	Very high	Moderate	Very low	1	Moderate
Trays					
Valve	High	Very good	High-medium	2	Moderate
Valve, caged	High	Excellent	High		High-moderate
Valve, fixed	High	Good	Medium		Moderate
Sieve	High	Good	Medium-low	2	Moderate-low
Dual flow	High	Very poor	Medium-low	3, 4	Moderate
Film	Low	Moderate	Low	5	High
Bubble cap	Medium	Excellent	High		High
Baffle (shed)	Very high	Poor	Very low	4, 6	Low

[a]Notes—1: Flexibility normally limited by liquid distribution system. 2: Most common types of tray. 3: No downcomer. 4: Very fouling services. 5: Vacuum services. 6: Very high liquid rates.

companies will choose definitions that favor specific marketing approaches. Literature values often define flooding based on very subjective factors. Even the same engineer may observe different flooding points for the same packing during different tests because of dynamic behavior of system hydraulics and the subjective definitions in use.

When quoting packing percentage of flood two different numbers are referenced: 1) percentage of flood at constant vapor-to-liquid (V/L) ratio and 2) percentage of flood at constant liquid (L) load. These concepts come from graphical methods of packing selection. Fig. 2 shows a packing flooding curve. Flooding vapor load is shown vs. liquid load. The Souders–Brown coefficient (c_{SB}) represents the vapor load:

$$c_{SB} = V_s\sqrt{\frac{\rho_V}{\rho_L - \rho_V}} \quad (1)$$

where V_s is the superficial velocity of the vapor across the cross section of the tower (in ft/sec). For the example shown, the constant V/L flood is $AB/AC = 20/25.2 = 0.784 = 78.4\%$. The constant L flood is $DB/DE = 0.2/0.278 = 0.719 = 71.9\%$. V/L flood is most applicable to distillation applications where increases in capacity are handled at constant V/L, equipment constraints permitting. Constant L flood is most applicable to absorption applications where increased vapor may not have a corresponding increased liquid rate. A constant vapor (V) load flood can also be stated; however, by common usage this is rarely, if ever, used.

Across the usable loading range, packing hydraulics go through two major zones. First, as vapor load increases the packing pressure drop increases close to proportionally to the square of the vapor rate. In this zone, the unloaded region, liquid holdup is essentially constant and the liquid volume fraction in the packing is low. The pressure drop comes from the vapor head losses as it flows through the packing. Once the loading point is reached, liquid holdup varies with gas flow rate. This is the loading region. Vapor kinetic energy is now being used to support the extra liquid retained in the packing, hence the pressure drop varies with gas flow rate to some power greater than 2. Fig. 3 shows a typical packing pressure drop curve. In this case the pressure drop is shown as inches of fluid per foot of packing vs. c_{SB} for parametric flow rates.

Eventually, higher gas rates prevent liquid from flowing down the packing. Visually, a continuous liquid surface forms across the top of the packed bed. This is another classical definition of flooding. Translation of the classical definition to observed system behavior is difficult and subjective. Most available packing capacity data come from vendor tests. Typically, each vendor uses a definition suitable for the best presentation of their product. Vendors have had few incentives to clarify equipment comparisons. Caution should be used in comparing flooding predictions based on different data.

Design Basis

The most commonly used design procedure is based on the Sherwood–Leva–Eckert packing correlation curve, commonly referred to as the general pressure drop (GPD) correlation curve. Multiple versions of the curve and associated correlations have been developed. They differ in the correlation parameters used in

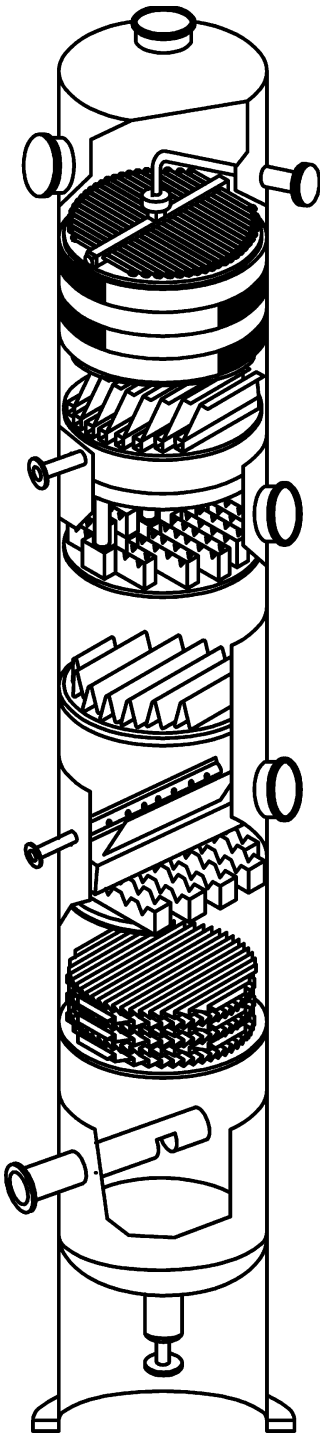


Fig. 1 Packed tower system. (Modified from Binkley, M.J.; Thorngren, J.T.; Gage, G.W.; Bonilla, J.A.; Beckman, D.H. Method of Downcomer-Tray Vapor Venting. U.S. Patent 5,106,556, Apr 21, 1992.)

defining the GPD curve axis. The GPD curve does not directly produce a ''flood'' load prediction. Rather, it predicts a pressure drop for a given loading. Then, the pressure drop is compared to a flooding pressure drop to get an approach to flooding.

The design method shown is a combination of the alternate form of the Norton method[1] for determining pressure drop coupled with the Kister and Gill[2] correlation for defining pressure drop at flood. The method works well for random packing. Limitations of the method and on the range of applicability are discussed at the end of the procedure. For structured packing, the apparent packing factor used in the equation varies with liquid and vapor load. Vendor curves should be obtained and used for proper evaluation of structured packing systems.

Both the design curve shown and the Kister and Gill correlation use a packing factor as the key definition of the packing's hydraulic capacity. Many researchers have attempted to define the packing factor by geometric analysis. However, the best definition of packing factor is that it is a correlating number that fits the packing performance to a pressure drop correlation curve. Because different pressure drop correlation curves are in use, packing factors reported from different sources may vary.[3–8] Test data are required to make fine distinctions between packing types.[9] Simple comparisons based on packing factors may be misleading.

Design Outline

The general design procedure for a new tower is:

1. Set design basis.
2. Selection of the packing type.
3. Detailed tower sizing for diameter and height: iterate on packing selection until overall design criteria are met.

Set design basis

Before starting work on the packed tower design obtain the following data:

1. Internal liquid and vapor traffic inside the column.
2. Vapor and liquid density, liquid viscosity, and surface tension.

The method shown does not take into account surface tension affects on capacity. Other detailed methods may include these.

The required liquid and vapor rates are the amount of liquid and vapor traversing the same plane through the packed bed. Owing to separation affects and heat transfer occurring concurrently with mass transfer the most loaded region of the packed bed may be at any point in the bed. Standard practice is to check

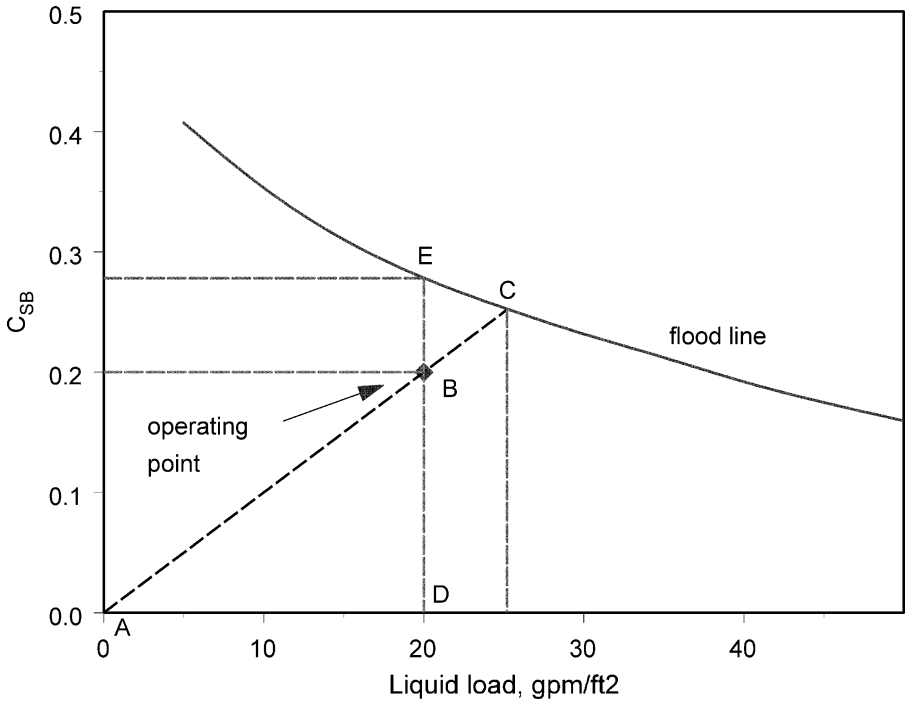


Fig. 2 Flooding capacity and constant V/L vs. constant L flood evaluation (graphical).

loading values from the bottom, middle, and top of the bed.

Today, engineers most often use simulation packages to calculate the vapor and liquid traffic. Care should be taken to extract the appropriate total stream rates and densities at operating conditions from the simulation. Many columns have failed as a result of taking incorrect flow rates or properties from the simulation.

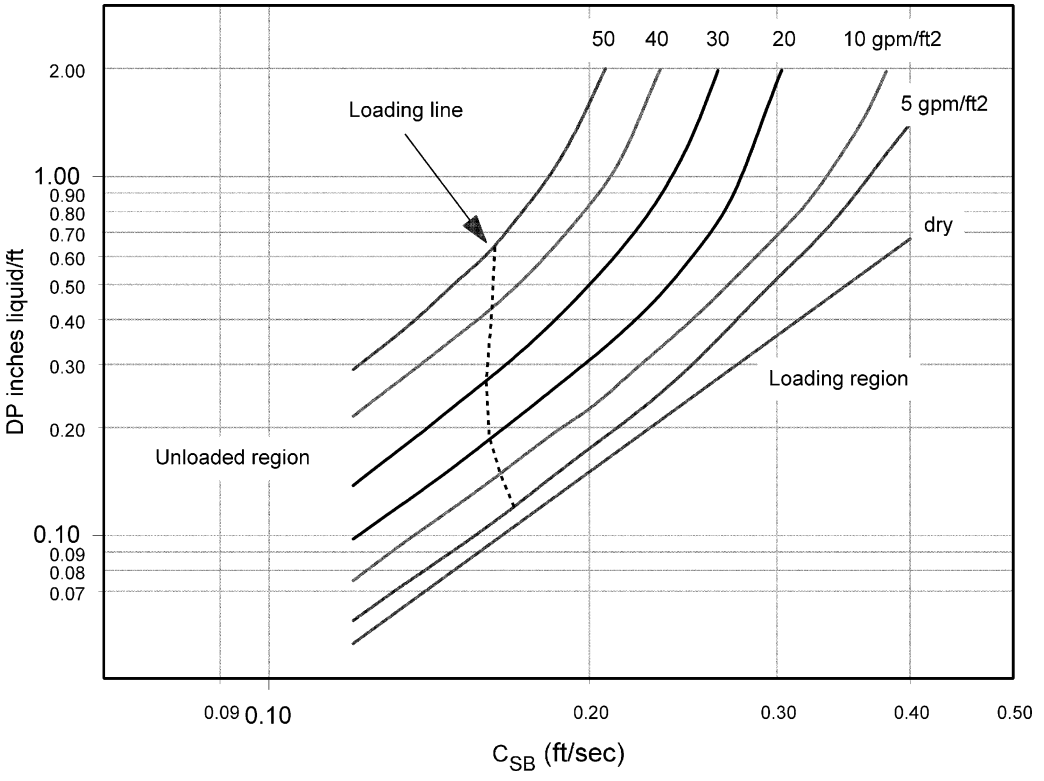


Fig. 3 Typical packing pressure drop curve.

Selection of packing type

Packing includes random packing, structured packing (sheet), grid, and wire-mesh packing. Fig. 4 shows the typical random packing available. Fig. 5 shows a typical structured packing brick. Fig. 6 shows a grid-type packing. Packing selection depends on balancing a variety of factors. Table 2 lists major packing selection criteria. Generally, most applications will call for either random packing or structured packing.

Type selection includes more than just hydraulic and efficiency evaluation. Other important factors include fouling resistance,[10] corrosion resistance, liquid holdup, ease of installation, ease of removal, and operability range.

Fouling arises from two mechanisms: introduction of foreign material in the unit feed streams and internal behavior and system characteristics. Packing is subject to fouling from both particulate and system sources. Liquid distributors are especially sensitive to particulate. Wire basket strainers should always be installed upstream of liquid feeds to packed columns. Fouling on the packing arises from materials being caught or sticking or is due to reactions at stagnant points in the packing. Random packing has more stagnant areas than structured packing or grid. In any given packing family, the more efficient the packing, the greater the tendency to foul. This is an important consideration in process selection of equipment for fouling services.

Packing is available in metal, polymers, and ceramics. In rare circumstances glass and wood packing has also been used. Materials selection depends on process requirements.

Metal packing is usually made of alloy material rather than carbon steel. The main reason for this is the susceptibility of carbon steel to corrosion during transport, storage, and installation. Unless the process requires carbon steel for specific reasons, it should never be used for packing.

Because of the thinness of the material, packing is very susceptible to damage during installation and inspection. Structured packing is extremely susceptible to handling damage. Damage to other internals may require removal of packing for access and support replacement. For this reason, in services where packing may need to be removed and reinstalled periodically, use of higher-grade materials and thicker materials in distributors and supports may be called for. Also, a minimum structured packing thickness of 0.008 in. (0.203 mm) should be specified if removal and reinstallation are required.

Random packing can be classed into first- through fourth-generation packing types. Raschig rings are the most widely known first-generation packing. The packing is a tube section. Second-generation random packing opens the interior of the tube to allow higher vapor flows. Pall rings are a traditional second-generation random packing. Third-generation random

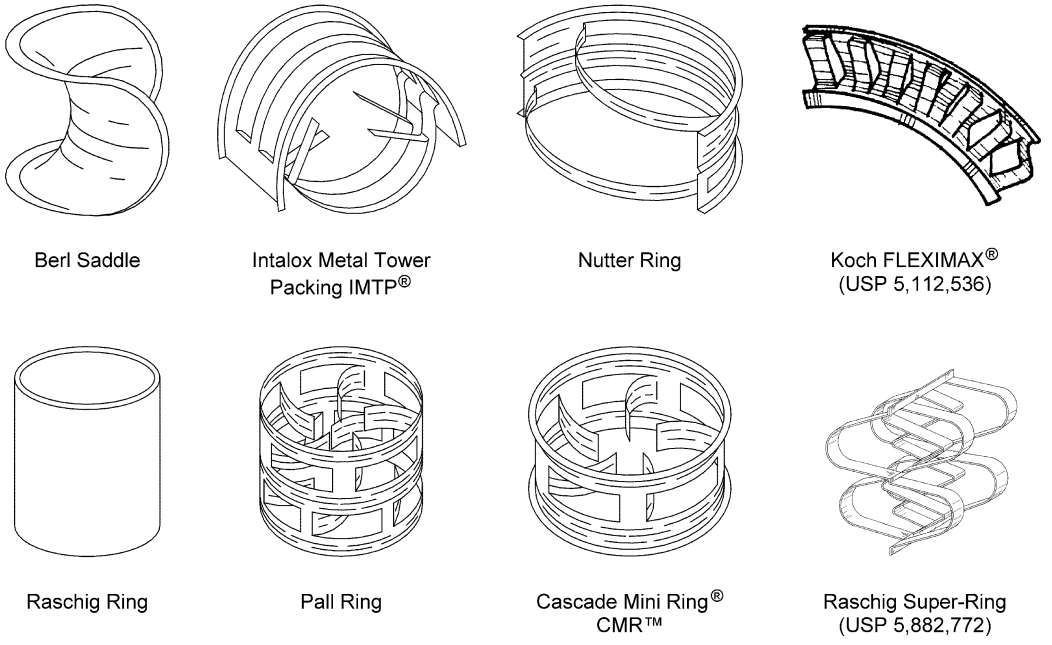


Fig. 4 Selection of random packing. (Illustration of Koch FLEXIMAX® from McNulty, K.J.; Yeoman, N.; Li-Hsieh, C. Random Packing Element and Method. U.S. Patent 5,112,536, May 12, 1992. Illustration of Raschig Super-Ring from Schultes, M. Packing Element for Use, in Particular, in Mass Transfer and/or Heat Transfer Columns or Towers. U.S. Patent 5,882,772, Mar 16, 1999.)

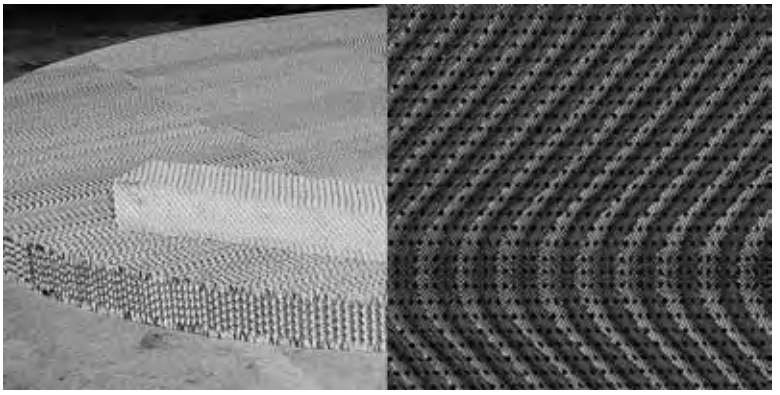


Fig. 5 Structured packing (Sulzer Mellapak® 252Y, courtesy of Sulzer Chemtech). Left shows overall packing layer. Right shows smooth flow direction change in center of packing layer used in this particular packing.

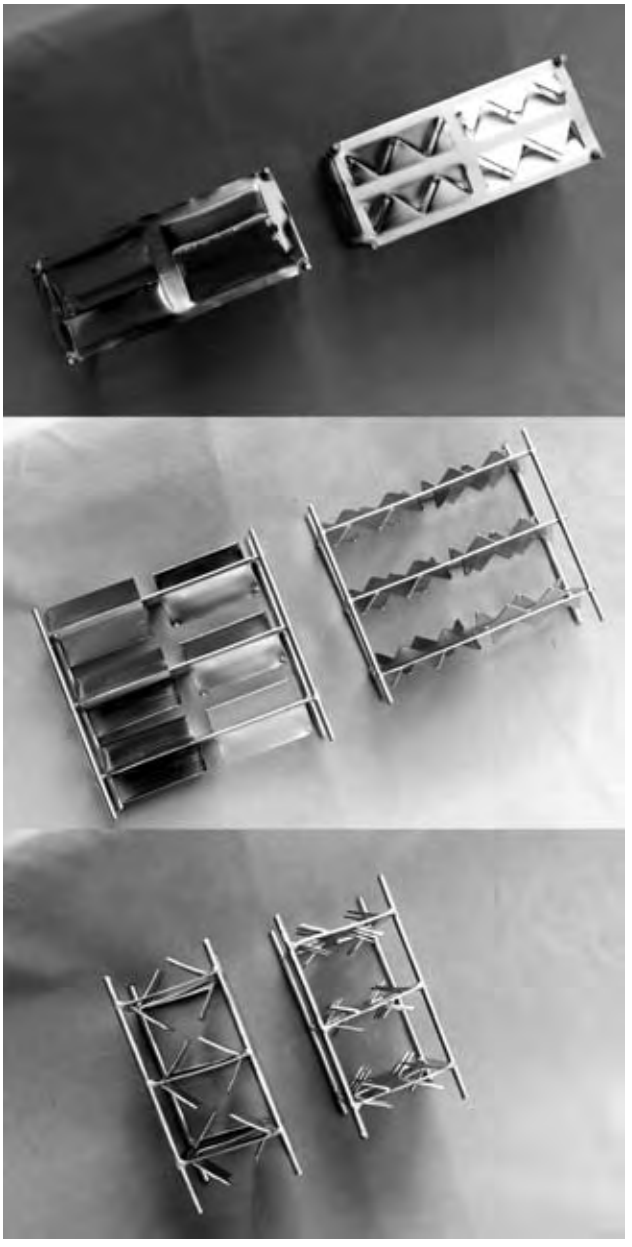


Fig. 6 Grid packing. Top, side, and end views of section of grid.

packing starts to use more complex shapes. The packing increases the number of potential drip points and film surface area for mass transfer. Third-generation random packing includes CMR™, Intalox® Rings, and Nutter Rings. Fourth-generation packing uses very complex shapes to optimize vapor and liquid mass transfer. These include the new Raschig SuperRing®. Overall, for any packing in a given generation, capacity and efficiency will be very similar for packing elements of the same dimension. Table 3 classifies commonly available packings.

At the present time, unless other factors are decisive, a third-generation random packing choice is the best starting point for packing installations. Situations where structured packing is the preferred choice include vacuum operation and where very high numbers of theoretical stages are required.

Detailed tower sizing

1. Determine the foaming tendency of the system. Pick a system factor, S_F, based on system experience or use typical values from Table 4.
2. Select a preliminary packing type. In general, select either a third-generation random packing or a structured packing.
3. For the desired packing, find its packing factor. Tables 5–8 list packing factors for some common packing types (author estimates[11–14]).
4. Calculate the pressure drop at flood for a nonfoaming system.

$$\Delta P_{\text{fnf}} = 0.115 F_{\text{PX}}^{0.7} \tag{2}$$

5. Calculate the pressure drop at flood for the actual system.

$$\Delta P_{\text{f}} = S_{\text{F}} \Delta P_{\text{fnf}} \tag{3}$$

Table 2 Initial packing selection

Factor	Favors	Reason
Liquid rate less than $0.5\,\text{gpm/ft}^2$ ($0.76\,\text{m}^3/\text{m}^2$)	Large-size packing	Liquid distributor design limits
Vacuum systems	Large-size packing Structured packing	Low-pressure drop
Expensive materials	Structured packing	Metal thickness
Fouling	Grid	Few dead spots in flow
High vapor rates	Large-size packing	Capacity vs. cost
High number of stages required	Small-size packing Structured packing	Minimize overall height
Low number of stages required	Large packing sizes	Minimum bed depth vs. height restrictions
High vapor–liquid density difference	Structured packing	Improved vapor–liquid separation
Low vapor–liquid density difference	Random packing	Known cause of structured packing hydraulic failures
Small diameter (revamps)	Small packing sizes Structured packing Woven wire-mesh packing	Natural distribution limits with random packing in small relative diameters

6. Determine the flow parameter

$$X = \frac{L}{V}\sqrt{\frac{\rho_V}{\rho_L}} \tag{4}$$

with L and V in consistent mass units and ρ_L and ρ_V in consistent units.

7. Using the modified GPD curve (Fig. 7), the flow parameter (X), and the flood pressure drop, find the allowable packing load parameter, L_F. Fig. 7 shows pressure drop isobars at flood for different flow parameters.
8. Determine the Souders–Brown coefficient at flood:

$$c_{SB} = \frac{L_F}{F_{PX}^{0.5}\ \nu_L^{0.05}} \tag{5}$$

with liquid kinematic viscosity, ν_L, in centistokes and with

$$\nu_L = \frac{\mu_L}{L_{SG}} \tag{6}$$

with μ_L in centipoise.

9. Determine the vapor superficial velocity V_s:

$$V_s = \frac{c_{SB}}{\sqrt{\frac{\rho_V}{\rho_L - \rho_V}}} \tag{7}$$

where V_s is in feet per second and densities are in consistent units.

10. Calculate the tower cross-sectional area required for operation at flood:

$$A_{CSF} = \frac{V}{\rho_V V_s} \tag{8}$$

where A_{CSF} is in square feet, V is in pounds per second, ρ_V is in pounds per cubic feet, and V_s is in feet per second.

11. Calculate the tower cross-sectional area required for operation at the design flooding fraction:

$$A_{CS} = \frac{A_{CSF}}{F_F} \tag{9}$$

The design flooding fraction is commonly 0.8 (80% of flood) for most new units. The value should never be higher the 0.95.

12. Find the tower diameter:

$$D = \sqrt{1.27 A_{CS}} \tag{10}$$

where A_{CS} is in square feet and D is in feet.

13. Calculate the design vapor velocity:

$$V_s = \frac{V}{\rho_V A_{CS}} \tag{11}$$

where V_s is in feet per second, V is in pounds per second, ρ_V is in pounds per cubic feet, and A_{CS} is in square feet.

Table 3 Packing families

Packing type	Name	Materials	Made by
First-generation random	Raschig Rings	Metal, plastic, ceramic	
Second-generation random	Ballast® Rings	Metal, plastic	Koch-Glitsch[a]
	Ballast® Plus™	Metal	Koch-Glitsch[a]
	Ballast® Saddle™	Plastic	Koch-Glitsch[a]
	Berl Saddles	Ceramic	
	Flexiring®	Metal, plastic	Koch-Glitsch[b]
	Flexisaddle™	Plastic, ceramic	Koch-Glitsch[b]
	Hiflow® Rings	Ceramic	Rauschert Industries
	Pall® Rings	Metal, plastic, ceramic	
	Ralu-Rings®	Metal, plastic	Raschig
	Super Intalox® Saddles	Ceramic	Koch-Glitsch[c]
Third-generation random	CMR™	Metal, ceramic	Koch-Glitsch[a]
	CMR	Plastic	Jaeger[a]
	Fleximax™	Metal	Koch-Glitsch[b]
	HcKp™	Metal	Koch-Glitsch[b]
	Hiflow Rings	Metal, plastic	Rauschert Industries
	Hy-Pak®	Metal	
	IMTP®	Metal	Koch-Glitsch[c]
	Intalox® Snowflake®	Plastic	Koch-Glitsch[c]
	Lanpac	Plastic	Lantec
	Nor-Pac®	Plastic	NSW Environmental Products
	Nutter Rings	Metal	Sulzer Chemtech[d]
	Q-Pac®	Plastic	Lantec
	Super Intalox Saddles	Plastic	Koch-Glitsch[c]
	Tellerette®	Plastic	Ceilcote
	Top-Pak®	Metal	Jaeger
	Tri-pack®	Metal, plastic	Jaeger
	VSP®	Metal	Jaeger
Fourth-generation random	SuperRings®	Metal, plastic	Raschig
Grid	Koch-Glistch® Grid	Metal, plastic	Koch-Glitsch
	Mellagrid®	Metal	Sulzer Chemtech
Structure packing (45° crimp)	Gempak-A®	Metal	Koch-Glitsch[a]
	Flexeramic®	Ceramic	Koch-Glitsch[b]
	Flexipak®	Metal	Koch-Glitsch[b]
	Intalox®	Metal	Koch-Glitsch[c]
	Max-Pak®	Metal	Jaeger
	Montz-Pak B	Metal	Montz
	Rombopak®	Metal	Kuhni
	Sulzer Mellapak-Y	Metal	Sulzer Chemtech
Structured packing (60° crimp)	Gempak-B	Metal	Koch-Glitsch[a]
	Montz-Pak B	Metal	Montz
	Sulzer Mellapak-X	Metal	Sulzer Chemtech

[a]Originally Glitsch.
[b]Originally Koch Engineering.
[c]Originally Norton Chemical Process Products (U.S. Stoneware).
[d]Originally Nutter Engineering.

14. Calculate the Souders–Brown coefficient for the design diameter:

$$c_{SB} = V_s\sqrt{\frac{\rho_V}{\rho_L - \rho_V}} \quad (1)$$

where c_{SB} is in feet per second, V_s is in feet per second, and density is in consistent units.

15. Calculate the design packing load factor, L_F:

$$L_F = c_{SB}F_{PX}^{0.5}\nu_L^{0.05} \quad (12)$$

Table 4 System factors

Service	System factor
General systems	
Nonfoaming	1.0
Low foaming	0.9
Moderate foaming	0.85
High foaming	0.70
Severe foaming	0.60
Foam stable	0.30
Specific systems	
C3–C6 hydrocarbons	1.0
Low-molecular-weight alcohols	1.0
Atmospheric crude petroleum	1.0
Oxygen stripper	1.0
Rectisol contactor	1.0
Halogens	0.90
Cryogenic gas plant absorbers	1.0–0.85
Oil absorbers	0.85
Amine regenerators	0.85
Glycol regenerators	0.85
Water strippers	0.70–0.80
Amine absorbers	0.70
Glycol absorbers	0.70
Methyl ethyl ketone (MEK) units	0.60
Caustic regenerators	0.30

with liquid kinematic viscosity ν_L in centistokes.

16. From Fig. 7 with the design packing load factor from Step 15 and the flow parameter X from Step 6 find the design pressure drop in inches of water.
17. Verify the packing size and diameter of the column. Column diameter should be a minimum of 10 times the packing size for random packing. Column diameters smaller than this criterion have a significant risk of excessive liquid and vapor channeling. Select a new packing size as needed.

LIMITATIONS AND APPLICABILITY

All packing correlation methods for hydraulics have systemic limitations. These limits arise from the underlying correlation forms and the difficulty in data interpretation. Table 9 lists applicability ranges for Fig. 7. Strigle and Kister and Gill both have excellent discussions of the limitations of the method shown. In summary, the limitations are:[1,2]

1. Over the entire range of data examined the pressure drop correlation fits ~40% of the data with an excellent fit, ~40% with a good fit, ~15% with a reasonable fit, and ~5% with a poor fit.
2. Flow parameters, X, from 0.05 to 0.3 give good fits. This is the normal operating range for atmospheric distillation systems.
3. At flow parameters greater than 0.3 with nonaqueous systems the pressure drop curves are optimistic. Actual pressure drops will typically be higher than that predicted for a given load. This operating region is typical of high liquid rates and high-pressure systems (vapor density >0.05 liquid density).
4. At flow parameters less than 0.05 the predicted pressure drop is nearly independent of the flow parameter. Pressure drop in this operating range tends to be higher than predicted. This operating range is typical of vacuum operation.

To correct for pressure drop at flow parameters greater than 0.3, the flooding pressure drop on the GPD correlation, the author modifies the predicted pressure drop at flood by:

$$\Delta P_{\text{fnfc}} = L_{\text{SG}}^{-0.33} 0.115 F_{\text{PX}}^{0.7} \tag{13}$$

This equation should be used in place of equation Eq. (2) and ΔP_{fnfc} is used in place of ΔP_{fnf} in Eq. (3). Care should be taken to understand this correction. The pressure drop at flood is actually the same [continue to use Eqs. (2) and (3) to determine the predicted pressure drop at flood]. The correction is to get a correlating factor that gives a more conservative definition of allowable vapor velocity from Fig. 7. This correction gives a good fit to hydrocarbon systems.

For conservative design at flow parameters less than 0.05, use the value of 0.02 for the flow parameter.

PACKING EFFICIENCY

Packing efficiency is even more difficult to predict than hydraulic capacity. For prudent design, data are mandatory from the actual operating system or a very similar system. Operating and test data have resulted in many, many surprises when comparing demonstrated vs. predicted efficiency.

Packing efficiency is an extremely complex function of system properties (density, diffusivity, relative volatility, surface tension, wetting tendency, composition, and other factors), packing geometry, and loading. No effective predictive method based on fundamentals is available. Even comparing different types of packing is difficult. In one service a specific packing may be more efficient than another. In a different service, the relative efficiency may be reversed. The only reliable general rule is that a physically smaller packing will have a higher efficiency than a geometrically identical, larger packing.

Table 5 Packing factors: random packing, metal

Packing	Packing factor, F_{PX}, for a given size							
		$\frac{5}{8}$ in.		1 in.		1.5 in.	2 in.	3.5 in.
Ballast Rings		81		56		40	27	18
							60 mm	
Ballast Plus™							26	
				#1		#1.5	#2	#3
CMR				40		29	22	14
							#2.5	#4
CMR							19	10
						#300	#400	
Fleximax						26	17	
				#1		#1.5	#2	#
HcKp				40		22	16	14
							2 in.	
Hiflow Rings							16	
				#1		#1.5	#2	#3
Hy-Pak				45		32	26	16
		#15		#25		#40	#50	#70
IMTP		51		41		24	18	12
				#1		#1.5	#2	#3
K-Pac®				45		32	26	16
		0.7		#1		#1.5	#2	#3
Nutter Rings		42		30		24	18	13
							#2.5	
Nutter Rings							16	
		$\frac{5}{8}$ in.		1 in.		1.5 in.	2 in.	3.5 in.
Pall Rings		81		56		40	27	18
	$\frac{1}{2}$ in.	$\frac{5}{8}$ in.	$\frac{3}{4}$ in.	1 in.	1.25 in.	1.5 in.	2 in.	3 in.
Raschig Rings ($\frac{1}{32}$ in.)	300	170	155	115				
Raschig Rings ($\frac{1}{16}$ in.)	410	290	220	137	110	83	57	32
				1 in.			2 in.	
Ralu-Rings				107			43	
								75 mm
Top-Pak								16
				25		40	50	
VSP				32		21	20	

Packings are grouped into columns of similar size.

The following method for predicting efficiency is based on the Norton method for IMTP packing.[15] It should be used with extreme caution if system-specific performance data are not known. It is no worse than other methods and has the benefit of simplicity. The method should only be used for rough evaluation of packing for checking the general feasibility of a design.

Packing height can be evaluated with either the transfer-unit or the height equivalent theoretical plate (HETP) approach. Traditionally, the transfer-unit approach has been used with absorption and stripping systems and the HETP approach for fractionation services. The method used here is the HETP approach. The method is applicable to distillation and reboiled stripping services for nonaqueous, nonreacting, nonionizing, systems with a relative volatility less than 3:

$$\text{HETP} = K_{21}\left[\frac{\sigma}{20}\right]^{-0.16}(1.78)^{\mu_L} \tag{14}$$

for $\mu_L < 0.4$ cP.

$$\text{HETP} = K_{22}\left[\frac{\sigma}{20}\right]^{-0.19}\left[\frac{\mu}{0.2}\right]^{0.21} \tag{15}$$

for $\mu_L > 0.4$ cP.

Table 6 Packing factors: random packing, plastic

Packing	Packing factor, F_{PX}, for a given size					
			1 in.		2 in.	3 in.
Ballast Saddle			40		28	18
			#1A		#2A	#3A
CMR			12		16	12
	$\frac{5}{8}$ in.		1 in.	1.5 in.	2 in.	3.5 in.
Flexiring	95		55	40	26	17
			1 in.		2 in.	3 in.
Flexisaddle			40		28	18
			1 in.		2 in.	3.5 in.
Hiflow Rings			42		20	9
					(2 in.)	
Intalox Snowflake					13	
					2.3 in.	3.5 in.
Lanpac					21	14
	$\frac{5}{8}$ in.	$\frac{3}{4}$ in.	1 in.	1.5 in.	2 in.	
Nor-Pac	38	32	25	17	12	
	$\frac{5}{8}$ in.		1 in.	1.5 in.	2 in.	3.5 in.
Pall Rings	95		55	40	26	17
						7 in.
Q-Pac						4
			1 in.	1.5 in.	2 in.	3.5 in.
Ralu-Rings			41	24	17	12
			25		50	70
Super Intalox Saddles			40		28	18
			#1		#2	#3
Tellerette			35		24	17
			#$\frac{1}{2}$	(1.25 in.)	#1 (2 in.)	#2 (3 in.)
Tri-pack			28	25	16	12

Packings are grouped into columns of similar size.

Tables 10–12 show K_{21} and K_{22} values for HETP (in in. and mm) and surface tension (in dyn/cm) for a selection of packings.

A design factor of 1.13 should be applied to these HETP values to account for the differences between HETP values found in test installations and those found in commercial units:

$$\text{HETP}_{\text{design}} = 1.13\text{HETP} \tag{16}$$

The total packing height needed for the service is then calculated by:

$$H_{\text{P}} = N_{\text{t}}\text{HETP}_{\text{design}} \tag{17}$$

The author's experience is that the HETP values calculated from this method tend to be systemically optimistic. A 10% extra safety margin should be applied to these values.

BED HEIGHT LIMITATIONS

Packing bed heights are limited for both process and mechanical reasons. Liquid maldistribution in packed beds reduces efficiency by having localized liquid ratios that differ from the average across the bed. The taller the bed, the more significant is the efficiency degradation. As a good practice, packed bed heights should be limited to 10 theoretical stages or less and 20 ft (6.1 m) or less.

All types of packing are subject to a limitation on bed height to avoid excessive crush loads on the bottom packing layers in a bed. The problem is most common in random packing and structured packing installations. To avoid crushing during installation and operation, aluminum materials should not be used in beds over 12 ft (3.66 m) deep or at operating temperatures in excess of 400°F (205°C). Ceramic packing should be limited to bed heights of 20 ft (6.10 m) or less. This avoids excessive breakage during installation and operation. Polypropylene packing

Table 7 Packing factors: random packing, ceramic

Packing	Packing factor, F_{PX}, for a given size							
	$\frac{1}{2}$ in.		$\frac{3}{4}$ in.	1 in.		1.5 in.	2 in.	
Berl Saddles	240		170	110		65	45	
				1 in.		1.5 in.	2 in.	3 in.
Flexisaddle				92		52	40	22
						1.5 in.	2 in.	3 in.
Hiflow Rings						37	29	15
	$\frac{1}{2}$ in.		$\frac{3}{4}$ in.	1 in.		1.5 in.	2 in.	3 in.
Intalox Saddles	200		145	92		52	40	22
				1 in.		1.5 in.	2 in.	
Pall Rings				107		55	43	
	$\frac{1}{2}$ in.	$\frac{5}{8}$ in.	$\frac{3}{4}$ in.	1 in.	1.25 in.	1.5 in.	2 in.	3 in.
Raschig Rings	580	380	255	179	125	93	65	37
				1 in.			2 in.	
Super Intalox Saddles				60			30	

Packings are grouped into columns of similar size.

should be restricted to operation at temperatures of 200°F (93°C) or lower. Care should be taken when polypropylene packing is to be specified that the tower will not be steamed out during shut-down or start-up operations. Some engineering plastics have higher temperature-resistant capabilities but are significantly more expensive.

For a typical design 2 ft (0.61 m) of extra height is required for each redistributor when using pan-orifice distributors and 4 ft (1.22 m) of extra height is required for each redistributor when using trough-orifice distributors.

INTERNALS REQUIRED WITH PACKING

Auxiliary equipment in packing installations includes:

- Bed supports
- Bed hold-downs
- Liquid collectors
- Liquid distributors and redistributors
- Vapor distributors and redistributors.

Design of this equipment is specialized. Standard references for these equipment types are included in the References. The following discussion covers the general issues involved.

Bed supports provide for safe support of the packed bed while not interfering with the flow of vapor and liquid through the grid. Additionally, the support grid must have sufficient strength to support the weight of packing in the bed, the inspection personnel on the bed, the liquid retained on the packing during operation, and the upset forces possible during sudden foaming and vaporization incidents. Selection of load criteria must take all these factors into account.

Random packing support grids use holes small enough to prevent random packing pieces from falling through the grid and down the tower. The most effective design uses a vapor injection style. The support grid has large corrugations that increase the overall surface area available for vapor flow. The hole area open for liquid and vapor flow should be equal to or greater than the tower cross-sectional area. The corrugation size is set by the tower diameter and manway access size. Fig. 8 shows a vapor injection support.

Structured packing and grid packing support grids use a flat subway grate construction to allow for proper placement of the structured packing bricks on the support. This allows for a high degree of strength with a minimum reduction in the area available for vapor and liquid transport up and down the tower.

During upset operation with random packing, sudden load increases can lead to packing displacement or fluidization of a random packing bed. Standard designs use a mesh-type structure or a subway-grating-type grill. The holes in the mesh or grill must be smaller than the packing size. As little as possible of the tower cross-sectional area should be taken up by the hold-down.

Standard hold-downs (such as subway grating) can be used with structured packing. As an alternative, rods can be used to tie a hold-down grid and packing support grid together.

Liquid collectors provide for withdrawal of products, mixing of new feeds with the internal liquid traffic, and remixing to avoid composition gradients after a certain height of packed bed has been traversed. Liquid collectors are often designed as one item to be integral with vapor distributors and liquid redistributors.

Table 8 Packing factors: structured packing and grid

Packing	Material	Packing factor, F_{PX}	Notes[a]
Glitsch Gempak® 4A	Metal	32	2
Glitsch Gempak 3A	Metal	21	2
Glitsch Gempak 2A	Metal	16	2
Glitsch Gempak 1.5A	Metal	12	2
Glitsch Gempak 1A	Metal	9	2
Glitsch Gempak 0.5A	Metal	6	2
Glitsch Grid® EF-25A	Metal	10	5
Glitsch Grid EF-25AP	Plastic	18	5
Jaeger Max-pak®	Metal	12	
Koch Flexeramic® #28	Ceramic	40	
Koch Flexeramic #40	Ceramic	24	
Koch Flexeramic #88	Ceramic	15	
Koch Flexigrid® #3	Metal	10	
Koch Felxigrid #2	Metal	4	
Koch Flexipac® #1	Metal	30	1
Koch Flexipac #2	Metal	13	1
Koch Flexipac #3	Metal	8	1
Koch Flexipac #4	Metal	6	1
Montz-Pak B1-300	Metal	33	
Montz-Pak B1-250	Metal	20	
Norton Intalox® 1T	Metal	20	3
Norton Intalox 2T	Metal	17	3
Norton Intalox 3T	Metal	13	3
Nutter Snap Grid #3	Metal	9	4
Rombopak 6M	Metal	18	
Sulzer Mellapak® 500Y	Metal	34	
Sulzer Mellapak 350Y	Metal	23	
Sulzer Mellapak 250Y	Metal	20	
Sulzer Mellapak 250Y	Plastic	22	
Sulzer Mellapak 2Y	Metal	14	
Sulzer Mellapak 170Y	Metal	12	
Sulzer Mellapak 125Y	Metal	10	
Sulzer Mellapak 500X	Metal	25	
Sulzer Mellapak 250X	Metal	8	
Sulzer Mellapak® 2X	Metal	7	
Sulzer Mellapak 170X	Metal	6	
Sulzer Mellapak 125X	Metal	5	

[a]1: Pre-1999. 2: No longer manufactured. 3: Now manufactured by Koch-Glitsch. 4: Now manufactured by Sulzer. 5: Now marketed as Koch-Glistch Grid.

Under normal service conditions, the liquid collector is a separate tray from any liquid distribution device. The basic collector is a chimney tray with a sump (Fig. 9). The chimneys provide passage for vapor rising through the tray. The sump(s) provides for liquid drainage into a draw nozzle (for product draw) or into transition piping to a redistributor. Unless fully welded, chimney trays will leak. In all critical services, chimney trays must be fully welded.

Passages on chimney trays should be designed so that liquid can easily flow into product draws and sumps. If blockage of liquid draws is a potential

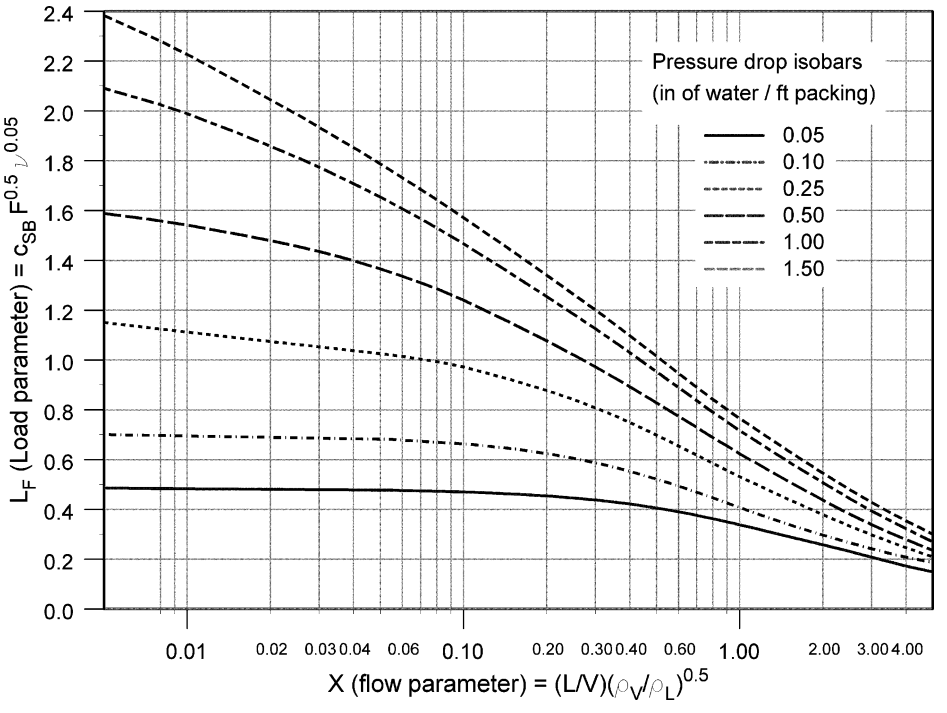


Fig. 7 General pressure drop curve.

problem, internal overflows can be added. This usually takes the form of emergency overflow down-pipes to prevent excessive buildup of liquid on the collector. Additionally, in fully welded designs, external draw nozzles and internal transfer piping should be located to enable full draining of the tray during shut-downs.

Liquid distributors function to spread an even distribution of liquid over the packing bed. Liquid distributors are commonly broken down by motive force into pressurized or gravity distributors. Pressurized distributors include pipe-orifice distributors and spray distributors. Gravity distributors include pan-orifice distributors and trough distributors.

Pressurized liquid distributors use pressure drop across an orifice to deliver a smooth liquid distribution to a packed bed. Two main types exist, pipe-orifice distributors and spray distributors. Of the two, spray distributors are the most common. In specifying either type, the main criteria to be set are the pressure drop available for use and the minimum orifice size. These must be set in conjunction with the definition of the flow rate required.

In pipe-orifice distributors, each orifice delivers a liquid stream to the packed bed. Pipe orifice distributors have the advantage that they are relatively compact and occupy a small vertical space in the tower. However, they tend to be very expensive. A separate orifice is needed for each drip point onto the packing. Typical designs call for between 5 and 15 drip points per square foot (50 and 150 drip points per square meter). The minimum orifice size is set by the fouling tendency of the system. Liquid velocities leaving the orifice should be kept to under 10 ft/sec (3 m/sec). This prevents excessive liquid splashing on the top of the packed bed.

In spray-header distributors (Fig. 10), a main header with a collection of spray nozzles is used to distribute liquid to the packed bed. Spray header distributors are well suited to liquid distribution at low liquid rates, of moderately fouling fluids, and for

Table 9 Sherwood–Leva–Eckert correlation data range

Property	Minimum	Maximum	Minimum	Maximum
System pressure	0.19 psia	470 psia	1.3 kPa	3240 kPa
Gas density	0.0036 lb/ft^3	5.2 lb/ft^3	0.057 kg/m^3	83 kg/m^3
Gas molecular weight	2	338	2	338
Liquid density	22.0 lb/ft^3	114 lb/ft^3	353 kg/m^3	1830 kg/m^3
Liquid viscosity (cP)	0.07	18	0.07	18
Surface tension (dyn/cm)	7	72	7	72

Table 10 Efficiency factors: random packing, metal

Packing	Size	K_{21} (in.)	K_{22} (in.)	K_{21} (mm)	K_{22} (mm)
Ballast Rings	0.625 in.	12.6	13.8	321	351
	1 in.	14.0	15.3	354	388
	1.5 in.	17.2	18.8	436	478
	45 mm	18.8	20.6	478	524
	2 in.	20.5	22.4	519	569
	60 mm	22.2	24.3	562	616
	3.5 in.	26.5	29	673	737
CMR	1	10.4	11.4	265	290
	1.5	12.9	14.2	328	360
	2	15.8	17.3	401	439
	2.3	17.3	19.0	440	482
	3	20.3	22.2	514	563
	4	25.9	28.3	657	719
	5	35.2	38.6	894	979
Flexiring	0.625 in.	14.4	15.8	366	401
	1	17.1	18.7	433	474
	1.5 in.	20.9	22.8	530	580
	2 in.	24.1	26.4	613	671
	3 in.	30.7	33.6	779	853
Hy-Pak	1	13.1	14.3	332	364
	1.5	19.7	21.6	500	548
	2	31.1	34.1	790	865
	3	34.7	37.9	880	963
IMTP	#15	10.7	11.7	272	296
	#25	13.0	14.4	330	366
	#40	15.8	17.3	401	439
	#50	20.7	22.8	526	579
	#60	24.2	26.7	616	678
	#70	29.9	32.6	758	827
K-Pac	30	14.3	15.7	364	398
	45	17.8	19.4	451	493
	60	20.5	22.4	519	569
	90	25.3	27.7	642	702
Nutter Rings	0.7	9.1	10.0	231	253
	1	11.0	12.0	278	305
	1.5	16.0	17.6	407	446
	2	17.3	19.0	440	481
	2.3	18.1	19.8	458	502
	3	20.7	22.6	524	574
Pall Rings	0.625 in.	11.9	13.0	301	330
	1 in.	13.1	14.3	332	364
	1.5 in.	17.6	19.3	447	489
	2.0 in.	19.5	21.4	496	543
	3.5 in.	24.1	26.4	613	671
Raschig Rings	0.5 in.	12.6	13.8	321	351
	0.625 in.	14.0	15.3	354	388
	0.75 in.	16.2	17.7	410	449
	1 in.	19.9	21.8	505	553
	1.25 in.	22.9	25.0	580	635
	1.5 in.	24.4	26.7	620	678
	2 in.	28.3	31.0	718	786
	3 in.	39.8	43.6	1010	1110
Top-pack	75	19.5	21.4	496	543
VSP	25	14.2	15.6	361	395
	40	15.2	16.7	387	423

Table 11 Efficiency factors: random packing, plastic

Ballast Rings	0.625 in.	9.8	10.7	248	271
	1 in.	12.6	13.8	319	349
	1.5 in.	17.6	19.3	447	489
	2 in.	20.9	22.8	530	580
	3.5 in.	27.2	29.8	690	756
Ballast Saddle	1 in.	7.7	8.4	195	213
	2 in.	13.8	15.1	350	383
	3 in.	20.1	22.0	510	558
CMR	1A	13.2	14.4	335	366
	2A	16.0	17.6	407	446
	3A	25.9	28.3	657	719
Flexiring	0.625 in.	9.0	9.8	227	249
	1 in.	12.6	13.8	319	349
	1.5 in.	17.6	19.3	447	489
	2 in.	20.9	22.8	530	580
	3 in.	27.2	29.8	690	756
Flexisaddle	1 in.	8.5	9.3	216	236
	2 in.	14.2	15.6	361	395
	3 in.	20.1	22.0	510	558
Pall Rings	0.625 in.	7.8	8.6	199	218
	1 in.	11.0	12.1	280	306
	1.5 in.	16.3	17.8	413	452
	2.0 in.	18.5	20.3	470	515
	3.5 in.	28.3	31.0	718	786
Super Intalox Saddle	1	8.5	9.3	216	236
	2	14.2	15.6	361	395
	3	20.1	22.0	510	558
Tri-pack	$\frac{1}{2}$ in.	9.3	10.2	237	260
	1	14.0	15.3	354	388
	2	17.2	18.8	436	478

Table 12 Efficiency factors: random packing, ceramic

Ballast Saddle	0.5 ft	10.7	11.8	273	298
	0.75 in.	15.3	16.8	389	426
	1 in.	20.3	22.2	514	563
Berl® Saddle	1.5 in.	26.9	29.4	682	746
	2 in.	30.3	33.1	768	841
	3 in.	43.0	47.1	1090	1190
CMR	1	10.4	11.4	265	290
	1.5	12.9	14.2	328	460
	2	15.8	17.3	401	439
	2.3	17.3	19.0	440	482
	3	20.3	22.2	514	563
	4	25.9	28.3	657	719
	5	35.2	38.6	894	979
Flexisaddle	0.5 in.	10.0	10.9	254	278
	0.75 in.	14.1	15.5	359	393
Intalox Saddle	1 in.	18.7	20.5	474	519
	1.5 in.	23.6	25.9	599	656
	2 in.	28.3	31.0	718	786
	3 in.	37.7	41.3	956	1050

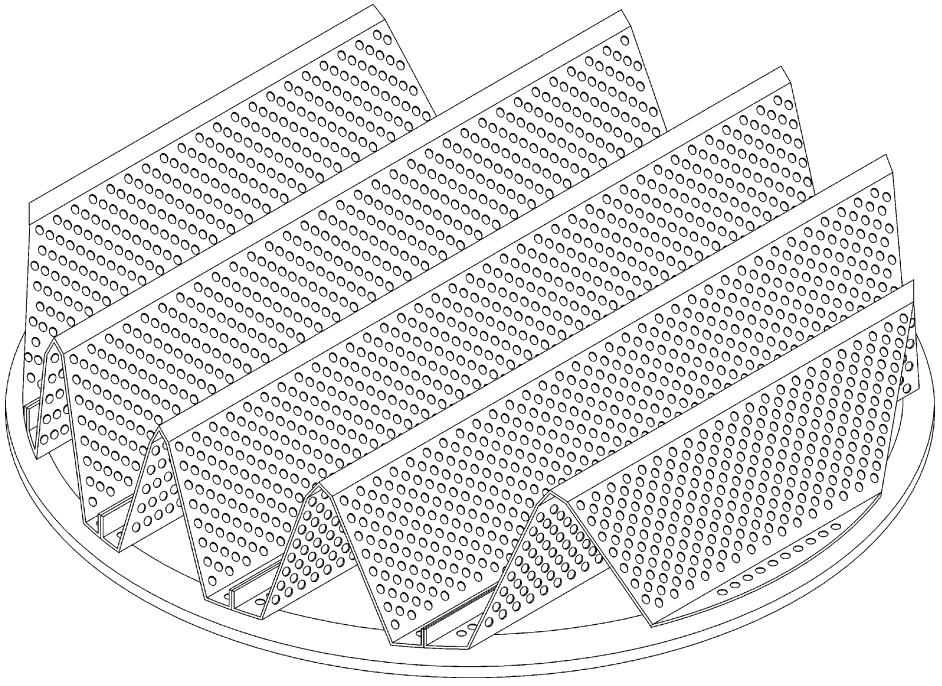


Fig. 8 Random packing vapor injection support.

heat-transfer services. Spray headers are much less expensive than pipe orifice headers because of the relatively small number of spray points. However, a spray nozzle tends to concentrate liquid along the outer periphery of the spray. For this reason, spray nozzles should never be used for distillation, absorption, or stripping services that require a large number of mass-transfer units per bed.

Spray nozzle spray angles and droplet size formation are functions of the pressure drop across a spray header. Effective pressure drop range for a spray nozzle distributor varies from 5 to 20 psi (from 34 to 138 kPa) for good design practice. Specific designs may operate outside this range.

Gravity liquid distributors use an internal head of liquid to flow through the distributor to the packed

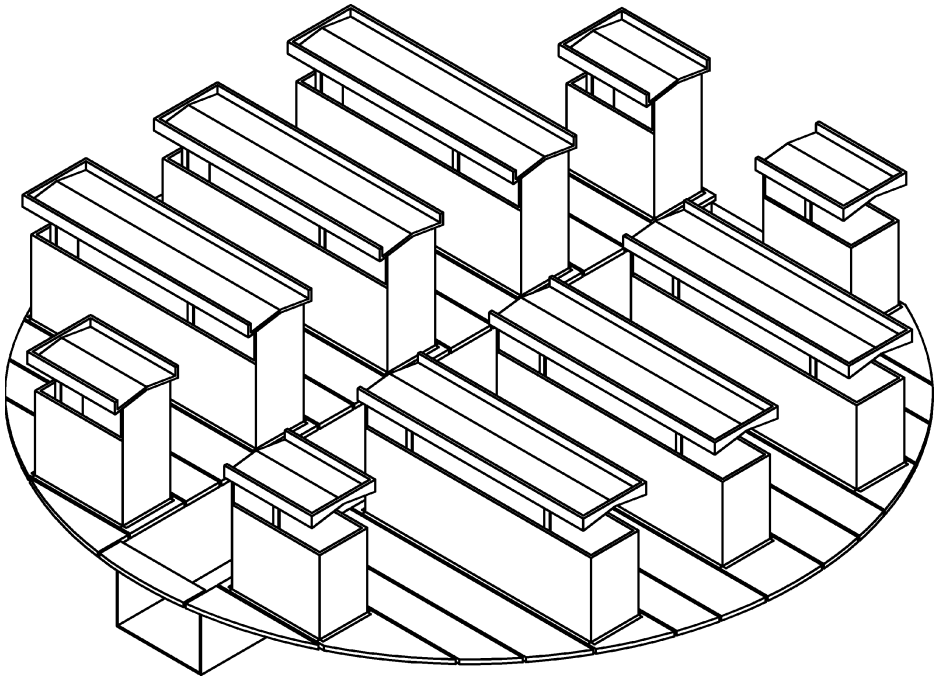


Fig. 9 Chimney collector tray.

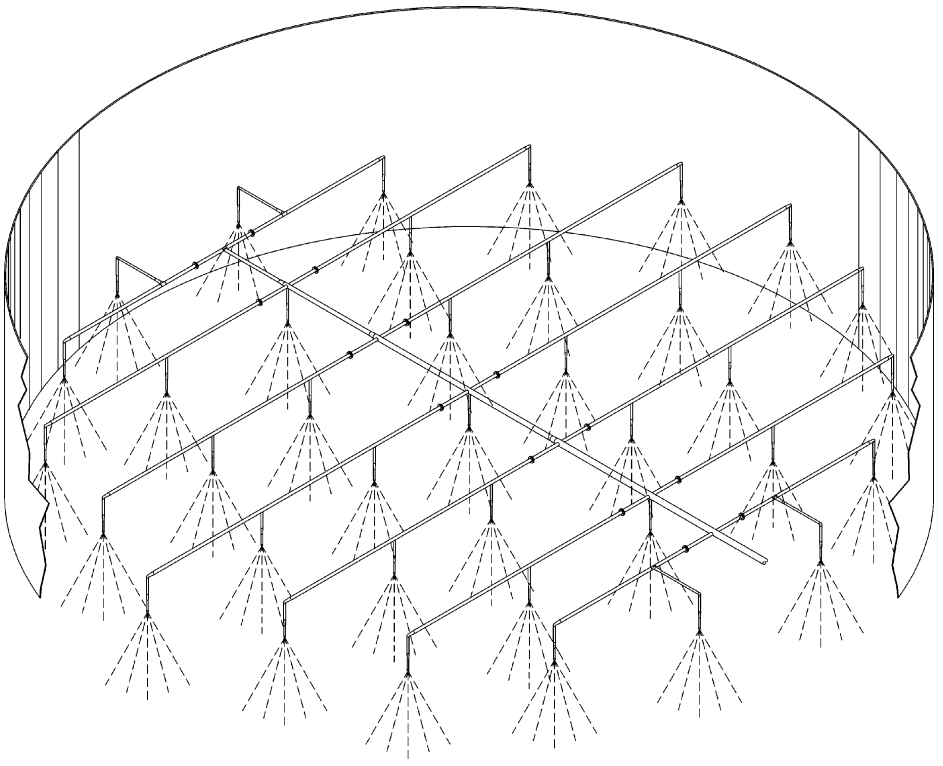


Fig. 10 Spray nozzle distributor.

bed. Two main types are available, pan-orifice distributors and trough distributors.

Pan-orifice distributors depend on liquid height on a pan to provide sufficient head for the liquid to flow through holes in the pan to the packed bed below. Pan-orifice distributors also include vapor chimneys for vapor to rise past the tray.

Of critical importance in deciding on the suitability of pan-orifice distributors is the degree of levelness expected in the installation and in the quality of distribution required. Flow through the orifice holes in the pan varies with the square root of the liquid height above the trap minus the pressure drop of the vapor through the distributor (expressed in height of liquid).[16] All liquid on the orifice pan is in one pool. When a liquid product is drawn from an orifice pan, a liquid gradient across the pan is set up. Moore and Rukovena[17] present the basic calculation methods for determining the liquid gradient on a pan-orifice distributor.

Whenever a product draw is taken from a collector pan, the liquid flow through any orifice hole depends on the local height. This guarantees that some degree of liquid maldistribution will take place. Additionally, orifice-pan distributors are installed on tray rings welded to the tower wall. If the tray ring is not level, the orifice pan will have liquid maldistribution caused by the out-of-level installation.

To avoid maldistribution due to out-of-level and by liquid draws, some manufacturers install tubes through the pan floor. This is intended to reduce flow maldistribution by guaranteeing a minimum liquid height for lateral flow toward the product draw on the pan during all operations. While this is an improvement, it is not an effective solution for the inherent limitations of this type of device.

Trough distributors (Fig. 11) use a preliminary distribution box (parting box) to meter a proportional flow to each of a series of troughs below the parting box. Each trough has a number of drip holes punched along the side. The liquid height required in both the parting box and the trough is a function of the flow range required. A typical flow range design for a trough distributor with round holes is a 2 : 1 turndown ratio. At a minimum height of 4 in. (100 mm) in the parting box (to avoid flow gradients along the box) a 2 : 1 flow range requires a 16 in. (400 mm) tall parting box. Hole areas above each trough are punched proportionate to the total area of the tower that the trough serves.

Slotted and v-notched distributor designs should be avoided. These designs are subject to large variations in local liquid distribution because of small manufacturing and installation errors.

Vapor distributors ensure a smooth vapor distribution before the vapor enters the packed bed. They fall into two classes. The first class consists of those that distribute the vapor flow profile by imposing a pressure drop on the vapor steam. The second class consists of devices that mechanically redistribute the vapor flow by means of vanes.

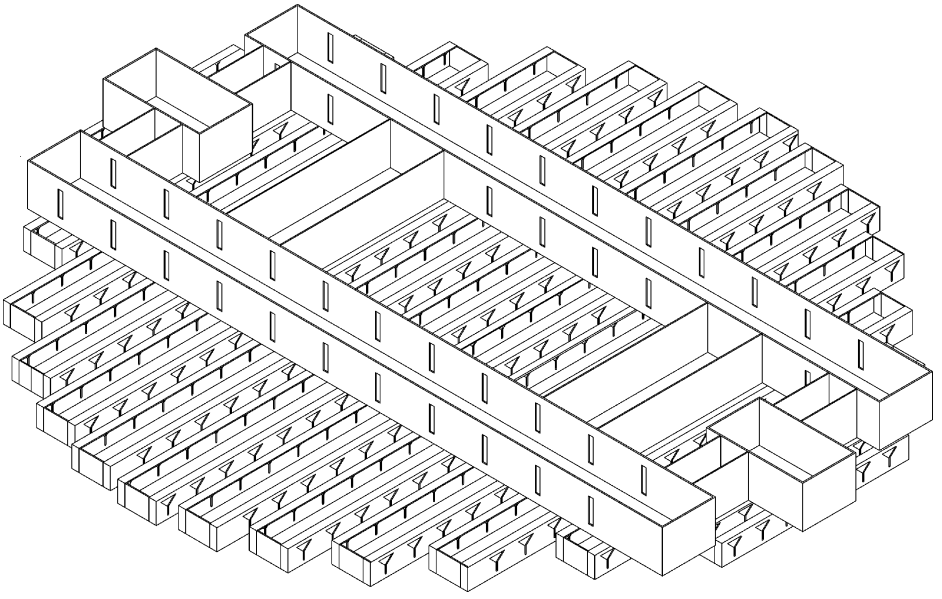


Fig. 11 Trough-orifice distributor.

Pressure vapor distribution devices are trays with orifice chimneys. They are usually constructed integrally with a liquid collector. The main difference between a liquid collector designed for liquid collection only and one designed for vapor redistribution as an additional function is in the open area available for vapor flow and the positioning of the vapor chimneys. Orifice plates welded to the tray deck under the vapor risers impose the specified pressure drop. Spacing between the top of the chimney hats and the packed bed above is set so that the angle from the hat to the bed required to give full diffusion of the rising vapor is 15° or less. Standard practice is to use chimney trays with a large number of small hats in vapor distribution service. This helps to ensure an even distribution of the vapor to the packed bed above.

Vane vapor distributors use directional vanes to ensure vapor distribution of a vapor stream. Vane-type distributors are commonly used for distribution of high-velocity, mixed-phase feeds. Various proprietary and patented designs are in use. Both radial entry and tangential entry designs are available. Tangential entry designs are the better choice for two-phase flows. However, radial entry designs will work acceptably for most applications.

CONCLUSIONS

Fractional distillation, absorption, and stripping are the most common separation unit operations. All these services can use packing to achieve their mass-transfer objectives. Effective hydraulic and mass-transfer design for packing involves selection of both the packing and the auxiliary internals. The other internals include packing supports, hold-downs, and liquid and vapor distributors. When properly selected, packing meets the combined process and mechanical requirements of the service. The process and mechanical requirements are not independent. The equipment must work, and it must work as part of a process to achieve plant objectives. Applications for packing cover the gamut of the petroleum refining, petrochemical, gas processing, and chemical industries. The general design criteria here cover standard design issues for new towers. Revamps often involve extremely complex balancing of mechanical possibilities vs. process risk to meet economic and operating targets.

NOMENCLATURE

A_{CS}	Total tower cross-sectional area
A_{CSF}	Tower cross-sectional area at flooded operation
c_{SB}	*c*-factor (Souders–Brown factor)
ΔP_f	Pressure drop at flood, packing
ΔP_{fnf}	Pressure drop at flood for a nonfoaming system, packing
F_F	Flooding fraction
F_{PX}	Factor, packing
H_P	Height of packing
L	Total liquid flow (mass or volume units as applicable)
L_F	Load parameter
L_{SG}	Liquid specific gravity
μ_L	Liquid absolute viscosity, (centipoises)

ν_L	Liquid kinematic viscosity, (centistokes)
N_t	Number of theoretical stages
ρ_L	Density, liquid
ρ_V	Density, vapor
σ	Surface tension (dyn/cm)
S_F	System factor
V	Total vapor flow (mass or volume units as applicable)
V_s	Velocity, superficial vapor velocity through the tower's entire cross section
X	Flow parameter

REFERENCES

1. Strigle, R.F. Jr. *Packed Tower Design*, 2nd Ed.; Gulf Publishing Company: Houston, TX, 1994.
2. Kister, H.Z.; Gill, D.R. Flooding and pressure drop prediction for modern random packings. AIChE Spring National Meeting, Orlando, FL, Mar 18–22, 1990.
3. Kaiser, V. Correlate the flooding of packed columns a new way. Chem. Eng. Prog. **1994**, *90* (6), 55–59.
4. Eckert, J.S. Design techniques for sizing packed towers. Chem. Eng. Prog. **1961**, *57* (9), 54–58.
5. Eckert, J.S. Tower packings ... comparative performance. Chem. Eng. Prog. **1963**, *59* (5), 76–82.
6. Eckert, J.S. Selecting the proper distillation column packing. Chem. Eng. Prog. **1970**, *66* (3), 39–44.
7. Eckert, J.S. How tower packings behave. Chem. Eng. **1975**, *82* (8), 70–76.
8. Robbins, L.A. Improve pressure-drop prediction with a new correlation. Chem. Eng. Prog. **1991**, *87* (5), 87–91.
9. Kister, H.Z.; Larson, K.F.; Gill, D.R. More interpolation charts for predicting packing flood and pressure drop. AIChE Spring National Meeting, Houston, TX, Mar 19–23, 1995.
10. Sloley, A.W.; Martin, G.R. Subdue solids in distillation. Chem. Eng. Prog. **1995**, *92* (1), 64–73.
11. Kister, H.Z. *Distillation Operation*; McGraw-Hill, Inc.: New York, 1990.
12. Kister, H.Z. *Distillation Design*; McGraw-Hill, Inc.: New York, 1992.
13. Norton Chemical Process Products. *Design Information for Packed Towers*, Bulletin DC-11; Norton Chemical Process Products: Akron, OH, 1976.
14. NSW Environmental Products. *Nor-Pac® High Performance Tower Packing;* NSW Environmental Products: Roanoke, VA, 2001.
15. Norton Chemical Process Products. *IMTP® High-Performance Random Packing*; Norton Chemical Process Products: Akron, OH, 1996.
16. Bonilla, J.A. Don't neglect liquid distributors. Chem. Eng. Prog. **1993**, *90* (3), 47–61.
17. Moore, F.; Rukovena, F. Liquid and gas distribution in commercial packed towers. 36th Canadian Chemical Engineers Conference, Calgary, Oct 1986.

Distillation Column Design: Trays

Andrew W. Sloley
VECO USA, Inc., Bellingham, Washington, U.S.A.

INTRODUCTION

The objective of mechanical layout and design of a distillation column is to construct and install equipment that creates the conditions that will meet the process design objectives of a distillation service. However, mechanical design considerations and process design objectives interact. Balancing mechanical convenience vs. process needs minimizes investment for any given process. Process design must take equipment limitations into account. Process design cannot require physically impossible equipment performance. Mechanical design must meet the process objectives. Mechanical convenience cannot override process necessities.

The general design guidelines here apply to liquid–vapor trays for heat-transfer and mass-transfer services. In addition to fractional distillation this includes many stripping, absorption, and other services.

This discussion covers a range of common equipment situations. The objective of the discussion is to allow for design sizing (overall height and diameter) of a straightforward, new distillation column. Complex columns and modified columns present many challenges requiring specialized application of design criteria. The references direct the engineer to an in-depth discussion of special requirements.

WHEN TO USE TRAYS

The main distillation devices are the tray and packing. Both trays and packing come in different types. Device selection requires balancing many competing factors. Table 1 lists the factors that favor each type of equipment. Final equipment selection is usually service specific.

For new tower design, without overriding process factors, the cheapest overall system (tower, foundation, internals) is normally a sieve or a valve tray. Both have essentially the same capacity. The driving force behind valve tray selection is the flexibility benefit of the valve tray at little additional cost. However, for larger diameters and higher liquid rates, the flexibility advantage of the valve tray is more illusory than factual.[1]

TRAY OPERATION

Basic Tray Operation

Tray contacting devices include both downcomer trays and downcomer-less trays. Trays with downcomers constitute the vast majority of tray installations. Fig. 1 shows a typical downcomer tray.

Liquid from the tray above descends a downcomer to the tray below. In the downcomer, the vapor in the liquid leaving the active area de-entrains. The downcomer inlet panel deflects the liquid sideways to the active area. Mass-transfer contacting devices on the active area intimately mix the liquid and rising vapor. The aerated liquid leaves the tray by entering the downcomer. In the space above the active area, the entrained liquid falls out of the vapors and returns to the tray deck. Fig. 1 shows the function of the tray zones.

The mass-transfer devices may be sieves (holes), fixed valves, moveable valves, or bubble caps. Fig. 2 shows a selection of mass-transfer devices.[2,3] The purpose of the device is intimate mixing of the vapor and liquid on the tray deck. An ideal device has high capacity, high flexibility, low leakage, low pressure drop, and low cost.

Vapor mixing into the liquid through the mass-transfer device creates the tray's active area. The active area may have either liquid as the continuous phase (froth regime) or vapor as the continuous phase (spray regime).

TRAY LIMITATIONS

Depending on tray type, at least 10 different operating modes can limit tray operation. Entrainment flood (jet flood) and downcomer filling set the upper capacity limit. Weeping and dumping set the lower operating limit. Other limits, including downcomer choking, system limit, blowing, and downcomer sealing constraints, restrict operation in specific situations.

Jet flooding includes situations where the rising vapor velocity is sufficient to prevent liquid droplets from disengaging from the vapor above the tray deck. Jet flooding can be reduced by increasing tray spacing.

Downcomer flooding includes situations where the liquid cannot get down the downcomer. Depending on the cause it can be reduced by increasing tray

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120041061

Table 1 Device selection criteria

Device	Capacity	Flexibility	Pressure drop	Notes[a]	Cost
Packing					
First-generation random	Low	Moderate	Medium	1	Inexpensive
Second-generation random	Medium	Moderate	Medium-low	1	Inexpensive
Third-generation random	High	Moderate	Low	1, 5	Moderate
Fourth-generation random	Very high	Moderate	Low	1, 5	Moderate
Structured	Very high	Moderate	Very low	1, 5	High
Grid	Very high	Moderate	Very low	1	Moderate
Trays					
Valve	High	Very good	High-medium	2	Moderate
Valve, caged	High	Excellent	High		High-moderate
Valve, fixed	High	Good	Medium		Moderate
Sieve	High	Good	Medium-low	2	Moderate-low
Dual flow	High	Very poor	Medium-low	3, 4	Moderate
Film	Low	Moderate	Low	5	High
Bubble cap	Medium	Excellent	High		High
Baffle (shed)	Very high	Poor	Very low	4, 6	Low

[a]1: Flexibility is normally limited by the liquid distribution system. 2: Most common types of tray. 3: No downcomer. 4: Very fouling services. 5: Vacuum services. 6: Very high liquid rates.

spacing, increasing downcomer volume, and reducing tray pressure drop.

Weeping and dumping set the lower operating limit of most tray designs. Weeping is liquid droplets bypassing the tray active area by falling through the contacting devices.

TRAY DESIGN

Introduction

Bubble-cap trays are rarely used in modern tower installations unless service specific reasons dominate the selection criteria. Based on the author's experience, valve tray installations outnumber sieve trays by approximately three to one. For these reasons, the tray design procedure given here is for valve trays. The most common valve tray methods are based on the Glitsch, Koch Engineering, and Nutter Engineering correlations.[4–6] Other, less commonly used methods are also available.[7]

For entrainment flooding (jet flood), all the correlations give useable results though the Glitsch correlation appears to be slightly more accurate than the Koch method. Both the Glitsch and the Koch correlations have a tighter fit to the data than the Nutter correlation. Given the subjective nature of judging the flood point, most statistical differences between the correlations are

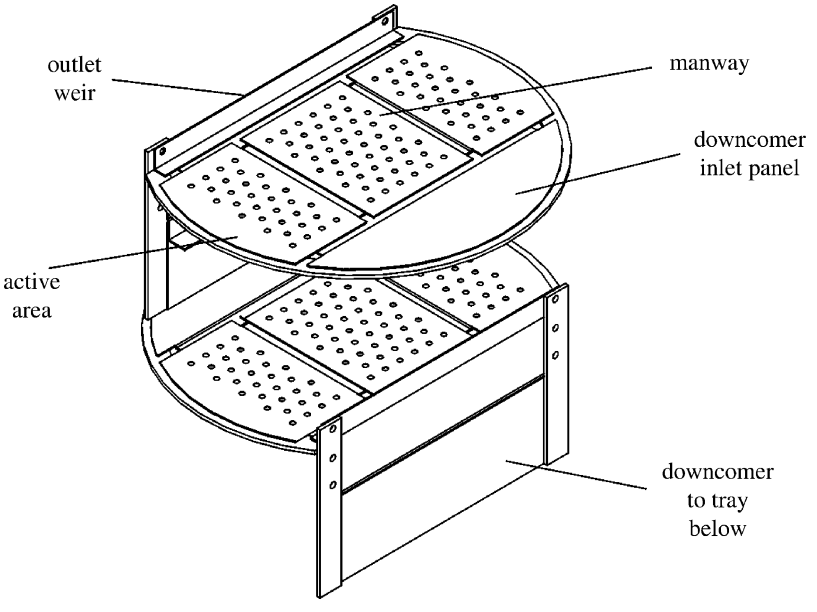


Fig. 1 Typical downcomer tray and functional areas.

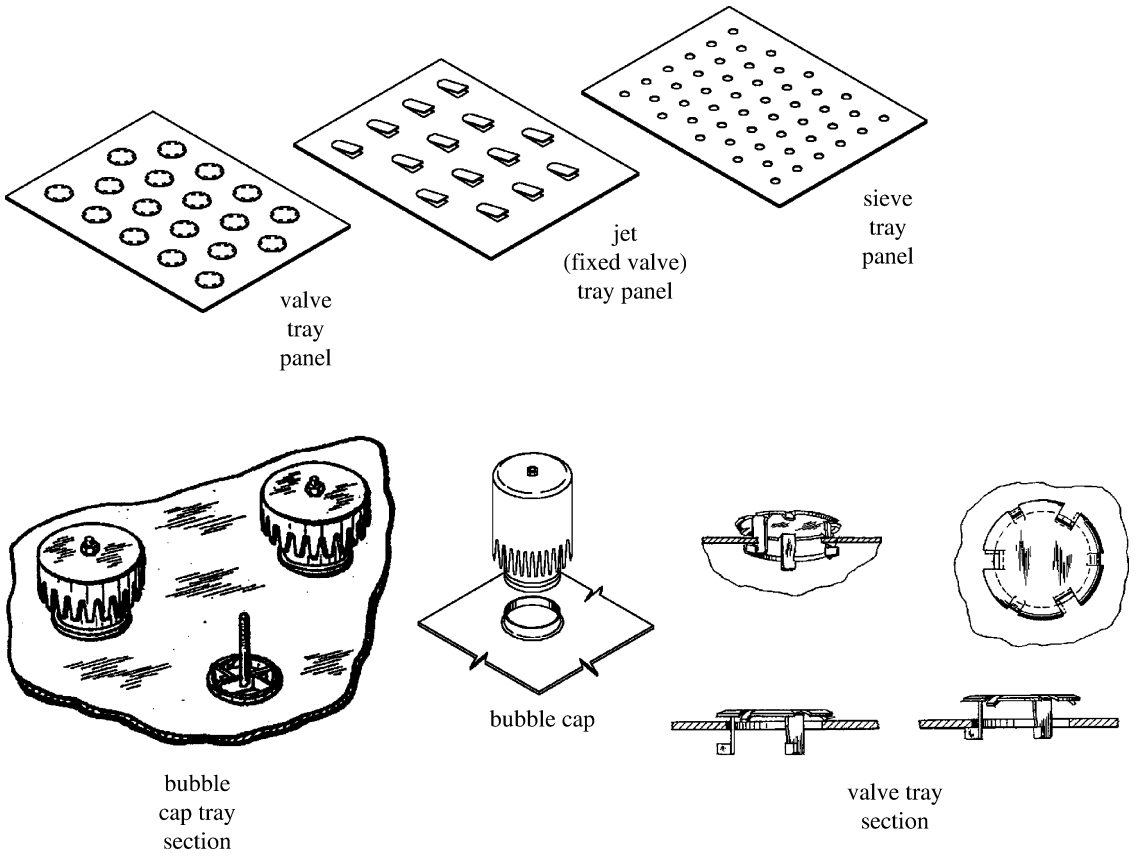


Fig. 2 Mass-transfer devices: bubble caps and valve sections. (From Refs.[2,3].)

not significant. Capacity differences between the different standard valve types are minimal.

In contrast, the literature methods for downcomer flooding have poor accuracy. Error bands for correlation results show poor results. Proprietary methods, including the FRI and TZF methods, give excellent results but are not generally available.[8]

The method shown is a modified version of the Glitsch method. Jet flood results for low-vapor-density operation ($<1.5\,lb/ft^3$, $<24\,kg/m^3$) are generally accurate. Application of the method is valid for tray spacing between 12 and 48 in. (30–122 cm).

Tray design is complex and design optimization involves many steps and decisions. Entire books have been written on the subject. The procedure here is for a general system with no special features. For those investigating the subject in detail the standard references are only a starting point for detailed discussions.[9–11]

Design Outline

The general design procedure for a new tower is:

1. Set design basis
2. Preliminary design of the tower: set diameter, tray spacing, and tray passes
3. Detailed design and pressure balance: iterative procedure using the preliminary design as a starting point.

Set Design Basis

Before starting work on the tray design obtain the following:

1. Internal liquid and vapor traffic inside the column
2. Vapor and liquid density.

The method shown does not take into account surface tension and viscosity effects. Other detailed methods may require these.

The required liquid and vapor rates are the amount of vapor entering the tray and the amount of liquid leaving the tray. This combination gives the correct pressure balance for the liquid inventory in the downcomer and the vapor pressure drop loss through the tray (Fig. 3). Today, engineers most often use simulation packages to calculate the vapor and liquid traffic. Care should be taken to extract the appropriate

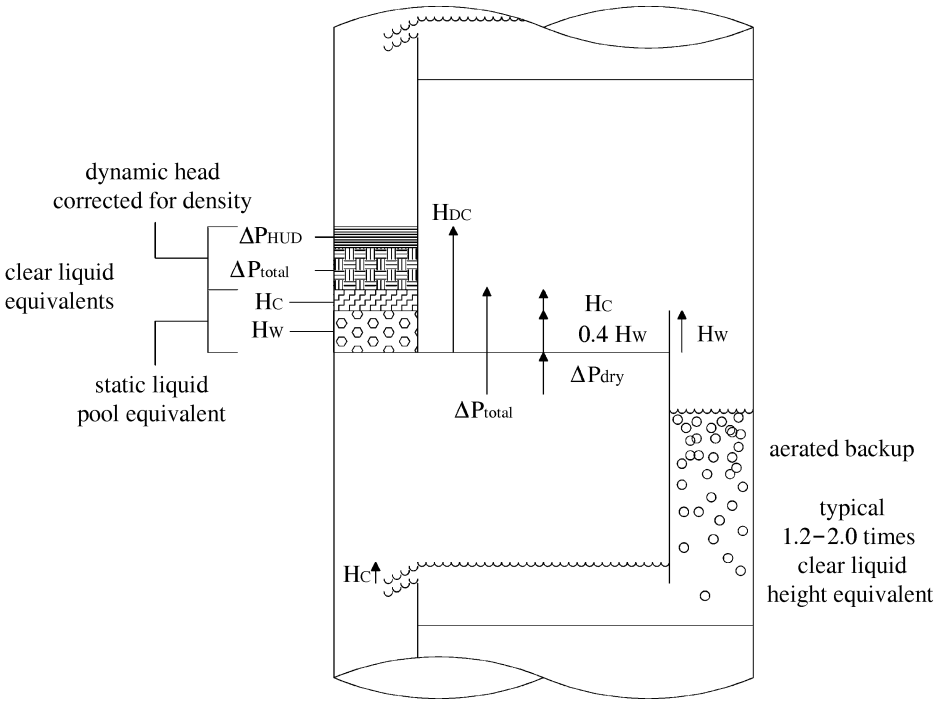


Fig. 3 Tray pressure balance—vapor entering the tray and liquid leaving the tray.

total stream rates and densities at operating conditions from the simulation. Many columns have failed as a result of taking incorrect flow rates or properties from the simulation.

Most columns use a 24 in. (61 cm) tray spacing for access room during inspection and installation. Use 24 in. (61 cm) as the starting point unless the column is expected to be very tall (over 100 stages), use very expensive materials of construction, or have a very large diameter. For tall columns made of expensive materials consider a shorter tray spacing and for columns with a very large diameter consider a taller tray spacing.

Preliminary Design

1. Determine the foaming tendency of the system. Pick a system factor, S_F, based on system experience or use typical values from Table 2. This factor will apply to both the downcomer and the active area design.
2. Estimate a design downcomer velocity using the following three equations and take the lowest value:

$$V_{DD} = K_1 S_F \quad (1)$$

$$V_{DD} = K_2 \sqrt{\rho_L - \rho_V} \quad (2)$$

$$V_{DD} = K_3 \sqrt{T_S} \sqrt{\rho_L - \rho_V} \quad (3)$$

Where V_{DD} is in feet per second, ρ is in pounds per cubic feet, T_S (tray spacing) is in inches, then $K_1 = 0.557$, $K_2 = 0.0914$, and $K_3 = 0.0167$. Where V_{DD} is in meters per second, ρ is in kilograms per cubic meter, T_S (tray spacing) is in centimeters, then $K_1 = 0.170$, $K_2 = 0.00696$, and $K_3 = 0.000799$.

3. Select a downcomer target flooding fraction, F_D. A typical design value for a new column would be 0.80 (80%). This value does not have to be the same as the jet flood design value.
4. Determine the minimum design downcomer area:

$$A_{DM} = \frac{L}{V_{DD} F_D} \quad (4)$$

where L and V_{DD} have consistent units of cubic feet per second and feet per second or cubic meters per second and meters per second, respectively. If L is in gallons per minute (gpm), multiply L by 0.00223 to convert to cubic feet per second.

5. Calculate the column vapor load:

$$V_{load} = V \sqrt{\frac{\rho_V}{\rho_L - \rho_V}} \quad (5)$$

where V_{load} and V are in cubic feet per second or cubic meters per second, and ρ is in consistent units.

Table 2 System factors

Service	System factor
General systems	
Nonfoaming	1.0
Low foaming	0.9
Moderate foaming	0.85
High foaming	0.70
Severe foaming	0.60
Foam stable	0.30
Specific systems	
C3–C6 hydrocarbons	1.0
Low-molecular-weight alcohols	1.0
Atmospheric crude petroleum	1.0
Oxygen stripper	1.0
Rectisol contactor	1.0
Halogens	0.90
Cryogenic gas plant absorbers	1.0–0.85
Oil absorbers	0.85
Amine regenerators	0.85
Glycol regenerators	0.85
Water strippers	0.70–0.80
Amine absorbers	0.70
Glycol absorbers	0.70
Methyl ethyl ketone (MEK) units	0.60
Caustic regenerators	0.30

6. Calculate the dry tray capacity factor:

$$c_{AFO} = T_S^{1/2}(K_4 - K_5\rho_V) \quad (6)$$

or

$$c_{AFO} = K_6 T_S^{0.65}\rho_V^{1/6} \quad (7)$$

if vapor density is less than $0.17\,lb/ft^3$ ($2.72\,kg/m^3$). Where c_{AFO} is in feet per second, T_S is in inches, and ρ is in pounds per cubic feet, then $K_4 = 0.0933$, $K_5 = 0.00293$, and $K_6 = 0.0833$. Where c_{AFO} is in meters per second, T_S is in centimeters, and ρ is in kilograms per cubic meter, then $K_4 = 0.0178$, $K_5 = 0.0000350$, and $K_6 = 0.00873$.

7. Select an active area target flooding fraction, F_A. A typical design value for a new column would be 0.80 (80%). This value does not have to be the same as the downcomer flood design value. Optimized designs use the same value, F_F, for both jet flood and downcomer flood.
8. Estimate a required active area using the following two equations and taking the highest value:

$$A_{AM} = K_7\frac{V_{load}}{S_F} \quad (8)$$

$$A_{AM} = \frac{V_{load} + K_8 L}{c_{AFO}F_A S_F} \quad (9)$$

Where A_{AM} is in square feet, c_{AFO} is in feet per second, and V_{load} and L are in cubic feet per second, then $K_7 = 4.00$ and $K_8 = 0.828$. If L is in gallons per minute, multiply L by 0.00223 to convert to cubic feet per second. Where A_{AM} is in square meters, c_{AFO} is in meters per second, and V_{load} is in cubic meters per second and L is in cubic meters per hour, then $K_7 = 13.1$ and $K_8 = 0.828$.

9. Check the minimum downcomer area. If A_{DM} is less than $0.11A_{AM}$ then modify the minimum downcomer area to the smaller of:

$$A_{DM} = 0.11A_{AM} \quad (10)$$

$$A_{DM} = 2A_{DM} \quad (11)$$

[from Eq. (4)].

10. Calculate the tower cross-sectional area.

$$A_{TM} = A_{AM} + 2A_{DM} \quad (12)$$

Calculate the tower diameter:

$$D = \sqrt{1.27A_{TM}} \quad (13)$$

Detail Design and Rating

This procedure is for approximate design and rating of one-pass or two-pass valve trays. Again, the procedure is based on the Glitsch design method. In design, tray sizes can be restricted by mechanical considerations for valve pattern and layout and minimum panel sizes for internal man-ways. Steps 1–3 establish the basic layout of the downcomer for designing cases. This starts with the areas calculated in the preliminary design procedure. For rating an existing tray, skip to Step 4 and use the existing tray dimensions.

1. Select the general pass configuration, one-pass or two-pass. Fig. 4 shows the tray layout and terminology. Higher liquid loads favor two-pass design. Acceptable liquid load ranges are 2–10 gpm/in. of weir length ($0.0267–0.134\,ft^3/sec/ft$, $8.94–44.7\,m^3/h/m$). Rates below 1 gpm/in. ($0.0134\,ft^3/sec/ft$, $4.47\,m^3/h/m$) and above 18 gpm ($0.240\,ft^3/sec/ft$, $80.5\,m^3/h/m$) should be strenuously avoided. Straight, chordal weir lengths can be determined from downcomer widths (rises) by:

$$W = 2\sqrt{2RH_1 - H_1^2} \quad (14)$$

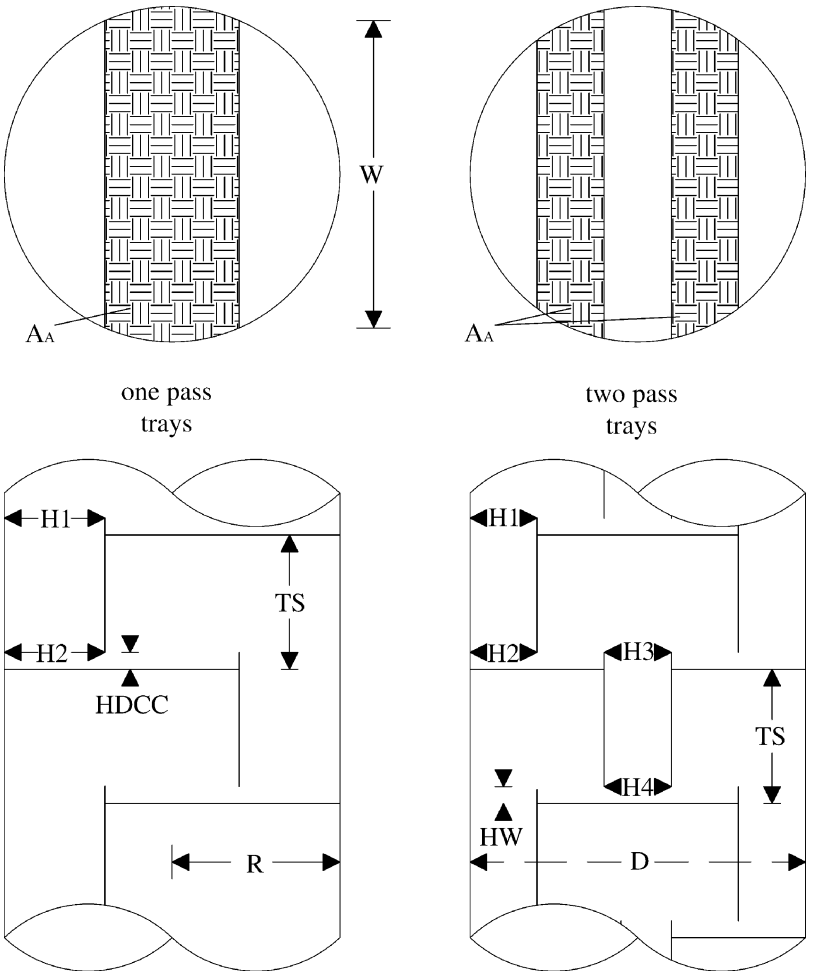


Fig. 4 Tray layout and terminology. One-flow-path and two-flow-path trays.

or downcomer areas by:

$$S - \sin S = \frac{2A_{DC}}{R^2} \tag{15}$$

and then

$$W = 2R \sin\left(\frac{S}{2}\right) \tag{16}$$

where S is in radians.

Eqs. (14–16) are for one-pass trays. For two-pass trays the total downcomer area is halved before calculating the weir length. The weir length is then doubled to account for both downcomers. Two-pass tray weir loads are limited by the liquid flow into the downcomers on the tower edge (outboard downcomers).

2. Dimension downcomers: Based on the areas and the downcomer, active areas and total areas determine the dimensions H_1, H_2, and L_{FP} for one-pass trays as:

$$S - \sin S = \frac{2A_{DC}}{R^2} \tag{15}$$

where S is in radians. Then,

$$H_1 = H_2 = R\left(1 - \cos\frac{S}{2}\right) \tag{17}$$

($H_1 = H_2$ for straight, vertical downcomers.)

$$L_{FP} = D - H_1 - H_2 \tag{18}$$

For two-pass trays, Eqs. (14) and (16) define H_1 for the outboard downcomer and half the total downcomer area A_{DC} is used. H_3, the center downcomer width, can be calculated by assuming the center downcomer is a rectangle. This gives good results for nearly all towers:

$$H_3 = H_4 = \frac{A_{DC}}{D} \tag{19}$$

($H_3 = H_4$ for straight, vertical downcomers.)

$$L_{FP} = \tfrac{1}{2}(D - H_1 - H_1 - H_3) \tag{20}$$

3. Check the downcomer and flow path length dimensions. Internal man-ways inside the tower require a minimum flow path length of 16 in. (40.6 cm). If the flow path length is less than

16 in. (40.6 cm) and internal man-ways are required, increase the tower dimensions to get a 16 in. (40.6 cm) F_{PL}. If the downcomer widths are less than 6 in. (15 cm) consider increasing the tower size to get a 6 in. (15 cm) wide downcomer. Update the flow path length, L_{FP}, and downcomer dimensions, and tower diameter as required.

4. Calculate the downcomer area. For outboard (side downcomers):

$$A_{DC1} = R^2 \cos^{-1}\left(\frac{R - H_1}{R}\right) - (R - H_1)\sqrt{2RH_1 - H_1^2} \quad (21)$$

An alternate approximation gives areas with a less than 1% error:

$$W = 2\sqrt{2RH_1 - H_1^2} \quad (14)$$

$$A_{DC1} \approx \frac{2}{3}WH_1 + \frac{H_1^3}{2W} \quad (22)$$

Use the entire area for one-pass trays; for two-pass trays, use half the total area for A_{DC}. For the inboard (center) area of two-pass trays use the approximation:

$$A_{DC3} = H_3 D \quad (23)$$

5. Calculate the active area:

$$A_A = A_{CS} - 2A_{DC1} \quad (24)$$

for one-pass trays or

$$A_A = A_{CS} - 2A_{DC1} - A_{DC3} \quad (25)$$

for two-pass trays.

6. Calculate vapor load:

$$V_{load} = V\sqrt{\frac{\rho_V}{\rho_L - \rho_V}} \quad (5)$$

where V_{load} and V are in cubic feet per second or cubic meters per second, and ρ is in consistent units.

7. Calculate the tray capacity factor, c_{AF}:

$$c_{AFO} = T_S^{1/2}(K_4 - K_5\rho_V) \quad (6)$$

or

$$c_{AFO} = K_6 T_S^{0.65}\rho_V^{1/6} \quad (7)$$

if vapor density is less than 0.17 lb/ft^3 (2.72 kg/m^3). Where c_{AFO} is in feet per second, T_S is in inches, and ρ is in pounds per cubic foot, then $K_4 = 0.0933$, $K_5 = 0.00293$, and $K_6 = 0.0833$. Where c_{AFO} is in meters per second, T_S is in centimeters, and ρ is in kilograms per cubic meter, then $K_4 = 0.0178$, $K_5 = 0.0000350$, and $K_6 = 0.00873$.

$$c_{AF} = c_{AFO}S_F \quad (26)$$

8. Calculate the jet flooding fraction and jet flooding percentage. This is equivalent to the Glitsch Equation 13[4] for flooding:

$$F_F = \frac{V_{load} + (LL_{FP}/K_9)}{A_A c_{AF}} \quad (27)$$

$$F_P = 100F_F \quad (28)$$

Where c_{AF} is in feet per second, V_{load} and L are in cubic feet per second, L_{FP} is in feet, A_A is in square feet, then $K_9 = 2.42$. Where c_{AF} is in meters per second, V_{load} is in cubic meters per second, L in cubic meters per hour, L_{FP} is in centimeters, A_A is in square meters, then $K_9 = 0.738$. Where c_{AF} is in feet per second, V_{load} is in cubic feet per second, L is in gallons per minute, L_{FP} is in inches, A_A is in square feet, then $K_9 = 13{,}000$.

9. Estimate the number of valves. Valve density varies between 12 and 14 per square foot depending on tray layout and geometry. Use an average of 13 valves per square foot (140 valves per square meter) or the actual number of standard valves ($1\frac{17}{32}$ in., 39 mm diameter) to rate a tray:

$$N_u = K_{10}A_A \quad (29)$$

Where A_A is in square feet, then $K_{10} = 13$. Where A_A is in square meters, then $K_{10} = 140$.

10. Calculate hole area in valves:

$$A_H = \frac{N_u}{K_{11}} \quad (30)$$

Where A_H is in square feet, then $K_{11} = 78.5$. Where A_H is in square meters, then $K_{11} = 845$.

11. Calculate hole vapor velocity:

$$V_H = \frac{V}{A_H} \quad (31)$$

12. Calculate dry tray pressure drop. The pressure drop is calculated for two cases, assuming valves full open and assuming valves part open. Use

the larger value of the two.

$$\Delta P_{\text{dry}} = K_{12} V_{\text{H}}^2 \frac{\rho_{\text{V}}}{\rho_{\text{L}}} \tag{32}$$

for valves full open

$$\Delta P_{\text{dry}} = K_{13} t_{\text{m}} \frac{\rho_{\text{M}}}{\rho_{\text{L}}} + K_{14} V_{\text{H}}^2 \frac{\rho_{\text{V}}}{\rho_{\text{L}}} \tag{33}$$

for valves part open.

Most trays are 10–14 gage for stainless steel and 12–16 gage for special alloys. Table 3 lists values of K_{12} for various tray deck thicknesses. Table 4 lists a selection of tray material densities and Table 5 lists conversions for gage in millimeters.

Where ΔP_{dry} is in inches of liquid at conditions, V_{H} is in feet per second, t_{m} is in inches, and ρ_{L}, ρ_{M}, and ρ_{V} are in consistent units, then the value of K_{12} is as given in Table 3, $K_{13} = 1.35$, and $K_{14} = 0.20$. Where ΔP_{dry} is in millimeters of liquid at conditions, V_{H} is in meters per second, t_{m} is in millimeters, and ρ_{L}, ρ_{M}, and ρ_{V} are in consistent units, then K_{12} is as given in Table 3, $K_{13} = 1.35$, and $K_{14} = 54.7$.

60. Check dry tray pressure drop. The dry tray pressure drop should be less than 20% of the tray spacing ($\Delta P_{\text{dry}} < 0.2 T_{\text{S}}$). If not, increase the tray spacing or increase the active area. In some cases, increasing the valve density may be acceptable.
61. Calculate the liquid crest over the outlet weir:

$$H_{\text{C}} = K_{15} \left(\frac{L}{W} \right)^{2/3} \tag{34}$$

Where H_{C} is in inches, L is in cubic feet per second, and W is in feet, then $K_{15} = 4.47$. Where H_{C} is in millimeters, L is in cubic meters per second, and W is in meters, then $K_{15} = 554$. Where H_{C} is in inches, L is in gallons per minute, and W is in inches then $K_{15} = 0.4$.

This is a version of the Francis weir formula without modification for weir height.[12]

Table 3 Dry tray pressure drop coefficients

Tray thickness (in.)	0.074	0.104	0.134	0.187	0.250
K_{12}	1.18	0.95	0.86	0.67	0.61
Tray thickness (mm)	1.88	2.64	3.40	4.75	6.35
K_{12}	323	260	235	183	167

Table 4 Metal densities

	Density	
Material	**(lb/ft^3)**	**(kg/m^3)**
Carbon steel	490	7,850
Stainless steel	500	8,010
Nickel	553	8,860
Monel	550	8,810
Titanium	283	4,530
Hastelloy$^{\text{TM}}$	560	8,970
Aluminum	168	2,690
Copper	560	8,970
Lead	708	11,340

Fig. 3 illustrates the pressure balance locations for the tray.

62. Calculate the total tray pressure drop. The total tray pressure drop is the dry tray pressure drop plus the height of the liquid on the tray deck. Aeration is assumed to reduce the tray liquid density to 40% of the nonaerated liquid density.

The weir height, H_{W}, usually varies from 1 to 3 in. (2.5–7.5 cm). Two inches (5 cm) is a typical value. For heights above 15% of the tray spacing reduce the effective tray spacing for jet flood by the excess over 15%:

$$\Delta P_{\text{total}} = \Delta P_{\text{dry}} + H_{\text{C}} + 0.4 H_{\text{W}} \tag{35}$$

ΔP_{total} is the height of liquid at conditions. To convert to other units:

$$\Delta P_{\text{total psi}} = \frac{\Delta P_{\text{total in. of liquid}} \rho_{\text{L}}}{1728} \tag{36}$$

$$\Delta P_{\text{total mmHg}} = \frac{\Delta P_{\text{total in. of liquid}} \rho_{\text{L}}}{33.3} \tag{37}$$

$$\Delta P_{\text{total kPa}} = \frac{\Delta P_{\text{total mm of liquid}} \rho_{\text{L}}}{101,900} \tag{38}$$

Table 5 Tray deck thickness

	t_{m}	
Gage	**(in.)**	**(mm)**
20	0.037	0.940
18	0.050	1.27
16	0.060	1.52
14	0.074	1.88
12	0.104	2.64
10	0.134	3.40

16. Calculate the downcomer escape area:

$$A_{DCE} = H_{DCC}W \tag{39}$$

For sloped downcomers adjust W (weir length) to the correct value for the downcomer bottom edge. Downcomer clearance, H_{DCC}, is typically a minimum of 1 in. (25.4 mm) and is usually 0.5 in. (12.7 mm) less than the outlet weir height to maintain a positive seal on the downcomer.

17. Calculate the velocity under the downcomer:

$$V_{UD} = \frac{L}{A_{DCE}} \tag{40}$$

18. Calculate the head loss under the downcomer:

$$\Delta P_{HUD} = K_{16}V_{UD}^2 \tag{41}$$

Where ΔP_{HUD} is in inches of liquid at conditions and V_{UD} is in feet per second, then $K_{16} = 0.65$. Where ΔP_{HUD} is in millimeters of liquid at conditions and V_{UD} is in meters per second, then $K_{16} = 178$. Alternately, ΔP_{HUD} may be calculated directly from liquid load and tray dimensions by:

$$\Delta P_{HUD} = K_{17}\left(\frac{L}{WH_{UD}}\right)^2 \tag{42}$$

Where ΔP_{HUD} is in inches of liquid at conditions, L is in gallons per minute, W and H_{UD} are in inches, then $K_{17} = 0.06$.

19. Calculate the downcomer backup:

$$H_{DC} = H_W + K_{18}\left(\frac{L}{W}\right)^{2/3} + (\Delta P_{total} + \Delta P_{HUD})\left(\frac{\rho_L}{\rho_L - \rho_V}\right) \tag{43}$$

Where H_{DC}, ΔP_{total}, and ΔP_{HUD} are in inches, L is in cubic feet per second, ρ is in consistent units, and W is in feet, then $K_{18} = 4.47$. Where H_{DC}, ΔP_{total}, and ΔP_{HUD} are in millimeters, L is in cubic meters per second, ρ is in consistent units, and W is in meters, then $K_{18} = 554$. Where H_{DC}, ΔP_{total}, and ΔP_{HUD} are in inches, L is in gallons per minute, ρ is in consistent units, and W is in inches, then $K_{18} = 0.4$.

The downcomer backup (clear liquid height basis) should be 40% or less of the tray spacing for high-vapor-density systems (>3.0 lb/ft^3, 48.1 kg/m^3), 50% or less for medium-vapor-density systems, and 60% or less for low-vapor-density systems (<1.0 lb/ft^3, 16.0 kg/m^3).

One of the major weaknesses in the method is that it does not account for varying vapor entrainment into the downcomer with changes in vapor rate.

20. Check design flexibility. The minimum vapor for no leakage through the tray is defined by the vapor kinetic energy to liquid density ratio, E:

$$E = V_H^2\frac{\rho_V}{\rho_L} \tag{44}$$

The acceptable value of E depends on the liquid height on the tray.

$$H_L = H_C + H_W \tag{45}$$

Table 6 shows the lower value of E for different liquid heights on the tray. This is the point at which liquid leakage through the valves begins. Values lower than that shown indicate leakage. This check does not calculate direct leakage rates.

Many trays function well even with some leakage. Leakage rates of up to 10% have a negligible effect on tray efficiency. Leakage up to 25% typically reduces tray efficiency by approximately 10%. Various strategies can be used to reduce leakage. These include changing valve types, using a mix of valves of different densities, and increasing tray spacing and using fewer valves (increase the hole velocity).

ADDITIONAL TRAY FEATURES

Specific service requirements may require additional tray features to improve operation. Many features result from either very high or very low liquid and vapor rates. High-liquid-rate design choices include

Table 6 Tray lower flexibility

Liquid level (in.)	1.0	1.5	2.0	2.5	3.0	3.5	4.0
E (ft^2/sec^2)	0.122	0.202	0.281	0.348	0.476	0.562	0.672
Liquid level (cm)	2.5	3.75	5.1	6.4	7.6	8.9	10.
E (m^2/sec^2)	0.0113	0.0188	0.0261	0.0323	0.0442	0.0522	0.0624

antijump baffles, modified arc and multichordal downcomers, and swept-back weirs. Low-liquid-rate design choices include picket fence weirs and splash baffles. High-vapor-rate choices include sloped downcomers and hanging downcomers. Low-vapor-rate trays may require careful allocation of holes or valves to prevent bypassing. Large flexibility range and solids handling also require many special design features.[13,14]

Multiple-pass trays use antijump baffles at high liquid rates to prevent downcomer choking (Fig. 5). On the center and offside downcomers for multiple-pass trays, large liquid jumps and vapor expansion in the outlet liquid can cause downcomer inlet choking. Deflecting the entering liquid prevents flooding. If the specific vapor load exceeds a tray liquid loading factor, then antijump baffles should be used.

$$W_{FP} = \frac{A_A}{L_{FP}} \tag{46}$$

If

$$\frac{V_{load}}{A_A} > K_{19} - K_{20}\left(\frac{L}{W_{FP}}\right) \tag{47}$$

use antijump baffles.

Where V_{load} is in feet per second, A_A is in square feet, L is in cubic feet per second, and W_{FP} is in feet, then $K_{19} = 0.336$ and $K_{20} = 7.18$. Where V_{load} is in cubic meters per second, A_A is in square meters, L is in cubic meters per second, and W_{FP} is in meters, then $K_{19} = 0.102$ and $K_{20} = 23.6$. Where V_{load} is in cubic feet per second, A_A is in square feet, L is in gallons per minute, and W_{FP} is in inches, then $K_{19} = 0.336$ and $K_{20} = 0.0192$.

Multichordal downcomers can increase weir length (Fig. 6). Fig. 6 shows a combined multichordal and stepped downcomer. The increased weir length decreases the liquid crest height over the weir. This decreases total tray pressure drop. The stepping is a variation of a sloped downcomer. The step increases the downcomer inlet area with a minimum decrease in tray active area.

Swept-back weirs increase the weir length and decrease the active area of the tray (Fig. 7). They are often the lowest-cost method to decrease weir load as long as sufficient vapor handling capacity is available.

Picket fence weirs are used in low-liquid-rate applications (Fig. 8). Picket fence weirs can serve two purposes at low liquid rates. First, they reduce the effective length of the weir for liquid flow increases the liquid height over the weir. This makes tray operation less sensitive to out-of-level installation. Second, pickets can prevent liquid loss (blowing) into the downcomer by spraying. This occurs at low liquid rates when the vapor is the continuous phase on the tray deck.[15] Picket fence weirs should be considered if the liquid load is less than 1 gpm per inch of weir ($0.0267\,ft^3/sec/ft$, $0.00248\,m^3/sec/m$). At liquid rates lower than 0.25 gpm per inch of weir ($0.00668\,ft^3/sec/ft$, $0.000620\,m^3/sec/m$) even picket fence weirs and splash baffles have a mixed record in improving tray efficiency. Operation at liquid rates this low strongly favors the selection of structured packing.

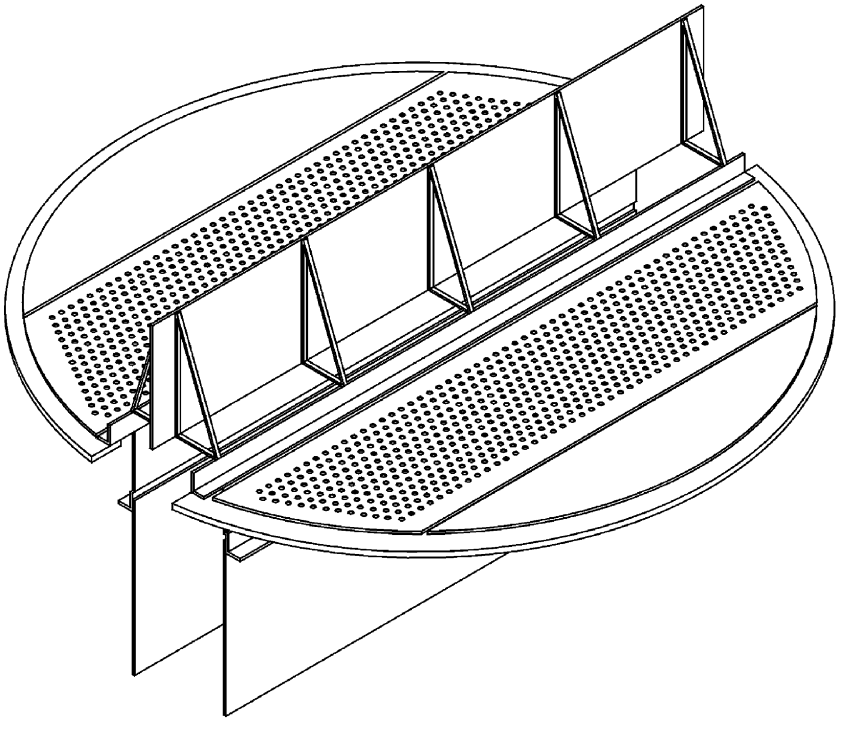


Fig. 5 Antijump baffle on sieve tray.

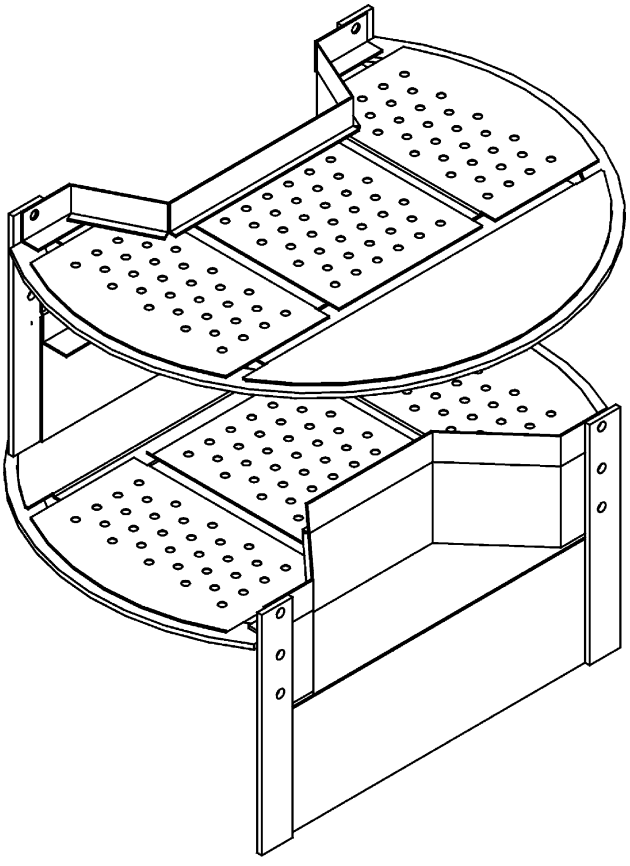


Fig. 6 Multichordal, stepped downcomer.

Picket fence weirs are also used to balance flow to the different passes on trays that have more than two flow paths. Splash baffles are an alternative to picket fence weirs when the major problem is blowing.

Sloped downcomers and modified arc-downcomers are traditional choices to increase the effective area of the tray for vapor–liquid contacting and increase vapor handling capacity (Fig. 9). In the last decade hanging downcomer trays have reached wide acceptance under the generic grouping of ''high-capacity'' trays. These trays can dramatically increase the capacity but are subject to more stringent design and operating limits.[16,17]

PASS SELECTION

Design criteria may force selection of multiple flow pass trays. Increasing the number of flow paths increases the weir length, hence decreasing the liquid load per length of weir. This reduces overall tray pressure drop for high-liquid-rate applications. Multiple flow pass trays also reduce the flow path length. Flow path lengths over 6 ft (1.83 m) start to create significant liquid gradients. Large liquid gradients have caused tower failures. No clearly accepted method for flow gradient prediction is available. Flow path lengths over 9 ft (2.74 m) should be avoided.

At the other extreme (small towers), mechanical limits constrain tray pass selection. Flow paths need to be long enough to install internal man-ways in the trays. Table 7 shows minimum tower diameters for different pass configurations.

TRAY EFFICIENCY

Tray efficiency is required to convert the number of theoretical stages identified as necessary to a number of real plates for installation in a tower. Many

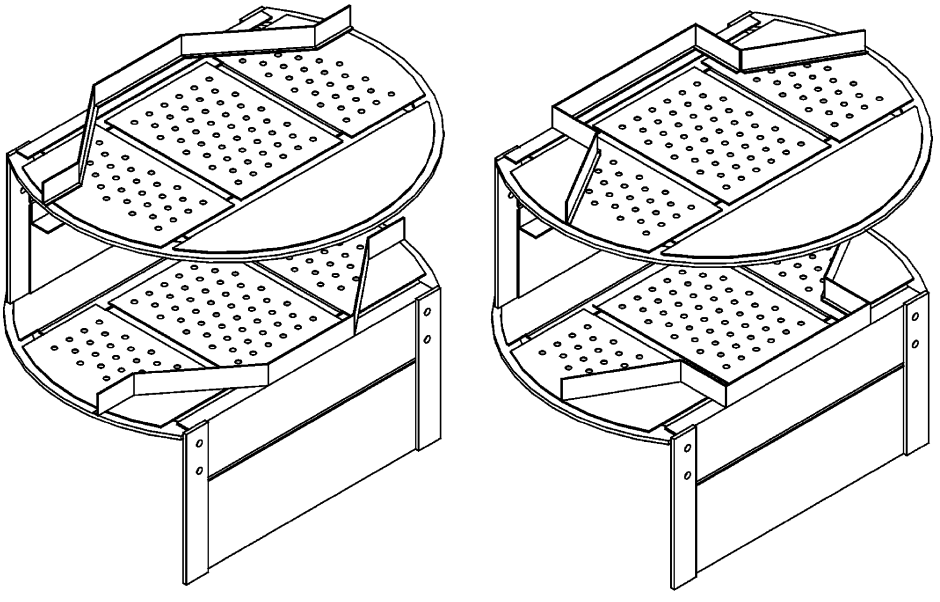


Fig. 7 Swept-back weirs.

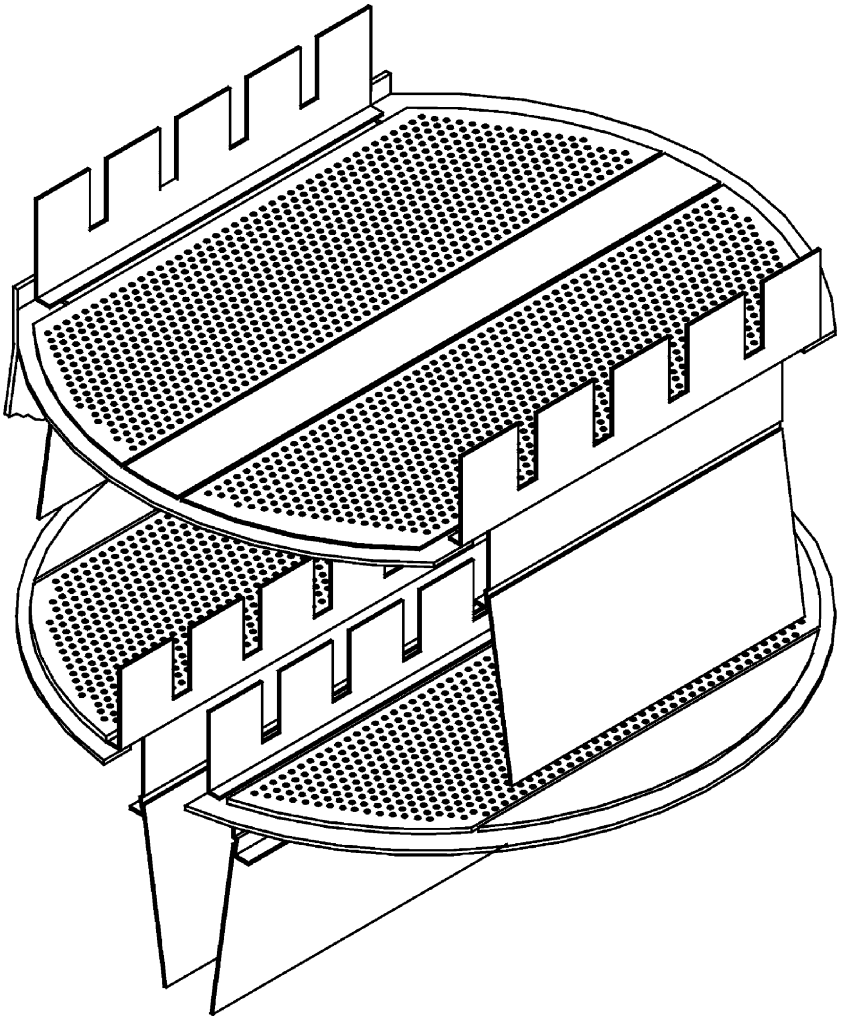


Fig. 8 Picket fence weirs.

theoretical and empirical methods have been proposed.[18–20] Apart from the O'Connell method, few have had wide acceptance for use in constructing towers.[12] The author's experience and the anecdotal experience of others shows the O'Connell method to give reasonable results, when used with a 10% safety factor, for a wide range of systems.

Multiple factors influence tray efficiency. These include the physical properties of the system, operating conditions, vapor and liquid loads, compositions, tray type, fabrication details of the tray, and many others. A range of experimental studies have examined tray efficiency under different conditions. These studies show often conflicting results for the influence of the mechanical and process conditions involved. Study results reveal few useful engineering guidelines for tray design and efficiency.

Two efficiency definitions are in common use for engineering design: overall tray efficiency and Murphree tray efficiency. Murphree tray efficiency basis can be set on either the vapor phase or the liquid phase. Most literature reports data on a vapor phase basis. Tray efficiency relations included here are on the vapor phase basis. Murphree point efficiencies are rarely used. However, some authorities recommend their use for specific situations and some literature includes point efficiency values.[21]

Overall tray efficiency is defined as the number of theoretical trays required for a service divided by the number of actual trays.

$$E_o = \frac{N_a}{N_t} \tag{48}$$

Murphree tray efficiency is defined as the approach to equilibrium on an individual tray by:

$$E_M = \frac{y_o - y_i}{y^* - y_i} \tag{49}$$

For an ideal separation at equilibrium and with straight operating lines, E_o and E_M are related by:

$$\lambda = m\frac{G_m}{L_m} \tag{50}$$

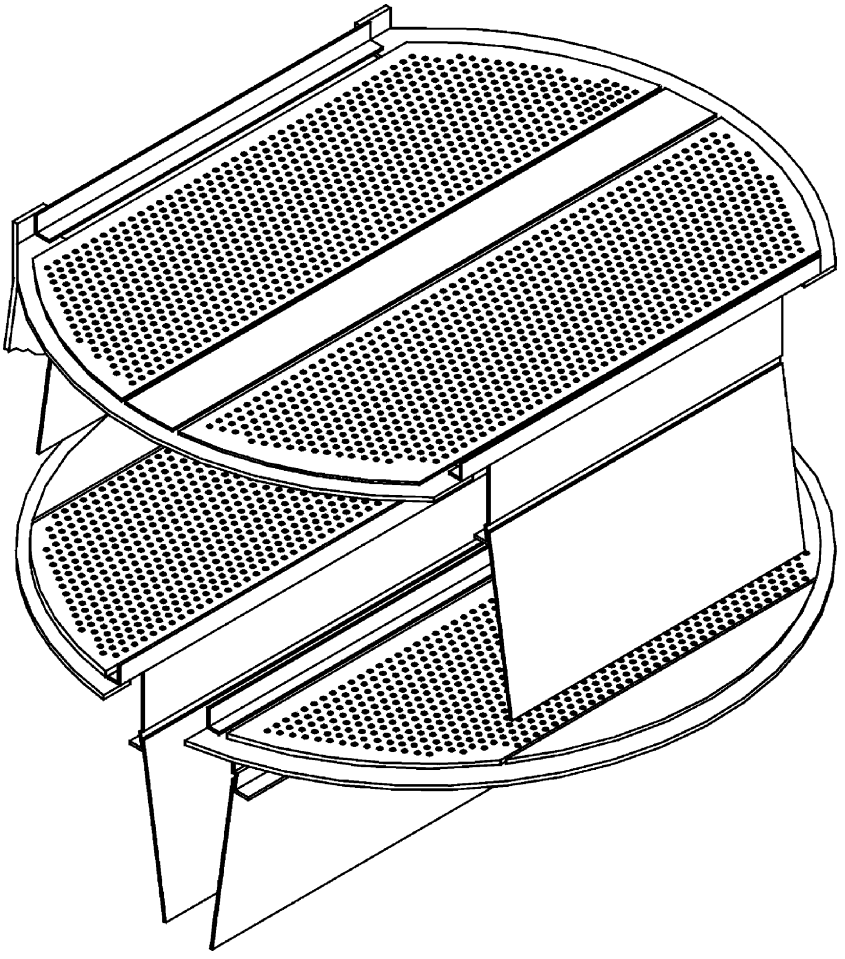


Fig. 9 Sloped downcomers.

(G_m and L_m are in consistent units.)

$$E_o = \frac{\ln(1 + E_M[\lambda - 1])}{\ln\lambda} \quad (51)$$

for $\lambda \neq 1$.

$$E_o = E_M \quad (52)$$

for $\lambda = 1$.

The O'Connell correlation relates tray efficiency to liquid viscosity and system relative volatility:

$$E_o = K_{21}(\mu_L\alpha)^{K_{22}} \quad (53)$$

Table 7 Minimum diameters and flow passes recommended

Number of flow passes	Minimum diameter (ft)	Minimum diameter (m)
2	5	1.52
3	8	2.44
4	10	3.05
5	13	3.96

Where μ_L is in centipoise, $K_{21} = 0.492$ and $K_{22} = -0.245$. The liquid viscosity and relative volatility are for the column feed at the average temperature and pressure of the column.

O'Connell derived his correlation from binary systems in distillation service with bubble-cap trays. Calculated values are slightly conservative for sieve and valve trays. Credit for the slight improvement in valve and sieve tray efficiency should be ignored and counted as a design margin. A separate correlation was developed for absorption services.

Efficiency in multiple component systems is far more complex. Every component may have a different efficiency at each stage. Components may interact and composition profiles may even reverse. The equilibrium slope for each component depends on many coupling and interaction effects. Therefore, overall efficiency cannot be easily derived from Murphree tray efficiencies. Most designs for both binary and multiple component systems use overall efficiency values based on operating data or test systems.

Table 8 lists tray efficiencies for common systems.[22] Performance is for "well-designed" trays in distillation operation. Table 9 summarizes tray design features,

Table 8 Tray efficiencies

System	Tray type	Tray efficiency (Murphree)	Overall
Nonhydrocarbon			
Acetic acid–water	Bubble cap	60	
Acetic acid–water	Sieve	75	
Acetone–methanol	Bubble cap	61	
Acetone–water	Sieve	80	
Aniline–water	Bubble cap	58	
Air–water	Bubble cap		83
Air–water–ammonia	Sieve	86	96
Air–water–ammonia	Valve	70	60
Air–water–triethylene glycol	Sieve	62	
Ammonia–water	Bubble cap	77	
Ammonia–water	Sieve	90	
Beer–water	Sieve		120
Benzene–methanol	Sieve	94	85
Benzene–propanol	Sieve	59	
Benzene–propanol	Valve	73	
Carbon dioxide–water	Bubble cap	80	125
Carbon dioxide–water	Sieve	80	125
CCl_4–benzene	Sieve		71
Deuterium–hydrogen	Bubble cap	50	44
Ethanol–water	Sieve	85	
Ethanol–water	Bubble cap	90	95
Ethanol–water	Valve		92
Ethanol–water–furfural	Sieve	80	
Ethylene dichloride–toluene	Bubble cap	95	
Isopropanol–water	Bubble cap	78	
Isopropnaol–water	Sieve		70
Methanol–isopropanol	Bubble cap	65	
Methanol–isopropanol–water	Bubble cap	68	
Methanol–water	Bubble cap	80	
Methanol–water	Sieve	92	
Methanol–water	Valve	88	
Methycyclohexane–water	Bubble cap	65	
Methylisobutylketone–water	Sieve		88
Naphtha–pinene–aniline	Bubble cap	90	
Naphtha–water	Bubble cap	65	
Oxygen–nitrogen	Bubble cap		76
Oxygen–water–ammonia	Sieve	75	
Toluene–propanol	Valve	78	
Toluene–propanol	Sieve	59	65
Hydrocarbon			
Atmospheric petroleum			
Light-heavy naphtha	Valve		75
Heavy naphtha–diesel	Valve		70
Diesel–gas oil	Valve		65
Wash	Valve		50
Stripping	Sieve		30
Side stripping	Valve		25
Benzene–toluene	Bubble cap	70	
Benzene–toluene	Sieve	82	
Benzene–toluene–xylene	Bubble cap	75	88
Benzene–toluene–xylene	Valve	69	
C_1-overhead	Sieve		100
$C_{2=}$–C_2	Sieve		100
C_2–C_3	Sieve		100

(Continued)

Table 8 Tray efficiencies *(Continued)*

System	Tray type	Tray efficiency (Murphree)	Overall
C_3–C_4	Valve		90
C_6–C_7	Valve		90
Cyclohexane–toluene	Sieve		70
Ethylbenzene–styrene	Sieve	75	
Ethylbenzene–styrene	Valve	85	80
iso-Butane–*n*-butane	Sieve	110	
iso-Butane–*n*-butane	Valve		80
Naphtha–pentanes	Valve		90r, 70s[a]
Naphtha splitter	Valve		90r, 70s
Naphtha stabilizer	Valve		90r, 70s
Natural gasoline stabilizers	Bubble cap	100	
n-Heptane–methylcyclohexane	Sieve	78	
n-Heptane–cyclohexane	Sieve	90	
n-Heptane–cyclohexane	Valve	100	96
n-Heptane–toluene	Sieve	53	
n-Heptane–benzene	Sieve	61	
n-Octane–toluene	Sieve		40
Propane–butane	Sieve	100	
Toluene–methylcyclohexane	Sieve	90	

Inconsistencies in data arise from compilation of multiple sources. Use with caution.
[a]r: rectification section; s: stripping section.
(Summary from Ref.[22] and from unpublished analysis by author.)

Table 9 Tray efficiency behavior

Feature	Behavior
Liquid viscosity	Low viscosity increases efficiency
Pressure	Froth systems: higher pressure slightly increases efficiency
Relative volatility	Low relative volatility increases efficiency
Surface tension	Minimal effect
Entrainment	Higher entrainment reduces efficiency. Impact minimal for entrainment below 5%
Flow path length	Longer flow paths improve efficiency. Flow paths longer than six deck devices have minimal impact unless composition profiles are created across tray (typically low relative volatility systems)
Hole area	Low hole area improves efficiency
Hole (valve) size	High-viscosity systems: small devices improve efficiency Froth systems: small devices slightly improve efficiency Spray systems: small devices slightly improve efficiency Efficiency changes in most systems are small. Most test data are from absorption systems. Limited test data from distillation systems have shown reduced efficiency with smaller tray devices in some cases
Multiple-pass tray flow maldistribution	More maldistribution reduces efficiency. L/V ratio variations under 1.2 between sections have minimal effect
Reflux ratio	Minimal effect
Stagnant regions on tray	Minimal efficiency reduction except in extreme cases
Weir height	Higher weirs improve efficiency in froth systems—moderate impact
Weeping	Higher weeping reduces efficiency. Large effect
Vapor-to-liquid ratio	Minimal effect

Froth systems: liquid continuous phase on tray; spray systems: liquid discontinuous phase on tray.

system characteristics, and general trends in tray efficiency. Kister has an excellent review of these factors.[11] Data from efficiency measurements show many inconsistencies. Pilot and plant tests are critical for determining the performance of a specific tray in a given system.

CONCLUSIONS

Fractional distillation, absorption, and stripping are the most common separation unit operations. All these services can use trays to achieve their mass-transfer objectives. Effective hydraulic and mass-transfer design for trays involves selection of the correct tray type. Industrially, most trays are either valve trays or sieve trays. Other types, including bubble-cap trays, have specialized uses. When properly selected and designed, trays meet the combined process and mechanical requirements of the service. The process and mechanical requirements are not independent. The equipment must work, and it must work as part of a process to achieve plant objectives. Applications for trays cover the gamut of the petroleum refining, petrochemical, gas processing, and chemical industries. The general design criteria here cover standard design issues for new towers. Revamps often involve extremely complex balancing of mechanical possibilities vs. process risk to meet economic and operating targets.

NOMENCLATURE

A_A	Active area
A_{AM}	Active area, design minimum
A_{CS}	Total tower cross-sectional area
A_{DC}	Area of the downcomer, actual or design
A_{DC1}	Area of the downcomer, top of outboard downcomer
A_{DC2}	Area of the downcomer, bottom of outboard downcomer
A_{DC3}	Area of the downcomer, top of inboard downcomer
A_{DC4}	Area of the downcomer, bottom of inboard downcomer
A_{DCE}	Downcomer escape area
A_{DM}	Area of the downcomer, design minimum
A_H	Area of the holes
A_{TM}	Total tower cross-sectional area, design minimum
α	Relative volatility
c_{AF}	Tray capacity factor
c_{AFO}	No liquid load capacity factor on tray free area
D	Tower diameter
ΔP_{dry}	Dry tray pressure drop
ΔP_{HUD}	Pressure drop (head loss) under the downcomer
ΔP_{total}	Total tray pressure drop
E	Vapor kinetic energy to liquid density ratio
E_o	Efficiency, overall
E_M	Efficiency, Murphree
F_D	Flooding factor, downcomer
F_A	Flooding factor, active area
F_F	Flooding fraction
F_P	Flooding percentage
G_m	Gas flow rate (moles per time)
H_1	Downcomer width at top of outboard downcomer (downcomer rise)
H_2	Downcomer width at bottom of outboard downcomer
H_3	Downcomer width at top of inboard (center) downcomer
H_4	Downcomer width at bottom of inboard (center) downcomer
H_C	Height of liquid crest over the outlet weir
H_{DC}	Downcomer backup height, clear liquid
H_{DCC}	Downcomer clearance height
H_L	Height of liquid on the tray
H_W	Height of outlet weir
λ	Ratio of the slope of the equilibrium curve and the slope of the component balance line.
L	Total liquid flow (mass or volume units as applicable)
L_m	Liquid flow rate (moles per time)
L_{FP}	Flow path length
m	Slope of equilibrium line on McCabe–Thiele diagram
μ_L	Liquid absolute viscosity (centipoise)
N_a	No. of actual trays
N_t	No. of theoretical trays
N_u	No. of valve units
ρ_L	Density, liquid
ρ_M	Density, tray material
ρ_V	Density, vapor
ρ	Tower radius
S	Angle for circular segment described by a downcomer
S_F	System factor
t_m	Valve thickness
T_S	Tray spacing
V	Total vapor flow (mass or volume units as applicable)
V_{DD}	Velocity, downcomer design (preliminary)
V_H	Velocity, vapor through holes
V_{load}	Vapor load
V_{UD}	Velocity under the downcomer

W	Weir length
W_{FP}	Active area to flow path length ratio
y_i	Mole fraction of light component in vapor to stage (inlet)
y_o	Mole fraction of light component in vapor from stage (outlet)
y^*	Mole fraction of light component in vapor if in equilibrium with liquid from stage (outlet)

REFERENCES

1. Sloley, A.W.; Fleming, B. Mechanical specification of mass-transfer equipment. Energy-Source Technology Conference, New Orleans, LA, Jan 24, 1994.
2. Glitsch, H.C Support Device for Bubble Caps. U.S. Patent 2,338,928, Jun 22, 1942.
3. Glitsch, H.C.; Winn, F.W. Flow Control Means U.S. Patent 3,080,155, Jan 18, 1960.
4. Glitsch, Inc. *Ballast Tray Design Manual*; Bulletin 4900, 6th Ed.; Glitsch, Inc.: Dallas, TX, 1993.
5. Koch Engineering. *Koch Flexitray Design Manual*; Wichita, KS, 1964.
6. Nutter Engineering. *Nutter Engineering Float Valve Design Manual*. (Rev. 1), Bulletin FV-1; Tulsa, OK, 1981.
7. Thorngren, J.T. Valve tray flooding generalized. Hydrocarbon Process. **1978**, *57* (8), 111–113.
8. Fleming, B.; Thorngreen, J. Tray zone flooding (TZF^{TM}) for valve and sieve trays. AIChE Spring Meeting, Houston, TX, Apr 23–27, 2001.
9. Lockett, M.J. *Distillation Tray Fundamentals*; Cambridge University Press: Cambridge, 1984.
10. Kister, H.Z. *Distillation Operation*; McGraw-Hill, Inc.: New York, 1990.
11. Kister, H.Z. *Distillation Design*; McGraw-Hill, Inc.: New York, 1992.
12. Weiland, R.H.; Resetarits, M.R. Finding new uses for old distillation equations. Pet. Technol. Q. **2002**, Autumn, 115–121.
13. Sloley, A.W.; Fleming, B. Successfully downsize trayed columns. Chem. Eng. Prog. **1994**, *90* (3), 39–47.
14. Sloley, A.W.; Martin, G.R. Subdue solids in distillation. Chem. Eng. Prog. **1995**, *91* (1), 64–73.
15. Sloley, A.W. Improve tray operations. Hydrocarbon Process. **2001**, *80* (6), 85–86.
16. Sloley, A.W. High-capacity distillation. Hydrocarbon Process. **1998**, *77* (8), 53–60.
17. Sloley, A.W. Should you switch to high-capacity trays? Chem. Eng. Prog. **1999**, *95* (1), 23–35.
18. O'Connell, H.E. Plate efficiency of fractionating columns and absorbers. AIChE Trans. **1946**, *42*, 741–755.
19. Fractionation Research, Inc. *Topical Report No. 18—Efficiency Comparisons AIChE Design Manual and FRI Experimental*; Fractionation Research, Inc.: Alhambra, CA, May 29, 1959.
20. Jacimovic, B.M.; Genic, S.B. Use a new approach to find Murphree tray efficiency. Chem. Eng. Prog. **1996**, *92* (8), 46–51.
21. Biddulph, M.W. Tray efficiency is not constant. Hydrocarbon Process. **1977**, *56* (10), 145–148.
22. Vital, T.J.; Grossel, S.S.; Olsen, P.I. Estimating separation efficiency part 2—plate columns. Hydrocarbon Process. **1984**, *63* (11), 147–153.

Drag Reducing Agents

Jacques L. Zakin
Ying Zhang
Department of Chemical Engineering, The Ohio State University, Columbus, Ohio, U.S.A.

Yunying Qi
Shell Global Solutions (US) Inc., Westhollow Technology Center, Houston, Texas, U.S.A.

INTRODUCTION

Drag reduction (DR), which was discovered more than 50 yr ago, is a turbulent flow phenomenon in which frictional energy losses can be significantly reduced by the addition of a small amount of a DR additive (DRA) to a turbulent flow system.[1,2] Drag reduction in internal flows is calculated using the following equation:

$$\%\mathrm{DR} = \frac{f_s - f}{f_s} \times 100 \quad (1)$$

where f is the measured friction factor of the solution with additive and f_s is that of the solvent at the same mean velocity.

Drag reduction additives in turbulent flowing systems can decrease system energy consumption, reduce pipe and pump size, increase flow rate or system length, and decrease heat transfer or a combination thereof.[3] The greatest use of DRA has been in crude and finished product petroleum pipelines as demonstrated in the 800 mi Alyeska Pipeline. Crude throughput in the 48 in.-diameter pipeline was increased by 30% by injecting a solution of high-polymer additive, at concentration as low as 1 ppm, downstream of the pumps at bottleneck sections of the line.[4] Other potential industrial applications include district heating and cooling (DHC) systems, oil well operations, sewage and floodwater disposal, hydraulic transport of solids, fire fighting, ships and submarines, jet cutting, irrigation, as well as heat tracer lines for cold North Sea crudes.[5–8]

Drag reduction systems can be either heterogenous (two phases) or homogenous (solution additives).[6] Heterogenous DR systems include solid-phase fibers or gas-phase microbubbles or even injection of concentrated polymer or surfactant solutions while polymers, aluminum disoaps, and surfactants are the main types of homogenous DRA.[9] Compared with heterogenous DRA, homogenous additives have the advantages of high effectiveness, low capital costs, and ease of operation. Homogenous polymer and surfactant systems have been the most widely studied DRAs and this entry will emphasize experimental findings with these two types of DRAs.

POLYMER DRA

Little progress was made in DR research until the early 1960s, but in the next 20 yr considerable progress was made.[6] Much of the early work focused on studies of poly(ethylene oxide) (PEO) in aqueous systems.[10] Research results on important solution variables affecting polymer DR, polymer structure, molecular weight (MW), concentration, and solvency, are described below as well as flow variables results.

Effect of Polymer Structure

Polymer DR effectiveness is related to molecular structure and polymer flexibility. Polymer backbone structure and side groups determine the flexibility of the molecule.[11] Liaw, Zakin, and Patterson showed that nonpolar polymers that are more flexible gave more effective DR.[12] They compared concentration effects of different polymer–solvent systems in a given tube in terms of a volume fraction, $C_c[\eta]$, where C_c is the critical concentration for disappearance of the laminar–turbulent transition in that tube and $[\eta]$ is the intrinsic viscosity. Values of $C_c[\eta]$ are very sensitive to the flexibility of the polymer molecules, increasing (becoming less DR effective) with increasing molecular rigidity. The most effective DR polymers, such as PEO and polydimethylsiloxane, are linear and flexible.[12,13] Other effective polymers include polyacrylic acid and polyacrylamide (PAM).[14,15] Mixtures of copolymers of linear α-olefins and 1-butene of very high MW are widely used as effective DRAs in hydrocarbon transport pipelines.[16]

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007743

Molecular Weight Effects

At a given polymer concentration, increased MW gives more effective DR. Polymers with MW lower than 100,000 are likely to be ineffective.[13] Liaw, Zakin, and Patterson, introduced a reduced MW, $m' = MW/M_c$, which takes into account the entanglement capacity of a polymer chain to compare MW effects of different polymers.[12] M_c is the minimum chain size for molecular entanglements to occur in polymer melts.

Molecular weight effects are illustrated in Figs. 1 and 2, showing the increasing DR effectiveness of higher-MW polystyrene samples at all concentrations regardless of solvent type.[17]

Hunston and Reischman showed that the contribution of 25 ppm of a 2×10^6 polystyrene in benzene solution to DR was negligible when added to 2.5 ppm of a 7×10^6 polystyrene (Fig. 3), demonstrating that a small amount of high-MW polymer dominates DR behavior in a mixture.[18]

Concentration

Drag reduction increases with polymer concentration until it reaches a maximum. The optimum concentration, and minimum torque in a turbulent Couette experiment with PEO in water decreased with increasing MW.[19] Hunston and Zakin studied the effect of concentration and MW on DR for polystyrene in toluene, a good solvent, and in toluene–isooctane, a poor solvent.[17] They observed increasing DR effectiveness with increasing concentration until offset by viscosity enhancement (Figs. 1 and 2). Liaw, Zakin, and Patterson found that increased concentration and decreased tube diameter reduced the onset Reynolds number (N_{Re}) leading to the disappearance of the laminar–turbulent transition, i.e., no observed onset, only a gradual departure from the laminar line (see section entitled "Onset" and Fig. 5).[12]

Virk et al. defined an intrinsic concentration for comparing DR effectiveness of homologous polymers of different MW and also different polymers:[20]

$$[C] = \frac{DR_m}{\lim_{C \to 0}(DR/C)} \tag{2}$$

where C = concentration, $\lim_{C \to 0}(DR/C) = [DR]$ is the intrinsic DR, i.e., a measure of DR per unit concentration at high dilution, and DR_m is the "maximum DR" as $C \to \infty$ for a given tube size and flow rate.[21]

Little simplified Eq. (2):[22]

$$\frac{DR}{DR_m} = \frac{C}{[C] + C} \quad \text{or} \quad \frac{C}{DR} = \frac{[C]}{DR_m} + \frac{C}{DR_m} \tag{3}$$

This equation predicts a linear relationship between C/DR and C and fits dilute aqueous PEO and PAM solutions also. $DR_m/[C]$ can be used to compare the effectiveness of different polymers at the same N_{Re} or to show the MW effect of a homologous series.[21]

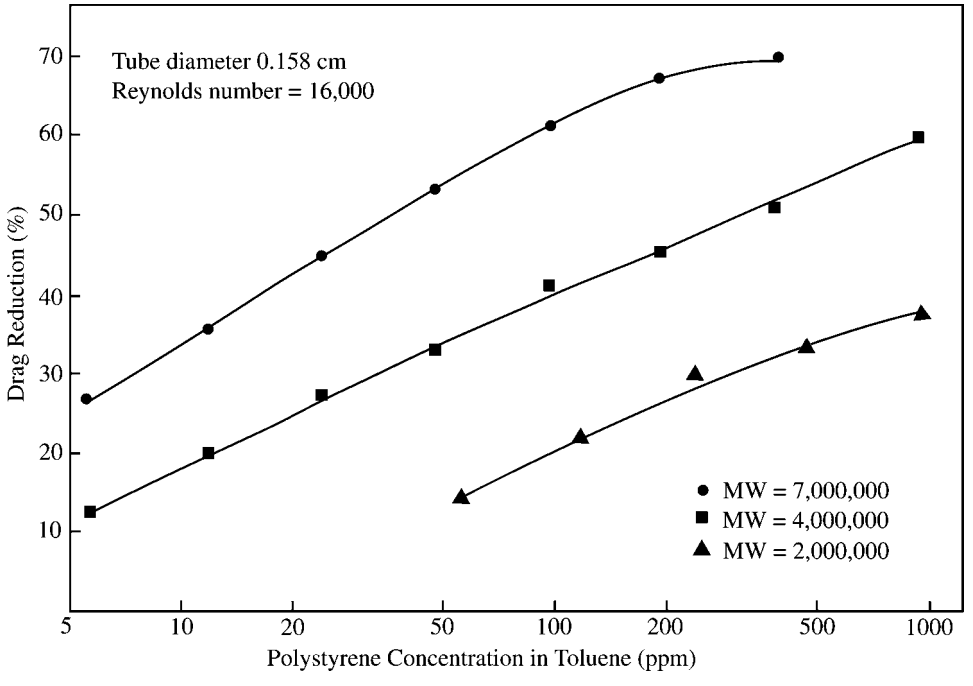


Fig. 1 Drag reduction effectiveness vs. polystyrene concentration in toluene, a good solvent. (From Ref.[17].)

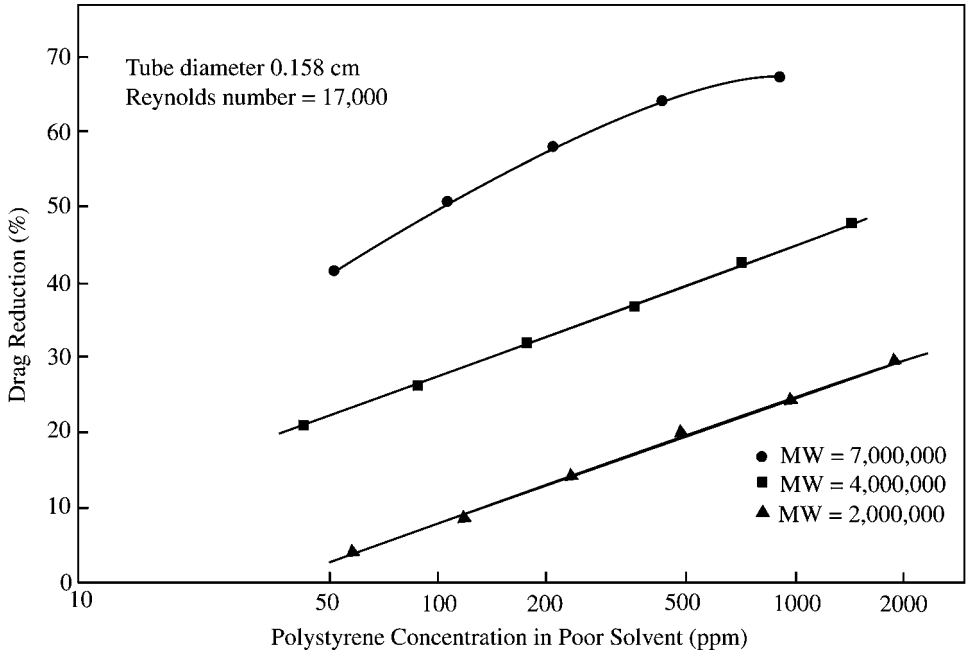


Fig. 2 Drag reduction effectiveness vs. polystyrene concentration in 58% toluene–42% isooctane, a poor solvent. (From Ref.[17].)

Solvency

For nonpolar linear polymers that form coils, Hershey and Zakin observed 40% less DR for polyisobutylene in benzene, a poor solvent, than in cyclohexane, a good solvent.[23] Hunston and Zakin showed that polystyrene in the good solvents, toluene or benzene, gave more DR than the same samples in a poor, mixed solvent of toluene and *iso*-octane as shown in Figs. 1 and 2.[24] These results and those of White and Gordon showed that expanded coil conformations, which enhance molecular flexibility, enhance DR effectiveness.[25]

Fig. 4 shows the pH effects on DR for a PAA-water system.[26] The untreated polymer solution (pH 5.0) has slight DR. Its DR ability increased slightly when pH was raised to 9.0, where the polymer molecules are fully ionized and extended (rod-shaped). Drag reduction also increased slightly as pH was lowered to 3 but increased abruptly at pH 2.1 and also at 1.0. They suggested that at these low pH values, the polymer chains undergo a second-order phase transition

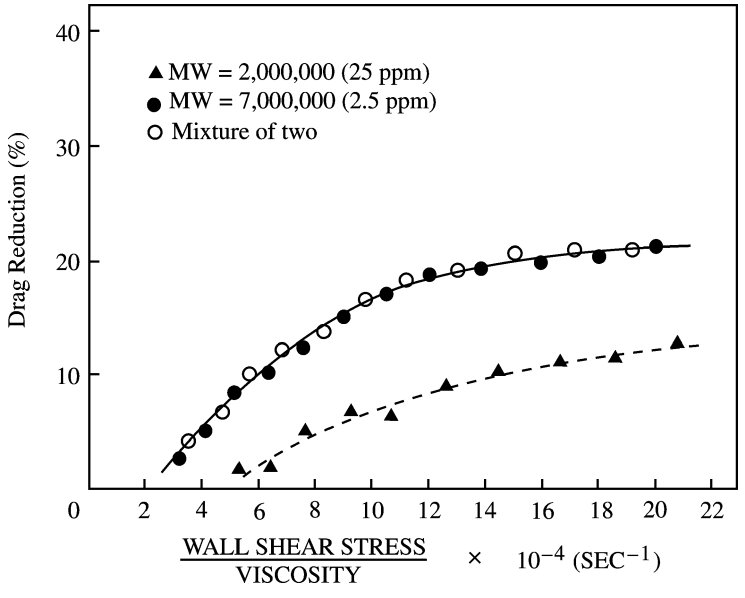


Fig. 3 Molecular weight effects on DR. (From Ref.[18].)

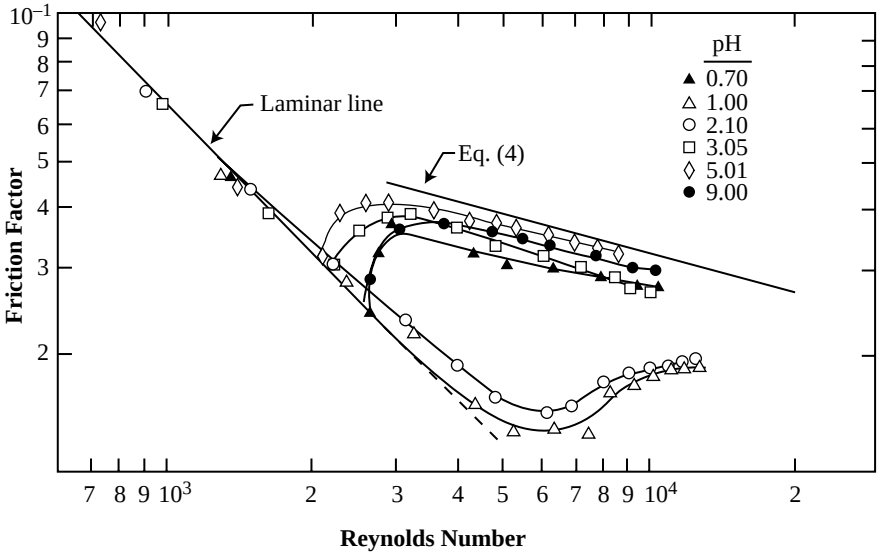


Fig. 4 pH effects on DR for a PAA–water system. (From Ref.[26].)

from extended rod to a more compact helical structure, which was very DR effective. Further increase in hydrogen ion concentration to pH of 0.7 led to reduction of DR, for which no explanation was provided.

For xanthan gum (XG) in water, Rochefort and Middleman observed that DR effectiveness is significantly reduced at high salt concentration.[27] Their rheological results indicated that as the ionic strength increased, XG molecules underwent a transition from highly extended, semiflexible rods with disordered backbone structure to collapsed, semirigid rods with ordered backbone structure.

Onset

For turbulent pipe flow of Newtonian, non-DR solutions, friction factors are predicted by the Prandtl-von Karman (PK) equation:

$$\frac{1}{\sqrt{f}} = 4\log_{10}\left(N_{\mathrm{Re}}\sqrt{f}\right) - 0.4 \tag{4}$$

which can be approximated by the Blasius equation:

$$f = 0.0791(N_{\mathrm{Re}})^{-1/4} \tag{5}$$

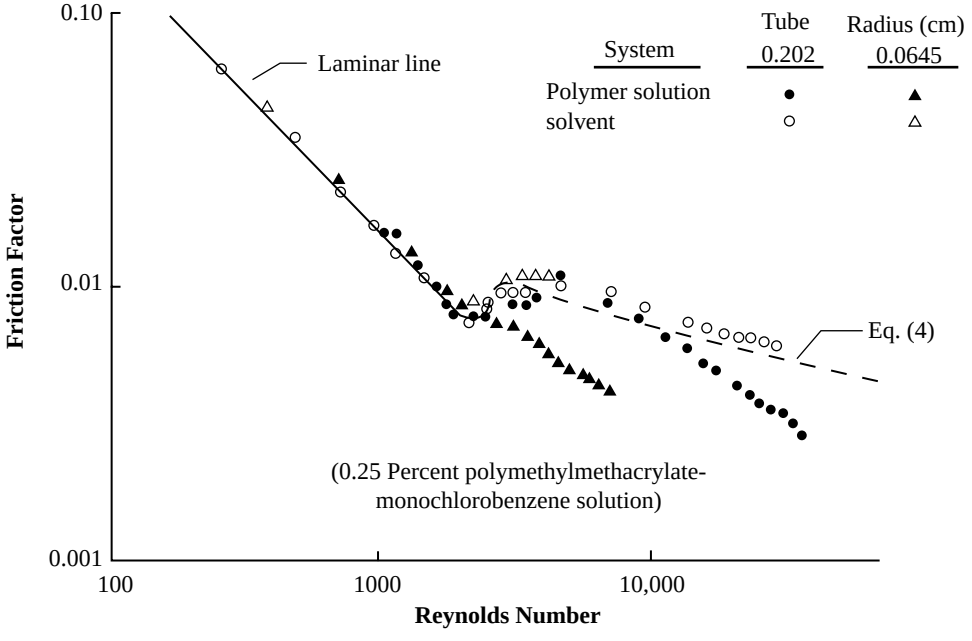


Fig. 5 An illustration of onset behavior. (From Ref.[1].)

Fig. 4 shows Toms' $f - N_{Re}$ data for both a Newtonian solvent and a DR polymer solution. Drag reduction onset occurs at $N_{Re} \approx 10{,}000$ in the 0.202 cm tube, while the critical conditions for DR are reached in the laminar region for the 0.0645 cm tube and there is no apparent onset. Rather, friction factors depart gradually from the laminar equation as N_{Re} increases (see also the section Concentration). Virk proposed that onset of DR in turbulent pipe flows, at which the friction factor departs abruptly from Eq. (4), occurs at a distinct wall-shear stress, τ_w, which does not depend on the pipe size and is only weakly dependent on concentration for a given polymer–solvent combination, but depends strongly on the radii of gyration, R_G, of the polymer molecules:

$$R_G{}^g \tau_w = \Omega \tag{6}$$

where Ω is an average onset constant characteristic of a given polymer species–solvent pair and g should vary between 2 and 3.[28] Virk's data for the PEO–water system showed that concentration hardly affects onset, but changes the slopes after onset.[28] However, concentration effects on onset were observed by Paterson and Abernathy for PEO–water and Hunston and Zakin for polystyrene in hydrocarbon solvent.[24,29] Paterson and Abernathy found that onset moved to higher N_{Re} as concentration decreased but that this dependence on concentration was considerably weaker than on the R_G. Hunston and Reischman showed that onset was determined by the highest-MW molecules in the polymer sample and occurred at lower N_{Re} with increase in MW.[18] Fabula, Lumley, and Taylor, and Hershey postulated that onset occurs when the product of wall-shear rate $\dot{\gamma} = \tau_w/\eta$, ($\eta$ = shear viscosity) and polymer molecule relaxation time, T_1, is of order 1.[30,31] This hypothesis was later supported by Berman and George from pipe flow experiments with dilute Polyox WSR N-80 (PEO) in glycerine–water solvents of different viscosities.[32] Hershey and Zakin proposed that the time scale theory might be applicable at values of $C[\eta] > 0.10$.[23] Hunston and Zakin observed that $\dot{\gamma} T_1$ was fairly constant ($\approx$10) for polystyrenes of different MW in a good solvent, but was larger in a poor solvent.[24] Later, Gyr and Bewersdorff hypothesized that DRAs are ineffective until the stress field is sufficiently strong to stretch and align the molecules.[33] Sreenivasan and White used an elastic theory to explain onset suggesting that onset is principally a near-wall effect (manifested through interference with turbulent bursting near the wall) and the polymers are only partially stretched at onset.[34]

Types A and B Drag Reduction

Virk and Wagger classified DR types based on the conformation of polymer molecules under flow.[35] Type A occurs in random-coiling systems and Type B with fully extended macromolecules. Type A DR is seen in Fig. 6, where a series of solutions of increasing concentration yield friction factor segments fanning outward from a common onset point on the PK line. Their slopes increase with increasing concentration

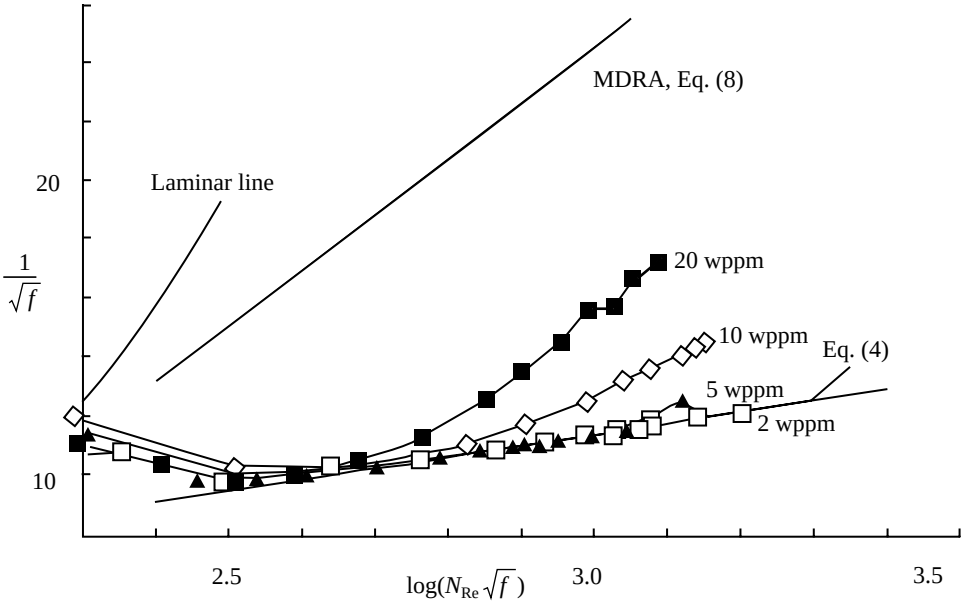


Fig. 6 Type A DR "fan" for coiled conformation. The polymer was partially hydrolyzed PAA of MW $\approx 15 \times 10^6$ at different concentrations. Pipe ID = 14.6 mm; Nacl = 0.1 mol/L. (From Ref.[35].)

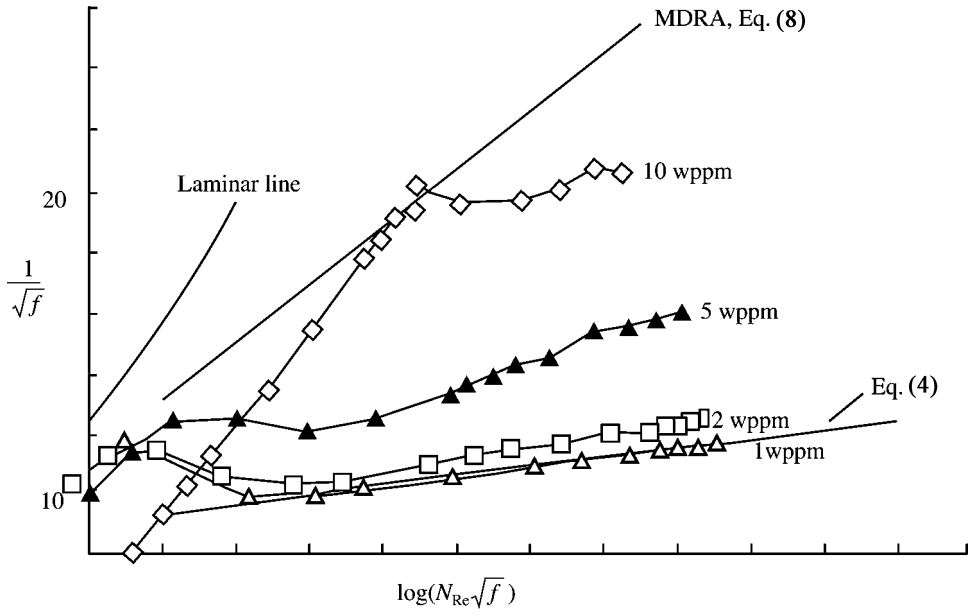


Fig. 7 Type B DR showing "ladder" for extended conformation. The polymer was partially hydrolyzed PAA of MW $\approx 15 \times 10^6$ at different concentrations. Pipe ID=14.6 mm; No salt added. (From Ref.[35].)

and DR increases with $N_{Re}\sqrt{f}$. They postulated that Type A DR is a result of the stretching or extension of random-coiled macromolecules by the turbulent flow field. In Type B DR (Fig. 7), the friction factor segments are parallel to, but displaced upward from the PK line as concentration varies from 2 to 10 ppm, with DR essentially independent of $N_{Re}\sqrt{f}$ but increasing with increasing concentration. Virk and Wagger observed a transition from Type A to Type B DR with decreasing salinity for a partially hydrolyzed PAM solution as salt concentration affected the molecular conformation.[36]

Shear Degradation

Polymer DRAs exposed to mechanical or extensional shear lose their DR effectiveness, as primary bonds are broken predominantly near the molecules' midpoints. Shear is particularly pronounced as solutions pass through a centrifugal pump, which presents a serious problem in long-distance pipeline transport, where booster pumps may be required every 50–100 mi. Pollert and Sellin reviewed in detail the mechanical degradation of DR polymers. The degradation rate increases with increasing MW and τ_w and decreases with concentration.[37] Patterson et al. studied 0.4% polyisobutylene (with a weight-average MW of 4.5×10^6) in cyclohexane, toluene, and *n*-heptane showing that the highest MW fractions of a polymer are most susceptible to shear degradation.[38] This was also observed by other researchers (for example, Refs.[39,40]).

Hunston and Zakin found that mechanical degradation rates are always larger in a poor solvent than in a good one.[24] Degradation is also related to molecular structure. Elata, Lehrer, and Kahanovitz noted that guar dissolved in water is more resistant to shear degradation than many other water-soluble polymer DRAs.[41] Patterson, Shen, and Chen reviewed reports on temperature effects on mechanical degradation, which showed, surprisingly, little effect or a negative effect.[42] That is, slower degradation was reported as the temperature rose, which was attributed to greater molecular flexibility. Thus, the factors governing shear degradation in DR include the stress level, concentration, molecular structure, molecular weight distribution, solvency, and temperature.

Initial degradation is relatively rapid and degradation can be represented by an exponential decay expression where time of exposure to shear is the independent variable. Minoura et al. showed that degradation could be characterized by a rate constant for molecular scission and a critical (minimum) MW below which no scission occurs at a given shear stress.[43] Similarly, Levinthal and Davison showed that for a given polymer chain length and strength of bond in laminar flow, there is a critical value of wall-shear stress below which shear degradation is negligible.[44] Based on this and the observation that a narrow high MW distribution polystyrene has greater resistance to degradation than a broader distribution polystyrene with a high MW tail, each initially of comparable DR ability, Hunston proposed a mechanism showing how the MW distribution affects shear

degradation.[45] He noted that chemical structure affects the critical lower MW for molecular scission and that MW distribution determines the fraction of DR effective molecules subject to scission.

SURFACTANT DRA

Introduction

Surfactant additives have been studied intensively in recent years because of the ''self-reparability'' or self-assembly of their microstructures after degradation by mechanical or extensional stresses. This ability has led to many studies of their applications in DHC recirculation systems. Classifications of surfactant DRAs and their self-assembly nature are described. Also discussed in this section are the main research results on microstructures, rheological properties, HTR of surfactant DR solutions, and approaches to enhance heat transfer coefficients. Significant field tests around the world are reviewed.

Classification of Surfactant DRAs

Surfactant DRAs can be roughly grouped into four types according to the charge properties of their hydrophilic head groups: cationic (positively charged) with anionic counterions, anionic (negatively charged), zwitterionic (with both positive and negative charges, zero net charge), and nonionic (uncharged) surfactants. Until now, cationic surfactants have received the most attention because of their lower costs compared with zwitterionic surfactants, broad DR temperature ranges (unlike nonionic surfactants with narrow temperature ranges), and insensitivity to the presence of calcium and magnesium ions in tap water, which cause DR anionic surfactants to precipitate. Typical cationic surfactants studied for DR are quaternary ammonium salts with one long alkyl chain (carbon number from 14 to 22) and methyl or hydroxyethyl groups in the other positions.[5] Recently, however, because of the unique DR properties mixed surfactant solutions exhibit, there has been increasing interest in mixed surfactant solutions, such as cationic and anionic, zwitterionic and anionic, and cationic surfactant solutions with different alkyl chain lengths.[46–53]

Self-Assembly of Surfactant DRAs

Primary bonds in high MW polymers subjected to high-shear or extensional flows, such as in pumps, contraction or expansion flows, break up and the degradation is permanent. The microstructures of surfactant solutions also break up under strong shear, but they self-assemble soon after the shear is released recovering their microstructures.[5]

Fig. 8 shows DR vs. N_{Re} of a typical DR cationic surfactant with counterion solution, $C_{17}H_{35}N(CH_3)_3Cl$/3,4-Cl-benzoate (5 mM/12.5 mM). Drag reduction reaches a maximum of 65%. In the effective temperature range (15–85°C), DR first increases with N_{Re} until a critical N_{Re} (critical wall shear stress) is reached above which it begins to lose its DR ability because of the

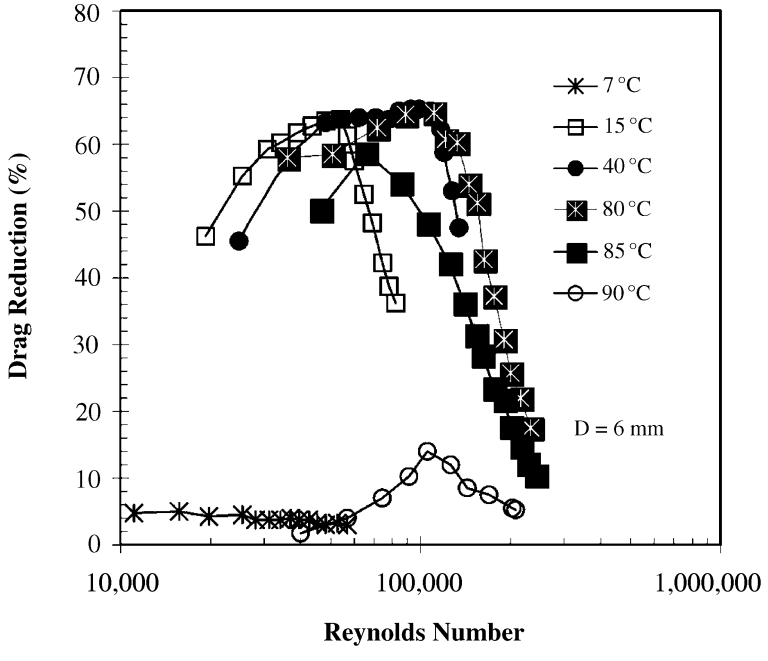


Fig. 8 Drag reduction vs. Reynolds number of a typical drag reducing cationic surfactant solution $C_{17}H_{35}N(CH_3)_3Cl$/3,4-Cl-benzoate (5 mM/12.5 mM).

high-shear stress breakup of the microstructure. If N_{Re} is decreased, the solution recovers its DR ability. That is, surfactant DR is mechanically recoverable. Surfactant DR is also temperature reversible. At 90°C, the surfactant microstructure no longer provides DR capability. However, if the solution temperature is reduced to 85°C, the solution will regain its DR ability. Each DR surfactant solution has an effective DR temperature range and, at each temperature and pipe diameter, an N_{Re} range.

Microstructures of DR Surfactant Solutions

Fig. 9 is a schematic phase diagram of a dilute aqueous cationic surfactant solution showing temperature and concentration effects on its microstructures. When the temperature is lower than the Krafft point [the temperature at which the solubility equals the critical micelle concentration (CMC)], the surfactant is partially in crystal or in gel form in the solution. At temperatures above the Krafft point and concentrations higher than the CMC, spherical micelles form in the surfactant solution. With further increase in concentration and/or on addition of counterions, the micelles form cylindrical rods or threads or worms with entangled thread-like and sometimes branched thread-like structures.

The concentration at which surfactants form rod-like micelles is called CMCII. Critical micelle concentration is almost independent of temperature while CMCII increases with temperature. Spherical micelles, rod-like or thread-like micelles, and vesicles are the three most common microstructures seen in dilute DR surfactant solutions.

An oppositely charged surfactant, organic counterions, and uncharged small compounds like alcohols are three common additives that induce the transformation of spherical micelles of cationic surfactants to thread-like micelles.[54] Addition of these additives gives a much smaller interfacial charge density on the micelle, shielding head group repulsions and allowing denser packing of surfactant molecules in the micelles.[55] Organic counterions, typically hydroxy- or halo-substituted benzoates at equimolar or higher concentrations, are the most effective and commonly used counterions for cationic DR surfactant solutions. They are small molecules, which can disperse both intermicellar repulsions and intramicellar repulsions and have organic portions that can penetrate into the lyophilic center of the micelle.[56] Fig. 10 shows a proposed elongated cylindrical micelle structure cartoon with a negatively charged organic counterion, 4-chlorobenzoate, inserted between positively charged ammonium cationic surfactant head groups.

Thread-like or worm-like micelles are believed to be necessary for surfactant solutions to be DR.[5,57] Many theoretical and experimental studies have investigated the kinetics, rheology, and microstructures of thread-like micelles. An important finding by Mendes et al. in their small-angle neutron scattering studies of microstructures and by Zheng et al. in their cryogenic transmission electron microscope studies of surfactant solutions is that vesicle microstructures can transform to thread-like micelles under shear.[58–60] This explains Zheng et al.'s surprising observation that a surfactant solution with vesicle structures in the quiescent state can also show very good DR ability.[60]

Such systems can form the thread-like micelles necessary for surfactant solutions to be DR under the shearing conditions in turbulent flows. The thread-like micelles align themselves along the flow direction causing DR of the solution.[61] Details of microstructures DR of aqueous cationic surfactant solutions vary with surfactant chemical structure and concentration, counterion chemical structure and

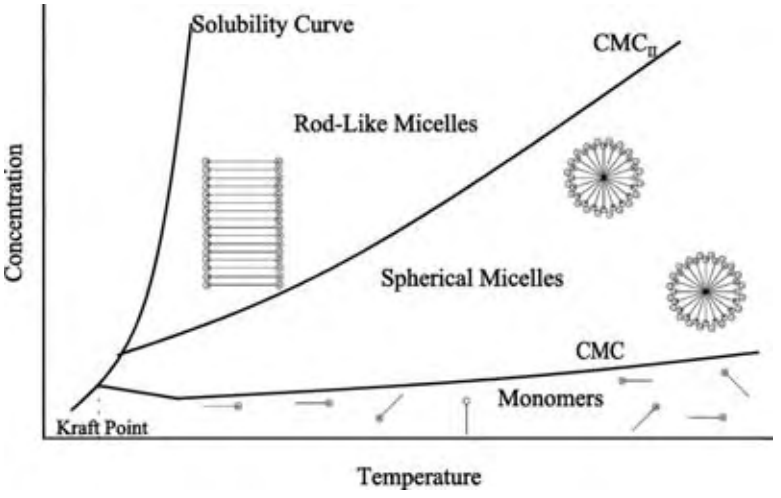


Fig. 9 Schematic phase diagram for dilute aqueous cationic surfactant solutions. (From Ref.[95].)

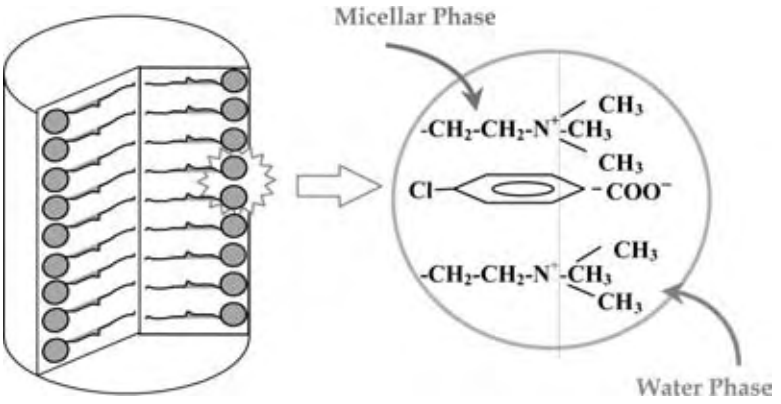


Fig. 10 Schematic structure of thread-like micelles for cationic surfactants. (From Ref.[63].) *(View this art in color at www.dekker.com.)*

concentration, temperature and pipe diameter, shear, and pH.

Rheological Properties of Surfactant DR Solutions

It has often been stated that DR of surfactant solutions is related to their rheological properties. A rise in shear viscosity at a critical shear rate, caused by a shear-induced structure (SIS), viscoelasticity (nonzero first normal stress difference, quick recoil, and stress overshoot), and high extensional viscosity/shear viscosity ratios (~100) are rheological properties found in many DR surfactant solutions. After reviewing the rheological behavior of many DR surfactant solutions, Qi and Zakin concluded that SIS and viscoelasticity are not always observed in DR surfactant solutions while high extensional/shear viscosity ratios may be a requirement for surfactant solutions to be DR.[57]

Enhancing Heat Transfer Ability of Drag Reducing Surfactant Solutions in Heat Exchangers

In addition to reduced friction factors (reduced momentum transfer), the heat transfer abilities of DR solutions are also greatly reduced. This may be caused by the thickened viscous boundary layers of DR flow and/or by reduced velocity fluctuations perpendicular to the flow.[62] Heat transfer reduction is defined as:

$$\%\mathrm{HTR} = \frac{Nu_s - Nu}{Nu_s} \times 100 \qquad (7)$$

where Nu_s and Nu are the Nusselt numbers measured for the solvent and the DR solution with additives.

Heat transfer reduction can be beneficial or detrimental to the practical use of DR solutions in DHC systems. On the beneficial side, the cost of insulation is reduced in pipelines in DHC systems. However, methods to locally enhance the heat transfer ability of DR surfactant solutions in heat exchangers to facilitate heat exchange while still maintaining their overall DR efficiency need to be developed. Three main approaches have been proposed: 1) modifying the inner surfaces of a heat exchanger to give a swirling motion and add disturbances to the viscous boundary layers of DR turbulent flows; 2) installing a pump, mechanical destructive devices, or an ultrasonic energy generator at the heat exchanger entrance to break up surfactant micelle structures temporarily; and 3) increasing the flow velocity of the fluid so that the critical wall-shear stress limit at the entrance or inside the heat exchanger is exceeded. In the second and the third method, surfactant solutions with broken microstructures will exhibit Newtonian flow behavior with reduced DR and HTR when passing through the heat exchanger. Surfactant solution microstructures are recovered downstream of the exchanger and the solution recovers its DR and HTR character.[63] The recovery time depends on surfactant species, temperature, system geometry, and on the solvent.[64]

Modifying the inner surface character of heat exchangers, especially fluted tube heat exchangers, is a very promising technique for enhancing heat transfer in new DHC systems with modest pressure drop penalty and low capital investment.[65] However, for existing DHC systems, it may not be practical. Installing mechanical destructive devices, such as static mixers or meshes, are practical for existing systems as they are cheap, easy to install, and easy to clean.[63,66,67] None of these techniques are useful for enhancing heat transfer of DR polymer solutions, however, as they would cause permanent mechanical degradation.

The effectiveness of different techniques to enhance the heat transfer ability of DR surfactant solutions depends on the surfactant solution, its concentration and temperature, and details of the circulating system. Minimum added costs and maximum heat transfer enhancement are the goals.[63]

Field Tests on Surfactant DRAs

The temperature reversibility and mechanical recoverability of surfactant DRAs are particularly important for a very promising application of DRAs, DHC systems, where hot or cold fluids are circulated in a loop between a waste heat source or a chiller and heat exchange stations in buildings in a district. District heating provides more efficient use of energy sources and reduces environmental pollution by combustion products. It is widely used in northern and eastern Europe and Japan and use in the United States, Canada, and Korea is expanding.[5] District cooling is often used in the United States, Canada, and Japan.

A field test in Herning, Denmark (1988–1991) demonstrated that Habon, a cationic DR surfactant, provided 70% DR from 40°C to 130°C, and its concentration was stable over 10 mo.[68] However, the slow biodegradability of Habon raised environmental concerns and prevented its commercial use in Denmark.

A similar ammonium cationic surfactant, Dobon-G, was tested in a larger-scale district heating system in Völklingen, Germany, from 1988 to 1990.[69] With the addition of sodium salicylate, up to 80% DR was achieved from 40°C to 130°C. Only slight HTR was observed in a plate heat exchanger, and, contrary to laboratory studies, the operation of the tube bundle heat exchanger was not influenced by using DRA, possibly because it was overdesigned. Another field test in the Czech Republic using Habon G also gave promising results.[70]

A district cooling test on a cationic surfactant solution, 2300 ppm Ethoquad T13/50 with 2000 ppm sodium salicylate, was conducted in a building-size system at Santa Barbara, CA, in 1995.[71] The results showed that 30% pumping energy saving can be achieved by using DRA for complex hydronic systems in buildings, with relatively low HTR. In a later study with a nonionic surfactant, the HTR in both evaporator and cooling coils was reduced by temporarily mechanically degrading the microstructure.[3] Even with total elimination of HTR, 12% saving on pumping in this small system was obtained at full load. Greater savings would be expected in larger systems.

Despite the efficiency of DR cationic surfactant solutions and several successful large-scale field tests, commercial use of these solutions has not been widespread largely because of their slow biodegradation properties.[5,51] A new series of biodegradable mixed zwitterionic/anionic surfactant solutions, which can be disposed of by normal sewage treatment, has been developed by Akzo Nobel with excellent field test results in the primary system of the same district heating system in Herning, Denmark.[51,72]

CHARACTERISTICS OF TURBULENT DRAG REDUCING SYSTEMS

Introduction

Mean flow behaviors and turbulence characteristics of DR solutions differ from those of Newtonian solutions. Topics covered in this section include the maximum DR asymptote (MDRA), mean velocity profiles, diameter scale-up, turbulence intensities, stress balance, and streak spacing in polymer and surfactant DR solutions as well as the HTR–DR relationship for surfactant DR solutions. Numerical simulations for polymer DR are noted and speculations on DR mechanisms are offered.

Maximum DR Asymptote

Both polymer and surfactant DRAs have MDRA. Drag reduction flow data lie between the PK line and MDRA. The MDRA proposed by Virk shown in Eqs. (8) and (9) is widely acknowledged as valid for high-polymer DR.[28]

$$f^{-(1/2)} = 19.0 \log\left(N_{\mathrm{Re}} f^{1/2}\right) - 32.4 \tag{8}$$

or

$$f = 0.58 N_{\mathrm{Re}}^{-0.58} (N_{\mathrm{Re}} = 4000 - 40,000) \tag{9}$$

Zakin, Myska, and Chara showed that some surfactant DR systems have friction factors over 40% below Virk's MDRA for high polymers and more than 90% below Newtonian solutions (Fig. 11).[73] Their MDRA for surfactants is

$$f = 0.32 N_{\mathrm{Re}}^{-0.55} \tag{10}$$

Later, Myska and Chara observed even lower surfactant friction factors indicating that the MDRA for surfactant solutions is even lower than in Eq. (10).[74]

Mean Velocity Profiles and Limiting Asymptote

The turbulent flow velocity profile for Newtonian fluids is arbitrarily divided into three regions: the viscous sublayer, the buffer layer, and the turbulent core. To represent velocity profiles in pipe flow, friction velocity defined as

$$u^* = \sqrt{\frac{\tau_{\mathrm{w}}}{\rho}} \tag{11}$$

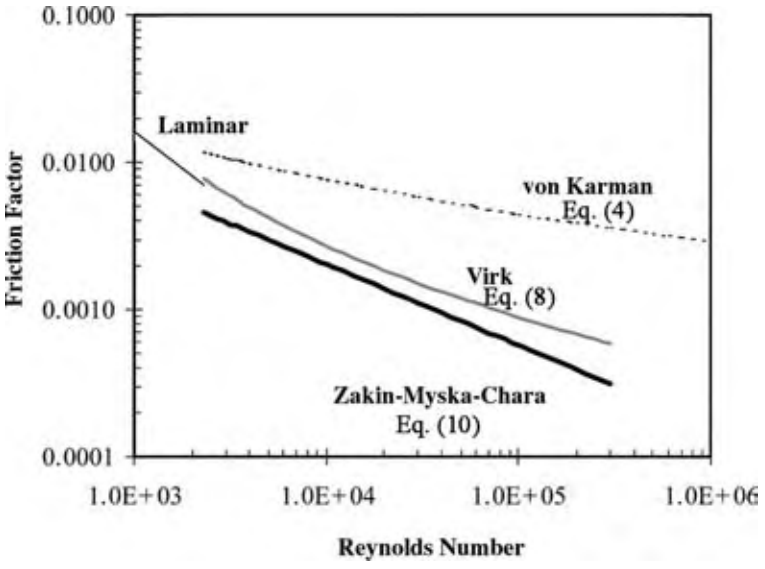


Fig. 11 Maximum DR asymptotes. *(View this art in color at www.dekker.com.)*

is used to form two nondimensional terms,

$$u^+ = \frac{u}{u^*} \tag{12}$$

$$y^+ = y\frac{u^*}{\nu} \tag{13}$$

where u is the local mean velocity, which varies with y, the distance from the wall, ν, is kinematic viscosity, and ρ is density. The velocity profile in the viscous sublayer of DR solutions is similar to that for Newtonian fluids. Fig. 12 shows turbulent core velocities for Newtonian fluids in pipe flow represented by:

$$u^+ = 2.5\ln(y^+) + 5.5 \tag{14}$$

while an additional term, ΔB, is needed for a parallel profile for fluids with moderate DR:

$$u^+ = 2.5\ln(y^+) + 5.5 + \Delta B \tag{15}$$

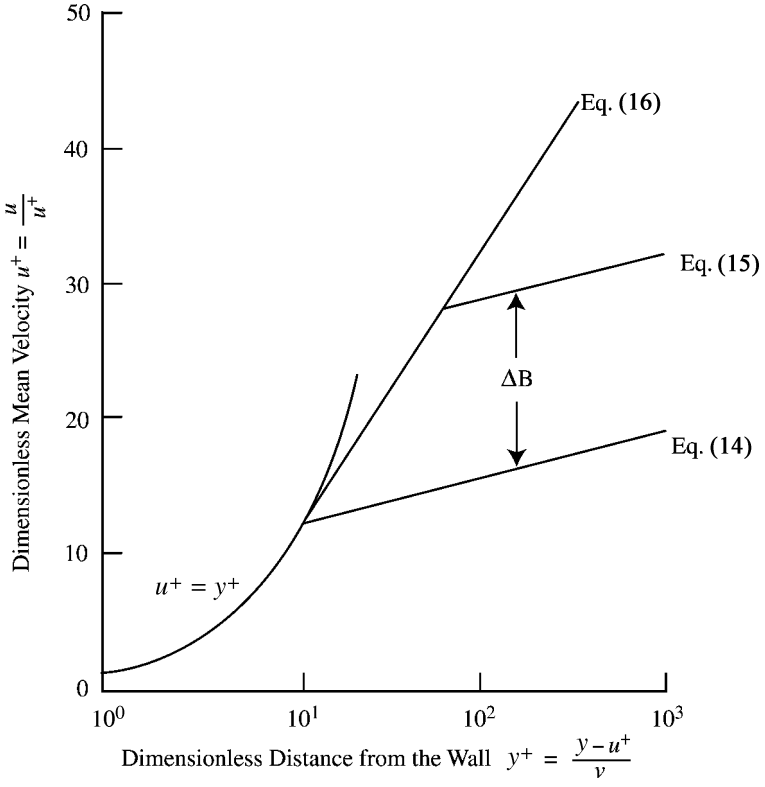


Fig. 12 Boundary layer and turbulent core velocity profile relationships for Newtonian and DR flows. (From Ref.[13].)

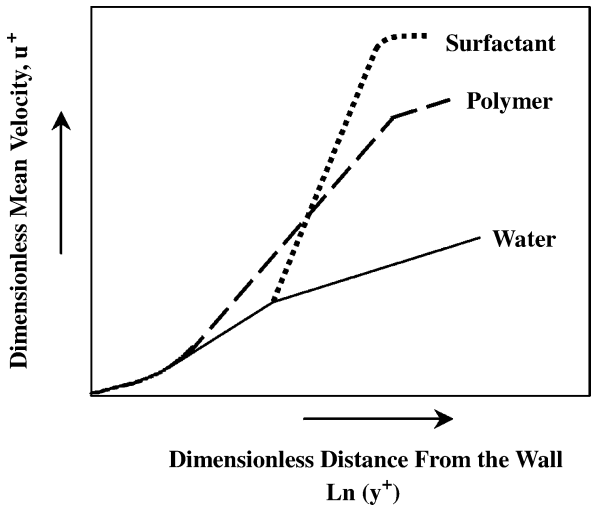


Fig. 13 Comparison of limiting turbulent mean velocity profiles.

However, for polymer and surfactant solutions close to their MDRA, limiting mean velocity profile asymptotes much steeper than Eqs. (12) and (13) are found in the turbulent core as shown in Eqs. (14) and (15) and Fig. 13.[28,73]

High-polymer asymptote:

$$u^+ = 11.7 \ln y^+ - 17.0 \tag{16}$$

Surfactant solution asymptote:

$$u^+ = 23.4 \ln y^+ - 65 \tag{17}$$

The limiting mean velocity profile for surfactant solutions is almost twice as steep as that for high polymers. These different limiting mean velocity profiles are related to the different MDRAs of their solutions.

Diameter Scale-Up

Diameter effects on DR are important for DR applications as practical economic pipe sizes are usually much larger than those in laboratory experiments. Virk showed that pipe size affects neither the value of τ_w for onset of DR nor the slope of the straight line describing the data on a PK plot.[28] As onset for small pipes occurs at lower N_{Re} than for larger pipes, Patterson, Zakin, and Rodriguez pointed out that the low velocities and τ_ws in larger commercial pipes may result in relatively low DR.[11]

Eq. (13) can be used for DR polymer solutions showing onset to scale from one pipe diameter to another. Techniques using the parameter ΔB assume similarity of velocity profiles for different pipe sizes or N_{Re} ranges. By measuring ΔB and u^* in one pipe size and assuming that they are independent of diameter, the friction factor of another pipe size can be obtained. However, this method gave only limited success in diameter scale-up with very small pipe sizes for polymer DR.[13]

Gasljevic and Matthys expressed DR as a function of both u^* and V.[75] They showed that polymer DR as a function of V is independent of diameter. However, diameter scale-up was only valid for Type A DR. Usui, Itoh, and Saeki suggested that the fluid relaxation time was essential to surfactant DR.[76] Their scaling law for DR as a function of either u^* or V (mean flow velocity) included the effect of shear thinning behavior of their surfactant solution.

Turbulence Intensities

A number of authors have observed lower root mean square time averaged turbulent intensities, v', normal to the flow direction for both high-polymer and surfactant DR systems compared with Newtonian solvents.[77–83] There is a subtle difference between high-polymer and surfactant solution data. The former peak is at y^+ values >100 while the latter is at $y^+ < 100$. Newtonian solvents peak at about $y^+ \approx 100$.[84] This, along with their different MDRAs and different limiting velocity profile slopes, suggests that their DR mechanisms may differ.

Warholic, Schmidt, and Hanratty showed that peak root mean square axial turbulent intensities, u', of a DR surfactant solution can be greater or lesser than solvent intensities depending on the N_{Re}—lower at low N_{Re} and higher at high N_{Re}.[82]

Stress Balance with Reduced Reynolds Stress

Several investigators have observed lower Reynolds stresses than expected for DR solutions caused by low v' and difference in phase between u' and v'. Recently, several researchers observed zero Reynolds stress profiles for high-DR surfactant solutions, clearly illustrating a major stress deficit in these systems.[74,79,82,83,85] Fig. 14 is a schematic illustrating the magnitude of the additional term, τ_{DR}, needed to satisfy the stress balance at any point for systems with zero Reynolds stresses.

$$\tau_{Total} = \tau_{viscous} + \tau_{Reynolds} + \tau_{DR} \tag{18}$$

$$\tau_{viscous} = \eta \, du/dy \tag{19}$$

$$\tau_{Reynolds} = -\rho \overline{u'v'} \tag{20}$$

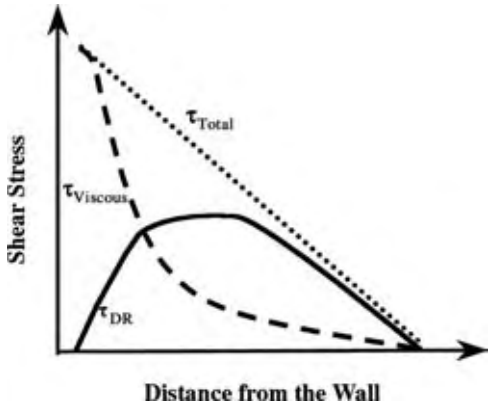


Fig. 14 Turbulent stress profiles for solution with zero Reynold stress. [See Eq. (18).]

Understanding the character of τ_{DR} and predicting it are major unresolved challenges.

Streak Spacing

Turbulent streaks in Newtonian flows have an average spacing of 100 in y^+ units. In polymer or surfactant DR flows, the spacing is increased and the streaks are more persistent and more stable.[86,87]

Relationship of HTR to DR

Aguilar studied the relationship of HTR in DR solutions to the reduction in drag.[88] He found that the ratio, HTR/DR, was greater than unity and relatively constant over a wide range of N_{Re} for both surfactants and high-polymer systems. A plot of the ratio for a number of surfactant solutions is shown in Fig. 15. He also showed that a slightly different ratio based on modified terms in the numerator and denominator of the HTR/DR ratio was even more constant. Thus, the more easily measured pressure drop data in straight tubes can be used to estimate the heat transfer coefficients.

Simulations of Polymer DR

A significant effort is under way to predict polymer DR in turbulent viscoelastic channel flow by direct numerical simulation using FENE, FENE-P, Gieskus, and other models (see, for example, Refs.[89–92]). While progress has been made by these and other authors, a review of these results is outside the scope of this entry.

Speculations on Surfactant and Polymer DR Mechanisms

Because of the interaction of the two complicated and not well-understood fields, turbulent flow and non-Newtonian fluids, understanding of DR mechanism(s) is still quite limited. Cates and coworkers (for example, Refs.[93,94]) and a number of other investigators have done theoretical studies of the dynamics of self-assemblies of worm-like micelles. Because these so-called "living polymers" are subject to reversible scission and recombination, their relaxation behavior differs from reptating polymer chains. An additional form of stress relaxation is provided by continuous breaking and repair of the micellar chains. Thus, stress relaxation in micellar networks occurs through a combination of reptation and breaking. For rapid scission kinetics, linear viscoelastic (Maxwell) behavior is predicted and is observed for some surfactant systems at low frequencies. In many cationic surfactant systems, however, the observed behavior in Cole–Cole plots does not fit the Maxwell model.[95,96]

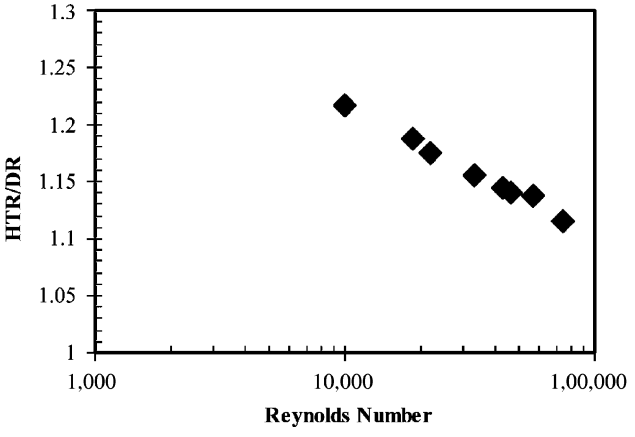


Fig. 15 Heat transfer reduction/DR ratio for a number of surfactant solutions. (From Ref.[88].)

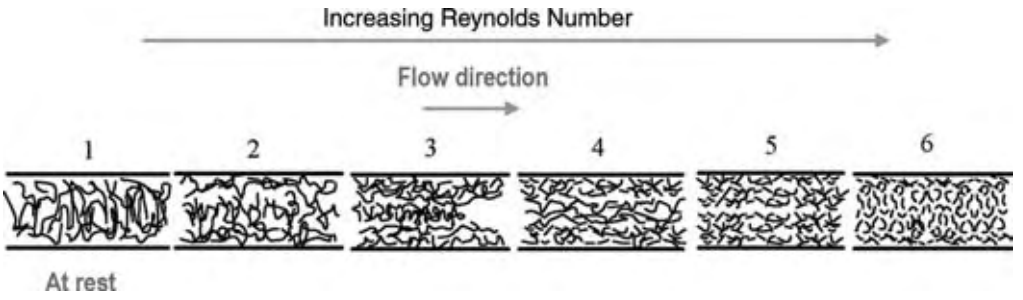


Fig. 16 Postulations on surfactant DR mechanism. 1) At rest, threadlike micelles distribute randomly in the solution. 2) As the Reynolds number increases, threadlike micelles near the wall get extended and start to align along the flow direction due to high wall shear stress. Micelle alignment decreases as the distance from the pipe wall increases and surfactant solution starts to show drag reduction as turbulent fluctuations in the radial direction which accompany the micelle alignment decrease. 3) With an increasing Reynolds number, more and more threadlike micelles get extended and align along the flow direction. Maximum drag reduction is observed when a maximum amount of long, extended threadlike micelles align along the flow direction. 4) At a higher Reynolds number, the wall shear stress becomes so large that the extended threadlike micelles near the wall are broken up. However, due to the non-uniform stress distribution in the pipe (maximum near the wall, zero at the center), the threadlike micelles away from the wall are not affected. Depending on how fast these micelles away from the wall can move to the wall, the maximum drag reduction of the surfactant solution can be maintained to even higher Reynolds number. 5) With further increase in the Reynolds number, threadlike micelles far from the wall are broken up resulting in loss of solution drag reduction ability. The corresponding shear stress at this critical Reynolds number is defined as the critical shear stress for surfactant drag reduction. 6) As the Reynolds number further increases, all of the micelles in the surfactant solution are broken down to a length which can not sustain drag reduction and the solution completely loses its drag reduction ability. (From Ref.[63].) *(View this art in color at www.dekker.com.)*

For DR surfactant solutions with thread-like micelles, Qi proposed a possible DR mechanism in pipe flow as shown in Fig. 16.[64] At rest, thread-like micelles are distributed randomly in the solution. As N_{Re} increases, the thread-like micelles near the wall are extended and start to align along the flow direction because of high wall-shear stress and the solution starts to show DR. Turbulent fluctuations decrease in the radial direction because of the micelle alignment and turbulent energy dissipation is reduced.

Shear stress and micelle alignment decrease as distance from the pipe wall increases, but with increasing N_{Re}, more and more thread-like micelles are extended and aligned along the flow direction. Maximum DR is observed when a maximum amount of long, extended thread-like micelles align along the flow direction. At higher N_{Re}, the wall-shear stress is so large that the extended thread-like micelles near the wall are broken up. However, because of the nonuniform stress distribution in the pipe (maximum near the wall, zero in the center), thread-like micelles away from the wall are still not affected. Depending on micelles' resistance to break up and how fast micelles away from the wall can move to the wall, the maximum DR of the surfactant solution may be maintained to large N_{Re}.

With further increase of N_{Re}, even thread-like micelles further from the wall are broken up and the small number of thread-like micelles in the center cannot reach the wall in sufficient number to maintain high DR. The wall-shear stress corresponding to this critical N_{Re} is defined as the critical shear stress for surfactant DR. With further increase in N_{Re}, more micelles in the surfactant solution are broken up, the remainder cannot sustain DR, and the solution loses its DR ability.

The amount, flexibility, and strength of the thread-like micelles determine the temperature range, critical wall-shear stress, and maximum DR ability of surfactant solutions.

The DR mechanism of high-polymer solutions may be similar to that of surfactant solutions with thread-like micelles being replaced by long polymer chains. However, the different MDRAs, limiting mean velocity profile slopes, and different locations of the transverse turbulent intensity peak positions in polymer and surfactant DR solutions suggest that their mechanisms may be different, probably because of the continual breakup and rapid self-reassembly of the micellar microstructures.

Other proposed DR mechanisms for both types of additives include high extensional viscosity providing additional resistance to vortex stretching and turbulent eddy growth and viscoelasticity, which dampens small turbulent eddies and stores turbulent fluctuation energy that would otherwise be dissipated.

CONCLUSIONS

Drag reduction in turbulent flows of dilute high-polymer of surfactant solutions is a striking

phenomenon. At concentrations of 1 ppm for high-MW polymer solutions and 200 ppm or less for surfactant solutions, significant DR can be attained, although there is little effect on the ordinary physical properties of the solvent. Heat transfer reduction of slightly greater magnitude accompanies DR. High-polymer DRAs have found a major application in reducing energy requirements in hydrocarbon pipeline transport while surfactant DRAs are highly promising for reducing energy requirements in DHC systems.

NOMENCLATURE

Roman Symbols

ΔB	Dimensionless shift in turbulent core velocity profile for DR solutions
% DR	Percent DR
% HTR	Percent HTR
$[C]$	Intrinsic concentration (ppm)
[DR]	Intrinsic DR (ppm^{-1})
C	Concentration (ppm)
C_c	Critical concentration for disappearance of the laminar-turbulent transition in a tube (g/dl)
DR_m	Maximum DR at infinite C
f	Friction factor in DR solutions
f_s	Friction factor in Newtonian solvents
g	Exponent of R_G in Eq. (6)
m'	MW/M_c
M_c	Minimum MW for entanglements to occur in polymer melts (g/mol)
MW	Molecular weight (g/mol)
N_{Re}	Reynolds number $= \rho Du/\eta$, where D is inner pipe diameter
Nu	Turbulent Nusselt number for DR solutions = hD/K, where h is the heat transfer coefficient and k is the thermal conductivity
Nu_s	Turbulent Nusselt number for Newtonian solvents
R_G	Radius of gyration of polymer molecules (M)
T_1	Polymer molecule relaxation time (sec)
u	Mean flow velocity (M/sec)
u^*	Friction velocity (M/sec)
u^+	Dimensionless velocity
u'	Root mean square fluctuating axial velocity (M/sec)
v'	Root mean square fluctuating velocity perpendicular to the flow direction (M/sec)
V	Mean flow velocity (M/sec)
y	Distance from the wall (M)
y^+	Dimensionless distance from the wall

Greek Symbols

$[\eta]$	Intrinsic viscosity (dl/g)
$\dot{\gamma}$	Wall-shear rate (Sec^{-1})
η	Shear viscosity (Pa/sec)
ν	Kinematic viscosity (M^2/sec)
ρ	Density (kg/m^3)
τ_{DR}	Additional stress for DR fluids (Pa)
$\tau_{Reynolds}$	Reynolds stress (Pa)
τ_{Total}	Total stress (Pa)
$\tau_{Viscous}$	Viscous shear stress (Pa)
τ_w	Wall-shear stress (Pa)
Ω	Average onset constant in Eq. (6)

REFERENCES

1. Toms, B.A. Some observations on the flow of linear polymer solutions through straight tubes at large Reynolds numbers. Proceedings of the International Congress of Rheology, North Holland: Amsterdam, 1949; Vol. 2, 135–141.
2. Mysels, K.J. Flow of Thickened Fluids U.S. Patent 2,492,173, Dec 27, 1949.
3. Gasljevic, K.; Hoyer, K.; Matthys, E.F. Field test of a drag-reducing surfactant additive in the hydronic cooling system of a building—Phase 2: heat transfer control. In *Applications of Heat Transfer in Equipment, Systems, and Education*. Proceedings of the 1998 ASME International Mechanical Engineering Congress and Exposition, Anaheim, CA, Nov 15–20, 1998; Swanson, L.W., et al., Eds.; ASME: New York, 1998; HTD-Vol. 361–363, PID-Vol. 3, 255–263.
4. Motier, J.F.; Chou, L-C.; Kommareddi, N. Commercial drag reduction: past, present, and future. In *Turbulence Modification and Drag Reduction*, Proceedings of the 1996 ASME Fluids Engineering Division Summer Meeting, San Diego, CA, Jul 7–11, 1996; Coleman, H., et al., Eds.; ASME: New York, 1996; FED-Vol. 237–242, 229–234.
5. Zakin, J.L.; Lu, B.; Bewersdorff, H.W. Surfactant drag reduction. Rev. Chem. Eng. **1998**, *14* (4–5), 253–320.
6. Manfield, P.D.; Lawrence, C.J.; Hewitt, G.F. Drag reduction with additives in multiphase flow a literature survey. Multiphase Sci. Technol. **1999**, *11*, 197–221.
7. Zohourian-Mashmoul, M.J-O-D.; Pourjavadi, A. Oil soluble drag-reducing polymers. J. Polymer Mater. **1994**, *11*, 239–247.

8. Sletfjerding, E.; Gladsø, A.; Oskarsson, H.; Elsborg, S. Boosting the heating capacity of oil production bundles using drag-reducing surfactants. Proceedings of the 12th European Drag Reduction Working Meeting, Herning, Denmark, Apr 18–20; The ERCOFTAC Special Interest Group (SIG): Eu, 2002.
9. Fontaine, A.A.; Deutsch, S.T.; Brungart, A.; Petrie, H.L.; Fenstermachker, M. Drag reduction by coupled systems: microbubble injection with homogeneous polymer and surfactant solutions. Exp. Fluids **1999**, *26*, 397–403.
10. Fabula, A.G. The Toms phenomenon in turbulent flow of very dilute polymer solutions. Proceedings of the 4th International Congress on Rheology, Brown University, Providence, Rhode Island, August 26–30, 1963; Lee, E.H., Copley, A., Eds.; Interscience: New York, 1965; Part 3, 445–479.
11. Patterson, G.K.; Zakin, J.L.; Rodriguez, J.M. Drag reduction: polymer solutions, soap solutions, and solid particle suspensions in pipe flow. Ind. Eng. Chem. **1969**, *61* (1), 22–30.
12. Liaw, G-C; Zakin, J.L.; Patterson, G.K. Effects of molecular characteristics of polymers on drag reduction. AIChE J. **1971**, *17* (2), 391–397.
13. Hoyt, J.W. Drag reduction. In *Encyclopedia of Polymer Science and Engineering*, 2nd Ed.; Kroschwitz, J., Mark, H.F., Bikales, N.M., Overberger, C.G., Menges, G., Eds.; John Wiley and Sons: New York, 1986; Vol. 5, 129–151.
14. Little, R.C.; Hansen, R.J.; Hunston, D.L.; Kim, O.-K.; Patterson, R.L.; Ting, R.Y. The drag reduction phenomenon. Observed characteristics, improved agents, and proposed mechanics. Ind. Eng. Chem. Fund. **1975**, *14* (4), 283–296.
15. Kim, O.K.; Little, R.C.; Ting, R.Y. Polymer structural effects in turbulent drag reduction. AIChE Symp. Ser. **1973**, *69* (130), 39–44.
16. Motier, J.F.; Carreir, A.M. Recent studies on polymer drag reduction in commercial pipelines. In *Drag Reduction in Fluid Flows: Techniques for Friction Control*; Sellin, R., Moses, R., Eds.; Ellis Horwood: West Sussex, U.K., 1989; 197–204.
17. Zakin, J.L.; Hunston, D.L. Mechanical degradation and drag reducing efficiency of dilute solutions of polystyrene. 2nd International Conference on Drag Reduction, Cambridge, U.K., Aug 31–Sep 2, 1977; BHRA Fluid Engineering: Cranfield, England, C5.
18. Hunston, D.L.; Reischman, M.M. The role of polydispersity in the mechanism of drag reduction. Phys. Fluids **1975**, *18* (12), 1626–1629.
19. Shin, H. Reduction of Drag in Turbulence by Dilute Polymer Solutions. Sc. D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1965.
20. Virk, P.S.; Merrill, E.W.; Mickley, H.S.; Smith, K.A.; Mollo-Christensen, E.L. The Tom's phenomenon: turbulent pipe flow of dilute polymer solutions. J. Fluid Mech. **1967**, *30* (2), 305–328.
21. Ting, R.Y.; Hunston, D.L. Polymeric additives as flow regulators. Ind. Eng. Prod. Res. Dev. **1977**, *16*, 129–136.
22. Little, R.C. Drag reduction in capillary tubes as a function of polymer concentration and molecular weight. J. Colloid Interface Sci. **1971**, *37* (4), 811–818.
23. Hershey, H.C.; Zakin, J.L. A molecular approach to predicting the onset of drag reduction in the turbulent flow of dilute polymer solutions. Chem. Eng. Sci. **1967**, *22*, 1847–1857.
24. Hunston, D.L.; Zakin, J.L. Effect of molecular parameters on the flow rate dependence of drag reduction and similar phenomena. In *Progress in Astronautics and Aeronautics*; Hough, S.R., Ed.; American Institute of Aeronautics and Astronautics: New York, 1980; Vol. 72, 373–385.
25. White, D., Jr.; Gordon, R.J. The influence of polymer conformation on turbulent drag reduction. AIChE J. **1975**, *21* (5), 1027–1029.
26. Hand, J.H.; Williams, M.C. Effect of secondary polymer structure on the drag-reducing phenomenon. J. Appl. Polym. Sci. **1969**, *13*, 2499–2502.
27. Rochefort, W.E.; Middleman, S. Relationship between rheological behavior and drag reduction for dilute xanthan gum solutions. In *Drag Reduction in Fluid Flows: Techniques for Friction Control*; Sellin, R., Moses, R., Eds.; Ellis Horwood: West Sussex, U.K., 1989; 69–76.
28. Virk, P.S. Drag reduction fundamentals. AIChE J. **1975**, *21* (4), 625–655.
29. Paterson, R.W.; Abernathy, F.H. Turbulent flow drag reduction and degradation with dilute polymer solutions. J. Fluid Mech. **1970**, *43* (4), 689–710.
30. Fabula, A.G.; Lumley, J.L.; Taylor, W.D. Some interpretations of the Toms effect. In *Modern Developments in the Mechanics of continua*, Proceedings of the International Conference on Rheology, Pinebrook, New York, Aug 23–27, 1965; Eskanazi, S., Ed.; Academic Press: New York, 1966; 145–164.
31. Hershey, H.C. Drag Reduction in Newtonian Polymer Solutions. Ph. D. dissertation, University of Missouri, Rolla, MO, 1965.
32. Berman, N.S.; George, W.K., Jr. Onset of drag reduction in dilute polymer solutions. Phys. Fluids **1974**, *17* (1), 250–251.
33. Gyr, A.; Bewersdorff, H.-W. *Drag Reduction of Turbulent Flows by Additives*; Kluwer Academic Press: Boston, MA, 1995.

34. Sreenivasan, K.; White, C.M. The onset of drag reduction by dilute polymer additives, and the maximum drag reduction asymptote. J. Fluid Mech. **2000**, *409*, 149–164.
35. Virk, P.S.; Wagger, D.L. Aspects of mechanisms in type B drag reduction. In *Structure of Turbulence and Drag Reduction*, Proceedings of the IUTAM Symposium, Zürich, Switzerland, Jul 25–28, 1989; Gyr, A., Ed.; Springer-Verlag: New York, 1990; 201–213.
36. Virk, P.S.; Wagger, D.L. The effect of salinity on turbulent drag reduction by polyelectrolyte additives at high Reynolds numbers. Addendum to the Proceedings of the International Symposium on Seawater Drag Reduction, Newport, RI, Jul 22–23; Office of Naval Research: Arlington VA, 1998.
37. Pollert, J.; Sellin, R.J. Mechanical degradation of drag reducing polymer and surfactant additives: a review. In *Drag Reduction in Fluid Flows: Techniques for Friction Control*; Sellin, R., Moses, R., Eds.; Ellis Horwood: West Sussex, U.K., 1989; 179–188.
38. Patterson, G.K.; Hershey, H.C.; Green, C.D.; Zakin, J.L. Effect of degradation by pumping on normal stresses in polyisobutylene solutions. Trans. Soc. Rheol. **1966**, *10* (2), 489–500.
39. Sylvester, N.D.; Kumor, S.M. Degradation of dilute polymer solutions in turbulent tube flow. AIChE Symp. Ser. **1973**, *69* (130), 69–81.
40. Yu, J.F.S.; Zakin, J.L.; Patterson, G.K. Mechanical degradation of high molecular weight polymers in dilute solution. J. Appl. Polym. Sci. **1979**, *23*, 2493–2512.
41. Elata, C.; Lehrer, J.; Kahanovitz, A. Turbulent shear flow of polymer solutions. Isr. J. Technol. **1966**, *4* (1–2), 87–95.
42. Patterson, G.K.; Shen, A.-M.; Chen, I.-M. Rates of shear degradation of drag reducing polymer. In *Drag Reduction,* 3rd International Conference on Drag Reduction, University of Bristol, Bristol, Jul 2–5, 1984; Sellin, R.H.J., Moses, R.T., Eds.; E.4-1–E.4-6.
43. Minoura, Y.; Kasuya, T.; Kawamura, S.; Nakano, A. Degradation of poly(ethylene oxide) by high-speed stirring. J. Polym. Sci. A-2 **1967**, *5*, 125–142.
44. Levinthal, C.; Davison, P.F. Degradation of deoxyribonucleic acid under hydrodynamic shearing forces. J. Mol. Biol. **1961**, *3*, 674–683.
45. Hunston, D. Effects of molecular weight distribution in drag reduction and shear degradation. J. Polym. Sci. Polym. Chem. Ed. **1976**, *14*, 713–727.
46. Koehler, R.D.; Raghavan, S.R.; Kaler, E.W. Microstructure and dynamics of wormlike micellar solutions formed by mixing cationic and anionic surfactants. J. Phys. Chem. B **2000**, *104*, 11,035–11,044.
47. Schubert, B.A.; Wagner, N.J.; Kaler, E.W. The rheology and microstructure of solutions of charged worm-like micelles. 13th International Congress on Rheology, Cambridge, U.K.; British Society of Rheology: Glasgow, 2000.
48. Marques, E.F.; Regev, O.; Khan, A.; Miguel, M.D.G.; Lindman, B. Vesicle formation and general phase behavior in the catanionic mixture SDS-DDAB-water the cationic-rich side. J. Phys. Chem. B **1999**, *103*, 8353–8363.
49. Salkar, R.A.; Mukesh, D.; Samant, S.D.; Manohar, C. Mechanism of micelle to vesicle transition in cationic-anionic surfactant mixtures. Langmuir **1998**, *14*, 3778–3782.
50. Li, F.; Li, G-Z.; Chen, J.-B. Synergism in mixed zwitterionic-anionic surfactant solutions and the aggregation numbers of the mixed micelles. Colloid Surface A **1998**, *145*, 167–174.
51. Hellsten, M.; Harwigsson, I.; Blais, C.; Wollerstrand, J. Drag reduction by N-alkylbetaines—a type of zwitterionic surfactants. In *Turbulence Modification and Drag Reduction*, Proceedings of the 1996 ASME Fluids Engineering Division Summer Meeting, San Diego, CA, Jul 7–11, 1996; Coleman, H., et al., Eds.; ASME: New York, 1996; FED-Vol. 237–242, 37–46.
52. Harwigsson, I. Surfactant Aggregation and Its Application to Drag Reduction. Ph.D. dissertation, Lund University, Sweden. 1995.
53. Lin, Z.; Chou, L-C.; Lu, B.; Zheng, Y.; Davis, H.T.; Scriven, L.E.; Talmon, Y.; Zakin, J.L. Experimental studies on drag reduction and rheology of mixed cationic surfactants with different alkyl chain lengths. Rheol. Acta **2000**, *39*, 354–359.
54. Rehage, H.; Hoffmann, H. Viscoelastic surfactant solutions; model for rheological research. Mol. Phys. **1991**, *74* (5), 933–973.
55. Bewersdorff, H.W. Rheology of drag reducing surfactant solutions. In *Turbulence Modification and Drag Reduction*, Proceedings of the 1996 ASME Fluids Engineering Division Summer Meeting, San Diego, CA, Jul 7–11, 1996; Coleman, H., et al., Eds.; ASME: New York, 1996; FED-Vol. 237–242, 25–29.
56. Prud'homme, R.K.; Warr, G.G. Elongational flow of solutions of rodlike micelles. Langmuir **1994**, *10* (10), 3419–3426.
57. Qi, Y.; Zakin, J.L. Chemical and rheological characterization of drag reducing cationic surfactant solutions. Ind. Eng. Chem. Res. **2002**, *41* (25), 6326–6336.
58. Mendes, E.; Narayanan, J.; Oda, R.; Kern, F.; Candau, S.J. Shear-induced vesicle to wormlike

micelle transition. J. Phys. Chem. B **1997**, *101*, 2256–2258.
59. Mendes, E.; Menon, S.V.G. Vesicle to micelle transitions in surfactant mixtures induced by shear. Chem. Phys. Lett. **1997**, *275*, 477–484.
60. Zheng, Y.; Lin, Z.; Zakin, J.L.; Talmon, Y.; Davis, H.T.; Scriven, L.E. Cryo-TEM imaging the flow-induced transition from vesicles to threadlike micelles. J. Phys. Chem. B **2000**, *104* (22), 5263–5271.
61. Qi, Y.; Littrell, K.; Thiyagarajan, P.; Talmon, Y.; Lin, Z.; Zakin, J.L. Small angle neutron scattering (SANS) study of shearing effects on drag reducing surfactant solutions. Manuscript in preparation.
62. Sellin, R.H.J.; Hoyt, J.W.; Scrivener, O. The effect of drag-reducing additives on fluid flows and their industrial applications, part I: basic aspects. J. Hydr. Res. **1982**, *20* (1), 29–68.
63. Qi, Y. Investigation of relationships among microstructure, rheology, drag reduction and heat transfer of drag reducing surfactant solutions. Ph.D. dissertation, The Ohio State University, Columbus, OH, 2002.
64. Blais, C.; Harwigsson, I.; Hellsten, M.; Wollerstrand, J. Drag reduction in district heating and cooling circuits-temporary disruption of micelles to preserve the heat exchanger capacity. Proceedings of the 4th World Surfactants Congress, Barcelona, Spain, Jun 3–7, 1996; Sanchez-Leal, J., Ed.; The Royal Society of Chemistry: U.K., 1996; Vol. 2, 424–438.
65. Qi, Y.; Kawaguchi, Y.; Lin, Z.; Ewing, M; Christensen, R.N.; Zakin, J.L. Enhanced heat transfer of drag reducing surfactant solutions with fluted tube-in-tube heat exchanger. Int. J. Heat Mass Trans. **2001**, *44* (8), 1495–1505.
66. Gasljevic, K.; Matthys, E.F. Experimental investigation of thermal and hydrodynamic development regions for drag-reducing surfactant solutions. Trans. ASME **1997**, *119*, 80–88.
67. Qi, Y.; Kawaguchi, Y.; Christensen, R.N.; Zakin, J.L. Enhancing heat transfer ability of drag reducing surfactant solutions with static mixers and honeycombs. Int. J. Heat Mass Trans. **2003**, *46*, 5161–5173.
68. Hammer, F. Demonstration of smooth water in the district heating system of Herning, Denmark. Proceedings of the International Symposium on Fluids for District Heating, Copenhagen, Denmark, Apr 10–11, 1991; Bøhm, B., Ed.; Laboratory of Heating and Air Conditioning: Technical University of Denmark, 1991; 139–150.
69. Steiff, A.; Althaus, W.; Weber, M.; Weinspach, P.M. Application of drag reducing additives in district heating systems—present state of investigations. In *Drag Reduction in Fluid Flows: Techniques for Friction Control*; Sellin, R., Moses, R., Eds.; Ellis Horwood: West Sussex, U.K., 1989; 247–254.
70. Pollert, J.; Zakin, J.L.; Myska, J.; Kratochvil, P. Use of friction reducing additives in district heating system field test at Kladno-Krocehlavy, Czech Republic. Proceedings of the International District Heating and Cooling Association, Seattle, WA, Jun 18–21, 1994; IDHCA: Seattle, WA, 1994; 141–156.
71. Gasljevic, K.; Matthys, E.F. Field test of a drag-reducing surfactant additive in a hydronic cooling system. Proceedings of the 1996 ASME Fluids Engineering Division Summer Meeting, San Diego, CA, Jul 7–11, 1996; Coleman, H., et al., Eds.; ASME: New York, 1996; FED-Vol. 237–242, 249–260.
72. Hammer, F.; Energiteknik, B.S.; Hellsten, M. Full-scale test of drag reducing additives in Herning, Denmark. Akzo Nobel Surface Chemistry AB Final Report, 2000.
73. Zakin, J.L.; Myska, J.; Chara, Z. New limiting drag reduction and velocity profile asymptotes for nonpolymeric additives systems. AIChE J. **1996**, *42* (12), 3544–3546.
74. Myska, J.; Chara, Z. New view of the surfactant drag reduction ability. Proceedings of the 11th European Drag Reduction Working Meeting, Prague, Czech Republic, Sep 15–17; Institute of Hydrodynamics, Academy of Sciences of the Czech Republic: Prague, 1999.
75. Gasljevic, K.; Aguilar, G.; Matthys, E.F. An improved diameter scaling correlation for turbulent flow of drag-reducing polymer solutions. J. Non-Newton. Fluid Mech. **1999**, *84*, 131–148.
76. Usui, H.; Itoh, T.; Saeki, T. On pipe diameter effects in surfactant drag-reducing pipe flows. Rheol. Acta **1998**, *37*, 122–128.
77. Gampert, I.; Rensch, A. Polymer concentration and near wall turbulence structure of chemical flow of polymer solutions. In *Turbulence Modification and Drag Reduction*, Proceedings of the 1996 ASME Fluids Engineering Division Summer Meeting, San Diego, CA, Jul 7–11, 1996; Coleman, H., et al., Eds.; ASME: New York, 1996; FED-Vol. 237–242, 129–136.
78. Chara, Z.; Zakin, J.L.; Severa, M.; Myska, J. Turbulence measurements of drag reducing surfactant systems. Exp. Fluids **1993**, *16*, 36–41.
79. Kawaguchi, Y.; Tawaraya, Y.; Yabe, A.; Hishida, K.; Maeda, M. Active control of turbulent drag reduction in surfactant solutions by wall heating. In *Turbulence Modification and Drag Reduction*, Proceedings of the 1996 ASME Fluids Engineering Division Summer Meeting, San Diego,

CA, Jul 7-11, 1996; Coleman, H., et al., Eds.; ASME: New York, 1996; FED-Vol. 237–242, 47–52.
80. Myska, J.; Zakin, J.L.; Chara, Z. Viscoelasticity of a surfactant and its drag-reducing ability. Appl. Sci. Res. **1996**, *55*, 297–310.
81. Park, S.P.; Suh, H.S.; Moon, S.H.; Yoon, H.K. Pump and temperature effects on and flow characteristics of drag reducing surfactants. In *Turbulence Modification and Drag Reduction*, Proceedings of the 1996 ASME Fluids Engineering Division Summer Meeting, San Diego, CA, Jul 7–11, 1996; Coleman, H. et al., Eds.; ASME: New York, 1996; FED-Vol. 237–242, 177–182.
82. Warholic, M.D.; Schmidt, G.M.; Hanratty, T.J. The influence of a drag-reducing surfactant on a turbulent velocity field. J. Fluid Mech. **1997**, *388*, 1–20.
83. Warholic, M.D.; Massah, H.; Hanratty, T.J. Influence of drag-reducing polymers on turbulence: effects of Reynolds number, concentration and mixing. Exp. Fluids **1999**, *27* (5), 461–472.
84. Zakin, J.L.; Myska, J.; Lin, Z. Similarities and differences in drag reduction behavior of high polymer and surfactant solutions. Proceedings of the International Symposium on Seawater Drag Reduction; Office of Naval Research: Arlington, VA, 1998; 277–280.
85. Saeki, T.; Guzman, D.; Manuel, R.; Morishima, H.; Usui, H.; Nishimura, T.A flow visualization study on the mechanism of turbulent drag reduction by surfactants. Nihon Reoroji Gakk. **2000**, *28* (1), 35–40.
86. Hetsroni, G.; Zakin, J.L.; Mosyak, A. Low-speed streaks in drag-reduced turbulent flow. Phys. Fluids **1997**, *9* (8), 2397–2404.
87. Luchik, T.S.; Tiederman, W.G. Turbulent structure in low-concentration drag-reducing flow. J. Fluid Mech. **1988**, *190*, 241–263.
88. Aguilar, G. An experimental study of drag and heat transfer reductions in turbulent pipe flows for polymer and surfactant solutions. Ph. D. dissertation, University of California, Santa Barbara, CA, Aug 1999.
89. Beris, A.N.; Dimitropoulos, C.D.; Sureshkumar, R.; Handler, R.D. Direct numerical simulations of polymer-induced drag reduction in viscoelastic turbulent channel flows. Proceedings of the International Congress on Rheology, Cambridge, U.K., Aug 20–25; British Society of Rheology: Glasgow, 2000; Vol. 2, 190–192.
90. Massah, H.; Hanratty, T.J. Added stresses because of the presence of FENE-P bead-spring chains in a random velocity field. J. Fluid Mech. **1997**, *337*, 67–101.
91. Dimitropoulos, C.D.; Sureshkumar, R.; Beris, A.N. Direct numerical simulation of viscoelastic turbulent channel flow exhibiting drag reduction: effect of the variation of rheological parameters. J. Non-Newton. Fluid **1998**, *79* (2–3), 433–468.
92. Nieuwstadt, F.T.M.; Ptasinski, P.K.; Boersma, B.J. Direct numerical simulations of maximum drag reduction by polymers, 12th European Drag Reduction Meeting, Herning, Denmark, Apr 18–20, 2002.
93. Turner, M.S.; Cates, M.E. Linear viscoelasticity of living polymers: a quantitative probe of chemical relaxation times. Langmuir **1991**, *7* (8), 1590–1594.
94. Turner, M.S.; Marques, C.; Cates, M.E. Dynamics of wormlike micelles: the ''bond-interchange'' reaction scheme. Langmuir **1993**, *9* (3), 695–701.
95. Lu, B. Characterization of drag reducing surfactant systems by rheology and flow birefringence measurements. Ph.D. dissertation, The Ohio State University, Columbus, OH, 1997.
96. Lin, Z. The effect of chemical structures of cationic surfactants or counterions on solution drag reduction effectiveness, rheology and micellar microstructure. Ph.D. dissertation, The Ohio State University, Columbus, OH, 2000.

Dust Explosion Hazard Assessment and Control

Vahid Ebadat
Chilworth Technology, Inc., Princeton, New Jersey, U.S.A.

INTRODUCTION

Statistics clearly show that fire/explosion hazards could exist in any plant/equipment that handles or processes a combustible dust. The consequences of a dust explosion can range from disruption of production to loss of plant and injury or fatality of personnel.

To eliminate/control potential dust cloud explosion hazards the following information is required:

- Understanding of the explosion characteristics of the dust(s).
- Identification of locations where combustible dust cloud atmospheres are or could be present during normal and abnormal operating conditions.
- Identification of potential ignition sources that could be present under normal and abnormal conditions.
- Proper plant design to eliminate and/or minimize the occurrence of dust explosions and protect people and plant against their consequences.

The intent of this document is to provide the readers with basic information on how to identify, assess, and eliminate/control potential dust explosion hazards in their facilities. More detailed information on the topics that are covered in this document may be obtained from Bibliography.

CONDITIONS REQUIRED FOR DUST EXPLOSIONS TO OCCUR

A number of conditions must exist simultaneously and in one location for a dust explosion to occur:

Dust must be combustible: (As far as dust clouds are concerned, the terms ''combustible,'' ''flammable,'' and ''explosible'' all have the same meaning and could be used interchangeably.)

The first stage of a dust explosion hazard assessment is to determine whether the dust will explode when dispersed as a cloud. The combustibility of a dust can be assessed by conducting an explosion classification test on a representative sample of the dust. In this laboratory test the observation of flame propagation determines whether or not a suspended dust is capable of initiating and sustaining an explosion. It should be noted that some powders that are noncombustible at ambient temperature conditions could become combustible at elevated temperatures. Therefore, if a noncombustible dust cloud is to be subjected to above ambient temperatures as a part of the process the explosion classification test should be conducted at the operating (elevated) temperature.

Dust must be airborne: If a layer of dust on a surface is ignited, the burning process will be relatively slow, releasing heat over a long period of time. However, if a sufficient concentration of finely divided dust particles are suspended in air and ignited, because of the availability of oxygen around each particle, the combustion process will be very rapid giving rise to a dust explosion.

Dust concentrations must be within the explosible range: If the concentration of a dust cloud is below the minimum explosible concentration (MEC) an explosion cannot propagate. The ease of ignition of a dust cloud and also the resulting explosion violence increase as the dust cloud concentration is increased above the MEC until an optimum concentration is reached causing the highest explosion violence. At higher dust cloud concentrations the explosion violence decreases. The MEC of dust clouds is typically in the range of 10–500 gm^{-3}.

The maximum explosible concentration—the concentration above which an explosion cannot propagate because of lack of sufficient oxidant—is not normally well defined for dust clouds.

The MEC of dust clouds can be determined by laboratory testing.

Dust must have particle size distribution capable of propagating flame: The sensitivity of a dust cloud to ignition and the resulting explosion violence (severity) increase with a decrease in particle size. This is because the combustion process involves chemical reaction at the solid–oxidant interface. Therefore, as the dust particles get finer the total surface area that is available for oxidation will increase. It should be noted that, in practice, very often dust clouds are made up of particles with sizes ranging from fine to coarse. As the fines become airborne more readily, they play a more prominent role in the initial ignition and explosion propagation.

The atmosphere in which the dust cloud is present must be capable of supporting combustion: A dust

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007763

cloud explosion will only occur if there is a sufficient amount of oxidant available. In practice, the oxygen in air is normally the most common oxidant. Other oxidants include chlorine, nitrous oxide, nitric oxide, and nitrogen tetraoxide.

Explosion prevention can be accomplished by the depletion of oxidant. The concentration of oxidant below which combustion cannot occur in a specific mixture is referred to as the limiting oxidant concentration (LOC).

Generally, combustible organic compounds are unlikely to propagate flame if the oxygen content of the atmosphere in which the dust cloud is present is below 8% v/v, using nitrogen or carbon dioxide as the inert gas. Other inert gases include argon, helium, steam, and flue gas (waste gas from on-site processes).

The LOC should be determined by laboratory testing.

An ignition source with sufficient energy to initiate flame propagation must be present: The ignition sources that have been found to be the cause of the majority of explosions in dust handling/processing plants include welding and cutting, heating and sparks generated by mechanical failure, sparks generated by mechanical impacts, flames and burning materials, self heating, electrostatic discharges, electrical sparks, etc.

The sensitivity of a dust cloud to ignition by different ignition sources could be determined through appropriate laboratory tests.

LABORATORY TESTING TO ASSESS EXPLOSION CHARACTERISTICS OF DUST CLOUDS

To assess the possibility of an explosion in a facility and to select the most appropriate basis of safety, explosion characteristics of the dust(s) that are being handled/processed in the facility should be determined.

The explosion characteristics of powders normally fall within one of two groups, ''likelihood of an explosion'' and ''consequences of an explosion.'' These two groups of tests are discussed below.

Laboratory Tests to Determine the Likelihood of an Explosion

The following tests provide information on the likelihood of a dust explosion:

Explosion classification test

The explosion classification test determines whether a dust cloud will explode when exposed to an ignition source, the test results in a material being classified as either combustible or noncombustible.

The explosion classification test is usually conducted in a modified Hartmann tube apparatus. The apparatus consists of a 1.2 L vertical tube mounted onto a dust dispersion system. Dust samples of various quantities are dispersed in the tube and attempts are made to ignite the resulting dust cloud by a 10 J electrical arc ignition source. If the material fails to ignite in the modified Hartmann tube apparatus, the testing is continued in the 20 L sphere apparatus. Dust samples of various quantities are dispersed inside the sphere and are exposed to a 10,000 J ignition source.

Minimum explosible concentration

The MEC test determines the lowest concentration of dust cloud in air that can give rise to flame propagation on ignition. The test involves dispersing a sample of the dust in a 20 L sphere apparatus and attempting to ignite the resulting dust cloud with an energetic ignition source. Trials are repeated for decreasing sample concentrations until the MEC is determined. The MEC of a given dust cloud is influenced by the size of the ignition source. An increase in the size of the ignition source will result in a lower MEC value.

The MEC test is performed in accordance with International Standards Organization (ISO) method 618411 or American Society for Testing and Materials (ASTM) E1515.

Minimum ignition temperature

The minimum ignition temperature (MIT) test determines the lowest temperature capable of igniting a dust dispersed in the form of a cloud. The MIT is an important factor in evaluating the ignition sensitivity of dusts to such ignition sources as heated environments, hot surfaces, electrical apparatus, and friction sparks.

The MIT test is performed in accordance with ASTM E-2021 and the European Standard 61241-2-1. Dust samples of various sizes are dispersed into the furnace and the minimum furnace temperature capable of igniting the dust cloud at its optimum concentration for ignition is determined.

The MIT value is influenced by particle size and moisture content of the dust. A decrease in particle size and moisture content of the dust particles results in a lower MIT value.

Minimum ignition energy

The minimum ignition energy (MIE) test determines the lowest electrostatic spark energy that is capable of igniting a dust cloud at its optimum concentration for ignition. The test is used primarily to assess the potential susceptibility of dust clouds to ignition by electrostatic discharges.

Minimum ignition energy test is performed in accordance with ASTM E2019, British Standard 5958, and International Standard: IEC 1241-2-3 using the modified Hartmann tube apparatus. Dust samples of various sizes are dispersed in a 1.2 L vertical tube and attempts are made to ignite the resultant dust cloud with discrete capacitive sparks of known energy.

The MIE value is influenced by particle size and moisture content of the dust and by process conditions, such as temperature and oxidant content. A decrease in particle size and moisture content of the dust particles results in a lower MIE value. An increase in the temperature of the atmosphere in which the dust cloud is suspended will result in a decrease in MIE.

Electrostatic volume resistivity

Volume resistivity is a measure of the electrical resistance for a unit volume of material and is the primary criterion for classifying powders as low, moderate, or high insulating. Insulating powders have a propensity to retain electrostatic charge and can produce hazardous electrostatic discharges when exposed to grounded plant, equipment, or personnel.

Volume resistivity can be measured in accordance with British Standard 5958. The method involves placing a powder sample into a standardized electrode cell. A voltage is applied to the cell and the current through the powder is measured. Volume resistivity is calculated using the known voltage, the measured current, and the geometrical relationship between the electrodes.

Because of the effect of atmospheric and absorbed moisture on volume resistivity, it is usually suggested that this test be performed at ambient and low relative humidity conditions.

Electrostatic chargeability

The concept of electrostatic chargeability refers to the propensity of powder particles to become charged when flowing through conveyances or when handled in containers. Electrostatic chargeability is measured by having samples flow through a pipe and measuring the resultant electrostatic charge. The test provides data that can be used to develop appropriate material handling guidelines.

Because of the effect of atmospheric and absorbed moisture on powder chargeability, this test is usually performed at ambient and low relative humidity conditions.

Limiting oxidant concentration

The LOC test determines the minimum concentration of oxygen (displaced by an inert gas such as nitrogen) capable of supporting combustion. An atmosphere having an oxygen concentration below the LOC is not capable of supporting combustion and thus cannot support a dust explosion. The LOC test is used to study explosion prevention or severity reduction involving the use of inert gases and to set oxygen concentration alarms or interlocks in inerted plants and vessels.

Limiting oxidant concentration testing can be performed using the 20 L sphere apparatus. Dust samples of various sizes are dispersed in the vessel and attempts are made to ignite the resulting dust cloud with an energetic ignition source. Trials are repeated for decreasing oxygen concentrations until the LOC is determined.

Limiting oxidant concentration of a given dust cloud is dependent on the type of the inert gas that is used to replace the oxidant of the atmosphere as well as some process conditions such as temperature. Therefore, LOC testing should simulate the process conditions and be performed by using an inert gas that is representative of the inert gas used in practice.

Laboratory Tests to Determine the Consequences of an Explosion

There are two laboratory tests (measurements) to determine the consequences of an explosion: maximum explosion pressure and maximum rate of pressure rise.

The maximum explosion pressure and maximum rate of pressure rise values are determined by using the 20 L sphere apparatus. The dust sample is dispersed within the sphere, ignited by chemical igniters, and the pressure of the resulting explosion is measured. The sample size is varied to determine the optimal dust cloud concentration. The maximum pressure and rate of pressure rise are measured and used to calculate the explosion severity (K_{st}) value of the dust cloud. These data can be used for the purpose of designing dust explosion protection measures.

Explosion severity testing is performed in accordance with the current ASTM Method E 1226, National Fire Protection Association (NFPA) Standard 68, German Society of Engineers (VDI) Method 3673, and ISO Method 6184/1.

BASIS OF SAFETY FROM DUST CLOUD EXPLOSIONS

Safety from potential dust cloud explosions could include taking measures to avoid an explosion (explosion prevention) or designing plant and equipment so that in the event of an explosion people and plant are protected (explosion protection). Selection of

explosion prevention and/or protection measures is usually based on:

- How much information is available on the sensitivity of the powder(s) to ignition and the resulting explosion severity.
- Nature of the processes and operations.
- Level of personnel's knowledge and appreciation regarding the consequences of a potential dust explosion and adherence to preventive measures.
- Environmental effects of a potential dust explosion.
- Business interruptions resulting from a potential dust explosion.

Explosion Prevention Measures

The risk of an explosion is removed when one of the following measures is taken.

An explosible dust cloud is never allowed to form

There are two main methods of ventilation for eliminating or controlling the spread of explosible atmospheres (fuel):

Dilution Ventilation. Dilution ventilation provides a flow of fresh air into and out of the building. This normally results in a reduction of the background concentration of the flammable atmosphere in the working area but there is no control of the flammable atmosphere at the source of release. This method is not practical for controlling the concentration of dust cloud atmospheres but is often used to control vapor concentrations.

Local Exhaust Ventilation. Local exhaust ventilation is designed to intercept the flammable atmosphere at the source of release and directs it into a system where air is safely separated from the fuel. Correctly designed local exhaust ventilation systems could be very effective in limiting the spread of dust cloud atmospheres beyond the source of release. Local exhaust ventilation is generally less expensive to run than dilution ventilation because less air is used.

A local exhaust ventilation system generally includes hood, ductwork, filter, and fan.

To ensure that the local exhaust system itself does not become an explosion hazard; the following suggestions should be considered:

- Exhaust air velocity in the ductwork should be high enough to prevent the powder from depositing in the ductwork.
- While extracting from several hoods or sources into a common duct, each branch duct should deliver the intended air volume at the required conveying velocity. It is preferable that the air balancing is achieved without using dampers. Air balancing without dampers safeguards the system against tampering with airflow rates, and also eliminates the possibility of dust accumulation behind the dampers.
- The degree of enclosure provided by the hood and the air capture velocity should be such that dust particles are prevented from entering the workplace.
- The air cleaning devices (air filters, dust collectors, cyclones, etc.) should be protected against the consequences of dust cloud explosions. This includes measures to prevent propagation of the explosion back through the exhaust ducting to other areas of plant and equipment.
- All the components of the local exhaust system should be made from conductive and/or static dissipative materials and electrically grounded. This includes the hoods, ductwork, filter housing, and metal support cages for filter socks (if any).

The atmosphere is sufficiently depleted of oxidant (normally the oxygen in air) so that it cannot support combustion

Safety may be based on reducing the oxidant concentration below a level that will no longer support combustion (LOC), by adding an inert gas. Nitrogen gas is perhaps the most commonly used inert gas. Other inert gases include carbon dioxide, argon, helium, steam, and flue gas (waste gas from on-site processes). Oxidant can also be removed by working under vacuum (safe vacuum pressure should be obtained by testing).

Limiting oxidant concentration for combustion is dependent on the type of dust and type of inert gas used. Once the LOC of the dust has been determined for the inert gas that will be used, the inert gas needs to be introduced into the vessel. Successful inert gas blanketing will only be possible if the entire volume of the vessel is inerted and the inert atmosphere is maintained throughout even when the vessel is opened during the addition of solids and/or liquids to the vessel.

The following techniques are commonly used to achieve an inert atmosphere in a vessel:

Pressure Purging. The vessel is pressurized with an inert gas and then relieved outside. This procedure is repeated until the desired oxygen concentration is reached. The number of pressure purges, n, required to achieve the desired LOC can be calculated by using the following equation:

$$n = \ln(21/\mathrm{LOC})/\ln P$$

where P is purge pressure (bar absolute) and n is the number of pressure purges.

In the above equation it is assumed that:

- Inert gas contains no oxygen.
- Initial oxygen concentration is 21%.

Vacuum Purging. The vessel is evacuated and then is increased to atmospheric pressure using the chosen inert gas. This procedure is repeated until the desired oxygen concentration is reached. The number of vacuum purges, n, required to achieve the desired LOC can be calculated by using the following equation:

$$n = \ln(\mathrm{LOC}/21)/\ln P$$

where P is vacuum pressure (bar absolute) and n is the number of vacuum purges.

In the above equation it is assumed that:

- Inert gas contains no oxygen.
- Initial oxygen concentration is 21%.

Flow-Through Purging. The vessel is purged with a continuous flow of inert gas. Flow-through purging will only be successful if there is a means of mixing the inert gas within the entire volume of the vessel. The required purge time, t, required to achieve the desired LOC is given by the following equation:

$$t = -V/\mathrm{Q}\log(\mathrm{LOC}/21)$$

where Q is purge gas flow rate (m^3/sec) and V is vessel volume.

All ignition sources capable of igniting the dust cloud are removed

The ignition sources that have been found to be the cause of the majority of explosions in dust handling/processing plants include welding and cutting, heating and spark generated by mechanical failure, mechanical impacts, flames and burning materials, self-heating, electrostatic discharges, and electrical sparks. This is not an exhaustive list.

Elimination of ignition sources involves the following steps.

Control of Heat Sources. Examples of heat sources are the following:

- External surfaces of hot process equipments such as heaters, dryers, steam pipes, and electrical equipment.
- Mechanical failure of equipment such as bearings, blowers, mechanical conveyers, mills, mixers, and unprotected light bulbs.
- Hot work.

A hot surface may directly ignite a dust cloud or first ignite a dust layer that may have settled on it and subsequently ignite a dust cloud.

Measures that may be considered for preventing a dust cloud ignition by heat sources include:

- Maintaining an effective housekeeping program to prevent/remove dust accumulations on potential hot surfaces.
- Maintaining the temperature of the processing equipment below the self-heating temperature of the powder.
- Providing regular inspection and maintenance of the processing plant to prevent overheating due to misalignment, loose objects, belt-slip/rubbing, etc.
- Preventing the overloading of processing equipment, such as grinders and conveyors. Consider installing overload protection devices on drive motors.
- Preventing ''foreign objects'' from entering the processing equipment by use of suitable separation devices, such as electromagnets or pneumatic separators.
- Isolating/shielding dust layers and clouds from hot surfaces.
- Using approved electrical equipment (correct temperature rating).

Control of Friction/Impact Sparks. The ability of friction/impact sparks to ignite flammable atmospheres is dependent, among other factors, on the composition of the impacting surfaces. In particular, incendive sparks could be expected in any one of the following conditions:

- Items constructed from light alloys strike a rusty steel surface.
- A rusty steel surface that has been coated with a layer of paint containing aluminum is struck by a hard object.
- Striking surfaces containing flint, rock, or grit with a hard object.

In any work where friction/impact sparks could be expected the measures that may be considered for preventing a dust cloud ignition include the following:

- Flammable (gas, vapor, and dust clouds) atmospheres should not be present.
- Hard surfaces, such as concrete, brick, or rock, should be kept wet with water. Alternatively, soft

rubber mats may be used to cover the surfaces and act as a cushion for the objects that might fall.

Welding, Cutting, and Similar Hot Work Operations. Flames and sparks that are present during welding/brazing/soldering and cutting and other similar operations could readily ignite dust layers and clouds.

To avoid fires and explosions, measures should be taken to prevent the formation of dust clouds and remove dust deposits from surfaces.

Electrical Equipment and Instruments. Electrical sparks produced during normal working of switches, contact breakers, motors, fuses, etc. can ignite dust layers and clouds.

The requirements for electrical equipment and wiring systems for use in locations where combustible dust clouds or layers may be present during normal and/or abnormal operation of the plant are provided by Article 500 of the National Electrical Code (NEC). The intent of Article 500 is to prevent electrical equipment from providing a means of ignition for an ignitable atmosphere.

Ensuring safety when electrical equipment and wiring systems are present normally involves:

1. Determining the explosion characteristics that are listed below for the dust, considered relative to those for Pittsburgh seam coal:
   a. Minimum explosible concentration, MEC.
   b. Minimum ignition temperature of the dust cloud, MIT.
   c. Minimum ignition energy of the dust cloud, MIE.
   d. Maximum explosion pressure, P_{max}.
   e. Maximum rate of pressure rise, dP/dt.
2. Calculating the ''ignition sensitivity'' and ''explosion severity'' as defined below:

$$\text{Ignition sensitivity} = (\text{MIT} \times \text{MIE} \times \text{MEC})_1 / (\text{MIT} \times \text{MIE} \times \text{MEC})_2$$

$$\text{Explosion severity} = (p_{max} \times dP/dt)_2 / (p_{max} \times dP/dt)_1$$

   Subscript 1 refers to the appropriate explosion characteristics for Pittsburgh seam coal (the standard dust used by the U.S. Bureau of Mines) and subscript 2 refers to the appropriate explosion characteristics for the specific dust present in the plant.
   Using the above equations, dusts having ignition sensitivity equal to or greater than 0.2 or explosion severity equal to or greater than 0.5 are considered combustible (for electrical area classification purposes only).
3. Conducting an electrical area classification: An electrical area classification involves identifying the extent (if any) of location(s) in a facility where combustible materials could be present during normal and/or abnormal conditions. According to Article 500 of the NEC, major categories of hazardous locations are:
   a. Class I, in which the combustible material is a gas or a vapor (NFPA 497).
   b. Class II, in which the combustible material is dust (NFPA 499).

Within Class II hazardous locations, Article 500 recognizes two degrees of hazard: Division 1 and Division 2.

Class II, Division 1 locations are considered to be any of the following:

- Combustible dust is in the air under normal operating conditions in quantities sufficient to produce explosible or ignitable mixtures.
- Mechanical failure or abnormal operation of machinery or equipment might cause such explosible or ignitable mixtures to be produced and might also provide a source of ignition through simultaneous failure of electrical equipment, operation of protection devices, or other causes.
- Group E dusts may be present in hazardous quantities. Group E dusts include combustible metal dusts, including aluminum, magnesium, and their commercial alloys, or other combustible dusts whose particle size, abrasiveness, and conductivity present similar hazards in the use of electrical equipment. The NEC does not recognize any Division 2 areas for such dusts.

Note: Dusts having a volume resistivity less than 1 Ωm are considered conductive and therefore Group E.

Class II, Division 2 are locations where combustible dust is not normally in the air in quantities sufficient to produce explosible or ignitable mixtures, and dust accumulations are normally insufficient to interfere with the normal operation of electrical equipment or other apparatus. However, combustible dust may be in suspension in the air as a result of infrequent malfunctioning of handling or processing equipment and where combustible dust accumulations on, in, or in the vicinity of the electrical equipment may be sufficient to interfere with the safe dissipation of the heat from electrical equipment or may be ignitable by abnormal operation or failure of electrical equipment.

Note: The quantity of combustible dust that may be present and the adequacy of dust removal systems are factors that merit consideration in determining the classification and may result in an unclassified area.

The following factors determine the extent of Class II locations:

- Combustible material involved [for example, the dust conductive (Group E)]
- Bulk density of the material
- Particle sizes of material
- Density of the particles
- Process or storage pressure
- Size of the leak opening
- Quantity of the release
- Dust collection system
- Housekeeping
- Presence of any flammable or combustible gas.

As discussed in this section, the intent of electrical area classification is to ensure that electrical equipment will not act as an ignition source. However, once the extent of a classified area is determined it is prudent to ensure that all potential ignition sources are eliminated or controlled.

Electrostatic Discharges. In this section it is assumed that the powder does not contain any flammable solvent and it is handled and processed in an atmosphere free from flammable gases and vapors.

Electrostatic charge generation: Although the magnitude and polarity of charge is usually difficult to predict, charge generation should almost always be expected whenever powder particles come into contact with another surface. It occurs, for example, during mixing, grinding, sieving, pouring, and pneumatic transfer. The chemical composition and the condition of the contacting surfaces can often influence the charging characteristics.

Electrostatic charge accumulation: Generally, powders are divided into three groups depending on their ability to retain static charge even if the powder is in contact with an electrically ground conductive object. This ability is known as volume resistivity:

1. Powders with volume resistivities up to about $10^6\,\Omega$m are considered conductive.
2. Powders with volume resistivities in the range 10^6–$10^9\,\Omega$m are of medium resistivity.
3. Powders with volume resistivities above $10^9\,\Omega$m are high-resistivity powders.

Charge will accumulate on a powder if the charge generation rate exceeds the rate at which the charge dissipates.

Electrostatic discharges: The accumulation and retention of charge on powder or equipment creates a dust explosion hazard only if the charge is suddenly released in the form of a discharge with sufficient energy to ignite the dust cloud. Potentially incendive discharges resulting from charged powder and equipment include spark discharges, brush discharges, propagating brush discharges, and cone (bulking) discharges.

General Precautions

Bonding and grounding

Spark discharges can be avoided by electrically grounding conductive items, such as metal plant, fiberboard drums, conductive liners, low-resistivity powders, and people.

Use of insulating materials

Where there could be high surface charging processes, nonconductive materials should not be used, unless the breakdown voltage across the material is less than 4 kV. Examples of nonconductive objects include plastic pipes, containers, bags, coatings, and liners.

Charge reduction by humidification

High relative humidity can reduce the resistivity of some powders and increase the rate of charge decay from bulk powder in grounded metal containers. However, in most cases this will only be effective if a relative humidity in excess of 65% is maintained.

Charge reductions by ionization

Localized ionization (corona discharges) from sharp, grounded, conducting probes or wires can on occasions be used to reduce the level of electrostatic charge from powder particles entering a vessel. Electrostatic ionization devices are not, however, without problems, and should only be used after consulting expert advice.

Explosion protection

In some powder handling processes it is not possible to avoid the simultaneous presence of an explosible dust cloud and a hazardous buildup of charge. In those situations, measures should be taken to protect against or prevent explosions. These include inerting, use of explosion-resistant equipment, explosion venting, or explosion suppression.

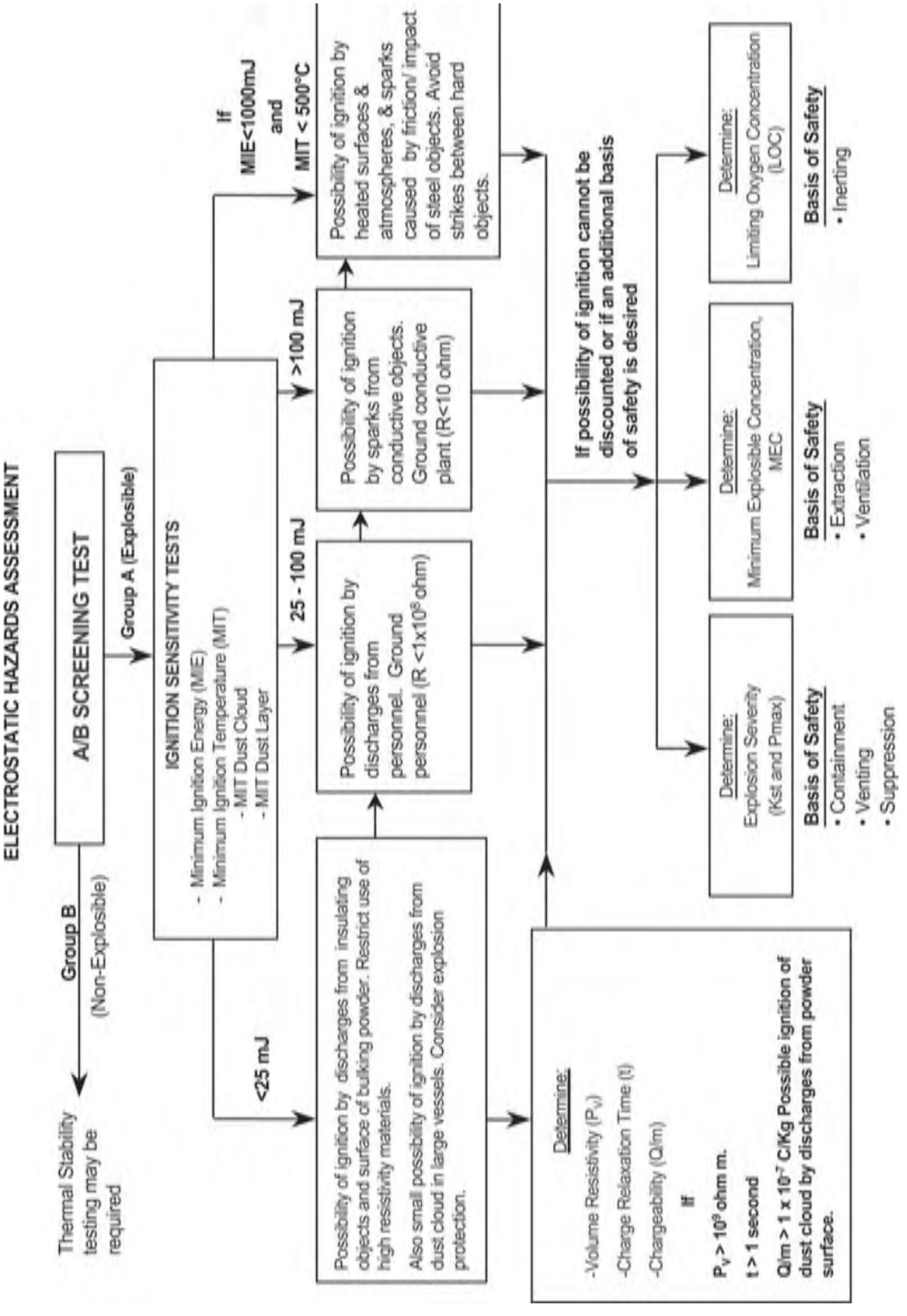


Fig. 1 Electrostatic hazards assessment.

The flow diagram in Fig. 1 provides a summary of the hazards and laboratory tests required to quantify electrostatic properties and measures that may be taken to ensure safety.

Finally, if the basis of safety from potential dust cloud explosions is the elimination of ignition sources, the answers to the following questions should be ''yes'':

- Can all ignition sources be identified?
- Is the sensitivity to ignition by these sources known for all process materials?
- Can all ignition sources be eliminated under normal and abnormal conditions?

If the answer to one or more of the above questions is ''no,'' then in addition to taking all reasonable steps to reduce the possibility of formation/spread of dust clouds and exclude potential ignition sources, other measures, such as exclusion of oxidant or explosion protection, should also be considered.

Explosion Protection Measures

If evolution or formation of an explosible atmosphere cannot be prevented and all sources of ignition cannot be reasonably eliminated or excluded, then the possibility of a dust cloud explosion persists. Under such conditions, explosion protection measures should be taken to protect people and minimize damage to the plant. It should be noted that explosion protection measures should be considered in addition to taking all reasonable steps to reduce the possibility of formation/spread of dust clouds and exclude potential ignition sources. Explosion protection measures include the following.

Explosion containment

Construct the plant to withstand the maximum explosion pressure resulting from the deflagration (propagation of a combustion zone at a velocity that is less than the speed of sound in the unreacted medium) of the dust present in the equipment.

The following guidelines should be applied when considering explosion containment as a basis of safety:

- Equipment should withstand the maximum explosion pressure that is expected for the dust present under the process (pressure, temperature, etc.) conditions.
- All interconnected pipes, flanges, covers, etc. should withstand the maximum explosion pressure of the dust being handled.
- If an explosion-resistant vessel fails, the pressure effects will be more severe than if an extremely weak vessel fails as a result of a dust explosion.
- Because of the relatively high cost of constructing plant to withstand the maximum explosion pressure (plus a safety factor), explosion containment is normally considered for smaller equipment.

Explosion suppression

Explosion suppression is detecting an explosion at an early stage and suppressing it with a suitable suppressant. Explosion suppression relies on early detection of an explosion and rapid injection of the suppressant. A typical explosion pressure at the moment of detection is 0.035–0.1 barg. Suppressant extinguishes the flame within 0.08 sec. An explosion suppression system normally includes explosion detector, control unit, suppressor, and a suitable suppressant.

To achieve explosion suppression, the following information is required:

- Type of powder
- Explosion severity (K_{st}) of the dust cloud
- Strength (P_{red}) of the vessel to be protected
- Dimensions and volume of the vessel to be protected
- Pressure and temperature of the vessel.

Explosion suppression has the following advantages and disadvantages compared to explosion relief venting:

- Extinguishing the flame.
- Reducing the risk of ejecting toxic and/or corrosive materials to the atmosphere.
- Process equipment does not need to be located in an area suitable for explosion relief venting.
- Suppression systems are generally more expensive to install and maintain than explosion relief venting.
- Some suppression systems are not suitable for powders with very high explosion severity (K_{st}) (generally above 300 bar/m/sec, although the limits are being extended all the time).

Explosion relief venting

Explosion relief venting is the process of relieving the explosion products (pressure and flame) from the plant to a safe location. The principle of explosion relief venting is that a dust explosion in a vessel causes vent(s) of sufficient area to open rapidly and relieve hot gases and dust to a safe location. In other words, the vessel ''fails'' in a predictable way such that people and plant are protected from the effects of the dust explosion.

Explosion relief venting has the advantage of being relatively inexpensive compared to other explosion protection options and is simple to install in many cases.

Some of the limitations of explosion relief venting include:

- Not suitable for toxic materials some of which could be released to the atmosphere in the event of explosion.
- Very strongly explosible dusts, such as explosives or industrial powders, with K_{st} values in excess of 600 bar m/sec would be expected to be very difficult to vent because of the lack of available space on the vessel to be protected.
- Venting of explosions to inside a building is not usually acceptable.

Explosion Relief Venting Requirements. Three elements are required for the design and specification of explosion relief vent(s) for a vessel:

- Information on the explosion severity of the airborne dust present is required. This is generally obtained by subjecting representative dust samples to a 20 L sphere explosion test. This measures the explosion severity as expressed by maximum explosion pressure and maximum rate of pressure rise (K_{st} index).
- Shape, volume, pressure rating (P_{red}), and location of the vessel in relation to outside walls, available vent area on the vessel, and vent activation pressures (P_{stat}).
- Vent sizing method based on the dust type and plant to be protected. In the United States, NFPA 68 provides the required information for vent sizing calculations.

Explosion Isolation Measures

Regardless of what type of explosion protection measure is considered, the dust cloud explosion should be prevented from propagating from the location where it originates to other locations in the plant. This is referred to as explosion isolation. Dust explosions can propagate through pipes, chutes, conveyors, etc. The first step in isolating an explosion is to avoid unnecessary connections. If this is not possible, barriers should be created in the path of the explosion.

There are two types of barriers that could be considered for isolating a dust explosion:

Mechanical barriers: Explosion propagation is prevented by some type of physical barrier. Mechanical barriers could include rotary valves that have a sufficient number of blades to form a barrier, screw feeders that are modified to continuously contain a plug of material, and fast acting shutoff valves.

Chemical barriers: Flame front or pressure wave is detected and a suitable suppressant is injected to extinguish the flame. Although chemical barriers extinguish the flame, they cannot prevent the explosion pressure from propagating. Downstream process equipment should therefore be able to withstand the resulting pressure.

CONCLUSIONS

The majority of powders that are used in the processing industries are combustible (also referred to as flammable and explosible). An explosion will occur if the concentration of the combustible dust that is suspended in air is sufficient for flame propagation when ignited by a sufficiently energetic ignition source.

A systematic approach to identifying potential dust cloud explosion hazards and taking measures to ensure safety against their consequences generally involves the following:

1. Determining the dust cloud's ignition sensitivity and explosion severity characteristics.
2. Identifying areas of the plant where combustible dust cloud atmospheres could exist under normal and/or abnormal conditions.
3. Identifying potential ignition sources.
4. Taking measures to eliminate/control
   a. Potential ignition sources
   b. Combustible dust cloud concentrations
   c. Oxidant concentrations.
5. Taking measures to protect against the consequences of potential dust cloud explosions. Explosion protection measures include:
   a. Explosion relief venting
   b. Explosion suppression
   c. Explosion containment
   d. Explosion isolation.

BIBLIOGRAPHY

Abbott, J.A. *British Materials Handling Board Survey of Dust Fires and Explosions in the United Kingdom 1979-1984*; Warren Spring Laboratory: Hertfordshire, U.K., 1986.

Bartknecht, W. *Dust Explosions Course, Prevention, Protection*; Springer-Verlag: Germany, 1989.

Barton, J. *Dust Explosion Prevention and Protection, a Practical Guide*; Institution of Chemical Engineers: Rugby, Warwickshire, U.K., 2002.

Bodurtha, F.T. *Industrial Explosion Prevention and Protection*; McGraw-Hill Book Company: New York.

Buschart, R.J. *Electrical and Instrumentation Safety for Chemical Processes*; Van Nostrand Reinhold: New York, 1991.

Eckhoff, R.K. *Dust Explosion in the Process Industries*, 2nd Ed.; Butterworth-Heinemann Linacre House: Jordan Hill, Oxford, 1997.

NFPA 77. *Recommended Practice on Static Electricity*; 2000.

NFPA 68. *Guide for Venting of Deflagrations*; 2002.

NFPA 69. *Explosion Prevention Systems*; 1992.

NFPA 91. *Standard for Exhaust Systems for Air Conveying of Vapors, Gases, Mists, and Noncombustible Particulate Solids*; 1999.

NFPA 497B. *Classification of Class II Hazardous (Classified) Locations for Electrical Installations in Chemical Process Areas*; 1991.

NFPA 497M. *Manual for Classification of Gases, Vapors, and Dusts for Electrical Equipment in Hazardous* (*Classified Locations*); 1991.

NFPA 654. *Standard for the Prevention of Fire and Dust Explosions from the Manufacturing, Processing and Handling of Combustible Particulate Solids*; 2000.

Van Laar. G.F.M. *Review of Incidents*; Europex Symposium, Dust Explosion Protection, Belgium, Sep 1989, Belgium.

Dynamic Mechanical Thermal Analysis

Kevin P. Menard
PerkinElmer Thermal Laboratory, Materials Science Department, University of North Texas, Denton, Texas, U.S.A.

INTRODUCTION

Dynamic mechanical analysis (DMA), or in the older literature dynamic mechanical thermal analysis (DMTA), is the technique of applying a stress or strain to a sample and analyzing the response to obtain phase angle and deformation data. This data allows the calculation of the damping or tan delta (δ) as well as complex modulus and viscosity data. Two approaches are used: a) forced frequency, where the signal is applied at a set frequency and b) free resonance, where the material is perturbed and allowed to exhibit free resonance decay. Most DMAs are the former type while the torsional braid analyzer (TBA) is the latter. In both approaches, the technique is very sensitive to the motions of the polymer chains, and it is a powerful tool for measuring transitions in polymers. It is estimated to be 100 times more sensitive to the glass transition than differential scanning calorimetry (DSC), and it resolves other more localized transitions not detected in the DSC. In addition, the technique allows the rapid scanning of a material's modulus and viscosity as a function of temperature or of frequency. Frequency scanning can be used to study a materials apply to damp or sustain vibrations. Despite the power of these techniques, only a few books have concentrated on them.[1] Modern DMA analyzers are either forced oscillation or free resonance: the former is mby for the more common technique. In a forced oscillation instrument, a load is applied to a sample sinusoidally and the sample will deform sinusoidally.

INSTRUMENTATION

One of the biggest choices made in selecting a DMA is to decide whether to choose stress (force) or strain (displacement) control for applying the deforming load to the sample. Because most DMA experiments run at very low strains ($\sim$0.5% maximum) to stay well within a polymers linear region, both analyzers give the same results.

Dynamic mechanical analyzers are normally built to apply the stress or strain in three ways (Fig. 1). One can apply force in a twisting motion; so one is testing the sample in torsion. Axial analyzers are normally designed for solid and semisolid materials, and apply a linear force to the sample. Finally, one can allow the sample to swing freely and damp the oscillations in a TBA.

APPLICATIONS

Thermoplastic Solids and Cured Thermosets

As mentioned above, the thermal transitions in polymers can be described in terms of either free volume changes[2] or relaxation times. A simple approach to looking at free volume, which is popular in explaining DMA responses, is the crankshaft mechanism,[3] where the molecule is imagined as a series of jointed segments. From this model, it is possible to simply describe the various transitions seen in a polymer. Other models exist that allow for more precision in describing the behavior; the best seems to be the Doi–Edwards Model.[4] Aklonis and Knight[5] give a good summary of the available models, as does Rohn.[6]

The crankshaft model treats the molecule as a collection of mobile segments that have some degree of free movement. This is a very simplistic approach, yet very useful for explaining behavior (Fig. 2). As the free volume of the chain segment increases, its ability to move in various directions also increases. This increased mobility in either side chains or small groups of adjacent backbone atoms results in a greater compliance (lower modulus) of the molecule. These movements have been studied, and Heijboer[7] classified β and γ transitions by their type of motions. The specific temperature and frequency of this softening help drive the end use of the material.

Moving from very low temperature, where the molecule is tightly compressed, to higher temperatures the first changes are the solid-state transitions. This process is shown in Fig. 3. As the material warms and expands, the free volume increases so that localized bond movements (bending and stretching) and side chain movements can occur. This is the gamma transition, T_γ, which may also involve associations with water. As the temperature and the free volume continue to increase, the whole side chains and localized groups of 4–8 backbone atoms begin to have enough space to move, and the material starts to develop some toughness.[8] This transition, called the beta transition,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007745

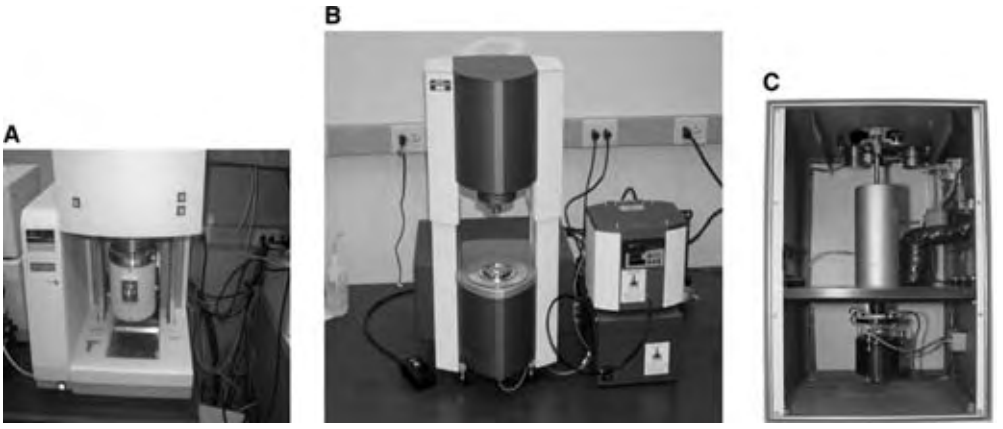


Fig. 1 Torsion vs. axial analyzers: The PerkinElmer Diamond DMA (A) is an axial analyzer while the ATS Rheo; (B) is a torsional instrument. Both are controlled stress but can act as strain controlled because of the feedback programmed in. Photos taken by the author of Equipment at the University of North Texas (C) shows the TBA. (Courtesy of Dr. John Enns of Polymer Network Characterizations, Inc.) *(View this art in color at www.dekker.com.)*

T_β, is not as clearly defined as described here. Often it is the T_g of a secondary component in a blend or a specific block in a block copolymer. However, a correlation with toughness is seen empirically.[9]

As heating continues, the T_g or glass transition appears when the chains in the amorphous regions begin to coordinate large-scale motions. One classical description of this region is that the amorphous regions have begun to melt. As the T_g only occurs in amorphous material, in a 100% crystalline material, there would not be a T_g. Continued heating drives the material through the $T_\alpha{}^*$ and T_{ll}. The former occurs in crystalline or semicrystalline polymer, and is a slippage of the crystallites past each other. The latter is a movement of coordinated segments in the amorphous phase that relates to reduced viscosity. These two transitions are not universally accepted. Finally, the melt is reached where large-scale chain slippage occurs and the material flows. This is the melting temperature, T_m. For a cured thermoset, nothing happens after the T_g until the sample begins to burn and degrade because the crosslinks prevent the chains from slipping past each other.

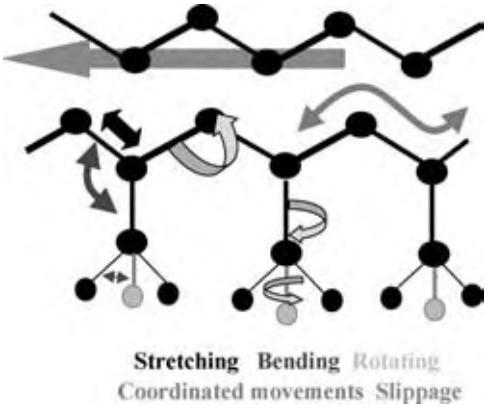


Fig. 2 The crankshaft mechanism is a simple way of considering the motions of a polymer chain permitted by increases in free volume. The molecule is visualized as a series of balls and rods, and these move as the free volume increases. *(View this art in color at www.dekker.com.)*

Sub-T_g transitions

The area of sub-T_g or higher order transitions has been heavily studied[10] as these transitions have been associated with mechanical properties. These transitions can sometimes be seen by DSC and TMA, but they are normally too weak or too broad for determination by these methods. DMA, DEA, and similar techniques are usually required.[11] Some authors have also called these types of transitions[12] as second order transitions to differentiate them from the primary transitions of T_m and T_g, which involve large sections of the main chains. Boyer reviewed the T_β in 1968[13] and pointed out that while a correlation often exists, the T_β is not always an indicator of toughness. Bershtien[14] has reported that this transition can be considered the ''activation barrier'' for solid-phase reactions, deformation, flow or creep, acoustic damping, physical aging changes, and gas diffusion into polymers as the activation energies for the transition and these processes are usually similar. The strength of these transitions is related to how strongly a polymer responses to those processes. These sub-T_g transitions are associated with the materials properties in the glassy state. In paints, for example, peel strength (adhesion) can be estimated from the strength and frequency dependence of the sub-ambient β transition.[15] Cheng[16] reports in rigid rod polyimides that the β

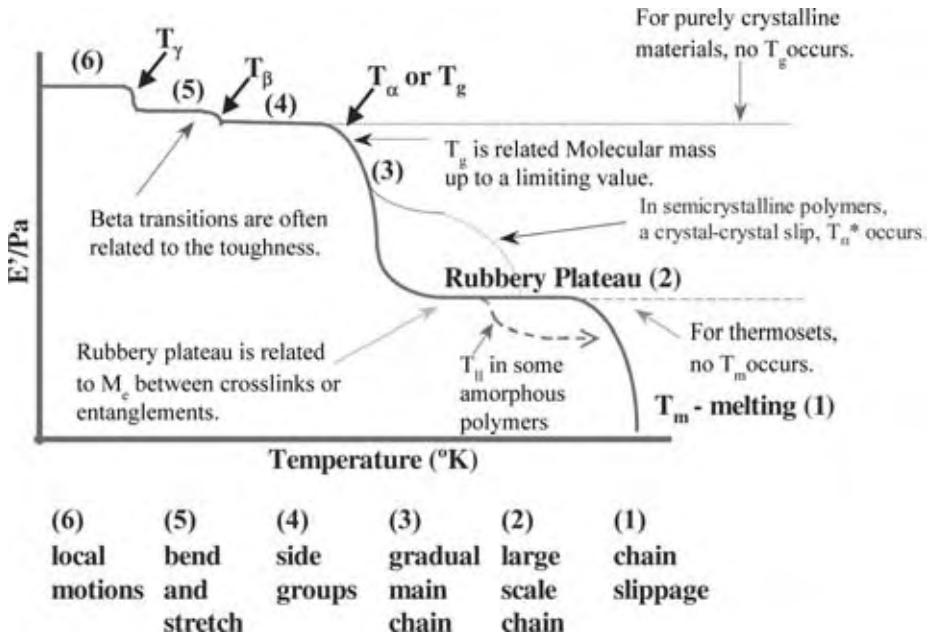


Fig. 3 Idealized temperature scan of a polymer: starting at low temperature, the modulus decreases as the molecules gain more free volume resulting in more molecular motion. This shows main curve as divided into six regions, which corresponds to: local motions (6), bond bending and stretching (5), movements in the side chain or adjacent atoms in the main chain (4), the region of the T_g (3), coordinated movements in the amorphous portion of the chain (2), and the melting region (1). Transitions are marked as described in the text. (From Ref.[1].) *(View this art in color at www.dekker.com.)*

transition is caused by the noncoordinated movement of the diamine groups although the link to physical properties was not investigated. Johari has reported in both mechanical[17] and dielectric studies[18] that both the β and γ transitions in bisphenol-A-based thermosets depend on the side chains and unreacted ends, and that both are affected by physical aging and postcure. Nelson[19] has reported that these transitions can be related to vibration damping. This is also true for acoustical damping.[20] In these cases, the strength of the β transition is taken as a measurement of how effectively a polymer will absorb vibrations.

Boyer[21] and Heijboer[7] showed that this information needs to be considered with care as not all β transitions correlate with toughness or other properties. This can be due to misidentification of the transition or that the transition does sufficiently disperse energy. A working rule of thumb[22] is that the β transition must be related to either localized movement in the main chain or very large side chain movement to sufficiently absorb enough energy. The relationship of large side chain movement and toughness has been extensively studied in polycarbonate by Yee[23] as well as in many other tough glassy polymers.[24]

Less use is made of the T_γ transitions and they are mainly studied to understand the movements occurring in polymers. Wendorff[25] reports that this transition in polyarylates is limited to inter- and intramolecular motions within the scale of a single repeat unit. Both Mc Crum[26] and Boyd[27] similarly limited the T_γ and T_δ to very small motions either within the molecule or with bound water. The use of what is called 2D-IR, which couples at FTIR and a DMA to study these motions, is a topic of current interest.[28]

The glass transition (T_g or T_α)

As the free volume continues to increase with increasing temperature, the glass transition, T_g, occurs where large segments of the chain start moving. This transition is also called the alpha transition, T_α. The T_g is very dependent on the degree of polymerization up to a value known as the critical T_g or the critical molecular weight. Above this value, the T_g typically becomes independent of molecular weight.[29] The T_g represents a major transition for many polymers, as physical properties change drastically as the material goes from a hard glassy to a rubbery state. It defines one end of the temperature range over which the polymer can be used, often called the operating range of the polymer. For where strength and stiffness are needed, it is normally the upper limit for use. In rubbers and some semicrystalline materials like polyethylene and polypropylene, it is the lower operating temperature. Changes in the temperature of the T_g are commonly used to monitor changes in the polymer, such as plasticizing by environmental solvents and increased crosslinking from thermal or UV aging.

The T_g of cured materials or thin coatings is often difficult to measure by other methods and more often than not the initial cost justification for a DMA is measuring a hard to find T_g. While the estimates of the relative sensitivity of DMA to DSC or DTA vary, it appears that DMA is 10–100 times more sensitivity to the changes occurring at the T_g. The T_g in highly crosslinked materials can easily be seen long after the T_g has become too flat and broad to be seen in the DSC. This is also a problem with certain materials like medical grade urethanes and very highly crystalline polyethylenes.

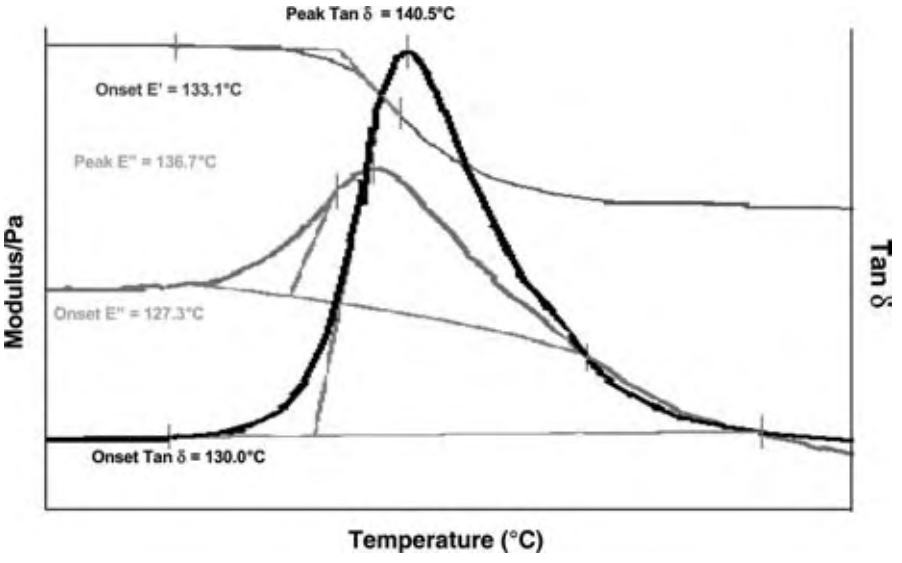


Fig. 4 Methods of determining the T_g are shown for the DMA. The temperature of the T_g varies by as much as 10°C in this example depending on the value chosen. Differences as great as 25°C have been reported. (From Ref.[1].) *(View this art in color at www.dekker.com.)*

The method of determining the T_g in the DMA can be a manner for disagreement as at least five ways are in current use (Fig. 4). Depending on the industry standards or background of the operator, the peak or onset of the tan δ curve, the onset of the E' drop, or the onset or peak of the E'' curve may be used. The values obtained from these methods can differ up to 25°C from each other on the same run. In addition, a 10–20°C difference from the DSC is also seen in many materials. In practice, it is important to specify exactly how the T_g should be determined. For DMA, this means defining the heating rate, applied stresses (or strains), the frequency used, and the method of determining the T_g. For example, the sample will be run at $10°C\,min^{-1}$ under 0.05% strain at 1 Hz in nitrogen purge ($20\,cc\,min^{-1}$), and the T_g determined from peak of the tan δ curve.

It is not unusual to see a peak or hump on the storage modulus directly preceding the drop that corresponds to the T_g. This is also seen in the DSC and DTA, and corresponds to a rearrangement in the material to relieve stresses frozen below the T_g by the processing method. These stresses are trapped in the material until enough mobility is obtained at the T_g to allow the chains to move to a lower energy state. Often a material will be annealed by heating it above the T_g and slowly cooling it to remove this effect. For similar reasons, some experimenters will run a material twice or use a heat-cool-heat cycle to eliminate processing effects.

The rubbery plateau, T_α^* and T_{ll}

The area above the T_g and below the melt is known as the rubbery plateau, and the length of it as well as its viscosity are dependent on the molecular weight between entanglements (M_e)[30] or crosslinks. The molecular weight between entanglements is normally calculated during a stress-relaxation experiment but similar behavior is observed in the DMA. The modulus in the plateau region is proportional to either the number of crosslinks or the chain length between entanglements. This is often expressed in shear as:

$$G' \cong (\sigma RT)/M_e \tag{1}$$

where G' is the shear storage modulus of the plateau region at a specific temperature, ρ is the polymer density, and M_e is the molecular weight between entanglements. In practice, the relative modulus of the plateau region shows the relative changes in M_e or the number of crosslinks compared to a standard material.

The rubbery plateau is also related to the degree of crystallinity in a material, although DSC is a better method for characterizing crystallinity than DMA.[31] Also as in the DSC, there is evidence of cold crystallization in the temperature range above the T_g (Fig. 5). This is one of the several transitions that can be seen in the rubbery plateau region. This crystallization occurs when the polymer chains have been quenched (quickly cooled) into a highly disordered state. On heating above the T_g, these chains gain enough mobility to rearrange into crystallites, which causes sometimes a dramatic increase in the modulus. Differential scanning calorimetry or its temperature-modulated variant, StepScan™ DSC, can be used to confirm this.[32] The alpha star transition, $T_\alpha{}^*$, the liquid–liquid transition, T_{ll}, the heat set temperature, and the cold crystallization peak are all transitions that can appear on the rubbery plateau. In some crystalline and semicrystalline polymer, a transition is seen called $T_\alpha{}^*$.[33] The α^* transition is associated with the slippage between crystallites, and helps extend the operating range of a material above the T_g.

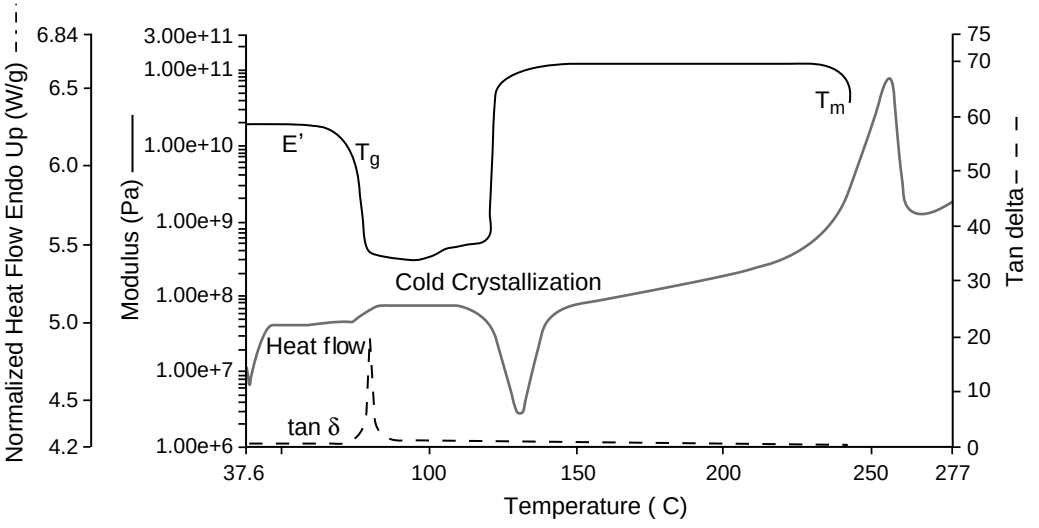


Fig. 5 Cold crystallization in PET caused a large increase in the storage modulus, E', above the T_g. A DSC scan of the same material is included. (From Ref.[1].) *(View this art in color at www.dekker.com.)*

This transition is very susceptible to processing induced changes and can be enlarged on decreased by the applied heat history, processing conditions, and physical aging.[34] Hence, the $T_\alpha{}^*$ has been used by fiber manufacturers to optimize the properties in their materials.

In amorphous polymers, the T_{ll} transition is seen instead, which is a liquid–liquid transition associated with increased chain mobility and segment–segment associations.[35] This order is lost when the T_{ll} is exceeded and regained on cooling from the melt. Boyer[36] reports that, like the T_g, the appearance of the T_{ll} is affected by the heat history. The T_{ll} is also dependent on the number average molecular weight, M_n, but not on the weight average molecular weight, M_w. Bershtien[37] suggests that this may be considered as quasi-melting on heating or the formation of stable associates of segments on cooling. While this transition is reversible, it is not always easy to see, and Boyer[38] spent many years trying to prove that it was real. It is still not totally accepted. Following this transition, a material enters the terminal or melting region.

The terminal region

On continued heating, the melting point, T_m, is reached. The melting point is where the free volume has increased, so the chains can slide past each other and the material flows. This is also called the terminal region. In the molten state, this ability to flow is dependent on the molecular weight of the polymer (Fig. 3). The melt of a polymer material will often show changes in temperature of melting, width of the melting peak, and enthalpy as the material changes[39] resulting from changes in the polymer molecular weight and crystallinity.

Degradation, polymer structure, and environmental effects all influence what changes occur. Polymers that degrade by crosslinking will look very different from those that exhibit chain scission. Very highly crosslinked polymers will not melt as they are unable to flow.

The study of polymer melts and especially their elasticity was one of the areas that drove the development of commercial DMAs. Although a decrease in the melt viscosity is seen with temperature increases, the DMA is most commonly used to measure the frequency dependence of the molten polymer as well as its elasticity. The latter property, especially when expressed as the normal forces, is very important in polymer processing.

Frequency dependencies in transition studies

The choice of a testing frequency or its effect on the resulting data must be addressed. A short discussion of how frequencies are chosen and how they affect the measurement of transitions is in order. Considering that higher frequencies induce more elastic-like behavior, there is some concern that a material will act stiffer than it really is, if the test frequency is chosen to be too high. Frequencies for testing are normally chosen by one of the three methods. The most scientific method would be to use the frequency of the stress or strain that the material is exposed to in the real world. However, this is often outside of the range of the available instrumentation. In some cases, the test method or the industry standard sets a certain frequency and this frequency is used. Ideally, a standard method like this is chosen so that the data collected on various commercial instruments can be shown to be compatible. Some of the ASTM methods for TMA

Table 1 ASTM tests for the DMA

D4065	Determining DMA properties terminology[a]
D4092	Terminology for DMA tests
D4440	Measurement of polymer melts
D4473	Cure of thermosetting resins
D5023	DMA in three-point bending tests
D5024	DMA in compression
D5026	DMA in tension
D5279	DMA of plastics in tension
D5418	DMA in dual cantilever
D5934	DMA tensile modulus under constant loading
E473–94	Terminology for thermal analysis
E1640–94	T_g by DMA
E1824–96	T_g by TMA in tension
E1867–97	Temperature calibration for DMA
E2254–03	Storage modulus calibration for DMA
E****-**	Loss modulus calibration for DMA (in balloting)

[a]This standard qualifies a DMA as acceptable for all ASTM DMA standards.

and DMA are listed in Table 1. Many industries have their own standards, so it is important to know whether the data is expected to match a Mil-spec, an ASTM standard, or a specific industrial test. Finally, one can arbitrarily pick a frequency. This is done more often than not, so that 1 Hz and 10 rad/sec are often used. As long as the data are run under the proper conditions, they can be compared to highlight material differences. This requires that frequency, stresses, and the thermal program be the same for all samples in the data set.

Lowering the frequency shifts the temperature of a transition to a lower temperature (Fig. 6). At one time, it was suggested that multiple frequencies could be used, and the T_g should then be determined by extrapolation to 0 Hz. This was never really accepted as it represented a fairly large increase in testing time for a small improvement in accuracy. For most polymer systems, for very precise measurements, one uses a DSC. Different types of transitions also have different frequency dependencies. If one looks at the slope of the temperature dependence of transitions against frequency, one sees that in many cases that the primary transitions like T_m and T_g have a different dependence on frequency than the lower temperature transitions. In fact, the activation energies are different for α, β, and γ transitions because of the different motions required and the transitions can be sorted by this approach.[1]

Polymer Melts and Solutions

A fluid or polymer melt responds to strain rate rather than to the amount of stress applied. The viscosity is one of the main reasons why people run frequency scans. As the stress–strain curves and the creep recovery runs show, viscoelastic materials exhibit some degree of flow or unrecoverable deformation. The effect is strongest in melts and liquids where frequency vs. viscosity plots are the major application of DMA. Fig. 7 shows a frequency scan on a viscoelastic material. In this example, the sample is a rubber above the T_g in three-point bending, but the trends and principles apply to both solids and melts. The storage modulus and complex viscosity are plotted on log scales against the log of frequency. In analyzing the frequency scans, trends in the data are more significant than specific peaks or transitions.

The zero shear plateau

On the viscosity curve, η^*, a fairly flat region appears at low frequency, called the zero shear plateau.[40] This is where the polymer exhibits Newtonian behavior and its viscosity is dependent on molecular weight, not the strain rate. The viscosity of this plateau has been shown to experimentally related to the molecular weight for Newtonian fluid:

$$\eta \propto cM_\nu^1 \tag{2}$$

for cases where the molecular weight, M_v, is less than the entanglement molecular weight, M_e and for cases where M_v is greater than M_e:

$$\eta \propto cM_\nu^{3.4} \tag{3}$$

where η_o is the viscosity of the initial Newtonian plateau, c is a material constant, and M_v is the viscosity

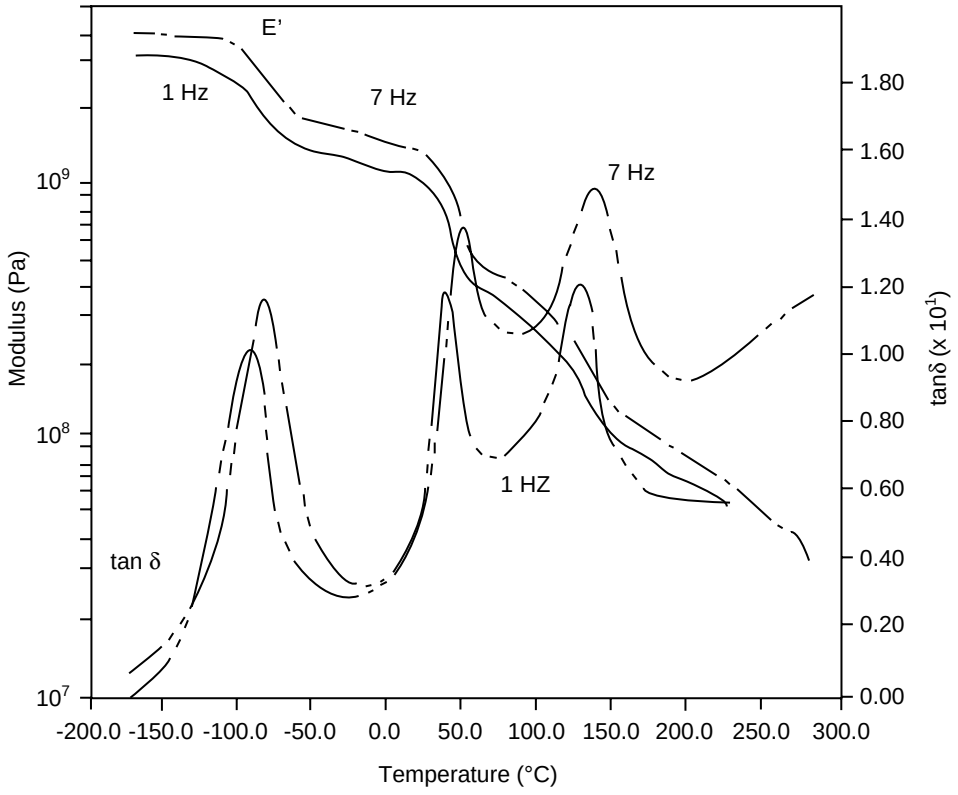


Fig. 6 Effect of frequency on transitions: (A) the dependence of the T_g in PMMA on frequency.

average molecular weight. This relationship can be written in general terms, replacing the exponential term with the Mark-Houwink constant, *a*. Eq. (3) can be used as a method of approximating the molecular weight of a polymer. The value obtained is closest to the viscosity average molecular weight obtained by osmometry.[41] In comparison with the weight average data obtained by Gel Permeation Chromatography, the viscosity average molecular weight would be between the number average and weight average molecular weights, but closer to the latter.[42] This was originally developed for steady shear viscosity but applies to complex viscosity as well.

The relationship between steady shear and complex viscosity is fairly well established. Cox and Merz[43] found that an empirical relationship exists between complex viscosity and steady shear viscosity when the shear rates are the same. The Cox–Merz rule is stated as follows:

$$|\eta(\omega)| = \eta(\dot{\gamma})|_{\dot{\gamma}=\omega} \tag{4}$$

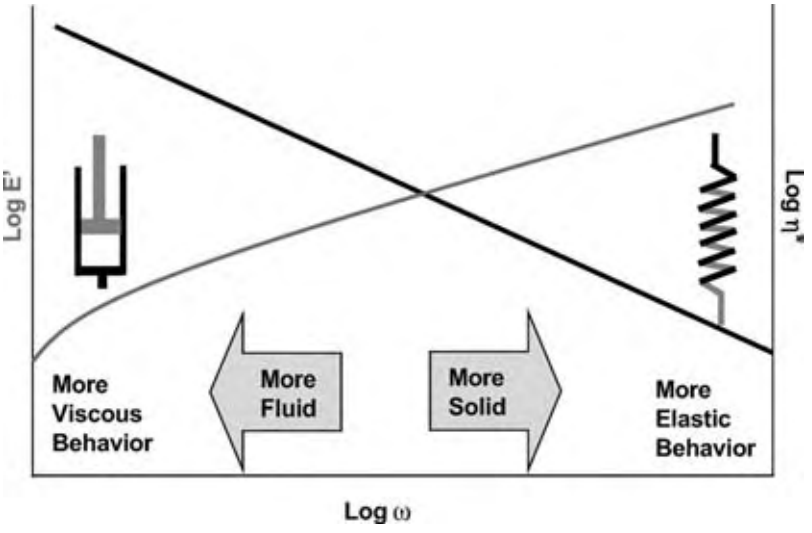


Fig. 7 An example of a frequency scan showing the change in a materials behavior as frequency varies. Low frequencies allow the material time to relax and respond, hence, flow dominates. High frequencies do not and elastic behavior dominates. (From Ref.[1].) *(View this art in color at www.dekker.com.)*

where η is the constant shear viscosity, η^* is the complex viscosity, ω the frequency of the dynamic test, and $d\gamma/dt$ the shear rate of the constant shear test. This rule of thumb seems to hold for most materials to within approximately $\pm 10\%$. Another approach is the Gleissele[44] Mirror Relationship that states the following:

$$\eta\dot{\gamma} = \eta^+(t)|_{t=1/\dot{\gamma}} \quad (5)$$

when $\eta^+(t)$ is the limiting value of the viscosity as the shear rate, , approaches zero.

The low frequency range is where viscous or liquid–like behavior predominates. If a material is stressed over long enough times, some flow occurs. As time is the inverse of frequency, this means materials are expected to flow more at low frequency. As the frequency increases, the material will act in a more and more elastic fashion. Silly Putty™, the children's toy, shows this clearly. At low frequency, Silly Putty™ flows like a liquid while at high frequency, it bounces like a rubber ball. This behavior is also similar to what happens with temperature changes.

A polymer becomes softer and more fluid as it is heated, and it goes through transitions that increase the available space for molecular motions. Over long enough time periods, or small enough frequencies, similar changes occur. As this implies, one can move a polymer across a transition by changing the frequency. This relationship is also expressed as the idea of time–temperature equivalence.[45] Often stated as low temperature is equivalent to short times or high frequency, it is a fundamental rule of thumb in understanding polymer behavior.

The power law region

As the frequency is increased in a frequency scan, the Newtonian region is exceeded and a new relationship develops between the rate of strain, or the frequency, and the viscosity of the material. This region is often called the power law region and can be modeled by:

$$\eta^* \cong \eta(\dot{\gamma}) = c\dot{\gamma}^{n-1} \quad (6)$$

where η^* is the complex viscosity, $\dot{\gamma}$ is the shear rate, and the exponent term n is determined by the fit of the data. This can also be written as:

$$\sigma \cong n(\dot{\gamma}) = c\dot{\gamma}^n \quad (7)$$

where σ is the stress and η is the viscosity. The exponential relationship is why the viscosity vs. frequency plot is traditionally plotted on a log scale. With modern curve fitting programs, the use of log–log plots has declined and is a bit anachronistic. The power law region of polymers shows shear thickening or thinning behavior. This is also the region in which the E'–η^* or the E'–E'' crossover point is found. As frequency increases and shear thinning occurs, the viscosity (η^*) decreases. At the same time, increasing the frequency increases the elasticity (E'). This is shown in Fig. 7. The E'–η^* crossover point is used an indicator of the molecular weight and molecular weight distribution.[46] Changes in its position are used as a quick method of detecting changes in the molecular weight and distribution of a material. After the power law region, another plateau is seen, the infinite shear plateau.

Infinite shear plateau

This second Newtonian region corresponds to where the shear rate is so high that the polymer no longer shows a response to increases in the shear rate. At the very high shear rates associated with this region, the polymer chains are no longer entangled. This region is seldom seen in DMA experiments and usually avoided in use because of the damage done to the chains. It can be reached in commercial extruders and causes degradation of the polymer, which causes the poorer properties associated with regrind.

As the curve in Fig. 7 shows, the modulus also varies as a function of the frequency. A material exhibits more elastic-like behavior as the testing frequency increases, and the storage modulus tends to slope upward towards higher frequency. The storage modulus' change with frequency depends on the transitions involved. Above the T_g, the storage modulus tends to be fairly flat with a slight increase with increasing frequency as it is on the rubbery plateau. The change in the region of a transition is greater. If one can generate a modulus scan over a wide enough frequency range, the plot of storage modulus vs. frequency appears like the reverse of a temperature scan. The same time–temperature equivalence discussed above also applies to modulus, as well as compliance, tan delta, and other properties.

The frequency scan is used for several purposes that is discussed here. One very important use, that is very straightforward, is to survey the material's response over various shear rates. Different properties are required at these regimes, and to optimize one property may require chemical changes that harm the other. Similarly, changes in polymer structure can show these kinds of differences in the frequency scan. Branching affects the response to different frequencies.[6]

For example, in a tape adhesive, sufficient flow under pressure at low frequency is desired to fill the pores of the material to obtain a good mechanical bond. When the laminate is later subjected to peel,

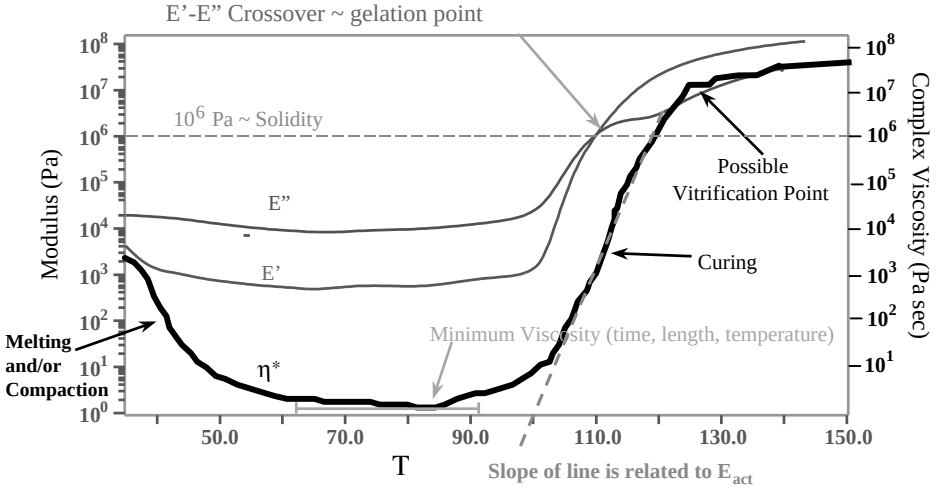


Fig. 8 The DMA cure profile of a two-part epoxy showing the typical analysis for minimum viscosity, gel time, vitrification time, and estimation of the action energy. (From Ref. [1].) *(View this art in color at www.dekker.com.)*

the material needs to be very elastic so it will not pull out of the pores.[47],[a] The frequency scan allows measurement of these properties in one scan, so ensuring that tuning one property does not degrade another. Viscosity vs. frequency plots are used extensively to study how changes in polymer structure or formulations affect the behavior of the melt. Often changes in materials, especially in uncured thermosetting resins and molten materials, affect a limited frequency range and testing at a specific frequency can miss the problem.

Thermosets

The DMA's ability to give viscosity and modulus values for each point in a temperature scan allow estimation of kinetic behavior as a function of viscosity. This has the advantage of describing how viscous the material is at any given time, so as to determine the best time to apply pressure, what design of tooling to use, and when the material can be removed from the mold. Recent reviews have summarized this approach for epoxy systems.[48]

Curing

The simplest way to analyze a resin system is to run a plain temperature ramp from ambient to some elevated temperature.[49] This ''cure profile'' allows collection of several vital pieces of information as shown in Fig. 8. Samples may be run ''neat'' or impregnated into fabrics in techniques that are referred to as ''torsional braid.'' Another problem area is paints and coatings,[50] where the material is used in a thin layer. This can be addressed experimentally by either a braid as above or by coating the material on a thin sheet of metal. The metal is often run first, and its scan subtracted from the coated sheet's scan to leave only the scan of the coating. This is also done with thin films and adhesive coatings.

From the cure profile seen in Fig. 8, it is possible to determine the minimum viscosity (η^*_{min}), the time to η^*_{min} and the length of time it stays there, the onset of cure, the point of gelation where the material changes from a viscous liquid to a viscoelastic solid, and the beginning of vitrification. The minimum viscosity is seen in the complex viscosity curve and is where the resin viscosity is the lowest. A given resin's minimum viscosity is determined by the resin's chemistry, the previous heat history of the resin, the rate at which the temperature is increased, and the amount for stress or strain applied. Increasing the rate of the temperature ramp is known to decrease the η^*_{min}, the time to η^*_{min}, and the gel time. The resin gets softer faster, but also cures faster. The degree of flow limits the type of mold design and when as well as how much pressure can be applied to the sample. The time spent at the minimum viscosity plateau is the result of a competitive relationship between the material's softening or melting as it heats, and its rate of curing. At some point, the material begins curing faster than it softens, and that is where the viscosity starts to increase.

As the viscosity begins to climb, an inversion is seen of the E″ and E′ values as the material becomes more solidlike. This crossover point also corresponds to where the tan δ equals 1 (since $E' = E''$ at the crossover). This is taken to be the gel point,[51] where the crosslinks have progressed to forming an ''infinitely'' long network across the specimen. At this point, the

[a]This is a very simplified version of adhesion. The reader is referred to Ref.[47] for a detailed discussion.

sample will no longer dissolve in solvent. While the gel point correlates fairly often with this crossover, it does not always. For example, for low initiator levels in chain addition thermosets, the gel point precedes the modulus crossover.[52] Temperature dependence for the presence of the crossover has also been reported.[49] In some cases, where powder compact and melts before curing, there may be several crossovers.[53] Then, the one following the $\eta^*_{\min}$ is usually the one of interest. Some researchers[54] believe that the true gel point is best detected by measuring the frequency dependence of the crossover point. This is done by either by multiple runs at different frequencies or by multiplexing frequencies during the cure. At the gel point, the frequency dependence disappears.[55] This value is usually only a few degrees different from the one obtained in a normal scan, and in most cases not worth the additional time. During this rapid climb of viscosity in the cure, the slope for η^* increase can be used to calculate an estimated E_a (activation energy).[56] This will be discussed below, but the fact that the slope of the curve here is a function of E_a is important. Above the gel temperature, some workers estimate the molecular weight, M_c, between crosslinks as:

$$G' = RT\rho/M_c \tag{8}$$

where R is the gas constant, T is the temperature in Kelvin, and ρ is the density. At some point, the curve begins to level off, and this is often taken as the vitrification point, T_{vf}.

The vitrification point is where the cure rate slows, because the material has become so viscous that the bulk reaction has stopped. At this point, the rate of cure slows significantly. The apparent T_{vf} however, is not always real: any analyzer in the world has an upper force limit. When that force limit is reached, the ''topping out'' of the analyzer can pass as the T_{vf}. Use of a combined technique like DMA–DEA[57],[b] to see the higher viscosities, or the removal of a sample from parallel plate and sectioning it into a flexure beam, is often necessary to see the true vitrification point. Vitirfication may also be seen in the DSC, if a modulated temperature technique like StepScan is used.[57],[b] A reaction can also completely cure without vitrifying and will level off the same way. One should be aware that reaching vitrification or complete cure too quickly could be as bad as too slowly. Often a overly aggressive cure cycle will cause a weaker material as it does not allow for as much network development, but gives a series of hard (highly crosslinked) areas among softer (lightly crosslinked) areas. On the way to vitrification, an important value is 10^6 Pa sec. This is the viscosity of bitumen,[58] and is often used as a rule of thumb for where a material is stiff enough to support its own weight. This is a rather arbitrary point, but is chosen to allow the removal of materials from a mold and the cure is then continued as a post-cure step. The cure profile is both a good predictor of performance as well as a sensitive probe of processing conditions. As discussed above under TMA applications, a volume change occurs during the cure.[59] This shrinkage of the resin is important and can be studied by monitoring the probe position of some DMA's as well as by TMA and dilatometry.

[b]DEA is dielectric analysis, where an oscillating electrical signal is applied to a sample. From this signal, the ion mobility can be calculated, which is then converted to a viscosity. See Mc Crum[3,26] for details. DEA will measure to significantly higher viscosities than DMA.

CONCLUSIONS

Dynamic mechanical analysis is an incredibly powerful technique for characterizing materials and predicting their in-service behavior. The technique is becoming better known and more commonly used for testing other than that just measuring very weak glass transitions, although it excels in that. As DMA becomes more common, it is being used to map modulus, viscosity, and damping as a function of temperature, time under adverse conditions (such as UV light, aggressive gases, or in solvent), or shear rate (frequency). This data can be obtained more quickly and cost effectively by DMA than by alternative methods. Current trends indicate increased interest in online methods that give similar results as well as extending the range of forces and frequencies to study nonlinear behavior.

NOMENCLATURE

δ	Phase angle
$\tan\delta$	Tangent of the phase angle, also called the damping
σ	Stress
γ	Shear strain
ε	Tensile strains
$\dot{\gamma}$	Shear strain rate
$\dot{\varepsilon}$	Strain rate
η	Viscosity
η^*	Complex viscosity
η'	Storage viscosity
η''	Loss viscosity
E^*	Complex modulus
E'	Storage modulus
E''	Loss modulus
J	Compliance
k	Deformation
T	Period

ρ	Density
G	Shear modulus
M_e	Entanglement molecular weight
M_c	Molecular weight between crosslinks
M_w	Molecular weight
f	Frequency
ω	Frequency in Hz
K	Rate constant
E_a	Activation energy
DMA	Dynamical mechanical analysis or analyzer
DMTA	Dynamic mechanical thermal analysis or analyzer
DEA	Dielectric analysis or analyzer
DSC	Differential scanning calorimeter
TBA	Torsional braid analyzer
TGA	Thermogravimetric analyzer
ν^f	Free volume
$T_{\alpha,\beta,\gamma}$	Transition (subscript type)
Λ	Logarithmic decrement
Γ	Torque

REFERENCES

1. (a) Menard, K. Dynamic Mechanical Analysis: A Practical Introduction. CRC Press: Boca Raton, 1999; (b) Sepe, M. *Dynamic Mechanical Analysis for Plastic Engineering*; Plastic Design Library: New York, 1998; (c) Murayama, T. *Dynamic Mechanical Analysis of Polymeric Materials*; Elsevier: New York, 1977.
2. Flory, P. *Principles of Polymer Chemistry*; Cornell University Press: Ithaca, 1953.
3. Mc Crum, N.; Read, B.; Williams, M. *Dielectric and Anaelastic Effects in Polymeric Solids*; Dover: New York, NY, 1991; 191–200.
4. Doi, M.; Edwards, S. *The Dynamics of Polymer Chains*. Oxford University Press: New York, 1986.
5. Knight, A. *Introduction to Viscoelasticity*; Wiley: New York, 1983.
6. Rohn, C.L. *Analytical Polymer Rheology*; Hanser-Gardener: New York, 1995.
7. Heijboer, J. Secondary loss peaks in glassy amorphous polymers. Int. J. Polym. Mater. **1977**, *6*, 11.
8. Boyer, R.F. Dependence of mechanical properties on molecular motion in polymers. Polym. Eng. Sci. **1968**, *8* (3), 161.
9. Rohn, C.L. *Analytical Polymer Rheology*; Hanser-Gardener: New York, 1995, 279–283.
10. (a) Heijboer, J. Secondary loss peaks in glassy amorphous polymers. Int. J. Polym. Mater. **1977**, *6*, 11; (b) Mangion, M.; Johari, G. Relaxations in thermosets. VI. Effects of crosslinking on sub-T_g relaxations during the curing and aging of epoxide-based thermosets. J. Polym. Sci: Part B Polym. Phys. **1991**, *29*, 437; (c) Johari, G.; Mikoljaczak, G.; Cavaille, J. Dynamic mechanical behavior and its dependence on preparation method in structural epoxide resins. Polymer **1987**, *28*, 2023; (d) Cheng, S.; Arnold, F.; Bruno, K.; Shen, D.; Eashoo, E.; Lee, C.J.; Harris, F. The origin of Beta relaxations in segmented rigid-rod polyimide and copolyimide polymers. Polym. Sci. Eng. **1993**, *33*, 1373; (e) Johari, G. Mechanical relaxations in polymer glasses. Lecture Notes in Physics **1987**, *277*, 90; (f) Daiz-Calleja, R.; Riande, E. Molecular motions in macromolecules under mechanical and dielectric force fields. Rheol. Acta **1995**, *34*, 58; (g) Boyd, R. Relaxation processes in crystalline polymers: experimental behavior—a review. Polymer **1985**, *26*, 323; (h) Bershtien, V.; Egorov, V.; Egorova, L.; Ryzhov, V. The role of thermal analysis in revealing the common molecular nature of transitions in polymers. Thermochim. Acta **1994**, *238*, 41.
11. Twombly, B. Determination of weak glass transitions in semi-crystalline polymers. Proc NATAS **1991**, *20*, 63.
12. (a) Rohn, C.L. *Analytical Polymer Rheology*; Hanser-Gardener: New York, 1995; (b) Heijboer, Secondary loss peaks in glassy amorphous polymers. J. Int. J. Polym. Mater. **1977**, *6*, 11.
13. Boyer, R. Determination of the packing density of solid polymer chains. Determination of the specific surface area of polymers by sorption of nitrogen vapor. Polym. Eng. Sci. **1968**, *8* (3), 161.
14. Bershtien, V.; Egorov, V. *Differential Scanning Calorimetery in the Physical Chemistry of Polymers*; Ellis Horwood: Chichester, 1993.
15. Coxton, B. Private communication.
16. Cheng, S.; Arnold, F.; Bruno, K.; Shen, D.; Eashoo, E.; Lee, C.J.; Harris, F. The origin of beta relaxations in segmented rigid-rod polyimide and copolyimide polymers. Polym. Sci. Eng. **1993**, *33*, 1373.
17. Johari, G.; Mikoljaczak, G.; Cavaille, J. Dynamic Mechanical behavior and its dependence on preparation method in structural epoxide resins. Polymer **1987**, *28*, 2023.
18. Mangion, M.; Johari, G. Relaxations in thermosets. VI. Effects of crosslinking on sub-T_g relaxations during the curing and aging of epoxide-based thermosets. J. Polym. Sci: Part B Polym. Phys. **1991**, *29*, 437.
19. (a) Nelson, F.C. A method for measuring the damping properties of acoustic materials. Shock and Vibration Digest **1994**, *26* (2), 11; (b) Nelson, F.C. Vibrational and acoustical damping

measurement in foams. Shock and Vibration Digest **1994**, *26* (2), 24.
20. Brostow, W. Private communication.
21. Boyer, R. Dependence of mechanical properties on molecular motion in polymers. Polym. Eng. Sci. **1968**, *8* (3), 161.
22. (a) Johari, G. Mechanical relaxations in polymer glasses. Lecture Notes in Physics **1987**, *277*, 90; (b) Heijboer, J. Secondary loss peaks in glassy amphorous polymers. Int. J. Polym. Mater. **1977**, *6*, 11; (c) Heijboer, J.; Bussink, J. Mechanical properties and molecular structure of organic polymers. In *Physics of Non-crystalline Solids*; Prins, J., Ed.; Interscience: New York, 1965; (d) Heijboer, J. Dynamic mechanical properties and impact strength. J. Polym. Sci. **1968**, *C16*, 3755; (e) Nielsen, L.; Morgan, T. Effects of solvent on dynamic mechanical properties of polystyrenes. J. Macromol. Sci. Phys. **1974**, *9*, 239.
23. Yee, A.; Smith, S. Molecular structure effects on the dynamic mechanical spectra of polycarbonates. Macromolecules **1981**, *14*, 54.
24. Gordon, G. Glass transitions in nylons. J. Polym. Sci. Part A-2 Polym. Phys. **1971**, *9*, 1693.
25. Wendorff, J.; Schartel, B. Dielectric investigations on secondary relaxations of polyarylates. Polymer **1995**, *36* (5), 899.
26. Mc Crum, N.; Read, B.; Williams, M. *Dielectric and Anaelastic Effects in Polymeric Solids*; Dover: New York, NY, 1991; 180–185.
27. Boyd, R.H. Relaxation processes in crystalline polymers: experimental behaviour—a review. Polymer **1985**, *26*, 323.
28. (a) Noda, I.; Smith, S.; Dowrey, A.; Gothaus, J.; Marcott, C. Dynamic IR studies on microdomain interphases of isotope-labeled block copolymers. Appl. Spectrosc. **1990**, *44* (4), 550; (b) Kien, V. Dynamic IR studies on block copolymers. Proceedings of the Sixth Symposium. Radiat. Chem. **1987**, *6* (2), 463.
29. Sperling, L.H. *Introduction to Physical Polymer Science,* 2nd Ed; Wiley: New York, 1992.
30. Macosko, C. *Rheology*; VCH: New York, 1994.
31. (a) Quinn, F.; Hatakeyama, T. *Thermal Analysis*; Wiley: New York, 1994; (b) Wunderlich, B. *Thermal Analysis*; Academic Press: New York, 1990.
32. (a) Schawe, J. Principles for the interpretation of modulated temperature DSC measurements. Part 1. Glass transition. Thermochim. Acta **1995**, *261*, 183; (b) Schawe, J. A comparison of different evaluation methods in modulated temperature DSC. Thermochim. Acta **1995**, *260*, 1; (c) Schawe, J. Simultaneous x-ray diffraction and differential thermal analysis of polymers. Thermochim. Acta **1995**, *271*, 1; (d) Wunderlich, B.; Boller, A.; Jin, Y. Heat capacity measurement by modulated DSC at constant temperature. J. Therm. Anal. **1994**, *42*, 949.
33. (a) Boyd, R.H. Relaxation processes in crystalline polymers: experimental behavior. Polymer **1985**, *26*, 323; (b) Boyd, R.H. Relaxation processes in crystalline polymers: molecular interpretation—a review. Polymer **1985**, *26*, 1123.
34. (a) Godber, S. Private communication; (b) Ahmed, M. *Polypropylene Fiber—Science and Technology*; Elsevier: New York, 1982.
35. Lobanov, A.; Frenkel, S. The nature of the so-called liquid-liquid transition in polymer melts. Polym. Sci. USSR **1980**, *22*, 1150.
36. (a) Boyer, R. Quasiharmonic treatment of infrared and raman vibrational frequency shifts induced by tensile deformation of polymer chains. II. Application to the polyoxymethylene and isotactic polypropylene single chains and the three-dimensional orthorhombic polyethylene crystal. J. Polym. Sci. Part B: Polym. Phys. **1992**, *30*, 1177; (b) Stadnicki, S.J.; Gilham, J.K.; Boyer, R.F. The T_{II} ($>T_g$) transition of atactic polystyrene. J. Appl. Polym. Sci. **1976**, *20*, 1245; (c) Enns, J.B.; Boyer, R. Differential Scanning Calorimetry (DSC) observations on the T_{II} transitions in amorphous polymers. Encyclopedia Polym. Sci. **1989**, *17*, 23–47.
37. Bershtien, V.; Egorov, V.; Egorova, L.; Ryzhov, V. The role of thermal analysis in revealing the common molecular nature of transitions in polymers. Thermochim. Acta. **1994**, *238*, 41.
38. Warner, C.M. Evaluation of the DSC for observation of the liquid–liquid transition. Master Thesis, Central Michigan State University. **1988**.
39. (a) Turi, E., Ed. *Thermal Characterization of Polymeric Materials*; Academic Press: Boston, 1981; (b) Turi, E., Ed. *Thermal Analysis in Polymer Characterization*; Heydon: London, 1981.
40. Rohn, C. *Analytical Polymer Rheology*; Hanser: New York, 1995.
41. Miller, M. *The Structure of Polymers*; Reinhold: New York, 1966; 611–612.
42. Rosen, S. *Fundamantal Principles of Polymeric Materials*; Wiley Interscience: New York, 1993; 53–77 and 258–259.
43. (a) Cox, W.; Merz, E. Correlation of dynamic and steady state viscosity. J. Polym. Sci. **1958**, *28*, 619; (b) Leblans, P.; Booij, H.; Palmen, J.; Taemersma-Thoona, G. Non-linear viscosity and Cox-Merz relationships for polymeric fluids. J. Polym. Sci. **1983**, *21*, 1703.
44. Gleissele, W. *Rheology*; Astarita, G., Ed.; Plenum Press: New York, 1980; Vol. 2, 457 pp.
45. Van Krevelin, D.W. *Properties of Polymers*; Elsevier: New York, 1987; 289 pp.
46. Macosko, C. *Rheology*; VCH Publishers: New York, 1996; 120–127.

47. (a) Lee, L.-H., Ed. *Adhesive Bonding*; Plenum Press: New York, 1991; (b) Lee, L.-H., Ed. *Fundamentals of Adhesion*; Plenum Press: New York, 1991.
48. Brostow, W.; Bilyeu, B.; Menard, K. J. Mater. Educ. **1999**, *21*, 281.
49. (a) Martin, G.; Tungare, A.; Gotro, J. Modeling rheological and dielectric properties during thermoset cure. In *Polymer Characterization*; ACS: Washington, DC, 1990; (b) Ryan, M.; Matsuoka, T. Theory and practice on nonlinear behavior of polymer solids. ANTEC Proc. **1973**, *31*, 187; (c) Gramelt, C. A new approach for determining mechanical properties of thermosetting resins during the cure. Am. Lab. **1984**, *26* (January); (d) Etoh, S.; Yamamato, Y. Preparation and characterization of thermoplastic/thermosetting blends. SAMPE J. **1985**, *3*, 6; (e) Hurwitz, F. Dynamic mechanical characterization of cure of a polyimide-graphite fiber composite. Polym. Compos. **1983**, *4* (2), 89.
50. (a) Roller, M.B. Thermoset and coatings technology: the challenge of interdisciplinary chemistry. Polym. Eng. Sci. **1979**, *19*, 692; (b) Roller, M. et al. Thermoset and coating technology: the challenge of an interdisciplinary science. J. Coat. Technol. **1978**, *50*, 57.
51. (a) Heise, M.; Martin, G.; Gotro, J. Gelation in thermosets formed by chain addition polymerization. Polym. Eng. Sci. **1990**, *30*, 83; (b) O'Driscoll, K.; Dionisco, J.; Mahabadi, H.; Abuin, E.; Lissi, E. High conversation polymers 4: Definition of the onset of the gel effect. J. Polym. Sci.: Polym. Chem. **1979**, *17*, 1891; (c) Okay, O. Kinetics of gelation in free radical crosslinking copolymerization. Polymer **1994**, *35*, 2613.
52. Hiese, M.; Martin, G.; Gotro, J. Gelation in thermosets formed by chain addition polymerization. Polym. Eng. Sci. **1990**, *30* (2), 83.
53. Wissbrun, K. etal. Model for correlation of rheological measurement with leveling and curing of thermosetting powder coating. J. Coat. Technol. **1976**, *48*, 42.
54. (a) Champon, F.; Winter, G. Linear viscoelasticity at the gel point of a crosslinking PDMS with imbalanced stoichiometry. J. Rheol. **1987**, *31*, 683; (b) Winter, H. Can the gel point of a cross-linking polymer be detected by the G'–G' crossover? Polym. Eng. Sci. **1987**, *27*, 1698; (c) Michon, C.; Cavelier, F.; Laurnay, B. Concentration and dependence of critical viscoelastic properties of gelatin at the gel point. Rheol. Acta **1993**, *32*, 94.
55. Michon, C.; Cavelier, F.; Laurnay, B. Concentration and dependence of critical viscoelastic properties of gelatin at the gel point. Rheol. Acta **1993**, *32*, 94.
56. Kalnin, I.; Hollands, K. Kinetics of gelation of some accelerated acid anhydride cured epoxy resins. In *Epoxy Resins*; ACS: Washington, DC, 1970.
57. Bilyeu, B.; Brostow, W.; Menard, K. *Materials Characterization by Dynamic and Modulated Thermal Analytical Techniques, ASTM STP 1402*; Riga, A.T., Judovits, L.H., Eds.; American Society for Testing and Materials: West Conshohocken, PA, 2001.
58. Barnes, H. *An Introduction to Rheology*; Elsevier: New York, 1989.
59. Snow, A.W.; Armistead, J. A simple dilatometer for thermoset cure shrinkage and thermal expansion measurements. J. Appl. Polym. Sci. **1994**, *52*, 401.

Education on Plant Design

Truman S. Storvick
Chemical Engineering Department, University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

The goal of any practicing chemical engineer is to visualize processes and assemblies of equipment that will produce products beneficial to mankind. These products result when the chemical compositions of the starting materials are transformed by chemical reactions to form the prized new compounds. The processes include the preparation of the materials that go to the reaction system and the isolation of the products that flow from it. It is this intentional change in the chemical composition of matter that distinguishes chemical engineering technology from the other engineering disciplines. What follows is a short presentation of the elements of plant design. Some of the tools available to the process engineer are identified. A discussion of topics that might work as an assignment for university students follows. The evolution of plant and process design over the professional lifetime of the practicing engineer will continue and require continuing education programs to renew skills. It is interesting to speculate what lies ahead.

DESCRIPTION OF PLANT DESIGN

The words ''plant design'' immediately bring to mind industrial applications. This makes it almost certain that it is an economic analysis that will provide the answer to the question ''Will this venture make a profit?'' It also implies that all of the science and technology used to select and assemble equipment to make the plant function must combine with the economic analysis to form a successful plant design. The economic analysis will include the cost of compliance with federal and state legislation and local ordinances, the policies written to ensure safety and compliance with local environmental regulations. Plant design brings into being the physical facility that must be located in contact with the civil and economic communities.

Process design is a team activity. Each person on the team brings everything they know to the design table. Persons who have special training and experience along with basic knowledge of science and technology related to the project are selected for assignment to the design team. Large firms may have those people on the payroll. Smaller firms and often large firms working on a design that requires specialized knowledge or skill will retain consultants to do part or all of the design. Process design success depends on the design team.

Plant design is both the object and the act of visualizing a chemical process or plant. Visualization is the mental activity that allows rapid scanning of many options from which an optimum will be selected. Plant design statements usually underspecify the design task. It is the plant designers who must ask the questions that must be answered to reach an optimum plant design. Douglas describes the process of conceptual design, a hierarchy of design decisions that quickly identifies the sequence of process steps that can lead to an optimum plant configuration.[1] These ''rules of thumb or heuristics'' are based on design experience. The lessons from the history of common process steps are used on a very large, complex design problem to decompose it into a series of small problems that are much easier to visualize and solve. The conceptual design method sorts through many processing sequence alternatives to select the one, two, or very few that show the most promise. Detailed design must be carefully done so it is important to quickly focus on the good alternative. The secret to successful process design is to spend time on the ''best'' option.

Plant design is an active part of the total lifetime of all chemical processes. The birth to death sequence of a chemical plant is characterized by the level of detail of the plant description required at each design step. Early death of a proposed process can occur at any step before it has been built and produces a product. Only a small fraction of the initial proposals survive beyond the second step. Design continues as the process is modified to keep the product profitable in the face of all competition. The stages of a process proposal can be described as follows:

1. Company management considers a move to manufacture a new product, take advantage of a new raw material, respond to a customer request for a product, reduce toxic emissions from the plant, buy or sell corporate technology from or to another firm.
2. Initial analysis leads to a recommendation that a design team be assembled to explore process options. This is where data are collected and the preliminary process steps are evaluated.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008036

There is a preliminary cost estimate that becomes the basis for the decision to move to the next step.

3. The detailed specifications for the process streams and the equipment are set. All of the drawings and documents required by construction contractors and equipment vendors are prepared. Another cost estimate determines whether to proceed with a request to the business and legal departments to call for bids for plant construction. Construction cost estimates present another decision to proceed with the project.
4. A new group of engineers who work for the construction contractor join to build the plant. The firm maintains supervisory control over construction and negotiates any construction change orders required to complete the project. There will be people writing operating manuals and setting the operator training schedules. This is the time when the equipment and construction fees are paid. The capital investment in the production facility is now committed.
5. The start-up and acceptance of the plant by construction contractors involves a new set of process engineers. Equipment must perform to specifications. There will be an agreement concerning product production rate and quality that must be met to complete the construction phase.
6. Plant operation and maintenance proceed throughout the life of the plant. There will be design of new equipment or procedures to improve the product quality, increase production, reduce energy requirements, etc. Process design never ends as long as the plant operates.
7. There comes a time when the plant is shut down and the equipment salvaged or scrapped. The engineering steps for equipment decommissioning can be as challenging as new construction. Important process design information is obtained by examining the equipment as it is moved out of service.

Many plant design assignments involve process changes that are made to improve an existing facility. A customer wants to change the purity or alter the composition of a product. Treatment of a by-product stream can eliminate a pollution problem and yield a saleable product. New separation technology improves the quality of a product and reduces energy costs. The sales force and the engineers who operate the production facilities play key roles in identifying these tasks assigned to the process engineer. Design changes in an operating plant advance through essentially the same steps as the design of a new plant.

New chemical products are usually produced on an existing plant site. The utilities, management support, maintenance staff, and operating personnel are there and can gradually be assigned to the new product facility as it gains commercial acceptance. Specialty chemicals and pharmaceutical plants are examples of manufacturing facilities that change rather quickly as demand for new products often, rises and falls over a few months rather than years.

There are examples of ''grassroots'' plants usually designed to produce an established product closer to a raw material source or to a product market Examples are fertilizer plants, fuel alcohol, and food processing plants. New petroleum and natural gas processing plants recently built in the Near East are set close to the huge crude oil reserves located there to provide fuel for local markets and to ship value-added products.

The time for completion of each design step is an important variable in the plant design sequence. Design problems by nature are underspecified, but the schedule to complete each step is fixed. More questions are answered at each step in the design sequence, but at every step there will be many unanswered questions. When the plant is well into production, there will be unanswered design questions. Each step in the lifetime of a design provides new insight and empirical data that can be used in the next design step or carried over as process information for the design of new production facilities.

COMPUTERS IN PROCESS DESIGN

Digital computers have been available to the educational community since the 1950s. Design of multicomponent distillation columns was one of the first industrial applications that were programmed. The tedious trial and error calculations for each equilibrium stage performed on desk calculators were replaced by this fast computing machine. Remarkably, a general solution to complex separations is not available and work to improve multicomponent distillation algorithms continues.

The computational speed and memory size of the digital computers increased in the 1960s. The computing power was sufficient to begin programming the mathematical representations of chemical processes. It was also possible to use the computers to record many (100 or 200) process variables (temperature and pressure) in real time while the plant was running. Samples were drawn from the process streams for chemical analysis and these were later matched to the timeline of the temperature and pressure measurements. These data gave a picture of the plant response to changes in the operating conditions. Data logging provided a way to select the settings for each unit in

the plant so that the overall plant performance was improved, justifying the cost of the computers to make the measurements.

Computers continued to get smaller, faster, and cheaper. The accuracy and speed of chemical analysis continued to improve. With real-time chemical analysis of process streams and a mathematical model for each unit in the process, the computer was programmed to set controllers on process streams using the new process control algorithms. With some additional programming and with the process control set points adjusted by the computer control system, the quality of the product improved. The physical property data for the chemicals in the process streams required for these model calculations were built into the process control program.

The past 20 yr has seen remarkable changes in the computational tools available to the process design engineer. The search for and access to data in the public domain and in company files has been reduced to minutes using communication software, global networks, and modern computers. Process simulator software provides graphical representation of the individual units in any process flow sheet. The simulators perform the tedious tasks of completing mass and energy balances on individual process steps and on the overall process sequence. It takes simple entry of process stream information to run the program. The simulators do an excellent job with equilibrium or steady-state calculations that use the exact differential forms of thermodynamics. The programs include an extensive file of thermodynamic data to make the computation algorithms work. New processes using new materials often require new data. These new data can be entered or one can elect a properties estimation procedure provided by the simulator to fill gaps in the properties data set.

It is more difficult to write the computer simulator programs that involve rates: the chemical reaction rate, thermal energy flow, diffusion, and the work required to move material through the equipment. There are no general mathematical solutions to the second-order partial differential equations that represent these physical rate processes. There are solutions for some of the important process steps and additional empirical functions are often used to complete the algorithms.[2] These empirical functions work when there are experimental data to fix the empirical constants. The programmers use information from available plant data as the basis for the algorithms. Simulating processes to produce chemical compounds similar to those used to develop the programs yields the best results. For example, processes involving compounds derived from petroleum are usually accurately represented. The results are less reliable when the process involves new materials. On adding water and inorganic compounds to the process streams, the accuracy of the calculations will be degraded unless provision is made to provide data for these complex mixtures.

The experience and judgment of the process engineer is an important part of the process design. The process simulator will give the correct numerical answer to the problem it has been programmed to solve. It takes experience with the simulator to know how it solves a problem and someone with plant design experience who can ''feel'' when the solution is right. It is easy to reject ridiculous numbers from the process simulator. It is much more difficult to detect a ''reasonable but wrong'' number. There is no substitute for a person with experience to check and verify the numbers the simulator generates.

The process simulators include options to estimate the size of the equipment for each process step. Historic equipment cost data have been correlated so that the equipment can be sized and cost estimated.

Selecting and sizing equipment to perform a single process step is an easy step in process design when the temperature, pressure, and composition of the streams are known. It is much more difficult to find the operating conditions of all the streams to and from each piece of equipment when they are connected to form a continuous flow process. The overall plant optimization calculations get progressively more difficult as the number of process steps increase and constraints on process streams are added to meet all the design objectives for the plant.

Firms often lease process simulators and customize them to model processes running in their plant. Proprietary data and process information are added to the programs and they become the plant process control system. Add another function to the program and it follows the cost and composition of feed materials and computes the process set points that will yield the distribution of products that maximize the economic return for the full array of products.

Writing a program to simulate the dynamic (time dependent) behavior of a process during normal startup and shutdown is more challenging than the steady-state calculations. It is yet another problem to design a program to monitor and control the stream conditions and the critical sequence of steps that must occur during an emergency plant shutdown. Each of these problems represents a special challenge for process design and computer programming.

Process simulators are very helpful but the difficult problem in design is always process selection, finding the best sequence of steps for that process. When a proposed process involves several steps (usually pieces of equipment), several starting materials to obtain several products, how do you decide what is the best way to arrange the necessary process steps? At what point in the process do you feed material and withdraw

product? One can visualize many configurations of the process equipment that will produce the desired products. Application of the conceptual design principles and process experience has been the primary source for selecting the process steps for a plant.

There are two parts to the design of an optimum process sequence that must be addressed to apply the full power of computers for process design. There must be a mathematical model of the proposed process and there must be a solution to that mathematical model. Most of the process simulators are programmed to solve the mathematical model. The more difficult problem is to program the computer to select the ''best'' sequence of process steps from raw materials to products, i.e., process synthesis.

A procedure for programming the synthesis of any process system was proposed in 1992.[3,4] This proposal is based on a reduction of general mathematical programming methods by using the unique features of the process structure to solve the process system selection problem. Specifications of a processing system can include the different operating units (feeders, mixers, reactor, separators, washer, etc.) and the raw materials, products, by-products, and waste materials that will be included in the process. A mathematical axiom system, graph theory, and combinatorial techniques are used to find all feasible process structures embedded in a maximal structure for a proposed process that can then be modified in response to technical, environmental, economic, and societal constraints. The representations of the process steps are mathematically exact. The mathematical structure of the solution is polynomial rather than combinatorial and it converges quickly to identify the possible solutions for complicated processing structures, and the program runs on modern personal computers.

This process synthesis procedure is versatile as demonstrated by the extent of the reported applications. There are computer simulated plant designs to control environmental waste.[5] Selecting the arrangement of operating units for azeotropic distillation has been reported.[6] Determining a best catalytic reaction mechanism and identifying preferred pathways for a biochemical reaction have been described.[7,8] An optimization of a retrofit design for steam supply to an existing manufacturing complex has been reported.[9] Optimum separation networks-reported in the literature were improved using this procedure.[10] Examples are cited where the program identified unnecessary steps in a process. Equally valuable, the program showed it was impossible to obtain the products when the proposal was incomplete. There will be continued development of this program and probably other algorithms will be developed that will search out optimum process sequences.

EDUCATION FOR PLANT DESIGN

Chemical process design is well established in the chemical process industries. The individual skills and problem solving methods have served a wide range of other industries. Education is important for sharpening the skills of the practicing engineer and essential to introduce young people to the methods of process design.

There are two areas where education on plant design for individuals takes place. Students working toward an academic degree in chemical engineering will surely see assignments that include process design principles. The second area is the practicing process engineer who must learn new design technology and how to use new design tools. The practicing design engineer and his or her supervisor will decide the content of this educational assignment. A process simulator vendor may provide tutorials or a software expert may be called in to teach the specialized technology. New technology may enter from in-house research, through a consultant, or from a course offered by a software specialist. The employee and the management will set the educational objectives, assign the resources, and provide the time that will be devoted to this learning experience.

Education on plant design for students studying in an academic institution presents a different and may be a more complex problem. The Accreditation Board for Engineering and Technology requires that a design problem be assigned in the courses in each of the 4 or 5 yr of all undergraduate engineering programs. There will be a comprehensive design course in the final year of the bachelor's degree program. Most of the exercises assigned in chemical engineering courses show how a mathematical function can be used to describe the chemical or physical changes that take place in some apparatus or equipment. Given the function and input data, everyone will get the same numerical answer. These exercises are completely specified and there is only one correct answer.

Design problems will always be underspecified. Given a design objective, the task is to select among operations (heating, mixing, filtering, etc.) and, select the equipment to produce the chemical and physical changes that will meet the objective. Each process description will produce a different numerical answer. It is a ''judgment call'' to rank the results to find the best design solution. Writing tighter objectives will usually drive the numerical answers closer together. In an industrial setting, the experience the process engineer brings to the design, the available facilities, and the financial commitment to the project will guide the design through the steps from concept to an operating facility. The academic setting brings each design

assignment to a new class with an experience level much like a ''clean slate.''

It is often difficult to find an instructor for an undergraduate design course. Usually, the instructor will be a late career teacher or someone with industrial experience who joined the faculty to fill the process design position. Effective design instructors need a general knowledge of chemical engineering technology and should be familiar with new developments. The computer and a wide range of software have become essential tools in process design so the instructor must be able to direct students as they use these resources.

There are textbooks and references available to ''coach'' the instructor and help the students in the academic setting. The following is a partial list of books written as teaching aids for students. There is a process flow diagram for many commercial chemicals and one-step biological transformations that can be useful in a design course.[10–12] Chemical reactions and reactor systems are important in process design.[13,14] Several recent textbooks on process design have been written.[15–21] Most of these books include CD-ROM tutorials with worked exercises using the process simulator software.

The academic setting limits the selection of plant design problems and the expectations for completion of a work assignment. Some considerations that apply to an academic setting are:

- Time is limited. The plant design course is one of three or four courses students take during the semester. There is a limit of about 150–200 hr each student has available during the semester.
- Students lack design experience. Each student has a plan for life after the B.S. degree. Many will choose employment as a chemical engineer. Some will attend a professional school (medicine, law, graduate school, and business are common). It is unusual to find two students in a class with common interests and similar career objectives.
- Students are usually assigned to work in small groups of three or four. This mimics the group effort of industrial design teams. Only those students with one or two semesters on an industrial internship or a summer job in industry bring industrial experience to the group. It is best to assume that there is no design experience in the class.
- Select a design problem that the students understand is ''real.'' Students do better work on a problem they recognize from their reading or hear about on broadcast media.
- Check that there are data and descriptions of the technology available to the students. They do not have time for a data ''treasure hunt.'' If information is really not there, the instructor must intervene and supply a ''data fix'' soon after assigning the problem.
- Monitor the progress of each group. Design is the process of formulating questions and finding the answers. Inexperience often leads students to do what they know how to do rather than what needs to be done to complete the design assignment.
- The results of the design project should be presented in a written report. Interim written and oral reports help students stay on track. Reporting is preparation for future employment opportunities.
- All design assignments are due at a specified date and time. Design problems are always underspecified and open ended so the design report never ends with QED. The conclusions will include recommendations for continuing the design. This report would be the basis for committing resources to the next design step in an industrial setting. Grades are assigned in the academic setting and the design sequence ends with the semester.

An academic design assignment must be a clear statement of the design objective. Examples include a change in the processing sequence of an existing plant, evaluating the advantage of a new catalyst in an existing plant to produce an improved product, assessing the economics of using microorganisms to decontaminate a waste water stream, producing hydrogen for transportation fuel. The list of possible assignments is long. The list of good class assignments is much shorter.

The library and the Internet are the prime sources for process information. Students should visualize the necessary process steps for their assignment, find what others have done that can be applied to their design. A feed cost and product revenue calculation is the first economic analysis. The early stage of the design process is where heuristics are helpful. A simple mass and energy balance is useful at this stage. Current textbooks suggest orderly ways of selecting process alternatives, sequencing process steps, and sizing equipment.[1,15,16,21] The best two or three process alternatives will usually produce one that will yield the final design report.

An academic process design course should include use of a process simulator. There are several simulators available with some offered at a high discount to academic institutions. The ''learning curve'' for these programs is rather steep, so it helps to start using the program in the sophomore and junior courses.

Students find it tempting to get on to simulator calculations before they have settled on a strategy or a process sequence for their design problem. It saves time to plan carefully what needs to be computed and what will be learned from those numbers. The numbers from the simulator have little value unless they provide

insight that supports the conclusions that become the final design report.

The advantage of process simulator calculations comes when the sequence of process steps has been set. The process simulator draws the process flow sheet specified and uses it to fix the computer algorithm for the process. There are optimization strategies to select the ''best'' conditions of the feed streams to the equipment and operating conditions for the whole process.

The American Institute of Chemical Engineers sponsors an annual process design contest for students. A problem statement is distributed and students are given 30 days to complete the design analysis and write their report to be eligible for judging in the competition. There is an individual and a small-group category in the contest. One or two of the best reports are selected from the class and submitted to the competition committee. Spending 30 days immersed in one design problem is a teaching strategy that can be a good learning experience. This is an opportunity for independent work, much like an industrial setting. Assign the problem early in the semester and use the remaining time to do the next iteration on the problem or select a new problem that covers another processing technology.

TRENDS IN CHEMICAL PROCESSING

The historic record of developments in chemical process technology indicates trends that point to the future. There are established industries that support our national infrastructure and there are new ideas for products that will be developed and become part of everyday commerce. Energy for transportation and electric power production represents large industrial investments in physical facilities. The energy industries generate huge revenues that make them powerful players in economics and politics on a global scale. Fossil fuels (petroleum, coal, and natural gas) are the source of most of the energy we use.

Petroleum refining is an example of a mature processing industry. It produces nearly all of the transportation fuels (cars, trucks, aircraft, and trains). A recent historic timeline of 100 yr of advances in chemical engineering technology was published and shows the dates of significant developments in refining.[22,23] As a mature industry, minor process changes reduced refining costs and the response to shifts in product demand yielded large financial savings. Major changes occurred when federal environmental policy mandated reduced air and water pollution. Tetraethyl lead used to increase the gasoline octane number was eliminated and this meant that the petroleum refiners had to learn to ''refine in'' higher octane numbers. This new fuel still had too low an octane number so the compression ratio of the automobile engines was reduced to efficiently use this fuel and meet the new exhaust emission standards. As the petroleum refining industry matured through the 20th, century it provided much of our modern chemical plant design technology.

The impact of greenhouse gases on local and regional climate is currently being debated. The average carbon dioxide concentration in the atmosphere has increased significantly over the past century with combustion of fossil fuels adding to the natural sources of this greenhouse gas. It does appear that international policy will be written for a reduction in the rate of release of carbon dioxide and other greenhouse gases. The response to such policies will require innovative fuel technology. Increasing the energy efficiency of automobiles and industrial processes that use fossil fuel is one way to reduce per capita carbon dioxide emissions and continue the trend of the past 30 yr.

One proposal to reduce pollution in cities is to use hydrogen as a transportation fuel. Hydrogen fuel cells produce electricity to power automobiles or to power televisions and they discharge water. However, most of the world's hydrogen is in hydrocarbons (methane, ethane, propane, ... petroleum) or in water. Extracting hydrogen from hydrocarbons releases carbon dioxide. Hydrogen can be obtained from water using electricity, but using coal to generate electricity releases carbon dioxide. So, use a solar cell to produce electricity or can a solar cell be made to produce hydrogen? Wind can produce electricity and yield hydrogen. Is electricity from thermal energy produced by nuclear fission (a nuclear power station) a possible source of hydrogen without producing carbon dioxide? All of these options are available today or can be made ready with additional scientific advances and new technology to produce fuel hydrogen. Switching from petroleum derived fuels to hydrogen requires large financial investments for production and distribution. Assuming there is the ''will'' to move to hydrogen fuel, it will probably take half of the 21st century to modify the petroleum fuel infrastructure and replace all of the vehicles that use petroleum based fuel.

Miniaturization and nanotechnology are becoming part of our technology tool kit. The discovery of the transistor revolutionized electronics and required new processing methods to produce the transistor chips. The circuits on a chip are now so small that quantum effects are important. ''Chemical analysis on a chip,'' which promises a medical doctor a complete blood analysis while interviewing a patient is being developed. Faster and cleaner chemical conversions may be possible with catalysts that use nanostructures. Process technology in miniature will follow these developments. Is there an answer to making new fuels concealed in ''small'' processing?

Technology development historically follows new science. Major research funding currently goes to the

biological and medical sciences. Basic studies seek to unravel the secrets that produce and maintain life. Plants are living chemical reactor systems that ''self-assemble'' from a seed, generate the enzymes required to function, and use solar energy to produce a remarkable array of chemical compounds. Animals have an even more dazzling array of attributes produced by a wider range of chemical reactions that run on energy extracted from those plant produced chemicals. All of this chemistry runs at nearly constant temperature (not far from room temperature) with mass and energy transport between cells and gels that make up the flesh and fluids of living systems. Yeast producing alcohol from sugars has been ''commercialized'' for centuries. Is there a new science and technology based on what we can learn from living organisms to collect energy from the sun? After all, the fossil fuels we use today are the naturally sequestered remains of living organisms.

FUTURE FOR PROCESS DESIGN

Education on plant design certainly has a future. Plant design in industry will always respond to the demands of maintaining production of commodity chemicals. Specialty chemicals are profitable if there is constant improvement in product quality and research continues to renew the list of products that come from scientific research and the new technology produces it to fill the new market demands.

The large investment in chemical production facilities usually keeps them in production for a long time. Most of the plants that are running today will be running 25 yr from now. There will be improvements in efficiency and response to the demands to reduce emissions. Process design will play a key role in maintaining these production facilities.

The economic competition that comes from the global economy will produce changes in the process industries and plant design.[24] Improved energy efficiency will be important in chemical processing. Fuel price fluctuations pass directly to the product costs are so designed for efficient energy use.

There will be a greater emphasis on safety. It is important for workers in the plant and essential to be a good industrial citizen in the community.

There will be new regulations limiting emissions from the production facility and from customer use of the chemical products. The optimum new process sequence will be designed so there is essentially zero discharge.

There will be a closer coupling between the new science and the new technology. Process modeling will begin as soon as there are results from the laboratory suggesting that there is a new product with commercial potential. New computer software will support this activity.

Process control of the manufacturing plants will continue to improve. There will be less material in process, the equipment will be smaller and of higher capacity so ''sharp'' control should be in place. Better control will give improved conversions and enhance product quality, especially for high-value-added products.

Current education on plant design emphasizes commodity chemicals in large manufacturing complexes. There is an increasing manufacturing response to customer demand for expensive, specialty products. This will lead to a shift from current process design toward product design.[24] These new production facilities will use general-purpose equipment that can be configured to make one specialty chemical and quickly switched to another product in response to seasonal or customer demand.

Education on process design does carry a bonus. Understanding how to work an open-ended process design problem is good citizenship training. Young people in process design courses will mature into leadership roles in industry and in their communities. Successful leaders spend every day working on open-ended problems. Process design is a discipline that provides the framework to select solutions to these problems.

CONCLUSIONS

The history of plant design spans more than 100 yr. In that time, the fingerprints of plant design specialists cover commercial developments that include extracting metals from ores, fixing nitrogen, bringing liquid fuels out of petroleum and coal, producing polymers for fibers, films, and solid plastics, processing foods, preparing pharmaceuticals, a long list of domestic and military products. Education on plant design focuses on the systematic procedure for visualizing the solution to a design proposal. The education of a process design specialist begins as an undergraduate student and continues over the professional lifetime. Digital computer software plays a key role in data collection, process control, and new plant design analysis. Plant design will move in the direction of product design with lower-volume, higher-value-added products using environment friendly processing. Biochemistry and biotransformations will be increasingly important in the chemical process industries.

REFERENCES

1. Douglas, J.M. *Conceptual Design of Chemical Processes*; McGraw-Hill: New York, 1988.
2. Rice, R.G.; Do, D.D. *Applied Mathematics and Modeling for Chemical Engineers*; John Wiley & Sons, Inc.: New York, 1995.

3. Friedler, F.; Tarjan, K.; Huang, Y.W.; Fan, L.T. Graph-theoretic approach to process synthesis: axioms and theorems. Chem. Eng. Sci. **1992**, *47* (8), 1973–1988.
4. Friedler, F.; Tarjan, K.; Huang, Y.W.; Fan, L.T. Graph-theoretic approach to process synthesis: polynomial algorithm for maximal structure generation. Comput. Chem. Eng. **1993**, *17* (9), 929–941.
5. Bumble, S. *Computer Simulated Plant Design for Waste Minimization/Pollution Prevention*; Lewis Publishers: New York, 2000.
6. Feng, G.; Fan, L.T.; Friedler, F.; Seib, P.A. Identifying operating units for the design and synthesis of azeotropic-distillation systems. Ind. Eng. Chem. Res. **2000**, *39* (1), 175–184.
7. Fan, L.T.; Bertok, B.; Friedler, F. A graph-theoretic method to identify candidate mechanisms for deriving the rate law of a catalytic reaction. Comput. Chem. **2002**, *26* (3), 265–292.
8. Seo, H.; Lee, D.-Y.; Park, S.; Fan, L.T.; Shafie, S.; Bertok, B.; Friedler, F. Graph-theoretic identification of pathways for boichemical reactions. Biotechnol. Lett. **2001**, *23* (19), 1551–1557.
9. Halasz, L.; Nagy, A.B.; Ivicz, T.; Friedler, F.; Fan, L.T. Optimal retrofit design and operation of the steam-supply system of a chemical complex. Appl. Therm. Eng. **2000**, *22* (8), 939–947.
10. Kovacs, Z.; Ercsey, Z.; Friedler, F.; Fan, L.T. Separation-network synthesis: global optimum through rigorous super-structure. Comput. Chem. Eng. **2000**, *17* (9), 929–941.
11. Speight, J.G. *Chemical Process and Design Handbook*; McGraw-Hill: New York, 2002.
12. Liese, A.; Seelbach, K.; Wandrey, C. *Industrial Biotransformations*; Wiley–VCH: Weinheim, Federal Republic of Germany, 2000.
13. Fogler, S.H. *Elements of Chemical Reaction Engineering*, 3rd Ed.; Prentice Hall PTR: Upper Saddle River, NJ, 1999.
14. Davis, M.E.; Davis, R.J. *Fundamentals of Chemical Reactor Engineering*; McGraw-Hill: New York, 2003.
15. Biegler, L.T.; Grossmann, I.E.; Westerberg, A.W. *Systematic Methods of Chemical Process Design*; Prentice Hall PTR: Upper Saddle River, NJ, 1997.
16. Seider, W.D.; Seader, J.D.; Lewin, D.R. *Process Design Principles Synthesis, Analysis and Evaluation*; John Wiley & Sons Inc.: New York, 1999.
17. Edgar, T.F.; Himmelblau, D.M.; Lasdon, L.S. *Optimization of Chemical Processes*, 2nd Ed.; McGraw-Hill: New York, 2001.
18. Hangos, K.M.; Cameron, I.T. *Process Modeling and Model Analysis*; Academic Press: New York, 2001.
19. Luyben, W.L. *Plantwide Dynamic Simulators in Chemical Processing and Control*; Marcel Dekker, Inc.: New York, 2002.
20. Peters, M.S.; Timmerhaus, K.D.; West, R.E. *Plant Design and Economics for Chemical Engineers*, 5th Ed.; McGraw-Hill: New York, 2000.
21. Turton, R.; Balie, R.C.; Whiting, W.B.; Shaeiwitz, J.A. *Analysis, Synthesis, and Design of Chemical Processes*, 2nd Ed.; Prentice Hall PTR: Upper Saddle River, NJ, 2003.
22. Chopey, N.P., Ed.; CE centenary: the flow of history. Chem. Eng. **2002**, Aug, 81–136.
23. Keller, G.E. II; Bryan, P.F. Process engineering: moving in new directions. Chem. Eng. Prog. **2000**, Jan, 41–50.
24. Cussler, E.L. *What happens to Chemical Engineering Education*? Phillips Petroleum Company Lecture Series; Oklahoma State University: Stillwater, OK, 2002.

Electrodeposition

André Avelino Pasa
Thin Films and Surfaces Group, Departamento de Física, Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, Brazil

Maximiliano Luis Munford
Group of Organic Optoelectronic Devices, Departamento de Física, Universidade Federal do Paraná, Curitiba, Paraná, Brazil

INTRODUCTION

Most important concepts and techniques of electrodeposition are introduced and described here, with examples of technological applications are given, with emphasis on the fabrication of nanostructures. A brief history of the process development is also given.

DEFINITIONS AND HISTORY

Electrodeposition is an electrochemical process that allows the preparation of solid deposits on the surface of conductive materials. It is a commercially highly relevant process, providing the basis for many industrial applications, such as electro-winning, refining, and metal plating. Metal plating is the process that has perhaps the closest contact with most people's everyday life, because we are surrounded by things that have a protective or decorative coating, such as watches, buttons, belt buckles, doorknobs, handlebars, etc. Additionally and more recently, as will be seen below, not only do the circuit boards and the packaging modules of computers, but also the recording and reading heads of their hard disk drives and the microprocessor chip itself may have plated material on them.

Electrodeposits are formed by the action of an electric current passing in an electrochemical cell, a device that consists of two conductive or semi-conducting electrodes immersed in an electrolyte. The electrodes are called the working electrode (cathode), consisting of the object where electrodeposition is planned, and the counter-electrode (anode), necessary to complete the electrical circuit. Electrolytes for electrodeposition are usually aqueous solutions containing positive and negative ions, prepared by dissolving metal salts. The electric current that flows between the two conductive electrodes in the presence of an external voltage is because of the motion of charged species, via migration and diffusion, towards the surfaces of the polarized electrodes. At the surface of the electrodes, the conduction mechanism must change from ionic to electronic, an interface process mediated by the occurrence of electrochemical reactions that promote the reduction or the oxidation (redox reactions) of the ionic species.

An electrochemical cell with a battery is illustrated in Fig. 1, where the motion of the ions toward the electrodes is also sketched. In this case, the metallic salt $NiSO_4$ (nickel sulfate) dissolved in water is a practical example of an electrolyte for Ni-plating metallic objects. In this example, the object to be plated is a key, placed as the working electrode. By applying an external voltage with the negative terminal of the battery connected to the working electrode, the Ni^{2+} move to this electrode, where deposition takes place, and the SO_4^{2-} toward the positively charged counter-electrode.

An essential characteristic of electrochemical reactions is that the exchange of charge does not occur between chemical species, as it does in a typical chemical reaction, but between chemical species and the electrode. The electrochemical reaction that is most important for the electrodeposition process is the one that occurs at the working electrode; i.e., for the example, in Fig. 1, it is the reduction reaction $Ni^{2+} + 2e^- \rightarrow Ni^0$, where the Ni ions are reduced by receiving two electrons (e^-) from the electrode. At the counter-electrode, the oxidation of the sulfate radical is too energetic to occur, and the most probable oxidation reaction for inert electrodes in an aqueous electrolyte is the electrolysis of the water, forming H^+ and O_2 through the reaction $H_2O \rightarrow 2H^+ + {}^1/_2O_2 + 2e^-$. This reaction occurs by donating two electrons to the anode, completing the electrical circuit and keeping the electric charge balanced. Frequently used inert electrodes include platinum and glassy carbon. If the counter-electrode is a metallic bar or foil (a non-inert electrode), the electrodissolution of the metal could provide electrons for the electrode and ions for the solution.

Historically, the discovery of electrodeposition is attributed to Luigi V. Brugnatelli, an Italian professor, who in 1805 was able to electrodeposit gold on the surface of a metallic object, from a solution containing dissolved gold, using a voltaic pile (battery). About 40 years later, John Wright, from Birmingham, England,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120037171

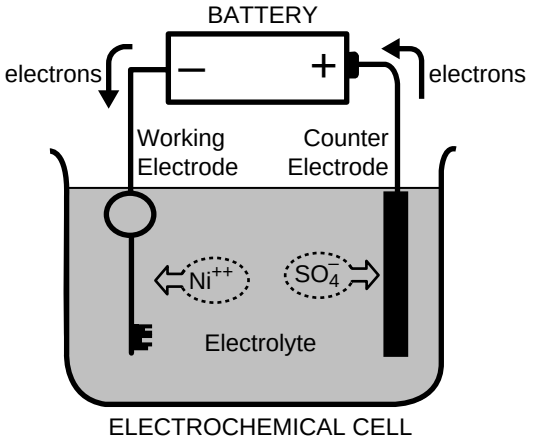


Fig. 1 Positively charged nickel ions in the electrolyte are attracted by the negatively charged key (working electrode). At the surface of the key they are reduced by gaining two electrons, and metal is deposited.

discovered that potassium cyanide was a suitable electrolyte for gold and silver electroplating. This discovery made electrodeposition an important commercial process for covering the surface of various kinds of metallic object with thin coatings of metals for corrosion protection and decorative purposes. Subsequently, baths for the deposition of other metals and alloys such as nickel, zinc, tin, and brass (an alloy consisting essentially of copper and zinc in variable proportions) were developed. For the next 100 years, the main idea was to use electrodeposition for covering the surface of inexpensive materials with a thin layer of a noble metal. By the 1940s, however, electrodeposition was rediscovered by the electronics industry. The electrodeposition of gold for electronic components was a totally different kind of application of electrodeposition techniques.

Over the years, electrodeposition became a highly developed process. Direct current (DC) power supplies were developed; anodic and cathodic reactions were described; new safer baths based on acid electrolytes, avoiding the earlier poisonous cyanide-based ones were discovered; models for the deposition process incorporating mass transport to the electrodes, charge transfer kinetics, and nucleation and growth at the working electrode were developed; and regulatory rules for waste water emission and waste disposal were created. Simultaneously, a gradual improvement in electrodeposition for large scale manufacturing processes took place.

ELECTRODEPOSITION APPARATUS AND CONCEPTS

Electrodeposition on the industrial scale requires an electrochemical cell and a DC current power supply. This approach is relatively simple and inexpensive, and is known as galvanostatic plating system, because the current between the electrodes is controlled (maintained constant). Another important deposition mode is the pontentiostatic one. This mode is a consequence of the development of electrochemical science, where electrochemical reactions at the surfaces of electrodes are carefully investigated. The electrochemist developed reference electrodes, in order to measure the potential drop near the surface of electrodes. Assuming that the electrolyte contains sufficient ions (has sufficiently high conductivity) to avoid any ohmic voltage drop, because of the resistance of the electrolyte between the electrodes, all the voltage applied by the battery (in Fig. 1) will appear near the surface of the electrodes, where a charged region is formed (usually named the double layer). It is very important to measure the voltage drop across these charged regions, because it controls the driving force for phase transformation from ion to reduced state. A simple metal foil could be used as a reference electrode; however, because of the need to have a standard electrode to measure potential drops at the surface of different types of working electrode in contact with different electrolytes, a hydrogen electrode was elected, and now all electrode potentials are quoted relative to this arbitrarily chosen reference electrode. What is always measured is the potential difference between two electrodes. By defining the potential of the hydrogen electrode as zero, it is possible to generate a table of all the possible redox reaction potentials relative to this electrode, and these potentials are called standard potentials. The standard hydrogen electrode (SHE), which is usually constructed by bubbling hydrogen gas over an immersed platinum foil, has its operation based on the redox reaction $H_2 \leftrightarrow 2H^+ + 2e^-$. Other reference electrodes that are robust, stable, and easily constructed than the SHE are frequently used in the laboratory, give potential measurements that can be converted to standard potentials by adding or subtracting a constant value. The most common are the calomel electrode (Hg/Hg_2Cl_2) and silver/silver chloride electrode (Ag/AgCl).

The potential of an electrochemical cell, also known as the cell potential or electromotive force (emf) is the sum of the potential drops at the cathode and anode, where the reduction and oxidation reactions occur. With the introduction of a reference electrode the potentials of these two electrodes can be measured independently, allowing the independent investigation of the reactions that are taking place at each electrode (working or counter). These redox reactions are called half-cell reactions or simply half-reactions. The half-reaction potential E^0 can be measured with a SHE electrode at standard conditions, i.e., at electrolyte concentrations of 1 M, gas pressures of 1 atm., and

Table 1 Standard electrode potentials in aqueous solution at 25°C

Cathode half-reaction	E^0(V)
Na^+ (aq) + e^- → Na (s)	−2.71
$2H_2O$ (l) + $2e^-$ → H_2 (g) + $2OH^-$ (aq)	−0.83
Fe^{2+} (aq) + $2e^-$ → Fe (s)	−0.41
Ni^{2+} (aq) + $2e^-$ → Ni (s)	−0.23
$2H^+$ (aq) + $2e^-$ → H_2 (g)	0.00
$2Cu^{2+}$ (aq) + $2e^-$ + $2OH^-$ → Cu_2O (s) + H_2O (l)	0.17
Cu^{2+} (aq) + $2e^-$ → Cu (s)	0.34
O_2(g) + $4H^+$ (aq) + $4e^-$ → $2H_2O$ (l)	1.23
$S_2O_8^{2-}$ (aq) + $2e^-$ → $2SO_4^{2-}$ (aq)	2.01

Note: aq., g and I denote aqueous, gas and liquid respectively.

temperature of 25°C, and tabulated. Table 1 shows a set of standard potentials for cathode half-reactions.

The introduction of the reference electrode led to a different experimental setup for electrochemical and electrodeposition experiments. Fig. 2A shows an electrochemical cell with three electrodes (working-, reference-, and counter-electrodes) and a potentiostat. The potentiostat is an electronic apparatus that maintains the potential difference between the working- and reference electrodes by controlling the potential difference between the working- and counter-electrodes. Fig. 2B shows a block diagram of the electronic circuitry of a potentiostat with an operational amplifier that keeps the voltage between reference electrode (RE) and working electrode (W) equal to the applied voltage E at the positive terminal, by regulating the cell potential between W and counter-electrode (CE). By convention W is connected to ground.

The three-electrode cell and the potentiostat enable the potentiostatic mode of deposition mentioned above. The potentiostatic mode means that the potential of the working electrode is kept constant during the experiment or deposition process, as depicted in Fig. 3A. An additional mode called pulsed deposition is also illustrated in Fig. 3B. In this mode, for pulsed potential, the potentiostat switches the working electrode potential between two values in order to have the potential varying as a square wave. For pulsed current deposition, a current source with a square wave output is sufficient.

The three-electrode cell and potentiostat is also a powerful experimental tool for electrochemical investigations, permitting the implementation of different techniques, such as voltammetry. This technique consists of applying a potential ramp to the working electrode, which is achieved by applying a potential ramp to the positive terminal of the operational amplifier (of Fig. 2B), and measuring the resultant cell current. When the applied potential starts at a defined level and comes back to the same value after a period of time, the technique is called cyclic voltammetry. When the applied potential starts at a level 1 and goes to a level 2, the resulting plot of the current versus the potential, is called a polarization curve or

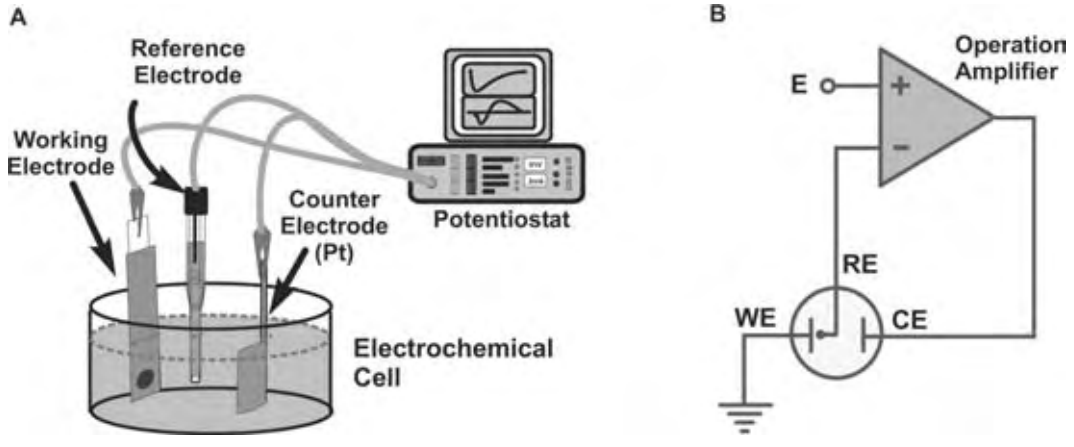


Fig. 2 (A) Electrochemical cell with three electrodes connected to a potentiostat. (B) Electronic sketch illustrating the mode of operation of a typical potentiostat.

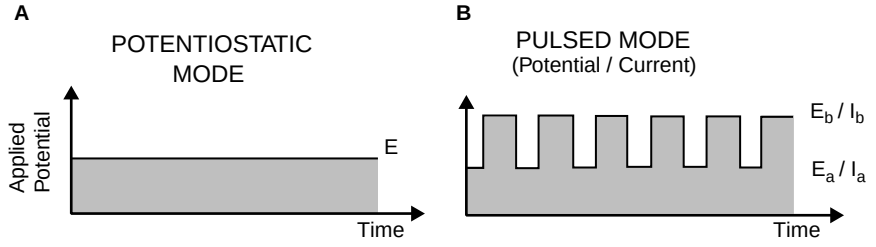


Fig. 3 Illustration of the potential as a function of time for potentiostatic and pulsed deposition modes.

simply voltammogram. Fig. 4A illustrates a potential ramp applied to the working electrode, and Fig. 4B the corresponding plot for the variation of the cathodic current as a function of the cathodic potential. This figure shows a typical polarization curve for the reduction of a metal at a conductive electrode. The onset of current, indicated by an arrow, corresponds to the minimum cathodic potential at which the reaction occurs (a fundamental value for electrodeposition purposes), the peak corresponds to the maximum current at a given rate of change of potential (also known as the reduction peak), while the third characteristic feature of the plot is the saturation of the current at more negative potentials. The intensity of the reduction peak depends on the potential scan-rate, this peak being not observed, in many cases, because of other reactions that occur simultaneously.

Potentiostatic Deposition

This section will describe the pontentiostatic mode using an electrolyte containing cobalt sulfate. By applying a potential ramp it is possible to obtain the polarization curve shown in Fig. 5A. In this figure the onset of Co reduction is about -0.8 V. By selecting a deposition potential negative than this value, it is possible to obtain a deposit. Fig. 5B illustrates a Co deposit that is very homogeneous, apart from a microscopic defect because of a hydrogen bubble, obtained at a deposition potential of -1.1 V. Fig. 5C shows a plot of the deposition current. This plot, known as a current transient, gives information about the deposition process and allows the calculation of the electrodeposited charge from the area below the curve.

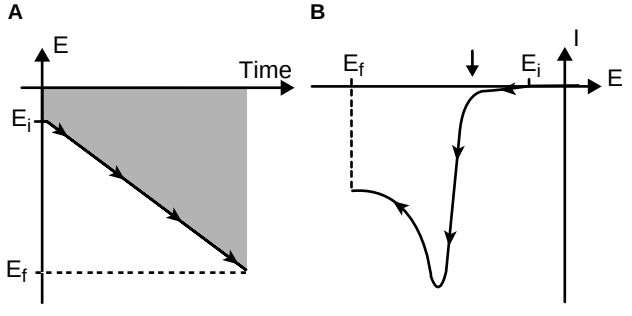


Fig. 4 Illustration of the applied potential ramp (A) to obtain a typical polarization curve (voltammogram) of metal deposition on a metal working electrode.

ELECTRODEPOSITION MECHANISMS

Electrodeposited Charge

Because the electrodeposition process involves the transfer of electrons to an electrode, by measuring the current in the electrochemical cell, it is in principle possible to calculate the amount of material deposited. If no other reaction occurs in parallel, then we can assume that the reaction at the working electrode in aqueous electrolyte is just the simple reduction of a metal (M)

$$M^{n+} + ne^- \rightarrow M^0, \quad (1)$$

where a metal ion M^{n+} is reduced to a metal atom M^0 after gaining n electrons. By assuming that all the metal ions reduced at the surface of the working electrode stick to this surface, the total amount of electrodeposited material can easily be calculated from the charge Q (in coulombs), which represents the product of the total amount of electrodeposited atoms N times the charge of n electrons, as given by the expression

$$Q = Nne, \quad (2)$$

where e is the charge of one electron, equal to -1.6×10^{-19} C. The charge Q is calculated from the current transient. If the deposition current is constant, Q can be calculated by simply multiplying the current I by the deposition time t,

$$Q = It. \quad (3)$$

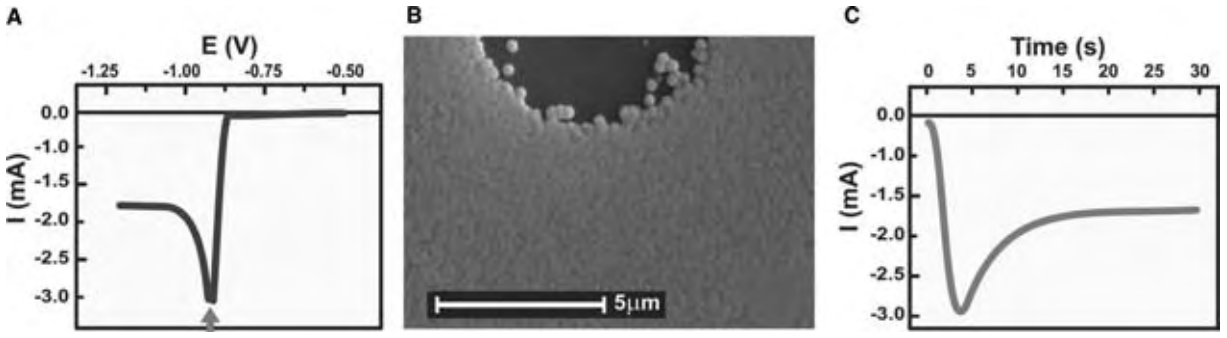


Fig. 5 (A) Polarization curve of an electrolyte containing cobalt sulfate, (B) scanning electron microscope (SEM) image of a Co deposit obtained at −1.1 V, and (C) the corresponding deposition current as a function of time (current transient). The reference electrode was saturated calomel and the working electrode semiconducting silicon. (From Ref.[1].)

However, if the current is varying during the deposition, as shown in Fig. 5C, Q can be calculated by integrating the current $I = I(t)$ as a function of time,

$$Q = \int I(t)\mathrm{d}t. \tag{4}$$

In order to calculate the thickness h (m) of the deposit on a known area A (m^2) of the surface of the electrode where deposition occurred, the quantity N can be expressed as

$$N = \frac{mN_a}{M}, \tag{5}$$

where m is the electrodeposited mass (g), N_a is Avogadro's number (the number of atoms in a mole, equal to 6.02×10^{23}) and M is the atomic weight. Using the density d (g/m^3), given by $d = m/V$, where the volume V (m^3) is given by the product of the area, and the thickness, $V = Ah$, Eq. (5) can be rewritten as

$$N = \frac{dAhN_a}{M}, \tag{6}$$

and Eq. (2) rewritten as

$$Q = \frac{ndAhN_a e}{M} = \frac{ndAhF}{M} \tag{7}$$

where F is Faraday's constant, defined as $F = N_a e$, equal to 96,485.34 C. By rearranging Eq. (7) and knowing the quantities M, Q, n, d, A and F, the thickness can be easily calculated from:

$$h = \frac{MQ}{ndAF}. \tag{8}$$

This calculation gives a deposit thickness in meters that has to be converted into more adequate units as microns ($1\,\mu m = 10^{-6}\,m$) or nanometers ($1\,nm = 10^{-9}\,m$). For a precise calculation of the thickness of deposits, it is necessary to take into account possible electrode reactions that occur simultaneously with the main reaction. One very common example is the hydrogen evolution reaction, $2H^+ + 2e^- \rightarrow H_2$. This reaction is so rapid, in conditions such as relatively high cathodic potentials in acidic electrolytes, that it dominates the exchange of electrons at the surface of the electrode. The contribution of hydrogen evolution to the cathodic current must be determined in order to obtain the efficiency of the plating process. This may be done indirectly by measuring the thickness of the deposit and calculating the amount of charge corresponding to ''missing'' metal. The presence of hydrogen during the deposition has several effects on the metallurgical properties of the deposits. One of these effects is the formation of gas bubbles that mask the surface of the electrode locally, introducing microscopic defects in the electrodeposited layers (see example of this effect in Fig. 5B).

Mass Transport

Electrodeposition has the ability to produce a relatively uniform distribution of metal upon a cathode of irregular shape. Though the uniformity depends on the distribution of electric fields inside the electrolyte toward the surface of the electrode, other important factors have to be considered. The addition of agents (additives) to the electrolyte, for example, can affect the microscopic mechanism of electrodeposition, reducing the roughness of the deposit and producing a visual effect known as brightening.

In order to obtain layers with a desired property, such as uniform thickness and low roughness, or an

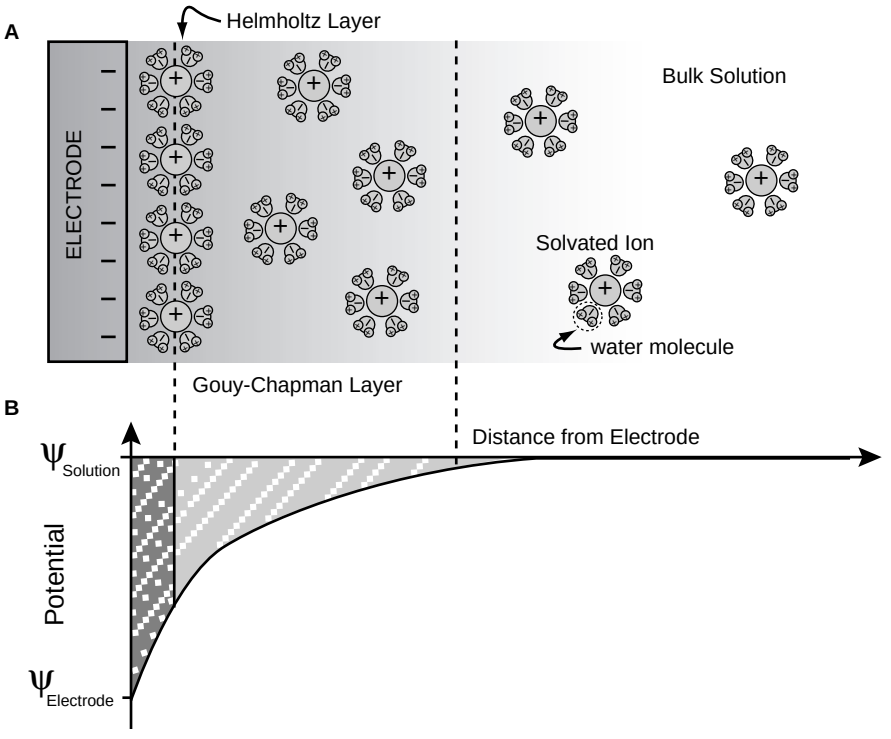


Fig. 6 Illustration of the electric double layer (A) and potential drop (B) near the surface of the electrode.

electrolyte with high filling capacity, i.e., with the ability to produce deposits inside holes or lith ographic features, one has to consider carefully the transport of different species inside the electrolyte and the reaction rates of these species on the surface of the electrode.

Basically, three mechanisms are responsible for mass transport inside an electrochemical cell: diffusion, migration, and convection. Diffusion is mass transport because of concentration gradients, i.e., variations in the concentration of a species with position. Diffusion occurs mainly near the electrode surface because of gradients created by the consumption of species that undergo redox reactions and are incorporated into the deposit. This incorporation process depletes the deposition species near the electrode, generating the concentration gradient.

The simple introduction of an electrode into an electrolyte will lead to an exchange of charge between electrolyte and electrode. An electric double layer, illustrated in Fig. 6, will be formed. Fig. 6A shows the double layer in greater detail. When the exchange process reaches equilibrium the double layer thickness depends on the physical and chemical properties of the electrode/electrolyte interface. In general, the description of the double layer considers the existence of two regions named the Helmholtz layer, a monolayer of ions on the surface of the electrode, and the Gouy-Chapman layer, a region that penetrates the bulk of the electrolyte with decreasing charge and concentration gradients. In the figure, the ions are shown with a sheath of water molecules. This solvation sheath is because of the electrostatic attraction of polar water molecules and the ionic species. Fig. 6B illustrates the potential drop near the surface of the electrode because of the presence of the double layer. In the bulk of the electrolyte the potential drop is

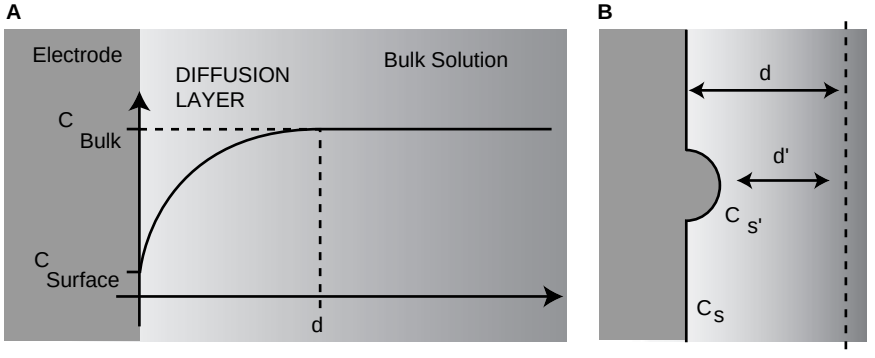


Fig. 7 (A) Idealized profile of the concentration gradient near the surface of the electrode and (B) an illustration of a deposition instability in the presence of an asperity.

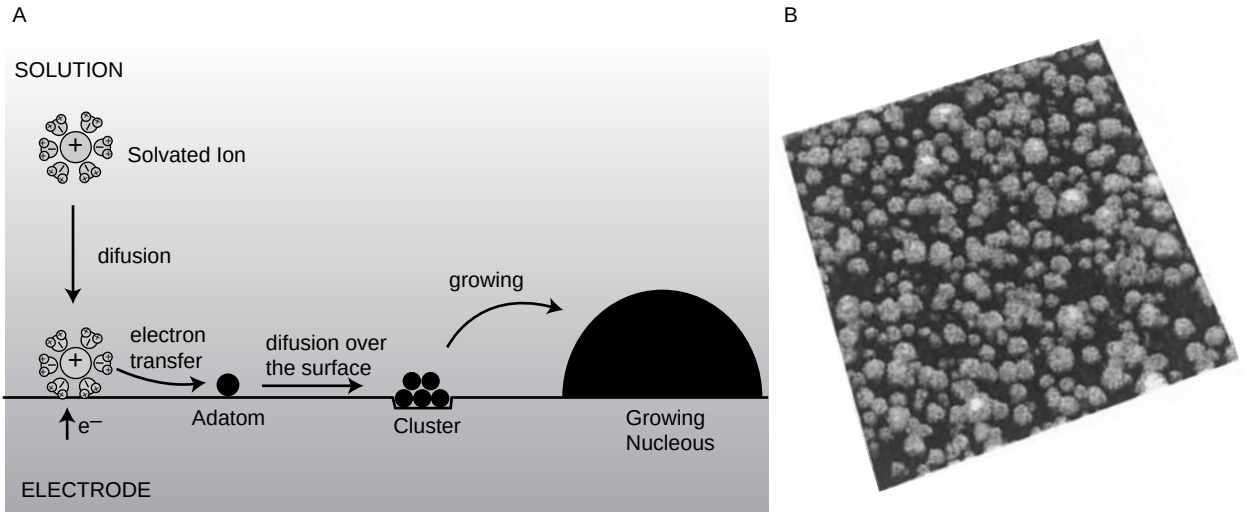


Fig. 8 (A) Mechanism of formation of electrodeposits on the surface of an electrolyte and (B) an atomic force microscope (AFM) image reveals the granular nature of a Co deposit. (From Ref.[1].)

normally very low, because this region is not depleted of ionic species.

Beyond the double layer, there is a depleted region named the diffusion layer with a thickness of microns, much wider than the double layer, formed during deposition by the consumption of a particular species. Fig. 7A is a plot of the concentration of an ionic species as a function of the distance from the surface of the electrode, showing the diffusion layer. The consumption of ions because of metal deposition generates a concentration gradient that, in steady-state conditions, is essentially determined by the redox reaction rate. If the consumption of ions arriving at the surface by diffusion is very high, the concentration of ions at the surface C_S is effectively zero, and the deposition process is controlled by diffusion. If the consumption is low, then the ion concentration at the surface is different from zero and the deposition is controlled by kinetics, i.e., by the velocity of the reaction.

Electrodeposition in the diffusion-limited regime is very sensitive to asperities on the surface of the electrode, as shown in Fig. 7B. These morphological structures reduce the size of the diffusion layer, thereby increasing the concentration gradient and hence the current. The local increase in current increases the deposition rate, favoring further growth of the asperities, and consequently an increase in roughness of the whole electrodeposited layer. This effect is known as deposition instability.

Inside the bulk of the electrolyte, mass transport is mainly because of migration, a mechanism of ionic motion caused by the presence of an applied electric field. In the electrochemical cell the potential drop creates an electric field that is much more intense in the regions near the surface of the electrodes, but is sufficiently intense in the bulk of the electrolyte to promote the migration of the ions to the border of the diffusion layers.

The third important mechanism of mass transport is convection. In this case, the fluid flows in an uncontrolled manner because of natural density gradients (gradients caused by concentration and temperature

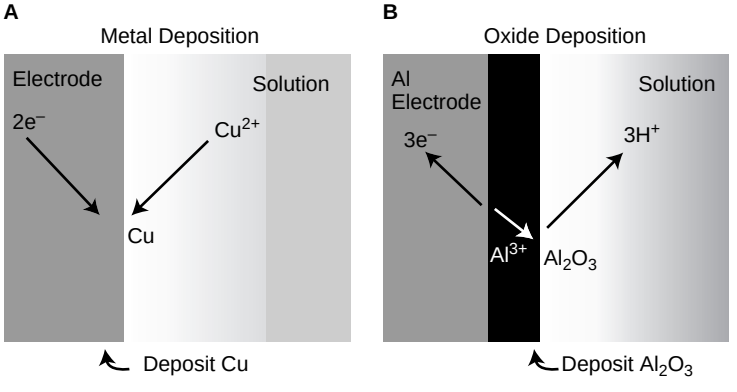


Fig. 9 Electrode reaction for (A) the cathodic electrodeposition of Cu and (B) the anodic electrodeposition of Al oxide.

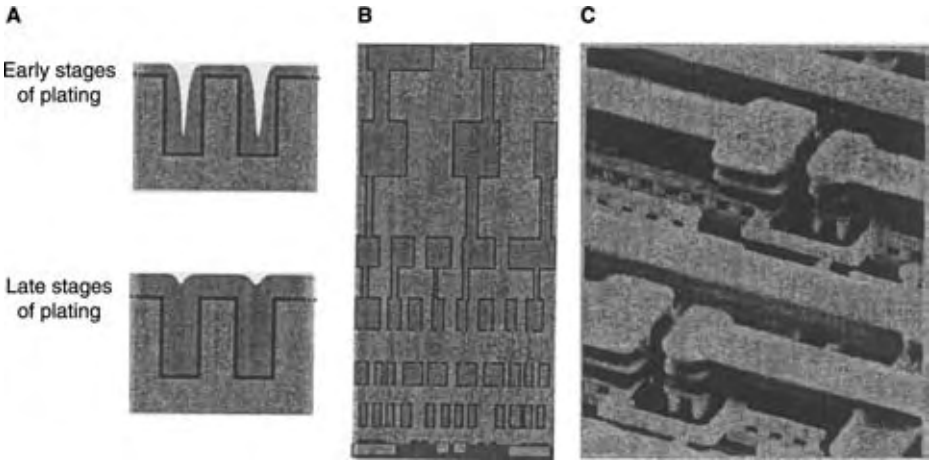


Fig. 10 (A) Sequence of filling of a trench profile for the fabrication of Cu interconnects, (B) a cross-section illustration of a six-level wiring structure, and (C) SEM view of IBM's first-to-market six level copper interconnect technology. (From Ref.[2].)

fluctuations). Convection can also be produced in a controlled manner by different methods, such as mechanically stirring the electrolyte.

Growth Mechanisms

To complete the explanation about the mechanisms of electrodeposition, it is very important to give an idea about the formation of the deposits. A model for the electrodeposition process considers a solvated ion going through the diffusion layer as a first step, loosening of the solvation sheath by transferring electrons with electrode and being adsorbed (adatom) as the second step, and surface diffusion and incorporation in an energetically favorable site as the third step. The deposition sites can be punctual or extended surface defects, such as vacancies or kinks, known in general as nucleation sites. The nucleation sites allow the formation of nuclei (islands) that evolve to grains, forming compact deposits that grow on top of the surface of the electrode. Fig. 8A depicts the mechanism of layer growth and Fig. 8B shows an atomic force microscopy (AFM) image of an electrodeposited layer from an electrolyte containing cobalt sulfate. The image clearly reveals the formation of Co grains on top of electrode surface.

Electrode Reactions

The current that flows at the working electrode may be divided into two kinds: faradaic and non-faradaic. The faradaic processes are the ones where charges are transferred across the liquid–solid interface. These processes are called faradaic because they follow Faraday's law, which says the amount of substance that undergoes oxidation or reduction at each electrode is directly proportional to the amount of electricity that passes through the cell. Two faradaic processes that are directly related to electrodeposition are shown in Fig. 9, where Fig. 9A represents simply the deposition of Cu by reduction of Cu^{2+} and Fig. 9B the growth of Al anodic oxide by oxidation of metallic aluminum, this being an example of anodic electrodeposition. Non-faradaic processes are

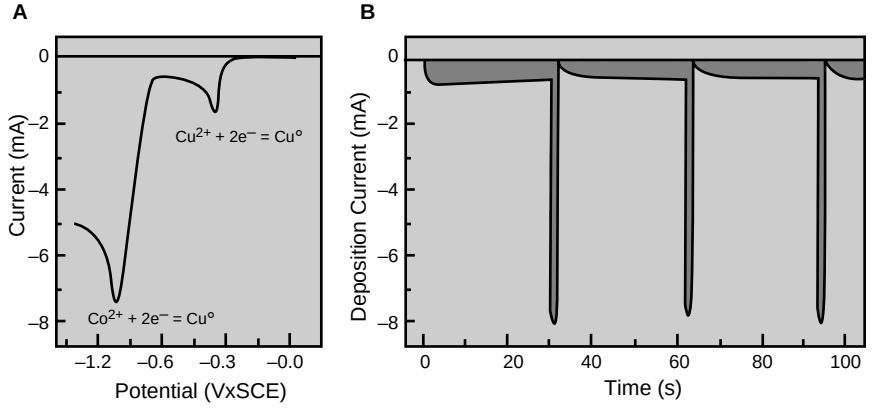


Fig. 11 (A) Voltammogram of an electrolyte containing two salts, Co and Cu sulfates and (B) pulses corresponding to the alternate deposition of a Cu/Co multilayer. The electrode is semiconducting Si. (Courtesy of L. Seligman, Universidade Federal de Santa Catarina).

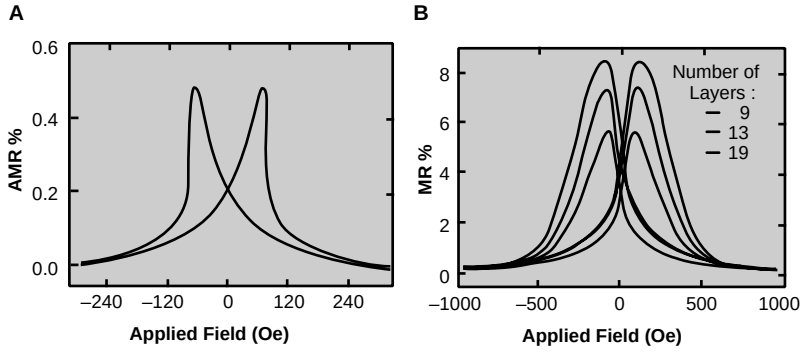


Fig. 12 (A) AMR of single layer of Co[1] and (B) GMR results depending on the number of layers. (From Ref.[5].) The substrate is semiconducting Si.

structural changes of the electrode–solution interface, such as absorption and desorption of species that change the potential of the electrode and solution composition without charge transfer.

ELECTRODEPOSITION IN MICROELECTRONICS

Recently, there has been a boom in the use of electrodeposition for microelectronics. The microelectronics industry came to the conclusion that the electrodeposition of Cu is the ideal manufacturing process for wiring for semiconductor logic and memory devices. Wiring is the network of wires that interconnects the devices (transistors) on integrated-circuit chips. Copper is a highly conductive metal and is relatively easy to electrodeposit. Since 1997, Cu has been successfully used for the production of interconnects.[2] Nowadays, such interconnects are electrodeposited in trenches with widths of 0.13 μm or less. The ability to fill trenches and vials with a plated material is called superfilling and is illustrated in Fig. 10A. The cross-section of the device illustrated in Fig. 10B depicts the multilevel wiring structure of an integrated circuit and Fig. 10C shows a real arrangement of electrodeposited interconnects in a device fabricated by IBM.

ELECTRODEPOSITION IN NANOSCIENCE AND NANOTECHNOLOGY

Simultaneously, with the rapid growth of electrodeposition in microelectronics, a new trend based on the electrodeposition of materials, structures, particles, devices, etc., generally called nano-objects, with dimensions below 100 nm commenced. Nano-objects are fundamental for nanoscience investigations and nanotechnology development. A nano-object is of particular interest if it has physical properties that differ from objects that have macroscopic sizes. Quantization of energy, for example, is observed in systems with greatly reduced size, such as atoms, molecules, and nanostructures.

Electrodeposition is an elegant and efficient technique for the production of nano-objects. Using the pulse deposition mode, it is possible to control the amount of atoms to be deposited with great precision. That is, pulsing with pulse durations of a few milliseconds to a few seconds, allows the deposition of clusters of atoms or layers with thickness of a few to hundreds of nanometers.

A typical example of an electrodeposited nanostructure is a multilayered structure. By having two salts in the electrolyte and applying two potentials in alternation, it is possible to deposit multilayer structures,

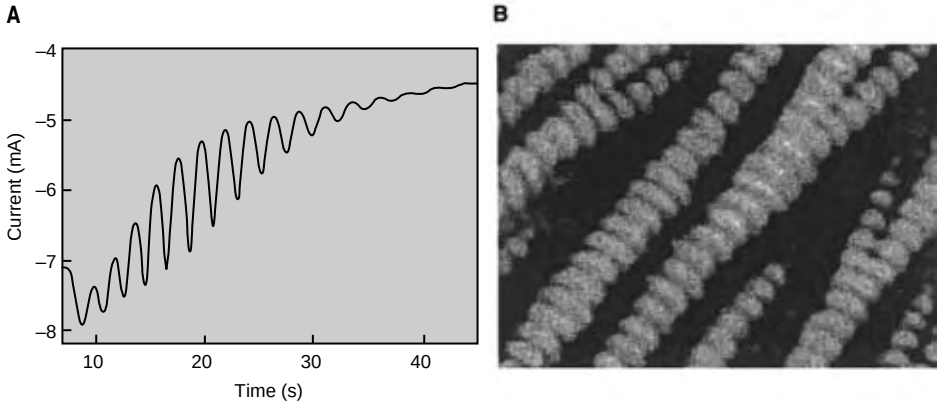


Fig. 13 (A) Spontaneous current oscillations during deposition at constant applied potential using an electrolyte containing Cu sulfate and lactic acid (Courtesy of R.G. Delatorre, Universidade Federal de Santa Catarina) and (B) SEM image of Cu/Cu_2O multilayered wires. (From Ref.[8].)

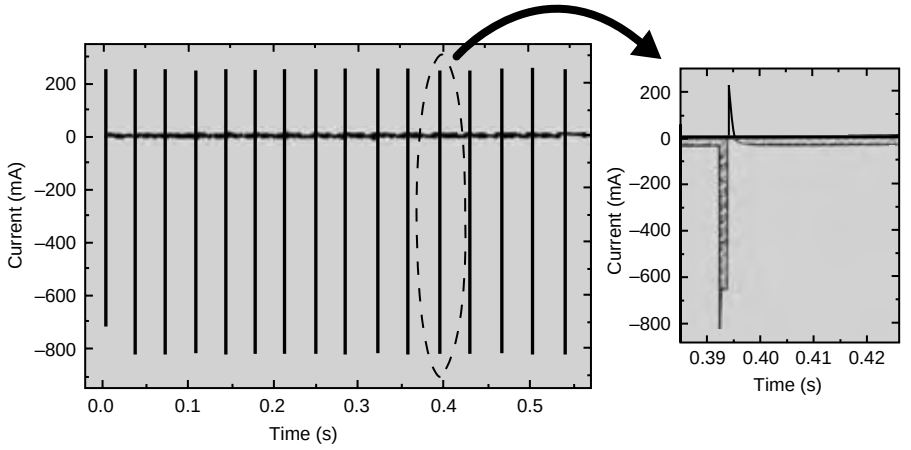


Fig. 14 Sequence of ultrafast pulses for the deposition of a nanostructured $Cu_{48}Ni_{52}$ alloy. (From Ref.[10].)

which are artificially fabricated materials that have application in the electronics industry. Fig. 11A shows the polarization curve for an electrolyte containing two salts, $CuSO_4$ and $CoSO_4$, where the reduction peaks of each metal are clearly seen. Pulsing the cathodic potential rapidly between a value at which only Cu is reduced, and one at which a Co-rich alloy is reduced, generates a deposit that is a Cu/Co multilayer with individual layers of nanometric thickness. The layer thickness may be controlled by integrating the current in real time and calculating the electrodeposited charge or, more simply, by controlling the deposition time. Fig. 11B shows typical current transients with characteristic peaks for each layer electrodeposited. The Co deposition current is much higher than the Cu one because to assure the deposition of a Co-rich alloy the concentration of Co sulfate in the electrolyte has to be much higher than the Cu sulfate concentration.

A multilayer structure with one of the repeating layers being a magnetic material is called magnetic multilayer. If the individual layers are only a few nanometers thick, which is easily achievable by electrodeposition, the electric resistance will vary with the magnetic field, an effect discovered very recently and known as giant-magnetoresistance (GMR).[3,4] The first observation of a magnetoresistive effect was by Lord Kelvin in 1857 by measuring the electrical resistance of magnetic alloys. Nowadays, the effect that he observed is called anisotropic magnetoresistance (AMR), and its magnitude does not exceed 6%. The GMR effect is about one order of magnitude greater than AMR and depends on many factors, such as the thickness and number of individual layers, the magnetic material used and the preparation method, and is observed also in non-layered granular structures. Magnetoresistive materials have been intensively used in the high-technology industry as magnetic sensors and reading heads for computer hard disk drives. Fig. 12A shows the magnetoresistance of a single layer of electrodeposited Co, similar to the one

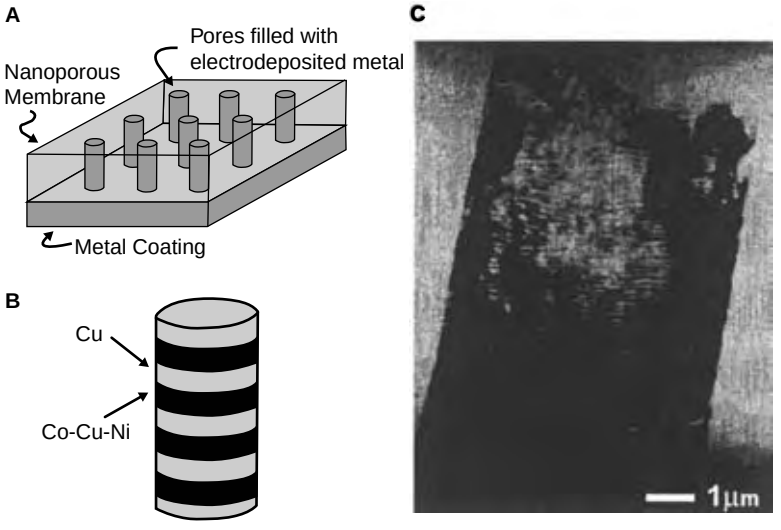


Fig. 15 (A) Alumina membrane with nanopores, (B) schematic view of a layered nanowire, and (C) TEM image revealing the layered structure of a Cu/CuCoNi nanowire grown in alumina nanopores. (From Ref.[11].)

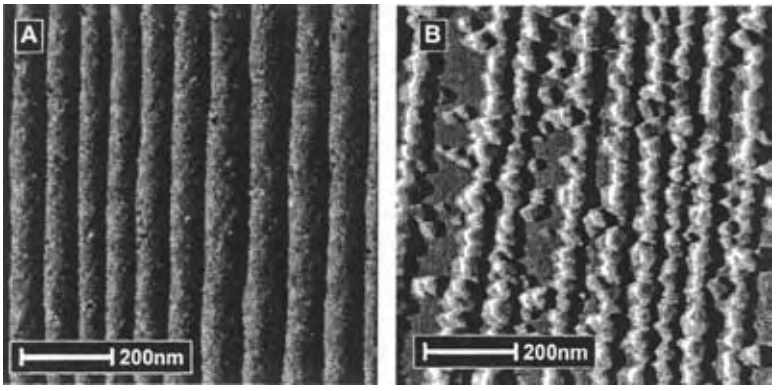


Fig. 16 (A) High-resolution AFM image of an atomically flat single crystal of Si showing large terraces and parallel steps and (B) nanowires of Au electrodeposited preferentially at the step edges. (From Ref.[12].)

depicted in Fig. 5, which shows an AMR effect of 0.5%. Fig. 12B illustrates the case of a magnetic multilayer, also electrodeposited, with the effect depending on the number of layers, reaching in this case a maximum of 8.5% for 15 Co/Cu layers.

An interesting achievement of electrodeposition in the preparation of nanostructures is the self-assembly of multilayers.[6] An electrolyte containing copper sulfate ($CuSO_4$) and lactic acid ($C_3H_6O_3$) is a standard example because under certain experimental conditions the cell current oscillates spontaneously leading to the growth of a nanometric Cu/Cu_2O multilayer. In Fig. 13A the spontaneous oscillations of the deposition current are illustrated, though a natural damping of the magnitude is observed. However, in stirred solutions the oscillatory behavior can be maintained for several days.[7] Fig. 13B shows a SEM image of filaments of spontaneously grown Cu/Cu_2O multilayers.[8]

The explanation for the spontaneous formation of multilayers lies in variations in the pH.[6,7,9] In the growth process, the electrodeposition of Cu_2O is favored, as it has an equilibrium potential more positive than that of Cu deposition (Table 1). However, the reaction $2Cu^{2+} + 2e^- + 2OH^- \leftrightarrow Cu_2O + H_2O$ depletes the OH^- species near the electrode, locally decreasing the pH and favoring the deposition of Cu. The above process is repeated, because the OH^- concentration is re-established during Cu^{2+} reduction.

Additionally, by having an experimental setup with a high-speed data acquisition system, it is possible to control deposition pulses with durations below milliseconds. This ultrafast pulsing method was called precision electrodeposition and allowed the deposition of sub-monolayer quantities of material.[10] Precision electrodeposition was demonstrated for the CuNi system, as shown in Fig. 14, where a sequence of ultrafast current pulses for the electrodeposition of a nanostructured CuNi alloy with a controlled composition of 48% Cu and 52% Ni is displayed. The duration of the pulses (tens of milliseconds), allows the deposition per cycle of 0.25 monolayer of Cu and 0.25 monolayer of Ni.

Nanowires

A special characteristic of electrodeposition is the fact the deposition occurs only where there is an electrical connection to the external circuit. This is a great advantage because it allows the deposition to be area selective. By covering the electrode surface with a patterned insulating layer, electrodeposition will occur only on the exposed areas. This makes electrodeposition an ideal method for growing materials on previously determined patterns and also for filling high-aspect ratio templates.

This advantage can be used for growing nanowires (wires with nanometric diameter). Nanoporous membranes that can be fabricated by the anodic oxidation of aluminum are appropriate templates. This process leads to the formation of an alumina layer with parallel nanopores, as shown in Fig. 15A, which can then be filled by electrodeposition. Fig. 15B shows a schematic view of a multilayer nanowire and Fig.15C a transmission electron microscopy image of a Cu/CuCoNi layered nanowire grown in the nanopores.

A different way to electrodeposit nanowires is by using the surface of a single crystal as a template. Fig. 16A shows an AFM image of a silicon surface, revealing large terraces with parallel steps. By electrodepositing Au at relatively low deposition rates, the steps act as deposition sites favoring the formation of wires of nanometric size along their edges, as shown in Fig. 16B.

CONCLUSIONS

Electrodeposition is a process widely used in industry. In this entry, emphasis was given to fundamental

aspects and to future potential applications of this technique.

ACKNOWLEDGMENTS

The authors wish to thank Prof. Walther Schwarzacher from Bristol University for reading carefully the manuscript and the Brazalian agencies CAPES, CNPQ, and FAPESC.

REFERENCES

1. Munford, M.L.; Sartorelli, M.L.; Seligman, L.; Pasa, A.A. Morphology and magnetic properties of Co thin films electrodeposited on Si. J. Electrochem. Soc. **2002**, *149* (5), C274–C279.
2. Andricacos, P.C. Copper on-chip interconnections: a breakthrough in electrodeposition to make better chips. Interface (Electrochem. Soc.) **1999**, *8*, 33–48.
3. Baibich, M.N.; Broto, J.M.; Fert, A.; Nguyen Van Dau, F.; Petroff, F.; Eitenne, P.; Creuzet, G.; Friederich, A.; Chazelas, J. Giant magnetoresistance of (001)Fe/(001)Cr magnetic superlattices. Phys. Rev. Lett. **1988**, *61*, 2472–2475.
4. Binasch, G.; Grünberg, P.; Saurenbach, F.; Zinn, W. Enhanced magnetoresistance in layered magnetic structures with antiferromagnetic interlayer exchange. Phys. Rev. B. **1989**, *39*, 4828–4830.
5. Seligman, L.; Sartorelli, M.L.; Pasa, A.A.; Schwarzacher, W.; Kasyutich, O.I. Co/Cu spin valves electrodeposited on Si. J. Magn. Magn. Mater. **2001**, *226*, 752–753.
6. Switzer, J.A.; Hung, C.J.; Huang, L.Y.; Miller, F.S.; Zhou, Y.C.; Raub, E.R.; Shumsky, M.G.; Bohannan, E.W. Potential oscillations during the electrochemical self-assembly of copper cuprous oxide layered nanostructures. J. Mater. Res. **1998**, *13* (4), 909–916.
7. Switzer, J.A.; Hung, C.-J.; Huang, L.-Y.; Switzer, E.R.; Kammler, D.R.; Golden, T.D.; Bohannan, E.W. Electrochemical self-assembly of copper/cuprous oxide layered nanostructures. J. Am. Chem. Soc. **1998**, *120*, 3530–3531.
8. Wang, M.; Zhong, S.; Yin, X.-B.; Zhu, J.-M.; Peng, R.-W.; Wang, Y.; Zhang, K.-Q.; Ming, N.-B. Nanostructured copper filaments in electrochemical deposition. Phys. Rev. Lett. **2001**, *86*, 3827–3830.
9. Zhang, M.-Z.; Wang, M.; Zhang, Z.; Zhu, J.-M.; Peng, R.-W.; Ming, N.-B. Periodic structures of randomly distributed Cu/Cu_2O nanograins and periodic variations of cell voltage in copper electrodeposition. Electrochim. Acta **2004**, *49* (14), 2379–2383.
10. Kazeminezhad, I.; Blythe, H.J.; Schwarzacher, W. Alloys by precision electrodeposition. Appl. Phys. Lett. **2001**, *78* (7), 1014–1016.
11. Schwarzacher, W. Metal nanostructures: a new class of electronic devices. Interface (Electrochem. Soc.) **1999**, *8*, 20–24.
12. Munford, M.L.; Maroun, F.; Cortes, R.; Allongue, P.; Pasa, A.A. Electrochemical growth of gold on well-defined vicinal H–Si(1 1 1) surfaces studied by AFM and XRD. Surf. Sci. **2003**, *537* (1–3), 95–112.

BIBLIOGRAPHY

Brad, A.J.; Faulkner, L.R. *Electrochemical Methods: Fundamentals and Applications*; Wiley: New York, 1980.

Schlesinger, M., Paunovic, M., Eds.; *Modern Electroplating*, 4th Ed.; Wiley: New York, 2000.

Electronic Chemical Sensors

Chung-Chiun Liu
Electronics Design Center and Department of Chemical Engineering, Case Western Reserve University, Cleveland, Ohio, U.S.A.

INTRODUCTION

Sensor is a device that recognizes an event or a change, and then converts this recognition through a transduction mechanism, resulting in an indication of the event or the change. Chemical sensors recognize the molecular species of an analyte, and transmit a signal as an indicator that can be related to the molecular species qualitatively and/or quantitatively. This transmitted signal or indication can be a color change, an electrical signal, such as current or potential, or optical output. Chemical sensors have found extensive applications in process control, environmental monitoring, personal safety and security, as well as biomedical applications, and many others. Research and development of a chemical sensor involves various disciplines and special skills, and is truly a multi-disciplinary scientific endeavor. Chemists, engineers of different background, and material scientists are often working together in chemical sensor development. In recent years, the advancement of silicon based microfabrication processing provides a new means for the development of chemical and biological sensors. The importance, and scientific and commercial potentials, of the microfabricated chemical and biochemical sensors will continue to be appreciated in the years to come. One should also realize that most of the biological or biomedical sensors operate on principles of chemical sensors with a bio-molecule recognition step. Therefore, the impact of chemical sensor development to biological and biomedical sensors is direct and relevant.

In this entry, we focus on the discussion of the platform technology for electrochemical sensors, metal oxide semiconductive (MOS) sensors, and piezoelectric based quartz crystal microbalance (QCM) sensors. There are other types of chemical sensors, such as optical sensors, Schottky diode based sensors, calorimetric sensors, field-effect transistor (FET) based sensors, surface acoustic wave sensors, etc. Information of these specific sensors can be found elsewhere[1,2] and in current journals on sensor technologies. Because of the increasing importance of microfabricated sensors, a brief discussion of microsensors is also given.

TERMINOLOGY AND BACKGROUND

Potential applications of chemical sensors are diverse and numerous, and the environment where the sensor is used varies. Therefore, chemical sensor often requires to be tailor-made or semi-tailor-made to meet the needs in the special circumstance. For example, sensing of oxygen in an automobile exhaust or in water or in blood can be accomplished by using an electrochemical-based sensor. However, the selection of electrolyte and a diffusion-limited layer or protective membrane will be different in each case. Therefore, the platform chemical sensor technology is discussed in general. Special applications of a chemical sensor under a particular circumstance need to be addressed separately.

Differing from a typical analytical or sensing instrument, chemical sensors are relatively small in size, mobile or portable. If possible, chemical sensor should require minimum or no additional reagents, and also minimum sample transportation or preparation. For many applications, the possibility of producing the sensors at modest cost, and making disposable sensor a reality, is extremely attractive. This is particularly true in many biological and medical sensing applications.

The essential properties of a chemical sensor include the sensitivity, selectivity, response time, long-term stability, and others. Sensitivity defines the detecting range of the sensor, and it can be in the needs of parts per billion (ppb) or in high concentration percentage. Because the sensor is often employed in a complex environment, potential interference by the other species presented can lead to error in quantifying the analyte. The need to have a reasonable response time for a sensor can vary from a fraction of a second to minutes depending on the applications. The long-term stability is also an important consideration for a chemical sensor to have practical application. For example, the common carbon monoxide sensor for home safety should be operable for a year, if not years, with good long-term stability. There are properties of the sensor, such as temperature dependence, operable temperature range, and others, which can be crucial. Obviously, a specific application of a chemical sensor defines the requirements of its sensitivity, selectivity, response time, and others. Consequently, this leads to adaptation or

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007698

modification of a specific platform technology for the chemical sensor development.

ELECTROCHEMICAL SENSORS

An electrochemical sensor is generally an electrochemical cell containing two electrodes, an anode and a cathode, and an electrolyte. Electrochemical sensors in general are classified, based on the mode of its operation, and they are conductivity sensors potentiometric sensors, and voltammetric sensors. Amperometric sensors can be considered as a special type of voltammetric sensors. The fundamentals of these sensors' operational principles are described exceptionally well in several excellent electroanalytical books.[3–5] In this entry, only the essential features are included.

Conductivity Sensor

In an electrochemical cell, the conductivity is inversely related to the resistance in the electrolyte/test medium. The presence of certain chemical or ionic species may affect the resistance of the electrochemical cell. This change in resistance or conductivity can then be used to quantify the amount of the analyte presented. Molar and equivalent conductivities are commonly used to express the conductivity in an electrochemical cell. The conductivity measurement can be made relatively straightforward, using a DC mode or with a potential or current excitation. However, any faradaic or change transfer process occurring at the electrode surface will affect the conductivity measurement in an electrochemical cell. Furthermore, conductivity measurement in general does not provide sufficient specificity or sensitivity to quantify the analyte. This limits the use of conductivity of an electrochemical cell for sensing applications.

Potentiometric Sensor

Potentiometric sensor is based on a redox reaction that occurs at the electrode–electrolyte interface in an electrochemical cell. If a redox reaction $Ox + Ze \rightarrow Red$ takes place at an electrode surface, it is called a half-cell reaction. In the above reaction, Ox is the oxidant, Red is the reduced product, e in the electron, and Z is the number of electrons transferred in the reaction. At thermodynamic quasiequilibrium conditions, the Nernst equation is applicable and can be expressed as:

$$E = E^{\circ} + \frac{RT}{FZ} \ln \frac{[a_{ox}]}{[a_{red}]}$$

where E and E° are the electrode potential and electropotential at standard state, respectively, a_{ox} and a_{red} are the activities of Ox and Red, respectively, F is the Faraday constant, R is the gas constant, and T is the operating temperature in absolute scale. The Nernst equation shows that the electrode potential E can be used to quantify the ratio of the activities of oxidant and the reduced product. The equation also suggests that at ambient temperature, 298°K (25°C), and $Z = 1$, the slope of the linear relationship of E versus the $\ln([a_{ox}]/[a_{red}])$ is predictable with a value of 59.8 mV.

In a potentiometric sensor, two half-cell reactions will take place simultaneously. However, only one of the two half-cell reactions should involve the sensing species of interest, whereas the other half-cell reaction should be preferably reversible and non-interference. Thus, a reference electrode, such as Ag/AgCl, is often used for the other half-cell electrode. The redox reaction for the reference electrode is highly reversible in an appropriate sensing environment and does not contribute interference to the sensing half-cell reaction. The sensing electrode of a potentiometric sensor can be inert or active. An inert electrode provides a surface for the electron transfer or a catalytic surface for the reaction, and the electrode does not actively engage in the half-cell reaction. On the other hand, an active electrode can be either an ion donor or an acceptor in the sensing reaction. There are three common types of active electrodes and they are the metal/metal ion, the metal/insoluble metallic salt or oxide, and the metal/metal chelate electrode.

Potentiometric sensors using inert electrode have been applied as pH sensors and electrochemical gas sensors including oxygen sensors in liquid and gas phases. Noble metals, such as platinum and gold, graphite, and glassy carbon have been used. In the potentiometric oxygen sensor, the sensor operates as a concentration cell. In this cell, one half-cell is in contact with the test medium, which the oxygen concentration is to be sensed, and the other is exposed to a known oxygen concentration reference gaseous mixture. Potentiometric sensors with active electrodes include ion-selective electrode type sensors. The active electrode often incorporates a specific membrane that may have ion-selective, ion-permeable, or ion exchange properties. These membranes minimize possible interference from other ions and provide the specificity for the desired sensing ions.

Measurement of the cell potential of a potentiometric sensor should be carried out under zero-current or quasiequilibrium conditions. A high input impedance electrometer is commonly used. The advantage of potentiometric sensor is that the sensor output and the cell potential do not depend on the electrode surface area. This can make the manufacturing of practical sensors simpler. A major disadvantage of

potentiometric sensor is its limited sensitivity. The Nernst equation shows a linear relationship between the cell potential and the logarithmic value of the activity of sensing species. Thus, for practical application of any potentiometric sensor, its advantages and limitation should be carefully assessed.

Voltammetric Sensor

A voltammetric sensor is characterized by the current and potential relationship of an electrochemical cell. Voltammetric sensor utilizes the concentration effect on the current-potential relationship. This relationship depends on the rate by which the reactant (commonly the sensing species) is brought to the electrode surface (mass transfer) and the kinetics of the faradaic or charge transfer reaction at the electrode surface. In an electrochemical reaction, the interdependence between the reaction kinetics and the mass transfer processes establishes the concentration of the sensing species at the electrode surface relative to its bulk concentration and, hence, the rate of the faradaic process. This provides a basis for the operation of the voltammetric sensor.

Unfortunately, simultaneous analytical solution of the mass transfer and kinetic equations of an electrochemical cell is usually complex. Thus, the cell is usually operated with definitive hydrodynamic characteristics. Operational techniques, relating to controlling either the potential or the current, have been developed to simplify the analysis of the electrochemical cell. Description of these operational techniques and their corresponding mathematical analyses are well discussed elsewhere.[6,7]

For a voltammetric sensor, the current or potential peak shift that may relate to the concentration of the sensing species is an important measurement. In a dynamic situation in which polarization characteristics are obtained, it is essential that the mass transfer characteristics are reproducible for both calibration and actual measurements. In the case of a stationary planar sensor, stagnant solution or steady flow conditions in a flow cell provides good reproducibility. Or in another case, a sufficiently high concentration of an electrolyte is used to maintain a constant ohmic drop in the cell, regardless of the concentration of the pertinent sensing component. Under these conditions, the mass transfer can be purely diffusional and adequately described by Fick's law of diffusion.

The cell current, a measure of the rate of the faradaic process at an electrode, usually increases with an increase in the electrode potential. The cell current approaches a limiting value, when the rate of the faradaic process at the electrode surface equals the maximum possible mass transfer rate. Consequently, the cell is defined as mass transfer limited, and the cell current is defined as diffusional limited current, which can be used to quantify the sensing species presented. However, in many cases, the cell current does not tend to a limiting value with an increase in the electrode potential. This is because as the electrode potential increases, other faradaic or nonfaradaic processes may become active, and the cell current represents the cumulative rates of all the active electrochemical cell is complex, and the limiting current approach becomes ineffective. The increase in the electrode potential may accelerate the cell reaction to an extent that it would lead to a substantial decrease in the bulk concentration of the sensing ions. Under such conditions, the current will reach a maximum and then decrease with further increase of the potential. In this case, the sensor should not be operated under limiting current condition.

Amperometric sensors are considered to be a sub-class of voltammetric modes. In this case, the current-sensing species concentration relationship is obtained at a fixed electrode potential. At this fixed potential, the cell current can then be used to quantify the amount of the sensing species presented. In addition to amperometric mode of operation, other modes of voltammetric sensor operation include linear sweep and cyclic voltammetry. Linear sweep voltammetry involves increasing the imposed potential linearly at a constant scanning rate from an initial pre-set value to a given upper limited potential. Cyclic voltammetry is similar to the linear sweep potential with the exception that the electrode potential is returned to its initial value at the same scanning rated. The cathodic and anodic sweeps normally produce two current peaks, corresponding to the redox couple in the simplest form of a voltammogram. However, the characteristics of a voltammogram can be complex. The effects of the scanning rate, diffusional properties of the sensing species, operating temperatures can also contribute to the voltammogram and its interpretation, and consequently its assessment of the quantity of the sensing species.

Voltammetric sensors, including amperometric sensors, are very effective and applicable in many sensing circumstances, if it is properly applied. It is recognized that voltammetric sensor is an important sensing element in electrochemical sensors. With appropriate assessment of the needs, platform electrochemical sensor technologies, including voltammetric sensors, can be used to meet our specific needs in sensing technology.

METAL OXIDE SEMICONDUCTIVE SENSORS

Metal oxide semiconductive sensors are an important class of chemical sensors particularly for gaseous sensing. Among the metal oxide semiconductive materials,

tin oxide is most widely studied, and commercial sensors based on tin oxide have been available over the years. Metal oxides are general *n*-type semiconductive materials, and they can be used as chemical and gaseous sensors based on the change of the electronic conductivity, when the oxides are exposed to the gases. The intrinsic conductivity comes from the nonstoichiometric composition of the metal oxide. The oxygen vacancies act as donor defect. Oxygen may be either physisorbed or chemisorbed as charged species on the metal oxide surface. The chemisorbed oxygen may interact with the physisorbed oxygen or interact directly with gaseous oxygen. The chemisorbed oxygen ionic species are dioxygen O_2^- and mono-oxygen, O^-. These ionic species are presented at room temperature. With increasing temperatures, the O_2^- can dissociate from O^- and O^- becomes the dominant species at higher temperature. It is also recognized that O^- is more reactive than O_2^-. When a reducing gas, such as H_2 or CO, reacts with a metal oxide semiconductive material, the O^- and the O_2^- will react with the reducing gas releasing the electrons from the O^- and O_2^-. Consequently, the conductivity of the metal oxide will increase. The change in conductivity can then be used to quantify the reducing gas presented.

The seminal research by Taguchi[8] and Seiyama et al.[9] led to the successful commercialization of tin oxide based gas sensors. The tin oxide can be sintered, forming a disk, on which electrical conductivity metal contacts can be formed. The MOS can also be formed by other techniques, such as sputtering or sol–gel techniques.

Different electron conduction mechanisms of a MOS have been proposed including neck controlled (bulk trap and surface trap limited) conduction, and barrier controlled (barrier limited) conduction.[10] A better understanding of the conduction mechanisms will be helpful in developing more effective MOS based sensors.

Metal oxide semiconductive materials exhibit a relatively low conductivity at ambient temperature. Thus, it will be very difficult to observe a small conductivity change because of its reaction with a reducing gas. Therefore, it is common to operate a MOS based sensor at elevated temperature. At higher temperature, the conductivity of the MOS increases substantially, and a change caused by the reaction with reducing gas is now observable.

Metal oxide semiconductive sensors are not limited to tin oxide only. Many other metal oxides, such as zinc oxide, tungsten oxide, and others can also be used for chemical and gas sensing. It is understandable that an incorporation of a selective catalyst or a dopant may enhance the selectivity of the MOS sensors. Palladium, platinum, and others have been used as catalytic dopants for these sensors. The processes of incorporating the catalysts and dopants are many. For instance, the selected dopant may be in the form of a precursor adding into the sol–gel solution of the MOS material.

It is recognized that the particle size and the size distribution of the particles of MOS materials can affect directly to the sensor performance. It is also suggested that the grain boundaries of the particles provide the sites for the reaction between the sensing gas and the metal oxide. Therefore, a MOS film that can provide large number of grain boundary will be more sensitive. Hence, the possibility of using nanocrystalline MOS materials as sensing films offers opportunity of developing higher sensitivity MOS sensors.

QUARTZ CRYSTAL MICROBALANCE

Among piezoelectric based sensors, QCM represents a major device. A quartz crystal with a noncentrosymmetric space group will have a dipole associated with the orientation of the atoms. When under stress, the crystal exhibits a charge separation because of the displacement of its atoms. The converse effect has also been proven. It has been demonstrated that an applied alternating electric field will cause a vibration in a quartz crystal that in turn results in the generation of acoustic standing waves. Also, the crystal shows a tendency to vibrate at a characteristic resonant frequency.

Also, it has been observed that a mass adsorption at the surface of a quartz crystal will lead to a linear shift of the resonant frequency. Sauerbrey[11] provides a theoretical basis of this phenomenon, and this results in using quartz crystal as a microbalance. It shows a high degree of sensitivity and can detect a very small quantity of the adsorped mass. In more recent years, special coatings are applied to the surface of the quartz crystal. The coating will specifically and selectively adsorb a species resulting in the increase in mass on the QCM, serving as a sensing mechanism for the adsorbed species. This approach has been proposed by King,[12] and much research has also been directed to develop various organic and inorganic coatings for this application.

Practical QCM devices are AT-cut alpha quartz crystals that have excellent mechanical and piezoelectric properties. The AT-cut that is cut at a 35°15 angle from the *Z*-axis, is commonly used because of its minimal temperature effect.

Quartz crystal microbalance is operated in typical resonant frequencies ranging from 1 to 10 MHz, with most of them operating in 5–10 MHz. A typical QCM is a disk in the size of 10–16 mm in diameter with a thickness of approximately 0.15 mm. A thin metal film, gold, aluminum, or others, is deposited onto the surface of the quartz serving as electrodes. The metal

film electrode is approximately 3–10 mm in diameter with a thickness of about 1000–3000 Å.

In the use of a QCM as a sensor, oscillator circuits are the most common electrical interface. A frequency counter can then measure the frequency output from the oscillator, which is identical to the resonant frequency of the quartz crystal. There are two general classifications for the oscillator circuits operation in series resonance and operation in parallel resonance. The performance of these circuits and the choice of the proper circuit are discussed extensively elsewhere.

Quartz crystal micro balance shows promise in its applications for chemical sensing. A proper and selective coating on the surface of a QCM can be used to selectively absorb a chemical species. The mass of the absorbed species, even in very minute quantity, can be used to quantify the species present. It is an effective means for chemical sensing under appropriate circumstance.

CONCLUSIONS

As mentioned, in addition to these described technologies for chemical sensor, there are other basic technologies that can be used for chemical sensing. It would not be feasible to describe each of these technologies, and researchers are encouraged to first carefully assess the needs of the desired chemical sensor, and then choose the potentially applicable platform technology.

In recent years, silicon based microfabrication and micromachining technology provide a new venue for the manufacturing of micro sized chemical sensors. This leads to new applications of chemical sensors in various scientific and commercial endeavors. Details of the microfabrication technologies are given elsewhere in this encyclopedia and references in microelectronic processing. Micro fabricated chemical sensors can be produced in geometrically well-defined highly uniform structure and at a modest cost. This can be important, for instance, for amperometric sensors that require uniform and reproducible electrode surface structure. The modest cost permits the sensor to be disposable. Micro chemical sensors can also be fabricated into an array, making multiple sensing feasible. Micro chemical sensor arrays will find increasing usage in various fields in the years to come.

REFERENCES

1. Sze, S.M. *Semiconductor Sensors*; Wiley-Interscience: New York, 1994.
2. Fraden, J. *AIP Handbook of Modern Sensors*; American Institute of Physics: New York, 1993.
3. Bard, A.J.; Faulkner, L. *Electrochemical Methods*; Wiley: New York, 1980.
4. Kissinger, P.; Heineman, W. *Laboratory Techniques in Electroanalytical Chemistry*, 2nd Ed.; Dekker: New York, 1996.
5. Wang, J. *Analytical Electrochemistry*, 2nd Ed.; Wiley-VCH: New York, 2000.
6. Newmann, J.S. *Electrochemical Systems*; Prentice Hall: New York, 1972.
7. Conway, B.E. *Theory and Principles of Electrode Processes*; Ronald Press, 1965.
8. Taguchi, N. Gas detection element and method of making it. U.S. Patent 3,644,795, Feb 22, 1972.
9. Seiyama, T.; Kato, A.; Fujiishi, K.; Nagatani, A. A new detector for gaseous components using semiconductive thin film. Anal. Chem. **1962**, *34*, 1502.
10. Williams, D.E. *Solid State Gas Sensors*; Moseley, P.T., Tofield, D.C., Eds.; Adam Higher: Philadelphia, 1987; 70–123.
11. Sauerbrey, G.Z. The use of quartz oscillators for weighing thin layers and for microweighing. Z. Phys. **1959**, *155*, 206.
12. King, W.H. Piezoelectric sorption detector. Anal. Chem. **1964**, *36*, 1735.

Electroplating

Helen H. Lou
Department of Chemical Engineering, Lamar University, Beaumont, Texas, U.S.A.

Yinlun Huang
Department of Chemical Engineering and Materials Science, Wayne State University, Detroit, Michigan, U.S.A.

INTRODUCTION

Electroplating is an electrodeposition process for producing a dense, uniform, and adherent coating, usually of metal or alloys, upon a surface by the act of electric current.[1] The coating produced is usually for decorative and/or protective purposes, or enhancing specific properties of the surface. The surface can be conductors, such as metal, or nonconductors, such as plastics. Electroplating products are widely used for many industries, such as automobile, ship, air space, machinery, electronics, jewelry, defense, and toy industries. The core part of the electroplating process is the electrolytic cell (electroplating unit). In the electrolytic cell (electroplating unit) a current is passed through a bath containing electrolyte, the anode, and the cathode. In industrial production, pretreatment and posttreatment steps are usually needed as well.

BACKGROUND

The workpiece to be plated is the cathode (negative terminal). The anode, however, can be one of the two types: sacrificial anode (dissolvable anode) and permanent anode (inert anode).[2] The sacrificial anodes are made of the metal that is to be deposited. The permanent anodes can only complete the electrical circuit, but cannot provide a source of fresh metal to replace what has been removed from the solution by deposition at the cathode. Platinum and carbon are usually used as inert anodes.

Electrolyte is the electrical conductor in which current is carried by ions rather than by free electrons (as in a metal). Electrolyte completes an electric circuit between two electrodes. Upon application of electric current, the positive ions in the electrolyte will move toward the cathode and the negatively charged ions toward the anode. This migration of ions through the electrolyte constitutes the electric current in that part of the circuit. The migration of electrons into the anode through the wiring and an electric generator and then back to the cathode constitutes the current in the external circuit. The metallic ions of the salt in the electrolyte carry a positive charge and are thus attracted to the cathode. When they reach the negatively charged workpiece, it provides electrons to reduce those positively charged ions to metallic form, and then the metal atoms will be deposited onto the surface of the negatively charged workpiece.

Fig. 1 illustrates a typical plating unit for plating copper from a solution of the metal salt copper sulfate ($CuSO_4$). The cathode, which is the workpiece to be plated, is charged negatively. Some of the electrons from the cathode bar transfer to the positively charged copper ions (Cu^{2+}), setting them free as atoms of copper metal. These copper atoms take their place on the cathode surface and copper plate it. Concurrently, the same number of sulfate ions $SO_4^{2-}$is discharged on the copper anodes, thereby completing the electrical circuit. In so doing, they form a new quantity of copper sulfate that dissolves in the solution and restores it to its original composition. This procedure is typical of ordinary electroplating processes with sacrificial anodes; the current deposits a given amount of metal on the cathode and the anode dissolves to the same extent (of the same electrical charge), maintaining the solution more or less uniformly.

ELECTROCHEMISTRY FUNDAMENTALS

When a direct electric current passes through an electrolyte, chemical reactions take place at the contacts between the circuit and the solution. This process is called electrolysis. Electrolysis takes place in an electrolytic cell. Electroplating is one specific type of electrolysis. Besides electroplating, electrolysis has also been widely used for preparation of halogens and notably chlorine, and refining of metals, such as copper and zinc. Understanding the electrochemical principles of electrodeposition is essential to the development of electroplating technologies. Some basic concepts are presented below.[3]

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007747

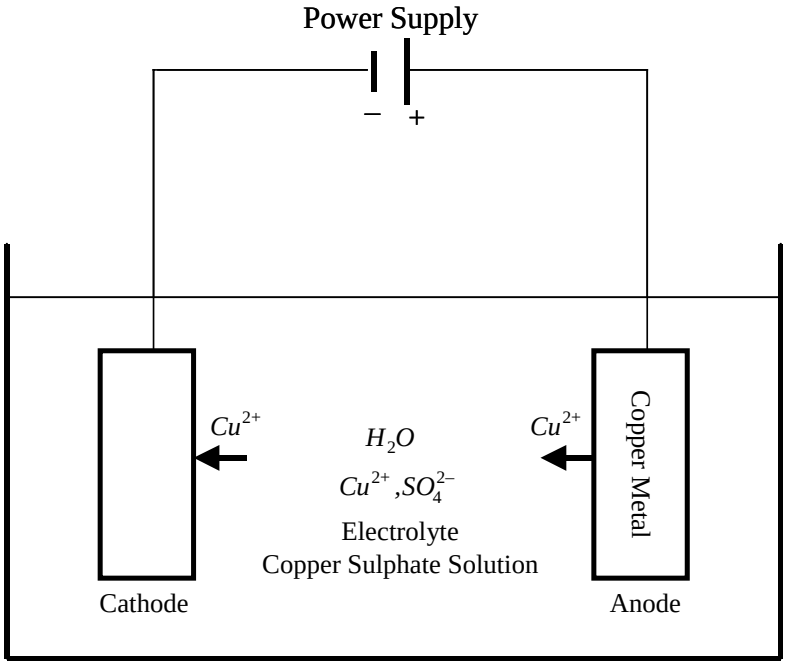


Fig. 1 Principle of electroplating.

Oxidation/Reduction

In a wider sense, all electron-transfer reactions are considered oxidation/reduction. The substance gaining electrons (oxidizing agent, or oxidant) oxidizes the substance that is losing electrons (reducing agent, or reductant). In the process, the oxidizing agent is itself reduced by the reducing agent. Consequently, the reduction process is sometimes called electronation, and the oxidation process is called ''de-electronation.'' Because a cathode is attached to the negative pole of the electric source, it supplies electrons to the electrolyte. On the contrary, an anode is connected to the positive pole of the electric source; therefore, it accepts electrons from the electrolyte. Various reactions take place at the electrodes during electrolysis. In general, reduction takes place at the cathode, and oxidation takes place at the anode.

Anode and Cathode Reactions

Electrodeposition or electrochemical deposition (of metals or alloys) involves the reduction of metal ions from electrolytes. At the cathode, electrons are supplied to cations, which migrate to the anode. In its simplest form, the reaction in aqueous medium at the cathode follows the equation:

$$M^{n+} + ne^- \rightarrow M \quad (1)$$

with a corresponding anode reaction. At the anode, electrons are supplied to the anions, which migrate to the anode. The anode material can be either a sacrificial anode or an inert anode. For the sacrificial anode, the anode reaction is:

$$M \rightarrow M^{n+} + ne^- \quad (2)$$

In this case, the electrode reaction is electrodissolution that continuously supplies the metal ions.

Faraday's Laws of Electrolysis

In 1833, the English scientist, Michael Faraday, developed Faraday's laws of electrolysis. Faraday's first law of electrolysis and Faraday's second law of electrolysis state that the amount of a material deposited on an electrode is proportional to the amount of electricity used. The amount of different substances liberated by a given quantity of electricity is proportional to their electrochemical equivalent (or chemical equivalent weight).

In the SI system, the unit quantity of electricity charge and the unit of electric charge are coulomb (C); one coulomb is equivalent to one ampere flowing for one second ($1\,C = 1\,A \cdot sec$). The electrochemical equivalent of an element is its atomic weight divided by the valence change involved in the reaction. For example, for the reaction, $Fe^{2+} \rightarrow Fe^0$, the valence change is 2, and the electrochemical equivalent of iron is $55.85/2 = 27.925$ in this reaction. Depending on the specific reaction, one element may have different equivalent weights, although it has only one atomic weight.

In detail, to reduce one mole of a given metal from a metal ion with the valence charge of $n+$, n moles of electrons are required. That is, the total cathodic charge used in the deposition, Q(C), is the product of the number of gram moles of the metal deposited, m, the number of electrons taking part in the reduction, n, Avogadro's number, N_a (the number of atoms in a mole), and the electrical charge per electron, Q_e(C). Thus, the following equation gives the charge required to reduce m moles of metal:

$$Q = mnN_aQ_e \tag{3}$$

The product of the last two terms in Eq. (3) is the Faraday constant, F. Therefore, the number of moles of the metal reduced by charge Q can be obtained as:

$$m = \frac{Q}{nF} \tag{4}$$

The Faraday constant represents the amount of electric charge carried by 1 mol, or the Avogadro's number of electrons. The Faraday constant can be derived by dividing Avogadro's number, or the number of electrons per mole, by the number of electrons per coulomb. The former is approximately equal to 6.02×10^{23} and the latter is approximately 6.24×10^{18}. Therefore,

$$F = \frac{(6.02 \times 10^{23})}{(6.24 \times 10^{18})} = 9.65 \times 10^4\,\mathrm{C/mol} \tag{5}$$

On the other hand, the total charge used in the deposition can be obtained as the product of the current, I(A), and the time of deposition, t(sec), if the deposition current is held constant. Or, if the current varies during the deposition,

$$Q = \int I\,\mathrm{d}t \tag{6}$$

The weight of the deposit, W(g), thus can be obtained by multiplying the number of moles of metal reduced with the atomic weight, M_w, of the deposited metal:

$$W = \frac{M_w}{nF}\int I\,\mathrm{d}t \tag{7}$$

Ideally, the deposition thickness, δ (cm), can be solved by:

$$\delta = \frac{W}{\rho A} = \frac{M_w}{nF\rho A}\int I\,\mathrm{d}t \tag{8}$$

where ρ is the density of the metal (g/cm^3) and A is the area of deposition (cm^2).

Current Efficiency, Current Density, and Current Distribution

Faraday's laws give theoretical prediction of electrodeposition in an ideal situation. In a real application, many factors influence the coating quantity and quality.[4]

Current efficiency

It is stated in Faraday's laws that the amount of chemical charge at an electrode is exactly proportional to the total quantity of electricity passing. However, if several reactions take place simultaneously at the electrode, side reactions may consume the product. Therefore, inefficiencies may arise from the side reactions other than the intended reaction taking place at the electrodes. Current efficiency is a fraction, usually expressed as a percentage, of the current passing through an electrolytic cell (or an electrode) that accomplishes the desired chemical reaction. Or,

$$\text{Current efficiency} = 100 \times W_{Act}/W_{Theo} \tag{9}$$

where W_{Act} is the weight of metal deposited or dissolved, and W_{Theo} is the corresponding weight to be expected from Faraday's laws [Eq. (7)] if there is no side reaction. Note that the cathode efficiency is the current efficiency applied to the cathode reaction, and the anode efficiency is the current efficiency applied to the anode reaction.

Current density

Current density is defined as current in amperes per unit area of the electrode. It is a very important variable in electroplating operations. It affects the character of the deposit and its distribution.

Current distribution

The local current density on an electrode is a function of the position on the electrode surface. The current distribution over an electrode surface is complicated. Current will tend to concentrate at edges and points, and unless the resistance of the solution is very low, it will flow to the workpieces near the opposite electrode more readily than to the more distant workpieces. It is desired to operate processes with uniform current distribution. That is, the current density is the same at all points on the electrode surface.

Potential Relationships

In electroplating, sufficient voltage should be provided by the power source. The voltage–current relationship

follows Ohm's law. The concepts of electrode potentials, equilibrium electrode potential, overpotential, and overvoltage are of fundamental importance.

The voltage–current relationship: Ohm's law

The current is driven by a potential difference, or voltage through the conducting medium, either electrolytic or metallic. The voltage necessary to force a given current through a conductor is given by Ohm's law:

$$E = IR \tag{10}$$

where E is the e.m.f (electromotive force) volts (V) and R the resistance of the conductor (Ω).

Electrode potentials

The electrode potential is the electrical potential difference between an electrode and a reference electrode. The absolute potential of an electrode is not directly measurable. Therefore, the electrode potential must always be referred to an arbitrary zero point that is defined by the potential of the reference electrode.

Equilibrium electrode potential

When a metal is immersed into a solution containing ions of that metal, equilibrium is set up between the tendency of the metal to enter solution as ions and the opposing tendency of the ions to lose their charge and deposit on or in the metal.

$$\mathrm{M} \leftrightarrow \mathrm{M}^{n+} + n\mathrm{e}^{-} \tag{11}$$

Depending on the conditions of the system, this can occur in either direction. At equilibrium, the driving forces for metal ions being discharged and metal atoms being ionized are equal. The potential difference between the metal and the solution phases under these conditions is the equilibrium potential difference.

The equilibrium electrode potential is the electrical potential of an electrode measured against a reference electrode when there is no current flowing through the electrode. It is also called open circuit potential (OCP). The equilibrium potential between a metal and a solution of its ions is given by the Nernst equation as follows:

$$E = E^0 + \frac{RT}{nF} \ln a \tag{12}$$

where E^0 is the standard electrode potential, which is a constant characteristic of the material of the electrode; R the gas constant ($8.3143\,\mathrm{J/k \cdot mol}$); T the absolute temperature (K); F the Faraday constant; n the valence change; a the activity of the metal ion. In approximation, the concentration of the metal ion can be used instead of the activity.

If numerical values are substituted for R and F, and T is at 25°C (298 K), and base 10 logarithm is used instead of base e, the Nernst equation can be expressed as:

$$E = E^0 + \frac{0.059}{n} \log a \tag{13}$$

In the above equation, if $a = 1$, then $E = E^0$. The standard potential of an electrode E^0 is the potential of an electrode in contact with a solution of its ions of unit activity. The standard potentials are always expressed against the standard hydrogen electrode (SHE), the potential of which is zero by definition. The standard potentials are a function of temperature; they are usually tabulated for 25°C. Standard electrode potential is also called normal electrode potential.

Overpotential and overvoltage

The equilibrium is dynamic with metal ions being discharged and metal atoms being ionized, but these two effects cancel each other and there is no net change in the system. For the realization of metal deposition at the cathode and metal dissolution at the anode, the system must be moved away from the equilibrium condition. An external potential must be provided for the useful electrode reactions to take place at a practical rate; this external potential may have several causes.

Overpotential is the difference in the electrode potential of an electrode between its equilibrium potential and its operating potential when a current is flowing. The overpotential represents the extra energy needed to force the electrode reaction to proceed at a required rate (or its equivalent current density). Consequently, the operating potential of an anode is always more positive than its equilibrium potential, while the operating potential of a cathode is always more negative than its equilibrium potential. The overpotential increases with increasing current density. The value of the overpotential also depends on the inherent speed of the electrode reaction. A slow reaction (with small exchange current density) will require a larger overpotential for a given current density than a fast reaction (with large exchange current density). Overpotential is also referred to as polarization of the electrode.

An electrode reaction always occurs in more than one elementary step, and there is an overpotential associated with each step. Even for the simplest case, the overpotential is the sum of the concentration overpotential and the activation overpotential.

Overvoltage is the difference between the cell voltage (with a current flowing) and the open-circuit voltage (OCV). The overvoltage represents the extra energy needed to force the cell reaction to proceed at a required rate. Consequently, the cell voltage of an electrolytic cell is always more than its OCV, while the cell voltage of a galvanic cell (e.g., a rechargeable battery during discharging) is always less than its OCV. Occasionally, it is also referred to as polarization of the cell.

The overvoltage is the sum of the overpotentials of the two electrodes of the cell and the ohmic loss of the cell. Unfortunately, the terms overvoltage and overpotential are sometimes used interchangeably.

SURFACE PREPARATION

Workpieces to be plated may be put through a variety of pretreating processes, including surface cleaning, surface modification, and rinsing.[4] A schematic flowsheet of a typical electroplating plant, including surface treatment and waste treatment, is depicted in Fig. 2.

The purpose of surface pretreatment is to remove contaminants, such as dust and films, from the substrate surface. The surface contamination can be extrinsic, composed of organic debris and mineral dust from the environment or preceding processes. It can also be intrinsic, such as a native oxide layer. Contaminants and films interfere with bonding, which can cause poor adhesion and even prevent deposition. Therefore, surface pretreatment is important to ensure plating quality. Most (metal) surface treatment operations have three basic steps: surface cleaning, surface treatment, and rinsing.

Surface Cleaning

Cleaning methods should be able to minimize substrate damage while removing the contaminants, dust, film, and/or debris. Cleaning processes are based on two approaches: chemical approach and mechanical approach.

Chemical approaches

A chemical approach usually includes solvent degreasing, alkaline cleaning, (soak cleaning), and acid cleaning (acid pickling).

Solvent Degreasing. Contaminants consist of oils and grease of various types, waxes, and miscellaneous organic materials. These contaminants can be removed by appropriate organic solvents, either by dipping the workpieces in the solvent or by vapor decreasing.

Alkaline Cleaning. Workpieces are immersed in tanks of hot alkaline cleaning solutions to remove dirt and solid soil. A special type of alkaline cleaning is electrocleaning. In electrocleaning, the workpiece can be either the cathode (namely direct cleaning) or the anode (reverse cleaning). Electrocleaning adds to the chemical action of the cleaner the mechanical action caused by plentiful gas evolution at the surface of the workpiece.

Acid Cleaning. Acid cleaning can move heavy scale, heat-treat scale, oxide, and the like. The most commonly used acids include sulfuric and hydrochloric. Pickling can also be combined with current to be more effective.

Mechanical approaches

Mechanical preparations include polishing, buffing, and some variations. Polishing is to remove small amounts of metal by means of abrasives. It produces a surface that is free of the larger imperfections left by grinding, and is a preliminary to buffing. Buffing is similar to polishing, but uses finer abrasives to remove very little metal. Buffing can produce an extremely smooth surface.

Surface Modification

Surface modification includes change in surface attributes, such as application of (metal) layer(s) and/or hardening.

Rinsing

In wet plating, when workpieces are transferred from one treating solution to another, or when they leave the final treating solution, they carry some of the solution in which it has been immersed. This solution is called drag-out. In most cases, this residue solution should be removed from the workpieces surface by rinsing before the workpieces enter the next step in sequence, or come out of the final processing solution. The dirty rinse water will be sent to the wastewater treatment facilities before being discharged to a public sewage system.

ELECTROLYTIC METAL DEPOSITION

There are three types of electrolytic metal deposition processes: direct current electrodeposition, pulse plating, and laser-induced metal deposition.[2]

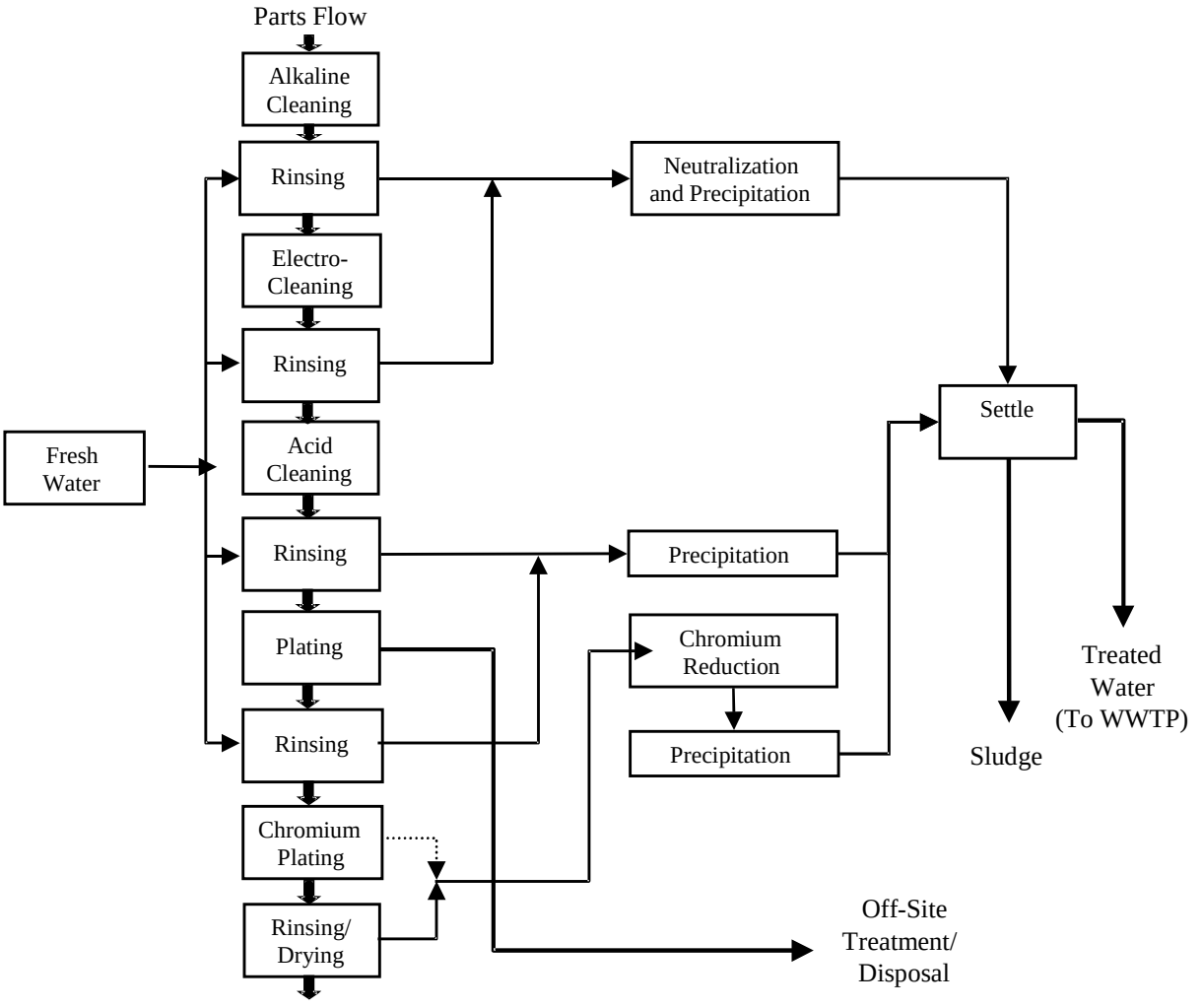


Fig. 2 Process flowsheet of a typical electroplating plant.

Direct Current Electrodeposition

In the direct current (DC) electrodeposition, the current source is a DC power source. A power source in the form of a battery or rectifier (which converts alternating current electricity to regulated low-voltage DC current) provides the necessary current.

Electroplating is performed in a plating unit. Electrodes, immersed in the electroplating bath (electrolyte), are connected to the output of a DC current source. The workpiece that is to be plated acts as a negatively charged cathode. The positively charged anode(s) completes the electric circuit. This type of circuit arrangement directs electrons (negative charge carriers) into a path from the power supply (rectifier) to the cathode (the workpiece to be plated).

The geometric shape and contour of a workpiece to be plated affect the thickness of the deposited layer. In general, workpieces with sharp corners and features will tend to have thicker deposits on the outside corners and thinner ones in the recessed areas. The cause of this difference in the resulting layer thickness is that the DC current flows more densely to sharp edges than to the less accessible recessed areas. In other words, the current distribution is not uniform. Therefore, a judicial placement of the anode(s) as well as modifications of the current density are required to overcome the thickness irregularity effects.

Pulse Plating

Electrodeposition using pulsed currents is known as pulse plating. The pulsed currents can be unipolar (on–off) or bipolar (current reversal). Pulses can be used along or be superimposed on a DC feed. By using the bipolar pulse, metal deposition occurs in the cathodic pulse period, with a limited amount of metal being

redissolved in the anodic period. This repeated deposition and partial redissolution could improve the morphology and the physical properties of the deposit.

Laser-Induced Metal Deposition

In laser-induced metal deposition, a focused laser beam is used to accelerate the metal deposition. Experiments have shown that the deposition rate can be increased by 1000 times. The plating equipment mainly consists of a laser head with focusing optics and the electrolytic cell. The focused laser beam can pass through a hole in the anode through the electrolyte and impinge on the cathode surfaces.

ELECTROLYTE

Different metals may need different types of electrolyte. The composition and properties of the electrolyte is very important for the coating quality.

Types and Components

Types of electrolytes include water solutions of acids, bases, or metal salts, certain pure liquids, and molten salts. Gases may act as electrolytes under conditions of high temperature or low pressure. In addition to metal salts, electrodeposition electrolytes usually contain a number of additives for various purposes. Some agents are used to increase electrolyte conductivity (supporting electolytes). Others may be used for increasing bath stability (stabilizers), activating the surface (surfactants or wetting agents), improving leveling or metal distribution (leveling agents), or optimizing the chemical, physical, or technology properties of the coating. These coating properties include corrosion resistance, brightness or reflectivity, hardness, mechanical strength, ductility, internal stress, wear resistance, or solderability.[2]

Properties of Electrolyte

The properties of electrolyte are usually characterized by electrolytic conductance, covering power, macrothrowing power, and microthrowing power.[4]

Electrolytic Conductance

Electrolytic conductance is different from electrical conductance in metal. Electronic conductance is called a ''Class I'' conductor, while electrolytic conductance is a ''Class II'' conductor. Both inorganic and organic salts, acids or alkalis can be used to increase the electrolytic conductance. The conductivity of an electrolyte is a function of the degree of dissociation, the mobility of the individual ions, the temperature and viscosity, and the electrolyte composition.

Covering Power

Covering power describes the extent to which an electrodeposition electrolyte can cover the entire surface of a workpiece being plated, with reasonable uniform thickness. Covering power is influenced by the nature of the substrate surface, the electrolyte composition, the temperature and viscosity, and the current density.

Macrothrowing Power

Macrothrowing power predicts the ability of an electrolyte to lay down as nearly as possible a uniformly thick deposit across the surface of a workpiece. A good covering power is a prerequisite for good macrothrowing power. Other factors that affect macrothrowing power include the current distribution and current density, electrolyte composition, electrolytic conductance, and electrolyte agitation.

Microthrowing Power

Microthrowing power indicates the extent to which metal electrodeposition occurs at the outer plane of the substrate or at the base of valleys or cracks. Microthrowing power can be improved by activating the surfaces at the base of valleys or cracks to promote electrodeposition there, while inhibiting the outer surfaces by using inhibitors preferentially. In many cases, microthrowing power is inversely related to macrothrowing power.

TYPES OF ELECTROPLATING PROCESSES

Depending on the size and geometry of the workpieces to be plated, different plating processes, including mass plating, rack plating, continuous plating, and in-line plating, may be adopted.[2]

Mass Plating

Mass plating is used for small workpieces to be plated in large quantities, such as nuts and bolts, but it is not used for delicate workpieces. The most widely used mass plating system is called barrel plating, where the workpieces are loaded into a plating barrel. Other mass plating containers include plating bells and vibratory units.

Rack Plating

Some workpieces cannot be mass plated because of their size, shape, or special features. Rack plating means workpieces are mounted on a rack for the appropriate pretreatment plating and posttreatments. Racks are fixtures suitable for immersion in the plating solution. Rack plating is sometimes called batch plating.

Continuous Plating

Continuous plating means the workpieces to be plated move continuously passing either one row or between two rows of anodes. Continuous plating is usually used for a workpiece of simple and uniform geometry, such as metal strip, wire, and tube.

In-Line Plating

In-line plating is used to integrate the plating and finishing processes into a main production line. The benefit of in-line plating includes exclusion of pretreatment steps and a significant reduction in material, chemical and energy consumption, and waste discharge.

TYPES OF METAL COATINGS

Plating metals can be roughly classified into the following categories with the typical applications.[4]

Sacrificial Coatings

Sacrificial coatings are primarily used for the protection of the base metal, usually iron and steel. Another name for sacrificial coating is anodic coating, because the metal coatings are anodic to the substrate metal, so the coatings sacrifice themselves to protect the base metal from corrosion. Zinc (Zn) and cadmium (Cd) coatings can be used as sacrificial coatings. Because of high toxicity, cadmium plating is now forbidden by law in many countries.

Decorative Protective Coatings

Decorative protective coatings are primarily used for adding an attractive appearance to some protective qualities. Metals in this category include copper (Cu), nickle (Ni), chromium (Cr), zinc (Zn), and tin (Sn).

Engineering Coatings

Engineering coatings (sometimes called functional coatings) are used for enhancing specific properties of the surface, such as solderability, wear resistance, reflectivity, and conductivity. Metals for engineering purpose include precious gold (Au) and silver (Ag), six platinum metals, tin, and lead (Pb). The six platinum metals are ruthenium (Ru), rhodium (Rh), palladium (Pd), osmium (Os), iridium (Ir), and platinum (Pt). These six metals are noble, i.e., with positive electrode potentials and they are relatively inert.

Minor Metal Coating

Minor metals here refer to iron (Fe), cobalt (Co), and indium (In). They are easily plated but have limited applications in electroplating.

Unusual Metal Coating

The unusual metals are rarely electroplated and can be divided into the following categories: 1) easily platable from aqueous solutions but not widely used, such as arsenic (As), antimony (Sb), bismuth (Bi), manganese (Mn), and rhenium (Re); 2) platable from organic electrolyte but not aqueous electrolyte, such as aluminum (Al); and 3) platable from fused-salt electrolyte but not aqueous electrolyte, including refractory metals (named because of their relatively high melting points), such as titanium (Ti), zirconium (Zr), hafnium (Hf), vanadium (V), niobium (Nb), tantalum (Ta), molybdenum (Mo), and tungsten (W). The periodic table in Fig. 3 shows that the metals that can be electrodeposited from aqueous solutions are those inside the frame.

Alloy Coatings

An alloy is a substance that has metallic properties and is composed of two or more chemical elements, at least one of which is a metal. The elements composing the alloy are not distinguishable by the unaided eye. Examples of alloy coating include gold–copper–cadmium, zinc–cobalt, zinc–iron, zinc–nickle, brass (an alloy of copper and zinc), bronze (copper–tin), tin–zinc, tin–nickle, and tin–cobalt. Alloy coatings are produced by plating two metals from the same solution.

Multilayered Coatings

Multilayered coatings are produced by plating different metals from the same solution at different potentials. A pulse train-shaped potential is enforced, resulting in the multilayer deposition. For example, multilayered coatings based on copper, nickle, chromium, in that order, can be applied to either metal or plastic components for visual appearance, corrosion and wear resistance, and weight saving.

Period	Group**																
	1 IA 1A																18 vIIIA 8A
1	1 H 1.0088	2 IIA 2A											13 IIIA 3A	14 IVA 4A	15 VA 5A	16 VIA 6A	17 VIIA 7A	2 He 4.003
2	3 Li 6.941	4 Be 9.012											5 B 10.81	6 C 12.01	7 N 14.01	8 O 16.00	9 F 19.00	10 Ne 20.18
3	11 Na 22.99	12 Mg 24.31	3 IIIB 3B	4 IVB 4B	5 VB 5B	6 VIB 6B	7 VIIB 7B	8 ------- VIII ---- --- ------- 8 -------	9	10	11 IB 1B	12 IIB 2B	13 Al 26.98	14 Si 28.09	15 P 30.97	16 S 32.07	17 Cl 35.45	18 Ar 39.95
4	19 K 39.10	20 Ca 40.08	21 Sc 44.96	22 Ti 47.88	23 V 50.94	24 Cr 52.00	25 Mn 54.94	26 Fe 55.85	27 Co 58.47	28 Ni 58.69	29 Cu 63.55	30 Zn 65.39	31 Ga 69.72	32 Ge 72.59	33 As 74.92	34 Se 78.96	35 Br 79.90	36 Kr 83.80
5	37 Rb 85.47	38 Sr 87.62	39 Y 88.91	40 Zr 91.22	41 Nb 92.91	42 Mo 95.94	43 Tc (98)	44 Ru 101.1	45 Rh 102.9	46 Pd 106.4	47 Ag 107.9	48 Cd 112.4	49 In 114.8	50 Sn 118.7	51 Sb 121.8	52 Te 127.6	53 I 126.9	54 Xe 131.3
6	55 Cs 132.9	56 Ba 137.3	57 La* 138.9	72 Hf 178.5	73 Ta 180.9	74 W 183.9	75 Re 186.2	76 Os 190.2	77 Ir 190.2	78 Pt 195.1	79 Au 197.0	80 Hg 200.5	81 Tl 204.4	82 Pb 207.2	83 Bi 209.0	84 Po (210)	85 At (210)	86 Rn (222)
7	87 Fr (223)	88 Ra (226)	89 Ac~ (227)	104 Rf (257)	105 Db (260)	106 Sg (263)	107 Bh (262)	108 Hs (265)	109 Mt (266)	110 --- ()	111 --- ()	112 --- ()		114 --- ()		116 --- ()		118 --- ()

Lanthanide Series*	58 Ce 140.1	59 Pr 140.9	60 Nd 144.2	61 Pm (147)	62 Sm 150.4	63 Eu 152.0	64 Gd 157.3	65 Tb 158.9	66 Dy 162.5	67 Ho 164.9	68 Er 167.3	69 Tm 168.9	70 Yb 173.0	71 Lu 175.0
Actinide Series~	90 Th 232.0	91 Pa (231)	92 U (238)	93 Np (237)	94 Pu (242)	95 Am (243)	96 Cm (247)	97 Bk (247)	98 Cf (249)	99 Es (254)	100 Fm (253)	101 Md (256)	102 No (254)	103 Lr (257)

Fig. 3 Periodic table (metals inside the frame can be electrodeposited from aqueous solutions). *(View this art in color at www.dekker.com.)*

Composite Coatings

Composite materials can be defined as coatings consisting of minute second-phase particles dispersed throughout a metal matrix. The size of the second phase particles may range from 10 μm down to nanoscale and the particles can be inorganic, organic, or occasionally metallic. The presence of fine particles in a metal matrix generally improves its mechanical and chemical properties, resulting in a wide range of applications. Composite coatings with an electrodeposited metal matrix and nonmetallic inclusions have excellent wear resistance and permit emergency dry running of machinery.

Conversion Coatings

Conversion coatings are formed by a reaction of the metal on the surface of the substrate with a solution.[5] For example, chromate coatings are formed by the reaction of water solutions of chromic acid or chromium salts. The chromate coatings can be applied to aluminum, zinc, cadmium, and magnesium. The coatings usually have good atmospheric corrosion resistance. Chromate coatings are widely used in protecting common household products, such as screws, hinges, and many hardware items with the yellow-brown appearance.

Anodized Coatings

Anodizing is produced by electrochemical conversion. In an anodizing process, the metal workpiece to be plated is the anode in a suitable electrolyte. With the electric current passing through the electrolyte, the metal surface is converted to a form of its oxide. An anodizing process is usually used on aluminum for protection and cosmetic purposes. The electrolyte provides oxygen ions that react with metal ions to form the oxide, and hydrogen is released at the metal or carbon cathode. Anodizing differs from electroplating in two aspects. In electroplating, the workpiece to be plated

is the cathode, and the metallic coatings are deposited on the workpiece. In anodizing, the workpiece is the anode, and its surface is converted to a form of its oxide.

RELATED PROCESSES

The related processes for metal deposition include electroless deposition, immersion plating, and electroforming.[4] They follow the basic principles of electrochemistry.

Electroless Deposition (Autocatalytic Plating)

A special type of electroplating is called electroless deposition, autocatalytic plating, or ''chemical deposition.'' In electroless plating, there is no external power source. The deposited metal is reduced from its ionic state in solution by a chemical reducing agent. The reducing agent supplies the electrons for the following reaction:

$$M^{n+} + ne^{-} \rightarrow M \quad (14)$$

This reaction takes place only on a catalytic surface. Therefore, once deposition is initiated, the metal deposited must itself be catalytic for the deposition to continue.

Not all metals can be plated autocatalytically. The reducing agents are usually more expensive electron sources as compared with the electric current. The major advantages of electroless deposition are as follows:

1. It can be used to deposit metal on nonconductive surfaces, such as plastics, glass, or ceramics. Some proper pretreatment steps are needed to activate these surfaces. The metallizing of printed circuit board is one such example.
2. The throwing power is perfect. Deposits are laid down on the surface with no excess buildup on projections or edges.

Immersion Plating

Immersion plating is the deposition of a metallic coating on a substrate by chemical replacement from a solution of salt of the coating metal. It requires no electric circuitry or source of power, but it differs from autocatalytic plating in not requiring a chemical reducing agent to reduce the metal ions to metal. Immersion deposition stops when the substrate is completely covered by a layer of coating.

The major advantages of immersion plating include simplicity, minor capital expense, and the ability to deposit in recesses and on the inside of the tubing. But, the applicability of immersion plating is limited.

Electroforming

Electroforming is to produce or reproduce a metal workpiece by electrodeposition in a plating bath over a base form (mold) or mandrel, which is subsequently removed. In some cases, the mandrel or mold may remain within the finished metal workpiece. A mandrel is a form used as a cathode in electroplating.

The advantage of the process is that it faithfully reproduces a form of mandrel exactly, to within one micrometer, without shrinkage and distortion associated with other metal forming techniques, such as casting, stamping, and drawing. Because the mandrel is machined as an outside surface, close dimensional tolerances and high surface finishes can be held and maintained on complex interior configurations. The disadvantage of electroforming includes slow production, relatively high cost, design limitations of the geometry, and the separation of workpieces from the mold or mandrel.

CONCLUSIONS

The electroplating industry has been experiencing continuous innovations and also facing significant challenges from economic and environmental perspectives. The purpose of electroplating is to produce a qualified coating with the desirable attributes. Based on the specifications of the coating and the substrate, one may select a specific electroplating process for a given application.

ACKNOWLEDGMENTS

This work is in part supported by NSF (DMI-0225843 and DMI-0225844) and EPA (R824732-01).

REFERENCES

1. ASTM International. In *B374-96 (2003) Standard Terminology Relating to Electroplating*; ASTM International: West Conshohocken, PA, 2003.
2. Kanani, N. *Electroplating: Basic Principles, Processes and Practice*; Elsevier Advanced Technology: Oxford, U.K., 2004.
3. Schlesinger, M. Electroplating. In *Electrochemistry Encyclopedia*; http://electrochem.cwru.edu/ed/encycl/ (accessed December 2004).
4. Lowenheim, F.A. *Electroplating*; State Mutual Book & Periodical Service: New York, 1997.
5. http://www.efunda.com/processes/surface/conversion_coatings.cfm (accessed Dec 2004).

Electrostatic Precipitation

Kenneth R. Parker
Ken Parker Consultant APC, West Midlands, U.K.

INTRODUCTION

Although electrostatic precipitators have been employed almost for a century to remove particulates, fumes, and mists from a wide range of industrial processes, the main approach for assessing the size and design of a new plant has been based on measurements from units operating under similar conditions. Recent developments using computerized fluid dynamics approaches have enabled the physics to be theoretically studied in greater detail, which leads to an improved understanding of precipitation. In this entry, the precipitator operation is reviewed from theoretical and practical aspects and some of the gas and particulate characteristics, which have to be considered when designing a precipitator for a specific duty, have been indicated. While it is mainly used currently for pollution control purposes, where efficiencies of 99.8%+ are required to comply with emission legislation, the precipitator can be used for the recovery of valuable materials, cleaning process gases for subsequent reuse, air cleaning duties, and the cleaning of process oil streams. The recent issue by the EPA of the ESPVI Series 4W precipitator program enables power station precipitators to be proactively studied to speedily evaluate low-cost enhancement scenarios for precipitators that require upgrading to meet more stringent emission regulations.

BACKGROUND

Electrostatic precipitators have been used commercially for almost a century for the collection of particulates present in the gas streams on many industrial processes. The early plants were installed to recover valuable material that, otherwise, would have been lost to atmosphere, rather than in preventing atmospheric pollution. These plants, mainly in the sulfuric acid, bullion, and nonferrous smelting industries, were designed and priced to produce a commercial ''pay back'' in the order of 3–5 years. In recent years, with the increasing awareness and worldwide recognition of the problems associated with atmospheric pollution, most industrialized nations have enacted legislation to limit emissions from all sources. This legislation is continuously reviewed and is becoming steadily more stringent, such that, in general, particulate emissions are presently controlled to a maximum of $50\,mg/Nm^3$ for inert materials and a maximum of $10\,mg/Nm^3$ for substances that are considered hazardous to health.

The features of the electrostatic precipitation process that make it an ideal vehicle for the removal of particulates are

1. Versatility and effective performance on a wide range of industrial processes. Can be designed to satisfy any required efficiency and gas flow rate.
2. Designs can be produced to cover a temperature range from ambient up to 850°C.
3. Can collect particles over the complete size range spectrum.
4. Material is usually recovered in its original state. (Plants can be designed to operate as a wet phase device if required.)
5. Low pressure loss; typically less that 1 mbar.
6. Acceptable electrical power consumption for required efficiency level.
7. Robust and reliable construction and a life expectancy of more than 20 years.
8. Low maintenance requirements.

BASIC OPERATING PRINCIPLES

The basic principle of operation is that the particulates are passed through an electric field where they initially receive an electric charge to become charged particles and are deflected across the field to be retained on the counter electrode. Most industrial precipitators are based on a single-stage approach where both charging and migration (precipitation) take place within the same set of electrodes, as illustrated in Fig. 1. Another type of unit, particularly used for air cleaning duties, is two-stage precipitation, where there is a separate charging field followed by a collection field, as illustrated in Fig. 2.

In general, the basic components of any precipitator consist of some form of small diameter discharging element and, for a parallel flow unit, a flat plate counter electrode, or, in the case of vertical flow mist type of installation, a large diameter tube.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007748

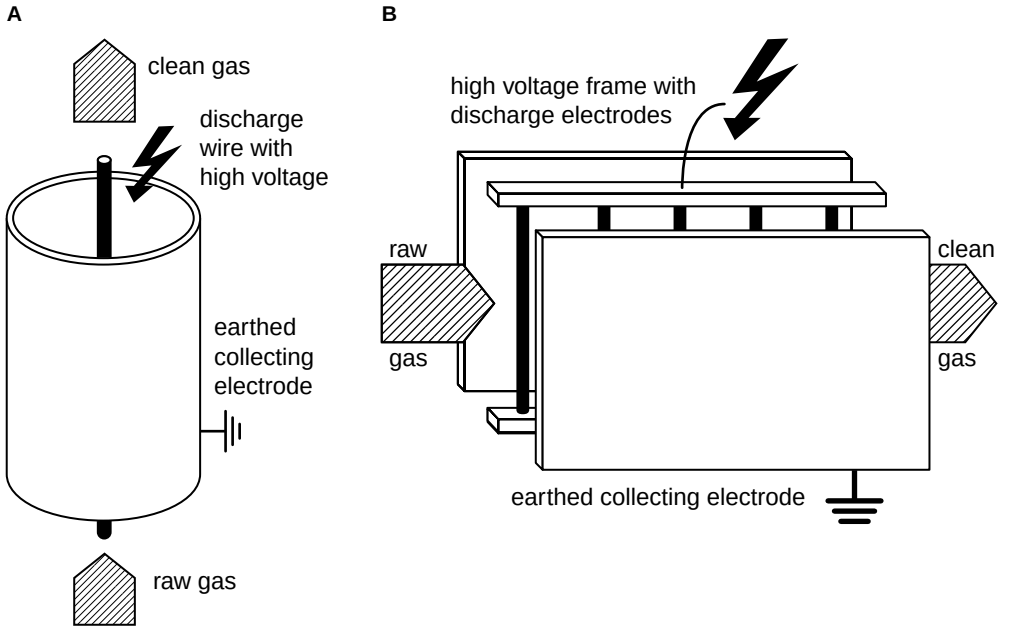


Fig. 1 Basic single-stage ESP arrangements: (A) a wire tube arrangement and (B) and parallel plate arrangement.

In practice, the form of discharge element ranges from a simple weighted wire, typical of tubular units, to specially designed controlled emission electrodes, used in a number of industries, not only to mitigate corona suppression from high concentrations of fine particulates but also to provide a long and trouble-free operating life.[1] The receiving, or collecting, electrodes range from stiffened flat plates to some form of channel construction for multifield high efficiency parallel flow applications. The size and the form of receiving electrode depend on the supplier's preference, but plates with a length of 5 m and a height of 15 m can be found in power station precipitators.[2] While for vertical flow mist type units, tubes with a diameter of 300 mm and a length of 6 m can be found.

Until the recent development of switched mode power supplies (SMPS), most industrial precipitators were energized from some type of 50–60 Hz rectified high voltage equipment. This basically comprised a suitably designed and insulated step-up transformer, the output from which can be controlled by employing silicon controlled rectifiers to modulate the incoming supply voltage. The transformer output is then rectified using silicon rectifiers to produce the high negative direct current (DC) voltage required for the necessary ionization for charging and the electric field for particle migration.

Recent developments in switched mode or high frequency direct current supplies have been applied to new and some older plants where performance enhancement is required.[3] In this form of equipment, the normal three-phase supply is first rectified and converted to a 25–50 kHz square wave, which can be modulated to supply the HT transformer, which, having a ferrite core, is much smaller and lighter than its conventional counter part. The output of the HT transformer is then rectified to produce the high voltage requirement. This form of supply has a number of advantages over the conventional (50–60 Hz) frequency supply, in that it produces almost a pure DC voltage, which enables the precipitator to operate at higher voltages and performance levels, while being much more efficient in terms of power conversion (95% as against 85%) and a power factor of 0.94 as

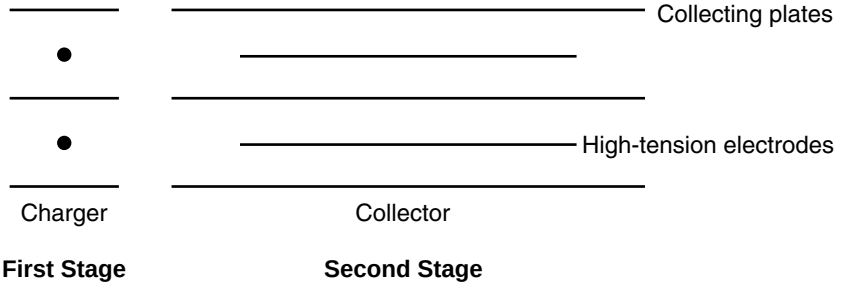


Fig. 2 Two-stage ESP arrangement.

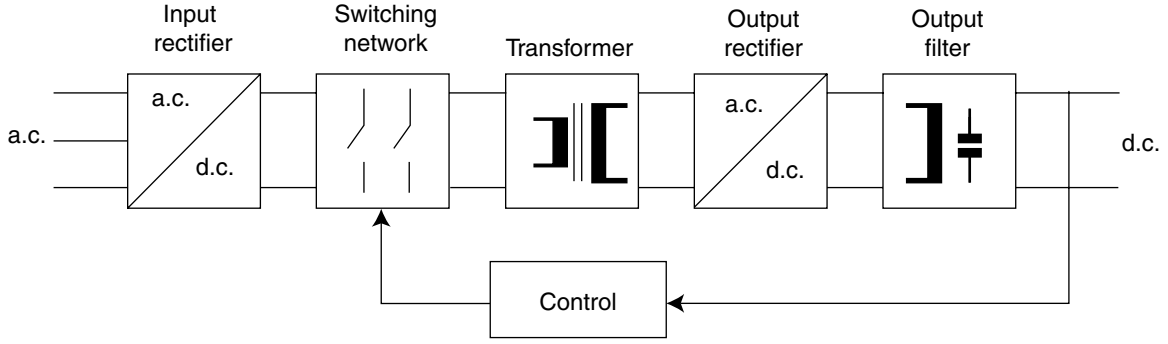


Fig. 3 Basic building block for a switch mode power supply.

against 0.63. The SMPS basic building block is shown in Fig. 3.

For air cleaning duties, although energization is normally derived from conventional supplies, the voltages are much lower, 6–12 kV, and typically positive, because of a lower ozone production, resulting from the reaction between the ions and the oxygen in the air.

In operation, the high electrical field adjacent to the energized electrode elements ionizes the gas molecules, forming both positive and negative ions. With negative energization, the positive ions are immediately captured, while the negative ions and any electrons, generally referred to as a corona discharge, migrate under the influence of the electric field into the interelectrode space. As the gas borne particles pass through the interelectrode space, the larger particles receive an electric charge either by collision with the ions/electrons or by diffusion charging for the smallest particles. The charged particles then move under the influence of the electric field and migrate to the collecting electrodes, where the charge subsequently leaks away to earth. In the interim, before becoming neutral, the charged particles are retained on the collector surface by a combination of van der Waals and electric forces.

In a dry precipitator to ensure that the collection process is continuous, after a period of time the collector electrodes are generally mechanically rapped to remove the deposited material. This time aspect, together with the frequency and intensity of the rapping, is important to minimize rapping re-entrainment and hence maximize collection efficiency. In wet precipitators, the particles are deposited on a film of running water or removed by flushing, while for mist precipitators, the droplets are usually selfdraining.

THE PHYSICS OF PRECIPITATION

Ion Production

Although there are various methods of particle charging, for example, triboelectric, ultraviolet, and radiation effects, for industrial precipitator applications corona charging is universally used, as it is the most efficient and cost-effective approach. The physics of corona or ion production therefore occupies an essential position in the practice of electrostatic precipitation. Investigations into the physics of ionization date back to the middle of the last century and are still ongoing, particularly since the development and application of computational fluid dynamics (CFD) using fast computers capable of deriving satisfactory solutions to the Poisson and La Place equations.

Early investigations were concerned with developing voltage–current relationships of corona discharge and the effects of positive and negative energization systems using wire and tube arrangements. Gaugain[4] found that the breakdown potential for a given outer diameter R depends on the diameter of the central discharge element r and derived the following relationship:

$$E = A + (C/r)^{1/3} \tag{1}$$

where E is the electrical breakdown field, r, the radius of inner electrode, and A and C are experimental constants (for $r \ll R$).

Rőentgen,[5] working with point/plane electrodes, found that a certain voltage had to be applied to initiate a corona current flow. This corona onset voltage was dependent on the sharpness of the point, gas pressure, and polarity of the point electrode. Further investigations produced a parabolic relationship for negative corona current flow as:

$$I = AV(V - M) \tag{2}$$

where I is the corona current flow, V the applied voltage, and A and M are experimental constants.

Townsend[6] noted that, with negative ionization of the discharge electrode, the appearance of the corona discharge was very different to when it was positively energized, when the corona appeared as a diffuse glow surrounding the emitter. With negative energization,

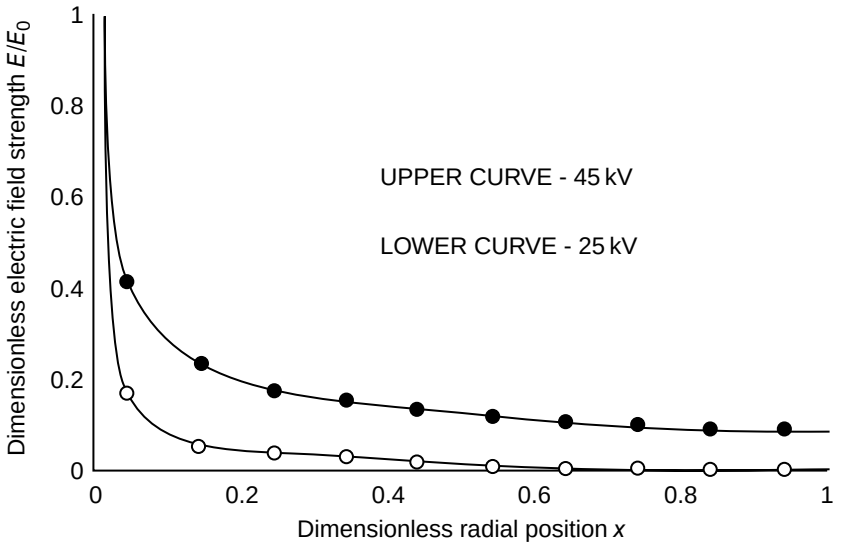


Fig. 4 Relative field strength between the electrodes of a tubular precipitator arrangement. *(View this art in color at www.dekker.com.)*

the corona appeared as bright flares, which tend to move across the surface of the electrode, and had a distinctive hissing sound, a lower initiation voltage, and a higher breakdown potential. The visual appearance of negative corona takes the form of ''tufts''—bright glow points—or Trichel pulses.[7] The number of tufts and their luminescence increase as the energizing voltage is raised, producing a higher corona current discharge.

The conduction of electricity through gases is fundamentally different to that in solids and liquids, which contain ''charge carriers'' that move under the influence of the electric field to produce the current flow. With gases, however, ions need to be provided from some outside force or agency to induce a current flow. For precipitators, this outside agency is high voltage applied across the electrode system. In Fig. 4, a typical electric field distribution between a small diameter emitter and a much larger passive electrode is shown (i.e., $r \ll R$).

This shows the electric field adjacent to the emitter is extremely high and it is this stress that excites any free electrons in the immediate vicinity. These fast moving electrons acquire sufficient energy from the applied electric field so as to collide with other gas molecules to produce further free electrons and positive ions. Townsend,[6] working in this area, proposed the concept of a chain reaction or electron avalanche, in which each new electron produced generates new electrons by ionization in ever increasing numbers.

The number of ions at a distance x from the active zone can be represented by:

$$n = n_0 e^{ax} \tag{3}$$

where n_0 represents the number of ions at a distance $x = 0$ and a is the Townsend ionization coefficient, which varies with the gas temperature, pressure, and electric field strength.

In general terms, the field strength varies with distance across the field, x, as in a precipitator. Hence the equation takes the integral form:

$$n = n_0 e^{\int_0^x a\,dx} \tag{4}$$

The term $1/a$ is the mean free electron path between collisions.

In Fig. 5, the relationship between the Townsend ionization coefficients (ion pair production) for air at atmospheric pressure at different temperatures and field strengths is given.

In the figure, it can be seen that, at 20°C, doubling of the field strength results in the number of ion pairs,

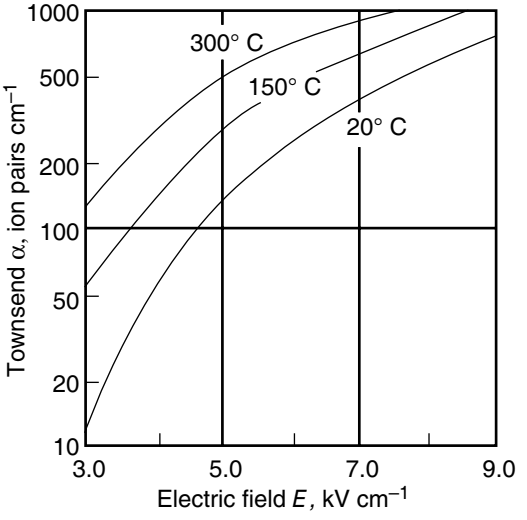


Fig. 5 Effect of electric field strength and temperature on ion pair production.

or Townsend coefficient, increasing by a factor of 20. The impact of a rise in gas temperature, which increases the mean free electron path and produces significantly increased ionization, is also indicated.

In practice, when the system is negatively energized, the electrons collide with and attach themselves to gaseous molecules to produce negative ions as they rapidly move across the field area, although there are a large number of ion pairs immediately adjacent to the discharge element. Concurrently, the positive ions are attracted toward the discharge element and, although during transit they produce further ion pairs on reaching the element take no further part in the process. As the distance from the element increases, because of attachment, the number of electrons decreases, and there is a corresponding increase in the number of negative ions. The net number of electrons at a distance x from the electrode is represented by the following equation:

$$n = n_0 e^{\int_0^x (a-\eta)dx} \tag{5}$$

where η is the coefficient of ion attachment or the Townsend second coefficient.

In the case of positive energization, normally found in air cleaning applications, the primary electrons produced at the boundary of the visible glow are attracted toward the emitter. In moving through the field, they collide with and produce new ion pairs by impact ionization, with the positive ions migrating toward the passive earthed electrode. Except for the collection of electrons, the emitter by itself plays little part in the ionization phenomenon, which is essentially a gas process with the primary electrons being released from the gas molecules through photoelectric effects in the plasma region.

Particle Charging

Particle charging occurs in the area between the active plasma region and the passive electrode surface. This area comprises a high space charge having neutral ions, negative ions, and some free electrons, all moving toward the passive electrode as a result of the electric field. As the gas borne particles enter the corona derived space charge region of the field, two charging mechanisms occur: the first is by ion attachment, i.e., field or impact charging, and the second by ion diffusion charging. The field or impact charging predominates for particles with a diameter greater than 1 μm ($1\mu m = 1 \times 10^{-6}$ m), while diffusion charging is essential for particles with a diameter less than 0.2 μm; both processes occur in the intermediate size range.

While field charging requires the presence of an electric field to drive the free mobile charge carriers, the diffusion process is based on randomly moving gas ions arising through temperature effects as described by the kinetic theory of gases, i.e., Brownian motion, which plays a significant and important role in the collection of particles in the submicron size range, in spite of their saturation charge being much smaller than that of the larger particles.

Over recent years, a great deal of numerical modeling work has been carried out using computational fluid dynamics to derive particle charging models;[8–10] however, the basic models need experimental support because the equations cannot be analytically solved. A reasonable alternative to modeling is proposed by Cochet,[11] who developed an equation, which appears to give reasonable correlation to actual precipitator measurements in the critical size range, as tested and reported by Hewitt.[12]

The particle saturation charge Q_p^{∞}, according to Cochet, is given by the following formula:

$$Q_p^{\infty} = \left\{ \left(1 - \frac{2\lambda}{d_p}\right)^2 + \frac{2}{1 + 2\lambda/d_p} \times \left(\frac{\varepsilon_r - 1}{\varepsilon_r + 2}\right) \right\} \times \pi \varepsilon_o d_p^2 E \tag{6}$$

where ε_r is the electrical permittivity of the particle, ε_o the electrical permittivity of the gas, d_p the particle diameter, E the electric field strength, and λ the mean free path electron path.

The equation can be rewritten in simplified terms:

$$Q_p^{\infty} = pEd_p^2 \tag{7}$$

where p is a constant, which varies between 1 and 3 fornonconductive particles and is 2 for conductive particles.

In practice, both field and diffusion charging occur simultaneously and are basically inseparable. It can be shown that all particles reach around 90% of their full saturation charge in less than 0.1 sec. Hence, for all practical conditions, it can be assumed that on entering the precipitation field, having a typical exposure/treatment time of 3–5 sec, the particles rapidly achieve their saturation charge.

Particle Migration

Within the precipitation field, a particle experiences the following forces acting upon it: a momentum force, $F_m = ma$; an electrical force, $F_e = Q_p E$; and a drag force, $F_d = \mathrm{Re}AC_o$ (Re is the Reynolds number and C_o the Cunningham coefficient).

Under a steady state condition

$$F_m + F_e + F_d = 0 \quad (8)$$

Prior to solving this equation, the drag force F_d has to be calculated.

In the case of low Reynolds number, the drag coefficient is given by

$$C_o = 24/\mathrm{Re} \quad (9)$$

As the particle size d_p reduces and approaches the region where the fluid loses its continuum (mean free path of molecules $= \lambda$), Stokes law needs to corrected by the Cunningham correction factor, C_o.

$$C_o = 1 + 1.246 \times 2\lambda/d_p + 0.42 \times 2\lambda/d_p \times e^{(0.87d_p/2\lambda)} \quad (10)$$

This relationship is plotted in Fig. 6, in which the significant correction factor that arises for submicron sized particles is shown.

The drag force (or Stokes law) can be written as

$$F_d = 3\pi\eta d_p \omega_{th} \times \frac{1}{C_o} \quad (11)$$

where η is the gas viscosity, d_p the particle diameter, C_o the Cunningham correction, and ω_{th} the theoretical particle migration velocity.

Assuming the fluid has no component acting toward the passive electrode and all particles achieve their saturation charge, the equation of motion of a charged spherical particle in an electric field is characterized by

$$\frac{d\omega_{th}}{dt} + \frac{3\pi\eta d_p \omega_{th}}{mC_o} = \frac{Q^{\infty}{}_p E}{m} \quad (12)$$

Taking $\omega_{th} = 0$ at $t = 0$, the solution of the above equation can be readily found, i.e.,

$$\omega_{th} = \frac{Q^{\infty}{}_p E}{3\pi\eta d_p} \times C_o \quad (13)$$

As $Q^{\infty}{}_p = pEd_p{}^2$,

$$\omega_{th} = E^2 \times d_p \times C_o \quad (14)$$

The significance of this relationship is

a. As the limiting charge on the particle is proportional to the radius squared, the migration velocity of the particle will increase with particle size.
b. As the electric field is proportional to the applied voltage, the migration velocity is proportional to the voltage squared.

It will be appreciated from this fundamental approach that the operation of an electrostatic precipitator requires a voltage high enough to produce an electric field to precipitate the particles and deliver sufficient corona current to satisfy ion production for initially charging the particles.

In Fig. 7, the relationship between the theoretical migration velocity ω_{th} against particle size for three different field strengths in a typical flue gas at 150°C, for particles having an electrical permittivity of 10, is given. Because of the logarithmic scale, the effect of

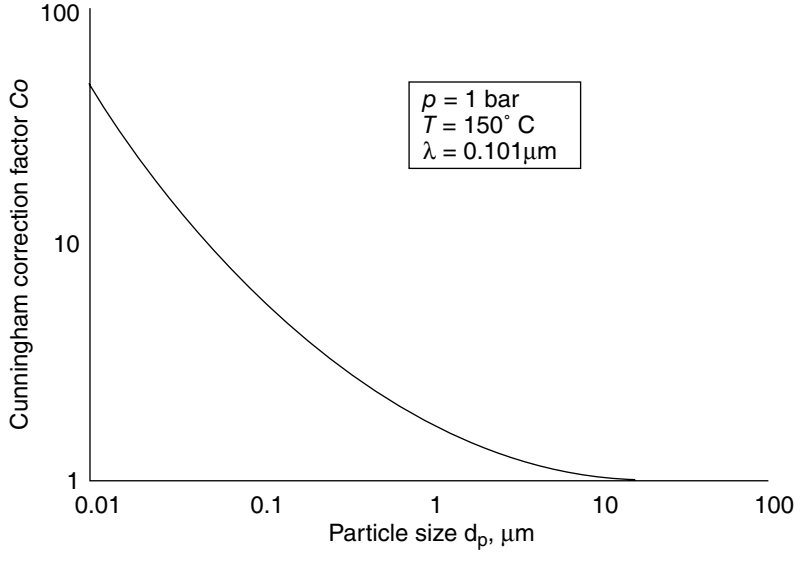


Fig. 6 Relationship between particle size and the Cunningham correction factor.

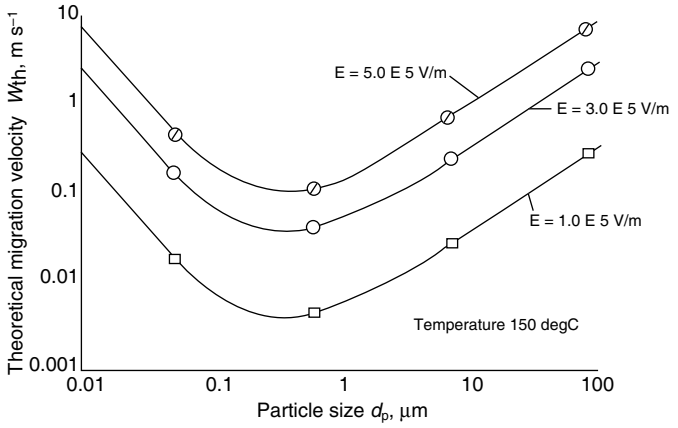


Fig. 7 Relationship between the theoretical particle migration velocity and particle size.

the field strength squared factor is not apparent, unless one examines the actual migration velocity figures.

Although the foregoing portrays the theoretical approach to the migration of charged particles through an electric field, the theoretical migration velocity should not be confused with the Deutsch ''effective migration velocity,'' which is derived from plant efficiency measurements and the specific surface area of the precipitator. The effective migration velocity derived from measured efficiencies and the specific collection area for the precipitator should more realistically be considered as a measure of a ''performance factor'' because it applies to the material that has been collected and not to the finer material that invariably forms the major part of the emission.

Regardless of which method is considered, the general trend of a much reduced migration velocity/efficiency at around 0.5 μm region is apparent. The subsequent increase in efficiency of smaller particles is the result of Brownian motion, which aids their charging and migration. To derive a fractional efficiency curve in practice, it is necessary to determine the particle sizings at both the inlet and the outlet of the precipitator and then evaluate separate grade efficiencies from the overall mass precipitator efficiency. The separate grade efficiencies can then be transposed into effective migration velocities from the Deutsch relationship.

In Fig. 8, a typical particle size penetration (1 – efficiency) relationship is illustrated through penetration curves based on both particle mass and number relationships; both demonstrate the high penetration window in the 0.5 μm region.

Particle Deposition and Removal from the Collector Electrodes

When considering the motion of a charged particle in an electric field, until the particle has been removed from the collector plate and transferred to the receiving hopper, it should not be considered as being collected. This is one of the many reasons why the theoretical values of migration velocity are usually far higher than those based on actual measured efficiency/plate area calculations. The theoretical approach also assumes that the particle on arriving at the collector losses its charge and plays no further part in the process. In practice, charged particulates arriving at the collector can cause a serious impact on the electrical operating conditions and hence the performance.

As the particle arrives at the collector, the first reaction is that it is held by a combination of van der Waals and electrical forces. Depending on the thickness of the deposited layer and the electric resistance of the layer, the charge on the particle leaks away to earth, leaving the particle to be retained only by van der Waals forces. Unless the deposit is periodically removed, the precipitation process slowly degrades, as the deposit has a deleterious impact on the electrical operating conditions. In the case of a dry precipitator, this is achieved by some form of

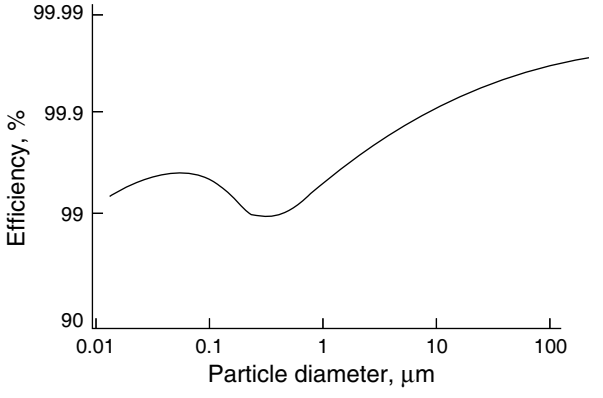


Fig. 8 Precipitator fractional efficiency curve.

mechanical rapping, and in the case of a wet precipitator, depending on the design, this is achieved by periodic water washing (flushing), a continuous spray irrigation system, or having a film of water continuously flowing over the collectors (film flow).

For a dry application, the frequency and the intensity of the rapping must be such as to shear the deposit as a layer from the plates, rather than attempting to disintegrate the layer completely, as this would lead to severe rapping re-entrainment. Ideally for the particles to reach the hopper and to overcome the horizontal component of velocity, the effective size of agglomerate should be in the order of at least 500 μm in diameter.[13]

As the rate of deposition varies along the length of the precipitator, the frequency of rapping reduces from the inlet to the outlet fields. This enables the deposited layer between blows to achieve a thickness, such that it can be sheared from the plates as an optimum sized agglomerate to reach the hoppers with the minimum of re-entrainment. Because the mean particle size of the deposited material decreases along the length of the precipitator, ideally, and because of the increased packing density and cohesion between particles, the intensity of the rapping blow should increase as the particle size becomes smaller toward the outlet.

Although the majority of the dust is deposited on the collectors, some is deposited on the discharge electrodes, and, to optimize corona production, the discharge electrodes should be kept deposit free. On a dry plant, this is achieved by rapping systems similar to those employed for the collectors, while on wet systems, water washing of the internals is usually effective in maintaining corona emission.

The relationship, indicated in Eq. (14), is important in practice, because to optimize performance it is important that the voltage applied across the electrode system is maintained at the maximum possible level. The actual operating voltage across the electrode system is largely dependent on the properties of the gas, the material suspended in the gas, the layer of deposited material, and the design of the electrode system. Though it would be ideal to consistently operate just below the electrode breakdown potential, as the complete system is dynamic and the conditions are constantly changing, an automatic voltage control system is used to maintain the voltage and, hence efficiency, as high as possible.

FACTORS AFFECTING THE DESIGN AND PERFORMANCE OF PRECIPITATORS

To obtain a better understanding of the factors that affect precipitator design and operation, one must examine the characteristics of both the gas and particulates as they are presented to the precipitator.

Gas Composition

The main requirement of the gas carrying the particulates is that it must be capable of maintaining as high an electric field as possible and permit the flow of corona current. The composition of some gases can, however, affect the electric operating conditions of the precipitator; generally the corona characteristics are modified by the presence of electropositive or electronegative gases, i.e., gases that readily absorb or reject negative ions.

Although one normally thinks of electrostatic precipitators only being applied for the removal of particulates from gas streams, the process has been satisfactorily applied for the removal of particulates from oil streams; again the main criterion is that the oil must be sufficiently insulating to maintain an electric field. The arrangement of the plant is somewhat different to a conventional gas application. Because the electrodes are normally perforated and the oil passes freely through them, the particles are contact charged before they migrate to the area of maximum field intensity between the electrode edge and casing wall. Here they are allowed to agglomerate to a sufficient size to fall under gravity into the base of the processing vessel from where they can be removed. Some installations use high voltage AC for initial treatment, followed by high voltage DC for final cleaning of the oil stream; others use DC treatment only.

Gas Temperature

The gas temperature has impacts mainly on the materials of construction; ordinary carbon steel is adopted as the cost-effective material for most applications, which limits the operating temperature to approximately 400°C, while for higher temperatures, stainless or high nickel alloys need to be used. In the case of plants operating close to acid dew point temperature, particularly for wet type environments, when choosing the material for fabrication, one must take corrosion into account to ensure durability.[14]

The main advantage of the dry type of precipitator is that the gas in most applications can be delivered directly to the precipitator from the process without the need for additional cooling or pretreatment. This means that the collected material is usefully captured in a dry state for subsequent disposal/reuse and that the cleaned gases are emitted from the chimney buoyant and usually at a temperature high enough to result in a steam free discharge.

Gas Pressure

Because most process plants operate close to ambient pressure conditions, gas pressure is not a major effect, except that one has to ensure the casing is ''gas tight'' and will withstand the operating conditions, i.e., to prevent either the egress of process gas or the in leakage of ambient air.

Gas Flow Rate

In the initial sizing of a precipitator, while it is important to have accurate knowledge of the total gas flow such that the correct contact time can be assessed to meet the required efficiency, there is also an optimum operating gas velocity to be considered. This optimum velocity is determined to some extent by the particle characteristics; too high a design velocity for a dry application can result in particle scouring and potential rapping re-entrainment, while too low a velocity will detract from the overall collection efficiency, as the deposition along the collector plate, most being collected close to the inlet, adversely distorts the electric precipitation field.

Generally with wet or mist applications, as the deposited material is retained on a wetted surface, operational gas velocities can be appreciably higher than that of a dry precipitator, as the risk of particle scouring is significantly reduced.

It is important that, to optimize the collection efficiency, the gas distribution across the frontal area of the precipitator must be as uniform as possible, although some recent work has indicated that for certain conditions a ''skewed'' velocity profile can be advantageous.[15] It is not easy to produce a typical operational gas velocity of around 1.5 m/s, which has been decelerated from approximately 15 m/s in the inlet approach ductwork, and an acceptable standard of distribution is an RMS deviation of 15%.[16] This standard can be achieved through field corrective testing, large-scale model tests, or the more recent CFD approaches.[17]

Viscosity and Density

These parameters are determined by the compositionand temperature of the gases and affect the precipitation process as follows:

a. As the charged particles migrate through the interelectrode space, they are retarded by the effects of gas viscosity and density.
b. As the temperature rises, although the gas density decreases assisting ion movement, the viscosity rises and hinders the particle transportation mechanism.

Particle Concentration

For many applications, the main effect of inlet fly ash concentration is minimal and normally has impact only on the overall removal efficiency requirement. Generally, increases in the inlet loading over that specified tend to arise from an ''upset condition'' on the process plant and are normally associated with the carryover of larger sized particles, which are more readily precipitated than the finer ones. Nevertheless, it is possible that the upset condition can modify other characteristics of the gas and result in an unacceptable emission transgression.

Particle Composition and Electrical Resistivity

For most dry process applications, the major impact of particle composition and precipitator performance is related to the electrical resistivity of the deposited material. The value of the electrical resistivity of the deposited particles can result in the required exposure/contact time within the precipitation fields varying by a factor of 4 or more.

For resistivities greater than approximately 10^{12} Ω cm, a phenomenon known as back or reverse ionization arises, which severely detracts from the overall efficiency. Because of the increasing resistance of the deposited material as the precipitation process proceeds, particles subsequently arriving only slowly lose their charge. Consequently, a voltage begins to build up on the deposit and in the worst case reaches a point where positive ions begin to be emitted from the surface of the layer. These positive ions not only neutralize any negative charge on the arriving particles but also considerably modify the electric field such that the overall efficiency is compromised. The onset of reverse ionization can be readily recognized by a significant increase in current flow and an apparent fall in average operating voltage.

With fly ash having a slightly lower dust resistivity, an alternate operational condition arises, where instead of the voltage continuing to build up on the surface of the deposited layer to initiate positive ion flow, it reaches a value where electrical break down occurs within the deposited material, producing a ''leader,'' which results in flashover across the interelectrode area. This condition produces an electrical condition typified by a slightly reduced operating voltage with a significant fall in corona current; any attempt to increase the current into the plant only results in further sparking and a decreasing performance level.

With conductive materials having electrical resistivities below 10^8 Ω cm, such as unburnt carbon or metallic particles, the particles are readily charged by the corona, and, as they reach the collector, they lose their charge so rapidly as to be repelled back into the gas stream. Although the charging and repelling process can occur several times during their transit through the precipitator, some particles can leave the plant without being captured.

An important thing as regards chemical composition and format is to know if the material is likely to produce a sticky deposit, or is in a liquid phase, when it reaches the collector electrode, such that the correct type of precipitator can be supplied. For wet or mist type applications, the materials of construction need to be corrosion resistant to protect it from either the gas or the particulate components.

Another difficulty, usually associated with particulate composition, is the cohesive strength of the deposited layer. If, on a dry plant, the material lacks cohesion and is only lightly bonded together, it is likely to be readily re-entrained during rapping and thus detract from the overall performance.

Particle Sizing

The electrostatic precipitator can effectively collect particulates having diameters from 0.01 μm to approximately 100 μm. The fractional efficiency, however, is not constant, as there is a higher penetration window in the 0.5 μm diameter range, which, coincides with the change from collision to induction charging of the particles (Fig. 8). This in itself, provided the particle sizing is available for a specific application/duty, means that the precipitator size needs to be increased to cater for particles falling in this range. This is of prime importance in enhancing existing precipitator performance levels since the emitted material, because the penetration window contains a major proportion of the finer particles.

Particle Shape

For most applications, the particle shape can be either granular resulting from the comminution of the feed material by grinding, milling, etc., or, in the form of a fume. These usually result from the material being initially volatilized in the high temperature zone of the process to subsequently condense into a spherical fume upon cooling. Neither of these forms, provided they are known or assessed beforehand and the size aspect has been considered in the design parameters, has a serious impact on the overall performance. For those processes producing a large mass of fine fume, space charge effects need to be addressed in the form of discharge element used to overcome potential corona suppression effects.

Operational problems can arise with platelet type materials, which have a fairly large surface area but virtually no thickness. These are very light in terms of mass and tend to attach themselves on to the collectors and each other, thereby reducing electrical clearance, which leads to flashover.

The presence or carryover of siliceous cenospheres, which sometimes arise during combustion, is understood to result from burning carbon particles, which are being trapped within a larger fused siliceous particle, producing sufficient gas to inflate the molten mass into a hollow sphere. This problem is not because of their shape, which is usually spherical, but because being hollow; they have little mass and poor cohesive properties, so can be readily re-entrained by rapping after being deposited.

Particulate Surface Properties

As far as precipitator performance is concerned, the surface condition/properties of the particle are generally more important than the chemical composition of the actual matrix. In many combustion applications using sulfur bearing carbonaceous fuels, the waste gases on cooling can pass through an acid dew point temperature, and any resultant sulfuric acid mist produced subsequently uses the particles as condensation nuclei depositing a thin layer of highly conductive material on their surface. At higher temperatures of operation, e.g., above 300°C, there is little possibility of reaching any dew point temperature and it is the chemical matrix of the particle that primarily governs the electrical resistivity.

Where high resistivity particles are met in practice, the electrical operating characteristics and the difficulties of the reverse ionization phenomenon can be mitigated by the injection of chemical reagents into the flue gases ahead of the precipitator to modify the electrical resistivity.[18] (Other reagents can be used to increase the cohesive properties of easily re-entrained materials if necessary to minimize re-entrainment.)

An alternative approach to flue gas conditioning is to modify the method of electrical operation, either by intermittent energization or by pulse charging. In these, instead of the continuous production of corona, the precipitator is pulsed, such that ions are only produced during the application of the voltage, thereby enabling the charge on the particles reaching the collector to discharge prior to the next set of ions arriving.[19] The capacitive component of the precipitator maintains the field voltage high enough to precipitate particles during the nonpulsing phase.

SIZING OF ELECTROSTATIC PRECIPITATORS

The development of computerized programs for resolving the Poisson and La Place equations has greatly assisted in gaining a better understanding of the physics of precipitation. To date, however, no computer program is commercially available for determining precipitator design/sizing parameters from first principles. In practice, most precipitators are sized by the suppliers from interpretation of test data from plant operating under similar duties.

Deutsch,[20] working in the mid-1920s, proposed that the performance or collection efficiency of a precipitator took the form of an exponential equation:

$$\text{Efficiency} = 1 - e^{-x} \tag{15}$$

where x is dependent on factors related to the precipitator design and physical properties of the gas and dust.

In deriving the above relationship Deutsch assumed, with infinite turbulence producing a homogeneous distribution of the particles within the gas stream, that the particles were fully charged and the gas velocity was uniform, as was the corona current distribution on the collector plates; none of which are strictly true in practice. During his investigations, Deutsch, by changing the gas velocity through the precipitator and keeping all other factors constant, derived the following efficiency relationship:

$$\text{Efficiency} = 1 - e^{-2l/v} \tag{16}$$

where l is the field length and v the gas velocity. It should be recognized that l/v is the reciprocal of the precipitation contact time.

The above relationship was later transposed by the precipitation industry and used for many years to relate the size of the precipitator, collection efficiency, and gas flow rate for plants operating on similar duties and inlet conditions. Although the formula is often assumed to be theoretically based, it is no more than a useful method of comparing precipitator performance levels.

The relationship is probably more recognizable in the following form:

$$\text{Efficiency} = 1 - e^{-(A\omega/V)} \tag{17}$$

where ω is the effective migration velocity (m/s), A the area of collecting plate (m^2), and V the gas flow rate, (m^3/s).

Rearranging

$$\omega = \log_e\left(\frac{1}{1 - \text{efficiency}}\right) \times V/A \tag{18}$$

It is important to note that the value ω, the effective migration velocity, derived from Deutsch, is not equal to the theoretical value ω_{th}, derived from Eq. (14).

PRACTICAL APPROACH TO INDUSTRIAL PRECIPITATOR SIZING

Although the Deutsch relationship was used for many years as a design tool for comparing and sizing precipitators to cater for different gas flow rates and efficiencies, it was not until the 1960s, when difficulties arose in meeting more stringent emission requirements, that modifications to the traditional Deutsch relationship were considered. The lower emission requirements demanding more of the finer, more difficult, particulates were collected.

One of the better known modifications was that derived by Matts and Onhfeldt,[21] who produced the following relationship, termed the modified Deutsch formula, for deriving an improved working figure:

$$\omega_k = \log_e\left(\frac{1}{1 - \text{efficiency}}\right)^2 \times V/A \tag{19}$$

where ω_k is termed the modified effective migration velocity.

An alternative approach is that proposed by Petersen,[22] who developed the following relationship:

$$\log_e\left(\frac{1}{1 - \text{efficiency}}\right) = (1 + b\omega A/V)^{1/b} \tag{20}$$

where b is an empirically derived constant (=0.22), which holds true for both the power and cement industry applications.

The practical implication of these modified approaches is that, for a change in efficiency, based on an installed plant whose operational data is known, the required plate area increase is significantly higher than that derived from the straight Deutsch formula, because the need to remove the finer particulates in the penetration window.

The recently updated and released Precipitator Computer Programme ESPVI Series 4W[23] for power plant applications can accurately track the performance of existing precipitators, because its algorithms use the established mathematical physics reviewed earlier. In its operation, the program divides the particle sizing into 27 discrete fractions and derives

the charging and migration velocity for each, before summating them to derive the overall efficiency. By using theplant's design and operating data as input, the program can be used to quickly investigate any upgrade scenario and hence derive a low-cost approach to enhance the performance of any existing installation.[24]

VERSATILITY OF THE ELECTROSTATIC PRECIPITATOR

The use and versatility of the electrostatic precipitator can be appreciated from the following generalized applications.

1. Steam raising: utility and industrial boilers; firing (e.g., anthracite, bituminous, sub-bituminous, lignite and brown coals, heavy and light fuel oils, orimulsion, petroleum coke, etc.); biomass (wood, straw, chicken litter, grain husks, etc.).
2. Iron and steel manufacturing: blast furnace gas cleaning for combustion (basic iron and ferromanganese units); steel converters, ladles, and electric arc furnaces; scarfing and deseaming machines; sinter plants and iron ore pelletizers; foundry and cupola applications.
3. Metallurgical process plant: dryers, smelters, roasters, and refining on nonferrous plants (e.g., copper, lead, nickel, and zinc); gold and silver bullion refining operations.
4. Coal and gas operations: coal drying, carbonization and treatment, milling, and grinding; detarring, coal gasification, and distillation processes.
5. Cement and lime manufacture: vertical and rotary kilns—wet, semidry, and dry processing; clinker coolers, limestone crushing, etc.; raw meal mills, grinders, and general feed preparation units.
6. Waste incineration: municipal, chemical, clinical, and hazardous waste disposal units; sewage sludge fired installations.
7. Pulp and paper manufacture: black liquor and Kraft recovery furnaces.
8. Chemical processing and production: dust and fume from roasters, crushers, and dryers; fumes arising during the production of fine chemicals and dyes; mist collection from sulfuric and phosphoric acid plants; gas cleaning ahead of acid production units.
9. Miscellaneous applications: carbon black collection; catalyst recovery on refinery "cat crackers;" penultimate gas cleaning stage on uranium reprocessing plant; capture and recovery of oil mists; "clean room" applications.

This list does not cover all possible applications but relates to many common industrial processes. As each application produces different waste gas flows, temperatures, particulate inlet loading, etc., the precipitator size and its design tend to be site specific for a given duty.

As indicated, the performance of the electrostatic precipitator, whether it is in material recovery or for prevention of pollution, can deliver emissions down to the mg/Nm^3 region and, for air cleaning, emissions in the low $\mu g/Nm^3$ region.

CONCLUSIONS

Electrostatic precipitation is perhaps the most versatile and cost effective of all particulate collecting devices and can be applied to any process where there is a need to remove solid particulate and mist of fume sized particles from the gas stream, whether it be for recovery or pollution control duties. It can be designed to deliver any efficiency for any gas flow rate and temperature and has a low pressure drop and a life span of more than 20 years.

The process of collection can be considered in the following stages:

- The production of a corona field to create ions.
- The charging of the particles by the ions.
- The migration of the charged particles through the field.
- The arrival of the charged particle at the receiving electrode.
- The removal of the deposited particles from the receiving electrode.

Each of these has been addressed from both theoretical and practical aspects, as they relate to the design and performance of any installation, so as to give even a nonspecialist an idea of the role played by electrostatic precipitation, particularly for controlling emissions from coal fired electricity generation plant where collection efficiencies in the order of 99.8%+ are now required to satisfy environmental legislation.

REFERENCES

1. Grieco, G.J. Electrostatic precipitator electrode geometry and collector plate spacing—selection considerations. Proceedings PowerGen 94 Conference, Orlando, FL, U.S.A., December 1994; Paper TP 94-58.

2. Riehle, C. Basic and theoretical operation of ESPs (Chapter 3). In *Applied Electrostatic Precipitation*; Parker, K.R., Ed.; Blakie Academic and Professional: London, 1997; ISBN 0-7514-0266-4.
3. Parker, K.R. High frequency energisation systems (Chapter 8). In *Electrical Operation of Electrostatic Precipitators*; Parker, K.R., Ed.; IEE (UK) Power and Energy Series 41; U.K., 2003; ISBN 0-85296-137-5.
4. Gaugain, J.M. On the disruptive discharge. Annales de Chimie et de Physique **1862**, *64*, 175.
5. Rőentgen, V. Gőttinger Nach. **1878**, 390
6. Townsend, J.S. *Electricity in Gases*; Oxford University Press: Oxford, U.K., 1915.
7. Trichel, G.W. The mechanism of the negative point-to-plane corona near onset. Phys. Rev. **1938**, *54*, 1078.
8. Murphy, A.T.; Adler, F.T.; Penny, G.W. A theoretical analysis of the effects of an electric field on the charging of fine particles. Trans. AIEE **1959**, *78*, 318–326.
9. Liu, B.Y.H.; Yeh, H.C. On the theory of charging aerosol particles in an electric field. J. Appl. Phys. **1968**, *39*, 1396–1402.
10. Liu, B.Y.H.; Kapadia, A. Combined field and diffusion charging of aerosols in the continuum regime. J. Aerosol Sci. **1978**, *9*, 227–242.
11. Cochet, R. Lois Charge des Fines Particules (submicroniques) Etudes Theoriques—Controles Recents Spectre de Particules. Coll. Int. la Physique des Forces Electrostatiques et Leurs Application (Centre National de la Recherche Scientifique, Paris) **1961**, *102*, 331–338.
12. Hewitt, G.W. The charging of small particles for electrostatic precipitation. Trans. AIEE **1957**, *76*, 300
13. Baylis, A.P.; Russell-Jones, A. Collecting electrode rapping designed for high efficiency electric utility boiler electrostatic precipitators. Proceedings of the Fourth EPA/EPRI Symposium on the Transfer and Utilization of Particulate Control Technology; Houston, TX, U.S.A., October 1982; EPRI: Palo Alto, CA.
14. Parker, K.R. WESPS and fine particle collection. Proceedings of the Eighth ICESP Conference, Birmingham, AL, U.S.A., May 2001; Paper B5-1.
15. Hein, A.G.; Gibson, D. Skewed gas flow technology improves precipitator performance—Eskom operating experience in South Africa. Proceedings of the Sixth ICESP Conference, Budapest, Hungary, 1996; 238–243.
16. ICAC Publication No. EP 7. *Gas Flow Model Studies*; Institute of Clean Air Companies, 1993.
17. Schwab, M.J.; Johnson, R.W. Numerical design method for improving gas distribution within electrostatic precipitators. Proceedings of the American Power Conference 56th Annual Meeting, Chicago, U.S.A., April 25–27, 1994.
18. Chandan, R.; Parker, K.R.; Sanyal, A. A review of flue gas conditioning technology in meeting particulate emission standards. Proceedings of the Electric Power 1999 Conference, Baltimore, U.S.A., 1999.
19. Parker, K.R. Alternative mains frequency energisation systems (Chapter 7). In *Electrical Operation of Electrostatic Precipitators*; Parker K.R.; IEE (UK) Power and Energy Series 41: U.K., 2003; ISBN 0-85296-137-5.
20. Deutsch, W. Bewegung und Ladung der Elektrizitatstrager im Zylinderkondensator. Ann. Der Phys. **1922**, *68*, 335–344.
21. Matts, S.; Onhfeldt, P.O. Efficient gas cleaning with SF electrostatic precipitators. Flakten **1963/4**, *1–12*, 93–110.
22. Petersen, H.H. A precipitator sizing formula. Proceedings of the Fourth ICESP Conference, Beijing. China, September 1990; Int. Academic Pub.: Beijing, 1993; 330–338.
23. Parker, K.R.; Plaks, N. *Electrostatic Precipitator Training Manual*. US Environmental Protection Agency Report No. EPA-600/R-04/072, July 2004.
24. Parker, K.R.; Plaks, N.; Zykov, A.M.; Kolchin, K.I.; Konovalov, V.K. ESP performance enhancement analysis by advanced modelling techniques. Proceedings Electric Power 2004 Conference, Baltimore, U.S.A., April 2004.

Emulsion Polymerization

Vincent G. Gomes
University of Sydney, Sydney, New South Wales, Australia

INTRODUCTION

A polymer [poly (many) + meros (particles)] consists of a large number of structural units formed by covalent bonding of monomer units to create larger molecular chains having relatively high molecular weights. The polymer industry is a multibillion dollar business globally (OECD countries alone produce over 100 MT/pa), growing at a relatively fast rate. Polymers are used in almost every major market sectors including construction, packaging, automobiles, aerospace, shipping, electrical/electronics, food, medical, sporting, and a variety of consumer goods, just to mention a few. Emulsion polymerization is a versatile process and is one of the several fundamental routes to produce nano- to microstructured materials. It is one of the principal techniques for polymerization among bulk, solution, suspension, and emulsion. In bulk and solution polymerization all reactions proceed in a single phase.

Bulk polymerization is the simplest, involving monomer and monomer-soluble initiator, without solvents or dispersion media. Monomer conversion causes rapid increases in viscosity, making reactor agitation and heat transfer inefficient. This may cause reactor thermal runaway. The difficulties associated with bulk polymerization can be overcome by solution polymerization, which is carried out with a solvent to reduce mixture viscosity, thereby ensuring better heat removal and mixing. However, chain transfer to solvent may reduce the chain length and, hence, the degree of polymerization. The complete removal of solvent from the final product is difficult and economically impractical. Thus, solution polymerization is mainly used where the polymer is required in a solution. An effective means of reducing the problems with bulk polymerization is to use emulsion polymerization, which is carried out in monomer droplets suspended by colloid stabilizers in water. The process facilitates heat and mass transfer and is used in industrial scale. Suspension polymerization has similar advantages as for emulsion; however, the particles generated are larger (10–1000 μm), while monomer conversion, and heat and mass transfer efficiencies are lower. An emulsion is a discontinuous liquid phase dispersed within a different continuous liquid phase. Emulsion polymerization is based on a system consisting of water as the continuous phase, one or more monomer(s), emulsifier(s) dispersed in immiscible monomer(s), and an initiator that initiates the reaction by producing highly reactive radicals in the aqueous phase. The end product of this process is a finely divided latex comprising polymer particles (0.05–0.3 μm) dispersed in a continuous phase. The particles form a colloid that can be used either in latex form (adhesives, surface coatings, etc.) or coagulated and dried for other applications (fibers, tires, etc.). In the related dispersion polymerization, the monomer and the initiator are soluble in the continuous phase and the polymer particles produced, precipitate from the mixture.

Emulsion polymerization is the basis of many industrial processes, and the production volume of latex technologies is continually expanding—a consequence of the many environmental, economic, health, and safety benefits the process has over solvent-based processes. A wide range of products are synthesized by emulsion polymerization, including commodity polymers, such as polystyrene, poly(acrylates), poly (methyl methacrylate), neoprene or poly(chloroprene), poly(tetrafluoroethylene), and styrene–butadiene rubber (SBR). The applications include manufacture of coatings, paints, adhesives, synthetic leather, paper coatings, wet suits, natural rubber substitutes, supports for latex-based antibody diagnostic kits, etc.[1]

Scope

Emulsion polymerization is in wide commercial use because of its substantial advantages:

- Heat generated by polymerization can be readily absorbed and dissipated by the aqueous phase, making it safer.
- The rate of polymerization is high compared to other processes, and a relatively high conversion can be achieved.
- The polymer is formed as low-viscosity latex rather than as the solid or viscous solution as in bulk and solution polymerization.
- Molecular weight of polymer is readily controlled in emulsion polymerization, where the average number of radicals, $\bar{n}$, can be kept constant over an extended period of the reaction, and higher molecular weights

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007749

can often be achieved because of the compartmentalization of growing chains.
- The process itself and the resulting polymer latex are water-based rather than solvent-based, which reduces both safety hazards and environmental impact.

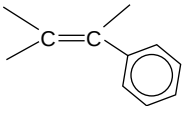
The disadvantages of emulsion polymerization are:

- Stabilizers and other additives may impair the quality of the final product.
- It may be necessary to separate the polymer from the water for further processing (e.g., by coagulation/dewatering) at increased expense.
- As the polymerization process is heterogenous and involves a minimum of two phases, mechanisms are complex and may pose operational challenges.

REAGENTS

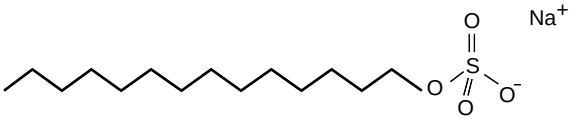
Most emulsion polymerization is based on free-radical reactions, involving monomers (e.g., styrene, butadiene, vinyl acetate, vinyl chloride, methacrylic acid, methyl methacrylate, acrylic acid, etc.), surfactant (sodium dodecyl diphenyloxide disulfonate), initiator (potassium persulfate), water (18.2 MΩ/cm), and other chemicals and reagents such as sodium hydrogen carbonate, toluene, eluent solution, sodium chloride, and sodium hydroxide.

Monomers must be bifunctional, e.g., styrene:



Surfactants or emulsifiers help stabilize the emulsion and are classified into four broad categories: anionic, cationic, nonionic, and zwitterionic. Surfactants are dissolved in water at low concentrations, where they form aggregates or micelles. At a concentration greater than their critical micelle concentration (CMC), all excess molecules form micelles.

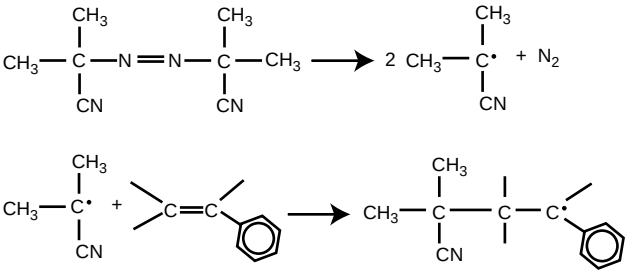
Typical surfactant: Sodium dodecyl sulfate (CMC of 6×10^{-3} mol/L at 80°C), a common alkyl sulfate of the anionic surfactant family.



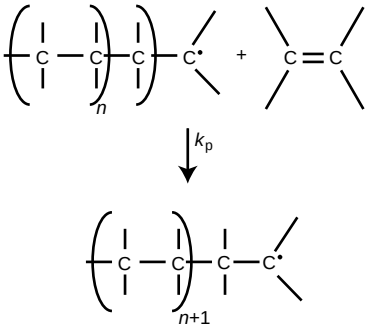
Typical initiator: $K_2S_2O_8$ or potassium persulfate.

Typical polymerization chemistry:

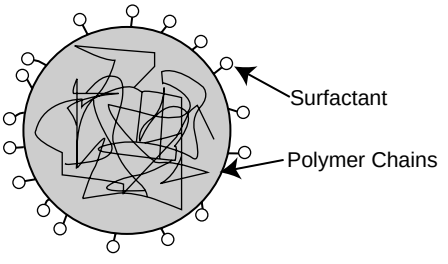
Free-radical polymerizations begin with initiation for example, for AIBN as initiator and styrene as monomer:



This is followed by propagation:



Product latex particle:



Magnified Latex Particle

A typical emulsion polymerization recipe includes specific proportions of the added ingredients, e.g. (in wt%): monomer, 100; water, 150; initiator, 0.5; surfactant, 5. Because the monomer has low water solubility, it is clear that there will be two separate phases referred to as the monomer phase and the aqueous phase. The aqueous phase, containing the surfactant in the form of micelles, can be considered as consisting of two phases, the micellar phase and the true aqueous phase. The emulsifier helps disperse the monomer in the aqueous phase with droplets in the order of a few micrometers in size. The hydrophobic interior of the micelles contains solubilized monomer, which is apportioned by diffusion out of the emulsified monomer droplets and through the aqueous phase.

REACTION KINETICS

The kinetics of emulsion polymerization is complex, involving a large number of species and at least two phases. The first quantitative approach to emulsion polymerization kinetics led to extensions by many others.[2] The important events to consider are: 1) the free-radical reactions of chain formation: initiation, propagation, chain transfer, and termination and 2) the phase transfer events that control particle formation: radical entry into particles from the aqueous phase, radical exit into the aqueous phase, radical entry into micelles, and the aqueous phase coil–globule transition. In free-radical emulsion polymerization, the fundamental steps are shown schematically in Fig. 1

Initiation

Initiation is the reaction step when free radicals are formed. In emulsion polymerization, the initiator is usually soluble in the aqueous phase, e.g., potassium persulfate. In the following reaction, the initiator decomposes to make two identical free-radical entities, denoted by $I^{\bullet}$:

$$I \xrightarrow{k_d} 2I^{\bullet}$$

where I stands for initiator and k_d is the rate coefficient for initiator decomposition. The initiation reaction is as follows:

$$I + M \xrightarrow{k_1} R_1^{\bullet}$$

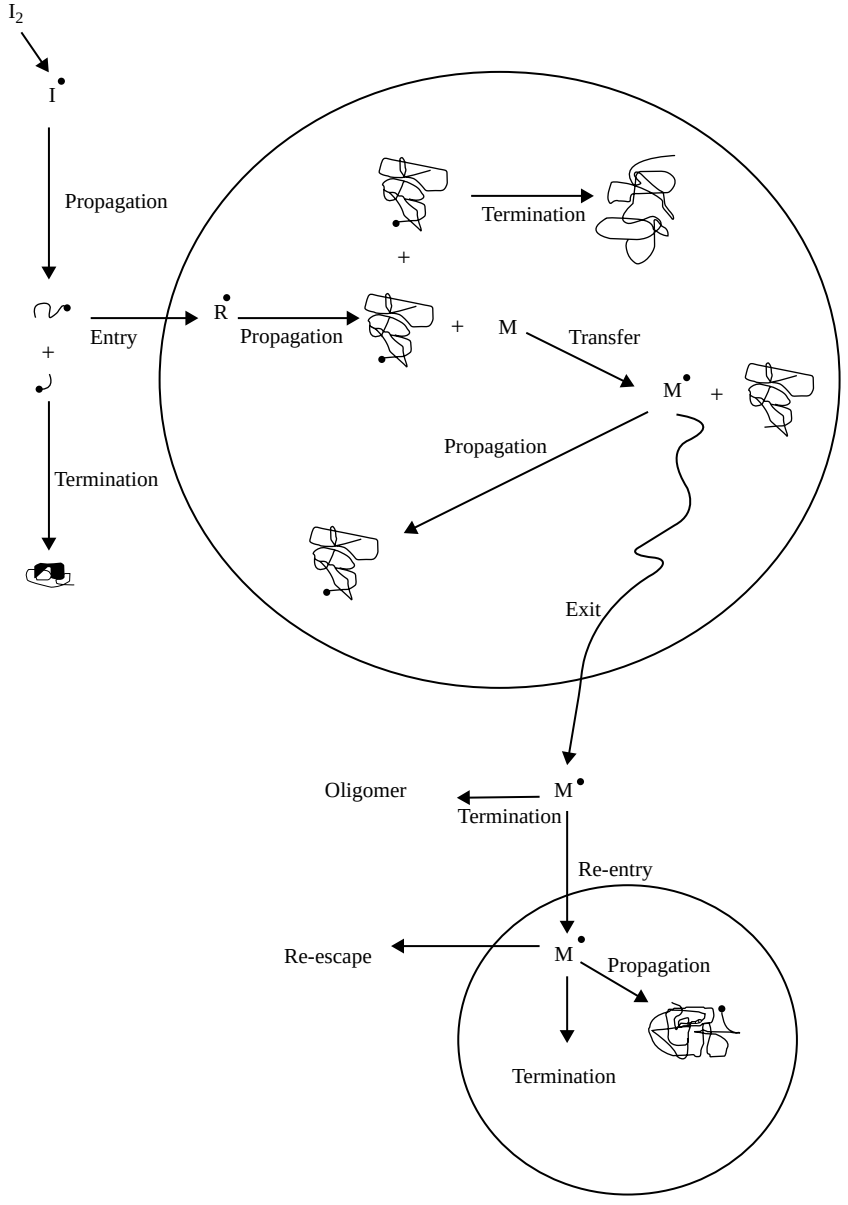


Fig. 1 Subprocesses occurring during emulsion polymerization. (From Ref.[3].)

where M is a monomer unit, $R_1^\bullet$ is a monomeric radical, and k_1 is the rate coefficient for the propagation of an initiator radical with a monomer molecule.

Propagation

Following the radical formation step, the propagation step takes place with the addition of polymer radicals to the monomers. The chemical equation for propagation is:

$$R_n^\bullet + M \xrightarrow{k_P^n} R_{n+1}^\bullet$$

where $R_n^\bullet$ is a radical of degree of polymerization n (i.e., a polymeric radical comprising n monomer units), and k_p^n is the rate coefficient for propagation of a polymeric radical having the degree of polymerization n.

Chain Transfer

Transfer occurs when radical activity is transferred from a polymeric radical to a monomer, polymer, initiator, solvent (or a chain transfer species). The mechanism of transfer to a monomer is given as:

$$R_n^\bullet + M \xrightarrow{k_{tr}} P_n + M^\bullet$$

where P_n is a dead polymer chain of degree of polymerization n, and k_{tr} is the rate coefficient for transfer to a monomer unit. The reaction takes place via hydrogen abstraction.

Termination

Termination occurs by two distinct mechanisms: combination and disproportionation. In combination, a polymeric radical of degree of polymerization n adds to a polymeric radical of degree of polymerization m, forming a dead polymeric chain of degree of polymerization $n + m$:

$$R_n^\bullet + R_m^\bullet \xrightarrow{k_{tc}} P_{n+m}$$

where k_{tc} is the rate coefficient for termination by combination.

The second mechanism, disproportionation, takes place as a result of a hydrogen atom abstraction from one radical to another. This results in two dead polymer chains:

$$R_n^\bullet + R_m^\bullet \xrightarrow{k_{td}} P_n + P_m$$

where k_{td} is the rate coefficient for termination by disproportionation.

Accurate kinetic models provide insight into the relationship between process variables and the product characteristics that they influence. Different emulsion polymerization theories, although controversial, were used over the years to aid in interpreting polymerization experiments. These models at the early stages focused only on the emulsion kinetics until recent research began to couple these kinetic models with reactor dynamics. These studies have helped to elucidate the various competing mechanisms involved, and led to mathematical models (with varying degrees of complexity) that are able to predict the key product attributes.[4–10]

The rate coefficient for propagation (k_p), essential for quantitative interpretation of rate data, is studied by pulsed laser polymerization, a technique that has now become well established. Termination coefficients are established by relaxation experiments performed with γ-radiolysis initiated free-radical polymerization with large latex particles, while analysis of the molecular weight distribution (MWD) under transfer-dominated conditions yields rate coefficients for chain transfer. Radical exit rate coefficients from small particles are measured by γ-radiolysis relaxation experiments, and in combination with the steady-state reaction rate and previously measured exit rate coefficients can be used to determine the rate of radical entry.[11] Pulsed laser polymerization uses periodic flashes of laser light to generate bursts of radicals. At each burst, a number of chains are initiated, and the growing chains are terminated by the small free-radical species. If a sufficient number of chains are initiated and terminated by successive laser pulses, the point of inflection on the gel permeation chromatography (GPC) trace of the polymer product corresponds to the molecular weight of these chains.[12]

Termination rate coefficients can be measured using the γ-radiolysis relaxation method. This involves initiation using γ-radiation, followed by removal of the reaction vessel from the γ-source. Conversion during the relaxation period is monitored by dilatometry, and the decay in polymerization rate over time is related to the rate of radical loss.[12] When large particles are used, radical loss is dominated by intraparticle termination, rather than exit into the aqueous phase, and the rate coefficient for termination can be determined from the decay curve. By using multiple insertions and removals, the termination rate coefficient is determined over a wide range of polymer mass fraction (w_p).

Most kinetic studies focus on batch emulsion polymerization. These studies enable estimation of important polymer properties. Only recently, the control of particle size and MWD described by population balance models has been achieved.[13–17] The differences between emulsion polymerization and copolymerization

behavior in batch and semibatch processes have remained unexplored in the literature until recently. It has been accepted that polymer particle formation and growth in batch or semibatch modes obey the same kinetics. No significant difference in average particle size has been reported.[18] However, recent work noted an increase in particle number with a decrease in the monomer feed rate in semibatch emulsion polymerization.[2,19] It was observed that increasing the feed rate results in a larger particle size and at sufficiently low rates, multimodal distributions are produced.

The early kinetic model by Smith and Ewart was based on Harkin's mechanistic understanding of the batch process.[11] The particle population balances were written for a stationary state assuming that the rate of formation of particles with n radicals equals the rate of their disappearance (see equation at the bottom of this page). Where R_i is the rate of radical entry into a particle (m^{-3}/sec); k_0 is the rate constant for radical exit (m/sec); S is the particle surface area (m^2); k_{tp} is the rate constant for bimolecular termination in the particles (m^3/sec) and υ is the particle volume. According to Smith and Ewart three limiting cases can be identified:

Case I: Where the rate of radical exit from the particle is greater than the rate of radical entry

$$\frac{k_0 S}{\upsilon} \gg \frac{R_i}{N_c}$$

In this case the number average radicals per particle $\bar{n}$ is less than unity:

$$\bar{n} = \frac{\sum nN_n}{\sum N_n}$$

Case II: The rate of termination is greater than the rate of entry and greater than the rate of exit

$$\frac{k_{tp}}{\upsilon} \gg \frac{R_i}{N_c} \gg \frac{k_0 S}{\upsilon}$$

This case gives $\bar{n} = 1/2$ on the basis that entry of a radical into an empty particle will result in a particle having one propagating chain, whereas the entry into a particle containing a radical will cause instantaneous termination. Therefore, the number of radicals in a particle can be either zero or one.

Case III: The particles contain more than one radical and

$$\frac{k_{tp}}{\upsilon} \leq \frac{R_i}{N_c} \gg \frac{k_0 S}{\upsilon}$$

Thus, the kinetics in the particles resembles bulk polymerization.

With these sets of equations the reaction rate, R_p, is calculated from:

$$R_p = k_p M_p \frac{N_c \bar{n}}{N_A}$$

where k_p is the reaction rate coefficient (m^3/mol/sec), N_A is Avogadro's number, and M_p is the monomer concentration in the particles (mol/m^3).

In a similar manner, the particle number N_c can be calculated via two models:

Model I: Where all the radicals generated by the initiator in the aqueous phase nucleate the micelles. Therefore, the rate of formation of particles equals the initiation rate. Consequently,

$$N_c = 0.53\left(\frac{R_{i,w}}{\mu}\right)^{2/5}(a_s[S])^{3/5}$$

where μ is the volume growth rate of the particle (m^3/sec), $R_{i,w}$ is the rate of initiator decomposition (mol/m^3/sec), a_s is the emulsifier covering capacity per mole of emulsifier, and [S] is the surfactant concentration (mol/m^3).

Model II: Where particles and micelles absorb radicals at a rate proportional to their current interfacial areas. Thus,

$$N_c = 0.37\left(\frac{R_{i,w}}{\mu}\right)^{2/5}(a_s[S])^{3/5}$$

Extensions of the early model employed mass balances across all components.[20] The model agreed well with experimental data for the production of polystyrene. However, as in most traditional models it was a particle number model, based on all particles being identical. These models helped analyze polymerization in terms of the number concentration of polymer particles containing "i" actively polymerizing radicals of number, $N_i(t)$. As a result, they contained no particle size

$$\underbrace{\frac{R_i}{N_c}N_{n-1}}_{\text{(entry)}} + \underbrace{\frac{k_0 S}{\upsilon}(n+1)N_{n+1}}_{\text{(exit)}} + \underbrace{\frac{k_{tp}}{\upsilon}(n+2)(n+1)N_{n+2}}_{\text{(termination)}} = \underbrace{\frac{R_i}{N_c}N_n}_{\text{(entry)}} + \underbrace{\frac{k_0 S}{\upsilon}nN_n}_{\text{(exit)}} + \underbrace{\frac{k_{tp}}{\upsilon}(n)(n-1)N_n}_{\text{(termination)}}$$

parameter and failed to take into account the volume dependence of the rate coefficients.

Later research extended the Smith–Ewart theory and developed models incorporating the complete particle size distribution (PSD).[21] In particular, the volume of latex particles was included, enabling computation of the density of polymer particles (containing "*i*" radicals) as a function of volume and time. The particles were then compartmentalized into discrete volumes studied over a discrete time interval. These models relied on mass balances over the concentration of particles of a specified diameter at a given time. They included terms for nucleation, growth, and coagulation of particles. The result was a series of coupled partial differential equations, which were solved numerically. Recent advances in the understanding of the physicochemical mechanisms have eliminated the controversy and uncertainties accompanying previous findings. These models incorporated the competing events, such as aqueous phase propagation and termination, and micellar and secondary nucleation.[4,22]

MECHANISTIC MODELS FOR DESIGN AND SCALE-UP

Mathematical modeling is an essential tool for design and scale-up. As emulsion polymerization involves a minimum of two phases, the kinetics is complex and makes it difficult to:

- Predict the effect of changes in process variables on product.
- Optimize the process to achieve the desired product characteristics and efficient use of raw materials.
- Control reaction in a safe and cost-effective manner.

The emulsion polymerization process is analyzed in terms of the stages or intervals it involves: typically represented by Intervals I, II, and III.

Interval I

Particle formation is the defining event of Interval I (Table 1, Fig. 2). Micelles may be present (if the surfactant concentration is sufficiently high for micelles to form), monomer droplets are present, and particle number and particle size are both increasing. Particles form by two competing mechanisms: micellar nucleation or homogenous nucleation. Micellar nucleation occurs when an oligomeric radical from a radical enters a micelle, and a particle forms. Homogenous nucleation occurs when an oligomeric radical is no longer soluble in the aqueous phase and collapses to form a particle. Largely because of the increasing number of particles the rate of polymerization increases. Interval I is the first stage of ab initio polymerization (i.e., where no particles are initially present) and can be avoided by using a preformed latex to give a seeded system, in which case, polymerization may begin in Interval II or III.

During Interval I, the initiator decomposes in the water phase to generate active radicals. The most commonly used initiator is potassium persulfate, $K_2S_2O_8$, which decomposes as follows:

$$S_2O_8^{2-} \rightarrow 2SO_4^{-\bullet}$$

The active radical propagates with the monomer dissolved in water to form oligomeric radicals, soluble in water

$$2SO_4^{-\bullet} + M \rightarrow {}^{\bullet}MSO_4^{-}$$
$${}^{\bullet}MSO_4^{-} + M \rightarrow {}^{\bullet}M_2SO_4^{-}$$

This may undergo termination with another radical:

$${}^{\bullet}MSO_4^{-} + {}^{\bullet}M_2SO_4^{-} \rightarrow M_3(SO_4^{-})_2$$

When an oligomeric radical enters a micelle, it propagates rapidly with solubilized monomer to form a polymer particle. Similarly, new polymer particles are continuously formed with the new particles being small compared to the larger early particles. Thus, the PSD is broad at this early stage. As the total interface area of micelles is much larger than that of the monomer droplets, the oligomeric radicals in the aqueous phase are much more likely to diffuse into a micelle swollen with monomer rather than into the monomer droplet. Polymerization thus occurs mainly in the micelles (which gradually transform into polymer particles) and the polymer latex particles, while

Table 1 Description of the three intervals of emulsion polymerization

Interval	Conversion range (%)	Micelles	Monomer droplets	Particle number	Particle size	Comments
I	0–10	Present	Present	Increases	Increases	Nucleation period
II	10–40	Absent	Present	Constant	Increases	C_p constant
III	40–100	Absent	Absent	Constant	Roughly constant	C_p decreases

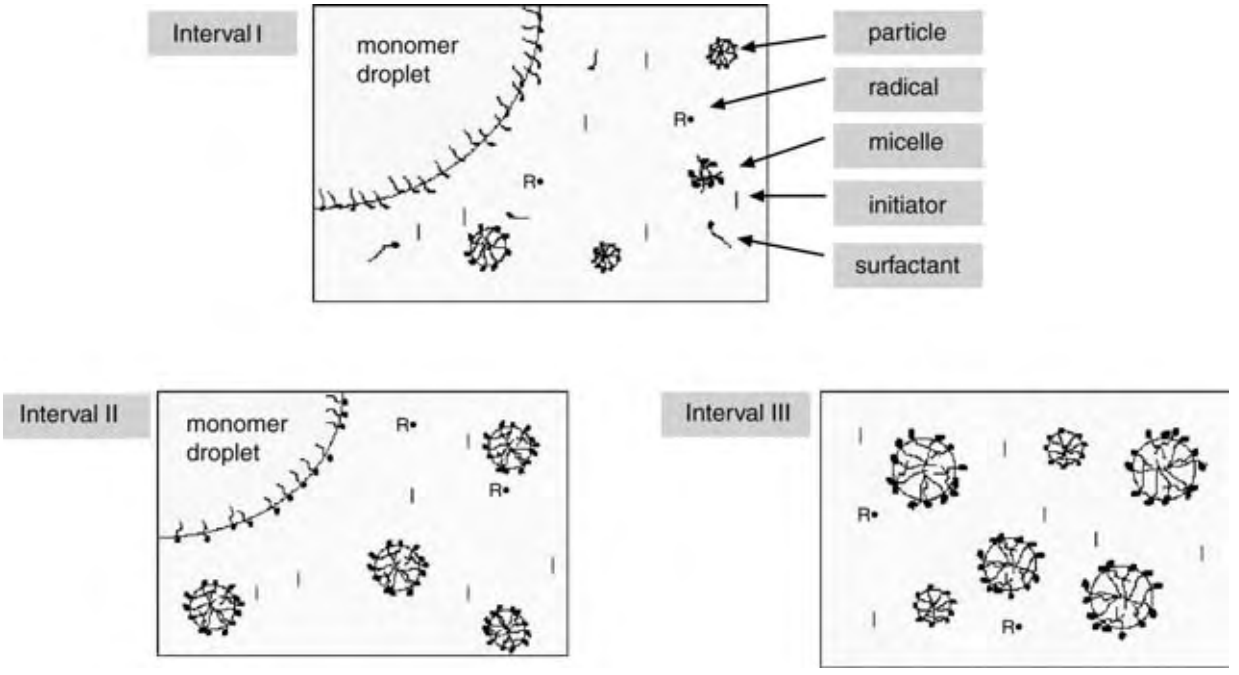


Fig. 2 Three intervals during emulsion polymerization. *(View this art in color at www.dekker.com.)*

continually consuming monomers, delivered via diffusion from the droplets through the aqueous phase. There are eventually a sufficient number of large-size particles so that all newly formed aqueous phase radicals enter preexisting particles rather than nucleating new ones. At the completion of this stage, all micelles are consumed by nucleation and the when surfactant concentration falls below the CMC, particle formation ceases and the number of latex particles stabilizes.

Interval II

When particle formation ceases, Interval II begins. At this point both surfactant-stabilized monomer droplets and polymer particles are present in the emulsion. Polymerization continues in the particles with the process of initiation and termination maintaining a stable or a slowly increasing $\bar{n}$. The particles continue to grow during this interval, as monomers continue to migrate from the monomer droplets into the particles to replenish that which is being polymerized. This migration of monomer normally maintains the concentration of monomer in the particles at the saturation level, C_P^{sat}. As C_P and N_c do not change in Interval II, this interval is associated with a relatively constant rate of polymerization.

The polymer chain within the particle keeps growing by propagation and eventually undergoes radical transfer forming a new monomeric radical:

$$M_n^{\bullet} + M \rightarrow M_n + M^{\bullet}$$

The monomeric radical continues to propagate, or escape from the particle into the aqueous phase. The exited radical may enter another particle and terminate the growth of a propagating chain in this particle, or propagate if the particle contains no growing chains. As a result of these processes of entry, transfer, and exit, the particle contains a large number of dead polymer chains with a high molecular weight (of the order of 10^6).

Interval III

When the monomer droplets are depleted, Interval III commences. The majority of the remaining monomer is in the latex particles with some dissolved in the aqueous phase. Particles are still able to grow, but the decreasing concentration of the monomer leads to a gradual decrease in the rate of polymerization (unless reduced diffusion of radicals because of the increasing weight fraction of polymers leads to an increase in $\bar{n}$). Polymerization will usually continue until all the monomer has reacted.

The heterogenous nature of emulsion polymerization processes, and the stepwise mechanisms involved add complications to the understanding of the kinetics. Despite the investigations conducted in emulsion polymerization over the last few decades, the scientific representation of these complex processes remains incomplete. With recent advances in fundamental chemistry and colloid science, model uncertainties have diminished.

KINETIC REGIMES

Two types of systems are identified in emulsion polymerization: the pseudo-bulk system and the zero–one system.[1]

Pseudo-Bulk System

In this system the number of radicals in a particle is relatively so high that the polymerization resembles bulk polymerization. The average number of radicals per particle, $\bar{n}$, is, almost, always greater than 0.5. Compartmentalization has no effect on the kinetics of a pseudo-bulk system, and termination, which is rate determining, is always diffusion controlled.

Zero–One System

The term zero–one designates that all latex particles contain either zero or one active free radical. The entry of a radical in a particle that already contains a free radical will instantaneously cause termination. Thus, the maximum value of the average number of radicals per particle, $\bar{n}$, is 0.5. In a zero–one system, compartmentalization plays a crucial role in the kinetic events of emulsion polymerization processes. In fact, a radical in one particle will have no access to a radical in another particle without the intervention of a phase transfer event. Two radicals in proximity will terminate rapidly; however, the rate of termination will be reduced in the process because of compartmentalization, as the radicals are isolated as separate particles. Consequently, the propagation rate is higher and the molecular weight of the polymer formed is larger than in the corresponding bulk systems. Which model is more appropriate depends primarily on the particle size. Small particles tend to satisfy the zero–one model, as termination is likely to be instantaneous.[1]

Mass balances on the aqueous radical species lead to the following set of rate equations:

$$\frac{d[I^\bullet]}{dt} = 2k_d[I] - k_{p1}[I^\bullet]C_w \quad (1)$$

$$\frac{d[IM_j^\bullet]}{dt} = k_{p1}[I^\bullet]C_w - k^1_{p,aq}[IM_1^\bullet]C_w - k_{t,aq}[IM_1^\bullet][T^\bullet] \quad (2)$$

$$\frac{d[IM_i^\bullet]}{dt} = k^{i-1}_{p,aq}[IM_{i-1}^\bullet]C_w - k^i_{p,aq}[IM_i^\bullet]C_w - k_{t,aq}[IM_i^\bullet][T^\bullet], \quad i < z \quad (3)$$

$$\frac{d[IM_i^\bullet]}{dt} = k^{i-1}_{p,aq}[IM_{i-1}^\bullet]C_w - k^i_{p,aq}[IM_i^\bullet]C_w - k_{t,aq}[IM_i^\bullet][T^\bullet] - k^i_e[IM_i^\bullet]\frac{N_{tot}}{N_A} - k^i_{e,micelle}C_{micelle}[IM_i^\bullet], \quad i = z, \ldots, j_{crit} - 1 \quad (4)$$

In the equations above, C_w is the monomer concentration in the water phase, N_{tot} is the total number of latex particle per litre of aqueous phase, and N_A is Avogadro's number.

Furthermore, the monomeric radicals formed by radical transfer to the monomer, $M^\bullet$may desorb from the particle with such "exited" radicals denoted by $E^\bullet$. Once such a desorbed radical meets the surface of a particle, it will immediately penetrate the particle because of its lipophilic nature. This adsorption–desorption process is reversible, as illustrated below:

$$E^\bullet + \text{particle} \underset{k_{dM}}{\overset{k_{cE}}{\rightleftarrows}} (\text{particle} - M^\bullet)$$

Hence, the total radical concentration in the aqueous phase $T^\bullet$ becomes:

$$[T^\bullet] = \sum_{i=1}^{z-1}[IM_i^\bullet] + [E^\bullet] \quad (5)$$

The rate of surfactant consumption determines the micelle concentration, $C_{micelle}$. The amount of adsorbed surfactant onto the particle surface is defined as

$$S_{ads} = \frac{4\pi r_s^2 N_{tot}}{N_A a_s} \quad (6)$$

and hence

$$C_{micelle} = \text{maximum}\left[0, \frac{[S_{added}] - S_{ads} - [CMC]}{n_{agg}}\right] \quad (7)$$

where $[S_{added}]$ is the total concentration of surfactant added, a_s is the area occupied by an adsorbed surfactant molecule, n_{agg} is the average number of surfactant molecules in a micelle (known as the micellar aggregation number), and [CMC] is the critical micelle concentration. Eq. (7) designates that particle formation via micelles stops when the surfactant concentration falls below the CMC. This equation also shows the critical relationship existing between the surfactant concentration and the number of particles, N_{tot}. Fig. 3 shows a schematic diagram of radical transfer between particles of unswollen volumes based on the zero–one kinetics.

Fig. 3 depicts the dominant mechanisms of micellar and homogenous nucleation, where an oligomer of degree greater than or equal to the critical degree of polymerization for entry, z, creates the latex particle containing a polymeric radical. The critical degree of polymerization is the degree in which the monomer while propagating in the aqueous phase acquires enough hydrophobicity to enter into micelles. For micellar nucleation a particle containing a radical of

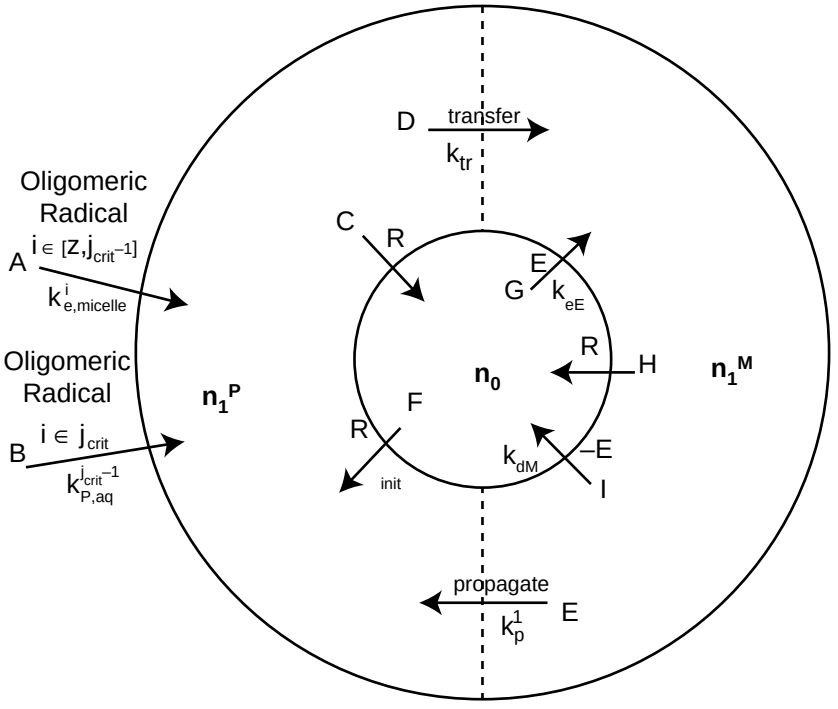


Fig. 3 Schematic representation of radical transfer between particles of unswollen volume based on zero–one kinetics.

degree $i \in [z, j_{crit}-1]$ is formed, with rate constant $k^{i}_{e,micelle}$(section A). Homogenous nucleation occurs on the formation of an oligomeric radical of degree j_{crit}, and hence forms with rate constant $k^{j_{crit}-1}_{p,aq}$ (section B).

Polymeric radical containing particles are also formed when an oligomeric radical enters an n_0-type particle with rate constant ρ_{init} (section F) and when an n_1^M-type particle propagates (rate constantk^1_p) to convert the monomeric radical into a polymeric radical (section E). Particles with polymeric radicals are consumed when either an exited or an oligomeric radical enters an existing n_1^P particle, causing termination (section C). This occurs with rate constant ρ, the pseudo first-order rate constant that accounts for all entry events. Alternatively, chain transfer to a monomer (section D) consumes them.

POPULATION BALANCES

The main assumptions in developing dynamic model equations are:[23]

- Polymerization occurs only within latex particles.
- Particles are formed through both homogenous and micellar nucleation.
- ''Competition'' occurs between the processes of particle growth, homogenous nucleation, and micellar nucleation.
- Three types of particles may be identified: 1) those containing no radicals—referred to as type n_0 particles; 2) those containing one monomeric radical—type n_1^M particles; and 3) those containing one polymeric radical—type n_1^P particles.
- Compartmentalization plays a crucial role in the overall kinetics, as a radical in one particle will have no access to a radical in another particle without the intervention of a phase transfer event.
- Two radicals in close proximity will terminate rapidly.
- The entry of a radical into a particle that already contains a free radical will instantaneously cause termination. Thus, the maximum value of the average number of radicals per particle, $\bar{n}$, is 0.5.
- Monomer diffusion inside a particle is a function of the polymer volume fraction.
- At low monomer concentration (that is, at high monomer conversion), entry, propagation, and termination become diffusion-controlled processes.
- Particle coagulation follows the DLVO theory.

Given these assumptions, the population balance equations for the three particles types (n_1^P, n_0, and n_1^M) are as follows:

$$\frac{\partial n_1^P(V,t)}{\partial t} = k_P^1 C_P n_1^M + \rho_{init} n_0 - \rho n_1^P - k_{tr} C_P n_1^P - \frac{\partial (K n_1^P)}{\partial V} + \delta(V - V_0) \times \left[C_{micelle} \sum_{i=z}^{j_{crit}-1} \left[k^i_{e,micelle} [IM_i] \right] - k^{j_{crit}-1}_{p,aq} C_w [IM_{j_{crit}-1}] \right]$$

$$+ \int_0^\infty B(V, V - V')[n_0(V')n_1^{\mathrm{P}}(V - V') + n_1^{\mathrm{P}}(V')n_0(V - V')]\mathrm{d}V' - n_1^{\mathrm{P}}(V) \times \int_0^\infty B(V, V')[n_0(V') + n_1^{\mathrm{P}}(V')]\mathrm{d}V' \tag{8}$$

$$\frac{\partial n_0(V,t)}{\partial t} = \rho[n_1^{\mathrm{P}} + n_1^{\mathrm{M}} - n_0] + k_{\mathrm{dM}}n_1^{\mathrm{M}} + \int_0^\infty B(V, V - V')[n_0(V')n_0(V - V') + n_1^{\mathrm{P}}(V')n_1^{\mathrm{P}}(V - V')]\mathrm{d}V' - n_0(V) \times \int_0^\infty B(V, V')[n_0(V') + n_1^{\mathrm{P}}(V')]\mathrm{d}V' \tag{9}$$

$$\frac{\partial n_1^{\mathrm{M}}(V,t)}{\partial t} = k_{\mathrm{tr}}C_{\mathrm{P}}n_1^{\mathrm{P}} + k_{\mathrm{eE}}[\mathrm{E}]n_0 - \left(k_{\mathrm{p}}^1 C_{\mathrm{P}} + k_{\mathrm{dM}} + \rho\right)n_1^{\mathrm{M}} \tag{10}$$

$$n(V,t) = n_1^{\mathrm{P}}(V,t) + n_0(V,t) + n_1^{\mathrm{M}}(V,t) \tag{11}$$

These coupled partial integrodifferential equations comprise the ''evolution equations'' for the latex particles in an emulsion polymerization process. In this mathematical description, an n_1^{P}-type particle is formed when an oligomeric radical enters an n_0-type particle and the monomeric radical propagates in an n_1^{M}-type particle to form a polymeric radical. Particles containing such polymeric radicals are consumed when an oligomeric radical enters an existing n_1^{P}-type particle, causing instantaneous termination. This process occurs with a rate constant ρ, the pseudo first-order rate constant, which describes all entry events. Alternatively, radical transfer to a monomer molecule transfers radical activity, and the resulting monomeric radical may undergo subsequent propagation or termination.

The population of n_0-type particles is increased by the entry of radicals (oligomeric or exited) into $n_1^{\mathrm{M}}$and n_1^{P}-type particles. They are also formed when monomeric radicals exit an n_1^{M}-type particle (a process that occurs with a rate constant k_{dM}). The population of n_0-type particles is decreased when oligomeric radicals enter an existing particle (to form an n_1^{P}-type particle), or when an exited radical enters to form an n_1^{M}-type particle.

Finally, n_1^{M}-type particles may be formed by chain transfer within particles containing a polymeric radical, and via the entry of exited monomeric radicals into n_0-type particles. These n_1^{M} particles are consumed through the propagation of radicals in existing particles, by the desorption of monomeric radicals from existing n_1^{M} particles, and by the entry of any type of radical into existing n_1^{M}-type particles.

Growth only significantly affects n_1^{P}-type particles, as radicals propagate without changing the particle's identity.

The entry processes, and hence the rate coefficients, k_{e}^i and $k_{\mathrm{e,micelle}}^i$, are assumed to be diffusion controlled (with an exponent of $\frac{1}{2}$ describing the chain length-dependent diffusion coefficient of small radicals). Given that oligomeric radicals (of degree greater than or equal to z) can only enter a micelle or preexisting particle, the entry coefficients can be calculated as follows:

$$k_{\mathrm{e}}^i(V) = 4\pi r_{\mathrm{s}}N_{\mathrm{A}}e_{\mathrm{e}}\frac{D_{\mathrm{W}}}{i^{1/2}} \quad i \geq z, \qquad k_{\mathrm{e}}^i(V) = 0 \quad i < z \tag{12}$$

$$k_{\mathrm{e,micelle}}^i(V) = 4\pi r_{\mathrm{micelle}}N_{\mathrm{A}}\frac{D_{\mathrm{W}}}{i^{1/2}} \quad i \geq z, \qquad k_{\mathrm{e,micelle}}^i(V) = 0 \quad i < z \tag{13}$$

In these equations, D_{W} is the diffusion coefficient for a monomeric radical in water, e_{e} is the entry efficiency, r_{micelle} is the radius of a micelle, and r_{s} is the radius of a latex particle swollen with monomer. Note that r_{s} is related to the unswollen radius, r, of a particle as follows:

$$\frac{r_{\mathrm{s}}}{r} = \left[\frac{d_{\mathrm{m}}}{d_{\mathrm{m}} - C_{\mathrm{P}}M_0}\right]^{1/3} \tag{14}$$

where d_{m} is the monomer density, M_0 is the monomer molecular weight, and C_{P} is the concentration of monomer in the latex particle. The entry coefficient for exited radicals can be defined in a similar fashion, as follows:

$$k_{\mathrm{eE}}(V) = 4\pi r_{\mathrm{s}}N_{\mathrm{A}}e_{\mathrm{e}}D_{\mathrm{W}} \tag{15}$$

The rate coefficient for desorption of monomeric radicals may be written as a function of the diffusivity of monomer both in the water phase and inside the particle, the aqueous monomer concentration, the monomer concentration in the particle, and the swollen radius:

$$k_{\mathrm{dM}}(V) = \frac{3D_{\mathrm{W}}D_{\mathrm{mon}}}{\left(\frac{C_{\mathrm{P}}}{C_{\mathrm{w}}}D_{\mathrm{mon}} + D_{\mathrm{W}}\right)r_{\mathrm{s}}^2} \tag{16}$$

The overall rate for entry, ρ, is derived from the following equation:

$$\rho = k_{\mathrm{eE}}[\mathrm{E}^\bullet] + \rho_{\mathrm{initiator}} \tag{17}$$

where $\rho_{\text{initiator}}$ is the rate of formation of ''z-mers'' obtained from

$$\rho_{\text{initiator}} = \sum_{i=z}^{j_{\text{crit}}-1} k_{\text{e}}^{i}[\text{IM}_i] \tag{18}$$

The propagational growth rate, $K(V)$, is defined by

$$K(V) = \frac{k_{\text{p}} M_0 C_{\text{P}}(V)}{N_{\text{A}} d_{\text{p}}} \tag{19}$$

where d_{p} is the density of the polymer.

The average number of radicals per particle, $\bar{n}$, is given by

$$\overline{n} = \frac{1}{\sum_{i=1}^{G} n(V_i)} \left(\sum_{i=1}^{G} n_1^{\text{P}}(V_i) + \sum_{i=1}^{G} n_1^{\text{M}}(V_i) \right) \tag{20}$$

while the total number of particles per unit volume, N_{tot}, can be calculated from

$$N_{\text{tot}} = N_{\text{A}} \int_0^{\infty} n(V)\mathrm{d}V \tag{21}$$

At high monomer conversion, the viscosity inside the polymer particles increases sharply and further polymerization becomes diffusion controlled. The particles are referred to as a glassy polymer and the kinetics for a zero–one system is no longer valid. To account for these changes, the propagation rate coefficient can be expressed as follows

$$k_{\text{p}} = 1/\left(1/k_{\text{p}_0} + 1/k_{\text{diff}}\right) \tag{22}$$

where k_{p_0} is the propagation rate coefficient at low conversion, while k_{diff} is the diffusion-controlled rate coefficient defined as

$$k_{\text{diff}} = 4\pi\sigma N_{\text{A}}(D_{\text{mon}} + D_{\text{rd}}) \tag{23}$$

where $D_{\text{rd}} = (1/6)\left(k_{\text{p}} C_{\text{p}} \alpha^2\right)$. The diffusion coefficient for the monomer (styrene in this case), D_{mon}, was expressed as a function of the polymer volume fraction inside the particle

$$\begin{aligned} D_{\text{mon}} &= 10\,\hat{}\,(0.417 - 29.51\phi_{\text{p}} + 53.14\phi_{\text{p}}^2 \\ &\quad - 36.03\phi_{\text{p}}^3) \quad \text{for } \phi_{\text{p}} < 0.8 \\ D_{\text{mon}} &= 9 \times 10^{-8} \exp(-19.16\phi_{\text{p}}) \quad \text{for } \phi_{\text{p}} \geq 0.8 \end{aligned} \tag{24}$$

An additional entry phenomenon can be brought forward at high ϕ_{p}, that is, entry of a z-mer into a glassy particle, is impeded because of slow diffusion through the interface. The hindered z-mer then desorbs back into the aqueous phase and may propagate to a j_{crit}-mer and cause secondary nucleation. Therefore, an empirical equation for the entry efficiency, $e_{\text{e}} = (1 - \phi_{\text{p}})^{C_{\text{p-sat}}}$, was used to model these changes. In a similar fashion, the termination rate coefficient is calculated from the empirical equation:

$$k_{\text{t}} = k_{\text{to}} \exp\left(-19\phi_{\text{p}}^{2.1}\right) \tag{25}$$

The instantaneous MWD can be calculated (over the conversion range) using:[24–27]

$$\begin{aligned} \frac{\partial \overline{P}(M)}{\partial t} &= P(M) = (k_{\text{tr}} C_{\text{P}} + \rho)\bar{n} \\ &\quad \times \exp\left(\frac{-(\rho + k_{\text{tr}} C_{\text{P}})}{k_{\text{P}} C_{\text{P}}} \frac{M}{M_0}\right) \end{aligned} \tag{26}$$

where M is the molecular weight of the polymeric chain.

For a semibatch process (as considered here), the rate of accumulation of monomer in the reaction vessel can be obtained from a molar balance equation

$$\frac{\mathrm{d}N_{\text{m}}}{\mathrm{d}t} = F_{\text{m}} - R_{\text{P}} V_{\text{r}} \tag{27}$$

where $R_{\text{P}} = k_{\text{p}} C_{\text{p}} (\bar{n} N_{\text{tot}}/N_{\text{A}})(V_{\text{w}}/V_{\text{r}})$ is the reaction rate and V_{r} is the total reaction volume, the latter being described by

$$\frac{\mathrm{d}V_{\text{r}}}{\mathrm{d}t} = M_0 R_{\text{p}} V_{\text{r}} (1/d_{\text{p}} - 1/d_{\text{m}}) + F_{\text{m}} M_0/d_{\text{m}} \tag{28}$$

Here, F_{m} is the monomer feed rate. At high feed rates, monomer droplets accumulate inside the reactor and the monomer concentration inside the particles reaches its saturation value. Therefore, C_{p} may be estimated as follows:

$$V_{\text{d}} = (N_{\text{m}} M_0 - C_{\text{w}} M_0 V_{\text{w}} - C_{\text{p}} M_0 V_{\text{p}})/d_{\text{m}}$$

where

$$\text{C}_{\text{p}} = \text{C}_{\text{p\_sat}} \quad \text{when } V_{\text{d}} > 0$$

and

$$\text{Cp} = N_{\text{m}}/V_{\text{p}} \quad \text{when } V_{\text{d}} = 0 \tag{29}$$

In Eq. (29), V_{d} is the droplet volume, the monomer concentration in the water phase is given by $C_{\text{w}} = C_{\text{p}} k_{\text{mwp}}$, while k_{mwp} is the partition coefficient of the monomer between the water and the polymer phases. In the same equation, V_{p} is the volume of the

polymer phase as calculated by

$$\frac{dV_p}{dt} = M_0 R_p V_r / d_p \quad (30)$$

Energy balance:

$$MC_p \frac{dT_e}{dt} = \sum_i F_i C_{p_i} (T_{amb} - T_e) + Q_f - Q_{loss} + Q_r + Q_s$$

where $\sum_i F_i C_{p_i}(T_{amb} - T_e)$ is heat flow due to feed input, Q_f is heat flux across the reactor wall, Q_{loss} is heat loss to the surroundings, Q_r is heat generation rate due to the chemical reaction, and Q_s is the heating due to stirring.

REACTOR OPERATION AND CONTROL

The reactor types commonly used for emulsion polymerization are:

- *Batch stirred tank reactor*: Low-volume products; flexibility to produce numerous grades (batch-to-batch variability); and homogenous (narrow MWD) or segregated (broad MWD).
- *Semibatch stirred tank reactor*: Similar to batch reactor but provides greater flexibility in operations and in obtaining the desired product characteristics.
- *Continuous reactor*: Single high-volume product; low operating costs; consistent quality (reduced flexibility); and broad MWD.

The current confidence in the design and scale-up of emulsion polymerization reactors is due to:

- The successful testing of a wide range of data with model predictions. For example, the models can successfully predict monomer conversions (see Fig. 4), molecular weight average and distribution (see Fig. 5), and particle size average and distribution (see Fig. 6) for batch and semibatch reactors. The results have also been verified for continuous reactors.
- The understanding of the dependence of particle diameter on initiator concentration and ionic strength below the CMC, and how the particle diameter varies with surfactant concentration above the CMC.
- The model has been used to analyze and improve performance for a range of emulsion processes.
- The substantial level of details (e.g., kinetics, PSD, MWD) that can be incorporated into reactor operation and control algorithms.[3]

OPTIMIZATION AND CONTROL OF EMULSION POLYMERIZATION

The challenges to be overcome in emulsion polymerization reactor optimization and control include the following considerations:[28–32]

- Multiphase media.
- Inadequate on-line measuring instruments.
- Effect of control on conversion and particle size are strongly interactive.
- Adaptive and decoupling control is needed.
- May exhibit limit cycle behavior.
- Control system must handle large, variable dead times, as well as severe nonlinearities.

Some of these challenges are overcome through the use of sophisticated models, complex algorithm application, and inferential measurements or use of soft-sensors.

MEASUREMENT OF VARIABLES

The relevant variables to be monitored in emulsion polymerization are temperature, flow rate, conversion, molecular weight, particle size, density, viscosity,

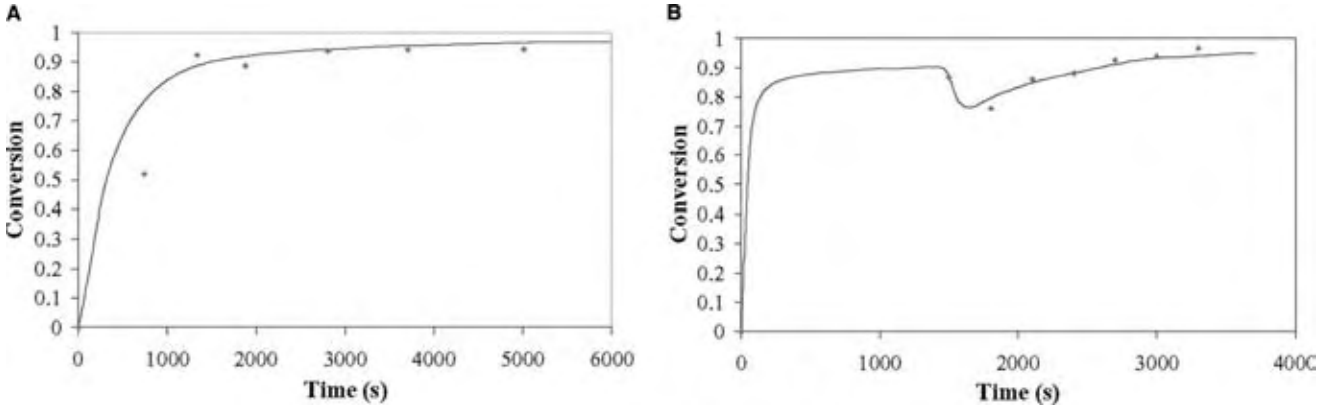


Fig. 4 Conversion profiles as a function of time for (A) batch and (B) emibatch operation. *(View this art in color at www.dekker.com.)*

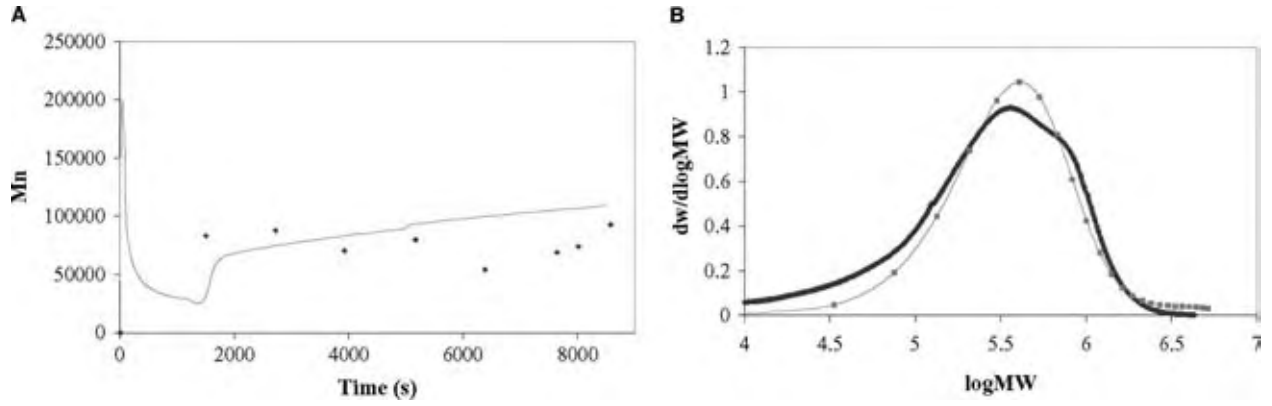


Fig. 5 Molar mass profiles for emulsion polymerizations: (A) number average and (B) distribution. *(View this art in color at www.dekker.com.)*

composition, etc. The measurement of an important process variable such as on-line conversion is a challenging task. Some of the measurement techniques tested, such as gravimetry (mass measurement), densitometry, chromatography, Raman spectroscopy, and NMR are off-line, suffer from inaccuracy, or are relatively expensive. Industrial and laboratory applications often rely on microcalorimetry, which does not share the above-mentioned disadvantages.

The measurement of particle size and molar mass distributions are typically carried out off-line with relatively expensive instruments. Techniques used for particle size include transmission electron microscopy, photon correlation spectroscopy, or capillary hydrodynamic fractionation and molar mass measurement with GPC.

MODEL PREDICTIVE CONTROL

The key polymer properties are heavily related to the polymer PSD and MWD, which are in turn highly sensitive to the process and kinetic history. Any disturbances in the operating conditions during normal operation may cause irreversible changes in the quality of the polymer formed. Therefore, adjustments may have to be made during reaction through an effective on-line control scheme. For this purpose the use of the dynamic matrix control (DMC) algorithm is studied in this entry for the on-line model based control of the PSD and MWD with different control formulations. An intelligent control hierarchy is formulated for such a distributed parameter system incorporating three different levels: off-line optimization, on-line DMC, and regulatory control successively.

At the heart of an model predictive control (MPC) application is the optimization of a variable subject to constraints. A typical MPC cost functional is given as follows:

$$J = \min_{\Delta u} \left[\sum_{i=1}^{P} \delta(y^{\text{setpoint}}(k+i) - y(k+i))^2 + \sum_{i=1}^{M} \lambda(\Delta u(k+M-i))^2 \right]$$

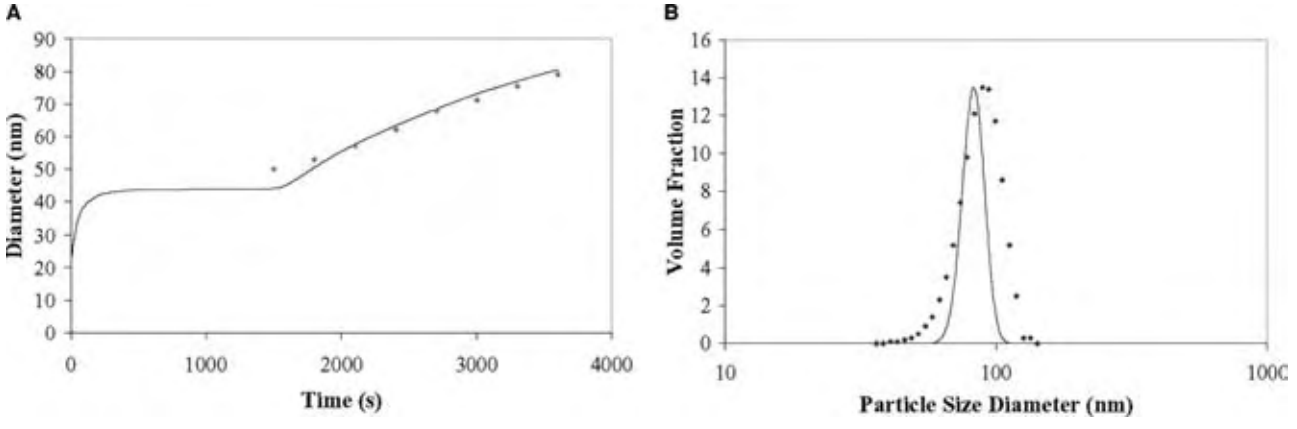


Fig. 6 Particle diameter for emulsion polymerization: (A) number average and (B) distribution. *(View this art in color at www.dekker.com.)*

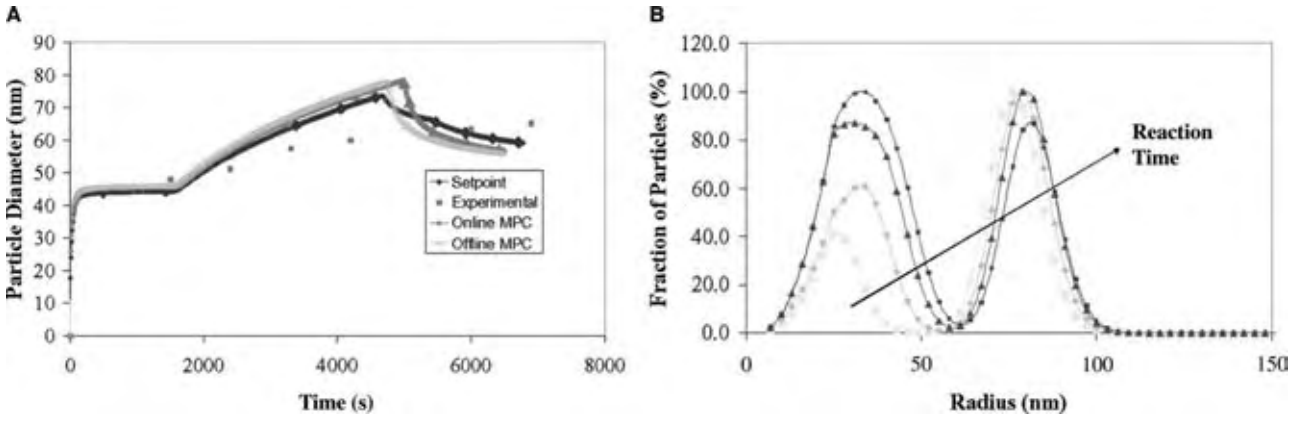


Fig. 7 Particle size control using on-line multivariable constrained MPC: (A) average particle diameter and (B) PSD. *(View this art in color at www.dekker.com.)*

subject to:

Constraints on input moves: $\Delta u(k) \in [\Delta u_{\min}, \Delta u_{\max}]$.

Constraints on input: $u(k) \in [u_{\min}, u_{\max}]$.

Constraints on output: $y(k) \in [y_{\min}, y_{\max}]$.

The weight on the error, δ, is a $P \times P$ diagonal matrix. Similarly, λ is the weight on the input (suppression move) and is an $M \times M$ diagonal matrix. The optimization problem is solved at every sampling time when a new prediction is updated by recent feedback measurements. Here, y^{setpoint} is a matrix of the set point values for the average radius, molar mass, and conversion.

The model predictions have been found to compare well with the experimentally controlled variables as shown in Figs. 7 and 8. Thus, the currently developed models show a high degree of accuracy and are suitable for process operation and control.

CONCLUSIONS

Emulsion polymerization is a powerful technique for developing products for a wide range of industry. The process can be used for the development of structured products in the nano- to microscales. Some of the key advantages of the process are the use of mild conditions, the near-complete conversion of monomers, the minimization of separation and recycling, the improved heat and mass transfer, and the improved environmental benefits due to the use of a water medium. However, the process is complex and requires careful modeling and scale-up. The prediction of process behavior, the key product properties, and the control of emulsion polymerization systems in particu-

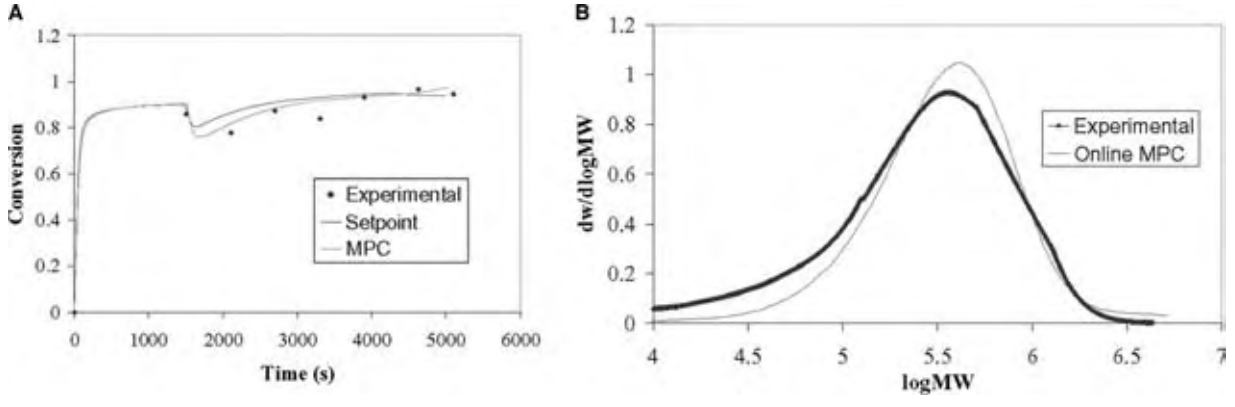


Fig. 8 Molar mass control using on-line multivariable constrained MPC: (A) conversion and (B) MWD. *(View this art in color at www.dekker.com.)*

lar require a thorough understanding of the mechanisms and the chemical subprocesses. In this respect, a good understanding of the reaction kinetics, the process stages or intervals, and the physicochemical transformations is essential.

Apart from the development of models suitable for predictive and control applications, some of the major challenges include accurate measurements of process variables that are particularly difficult to accomplish on-line. In this respect, the PSD and the MWD are particularly important, as they are factors that influence some of the most important properties of polymers, such as the surface coverage of coatings and the mechanical strength of the polymer product. However, the use of soft-sensors, the application of model predictions, and process control implementation for a varied number of systems have shown positive results. Several studies in the literature have demonstrated the efficacy of the model in adequately predicting the full PSD and the MWD of the polymer formed, over a wide range of reactor operating conditions. Further, the PSD and the MWD are controllable by varying the operating regime and the process conditions, such as temperature and flow rates of the monomers, initiators, and surfactants. Thus, the advanced operation and control of the process and some of the key products are now feasible. Consequently, on-line optimal control is gradually being implemented for the process.

NOMENCLATURE

a_s	Surface area of surfactant molecule (dm^{-2})
$B(V,V')$	Rate coefficient for coagulation of two particles of volume V and V and V′ (L/mol/sec)
CMC	Critical micelle concentration (mol/L)
$C_{micelle}$	Concentration of micelles (mol/L)
C_P	Monomer concentration in latex particles (mol/L)
C_{p_sat}	Saturated monomer concentration in latex particles (mol/L)
C_W	Monomer concentration in the water phase (mol/L)
C_{W_sat}	Saturated monomer concentration in the aqueous phase (mol/L)
d_m	Density of monomer (kg/L)
d_p	Density of polymer (kg/L)
D_W	Diffusion coefficient for monomer in water (dm^2/sec)
[E]	Aqueous phase concentration of desorbed radicals (mol/L)
F_m	Monomer feed to the reactor (mol/sec)
$[IM_i]$	Aqueous phase concentration of oligomeric radicals of degree i (mol/L)
j_{crit}	Critical degree of polymerization for homogenous nucleation
K	Rate of propagation (volume growth per particle) (L/sec)
k_d	Rate coefficient for initiator decomposition (s^{-1})
k_{dM}	Rate coefficient for desorption of monomeric radicals from particles (s^{-1})
k_e^i	Rate coefficient, entry of oligomeric radical of degree i into existing particle (L/mol/sec)
k_{eE}	Rate coefficient for reentry of desorbed radicals (L/mol/sec)
$k_{e,micelle}^i$	Rate coefficient for entry of an oligomeric radical of degree i into a micelle (L/mol/sec)
k_p	Long chain limiting propagation rate coefficient in the latex phase (L/mol/sec)
k_p^1	Monomeric radical propagation rate coefficient in the latex phase (L/mol/sec)
$k_{p,aq}^i$	aqueous phase propagation rate coefficient for oligomeric radicals of degree i (L/mol/sec)
k_{mwp}	Partition coefficient for styrene between the water phase and polymer phase
$k_{t,aq}$	Rate coefficient for termination reaction between oligomeric radicals in the aqueous phase (L/mol/sec)
k_{tr}	Rate coefficient for radical transfer to monomer (L/mol/sec)
M_0	Monomer molecular weight (kg/mol)
$\bar{n}$	Average number of radicals per particle
n_{agg}	Surfactant agglomeration number
N_m	Number of moles of monomer
N_{tot}	Total number concentration of particles (1/L)
r	Radius of latex particle (dm)
R_P	Polymerization reaction rate (mol/L/sec)
r_s	Swollen radius of latex particle (dm)
V_d	Volume of monomer in droplets (L)
V_M	Volume of monomer in system (L)
V_P	Volume of polymer in system (L)
V_r	Total reaction volume (L)
V_W	Volume of water in the system (L)
z	Critical degree of polymerization for entry
σ	van der Waals radius of a monomer unit (dm)

ACKNOWLEDGMENTS

The author gratefully acknowledges support from Zeaiter, Alhamad, the Australian Postgraduate Award

scheme and the Key Centre for Polymer Colloids, which is established and supported under the Australian Research Council's Research Centres Program.

REFERENCES

1. Gilbert, R.G. *Emulsion Polymerization*: *A Mechanistic Approach*; Academic Press: London, 1995.
2. Smith, W.V.; Ewart, R.H. Kinetics of emulsion polymerization. J. Chem. Phys. **1948**, *16*, 592.
3. Zeaiter, J.; Romagnoli, J.A.; Gomes, V.G.; Gilbert, R.G. Operation of semi-batch emulsion polymerisation reactors: modelling, validation, and effect of operating conditions. Chem. Eng. Sci. **2002**, *57*, 2955–2969.
4. Coen, E.M.; Gilbert, R.G.; Morrison, B.R.; Leube, H.; Peach, S. Modelling particle size distributions and secondary particle formation in emulsion polymerization. Polymer **1998**, *39*, 7099–7112.
5. Dougherty, E.P. The scope of dynamic-model for emulsion polymerization. 1. Theory. J. Appl. Polym. Sci. **1986**, *32* (1), 3051–3078.
6. Gugliotta, L.M.; Brandolini, M.C.; Vega, J.R.; Iturralde, E.O.; Azum, J.L.; Miera, G.R. Dynamic-model of a continuous emulsion copolymerization of styrene and butadiene. Polym. React. Eng. **1995**, *3* (3), 201–233.
7. Liotta, V.; Sudol, E.D.; El-Aasser, M.S.; Georgakis, C. On-line monitoring, modeling, and model validation of semibatch emulsion polymerization in an automated reactor control facility. J. Polym. Sci. Pt. A: Polym. Chem. **1998**, *36* (10), 1553–1571.
8. Penlidis, A.; Macgregor, J.F.; Hamielec, A.E. Mathematical-modeling of emulsion polymerization reactors—a population balance approach and its applications. ACS Symp. Ser. **1986**, *313*, 219–240.
9. Richards, J.R.; Congalidis, J.P.; Gilbert, R.G. Mathematical modeling of emulsion copolymerization reactors. J. Appl. Polym. Sci. **1989**, *37*, 2727–2756.
10. Li, B.; Brooks, B.W. Semibatch processes for emulsion polymerization. Polym. Int. **1992**, *29* (1), 41–46.
11. van Herk, A.M. Pulsed initiation polymerization as a means of obtaining propagation rate coefficients in free-radical polymerization. J. Macromol. Sci. Rev. Macromol. Chem. Phys. **1997**, *C37*, 633
12. Beuermann, S.; Buback, M.; Davis, T.P.; Gilbert, R.G.; Hutchinson, R.A.; Kajiwara, A.; Klumperman, B.; Russell, G.T. Critically evaluated rate coefficients for free-radical polymerization. Macromol. Chem. Phys. **2000**, *201*, 1355
13. Hansen, F.K.; Ugelstad, J. Eds. *Particle Formation Mechanisms. Emulsion Polymerization*; Academic Press: New York, 1982.
14. Fitch, R.M. *Polymer Colloids: A Comprehensive Introduction*; Academic Press: New York, 1995.
15. Dotson, N.A.; Galvan, R.; Laurence, R.L.; Tirrell, M. *Polymerization Process Modeling*; VCH Publishers: New York, 1996.
16. Saldivar, E.; Dafniotis, P.; Ray, W.H. Mathematical modeling of emulsion copolymerization reactors. I. Model formulation and application to reactors operating with micellar nucleation. J. Macromol. Sci. Rev. Macromol. Chem. Phys. **1998**, *C38* (2), 207–325.
17. Saldivar, E.; Araujo, O.; Giudici, R.; Guerrero-Sánchez, C. Modeling and experimental studies of emulsion copolymerization systems. III: acrylics. J. Appl. Polym. Sci. **2002**, *84*, 1320–1338.
18. Gerrens, H. On semicontinuous emulsion polymerization. J. Polym. Sci. Pt. C **1969**, *27* (N7), 77.
19. Sajjadi, S. Particle formation under monomer-starved conditions in the semibatch emulsion polymerization of styrene. I. Experimental. J. Polym. Sci. Pt. A: Polym. Chem. **2001**, *39*, 3940–3952.
20. Harada, M.; Nomura, M.; Eguchi, W.; Nagata, S. Studies of the effect of polymer particles on emulsion polymerization. J. Chem. Engng. Jpn. **1971**, *4*, 54–60.
21. Lichti, G.; Gilbert, R.G.; Napper, D.H Molecular weight distributions in emulsion polymerizations. J. Polym. Sci. Pt. A: Polym. Chem. **1980**, *18*, 1297.
22. Clay, P.A.; Gilbert, R.G. Molecular weight distributions in free-radical polymerizations. I. Model development and implications for data interpretation. Macromolecules **1995**, *28* (2), 552–569.
23. Lovell, P.A.; El-Aasser, M.S. *Emulsion Polymerization and Emulsion Polymers*; Wiley: London, 1997.
24. Forcada, J.; Asua, J.M. Modelling the microstructure of emulsion copolymers. J. Polym. Sci. Polym. Chem. Ed. **1985**, *23*, 1955–1962.
25. Van Doremaele, G.H.J.; van Herk, A.M. Modelling of emulsion copolymer microstructure. Polym. Int. **1992**, *27* (2), 95–108.
26. Echevarria, A.; Leiza, J.R. Molecular-weight distribution control in emulsion polymerization. AIChE J. **1998**, *44* (7), 1667–1679.
27. Tobita, H.; Takada, Y.; Nomura, M. Simulation model for the molecular weight distribution in

emulsion polymerization. J. Polym. Sci. Pt. A: Polym. Chem. **1995**, *33*, 441–453.
28. Sayer, C.; Arzamendi, G.; Asua, J.M.; Lima, E.L.; Pinto, J.C. Dynamic optimization of semicontinuous emulsion copolymerization reactions: composition and molecular weight distribution. Compos. Chem. Eng. **2001**, *25* (4–6), 839–849.
29. Show-An, C.; Kuo-Wei, W. Emulsion polymerization: theory and particle size distribution in copolymerization system. J. Polym. Sci. Pt. A: Polym. Chem. **1988**, *26*, 1487–1506.
30. Storti, G.; Polotti, G.; Cociani, M.; Morbidelli, M. Molecular weight distribution in emulsion polymerization. I. The homopolymer case. J. Polym. Sci. Polym. Chem. **1992**, *30*, 731–750.
31. Vicente, M.; Sayer, C.; Leiza, J.R.; Arzamendi, G.; Lima, E.L.; Pinto, J.C.; Asua, J.M. Dynamic optimization of non-linear emulsion copolymerization systems open-loop control of composition and molecular weight distribution. Chem. Eng. J. **2002**, *85* (2–3), 339–349.
32. Crowley, T.J.; Meadows, E.S.; Kostoulas, E.; Doyle, F.J. Control of particle size distribution described by a population balance model of semibatch emulsion polymerization. J. Process. Control **2000**, *10* (5), 419–432.

Enhanced Oil Recovery

Kishore K. Mohanty
Department of Chemical Engineering, University of Houston, Houston, Texas, U.S.A.

Gerard T. Caneba
Department of Chemical Engineering, Michigan Technological University, Houghton, Michigan, U.S.A.

INTRODUCTION

More than half the original oil typically remains in oil reservoirs after primary and secondary recovery operations. Primary recovery refers to production of oil because of its natural energy; fluids expand as pressure falls to push out some oil and gas. Expansion of associated aquifers and gas caps also help in pushing out oil. Primary recovery efficiency varies greatly from reservoir to reservoir and is typically in the range of 5%–20%. Secondary recovery refers to injection of immiscible fluids, such as water and gas, to recover oil. These fluids displace oil from the pore space immiscibly. Secondary recovery efficiency is typically another 10–20%. Oil is left behind in bypassed regions as well as in swept zones. Oil is bypassed in certain zones of the reservoir because of permeability heterogeneity, lack of conformance at the wells, pattern orientation, and sometimes-viscous fingering. Oil is also left behind in the swept zones because of capillary forces in immiscible displacements during secondary recovery. Tertiary recovery techniques (also called enhanced oil recovery (EOR) techniques) are needed to recover additional oil from existing fields.

MISCIBLE GAS INJECTION PROCESSES

Process Description

Miscible flooding is one of the commercially successful improved recovery methods. It constitutes the injection of CO_2, hydrocarbon gases, and even nitrogen or flue gas.[1] Typically 10–50% PV of the injectant is injected in the case of CO_2 or hydrocarbon gases. A much larger amount of nitrogen or flue gas can be injected because they are cheaper. These gases can be injected in different modes: miscible gas injection followed by dry gas injection, miscible gas injection followed by water injection or water–alternating-gas (WAG) injection. The individual slug size in the WAG injection is in the range of 1–5% and typical WAG ratio is 0.5–5. Incremental production cost for CO_2 flooding is estimated to be between \$2 and \$8 per barrel of oil.[a]

Recovery Mechanisms

Gas flooding recovers oil by reducing residual oil saturation and oil viscosity, and oil swelling.[2] When two fluids are miscible, interfacial tension disappears and capillary forces do not exist. Thus, residual saturation of a displaced phase is zero in completely miscible floods. Oils can be solubilized by solvents and recovered. When the solvent is a gas, e.g., CO_2 or enriched gas, it develops miscibility at high pressures, often greater than 1200 psi at reservoir temperatures. The pressure needed for miscibility must be lower than the fracture pressure of the formation. The reservoir should be deep enough to accommodate this pressure. Availability of suitable miscible solvents is often a constraint. Large deposits of CO_2, present in the West Texas and mid-continental areas, have promoted many CO_2 floods in that region. Miscible hydrocarbon solvents are available in some reservoirs to a limited extent from its associated gas or gas in neighboring fields, and hydrocarbon gasflood processes are used in those reservoirs. Air and nitrogen are available widely but need very high pressure or depth to develop miscibility.

Gas injection can also recover oil by reducing oil viscosity and residual oil saturation,[2] even when miscibility is not achieved. Reduction in viscosity is more significant if the oil viscosity is large, and this process is attractive in viscous or semiviscous reservoirs, especially when accompanied by some other improved recovery mechanism. Residual oil saturation in three-phase flow in water-wet rock is very low (essentially zero), even at very low capillary numbers.[3] Two main problems in such a process are the low relative permeabilities and sweep efficiencies. This process can be implemented in a highly dipping reservoir to take

[a]Taken from the website wysiwyg://256http://www.fe.doe.gov/oil_gas/res_efficiency/res_progareas.shtml

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007753

advantage of gravity. Three-phase immiscible residual gasfloods are not that low for other wettability and spreading conditions.[4] The key advantage of immiscible processes is the cost of the solvent, as solvents are cheap.

Phase Behavior and Miscibility

Mixtures of oil and gases form one, two, or even three hydrocarbon fluid phases.[1] This does not include a solid phase or asphaltene, which can also be present. When the mixture is one phase irrespective of the ratio of the oil and gas, the solvent is called first contact miscible (FCM) with the oil. This happens with most gases at very high pressures. If the pressure or enrichment is lower, solvents may not be first-contact miscible, but can still displace the oil in a one-dimensional tube by forming a transition zone of fluid compositions that range from the oil to injection fluid composition, all within a single phase. One of the tielines shrinks to zero length in this composition path. Such a solvent is called multiple-contact miscible (MCM). Multi-contact miscibility can be developed by vaporizing, condensing, or condensing–vaporizing mechanisms.[5] If oil is displaced by a solvent at a lower pressure or enrichment, multiple phases exist at all points of the transition zone, and these are termed immiscible.

The multicontact miscibility development by a vaporizing mechanism is illustrated in Fig. 1. The pseudoternary diagram shows light (L) components (CO_2, CH_4), intermediate (I) gases (C_2–C_5) and heavy (H) components (C_{6+}) in the vertices. The critical point on the two-phase diagram is denoted by C. Oil and solvent compositions are denoted by O and S, respectively. As oil and solvent mix at oil–solvent front, two phases form, as shown by L1 and G1. The gas moves ahead being less viscous and mixes with original oil. This produces a two-phase mixture that forms compositions shown by L2 and G2. G2 moves ahead and mixes with oil. This process repeats itself and the gas composition moves along the phase diagram along the red line, which eventually is first contact miscible with the oil near the critical point. This is possible if the pressure is high enough. The MCM development by a condensing mechanism is illustrated in Fig. 2. Multi-contact miscibility can be developed by a vaporizing or a condensing mechanism if the oil composition and the solvent composition lie on opposite sides of the critical tie-line shown by CA'. This is possible by high enough pressure or enough enrichment of the solvent.

The enrichment (or the pressure) needed to develop miscibility between the injectant and the oil is determined experimentally in one-dimensional slim-tube tests.[1] As the enrichment (or pressure) increases, the slim-tube recovery reaches a plateau long before first contact miscibility is developed. This enrichment is called the minimum miscibility enrichment, or MME, which is a function of reservoir pressure, temperature, and contaminants in the solvent. Similarly, a pressure called the minimum miscibility pressure, or MMP, can be identified for any solvent. Other experimental methods (e.g., rising bubble method) are also available to determine MMP or MME. In vaporizing three component systems, MMP (or MME) corresponds to the pressure (or enrichment) at which the critical tie line passes through the crude oil composition. In condensing three component systems, MMP (or MME) corresponds to the pressure (or enrichment) at which the critical tie line passes through the solvent composition.[5]

Several correlations have been developed in the past to estimate MME and MMP from the compositions of oil and solvent and the reservoir temperature.[1,6] Generally, the MMP increases with the temperature and the molecular weight of the oil heavy fraction. Carbon dioxide must achieve a certain density to develop with a crude oil. This required density is a function of the heavy fraction content. Nitrogen and CH_4 increase CO_2 MMP; H_2S decreases it. The solvency is increased if CO_2 is diluted with an impurity, whose critical

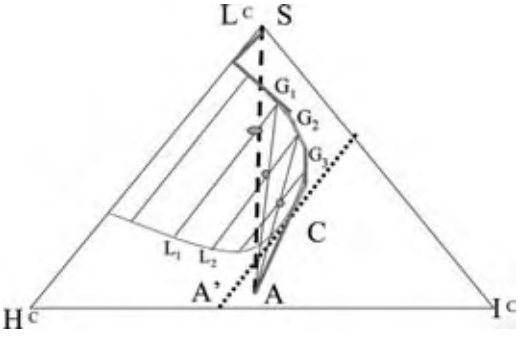


Fig. 1 Compositional path in vaporizing multi-contact miscible flood. *(View this art in color at www.dekker.com.)*

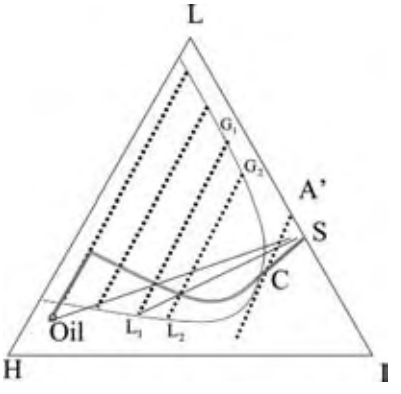


Fig. 2 Compositional path in condensing multi-contact miscible flood. *(View this art in color at www.dekker.com.)*

temperature is higher than that of CO_2. Sebastian[7] has developed a correction to CO_2 MMP in terms of the critical temperatures of the impurities. Benham et al.[8] give correlations of MME or MMP for hydrocarbon gases. The MME depends on the temperature, pressure, and molecular weight of the intermediate components in the solvent and molecular weight of the C_{5+} fraction in the crude oil.

Many relatively shallow reservoirs, especially those in Alaska and some in West Texas, are at temperatures below 50°C. In such reservoirs, three hydrocarbon fluid phases (L1-L2-V) form when oil and solvents mix.[9] Slim tube recoveries greater than 90% can be observed in such fluids with hydrocarbon solvents, even when three fluid phases form. Although true miscibility did not develop, high oil recovery rates were observed through the condensing–vaporizing mechanism. These solvents also had high coreflood recoveries.

Bypassing in Multidimensional Floods

The mobility ratio and density contrast is large in most solvent floods. Sweep efficiency can be low because of rock heterogeneity, viscous fingering, and gravity override, and it plays a crucial role in determining the overall recovery efficiency.[1] To understand this process, one needs to model the amount of bypassing and the resulting mass transfer between bypassed and fingered fluids. In one-dimensional displacements, MME and MMP are the optimum levels of enrichment or pressure for the solvent. However, reservoir flow is three-dimensional and the optimum enrichment (or pressure) can be different from that of the one-dimensional flood in a slim-tube.[10] Injected solvent composition (or pressure) affects not only the local displacement efficiency, i.e., that evaluated by one-dimensional corefloods, but also the sweep efficiency.

Bypassing must be reduced in all gasfloods. There are several techniques to reduce bypassing—WAG, foams, and polymers.[11] Field tertiary gasfloods are often conducted with alternating water, called WAG scheme. Injection of water and solvent slows both the solvent and the water, as they block each other at the pore-scale. Efficiencies of WAG schemes depend on the WAG ratio, slug size, and rock wettability. Optimum WAG ratios are often determined from fractional flows in homogeneous cores. The traditional fractional flow analysis for WAG ratio is not valid in heterogeneous rocks. Stern[12] found that CO_2 WAG displacements were quite effective at high WAG ratios in mixed-wet rocks.

Foams and polymers have been studied in the past without much field-scale success. Surfactant water is injected into the reservoir alternating with the miscible injectant to produce foams. The foams used in the near-wellbore region are designed to be strong and do work. It has been very difficult to design weak foams that propagate through the reservoir and move the solvent banks in a stable manner. Surfactant loss (because of adsorption and otherwise) and cost are some of the major problems. Polymers or other chemicals that are directly soluble in the solvent have been used to increase the viscosity and the density of the solvent. However, cost is the key limiting factor.

Screening Criteria

CO_2 flooding is applicable to reservoirs with relatively light oil (gravity >25 API, viscosity <15 cp, and high amounts of C_5–C_{20} in oil).[13] The formation can be either sandstones or limestones. The reservoir pressure should be high (>1200 psi for high gravity oil to >4500 psi for low gravity oil). This implies that the reservoir should be deep enough. The reservoir should be relatively thin or dipping. A high degree of heterogeneity lowers the efficiency of the process. The temperature is not that critical, but the minimum pressure required for miscibility increases with pressure. Hydrocarbon gas injection works in similar reservoirs except with slightly lighter oils (gravity >35 API). Nitrogen and flue gas require much higher pressure to develop miscibility. Thus, they are applied to deeper reservoirs (>4500 ft) with light oils (>35 API for nitrogen and 25 API for flue gas).

Field Experience

CO_2 and hydrocarbon gas flooding remain popular methods to recover oil. Between 1982 and 1992, oil production from gas floods increased from 72,028 to 296,020 b/d in the United States.[14] The latest data can be found in the review of EOR projects published by *Oil & Gas Journal* every two years. In 1992, there were 45 active CO_2 miscible projects in the USA.[15] About 2.4 BCF/D of CO_2 was being injected into these reservoirs. About 40% of this gas was being produced and reinjected. The incremental production response was about 142,000 b/d of oil. The industry had booked about 1.9 billion barrels of oil because of CO_2 flood. Hydrocarbon flooding has been used in Prudhoe Bay field in Alaska, Block 31 field in West Texas, and Hassi-Messaoud field in Algeria, to give a few examples.

One of the key factors for feasibility of a CO_2 flooding project is the availability of large amounts of CO_2. The 45 projects discussed here are located near large natural deposits of CO_2 gas. The main area of activity has been the Permian Basin area of the West Texas and New Mexico. The main sources of CO_2 supply for these projects have been McElmo Dome, Sheep Mountain, and Bravo Dome, which contain between 20 and 40 TCF of

CO_2. In northwest region, LaBarge field, which supplies to several CO_2 floods, contains about 20–60 TCF of CO_2.

Oil recovery has been significantly enhanced by the injection of CO_2 and hydrocarbon gases. The typical incremental oil is about 8–14% OOIP for tertiary injection. For example, the SACROC field quadrupled the oil production rate and produced more than 10% of the OOIP. In secondary injections, for example, in Block 31 field in Texas, the oil recovery is close to 60% of the OOIP. Improved reservoir management in the preparation of CO_2 flood also added to improved recovery. For example, in the Means field, well spacing was changed from 20 acre/pattern to 10 acre/pattern and five-spot to inverted nine-spot. These changes increased oil production rate from 7,000 to 10,000 b/d, even before CO_2 injection. 35% of the EOR oil was produced because of improved reservoir management.

CO_2 injection leads to early breakthrough, but the oil production continues beyond breakthrough. CO_2 utilization was about 7–10 MCF/BBL. Two-thirds of the injected CO_2 was recovered. CO_2 production can be curtailed by the use of WAG and polymers and foams. CO_2 utility rate in the field are often in line with the simulation indicating that CO_2 displaces the trapped residual oil. CO_2 injectivity in WAG floods was an industry concern. The field tests have shown that loss of injectivity is minor and manageable. Post CO_2 flood-water injectivity was higher than the original water injectivity in several reservoirs. Water blocking is not found to be an issue in WAG floods. Corrosion problems are manageable if separate CO_2 injection line, SS wellheads and fiberglass gathering system are used. Asphaltene problems are negligible, except during early CO_2 breakthrough.

THERMAL METHODS

Process Description

The two main thermal recovery processes are steam injection and in situ combustion. In the steam injection process, steam of 80% quality is injected into the reservoir to displace oil. The steam can be injected continuously (called ''steam drive'') or intermittently (called ''cyclic steam injection''). In the steam drive process, steam is injected in several injection wells and the oil is produced in several production well. In cyclic injection process, steam is injected in several (previously producing) wells for 2–6 weeks, soaked for 3–7 days, and produced back for a few weeks or months. This cycle can be repeated several times. Often, the steam flood is preceded by cyclic injection. Steam injection has been used commercially for several decades. In fire flooding or in situ combustion process, air is injected and ignited inside the reservoir. A combustion front propagates through the reservoir pushing the oil out through production wells. A part of the oil is burnt to provide the energy to push the oil out. The variations of this process are wet combustion (where water is injected in the injection well after establishment of the combustion front) and reverse combustion (where the injection and production wells are switched after the establishment of the combustion front). The main attraction of this process is that it uses the most available and the cheapest fluid (air) as the injectant. The in situ combustion method has been field tested, but not commonly used commercially. Incremental production cost for steam injection and in situ combustion are estimated to be \$3–6 and \$5–10 per barrel of oil, respectively.[b]

Recovery Mechanisms

In steam injection, temperature of the oil is increased, which lowers the oil viscosity, thereby making it easier for oil to flow.[16] In fact, four distinct zones develop during steam flooding. The zone close to the injection well is the steam-saturated zone; oil saturation is reduced here to residual oil saturation (typically less than 0.2). The zone just ahead of the steam zone is the condensing zone; hot water at the boiling temperature is formed from condensation of steam. Ahead of this zone is the hot water zone, where the temperature decreases upstream from the boiling temperature to the reservoir temperature. Ahead of the hot water zone is the oil bank at its original reservoir temperature. Some amount of gravity override takes place because steam is lighter than water and oil. Steam distills lighter components from the oil, which move downstream and condense into oil as temperature falls enriching the downstream oil. This process also reduces the viscosity. The oil also swells in the heated zones, thus, reducing the amount of oil left behind. Cyclic steam stimulation increases oil recovery by reduction of viscosity, wellbore cleanup of asphaltic/paraffinic material, and pressure gradient because of the presence of injected steam.

In situ combustion recovers oil by reduction of viscosity, upgradation of the oil, reduction of residual oil, in situ steam flooding, in situ miscible flooding, thermal expansion, and wettability changes.[17] The oil is ignited with the injected air or oxygen. The combustion zone moves toward the production well as long as oxygen is supplied to it. Temperature increases to 600–1200°F, deposited coke is burnt by the oxygen, and hot gases with steam are formed. Downstream of the combustion zone is a vaporization zone, where water and lighter hydrocarbons from oil are vaporized,

[b]Taken from the website wysiwyg://256http://www.fe.doe.gov/oil_gas/res_efficiency/res_progareas.shtml

heavier hydrocarbons are thermally cracked, and coke is deposited. Steam, CO_2, and hydrocarbon gases move forward because of their lower viscosity, cool down by the existing oil/reservoir, and form a condensation zone ahead of the vaporization zone. The oil is enriched by the light hydrocarbon gases/CO_2 and its viscosity is lowered. The residual oil is reduced because the oil is enriched. Water saturation increases because of steam condensation. This water bank pushes the oil downstream. A part of the oil is burnt in the process and is never recovered.

The two main problems in these processes are bypassing and heat loss. Both in situ combustion and steam flooding are affected by gravity override because the injected fluid (air or steam) is lighter than the existing oil. The bypassing problem is more severe in the case of in situ combustion. Condensation has a stabilizing effect on the steam fronts. Steam foams have been developed to reduce bypassing for steam floods. Heat losses from the surface equipment, in the wellbore, to the surrounding formation and the water represent significant energy loss in steam floods. To reduce heat loss, the equipment and well bore should be insulated. By increasing the injection rates, the amount of heat loss as fraction of the total heat input can be reduced. This implies the patterns should be relatively small. Also the application of steam floods is restricted to shallow wells (<1000 m).

Screening Criteria

Steam flooding is applicable for viscous oil reservoirs. The oil gravity should be less than 25°API and the viscosity greater than 20 cp.[13] Oil composition is not critical, but having some light ends helps in steam distillation at the back of the displacement front. The reservoir should be thick (>20 ft) and the oil saturation should be high (>50% PV). Sand or sandstone is preferred with high porosity and high permeability (>200 md). The depth should be low (<4000 ft) to minimize the heat loss in the wells. The temperature is not that critical for steamflooding. For in situ combustion, the oil gravity should be between 10 and 40°API. Viscosity should be less than 1000 cp. Having some asphaltic components in the oil helps in coke deposition. The reservoir should be thicker than 10 ft and the oil saturation should be high (>50% PV). Sand or sandstone with high porosity and high permeability (>100 md) is preferred. The depth should be greater than 500 ft. The temperature should be higher than 150°F.

Field Experience

Steam flooding has been practiced in oil fields for many years. Between 1982 and 1992, oil production from steam flood increased from 288,396 to 454,009 b/d.[14] The number of projects remained almost flat at about 118. In situ combustion was much less popular. The oil production from in situ combustion decreased from 10,228 to 4,705 b/d. In Kern River field, California, cyclic steam injection, steam flood, and steam foam have been applied quite successfully.[18] This reservoir is relatively shallow (<1000 ft) and thick (70 ft). The porosity is high (~35%) and so is the permeability (~7600 md). The oil is heavy (13 API). The recovery was 6.8% with primary production, 15% with cyclic steam injection, 28.7% with steam flood, and additional 14% with subsequent steam foam injection. This test showed that steam flooding could be successful; gravity override is significant, but can be curtailed by foam injection. In the Midway-Sunset field, California, in situ combustion process has been tested.[19] This reservoir is an anticline with a dip of 20–45. It is about 130 ft thick at a depth of 2100–2700 ft. The oil is 14.5 API. The primary production is about 17% of OOIP. Ten years of air injection produced 24% OOIP. Gravity stabilization helped stabilize the in situ combustion process. Other technically successful pilots have been conducted elsewhere.

CHEMICAL METHODS

Chemical methods of EOR are usually carried out by using water or brine as the carrier fluid. The exception is foam flooding, which also involves an immiscible gas as the driving fluid. Although various chemical methods can be used by themselves, studies have shown synergism in their combined application.

Polymer Flooding

The idea behind this method is to reduce the driving fluid mobility relative to the oil using high molecular weight polymers.[20] Bypassed oil is reduced, thereby increasing production. From the early studies in the 1960s, it was evident that compared to water flooding, the addition of certain high molecular weight polymers in the driving fluid (water or brine) at relatively low concentrations (in the hundreds of ppm only) reduces the occurrence of fingering or channeling in both areal and vertical sweeps. Essentially, when the polymer solution enters the high permeability zones, the mobility decreases.[21–23] This allows the flow to occur within the low permeability zones.

Materials that are popularly used in polymer flooding are based on ionic polyacrylamides (PAM) and xanthan.[20] Ionic PAM is obtained either by free-radical polymerization of acrylamide monomer with subsequent hydrolysis or by free-radical copolymerization

of acrylamide and acrylic acid monomers followed by neutralization of the acrylic acid segments with sodium hydroxide. The degree of hydrolysis or fraction of ionizable groups ranges from 15% to 35%. Weight average molecular weights are in the order of 10^7 Daltons, and polydispersity indices are in the 2–3 range. In solution, these polymers attain the conformation of isolated chains in a dilute concentration, overlapping chains in the semidilute regime, and entangled network in concentrated regime. Xanthan, a biopolymer, is produced by fermentation of the bacterium *Xanthamonas campestris*. Weight-average molecular weights are in the $4–5 \times 10^6$ Dalton range, and they are monodisperse with polydispersity indices between 1.3 and 1.5. In water or brine solution, xanthan attains a double-stranded conformation stabilized by hydrogen bonds.

Flow characteristics within cores are both shear and elongational because of the presence of converging and diverging zones in the rock matrix. Polymer solutions, under these conditions, will exhibit both shear thinning and shear thickening behavior. Below a certain critical shear rate, the behavior is purely shear thinning. However, above the critical shear rate, shear thickening becomes dominant because of elongational effects. This is borne by the fact that elongational viscosities have been found to be as high as 10^4 times zero-shear viscosities.

Retention of the polymer within the rock matrix can occur with varying effects. There are three known retention mechanisms in the absence of polymer degradation: adsorption, mechanical entrapment, and hydrodynamic retention.[20] Thermodynamically, adsorption involves the removal of water molecules and small ions from the rock surface, and replaced by the polymer. This will occur if the Gibbs free energy change, ΔG, is negative. As ΔG depends on the enthalpy change ΔH, the temperature T, and entropy change ΔS, based on the following equation

$$\Delta G = \Delta H - T\Delta S \quad (1)$$

a decrease in the entropy must be offset by an increase in enthalpy to prevent adsorption. The positive entropy change occurs because of the removal of water molecules and small ions from the surface; adsorption of polymer chains results in only a slight loss in chain conformation. For a positive enthalpy change, anionic polymer species are needed if the rock surface is also anionic. This is the case with ionic PAM in most sandstone formations. In terms of adsorption rate, lower molecular weight polymer species are favored to reach the rock surface. Eventually, higher molecular weight species reach the surface and tend to adsorb more than their lower molecular weight counterparts. Usually, adsorption is delayed by minutes to weeks because of the time it takes for polymer species to reach the solid surface.

Mechanical entrapment occurs when larger polymer molecules cannot pass through smaller pore throats. On the other hand, hydrodynamic retention occurs when the polymer concentration increases at pore entrances, thereby reducing the polymer translational diffusion rate. In severe circumstances when the polymer solution contains microgels, hydrodynamic retention can result in cake formation leading to polymer plugging.

Polymer degradation within the rock matrix can occur by chemical-, mechanical-, and biological-based mechanisms. Their detriment is viscosity reduction. For ionic PAM and xanthan polymers, chemical degradation can occur through free radical and hydrolytic attack. Ionic PAM is usually prone to free radical-based chain scission degradation because of its relatively high molecular weight. Free radicals are formed from reducing agents, traces of transition metal catalysts, and oxygen. Usually, hydroperoxides are formed, which can also cause propagation reactions. Hydrolytic attack occurs in the acrylamide segments of ionic PAM at high and low pH values, especially at relatively high temperatures. Even at neutral pH at relatively high brine concentrations, ionic PAM precipitates because of the exchange of sodium ions on the polymer with calcium and magnesium ions in the brine.

Mechanical degradation is based on the elongational flow of the polymer solution. It occurs in ionic PAM because of its relatively high molecular weight, while xanthan seems to be stable to mechanical degradation.

Bacterial-based degradation of polymers occurs during storage or in cooler regions of the reservoir. Usually, biocides are used, but they readily adsorb on rock surfaces.

Based on overall performance, as well as problems encountered, incremental production cost of polymer flooding is cited to be between \$5 and \$10 per barrel of oil.[c]

Alkaline Flooding

As the presence of multivalent salts normally results in flocculation of the polymers used in chemical-assisted EOR, the introduction of a monovalent alkaline solution has been employed. As a separate downstream slug, alkaline solutions are normally used as sacrificial materials to obtain synergistic effects with surfactant and polymer flooding methods. Alkaline materials that

[c]Taken from the website wysiwyg://256http://www.fe.doe.gov/oil_gas/res_efficiency/res_progareas.shtml

Table 1 Trends for material and external variables that promote Type III surfactant behavior

Material and external variables	Trends that promote Type III behavior
Oil	Decrease in molecular weight Decrease in branching (aromaticity)
Brine	Increase in salinity Increase in divalent/monovalent ratio Decrease in pH (for carboxylated surfactants)
Amphiphiles	Increase in molecular weight (or branching) of surfactant-hydrophobic part Decrease in polarity of surfactant-hydrophilic part Increase in long-chain alcohol concentration Decrease in short-chain alcohol concentration
External variables	Decrease in temperature (for anionic surfactants) Increase in temperature (for nonionic surfactants) Decrease in pressure

(From Ref.[27].)

are commonly used include.[24] sodium hydroxide, potassium hydroxide, sodium orthosilicate, sodium metasilicate, sodium carbonate, and potassium carbonate.

At the outset, it was noted that alkaline agents lower the interfacial tension (IFT) within the reservoir.[25] The accepted IFT-lowering mechanism is the neutralization of carboxylic acid-containing compounds (asphaltenes and resins) in the oil to form surface-active agents.[26] Mobility control can be affected by the design of a combination of monovalent and divalent cations resulting from the action of the alkaline agent solution. Surfactants produced from monovalent alkaline cations tend to be oil-in-water (O/W) types, while divalent alkaline cations result in water-in-oil (W/O) surfactants. The optimum is somewhere in between, which forms a bicontinuous micellar structure. This condition is difficult to uniformly attain across the reservoir because of consumption of monovalent cation alkaline material through ion exchange with hard brine, formation of divalent cations from the rock surfaces at high pH, and precipitation followed by coalescence of divalent-based surfactants produced from the oil. Such control difficulties make alkaline flooding design site specific at best, and it is better to effect mobility control by some other means, such as the use of a surfactant slug. Eventually, alkaline agents have been found to be more conveniently used as a preflush or as a leading slug before the surfactant slug.

Surfactant Flooding

This method is also known as micellar flooding, microemulsion flooding, or low tension water flooding.[27] The primary effect of the use of surfactants is the lowering of the interfacial tension between the driving fluid and the oil. More formally the capillary number, N_c, is defined as

$$N_c = \frac{\mu V}{\sigma} \tag{2}$$

where V is the interstitial fluid velocity, μ is the fluid viscosity, and σ the interfacial tension, must be increased by three or four orders of magnitude from a critical $N_c = 10^{-6}$. Amphiphilic surfactant molecules partition themselves at the oil–water interface to achieve this effect; their hydrophilic heads locate themselves in the water side, while their hydrophobic tails locate themselves in the oil side of the interface. Surfactants used in EOR normally have anionic hydrophilic heads, which tend to have lower adsorption on reservoir rock. Also, anionic surfactants are normally available in large quantities at a lower cost when compared to nonionic and cationic surfactants. Chemically, anionic surfactants commonly used in micellar EOR are sodium alkylsulfates or sodium alkylarylsulfonates. One drawback of anionic surfactants is their salt sensitivity, especially to multivalent cation salts. In hard brines, ion exchange between the sodium counterion in anionic surfactants and calcium (or magnesium and iron) in the brine causes the surfactant to coagulate. Thus, nonionic surfactants and anionic surfactants with ethoxy and propoxy groups have been studied extensively for EOR applications because of their higher salt tolerance when compared to anionic surfactants.

From a morphological standpoint, surfactants are oil-in-water (O/W) type, water-in-oil (W/O) type, or somewhere in-between.[27] O/W surfactants form micellar domains in water, within which contain oil particles; this is called the Windsor Type I structure. W/O surfactants form micellar domains in oil, within which contain water particles; this is called the Windsor Type II structure. The so-called Windsor Type III structure is bicontinuous in nature, and it is preferred

because of its relatively low interfacial tension and relatively high viscosity. Table 1 shows the various factors that promote Type III behavior.

There are two possible modes of injecting the surfactant slug.[28] One is a relatively high % PV slug (15–60% PV or more) with relatively low surfactant concentration. This mode of injection has been found to yield a more sustained production at lower oil proportion in the receiver wells. The other approach is the injection of low 3–20% PV slug at a relatively high surfactant concentration, which has been found to yield higher oil proportions at less sustained levels of production. Normally, the surfactant slug is followed by a polymer solution slug, which is in turn followed by a slug of water or brine. The use of an alkaline slug ahead of the surfactant has been found to be beneficial as a preflush,[29] which has the effect of reducing the adsorption of the surfactant onto the rock surfaces.

From an economic standpoint, a surfactant-based EOR has been found to involve an incremental production cost of $8–12 per barrel of oil.[d]

MICROBIAL METHODS

Microbial enhanced oil recovery (MEOR) typically involves the use of bacterial cells for in situ production of chemicals in an oil well. The method has been found to be appropriate in recovering more oil from stripper wells (producing less than 10 barrels per day);[30] thus, economic constraints result in restricted operating procedures. The organisms, nutrients, such as molasses, and air (for aerobic operation), are fed into the wellbore. Then, the well is closed for up to a few weeks to allow the cells to multiply and produce EOR chemicals, which can include gases, solvents, acids, biomass, and natural polymers.

Incremental production cost for MEOR was estimated to be $1–8 per barrel of oil.[d]

CONCLUSIONS

Enhanced oil recovery has been shown to be a viable approach toward the removal of additional oil from partially exploited fields. This is more evident as domestic production continues to gradually drop while demand keeps on increasing. Preference to specific methods depends on various factors, such as the nature of the field, economies of raw materials and utilities, and government codes and regulations.

[d]Taken from the website wysiwyg://256http://www.fe.doe.gov/oil_gas/res_efficiency/res_progareas.shtml

REFERENCES

1. Stalkup, F.I. Miscible displacement. In *SPE Monograph*; SPE of AIME: Dallas, 1983; Vol. 8.
2. Lake, L.W. *Enhanced Oil Recovery*; Prentice Hall: Englewood Cliffs, NJ, 1989.
3. Chatzis, I.; Kantzas, A.; Dullien, F.A.L. On the investigation of gravity-assisted inert gas injection using micromodels, long berea sandstone cores, and computer-assisted tomography. Proceedings of the SPE of AIME, Houston, TX, Oct 2–5, 1988; SPE 18284.
4. Mani, V.; Mohanty, K.K. Effect of spreading coefficient on three-phase flow in porous media. J. Coll. Interf. Sci. **1997**, *187*, 45–56.
5. Wang, Y.; Orr, F.M., Jr. Analytical calculation of minimum miscibility pressure. Fluid Phase Equilibria **1997**, *139*, 101–124.
6. Lang, E. Correlation and prediction of residual oil saturation for gas injection EOR processes. Proceedings of the SPE/DOE EOR Symposium, Tulsa, OK, Apr 21–24, 1996; SPE/DOE 35425.
7. Sebastian, H.M.; Winger, R.S.; Renner, T.A. Correlation of minimum miscibility pressure for impure CO_2 streams. Proceedings of the SPE/DOE EOR Symposium, Tulsa, OK, Apr 21–24, 1984; SPE/DOE 12648.
8. Benham, A.L.; Dowden, W.E.; Kunzman, W.J. Miscible fluid displacement-prediction of miscibility. Trans. AIME **1961**, *219*, 229–237.
9. Mohanty, K.K.; Masino, W.H., Jr.; Ma, T.D.; Nash, L.J. Role of three-hydrocarbon-phase flow in a gas displacement process. SPE Reserv. Eng. **1995**, *10*, 214–211.
10. Pande, K.K. Effect of gravity and viscous crossflow on hydrocarbon miscible flood performance in heterogeneous reservoirs. 67th Annual Conference of SPE, Washington D.C., Oct 4–7, 1992; SPE 24935.
11. Heller, J.P.; Lien, C.L.; Kintamukula, M.S. Foam-like dispersions for mobility control in CO_2 flood. Proceedings of the SPE of AIME, New Orleans, LA, Oct 5–8, 1982; SPE 11233.
12. Stern, D. Mechanisms of miscible oil recovery: effects of pore-level fluid distribution. Proceedings of the 66th Annual Conference of SPE, Dallas, TX, Oct 6–9, 1991; SPE 22652.
13. Taber, J.J.; Martin, F.D.; Seright, R.S. EOR screening criteria revisited. SPE Reserv. Eng. **1997**, *11*, 189–198.
14. Moritis, G. EOR survey. Oil & Gas J. **1992**, 51–79.
15. Hadlow, R.E. Update of industry experience with CO_2 injection. Proceedings of ATCE of SPE, 1992; SPE 24928.
16. Prats, M. Miscible displacement. In *SPE Monograph*; SPE of AIME: Dallas, 1982; Vol. 7.

17. Carcoana, A. *Applied Enhanced Oil Recovery*; Prentice Hall: Englewood Cliffs, NJ, 1991.
18. Patzek, T.W.; Koinis, M.T. Kern river steam foam pilots. Proceedings of the SPE/DOE EOR Symposium, Tulsa, OK, Apr 17–20, 1988; SPE/DOE 17380.
19. Gates, C.F.; Sklar, I. *SPE Reprint Series No. 7*; SPE: Richardson, TX, 1985; 385–394.
20. Chauveteau, G.; Sorbie, K.S. Mobility control by polymers. In *Basic Concepts in Enhanced Oil Recovery—Chapter 2*; Critical Report on Applied Chemistry; Baviere, M., Ed.; Elsevier: London, New York, 1991; Vol. 33.
21. Sandiford, B.B. Laboratory and field studies of well floods using polymer solutions to increase oil recoveries. Pet. Trans. AIME **1964**, *231*, 917–929.
22. Gogarty, W.B. Rheological properties of pseudoplastic fluids in porous media. Soc. Pet. Eng. J. **1967**, *7* (2), 161–170.
23. Szabo, M.T. Laboratory investigations of factors influencing polymer flood. Soc. Pet. Eng. J. **1975**, *15* (4), 338–346.
24. Labrid, J. The use of alkaline agents in enhanced oil recovery processes. In *Basic Concepts in Enhanced Oil Recovery—Chapter 4*; Critical Report on Applied Chemistry; Baviere, M., Ed.; Elsevier: London, New York, 1991; Vol. 33.
25. Wagner, O.R.; Leach, R.O. Improving oil displacement by wettability adjustment. Soc. Pet. Eng. J. **1966**, *6* (4), 335–147.
26. Huh, C. Interfacial tensions and solubilizing ability of a microemulsion phase that coexists with oil and brine. J. Colloid Interface Sci. **1979**, *71* (2), 408–425.
27. Pope, G.A.; Baviere, M. Reduction of capillary forces by surfactants. In *Basic Concepts in Enhanced Oil Recovery—Chapter 3*; Critical Report on Applied Chemistry; Baviere, M., Ed.; Elsevier: London, New York, 1991; Vol. 33.
28. Gogarty, W.B. Oil recovery with surfactants: history and current appraisal. In *Improved Oil Recovery by Surfactant and Polymer Flooding*; Shah, D.O., Schechter, R.S., Eds.; Academic Press: New York, 1977; 27–54.
29. Wilson, L.A., Jr. Physico-chemical environment of petroleum reservoirs in relation to oil recovery. In *Improved Oil Recovery by Surfactant and Polymer Flooding*; Shah, D.O., Schechter, R.S., Eds.; Academic Press: New York, 1977; 1–26.
30. Stosur, G.J. Unconventional EOR concepts. In *Basic Concepts in Enhanced Oil Recovery—Chapter 11*; Critical Report on Applied Chemistry; Baviere, M., Ed.; Elsevier: London, New York, 1991; Vol. 33.

Environmental Chemodynamics

Louis J. Thibodeaux
Louisiana State University, Baton Rouge, Louisiana, U.S.A.

INTRODUCTION

Environmental chemodynamics (EC) is a multiscience and engineering subject used by natural scientists and applied professionals for the conceptual understanding and quantitative tracking of anthropogenic and natural chemicals in their movements from the places of origin, along the various pathways of migration within and across media interfaces of the earth's geosystems (i.e., air, water, and soil) to points far and near where elements of the ecosystem are impacted. The participants in this multidisciplinary field include environmental chemists and engineers, geochemists, geologists, geophysicists, agricultural chemists, chemical engineers, environmental toxicologists and biologists, soil scientists, public health professionals, and other environmental scientists, practitioners, and specialists. One particular important application area in the field is the tracking of substances with hazardous and toxic properties targeting their contact with sensitive biological species including humans. The engineering design aspect of EC consists of defining pathways, quantifying chemical mobility, and using mathematical descriptions of key chemical, physical, and biological processes in forecasting their presence in space and time within air, water, and soil media primarily. This chemical presence is manifest as fluxes across media interfaces, and concentrations and residence times within the various media. The purpose of this entry is to define EC, describe the role it plays in chemical engineering and in environmental engineering science in general, and give examples of its application.

USER GROUPS

The chemical manufacturing process has an EC connection. Commencing with the extraction of raw materials from the earth to the delivery and use of the final products it is inevitable that hazardous substances will be formed. This is the first law (refer to Table 1 for the five natural laws of hazardous waste).[1,2] The use of ''green engineering'' and product substitution will substantially reduce the mass of hazardous materials produced but will not eliminate the problem entirely. The second law concerns recycle and reuse, but it can only be a partial solution, so the manufacturing engineer must have a basic understanding of the expected behavior of these chemical residues being discharged in the media adjoining the facility. Environmental chemodynamics provides a basis for obtaining a qualitative understanding of the relevant transport and transformation processes acting upon these substances as they enter the natural media (i.e., air, water, soil, and sediments). In addition, models are available in various levels of complexity, from vignette models to supercomputer versions, for use by the engineer to quantitatively track these residues in a real-time sense as they are released and move away from the plant-site. An example is the accidental spill of a quantity of chemical onsite. Regulatory reporting requirements increasingly demand an assessment of the mass fractions evaporated to air entering surface waters, percolating to groundwater and retained on the soil solids. Environmental chemodynamics concepts and models make it possible to understand the chemical behavior patterns in this case and provide tools in the form of mathematical algorithms by which to quantify the respective mass distributions.

The treatment of waste streams generated within industrial and commercial manufacturing, by municipalities, agricultural operations, etc. also has an EC connection. Typically environmental (civil) engineers or the equivalent applied environmental science professionals need to assess the multimedia impacts of these waste streams during the design and operation of the waste treatment facilities. These facilities are designed primarily to produce dischargeable residues, which are compatible with earthen materials (natural law no. 3 in Table 1). Other uses of EC by this group are much the same as those of the manufacturing process engineer.

The combined efforts of the manufacturing engineers, the waste treatment engineers, and pollution control applied scientists not withstanding, residual leaks are inevitable (natural law no. 4 in Table 1). It is clear that quantities of hazardous substances will always enter the natural environment. The earth forces that result in volcanic chemical emissions and the hazardous substances formed by many plants and animals for example have been ever present. However, it is the impact of human activities that has been the most

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030913

Table 1 The natural laws of hazardous waste

"I am, therefore I pollute;" undeniably, the production of some waste by beings and machines is not preventable
"Recycle, reuse, and minimization" are only partial solutions to waste production
"Convert remaining waste" to earthen-like materials that are environmentally compatible
"Small waste leaks are unavoidable" and acceptable
"Nature sets the standards" for the earthen-like forms and acceptable leak quantities

persistent and constant generator, increasing the quantities of hazardous substance in the last 10,000 years. The impact of these activities has been and is being monitored by environmental chemists, geochemists, biologists, toxicologists, and other environmental professionals—both scientists and engineers. Observing and cataloging chemical concentration levels in media and biota compartments and measuring organism response is the main objective of these skilled professionals. This knowledge coupled with EC protocols that are concerned with quantifying the chemical kinetics of sources and sinks, movement routes, partitioning, transport, etc., can provide needed key information relevant to the forensic aspects of establishing the cause and effect of the impacts encountered during monitoring. In this regard, EC knowledge is applied in a qualitative way aimed at understanding and explaining pathways and routes. In a more quantitative way by using predictive models, EC offers independent evidence of cross-media transport, media concentration levels, and exposure times for organism uptakes. In addition to elucidating the exposure pathway scenario, the EC analysis also provides a theoretical basis and a means for designing systems barriers for eliminating or deflecting the exposure pathway. An example of this is the use of clean soils or other natural material to cap a contaminated aquatic sediment bed surface; the imposed barrier significantly (i.e., 99%) reduces the chemical flux to water.

The fifth and final natural law of hazardous waste in Table 1 acknowledges that "nature" sets the standards for what is acceptable as to the presence of small quantities of hazardous substances in the media. Nature in this instance means the biological forms that encounter and are impacted by these substances in air, water, soils, etc. Regulators have been using EC protocols as one of the many needed in assessing chemical exposure to humans and other key biological species. These efforts have culminated recently in the design of the first multimedia, multipathway and multireceptor risk assessment tools by the USEPA.[3] Monitoring the presence of and movement of small quantities in the media is a difficult task which is both time consuming and expensive. Environmental chemodynamics modeling provides an economical means of extending this hard to get data. The advantage EC models have over monitoring is their ability to make predictions in time and space that extend and enhance the laboratory and field measurements. Increasingly, more sophisticated EC models will need to be developed, tested, and used. Historical evidence abounds as to the impact the physical activities by humans have had on altering the earth's biological status.[4] The mass production of chemical devices with hazardous properties which commenced in earnest in the early twentieth century has the potential for massive negative impacts as well.

THE SOURCE TERM FOR HAZARDOUS SUBSTANCES

Flux is the unique EC contribution; it connects chemical sources of the adjoining media. Primary sources that place hazardous substances in the media are end-of-pipe entry that include "smoke-stack" emissions to air from factories and automobile exhaust pipes, liquid and solid slurry sewer pipe outfalls to receiving waterbodies, the placement of pesticides onto agricultural soils, and the spreading of solid waste onto the land, in addition to the numerous inadvertent spills onto the land and water or release to air. These primary sources are always the precursors of numerous secondary chemical sources. The formation of secondary sources occurs for example when waste quantities accumulate and remained in areas adjacent to or in close proximity to many of these primary source points of entry. An example is fall-out of particles from the stack emission plumes deposited on the nearby soils forming a contaminated surface layer. In water bodies near sewer out-falls, particle settling and nonaqueous liquid discharges have formed layers of contaminated sediments. An example for the land application of waste was noted previously. These are just a few examples of media locales that become secondary hazardous substance sources. A unique focus of EC is to formulate the mechanism of release so as to realistically quantify the chemical flux from the secondary sources to the adjoining media.

Chemical transport and transformation have been a part of environmental science and engineering for decades. Air pollutant plume dispersion modeling and surface water quality stream modeling are mature elements of EC, commonly termed chemical fate and

transport (CFaT) models. They were developed for the tracking of chemicals and particles into air and water from the primary emission sources. It became clear to these professionals that the primary emission sources were giving rise to secondary source terms. In recent decades, ground water pollution has extended CFaT modeling into the subterranean, where both primary and secondary source terms were often involved. It was also becoming apparent that many so-called waste management practices were also creating secondary waste sources. Waste water treatment operations that emitted volatile chemicals to the air and waste piles that leached soluble chemicals to surface or ground water are well-known examples.

These secondary source terms are a particularly tough technical challenge to the air, water, and subterranean CFaT modeling efforts. Early attempts treated these as constant or as time varying forcing functions for describing the input rates or fluxes into the adjoining media. In doing so, the problem of quantifying the chemical inputs from the adjoining phase was grossly simplified and the multimedia nature of the problem was side-stepped. For example, in air modeling of these secondary sources, a time-series of chemical fluxes was selected so as to match the model projected downwind concentrations with the measured ones. Although the technique is operationally effective and is still being used, it nonetheless often skirts the hard problem of understanding and realistically quantifying the secondary source emission process. The technical problem of modeling the secondary sources is typical of all multimedia CFaT models. Developments in the waste management field now focus on the notion that chemicals in the environment needed a multimedia approach. Attempting to model and manage chemicals on a single media basis is a very unrealistic approach in most instances.

The unique contribution of cross-media flux assessment brought about by the EC approach allows the full and proper scientific attention to the mechanistic processes that occur within the secondary source media. Adding the source term to the existing suite of single media CFaT models essentially closes the chemical loop, technically speaking. With the development of science-based concepts, realistic chemical process models have become available for describing the cross-media transfers. With this piece in place, the way was clear to begin developing truly multimedia environmental models. The fugacity approach to multimedia modeling using compartments is possibly the best known of this genre.[5]

ORIGINS OF EC

Currently, chemical engineering is the applied science field that may best address the needs of environmental chemodynamics. It contained several of the ''right stuff'' basic science components: they were mathematics, chemistry, and physics. The ''science'' of chemical engineering is that body of knowledge that provides for the integrating of these basic sciences into a mechanistic process for the understanding of and developing quantitative design tools for predicting, controlling, and manipulating chemical behavior for product manufacture. ''Classical'' process design for chemical manufacturing was a mature engineering activity in the 1960s. Beyond manufacturing, the chemical engineering science paradigm provides the theoretical and practical wherewithal for developing environmental chemodynamics processes. However, to be completely applicable to the new EC field, it was necessary to assimilate large doses of the biological sciences and geosciences (i.e., geology, geophysics, and geochemistry). The processing paradigm of chemical engineering could naturally assimilate these additional sciences; however, the certainty the classical approach enjoyed in specifying design and control was compromised. Classical chemical engineering manufacture requires strict process design and control over vessels. The waste produced and the manufactured products released pose a particular challenge when they enter the natural system; this chemical processing system was not designed by scientists or engineers and it is in practical terms not under their control. The addition of the biological and geoscience components so necessary to environmental chemodynamics introduced additional complexities in the original. The net result was an increased level of uncertainties to the ''designs'' required of the new field. These ''designs'' in the context of EC included forecasting of chemical concentrations in time and space and predicting the chemical fluxes and time-of-persistence in the media compartments. Although unsettling for the novice chemical engineer working in the EC area, high levels of uncertainties doing predictive ''designs'' in natural settings are a reality and must be accommodated in the geoscience field.

AN EC SYLLABUS

The EC subject matter was selected and developed for use as a technical elective course for the curriculum in chemical engineering. The course contents appear in Table 2. Although developed for seniors and first year graduate students, it has been widely accepted and adopted in university departments of civil and environmental engineering and environmental chemistry. In addition, it is used widely in the geosciences, agricultural chemistry, agronomy, and environmental science departments at the graduate level. It contains introductory material, two theoretical chapters, and four applied chapters. Only a working knowledge of

Table 2 Environmental chemodynamics course content

Introduction	
1.1	Introduction to environmental chemistry and engineering
1.2	Illustration of objectives and content: re-aeration of natural streams
Equilibrium at environmental interfaces	
2.1	Chemical equilibrium at environmental interfaces
2.2	Thermal equilibrium at environmental interfaces
2.3	Chemical equilibrium models for environmental compartments
Transport fundamentals	
3.1	Diffusion and mass transfer
3.2	Turbulence in the environment
3.3	Other transport topics
Chemical exchange between air and water	
4.1	Desorption of gases and liquids from aerated basins and rivers
4.2	Exchange of chemicals across the air–water interface of lakes and oceans
4.3	Heat transfer across the air–water interface
Chemical exchange between water and adjoining earthen material	
5.1	Chemical transport at bottom of flowing streams
5.2	Chemical movement at bottoms of ponds, lakes, and quiescent water bodies
5.3	Chemical movement at bottoms of estuaries and oceans
5.4	Thermal energy movement across the sediment–water interface
Chemical exchange between air and soil	
6.1	Thermal turbulence above air–soil interface
6.2	Chemical flux rates through lower layer of atmosphere
6.3	Chemical flux rates through upper layer of earthen material
6.4	Heat transfer at air–soil interface
Intraphase chemical transport and fate	
7.1	Chemical transport and fate in surface waters
7.2	Chemical transport and fate within atmospheric boundary layer
7.3	Chemical transport and fate within subterranean media

algebra is needed for the use and interpretation of the many vignette environmental mathematical models presented. Some knowledge of calculus and differential equations is convenient for understanding the development of selected models and for developing alternatives. Numerous worked examples illustrate the application of the EC models.

The EC contribution to the science of chemical behavior in the geospheres was in providing theory-based general concepts and a practical understanding for the means of quantifying chemical exchanges between adjoining media (i.e., air-to-water, water-to-sediment, and soil-to-air). The necessary thermodynamic and transport theories appear in chapters 2 and 3 (Table 2). A focus on the interface regions between phases allow for the closing of the chemical loop and connecting the adjoining phases, thus allowing the development of truly integrated multimedia environmental models. The importance of this element was noted above in discussing secondary sources of chemical emissions into the media. The cross-media material course topics, the heart of the subject, appear in chapters 4, 5, and 6 (Table 2). By focusing on mechanisms on either side of the interfaces in turn, algorithms for interphase chemical movement are developed. A qualitative understanding of processes in each phase near each interface is developed prior to offering qualitative descriptions and the final algorithms.

The first section (i.e., 1 in Table 2) serves as an introduction and defines the scope of the subject. As implied in the title, it is one of chemodynamics or the movement of chemicals. Chemical transport is the primary focus of the material. Critics have noted that production and degradation rates of chemical reactions are all but absent in the course syllabus. Environmental reaction is a very important but is also a very broad subject and its inclusion at even a basic technical level into EC would detract from the transport message. Two basic subjects are necessary for understanding transport. These are chemical equilibrium at interfaces and the fundamentals of transport phenomena. Highly condensed material on these two key subjects are presented in chapters 2 and 3. The last chapter, number 7, is on the fate and transport in water, air, and soil. These are the traditional subjects of environmental modeling which treat each of the three media separately and as isolated units from a multimedia perspective. Nevertheless, this approach is very appropriate for numerous EC applications. The section stresses the commonalities of fate and transport in the three media; however, the brief coverage offered on each belies the importance of these respective intraphase transport topics.

An example follows that is used to illustrate some of the details of the EC analysis process and its forecasting characteristics. The general reader may skip over this technical section without loss of continuity of the subject. It involves an important and very common EC problem and illustrates the effective use of the flux concept in connecting the interphase chemical movement in a multimedia context.

AN EC EXAMPLE

Methods of measuring and modeling volatile chemical emissions to air from a surface soil source are common and important EC ''design'' application. The examples

are many. Spills of organic and inorganic liquids and solids will continue to occur. The direct placement of pesticides on to surface soils and the creation of soil-like solid waste piles containing volatiles are necessary activities. The latter includes the disposal of dredging materials (DM) derived from remediation contaminated bed sediment of harbors, bayous, lakes, estuaries, etc. Large areas of soil-like drying mud flats are formed that become sources of volatiles. In the case of the planned dredging of the Indiana Harbor and Canal near Chicago, U.S.A., the DM disposal operation will involve a 68 ha confined disposal facility (CDF) to be operated for 30+ years. Prior to building, the CDF estimates of chemical fluxes to air including naphthalene are needed. This and other polycyclic aromatic hydrocarbons will contribute to degrading the local air quality in addition to creating a breathing air risk to nearby residents. In this case, the scale of the naphthalene emission problem is a local one. On the other hand, volatile mercury problem involves a similar emission process, but its impact is on a much larger scale. Its flux from naturally enriched surface soils plays a significant role in its global chemodynamics as an intrinsic component of regional and global Hg budgets.[6] Examples of other types of common air emission include accidental spills and direct application to soils. All have a common EC process base both in the technologies used for measuring their magnitudes and in the multimedia models used to forecast the emission through time. The following two sections contain brief EC designs for both.

Zhang et al.[6] observed that the emission flux of a gaseous substance has to be estimated using a combination of theoretical models and data collection approaches because it is not a physical quantity that can be measured directly. The three commonly used estimation methodologies include in-air micrometeorological methods, spatial mass balance methods, and enclosed (flux chamber) methods. The latter is the most flexible technology and useful for exchange process studies; of these the dynamic flux chamber (DFC) is widely used.[6,7] Its operation involves a box-type enclosure placed upon a soil area A (m^2) with flushing air continuously pulled (or pushed) through to capture and deliver the emitted gases or vapors to an appropriate adsorbent trap for eventual chemical identification and quantification. The results of a series of steady-state DFC experiments with a naturally enriched Hg soil appear in Fig. 1. Here it is observed that as the air flushing rate through the chamber, Q (m^3/hr), increases, so does the flux of total gaseous mercury (TGM), n (ng/m^2 hr), from the soil. Both the collected flux data and the theoretical model flux developed for the process demonstrate this increase with an asymptotic approach to a maximum flux, n_{mx}. The important

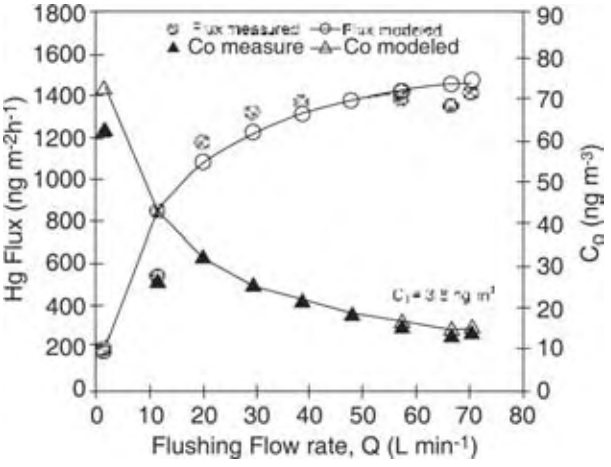


Fig. 1 Steady-state flux chamber measurements and model results of TGM moving from soil to air. (From Ref.[7].)

model equation is

$$n = \frac{n_{mx}}{1 + (\beta Sk_s + Ak_s)/Q} \quad (1)$$

where $\beta = 31$ is a consonant, $S(m^2)$ is the vertical chamber area, and k_s (m/hr) is the mass transfer coefficient of TGM in the soil surface layer. The decreasing data set shown in Fig. 1 is the outlet concentration of TGM from the chamber. Increasing the air flushing rate dilutes the constant the flux, resulting in decreasing outlet concentrations. Besides providing realistic estimates of the TGM flux to air from representative soil samples, such EC experiments establish the dominant processes that control the chemical-to-air releases. In this case, it is the soil source strength as total soil Hg concentration and the transport kinetic parameters on the soil-side and the air-side of the interface. With this information it is now possible to more confidently build field-scale EC models for use in estimating chemical emissions forecast now and in the future.

Volatile and semivolatile organic chemicals placed on surface soils have very similar chemical process behavior patterns as those shown above for Hg. Numerous laboratory and a few field flux measurements have been performed on this class of chemicals as documented in a recent review.[8] The review was performed to assess the quantity and quality of measured flux data and the availability of process-based emission models. The basic difference between the flux chamber model and the field-scale model is that the former involves a steady-state flux process and the latter is a transient one. The EC models needed for the large time-scale involved with most field applications must consider the so-called ''weathering'' effect which characterizes the depletion of the available-for-volatilization chemical mass in the

surface layers. Traditional chemical engineering modeling approaches to the drying of porous solids including a constant rate (i.e., flux) period and a falling rate period. As ''drying'' time increases, the effective position of the plane of origin of the volatile chemical recedes further into the soil column and thereby it encounters an increasingly larger diffusion pathway length to the air interface.

Decades of research by numerous individuals has uncovered numerous minute facts about this complex soil-side and air-side coupled chemical release process. While lab.-specific and controlled conditions allow sophisticated theoretical mathematical models to be developed that can match measured flux data very well, such elaborate protocols are of little practical use in most field applications. Vignette EC models that capture the theoretical essence of the significant processes and contain a minimum number of parameters that can be transparently adjusted for fine-tuning the model to the site data appears to be the most realistic approach. The following equation is a transient EC flux model that is simple but is theoretically consistent with the known major processes and capable of making quantitative predictions using a few key parameters:

$$n = (W_a H/K_d - C_a) \Big/ \left(\sqrt{\frac{\pi t}{D(\varepsilon_a + H/K_d \rho_b)}} + \frac{1}{k_g} \right) \quad (2)$$

In the equation W_a is the chemical loading onto the soil (mg/kg), H is Henry's constant, and K_d (m^3/kg) is the chemical soil-to-water partition coefficient. C_a is the concentration in the background air. The transport within the soil is dominated by air-filled pore spaces ε_a and is quantified by an effective diffusivity, D (m^2/hr). k_a (m/hr) is the air-side mass transfer coefficient. The independent variable is time t (hr): it represents the increasing time period that commences with the placement or spreading of the soil-like waste material on the land. The result allows time projection estimates of the chemical flux to air. At a certain time period, a shift occurs in the flux process; the vaporization process changes from being air-side resistance controlled, through the $1/k_g$ parameter, to being increasingly soil-side resistance controlled by the group under the square root symbol. In Fig. 2, measured flux data for dibenzofuran[9] and model results using Eq. (2) are shown. However, some challenges remain applying the model to a porous media that undergoes dramatic particle consolidation and volume changes as water is lost by evaporation and percolation.

Numerical values of chemical concentration and time-of-persistence in the media convey the information necessary for quantifying the related health risk.

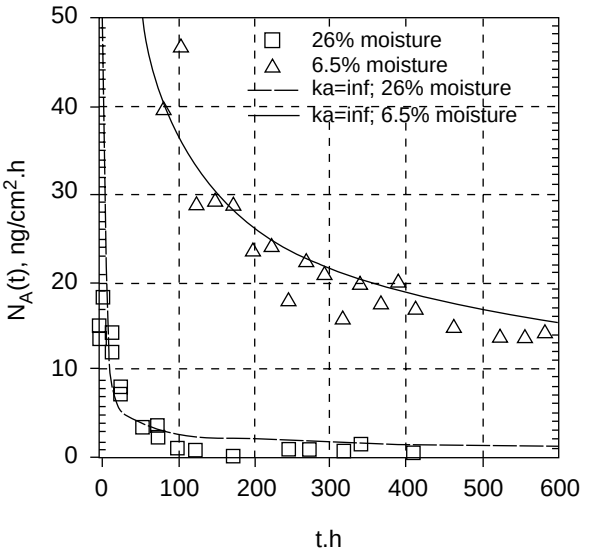


Fig. 2 Transient flux of dibenzofuran from soil to air in laboratory apparatus demonstrating the effect of soil moisture. (From Ref.[9].)

Flux is a somewhat abstract entity and does not relate this information directly. It quantifies the chemical movement rate across an interface plane into a receiving media such as the air boundary layer (BL) in the above example. Only when it is coupled with an air dispersion model does it produce concentrations in air. In the case of a large soil surface area source, a simple relationship exists between flux and concentration. For neutral air stability conditions in the atmospheric BL with steady-state wind speed v (m/sec), the concentration in air, c (mg/m^3), can be approximated by

$$c = \frac{c_a + 27n}{v} \quad (3)$$

where c_a is the concentration in the background air approaching the area source. According to Eq. (2), the flux decreases with increasing time as $n = k/t^{1/2}$

For a constant wind speed, Eq. (3) indicates that the concentration in air above the source decreases with time in a similar fashion.

From an air pathway exposure perspective, the time-of-persistence is also an important factor. In the case of a volatile chemical moving from a soil surface source a time-period, tp (sec), may be defined as the duration during which the concentration level in the air is at or above some predetermined safe level C^*. Using C^* and the simplified flux expression in Eq. (3), the time can be calculated:

$$\text{tp} = \left[\frac{27k}{v(C^* - C_a)} \right]^2 \quad (4)$$

where numerical values of k can be estimated from Eq. (2). This result indicates that for substances with small safe concentration numerical values, the time-of-persistence is very long. For those with high safe concentration levels, the tp can be very short. Taken together the Eqs. (3) and (4) are design algorithms which are useful for estimating the concentrations in air and the time-of-persistence needed for quantifying chemical exposures to biota located in the vicinity. This completes the presentation of EC flux concept in assessing chemical behavior and release as it applies to this important class of secondary emission sources.

CONCLUSIONS

A Brief History

Environmental chemodynamics was 35 years old in 2005. The widespread appearance of pesticide residues in soil, water, and air was the first instance in observing the multiphase spread of an anthropogenic substance. This was attributed to the development of analytical chemistry methods in determining minute amounts of pesticides in various media. Evidence of the widespread presence of chemicals was contained in an American Chemical Society (ACS) report on the cleaning of our environment.[10] 1,1,1-trichloro-2,2-bis(*p*-chlorophenyl)ethane (DDT) is an insecticide used in large quantities since 1942 and its residues attained worldwide distribution.[11] Global modeling of the primary reservoirs—the land surface, the troposphere, the mixed layer of the ocean, and the abyss—attempted to answer the question as to where does it go. This substance played a sentinel role in our appreciation of chemical hazards in the biosphere. A working conference of the National Academy of Sciences and National Academy of Engineering on principles and protocols for evaluating chemicals in the environment was convened in 1973. In the report,[12] which appeared two years later, the term ''chemodynamics'' was first defined. It was lengthened to environmental chemodynamics by Hague and Freed in 1975 upon publishing the proceedings of the 1974 ACS Conference containing papers on the behavior of chemicals in air, water, soil, and biota and included topics on their modeling in the environment, photochemical behavior, adsorption, leaching, and breakdown, vapor loss, interaction with biological macromolecular, and detoxification by biota.[13] These developments essentially launched the new environment field focused primarily on anthropogenic chemicals, but require support from nearly all the basic and applied science fields. Two textbooks[14,16] and a textbook chapter[15] on the subject appeared later in the decade of the 1970s. Fifteen years later numerous books have appeared. Most bear titles using the alternative descriptive phrases such as chemical or pollutant fate and transport in water, air, and soil, environmental chemical modeling, etc.[16–24]

The Future of EC, New Field or Just New Fashion?

The subject is an important concern to the profession of chemical engineering. The material was chosen and organized as a university level course with this targeted audience in mind. Groups that accepted, adopted, and used EC dwarfed the targeted audience whose primary mission remains chemical and materials manufacturing. The nontarget group consisted primarily of environmental scientists and engineers. The EC concept appeared at a key time-period in the growth of the environmental science and engineering movement. Measurable quantities of many hazardous chemicals were being found in all the natural media and existed in at elevated levels in many key biota. Laws regulating waste treatment, pollution control, and discharge limits were proposed, passed, and implemented.

The concern about chemicals in the environment began to move away from being a series of single media problems and approaches to being viewed as a multimedia one. The necessity of and growing reliance of mathematical models to track chemical movements and reactions in nature elevated the need to formulate science-based process algorithms. Surface water quality models and air dispersion models were in common use; later groundwater contaminant models appeared. Together these were generally termed CFaT models. Environmental chemodynamics provided the unifying concept as the means of connecting the single phase models. The fugacity-type models and a host of other multimedia, multiphase, and multicomponent risk assessment models were formulated and are continuing to develop in sophistification and are finding widespread applications.

Whether EC is a new subfield of environmental science/engineering field is irrelevant; the continuing use of the term, first offered by Freed and Hague, may be transitory as is fashion. However, it is unlikely that the subject will diminish in importance as long as humans continue to rely on chemical ''devices'' for enhancing their quality of life. A subfield by any other name that focuses on discovering and understanding the nuances of chemical mobility in nature plus providing quantitative designs on pathways, movement rates, etc. will not go out of fashion.

REFERENCES

1. Thibodeaux, L.J. Hazardous material management in the future. Environ. Sci. Technol. **1990**, *24* (4), 456–459.
2. Constant, W.D.; Thibodeaux, L.J. Integrated waste management via the natural laws. Environmentalist **1993**, *13* (4), 245–253.
3. Marin, C.M.; Guvanasen, V.; Saleem, Z.A. The 3MRA risk assessment framework—a flexible approach for performing multimedia, multipathway and multireceptor risk assessment under uncertainty. Human Ecol. Risk Assess. **2003**, *9*, 1–22.
4. Ponting, C. *A Green History of the World*; Penguin Press: New York, U.S.A., 1991.
5. Mackay, D. *Multimedia Environmental Models—A Fugacity Approach*; Lewis Publishers, Inc.: Chelsea, MI, U.S.A., 1991.
6. Zhang, H.; Lindberg, S.E.; Barnett, M.O.; Vette, A.F.; Gustin, M.S. Dynamic flux chamber measurements of gaseous mercury emissions over soils. Part 1. Simulation of gaseous mercury emissions from soils using a two-resistance exchange interface model. Atmos. Environ. **2002**, *36*, 835–846.
7. Lindberg, S.E.; Zhang, H.; Vette, A.F.; Gustin, M.S.; Barnett, M.O.; Kuiken, T. Dynamic flux chamber measurement of gaseous mercury emissions fluxes over soils. Part 2. Effect of flushing flow rate on verification of a two resistance exchange interface simulation model. Atmos. Environ. **2002**, *36*, 847–859.
8. Thibodeaux, L.J.; Ravikrishna, R.; Valsaraj, K.T. Volatilization Rates from Dredged Material and Soils—A Literature Review. Final Report. U.S. Army Corp of Engineers, Chicago District. Monitored by USA ERDC, WES 3909 Halls Ferry Road, Vicksburg, MS, U.S.A.; February 1, 2002.
9. Valsaraj, K.T.; Ravikrishna, R.; Choy, B.; Reible, D.D.; Thibodeaux, L.J.; Price, C.B.; Yost, S.; Brannon, J.M.; Myers, T.E. Air emissions from exposed contaminated sediments and dredged material. Environ. Sci. Technol. **1999**, *33*, 142–149.
10. Long, F.A.; Price, C.C. Cleaning our environment—the chemical basis for action. Chem. Eng. News **1969**, September 8, 58–69.
11. Woodwell, G.M.; Craig, P.P.; Johnson, H.A. DDT in the biosphere: where does it go? Science **1971**, *174*, 1101–1108.
12. Nelson, N., Chairman. *Principles for Evaluating Chemicals in the Environment*; National Academy of Sciences: Washington, DC, 1975.
13. Hague, R.; Freed, V.H. *Environmental Dynamics of Pesticides*; Plenum Press: New York, U.S.A., 1975.
14. Thibodeaux, L.J. *Chemodynamics—Environmental Movement of Chemicals in Air, Water, and Soils*; John Wiley and Sons, Inc.: New York, U.S.A., 1979.
15. Tinsley, I.J. *Chemical Concepts in Pollutant Behavior*; John Wiley and Sons, Inc.: New York, U.S.A., 1979.
16. Keely, W.B. *Chemicals in the Environment*; Marcel Dekker, Inc.: New York, U.S.A., 1980.
17. Hemond, H.F.; Fechner, E.J. *Chemical Fate and Transport in the Environment*; Academic Press: San Diego, CA, U.S.A., 1994.
18. Schnoor, J.L. *Environmental Modeling—Fate and Transport of Pollutants in Water, Air, and Soil*; John Wiley and Sons, Inc.: New York, U.S.A., 1996.
19. Thibodeaux, L.J. *Environmental Chemodynamics*, 2nd Ed.; John Wiley and Sons, Inc.: New York, U.S.A., 1996.
20. Trapp, S.; Matthies, M. *Chemodynamics and Environmental Modeling*; Springer: Berlin, Germany, 1998.
21. Reible, D.D. *Fundamentals of Environmental Engineering*; Lewis Publishers: Boca Raton, FL, U.S.A., 1998.
22. Logan, B.E. *Environmental Transport Processes*; John Wiley and Sons, Inc.: New York, U.S.A., 1999.
23. Valsaraj, K.T. *Elements of Environmental Engineering*, 2nd Ed.; Lewis Publishers: Boca Raton, FL, U.S.A., 2000.
24. Ramaswami, A.; Milford, J.B.; Small, M.J. *Integrated Environmental Modeling—Pollutant Transport, Fate, and Risk in the Environment*; John Wiley: New York, 2005.

Environmental Law and Policy

Don C. Haddox
Waterbury, Vermont, U.S.A.

Teresa J. Cutright
Department of Civil Engineering, University of Akron, Akron, Ohio, U.S.A.

INTRODUCTION

The subject of environmental law and policy is immense, far too large to be adequately covered within the confines of a small entry. Any single aspect of environmental law could easily encompass an entire encyclopedia but still be found to be lacking in one or more areas. With that caveat, an overview of key environmental policies can be discussed, knowing that not all intricacies of any one program can be fully investigated or developed. Resource Conservation and Recovery Act (RCRA), Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), Safe Drinking Water Act (SDWA), Clean Air Act (CAA), Clean Water Act (CWA), Toxic Substances Control Act (TSCA), Superfund Amendments and Reauthorization Act (SARA), Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), National Environmental Policy Act (NEPA), Wild and Scenic Rivers Act, Endangered Species Act—the list can continue for pages before running out of federal acts that govern the use and protection of the environment. In many cases, individual states have, or are seeking, the authority to enforce the laws as they see fit; provided their implementation is at least as strict as that mandated by the federal statutes. This means that a state cannot implement a law that counters a provision of the federal act; however, it can utilize regulation that is more stringent than those measures called for by the federal law. Many states desire this level of autonomy to shape the environmental direction of their state. Difficulties can arise, from not only the complexity of the legal system, but also the unique obstacles associated with the interdisciplinary nature of the ''environment.'' Chemistry, biology, physics, engineering, health and safety, and long- or short-term sustainability along with the inherent uncertainty of many sciences complicate the ''fact'' vs. ''conjecture'' arguments necessary to provide for legal prosecution or defense.

Environmental law provides the user with an ''organized way of using all of the laws in our legal system to minimize, prevent, punish or remedy the consequences of actions which damage or threaten the environment, public health and safety.''[1] Implementation of the environmental law system entails much more than a compendium of information contained under the pretext ''Environmental Law'' found in a law library. In fact, it can and does involve using any law for the betterment of the environment and/or human health and safety. The practice of environmental law includes all of the following:

- Laws, both federal and state statutes and local ordinances.
- Regulations promulgated by federal, state, and local agencies.
- Court decisions interpreting these laws and regulations.
- The common law.
- The United States Constitution and state constitutions and treaties.[1]

For a bill to become an environmental law it follows the same legislative process as any other bill. That is, it is introduced in either the U.S. House of Representatives or the Senate, and then referred to congressional committees for consideration. A recommendation on whether the bill should pass is then provided. If the bill passes both the House of Representatives and the Senate it becomes an act. The act then becomes law if it is: 1) signed by the President of the United States or 2) not vetoed within 10 days. An additional complication with environmental bills, although not unique to the environmental field, is that the House and Senate generally pass different bills. This then generates the need to have a congressional congress of representatives from both sides to settle the disparity.[1] Resolution of the differences between the groups often results in complex and intricate wording of the bill to appease everyone.

The text of this entry will emphasize a summary of select acts that are generally regarded as important tools in the system of environmental regulation. The main body of each act will be summarized, along with key differences in enforcement authority between federal and state implementation if applicable. Finally, any recent additions or new interpretations along with

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120022526

current actions taken with regard to the act will be explored.

NATIONAL ENVIRONMENTAL POLICY ACT

The National Environmental Policy Act was passed in 1969 and signed into law on January 1, 1970, as public law 91-190 (now found in 42 U.S.C. §§ 4321-4370c). It was, and still remains, the first federal statute that required agencies to evaluate the impact of implementing major programs and alternatives in the early planning stages. This approach makes it the first, truly proactive environmental law.[2] There are two titles in NEPA. Title I contains the declaration of National Environmental Policy while Title II establishes the Council on Environmental Quality (CEQ).

Title I

The primary objectives of NEPA are to encourage the productive and enjoyable harmony between man and environment, promote preventative efforts that eliminate damage to the environment, advance the understanding of ecological systems and natural resources important to the Nation, and establish the CEQ. To achieve these goals, all federal agencies must file a formal Environmental Impact Statement (EIS) for any project or action that is expected to have a ''significant effect'' on the quality of the human environment. The specific requirements and timeline for completing an EIS is contained in Section 102(2)C of NEPA. In brief, a detailed assessment is required for: 1) environmental impact of the proposed action; 2) any adverse environmental effect that cannot be avoided if the action is implemented; 3) potential alternatives and the associated impacts; 4) the relationship between local short-term environmental uses and long-term productivity; and 5) irreversible allocation of natural resources should the proposal be implemented.[3] The assessment of natural resources served as the cornerstone to the future legislation regarding waste minimization.

Title II

The Council on Environmental Quality was established to oversee the completion of Title I. The council comprises three presidential appointees. The council also employs environmental lawyers and scientists to conduct thorough risk-based evaluations of the proposed activities so that the council recommendations provided are sound. Although the CEQ does not have any enforcement provisions, the council has been very influential in its advisory capacity. Currently, enforcement of NEPA has come through suits from private citizens filed under the Administrative Procedure Act. Part of the NEPA's success has been attributed to the successful filing of such suits. However, civil actions have also identified the key drawback with NEPA—the identification of ''significant impact.'' The use of the word significant can be open to interpretation as some of the effects are intangible and different for each person. For instance, some individuals view acid rain deposition on historical buildings as a significant adverse impact, whereas others simply view it as a nuisance.

RESOURCE CONSERVATION AND RECOVERY ACT

The RCRA, enacted in 1976, is found in 42 U.S.C. § 6901 et seq. and is also referred to as the Solid Waste Disposal Act (SWDA). Amended in 1984 by the Hazardous and Solid-Waste Amendments (HSWA), RCRA is designed to control the generation, transportation, storage, and disposal of hazardous materials from the ''cradle-to-grave.'' As of September 1999, approximately 40.7 million tons of hazardous waste is generated annually in the United States and governed by the RCRA.[4] The RCRA additionally provides the means to regulate underground storage tanks (USTs) and manage solid waste (including non-hazardous and/or medical waste); and a national research, development, and demonstration program for innovative or improved waste management and resource conservation.[5]

The RCRA is divided into 10 subtitles, lettered A–J, that assign different aspects of the program (Table 1). Of these 10, subtitles C, D, and I provide the skeleton for the most significant programs in RCRA.

Subtitle C establishes the hazardous-waste management system, which includes defining hazardous. According to RCRA, solid waste is determined to be hazardous because its quantity, concentration, or characteristics (referring to ignitability, corrosivity, reactivity, or toxicity) may:

1. Cause or significantly contribute to an increase in mortality or an increase in serious irreversible or incapacitating reversible illness.
2. Pose a substantial present or potential hazard to human health or the environment when improperly managed.[4]

Besides the obvious characteristics of a given contaminant establishing it as hazardous, a solid waste

Table 1 Subtitles of the RCRA

Subtitle A	Introduction and preamble
Subtitle B	Office of solid-waste; authorities of the Administrator
Subtitle C	Hazardous-waste management
Subtitle D	State or regional solid-waste plans
Subtitle E	Duties of the secretary of commerce in resource and recovery
Subtitle F	Federal responsibilities
Subtitle G	Miscellaneous provisions
Subtitle H	Research, development, demonstration, and information
Subtitle I	Regulation of UST
Subtitle J	Medical waste tracking act

Credit: U.S. Code Title 42, Chapter 82.

may also be listed as a hazardous waste because of its identification with a process or use that has been established by the Environmental Protection Agency (EPA). Three lists have been created: ''F'' for non-specific sources, ''K'' for waste from a specific source, and ''U'' for commercial chemical products. Full investigation and explanation of the listing process (and delisting) are beyond the scope of this entry and the reader is encouraged to seek further information on hazardous-waste identification from outside of this entry.

The other key aspects of Subtitle C outline the standards for generators and transporters of hazardous waste; permitting programs for treatment, storage, and disposal facilities (TSDFs); administration authorization to states; and provisions for enforcement.[6] Classification of TSDFs takes into account the types and amounts of hazardous material, as well as the length of time the material is stored at the facility, before identifying them as large or small generators. Small-quantity generators have 100–1000 kg of hazardous waste and/or <1 kg of acutely hazardous, for a maximum of 90 days. Storing more than 1000 kg of hazardous waste or more than 1 kg of acutely hazardous waste for longer than 180 days triggers a large-quantity TSDF label.[5,6]

The RCRA's Subtitle D establishes the guidelines for instituting solid-waste management programs including state and regional solid-waste plans. Management of sanitary landfills including federal requirements and specific criteria for approval of plans is included within this section. This includes the disposal of bulk, noncontainerized liquids, which is banned, and disposal of containerized liquids, which is on a case-by-case basis. Of interest within Subtitle D is the provision that EPA has no enforcement authority with respect to open dumps. The RCRA grants EPA the ability of enforcement only with respect to hazardous waste. All enforcement regarding nonhazardous materials is left to the discretion of the state.[4] The one control that RCRA imposes on states within Subtitle D is that a state may not impose a ban on importation of wastes for treatment, storage, or disposal; even practices that would significantly deter this from occurring are prohibited.[4] However, as EPA maintains very limited enforcement authority, the enforcement of this requirement may be difficult.

Subtitle I provides for the regulation of USTs that store regulated substances, where 10% or more of the regulated substance is below ground.[4] In 1992, it was estimated that there were over 2 million UST systems in existence and that as much as 20% of these systems were leaking.[7,8] The main objective of Subtitle I is the prevention of further environmental destruction because of leaking USTs. This is accomplished through the identification of existing tanks, which are then required to meet current standards of construction or be closed and/or removed. New tanks are held to strict standards during construction and installation, which includes government notification of any new installations. Minimum standards are established by federal law, but authority may be delegated to states that adopt laws at least as stringent as the federal standards.

COMPREHENSIVE ENVIRONMENTAL RESPONSE, COMPENSATION, AND LIABILITY ACT

The CERCLA, found in 42 U.S.C. § 9601 et seq., has five basic objectives:[9]

- Provide the enforcement agency the authority to impose cost and liabilities for principal responsible parties (PRPs) of the past and the future release of hazardous chemicals.
- Establish a fund for cleanup and prioritization of sites.
- Authorize EPA to initiate cleanup for emergency and remedial response.
- Establish a postliability fund in conjunction with the RCRA for site closures.
- Advance science in technology in all aspects of hazardous-waste management, treatment, and disposal.

The CERCLA's creation was a response to the growing concern over escalating dangers associated with inappropriately disposed hazardous waste. The most commonly cited incident being the Niagara Falls, New York's ''Love Canal,'' where then-president Jimmy Carter declared the former landfill a federal

emergency on August 7, 1978.[10] The CERCLA's 1980 enactment was the next legal step in an evolving scheme of environmental protection.

Passed on December 11, 1980, the CERCLA amends the SWDA and is commonly referred to as Superfund because of the funding objective designed into its overall principles. This trust fund was generated through the creation of a tax applied directly to the chemical and petroleum industries. In the first 5 yr $1.6 billion was collected to provide for cleanup of hazardous-waste sites, thereby meeting the second overall objective.[9]

The appropriation of the funds kept in the Superfund trust is governed by the Act itself. In addition to payment of EPA's cleanup and enforcement costs, certain monies can be recovered by private parties. This entitlement includes repayment for completion of an EPA-approved cleanup. Reimbursement claims may also be filed if a private party has completed a cleanup for which the facility owners refuse to pay, or if EPA has administratively required a cleanup that is considered capricious or for which they were not liable.[11] It is important to note that these reimbursements are not guaranteed entitlements and very rarely extend to those parties responsible for the negligent disposal practices that caused the hazardous condition. In fact, the vast majority of litigation under CERCLA is brought pursuant to cost recovery actions.[11] One noteworthy exception to the list of eligible reimbursements is the exclusion of federal facilities from having remediation costs financed through Superfund. These site numbers are significant with the majority of them being a direct result of Department of Energy or Department of Defense activities during World War II and stretching into the Cold War era.[11] Both departments have extensive long-term cleanup plans that must seek funding through alternative sources.

As of the end of the 2003 fiscal year, Superfund was officially bankrupt. Current legislation plans on using general fund appropriations (i.e., taxpayers) to revitalize the fund, requiring their contribution to be substantially over the $676 million previously collected. Other sources (called offsetting collections) that will be used include individual states, which pay a small portion of the cleanup, and other federal agencies paying for EPA services. In 2002, monies collected from offsetting collections and PRPs were $40 million and $130 million, respectively, while expenditures were $1.3–1.7 billion.[12] As a result, PRPs will be forced to bear the burden of cleanup. Sites that have no readily identifiable PRP to pay for the cost of remediation will be paid for from the $1.1 billion that the EPA has requested from the 2004 fiscal year general appropriations. In the interim, the EPA has requested a subcommittee to be established by the National Advisory Council for Environmental Policy and Technology to address how Superfund prioritized sites and future monetary issues.[12]

In the same way that RCRA governs new hazardous waste from the cradle to the grave, CERCLA creates a complementary Act that regulates waste generated and disposed of prior to the 1976 RCRA regulations. Hazardous wastes, for the benefit of CERCLA, are defined by other environmental acts (i.e., CWA, TSCA, RCRA, etc.) and can be expanded to include other materials as necessary by the enforcing agency. The sites identified as being the most problematic according to a hazardous ranking system are compiled on the National Priority List (NPL), which is maintained along with the current status of each site by the USEPA. Sites are placed on the NPL as a proposed listing. Once on the "final list," a remedial investigation/feasibility study (RI/FS) must be completed within 6 mo, followed by a proposed cleanup strategy. The USEPA then selects the appropriate treatment strategy and issues a Record of Decision (ROD) setting forth all the details of the proposed remedy. The ROD is then placed into the administrative record for a public response period. Ultimately, the ROD leads to a remedial design (RD) phase followed by a remedial action (RA) phase. Following satisfactory completion of all aspects of the ROD, a site may ultimately become "delisted," at which point it is removed from the NPL. At the end of the 2002 fiscal year, only 265 sites have been removed from the NPL. This leaves 1233 on the NPL, with 62 new proposed sites.[12]

The CERCLA is almost exclusively the domain of federal legislation via USEPA. State involvement is limited in most areas as an assistant in the provision of future maintenance of the site. A state may enter into an agreement with USEPA to take response actions and then seek finances from the Superfund to help defray the costs of the remedial action.[6] In an alternate scheme, where USEPA is the lead regulator, a state must agree with the selected remedy as set forth in the ROD before the RD/RA phases of the project may proceed. From the State's view the largest involvement also becomes a major stumbling block in some Superfund sites, the unwillingness or inability of a state to fund 10% of the required RA cost.[11] This figure includes all future maintenance of the site.

SUPERFUND AMENDMENTS AND AUTHORIZATION ACT

On October 17, 1986, the SARA was created to extend the reach of CERCLA. Found at 42 U.S.C. §11001 et seq., this amendment increased the Superfund trust to $8.5 billion, but more importantly added legislation stressing the importance of community involvement, knowledge, and well-being, while simultaneously increasing the emphasis on permanence in solutions to hazardous-waste sites. As a result, the most well-known

aspect of SARA is Title III, the Emergency Planning and Community Right to Know Act. The emergency planning provision requires a detailed plan (personnel training, evacuation procedures, hospital routes, etc.) in the event of an accident. The community right to know has two subtitles. Subtitle A requires facility owners and operators to provide information to the public. Subtitle B states the specific type of information required. For instance, material safety data sheets (MSDS) must be made available on request, emergency and hazardous chemical forms are submitted annually to the state, etc.

The RCRA's Overlap with CERCLA

As outlined above, RCRA is a preventative legislature regulating current/future hazardous-waste management. Conversely, CERCLA's goal is to rectify problems caused by the previous mismanagement of hazardous waste (i.e., it addresses past sins). There are three general areas in which RCRA and CERCLA statutes overlap: disposal of CERCLA-waste sites; reactions to imminent hazards; and corrective actions (CA) for releases from TSDFs.

Waste disposal from CERCLA sites must be managed in accordance with regulations outlined in subtitle C of RCRA. For instance, generator/regulator statutes would regulate wastes taken off-site for treatment. This would entail the proper labeling of all material as well as hazardous-waste manifests. In addition, any waste treated or stored on-site must be handled in accordance with TSDFs.

An imminent hazard is a clearly identifiable threat. Imminent hazards from either current or previously existing (whether still operational or abandoned) TSDFs are regulated by section 7003 of RCRA as well as section 106 of Superfund. Both statutes require the PRPs to take immediate action for cleanup. As Superfund is a comprehensive authority, it is the primary authority for imminent hazards. Duplication in RCRA is used to bolster government action for maintaining the statutes and enforcement of penalties.

Previously, unless a release from a TSDF posed an imminent threat, a ''remedy'' was not required. As of 1984, CA are required for all releases from TSDFs. Prior to this date, CERCLA was the only legislative body that provided the EPA authority for requiring CAs. Treatment, storage, and disposal facilities fall under RCRA subtitle C, which only mandated a remedial-site investigation. Often, action ceased after the site investigation because regulation at these sites was not really enforceable. Providing an overlap between the two legislative bodies required TSDFs to move from simply completing a remedial-site investigation to actually implementing CA.

CLEAN AIR REGULATIONS

Overview of Air Regulations

The legal definition of air pollution is ''the presence in the outdoor atmosphere of any one or more substances in quantities, which may or are harmful/injurious to human health or welfare, animal or plant life, or property, or interfere with the enjoyment of life, property or outdoor activity.''[13]

Based on this definition, the first documented case of air pollution occurred in 1272 when thousands of people died as a result of volcanic ash. During the early to mid-1900s, several air pollution episodes resulted in either unacceptable illness levels or death. For instance, in 1948 the mill town of Donora, PA, was subjected to a 4-day fog from the uncontrolled steel mill emissions reacting with other constituents in the atmosphere. The fog left 14,000 people (half the town) sick and 20 dead.

Events such as the Donora fog led to the establishment of air pollution regulations. Fig. 1 contains a timeline of the key federal air pollution regulations. As shown in Fig. 1, the clean air legislation, CAA, and Motor Vehicle Act were the first three key legislative documents for reducing the release of harmful compounds into the atmosphere. However, it was not until the Air Quality Act of 1967 that actual air quality standards were formulated. The standards were constructed in an effort to reduce emissions in the newly formed air quality control regions (AQCRs), i.e., the most heavily contaminated areas. Each state in an AQCR had to develop a State Implementation Plan (SIP) outlining how, and by when, they would achieve the necessary reductions. To generate the air quality standards and realistic SIPs, the contaminants were identified as either a primary or a secondary pollutant. A primary pollutant is defined as a compound that is directly emitted to the atmosphere as a contaminant (i.e., CO, SO_2, etc.). Conversely, a secondary pollutant is a contaminant that is formed in the lower atmosphere by a chemical reaction with other primary pollutants.

Clean Air Act Amendments 1970–1990

The establishment of ACQRs and SIPs started to move air pollution control in the right direction. However, the EPA still lacked the ''teeth'' needed to enforce the regulations. This was partly owing to the fact that when generating the SIPs, the states set their own air quality standards, therefore no one was using the same constraints. One of the key provisions of the CAA Amendments (CAAA) in 1970 was to establish National Ambient Air Quality Standards (NAAQS). Anything above the NAAQS was considered a pollutant.

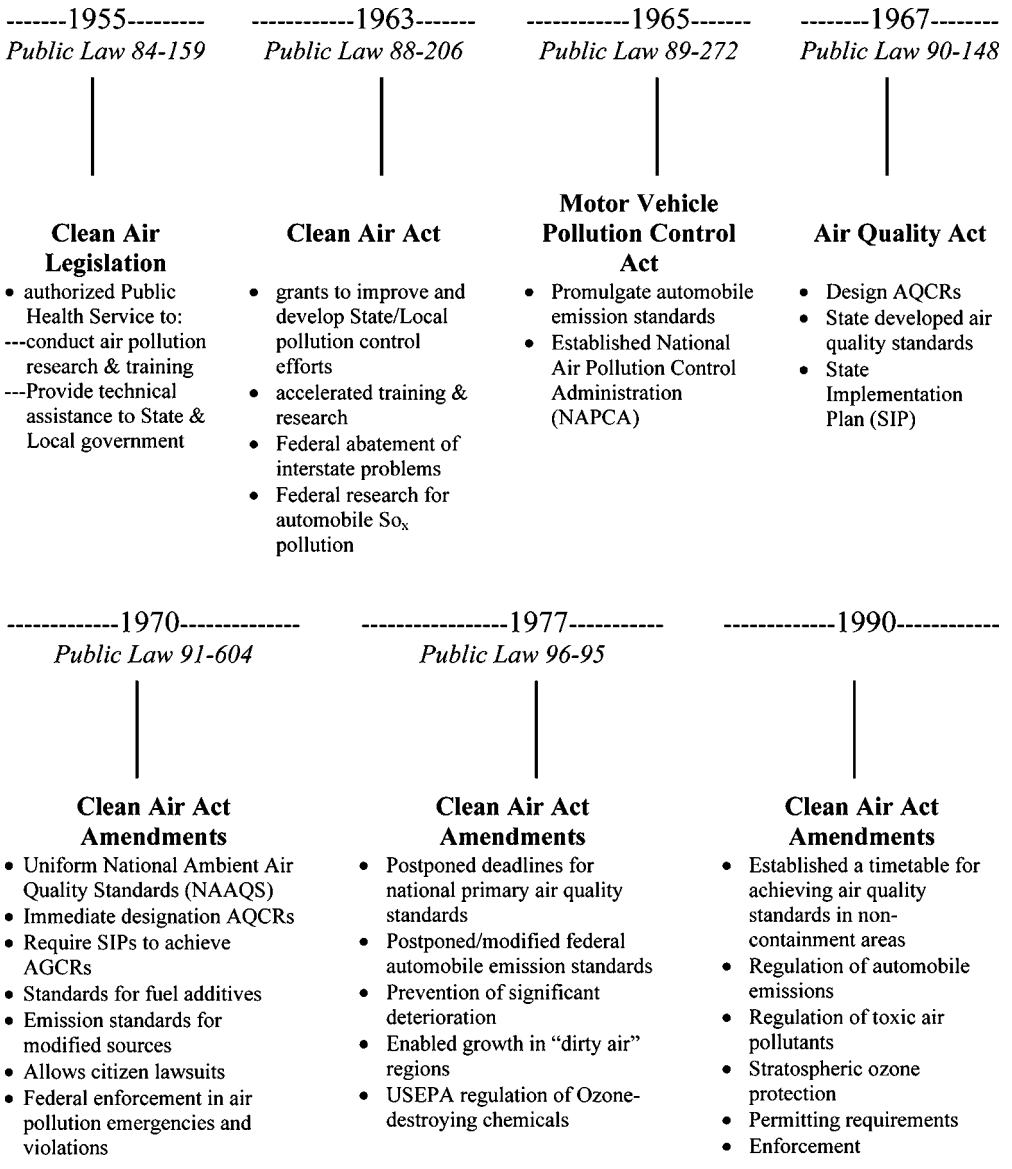


Fig. 1 Timeline of key federal air pollution regulations.

After the new standards were set, states had to resubmit SIPs. This time, the SIP was required to establish the approach for attaining the NAAQS within 3 yr. The NAAQS were more aggressive than previous legislation. In addition, uniform standards also stifled interstate competitions. Table 2 lists the NAAQS of key contaminants for the most commonly used sampling intervals. Refer to 42 USC § 7408(a)(1) for other sampling intervals. To provide more uniform enforcement of the new regulations, the National Air Pollution Control Administration was eliminated and all of the responsibilities transferred to the EPA.

The CAAA of 1977 realized that the aggressive approach of CAAA 1970 could not be achieved within an economically feasible manner. Therefore, deadlines were postponed. The key initiative of the amendment was the regulation of ozone destroying chemicals. This was the start of the phasing out of chlorofluorohydrocarbons. In 1990, amendments were passed that regulated at the federal level the control of minor point source emissions as well as nonpoint source emissions. Previous regulations only focused on boilers, incinerators, and flares. Major changes, such as regulation of hazardous and toxic air pollutants, and stratospheric ozone protection, were made to address global problems. Incorporation of such provisions was a major political achievement.

Emission Trading

Emission trading is based on the concept that the best way to improve overall air quality is to create a genuine economic incentive for industrial facilities to decrease emissions. To move away from the previous

Table 2 National ambient air quality standards

Pollutant	Primary standard ($\mu g/m^3$)	Secondary standard ($\mu g/m^3$)	Sampling time	Health and environmental effects when levels are exceeded
Carbon monoxide	10,000 (9 ppm)	NA	8 hr	Headaches, diminished alertness, heart damage, death, and smog
Nitrogen dioxide	100 (0.05 ppmv)	100 (0.05 ppmv)	1 yr	Lung damage, acid rain, and formation of ozone and secondary pollutants
Sulfur dioxide	365 (0.14 ppmv)	1300 (0.5 ppmv)	24 hr	Eye irritant, lung damage, kills aquatic life, acid rain, and formation of secondary pollutants
Nonmethane hydrocarbons	160 (0.24 ppmv)	160 (0.24 ppmv)	8 hr	Cancer, respiratory disorders, tissue damage, and dead vegetation
Ozone	157 (0.08 ppmv)	157 (0.08 ppmv)	8 hr	Eye irritant, lung damage, and respiratory problems
PM10	50 150	50 150	1 yr 24 hr	Eye irritant, lung damage, smog, crop damage, and discolored buildings
PM2.5	15 65	15 65	1 yr 24 hr	Eye irritant, lung damage, smog, crop damage, and discolored buildings
Lead (Pb)	1.5	1.5	3 mo	Brain damage, kidney damage, smog, and crop contamination

NA, not applicable, only primary standard.

command–control-type approach, the EPA developed several different emission-trading concepts. The three most commonly used approaches are the open market trading, offset trading, and allowance trading.

Open market trading encourages sources to voluntarily reduce emissions below the required standards. Other companies that are not in compliance can purchase the reductions to comply with regulatory limits and/or increase operational flexibility. Credits created under open market trading must meet strict criteria to ensure that the process yields continuous environmental improvement as well as meets the intent of air regulations.

Offset trading was developed under the New Source Review program to allow new or significantly modified sources of air polluters to be built in nonattainment areas without adversely affecting air quality. Previously, new sources could not be built in nonattainment areas regardless of the air pollution control devices used. With offset trading, contemporous emission rates are transferred between sources in the same area. New sources with the new advanced technology can achieve a greater reduction in emission rates than that required in the nonattainment area, thereby providing a permanent reduction in emissions.

Allowance trading is an annual, area-wide emission cap (i.e., budget) established over a defined area of air pollution sources. There is a specified time-dependent reduction schedule with allowances allocated to each party based on their historical emission rate. Each participating source has the option of meeting the annual compliance through a combination of emission control, use of assigned allowances, or purchase of allowances sold by other sources within the area. The purchase of another company's allowance makes allowance trading very similar to offset trading. The most well-known allowance trading has been with SO_2 for the reduction of acid rain.[14]

Success with the EPA's SO_2 trading has led to the proposal of global emission trading of CO_2. The key concepts for implementation are the level of emission per country and issuing of permits to each country and the associated industries. Although global trading is a good theory, too many obstacles remain. For instance, how will the quotas per country be determined? How will monitoring and emissions be enforced?

CLEAN WATER ACT AND OTHER WATER REGULATIONS

Prior to the CWA, the Treasury Department was the overseer for the United States Public Health Service (USPHS). From 1914 to 1962 the USPHS established comprehensive reviews of drinking water concerns and

established standards for over 28 compounds within each state. This was followed by the Water Quality Act of 1965, which required states to improve water quality, designate the intended use for interstate waterways, and adopt more uniform water quality standards.

In 1972, the Water Pollution Control Act, often referred to as the CWA, was passed. The primary provisions of the CWA are contained in 33 USC § 1251 et seq. The overall objective is to restore and maintain the chemical, physical, and biological integrity of the Nation's waters. Seven national goals were established to achieve this aim:

- Eliminate discharge of pollutants into navigable waters by 1985.
- Set interim goals of water quality to protect fish and wildlife and provide recreation by July 1, 1983.
- Prohibit discharge of toxic pollutants in concentrations, which may adversely affect the environment.
- Construct publicly owned treatment works (POTW) facilities with $5 billion/yr federal assistance.
- Establish waste treatment management plans with each state.
- Establish technology to eliminate discharge of pollutants.
- Develop and implement programs to control nonpoint sources.

Prior to 1972, water quality regulations were limited to point sources. A point source is a single source, such as a commercial facility or municipal wastewater treatment, that discharges effluent directly into a waterway. With the CWA, water quality was expanded to encompass nonpoint sources. Nonpoint sources often associated with waste from runoff (agricultural, mine tailings, etc.) are those contaminant sources whose origin cannot be identified.

National Pollutant Discharge Elimination System and Pretreatment Requirements

The National Pollutant Discharge Elimination System (NPDES) permits were based on effluent limitations in a staggered fashion to achieve the goals set forth in the 1983 CWA amendments. The permits are issued to individual wastewater treatment plants to reduce the number of pollutants released into nearby waterways. The permits contain standard conditions applicable to all POTWs, site-specific constraints, compliance monitoring, and reporting requirements.

The NPDES is a voluntary process that covers priority pollutants, conventional wastewater (BOD, TSS, fecal coliforms, etc.), and nonconventional pollutants (COD, ammonia, priority toxics, etc.). Once established, the monitoring plans revealed that several nonconventional constituents in industrial effluents either interfered with municipal treatment or passed through unaltered. Interference with treating conventional wastes could result in a loss of compliance of the regulated, conventional pollutants. Compounds passing through unaltered violate the intent of clean water statutes. This led to the need for industrial pretreatment to ensure that POTWs could handle the incoming waste streams. Industries requiring pretreatment prior to effluents being released for municipal treatment include pulp and paper, metal plating, electroplating, foundries, pharmaceuticals, petroleum refining, tanneries, etc. Pretreatment standards for each industry must be met before the facility can discharge to a POTW.

SAFE DRINKING WATER ACT

The SDWA (public law 93-523,) was passed on December 16, 1974, to expand the nation's goals to include drinking water. Within 6 mo of legislation, National Interim Primary Drinking Water Regulations (NIPDWRs) were enacted to ensure that serious efforts were being made. Every 2.5 yr the USEPA would review the current water quality and publish recommended maximum contaminant levels (MCLs) for potentially threatening chemicals, thereby generating new NIPDWRs. The continual review of NIPDWRs led to the establishment of national primary drinking water standards (Table 3).

In 1986, the most significant amendments were made to the SDWA. Some of the key provisions of this amendment include mandatory standards for 83 contaminants by June 1989, mandatory regulation of 25 contaminants on a triennial monitoring schedule, disinfection of all public community water supplies, and monitoring of contaminants that are not regulated. In 1996, the SWDA was amended to include National Secondary Drinking Water Standards (NSDWRs). The secondary standards are nonenforceable MCLs that deal with the aesthetics of the water (taste, color, odor, etc.), but whose exceedance would not cause a public health risk. While unenforceable, some states have adopted secondary MCLs that are higher or lower than those established in the NSDWRs as standards that are enforceable. Finally, there are MCL goals (MCLGs) that establish a nonenforceable, health-based goal at which no known or anticipated adverse effects on human health occur and that allow for an adequate margin of safety.[15] Maximum contaminant level goals are often used as a guide in establishing actual MCLs. As a result of the SDWA and its amendments, more than 240,000 publicly owned water systems have been established that serve over 300 million people.

Table 3 Primary and secondary drinking water standards for select contaminants

Primary standards		Secondary standards	
Inorganic	(mg/L)	*Inorganic*	(mg/L)
Antimony	0.006	Aluminum	0.05–0.2
Arsenic	0.05	Chloride	250
Barium	2.0	Copper	1.0
Berrylium	0.004	Fluoride	2.0
Cadmium	0.005	Iron	0.3
Chromium, total	0.10	Manganese	0.05
Copper	1.3	Silver	0.1
Cyanide	0.20	Zinc	5
Lead	0.015		
Mercury	0.002	*Others*	
Nickel	0.1	Color	15 units
Nitrate-N	10	Corrosivity	Noncorrosive
Nitrite-N	1.0	pH	6.5–8.5
Selenium	0.05	Foaming agents	0.5
Thallium	0.002	Odor	15 color units
Select VOCs		Sulfate	250 mg/L
Benzene	0.005	Total dissolved solids	500 mg/L
Carbon tetrachloride	0.005		
Dichloromethane	0.005		
Ethylbenzene	0.7		
Styrene	0.1		
Tetrachloroethylene	0.005		
Vinyl chloride	0.002		
Xylenes, total	10.0		
Select synthetics			
Alachlor	0.002		
Atrazine	0.003		
Benzo(a)pyrene	0.0002		
Chlordane	0.002		
Diethylhexyl phthalate	0.006		
Dinoseb	0.007		
Endrin	0.002		
Heptachlor	0.0004		
Lindane	0.0002		
Methoxychlor	0.04		
PCBs	0.0005		

TOXIC SUBSTANCE CONTROL ACT

The Toxic Substance Control Act (TSCA) was passed in 1976 and is contained in 15 USC § 2601 et seq. It regulates toxic substances, other than wastes, that are not adequately covered by other statutes. Promulgation of TSCA gave the EPA the authority to manage chemicals from production to final disposal. In fact, if a compound (e.g., carbon tetrachloride) is found to pose unreasonable risk to human health or the environment, the EPA can immediately remove it from commerce.[3] The TSCA's requiring a premanufacture notification (PMN) 90 days prior to manufacture has minimized the occurrence of immediate emergency removal of toxic chemicals. The PMN contains information on the physical/chemical properties, health and environmental effects, quantities, intended uses, and potential exposures. If the chemical cannot be processed, used, or disposed of without posing unreasonable risk, the EPA can ban its production. In situations where the PMN does not contain sufficient information to ascertain potential risks, additional testing can be ordered. The testing could be required from the manufacturer, processor, and/or distributor of the chemical in question.

In addition to the PMN, section 4 of TSCA requires the formation of a review committee. Each year the committee selects 50 compounds that are currently produced and/or processed to be retested. To ensure that the intent and guidelines of TSCA are being met, each industry must keep meticulous records on the quantities of chemicals imported, manufactured,

and used. Every 4 yr the list is submitted to TSCA for review. The quadrennial review led to four of the key provisions of TSCA: establishment of asbestos emergency response, ban of lead-base paints, restriction of hexavalent chromium in wastewater, and prohibition of the production of polychlorinated biphenyls (PCBs) past 1979.[6]

CONCLUSIONS

Environmental law is an ever-changing entity evolving because of a complex interaction between the protection of human health and the environment and the demands of an industrialized nation. It would be difficult, at best, to find an individual or company that would admit to a blatant disregard for environmental protection; however, the economic constraints that drive the day-to-day operations of a business will often govern a short-term decision. Corporations are, by and large, in business to make money. The easiest way to make money may be in direct conflict with what is in the best interest of the environment. An example of this was in the September 8, 2003, issue of C&E News, which reported the sale of PCB contaminated properties without requiring owners to clean the soil to threshold constraints. The institution of laws and regulations, like those found in this review, are the result of public outcry over decisions historically made with the protection of the environment as nothing more than an afterthought. The careful enactment of environmental laws places the burden of protection on those who have been historically responsible for the destruction of environmental resources, while still allowing for the careful growth of industry that promulgates a healthy economy.

For more in-depth coverage of any law highlighted during this review or not specifically noted, the reader is encouraged to seek the guidance of one of the many books devoted specifically to the law of interest. The texts used as reference for this review are all good sources of additional material. Additionally, the actual CFR, if applicable, may be retrieved with little difficulty. Volumes of material on each law can give the researcher a very detailed view in as broad or as limited a scope as desired and should be used whenever a more complete picture of a given law is necessary.

NOMENCLATURE

AQCRs	Air quality control regions
BOD	Biological oxygen demand
CA	Corrective action
CAA	Clean Air Act
CAAA	Clean Air Act Amendments
CERCLA	Comprehensive Environmental Response, Compensation, & Liability Act
CEQ	Council on Environmental Quality
CFR	Code of Federal Regulations
COD	Chemical oxygen demand
CWA	Clean Water Act
EIS	Environmental Impact Statement
EPA	Environmental Protection Agency
FIFRA	Federal Insecticide, Fungicide, Rodenticide Act
FS	Feasibility study
HSWA	Hazardous and Solid-Waste Amendments
MCL	Maximum contaminant level
MCLG	Maximum contaminant level goal
MSDS	Material safety and data sheet
NAAQs	National Ambient Air Quality Standards
NAPCA	National Air Pollution Control Administration
NEPA	National Environmental Policy Act
NIPDWRs	National Interim Primary Drinking Water Regulations
NPDES	National Pollutant Discharge Elimination System
NPL	National Priority List
NSDWR	National Secondary Drinking Water Regulations
PCB	Polychlorinated Biphenyls
PMN	Premanufacture notification
POTW	Publicly owned treatment works
PRP	Principal responsible party
RA	Remedial action
RCRA	Resource, Conservation, and Recovery Act
RD	Remedial design
RI	Remedial investigation
ROD	Record of decision
SARA	Superfund Amendments and Reauthorization Act
SDWA	Safe Drinking Water Act
SWDA	Solid-Waste Disposal Act
SIP	State Implementation Plan
TSCA	Toxic Substance Control Act
TSDF	Treatment, storage, and disposal facility
TSS	Total suspended solids
USPHS	United States Public Health Service
UST	Underground storage tank

REFERENCES

1. Sullivan, T.F.P.; Steinway, D., Ed. Fundamentals of environmental law. In *Environmental Law*

Handbook, 16th Ed.; Government Institutes, ABS Group: Rockville, MD, 2001; 1–53.
2. Salzman, J.; Thompson, B.H., Jr. *Environmental Law and Policy*; Foundation Press: New York, 2003.
3. Jain, R.K.; Urban, L.V.; Stacey, G.S.; Balbach, H.E. *Environmental Assessment*; McGraw-Hill: New York, 1993.
4. Case, D.R. Resource Conservation and Recovery Act. In *Environmental Law Handbook*, 16th Ed.; Government Institutes, ABS Group: Rockville, MD, 2001; 109–156.
5. La Grega, M.D.; Buckingham, P.L.; Evans, J.C. *Hazardous Waste Management*; McGraw-Hill: New York, 2001.
6. Jain, R.K. Environmental legislation and regulations. In *Standard Handbook of Environmental Engineering*; Corbitt, R.A., Ed.; McGraw-Hill: New York, 1990; 2.1–2.32.
7. Cookson, J.R., Jr. *Bioremediation Engineering: Design & Application*; McGraw-Hill: New York, 1995.
8. Dean, N.; Kremer, F. Advancing research for bioremediation. In *Environmental Protection*; Elseveir; Sep 19–25, 1992.
9. USEPA. *CERCLA Overview*; http://www.epa.gov/superfund/action/law/cercla.htm (accessed Nov 18, 2003).
10. State University of New York (SUNY) at Buffalo. Background on the Love Canal; http://ublib.buffalo.edu/libraries/projects/lovecanal/backgorund_lovecanal.html (accessed Nov 18, 2003).
11. Cardwell, R.E. Comprehensive Environmental Response, Compensation, and Liability Act. In *Environmental Law Handbook*, 16th Ed.; Government Institutes, ABS Group: Rockville, MD, 2001; 417–482.
12. GAO. *Superfund Program: Current Status and Future Fiscal Challenges*, Report to U.S. Senate, GAO-03-850, Jul 2003.
13. Wark, K.; Warner, C.F. *Air Pollution: Its Origin & Control*, 2nd Ed.; Harper Collins: New York, 1981; 12–23.
14. EPA. Acid Rain Program; www.epa.gov/docs/acidrain/overview.html.
15. Kucera, D.J. Safe Drinking Water Act. In *Environmental Law Handbook*, 16th Ed.; Government Institutes, ABS Group: Rockville, MD, 2001; 371–416.

Epoxy Resins

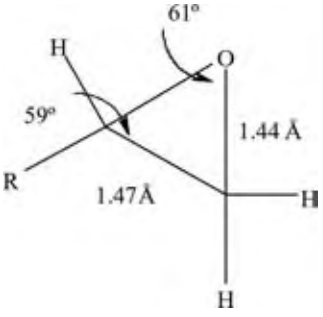
Ian Hamerton
University of Surrey, Surrey, U.K.

INTRODUCTION

Epoxy resins are a technologically important family of cross-linked, thermosetting polymers, which offer a good balance of thermal, mechanical, and electrical properties. The cured polymers are characterized by their good adhesion to a variety of substrates, hardness, chemical inertness (in the cured state), and thermal resistance. In this entry, the structure of the epoxy group, routes to commercial manufacture, physical properties of monomers and oligomers, and a variety of methods used in the characterization of these materials are addressed. Prior to cure, the monomers are typically formulated, and the range of modifiers available is discussed along with an indication of the property enhancements that may be achieved in this way. Polymerization may occur either in the absence or in the presence of curing agents, and the effect of the nature of these compounds on the cured properties is examined with particular attention to fracture toughness and flame retardancy. Finally, the technological importance of the materials is underlined with a discussion of the principal applications to which they are routinely put.

EPOXY RESINS

Epoxy resins have been in existence for over a century, with reports of the first synthetic work being carried out in 1891, although the first commercial products appeared only during the 1940s. These early materials, offering low-pressure molding resins, were based on the phthalic anhydride-cured diglycidyl ether of bisphenol A (DGEBA), which remains the mainstay of the epoxy industry. Epoxy resins are formed from monomers containing at least two epoxy (oxirane) groups.

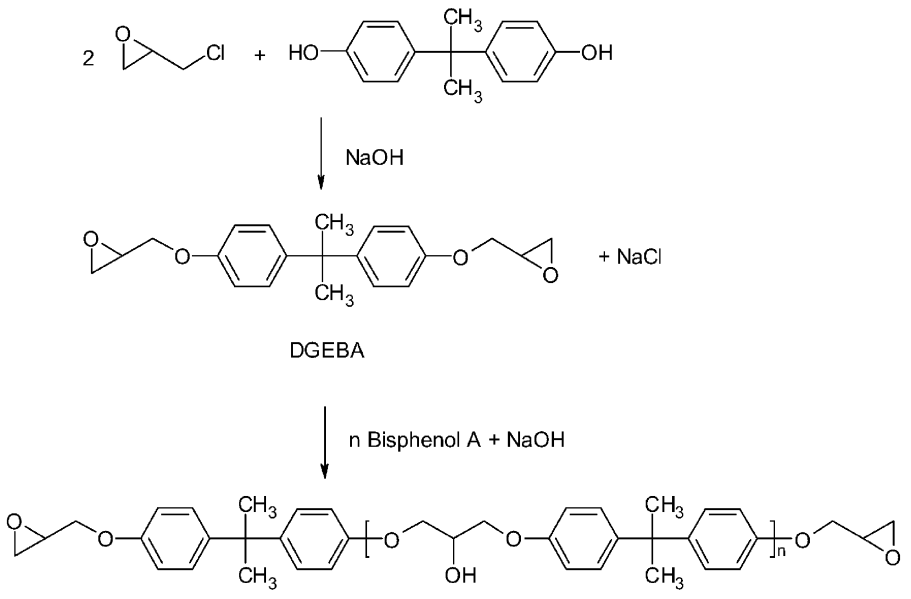


The three-membered ring is planar and may be bonded to a variety of substituents, R = $-CH_2OPh$, O-alkyl, CH_2-N-aryl, etc. The structure confers considerable ring strain, as the internal bond angles of the ring deviate substantially from the ''ideal'' angle of 109° for a tetrahedral sp^3-hybridized carbon. This makes the geometry similar to that of cyclopropane, but the marked electronegativity of the oxygen accounts for the inequality in the internal ring bond angles and lengths. The second, arguably more important, consequence of this geometry is the effect on the chemical reactivity of the functional group, with particular ramifications for the cure chemistry. In contrast with ''conventional'' noncyclic ethers (i.e., R-O-R′, Ph-O-Ph, or Ph-O-R), in which the ether bridge resists attack by alkalis, ammonia, or amines, the epoxy ring is susceptible to both nucleophilic attack (on one of the ring carbons) and electrophilic attack (on the ring oxygen), and mixtures of epoxy monomers and aliphatic amines will follow a nucleophilic mechanism at room temperature.

Commercial Manufacture of Epoxy Resins

The first commercial epoxy resins were structurally and chemically very similar and followed patent applications by Pierre Castan for denture bases (the technology was later licensed to CIBA, Switzerland)[1] and Sylvan Greenlee for damage tolerant surface coatings (Devoe and Raynolds, U.S.A.).[2] Whilst a number of synthetic routes to manufacture epoxy monomers exist,[3] two commercial paths prevail, depending on the desired structure (e.g., glycidyl ether or ''ene oxide'' monomer) and proposed application. Most commercial epoxies are produced via the dehydrohalogenation of halohydrins (in practice, this is usually epichlorohydrin) in the presence of a stoichiometric amount of a strong base catalyst (e.g., NaOH) to form a glycidyl ether (shown for the most common precursor, bisphenol A, in Scheme 1. The stoichiometry and reaction time can be controlled very carefully to produce the monomer shown, DGEBA ($n = 0$), a particular grade of epoxy, which can be used in electrical applications where high purity is necessary. More commonly, excess bisphenol A is added and the reaction is allowed to proceed to yield oligomeric polyethers, of varying chain length with $0.2 \leq n \leq 12$, containing hydroxyl groups along the backbone. Epoxies of this

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007755



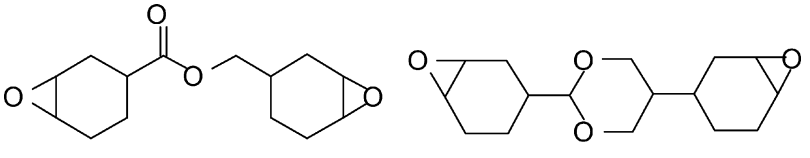
Scheme 1 Commercial preparation of the diglycidyl ether of bisphenol A, DGEBA, and higher oligomers.

general type (and those formed from glycidyl amines) feature heavily like the resin matrix in advanced composite applications, adhesives, and coatings. However, examination of the reaction scheme reveals that the use of stoichiometric quantities of alkali produces significant quantities of halide ions, particularly chloride, associated with this route (arising from incomplete epichlorohydrin utilization), even after rigorous washing. Commercial oligomers may contain anything up to 1% chloride. In many electrical applications, the presence of even low (parts per million) quantities of chloride is simply not acceptable as it alters the electrical conductivity of the cured resin (normally a good insulator) and accelerates component corrosion in use. Furthermore, the presence of chlorine may also impart a colored taint to the resin and also reduce its reactivity (active chlorine blocks the reaction of less basic catalysts such as commonly used tertiary amines).

In some microelectronic or higher performance structural applications (e.g., satellites in which reduced levels of UV absorption are desirable, to withstand the harsh space environment), an alternate preparative procedure is employed involving epoxidation via Prileschaiev's reaction, e.g., using peracetic acid.[4] Two commercial structures depicting the resulting diepoxides are shown in Fig. 1 and the lack of strong UV-absorbing groups (chromophores) is immediately apparent. The other important characteristic, particularly in microelectronic applications, is the lack of hydrolyzable chloride ions and inorganic salts (ash). The latter is a direct result of the synthetic route employed, as the starting materials (typically buta-1,3-diene or acrolein, which acts as a dienophile) undergo a Diels–Alder reaction to yield adducts, containing cyclohexene moieties.

Physical Properties of Epoxy Resins

The monomeric form of DGEBA (shown in Scheme 1, where $n = 0$) may exist as a waxy crystalline solid (m.p. 43°C) even in some mixtures with other species of higher molecular weight, but it is unusual. The increasing chain length of the oligomer tends to yield viscous amber liquids at lower molecular weights or brittle solids as n increases (see Table 1 for representative melting temperatures). Epoxy resins are most likely to form glassy, amorphous solids, although liquid crystalline thermoset epoxies, containing highly rigid structural

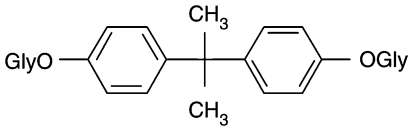
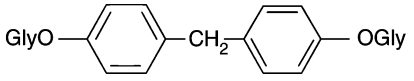
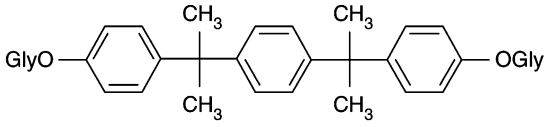
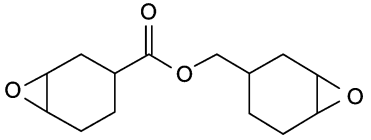


Union Carbide ERL-4221
Ciba-Geigy CY-179

Union Carbide ERL-4234
Ciba-Geigy CY-175

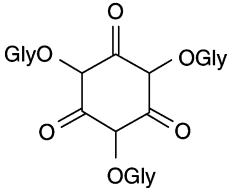
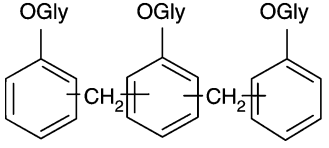
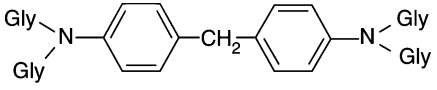
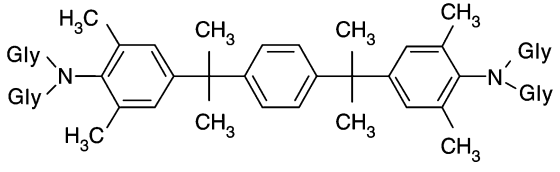
Fig. 1 Selected commercial cycloaliphatic epoxy monomers.

Table 1 Structure and physical properties and technological applications of selected epoxy monomers

Epoxy monomer	EEW (g/epoxide)	Viscosity (250°C) cP	Melting point (°C)	T_g (°C)	Typical applications
	Liquids				
	95–450	5,000–30,000	—	110–150	
	185–200	2,000–7,000	—		
	Solids				
	500–6,000	—	60–180		Coatings, especially powder coatings, tube coating primers, stoving enamels, castings, prepregs
	170 ($n = 0$)	—	43		
	908 ($n = 2$)	—	70		
	908 ($n = 2$)	—	70		
	1,590 ($n = 3$)	—	95–100		
	158–175	5,000–7,000	—	—	Additive for bisphenol A resins in adhesives, casting, tooling, composites, coil coatings, marine composites, potting, coating, and flooring
	240–255	1,195 (30°C)	—	—	
	131–143	250–450	—	≤200	General purpose casting resin, filament winding, acid scavenger, plasticizer

(Continued)

Table 1 Structure and physical properties and technological applications of selected epoxy monomers *(Continued)*

Epoxy monomer	EEW (g/epoxide)	Viscosity (250°C) cP	Melting point (°C)	T_g (°C)	Typical applications
	101–111	—	85–110	160–180	Powder coatings, casting compounds
	Liquids			198	
	170–190	30,000–90,000			Coatings, encapsulators, laminating, molding, adhesives
	Solids				
	190–220		90–100		Molding powder
	275–330		90–95		Adhesives, coatings
	110	3,000–6,000	—	262	Aerospace composites
	117–134	8,000–18,000 (50°C)			
	185–205	—	65	258	Aerospace composites, adhesives

where Gly =  , EEW = equivalent epoxide weight, T_g = glass transition temperature (dry), n = degree of polymerization (see Scheme 1).

Data adapted from Ellis, B., Ed.; *Chemistry and Technology of Epoxy Resins*; Blackie Academic and Professional: Glasgow, 1993; Hamerton, I. *Recent Developments in Epoxy Resins*; RAPRA Review Report Number 91, 1996; *8* (7); Hamerton, I.; Hay, J.N. Structural polymers, Chapter 8. In *Specialty Polymers*, 2nd Ed.; Dyson, R.W. Ed., Blackie Academic and Professional: London, 1998, 200–250; Pilato, L.A.; Michno, M.J. *Advanced Composite Materials*; Springer-Verlag: Berlin, 1994.

motifs are known to form nematic phases.[5,6] Polymerization of these monomers, e.g., diglycidyl ether of 4,4′-dihydroxy-α-methylstilbene, using 4,4-*bis*(aminophenyl) methane (DDM), as a curing agent preserves the anisotropic order in the thermoset network, leading to high cross-linked organization of mesogens, which is claimed to have improvements in modulus, reductions in the coefficient of thermal expansion, and increased fracture toughness.

Characterization of Epoxy Monomers and Polymers

Many analytical techniques may be applied to the characterization of uncured epoxy resins, from simple chemical analysis to more sophisticated forms of spectroscopy. In the uncured form, the solubility of monomeric and oligomeric species renders them amenable to analysis. Common analytical measurements made on uncured epoxy monomers include determination of the epoxide content (achieved by the production of halohydrin from the cleavage of the epoxy rings using haloacids), hydroxyl content (using near infrared spectroscopy or assay with lithium aluminum hydride to yield hydrogen determined volumetrically or using gas liquid chromatography), and chlorine content. The latter may take the form of organically bound chlorine, either hydrolyzable (resulting from incomplete epichlorohydrin) or inactive (resulting from the reaction of epichlorohydrin with a secondary alcohol or phenol). The total chlorine content is commonly determined by oxidizing the material using a Parr bomb and quantifying the chloride content as silver chloride. Hydrolyzable (saponifiable) chlorine may be determined by treatment with excess caustic soda, followed by back titration with standard hydrochloric acid.

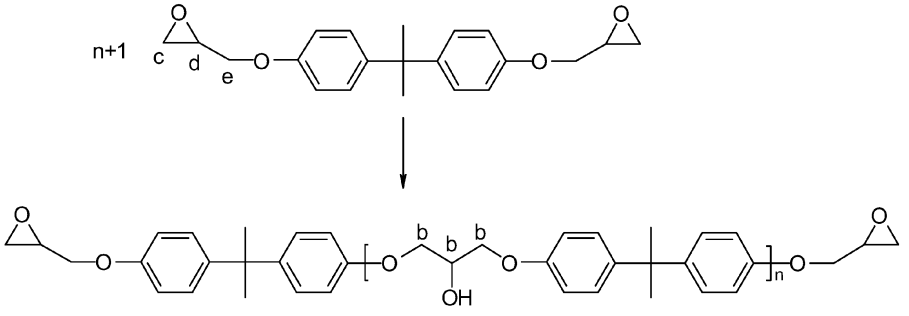
Conventional (spectroscopic) analysis techniques are also readily applied to determine structural details (e.g., the epoxy ring and the repeat unit) and can also be applied to determine the degree of polymerization, particularly where the poor solubility of the oligomer may hamper solution-based techniques. Mertzel and Kœnig[7] presented a thorough spectral characterization of EPON 828 (a commercial grade of DGEBA).

Thus, infrared spectroscopy reveals that the characteristic epoxy ring breathing vibrations are located at ca. 863 and 915 cm^{-1}, and these are useful for kinetic analysis and quantification, with an overtone band evident at 4535 cm^{-1}, whilst ^{1}H NMR spectroscopy reveals that the methylene and methyne ring protons of the oxirane ring resonate at ca. 2.67, 2.87, and 3.34 ppm. ^{13}C NMR is not only useful for yielding structural information (e.g., the glycidyl methylene at ca. 65 ppm, the methyne ring proton at ca. 50 ppm, and the methylene ring proton at ca. 45 ppm), but also epoxy equivalent weight (EEW). The latter is achieved by calculating the ratio between the integrals of the chemical shifts for the terminal ether carbons (C_c, C_d, and C_e; Scheme 2 and the bridge carbons (C_b).[8] The agreement between ^{13}C NMR and chemical analysis in this study was quite reasonable, although where discrepancies arose (e.g., 171 g/epoxide cf 193 g/epoxide or 971 g/epoxy cf 1086 g/epoxy), the spectroscopic analysis consistently gave a lower value of EEW.

Formulation of Epoxy Resins

Epoxy resins are amenable to modification in many ways and commercial epoxy formulations often contain a number of materials whose presence serves to modify the properties and characteristics of both uncured and cured epoxy. These additives fall into the following general categories:

i. Reactive and nonreactive diluents (to reduce viscosity and aid general processability, such as wetting characteristics and allow the incorporation of other fillers). Reactive diluents are often monoepoxides.
ii. Fillers (see Table 2 for a selected list).
iii. Plasticizers may be long-chain, nonreactive molecules, e.g., poly(vinyl chloride) or 2-hydroxyethylmethacrylate, entrapped within the epoxy network, whilst ''flexibilizers'' react with the epoxy group during cure. Both are employed to improve properties such as strain tolerance, impact strength, low temperature crack resistance, and adhesive properties such as lap shear,



Scheme 2 Polymerization of the DGEBA monomer to the corresponding epoxy polymer, showing the terminal ether carbons (C_c, C_d, and C_e) and the bridge carbons (C_b).

Table 2 Selected fillers and potential property modifications

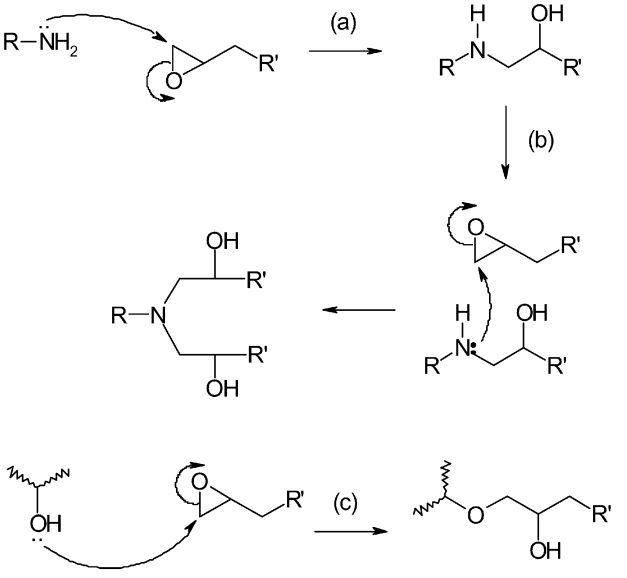
Filler	Resin property modification
Alumina	Abrasion resistance, dimensional stability, electrical resistivity, thermal conductivity, toughness
Aluminum	Impact resistance, machinability, mechanical properties, thermal conductivity
Aluminum silicate	Chemical resistance, dimensional stability, extender, pigmentation
Aluminum trioxide	Flame retardancy
Beryllium oxide	Thermal conductivity
Calcium carbonate	Dimensional stability, extender, machinability, mechanical properties, pigmentation
Calcium sulfate	Dimensional stability, extender
Carbon black	Electrical conductivity, pigmentation, reinforcement, thermal conductivity, thermal resistance
Copper	Electrical conductivity, thermal conductivity, thermal resistance
Glass microballoons	Density reduction
Mica	Chemical resistance, dielectric properties, electrical conductivity, lubricity, moisture resistance, toughness
Sand	Abrasion, thermal conductivity
Silver	Electrical conductivity, thermal conductivity
Titanium dioxide	Dielectric properties, extender, pigmentation
Zirconium silicate	Arc resistance

(From Shaw, S.J. Additives and modifiers for epoxy resins. In *Chemistry and Technology of Epoxy Resins*; Ellis, B., Ed.; Blackie Academic and Professional: Glasgow, 1993.)

and particularly peel strength, but they may also reduce chemical and solvent resistance.

Curing Agents for Epoxy Resins

Epoxy resins undergo cure to yield three-dimensional networks from a liquid or thermoplastic monomer or oligomer and the resulting product is a hard, insoluble, and infusible product. Epoxy monomers with higher functionality (e.g., trifunctional monomers, such as triglycidyl isocyanate or the tetrafunctional *N*,*N*,*N*′,*N*′-tetraglycidyldiaminomethane, TGDDM) are often more reactive than difunctional analogs. During the cure, the epoxy and curative (e.g., curing agent, accelerator, and a variety of additives, discussed later) undergo conversion to a solid via gelation and perhaps vitrification (which occurs as the glass transition temperature, T_g, approaches the cure temperature). The cure reaction proceeds via two general routes: homopolymerization to yield a polyether structure following initiation by a catalytic curing agent; copolymerization or polyaddition involving a multifunctional curing agent to produce a more highly cross-linked network (shown in Scheme 3 for polyamines). These



Scheme 3 General mechanism for amine cure of epoxy resins involving (a) primary and (b) secondary amines; (c) hydroxyl catalyzed polyetherification reaction. This is shown for one functional group; R and R′ denote the other half of an aryl or alkyl backbone.

step-growth mechanisms do not generate volatile condensates and consequently the production of void-free moldings or artifacts can be achieved relatively easily without the need for applied pressure. The characteristic structure and consequent reactivity of the epoxy ring ensure that epoxy resins are extremely versatile and may undergo cure with a wide variety of commercial curing agents (see Fig. 2 for typical structures).

Epoxy Resin Cure

The versatility of the epoxy group has already been highlighted: reaction may occur with a number of different curatives via different reaction mechanisms with associated kinetic expressions. Furthermore, the topic is a complex one and falls outside the scope of this entry, but some general observations may be made. In the initial stages of cure, when the cure temperature may greatly exceed the glass transition temperature of the uncured epoxy mixture, the reaction mixture comprises low molecular weight species (e.g., monomers, dimers, trimers, etc.). As cure proceeds, the epoxy initially undergoes chain extension, before further branching takes place to yield large, highly branched molecules; during these early stages, the reaction is under chemical control. With further branching, the reaction mixture experiences a rapid increase in viscosity and a critical point is reached when gelation occurs: the formation of an incipient three-dimensional network. If the cure temperature is too low, then vitrification may occur before gelation occurs; further reaction is inhibited as the glass transition temperature of the curing mixture reaches the cure temperature and molecular mobility is limited. Thus as cure proceeds, the increasing viscosity ensures that the kinetics of the reaction mixture become increasingly subject to diffusion control in determining the rate and extent of reaction. Prior to gelation, the reaction mixture contains a significant proportion of soluble material (the "sol"); after this point, the weight fraction of sol decreases dramatically (essentially approaching zero) as the gel fraction grows. Furthermore, the incipient network undergoes further growth as the cross-link density increases to tie polymer chains together, and it is often in these later stages of reaction that the development of the thermomechanical properties, necessary for application of the epoxy resins, occurs. The relationship between cure, C, temperature, T, and elastic property, P, may be depicted using a method proposed by Gillham and coworkers (Fig. 3).

In general, the cure may be monitored using a variety of techniques:

i. Direct assay of the concentration of functional groups.
ii. Thermal analysis.
iii. Rheological changes during cure.

During the early stages of reaction, solution-based analysis techniques (e.g., "wet" chemical tests, such as titrimetry or chromatography: HPLC or GPC) may be used to determine the extent of network growth. However, when the cure reaches the gel stage, the reduction in solubility renders such methods ineffective, necessitating the use of more exotic, nonsolution-based

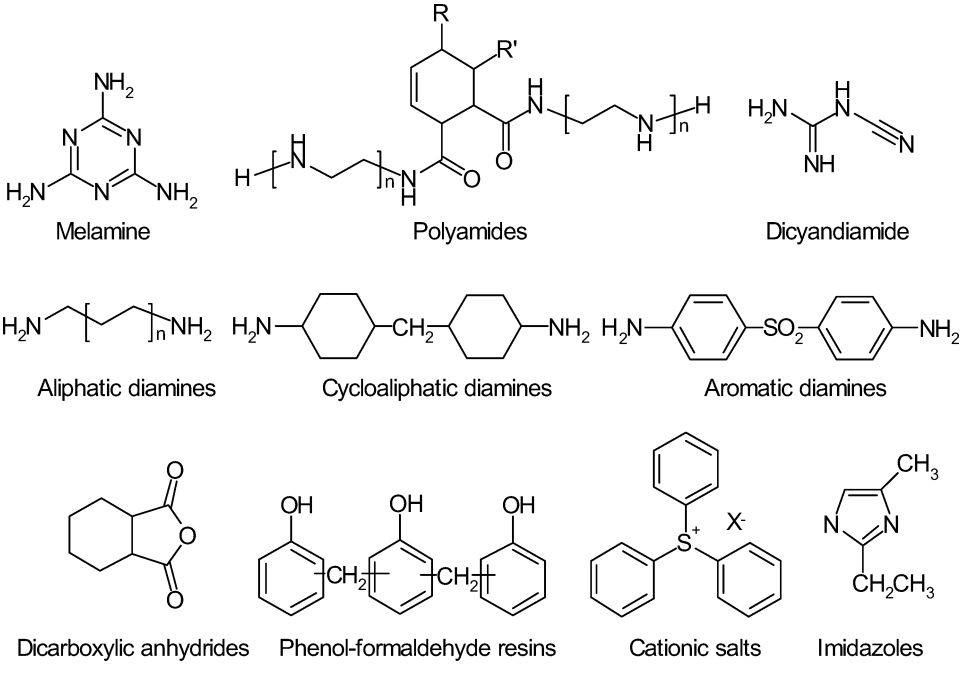


Fig. 2 Structures of selected catalysts and curing agents commonly used for the cure of epoxy resins.

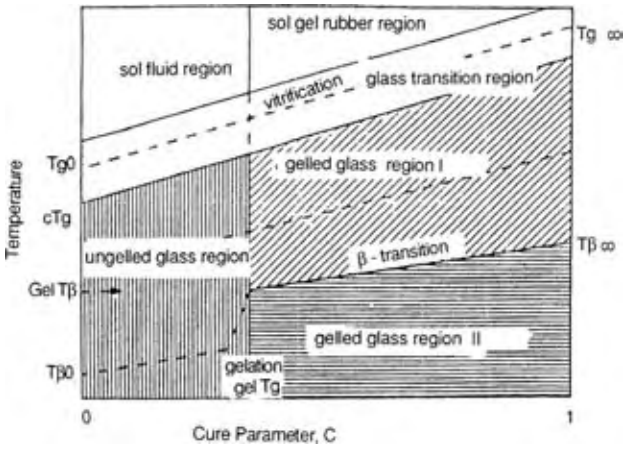


Fig. 3 The CTP diagram, a modified version of the T_gTP proposed by Wang and Gillham.[9] $C = (T_g - T_{g0})/(T_{g\infty} - T_{g0})$, where T_{g0} is the initial glass transition temperature of the uncured resin; $T_{g\infty,}$ is the glass transition temperature of the fully cured resin; T_g, is the glass transition temperature of the partially cured resin. (From Ref.[3].)

techniques, such as cross-polarization-magic angle spinning nuclear magnetic resonance (CP-MAS NMR) spectroscopy to obtain structural information.

Cured Epoxy Resin Properties

Undoubtedly, DGEBA remains the most widely researched and commercially most important epoxy resin, because of the impressive combination of its relatively low melt viscosity, physical strength after curing, and its relatively low production cost. However, the presence of glycidyl ether groups in the polymer not only imparts molecular flexibility and improves processability (by reducing the melt viscosity of the uncured resin), but they also decrease the heat resistance of the cured epoxy resin. This, coupled with relatively poor thermomechanical properties ($T_g < 120°C$), renders the polymer unacceptable for high performance applications (unless special aromatic curing agents are formulated into the final polymer). Where high performance is required, the tetrafunctional, predominantly aromatic TGDDM offers a good balance of properties (e.g., good long-term high-temperature properties, high mechanical strength retention, very low cure shrinkage, fine chemical resistance, etc.) at relatively low cost. The aerospace industry makes widespread use of TGDDM in mixtures containing the curing agent diaminodiphenylsulphone (DDS), for advanced structural composites and structural adhesives, etc. because of their high strength : weight ratio and relative ease of cure, although it is common to use a small quantity of a boron trifluoride ethanolamine complex as an accelerator. The ease of processing is also very attractive: the TGDDM (e.g., XU MY 722) may be heated to 140°C and DDS may be stirred in slowly and quite easily at this temperature to yield a clear, homogeneous blend before being degassed at 0.1 MPa for 15–30 min. The storage life of the degassed melt is exceptionally good and may be refrigerated in this state for several months. This is a key issue, because the production of complex composite components using lay up procedures may take many days to achieve, during which advancement in the prepreg is undesirable. Cure proceeds at 177°C via a combination of primary amine-epoxide, secondary amine-epoxide, and epoxide-hydroxyl reactions; whilst, according to infrared spectroscopy, ether formation occurs to a significant degree in the later stages. Table 3 displays a selection of cured resin properties for mixtures containing near stoichiometric quantities of DDS, from which the effects of both mixture composition and curing schedule may be appreciated.

STRATEGIES FOR TOUGHENING EPOXY RESINS

Conventional polyfunctional epoxy monomers, in common with most high performance thermoset polymers for load bearing applications, can be classed as brittle in fracture (i.e., typically offering G_{IC} values of $300\,J/m^2$ or less). The inevitable combination of high aromatic content (i.e., comprising rigid rings) and high cross-link density following cure not only leads to the desired elevation in glass transition temperature, stiffness, and chemical resistance in the polymer, but also to deleterious effects on its fracture toughness. The latter reflects the reduction in the molecular mobility of the matrix and its ability to yield under stress, and thus absorb energy. Consequently, for some 20 years or more, much research effort has been directed toward the development of different approaches to increase the toughness of epoxy resins. It is now generally accepted that the principal microdeformation process taking place in loaded, cross-linked epoxy resins is localized by shear yielding in areas of high stress concentration. Several general approaches have been examined in detail for improving this most important performance parameter and is discussed individually.

Mineral Filler Modification

The incorporation of glass microballoons has already been mentioned in the context of density reduction.

Table 3 Selected properties of cured TGDDM-DDS mixtures

	Cure schedule		
Composition, phr	**A**	**A**	**B**
TGDDM			
Araldite MY 720	100	—	—
Araldite MY 721	—	100	
Araldite XU MY 722	—	—	100
4,4′-DDS	44	49	50
Tensile strength, MPa			
25°C	58.9	48.3	58.1
150°C	44.5	51.7	—
Tensile modulus, GPa			
25°C	3.7	3.9	4.2
100°C	2.6	2.6	—
Tensile elongation, %			
25°C	1.8	1.3	1.6
100°C	1.9	2.3	—
Flexural strength, MPa, 25°C	89.6	126.9	124.5
Flexural modulus, GPa, 25°C	3.4	3.6	4.0
Compression properties			
Ultimate strength, MPa, 25°C	234.4	—	—
Yield strength, MPa, 25°C	199.9	—	—
Modulus, GPa, 25°C	1.9	—	—
Heat deflection temperature,°C	238	—	—
Glass transition temperature,°C	177	265	240
Water uptake, %	—	—	3.7

TGDDM: *N,N,N′,N′*-tetraglycidyl-4,4′-diaminophenylmethane; DDS: diaminodiphenylsulfone; phr: parts per hundred of resin; A: 80°C (2 hr), 100°C (1 hr), 150°C (4 hr), 200°C (7 hr); B: 180°C (2 hr), 210°C (2 hr).
(Adapted from Table 8.4 in: Lin, S.-C.; Pearce, E.M. High-performance epoxy resins, Chapter 8. In *High-Performance Thermosets, Chemistry, Properties, Applications*; Hanser: New York, 1998, p. 259.)

However, it has been known for some 20 years that significant enhancements in the modulus and toughness of epoxy resins can be made through the introduction of small glass spheres (microspheres) or silica. The results are impressive: the addition of ca. 40 vol.% of glass beads can lead to a doubling in Young's modulus and an increase in the fracture toughness (K_{IC}) of some 400% with no reduction in tensile strength.[10,11]

Rubber (Elastomer) Modification

The use of rubbers (particularly epoxy-terminated butadiene nitrile, ETBN, rubber or carboxy-terminated butadiene acrylonitrile, CTBN, rubber) to toughen thermoset polymers is perhaps the most widely explored method and has been applied with some measure of success in epoxy resins. Phase separation of the second rubbery phase occurs during cure and its incorporation in the epoxy matrix can significantly enhance the fracture toughness of the thermoset. Although the rubber has a low shear modulus, its bulk modulus is comparable to the value measured for the epoxy, ensuring that the rubber inclusions introduced during modification can play a full role in bearing a significant load within the polymer. Several mechanisms have been proposed to account for the observed increase in fracture toughness for rubber modified epoxy resins: shear yielding, void growth, and rubber bridging, although the contribution that each phenomenon makes to the mechanism is the subject of both research and debate. Huang and Kinloch concluded that at lower temperatures shear yielding represents the principal energy-absorbing mechanism (and that rubber bridging made an appreciable contribution at subzero temperatures), whilst at elevated temperatures shear yielding and void growth become equally important.[12] The incorporation of a dispersed rubber phase, with each particle acting as a stress concentration, offers numerous sites at which plastic deformation can occur and, where stress concentration occurs at the tip of a crack, the rubber particles undergo failure and cavitation. The latter leads to the production of voids within the polymer matrix and these grow in response to further loading, thus dissipating energy.

Apart from the nature of the actual rubber employed, the degree of toughness enhancement is influenced by a variety of factors, including the volume

fraction of elastomer incorporated, the molecular weight of the resin (Fig. 4), the size and size distribution of the rubber particles, and the magnitude of the interfacial bonding. In general, for CTBN-toughened epoxies, beneficial enhancements in fracture toughness are found for rubber loadings up to ca. 18–20 wt.%, whilst toughened epoxies typically contain many particles in the size range 1–5 μm, which contain both rubber and resin with smaller phase-separated particles (<0.2 μm). It is clear from a review of the literature in this area that there is some disagreement as to whether submicron or larger particles are more efficient toughness enhancers. Contradictory results have been reported, although the difference in particle size may lead to differences in the mechanism through which toughness is enhanced. Thus, larger particles may be acting as bridging particles, whereas smaller particles may be undergoing cavitation within the crack tip zone.[13]

Core-Shell Rubber Modification

A variant of rubber toughening involves the use of preformed core-shell rubbers comprising a highly cross-linked polybutadiene core with a grafted shell of a vinylic polymer. In this case, the particles are small, typically ca. 0.1 μm, and thus have little effect on the observed viscosity of the epoxy. One of the principal advantages of this over simple rubber toughening is the ability to produce predetermined controllable morphology in the cured polymer.[14]

Thermoplastic Modification

The incorporation of engineering thermoplastics, such as poly(arylene ether sulfone)s, polysulfone, poly(phenylene oxide), or poly(ether imide) is a more recent innovation[15,16] and has also been shown to yield significant improvements in the toughness of cured epoxy resins.[17,18] Importantly, however, unlike their rubber modified analogs, the introduction of engineering thermoplastics into epoxy resins does not result in a reduction in modulus, high-temperature properties, or creep resistance. Whilst original attempts to toughen epoxy resins in this manner yielded only modest improvements in fracture toughness, it has been shown that enhancements close to those obtained for CTBN can be achieved.[17] Bucknall and Patridge[19] have demonstrated this for a number of epoxies, and is shown in Fig. 5 that the incorporation of PEI serves to increase the toughness of the cured epoxy markedly, although the T_g of the PEI phase is always lower than that of the neat thermoplastic because of plasticization of PEI by low molecular weight species.

The approach taken by Sefton et al.[18] for the synthesis of novel thermoplastics designed to undergo phase separation from the thermoset epoxy matrix is complementary to the method employed by Bucknall and Patridge,[19] in which the nature of the thermoset is varied to achieve a similar result.

Several basic morphologies are observed in thermoplastic-modified epoxies and, indeed, other thermosets. Homogeneous [Fig. 6(A), in which no phase separation is observed] and particulate [Fig. 6(B), in which the modifier phase separates to produce small domains] morphologies occur at low concentrations of modifier. In these cases, the thermoplastic modifier is encapsulated within a thermoset matrix, whereas in the phase-inverted morphology, [Fig. 6(C)] the minor thermoplastic component is the continuous phase surrounding large, discontinuous domains of the major

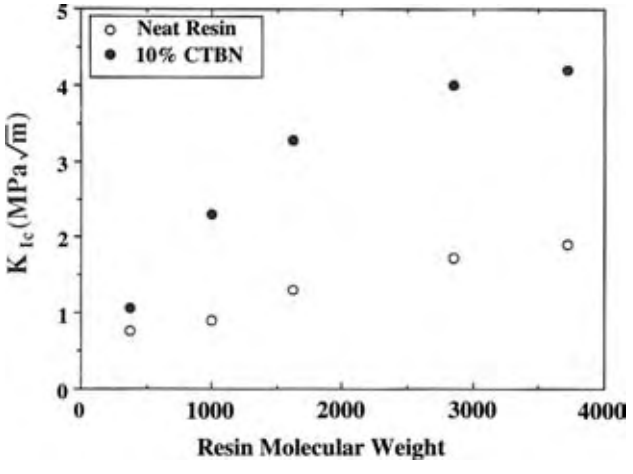


Fig. 4 Plot of fracture toughness vs. resin molecular weight for neat and CTBN-modified DGEBA-DDS.[13] (From Ref.[3].)

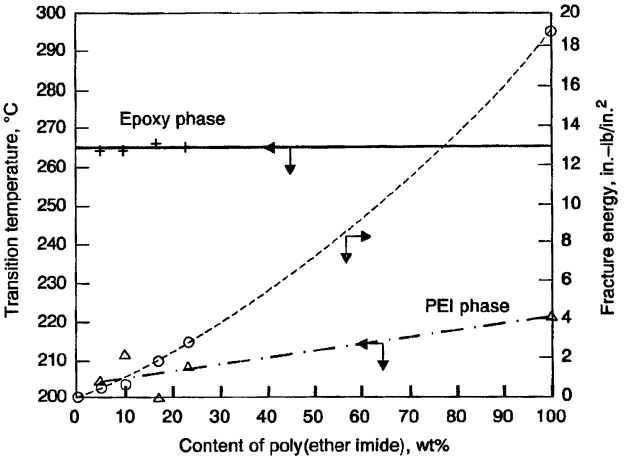


Fig. 5 Fracture energy and glass transition temperature of poly(ether imide) PEI, *N,N,N′,N′*-tetra-glycidyl-4,4′-diaminodiphenylmethane, and diaminodiphenylsulphone. (From Ref.[20].)

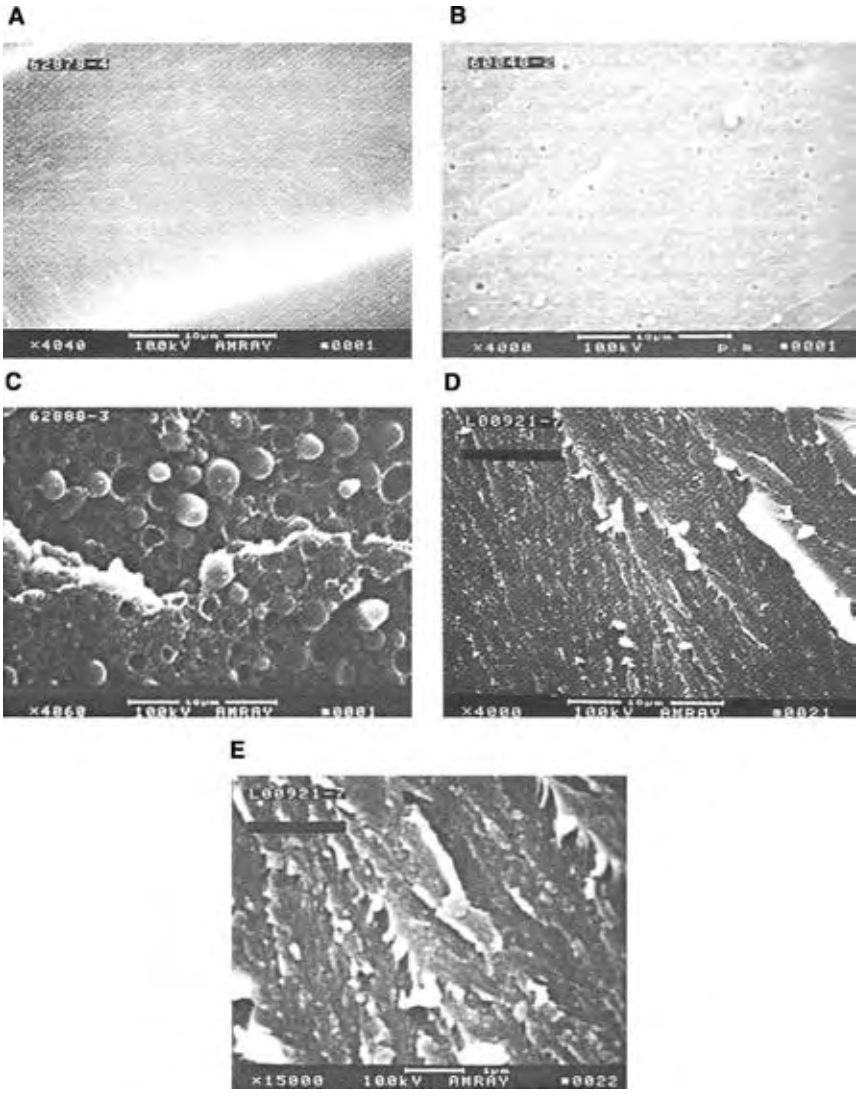


Fig. 6. Scanning electron micrographs of various morphologies produced with tailor-made thermoplastics toughened epoxy resins: (A) homogeneous, (B) particulate, (C) phase inverted, (D) and (E) cocontinuous. (From Hamerton, I., Ed.; *Chemistry and Technology of Cyanate Ester Resins*; Blackie Academic and Professional: Glasgow, 1994.)

thermoset component. A final morphology, designated ''cocontinuous'' [Figs. 6(D) and 6(E)], occurs in semi-interpenetrating networks and results in both phases remaining continuous in three dimensions. Practically, the larger phase-inverted morphologies may be observed using scanning electron microscopy (SEM; see Fig. 6), whereas the smaller scale of the cocontinuous morphologies, in which the phase sizes reflect the higher degree of compatibility between thermoset and thermoplastic, requires the use of transmission electron microscopy (TEM; see Fig. 7).

Flame Retardant Epoxy Resins

The demand for flame retardant (FR) epoxies is extremely high owing to their widespread use in adhesives, coatings, advanced composites, and the electronics industry. Currently, extensive use is made of bromine to impart flame retardancy and a typical laminating resin for printed wiring boards is an FR4 epoxy resin. This comprises DGEBA advanced with tetrabromobisphenol A (Fig. 8) and a cross-linking agent package based on a styrene–maleic anhydride copolymer and an optionally brominated bisphenol A and/or an optionally brominated DGEBA. In 2004, the typical price for a brominated epoxy (in liquid, solid, and solid solution forms) was around 220 €/tonne.

However, both the European Community (EC) and the United States (U.S.) Government have expressed concern about the use of FRs, particularly those containing halogens, because of the toxicity and perceived adverse environmental impacts of these compounds. Thus, the EC has proposed to restrict the use of brominated diphenyl oxide FRs because highly toxic and potentially carcinogenic brominated furans and dioxins may form during combustion.[21] The World Health Organization and the U.S. Environmental Protection Agency have also recommended exposure limit and risk assessment of dioxins and similar compounds.

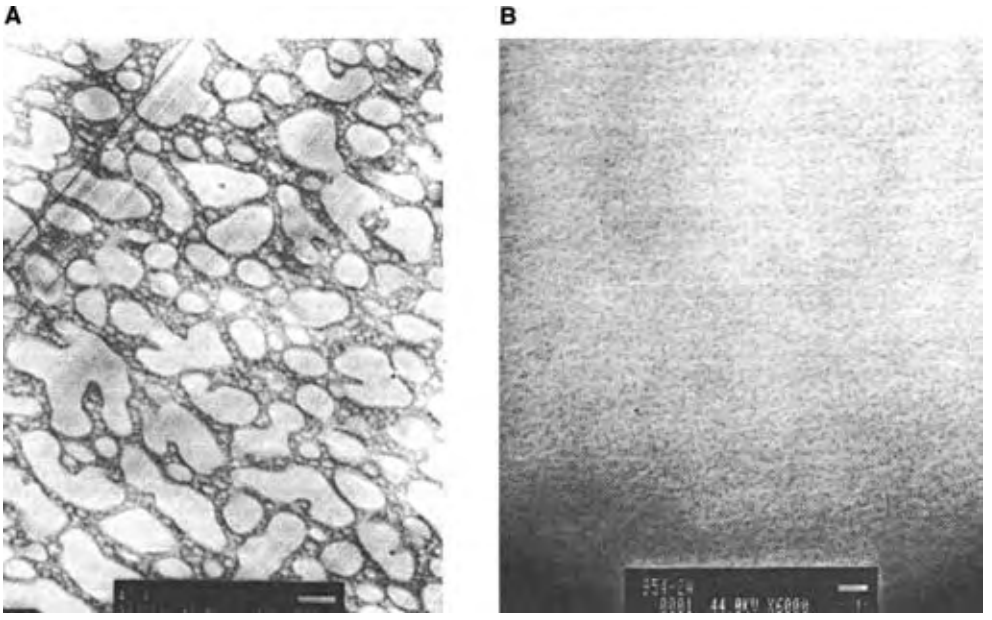


Fig. 7 Transmission electron micrographs of morphologies produced with tailor-made thermoplastics toughened epoxy resin (Fiberite 954-2A): (A) phase inverted and (B) cocontinuous. (From Hamerton, I., Ed.; *Chemistry and Technology of Cyanate Ester Resins*; Blackie Academic and Professional: Glasgow, 1994.)

Consequently, nonhalogenated FRs are being examined as alternatives to existing compounds, and an extensive literature already exists in this area.[22] For epoxy resins, the most effective route to improving the flame retardancy is the reactive approach, wherein inherently, FR groups are incorporated into the polymer backbone or side chain. Organophosphorus groups, such as dialkyl or diaryl phosphates (Scheme 4), are readily introduced via reaction with the epoxide group.

In this way, a range of polymerizable diepoxides or organophosphorus-containing curing agents based on phosphine oxide, phosphate, or 9,10-dihydro-9-oxa-10-phosphaphenanthrene-10-oxide (DOPO) have been prepared and evaluated (Fig. 9).

The incorporation of DOPO derivatives in DGEBA has been found to increase not only flame retardancy but also thermal stability. The DOPO-based diamine shown in Fig. 9, having a reactivity toward DGEBA intermediate between DDS and DDM, has also proved to be an effective FR. It displays a char yield (in a DEGBA/DOPO-amine) polymer of 32% at 700°C under nitrogen, compared with 13–15% for a phosphorus-free epoxy. Tricyclophosphazenes are already known to impart flame retardancy,[23] and whilst polyepoxides containing the tricyclophosphazene structure have been synthesized, the preparative route is perhaps less cost effective than that of the corresponding polyamine (Fig. 9, top left). Incorporation of the tricyclophosphazene moiety might be achieved more readily through coreaction with a conventional commercial epoxy monomer. Silicon, which shares many aspects of its chemistry with carbon, is another element that has been examined widely as a potential FR. Research has shown that the addition of relatively small amounts of silicon can improve the flame retardancy of polymers significantly, both through char formation in the condensed phase and the trapping of active radicals in the vapor phase. Incorporation can be achieved via epoxy- or amino-functionalized linear silanes or oligosiloxanes (see Fig. 10).[24] The latter can be polymerized in their own right (with or without curing agents), but also blended and copolymerized with conventional epoxy monomers. Incidentally, synergistic effects may be obtained by combining phosphorus- and silicon-containing epoxies or curing

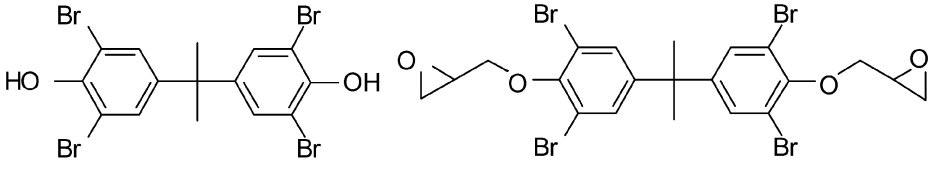
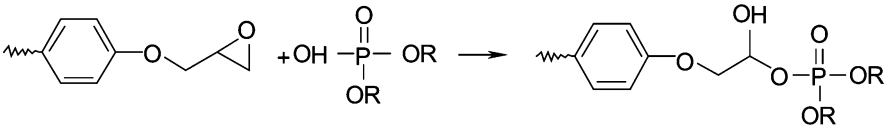


Fig. 8 Tetrabromobisphenol A (left) and brominated DGEBA (right).



Scheme 4 Dialkyl (or diaryl) phosphate modified epoxy resin, where R is methyl, ethyl, butyl, or phenyl. (Redrawn from Ref.[22].)

agents: a polymer containing 4.8% P and 12.7% Si achieves a limiting oxygen index (LOI) of 41 (where a value of at least 26 is required for a material to be qualified as self-extinguishing).

Polyhedral oligomeric silsesquioxane (POSS) reagents are becoming an increasingly important route to the preparation of nano-reinforced polymers (or ''nanocomposites''). POSS molecules are cube-shaped hybrid inorganic–organic structures comprising silicon atoms at the vertices, sandwiching oxygen atoms. Each of the silicon atoms may be functionalized, using a wide variety of chemical transformations, so that mono-, di-, or poly-functionalized POSS reagents may be prepared in high yield and purity.[25] In some cases, highly functionalized POSS-epoxides may present problems because of high crystallinity (making liquid incorporation difficult) or high cross-link density (in which high degrees of conversion may be difficult to achieve, for example, conversions of 78–90% have been reported for such materials)[26] and mono-substituted epoxy-POSS reagents may offer the best route for producing nanocomposites bearing pendant POSS structures.

TECHNOLOGICAL APPLICATIONS FOR EPOXY RESINS

A wide range of commercial epoxy monomers are available and cured epoxy resins find application in areas as diverse as protective coatings, electrical applications, reinforced resins (composites), bonding and adhesives, flooring, and tooling and casting. Whilst a full discussion is beyond the scope of this entry, several of the principal applications are addressed.

Surface Coatings

Along with dental adhesives, the development of applied epoxy resins for protective or decorative purposes was one of the principal reasons for the original development of epoxy resins in the U.S.A. Some 50 years

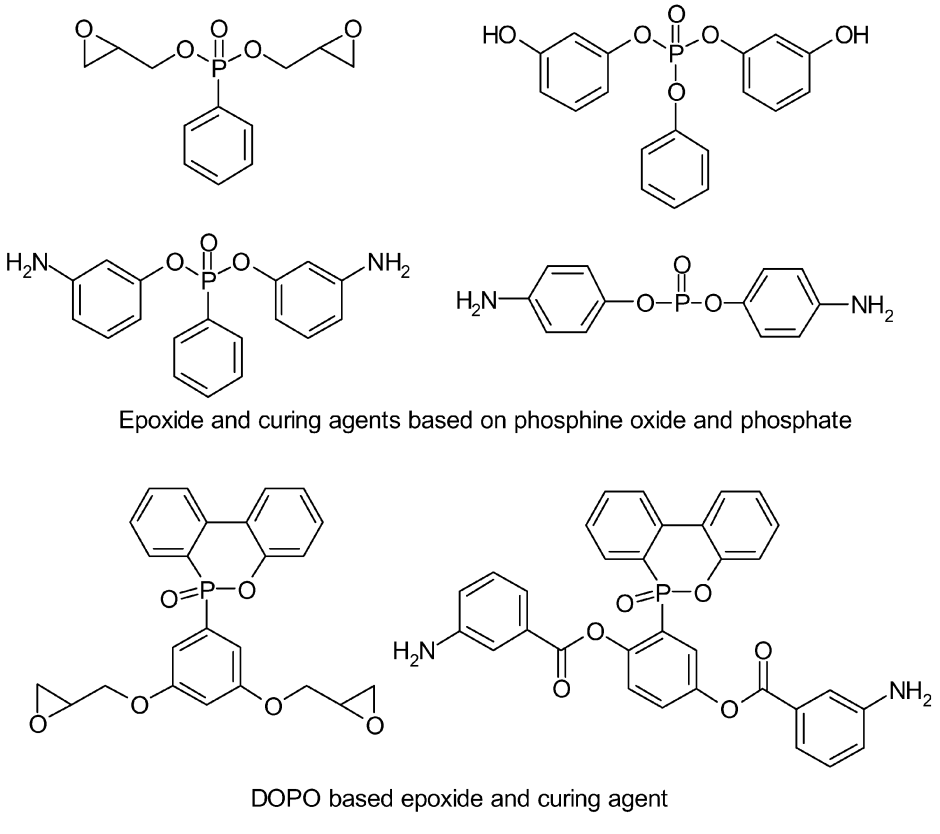


Fig. 9 Selected phosphine oxide-, phosphate-, and DOPO-based epoxides and curing agents. (Redrawn from Ref.[22].)

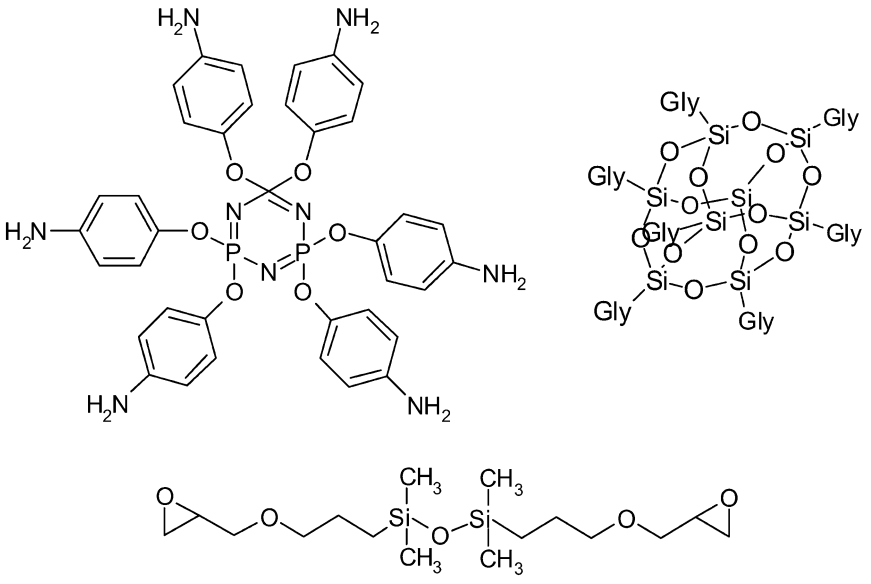


Fig. 10 Selected tricyclophosphazene (top left), POSS- (top right), and silane-based (bottom) epoxides and curing agents.

later, during the early 1990s, around half of the epoxy resins in Western Europe and the U.S.A. was used in protective coatings (often in the construction industry), which was by far the largest single category (Table 4). This is an unusually high figure for a family of polymers that is generally more expensive than their market rivals for coatings or laminates (e.g., phenolic resins). Consequently, it is of paramount importance that for an epoxy to be used to supplant a phenolic, it should have markedly superior properties to offset the cost.

Electrical Properties of Epoxy Resins

Epoxy resins are finding increasing use in many applications associated with the electronics industry (e.g., encapsulation, potting, thin film coating, embedding or packing of electronic circuits, etc.). This is partly attributable to their versatility discussed previously (i.e., ease of processing, a wide choice of curing agents, etc.).

Table 4 Uses of epoxy broken down by market sector

	1967 (%)	1990 (%)	1991 (%)
Protective coatings	55	49	51
Electronic applications		14	13
Reinforced resins	20	7.5	8
Bonding and adhesives	5	7.2	7.25
Flooring		6.5	6.25
Tooling and casting	10	7.5	7.25
Other	10	8.3	7.25
Total	100	100	100

(Adapted from Modern Plastics International (**1992**, Jan): 48.)

Cured epoxy resins have a combination of electrical properties that make them attractive in electrical and electronic applications:

- Dielectric permittivity of ca. 3–6 at ambient temperature and low frequencies.
- Dissipation factor, $\tan\delta$, of 10^{-3} to 10^{-2} for 60–1000 Hz.
- Dielectric strength ca. 120–180 kV/mm.
- Volume resistivity of 10^{19}–$10^{12}\,\Omega$/m.
- Good adhesion to metals and silicon.
- Low permeability or absorption of water (typically 0.1–0.25% for certain epoxies or epoxy formulations).
- Low viscosity prior to cure.
- Low stress resistance.
- Low heat of reaction.

Although some competitor resins (e.g., polyimides and cyanate esters) are replacing epoxy resins in some more demanding applications, in which the superior glass transition temperatures or lower dielectric permittivity/low dielectric loss are preferred, brominated epoxies are still widely used.

Epoxy-Based Composite Materials

Epoxy resins have found application in carbon fiber reinforced composites for some 30 years or more and the benefits are well documented (Table 5). The traditional limitations are also summarized simply in the same table.

Epoxy resins are by far the most widely used polymer matrices for advanced structural composites and, if carbon, glass, and aramid fiber reinforced epoxies

Table 5 General characteristics of epoxy-based carbon fiber reinforced plastics (CFRPs)

Benefits conferred by epoxy-based CFRPs	Traditional limitations epoxy-based CFRPs
Weight saving over aluminum alloys (high specific stiffness and strength)	Susceptibility to operational impact damage
Tailored directional mechanical properties	Restricted environmental stability in terms of temperature and moisture absorbency
Reduced component count over metallic equivalents	Excessive localized damage through lightning strikes
Modified radar response compared with metallics	Uncertainties on repair techniques
Resistance to corrosion in saline environments	Cost compared with metallics
Excellent fatigue resistance	Labor-intensive and time-consuming manufacture process
Dimensional stability	

are considered together, they accounted for over 90% of the aircraft market alone in 1990.[27] Table 6, shows the wide diversity of applications to which epoxy-based advanced composite materials have been put in to use.

In Table 7, typical key laminate properties for thermoset and thermoplastic composites are given. It is immediately apparent that suitable toughening (discussed previously) can achieve superior compression after impact and comparable edge delamination strength to composites prepared from a thermoplastic matrix.

The fabrication of components from advanced composite materials also requires suitable tooling to shape and conform the prepreg during lay-up and epoxy have a role to play in this operation as well. Whilst less durable than conventional tooling constructed from steel or aluminum, tooling prepared from carbon fiber reinforced epoxy resins offers lightweight and, most importantly, a good match with the

Table 6 Selected applications for epoxy-based advanced structural composites

Aircraft	
Structural	CF/epoxy suitable in less thermally critical areas
Interior components	Cargo liners – monolithic laminate of aluminum, E or S-2 glass, of Kevlar with epoxy, polyester or phenolic
Space	
Launch systems	CF/epoxy composite motor cases
Communication Satellites	Antennae and large bus (housing) structures
Spacecraft	CF/epoxy metering truss (Hubble space telescope)
Space shuttle	CF/epoxy payload bay doors
Sports/leisure	
Golf	Medium or high modulus CF/epoxy club shafts
Tennis racquets	Modified epoxy/CF, boron, Aramid, and Spectra head/frame
Bicycles	CF/epoxy frame and spokeless/disk wheels
Skis	CF/epoxy skis and poles—high stiffness and compressive strength
Fishing rods	Lightweight CF/glass/epoxy hybrid rods
Archery	CF/thermoset bow (prepreg or pultruded) and pultruded CF arrow (resist bending and shattering)
Marine/naval vessels	CF/epoxy submarine hulls
Construction	Filament wound epoxy pipes
Tooling	CF/epoxy composite—fine CTE match with prepreg
Automotive	Composite drive shaft - CF/epoxy/CF stiffened aluminum shaft, passenger car floors, epoxy/GFRP springs
Racing cars	Oriented CF/epoxy composite chassis—improved energy absorption
Transportation	Epoxy/glass faced aramid honeycomb for monorail
Masts, antennae, radomes	Improved strength and dielectric properties over GFRP
Musical instruments	Epoxy/CF sound boards as facings for violins and guitars—more reliable than wood, retains tuning, improved sound projection

Table 7 Typical laminate properties on intermediate modulus carbon fiber at a fiber volume of ca. 60%

Test	TGDDM epoxy	Toughened epoxy	Amorphous thermoplastic
0° Tensile strength (MPa)	2,400	2,400	2,400
±45° Tension			
Modulus (GPa)	21	18	19
Strength (MPa)	186	255	248
Shear modulus	5.6	6.3	4.8
0° Compression			
Modulus (GPa)	152	152	152
Strength (MPa)	1,700	1,700	1,300
OHT strength (MPa)	480	449	483
OHC strength (MPa)	310	304	269
CAI (MPa) at 6,700 J/m	160	311	304
Compression interlaminar (MPa)	90	76	69
EDS (MPa)	214	290	311

OHT: open-hole tensile; OHC: open-hole compression; CAI: compression strength after impact; EDS: edge delamination strength. (Adapted from Table 4 in: Pilato, L.A.; Michno, M.J. *Advanced Composite Materials*; Springer-Verlag: Berlin, 1994; 118.)

linear coefficients of thermal expansion for both carbon fiber prepreg (2.5–3.6 ppm/°C) and carbon fiber wet lay-up (5.5 ppm/°C). This contrasts sharply with, e.g., mild steel (11.0 ppm/°C) or aluminum (23.0 ppm/°C), the expansion behavior of which varies widely from the prepreg undergoing cure.

Epoxy Resin Adhesives

Originally produced during the two decades following the Second World War, epoxy resin adhesives have grown in popularity such that they are now used widely in many commercial and domestic settings. Over the years, epoxy resins have gained a large measure of popularity as adhesives for a combination of favorable characteristics:

- Excellent adhesion to a variety of substrates including most metallic alloys.
- Short- and longterm operation up to ca. 150°C.
- Highly versatile and able to achieve a wide range of processing, cure, and properties characteristics.
- No volatiles evolved during cure—no requirement for high pressure bonding operations.
- Good wetting properties on well prepared surfaces.
- Relatively low cure shrinkage.

DGEBA is the most commonly used epoxy in adhesive formulations. Although they may range from very simple, low molecular weight, primarily aliphatic resins to more complex, multifunctional, primarily aromatic materials. A wide choice of curing agents is available, in common with other applications. Schmidt and Bell[28] published a very thorough discussion of the adhesion of epoxy resins to metal surfaces. Initial adhesion strength values achieved by epoxy adhesives and coatings are principally because of the inherent epoxy chemistry: ring opening leads to the generation of aliphatic hydroxyl and ether groups in the cured polymer. This high polarity offers sites for the formation of strong hydrogen bonds between epoxy molecules and metal oxides (with typical bond energies of 5–10 kcal/mol).

CONCLUSIONS

Epoxy resins undergo cure to yield three-dimensional networks from a liquid or thermoplastic monomer or oligomer, and the resulting product is a hard, insoluble, and infusible product. Cured epoxy resins offer a good balance of thermal, mechanical, and electrical properties, as well as good adhesion to a variety of substrates, hardness, chemical inertness (in the cured state), and thermal resistance. Additionally, epoxies may be modified easily with a range of additives to improve fracture toughness and flame resistance. Commercial epoxy resins find application in many technological applications, often as market leaders: protective coatings, electrical applications, reinforced resins (composites), bonding and adhesives, flooring, and tooling and casting.

REFERENCES

1. Castan, P. (for De Trey AG) CH 211116 (1938); DRP 749512 (1938); GB 518057 (1938).

2. Greenlee, S. (for DeVoe and Raynolds) USP 2456408 (1948).
3. Ellis, B. Introduction to the chemistry, synthesis, manufacture and characterization of epoxy resins. In *Chemistry and Technology of Epoxy Resins*; Ellis, B., Ed.; Blackie Academic & Professional: Glasgow, 1993; 1–36.
4. Tanaka, Y. Synthesis and characterization of epoxides. In *Epoxy Resins—Chemistry and Technology,* 2nd Ed.; May, C.A., Ed.; Marcel Dekker: New York, 1988; 9–283.
5. Mormann, W. Liquid crystalline thermosets. Trends Polym. Sci. **1995**, *3*, 255–261.
6. Barclay, C.G.; McNamee, S.G.; Ober, C.K.; Papathomas, K.I.; Wang, D.W. Liquid crystalline epoxy thermosets: mechanical and magnetic alignment. J. Polym. Sci. Polym. Chem. **1992,** *30*, 1845–1853.
7. Mertzel, E.; Kœnig, J.L. Application of FT-IR and NMR to epoxy resins. Epoxy resins and composites 2. In *Advances in Polymer Science*; Dŭsek, K., Ed.; Springer-Verlag: Berlin, 1986; Vol. 75, 73–112.
8. Moniz, W.B.; Poranski, C.F., Jr. Epoxy resin chemistry (Chapter 7). In *ACS Symposium Series*; No. 114; American Chemical Society: Washington, DC, 1979.
9. Wang, X.; Gillham, J.K. Tg—temperature property (TgTP) diagram for thermosetting systems: anomalous behaviour of physical properties vs. extent of cure. J. Coatings Tech. **1992,** *64*, 37–45.
10. Moloney, A.C.; Kausch, H.H.; Stieger, H.R. The fracture of particulate-filled epoxide resins. J. Mater. Sci. **1984,** *19*, 1125–1130.
11. Spanoudakis, J.; Young, R.J. Crack-propagation in a glass particle-filled epoxy resin. 1. Effect of particle-volume fraction and size. J. Mater. Sci. **1984,** *19*, 473–486.
12. Huang, Y.; Kinloch, A.J. The role of plastic void growth in the fracture of rubber-toughened epoxy polymers. J. Mater. Sci. Lett. **1992**, *11*, 484–489.
13. Cantwell, W.J.; Kausch, H.H. Fracture behaviour of epoxy resins. In *Chemistry and Technology of Epoxy Resins*; Ellis, B., Ed.; Blackie Academic & Professional: Glasgow, 1993.
14. Becu, L.; Sautereau, H.; Maazouz, A.; Gerard, J.F.; Pabon, M.; Pichot, C. Synthesis and structure–property relationships of acrylic core-shell particle-toughened epoxy networks. Polym. Adv. Technol. **1995,** *6*, 316–325.
15. Hedrick, J.L.; Yilgör, I.; Wilkes, G.L.; McGrath, J.E. Chemical modification of matrix resin networks with engineering thermoplastics. I. Phenolic hydroxyl terminated poly(aryl ether sulfone)-epoxy systems. Polym. Bull. **1985**, *13*, 201–208.
16. Bucknall, C.B.; Patridge, I.K. Polym. Eng. Sci. **1986,** *26*, 54–60.
17. Pearson, R.A.; Yee, A.F. Proceedings of the Eighth International Conference on Deformation, Yield and Fracture of Polymers; PRI Paper 40, 1991.
18. Sefton, M.S.; McGrail, P.T.; Peacock, J.A.; Wilkinson, S; Crick, R.A.; Davies, M; Almen, G. 19th International SAMPE Technical Conference, 1987; Vol. 19, pp. 700.
19. Bucknall, C.B.; Patridge, I.K. Phase separation in epoxy resins containing polyethersulphone. Polymer **1983,** *24*, 639–644.
20. Gilbert, A.H.; Bucknall, C.B. Epoxy resins toughened with thermoplastic. In *Makromol. Chem. Macromol. Symp*; Vol. 45, 289–298.
21. Kracklauer, J. *Flame-retardant Polymeric Materials*; Lewin, M., Atlas, S.M., Pearce, E.M., Eds.; Plenum Press: New York, 1978; Vol. 2.
22. Lu, S.-Y.; Hamerton, I. Recent developments in the chemistry of flame retardant composite matrices. Prog. Polym. Sci. **2002**, *27*, 1661–1712.
23. De Jaeger, R.; Gleria, M. Poly(organophosphazene)s and related compounds: synthesis, properties and applications. Prog. Polym. Sci. **1998**, *23*, 179–226.
24. Wang, W.J.; Perng, L.H.; Hsiue, G.H.; Chang, F.C. Characterization and properties of new silicone-containing epoxy resin. Polymer **2000**, *41*, 6113.
25. Provatas, A.; Matisons, J.G. Silsesquioxanes: synthesis and applications. Trends Polym. Sci. **1997**, *5*, 327–332.
26. Crivello, J.V.; Malik, R. Synthesis and photoinitiated cationic polymerization of monomers with the silsesquioxane core. J. Polym. Sci. Polym. Chem. **1997**, *35*, 407.
27. Bashford, D. The potential of composite materials. In *Composite Materials in Aircraft Structures*; Middleton, D., Ed.; Longman Scientific and Technical: Harlow, 1990; 9–16.
28. Schmidt, R.G.; Bell, J.P. Epoxy resins and composites 2. In *Advances in Polymer Science*; Dŭsek, K., Ed.; Springer-Verlag: Berlin, 1986; Vol. 75, 33–71.

BIBLIOGRAPHY

Bauer, R.S., Ed.; *Epoxy Resin Chemistry II*; American Chemical Society: Washington, D.C, 1983.
Bruins, P.F., Ed.; *Epoxy Resin Technology*; Interscience Publishers: New York, 1968.
DiStasio, J.I., Ed.; *Epoxy Resin Technology*; Noyes Data Corporation: Park Ridge, 1982.
Dŭsek, K., Ed.; *Epoxy Resins and Composites 1, Advances in Polymer Science*; Springer-Verlag, Berlin, 1985; Vol. 72.

Dŭsek, K., Ed.; *Epoxy Resins and Composites 2, Advances in Polymer Science*; Springer-Verlag, Berlin, 1986; Vol. 75.

Dŭsek, K., Ed.; *Epoxy Resins and Composites 3, Advances in Polymer Science*; Springer-Verlag, Berlin, 1986; Vol. 78.

Dŭsek, K., Ed.; *Epoxy Resins and Composites 4, Advances in Polymer Science*; Springer-Verlag, Berlin, 1986; Vol. 80.

Flick, E.W. *Epoxy Resins, Curing Agents, Compounds and Modifiers. An Industrial Guide*; Noyes Data Corporation; Park Ridge, 1987.

Lee, H., Neville, K., Eds.; *Handbook of Epoxy Resins*; McGraw-Hill, New York, 1967.

May, C.A., Ed.; *Epoxy Resins – Chemistry and Technology;* 2nd Ed.; Marcel Dekker New York, 1988.

Potter, W.G. *Epoxide Resins*; Iliffe Books: London, 1970.

Ramney, M.W. *Epoxy Resins and Products*; Noyes Data Corporation: Park Ridge, 1977.

Saunders, K.J. *Organic Polymer Chemistry. An Introduction to the Organic Chemistry of Adhesives, Fibres, Paints, Plastics and Rubbers*, 2nd Ed.; Chapman and Hall: London, Chapter 18, Epoxies, 1988; 412–435.

Seymour, R.B., Kirshenbaum, G.S., Eds.; *High Performance Polymers: Their Origin and Development*; Elsevier: New York, 1986.

Ethylbenzene

Guy B. Woodle
UOP LLC, Des Plaines, Illinois, U.S.A.

INTRODUCTION

Ethylbenzene (EB) is a single-ring alkylaromatic compound that is used almost exclusively as an intermediate in the production of styrene monomer. On a commercial scale, essentially all EB is produced by alkylating benzene with ethylene. As of April 2004 approximately 30% of the worldwide EB production is carried out by liquid phase alkylation using a homogenous aluminum chloride catalyst in a process that was first commercialized in the 1930s.[1] Ethylbenzene was first produced commercially using a zeolite catalyst in a vapor phase reactor in 1980, which was a significant improvement over the aluminum chloride process. However, it was the development and commercialization of liquid phase and mixed liquid–vapor phase technologies in the 1990s that allowed for efficient production of high-purity EB. In 2003 the annual world EB production capacity was around 28.6 million metric tons, and EB production capacity is forecast to grow at an annual rate of just under 5% from 2004 to 2014.[2]

PHYSICAL AND CHEMICAL PROPERTIES

Ethylbenzene is a colorless aromatic liquid. It is only slightly soluble in water, but infinitely soluble in alcohol and ether. Additional properties are listed in Table 1. Ethylbenzene is chemically reactive with the most important reaction being its dehydrogenation to form styrene. Styrene is used to produce polystyrene, which is used in the manufacture of many commonly used products such as toys, household and kitchen appliances, plastic drinking cups, housings for computers and electronics, foam packaging, and insulation. In addition to polystyrene, styrene is used to produce acrylonitrile–butadiene–styrene polymer (ABS), styrene–acrylonitrile polymer (SAN), and styrene–butadiene synthetic rubber (SBR).

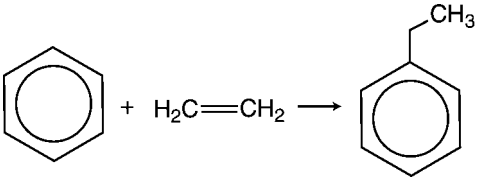
Ethylbenzene can also be oxidized to form ethylbenzene hydroperoxide, an intermediate in a process to produce propylene oxide.

REACTION KINETICS AND THERMODYNAMICS

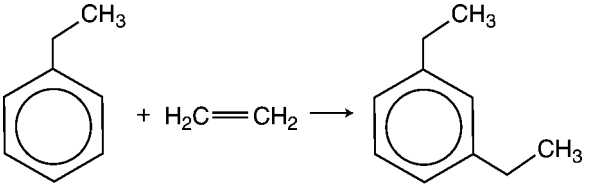
Commercially produced EB is based on alkylating benzene with ethylene.

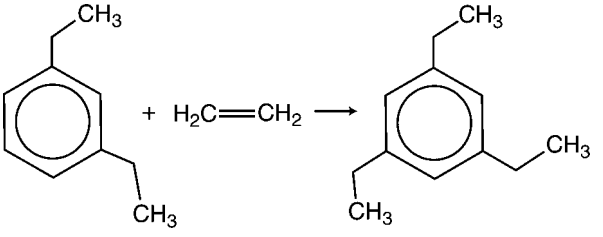
Benzene Alkylation with Ethylene

In the production of EB, alkylation is the reaction of ethylene with benzene according to the equation:



Successive alkylation reactions occur to a limited extent resulting in the formation of diethylbenzene and other higher ethylated benzenes, commonly called polyethylbenzene (PEB).





Kinetic reaction rate constants increase with the number of ethyl groups alkylated on the benzene ring. For example, the relative rate constant for alkylation of EB is roughly twice that for the alkylation of benzene. Reaction rate constants continue to increase with each successive alkylation reaction until a limitation is reached, such as steric hindrance. The formation of penta-EB and hexa-EB proceeds very slowly for this and other reasons so that only trace quantities are formed.

Alkylation reactions are exothermic. The initial alkylation of ethylene to benzene and each successive alkylation reaction generate roughly the same amount of heat (Table 2).

In addition to alkylation reactions, transalkylation reactions play a significant role in EB production.

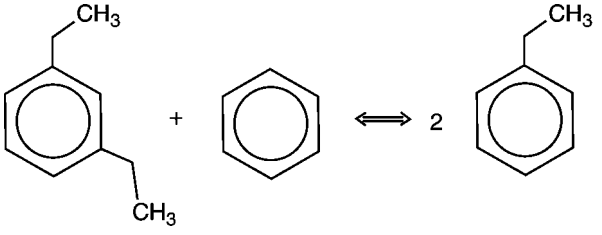
Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016171

Table 1 Physical properties of EB

Molecular weight	106.169
Specific gravity[a]	0.867
Melting point (°C)	–94.975
Boiling point (°C)	136.19
Critical temperature (°C)	346.4
Critical pressure (atm)	37
Vapor pressure[b] (mmHg at T°C)	
T (°C)	Vapor pressure (mmHg)
1	−9.8
5	13.9
10	25.9
20	38.6
40	52.8
60	61.8
100	74.1
200	92.7
400	113.8
760	136.2

[a]From *Perry's Chemical Engineers' Handbook*, 6th Ed.; p. 3–34. Density is at 20°C referred to water at 4°C.
[b]From *Perry's Chemical Engineers' Handbook*, 6th Ed.; p. 3–56.
(From *Ethylene and Its Industrial Derivatives*; S.A. Miller, Ed.; p. 900.)

Commercially it is found to be economically attractive to transalkylate all the PEB formed as a result of successive alkylation reactions with benzene in a separate transalkylation reactor to produce additional EB.



Transalkylation reaction rates are relatively slow and conversion is generally limited by equilibrium. For transalkylation reaction, the heat of reaction is essentially zero, which leads to a reactor that operates nearly isothermally.

The occurrence of both alkylation and transalkylation reactions results in a reaction chemistry that is affected by equilibrium. The equilibrium has been studied and is illustrated in Fig. 1. The horizontal axis is the ratio of ethyl groups to benzene rings and is often referred to as the ethyl-to-phenyl ratio. In the case of an alkylation reaction, the ethyl-to-phenyl ratio is the moles of ethylene to moles of benzene. Similarly, for a transalkylation reaction, the ethyl-to-phenyl ratio is equivalent to the moles of ethyl groups contributed by PEB to the moles of phenyl groups contributed by PEB plus benzene. At ethyl-to-phenyl ratios above about 0.6, the equilibrium EB concentration is relatively constant at about 48 wt% whereas the PEB concentration continues to increase as the ratio approaches 1.0. Most commercial reactors operate with ethyl-to-phenyl ratios less than 0.6. The equilibrium composition varies only slightly across the temperature range of commercial interest.

Table 2 Benzene ethylation thermodynamics

Alkylation reaction	**ΔH (500 K) (kcal/mol)**
Ethylene + benzene → ethylbenzene (EB)	−36.7
Ethylene + EB → di-ethylbenzene (DEB)	−28.3
Ethylene + DEB → tri-ethylbenzene (TEB)	−25.9
Ethylene + TEB → tetra-ethylbenzene	−25.5

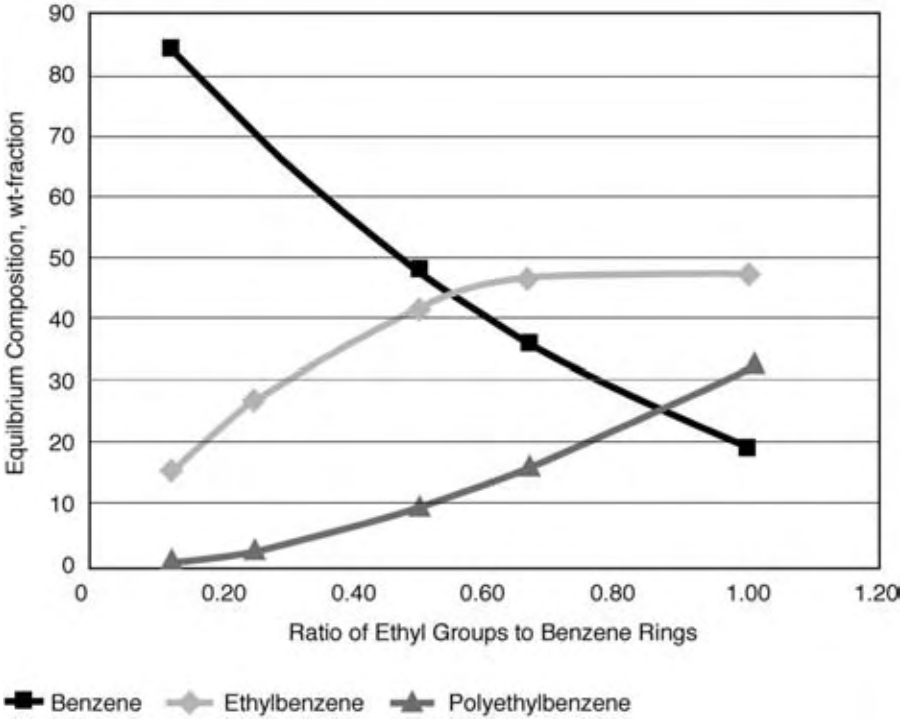


Fig. 1 Ethylbenzene equilibrium composition.

Other reactions, such as oligomerization, cracking, and isomerization, can also occur resulting in the formation of compounds such as cumene, butylbenzene, xylene, diphenylethane, and other high-boiling compounds. The formation of these by-products is impacted by the alkylation reaction conditions, in particular whether the reactions occur in the vapor phase or the liquid phase. For example, EB isomerization to xylene typically only occurs under vapor phase reaction conditions where reaction temperatures are relatively high. Isomerization does not occur to a great extent in the liquid phase because of the lower operating temperatures. The formation of by-products is also affected by the type of catalyst.

Acid catalysts are used to promote the alkylation of ethylene to benzene. Acid catalysts suitable for benzene alkylation include protonic acids (i.e., H_2SO_4, HF, and H_3PO_4), Friedel–Crafts catalysts (i.e., $AlCl_3$ and BF_3), and more recently, solid acid catalysts. Solid acid catalysts used for the commercial manufacture of EB are typically zeolitic molecular sieves and materials such as ZSM-5, faujasite, MCM-22, and zeolite beta.[3] Zeolites' physical and chemical properties can be modified to optimize the activity, selectivity, and stability of the catalysts. This flexibility of zeolites has made them the preferred catalyst of choice.

Many zeolites occur naturally as minerals. Some of these are natrolite, chabazite, sodalite, faujasite, and mordenite. Several of these naturally occurring zeolites can be produced synthetically, which makes them suitable for commercial application. In particular, faujasite-type structures zeolite X and Y have been broadly used in the petrochemical and chemical industries. A large number of new zeolite materials have been discovered and developed that cannot be found in nature. These specialty synthetic materials include MCM-22 and zeolite beta. The use of these synthetic zeolites has enabled the production of EB to become the highly efficient process it is today.[4]

Because the liquid phase process is predominantly used for new EB plants, the critical operating and design parameters for liquid phase benzene alkylation are discussed below.

Alkylation Benzene-to-Ethylene Ratio

The benzene-to-ethylene molar ratio (B/E) for the alkylation reaction section is the most important parameter for design and operation of an EB plant. A high B/E is beneficial from an equilibrium and catalyst selectivity standpoint, but the large molar excess of benzene relative to ethylene requires substantial energy to recover and recycle. The first liquid phase plants were designed and operated with B/E equal to 6 or greater. Over time, as improved catalysts were developed, the B/E has steadily decreased. In 2004, units are typically designed with B/E in the range of approximately 3.0–3.5. This substantial decrease in B/E has resulted in significantly lower capital costs, because nearly three-fourths of a plant's equipment cost is associated with recovery and recycle of

the excess benzene used in the reactors. Generally, a commercial plant is operated at or very close to the design B/E and is not frequently adjusted. With further improvements in the near future, commercial units will operate at even lower B/E ratios with B/E ratios possibly approaching 1.5 being made possible.

Alkylation Reaction Temperature

For liquid phase alkylation, the reaction temperature is normally in the range of 170–270°C to achieve an acceptable reaction rate using typical commercial catalysts. Excessive temperatures can increase by-product formation, so reactor design temperatures are often set with this in mind. Temperature also affects the reactor operating pressure and hence the equipment cost. Varying temperature within a relatively small range has little impact on the alkylation reaction overall, so generally, a commercial scale plant is maintained at a constant temperature near the design value throughout its operation cycle.

Reactor design plays a significant role in temperature control. Multiple ethylene injection points and heat removal stages are incorporated into the reactor section design to allow reaction temperature to be maintained in the desired range. A common design for reactor sections contains four ethylene injections with one heat removal stage. Another design option uses six ethylene injections with two heat removal stages. Other configurations have been used commercially.

Alkylation Reaction Pressure

Reaction pressure is sufficiently high so as to prevent any components from vaporizing in the alkylation reactor section. Alkylation reactors are typically operated at about 35–40 bar to maintain the reactor catalyst outlet streams in the liquid phase even at the maximum operating temperature. The ethylene injected into the reactor dissolves into the liquid hydrocarbon mixture such that the alkylation catalyst beds are always in the liquid phase. Reaction pressure is not normally varied during operation.

Transalkylation Benzene-to-Polyethylbenzene Ratio

To obtain an economically viable PEB conversion in the transalkylation reactor, a molar excess of benzene relative to PEB is needed. A high benzene-to-PEB ratio (Bz/PEB) results in high-equilibrium PEB conversion, but at the expense of increased capital cost and operating cost associated with the recovery and recycle of the excess benzene. As the Bz/PEB is decreased, these capital and operating costs decrease, but the PEB conversion level declines and formation of heavy by-products increases.

Similar to the B/E in the alkylation reactor section, the first liquid phase plants were designed and operated with Bz/PEB close to 10 or higher. Over time, transalkylation catalyst system stability has been improved and the Bz/PEB has steadily decreased and in 2004 plants are typically designed with Bz/PEB in the range of 2.0–4.0. Generally, a commercial plant is operated at or very close to its design Bz/PEB and the operating Bz/PEB is not frequently adjusted.

Transalkylation Reaction Temperature

The reaction temperature is the key variable for controlling the operation and performance of the transalkylation reactor. Transalkylation reactors are designed to operate across a relatively wide temperature range. During initial operation when catalyst activity is high, relatively low reaction temperatures are sufficient to obtain the desired conversion of polyalkylated compounds. As the catalyst ages and loses activity, the temperature is increased to maintain PEB conversion at or near the desired level. Liquid phase transalkylation reactors typically operate between 170°C and 270°C.

Catalyst Poisons

Because of the acidic nature of the zeolite catalysts used in the production of EB, a number of materials can interact with the zeolite and negatively impact its performance. These compounds are referred to as catalyst poisons and mostly impact the catalyst activity, although selectivity can also be affected.

Any basic or alkaline material can react with a zeolite to effectively neutralize the acidic active sites, which generally results in irreversible loss of catalyst activity. Basic compounds found in the ethylene or benzene feedstocks can include amines, amides, nitriles, and trace metal cations such as sodium and potassium. Of particular concern are nitrogen-containing organic compounds typically present in the benzene feed.

There are different types of nitrogen compounds that have been identified as benzene feed contaminants. The most common ones are *N*-formyl morpholine (NFM), *N*-methyl pyrolidone (NMP), morpholine, monoethanol amine (MEA), diethanol amine (DEA), and acrylonitrile (ACN). Both NFM and NMP are common aromatic extraction solvents that are used to

purify benzene. They are used commonly in Europe and less frequently in the rest of the world. Morpholine is a decomposition product of NFM and has also been identified in various benzene feeds. Corrosion inhibitors, including MEA and DEA and other similar compounds, are used in benzene recovery columns and can be found in the feed to an EB plant. Transoceanic shipment containers sometimes alternate loads between benzene and ACN, which can lead to nitrile contamination. In addition to these most common compounds, other basic nitrogen compounds can be present in the feed depending on its origin, processing, and handling.

Feed Treatment to Remove Catalyst Poisons

Even trace quantities of catalyst poisons can lead to significant catalyst deactivation and have a large impact on commercial production. There are varieties of adsorbent materials that are capable of removing basic nitrogen compounds from the feed streams. The type of nitrogen species in the feed is an issue that affects the choice of material as well as the guard bed design. Additionally, it is necessary to consider both the equilibrium capacity of these materials and their mass transfer properties when selecting a guard bed material and design.

The most common guard bed materials are acidified resin, clay, and zeolite. The choice of optimum guard bed material is generally a function of several variables, including nitrogen adsorption capacity, mass transfer properties, disposal methods, regenerability, and cost. These variables also influence the design of the guard bed.

COMMERCIAL PRODUCTION

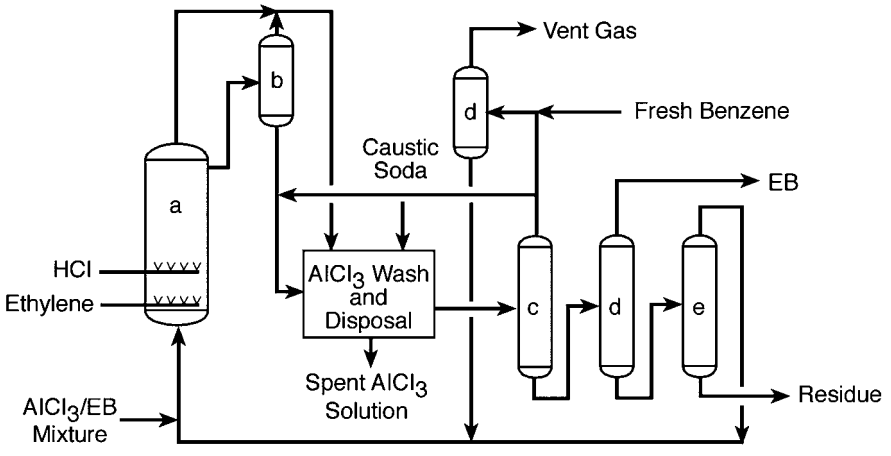
Liquid Phase Aluminum Chloride Catalyst Process

The primary means of producing EB from the 1930s to about 1980 was the liquid phase aluminum chloride catalyst process. Although the aluminum chloride catalyst process is still in use at many plants, its share of worldwide EB production is diminishing as all new plants use a solid acid catalyst.

A flow diagram of a typical aluminum chloride catalyzed EB plant, based on the widely used Monsanto/Lummus technology, is shown in Fig. 2. The Monsanto/Lummus technology is used in the majority of aluminum chloride catalyst plants because it significantly reduces aluminum chloride catalyst consumption by operating at higher reaction temperatures than competing processes.

In the latest version of aluminum chloride plant designs, the alkylation reactions occur in a homogenous liquid phase at 160–180°C. The conditions of the alkylation reactor prohibit the recycle of PEB to the reactor. As a result, these plants have a separate transalkylation reaction zone. The recycle PEB stream is mixed with the alkylation reactor effluent and fed to the transalkylation reaction zone. The aluminum chloride present in the alkylation reactor effluent catalyzes the transalkylation reactions.

The effluent stream from the alkylation–transalkylation reaction section is cooled, washed, and neutralized to remove and recover the $AlCl_3$ catalyst. The washed hydrocarbon stream contains unconverted benzene, EB, PEB, and other minor reaction byproducts. It is separated into product and recycle



a. Reactor
b. Flash Drum
c. Benzene Column
d. EB Column
e. PEB Column
f . Drying Column

Fig. 2 Monsanto/Lummus aluminum chloride catalyzed EB process.

streams by fractionation in a series of three distillation columns. The first column recovers unconverted benzene in the overhead stream, which is dried in a drying column before being recycled to the alkylation reactor section. The second column separates the product EB as the overhead stream. The last column recovers PEB from the high-boiling, heavy by-product tar components. The PEB is recycled to the transalkylation reaction section.

The handling and disposal of the aluminum chloride catalyst and waste has become increasingly more costly and complicated because of environmental considerations. Equipment and piping corrosion and fouling along with the related environmental issues led to the development of EB processes based on solid acid, heterogenous catalysts. These are the main reasons why new plants are not based on the Friedel–Craft-type catalysts. Major equipment pieces needed to be replaced on a regular schedule because of corrosion. This resulted in extensive turnarounds, poor plant onstream efficiency, and thus, are primary contributors to the high operating costs associated with the aluminum chloride process.

Vapor Phase Zeolite Catalyst Process

The first commercial plant based on the Mobil/Badger vapor phase technology was commissioned in 1980.[5] From 1980 until the early 1990s, use of the vapor phase process gained in popularity because it offered several advantages over the aluminum chloride process. A major benefit of the vapor phase process was the use of a zeolite catalyst that eliminated the issues associated with corrosion and waste disposal of aluminum chloride.

The alkylation of benzene is performed in a vapor phase, fixed-bed reactor using a ZSM-5 based catalyst. ZSM-5 is an aluminosilicate zeolite with a high silica and low aluminum content. ZSM-5, a highly porous material, is considered a medium-pore zeolite with two types of pores, both formed by 10-membered oxygen rings. The first type of pore is straight and elliptical in cross section and the second type of pore is circular in cross section and intersects the straight pores at right angles in a zigzag pattern. Therefore, throughout its crystalline structure, ZSM-5 has an intersecting two-dimensional pore structure.

The original vapor phase design accomplished the alkylation and transalkylation reactions in a single reactor. Subsequent designs performed the transalkylation reactions in a vapor phase, secondary reactor that was separate from the alkylation reactor. Fig. 3 shows a flow diagram for the latest publicly disclosed version of the process, sometimes referred to as the third-generation EB process. The alkylation and transalkylation reaction section consists of a fired reactor feed heater, a multibed alkylation reactor, a transalkylation reactor, and various heat exchange equipment. Both the alkylation and the transalkylation reactors are vapor phase, operating at temperatures in the range of 370–420°C and pressure in the range of 0.69–2.76 MPa.

The third-generation process is capable of achieving an EB yield greater than 99%. However, the high-temperature vapor phase operation of the reactors is not trouble free. The significant extent of the isomerization reactions and the catalyst deactivation by deposition of carbonaceous material are the most important problems associated with the high temperature. Any xylene formed because of isomerization cannot be separated and therefore ends up contaminating the EB product. While the xylene impurity is not a significant problem in the vapor phase EB plant, it is not desired because it results in higher operating cost in the downstream styrene plant. Catalyst deactivation occurs at a rate that requires periodic catalyst regeneration. The length of time between regeneration can vary from as little as 2 mo to slightly more than 1 yr depending on the specific plant design and operating conditions. Because the reactors must be taken off-line for regeneration, the onstream efficiency can be low, resulting in high operating costs for a vapor phase plant. Additional equipment may be required for the regeneration procedure, depending on the specific plant design, which adds capital cost to the plant.

The Mobil/Badger vapor phase process includes four distillation columns. The first major separation is in a benzene recovery column where unconverted benzene is recovered as an overhead product for recycle to the alkylation and transalkylation reactors. The bottom stream is fed to an EB recovery column where EB product is separated from cumene, the PEB, and other heavy components. The cumene, PEB, and other heavy by-products are further separated in the PEB recovery column. The heavy residue is typically used as fuel in the reactor feed heater. The PEB fraction is recovered in the overhead stream and recycled to the transalkylation reactor where it reacts to form additional EB. A fourth column is used as a stabilizer column to vent any light components and to remove water from the system.

Liquid Phase Zeolite Catalyst Processes

EB*One*™ process offered by Lummus/UOP

One of the shortcomings of the vapor phase zeolitic EB process is the occurrence of side reactions that can lead to high levels of contaminants in the EB product.

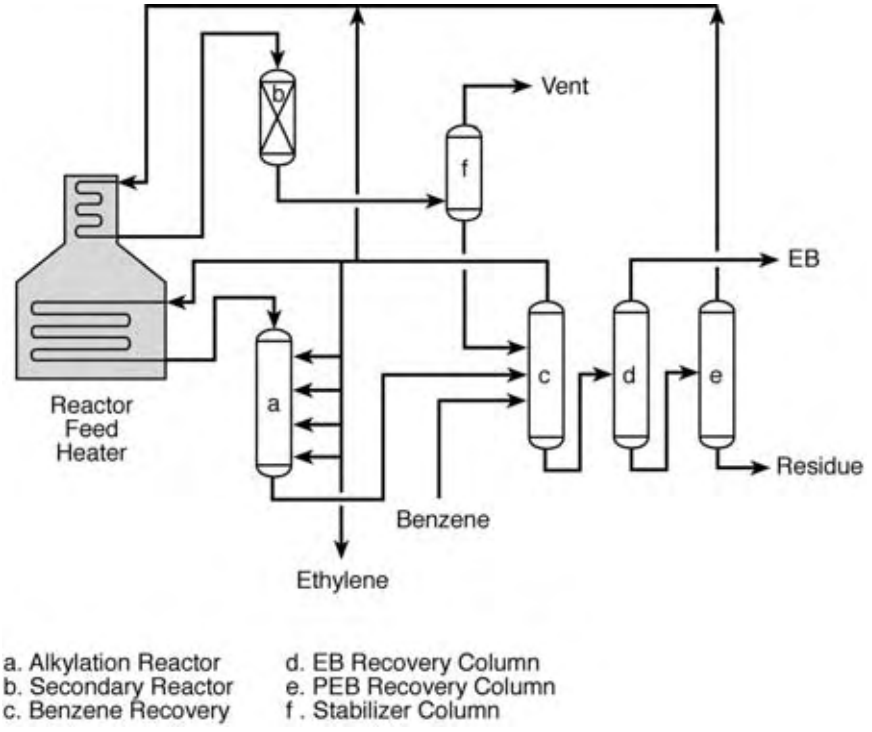


Fig. 3 Mobil/Badger vapor phase EB process.

The commercialization of liquid phase processes, which operate at substantially lower temperatures, decreased the side reactions dramatically, resulting in ultra-high-purity EB product. This improvement alleviated problems previously encountered in the downstream styrene plant. The first liquid phase commercial plant based on the Lummus/UOP process was commissioned in 1990. Since then, more than 25 projects have been licensed with more than 17 plants in commercial operation as of 2004. The liquid phase plants typically achieve high onstream efficiency, often greater than 99%, which results in low turnaround and maintenance costs. This technology is now licensed by UOP LLC and ABB Lummus Global.

The first liquid phase plants used a zeolite Y based catalyst for both the alkylation and the transalkylation reactions. A significant improvement in the process occurred in the mid-1990s when EBZ-500^{TM} catalyst was developed and put into commercial operation.[6] EBZ-500 catalyst is based on zeolite beta, which has unique characteristics that make it highly suitable for benzene alkylation. Zeolite beta has a tetragonal crystal structure with straight 12-membered ring channels with crossed 10-membered ring channels. This crystal arrangement gives zeolite beta a unique three-dimensional structure that results in high catalyst activity, an important feature in the relatively low-temperature liquid phase process. Furthermore, benzene alkylation in the liquid phase is typically limited by diffusion, so zeolite beta with its relatively large pore dimensions is well suited for the application.

A typical EB*One* plant flow diagram is shown in Fig. 4. The alkylation reactor is maintained in the liquid phase and uses multiple catalyst beds and ethylene injections to improve the reaction selectivity. Dividing the ethylene into multiple feed streams keeps the alkylation catalyst deactivation rate very low. In some plants using EBZ-500 catalyst, operating lengths of more than 8 yr have been obtained without catalyst regeneration. The ethylene conversion is essentially 100% in the alkylation reactors, and the reactors operate nearly adiabatically. The exothermic heat of reaction is recovered and used within the process to heat internal process streams or to generate steam.

In the few instances when EBZ-500 catalyst has been regenerated, it has been restored to essentially the same activity and selectivity as fresh catalyst. The regeneration is a mild carbon burn procedure that is relatively inexpensive. If required, in situ regeneration equipment can be incorporated into the design. This is not common and is usually considered only for locations were ex situ regeneration facilities are not readily accessible.

The transalkylation reactor is also maintained in the liquid phase but uses EBZ-100^{TM} catalyst, which is made using zeolite Y. Transalkylation reaction is nearly thermo-neutral, so it operates essentially isothermally. The reactor temperature is generally adjusted to provide the desired level of PEB conversion. While a high temperature results in high PEB conversion that closely approaches equilibrium composition, these conditions can result in undesired side reactions.

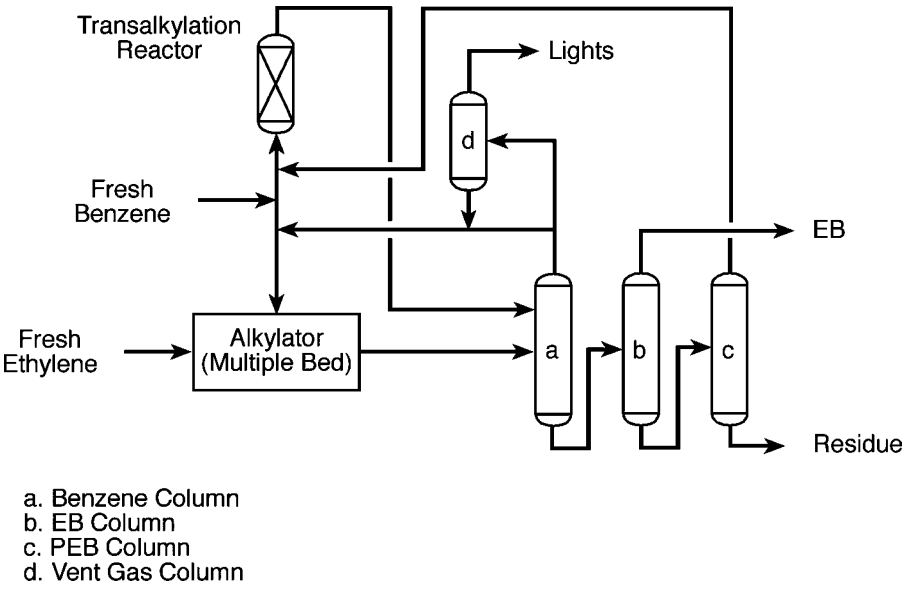


Fig. 4 Lummus/UOP's EB*One*™ process.

Deactivation of EBZ-100 catalyst is rare, usually only occurring because of unusual upsets or operation of the transalkylation reactor. Plants have operated for approximately 10 yr without regenerating the transalkylation catalyst. If EBZ-100 catalyst requires regeneration, an inexpensive, mild carbon burn procedure is used.

The alkylation and transalkylation reactor effluent streams are sent to the distillation section, which consists primarily of three fractionation columns. The first column is the benzene column. It separates unconverted benzene into the overhead stream for recycle to the reactors. The benzene column bottom stream is the feed to the EB column. The EB column recovers the EB product in an overhead stream at purities as high as 99.98 wt%. The bottom stream of the EB column feeds a relatively small PEB column where PEB is fractionated overhead and recycled to the transalkylation reactor. The bottom stream of the PEB column, referred to as flux oil, is generally used as fuel in an integrated styrene complex.

Ethylbenzene yields greater than 99.5% can be achieved by the Lummus/UOP technology.

EBMax™ process offered by Mobil/Badger

The EBMax process offered by Mobil/Badger is a liquid phase alkylation reaction using a catalyst based

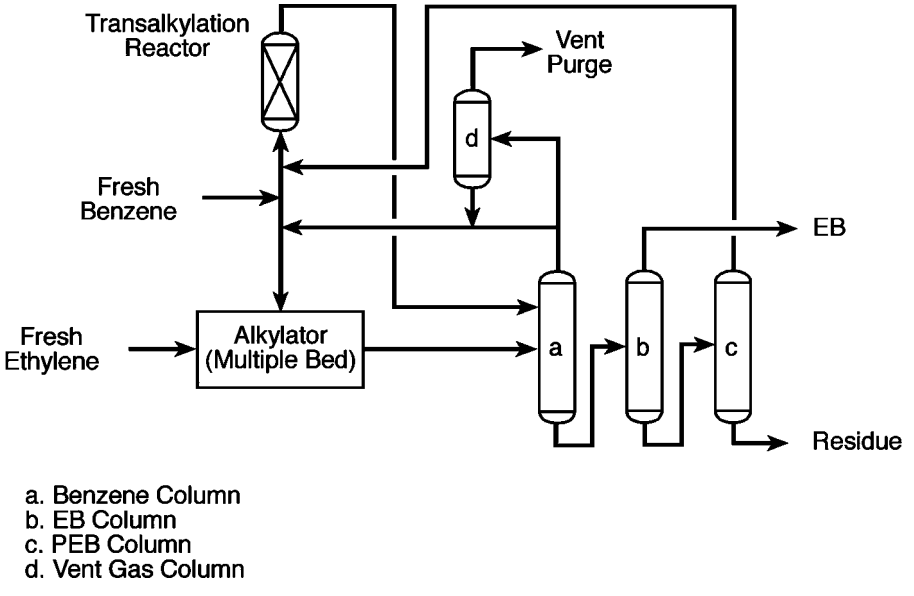


Fig. 5 Mobil/Badger's EBMax™ process.

on MCM-22. A commercial plant based on the EBMax technology was commissioned in 1995 at Chiba Styrene Monomer Company.[7]

MCM-22 is classified as a medium-pore zeolite consisting of two independent, nonintersecting, 10-membered ring channels. One of the channels contains "super cages'' that have a diameter defined by 12-membered rings. The MCM-22 crystal surface is covered with 12-membered ring pockets with each pocket being half of a "super cage.'' It is within these surface pockets that the alkylation reactions are thought to occur.

A typical EBMax plant flow diagram is shown in Fig. 5. The alkylation reactor is maintained in the liquid phase and uses multiple catalyst beds and ethylene injections. The ethylene conversion is essentially 100% in the alkylation reactors, and the reactors operate nearly adiabatically. The exothermic heat of reaction is recovered and used to generate steam, heat reactor feed streams, or as heat duty in the distillation columns.

The transalkylation reactor in an EBMax plant can be either vapor phase or liquid phase. More recently, the transalkylation reactor has been designed as liquid phase because of its improved energy efficiency. The transalkylation reaction is conducted in the liquid phase using Mobil TRANS-4^{TM} catalyst.

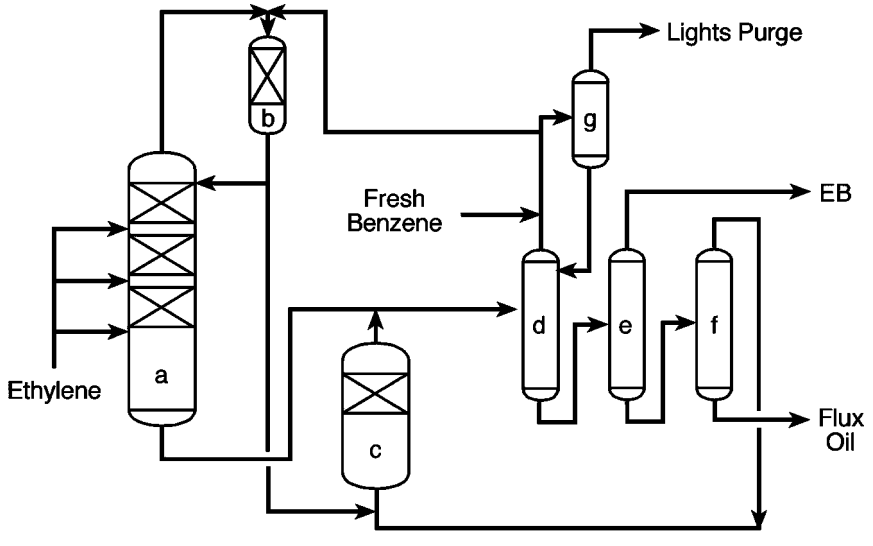
The alkylation and transalkylation reactor effluent streams are sent to the distillation section, which consists primarily of three fractionation columns. The first column is a benzene column and it separates unconverted benzene into the overhead stream for recycle to the reactors. The benzene column bottom stream feeds the EB column. The EB column recovers the EB product in the overhead stream, and the bottom stream of the EB column feeds the PEB column where PEB is fractionated overhead and recycled to the transalkylation reactor. The bottom stream of the PEB column is removed as a residue stream and is generally used as fuel in an integrated styrene complex.

Mixed Liquid–Vapor Phase Zeolite Catalyst Process

The CDTECH EB^{TM} process is based on a mixed liquid–vapor phase alkylation reactor section. The design of a commercial plant is similar to the liquid phase technologies except for the design of the alkylation reactor, which combines catalytic reaction with distillation into a single operation.[8]

Theoretically, catalytic distillation can overcome limitations in a typical two-step process consisting of reaction followed by distillation or separation. Often, a two-step process is limited by chemical equilibrium, heat transfer, mass transfer, or some combination of these. Catalytic distillation can overcome many of these constraints by simultaneously separating products from reactants, maintaining nearly isothermal operation and lowering the external ratio of reaction diluents.

The CDTECH alkylation reactor consists of two main sections—a catalytic distillation section and a standard distillation section—as shown in Fig. 6. Benzene is fed to the top of the alkylation reactor and ethylene is fed as a vapor below the catalytic distillation section, creating a countercurrent flow of



a. Alkylator
b. Finishing Reactor
c. Transalkylator
d. Benzene Column
e. EB Column
f. PEB Column
g. Lights Column

Fig. 6 CDTECH EB^{TM} process.

the alkylation reactants through the catalytic distillation section. Throughout the catalytic distillation section, a vapor–liquid equilibrium is established with ethylene largely concentrated in the vapor phase. The ethylene that dissolves into the liquid phase rapidly alkylates benzene on the catalyst active sites to produce EB. The rapid reaction of ethylene in the liquid phase creates a driving force for additional ethylene to dissolve into the liquid phase where the alkylation reaction occurs on the catalyst active sites. The exothermic heat of reaction creates the vaporization necessary to effect the distillation. The alkylation reaction products, mainly EB, diethylbenzene, and smaller amounts of other by-products, are continuously fractionated and removed from the catalytic distillation section. In the lower section of the alkylation reactor, standard distillation occurs and the reactor bottom stream contains primarily EB, PEB, and other high-boiling by-products.

The catalytic distillation section uses a zeolite catalyst that is packaged into specially engineered bales. The catalyst bales function similarly as typical column structured packing and are designed to optimize both the distillation and the chemical reaction processes that occur in this portion of the alkylator.[9]

The alkylator typically does not achieve 100% conversion of the ethylene, so the overhead stream from the alkylation reactor contains some unconverted ethylene and benzene. This overhead stream is fed to a finishing reactor were the unconverted ethylene is fully reacted. The finishing reactor is a fixed-bed reactor that operates in the liquid phase.

One particular advantage of the CDTECH process is the ability of the alkylation reactor to accept a dilute ethylene feed. Because the alkylator operates in a mixed vapor–liquid phase, it is capable of utilizing dilute ethylene feeds, for example, offgas from a fluid catalytic cracking plant or dilute ethylene from a steam cracker plant. In general, ethylene feed streams containing significant amounts of hydrogen, methane, or ethane do require some pretreatment and cannot be used directly in the straight liquid phase technologies.

ECONOMICS

Although there are several different commercial technologies in use; the economic information described below relates only to the liquid phase technology. The cost of EB production consists of three main components: raw materials, utilities, and the fixed cost associated with the plant. The cost of utilities includes fuel, electricity, steam, cooling water, catalyst, and chemical costs required to operate the plant. Ethylbenzene plants typically have a small net negative utilities cost because the credit value of steam generated usually exceeds the cost of other utilities used throughout the plant.

The major cost components for EB production using the liquid phase process are listed in Table 3. The major cost of production is the cost of ethylene and benzene raw materials, which accounts for more than 95% of the total cost of production because of the extremely high product yield of the commercial processes. As seen in Table 3, more than 95% of the total cost of production comes from the raw material costs of the ethylene and benzene feedstocks. The remaining cost, less than 5%, comes from fixed and utilities costs. The utilities costs are zero or slightly negative because all heat input and the heat of the alkylation reaction is recovered as low-pressure steam, which is valuable for a downstream styrene plant. The efficiency of this liquid phase process delivers extremely high commercial yields. The benzene cost is the largest cost component, so the economics of EB production is highly dependent on benzene price.

Table 3 Typical economics for an EB liquid phase process[a]

	Unit	Quantity Unit/MT	Price $/Unit	Cost $/MT[b]
Product				
Ethylbenzene	MT	1.0000	530	530.0
Raw materials				
Ethylene	MT	0.2653	629	166.9
Benzene	MT	0.7387	453	334.6
By-product credits				
Flux Oil	MT	0.0030	125	(0.4)
Net feedstock cost				501.1
Utilities cost				(0.6)
Fixed cost				12
Total cost of production				512.5

[a]North America, 2003.
[b]MT, metric ton.

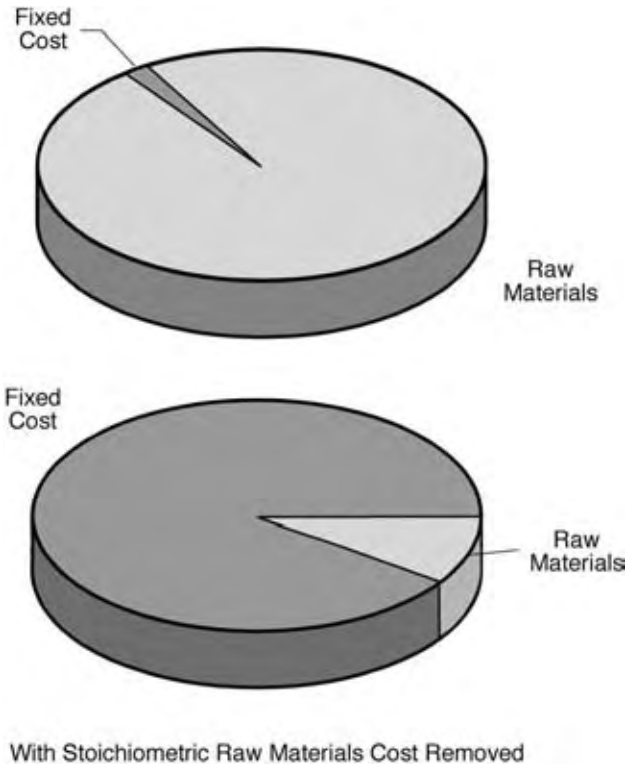


Fig. 7 Distribution of EB cost components.

The raw material cost has two components—one dictated by the stoichiometry and one caused by yield losses occurring as a result of the process technology. If the unalterable stoichiometric raw material consumption is removed from the cost of production, the resultant distribution of cost components appears very different, as illustrated in Fig. 7. From this perspective, the raw material cost is only about 10% of the incremental cost of production and the fixed costs become dominant. Recent catalyst and process design improvements have reduced the variable costs of EB production, while ever-increasing plant complexity and more stringent environmental regulations have greatly increased the fixed costs. Other recent trends, such as globalization of the EB–styrene market, have also resulted in higher fixed costs.

The result of the shift in focus from variable costs to fixed costs is that plants are being designed for larger capacities. For example, in 2003, typical new EB plants in the Asia Pacific Region produced an average of 368 KMTA EB per year, nearly double the capacity of typical plants started up just 5 yr earlier.

CONCLUSIONS

Since its first commercial production in the 1930s, EB, mainly through its role as an intermediate in the production of polystyrene, has become an important feedstock for products that are used in everyday life. Most people come in contact with numerous products produced from styrene throughout the course of a normal day. Because of its close link with styrene production, the demand for EB is expected to continue growing at a rate comparable to the demand growth rate of styrene, which is nearly equal to the gross domestic product (GDP) growth rate.

The chemical processing technologies that have been developed are sophisticated and produce EB to meet that demand at the lowest possible cost. Research and development aimed at discovering further improvements in existing technologies and identifying new technologies for EB production remains an area of great focus with strong potential for application in the marketplace.

REFERENCES

1. Lentz, R.; Gami, A. Recent EB capacity expansion; creative approaches to revamp projects. *Styrene Conference General Session*; Prague, Czech Republic, Jun 22–25, 2003.
2. CMAI Capacity Database Report for World Ethylbenzene Capacity; http://www.cmaiglobal.com (accessed May 2004).
3. Sherman, J. Synthetic zeolites and other microporous oxide molecular sieves. National Academy of Sciences Colloquium "Geology, Mineralogy, and Human Welfare,'' Irvine, CA, Nov 8–9, 1998; Proc. Natl. Acad. Sci. USA **1999**, *96*, 3471–3478.
4. Schmidt, R.J.; Zarchy, A.S.; Peterson, G.A. New developments in cumene and ethylbenzene alkylation. *AIChE Spring Meeting*; New Orleans, LA, Mar 10–14, 2002.
5. Degnan, T.F. Jr.; Smith, C.M.; Venkay, C.R. Alkylation of aromatics with ethylene and propylene: recent developments in commercial processes. Appl. Catal. **2001**, *221*, 283–294.
6. Woodle, G.B.; Zarchy, A.S.; Morita, M.; Shinohara, K. Leading-edge ethylbenzene production, Lummus/UOP liquid phase EB process. *1998 International Styrene Symposium*; Sappora, Japan, Jun 14–18,1998.
7. Bhandarkar, M.; Lewis, P.J.; Dandekar, A.B.; Venkat, C.R.; Degnan, T.F. EBMax: liquid phase alkylation of benzene with ethylene over zeolite MCM-22. In *European Petrochemicals Technology Conference*; London, U.K. Jun 21–22, 1999.
8. Phanse, G. Catalytic distillation: the next step in aromatic alkylation. *AIChE Spring Meeting*; New Orleans, LA, Mar 10–14, 2002.
9. Smith L.A., Jr. Alkylation of Organic Aromatic Compounds, US. Patent 5,446,223, Aug 29, 1995.

Fermentation Processes

Kimberly Ogden
Department of Chemical and Environmental Engineering, University of Arizona, Tucson, Arizona, U.S.A.

INTRODUCTION

Fermentation technology started with wine and brewing industries. Throughout the centuries, methodology continuously improved and the technology was transferred to the pharmaceutical industry for the production of antibiotics in the 1930s. Throughout the 1980s and 1990s there was a huge increase in the number of products manufactured from live cells. The two major scientific discoveries that contributed to the industrial expansion were recombinant DNA techniques and hybridoma cell technology. Molecular biologists discovered how to isolate, purify, modify, and reinsert DNA from one species to another and produce significant amounts of a protein product in the host cell. Hybridoma cells are a cross between a cancer cell and a mammalian cell that allows the cell line to reproduce longer and produce mammalian enzymes and antibodies.

CELL TYPES—CHOOSING A HOST/VECTOR SYSTEM

All fermentation or bioprocesses start with the selection of the cell type. The choice is dependent on the product synthesized, the amount of product needed, the purity of the required product, and of course the cost. A biopharmaceutical such as an antibiotic, antibody, or protein that is directly injected into a human requires a highly pure product, whereas food products require intermediate purity, while commodity products such as ethanol produced for alternative fuel usage do not require as stringent sterile methodology. The advantages and disadvantages of the major classes or cell types are summarized in Table 1.

Bacterial cells have been studied for many decades and are well understood. The most common strain used for early recombinant DNA work was *Escherichia coli*. It is still used frequently today for laboratory studies. The common gram-positive strain is *Bacillus subitilis*. Other bacterial cells are studied to determine their potential to destroy environmental pollutants or to synthesize specialty chemicals. Bacteria are not extremely complex, which is advantageous from a DNA manipulation standpoint, but a challenge in terms of the organism's ability to synthesize a human protein product. At the other end of the spectrum are mammalian cells. These cells make proteins, antibodies, and antigens that are very similar to humans, but the cells grow slowly, require expensive media, and are typically mortal.

In addition to bacteria and mammalian cells, yeast, plant cells, and insect cells have been studied for synthesis of bioproducts. The yeast strain, *Saccharomyces cerevisiae*, is used in the food industry and the genetics are reasonably well understood; however, it tends to express only low levels of a foreign protein and hyperglycosylates the product. Plants themselves offer the advantage of diversity. As many as 25% of today's pharmaceuticals (primarily nonprotein products) are extracted from plants. Plant cell cultures do allow for more control than using an intact plant; however, genetic knowledge is less than what is known for bacterial and animal cells, and product expression levels are low. Only a few plant cell systems are used commercially in Japan and Germany. The most well-known product made in plant cell culture is the anticancer agent, paclitaxel or Taxol. The insect cell–baculovirus system is used primarily as a research tool and for small-scale studies (100 L). The advantages of this system are that it does allow for high expression of foreign protein and offers potential safety advantages because viruses that affect insect cells do not affect humans. However, the insect system does not quite mimic the mammalian cell system, in that the protein product may have slight structural modifications, which are useful when making vaccines but not necessarily so for complex products.

A technology that has made great strides in the past decade is the use of transgenic animals and plants for production of proteins. The living animal or plant becomes the ''bioreactor.'' The use of transgenic animals is less developed than that of plants. Animals have the advantage of performing complex posttranslational processing steps that cannot be done in animal cell culture. New genetic information is inserted into the embryo of the animal and the nontoxic protein is expressed by the mature animal—typically in the milk. The primary concern is safety. Besides the possibility of having a negative health effect on the animal, there is fear regarding the mutation of animal viruses that jump species and become a serious health issue

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120018064

Table 1 Comparison of cell host–vector systems for product synthesis

Cell type	Advantages	Disadvantages
Escherichia coli	High growth rate Wide range of host backgrounds Wide range of vectors and promoters High expression levels and cell densities Low-cost media	Formation of inclusion bodies Misfolded protein products No ability to glycosylate Proteolytic activity Usually low levels of product excretion No posttranslational processing Must ensure removal of endotoxins during purification
Bacillus subitilis	Second best studied bacteria Gram positive—no outer membrane—excretes proteins	Problem excreting some foreign proteins Makes large amounts and varieties of proteases (degrades proteins rapidly) More difficult to manipulate genetically owing to limited vectors and promoters Instability of plasmids more problematic than *E. coli*
Other bacteria	Typically studied or used to produce one product of interest. One example is *Zymmonas mobilis* for producing ethanol from glucose	Genetics less understood than *E. coli*
Yeast Most common—*Saccharomyces cerevisiae*	Food and beverage industry Larger than bacteria, harvested easier Grow only 25% slower than bacteria to high cell densities Glycosylation of proteins (add sugars) Secrete proteins Easier for regulatory approval	Difficulty with high protein expression Difficulty with good excretion Range of genetic systems is limited Inefficient if extensive posttranslational processing is needed
Fungi	Better secretion than yeast Natural producer of antibiotics	Filamentous growth difficulty for large-scale cultivation (but problem has been addressed) Normally produce low levels of desired protein (except for penicillin studied in detail) Inefficient if complex posttranslational modification needed
Plant cells	Used to make pharmaceuticals currently derived from plants (25% of all pharmaceuticals currently extracted from plants) Food flavors, fragrances, insecticides dyes Taxol™—anticancer drug	Slow growth Plant genetics poorly understood Products are often not proteins and chemically complex Expensive to culture so need high-value products like Taxol
Mammalian and hybridoma cells	Availability of host and vector systems Excretion of product Ability to produce authentic proteins (correct folding, glycosylation, and posttranslational processing) Low proteolytic activity	Slow growth Expensive, complex media Low protein expression levels Shear sensitive cells Vectors are derived from virus and there is fear of reversion to pathogenic form Must ensure removal of nucleic acids from hybridoma lines Most strains are mortal so dead cells must be removed
Transgenic plants Most common—corn	Easy separation of product Environment friendly Dow, Monsanto, Sigma–Aldrich, Prodigen, Epicyte Pharmaceutical developing products No virus transmission	Movement from plant to plant even though contained Fear of getting into food chain
Transgenic animals	Complex posttranslational processing not done by cell culture Sheep, goats, and pigs	Development in its infancy Regulation issues Transfer of disease/virus

for humans. The fear is real because some viruses that still lack effective treatment are believed to have originated in animals, such as AIDS and SARS. Screening of animals for virus is an extremely expensive undertaking and may limit the cost-effectiveness of their use for production.

Transgenic plant technology holds great promise. Several crops including corn, tobacco, potato, and soybean have been studied for the production of antibodies and proteins. The product, especially antibodies, concentrates in the corn kernels, and is easily separated. The systems are safe in terms of virus transmission, because plant viruses do not survive in animals. The one concern is the ability to keep the plants isolated from other crops. The pharm plant products, as they are called, cannot be allowed into the food chain. So currently, these plants are not grown in any food producing areas.

Optimization of recombinant protein production is a continuing objective of industrial and academic research. Engineering and science work together to reach this objective. For instance, a common problem is genetic instability (segregational, structural, host cell gene mutation, and/or growth rate dominated), because the overproduction of foreign proteins is always detrimental to cell growth and survival. Cells have been genetically manipulated to alleviate this problem by inserting antibiotic resistance genes into the plasmids and supplying antibiotics to the medium. Research continues to find the ''ideal'' host–vector system that rapidly grows and expresses high levels of foreign protein that are excreted while minimizing production of impurities. The ideal system also does not produce proteases and folds and glycosylates the protein product as required.

The disadvantages shown in Table 1 continue to be areas for research and development, and progress is made everyday toward overcoming these challenges. One example of the amount of research and development in the area of strain development is the number of patents found using Scifinder. In the last 20 yr, approximately 340 bacterial strain patents, 150 yeast strain patents, 1300 plant cell patents, 4300 hybridoma cell patents, and 2700 transgenic plant patents have been issued internationally. The number of journal articles published is phenomenal.

GROWTH KINETICS

Cells growth is described as an autocatalytic reaction. All growth requires an initial seed or inoculum of cells. As the cells are given nutrients, they metabolize this food to make more cells. The nutrient requirements vary considerably with cell type, but all cells undergo the growth phases shown in Fig. 1—stationary, exponential, deceleration, stationary, and death phases. During lag phase, the cells are adjusting to their environment. No cell division occurs but the cells may increase in size as they ''turn on'' metabolic processes to begin cell division. The length of this phase varies significantly from minutes to hours to days depending on both the cell type and the environment. The shortest lag phase is achieved by transferring rapidly dividing cells into the same environment from which they came.

The exponential growth phase is the time when the cells divide at a relatively constant rate. Products that are growth associated, which include many proteins and antibodies, are synthesized during this growth phase. As the cells begin to run out of a particular nutrient or if they become inhibited by too much product, their growth rate decreases and this is the deceleration phase. Cells tend to ''turn off'' some enzymatic processes associated with cell division and ''turn on'' different enzymes associated with cell maintenance. After this, the cells enter the stationary phase. This phase is defined as the phase where there is no net increase or decrease in cell mass. However, the cells may be slowly dividing and dying. Secondary metabolites such as antibiotics are produced. Some products such as ethanol may be produced during both exponential and stationary growth, but at different rates. The final phase is death phase, when the cells die. This phase is very prevalent in mammalian cell cultures that are ''mortal,'' whereas bacterial cells may remain in stationary phase for weeks with very little death.

A variety of cell measurement techniques are commonly used to monitor growth. These include:

- Direct counting using a microscope and hemacytometer
- Counting by diluting and plating the cells on solid medium
- Particle counters
- Dry cell weight

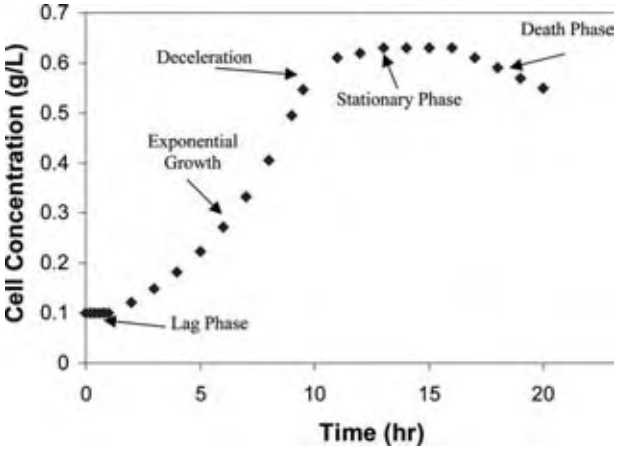


Fig. 1 Growth phases for a cell culture.

- Cell volume by centrifugation
- Turbidity or optical density—560–600 nm
- Measurement of product formed or amount of substrate consumed
- Measurement of cellular protein or DNA

This list includes both direct and indirect methods. The latter requires performing calibration curves to relate the indirect measurement to cell number. Indirect measurements are typically less time-consuming and allow for quick results; however, the danger is that something other than cells is being measured. For example, when optical density is being measured, other particles or color from nutrient components also absorb light and may affect results. Also, some indirect measurements do not differentiate between alive and dead cells. Thus, calibration curves must repeatedly be performed. Often consumption of the limiting nutrient is monitored or the amount of a product formed. Combining cell measurements and chemical measurements is the most accurate.

Various models ranging from unstructured, nonsegregated to structured, segregated are used to describe cell growth. The most common kinetic expression is the Monod equation:

$$\mu = \frac{\mu_{max} S}{K_s + S} \quad (1)$$

where μ is the specific growth rate, μ_{max} is the maximum specific growth rate, K_s is the saturation constant, and S is the substrate. This growth expression is inserted into batch, continuous, biofilm, scaffold, or immobilized reactor equations to predict the rate of utilization of substrate, cell concentration, and growth-associated product concentration. The semiempirical expression assumes that a single chemical species, S, is limiting, while changes in other nutrient concentrations have no effect; and that a single enzyme system with Michaelis–Menten kinetics is responsible for the uptake of S. Although this premise is seldom true, the Monod equation is used routinely to describe bioreactor behavior of everything from well-defined recombinant bacterial or mammalian systems to wastewater treatment systems. Growth rates are affected by temperature, pH, and media composition.

Chemically structured models provide a more general approach with greater predictive power by relating cell growth and product production not to just one substrate, but to nitrogen, carbon, and oxygen uptake, and also include expressions that relate important kinetic interactions among cellular subcomponents such as RNA, DNA, lipids, and proteins.[1] These more sophisticated models predict growth rates better, and thus reactor behavior, but consist of between 4 and 40 equations. Segregated models typically differentiate between productive and nonproductive cells within one reactor system.[2,3] The very complex models are not typically used by industry but are used in research laboratories. There are future challenges for modeling efforts.[4]

REACTOR TYPES

Fermentation processes are run in batch, fed-batch, continuous, or immobilized systems (Fig. 2). The choice of reactor is extremely important and determines the amount of product synthesized, the number of impurities, the yield, stability, and reliability. Some of the factors that must be considered when choosing the reactor system include production temperature, type of cells, genetic instability, growth associated product vs. secondary metabolite, and the amount of product needed (e.g., small quantities of a high-value medicine vs. large quantities of ethanol for fuel). Each type of reactor is discussed here and the advantages and disadvantages are given in Table 2.

A batch system (Fig. 2) is one in which all of the nutrients needed to grow a cell culture are added to a

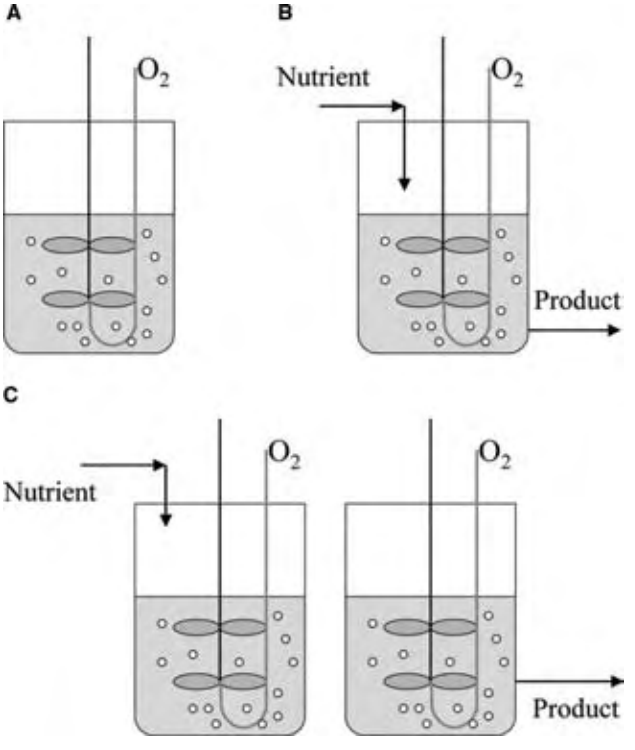


Fig. 2 Reactor types used for fermentation processes: (A) batch reactor, (B) continuous reactor or chemostat, and (C) fed-batch reactors with either a feed stream or a product withdrawal stream.

Table 2 Comparison of reactor systems for product synthesis

Reactor type	Advantages	Disadvantages
Batch	Less generations, thus less genetic instability Easier to keep sterile Flexibility in number of products produced Allows for production of secondary metabolites Used for high-value products	Can have large variability from batch to batch. Lower productivity than continuous
Fed-batch	Allows for intermittent removal of inhibitory product Allows for addition of inducers Allows for low feed rate for production of secondary metabolites Used for high-value products	Lower productivity than continuous Maintenance of sterility more difficult than batch
Continuous (CSTR or chemostat)	High productivity of growth-associated products Used for waste treatment, ethanol production, and other large-volume products Powerful experimental tool	Maintenance of sterility Difficult to produce secondary metabolites Genetic instability limits productivity High power consumption
Immobilized (batch or continuous)	High cell concentrations Cell reuse and eliminates processes of cell recovery and cell recycle Eliminates cell washout at high dilution rates Combination of high cell concentrations and high cell flow rates leads to high volumetric productivities May provide favorable microenvironmental conditions (cell–cell contact, nutrient-product gradients, pH gradients) resulting in better performance Can provide genetic stability—plasmid maintenance Can provide protection against shear damage (mammalian cells) Less moving parts	Product should be excreted by cells for continuous operation Diffusion limitation of nutrients Control of microenvironments is a challenge Gas evolution and cell growth can disrupt cell matrix Cells may be killed during immobilization Limited to low-viscosity systems Excessive foaming
Continuous with recycle	Operating at dilution rates greater than the maximum specific growth rate Increased productivity	Maintenance of sterility High energy costs
Two continuous reactors in series	Separate production and growth operations	Maintenance of sterility

large well-mixed reactor typically having temperature, pH, and other controllers attached to it. A few cells or an innoculum is added to the vessel and then after the culture has reached stationary or death phase, the cells are harvested and the product is purified and packaged for sale. Batch reactors are very common in industry. These reactors are flexible, can be used for multiple products in a given plant, and are used for both growth associated and nongrowth associated products, and genetic stability is more easily controlled.

Continuous reactors (Fig. 2) have both a continuous feed of nutrients and withdrawal of cells and product. The ideal continuous reactor is a stirred tank (CSTR) where the system is perfectly mixed and pH, temperature, and dissolved oxygen are controlled. When cells are grown continuously in a CSTR the reactor is known as a chemostat, a reactor with a constant chemical environment in which typically one nutrient is limiting while others are supplied in excess. The flow into the reactor is equal to the flow out of the reactor, and thus the volume remains constant. The term dilution rate is used to describe chemostat behavior. It is the flow rate into the reactor divided by the reactor working or liquid volume. The maximum flow rate of nutrients is limited by the growth rate of the cells. The cells have to be retained in the reactor long enough to divide, otherwise a state termed ''washout'' occurs. In other words, the dilution rate has to

be less than the growth rate of the cells to maintain a productive reactor.

A fed-batch reactor (Fig. 2) is one in which either nutrients only are fed to the reactor or products only are withdrawn. The feed rate or withdrawal rate may be discontinuous as well. One common example of a fed batch reactor is when cells are grown batch-wise until the late exponential phase, then a small amount of feed is added to the reactor continuously to provide just enough nutrients to allow for the production of secondary metabolites. These secondary products are then removed at given time intervals and purified. Other examples of fed-batch reactors include the discontinuous removal of an inhibitory end product such as ethanol from a fermentor, or the timed addition of an inducer to ''turn on'' production of a growth associated product after a cell culture has entered midexponential growth.

Immobilized systems (Fig. 3) are the ones in which the cells are confined to a space and are either chemically or physically adhered to a surface such as activated carbon or entrapped within a matrix.[5,6] Cells bind to a surface by: 1) electrostatic forces, when the surface is charged like ion-exchange resin; 2) covalent bonding, when a coupling agent like gluteraldehyde or metal oxides are used to treat the surface prior to cell attachment; 3) hydrogen bonding; and/or 4) van der Waals forces. Hydrogen bonding and van der Waals attachment form weaker bonds and thus large amounts of desorbed cells often result. When cells are entrapped, a variety of matrices are used including, but not limited to, gelatin, agar, Ca-alginate, Al-alginate, k-carrageenan, chitosan polyphosphate, polystyrene, cellulose triacetate, and collagen.

The major advantage of immobilized systems is that the cells are maintained in the reactor vessel and thus higher cell concentrations can be achieved compared to a chemostat reactor. The reactor system can be operated at dilution rates greater than the maximum specific growth rate of the culture. These reactors can also be used in batch mode if the product is not excreted by the cells.

In addition to gel-type matrices and activated carbon, which typically result in spherical support systems, immobilized systems sometimes consist of a membrane surface for cell adhesion and cells are immobilized on long fibers of tubes (Fig. 3). The reactor then resembles a shell and tube heat exchanger. The membranes are made of materials such as nylon, polystyrene, cellulose acetate, or ethyl cellulose. These membranes are semipermeable membranes. Cells are immobilized on the shell side; growth nutrients are pumped in and diffuse through the membrane while metabolic products diffuse back across the membrane, and are removed from the reactor system.

The major disadvantage of immobilized systems is diffusion limitations. As the immobilized cells grow, it becomes more difficult to transport nutrients to them within the cell matrix and for products to diffuse out of the matrix. Many times the cells contained in the inner part of the immobilized support die because of these transport limitations.

Reactors with a recycle stream have been used in the chemical industry for years to increase productivity in continuous operations and avoid wasting feedstocks. For bioreactors, recycle reactors are the most effective if the biomass is concentrated and recycled while the liquid stream is removed. The system shown in Fig. 4 contains an inclined settler that allows for sedimentation, concentration, and recycle of the cells while the product is removed from the top of the settler.[7] A continuous filter or centrifuge can also be used to concentrate the cells. Recycle reactors, like immobilized systems, allow systems to be operated at dilution rates greater than the maximum specific growth rate because of the increased cell mass. The disadvantages are the same as for chemostat reactor; the most severe being the sterility issues and increased energy costs especially if centrifuges or filters are used for cell concentration.

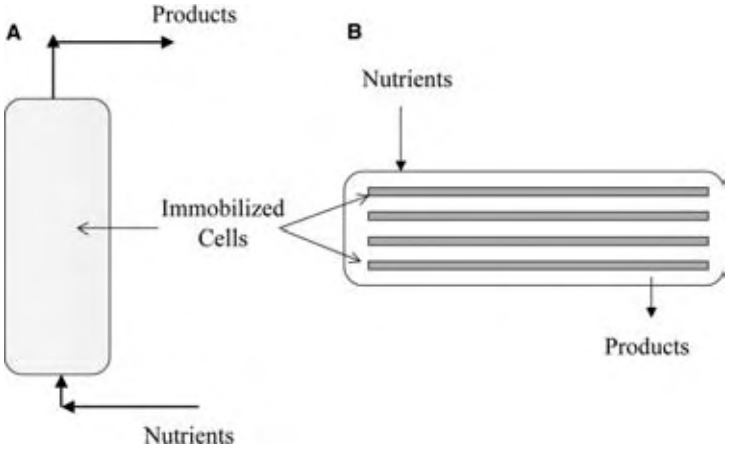


Fig. 3 Immobilized reactor systems: (A) packed bed and (B) hollow fiber reactor.

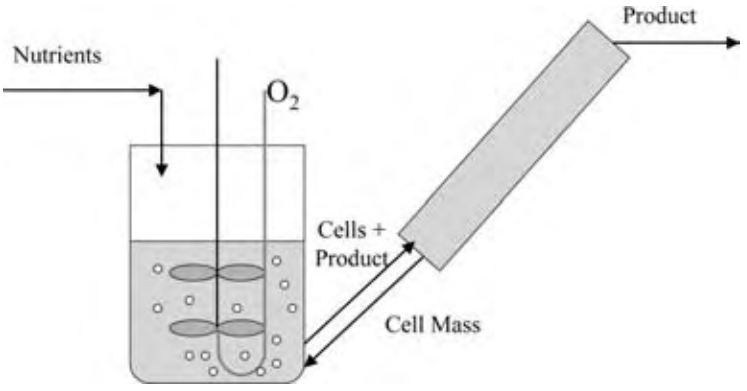


Fig. 4 Bioreactor with inclined settler to recycle concentrated biomass to the reactor.

Another type of reactor strategy that has been investigated in detail is the use of two continuous reactors in series. The advantage of this type of system is that it allows for increased productivity compared to a batch system or one CSTR. It also allows for the separation of growth and product synthesis. For example, the first reactor is run to maximize cell growth and consume the majority of the substrate, then the dilution rate in the second reactor is adjusted to allow for production of non-growth-related products or secondary metabolites.[8] The second reactor may have an additional fresh feed stream (Fig. 5) to aid in controlling the dilution rate and supply any required additional nutrients. Finally, if recombinant products are produced, the first reactor is again used for growth, and the inducer to turn on product synthesis is added to the second reactor to maximize production while minimizing the amount and, thus, cost of the inducer.

SCALE-UP

Effective scale-up is essential for successful bioprocessing. While it is desirable to keep as many operating parameters constant as possible during scale-up, the number of constant parameters realizable is limited by the degrees of freedom in designing the large-scale operation. Scale-up of aerobic fermentors is often carried out on the basis of a constant O_2 transfer coefficient, kLa, to ensure the same O_2 supply rate to support normal growth and metabolism of the desired high cell concentration populations. Propeller design, stirrer geometry, and agitation rate as functions of temperature and pressure are typically varied.[9] The challenge is greater when shear-sensitive cultures are used. In this case, pneumatically agitated airlift systems, stirred tank geometries, and hybrid bioreactor configurations combining the actions of an internal airlift draft-tube and axial-flow propeller have been studied.[10] The overall goal is to examine all aspects of bioreactor design while improving productivity and maintaining product quality. The majority of the work in this area was published in the 1980s and early 1990s.[11]

More recently, an economic study was done.[12] This study showed that multiple bioreactor approach to scale-up increases the return on investment (ROI) of the entire plant when compared to one large bioreactor to make high-value products. The increase in ROI results from the smaller size of the downstream units compared to the base case, because downstream processing accounts for about 80% of the total cost for high-value products like tissue plasmid activator.

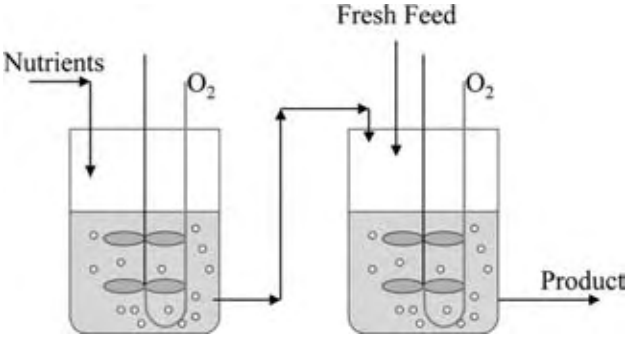


Fig. 5 Two continuous reactors in series. Typically, the first one is used for cell mass production and the second reactor is used for product synthesis.

CONTROL AND MONITORING

The measurement of analyte concentration is a critical part of successful bioreactor monitoring. Although strategies exist for measuring the majority of relevant analytes, industrial online bioreactor control is carried out primarily by measurement and control of temperature, pH, dO_2, CO_2, and, in some cases, cell density. This is because the available technology cannot be easily and inexpensively adapted to measure the analyte in an aseptic manner and measurement is not often achieved in real time so that online control is challenging. Biosensors for a variety of applications for

detecting contaminants in soil and water, and biowarfare agents are being studied extensively, but the current literature (2000 or later) does not contain too many articles or patents on bioreactor control. One book published in 1990 discusses on-line measurement and parameter estimation in bioreactors.[13] Recently, a research group is investigating the use of near-infrared (NIR) spectra in the 700–1800 nm region to obtain fermentation data on biomass, glucose, lactic acid, and acetic acid. Batch, repeated batch, and continuous fermentations were monitored and automatically controlled by interfacing the NIR to the bioreactor control unit.[14]

The long range goal continues to be to relate the metabolic activities of cell cultures to bioreactor control. Early studies focused on measuring the respiration quotient and rate of ammonia addition for pH control and combining these measurements with elemental and microscopic balances to predict behavior of batch, fed-batch, and continuous yeast cultures.[15] Cybernetic models continue to be studied with the goal of using these models to control bioreactors.[16]

CONCLUSIONS

There are many articles and textbooks available for further information about fermentation processes and bioreactor engineering.[17–19] The technology continues to be an area of study to improve production of pharmaceuticals, specialty chemicals, antibodies, and food products. Choosing the correct cell type and reactor strategy is of utmost importance. This choice also effects the downstream processing or separation sequence used to purify the product; thus, a systems approach to fermentation is extremely valuable.

Current and future areas of bioreactor research and development are the growth of tissues in biofilm or scaffold reactors for regeneration of tissues. Great strides continue to be made. For example, it is now possible for a patient's cells to be shipped to a company, cultured in a scaffold bioreactor, and subsequently reimplanted. Regenerative and rejuvenating therapies may become common in the near future. Human genes, proteins, antibodies, and cells are used in combination to replace, repair, and restore tissue that is damaged by disease, injury, or old age. Tissue cell reactors for these purposes require precise control compared to those used for the production of food or pharmaceutical products, but the principles of operation build upon the knowledge learned from the existing whole-cell bioreactors. A more controversial area of research is the use of stem cells for medicinal purposes. The ethical issues surrounding this type of work may slow down its development but these areas will be an important part of the next few decades of research.

Another area is the production of chemical intermediates from renewable feedstocks. Cargill–Dow and Dupont are just two of the companies beginning to market biobased polymers and plastics to replace petroleum based polymers. Again, the fermentation fundamentals originally developed for food manufacturing continue to apply to a wide variety of products.

REFERENCES

1. Ataai, M.M.; Shuler, M.L. Simulation of CFSTR through development of a mathematical model for anaerobic growth of *Escherichia coli* cell population. Biotech. Bioeng. **1985**, *27*, 1051–1055.
2. Bentley, W.E.; Mirjalili, N.; Andersen, D.C.; Davis, R.H.; Kompala, D. The principal factor in the metabolic burden associated with recombinant bacteria. Biotech. Bioeng. **1990**, *35*, 668–681.
3. Seo, J-H.; Bailey, J.E. Effects of recombinant plasmid content on growth properties and cloned gene product formation in Escherichia coli. Biotech. Bioeng. **1985**, *27*, 1668–1674.
4. Bailey, J.E. Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities. Biotech. Prog. **1998**, *14*, 8–22.
5. Characklis, W.G.; Marshall, K.C. Biofilms: a basis for an interdisciplinary approach. In *Biofilms*; Characklis, W.G., Marshall, K.C., Eds.; John Wiley and Sons: New York, 1990; 4–16.
6. Moo-Young, M. *Bioreactor Immobilized Enzymes and Cells: Fundamentals and Applications*; Elsevier Science Publishing, Inc.: New York, 1998.
7. Ogden, K.L.; Davis, R.H. Plasmid maintenance and protein overproduction in selective recycle reactors. Biotech. Bioeng. **1991**, *37*, 325–333.
8. De Gooijer, C.D.; Bakker, W.A.M.; Beeftink, H.H.; Tramper, J. Bioreactors in series: an overview of design procedures and practical applications. Enzyme Micro. Technol. **1996**, *18*, 202–219.
9. Bourne, J.R.; Zurita, E.P.; Heinzle, E. Bioreactor scale-up for the oxygen-sensitive culture *Bacillus subtilis*: the influence of stirrer shaft geometry Biotechnol. Prog. **1992**, *8* (6), 580–582.
10. Moo-Young, M.; Chisti, Y. Bioreactor design for aeration of shear-sensitive fermentation cultures. 8th International Biolotechnology Symposium; Durand, G., Bobichon, L., Florent, Eds.; Paris, France: Societé Française de Microbiologie, 1988; Vol. 1, 454–466.
11. Murhammer, D.W. Review and patents and literature. The use of insect cell cultures for recom-

binant protein synthesis: engineering aspects. Appl. Biochem. Biotechnol. **1991**, *31* (3), 283–310.
12. Rouf, S.A.; Moo-Young, M.; Scharer, J.M.; Douglas, P.L. Single versus multiple bioreactor scale-up: economy for high-value products. Biochem. Eng. J. **2000**, *6* (1), 25–31.
13. Bastin, G.; Dochain, D. *Process Measurement and Control, 1: On-line Estimation and Adaptive Control of Bioreactors*; Elsevier: Amsterdam, Netherlands, 1990.
14. Tosi, S.; Rossi, M.; Tamburini, E.; Vaccari, G.; Amaretti, A.; Matteuzzi, D. Assessment of in-line near-infrared spectroscopy for continuous monitoring of fermentation processes. Biotechnol. Prog. **2003**, *19* (6), 1816–1821.
15. Stephanopoulos, G. Application of macroscopic balances and bioenergetics of growth to the on-line identification of biological reactors. Ann. N. Y. Acad. Sci. **1986**, *469* (Biochem. Eng. 4), 332–349.
16. Ramkrishna, D. On modeling of bioreactors for control. J. Process Control **2003**, *13* (7), 581–589.
17. Shuler, M.L.; Kargi, F. *Bioprocess Engineering, Basic Concepts*; Prentice Hall: Englewood Cliffs, NJ, 2002.
18. Lee, J.M. *Biochemical Engineering*; Prentice Hall: Englewood Cliffs, NJ, 1992.
19. Bailey, J.E.; Ollis, D.F. *Biochemical Engineering Fundamentals*; McGraw-Hill, Inc.: New York, 1986.

Fermenter Design

Kishore K. Kar
Fluid Mechanics and Mixing Group, The Dow Chemical Company, Midland, Michigan, U.S.A.

Juergen Lueske
Fluid Mechanics and Mixing Group, The Dow Chemical Company, Niedersachsen, Germany

Richard F. Cope
Fluid Mechanics and Mixing Group, The Dow Chemical Company, Midland, Michigan, U.S.A.

INTRODUCTION

Fermentation and separation processes are vital unit operations in the production of biological products using microbial and mammalian cells in the pharmaceutical, food, and chemical industries. Although primary metabolites leading to such industrial chemicals as alcohols and acids often reach yields close to their theoretical maximum, the fermentation of secondary metabolites such as antibiotics and enzymes yields significantly lower values. In recent years, while microbiologists have focused their research on new strain developments, biochemical engineers have devoted their research to the advancement of fermentation and separation technologies to address productivity and environmental or regulatory issues.

BACKGROUND

Industrial fermentation processes are either aerobic, where aeration is a must for the growth and production of the microorganisms, or anaerobic, where aeration is irrelevant to the process. For example, in the presence of the *Acetobacter* species acetic acid is produced by the oxidation of ethanol at a concentration of up to 15% (wt/vol) (g/100 ml). Full aeration is required; interruption of the air supply for only 30 sec can kill acetic-acid-producing bacteria. On the other hand, lactic acid is the end product of dextrose glycolysis using species of the genus *Lactobacillus*, especially *L. delbrueckii.* The specific rate of lactic acid synthesis is strongly dependent on the media pH but independent of aeration. In this case, adjustment of media normality is essential. Localized low-pH zones seriously hinder the bacterial growth and consequently the yield. Fast dispersion and dissolution of the pH adjusters [alkali—$Ca(OH)_2$] in the glucose media are critical factors for successful production of lactic acid. Because the design of lactic acid (anaerobic) fermenters is akin to that of typical large-scale stirred tank reactors, this entry will be limited to the design of aerobic fermenters only.

Biochemical engineers are often challenged when scaling up fermentations demonstrated by microbiologists on a laboratory scale. Those skilled in the art have realized the difficulty and ambiguity involved in scaling up multiphase mixing systems from a biochemist's 1 L flask to a cubic-meter scale. Hence, tests in an intermediate pilot scale process are usually completed to understand process feasibility and to identify parameters that control yield. It is pilot samples of the therapeutics that are submitted for pharmacological and toxicological studies during preclinical and clinical trials. Business strategies such as speed to market, early cash flow, and capturing value by extending sales under patent protection often necessitate engineering-only authorization or toll manufacturing while the clinical studies by the regulatory authorities are under way. Irrespective of the approach, biochemical engineers play a crucial role in the design of pilot plants, production fermenters, and a web of associated peripherals.

Multipurpose batch fermenters are typical installations in the toll manufacturing and pharmaceutical industries. The batch fermentation process provides economic and operational advantages because a single fermenter can produce multiple products. Although most of the fermentation is carried out using fed-batch processes, biological wastewater treatment commonly utilizes large-scale continuous processing to treat a vast amount of water. In this case, a series of configurations of continuous stirred tank fermenters possesses advantages such as sequential treatment of multiple substrates, reasonable residence times for high conversion efficiency under flow-through conditions, and

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120025014

improved volumetric reaction rates due to high substrate concentration.

TYPES OF FERMENTERS

Mechanically agitated and pneumatically agitated fermenters are commonly used in the bioprocessing industries. Figs. 1–5 and Table 1 show the schematics of these bioreactors and their specific usage based on fermentation criteria outlined by Storhas.[1]

Mechanically Agitated Fermenters

A stirred tank fermenter consists of a centrally mounted agitation system inside a cylindrical vessel. Typically, the agitation system is composed of either multiple radial flow impellers (see Fig. 4A) or a combination of radial and axial flow impellers as shown in Fig. 4B. A gas sparger is located below the radial gas dispersing impeller. The role of the bottom radial impeller is to break the sparged gas stream into smaller bubbles. Axial, down pumping impellers establish a synergistic flow pattern with the radial impeller that includes upward flow near the wall and downward flow in the core region of the tank. Multiple radial flow impellers create undesirable compartmentalized flow regions in the tank.

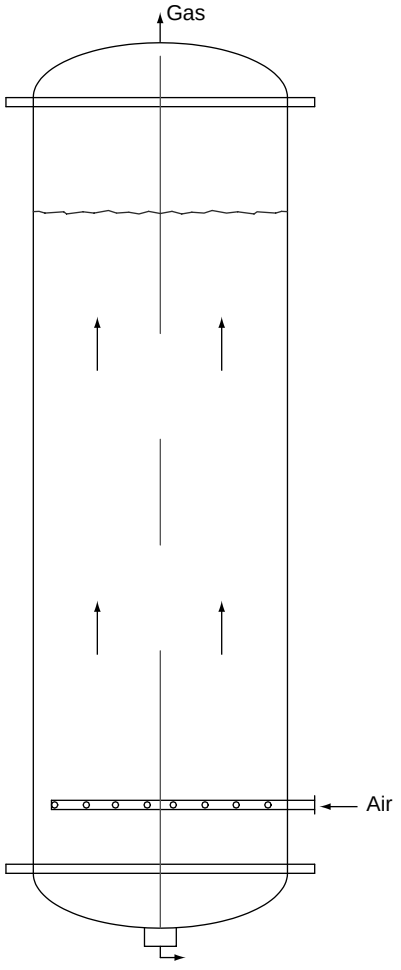


Fig. 1 Bubble column fermenter. *(View this art in color at www.dekker.com.)*

Stirred tanks are commonly used as the most efficient and reliable gas–liquid contactors for mass transfer limited aerobic processes in the biochemical industries. More than 90% of today's bioproducts are produced in stirred tank fermenters. Typical applications range from the large-scale production of such industrial enzymes as amylases and lipases, antibiotics or amino acids to the smaller-scale production of value-added specialty recombinant DNA therapeutic proteins and vaccines. The key advantages of stirred tank fermenters are enhanced mass transfer due to intimate gas–liquid turbulent mixing, improved heat removal through external heat transfer jackets and internal heat exchangers, and shear induced thinning of the pseudoplastic viscous media. Stirred tank fermenters are easily scaleable from the pilot to the large industrial scale. However, stirred tanks are not ideal for shear sensitive microorganisms and cannot be aerated at high rates because of impeller flooding. Poorly designed, tall, stirred tank fermenters may create regions of slow flow or stagnation, which lead to zones of poor oxygen transfer.

For low-viscosity media, the draft-tube–impeller combination in a stirred tank achieves better mass transfer performance than a conventional stirred tank fermenter. A cocurrent flow of gas and liquid is induced down through the draft tube and exits at its bottom. This flow pattern provides intimate gas–liquid contact and a high gas–liquid mass transfer rate. Consequently, bioreactors can consist of a tank, a draft tube, and an agitation system with axial and radial flow impellers as shown in Fig. 5. Dual axial flow impellers are located inside the draft tube and pump the fluid down toward the radial flow impeller. The radial gas dispersing impeller is located above the gas sparger ring. This configuration confines aeration to the annular zone and establishes a highly directional flow pattern in the tank. An optimized design draft-tube–impeller agitation system allows higher aeration rates before the onset of flooding.

Pneumatically Agitated Fermenters

The airlift fermenter relies on air as the principal means of achieving various transport requisites. It is composed of a vertical cylindrical vessel and a concentric draft tube into which air or other suitable oxidation gases are injected (see Fig. 2). Reduced bulk density causes the contents of the central riser to move upward. This action displaces the contents of the

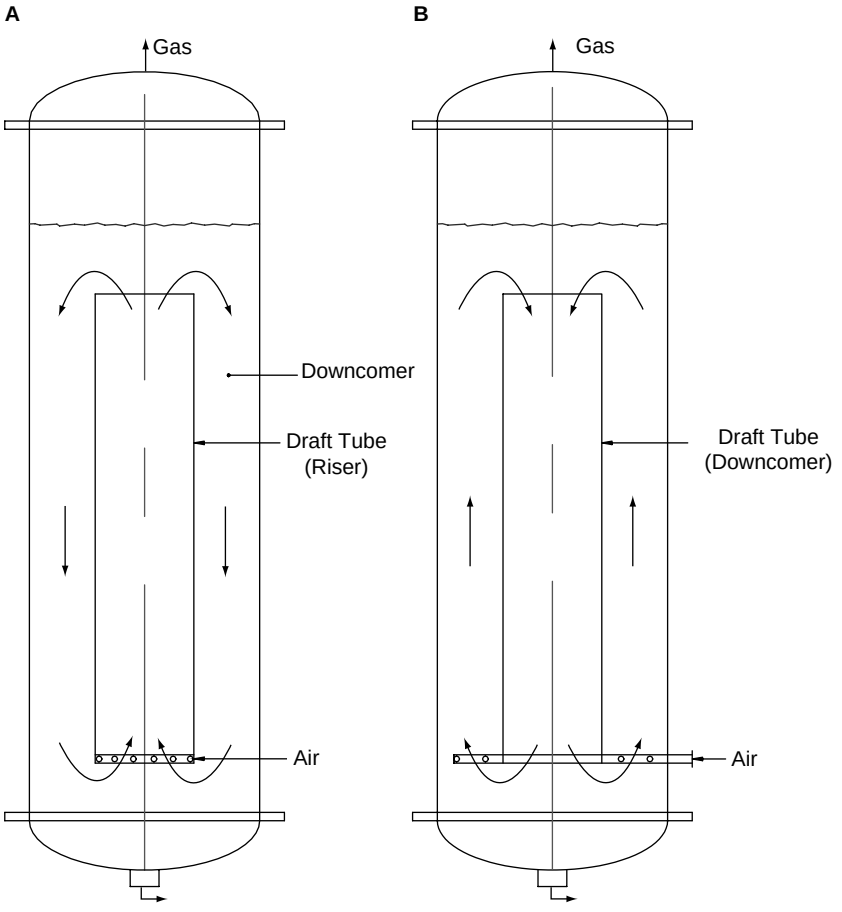


Fig. 2 Concentric tube airlift fermenters: (A) draft-tube sparged and (B) annulus sparged. *(View this art in color at www.dekker.com.)*

surrounding downcomer and moves them downward to complete the fermenter circulation cycle.

Airlift bioreactors have the advantages of generating liquid mixing and gas transfer without the use of a mechanical agitator. The special characteristics of airlift fermenters make them most suitable for some processes, especially those cultivating shear sensitive cultures such as mycelial fungi, filamentous bacteria, and certain mammalian cell lines. The system fluid dynamics ensure satisfactory mass and heat transfer together with an even distribution of shear stress. This offers advantages for both high-volume, low-cost processes and low-volume, high-value-added processes like those producing animal and cell cultures. Airlift fermenters are basically simple, low-capital-investment units.

There have been many simple modifications to airlift bioreactors for specific applications. For example, a novel airlift loop fermenter (schematic unavailable in the literature) utilizes a side arm. The external loop in this integrated system overcomes the problem of ethanol inhibition by continuously stripping ethanol from the fermentation broth and recovering it by condensation. This is suitable for the simultaneous production and recovery of ethanol.[2]

Mechanically stirred hybrid airlift reactors (see Fig. 6) are well suited for use with shear sensitive fermentations that require better oxygen transfer and mixing than is provided by a conventional airlift reactor. Use of a low-power axial flow impeller in the downcomer of an airlift bioreactor can substantially enhance liquid circulation rates, mixing, and gas–liquid mass transfer relative to operation without the agitator. This enhancement increases power consumption disproportionately and also adds other disadvantages of a mechanical agitation system.

Fig. 3 shows a modified airlift fermenter consisting of an external loop where gas is injected into the recycled broth through a gas-dispersing nozzle. In the working media inside the downcomer, two-phase discharge from the nozzle creates a jet of microbubbles. Jet-nozzle airlift fermenters provide better mass transfer but the accompanying high-shear bubble breakup is detrimental to the growth of most microorganisms.

The bubble column in Fig. 1 is the simplest of the pneumatically agitated bioreactors. Typical installations provide low-cost oxygen transfer in low-viscosity media where mass transfer is not limiting. Bubble columns are ideal for extremely shear sensitive microorganisms

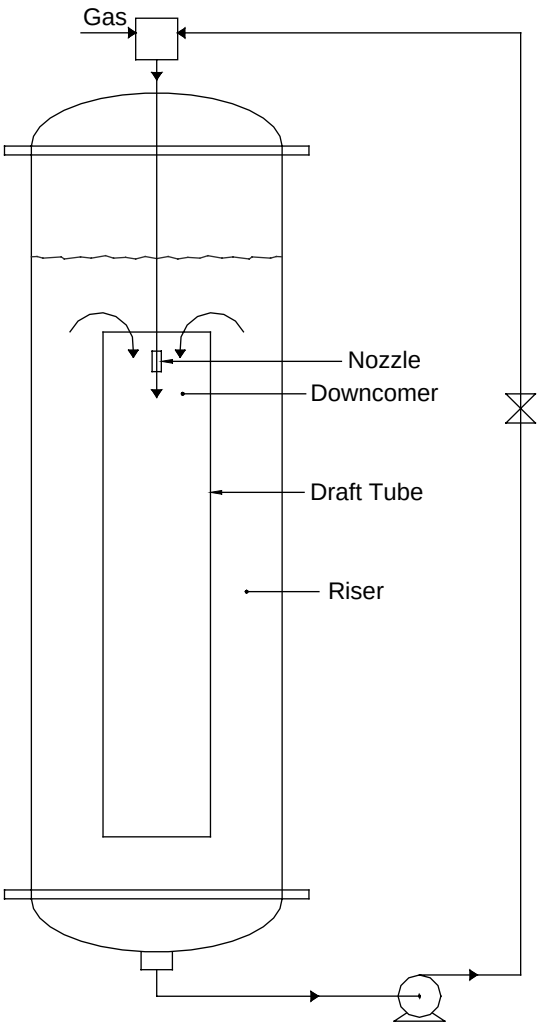


Fig. 3 External loop-jet nozzle airlift fermenter. *(View this art in color at www.dekker.com.)*

because their simplicity makes sterilization and maintenance of an aseptic environment simple. Because most bioprocessing applications are known to be mass transfer limited, bubble columns are rarely used for fermentation despite their obvious advantages.

Tubular fluidized and fixed bed fermenters are deviations from the simple bubble column fermenter. Often utilized in producing beer and ciders, these fermenters contain immobilized microorganisms or microbial films on support surfaces. Microbes lost with the product are continuously replenished by adding fresh microorganisms into the packed bed fermenters. In the fixed bed case, slow downward flow of the medium significantly reduces the shear removal (mobilization) of the microbes from the support materials and increases the residence time in the packed column. This is a typical characteristic of the trickle bed fermenter for continuous operation. Readers are referred to the packed bed reactor entry in this volume for a more comprehensive understanding of packed bed fermenter operations as they are not discussed in this entry.

SCOPE

The focus of this entry is limited to the design of aerobic fermenters, i.e., stirred tank and concentric tube airlift fermenters, which are commonly utilized in the bioprocessing industries. Design principles and the basic calculations are described with a couple of industrial examples. Readers with a limited background in mixing technology are referred to the gas–liquid contactor entry of this encyclopedia for a more comprehensive understanding of fluid flow and mass transfer characteristics in stirred tank reactors and bubble columns.

DESIGN PRINCIPLES

Oxygen Transfer Rate

Gas–liquid mass transfer is commonly modeled in terms of a gas film (between the bulk gas and interface) and a liquid film (between the interface and bulk liquid). Hindrance to mass transfer causes soluble gas (e.g., O_2) concentrations to decrease across these films. The highest mass transfer resistance usually exists in the liquid film; therefore, it controls the overall oxygen transfer rate (OTR). In aerobic fermentation, an effective fermenter design achieves an efficient OTR through intimate gas–liquid contact. OTR is described in terms of oxygen concentration and characteristics of the gas–liquid interface, as follows:

$$\mathrm{OTR} = k_{\mathrm{L}}a\Delta c = k_{\mathrm{L}}a(c^* - c_{\mathrm{L}}) \tag{1}$$

In this equation, k_{L} is the mass transfer coefficient in the liquid film, a is the gas–liquid interfacial area per unit liquid volume, c^* is the oxygen concentration near the gas–liquid interface, and c_{L} is the oxygen concentration in the bulk liquid.

Henry's law (i.e., $c^* = p_{\mathrm{G}}/H$) relates the equilibrium oxygen solubility in the liquid (c^*) to oxygen partial pressure in the gas (p_{G}) by the corresponding Henry's law constant (H). Because the liquid film mass transfer coefficient, k_{L}, is difficult to measure independently, $k_{\mathrm{L}}a$ (i.e., k_{L} times a) is used. It is a lumped parameter known as the volumetric mass transfer coefficient to characterize the overall mass transfer rate.

Because c^* is a function of hydrostatic head, a log-mean concentration difference generally replaces the Δc term used in the earlier equation:

$$\mathrm{OTR} = k_{\mathrm{L}}a\Delta c_{\mathrm{ln}} \tag{2}$$

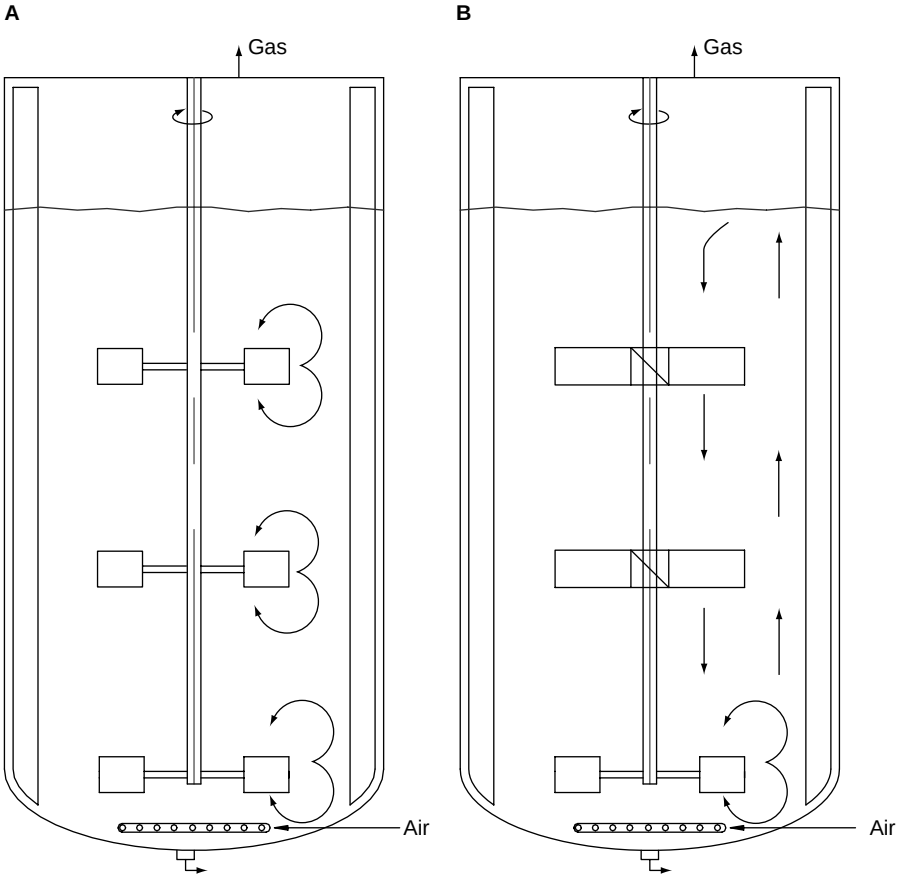


Fig. 4 Stirred tank fermenters: (A) with multiple radial impellers and (B) with axial and radial impellers. *(View this art in color at www.dekker.com.)*

where

$$\Delta c_{\ln} = \frac{(c_1^* - c_L) - (c_2^* - c_L)}{\ln((c_1^* - c_L)/(c_2^* - c_L))} \quad (3)$$

The average logarithmic concentration difference is calculated with the oxygen gas concentrations in the bulk liquid phase and saturation conditions at the inlet (1) and outlet (2). In this application the value of c_L is commonly assumed to be 0 ($c_L \ll c^*$).

Several simple theories of interaction at the gas–liquid interface are pertinent to the design of pneumatically and mechanically agitated fermenters, and are now discussed.[3]

The first is the penetration theory of Higbie (1935). If the liquid immediately adjacent to a rising bubble is assumed to rise with the bubble, i.e., the relative velocity between the bubble and the liquid is 0, the mass transfer conditions are those of unsteady-state molecular diffusion. The mathematical solution of this problem leads to

$$k_L = 2\sqrt{\mathscr{D}_L/\pi\theta} \quad (4)$$

where k_L is the liquid film mass transfer coefficient over the bubble lifetime (θ), and $\mathscr{D}_L$ is the coefficient of molecular diffusion.

The second is the surface renewal theory of Danckwerts (1951). If turbulent mixing occurs in liquid near a bubble, it is likely that unsteady molecular diffusion will take place in individual eddies. In these conditions, mathematical analysis leads to

$$k_L \propto \sqrt{\mathscr{D}_L s} \quad (5)$$

where s is the rate of surface renewal, i.e., the rate at which eddies sweep the interfacial boundary layer. Because the surface renewal rate incorporates the eddy lifetime (θ), it increases with increased turbulence intensity just as θ does in Higbie's penetration theory.

Because of the absence of mechanical agitators, airlift fermenters are generally unable to produce the turbulence necessary for high mass transfer rates. Superficial gas velocity and liquid circulation rate in the risers determine the relative velocities of rising bubbles and adjacent liquid and, in turn, the volumetric mass transfer coefficient, $k_L a$. Increasing the gassing rate increases the liquid circulation rate indirectly by increasing the hydrostatic pressure on top of the downcomer. Decreasing the downcomer pressure drop by increasing the ratio of downcomer to riser cross-sectional area also increases the circulation rate. Unfortunately, other challenges arise if higher liquid

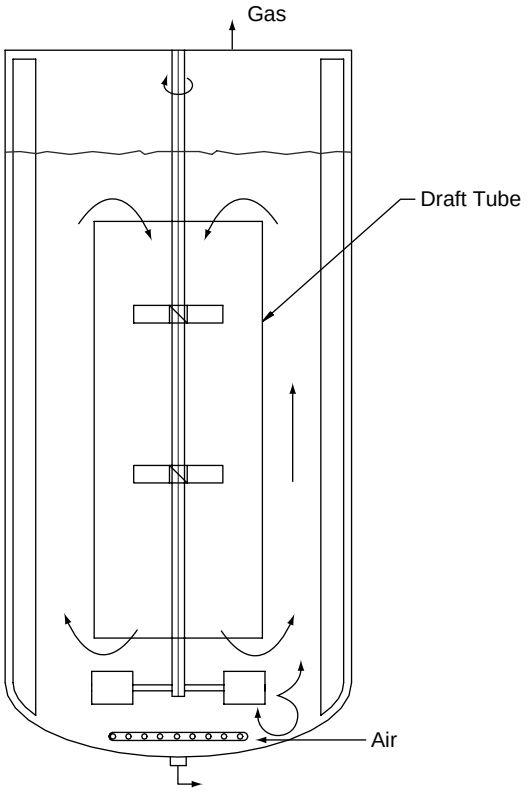


Fig. 5 Draft-tube–impeller stirred tank fermenter. *(View this art in color at www.dekker.com.)*

circulation rates shift the liquid conditions from bubbly flow to churn flow. Although this transition contributes to turbulence and bubble surface transience, bubble surface renewal also increases coalescence and bubble streaming along the wall, thus reducing specific interfacial area, a, and $k_L a$. Particularly in concentric draft-tube airlift fermenters, designers must maintain bubbly flow in the riser by optimizing gassing rate and liquid circulation rate. Typically, industrial fermenters have an a_r/a_d ratio of 0.8–1.2 (preferably $d_t/d_c = 0.7$) and a superficial gas velocity, u_g, of 0.02–0.10 m/sec.

The gas–liquid interfacial area per unit volume, a, increases as bubble size decreases. Hence, the design and location of gas spargers (aerators) are critical to the success of aerobic fermenters, particularly those with pneumatic agitation. Ring spargers, rake spargers, and porous plates are used as aerators in airlift fermenters. In the case of stirred tank fermenters, energy dissipation from the gas dispersing radial impellers normally controls bubble size. Ring spargers are typical aerators in mechanically agitated fermenters. Irrespective of the sparger holes, bubble size distribution depends on the media rheology, interfacial tension, and turbulent intensity. Increased broth viscosity reduces diffusivity, increases bubble size (i.e., under turbulent conditions Sauter mean bubble diameter, d_{b32}, is proportional to $\mu_L^{0.1}$), and decreases bubble-specific surface area a.[4] Therefore, it significantly decreases $k_L a$. Generally, $k_L a$ decreases with decreases in pressure and temperature and with increases in gas–liquid interfacial tension and liquid viscosity.

Designing fermenters that produce high yields with such viscous media as polysaccharides is a challenge. In batch fermenters, for example, agitation initially disperses the sparged gas as tiny, well-distributed bubbles. The viscous media stabilize these bubbles and extend their lifetime beyond that required for oxygen mass transfer. The extended lifetime of these bubbles hinders subsequent oxygen transfer, and thus fermenter productivity, by dampening agitation and its ability to continue dispersing fresh sparged gas. The resulting mass transfer coefficient falls off dramatically with continued batch time. A much higher energy input is needed to generate fine bubbles from the sparged gas and prevent them from coalescing.

Microbial Metabolic Heat

In some cases, heat is added to the initial broth to promote the desired microbial activity. During metabolism, however, thermodynamics of the overall microbial activity generates heat in such quantities that, if not removed, it would raise broth temperature beyond the system optimum. Elevated temperatures affect the fermenter by producing everything from a slight productivity decrease to widespread microbe death.

Metabolic heat evolution is determined from small-scale laboratory experiments and used for the appropriate design of heat transfer elements. When data are not available, heat evolved during bacterial fermentation can be estimated using the following correlation:[5]

$$\dot{Q}_f \cong 0.12 \times \text{OUR} \tag{6}$$

where $\dot{Q}_f$ is heat evolution (kcal/L/hr) and OUR is the oxygen uptake rate (mmol O_2/L/hr).

Addition and removal of heat is achieved by such means as a fermenter jacket, internal heat transfer coils (horizontal or vertical), and/or an external heat exchanger. A fermenter jacket provides insufficient heat removal by itself, especially in large microbial fermenters. Internal cooling parts may be particularly necessary in slender tanks despite their impedance to cleaning, sterility, and intermixing. Appropriately designed vertical heat transfer coils often serve as baffles in mechanically agitated fermenters, thus minimizing extraneous structures inside the vessel. Sterilization difficulties and asepsis also make external heat exchangers with multiphase pumping a last resort for

Table 1 Critical matrix for fermenter selection

Criteria	Units	Criteria and preferred bioreactor type					
Media viscosity	Pa s	>2.0	<2.0	<0.4	<0.1		
	Reactor types[a]	4, 5	4, 5, 3	4, 5, 3, 2	5, 4, 3, 2, 1		
Ease of pumping		Very poor	Poor	Good	Very good		
	Reactor types[a]	4, 5	4, 5, 3	4, 5, 3, 2	5, 4, 3, 2, 1		
Susceptibility to shear stress		No	Little	Yes			
	Reactor types[a]	4, 5, 3	4, 2	1, 2			
Maximum reactor size	m^3	>500	<400	<300	<100	<10	<5
	Reactor types[a]	1	4, 1	5, 4, 2	5, 4, 3, 2	5, 4, 3, 2	4, 3, 2
Media solid content		Very high	⟶		Very low		
	Reactor types[a]	4	4, 5	4, 5, 2, 3	4, 5, 2, 3, 1		
Tendency to foam		Very strong	⟶		Very weak		
	Reactor types[a]	4	4, 5	4, 5, 2, 1	4, 5, 2, 1, 3		
Difficulty in achieving media homogeneity		High ⟶			Low		
	Reactor types[a]	5, 4	5, 4	5, 4, 3, 2	5, 4, 3, 2, 1		
Mass and heat transfer required		High ⟶			Low		
	Reactor types[a]	5, 4	5, 4	5, 4, 3, 2	5, 4, 3, 2, 1		
Sterilization required (cleaning in process)		High ⟶			Low		
	Reactor types[a]	1, 2	1, 2, 3	1, 2, 3, 4	1, 2, 3, 4, 5		
Biological safety		High ⟶			Low		
	Reactor types[a]	1, 2	1, 2, 3	1, 2, 3, 4	1, 2, 3, 4, 5		
Example case							
Criteria	0.4–2.0 Pa s	<300 m^3	30% solids	High mass transfer	High sterility		
Bioreactor types	4, 5, 3	5, 4, 2	4, 5	5, 4	1, 2		
Preferred type	4 or 5						

[a]Refers to the bioreactor types shown in Figs. 1–5.
(From Ref.[1].)

heat removal. When an external heat exchanger is implemented, its design and materials of construction must allow it to withstand thermal stress cycling and chemical cleaning.

In the case of stirred tank fermenters, heat that must be dissipated includes not only that generated by microbial metabolic activity but also that evolved from agitation power (i.e., 2500 Btu/shp-hr) and expansion of sparged gas. This can lead to scale-up problems because vessel volume is proportional to vessel diameter cubed, while heat transfer area is proportional only to vessel diameter squared.

Sterilization and Asepsis

For most aerobic fermentation processes, maintenance of absolute sterility is critical to cell growth and biomass production. Culture growth rate determines susceptibility to culture contamination. Mammalian cells divide in a day, while microbials divide in an hour or less. The difference in growth rate makes slower-growing mammalian cell cultures more susceptible to contamination than faster-growing bacteria or yeast cultures, which can out-compete contaminating microorganisms. Substrates that include such heat-labile nutrients as vitamins or enzymes are often added to the media as sterilizing agents and growth enhancers.[6] Prior to fermenter operation, standard operating practice usually includes a caustic wash and/or superheated steam cleaning of the vessel and gas sparger.

Typical sources of contamination include the steady bearing, mechanical seal, sparger, and fasteners. Because bushing interfaces and the shaft wear-sleeve of a steady bearing are prone to biomass deposition, their design is given special attention. Use of steam-purged or dry seals minimizes biomass contamination. Type 316L stainless steel is the most common construction material for bioreactor applications. Because microcrevices are known to harbor bacteria and contaminants, the surface of the tank walls and internals are electro-polished to a minimum surface finish of 0.6 μm (average roughness). To prevent the influx of air, which can compromise aseptic conditions, fermenters normally operate with a positive internal pressure.

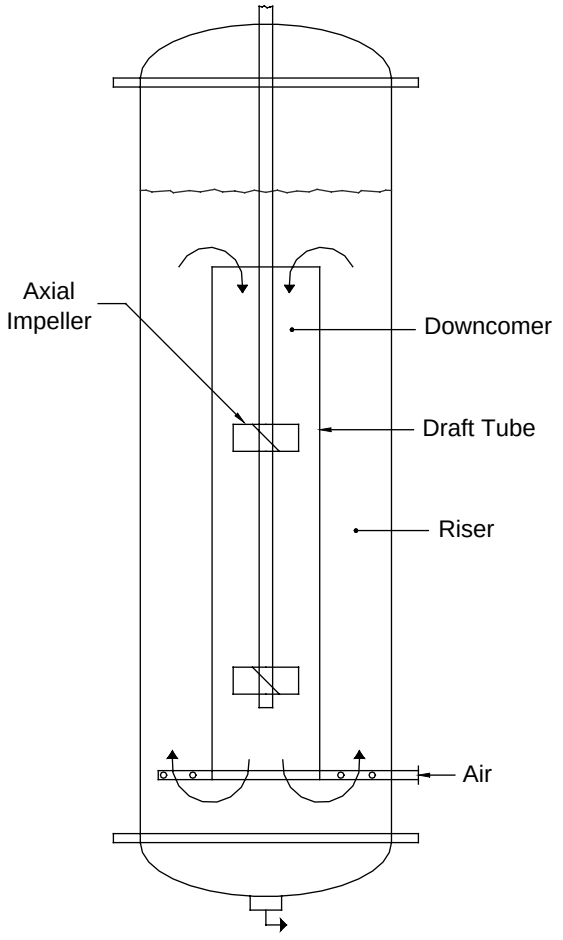


Fig. 6 **M**echanically agitated hybrid airlift fermenter. *(View this art in color at www.dekker.com.)*

A high level of positive pressure also has the advantage of increased oxygen partial pressure and, consequently, mass transfer rate.

Turbulence and Shear Tolerance

Prior knowledge of cell sensitivity to shear fields and turbulence is critical in designing mechanically agitated fermenters. Morphology dictates the tolerance of microorganisms to turbulence intensity and shear stress though most microorganisms are too small (typical sizes: bacteria 1–5 μm) to be destroyed by turbulent microscale eddies or local shear forces. Secondary metabolites with spherical, ellipsoidal, and rod-like structures are not susceptible to mechanical degradation, while filamentous microbes such as *Spinosad* mycelia and fungi are degraded by shear. Mammalian cells, because of their lack of cell wall and larger size (>5 vs. 1 μm), are more shear sensitive than microbial cells. Hydrodynamic shear stresses or shear rates also affect the production of extracellular microbial polysaccharides. During pilot scale fermentation, potential cell damage due to high specific energy must be considered. Intense shear fields can disrupt even the most robust microorganisms, but the likelihood of cell damage can be safely ignored if the Kolmogoroff length of a microvortex is greater than the effective size of the microbe in question.

According to Kolmogoroff, the size of the microvortex, λ', can be determined using the following equation:

$$\lambda' = \left(\frac{\nu_L^3}{\varepsilon_{loc}}\right)^{1/4} \tag{7}$$

Media Rheology

Flow properties of biofluids and slurries fall into two categories:

1. Newtonian fluids such as water, and most bacterial and yeast fermentation broths.
2. Non-Newtonian media such as polysaccharide fermentations and broths of *Streptomyces, Aspergilli*, and *Penicillia*.

The viscosity of Newtonian suspensions of yeast and bacteria growing as individual cells in a water-like medium can be estimated from the following empirical equation proposed by Thomas.[7]

$$\mu_{susp} = \mu_L\left(1 + 2.5c_s + 10.05c_s^2 + 0.00273\,e^{16.6c_s}\right) \tag{8}$$

where c_s is the volume fraction of solids.

While working media can be Newtonian or non-Newtonian, most high-viscosity non-glucose-based broths containing microorganisms demonstrate pseudoplastic or shear thinning behavior. For fungal suspensions (2–10 kg/m^3 dry biomass) such as *Aspergillus niger*, in the power law model, $\tau = k\dot{\gamma}^n$, the flow behavior index n varies from about 0.6 to 0.16, while the consistency index k increases from 0.17 $Pa\,sec^{0.6}$ to 9.1 $Pa\,sec^{0.16}$.[8] While *Acetobacter* fermentation is highly aerobic, the fermentation broth itself is viscous, even viscoelastic in certain instances. Although the fluid's turbulence does not seem to significantly affect the biomass or its productivity, the quality of the polymer produced is distinctly shear stress dependent. To ensure sufficiently long polymer microfibrils, the shear must be kept low.[9]

For high-viscosity applications, lack of turbulence adjacent to the bubble, and the general difficulty in pumping around the draft tube (riser), prevent the use of simple pneumatically agitated airlift fermenters. Stirred tank fermenters are typically used for media with a viscosity exceeding 100 cP. As noted before, an

increase in viscosity results in a substantial decrease in the mass transfer coefficient, all other parameters being equal. Fig. 7, a graph of k_La vs. apparent viscosity for a penicillin slurry, illustrates this trend.[10] To illustrate further, achieving a reasonable mass transfer coefficient ($>0.1\,sec^{-1}$) typically requires two to three times more specific power in a high-viscosity broth ($>1000\,cP$) than in low-viscosity ($<100\,cP$) media. In viscous broth fermentation, yield is often limited by oxygen starvation.

Blending and Homogeneity

Uniformly distributing air bubbles, nutrients, and microbes within a broth are of paramount importance for the successful cultivation of microorganisms. Airlift fermenters are ideal for achieving homogeneity in broths with a low solids content, while mechanically agitated fermenters are suitable for high solids (slurry) broths. In spite of the possibility of localized non-homogeneity due to slow flow regions in stirred tanks, mechanically agitated fermenters are preferred over pneumatically agitated ones. An appropriately designed draft-tube–impeller agitation system (shown in Fig. 5) overcomes the uniform blending deficiency of the standard stirred tank fermenter. Configuration and location of tank internals such as agitators, baffles, feed pipes, and heat transfer coils/plates are critical for blending and efficient heat transfer. Fermentation yield depends strongly on the homogeneity of pH and temperature in the working media.

Fed batch fermentation is widely practiced in the industry. Substrates including nutrients such as glucose are continuously fed through dip pipes as the fermentation proceeds. For instantaneous or fast dispersion, these ingredients must be added in high-turbulence regions in the vessel. It must be noted that a local surplus of glucose can cause excessive biomass or by-products to be produced, which can negatively impact the yields of the desired products, e.g., antibiotics. In this regard, selection of the broth, its feed point, and the feeding rate strategy are important.

Gas Holdup and Foam Formation

In vessel design work, size and geometry must account for the fact that aerobic fermenters experience significant gas holdup owing to finely dispersed bubbles in the broth. Gas holdup can be in the range of 10–30% depending on the media viscosity and the presence of surface-active substrates. Because of much higher turbulent intensity, mechanical agitated fermenters have a much higher gas holdup than pneumatically agitated fermenters. As with the mass transfer coefficient, gas holdup in a given media depends on the specific power input ε and the superficial gas velocity u_g.

From the mass transfer point of view, airlift fermenters should be designed and operated in such a way that carryover of air from the riser into the downcomer is kept as low as possible. Gas in the downcomer liquid contributes little to oxygen transfer. It reduces the effective density difference between the contents of the riser and downcomer, however, which reduces the liquid circulation rate and also impairs the mixing performance of the fermenter. In fermentations where even a momentary lack of oxygen can very seriously affect productivity, some air is essential to maintain aerobic conditions and sustain fermentation in the downcomer.

Foaming is a major concern in aerobic cell cultures because excessive foaming significantly reduces broth volume and impedes bioreactor productivity. Normally, silicon-based or oil-based antifoam agents (e.g., corn oil, silicone antifoam emulsions such as GE AF-72, Dow Corning's FG-10, DC-1510/1520-US, etc.) are added to suppress foaming. In choosing a foam inhibitor, one must be aware of possible adverse impact on the cells, products, or downstream processes and on pertinent regulatory (Food and Drug

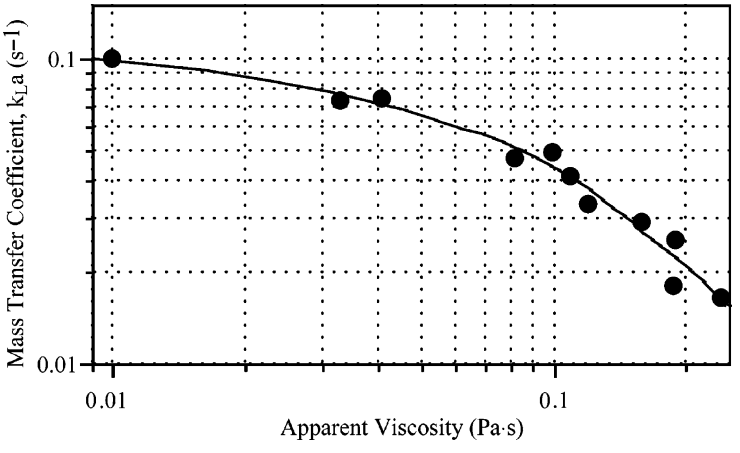


Fig. 7 Mass transfer coefficient (k_La) as a function of apparent viscosity of penicillin broth at a specific energy of 3 W/kg. (From Ref.[10].)

Administration) and/or kosher requirements. High molecular weight antifoam agents, when used in the low-viscosity broth, typically reduce the gas–liquid mass transfer coefficient.

Biological Safety

In the fermentation plant, the vent gas system commonly contains a backpressure valve to maintain positive pressure on the bioreactor. The vent is usually piped to a scrubber, which scavenges microbes from the fermenter off-gas. In addition to the scrubber, fermenters are equipped with an overhead safety relief device for potential loss of containment. With the advent of genetic engineering and the targeting of proteins from mammalian cells, biological safety standards have been extended to overcome unforeseen health issues. In this case, all U.S. large-scale bioreactors must comply with the Good Large-Scale Practice (GLSP) guidelines set forth by the U.S. National Institutes of Health (NIH). When using an organism covered by Biological Safety Level (BSL) classifications, BSL-1 or BSL-2, it is necessary to ensure that the cell is contained within the bioreactor's hygienic envelope. BSL-1 containment commonly requires inactivation of the organism by means of heat or chemicals (pH kill) in a separate tank.[5,6]

DESIGN GUIDELINES

The design basis is the most important consideration when determining the size of a fermenter. Factors to consider are the final product, the growth rate and oxygen requirement of the microorganism, the product expression concentration or titer, and the nature of the product expression.

Choosing an optimal bioreactor configuration for a given fermentation depends on a number of factors including oxygen transfer and mixing requirements, and the acceptable range of shear rates. Some of these factors are mutually contradictory. For example, while a given fermentation may require a very high gas–liquid OTR, which is generally increased by increasing turbulence and shear, its tolerance to such shear may be low. In such cases, successful bioreactor design obviously involves a careful balance of these contradictory factors.

The two sections that follow describe the designs of a stirred tank fermenter and a concentric tube airlift fermenter. Each design is for a given range of operating variables and includes the selection of geometric parameters. Important correlations and formulae are presented and followed by an example design case. The design process actually requires multiple iterations to simultaneously satisfy both geometric and mass transfer requirements. Such iteration has become almost trivial, however, with the advent of such tools as spreadsheet-based "goal–seek" computations.

Stirred Tank Fermenter

This, the first of two fermenter design examples, deals with a stirred tank fermenter. Nondimensional parameters that are applied in this example are defined as follows.

Nondimensional parameters

N_P	power number, $P/(\rho N^3 D^5)$
Na	aeration number, $\dot{V}_g/ND^3$
Fr	Froude number, N^2D/g
K	k-factor, P_g/P_{ug}
R	diameter ratio, D/T
S	impeller spacing ratio, s/D
C	impeller bottom clearance ratio, c_{btm}/D
Φ	gas holdup, $V_g/(V_g + V_L)$
μ_{sr}	suspension viscosity ratio, μ_{susp}/μ_w
Nu	Nusselt number, $\alpha_w T/\lambda$
Re	Reynolds number, $N_g D^2/\nu_L$
Re_o	orifice Reynolds number, $v_o d_o/\nu_g$
Pr	Prandtl number, $\mu c_p/\lambda_w$

Mass transfer

The relationship between the volumetric mass transfer coefficient, specific aerated shaft power, superficial gas velocity, and suspension viscosity ratio is given as follows:

$$k_L a_{20} = \alpha\, \varepsilon_g^{\beta}\, u_g^{\gamma} \left(\frac{\mu_{susp}}{\mu_w}\right)^{\kappa} \tag{9}$$

The following model parameters are specific to proprietary glucose-based fermentation media that were correlated with the above $k_L a$ model for $1.5 \leq \varepsilon_g \leq 3.5\,\mathrm{W/kg}$:

$\alpha = 0.258$, $\beta = 0.562$, $\gamma = 0.232$, and $\kappa = -0.333$

Using Eq. (8), the viscosity of the suspension is determined from the broth viscosity and volumetric concentration of solids.

The $k_L a$ value determined with the above model is valid at a constant temperature of 20°C. The mass transfer coefficient at actual operating temperatures is given by

$$k_L a = k_L a_{20} \frac{1}{1.024^{(20-t)}} \tag{10}$$

Stirrer power under gassed conditions is shown below:

$$P_g = \varepsilon_g m_{Broth} = \varepsilon_g \rho_{Broth} V_{Broth} \quad (11)$$

Gas holdup

Similar to the mass transfer model, gas holdup may also be expressed as a function of the specific power ε_g, the superficial gas velocity $\bar{u}_g$, and the viscosity ratio. Model parameters (α, β, γ, and κ) obtained from the gas holdup expression are neither related to nor interchangeable with those obtained in the mass transfer model:

$$\Phi_g = \alpha \varepsilon_g^{\beta} \bar{u}_g^{\gamma} \left(\frac{\mu_s}{\mu_w}\right)^{\kappa} \quad (12)$$

The following parameters are specific to the broth used in fitting the above gas holdup model. In this example these parameters have values of:

$\alpha = 0.500, \quad \beta = 0.143,$
$\gamma = 0.197,$ and $\quad \kappa = -0.333$

The k-factor

A common way to determine the agitator power under gassed conditions is to use the k-factor, sometimes called the relative power demand, defined as the ratio of gassed to ungassed shaft power:

$$\text{k-factor: } K \equiv \frac{P_g}{P_{ug}} \quad (13)$$

For a stirred system with multiple impellers of different types, the overall k-factor can be experimentally determined as in Fig. 8 (k-factor vs. aeration number at different Reynolds numbers), or can be approximated by adding weighted k-factors for the individual impellers, as follows:

$$K = \frac{N_{P_R}}{N_P} K_R + \frac{N_{P_A}}{N_P} K_A \quad (14)$$

The first term on the right represents the radial, bottom impeller and the second term on the right applies to axial, down-flow impeller(s). The following relationships are used in the design:[11]

$$K_R = 1 - (b' - a'\mu_{b'}) Fr_R^{d'} \tanh(c' \mathrm{Na}_R) \quad (15)$$

Values of the constant parameters a', b', c', and d', as shown in Table 2, depend on the radial impeller type.

$$K_A = 1 - (a'' + b'' Fr_A) \mathrm{Na}_A^{c'' + 0.04 Fr_A} \quad (16)$$

The parameters a'', b'', and c'' in the K_A relationship are not constant, but depend on the impeller to tank diameter ratio D/T:

$$a'' = 5.3\, e^{-5.4(D_A/T)}$$
$$b'' = 0.47(D_A/T)^{1.3}$$
$$c'' = 0.64 - 1.1(D_A/T)$$

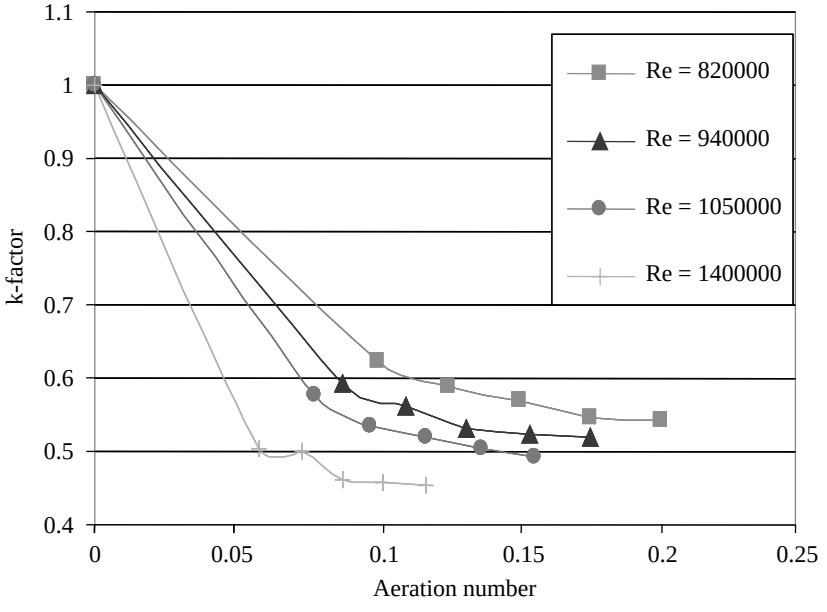


Fig. 8 k-Factor of multiple impeller system (Rushton + PBT) as a function of aeration number. *(View this art in color at www.dekker.com.)*

Table 2 Constants for Eq. (15)

Blade type	a'	b'	c'	d'
Flat	0.72	0.72	24	0.25
Concave	0.12	0.44	12	0.37

(From Ref.[11].)

Eqs. (15) and (16) are valid for

$$0.40 < D/T < 0.65$$
$$0.50 < Fr < 2.0$$
$$0.05 < Na < 0.35$$

$$P_{ug} = \frac{1}{K} P_g \tag{17}$$

Motor and gear box design normally assumes an 80% drive system efficiency ($\eta_m = 0.8$).

$$P_m = \frac{P_g}{\eta_m} \tag{18}$$

At a fixed agitation rate, the gassed power, P_g, is generally less than the ungassed power, P_{ug}. Consequently, the ungassed agitation rate is typically lower than the gassed agitation rate, i.e., $N_{ug} < N_g$. If gas feed suddenly stopped while operation continued at the gassed agitation rate, the mixer drive system would likely be overloaded and damaged.

Gas flooding

A stirred tank fermenter has power introduced through both the agitation system and the expanding gas. With the assumption of fast temperature equalization, an isothermal expansion of the gas (air) takes place. The gas power is:

$$P_{gas} = \dot{m}_{gas} \frac{RT}{M_{gas}} \ln\left(\frac{p_1}{p_2}\right) \tag{19}$$

where R is the gas constant (8.314 kJ/kmol/K), T is the absolute temperature at operating conditions, and M_{gas} is the molecular weight of gas. The pressure ratio is given at inlet (1) and outlet (2) tank locations.

For high aeration values, gas expansion power P_{gas} can reach the magnitude of the aerated agitator power P_g. If gas expansion power is greater than or equal to the stirrer power, there is the likelihood of impeller flooding, which limits gas dispersion to the central region of the tank.

To avoid potential flooding, it is recommended that

$$(P_g/P_{gas})_{design} \geq 3 \tag{20}$$

Design example

The design of a stirred tank fermenter for the production of an industrial enzyme at an annual rate of 800 ± 5% mt/yr is illustrated below. Product recovery efficiency is 80% and the expected yield (product concentration) is 75 kg/m^3. Maximum oxygen uptake demand is 185 mmol O_2/L/hr. Operational parameters and media physical properties are listed as follows.

Batch length = 3 days.
Fermenter turnaround rate = 2 days.
Number of operating days per year = 320.
Annual fermenter output = required production/recovery efficiency = 800 mt/yr/0.8 = 10^3 mt/yr.
Required batch size = (annual fermenter output/titer) × (total batch duration/yearly operation) = (10^6/75) (5/320) = 208 m^3.
Number of bioreactors = 2.
The required batch size per fermenter = 208/2 = 104 m^3.
Assuming a tank volume utilization of 80% the size of the bioreactor is 130 m^3.
The nominal tank size is 140,000 L.

Broth properties of Newtonian fluids

Effective glucose broth (liquid) viscosity = 5×10^{-3} Pa sec.
Effective media (suspension) viscosity = 15.3×10^{-3} Pa sec.
Density = 993 kg/m^3.
Interfacial tension (gas–liquid) = 7.1×10^{-2} N/m.
Thermal conductivity = 0.617 W/m/K.
Specific heat capacity = 4.118 kJ/kg/K.
O_2 broth diffusion coefficient = 4.3×10^{-9} m^2/sec.

Additional conditions for fermentation and heat removal, as determined at the pilot scale, are presented in Table 3. The state-of-the-art aerobic fermenter agitation system design recommendations are shown in Table 4. Based on the required fermentation batch size and agitator design recommendations listed in Table 2, dimensions of the fermenter and its internals are summarized in Tables 5 and 6, respectively. Drive system specifications for satisfactory OTR are calculated using pertinent design equations and correlations (see Table 7). Fig. 9 shows a schematic of the designed production fermenter. For details of the mechanical design, readers are referred to the publication by Charles and Wilson.[12]

Gas sparger design

Some basic guidelines for the sparger design are presented in the ''Gas Liquid Contactor'' entry.

Table 3 Production fermenter design parameters

Batch operation parameters	
Temperature (°C)	32
Head space pressure [bar (gage)]	0.689
VVM i.e. V_g/V_L (standard conditions) (1/min)	1.3
Solid concentration (vol%)	30
Respiration ratio (mol CO_2/mol O_2)	0.8
Heat transfer conditions	
Reaction heat liberation (kJ/mol O_2)	500
Chilled water supply temperature (°C)	12
Max. chilled water flow (m^3/h)	650
Thermal conductivity (W/m/K)	16
Fouling coefficient on product side (m^2 K/W)	0

The recommended sparge ring to impeller diameter ratio is 0.65–0.8, while the ring sparger orifice diameter is in the range of 2–10 mm, and the gas velocity through the orifice is between 15 and 30 m/sec.[13]

Pressure drop across a perforated disperser or orifice is estimated by the conventional orifice equation:

$$\Delta_{P_o} = \rho(\nu_o/c_v)^2/2 \quad (21)$$

The above equation neglects velocity of approach to the orifice. Typically, the coefficient of velocity, $c_{v,}$ is assumed to be 0.8.

To ensure uniform flow through all the orifices or holes, rules of thumb for the gas sparger design call for satisfaction of the following criteria.[14] First, the pressure drop ratio between the total pressure loss along the pipe, Δp_p, and the pressure loss across each hole, Δp_o, should be:

$$\Delta_{P_p}/\Delta_{P_o} \leq 0.1 \quad (22)$$

where

$$\Delta_{P_p} = [(4fL_p/3d_p) - 1](\rho u_{in}^2/2) \quad (23)$$

and f is the Fanning friction factor.

Second, the ratio of kinetic energy of the gas at the inlet of the pipe, and the pressure drop across each hole should be:

$$KE_{in}/\Delta_{P_o} \leq 0.1 \quad (24)$$

where

$$KE_{in} = 1.05(\rho u_{in}^2/2) \quad (25)$$

Details of the ring sparger design are summarized in Table 8. A schematic of the sparger is shown in Fig. 10.

Table 4 Agitator design recommendations

Impeller type	Radial	Axial
Description	Gas dispersing disk turbine	Pitched blade turbine
Number of impellers	1	2 (0–3)
Impeller diameter, D/T	0.4–0.6	0.4–0.6
Impeller spacing, s/D	—	1.0–1.5
Impeller bottom clearance, c_B/T	0.25	—
Material of construction	SS-316	

Table 5 Fermentation tank data

Material of construction	SS-316
Outside diameter (m)	4.30
Height (cylindrical section) (m)	8.25
Wall thickness (m)	0.025
Bottom head shape	2 : 1 semielliptical
Total fill volume (m^3)	130
Nominal tank volume (L)	140,000
Number of baffles (flat)	4
Baffle width (m)	0.43
Baffle clearance to tank wall (m)	0.09
Baffle height (m)	7.425

Heat transfer calculations

The first step in the design calculations is to determine the amount of heat that has to be removed to maintain

Table 6 Agitation system design data

Number of radial flow impellers (bottom)	1
Impeller type	Smith turbine
Diameter (m)	2.15
D/T ratio	0.50
Number of axial flow impellers second to fourth	2.00
Impeller type	Hydrofoil: A-320™
Diameter (m)	2.15
D/T ratio	0.50
Distance between bottom and middle impellers (m)	2.58
Distance between middle and top impellers (m)	2.58
Distance between top impeller and ungassed liquid (m)	1.30
Distance between top impeller and gassed liquid (m)	2.45
Critical speed ($N_{cr} > 1.25N_g$)	Contact vendor
Shaft diameter	Contact vendor

Table 7 Results of stirred tank fermenter design calculations

Variable	Equation	Value
Required $k_L a$ (1/sec)	2	0.12
VVM (1/min)		1.30
Standard volume flow (m^3/hr)		8124
Superficial gas velocity (m/sec)		0.072
Required aerated power per mass (W/kg)	9	2.3
Aerated power (kW)	11	238
k-Factor (P_g/P_{ug})	14–16	0.75
Ungassed power (kW)	17	320
Shaft power (aerated power) (kW)	11	238
Gassed speed (from ungassed power) (rpm)		69/AGMA-68
Ungassed speed (from shaft power) (rpm)		63/AGMA-56
Motor power (kW)	18	298
Gas expansion power (kW)	19	80
Shaft power/gas power (check)	20	>3.0
Aeration number		0.091
Gas holdup (%)	12	13.80
Ungassed height (m)		7.55
Gassed height (m)		8.70

isothermal conditions in the tank. As mentioned earlier, a gas–liquid stirred tank bioreactor has three sources for heat generation. These are the gas expansion power, the agitation power under gassed conditions, and the heat liberated by the bioreaction. The amount of heat to be removed can be expressed as:

$$\dot{Q} = P_{gas} + P_g + \dot{Q}_f \tag{26}$$

The last term, $\dot{Q}_f$, is the heat of formation in the fermenter. It is directly proportional to the OUR and normally much higher than the gas expansion and agitation power. For enzyme production, the relationship between heat of formation and oxygen demand, $\dot{n}_{O_2}$, is:

$$\dot{Q}_f \quad (\text{kW}) \approx 500\Delta\dot{n}_{O_2} \quad (\text{kmol/sec}) \tag{27}$$

Whether the heat duty can be removed depends on the different heat resistances discussed in the following section.

The heat duty is described as:

$$\dot{Q} = UA\Delta T_{ln} \tag{28}$$

where U is the overall heat transfer coefficient and A is the heat transfer area. ΔT_{ln} is the average logarithmic temperature difference defined as follows.

$$\Delta T_{ln} = \frac{\Delta T_1 - \Delta T_2}{\ln(\Delta T_1/\Delta T_2)} \tag{29}$$

ΔT_1 and ΔT_2 are the temperature difference between the cooling medium and tank contents at the inlet (1) and the outlet (2) of the cooling or utility flow. The overall heat transfer coefficient, U, is determined as follows:

$$\frac{1}{U} = \frac{1}{\alpha_p} + f_p + \frac{\delta_w}{\lambda_w} + f_j + \frac{1}{\alpha_j} \tag{30}$$

where, δ_w is the wall thickness, and α_p, and α_j are the heat transfer coefficients for the process side and jacket side, respectively.

For large-scale fermenters with multiple impellers, there are very few heat transfer correlations available. Most of the experimental investigations have been done with Newtonian liquids with a single, standard impeller. The following ungassed liquid phase Nusselt number correlation for heat transfer at the tank wall, in different vessel geometries with four standard baffles and a Rushton turbine, was developed by Strek.[15]

$$Nu_{pw-ug} = 1.01\left(\frac{D}{T}\right)^{0.13}\left(\frac{c_{bot}}{T}\right)^{0.12} Re^{2/3} Pr^{1/3}\left(\frac{\mu_L}{\mu_w}\right)^{0.14} \tag{31}$$

where pw represents process side locations near the wall (i.e., heat transfer surface).

Under gassed conditions, using a ring sparger in a similar system, Kurpiers and Steiff developed the

Under gassed conditions, using a ring sparger in a similar system, Kurpiers and Steiff developed the following correlation:[16]

$$Nu_{\mathrm{pw-g}} = 0.1\left\{\xi\left(\frac{T}{D}\right)\left(\frac{T^3}{V_{\mathrm{broth}}}\right)N_{P,\mathrm{g}}Re^3 \left[1 - 2\left(\frac{D}{T}\right)\left(\frac{D}{H_{\mathrm{ug}} - H_{\mathrm{sp}}}\right)N_{P,\mathrm{g}}^{0.5}Fr\right]\right\}^{0.25} \mathrm{Pr}^{0.4}\left(\frac{\mu_{\mathrm{L}}}{\mu_{\mathrm{W}}}\right)^{0.23} \quad (32)$$

The constant ξ changes linearly from $0.2(T/D = 3)$ to $0.37(T/D = 1.9)$.

Eq. (32) is used to determine the internal heat transfer coefficient ($\alpha_{\mathrm{pw-g}}$) under gassed conditions in a jacketed stirred tank. After appropriate corrections, it can also be used to approximate the process side heat transfer coefficient ($\alpha_{\mathrm{p-g}}$) in the presence of an internal heat exchanger. A multiplying correction factor, the ratio of measured heat transfer coefficients under ungassed conditions ($\alpha_{\mathrm{p-ug}}/\alpha_{\mathrm{pw-ug}}$), is used as indicated in Eq. (33). In this case, $\alpha_{\mathrm{p-ug}} = \alpha_{\mathrm{pw-ug}}$.

$$\alpha_{\mathrm{p-g}} = \alpha_{\mathrm{pw-g}}\frac{\alpha_{\mathrm{p-ug}}}{\alpha_{\mathrm{pw-ug}}} \quad (33)$$

The biggest challenge in using Eq. (28) is the lack of reliable values of the overall heat transfer coefficient U. It depends on a number of factors including rheology of the broth, thermal properties of the broth and utility fluid, aeration and agitation levels, and the nature of the heat transfer surfaces (e.g., jacket, dimpled jacket, coils, or tube bundles). Because so few applicable correlations exist for aerated fermenters, experience and a conservative approach are essential for their design. Design guidelines for a typical half-pipe jacket are discussed here. Based on the fermentation heat duty (see Table 9), half-pipe coil design results are summarized in Table 10.

Heat transfer in a half-pipe jacket

According to Stein and Schmidt a few adaptations make it possible to use correlations for horizontal tube coils to estimate heat transfer in a vessel with a half-pipe jacket.[17] Necessary adaptations include using the thermic diameter $d_{\mathrm{th}} = (\pi/2)d_{\mathrm{i}}$ instead of the tube inner diameter d_{i} to calculate the Reynolds and Nusselt numbers, and replacing the bending ratio ($d_{\mathrm{be}}/D_{\mathrm{b}}$) with the ratio of $d_{\mathrm{i}}/2(T + 2\delta_{\mathrm{w}})$.[1,2]

The geometry with the main dimensions of the coil is shown in Fig. 11. Here, d is the inside diameter of the coil tube and h is the pitch of the coil. D_{w} is the average diameter of the coil.

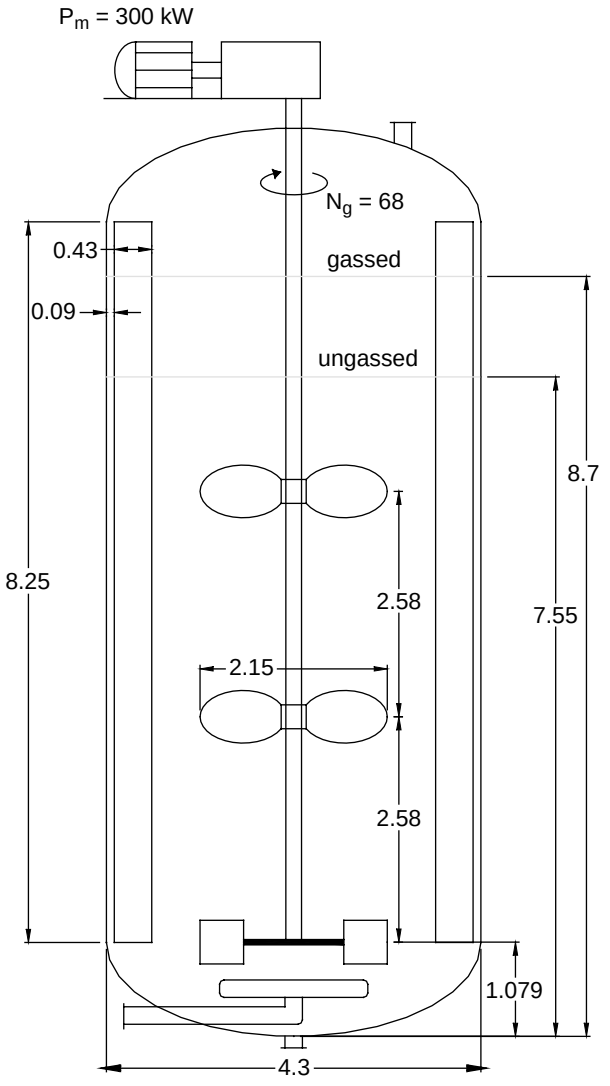


Fig. 9 Schematic of the production fermenter (dimensions in meters). *(View this art in color at www.dekker.com.)*

Another characteristic parameter used to describe the coil geometry is the bending ratio $d_{\mathrm{hc}}/D_{\mathrm{b}}$, with D_{b} as the average bending diameter of the coil. The calculation of D_{b} and that of the heat transfer coefficient are described next.

The average turning diameter D_{t} of a coil can be determined for a tube with length l, and n turns of slope h:

$$D_{\mathrm{t}} = \frac{l}{n\pi} \quad (34)$$

With a right triangle configuration of the coil, the average diameter of the coil (shown in Fig. 11) is calculated to be

$$D_{\mathrm{w}} = \sqrt{D_{\mathrm{t}}^2 - \left(\frac{h}{\pi}\right)^2} \quad (35)$$

Table 8 Gas sparger design results

Variable	Equation	Value
Diameter of sparge ring (m)		1.7
Sparge pipe inside diameter (m)		0.20
Sparger to bottom impeller diameter ratio		0.79
Sparger bottom clearance (m)		0.55
Number of holes		570
Hole diameter (mm)		8.00
Velocity through each orifice (m/sec)		36.55
Number of drain holes		16
Size of drain holes (mm)		16
Pressure drop per orifice (Pa)	21	4500
Pressure drop in the pipe ($f = 0.005$) (Pa)	23	127
Sparge pipe velocity (m/sec)		16.03
Sparge feed pipe inside diameter (m)		0.33
Sparge feed pipe velocity (m/sec)		11.86
Pressure drop ratio pipe to hole	22 (check)	0.03

The average bending diameter D_b of the coil that is used in calculating the bending ratio d_{hc}/D_b is obtained from

$$D_b = D_w\left[1 + \left(\frac{h}{\pi D_w}\right)^2\right] \tag{36}$$

In contrast to a straight pipe, the critical Reynolds number for the transition from laminar to turbulent flow in a coiled pipe increases with increasing bending ratios. For the critical Reynolds number, Schmidt found that

$$Re_{crit} = 2300\left[1 + 8.6\left(\frac{d_{hc}}{D_b}\right)^{0.45}\right] \tag{37}$$

Based on the value of the critical Reynolds number or flow regime, different heat transfer correlations are applied. For laminar flow with $Re < Re_{crit}$, the Nusselt number is correlated to:

$$Nu_{lam} = \left(3.66 + 0.08\left[1 + 0.8\left(\frac{d_{hc}}{D_b}\right)^{0.9}\right]Re^m Pr^{1/3}\right)\left(\frac{Pr}{Pr_w}\right)^{0.14} \tag{38}$$

with

$$m = 0.5 + 0.2903\left(\frac{d_{hc}}{D_b}\right)^{0.194} \tag{39}$$

Here, previous definitions have been slightly altered as follows: $Nu = \alpha_c d_{hc}/\lambda_{cool}$, $Re = w d_{hc}\rho_{cool}/\mu_{cool}$, and $Pr = \mu_{cool} c_{p,cool}/\lambda_{cool}$.

For heat transfer in the turbulent flow regime, i.e., $Re > 2.2 \times 10^4$, the Nusselt number is determined from the following equation:

$$Nu_{turb} = \frac{\xi/8 Re \times Pr}{1 + 12.7\sqrt{\xi/8}(Pr^{2/3} - 1)}\left(\frac{Pr}{Pr_w}\right)^{0.14} \tag{40}$$

with

$$\xi = \frac{0.3164}{Re^{0.25}} + 0.03\left(\frac{d_{hc}}{D_b}\right)^{0.5} \tag{41}$$

The installed area of half-pipes can be calculated as:

$$A_{inst} = n_{coil} d_o L_{coil} = n_{coil} d_o n_{tc}\pi(T + 2\delta_w) \tag{42}$$

where d_o is the outside diameter of the half-pipe.

The effective area for heat transfer is linearly proportional to the gassed height of the cylindrical portion in the tank:

$$A_{eff} = A_{inst}[(H_g - H_{btm})/n_{coil} h_{coil}] \tag{43}$$

The heat removal duty from the tank is:

$$\dot{Q} = UA_{eff}\Delta t_{ln} = UA_{eff}\frac{(t - t_1) - (t - t_2)}{\ln((t - t_1)/t - t_2)} \tag{44}$$

$$\dot{Q} = \dot{m}_{cool} c_{p,cool}(t_2 - t_1) \tag{45}$$

The cooling medium volumetric flow is:

$$\dot{V}_{cool} = w\left(\frac{\pi}{4} d_i^2 n_{coil}\right) \tag{46}$$

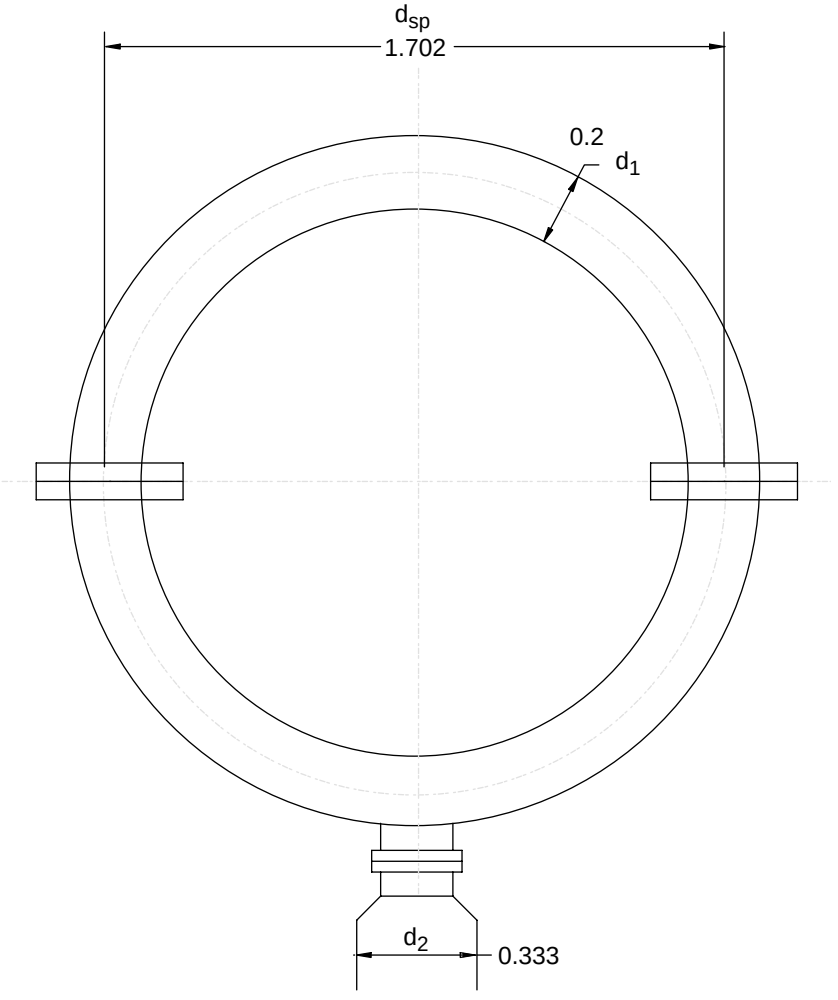


Fig. 10 Schematic of sparger ring (dimensions in meters). *(View this art in color at www.dekker.com.)*

where w (assumed 2 m/sec) is the velocity of the utility fluid in the half-pipe. The cooling medium return temperature t_2 is:

$$t_2 = \frac{UA_{\text{eff}}\Delta t_{\text{ln}}}{\dot{m}_{\text{cool}}c_{\text{p,cool}}} + t_1 \tag{47}$$

Concentric Tube Airlift Fermenter

This, the second of two fermenter design examples, deals with a concentric tube airlift fermenter. Nondimensional parameters applicable in this example are defined as follows.

Nondimensional parameters

R	diameter ratio, d_c/d_t
S	slenderness ratio, l_t/d_c
M	bubble separation group, $d_s/4d_c$
X_t	draft-tube bottom clearance ratio, $c_{t\text{-}b}/d_t$
Y_t	draft-tube top clearance ratio, $c_{t\text{-}t}/d_t$
Φ_g	gas holdup, $V_g/(V_g + V_L)$
DR	disengagement ratio, $(d_c/d_s)\ \{(R^2 - 1)/R\}$
Eu'	modified Euler number, $\Delta P/\rho u_g^2$
Fr	Froude number, $u_g/(gd_c)^{0.5}$ [also defined as $u_g^2/(gd_c)$]
Ga	Galileo number, $g\rho_L^2 d_c^3/\mu^2$
Gr	Grashoff number, $g\rho_L(\rho_L - \rho_g)d_b^3/\mu_L^2$
Sc	Schmidt number, $\mu_L/(\rho_L D_L)$
Sh	Sherwood number, $k_L a d_c^2/D_L$
We_o	Weber number for the orifice, $v_o^2 d_o \rho_o/\sigma$

Mass transfer correlations

The following nondimensional mass transfer correlation, represented by the Sherwood number, is valid for the draft-tube sparged concentric tube airlift fermenter.[18]

$$Sh = 17.3 \times 10^3 Fr^{0.9} M^{-3.4} Ga^{0.13} X_t^{-0.07} Y_t^{-0.18} \tag{48}$$

The influence of apparent viscosity, represented by Ga, reflects the expected decrease in mass transfer rate with

Table 9 Fermentation heat duty

Variable	Equation	Value
Agitation power (kW)	From Table 7	238
Gas expansion power (kW)	From Table 7	80
Reaction heat liberation (kJ/mol O_2)		500
Reaction heat liberation (kW)	27	2676
Total heat to be removed (MW)	26	3

an increase in viscosity. Both draft-tube bottom and draft-tube top clearances, i.e., X_t and Y_t affect the mass transfer rate only modestly while the bubble separator to column diameter ratio, M, is obviously ascribed the greatest influence.

Gas holdup correlations

The following correlations are recommended for calculating gas holdup in the riser, downcomer, and separator, respectively.

$$\Phi_r = 1.5Fr^{0.87}M^{-0.4}X_t^{-0.19}Y_t^{-0.2} \quad (49)$$

$$\Phi_d = 4.76Fr^{1.3}Ga^{-0.09}M^{-3.8}X_t^{0.65} \quad (50)$$

$$\Phi_{to} = \Phi_s = 0.29Fr^{1.05}M^{-2.5}X_t^{0.1}Y_t^{-0.07} \quad (51)$$

The main variable is Fr, which represents the influence of gas input rate (and also the energy input rate). Its effect on holdup is much stronger in the downcomer than in the riser, and the balance between the two gives an exponent of almost 1 (1.05) for the total holdup. Interestingly, the behavior of holdup in the gas separator is very close to that of total holdup. As draft-tube bottom clearance c_{t-b} increases, Φ_r decreases.[18]

Dynamic pressure drop correlations

According to Merchuk et al., the following correlation is satisfactory for the prediction of pressure drop in

Table 10 Results of half-pipe external heat transfer jacked design

Variable	Equation	Value
Half-pipe jacket design		
Vessel wall thickness (m)		0.032
Thermal conductivity tank wall (W/m/K)		16
Number of jacket coils		3
Nominal tube size (m, in.)		0.1, 4
Tube outside diameter (m)		0.114
Schedule		Sch. 10
Tube thickness (m)		0.003
Tube inside diameter (m)		0.108
Number of coil turns		16
Coil tube pitch (m)		0.149
Height of coil section (m)		2.377
Distance between coil section (m)		0.114
Total surface area of half pipes (m^2)		71.20
Velocity (m/sec)		2.0
Total chilled water (CW) flow (m^3/hr)	46	100.0
Reynolds number		2.94×10^5
Zeta	41	0.0183
Nusselt number	40	1,922
Process side heat transfer coefficient ($W/m^2/K$)	31–33	1,764
Utility side heat transfer coefficient ($W/m^2/K$)		10,608
Fouling process side (m^2 K/W)		0.00000
Fouling utility side (m^2 K/W)		0.00017
Overall heat transfer value ($W/m^2/K$)	30	355
Process side temperature (°C)		32
Log mean temperature difference (K)	28, 29	18
Effective heat transfer area (m^2)	42, 43	71.2
Actual heat duty (MW)	26	0.455
CW supply temperature (°C)		12.0
CW return temperature (°C)	43, 44, 46, 47	15.9

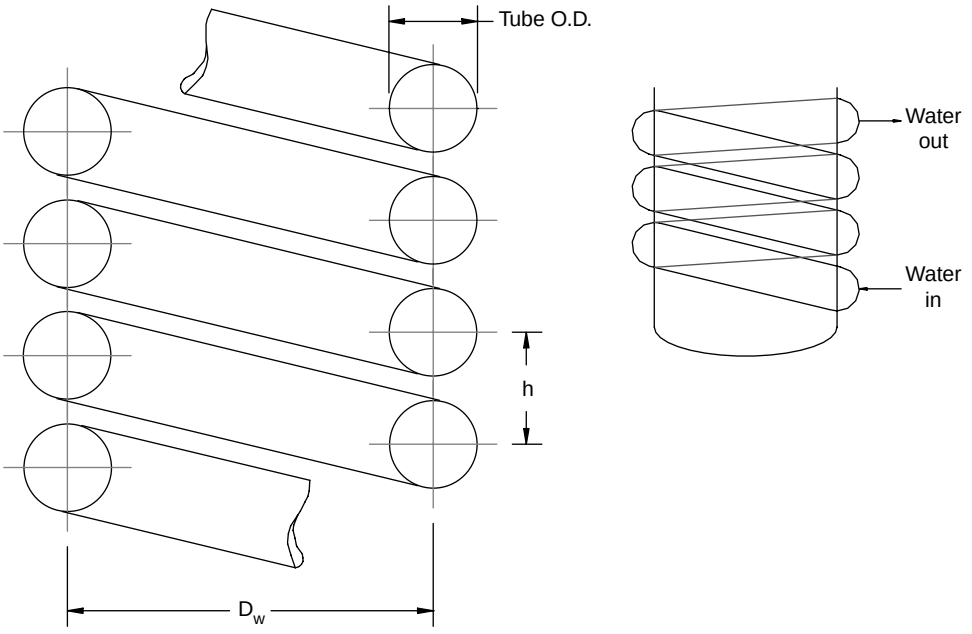


Fig. 11 Schematic of half-coil jacket. *(View this art in color at www.dekker.com.)*

concentric tube airlift fermenters over a wide range of physicochemical, geometrical, and operating variables.[18]

$$Eu' = 0.858 Fr^{-1.51} Ga^{0.06} X_t^{-1.1} M^{4.2} \quad (52)$$

Substituting the definitions of Eu' and Fr into this expression shows pressure drop to be a function of $u_g^{0.5}$. The bottom clearance ratio, X_t, and the bubble separation group, M, have a strong influence on the pressure drop because of their effect on liquid velocity.

Note that all these correlations (mass transfer, gas holdup, and dynamic pressure drop) are valid only within the following operating ranges:

$$6 \times 10^{-4} < Fr < 3.5 \times 10^{-2}$$
$$3 \times 10^{7} < Ga < 6 \times 10^{11}$$
$$0.04 < X_t < 0.4$$
$$0.22 < M < 0.34$$

Shear rate correlations

The rate of shear in the airlift bioreactor remains the subject of speculation. The disparity in "average shear rate'' ($\dot{\gamma}$), as calculated with several available correlations, is shown by Chisti.[8] Several investigators (Nakanoh and Oshida) have accepted the following conservative expression for use in airlift bioreactor:[19]

$$\gamma = 5000\, u_{gr} \text{ for } 0.04 \leq u_{gr} \leq 0.1\,\text{m/sec} \quad (53)$$

However, Schumpe and Deckwer and Moo-Young et al. proposed the following equations, respectively, for average shear rate:[20,21]

$$\dot{\gamma} = 2800 u_{gr} \quad \text{for } u_{gr} \leq 0.08\,\text{m/sec} \quad (54)$$

$$\dot{\gamma} = 1000 u_{gr} \quad \text{for } u_{gr} \leq 0.04\,\text{m/sec} \quad (55)$$

These expressions show shear rate dependence on superficial gas velocity, but not as a function of specific energy, which is responsible for turbulence. In an airlift bioreactor, pneumatic power input is given by the following equation:[22]

$$P_g/V_L = \rho_L g u_{gr} [a_r/(a_r + a_d)] \quad (56)$$

Incorporating this definition of pneumatic power into Eqs. (53) and (55) produces the following modified

Table 11 Maximum gas flow rates for separate bubbling

Orifice diameter, $d_o \times 10^3$ (m)	Orifice area, $a_o \times 10^6$ (m^2)	Gas rate/orifice, $q_o \times 10^6$ (m^3/sec)
0.25	0.05	0.275
0.51	0.20	0.471
1.02	0.82	0.864
2.54	5.06	1.807
5.08	20.26	3.300
10.16	81.03	6.679

and corrected shear rate correlations:[23]

$$\dot{\gamma} = [5000\,P_g/(V_L\rho_L g)][(a_r + a_d)/a_r] \quad \text{for } u_{gr} \geq 0.04\,\text{m/sec} \tag{57}$$

$$\dot{\gamma} = [1000\,P_g/(V_L\rho_L g)][(a_r + a_d)/a_r] \quad \text{for } u_{gr} < 0.04\,\text{m/sec} \tag{58}$$

Sparger design equations

For satisfactory gas distribution from a ring sparger,[24] Weber number for the orifice,

$$We_o = \left(v_o^2 d_o \rho_o/\sigma\right) > 2 \tag{59}$$

In the case of gas emerging from a simple submerged orifice at very low flow rates, bubbles periodically form at the orifice, grow to a certain size, and break away. For media with a reasonably low viscosity, bubble diameter is a function of the interfacial tension and density as indicated below:

$$d_b = 1.82\left[\frac{d_o\sigma g_c}{(\rho_L - \rho_g)g}\right]^{1/3} \tag{60}$$

This relationship holds for "slow" bubbling in the "constant volume" region, although the exponent may be nearer $\frac{1}{4}$ when static surface tension data are

Table 12 Airlift fermenter design specification

Variable	Equation	Value
Column diameter (m)		1
Unaerated liquid height (m)		5
Ratio of column to draft-tube area		2
Draft-tube diameter (m)		0.707
Slenderness ratio		4.5
Draft-tube length (m)		4.5
Diameter of gas separator (m)		1.25
Disengagement ratio		0.566
Draft-tube bottom clearance ratio		0.25
Draft-tube top clearance ratio		1.25
Bubble separation group		0.313
Type of bottom head		ASME flange-dished head
Nominal unaerated media volume (m^3)		4.0
Aeration rate/media volume (VVM) (min^{-1})		0.5
Actual aeration rate (m^3/min)		2.0
Superficial gas velocity (m/sec)		0.043
Mass transfer efficiency		0.80
Required mass transfer coefficient (sec^{-1})	1	0.0028
Superficial gas velocity (m/sec)		0.043
Pressure drop (Pa)		171
Nondimensional parameters		
Sherwood number		6.5×10^5
Galileo number		3.9×10^{11}
Froude number	48	0.014
Schmidt number		1.2×10^3
Grashoff number		4.4×10^4
Bond number		1.4×10^5
Euler number	52	93
Gas holdup		
In the riser	49	0.072
In the downcomer	50	0.055
In the airlift fermenter	51	0.051
Aerated height (m)		5.4
Specific power input (W/m^3)	56	417
Shear rate in the riser (sec^{-1})	57	425

used. Bubble size is independent of flow rate while bubble frequency is proportional to it. At higher gas flow rates, the frequency levels off and bubble volume increases. Finally, a point is reached where distinct bubble formation is replaced with a chain bubbling process. This undesirable phenomenon is avoided by keeping the overall gas rates below those shown in Table 11.[24] The turbulent churn flow regime typically sets in as superficial gas velocity, u_g, exceeds 0.08 m/sec.

As cited by Bhavaraju, Russel, and Blanchard, Davidson obtained the following correlation for bubble diameter using orifices ranging in diameter from 0.001 to 0.01 m:[25]

$$d_b = 0.19 d_o^{0.48} Re_o^{0.32} \quad (61)$$

Design example

The design of a concentric tube internal loop airlift fermenter for the production of a biocatalysis enzyme at an annual rate of 8000 ± 5% kg/yr is illustrated below. Product recovery efficiency is 80% and the titer (product concentration) is 20 g/L. Maximum oxygen uptake demand is 10 mmol O_2/L/hr. Operational parameters and media physical properties are as follows:

Batch length or growth rate = 5 days
Bioreactor turnaround rate = 2 days
Number of operating days in a year = 300.
Annual fermenter output = required production/recovery efficiency = 10,000 kg/yr.
Required batch size = (annual fermenter output/titer) (total batch duration/yearly operation) = 11,667 L = 12,000 L.
Number of bioreactors = 3 (assumed).
Size of each bioreactor = 4000 L.

Broth properties of Newtonian fluid

Effective viscosity = 5×10^{-3} Pa sec (glucose media Newtonian).
Density = 1000 kg/m^3.
Interfacial tension (gas–liquid) = 7.2×10^{-2} N/m.
O_2 broth diffusion coefficient = 4.3×10^{-9} m^2/sec.

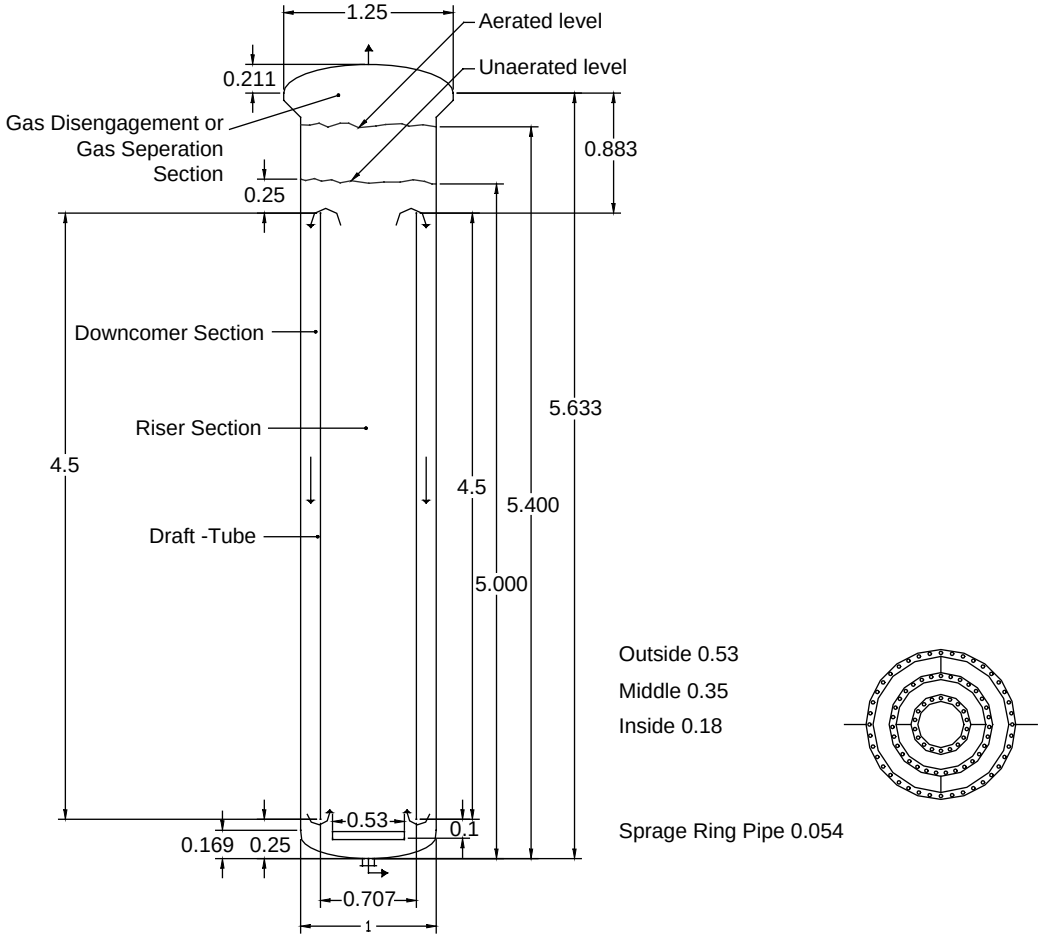


Fig. 12 Schematic of the airlift fermenter and its triple ring sparger (dimensions in meters). *(View this art in color at www.dekker.com.)*

Table 13 Sparger design results

Variable	Equation	Value
Type: three-ring air sparger		
Actual volume of gas (m^3)		0.23
Ratio of column to draft-tube area		2
Superficial gas velocity in the riser (m/sec)		0.085
Number of orifices		450
Orifice diameter (m)		2.54×10^{-3}
Velocity through orifice (m/sec)		14.6
Orifice Weber number	59 (check)	9
Total cross-sectional area of the orifices (m^2)		2.28×10^{-3}
Minimum sparger pipe diameter (m)		0.054
Orifice velocity coefficient		0.8
Pressure drop across the orifice (Pa)	21	201
Bubble size (noncoalescing media) (m)	61	0.005
Orifice Reynolds number		2.5×10^3
Outer sparger diameter (m) and number of orifices		0.53, 225
Middle sparger diameter (m) and number of orifices		0.35, 150
Inner sparger diameter (m) and number of orifices		0.18, 75

Based on the above fermentation volume, important fermenter dimensions are initially estimated. The required mass transfer coefficient is calculated from the oxygen uptake rate (OUR). Superficial gas velocity is computed from the Sherwood number–Froude number correlation [Eq. (48)]. Final column dimensions are determined iteratively while ensuring that bubbly flow (as indicated by superficial gas velocity) is maintained within the riser. After finalizing column dimensions, the required volume is calculated and the sparger is designed according to accepted rules of thumb. These and other details of the iterative solution procedure are tabulated in Table 12. Fig. 12 shows a schematic of the production scale concentric tube airlift fermenter. The detailed design specification for the triple-ring sparger is presented in Table 13.

EMERGING UNCONVENTIONAL FERMENTERS

The yield of secondary metabolites in a large-scale fermenter typically ranges from 0.1 to 10 g/L of broth. Such poor yield leads to cumbersome and expensive processes for both product separation and broth disposal. Within the last decade, several novel bioreactors have been developed for the intensification of fermentation processes. Examples include a centrifugal bioreactor, a rotating packed bed fermenter, and a sonobioreactor.[26,27] Most of these, however, are yet to be implemented on a production scale because they generally lack practicality and well-defined scale-up criteria.

CONCLUSIONS

Even though the bioreactor itself may not be the most costly piece of process equipment, it determines the technical feasibility of the entire process. The attention paid to bioreactor design, scale-up, and operating parameters reflects its central role in successful commercial ventures.

This chapter discusses important controlling parameters of aerobic fermenters, and provides design methodology and appropriate correlations for commonly practiced mechanically agitated and concentric tube airlift fermenters. Two complete design examples for the production of biocatalysis enzymes are presented. Aspects that were not discussed in detail include mass transfer and gas holdup correlations for high-viscosity or pseudoplastic fluids. In this regard, readers are referred to the open literature, which includes such works as those of Kilonzo and Margaritis and Kawase and coworkers[4,28–30]

NOMENCLATURE

a	Cross-sectional area (m^2)
c	Clearance (m)
d	Diameter (m)
D	Impeller diameter (m)
$\mathscr{D}$	Molecular diffusion coefficient (m^2/sec)
f	Wall heat transfer fouling resistance (m^2 K/W)
g	Acceleration due to gravity (m/sec^2)

h Pitch (m)
H Height (m)
K k-factor (–)
l Height/length (m)
L Length (m)
m Mass flow rate (kg/sec)
n Number (–)
$\dot{n}$ Mole flow rate (kmol/sec)
N Impeller speed (sec^{-1})
p Pressure (Pa)
P Power (W)
$\dot{Q}$ Heat flux rate (W)
s Impeller spacing (m)
t Temperature (°C)
T Tank diameter (m)
u Superficial gas velocity (m/sec)
$\bar{u}$ Average superficial gas velocity (m/sec)
U Overall heat transfer coefficient ($W/m^2/K$)
V Volume (m^3)
$\dot{V}$ Volumetric flow rate (m^3/sec)
w Coolant velocity in the pipe (m/sec)

Greek Letters

α Heat transfer coefficient ($W/m^2/K$)
$\dot{\gamma}$ Average shear rate (sec^{-1})
ε Power per mass (W/kg)
ε_g Gas power per volume (W/m^3)
ρ Density (kg/m^3)
Φ_g Gas holdup in the fermenter (–)
μ Dynamic viscosity (Pa sec)
η Efficiency (–)
σ Interfacial tension (N/m)
ν Kinematic viscosity (m^2/sec)
λ Thermal conductivity (W/m/K)
δ Tank wall thickness (m)
τ Shear stress (Pa)

Subscripts

A Axial impeller
b Bubble
btm Bottom
c Column
cool Coolant
g Gassed
gas Gas expansion
hc Horizontal coil
loc Local
L Liquid
m Motor
ug Ungassed
r Riser
R Radial impeller
d Downcomer
susp Suspension (slurry)
t Draft tube
t-t Draft-tube top
t-b Draft-tube bottom
tc Turns per coil
s Gas separator
to Total
rs Ring sparger
sp Sparger pipe
rsi Inner ring sparger
rsm Middle ring sparger
rso Outer ring sparger
o Sparger orifice
p Process side
j Jacket side
w Tank wall
i Inner
in Inlet
inst Installation
eff Effective

ACKNOWLEDGMENTS

The authors would like to thank the Dow Chemical Company for permitling external release of the information in this chapter, Erik Gabiewski for CAD drawings in the figures, and Dr. Lawrence Chew for the technical review.

REFERENCES

1. Storhas, W. *Bioreaktoren und Periphere Einrichtungen*; Vieweg Verlag: Wiesbaden, 1994; 657.
2. Krishnan, M.S.; Du, J.X.; Cao, N.J.; Gong, C.S.; Tsao, G.T. Simultaneous production and recovery of fuel ethanol by gas stripping in a novel airlift loop fermenter; improved biofuel production by corn fiber simultaneous saccharification and fermentation in an airlift-loop fermenter with side arm. In 213th National Meeting of the American Chemical Society, San Francisco, CA, Apr 13–17, 1997; American Chemical Society: Washington, DC, 1997; BIOTECH-239.
3. Atkinson, B.; Mavituna, F. Gas-liquid mass transfer. In *Biochemical Engineering and Biotechnology Handbook*; The Macmillan Publishers Ltd: U.K., 1983; 801.
4. Kilonzo, P.M.; Margaritis, A. The effect of non-Newtonian fermentation broth viscosity and small bubble segregation on oxygen mass transfer in gas-lift bioreactors: a critical review. Biochem. Eng. J. **2004**, *17*, 27–40.
5. Bartow, M.V. Supersizing the aerobic fermenter. Chem. Eng. **1999**, Jul, 70–75.

6. Bartow, M.; Erin, S. Bioreactor design for mammalian cell cultures. Chem. Eng. **2004**, Jan, 49–54.
7. Govier, G.W.; Aziz, K. The flowproperties of fluids. In *The Flow of Complex Mixtures in Pipes*; Robert E. Krieger Publishing Company: Malbar, 1982; 98.
8. Chisti, M.Y. Airlift reactor: current technology. In *Airlift Bioreactors*; Elsevier: Amsterdam, 1989; 74–79.
9. Moo-Young, M.; Chisti, Y. Considerations for designing bioreactors for shear sensitive culture. Biotechnology **1988**, Nov, 1291–1296.
10. Ryu, D.Y.; Oldshue, J.Y. A re-assessment of mixing cost in fermentation process. Biotech. Bioeng. *XIX* **; 1977**, 621–629.
11. Bakker, A.; Smith, J.M.; Myers, K.J. How to disperse gases in liquids. Chem. Eng. **1994**, Dec, 98–104.
12. Rehm, H.J.; Reed, G.; Puehler, A.; Stadler, P. Design of aseptic, aerated fermentors. In *Biotechnology*; VCH Verlagsgesellschaft: Weinheim, 1993; Vol. 3, 418–425.
13. Weiss, S. Stoffvereinigen in fluiden phasen. In *Verfahrenstechnische Berechnungsmethoden*; VCH Verlagsgesellschaft: Weinheim, 1988; 186.
14. Lee, S.; Tsui, Y. Succeed at gas/liquid contacting. Chem. Eng. Prog. **1999**, Jul, 23–49.
15. Strek, F. Heat transfer in liquid mixers—study of a turbine agitator with six flat blades. Int. Chem. Eng. **1963**, *3*, 533–556.
16. Kurpiers, P.A.; Steiff, P.M. Weinspach: zum wärmeübergang in gerührten mehrphasensystemen. Chem. Ing. Techn. **1984**, *56* (3), 234–235.
17. Stein, W.A.; Schmidt, W. Heat transfer in agitated tank jackets with liquid media. Ger. Chem. Eng. **1986**, *9* (6), 362–371.
18. Merchuk, J.C.; Ladwa, N.; Cameron, A.; Bulmer, M.; Pickett, A. Concentric-tube airlift reactors: effects of geometrical design on performance. AIChE J. **1994**, *40* (7), 1105–1117.
19. Nakanoh, M.; Yoshida, F. Gas absorption by Newtonian and non-Newtonian liquids in a bubble column. Ind. Eng. Chem. Process Des. Dev. **1980**, *19*, 190–195.
20. Schumpe, A.; Decker, W.D. Viscous media in tower bioreactors: hydrodynamic characteristics and mass transfer properties. Bioprocess. Eng. **1987**, *2*, 79–94.
21. Moo-Young, M.; Halard, B.; Allen, D.G.; Burrell, R.; Kawase, Y. Oxygen transfer to mycelial fermentation broth in an airlift fermenter. Biotechnol. Bioeng. **1987**, *30*, 746–753.
22. Chisti, Y.; Moo-Young, M. Communication to the editor on the calculation of shear rate and apparent viscosity in airlift and bubble column bioreactors. Biotechnol. Bioeng. **1989**, *34*, 1391–1392.
23. Kilonzo, P.M.; Margaritis, A. The effects of non-Newtonian fermentation broth viscosity and small bubble segregation on oxygen mass transfer in gas-lift bioreactors: a critical review. Biochem. Eng. J. **2004**, *17*, 27–40.
24. Fair, J.R. Designing gas-sparged reactors. Chem. Eng. **1967**, Jul, 67–74.
25. Bhavaraju, S.M.; Russel, T.W.F.; Blanchard, H.W. Design of gas sparged devices for viscous liquid system. AIChE J. **1978**, *24* (3), 454–466.
26. Voit, H.; Goetz, F.; Mersmann, A.B. Overproduction of lipase with Staphylococcus carnosus (pLipPS1) under modified gravity in a centrifugal field bioreactor. Chem. Eng. Technol. **1989**, *12*, 364–373.
27. Chisti, Y. Sonobioreactors: using ultrasound for enhanced microbial productivity. Trends Biotechnol. **2003**, *21* (2), 89–93.
28. Kawase, Y.; Moo-Young, M. Volumetric mass transfer coefficients in aerated stirred tank reactors with Newtonian and non-Newtonian media. Chem. Eng. Res. Des. **1988**, *66*, 284–288.
29. Kawase, Y.; Moo-Young, M. Theoretical prediction of gas hold-up in bubble columns with Newtonian and non-Newtonian fluids. Ind. Eng. Chem. Res. **1987**, *26*, 933–937.
30. Kawase, Y.; Hallard, B.; Moo-Young, M. Theoretical prediction of volumetric mass transfer coefficients in bubble columns for Newtonian and non-Newtonian fluids. Chem. Eng. Sci. **1987**, *42* (7), 1609–1617.

Fluid Flow

T. Reg. Bott
School of Engineering, Chemical Engineering, University of Birmingham, Birmingham, U.K.

INTRODUCTION

Most processes involve the flow of fluids (gases and liquids) either as material being processed or as an agent used to transport particles or droplets that are themselves being processed in the equipment making up the plant. Because the fluid is contained within the system in pipe lines, mixers, heat exchangers, reactors, and so on, the confining walls generally influence the character of the flow. Furthermore, the physical properties of the fluid, particularly the viscosity, also have a profound influence on the behavior of the fluid in the process plant.

The movement of the fluid on its journey through the various items of equipment requires energy that usually, although not always, is applied through some sort of pump. The amount of energy involved will again be dependent on the physical properties of the fluid and the design of the equipment through which it flows. A major cause of energy dissipation is through friction, either within the fluid itself or through its contact with the solid confining wall. It is essential that, at the process plant design stage, a reliable estimate of the likely pressure loss can be made, so that the pump capacity and power requirement can be calculated. The energy consumption involved could represent a substantial constituent of the future plant operating costs. The quality of the fluid flow can also have a profound effect on the efficiency of the various pieces of equipment operating in the process plant. For instance, the efficiency of heat transfer is strongly dependent on the characteristics of the flow within the heat exchanger involved. In distillation and absorption columns effective mass transfer depends on the characteristics of the fluid flow within the equipment.

There are two definitions that are usually applied to fluids. They are ''compressible'' and ''incompressible'' fluids. If the volume of a fluid is not dependent on the applied pressure and its temperature, the fluid is said to be incompressible. Although not strictly true, it is usual to assume that liquids are incompressible, when estimates involving fluid flow are made. On the other hand gases are usually considered to be compressible, because large changes in volume are apparent with changes in pressure and temperature. For calculation simplification, however, for small changes in pressure and temperature, a gas can often be regarded as incompressible.

FLOW IN RELATION TO SURFACES

When a fluid flows across a solid surface, the drag (the resistance to flow) produced by the presence of the surface in contact with the fluid creates a velocity distribution through the bulk fluid. The fluid velocity increases with the distance from the solid surface. The flow may be visualized by so-called streamlines that represent the paths of parcels of fluid. Each streamline represents fluid with the same velocity. Fig. 1 shows the velocity profile with a fluid flowing across a solid surface. The difference in flow rate between adjacent streamlines is always the same when the streamlines are equidistant.

Fig. 2 shows increased velocity around a submerged object by streamlines that are closer together compared to the region where there is no submerged object. The concept of streamlines can help to define two different flow regimes:

1. Laminar flow where movement of particles of fluid between streamlines is solely by molecular diffusion.
2. Turbulent flow in which the transfer of fluid in the direction perpendicular to the flow occurs as a result of the physical movement of fluid, although the time average of flow rate remains constant.

FLOW IN PIPES

Pioneering work by Osborne Reynolds in the late 19th century[1] added considerably to the understanding of fluid flow in relation to surfaces, and established concepts on which subsequent theoretical, empirical, and practical work could be based. The principal finding was in connection with fluid flow in pipes, visually demonstrating the difference between laminar and turbulent flow. Reynolds discovered that the dimensionless number that now bears his name, the Reynolds number (Re), defined the flow condition in a tube,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120019080

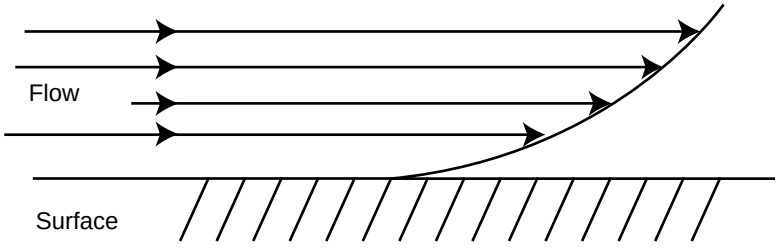


Fig. 1 Velocity profile of a fluid flowing across a surface.

where

$$Re = \frac{Dv\rho}{\mu} \quad (1)$$

where D represents the tube internal diameter, v is the fluid velocity, and ρ and μ are the fluid density and viscosity, respectively.

In general terms the Reynolds number can be considered a ratio of the forces acting on the fluid, i.e.,

$$\frac{\text{Momentum forces}}{\text{Viscous forces}}$$

It has been found that for values of the Reynolds number less than 2000 the flow regime is usually laminar. Turbulent conditions are usually encountered when the Reynolds number is greater than 4000. The transition from laminar to turbulent flow occurs gradually between these two values of Reynolds number. The point when it can be stated with confidence that the transition is complete will depend on the pipe geometry and the roughness of the pipe surface in contact with the flowing fluid. Exceptions to the establishment of turbulent flow at a Reynolds number of 4000 can occur if the fluid velocity in a given pipe is very gradually increased, say by a gradual reduction of the flow diameter of an exceptionally smooth surfaced tube. Under these conditions the streamline flow could be regarded as unstable, and a small disturbance in the flow conditions makes the system revert rapidly to turbulent flow.

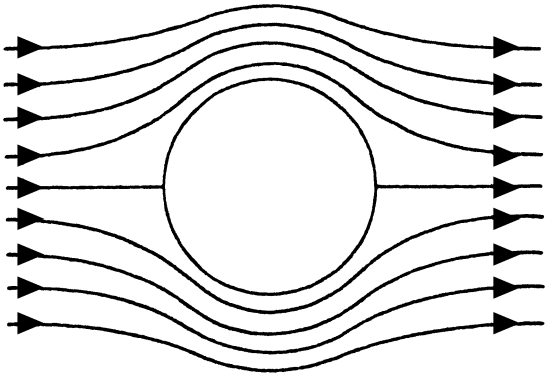


Fig. 2 Flow around a submerged object.

In pipe flow, the velocity profiles along a diameter for laminar and turbulent conditions are quite different as represented in Figs. 3 and 4, respectively. The profile can be seen to be much "flatter" for turbulent compared with laminar flow. It has to be stressed, however, that even under high turbulent conditions the fluid in direct contact with a confining surface is at rest, owing to the viscous drag on the fluid, and the adjacent layers of fluid will also be retarded in their flow. The velocity of the fluid in the neighborhood of a flat solid surface will change in a direction at right angles to the surface, increasing with distance from the solid surface. Under general turbulent conditions, eventually as the distance into the bulk of the fluid is increased, the fluid as a whole will be turbulent, but there remains a slow moving layer adjacent to the surface. Usually, this fluid layer is referred to as the "viscous sublayer." Because the mixing under laminar conditions at right angles to the flow direction is due to molecular diffusion alone, the viscous sublayer acts as a resistance to both mass and heat transfer. It will be appreciated that the magnitude of the resistance will depend on the thickness of the slow moving viscous sublayer, which itself is dependent on the flow conditions as defined by the Reynolds number. In general, the higher the Reynolds number of the bulk fluid, the thinner the viscous sublayer. Turbulent conditions are to be preferred where transport processes are required to be as effective as possible, i.e., increasing fluid velocity for a given pipe diameter and given fluid, although the choice of velocity has to be made in the light of the desired pressure drop through the equipment in any particular application.

The conditions referred to in the previous paragraph, however, are not instantly attained. As a fluid with a uniform velocity encounters a solid surface, a velocity gradient is established at right angles to the surface, owing to the influence of the viscous forces. The conditions referred to in the previous paragraph, however, are not immediately attained but a rapid change in velocity occurs near the surface, as illustrated in Fig. 5.

As the fluid continues its flow across the surface, the distance at right angles from the surface to where

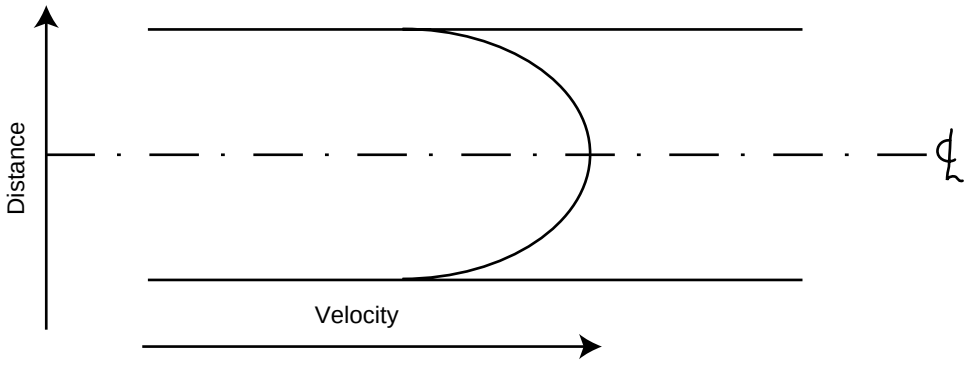


Fig. 3 Velocity profile of a fluid flowing in a tube—streamline flow.

the fluid can be considered to have reached the fully developed flow velocity v_f gradually increases. The locus of points where this occurs can be considered to divide the flow into two regions. In Fig. 5 the upper region can be considered to flow near or at the flow velocity v_f. The lower region where the velocity changes from 0 at the solid surface to the fully developed velocity v_f is called the "boundary layer." The mainstream velocity is approached asymptotically, and so it could be argued that the boundary layer has no exact limit. It can be assumed, however, that after a relatively short distance of flow across the surface, at the outer limit of the boundary layer, the fluid velocity has attained the stable flow velocity v_f. For this reason in most engineering calculations, for instance, the design of a tubular heat exchanger, the establishment of the boundary layer is not taken into account.

FLOW OF FLUIDS IN RELATION TO APPLIED STRESS

The viscosity of many fluids, μ, may be defined in terms of the shear stress τ and the rate of shear (dv/dy) where y is the distance from the surface. The relationship may be understood by reference to Fig. 6.

$$\tau = \mu \frac{dv}{dy} \tag{2}$$

$$\mu = \tau/(dv/dy) \tag{3}$$

For many fluids encountered in the process industries, the shear rate increases linearly with the shear stress over a wide range of shear and shear rate, and the viscosity is constant, as shown in Fig. 6, provided that there is no variation in temperature and pressure. Such fluids are usually referred to as "Newtonian fluids."

There are some industrial fluids that do not display the simple relationship between shear stress and rate of shear, that are classified as "non-Newtonian," and for which viscosity is not constant. Two examples are given in Table 1.

For some substances their viscosities may show time dependent changes. It may decrease with time under shear but return to its original value when the shear stress is removed. Such material is termed thixotropic. "Negative thixotropic" refers to substances that display the opposite behavior. Other fluids are solid-like and will not flow till some critical yield stress is exceeded. Such fluids are known as "viscoplastic."

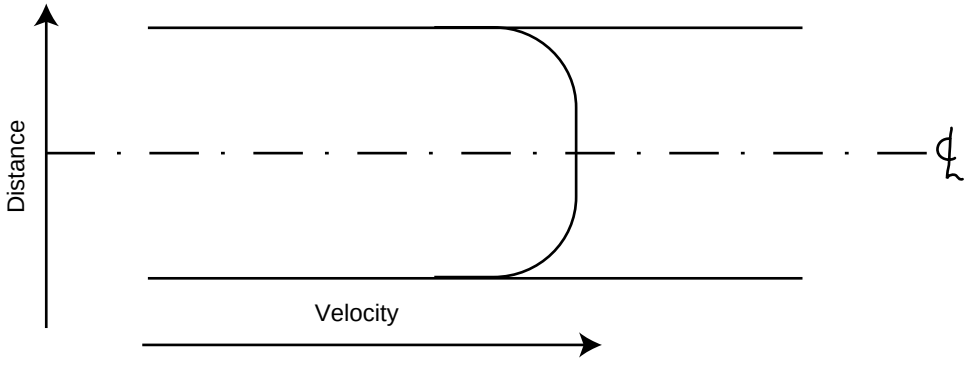


Fig. 4 Velocity profile of a fluid flowing in a tube—turbulent flow.

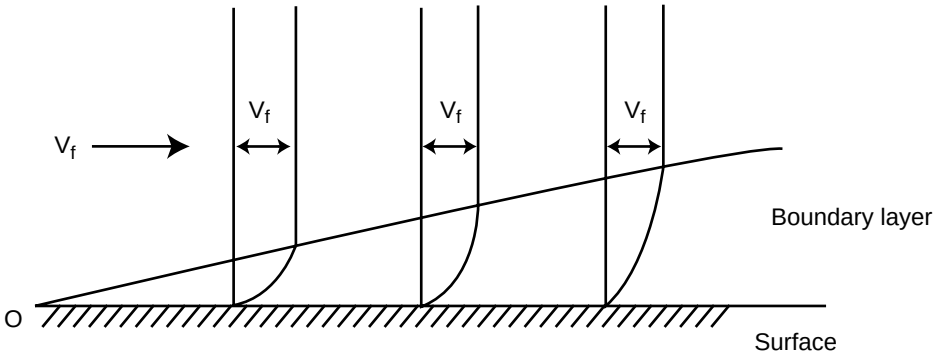


Fig. 5 The development of a boundary layer.

Other fluids display properties of elasticity and they are termed ''viscoelastic.''

The shearing characteristics of non-Newtonian fluids are illustrated in Fig. 7. Curves A and B represent viscoelastic behavior. Curve C illustrates the behavior if the fluid ''thins'' with increasing shear, generally referred to as ''shear thinning'' or ''pseudoplasticity.'' The opposite effect of ''shear thickening'' or ''dilatancy'' is shown as curve D.

PRESSURE LOSS THROUGH TUBES

As mentioned earlier, to estimate the energy requirements associated with a particular process plant in which fluids are moving, it is necessary to be able to make an estimate of the pressure loss through the equipment and in particular, through the linking pipe work. Research, in many respects as significant as that of Reynolds, was carried out by Stanton and Pannell early in the last century.[2] The results of the work provided the opportunity to make an estimate of the pressure loss due to friction, in a fluid flowing through a pipe. It involves the use of a dimensionless number plotted as a function of Reynolds number. The dimensionless number is termed the ''friction factor'' (Φ) defined as

$$\frac{R}{\rho v^2}$$

where R is the shear stress at the pipe wall.

A plot of $R/\rho v^2$ against $Dv\rho/\mu$ is presented in Fig. 8. There are two major regions on the diagram corresponding to laminar ($Re < 2000$) and turbulent ($Re > 4000$) flow conditions, with a transition region ($2000 < Re < 4000$) linking the two. It is apparent from Fig. 8 that in the turbulent region the surface roughness, as would be expected, affects the magnitude of the friction factor. In Fig. 8 it is taken into account by the dimensionless ratio e/D, where e is the height of the roughness and D the tube internal diameter. At very high Reynolds numbers the friction factor is independent of Reynolds number.

Making a force balance on the fluid flowing in a pipe of diameter D over a length L yields

$$\Delta P\left[\frac{\pi D^2}{4}\right] = R\pi DL \tag{4}$$

where ΔP is the loss of pressure. That is,

$$\Delta P = \frac{4RL}{D} \tag{5}$$

$$= \frac{4R}{\rho v^2}\frac{L}{D}\rho v^2 \tag{6}$$

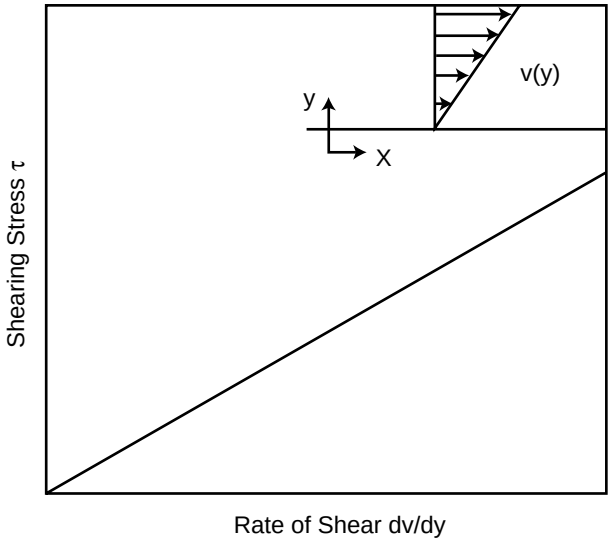


Fig. 6 The relationship between shearing stress τ and rate of shear (velocity gradient).

Table 1 Fluid viscosity change with increasing shear rate

Name	Viscosity
Pseudoplastic	Decreases
Dilatant	Increases

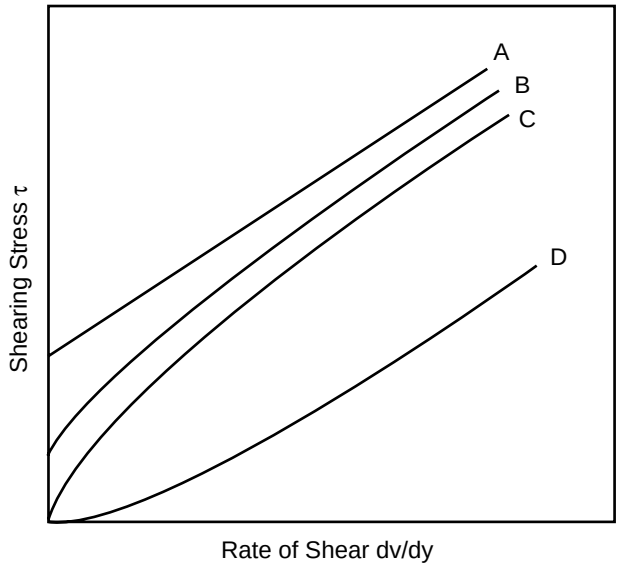


Fig. 7 The shearing characteristics of non-Newtonian fluids.

Assigning a value to e/D and calculating the Reynolds number, it is possible to obtain the value of $R/\rho v^2$ for inclusion in Eq. (6) with the known values of D, L, v, and the density of the fluid, ρ.

The pressure loss in terms of the height of the fluid (h_f) necessary to drive the fluid through the system at the stated velocity is given by:

$$h_f = \frac{\Delta P}{\rho g} = \frac{4R}{\rho v^2}\frac{L}{D}\frac{\rho v^2}{\rho g} \tag{7}$$

$$= 8\left[\frac{R}{\rho v^2}\right]\left[\frac{L}{D}\right]\left[\frac{v^2}{2g}\right] \tag{8}$$

The Bernoulli Equation

To calculate the pressure loss for a fluid flowing through a pipe work system, it is necessary to make an energy balance between the two points in the system between which pressure loss has to be determined. Neglecting forms of energy, such as magnetic or electrical energy, that are not generally associated with the flow of fluids, the energy contained in the system has three components:

1. Kinetic energy by virtue of the velocity of the fluid.
2. Potential energy due to the fluid being in the earth's gravitational field, i.e., the work that has to be done, against gravity, from some arbitrary datum (usually sea level) to transport it to its present position.

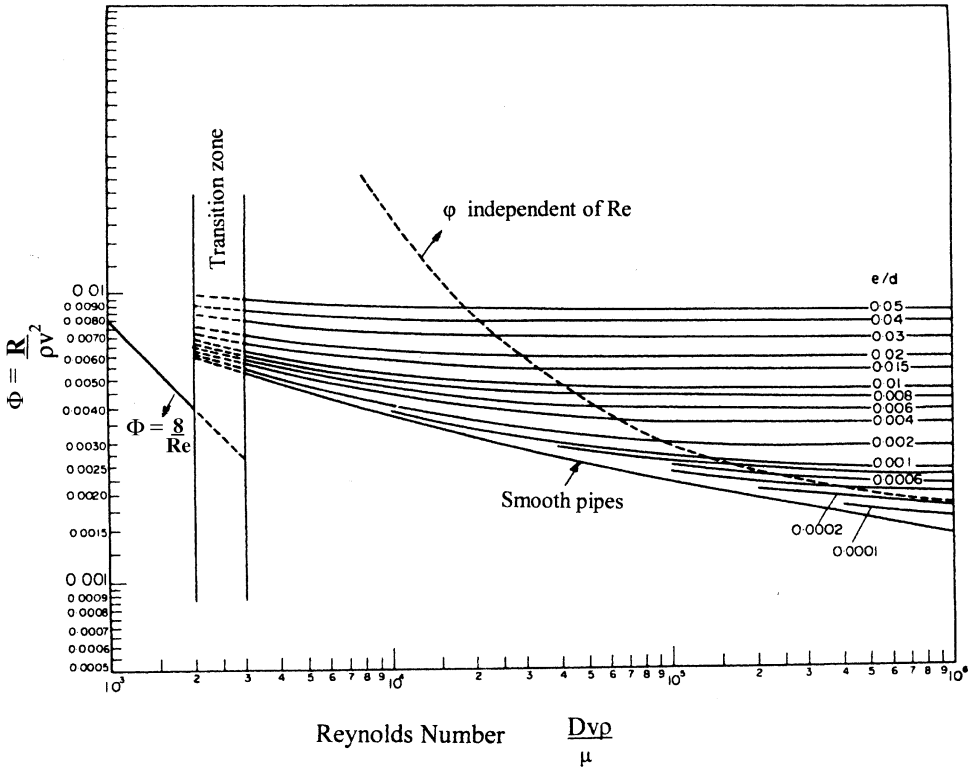


Fig. 8 Friction in pipes related to Reynolds number. (From Ref.[8].)

3. Pressure energy, i.e., the work done to bring the fluid into the system without change in volume.

In the light of the conservation of energy concept, the Bernoulli equation states that the sum of these three energies is constant. Generally, the three terms are referred to as "heads" having the dimensions of length.

$$\frac{PV}{g} + \frac{v^2}{2g} + Z = \text{a constant} \tag{9}$$

where P is the pressure, V is the volume of unit mass, v is the velocity, and Z is the height above the datum.

Under ideal conditions, as illustrated in Fig. 9, the Bernoulli equation written between points 1 and 2 yields

$$\frac{P_1V}{g} + \frac{v_1^2}{2g} + Z_1 = \frac{P_2V}{g} + \frac{v_2^2}{2g} + Z_2 \tag{10}$$

Under real conditions, however, rather than the idealized conditions of the above Eq. (10), energy is lost as a result of friction and energy could be added, say, by a pump. The energy balance then becomes:

$$\frac{P_1V}{g} + \frac{v_1^2}{2g} + Z_1 + W = \frac{P_2V}{g} + \frac{v_2^2}{2g} + Z_2 + F \tag{11}$$

where W represents the work done on the system, say, by the pump and F represents the frictional losses.

From this equation, knowing the frictional losses by calculation using Eq. (8) and the other terms, it is possible to estimate the energy required to transport the fluid from points 1 and 2 on a continuous basis. It should be noted that the term W has the units of length. To calculate the power required, W has to be multiplied by the flow of fluid and the acceleration due to gravity, g.

Pressure Loss Through Fittings

It is seldom that a fluid flows only along a straight length of pipe. The pipe run may involve many changes in direction and velocity resulting from changes in flow area, due to the effects of the presence of fittings that could include such items as bends, valves, tee junctions, expansion and contraction joints. As a result, a higher pressure loss than that associated with the straight length of pipe is encountered. In general, for turbulent flow, the effect of fittings is taken into account in terms of equivalent lengths of straight pipe that would give the same pressure loss as the fitting, under the same conditions. It is possible then in a practical example to add together all the equivalent lengths of pipe for the fittings and the straight runs of pipe, so that an estimate of the total pressure drop along the entire length of pipe between two chosen points may be made. Because the effect is related to pipe diameter, it is usual to quote the equivalent length in terms of the number of pipe diameters. To provide easy access to the Bernoulli equation it is possible as an alternative to quote the energy loss in terms of velocity heads. Table 2 is an adapted version of a table provided by Coulson and Richardson.[3]

Measurement of Flow Through Pipes

The loss of pressure through constrictions (where in addition to the loss of energy due to friction, there is conversion of pressure energy into kinetic energy) provides a method of measuring flow rate. Special devices have been used to measure flow. These include venturi meter, orifice plate, and nozzle meter. The pressure loss through the device may be measured by

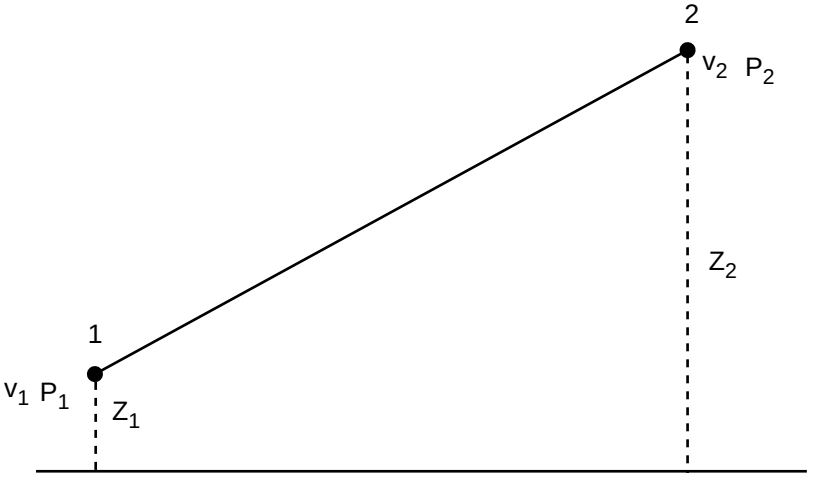


Fig. 9 Energy balance in a fluid flowing between two points.

Table 2 Friction losses in pipe fittings

Fitting	Equivalent length (number of pipe diameters)	Number of velocity heads
45° elbow	15	0.3
90° bend (standard radius)	30–40	0.6–0.8
90° bend (square)	60	1.2
Entry from leg of tee	60	1.2
Entry into leg of tee	90	1.8
Unions and pipe couplings	Negligible	Negligible
Fully open globe valve	60–300	1.2–6.0
Fully open gate valve	7	0.15

(From Ref.[3].)

a simple manometer, which is common in laboratories, or by a more sophisticated electronic technique used on process plant. Other devices are available and the choice will depend very much on the application and the cost.

Noncircular Pipes

In some process plants, fluid flows through noncircular pipes or ducts. For example air in air-conditioning systems is often made to flow through ducting of square cross section. A simple heat exchanger may comprise a double-walled pipe, with one fluid flowing through the annulus. In calculations involving these cross sections, the hydraulic mean pipe diameter may be used instead of the pipe diameter, in formulae normally associated with pipe flow. For instance, the Reynolds number would be $D_m v\rho/\mu$, where the hydraulic mean diameter D_m is defined as:

$$D_m = \frac{4 \times \text{cross-sectional flow area}}{\text{wetted perimeter}} \tag{12}$$

For a duct of rectangular cross section $l_1 \times l_2$

$$D_m = \frac{4l_1l_2}{2(l_1 + l_2)} = \frac{2l_1l_2}{(l_1 + l_2)} \tag{13}$$

For an annular section where D_o is the outer diameter of the inner pipe and D_i is the inner diameter of the outer pipe

$$D_m = \frac{4\left(\frac{\pi D_i^2}{4}\right) - \left(\frac{\pi D_o^2}{4}\right)}{\pi D_i + \pi D_o} \tag{14}$$

$$= \frac{(D_i^2 - D_o^2)}{(D_i + D_o)} \tag{15}$$

$$= \frac{(D_i - D_o)(D_i + D_o)}{(D_i + D_o)} \tag{16}$$

$$= D_i - D_o \tag{17}$$

Open Channels

Liquid flow in open channels is not common in a process plant, because of health and safety considerations, apart from the potential risk of contamination of the liquid itself. Open channels, however, are sometimes to be found in cooling water systems where large volumes of water are involved. Nevertheless, for the sake of completeness a brief treatment of the subject is included in this entry.

As with noncircular ducts the hydraulic mean diameter is employed in formulae that involve diameter. If a channel has a height of a and a width b, the flow area of the channel is ab. In the calculation of the wetted perimeter the free surface is not included so that the wetted perimeter is $2a + b$, and the hydraulic mean diameter

$$D_m = \frac{4ab}{(2a + b)} \tag{18}$$

In a simple system where the open channel is inclined, the difference in the head between two points will equal the friction losses for a uniform cross section between the two points, which will also dictate the flow rate.

TWO-PHASE FLOW

Two-phase flow concerns the interacting flow of two phases involving mixtures of solid/liquid, solid/gas, gas/liquid. The interface between the phases is affected by the motion of the phases. In general, two-phase flow consisting of liquid/gas (or vapor) can be considered to be the most common in the processing industries,

although the behavior of droplets in a gas stream has similarities to that of solid particles, except that droplets are susceptible to deformation under the influence of the operating conditions. Vapor/liquid two-phase flow is fundamental to the following industrial operations:

Boilers
Nuclear reactors
Flash evaporators
Condensers
Distillation columns
Absorption columns

The patterns, or regimes, that can be identified when gas (or vapor) and liquid flow together in a pipe will depend on not only the ratio of gas to liquid but also whether the flow is horizontal or vertical. It is to be anticipated that under vertical flow conditions, there will be dispersion of the gas within the liquid, the pattern dependent on the ratio of gas/liquid and the flow conditions. It will be appreciated that under horizontal flow conditions, there is a possibility of separation between the phases, with a tendency for the gas to flow above the flowing liquid. Again, the resulting pattern will depend on the ratio of the two phases in the mixture and the flow conditions.

Diagrams of flow patterns have been presented by Hewitt and Hall Taylor[4], and these are reproduced as Figs. 10 and 11.[4] These authors also describe the flow patterns. In consideration of Fig. 10 for vertical flow:

Bubble flow (sometimes referred to as bubbly flow): The gas phase is distributed as discrete bubbles throughout a continuous phase of liquid.

Slug flow (sometimes referred to as plug flow): Much of the gas phase is distributed in large bubbles that virtually have the same cross section as the pipe or channel in which they occur. The liquid between the large bubbles may contain a dispersion of smaller bubbles.

Churn flow: If the velocity of a two-phase mixture in slug flow is increased, the large slugs of gas will tend to become unstable, with the possibility of breakup. The result is the destruction of the slug flow pattern, with an oscillating characteristic being established.

Annular flow: The liquid phase flows along the pipe or channel walls, as a more or less continuous stream, with the gas phase acting as a "core." The gas phase may carry droplets of liquid that may be generated by the breakup of waves on the surface of the liquid film. Some liquid drops may fall back into the liquid phase, so that there may be a continuous liquid interchange between the continuous liquid phase and the gas phase. Furthermore, the liquid may contain entrained gas bubbles. The pattern detail will depend very strongly on the flow conditions in the system. Hewitt and Hall Taylor describe a subpattern of annu-

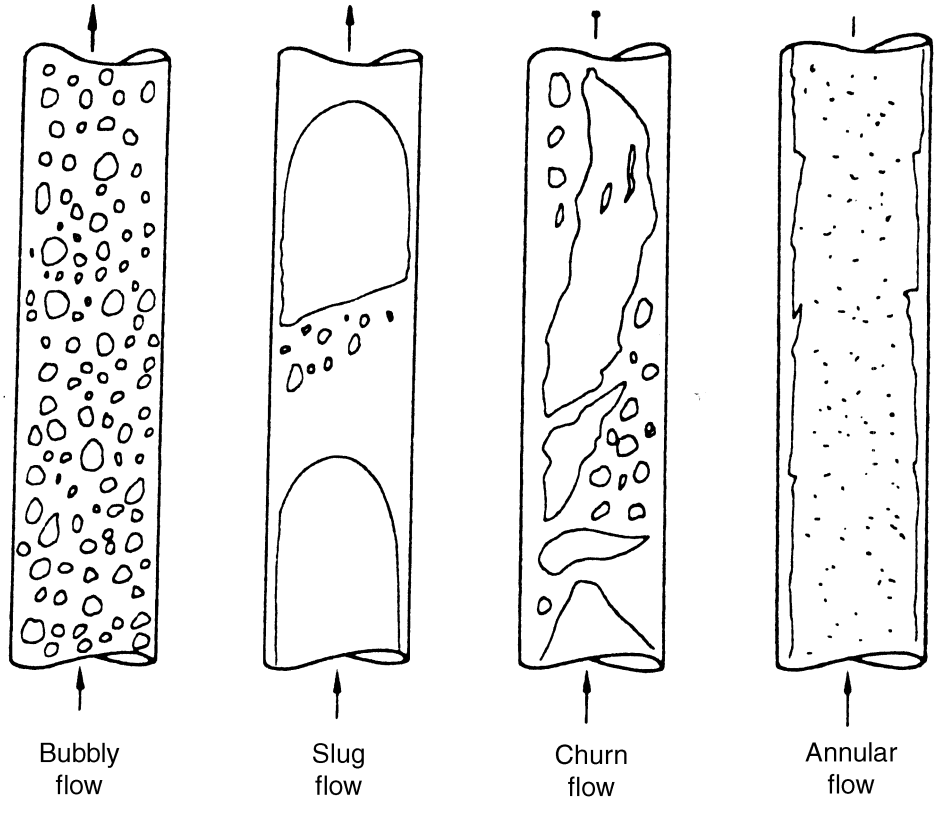


Fig. 10 Flow patterns in vertical two-phase flow. (From Ref.[4].)

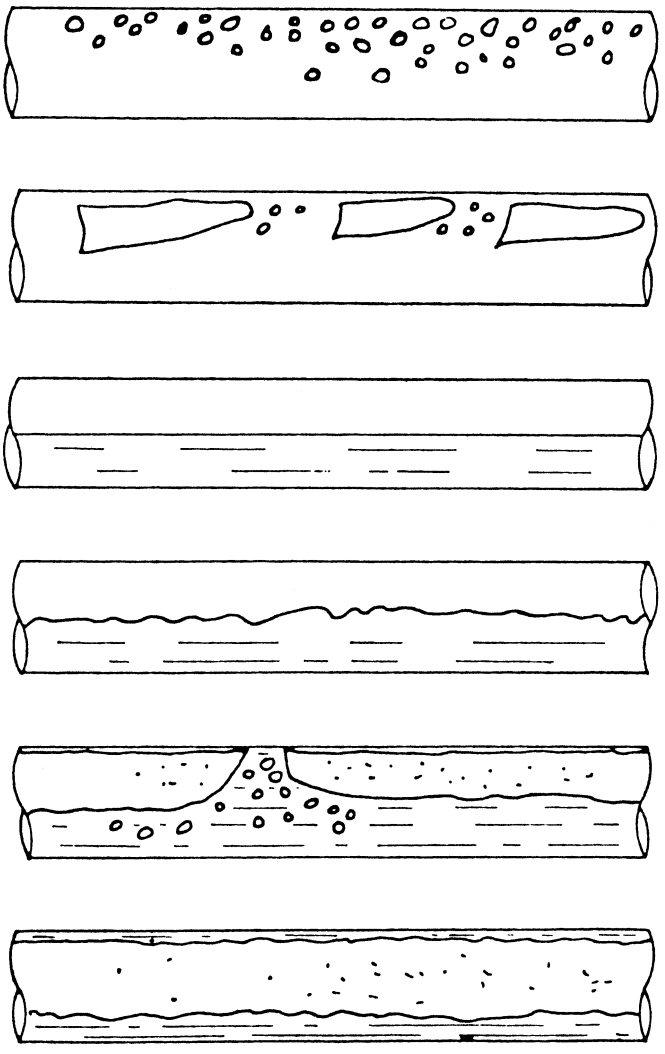


Fig. 11 Flow patterns in horizontal two-phase flow. (From Ref.[4].)

lar flow, which they call "wispy annular" flow.[4] In this pattern the entrained phase flows in large "lumps" or "wisps." The size of these lumps is gradually reduced if the gas velocity is increased.

Fig. 10 illustrates upward vertical flow through channels. It is possible of course that vertical downward flow may be encountered. It would be anticipated that the regimes would be modified under these conditions.

Fig. 11 defines flow patterns for horizontal (and inclined) flow:

Bubbly flow: Similar to that experienced in vertical flow, but with a tendency for the bubbles to accumulate in the upper part of the flow channel. It is possible for some of the bubbles to collide and form larger bubbles.

Plug flow: Again, similar to the pattern for vertical flow, but with the liquid film at the bottom of the channel tending to be thicker than at the top of the channel.

Stratified flow: As the name suggests, there is complete separation between the two phases: the liquid being in the lower part of the channel.

Wavy flow: With increased velocity in stratified flow, waves appear on the surface of the liquid layer. It is possible for droplets to be torn from the tops of the waves to be carried forward in the gas phase. Some droplets may coalesce and fall back into the liquid layer.

Slug flow: At higher velocities, the waves generated are so large that their crests can touch the upper surface of the channel, and it is possible that this surface will become wetted by the liquid under these conditions.

Annular flow: At still higher velocities the liquid is dispersed to the channel walls, where it forms a film that is thicker at the bottom of the channel. The gas flows as a "core."

Flow in inclined channels is not common, but it may be stated that the flow patterns of two-phase flow systems in inclined channels fall between those for horizontal and vertical flow. The detail of the pattern depends on the angle of inclination of the tube.

The discussion so far suggests that the flow pattern in two-phase flow, as well as being related to the physical properties of the phases, will be strongly

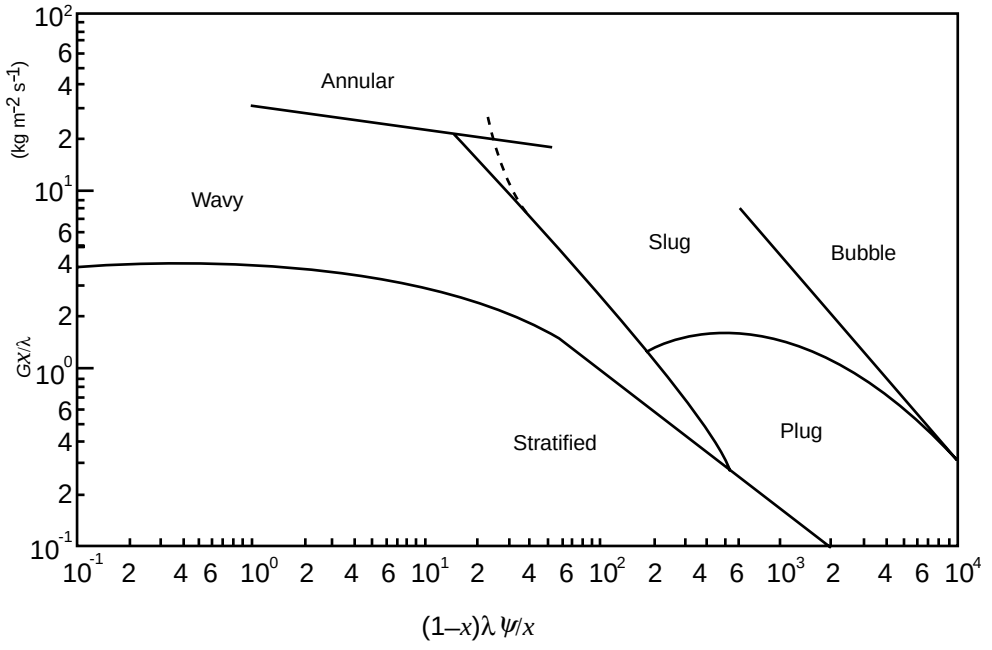


Fig. 12 Flow pattern map for horizontal two-phase flow. (From Ref.[6].)

influenced by the amounts of each of the phases present, and the velocity of the mixture through the channel. This has led to the use of empirical parameters to provide a ''chart'' for the determination of where the different flow patterns occur.

One of the earliest flow diagrams for horizontal flow was devised by Baker.[5] An example is given in Fig. 12 based on the work of Schicht from Butterworth and Hewitt.[6,7] The vertical axis is Gx/λ and the horizontal axis is $(1 - x)\lambda\psi/x$, where G is the total mass flow rate and x is the ''quality,'' i.e., the gas or vapor phase mass flow fraction:

$$\lambda = \left[\left[\frac{\rho_G}{\rho_A}\right]\left[\frac{\rho_L}{\rho_W}\right]\right]^{0.5} \tag{19}$$

$$\psi = \left[\frac{\sigma_W}{\sigma}\right]\left[\left[\frac{\mu_L}{\mu_W}\right]\left[\frac{\rho_W}{\rho_L}\right]^2\right]^{1/3} \tag{20}$$

where ρ is the density, σ is the surface tension, subscripts G and L refer to gas and liquid, respectively, and subscripts A and W refer to the corresponding values for air and water at atmospheric pressure.

A similar chart for vertical flow is given in Fig. 13, although the axes are simplified: the vertical and horizontal axes are $\rho_G v^2$ and $\rho_L v^2$, respectively, and ν_G and v_L are the superficial gas and liquid velocities, respectively.

Two-Phase Flow Pressure Gradients

The method that is usually adopted for the calculation of pressure drop in two-phase flow systems is based on empirical correlations, and involves putting parameters ϕ_G and $\phi_L$against the variable x as defined by the following equations:[4,9]

$$\varphi_{GorL} = \left[\frac{dp_F/dz}{dp_F/dz_{GorL}}\right]^{1/2} \tag{21}$$

and

$$x = \left[\frac{[dp_F/dz]_L}{[dp_F/dz]_G}\right]^{1/2} \tag{22}$$

where $[dp_F/dz]_G$ and $[dp_F/dz]_L$are the pressure gradients for the gas and the liquid phases, respectively, flowing alone in single-phase mode in the channel.

Four sets of data are obtained depending on whether the phases would be flowing under laminar or turbulent conditions.

Space limitations prevent an extensive review and discussion of two-phase flow but the reader is referred to Butterworth and Hewitt and Hewitt and Hall Taylor for further reading.[4,7]

SPECIALIST FLUID FLOW

The following provides a list of the operations that involve flow of fluids and require specialized individual

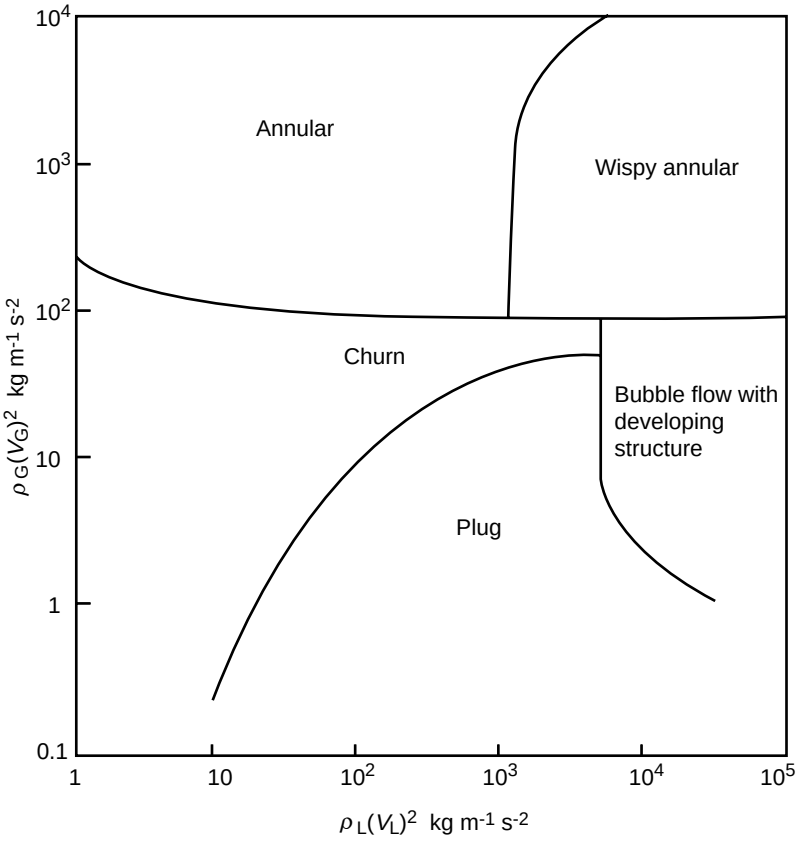


Fig. 13 Flow pattern map for vertical two-phase flow. (From Ref.[8].)

mathematical treatment and that are outside the scope of this encyclopedia.

The reader is referred to specialist publications in respect of these topics:

Flow through
Packed beds
Distillation columns
Fluidized beds
Filters
Membranes
Turbulence promoters
Compact heat exchangers
Cyclones
Chimneys
Ovens
Dryers
Settlers
Pumps and blowers
Mixers
Flow-in
Falling films
Jets, nozzles, and sprays
Plumes

CONCLUSIONS

The flow of fluids is fundamental to many processes, and a thorough knowledge of the behavior of fluids under flowing conditions is vital for the safe and economic operation of a process plant. Furthermore, to make a reliable plant design and an accurate assessment of the energy requirements that will be needed to move fluids around the plant when it is built, an engineer relies on the accurate prediction of fluid behavior in the various items of equipment that constitute the plant. The discussion in this section is aimed at providing an indication of the principles of fluid flow. The limitations of space, however, have of necessity curtailed that discussion to the fundamentals. The reader is directed to the many books that have been published on the topic; some of them are referred to in the text.

REFERENCES

1. Reynolds, O. An experimental investigation of the circumstances which determine whether the motion of water shall be direct or sinuous and the law of

resistance in parallel channels. Pap. Mech. Phys. Subj. **1881–1901**, *2*, 51.
2. Stanton, T.; Pannell, J. Similarity of motion in relation to the surface friction of fluids. Philos. Trans. R. Soc. **1914**, *214*, 199.
3. Coulson, J.M.; Richardson, J.F. *Chemical Engineering*, 3rd Ed.; Pergamon Press: Oxford, U.K., 1977; Vol. 1, 60.
4. Hewitt, G.F.; Hall Taylor, N.S. *Annular Two Phase Flow*; Pergamon Press: Oxford, U.K., 1970; 4–7.
5. Baker, O. Multiphase flow in pipelines. Oil Gas J. Prog. Rep. **1958**, 156–167.
6. Schicht, H.H. Flow patterns for an adiabatic two phase flow of water and air within a horizontal tube. Verfahrens-technik **1969**, *3* (4), 153–161.
7. Hewitt, G.F. Flow patterns. In *Two Phase Flow*; Butterworth, D., Hewitt, G.F., Eds.; Oxford University Press: Oxford, U.K., 1977; 18–39.
8. Hewitt, G.F.; Roberts, D.N. *Studies of Two Phase Flow Patterns by Simultaneous X-ray and Flash Photography*, U.K.A.E.A. Report No. A.E.R.E.-M 2159.
9. Lockhart, R.W.; Martinelli, R.C. Proposed correlation of data for isothermal two-phase, two component flow in pipes. Chem. Eng. Prog. **1949**, *45* (1), 39–48.

Fluid Transport in Porous Media

Michael C. Brooks
U.S. Environmental Protection Agency, Kerr Research Center, Ada, Oklahoma, U.S.A.

INTRODUCTION

Fluid transport through porous media is a relevant topic to many scientific and engineering fields. Soil scientists, civil engineers, hydrologists, and hydrogeologists are concerned with the transport of water, gases, and non-aqueous phase liquid (NAPL) contaminants through porous earth materials; petroleum engineers are concerned with the flow of water, gas, and oil through production reservoirs; chemical engineers use principles of fluid transport through porous media in the design and operation of chemical reactors, and material engineers design and manufacture a variety of porous materials. The topic in general requires an understanding of concepts from a number of disciplines, such as physics, chemistry, engineering, fluid mechanics, microbiology, and geology. Albeit a complex topic, the following four questions can provide a conceptual framework within which to understand the fundamentally important components:

- Which phases are present?
- How many phases are in motion?
- Do chemical species of interest partition between phases?
- Are the species of interest produced or consumed by reactive processes?

The first question is addressed in the second section through a discussion of porous media characteristics and the distribution of fluids in porous media. Fluid motion in porous media is discussed in the third section, and partitioning and reactive processes are discussed in the fourth section.

There is a wealth of literature available on transport in porous media; consequently, the goal herein is to provide the reader with a general overview of important topics relevant to the framework provided by the four general questions above. To limit the scope of this work, the topics herein are limited to those associated with isothermal flow in rigid porous media with unchanging pore structure.

POROUS MEDIA

A porous medium is simply a solid material (i.e., solid phase) through which a significant void volume extends. Typically, the term is used when the void size in the media is large relative to the fluid or gas molecules contained therein, yet small enough such that capillary forces are significant.[1] Examples include ceramic coasters on which drinking glasses are placed, activated carbon for water treatment, and aquifers used to supply drinking water. Fig. 1 lists six fundamental properties used to characterize porous media. Because porous media are often composed of solid particles, the particle-size distribution is one such fundamental characteristic of porous media, and other properties are often correlated to the particle size distribution. Another important characteristic of porous media is its porosity (η), which is defined as:

$$\eta = \frac{V_V}{V_S} \tag{1}$$

where V_V is the volume of voids [L^3] and V_S is the total sample volume [L^3]. For transport to occur, the voids must be interconnected. The term effective porosity (η_e) is used to describe the interconnected pore space, and is defined as the volume of interconnected pore volume divided by the total sample volume. The total sample volume is not arbitrary and must be of sufficient size to meet the requirements of a representative elementary volume (REV). This ensures the sample volume is large enough for a value representative of the bulk porous media properties, yet small enough for mathematical treatment on a continuum basis. Fig. 2 illustrates porosity as a function of sample volume to illustrate the concept. The REV concept is used often in the study of porous media because it is impractical to fully describe the actual pore space over most domain sizes of interest.

The voids within the porous media are occupied by gas and liquid. The terms used to describe the phase distribution in the pore space include saturation and fluid content. The saturation (S_α) is defined as the volume occupied by phase α divided by the pore volume, while the fluid content (θ_α) is defined as the volume of fluid α divided by the sample volume. The sum of all phase saturations is equal to 1, and the sum of all fluid contents is equal to the porosity.

The pore-size distribution of porous media is inherently important, as it is directly related to such characteristics as fluid flow and storage, and multiphase fluid distribution and displacement. The methods used to measure the pore-size distribution include image

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120019072

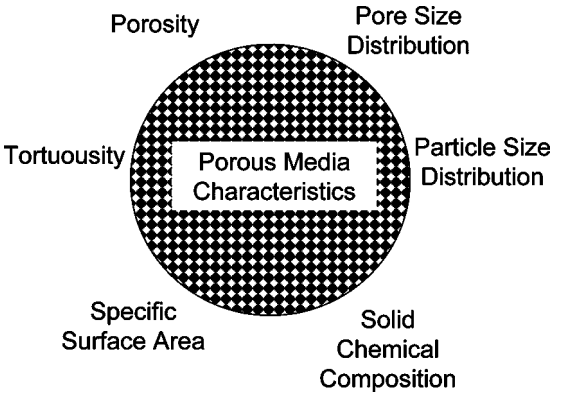


Fig. 1 Characteristic properties of porous media.

analysis techniques, geophysical techniques, and fluid displacement techniques.[2] The pore-size distribution directly affects the capillary pressure–saturation relationship. Capillary pressure–saturation curves can be measured by mercury intrusion or water drainage (or retention) methods. Both methods are based on the measurement of displaced fluid volume as a function of capillary pressure, and differ in that the former measurement is based on the intrusion of a nonwetting fluid (mercury) into the media, and the latter measurement is based on the drainage of a wetting fluid (typically water) from the media.

The terms wetting and nonwetting refer to the relative affinity of a fluid to the porous media, and are formally defined based on the contact angle; wetting fluids have contact angles less than 90° and nonwetting fluids have contact angles greater than 90° (Fig. 3). The capillary pressure is the pressure difference across the nonwetting–wetting fluid interface, and is defined by the Young–Laplace equation of capillarity:

$$P_c = P_{nw} - P_w = \sigma\left(\frac{1}{r_1} + \frac{1}{r_2}\right) \quad (2)$$

where P_c is capillary pressure [$ML^{-1}T^{-2}$], P_{nw} is the nonwetting fluid pressure [$ML^{-1}T^{-2}$], P_w is the wetting fluid pressure [$ML^{-1}T^{-2}$], σ is interfacial tension [MT^{-2}], and r_1[L] and r_2 [L] are the radii of curvature for the wetting–nonwetting interface.[3] Considering a pore space filled with a wetting fluid, it will remain saturated with the wetting fluid until changes in P_{nw} or P_w result in a capillary pressure that exceeds the pore space entry pressure (i.e., the capillary pressure based on interfacial tension and pore space geometry). At that point, the wetting fluid will be displaced by the nonwetting fluid. The inverse relationship between capillary pressure and radii of curvature in Eq. (2) indicates that the nonwetting phase will displace the wetting phase in the larger pores first.

Several analytical equations have been proposed to describe capillary pressure–saturation relationships in porous media, and Fig. 4 presents one of the equations commonly used, the van Genuchten model.[4] The water release curve shown is based on the data presented by Schroth et al. for 20/30 and 40/50 silica sand.[5] The saturation–capillary pressure curve can be viewed as a cumulative distribution of pore sizes,

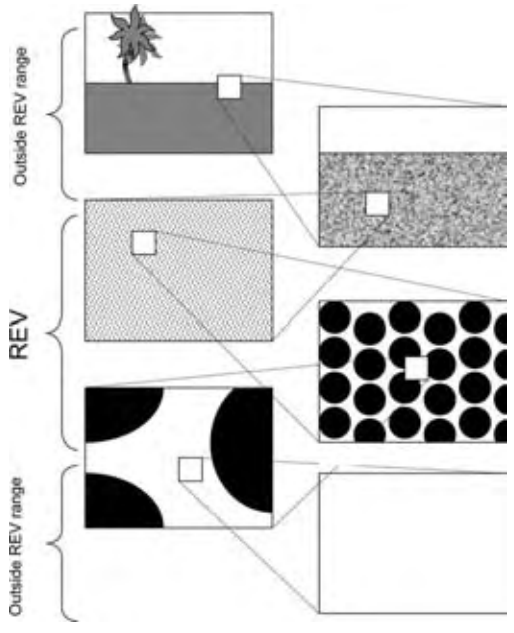


Fig. 2 Porosity as a function of sample volume; an illustration of the REV concept. If the sample volume (represented by the squares) becomes too large or too small, the porosity is no longer representative of the bulk porous media.

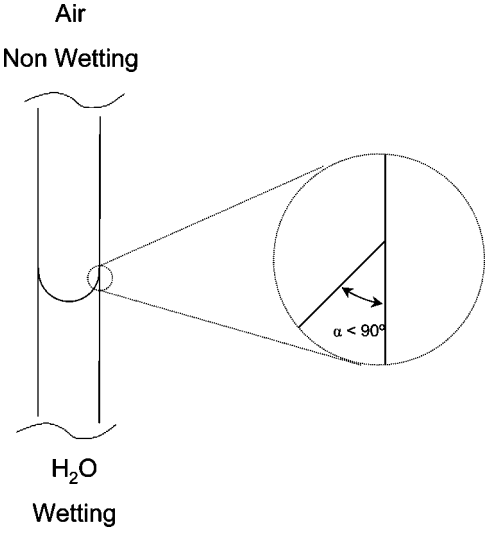


Fig. 3 Wetting vs. nonwetting fluids in porous media, illustrated by analogy to fluids in a capillary tube. Wetting fluids have contact angles less than 90°, and nonwetting fluids have contact angles greater than 90°.

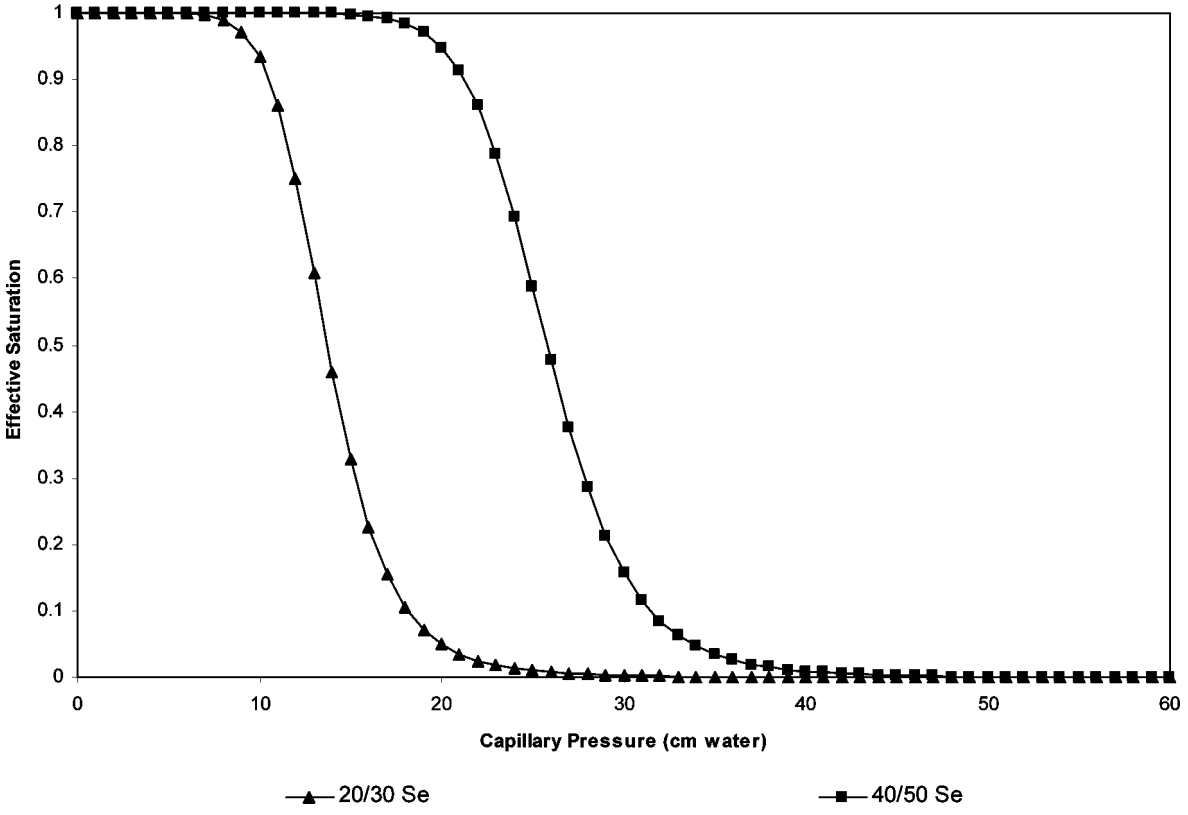


Fig. 4 Effective saturation (a normalized saturation value) as a function of capillary pressure, as described by the van Genuchten equation. The curves shown are based on data reported by Schroth et al. (From Refs.[4,5].)

and its derivative then represents the pore-size distribution.[6]

Other important characteristics of porous media include the specific surface area and the tortuosity factor. The specific surface area is defined as the surface area of the solid phase divided by the sample volume (and hence has units of L^{-1}), and is important to solid–fluid interactions (for example, this is an important parameter for activated carbon characterization). The tortuosity factor τ in porous media is defined as:

$$\tau = \left(\frac{l^*}{l}\right)^2 \tag{3}$$

where l^* [L] is the tortuous path length between two points traveled by a fluid particle, and l [L] is the straight line distance between the points.[2] Because of the tortuous path through porous media, τ is greater than unity. Finally, the chemical composition of the solid phase may also be important in fluid transport through porous media, and may affect such factors as the fluid's wetting nature (i.e., whether it acts as a wetting or nonwetting fluid), sorption of chemical species in the fluids, and reactions with chemical species in the fluids.

FLUID MOTION IN POROUS MEDIA

There are two transport processes associated with fluid motion through porous media: advection and dispersion. The former can be attributed to the bulk motion of the fluid, and the latter can be attributed, in general, to variations in concentration and velocity.

Advection

One of the most common equations for fluid flow through porous media is Darcy's law, which states that fluid flux in porous media is linearly proportional to the hydraulic gradient:

$$\mathbf{q} = \frac{Q}{A} = -K\frac{\mathrm{d}h}{\mathrm{d}x} \tag{4}$$

where $\mathbf{q}$ is the Darcy flux (the bold format indicates a vector) [LT^{-1}], Q is the volumetric flow rate [L^3T^{-1}], K is the hydraulic conductivity [LT^{-1}], A is the cross-sectional area [L^2], h is the hydraulic head [L], and x is the linear flow distance [L]. Darcy's law is generally valid when the Reynolds number (defined as $qd\rho/\mu$ where d is a characteristic length scale of the porous media [L], ρ is the fluid density [ML^{-3}], and μ is the

viscosity [$ML^{-1}T^{-1}$]) is less than 10.[7] The hydraulic head h is the sum of gravitational and pressure potentials, as velocity potentials are negligible under conditions applicable for Darcy's law:

$$h = z + \frac{P}{g\rho} \tag{5}$$

where z is height above a datum [L], P is pressure [$ML^{-1}T^{-2}$], ρ is density [ML^{-3}], and g is the gravitation constant [LT^{-2}]. The hydraulic conductivity K is a function of both the porous media and the fluid moving through it. The term hydraulic conductivity is used when the fluid is water. By convention, positive flow is in the direction of the negative hydraulic gradient, and hence a negative is included on the right-hand side of the equation.[4] Because a portion of the cross-sectional area A is occupied by solid material (i.e., $\{1 - \eta\}^* A$ is occupied by solid material), the fluid velocity through the pores is higher than the Darcy flux $\mathbf{q}$. This velocity is called the seepage velocity or pore water velocity, v [LT^{-1}], and is defined as $\mathbf{q}/\eta$. Although it was originally derived from an empirical basis, it has been shown that Darcy's law can be derived from momentum conservation principles, and is a special case of the Stokes equation when inertial effects can be neglected, e.g., Refs.[8,9]. Under these constraints, it is applicable to homogenous laminar liquid flow; within a rigid, immobile, and homogenous solid matrix, averaged over a sample volume of sufficient size such that viscous forces can be treated as a body force. A more general form of Darcy's law is as follows:

$$\mathbf{q} = -\frac{k k_r \rho g}{\mu} \frac{\partial \phi}{\partial x} \tag{6}$$

where k is the intrinsic permeability [L^2], k_r is the relative permeability [dimensionless], and ϕ is the potential.[10] In cases where density is pressure dependent, as possibly in gas flow, the hydraulic head is replaced by the more general potential term ϕ, given by:

$$\phi = z + \frac{1}{g} \int_{P_0}^{P} \frac{dp}{\rho(P)} \tag{7}$$

where P is pressure [$ML^{-1}T^{-2}$], P_0 is a reference pressure [$ML^{-1}T^{-2}$], and $\rho(p)$ is pressure-dependent density [ML^{-3}]. The conductivity K in Eq. (4) was replaced in Eq. (6) by separate terms for conductance as a function of the porous media (i.e., intrinsic permeability k) and fluid properties (i.e., ρ/μ). The relative permeability k_r accounts for variations in permeability when multiple fluids are present in the porous media and is a function of saturation. It is assumed in multiphase fluid applications of Darcy's law that the pressure distribution in one fluid does not impact the pressure distribution in the other phases.

It is worth mentioning that while Darcy's law is often used to describe gas and liquid flow under heterogenous conditions, its validity under extended conditions is not accepted without debate (e.g., Refs.[11,12]). Consideration should therefore be given to the particular conditions under investigation to ensure the applicability of Darcy's law to describe fluid flux.

Flow Equations

Darcy's law describes fluid flux in porous media, and must be combined with the continuity equation to develop flow equations. From the flow equations, the spatial and temporal pressure and velocity distributions can be estimated that are needed for the transport equations. The derivation of flow equations starts with the continuity equation, which states that the change in mass or volume within a control volume equals the net flux across the control volume boundary, plus sources and sinks within the control volume. For water within porous media, the continuity equation on a mass basis is:

$$\frac{\partial(S_W \rho_W \eta)}{\partial t} = -\nabla \cdot (\rho_W \mathbf{q}) \pm \phi \tag{8}$$

where S_W is the water saturation [dimensionless], ρ_W is the density of water [ML^{-3}], t is time [T], and ϕ is mass sources or sinks [MT^{-1}]. To solve Eq. (8), both sides must be expressed as a function of one variable, which is chosen to be pressure in most cases. To express the left-hand side of Eq. (8) in terms of pressure, the chain rule is used:

$$\frac{\partial(S_W \rho_W \eta)}{\partial t} = \left(\rho_W \eta \frac{\partial S_W}{\partial P} + S_W \eta \frac{\partial \rho_W}{\partial P} + S_W \rho_W \frac{\partial \eta}{\partial P}\right) \frac{\partial P}{\partial t} \tag{9}$$

It is from the three terms inside the parentheses on the right-hand side of Eq. (9) that the various storage terms used in flow equations are derived. Simplification of Eq. (9) depends on the nature of the particular flow problem (or an a priori assumption of flow behavior). For example, within confined aquifers (i.e., fully saturated), $\partial S_W/\partial P = 0$, and $S_W = 1$, and the terms within the parentheses on the right-hand side of Eq. (9) represent the specific storage of the confined aquifer.[7] Or, in partially saturated conditions, changes in water density and porosity are negligible compared to changes in saturation as a function of pressure,

and the latter represents the storage capacity for partially saturated systems. Darcy's law is used to express the right-hand side of Eq. (8) as a function of hydraulic head (and hence, pressure) to complete the flow equation.

As a final note to the flow equation, it should be evident that porous media properties can vary with location. In this case, the medium is considered heterogenous with respect to the property of interest (as opposed to homogenous media where the porous media property of interest does not vary with location). Likewise, the medium is characterized as either anisotropic or isotropic depending on whether the property is a function of direction or independent of direction, respectively. For example, if the vertical hydraulic conductivity is less than the horizontal conductivity (as is typically the case in aquifers because of sedimentation processes), the medium is considered to be anisotropic. Simplification of the flow equations will depend on the heterogeneity and the anisotropic nature of the porous media.

Multiphase Flow

Multiphase flow occurs when there are two immiscible fluids moving in the porous media, or when there are two miscible phases provided the viscosities are sufficiently different. As presented above, Darcy's law can be extended to describe multiphase flow. However, a wide range of flow patterns are possible because of the instabilities that may result as one fluid displaces another. For example, Tung and Dhir report observations of bubbly, slug, and annular two-phase flow patterns in two-phase flow experiments using air, water, and porous media composed of particles with diameters greater than or equal to 6 mm, while Lenormand, Touboul, and Zarcone reported viscous fingering (i.e., instabilities at the interface between the fluids) two-phase flow patterns using oil and air in micromodels with rectangular pores, 1 mm in depth by 0.1–0.6 mm in width.[13,14]

The major forces controlling multiphase flow patterns are capillary forces, viscous forces, and buoyancy forces.[15] These forces are often compared using three key dimensionless numbers: the ratio of the displaced to the invading fluid viscosities,

$$M = \frac{\mu_{\text{displaced}}}{\mu_{\text{invading}}} \tag{10}$$

the capillary number, which is the ratio of the viscous to the capillary forces,

$$Ca = \frac{(\nu\mu)_{\text{displaced}}}{\sigma} \tag{11}$$

and the Bond number, which is the ratio of buoyancy to capillary forces.

$$Bo = \frac{g\Delta\rho d^2}{\sigma} \tag{12}$$

For example, an unstable interface is observed when a more viscous fluid is displaced by a less viscous fluid in porous media (i.e., as M increases). However, capillary forces can offset the effects of viscosity differences, but as the speed of displacement increases (i.e., as Ca increases), instabilities are more likely. Likewise, as Bo increases, instabilities are more likely because of fluid density differences.

Dispersion

Advective transport alone does not account for all observed transport behavior. Transport observations in porous media exhibit characteristics indicative of a phenomenon beyond that described by advection only, such as breakthrough prior to and tailing after the advective front. This additional transport phenomenon is attributed to hydrodynamic dispersion, which is the sum of diffusive and mechanical dispersive processes. The former is attributed to concentration gradients, and the latter is attributed to variations in velocity at both micro and macro spatial scales.

Diffusive transport results from random molecular motion, which produces net molecular motion from areas of high concentration to low concentration. Fick's law states that diffusive flux is proportional to the concentration gradient, and for porous media applications, is expressed as:

$$\mathbf{J} = -\theta\tau D_{\text{mo}} \frac{\partial C}{\partial x} \tag{13}$$

where $\mathbf{J}$ is flux [$\text{ML}^{-2}\text{T}^{-1}$], D_{mo} is the molecular diffusion coefficient [L^2T^{-1}], C is concentration [ML^{-3}], and x is distance [L]. Under most conditions, the diffusion coefficient can be assumed to be a constant independent of concentration; however, there are conditions in which the diffusion coefficient is a function of concentration, as, for example, in the case of transport of high ionic strength solutions.[16] The diffusion coefficient for a chemical species in porous media is less than the diffusion coefficient for that species in an open fluid because of porous media tortuosity and partial saturation.[17] The diffusion coefficient in porous media is referred to as the effective diffusion coefficient (D_{eff}), and is taken as the product τD_{mo}.

In contrast to diffusion, mechanical dispersion is attributed to variations in advective velocities over a wide range of spatial scales. On the microscale, velocity

variations within pores result in a transport-mixing phenomenon; at an intermediate scale, velocity variations between pores result in a transport-mixing phenomenon; and on the macroscale, variations in porous media characteristics (such as permeability) result in a transport-mixing phenomenon. The magnitude of mechanical dispersion is consequently dependent on the scale of the system under consideration, and can vary by orders of magnitude between laboratory- and field-scale systems.

Mechanical dispersion is assumed to mathematically follow a Fickian diffusion formulation, i.e., the mechanical dispersive flux is assumed to be linearly proportional to the concentration gradient. As such the hydrodynamic dispersion is the sum of diffusive and mechanical dispersive terms, and the total dispersive flux is written as:

$$\mathbf{J}^{\mathrm{D}} = -\theta D_{\mathrm{H}} \frac{\partial C}{\partial x} \tag{14}$$

where D_{H} [$\mathrm{L}^2\mathrm{T}^{-1}$] is the sum of molecular diffusion and mechanical dispersion. A more detailed discussion of hydrodynamic dispersion can be found in Bear.[7]

Dispersive transport behavior in porous media is not always adequately described using the Fickian dispersive model as given by Eq. (14). Nonideal (i.e., non-Fickian) behavior can result from a number of factors, such as extensive heterogeneity in porous media properties and nonequilibrium partitioning processes. Other models have been proposed to relate the dispersive flux to the concentration field (e.g., Ref.[18]). Schumer et al. proposed an alternative formulation in which the dispersive flux is proportional to a fractional derivative of concentration:[19]

$$\mathbf{J}^{\mathrm{D}} = \eta_{\mathrm{e}} D_{\mathrm{H}} \Delta^n(C) \tag{15}$$

where $\Delta^n(C)$ is the nth ($0 < n \leq 1$) derivative of concentration, with n closer to 0 for a very heterogenous medium and equal to 1 for a homogenous medium. The resulting transport equation is reportedly better able to match transport behavior with large tailing phenomena (i.e., prolonged observations of concentration after the advective front).

Advective-Diffusion Equation

A complete derivation of the mass transport equation is presented in detail elsewhere, and an abbreviated derivation is outlined here for one-dimensional transport of a chemical species in porous media saturated with water (e.g., Ref. [9,20]). As with the flow equation, the transport equation begins with the mass conservation principle: the change in mass within a control volume is equal to the net mass flux across the boundary of the control volume, plus sources or sinks within the control volume. The total mass flux, $\mathbf{J}^{\mathrm{T}}$ [$\mathrm{ML}^{-2}\mathrm{T}^{-1}$], is equal to the sum of the advective and dispersive fluxes:

$$\mathbf{J}^{\mathrm{T}} = \mathbf{J}^{\mathrm{A}} + \mathbf{J}^{\mathrm{D}} = \mathbf{q}C - \theta D_{\mathrm{H}} \frac{\partial C}{\partial x} \tag{16}$$

where the negative sign preceding the dispersive flux term indicates positive flux in the direction of negative concentration gradient. In the limit as the finite control volume approaches zero:

$$\frac{\partial m}{\partial t} = \nabla \cdot \mathbf{J}^{\mathrm{T}} \pm \Phi \tag{17}$$

where m is the total mass per sample volume [ML^{-3}] of a species of interest, $\mathbf{J}^{\mathrm{T}}$ is the total mass flux [$\mathrm{ML}^{-2}\mathrm{T}^{-1}$], and Φ refers to mass sources or sinks [MT^{-1}]. The derivation can likewise be extended to three dimensions. The mass term m represents the total mass in the control volume, and accounts for the mass in all phases present. For the simple system under consideration, mass of the species may exist on the solid and in the water phase:

$$m = \rho_{\mathrm{b}} C_{\mathrm{S}} + \theta_{\mathrm{w}} C_{\mathrm{w}} \tag{18}$$

where ρ_{b} is the bulk soil density [ML^{-3}], C_{s} is the solid concentration [MM^{-1}], θ_{w} is the water content [$\mathrm{L}^3\mathrm{L}^{-3}$], and C_{w} [ML^{-3}] is the water concentration. To complete the transport equation, a relationship describing mass partitioning between the water and solid phase is needed. Assuming linear, reversible, and equilibrium partitioning (i.e., $C_{\mathrm{s}} = K_{\mathrm{d}} C_{\mathrm{w}}$), Eq. (18) becomes:

$$m = \left(1 + \frac{K_{\mathrm{d}} \rho_{\mathrm{b}}}{\theta_{\mathrm{w}}}\right) C_{\mathrm{w}} \tag{19}$$

and the transport equation can be written as follows:

$$RC = D \frac{\partial^2 C}{\partial x^2} - v \frac{\partial C}{\partial x} \tag{20}$$

where R is the retardation factor (i.e., $1 + K_{\mathrm{d}} \rho_{\mathrm{b}} / \theta_{\mathrm{w}}$).

Under complex conditions, numerical approximation techniques are required to solve the transport equation, and the solution to the advective-diffusion equation may have to be coupled to numerical solutions to the flow equations as well. However, analytical solutions can be obtained in simpler systems. The solution to the one-dimensional advective–dispersive transport equation with constant pore velocity, subject to the initial condition of $c(x, 0) = 0$ for $x \geq 0$, and the boundary conditions of $c(0, t) = c_0$ for $t \geq 0$,

and $c(\infty, t) = 0$ for $t \geq 0$, in nondimensional form is

$$c = \left(\frac{1}{2}\right)\left\{\text{erfc}\left[\sqrt{\frac{Pe}{4R\tau_{pv}}}(R - \tau_{pv})\right] + \exp(Pe)\text{erfc}\left[\sqrt{\frac{Pe}{4R\tau_{pv}}}(R + \tau_{pv})\right]\right\} \quad (21)$$

where c is the dimensionless concentration (C/C_0), τ_{pv} is the dimensionless pore volume ($\tau = vt/L$, t = time [T], and L = linear extent of the flow domain [L]), and Pe is the Peclet number [dimensionless].[21] Eq. (21) represents an analytical solution based on the step change in injected concentration (i.e., $c = 0$ for $t \leq t^*$, $c = c_0$ for $t > t^*$), and can be used to generate a solution based on a finite pulse, i.e., $c(0, 0 \leq t \leq t_p) = c_0$, $c(0, t > t_p) = 0$, where t_p is the pulse duration [T], by superposition.

The Peclet number (Pe) is defined as

$$Pe = \frac{vl}{D_H} \quad (22)$$

where v is velocity [L^2T^{-1}], l is a characteristic length of the system [L], and D is dispersion coefficient [LT^{-2}]; and is a nondimensional measure of the ratio of advective to dispersive transport processes. When velocity is sufficiently low, diffusion is the dominant process in hydrodynamic dispersion, but as velocity increases, mechanical dispersion becomes the dominant process. It is instructive to note the limiting cases of $Pe = 0$ and $Pe = \infty$. In the former case, diffusive processes dominate transport and the solution resembles a continuously stirred reactor tank. In the latter case, advective transport dominates, and the solution resembles a plug flow solution.

PHASE PARTITIONING AND REACTIVE PROCESSES

Phase partitioning is of course a very important transport process in porous media, and phase partitioning processes in porous media are summarized in Fig. 5. Numerous treatment processes are based on phase partitioning, such as liquid and gas treatment by activated carbon (sorption), or packed tower air stripping (volatilization). Conversely, partitioning plays a direct role in water resource contamination, as in the case of groundwater pollution resulting from leaky gasoline storage tanks (dissolution).

In the analysis presented in the previous section, the porous media were saturated with water, and partitioning was described using a linear equilibrium reversible relationship. Considering a more complex system, if the pore space is filled by three phases (water, gas, and an NAPL), then Eq. (18) becomes:

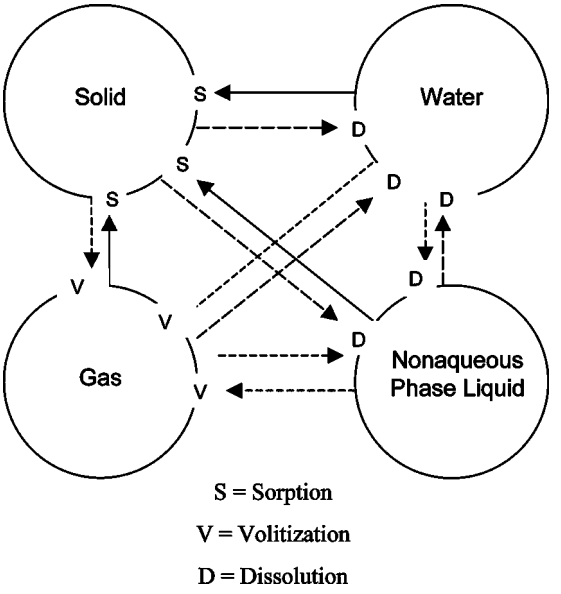


Fig. 5 Conceptual figure of partitioning in porous media.

$$m = C_w\theta_w + \rho_b C_s + \theta_g C_g + \theta_N C_N + a_{g/N}C_{g/N} + a_{N/w}C_{N/w} + a_{g/w}C_{g/w} \quad (23)$$

where the subscripts g, N, g/N, N/w, and g/w refer to gas, NAPL, gas/NAPL interface, NAPL/water interface, and gas/water interface, respectively. The partitioning represented by the last three terms accounts for chemical species that partition at fluid interfaces, such as surfactants for water–air or water–NAPL systems, or long-chain alcohols for water–air systems (e.g., Refs.[22–24]). To complete the equation, relationships describing mass partitioning must be used; assuming linear, reversible, and equilibrium partitioning, Eq. (23) becomes:

$$m = \left(1 + \frac{K_d\rho_b}{\theta_w} + \frac{H\theta_g}{\theta_w} + \frac{K_N\theta_N}{\theta_w} + \frac{\Gamma_1 a_{g/N}}{\theta_w} + \frac{\Gamma_2 a_{N/w}}{\theta_w} + \frac{\Gamma_3 a_{g/w}}{\theta_w}\right)C_w \quad (24)$$

where K_N is the NAPL-water partitioning coefficient, and Γ_1, Γ_2, and Γ_3 represent interfacial partitioning coefficients. Of course, other partitioning behavior is possible besides linear-reversible-equilibrium partitioning, and the following sections discuss phase partitioning in more detail.

Sorption

Chemical species may sorb onto the porous media from the liquid or gas phases, owing to one or more

chemical, electrostatic, or physical forces.[25] In general, chemicals that are hydrophobic in nature tend to sorb more strongly than hydrophilic chemicals. As presented in the section on fluid motion, sorption was assumed to be linear and reversible, and was assumed to be an equilibrium process. However, sorption may be nonlinear and irreversible, and may occur slow enough relative to transport to be a nonequilibrium process. Furthermore, sorption characteristic may vary with location or change with time.

Equilibrium sorption is most often described using a linear, Freundlich or Langmiur relationship. The Freundlich sorption isotherm is:

$$C_s = k_f C_w^{1/n} \tag{25}$$

where k_f is the partitioning coefficient $[(ML^{-3})^{-1/n}]$. When $n = 1$, the equation reduces to the linear partitioning equation:[25]

$$C_s = k_d C_w \tag{26}$$

where k_d is the distribution coefficient $[L^3M^{-1}]$. The Langmuir sorption equation is:

$$C_s = \frac{C_s^{Max} b C_w}{1 + bC_w} \tag{27}$$

where b is the Langmuir sorption constant, and C_s^{Max} is the maximum number of sorption sites.

If transport occurs much faster than sorption, sorption processes may not reach equilibrium conditions. Nonequilibrium sorption may result from physical causes such as intraparticle rate-limited diffusion, chemical causes such as rate-limiting reaction kinetics, or a combination of the two. One approach used to model rate-limited sorption is bi-continuum models consisting of one region where transport is described by the advection–dispersion equation with equilibrium sorption, and another region where transport is diffusion limited with equilibrium sorption, or another region where sorption is chemically rate limited.

Dissolution

Dissolution refers to the partitioning of chemical species between gas–liquid and liquid–liquid phases. Assuming local equilibrium partitioning, the maximum amount of a species that will dissolve into a liquid at a given temperature is given by its solubility. When local equilibrium is not achieved, a rate-limited mass transfer expression is used. For example, Miller, Poirzer-McNeill, and Mayer present a common rate-limited expression for dissolution:[26]

$$J = \kappa\left(C_{aq}^* - C_{aq}\right) \tag{28}$$

where J = mass flux across the interface $[MT^{-1}L^{-2}]$, κ = dissolution mass transfer coefficient $[LT^{-1}]$, and C_{aq}^* = equilibrium aqueous phase concentration $[ML^{-3}]$. The product of the mass transfer coefficient and the specific interfacial area is the mass transfer rate coefficient (i.e., $K_a = \kappa a$). Extensive research on the mass transfer coefficient and the mass transfer rate coefficient has been conducted in several fields of study, and results are often expressed by empirical relationships in the form of $Sh = f(\eta, S, Re, Sc, Pe)$, where Sh is the Sherwood number:

$$Sh = \frac{\kappa l}{D} \tag{29}$$

which represents the ratio of mass transfer rate to diffusion rate, and Sc is the Schmidt number:

$$Sc = \frac{\mu}{\rho D} \tag{30}$$

which represents the ratio of momentum to mass diffusion. Ramachandran and Chaudhari present summaries of empirical relations developed by researchers investigating κ, a, and K_a.[27] For the dissolution of NAPLs into groundwater, Miller, Poirzer-McNeill, and Mayer, Mercer and Cohen, and Powers et al. review factors affecting dissolution and summarize mass transfer coefficient correlations reported in the literature.[26,28,29]

Volatilization

If a gas phase is present, chemical species may volatilize from the liquid or solid phase, which is an important partitioning process in a variety of circumstances (e.g., transport in the unsaturated zone, or for treatment processes). The equilibrium vapor pressure can be used with the ideal gas law to estimate the mass in a given volume and temperature under equilibrium conditions. For solutions with more than one component, Raoult's law can be used to quantify the vapor pressure of each component. For dilute aqueous solutions, Henry's law describes the equilibrium relationship between dissolved chemicals and their vapor pressure:

$$P_i = H^* X_i \tag{31}$$

where P_i = vapor pressure of species $i[ML^{-1}T^{-2}]$, H = Henry's law constant for species $i[ML^{-1}T^{-2}]$, and X_i = mole fraction of the ith species in solution.

For nonequilibrium conditions, a rate-limited mass-transfer expression similar to Eq. (28) is typically assumed:

$$J = k\left(C_g^* - C_g\right) \tag{32}$$

where J is the mass flux across the gas–liquid interface [$MT^{-1}L^{-2}$], k is the mass transfer coefficient [LT^{-1}], C_g^* is the equilibrium gas concentration [ML^{-3}], and C_g is the bulk gas concentration [ML^{-3}].

Reactions

Abiotic and biotic reactions may occur that can produce or consume chemical species in porous media, and transport equations must therefore include descriptions of such processes for prediction accuracy. For example, microorganisms may use ethanol as a carbon source, and consume ethanol as it is transported through the porous system. Examples of other reactive processes include radioactive decay and abiotic degradation. Reactions of chemical species in porous media are often expressed using:

$$\frac{dC}{dt} = -kC^n \tag{33}$$

where k is a reaction constant, and n is the order of the reaction.[30] For a zero-order reaction:

$$C = -kt + C_0 \tag{34}$$

and for a first-order reaction:

$$C = C_0 e^{-kt} \tag{35}$$

For inclusion in the transport equation, the right-hand side of Eq. (33) would be used in place of Φ in Eq. (17).

CONCLUSIONS

A general overview was provided of porous media characteristics, fluid flow in porous media, advection and dispersion in porous media, and phase partitioning and reactive processes in porous media. Four questions were posed in the introduction, and it was suggested that the answers to those questions could be used to highlight the important features of a particular porous media transport problem. By the very nature of porous media, the answer to the first question (i.e., which phases are present?) will at least include a solid phase and one fluid phase, composed of either a liquid or a gas. If multiple phases are present in the void space, then the distribution of the liquid and gas in the pore space will be a function of capillary pressure. The second question (i.e., how many phases are in motion?) indicates the nature of the transport phenomenon. In the simplest case, no phases are in motion, and the only transport mechanism to consider in the static system is diffusion. If one or more phases are in motion, then advection and dispersion must also be considered for each phase in motion. The third question (i.e., do chemical species of interest partition between phases?) depends on the chemical properties of the transporting species and porous media. In the simplest case, the species of interest does not partition between phases. Or when partitioning does occur, the simplest case is partitioning described by reversible, linear equilibrium partitioning. Beyond that, a wide range of partitioning behavior can occur. And finally, the last question (i.e., are the species of interest produced or consumed by reactive processes?) addresses changes that may occur due to radioactive decay, microbiological activity, or abiotic reactions.

REFERENCES

1. Corey, A.T. *Mechanics of Immiscible Fluids in Porous Media*; Water Resources Publication: Littleton, CO, 1990.
2. Dullien, F.A.L. *Porous Media Fluid Transport and Pore Structure*; Academic Press, Inc.: San Diego, 1992.
3. Adamson, A.W.; Gast, A.P. *Physical Chemistry of Surfaces*, 6th Ed.; John Wiley and Sons: New York, 1997.
4. van Genuchten, M. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. **1980**, *44* (5), 892–898.
5. Schroth, M.H.; Ahearn, S.J.; Selker, J.S.; Istok, J.D. Characterization of Miller—similar silica sands for laboratory hydrologic studies. Soil Sci. Soc. Am. J. **1996**, *60* (5), 1331–1339.
6. Ritter, H.L.; Drake, L.C. Macropore-size distributions in some typical porous substances. Ind. Eng. Chem. **1945**, *17* (12), 787–791.
7. Bear, J. *Hydraulics of Groundwater*; McGraw-Hill, Inc.: New York, 1979.
8. Li, J.; Helm, D. Viscous drag, driving forces, and their reduction to Darcy's law. Water Resour. Res. **1998**, *34* (7), 1675–1684.
9. Kolditz, O. *Computational Methods in Environmental Fluid Mechanics*; Springer: Berlin, 2002.
10. Parker, J.C. Multiphase flow and transport in porous media. Rev. Geophys. **1989**, *27* (3), 311–328.
11. Hassanizadeh, S.M.; Gray, W.G. Toward an improved description of the physics of two-phase flow. Adv. Water Resour. **1993**, *16* (1), 53–67.

12. Rose, W. Myths about later-day extensions of Darcy's Law. J. Petrol. Sci. Eng. **2000**, *26* (1–4), 187–198.
13. Tung, V.X.; Dhir, V.K. A hydrodynamic model for two-phase flow through porous media. Int. J. Multiphase Flow **1988**, *14* (1), 47–65.
14. Lenormand, R.; Touboul, E.; Zarcone, C. Numerical models and experiments on immiscible displacements in porous media. J. Fluid Mech. **1988**, *189* (7), 165–187.
15. Conrad, S.H.; Wilson, J.L.; Mason, W.R.; Peplinski, W.J. Visualization of residual organic liquid trapped in aquifers. Water Resour. Res. **1992**, *28* (2), 467–478.
16. Carey, A.; Wheatcraft, S.W.; Glass, R.J.; O'Rourke, J.P. Non-Fickian ionic diffusion across high-concentration gradients. Water Resour. Res. **1995**, *31* (9), 2213–2218.
17. Hu, Q.; Wang, J.S.Y. Aqueous-phase diffusion in unsaturated geologic media: a review. Crit. Rev. Environ. Sci. Technol. **2003**, *33* (3), 275–297.
18. Cushman, J.H.; Ginn, T.R. Fractional advection-dispersion equation: a classical mass balance with convolution-Fickian flux. Water Resour. Res. **2000**, *36* (12), 3763–3766.
19. Schumer, R.; Benson, D.A.; Meerschaert, M.M.; Wheatcraft, S.W. Eulerian derivative of the fractional advection–dispersion equation. J. Contam. Hydrol. **2001**, *48* (1–2), 69–88.
20. Fetter, C.W. *Contaminant Hydrology*; Prentice Hall: Englewood Cliffs, NJ, 1999.
21. Lapidus, L.; Amundson, N.R. Mathematics of adsorption in beads. VI. The effect of longitudinal diffusion in ion exchange and chromatographic columns. J. Phys. Chem. **1952**, *56* (8), 984–988.
22. Hoff, J.T.; Mackay, D.; Gillham, R.; Shiu, W.Y. Partitioning of organic chemicals at the air-water interface in environmental systems. Environ. Sci. Technol. **1993**, *27* (10), 2174–2180.
23. Saripalli, K.P.; Kim, H.; Rao, P.S.C.; Annable, M.D. Measurement of specific fluid-fluid interfacial areas in immiscible fluids in porous media. Environ. Sci. Technol. **1997**, *31* (3), 932–936.
24. Kim, H.; Rao, P.S.C.; Annable, M.D. Gaseous tracer technique for estimating air–water interfacial areas and interface mobility. Soil Sci. Soc. Am. J. **1999**, *63* (6), 1554–1560.
25. Weber, W.J., Jr.; McGinley, P.M.; Katz, L.E. Sorption phenomena in subsurface systems: concepts, models, and effects on contaminant fate and transport. Water Resour. **1991**, *25* (5), 499–528.
26. Miller, C.T.; Poirzer-McNeill, M.M.; Mayer, A.S. Dissolution of trapped nonaqueous phase liquids: mass transfer characteristics. Water Resour. Res. **1990**, *26* (11), 2783–2796.
27. Ramachandran, P.A.; Chaudhari, R.V. *Three-Phase Catalytic Reactors*; Gordon and Breach Science Publishers: New York, 1983.
28. Mercer, J.W.; Cohen, R.M. A review of immiscible fluids in the subsurface: properties, models, characterization and remediation. J. Contam. Hydrol. **1990**, *6* (2), 107–163.
29. Powers, S.E.; Loureiro, C.O.; Abriola, L.M.; Weber, W.J., Jr. Theoretical study of the significance of nonequilibrium dissolution of nonaqueous phase liquids in subsurface systems. Water Resour. Res. **1991**, *27* (4), 463–477.
30. Fogler, H.S. *Elements of Chemical Reaction Engineering*, 3rd Ed.; Prentice Hall: Englewood Cliffs, NJ, 1999.

Fluidization

A.-H. Park
L.-S. Fan
Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio, U.S.A.

INTRODUCTION

Fluidization refers to the state of solid particles in a suspended condition owing to the flow of fluid, gas, and/or liquid. Contact schemes of fluidized bed systems can be classified on the basis of the states of solid motion. For a batch-solids system, the fluid at a low velocity merely percolates through the voids between packed particles, while the particles remain motionless. The solids in this case are in the fixed bed state. With an increase in the fluid velocity, particles move apart and become suspended; the bed then enters the fluidization state. The fluidization characteristics vary, depending on whether gas, liquid, or gas-liquid is the fluidizing medium.

Fluidized beds generally possess the following properties that promote their use in reactor applications:

1. Capability of continuous operation and transport of solids in and out of the bed.
2. High heat transfer rates from bed to surface and from gas to particle leading to temperature uniformity in the bed.
3. High mass transfer rates from gas to particle.
4. Applicability over a wide range of particle properties and high solids mixing rates.
5. Simplicity in geometric configuration and suitability for large-scale operation.

Fluidized beds have been used extensively for physical operations (e.g., adsorption and heat exchanger), chemical synthesis (e.g., acrylonitrile synthesis and maleic anhydride synthesis), metallurgical and mineral processes (e.g., roasting of sulfide ores, resid hydrotreating, and reduction of iron oxide), and other applications, such as coal combustion and microorganism cultivation.

In what follows, the fundamental properties of two-phase fluidization are described. Unless otherwise noted, the properties refer to gas–solid fluidization. The books that can serve as general references for two-phase (gas–solid or liquid–solid) or three-phase (gas–liquid–solid) fluidization subjects are Refs.[1–9].

PARTICLE AND REGIME CLASSIFICATION

Classification of Fluidized Particles

Particles can be classified into four groups (i.e., Groups A, B, C, and D) as shown in Fig. 1, based on their fluidization behavior.[10] Group C comprises small particles ($d_p < 20\,\mu m$) that are cohesive. Group A particles, with a typical size range of 30–100 μm, are readily fluidized. No maximum stable bubble size exists for Group B particles. Group D comprises coarse particles ($d_p > 1\,mm$) that are commonly processed by spouting.

Fluidization Regime

Regime classification for dense- and lean-phase fluidization, in general, can be based on bubble or solid flow behavior. Dense-fluidization regimes include particulate fluidization, bubbling fluidization, and turbulent fluidization. In a broad sense, ''dense-phase fluidization'' also encompasses the slugging, spouting, and channeling conditions of operation. Lean-phase fluidization includes fast fluidization and dilute transport. The fundamental distinction between these regimes or conditions of operation is briefly described next. Fig. 2 illustrates various flow regimes.

Minimum fluidization and particulate fluidization

The state of fluidization begins at the point of minimum or incipient fluidization. At the minimum fluidization point, the pressure drop for a fixed bed and that for a fluidized bed are equivalent. This relationship is used as the basis for the formation of the predictive equation for the minimum fluidization velocity. The pressure drop in the fixed bed can be described by the Ergun equation. Under the minimum fluidization condition, the Ergun equation can be expressed as

$$\frac{\Delta p_b}{H_{mf}} = 150\frac{(1-\alpha_{mf})}{\alpha_{mf}^3}\frac{\mu U_{mf}}{\varphi^2 d_p^2} + 1.75\frac{(1-\alpha_{mf})}{\alpha_{mf}^3}\frac{\rho U_{mf}^2}{\varphi d_p} \quad (1)$$

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120006805

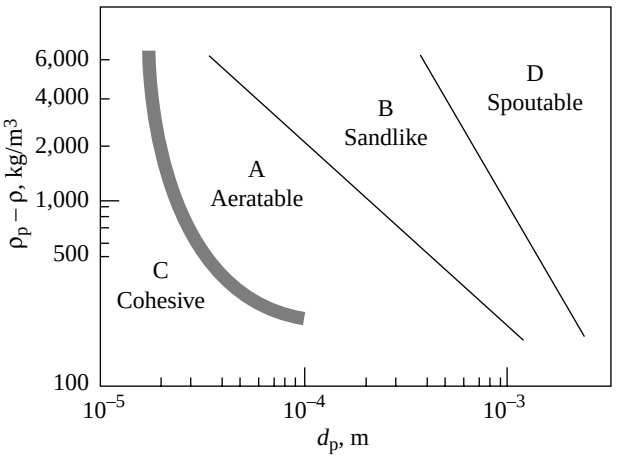


Fig. 1 Geldart's classification of fluidized particles. (From Ref.[10].)

In a fully fluidized bed, the pressure drop (cross-sectionally averaged) counterbalances the weight of the pseudocontinuum of the gas–solid mixture, which yields

$$-\frac{\mathrm{d}p_\mathrm{d}}{\mathrm{d}H} = (\rho_\mathrm{p} - \rho)(1 - \alpha)g \tag{2}$$

Under the minimum fluidization condition, Eq. (2) gives

$$\frac{\Delta p_\mathrm{b}}{H_\mathrm{mf}} = \frac{\Delta p_\mathrm{d}}{H_\mathrm{mf}} = (\rho_\mathrm{p} - \rho)(1 - \alpha_\mathrm{mf})g \tag{3}$$

Equating Eqs. (1) and (3) results in

$$\mathrm{Ar} = 150\frac{(1 - \alpha_\mathrm{mf})}{\alpha_\mathrm{mf}^3\varphi^2}\mathrm{Re}_\mathrm{pmf} + \frac{1.75}{\alpha_\mathrm{mf}^3\varphi}\mathrm{Re}_\mathrm{pmf}^2 \tag{4}$$

and

$$\mathrm{Re}_\mathrm{pmf} = \frac{\rho U_\mathrm{mf} d_\mathrm{p}}{\mu} \tag{5}$$

A semiempirical correlation based on Eq. (4) can be given by[11]

$$\mathrm{Re}_\mathrm{pmf} = \sqrt{(33.7)^2 + 0.0408\mathrm{Ar}} - 33.7 \tag{6}$$

The above analysis for minimum fluidization is applicable to both, the gas–solid and liquid–solid systems.

For a bed with Group A particles, bubbles do not form when the gas velocity reaches U_mf. The bed enters the particulate fluidization regime under this condition. This is also the regime under which liquid fluidization is operated. The operation under the particulate fluidization regime is characterized by a smooth bed expansion with an apparent uniform bed structure without bubbles for $U_\mathrm{mf} < U < U_\mathrm{mb}$, where U_mb is the superficial gas velocity at the minimum bubbling condition. In particulate fluidization, all the gas passes through the interstitial space between the fluidizing particles without forming bubbles. The bed appears grossly homogeneous. This regime exists only in a bed with Group A particles, under a narrow operating range of gas velocities. At high pressures or with gases of high density, the operating range of this regime expands.

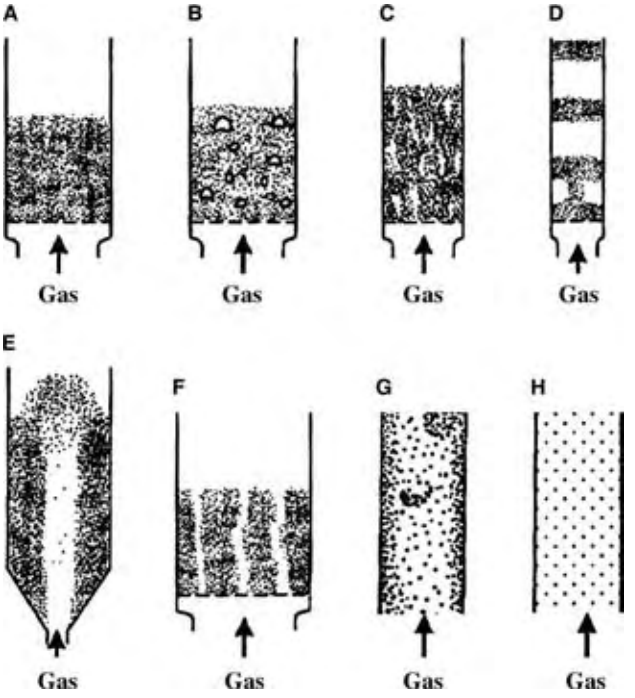


Fig. 2 Various flow regimes or patterns in dense-phase fluidization: (A) particulate fluidization; (B) bubbling fluidization; (C) turbulent fluidization; (D) slugging; (E) spouting; (F) channeling; (G) fast fluidization. (From Ref.[7].)

For liquid fluidization, the Richardson–Zaki equation[12] as given below can be used to describe the bed expansion:

$$\varepsilon = \left(\frac{U_\mathrm{l}}{U_i}\right)^{1/n} \tag{7}$$

where U_i is the extrapolated liquid velocity as the bed voidage approaches 1, and n is the Richardson–Zaki index (see Table 1 for estimation).

Bubbling fluidization

Bubbles are formed as a result of the inherent instability of gas–solid systems. The instability of a gas–solid fluidized bed is characterized by fast growth in local voidage in response to a system perturbation. Because

Table 1 Index for the Richardson–Zaki equation

$n = 4.65 + 20d_p/D_c$	$Re_t < 0.2$
$n = (4.4 + 18d_p/D_c)Re_t^{-0.03}$	$0.2 < Re_t < 1$
$n = (4.4 + 18d_p/D_c)Re_t^{-0.1}$	$1 < Re_t < 200$
$N = 4.4Re_t^{-0.1}$	$200 < Re_t < 500$
$N = 2.4$	$500 < Re_t$

$Re_t = U_t d_p \rho_l / \mu_l$.
(From Ref.[13].)

of the instability in the bed, the local voidage usually grows rapidly into a shape resembling a bubble.[13] Although it is not always true, the initiation of the instability is usually perceived to be the onset of bubbling, which marks the transition from particulate fluidization to bubbling fluidization. The theoretical expansion of the physical origin behind and prediction of the onset of the instability of gas–solid fluidized beds has been attempted (e.g., Refs.[14–16]). The efforts have been focused on the primary forces behind the stability among the interparticle contact forces, particle–fluid interaction forces, and particle–particle interaction via particle velocity fluctuation.

Most bubbles in bubbling fluidized beds are of spherical cap or ellipsoidal cap shape. Configurations of two basic types of bubbles, fast bubble (clouded bubble) and slow bubble (cloudless bubble), are schematically depicted in Fig. 3. The cloud is the region established by the gas, which circulates in a closed loop between the bubble and its surroundings. The cloud phase can be visualized with the aid of a color tracer gas bubble. The bubble wake shown in Fig. 3 plays an important role in solid movement or -mixing in the bed and the freeboard. A bubble wake in a single-phase fluid is defined as the streamline-enclosed region beneath the bubble base. In a gas–solid fluidized bed, the emulsion phase can be treated as a pseudosingle-phase fluid. Hence, a bubble wake is defined as the region enclosed by streamline of the pseudofluid behind the bubble base. In the bed, the wake rises with the bubble and thereby provides an essential means for global solid circulation and induced axial solids mixing.[17] In the freeboard, particles carried along the bubble wake are the primary origin of the particles there.

When gas enters the orifice of the distributor, it can initially form bubbles or jets. The formation of bubbles or jets depends on various parameters including types of particles, fluidization conditions around the orifice, orifice size, and the presence of internals in the bed. The initial bubble or jet is then transformed into a chain of bubbles. The jet is defined as an elongated void, which is appreciably larger than a bubble; it extends to some distance from the orifice into the inner bed. In general, bubbles tend to form in the presence of small particles, such as Group A particles; jets tend to form in the presence of large particles, such as Group D particles, when the emulsion phase is not sufficiently fluidized, or when internals are present and disrupt the solid flow to the orifice region.[18]

In bubbling fluidized beds under ambient pressure and low gas velocity conditions where the bubble size increases with the gas velocity, the bubble size may be estimated by various correlation equations, such as those developed by Mori and Wen,[19] Darton et al.,[20] and Cai et al.[21] There is a similarity in the rise behavior of a single bubble in gas–solid and liquid media. The rise velocity of a single spherical cap bubble in an infinite liquid medium can be described by the Davies and Taylor equation.[22] Experimental results indicate that the Davies and Taylor equation is valid for large bubbles with bubble Reynolds numbers greater than 40.[23] Whereas for bubbles in fluidized beds, the bubble Reynolds numbers are typically of the order of 10 or less. By analogy, the rise velocity of an isolated single spherical cap bubble in an infinite gas–solid medium can be expressed in terms of the volume bubble diameter by[1]

$$U_{b\infty} = 0.71\sqrt{g d_{b\infty}} \tag{8}$$

In a free bubbling bed, the average bubble-rise velocity U_{bb} can be described by[1]

$$U_{bb} = U - U_{mf} + 0.71\sqrt{g d_{bb}} \tag{9}$$

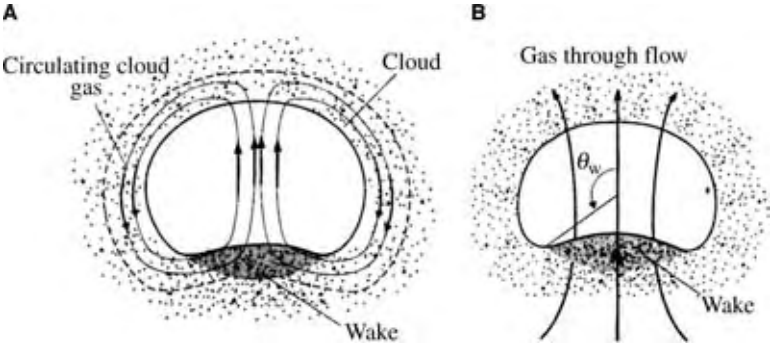


Fig. 3 Bubble configurations and gas flow patterns around a bubble in gas–solid fluidized beds: (A) fast bubble (clouded bubble) $U_b > U_{mf}/\alpha_{mf}$; (B) slow bubble (cloudless bubble) $U_b < U_{mf}/\alpha_{mf}$. (From Ref.[7].)

The distribution of gas flow in the fluidized bed is important for the analysis of the fundamental characteristics of transport properties in the bed. One common method to estimate the superficial gas flow division is based on the two-phase theory of fluidization, which considers the division of the superficial gas flow in the bed into two subflows—the bubble-phase and the emulsion-phase flow. According to the theory, the flow velocity can be generally expressed as

$$U = \alpha_b U_{bb} + U_{em} \quad (10)$$

where U_{em} and $\alpha_b U_{bb}$ are the superficial gas velocities for the emulsion-phase flow and for the bubble-phase flow, respectively. The variation of the relative magnitude of each flow depends on the gas velocity as well as on the gas and particle properties. A simple two-phase theory is to express U_{em} as U_{mf} and $\alpha_b U_{bb}$ as $(U - U_{mf})$. However, $U - U_{mf}$ may overestimate the actual bubble phase flow in a fluidized bed, as a result of two prevailing effects of the gas flow pattern: 1) significant emulsion-phase flow or invisible gas flow through the bubble, and 2) larger interstitial gas flow in the emulsion phase than U_{mf}.[24]

Turbulent fluidization

The turbulent regime is often regarded as the transition regime from bubbling fluidization to lean-phase fluidization. At relatively low gas velocities, bubbles are present in the turbulent regime, while at relatively high gas velocities in the turbulent regime, the clear boundary of bubbles disappears and the nonuniformity of solid concentration distribution yields distinct gas voids, which become less distinguishable as the gas velocity further increases toward lean-phase fluidization. The onset velocity of the transition to the turbulent regime is commonly defined, based on the variation of the amplitude of the pressure fluctuation with the gas velocity, as shown in Fig. 4, to be the gas velocity corresponding to the peak, U_c, whereas the leveling point, U_k, may be recognized as the onset of the turbulent regime proper.[25]

Entrainment and elutriation

The freeboard region in a fluidized bed accommodates particles that are being entrained from the dense bed. Entrainment refers to the ejection of particles from the dense bed into the freeboard by the fluidizing gas. Elutriation refers to the separation of fine particles from a mixture of particles, which occurs at all heights of the freeboard, and their ultimate removal from the freeboard. The terms entrainment and elutriation are sometimes used interchangeably. The carryover rate relates to the quantities of the particles leaving the freeboard. Coarse particles with a particle terminal velocity that is higher than the gas velocity eventually return to the dense bed, while the fine particles eventually exit from the freeboard. The freeboard height required in design consideration is usually higher than the transport disengagement height, defined as a height beyond which the solids holdup, solids entrainment, or carryover rate remains nearly constant.

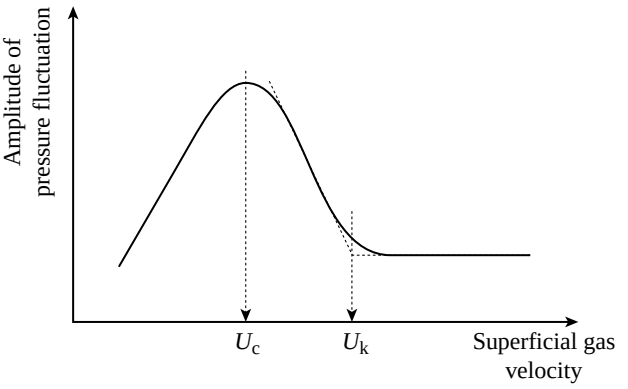


Fig. 4 Variation of pressure fluctuation with the gas velocity for dense-phase fluidized bed with FCC particles. (From Ref.[23].)

Slugging beds

When bubbles grow to sizes comparable to the bed diameter, slugging occurs. Slugging is most frequently encountered in small-diameter beds with large bed heights, especially when large/heavy particles are employed. There are requirements of a minimum slugging velocity and a minimum bed height for a slug flow to take place. The slugs may appear in different forms including the round-nosed slug, which occurs in systems of fine particles, the wall slug (also known as the half slug), which takes place in the beds with rough walls, and the square-nosed slug, which appears in coarse particle systems where the particle bridging effect is significant.

Spouted beds

In a spouted bed, gas enters the bed through a jet nozzle forming a spout. The surrounding annular region forms a downward-moving bed. Particles are entrained into the spout from the bottom and sidewall of the spout. A part of the gas seeps into the annular region through the spout wall, whereas the other part leaves the bed from the top of the spout. The particles carried into the spout disengage from the gas in a solid-disengagement fountain, just above the bed, and then return to the top of the annular region.[26] Group D particles are commonly used for the spouted bed operation.

Fast fluidization and dilute transport

The fast fluidization regime is characterized by the clustering phenomenon with a core-annular heterogeneous flow structure, whereas the dilute transport regime is characterized by a homogeneous flow structure. Lean-phase fluidization that encompasses these regimes is carried out in a circulating fluidized bed (CFB), where solid particles circulate between the riser and the downer. The fast fluidization regime is the principal regime under which the CFB is operated. The operating variables for a CFB system include both the gas flow rate and the solid circulation rate, whereas in a dense-phase fluidized bed system, only the gas flow rate is the operating variable. The solid circulation is established by a high velocity of gas flow.

The fast fluidization regime is represented by a dense region at the bottom of the riser and a dilute region above it. The inter-relationship of the fast fluidization regime with other fluidization regimes in dense-phase fluidization and with the dilute transport regime is reflected in the variations of the pressure drop per unit length of the riser, Δ_p/Δ_z, gas velocity, and solid circulation rate as given in Fig. 5.[27]

The transport velocity, U_{tr}, marks the lower limit of the gas velocity for fast fluidization operation. The transport velocity can be evaluated from the variations of the local pressure drop per unit length (Δ_p/Δ_z) with respect to the gas velocity and the solid circulation rate, J_p. An example of such a relationship is shown in Fig. 6. It is seen in the figure that, along the curve AB, the solid circulation rates are lower than that of the saturation carrying capacity of the flow. Particles with low particle terminal velocities are carried over from the riser. With an increasing solid circulation rate, more particles accumulate at the bottom. At point B in the curve, the solids fed into the riser are balanced by the saturated carrying capacity. A slight increase in the solid circulation rate yields a sharp increase in the pressure drop. This behavior reflects the collapse of the solid particles into a dense-phase fluidized bed and is noted as choking. When the gas velocity is equal to or higher than the

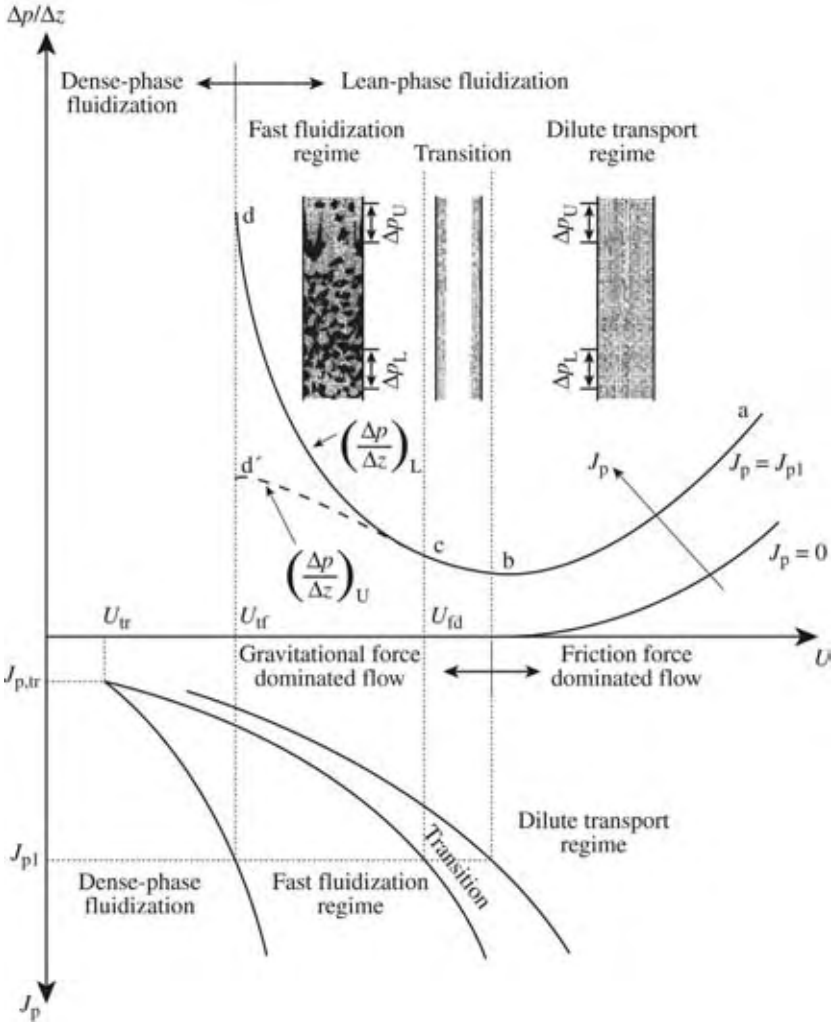


Fig. 5 Variation of pressure drop per unit riser length with solid circulation rate and gas velocity for various fluidization regimes. (From Ref.[27].)

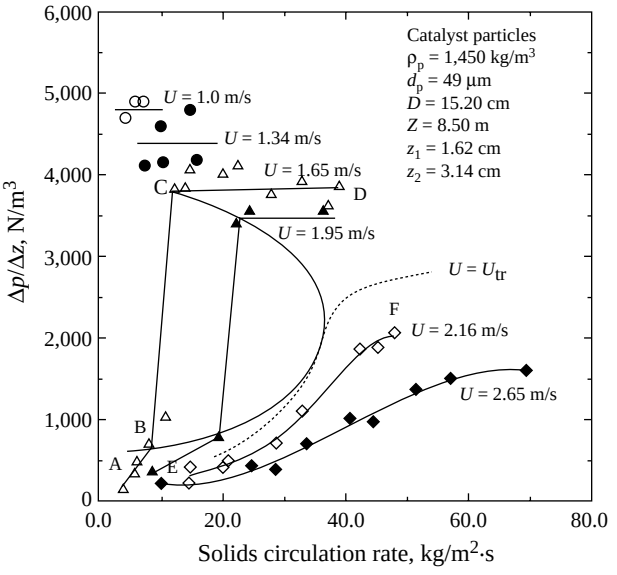


Fig. 6 Local pressure drop as a function of solid circulation rate at various gas velocities. (From Ref.[23].)

transport velocity (e.g., curve EF in Fig. 6), there is no longer a sharp increase in the slope of the J_p vs. (Δ_p/Δ_z) relationship. Thus, U_{tr} is characterized as the lowest gas velocity at which the discontinuity in the curve of J_p versus (Δ_p/Δ_z) disappears.[25] U_{tr} varies with gas and solid properties as correlated by Bi and Fan.[28]

OTHER SYSTEMS

Downer

In contrast to the riser, the downer involves downward flow of gas and solid in the gravitational direction at a high velocity. The downer is characterized by the absence of the minimum fluidization velocity, which allows a higher solid/gas loading ratio and a shorter flow development zone providing near plug flow conditions.[29] Thus, the downer has the benefits of uniform axial and radial gas–solid flow structures, thereby enhancing the reactant conversion for reactions in favor of the plug flow pattern.

Three-Phase Fluidized Bed

In gas–solid–liquid fluidization systems, the gas can be a continuous phase or discrete bubbles, while the liquid can be a continuous phase, a film or droplets, and the solids can be either in a continuous flow or a batch.[6]

	Mode Designation	E-I-a-1	E-I-a-2	E-I-b	E-II-a-1	E-II-a-2	E-II-b	E-III-a	E-III-b
Expanded Bed Regime in Gas-Liquid-Solid Fluidization	Schematic Diagram	G; L; (S→); (→S); L G	G; L(□S)⇐; ⇐L(+S); G	G; L; (S→); (→S); L G	LG; (S→); (→S); L G	LG; (S→); (→S); L G	LG; (S→); (→S); L	G; G	G; G
	Continuous Phase	Liquid		Gas	Liquid		Gas	Liquid	Gas
	Flow Direction	Cocurrent Up-Flow			Countercurrent Flow			Gas Up-Flow Liquid-Batch	
	References (Chapters)	1,2,3,6,7, 8,10,11,A	1,4,6,7,10,11,A	1,2,11	1,5,6,7,8	5,9,11	1,5,7,9,11	1,4,7,11	1,5
	Mode Designation	T-I-a-1	T-I-a-2	T-I-b	T-II-a	T-II-b	T-III-a		T-III-b
Transport Regime in Gas-Liquid-Solid Fluidization	Schematic Diagram	G L+S; G L□S	G; L+S⇐; ⇐L+S; G	G L+S; G L+S	L+S G; L+S G	L+S G; L+S G	G L+S; G L+S		L+S G; L+S G
	Continuous Phase	Liquid		Gas	Liquid	Gas	Liquid		Gas
	Flow Direction	Cocurrent Up-Flow			Countercurrent Flow		Cocurrent Down-Flow		
	References (Chapters)	1,6,7,8,11	1,4,10,11,A	1,6	1,9,11	1,9	1,6,11		1,6,9

(S→) Independent Charge of Solids from Fluid (→S) Independent Discharge of Solids from Fluid (+S) Charge or Discharge of Solids with Fluid
A: Appendix A

Fig. 7 Basic classification of gas–liquid–solid fluidization systems. (From Ref.[6].)

The operational modes of the three-phase fluidized beds can be classified by the directions of the gas and liquid flows: cocurrent upward, cocurrent downward, countercurrent, or crosscurrent as shown in Fig. 7.[30]

The dispersed (or homogeneous) bubble and coalesced bubble (or churn-turbulent) flow regimes are the most common in operating the three-phase fluidized bed systems.[9] In the dispersed bubble flow regime, the bubbles with a fairly uniform size distribution rise without significant bubble coalescence. The vortical-spiral flow condition occurs in the coalesced bubble flow regime at lower gas velocities and marks a transition between the dispersed bubble regime and the churn-turbulent flow.[31] As the gas velocity further increases, the turbulent flow condition is developed, as large bubbles are generated by intensive bubble coalescence. The addition of fine particles leads to larger bubble sizes, and thus accelerates the transition.[32]

High-pressure and high-temperature operation of three-phase fluidized beds is commonly encountered in most industrial applications of commercial interest.[9] The flow characteristics of reactors at high pressure and temperature are distinctly different from those in ambient conditions. For example, elevated pressure leads to higher gas holdup and smaller bubble size in the system and, thus, dramatically affects the transport phenomena, including heat and mass transfer, and phase mixing. The effect of the operating pressure on the regime transition has been examined by many researchers in bubble columns,[32–40] in three-phase fluidized beds,[41] and in slurry bubble columns.[42]

The effects of pressure and temperature on fluid dynamics and transport properties are mainly due to variations in bubble characteristics, such as bubble size and bubble size distribution, and changes in the physical properties of fluid phases. The bubble size and distribution in the bed are closely associated with the initial bubble size, bubble coalescence rates, and bubble breakup rates. Under high-pressure conditions, bubble coalescence is suppressed and bubble breakup is enhanced, yielding smaller bubble sizes and narrower bubble size distributions. Thus, the flow regime transition is delayed in high-pressure bubble columns and slurry bubble columns.[42] Increasing temperature also delays the regime transition. In the coal liquefaction reactors, a gas holdup as high as 0.50 with the bubble size that may be as small as 0.7 mm was reported.[33]

For three-phase fluidization systems, two distinct phenomena pertaining to macroscopic hydrodynamic behavior—bed contraction and moving packed bed flow—are noted below.

Bed contraction phenomenon

Three-phase fluidized beds using small particles display unique bed expansion characteristics. Upon the initial introduction of the gas into the liquid–solid fluidized bed, contraction, instead of expansion, of the bed occurs.[43] An increasing gas flow rate causes further contraction up to a critical gas flow rate beyond which the bed expands.[44,45] A quantitative elucidation of the bed contraction phenomenon was reported by Stewart and Davidson[46] and El-Temtamy and Epstein.[47] Basically, bed contraction is caused by the behavior of the bubble wake, which entraps liquid and particles and, therefore, is associated with large bubble systems. The entrainment of the liquid and particles by the bubble wake reduces the effective amount of liquid in the bed used to fluidize the remaining particles. The bed contraction phenomenon has been extensively studied under ambient fluidization conditions and has also been observed at high pressure.[48]

On the basis of the generalized wake model of Bhatia and Epstein,[49] a criterion for the bed contraction was developed.[49,50] In the generalized wake model, the three-phase fluidized bed is assumed to consist of three regions, the gas bubble region, the wake region, and the liquid–solid fluidized region. Bed contraction will occur when the following criterion is satisfied, $\psi < 0$ where

$$\Psi = \left(\frac{n}{n-1} + k\right)\frac{U_l}{\varepsilon_l} + \frac{xk\left(U_g/\varepsilon_g\right)}{n-1} - \left[(1+k)U_l + \frac{k}{n-1}\left(\frac{U_g}{\varepsilon_g} - \frac{U_l}{\varepsilon_l}\right)\right] + xk\left(U_l - \frac{U_l}{\varepsilon_l}\right)\left(\frac{n}{n-1}\right) \quad (11)$$

Here, n is the Richardson–Zaki index, k is the ratio of wake size to bubble size, and x is the ratio of solid concentration in the wake region to that in the liquid–solid fluidized region.

Moving packed bed phenomenon

The moving packed bed flow is characterized by the motion of solids in a piston flow in a three-phase fluidized bed. Moving packed bed flow, which usually occurs during the start-up, depends not only on the gas and liquid velocities, but also on how they are introduced into the bed. Moving packed bed flow is caused by the surface phenomena involving fine bubbles attached onto particles and subsequent formation of a fine bubble blanket under the packed solids; a liquid flow would move the entire bed upward. Thus, this phenomenon is associated with the small bubble system. The moving packed bed flow phenomenon in a three-phase fluidized bed is a known, anomalous event in the resid hydrotreating industry. This was observed in the 1960s, in the bench and pilot units, during the development and commercialization of the

resid hydrotreating process.[51] The reactor was typically operated at pressures between 5.5 and 21 MPa and temperatures between 300°C and 425°C. In the early 1970s, moving packed bed flow was observed in a commercial three-phase fluidized bed reactor. The occurrence of a moving packed bed in a three-phase fluidized bed could simply be circumvented by utilizing a start-up procedure that involves degassing the bed first and then introducing liquid flow to expand the bed prior to commencing the gas flow. Commercial operators of three-phase fluidized bed reactors have long recognized and undertaken a proper start-up procedure of this nature since observing this anomalous event. As small bubbles can also be generated under ambient conditions using surfactants in an air–water system, the moving packed bed flow phenomenon was reported in open literature first by Saberian-Broudjenni et al.[52] and later by Bavarian and Fan[53,54] in small columns with small bubbles generated in the same manner.

COMPUTATIONAL FLUID DYNAMICS

The discrete phases including particles, bubbles, and droplets in multiphase fluidization systems can be simulated numerically using the Eulerian continuum, and Lagrangian and direct simulation for solid particles, or the front tracking method for deformable surfaces, such as bubbles and droplets. In the continuum method, the individual phases are treated as pseudocontinuous fluids, each being governed by conservation laws expressed in terms of the volume per unit time or ensemble-averaged properties.[55–60] Numerical simulations using the continuum method have been extensively conducted for gas–solid,[61–64] gas–liquid,[65–68] and gas–liquid–solid flows.[68–71]

There are two main approaches for the numerical simulation of the gas–solid flow:[72] 1) Eulerian framework for the gas phase and Lagrangian framework for the dispersed phase (E–L) and 2) Eulerian framework for all phases (E–E). In the E–L approach, trajectories of dispersed phase particles are calculated by solving Newton's second law of motion for each dispersed particle, and the motion of the continuous phase (gas phase) is modeled using an Eulerian framework with the coupling of the particle–gas interaction force. This approach is also referred to as the distinct element method or discrete particle method when applied to a granular system.[73] The fluid forces acting upon particles would include the drag force, lift force, virtual mass force, and Basset history force.[74] Moreover, particle–wall and particle–particle collision models (such as hard sphere model, soft sphere model, or Monte Carlo techniques) are commonly employed for this approach. In the E–E approach, the particle cloud is treated as a continuum.[69] Local mean variables are utilized instead of point variables to describe the motions of both phases on the basis of the local average technique of Anderson and Jackson.[55] This approach is more suitable for modeling the dense multiphase system with a significant volume fraction for the dispersed phase (>10%).

Because the effects of turbulence on particle motion are significant for most gas–solid fluidization systems, the numerical modeling of the carrier flow turbulence and motion of particles dispersed in a turbulent flow have been the subjects of extensive research. Direct numerical simulation (DNS) resolves the smallest scale structures (Komolgorov length scale) and can provide detailed point information for the flow field. No empirical closure models are needed in the DNS scheme. Large Eddy simulation (LES) involves both direct simulation and the Reynolds-averaging method. For the LES, large-scale motions are resolved rigorously, and the small-scale (sub-grid) motion is assumed to be homogeneous and independent of flow geometry.

On computation of gas–liquid bubble columns and gas–liquid–solid fluidized beds, numerical simulations provide a useful description of the hydrodynamics of bubble flows in liquids.[67,70,75–78] In the E–E method, both the continuous liquid phase and the dispersed bubble phase are treated as interpenetrating continua, occupying the same space based on the concept of volume averaging, with different velocities and volume fractions for each phases. In the E–L method,[75,77] on the other hand, liquid is treated as a continuous phase as described in the Eulerian mode, and bubbles are treated as a dispersed phase as tracked in the Lagrangian mode. Both the E–E and the E–L methods have proven to be more effective in modeling the gas–liquid flow in the homogeneous regime than in the heterogeneous regime.

More recently, the level-set technique has been developed for computing interfacial motion in two or three dimensions.[79] It is especially effective in simulating large topological changes such as bubble breaking and merging.

CONCLUSIONS

The fluidized bed systems have been utilized extensively in many physical, chemical, petrochemical, electrochemical, and biochemical processes. Successful applications of the fluidization systems lie in a comprehensive understanding of hydrodynamics, heat and mass transfer properties, and mixing. Various nonintrusive measurement techniques, such as electric capacitance tomography and radioactive particle tracking technique, are available to advance the fundamental understanding of the microscopic and macroscopic phenomena of fluidization. Till date, the

design of the fluidized beds, however, still relies heavily on empirical relationships and engineering models. The computational fluid dynamics approach to fluidized bed simulation has gained considerable attention in recent years. However, many challenges remain, such as the formulation of closure relationships on turbulence. Fluidization of nanoparticles (<100 nm) is an area of appreciable fundamental and applied interest for future research and development efforts.

NOMENCLATURE

Ar	Archimedes number
d_{bb}	Averaged volume diameter of bubble in the bed
$d_{b\infty}$	Volume diameter of an isolated bubble
d_p	Particle diameter
g	Gravitational acceleration
H	Vertical distance from gas distributor
H_{mf}	Bed height at minimum fluidization
p_d	Dynamic pressure
Re_{pmf}	Particle Reynolds number at minimum fluidization velocity
U	Superficial gas velocity
U_{bb}	Average bubble-rise velocity in the whole bed
U_{mf}	Superficial gas velocity at the minimum fluidization
α	Bed voidage
α_b	Volume fraction of the bed occupied by bubbles
α_{mf}	Bed voidage at minimum fluidization
ϵ	Bed voidage
Δp_b	Pressure drop across the bed
μ	Dynamic viscosity
ρ	Density of fluid
ρ_p	Density of particle
ϕ	Sphericity of particles

REFERENCES

1. Davidson, J.F.; Harrison, D. *Fluidized Particles*; Cambridge University Press: Cambridge, 1963.
2. Davidson, J.F.; Clift, R.; Harrison, D., Eds.; *Fluidization*, 2nd Ed.; Academic Press: London, 1985.
3. Kunii, D.; Levenspiel, O. *Fluidization Engineering*, 2nd Ed.; Butterworth-Heinemann: Boston, 1991.
4. Rietema, K. *The Dynamics of Fine Powders*; Elsevier Applied Science: London, 1991.
5. Gidaspow, D. *Multiphase Flow and Fluidization: Continuum and Kinetic Theory Descriptions*; Academic Press: New York, 1994.
6. Fan, L.S. *Gas–Liquid–Solid Fluidization Engineering*; Butterworths: Boston, 1989.
7. Fan, L.S.; Zhu, C. *Principles of Gas–Solid Flows*; Cambridge University Press: Cambridge, 1998.
8. Jackson, R. *The Dynamics of Fluidized Particles*; Cambridge University Press: Cambridge, 2000.
9. Grace, J.R.; Aridan, A.A.; Knowlton, T.M., Eds.; *Circulating Fluidized Beds*; Blackie Acasdemic and Professional: New York, 1997.
10. Geldart, D. Types of gas fluidization. Powder Tech. **1973**, *7*, 285.
11. Wen, C.Y.; Yu, Y.H. Mechanics of fluidization. Chem. Eng. Prog. Symp. Ser. **1966**, *62* (62), 100.
12. Richardson, J.F.; Zaki, W.N. Sedimentation and fluidization, part I. Trans. Inst. Chem. Eng. **1954**, *43*, 35.
13. Fan, L.S.; Tsuchiya, K. *Bubble Wake Dynamics in Liquids and Liquid–Solid Suspensions*; Butterworths: Boston, 1990.
14. Anderson, T.B.; Jackson, R.A. Fluid mechanical description of fluidized beds: stability of the state of uniform fluidization. I&EC Fundam. **1968**, *7*, 12.
15. Verloop, J.; Heertjes, P.M. Shock waves as a criterion for the transition from homogeneous to heterogeneous fluidization. Chem. Eng. Sci. **1970**, *25*, 825.
16. Rietema, K.; Piepers, H.W. The effect of interparticle forces on the stability of gas-fluidized beds: I. Experimental evidence. Chem. Eng. Sci. **1990**, *45*, 1627.
17. Clift, R.; Grace, J.R.; Weber, M.E. *Bubbles, Drops and Particles*; Academic Press: New York, 1978.
18. Massimilla, L. Gas jets in fluidized beds. In *Fluidization*, 2nd Ed.; Davidson, J.F., Clift, R., Harrison, D., Eds.; Academic Press: London, 1985.
19. Mori, S.; Wen, C.Y. Estimation of bubble diameter in gaseous fluidized beds. AIChE J. **1975**, *21*, 109.
20. Darton, R.C.; La Nauze, R.D.; Davidson, J.F.; Harrison, D. Bubble growth due to coalescence in fluidized beds. Trans. Inst. Chem. Eng. **1977**, *55*, 274.
21. Cai, P.; Schiavetti, M.; DeMichele, G.; Grazzini, G.C.; Miccio, M. Quantitative estimation of bubble size in PFBC. Powder Tech. **1994**, *80*, 99.
22. Davies, L.; Taylor, G.I. The mechanics of large bubbles rising through extended liquids and through liquids in tubes, Proc. R. Soc. Lond. **1950**, *A200*, 375.
23. Clift, R. Hydrodynamics of bubbling fluidized beds. In *Gas Fluidization Technology*; Geldart, D., Ed.; John Wiley & Sons: New York, 1986.
24. Clift, R.; Grace, J.R. Continuous bubbling and slugging. In *Fluidization*, 2nd Ed.; Davidson, J.F., Clift, R, Harrison, D., Eds.; Academic Press: London, 1985.

25. Yerushalmi, J.; Cankurt, N.T. Further studies of the regimes of fluidization. Powder Tech. **1979**, *24*, 187.
26. Mathur, K.B.; Epstein, N. *Spouted Beds*; Academic Press: New York, 1974.
27. Bai, D.; Jin, Y.; Yu, Z. Flow regimes in circulating fluidized beds. Chem. Eng. Techno. **1993**, *16*, 307.
28. Bi, H.T.; Fan, L.-S. Existence of turbulent regime in gas–solid fluidization. AIChE J. **1992**, *38*, 297.
29. Jin, Y.; Zheng, Y.; Wei, F. State-of-the-art review of downer reactors. In *Circulating Fluidized Bed Technology VII*, Proceedings of the 7th International Conference on Circulating Fluidized Beds, Niagara Falls, Ontario, Canada, May 5–8, 2002; Grace, J.R., Zhu, J.-X., de Lasa, H., Eds.; Canadian Society of Chemical Engineers: Ottawa, 2002; 40–60.
30. Fan, L.S. *Gas-Liquid-Solid Fluidization Engineering*; Butterworth: Boston, 1989.
31. Chen, R.C.; Fan, L.S. Particle image velocimetry for characterizing the flow structure in three-dimensional gas–liquid–solid fluidized beds. Chem. Eng. Sci. **1992**, *47*(13,14), 3615–3622.
32. Clark, K.N. The effect of high pressure and temperature on phase distributions in a bubble column. Chem. Eng. Sci. **1990**, *45*, 2301.
33. Tamy, B.; Chang, M.; Coulaloglou, C.; Ponzi, P. Hydrodynamic characteristics of three phase reactors. Chem. Eng. **1984**, *407*, 18.
34. Krishna, R.; Wilkinson, P.M.; Van Dierendonck, L.L. A model for gas holdup in bubble columns incorporating the influence of gas density on flow regime transitions. Chem. Eng. Sci. **1991**, *46*, 2491.
35. Krishna, R.; Swart, J.W.A.; Hennephof, D.E.; Ellenberger, J.; Hoefsloot, C.J. Influence of increased gas density on hydrodynamics of bubble-column reactors. AIChE J. **1994**, *40*, 112.
36. Wilkinson, P.M.; Sper, A.P.; Van Dierendonck, L.L. Design parameters estimation for scale-up of high-pressure bubble columns. AIChE J. **1992**, *38*, 544.
37. Hoefsloot, H.C.J.; Krishna, R. Influence of gas density on the stability of homogeneous flow in bubble columns. Ind. Eng. Chem. Res. **1993**, *32*, 747.
38. Reilly, I.G.; Scott, D.S.; de Bruijn, T.J.W.; MacIntyre, D. The role of gas phase momentum in determining gas holdup and hydrodynamic flow regimes in bubble column operations. Can. J. Chem. Eng. **1994**, *72*, 3.
39. Letzel, H.M.; Schouten, J.C.; Van den Bleek, C.M.; Krishna, R. Influence of elevated pressure on the stability of bubbly flows. Chem. Eng. Sci. **1997**, *52*, 3733.
40. Lin, T.J.; Tsuchiya, K.; Fan, L.S. Bubble flow behavior at high pressure bubble column. Can. J. Chem. Eng. **1999**, *77*, 370–374.
41. Fan, L.-S.; Yang, G. Gas–liquid–solid three-phase fluidization. In *Handbook of Fluidization and Fluid-Particle Systems*; Yang, W.-C., Ed.; Marcel Dekker: New York, 2003; Chapter 27, 765–810.
42. Fan, L.S.; Yang, G.Q.; Lee, D.J.; Tsuchiya, K.; Luo, X. Some aspects of high pressure phenomena of bubbles in liquids and liquid-solid suspensions. Chem. Eng. Sci. **1999**, *54*, 4681.
43. Massimilla, L.; Majuri, N.; Signorini, P. Sull'assorbimento Di Gas in Sistema: Solido-Liquido Fluidizzato. La Ricerca Scientifica **1959**, *29*, 1934.
44. Turner, R. Fluidization. Soc. Chem. Ind. Lond. **1964**, 47.
45. Ostergaard, K. Fluidization. Soc. Chem. Ind. Lond. **1964**, 58.
46. Stewart, P.S.B.; Davidson, J.F. Three-phase fluidization: water, particles and air. Chem. Eng. Sci. **1964**, *19*, 319.
47. El Temtamy, S.A.; Epstein, N. Contraction or expansion of three-phase fluidized beds containing fine/light solids. Can. J. Chem. Eng. **1979**, *57* (4), 520–522.
48. Jiang, P.; Lin, T.J.; Fan, L.-S. High temperature and high pressure three-phase fluidization—bed expansion phenomena. Powder Tech. **1997**, *90*, 103.
49. Bhatia, V.K.; Epstein, N. Three-phase fluidization: generalized wake model. In *Fluidization and Its Applications*, Proceedings of the International Symposium, Toulose, October 1–5, 1973; Angelion, H.; Coudere, J.P.; Gibert, H.; Laguerie, C., Eds.; Cepadues-Editions: Toulose, 1974, 380–392.
50. Jean, R.H.; Fan, L.S. Bed contraction criterion for three phase fluidization. Can. J. Chem. Eng. **1987**, *65*, 351–352.
51. Fan, L.S. Moving packed bed phenomenon in three-phase fluidization. Powder Tech. **1999**, *103*, 300.
52. Saberian-Broudjenni, M.; Wild, G.; Charpentier, J.C.; Fortin, Y.; Euzen, J.P.; Patoux, R. Contribution à L'ètude Hydrodynamique Des Rèacteurs à Lit Fludisè Gaz-Liquide-Solide. Entropie **1984**, *120*, 30.
53. Bavarian, F.; Fan, L.-S. Mechanisms of hydraulic transport of a packed bed at the start-up of a three-phase fluidized bed. Chem. Eng. Sci. **1991**, *46*, 3081.
54. Bavarian, F.; Fan, L.S. Hydraulic transport of a packed bed during the start-up of a three-phase fluidized bed with large gas holdups. Ind. Eng. Chem. Res. **1991**, *30*, 408.

55. Anderson, T.B.; Jackson, R. A fluid mechanical description of fluidized beds. I&EC Fundam. **1967**, *6*, 527–539.
56. Ishii, M. *Thermo-Fluid Dynamic Theory of Two-phase Flows*; Eyrolles: Paris, 1975.
57. Drew, D.A. Averaged field equations for two phase media. Stud. Appl. Math. **1971**, *50*, 133.
58. Nigmatulin, R.I. Spatial averaging in the mechanics of heterogeneous and dispersed systems. Int. J. Multiphase Flow **1979**, *5*, 353.
59. Zhang, D.Z.; Prosperetti, A. Averaged equations for inviscid disperse two-phase flow. J. Fluid Mech. **1994**, *267*, 185.
60. Jackson, R. Locally averaged equations of motion for a mixture of identical spherical particles and a Newtonian fluid. Chem. Eng. Sci. **1997**, *52*, 2457.
61. Sinclair, J.L.; Jackson, R. Gas–particle flow in a vertical pipe with particle–particle interactions. AIChE J. **1989**, *35*, 1473.
62. Ding, J.; Gidaspow, D. A bubbling fluidization modeling using kinetic theory of granular flow. AIChE J. **1990**, *36*, 523.
63. Pita, J.A.; Sandaresan, S. Developing flow of a gas–particle mixture in a vertical riser. AIChE J. **1993**, *39*, 541.
64. Dasgupta, S.; Jackson, R.; Sundaresan, S. Turbulent gas–particle flow in vertical risers. AIChE J. **1994**, *40*, 215.
65. Torvik, R.; Svendsen, H.F. Modeling of slurry reactors: a fundamental approach. Chem. Eng. Sci. **1990**, *45*, 2325.
66. Svendsen, H.F.; Jakobsen, H.A.; Torvik, R. Local flow structures in internal loop and bubble column reactors. Chem. Eng. Sci. **1992**, *47*, 3297.
67. Sokolichin, A.; Eigenberger, G. Gas–liquid flow in bubble columns and loop reactor: part I. Detailed modeling and numerical simulation. Chem. Eng. Sci. **1994**, *49*, 5735.
68. Boisson, N.; Malin, M.R. Numerical prediction of two-phase flow in bubble columns. Int. J. Numer. Methods Fluids **1996**, *23*, 1289.
69. Gidaspow, D.; Bahary, M.; Jayaswal, U.K. In *Hydrodynamic Models for Gas–Liquid–Solid Fluidization: Numerical Methods in Multiphase Flows*; FED, ASME: New York, 1994; Vol. 185, 117–124.
70. Grevskott, S.; Sannas, B.H.; Dudukovic, M.P.; Hjarbo, K.W.; Svendsen, H.F. Liquid circulation, bubble size distributions, and solids movement in two- and three-phase bubble columns. Chem. Eng. Sci. **1996**, *51*, 1703.
71. Mitra-Majumdar, D.; Farouk, B.; Shah, Y.T. Hydrodynamic modeling of three-phase flows through a vertical column. Chem. Eng. Sci. **1997**, *52*, 4485.
72. Crowe, C.; Sommerfeld, M.; Tsuji, Y. *Multiphase Flows with Droplets and Particles*; CRC Press LLC: Boca Raton, 1998.
73. Xu, B.H.; Yu, A.B. Numerical simulation of the gas–solid flow in a fluidized bed by combining discrete particle method with computational fluid dynamics. Chem. Eng. Sci. **1997**, *52* (16), 2785.
74. Ranade, V.V. In *Computational Flow Modeling for Chemical Reactor Engineering*; Process Systems Engineering; Academic Press: San Diego, 2002; Vol. 5.
75. Lapin, A.; Lűbbert, A. Numerical simulation of the dynamics of two-phase gas–liquid flows in bubble columns. Chem. Eng. Sci. **1994**, *49*, 3661–3674.
76. Sokolichin, A.; Eigenberger, G. Applicability of the standard k-e turbulence model to the dynamic simulation of bubble column: part 1. Detailed numerical simulation. Chem. Eng. Sci. **1999**, *54*, 2273–2284.
77. Delnoij, E.; Kuipers, J.A.M.; Van Swaaij, W.P.M. Computational fluid dynamics applied to gas–liquid contactors. Chem. Eng. Sci. **1997**, *52*, 3623–3638.
78. Mudde, R.F.; Simonin, O. Two- and three-dimensional simulations of bubble plume using a two-fluid model. Chem. Eng. Sci. **1999**, *54*, 5061–5069.
79. Chen, C.; Fan, L.-S. Discrete simulation of gas–liquid bubble columns and gas–liquid–solid fluidized beds. AIChE J. **2004**, *50* (2), 288.

Fluidized Bed Reactor

John R. Grace
University of British Columbia, Vancouver, British Columbia, Canada

Jamal Chaouki
Ecole Polytechnique, Montreal, Quebec, Canada

Todd Pugsley
University of Saskatchewan, Saskatoon, Saskatchewan, Canada

INTRODUCTION

This entry covers key features of fluidized bed reactors for catalytic and noncatalytic reactions involving solid particles and one or more fluids, most commonly a gas, and also liquids and gas–liquid mixtures. Fluidized beds find wide application in several industries—chemical processes, petrochemicals, polymers, mineral processing, pharmaceuticals, and food processing—because of their unique features that are advantageous in a number of applications. Enough background is provided for the reader to understand the major advantages and disadvantages of fluidized bed reactors, the most important design challenges and considerations, and the principal applications. Emphasis is placed on gas–solid-fluidized beds, but liquid–solid and gas–liquid–solid (three-phase) fluid bed reactors are also treated in brief. For more extensive coverage, the reader should consult standard reference works on fluidization.[1–6]

KEY FEATURES OF GAS–SOLID-FLUIDIZED BED REACTORS

The advantages of gas–solid-fluidized bed reactors relative to packed bed reactors are:

- Greatly improved bed-to-wall and bed-to-immersed-surface heat transfer
- Reduced axial and lateral temperature gradients, minimizing the probability of hot spots, catalyst sintering, and unwanted side reactions
- Ability to add or remove particles continuously or intermittently, without shutting down the process
- Reduced pressure drops (The pressure drop across the bed, once fluidized, essentially remains equal to only that required to support the weight of the bed.)
- Smaller catalyst particles, leading to improved catalyst effectiveness factors
- Ability to introduce (usually as a spray) modest quantities of liquid reactants that vaporize before reacting or yield solid products upon reaction inside the bed.

Fluidized beds, however, have some important disadvantages relative to packed (fixed or moving) bed reactors:

- Substantial axial gas mixing, causing much larger deviations from plug flow than for packed bed reactors, thereby adversely affecting conversions and selectivities
- For reactions where the particles themselves react, substantial particle mixing, greatly broadening solid residence time distributions relative to moving beds
- Particle attrition because of particles colliding with each other and with fixed surfaces
- Wear on immersed tubes and other interior surfaces because of the particle impingement
- Entrainment of particles, causing loss of catalyst and/or solid product, contributing to air pollution and requiring gas–solid separation equipment
- Increased risk because of complex hydrodynamics and difficulties in characterizing and predicting reactor performance

Gas–solid contacting and axial dispersion of both gas and particles depend on the particle properties, operating conditions, column geometry, and scale, all of which affect the motion of gas and solids, commonly referred to as hydrodynamics. For extensive coverage of hydrodynamics[2–6] may be consulted. Here we summarize the key features that affect the performance of gas–solid-fluidized beds as chemical reactors.

Overall Configuration: Common configurations for fluidized bed reactors are shown in Fig. 1. Bubbling beds and turbulent beds [Fig. 1(a)] usually operate at excess superficial gas velocities ($U - U_{mf}$) of $<0.25\,m/s$ and between ~ 0.3 and $1.2\,m/s$, respectively,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007772

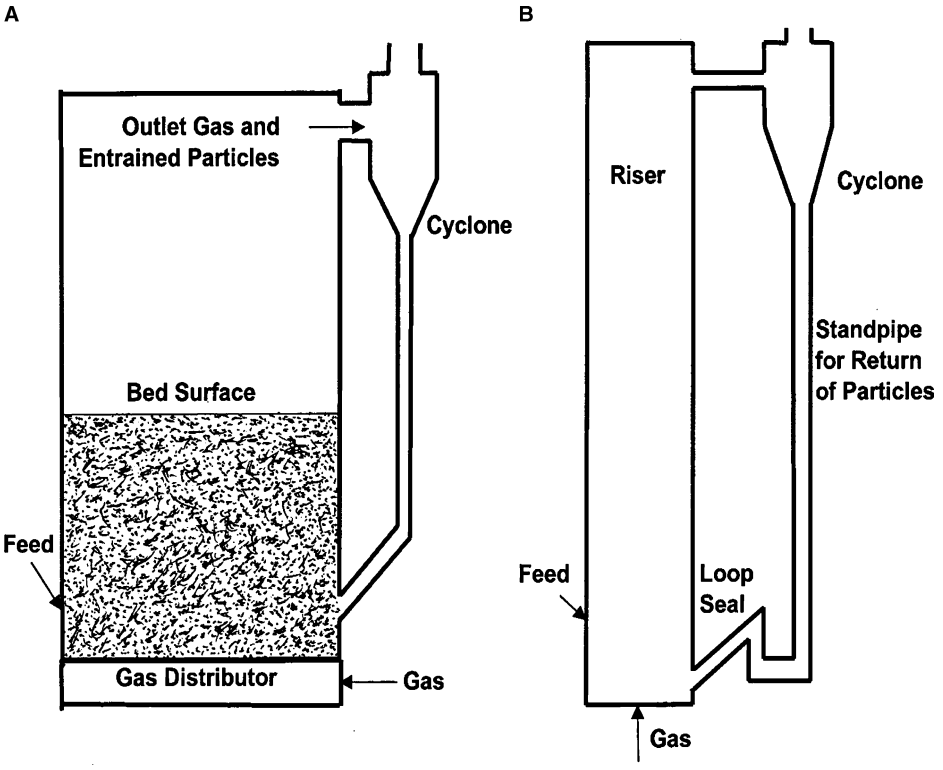


Fig. 1 Typical equipment configurations for: (A) dense (bubbling, slugging or turbulent) fluidized bed reactors; (B) circulating fluidized bed reactors. Heat transfer surfaces or baffles that might be present in (A) or (B) are not shown.

whereas circulating beds (Fig. 1B) generally feature superficial gas velocities of 2–12 m/s.

Particle Properties: Particles should be free-flowing (not sticky and with rounded shapes), ideally within either group A or group B of the powder classification scheme of Geldart.[7] Catalyst particles primarily belong to group A (mean diameter 50–100 μm), whereas particles that react in the bed fall into group B or D (mean diameter 200–2000 μm). Particle size distributions should be reasonably broad, e.g., covering a range from 10 to 200 μm. Particle densities should, however, be as uniform as possible. The particles must be able to withstand frequent collisions without breaking. Ideally, they should also not be subject to major electrostatic charging effects.

Operating Pressure and Temperature: Fluidized beds generally operate more smoothly with increasing absolute pressure, so that elevated pressures do not present a problem from a fluidization point of view. Many fluidized bed reactors also operate at high temperatures. The effects of temperature differ from system to system, probably because temperature affects particle properties (e.g., surface hardness and stickiness) in addition to causing well-characterized changes in gas properties. The effects of pressure and temperature have been reviewed by Yates.[8]

Gas Distributor: The gas distributor plate at the bottom of the reactor must introduce gas uniformly, prevent hole-plugging or weeping of solids, promote good gas–solid contacting, and support the weight of the bed when it is defluidized. During steady state operation, it is usually important to ensure that the pressure drop across the distributor is at least 30% of that across the bed itself. Gas introduction points should not be more than 0.3 m apart. Many different distributor geometries are found in industry, including tuyères, bubble caps, pipe grids with orifices oriented obliquely downward, and multiorifice plates (see Ref.[9] for more details).

Internals: Fluidized bed reactors commonly contain heat transfer surfaces, and baffles are sometimes added to improve gas–solid contacting and/or to reduce axial dispersion. Surfaces should, whenever possible, be horizontal or vertical, not inclined at other angles. The minimum gap between adjacent surfaces should be at least 20–30 mean particle diameters and several times the maximum particle dimension to prevent bridging and defluidization. Internal surfaces are subjected to buffeting, buoyancy forces, and impingement of particles, causing wear and/or particle attrition. These issues need to be considered in the design process. To prevent fouling and build-up of stagnant solids on top of the surfaces, flat surfaces should be inclined at 60° or more to the horizontal.

Particle Feeding: Many fluidized bed reactors require that particles be fed continuously or intermittently. Pneumatic feeding and screw feeders are most common. Because lateral mixing is much less than axial mixing in fluidized beds, it may be essential to introduce solids at a number of different positions

around the periphery of the reactor to minimize horizontal gradients. Lock hoppers may be needed if the reactor operates at elevated pressures. Preventing backflow of gases from the reactor may be essential for reasons of safety.

Entrainment: Most fluidized bed reactors employ one or more cyclone, either inside the freeboard region at the top of the vessel or located externally, to capture entrained solids that are then returned continuously to the base of the fluidized bed via a standpipe and a mechanical (e.g., slide) valve or aerated nonmechanical valve (see Ref.[10] for details of solid return systems). A ''flapper'' gate, acting as a check valve, is commonly employed to prevent backflow of gas up the standpipe. While cyclones are by far the most popular, other gas–solid separators like impingement separators, electrostatic precipitators, filters, and scrubbers are sometimes provided, especially as second- or third stage separators.

Flow Regime: Fluidized beds may operate in several different distinct flow regimes—bubbling, slug flow, turbulent, fast-fluidization, or dense suspension upflow. In principle, they may operate in any of these flow regimes. In practice, however, slug flow is seldom experienced in commercial reactors because of their large diameter and limited H/D, although slugging is common in laboratory and pilot-scale fluid bed reactors. Industrial reactors used for solid-catalyzed reactions seldom operate in the bubbling regime because the interphase mass transfer resistance would then severely limit gas–solid contacting. However, bubbling is often relevant for the larger particles found in gas–solid reactions. Catalytic reactors most often operate in the turbulent or dense suspension upflow regime, because of favorable gas–solid contacting combined with limited gas axial mixing, whereas the turbulent and the fast-fluidization regimes are common for gas–solid reactions.

FLUIDIZED BED CATALYTIC REACTOR PROCESSES

The advantages of fluidized bed reactor technology, discussed earlier, have led to several commercial fluidized bed catalytic processes. The key ones are surveyed here.

Fluid Catalytic Cracking

Fluid catalytic cracking (FCC) represents the major industrial application of fluidized bed technology. As of 1997, there were over 350 units operating worldwide, processing more than 16 million barrels of feed per day.[11] For a summary of typical FCC process conditions, see Ref.[11].

The process is operated as a circulating fluidized bed, with fine catalyst particles continuously circulated between the riser, where the cracking reactions occur and catalyst is deactivated because of coke deposition, and the regenerator, where the coke on the catalyst particles is combusted. Matsen[12] reported the slip factor in FCC risers to be approximately 2, which is consistent with radial segregation of the catalyst, leading to radial variations in gasoline yield.[13,14] Data[15] indicate that the catalyst particles travel upward at the wall of the riser because of the high gas velocity and solid flux. This results in reduced gas backmixing, thereby minimizing overcracking of gasoline.

In early FCC process designs, the regenerator was operated as a bubbling or a turbulent fluidized bed combustor. Mobil and UOP developed riser regenerators in the 1970s.[11] The riser regenerator is advantageous in operations in which there is no external CO recovery boiler. With today's trend of processing heavier and more variable feedstocks in FCC units, the preference is to operate in a partial combustion mode in the regenerator, in which case riser regenerators are of little benefit. Most modern FCC systems use a turbulent fluidized bed regenerator.

Partial Oxidation Reactions

This class of reactions, carried out in fluidized beds, involves parallel and series reactions, with reaction intermediates being the desired products. Industrial examples include partial oxidation of *n*-butane to maleic anhydride and *o*-xylene to phthalic anhydride. The vigorous solid mixing of fluidized beds is valuable for these reactions because they are highly exothermic. However, gas backmixing must be minimized to avoid extended gas residence times that lead to the formation of products of total combustion (i.e., CO_2 and H_2O). For this reason, fluidized bed catalytic partial oxidation reactors are operated in the higher velocity regimes of turbulent and fast-fluidization.

Most fluidized bed partial oxidation processes are operated in the turbulent flow regime of fluidization. However, DuPont operated a circulating fluidized bed catalytic reactor process for maleic anhydride production in Spain,[16] featuring regeneration of the catalyst (by oxidation) on the downcomer side of the circulating system.

Fischer–Tropsch Synthesis

Fischer–Tropsch reactions convert synthesis gas into hydrocarbons. The fluidized bed process for this reaction, known as the Synthol process, was developed

by Sasol of South Africa to produce synthetic gasoline from coal. Duvenhage and Shingles[17] describe the development of the Sasol reactor technology from pilot studies in the 1940s to today's commercial operation. Sasol 1A and 1B circulating fluidized bed reactors started up in 1955 and were fraught with problems: the planned run length of 340 days was in actuality more like 40 days. Many of the problems were resolved over 20 years,[17] and two additional trains of larger circulating fluidized bed reactors were commissioned uneventfully in the early 1980s. In the late 1980s, Sasol began pilot studies that led to the installation of two large turbulent fluidized bed reactors in 1995 and 1999. Among the advantages of this design are lower installed and operating costs associated with smaller vessels, reduced bed pressure drops, and lower catalyst usage.[17] The Sasol reactors rely on immersed heat exchangers to remove the heat of reaction. Plugging and cutting of the tube sheet coolers in the risers of the Sasol 1A and 1B reactors proved to be a problem.[17] The tube sheets were eventually replaced with cooling coils because of the tendency for tube sheet coolers to plug with solids.[18]

Oxychlorination of Ethylene

Vinyl chloride monomer, the basic building block of polyvinylchloride (PVC), is commercially manufactured by dehydrochlorination of 1,2-dichloroethane. The modern process for producing 1,2-dichloroethane involves oxychlorination of ethylene in a fluidized bed catalytic reactor:

$$C_2H_4 + 2HCl + \frac{1}{2}O_2 \rightarrow Cl_2C_2H_4 + H_2O \quad (1)$$

This is a highly exothermic reaction that takes place in the gas phase in the presence of a $CuCl_2$ catalyst. Small amounts of alkali and rare-earth metals in the catalyst inhibit by-product formation. The reactor is operated at 220–240°C and 2–4 bar. Immersed heat exchangers again remove the excess exothermic heat of reaction.

Propylene Ammoxidation

Ammoxidation refers to the catalytic oxidation of a feedstock with ammonia. When propylene is the feedstock, acrylonitrile is produced. Most of the world's acrylonitrile is based on the Sohio (now BP) process in which stoichiometric amounts of propylene and ammonia are reacted with a slight excess of air in a fluidized bed operated in the turbulent fluidization flow regime. The reactor temperature and pressure are 450°C and 1.5 bar, respectively. The reaction usually takes place over a bismuth molybdenum oxide catalyst; other metal oxides (iron–antimony, uranium–antimony, and tellurium–molybdenum) have also been used. The catalyst contains iron compounds that increase selectivity to acrylonitrile. The exothermic heat of reaction is removed by immersed heat exchangers in which boiler feed water is circulated and high-pressure steam is produced.

Polymerization

The gas-phase polymerization of ethylene to produce linear low-density polyethylene (LLDPE) is often carried out in fluidized bed reactors. The Unipol process technology is licensed worldwide by Union Carbide. It can also be used to make high-density polyethylene (HDPE). In the early years, ethylene was fed to the reactor with a 1-butene comonomer. The process was later modified to introduce higher α-olefins such as 1-hexene, yielding stronger LLDPE. The LLDPE resin is produced with Ziegler catalysts in large fluidized bed reactors (~5 m tall and 3–4 m in diameter) operated at pressures of 15–25 bar and temperatures of 70–90°C. Unipol reactors employ flared sections at the top to promote gas–solid disengagement. The monomer/comonomer mixture, together with nitrogen and hydrogen, is fed to the reactor in which the solid resin is vigorously fluidized. The gaseous feed reacts and condenses on the solid product. Particles continue to grow; large resin particles sink to the bottom of the reactor and are continuously withdrawn. The heat of polymerization is removed by immersed heat transfer surfaces. Per pass conversion is only ~2%. The unconverted gaseous feed exiting the top of the reactor is compressed, cooled, and recycled. In the condensing mode of operation, the recycled gas mixture is cooled below its dew point so that a portion of the recycled olefins enters the reactor in the liquid phase. The liquid droplets rapidly evaporate, helping to remove the heat of polymerization. This strategy increases the rate of production from fluidized bed polymerization reactors.

Catalytic Reforming

Catalytic reforming is the key refinery process for converting low-octane paraffinic hydrocarbons in the gasoline range (i.e., naphtha) to high-octane naphthenes and aromatics. The noble metal catalysts and promoters (e.g., platinum and rhenium) deactivate because of coke deposition and must be regenerated. The first generation catalytic reforming technology developed in the 1950s employed fixed bed reactors operated at pressures of 30 bar or more. The rate of coke formation is reduced at high pressure, facilitating longer run times before the reactors are shut down for

regeneration (after ~6 months). However, reformate yield and octane number are adversely affected by high pressure. Continuous catalytic reforming (CCR) technology, featuring four reactors in series, has been developed to overcome this problem. The reforming catalyst forms a slowly moving bed that descends through each reactor. In the UOP stacked design, the catalyst is transferred between reactors because of gravity. In the IFP design, the reactors are side by side, and the catalyst leaving the bottom of one reactor enters a lift-pot from which it is transported to the top of the next reactor. From the lift-pot at the base of the fourth reactor, the deactivated catalyst is sent back to the regenerator, located upstream of the first reactor. Modern CCRs operate at ~5 bar, substantially increasing the reformate yield and octane number. The CCR process is similar to the FCC process in that the catalyst is continuously circulated. However, the solid circulation flux is much lower in the CCR process. Furthermore, the regenerators of FCC units provide combustion environments to remove coke from the catalyst. Combustion also occurs in the CCR regenerator, but oxygenation and chlorination take place to restore the acidity of the bifunctional catalyst.

GAS–SOLID REACTIONS

This section concerns noncatalytic gas–solid reactions:

$$\mathrm{A\,(gas)} + \mathrm{B\,(solid)} \rightarrow \mathrm{products\,(gas\ and/or\ solid)} \quad (2)$$

Table 1 describes the nature of gases and solids depending on the reaction. These reactions include combustion, gasification, pyrolysis, calcination, and roasting. Some specific information is also provided for these types of reaction. Many noncatalytic gas–solid reactions use coal, biomass, and municipal solid wastes as feedstock. Table 2 summarizes their chemical compositions. These compositions differ in many ways, including the organic/inorganic content and chemical and physical properties.[19] The most important noncatalytic gas–solid reactions using fluidized bed technology are discussed here.

Combustion

Fluidized bed steam generators burning solid fuels have operated in the energy industry since 1980. This technology penetrated the energy market surprisingly quickly. Outstanding features of fluidized bed steam generators are high fuel flexibility, integrated emission control, flexibility in meeting the operating requirements of various industries, and a proven record of reliability. Fuel flexibility has become an important criterion in selecting boiler technology for new utility power plants: the type and the grade of fuels also affect the solid handling provisions, energy recovery, emissions control, and solid waste handling.[20] For example, the 200 MW_e Tonghae thermal power plant CFB boiler (2 units) began commercial operation in 1998–1999 by firing low-quality Korean anthracite for electrical power generation.

Although combustion efficiency is a function of many variables (the most important being combustion

Table 1 Main characteristics of fluidized bed combustion, gasification, pyrolysis, calcination, and roasting

Reaction type	A	B	Main products	Other products	Temperature range (°C)	Solid residence time (sec)
Combustion	Air	Coal, MSW, or biomass	CO_2, H_2O, ash	SO_x, NO_x	850–930	3–500
Gasification	Air, oxygen, or steam	Biomass, MSW, or coal	CO, H_2, CH_4, char	Tar, SO_x, NO_x	650–850	1–600
Pyrolysis or fast pyrolysis	Recycle gases with or without air	Biomass, hydrocarbons	Liquid fuel (bio-oil), H_2O, char, coke	Lower molecular weight hydrocarbons	450–550	0.5–500
Calcination	Air	Limestone, dolomite, alumina, phosphates, etc.	Oxides	SO_x, NO_x	600–800	1–60
Roasting	Air or oxygen-enriched air	Sulfide ores	Oxides	SO_2	600–1000	30–120

Table 2 Physical, chemical, and fuel properties of biomass, coal, and municipal solid waste

Property	Biomass	Coal	MSW
Fuel density (kg/m^3)	~500–800	~1300–1500	~700–800
Mean particle size (mm)	~3	~0.1–3	~3–5
C content (wt.% of dry fuel)	42–54	65–85	25–30
O content (wt.% of dry fuel)	35–45	2–15	20–25
S content (wt.% of dry fuel)	Max. 0.5	0.5–7.5	Max. 0.25
SiO_2 content (wt.% of ash)	23–49	40–60	10–40
K_2O content (wt.% of ash)	4–48	2–6	1.5–9
Al_2O_3 content (wt.% of ash)	2.4–9.5	15–25	2–10
Fe_2O_3 content (wt.% of ash)	1.5–8.5	8–18	2–6
Ignition temperature (°C)	418–426	490–595	400–450
Friability	Low	High	Medium
Dry heating value (MJ/kg)	14–21	23–28	10–13[a]

[a]Dry heating value could be as low as 3 MJ/kg for developing countries.

temperature, excess air, residence time, feed size, cyclone separation efficiency, and mixing of gas and solids), carbon burnout efficiency is high (usually 99% or more) in fluidized bed combustion (FBC) systems. Many comprehensive FBC simulation models are available in the literature.[21] One of them has the advantage of being integrated with Aspen Plus, an advanced computer based simulation package for process engineering that is able to simulate a variety of processes ranging from single unit operations to complex multiunit processes.[21]

Several clean coal municipal waste or biomass technologies [bubbling fluidized bed (BFB) and circulating fluidized bed (CFB)] at atmospheric or at higher pressure, typically 6–15 bar) have been developed that reduce pollutant emissions.[22] Limestone is employed as a sorbent to capture sulfur in the bed, where it undergoes calcination followed by sulfation. The Ca to S molar feed ratio must typically be 2 to 3 for 90–95% retention of sulfur (SO_2 emission of ~200 mg/MJ). Unfortunately, the limestone utilization is normally low, typically 30–40%. NO_x emissions are low for FBC boilers relative to those of conventional pulverized coal boilers because of lower operating temperatures (typically 830–880°C). Compared with BFB combustors (135–1800 mg/MJ), CFB combustors tend to have lower NO_x levels (135–180 mg/MJ) because of air staging. Due to the catalytic effect of SO_2 and limestone, N_2O emissions show a maximum for a bed temperature of ~800°C. Ultralow emissions of NO_x (45 mg/MJ) can be obtained by combining FBC, and selective noncatalytic reduction with ammonia or selective catalytic reduction (over Pt, Au or V_2O_5) is commonly used. However, achieving ultralow NO_x emissions may compromise sulfur oxide reduction and/or carbon utilization. Compared with BFB combustors, CFBs also produce less CO (~40 mg/MJ) because of the more intensive gas turbulence and better mixing, although lateral solid mixing remains an issue. Before discharge to the atmosphere, the flue gas is passed through a baghouse to remove most of the entrained dust. Particulate emissions from FB combustors are, on average, ~10 mg/MJ. Cofiring biomass and coal, compared with firing coal alone, helps reduce both NO_x and SO_2 and also fuel costs, minimize waste, and reduce soil and water pollution, depending upon the chemical composition of the biomass.[22]

Frequent operational problems have been encountered in biomass combustors. One common issue is agglomeration of bed material. Biomass ash is often relatively rich in alkali and alkaline-earth metals, causing it to melt at relatively low temperatures. Fouling of the heat exchangers may cause additional problems.

Several countries have introduced stringent emission limits (0.1 ng-TE/Nm^3) for chlorinated dioxins and furans emitted from combustion sources, in particular solid waste incinerators, because of concerns over their adverse health effects. Technologies for reducing their formation and emission in incineration processes have been studied extensively and can be applied in modern incineration plants.[23] Activated carbon injection and fabric filtration are currently practiced in many installations. However, to minimize capital cost, a more fundamental approach is needed to control and limit formation of these pollutants in incineration processes, e.g., involving the postcombustion zone, the combustion chamber, and waste feeding.[23]

Gasification

Gasification is the conversion of coal, municipal solid waste (MSW), or biomass to a gaseous fuel by heating in the presence of oxygen and/or steam. The resulting

fuel gas is more versatile than the original solid reactant. For example, it could be used to power gas engines and gas turbines, or as a chemical feedstock to produce liquid fuels. Air-blown processes produce low-calorific-value gases with a typical higher heating value (HHV) of 4–7 MJ/m^3, while oxygen- and steam-blown processes result in gases with an HHV of 10–18 MJ/m^3.[24] The reactions in the gasifier include partial oxidation and complete oxidation of the carbon content, steam–carbon reaction, Boudouard reaction, shift reaction, and methane formation.[25] A number of reactor configurations have been used including BFBs and CFBs.

Lurgi's atmospheric CFB technology, originally developed for alumina calcination, was later adapted to coal combustion. It has since been applied to biomass gasification. Recently, a consortium of Japanese utilities announced the construction of a commercial-scale 250 MW (1700 t/d) integrated gasification combined cycle (IGCC) plant at Nakoso, Japan, based on two stages. In the first (combustion) stage, the high temperature simplifies separation of liquid slag from the gas. In the second (reduction), only coal is introduced, in the absence of additional oxidant. The temperature drop over the reductor stage is 700°C, with a reactor outlet temperature of ~1000°C.[25] Many fluidized bed gasification processes have been and are being developed, but none incorporates a heat carrier in such a way that the tars are combusted and char reacts with the gasifying agent to produce synthesis or fuel gas. Such a system could produce pure syngas, with the gas free of tars and carbon conversion virtually complete.[26]

The major operational difficulty is gas cleaning, which is still the major bottleneck, limiting the use of biomass for electricity generation. The gas formed contains a number of impurities: particulates (ash and char), tar, and nitrogen, sulfur, and alkali compounds. The tar, consisting of high molecular weight compounds that typically condense at temperatures below 450°C, can cause blockages and corrosion, while also reducing overall efficiency. Moreover, impurities (such as methane) can affect the end usage of syngas.

As a first approximation, a nonstoichiometric equilibrium model based on direct minimization of Gibbs free energy may be applied to predict gasifier performance, although experimental data deviate somewhat from model predictions.[24] Extensive research is being carried out to develop a process to produce a low-tar or tar-free gas. To reach these goals, two strategies are applied: improvement of the gasification technologies (updraft gasifier and cocurrent moving bed gasifier with throat[26]) and development of tar-removal processes. Since the mid-1980s, interest has grown with respect to catalytic biomass gasification.[27] The catalysts may be divided into two groups depending on the position of the catalyst reactor relative to the gasifier. In the first, the catalyst is added directly to the biomass prior to gasification. These catalysts consist of cheap disposable materials, e.g., dolomite ($MgCO_3 \cdot CaCO_3$), calcite (CaO), or magnesite (MgO), which reduce tar content. The second group of catalysts is placed in a secondary reactor downstream from the gasifier. Commercially available and relatively inexpensive, nickel reforming catalyst is highly effective at removing hydrocarbons and adjusting the gas composition to syngas quality. It is most active and has a longer lifetime when operated at 780°C in a fluidized bed.[27]

Pyrolysis

Pyrolysis is a moderate temperature (~500°C) process in which biomass, coal, or MSW is decomposed in the absence of an oxidizing agent at very high heating rates and vapor residence times typically less than 1 sec.[28] The product is collected as a liquid (bio-oil) and a solid char after rapid cooling. For fast pyrolysis, BFBs and CFBs have been used extensively. They are designed and operated to maximize the liquid fraction at up to 75 wt.% on a dry feed basis. The bio-oil can be stored and transported. Its typical higher heating value is ~18 MJ/kg, and its chemical composition differs significantly from that of petroleum-derived oils. Several fast pyrolysis technologies have reached near-commercial status. Six CFB plants have been constructed by Ensyn Technologies, the largest with nominal capacities of 50 t/day. DynaMotive demonstrated a BFB process at 10 t/day of biomass and is scaling up the plant to 110 t/day.[28] Bio-oil has been successfully burned in boilers and furnaces, although some problems have been reported with char levels which can block filters and atomizers. Other problems have also been encountered—the alkaline nature of the ash in turbines, the high viscosity of the bio-oil product (40–100 cP at 40°C), water content with possible phase separation, and corrosion from its low pH (pH 2–3). Comparison of three routes from biomass to electricity—combustion, gasification, and fast pyrolysis—based on cost and performance indicates that fast pyrolysis is expensive up to 5 MW_e. However, competition will be fierce as electricity production costs vary only slightly among these options.[29]

Upgrading bio-oil (oxygen content typically 35–40 wt.%) to quality transport liquid fuel requires full deoxygenation, e.g., by hydrotreating or catalytic vapor cracking over zeolite catalysts. Neither of these processes appears to be economically viable. The hydrotreating process is realized at high pressure and high temperature with sulfided CoMo or NiMo catalysts supported on alumina.

Very few chemicals produced from the fractionation of the bio-oil are already available. The most promising

are glycoaldehyde as a meat-browning agent and levoglucosan, which has potential in the synthesis of antibiotics and flavor compounds. Some phenols were even found in coal pyrolysis, which may produce high value specialty chemicals. Specialty chemicals generally require further work to develop reliable low-cost separation procedures.

Fluid coking is a thermal cracking process developed by Exxon (now ExxonMobil) in which heavy residual hydrocarbons, heavy oils, or bitumen derived from oil sands are upgraded in fluidized bed reactors at high temperatures and moderate pressures. The hydrocarbon to be upgraded is fed as a liquid. The lower-molecular-weight product is withdrawn at the top, while heavy ''coke'' coats the particles. The coke particles are then extracted near the bottom, after being passed through a stripper in which steam removes liquid hydrocarbon from the surfaces of the particles. The coke particles are circulated to a combustion chamber where the exothermic heat of reaction heats the coke particles, some of which are then recirculated to the fluid coker to provide the heat needed by the endothermic thermal pyrolysis reactions.

Calcination

An early fluid bed calciner for limestone and dolomite, built by England Lime Company in 1949, was 4 m in diameter, 14 m tall, and operated at 1000°C. As the reactions are highly endothermic and both the gas and solid exit at high temperature, several attempts were made to save energy, including multistaging and combining a calciner with suspension preheating.[4]

In the late 1960s, Lurgi adopted the CFB for calcination of alumina trihydrate. Design studies have shown that CFBs can be applied for capacities as high as ~1000 t/day. A high degree of automation, economy of space, and a simplified process flowsheet are important design criteria.

Apatitic rock phosphates (a poor-quality phosphate) are commonly calcined in fluidized beds (with one, two, or even three stages) to remove impurities (carbonates, water, and organic matter) and to increase phosphorus concentration. The hydrocarbon content (up to 5%) of this phosphate provides most of the heat needed for calcination. Increasing the process temperature improves the upgrading of the ore, but decreases the agronomic effectiveness in direct application to soils. This operation has a maximum temperature of ~900°C.[30]

Roasting

Roasting is often used in the metallurgical industry to convert mineral sulfides into oxides or other compounds such as sulfates suitable for further treatment and recovery of metals. Oxides and sulfates are readily dissolved in leach solutions, while sulfides only dissolve with great difficulty. In these pyrometallurgical processes, sulfide ores as fine particles (100–500 μm) are normally roasted to oxides in fluidized beds at 600–1000°C. The residence time of the solids is around a minute, and short circuiting of solids must be avoided. Fluidized bed roasters are widely used, especially for zinc sulfide ores, which account for >80% of all zinc production. As the reactions are exothermic, the temperature is kept constant by direct (water addition) or indirect (immersed cooling tubes) heat exchangers. In the late 1950s, Dorr-Oliver designed the first fluidized bed roaster producing SO_2 from pyrite, zinc blend, and sulfide ores. The roaster was 5.5 m in diameter and 7.6 m high; it operated at 700°C and superficial gas velocities of 0.45–0.5 m/s.

Pyrometallurgical processes operating at high temperatures are usually associated with expensive emissions control, as dioxins, chloride compounds, and mercury can be generated. Competitive technologies include hydrometallurgical operations. In such processes, pressure reactors (autoclaves) are gaining acceptance for the leaching of ores and concentrates, and in the recovery of metals from leach solutions. To continue to be attractive, pyrometallurgy faces two major challenges: reducing environmental emissions and increasing the capacity of existing plants. Although many plants attempt to increase the roasting capacity by applying an oxygen-enriched roasting condition, optimization of the concentrate flow and the process hydrodynamics is an alternative way to improve capacity, versatility, and stability. Improvements in mass/heat transfer in conventional fluidized beds have led to numerous innovative reactor configurations. One example is the toroidal fluidized bed developed by Torftech for sulfide ore roasting. This roaster has a fixed annular gas distributor consisting of angled blades, which force the particles into a toroidal motion that enhances heat/mass transfer, even at low gas velocities.[31]

Other Processes

Ultrapure silicon and iron ore reduction

Silicon of exceptional purity for semiconductor and photovoltaic industries is obtained from metallurgical-grade silicon or liquid silicon tetrachloride. At least four of six steps involve fluidized beds. In Osaka Titanium's process, step 1 (obtaining $SiHCl_3$ from Si and HCl) takes place at 300°C, whereas in the Union Carbide process, $SiHCl_3$ is obtained by hydrogenating gaseous $SiCl_4$ over $CuCl_2$ catalyst at 500°C. Other

noncatalytic gas–solid hydrogenation reactions are also carried out in fluidized beds. For example, Bethelem Steel developed the H-Iron process. Alan Wood Steel, Kawasaki Iron and Steel, and others have developed processes that produce iron by reducing fine iron oxides in a single stage at ~700°C. Other processes for iron ore reduction with hydrogen include the Stelling process to form cementite and the Armco or Exxon process with multistage reactors.[4]

Chlorination and fluorination of metal oxides

These reactions are in some cases the only way to obtain highly purified metals such as titanium and U-235. The chloride process, in which $TiCl_4$ is produced from titaniferous feedstocks, chlorinated in the presence of solid carbon at 950–1150°C in a BFB, is of significant importance in producing titanium metal and TiO_2 pigments. A recent techno-economic study identified the potential of a CFB reactor for the chlorination of fine slag, which can be obtained at lower cost than standard chloride-grade slag.[32]

Nano- and ultrafine particles

Several fluid bed processes are under development for production and encapsulation of nanoparticles, for example, WC-Co composites, bioceramics (i.e., calcium phosphate hydroxyapapite), carbon encapsulation of iron magnetic nanoparticles, and carbon nanotubes. These nano- or ultrafine powders have broad industrial and pharmaceutical applications. Production processes usually include solution preparation (sol–gel), drying, calcination, and sintering. The last three steps may be realized in a fluidized bed, but fluidization of nano- and ultrafine powders is very difficult because of strong interparticle forces.

LIQUID–SOLID REACTIONS IN FLUIDIZED BEDS

Liquid-fluidized beds predate gas-fluidized beds, but they have considerably fewer applications because of a smaller number of advantages. Most applications are physical,[33] with bioreactors being the sole significant reactor application. Much of the recent attention has focused on aerobic wastewater treatment and fermentation processes, e.g., with methane as the organic substrate (see Refs.[34,35] for more details). In these processes, microbial cells are attached to the surface of inert particles (e.g., sand or activated carbon) as a biofilm, or trapped within the pores or interior of particles, causing the particle size and/or density to vary with time. Loaded particles therefore have different fluidization characteristics, and this difference can be used to separate product from fresh particles to recover the biomass.

Advantages of liquid fluidization include lack of plugging, reduced pressure drops, good heat and mass transfer, favorable surface area, and ability to readily separate excess biomass. The fluidization may be upward, or downward (referred to as ''inverse fluidization'') when the particles have a lower density than the liquid. Tapering may be used to reduce entrainment. Various configurations, including circulating beds,[36] are possible.

GAS–LIQUID–SOLID (THREE-PHASE) CHEMICAL PROCESSES IN FLUIDIZED BEDS

In gas–liquid–solid (three-phase) fluidized beds, solid particles are simultaneously contacted with both gas and liquid. The gas and liquid may flow cocurrently upward, or the liquid may descend, while the gas rises. The liquid usually forms the continuous phase in which the solid particles and gas bubbles are dispersed. The bubbles are larger when the particles are smaller, and bed contraction can occur when gas is introduced into a liquid-fluidized bed of fine particles. Higher pressures lead to smaller bubbles and increased gas hold-ups.

Advantages of three-phase fluidized beds over trickle beds and other fixed bed systems are temperature uniformity, high heat transfer, ability to add and remove catalyst particles continuously, and limited mass transfer resistances (both external to the particles and bubbles, because of turbulence and limited bubble size, and inside the particles owing to relatively small particle diameters). Disadvantages include substantial axial dispersion (of gas, liquid, and particles), causing substantial deviations from plug flow, and lack of predictability because of the complex hydrodynamics. There are two major applications of gas–liquid–solid-fluidized beds: biochemical processes and hydrocarbon processing.

Biochemical Processes

Three-phase fluidized beds can be used as bioreactors for aerobic biochemical processes, including both fermentation processes and wastewater treatment. The gas phase is air, required for biological growth, while the solid particles provide immobilized surfaces on or in which cell growth can occur. The aqueous liquid phase provides the culture medium needed for the growth and maintenance of the cells. Air may be introduced separately from the liquid, or be premixed with the aqueous medium. The liquid medium may exhibit non-Newtonian rheology. A disadvantage of three-phase

fluidization for some processes, e.g., those involving mammalian cells, is that fluidization results in significant shear stresses, which may disrupt or damage the cells. Extensive background information, literature references, and a summary of recent advances in understanding three-phase hydrodynamics are provided by Fan[34,37] and Wright and Raper.[38] Various geometries are possible, for example, tapered columns, draft tubes, and external circulation.[34]

Wastewater treatment is relatively simple in that the sole purpose is the degradation of all the organic species present in the liquid to remove both biological oxygen demand (BOD) and chemical oxygen demand (COD). Three-phase fluidization is also of interest for bioremediation of contaminated soils. Production of alcohols by fermentation (e.g., ethanol from glucose) has been practiced commercially. Other fermentation processes have been examined for production of enzymes, acetic acid, stem cells, monoclonal antibodies, antibiotics, and other pharmaceutical products.

Hydrocarbon Processing

A number of energy-related processes have been developed employing three-phase fluidization. Early examples were in coal liquefaction and hydrogenation. In the 1960s and 1970s, there was significant activity to develop three-phase fluidization processes, including the H-Oil, H-Coal, and LC-Fining processes. The latter remains in commercial operation, and practical aspects have been described in the literature.[34,39] The reactor in this case is called an ''ebullated bed.'' Hydrogen gas is bubbled through a heavy hydrocarbon liquid (e.g., a residue, heavy oil, or bitumen) at elevated temperature ($>400^{\circ}C$) and pressure ($>10\,MPa$) in the presence of solid catalyst particles. Complex reactions including thermal cracking, hydrocracking, and hydrodesulfurization occur, leading to an upgraded product with a higher H/C ratio. Fresh catalyst is added periodically to maintain the overall activity level. Efforts to resolve practical problems in a large industrial unit have been described by McKnight et al.[39] A more recent development is the liquid-phase methanol synthesis process where syngas is reacted to produce methanol in a catalyst-inert oil slurry.

FLUIDIZED BED REACTOR MODELING

Considerable effort has been expended to devise models for fluidized bed reactors. These models differ greatly in complexity. It is important to adopt a model with the optimum degree of sophistication, complex enough to capture the mechanistic elements affecting the principal aspects of concern, but not so complex that they contain features whose inclusion is not justified in view of the uses to which the model will be put.

Models that are entirely empirical are seldom of much value for fluidized bed reactors, given the hydrodynamic complexity and the large number of variables. The simplest mechanistic models treat the reactor as simple single-phase vessels (well-mixed or plug flow) and have little relation to reality. Two-phase models were developed beginning in the 1950s and 1960s, with bubbling beds primarily in mind. In such models, reviewed in a number of studies (e.g., in Refs.[1,4,40]), the bed is divided into two parallel one-dimensional paths, one of high voidage (e.g., to represent the bubble phase) and the other of much lower voidage (representing the dense phase). Mole balances are then written for each phase, incorporating interphase mass transfer between the phases, as well as terms to account for the reaction(s). The predicted performance is then dependent on the interphase transfer, as well as the fraction of solids, total volume of each phase, gas flow assigned to each phase, and chemical kinetics. Such models are widely used and have been quite successful for steady state fluidized bed gas-phase solid-catalyzed reactions.

The two-phase models have been extended to gas–solid reactions (e.g., in Refs.[1,4,40]), requiring that the mole balances also account for reaction of the solid phase and that the stoichiometry of the reaction be satisfied. Population balances may then be added to account for changes in particle size distribution. Single-particle reaction models (e.g., shrinking core model) may also be needed to account for changes in particle properties resulting from the reaction(s). Two-phase models have also been extended to other flow regimes—slugging beds (e.g., in Ref.[41]), turbulent beds (e.g., in Ref.[42]), and fast fluidization.[5] They also form the basis of a more generic probabilistic approach,[43] which covers the bubbling, turbulent, and fast-fluidization flow regimes.

In recent years, there has been considerable effort to develop computational fluid dynamic (CFD) models to predict the hydrodynamics and performance of fluidized beds. While this approach will no doubt yield valuable tools in the future, CFD models are not yet at the point where they can be used with confidence for design and scale-up of fluidized bed processes.

CONCLUSIONS

Few fluidized bed reactors in practice operate in the bubbling regime.[18] The trend is toward higher gas velocities. Numerous potential applications of circulating fluidized bed technology as a catalytic reactor have been listed.[44] Improved understanding of the flow inside fluidized beds is reducing the risk associated with scale-up and leading to new processes.

A promising area involves integration of product separation with reaction by using fluidized bed membrane reactors, where semipermeable membranes, immersed in the bed, allow selective withdrawal of products like hydrogen, or distributed introduction of reactants like oxygen. Continuous product removal can overcome reaction equilibrium limitations of industrially important reactions such as steam methane reforming.[45] Controlled addition of oxygen through membranes can also enhance selectivity to desired products and reduce the risk of explosions in catalytic partial oxidation reactions.[46]

Other areas for future development captured in this entry are nanoparticle processes, the development and validation of CFD, and other advanced reactor models for fluidized beds. In addition, incremental improvements in fluidized bed processes will be made, helping to ensure that fluidized beds continue to gain ground as reactors in many industries.

ARTICLE OF FURTHER INTEREST

Fluidization, p. 997.

REFERENCES

1. Yates, J.G. *Fundamentals of Fluidized-Bed Chemical Processes*; Butterworths: London, 1983.
2. Davidson, J.F.; Clift, R.; Harrison, D. *Fluidization*, 2nd Ed.; Academic Press: London, 1985.
3. Geldart, D., Ed.; *Gas Fluidization Technology*; Wiley and Sons: Chichester, U.K., 1986.
4. Kunii, D.; Levenspiel, O. *Fluidization Engineering*; 2nd Ed.; Butterworth-Heinemann: Boston, 1991.
5. Grace, J.R.; Avidan, A.A.; Knowlton, T.M. *Circulating Fluidized Beds*; Chapman and Hall: London, 1997.
6. Yang, W.C., Ed.; *Handbook of Fluidization and Fluid-Particle Systems*; Marcel Dekker: New York, 2003.
7. Geldart, D. Types of gas fluidization. Powder Technol. **1973**, *7*, 285–292.
8. Yates, J.G. Effect of temperature and pressure. In *Handbook of Fluidization and Fluid-Particle Systems*; Yang, W.C., Ed.; Marcel Dekker: New York, 2003; 129–154.
9. Karri, S.B.R.; Werther, J. Gas distributor and plenum design in fluidized beds. In *Handbook of Fluidization and Fluid-Particle Systems*; Yang, W.C., Ed.; Marcel Dekker: New York, 2003; 155–170.
10. Knowlton, T.M. Standpipes and non-mechanical valves. In *Handbook of Fluidization and Fluid-Particle Systems*, Yang, W.C., Ed.; Marcel Dekker: New York, 2003; 599–617.
11. Avidan, A.A. Fluid catalytic cracking. In *Circulating Fluidized Beds*; Grace, J.R., Avidan, A.A., Knowlton, T.M., Eds.; Chapman and Hall: London, U.K., 1997; 466–488.
12. Matsen, J.M. Some characteristics of large solids circulation systems. In *Fluidization Technology*; Keairns, D.L., Ed.; Hemisphere: Washington, 1976; Vol. 2, 135–149.
13. Fligner, M.; Schipper, P.H.; Sapre, A.V.; Krambeck, F.J. Two-phase cluster model in riser reactors impact of radial density distribution on yields. Chem. Eng. Sci. **1994**, *49*, 5813–5818.
14. Derouin, C.; Nevicato, D.; Forissier, M.; Wild, G.; Bernard, J. Hydrodynamics of riser units and their impact on FCC operation. Ind. Eng. Chem. Res. **1997**, *36*, 4504–4515.
15. Karri, S.B.R.; Knowlton, T.M. A comparison of annulus solids flow direction and radial solids mass flux profiles at low and high mass fluxes in a riser. In *Circulating Fluidized Bed Technology VI*; Werther, J., Ed.; Dechema: Frankfurt, Germany, 1999; 71–76.
16. Contractor, R.M. DuPont's CFB technology for maleic anhydride. Chem. Eng. Sci. **1999**, *54*, 5627–5632.
17. Duvenhage, D.J.; Shingles, T. Synthol reactor technology development. Catal. Today **2002**, *71*, 301–305.
18. Bolthrunis, C.O.; Silverman, R.W.; Ferrari, D.C. Rocky road to commercialization: breakthroughs and challenges in the commercialization of fluidized bed reactors. In *Fluidization XI: Present and Future for Fluidization Engineering*; Arena, U., Chirone, R., Miccio, M., Salatino, P., Eds.; Engineering Foundation: New York, 2004; 547–554.
19. Demirbas, A. Combustion characteristics of different biomass fuels. Prog. Energy Combustion Sci. **2004**, *30*, 219–230.
20. Sondreal, E.A.; Benson, S.A.; Hurley, J.P.; Mann, M.D.; Pavlish, J.H.; Swanson, M.L.; Weber, G.F.; Zygarlicke, C.J. Review of advances in combustion technology and biomass cofiring. Fuel Proces. Technol. **2001**, *71*, 7–38.
21. Sotudeh-Gharebaagh, R.; Legros, R.; Chaouki, J.; Paris, J. Simulation of circulating fluidized bed reactors using ASPEN PLUS. Fuel **1998**, *77* (4), 327–337.
22. Leckner, B.; Lyngfelt, A. Optimization of emissions from fluidized bed combustion of coal, biofuel and waste. Int. J. Energy Res. **2002**, *26*, 1191–1202.

23. Buekens, A.; Huang, H. Comparative evaluation of techniques for controlling the formation and emission of chlorinated dioxins/furans in municipal waste incineration. J. Hazardous Mater. **1998**, *62*, 1–33.
24. Li, X.T.; Grace, J.R.; Lim, C.J.; Watkinson, A.P.; Chen, H.P.; Kim, J.R. Biomass gasification in a circulating fluidized bed. Biomass Bioenergy **2004**, *26*, 171–193.
25. McKendry, P. Energy production from biomass (part 3): gasification technologies. Bioresource Technol. **2002**, *83*, 55–63.
26. Higman, C.; Van der Burgt, M. *Gasification*; Gulf Professional Publishing: Amsterdam, 2003.
27. Sutton, D.; Kelleher, B.; Ross, J.R.H. Review of literature on catalysts for biomass gasification. Fuel Proces. Technol. **2001**, *73*, 155–173.
28. Czernik, S.; Bridgwater, A.V. Overview of applications of biomass fast pyrolysis oil. Energy Fuel **2004**, *18*, 590–598.
29. Bridgwater, A.V.; Toft, A.J.; Brammer, J.G. A techno-economic comparison of power production by biomass fast pyrolysis with gasification and combustion. Renew. Sustain. Energy Rev. **2002**, *6*, 181–248.
30. Lim, H.H.; Gilkes, R.J. Beneficiation of apatite rock phosphates by calcination: effects on chemical properties and fertilizer effectiveness. Aust. J. Soil Res. **2001**, *39* (2), 397–402.
31. Shu, J.; Lakhmanan, V.I.; Convey, J. Sintering and ferrite formation during high temperature roasting of sulfide concentrates. Can. Metall. Quart. **1999**, *38* (4), 215–225.
32. Luckos, A.; den Hoed, P. The carbochlorination of titaniferous feedstocks in a fluidized bed. In *Fluidization XI*; Arena, U., Chirone, R., Salatino, P., Eds.; Engineering Conferences International: New York, 2004.
33. Epstein, N. Applications of liquid–solid fluidization. Int. J. Chem. Reactor Eng. **2003**, *1*, Review 1.
34. Fan, L.-S. *Gas–Liquid–Solid Fluidization Engineering*; Butterworths: Boston, 1989.
35. Grady, C.P.L.; Daigger, G.T.; Lim, H.C. *Biological Wastewater Treatment*; Marcel Dekker: New York, 1999.
36. Zhu, J.-X.; Zheng, Y.; Karamanev, D.G.; Bassi, A.S. (Gas–)liquid–solid circulating fluidized beds and their potential applications to bioreactor engineering. Can. J. Chem. Eng. **2000**, *78*, 82–94.
37. Fan, L.-S. Advances in gas–liquid–solid fluidization. In *Fluidization XI*; Arena, U., Chirone, R., Miccio, M., Salatino, P., Eds.; Engineering Conferences International: New York, 2004; 1–20.
38. Wright, P.C.; Raper, J.A. A review of some parameters involved ion fluidized bed bioreactors. Chem. Eng. Technol. **1996**, *19*, 50–64.
39. McKnight, C.A.; Hackman, L.P.; Grace, J.R.; Macchi, A.; Kiel, D.; Tyler, J. Fluid dynamic studies in support of an industrial fluidized bed hydroprocessor. Can. J. Chem. Eng. **2003**, *81*, 338–350.
40. Grace, J.R. Fluid beds as chemical reactors. In *Gas Fluidization Technology*; Geldart, D., Ed.; Wiley: Chichester, U.K., 1986; 285–339.
41. Hovmand, S.; Freedman, W.; Davidson, J.F. Chemical reaction in a pilot scale fluidized bed. Trans. Inst. Chem. Eng. **1971**, *49*, 149–162.
42. Chaouki, J.; Gonzalez, A.; Guy, C.; Klvana, D. Two-phase model for a catalytic turbulent fluidized bed reactor: application to ethylene synthesis. Chem. Eng. Sci. **1999**, *54*, 2039–2045.
43. Abba, I.A.; Grace, J.R.; Bi, H.T.; Thompson, M.L. Spanning the flow regimes: generic fluidized-bed reactor model. AIChE J. **2003**, *49*, 1838–1848.
44. Berruti, F.; Chaouki, J.; Godfroy, L.; Pugsley, T.S.; Patience, G.S. Hydrodynamics of circulating fluidized bed risers: a review. Can. J. Chem. Eng. **1995**, *73*, 579–602.
45. Adris, A.M.; Grace, J.R. Characteristics of fluidized-bed membrane reactors: scale-up and practical issues. Ind. Eng. Chem. Res. **1997**, *36*, 4549–4556.
46. Ramos, R.; Pina, M.P.; Menendez, M.; Santamaria, J.; Patience, G.S. Oxidative dehydrogenation of propane to propene: simulation of a commercial inert membrane reactor immersed in a fluidized bed. Can. J. Chem. Eng. **2002**, *79*, 902–912.

Fluorescent Coatings for High Temperature Phosphor Thermometry

S. W. Allison
Engineering Science and Technology Division, Oak Ridge National Laboratory (ORNL), Oak Ridge, Tennessee, U.S.A.

W. A. Hollerman
Department of Physics, University of Louisiana at Lafayette, Lafayette, Louisiana, U.S.A.

S. M. Goedeke
M. R. Cates
Engineering Science and Technology Division, Oak Ridge National Laboratory (ORNL), Oak Ridge, Tennessee, U.S.A.

T. J. Bencic
Optical Instrumentation Technology Branch, NASA John H. Glenn Research Center at Lewis Field, Cleveland, Ohio, U.S.A.

INTRODUCTION

Fluorescence provides a good noncontact thermometry technique in hostile environments such as those found at high temperatures. Phosphors are typically rare earth-doped ceramics that emit light when excited. The intensity, rise time, decay time, and wavelength shift of this emitted light can be temperature dependent. When thermographic phosphors are applied to a surface, with an excitation source and a method to characterize the emission provided, it is possible to determine the surface temperature. One of the simpler methods to apply these coating is through the use of temperature sensitive paints (TSPs). These TSPs are created by mixing phosphor with a binder material to form a sprayable coating, which can be easily and economically applied to a large area. Ideally, these phosphor paints need to survive at the limit of the existing decay time data, or 1700°C. The survivability of phosphor paint depends on the physical characteristics of the binder. The goal of this research is to discover binders that will allow phosphor paints to survive at high temperatures. Suitable binders will allow the construction of noncontact measurement devices useful in environments that are not suited for more common thermocouple or infrared devices. For a phosphor paint to be useful at a selected temperature, it must fluoresce when excited and have a measurable decay time.

BACKGROUND

Phosphors are fine powders that are doped with trace elements that give off visible light when suitably excited. Many of these phosphors have a ceramic base and can survive and function at high temperatures such as those present during combustion. When phosphor is applied as a thin coating, it quickly equilibrates to the ambient environment and can be used to measure the surface temperature.

The basic physics of thermographic phosphors is well established, and researchers at Oak Ridge National Laboratory (ORNL) have demonstrated several useful applications.[1–14] The thermometry method relies on measuring the rate of decay of the fluorescence yield as a function of temperature. Having calibrated the phosphor over the desired temperature range, a small surface deposit is excited with a pulsed light source. The resulting fluorescent decay (typically in less than 1 ms) time is measured to calculate the temperature of the substrate. In many instances, (e.g., in a continuous steel galvanneal process) a simple puff of powder onto the surface provides an adequate fluorescent signal.[1,2]

Often temperature measurements are made using thermocouples or optical pyrometry. However, in situations where rapid motion or reciprocating equipment is present at high temperatures, it is best to use other techniques. For many phosphors, the prompt fluorescence decay time (τ) varies as a function of temperature and is defined by:

$$I = I_0 \exp\left(-\frac{t}{\tau}\right) \quad (1)$$

where I is the fluorescence light intensity (arbitrary units), I_0 the initial fluorescence light intensity

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008045

(arbitrary units), t the time since cessation of excitation source (s), and τ the prompt fluorescence decay time (s).[1,2]

The time needed to reduce the light intensity to e^{-1} (36.8%) of its original value is defined as the prompt fluorescence decay time. An example of this quantity for several thermographic phosphors is shown in Fig. 1. Notice the fluorescence decay time decreases by four orders of magnitude when the temperature increases from 600 to 1100°C.

This entry gives an overview into research to find binder and phosphor combinations that can emit fluorescence and remain mechanically viable at temperatures as large as 1700°C. Emphasis will be placed on developing procedures and techniques for the application and the pretreatment of candidate binder and phosphor combinations.

HISTORY

In May 2001, a research program was initiated to evaluate temperature sensitive paints (TSPs) that could be used in high temperature thermometry applications. This three-year program was funded by the NASA John H. Glenn Research Center in Cleveland, Ohio. A research team lead by ORNL was assembled to locate binder and phosphor combinations that can emit fluorescence and maintain structural integrity at high temperatures. The ultimate goal for this research was to locate TSPs that could be used to measure the heat flux and temperature present inside a common turbine or rocket engine. The phosphors in question are typically rare earth compounds that emit copious

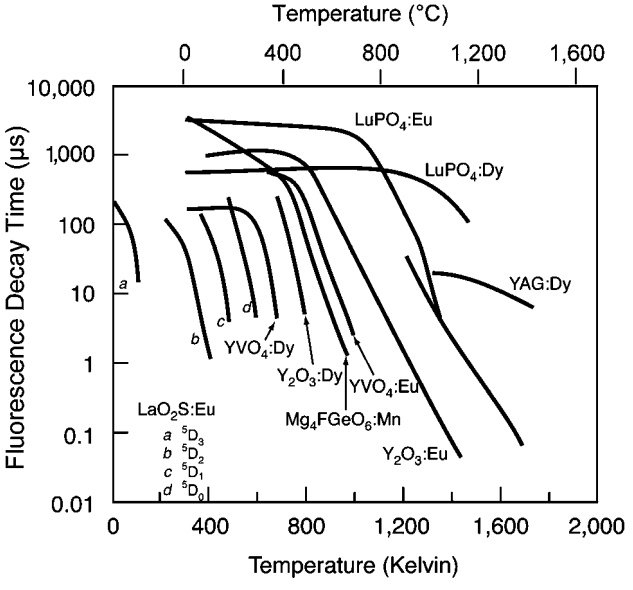


Fig. 1 Prompt fluorescence decay time for a selection of phosphors.

fluorescence and have grain sizes of less than 10 μm. Binders used for this purpose must: 1) be easy to apply with an airbrush; 2) set to a temperature resistant inorganic finish; and 3) have minimal reaction to the phosphor material.

During the first year of the program, research was completed to determine if heat flux could be measured using a phosphor coating. From the second to the fourth year, research was completed to find phosphor and binder combinations that can emit light and remain mechanically viable at high temperatures.

HEAT FLUX MEASUREMENTS

Measurement of prompt fluorescence decay time can be correlated with temperature to determine the amount of heat flux present in the sample environment. In the simplest sense, heat flux is proportional to the temperature difference on both sides of a homogeneous slab. By measuring these temperatures, it is a simple matter to determine the heat flux across any gradient. To measure the heat flux passing through a selected region, it will be necessary to simultaneously excite the phosphor on both surfaces of the material slab. The best excitation source for this task is the pulsed laser, since it emits a desired wavelength of coherent light with a large power density. Lasers should have sufficient power to excite the phosphor paint through the thickness of the gauge material.

This initial effort was designed to determine if fluorescence from two candidate phosphors could be detected through several single crystal and polycrystalline yttrium stabilized zirconia (YSZ) samples. YSZ samples in this research are divided into crystal and polycrystalline forms. The YSZ compound is a combination of approximately 90% zirconium oxide (ZrO_2) and 10% yttrium oxide (Y_2O_3). The single crystal YSZ samples have a density of 5.8 g/cm^3 with a lattice constant of 0.541 nm. The polycrystalline YSZ samples have a density of 6.6 g/cm^3 with thickness of 0.11, 0.16, and 0.20 mm. A thin layer of graphite was applied to one side of each polycrystalline YSZ sample to add strength. After application of the phosphor paint, each polycrystalline sample was heated to 800°C for 8 hr to remove the graphite. This heating cycle also cured the water-based phosphor paint. Detailed information on this research can be found in Refs.[6–8].

Yttrium aluminum garnet ($Y_3Al_5O_{12}$–YAG) doped with europium (Eu) and cerium (Ce) was used as the active phosphors for this research. The YAG : Eu phosphor glows bright orange when irradiated with a standard ultraviolet light. Emission peaks at wavelengths of 592, 610, 631, 697, and 710 nm are clearly visible in the YAG : Eu spectrum. Conversely, YAG : Ce glows bright green when irradiated with

ultraviolet radiation. The fluorescence emission is centered at 510 nm with a full width at half maximum of about 150 nm.

A mixture of 50% HPC® and 50% LK® (by volume) was used as a binder for each phosphor paint. Both the HPC and LK binders are manufactured by Zyp Coatings, Inc., Oak Ridge, Tennessee. When cured, HPC is not reactive and is composed of magnesium aluminum silicate. Engineers at Zyp indicate that HPC will withstand temperatures of about 1400°C. HPC is water based and has a shelf life of 12 months. The LK binder is not reactive and is composed 75% SiO_2, 20% K_2O, and 5% Li_2O. Engineers at Zyp indicate that LK will withstand temperatures of about 1100°C if additional fillers are added to the mix. LK is also water based and has a shelf life of 12 months. The selected HPC and LK paint is a good combination for this research. By itself, HPC is thick and difficult to spray. Conversely, LK by itself cannot be used safely above 1000°C. The selected combination is easy to spray and provides consistent results to temperatures in excess of 1000°C.

The sample matrix used in this research can be found in Table 1. Both the single crystal (set I) and polycrystalline (set II) samples were coated with a paint containing about 20% YAG : Eu by volume. The set III polycrystalline samples were coated with a paint containing approximately 20% YAG : Ce by volume. Fluor pigments and the HPC/LK binder was well mixed before application. Paint was applied to the each YSZ sample using a standard airbrush. Each sample was kept warm on a hotplate during the spraying process to help evaporate the water from the binder. Paint uniformity was checked using an ultraviolet lamp. The paint thickness for this research was estimated to be approximately 0.025 mm (1 mil). At the conclusion of the spraying process, samples were heated to 800°C to set the binder.

Opposite halves of each YSZ crystal piece were coated with YAG : Eu paint mixture. This procedure was accomplished to allow the YAG : Eu to be excited through the thickness of the YSZ. The number of edges visible on the painted phosphor surface can identify the polycrystalline samples. A simple cardboard and tape mask was used during the spraying process to make a straight edge on the YSZ surface. When illuminated with an ultraviolet lamp, both sets of samples clearly show two, one, and zero edges. These edges correspond to YSZ thicknesses of 0.20, 0.16, and 0.11 mm respectively. The surface of the painted YAG : Eu samples was bright white. However, the surface of the painted YAG : Ce samples was light green. The surface of the unpainted YSZ was light beige in color.

A PerkinElmer LS-50B spectrophotometer was used to determine if fluorescence emission could be imaged through two different thicknesses of YSZ. Painted crystal samples were imaged using an excitation light source with wavelengths of 355, 405, and 532 nm. Emission lines from YAG : Eu have the highest intensity for the 405 nm excitation source. All the YAG : Eu emission peaks are clearly visible using the 405 nm excitation source. The same YAG : Eu emission peaks are also faintly observed using the 355 nm excitation source. No peaks were observed using the 532 nm excitation source.

The goal for this analysis was to measure the prompt fluorescence decay time for each polycrystalline YSZ thickness. Results from this analysis are shown in Table 2. Visible fluorescence was observed through all three YSZ thicknesses at each of the tested laser wavelengths. It was not possible to measure the prompt fluorescence decay time for YAG : Eu because the reduction of light intensity did not follow a simple exponential curve. This phenomenon is most likely caused by electrical noise, stray light, or the absorption and re-emission of fluorescence from the YSZ.

The average prompt fluorescence decay time for YAG : Ce was measured to be 62.7 ± 2.9 ns. The accepted value for YAG : Ce of 65 ns falls within the uncertainty of these measurements.[15] It should be noted that the prompt fluorescence decay time for YAG : Ce was measured through several thicknesses of YSZ.

Table 1 YSZ and phosphor sample test matrix

Set	YSZ sample	Sample identifier	YSZ thickness (mm)	Phosphor	Phosphor ratio (vol.%)
I	Single crystal	One	1.00	YAG : Eu	20
		Two	0.25		
II	Polycrystalline	Zero	0.11	YAG : Eu	20
		One	0.16		
		Two	0.20		
III	Polycrystalline	Zero	0.11	YAG : Ce	20
		One	0.16		
		Two	0.20		

Table 2 Polycrystalline YSZ results

Phosphor paint	Excitation laser wavelength (nm)	YSZ thickness (mm)	Observed fluorescence (YSZ side)	Measured fluorescence decay time (τ)
YAG : Eu	405	0.11	Very faint	None
		0.16	Very faint	None
		0.20	Very faint	None
	532	0.20	Strong	Not valid
YAG : Ce	355	0.11	Faint	57.5 ± 0.2 ns
		0.16	Faint	None
		0.20	Faint	None
	532	0.16	Strong	63.1 ± 0.2 ns
		0.20	Strong	67.5 ± 0.3 ns

ZYP COATINGS SURVIVABILITY RESEARCH

Three binders, HPC, LK, and ZAP®, from Zyp Coatings, Inc., Oak Ridge, Tennessee, were used during the second year of the program. Each of these three binders was designed for use in high temperature applications and are available in the commercial market. Three other binders were tried during the initial stages of this research. None of these other binders were found to be viable above 1200°C and were dropped from future trials. Detailed information on this research can be found in Refs.[9–12].

When cured, the ZAP binder coating is composed of 53.5% P_2O_5 and 46.5% Al_2O_3. The maximum recommended use temperature for this formulation is 1800°C. It is water and alcohol based with a shelf life of 12 months. The sprayed coating dries at room temperature in 15–20 min after application. ZAP is a hard coating that can be used in any atmosphere. Heating to 900°C for 60 min cures the ZAP binder without organic ash.

A selection of rare earth compounds is used for the phosphor powders in this research: yttrium oxide doped with europium (Y_2O_3 : Eu), and YAG doped with dysprosium (YAG : Dy), thulium (YAG : Tm), and europium (YAG : Eu). The dopant concentrations are typically a few percent by mass. A selected binder and phosphor powder are mixed together to create a strong and durable paint. A maximum of 50% phosphor powder by volume was added to each mixture. Paint mixtures were sprayed on a clean 25 mm × 25 mm (1 in. × 1 in.) ceramic card and cured based on specific binder instructions. Paint uniformity was checked using black light inspection during application. All samples were inspected after the cure cycle before being exposed to the high temperature environment for 1 hr. After curing and before thermal cycling, a PerkinElmer LS-50B spectrophotometer was used to determine the baseline fluorescence spectrum for each paint sample. After each thermal cycle the paint sample was reanalyzed using the LS50-B.

Results from these measurements can be found in Table 3. Each paint combination was exposed to the high temperature listed in Table 3 for 1 hr and then slowly cooled. The heating rate was kept small to minimize effects because of the difference in expansion coefficient between the paint and the ceramic substrate.

Results shown in Table 3 show that most of the HPC/LK bound samples survived heating to 1400°C. Most of the paints were removed from the surface of the ceramic at 1500°C. These data also shows that paint consisting of a 100% ZAP binder and 30% YAG : Dy powder by volume is intact and emits fluorescence after heating to 1500°C. This paint surface did look bumpy or mottled after heating. The other ZAP paints were also intact, but with reduced fluorescence. In fact, the 100% ZAP and 30% by volume YAG : Dy paint was also found to emit fluorescence after heating to 1600°C.

COTRONICS COATINGS SURVIVABILITY RESEARCH

During the third and final year of the NASA-Glenn research, several Resbond® ceramic binders manufactured by Cotronics Corporation of Brooklyn, New York, were evaluated using a powdered Y_2O_3 : Eu phosphor. The main fluorescence emission from Y_2O_3 : Eu occurs at a wavelength of 611 nm. Selected material and preparation properties for the tested Resbond binders are given in Table 4. The Cotronics series of ceramic binders includes six different formulations. Of these, only the four varieties listed in Table 4 were tested. During initial research, Cotronics Resbond 795 did not adhere to the surface. Since the other two binders have similar

Table 3 Binder and phosphor paint survivability results

	Paint			Phosphor emission				
Binder (composition)	Fraction(vol.%)	Phosphor	Emission lines (nm)	1200°C	1300°C	1400°C	1500°C	Comments
100% HPC	20	Y_2O_3 : Eu	611			Yes	Yes	Paint mostly gone after heating to 1500°C
100% HPC	10	Y_2O_3 : Eu	611			Yes	Yes	Paint mostly gone after heating to 1500°C
75% HPC 25% LK	20	Y_2O_3 : Eu	611				No	All of paint gone after heating to 1500°C
75% HPC 25% LK	10	Y_2O_3 : Eu	611				Yes	Most of coating gone after heating to 1500°C
50% HPC 50% LK	20	Y_2O_3 : Eu	611			Yes	Yes	Paint mostly gone after heating to 1500°C
50% HPC 50% LK	10	Y_2O_3 : Eu	611			Yes	No	No 611 nm peak after heating to 1500°C
100% ZAP	50	Y_2O_3 : Eu	611	Yes	Yes	Yes	Yes	Paint still intact with diminished fluorescence after heating to 1500°C
100% ZAP	30	YAG : Dy	585	Yes	Yes	Yes	Yes	Paint still intact after heating to 1500°C
100% ZAP	30	YAG : Tm	420 480				Yes	Coating intact but poor fluorescence after heating to 1500°C
100% ZAP	30	YAG : Eu	595 611				Yes	Coating intact but poor fluorescence after heating to 1500°C

properties, it was considered more productive to focus on the three remaining binders. Detailed information concerning this research can be found in Refs.[13,14].

SAMPLE PREPARATION

To create the TSPs, phosphor powder is mixed with ceramic binder. The compositions for paint samples used in this research are shown in Table 5. Typically, the phosphor powder accounts for 20 vol.% of the paint mixture. In some cases, 10 vol.% of a second powder such as magnesium oxide (MgO_2) is added displacing the binder. This addition is an attempt to increase the thermal conductivity. The paint preparation procedure began by measuring the powdered solids in a graduated capped vial. Individual tubes were vibrated to insure proper volume measurement.

Table 4 Properties for selected Cotronics Resbond ceramic binders

Resbond composition	791 Silicate glass	792 Silicate glass	793 Silica oxide	795 Alumina oxide
Applications	Adhesives Coatings	Adhesives Coatings Electronics	Bonds Fibrous materials	High-purity binder
pH	Mildly basic	Mildly basic	Mildly basic	Mildly acidic
Maximum temperature (°C)	1650	1650	1760	1870
Density (g/cm^3)	1.40	1.20	1.42	1.40
Viscosity (cps)	1500	150	50	500
Cure time (hr)	24	24	24	2
Cycling instructions	16 hr at 25°C or 2 hr at 120°C	16 hr at 25°C or 2 hr at 120°C	2 hr at 25°C or 2 hr at 175°C	2 hr at 95°C and 2 hr at 175°C

Table 5 Tested paint sample compositions (fractions by volume)

Resbond binder		Phosphor			
Product	**Quantity (%)**	**Compound**	**Quantity (%)**	**Water quantity (%)**	**MgO_2 quantity (%)**
791	35	Y_2O_3 : Eu	20	35	10
	40			40	0
	70			0	10
	80			0	0
792	70	Y_2O_3 : Eu	20	0	10
	80			0	0
793	80	Y_2O_3 : Eu	10	0	10
			20	0	0

Liquids were then added in predetermined amounts. After all the components were added to the cylinder, they were mixed by vigorously by shaking the capped vial. The mixture was applied to the surface of a cleaned alumina substrate using a commercial air-brush. The TSPs were allowed to cure according to the manufacturer's specifications as shown in Table 6. In the case of Resbond 791, water was added to thin the mixture to make it easier to apply using the air-brush. However, the undiluted Resbond 791 paints were applied using a brush, since spraying was difficult and it did not give the desired uniform coating.

EXPERIMENTAL METHOD

To determine the survivability for a TSP, samples were thermally cycled in a Thermoline 46200 high-temperature furnace. The furnace controls were set to raise the temperature to some predetermined value. The sample was then allowed to remain at this high temperature for 1 hr or more, followed by cooling to ambient room conditions. To quantify the fluorescence efficiency of phosphor suspended in the TSP, a Perkin Elmer LS-50B spectrophotometer was used to measure the emission spectrum of the sample after each thermal cycle. For each TSP mixture, two samples were made. The first sample was heated through the curing cycle and used as a control. The second was thermally cycled by heating to a set temperature and allowing it to cool back to room temperature. After the sample cooled, the emission spectrum was measured. Fig. 2 shows the variation in the emission spectrum for a Resbond 793 and Y_2O_3 : Eu-based TSP after thermal cycling at several different temperatures. It is quite obvious that fluorescence intensity for the 611 nm emission line decreases as a function of temperature. In fact, the 611 nm fluorescence emission is reduced to the background level after thermal cycling at 1600°C. The TSP fluorescence spectrum shown in Fig. 2 contained 80% Resbond 793 and 20% Y_2O_3 : Eu by volume. After thermal cycling, TSP samples were analyzed under a ultraviolet light to qualitatively estimate the intensity of the remaining 611 nm fluorescence from Y_2O_3 : Eu.

QUALITATIVE RESULTS

Qualitative results for the Resbond binder and Y_2O_3 : Eu TSP samples are shown in Table 6. Notice that all of the tested TSP samples showed some fluorescence after thermal cycling to 1300°C. Fluorescence was also observed for all but one TSP sample at 1400°C. Only three TSPs made with Resbond 792 and 793 showed significant fluorescence at 1500°C. Finally, fluorescence was observed after exposure to 1600°C for one TSP containing 20% Y_2O_3 : Eu and 80% Resbond 793 by volume. These results are consistent with

Table 6 Qualitative results for the Resbond binder and Y_2O_3 : Eu TSP samples

Resbond binder		Remaining paint components			Phosphor emission at given cycling temperature				
Type	**Amount (%)**	**Y_2O_3 : Eu** *(%)*	**Water (%)**	**MgO_2 (%)**	**1200°C**	**1300°C**	**1400°C**	**1500°C**	**1600°C**
791	35	20	35	10	Yes	Yes	Yes	No	No
	40		40	0	Yes	Yes	Yes	No	No
	70		0	10	Yes	Yes	Yes	No	No
	80			0	Yes	Yes	Yes	No	No
792	70	20	0	10	Yes	Yes	Yes	Yes	No
	80			0	Yes	Yes	Yes	Yes	No
793	70	20	0	10	Yes	Yes	No	No	No
	80			0	Yes	Yes	Yes	Yes	Yes

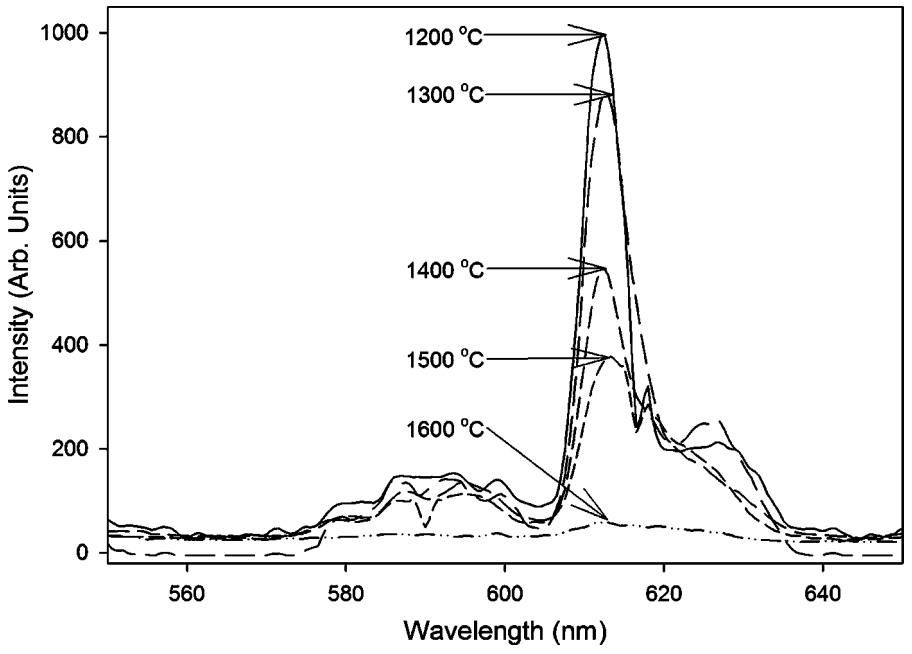


Fig. 2 Resbond 793 and Y_2O_3 : Eu TSP emission spectrum after thermal cycling.

earlier TSP research completed during the second year of the NASA-Glenn program.[9–12] As a group, it appears that TSPs made with Zyp binders were better able to withstand slightly higher temperatures compared to the ones made using any of the three tested Cotronics Resbond formulations. However, selected Zyp and Cotronics based TSPs were both able to survive to 1500°C.

QUANTITATIVE RESULTS

A quantitative determination of fluorescence intensity as a function of cycling temperature is more complex. It was decided to use a ratio of 0.2 (20%) of the maximum emission intensity as the criteria to determine the viability of fluorescence for a given TSP sample. If the fluorescence emission is small, it will be difficult to measure the decay time and obtain a corresponding surface temperature. There will come a point in intensity where a phosphor system cannot be used to measure temperature. The decision ratio of 0.2 was completely arbitrary and was based on the observation that the apparent fluorescence measurement uncertainty was about ±10% (intensity fraction of 0.1), which was two times the measured error for the 611 nm line for Y_2O_3 : Eu.

For each tested TSP, the spectral fluorescence intensity was measured using a PerkinElmer LS-50B spectrophotometer and an ultraviolet excitation source. In each case, the fluorescence was normalized to the Y_2O_3 : Eu maximum emission line at 611 nm. Samples were held at temperature for 1 hr, unless otherwise indicated.

Resbond 791

The evaluation of the Resbond 791 binder was complicated by its large viscosity. This required modifying the application procedure by either applying the paint with a standard bristle brush, increasing the supply pressure to the airbrush, or by thinning the binder. Since the other evaluated Resbond binders had smaller viscosities, it was felt that applying the thick paint with a brush and thinning of binder would provide a more equitable comparison. As such, the Resbond 791 binder was thinned in a one-to-one ratio with water, which aided in the application of the paint. Quantitative results for the Resbond 791 binder TSP are shown in Fig. 3.

Tests were performed using brushed-on paints. These coatings were thicker than the airbrushed coatings, but appeared to have the same maximum cycling temperature. The plot on the left side of Fig. 3 (labeled A) shows the normalized fluorescence emission (611 nm) for the brushed-on Resbond 791 paint vs. cycling temperature. The brushed-on TSP maintained a normalized peak emission greater than 0.2 to cycling temperatures up to 1600°C. The emission intensity decreased consistently from 1300 to 1500°C.

The plot on the right side of Fig. 3 (labeled B) shows the normalized fluorescence emission (611 nm) for the airbrushed (diluted) Resbond 791 paint vs. cycling temperature. The normalized emission spectra for the diluted sample was similar to that measured for the brushed-on TSP. It is interesting to note that the intensity of the brushed-on samples decreases slightly faster than the diluted TSPs, which could be partially caused

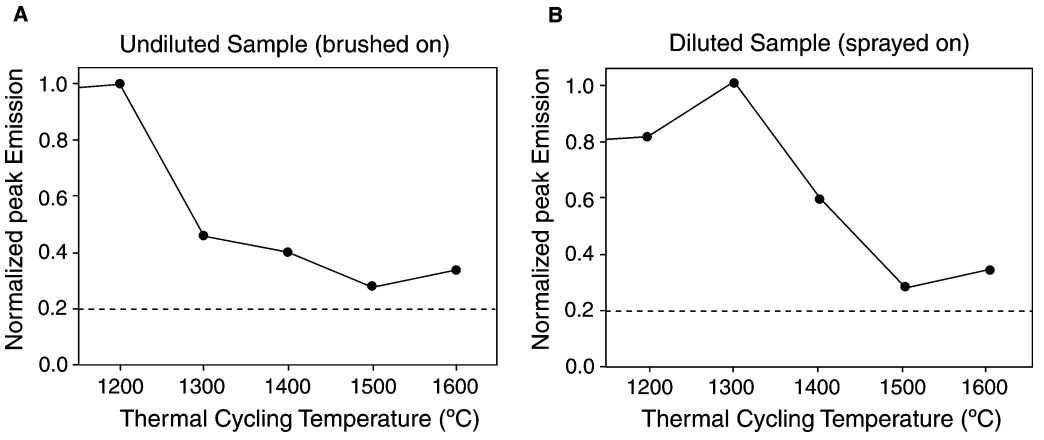


Fig. 3 Normalized emission for brushed-on and sprayed on Resbond 791 TSPs as a function of cycling temperature.

by differences in application method. The undiluted samples were brushed-on to the substrate surface, which typically does not provide even a coating. The uneven thickness could cause uneven heating and enhance flaking of the TSP.

Resbond 792

For all tested cases, the Resbond 792 TSPs falls below the 0.2 intensity criteria after the first thermal cycle (1100°C). Visual inspection of the samples in ultraviolet light indicates that there was a significant drop in emission intensity after 1100°C. These inspections also showed discernable fluorescence emission to 1300°C and limited fluorescence up to 1500°C. The dramatic drop in fluorescence intensity would tend to indicate that the Resbond 792 is not suitable for high-temperature TSP applications.

Resbond 793

Quantitative results for the Resbond 793 binder TSP are shown in Fig. 4. The Resbond 793 TSP maintained a normalized peak emission greater than 0.2 to cycling temperatures up to 1500°C. The corresponding 611 nm normalized peak emission decreased uniformly from 1200 to 1600°C. Visual inspection of the samples indicates that the coating remains relatively bright to 1400°C. At 1500°C, the coating is still bright, but starts to flake away from the surface. By 1600°C, almost all of the coating has flaked off of the surface, but the remaining portion was still discernable. Resbond 793

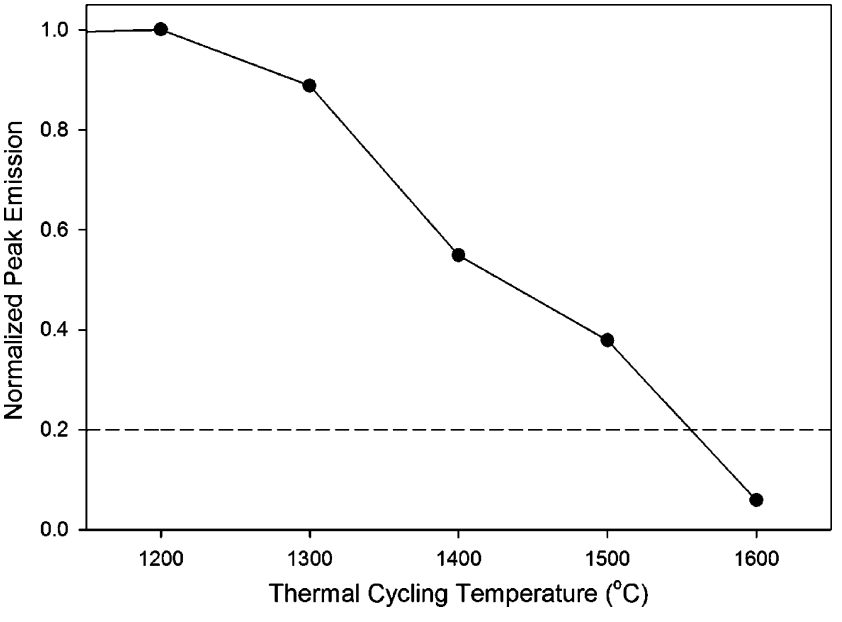


Fig. 4 Normalized emission for Resbond 793 TSP as a function of cycling temperature.

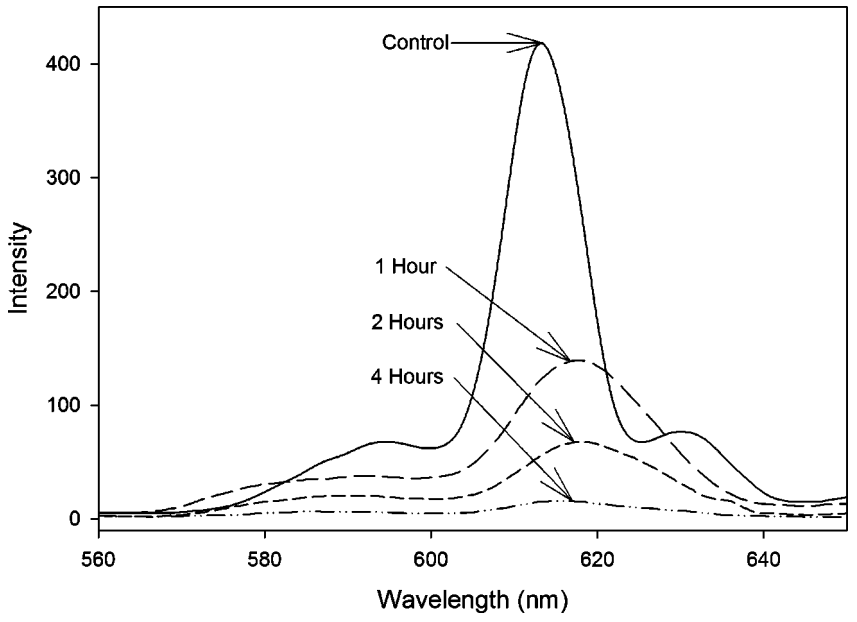


Fig. 5 Changes in fluorescence intensity for a Resbond 793 and Y_2O_3 : Eu TSP as a function of thermal exposure time.

has a low viscosity, which makes it easy to spray, and has the highest temperature survivability of any of the tested Cotronics binders. It appears to be the best of the tested materials for use in a TSP.

Temperature Exposure Time Dependence

The exposure time was originally selected because the anticipated test length was only a few seconds. It was felt that an hour of temperature exposure would well exceed the lifetime of any experimental test series. This exposure time was selected to make this work consistent with earlier research.[6–12] To date, no research had been completed comparing the effects of exposure time at temperature to the magnitude of fluorescence emission. It was decided to complete a series of tests with a single binder to determine the relationship between temperature exposure time and normalized emission intensity for a Y_2O_3 : Eu and Resbond 793 TSP.

To complete these tests, four new samples were prepared and cured as described in earlier sections. The first sample was held in reserve as the post cure control. The remaining samples were heated to 1400°C for 1, 2, and 4 hr. The 1400°C temperature was selected because nearly half of the room temperature fluorescence was emitted for this TSP in the previous test series. A separate sample was used in each duration test. The resulting emission spectra (ultraviolet excitation) for Y_2O_3 : Eu in a Resbond 793 binder with varied thermal cycling duration are shown in Fig. 5. After a 1-hr exposure, there is a noticeable decrease in fluorescence intensity. In fact, the decrease in intensity in the first hour of the test is surprising, as the previous results showed a 40% decrease, while this test had a nearly 70% decrease from the control. After 2 hr, there is further decrease in intensity to greater than 80% of the control, which indicates that the coating is marginal after 2 hr of exposure to 1400°C. After 4 hr, almost all the fluorescence is gone.

CONCLUSIONS

Recent research has shown that phosphor-based TSPs could be used for high temperature thermometry applications. Fluorescence from YAG : Eu and YAG : Ce could be detected through several thin YSZ samples. The average prompt fluorescence decay time for YAG : Ce was measured to be 62.7 $\pm$ 2.9 ns, which is close to the accepted value of 65 ns. This result can be used directly to develop an operational high temperature heat flux gauge.

Results also indicate that a 100% ZAP binder (Zyp Coatings) and 30% YAG : Dy (by volume) TSP will emit useful fluorescence to 1600°C. Other phosphors using the ZAP binder were found to be viable to 1500°C with reduced quantities of light emission. The Y_2O_3 : Eu phosphor and HPC/LK binder paint combinations were found to emit fluorescence to 1400°C. All of the tested Cotronics Resbond-based TSPs showed some fluorescence after thermal cycling to 1300°C. Fluorescence was also observed for all but one TSP sample at 1400°C. Only three TSPs made with Resbond 792 and 793 showed significant fluorescence at 1500°C. Fluorescence was observed at 1600°C for one TSP containing Resbond 793. In addition, Resbond 793 appears to have the best TSP characteristics of

the Cotronics binders, and it is one of the few binders that survives above 1500°C.

The temperature exposure time was found to be an important consideration for designing a TSP. After 2 hr of exposure at 1400°C, the fluorescence intensity decreases by 80%, which indicates the coating is of marginal use as a TSP. It appears that TSPs made with Zyp binders were better able to withstand slightly higher temperatures as compared to the ones made using any of the three tested Cotronics Resbond formulations. However, selected Zyp and cotronics based TSPs were both able to survive to 1500°C. Additional research is needed to further quantify these results.

REFERENCES

1. Allison, S.W.; Cates, M.R.; Goedeke, S.M.; Hollerman, W.A.; Womack, F.N.; Gillies, G.T. Remote thermometry with thermographic phosphors: instrumentation and applications, Chapter 4. In *Handbook of Luminescence, Display Materials, and Devices, Volume 2: Inorganic Display Materials*; Nalwa, H.S., Rohwer, L.S., Eds.; American Scientific Publishers, 2003; 187–250.
2. Allison, S.W.; Gillies, G.T. Remote thermometry with thermographic phosphors instrumentation and applications. Rev. Sci. Instrum. **1997**, *68* (7), 2615–2650.
3. Cates, M.R.; Allison, S.W.; Franks, L.A.; Borella, H.M.; Marshall, B.R.; Noel, B.W. Laser-induced fluorescence of europium-doped yttrium oxide for remote high-temperature thermometry. Proc. Laser Inst. Am. **1985**, *142*, 49–51.
4. Allison, S.W.; Boatner, L.A.; Gillies, L.A. Characterization of high-temperature thermographic phosphors: spectral properties of $LuPO_4$:Dy (1%) Eu(2%). Appl. Opt. **1995**, *34,* 5624.
5. Lopez, O.A.; McKittrick, J.; Shea, L.E. Fluorescence properties of polycrystalline Tm^{+++}-activated $Y_3Al_5O_{12}$ and Tm^{+++}–Li^{+} co-activated $Y_3Al_5O_{12}$ in the visible and near IR ranges. J. Luminesc. **1997**, *71*, 1–11.
6. Allison, S.W.; Beshears, D.L.; Bencic, T.; Hollerman, W.A.; Boudreaux, P. Development of temperature-sensitive paints for high temperature aeropropulsion applications. Proceedings of the American Institute of Aeronautics and Astronautics Propulsion Conference, 2001; AIAA-2001–3528, 2001.
7. Allison, S.W.; Beshears, D.L.; Gadfort, T.; Bencic, T.; Eldridge, J.; Hollerman, W.A.; Boudreaux, P. High temperature surface measurements using lifetime imaging of thermographic phosphors:bonding tests. Proceedings of the 19th International Congress on Instrumentation in Aerospace Simulation Facilities, August 27–30, 2001.
8. Hollerman, W.A.; Allison, S.W.; Beshears, D.L.; Guidry, R.F.; Bencic, T.J.; Mercer, C.R.; Eldridge, J.I.; Cates, M.R.; Boudreaux, P.; Goedeke, S.M. Development of fluorescent coatings for high temperature aerospace applications. Proceedings of the 2002 Core Technologies for Space Systems Conference*,* November 19–21, 2002; Colorado Springs, CO.
9. Hollerman, W.A.; Allison, S.W.; Beshears, D.L.; Guidry, R.F.; Bencic, T.J.; Mercer, C.R.; Eldridge, J.I.; Cates, M.R.; Boudreaux, P.; Goedeke, S.M. Development of inorganic fluorescent coatings for high temperature aerospace applications. Proceedings of the 49th International Instrumentation Symposium, Orlando, FL, May 5–9, 2003.
10. Hollerman, W.A.; Allison, S.W.; Goedeke, S.M.; Boudreaux, P.; Guidry, R.; Gates, E. Comparison of fluorescence properties for single crystal and polycrystalline YAG:Ce. IEEE Trans. Nuclear Sci. **2003**, *50* (4), 754–757.
11. Allison, S.W.; Goedeke, S.M.; Beshears, D.L.; Cates, M.R.; Hollerman, W.A.; Womack, F.N.; Bergeron, N.P.; Bencic, T.J.; Mercer, C.R.; Eldridge, J.I. Advances in high temperature phosphor thermometry for aerospace applications. Proceedings of the 39th AIAA/ASME/SAE/ASEE Joint Propulsion Conference, 2003; AIAA-2003–4584.
12. Hollerman, W.A.; Guidry, R.F.; Womack, F.N.; ; Bergeron, N.P.; Allison, S.W.; Goedeke, S.M.; Beshears, D.L.; Cates, M.R.; Bencic, T.J.; Mercer, C.R.; Eldridge, J.I. Use of phosphor coatings for high temperature aerospace applications. Proceedings of the 39th AIAA/ASME/SAE/ASEE Joint Propulsion Conference, 2003; AIAA-2003–4585.
13. Goedeke, S.M.; Hollerman, W.A.; Bergeron, N.P.; Allison, S.W.; Cates, M.R.; Bencic, T.J.; Mercer, C.R.; Eldridge, J.I. Study of Resbond® ceramic binders used for high temperature noncontact thermometry. Proceedings of the 106th Annual Meeting of the American Ceramic Society. Indianapolis, IN, 2004.
14. Bergeron, N.P.; Hollerman, W.A.; Goedeke, S.M.; Allison, S.W.; Cates, M.R.; Bencic, T.J.; Mercer, C.R.; Eldridge, J.I. Effect of adding MgO_2 to a selection of Resbond® thermally sensitive paints. Proceedings of the 50th Annual International Instrumentation Symposium, San Antonio, TX, May 9–13, 2004.
15. Shionoya, S.; Yen, W.M., Eds. *Phosphor Handbook*; CRC Press, 1998.

Fluoropolymers

Sina Ebnesajjad
DuPont Fluoroproducts, Chestnut Run Plaza, Wilmington, Delaware, U.S.A.

INTRODUCTION

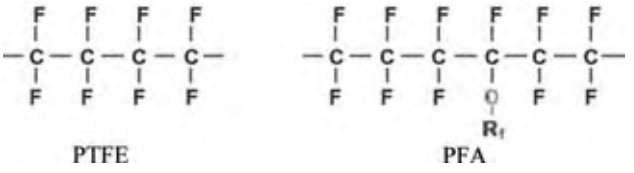
In this entry, fluoropolymer means a polymer that consists of partially or fully fluorinated olefinic monomers, such as vinylidene fluoride (CH_2=CF_2) and tetrafluoroethylene (CF_2=CF_2). Commercial fluoropolymers include homopolymers and copolymers. Homopolymers contain 99 wt.% or more one monomer and 1 wt.% or less of another monomer according to the convention by American Society for Testing Materials. Copolymers contain 1 wt.% or more of one or more comonomers. The major commercial fluoropolymers are based on tetrafluoroethylene, vinylidene fluoride, and to a lesser extent chlorotrifluoroethylene. Examples of comonomers include perfluoromethyl vinyl ether (PMVE), perfluoroethyl vinyl ether (PEVE), perfluoropropyl vinyl ether (PPVE), hexafluoropropylene (HFP), chlorotrifluoroethylene (CTFE), and perfluorobutyl ethylene (PFBE).

Fluoropolymers discussed include polytetrafluoroethylene (PTFE), perfluoroalkoxy polymer (PFA), fluorinated ethylene–propylene polymer (FEP), ethylene–tetrafluoroethylene copolymer (ETFE), ethylene–chlorotrifluoroethylene copolymer (ECTFE), polychlorotrifluoroethylene (PCTFE), polyvinylidene fluoride (PVDF), and polyvinyl fluoride (PVF).

In this entry, the classification, preparation, properties, fabrication, safety considerations, and economics of fluoropolymers are discussed. Monomer synthesis and properties have also been discussed. Increasing the fluorine content of a polymer increases chemical and solvent resistance, flame resistance, and photostability, improves electrical properties, such as dielectric constant, lowers coefficient of friction, raises melting point, increases thermal stability, and weakens mechanical properties.

FLUOROPOLYMER CLASSIFICATION

The era of fluoropolymers began with the serendipitous discovery of PTFE by Roy Plunkett of DuPont Company[1] while conducting research to find new refrigerants. A number of fluoroplastics have been developed since the discovery of PTFE. They are divided into two classes of perfluorinated and partially fluorinated polymers. Perfluorinated fluoropolymers are homopolymers and copolymers of tetrafluoroethylene (TFE). Some of the comonomers may contain a small amount of elements other than C or F. For example, PFA is a copolymer of TFE and perfluoroalkyl vinyl ether that contains oxygen. R_f is a perfluoroalkyl group of C_1 to C_4.



Partially fluorinated fluoropolymers contain hydrogen (H) or other atoms such as chlorine, in addition to fluorine and carbon. The most significant are homopolymers and copolymers of vinylidene fluoride (VDF). There are also copolymers and homopolymers of CTFE, although some have elastomeric properties. Other significant fluoroplastics include ETFE and PVF.

POLYMER DEVELOPMENT

Because of its high viscosity (10^{10}–10^{12} poise at 380°C), PTFE cannot be fabricated by melt-processing techniques. Melt-processible fluoropolymers have been developed by copolymerization of TFE, and FEP, a copolymer of TFE and HFP, has a lower maximum continuous use temperature than PTFE (200°C vs. 260°C) because of the deterioration of mechanical properties. Whereas, PFA, a copolymer of TFE with PPVE or PEVE, offers thermal stability, melt-processibility, and a maximum continuous use temperature of 260°C. Both FEP and PFA are considered perfluoropolymers.

Copolymers of ethylene with tetrafluoroethylene (ETFE) and chlorotrifluoroethylene (ECTFE) are mechanically stronger than perfluoropolymers, with some reduction in their chemical resistance and continuous use temperature and an increase in the coefficient of friction.

Amorphous copolymers of TFE are soluble in special halogenated solvents and can be applied to surfaces as a polymer solution to form thin coatings. The dried coating is as resistant to almost as many chemicals as PTFE is.[2]

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007773

MONOMER SYNTHESIS

Synthesis of Tetrafluoroethylene

The first reliable and complete description of synthesis was published in 1933 by Ruff and Bretschneider[3] in which they demonstrated the preparation of TFE (CF_2=CF_2, CAS number *116-14-3*) from the decomposition of tetrafluoromethane in an electric arc. Then TFE was obtained by bromination and separation of the dibromide (CF_2Br–CF_2Br) from the other reaction products. Dehalogenation with zinc was the next step for obtaining pure TFE. Commercially significant techniques for TFE preparation list Fluorspar (CaF_2), hydrofluoric acid, and chloroform as the starting ingredients,[4–11] as shown in the reaction sequence in Fig. 1.

Among other compounds produced are hexafluoropropylene and a small amount of highly toxic perfluoroisobutylene.

Sherratt[12] has provided a description of the preparation of TFE. The overall yield of TFE production depends on the pyrolysis reaction. The products of pyrolysis are cooled, scrubbed with a dilute basic solution to remove HCl, and dried. The resulting gas is compressed and distilled to recover the unreacted $CHClF_2$ and to recover high purity TFE. Polymerization of tetrafluoroethylene to a high molecular weight requires extreme purity, hence the removal of all traces of telogenic hydrogen or chlorine-bearing impurities. Tetrafluoroethylene can autopolymerize if it is not inhibited with terpenes, such as α-pinene, Terpene B, and D-limonene.[13]

HF preparation:

$$CaF_2 + H_2SO_4 \rightarrow 2HF + CaSO_4$$

Chloroform preparation:

$$CH_4 + 3Cl_2 \rightarrow CHCl_3 + 3HCl$$

Chlorodifluoromethane preparation:

$$CHC1_3 + 2HF \rightarrow CHClF_2 + 2HCl$$
(SbF_3 catalyst)

TFE synthesis:

$$2CHC1F_2 \rightarrow CF_2{=}CF_2 + 2HCl$$
(pyrolysis)

Fig. 1 Synthesis reactions of tetrafluoroethylene. (Courtesy of William Andrew Publishing, Inc.)

Synthesis of Hexafluoropropylene

Hexafluoropropylene (CF_3CF=CF_2, CAS number *116-15-4*) was first prepared by Downing et al.[14] by pyrolysis. The full synthesis and identification of HFP was conducted by Henne and Woalkes.[15] A six-step reaction scheme starting with the fluorination of 1,2,3-trichloropropane led to 1,2-dichlorohexafluoropropane, which was dehalogenated with zinc in boiling ethanol to yield hexafluoropropylene.

Synthesis of Perfluoroalkylvinylethers

Perfluoroalkylvinylethers, such as perfluoropropylvinylether (CF_2=CF–O–C_3F_7, CAS number *1623-05-8*) are synthesized according to the steps shown in Fig. 2. There are also electrochemical processes for the production of perfluoro-2-alkoxy-propionyl fluoride.[16]

Synthesis of Chlorotrifluoroethylene

This monomer is fairly simple to manufacture compared with the perfluorinated monomers.[17,18] The commercial process for the synthesis of CTFE (CF_2=$CClF$, 79-38-9) begins with 1,1,2-trichloro-1,2,2-trifluoroethane (TCTFE). It is dechlorinated by pyrolysis at 500–600°C in vapor phase. An alternative method for the preparation of TCTFE is catalytic dechlorination:

$$CCl_3-CCl_3 + HF \rightarrow CCl_2F-CClF_2 + 2HCl$$
(catalyst $SbCl_xF_y$)

$$CCl_2F-CClF_2 + Zn \rightarrow CFCl=CF_2 + ZnCl_2$$
(at 50–100°C in methanol)

The reaction products are put through a number of purification and distillation steps to remove the gaseous and liquid contaminants. Chlorotrifluoroethylene is further purified by the removal of methyl chloride, dimethyl ether, and water by passing the gas stream through sulfuric acid. Water and hydrochloric acid are removed by passing the CTFE through an alumina column before condensing it into a liquid.

Synthesis of Vinylidene Fluoride

There are numerous ways to prepare VDF (CF_2=CH_2, CAS number *75-38-7*). Two methods, including the popular commercial technique for VDF production, are described. Conversion of 1,1,1-trifluoroethane[19] begins by passing this gas through a platinum-lined Inconel tube, which is heated to 1200°C. Contact time is about 0.01 sec. The exit gases are passed through a sodium fluoride bed to remove the hydrofluoric

1. Hexafluoropropylene is converted hexafluoropropylene epoxy (HFPO) reacting HFP with oxygen under pressure in the presence of an inert diluent at 50-250°C or with an oxidizer such as hydrogen peroxide in a basic solution:[43,44]

$$CF_3CF{=}CF_2 + H_2O_2 \rightarrow CF_2{-}CF{-}CF_2 + H_2O \quad \text{(Basic Solution)}$$

HFPO

2. HFPO is reacted with a perfluorinated acylfluoride to produce perfluoro-2-alkoxy-propionyl fluoride:

$$CF_2{-}CF{-}CF_2 + R_f{-}C{=}O \rightarrow R_fCF_2OCF{-}C{=}O$$

Perfluoro-2-alkoxy-propionyl fluoride

3. Perfluoro 2-alkoxy-propionyl fluoride is reacted with the oxygen containing salt of an alkali or alkaline earth metal at an elevated temperature which depends on the type of salt. Examples of the salts include sodium carbonate, lithium carbonate, and sodium tetraborate:[45]

$$R_fCF_2OCF{-}C{=}O + Na_2CO_3 \rightarrow R_fCF_2OCF{=}CF_2 + 2CO_2 + 2NaF$$

Fig. 2 Synthesis of perfluoroalkylvinylethers. (From Ref.[16].)

acid and are then collected in a liquid nitrogen trap. Vinylidene fluoride is separated by low temperature distillation. Unreacted trifluoroethane is removed at −47.5°C and is recycled.

The commercial method begins with hydrofluorination of acetylene followed by chlorination,[20] by hydrofluorination of trichloroethane,[21] or by hydrofluorination of vinylidene chloride.[22] In each case, the final product, 1-chloro-1,1-difluoroethane, is stripped of a molecule of hydrochloric acid to yield vinylidene fluoride.

Synthesis of Vinyl Fluoride

Vinyl fluoride[23,24] was first prepared by the reaction of 1,1-difluoro-2-bromoethane [*359-07-9*] with zinc. Most approaches to vinyl fluoride synthesis involve reactions of acetylene [*74-86-2*] with hydrogen fluoride (HF) either directly or utilizing catalysts. Other routes involve ethylene [*74-85-1*] and HF, pyrolysis of 1,1-difluoroethane [*624-72-6*] and fluorochloroethanes, reaction of 1,1-difluoroethane with acetylene, and halogen exchange of vinyl chloride [*75-01-4*] with HF.[25–27]

MONOMER PROPERTIES

Properties of Tetrafluoropropylene

Tetrafluoroethylene is a colorless, odorless, tasteless, and nontoxic gas, which boils at −76.3°C and melts at −142.5°C. Critical temperature and pressure of tetrafluoroethylene are 33.3°C and 3.92 MPa. It is stored as a liquid; vapor pressure at −20°C is 1 MPa. Its heat of formation is reported to be −151.9 kcal/mol. Polymerization of tetrafluoroethylene is highly exothermic and generates 41.12 kcal/mol heat. The extent of which can be compared to the heats of polymerization of vinyl chloride and styrene, 23–26 and 16.7 kcal/mol, respectively.[28]

Safe storage of TFE requires its oxygen content to be less than 20 ppm. Temperature and pressure should be controlled during its storage. Increasing the temperature, particularly at high pressures, can initiate deflagration in the absence of air (TFE degrades into carbon tetrafluoride). In the presence of air or oxygen, TFE forms explosive mixtures in the molar percentage range of 14–43%. Detonation of a mixture of tetrafluoroethylene and oxygen can increase the maximum pressure to 100 times the initial pressure.

Properties of Hexafluoropropylene

Hexafluoropropylene is a colorless, odorless, tasteless, and relatively low toxicity gas, which boils at −29.4°C and freezes at −156.2°C. In a 4-hr exposure, a concentration of 3000 ppm corresponded to LC50 in rats.[29] Critical temperature and pressure of hexafluoropropylene are 85°C and 3254 MPa. Unlike tetrafluoroethylene, HFP is extremely stable with respect to autopolymerization and may be stored in liquid state without the addition of telogen.

Hexafluoropropylene is thermally stable up to 400–500°C. At about 600°C under vacuum, HFP decomposes and produces octafluoro-2-butene ($CF_3CF{=}CFCF_3$) and octafluoroisobutylene.[30]

Properties of Perfluoroalkylvinylethers

Perfluoroalkylvinylethers (PAVE) form an important class of monomers in that they are comonomers of choice for the ''modification'' of the properties of homofluoropolymers in addition to their broad use in the structure of copolymers of TFE. The advantage of PAVE as modifiers over hexafluoropropylene is their remarkable thermal stability. A commercially significant example is PPVE. It is an odorless and colorless liquid at room temperature. It is extremely flammable and burns with a colorless flame. It is less toxic than hexafluoropropylene.[16]

Properties of Chlorotrifluoroethylene

Chlorotrifluoroethylene is a colorless gas at room temperature and pressure. It is fairly toxic with an LC50 (rat), 4 hr of 4000 ppm.[31] It has a critical temperature and pressure of 105.8°C and 4.03 MPa. Oxygen and liquid CTFE react and form peroxides at fairly low temperatures. A number of oxygenated products are generated by the oxidation of chlorotrifluoroethylene, such as chlorodifluoroacetyl fluoride.[32] The same reaction can occur photochemically in the vapor phase. Chlorotrifluoroethylene oxide is a by-product of this reaction. The peroxides act as initiators for the polymerization of CTFE, which can occur violently.

Properties of Vinylidene Fluoride

Vinylidene fluoride (VDF), $CH_2{=}CF_2$, is flammable and is a gas at room temperature. It is colorless and almost odorless and boils at −84°C. Vinylidene fluoride can form explosive mixtures with air. Polymerization of this gas is highly exothermic and takes place above its critical temperature and pressure.[33]

Properties of Vinyl Fluoride

Vinyl fluoride (VF) [*75-02-5*] (fluoroethene) is a colorless gas at ambient conditions.[23] Vinyl fluoride is flammable in air between the limits of 2.6 and 22 vol.%. Minimum ignition temperature for VF and air mixtures is 400°C. Adding a trace amount (<0.2%) of terpenes is effective to prevent spontaneous polymerization of vinyl fluoride. Inhibited vinyl fluoride has been classified as a flammable gas by the U.S. Department of Transportation.

POLYMERIZATION AND FINISHING

Polytetrafluoroethylene [*9002-84-0*]

PTFE is produced[28,34] by free-radical polymerization mechanism in an aqueous media via addition polymerization of tetrafluoroethylene in a batch process. The initiator for the polymerization is usually a water-soluble peroxide, such as ammonium persulfate or disuccinic peroxide. A redox catalyst is used for low temperature polymerization. PTFE is produced by *suspension* (or *slurry*) *polymerization* without a surfactant to obtain granular resins or with a perfluorinated surfactant (*emulsion polymerization*) to produce fine powder and dispersion products. Polymerization temperature and pressure usually range from 0 to 100°C and 0.7 to 3.5 MPa.

Granular PTFE is produced by polymerizing tetrafluoroethylene alone or by using trace amounts of comonomers. A peroxide initiator, little or no surfactant, and other additives may be present in the aqueous polymerization medium that is vigorously stirred and sometimes buffered by an alkaline solution. Most of the polymer is formed in the gas phase in the shape of stringy and irregularly shaped particles. The particles are comminuted to different sizes, depending on the powder properties required by the fabrication process. For example, a smoother surface part requires smaller particle size while good flow is improved by larger particle size.

Fine powder PTFE is produced by polymerization of TFE in an aqueous medium in the presence of an initiator and surfactant. The polymerization does follow a conventional emulsion mechanism but some of the principles which apply. The stability of the dispersion during the polymerization, to avoid premature coagulation, is balanced against the need to break the emulsion to recover the PTFE. Low shear rate agitation is maintained during the polymerization using surfactant levels below the critical micelle concentration. The rate of polymerization and particle shape and size are affected by the concentration of the surfactant. Majority of the particles is generated in the early part of polymerization and grows as the cycle proceeds. Molecular weight and composition of within the particle can be controlled using the polymerization ingredients and conditions.

The same polymerization process makes aqueous dispersions of PTFE as fine powder. The dispersion is concentrated and stabilized using a variety of ionic and nonionic surfactants. Several concentration methods have been reported including electrodecantation, evaporation and thermal concentration. Chemical additives to match them with the fabrication process or part property requirements can modify the final PTFE dispersion.

Filled compounds of PTFE are produced from all three forms of using fillers, such as glass fiber, graphite, metal powder, carbon fiber, and others.[34]

Perfluoroalkoxy Polymer [*26655-00-5*]

PFA is a copolymer of TFE and a perfluoroalkyl vinyl ether, such as perfluoropropyl vinyl ether PPVE. Copolymerization of perfluoroalkylvinyl ethers with tetrafluoroethylene can be done in a halogenated solvent,[35] in an aqueous phase[36] sometimes containing some halogenated solvent usually in the absence of a surfactant.[37] Terpolymers of this class contain other monomers like hexafluoropropylene HFP.

Commercially, PFA is polymerized by free-radical polymerization mechanism usually in an aqueous media via addition polymerization of TFE and perfluoropropyl vinyl ether. The initiator for the polymerization is usually water-soluble peroxide, such as ammonium persulfate. Chain transfer agents such methanol, acetone and others are used to control the molecular weight of the resin. Generally, the polymerization regime resembles that used to produce PTFE by emulsion polymerization. Polymerization temperature and pressure usually range from 15 to 95°C and 0.5 to 3.5 MPa.

End groups are stabilized by treating the PFA with methanol, ammonia, amines and elemental fluorine that produces CF_3 end groups.[38–42] The polymer is recovered, dried and melt-extruded into cubes for melt fabrications processes. PFA is also available in bead (as polymerized), dispersion and fine powders forms.

Fluorinated Ethylene–Propylene Copolymer [*25067-11-2*]

FEP is a random copolymer of TFE and HFP, which can be polymerized, in an aqueous or a nonaqueous media.[37] Terpolymers of this class contain other monomers, such as perfluoroalkyl vinyl ether (e.g., PPVE) to improve stress crack resistance.

Commercially, it is polymerized by free-radical polymerization mechanism, usually in an aqueous (or nonaqueous) media via addition polymerization of TFE and hexafluoropropylene. The initiator for the polymerization is usually water-soluble peroxide, such as potassium persulfate. Chain transfer agents could be used to control the molecular weight of the resin. In general, the polymerization regime and conditions resemble those used to produce PTFE by emulsion polymerization. For melt fabrication processes, FEP is recovered, dried, and melt-extruded into cubes. It is also available in dispersion form.

Ethylene-co-tetrafluoroethylene Polymers [*68258-85-5*]

This plastic is a partially fluorinated straight-chain polymer with a very high molecular weight.[37] It is produced by free-radical polymerization mechanism in a solvent or a hybrid (a solvent/aqueous mixture) media, using an organic peroxide initiator. Copolymerization of tetrafluoroethylene and ethylene ($CH_2{=}CH_2$, molecular weight 28, CAS number *74-85-1*) proceeds by an addition mechanism.

Copolymers of tetrafluoroethylene and ethylene are highly crystalline and fragile at elevated temperatures and are modified by a third monomer. Production of ETFE terpolymers having improved high temperature mechanical (especially tensile) properties has been demonstrated.[43] They comprise of 40–60 mol% ethylene, 40–60 mol% tetrafluoroethylene, and a small amount of a polymerizable vinyl termonomer, such as perfluoroisobutylene, perfluoropropyl vinyl ether, and hexafluoropropylene.

Ethylene-co-chlorotrifluoroethylene Polymers [*25101-45-5*]

Ethylene and chlorotrifluoroethylene have been polymerized[18] in aqueous and solvent mediums using organic peroxides and oxygen-activated triethylboron. Typically, polymerization is done at 60–120°C and at a pressure of 5 MPa or higher. The polymerization reaction can also be initiated by radiation, such as by gamma rays. The most effective catalyst is tri-*n*-butylboron which produces an ECTFE with an alternating 1 : 1 ethylene : TFE ratio. To control the molecular weight of the resin, chain transfer agents, such as chlorinated compounds, alcohols, and ketones were required.

Polychlorotrifluoroethylene [*9002-83-9*]

Bulk, suspension, and emulsion techniques are used to polymerize CTFE .[44] Bulk polymerization takes place using halogenated acyl peroxide catalysts or UV and gamma rays. Suspension polymerization is carried out in aqueous medium using inorganic or organic peroxide catalysts. Emulsion polymerization yields a polymer with a normal molecular weight distribution and a molecular weight–melt viscosity relationship, similar to that of bulk polymerized polymer. Inorganic peroxy catalysts initiate the reaction in the presence of halogenated alkyl acid salt surfactants. Emulsion polymerization produces the most thermally stable PCTFE.

Polyvinylidene Fluoride [*24937-79-9*]

The first successful aqueous polymerization of vinylidene fluoride was reported in 1948[45], using a peroxide initiator in water at 50–150°C and 30 MPa. No surfactants or suspending agents were present in the polymerization recipe. Polyvinylidene fluoride has been polymerized by a number of methods including emulsion, suspension, solution, and bulk. Later, copolymers of vinylidene fluoride with ethylene and halogenated ethylene monomers were also produced.[46] In 1960, a manufacturing process was developed and PVDF was first introduced to the market. Reaction temperature ranges from 10 to 150°C at a pressure of 1 MPa or higher. Similar to that of TFE, emulsion polymerization of vinylidene fluoride requires a stable fluorinated surfactant and an initiator, such as peroxide or persulfate. Suspension polymerization is conducted in an aqueous medium sometimes in the presence of a colloidal dipsersant like a hydroxy cellulose. Solution polymerization of VDF is conducted in solvents using free-radical initiators. PVDF is commercially produced by aqueous emulsion or suspension process.

Polyvinyl Fluoride [*24981-14-4*]

Vinyl fluoride undergoes free-radical polymerization.[23,24] The first polymerization involved heating a saturated solution of VF in toluene at 67°C under 600 MPa for 16 hr. A wide variety of initiators and polymerization conditions have been explored. Examples of bulk and solution polymerizations exist; however, aqueous suspension or emulsion method is generally preferred. Copolymers of VF and a wide variety of other monomers have been prepared. More recently, interpolymers of VF have been reported with tetrafluoroethylene and other highly fluorinated monomers, such as hexafluoropropylene, perfluorobutylethylene, and perfluoroethylvinylether.

STRUCTURE–PROPERTY RELATIONSHIP

A way to understand the impact of fluorine is to explore the differences between linear polyethylene (PE) and PTFE.[37] There are important differences between the properties of PE and PTFE:

1. PTFE has one of the lowest surface energies in polymers.
2. It is the most chemically resistant polymer.
3. It is one of the most thermally stable polymers.
4. Its melting point and specific gravity are more than double those of PE.

The differences between PTFE and PE are attributable to the differences of C–F and C–H bonds. The differences in the electronic properties and sizes of F and H lead to the following observations:

1. Fluorine is the most electronegative of all elements (4 Paulings).
2. It has unshared electron pairs.
3. It is more easily converted to F^-.
4. Bond strength of C–F is higher than that of C–H.
5. It is larger than hydrogen.

The electronegativity of carbon at 2.5 Paulings is somewhat higher than that of hydrogen (2.1 Paulings) and lower than that of fluorine. Consequently, the polarity of the C–F bond is opposite to that of the C–H bond, and the C–F bond is more highly polarized. In the C–F bond, the fluorine end of the bond is negatively charged when compared with the C–H bond in which the carbon end is negatively charged.

The difference in the bond polarity of C–H and C–F affects the relative stability of the conformations of the two polymer chains. Crystallization of polyethylene takes place in a planar and trans conformation. At extremely high pressure, PTFE can be forced into such a conformation.[47] Below 19°C, PTFE crystallizes as a helix with 0.169 nm per repeat distance; it takes 13 carbon atoms for a 180° turn to be completed. Above 19°C, the repeat distance increases to 0.195 nm, which means that 15 carbon atoms are required for a 180° turn. At above 19°C, the chains are capable of angular displacement, which increases above 30°C until reaching a melting point (327°C).

The substitution of F for H in the C–H bond substantially increases the bond strength from 99.5 kcal/mol for the C–H bond to 116 kcal/mol for the C–F bond. Consequently, thermal stability and chemical resistance of PTFE are higher than those of PE because more energy is required to break the C–F bond. The polarity and the strength of the C–F bond render F atom abstraction mechanism for branching difficult. In contrast, highly branched polyethylene (>8 branches per 100 carbon atoms) can be synthesized. Branching mechanism as a tool to adjust crystallinity is not practical for PTFE. Instead, comonomers with pendent groups have to be polymerized with TFE.

Crystallinity of never-melted PTFE is in the range of 92–98%,[28] consistent with an unbranched chain structure; while FEP, a copolymer of tetrafluoroethylene and hexafluoropropylene, has an as-polymerized crystallinity of 40–50%. In FEP, the pendent CF_3 group is bonded to a tertiary carbon that is less thermally stable than primary and secondary carbon atoms. Degradation curves (Fig. 3) indicate degradation onset temperatures of 300°C for FEP (0.02% weight loss) and 425°C for PTFE (0.03% weight loss).

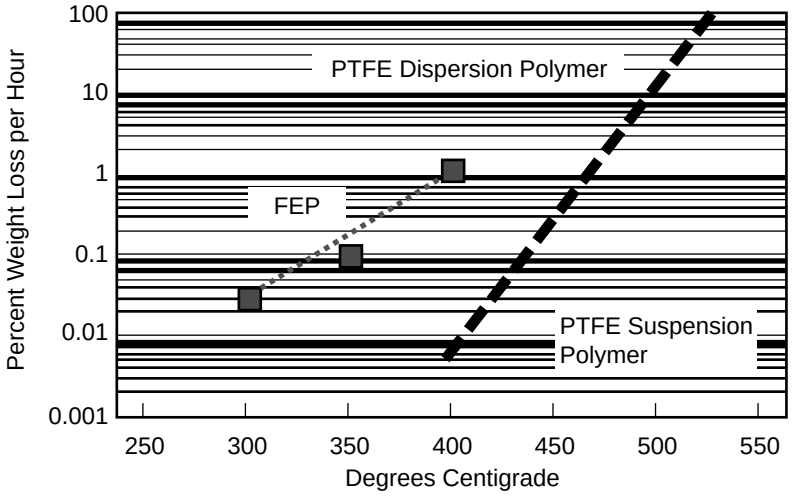


Fig. 3 A comparison of thermal degradation of FEP and PTFE in air. (From Ref.[48].) *(View this art in color at www.dekker.com.)*

POLYMER PROPERTIES

PTFE

It has excellent properties, such as chemical inertness, heat resistance (both high and low), electrical insulation properties, low coefficient of friction (static 0.08 and dynamic 0.01), and nonstick property over a wide temperature range (−260 to + 260°C). It has a density in the range of 2.1–2.3 g/cm^3 and melt viscosity in the range of 1–10 GPa sec.[28] Molecular weight of PTFE cannot be measured by standard methods. Instead, an indirect approach is used to judge molecular weight. Standard specific gravity (SSG) is the specific gravity of a chip prepared according to a standardized procedure. The underlying principle is that lower molecular weight PTFE crystallizes more extensively, thus, yielding higher SSG values.[49]

PTFE that has not been previously melted has a crystallinity of 92–98%, indicating a linear and nonbranched molecular structure. Upon reaching 342°C, it melts, changing from a chalky white color into a transparent amorphous gel. The second melting point of PTFE is 327°C because it never re-crystallizes to the same extent as prior to its first melting.

First order and second order transitions have been reported for PTFE. The transitions that are close to room temperature are of practical interest because of the impact on processing of the material. Below 19°C, the crystalline system of PTFE is a nearly perfect triclinic. Above 19°C, the unit cell changes to hexagonal. In the range of 19–30°C, the chain segments become increasingly disorderly and the preferred crystallographic direction disappears, resulting in a large expansion in the specific volume of PTFE (1.8%),[50] which must be considered in measuring the dimensions of articles made from this plastics.

Polytetrafluoroethylene is by far the most chemically resistant polymer among thermoplastics. The exceptions include molten alkali metals, gaseous fluorine at high temperatures and pressures, and few organic halogenated compounds, such as chlorine trifluoride (ClF_3) and oxygen difluoride (OF_2). A few other chemicals have been reported to attack PTFE at or near its upper service temperature, and PTFE reacts with 80% sodium or potassium hydroxide and some strong Lewis bases including metal hydrides.

Mechanical properties of PTFE are generally inferior to that of engineering plastics at the room temperature. Compounding with fillers has been the strategy to overcome this shortage. In the normally used temperature range, PTFE has useful mechanical properties.

Also, PTFE has excellent electrical properties, such as high insulation resistance, low dielectric constant (2.1), and low dissipation factor. Dielectric constant and dissipation factor remain virtually unchanged in the range of −40 to 250°C and 5 Hz to 10 GHz. Dielectric breakdown strength (short term) is 47 kV/mm for a 0.25 mm thick film (ASTM D149). Dielectric breakdown strength is enhanced with decrease in voids in PTFE, which is affected by the fabrication process. In air, PTFE is attacked by radiation and degradation, beginning at a dose of 0.02 Mrad.

PFA

These polymers are fully fluorinated and melt-processible.[51] They have chemical resistance and thermal stability comparable to those of PTFE. Specific gravity of PFA resins is in the range of 2.12–2.17. It has an upper continuous use temperature of 260°C.

Crystallinity and specific gravity of PFA parts decrease when the cooling rate of the molten polymer

is increased. The lowest crystallinity obtained by quenching molten PFA in ice was 48% (specific gravity 2.123).

Similar to that of PTFE, the molecular weight of PFA cannot be measured by conventional techniques. An indirect factor called melt flow rate (MFR), also called melt flow index (MFI), is used, which is defined as the amount of polymer melt that would flow through a capillary rheometer at a given temperature under a defined load (usually, grams in 10 min). It is inversely proportional to viscosity; viscosity is directly proportional to molecular weight of the polymer.

In contrast to the two transition temperatures for PTFE at 19 and 30°C, PFA exhibits one first order transition at −5°C. It has three second-order transitions at −100, −30 and 90°C.[18] It has excellent electrical properties , such as high insulation resistance, low dielectric constant (2.1), and low dissipation factor. Dielectric constant and dissipation factor remain virtually unchanged in the range of –40 to 250°C and 10^2 to 2.4×10^{10} Hz. Dielectric breakdown strength (short term) is 80 kV/mm for a 0.25 mm thick film (ASTM D149). Chemical properties of PFA are similar to those of PTFE. In air, PFA is attacked by radiation and degradation, beginning at a somewhat higher dose than that of PTFE.

FEP

Fluorinated ethylene–propylene copolymers are fully fluorinated and melt-processible.[52] They have excellent chemical resistance and thermal stability. Specific gravity of FEP resins is in the range of 2.13–2.15. It has an upper continuous use temperature of 200°C.

Similar to that of PTFE, molecular weight of FEP cannot be measured by conventional techniques. As in the case of PFA, MFR is used to characterize the molecular weight of FEP. Molecular weight distribution is determined by measuring the dynamic moduli of the polymer melt, using rheological analyses. Crystallinity of virgin (unmelted) FEP is 65–75%. It exhibits a single first order transition that is its melting point. Relaxation temperature of FEP increases with hexafluoropropylene content of the copolymer. It has a dielectric transition at −150°C which is unaffected by the monomer composition or crystallinity (specific gravity). Chemical properties of FEP are similar to those of PTFE and PFA. In air, FEP is attacked by radiation and its degradation begins at a dose of 0.2 Mrad (10 times higher than that of PTFE).

PCTFE

It is a semicrystalline polymer[44] with a helical polymer chain and a pseudohexagonal crystal. Crystal growth is spherulitic and consists of folded chains. The large size of chlorine constrains recrystallization after melting during the processing. This resin has good properties at cryogenic temperatures, though it is inferior to all fluoropolymers except PVDF.

In addition, PCTFE has exceptional barrier properties and superb chemical resistance. It is attacked by a number of organic solvents. It has low thermal stability and degrades upon reaching its melting point, requiring special care during processing.

ETFE

Equimolar ETFE and PVDF are isomers, but the former has a higher melting point and a lower dielectric loss than the latter. It crystallizes into unit cells believed to be orthorhombic or monoclinic.[53] The molecular conformation of ETFE is an extended zigzag. This polymer dissolves in some boiling esters at above 230°C, thus allowing the determination of molecular weight (weight-average) by light scattering. It has several transitions, alpha relaxation at 110°C (shifts to 135°C at higher crystallinity), beta at −25°C, and gamma relaxation at –120°C.

Terpolymers of ETFE have good mechanical properties including tensile and cut-through resistance and lower creep than perfluoropolymers. It is more resistant to radiation than perfluoropolymers (modestly affected up to 20 Mrad) and can be cross-linked by radiation, such as electron beam. Cross-linking is used to strengthen cut-through resistance of ETFE wire insulation. It has a dielectric constant of 2.6–3.4 and a dissipation factor of 0.0006–0.010 as frequency increases from 10^2 to 10^{10} Hz. Terpolymers of ETFE are resistant to stress cracking and chemical attack by most compounds. Strong oxidizing acids, concentrated boiling sulfonic acids, organic bases (amines), and any chemical that affects PTFE, PFA, and FEP attack ETFE.

ECTFE

It is semicrystalline (50–60%) and melts at 240°C (commercial grade).[44] It has an alpha relaxation at 140°C, a beta at 90°C, and gamma relaxation at −65°C. Conformation of ECTFE is an extended zigzag in which ethylene and CTFE alternate. The unit cell of ECTFE's crystal is hexagonal.

As in the case of ETFE, ECTFE terpolymers (same termonomers) have better mechanical, abrasion, and radiation resistance than those of PTFE and other perfluoropolymers. Dielectric constant of ECTFE is 2.5–2.6, and it is independent of temperature and frequency. Dissipation factor is 0.02 and much larger than ETFE's. ECTFE is resistant to most chemicals except for hot polar and chlorinated solvents. It does

not stress crack dissolve in any solvents. It has better barrier properties to SO_2, Cl_2, HCl, and water than FEP and PVDF.

PVDF

Polyvinylidene fluoride is a semicrystalline polymer (35–70% crystallinity) with an extended zigzag chain.[33] Head-to-tail addition of VDF dominates, but there are head-to-head or tail-to-tail defects that affect crystallinity and properties of PVDF. It has a number of transitions, and its density alters for each polymorph state. There are four known states, named as α, β, γ, and δ, and a proposed state. The most common phase is α-PVDF, which exhibits transitions at −70°C (γ), −38°C (β), 50°C (α″), and 100°C (α′).

It resists most organic and inorganic chemicals including chlorinated solvents. Strong bases, amines, esters, and ketones attack this resin. The impact ranges from swelling to complete dissolution in these solvents, depending on the conditions. It exhibits compatibility with a number of polymers. Commercially useful blends with acrylics and methacrylics have been developed. Just as ETFE, PVDF readily cross-links as a result of exposure to radiation. Radiation (gamma rays) has modest effect on the mechanical properties of PVDF.

PVF

Poly(vinyl fluoride) is a semicrystalline polymer with a planar, zigzag conformation.[23,24] The degree of crystallinity can vary significantly from 20–60% and is a function of defect structures. Commercial PVF is atactic, contains approximately 12% head-to-head linkages, and displays a peak melting point of about 190°C (52,53,62,63). Poly(vinyl flouride) displays several transitions below the melting temperature. Lower T_g occurs at −15 to −20°C and upper T_g is in the 40–50°C range. Two other transitions at −80 and 150°C have been reported.

Below about 100°C, PVF has low solubility in all solvents. Polymers with greater solubility have been prepared using 0.1% 2-propanol polymerization modifier and were characterized in *N*,*N*-dimethylformamide solution containing 0.1 N LiBr. M_n ranged from 76,000 to 234,000 (osmometry), and M_s ranged from 143,000 to 654,000 (sedimentation velocity). High molecular weight PVF is reported to degrade in an inert atmosphere, with concurrent HF loss and backbone cleavage occurring at about 450°C. In air, HF loss occurs at about 350°C, followed by backbone cleavage around 450°C.

It is transparent to radiation in the UV, visible, and near IR regions, transmitting 90% of the radiation from 350 to 2500 nm. It becomes embrittled upon exposure to electron-beam radiation of 1000 Mrad, but resists breakdown at lower doses. It retains its strength at 32 Mrad, while polytetrafluoroethylene is degraded at 0.2 Mrad. The self ignition temperature of PVF film is 390°C. The limiting oxygen index (LOI) for PVF is 22.6%. Hydrogen fluoride and a mixture of aromatic and aliphatic hydrocarbons are generated from the thermal degradation of PVF.

FABRICATION TECHNIQUES

With the exception of two fluoropolymers, PVF and PTFE, the rest of the resins described in this entry can be processed by standard melt-processing techniques, such as injection, transfer and blow molding, extrusion, and rotational molding. Process equipment for fluoropolymers must be made from corrosion resistant alloys because of the corrosive compound that may be produced when fluoropolymers are heated above their melting points. Higher melt viscosity of these resins may require more powder and higher pressure rating equipment.

Metal powder processing techniques in which a preform is molded and ''sintered'' are used to process PTFE. Compression molding may also be used to fabricate PTFE parts. Its dispersions are applied by similar techniques to other coatings. Paste extrusion in which PTFE is blended with a hydrocarbon, prior to molding a preform, is used to continuously fabricate PTFE into tubes, tapes, and wire insulation. The hydrocarbon is vaporized before the parts are sintered. PVF is dispersed in a polar latent solvent such as dimethyl acetamide and is melt-extruded as a plastisol, followed by solvent removal by drying.

APPLICATIONS

Properties of fluoropolymers that have led to applications include chemical resistance, thermal stability, cryogenic properties, low coefficient of friction, low surface energy, low dielectric constant, high volume and surface resistivity, and flame resistance. Fluoropolymers are used as liners (process surface) because of their resistance to chemical attack. They provide durable, low maintenance and economical alternatives to exotic metals for use at high temperatures without introducing impurities. Electrical properties make fluoropolymers highly valuable in electronic and electrical applications as insulation, e.g., FEP in data communications.

Mechanical properties of fluoropolymers are beneficial in low-friction bearings and seals that resist attack by hydrocarbons and other fluids in automotive and

office equipment. In food processing, the Food and Drug Administration approved grades are fabrication material for equipment. In houseware, fluoropolymers are applied as nonstick coatings for cookware and appliance surfaces. Medical articles, such as surgical patches and cardiovascular grafts rely on the long-term stability of fluoropolymers, as well as on their low surface energy and chemical resistance.

For airports, stadiums, and other structures, glass fiber fabric coated with PTFE is fabricated into roofing and enclosures, where it provides excellent resistance to weathering, including exposure to UV rays in sunlight, flame resistance for safety, and low surface energy for soil resistance and easy cleaning.

SAFETY

Fluoropolymers are chemically stable and inert or relatively unreactive. Reactivity, generally, decreases as fluorine content of the polymer increases. Fluorine induces more stability than chlorine. Fluoropolymers can produce toxic products if overheated. Precautions should be taken to exhaust any degradation fragments produced during the processing and fabrication of parts from fluoropolymers.

This family of plastics has low toxicity and almost no toxicological activity. No fluoropolymers have been known to cause skin sensitivity and irritation in humans. It has been shown that PVF does not cause any skin reaction in human beings.[1] Excessive human exposure to fluoropolymer dust resulted in no toxic effects, although urinary fluoride content increased.[2]

ECONOMY

Fluoropolymers are more costly to produce than polyolefins and many other plastics because of the capital costs and the cost of fluorine. Polymerization and finishing of these resins require processing of highly flammable hazardous materials, thus, mandating the use of expensive construction material and elaborate equipment. In early 2004, the cost ranges from US$6 per kg for mechanical grade PTFE to US$70 per kg for specialty grade PFA. Soluble perfluoropolymers cost $10–20 per gram and are only used in high value applications.

CONCLUSIONS

Commercial fluoropolymers are based on tetrafluoroethylene, vinylidene fluoride, and, to a lesser extent, chlorotrifluoroethylene. Examples of comonomers include perfluoromethyl vinyl ether, perfluoroethyl vinyl ether, perfluoropropyl vinyl ether, hexafluoropropylene, chlorotrifluoroethylene, and perfluorobutyl ethylene. The general consequences of substitution of fluorine for hydrogen in a polymer include increased chemical and solvent resistance, enhanced electrical properties, such as lower dielectric constant, lower coefficient of friction, higher melting point, increased photostability and thermal stability, improved flame resistance, and weakened mechanical properties. The ultimate properties are achieved when a polymer is entirely fluorinated.

REFERENCES

1. Plunkett, R.J. The history of polytetrafluoroethylene: discovery and development. In *High Performance Polymers: Their Origin and Development*, Proceedings Symp. Hist. High Perf. Polymers at the ACS Meeting in New York, April, 1986; Seymour, R.B. Kirshenbaum G.S., Eds.; Elsevier: New York, 1987.
2. Teflon® AF; www.dupont.com/Teflon/AF (accessed 2003).
3. Ruff, O.; Bretschneider, O.Z. Anorg. Che. **1933**, *210*, 73.
4. Park, J.D. et al. Synthesis of tetrafluorethylene—pyrolisis of monochlorodifluoromethane. Ind. Eng. Chem. **1947**, *39*, 354.
5. Hamilton, J.M. In *Advances in Fluorine Chemistry*; Stacey, M., Eds.; Butterworth & Co., Ltd.: Kent, U.K., 1963; Vol. 3, 117.
6. Edwards, J.W.; Small, P.A. Nature **1964**, *202*, 1329.
7. Gozzo, F.; Patrick, C.R. Nature. **1964**, *202*, 80.
8. Hisazumi, M.; Shingu, H. Japanese Patent 60 15,353.
9. Scherer, O. et al US Patent 2,994,723, Assigned to Farbewerke Hoechst, August 1, 1961.
10. Edwards, J.W.; Sherratt, S.; Small, P.A. British Patent 960,309; Assigned to ICI, June 10, 1964.
11. Ukahashi, H.; Hisasne, M. US Patent 3,459,818; Assigned to Asahi Glass Co, August 5, 1969.
12. Sherratt, S. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 2nd Ed.; Standen, A., Ed.; Interscience Publishers, John Wiley & Sons: New York, 1966; Vol. 9, 805–831.
13. Dietrich, M.A.; Joyce, R.M. US Patent 2,407,405, Assigned to DuPont, September 10, 1946.
14. Downing, F.B.; Benning, A.F.; McHarness, R.C. US Patent 2,384,821, Assigned to DuPont, September 18, 1945.
15. Henne, A.L.; Woalkes, T.P. Fluorinated derivatives of propane and propylene. VI. J. Am. Chem. Soc. **1946**, *68*, 496.

16. Brice, T.J.; Pearlson, W.H. US Patent 2,713,593, Assigned to 3M Co, July 1955.
17. Carpenter, C.P.; Smyth, H.F.; Pozzani, U.C.J. Ind. Hygiene **1949**, *31*, 343.
18. Chandrasekaran, S. Chlorotrifluoroethylene polymers. In *Encyclopedia of Polymer Science and Engineering*, 2nd Ed.; John Wiley & Sons: New York, 1989; Vol. 3, 463–480.
19. Hauptschein, A.; Feinberg, A.H. US Patent 3,188,356, Assigned to Pennsalt Chemicals Corp, June 8, 1965.
20. Schultz, N.; Martens, P.; Vahlensieck, H.J. German Patent 2,659,712; Assigned to Dynamit Nobel AG, July 6, 1976
21. McBee, E.T. et al. Fluorinated derivatives of ethane. Ind. Eng. Chem. **1947**, *39* (3), 409–412.
22. Kaess, F.; Michaud, H. US Patent 3,600,450, Assigned to Sueddeutsche Kalkstickstoff-Werke AG, August 17, 1971
23. Ebnesajjad, S.; Snow, L.G. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th Ed.; John Wiley & Sons, Inc.: New York, 1994; Vol. 11, 683–694.
24. Brasure, D.E.; Ebnesajjad, S. Vinyl fluoride polymers. In *Encyclopedia of Polymer Science and Engineering*, 2nd Ed.; John Wiley & Sons, Inc.: New York, 1989; Vol. 17, 468–491.
25. Coffman, D.D.; Cramer, R.; Rigby, G.W. Synthesis of chlorofluoropropanes. J. Am. Chem. Soc. **1949**, *71*, 979–980.
26. Coffman, D.D.; Raasch, M.I.; Rigby, G.W.; Barrich, P.L.; Hanford, W.E. Addition reactions of tetrafluoroethylene. J. Org. Chem. **1949**, *14*, 747–753.
27. Pajaczkowski, A.; Spoors, J.W. Chem. Ind. **1964**, *16*, 659.
28. Gangal, S.V. Polytetrafluoroethylene, homopolymers of tetrafluoroethylene. In *Encyclopedia of Polymer Science and Engineering*, 2nd Ed.; John Wiley & Sons: New York, 1989; Vol. 16, 577–600.
29. Clayton, J.W. Occup. Med. **1962**, *4*, 262–273.
30. Gibbs, H.H.; Warnell, J.J. British Patent 931,587; Assigned to DuPont, July 17, 1963
31. Carpenter, C.P.; Smyth, H.F.; Pozzani, U.C. The assay of acute vapor toxicity, and the grading and interpretation of results on 96 chemical compounds. J. Ind. Hygiene **1949**, *31*, 343.
32. Haszeldine, R.N.; Nyman, F.J. Chem. Soc. **1959**, 1085.
33. Dohany, J. Poly(vinylidene fluoride). In *Kirk-Othmer Encyclopedia Chemical Technology*, 4th Ed.; John Wiley & Sons: New York, 1994; Vol. 11, 694–712.
34. Ebnesajjad, S. Non-Melt Processible Fluoroplastics: The Definitive User's Guide and Data Book, *Plastics Design Library*. William Andrew Publishing: Norwich, NY, 2000.
35. Bro, M.I. US Patent 2,952,669, Assigned to DuPont Co, September 13, 1960.
36. Berry, K.L. US Patent 2,559,752, Assigned to DuPont Co, July 10, 1951.
37. Ebnesajjad, S. Melt Processible Fluoroplastics: The Definitive User's Guide and Data Book, *Plastics Design Library*. William Andrew Publishing: Norwich, NY, 2002.
38. Carlson, D.P. US Patent 3,674,758, Assigned to DuPont Co, July 15, 1972.
39. Carlson, D.P. Heat stable tetrafluoroethylene-perfluoro(alkyl vinyl ether) copolymers US Patent 4,599,386, Assigned to DuPont Co., July 8, 1986.
40. PCT Int. Appl. WO89,11,495, M.D. Buckmaster to DuPont Co., 1989.
41. Imbalzano, J.F.; Kerbow, D.L. Stable tetrafluoro thylene copolymers US Patent 4,743,658; Assigned to DuPont Co., May 10, 1988.
42. Goodman, J.; Andrews, S. Fluoride contamination from fluoropolymers in semiconductor manufacture. Solid State Technol., June 1990.
43. Carlson, D.P. US Patent 3,624,250, Assigned to DuPont Co, November 30, 1971.
44. Miller, W.A. Chlorotrifluoroethylene-ethylene copolymers. In *Encyclopedia of Polymer Science and Engineering*, 2nd Ed.; John Wiley & Sons: New York, 1989; Vol. 3, 480–491.
45. Ford, T.A.; Hanford, W.E. US Patent 2,435,537, Assigned to DuPont Co, February 3, 1948.
46. Ford, T.A. US Patent 2,468,054, Assigned to DuPont Co, April 26, 1949.
47. England, D.C. et al.; Proceedings of Robert A. Welch Conference on Chemical Research XXVI, R.A. Welch Foundation; 1982; 193–243.
48. Baker, B.B.; Kasprzak, D.J. Thermal degradation of commercial fluoropolymer in air. Polym. Degrad. Stabil. **1994**, *42*, 181–188.
49. Sperati, C.A., Starkweather, H.W., Jr. Adv. Polym. Sci. **1961**, *2*, 465.
50. McCrum, N.G. An internal friction study of polytetrafluoroethylene. J. Polym. Sci. **1959**, *34*, 355.
51. Gangal, S.V. Tetrafluoroethylene-perfluorovinyl ether copolymer. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th Ed.; John Wiley & Sons: New York, 1994; Vol. 11, 671–683.
52. Gangal, S.V. Perfluorinated ethylene-propylene copolymer. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th Ed.; John Wiley & Sons: New York, 1994; Vol. 11, 644–656.
53. Gangal, S.V. Tetrafluoroethylene-ethylene copolymers. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th Ed.; John Wiley & Sons: New York, 1994; Vol. 11, 657–671.

Fouling of Heat Exchangers

T. Reg. Bott
School of Engineering, Chemical Engineering, University of Birmingham, Birmingham, U.K.

INTRODUCTION

During the course of operation of heat exchangers, which is usually 24 hr per day and 365 days per year, their heat transfer surfaces generally become dirty. This is generally known as fouling. The possible origins of the unwanted deposit are numerous, and almost any fluid being processed can give rise to deposit formation, depending on operating conditions. The presence of a deposit causes an added resistance to the transfer of heat, reducing the effectiveness of the exchanger and, therefore, the overall efficiency of the process of which the exchanger forms a part. The shortfall in energy recovery has to be made up usually, but not always, from the combustion of fossil fuel, with implications for the environment. Furthermore, the presence of the deposit increases the pressure loss through the exchanger. It also leads to increased maintenance, particularly in respect of cleaning the exchanger when the increasing backpressure can no longer be tolerated. The higher upstream pressure, as a result of the restrictions to flow imposed by the deposit, can also lead to leaking joints in pipe work and greater wear and tear on pumping equipment. In some examples of heat exchanger fouling, the presence of the deposit can lead to problems of corrosion of the heat exchanger surfaces beneath the deposit.

The loss of production during maintenance and reduced throughput because of the reduced capacity of the process plant caused by unwanted deposits can also add substantially to operating costs.

To cope with the presence of the fouling and its effect on the exchanger efficiency, additional heat transfer area is usually incorporated into the design of the exchanger. The additional area may be quite substantial, depending on the anticipated level of fouling. The result is an addition to the capital cost of the process plant.

BASIC MECHANISMS OF DEPOSITION

Three basic steps can be visualized in relation to the accumulation of deposits on heat exchanger surfaces. They are

1. The transport of the foulant material or its precursors from the flowing fluid, across the boundary layers adjacent to the heat exchanger surface.
2. The adhesion of constituents of the deposit on the heat transfer surface.
3. The transport of material that may become detached away from the surface.

The summation of these steps results in the growth of a deposit on the surface. The rate of accumulation of deposit can be simply visualized as the difference between the rate of deposition and the rate of removal. It may be expressed in mathematical terms as:

$$\text{rate of deposit growth} = \phi_D - \phi_R \quad (1)$$

where ϕ_D and ϕ_R are the rates of deposition and removal, respectively.

Curve A in Fig. 1 provides an idealized picture of the growth of foulant on a surface with time. The curve is asymptotic in shape. Three regions can be identified:

1. Adhesion, where the growth rate is relatively small and where there could be a period of initiation before any substantial deposit is present on the surface.
2. A period of relatively rapid growth before the rate of growth begins to slow.
3. A tendency to produce a plateau where the growth is reduced to zero.

In Fig. 1, an indication of two other variations that can occur is also given. The curves show a linear increase (C) and a falling rate (B) of increase in deposit attached to the surface. It is not easy to decide whether these are distinctly different types of curve or simply parts of a different, but still asymptotic curve.

Fig. 2 is what might be regarded as a ''practical'' version of Fig. 1A. It demonstrates that as the deposit grows weaknesses in its structure allow pieces of the deposit to be removed, so that the general shape is retained, but modified by the saw-tooth appearance. The weaknesses in the deposit structure are generally because of some inconsistencies in the depositing material either chemically or physically, or both.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120034101

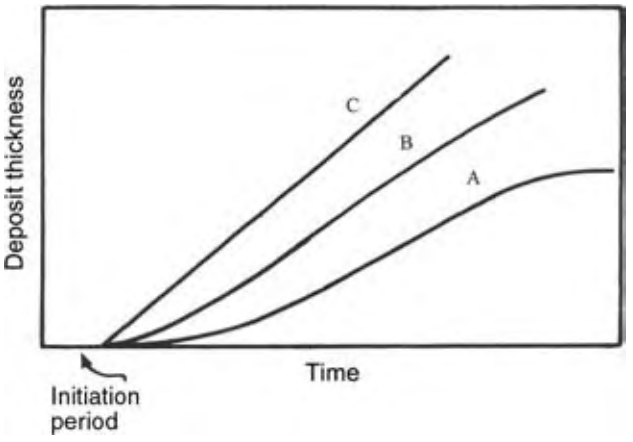


Fig. 1 Idealized deposition curves. (From Ref.[2].)

Deposition of Particles

The deposition of particles can occur in heat exchangers both in liquid streams and in processes involving the flow of gases. The origin of the particles is extremely diverse. They may already be in the fluid being processed, unintentionally produced as a result of the process itself, or corrosion products from upstream.

Basic concepts

In general, two processes are responsible for the deposition of particles on surfaces. They are diffusion transport and settling under the influence of gravity. Both these mechanisms are possible causes of particulate fouling of heat exchangers, although particle transport from a moving fluid is by far the most common cause and will form the basis of this discussion; settling usually occurs in stationary or slow-moving fluids.

Particle deposition from a moving fluid involves two aspects. First the individual particles have to be transported to the surface by one or a combination of mechanisms, including Brownian motion, turbulent diffusion, and by virtue of the momentum possessed by the particle, as it is carried in the fluid stream. It will be appreciated that the size of the particle will exert a great influence on the dominant mechanism. Larger particles would be expected to move toward a surface, as a result of the momentum they possess. Finer particles with relatively low momentum can only approach the surface across the boundary layers, by Brownian or eddy diffusion. Having reached the surface to form a part of the foulant layer, the particle has to ''stick'', but it may be removed from the surface by the shear forces produced by the flowing fluid [see Eq. (1)].

In Fig. 3, a particle concentration profile in a turbulent flowing fluid is shown. The concentration is virtually uniform outside the viscous sublayer, i.e., c_b, and the concentration at the surface is c_s. The transport of particles is given by

$$\phi_D = K_t(c_b - c_s) \tag{2}$$

where K_t is the appropriate transport coefficient.

It is possible to consider that the concentration of particles at the surface is zero as the particles at the surface are taken out of the fluid and incorporated into the deposit. It is possible, therefore, to rewrite Eq. (1) as

$$\phi = K_d c_b \tag{3}$$

where K_d is the deposition coefficient and it is only identical with K_t when all the particles that arrive at the surface remain there. When some of the particles that arrive at the surface do not stay, either they

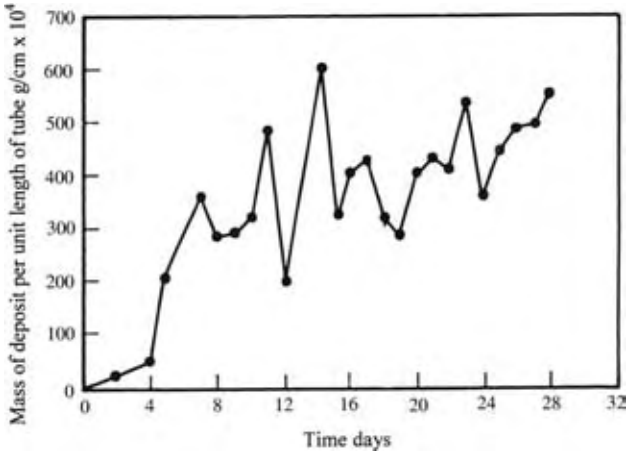


Fig. 2 ''Practical'' deposition curve.

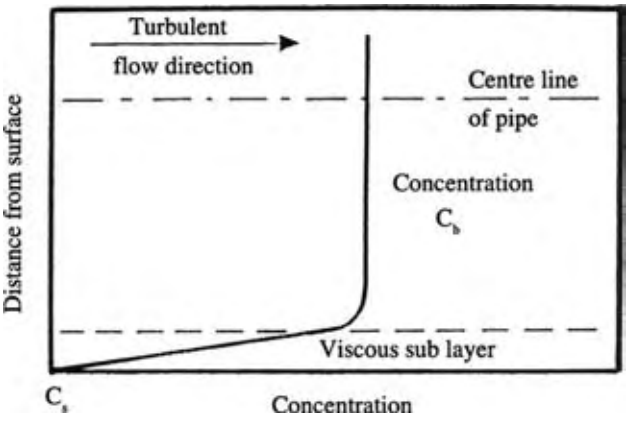


Fig. 3 Particle concentration distribution in a turbulent flowing fluid in a pipe.

rebound back into the fluid or they are swept from the surface by the movement of the fluid, then $K_d < K_t$.

$$K_d = PK_t \quad (4)$$

The surface chemistry of the particles and the surface onto which the particles deposit will also affect the accumulation of the deposit. In addition, the transport properties of the particles will be influenced if agglomeration takes place. Deposition and adhesion are complex phenomena.

In Fig. 4, some data on the deposition of the corrosion product haematite from water flowing through 316 stainless steel tubes at different pH values is provided.[1] Conditions under which these data were obtained include particle size of approximately 0.2 μm, particle concentration of 100 mg/kg, and a Reynolds number of 11,000.

It will be seen that the curves of deposit accumulation with time follow the general pattern of deposition discussed earlier. The effect of water pH is very marked with a maximum deposition occurring at a pH of around 6. The pH controls the magnitude and sign of the charge on the surface to which the particles attach, and the surface of the particles themselves. It might be possible to control particulate deposition by pH adjustment, but the additional problem of corrosion would have to be taken into account.

In combustion systems, it is very likely that the mineral particles going forward with the flue gases will be molten, because of the high temperature in the combustion zone. For this reason, the sticking ability of these particles will be enhanced. It is also possible that the presence of certain combinations of minerals will give a reduced melting point, which will aggravate the fouling problem.

The foregoing discussion applies in general, to isothermal conditions. If particles in suspension are located in a temperature field, there is a tendency for the particles to move down the temperature gradient. The phenomenon is referred to as thermophoresis, but sometimes it is called ''thermal diffusion.'' In heat exchangers, clearly there is a temperature gradient between the hot and the cold fluids, but in many heat exchangers this does not give rise to any significant particle movement, because the temperature difference is relatively small. In addition, the fluid systems are likely to be liquids that exert a resistance, because of their viscosities, on the movement of particles. Thermophoresis effects are more significant in combustion systems, where there is likely to be greater differences between the high temperature flue gases and, say, the surface of heat exchangers in which steam is being generated. In addition, the viscosity of the flue gases is very much lower than that of liquids, so that the viscous drag on the particles is relatively low with the attendant low resistance to particulate transport.

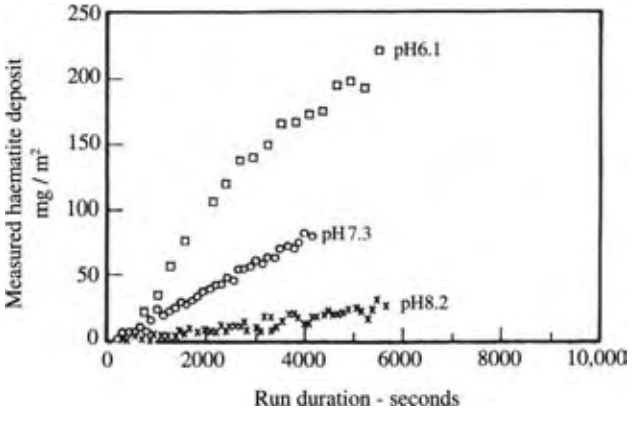


Fig. 4 The effect of pH on hematite deposition inside a tube. (From Ref.[1].)

Scale Formation

The deposition of crystalline salts, which form a scale on heat exchanger surfaces, is a common problem, where naturally occurring waters are used for cooling and in water desalination using evaporators. The deposition of salts from solution in water can take place either on the heat transfer surface itself, or in the bulk water, followed by the migration of the resulting crystals or crystallites to the surface. It will be seen that the distinction between these alternatives is dependent on the temperature distribution in the particular system. Broadly two forms of deposit are possible: a tenacious, ''hard to remove'' accumulation, generally referred to as ''scale,'' and a soft deposit that may be described as ''sludge.''

The origin of the water will have a profound influence on its quality in terms of dissolved salts. Much will depend on the chemical composition of the rock, sand, and soil, which the water has encountered after falling as rain or snow, as it flows into a river, lake, or canal. Geothermal water is often used as a source of heat energy in places where it is accessible, and it can give rise to serious fouling problems during utilization. Wherever a saturated solution, whether organic, inorganic, aqueous, or non-aqueous, is passed through a heat exchanger, there could be a risk of deposition occurring.

Basic concepts

A solution of a soluble salt in contact with its solid phase is said to be saturated with respect to that salt at the temperature of the solution. Under certain conditions, however, it is possible to have solutions of a concentration higher than the saturated concentration at the particular temperature; the condition is generally

known as ''supersaturation.'' It is generally accepted that a degree of supersaturation must be present, before scaling can occur on heat transfer surfaces.

Two different solubility/temperature relationships are possible depending on the chemical make up of the salt under consideration. Many salts have greater solubility as the temperature of the solution is raised. Such salts are generally referred to as normal solubility salts, e.g., NaCl and $NaNO_3$. In Fig. 5, it is demonstrated how the initial concentration of the salt at point A is affected as the temperature of the solution is reduced. Eventually, the temperature reaches T_1 on the solubility curve. Further cooling brings the temperature to T_2 at point C and the solution can be regarded as supersaturated and precipitation occurs. Point C is a metastable point. Further cooling brings the solution temperature to T_3 and point D on the saturation curve. The production of crystals in the cooling process could be responsible for fouling in any associated heat exchanger.

In contrast, inverse solubility salts are less soluble as the temperature of the solution is raised. Examples of inverse solubility salts are $CaCO_3$ and $CaSO_4$. It will be readily appreciated that water containing inverse solubility salts used for cooling purposes is likely to cause fouling problems because as heat is abstracted by the water, its temperature will rise, and if it is saturated with inverse solubility salts, precipitation of the salts will occur. In Fig. 6, the sequence of events that occurs when an inverse solubility salt is heated is illustrated. The curve represents the solubility variation with temperature. The solution of the salt at point A is not saturated. As the solution represented by A is heated it will eventually reach the solubility curve at

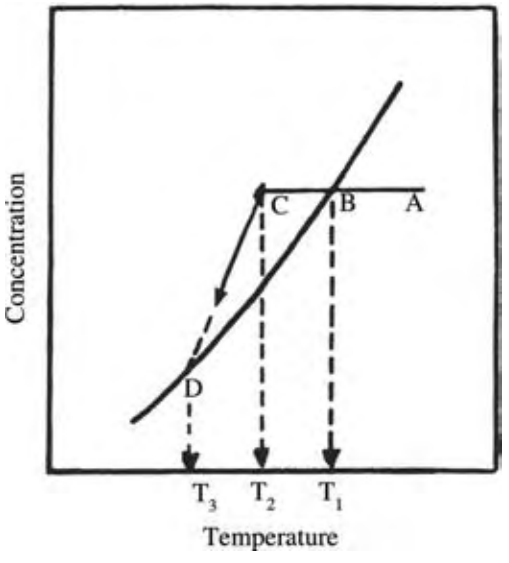


Fig. 5 Cooling a ''normal'' solubility salt solution. (From Ref.[2].)

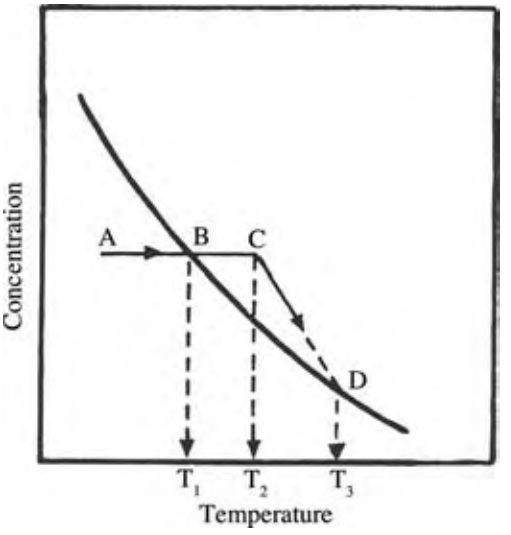


Fig. 6 Cooling an ''inverse'' solubility salt solution. (From Ref.[2].)

B when its temperature is T_1. Further heating brings the solution to C. At this point the solution is supersaturated at a temperature of T_2. It is in a metastable condition and precipitation can be expected. As heating is continued, precipitation reduces the salt concentration till it falls onto the solubility curve at temperature T_3.

The extent of the supersaturation will in general provide the ''driving force'' for the crystallization process. Three steps may be envisaged:

1. Supersaturation is produced.
2. Crystal nuclei and crystallites are formed.
3. Crystal growth is initiated either in the bulk liquid or on the surface.

In general, the initiation of the precipitation process may result from the presence of particulate matter in the bulk water that seeds the crystallization. The process is usually termed ''heterogeneous'' nucleation. It is possible for ''homogeneous'' nucleation to occur when the nucleation is spontaneous. Once nucleation has occurred, crystals can grow, provided that the solution is supersaturated. Suitable nucleation points on the heat transfer surface facilitate deposit formation on the surface. In turbulent flow, it is possible that crystallites that are formed in the bulk fluid may be carried into regions, where they can redissolve.

The deposit resulting from the precipitation from solution would be expected to be crystalline in character, but it is usually accepted that for one reason or another, the structure is not uniform and will contain imperfections. Much will depend on the manner by which the crystals are laid down from the flowing water. If the crystals are formed in situ on the heat

transfer surface, the crystalline structure is likely to be different in detail from crystals formed from crystallites that are produced in the bulk solution and migrate to the surface to be incorporated into the accumulating deposit. A common defect is a missing unit from the crystal lattice leaving a ''hole'' in the structure. The vacant site could be subsequently occupied by an impurity. Furthermore, gaps between crystal units could be filled with ''foreign bodies'', such as solid particles, which are very different from the basic crystal structure. It is also possible to incorporate ions that are different from the ions that form the basic structure of the deposit. All these modifications to the pure crystal habit will introduce imperfections that will affect the robustness of the deposit. In addition, because of the random way in which the structure is formed, there may be ''line imperfections'' where adjacent structures are growing and are unable to blend into a regular crystal structure. As a result, shearing within the deposit is possible under a suitable applied force.

The discussion has concerned the transfer of sensible heat and the associated scale formation. In addition, an important area for processing is when fouling accompanies boiling heat transfer. The problem is encountered in such processing as the crystallization of final products, such as sugar and table salt or in the desalination of seawater by flash evaporation. There is a potential problem of scale formation in boiler plant, but the technology is such that in modern boiler plant, the problem is well controlled by the use of chemical additives.

Circular ''volcano''-shaped deposits can occur as a result of bubble formation in connection with boiling heat transfer and these deposits can act as nucleation points for further evaporation and deposition. As the process of deposition continues, the surface conditions will be modified and the scale may contain fissures, so that steam formation may occur within the deposit. The process is generally referred to as ''wick boiling'' that may influence the mechanism of heat transfer. The process of scale formation under boiling conditions is complex.

The deposition of organic compounds on cooled heat transfer surfaces could be regarded as crystallization, although it might also fall into the category of ''freezing fouling.'' The phenomenon is often associated with the cooling of crude oil or petroleum fractions that contain a mixture of molecules with different molecular weights. The problem is more likely to be a nuisance in the transport of waxy crude oils rather than in heat transfer equipment.

Freezing of a Flowing Liquid

Where a flowing liquid is being cooled, for instance, the production of chilled water, it is possible for solid to be formed on the heat transfer surface, if that surface is below the freezing point of the liquid. The presence of this frozen layer represents a resistance to heat removal from the flowing liquid. The occurrence of liquid solidification as a fouling mechanism is not common, but it can be a problem in certain specific processes. Freezing fouling may occur, for instance, in the condensation of some products, where the surface of the condenser is below the freezing point of the condensate. Clearly this is a complex phenomenon.

Basic concepts

In freezing fouling, the deposit would be expected to be crystalline and this would suggest that it could be regarded as being similar to crystallization. There is a critical difference, however, in that the deposit-forming species is already in contact with the surface, so there is no diffusion required to bring the foulant precursor to the surface prior to deposition. It would be expected that the process of deposit formation would be quite rapid, particularly if there is a relatively low temperature on the other side of the heat transfer surface on which deposition is occurring. The thickness of the deposit will be determined by the limitation imposed by the fact that the outer surface of the deposit will eventually be the freezing point of the flowing liquid. It will be apparent that the lower the temperature of the coolant stream, the thicker will be the accumulation of solidified liquid. The thickness is unlikely to be uniform, however, because the distribution of temperature difference is likely to vary along the length of the heat exchanger. Control of freezing fouling is relatively simple, by adjusting the temperature distribution in the heat exchanger or allowing the temperature to rise on a periodic basis for purposes of ''cleaning.''

Corrosion Fouling

Because of the processing of aggressive fluids or of the conditions under which a heat exchanger is operated, the heat transfer surface may be subject to corrosion, which may be defined as the deterioration and loss of material because of some form of chemical attack. The products of corrosion on the surface offer a resistance to heat transfer. The corrosion is often because of the impurities in the fluid being processed, and it may be accelerated by the presence of other deposits, such as scale or biofilm. It has to be remembered, however, that in many instances a layer of corrosion product, e.g., a thin oxide layer, is necessary to protect the metal from further deterioration. If this protective layer is removed, there could be serious implications for the continued integrity of the equipment involved.

Prevention of corrosion is, of course, possible by the choice of corrosion resistant material, for the construction of the processing equipment, but on the whole, such sophisticated materials are expensive and hence may not be acceptable.

Corrosion is a vast subject in its own right, and it is not possible here to enter into a detailed discussion of corrosion science. It is useful, however, in the consideration of corrosion fouling to provide some basic details that are generally related to iron immersed in water, which provides the background for corrosion fouling.

Basic concepts

Reddish brown rust (ferric hydroxide) is usually evident on the surface of iron or steel subject to the presence of oxygen in the atmosphere and in contact with water. It is produced by the following chemical reaction:

$$4Fe + 3O_2 + H_2O \rightarrow 4Fe(OH)_3$$

The reaction is an electrochemical process, essentially oxidation of iron, which necessitates the removal of electrons from the metal.

$$Fe \rightarrow Fe^{2+} + 2e$$

where e represents an electron,and

$$Fe^{2+} \rightarrow Fe^{3+} + e$$

The removal of metal from the anodic site on the surface of the metal gives rise to metallic ions in solution and an accumulation of excess electrons on the metal surface. The presence of these electrons makes it possible for reactions to occur on an associated cathodic site. A common example is the reduction of dissolved oxygen to hydroxyl ions.

$$O_2 + 2H_2O + 4e \rightarrow 4OH^-$$

The electrochemical corrosion of a metal is illustrated in Fig. 7.

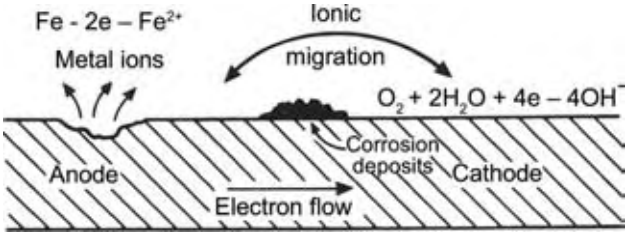


Fig. 7 Electrochemical corrosion of a metal. (From Ref.[2].)

Other reduction reactions are possible for solutions with pH below 7, i.e., under acidic conditions, such as

$$2H^+ + 2e \rightarrow H_2$$

Both anodic and cathodic reactions occur uniformly over the surface in simple examples of corrosion, but the individual sites will depend very much on the surface characteristics such as grain boundaries, crevices and, of course, discontinuities in any protective oxide layer, as the protective film acts as an electrical barrier. The protection realized will depend on the uniformity of the protective layer and the intimacy of its contact with the underlying metal. The potential for corrosion will depend on the pH of the liquid in contact with the metal.

The discussion so far has concentrated on basic corrosion reactions, but enhanced corrosion can occur for a number of reasons.

An expensive special alloy may be employed to eliminate corrosion of the tubes of a shell and tube exchanger from an aggressive liquid flowing through the tubes, but to save capital cost the shell of the exchanger might be fabricated from mild steel. The combination of these metals in the heat exchanger could give rise to corrosion. The less resistant metal of the two is prone to corrode or corrode more rapidly. In Fig. 8, the ionic and electron flow in relation to the two metals in contact in an aqueous environment is shown.[2] Electrons flow across the boundary between metals A and B, while ions are released into solution, and corrosion takes place at C.

Enhanced corrosion can occur for several other reasons:

1. Crevice corrosion is localized attack where there is a crack in the protective oxide layer.
2. Pitting corrosion is localized attack, which could be initiated by occasional fluctuations in

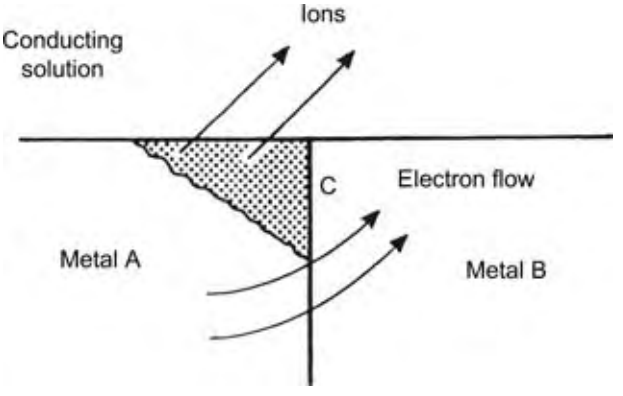


Fig. 8 Electrochemical corrosion with dissimilar metals in contact. (From Ref.[2].)

operating conditions. Oxygen depletion in a pit in the metal will aid the corrosion process.

3. Stress corrosion cracking involves the failure of some alloys in very specific environments under tensile stress conditions present in the heat exchanger.
4. Selective leaching involves the preferential leaching of one of the constituents of an alloy, for instance, the removal of zinc from brass.
5. Corrosion fatigue cracking is the result of a combination of corrosive environment and repeated working, such as the flexing of tubes in a shell and tube heat exchanger.
6. Impingement attack is because of the presence of particulate matter in a fluid passing through a heat exchanger that removes the protective oxide layer.

Corrosion in Gas Streams

The majority of examples of fouling in gas streams generally involve flue gases derived from the combustion of fossil or other fuels. Two different areas, where the problem can arise, are possible based on temperature, but the boundary between the two is not precise.

At relatively low temperatures, the problem of corrosion fouling in heat exchangers is likely to be because of a combination of condensation of water and acidic components from the flue gases, such as SO_2 and SO_3, giving solutions of relatively low pH.

$$SO_2 + H_2O \rightarrow SO_3^{2-} + 2H^+$$

and

$$SO_3 + H_2O \rightarrow SO_4^{2-} + 2H^+$$

The problem will usually occur where the flue gas temperature is reduced to relatively low values in the order of around 120°C, such as in the preheating of the combustion air. To avoid this corrosion problem, it is usual to operate the exchangers involved above the acid dew point temperature, and thus avoiding acid condensation.

Corrosion can occur on the heat transfer surfaces, where heat is being extracted from the combustion gases at high temperatures, e.g., in steam raising plant. In many instances of fossil fuel utilization, it is the impurities in the fuel that give rise to the problem.

In general, high temperature corrosion associated with combustion systems is related to other foulants, such as particulate matter that may be solid or molten. Often corrosion under deposits may be attributed to the presence of alkali sulfates and may cause a severe problem during the combustion of coal. It is possible for sulfates or polysulfates of sodium or potassium to form ionic melts on the heat transfer (or other) surfaces, when the surface temperature is higher than the corresponding melting point.

The molten condition of the deposit can allow electrochemical corrosion cells to be established. Furthermore, the problem may be accompanied by a lowering of the melting point of the deposit because of the presence of fluxing agents. It is at the anodic area that corrosion occurs:

$$Fe \rightarrow Fe^{3+} + 3e$$

At the cathodic site, the electrons are consumed by the reduction:

$$O_2 + 4e \rightarrow 2O^{2-}$$

The molten ash with its ion content provides the pathway that completes the electrical circuit.

As corrosion of metals depends on the presence of aggressive agents in contact with the metal, the availability of these chemical agents is of prime importance. As with other fouling mechanisms, high velocities of the fluid favor diffusion across the boundary layers in contact with the solid surface. The availability of O_2 for the utilization of electrons on cathodic sites will affect the corrosion rate. It could be anticipated that an increased flow rate would raise the rate of corrosion, till it becomes limited by the chemical reaction rate. It is possible also that increased availability of oxygen could improve the quality of the protective oxide layer. The detail of the conditions will determine what occurs at the metal surface. It is probably true to say that each example of corrosion fouling in processing operations is unique.

Chemical Reaction Fouling

Chemical reaction fouling involves the production of insoluble substances in a fluid flowing through a heat exchanger, generally associated with organic substances. It is possible that the metallic surface of a heat exchanger could act as a catalyst to enhance some chemical reactions that lead to surface fouling. The chemical reactions are often complex and may involve autoxidation, polymerization, cracking, and coking reactions. Chemical reaction fouling is to be found principally, although not exclusively, in the mineral oil refining and processing industries, but it may occur in food processing industry also. Because chemical reactions are generally sensitive to temperature, usually temperature is the dominant influence on the extent of any associated chemical reaction fouling in addition to the composition of the material being processed.

The problem of fouling as a result of chemical reaction in petroleum processing can occur in almost all operations, but much will depend on the quality of the feedstock and the intermediates. In petroleum processing, in primary distillation, the crude oil preheaters are very prone to fouling, where the temperature is raised from ambient to around 250°C. Following preheating, the temperature is raised still further in a directly fired pipe still or some similar piece of equipment. Fouling is also prevalent in cracking operations that are used to reduce the molecular weight of the components in the feed stream; the purpose being to provide ''building blocks'' for further processing to meet the market demands.

Chemical reaction fouling can also occur in the pasteurization of milk. Complex chemical deposits appear after a relatively short time so that, apart from overcoming the operational difficulties resulting from the accumulation of deposits, cleaning becomes essential to maintain hygienic conditions in the equipment.

In addition to the organic character of fouling because of the chemical reactions in the petroleum industry, there might be a significant inorganic content including:

1. Corrosion residues from upstream equipment that have been detached and subsequently trapped in the deposits produced by chemical reaction.
2. Organic molecules that contain heteroatoms. Sulfur in some form is contained in most crude oils, incorporated in mercaptans or other organosulfur compounds.
3. Vanadium and nickel are common elements found in certain crude oils existing in combination with complex organic structures.
4. Crude oil may contain an aqueous phase that is likely to contain inorganic salts in solution, despite the inclusion of a ''desalter'' in the process design.

The chemical reactions responsible for the production of deposits that accumulate on heat transfer surfaces may take place on the surface itself. The deposit precursors migrate to the surface from the bulk fluid across the boundary layers by simple diffusion or diffusion brought about by the eddy motion in the fluid stream. An alternative route to the production of deposits from chemical reaction is that the reaction takes place in the bulk and particulate solid products are transported to the surface by mass transfer mechanisms.

As the fouling process is dynamic, the thickness of the deposit will change with time, as indicated in the earlier discussion. As a result, the temperature distribution will change, and if heat passes from the heat transfer surface to the bulk fluid, the surface/deposit interface temperature will rise. The change in temperature may result in a change in the character of the deposit through further chemical reaction, perhaps producing a ''coke like'' dense substrata, whereas the outer layers at the lower temperature remain porous. Over a period of time, the dense portion of the deposit gradually moves outward with a changing influence on the character of the deposit. Released gases from the reactions may produce a porous structure. In general terms, each heat exchanger experiencing chemical reaction fouling will be unique, because of the complexities of the interaction of temperature and process fluid composition. Some observations made as long ago as in 1988 by Crittenden[3] are still relevant:

1. Organic liquids with high molecular weight have a greater tendency to form deposits compared with those with lower molecular weight.
2. Increased deposition rates are apparent for branched chain molecules.
3. Thermally decomposed liquid streams are susceptible to gum formation, but the chemical reactions involved may be modified by the presence of other organic compounds.
4. Dissolved oxygen and its concentration can have a profound effect on chemical reaction fouling rates.
5. Catalytic action by dissolved metallic complexes can enhance the deposition process.
6. Chemicals added to act as corrosion inhibitors might increase the rate of deposit formation.

The effects of operating variables

In common with all fouling mechanisms, two process variables can have a profound influence on the extent and the rate of the fouling process: temperature and fluid velocity across the heat transfer surface.

It is well known that an increase in temperature usually favors chemical reactions, with an exponential increase in the rate constant (r_c). The Arrhenius equation states:

$$r_c \propto \exp(-E/RT)$$

where E is the activation energy, R the universal gas constant, and T the absolute temperature.

A wide range of activation energies has been reported varying between 15 and 120 kJ/mol.[4]

Crittenden[3] observes that there are generally a minimum temperature below which fouling will not be observed and a critical upper temperature above which the deposition rate falls away because of the changes in the underlying reaction. Two aspects of fluid velocity have to be appreciated in connection with the effects of velocity on chemical reaction fouling: heat and mass transfer, both of which are dependent

on the level of turbulence, which is itself dependent on the fluid velocity. Increased turbulence (as a result of increased velocity) will assist not only the rate of arrival of reactants (or intermediates and precursors), but also the removal of products of the reactions from the vicinity of the heat transfer surface. If the production of a foulant on the heat transfer surface is controlled by reaction rate, then an increase in mass transfer, brought about by increased turbulence, will not affect the rate of accumulation of deposit. The effect of the level of turbulence on the viscous sublayer will also affect the temperature of the heat transfer surface with the attendant effect on the rate of reaction and hence the rate of deposition.

It is not possible to provide ''rule of thumb'' indications of the expected effect of changes in the variables such as velocity, temperature, or impurity and its concentration, as the interaction between them can be seen to be quite complex. It is probably for this reason that there are apparent contradictions in the literature, regarding the effects of velocity on deposit accumulation resulting from chemical reactions.

Biofouling

The accumulation of living matter on heat transfer surfaces may be divided into two groups depending on the size of the organisms involved: microorganisms such as bacteria, algae, and fungi and macro-organisms that include mussels, barnacles, hydroids, and serpulid worms, and vegetation such as seaweed. This discussion will focus on macro-organisms on surfaces, as the accumulation of micro-organisms on surfaces was discussed in the entry dealing with biofilms.

In general, the problems associated with macro-organisms relate to water systems, almost exclusively to cooling operations using seawater, in particular to the operation of power plants. The employment of sea or brackish water for cooling operations is usually based on the ''once through'' concept, i.e., the water is abstracted, put through the coolers, and returned to its source. Fouling often occurs in intakes and open culverts, and elaborate screening and filtering techniques are employed to reduce the incidence, or eliminate altogether, the build up of deposits on surfaces.

Restrictions because of filtering and the accumulation of waterborne material (living and dead) may cause restrictions in water flow that might impair the operation of the whole system, including heat exchangers.

In general, conditions favor the development of macrofouling communities, with a continuous flow of aerated seawater that brings food with it to sustain the colonies of living matter. In addition, there is a lack of competition from algae in culverts owing to restricted light conditions and a reduced presence of predators. The waste products from this marine life, together with the decomposition products of dead materials, are potential nutrients for micro-organisms. In the fouling process, it is generally accepted that, initially, the surfaces are colonized by bacteria, forming slime layers that condition the surface to facilitate the attachment of macro-organisms.

As far as the heat exchanger fouling is concerned, problems can occur as a result of the small larvae of mussels and barnacles being carried forward with the water and settling in the parts of heat exchangers where the velocities are low. Partially blocked water tubes of shell and tube heat exchangers may encourage settlement in those tubes. In turn, this can lead to pitting corrosion where growth occurs and also for problems of erosion-corrosion. Accumulation of living material in the header boxes of shell and tube heat exchangers may cause water distribution problems in the heat exchanger.

Mixed Fouling

The discussion has of necessity centered on recognized mechanisms for specific examples of fouling in heat exchangers. It will be appreciated that in most systems, individual mechanisms are not responsible for fouling problems. Combinations of fouling mechanisms are present in most industrial processing operations. The combinations encountered will be very dependent on the operating conditions and the origin and quality of the fluids passing through the heat exchanger. A complex situation exists where it is possible that the different sources and mechanisms of fouling interact. It is impossible to provide specific information on the likely interaction; each example of mixed fouling is likely to produce different effects. The approach would have to be empirical—based on the individual examples.

All kinds of interactions and combinations are possible. In cooling water, for instance, the deposits are likely to include micro-organisms, particles, scale, and corrosion products. In flue gas systems, the fouling may be because of particle deposition, chemical reaction, and corrosion. In crude oil processing, the deposits may be products of reaction, combined with particle deposition and corrosion.

CONCLUSIONS

Where heat is being transferred across a surface, the accumulation of unwanted deposits is possible. The existence of such a deposit impairs the heat transfer and leads to inefficient operation. Furthermore, because of the restrictions to flow imposed by the

deposit, the energy required for pumping for a given flow rate is increased. Combating the problem of fouling on-line may be a necessity to maintain plant operation and, in addition, there is often a need to clean heat exchangers periodically to restore efficient operation. All these detrimental effects increase operating costs.

Six mechanisms have been identified as being responsible for heat exchanger fouling, but it is seldom that one mechanism is responsible for a particular fouling problem; a combination of mechanisms is much more likely, although one mechanism may be dominant. The variables that have the greatest influence on these mechanisms, apart from the nature of the fluids being processed, are temperature and flow rate.

ARTICLE OF FURTHER INTEREST

Biofilms, p. 111.

REFERENCES

1. Newson, I.H.; Miller, G.A.; Haynes, J.W.; Bott, T.R.; Williamson, R.D. Particulate fouling; studies of deposition, removal and sticking mechanism in a haematite/water system. Proceedings of the Second U.K. National Conference on Heat Transfer, Glasgow, U.K., Sept 14–16, 1988; Int. Mech. Eng., London, U.K., 1988; 137–160.
2. Bott, T.R. *Fouling of Heat Exchangers*; Elsevier: Amsterdam, Holland, 1995.
3. Crittenden, B.C. Basic science and models of reaction fouling. In *Fouling Science and Technology*; Melo, L.F., Bott, T.R., Bernardo, C.A., Eds.; Kluwer Academic Publishers: Dordrecht, Holland, 1988.
4. Watkinson, A.P. *Critical Review of Organic Fluid Fouling*, Final Report, AWL/CNSV-TM-208; Argonne National Laboratory: Argonne, 1988.

Fractal Geometry: Applications

Douglas K. Ludlow
Chemical and Biological Engineering, University of Missouri–Rolla, Rolla, Missouri, U.S.A.

INTRODUCTION

The concept of the non-Euclidian geometry, known as fractal geometry was introduced in the late 1970s. During the 1980s and 1990s there was an explosion of scientific and engineering research and journal publications involving fractals as the concept of nature inspired geometry was tried in numerous applications. The beauty of the underlying mathematics and the ability to describe complex spatial and temporal structures utilizing the concepts of fractal geometry inspired a generation of researchers seeking applications and uses of fractal geometry. Originally, it was thought to be able to fill the need and provide the ability to model complex systems and morphologies and was nearly considered a panacea. Time and maturity of the topic has tempered some of the initial exuberance; however, numerous applications have been found in areas of science, engineering, economics, etc. Because many of the systems important to chemical processing occur at the molecular scale where morphologies are fractal over the length scale range of interest, there have been several important applications of fractal geometry in chemical systems.

FRACTAL GEOMETRY

Benoit Mandelbrot's *The Fractal Geometry of Nature* made the world aware of the underlying universal order that can be found in many spatial and temporal phenomena.[1] Prior to that, such phenomena at best could be referred to as ''complex'' and described in terms such as ''rough,'' ''ramified,'' or ''fragmented.'' The presence of scaling laws had been recognized for a number of individual objects and systems, but the significance of the concepts of fractal geometry was to unify these and to recognize the importance of the underlying symmetry. Many objects and phenomena scale with noninteger exponents. The concept of fractal geometry implies the invariance of details when magnified. This symmetry is essential in understanding the object or phenomena, and helps to describe, characterize, and to measure. Many objects in nature exhibit self-similar features and are of fractal shape. A mathematical fractal object has details at all scales, whatever the magnification, and is invariant under its generating transformation. These objects show both self-similarity and self-affinity. The mathematics of fractal geometry is yet to be fully developed. Until now the use of fractals in applied sciences and engineering has been mostly limited to phenomenological descriptions related to the empirical discovery of power laws. Nearly a decade after Mandelbrot's seminal work, scientists began reporting applications for these concepts. *The Fractal Approach to Heterogeneous Chemistry* described the application of several fractal concepts to chemistry.[2] A few years later, *Fractals in Chemistry, Geochemistry and Biophysics* made further inroads in the application of fractal geometry concepts to chemistry and chemical processes.[3] Half a decade later *Fractals in Chemistry* followed and had the advantage in that the field had matured somewhat, so that critical reviews and opinions could be expressed concerning at least some of the aspects of the use of fractals in chemistry in the chemical processing industry.[4]

During the later part of the 1980s, as Mandelbrot's concepts reached a wider audience, sessions in the American Physical Society, the American Chemical Society, and the American Institute of Chemical Engineers gave many optimistic hopes for the application of the relationship:

$$Property \propto Scale^{\beta}$$

to numerous different phenomena. It almost appeared that many thought the formalism of fractals and the exponential scale β (slope of power–law relationship) was a panacea to reduce many complex situations to one simple explanation, a fractal dimension. However, deeper thought and time have shown that in many cases the exponential scaling factor β is not truly a fractal dimension, and that fractals do not answer well all of the questions initially thought to be accessible through their formalism. *Nature Inspired Chemical Engineering* by Coppens points out not only some of the limitations but also many of the exciting vistas that can occur through the application of the concepts of fractal geometry to applied physical science and engineering.[5] Indeed, fractals are everywhere, but not everything is fractal: A noninteger exponent is not necessarily a fractal dimension. The distinction is not critical if the noninteger scaling exponent is to be

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120025846

used only for phenomenological descriptions, but is crucial in more advanced, deductive modeling and in building new applications.

The key to applying fractal geometry to any field, including chemical engineering, is to understand the underlying assumptions. Some property X vs. some scale δ, may be plotted as $\log(X)$ vs. $\log(\delta)$ and can be fitted to a straight line, so that $X \sim \delta^{\beta}$. Typically, this scaling relationship is only determined over a fairly narrow range of scales, because is it difficult to get experimental data over a wide range of scales. No object or phenomenon in nature is scaling over an infinite range of scales as is the case in a mathematical fractal object. The limit of $\delta \rightarrow 0$, necessary in mathematical theorems involving fractal measures, is impossible to take in practice. In the nonmathematical world, scaling can only be measured within a finite range, and if the range of the measurements is wide enough, the curve of $\log(X)$ vs. $\log(\delta)$ will deviate from a straight line outside of the interval of $[\delta_{\min}, \delta_{\max}]$, the fractal scaling range. When this concept is ignored in applications of fractal geometry the results are misunderstood and even misused. If the experimentally available interval is too narrow, say less than a factor of 10, then there may be a noninteger exponent that might be wrongly considered a fractal dimension when, owing to the limited experimental range, it is unclear whether there is an underlying self-similarity or self-affinity to the phenomena.

Fractal Surfaces[6]

One place where the concepts of fractal geometry have found use is in the quantitative description of surfaces and surface ruggedness. In nature, most surfaces are not smooth except over some limited ranges. In the classical, ''what is the length of the coastline of England?'' problem, it is evident, that the total length of the coastline determined will be dependent on the length of the ruler that you use to measure it. With the ruggedness of the actual coastline, if your ruler length is 10 km, many features that are smaller will be disregarded. If a smaller ruler is used, more of the detail will be followed. It is found that the length of the coastline measured scales to the length of the ruler in a way that can be related to a fractal dimension. This same concept can be applied to surfaces. If a rugged surface area is to be determined by measuring the number of equal sized particles that just cover it (monolayer coverage), then the surface area determined will depend on the size of the particles used as the yardstick. Thus, the fractal dimension of a surface can be estimated by determination of the number of objects (molecules) of a given size that would completely cover the surface for a series of different sized objects. This concept has been used in several different techniques to determine the fractal dimension of a surface, such as gas adsorption using a homologous series of adsorbates, liquid adsorption of a series of different sized polymers or latex beads, etc.

The surface property of a solid is characterized by the nature of the surface boundary. The surface boundary is expected to be related to the underlying geometric nature of the surface, hence its fractal dimension. Many properties of the solid depend on the scaling behavior of the entire solid and of the pore space. The distribution of mass in the porous solid and the distribution of pore space may also reflect the fractal nature of the surface. If the mass and the surface scale are alike, that is, have the same power–law relationship between the radius of a particle and its mass, then the system is referred to as a mass fractal. In a similar manner if the pore volume of porous material has the same power–law relationship between the pore volume and radius as that of the surface, then it is described as a pore fractal.

Powdered and porous materials can also be characterized from additional techniques that do not depend on probing the surface with a molecule. Image analysis techniques (but still using the concept of varying the length of the ruler) have been developed. Techniques have also been developed to determine a surface fractal dimension from scattering experiments using neutrons, visible light, and x-rays. Although no solid surface is fractal over all scales in a mathematical sense, many surfaces in the range where molecules interact (single-surface molecules up through macropore networks in the particles) can be described and characterized with a fractal dimension. This has been one of the useful applications of fractal theory, being able to quantitatively describe surface/interface ruggedness by the fractal dimension. Prior to fractal analysis, surfaces could be described as being more rugged, or more ramified, or more pitted than other surfaces. Using fractal analysis, the fractal dimension of the surface can be measured and quantified.

Aggregates

Another area where the application of fractal geometry concepts has provided insight into physical phenomena is in the characterization and understanding of aggregation and growth. Aggregates are formed from smaller, nearly identical particles. Because mechanisms for formation contain random elements, we can expect three main structural phenomena arising in aggregated objects. First, mass fractality, because the process of formation often leads to more or less diaphanous aggregates. Second, fractal pore domains consisting of interstitial void elements of a wider range of linear

size. Third, surface fractality from the interfaces between pore and mass regions.

Fractal geometry has been found to be very useful in describing aggregates.[2–4,7] For several decades it has been known that diffusion and randomness play an important part of aggregate growth. In the simplest form diffusion-limited aggregation (DLA), a randomly drifting/diffusing particle is in a system with a preexisting nucleus/aggregate. When the diffusing particle hits the nucleus/aggregate, then there is an instantaneous and irreversible attachment. As more and more particles stick to the growing nucleus a branching, tree-like structure is formed. For a mass–distance scaling relationship, the number of particles, N, within a given radius, R, scales by:

$$N \sim R^D$$

Image analysis of two-dimensional representations of aggregates, such as planar projections can be used to determine a fractal dimension using the technique above. Numerous computational and theoretical studies of DLA clusters generated on computers have found that in DLA structures in which there is a high sticking probability, the fractal dimension, D, of a DLA cluster formed in a two-dimensional lattice is close to 1.7 and for a three-dimensional lattice, it is close to 2.5. As the sticking probability is decreased and as relaxation of the structures is allowed, the numerical models indicate for the two-dimensional lattice systems that the scaling relationship will approach 2.

Many experimental aggregate structures show this branch-like nature. Besides being formed in cases of diffusion limited growth, there will also be varying kinetics of aggregation, which can be classified into two kinds, slow and fast aggregation, each with a different rate-limiting physics. Systems that demonstrate either diffusion limited and/or reaction limited aggregation form fractal particles. Aggregation of solid particles or monomers in the liquid or in the gas phase far from saturation forms mass fractals. Of course if there is some form of relaxation or equilibrium of the structures formed the fractal nature of the particles will change. Aggregation conditions closer to saturation lead to denser structures, with a smooth or fractally rough surface. In the latter case these are surface fractals. It has been demonstrated that when the planar projections of numerous different aggregate particles are analyzed using the scaling relationship, the dimension determined is in the range of $D \sim 1.77$. If the same types of particles are formed under conditions where the kinetics of aggregation are slow, then more compact particles are formed and the planar projections give a scaling close to 2. Hence, the mass scaling relationship with the size of aggregates gives one indication of the kinetics of formation. There are several books dedicated to the study and aspects of fractal growth of aggregates and this has been one area where the concepts of fractal geometry have found many applications.

Techniques have been developed to determine the fractal dimension of experimental aggregate particles in solution using small-angle scattering techniques, from x-ray, light, or even neutron sources. In these techniques the scattering intensity, $I(q)$, is proportional to the scattering vector, q, raised to the mass fractal dimension by:

$$I(q) \sim q^{-D_m}$$

This technique has been successfully applied in numerous studies to determine the fractal dimension of aggregate particles.

Closely related to the aggregates formed in colloidal aggregation in gels and sols is the deposition of aggregate particles during electrodeposition. The complex structures formed from electrodeposition have been analyzed using fractal methods. There have been several studies in which dentritic structures are formed that are statistically simple and self-similar, i.e., fractal in nature. These are typically formed under conditions of diffusion-limited growth. The growth of deposits on the cathodes of batteries is often the main factor limiting the lifetime of the batteries and other electronic parts. There have been studies that show that the fractal nature of the electrodes and/or deposits on the electrodes affects the efficiency of the battery.

Dissolutions and Etchings

The chemical reaction between a solid and a reactive fluid is of interest in many areas of chemical engineering.[8] The kinetics of the phenomenon is dependent on two factors, namely, the diffusion rate of the reactants toward the solid/fluid interface and the heterogenous reaction rate at the interface. Reactions can also take place within particles, which have accessible porosity. The behavior will depend on the relative importance of the reaction outside and inside the particle. Fractal analysis has been applied to several cases of dissolution and etching in such natural occurring caves, petroleum reservoirs, corrosion, and fractures. In these cases fractal theory has found usefulness for quantifying the shape (line or surface) with only a few parameters: the fractal dimension and the cutoffs. There have been some attempts to use a fractal dimension for reactivity as a global parameter. Finally, fractal concepts have been used to aid in the interpretation of experimental results, if patterns quantitatively similar to DLA are obtained.

Diffusion and Reactions

One area that has seen much interest by chemists and chemical engineers is the study of diffusion and reaction on fractals. Many of the heterogenous reaction systems contain structures that have a fractal-like nature. Several numerical studies of diffusion and transport on fractals have been completed with the aim of being able to describe the effect of fractal structure on flow and diffusion of species. Utilizing concepts of fractal lattices as models for more realistic porous media, several numerical studies have been carried out to characterize the flow and diffusion through porous media. Numerical studies of the diffusion and reaction of reactant systems occurring in a fractal space vs. a traditional Euclidean space has led to predictions of anomalous diffusion and reaction kinetics. The application of fractal concepts has been useful in modeling permeation and percolation through porous structures.

Nature Inspired Chemical Engineering

One of the positive by-products of the application of fractal theory to various different physical phenomena has been the realization that many structures found in nature demonstrate a fractal nature over a limited range and more importantly that these self-similar and self-affine structures have a natural optimization. For instance a tree has to control the essential processes at the scale of the elementary microunits such as the cells in the leaves and this is realized in the specific structure of the veins within the leaves. Yet, the tree can only function well if the nutrients can access the leaves easily, thus there is a different fractal structure of the trunk and branches. We find similar behavior in lungs and kidneys. Through the fractal structure of the tree branches, there is a fast and uniform transport to the microunits or cells, covering a high surface area or volume, from one or a few points (the stem). These hierarchical structures are found throughout nature because they lead to easy transport of molecules and energy and also give mechanical strength and flexibility while making efficient use of available materials. Researchers have successfully applied such concepts to the design of fluid injectors for fluidized bed with the injector having a fractal shape. It was found that the fractal, nature inspired, injectors gave more uniform fluid flow in the solids with greater mixing and contact.

CONCLUSIONS

In the mid-1980s, there was an explosion of research seeking applications for fractal theory to various aspects of physical phenomenon in physics, chemistry, and engineering. The sessions at the professional conference that dealt with fractals would leave one almost breathless with the vast variety of topics and experimental results that were being analyzed in terms of fractal concepts. It seemed that everywhere we looked we found fractals and that many phenomena were scaled by a power–law relationship. It has been said that indeed, exuberance sort of ran away from sober reflection, and that perhaps the scaling exponent was taken too frequently for a fractal dimension without deeper investigation.[4] Perhaps, the concept of fractals was just filling the human need to be able to explain complex and seemingly intractable systems in a simple manner. Time has found that fractals are neither simple nor have they answered all of the questions initially thought to be accessible through their formalism. Nevertheless, a fractal approach to the modeling of many phenomena has been successful and helpful in our understanding of these phenomena. At the same time, and despite many theoretical highlights, real technological applications of fractals are still strikingly scarce for such a fundamental, universal concept. Many natural objects and phenomena, such as trees and turbulence, are scaling within a finite scaling range. Fractals become useful when recognizing this finite range, and measuring the scaling range by correctly interpreting the measurement results, remembering that too narrow of a range may lead to a bias on the value of the dimension determined. Fractal geometry, like Euclidean geometry, is useful within the range it applies to. Many important physical, chemical, biological, geological, economical, and other phenomena occur exactly within the range where the underlying structure is geometrically scaling and hereby fractal.

REFERENCES

1. Mandelbrot, B.B. *The Fractal Geometry of Nature*; W.H. Freeman & Company: New York, 1977.
2. Avnir, D. Ed. *The Fractal Approach to Heterogeneous Chemistry, Surfaces, Colloids, Polymers*; John Wiley & Sons: New York, 1989.
3. Birdi, K.S. *Fractals in Chemistry, Geochemistry, and Biophysics: An Introduction*; Plenum Press: New York, 1993.
4. Rothschild, W.G. *Fractals in Chemistry*; John Wiley & Sons: New York, 1998.
5. Coppens, M.-O. *Nature Inspired Chemical Engineering*; Delft University Press: Delft, Netherlands, 2003.
6. Russ, J.C. *Fractal Surfaces*; Plenum Press: New York, 1994.
7. Vicsek, T. *Fractal Growth Phenomena*; 2nd Ed.; World Scientific: Singapore, 1992.
8. Giona, M.; Biardi, G. *Fractals and Chaos in Chemical Engineering*; World Scientific: Singapore, 1997.

Free-Radical Polymerization

Yadunandan Lal Dar
Corporate Research, National Starch and Chemical Company, Bridgewater, New Jersey, U.S.A.

Rajeev Farwaha
Celanese Polymers, Bridgewater, New Jersey, U.S.A.

Gerard T. Caneba
Department of Chemical Engineering, Michigan Technological University, Houghton, Michigan, U.S.A.

INTRODUCTION

Free-radical polymerization is an important industrial process to manufacture synthetic high polymers. It is a chain polymerization process utilizing free-radical reactions for chain growth. This process utilizes monomers with carbon–carbon double bonds that are susceptible to free-radical attack and cleavage. The attack of a free radical leads to the cleavage of a double bond with the monomer unit being attached to the previous radical at one of the cleavage sites and the radical being stabilized at the other cleavage site.

Many industrial free-radical polymerization processes have been developed and commercialized for a variety of monomer and polymer types. A variety of monomers and polymerization processes have been commercially exploited. Further, there have been significant developments in synthesizing polymers with controlled architectures using free-radical polymerization during the last two decades. Materials synthesized include block copolymers, graft copolymers, and radial polymers. These materials find use in many common industrial and household applications such as adhesives, paints and coatings, textiles, nonwoven fabrics, personal care products, wallpaper, construction materials, specialty additives, and many other areas.

This entry on free-radical polymerization discusses relevant aspects from a chemical processing perspective. It briefly discusses the chemistry and mechanism, the types of reactors and processes typically used in the industry, and other relevant reaction engineering and processing aspects. This should provide an introduction to the process with more detailed information included in the references herein.

MECHANISM FOR FREE-RADICAL POLYMERIZATION

The mechanism of free-radical polymerization has been studied extensively.[1,2] Most monomers that have a double bond can undergo free-radical polymerization (with certain notable exceptions including vinyl ethers). The double bond is attacked by a free radical, either on the end of a polymer chain or newly formed by an initiating species. Four main types of reactions occur in this system: chain initiation, propagation, termination, and transfer reactions. Each of these steps has reaction rate constants (K_{reaction}) that follow the Arrhenius relationship as shown in Eq. (1):

$$K_{\text{reaction}} = A\text{e}^{-(E_a/RT)} \tag{1}$$

where E_a is the activation energy for each reaction considered. The subscript ''a'' and the subscript for K_{reaction} may be replaced by ''d'', ''p'', ''t'', or ''tr'' representing decomposition (for initiator), propagation, termination, or transfer, respectively. T is the absolute temperature, R is the universal gas constant, and A is a constant.

Initiation

Free-radical polymerization is typically initiated by a compound, called an initiator or catalyst, that undergoes chemical change to yield free radicals. Many types of initiators have been used depending on the polymerization process. Initiation mechanisms include free-radical forming agents activated by temperature, ultraviolet radiation, gamma radiation, redox reactions, ultrasonic energy, or other high-energy producing techniques.

A thermal initiator is typically either an azo compound or a peroxide based initiator, e.g., 2,2′-azobis-isobutyronitrile. Each initiator molecule can decompose to give one or two primary radicals ($R^{\bullet}$) and possibly inert by-products. Initiation rate constants for several common thermal initiators are listed in Table 1. Each primary radical can attack a monomer molecule (M) to initiate a polymer chain ($P_1^{\bullet}$). The recombination

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007776

Table 1 Rate constants for common thermal initiators in various solvents

Initiator	Solvent	Temperature (°C)	K_d (sec^{-1})	E_d (kcal/mol)
2,2′-Azo-bis-isobutyronitrile (AIBN)	Benzene	40	5.4×10^7	30.7
Phenyl-azo-triphenylmethane	Benzene	25	4.3×10^7	26.8
tert-Butyl peroxide	Benzene	80	7.8×10^6	34
Cumyl peroxide	Benzene	115	1.6×10^5	40.7
Acetyl peroxide	Benzene	35	9.5×10^5	32.3
Benzoyl peroxide	Benzene	30	4.8×10^8	27.8
Lauryl peroxide	Benzene	30	2.6×10^7	30.4
tert-Butyl hydroperoxide	Benzene	154	4.3×10^6	40.8
tert-Butyl perbenzoate	Benzene	100	1.1×10^5	34.7

(From Ref.[3].)

of initiator molecules and the resulting inefficiency are the result of the so-called cage effect. An initiation efficiency, f, is used to account for this loss of free radicals.

$$I \rightarrow R^{\bullet}$$

$$R^{\bullet} + M \rightarrow P_1^{\bullet}$$

$$2R^{\bullet} \rightarrow I' \quad \text{(recombined inactive initiator molecule)}$$

Redox reactions can be used to initiate free radicals as well. Typical examples include persulfate/bisulfite systems. Redox initiation can be used at lower temperatures than thermal initiation reactions. This is useful for polymerization systems that are unstable or show undesirable side reactions at higher temperatures.

Propagation

Each primary polymer radical ($P_1^{\bullet}$) can add a monomer molecule to produce a dimer ($P_2^{\bullet}$) and then a trimer ($P_3^{\bullet}$), and so on. This series of successive additions of monomer molecules to a polymer chain is called chain propagation. The propagation reaction step is very fast and almost as soon as a radical is initiated it can lead to the formation of very high molecular weight polymer.

$$P_n^{\bullet} + M \rightarrow P_{n+1}^{\bullet}$$

Termination

The loss of the propagating free radical at the end of a polymer chain ends chain growth. This is called chain termination. Polymer chains can mutually end their propagation in two ways. Termination by combination occurs when two polymer radicals react with each other to produce one polymer chain. Termination by disproportionation is the result of the abstraction of an atom, typically hydrogen, from one chain by the other. This results in one chain with a double bond at the end and another with a saturated carbon atom.

$$P_n^{\bullet} + P_m^{\bullet} \rightarrow D_n + D_m \quad \text{(Disproportionation)}$$

$$P_n^{\bullet} + P_m^{\bullet} \rightarrow D_{n+m} \quad \text{(Combination)}$$

Termination can also occur by the reaction of polymer free radicals with primary initiator radicals (called primary termination) or free-radical scavenging species, especially oxygen. Activation energies for propagation and termination for some typical monomers are listed in Table 2.

Chain Transfer

A polymer radical can take part in a chain transfer reaction, which results in a terminated polymer chain and a radical derived from the molecule that accepted the transfer. Chain transfer can take place to monomer, solvent, or other polymer molecules. It could also take place to impurities in the system. The general transfer reaction can be represented as

$$P_n^{\bullet} + CT \rightarrow D_n + CT^{\bullet}$$

Table 2 Energies of activation for propagation (E_p) and termination (E_t) in free-radical polymerization

Monomer	E_p (kcal/mol)	E_t (kcal/mol)
Methyl acrylate	7.1	5.3
Acrylonitrile	4.1	5.4
Butadiene	9.3	—
Ethylene	8.2	—
Methyl methacrylate	6.3	2.8
Styrene	7.8	2.4
Vinyl acetate	7.3	5.2
Vinyl chloride	3.7	4.2

(From Ref.[4].)

where CT is the chain transfer agent, which could be either of the agents mentioned above. The rate of chain transfer is determined by the chain transfer constant C_{tr}, which is defined as the ratio of the rate of chain transfer to the rate of propagation for two different molecules (Table 3).

Inhibition

Some chemicals retard or suppress free-radical polymerization by reacting with primary radicals or macroradicals to yield radicals that are very stable to further reaction or yield nonradical products. These materials could be retarders or inhibitors. Retarders slow down the formation of polymer but inhibitors completely eliminate it. Oxygen is one of the most commonly known inhibitors for vinyl polymerization and good practice requires the removal of air from the reactor vessels before the reaction is started. It combines with active radicals to form unreactive peroxy radicals.

Commercial monomers are sold with added inhibitor to prevent their premature polymerization during storage and transportation. These inhibitors are either removed or extra initiator is added to reactor vessels to reactively consume the inhibitors. Retarders and inhibitors differ only in the frequency with which propagating radicals react with them rather than monomer and also in the ability of these resulting macroradicals to reinitiate. Fig. 1 schematically illustrates the difference in polymerization kinetics in the presence of retarders and inhibitors.

Copolymerization

Copolymerization of two or more monomers during free-radical polymerization is an effective way of altering the properties of polymers. The discussion here is based on the copolymerization of two monomers, but it is possible to extend it to multiple monomers as well. If the reactive species are M_1 and M_2, the composition of the resulting polymer is not a mixture or blend of homopolymers of M_1 and M_2, but a statistical copolymer of both the monomers. Commercial polymer products can be copolymers of several monomers that provide a unique property or functionality to the product. The reactivity of monomers varies with the type and concentration of other species in the reaction mixture feed.

Theoretical studies of the polymerization of two or more monomers have been conducted on various reaction systems since Dostal first proposed the concept of the terminal model in 1936.[5] In the terminal model, it is assumed that the reactivity of the growing polymer chain is determined merely by the last added monomer unit (i.e., terminal unit), independently of the chain length and composition. For a two-component (M_1 and M_2) copolymerization, the terminal model leads to the following four propagation reaction equations:

$$\sim\sim\sim M_1^{\bullet} + M_1 \xrightarrow{k_{11}} \sim\sim\sim M_1^{\bullet}$$

$$\sim\sim\sim M_1^{\bullet} + M_2 \xrightarrow{k_{12}} \sim\sim\sim M_2^{\bullet}$$

$$\sim\sim\sim M_2^{\bullet} + M_1 \xrightarrow{k_{21}} \sim\sim\sim M_1^{\bullet}$$

$$\sim\sim\sim M_2^{\bullet} + M_2 \xrightarrow{k_{22}} \sim\sim\sim M_2^{\bullet}$$

in which $\sim\sim\sim M_1^{\bullet}$ and $\sim\sim\sim M_2^{\bullet}$ are the growing chains with M_1 and M_2 as the terminal units, respectively, and k_{11}, k_{12}, k_{21}, k_{22} are the reaction constants. Based on the propagation equations, the consumption

Table 3 Chain transfer constants (C_{tr}) of solvents to styrene in free-radical chain polymerization at 60°C

Transfer agent	$C_{tr} \times 10^4$	Transfer agent	$C_{tr} \times 10^4$
Acetic acid	2.0	1-Dodecanethiol	148,000
Benzene	0.01	Hexane	0.9
Butyl alcohol	0.06	*N,N*-dimethylaniline	12
tert-Butyl alcohol	6.7	1-Naphthalenethiol	1,500
Butyl disulfide	0.24	1-Octanethiol	190,000
Carbon tetrabromide	18,000	*p*-Methoxyphenol	260
Carbon tetrachloride	84	Phenol	8.1
Chloroform	0.5	Triethylamine	1.4
O-Chlorophenol	6.0	Toluene	0.105
2,6-Ditertiary-butyl phenol	49	Water	0

(From Ref.[4].)

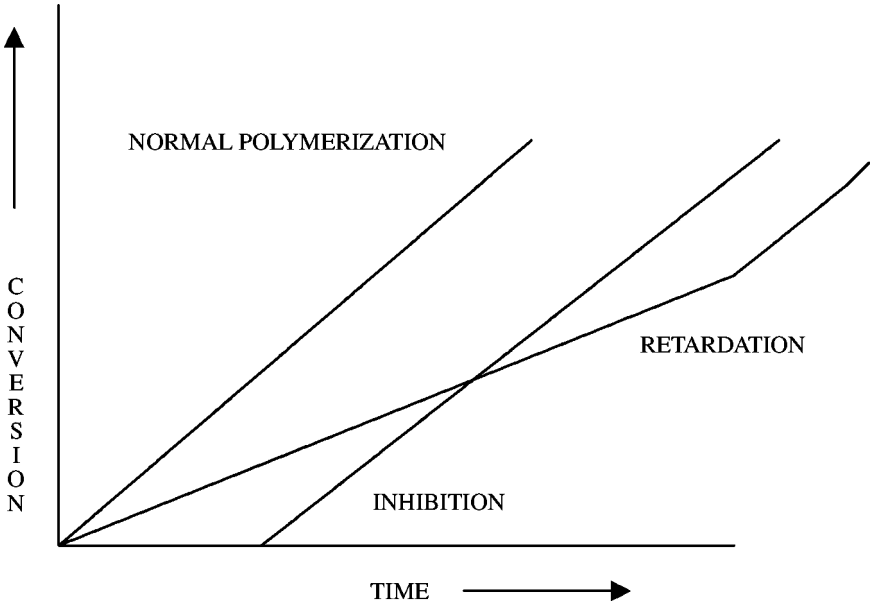


Fig. 1 Conversion vs. time for normal polymerization as well as inhibited and retarded free-radical polymerizations. (From Ref.[3].)

rates of M_1 and M_2 can be given by

$$-\frac{d[M_1]}{dt} = k_{11}[\sim\sim\sim M_1^*][M_1] + k_{21}[\sim\sim\sim M_2^*][M_1] \quad (2)$$

$$-\frac{d[M_2]}{dt} = k_{12}[\sim\sim\sim M_1^*][M_2] + k_{22}[\sim\sim\sim M_2^*][M_2] \quad (3)$$

It should be noted that the monomer consumption in the initiation step is neglected in the above equations owing to the small amount of initiator used in most systems. Furthermore, Mayo and Walling suggested that radical concentrations (i.e., $\sim\sim\sim M_1^*$ and $\sim\sim\sim M_2^*$) should remain constant at steady state because of the relatively short lifetime of the radicals.[6] This means that the rates of the conversion between $\sim\sim\sim M_1^*$ and $\sim\sim\sim M_2^*$ should be equal, which is given by the following equation:

$$k_{21}[\sim\sim\sim M_2^*][M_1] = k_{12}[\sim\sim\sim M_1^*][M_2] \quad (4)$$

Rearrangement of Eqs. (2)–(4) to eliminate the terms of radical concentrations yields[7]

$$\frac{d[M_1]}{d[M_2]} = \frac{[M_1](r_1[M_1] + [M_2])}{[M_2]([M_1] + r_2[M_2])} \quad (5)$$

where

$$r_1 = \frac{k_{11}}{k_{12}}, \quad r_2 = \frac{k_{22}}{k_{21}}$$

r_1 and r_2 are known as monomer reactivity ratios, which are the ratios of the reaction rates of homo-polymerization to cross-polymerization for individual radicals.

If we let

$$f_1 = 1 - f_2 = \frac{[M_1]}{[M_1] + [M_2]}$$

$$F_1 = 1 - F_2 = \frac{d[M_1]}{d[M_1] + d[M_2]}$$

in which f_i is the molar fraction of M_i in the reaction medium, and F_i is the corresponding molar fraction of M_i unit in the instantaneously formed copolymer, where i = 1 or 2. Substitution of f_1 and F_1 in Eq. (5) gives the copolymer composition equation

$$F_1 = \frac{r_1 f_1^2 + f_1 f_2}{r_1 f_1^2 + 2 f_1 f_2 + r_2 f_2^2} \quad (6)$$

This is the fundamental equation, upon which various methods were suggested to determine the monomer reactivity ratios of M_1 and M_2. These methods will be discussed in a later section.

Various methods have been used to determine monomer reactivity ratios for various copolymerization systems. According to Eq. (6), a series of monomer molar fractions (f_1) and their corresponding instantaneous copolymer compositions (F_1) are required to obtain monomer reactivity ratios. The most common experimental procedure is to carry out a set of copolymerizations with the feed ratios of M_1 to M_2 ranging from 0.1 to 0.9, and the conversions are kept below 5%. It is noted that owing to the relatively low conversion, the instantaneous copolymer composition can be approximated by the composition of the final copolymer

measured by IR, UV, or NMR. Even though this treatment is defective, errors in calculated reactivity ratios are tolerable enough for such a simple method.

In 1944, Mayo and Lewis proposed a method called ''intersection method,'' based on a rearrangement of Eq. (5):[7]

$$r_2 = \frac{F^2}{f} r_1 + F\left(\frac{1}{f} - 1\right) \tag{7}$$

where

$$F = \frac{d[M_1]}{d[M_2]}, \quad f = \frac{M_1}{M_2}$$

For each pair values of F and f, plotting r_2 against r_1 should generate a straight line. In an ideal case, all the lines should intersect at one specific point, which represents the monomer reactivity ratios. However, the inevitable experimental errors could lead to an intersection area instead of a point. Consequently, which point in the area should be chosen would be a problem.

A useful method for calculating reactivity ratio values for free-radical polymerizations called the Q–e scheme was proposed by Alfery and Price in 1947. For a system of two monomers with reactivity ratios r_1 and r_2 as exhibited in Table 4, Q_1 and Q_2 are related to reactivity and e_1 and e_2 are related to the polarity of monomers M_1 and M_2, respectively. Styrene with Q and e values of 1 and -0.8 is used as a comparative standard. Higher Q values indicate greater resonance stability, and higher e values (less negative) indicate greater electron withdrawing power of the alpha substituent on the vinyl monomer.

$$k_{11} = P_1 Q_1 e^{-e_1^2} \tag{8}$$

$$k_{12} = P_1 Q_2 e^{-e_1 e_2} \tag{9}$$

$$r_1 = \frac{k_{11}}{k_{12}} = \frac{Q_1}{Q_2} e^{-e_1(e_1 - e_2)} \tag{10}$$

$$k_{22} = P_2 Q_2 e^{-e_2^2} \tag{11}$$

$$k_{21} = P_2 Q_1 e^{-e_1 e_2} \tag{12}$$

$$r_2 = \frac{k_{22}}{k_{21}} = \frac{Q_2}{Q_1} e^{-e_2(e_2 - e_1)} \tag{13}$$

$$r_1 r_2 = e^{-(e_1 - e_2)^2} \tag{14}$$

In 1950, Fineman and Ross suggested another method called ''FR method'' or ''linear least square method,'' which is based on Eq. (7).[8] For a set of values of F and f from experiments, a plot of $(1 - f)/F$ vs. f/F^2 should be a straight line with the slope r_2 and the intercept $-r_1$. Similarly, owing to experimental errors, all points may not sit on one line but be scattered instead. To obtain the optimal values for r_2 and r_1, the linear least square method should be used.

It is noted that for both methods, the instantaneous copolymer composition, F, is approximated by the composition of the final copolymer, and the actual drift in the monomer feed ratio, f, is neglected because of the low conversions for the experiments. Therefore, the reactivity ratios obtained from both methods might be inaccurate for high conversion.

To minimize the effect of the drift in the monomer feed ratio, Kelen and Tüdõs proposed a new method.[9] In their work, the following equation was introduced, based on Eq. (5):

$$\frac{G}{\alpha + H} = \left(r_1 + \frac{r_2}{\alpha}\right)\frac{H}{\alpha + H} - \frac{r_2}{\alpha} \tag{15}$$

where

$$x = \left(\frac{[M_1]_f}{[M_2]_f} + \frac{[M_1]_0}{[M_2]_0}\right)\Big/2, \quad y = \frac{dM_1}{dM_2},$$

$$G = \frac{x(y - 1)}{y}, \quad H = \frac{x^2}{y}$$

$[M_1]_f$ and $[M_1]_0$ are the final and initial molar concentrations of M_1, respectively. If we let

$$\eta = \frac{G}{\alpha + H} \quad \text{and} \quad \xi = \frac{H}{\alpha + H}$$

Eq. (15) can be written as

$$\eta = \left(r_1 + \frac{r_2}{\alpha}\right)\xi - \frac{r_2}{\alpha} \tag{16}$$

where α is an arbitrary constant to make the variables η and ξ lie in the interval (0, 1). Kelen and Tüdõs suggested the following optimal value for α:

$$\alpha = \sqrt{H_m H_M} \tag{17}$$

where H_m and H_M are the smallest and biggest values of H, respectively. Plotting η vs. ξ should give a straight line, from which r_1 and r_2 can be obtained. Additionally, the following equations are given for r_1 and r_2, using the least squares method:

$$r_1 = \frac{\sum \eta\xi(n - \sum \xi) - \sum \eta\left(\sum \xi - \sum \xi^2\right)}{n\sum \xi^2 - \left(\sum \xi\right)^2}$$

$$r_2 = \frac{\sum \eta\xi \sum \xi - \sum \eta \sum \xi^2}{n\sum \xi^2 - \left(\sum \xi\right)^2} \tag{18}$$

where n is the number of a set of experimental data.

Table 4 Typical free-radical chain-copolymerization reactivity ratios

M_1	M_2	r_1	r_2	r_1r_2
Vinyl acetate	Vinyl chloride	0.23	1.68	0.39
	Vinyl laurate	1.4	0.7	0.98
Styrene	Vinyl acetate	55	0.01	0.55
	Vinyl chloride	17	0.02	0.34
Methyl acrylate	Styrene	0.18	0.75	0.14
	Vinyl acetate	9	0.1	0.90
Butadiene	Styrene	1.39	0.78	1.08

(From Ref.[5].)

To overcome the errors brought in by the approximation for instantaneous copolymer composition and the drift in monomer feed ratio, the integration equation for Eq. (5) was used to determine the monomer reactivity ratios, which was given as[10]

$$\ln x_2 = r_2 \ln x_1 + \frac{1 - r_1 r_2}{1 - r_1} \times \ln\left[\frac{(1 - r_2)x_2 - (1 - r_1)x_2 x}{(1 - r_2)x_2 - (1 - r_1)x_1 x}\right] \qquad (19)$$

in which

$$x_i = \frac{[M_i]}{[M_i]_0}, \quad x = \frac{[M_1]_0}{[M_2]_0}$$

where $[M_1]_0$ is the initial molar concentration of M_1.

Besides the above methods, Tosi suggested a new method called ''method of grouping,'' which was claimed to have minimal computation difficulty, but only provides approximate values.[11] Tidwell and Mortimer proposed a nonlinear least square method by minimizing the difference between the observed and calculated copolymer compositions.[12] This method was claimed to circumvent the subjective judging of experimental data and lead to better results compared to other methods, although its computation process is quite complicated.

The methods described here as well as other methods proposed can typically be divided into two groups: one is based on the differential composition equation [Eq. (5)], while the other employs the integration composition equation [Eq. (19)]. The latter is preferred because it is free of the limitation on conversion and can produce more accurate values. However, the former methods based on the differential composition equation were widely used because of their simplicity and tolerable errors.

The values of the calculated monomer reactivity ratios can be affected by various factors, such as the estimation method employed, the accuracy of the experimental data, and even the reaction medium (solvent).[13] Reactivity ratios for some common monomers are listed in Table 4.

Controlled Free-Radical Polymerization

Free-radical polymerization is an inherently statistical process with the probability of the reactions being discussed in the earlier sections being dictated by their rate coefficients. However, during the last 10 yr, a lot of effort has been directed toward controlling the statistical nature of this process.[14]

A number of approaches have been proposed that reduce the probability of termination by adding reversible termination agents, highly efficient chain transfer agents, or by reducing the diffusional mobility of polymers in poor solvents. The processes developed in recent years can be used to control polymer microstructure and composition with high yield, efficiency, and specificity. The ability to do this has led to a variety of novel materials, properties, and applications.

Controlled free-radical polymerization (CFRP) has been used successfully to produce block, graft, and other controlled architecture copolymers within the last decade for a variety of free radically polymerizable monomers. The main techniques include reversible addition fragmentation and transfer (RAFT) polymerization, stable free-radical polymerization (SFRP) mediated by nitroxide/alkoxyamine based radicals, atom transfer radical polymerization (ATRP), diphenyl ethylene (DPE) mediated polymerization, and novel precipitation/emulsion polymerization based methods like free-radical retrograde precipitation polymerization (FRRPP).[15]

Nitroxide mediated SFRP, DPE mediated polymerization, ATRP, RAFT polymerization, etc. achieve polymerization control through the use of kinetic mediators or transfer agents, which protect a propagating free radical from undesirable transfer and termination reactions. The emulsion block copolymer method is unique in that it does not require the use of any chemical mediators to achieve this control. Polymerization control is achieved by physically trapping radicals by

precipitation. This minimizes the reaction of the radicals with each other by severely restricting their mobility in the reaction medium.

The ability of these methods in delivering block copolymer structures has been well demonstrated. The ATRP, RAFT, and SFRP methods could all be used to make diblock and triblock copolymers, as well as radial polymers using multiarm initiators. Because these methods are based on free-radical polymerization, they give access to a wider variety of monomer systems than are currently available through non-free-radical polymerization based techniques. They can also lead to controlled polymerization under more industrially practicable conditions as compared to ionic polymerization.

Each of these techniques has unique capabilities and drawbacks that make it extremely useful for certain types of monomers, but not that effective for others. Table 5 shows a summary of the available technology options.

COMMON TYPES OF POLYMERIZATION PROCESSES[16,17]

Radical polymerization can be carried out under homogenous as well as heterogenous conditions. This difference is classified based on whether the initial mixture and/or final product are homogenous or heterogenous. Some homogenous mixtures become heterogenous as polymerization proceeds as a result of insolubility of the resulting polymer in the reaction media. There are many other specialized processes that are used to synthesize materials via free-radical polymerization. These include interfacial polymerization, gas phase reactions (''popcorn polymerization''), as well as the use of specialized media like supercritical fluids.[18] Current research efforts include the study of such processes to understand and control them. Table 6 summarizes and compares some common processes.

Table 5 Summary of CFRP

Name	Technology
ATRP	Transition metal mediators
RAFT	Thioesters, thiophosphates, xanthates, and other chain transfer agent mediators
SFRP	Nitroxides/alkoxyamines, functional nitroxides, imidazoline, piperizinones, morpholones, and related mediators
DPE	Diphenyl ethylene mediator
FRRPP	Physical immobilization of free radicals to reinitiate polymerization
Others	Telomerization, macromers by chain transfer, bimetallic chain transfer, electron donors, alkyl iodide mediators

Bulk Polymerization

Many polymers of industrial importance are made by bulk polymerization. This is the simplest of all polymerization processes based on the ingredients used. Typical ingredients are just monomer and initiator. The polymer formed leads to great changes in viscosity and volume in the reaction vessel. For this reason, most bulk polymerizations are either not carried out to complete conversion or are conducted in specialized equipment like polymer extruders.

Solution Polymerization

Solution polymerization of vinyl monomers is carried out in large agitated reactors in organic solvents. Both propagation and termination rates are impacted by the nature of organic solvent used. The rate of initiator decomposition is independent of the solvent used. The auto acceleration impact (Trommsdorff effect) is less pronounced in solution polymerization than in bulk or suspension polymerization because of dilution and lower viscosity of the continuous phase. To prevent a runaway reaction, the reactants are added gradually to the reactor. The polymer molecular weight is controlled through the proper choice of chain transfer agent, and by initiator concentration and amount. The other contributors are solvent type, monomer type, and reaction temperature.

A typical recipe is shown in Table 7. The solution polymerization is carried out at 140°C by adding the monomer and initiator mixture uniformly for over 4 hr. After the addition of initiator, the reaction is continued over 2 hr.

Emulsion Polymerization

Emulsion polymerization is a heterogenous reaction process in which unsaturated monomers are dispersed in continuous phase with the aid of emulsifiers and polymerized by free-radical initiators. The resulting product is a dispersion of polymer particles, typically smaller than 1 μm in size, in water and is referred to as polymer latex.

The emulsion polymerization process has various advantages as compared to bulk or solution polymerization as the reaction proceeds at low viscosity. The low viscosity during the process allows adequate removal of heat of polymerization and the production of higher solid latexes in combination with high monomer conversion and short cycle time. This process is

Table 6 Comparison of different types of free-radical polymerization processes

Type	Advantages	Disadvantages
Bulk (batch)	Simpler equipment No diluent impurities	May require solution and subsequent precipitation for purification and/or fabrication May require reduction to usable particle size Broad molecular weight distribution Inefficient heat and mass transport at higher conversion
Bulk (continuous)	Easier to control heat and mass transport Narrower molecular weight distribution	Reactant recycling May require solution and subsequent precipitation for purification and/or fabrication May require reduction to usable particle size Complex equipment
Solution	Easier to control heat and mass transport Wide range of accessible molecular weights Easier to transport reagents and products	Needs agitation Solvent removal and recycling Chain transfer to solvent may lead to undesirable effects Inefficient heat and mass transport at high conversions and/or concentrations
Emulsion	Easy heat control, easy agitation, latex may be directly usable, high polymerization rates, molecular weight control, usable small particle size,usable in producing tacky soft solid products, access to higher molecular weight ranges	Polymer may require additional cleanup and purification Difficult to eliminate entrenched coagulants, emulsifiers, surfactants, etc. Often requires rapid agitation at different stages
Precipitation/ dispersion	Molecular weight and molecular weight distribution control via environment, ease of isolation	May require solution and subsequent precipitation for purification and/or fabrication Precipitation may limit molecular weight
Suspension	Easy agitation, higher-purity product when compared to emulsion	Sensitive to agitation Particle size difficult to control Difficult to eliminate residual reagents

(From Ref.[2].)

applied on an industrial scale to produce a variety of products such as paints, paper coatings, adhesives, carpet and textile coating binders, wall paper ground coat, nonwoven binders, glass fiber sizing, personal care, pressure sensitive adhesives, etc.

Classical emulsion polymerization is divided into three kinetic stages. At the start of the process, the unsaturated monomers are dispersed into small droplets, stabilized with surfactants. Additional surfactant aggregates into micelles. These micelles are very small ($\sim$10 nm) relative to monomer droplets ($\sim$1–10 μm). During stage 1 the initial formation of polymer particles takes place by initiation within micelles. In stage 2 the polymerization precedes with a constant supply of new unsaturated monomer, which results in the growth of polymer particles. At the end of stage 2 the monomer supply ceases and subsequently the rate of polymerization decreases gradually to yield latex particles.

A recipe to synthesize a vinyl acrylic latex for paint formulation is shown in Table 8.

Table 7 Typical recipe for methyl methacrylate solution polymerization

Ingredient	Function	Parts (by weight)
Xylene	Solvent	28
Ethoxyethanol	Solvent	14
Methyl methacrylate	Monomer	60
t-Butyl perbenzoate	Initiator	1.1

Table 8 Typical emulsion polymerization recipe

Components	Parts (by weight)	Function
Vinyl acetate	80	Monomer
Butyl acrylate	20	Monomer
Water	180	Continuous phase
Surfactant	2–4	Stabilizer
Sodium bicarbonate	0.05	pH adjuster
Sodium persulfate	0.1	Initiator
Ascorbic acid	0.01	Reductant

(From Ref.[19].)

Water is used as a dispersing medium and a heat sink for the heat of polymerization. Surfactant is an essential part of the emulsion polymerization process. It provides the large surface area needed for the process and stabilizes the final product. The initiator used is typically water soluble and initiation usually takes place within the aqueous phase.

Emulsion polymerization typically refers to the polymerization of a nonaqueous material in water. The polymerization of a water-soluble material in a nonaqueous continuum has been called inverse emulsion polymerization. The inverse emulsion polymerization technique is used to synthesize a wide range of polymers for a variety of applications such as wall paper adhesive, waste water flocculant, additives for oil recovery fluids, and retention aids. The emulsion polymerization technique involves water-soluble polymer, usually in aqueous solution, emulsified in continuous oil phase using water in oil emulsifier. The inverse emulsion is polymerized using an oil- or water-soluble initiator. The product is a colloidal dispersion of submicroscopic particles with particle size ranging from 0.05 to 0.3 μm. The typical water-soluble monomers used are sodium *p*-vinyl benzene sulfonate, sodium vinyl sulfonate, 2-sulfo ethyl acrylate, acrylic acid, and acrylamide. The preferred emulsifiers are Sorbitan monostearate and the oil phase is xylene. The proposed kinetics involve initiation in polymer swollen micelles, which results in the production of high molecular weight colloidal dispersion of water-swollen polymer particles in oil.

Another type of emulsion-like process is called miniemulsion polymerization. Miniemulsions are stable oil in water emulsions with average droplet diameter of 80–400 nm, prepared using a mixture of an anionic emulsifier and a cosurfactant such as a long-chain fatty alcohol or *n*-alkane. The polymer latexes are prepared by initiation of polymerization in the miniemulsion droplets.

Microemulsion polymerization is a term used to describe thermodynamically stable surfactant-rich microemulsions and polymerization therein.

Precipitation/Dispersion Polymerization

Dispersion polymerization is a technique that permits the preparation of polymer particles typically in the range of 1–15 μm in a rapid process with high conversion. In a dispersion polymerization process, the continuous phase is chosen to be a solvent for monomer to be polymerized, a nonsolvent for the resultant polymer. A steric stabilizer is used to produce a colloidally stabilized dispersion. In the absence of stabilizer, the polymerization produces macroscopic particles of polymer of an uncontrolled size, and is called precipitation polymerization.

The minimum requirements for a dispersion polymerization are monomer, solvent/nonsolvent, initiator, and steric stabilizer. The monomer must be soluble in the reaction mixture and its polymer, insoluble. The monomers used in systems of commercial interest are methyl methacrylate, vinyl chloride, vinylidene chloride, vinyl esters, hydroxyl alkyl acrylates. A typical recipe for dispersion polymerization is shown in Table 9.

Uses for dispersion polymers include surface coatings for metal coatings, particularly for the automotive industry, food can coating, chromatographic medium, electrophotographic toners, and pressure sensitive adhesives.

Suspension Polymerization

The term suspension polymerization (also termed as bead or pearl polymerization) refers to polymerization in an aqueous phase with monomer as a dispersed phase, resulting in polymer as a dispersed solid phase. In general, suspension polymerization is carried out by suspending the polymer as droplets (0.001–1 cm in diameter) in water as a continuous phase. In a typical suspension polymerization, the initiator is dissolved in the monomer phase. These types of initiators are referred to as oil-soluble initiators. Each monomer droplet in a suspension polymerization is like a small bulk polymerization reactor. The suspension is maintained by agitation and the use of polymeric stabilizers. The suspension polymerization method is not used with monomers whose polymers are highly water soluble or whose homo- or copolymers have a very high glass transition temperature. The suspension polymerization technique is used to prepare polymers such as polystyrene, poly(methyl methacrylate), poly(vinyl acetate), and poly(vinyl nitrile). In other terms, suspension polymerization is essentially equivalent to bulk polymerization, but is carried out in reaction medium in which the monomer is insoluble and dispersed in droplets, with a catalyst system that generates radicals within the droplets. The highest-volume polymer made

Table 9 Typical dispersion polymerization recipe

Material	Parts (by weight)	Function
Methyl methacrylate	10	Monomer
Polyvinyl pyrrolidone (PVP)	0.55	Stabilizer
Anionic surfactant	0.06	Stabilizer
Methanol	62.5	Dispersion medium
Water	26.8	Dispersion medium
Benzoyl peroxide	0.1	Initiator

Table 10 Typical recipe: PVC polymerization by suspension polymerization

Ingredients	Amount (per 100 parts monomer)
0.03% aq. dispersing agent	200
Ferrous sulfate	0.15
Vinyl chloride	100
Lauroyl peroxide	0.2
pH adjuster (0.5% sodium hydroxide)	1.7

by suspension polymerization is polyvinyl chloride (PVC). A typical recipe for PVC polymerization by suspension polymerization is described in Table 10.

ADVANTAGES AND DISADVANTAGES OF FREE-RADICAL POLYMERIZATION TECHNIQUES

The advantages and disadvantages of several free-radical techniques discussed in the previous section are discussed in Table 11. Some common polymers produced using these processes in the chemical industry are also listed in the table.

REACTION ENGINEERING

The reaction kinetics and reaction processes described in the previous sections lead to a variety of industrial implementations and related issues with pre- and postprocessing of materials. The heat, mass, and momentum transport in the reactors, the thermodynamics of various reaction systems and their impact on the reaction kinetics have been the subject of intense study.[20]

There are three broad stages of any industrial free-radical polymerization. These include the preparation stage, the polymerization stage, and the postprocessing stage. In each of these stages, there are significant issues that need to be dealt with depending on the type of polymerization process being used.

The Preparation Stage

The preparation stage involves purification of reagents and preparation of reagent streams for addition to the reactor. The unit operations may involve removal of impurities from reagents including purging with inert gases to remove oxygen (an inhibitor). Reagent streams may be premixed by operations like stirring, homogenization, and gas sparging, depending on the type of polymerization process under consideration. For example, it is common to add certain monomer streams homogenized with a concentrated aqueous surfactant mixture called a pre-emulsion or premix. The preparation steps may also include heating or cooling certain system parts or reagents to predetermined temperatures. In addition, there are the usual setup and preparation steps to operate the banks of pumps, heat exchangers, utility streams, and other processes that go hand in hand with running an industrial polymerization process.

The Polymerization Stage

The polymerization stage deals with converting the monomer(s) to polymer and controlling the level and rate of conversion as well as the properties of the product to within desirable specifications. There are three very important parameters that change significantly with most polymerization processes. These are viscosity, density, and solubility. An increase in the amount of polymer in the reactor typically leads to higher viscosity, higher density, and lower solubility in the reaction medium. These have different impacts on different polymerization processes. There have been a variety of models and correlations that can effectively describe these changes and even relate them with some degree of confidence to the transport phenomena and the resulting impact on reaction kinetics.[20]

Higher viscosity leads to significant changes in the effectiveness of heat, mass, and momentum transport. These can result in issues like compartmentalization and localized autoacceleration in reactors resulting in localized charring or undesirable variation in product properties. Higher density leads to volume shrinkage and could have a significant impact on design issues like the location of feed lines, sampling lines, heat addition and removal equipment, and agitator configurations.

These factors have led to a variety of reactor types and designs that help control the rates of heat, mass, and momentum transport, residence time distribution of the reactor contents, and as a result the quality and consistency of the reactor product. The basic reactor types, the processes they are used for, and some of the polymers they are used to produce are listed in Table 12.

The safety of the equipment and the people operating it, as well as the environmental impact, need to be considered for the design of the equipment and processes. Other important details include the installation and utilization of appropriate sensors and measurement tools, the establishment and implementation of

Table 11 Comparison of free-radical polymerization processes

Process	Advantages	Disadvantages	Polymers
Bulk—polymerization in the absence of a dispersing or diluting medium	High reactor utilization, low separation cost, high product purity, no transfer reactions	High viscosity during the course of polymerization	Low-density polyethylene, PVC, methyl methacrylate
Solution—polymerization of monomers dissolved in solvents in which the polymer products are also soluble	Lower viscosity than bulk and better heat transfer and mixing; direct application of solution; less reactor fouling	Smaller reactor capacity than bulk; high separation cost; often inflammable and toxic solvents; lower molecular weights	SBR, polyvinyl acetate, polystyrene, acrylics
Suspension—polymerization of monomers dispersed in an inert phase with monomer-soluble initiator	Low dispersion viscosity compared to bulk; good heat transfer; high polymerization rate and high molecular weight; direct application of the latex	Smaller reactor capacity than bulk; reactor wall fouling; wastewater problems	Polystyrene, PVC, polypropylene
Emulsion—formation of small particles via micellar or homogenous nucleation	Low dispersion viscosity compared to bulk; high molecular weight; good heat transfer	Emulsifier impurity; reactor wall fouling; emulsifier impurity	SBR, polyvinyl acetate, vinyl acrylics, etc.

(From Ref.[2].)

Table 12 Reactor types classified according to their operation modes

Operations	Reactor types	Polymers
Bulk and semibatch	Stirred tank reactor	PVC, PVC (B, S, E), polyester (B), polyurethane (B)
Continuous	Continuous stirred tank reactor: loop reactor, stirred tank reactors, fluidized reactors, tubular reactors	Polyvinyl acetate, styrene–butadiene, PVC (E), polystyrene (S), low-density polyethylene (B)

B, bulk; S, solution; E, emulsion.
(From Ref.[2].)

appropriate process control schemes, and the monitoring of appropriate parameters to ensure that the polymerization process is proceeding according to specification.

Changes in reactor conditions may lead to changes in several important product properties, such as molecular weight and molecular weight distribution, copolymer composition, and the level and type of branching and cross-linking. These may have a substantial impact on the application properties of the polymer and hence need to be controlled to ensure quality.

The Postprocessing Stage

Once the reaction has been completed the reactor product typically consists of polymer with other inert or unreacted reagents. The extraction of polymer in a form that can be further utilized needs special attention. The postprocessing may be relatively simple as in the addition of residual monomer scavenging systems to solution or emulsion polymer products to get close to 100% conversion and remove residuals that could be harmful to eventual users or the environment. It may also involve complex processes like controlled coagulation of styrene–butadiene emulsions to get styrene–butadiene rubber (SBR).

The physical transport of the reactor products may also need sufficient processing. Bulk polymerizations in high-pressure extruders may be followed by dicing the product into small pellets suitable for pneumatic transport. Emulsion and solution products are usually transported as obtained and frequently used in the same mode. The design of suitable containers and transportation protocols is very important to avoid harming the product during transportation and storage. For example, special containers may be required for air-sensitive materials, or to avoid solvent or water loss from solutions and emulsions, respectively.

In some instances, undesirable products may result because of loss of reactor control or unforeseen events during the reaction. These may have a varied impact on the polymerization product and could lead to significant costs, if they occur. The design of polymerization processes, equipment, and related pre- and postprocesses needs to consider all of these factors to ensure success.

CONCLUSIONS

Free-radical polymerization continues to be one of the principal industrial routes to synthetic polymers. Every year, billions of pounds of industrial polymers are synthesized using this technique for a variety of applications. The advent of CFRP techniques promises to enhance the value of free-radical polymerization significantly from an industrial as well as an academic perspective. It promises to deliver products that will help in satisfying the increasing demands of high-performance applications in existing and emerging market segments.

REFERENCES

1. Odian, G. *Principles of Polymerization*; John Wiley and Sons: New York, 1991.
2. Allen, G., Ed. *Comprehensive Polymer Science*; Pergamon Press: New York, 1989; Vol. 4.
3. Rudin, A. *The Elements of Polymer Science and Engineering*; Academic Press: New York, 1982.
4. Immergut, E.H.; Brandrup, J. *Polymer Handbook*, 3rd Ed.; John Wiley and Sons: New York, 1989.
5. Dostal, H. Isomerization by bimolecular nucleus formation. Monatsh. Chem. **1935**, *67* (1–9), 222.
6. Mayo, F.R.; Walling, C. Copolymerization. Chem. Rev. **1950**, *46*, 191–287.
7. Mayo, F.R.; Lewis, F.M. Copolymerization. I. A basis for comparing the behavior of monomers in copolymerization; the copolymerization of styrene and methyl methacrylate. J. Am. Chem. Soc. **1944**, *66*, 1594–1601.
8. Fineman, M.; Ross, S.D. Linear method for determining monomer reactivity ratios in copolymerization. J. Polym. Sci. **1950**, *5*, 259–262.
9. Tüdõs, F.; Kelen, T.; Foldes-Berezsnich, T.; Turcsanyi, A. Evaluation of high conversion

copolymerization data by a linear graphical method. Reaction Kinetics and Catalysis Letters **1975**, *2* (4), 439–447.
10. Chee, K.K.; Ng, S.C. Estimation of monomer reactivity ratios by the error-in-variable method. Macromolecules **1986**, *19* (11), 2779–2787.
11. Tosi, C. Determination of monomer reactivities in copolymerization. Eur. Polym. J. **1973**, *9* (4), 357–365.
12. Tidwell, P.W.; Mortimer, G.A. An improved method of calculating copolymerization reactivity ratios. J. Polym. Sci.: Part A **1965**, *3* (1), 369–387.
13. Kim, S. Effect of Solvent on Free Radical Initiated Copolymerization of Styrene and Acrylic Acid, Ph.D. dissertation, University of Akron, 1990.
14. Matyjaszewski, K. *Advances in Controlled/Living Radical Polymerization*; ACS Symposium Series 854; American Chemical Society: Washington, DC, 2003.
15. Dar, Y.L.; Caneba, G.T. Transport phenomena aspect of the free-radical retrograde-precipitation polymerization process. Chem. Eng. Commun. **2002**, *189* (5), 751–607.
16. Mishra, M.; Yagci, Y. *Handbook of Radical Vinyl Polymerization*; Marcel Dekker Inc.: New York, 1998.
17. Lovell, P.A.; El-Aasser, M.S. *Emulsion Polymerization and Emulsion Polymers*; John Wiley and Sons: New York, 1997.
18. Long, T.E.; Hunt, M.O. *Solvent Free Polymerizations and Processes*; ACS Symposium Series 713; American Chemical Society: Washington, DC, 1988.
19. VeOVA Monomers Technical Information System. Resolution Performance Products: Rotterdam, The Netherlands, 2004.
20. Dotson, N.A.; Galvan, R.; Laurence, R.L.; Tirrell, M. *Polymerization Process Modeling*; VCH Publishers: New York, 1996.

Friction Materials

Sunil Kesavan
Xinming Shao
Akebono Corporation, Farmington Hills, Michigan, U.S.A.

INTRODUCTION

Friction materials are used in a variety of automotive and nonautomotive applications to control vehicle movement or for the transmission of power. This entry will limit its scope to automotive brake friction materials. Automotive friction materials are composites that are incorporated into a disk or drum brake system (a) for stopping or reducing the speed of a vehicle by converting kinetic energy into frictional heat or (b) to prevent the movement of parked vehicles.

Over the past century, a variety of brake configurations have evolved to accommodate the increasing demands placed on them. Modern vehicles use friction materials shaped for use in disk or drum brakes for dynamic braking or parking purposes. Disk brakes employing two friction pads hydraulically pressed on the faces of a rotating solid or ventilated pearlitic gray cast iron disks are used for the front or rear corners of vehicles (Fig. 1). Drum brakes use two curved brake linings pressing outward on the inside surface of a cast iron drum and are used for the rear service or parking brakes on vehicles (Fig. 2).

Original equipment (sold on new vehicles) friction materials are formulated to help the brake system meet three sets of requirements:

1. *Legal requirements*: Government requirements like the Federal Motor Vehicle Safety Standard (FMVSS) 105 and 135 in the United States and ECE R13 in Europe and South America are normally specified to cover vehicle stopping distance under cold, hot, high-speed, thermal and water fade and recovery conditions, and inoperative power assist. Additional requirements are imposed on the selection of raw materials by health, safety, and environmental regulations. These are the minimum requirements that need to be satisfied by a brake system.
2. *Vehicle manufacturer requirements*: These requirements are defined by the vehicle manufacturers and cover performance under different conditions, product piece and life cycle costs, ease of installation and servicing, health, safety, liability, and warranty considerations.
3. *Customer requirements*: Customers expect stable, smooth, responsive, reliable, and noise-free performance from their brakes in a variety of real-life terrains and seasons. In addition to performance considerations, they desire long pad/rotor life and low maintenance costs.

No single friction material can completely meet all these requirements for every vehicle. The objective of friction material formulation development is to custom design application-specific formulations whose performance is a compromise between conflicting requirements. In several countries, including North America, there are no regulations or standards governing the performance or durability of replacement friction products installed on vehicles after sale.

TYPES OF FRICTION MATERIALS

Asbestos Friction Materials

Asbestos friction materials introduced in 1908 contained up to 65% of short chrysotile asbestos fibers that served as a cost-effective, heat-resistant, insulating reinforcement and friction enhancer. Asbestos formulations were widely employed in drum and disk formulations for several decades until health and safety concerns triggered efforts to find suitable substitutes. These materials are being universally phased out in favor of nonasbestos formulations.

Semimetallics

A semimetallic formulation typically contains over 50 wt% of ferrous metal in the form of steel fiber and porous iron powder. Other typical ingredients include binder resins, abrasives, graphite and coke lubricants, rubber particles, and cashew friction dust. These materials were developed in the 1960s to meet the demands of heavy-duty and high-temperature braking. The use of semimetallic formulations for automobiles increased in popularity in the 1970s with the imposition of Federal Motor Vehicle Safety Standard 105 (FMVSS105) and the arrival of front wheel drive vehicles with their higher brake operating temperatures. Semimetallics have good wear

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007778

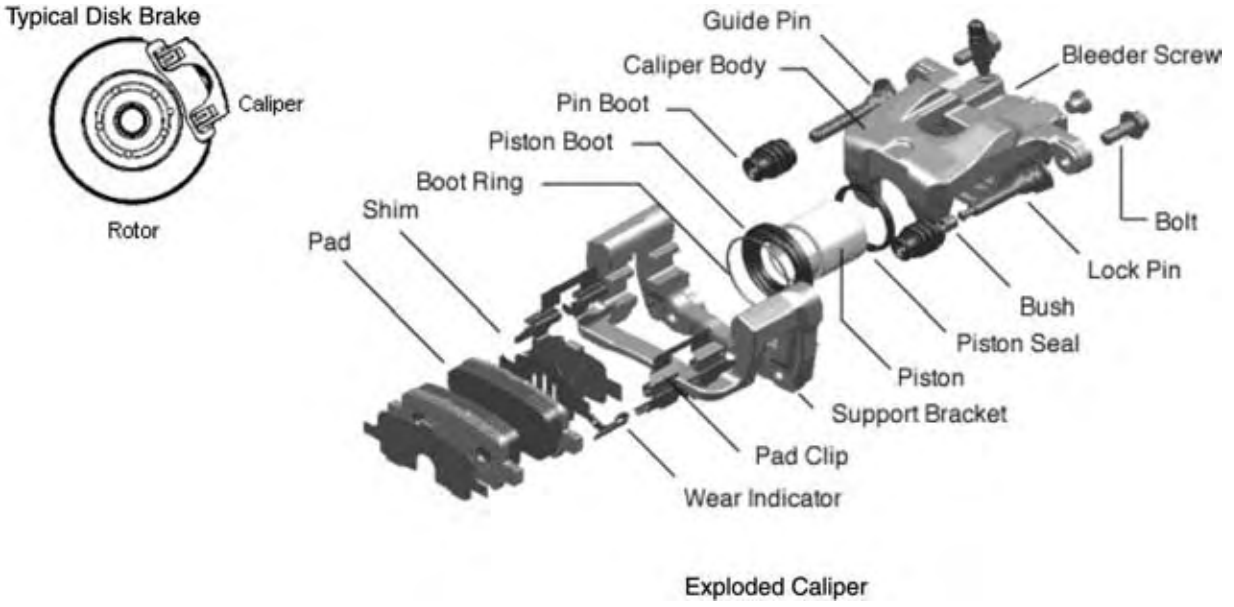


Fig. 1 Schematic of a typical disk brake showing the caliper and rotor. The exploded view of a floating single piston caliper is also shown. (Courtesy of the Akebono Corp.)

characteristics under heavy-duty and higher temperature use, but have high wear at lower temperatures and high speeds. Wear life of semimetallics in actual use is very dependent on the conditioning of the friction couple. These materials have poor friction performance under cold conditions and can also suffer from noise, vibration, and harshness (NVH) and corrosion problems. Semimetallic formulations have a relatively high thermal conductivity, which can lead to problems with brake fluid boil. Backing layers (a thin layer of friction material between the backing plates and bulk friction material) with lower thermal conductivity are sometimes used to partially offset the effect of higher thermal conductivity.

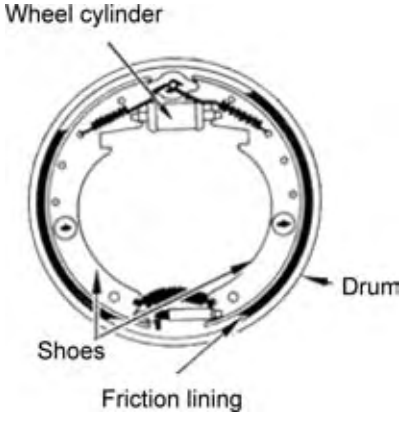


Fig. 2 Schematic of a typical drum brake.

Nonasbestos Organics

As the name indicates, these are asbestos-free formulations that use a fiber cocktail of aramid pulp, ceramic fiber, rockwool, and other fibers or whiskers for reinforcement. These fibers are mixed with binders, abrasives, lubricants, fillers, and other property modifiers. This type of formulation is available in friction levels up to about 0.40 and is best suited for light and medium duty applications. Nonasbestos organics (NAOs) show excellent wear characteristics at usage temperatures up to about 250°C. Above this temperature, pad and rotor wear tend to increase dramatically. Nonasbestos organics started replacing semimetallic formulations on automobiles starting in the early 1990s because of their superior friction stability, creep groan, NVH behavior, and low wheel dust. The so-called ceramic formulations currently marketed by some manufacturers are essentially NAOs containing heat-resistant ceramic ingredients.

Low-Metallics

These friction materials contain both ferrous and nonferrous metals and are mostly used in European friction applications. Low-metallics normally have a high friction level with the friction coefficient in the 0.40–0.50 range and good high-temperature and high-speed performance to meet European legal requirements (ECE R13). A combination of carbonaceous and

sulfide lubricants and aggressive abrasives like alumina and silicon carbide are used to achieve the desired performance. Given the aggressiveness of low-metallics, pad and rotor life are limited, with a concomitant increase in wheel dust formation. Table 1 summarizes the relative characteristics of semimetallic, low-metallic, and NAO friction materials.

Sintered Metallics and Cerametallics

These are iron- and copper-based inorganic friction materials used for severe applications like tractor clutches, train, and aircraft brakes. Sintered metallic formulations contain lubricants like graphite, abrasives, and other property modifiers held together by metal binders like copper and iron. Sintered friction materials are normally fabricated as buttons that are riveted or fused to a carrier plate or cup. These friction materials are well suited for heavy-duty applications and find use in either dry or oil-immersed applications.

Carbon–Carbon Friction Materials

These expensive carbon–carbon fiber composites are used for thermally demanding applications like aircraft and racecar brakes.[1] Typically, both pads and rotors are made of carbon composite. These friction couples have a relatively low friction level, which is insensitive to the high temperatures experienced in operation. Friction couple components are expensive to manufacture as they are fabricated by time-consuming chemical vapor deposition techniques. Carbon–carbon systems suffer from low cold friction and are best suited for hot operation.

Ceramic Friction Materials

Ceramic friction couples have been developed for use in race and high-performance Porsche and Mercedes car applications. These brakes use expensive heat and fade-resistant ceramic disks which weigh about half of conventional cast-iron disks. The disks are made from specially treated carbon fibers that are siliconized at 1700°C in a vacuum. The use of these lightweight brakes improves braking performance as well as ride quality owing to lower unsprung weight. However, their low heat capacity and thermal conductivity places high demands on the friction pads. Development of suitable friction materials for use with ceramic rotors is still work in progress.

TRIBOLOGY

Friction Coefficient

The simplest way to characterize friction materials is to use the so-called friction coefficient μ. This coefficient is defined by Amontons' law:

$$F = \mu P$$

where F is the friction force at the sliding interface acting in a tangential direction relative to the normal force P. According to the Amontons' law, friction coefficient μ is a dimensionless constant and is an inherent property of materials in contact. However, in real-world applications such as an automobile brake system, the friction coefficient is almost never a constant. In fact, it is very much dependent on the brake system and its operating conditions. Fig. 3 depicts the complex interdependence of friction behavior on various factors. This figure clearly shows why it is almost impossible to satisfactorily model the friction behavior of a brake system.

Because of the continuously changing nature of the friction process, many different mechanisms of friction and wear may operate simultaneously with one or two mechanisms dominating. The dominant mechanism is different from one friction system to another. In addition, for a given friction system, the dominant

Table 1 Comparison of the performance characteristics of various classes of friction materials

				Wear		NVH				
Friction materials	Friction stability	Cold friction	Fade	Low temperature	High temperature	Roughness	Squeal	Groan	Corrosion	Thermal transfer
NAOs	★	★	◯	★	△	★	★	◯	★	★
Semimets	◯	△	◯	△	★	△	◯	△	△	△
Low steel	★	★	★	◯	◯	△	△	△	◯	◯

★ Excellent; ◯ good; △ fair.

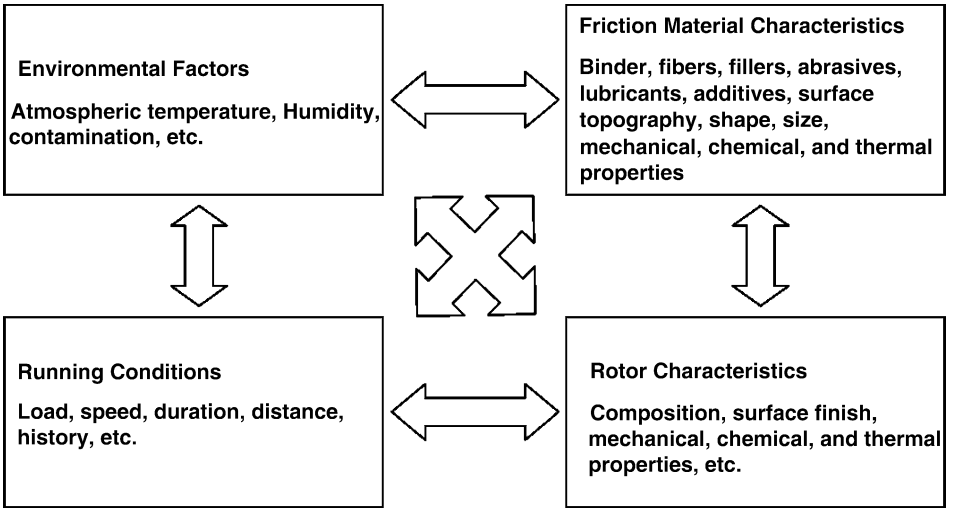


Fig. 3 Representation of the complex interdependency of friction characteristics on various factors.

mechanism can change depending on operating conditions. The key for engineering a successful friction system is to identify and analyze the dominant friction and wear mechanisms for a particular system under the operating conditions of interest to the end user.

The heat energy generated during friction is due to the work done by the friction force. In addition to energy dissipation as heat, energy is also spent to create new chemicals, form new surfaces, and in the work of plastic and viscous deformation and dislocation of surface materials. An important energy dissipation mechanism is the creation of vibration or sound energy. In automobile brake systems, generation of friction-induced vibration, groan, squeal, and other unpleasant noises has often been the most important and difficult engineering problems to solve.

Wear Mechanisms

In sliding friction systems, one or more of the following wear mechanisms come into play: abrasive wear, adhesion and tearing, thermal or ablative wear, macroshear, and fatigue.[2,3] Abrasive wear is a result of the ploughing action of surface asperities (two-body) or the hard wear particles trapped between the wear surfaces (third body; Fig. 4). Adhesive wear involves the formation and rupture of interfacial adhesion bonds. Repeated thermal and mechanical stressing of friction material surfaces results in fatigue wear. Macroshear wear results when a thermally degraded friction surface undergoes mechanical failure under severe mechanical stress. Fig. 5 shows the wear mechanisms operating under different driving conditions for automobile braking systems.

Fig. 4 schematically represents transfer layers that usually form on the friction and mating surfaces of friction couples. These complex layers are formed from inorganic and organic constituents derived from the mating surfaces. The transfer layers are formed and destroyed in dynamic fashion and contribute greatly to the friction, wear, and noise performance of friction couples. Some of the wear generated by transfer layers is of a redundant nature as material gets transferred back and forth between the mating surfaces.

In simple terms, the objective of designing friction materials is to maintain a desirable friction level while keeping friction couple wear and noise/vibration to a minimum.

PERFORMANCE CHARACTERISTICS OF FRICTION MATERIALS

Friction materials are often characterized in terms of friction, wear, and noise performance. However, as explained earlier, these are characteristics of the complete brake system and not of the friction couple alone. Friction material characterization techniques

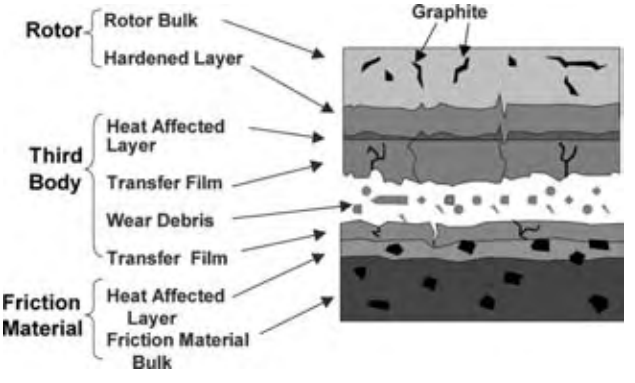


Fig. 4 Representation of the interface between friction materials and rotor/drum counterface. *(View this art in color at www.dekker.com.)*

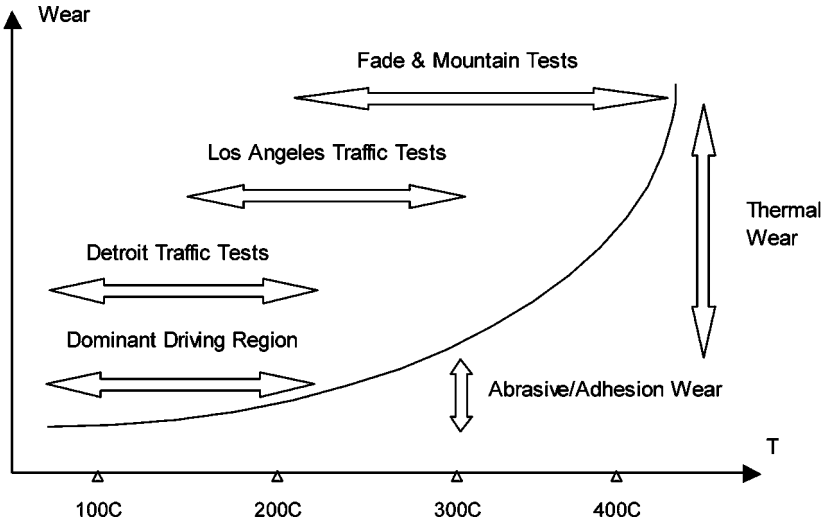


Fig. 5 Operating temperature and wear regimes for various driving conditions.

are developed based on the end-use application. As a result, there are different performance characteristics and related test techniques for automobile brake friction materials.

A typical standard friction performance (effectiveness) test for a passenger car brake usually includes the following measurements: green and burnished effectiveness; thermal fade and recovery; speed, pressure, and temperature sensitivity; water fade/recovery; moisture sensitivity; cold and static friction. These aspects of friction performance are discussed elsewhere in some detail.[4] The performance of an ideal friction material will be immune to changes in braking conditions.

Wear: Wear behavior of friction materials and their mating surfaces is an economic consideration. Although it is desirable to minimize wear, a certain amount of wear is needed to satisfactorily remove corrosion products and material transfer at the interface. As in the case of the friction coefficient, wear depends on brake system and operating characteristics. Accelerated dynamometer tests are routinely used to characterize the temperature dependence of wear rate. These tests are best suited for judging the relative wear performance of materials. More meaningful predictions of lining and rotor life are derived from vehicle road tests that include a large number of stops designed to emulate real-life driving conditions. For most friction materials, wear rate increases with temperature (Fig. 5).

NVH: As mentioned in the section on Tribology, some braking energy is converted into vibration and sound energy. Just as in the tests used for wear characterization, several dynamometer and vehicle tests have been developed to characterize the NVH-generating tendency of brake systems. Dynamometer tests are conducted with the brake corner or a whole vehicle mounted in a controlled environmental chamber. Noise and vibration measurements are made with microphones and accelerometers. Sometimes, sophisticated instruments like laser vibrometers are used for NVH work. Common noise fixes include cutting slots and/or chamfers on the brake pads and linings, the application of constrained layer damping noise insulators to the back of the brake pad shoe plates, and detuning the resonant frequency of mechanical brake corner components.

EVALUATION EQUIPMENT

Laboratory evaluation equipment for testing brakes are designed to evaluate brakes under the widest possible sets of simulated real-life conditions to look for performance flaws or behavioral inconsistencies that can lead to end-user dissatisfaction. Laboratory evaluation equipment can be tailored for either the component or the system level. For component level evaluations, a variety of laboratory evaluation methods have been devised. The custom-designed equipment for this kind of work ranges from small-scale evaluation equipment to large, full-scale inertial dynamometers. Of all available evaluation equipment, full-scale inertial dynamometers come the closest to simulating real-life use. Fig. 6 is a schematic representation of a full-brake dynamometer that is used for evaluation of friction materials. These machines simulate vehicle braking by testing a complete brake-assembly mounted on a rotating shaft loaded with inertia weights chosen to simulate actual road inertia. Table 2 summarizes the commonly used laboratory-scale testing machines.

In addition to laboratory tests, vehicle road tests are used for the evaluation of the whole brake system in various terrains and seasons. The evaluation system

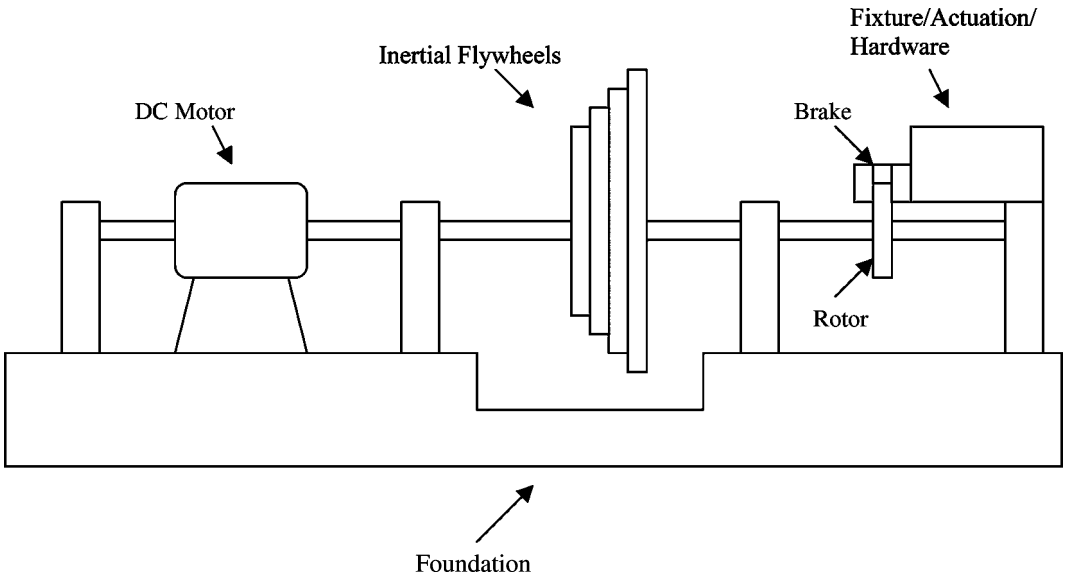


Fig. 6 Schematic representation of a brake dynamometer that is used for evaluation of friction materials.

should include the vehicle with the appropriate end-use brake configuration. The test vehicles are instrumented to acquire information on brake temperature, line pressure, deceleration, noise, vibration, etc. Vehicles can either be driven on roads or run automatically on a vehicle dynamometer.

In addition to the tests mentioned above, there are many standardized pieces of equipment for physico-chemical characterization of friction materials. Table 3 summarizes the different physical and chemical tests that are generally used to characterize friction materials for performance evaluation/prediction or product quality control. Additional information on some of these test procedures is given elsewhere.[5–7]

FORMULATION DEVELOPMENT

As no single raw material can meet all the requirements expected of a friction material, friction linings are

Table 2 Sample and full-scale dynamometers used for characterization of friction materials

Equipment	Characteristics	Application
FAST (friction assessment and screening test) machine	Small friction material specimens dragging for 90 min on the circumference of a test ring at constant torque	Used for quality assurance only
Chase machine	Friction sample (~25.4 mm square pad) rubbing against the inner diameter of a rotating drum	Used to conduct SAE J661 tests for friction edge code determination
Subscale custom testing machines	Usually custom-built for a specific purpose (vibration, fade, thermal stability, friction, wear stability, etc.) with well-designed control systems	Cost-effective tools for simulation and studying the fundamental performance of friction materials
Full-size inertial dynamometers	Uses a motor and shaft-mounted inertial weights to provide energy for friction testing. Brake-specific hardware is mounted on these machines for evaluation	Extensively used for evaluation and problem solving in friction material and brake component development

Table 3 Summary of various tests used in friction material manufacturing for raw material and finished product quality assurance

Raw material tests	Finished product tests
Visual checks	Specific gravity
Specific gravity	Compressibility
Particle size distribution	Hardness
Bulk density	Sonic test (grindosonic)
Moisture content	pH
Flow tests for resin	Flatness and parallelism
Viscosity and solid content of liquid resin	Dimensional tests
Chemical compositions as needed	Swell and growth
Pyrolytic gas chromatography	Shear strength
Infrared analysis	Visual inspection for
X-ray diffraction	Cracks and chips
TGA/DSC	Delaminations
	Contamination
	Dispersion
	Dynamometer tests
	SAE J661, FAST, full scale
	Chemical composition as needed
	Pyrolytic gas chromatography
	Infrared analysis
	X-ray diffraction
	TGA/DSC
	Electron microprobe analysis (EMPA)
	SEM

TGA, thermogravimetric analysis; DSC, differential scanning calorimetry; EMPA, electron microprobe analysis; SEM, scanning electron microscopy.

complex composites that contain anywhere from about 6 to over 20 raw materials. Friction material companies use iterative development techniques to develop proprietary formulations customized to their end-use applications. Formulation development is both an art and a science as very little is known about fundamental tribological processes and their dependence on ingredient properties. Intuition and past experience have proven to be the best tools in the evolutionary process of friction material development.

In addition to the characteristics of the myriad raw materials, processing parameters also have an important role in determining the performance of the compounded product. Given the limitless ingredient and process combinations possible, formulation by trial and error proves to be a laborious and expensive process. Some effort has been made in recent years to use artificial intelligence tools to shorten the development cycle time and to remove human bias from the development process.

RAW MATERIALS

Raw materials can be classified under four broad categories with some overlap: binders, reinforcements, friction modifiers, and fillers.

Binders

Both organic and inorganic binders are used to hold the various ingredients together in a friction material formulation. Unmodified and organic/inorganic modified phenolic resins are the most widely used binders in friction materials because of their easy processing, heat stability, high char yield, and relatively low cost. Organic modifiers typically used for phenolic resins are nitrile, silicone and chlorinated rubbers, cashew and vegetable oils, epoxy, cresol, alkylphenol, and phenoxy groups. Typical inorganic modifiers are phosphorus, boron, and silicon. Liquid resole resins are used either alone or in combination with powdered phenolic resins. Water, alcohol, and other solvents are sometimes used to intimately disperse the powdered resins in friction material mixes. The physicochemical characteristics of binder resins like flow distance, B-time (time needed for the resin to pass through a rubbery stage where the resin is insoluble and almost infusible), hexamethylene tetramine content, and type of modification can be tailored to obtain the required processing and end-use friction material characteristics.

Unmodified or rubber-modified liquid cashew resins have been used to make wet drum lining mixes. These resins are inexpensive and help to improve product

flexibility, but do not confer the mechanical properties and heat resistance provided by powdered phenolic resins.

Green rubber is sometimes used as a primary or secondary binder in disk and drum formulations. Nitrile, styrene–butadiene and chlorinated rubbers have been used for this purpose. These rubbers are intimately mixed with curatives and other ingredients in intensive mixers.

Metal powders and fibers like iron, tin, copper, brass, and bronze are used as binders in sintered formulations. Powder metallurgy processes are used to fabricate these friction materials.

Reinforcements

Chrysotile asbestos fibers that were widely used as reinforcement in friction materials in the past are being replaced by a combination of fibers. Although being phased out for health and safety reasons, asbestos is well suited for use in friction materials because of its strength, thermal stability, flexibility, briquettability, ease of processing, and low cost.

No single fibrous reinforcement embodies all the desirable characteristics of asbestos fibers. Combinations of organic and inorganic fibers are used in friction materials to achieve the desired effect. Given the past experience with asbestos, end-users of fibrous reinforcements have to insure that the materials do not cause cancer, fibrotic lung disease, or pulmonary dysfunction.

Organic fiber reinforcements

Fibrillated aramid (Kevlar® and Twaron®) pulps are widely used as a moderately heat-resistant reinforcement in friction formulations. Aramid pulps provide briquettability for preforms, improve friction and wear characteristics, help modify compressibility, and aid in noise damping. However, these fibers are one of the most expensive raw materials used in friction materials.

Polyacrylonitrile pulp is used in some friction materials as a processing aid and to help improve the briquettability, ingredient retention, and green strength of preforms. Various forms of cellulose pulp or fiber derived from recycled clothing, printed matter, and natural sources (flax, hemp, and sisal) are also used in friction materials. Cellulose does not possess heat resistance but helps to hold ingredients together. Sometimes cellulose fiber is added to formulations to provide porosity when burned off during oven curing.

Inorganic fiber reinforcements

Chopped strands (1/8 to 1/4 in.) of silanized E-glass fibers are used as reinforcements in some formulas. Glass fibers provide excellent thermal reinforcement but tend to be aggressive on the mating surfaces. Dispersion of glass fibers in dry mixes can be problematic as overmixing of fibers can result in the formation of undesirable fiber balls.

Man-made mineral fibers (MMMFs) are widely used to provide high-temperature reinforcement for friction materials. Mineral wool fibers are short (with diameters around 5–6 μm and lengths up to 650 μm) calcium–alumino silicate fibers produced from raw materials like slag, rock, and basalt. Mineral wool is not as effective as chopped glass in providing thermal reinforcement but provides sufficient reinforcement in a cost-effective manner. Also, low-shot (shot is the term used for unconverted raw material present in the final fiber product in the form of small beads) mineral fiber can cause less rotor/drum wear than glass fibers. Another type of MMMF is vitreous ceramic fiber based on combinations of calcia, alumina, magnesia, silica, and zirconia that have a higher melting and softening temperature than mineral wool. Ceramic fibers are generally more expensive than mineral wool. Recently, fiber manufacturers have introduced biosoluble grades of MMMFs to address health and safety concerns associated with respirable fibers.

Chopped carbon fibers derived from either pitch or polyacrylonitrile are also used to provide reinforcement in friction materials but their lubricating tendency lowers the friction level. Although prices have come down in recent years, these reinforcements are quite expensive as a friction material ingredient.

Different grades of chopped low-carbon steel fibers are used as reinforcements and friction enhancers in formulations. Steel fibers have good reinforcing properties and thermal resistance but high density, poor corrosion resistance, and high thermal conductivity. Chopped stainless steel fibers have found limited application in friction materials and backing layers. Some manufacturers have developed annealed and softer grades of steel fiber to reduce rotor and drum wear.

Wollastonite is a natural calcium metasilicate mineral with an acicular structure. Silanized and uncoated versions with maximum fiber lengths up to 25 μm are used as inexpensive secondary reinforcements in friction materials. Short fiber attapulgite clays have also been developed as secondary reinforcements.

Potassium titanate whiskers were found to improve the tribological properties of NAO friction materials.[8] These soft, synthetic refractory materials stabilize friction level, reduce pad wear, noise, fade, thermal conductivity, control porosity, and provide thermal reinforcement. Fibrous potassium titanates are being replaced by platelet and powder versions because of health concerns associated with respirable fibers.

These platy and powder titanates help to improve friction stability, wear and noise performance, but do not provide the thermal crack resistance or porosity control characteristics of fiber versions of titanate.

Friction Modifiers

Friction modifiers are ingredients used in friction materials to alter the friction, wear, and noise behavior. Some friction modifiers also serve as fillers for reducing product cost.

Abrasives are universally used in friction materials to provide the desired friction and wear characteristics. Abrasives also help to clean up corrosion products and other deposits that form on the rotor. The most widely used abrasives are metal oxides and silicates. Examples of abrasives are finely divided magnesium oxide, iron oxides, alumina, chromia, zircon, zirconia, mullite, tin oxide, iron chromite, and silicon carbide. Harder and larger size abrasive particles are used in aggressive formulations that exhibit higher friction level. Finer and softer abrasives are gentler to the rotor and, in general, generate lower friction levels, wear, and noise. In addition to particle size and hardness, morphology also plays a large role in determining the function of an abrasive. Many friction formulations use a combination of abrasives to achieve the desired end-use performance.

Carbonaceous lubricants are used in most friction materials as a cost-effective way to control friction material performance, wear, and NVH characteristics. Commonly used carbonaceous lubricants are natural and synthetic graphites and metallurgical and petroleum cokes.

Fibers and chips of nonferrous metals and alloys like copper, brass, aluminum, zinc, and bronze are used in friction materials to provide the desired friction and wear performance. Porous and solid metal powders like tin, brass, copper, bronze, and iron are also used in some products. As mentioned earlier, some metal powders are also used as binders in sintered formulations.

A variety of metal sulfides and their mixtures have been employed to improve the friction stability, wear, and noise characteristics of friction materials.[9] Commonly used sulfides are antimony trisulfide, molybdenum disulfide, tin sulfides, zinc sulfide, copper sulfides, titanium sulfide, and bismuth trisulfide. Sulfide lubricants are usually used to complement carbonaceous lubricants. Fluoride and molybdate additives are sometimes used to enhance the action of carbonaceous and sulfide lubricants. Examples of fluorides are calcium fluoride and cryolite. Sodium and calcium molybdates have been used in small amounts.

Friction dusts are cured coarse powders composed of polymerized cashew nut shell liquids. These materials are added to friction materials to confer friction stability, resiliency, and noise damping. Rubber and abrasive modified friction dusts are also available. Friction dusts tend to absorb solvents, causing the friction material to spring back after forming. This limits the use of cashew friction dusts to dry mixes or solvent-free wet mixes.

Rubber particles of various origins are used in friction materials to provide friction and wear improvement, noise damping, and as a cost-reducing filler. Powders made from tire peel and other recycled rubber products and partially cross-linked acrylonitrile powders are the most widely used products in friction materials.

Examples of other property modifiers are silicone powder and liquids, fluoropolymer powders, Promaxon® (processed calcium silicate), mica, vermiculite, and waxes.

Fillers

Inexpensive, finely ground minerals like barium sulfate (barytes), dolomite, limestone (whiting), clays, and silica are widely used to provide bulk and reduce the cost of friction material formulations. These materials also act as friction modifiers and alter the performance of the end product. Other less commonly used fillers are hollow and solid organic and inorganic microspheres and fly ash.

MANUFACTURING

All friction material manufacturing processes can be subdivided into the following stages: mixing of proportioned raw materials, forming of friction material under heat and pressure, oven curing to complete cross-linking of thermosetting binder resin, mechanical finishing to final configuration, and packaging.

Mixing

In the mixing process, the proportioned friction material ingredients are intimately mixed under wet or dry conditions. A variety of mixing equipment is used in the friction industry depending on the nature of the ingredients and the desired final product. Several suppliers have developed custom-made mixers suited for the friction industry. Mixers supplied by Littleford (Fig. 7), Lödige, Eirich (Fig. 8), Henschel, and MTI are widely used in the friction industry. These mixers are basically static or rotating cylindrical containers with heavy-duty mixing and chopping blades designed to open fibers and disperse ingredients uniformly.

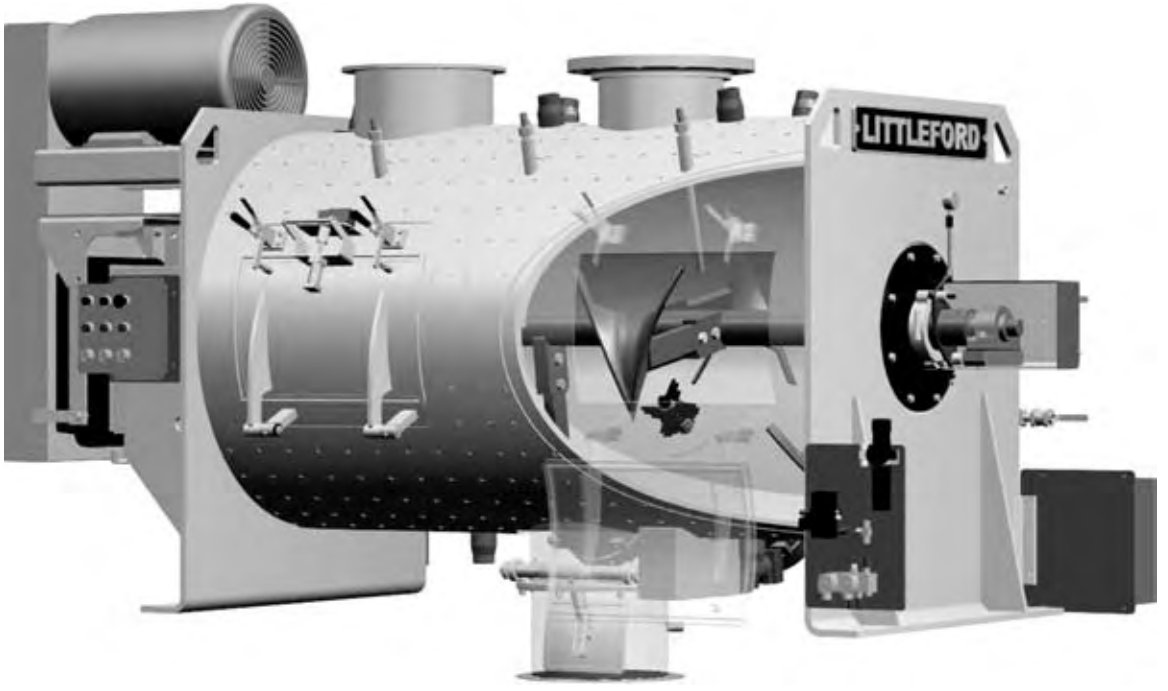


Fig. 7 Cutaway view of a Littleford ploughshare mixer. (Courtesy of Littleford Inc.)

Some mixers are provided with temperature control jackets for heating or cooling. Solvents or liquid resins can be introduced into the mixers through spray nozzles during the mixing cycle. Littleford and Lödige ploughshare mixers use mixing ploughs mounted on a horizontal shaft running in a cylindrical shell. A chopper with a variety of different blades is mounted on the mixer wall. The Eirich mixer uses a mixer blade that rotates in a tilted cylindrical shell. The MTI mixer uses a mixer blade in the bottom of a vertical cylindrical mix chamber with a chopper mounted in the wall.

Various mix parameters can be changed to control the dispersion of ingredients and the ''fluffiness'' of the mix: 1) mixing time: good opening of fibers and dispersion of ingredients requires an appropriate mixing duration—mixing for too long can cause agglomeration and/or degradation of fiber length; 2) adjusting mixer load; 3) changing chopper configuration and speed; and 4) order of addition of ingredients.

Fig. 8 Cutaway view of an Eirich mixer showing mixing action. (Courtesy of Eirich Machines.)

Rubber premixes made using bale rubber are processed in high-intensity Banbury mixers that are widely used in the rubber processing industry. The mixer uses two heavy-duty counterrotating mixer blades to intimately mix fillers and fibers into the rubber. The product of Banbury mixers are either directly formed in a cure press or used as an ingredient in a friction formulation after size reduction.

Preforming

Many manufacturers preform the dry or wet mix into briquettes prior to press curing. The preforms are usually made by briquetting mixes at room temperature. Although an additional step in the manufacturing process, preforming is used to improve utilization of the cure presses. Some manufacturers bypass the preforming step by producing a granular, free-flowing mix that can be directly metered into the cure press molds.

Friction Material Forming

Disk pads

In the integral molding process for fabricating disk pads, the preforms or loose mix is molded onto a shoe

or metal backing in a press under high pressure and at temperatures ranging from room temperature to around 150°C. Prior to molding, the friction material side of the shoe or metal carrier plate surface is coated with an adhesive after a surface treatment designed to promote adhesion. Some products designed for heavy-duty applications use a mechanical attachment technique instead of adhesive attachment. Depending on the temperatures used, the pressing process is classified as cold, warm, or hot pressing. In the past, disk pad shapes molded without a backing plate were riveted onto a backing plate. This process is generally being replaced by the integral molding process.

Drum linings

For riveted and bonded linings, the drum friction material segment is first formed into shape by either positive molding or extrusion techniques. In the positive molding process, drum linings are formed under heat and pressure from a dry mix in either flat or curved molds. If a flat sheet fabrication method is used, the formulation is only partially cured so as to allow subsequent forming on a mandrel into a curved shape. Extrusion forming is a low-cost manufacturing process for bondable drum linings where a granulated mix is extruded through a rectangular orifice. Liquid phenolic or cashew nut shell resins and rubbers with or without solvents are used to produce an extrudable mix with sufficient tack. The mix is forced into the nip between forming rollers that extrude the friction material in strip form. Some manufacturers embed a wire backing into the extruded strip to provide strength during process handling and service use. The strips are then either formed into rolls or cut into the desired length and cured in confined preforms to confer the curved shape required in use.

Postcuring

A final oven cure cycle is needed to complete the cure of the binder resin started in the molding process and also to develop the maximum heat resistance of the friction material. Both batch and continuous ovens using convection or infrared heating are used in curing friction materials. Typical oven cure cycles can range from 1 to 16 hr depending on the cured product. Both disk and drum linings can be cured either free or confined by steel plates under pressure. Although confined curing is more expensive than free curing, it reduces the chance of delamination or debonding during oven cure. After postcuring, the inner and outer surfaces of the formed linings are finish ground to the desired thickness. For bonded applications, adhesive is applied on the inner surface before the lining is bonded under pressure and heat onto the carrier shoe. In riveted

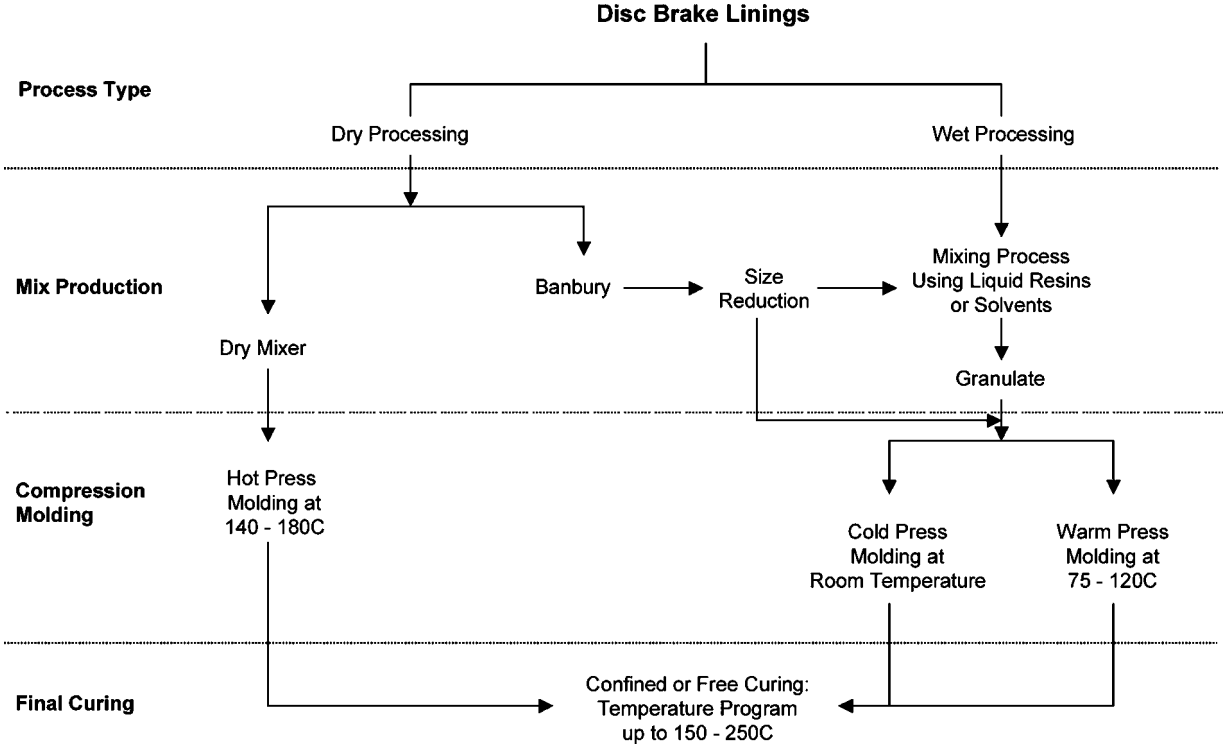


Fig. 9 Disk brake pad production processes.

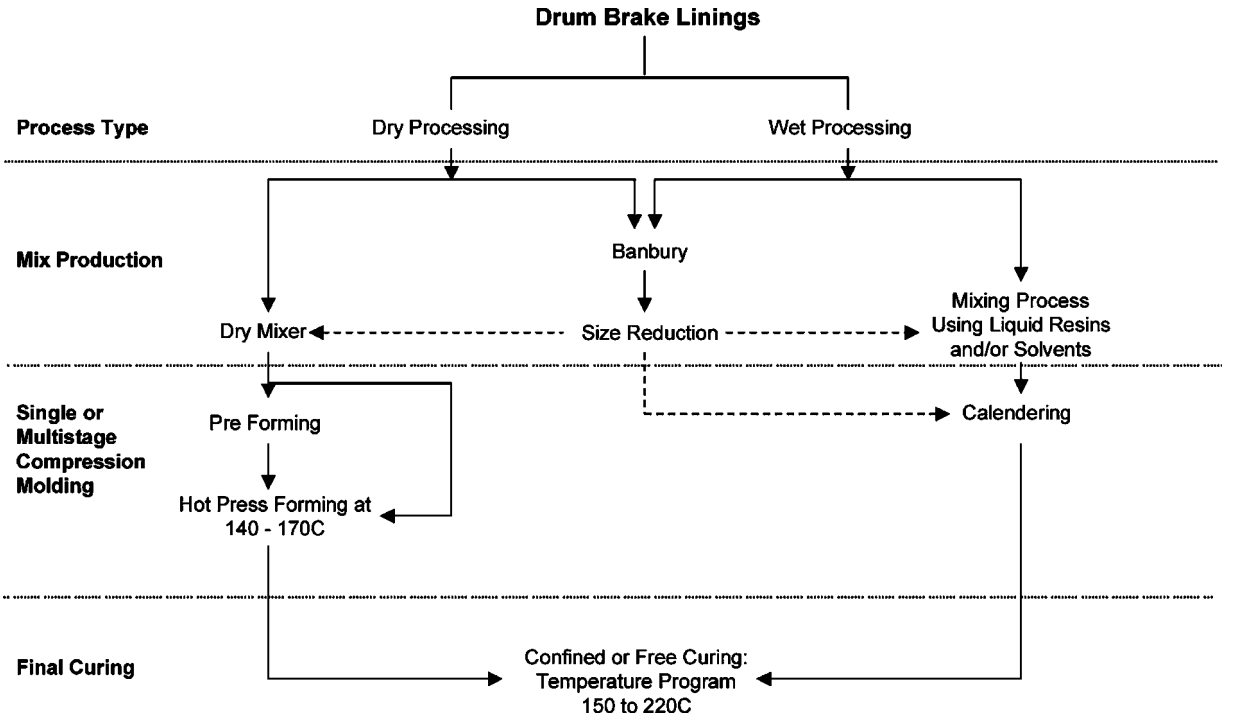


Fig. 10 Drum brake lining production processes.

applications, the drilled linings are riveted onto the carrier shoe with multiple metal rivets. Some manufacturers use integral molding onto a carrier shoe to produce drum linings designed for thermally demanding applications. Figs. 9 and 10 summarize the various disk pad and drumlining manufacturing processes.

Postcure finishing

Disk brake pads and drum linings are finish ground to the final dimensions after the oven cure process. Sometimes, the surface of disk pad products are scorched or seared under the influence of heat applied in the 250–700°C range. This burns off the organics in the pad surface and helps the initial friction properties of brake pads. Scorching can be accomplished by a variety of methods—direct gas flame, contact with an electrically heated hot plate, or radiant heat from a heated surface. Subsequently, the parts are painted and stenciled with product information prior to packaging.

CONCLUSIONS

Friction material technology has gradually evolved over the past century driven primarily by increasing demands on end-use performance or environmental considerations. The recent introduction of FMVSS 135 has placed increased demands on automobile brake friction materials. In addition to higher friction performance, a more general requirement exists for increasing product life to help extend maintenance periods. These requirements are in addition to the basic requirements of excellent NVH performance at the lowest product cost.

The main limitation in developing friction materials that satisfactorily meet customer needs is the lack of basic understanding of complex relationships between various friction material structures and compositions, and their interactions with brake system hardware and operating conditions. These areas offer major opportunities for research in addition to an understanding of basic third-body tribological mechanisms. A better understanding of the correlation between friction material composition and characteristics and dynamometer/vehicle performance will help in increasing the efficiency of the friction product development process. Artificial intelligence and other computational productivity tools have been playing an increasing role in complementing conventional compounding techniques to give companies a competitive edge.[10,11] Improved and efficient bench tests, which better correlate with vehicle performance, are needed to speed up screening of formulations.

In addition to product improvements, continuous improvement of processes to reduce scrap, improve quality, and reduce production time and energy requirements is essential to insure future competitiveness

among friction material manufacturers. Several companies are aggressively applying lean manufacturing and six-sigma techniques to stay competitive.

In the longer term, alternative automotive braking methods may play an increasing role. Examples of such technologies are electric and/or magnetic brakes. The increased adoption of electric and hybrid propulsion systems will greatly influence braking requirements. The need to develop cost-effective friction material solutions for these applications will pose further challenges.

REFERENCES

1. Murdie, N.; Ju, C.P.; Don, J.; Wright, M.A. *Carbon-Carbon Matrix Materials in Carbon-Carbon Materials and Composites*; Buchley, J.D., Edie, D.D., Eds.; Noyes Publications: New Jersey, 1993.
2. Spurgeon, W.M.; Spencer, A.R. Reliability and durability of automotive friction materials. Bendix Tech. J. **1969**, *2* (3), 57.
3. Burwell, J.T., Jr. Survey of possible wear mechanisms. Wear, **1958**, *1*, 119.
4. Anderson, A.E. Friction material performance issues. Symposium on Fibers in Friction Materials, Atlantic City, NJ, Oct 7–8, 1987.
5. Sibilia, J.P., Ed. *Materials Characterization and Chemical Analysis*; VCH Publishers: New York, 1996.
6. Shibata, K.; Goto, A.; Yoshida, S.; Azuma, Y.; Nakamura, K. Development of brake friction material. International Congress and Exposition, Detroit, Ml, Mar 1–5, 1993.
7. Abendroth, H. Advances in brake NVH test equipment. Automotive Eng. Int. **1999**, Feb, 60–63.
8. Halberstadt, M.L; Mansfield, J.A.; Rhee, S.K. Effect of potassium titanate fiber on the wear of automotive brake linings. International Conference on Wear of Materials, St. Louis, MO, Jul 2–6, 1977.
9. Melcher, B.; Faullant, P. A comprehensive study of chemical and physical properties of metal sulfides. Proceedings of the 18th Brake Colloquium and Engineering Display, San Diego, CA, Oct. 1–4, 2000.
10. Shao, X. Artificial intelligence for friction material development. Brake Colloquium and Engineering Display, San Francisco, CA, Oct, 1998.
11. Kato, T.; Soutome, H. Friction material design for brake pads using databases. STLE Tribol. Trans. **2001**, *44*, 137–141.

Fuel Cell Membranes

Andrew M. Herring
Department of Chemical Engineering, Colorado School of Mines, Golden, Colorado, U.S.A.

INTRODUCTION

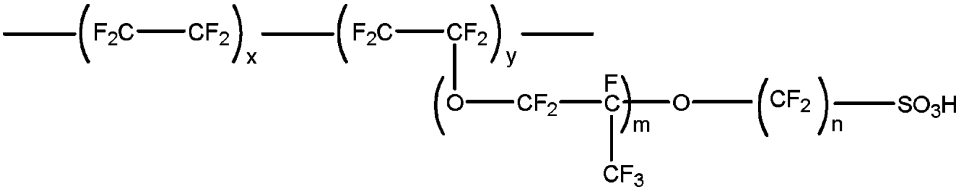
One of the driving forces of the recent interest in proton exchange membrane (PEM) fuel cells (FCs), is the dramatic improvement in performance that has been achieved in commercially available PEM. The membrane in PEM FCs must allow the transport of a high flux of protons while having low reactant crossover, be stable to oxidative attack, be electrically insulating, and be of desirable mechanical properties. Fast start-up and high current densities of several amperes per square centimeter are easily achieved with lifetimes of over 60,000 hr of operation. Platinum must currently be used as the catalyst for both the anode, fuel oxidizing, and the cathode, oxygen reducing electrodes, at these low operating temperatures. Because some oxygen crossover is inevitable, oxygen reacts at both the electrodes to form hydroxyl radical as a by-product. The resultant hydrogen peroxide produced at the membrane electrodes is thus significant and so the materials from which the membrane can be constructed are limited to oxidatively stable fluorinated and fully aromatic polymers. The current state-of-the-art membranes are fabricated from perfluorinated sulfonic acid (PFSA) polymers, such as Nafion®. High proton fluxes are obviously achieved by manufacturing membranes as thin as possible, and thus most state-of-the-art membranes are reinforced with inert fillers or scaffolds. These materials only have useful, high proton conductivities when fully hydrated limiting their usefulness to temperatures below 100°C. These hydrated sulfonated systems are well understood in terms of proton transport, water diffusion, and electro-osmotic drag (transport of water with the protons under FC operation). Ion transport between sulfonic acid groups is highly dependent on the bound and free water associated with those sites and may be described in terms of both a proton hoping, Grotthus mechanism and a Vehicle mechanism in which the protons move in association with one or more water molecules leading to electroosmotic drag. Water management issues with these materials are therefore crucial to a successful FC design as water is being constantly transported from the anode, which tends toward dry-out, to the cathode where it is being produced by the electrochemical reaction leading to flooding. Other desirable properties of an ideal PEM would include the ability to operate at moderately elevated temperatures, 120–200°C, with no external humidification, and to be manufactured inexpensively. Operation of the PEM FC at temperatures above 100°C is desirable as the platinum catalyst is severely poisoned by carbon monoxide, a by-product of reforming hydrocarbon fuels to hydrogen, below 100°C. Current research therefore is focused on developing PEM capable of operation at these moderately elevated temperatures in the absence of external humidification. Other active areas of research at the current time are concerned with maximizing the lifetimes of PEM and reducing their cost of fabrication. Further resources describing this topic area are available.[1–6]

MEMBRANE CHARACTERISTICS

The initial membranes developed for PEM FC were for space applications where the emphasis was solely on performance and durability.[7] The recent excitement with PEM FC is that they may one day be widely be applied in automotive applications, where in addition to high performance and durability, the membrane must also be inexpensive.[4,5] The current well-developed PEM FC technology is based on PFSA polymer membranes as the electrolyte. All the currently used PFSA membranes are composed of carbon–fluorine backbones with pefluorinated ether side chains containing sulfonic acid groups. (Table 1).[8] The processes occurring in a PEM FC are shown in Fig. 1. The power produced by a FC is a product of the cell voltage and the current produced. The ideal standard voltage of a H_2/O_2 FC is 1.229 V at 298 K. In reality various losses determine the current that can be drawn at useful power densities (Fig. 2). The biggest single loss in the system is attributed to slow reaction rates at the platinum cathode. The catalytic activity of oxygen reduction is high in the PFSA electrolyte because the sufonic acid groups do not adsorb on the platinum surface.[8] Fast reaction kinetics at both the electrodes is also enhanced as long as the coating electrolyte film is optimized because the solubility of the reactant gases, H_2 and O_2, are high in the electrolyte. Losses attributable to the membrane are reactant crossover and short-circuit voltage losses as

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120025529

Table 1 Commercial PFSA membranes by producer



Structure parameter	Trade name and type	Equivalent weight	Thickness (μm)
$m = 1$; $x = 5$–13.5;	Du Pont		
$n = 2$; and $y = 1$	Nafion 120	1200	260
	Nafion 117	1100	175
	Nafion 115	1100	125
	Nafion 112	1100	80
$m = 0$, 1 and $n = 1$–5	Asahi glass		
	Flemion®-T	1000	120
	Flemion®-S	1000	80
	Flemion®-R	1000	50
$m = 0$; $n = 2$–5;	Asahi Chemicals		
and $x = 1.5$–14	Aciplex-S	1000–1200	25–100
$m = 0$; $n = 2$;	Dow Chemicals		
$x = 3.6$–10	Dow	800	125

(From Ref.[8].)

well as ohmic losses associated with membrane and membrane catalyst's interface resistances.

Proton Transport

In an idealized PEM FC only protons would be transported across the membrane (Fig. 1). Proton conductivities of membranes are measured in siemens per centimeter and are typically on the order of 0.1 S/cm for fully hydrated PFSA membranes.[2,3] High proton conductivities of membranes depend on a high ion exchange capacity, which is expressed usually as the inverse equivalent weight that ranges from 1100 to 1000 or less; this is equivalent to an ion exchange capacity of 0.8–1.1 meq/g. The other critical factor for high proton conductivity is the membrane thickness, the thinner the membrane the higher the flux of protons. A typical unsupported PFSA membrane used in an FC membrane electrode assembly (MEA) is 150–170 μm in thickness. Water content is also critical, the water content of the membrane is expressed as λ, the moles of water per mole of acid group, and is typically 21–23 for Nafion.[9] Two mechanisms of proton transport have been differentiated: in the first, the so-called, Grotthus mechanism, protons hop between fixed acid sites; in the second, the so-called, Vehicle mechanism, protons move through the membrane with associated water molecules. It is this Vehicle mechanism which is important to fully hydrated PFSA membranes.

Water Management

The transport of the water associated with protons is called electroosmotic drag and typical coefficients of H_2O/H^+ for PFSA membranes are 3 and 4 at 80°C (Fig. 1).[10] Electroosmotic drag is independent of the membrane thickness, proportional to the current density, and induces a water concentration difference between the anode and the cathode.[11] The electroosmotic drag coefficient, K_{drag}, increases with the membrane water content (Fig. 3).[6] At the high current densities at which PEM FC are run; therefore, significant amounts of water are being removed from the anode and transported to the cathode. As the PFSA membrane must be fully hydrated, this leads to significant water-management issues. One advantage of thinning the membrane is that back diffusion of water from the relatively wet cathode to the relatively dry anode will increase (Fig. 1). The diffusional flux of water, j_{H_2O}, from cathode to anode may be derived from Fick's law

$$j_{H_2O} = -D_{H_2O} \frac{C^c_{H_2O} - C^a_{H_2O}}{\Delta x}$$

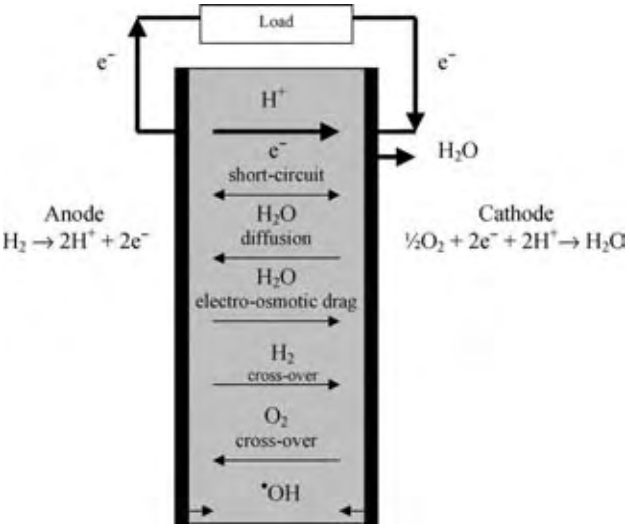


Fig. 1 Membrane processes occurring in a polymer electrolyte FC.

where $C^{c}_{H_2O}$ and $C^{a}_{H_2O}$ are the membrane water content at the anode and the cathode, respectively, Δx is the membrane thickness, and D_{H_2O} is the water diffusion coefficient in the membrane.[11] D_{H_2O} can be assumed to be a linear function of the water content λ, so that $D_{H_2O} = 7.1 \times 10^{-7} \times \lambda\,cm^2/sec$.[11] The thinness of FC membranes cannot be solely designed based on the need for high proton fluxes and high back diffusion, as thinner membranes will not be structurally robust, although state-of-the-art membranes are reinforced. As the membrane thickness is decreased reactant crossover increases. The permeability of oxygen and hydrogen is 10^{-11} and 10^{-10} mol/cm/sec/atm, respectively, corresponding to a loss of 1–10 mA/cm^2 or 1% for Nafion 117.[3,8] Both crossover and electrical shortening lead to decreases in cell efficiency, which are manifested in decreases in the open circuit voltage of the cell. These factors limit the practical thickness of a PEM. It is therefore necessary to humidify the PFSA membrane by methods other than back diffusion. The most common method of humidifying the membrane is to humidify the reactant gases using bubble humidifiers. Care must be taken to not over humidify the cathode, as too much water will result in liquid water flooding the electrode leading to transport losses (Fig. 2). Various internal humidification schemes have been proposed to transport water from the cathode to the anode, such as using porous carbon bipolar plates, current collectors, as were used in the phosphoric acid FC, and sewing wicks into the membrane to wick the water back to the anode. It has also been proposed to impregnate the membrane with Pt nanoparticles, which would catalyze the production of water from the crossover H_2 and O_2.[12] This has the added advantage of increasing the open circuit potential of the cell, but not the cell's efficiency. The Pt nanoparticles might also decompose hydroxyl radical (see below).

The most widely used model of the structure of water-swelled Nafion is that of Gierke.[13] In this model the polymer is assumed to form reverse micelles of radius 20 Å connected by small cylindrical pores 10 Å in length and 5 Å in radius. This model breaks down for all swelling states of the membrane and although modified to include more ordered structures, it has only been recently that a new more comprehensive picture has emerged from extensive small angle X-ray scattering (SAXS) and small angle neutron scattering (SANS) studies. It is now proposed that the structure comprises cylindrical or ribbon-like polymeric aggregates surrounded by ionic groups and water molecules.[14] In general, compared to other ionomers, Nafion has relatively wide, more separated, and less branched channels with good connectivity, small sulfonate separation, and a pK_a of ~6 (Fig. 4).

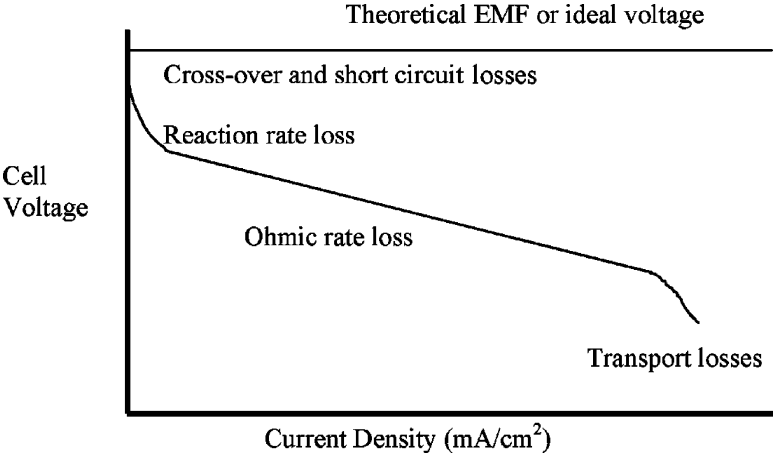


Fig. 2 Polarization curve for a realistic polymer electrolyte FC.

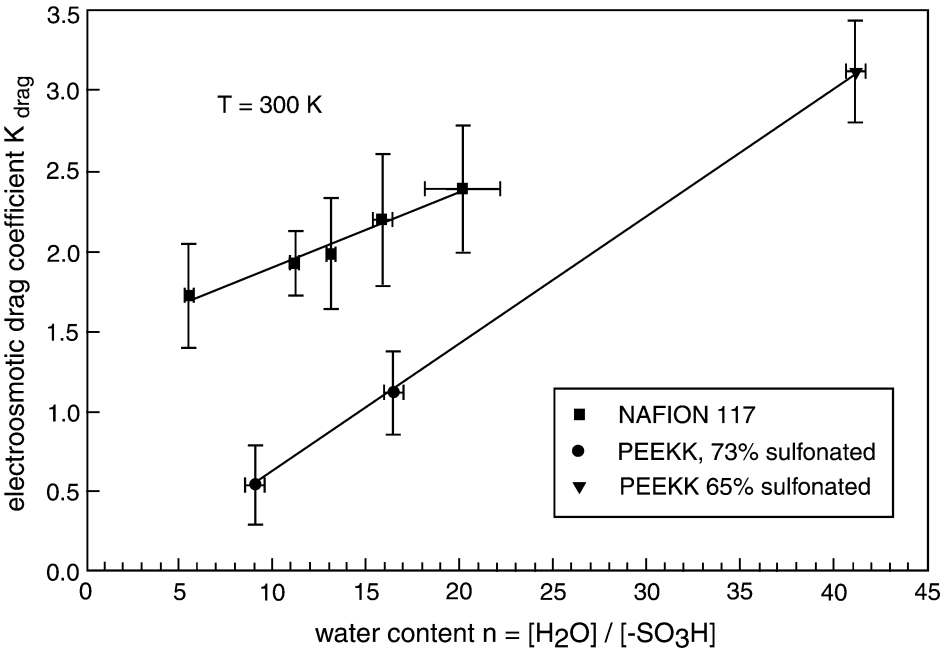


Fig. 3 Electroosmotic drag co-efficient obtained by electrophoretic-NMR as a function of the water content, $n = [H_2O]/[-SO_3H]$. (From Ref.[6].)

Oxidative Stability

Although the anode side of a PEM FC may be considered to be operating under reducing conditions, the cathode side is severely oxidizing. Not only is oxygen present and activated by the catalyzed cathode electrode, but a key intermediate in the reduction of oxygen, hydroxyl radical, is released at the cathode; additionally, as oxygen does diffuse across the membrane, hydroxyl radical is also formed at the anode. The PEM is therefore subjected to very harsh oxidizing and acidic conditions. This is the biggest limitation on the choice of materials that may be used to fabricate a PEM FC. A PFSA membrane can be imagined to decompose in the presence of hydroxyl radical by the following three idealized steps:[15]

$$R_f{-}CF_2COOH + {}^{\bullet}OH \rightarrow R_f{-}CF_2^{\bullet} + CO_2 + H_2O$$

$$R_f{-}CF_2^{\bullet} + {}^{\bullet}OH \rightarrow R_f{-}CF_2OH \rightarrow R_f{-}COF + HF$$

$$R_f{-}COF + H_2O \rightarrow R_f{-}COOH + HF$$

It should be noted that the products of this decomposition are water, carbon dioxide, and HF. While PFSA membrane FCs have been demonstrated for many thousands of hours, the flux of HF is significant enough so that uncoated metallic bipolar plates are precluded. Hard to machine graphite bipolar plates must be used or an electrically conducting corrosively resistive coating must be developed for easily fabricated metal bipolar plates. Lifetime studies of PEM FC membranes are usually expressed in terms of the decay in cell voltage while under operation. As the membrane is corroded by the harsh conditions of the cell, pinholes will inevitably develop leading to reactant crossover and decay in the cell voltage. Decay rates of less than $6\,\mu V/hr$ can now be obtained with state-of-the-art PFSA membranes.[16] Metal ion impurities in the membrane, such as iron or cobalt, dramatically affect the life of the membrane, as supplied Nafion is always boiled in nitric or sulfuric acid, to also ensure it is in the acid form, hydrogen peroxide, and DI water to clean the material prior to use. To test for oxidative stability, membranes are routinely boiled in Fenton's reagent in which an iron compound is used to generate a large concentration of hydroxyl radical. While PFSA membranes survive this test, many experimental membranes currently under consideration do not, and while the test is useful, it represents an unrealistically high concentration of hydroxyl radical than would be found in a FC. Another related failure mode of anhydrous Nafion is that it loses sulfonic acids groups above 165°C.

Mechanical Attributes

The membrane must be mechanically robust and also suitable for manufacture. One of the consequences of the need for a fully hydrated membrane is that the membrane will shrink and swell between operation and resting. This can have dire consequences for an FC stack that must be sealed if the membrane dimensions are radically changing between on and off conditions. So, ideally the membrane should not

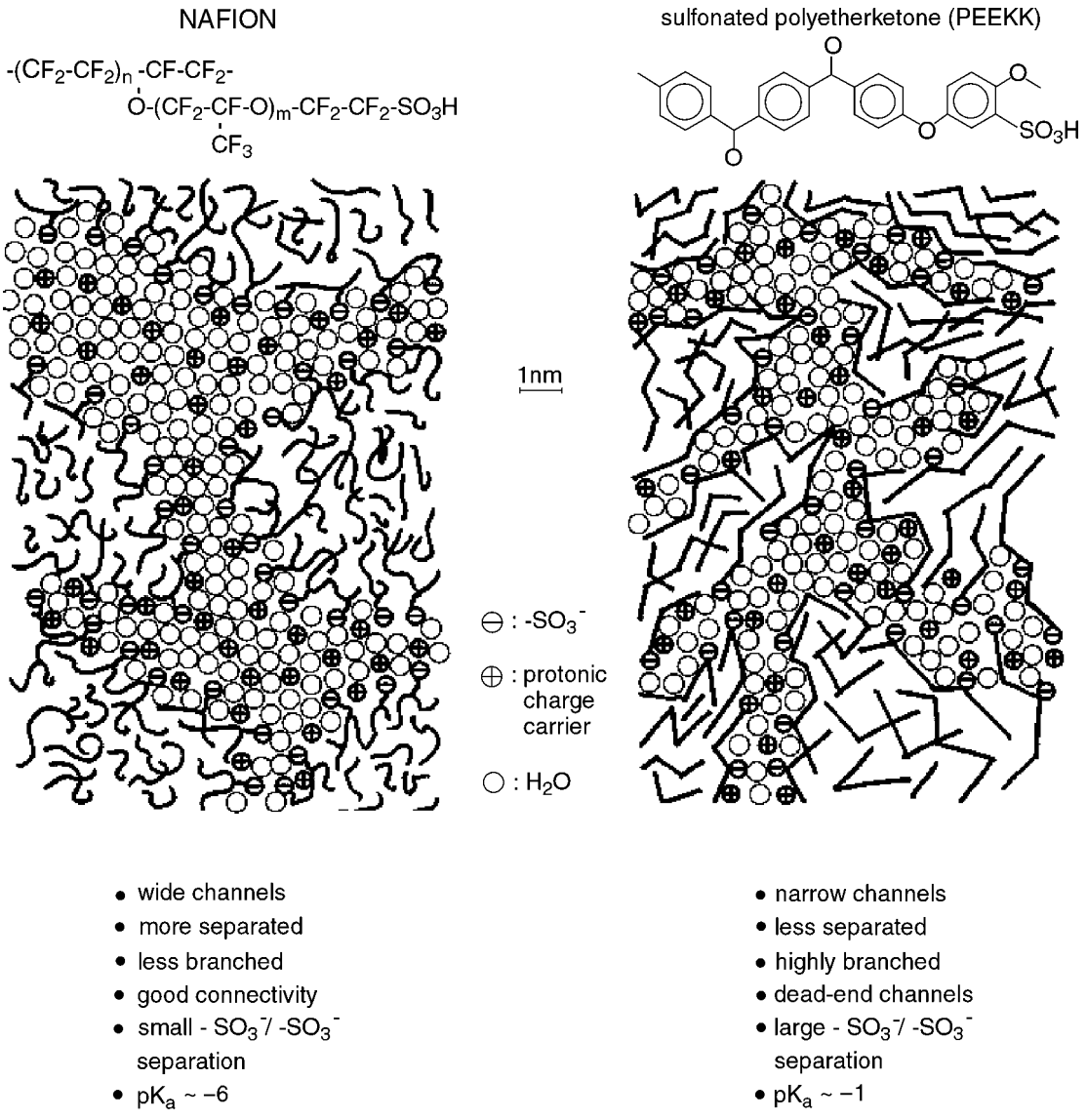


Fig. 4 Schematic representation of the microstructures of Nafion and a sulfonated PEK (derived from SAXS experiments) illustrating the less pronounced hydrophobic/hydrophilic separation of the latter compared to that of the former. (From Ref.[6].)

change dimensionally between water swollen and dry states. Nafion 117 typically shrinks 10% between wet and dry states.[17] Water also acts as a plasticizer in PFSA lowering the T_g and the tensile strength of the material. Nafion loses half of its tensile strength between dry and fully hydrated states.

HISTORY OF PEM FC

The proton membrane FC was discovered in the late 1950s at GE; the first PEM FC was constructed using phenol–sulfonic ion exchange membranes. The cells operated at 15 mA/cm^2 at 0.6 V for 500 hr.[7] The next significant breakthrough that came in the 1960s for the Gemini space program was the development of sulfonated polystyrene resins. After optimization, these membranes were capable of 300 mA/cm^2 at 0.6 V and 4000 hr of operation.[7] These membranes were limited in the oxidizing environment of the FC as the methylene CH bonds were too susceptible to oxidation. The next improvement was the introduction of Du Pont's Nafion PFSA membrane, which gave an operating time in excess of 10,000 hr. NASA funding for PEM FCs ended in the mid-1970s and while interest continued to grow in the intervening decades, large increases in performance have only begun to be achieved in recent years as the emphasis and interest has switched to the potential for PEM FCs in automotive applications. Currently, the biggest barrier to commercialization is cost, while no FCs are currently mass produced and so it is hard to get a true idea of what a realistic cost would be,

FCs must cost around \$45/kW to be competitive with the internal combustion engine. Perfluorinated sulfonic acid membranes currently cost \$100–200/KW resulting in PEM FC prices estimated to be of the order of \$1000/KW.

MEMBRANES FOR HIGH-TEMPERATURE AND HIGH-PERFORMANCE OPERATION

There is considerable interest in developing PEM that will allow FC operation at elevated temperatures, 120–200°C. The two primary reasons are to allow the use of carbon monoxide contaminated H_2 from reforming hydrocarbons and to facilitate heat transfer. The platinum catalyzed anode is poisoned by as little as 20 ppm of carbon monoxide at 80°C, but can tolerate 1000 ppm at 130°C and 3% at 200°C.[3,8] Very pure hydrogen for PEM FC operating at 80°C is produced via a costly multistep process. In fact, the reformer costs as much as the FC. Raising the temperature of operation of the FC to 120°C or above will dramatically decrease the complexity and expense of the reformer. An FC run at 120°C will seamlessly fit into the existing automotive radiator technology facilitating cooling. At these and higher temperatures, stationary combined heat and power applications become attractive. Other advantages of higher-temperature operation include faster electrode kinetics because of enhanced gas transport, and higher reliability due to system simplification with only one phase of water. As can be seen from Fig. 5, no polymer seriously outperforms Nafion below 100°C and only phosphoric acid-doped benzimidazole and polybenzimidazole (PBI) have appreciable proton conductivity at higher temperatures. The field of emerging materials for higher PEM FC operation has been recently reviewed.[2,6,8] In addition to the high temperature limitations of PFSA, the cost of fluorinating the polymer has often been invoked as the reason for the materials relative expense. The manufacturers obviously dispute this quoting the relatively low volumes of material as the reason for the high cost. Fluorination remains a process of environmental concern and the search is on for a greener PEM.

Reinforced Perfluorosulfonic Acid Membranes

Perfluorinatedsulfonic acid membranes are limited to temperatures of operation below 100°C because they must be fully hydrated. By reinforcement, PFSA membranes may be made much thinner allowing enhanced back diffusion of water. Perfluorinatedsulfonic acid membranes enforced with polytetrafluoroethylene (PTFE) or Teflon® are used in industrial electrochemical membranes, but these are much too thick for FC applications. Goretek® is stretched PTFE, which forms the basis for the Gore-select® membranes in which Goretek® is impregnated with Nafion. Gore-select® membranes are transparent and have ionic conductivities

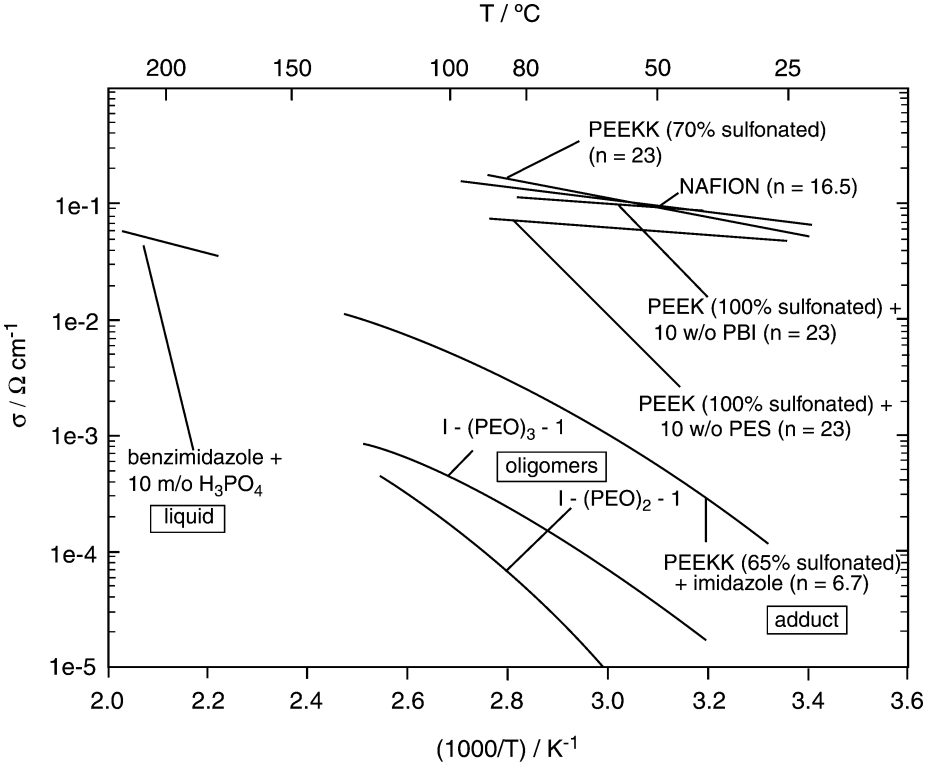


Fig. 5 Proton conductivity of different fully hydrated acidic polymers and a liquid, an adduct, and an oligomer containing heterocycles as proton solvent. (From Ref.[6].)

and equivalent weights comparable to Nafion while being an order of magnitude thinner, on the order of 5–20 μm. In addition, the gore-select membranes are dimensionally stable between wet and dry states shrinking only 2% and losing little tensile strength.[17] Recent advances with these membranes resulted in significant stabilization to peroxide attack. The membranes do indeed perform well under relatively dry conditions showing little degradation in performance at 80°C or 95°C with relative humidities decreased from 100% to 30%. A Gore MEA has been demonstrated at 120°C giving a current density of 250 mA/cm^2 at 0.6 V compared to 20 mA/cm^2 for Nafion 117 under the same conditions. Asahi Glass has developed a reinforced Flemion using 2–5 wt% dispersed PTFE fibrils.[3,8] This reinforcement again allowed improved mechanical strength of the membrane and decrease in thickness of the membrane from 100 to 50 μm. Du Pont has made available a PTFE reinforced PFSA, Nafion 324 and 417, for many years for industrial electrochemical applications but these materials are too thick for FC applications. Du Pont is now offering a strengthened thinner Nafion for FC applications, but little is known about this new ionomer.

Modified PFSA Polymers

Various additives have been added to PFSA membranes to either replace the water or to retain the water at higher temperatures. Nafion has been swollen with phosphoric acid and a conductivity of 0.05 S/cm has been achieved at 150°C. Unfortunately, the anode fails after a short period of time in these membranes so no successful FC tests have been run.[8] 1-Butyl,3-methyl imidazolium triflate and tetrafluoroborate have also been used to swell Nafion giving a conductivity of 0.1 S/cm at 180°C. Nafion has also been swollen with heterocycle solutions, imidazole, and imidazolium salts in trifluoroacetate and trifluoro methane sulfonate solution, although the reported conductivities are more modest, 10^{-3} S/cm at 100°C.[8] Acetic acid and tetra-*n*-butylammonium chloride solutions of the heteropoly acid (HPA), 12-phosphotungstic acid (PTA), have all been used with Nafion giving improved FC performance at 110°C and 120°C, respectively, vs. the undoped PFSA.[18]

As mentioned above, platinum nanoparticles have been used to generate water in the membrane in situ from hydrogen and oxygen crossover. TiO_2 and 3 wt% SiO_2 nanoparticles have also been used to retain water in the membrane under hot and dry conditions. There has been considerable interest in composite recast PFSA membranes using inorganic super acids and oxides. The HPAs are a large and diverse set of super acid inorganic oxides of which a few can be obtained commercially and have been extensively studied. It should be pointed out that the generalizations in the literature to date are only based on the three or four commercially available HPAs and not on the many hundreds of structures that have been reported in the literature since the late 19th century. The common commercially available HPAs, PTA, 12-sillicotungstic acid (STA), and 12-phosphomolybdic acid (PMA), have the Keggin structure (Fig. 6).[19] The theory is that if the pK_a of the nonvolatile acid is higher than the pK_a of the PFSA electrolyte, it will solvate the proton of the stronger sulfonic acid, creating the ionic network clusters for conduction.[2] In addition, the free acids in the electrolyte pores will also conduct protons through self-ionization. Recast HPA-composite membranes are fabricated as follows. A commercial 5% Nafion solution is reduced by 50% and mixed with the appropriate concentration of HPA. Cast membranes are then produced by evaporation of the solvent from various amounts of solution to give membranes of thickness 15–500 μm. The films are then dried at 45–60°C for 24 hr and then in an oven at 130–170°C for 4 hr. The resultant films were peeled off the casting plate and washed and stored in deionized water.[2] All the HPA/Nafion composite membranes swell more in water than in Nafion in the order Nafion 117 (27%) < Nafion/STA (60%) < Nafion/PTA (70%) < Nafion/PMA (95%) at 90°C. The ionic conductivites are also greater than for Nafion in the order Nafion 117 (1.3×10^{-1} S/cm^2) < Nafion/PMA (1.5×10^{-1} S/cm^2) < Nafion/PTA (2.5×10^{-1} S/cm^2) < Nafion/STA (9.5×10^{-1} S/cm^2) at 90°C. The FC currents at 0.6 V increase in the order Nafion 117 (640 mA/cm^2) < Nafion/STA (695 mA/cm^2) < Nafion/PTA (810 mA/cm^2) < Nafion/PMA (940 mA/cm^2) at 90°C. These membranes show less deformation than Nafion, but the tensile strength of the membranes is, however, compromised.[2] Further improvements have been achieved in the Nafion/STA by the use of thiophene. Heteropoly acids have also been immobilized in silica; a PMA/silica Nafion-composite membrane has been reported to have a proton conductivity >0.3 S/cm^2 at 90°C.[8]

Another inorganic proton conductor, zirconium phosphate, $Zr(HPO_4)$ has also attracted considerable interest in Nafion composite membranes. Zirconium phosphate, a very hydroscopic insoluble solid, has been electrochemically and chemically precipitated in the pores of PFSA membranes as well as recast from Nafion solution containing zirconium ions.[2] The preparation of recast Nafion zirconium phosphate membrane involves swelling a membrane with a 1 : 1 methanol–water solution, which is then dipped in zirconylchloride followed by rinsing and storing in 1 M phosphoric acid. Current densities using Nafion 115 of 1.5 A/cm^2 at 130°C and 3 bar pressure using H_2/O_2 have been achieved with these membranes. These

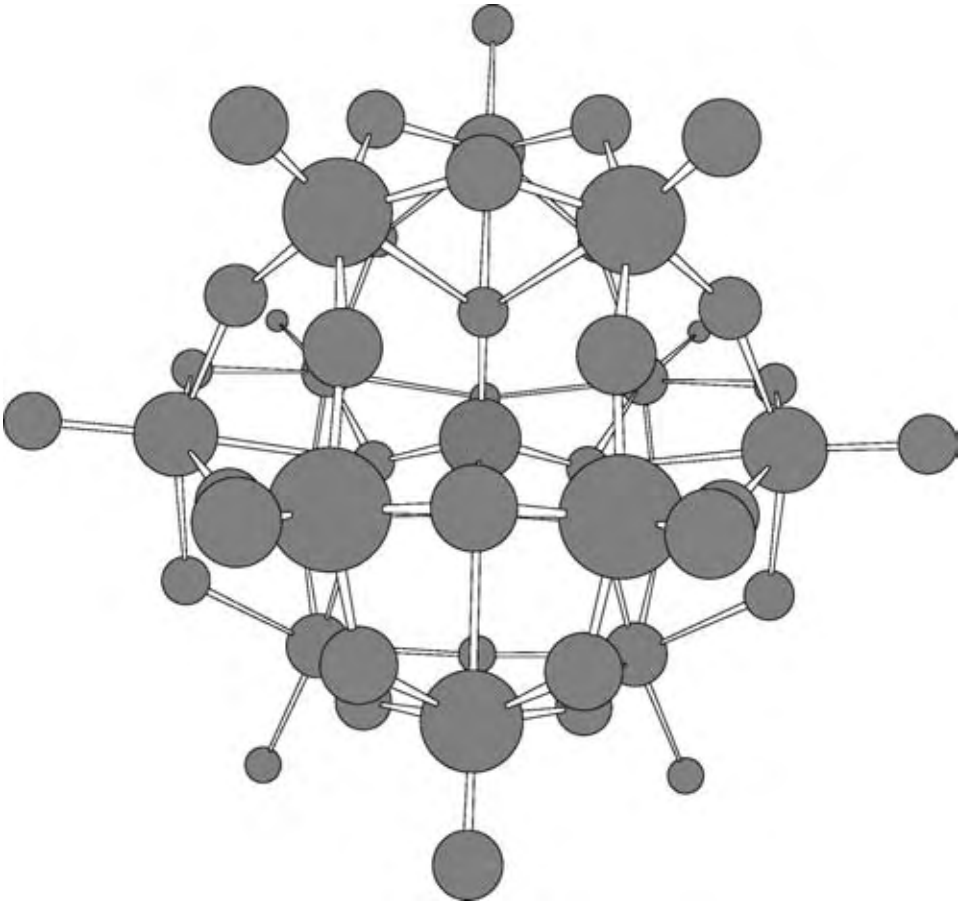


Fig. 6 Keggin anion structure, $[XM_{12}O_{40}]^{(3 \text{ or } 4)-}$ (X = Si(4–) or P(3–), M = W or Mo), atomic co-ordinates. (From Ref.[28].) *(View this art in color at www.dekker.com.)*

improvements have been attributed to the hydroscopy of the zirconium phosphate and the reduction of free spaces in the pores of the PFSA promoting capillary condensation and thus water retention.

A major modification to the PFSA electrolyte polymers is the replacement of the sulfonic acid group with a bis[(perfluoroalkyl)sulfonyl] imide acid group, i.e., $-SO_3H$ is replaced with $-SO_2NHSO_2CF_3$. This modified acid group is more thermally stable and has greater gas phase acidity than the sulfonic acid group. This ionomer can be copolymerized with tetrafluoroethylene to produce a structure very similar to Nafion except with a different proton conducting functionality. The observed dependencies on water absorption and ionic conductivity on RH and EW for the bis[(perfluoroalkyl) sulfonyl] imide polymers are similar to Nafion.[2]

Sulfonated Aromatic Polymers

Partially perfluorinated polymers

The first PEM were constructed from sulfonated polystyrene, but these materials were not of sufficient oxidative stability to survive the FC environment. Chemically, the bond strength for C–F is 485 kJ/mol higher than for C–H bonds (typically 350–435 kJ/mol). Ballard has developed a series of membrane materials based on sulfonated copolymers incorporating α,β,β-trifluorostyrene known as Ballard advanced membrane third generation (BAM3G). These materials have low EWs 320–920 g/mol and have very high water retentions. Composite BAM3G membranes have been fabricated using sulfonated BAM3G incorporated into a porous film with ion exchange and ethylene-based monomeric units. This should improve the membranes mechanical and dimensional stability but no information is available.[2] Perfluorinated grafted membranes may be prepared by radiation grafting. The perfluorinated film is irradiated and reacted with a sulfonated polystyrene to produce the electrolyte membrane. In a similar approach, trifluorovinyl aromatic monomers may be grafted to a preformed polymeric base film and modified to incorporate ion exchange groups.

Nonperfluorinated composite membranes

The basic idea of these membranes is to take a polymer that has been developed for use in extreme temperatures, atmospheres, or corrosive environments and to convert it into a PEM usually by sulfonation (Fig. 7). Sulfonation may be accomplished as follows: 1) by

direct sulfonation in sulfuric acid or chlorosulfonic acid; 2) by lithiation–sulfonation–oxidation; 3) by chemically grafting a group containing a sulfonic acid to a polymer; 4) by graft copolymerization using a high-energy radiation source; and 5) by synthesis using monomers bearing sulfonic acid groups. Purely hydrocarbon polymers, such as poly-*p*-phenylene, have excellent resistance to oxidation and thermal stability but are stiff-ridged rod polymers that are hard to process. However, polyelectrolytes based on poly-*p*-phenylene have recently appeared.[8] More often, aromatic polymers with –O–, polyether; –(C=O)–, polyketone; –S–, polysulfide; or –(SO_2) –, polysulfone linkages are employed, as these materials are more easily processed into membrane films (Figs. 7A–D). Ether linkages are particularly flexible and are the most widely used. Polyetheretherketones (PEEK) are thermoplastic polymers with an aromatic backbone in

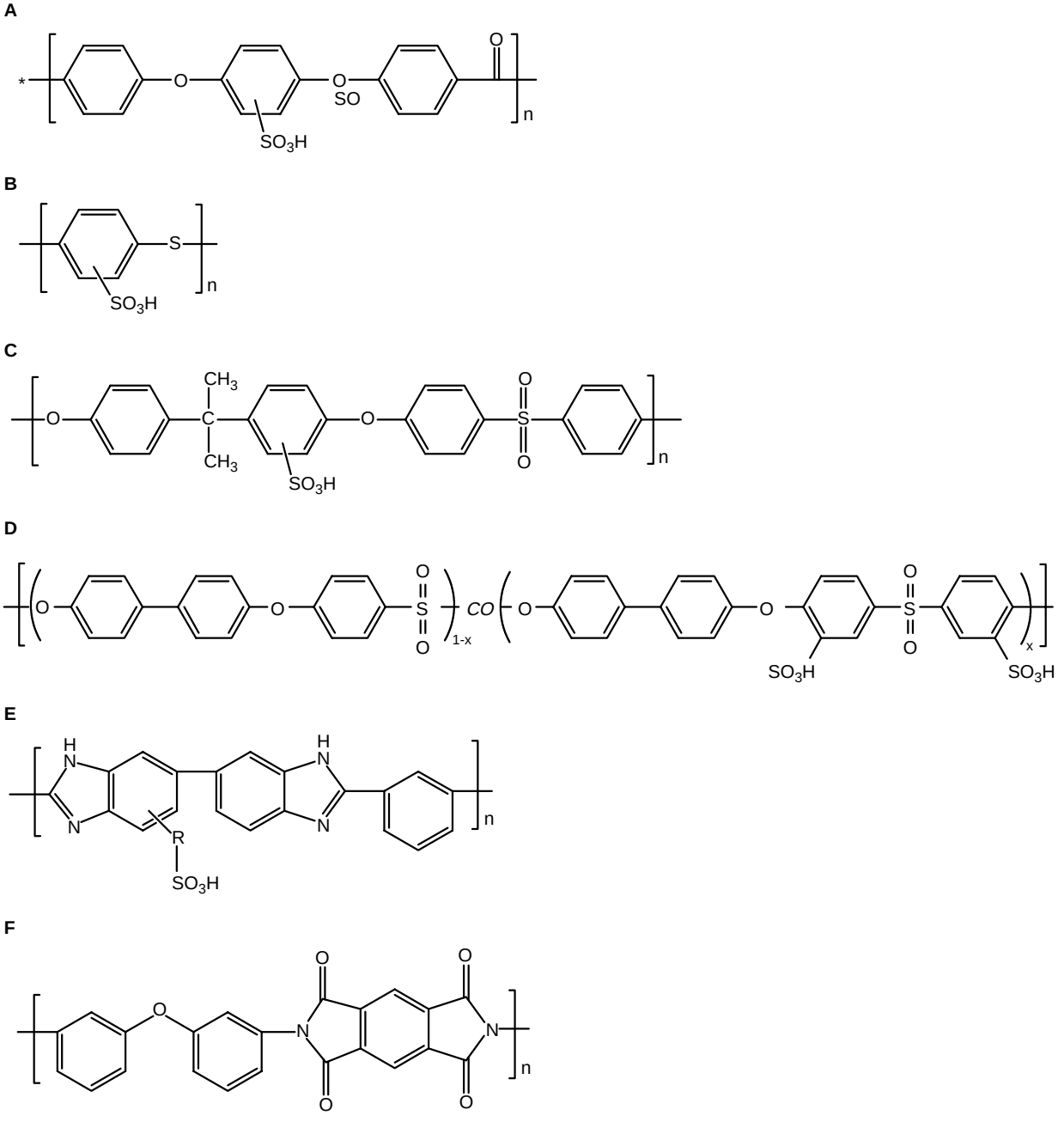


Fig. 7 Chemical structures of some sulfonated polymers and a polyimide: (A) sulfonated polyetheretherketone, PEEK, PSF; (B) sulfonated polyphenylenesulfide, PPS; (C) sulfonated polysulfone; (D) poly(4,4′-biphenol) (4,4′-dichlorodiphenyl sulfone), BPSH–XX (XX is mol% of disulfdonated units); (E) sulfonated polybenzimidazole, PBI; (F) polyimide.

which 1,4-disubstituted phenyl groups are separated by –O– and –CO– linkages.[2] Various other configurations, such as polyetherketone (PEK), polyetherketoneetherketone, or polyetheretherketoneketone (PEEKK), are also available with similar properties. These polymers are easily sulfonated up to 60% by treating, e.g., PEEK, with concentrated sulfuric acid with a SO_3 concentration of 82.4% at room temperature to give sulfonated-PEEK (SPEEK). Membranes can be cast from a *N*-methylpyrrolidone solution prepared at 130°C under vacuum. The membranes are reinforced with glass fiber. The conductivities of SPEEKK and SPEEK approach those of Nafion (Fig. 5). Sulfonated-PEEK loses water and when dehydrated, like all sulfonated polymers is susceptible to loss of sulfonic acid groups, limiting its use in high-temperature FC. A current density of 620 mA/cm^2 at 0.6 V has been achieved for a woven SPEEK membrane at 110°C.[2] In general, comparison to Nafion, SPEEK has relatively narrow, less separated, and highly branched channels with many dead-end channels, large sulfonate separation, and a pK_a of ~1 (Fig. 4). Composite SPPEK/inorganic membranes have been prepared using HPAs, boronphosphate, zirconium phosphate, sulfenylphosphonate, and aminophenyl functionalized silica in an effort to increase the conductivity of the material.[2] Proton conductivities of 0.1 S/cm at 100°C have been achieved with SPEEK–HPA composite films. Poly-*p*-phenylene sulfone has too high a softening point, 520°C to be processable, but the use of an ether linkage as in the Udel polysulfone brings the T_g to 190°C (Fig. 7C). There is a lot of recent interest in a class of block copolymer polysulfones in which a sulfonated aromatic group is used as a monomer to form a copolymer, such as BPSH-40 (Fig. 7D). The reason for this is that postsulfonation occurs at the most reactive, least stable site, so by using a sulfonated monomer in which the sulfonate group is in the most stable least reactive position the hope is to make more stable sulfonated polysulfones for high-temperature PEM applications.[20] Furthermore, the block polymers morphology is controlled such that there are hydrophobic regions and hydroscopic regions. This morphological control is used to design a polymer with desirable mechanical properties and greater oxidative stability. Proton conductivities as high as 0.17 S/cm have been obtained with sulfonation rates as high as 60%, but these materials swell excessively in water so the optimum 40% sulfonation, hence BPSH-40, is being further developed, which takes up 40 wt% water. Composites of both PTA and zirconium hydrogen phosphate have been made for high-temperature applications.[21] While these inorganic super acids improve the materials morphological, dimensional, and thermal stability, only PTA improves the materials proton conductivity, but its retention in the structure is not optimum.

Polybenzimidazole is a polymer used to fabricate fire resistant clothing as well as a membrane for use in blood dialysis and reverse osmosis in high-temperature and harsh environments (Fig. 7E). The material is a basic polymer, pK_a = 5.5, with excellent thermochemical stability, mechanical properties, low hydrogen permeability, and is of relatively low cost. Direct sulfonation of PBI results in a brittle material. Polybenzimidazole may be sulfonated by first deprotonanting a nitrogen in a benzimidazole ring followed by reaction with aryl or alkyl sulfones. The longer the alkyl linking group the higher the water uptake of the sulfonated PBI; when the alkyl chain is propyl, the conductivity of the material is maintained above 100°C, but, of course, the methylene hydrogens would not be expected to survive a working FC environment. Polyimides are also basic thermoplastic polymers that may be sulfonated (Fig. 7F). These materials show very little sensitivity to external conditions, the number of water molecules per ionic group is independent of equivalent weight, the conductivity is not related to the water content, the monomers contain an internal structure of ionic domains, and the microstructure of the polymer is not spherical ionic domains as in other ionomers.[2]

Acid Doped Basic Polymers

Another way to create a polymer PEM from a basic polymer is to dope it with an amphoteric acid, which acts as both a donor and an acceptor in proton transfer and therefore allows proton migration.[8] The most widely studied of these materials are the phosphoric acid-doped PBI membranes. Phosphoric acid-doped PBI has been shown to exhibit high conductivities up to 200°C, low gas permeabilities, excellent oxidative and thermal stability, and nearly zero electroosmotic drag as the proton conduction is by the Grotthus, proton hoping, mechanism. H_3PO_4–BPI membranes are cast from phosphoric acids solution to give 5 mol of H_3PO_4 per PBI repeat unit.[8] H_3PO_4 interacts only with the imidazole group but does not protonate it.[2] Polybenzimidazole may be doped with other acids the conductivity follows the order $H_2SO_4 > H_3PO_4 > HCO_4 > HNO_3 > HCl$. Recently, new H_3PO_4–BPI materials have been synthesized by the use of polyphosphoric acids (PPA) as both solvent and polycondensation reagent in a process in which the PAA polymerization solution is directly cast without isolation are followed by a sol–gel transition induced by the hydrolysis of the PPA into H_3PO_4.[22] These PPA–PBI materials have 15–25 mol of H_3PO_4 per PBI repeat unit, yet still have 1.0–2.5 MPa tensile strength. All acid-doped PBIs have high proton conductivities at higher temperatures, >140°C, but relatively low conductivities at ambient temperatures, the conductivity rapidly rising

between 80°C and 140°C. This apparent slow cold start problem and the still not fully understood effects of water on the material are major hurdles to the widespread use of these membranes in PEM FC; even so the performance of the PPA–PBI as shown by proton conductivity of 0.21 S/cm and 500 mA/cm at 0.6 V in a H_2/O_2 FC is impressive.

Inorganic Proton Conductors

There have been an increasing number of reports of using HPAs as the proton-conducting element in PEM. Solid HPA exhibit proton conductivities of 0.2 S/cm at room temperature, but this value rapidly falls with increasing temperature for HPA, which do not readily retain water. The HPA are water soluble and so must be immobilized in a matrix. Sol–gel approaches using hybrid organosilanaes have been used to produce membranes with conductivities as high as 10^{-2} S/cm at temperatures up to 160°C and temperature stability up to 400°C.[23] The group of compounds $M^{IV}(RXO_3)_2 \cdot n$S where M is a tetravalent metal such as Zr, Ti, Ce, Th, or Sn; R is an inorganic or organic group; and S is a solvent of which $Zr(HPO_4)_2$ is a member and exhibits good proton conductivity up to 300°C.[8] These materials may be cast as glassy plates or films and the conductivities may be enhanced with silica or alumina. Organic derivatives of α- or γ-zirconium phosphates where the HPO_4 group of the α- is replaced with a O_3POR group and the $O_2P(OH)_2$ group of the γ- is replaced with a O_3POR or $O_2PR'R$ group are being developed for PEM FC applications when the R group contains a proton-generating function, such as –COOH, $–PO_3H$, $–SO_3H$, or NH_3^+.[24] Conductivities of 2×10^{-2} S/cm at 105°C and 85% RH have been achieved with α-$Zr(O_3PC_6H_4SO_3H) \cdot 3.6H_2O$.[24] Another group of inorganic proton conductors that has recently been demonstrated impressively in an elevated PEM FC are the hydrogen sulfates, $MHXO_4$, where M is a large alkali metal, such as Rb, Cs, or NH_4^+, and X is S, Se, P, or As.[1,25] $CsHSO_4$ exhibits an ordered arrangement of hydrogen bonds at room temperature and on slight heating becomes structurally disordered and transforms from a monoclinic to a tetragonal structure at 141°C. Accompanying this transformation is an increase in proton conductivity of two to three orders of magnitude reaching values as high as 10^{-2} S/cm.[25] A current density of 50 mA/cm^{-2} at 0.6 V has been achieved at 235°C over 100 hr of operation using humidified H_2 and O_2.[25]

MEMBRANES FOR DIRECT METHANOL FC

Methanol as a liquid is an ideal fuel for an FC, but unfortunately its reaction at the anode is slow so power densities from DMFC are typically low. However, the lower temperature DMFCs are attracting considerable interest for portable applications. The anode reaction of the DMFC is:

$$CH_3OH + H_2O \longrightarrow 6H^+ + 6e^- + CO_2$$

The reaction produces $6e^-$ for every mole of methanol, but the methanol must be mixed with water. Water management in DMFCs is challenging because electroosmotic drag will take substantial amounts of water across the membrane from the anode to the cathode. In a portable device this water must be pumped or wicked back from the cathode to the anode to ensure that concentrated aqueous methanol is used so that the device can achieve a high power density. Fuel crossover is also much worse with methanol than hydrogen as the fuel is absorbed into the membrane resulting in further efficiency degradations.[26] In PFSA membranes the methanol crossover rate is 10^{-6} mol/cm^2/sec corresponding to a loss of current density of 50–100 mA/cm^2.[8] Much effort in commercializing PFSA-based materials as the electrolyte for DMFC is currently under way, which primarily involves engineering around these problems. Membrane development for DMFC focuses on minimizing crossover and electroosmotic drag. A number of modifications to PFSA membranes are under consideration. By operating the FC at 140°C and using a silica/Nafion composite membrane, methanol crossover can be diminished.[5] Nafion has been treated with phosphoric acid, doped with cesium ions, and impregnated with polymer precursors to produce polymers, such as poly(1-methylpyrrole), in situ.[27] A number of recent publications describe barrier strategies to the development of new polymer electrolyte membranes, including using plasma polymerization, layering solution cast PBI films with Nafion, using a sulfonated polyvinyl alcohol barrier, creation of a three-layered laminar electrolyte, combining a palladium membrane with Nafion, use of low-dose electron beam exposure to modify the Nafion surface structure, and creating a thin methanol impermeable barrier at the film surface.[27,28] While PFSA remains the benchmark for DMFC a number of other materials are being developed for DMFC, such as SPEEK, H_3PO_4–PBI, zirconium phosphate doped BPSH-40, and sol–gel HPA membranes.

CONCLUSIONS

Fuel cell membranes are close to optimization with PFSA materials. These materials give excellent performance when fully humidified and have adequate oxidative stability. These materials limit the temperature of operation of the FC to <100°C and have unacceptably high methanol crossover for direct methanol

FCs. For the further development of FCs, membranes must be fabricated that allow FC operation at temperatures at or above 120°C. While direct methanol FCs are close to commercialization, membranes with lower methanol crossover and electroosmotic drag coefficients would greatly facilitate their wider use.

ACKNOWLEDGMENTS

The author would like to thank Dr. Steven Hamrock of 3M and Dr. John Turner of the National Renewable Energy Laboratory for useful discussions.

REFERENCES

1. Haile, S.M. Fuel cell materials and components. Acta Materialia **2003**, *51*, 5981.
2. Savadogo, O. Emerging membranes for electrochemical systems: part II. High temperature composite membranes for polymer electrolyte fuel cell (PEFC) applications. J. Power Sources **2004**, *127*, 135.
3. Vielstich, W., Lamm, A., Gasteiger, H., Eds.; *Handbook of Fuel Cells—Fundamentals, Technology, Applications*; John Wiley & Sons: West Sussex, U.K., 2003.
4. Costamagna, P.; Srinivasan, S. Quantum jumps in the PEMFC science and technology from the 1960s to the year 2000: part II. Engineering, technology development, and application aspects. J. Power Sources **2001**, *102*, 253.
5. Costamagna, P.; Srinivasan, S. Quantum jumps in the PEMFC science and technology from the 1960s to the year 2000: part I. Fundamental scientific aspects. J. Power Sources **2001**, *102*, 242.
6. Kreuer, K.D. On the development of proton conducting polymer membranes for hydrogen and methanol fuel cells. J. Membr. Sci. **2001**, *185*, 29.
7. Kordesch, K.V. 25 years of fuel cell development (1951–1976). J. Electrochem. Soc. **1978**, *125*, 77C.
8. Li, Q.; He, R.; Jensen, J.O.; Bjerrum, N.J. Approaches and recent development of polymer electrolyte membranes for fuel cells operating above 100°C. Chem. Mater. **2003**, *15*, 4896.
9. Slade, S.; Campbell, S.A.; Ralph, T.R.; Walsh, F.C. Walsh, ionic conductivity of an extruded nafion 1100 EW series of membranes. J. Electrochem. Soc. **2002**, *149*, A1556.
10. Ren, X.; Gottesfeld, S. Electro-osmotic drag of water in poly(perfluorosulfonic acid) membranes. J. Electrochem. Soc. **2001**, *148*, A87.
11. Buchi, F.N.; Scherer, G.G. Investigation of the transversal water profile in Nafion membranes in polymer electrolyte fuel cells. J. Electrochem. Soc. **2001**, *148*, A183.
12. Uchida, H.; Ueno, Y.; Hagihara, H.; Watanabe, M. Self-humidifying electrolyte membranes for fuel cells. J. Electrochem. Soc. **2003**, *150*, A57.
13. Hsu, W.Y.; Gierke, T.D. Ion transport and clustering in Nafion perfluorinated membranes. J. Membr. Sci. **1983**, *13*, 307.
14. Rubatat, L.; Rollet, A.L.; Gebel, G.; Diat, O. Evidenvce of elongated polymeric aggregates in nafion. Macromolecules **2002**, *35*, 4050.
15. Curtin, D.E.; Lousenberg, R.D.; Henry, T.J.; Tangeman, P.C.; Tisack, M.E. Advanced materials for improved PEMFC performance and life. J. Power Sources **2004**, *131*, 41.
16. Cleghorn, S.; Kolde, J.; Reid, R.; Teller, O. New MEAs for low cost system design. In *2003 Fuel Cell Seminar Abstracts*; Courtesy Associates: Washington, DC, 2003.
17. Kolde, J.A.; Bahar, B.; Wilson, M.S.; Zawodzinski, T.A.; Gottesfeld, S. Advanced composite polymer electrolyte fuel cell membranes. In *Proton Conducting Membrane Fuel Cells I*; The Electrochemical Society: Chicago, IL, 1995.
18. Malhotra, S.; Datta, R. Membrane-supported nonvolatile acidic electrolytes allow higher temperature operation of proton-exchange membrane fuel cells. J. Electrochem. Soc. **1997**, *144*, L23.
19. Brown, G.M.; Noe-Spirlet, M.-R.; Busing, W.R.; Levy, H.A. Dodecatungstophosphoric acid hexahydrate, $(H_5O_2^+)_3(PW_{12}O_{40}^{3-})$. The true structure of Keggin's 'pentahydrate' from single-crystal X-ray and neutron diffraction data. Acta Cryst. **1977**, *B33*, 1038.
20. Kim, Y.S.; Dong, L.; Hickner, M.A.; Glass, T.E.; Webb, V.; McGrath, J.E. State of water in disulfonated poly(arylene ether sulfone) copolymers and a perfluorosulfonic acid copolymer (Nafion) and its effect on physical and electrochemical. Macromolecules **2003**, *36*, 6281.
21. Kim, Y.S.; Wang, F.; Hickner, M.; Zawodzinski, T.A.; McGrath, J.E. Fabrication and characterization of heteropolyacid $(H_3PW_{12}O_{40})$/directly polymerized sulfonated poly(arylene ether sulfone) copolymer composite membranes for higher temperature fuel cell applications. J. Membr. Sci. **2003**, *212*, 263.
22. Lixiang, X.; Zhang, H.; Choe, E.-W.; Scanlon, E.; Ramanathan, L.S.; Benicewicz, B.C. Synthesis and characterization of pyridine-based polybenzimidazoles as novel fuel cell membrane materials. Fuel Chem. Div. Prepr. **2003**, *48*, 447.
23. Honma, I.; Nakajima, H.; Nishikawa, O.; Sugimoto, T.; Nomura, S. Family of high-temperature

polymer-electrolyte membranes synthesized from amphiphilic nanostructured macromolecules. J. Electrochem. Soc. **2003**, *150*, A616.
24. Alberti, G.; Casciola, M.; Palombari, R. Inorgano-organic proton conducting membranes for fuel cells and sensors at medium temperatures. J. Membr. Sci. **2000**, *172*, 233.
25. Boysen, D.A.; Uda, T.; Chisholm, C.R.I.; Haile, S.M. High-performance solid acid fuel cells through humidity stabilization. Science **2004**, *303*, 68.
26. Heinzel, A.; Barragan, V.M. A review of the state-of-the-art of the methanol crossover in direct methanol fuel cells. J. Power Sources **1999**, *84*, 70.
27. Hobson, L.J.; Ozu, H.; Yamaguchi, M.; Hayase, S. Modified Nafion® 117 as an improved polymer electrolyte membrane for direct methanol fuel cells. J. Electrochem. Soc. **2001**, *148*, A1185.
28. Shao, Z.-G.; Hsing, I.-M. Nafion® membrane coated with sulfonated poly(vinyl alcohol)-Nafion® film for direct methanol fuel cells. Electrochem. Solid-State Lett. **2002**, *5*, A185.

Functional Biomaterials

Chun Wang
Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.

INTRODUCTION

The term ''functional biomaterials'' is a broad definition of biomaterials that carry biologically relevant function(s). These may include all the contemporary biomaterials used in medical implants and devices and are classified traditionally as metals, ceramics, composites, and polymers. As the development of biomaterials relies more and more heavily on the understanding and adaptation of principles in biology, functional biomaterials often refer to materials that combine biological molecules, such as proteins, peptides, and nucleic acids, forming systems that actively interact with biological entities (such as cells) and modulate biological processes. Functional biomaterials are also responsive materials that are able to recognize signals in biological environment, change their structures, and carry out their functions accordingly.

The layout of this entry is structured around ''biologically relevant functions,'' which specifically refer to the ability of recognizing cells and regulating cellular activities and the ability of delivering biologically active substances precisely. Topics discussed in detail include material surfaces functionalized with biomolecules that bind to cells and support cell growth in 2-D, material matrices that encapsulate and support cell and tissue growth in 3-D, ''intelligent'' materials and material surfaces that respond to biological signals, and materials that control specific delivery of genetic therapeutics to cells in vivo. Preparation, processing, and functionalization strategies are described for synthetic polymeric biomaterials, synthetic systems combined with biomolecules, and biopolymers that are derived naturally or synthesized de novo.

FUNCTIONAL BIOMATERIALS THAT PROMOTE CELL ADHESION

Inside the human body, cell adhesion to foreign biomaterial surfaces is mediated by a layer of proteins found in the blood or serum. Biomaterials that are able to control the adsorption of blood proteins will be able to control selectively the adhesion of cells. This function underlies the so-called ''biocompatibility'' of biomaterials and is ultimately responsible for the success or failure of medical implants and devices in vivo.[1]

Minimizing Nonspecific Cell Adhesion

Functional biomaterial surfaces that absorb proteins minimally are desirable in prolonging the lifetime of medical implants and providing a clean background for introducing specific cell adhesion functionalities.[1] Nonspecific protein adsorption occurs in various degrees to all surfaces, but more readily to hydrophobic and positively charged surfaces. To date, the most effective way to minimize nonspecific protein and cell adhesion is to use surfaces comprised of chains of polyethylene oxide (PEO; also named polyethylene glycol, or PEG).[1]

PEG chains are extremely flexible and hydrophilic. Protein adsorption to a PEG-surface restricts the freedom of the PEG chains and is therefore thermodynamically unfavorable. The effectiveness of reducing protein adsorption depends on the length of PEG—longer chains are more effective than short ones at any given surface coating density. PEG chains can be chemically coupled at one end to many biomaterial surfaces bearing reactive groups such as amines and carboxylates. Block copolymers of PEG, such as di- or tri-block copolymers of PEG and polypropylene oxide (PPO), known as Pluronics® or Poloxamers®, and PEG graft copolymers, can be used to physically coat hydrophobic material surfaces such as polystyrene beads or dishes. PEG-based polymer networks, or hydrogels, are well known to absorb proteins minimally and are often used to coat otherwise protein-absorbing surfaces. Gold surfaces can be similarly passivated through forming self-assembled monolayers (SAMs) of alkanethiols terminated with oligoethylene glycol chains. Recently, a new approach to more stable surface coating was developed[2] combining thiol–gold interaction with PEG block copolymers (PEG-polypropylene sulfide-PEG). The free-ends of surface-bound PEG chains can then be used to attach cell adhesion motifs to facilitate specific interactions with cells on top of a nonadhesive ''low-noise'' background.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120037175

In addition to PEG, certain polysaccharides, such as hyaluronic acid (HA) and dextran, have been used as low protein-adsorption, low cell adhesion surface coatings. Synthetic polymer surfactants consisting of poly(vinyl amine) with dextran and alkanoyl side chains, which mimicks the glycocalyx—negatively charged sugar layer outside cell membrane, are also shown to reduce protein adsorption to hydrophobic graphite surface.[3]

Surface Immobilization of Cell Adhesion Motifs

Early attempts to functionalize biomaterial surfaces with biological molecules[1] were focused on improving blood compatibility of cardiovascular devices, such as the artificial heart and synthetic blood vessels, by immobilizing heparin or albumin on polyurethane or Dacron®. To enhance cell adhesion to biomaterial surfaces, entire extracellular matrix (ECM) proteins, such as fibronectin and laminin, have been used directly as coatings. However, because of the nonspecific manner of whole protein adsorption, most of the cell binding capability is often lost. Using a molecular templating technique, it may be possible to select which protein(s) to absorb on biomaterial surfaces.[4]

It was found that ECM proteins contain short stretches of amino acids (peptides) that bind specifically to a group of receptors on most cell surface, the integrin receptors, which triggers intracellular events that lead to various kinds of cell behavior. These cell adhesion peptides have been immobilized on synthetic biomaterial surfaces to promote specific attachment of selected cell types, which is desirable in many tissue engineering applications. Comparing with ECM protein adsorption, immobilizing cell adhesion peptides can be done in a much better controlled fashion with higher cell binding functionality.[5]

A common method to graft cell adhesion peptides on biomaterial surface is chemical end-coupling, attaching reactive amino acid residues in the peptides to reactive groups on the surface. If the biomaterial surface does not naturally contain reactive groups, they can be generated either chemically or by radiation (UV, plasma gas discharge, etc.). A linker or ''arm'' is often bridging the peptide and the surface to enhance the freedom and activity of the adhesion motif.[1]

One of the most prevalent cell adhesion peptides is the tripeptide arginine–glycine–aspartate (RGD) from fibronectin.[5] This tripeptide has been used extensively to promote adhesion and spreading of many cell types. For example, polyurethane surface can be activated and coupled to RGD via the carboxy terminus. Surfaces containing photo-activable groups, such as benzophenone or aryl azide, can be modified easily with RGD by UV radiation. This tripeptide and other similar adhesion peptides have also been grafted to activated PEG surface, which supported long-term highly specific cell adhesion and spreading. Through copolymerized lysines, RGD has been immobilized on biodegradable polymers of the poly(glycolic acid) and poly(lactic acid) families. Glass surface can be silylated by trialkoxysilanes to couple with adhesion motifs bearing amines, hydroxyls, or carboxyls. In addition to peptide motifs, monosaccharides, such as glucose and lactose have been used to functionalize polystyrene surfaces to enhance hepatocyte binding.[5]

Immobilization of cell adhesion motifs using physical interactions is more straightforward than chemical attachment.[5] One of the earlier reports on the practical application of RGD involved peptide adsorption to many polymer and inorganic surfaces via a hydrophobic tail of oligoleucine. Polymeric surfactants such as the PEG block copolymers can be modified to display RGD on the end of the hydrophilic block and coat hydrophobic substrates. Alkanethiols bearing cell adhesion peptides form SAMs on gold surfaces. Peptide amphiphiles containing RGD and other sequences on one end, and long alkyl ester lipid tails on the other, have been shown to self-assemble on hydrophobic surfaces and used to probe specific cell adhesion.[6]

The strength and the specificity of cell adhesion by surface-functionalized biomaterials are dependent on the density, composition, and 2-D distribution of the immobilized adhesion motif.[5] It is generally accepted that there is a trade-off between cell affinity and cell mobility, with the optimal cell mobility achieved at an intermediate ligand density. The explanation is that low ligand density does not provide enough adhesion sites and strength, whereas high density creates too much adhesion, limiting mobility. Cell adhesion motifs with different chemical composition mediate adhesion and spreading of different cell types. For example, a peptide derived from laminin, isoleucine-lysine-valine-alanine-valine (IKVAV), mediates neurite extension, whereas the RGD peptide attaches to most cell types. The combination of cell adhesive surfaces with non-adhesive surfaces makes it possible to generate 2-D patterns that lead to cell attachment to selected and predetermined areas on biomaterials. Micropatterning techniques such as photolithography and microcontact printing (or soft lithography) have been applied[7] either to create patterns of large populations of one or multiple cell types or to control the shape of cells, which adopt the shape of the ''adhesive islands.''[8] The ability of organizing cells spatially is of interest in tissue engineering applications, and cell shape control serves an excellent model for elucidating the biology of mechanotransduction.

It has long been realized that the surface topography of biomaterials plays a very important role in addition to the cell adhesive motifs effecting cell

attachment and spreading. Texture of biomaterial surfaces, such as micrometer-size grooves, influences cell behavior through contact guidance.[5] Recently it became clear[9] that nanometer-size topographic features are more important than micrometer-size features; however, the exact mechanism of contact guidance remains unclear. The design of next generation functional biomaterial surfaces for controlling cell behavior must take into account both the chemical signals (the cell adhesion motifs) and the topographic cues.

Besides solid material surfaces, functional water-soluble polymers have also been created to promote cell–cell adhesion in aqueous media. For example, cell adhesion peptides of RGD or tyrosine–isoleucine–glycine–serine–arginine (YIGSR) have been attached to the ends of biofunctional PEG and shown to induce neural cell aggregation in culture media.[10]

Responsive Material Surfaces

Functional biomaterial surfaces have been created to change between being hydrophobic and hydrophilic, in response to external signals, such as differences in temperature, solvent environment, light, or electrical current.[11] Temperature-sensitive poly(*N*-isopropylacrylamide) (PNIPAm)-coated substrate is hydrophobic at a cell culture temperature of 37°C that favors cell adhesion and becomes hydrophilic at a lower temperature of 20°C, causing the detachment of cell sheets.[12] These reversible responsive surfaces have been used to culture and harvest layers of endothelial, epithelial, lung, liver, cardiac, and kidney cells, and could eventually enable assembly of complex tissues and organs.

FUNCTIONAL BIOMATERIALS THAT PROMOTE CELL AND TISSUE GROWTH IN 3-D

Functional biomaterials that support growth of cells and tissues in 3-D are divided into two categories: polymeric scaffolds and hydrogels. These structures not only provide mechanical support of cells, but also provide necessary chemical and biological signals to allow cell attachment, migration, proliferation, and differentiation.

Material Processing for Cell Macroencapsulation

The inclusion of cells into functional biomaterial scaffolds carries several critical requirements. During and after the macroencapsulation process, cells must remain viable and physiologically functional. It is desirable to encapsulate as many cells as possible and have them distributed evenly throughout the scaffolds, because the mass and the structure of the tissue product depend on the initial cell number and distribution. For in vivo applications, it is often necessary to perform cell encapsulation quickly inside body cavities with irregular shapes and sizes. These requirements demand special processes for material preparation and scaffold formation.

There are generally two approaches for cell macroencapsulation. One is to fabricate the scaffold first, followed by cell seeding. The other approach is to combine cells with a mixture of precursors in a liquid state and form the scaffold structure around cells, or in situ. The first cell seeding approach requires scaffolds with high porosity, so that large numbers of cells could be introduced into the scaffold interior with ease. One way of generating porous polymeric scaffolds is by solvent casting in the presence of porogens such as PEG or salt crystals.[13] The polymer is first dissolved in an organic solvent and cast into a mold. The solvent is then removed, followed by the removal of the porogens, leaving pores in the polymer bulk. Other pore-generating techniques for water-insoluble polymer scaffolds such as poly(lactic acid) (PLA) have also been developed, using gas-foaming and supercritical carbon dioxide.[13] Polymer fibers can also be woven into porous structures as cell scaffolds. The in situ encapsulation approach usually requires the chemistry of scaffold formation to be triggerable externally, such as photo-initiated polymerization and gelation of precursor molecules, temperature-controlled reversible sol–gel transformation, or mixing-controlled chemical coupling of two activated polymer precursors. Scaffold formation and cell encapsulation can also be driven by self-assembly of molecules such as block copolymers, peptides, proteins, and polysaccharides, into supra-molecular structures such as fibers and gels.

Synthetic Biodegradable Polymers

Biological functionalization of synthetic biodegradable polymers as 3-D scaffold for cells involves grafting ECM proteins or cell adhesion motifs such as the RGD peptide. One example is introducing reactive groups such as amines into the polymer backbone through copolymerization and using these groups later for peptide conjugation, as in the case of poly(lactic acid-*co*-lysine).[14] Other scaffolds of biodegradable polymers may be treated to reveal reactive groups by plasma, by surface-restricted controlled hydrolysis (by, e.g., sodium hydroxide), or may be blended with another functional polymer near the surface for peptide conjugation. Surface treatment involving breaking polymer chains should be done carefully, because of the risk of compromising the mechanical strength

of the scaffold as a result of excessive polymer chain scission.[1]

Most of the commonly used degradable polymer scaffolds are mechanically strong, but for certain applications such as engineering muscles and tendons, which require considerable elasticity, these polymers are not optimal. Novel biodegradable polyesters have been developed with superior elasticity and strength that resemble vulcanized rubber and are hence termed as "biorubber."[15] Scaffolds made with these mechanically functional materials may be useful especially in engineering elastic tissue such as muscular-skeletal tissues and blood vessels.

Synthetic Polymer Hydrogels

Because of the excellent biocompatibility and hydrophilicity of PEG and well-known chemical derivatizations of PEG end groups, PEG and PEG-based copolymer hydrogels are used extensively to encapsulate cells in tissue engineering.[16] PEG containing terminal acrylates and α-hydroxy acid can be photopolymerized to form hydrogels. Cross-linkers containing peptide sequences have been introduced into these gels, which are susceptible to degradation by specific enzymes secreted by cells. These systems have been explored to mimic the cell-initiated remodeling of natural ECM, allowing cell migration, degradation, secretion of natural ECM, and releasing growth factors on demand.[17] They have also been used to form protective layer of endothelium to block restenosis, the unwanted clogging of blood vessels after balloon angioplasty.[5] Another synthetic pathway of PEG hydrogels uses Michael-type addition reaction between a nucleophile such as a thiol group and an acrylate or acrylamide. Peptides containing terminal cysteines can crosslink branched or multiarmed (or star-shaped) PEG upon simple mixing and form hydrogels in minutes with encapsulated cells under room temperature.[18] In addition to peptide-cross-linkers allowing biodegradation, PEG hydrogels can also be modified with sugar residues to enhance adhesion to certain cell types. Extensive work has been reported on tailoring the bioadhesive, mechanical, and degradation properties of PEG hydrogels to achieve better tissue engineering outcomes.[19]

PEO-PPO-PEO triblock copolymers (Pluronics or Poloxamers) form reversible physically cross-linked hydrogels under certain concentration range and temperature. The use of this system in tissue engineering is scarce because of its inability to degrade. Di- or tri-block copolymers of PEG with PLA have been developed to overcome this problem. Multiple blocks of PEG and PLA, synthesized by condensation reaction of L-lactic acid in the presence of succinic acid, can also form reversible hydrogels near body temperature. Cells may be combined with the block copolymers in the liquid state at a lower temperature and can be subsequently injected into the body where gels are formed. Biodegradation occurs later through the PLA segments.[19]

Poly(vinyl alcohol) (PVA) has tunable hydrophilicity and water-solubility through controlling the extent of hydrolysis of its precursor, poly(vinyl acetate), and its molecular weight. It can be chemically cross-linked into gels by glutaraldehyde or epichlorohydrin, which are highly toxic small molecules. Alternative methods of gelation have been developed using a repeated freeze/thaw method or using an electron beam. Because of its low degradability in vivo, PVA has been primarily used as a long-term or permanent scaffold material or as blends with other degradable polymers.[19]

Polyphosphazene is a class of organometallic polymer that contains alternating phosphorus and nitrogen atoms with two side groups attached to each phosphorus atom. It is synthesized by a substitution reaction between poly(dichlorophosphazene) and alcohols or amines. The backbone of the polymer is hydrophilic. The two side groups on the phosphorus atom can take on various structures and properties, which potentially can be modified to create biofunctionalities. Cross-linked polyphosphazene gels have been investigated for skeletal tissue regeneration.[19]

Poly(2-hydroxyethylmethacrylate) (HEMA) hydrogels were developed initially as soft contact lenses. Macroporous polyHEMA gels have been prepared by freeze/thaw method or salt leaching technique and used for nondegradable cartilage replacement. The modification of polyHEMA gels by dextran enables enzymatic degradation of the scaffolds. A novel cross-linking chemistry of polyHEMA with grafted enantiomeric oligo(L-lactide) and oligo(D-lactide) has been developed, allowing in situ formation of polyHEMA hydrogels, which are potentially useful in cell encapsulation.[19]

Owing to their temperature-sensitive sol–gel transition behavior, PNIPAAm and copolymers with acrylic acid or acrylamide can be used to encapsulate cells. Chondrocytes have been successfully encapsulated without using any organic solvent or toxic compound. The application of these systems as injectable bioadhesive (RGD) scaffold for cartilage tissue engineering is being explored. Degradability can be introduced into the otherwise nondegradable polymer system by peptide- or polysaccharide-based cross-linkers.[19]

Poly(propylene fumarate-*co*-ethylene glycol) (PF-*co*-EG) is a hydrophilic block copolymer that can be cross-linked chemically or by UV light. When used as an injectable cell scaffold in bone and vascular tissue engineering, this block copolymer degrades through the ester bonds in the PF blocks.[20]

To afford biological functionality to the otherwise inert synthetic polymer hydrogels, cell adhesion peptides are either incorporated as cross-links or attached to pendant polymer chains in the bulk or on the surface. Biologically active growth factors, such as TGF-β, bone morphogenetic proteins (BMPs), and VEGF, have been incorporated in hydrogels through covalent tethering, physical tethering, or entrapment, as stimuli for cell growth and differentiation.[20]

Natural Polymer Hydrogels

Collagen is a major component of the ECM in many tissues, such as bone, cartilage, and skin. Collagen forms thermal reversible gels physically and gels cross-linked by dialdehyde or diazide chemically. Collagen contains numerous cell adhesion motifs and sequences that are degradable by specific cell proteases. Collagen gels are often modified to include other biomolecules such as fibronectin, chondroitin sulfate, or hyaluronic acid. Although there are large variations of collagen extracted from natural sources, it serves as an excellent functional biomaterial matrix for engineering tissues and organs such as liver, skin, blood vessel, and small intestine. Single-stranded partial degradation product of collagen is gelatin, which also forms gels and is used in various tissue engineering applications.[19]

Silkworm silk has a long history of medical use as degradable sutures with good biocompatibility. Recently, the mechanism of natural processing to create high strength silk fibers was elucidated. Water-soluble silk fibroin protein extracted from silkworm silk has been processed into hydrogels with or without PEG by varying temperature, pH, and salt concentration. Because of its long degradation time, silk protein may be especially useful in reconstructing tissues requiring relatively stable scaffolds.[21]

Hyaluronic acid (or hyaluronate) is a giant molecule of glycosaminoglycan found in natural ECM and is especially important in wound healing. Hydrogels of HA can be prepared by cross-linking HA chains with hydrazide derivatives or by polymerization of glycidyl methacrylate-containing HA macromers, and HA gels are degraded by a natural enzyme in the body, hyaluronidase. Applications of HA gels include artificial skin, wound dressing, and soft tissue augmentation. Blending with synthetic polymers is often necessary to enhance the relatively weak mechanical properties of pure HA gels.[19]

Fibrin is the major protein component of blood clots and is formed by enzymatic cleavage and polymerization of fibrinogen. Because of its important role in natural wound healing process, fibrin is an attractive functional material for tissue engineering. Additional cell binding peptides, such as RGD, have been incorporated into fibrin using transglutaminase factor XIIIa activity. The resulting fibrin gels have been shown to promote neurite extension. Other applications of fibrin gels include engineering skeletal muscles, smooth muscles, and cartilage.[19]

Alginate is a natural polysaccharide obtained from brown algae. Divalent ions such as Ca^{2+} cross-link alginate into physical gels, which are used for microencapsulation of chondrocytes, hepatocytes, and islet cells to treat diabetes. Ionically cross-linked alginate gels degrade in vivo by losing the ions, and the degradation is difficult to control. Covalent cross-link of alginate has been achieved using adipic dihydrazide or bifunctional PEGs. It is possible to control alginate-PEG hydrogel formation, swelling properties, mechanical properties, and degradation. To incorporate cell binding ability, alginate gels have been modified with sugar-binding lectin and RGD and have been used to culture skeletal muscle cells.[19]

Agarose is another polysaccharide extracted from algae, and it forms thermally reversible gels. The concentration of agarose determines the pore size and stiffness of agarose gels and has been shown to influence cell migration. Agarose gels covalently modified with chitosan have been used to promote neurite outgrowth. Certain cell adhesion peptides have also been coupled to agarose gels.[19] Recently a method was developed to enable 3-D spatially controlled coupling of cell adhesive RGD within agarose gels, and directed growth of neurites has been demonstrated.[22]

Chitosan is a cationic polysaccharide derived from chitin, which is most abundant in crab shells. Because of its insolubility in neutral buffers and most organic solvents, chitosan has to be chemically modified to increase its solubility. Chitosan gels can be prepared by ionic crosslinking or chemical crosslinking by glutaraldehyde. Azide derivatives of chitosan can also be gelled by UV light. Chitosan gels have been modified with sugar moieties to interact with hepatocytes and with natural proteins such as collagen, gelatin, and albumin for neural regeneration.[19]

Genetically Engineered Protein Polymers

Recombinant DNA technology and bacterial fermentation techniques have enabled the design of artificial genes and the production of artificial protein materials. Typically a peptide repeating sequence is designed and translated into a DNA sequence. The DNA monomer is chemically synthesized and enzymatically polymerized and cloned into a bacterial vector or a plasmid, which is later transferred into a bacterial host to express the coded protein polymer. Because the protein products are genetically coded in the DNA sequences,

it is possible to produce protein-based biomaterials with precisely defined molecular weight, composition, biologically functional domains, and molecular assembly properties.[5,14]

Natural silk proteins constitute long repeated sequences of alanine–glycine (AG), which form aligned, stacked, antiparallel β-sheets. Hydrophilic amino acids such as glutamic acid have been placed with regular distances from one another, creating β-turns. It has been shown that the thickness of the β-sheet crystals can be precisely controlled and that non-natural amino acid residues can be placed precisely at one face of the crystals. Cell adhesion peptides, such as RGD and YIGSR, have been incorporated into these structures to be exposed at a particular crystal surface. These materials have been used to coat polystyrene substrates, exposing the cell adhesion motifs as support for cell culture.[5]

Elastin is an amorphous natural protein that affords elasticity to the ECM and connective tissues. It contains a five-residue repeat of glycine–valine–glycine–valine–proline (GVGVP), which has been polymerized genetically to yield elastin-analogs. These protein materials display distinct temperature-dependent phase transition, and the transition temperature is determined by the sequence of the elastin repeating unit and the substitution of the proline residue. Elastin polymer gels have been prepared by gamma-radiation or chemical cross-linking at precisely positioned lysine residues. Cell adhesion peptides such as RGD and others have been incorporated genetically into the elastin polymer at predetermined places. Elastin hydrogels are being evaluated for engineering vascular graft and cartilage tissue. Hydrogels based on elastin polymer grafted with PEG can be formed by photopolymerization, which, along with the temperature-triggered sol–gel transition behavior, is being exploited to encapsulate cells in situ.[5,14]

Block protein polymers containing hydrophilic segments and self-associating α-helical segments form reversible hydrogels spontaneously under certain temperature and pH conditions.[23] Natural and synthetic α-helical coiled coil protein segments have been combined with a synthetic copolymer of *N*-hydroxypropylmethacrylamide (HPMA) to create hybrid gels with unique responsive properties.[24] These gel systems have the potential of acquiring biofunctionalities that are relevant to tissue engineering, and they are being developed to control cell behavior.

Self-Assembling Peptides

Short peptide sequences containing alternating positively and negatively charged amino acid residues have been discovered serendipitously from a natural DNA-binding protein and found to self-assemble in salt-containing water into nanofibers, membranes, and gels.[25] Because of the high hydrophilicity of the sequences and the nanofiber-woven structure, these peptide hydrogels contain extremely high water-content (close to 99.5%). Salt-triggered gelation enables in situ encapsulation of cells into hydrogels, which have been shown to support growth of several types of cells and tissues including cartilage, liver cells, and neuronal cells.[25] Similar high water-content hydrogels have been created from chemically synthesized amphiphilic block copolypeptides. Toxicity of such materials to cells depends on the amino acid sequences of the polypeptides.[26]

Peptide amphiphiles containing an amino-terminal nonpolar hydrocarbon tail and a hydrophilic peptide head self-assemble into similar nanofibers, which subsequently form 3-D hydrogels. By incorporating a phosphorylated serine residue in the peptide region of the amphiphiles, calcium deposition and mineralization of hydroxyapatite can be initiated, forming bone-like structures. More recently, by incorporating cell adhesion peptide IKVAV into the amphiphile, neural progenitor cells cultured within these hydrogels differentiated into neurons, instead of astrocytes, owing to the specific binding and signaling of the IKVAV peptide.[27] This initial demonstration of how biomaterial matrix alone can be tailored to steer the fate of progenitor cells suggests the possibility of using these materials in vivo to promote regeneration of damaged tissues.

FUNCTIONAL BIOMATERIALS IN GENE DELIVERY

Functional Requirements for Biomaterials in Gene Delivery

Gene therapy is broadly defined as supplying the patient's cells with genetic materials, which prevent or correct a disease directly, or do so through their protein products. These genetic therapeutics include DNA (both large and small) and RNA. Genetic materials such as DNA plasmids can also be used as vaccines, if the plasmids encode antigenic proteins. The ability to deliver genetic therapeutics and vaccines specifically to target tissues and cells is crucial for the success of gene therapy, which promises to revolutionize modern medicine.[28]

There are two classes of gene delivery vehicles, or vectors, that are based on viruses or nonviral biomaterials. The viral delivery approach is limited by potential toxicity, immunogenecity, lack of targeting, production and packaging problems, and high cost. Nonviral vectors, on the other hand, have the potential to

overcome these limitations of viral vectors.[28] Unlike delivery of small molecular drugs, peptide, or protein drugs, gene delivery presents unique challenges and demands biomaterials assisting delivery to be equipped with specific biological functions.

Barriers to gene delivery are categorized into two levels: systemic and cellular. The systemic barrier is the reticular endothelial system (RES) of the body that includes liver, spleen, and phagocytic cells, which capture and clear any foreign material such as an exogenous DNA. Unwanted delivery to nontarget tissues and cells is also undesirable. At the cellular level, there are multiple barriers for DNA uptake, escaping endosomes following endocytosis, nuclear targeting/entry, and DNA unpackaging to allow transcription.[28] Functional biomaterials are expected to overcome these barriers, in addition to having excellent biocompatibility.

Targeted Gene Delivery to Cells

One way to localize gene delivery to a particular site in the body is using controlled release materials, mainly biodegradable polymers.[28] Microparticles of polymers based on poly(lactic-*co*-glycolic acid) (PLGA) have been extensively used to deliver DNA in vivo after local injections. By tuning particle size to the micrometer range, it is possible to passively target phagocytic cells such as macrophages and dendritic cells, which could be advantageous in DNA vaccine delivery. Recently, microparticles of novel poly(ortho ester) polymers have been developed[29] to accelerate degradation and DNA release in response to intracellular pH. DNA can also be tethered to or entrapped in a solid substrate, such as a tissue engineering scaffold, and be released to specific cells upon cell-initiated substrate degradation.

Prolonged systemic delivery of DNA can be achieved by coating the delivery vehicles with PEG, which avoids nonspecific cell adhesion by being ''stealth.'' Various biologically specific targeting ligands have also been coupled to delivery vectors.[30] These include transferrin, monoclonal antibody, mannose, galactose, lactose, folic acid, low-density lipoproteins, and RGD peptides. By binding to cell surface receptors, these targeting strategies enable DNA to enter target cells through receptor-mediated endocytosis. Recently, it is discovered that certain cationic peptides, such as a segment of the HIV Tat protein, when conjugated to a gene delivery vector, can enter cells directly and rapidly without relying on endocytosis. The benefit of direct cell entry by a gene delivery vector is that it bypasses the barrier of endosomal membrane.

Gene Condensation

Genomic DNA in the cell nucleus is highly condensed by cationic proteins such as histone, so that it can be packaged within a compacted organelle. Gene delivery also requires condensation of the exogenous DNA, because for the highly negatively charged hydrophilic DNA molecules it will otherwise be extremely difficult to cross the cell membrane. This is often accomplished by cationic vectors that neutralize the negative charges on the DNA, resulting in small complexes less than 200 nm in size.

Since the early 1970s, one of the first gene delivery systems useful in vivo was developed based on lipids of cationic *N*[1-(2,3-dioleyloxy)propyl]-*N,N,N*-trimethylammonium chloride) (DOTMA) and dioleoylphosphatidylethanolamine (DOPE), known as Lipofectin. Other commonly used cationic lipids include 2,3-dioleyloxy-*N*-[2-(spermine-carboxamido) ethyl]-*N,N*-dimethyl-1-propanaminium trifluoroacetate (DOSPA), dioctamido-decylamidoglycylspermine (DOGS), 1,2-dimyristyloxypropyl-3-dimethylhydroxyethyl ammonium bromide (DMRIE), 1,2-bis(oleoyloxy)-3-(trimethylammonio) propane (DOTAP), and 3β-(*N,N*-dimethylaminoethane) carbamoyl cholesterol (DC-Chol).[30] DNA is compacted and encapsulated in these cationic lipid vesicles, or liposomes, which enter cells by endocytosis.

Cationic polymers, such as poly(L-lysine) (PLL), polyethylenimine (PEI), chitosan, polyamidoamine (PAMAM) dendrimers, poly(2-dimethylamino) ethyl methacrylate, and polyphosphoesters, condense DNA to form compacted polyplexes.[30] The size and the stability of polyplexes depend on the ratio of cations vs. anions, temperature, ionic strength, and the solvent. Stability of polyplexes can be enhanced by conjugating PEG to the polycations or by using PEG-containing block or graft polymers that form micelles. Small cationic peptides are also able to condense DNA, however, six-consecutive-cations is the minimal requirement to achieve this effectively.

Subcellular Transport Processes

Endocytosed DNA/vector complexes end up in the endosomal compartments, which are filled with digestive enzymes in a slightly acidic microenvironment. Therefore, it is crucial for the DNA to be removed from that environment, which can result in DNA degradation and loss of activity. Polymers such as PEI and polyhistidine are thought to mediate the endosomal escape process by a hypothesized ''proton sponge'' mechanism—tertiary amine groups in the polymers buffer the acidic pH in endosome, causing it to burst or leak owing to osmotic effect.[30] Small

''fusogenic'' peptide sequences derived from virus such as influenza are known to change their conformation at endosomal pH and fuse with endosomal membrane. Similar peptides have been conjugated to cationic polymers and shown to enhance transfection efficiency of the exogenous DNA, presumably because of enhanced endosomal escape capability.[29] Synthetic anionic polymers such as poly(2-alkylacrylic acid) behave similarly to the fusogenic peptides in mediating pH-triggered release of DNA from endosome into the cytoplasm.

Once in the cytoplasm, the DNA or DNA/vector complex has to be transported to and enter the cell nucleus for gene transcription. This process is thought to be facilitated by specific peptide sequences termed nuclear localization signals (NLS).[29] In fact, it was shown that one NLS peptide appears sufficient for transporting a DNA plasmid to the nucleus. NLS peptides are frequently coupled to cationic vectors to aid in nuclear transport of DNA/vector complexes. Interestingly, it is reported recently that PEI/DNA complexes can be localized readily to the cell nucleus without using any NLS. Further elucidation of the transport mechanism inside cells should lead to better strategies of designing functional gene delivery materials.

Gene Unpackaging

To be transcriptionally active, the DNA must recover from its condensed form and be separated from its cationic carrier.[29] Disintegration of DNA/vector complex can be engineered by using disulfide crosslinks that are reduced in the cytoplasm by glutathione activity. The cationic polymer carriers (such as poly[α-(4-aminobutyl)-L-glycolic acid], and poly(4-hydroxy-1-proline ester)) can be programmed to degrade overtime by hydrolysis, releasing uncomplexed DNA for transcription.[30]

CONCLUSIONS

This entry focuses on biomaterials that carry biologically relevant functions. Methods are introduced for modifying biomaterial surface to minimize nonspecific cell adhesion and enhance specific cell adhesion and spreading by incorporating biomolecules. Preparation and processing of biomaterials that form 3-D scaffolds for supporting cell and tissue growth and differentiation are described. Biodegradable (synthetic and natural) polymers, hydrogels, and responsive biomaterials play pivotal roles in performing these biological functions. Finally, the development of functional biomaterials to facilitate gene delivery is discussed in the context of overcoming various biological barriers inside and outside cells.

REFERENCES

1. Ratner, B.D.; Hoffman, A.S.; Schoen, F.J.; Lemons, J.E. *Biomaterials Science: An Introduction to Materials in Medicine*; Academic Press: San Diego, 1996.
2. Bearinger, J.P.; Terrettaz, S.; Michel, R.; Tirelli, N.; Vogel, H.; Textor, M.; Hubbell, J.A. Chemisorbed poly(propylene sulphide)-based copolymers resist biomolecular interactions. Nat. Mater. **2003**, *2*, 259–264.
3. Holland, N.B.; Qiu, Y.; Ruegsegger, M.; Marchant, R.E. Biomimetic engineering of non-adhesive glycocalyx-like surfaces using oligosaccharide surfactant polymers. Nature **1998**, *392* (6678), 799–801.
4. Shi, H.; Tsai, W.B.; Garrison, M.D.; Ferrari, S.; Ratner, B.D. Template-imprinted nanostructured surfaces for protein recognition. Nature **1999**, *398* (6728), 593–597.
5. Hubbell, J.A. Biomaterials in tissue engineering. Biotechnology **1995**, *13*, 565–576.
6. Tirrell, M.; Kokkoli, E.; Biesalski, M. The role of surface science in bioengineered materials. Surf. Sci. **2002**, *500*, 61–83.
7. Whitesides, G.M.; Ostuni, E.; Takayama, S.; Jiang, X.; Ingber, D.E. Soft lithography in biology and biochemistry. Annu. Rev. Biomed. Eng. **2001**, *3*, 335–373.
8. Chen, C.S.; Mrksich, M.; Huang, S.; Whitesides, G.M.; Ingber, D.E. Geometric control of cell life and death. Science **1997**, *276* (5317), 1425–1428.
9. Curtis, A.S.; Wilkinson, C.D. Reactions of cells to topography. J. Biomater. Sci. Polym. Ed. **1998**, *9* (12), 1313–1329.
10. Dai, W.; Belt, J.; Saltzman, W.M. Cell-binding peptides conjugated to poly(ethylene glycol) promote neural cell aggregation. Biotechnology **1994**, *12* (8), 797–801.
11. Russell, T.P. Surface-responsive materials. Science **2002**, *297*, 964–967.
12. Shimizu, T.; Yamato, M.; Kikuchi, A.; Okano, T. Cell sheet engineering for myocardial tissue reconstruction. Biomaterials **2003**, *24* (13), 2309–2316.
13. Lanza, R.P., Langer, R., Chick W.L., Eds.; *Principles of Tissue Engineering*; Academic Press: San Diego, 1997.
14. Langer, R.; Tirrell, D.A. Designing materials for biology and medicine. Nature **2004**, *428*, 487–492.
15. Wang, Y.; Ameer, G.A.; Sheppard, B.J.; Langer, R. A tough biodegradable elastomer. Nat. Biotechnol. **2002**, *20* (6), 602–606.

16. Hoffman, A.S. Hydrogels for biomedical applications. Adv. Drug Delivery Rev. **2002**, *54* (1), 3–12.
17. Nguyen, K.T.; West, J.L. Photopolymerizable hydrogels for tissue engineering applications. Biomaterials **2002**, *23* (22), 4307–4314.
18. Lutolf, M.P.; Hubbell, J.A. Synthesis and physicochemical characterization of end-linked poly-(ethylene glycol)-co-peptide hydrogels formed by Michael-type addition. Biomacromolecules **2003**, *4* (3), 713–722.
19. Lee, K.Y.; Mooney, D.J. Hydrogels for tissue engineering. Chem. Rev. **2001**, *101* (7), 1869–1879.
20. Drury, J.L.; Mooney, D.J. Hydrogels for tissue engineering: scaffold design variables and applications. Biomaterials **2003**, *24*, 4337–4351.
21. Altman, G.H.; Diaz, F.; Jakuba, C.; Calabro, T.; Horan, R.L.; Chen, J.; Lu, H.; Richmond, J.; Kaplan, D.L. Silk-based biomaterials. Biomaterials **2003**, *24* (3), 401–416.
22. Luo, Y.; Shoichet, M.S. A photolabile hydrogel for guided three-dimensional cell growth and migration. Nat. Mater. **2004**, *3*, 249–253.
23. Petka, W.A.; Harden, J.L.; McGrath, K.P.; Wirtz, D.; Tirrell, D.A. Reversible hydrogels from self-assembling artificial proteins. Science **1998**, *281*, 389–392.
24. Wang, C.; Stewart, R.J.; Kopecek, J. Hybrid hydrogels assembled from synthetic polymers and coiled-coil protein domains. Nature **1999**, *397*, 417–420.
25. Zhang, S. Fabrication of novel biomaterials through molecular self-assembly. Nat. Biotechnol. **2003**, *21* (10), 1171–1178.
26. Pakstis, L.M.; Ozbas, B.; Hales, K.D.; Nowak, A.P.; Deming, T.J.; Pochan, D. Effect of chemistry and morphology on the biofunctionality of self-assembling diblock copolypeptide hydrogels. Biomacromolecules **2004**, *5* (2), 312–318.
27. Silva, G.A.; Czeisler, C.; Niece, K.L.; Beniash, E.; Harrington, D.A.; Kessler, J.A.; Stupp, S.I. Selective differentiation of neural progenitor cells by high-epitope density nanofibers. Science **2004**, *303* (5662), 1352–1355.
28. Luo, D.; Saltzman, W.M. Synthetic DNA delivery systems. Nat. Biotechnol. **2000**, *18*, 33–37.
29. Wang, C.; Ge, Q.; Ting, D.; Nguyen, D.; Shen, H.-R.; Chen, J.; Eisen, H.N.; Heller, J.; Langer, R.; Putnam, D. Molecularly engineered poly(ortho ester) microspheres for enhanced delivery of DNA vaccines. Nat. Mater. **2004**, *3*, 190–196.
30. Han, S.; Mahato, R.I.; Sung, Y.K.; Kim, S.W. Development of biomaterials for gene therapy. Mol. Ther. **2000**, *2* (4), 302–317.

Gas Explosion Hazard: Prevention and Protection

Dehong Kong
Chilworth Technology, Inc., Monmouth Junction, New Jersey, U.S.A.

INTRODUCTION

Controlling gas/vapor explosion hazards is often a necessity of chemical process safety. To control gas explosion hazards, one has to

- Determine the conditions for forming flammable gas–oxidant mixtures.
- Identify the ignition sources.
- Design plants to eliminate or minimize the likelihood of gas explosions, or provide safety devices to protect people and plant from the harms of gas explosions.

This entry introduces readers to the fundamentals of gas explosions, and basic information on how to identify, assess, and eliminate or control gas explosion hazards in their facilities. Any further detailed information regarding the subjects that are covered in this entry may be obtained from the references.

FUNDAMENTALS OF GAS EXPLOSIONS

Basic Concepts

The term ''explosion'' is best defined as a process that involves a sudden release of energy resulting in a rapid and significant buildup of overpressure. Explosions can be categorized into physical/mechanical and chemical explosions. For example, an explosion caused by a sudden release of compressed gas is a physical explosion. A chemical explosion is caused by a chemical reaction(s), which could be combustion, exothermic decomposition or exothermic reaction. Chemical explosions can occur in gas, liquid or solid phase. Chemical explosions that occur in liquid and solid phases are sometimes called condensed phase explosions. Explosive explosions fall in this category.

Gas/vapor explosions are caused by chemical processes, such as combustions or exothermic decompositions of unstable chemicals like acetylene. Because the mechanisms of gas and vapor explosions are same, ''gas explosion'' means both gas explosion and vapor explosion throughout this entry. Combustion is an oxidation accompanied by a production of heat and light. Three elements must be present to support combustion—a combustible chemical called fuel, an oxidant—usually oxygen in air, and a sufficiently energetic stimulus called ignition source. These three elements are conventionally referred to as the fire triangle.

Fire is referred to as a combustion process in which the fuel and oxidant are transported separately to the reaction zone during the combustion. Because the rate of combustion in fires is limited by the rate of transportation of fuel and oxidant molecules, which are usually slower than the rate of oxidation, fires do not cause rapid pressure buildup. Gas explosions are combustions of premixed fuel–oxidant mixtures in confined spaces.

Flash fire is referred to as combustion of gas–oxidant mixture in which the flame propagates through the mixture, in a manner such that negligible or no damaging overpressures are generated. Although gas explosions and flash fires are different in the magnitude of overpressures and the mechanical damages, they are all combustions of premixed fuel–oxidant mixtures. Therefore, the basic mechanisms, prevention and protection methods for gas explosions are also applicable to flash fires.

Two velocities are used to express the speed of flame propagation. The first is the velocity of the flame front relative to the unburned gas mixture, which is called burning velocity (u). The second is the velocity of a flame front with respect to a stationary observer, which is called flame speed (S). The relationship between these two velocities is:

$$S = u \pm U \tag{1}$$

where U is the velocity of the unburned gas mixture relative to a stationary observer. The sign in front of U will be ''+'' when flame propagates in the same direction of the unburned gas flow velocity; and ''−'' when the flame propagates in the direction against the flow velocity.

When a premixed gas–oxidant cloud is ignited, the flame can propagate in two different modes through the gas mixture—deflagration and detonation. Deflagrations propagate at subsonic speeds relative to the unburned mixture and the heat and mass are transported by conduction, diffusion, and convection. Gas mixture detonations propagate at speeds faster than the local sound speed of the unburned gas. In a gas mixture detonation, a shock wave is sustained

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120024364

by the combustion of the flammable gas. The gas mixture at the shock front is compressed, such that the gas temperature exceeds its auto-ignition temperature causing auto-ignition just behind the shock wave. Therefore, the combustion wave and the shock wave propagate at the same speed.

Flammability Limits

Flammability limits and influential factors

A premixed fuel–oxidant mixture can burn only if the mixture composition falls into a certain range, which is referred to as the flammable range. A useful way to express the mixture composition is the triangle coordinates of fuel gas, oxidant and inert diluent, known as the flammability diagram (Fig. 1). If a gas mixture contains multiple fuel gases, all the combustibles can be considered as one fuel gas. For example, if a gas mixture includes methane, propane, and hydrogen with a given composition, the three gases can be considered as a single fuel gas as long as the composition of the mixture of these gases remains constant. The same applies to oxidants and diluents.

The envelope line, which separates the flammable and nonflammable gas compositions, is called the flammability limit. The area enclosed by this line contains all the flammable mixture compositions and the area outside the line contains all nonflammable mixture compositions. The lower section of the line toward the oxidant axis is called the lower flammability/explosible limit (LFL/LEL). The upper section of the line is called the upper flammability/explosible limit (UFL/UEL). Most flammability limits reported in the literature are for fuel–air mixtures, which are the two intersections of the flammability limits line and the constant oxidant/inert (=0.21/0.79) line. The two intersection points of the zero-inert-gas line and the fuel axis are the LFL and UFL in 100% oxidant.

There exists an oxidant concentration below which the fuel–oxidant mixture is not capable of propagating flame. This oxidant concentration is called the limiting/minimum oxidant concentration (LOC/MOC). LOC is dependent on the type of inert gas used.

LOC in air/nitrogen mixture can be estimated using:

$$\mathrm{LOC} = n_{\mathrm{oxygen}} \times \mathrm{LFL} \tag{2}$$

where n_{oxygen} is the moles of oxygen required to burn completely 1 mole of the combustible gas. LFL is

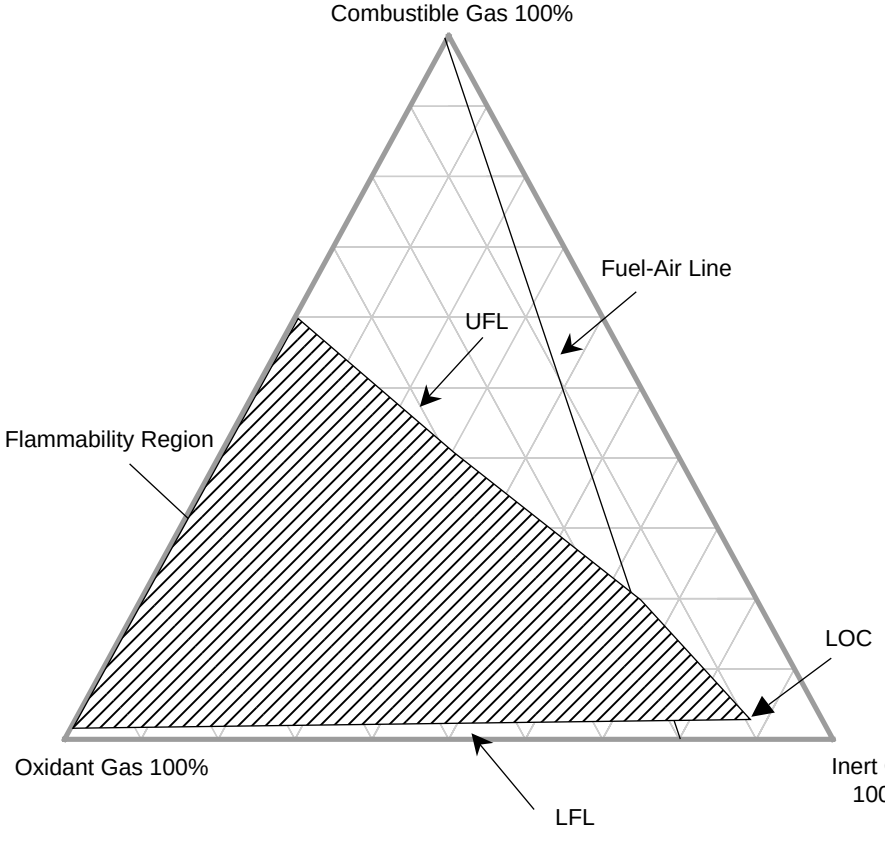


Fig. 1 Flammability diagram. *(View this art in color at www.dekker.com.)*

affected slightly by oxygen concentration. This equation can be used to calculate LOCs with good accuracy provided the LFL measurements are accurate.

To form a flammable vapor–air mixture above the surface of a liquid, the temperature of the liquid must be sufficiently high, and the lowest temperature at which a flammable gas–air mixture is formed at one atmosphere is defined as the flashpoint.[1]

The flammability limits for a vapor–liquid equilibrium system is depicted in Fig. 2.[2] The vapor concentration at the point at which the LFL line intersects the vapor concentration curve is the LFL and the corresponding temperature is referred to as the lower temperature limit of flammability (LTL),[2,3] which is identical to flashpoint. However, the reported LTL and flashpoint can be slightly different because of the different methods used in the tests.

The temperature at which the vapor concentration is equal to the UFL is referred to as the upper temperature limit of flammability (UTL).[2] Above the UTL, the saturated vapor concentration is greater than the UFL, i.e., the gas mixture becomes nonflammable.

Flammability limits are influenced by factors such as pressure, temperature, and oxidant concentration. As pressure increases, the LFL is decreased slightly while the UFL is increased significantly. The effect of higher pressure on LOC is comparable to that on the LFL. For several saturated hydrocarbons at a pressure within 0.1–20.7 MPa (1–204 atm), the effect of elevated pressure on UFL can be estimated using:

$$\mathrm{UFL}_P = \mathrm{UFL}_{P_0} + 20.6 \times (\log P + 1) \quad (3)$$

where UFL_{P_0} and UFL_P (%v/v) are the UFLs at 1 atm (0.1 MPa) and pressure P (MPa), respectively.[4]

Gases such as acetylene that have positive heat of formation can decompose explosively at high ressures in the absence of air when exposed to a sufficiently strong ignition source and, thus, have no UFLs.

A decrease in pressure below one atmosphere has little effect on the LFL as long as the pressure is higher than a limiting level (approximately 5 kPa), below which the mixture becomes nonflammable.[4]

Flashpoint/LTL is related to LFL as:

$$\mathrm{LFL} = \frac{P_{\mathrm{FP}}}{P_{\mathrm{total}}} \quad (4)$$

where P_{FP} and P_{total} are the vapor pressure at the flashpoint/LTL and total gas pressure in the vessel, respectively. As the total gas pressure is decreased, a lower vapor pressure and, thus, a lower temperature (flashpoint) will be required to form a vapor concentration equal to the LFL. Notably, the published flashpoint/LTL data, which are obtained at one atmosphere, do not apply to pressures above or below one atmosphere. Therefore, it is useful to broaden the definition of flashpoint/LTL such as this—the flash point/LTL is the lowest temperature at which the vapor–oxidant mixture at equilibrium with the liquid at a given pressure can be ignited by an incendive ignition source.

As the temperature increases, LFL is decreased and UFL is increased. The effect of temperature T(°C) on LFL can be estimated using the flammability

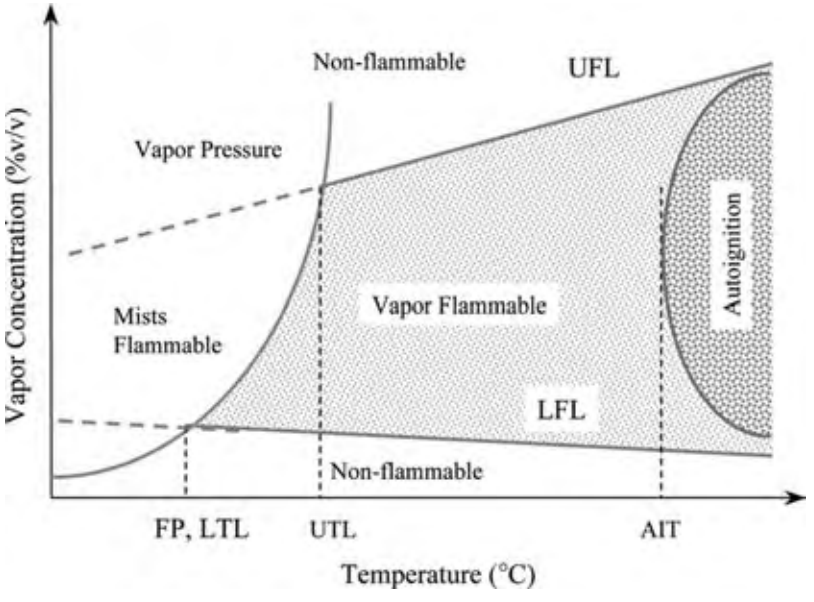


Fig. 2 Effect of temperature on flammability limits of a combustible vapor in air at constant initial pressure. (From Ref.[2].) *(View this art in color at www.dekker.com.)*

limits for 25°C:[2]

$$\mathrm{LFL}_T = \mathrm{LFL}_{25} \times \left[1 - \left(\frac{0.75}{H_c}\right) \times (T - 25)\right] \quad (5)$$

$$\mathrm{UFL}_T = \mathrm{UFL}_{25} \times \left[1 + \left(\frac{0.75}{H_c}\right) \times (T - 25)\right] \quad (6)$$

where H_c is the net heat of combustion (kcal/mol), and 0.75 is essentially a molar heat capacity (kcal/mol) times 100. For many flammable gases and vapors, LFL decreases and UFL increases by approximately 8% as temperature increases by 100°C.[4]

The presence of oxidants such as chlorine, nitrous oxide, and nitric oxide can affect the limits of flammability. As the oxidant concentration is increased, UFL is markedly increased, while LFL is changed little.

The effect of oxygen concentration on flashpoint is similar to that on LFL. Experimental results of Kong et al.[5] using an equilibrium closed bomb (ECB) showed that flashpoint of the flammable liquids in air and in oxygen are essentially the same. The flashpoint of halogenated chemicals such as dichloromethane is dependent on oxygen concentration and the strength of the ignition source. Using the ECB apparatus, the flashpoint of dichloromethane in oxygen was measured to be −7.1°C, however, its flashpoint in air was found to be higher and a more energetic ignition source was required for ignition.[5]

When two or more flammable materials are present, the flammability limits of the mixture can be estimated using the LeChatelier's rule, which is expressed as follows:

$$\mathrm{LFL}_{\mathrm{mix}} = \frac{100}{\sum_{i=1}^{N} \frac{y_i}{\mathrm{LFL}_i}} \quad (7)$$

where subscripts "i" and "mix" refer to the ith fuel component and the fuel mixture; y_i (%v/v) is the concentration of the ith fuel component in the fuel mixture, excluding oxygen and inert-gas components like nitrogen; N is the number of fuel components. This equation works well for some flammable mixtures, but not particularly well for others. It may not work well for chemically dissimilar materials and should, therefore, be applied with discretion. However, this relationship reveals one important nature of mixtures of combustible gases/vapors—a mixture of more than one combustible gas can form flammable mixtures, even when the concentrations of individual combustibles are lower than their respective LFLs.

There are no reliable theoretical equations for the combined effects of pressure, temperature, oxidant type and concentration, and fuel mixture composition on the limits of flammability. However, chemical processes are often operated at elevated temperatures and pressures and at times in oxidant enriched atmospheres. Flammability limits should be measured at actual process conditions with adequate test methods.

Measurements of flammability limits

ASTM E681[6] is a widely used standard for measuring LFLs and UFLs of gases and vapors at one atmosphere or lower. A 5-L and 12-L glass flasks are recommended for testing ordinary and difficult-to-ignite materials, respectively. Ignition is defined as an upward and outward development of the flame front from the ignition source to a certain distance to the vessel wall. Electric spark or fuse wire with sufficient energy can be used as the ignition source.

According to ASTM E918,[7] high-pressure vessels with a minimum volume of 1-L and a minimum diameter of 76-mm should be used for flammability tests at high pressures. A prescribed fuse wire is recommended as the ignition source. A 7% pressure rise is defined as an ignition.

For measuring LOCs for gases/vapors, ASTM E2079[8] recommends that an almost spherical closed vessel, with internal volume not less than 4 L, be used. Several ignition sources are recommended. A 7% pressure rise is defined as an ignition.

Usually, flashpoint is measured in air. Open cup methods may overestimate the flashpoint for liquids containing multiple components because of the loss of more volatile components during testing. Nevertheless, open cup testers can provide flashpoint for situations of open vessels and spills. In contrasts, closed cup techniques prevent the loss of volatile components by keeping the sample enclosed until the ignition source is introduced, and therefore, closed cup data are more conservative than and generally preferred to open cup data.

For LTL tests, ASTM E1232 recommends that a 5-L glass flask equipped with a magnetic stirrer bar be used, which should be placed in a thermally insulated chamber. An electrical arc or fuse wire is used as the ignition source. The equilibrium closed bomb test method[5] has been successfully used to measure flashpoint in air and in oxygen under atmospheric, vacuum, and high pressures.

Ignition of Gases

When a flammable gas–oxidant mixture is heated to a sufficiently high temperature, the mixture ignites

without the presence of external energy sources such as electric sparks and this ignition process is called auto-ignition or spontaneous ignition. The lowest temperature at which a material auto-ignites is called the minimum auto-ignition temperature (AIT).

AIT is affected by the heat loss rate among other factors. The lower the heat loss rate, the lower will be the AIT. The surface (S) to volume (V) ratio affects heat loss rate and, thus, the AIT as shown by Ref.[9]:

$$\ln\left(\frac{V}{S}\right) = \frac{A}{T_{ig}} + B \tag{8}$$

where A and B are constant, and T_{ig} is the AIT. Therefore, a material may autoignite at different temperatures in different test apparatus and the larger the volume the lower the AIT.

Raising the gas mixture pressure causes the mixture to ignite at lower temperatures as shown by Ref.[9]:

$$\ln\left(\frac{P}{T_{ig}^2}\right) = \frac{E}{2RT_{ig}} + \text{constant} \tag{9}$$

For example, as the pressure increases from 1 to 10 atm, the AIT of a mineral oil is lowered from 350°C to 236°C.[2]

Chlorine affects AIT significantly, e.g., the ignition temperature of 18%(v/v) hydrogen in chlorine is 227°C, while the AIT of hydrogen in air is 400°C.[4] The surface material affects AIT by acting as a catalyst or by changing the heat transfer rate at the surface.

The AIT of combustible liquid materials can be measured using the ASTM E659 method,[10] which employs a uniformly heated 500 ml glass flask as the test vessel. A small, metered sample is introduced. The contents of the flask are observed for 10 min following the introduction of the sample, or until auto-ignition occurs. Auto-ignition is evidenced by a sudden appearance of a flame inside the flask.

Flammable gas–oxidant mixtures can be ignited by various ignition sources. Electrical devices can produce heat, which raises the surface temperature. They can also produce sparks or arcs, which concentrate the energy in small volumes lasts as short as a fraction of a microsecond. Electric sparks can be divided into two types—high voltage electrostatic sparks, and low voltage break sparks. The energy of electrostatic spark can be calculated using:

$$E = \frac{1}{2}CU^2 \tag{10}$$

where E, C, and U are the stored energy (J), capacitance (F), and voltage (V) of the conductor, respectively.

Break sparks are formed when current-carrying electrical circuits are abruptly interrupted or interrupted circuits are closed, which results in collapsing or establishing an electric field. These events result in losses of electric energy in the form of sparks. The spark energy is given by:

$$E = \frac{1}{2}LI^2 \tag{11}$$

where E is the stored energy (J), L is the inductance of the circuit (H), and I is the current before the circuit is interrupted or after a broken circuit is closed (A).

The spark energy for igniting a gas–oxidant mixture is a strong function of fuel-gas concentration (Fig. 3).[11] There exists an optimal fuel-gas concentration at which the energy for ignition is minimal and this energy is called minimum ignition energy (MIE). While the MIE may be only a few tenths of a millijoule, the energy for ignition tends to be infinitely large as the fuel-gas concentration approaches LFL and UFL.

The MIE occurs at an optimal spark gap width and varies with temperature, pressure, and oxidant concentration. The MIEs of flammable gases are usually 0.01–1.0 mJ in air and 0.002–0.1 mJ in oxygen. The MIE of gases and vapors can be determined using the method of ASTM E582.[12]

Explosion Consequences

Because of the large amount of heat release from combustion, gas explosions always involve high temperature rise. For example, the maximum flame temperatures for hydrogen and methane are 2045°C and 1875°C, respectively.[13] Even for weak deflagrations in fuel-lean mixtures near the LFL, the flame temperatures of hydrocarbons are in the range of 1300–1350°C (p. 330 in Ref.[13].). This is why even ''weak'' deflagrations such as flash fires can cause severe burn injuries.

The temperature rise and the possible increase in gas molecules as a result of combustion causes the pressure in a closed vessel to rise. The violence of a gas explosion can be expressed by the maximum explosion pressure, P_{max}, and the maximum rate of pressure rise, $(dP/dt)_{max}$, measured in a closed test vessel.[14] The product of the maximum rate of pressure rise and the cubic root of vessel volume:

$$K_G = \left(\frac{dP}{dt}\right)_{max} \times V^{\frac{1}{3}} \tag{12}$$

is called the deflagration index. Potential structural damages by explosions are determined not only by the explosion violence, but also by the strength of the structure. While P_{max} is affected little by the vessel

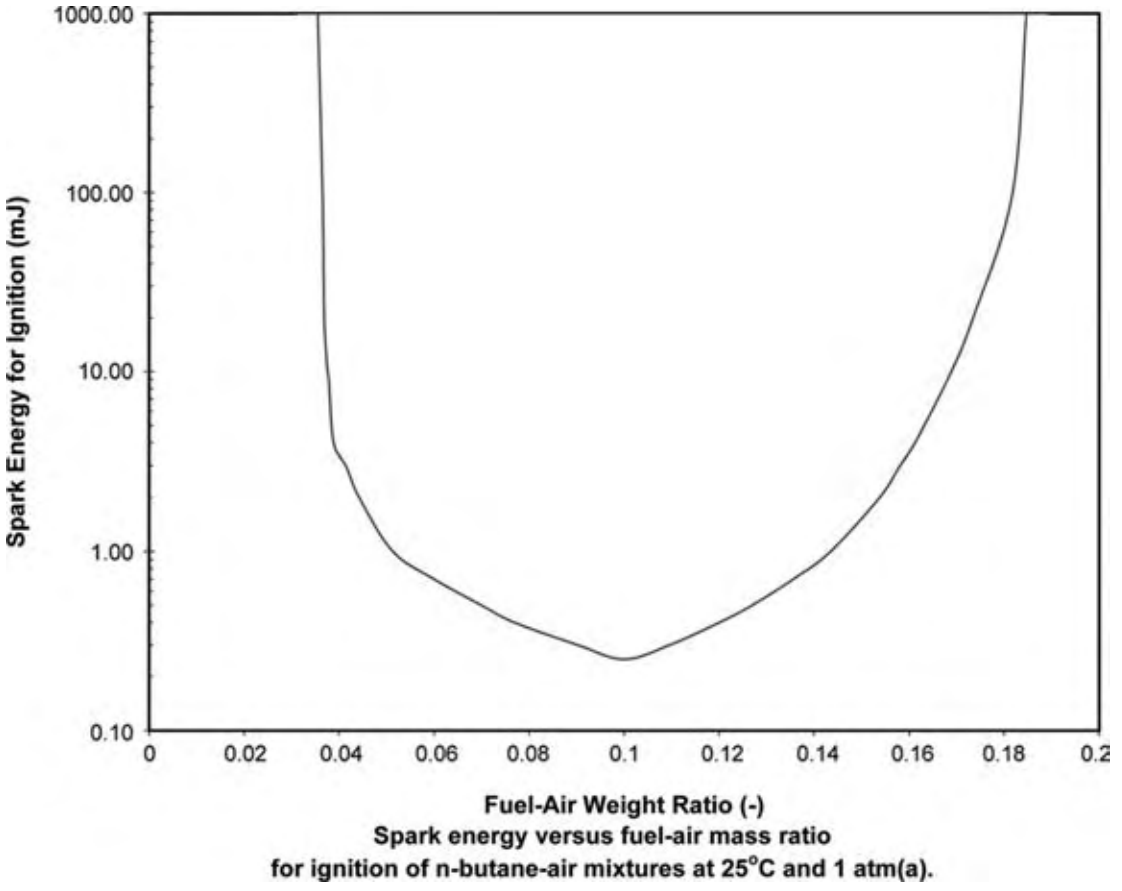


Fig. 3 Influence of fuel-gas concentration on MIE. (From Ref.[11].) *(View this art in color at www.dekker.com.)*

volume, $(dP/dt)_{max}$ and K_G are dependent on many factors, such as gas turbulence, geometry, and size of the vessel. So far, there is no standard test method for K_G measurements. K_G values have been measured using different test apparatus and the results have been used in the development of deflagration venting and suppression techniques. In order to use these guidelines, the measured K_G should be adjusted to be consistent with those data used for developing these guidelines.[15]

PREVENTION AND PROTECTION

Preventive Measures

Elimination/control of flammable gas concentration

Gas explosions can be prevented by keeping the concentration of combustibles below the LFL. For liquid combustibles, this can be achieved by controlling the process temperature below the flashpoint, provided no mists are present. For combustible gases and liquids at temperatures above their flashpoint, ventilation has to be used to control the gas concentration, usually, below 25% of the LFL.[16] Although gas explosions can be prevented by keeping the fuel-gas concentration above the UFL, there exists a risk of gas explosion as the fuel-gas concentration can be diluted to the flammable range when the process conditions are changed.

The release of flammable gases should be minimized with local exhaust ventilation, and the overall gas concentration in process buildings should be controlled with room ventilation. In addition to the common engineering guidelines for ventilation system designs, the following safety precautions should be considered:

- Ensure effective capture of combustibles at the sources.
- Minimize accumulations of condensed liquids.
- Balance airflow without dampers to prevent unwanted change of airflow rates, and eliminate accumulations of condensed liquids and solids.
- Provide air-cleaning devices (air filters, scrubbers, etc.) with protection measures against the consequences of gas explosions (refer to the section titled ''Protective Measures'').
- All ventilation components should be made from conductive and/or static dissipative materials and electrically grounded.

Elimination/control of oxidant concentration

Gas explosions can be prevented by keeping the oxidant concentration below the LOC by adding an inert gas such as nitrogen, carbon dioxide, argon, helium, steam, and flue gas. This method is called inerting.

There are three purging methods. One is called pressure purging, in which a vessel is repeatedly pressurized with an inert gas, and then exhausted to one atmosphere until the desired oxygen concentration is reached. Another method is vacuum purging, in which a vessel is repeatedly evacuated and then pressurized to 1 atm using an inert gas until the desired oxygen concentration is reached. The other type is called flow-through purging, in which a vessel is purged with a continuous flow of an inert gas for a sufficient time, with a given flow rate of the inert gas, until the desired oxygen concentration is reached. The necessary number of purges for pressure and vacuum purging, and the time for flow-through purging can be calculated using the equations given in NFPA 69.[16]

The following precautions should be exercised:

- The inert gas should be allowed to pass through the whole volume of the vessel.
- The design and operation of the protected process vessels should not compromise the inerting system.
- Oxygen sensors should be provided to monitor the oxygen concentration in the protected vessels.
- Flow meters should be provided to measure the inert-gas flow continuously, and equipped with an alarm to signal a low inert-gas flow.

Elimination/control of ignition sources

It suffices to say that all ignition sources should be excluded from areas where flammable gases or vapors may be evolved. Ignition sources that can ignite gas mixtures include but are not limited to:

- Flames and burning materials.
- Welding and cutting.
- Hot-surfaces and sparks generated by frictions.
- Heating by compression.
- Self-heating.
- Electrical sparks.
- Electrostatic discharges.
- Lightning.

Many of these ignition sources can be controlled through effective housekeeping, equipment maintenance, hot work permit, and general safety programs. Only some of the ignition sources are discussed in greater detail in this entry.

Electrical Equipment and Instruments. The ignition hazards of electrical equipment can be eliminated by:

- "Explosion-proof"—internal explosion is stopped by strong enough structures, and narrow enough openings.
- "Intrinsically safe"—the maximum energy of sparks is less than the MIE of gases.
- Placing sparking-equipment in gas-purged enclosures.

In either case, the surface temperature must not exceed the ignition temperature of the gases or vapors. To minimize the cost of using special electrical equipment, the electrical equipment is selected in such a way that the higher the likelihood of forming a flammable mixture in an area, the lower should be the risk of causing ignition by the electrical equipment.

The process of evaluating the likelihood of forming flammable atmospheres in plant areas is commonly referred to hazardous area classifications (HAC). HAC is usually conducted in accordance with regulations, codes, and standards such as the NEC/NFPA 70[17] and NFPA 497.[18] NEC/NFPA 70 classifies hazardous areas into three categories—Class I, II, and III, in which the combustible materials are gas, dust, and fibers, respectively. The degree of hazard is divided into two levels: Division 1 and Division 2.

Static Electricity. When two solid surfaces, one of which or both are electrically insulating, are brought in contact and separated, there is a rearrangement of electrons at the surfaces. One surface is left with a surplus of electrons and is said to be negatively charged, while the other is left in electron deficit and is said to be positively charged. This process of charge separation is called frictional charging or triboelectrification.

Ionic charge-carrying species in liquids are adsorbed nonuniformly at the interfaces with solids, such that charge of one polarity predominates in a tightly held "fixed thin layer," while the ions of opposite polarity within the liquid are attracted loosely by the charge on the "fixed thin layer" forming a layer near the interface. This layer is thin for conductive liquids and more diffused for less conductive liquids because of reduced ion mobility. As the liquid flows relative to the interface, the charge in the loose layer are convected away, which increases the electric potential within the liquid.

Electrification of gases flowing along solid surfaces is negligible at normal temperatures. However, liquid or solid aerosols contained in a flowing gas stream can be charged to a significant level.

As an ungrounded conductor is exposed to an electric field, some electrons travel to direction of the positive pole of the external field, leaving excessive

positive charge on the opposite side of the conductor. This process of charge separation is called induction charging. A spark will occur if the conductor is brought into contact with a grounded object, which leaves a net charge on the conductor. If the conductor is removed from the electric field, the net charge existing in the ungrounded conductor results in an elevated potential relative to ground, and a second spark may occur upon a contact with a grounded object.

Upon separation of static charges, unlike charges start immediately to neutralize each other. The actual charge buildup on the material is the balance between the separated charge and the neutralized charge. As the charges have to physically reach each other, the rate of charge relaxation is determined mainly by the electrical resistivity of the material. Obviously, the more insulating a material is, the more prone will be the material to retain electrostatic charge. Solid and liquid materials are classified into conductive, antistatic (semi-conductive), or insulating based on the electrical resistivity (ρ) and conductivity (γ) (reciprocal of resistivity), respectively as follows:[19]

Solids	
$\rho < 10^8\ \Omega$m	Conductive
$10^8 < \rho < 10^{10}\,\Omega$m	Antistatic
$\rho > 10^{10}\,\Omega$m	Insulating
Liquids	
$\gamma > 10^4$ pS/m	Conductive
10^2 pS/m $< \gamma < 10^4$ pS/m	Antistatic
$\gamma < 10^2$ pS/m	Insulating

The propensity of charge accumulation can also be expressed by the charge relaxation time, which is defined as the time required for a charge on a material to decay to a certain percentage (typically 37% ($=e^{-1}$)) of its initial value. For conductive and semiconductive materials the charge relaxation time is typically less than 1 sec.

The discharges of static electricity that occur in common industrial processes can be classified as spark, corona, brush discharges, propagating brush discharges, and cone discharges.[19]

The assessment and control of electrostatic hazards associated with process or operation requires a systematic analysis. General precautions include:

- Ground all conductive and antistatic parts.
- Avoid or minimize using insulating materials.
- Add antistatic agents to insulating liquids.
- Keep high relative humidity (usually >65%).
- Use electrostatic deionization/neutralization devices. However, these devices are not without problems, and should only be used after consulting expert advice.

Frictions and Impacts. Friction produces hot surfaces and emits small luminous objects called mechanical sparks, both of which can ignite flammable gases. Impact can cause friction between the two surfaces. Impact without friction can cause ignition of shock sensitive materials.

Friction and impact are complex mechanical processes at the surfaces where a portion of the kinetic energy is converted to heat. Whether a mechanical spark will be incendive toward a flammable gas is determined on a number of factors, which include:[9]

- Type of gas.
- Anvil material.
- Striker material.
- Geometry and roughness of the impacting surfaces.
- Angle of contact between the striker and anvil.
- Frequency of striking.
- Length of strike contacting time.
- Pressure applied at the interface.
- Energy expended at the point of striking.
- Direction of travel of sparks.
- Nature of the sparks—size, temperature, and possibility of exothermic.

In general, the higher the pressure or rubbing speed, the more incendive will be the sparks. However, quantitative predictions can be done only within narrowly defined experiments.

The temperature rise at the surface and sparks is limited by the melting point of the lower-melting-point solid. This limit does not apply if one of the materials can be easily oxidized and the oxidation is highly exothermic. Apart from direct heating because of friction, substances that can undergo a local oxidation sustain additional heating, and can build up much higher temperatures. Thermite reactions are special oxidations of metals by metal oxides. The most common thermite reaction is between aluminum and rust:[4]

$$2Al + Fe_2O_3 = 2Fe + Al_2O_3 - 15.56\text{MJ/kgAl} \quad (13)$$

This reaction releases a large amount of heat, which elevates the temperature to the order of 2200°C. Other metals, such as magnesium and titanium, can also undergo thermite reactions.

The following precautions are recommended:

- Use spark-resistant materials for moving/rotating parts.
- Avoid using aluminum tools, paints, and ladders at carbon steel surfaces.

- Avoid using aluminum paints on rusty steel surfaces.
- Avoid striking surfaces containing flint, rock, or grit with a hard object.

Protective Measures

If gas explosion hazards cannot be eliminated with preventive measures, the plants and personnel should be protected against the damaging effects of gas explosions by using protective measures as introduced below.

Explosion containment

Containment means to construct a vessel or equipment to withstand the maximum explosion pressure resulting from gas explosions. The following points should be considered:

- All interconnected pipes, flanges, covers, etc., should withstand the maximum explosion pressure.
- If an explosion resistant vessel fails, the pressure effects will be more severe than if an extremely weak vessel fails in a gas explosion.
- Explosion containment is normally considered for small-size equipment because of its relatively high cost.

Explosion venting

The principle of explosion venting is that a gas explosion in an enclosure causes the vent(s) of sufficient area to open rapidly at a pressure well below the enclosure's strength, releasing hot gases to a safe location. In other words, the vessel ''fails'' in a predictable way such that people and plant are protected from the pressure effects of the explosion.

In the USA, NFPA 68[15] provides comprehensive guidelines on deflagration venting. The required vent area is dependent on:

- K_G of the gas.
- Shape, volume, pressure rating, and available vent area of the vessel.
- Location of the vessel in relation to outside walls.
- Vent activation pressures.

Explosion venting is not suitable for materials whose combustion products are toxic, or having other adverse effects on human health or environment. It is not acceptable to vent combustion products to the inside of a building without quenching the flame. No venting recommendations are currently available for fast-burning gases with fundamental burning velocities greater than 1.3 times that of propane, such as hydrogen.[15]

Explosion suppression

Explosion suppression systems restrict and confine the fuel combustion in an early stage of an explosion by rapidly supplying flame suppressant into the fuel–air mixture. An explosion suppression system consists of the following elements:

- Explosion sensor.
- Control unit.
- High rate discharge (HRD) explosion suppressors.
- Suppressant.

The sensing and control system detects an incipient explosion and send an electrical signal to activate the valves of the pressurized HRD suppressors. Typically, the membranes of the valves are blown by explosive charges. Before the explosion suppression is achieved, combustion can develop producing modest overpressures.

Optical flame sensors can be used for detecting the initial explosion, provided interference of environmental conditions can be reliably prevented. Pressure transducers are often used because the pressure wave travels at the speed of sound and can be detected at various angles. Vibrations and other mechanical movements can interfere pressure sensors. Therefore, efforts have to be made to minimize the influence of these interferences. The suppressant must be effective in flame quenching and compatible with product and the material of the plant. The suppressors must be designed and arranged adequately, so that the suppressant is rapidly and uniformly injected to the gas mixture in the protected enclosure.

Flame/explosion isolation

Openings in and pipe connections between process plants provide paths for flame propagations. If a path for flame propagation is sufficiently long, deflagration can develop to detonation. Such a flame path should be fitted with explosion isolation devices.[16] Fast-acting shut-off valves can be closed quickly to stop flame propagation of gas explosions. Such valves are typically equipped with pneumatic actuators, which are activated upon receiving a signal of explosions that are sent from an explosion detection and control unit. The distance between a detector and the fast-acing valve should be based on the maximum flame speed expected in the duct, and the response time of the detector, valve, and the actuator circuitry.

Chemical barriers are similar to deflagration suppression systems. Typically, optical sensors are installed in upstream locations of the pipeline in which the flame is to be stopped. Upon a detection of flame, the sensor sends a signal to the control unit, which amplifies the signal and triggers the detonator-controlled valve in a suppressant bottle. The extinguishing agent is injected into the pipeline through a suitable nozzle. Pressure

sensors are not suitable for chemical barriers, as there is no clear correlation between the pressure wave front and the flame front. The distance between the sensor and the suppressant discharge point should be based on the maximum flame speed expected in the pipe, and the response time of the sensor and the discharge rate.

A liquid seal prevents the passage of flame by forcing the hot gas through a liquid. The liquid absorbs the pressure wave and quenches the flame. The liquid seal should be designed for the gases being handled at the flow velocity range in the system and to withstand the maximum anticipated deflagration pressure. Liquid seals may fail to stop flame propagation through the liquid seal for high velocity gas mixtures under certain conditions.[14,20]

Solid flame arresters are built based on the principle of flame quenching by narrow passages. If the spacing of a passage formed by solid surfaces is sufficiently small, flame can be quenched. The minimum characteristic spacing of such a passage is referred to as quenching distance. Solid flame arresters are made in a variety of forms depending on material and geometry of the passages in the flame arresters.[20]

CONCLUSIONS

A substantial amount of information on gas/vapor explosion hazards and prevention and protection techniques has been made available. However, a successful use of the available prevention and protection techniques to control gas/vapor explosion hazards can be achieved, only if the explosion hazard characteristics of the gases/vapors are determined adequately.

Most of the published explosion hazard data were obtained under standard conditions. If the actual process conditions are different from the standard conditions, the published data may not represent the actual gas explosion hazards, and the ignition sensitivity, flammability, and explosion severity/violence of the combustible gases should be measured under the actual process conditions with adequate test methods.

REFERENCES

1. ASTM D93-02. *Standard Test Methods for Flashpoint by Pensky–Martens Closed Cup Tester*; ASTM International: West Conshohocken, PA, USA, 2002 Edition.
2. Zabetakis, M.G. *Flammability Characteristics of Combustible Gases and Vapors, Bulletin 627*; Bureau of Mines: Pittsburgh, 1965.
3. ASTM E1232-02. *Standard Test Method for Temperature Limit of Flammability of Chemicals*; ASTM International: West Conshohocken, PA, USA, 2002 Edition.
4. Bodurtha, F.T. *Industrial Explosion Prevention and Protection*; McGraw-Hill Book Company, 1980.
5. Kong, D.; am Ende, D.J.; Brenek, S.J.; Weston, N.P. Determination of flash point in air and pure oxygen using an equilibrium closed bomb apparatus. J. Hazard. Mater. **2003**, *102* (2/3), 155–165.
6. ASTM E681-01. *Standard Test Method for Concentration Limits of Flammability of Chemicals (Vapors and Gases)*; ASTM International: West Conshohocken, PA, USA, 2001 Edition.
7. ASTM E918-83. *Standard Practice for Determining Limits of Flammability of Chemicals at Elevated Temperature and Pressure,* 1999 Edition.
8. ASTM E2079-01. *Standard Test Methods for Limiting Oxygen (Oxidant) Concentration in Gases and Vapors*; ASTM International: West Conshohocken, PA, USA, 2001 Edition.
9. Babrauskas, V. *Ignition Handbook*; Fire Science Publishers: Issaquash, Washington, 2003.
10. ASTM E659-78. *Standard Test Method for Auto-ignition Temperature of Liquid Chemicals*; ASTM International: West Conshohocken, PA, USA, 2000 Edition.
11. Kuchta, J.M. *Investigation of Fire and Explosion Accidents in the Chemical, Mining, and Fuel-Related Industries, Bulletin 680*; Bureau of Mines: Pittsburgh, 1985.
12. ASTM E582. *Standard Test Method for Minimum Ignition Energy and Quenching Distance in Gaseous Mixtures*; ASTM International: West Conshohocken, PA, USA, 1999 Edition.
13. Lewis, B.; von Elbe, G. *Combustion, Flames and Explosion of Gases*, 3rd Ed.; Academic Press: Orlando, FL, 1987.
14. Bartknecht, W. *Explosions. Course, Prevention, Protection*, 2nd Ed.; Springer-Verlag: New York, 1981.
15. NFPA 68. *Guide for Venting of Deflagrations*; National Fire Protection Association: Quincy, MA, USA, 2002 Edition.
16. NFPA 69. *Standard on Explosion Prevention Systems*; National Fire Protection Association: Quicy, MA, USA, 1997 Edition.
17. NFPA 70. *National Electrical Code*; National Fire Protection Association: Quicy, MA, USA, 2002 Edition.
18. NFPA 497. *Recommended Practice for the Classifications of Liquids, Gases or Vapors and of Hazardous Locations for Electrical Installations*; National Fire Protection Association: Quicy, MA, USA, 1997 Edition.
19. NFPA 77. *Recommended Practice on Static Electricity*; National Fire Protection Association: Quicy, MA, USA, 2000 Edition.
20. Grossel, S.S. *Deflagration and Detonation Flame Arresters*; Center for chemical Process Safety (CCPS) of the American Institute of Chemical Engineers: New York, 2002.

Gas–Liquid Contactors

Kishore K. Kar
Richard F. Cope
Fluid Mechanics and Mixing Group, The Dow Chemical Company, Midland, Michigan, U.S.A.

Juergen Lueske
Fluid Mechanics and Mixing Group, The Dow Chemical Company, Niedersachsen, Germany

INTRODUCTION

Gas–liquid contacting operations, which transfer one or more components between a gas phase and a liquid phase, are important to numerous industrial chemical processes. Their significance is reflected in the abundance of different contactor designs and review articles.[1,2] The importance of these operations to the chemical industry is affirmed by their global prevalence and involvement in annually producing hundreds of millions of tons of basic chemicals. The various gas–liquid contactor designs attempt to optimize controlling parameters or such specific domains as the gas–liquid interface or continuous-phase residence time.

BACKGROUND INFORMATION

To illustrate the various gas–liquid contactor designs, consider worldwide phenol production which totaled 6.1 million tonnes in 2001.[3] Bubble columns are the gas–liquid contactor typically used in a common phenol process that oxidizes cumene (at ~7 bar and ~110°C) to produce intermediate cumenehydroperoxide. The compressed air introduced into the base of the oxidizer provides oxygen for chemical reaction, as well as the agitation needed to keep the alkali and organic phases intimately mixed.

Another process example, that of fiber grade terephthalic acid (TPA), had a worldwide production of 22.7 million tonnes in 2001.[4] Stirred tank reactors (STRs) operating at 14 bar and 200°C are a preferred gas–liquid contactor in at least one continuous, multistep TPA process. This particular liquid-phase air blown process oxidizes *p*-xylene to crude terephthalic acid (CTA) in the presence of acetic acid, a bromine promoter, and a cobalt catalyst. Heat from the highly exothermic oxidation is removed by condensing and refluxing solvent, *p*-xylene, and water vapor. Another gas–liquid contactor located downstream, a packed bed hydrogenation reactor, uses hydrogen to remove color impurities (4-CBA) from aqueous CTA. Hydrogen in the reactor headspace is absorbed into a thin film of liquid pooled atop the catalyst packing. As this liquid film moves down through the packing, the absorbed hydrogen reacts with the trace liquid impurities.

More fundamentally, gas–liquid systems promote reaction in the liquid phase by utilizing the following three possible modes of gas–liquid contact: 1) gas bubbling through a liquid (as in bubble columns and stirred tanks); 2) gas contacting a thin film of flowing liquid (as with packed beds or columns); and 3) liquid droplets dispersed in gas (as in spray columns and venturi scrubbers).

The best possible mode of gas–liquid contact for a given process depends upon a combination of effects, including hydrodynamics, mass transfer, and chemical kinetics. In treating this combination, a dimensionless parameter "β" has been defined as the ratio of "total volume of the liquid phase" to "volume of the liquid diffusion layer." Krishna and Sie[5] reported general values of β to be 10–40 for thin liquid films and liquid sprays, and 10^3–10^4 for gas bubbles within a continuous liquid. The relative rates of mass transfer and chemical reaction show whether high values or low values of β best utilize available reactor volume.

For example, when mass transfer limits the overall reaction rate, reaction occurs near the gas–liquid interface to make significant liquid volume unnecessary. To minimize the value of β, bulk liquid volume is minimized, while interfacial area is maximized. As intense mixing and turbulence throughout a large liquid volume would be of limited benefit, the gas–liquid contactor of choice is a packed bed or a spray column as in the earlier TPA hydrogenation reactor example.

Conversely, if chemical reaction in the liquid phase limits the overall reaction rate, a large value of β suggests minimizing the gas–liquid interface and maximizing both liquid volume and turbulent mixing. The best contactor would disperse discrete bubbles in a continuous liquid phase. Both the bubble column

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016177

mentioned in the phenol example and the stirred reactor in the CTA example utilize this gas–liquid contacting mode.

SCOPE

After introducing the theoretical basis of gas–liquid contacting, the remainder of this entry focuses on gas–liquid contact through gas bubbles dispersed within a continuous liquid, i.e., the most common mode of gas–liquid contact within the chemical industry. Finally, the gas–liquid contactor design procedure is presented, followed by an example involving an industrial-scale STR. Further discussion of bubble columns, packed beds, thin films, and venturi scrubbers is found in related entries in this encyclopedia.

GAS–LIQUID CONTACTING IN STR

STRs create the intimate gas–liquid contact needed for mass transfer in aerobic fermentation, hydrogenation, phosgenation, neutralization, chlorination, organic oxidation, and numerous other chemical processes. Extensive experimental and computational (e.g., computational fluid dynamics) effort has been expended to develop and apply gas–liquid contacting competencies to manufacturing practices. STR design significantly affects bubble dispersion, bubble size distribution, bubble surface transients, and consequently the coefficient of mass transfer. In gas dispersion applications, an ideal agitation system performs the following tasks:

- Increases the interfacial area for mass transfer by dispersing the gas as small bubbles.
- Decreases mass transfer resistance by providing shear and turbulence to thin films and constantly renews the liquid film around the bubbles.
- Distributes the gas bubbles throughout the liquid by creating synergistic axial flow.
- Achieves homogeneous temperature and concentrations by blending the liquid.

GAS–LIQUID CONTACTOR THEORY

Film theory says that the gas–liquid interface is separated from the bulk gas and liquid phases by a gas film on one side and a liquid film on the other. While Fig. 1 shows this specifically for bubble or droplet systems, it is easily extended to liquid film contactors as well. Resistance in both films decreases the concentration of species diffusing across the gas–liquid interface irrespective of whether that diffusion is from gas to liquid (as depicted in Fig. 1) or from liquid to gas.

Because diffusion in the liquid phase is much slower than in the gas phase, the liquid film presents most of the resistance to mass transfer. Gas–liquid film theory defines the mass transfer rate (MTR) in terms of the liquid film, as follows

$$\begin{aligned} \text{MTR} &= k_{\text{L}}a(C^* - C_{\text{L}}) \\ &= k_{\text{L}}a(C_{\text{g}}H_{\text{e}} - C_{\text{L}}) \ (\text{mol/m}^3\,\text{s}) \end{aligned} \tag{1}$$

where k_{L} is the liquid film mass transfer coefficient (m/s), a is the interfacial area per unit volume (m^{-1}), C^* is the saturation concentration at the gas–liquid interface (mol/m^3), C_{L} is the bulk concentration in the liquid phase (mol/m^3), C_{g} is the transfer gas concentration in the bubble (mol/m^3), H_{e} is Henry's law constant [$H_{\text{e}} = C^*RT/p_{\text{g}}$], R is the universal gas constant [$\equiv$ 8.314] (J/mol K), T is the temperature at the gas–liquid interface (K), and p_{g} is the transfer gas pressure in the bubble (Pa).

Values of k_{L} typically average $0.02\,\text{cm/s} \pm 0.02$[5] depending on diffusivity of the dissolved gas and relative velocities of the gas and liquid phases. The value of "a" can vary by several orders of magnitude especially in bubble systems where it is inversely proportional to bubble size. Because of difficulties in measuring k_{L} and "a" individually, the lumped parameter or volumetric mass transfer coefficient $k_{\text{L}}a$ (i.e., k_{L} times "a") is typically used to determine MTR. The value of $k_{\text{L}}a$ can be raised by increasing superficial gas velocity and power input. A higher $k_{\text{L}}a$ can cut capital costs by producing a comparable MTR with a smaller vessel, a shorter batch process time, or a higher continuous process throughput.

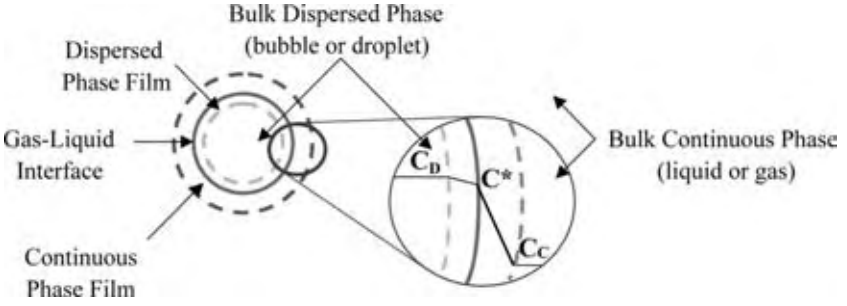


Fig. 1 Gas–liquid film theory for diffusion from the dispersed phase to the continuous phase (with maximum diffusion resistance in the continuous phase film). *(View this art in color at www.dekker.com.)*

EXPERIMENTAL DETERMINATION OF k_La

Values of k_La can be evaluated for a diffusing gas (e.g., O_2) as a function of operating conditions by using an appropriate dissolved gas probe, and either a transient or a steady-state experimental measurement method. While these methods apply to all diffusing gas species, they will be confined to oxygen for the sake of this discussion. The transient method of determining k_La for oxygen at a given set of operating conditions begins by sparging an inert gas (e.g., N_2) into the liquid to purge it of dissolved oxygen (DO). After the DO probe reading indicates essentially zero concentration, the flowing N_2 is turned off and replaced with air. The increasing concentration of DO in the liquid phase "$C(t)$" is recorded as a function of time. As the gas dissolution process can be described by Eq. (2), the corresponding value of k_La is obtained by regressing the experimental data as suggested in Eq. (3)

$$\ln\left[\frac{C^* - C_L}{C^* - C(t)}\right] = k_La \cdot t \tag{2}$$

$$C(t) = C^* + (C_L - C^*)\exp(-k_La \cdot t)\ (\text{mol/m}^3) \tag{3}$$

While these equations include interfacial concentration of the DO (C^*), its value need not be known to determine k_La from the "$C(t)$ vs. t" data regression shown in Eq. (3). Accurate determination of k_La requires a DO probe with adequate response time.

The steady-state k_La method uses chemical reaction in the liquid phase to generate oxygen at a "steady-state" rate. Oxygen is transferred from the bulk liquid into the gas bubbles, an exact reversal of the transfer direction employed in the transient k_La measurement method. For example, the NEL/Hickman[6] steady-state method for measuring oxygen mass transfer continuously feeds hydrogen peroxide solution (H_2O_2) and dry air into the liquid at known rates. Bovine catalase in the liquid decomposes H_2O_2 into H_2O and O_2.

The DO concentration in the liquid is monitored with a DO probe and recorded after it reaches a steady-state (i.e., constant) value. The oxygen generation rate (OGR) in the liquid equals the oxygen transfer rate (OTR), and is related to the hydrogen peroxide mass feed rate ($Q_{H_2O_2}$) as shown in Eq. (4)

$$\text{OGR} = \left[\frac{Q_{H_2O_2}}{2 \cdot \text{MW}_{H_2O_2}} \cdot \frac{1}{V_L}\right] = \text{OTR} = k_La \cdot (C_L - C^*)\ (\text{mol/m}^3\text{/s}) \tag{4}$$

where $\text{MW}_{H_2O_2}$ is the H_2O_2 molecular weight (34 g/mol), V_L is the volume of liquid phase (corrected for H_2O_2 addition) (m^3), and C^* is the average saturation concentration at the gas–liquid interface (mol/m^3).

The H_2O_2 feed rate is divided by 2 because of the stoichiometry of oxygen formation. As liquid temperature and salinity change during a measurement, oxygen concentrations in the liquid must be corrected accordingly. In addition, use of this method requires the following assumptions:

- Gas phase is ideal gas.
- Liquid phase is well mixed.
- Oxygen concentration at the gas–liquid interface is in equilibrium with that in the bulk gas.
- Oxygen partial pressure in air is fixed at 0.2095 atm.

The presence of H_2O_2 and salt in the liquid phase generally elevates the associated k_La values above those obtained by the transient method.[7] A slight modification of this steady-state method allows the authors to use ferric chloride instead of bovine catalase in large-scale experiments.

MASS TRANSFER COEFFICIENT

The mass transfer coefficient, k_La, is correlated with power per mass, superficial gas velocity, and media viscosity as follows

$$k_La_{20} = \alpha(\varepsilon)^{\beta}(V_{sg})^{\gamma}(\eta_r)^{\kappa}\ (\text{s}^{-1}) \tag{5}$$

$$k_La_{20} = k_La(1.024)^{(20-t)}\ (\text{s}^{-1}) \tag{6}$$

where ε is the specific energy (shaft power per unit mass) under gassed conditions (W/kg), V_{sg} is the superficial gas velocity [vol. gas flow rate (m^3/s)/tank cross-sectional area (m^2)] (m/s), η_r is the relative viscosity, i.e., viscosity of working media/viscosity of water, and k_La_{20} is the mass transfer coefficient at 20°C.[8]

Example values of the coefficient α and exponents β, γ, κ are shown in Table 1[9] for an air–water system that was extended to higher viscosities as noted. The value of α can change significantly for different working media. For example, $k_La_{\text{saltwater}} = 1.5\text{–}2 \times k_La_{\text{purewater}}$. Furthermore, k_La is less dependent on specific power at higher levels of power input, such as $\varepsilon > 3$ W/kg (see Table 2). In practice, appropriate values for each variable in the k_La correlation should be determined for specific process or simulation fluids of interest.

Effect of Surface Active Agents

The presence of low molecular weight surfactants in the working media reduces interfacial tension and

Table 1 Fitting parameters for a coalescing gas–liquid system ($\varepsilon < 3\,W/kg$)

Source	Working media	Interfacial characteristics	Viscosity ratio	α	β	γ	κ
van't Riet/Dow	Air–water	72 dynes/cm	1	0.412	0.4	0.5	−0.46[a]

[a]Extended correlation using k_La data from noncoalescing 5 cP glucose broth, 10–300 cP penicillin cultures, and 2000 cP polyglycol.

resulting bubble sizes. Consequently, k_La increases significantly. In one instance, the presence of 1 wt% formic acid (free surface ions) in an agricultural intermediate surprisingly yielded the 34% increase in k_La seen in Table 2.[10] High concentrations of high molecular weight surfactants, however, tend to decrease k_La by increasing the viscosity of liquid film surrounding the bubble.

Effect of Solids on the Mass Transfer Coefficient

In gas–liquid mixing applications, solids are often present as a catalyst or a product. Inherently having small solids present on the bubble surface can diminish the effective intrinsic bubble surface area and significantly decrease the mass transfer coefficient. Though others report just the opposite effect, an illustrative example from Dow is shown in Table 3.[10]

GAS HOLDUP

Besides the mass transfer coefficient, k_La, other parameters help quantify gas–liquid contactor performance in bubble systems. One such parameter is gas holdup (ε_g), which is the ratio of gas volume within the media to the mixture volume as shown in Eq. (7)

$$\varepsilon_g = (V_g)/(V_g + V_L) \tag{7}$$

where V_g is the gas-phase volume (m^3) and V_L is the liquid-phase volume (m^3).

Sometimes gas holdup information is expressed as the fluid volume expansion factor, or $1/(1 - \varepsilon_g)$. Gas holdup is correlated with power per mass (ε) and superficial gas velocity (V_{sg}) as follows

$$\varepsilon_g = \alpha_1(\varepsilon)^{\beta_1}(V_{sg})^{\gamma_1} \tag{8}$$

Table 2 Effect of surface active agents (free surface ions) on k_La ($\varepsilon > 3\,W/kg$, at 40 °C)

Working medium	α	β	γ
Dilute slurry + SARAN particles	0.945	0.05	0.47
Dilute slurry + SARAN particles + formic acid	1.265	0.05	0.47

For the air–water (coalescing) system at specified conditions (8), $\alpha = 0.65$, $\beta = 0.20$, and $\gamma = 0.55$.

Gas holdup is an important parameter for fill volume calculations and reactor sizing. In the case of aqueous coalescing systems, gas holdup normally ranges between 10% and 20%. This is also typical of most aerobic bioreactors. In aqueous noncoalescing systems or viscous systems, gas holdup can be appreciably higher (e.g., 50%). For example, gas holdup in a ~12% acetic acid solution is in the range of 45–50% depending on power level and superficial gas velocity. Significantly, a typical coalescing medium such as water may often be converted into a noncoalescing medium with the addition of salt or other electrolytes.

Higher gas holdup generally produces significant increases in the volumetric mass transfer coefficient, k_La. An exception, however, is high viscosity media where initially many tiny (<2 mm) noncoalescing bubbles and a few relatively large (50–300 mm) coalescing bubbles form. As viscous forces dominate over local turbulence (inertial forces), the tiny bubbles are trapped in the liquid phase and have little-to-no relative velocity with respect to the bulk liquid motion (i.e., the two-phase media behave as a single liquid phase). This phenomenon limits bubble surface transience and gas disengagement, and consequently, the volumetric MTR. Simultaneously, the large bubbles quickly disengage (burp) from the media with little contribution to the mass transfer.

AGITATION SYSTEM

A modern high efficiency gas–liquid STR commonly consists of a radial gas dispersing impeller in combination with one or more axial down-pumping impellers, as seen in Fig. 2. The type and number of open impellers used in the tank depend on the gassed liquid height and the media viscosity (see Table 4). Recommended axial impellers are 45° four-bladed pitch blade turbines

Table 3 Effect of solid SARAN particles on k_La ($0.4 < \varepsilon < 3\,W/kg$, at 22 °C)

Working medium	α	β	γ
Slurry	0.476	0.52	0.46
Slurry + SARAN	0.378	0.53	0.44

Fig. 2 A typical high efficiency gas–liquid mixing system. *(View this art in color at www.dekker.com.)*

(PBTs) for lower viscosity (<200 cP) working media. High efficiency low solidity three-bladed hydrofoils such as Lightnin A-320s and Chemineer Maxflow are suggested for higher (≥200 cP) viscosity working media. Possible radial impellers include the Rushton turbine and such concave-blade turbines as the Smith turbine (see Fig. 3). The authors recommend six-bladed Smith turbines over Rushtons. Normally the impeller diameter to tank diameter ratio ranges between 0.4 and 0.54.

Table 4 shows the recommended numbers of impellers used in low viscosity working media. In the case of high viscosity media, the impeller spacing, S, normally needs to be decreased, i.e., $S = 0.75$–$1.0\,D$. Multiple axial impellers are desirable for a synergistic flow pattern.

Down-Pumping vs. Up-Pumping

In recent years, some mixer manufacturers have promoted agitation systems that are up-pumping as opposed to down-pumping.[11] To verify their claims, the authors conducted several experiments in both low viscosity medium (water) and high viscosity medium (polyglycol) using the up-pumping configuration shown in Fig. 4A and the corresponding down-pumping configuration. Results from these experiments (Fig. 4B) show the down-pumping k_La to be 20% higher than the corresponding up-pumping value.

Power Draw

The power draw characteristics of open impellers are expressed in terms of power number (N_P) vs. Reynolds number (N_{Re}). By definition,

Table 4 Multiple impeller arrangement

Gassed height/ tank diameter (Z_g/T)	<0.9	0.9–1.7	1.7–2.5	2.5–3.3
Number of impellers	1	2	3	4

$$N_p = P/(\rho \cdot N^3 \cdot D^5) \tag{9}$$

$$N_{Re} = \rho \cdot N \cdot D^2/\mu \tag{10}$$

where P is the shaft (hydraulic) horse power (W), ρ the density of the working media (kg/m^3), and μ the viscosity of the working media (kg/m s or Pa s).

Impeller power number is fairly constant for the fully turbulent ($N_{Re} > 10^4$) flows encountered in most production-scale vessels working with a low viscosity media. For various reasons including synergy between impellers, the N_P vs. N_{Re} relationship approaches an asymptote more slowly in multiple impeller systems than in single impeller systems. Fig. 5 shows power number vs. Reynolds number characteristics for two different dual impeller systems, i.e., Rushton turbine + PBT and concave disk turbine + PBT. Slopes of Fig. 5 curves flatten as $N_{Re} > 10^6$ (beyond the fully turbulent flow regime), and increase significantly where $N_{Re} < 75{,}000$ (the transition flow regime). The ungassed fully turbulent flow power numbers for the standard Rushton, Smith, and PBTs are 4.6, 3.2, and 1.27, respectively.

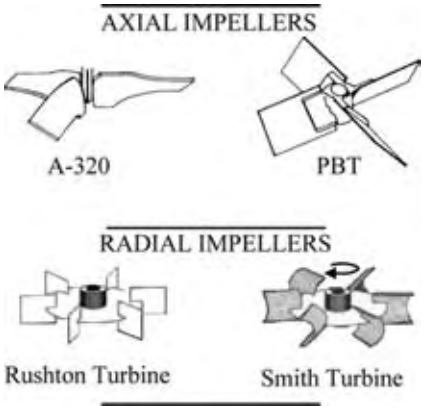


Fig. 3 Examples of common axial and radial impellers. *(View this art in color at www.dekker.com.)*

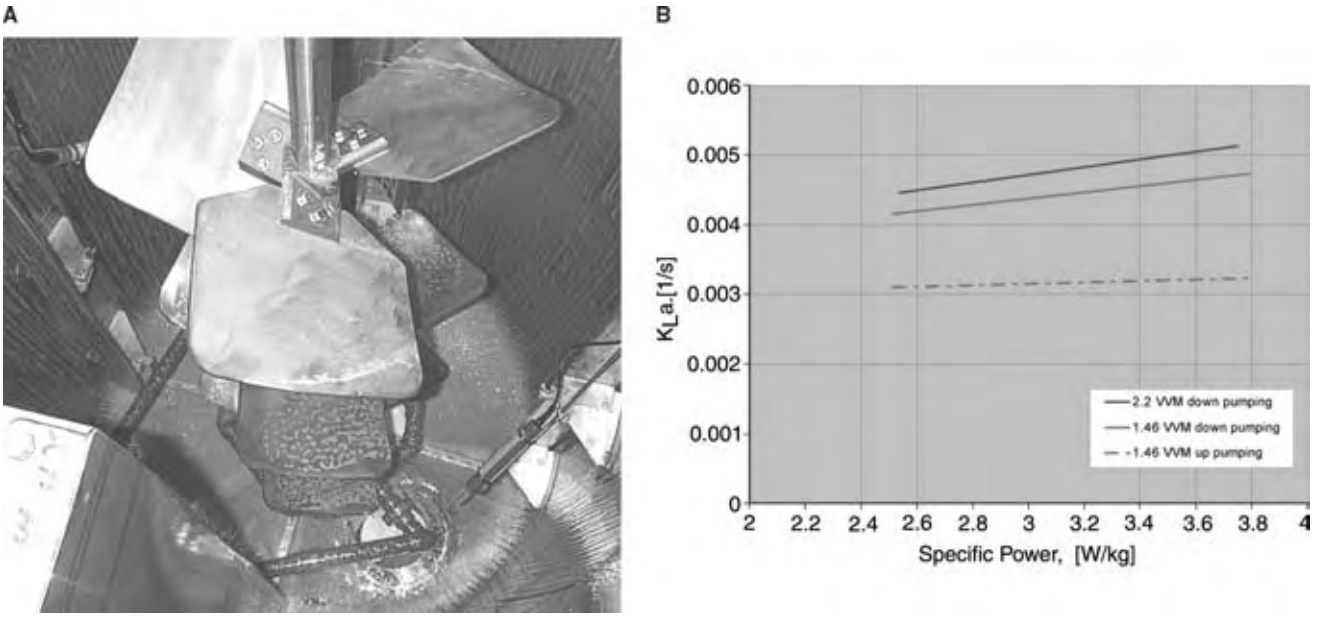


Fig. 4 Up-pumping (A-320) agitation system (A) photograph (counterclockwise rotation), and (B) experimental k_La results in a polyglycol (2000 cP) medium. *(View this art in color at www.dekker.com.)*

The *k*-Factor

The k-factor is the ratio of the power drawn under gassed conditions to that under ungassed conditions, i.e., P_g/P_{ug}. Its value is generally less than unity and varies with changes in such operating conditions as gas flow rate, impeller type, and impeller speed. The k-factor decreases with increasing aeration number N_a as shown in Fig. 6. While lower power draw under gassed conditions is generally not considered a problem, it becomes one if the system motor power is sufficient under gassed conditions but insufficient during system startup under ungassed conditions. In the limit of fully turbulent gassed conditions, the power draw of all the gas dispersing impellers approaches a similar value.

Fig. 6 shows k-factor vs. aeration number curves for the two dual agitator combinations discussed previously. As the power draw (and the power number) is higher under ungassed conditions, agitators typically operate at lower speeds when ungassed. The change in power draw with increasing gas flow is managed by a two-speed motor or a variable frequency drive (VFD) in the drive system.

For design purposes, one could use the k-factors listed in Table 5 for open impellers operating at $N_{Re} \geq 100{,}000$ and $N_a \geq 0.03$.

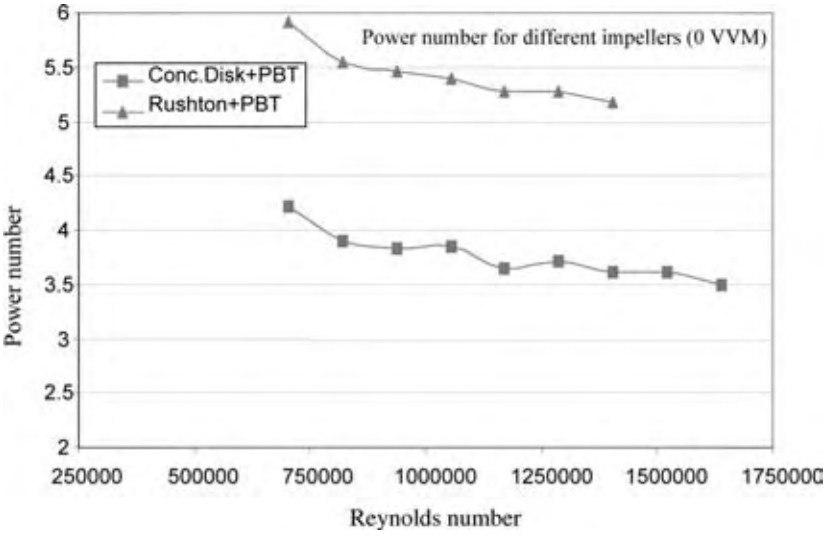


Fig. 5 Power number vs. Reynolds number for two agitation systems. *(View this art in color at www.dekker.com.)*

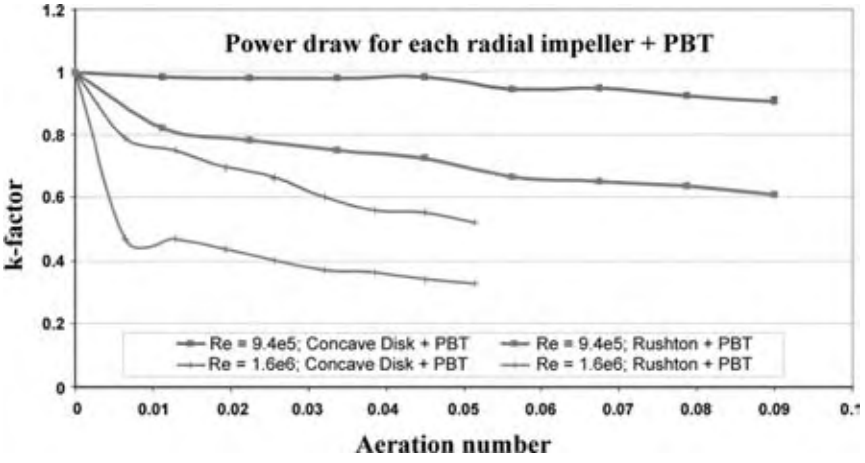


Fig. 6 *k*-Factor vs. aeration number for the two dual agitation systems. *(View this art in color at www.dekker.com.)*

Criteria for Impeller Flooding

The Froude number (N_{Fr}) characterizes gas flow regimes using the following ratio of inertial forces to gravitational forces

$$N_{Fr} = N^2 D/g \tag{11}$$

where g is the acceleration due to gravity (9.81 m/s^2).

The aeration number (N_a) is the ratio of gas flow rate to impeller discharge flow and is written as

$$N_a = Q_g/ND^3 \tag{12}$$

where Q_g is the actual gas flow rate (m^3/s), N is the agitator speed (s^{-1}), and D is the gas dispersing impeller diameter (m).

For a given gas flow rate, the dispersion pattern of gas bubbles in the working media depends on the interplay between the isothermal expansion energy of the gas and the hydraulic power of the impeller. One of three bubble dispersion scenarios is typically expected, i.e., flooding, dispersion in the upper region only, or complete dispersion. Fig. 7[12] illustrates each of these flow regimes.

In water, the maximum aeration number at which flooding occurs ($N_{a,FL}$) with one of several different impellers is correlated[12] as:

$$N_{a,FL} = C_{FL} N_{Fr} (D/T)^{3.5} \tag{13}$$

Table 5 *k*-Factors for different impellers

Impellers	*k*-Factors
Rushton	0.35
CD-6	0.7
PBT	0.9
A-320	0.9

where Rushton $C_{FL} = 30$ and Smith turbine $C_{FL} = 70$.

In the same water system, gas is completely dispersed at transition aeration numbers ($N_{a,CD}$) given by the following correlation:

$$N_{a,CD} = C_{CD}\, N_{FR}^{0.5} (D/T)^{0.5} \tag{14}$$

where Rushton $C_{CD} = 0.2$ and Smith turbine $C_{CD} = 0.4$.

A rule of thumb is that one can expect complete dispersion when the desirable gassed shaft power is greater than three times (minimum two times) the isothermal gas expansion power [defined in Eq. (15)]. A design check for adequate shaft power is absolutely necessary to avoid an undesirable flooding situation

$$P_{gas} = p_1 Q_1 \ln(p_1/p_2) \tag{15}$$

where P_{gas} is the power delivered by the isothermally expanding gas (W), p_1 is the gas pressure (hydrostatic + back pressure) at the sparger (Pa), p_2 is the back pressure (Pa), and Q_1 is the actual volumetric gas flow rate at the sparger (m^3/s).

Gas Sparger

The gas sparger is a critical internal component of STRs used in gas–liquid applications. The type of sparger used depends on the ratio of volumetric gas flow to liquid volume (VVM), and the sizes of any particles present.

Fig. 8 shows the different types of spargers used in STRs. For systems containing large (>50 microns) settling particles, tangential sparger nozzles are more appropriate than common ring spargers. Open pipe spargers provide adequate gas delivery under low (<0.2 min^{-1}) VVM conditions. For high gas flow

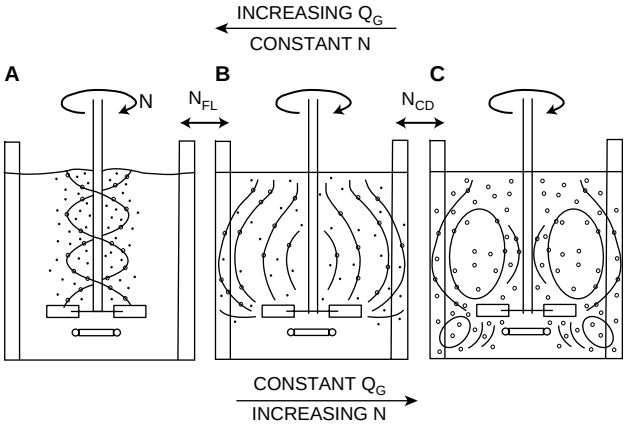


Fig. 7 Gas bubble dispersion in an STR.

rates (high VVM, i.e., $>2.0\,\text{min}^{-1}$), plate spargers are typical, though an alternative approach uses a draft-tube impeller agitation system. Sometimes, gas is self-induced through the rotation of hollow shafts and blades.[13] Disadvantages of self-induction include susceptibility to plugging in sticky polymers and variation of gas flow with hydrostatic head pressure.

Tickler (Kicker)

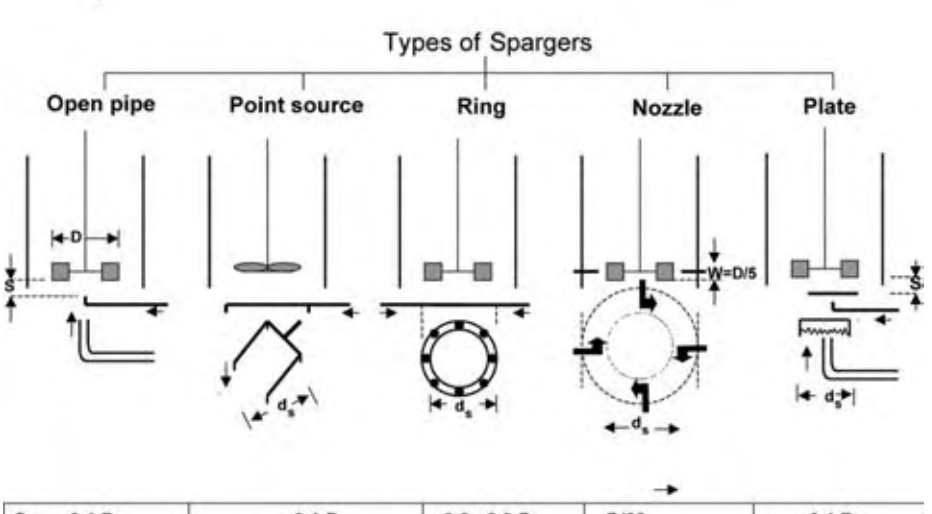
A tickler is a relatively smaller impeller placed at close proximity to the vessel bottom. Often in gas–liquid mixing systems that contain some solids, an axial- (up- or down-pumping) or a radial tickler is used to overcome the solid suspension deficiency of radial gas dispersing impellers. Pitched blade or flat blade turbines are normally used as ticklers. In the case of gas–liquid–solid systems where the slurry needs to be drained through a central nozzle, the authors recommend the Dow-designed KT-3 tickler. Use this tickler prevents pump starvation and facilitates solid removal.

Steady Bearing

Gas–liquid mixing systems typically operate at high power levels, i.e., 1–4 W/kg, and often involve tall vessels ($H/T > 2$) with long mixing shafts. Most tall vessels incorporate a steady bearing or a limit ring to prevent the shaft deflection caused by severe transient loading of the agitation system. Unfortunately, steady bearings wear out (requiring regular maintenance), are susceptible to fouling in sticky polymers, and are often recognized as a contamination source in bioreactors. Thus, caution must be exercised and vendors should be consulted during the consideration and mechanical design of steady bearings.

Baffles

In stirred tanks for gas–liquid mixing, baffles are essential to create axial flow and maintain overall



S = 0.1 D	0.1 D	0.2 - 0.3 D	- D/20	0.1 D
d_s = -	1.3 - 1.4 D	0.7 - 0.8 D	1.1 D	0.7 D
V_p = 11 - 23 (~18)	11 - 23	11 - 23	11 - 23	11 - 23
EX: Low VVM	Axial Impellers + Solids	+ Fine light solids	+Solids	High VVM

V_p = Gas velocity in the pipe, m/s

Fig. 8 Sparger types and design rules. *(View this art in color at www.dekker.com.)*

uniformity in the mixture. Normally, four simple flat baffles are recommended. The baffle width and baffle clearance from the wall are usually in the range of $^1/_{12}$ – $^1/_8$ and $^1/_{64}$ – $^1/_{48}$ times the tank diameter, respectively. Profiled baffles are preferred for sticky high viscosity or shear-sensitive materials.

STIRRED TANK DESIGN METHODOLOGY

As discussed earlier, stirred tanks are desirable when the overall reaction rate is limited by liquid-phase reaction. Bubbling gas through well-mixed liquid in a stirred tank should minimize the thickness of the gas–liquid interface while maximizing the availability of the liquid volume ($\beta = 10^3$–10^4). Consequently, process conditions that achieve this objective are the first consideration in designing stirred tank gas–liquid contactors.

The optimal configuration depends on numerous factors, including the required gas transfer and mixing rates, and an acceptable range of shear rates. Some factors may be mutually contradictory as in a crystallizer where the required gas–liquid transfer rate is best achieved by increasing turbulence and shear; yet the crystals are shear sensitive. In these cases, successful design carefully balances the contradictory factors.

This section describes the design of a stirred tank contactor for a given range of operating variables and includes the selection of geometric parameters. Important correlations and formulae are presented and followed by an example design case. The design process actually requires multiple iterations to simultaneously satisfy both geometric and mass transfer requirements, but such iteration has become almost trivial when using tools such as spreadsheet-based ''goal-seek'' computations.

Process conditions that must be considered early in the design process include:

- Desired batch size (required tank dimensions) based on required yield.
- Physical and thermal properties of the working media.
- Volumetric gas demand based on reaction kinetics.

Gas demand influences mass flow meter selection, as indicated in the following calculations:

- Actual cubic meters per minute (ACMM) of the reacting gas.
- ACMM of (reacting + carrier) gas based on percentage of reacting gas consumed.
- VVM (gas volumetric flow rate per minute/volume of liquid).
- Standard cubic meters per minute (SCMM) in selecting a mass flow meter.

Design of the gas sparger and agitation system can proceed at this point:

1. Sparger:
   a. Choose the sparger type.
   b. Determine superficial gas velocity from the required ACMM at the sparger.
2. Agitation system (radial + axial impellers):
   a. Calculate C^* and required bulk mean gas concentration in the liquid phase.
   b. Determine k_La from the gas transfer rate.
   c. Calculate the overall transfer rate $\text{OTR} = k_La(C^* - C_L)$ or $k_La\ (\Delta C_{lm})$.
   d. Predict gassed power per mass using k_La correlation (α, β, γ, and κ).
   e. Determine gassed power, P_g.
   f. Determine ungassed power per mass using k-factor.
   g. Calculate gas isothermal expansion energy.
   h. Check the P_g/P_{gas} ratio (>2 as a minimum; >3 to ensure full dispersion).
   i. Check the flooding conditions for the gas dispersing impeller.
   j. Check for gas holdup and recheck tank dimensions.

Table 6 Specification of reactor dimensions, sparger type, and mass flow meter

Tank diameter (m)	3.80
Tank height: tangent–tangent (m)	4.66
Nominal tank volume (m^3)	60
Working media volume (m^3)	45
Unaerated liquid height (m)	4.3
Liquid height/tank diameter	1.13
Volumetric gas flow rate (m^3s^{-1})	0.21
Midpoint pressure (Pa)	1.57×10^5
Pressure at the sparger (Pa)	1.78×10^5
Required ACMM of oxygen (m^3min^{-1})	3.0
Required ACMM of air (m^3min^{-1})	14.4
VVM (min^{-1})	0.37
Required SCMM (m^3min^{-1})	23.6, specify mass flow meter
Superficial gas velocity at the sparger (m^3s^{-1})	0.031
Type of sparger	Ring

Table 7 Design of agitation system

C^* ($mol\,m^{-3}$)	1.72	
C_L ($mol\,m^{-3}$)	1.0	
Henry's law constant ($mol\,m^{-3}\,Pa^{-1}$)	1.26×10^{-5}	
Volumetric mass transfer coefficient (s^{-1})	0.12	Eq. (4)
Power per mass (W/kg)	1.82	Eqs. (5) and (6)
Gassed power (kW)	82	Eq. (9)
k-Factor	0.7	Table 5
Ungassed power (kW)	117	
Gas isothermal expansion power (kW)	14.4	Eq. (15)
Gassed agitation power/gas expansion power	5.7	>3 Check
Aeration number for flooding conditions	2.10	Eqs. (11)–(14) Check
Gas holdup	0.11	Eq. (8)
Aerated liquid height (m)	5.0	Recheck tank dimensions

k. Specify agitation system:
   i. Motor power.
   ii. Gearbox.
   iii. VFD or two-speed motor.
   iv. Operating conditions.

The optimal stirred tank design is identified through iterating until both geometric and mass transfer requirements are simultaneously satisfied.

Design Example

The design of an STR for a fast reacting oxidation process is illustrated in the following section. Operational parameters include a process oxygen demand of $250\,mmol\,l^{-1}\,hr^{-1}$ and a required batch size of $45\,m^3$. Reactor absolute pressure and temperature are 1.32 bar and 30°C, respectively. Minimum required oxygen concentration in the liquid is $1.0\,mol\,m^{-3}$. Physical properties of this Newtonian working media are: viscosity, 1×10^{-3} Pas; density, $1000\,kg\,m^{-3}$; and interfacial tension (gas–liquid), $7.1 \times 10^{-2}\,N\,m^{-1}$.

Design Worksheet

Based upon product batch requirements, preliminary vessel dimensions are listed in Table 6. Gas flow rate (ACMM) is determined to obtain the mass transfer required by reaction kinetics. A ring sparger is selected as discussed in conjunction with Fig. 8. Table 7 summarizes the calculation of the required specific power and gas holdup based on mass transfer and the desired oxygen concentration in the liquid. It must be confirmed that vessel dimensions accommodate the aerated liquid height. An appropriate agitation drive system is recommended in Table 8. Fig. 9 is a schematic of the reactor in this design example.

CONCLUSIONS

This entry briefly presents the mass transfer fundamentals of gas–liquid contacting followed by a series of experimental correlations for the design of STRs. Discussions are intended to familiarize practicing engineers with how to design, optimize, and troubleshoot

Table 8 Specification of agitation system

Number of impellers	2	
Radial impeller: Smith turbine	1	$N_p = 3.2$
Axial impeller(s): four-blade-45° PBT	1	$N_p = 1.27$
Impeller diameter (m)	1.7	$D/T = 0.50$
Off-bottom clearance (m)	0.85	$C/T = 0.25$
Impeller spacing (m)	1.7	$S/T = 1.0$
Shaft speed under gassed condition (rpm)	74 (AGMA = 84)	Eq. (9)
Shaft speed under ungassed condition (rpm)	66 (AGMA = 56)	Eq. (9)

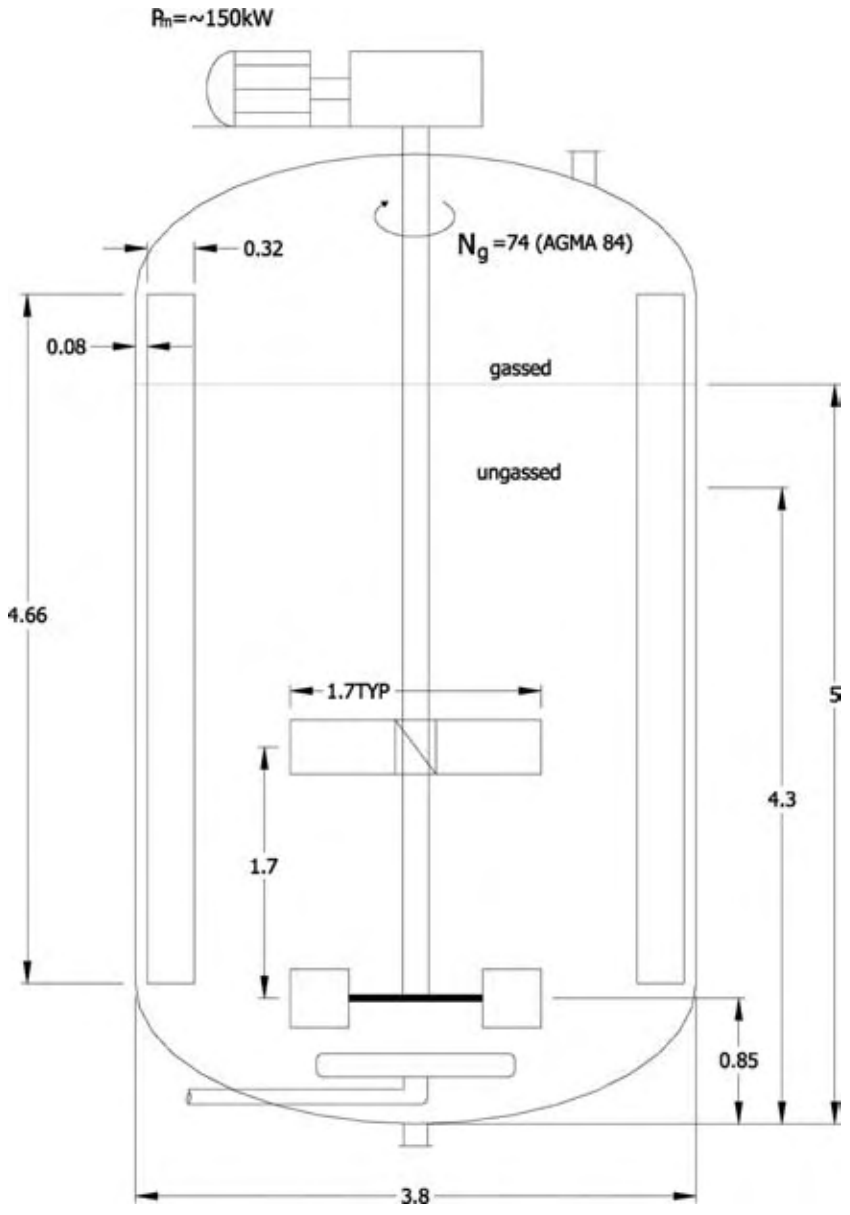


Fig. 9 Schematic of the production stirred tank reactor (dimensions in meters).

stirred tank gas–liquid contactors. To illustrate the design methodology, the design of a typical industrial-scale stirred tank oxidation reactor is presented in a worksheet. Gas–liquid contact in bubble columns, packed beds, thin films, and venturi scrubbers is also discussed in related chapters.

ACKNOWLEDGMENTS

The authors would like to thank the Dow Chemical Company for permitting external release of the information in this chapter, Erik Gasiewski for CAD drawings the figures, and Jan Tuinstra for her final review. Gratitude is also expressed to Tim Geiger of Mattoon & Lee Equipment, Inc. for valuable feedback on this topic.

NOMENCLATURE

a	Gas–liquid interfacial area per unit volume (m^{-1})
c	Impeller off-bottom clearance (m)
C	Concentration of diffusing gas ($mol\,m^{-3}$)
C^*	Saturation concentration at the gas–liquid interface ($mol\,m^{-3}$)
C_L	Bulk concentration in the liquid phase ($mol\,m^{-3}$)

C_g	Transfer gas concentration in the bubble ($mol\,m^{-3}$)
k_L	Liquid film mass transfer coefficient (m/s)
D	Impeller diameter (m)
g	Acceleration due to gravity ($m\,s^{-2}$)
H	Tank height (m)
H_e	Henry's law constant ($H_e = C^*RT/p_g$)
MW	Molecular weight ($g\,mol^{-1}$)
N	Impeller speed (rpm)
p	Pressure (Pa)
P	Power (W)
Q	Volumetric gas flow rate ($m^3\,s^{-1}$)
R	Universal gas constant (= 8.314) ($m^3\,Pa\,mol^{-1}\,K^{-1}$)
S	Impeller spacing (m)
t	Time (s)
T	Temperature at the gas–liquid interface (K)
T	Tank diameter (m)
V	Volume (m^3)
V_{sg}	Superficial gas velocity ($m\,s^{-1}$)

Greek Letters

ε	Power per mass ($W\,kg^{-1}$)
ε_g	Gas holdup
ρ	Density ($kg\,m^{-3}$)
μ	Dynamic viscosity (Pa s)
ν	Kinematic viscosity ($m^2\,s^{-1}$)
η_r	Viscosity ratio of the liquid phase to that of water

Subscripts

g	Gassed or gas phase
gas	Gas expansion
ug	Ungassed
L	Liquid phase

ARTICLE OF FURTHER INTEREST

Fermenter Design, p. 951.

REFERENCES

1. Middleton, J.; Smith, J. Gas–liquid mixing in turbulent systems. In *Handbook of Industrial Mixing: Science and Practice*; Paul, E.L., Atiemo-Obeng, V.A., Kresta, S.M., Eds.; John Wiley and Sons: New York, 2003; 585–638.
2. Lee, S.; Tsui, Y. Succeed at gas/liquid contacting. Chem. Eng. Prog. **1999**, 23–49.
3. SRI-International, Phenol. In *Chemical Economics Handbook*: 2002; 4.
4. World Tables, *p*-xylene & Derivatives World Supply & Demand Report, 2002; 52.
5. Krishna, R.; Sie, S.T. Strategies for multiphase reactor selection. Chem. Eng. Sci. **1994**, *49*, 4029–4065.
6. Cooke, M.; Dawson, M.K.; Nienow, A.W.; Moody, G.W.; Whitton, M.J. Mass transfer in aerated agitated: vessels assessment of the NEL/Hickman steady state method. Proceedings of Seventh European Mixing Conference. Brugge, Belgium, Bruxelmane, M., Froment, G., Eds.; KVIV: Belgium, 1991; 409–418.
7. Coyne, J.; Kaufman, P.; Post, T.A. Key parameters for consideration in up-pumping technology in fermentation. Proceedings INTERPHEX East, ISPE, Philadelphia, PA, March 19, 1998.
8. Stenstrom, M.K.; Gilbert, R.G. Effects of alpha, beta, theta factor upon the design, specification and operation of aeration systems. Water Res. **1981**, *15*, 643–654.
9. Van't Reit, K. Turbine Agitator Hydrodynamics and Dispersion Performance. Ph.D thesis, Delft University of Technology, The Netherlands, 1975.
10. Kar, K.K.; Cope, R.F. *Prediction of the Oxygen Transfer Rate for the Catalytic Oxidation ... in a Stirred, Batch Reactor*, Internal Dow Chemical Report ERPD 99-5, 1999.
11. Kaufman, P.; Post, T.A.; Preston, M. Up-pumping mixing technology: breaking the mold of traditional systems. Chem. Process. **1998**, 63–66.
12. Bakker, A.; Smith, J.M.; Myers, K.J. How to disperse gases in liquids. Chem. Eng. **1994**, *101*, 98–104.
13. Saravanan, K.; Mundale, V.D.; Joshi, J.B. Gas inducing type of agitated contactors. Ind. Eng. Chem. Res. **1994**, *33*, 2226–2241.

Gas–Liquid Mixing in Agitated Reactors

John C. Middleton
BHR Group Ltd., Cranfield, Bedfordshire, U.K.

John M. Smith
School of Engineering, University of Surrey, Guildford, Surrey, U.K.

Piero M. Armenante
Otto H. York Department of Chemical Engineering, New Jersey Institute of Technology, Newark, New Jersey, U.S.A.

INTRODUCTION

Gas–liquid contacting is important in many reactive processes. In gas–liquid operations, a gas must be effectively and efficiently contacted with the liquid to provide mass transfer. Different contexts bring different challenges. Fermentations and effluent treatment can be at a very large scale, but the product value and workup tend to be comparatively low, so mixer capital and energy are important, whereas mass transfer requirements can be modest (which is fortunate, if the microorganisms are shear-sensitive). Gas–liquid reactions in low-viscosity liquids are often conducted at large scale, have reaction selectivity issues involving the dissolved gas concentration, and have rapid reactions with large exotherms. So for these, scale-up, liquid mixedness, and mass and heat transfer are important while impeller capital and energy cost are not.

Absorption of a gas into a liquid to produce a chemical reaction is often a particularly critical duty. Chlorinations and sulfonations tend to be fast reactions with soluble gases, so high mass transfer intensity with short contact time is efficient. With oxidations, the gas is less soluble, but selectivity is often critical. Hydrogenations involve longer contact times, often with gas recycling (where compression safety can become an issue) and solid particles to be kept in suspension. [1]

Many equipment possibilities exist for gas–liquid operations. They are outlined in Table 1 with their main operational characteristics (at least in air–water systems) presented in Table 2. However, the remainder of this section will deal exclusively with agitated vessels containing low-viscosity liquids in which turbulent flow is achieved ($Re = \rho_L N D^2/\mu > \sim 10^4$).

SELECTION AND CONFIGURATION OF GAS–LIQUID EQUIPMENT

In stirred vessels, the gas should preferably be fed beneath the impeller such that the impeller will "capture" the rising gas plume. [Note that in some cases the gas is drawn in from the headspace using a special self-inducing impeller system (see Ref.[1])]. With radial or upward flow impellers it is sufficient to use a sparger ring with a diameter smaller than the impeller itself ($D_{sparger} = \sim 0.75D$ is recommended). This ensures that the gas can be dispersed into fine bubbles by the impeller, thus providing a high gas–liquid contact area (Figs. 1 and 2). To provide maximum gas contact time the impeller should be near the base of the vessel, but not so near as to inhibit its liquid pumping action: a clearance of $T/4$ is recommended. The bubble breakup mechanism relies upon a high relative velocity between the impeller blades and the liquid, so wall baffles are necessary to restrict the circumferential motion of the liquid. Baffles also enhance the vertical motion of the liquid and hence the mixing of the liquid bulk and the recirculation of liquid and gas back to the impeller, increasing the gas holdup. For any single impeller this recirculation is favored by an aspect ratio (liquid height/vessel diameter $= H/T$) of about one. A vessel of larger aspect ratio (with multiple impellers) may be required to provide a longer contact time for the gas.

An impeller that produces power dissipation under gassed conditions as close as possible to the ungassed power level will provide maximum stability and minimal scale-up difficulties. In the past, flat-blade disk turbines (Rushton turbines) (Fig. 3) have been extensively used in gas–liquid operations. However, their power dissipation is drastically affected by the gas, and they are no longer considered optimal for gas–liquid applications. Recommended impeller types (Fig. 3) include, for radial flow, "hollow-blade" (or "concave-blade") disk turbine designs such as the Scaba SRGT, Chemineer CD6 or BT6, and Lightnin R-130, and, for axial flow, upward-pumping wide-blade hydrofoils such as the Lightnin A345 or A340, and the Prochem-Chemineer Maxflo. Down-pumping hydrofoils or pitched-blade turbines may be unstable during gas–liquid operation because the liquid flow induced by such impellers is opposed to the natural tendency of buoyant gas to rise.[2–4]

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120041245

Table 1 General classification of gas–liquid reactors

Contactors in which the liquid flows as a thin film
Packed columns
Trickle bed reactors
Thin film reactors
Rotating disk reactors
Contactors in which gas is dispersed into the liquid phase
Plate columns (including control cycle reactors)
Mechanically agitated reactors (principally stirred tanks)
Bubble columns
Packed bubble columns
Sectionalized bubble columns
Two-phase horizontal contactors
Cocurrent pipeline reactors
Coiled reactors
Plunging jet reactors, ejectors
Vortex reactors
Contactors in which liquid is dispersed in the gas phase
Spray columns
Venturi scrubbers

(From Ref.[1].)

In vessels where $H > 1.2T$ multiple impellers may be required. These improve liquid mixing, and help redisperse and redistribute the gas. Generally, spacing between impellers should be larger than their diameter D. Multiple radial impellers tend to generate zoned or compartmentalized flow fields in contrast with the better top to bottom circulation generated by multiple axial flow configurations. A combination of a lower radial flow impeller to produce dispersion together with one or more axial flow impellers placed above the radial impeller is often recommended (Fig. 2). Many operators use upward-pumping wide-blade hydrofoils ($D/T = \sim 0.6$) even though there is a tendency for these to develop regions of very high gas fraction in the upper part of the vessel.[5]

TURBULENT MECHANISMS OF GAS DISPERSION

The processes of liquid mixing, generation of interfacial area, and gas–liquid mass transfer in turbulent systems are primarily controlled by the power dissipated in the fluids and the retained gas volume fraction (or gas holdup), ϕ. The power and the fluid properties influence the bubble size. The gas is broken up into a dispersion of bubbles in a so-called ''high-shear'' zone, such as at the impeller-blade tips in an agitated vessel. The power dissipated in this zone controls the bubble breakup process. However, with agitated vessels the design correlations are commonly based on the average energy dissipation per unit mass in the vessel, ε_{avg}. The power in this term is the sum of the shaft power and the, mainly potential, energy introduced as a result of injecting the gas at depth.[6] The ratio of local to average energy dissipation rates can be large and will differ between impeller types.

The bubbles may or may not subsequently recoalesce to some extent, depending on the local fluid dynamics and the interfacial behavior. The unpredictability of this phenomenon rules out a priori prediction of bubble size and interfacial area in general, so design via scale-up from experiments is preferred.

GAS FLOW PATTERNS AND OPERATING REGIMES

The gas flow pattern depends on the regime of gas–impeller interaction. For 6-blade disk turbines and similar impellers, three regimes of flow in the vessel can be defined, as shown in Fig. 4.

- *Flooding*, in which the impeller is overwhelmed by the gas. Gas–liquid contact, mixing and mixing-controlled phenomena (e.g., mass transfer) are very poor.
- *Loading*, in which the impeller disperses the gas throughout the upper portion of the vessel, and the bubbles reach the tank walls but do not recirculate below the impeller.
- *Complete gas dispersion*, in which gas bubbles are distributed throughout the vessel, and a significant amount of gas can be recirculated back to the impeller.

Which regime exists will depend primarily on the impeller agitation speed and the magnitude of the gas flow rate (Fig. 4). These regimes are closely related to the type of gas–impeller interaction: as more gas is fed to the impeller (or the impeller speed is decreased), the impeller becomes increasingly unable to disperse the gas effectively, and there is a greater tendency for the gas to accumulate in the low-pressure regions behind the blades, forming ventilated ''cavities'' (Fig. 5). When such cavities are large they cause a significant reduction in the power number of the impeller (related to their obstruction of the liquid discharge from the impeller; Fig. 6), and hence in its performance for mixing, mass, and heat transfer. This is particularly important for flat-blade disk turbines with four, six, or eight blades.

The transitions between the various regimes generated by a gassed disk turbine can be quantified using dimensionless numbers such as the gas flow number, $Fl_G(=Q_G/ND^3)$, the impeller Froude number, Fr $(=N^2D/g)$, and system geometry ratios, such as the impeller diameter to tank diameter ratio (D/T).[7] For 6-blade disk turbines the cavity regime as well as other regimes are best obtained from the flow

Table 2 Characteristics of gas–liquid contacting equipment

Type of absorber	Typical gas velocity ($\times 10^2$ m/sec)	Residence time distribution		Residence time of liquid	Fractional liquid holdup	$k_L \times 10^4$ m/sec	a_v (m^2/m^3)	$k_L a_v \times 10^2$ sec^{-1}
Film type								
Packed column and trickle bed reactors	10–100	Plug	Plug	Very low	0.05–0.1	0.3–2	2–35	0.06–7
With gas dispersed as bubbles in liquids								
Bubble columns	1–30	Plug	Mixed	Unlimited	0.6–0.8	1–4	2.5–100	0.25–40
Packed bubble columns	1–20	Plug	Mixed	Unlimited	0.5–0.7	1–4	10–30	1–12
Bubble cap plate columns	50–200	Plug	Mixed	Unlimited	0.7–0/7	1–4	10–40	1–16
Plate columns without downcomers	50–300	Plug	Mixed	Limited variation	0.5–0.5	1–4	10–20	1–8
Mechanically agitated contactors	0.1–2	Mixed	Mixed	Unlimited	0.5–0.8	1–5	20–100	2–50
Horizontal pipeline contactors	5–300	Plug	Plug	Low	0.1–0.8	2–6	10–40	2–24
Static mixers	0.05–20	Plug	Plug	Low	0.01–0.99	1–20	10–100	10–200
With liquid dispersed in gas								
Spray columns	5–300	Mixed	Plug	Very low		0.5–1.5	2–15	0.1–2.25
Sieve plate in spray regime	100–300	Plug	Mixed	Unlimited	—	1–3	5–20	0.5–6

Note: These are comparative values only, based on air and water.
(From Ref.[1].)

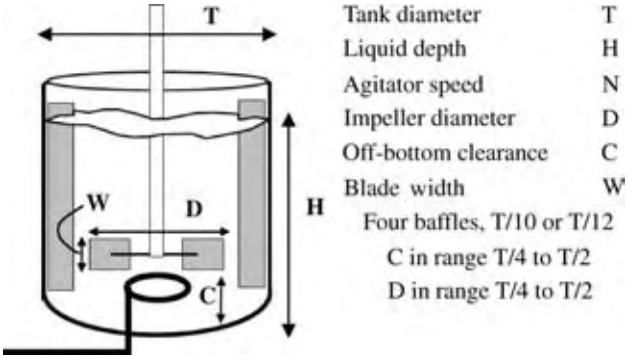


Fig. 1 Standard vessel geometry (single impeller, $H \approx T$). (From Ref.[1].)

regime maps of Warmoeskerken and Smith (Fig. 7) (also summarized by Middleton) valid only for a disk turbine with $D = 0.4T$.[8,9] For other D/T ratios, the regime boundaries should be adjusted using the appropriate correlations given here for flat-blade disk turbines and concave-blade disk turbines:

- Below a certain minimum agitation speed, the impeller has no discernible action on gas dispersion, irrespective of any other factor, including blade type. This occurs approximately when:

$$Fr < 0.04 \tag{1}$$

  which is represented by a horizontal line in the Fr vs. Fl_G. flow map.
- As the impeller speed is increased the impeller starts interacting with the gas. However, when the impeller speed is too low (at a given gas flow rate) the impeller is not capable of dispersing the gas effectively, and flooding will result. For a flat-blade disk turbine flooding will occur when:

$$Fl_G > 30Fr\left(\frac{D}{T}\right)^{3.5} \tag{2}$$

  For a concave-blade disk turbine, the multiplying constant in this equation is 70 instead of 30.[10] The values of this constant clearly show that the concave-blade disk turbine can handle more than twice as much gas flow rate as the flat-blade disk turbine before becoming flooded. Eq. (2) also shows that, for a given D/T ratio, a linear dependence exists between the Froude number and the gas flow number at flooding. Eq. (2) creates two regions separated by a straight line with a slope of 1 in the logarithmic Fr vs. Fl_G flow map. From Eq. (2), it also must be that $Q_G \propto N^3$ at flooding, for a given system.
- Bakker, Smith, and Myers combined the results of Nienow and those of Warmoeskerken and Smith to establish an equation for the determination of the transition point at which the gas becomes completely dispersed in the liquid (in the regions both below and above the impeller). For flat-blade disk turbines, the expression that they found is:[10–12]

$$Fl_G = 0.2Fr^{0.5}\left(\frac{D}{T}\right)^{0.5} \tag{3}$$

  The multiplying constant for concave-blade disk turbines is 0.4 instead of 0.2. Again, these numbers show that concave-blade turbines completely disperse twice as much a gas flow rate at the same agitation speed. This equation delimits two regions separated by a straight line with a slope of 2 in the logarithmic Fr vs. Fl_G flow map.
- Large cavities are developed by a flat-blade disk turbine when:

$$Fl_G > \sim 0.025\left(\frac{D}{T}\right)^{-0.5} \tag{4}$$

  The multiplying constant for concave-blade disk turbines is 0.058 instead of 0.025. This constant has a weak dependence (to the power of about 0.2) on the scale of the equipment. This inequality defines two regions delimited by a vertical line in the Fr vs. Fl_G flow map, for a given value of the D/T ratio.
- Nienow, Wisdom, and Middleton developed a relationship to predict the agitation speed of a flat-blade disk turbine at which gas recirculation occurs for a given gas rate.[13] This expression

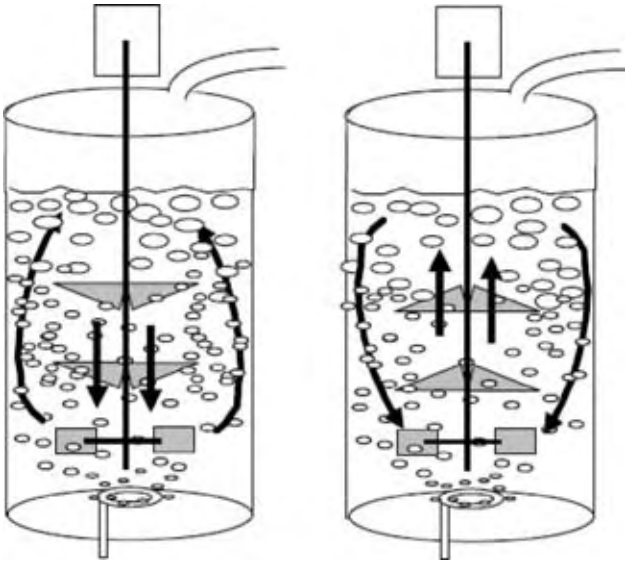


Fig. 2 Multiple impeller agitators: down- and up-pumping hydrofoils above a radial dispersing impeller. (From Ref.[1].)

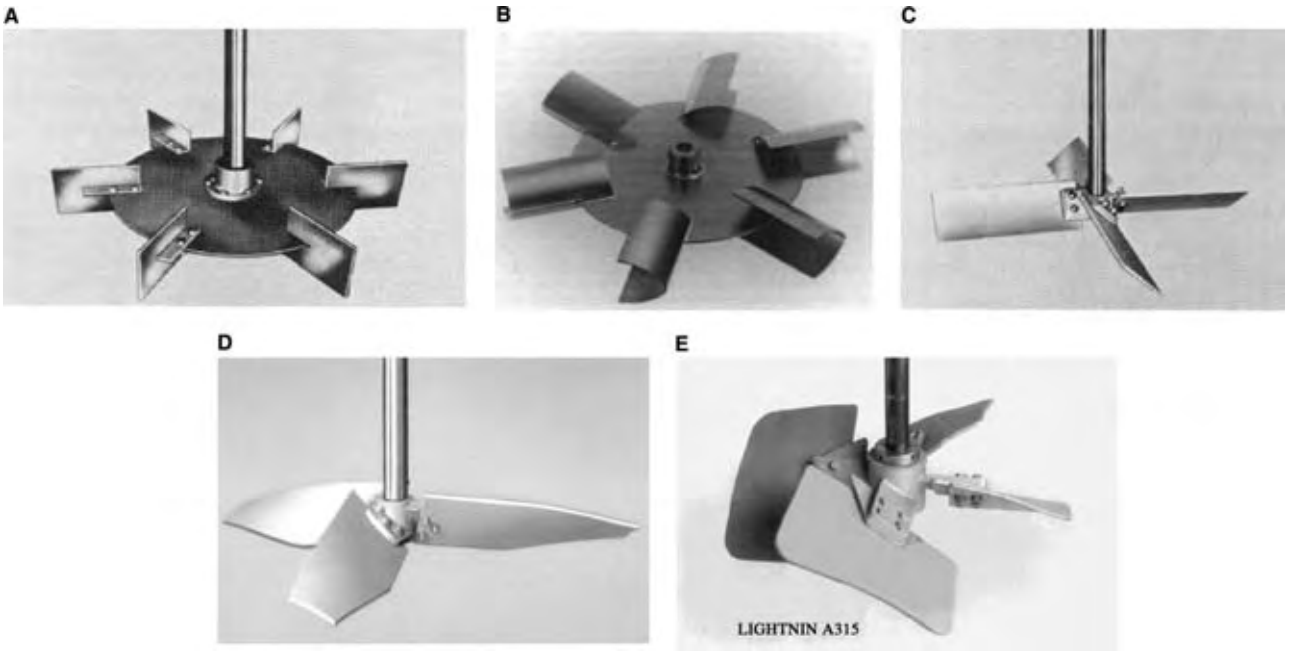


Fig. 3 Various impellers: (A) Rushton disk turbine, (B) hollow-blade turbine, (C) pitched-blade turbine, (D) narrow-blade hydrofoil (Lightnin A310), and (E) wide-blade hydrofoil (Lightnin A315). (From Ref.[1].)

can be reformulated and expressed as:

$$Fl_G < 13Fr^2\left(\frac{D}{T}\right)^{5.0} \quad (5)$$

which defines two regions delimited by a straight line of slope 0.5 in the logarithmic Fr vs. Fl_G flow map.

The equations above allow one to predict the operating conditions in any coalescing gas–liquid system equipped with a flat-blade disk turbine. A similar analysis and flow map have been produced for a concave-blade impeller such as the CD6.[12]

With axial-flow impellers in down-pumping mode, two important regimes are identified: ''direct'' and ''indirect'' loading (Fig. 8).[14] At lower gas flow rates and higher impeller speeds, the downflow from the impeller dominates and gas enters the impeller from above; this is known as indirect loading. If the gas buoyancy dominates, the gas loads the impeller ''directly,'' and the impeller now pumps radially with a much diminished power number (Fig. 9). Operation near the transition is to be avoided because the regime can flip unstably, giving rise to serious mechanical and operational problems. It is preferable to avoid this possibility altogether by operating in an upward-pumping

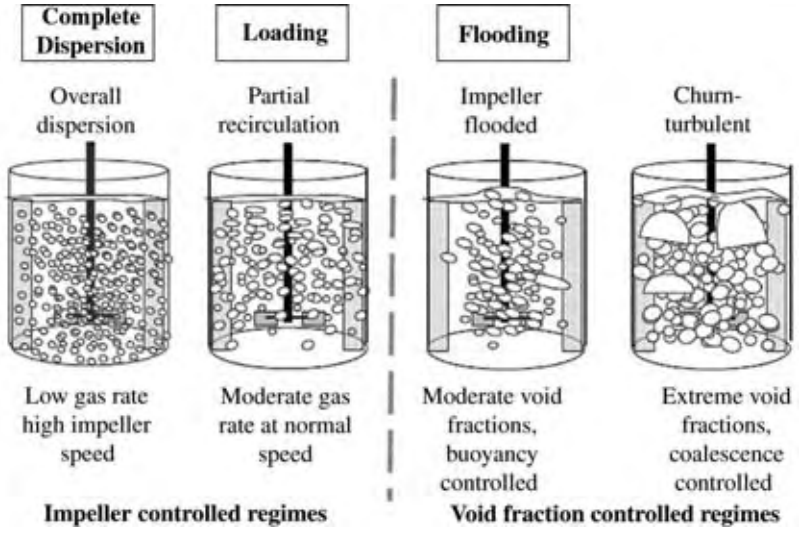


Fig. 4 Typical void distributions in vessels with a single impeller. (From Ref.[1].) *(View this art in color at www.dekker.com.)*

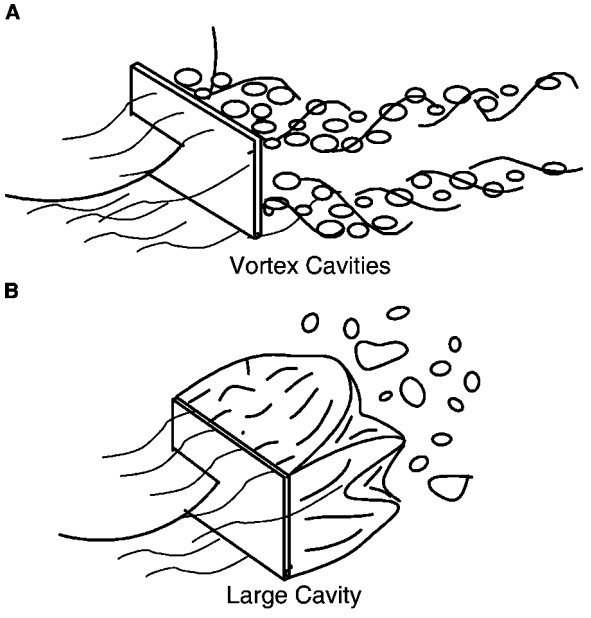


Fig. 5 Ventilated gas cavity forms on turbine blades. (From Ref.[1].)

mode. Here, the gas and liquid flows are not in conflict, and the power curve with gassing is stable.

POWER DISSIPATION IN GAS–LIQUID STIRRED REACTORS

Under turbulent flow conditions, power dissipation is the controlling factor for mass transfer and phase dispersion. The power drawn by a single impeller in a liquid–gas system is typically lower than that drawn by the same impeller in liquid alone. The presence of the gas reduces the average density of the mixture, and the gas flow regime (e.g., flooding) may cause the impeller blades to be locally surrounded by a higher fraction of low-density gas than other locations in the vessel. The power dissipated by an impeller under gassed conditions can be calculated from:

$$P_G = K_{G/U}P_U = K_{G/U}N_p\rho_L N^3 D^5 \quad (6)$$

where P_U (equal to P in the liquid alone) is the power drawn at the same impeller speed under ungassed conditions, which is given by:

$$P_U = N_p\rho_L N^3 D^5 \quad (7)$$

In Eq. (6) $K_{G/U}$ ($=P_G/P_U$) is the "gassing factor" or relative power demand (RPD) between (otherwise identical) gassed and ungassed systems. $K_{G/U}$ depends on the impeller type, Q_G, N, and D, and generally decreases with increasing gas flow number, Fl_G. A summary of typical $K_{G/U}$ values for popular impellers at high gas flow rates ($Fl_G = \sim 0.1$) is given in Table 3.

The value of $K_{G/U}$ in Eq. (6) can easily fall as low as 0.4 for 6-blade flat-blade disk turbines (Fig. 6) and downflow pitched-blade turbines and hydrofoils (Table 3). However, for modern impellers with parabolic concave blades (e.g., Scaba SRGT, Chemineer BT6, Lightnin R-130), $K_{G/U}$ typically falls to only about 0.9 (and then only at high Fl_G values). The ability to deliver high power makes these impellers highly suitable for gas–liquid operations.

If higher power numbers are required, flat-blade turbines with more than six blades (preferably 12 or 16) can be used, for which $K_{G/U}$ eventually drops to about 0.4, but not until much higher flow numbers than for six flat blades. $K_{G/U}$ for up-pumping wide-blade hydrofoils remains close to 1.0. Axial-flow impellers (e.g., Lightnin A345, Prochem Maxflo W, APV B6), if used for gas–liquid duty, should be preferentially operated in the upflow direction, when they are stable and maintain more than 70% of their ungassed power draw

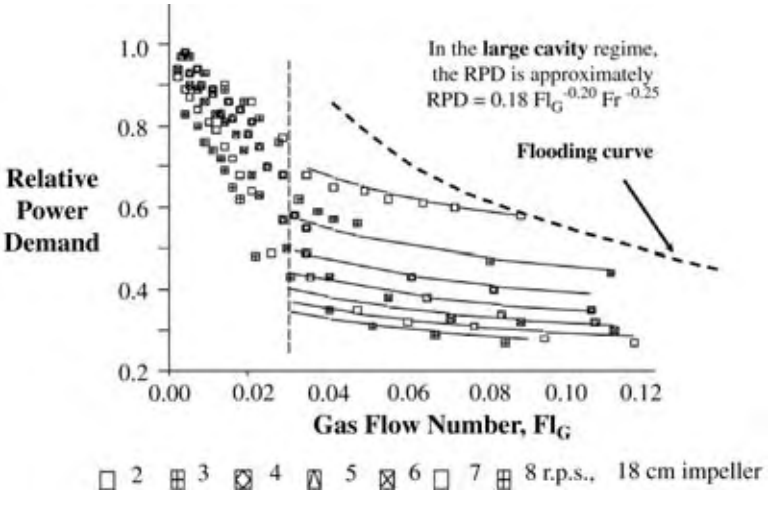


Fig. 6 Relative power demand for a gassed Rushton turbine ($D/T = 0.4$). (Data from Refs.[1,20].)

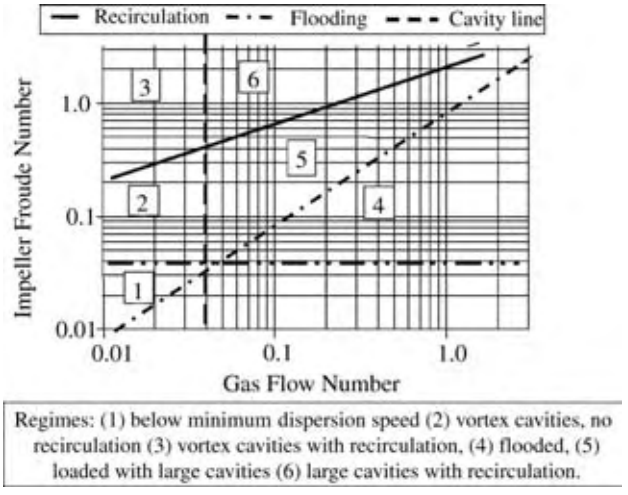


Fig. 7 Flow map for single Rushton turbine (T/D = 2.5). (From Ref.[1].)

on gassing. Down-pumping axial impellers, especially pitched-blade turbines and narrow-blade hydrofoils, have a seriously unstable operating regime in gassed systems, and suffer a drastically sharp fall in $K_{G/U}$ under particular conditions (the direct–indirect loading transition) with dire consequences such as fluctuating process performance, rapid seal and bearing wear, and high risk of shaft failure. However, wide-blade hydrofoils can be quite effective, especially as the upper impellers in multiple impeller agitators. High gas–liquid mass transfer requires high P_G values. Hence, impellers with a combination of high N_p and $K_{G/U}$ are best suited for gas–liquid duties.

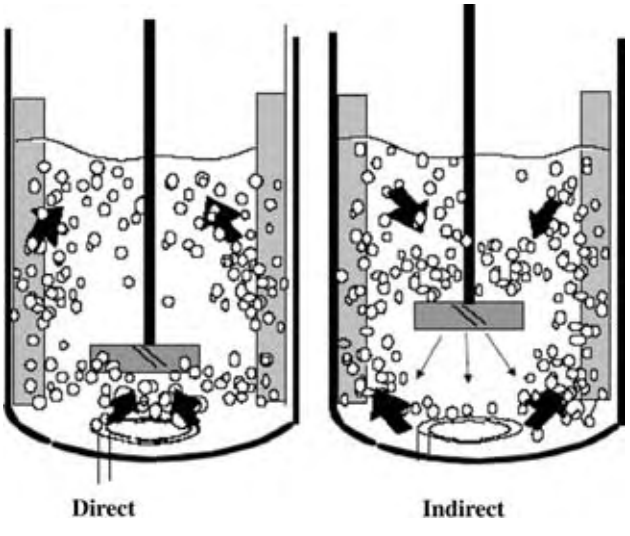


Fig. 8 Direct and indirect loading of a downward pumping axial flow impeller. (From Ref.[1].)

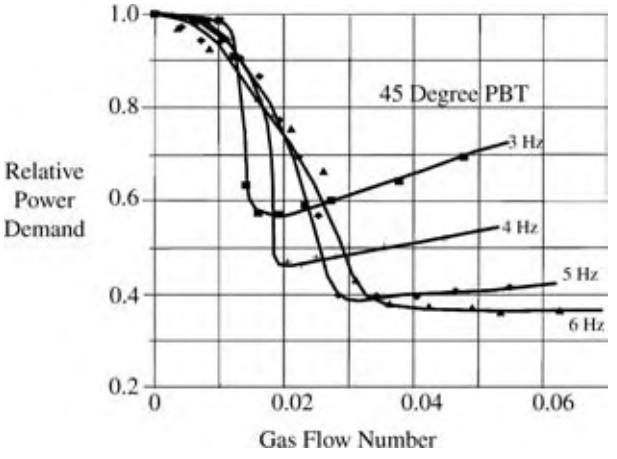


Fig. 9 Relative power demand for a down-pumping 45° pitched-blade turbine. (From Refs.[1,14].)

Correlations have also been developed for P_G. Bakker, Smith, and Myers report the following equation for a single gassed impeller:[10]

$$\frac{P_G}{P_U} = [1 - (b - a\mu)Fr^d]\tanh(cFl_G) \quad (8)$$

where the values of the constants a, b, c, and d are given in Table 4 for two different types of impellers, i.e., flat-blade disk turbine and concave-blade disk turbine. The use of this equation clearly shows that for the same Fl_G value the P_G/P_U ratio is much higher for the concave-blade disk turbine than for the flat-blade disk turbine. This confirms that the concave-blade design is superior to the flat-blade disk turbine in being able to deliver high power at high gas flow. However, it should also be remarked that the power number under ungassed conditions is quite different for the two impellers (5.0 for the flat-blade disk turbine vs. 3.2 for the concave-blade design). Nevertheless, the concave-blade turbine typically delivers more power under most gassed conditions than a straight-blade impeller such as the flat-blade disk turbine.

In the large-cavity regime of gassed aqueous systems, a good approximation for the gassed power of a single flat-blade disk turbine ($D/T = 0.4$) is given by the following expression:[1]

$$\frac{P_G}{P_U} = 0.18Fl_G^{-0.20}Fr^{-0.25} \quad (9)$$

Multiple impellers are often used in gas–liquid operations. Assuming that the lowest impeller is used for the primary gas dispersion, the upper ones are not loaded by all the gas entering through the sparger.[7,10] For the purpose of power demand estimation, it can be assumed that upper impellers experience about half the total gas rate. Correlations to estimate

Table 3 $K_{G/U}$ value for various impellers

Impeller type	N_p	$K_{G/U}$ ($Fl_G = 0.1$)
Radial flow		
6-Blade disk turbine[a] $D = T/3$	5	0.4
12-Blade disk turbine $D = T/3$	10	0.6
18-Blade disk turbine $D = T/3$	12	0.7
Chemineer CD6	2.3	0.8
Chemineer BT6	2.0	0.9
Scaba 6-SRGT	1.5	0.9
Axial upflow		
4-Pitched-blade turbine $D = T/3$, C $= T/3$	1.3	0.75
6-Pitched-blade turbine $D = T/3$, C $= T/3$	1.7	0.75
Lightnin A345, $D = 0.4T$	0.8	0.75
Axial downflow		
4-Pitched-blade turbine, $D = T/3$, $C = T/3$	1.3	0.3
6-Pitched-blade turbine, $D = T/3$, $C = T/3$	1.7	0.4
Prochem Maxflo 5, $D = 0.45T$	1.3	0.7
Lightnin A315 $D = 0.4T$	0.8	0.7

[a]This is actually a function of scale.[15]
(From Ref.[1].)

the power dissipated in multiple impeller systems are also available.[1,10,16,17]

GAS HOLDUP

Gas holdup (or retained gas volume fraction), ϕ, is the volumetric fraction of gas in the vessel with respect to the total volume of the gas–liquid mixture, i.e.,

$$\phi = \frac{V_G}{V_L + V_G} \tag{10}$$

In this expression, V_G is the gas volume in the tank under (dynamic) gassed conditions. Experimentally, the gas holdup can be obtained by taking the ratio of the observed incremental height of the gas–liquid mixture (with respect to the height of the liquid when no gas is present) to the total height of the gas–liquid mixture under gassed conditions.

Gas holdup in agitated vessel is difficult to predict a priori, but scale-up can be made reasonably accurately using empirical correlations of the form:

$$\phi = \alpha' \left(\frac{P_G}{\rho V_L}\right)^{\beta'} (v_s)^{\gamma'} \tag{11}$$

where the superficial gas velocity, v_s, is defined as the velocity of the gas in the empty vessel:

$$v_s = \frac{Q_G}{(\pi/4)T^2} \tag{12}$$

The constants α', β', and γ' are independent of scale. The value of α' depends on the physical properties of the liquid in a way that is in general difficult to predict (hence, the recommendation to obtain at least one measurement at the pilot plant scale during process development and use the above correlation for scale-up only). Published data are for aqueous systems, in which the addition of any solute exhibiting surface activity (including electrolytes, alcohols, and surfactants) has a large impact on the gas fraction. For example, in a given system the value of ϕ may be, say, 0.1 if water is used, whereas ϕ may be 0.25 with a solution of a simple electrolyte. For engineering purposes, the situation has been simplified to cover two ''extreme'' classes of liquid system, the so-called ''coalescing'' and ''noncoalescing'' systems, with separate correlations for each for gas fraction and mass transfer.

Values of β' and γ' vary in the literature between 0.2 and 0.7. Generally, $\beta' = 0.48$ and $\gamma' = 0.4$ are quite reliable.[18] More recent work has led to the equation

Table 4 Values of the constants a, b, c, and d in Eq. (8)

Impeller type	a	b	c	d
Rushton turbine	0.72	0.72	24	0.25
Concave-blade turbine (CD6)	0.12	0.44	12	0.37

(From Ref.[10].)

(expressed in W, m, sec units):[19]

$$\phi = 0.9\left(\frac{P_G}{\rho_L V_L}\right)^{0.20}(v_s)^{0.55} \quad (13)$$

for holdup in vessels with multiimpeller agitators dispersing air in water. This equation is for operation at ambient temperature. In the fully turbulent regime, there is a dependence of void fraction on temperature, which results in $\phi \propto \mu^{0.55}$.

INTERFACIAL AREA AND MASS TRANSFER COEFFICIENT IN GAS–LIQUID DISPERSIONS

Because of the difficulty of evaluating independently the specific gas–liquid interfacial area and the mass transfer coefficient, the group $k_L a_v$ is often determined experimentally and correlated to other variables, such as power dissipation under gassed conditions and superficial gas velocity. Correlations for $k_L a_v$ under turbulent conditions are typically expressed in the form:

$$k_L a_v = \alpha''\left(\frac{P}{\rho_L V_L}\right)^{\beta''}(v_s)^{\gamma''} \quad (14)$$

where P includes both the shaft power and the gas buoyancy power ($= QHg[\rho_L - \rho_G]$), but not the gas kinetic energy.[6] Typical values for the air–water system at 20°C are $\alpha'' = 1.2$, $\beta'' = 0.7$, and $\gamma'' = 0.6$, with P (W), V_L (m^3), v_s (m/sec), $k_L a_v$ (sec^{-1}) and α'' dimensioned appropriately.[9] However, it has been found (for example, by Smith, Middleton, and van't Riet) that whereas the indices β'' and γ'' do not change with liquid type, impeller type, or scale, α'' is a strong function of liquid type and properties, with the noncoalescing value of α'' being about twice that for coalescing systems.[18] Thus, such correlations can be used for scale-up purposes, but not for general prediction.

Although it is commonly assumed that when agitation conditions are sufficiently intense for effective gas–liquid dispersion, the liquid mixedness will be good, it is worth checking this, particularly for large vessels. If the mixing time is longer than mass transfer time (90% mass transfer time $= 2.3/k_L a_v$), then the liquid mixing should be improved, or otherwise a more complex design calculation with interlinked zones of different driving force (and even perhaps local values for $k_L a_v$) will be necessary.

The gas–liquid mass transfer rate can be calculated from:

$$\dot{m} = k_L a_v V_L \Delta C \quad (15)$$

where ΔC is the concentration driving force across the phases. This equation shows that knowledge of $k_L a_v$ [through Eq. (14)] is needed to calculate the mass transfer rate. In general, V_L can be taken as the entire volume of the gas–liquid system, provided that the agitation speed is high enough to ensure that the complete dispersed regime has been achieved.

CONCLUSIONS

Mixing is a critical component in the analysis of any gas–liquid reactor system. Mixing significantly affects mass transfer, which, in turn, can have a profound effect on the final yield and selectivity of gas–liquid reactions, especially if the reaction step is faster than the mass transfer step. The complexity of a typical gas–liquid system is such that experimentation should always be incorporated in any development and scale-up work. As in many other multiphase mixing systems, the designer should be aware that the selection of the mixing system and the determination of the operating conditions will have a significant effect on the attainment of adequate gas dispersion and the generation of high interfacial area and a large mass transfer coefficient. The material presented here contains some fundamental knowledge, which, when integrated with practical insight and engineering judgment, can help solve problems associated with industrial gas–liquid applications.

NOMENCLATURE

A	Interfacial area between phases (m^2)
a_v	Interfacial area per unit volume of liquid (m^2/m^3)
B	Baffle width (m)
C	Impeller clearance measured from the impeller centerline to the vessel bottom off the vessel bottom (m)
ΔC	Concentration driving force for mass transfer (mol/m^3)
D	Impeller diameter (m)
$D_{sparger}$	Diameter of sparger ring (m)
g	Gravitational acceleration (m/sec^2)
H	Height of liquid in vessel (m)
$K_{G/U}$ ($=P_G/P_U$)	Gassing factor, i.e., ratio of gassed to ungassed power dissipations
k_L	Mass transfer coefficient (m/sec)
m_L	Mass of liquid (kg)

N	Impeller speed (rotations/sec)
P	Power dissipation (W)
P_G	Power dissipation under gassed conditions (W)
P_U	Power dissipation under ungassed conditions (W)
Q_G	Mean gas volumetric flow rate (m^3/sec)
RPD	Ratio of gassed to ungassed power for impeller (–)
t	Time (sec)
T	Vessel diameter (m)
v_s	Superficial gas velocity, i.e., volumetric gas flow rate/cross-sectional area of vessel (m/sec)
V_G	Volume of gas in vessel (m^3)
V_L	Volume of liquid in vessel (m^3)
W	Width of an impeller blade (m)

Greek Symbols

α', β', γ'	Empirical constants in holdup correlations (–)
α'', β'', γ''	Empirical constants in gas–liquid mass transfer correlations (dimensioned appropriately)
ε	Energy dissipation rate (or power draw) per mass of mixture (W/kg)
ε_{avg}	Average energy dissipation rate (or power draw) per mass of mixture per mass (W/kg)
ϕ	Gas holdup (retained volume fraction of gas phase), i.e., ratio of volume of gas phase to total volume of the gas–liquid mixture (–)
μ	Viscosity (Pa sec)
ν	Kinematic viscosity (m^2/sec)
ρ_L	Density of liquid (kg/m^3)
ρ_G	Density of gas (kg/m^3)
$\bar{\rho}$	Average density of mixture (kg/m^3)
σ	Interfacial tension (N/m)

Dimensionless Groups

Fr	Impeller Froude number, $=N^2D/g$
Fl_G	Gas flow number, $=Q_G/ND^3$
N_p	Power number, $=P/(\rho_L N^3 D^5)$
P_G/P_U	Ratio of gassed to ungassed power for impeller
Re	Impeller Reynolds number, $=\rho_L ND^2/\mu$

REFERENCES

1. Middleton, J.C.; Smith, J.M. Gas-liquid mixing in turbulent systems. In *Handbook of Industrial Mixing: Science and Practice*; Paul, E.L., Atiemo-Obeng, V.A., Kresta, S.M., Eds.; John Wiley & Sons: New York, 2004; 543–584.
2. Chapman, C.M.; Nienow, A.W.; Middleton, J.C.; Cooke, M. Particle–gas–liquid mixing in stirred vessels. Chem. Eng. Res. Des. (Trans. Inst. Chem. Eng. A) **1983**, *61*, 71–82 and 167–182.
3. Nienow, A.W.; Konno, M.; Bujalski, W. Studies on three phase mixing: a review and recent results. Chem. Eng. Res. Des. (Trans. Inst. Chem. Eng. A) **1986**, *64*, 35–42.
4. Hari-Prajitno, H.; Mishra, V.I.; Takemaka, K.; Bujalski, W.; Nienow, A.W.; McKemmie, J. Gas–liquid mixing studies with multiple up and down pumping hydrofoil impellers: power characteristics and mixing times. Can. J. Chem. Eng. **1998**, *76*, 1056–1068.
5. Smith, J.M.; Gao, Z.; Müller Steinhagen, H. Void fraction distributions in sparged and boiling reactors with modern impeller configurations. Chem. Eng. Process. **2001**, *40*, 489–497.
6. Middleton, J.C.; Cooke, M.; Litherland, L. The role of kinetic energy in gas–liquid dispersion. In *Institute of Chemical Engineering Symposium Series*, No. 136, Proceedings of the 8th European Mixing Conference, Cambridge, 1994; 595–602.
7. Smith, J.M.; Warmoeskerken, M.M.C.G.; Zeef, E. Flow conditions in vessels dispersing gases in liquids with multiple impellers. In *Biotechnology Processes*; Ho, C.S., Oldshue, J.Y., Eds.; AIChE 1987, Vol. 11, 107–115.
8. Warmoeskerken, M.M.C.G.; Smith, J.M Flow regime maps for rushton turbines. In 3rd World Congress of Chemical Engineering, Tokyo, 1986; Paper W-624.9.
9. Middleton, J.C. Gas-liquid dispersion and mixing. In *Mixing in the Process Industries*; Harnby, N., Nienow, A.W., Edwards, M.F., Eds.; Butterworth–Heinemann: Oxford, 1997.
10. Bakker, A.; Smith, J.M.; Myers, K.J. How to disperse gases in liquids. Chem. Eng. **1994**, *101* (12), 98–104.
11. Nienow, A.W. Gas dispersion performance in fermentor operation. Chem. Eng. Progr. **1990**, 61–71.
12. Warmoeskerken, M.M.C.G.; Smith, J.M. Hollow blade turbine impellers for gas dispersion and mass transfer. Chem. Eng. Res. Des. (Trans. Inst. Chem. Eng. A) **1989**, *67A*, 193–198.
13. Nienow, A.W.; Wisdom, D.J.; Middleton, J.C. The effect of scale and geometry on flooding, recirculation and power in gassed stirred vessels.

Proceedings of the 2nd European Mixing Conference, Cambridge; BHRA: Cranfield, U.K., 1977; 17–34.
14. Warmoeskerken, M.M.C.G.; Speur, J.; Smith, J.M. Gas–liquid dispersion with pitched blade turbines. Chem. Eng. Commun. **1984**, *25*, 11–29.
15. Bujalski, W.A.; Nienow, A.W.; Chatwin, S.; Cooke, M. Power consumption with aerated rushton turbines. Chem. Eng. Sci. **1987**, *42*, 317–326.
16. Hudcova, V.; Machon, V.; Nienow, A.W. Gas-liquid dispersion with dual rushton turbine impellers. Biotechnol. Bioeng. **1989**, *34*, 617–628.
17. Tatterson, G.B. *Fluid Mixing and Gas Dispersion in Agitated Tanks*; McGraw-Hill: New York, 1991.
18. Smith, J.M.; Middleton; van't Riet, K. Scale up of agitated gas–liquid reactors for mass transfer. Proceedings of the 2nd European Conference Mixing; BHRA: Cambridge, 1977; F4-51–F4-66.
19. Gao, Z.; Smith, J.M.; Muller-Steinhagen, H. The effect of temperature on the void fraction in gas–liquid reactors. Proceedings of the 5th Symposium on Gas–Liquid–Solid Systems, Melbourne, 2001.
20. Warmoeskerken, M.M.C.G.; Smith, J.M. Description of the power curves of turbine stirred gas dispersions. Proceedings of 4th European conference on Mixing; Noordwijkerhout: The Netherlands, 1982; 237–246.

Gas-Phase Lubrication of MEMS Devices: Using Alcohol Vapor Adsorption Isotherm for Lubrication of Silicon Oxides

Kenneth Strawhecker
David B. Asay
Department of Chemical Engineering, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.

Seong H. Kim
Department of Chemical Engineering, Materials Research Institute, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.

INTRODUCTION

The gas adsorption isotherm can be used as an effective way of continuously delivering and replenishing lubricant molecules on solid surfaces when lubrication boundary layers approach molecular length scales. As the gas has essentially negligible viscosity, this approach is favorably applicable to micromechanical systems where power dissipation by the lubricant media is a serious problem. Gas adsorption is fully conformal to all regions of the device when the mean free path of the adsorbate is less than the distance between lubrication faces. This approach is suitable for the lubrication of both buried and exposed surfaces that exist in micro-electromechanical system (MEMS) devices. This entry describes the lubrication effects of alcohol adsorption for native silicon oxide surfaces. The need and advantage of gas-phase lubrication and the molecular basis for alcohols being a good gas-phase lubricant are explained in the background section. The use of atomic force microscopy (AFM) and quartz crystal microbalance (QCM) for gas-phase lubrication measurements are described in the experimental section. The adhesion and the friction of silicon oxide surfaces measured in ambient air and alcohol vapor containing environments are then compared. The alcohol molecules such as ethanol, *n*-propanol, *n*-butanol, and *n*-pentanol adsorb on the substrate surface from the gas phase and form a lubricating film. The resulting film decreases both the adhesion and friction forces between two silicon oxide surfaces. The possible lubrication mechanism and its implication to micromechanical devices are discussed.

BACKGROUND

One of the issues in the design of micromachines is the lubrication of device surfaces and the elimination of wear and stiction between moving parts.[1] As critical dimensions decrease, adhesion and friction become larger problems relating to the reliable operation of these devices.[2] Current approaches to solve the friction and wear problems in MEMS utilize mainly solid-phase lubrication films, which can be deposited on the device surface during the device fabrication.[1,2] Although these approaches have made significant improvements in yielding conformal coating and reducing friction as well as wear in laboratory tests, they have technical limitations relating to the durability of what is essentially a one-time loaded coating. Solid lubricant coatings and organic self-assembled monolayers have been shown to wear off during MEMS operation.[3,4] What is absent in these thin film lubrication systems is the capability for continuous replenishment of the lubricating layer. In fact, the continuous replenishment of lubricating layers has been the key feature of the conventional lubrication approaches typically using viscous liquids. Unfortunately, the viscous nature of the liquid lubricant causes power dissipation, making their usage in MEMS devices undesirable and impractical.

The surface of all inorganic materials exposed to ambient (humid) air is always covered with a thin layer of water adsorbed from the gas phase.[5] The thickness of the adsorbed water layer varies with the humidity and surface chemistry.[6] This water layer has been shown to reduce wear in MEMS operation.[7] However, the high surface tension of the water film can cause an in-use stiction problem. The gas-phase lubrication concept discussed here employs the same equilibrium adsorption principle as the water adsorption in humid environments. The difference is that our approach utilizes a ''surfactant-like'' molecule that can provide low adhesion and good lubrication. The entry summarizes the advantages of gas-phase lubrication for MEMS devices and discusses the effect of alcohol adsorption on the adhesion and lubrication of silicon oxide surfaces.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120041047

ALCOHOL AS A GAS-PHASE LUBRICANT FOR MEMS DEVICES

Alcohols are chosen as a gas-phase lubricant candidate because they have high enough vapor pressure for gas-phase mass transport to the buried surfaces at room temperatures, are environment friendly, and can readily dissolve into the water film present on inorganic solid surfaces in ambient conditions. The vapor pressures of different *n*-alcohols are shown in Fig. 1. The solubility of alcohols in water is shown in Fig. 2. The diffusivity of alcohols in the air is quite high, and their mean free path at room temperature and ambient pressure is much less than device part dimensions. This ensures the lubrication of sidewalls and buried surfaces. Simple alcohols can be used for lubrication of MEMS devices that operate at as low as –100°C and as high as 150°C depending on the number of carbons in the alkyl group (Table 1). The thickness of the adsorbed alcohol layer can be varied by adjusting the alcohol vapor pressure and the substrate temperature.[8,9] In bulk liquid lubrication studies, alcohols have shown good lubrication and antiwear properties for silicon oxide and silicon nitride surfaces in the liquid phase.[10] As in the case of water molecules in the humid air, the presence of the gas-phase alcohol molecules and a few layers of alcohol molecules at the surface will not interfere with optical reflection or electrical contact of devices.

One of the significant problems in MEMS operation under humid conditions is the high stiction forces owing to adsorbed water molecules on the device surfaces. The thickness of the condensed water layer varies with the relative humidity and the surface chemistry (hydrophilicity, hydrophobility, organic contamination, etc.). As water has a very high surface tension (γ_{water} = 72 dyn/cm), the water meniscus formed between two

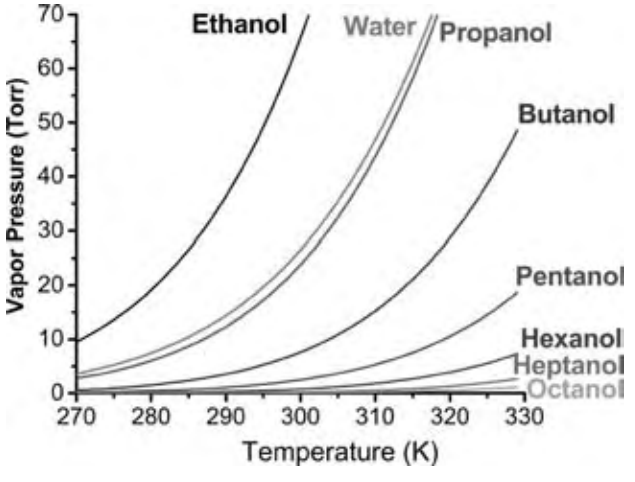


Fig. 1 Vapor pressure of water and 1-alcohols calculated from Antoine equation. *(View this art in color at www.dekker.com.)*

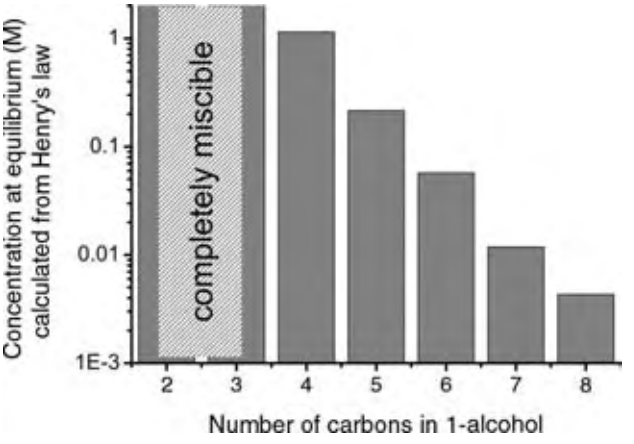


Fig. 2 Solubility of alcohols in water calculated from Henry's law. *(View this art in color at www.dekker.com.)*

surfaces separated by a distance d and having a contact angle θ generates a large Laplace pressure, P_L:

$$P_L = (2\gamma_{water}\cos\theta)/d$$

On a high surface energy substrate, the water contact angle is small and the capillary forces thus generated are sufficiently strong to collapse a typical suspended microstructure to the substrate.[11]

This problem can be mitigated by adding alcohol molecules at the interface. The mass accommodation coefficient for alcohol adsorption at the vapor/water interface varies in the range of 0.1–0.98, depending on the types of alcohols, temperature, and partial pressure.[12,13] Taking into account the high collision frequency of vapor-phase molecules at ambient pressure (in the order of 10^7–10^9collisions/site/sec), the measured accommodation coefficients of alcohol molecules indicate that the dissolution of gas-phase alcohol into water takes place very fast.

The surface tension of aqueous organic solution is lower than that of water.[14] The Gibbs isotherm predicts surface excess of organic molecules at the gas/liquid interface, leading to a nonlinear decrease of the surface tension with respect to the bulk mole fraction of organic molecules in the solution.[15,16] The same effect is expected to occur in the condensed liquid layer of water and alcohol on the substrate surface. The saturated surface concentration of small alcohols reaches $\sim 10^{14}$ molecules/cm^2, resulting in an almost complete coverage of the gas/liquid interface with organic molecules.[17,18] The amphiphilic nature of the molecular structure induces orientational ordering at the gas/liquid interface such that hydrophobic hydrocarbon groups are pointing toward the gas phase and the hydroxyl groups being fully immersed in water.[19,20] This surface ordering is driven by an

Table 1 Melting and boiling points of linear *n*-alcohols

Alcohols	Melting point (°C)	Boiling point (°C)
1-Octanol	−15	196
1-Heptanol	−36	176
1-Hexanol	−52	156
1-Pentanol	−78	137
1-Butanol	−90	118
1-Propanol	−127	97
Ethanol	−117	78

excellent match between water and alcohol OH groups. This reduces the large stiction forces, achieving a film with a lower surface tension because of the hydrocarbon groups of the alcohol molecules, which results in a significant reduction of the free energy.[21,22]

EXPERIMENTAL

Fig. 3 is a basic schematic of the apparatus used for adhesion and friction measurements as a function of the alcohol vapor pressure. Linear *n*-alcohols were used as received, including absolute ethanol (C_2H_5OH, 99.5%), 1-propanol (C_3H_7OH, 99.95%), 1-butanol (C_4H_9OH, 99.9%), and 1-pentanol ($C_5H_{11}OH$, 99+%). The gas composition is controlled by adjusting the ratio of dry argon flowrate to the saturated alcohol/argon flowrate. The mixture of argon and saturated alcohol vapor is produced by flowing dry argon gas through a heated bubbler containing alcohol and a condenser held at the experimental temperature. The purpose of the condenser is to remove any supersaturated alcohol vapor from the argon gas flow, leaving only argon with saturated alcohol vapor at the experimental temperature. The fully saturated vapor experiment was done in a small environment chamber (not shown in Fig. 3) that contained a source of liquid alcohol and concealed the sample compartment of the AFM. The AFM experiments were performed using a Molecular Imaging PicoSPM head and an RHK Technology controller. All measurements were made at room temperature. The substrate was cleaned with a piranha solution followed by UV-ozone cleaning and heating to 300°C. Silicon tips (force constant = 2.2 and 4.5 N/m) covered with a native oxide layer were cleaned with the UV-ozone treatment. The nominal tip diameter characterized by the manufacturer was 20 nm; however, the tip sharpness was readily reduced in our experiment conditions (applied normal force up to 10–100 nN). Adhesion force was determined from the difference between the pull-off point and the free standing location on the force–distance (*f*–*d*) curve. Friction force was determined from the lateral force hysteresis in the left and right scans in contact-mode operation. To minimize experimental artifacts because of the uncertainty in tip diameter and AFM tip aligning, the measured adhesion and friction forces were normalized to the initial adhesion and friction force values measured with dry argon flowing previous to each experiment. A QCM was used to measure the approximate thickness of the alcohol on a gold substrate.

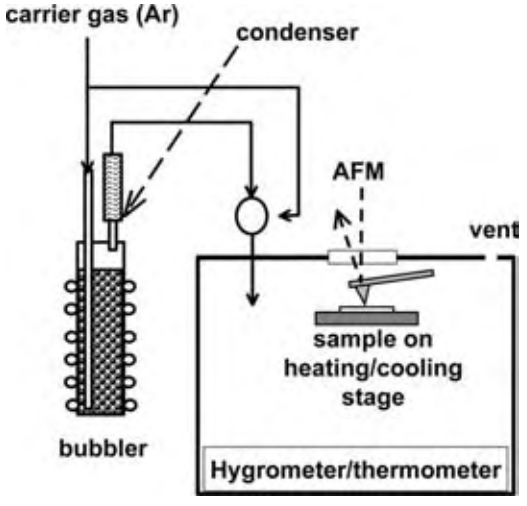


Fig. 3 Schematic diagram of the gas-handling system and the environment-controlled AFM chamber. The alcohol partial pressure is controlled by varying the ratio of dry Ar and alcohol saturated Ar gas flowrates. The alcohol saturated Ar gas stream is generated by flowing Ar through a heated bubbler followed by a condensor held at room temperature.

RESULTS AND DISCUSSION

Adsorption Isotherm of *n*-Propanol on QCM Electrode

Fig. 4 shows the adsorption isotherm of *n*-propanol on the QCM sensor at room temperature. The observed trend of the *n*-propanol film thickness as a function of partial pressure is consistent with the general characteristic of the alcohol adsorption isotherm observed for other systems.[8,9] The inset in Fig. 4 gives the approximate thickness of the adsorbed alcohol layer on the QCM sensor measured at a partial pressure of 90 ± 10% to the saturation pressure of each alcohol. The actual alcohol thickness on the clean, hydrophilic silicon oxide surface would be slightly larger than that on gold.

Formation of the Condensed Liquid Layer at the Saturation Vapor Pressure

At fully saturated vapor pressure, condensation occurs forming a thick liquid layer. The condensed alcohol

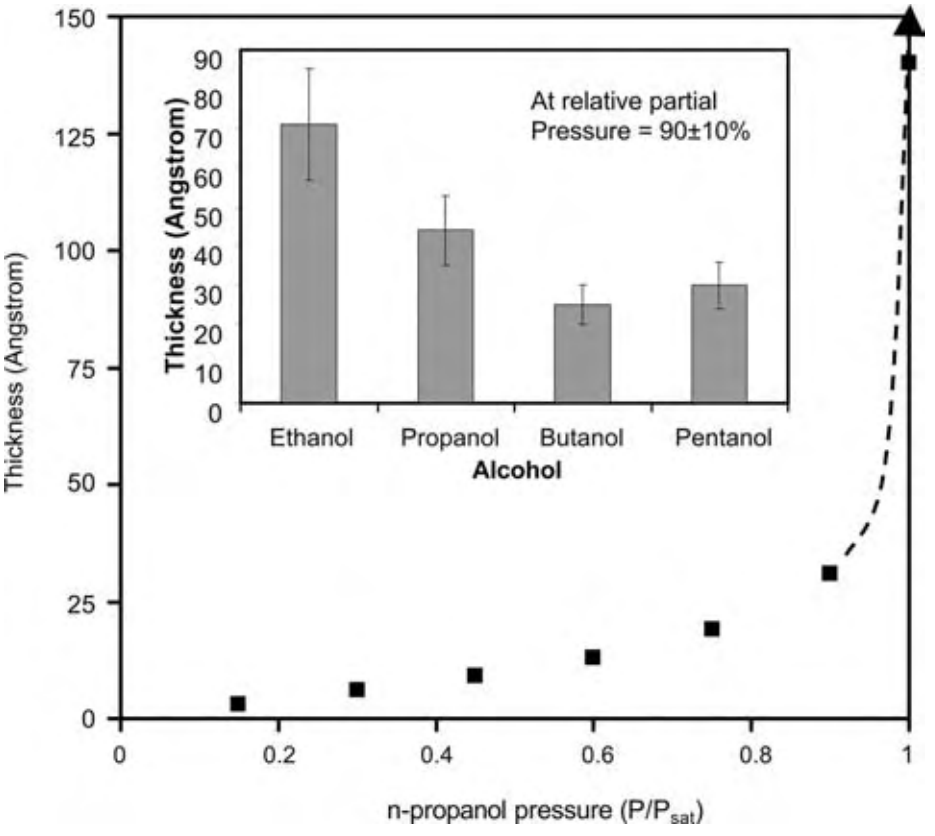


Fig. 4 Thickness of the adsorbed *n*-propanol layer measured with QCM as a function of *n*-propanol partial pressure. It should be noted that the QCM measures the thickness on a gold surface. The 100% saturation data point contains large error because a slight fluctuation in the substrate temperature causes a large change in the condensed layer thickness. The saturation vapor pressure on *n*-propanol is 21.2 Torr at room temperature. The inset shows the adsorbed layer thickness of different alcohols at partial pressures 90 ± 10% to the saturation. *(View this art in color at www.dekker.com.)*

layers are detected in the AFM *f–d* measurements. Fig. 5 shows the approach curve of the *f–d* measurement in the saturated butanol vapor as a function of time. At time zero, the environment suddenly changes by moving the AFM from the ambient air to the *n*-butanol saturated chamber. In ambient air, the AFM tip is snapped into the substrate surface at ~9 nm from the surface. When the environment is changed to saturated butanol vapor, the tip snaps in at a larger distance from the substrate surface. As the condensed butanol layer forms at both substrate and tip surfaces, capillary necking is formed between these condensed layers and pulls the cantilever downward (negative force in Fig. 5). The cantilever remains bent downward because of the capillary action of the condensed butanol layer until the tip finally touches the substrate surface at which the cantilever bends upward upon further movement toward the substrate. As the condensed layer forms on both the tip and the substrate surfaces, half of the distance between the initial snap-in and the substrate contact roughly corresponds to the thickness of the condensed butanol layer (~13 nm in Fig. 5). The calculation of the exact thickness will require detailed knowledge of the curvature dependence of the liquid layer thickness and the capillary neck formation dynamics, which are beyond the scope of this entry.

The snap-in behavior because of the capillary necking of the condensed layer is observed for all alcohols tested. The rate of the condensed layer formation and the equilibrium thickness of the condensed layer are very sensitive to small differences between the initial substrate temperature and gas temperature in the environment chamber. When the substrate

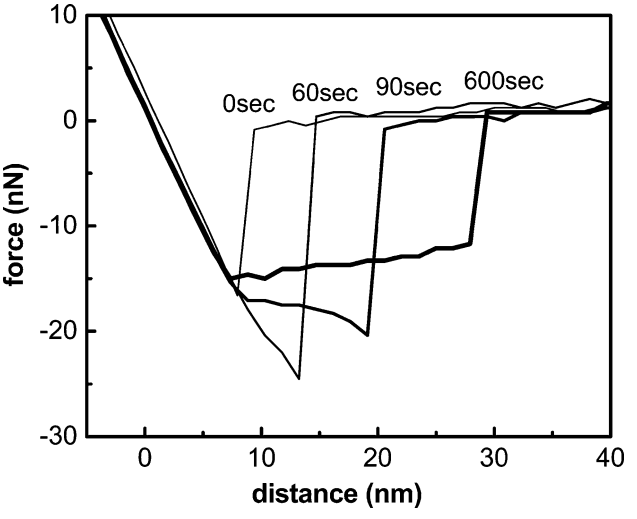


Fig. 5 Approach curve of *f–d* measurement as a function of time upon environment change from ambient air to saturated butanol vapor. The force constant of the cantilever is 2.4 N/m. A blunt tip is used for this measurement.

temperature is lower than the environment chamber temperature, a thick condensed layer (up to ~100 nm) is formed.

The presence of the condensed alcohol layer affects tip–substrate separation processes in *f–d* measurements. Fig. 6 shows the change in adhesion force over time as the environment changes from ambient air to saturated propanol vapor. The pull-off force shown in this figure was measured using a blunt (used) tip to minimize the tip damage/wear effect in subsequent measurements. Although it makes quantitative calculation difficult, the use of a blunt tip is advantageous when taking a series of measurements under high load conditions. The relative change of the contact area occurring at each contact with the substrate surface is assumed to be negligible for a blunt tip. The data obtained with a single blunt tip can then be directly compared without a contact area calibration. In ambient humid air, there is only one pull-off from the substrate. The adhesion force is very large because of the large contact area and the presence of water molecules at the interface. Upon introduction of saturated *n*-propanol vapor, the force required to pull the tip off the surface decreases by ~70% within the first 30 sec. At equilibrium, the adhesive force is reduced to ~90% of the original value. It should also be noted that there is a second snap-off in the retraction curves taking place in saturated alcohol environments. This second snap-off is because of the presence of condensed alcohol formed on the tip and the substrate surfaces. In the first snap-off, the tip is separated from the substrate, but it is still immersed in the condensed layer. As the tip is pulled further away from the surface, it eventually detaches from the condensed layer. As the thickness of the condensed alcohol layer increases, the second pull-off occurs further from the substrate surface.

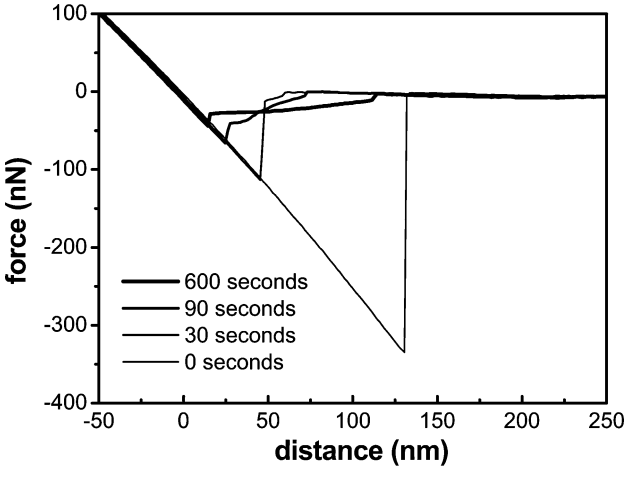


Fig. 6 Retraction part of the *f–d* measurement shown in Fig. 5.

Reduction of Adhesion Force by Alcohol Adsorption

The *n*-propanol adsorption from the gas phase significantly decreases the adhesion force between the silicon oxide surfaces measured with AFM. Fig. 7 illustrates the adhesion force measured with a single tip (2.2 N/m) at various partial pressures of *n*-propanol. The adhesion force decreases 40% compared with the dry Ar case upon initial introduction of *n*-propanol partial pressure. This reduction is not as drastic upon further increase of the *n*-propanol partial pressure. A sudden change in adhesion with only 10% partial pressure indicates that a few monolayer thick *n*-propanol film, as shown in Fig. 4, is sufficient to reduce the adhesion between the silicon oxide surfaces. This behavior is in sharp contrast to the relationship between the adhesion force and the water adsorption isotherm. In the case of water, the adhesion force increases several fold when the relative humidity increases from zero to 50%.[23,24]

This difference could be attributed to the amphiphilic nature of the alcohol. In the case of water, the free OH groups at the surface of the adsorbed water layer on the substrate can form hydrogen bonding with their counterparts on the AFM tip causing high adhesion force. In the case of *n*-propanol, the OH groups have strong interaction with the substrate surface,[25] and the hydrophobic propyl groups are likely to be exposed at the air interface of the adsorbed layer. In this structure, the hydrogen bonding interaction between the two *n*-propanol layers adsorbed on the substrate and the tip will be very small, leading to much smaller adhesion force.

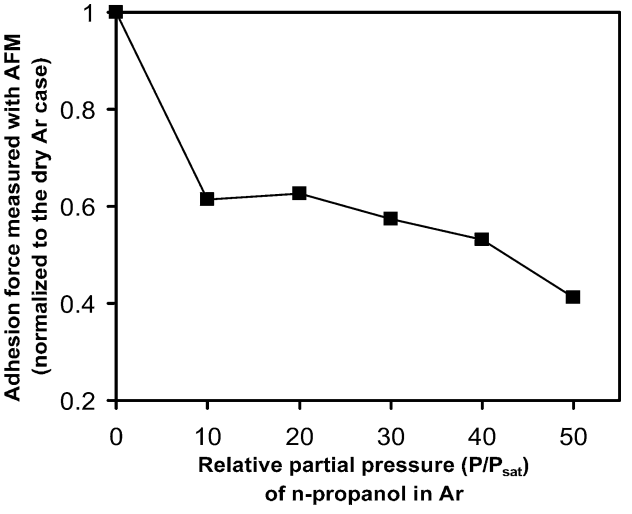


Fig. 7 Adhesion force measured with AFM as a function of *n*-propanol partial pressure. The data are normalized to the dry Ar case, which is measured before the *n*-propanol exposure. Data are acquired with a single tip (2.2 N/m).

Fig. 8 shows the variation of the force required to pull off the tip from the substrate surface with time as the environment is changed from ambient air to the saturated *n*-propanol environment and back to ambient air. Once an equilibrium was reached, the tip was scanned over a $2\,\mu m \times 2\,\mu m$ region of the substrate for further data analysis. The pull-off force decreases immediately upon the exposure of the substrate and tip to the saturated *n*-propanol vapor and increases upon the return of the substrate to ambient humid air. When the substrate is returned to ambient humid air, the pull-off force is initially higher than the equilibrium value and then gradually decreases to the equilibrium with some scatter. The transient behavior is somewhat different from run to run, but the overall trend is quite reproducible. Similar behavior is observed for all four alcohols studied. The origin of this transient overshoot is not clear at this moment. It might be related to the fast evaporation of alcohol from the condensed layer. As the alcohol evaporates, the substrate is cooled locally, prompting large adsorption of water at the cold surface. As the substrate temperature recovers to the temperature equilibrated with the surroundings, the excess water will desorb and the surface reaches equilibrium.

Lubrication by Adsorbed Alcohol Molecules

The change in friction force upon introduction of alcohol vapors is monitored with the lateral force signal during contact-mode scanning. The lateral force signal was measured for several seconds in ambient air to find the initial equilibration value. The substrate and AFM tip were then enclosed in the saturated alcohol vapor environment while scanning, and the lateral force signal was monitored. Fig. 9 shows the magnitude of friction force decrease for each alcohol vapor at saturation environments. To make qualitative comparisons between different alcohols without experimental artifacts caused by the tip curvature difference in each run, all of the data in Fig. 9 are normalized to the initial lateral force signal measured on native silicon oxide in ambient air. The inset in the figure illustrates the change in lateral force signal vs. time. At $t = 0$ sec, the AFM is moved from ambient air to propanol saturated environments and at $t = \sim 1000$ sec, it is returned back to ambient air. The spike in friction force measured upon return to the air might be related to the transient cooling because of evaporation of condensed alcohol. These results illustrate that the lateral force decreases significantly in all the cases studied.

Fig. 10 displays the friction coefficient as a function of the *n*-propanol partial pressure. The friction coefficient is determined from the slope of the lateral force vs. normal load plot. It illustrates that the formation of the *n*-propanol layer on the silicon oxide surface significantly reduces the friction coefficient. It should be noted that the decreasing friction coefficient trend

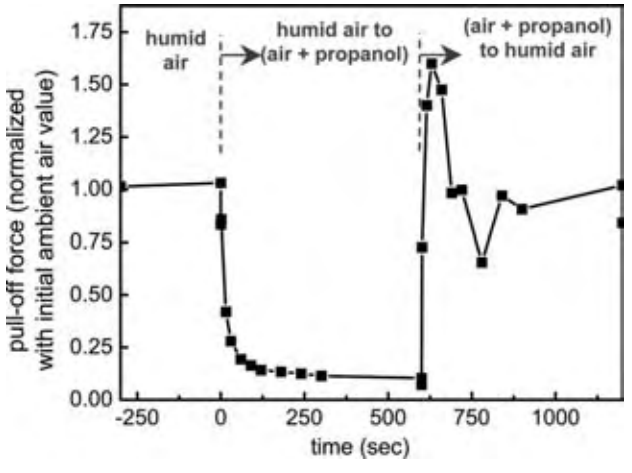


Fig. 8 Pull-off force changes upon environment change from ambient humid air to saturated *n*-propanol vapor and back to ambient air. The $1\,\mu m \times 1\,\mu m$ surface was scanned between –300 and 0, 300 and 600, and 900 and 1200 sec with a normal force of 50 nN. The tip is blunt because of wear in contact-mode scanning made before each *f–d* measurement. The force constant of the cantilever is 2.2 N/m. *(View this art in color at www.dekker.com.)*

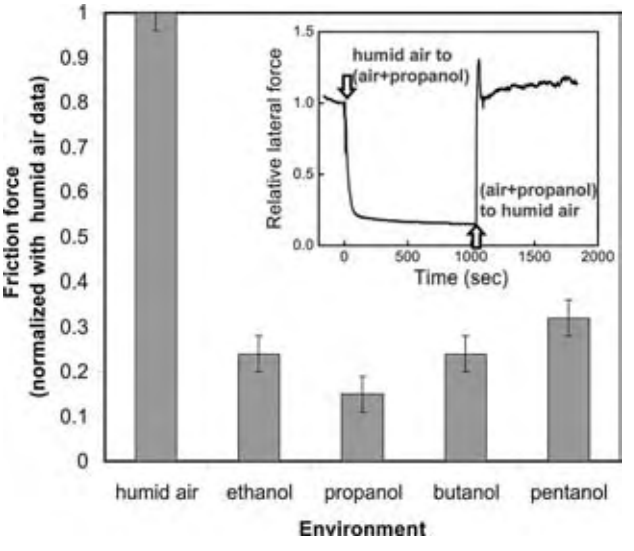


Fig. 9 Decrease of friction force upon environment change from ambient humid air to saturated alcohol vapor. The friction signals are normalized to the initial friction signals measured in ambient air for each experiment to minimize errors because of tip diameter variance in each measurement and emphasize the relative change because of condensed alcohol layer formation. The inset shows temporal changes of friction force upon sudden environment changes. The applied normal signal is kept constant at 75 nN. *(View this art in color at www.dekker.com.)*

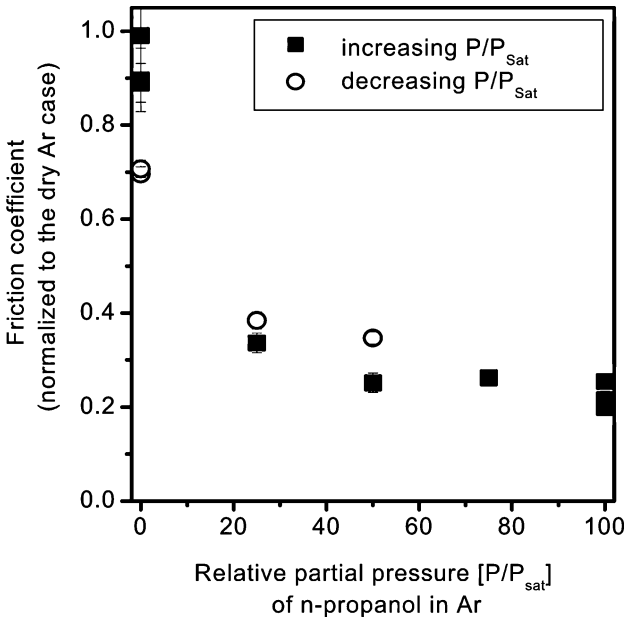


Fig. 10 Friction coefficient measured in contact-mode AFM scan as a function of *n*-propanol partial pressure. Filled squares are data taken at increasingly larger constant partial pressures, and open circles are those taken at decreasingly smaller constant partial pressures.

is very close to that of the adhesion force (Fig. 7). Both the friction coefficient and the adhesion force show large initial decreases upon introduction of *n*-propanol vapor followed by a small decrease with further increase of *n*-propanol vapor. The friction coefficient was also measured while the *n*-propanol partial pressure was decreased from the saturated vapor condition to the dry argon condition. Upon decrease of the *n*-propanol partial pressure from the saturation vapor pressure to zero, the friction coefficient increases close to its initial values.

Lubrication Mechanisms

A qualitative model for lubrication of the condensed alcohol layer is not fully developed yet. We present here the best qualitative model based on the literature. As the applied pressure is very large and the scanning speed is very low in AFM, one can rule out the hydrodynamic lubrication even in the presence of condensed alcohol layer on the substrate. The AFM tip and substrate interface must be in the boundary lubrication regime. Figs. 7–10 show that the adhesion force reduction upon alcohol adsorption is accompanied by reduction of the friction force. Part of the adhesion force reduction is because of the surface tension decrease of the water layer on the substrate. When alcohol is dissolved in water, the surface tension decreases significantly because of segregation of alcohol molecules to the liquid–air interface.[15] Considering purely mechanical components, one can relate this to the JKR theory, which predicts a smaller contact area at a lower adhesion force ($A \propto F_{ad}^{1/3}$). In asperity contact, the friction force is proportional to the contact area as well as the friction coefficient ($F = sA + \mu F_{normal}$ where s is the critical shear modulus and μ is the friction coefficient).

The main question from a chemical aspect is the identification of the species that are responsible for reduction of adhesion and friction at the interface. It is speculated that the alkoxide groups are formed at the substrate. In the self-assembled monolayer study, it is well known that alkyl groups at the interface reduce the adhesion and friction forces.[26] Gates and Hsu pointed out the alkoxide formation at the alcohol and silicon oxide interface and demonstrated their lubrication and antiwear properties.[10] These alkoxide groups can effectively prevent the chemical wear process associated with water. In the presence of water molecules and surface Si–OH groups, the Si–O–Si bonds can be easily formed and dissociated at the interface.[27] This bond formation and dissociation would cause large adhesion and friction. If the reaction proceeds to the formation of $Si(OH)_4$ or other easily removable species, these reaction products would be released from the substrate surface, producing wear. In the presence of alkoxide at the surface, the water-mediated chemical reactions will not take place, which will result in lower adhesion and friction and prevention of wear. The alkoxide group is of course susceptible to wear under extreme pressure. In our approach, the surface sites exposed by the loss of alkoxide groups will be replenished immediately with alcohol layers. It is this continuous replenishment that makes the adsorption of gas-phase alcohol molecules work as an effective lubrication process. Without continuous replenishment, the alcohol layer lubrication will eventually fail as do other self-assembled monolayer approaches.

CONCLUSIONS

The adhesion and friction forces between silicon oxide surfaces decrease drastically and immediately upon exposure of native silicon oxide surface to vapors of short-chain linear alcohols. The adhesion and friction force changes appear to have reversibility. Re-exposure of the substrate to ambient (humid) air causes the adhesion and friction force to increase back to the original values. The large decrease in friction force in an alcohol environment is in agreement with the decrease in adhesion force. These results indicate the potential of the gas-phase lubrication for MEMS applications. The gas-phase process can be used as an in-use lubrication in which lubricant molecules are supplied to the working

interface while the device is in operation. Furthermore, it can be used in combination with other solid lubricants to enhance their performance.

REFERENCES

1. Hsu, S.M.; Ying, Z.C. *Nanotribology: Critical Assessment and Research Needs*; Kluwer Academic Publishers: Boston, 2003.
2. Romig, A.D.; Dugger, M.T.; McWhorter, P.J. Materials issues in microelectromechanical devices: science, engineering, manufacturability and reliability. Acta Mater. **2003**, *51* (19), 5837.
3. Sundararajan, S.; Bhushan, B. Micro/nanotribology of ultra-thin hard amorphous carbon coatings using atomic force friction force microscopy. Wear **1999**, *229*, 678.
4. Eapen, K.C.; Patton, S.T.; Zabinski, J.S. Lubrication of microelectromechanical systems (MEMS) using bound and mobile phases of Fomblin Zdol (R). Tribol. Lett. **2003**, *12* (1), 35.
5. Maboudian, R.; Ashurst, W.R.; Carraro, C. Tribological challenges in micromechanical systems. Tribol. Lett. **2002**, *12* (2), 95.
6. Pashley, R.M.; Kitchener, J.A. Surface forces in adsorption multilayers of water on quartz. J. Colloid Interface Sci. **1979**, *71*, 491.
7. Patton, S.T.; Zabinski, J.S. Failure mechanisms of a MEMS actuator in very high vacuum. Tribol. Int. **2002**, *35* (6), 373.
8. Kittaka, S.; Umezu, T.; Ogawa, H.; Maegawa, H.; Takenaka, T. Interaction of alcohols with surface hydroxyls on chromium(III) oxide. Langmuir **1998**, *14* (4), 832.
9. Adams, M.J.; Briscoe, B.J.; Law, J.Y.; Luckham, P.F.; Williams, D.R. Influence of vapor condensation on the adhesion and friction of carbon–carbon nanocontacts. Langmuir **2001**, *17* (22), 6953.
10. Gates, R.S.; Hsu, S.M. Silicon-nitride boundary lubrication–lubrication mechanism alcohols. Trib. Trans. **1995**, *38* (3), 645.
11. Maboudian, R.; Carraro, C. Surface engineering for reliable operation of MEMS devices. J. Adhesion Sci. Technol. **2003**, *17* (4), 583.
12. Muller, B.; Heal, M.R. Mass accommodation coefficients of phenol, 3-nitrophenol, and 3-methylphenol over the temperature range 278–298 K. J. Phys. Chem. A **2002**, *106* (20), 5120.
13. Li, Y.Q.; Davidovits, P.; Shi, Q.; Jayne, J.T.; Kolb, C.E.; Worsnop, D.R. Mass and thermal accommodation coefficients of $H_2O(g)$ on liquid water as a function of temperature. J. Phys. Chem. A **2001**, *105* (47), 10627.
14. Vezenov, D.V.; Zhuk, A.V.; Whitesides, G.M.; Lieber, C.M. Chemical force spectroscopy in heterogeneous systems: Intermolecular interactions involving epoxy polymer, mixed monolayers, and polar solvents. J. Am. Chem. Soc. **2002**, *124* (35), 10578.
15. Lavi, P.; Marmur, A. Adsorption isotherm for concentrated aqueous-organic solutions (CAOS). J. Colloid Interface Sci. **2000**, *230* (1), 107.
16. Chattoraj, D.K.; Moulik, S.P. Binding of water and organic liquid components (ethyl alcohol, methyl alcohol, propyl alcohol, formic acid, acetone, pyridine + glycerol. Ind. J. Chem. **1977**, *15a*.
17. Donaldson, D.J.; Anderson, D. Adsorption of atmospheric gases at the air–water interface. 2. C1–C4 alcohols, acids, and acetone. J. Phys. Chem. A **1999**, *103* (7), 871.
18. Li, Z.X.; Lu, J.R.; Styrkas, D.A.; Thomas, R.K.; Rennie, A.R.; Penfold, J. The structure of the surface of ethanol-water mixtures. Mol. Phys. **1993**, *80* (4), 925.
19. Wolfrum, K.; Laubereau, A. Vibrational sum-frequency spectroscopy of an adsorbed monolayer of hexadecanol on water—destructive interference of adjacent lines. Chem. Phys. Lett. **1994**, *228* (1–3), 83.
20. Miranda, P.B.; Pflumio, V.; Saijo, H.; Shen, Y.R. Chain–chain interaction between surfactant monolayers and alkanes or alcohols at solid/liquid interfaces. J. Am. Chem. Soc. **1998**, *120* (46), 12092.
21. Stewart, E.; Shilds, R.L.; Taylor, R.S. Molecular dynamics simulations of the liquid/vapor interface of aqueous ethanol solutions as a function of concentration. J. Phys. Chem. B **2003**, *107* (10), 2333.
22. Chen, B.; Siepmann, J.I.; Klein, M.L. Vapor–liquid interfacial properties of mutually saturated water/1-butanol solutions. J. Am. Chem. Soc. **2002**, *124* (41), 12232.
23. He, M.; Blum, A.S.; Aston, D.E.; Buenviaje, C.; Overney, R.N.; Luginbuhl, R. Critical phenomena of water bridges in nanoasperity contacts. J. Chem. Phys. **2001**, *114* (3), 1355.
24. Qian, L.; Tian, F.; Xiao, X. Tribological properties of self-assembled monolayers and their substrates under various humid environments. Tribol. Lett. **2003**, *15* (3), 169.
25. Mizukami, M.; Moteki, M.; Kurihara, K. Hydrogen-bonded macrocluster formation of ethanol on silica surfaces in cyclohexane. J. Am. Chem. Soc. **2002**, *124* (43), 12889.
26. Noy, A.; Vezenov, D.V.; Lieber, C.M. Chemical force microscopy. Annu. Rev. Mater. Sci. **1997**, *27*, 381.
27. Maw, W.; Stevens, F.; Langford, S.C.; Dickinson, J.T. Single asperity tribochemical wear of silicon nitride studied by atomic force microscopy. J. Appl. Phys. **2002**, *92* (9), 5103.

Gas–Solid Reactions

Douglas P. Harrison
Gordon A. and Mary Cain Department of Chemical Engineering, Louisiana State University, Baton Rouge, Louisiana, U.S.A.

INTRODUCTION

Noncatalytic gas–solid reactions are discussed from the perspectives of single-pellet kinetics and integral reactor design. The single-pellet kinetics section includes a description of the traditional unreacted core and grain models followed by a brief discussion of the importance of solid structural property changes resulting from reaction and/or sintering. The basic equations used to describe noncatalytic gas–solid reactions in fixed-bed, fluidized-bed, and moving-bed reactors are then developed. The mathematical analysis is limited to isothermal reactions. This entry concludes with brief descriptions of selected commercial processes in which noncatalytic gas–solid reactions play a major role. Traditional processes such as the production of iron from its ore in the blast furnace and new processes such as the production of silicon for semiconductor applications are included. The processes selected represent each of the integral reactor types.

BACKGROUND

Gas–solid noncatalytic reactions are encountered in a broad range of industrial processes, including coal combustion and gasification, mineral processing, the capture of contaminants from gas streams, and the regeneration of coked catalysts. In applications such as combustion and gasification, the solid phase disappears as the reaction proceeds, while in others such as the capture of sulfur from gas streams, the solid reactant is converted into a solid product.

Although many similarities exist between gas–solid catalytic and gas–solid noncatalytic reactions, the noncatalytic systems, particularly when a porous reactant is converted to a porous product, are more complex. Both occur as the result of a number of series–parallel steps. Mass transfer of reacting gas from the bulk gas to the exterior of the solid and that of gas product from the solid to the bulk gas are involved in each. Diffusion of the reacting gas from the exterior surface into a porous catalyst or porous solid reactant and that of gas product from the pores to the exterior surface are also common to the two types of reactions. Adsorption of reacting gas, surface reaction, and desorption of product gas on catalyst active sites and on solid reactant are similar. However, the noncatalytic gas–solid system is inherently unsteady state as the solid is a stoichiometric participant in the reaction. Catalytic reactions, except when catalyst deactivation is rapid, are generally considered to be steady state. While catalysis is limited to surface active sites, the noncatalytic reaction may occur throughout the solid volume, which adds a product layer diffusion step. In addition, structural properties such as surface area and pore size distribution of catalysts are generally constant and they may vary in a noncatalytic reaction because of differences in the molar volumes of solid reactant and product. While sintering can cause structural properties to change in both a catalyst and solid reactant, catalytic reaction conditions are chosen so that sintering is, at worst, a slow process. In contrast, noncatalytic reactions may occur at such high temperatures that sintering cannot be avoided.

The desired product from gas–solid reactions may be a gas, such as in the production of synthesis gas during coal gasification, a solid, as encountered in the roasting (oxidation) of sulfide ores, "heat" from the combustion of a solid fuel, or some combination. The solid product may be used directly or may require further processing. When the solid is used to capture contaminants from the gas stream, it may constitute a waste material such as occurs during the formation of $CaSO_3$ and $CaSO_4$ in flue gas desulfurization processes, or the solid reactant may be regenerated for multicycle use as in the fluidized-bed catalytic cracking process.

The types of reactors used for catalytic and noncatalytic gas–solid reactions are also often similar. Moving-bed reactors are used in blast furnaces and cement kilns. Fluidized-bed reactors are used for the roasting of sulfide ores and regeneration of catalytic cracking catalyst, and fixed-bed reactors are used to remove sulfur compounds from ammonia synthesis feed gas. When regeneration of the solid reactant is desired, two or more reactors operating in parallel are required if continuous, steady-state operation is to be achieved.

This entry examines noncatalytic gas–solid reactions beginning with the behavior of single pellets, moving to the performance of integral reactors, and ending with a description of selected noncatalytic gas–solid reaction processes.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007786

SINGLE-PELLET REACTIONS MODELS

Numerous mathematical models describing the reactions of single solid pellets were proposed in the period ranging roughly from 1960 to 1980. While more sophisticated variations have been presented subsequently, the early models established the basic mathematical framework. The appropriate kinetic model depends on whether the solid reactant is porous or nonporous and whether a solid product is formed by the reaction. With no solid product, the pellet gradually disappears as the reaction proceeds, while if a solid product is formed it is normally assumed that the pellet is of constant diameter. These cases are represented by the generic equations:

$$\mathrm{A}(g) + b\mathrm{B}(s) \leftrightarrow c\mathrm{C}(g) \tag{1}$$

$$\mathrm{A}(g) + b\mathrm{B}(s) \leftrightarrow c\mathrm{C}(g) + d\mathrm{D}(s) \tag{2}$$

The following discussion is limited to pellets of constant diameter. The so-called unreacted core and grain models, which are developed in the following sections, provide a clear picture of the overall processes that occur in the reactions of nonporous and porous particles of constant diameter. Mathematical analysis is limited to isothermal reactions. Models that include the effects of property variations are then described briefly.

The mathematics is based on a single spherical pellet. Szekely et al.[1] present a similar approach for other geometries (cylinders and flat plates).

THE UNREACTED CORE MODEL

This model is applicable to the reactions of nonporous pellets and to porous pellets when the global rate is controlled by pore diffusion. Reaction is limited to a surface separating the solid reactant at the core of the pellet surrounded by a porous layer of solid product. It occurs initially on the external surface of the pellet, and the thickness of the product layer increases as the reaction proceeds, as illustrated in Fig. 1. The global reaction rate is determined by three resistances—mass transfer from bulk gas to particle surface, diffusion through the product layer, and the intrinsic chemistry at the reaction surface.

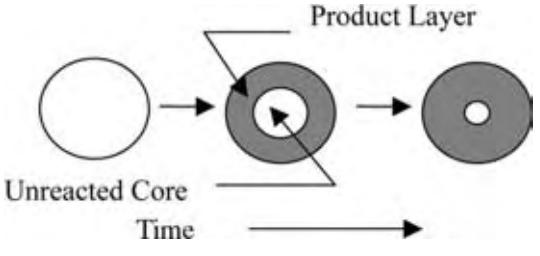


Fig. 1 The unreacted core model.

The mass balance equation and boundary conditions for gas reactant A are:

$$\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r^2\mathrm{d}C_\mathrm{A}}{\mathrm{d}r}\right) = 0 \quad \text{for } r_\mathrm{s} > r > r_\mathrm{c} \tag{3}$$

$$D_\mathrm{eA}\left(\frac{\mathrm{d}C_\mathrm{A}}{\mathrm{d}r}\right) = -R_\mathrm{A} \quad \text{at } r = r_\mathrm{c} \tag{4}$$

$$D_\mathrm{eA}\left(\frac{\mathrm{d}C_\mathrm{A}}{\mathrm{d}r}\right) = k_\mathrm{mA}(C_\mathrm{Ao} - C_\mathrm{As}) \quad \text{at } r = r_\mathrm{s} \tag{5}$$

The intrinsic rate equation for first-order kinetics is

$$R_\mathrm{A} = k_\mathrm{s}C_\mathrm{So}C_\mathrm{Ac} \tag{6}$$

The position of the unreacted core as a function of time is given by the solid reactant material balance and initial condition

$$\frac{\mathrm{d}r_\mathrm{c}}{\mathrm{d}t} = bk_\mathrm{s}C_\mathrm{Ac} \tag{7}$$

$$r_\mathrm{c} = r_\mathrm{s} \quad \text{at } t = 0 \tag{8}$$

The fractional solid conversion and the radius of the unreacted core are related by

$$x = 1 - \frac{r_\mathrm{c}^3}{r_\mathrm{s}^3} \tag{9}$$

All terms are defined in the nomenclature section. The above equations can be solved analytically for an isothermal reaction to obtain the following algebraic relationships for C_A as a function of r_c and x as a function of t.

$$\frac{C_\mathrm{A}}{C_\mathrm{Ao}} = \frac{\left[\left(1 - \frac{D_\mathrm{eA}}{K_\mathrm{s}C_\mathrm{So}r_\mathrm{c}}\right)\frac{1}{r_\mathrm{c}} - \frac{1}{r}\right]}{\left[1 - \frac{D_\mathrm{eA}}{k_\mathrm{s}C_\mathrm{So}r_\mathrm{c}}\right]\frac{1}{r_\mathrm{c}} - \left[1 - \frac{D_\mathrm{eA}}{k_\mathrm{mA}r_\mathrm{s}}\right]\frac{1}{r_\mathrm{s}}} \tag{10}$$

$$t = \frac{r_\mathrm{s}C_\mathrm{So}}{(-b)C_\mathrm{Ao}}\left[\frac{x}{3k_\mathrm{mA}} + \frac{r_\mathrm{s}}{6D_\mathrm{eA}}\{1 + 2(1 - x) - 3(1 - x)^{2/3}\} + \frac{1}{k_\mathrm{s}C_\mathrm{so}}\{1 - (1 - x)^{1/3}\}\right] \tag{11}$$

Eq. (11) shows the relationship between the three reaction resistances. The first term represents the external mass transfer resistance, the second the resistance associated with diffusion through the product layer, and the third the chemical reaction resistance at the reactant-product interface.

In spite of the existence of the three resistances, it has been common practice to interpret experimental data on the basis of a single controlling resistance. When external mass transfer resistance controls, $k_{mA} \ll D_{eA}/r_s$ and $k_{mA} \ll k_s C_{So}$, so that Eq. (11) reduces to

$$t = \frac{r_s C_{So} x}{3(-b) C_{Ao} k_{mA}} \tag{12}$$

A linear relationship between t and x is used to conclude that mass transfer is the controlling resistance and to evaluate the mass transfer coefficient. In practice, the relative importance of the mass transfer resistance should decrease with time as the reaction proceeds and the global rate decreases. Hence, it is unlikely that mass transfer resistance would be the rate-controlling step throughout the entire reaction period.

When diffusion through the product layer controls the global rate, $D_{eA}/r_s \ll k_{mA}$ and $D_{eA}/r_s \ll k_s C_{So}$, Eq. (11) becomes

$$t = \frac{r_s^2 C_{So}}{6(-b) C_{Ao} D_{eA}} \left[1 + 2(1 - x) - 3(1 - x)^{2/3}\right] \tag{13}$$

A linear relationship between t and the term in brackets is indicative of product layer diffusion control. However, the product layer diffusion resistance must be zero initially when no product layer exists. Product layer diffusion resistance may increase rapidly as the product layer thickness increases, and, particularly at high temperature, can become rate-controlling during a large portion of the reaction.

Finally, when the surface reaction resistance controls the global rate, $k_s C_{So} \ll k_{mA}$ and $k_s C_{So} \ll D_{eA}/r_s$, Eq. (11) becomes

$$t = \frac{r_s}{(-b) k_s C_{Ao}} \left[1 - (1 - x)^{1/3}\right] \tag{14}$$

The surface reaction resistance is most likely to be important at lower reaction temperatures and is the only resistance that can truly be rate controlling throughout the entire reaction.

A similar treatment may be used to describe the gas–solid reaction rate when there is no solid product formed such as encountered in solid fuel combustion. However, in that case, the product layer diffusion resistance is always zero.

A classic example of the use of the unreacted core model is that of Weisz and Goodwin[2] who studied the regeneration of fluidized-bed cracking catalyst. Although the cracking catalyst was porous, at a reaction temperature of 700°C the global rate was controlled by diffusion through the coke-free layer of catalyst and the relation between fractional regeneration and time agreed with Eq. (13). In addition, an examination of the radial profile of coke concentration in partially regenerated catalyst showed a high coke concentration near the center surrounded by a layer of essentially zero coke.

In another example, Beveridge[3] studied the oxidation of zinc sulfide spheres and reported that the global rate was, in turn, controlled by the surface chemical reaction at low temperatures, diffusion through the zinc oxide product layer at intermediate temperatures, and external mass transfer resistance at higher temperatures.

Costa and Smith[4] studied the hydrofluorination of nonporous uranium dioxide pellets under conditions where external mass transfer resistance was negligible. The global rate was initially controlled by the surface chemical reaction resistance, but switched to product layer diffusion as the reaction progressed and the product layer thickness increased.

For nonisothermal reactions, the energy balance must also be considered. The mass and energy balances are coupled through the reaction rate term and an analytical solution is no longer possible. Temperature increases for exothermic reactions may sometimes be quite large. For example, Themelis and Yannopoulos[5] studied the reduction of cupric oxide with hydrogen and, by imbedding very fine thermocouples in the pellet, found that the temperature increased from about 250 to 900°C in 1 min. Luss and Amundson[6] developed a model to calculate the maximum temperature rise in a spherical pellet by assuming a uniform temperature within the pellet.

THE GRAIN MODEL

The grain model is perhaps the simplest of a number of so-called structural models. The porous spherical pellet is assumed to be composed of a large number of spherical grains, each of which reacts according to the unreacted core model. A diagram of this model is shown in Fig. 2. The magnified section shows completely reacted grains near the outside of the pellet, unreacted grains toward the center, and partially reacted grains between. Grain size may be estimated from measured surface area, while diffusion properties

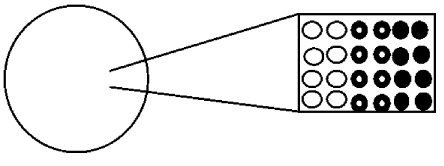


Fig. 2 The grain model.

between the grains may be estimated from porosity and pore size distribution measurements.

The grain model contains four reaction resistances: mass transfer of reactant gas from the bulk gas to the pellet surface, diffusion between the grains, diffusion through a solid product layer surrounding individual grains, and reaction at the interface of each grain. Product layer diffusion may occur by a solid-state mechanism or by diffusion through micropores. When chemical reaction at the grain interface is the controlling resistance, the gas concentration will be constant at all radial positions and the grains will react uniformly. When diffusion between grains is the controlling mechanism, the thickness of the reaction zone approaches zero and the grain model becomes identical to the unreacted core model.

The grain model gas phase material balance and boundary equations for a first-order reaction are

$$\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left[\frac{D_{\mathrm{eA}}r^2\mathrm{d}C_{\mathrm{A}}}{\mathrm{d}r}\right] + \frac{(-3)r_{\mathrm{c}}'^2k_{\mathrm{s}}'C_{\mathrm{So}}C_A}{r_{\mathrm{s}}'^3\left[1 - \frac{k_{\mathrm{s}}'C_{\mathrm{So}}r_{\mathrm{c}}'}{D_{\mathrm{eA}}'}\left(1 - \frac{r_{\mathrm{c}}'}{r_{\mathrm{s}}'}\right)\right]} = 0 \tag{15}$$

$$D_{\mathrm{eA}}\left(\frac{\mathrm{d}C_{\mathrm{A}}}{\mathrm{d}r}\right) = k_{\mathrm{mA}}(C_{\mathrm{Ao}} - C_{\mathrm{As}}) \quad \text{at } r = r_{\mathrm{s}} \tag{16}$$

$$\frac{\mathrm{d}C_{\mathrm{A}}}{\mathrm{d}r} = 0 \quad \text{at } r = 0 \tag{17}$$

The concentration at the grain surface, C_{As}', is assumed to be the same as that within the pores at the same radial position.

The radius of the unreacted core within each grain as a function of time is given by the solid reactant material balance and initial condition:

$$\frac{\mathrm{d}r_{\mathrm{c}}'}{\mathrm{d}t} = \frac{bk_{\mathrm{s}}'C_{\mathrm{A}}}{\left[1 - \frac{k_{\mathrm{s}}'C_{\mathrm{So}}'r_{\mathrm{c}}'}{D_{\mathrm{eA}}'}\left(1 - \frac{r_{\mathrm{c}}'}{r_{\mathrm{s}}'}\right)\right]} \tag{18}$$

$$r_{\mathrm{c}}' = r_{\mathrm{s}}' \quad \text{at } t = 0 \tag{19}$$

Numerical solution is required even for an isothermal reaction. However, an approximate solution based on adding the times required to reach a given fractional conversion if each separate step in the process is assumed to be rate limiting has been published[7] that agrees quite well with the numerical solution.

A number of studies have used the grain model as the basis for comparison with experimental data. For example, Szekely and Evans[8] studied the reduction of nickel oxide pellets with hydrogen. Their grain model analysis considered the effects of external mass transfer, diffusion between grains, and surface reaction, but assumed that diffusion resistance through the product layer was negligible. Reasonable agreement was obtained between model predictions and experimental results except at reaction temperatures above 650°C, where the product solid experienced significant sintering.

Gibson and Harrison[9] used the grain model to analyze data on the reaction of hydrogen sulfide with zinc oxide. All model parameters with the exception of the grain diffusion coefficient were directly measured or estimated from literature correlations. Good agreement between the model and experimental results was found in the temperature range of 600–700°C when the grain product layer resistance was neglected. However, at lower temperatures, the grain model without product layer resistance was unable to predict the reaction "die-off" that occurred well before complete conversion of oxide to sulfide. Qualitative agreement was obtained when product layer diffusion resistance was included with the grain diffusion coefficient treated as a fit parameter. At higher temperatures, the failure of the model to agree with experimental results was attributed to solid structural property changes associated with sintering.

Other Single-Pellet Kinetic Models

The need to include solid structural property variations in the single-pellet models was illustrated by the high-temperature results from the two examples cited above. In those cases, the structural property variations were attributed to sintering of the solid product and/or reactant. Structural property variations may also occur at lower temperatures where sintering would not be expected because of the differences in molar volumes of solid reactant and product.

One of the early models that considered property variations owing to molar volume differences was the single pore model of Ramachandran and Smith.[10] The solid reactant initially consisted of a number of cylindrical pores of fixed length and diameter. The reaction that occurs along the walls of the pores causes a decrease in pore diameter when the molar volume of the product is larger than that of the reactant and an increase in pore diameter for the reverse case. The global reaction rate is determined by four resistances: external mass transfer, pore diffusion, solid product layer diffusion, and surface chemical reaction. Fig. 3 illustrates a case in which pore diffusion resistance is important and the molar volume of the product is greater than that of the reactant. The reaction tends to occur first near the entrance to the pore, causing the pore radius to be smaller near the pore mouth. Depending on the relative molar volumes of solid

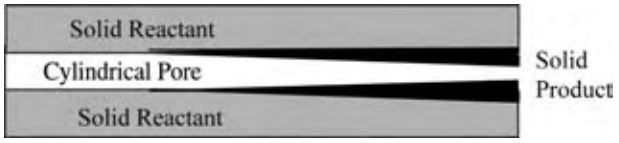


Fig. 3 Single pore model.

reactant and product, complete pore mouth closure may occur, thus preventing reactant gas from entering the pore and causing reaction "die-off" well before complete conversion of the solid. If the pore diffusion resistance is negligible, the gas concentration will be uniform throughout the length of the pore and pore radius change will be independent of pore length.

Ramachandran and Smith obtained satisfactory agreement with experimental results on the reduction of nickel oxide with carbon monoxide (pore opening case) by considering the product layer diffusion coefficient as an adjustable parameter. Similarly, the model predicted pore closure and reaction "die-off" for the reaction of calcium oxide with sulfur dioxide, where the molar volume of calcium sulfate product is about three times that of the calcium oxide reactant.

Hartman and Coughlin[11] modified the grain model to consider structural changes because of the differences in molar volume and also applied the model to the reaction between calcium oxide and sulfur dioxide. Although the grain radius was assumed to remain constant, the pore structure of the solid was allowed to change as the reaction progressed. As a result, the diffusion coefficient between grains was a function of radial position in the pellet. Again the product layer diffusion coefficient was treated as a fit parameter. The predicted fractional solid conversion corresponding to reaction "die-off" closely matched experimental values for a number of pellet sizes.

In a subsequent paper, Ramachandran and Smith[12] again used the grain model as the basis and considered the effects of both reaction and sintering on solid physical properties. As reaction proceeds, the grains expand, contract, or remain the same size due to differences in molar volumes and sintering. Sintering results in a decrease in product layer porosity that causes a reduction in product layer diffusion coefficient. By using best-fit model parameters, there was good agreement with experimental data on the hydrofluorination of uranium dioxide. It is important to note, however, that although the model was capable of including sintering, the effect was ignored in this case as sintering was thought to be negligible at the reaction temperature tested.

Ranade and Harrison[13] also modified the grain model to account for structural changes because of the combined effects of reaction and sintering. As the reaction proceeds, the individual grains expand or contract owing to the different molar volumes of solid reactant and product. Also, because of sintering, adjacent grains combine to give larger grains as shown in Fig. 4. The large number of small grains initially present is replaced by a smaller number of larger grains. For mathematical simplicity, a single unreacted core is assumed to remain at the center of sintered grains. In contrast to the standard grain model in which the grain core radius decreases monotonically as the reaction proceeds, sintering may actually lead to an increase in the unreacted core radius within the larger grains.

Solid structure was correlated with surface area measurements as a function of solid conversion under conditions where no sintering occurred, as a function of sintering time and temperature in the absence of reaction, and at conditions where reaction and sintering occurred simultaneously. The model was then tested using the experimental data of Gibson and Harrison[9] for the reaction between hydrogen sulfide and zinc oxide referred to in the previous section. Although sintering was quite severe at temperatures above 500°C, its inclusion in the model had a relatively small effect on predicted time–solid conversion performance.

INTEGRAL REACTOR MODELS

Gas–solid reactions are carried out on a commercial basis using fixed-bed, moving-bed, and fluidized-bed reactors. The fixed-bed reactor is an unsteady-state system as reactive gas is fed on a continuous basis through the reactor that is packed with a finite quantity of solid reactant. The solid is depleted and breakthrough of the gas reactant occurs after a certain reaction time. In the moving-bed reactor, both solid and gas are fed on a continuous basis and overall operation is steady state. The fluidized-bed reactor, where small solid particles are fluidized by upward flow of gas, also operates in a steady-state manner. Diffusional reaction resistances are reduced because of the small solid particles while solid backmixing reduces solid concentration gradients and promotes isothermal operation.

At least two reactors are needed with each reactor type if the solid is to be regenerated for multicycle operation. In the fixed-bed operation, the primary

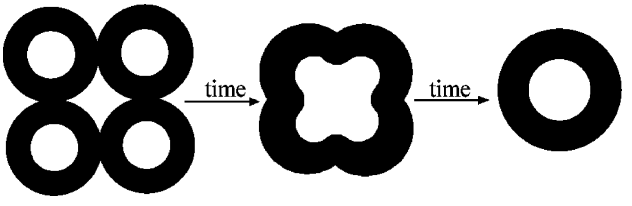


Fig. 4 The changing grain size model.

reaction is carried out in one reactor while regeneration occurs in the second. Gas flows are reversed when breakthrough occurs in the primary reactor and the functions of the two reactors are reversed. In the moving-bed system with solid regeneration, the first reactor is always used for the primary reaction and the second for regeneration. Solids are transported between the two reactors, perhaps by a system that resembles a bucket elevator, but gas flows always remain constant. The functions of the two reactors remain constant in the fluidized-bed scheme as well, where the small solid particles are pneumatically transferred between reactors.

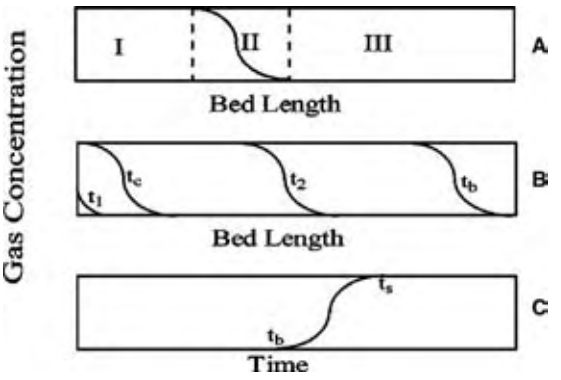


Fig. 6 Fixed-bed reactor performance.

Fixed-Bed Reactor System

A schematic diagram of the two-reactor fixed-bed system is shown in Fig. 5. The primary reaction occurs in the reactor on the left while regeneration occurs on the right. Dashed lines indicate gas flow paths when the functions of the two reactors are reversed. Both gas and solid concentrations and temperatures (for non-isothermal reactors) are functions of time and axial position.

The progress of the reaction is shown in Fig. 6. In Fig. 6A the reactor is divided into three regions. In region I, the solid has been fully reacted and the gas concentration is equal to the inlet gas concentration. Reaction occurs in region II where the gas concentration is reduced from its inlet value to near zero or to near the equilibrium value. Solid reactant in region III has not yet been exposed to reactant gas. Fig. 6B illustrates the movement of the reaction front through the bed with time. To a good approximation, once the reaction front is fully established at time, t_c, it moves through the bed at a constant velocity (known as constant-pattern behavior) until the leading edge of the reaction zone just corresponds to the exit of the packed bed at t_b, the breakthrough time. After the constant-pattern behavior has been established, the solid fraction conversion profile and the gas concentration profile shown in Fig. 6B are identical. Fig. 6C shows the product gas concentration as a function of time. Before breakthrough, the exit gas concentration is near zero. After breakthrough, the reactant gas concentration increases with time until all solid has reacted at t_s. Inlet and outlet gas concentrations are then equal.

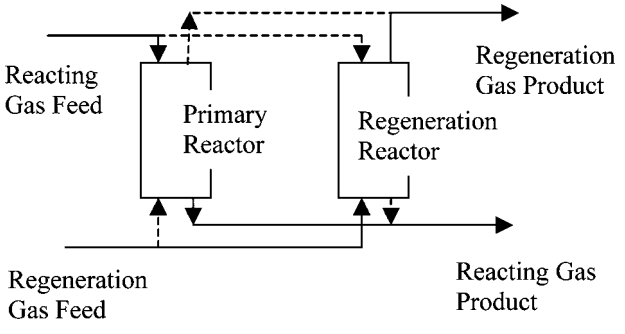


Fig. 5 The fixed-bed reactor system.

The performance of an isothermal fixed-bed reactor may be described by solving appropriate material balances on the gaseous and solid reactants. If plug flow of the gas phase is assumed, the gas reactant material balance and boundary condition are

$$U_o \frac{dC_A}{dz} = -k_{mA} a_v (C_{Ao} - C_{As}) = R_A^v \quad (20)$$

$$C_A = C_{Ai} \quad \text{at } z = 0 \quad (21)$$

Similarly, the solid phase material balance and initial condition are

$$(1 - \varepsilon_v) \frac{dC_s}{dt} = b k_{mA} a_v (C_{Ao} - C_{AS}) = b R_A^v \quad (22)$$

$$C_s = C_{So} \quad \text{at } t = 0 \quad (23)$$

The expression for R_A^v is obtained from the appropriate single-pellet model.

Wang et al.[14] compared the model predictions to experimental data on the desulfurization of simulated coal gas from a laboratory-scale fixed-bed reactor and from a process development reactor operated on actual coal gas. The solid reactant was formed from cylindrical pellets of zinc and titanium oxides. Data from six experiments using different temperatures, pressures, feed gas flow rates, and feed gas H_2S concentrations were available. Product gas concentrations were measured as a function of time, and the axial distribution of sulfur within the reactor was determined at the conclusion of the test.

The unreacted core model, suitably modified for cylindrical geometry, was used to describe the behavior

of a single pellet. All fixed-bed model parameters with the exception of the diffusion coefficient through the product layer, D_{eA}, associated with each pellet were established a priori. D_{eA} was chosen to provide good fit with experimental gas product concentrations during breakthrough. Variations in best-fit values of D_{eA} from run-to-run were qualitatively consistent with those in reactor temperature and pressure. In addition, the model predictions of axial sulfur concentration within the bed at the conclusion of the test using the previously determined value of D_{eA} were in good agreement with the measured sulfur distributions.

Fluidized-Bed Reactor System

A fluidized-bed reactor system may be operated with dual circulating beds as shown in Fig. 7, when the solid is to be regenerated or in a single bed if regeneration is not required. In the dual-bed mode, the solid is pneumatically transferred between reactors to provide for steady-state operation. The roasting of sulfide ores is an important example of the use of a single fluidized bed, while the dual circulation fluidized beds are used in the catalytic cracking of gas oil and regeneration of the coked catalyst.

The three-phase bubble model developed by Kunii and Levenspiel[15] shown in Fig. 8 is used to illustrate fluidized-bed reactor modeling. The bed is composed of bubble, cloud plus wake, and emulsion phases. The gas in the bubble phase is assumed to be in plug flow, and gas flow through the cloud and emulsion phases is neglected. The bubble is assumed to be free of solids so that reaction is restricted to the cloud and emulsion phases. Because of extensive solid backmixing, it is assumed that no solid concentration gradients exist (or temperature gradients in the case of a nonisothermal system). As the particles are much smaller than those used in fixed or moving beds, there is little, if any, mass transfer or intraparticle diffusion limitation on the global reaction rate, which leaves the intrinsic reaction rate as the dominant reaction resistance.

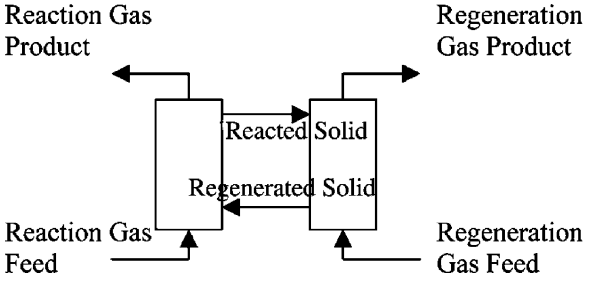


Fig. 7 The fluidized-bed reactor system.

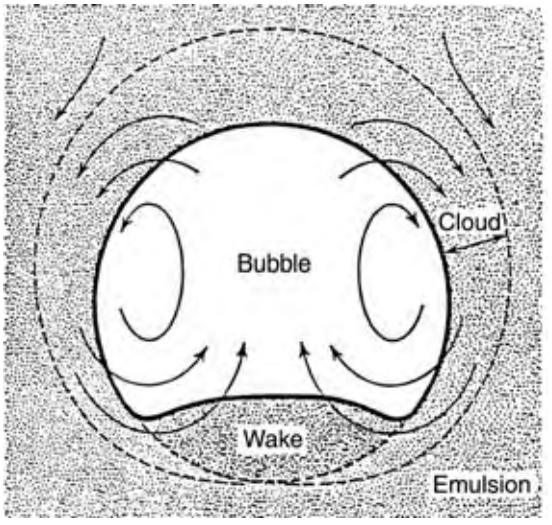


Fig. 8 The three-phase fluidized-bed reactor model.

With these simplifications, the material balance equation for reacting gas is

$$-U_b \frac{dC_{Ab}}{dz} = K_r C_{Ab} \tag{24}$$

K_r is a pseudo first-order rate constant. As the reactor is isothermal and the solid concentration is constant, Eq. (24) can be integrated to give an algebraic relationship between bed height and gas phase concentration in the bubble. Likewise, an overall solid reactant material balance yields an algebraic expression for the concentration of solid reactant in the reactor product.

$$C_{Ab} = C_{Ai} \exp\left(\frac{-K_r z}{U_b}\right) \tag{25}$$

$$C_{sf} = C_{So} + \frac{bV_o}{Q_{So}}(C_{Ao} - C_{Af}) \tag{26}$$

The pseudo first-order rate constant includes the effects of gas phase transfer between the bubble, cloud, and emulsion phases. Correlations for estimating model parameters may be found in the original reference.[15]

$$K_r = k_i C_{Sf} \left\{ \rho_b + \left[\frac{1}{\frac{k_i C_{Sf}}{k_{bc}} + \left(\frac{1}{\frac{\rho_c V_{jz}}{V_b} + \frac{1}{\frac{k_i C_{Sf}}{k_{ce}} + \frac{f}{\rho_c (1-f)}}} \right)} \right] \right\} \tag{27}$$

Matsui et al.[16] compared experimental data on the steam gasification of coal char from a laboratory-scale fluidized-bed reactor with the Kunii and Levenspiel model. Good agreement was reported when the bubble diameter was treated as a fitted constant.

Moving-Bed Reactor System

The moving-bed reactor has features common to both the fixed-bed and fluidized-bed reactors. Like the fluidized bed, the moving-bed reactor operates in a steady-state manner by continuous feed of both the solid and gas phases. Flow directions may be countercurrent, cocurrent, or crossflow as shown in Fig. 9. However, the particle sizes used in moving-bed reactors are similar to those used in fixed-bed reactors so that the resistances associated with mass transfer and intraparticle diffusion must be considered. Solids move slowly through the moving-bed reactor without the vigorous mixing encountered in the fluidized-bed system.

In the following analysis, both the solid and gas phases are assumed to move in countercurrent plug flow so that both concentrations (and temperature in nonisothermal reactors) vary with axial position. The material balance with boundary condition for the gas phase reactant is

$$\frac{U_o dC_A}{dz} = -k_{mA} a_v (C_A - C_{As}) = -R_A^v \tag{28}$$

$$C_A = C_{Ao} \quad \text{at } z = 0 \tag{29}$$

The material balance and boundary condition for the solid reactant are

$$U_S(1 - \varepsilon_v)\frac{dC_S}{dz} = -bk_{mA} a_v (C_A - C_{As}) = -bR_A^v \tag{30}$$

$$C_S = C_{So} \quad \text{at } z = L \tag{31}$$

The unreacted core or other appropriate single-pellet model may be used to describe the reaction rate.

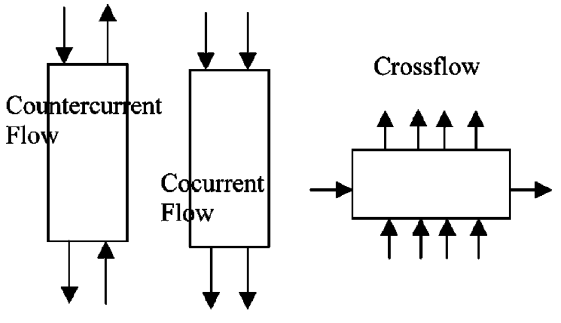


Fig. 9 Flow directions in moving-bed reactors.

PROCESSES INVOLVING NONCATALYTIC GAS–SOLID REACTIONS

This section provides brief descriptions of industrial processes in which noncatalytic gas–solid reactions play a major role. Although by no means complete, the discussion includes both traditional processes, such as the blast furnace for the production of iron from ore and the regeneration of fluidized-bed catalytic cracking catalyst, and newer processes such as the dry capture of SO_2 from flue gas and the production of silicon for semiconductor applications. Each of the three primary reactor types is represented in the processes described.

Blast Furnace

The iron blast furnace, which is the first step in the production of steel, provides an example of a countercurrent flow moving-bed reactor in which numerous gas–solid reactions occur. Iron ore, coke, and limestone are charged to the top, while hot air is fed to the bottom of a refractory-lined reactor vessel. The solids have a residence time of 6–8 hr, whereas the air residence time is only 6–8 sec. The iron ore may contain 50–70% iron, while the molten iron product will typically contain about 95% iron, 4% carbon, and 1% of a number of compounds including silicon, manganese, titanium, phosphorus, and sulfur.

Although large, modern blast furnaces are capable of producing more than 10,000 tons per day of iron, the basic process has not changed since iron was first produced some 700 years ago. Heat is supplied by the oxidation of coke, primarily to carbon monoxide. Both coke and carbon monoxide act as reducing agents for the reduction of iron oxide to metallic iron. The blast furnace temperature is sufficiently high that product iron is a liquid that flows from the bottom of the reactor. Limestone is decomposed to calcium oxide, which reacts with sulfur compounds present in the ore and also combines with other impurities such as alumina and silica to form a low-melting slag that floats on top of the molten iron. Dirty gases emerging from the top of the blast furnace contain a large quantity of carbon monoxide and are burned to preheat the inlet air and/or to generate steam in the plant boiler.

Direct Reduction Iron (DRI)

While the traditional blast furnace is preferred for large-scale operations, processes for the direct reduction of iron ore using carbon monoxide and/or hydrogen are becoming increasingly important. Such processes have the ability to use lower-grade iron ores, particularly fine ore particles, and fuels such as natural gas that are unacceptable for the blast furnace. The iron

product is used as a substitute for scrap steel in the electric arc steelmaking process, and in powder metallurgy and the briquetting industries.

A schematic diagram of the H-iron process[17] is shown in Fig. 10. The primary chemical reaction is

$$Fe_2O_3(s) + 3H_2(g) \rightarrow 2Fe(s) + 3H_2O(g) \quad (32)$$

Iron ore is charged to the top and hydrogen is fed to the bottom of a three-stage fluidized-bed reactor where the ore is reduced to elemental iron at about 500°C. Only about 5% of the hydrogen is utilized in the reaction, so that unreacted hydrogen is recycled after removal of water. The iron product, which is pyrophoric as it emerges from the reactor, is treated with nitrogen at about 850°C for safe storage.

Catalytic Cracking Catalyst Regeneration

Catalytic cracking is widely used in modern petroleum refineries to increase the yield of gasoline by cracking heavy gas oil. Coke, which destroys the catalyst activity, is rapidly deposited on the catalyst during the cracking step. Burning of the coke in the regeneration reactor to restore catalytic activity is an important example of a noncatalytic gas–solid reaction.

Most catalytic cracking units operate in a recirculating fluidized-bed mode as shown in Fig. 7. The coked and regenerated catalyst are continually transferred between the cracking and regeneration reactors to provide steady-state operation. The regenerator can be operated to produce a gas containing both carbon monoxide and carbon dioxide or carbon dioxide alone. Regenerators typically operate in the temperature range of 650–800°C, and careful temperature control is required to prevent destroying the catalyst activity. The primary reactions are

$$C(s) + \frac{1}{2}O_2(g) \leftrightarrow CO(g) \quad (33)$$

$$C(s) + O_2(g) \leftrightarrow CO_2(g) \quad (34)$$

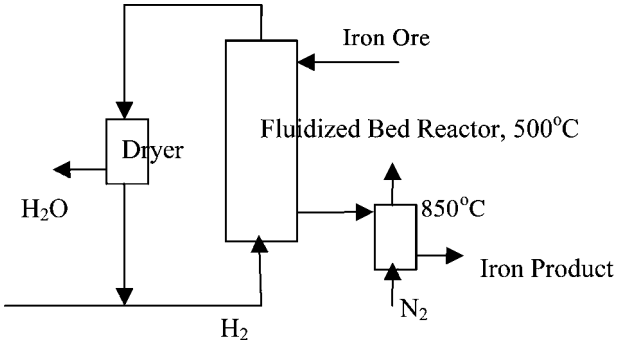


Fig. 10 The H-iron process.

Because the primary cracking reactions are endothermic and the catalyst regeneration reactions are exothermic, the energy associated with the high-temperature regenerated catalyst is used to vaporize gas oil entering the cracking reactor. When flue gas leaving the regenerator contains significant quantities of carbon monoxide, it is burned in a CO boiler to recover the available energy.

Desulfurization of Ammonia Synthesis Feed Gas

Natural gas or light hydrocarbons that serve as feed gas in the synthesis of ammonia contain sulfur compounds that act as poisons for the nickel catalyst used in synthesis gas production. Hydrogen sulfide and mercaptans are the dominant sulfur species in natural gas, while the light hydrocarbons may contain higher boiling sulfur species. A fixed-bed reactor containing zinc oxide is often used to desulfurize the feed gas. The chemical reaction with hydrogen sulfide is

$$ZnO(s) + H_2S(g) \leftrightarrow ZnS(s) + H_2O(g) \quad (35)$$

The reactor operates in the temperature range of 350–400°C and the hydrogen sulfide concentration is reduced from about 50 ppm to less than 1 ppmv in the product gas. Zinc sulfide is not normally regenerated, because regeneration is difficult and the low inlet sulfur concentrations make once-through use economical.

Considerable effort has been expended over the last few years to extend the use of zinc oxide to the desulfurization of coal-derived gases. Harrison[18,19] has described these efforts. The much higher inlet sulfur concentrations in this application require that the sorbent be regenerated so that it can be used for multiple cycles.

Semiconductor-Grade Silicon

High purity silicon is made by the direct decomposition of silane through the reaction

$$SiH_4(g) \leftrightarrow Si(s) + 2H_2(g) \quad (36)$$

Note that this is a special case where the only solid species is a reaction product.

In the process shown in Fig. 11,[20] hydrogen and silane are fed to the bottom of a fluidized-bed reactor, while high purity silicon seed crystals are charged to the top. Hydrogen utilization is quite low so that most of the hydrogen is recycled. The reactor operates at 600–800°C and slightly above atmospheric pressure. The desired decomposition reaction occurs on the surface of the seed crystals and results in particle

size growth. Homogeneous decomposition that occurs in the gas phase produces submicron to 10-μm size particles that tend to be elutriated from the reactor. Product particles in the 700–1000 μm size range are removed from the bottom of the reactor where they are cooled and transferred to a product hopper. Special silicon-lined vessels and piping are required to achieve the high purity required for the semiconductor applications.

Flue Gas Desulfurization

While most commercial flue gas desulfurization processes use wet scrubbing technology, there are economic incentives to develop dry sorption-based processes. The dry processes offer potential for lower capital and operating costs, but are limited primarily by low sorbent utilization.[21]

All dry processes under development use a calcium-based sorbent to react with SO_2 to form $CaSO_3$ and/or $CaSO_4$. All processes are ''once-through'' and the sulfated sorbent is normally land filled along with coal ash. The processes are classified as high-, medium-, and low temperature.

In the high-temperature process, either $CaCO_3$, $Ca(OH)_2$, or CaO is injected directly into the furnace where reaction occurs at 1000–1300°C. Both $CaCO_3$ and $Ca(OH)_2$ decompose to CaO at these temperatures so that the actual desulfurization reaction is

$$CaO(s) + SO_2(g) + \frac{1}{2}O_2(g) \leftrightarrow CaSO_4(s) \qquad (37)$$

The residence time at these temperatures is only about 100 ms so that the reaction must be quite fast if efficient SO_2 capture is to be achieved. Furnace injection is capable of achieving 40–50% SO_2 capture at a molar ratio of Ca/S of 1, and 65–75% capture with Ca/S equal to 2. Sorbent conversion is often limited to about 25% or less because of pore mouth plugging of the particles and sintering at the high temperatures.

In medium-temperature desulfurization, the sorbent is injected into the economizer section of the boiler and reaction occurs at 500–600°C. This temperature is insufficient for $CaCO_3$ calcination to occur so that the feed is limited to $Ca(OH)_2$ or CaO. Finally, the low-temperature process involves sorbent injection at some point in the system where the temperature is below about 350°C. At this low temperature, it is necessary to inject liquid water with the sorbent, either $Ca(OH)_2$ or CaO, to promote the reaction.

In all options, the sorbent is collected along with fly ash in the electrostatic precipitator or bag house, thereby adding to the particulate load on these units.

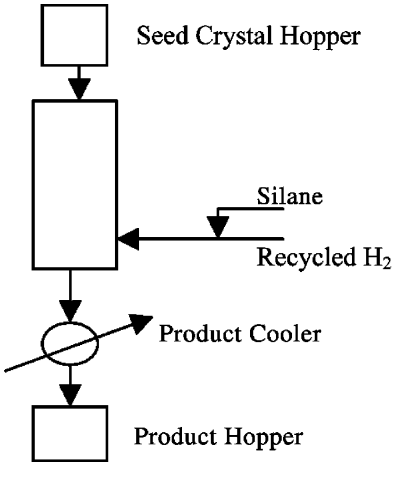


Fig. 11 Fluidized-bed process for silane manufacture.

CONCLUSIONS

The basic framework for modeling noncatalytic gas–solid reactions was established in the 1960–1980 time period. Models include those based on the effects of resistances associated with mass transfer from bulk gas to solid surface, diffusion within the solid pores and through the solid product layer, and the surface chemical reaction. Solid structural property variations associated with reaction and/or solid sintering have been included in the models. The single-pellet models are incorporated into integral models for the design of fixed-bed, fluidized-bed, and moving-bed reactors. Commercial applications of noncatalytic gas–solid reactions include the combustion and gasification of solid fuels, mineral processing, catalytic cracking catalyst regeneration, removal of pollutants from gas streams, and the production of silicon for semiconductor applications.

NOMENCLATURE

a_v	Gas–solid interfacial area per unit bed volume, l^2/l^3
A	Gas reactant
b	Stoichiometric coefficient of the reacting solid, mol B/mol A ($b < 0$)
B	Solid reactant
c	Stoichiometric coefficient of the product gas, mol C/mol A ($c > 0$)
C	Gas product
C_A	Molar concentration of reactant gas A, mol A/l^3; C_{Ab} in the bubble; C_{Ac}, at the interface of the unreacted core; C_{Af}, in reactor gas product; C_{Ai}, in reactor gas feed; C_{Ao}, in the bulk gas; C_{As}, at the external surface of the solid

C_S Molar concentration of reactant solid, mol B/l^3; C_{Sf}, in the fluidized-bed reactor product; C_{So}, initial concentration in the pellet; C'_{So}, initial concentration in the grain
d Stoichiometric coefficient of product solid, mol D/mol A ($d > 0$)
D Solid product
D_{eA} Effective diffusivity of A through the pellet solid product layer, l^2/t
D'_{eA} Effective diffusivity of A through the grain solid product layer, l^2/t
f Volume fraction of bubble phase
k_{bc} Mass transfer coefficient between bubble and cloud, t^{-1}
k_{ce} Mass transfer coefficient between cloud and emulsion, t^{-1}
k_i Reaction rate constant in fluidized-bed reactor, l^6/m mol B t
k_{mA} Mass transfer coefficient of A from bulk gas to pellet surface, l/t
k_s Reaction rate constant based on the surface area of the pellet unreacted core, l^4/mol · t
k'_s Reaction rate constant based on the surface area of the grain unreacted core, l^4/mol · t
K_r Pseudo first-order rate constant in the fluidized-bed reactor, defined by Eq. (27), t^{-1}
L Reactor length, l
Q_{So} Solid volumetric feed rate, l^3/t
r Radial coordinate, l
r_c Radius of the pellet unreacted core, l
r'_c Radius of the grain unreacted core, l
r_s Pellet radius, l
r'_s Grain radius, l
R_A Molar rate of formation of A per unit surface area, mol A/l^2 · t
R^v_A Molar rate of formation of A per unit bed volume, mol A/l^3 · t
t Reaction time, t
U_b Bubble rise velocity in the fluidized bed, l/t
U_o Superficial gas velocity in the fixed-bed and moving-bed reactors, l/t
U_S Superficial solid velocity in the moving-bed reactor, l/t
V_b Bubble volume, l^3
V_{iz} Cloud volume, l^3
V_o Volumetric gas feed rate, l^3/t
x Fractional conversion of solid reactant
z Axial coordinate in the reactor, l
ε_v Bed porosity
ρ_b Mass solid density in bubble, m/l^3
ρ_c Mass solid density in cloud phase, m/l^3
ρ_e Mass solid density in emulsion, m/l^3

REFERENCES

1. Szekely, J.; Evans, J.; Sohn, H. *Gas-Solid Reactions*; Academic Press: New York, 1976.
2. Weisz, P.; Goodwin, R. Combustion of carbonaceous deposits within porous catalysts. I. Diffusion controlled kinetics. J. Catal. **1963**, *2*, 397–404.
3. Beveridge, G. The oxidation rates of zinc sulfide spheres. In *Agglomeration*; Knepper, W., Ed.; Interscience Publishers: New York, 1962; 303–349.
4. Costa, E.; Smith, J. Kinetics of noncatalytic, nonisothermal gas–solid reactions: hydrofluorination of uranium dioxide. AIChE J. **1971**, *17*, 947–958.
5. Themelis, N.; Yannopoulos, J. Mass- and heat-transfer phenomena in the reduction of cupric oxide by hydrogen. Trans. Metall. Soc. AIME **1966**, *236*, 414–420.
6. Luss, D.; Amundson, N. Maximum temperature rise in gas–solid reactions. AIChE J. **1969**, *15*, 194–199.
7. Szekely, J.; Sohn, H. The effect of intragrain diffusion on the reaction between a porous solid and a gas. Chem. Eng. Sci. **1974**, *29*, 630–634.
8. Szekely, J.; Evans, J. A Structural model for gas–solid reactions with a moving boundary. II. The effect of grain size, porosity, and temperature on the reaction of porous pellets. Chem. Eng. Sci. **1971**, *26*, 1901–1913.
9. Gibson, J.; Harrison, D. The reaction between hydrogen sulfide and spherical particles of zinc oxide. Ind. Eng. Chem. Proc. Des. Dev. **1980**, *19*, 231–237.
10. Ramachandran, P.; Smith, J. A single-pore model for gas–solid noncatalytic reactions. AIChE J. **1977**, *23*, 353–361.
11. Hartman, M.; Coughlin, R. Reaction of sulfur dioxide and the grain model. AIChE J. **1976**, *13*, 490–498.
12. Ramachandran, P.; Smith, J. Effect of sintering and porosity changes on rates of gas–solid reactions. Chem. Eng. J. (Lausanne) **1977**, *14*, 137–146.
13. Ranade, P.; Harrison, D. The grain model applied to porous solids with varying structural properties. Chem. Eng. Sci. **1979**, *34*, 427–432.
14. Wang, J.; Groves, F.; Harrison, D. Modeling high temperature desulfurization in a fixed-bed reactor. Chem. Eng. Sci. **1990**, *45*, 1693–1701.
15. Kunii, D.; Levenspiel, O. *Fluidization Engineering*; Wiley: New York, 1969.
16. Matsui, I.; Kojimas, T.; Furusawa, T.; Kunii, D. Gasification of coal char in a continuous fluidized bed reactor. In *Fluidization*; Kunii, D., Toei, R.,

Eds.; Engineering Foundation: New York, 1984; 655–662.
17. Gupta, C.; Sathiyamoorthy, D. *Fluid Bed Technology in Materials Processing*; CRC Press: Boca Raton, 1999.
18. Harrison, D. Performance analysis of ZnO-based sorbents in removal of H_2S from fuel gas. In *Desulfurization of Hot Coal Gas*; Atimtay, A., Harrison, D., Eds.; Springer: Berlin, 1998; 213–242.
19. Harrison, D. Regeneration of sulfided sorbents and direct production of elemental sulfur. In *Desulfurization of Hot Coal Gas*; Atimtay, A., Harrison, D., Eds.; Springer: Berlin, 1998; 331–364.
20. Jazayeri, B. Applications for chemical production and processing. In *Handbook of Fluidization and Fluid-Particle Systems*; Yang, W., Ed.; Dekker: New York, 2003; 421–444.
21. Kakmi, J.; Adler, R.; Prudich, M.; Fan, L.; Raghunathan, K.; Khang, S.; Keener, T. Flue gas desulfurization for acid gas control. In *Dry Scrubbing Technologies for Flue Gas Desulfurization*; Toole-Oneil, B., Ed.; Kluwer Academic Publishers: Boston, 1998; 1–113.

Gas-to-Liquid Mass Transfer

Huu D. Doan
Simant R. Upreti
Ali Lohi
Department of Chemical Engineering, Ryerson University, Toronto, Ontario, Canada

INTRODUCTION

Gas-to-liquid mass transfer is a transport phenomenon that involves the transfer of a component (or multiple components) between gas and liquid phases. Gas–liquid contactors, such as gas–liquid absorption/stripping columns, gas–liquid–solid fluidized beds, air-lift reactors, gas bubble reactors, and trickle-bed reactors (TBRs) are frequently encountered in chemical industry. Gas-to-liquid mass transfer is also applied in environmental control systems, e.g., aeration in wastewater treatment where oxygen is transferred from air to water, trickle-bed filters, and scrubbers for the removal of volatile organic compounds. In addition, gas-to-liquid mass transfer is an important factor in gas–liquid emulsion polymerization, and the rate of polymerization could, thus, be enhanced significantly by mechanical agitation.

The objective of this entry is to introduce the readers to the fundamental principles of gas-to-liquid mass transfer, as well as its major applications. Therefore, the first section of the entry is on the three fundamental mechanisms of gas-to-liquid mass transfer: the film theory, the penetration theory, and the surface renewal theory followed by the applications of gas-to-liquid mass transfer in unit operations that are widely used in various chemical processes. There is a vast pool of reported literature on different aspects of gas-to-liquid mass transfer processes, all of which is impossible to be included in this entry. Therefore, only typical gas-to-liquid mass transfer processes are presented here.

PRINCIPLES OF GAS-TO-LIQUID MASS TRANSFER

Gas-to-liquid mass transfer can take place from a gas phase to a liquid phase (and vice versa) with or without chemical reactions. The concentration gradient of a transferred component in the bulk fluid and in the fluid at the interface is the driving force for mass transfer. When the mass transport occurs in a phase that is moving, the transport of the component is known as convective mass transfer. Convective mass transfer is usually expressed in a standard flux equation as:

$$N_A = k_c(C_{Ab} - C_{Ai}) \tag{1}$$

where N_A is the molar flux of component A from gas to liquid, C_{Ab} and C_{Ai} are the concentrations of component A in the bulk gas and in the gas phase at gas–liquid interface, respectively, and k_c is the convective mass transfer coefficient.

Eq. (1) is applicable to both pure diffusion and convective transfer in a laminar or turbulent flow. For a binary system, the total molar flux, which takes into account mass transfer by both molecular diffusion and convection because of bulk flow, can be expressed as:

$$N_A = \frac{k'_c}{\alpha}(C_{Ab} - C_{Ai}) \tag{2}$$

where α is the correction factor for the bulk flow, which can be written as:

$$\alpha = \frac{\left(\frac{N_A}{N_A + N_B} - \frac{C_{Ai}}{C}\right) - \left(\frac{N_A}{N_A + N_B} - \frac{C_{Ab}}{C}\right)}{[N_A/(N_A + N_B)]\ln\{[N_A/(N_A + N_B) - C_{Ai}/C]/[N_A/(N_A + N_B) - C_{Ab}/C]\}} \tag{3}$$

The value of α varies with the system under consideration. For example, in equimolar counter diffusion, N_A and N_B are of the same magnitude, but in opposite direction. As a result, α is equal to 1; and hence, Eq. (2) reduces to Eq. (1), where k'_c is equal to k_c.

Convective mass transfer coefficients are used in the design of mass transfer equipment. However, in most cases, these coefficients are extracted from empirical correlations that are determined from experimental data. The theories, which are often used to describe the mechanism of convective mass transfer, are the film theory, the penetration theory, and the surface renewal theory.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007787

FILM THEORY IN MASS TRANSFER

Proposed by Whitman,[1] the film theory is based on the assumption that for a flowing fluid, there is a fictitious stagnant fluid film at the phase boundary, in which the entire resistance to mass transfer resides and the mass transport is, thus, completely by molecular diffusion. Therefore, the mass transfer coefficient is proportional to the ratio of the diffusivity to the thickness of the fictitious film.

Dependent upon the mode of mass diffusion, i.e., diffusion through a nondiffusing layer, counter-diffusion, or equimolar counter-diffusion, the relationship between the mass transfer coefficient and the diffusivity may take different forms.[2] For example, under the condition of equimolar counter diffusion for a binary system, the molar flux in x-direction can be expressed in terms of the diffusivity as:

$$N_A = D\frac{dC_A}{dx} = \frac{D}{\delta}(C_{Ab} - C_{Ai}) \quad (4)$$

From Eqs. (1) and (4), the relationship of the mass transfer coefficient k_c, with the diffusivity D, and the thickness of the fictitious film δ can be written as:

$$k_c = \frac{D}{\delta} \quad (5)$$

The thickness of the fictitious film can neither be predicted nor measured experimentally. This limits the use of the film theory to directly calculate the mass transfer coefficients from the diffusivity. Nevertheless, the film theory is often applied in a two-resistance model to describe the interphase mass transfer between the two contacting phases (gas and liquid). This model assumes that the resistance to mass transfer only exists in gas and liquid films. The interfacial concentrations in gas and liquid are in equilibrium. The interphase mass transfer involves the transfer of mass from the bulk of one phase to the interfacial surface, the transfer across the interfacial surface into the second phase, and the transfer of mass from the interface to the bulk of the second phase. This process is described graphically in Fig. 1.

The molar flux of a component A can be written for individual phases. For the gas film

$$N_A = k_G(P_{AG} - P_{Ai}) = k_y(y_{AG} - y_{Ai}) \quad (6)$$

where k_G is the mass transfer coefficient in the gas film, P_{AG} the partial pressure of A in the bulk gas, and P_{Ai} the partial pressure of A in gas at the gas–liquid interface. k_y is the mass transfer coefficient defined in terms of the mole fraction of the transferred species, y_{AG} the mole fraction of A in the bulk gas, and y_{Ai} the mole fraction of A in gas at the gas–liquid interface.

For the liquid film

$$N_A = k_L(C_{Ai} - C_{AL}) = k_x(x_{Ai} - x_{AL}) \quad (7)$$

where k_L is the mass transfer coefficient in liquid film, C_{Ai} is the concentration of A in liquid at the gas–liquid interface, and C_{AL} is the concentration of A in the bulk liquid. k_x is the mass transfer coefficient in terms of the mole fraction of A in liquid, x_{Ai} the mole fraction of A in liquid at the gas–liquid interface, and x_{AL} the mole fraction of A in the bulk liquid.

The inability to measure interfacial concentrations limits the use of film molar flux equations to determine film mass transfer coefficients. To overcome this shortcoming, the concept of the overall mass transfer coefficient is used. The molar flux equation can be rewritten for the two phases as below.

For the gas phase:

$$N_A = K_y(y_{AG} - y_A^*) \quad (8)$$

where K_y is the overall gas-phase mass transfer coefficient, and y_A^* is the mole fraction of A in the gas phase in equilibrium with the concentration of A in the bulk liquid.

For the liquid phase:

$$N_A = K_x(x_A^* - x_{AL}) \quad (9)$$

where K_x is the overall liquid-phase mass transfer coefficient, and x_A^* is the mole fraction of A in the liquid phase in equilibrium with the concentration of A in the bulk gas.

By combining and rearranging Eqs. (6)–(9), the relationship between the overall mass transfer coefficients

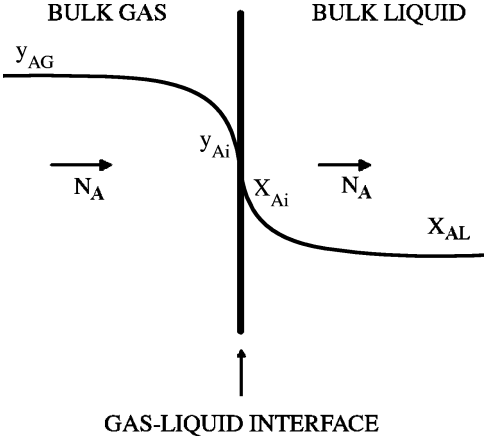


Fig. 1 Solute concentration profile for two-resistance model.

and the film mass transfer coefficients can be written as:

$$\frac{1}{K_y} = \frac{1}{k_y} + \frac{m'}{k_x} \tag{10}$$

$$\frac{1}{K_x} = \frac{1}{k_x} + \frac{1}{m''k_y} \tag{11}$$

where the slopes of the equilibrium curve m' and m'' (see Fig. 2) can be expressed as:

$$m' = \frac{y_{Ai} - y_A^*}{x_{Ai} - x_{AL}} \tag{12}$$

$$m'' = \frac{y_{AG} - y_{Ai}}{x_A^* - x_{Ai}} \tag{13}$$

In a gas–liquid contactor, the film mass transfer coefficient is independent of the solute concentration unless it changes significantly with position in the contactor such that the physical properties of the fluid are changed. On the other hand, as indicated in Eqs. (10) and (11), the overall mass transfer coefficients depend on both the film mass transfer coefficients and the slope of the equilibrium curve over the concentration range of interest. Therefore, experimentally obtained overall mass transfer coefficients should be used for the design of a gas–liquid contactor, only with conditions similar to those under which they were measured. However, when the equilibrium line is straight, m' and m'' are constant, the overall mass transfer coefficients then become independent of the solute concentration.

The film mass transfer coefficient can be extracted from the overall mass transfer coefficient when the experiments are carefully designed, so that the gas–liquid mass transfer is predominantly controlled by either the gas or liquid phase. The gas-phase-controlled mass transfer occurs when the solute is highly soluble in the liquid phase. Almost entire resistance to the mass transfer is within the gas phase and the slope m' is very small. Assuming that the magnitude of the liquid film mass transfer coefficient k_x is comparable to that of the gas film mass transfer coefficient k_y, the overall gas-phase mass transfer coefficient approximates the gas film mass transfer coefficient because the term m'/k_x in Eq. (10) becomes nonsignificant when compared to $1/k_y$. On the other hand, when the solute is only slightly soluble in the liquid phase, almost entire resistance to mass transfer resides in the liquid phase, and the mass transfer process is under the liquid-phase-controlled condition. The slope m'' is large; hence the overall liquid-phase mass transfer coefficient is approximately equal to the liquid film mass transfer coefficient, as shown by Eq. (11). For cases with a significant difference between k_x and k_y, the magnitudes of both the ratio k_x/k_y and m' or m'' must be considered in determination of the controlling phase.

The determination of the controlling phase in a gas–liquid mass transfer operation is of help in identifying the suitable operational parameters that can be manipulated to improve the performance of the operation. For a gas-phase-controlled process, mass transfer can be improved by increasing the gas velocity. A higher velocity causes an increase in the concentration gradient of the solute in the gaseous film; hence a higher mass transfer coefficient is achieved. Similarly, in a liquid-phase-controlled process, increasing liquid flow rate enhances the mass transfer coefficient, thereby improving the mass transfer rate. In addition, for a given solute concentration in the gas phase, when the liquid temperature is decreased, the solubility of solute in liquid increases; hence the equilibrium concentration of the solute in the liquid phase (at the gas–liquid interface) increases. As a result, the concentration gradient of the solute in the liquid film increases and so does the rate of mass transfer.

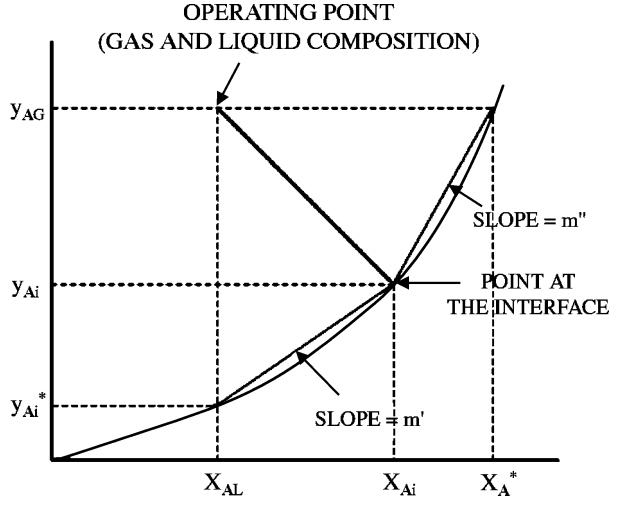


Fig. 2 Equilibrium curve and its slopes—two-resistance model.

PENETRATION THEORY IN MASS TRANSFER

The thickness of the fictitious film in the film theory can never be measured. The film theory predicts that the convective mass transfer coefficient k_c is directly proportional to the diffusivity whereas experimental data from various studies show that k_c is proportional to the two-third exponent of the diffusivity. In addition, the concept of a stagnant film is unrealistic for a fluid–fluid interface that tends to be unstable. Therefore, the penetration theory was proposed by Higbie[3] to better describe the mass transfer in the liquid phase

during gas absorption. In this theory, the liquid surface is assumed to consist of small fluid elements that contact the gas phase for an average time, after which they penetrate into the bulk liquid. Each element is then replaced by another element from the bulk liquid. The mass transfer into the liquid phase is under unsteady-state diffusion with zero concentration at an infinite depth in the bulk liquid. Thus, Fick's second law of diffusion in one dimension can describe this transfer process, from which the average molar flux can be obtained:

$$N_A|_{avg} = 2\sqrt{\frac{D}{\pi t_{exp}}}(C_{Ai} - C_{Ab}) \quad (14)$$

where t_{exp} is the exposure time of the fluid elements. Eq. (14) indicates that the mass transfer coefficient is proportional to the square root of the diffusivity.

One of the boundary conditions of the penetration theory is that the concentration at an infinite distance from the interface is equal to zero. This restricts the applicability of the penetration model to systems with a sufficiently large bulk of liquid. In a gas–liquid contactor, such as a packed column, there only exists a thin layer of liquid flowing over packing particles. Hence, the boundary conditions of the penetration theory are modified to have zero solute concentration at the solid surface of the packing.[4]

Although the penetration theory better describes the gas–liquid mass transfer than the film theory, its advantage is only significant with physical absorption. For gas absorption with fast chemical reactions in the liquid phase, the entire mass transfer process is gas-phase-controlled rendering the penetration theory inapplicable.[5]

SURFACE RENEWAL THEORY

The exposure times of all liquid elements are usually not the same as assumed in the penetration model. To address this matter, Danckwerts proposed the surface renewal theory. In this theory, an average exposure time, which is determined from an assumed time distribution, is used in place of the constant exposure time in the penetration theory.[6] The average molar flux is given as below:

$$N_A|_{avg} = \sqrt{\frac{D}{\pi}}(C_{Ai} - C_{Ab}) \int_0^\infty \frac{\tau(t)}{t^{0.5}} dt \quad (15)$$

where $\tau(t)\,dt$ represents the fraction of the liquid surface that consists of liquid elements with an exposure time span from t to $(t + dt)$. Therefore, the sum of the fractions is equal to unity:

$$\int_0^\infty \tau(t) dt = 1.0 \quad (16)$$

Dankwerts proposed the exposure time distribution function as:

$$-\frac{d\tau}{dt} = s\tau \quad (17)$$

where s is the surface renewal rate, which is inversely proportional to the exposure time of the liquid elements.

By integrating Eq. (17), substituting the resulting expression for τ in terms of s into Eq. (16), and integrating the resultant equation, the following relationship is obtained:

$$\tau = s e^{-st} \quad (18)$$

and the average molar flux is then given by:

$$N_A|_{avg} = \sqrt{\frac{D}{\pi}}(C_{Ai} - C_{Ab}) \int_0^\infty \frac{s e^{-st}}{t^{0.5}} dt = \sqrt{sD}(C_{Ai} - C_{Ab}) \quad (19)$$

The values of s are generally not known. However, they can be determined from the experimental data of the mass transfer coefficients for various gas–liquid systems.

APPLICATIONS OF GAS-TO-LIQUID MASS TRANSFER PROCESS

Gas–liquid mass transfer is applied in a wide variety of processes in the chemical industry, biochemical processes and wastewater treatment systems. The following widely used processes are reviewed:

- Gas absorption in a packed column.
- Gas–liquid mass transfer in a three-phase fluidized bed reactor.
- Gas–liquid mass transfer in an airlift reactor.
- Liquid–gas bubble reactors.
- Trickle-bed reactors.

Distillation is also a common transport process involving gas-to-liquid mass transfer.

GAS ABSORPTION IN A PACKED COLUMN

Packed columns have been used in chemical industry for more than a century. The packing in a packed column provides a large surface area over which the gas contacts the liquid and the gas-to-liquid (and vice versa) mass transfer occurs (Fig. 3). As a result, the gas–liquid mass transfer process in a packed column is very efficient. Research and development on packed column has received much attention. One of the earlier and original literatures reported by Leva[7] covers the principles of gas–liquid mass transfer in a packed bed, as well as packed column design and column internals. Recent reported literatures on gas absorption in a packed column from Fair et al.,[5] Strigle,[8] and Billet[9] are excellent references on transport phenomena in a packed column and packed column design.

The performance of a packed column is determined by the ultimate contact of gas and liquid in the column. The effective surface area of the packing (hydraulic area or wetted area), where mass transfer may occur, is usually lower than the physical surface of area of the packing (specific area, a). Liquid distribution affects the effective surface area in a packed bed. Several research groups have measured liquid distribution for various packing heights, packing materials and sizes, and liquid distributors.[10,11] The quality of liquid distribution is dependent on the initial liquid distribution (from the liquid distributor), liquid flow rate, packing type and size, and packing height. Poor initial liquid distribution because of a small number of liquid nozzles on a liquid distributor or a poor liquid distributor design causes dry packing zones in the top section of the packed bed. This is reflected by a higher mass transfer variation index, which is defined as an average deviation of local mass transfer coefficients in the packed bed, for a single-point distributor as compared to that for a multiple-point liquid distributor.[12] Smaller packing has a tendency to spread out liquid quicker; hence, liquid distribution with smaller packing is better than that with larger packing. It is a common practice to use the packing of a size less than one-tenth of the tower diameter, to avoid severe wall flow. For most industrial applications of packed columns, liquid flow rate is usually sufficiently high, such that liquid distribution does not change with the liquid flow rate significantly. Liquid distribution has also been shown to be independent of the gas flow rate below loading point. In general, liquid distribution is improved when the packing height is increased. For 2 in. metallic Pall rings, liquid distribution is improved with increases in the ratio of packing height to tower diameter, x/D, up to 5. For x/D in the range from 5 to 6.5, liquid distribution reaches a steady pattern.[13] Nevertheless, it has been reported that when the packing height is increased to about 20 ft, liquid maldistribution may occur; hence the use of liquid redistributors is recommended.[14]

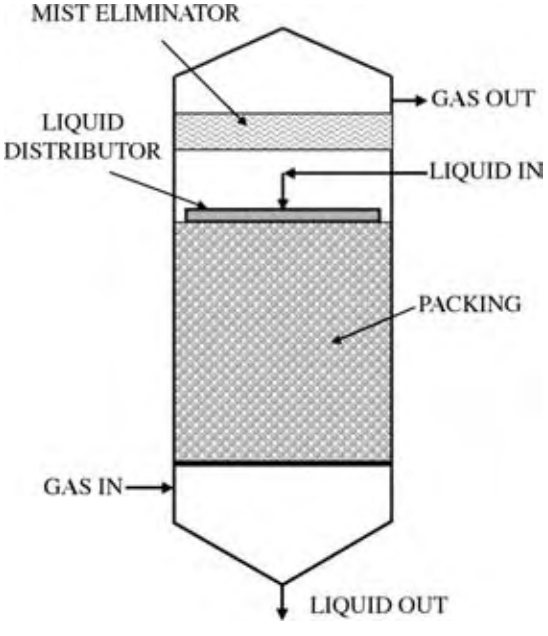


Fig. 3 Schematic diagram of a packed column.

The relationship between the specific area a and the hydraulic area a_h of several packing types was studied and the following correlations were reported:[9]

$$\frac{a_h}{a} = C_h \mathrm{Re}_L^{0.15} \mathrm{Fr}_L^{0.1} \quad \text{for } \mathrm{Re}_L < 5 \tag{20}$$

$$\frac{a_h}{a} = 0.85 C_h \mathrm{Re}_L^{0.25} \mathrm{Fr}_L^{0.1} \quad \text{for } \mathrm{Re}_L \geq 5 \tag{21}$$

where C_h is the packing constant as shown in Table 1, Re_L the liquid Reynolds number that is defined as $(u_L \rho_L)/(\mu_L a)$, Fr_L the Froude number that is defined as $(u_L{}^2 a)/g$, a the specific area of packing, g the gravitational acceleration, u_L the liquid superficial velocity, ρ_L the liquid density, and μ_L the liquid viscosity.

However, there is significant deviation among the values of a_h estimated from different correlations. Therefore, the mass transfer coefficients, k_L and k_G, are often lumped with the specific area a in most reported literatures. Published data have often been obtained from investigating mass transfer in small columns (less than 0.254 m diameter) in which gas flow patterns may be different from those in a large industrial-scale column.[15] In addition, the entrance section below the packing support can be accounted for about 17% of the overall mass transfer over the entire packed column. This may vary with different mechanical designs of packed columns.[16] The uncertainty in scaling up laboratory and pilot-plant data to large towers is significant. Therefore, mass transfer

Table 1 Packing factor, C_h, for some common random packing

Packing	Material	Size (mm)	a (m^2/m^3)	C_h
Pall ring	Metal	50	112.6	0.784
		35	139.4	0.644
		25	215.0	0.719
		15	368.4	0.590
	Plastic	50	102.0	0.593
		35	148.3	0.718
		25	225.0	0.528
	Ceramic	50	121	0.1335
Raschig ring	Metal	15	378.4	0.455
	Ceramic	25	191.9	0.577
		15	310.2	0.648
		10	492.1	0.791
	Carbon	25	205.7	0.623
Berl saddle	Ceramic	25	205.4	0.620
		13	436.4	0.833
Tellerette	Plastic	25	180.0	0.588
Hiflow ring	Metal	50	92.3	0.876
		25	197.5	0.799
	Plastic	50	120.2	1.039
	Ceramic	20, 4 ribs	239.3	1.167

(From Ref.[9].)

coefficient correlations available in literature should be used with care and reservation. There are literally hundreds of equations. However, the most widely used correlation to predict $k_L a$ is still that developed by Sherwood and Holloway[17] as below:

$$\frac{k_L a}{D} = \alpha\left(\frac{L}{\mu_L}\right)^{1-n}\left(\frac{\mu_L}{\rho_L D}\right)^{0.5} \quad (22)$$

where D is the diffusivity of the solute in liquid, and L is the liquid mass flux. The constants α and n for various types of packing are given in Table 2. Au-Yeung and Ponter[18] reported a comprehensive review of empirical and theoretical correlations for $k_L a$, published since 1940.

GAS–LIQUID MASS TRANSFER IN THREE-PHASE FLUIDIZED BED REACTORS

Three-phase fluidized bed reactors are widely used in petrochemical, metallurgical, environmental, and coal liquefaction processes. In some cases, such as liquid-phase methanol process, these reactors are also referred to as ebulliated reactors with suspended catalyst particles in inert liquid. Three-phase catalytic reaction is also applied to liquid-entrained reactors where a slurry of small catalyst particles and inert liquid is pumped to the reaction zone, where gas is fed separately. Liquid-entrained reactors alleviate catalyst attrition, which often occurs in a three-phase fluidized bed reactor, especially under high solid loading.[19,20] In three-phase fluidized bed reactors, solid particles catalyze reactions involving species in gas and liquid phases. The particles are fluidized by the upward flow of these phases (Fig. 4). Gas bubbles are formed and dispersed in the continuous liquid phase.[21] The gas–liquid mass transfer is affected by the presence of solid particles, and is controlled by the liquid phase. In general, the volumetric gas–liquid mass transfer coefficient increases with gas as well as liquid velocity, particle size, and density. On the other hand, the transfer coefficient decreases with the surface tension and viscosity of liquid.

Table 2 Constants α and n for use in the Sherwood and Holloway correlation

Packing	Size (in.)	α	n
Rings	2.0	341	0.22
	1.5	383	0.22
	1.0	426	0.22
	0.5	1392	0.35
	3/8	3116	0.46
Saddles	1.5	732	0.28
	1.0	778	0.28
	0.5	686	0.28
Tile	3.0	503	0.28

(From Ref.[17].)

The mass transfer coefficient increases with gas velocity because the gas–liquid interfacial area and the mixing of solid particles increase with gas velocity. However, skewed bubble-size distribution occurs when gas velocity is increased beyond a certain value, and the interfacial area does not increase further. The solid mixing also approaches a constant level. The transfer coefficient has been reported to be proportional to gas velocity to an exponent in the range of 0.44–0.98.[22]

The increase in liquid velocity enhances turbulence that helps in breaking up gas bubbles and creating more smaller-sized gas bubbles by solid particles. This leads to increases in the gas–liquid interfacial area, thereby increasing the mass transfer rate and the volumetric mass transfer coefficient. On the other hand, the holdup of solid particles and bubbles decreases with increases in liquid velocity, which has a negative effect on gas–liquid mass transfer. Beyond a critical value,

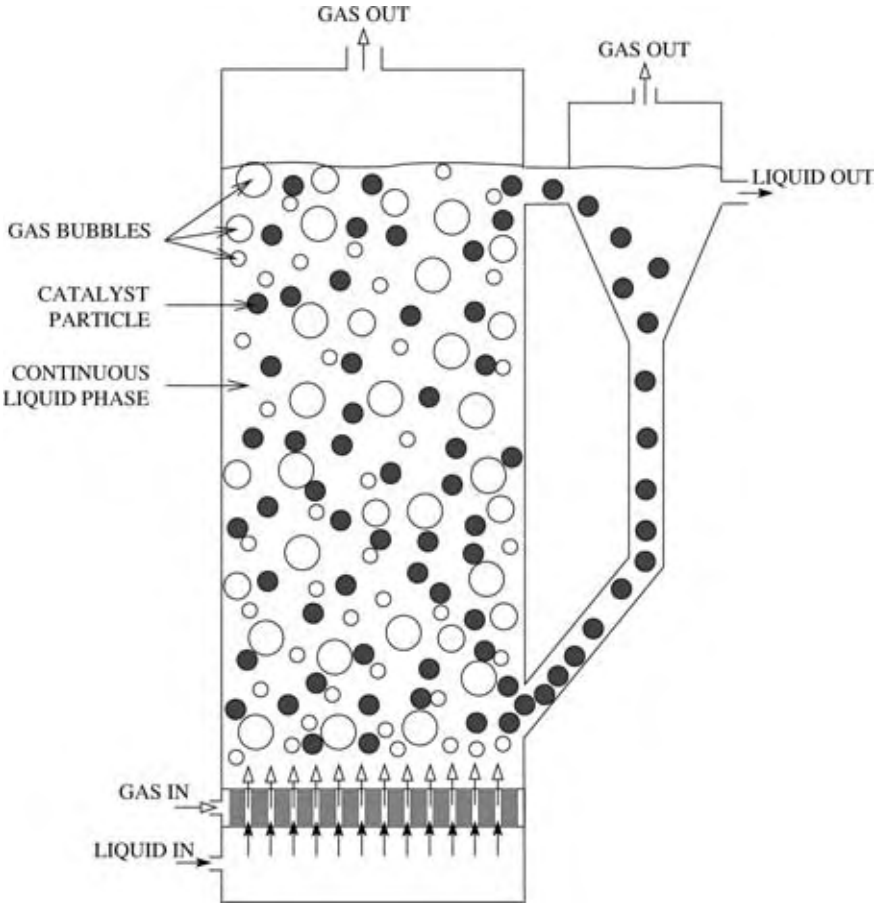


Fig. 4 Schematic diagram of a three-phase fluidized bed.

the transfer coefficient begins to decrease with increases in liquid velocity. The transfer coefficient has been reported to be proportional to liquid velocity to an exponent in the range of 0.42–0.97.[22]

The effect of particle size and particle density on the volumetric gas–liquid mass transfer coefficient is more complicated. Although larger particles promote the breakage of large gas bubbles into small bubbles, a higher minimum fluidization velocity is required, resulting in the reduction of gas and solid holdups and changes of bubble regime. The transfer coefficient increases with particle size, as long as there are enough holdups of gas bubbles and solid particles in the bubble disintegration regime. Nevertheless, for porous solid that is used to provide higher surface area for the catalytic reaction in a three-phase fluidized bed reactor, larger particles may have a negative effect on the reaction rate because intraparticle diffusion is usually slower because of long porous paths of larger particles when compared to smaller particles. The effect of particle density in this regime is insignificant. However, in bubble coalescence regime, the transfer coefficient has been found to increase with decreases in particle density because lighter particles circulate more vigorously, and break up more gas bubbles.

The increase in liquid viscosity reduces the gas holdup and restricts the movement of gas bubbles, thereby decreasing the mass transfer coefficient. The transfer coefficient has been reported to be inversely proportional to the liquid viscosity to an exponent in the range of 0.34–0.67.[22] The transfer coefficient has also been found inversely proportional to the liquid surface tension to an exponent of 1.532 and 2.273 in the bubble disintegration and coalescence regimes, respectively because higher liquid surface tensions hinder bubble formation.[23]

Based on different experimental studies for superficial gas velocities from 0.005 to 0.19 m/s, superficial liquid velocities from 0.008 to 0.175 m/s, average solid particle diameters from 0.76 to 8 mm, and liquid viscosities from 0.001 to 0.0967 Pa s; Kim and Kang reported the following correlation for gas–liquid mass transfer in a three-phase fluidized bed:[22]

$$\mathrm{Sh} = 2.694\mathrm{Sc}^{1.3}\mathrm{Re}_{\mathrm{b}}^{0.196} \tag{23a}$$

The Sherwood number Sh in Eq. (23a) is defined as:

$$\mathrm{Sh} = \frac{(k_{\mathrm{L}}a)d_{\mathrm{b}}^2}{6\varepsilon_{\mathrm{G}}D} \tag{23b}$$

where $k_L a$ is volumetric gas–liquid mass transfer coefficient, d_b the mean bubble diameter, ε_G the gas phase holdup, and D the molecular diffusivity of gas in liquid. The Schmidt number Sc is defined in terms of kinematic viscosity of liquid, ν_L:

$$\mathrm{Sc} = \frac{\nu_L}{D} \tag{23c}$$

and the Reynolds number, Re_b, is defined as:

$$\mathrm{Re}_b = \frac{E d_b^4}{\nu_L^3} \tag{23d}$$

where E is the energy dissipation rate per unit mass of liquid phase.

GAS–LIQUID MASS TRANSFER IN AIRLIFT REACTORS

An airlift reactor is a column divided into two sections, a riser and a downcomer, interconnected at top and bottom (Fig. 5). Gas entering the bottom of the riser lifts the liquid and any accompanying solid particles (for a three-phase reactor) upward, which subsequently enter the top of the downcomer and then flow downward to the bottom of the riser again.[24] This circulation within the reactor enhances mixing and mass transfers without the use of mechanical agitators. While an internal-loop airlift reactor has the riser and the downcomer in the same column, an external-loop airlift reactor has separate tubes for the riser and the downcomer. Airlift reactors are used in multiphase chemical reactions and biotechnological processes.

Compared to bubble columns, airlift reactors have better liquid circulation but lower rates of mass transfer and mixing. These rates are enhanced in modified airlift reactors with perforated single or coaxial draft tubes.[25] This enhancement is because of the breakup of gas bubbles into smaller bubbles when crossing perforated tubes. The gas–liquid interfacial area and the gas–liquid mass transfer coefficient increases. Similar effect can be achieved with the addition of packing to the riser.[26]

The volumetric gas–liquid mass transfer coefficient in an airlift reactor increases with gas velocity and its holdup. Based on the penetration theory and the isotropic turbulence theory, the following theoretical

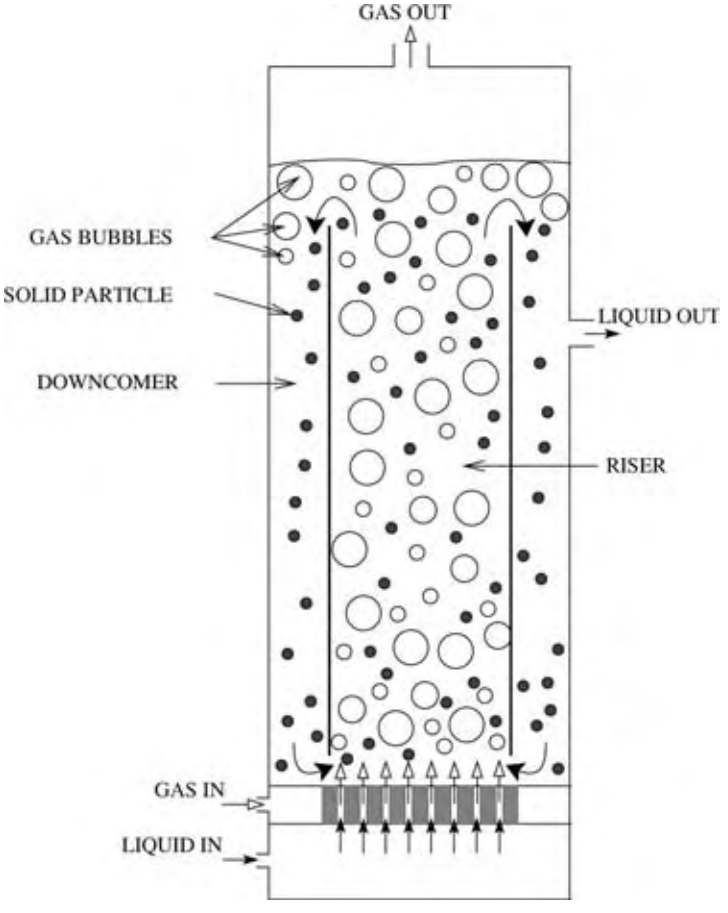


Fig. 5 Schematic diagram of an airlift reactor.

model was developed for air–water system:[27]

$$k_L a = 2\sqrt{\frac{D}{\pi}}\left(\frac{v_G \rho_L g \varepsilon_G}{\mu_L}\right)^{0.25} \frac{6\varepsilon_G}{d_B(1 - \varepsilon_G)} \quad (24)$$

where $k_L a$ is the volumetric gas–liquid mass transfer coefficient, D the diffusivity, v_G the slip gas velocity, g the gravity, ε_G the gas holdup, μ_L and ρ_L are the viscosity and density, respectively, of liquid, and d_B is the average diameter of gas bubbles.

However, for a given gas velocity, any change in gas or liquid properties, downcomer and riser geometry, phase separation conditions, liquid volume, reactor height, or gas distribution causes changes in liquid velocity and gas holdup. Therefore, no generalized model or correlation for the volumetric gas–liquid mass transfer coefficient in airlift reactors exists.

LIQUID–GAS BUBBLE REACTORS

Liquid–gas bubble reactors are divided into two main groups: gas–liquid bubble columns and gas–liquid continuous stirred reactors (Figs. 6 and 7). Gas–liquid bubble columns are operated under either of the two characteristic modes: particulate or homogeneous regime, and aggressive or heterogeneous regime. The rates of mass transfer and mixing are quite different in these two regimes. Agitation is done to enhance mass transfer in stirred reactors in a similar fashion as in mechanically agitated slurry reactors where the mass transfer from liquid or gas to solid particles increases with the mechanical agitation.

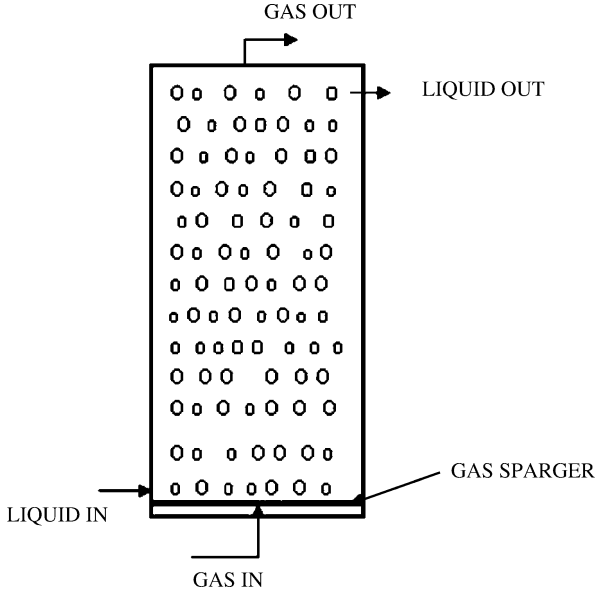


Fig. 6 Schematic diagram of a liquid–gas bubble column.

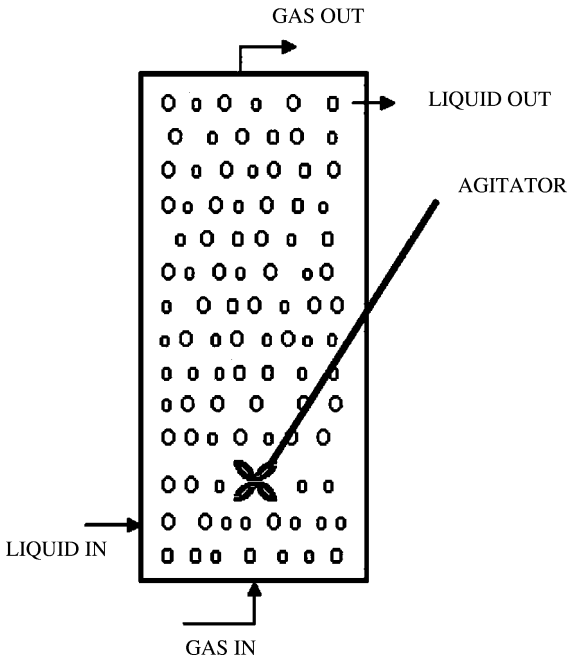


Fig. 7 Schematic diagram of a gas–liquid continuous stirred reactor.

In bubble columns, gas is dispersed in a continuous liquid phase. Uniform bubble size and bubble concentration characterize the homogeneous regime, particularly in the traverse direction indicating the absence of bulk liquid circulation. In contrast, the heterogeneous regime is characterized by a nonuniform bubble concentration, especially in the traverse direction, because of liquid circulation.

A bubble coming out of a hole in a sparger grows until the buoyant force overcomes the surface tension force, and the bubble breaks away from the sparger. The bubble expands as it rises in liquid because of the reduction of the hydrostatic pressure. The bubble diameter may also increase because of mass transfer from liquid to gas and coalescence or decrease because of breakage. In the homogeneous regime with a low gas flux, gas is sparged uniformly at the bottom of the column, and it remains uniformly distributed all over the column. Eventually, all bubbles rise vertically with minor traverse and axial oscillations, and particularly, there is no coalescence or dispersion. The size of the bubbles in the homogeneous regime is, thus, almost entirely dictated by the type and design of the sparger and the physical properties of the system. On the other hand, for the heterogeneous regime, there exists a wide bubble size distribution in the column. The average bubble size is governed by coalescence and redispersion phenomena, which in turn are controlled by the energy dissipation rate in the bulk liquid. The intensity of turbulence is also much higher. Such highly turbulent recirculation results in substantially high values

of eddy diffusivity for mass transfer. As a result, the rates of mass transfer and mixing are quite different between homogeneous and heterogeneous regimes. Therefore, it is important to know the range of physical properties and operating parameters over which a particular regime prevails. As the transition from the homogeneous to the heterogeneous regime starts, there is an onset of liquid circulation that is upward in the central region and downward near the column wall. More bubbles enter the central region because it offers lower resistance to the flow. As a result, a transverse holdup profile begins to build up, which in turn intensifies liquid circulation. Therefore, the beginning of transition regime is very important, and further development is self-propagating.[28,29]

In bubble columns and gas–liquid stirred reactors, the estimation of parameters is more difficult than in gas–solid or liquid–solid fluidized beds. Solid particles are rigid, and hence the fluid–solid interface is nondeformable, whereas the gas–liquid interface is deformable. In addition, the effect of surface-active agents is much more pronounced in the case of gas–liquid interfaces. This leads to uncertainties in the prediction of all major parameters, such as the terminal bubble rise velocity, the bubble diameter, the gas holdup, and the relation between the bubble diameter and the terminal bubble raise velocity.

TRICKLE-BED REACTORS

Trickle-bed reactor is a type of three-phase reactor in which a gas and liquid stream flow concurrently, usually downward, over a fixed bed of catalyst particles at such a velocity that the liquid trickles down at the surface of the solid matrix while the gas fills the available porous space in the bed. Downward operation is normally preferred because of mechanical stability, lower axial mixing and less flooding, which can support higher liquid and gas flow rates, i.e., higher reactor capacity. Ideally, fluid stream should be a plug flow for better mass transfer rate and reaction rate. However, some axial intermixing is inevitable in practice. In some cases, a concurrently upward liquid and gas flow is used to provide better radial and axial mixing. This can be an advantage in the case with highly exothermic reactions.

In general, when a fixed bed is selected, the issue whether to employ a concurrent upflow or downflow operation must be considered. Operating a randomly packed bed reactor in the countercurrent mode is usually not feasible because flooding occurs at gas velocities far below industrial relevance. In a concurrent upflow, complete catalyst wetting is obtained at the expense of much larger liquid holdup compared to a concurrent downflow. High liquid holdup increases the liquid film mass transfer resistance for gaseous reactants and is undesirable if homogeneous liquid phase side-reactions occur. Because of complete catalyst wetting and a higher liquid holdup, heat transfer characteristics are much better (more uniform heat transfer rate and no hot spots) in a concurrent upflow operation. However, in a concurrent upflow, the flow may induce local vibration or movement of the particles, which possibly results in the attrition of the catalyst.[30]

Low liquid holdup, low pressure drop, and lack of foaming in high liquid flow rates are some of the advantages of downward trickle-bed reactors. For downward type reactors, the modulation of both gas and liquid flow rates has a profound effect on the hydrodynamics of the reactor. The hydrodynamics of the reactor in turn can influence the kinetics of the reaction.[31] On the other hand, poor liquid distribution, partial wetting of catalyst, poor mixing, difficulties in controlling the temperature of highly exothermic reactions, and low gas–liquid interaction resulting in lower mass transfer coefficients are some of the disadvantages of downward trickle-bed reactors.

Trickle-bed reactors can be operated in several flow regimes. The four main flow regimes are trickle flow, pulsing flow, mist flow, and bubble flow. Trickle flow regime prevails at relatively low gas and liquid flow rates. Liquid flows as a laminar film and/or in rivulets over the packing, while gas passes through the remaining void space. At high gas and low liquid flow rates, transition to mist flow occurs, where liquid mainly travels down the column as droplets entrained by the continuous gas phase. Bubble flow regime appears at high liquid flow rates and low gas flow rates, and is opposite in structure to mist flow. Liquid is the continuous phase and gas moves in the form of dispersed bubbles. At moderate gas and liquid flow rates, pulsing flow regime is obtained, which is characterized by the successive passage of liquid-rich and gas-rich regions through the bed. Pulsing flow is a kind of self-organization through which the bed is periodically run through with waves of liquid followed by relatively quiet periods of continuous gas and liquid flow. The pulses are characterized by high particle-liquid mass and heat transfer rates, large gas–liquid interfacial areas, complete catalyst wetting, mobilization of stagnant liquid, and diminished axial dispersion.

Trickle-bed reactors are widely used in hydrotreating processes, i.e., hydrodesulfurization of gasoline and diesel fuel, in petroleum refining, chemical, petrochemical, and biochemical processes. The knowledge of hydrodynamic parameters is vital in the design of a TBR because the conversion of reactants, reaction yield, and selectivity depend not only on reaction kinetics, operating pressure, and temperature, but also on the hydrodynamics of the reactor. Special care is also required to prevent flow maldistribution, which can cause incomplete catalyst wetting in some parts

of the bed.[32] This may result in reduced overall production rates and poorer selectivity. For strong exothermic reactions, more severe consequences can occur because of hot spot formation and possibly even runaway reaction.

CONCLUSIONS

The fundamental principles of the gas-to-liquid mass transfer were concisely presented. The basic mass transfer mechanisms described in the three major mass transfer models: the film theory, the penetration theory, and the surface renewal theory are of help in explaining the mass transport process between the gas phase and the liquid phase. Using these theories, the controlling factors of the mass transfer process can be identified and manipulated to improve the performance of the unit operations utilizing the gas-to-liquid mass transfer process. The relevant unit operations, namely gas absorption column, three-phase fluidized bed reactor, airlift reactor, liquid–gas bubble reactor, and trickled bed reactor were reviewed in this entry.

REFERENCES

1. Whitman, W.G. Preliminary experimental confirmation of the two-film theory of gas absorption. Chem. Metall. Eng. **1923**, *29*, 146–148.
2. Hines, A.; Maddox, R.N. *Mass Transfer Fundamentals and Applications*; Prentice-Hall PTR: Englewood Cliffs, NJ, 1985; 140–143.
3. Higbie, R. The rate of absorption of pure gas into a still liquid during short periods of exposure. Trans. Am. Inst. Chem. Eng. **1935**, 36–60.
4. van Elk, E.P. Gas–Liquid Reactions—Influence of Liquid Bulk and Mass Transfer on Process Performance, Ph.D. Thesis; Twente University, The Netherlands, 2001.
5. Fair, J.R.; Steinmeyer, D.E.; Penney, W.R.; Crocker, B.B. Liquid–gas systems. In *Perry's Chemical Engineers' Handbook*, 6th Ed.; Perry, R.H., Green, D.W., Maloney, J.O., Eds.; McGraw-Hill, Inc.: New York, 1984.
6. Danckwerts, P.V. Significance of liquid-film coefficients in gas absorption. Ind. Eng. Chem. **1951**, *43*, 1460–1467.
7. Leva, M. *Tower Packings and Packed Tower Design*, 2nd Ed.; The U.S. Stoneware Co.: Akron, OH, 1953.
8. Strigle, R.F., Jr. *Random Packings and Packed Towers Design and Applications*; Gulf Publishing Company: Houston, TX, 1987.
9. Billet, R. *Packed Towers in Processing and Environmental Technology*; VCH Publishers, Inc.: New York, 1995.
10. Hoek, P.J.; Wesselingh, J.A.; Zuiderweg, F.J. Small scale and large-scale liquid maldistribution in packed columns. Chem. Eng. Res. Des. **1986**, *64*, 431–449.
11. Yin, F.; Wang, Z.; Afacan, A.; Nandakumar, K.; Chuang, K.T. Experimental studies of liquid flow maldistribution in a random packed column. Can. J. Chem. Eng. **2000**, *78*, 449–457.
12. Gostick, J.; Pritzker, M.; Lohi, A.; Doan, H.D. Mass transfer variation within a packed bed and its relation to liquid distribution. Chem. Eng. J. **2004**, *100*, 33–41.
13. Zhu, Y.; Doan, H.D.; Lohi, A. *Liquid Distribution and Its Effect on Local Mass Transfer in a Packed Column of Pall Rings*; Paper presented at the 56th CSChE Conference, October 3–6, 2004, Calgary, Alberta.
14. Treybal, R.E. *Mass-Transfer Operations*, 4th Ed.; McGraw-Hill, Inc.: New York, 1985.
15. Bolles, W.; Fair, R.J. Improved mass transfer model enhances packed column design. Chem. Eng. **1982**, *89* (14), 109–116.
16. Doan, H.D.; Fayed, M.E. Entrance effect and gas-film mass-transfer coefficient in a large-diameter packed column. Ind. Eng. Chem. Res. **2000**, *39* (4), 1039–1047.
17. Sherwood, T.K.; Holloway, F.A.L. Performance of packed towers—liquid film data for several packings. Trans. Am. Inst. Chem. Eng. **1940**, *36*, 39–70.
18. Au-Yeung, P.H.; Ponter, A.B. Estimation of liquid film mass transfer coefficients for randomly packed absorption columns. Can. J. Chem. Eng. **1983**, *61*, 418–493.
19. Vijayaraghavan, P.; Kulik, C.J.; Lee, S. Modelling of a liquid entrained reactor for liquid phase methanol synthesis process. Fuel Sci. Technol. Int. **1992**, *10* (9), 1501–1521.
20. Air Products and Chemicals, Inc. *Liquid-Entrained Catalyst Operations at LaPorte Pilot Plant for Liquid-Phase Methanol Process, 1984–1985*; Final report to U.S. Department of Energy, AP-5049 Research Project 317-3, February 1987; 1–19.
21. Fan, L.S.; Yang, G. Gas–liquid–solid three-phase fluidization. In *Handbook of Fluidization and Fluid-Particle Systems*; Yang, W.C., Ed.; Marcel Dekker: New York, 2003.
22. Kim, S.D.; Kang, Y. Heat and mass transfer in three-phase fluidized-bed reactors—an overview. Chem. Eng. Sci. **1997**, *52* (21/22), 3639–3660.
23. Lee, D.H.; Kim, J.O.; Kim, S.D. Mass transfer characteristics and phase holdup in three

phase-fluidized beds. Chem. Eng. Commun. **1993**, *119*, 179–196.
24. Chisti, M.Y.; Moo-Young, M. Airlift reactors: characteristics, applications and design considerations. Chem. Eng. Comm. **1987**, *60*, 195–242.
25. Fu, C.C.; Wu, W.T.; Lu, S.Y. Performance of airlift bioreactors with net draft tube. Enzyme Microb. Technol. **2003**, *33*, 332–342.
26. Meng, A.X.; Hill, G.A.; Dalai, A.K. Hydrodynamic characteristics in an external loop airlift bioreactor containing a spinning sparger and a packed bed. Ind. Eng. Chem. Res. **2002**, *41*, 2124–2128.
27. Tobajas, M.; García-Calvo, E.; Siegel, M.H.; Apitz, S.E. Hydrodynamics and mass transfer prediction in a three-phase airlift reactor for marine sediment biotreatment. Chem. Eng. Sci. **1999**, *54*, 5347–5354.
28. Joshi, J.B.; Vitankar, V.S.; Kulkarni, A.A.; Dhotre, M.T.; Ekambara, K. Coherent flow structures in bubble column reactors. Chem. Eng. Sci. **2002**, *57* (16), 3157–3183.
29. Thorat, B.N.; Joshi, J.B. Regime transition in bubble columns: experimental and predictions. Exp. Thermal Fluid Sci. **2004**, *28* (5), 423–430.
30. Dudukovic, M.P.; Larachi, F.; Mills, P.L. Multiphase catalytic reactors: a perspective on current knowledge and future trends. Catal. Rev. Sci. Eng. **2002**, *44* (1), 123–246.
31. Gianetto, A.; Specchia, V. Trickle-bed reactors. State of art and perspectives. Chem. Eng. Sci. **1992**, *47* (13/14), 3197–3218.
32. Kundu, A.; Nigam, K.D.P.; Verma, R.P. Catalyst wetting characteristics in trickle-bed reactors. AIChE J. **2003**, *49* (9), 2253–2263.

Geothermal Energy

Sunggyu Lee
H. Bryan Lanterman
Department of Chemical Engineering, University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

The natural heat of the earth is called geothermal energy. ''Geothermal'' is a hybrid word combining geo(earth) and thermal (of heat). In total, geothermal energy includes all the thermal energy contained in the entire volume of the earth except for a relatively thin, comparatively cool outer surface. This represents about 260 billion cubic miles of rocks and metallic alloys at or near their melting temperatures. Geothermal resources are available ranging from shallow ground to hot rock and water several miles below the earth's outer surface, and even farther down toward the earth core to the region of extremely high temperatures from molten rock called magma.

Geothermal energy is the second largest source of heat to the earth after solar energy. Geothermal energy that is available using current technology is concentrated in underground reservoirs, usually in the forms of steam, high-temperature water, and/or hot rocks. The three applicable technology categories include geothermal heat pumps (GHPs), direct-use applications, and electric power plants. GHPs use the earth near its surface as a heat sink and heat source for heating and cooling. Direct-use applications utilize naturally occurring geothermally heated water for heating. Electric power plants use electric turbines fed by geysers to generate electricity. Similar to solar energy, the ability to use geothermal energy is hampered by its distribution over the earth's surface in amounts that are too small or of too low intensity.[1] This is especially true for the generation of electricity.

The most obvious forms of geothermal energy are geysers, boiling pools of mud, fumaroles, and hot springs. However, a greater potential exists in regions not yet deemed valuable for their energy possibilities—these are hot dry rocks (HDRs).

Besides the vast resource availability and the unique distribution pattern of the resources, geothermal energy is very clean and environment friendly. The utilization or conversion of geothermal energy generates no or little greenhouse gases, since the conversion or utilization process does not involve any chemical reaction, in particular combustion. Geothermal fields produce only about one-sixth of the carbon dioxide that a natural-gas-fueled power plant produces, and very little, if any, of the nitrous oxide or sulfur-bearing gases. Furthermore, geothermal energy is available 24 hr a day and 365 days a year, independent of the outside weather conditions. This is in sharp contrast to other green energy technologies such as wind and solar. In fact, geothermal power plants typically have average availabilities of 95% or higher, much higher than most coal and nuclear plants. Even this high availability can be further enhanced to a level that is practically near 100%, with advances and enhancements in the process technology.

GEOTHERMAL ENERGY AS RENEWABLE ENERGY

Need of Geothermal Energy

The development and importance of new clean energy sources such as geothermal energy have become intensified not only due to the depletion of petroleum resources, but also due to the environmental problems involved with conventional energy processes. Environmental problems associated with the utilization of fossil fuel sources involve (a) emission of greenhouse gases such as CO_2, CH_4, and N_2O; (b) emission of SO_x and H_2S; (c) discharge of nitrogen oxides; (d) potential emission of mercury and selenium; (e) emission of volatile and semivolatile organic compounds; (f) emission of particulate matters; and (g) contamination of soils and ground water resources with hazardous wastes.

At a depth of about six miles from the earth surface, the temperature is higher than 100°C; thus the total amount of geothermal energy in storage far exceeds, by several orders of magnitude, the total thermal energy accountable in all forms of nuclear and fossil fuel resources of this planet. Solar energy is the only comparable resource of such vast quantities of energy. Therefore, it is very logical, if not imperative, that our energy priorities must incorporate a resource such as geothermal energy.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007789

Renewability and Sustainability of Geothermal Energy

The U.S. Department of Energy (DOE) classifies geothermal energy as renewable. Its source is the continuously emanated thermal energy generated by the earth's core. Each year rainfall and snowmelt supply new water source to geothermal reservoirs, and production from individual geothermal fields can be sustained for decades and perhaps for centuries. An accurate prediction for sustainable service life of each field is very difficult.

Occurrence of Geothermal Energy

a. The occurrence of geothermal heat (also known as geoheat) can be explained by one of the following theories:[2] The first theory is that about six billion years ago, the earth was a hot molten mass of rock and this mass has been cooling through the epochs of time, with the outer crust formed as a result of a faster cooling rate.
b. The second theory presupposes that the earth is like a giant furnace. The decaying of radioactive materials within the earth provides a constant heat source.
c. The third theory is based on the presumption that geothermal heat originates from the earth's fiery consolidation of dust and gas over four billion years ago.

Even though a generally agreeable explanation for the natural occurrence of geothermal energy is unavailable, some sort of combination of the aforementioned theories are widely offered.

The interior of the earth consists of a molten fluid of rocks at extremely high temperatures, called magma. This magma is cooling and/or expelling heat to the earth's surface according to the second law of thermodynamics. The flow of heat is from the hot source (earth core) toward the cold sink (earth surface). The cold sink (i.e., heat sink in thermodynamics) consists of the earth's crust, surface, and atmosphere. This may be regarded as a very slow process of heat transfer.

Fig. 1 shows a typical geological setting of a geothermal energy source. Thermal energy from the earth core continuously flows outward. The heat transfer from the core to the surrounding layers of rock, the mantle, is principally via conduction. As the temperature and pressure of the system become high enough, some mantle rocks melt and form magma. Since the magma as a liquid phase is less dense and more fluid-like than the surrounding rock, it slowly rises and moves toward the earth's crust, thus convecting the heat from the core. This is why a slow convective heat transfer often represents the overall heat transfer process. Sometimes, the hot molten magma reaches all the way to the earth's surface, where it is known as lava. However, in most cases, the magma remains well below the earth's crust, heating neighboring rocks and water that originates from rainwater seeped deep into the earth. The temperature of the water can be as hot as 380°C, which is even higher than the critical temperature of water, 374°C.

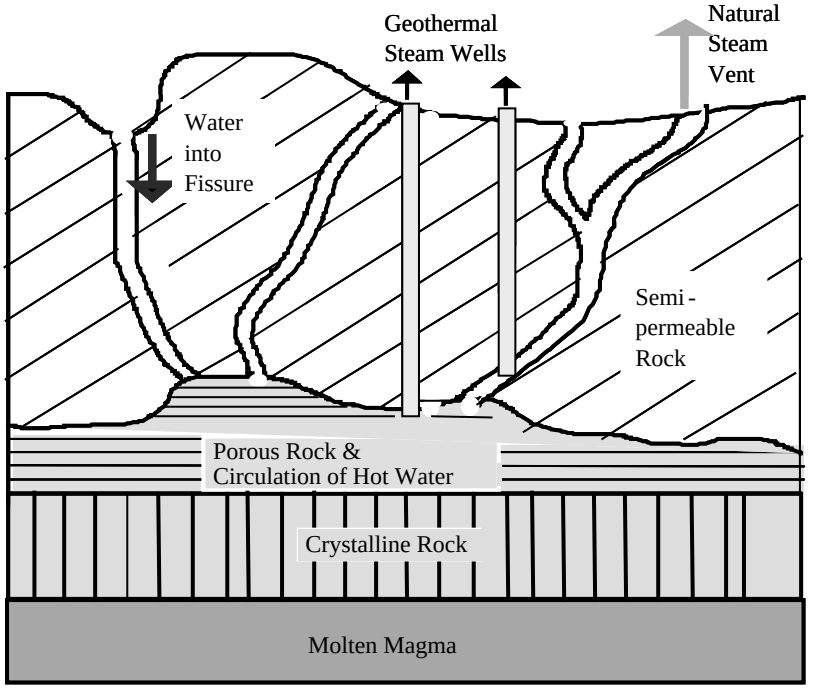


Fig. 1 A geothermal energy source. *(View this art in color at www.dekker.com.)*

The water at this depth is subjected to high pressure and temperature. Depending upon the imposed conditions, water may exist as a supercritical fluid. It then rises to the surface through fissures as a result of its density change and in effect, it vents from the system, thereby reducing the pressure of the system. When the pressure decreases, the water boils, turns into steam, and rises to the surface through fissures and/or wells. Some of the best-known examples of hot geothermal water are hot springs or geysers. However, most of the hot geothermal water remains deep underground, typically trapped in cracks and porous rocks. This natural collection of hot underground water is called a geothermal reservoir.

The immenseness of the geothermal energy of the earth cannot be properly determined. The total energy content of the rocks above the depth of 10 km has been estimated to be 3×10^{26} cal. At the earth core, i.e., 6400 km deep, the temperature may reach over 5000°C. However, based on current technology, only a fraction of this heat is available as a recoverable resource. A mere 0.03% or 10^{23} cal of this energy is considered hot enough and near enough to the earth's surface to be recoverable, the rest of the energy is too dispersed throughout the crust or too deep to be practical.

Geothermal resources may be classified into several categories principally based on their phases and forms, as shown in Table 1.

While all resources are not viable to produce electricity, they are still useful in many industrial, agricultural, and domestic needs. Research conducted by the U.S. Geological Survey identifies the significant research areas that need to be accomplished for full utilization of the geothermal resources.[3]

Table 1 Classification of geothermal resources

Types of geothermal resources	Temperature (approx. °C)
(a) *Convective hydrothermal resources*	
Vapor dominated	240
Hot water dominated	30–350
(b) *Other hydrothermal resources*	
Sedimentary basins/regional aquifers (hot fluid in sedimentary rocks)	30–150
Geopressured (hot fluid under pressure that is greater than hydrostatic)	90–200
Radiogenic (heat generated by radioactive decay)	30–150
(c) *Hot rock resources*	
Part still molten (magma)	>600
Solidified (hot dry rock)	90–650

Advantages of Geothermal Energy

Geothermal energy resources are continuous and reliable, sustainable, clean, and can be cost-competitive in meeting baseload capacity needs. Specific advantages of geothermal systems include:

1. Indigenous energy—geothermal energy helps reduce dependence on fossil or nuclear fuels and as such helps keep the economic benefits in the region.
2. Clean energy—use of geothermal energy helps reduce combustion-related emissions.
3. Diversity of use—geothermal energy has three common economic uses—electricity generation, direct use of heat, and GHPs.
4. Long-term resource potential—with optimum development strategies, geothermal energy can provide a significant portion of a nation's long-term energy needs.
5. Flexible system sizing—current power generation projects range in capacity from a 200 kW system in China to 1200 MW at The Geysers in California. Additional units can be installed in increments depending upon the growth of the electricity demand.
6. Power plant longevity—geothermal power plants are designed for a life span of 20–30 yr. With proper resource management strategies, life spans can exceed design periods.
7. High availability and reliability—''Availability'' is defined as the percentage of time that a system is capable of producing electricity. Availability of 95–99% is typical for modern geothermal plants compared to maximum 80–85% for coal and nuclear plants.[6]
8. Combined use—geothermal energy can be simultaneously used for both power generation and direct-use applications.
9. Low operating and maintenance costs—the annual operation and maintenance costs of a geothermal electric system are typically 5–8% of the capital cost.
10. Land area requirement—the land area required for geothermal power plants is smaller per megawatt than that for almost every other type of power generation plant.
11. Enhanced standard of living—geothermal systems can be installed at remote locations without requiring other industrial infrastructure. The region can grow without pollution.

Global Geothermal Energy

The current total installed capacity of geothermal power stations throughout the world is over

8200 MW.[4] The United States remains the biggest producer of electricity from geothermal energy as shown in Table 2. The developing countries accounted for 35% of the total for 1995 and 46% for 1999. About 2850 MW of electricity generation capacity is available from geothermal power plants in the western United States. The major geothermal fields along with their capacities are shown in Table 3.[5] Direct-use geothermal technologies utilize naturally hot geothermal water for commercial greenhouses, crop dehydration, fish farming, bathing, and district community heating. GHPs use the constant temperature of the top 50 ft of earth's surface to heat buildings in winter and cool them in summer. Table 4 shows the direct use of geothermal heat in various categories in the United States.[6]

Fossil fuels, namely oil, coal, and gas, provide 85% of all the energy used in the United States. Renewable energy sources supply just 8%, most of which comes from hydropower and the burning of biomass while only 4% comes from geothermal sources. Fig. 2 shows the history and projections of the U.S. energy consumption by fuel sources for the period of 1970–2020.

Table 2 Installed geothermal electricity generation capacity

Country	1990	1995	1998
Argentina[a]	0.67	0.67	0
Australia	0	0.17	0.4
China	19.2	28.78	32
Costa Rica	0	55	120
El Salvador	95	105	105
France (Guadeloupe)	4.2	4.2	4.2
Greece[a]	0	0	0
Guatemala	0	0	5
Iceland	44.6	49.4	140
Indonesia	144.75	309.75	589.5
Italy	545	631.7	768.5
Japan	214.6	413.7	530
Kenya	45	45	45
Mexico	700	753	743
New Zealand	283.2	286	345
Nicaragua	70	70	70
Philippines	891	1191	1848
Portugal (Azores)	3	5	11
Russia	11	11	11
Thailand	0.3	0.3	0.3
Turkey	20.4	20.4	20.4
U.S.A.	2774.6	2816.7	2850
Totals	5866.72	6796.98	8240

[a]Argentina and Greece closed their pilot plants.

Table 3 Major geothermal power plants in the United States

Location	Capacity installed (MW)
The Geysers, CA	2115
East Mesa, CA	119
Salton Sea	198
Heber, CA	94
Mammoth, CA	7
Coso, CA	225
Amadee, CA	2
Wendel, CA	0.6
Puna, HI	18
Steamboat, NV	31
Beowave, NV	17
Brady, NV	6
Desert Peak, NV	9
Wabuska, NV	1.2
Soda Lake, NV	3.6
Stillwater, NV	14
Empire Farms, NV	4.8
Roosevelt, UT	20
Cove Fort, UT	4.2
Total	2889.4

(From Ref.[5].)

As shown in Fig. 2, the projected future growth in the area of nonhydro and nonnuclear renewable energies is substantial. As the relative cost of electricity generation from geothermal sources decreases, the popularity of the geothermal power generation will undoubtedly increase.[7] Fig. 3 shows a cost comparison among various modes of power generation.

HISTORY OF GEOTHERMAL ENERGY DEVELOPMENTS

Ancient people regarded the depths of the earth with horror, as the seat of hell and of malignant Gods, due to natural phenomena like earthquakes and volcano eruptions. In ancient times the Romans and in modern times the Icelanders, Japanese, Turks, Koreans, and others have used it for baths and for space heating.

The Larderello field in Tuscany, Italy, first began to produce electricity in 1904 and developed over the next 10 yr to a capacity of 250 KW. In Japan, Beppu was the first site for experimental geothermal work in 1919 and these experiments effected a pilot plant in 1924 producing 1 KW of electricity. Somewhat earlier than this the Japanese began to use geoheat to heat their greenhouses. In Iceland, municipal heating was

Table 4 U.S. geothermal direct-use projects

Application	No. of sites	Thermal capacity (MW)	Annual energy (GWh)
Geothermal heat pumps	Most states	2072	2402
Space and district heating	126	188	433
Greenhouses	39	66	166
Aquaculture	21	66	346
Resorts/pools	115	68	426
Industrial processes	13	43	216
Totals		2503	3969

provided using hot thermal waters in the 1930s and is still the major source of heating today.

It was not until the early 1920s that the United States examined the possibilities of the commercial usage of geothermal steam. However, the competition from hydroelectric power was too keen to promote further development at that time. Today, the largest geothermal power plant in the world is located in California at The Geysers, which is probably the largest reservoir of geothermal steam in the world.[8] In recent years, U.S. DOE's GeoPowering the West (GPW) program has been working to further geothermal energy efforts.[9]

Growing concerns about the global effects of increasing CO_2 and methane in the atmosphere are working to enhance the role of geothermal resources worldwide. Hence the utilization of geothermal energy to generate electric power dominates all other applications.

A number of factors that have boosted the production of geothermal energy are:[10]

i. The economics of geothermal energy became more favorable due to the increase in petroleum and natural gas prices.
ii. The cost of producing geothermal energy decreased from 1980 to 2000.
iii. Legislative actions and measures encouraging geothermal developments have been in place for many countries. Examples in the Unites States include the Energy Policy Act and the National Geologic Mapping Act in the early 1990s.[11]
iv. The implementation of the Clean Air Act Amendments of 1990 also provides an economic benefit because of the well-developed technology for control of gas emissions from geothermal power plants.
v. Amendment of the Public Utilities Regulatory Act removed the 80 MW limit from independent power plants selling electricity to utilities and is expected to help competitiveness of geothermal energy.

GEOTHERMAL PROCESSES AND APPLICATIONS

Geothermal Power Plants

Geothermal resources may be described as hydrothermal, HDR, or geopressured. Hydrothermal resources contain hot water, steam, or a mixture of water and

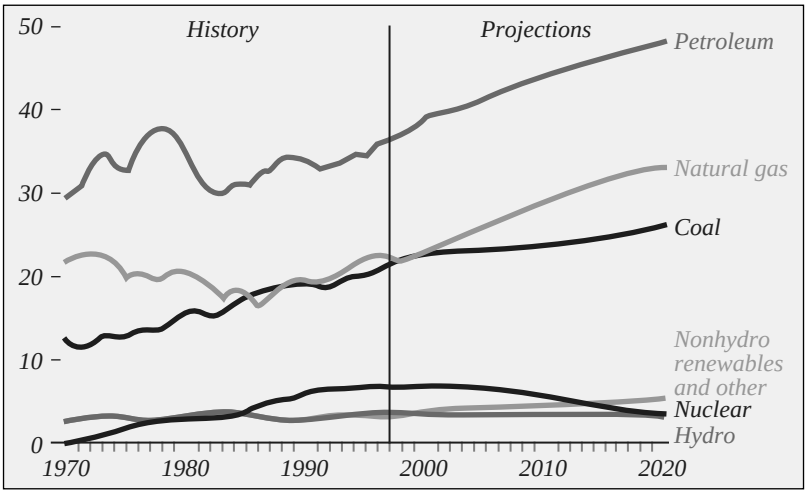


Fig. 2 U.S. energy consumption by fuel sources; past, current, and future forecast. Unit used is in quadrillion BTUs. (U.S. Department of Energy,[11] 1999.) *(View this art in color at www.dekker.com.)*

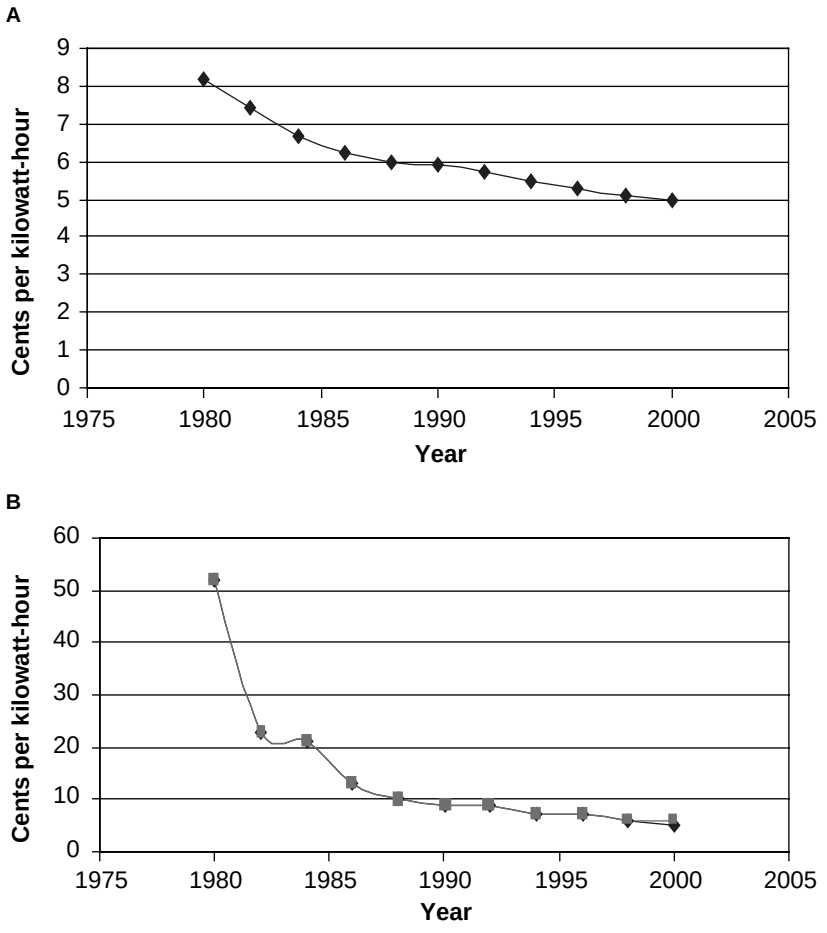


Fig. 3 Cost comparison of electricity generation between (A) geothermal energy and (B) solar thermal energy. *(View this art in color at www.dekker.com.)*

steam. Although research and development continue to look for ways to efficiently extract and use the energy contained in HDR and geopressured resources, virtually all current geothermal power plants operate on hydrothermal resources.

The characteristics of the hydrothermal resource determine the power cycle of the geothermal power plant. A resource that produces dry steam uses a direct steam cycle. A power plant for a liquid-dominated resource with a temperature above 165°C typically uses a flash steam cycle. For liquid-dominated resources with temperatures below 165°C, a binary cycle is the best choice for power generation. Power plants on liquid-dominated resources often benefit from combined cycles, using both flash and binary energy-conversion cycles.

Direct steam cycle[12]

Direct steam is also referred to as dry steam. As the term implies, steam is routed directly to the turbines, thus eliminating the need for the boilers used by conventional natural gas and coal power plants. Fig. 4 shows a schematic of a direct steam cycle power generation process. In a direct steam cycle power plant, a geothermal turbine can operate with steam that is far from pure. Chemicals and compounds in solid, liquid, and gaseous phases are transported with the steam to the power plant. At the power plant, the steam passes through a separator that removes water droplets and particulate before it is delivered to the steam turbine. The turbines are of conventional design with special materials, such as 12Cr steel and precipitation-hardened stainless steel, to improve reliability in geothermal services.

The other components present in the direct steam geothermal cycle include:

a. A condenser used to condense turbine exhaust steam. Both direct contact and surface condensers are used in direct steam geothermal power plants.
b. Noncondensable gas-removal system, to remove and compress the noncondensable gases. A typical system uses two stages of compression. The first stage is a steam jet ejector. The second stage is another steam jet ejector, a liquid ring vacuum pump, or a centrifugal compressor.

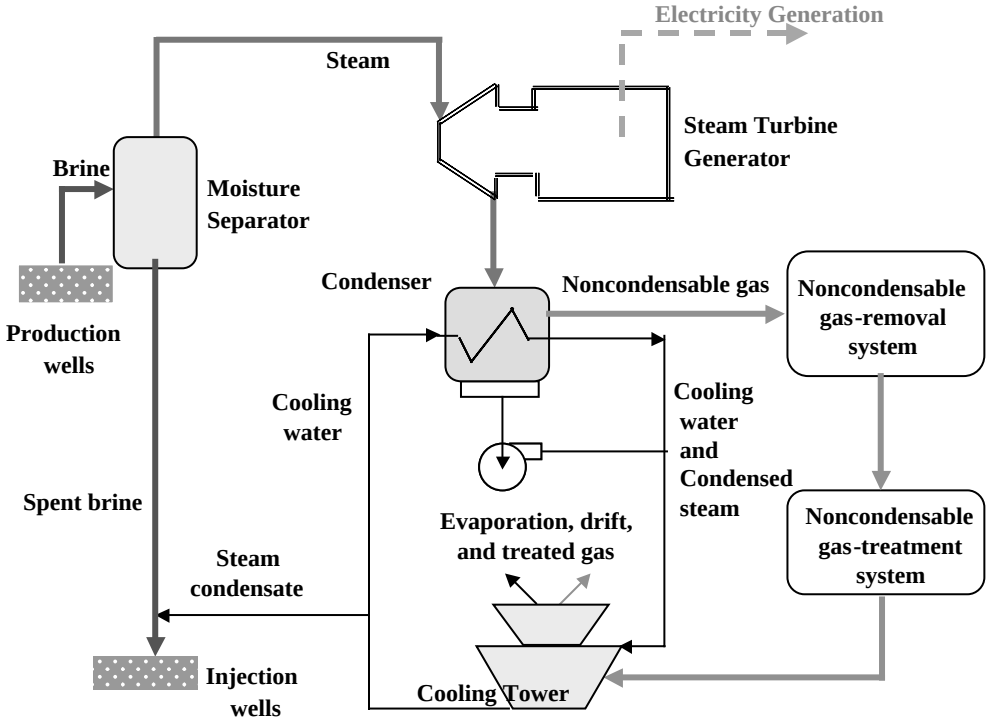


Fig. 4 Direct steam cycle geothermal power plant. *(View this art in color at www.dekker.com.)*

c. Cooling tower, which is a multicell wet mechanical draft design. Cooling is accomplished primarily by evaporation. Water that is lost from the cooling system to evaporation and drift is replaced by steam condensate from the condenser.
d. Excess water is returned to the geothermal resource in an injection well.

The direct steam cycle is typical of power plants at The Geysers in northern California, the largest geothermal field in the world. The primary operator of The Geysers is Pacific Gas and Electric.

Flash steam cycle[12]

Flash steam is the steam produced when the pressure on a geothermal fluid is reduced. A flash steam cycle for a high-temperature liquid-dominated resource is shown in Fig. 5. This dual-flash cycle is typical of most larger flash steam geothermal power plants. Single-flash cycles are frequently selected for smaller facilities.

Geothermal brine or a mixture of brine and steam is delivered to a flash vessel at the power plant by either natural circulation or pumps in the production wells. At the entrance to the flash vessel, the pressure is reduced to produce flash steam. The steam is delivered to the high-pressure inlet to the turbine. The remaining brine drains to another flash vessel where the pressure is again reduced to produce low-pressure flash steam.

The other components present in double-flash steam geothermal cycle include:

a. Direct contact condenser, as hydrogen sulfide is not produced in large quantities.
b. A cooling tower. This cycle also uses a multicell wet mechanical draft cooling tower. The water lost to evaporation and drift is replaced by steam condensate.
c. The excess water and spent brine from the flash vessels are injected back into the geothermal resource in an injection well.

Binary cycle[4]

A binary-cycle geothermal power plant employs a closed-loop heat exchange system in which the heat of geothermal fluid (''primary fluid'') is transferred to a lower-boiling heat transfer fluid (''secondary fluid'') that is thereby vaporized and used to drive a turbine/generator set. In other words, a binary cycle uses a secondary heat transfer fluid instead of steam in the power generation equipment. Binary geothermal plants have been in service since the late 1980s. A binary cycle is the economic choice for hydrothermal resources with temperatures below approximately 165°C. A typical binary cycle is shown in Fig. 6.

The binary cycle shown in Fig. 6 uses isobutane (i-C_4H_{10}) as the binary heat transfer fluid. Heat from

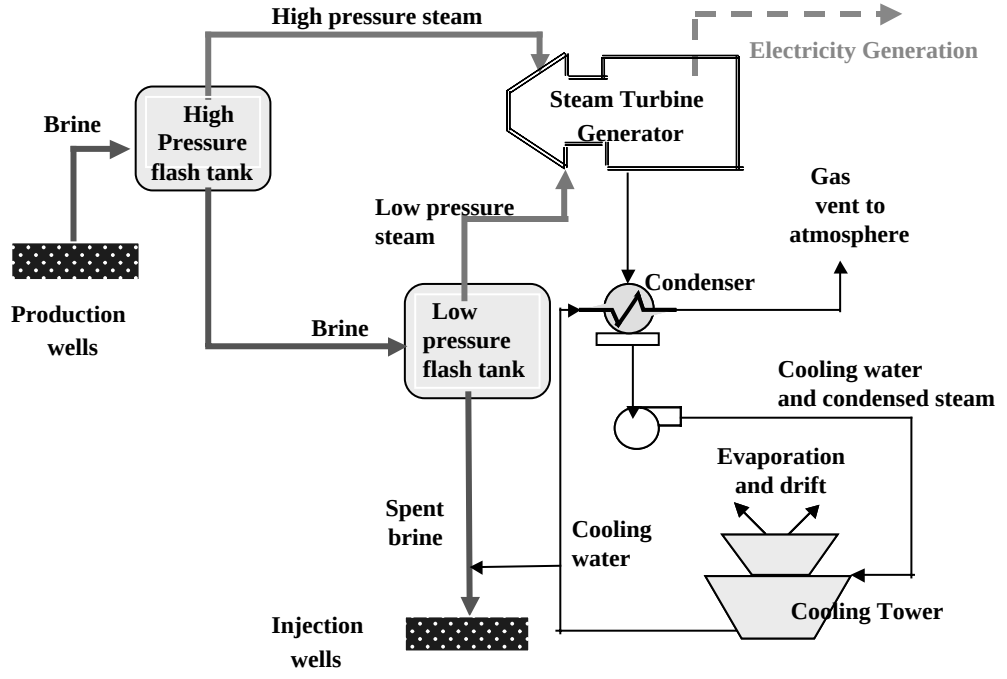


Fig. 5 Double-flash steam cycle geothermal power plant. *(View this art in color at www.dekker.com.)*

geothermal brine vaporizes the binary fluid in the brine heat exchanger. Spent brine is returned to the resource in injection wells, and the binary fluid vapor drives a turbine generator. The turbine exhaust vapor is delivered to an air-cooled condenser, where the vapor is condensed. Liquid binary fluid drains to an accumulator vessel before being pumped back to the brine heat exchangers to repeat the cycle. The brine heat exchangers are typically shell-and-tube units fabricated from carbon steel.

HDR (dry geothermal sources) systems[2,13,14]

Since the vast majority of the geothermal heat resources of the world exists in forms of HDR sources rather than water (hydrothermal) systems, it is only

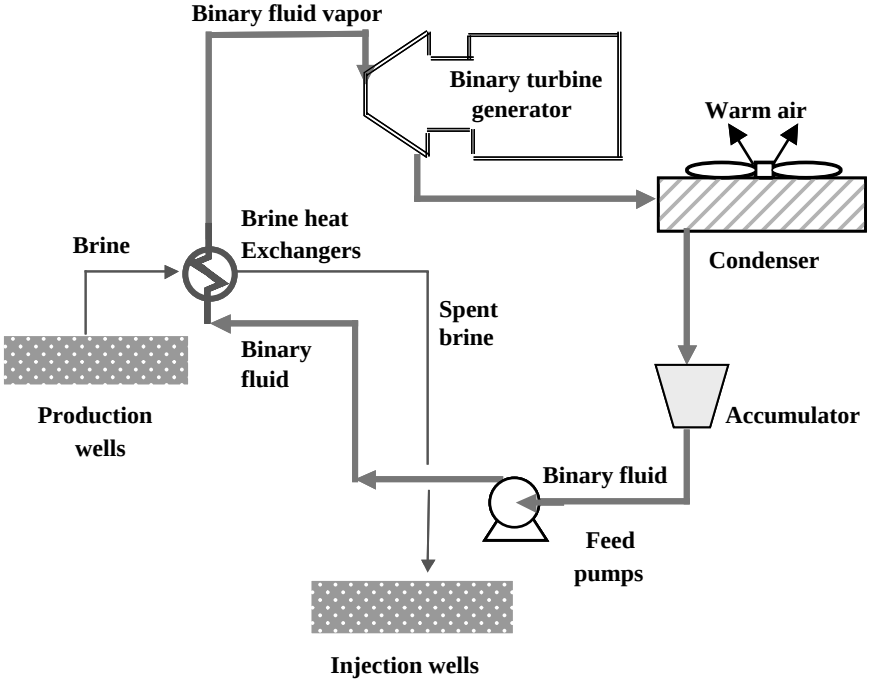


Fig. 6 Binary cycle geothermal power plant. *(View this art in color at www.dekker.com.)*

natural for this energy source to receive more attention from geothermalists. The more accessible HDR resources in the United States alone would provide an estimated 650,000 quads of heat, one quad (one quadrillion BTUs) being equivalent to the amount of energy contained in 171.5 million barrels of oil. Since the annual U.S. energy consumption is approximately 84 quads, whoever figures out how to economically tap even a fraction of the potential in HDR could earn a place in history.

HDR is a deeply buried crystal rock at a usefully high temperature. Current engineering designs plan to tap its heat by drilling a welbore, fracturing or stimulating pre-existing joints around the welbore, and directionally drilling another welbore through the fracture network. Cold water then flows down one welbore, pushes through the fractured rock, warms, returns up the other welbore, and drives a power plant. The major technical uncertainty is establishing the fracture network between the two welbores. If adequate connectivity can be established and sufficiently large fracture surface area can be exposed between the two welbores, HDR can be a very competitive source of energy.

Fig. 7 is a schematic diagram of the experimental Los Alamos System in New Mexico. Water at 65°C and 1000 psia is pumped into hydraulic fracture network, approximately 3000 ft in diameter and circulated at 7500 ft where temperatures range between 260°C and 320°C. The pressure is around 4100 psia. The water is then pumped out of the ground and when it reaches the surface its temperature is 230°C at 1250 psia.

In this experimental system, the hot water is circulated through an air-cooled heat exchanger with the extracted heat dissipated to the atmosphere.

Fresh water production

Less than 2% of the earth's retained water supply is available for drinking. The oceans, atmosphere, rocks or rock formations, and polluted resources contain the remaining 98%. From the standpoint of water shortage, all the systems recognized to date (desalination, recycling, and/or transportation over long distances) consume enormous amounts of energy and have also proven to be uneconomical. Geothermal resources, on the other hand, contain vast reservoirs of hot water and steam, and some of these are producing electricity and fresh water as a byproduct. The geothermal resource satisfies two main criteria for alleviating water shortages, viz.

a. An energy source for distillation process such as multistage flash and vertical tube evaporator.
b. An ample supply of water.

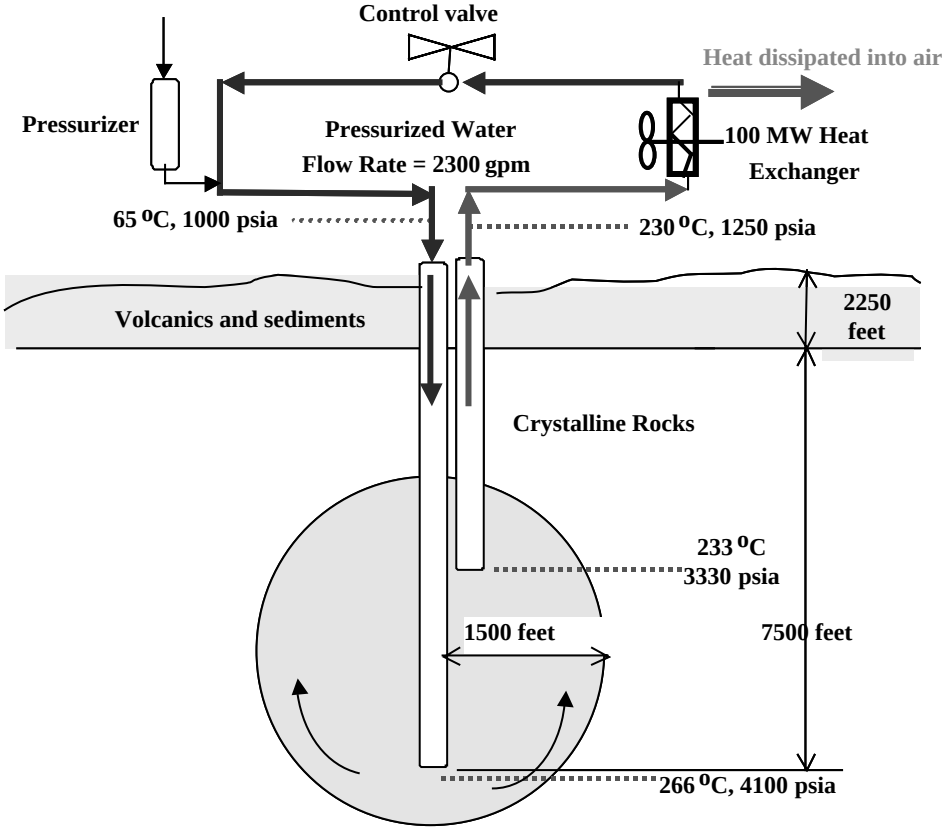


Fig. 7 Experimental configuration and operating conditions Los Alamos hydraulic fracture network. *(View this art in color at www.dekker.com.)*

Direct Use of Geothermal Heat

Direct heat use is one of the oldest, most versatile, and also the most common form of utilization of geothermal energy. The term, direct use, means that geothermal heat is used directly without first converting it to electricity. The warm water or steam exiting the ground will be piped into the dwelling or structure to provide warmth. Direct heating is obviously an older technology than geothermal power generation and is widely practiced.

Space and district heating

Direct heating can be applied in what are known as either district or space heating systems, the distinction being that space heating systems serve only one building, while district heating systems serve many structures from a common set of wells. Direct heating has made the greatest progress and development in Iceland, where the total capacity of the operating geothermal district heating system is 800 MW.[15] Fig. 8 shows an example of a district heating system.

Each system has to be adapted to the local situation, depending on the type of geothermal resource available, the population density of the area and the predicted population growth, the type of buildings requiring heating/cooling and, above all, the local climate. Geothermal district heating pumps are capital intensive in the early stages. The principal costs are initial investments for production and injection wells, down-hole and circulation pumps, heat exchangers, and pipelines, as well as the distribution network. A high load density usually makes district heating economically more feasible, because the cost of the distribution network transporting hot water to consumers is shared. Importantly, operating costs are comparatively low, thus making the long-term cost much more favorable. Geothermal district heating systems offer significant life cycle cost savings to consumers, as much as 30–50% of the cost of using natural gas or oil.

At present, a very successful district heating system exists in San Bernardino, California. The water production system, consisting of two wells, yields an average flow of 5200 L/min at 54°C water. The system currently serves 33 buildings including government centers, a prison, a new blood bank facility, and other private buildings.

Agricultural applications

One specific application of direct heating is greenhouse heating. This is one of the most common worldwide applications of geothermal energy. Fruits, vegetables, flowers, and ornamental plants are successfully grown year-round, in geothermally heated greenhouses using low-temperature sources (<38°C). Geothermal energy can extend short growing seasons and significantly reduce fuel costs. One example is a 650 m^2 greenhouse in California utilizing a geothermal well 150 m deep that supplies 67°C water. The well is capable of supplying heat for an additional 1800–3700 m^2 of greenhouse.

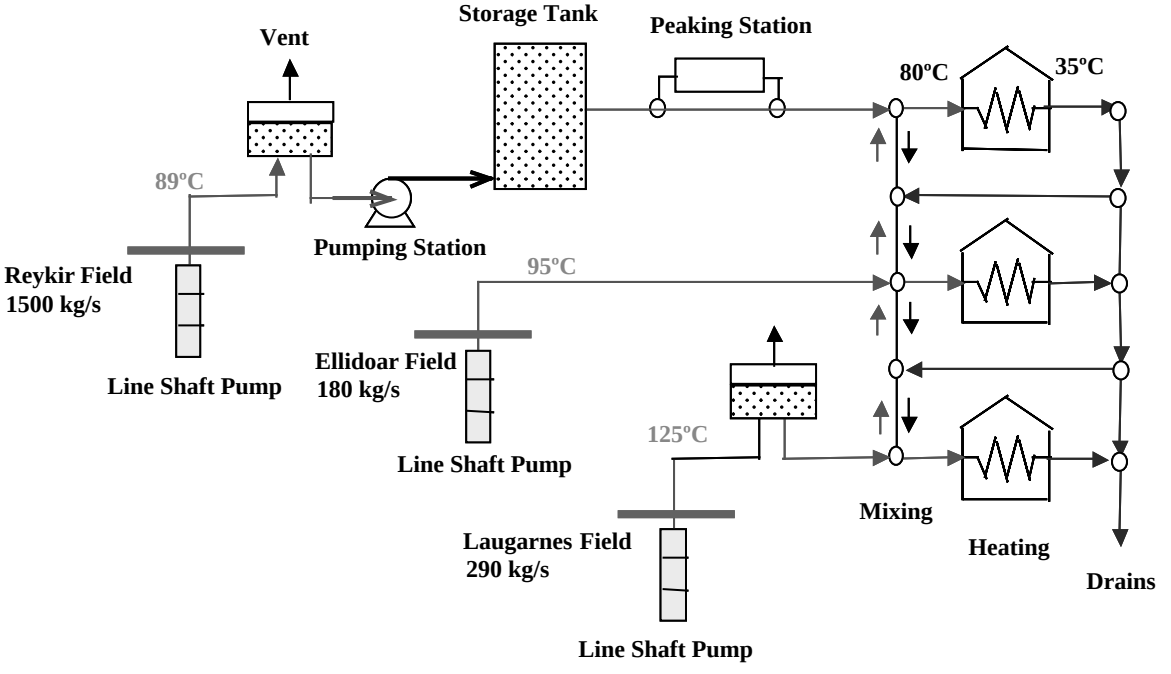


Fig. 8 A schematic of Hitaveita Reykjavikur (Reykjavik district heating system). *(View this art in color at www.dekker.com.)*

It should be noteworthy that the energy crisis experienced in California in 2001 rendered a very severe threat to greenhouse farmers who relied on electricity or natural gas heating.

Another direct heating application involves aquaculture, which is the raising of freshwater or marine organisms in a controlled environment. Geothermally heated water produces excellent yields of high quality fish and crustacean under accelerated growth conditions. Furthermore, geothermal aquaculture permits breeding in the winter, allowing fish farmers to harvest their products when product availability is low and market prices are high. California has six geothermal aquaculture operations as of late 1990s. The largest, the Hot Creek Hatchery near Mammoth, uses water from four springs with temperatures of 11–16°C.

Balneology

Balneology involves the use of geologically heated water/brine/mud sources for bathing purposes while also being alleged to possess healing and prophylactic properties. Balneology is centuries old and has been practiced by Etruscans, Romans, Greeks, Turks, Mexicans, Japanese, Koreans, Americans, and undoubtedly others.

Industrial process heat

Industrial processes can be heat intensive, and they commonly use either steam or superheated water with temperatures of 150°C or higher. This makes industrial processes the highest-temperature users of geothermal direct-heat applications. However, lower temperatures can suffice in some cases, especially for some drying applications. Two of the largest industrial users of geothermal heat are a diatomaceous-earth drying plant in Iceland, and a paper and pulp processing plant in New Zealand.

Geothermal Heat Pumps

The GHP, also referred to as the ground source heat pump (GSHP), uses the earth as a heat source for heating and as a heat sink for cooling. The GSHP uses a reversible refrigeration cycle combined with a circulating ground loop to efficiently provide either heating or cooling from electricity. The basic mechanism is the same as that of an air-source heat pump, but operates more efficiently since the temperature of the ground is more favorable than that of the air, i.e., the ground is warmer in the winter and cooler in the summer than the air. Additionally, the ground temperature is fairly constant throughout the year, even at depths of as little as 5–10 ft.

The typical components of a residential GSHP during heating and cooling cycles are shown in Figs. 9 and 10. The major components of the system are the ground loop and a refrigeration unit composed of the compressor, primary heat exchanger, expansion valve, and secondary heat exchanger. The refrigeration cycle utilizes the same unit operation steps as that of an air-source heat pump or typical home air-conditioning unit. In what is referred to as a closed-loop system, a water and antifreeze mixture circulates through a pipe buried in the ground and transfers thermal energy between the ground and the primary heat exchanger in the heat pump. Depending on the mode of operation, either heating or cooling is provided based on the reversible valve that allows the refrigerant to reverse the order of the operations of the cycle. Therefore, the primary heat exchanger, which consists of a water-to-refrigerant loop, can act as an evaporator or a condenser. Also included in this system is a heat exchanger following the compressor that provides heat to a hot water heater, this is often referred to as a de-superheater. The Geothermal Heat Pump Consortium, Inc., offers technical, educational, and promotional support for Geoexchange systems.[16]

GHPs offer a distinct advantage over the use of air as a source or sink, since the ground is at a more favorable temperature. Compared to atmospheric air, the ground is warmer in winter and cooler in summer. Therefore, GHPs demonstrate better performance over air-source heat pumps. They also reduce electricity consumption by approximately 30% compared to air-source heat pumps. Aided by utility sponsored programs, GHPs are becoming increasingly popular throughout the world. In the U.S., the GHP industry is expanding at a growth rate of 10–20% annually. As of 2004, more than 200,000 GHPs are being operated in U.S. homes, schools, and commercial buildings.[17]

The industrial and other potential applications of geothermal energy suggest that the great economic advantages could be gained from dual- or multipurpose plants combining power production with one or more other applications. Such plants would enable the costs of exploration drilling and certain other items to be shared among two or more end users.

SCIENTIFIC AND TECHNOLOGICAL DEVELOPMENTS

Major Research Efforts

The following major activities are examples of U.S. Government funded research being conducted in accordance with its R&D strategy:[6,9]

a. Advanced techniques to detect and delineate hidden geothermal resources are being developed,

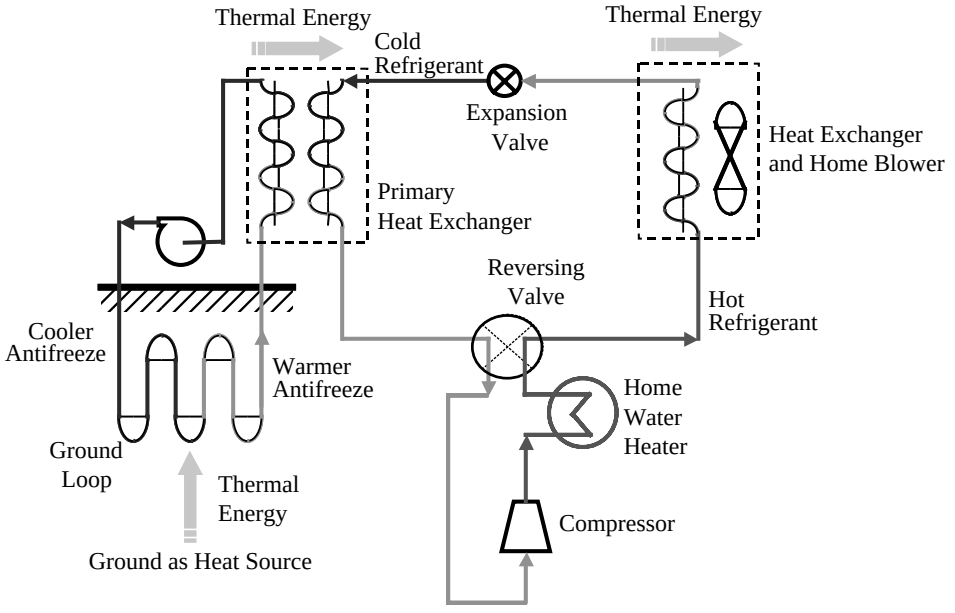


Fig. 9 GSHP during heating cycle. *(View this art in color at www.dekker.com.)*

including remote sensing techniques and improvements of various electric and acoustic methods.

b. Slim-hole drilling/coring, a cost effective option for exploratory drilling, needs to be improved. This research includes developing slim-hole reservoir engineering techniques and logging tools.
c. Improved materials that are capable of withstanding the high temperature and corrosive nature of geothermal brines are being developed.
d. Methods to increase the net brine effectiveness of geothermal power plants are being pursued, as are ways to reduce power plant costs.

Significant research activities ongoing at National Laboratories and Universities in the United States and the world include those of, Sandia National Laboratory,[9,18] Lawrence Berkeley National Laboratory,[9] Brookhaven National Laboratory,[4,9] National Renewable Energy Laboratory,[17] Los Alamos National Laboratory,[19–21] The Geysers,[10] Camborne School of Mines,[4] European Hot Dry Rock Industries,[4] Stanford University and Leningrad's Mining Institute,[4] Electric Power Research Institute (EPRI), Geo-Heat Center of the Oregon Institute of Technology, Southern Methodist University Geothermal

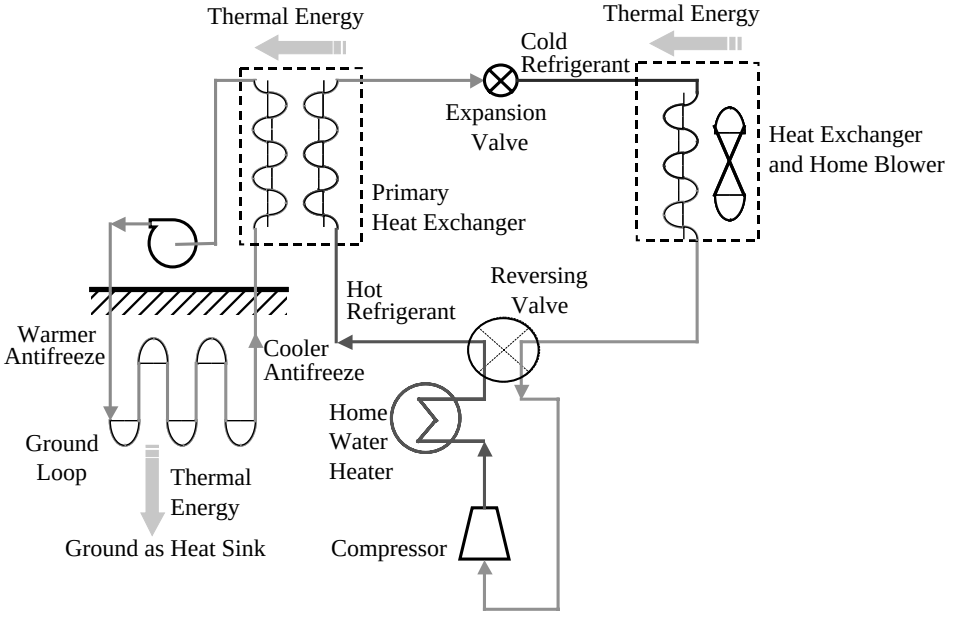


Fig. 10 GSHP during cooling cycle. *(View this art in color at www.dekker.com.)*

Laboratory, and Virginia Polytechnic Institute and State University.

Technology Updates

The development of a successful geothermal energy project relies on a variety of specialized technologies as well as their cost effectiveness.

Exploration technology

Exploration is a key to the discovery of new geothermal resources. It identifies geothermal resources, estimates resource potential, and establishes resource size, depth, and potential production. It relies on surface measurements of subsurface geological, geochemical, and geophysical conditions to develop a conceptual model of the system. Geothermal exploration of unmapped regions typically proceeds in two basic phases, reconnaissance and detailed exploration. During the reconnaissance phase, regional geology and fracture systems are studied, such as young volcanic features, tectonically active fault zones (as deduced from seismic information), and overt or subtle geothermal manifestations. If the reconnaissance phase confirms that the province has geothermal potential and that specific sites in the province should be explored further, the second phase focuses on one or more individual prospects covered in the reconnaissance phase.

Brine handling technology

Brine is a geothermal solution containing appreciable concentrations of sodium chloride or other salts. The chemical composition including the salinity of geothermal fluids varies greatly from one reservoir to another. Variations in chemistry and salinity affect the design, maintenance, and longevity of wells and surface equipments. Recent advances in this area include:[6]

i. Use of scale-inhibiting chemicals to reduce carbonate scaling of flashing wells.
ii. Development of pH modification to control silica scaling in power plants.
iii. Development of highly effective computer programs to estimate and predict chemistry effects in geothermal systems.
iv. Continued development of polymeric cement coating to reduce corrosion in heat exchangers and process piping.

Environmental issues of geothermal energy utilization

Even though geothermal energy is one of the cleanest and safest means for generating electric power, its effects on water resources, air quality, and noise during geothermal development and operation must be understood and mitigated. Among these are emissions to air (particularly of hydrogen sulfide), land use, and disposal of solid wastes. Effects can vary greatly from site to site.

Steam and flash plants emit mostly water vapor (steam). Binary power plants run on a closed-loop system, therefore zero discharge of gases is accomplished. The geothermal industries have developed advanced technologies to recycle minerals in geothermal fluid so that little or no disposal or emissions occur. The examples are found from The Geysers power plants in northern California that separate and use sulfur for sulfuric acid production, and also from the Salton Sea power plants in southern California recycling salts from geothermal brine, recovering silica from mineralized brine for use as fillers in concrete, and extracting zinc for additional plant profitability.

CONCLUSIONS

Geothermal resources are continuously renewable sources of energy regardless of climate or weather conditions, unlike wind or solar energy. Reliability, sustainability, and cleanness make geothermal energy especially attractive as a source for baseload electricity generation or for direct-use applications that need constant heat or energy. Geothermal power plants compete well economically with coal, oil, and nuclear plants in meeting baseload capacity needs with significant environmental advantages. The next generation of geothermal power plants will be designed using long-term projections for resource production as the basis for cycle selection, optimization, and system design.

Current HDR technology is competitive with modern coal-fired plants in regions with geothermal gradients exceeding 60°C/km.[13] Reasonable improvements in reservoir performance or reductions in drilling and completion costs may substantially lower the effective cost of HDR power. In areas with steep geothermal gradients, the use of HDR may demonstrate a substantial cost advantage over coal. This advantage may increase over time allowing the use of HDR to produce a significant portion of the future electricity of the world.

Due to its practicality and low operating costs, direct application of geothermal energy is expected to grow in popularity, especially in geothermally favored regions. Diverse applications are expected to be developed in this field and more advances in GHP technology are also expected. Advances in materials, process integration and design, resource management, instrumentation, and drilling technology will undoubtedly enhance the global utilization of geothermal energy.

REFERENCES

1. Geothermal Resource Group. *Geothermal Resource and Technology in the United States*; National Academy of Sciences: Washington, DC, 1979; Vol. 1.
2. Chermisinoff, P.N.; Morresi, A.C. *Geothermal Energy Technology Assessment*; Technomic Publishing Co., Inc.: Connecticut, 1970.
3. Griffin, R.D. Can renewable energy sources replace fossil fuels? CQ Researcher **1992**, *2* (25), 575–588.
4. McLarty, L.; Reed, M.J. The United States geothermal industry—3 decades of growth. Energy Sources **1992**, *14* (4), 443–455.
5. Wright, P.M. Exploration potential for new hydrothermal resources for electrical power generation in the 48 contiguous United States. Transactions—Geothermal Resources Council **1991**, *15*, 217–28.
6. DOE Office of Energy Efficiency and Renewable Energy (EERE). *United States Geothermal Technology—Equipment and Services for Worldwide Application*; DOE/EE-0044; U.S. Department of Energy: Washington, D.C., 1995; 1–50.
7. National Renewable Energy Laboratory. *Geothermal Energy: 1992 Program Overview*; DOE-CH-10093-182; U.S. Department of Energy: Washington, D.C., 1993; 1–17.
8. Hadfield, P. The geothermal potential of a hot spring bath. New Scientist **1990**, *125* (February), 58.
9. Website of Geothermal Technologies Program, Energy Efficiency and Renewable Energy, U.S. Department of Energy; http://www.eere.energy.gov/geothermal/deployment_gpw.html (accessed 2004).
10. Reed, M.J. Geothermal energy. Geotimes **1991**, *36* (2), 16–18.
11. Reed, M.J. Geothermal energy. Geotimes **1993**, *38* (2), 12–13.
12. Phair, K.A. Getting the most out of geothermal power. Mechanical Engineering **1994**, *116* (9), 76.
13. Haraden, J. The status of hot dry rock as an energy source. Energy **1992**, *17* (8), 777.
14. Tenenbaum, D. Tapping the fire. Technol. Rev. **1995**, *98*, 38–47.
15. Dickson, M.H.; Fanelli, M. Small geothermal resources—a review. Energy Sources **1994**, *16* (3), 349376.
16. Website of Geothermal Heat Pump Consortium, Inc.; http://www.geoexchange.org/about/how.html (accessed 2004).
17. Website of National Renewable Energy Laboratory; http://www.nrel.gov/documents/geothermal_energy.html (accesses 2005).
18. Website of Sandia National Laboratory Geothermal Research; https://cfwebprod.sandia.gov/cfdocs/GPI/program (accessed 2005).
19. Joyce, C. Keeping the world cool with deep heat. New Scientist **1989**, *121* (February), 58.
20. Anderson, I. Drilling deep for geothermal power and science. New Scientist **1986**, *111* (July), 22.
21. Coveney, S.M. Geothermal energy goes universal. Feature Article, Mechanical Engineering **1991**, *113* (January), 10.

Greenhouse Gas Management for Multiplant Complexes

Ralph W. Pike
Department of Chemical Engineering, Louisiana State University, Baton Rouge, Louisiana, U.S.A.

INTRODUCTION

The relationship between greenhouse gases and climate changes has been studied extensively.[1,2] The conclusion is that greenhouse gases can induce global climate warming and emissions should be reduced, especially of carbon dioxide, which accounts for 83% of the total emissions as shown in Fig. 1.[3]

A summary of the two sources of carbon dioxide and the carbon dioxide cycle is illustrated in Fig. 2. Natural sources from plants and soils emit 1.6 Gt C/yr, and anthropogenic sources account for 5.5 Gt C/yr, mainly from the burning of fossil fuels and deforestation as shown in Table 1.[4] The CO_2 emissions from the consumption and flaring of fossil fuels in the United States in 1999 have been estimated as 1526 million metric tons carbon equivalent accounting for 24% of worldwide emissions. The carbon emissions from industrial sources in the United States were reported to be 630 million metric tons carbon equivalent, with 524 and 473 million metric tons carbon equivalent from buildings and transportation, respectively, as shown in Table 1.[5] The manufacturing industries' total was reported to be 402 million metric tons carbon equivalent, and the petroleum and coal products industry and the chemical industry had 175 million metric tons carbon equivalent or 43% of the total.[3] British Petroleum reported that 90% of the CO_2 emissions were from energy generation, 7% from flaring, and 3% from process (noncombustion) sources in refineries.[6] Worldwide, CO_2 emissions from the consumption and flaring of fossil fuels were estimated to be 6320 million metric tons carbon equivalent per year.[9]

There have been extensive discussions about separating, capturing, and converting carbon dioxide to useful products from industrial emissions. These evaluations include sequestering carbon dioxide in geological formations, oceans, and natural systems.[10] These concepts have been summarized by Kim and Edmonds, and they estimated the sequestering costs to range from US $120 to US $340 per metric ton of carbon equivalent.[11] Also, they estimated that this cost could drop to US $50 per ton of carbon equivalent by 2015.

Approximately, 110 million metric tons of carbon dioxide per year (30 million metric tons carbon equivalent) are used as a raw material for the production of urea, methanol, polycarbonates, cyclic carbonates, and specialty chemicals.[8] The largest use is for urea production, which reached about 90 million metric tons per year in 1997.[12]

Ammonia production consumes hydrogen that is obtained from synthesis gas after separating carbon dioxide. However, ammonia plants in the United States produce about 6.8 million tons of carbon dioxide per year, and urea and methanol plants only consume 4.0 million tons per year.[13] This leaves an excess of 2.8 million metric tons of high-purity carbon dioxide per year that is vented to the atmosphere in the United States as shown in Table 1.

Also, there is approximately another 19 million metric tons of relative high-purity carbon dioxide from refineries and other chemical plants in the United States that use hydrogen from synthesis gas and discharge carbon dioxide to the atmosphere.[13,14]

A potential upper limit of carbon dioxide used as a raw material has been estimated by Song.[15] This total of 650 million metric tons of carbon dioxide included traditional processes for urea and methanol in addition to plastics, fibers, rubber, and other uses. This total is comparable to carbon dioxide emissions from fossil fuel power plants.

In the chemical production complex in the lower Mississippi River corridor shown in Fig. 3, there are about 150 plants that consume 1.0 quad (10^{15} BTU/yr) of energy and generate about 215 million pounds of pollutants per year.[16] There is a carbon dioxide pipeline that connects plants. Currently, there is approximately an excess 1.0 million metric tons of high-purity carbon dioxide per year from ammonia production that is being vented to the atmosphere. The cost of carbon dioxide as a raw material is essentially the pumping cost to a plant, about US $2–3 per metric ton.[17]

The chemical production complex in the lower Mississippi River corridor is one of the several worldwide complexes that can benefit from using carbon dioxide as a raw material and the resulting reduced energy consumption. In Fig. 4, a list of some of these

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120018372

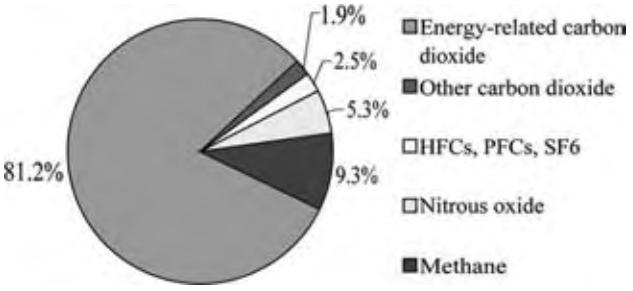


Fig. 1 Greenhouse gas composition in the United States 2002, revised from Energy Information Administration (EIA), 2001. (From Ref.[3].) *(View this art in color at www.dekker.com.)*

complexes is given, and the one in Houston, TX, is the largest complex in the world.

GREENHOUSE GASES AS RAW MATERIALS

There have been a book, five international conferences, and numerous articles in the past 20 yr on carbon dioxide reactions that consider using it as a raw material.[12,15,18–21] The diagram in Fig. 5 is a convenient way to show the range of reactions for carbon dioxide. It can be used as a whole molecule in reactions and can be used as a carbon source or as an oxygen source.

In Fig. 6 carbon dioxide reactions are categorized by industrially important products. Hydrogenation reactions produce alcohols, hydrocarbon synthesis reactions produce paraffins and olefins, and amine synthesis produces methyl and higher-order amines. Hydrolysis reactions can produce alcohols and organic acids. Carbon dioxide serves as an oxygen source in the ethyl benzene to styrene reaction. It can be used in dehydrogenation and reforming reactions.

For the reactions shown in Fig. 6, there are nearly 100 reports of new methods and catalysts to produce these commercially important products.[7] An important reaction is the direct catalytic reaction of carbon dioxide and methane to produce acetic acid, and this reaction has been used in a new process that was evaluated using the Chemical Complex and Cogeneration Analysis System, which is described in the next section.

New processes using carbon dioxide will provide alternate methods for manufacturing chemical intermediates used to produce building materials, motor vehicles, tires, paper, plastics, apparel, fertilizer, furniture, and appliances.[22] Essentially, all of these products are sequestered because at the end of their useful life they end up in landfills. The carbon from fertilizers ends up in plant tissue and will eventually be released to the atmosphere.

EVALUATION OF NEW PROCESSES FOR CARBON DIOXIDE USE

Chemical Complex and Cogeneration Analysis System

The Chemical Complex and Cogeneration Analysis System is an advanced technology for energy conservation and pollution prevention. The System has been developed by industry–university collaboration and used by corporate engineering groups for regional economics, energy, environmental and sustainable development planning to design energy efficient and environmentally acceptable plants and produce new products from greenhouse gases. The System assists in overcoming growth and productivity limitations in the chemical industry by inefficient power generation and greenhouse gas emission constraints. Results from using the System demonstrate how new processes can be integrated into existing chemical complexes to convert greenhouse gases into useful products and to reduce energy consumption and emissions by cogeneration.

This System combines the Chemical Complex Analysis System with the Cogeneration Design System. The Chemical Complex (Multi-Plant) Analysis System is a new methodology to determine the best configuration

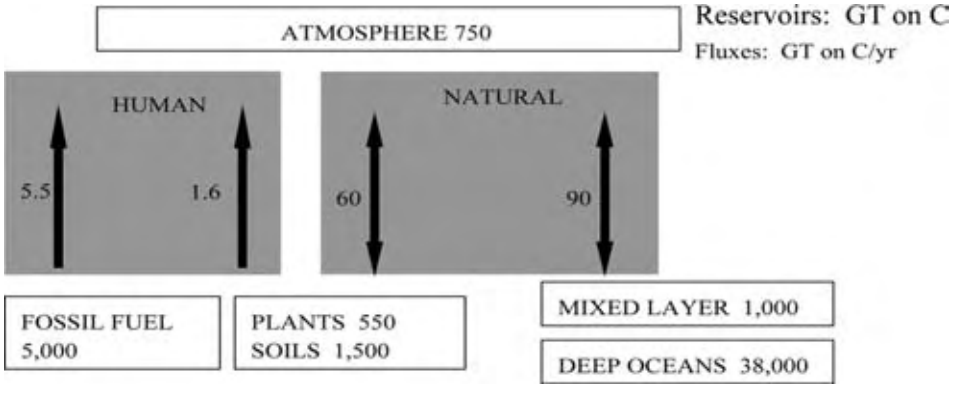


Fig. 2 The carbon cycle, from Intergovernmental Panel on Climate Change (IPCC). (From Ref.[1].)

Table 1 Carbon dioxide emissions and utilization (million metric tons carbon equivalent per year)

CO_2 emissions and utilization		Reference
Total CO_2 added to atmosphere		
Burning fossil fuels	5500	IPCC[1]
Deforestation	1600	
Emissions of CO_2 for several countries and world total		EIA[3]
United States	1526	
China	792	
Russia	440	
Japan	307	
All others	3258	
Total	6323	
U.S. CO_2 emissions		Stringer[5]
Industry	630	
Buildings	524	
Transportation	473	
Total	1627	
U.S. industry (manufacturing)		EIA[3]
Petroleum, coal products, and chemicals	175	
Chemical and refinery (British Petroleum)		McMahon[6]
Combustion and flaring	97%	
Noncombustion direct CO_2 emission	3%	
Excess high-purity CO_2 from ammonia production		Hertwig et al.[7]
U.S.A.	0.76	
Lower Mississippi River corridor	0.18	
CO_2 used in chemical synthesis	30	Arakawa et al.[8]

of plants in a chemical complex based on the AIChE Total Cost Assessment (TCA) for economic, energy, environmental, and sustainable costs, and incorporates the Environmental Protection Agency (EPA) Pollution Index methodology (WAR) algorithm. The Cogeneration Design System examines corporate energy use in multiple plants and determines the best energy use based on economics, energy efficiency, regulatory emissions, and environmental impacts from greenhouse gas emissions. It uses sequential layer analysis to evaluate each plant's current energy use as at an acceptable level or cost-effective improvements if possible. It includes cogeneration as a viable energy option and evaluates the cogeneration system operating optimally.

The structure of the System is shown in Fig. 7, and the System output includes evaluating the optimum configuration of plants in a chemical production complex by determining the maximum profit and minimum energy use and emissions. The input includes incorporating new plants that use greenhouse gases as raw materials in the existing complex of plants. The integrated cogeneration sequential layer analysis determines cost-effective improvements for individual plants using heat exchanger network analysis and cogeneration opportunities. Then, these results are used to determine the optimum complex configuration and utilities integrated with the plants.

The System has an interactive Windows program that integrates existing programs. All interactions with the System are through a graphical user interface designed and implemented with Visual Basic. As shown in Fig. 7, the process flow diagram for the complex is constructed, and equations for the process units and variables for the streams connecting the process units are entered and stored in an Access database using interactive data forms. Material and energy balances, rate equations, and equilibrium relations for the plants are entered as equality constraints using the format of the GAMS programming language, which is similar to Excel, and stored in the database. Process unit capacities, availability of raw materials, and demand for product are entered as inequality constraints and stored in the database. The System takes the equations in the database and writes and runs a GAMS program to solve the mixed integer nonlinear programming problem for the optimum configuration of the complex. Then, the important information from the GAMS solution is presented to the user in a convenient format, and the results can be exported to Excel, if desired. Features for developing flowsheets include adding, changing, and deleting the equations that describe units and streams and their properties. Usual Windows features include cut, copy, paste, delete, print, zoom, reload, update, and grid, among others.

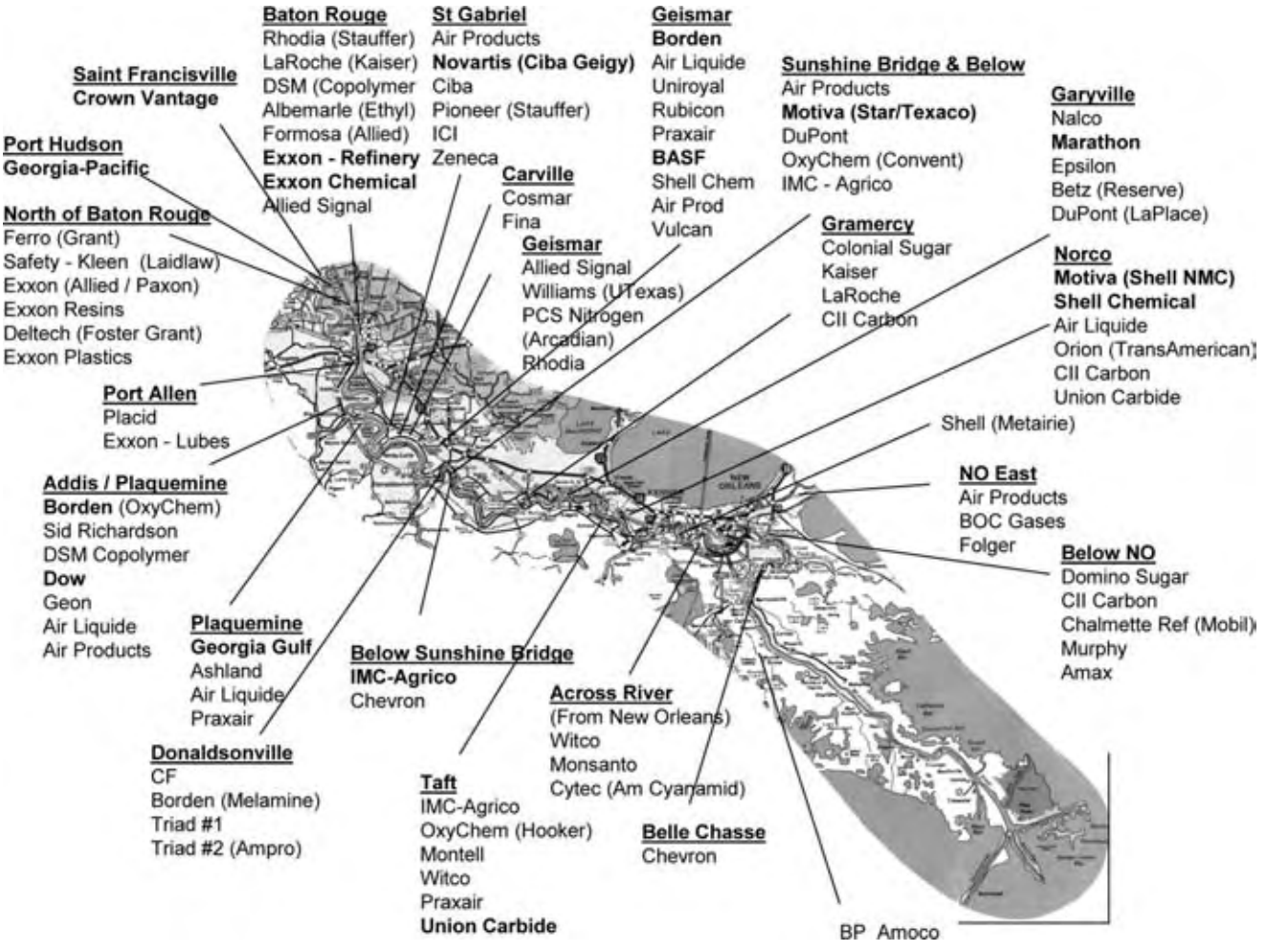


Fig. 3 Plants in the lower Mississippi River corridor. (From Ref.[16].)

A typical window for entering process information is shown in Fig. 8, and in this figure a material balance equation for the acetic acid process, U15, has been entered as an equality constraint. Typical output from the cogeneration analysis is shown in the diagram in Fig. 9 for the results from the prototype. A detailed description of these operations is provided in an interactive user's manual with help files and a tutorial. All of this is available from the Louisiana State University Minerals Processing Research Institute's web site www.mpri.lsu.edu.

The system has been applied to the agricultural chemical production complex in the lower Mississippi River corridor, and the process flow diagram of plants in this complex is shown in Fig. 10 and is called the base case of existing plants. There are 11 production units plus associated utilities for power, steam and cooling water, and facilities for waste treatment. A production unit contains more than one plant; for example, the sulfuric acid production unit contains five plants owned by two companies.[17] This base case includes a standard acetic acid plant with methanol as feedstock, and this is a step to extend the agricultural chemical complex into the petrochemical complex focusing on the CO_2 reuse. For this base case there were 334 equality constraint equations describing the material and energy balances and chemical conversions. Also, there were 19 inequality constraint equations describing the demand for product, availability of raw materials, and range of capacities of the individual plants in the complex.

A diagram comparable to Fig. 10 showing the superstructure of all plants is too detailed to include here, and a convenient way to show the plants in the base case and the plants added to form the superstructure is given in Fig. 11. The superstructure included three options for producing phosphoric acid. Acetic acid can be produced by the conventional process in the base case, and a new catalytic process is included in the superstructure. There are new plants to produce

Continent	Name and Site	Notes
North America	•Gulf coast petrochemical complex in Houston area (U.S.A.) and •Chemical complex in the Baton Rouge-New Orleans Mississippi River Corridor (U.S.A.)	•Largest petrochemical complex in the world, supplying nearly two-thirds of the nation's petrochemical needs
South America	•Petrochemical district of Camacari-Bahia (Brazil) •Petrochemical complex in Bahia Blanca (Argentina)	•Largest petrochemical complex in the southern hemisphere
Europe	•Antwerp port area (Belgium) •BASF in Ludwigshafen (Germany)	•Largest petrochemical complex in Europe and world wide second only to Houston, Texas •Europe's largest chemical factory complex
Asia	•The Singapore petrochemical complex in Jurong Island (Singapore) •Petrochemical complex of Daqing Oilfield Company Limited (China) •SINOPEC Shanghai Petrochemical Co. Ltd. (China) •Joint-venture of SINOPEC and BP in Shanghai under construction (2005) (China) •Jamnagar refinery and petrochemical complex (India) •Sabic company based in Jubail Industrial City (Saudi Arabia) •Petrochemical complex in Yanbu (Saudi Arabia) •Equate (Kuwait)	•World's third largest oil refinery center •Largest petrochemical complex in Asia •World's largest polyethylene manufacturing site •World's largest & most modern for producing ethylene glycol and polyethylene
Oceania	•Petrochemical complex at Altona (Australia) •Petrochemical complex at Botany (Australia)	
Africa	petrochemical industries complex at Ras El Anouf (Libya)	one of the largest oil complexes in Africa

Fig. 4 Some major chemical production complexes in the world. (From Ref.[7].)

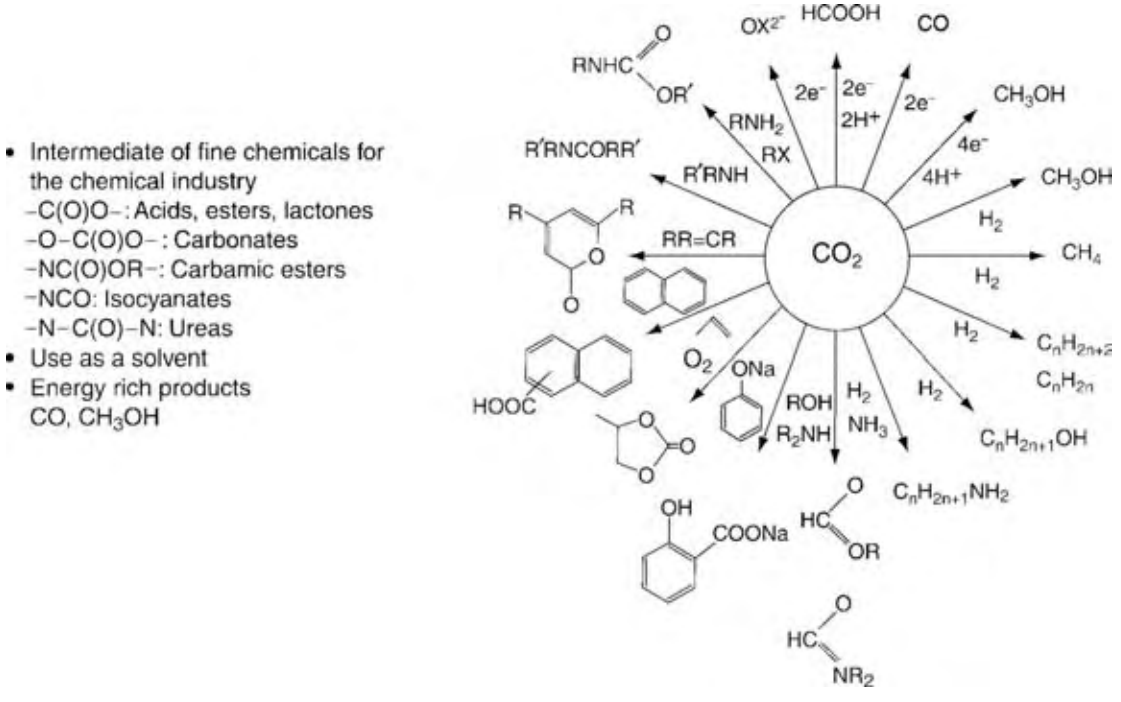


Fig. 5 Utilization of carbon dioxide in synthetic chemistry. (From Ref.[12].)

Hydrogenation

$CO_2 + 3H_2 \rightarrow CH_3OH + H_2O$ methanol
$2CO_2 + 6H_2 \rightarrow C_2H_5OH + 3H_2O$ ethanol
$CO_2 + H_2 \rightarrow CH_3\text{-}O\text{-}CH_3$ dimethyl ether

Hydrolysis and Photocatalytic Reduction

$CO_2 + 2H_2O \rightarrow CH_3OH + O_2$
$CO_2 + H_2O \rightarrow HC{=}O\text{-}OH + 1/2O_2$
$CO_2 + 2H_2O \rightarrow CH_4 + 2O_2$

Hydrocarbon Synthesis

$CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O$ methane and higher HC
$2CO_2 + 6H_2 \rightarrow C_2H_4 + 4H_2O$ ethylene and higher olefins

Carboxylic Acid Synthesis

$CO_2 + H_2 \rightarrow HC{=}O\text{-}OH$ formic acid
$CO_2 + CH_4 \rightarrow CH_3\text{-}C{=}O\text{-}OH$ acetic acid

Graphite Synthesis

$CO_2 + H_2 \rightarrow C + H_2O$
$CH_4 \rightarrow C + H_2$
$CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O$

Other Reactions

CO_2 + ethylbenzene $\rightarrow$ styrene
dehydrogenation of propane
$CO_2 + C_3H_8 \rightarrow C_3H_6 + H_2 + CO$
reforming
$CO_2 + CH_4 \rightarrow 2CO + H_2$

Amine Synthesis

methyl amine and higher amines

Fig. 6 Some catalytic reactions of CO_2 from various sources. (From Ref.[7].)

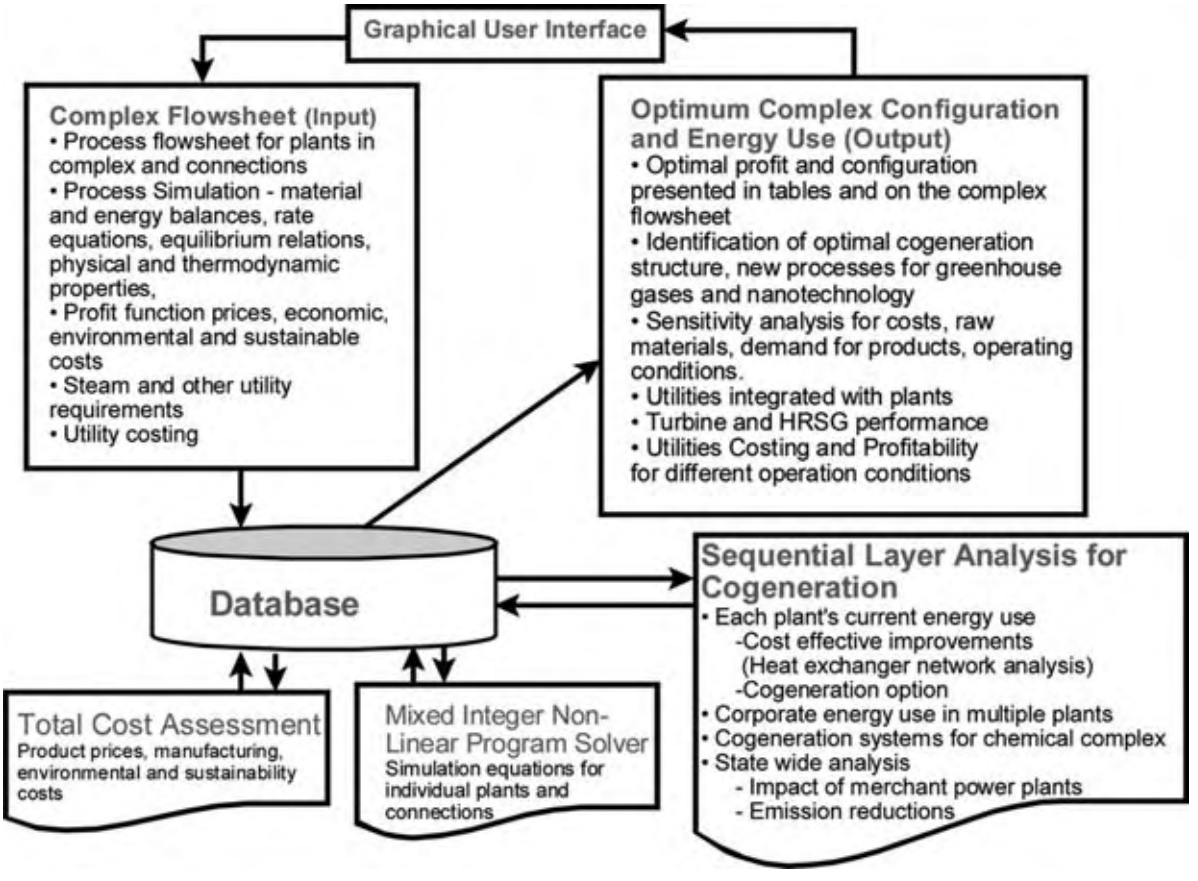


Fig. 7 Structure of the chemical complex and cogeneration analysis system. (From Ref.[17].) *(View this art in color at www.dekker.com.)*

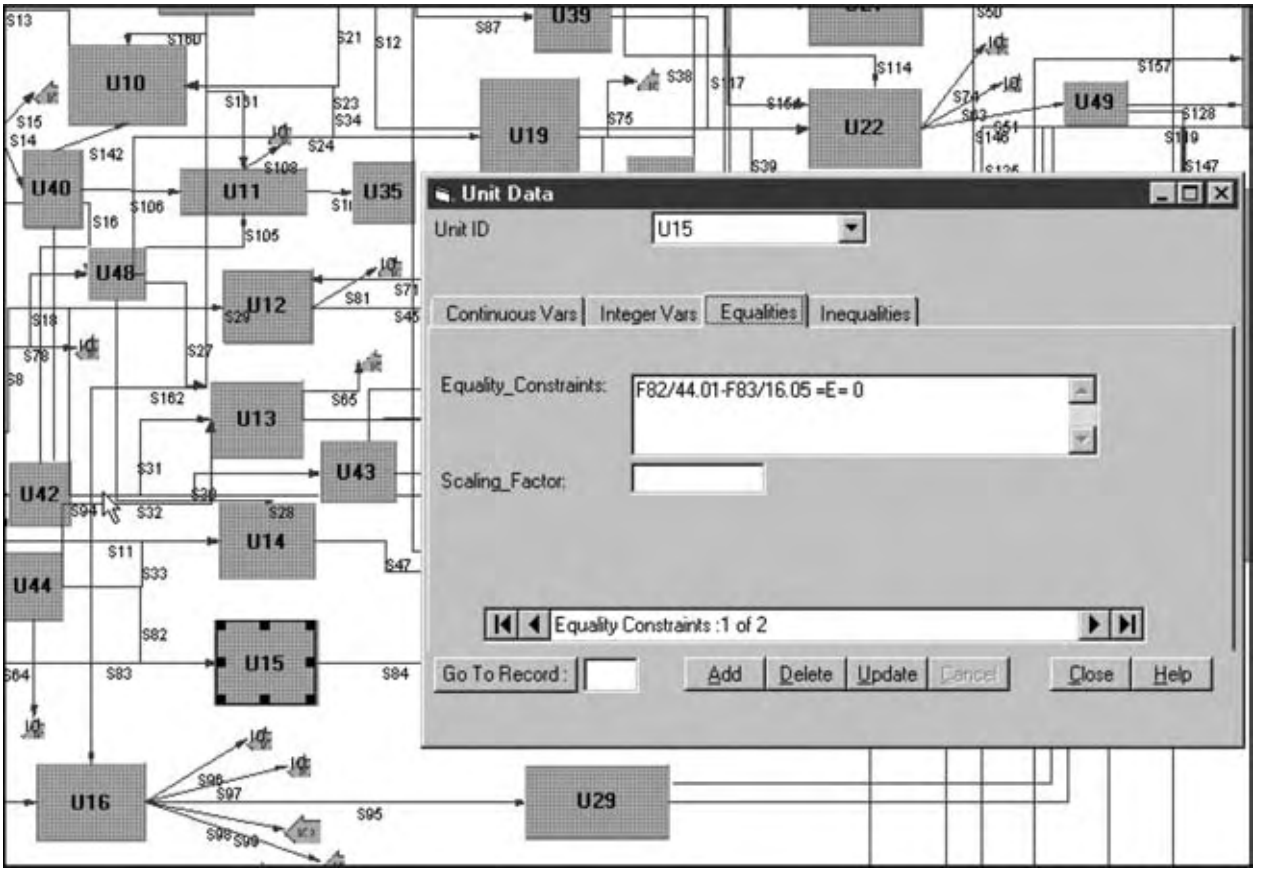


Fig. 8 Illustration of the input to the system for unit data. (From Ref.[17].) *(View this art in color at www.dekker.com.)*

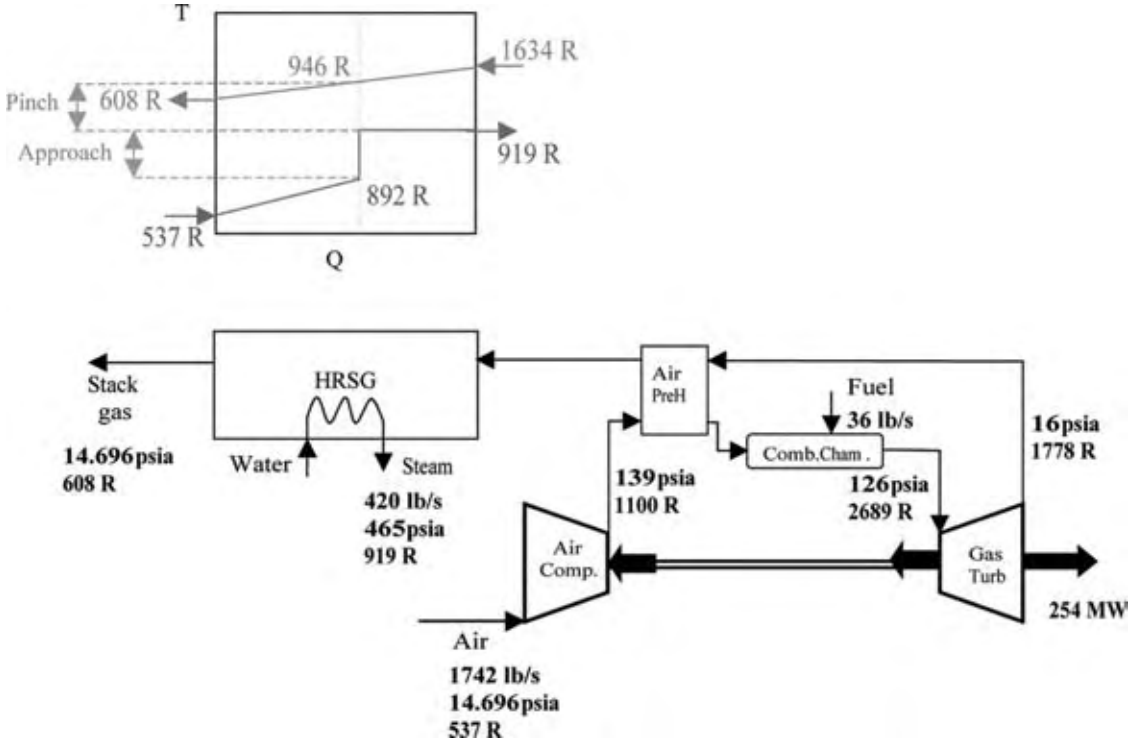


Fig. 9 Typical output from the cogeneration analysis. (From Ref.[17].) *(View this art in color at www.dekker.com.)*

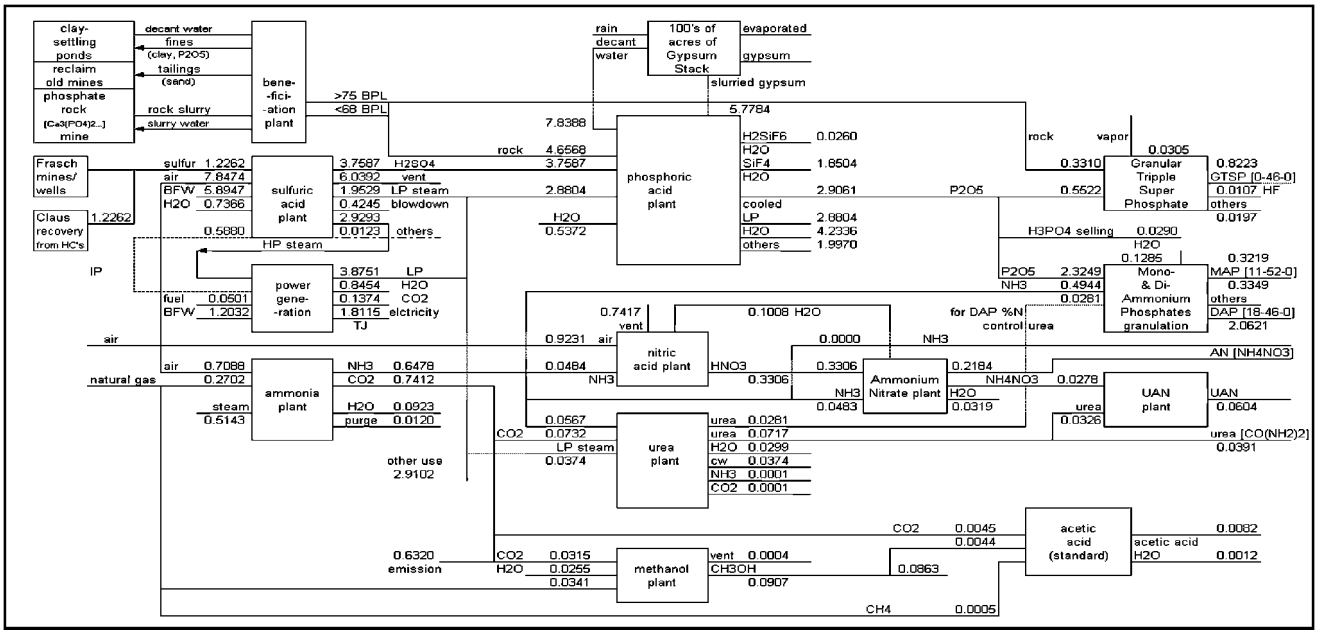


Fig. 10 Base case of plants in the chemical production complex in the lower Mississippi River corridor. (From Ref.[22].)

ammonium sulfate and to recover sulfur and sulfur dioxide. The model of the superstructure has 594 continuous variables, 7 integer variables, 505 equality constraint equations for material and energy balances, and 27 inequality constraints for availability of raw materials, demand for product, and capacities of the plants in the complex.

The System solves the mixed integer nonlinear programming problem using the economic data shown in Fig. 12. This table gives the sale prices for products and costs of raw material that were used in the economic model of the complex. Also shown are sustainable costs and credits. Environmental costs were estimated as 67% of the raw material costs, which is based on the data provided by Amoco, DuPont, and Novartis in the American Institute of Chemical Engineers/ Center for Waste Reduction Technologies (AIChE/ CWRT) report.[23] This report lists environmental costs as approximately 20% of the total manufacturing costs and raw material costs as approximately 30% of total manufacturing costs. Sustainable costs were estimated from the results given for power generation in the AIChE/CWRT report where carbon dioxide emissions had a sustainable cost of US $3.25 per ton of carbon dioxide. A cost of US $3.25 per ton was charged as a cost to plants that emit carbon dioxide, and plants that consume carbon dioxide were given credit of twice this cost or US $6.50 per ton. This credit was included for steam produced from waste heat by the sulfuric acid plant displacing steam produced from a package boiler firing hydrocarbons and emitting carbon dioxide. These costs are arbitrary but a conservative approach. Emissions trading costs of carbon dioxide are about US $50 per ton, and costs to sequester carbon dioxide have been estimated to be US $120–340 per ton.

Results from the System are shown in Fig. 13 as a comparison of the base case and optimal structure

Processes in Superstructure	
Processes in Base Case	Electric furnace process for phosphoric acid
Ammonia	HCl process for phosphoric acid
Nitric acid	Ammonium sulfate
Ammonium nitrate	SO_2 recovery from gypsum process
Urea	S & SO_2 recovery from gypsum process
UAN	Acetic acid–new catalytic process
Methanol	
Granular triple super phosphate	
MAP & DAP	
Power generation	
Contact process for Sulfuric acid	
Wet process for phosphoric acid	
Acetic acid-standard method	

Fig. 11 Processes in the base case and superstructure of the chemical production complex in the lower Mississippi River corridor. (From Ref.[22].)

Raw Materials	Cost ($/T)	Raw Materials	Cost ($/T)	Products Price ($/T)	
Natural Gas	245	Market cost for short term purchase		Ammonia	190
Phosphate Rock				Methanol	96
wet process	27	Reducing gas	1394	Acetic Acid	623
electrofurnace	24	Wood gas	634	GTSP	142
HCl process	25	Sustainable Costs and Credits		MAP	180
GTSP process	30	Credit for CO_2 Consumption	6.50	DAP	165
HCl	50	Debit for CO_2 Production	3.25	NH_4NO_3	153
Sulfur		Credit for HP Steam	10	UAN	112
Frasch	42	Credit for IP Steam	6.4	Urea	154
Claus	38	Credit for gypsum Consumption	5	H_3PO_4	320
C electrofurnace	760	Debit for gypsum Production	2.5	$(NH_4)_2SO_4$	187
		Debit for NO_x Production	1025		

Fig. 12 Raw material costs, product prices, and sustainable costs. (From Green Market Sheet, AIChE/TCA Report, and Chemical Market Reporter.)

for the chemical production complex. The profit that includes the economic, environmental, and sustainable costs increased about 66.74% from the base case to the optimal solution. Also, environmental cost decreased about 31.27%, and sustainable costs increased about 9.18%. Energy requirements decreased from 2912 to 1344 TJ/yr. The standard acetic acid plant consuming more energy in the base case was replaced by the new acetic acid plant in the optimal solution to reduce energy consumption. The system selected plants for the complex with less energy requirement. Production rates for the products in the optimal solution were constrained by their capacity limit. It was optimal to operate the ammonium sulfate. The energy requirement of the ammonium nitrate plant based on the unit product in the optimal structure was different from the base case. The reasons are the different production rates of two types of ammonium nitrate, ammonium nitrate solution and granular ammonium nitrate.

SELECTION OF NEW PROCESSES FOR CARBON DIOXIDE USE

For the reactions shown in Fig. 6, there are nearly 100 published articles of laboratory experiments for new methods and catalysts to produce these commercially important products.[7] These articles summarize the types of reactions with values for the heat of reaction, ΔH°, and the Gibbs free energy, ΔG°. Negative values for ΔH° indicate that a reaction is exothermic, i.e., heat is released; and positive values indicate that a reaction is endothermic, i.e., heat is absorbed. The Gibbs free energy, ΔG°, is related to the equilibrium constant for a chemical reaction, K, by the equation, $\Delta G^\circ = -RT\ln K$, where $\Delta G^\circ = \Sigma \nu_i G_i^\circ$ is the difference between the Gibbs energies, G_i°, of the products and reactants (weighted by their stoichiometric coefficients, ν_i). The reactants and products are in their standard state as a pure substance at the standard-state pressure but at the system temperature.[24]

In an ideal solution, for example, the equilibrium constant $K = \pi(x_i)^{\nu_i}$, where x_i is the mole faction of the products and reactants. This equation is equivalent to writing the concentration of the products in the numerator and the concentration of the reactants in the denominator of a fraction. If $\Delta G^\circ = 0$, then $K = 1$, and the reaction proceeds to a considerable extent before equilibrium is reached. The extent of reaction is smaller at equilibrium if ΔG° increases in a positive direction [$K = \exp(-\Delta)G^\circ/RT$)]. According to Dodge (1944) reactions are said to be less feasible as ΔG° increases in a positive direction, but there is no definite value that can be chosen as clearly indicating that a reaction is not feasible from the standpoint of industrial operations. For example, the methanol synthesis reaction is +11,000 cal/gm mol at 600 K, and this reaction is certainly feasible. Dodge provides the following guidelines for the purpose of ascertaining quickly and only approximately if any given reaction is promising at a given temperature:[25]

- The $\Delta G^\circ > 0$ reaction is promising.
- The $0 < \Delta G^\circ < 10{,}000$ cal/gm mol reaction is doubtful but warrants further study.

		Base case		Optimal structure	
Profit (U.S.$/year)		148,087,243		246,927,825	
Environmental cost (U.S.$/year)		179,481,000		123,352,900	
Sustainability cost (U.S.$/year)		-17,780,800	energy	-16,148,900	energy
Plant name	Capacity (t/year)	Capacity	requirement	Capacity	requirement
	(upper-lower bounds)	(t/year)	(TJ/year)	(t/year)	(TJ/year)
Ammonia	329,030-658,061	647,834	3,774	658,061	3,834
Nitric acid	0-178,547	178,525	-649	89,262	-324
Ammonium nitrate	113,398-226,796	226,796	116	113,398	26
Urea	49,895-99,790	99,790	127	49,895	63
Methanol	90,718-181,437	90,719	1,083	90,719	1,083
UAN	30,240-60,480	60,480	0	60,480	0
MAP	0-321,920	321,912		160,959	
DAP	0-2,062,100	2,062,100	2,127	1,031,071	1,063
GTSP	0-822,300	822,284	1,036	411,150	518
Contact process sulfuric acid	1,851,186-3,702,372	3,702,297	-14,963	2,812,817	-11,368
Wet process phosphoric acid	697,489-1,394,978	1,394,950	7,404	697,489	3,702
Electric furnace phosphoric acid	697,489-1,394,978	na	na	0	0
HCl to phosphoric acid	697,489-1,394,978	na	na	0	0
Ammonium sulfate	0-2,839,000	na	na	1,295,770	726
Acetic acid (standard)	0-8,165	8,165	268	0	0
Acetic acid (new)	0-8,165	na	na	8,165	92
SO_2 recovery from gypsum	0-1,804,417	na	na	0	0
S & SO_2 recovery from gypsum	0-903,053	na	na	0	0
Ammonia sale		0		0	
Ammnium Nitrate sale		218,441		105,043	
Urea sale		39,076		3,223	
Wet process phosphoric acid sale		13,950		6,975	
Methanol sale		86,361		90,719	
Total energy requirement from fuel gas			2,912		1,344

Fig. 13 Comparison of the base case and optimal structure for the chemical production complex. (From Ref.[22].)

- The $\Delta G^\circ < 10{,}000$ cal/gm mol reaction is unfavorable and would be feasible only under unusual circumstances.

For reactions with a positive ΔG° there are ways to increase the conversion. One standard method is to remove products in an intermediate step. This procedure is used in the double absorption contact process for sulfuric acid where sulfur trioxide is removed after the gas passes through two-packed bed reactors before entering the last two reactors.

The basis for this analysis is given in Fig. 14 and includes conversion and selectivity, cost and life (time onstream to deactivation), and methods to regenerate catalysts; hydrogen consumption for hydrogenation reactions; energy requirement for reactions, ΔH°, and equilibrium conversion based on Gibbs free energy, ΔG°. Moreover, the following are considered: raw material costs, energy, environmental, sustainable, and other manufacturing costs, product sales, and market penetration, among others.

The procedure to evaluate a potential process for incorporation into the Chemical Complex Analysis System is shown in Fig. 15. A flowsheeting program, HYSIS or other comparable one, is used to develop the process flow diagram. The flowsheeting program determines the operating conditions and the utilities required—steam and cooling water. Then, a value-added economic analysis is performed to estimate the profitability of the plant. If the profitability is acceptable,

Catalyst
- **conversion and selectivity**
- **life (time on stream to deactivation)**
- **methods to regenerate**
- **hydrogen consumption for hydrogenation reactions**
- **energy requirement for reactions,ΔH°**
- **equilibrium conversion based on Gibbs free energy, ΔG°**

Sales
- **product sales and market penetration**

Costs
- **catalyst, raw materials, energy, environmental, sustainable and other manufacturing costs**

Fig. 14 Bases for analysis of new energy efficient processes using carbon dioxide as a raw material. (From Ref.[7].)

- Simulate process using HYSYS.
- Estimate utilities required.
- Perform economic analysis.
- Obtain process constraint equations from HYSIS simulation.
- Maximize the profit function to find the optimum process configuration with the system.
- Incorporate into superstructure.

Fig. 15 Procedure for new process evaluation for use in the Chemical Complex and Cogeneration Analysis System.

then the process is entered in the System using the material and energy balances, rate equations, and equilibrium relations. These are entered as equality constraints, and demand for product, availability of raw material, and capacities of the process units are entered as inequality constraints. Results from the System give the optimum configuration of the process. Then, this information is included in the superstructure of the complex.

There are four areas for potentially significant energy savings that can be used for selecting designs to evaluate new energy-efficient processes for carbon dioxide consumption:

1. Potential energy savings through improved catalyst using Pacific Northwest National Laboratory (PNNL) estimates for 26 commercial chemicals with 14 of these chemicals shown in Fig. 16.[26]
2. Actual vs. theoretical energy consumption for 16 chemicals and 3 industry groups as shown in Fig. 17 and energy performance levels in Fig. 18.[26,27]
3. Replacement of hazardous substances: An example is a new route to dimethyl carbonate by reacting carbon dioxide and methanol. The current route that reacts carbon monoxide and chlorine to produce phosgene, an extremely toxic intermediate, which is then reacted with methanol to produce dimethyl-carbonate and hydrochloric acid.[15]
4. Other potential uses: For example, applications as a solvent in separations and reactions, expanding markets for chemicals from carbon dioxide, enhanced recovery of oil, natural gas, and coal bed methane.

For the first area, the Chemical Industries of the Future (IOF) Report identified the potential energy savings though improved catalysts for 26 chemicals. This information is given in Fig. 18, and it shows the rank and energy savings.[26] Acetic acid is on this list with a potential savings of 2.0 trillion BTU/yr. Based on this potential, a new acetic acid process has been evaluated as discussed below, and energy savings were found to be 275 billion BTU/yr based only on heat and power savings for a 100 million pound per year plant. Using this new process to replace the 5540 million pounds per year of U.S. acetic acid capacity could generate 15 trillion BTU/yr in energy savings.[28]

The second method of evaluation for actual vs. theoretical energy requirements will use the results in the Chemical IOF Report and the AIChE/CWRT report on Energy Level Performance by the Chemical Industry (Figs. 17 and 18).[26,27] For example, the IOF report states that production of ethylene glycol, vinyl chloride, and styrene all require 5–10 times the theoretical energy, and by industry groups, organic chemicals require about three times the theoretical energy, as shown in Fig. 17. In the energy efficiency evaluations in Fig. 18, it was shown that optimized energy integration can reduce energy consumption by 10% and aggressive process redesign can reduce energy consumption by 10–50%.

	Energy Consumed in Production				Potential Energy Savings	
	Heat & Power Feedstock					
Chemical	Rank	Energy	Energy	Rank	Savings	Energy
Ammonia	2	290	460	1	294	
Propylene	3	160	540	2	98	
p-Xylene	13	60	140	3	94	
Butadiene	20	70	80	4	81	
Vinyl Chloride	5	100	300	5	44	
Methanol	–	–	–	6	37	
Ethylene Oxide	18	0	160	7	29	
Styrene	7	50	280	10	20	
Phenol	15	20	160	13	12	
Ethylene Dichloride	9	30	290	14	11	
Formaldehyde	–	–	–	17	6	
Ethylbenzene	–	–	–	18	4	
Acetic Acid	–	–	–	20	2	
Ethylene Glycol	–	–	–	24	1	

Fig. 16 Energy consumed in the production of major chemicals and potential energy savings through improved catalysts for potential products from carbon dioxide (trillion BTUs). (From Ref.[26].)

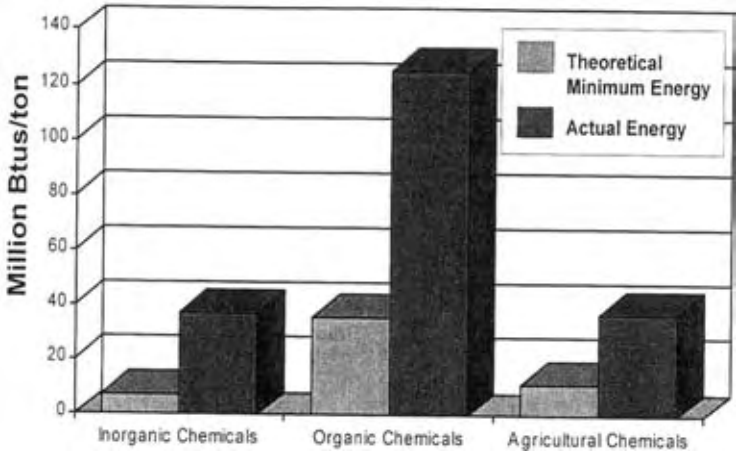


Fig. 17 Theoretical energy minima vs. operating practice in selected chemical industry segments. (From Ref.[26].)

CONCLUSIONS

These results illustrated the capability of the Chemical Complex and Cogeneration Analysis System to select an optimum configuration of plants in an agricultural chemical complex and incorporate economic, environmental, and sustainable costs. With this System, engineers have a new capability to consider projects in depths significantly beyond current capabilities to convert the company's goals and capital into viable projects that are profitable and to meet energy and environmental requirements by developing and applying a regional methodology for cogeneration, and conversion of greenhouse gases to saleable products. The System is available at no charge from the Louisiana State University Minerals Processing Research Institute's web site www.mpri.lsu.edu along with a user's manual with help files and a tutorial.

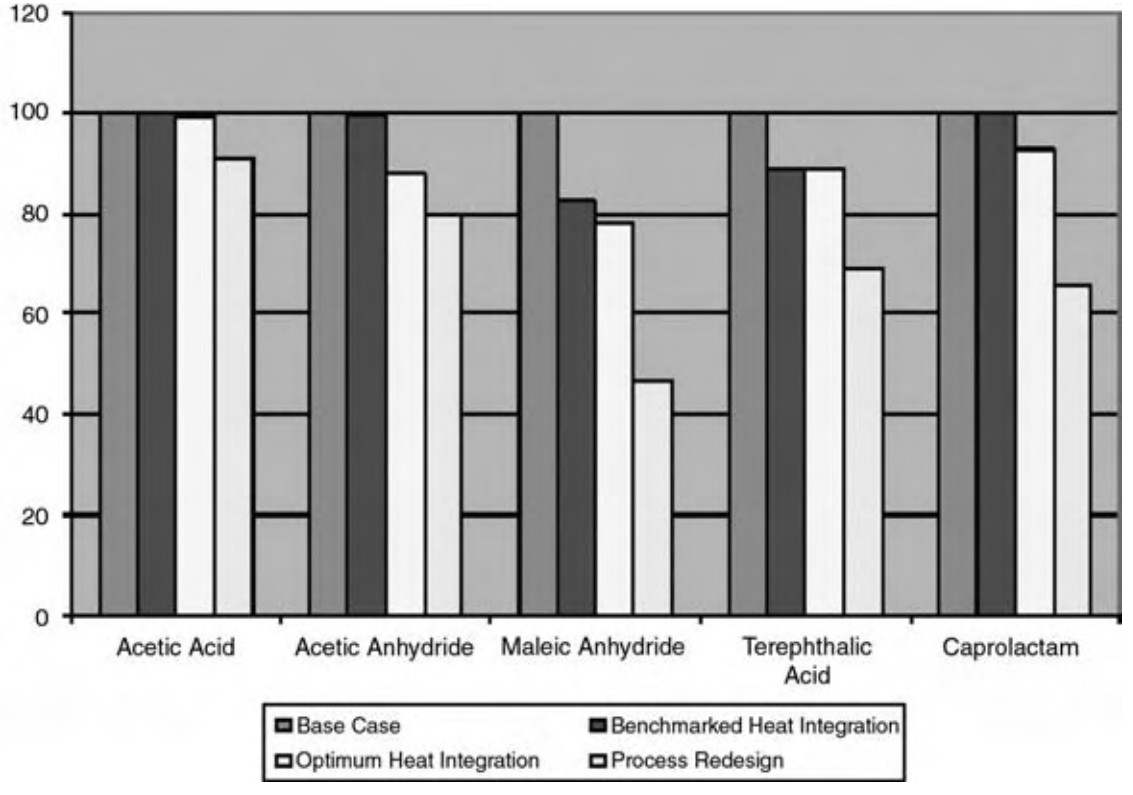


Fig. 18 Energy performance levels for four processes. (From Ref.[27].) *(View this art in color at www.dekker.com.)*

REFERENCES

1. IPCC. *Climate Change 2001, The Scientific Basis*; Intergovernmental Panel on Climate Change. Cambridge University Press: Cambridge, U.K., 2001.
2. NRC. *Carbon Management: Implications for R & D in the Chemical Sciences and Technology*: National Research Council. National Academy Press: Washington, DC, 2001.
3. EIA. *Emissions of Greenhouse Gases in the United States 2000*, Department of Energy Report DOE/EIA-0573. EIA: Washington, DC, 2001.
4. Parsons, M.L. *Global Warming*; Plenum Press: New York, 1995.
5. Stringer, J.C. *Opportunities for Carbon Control in the Electric Power Industry, Carbon Management: Implications for R & D in the Chemical Sciences and Technology*, A Workshop Report to the Chemical Sciences Roundtable; National Academy Press: Washington, DC, 2001.
6. McMahon, M. *Technical Aspects of Measuring Emissions in the Petroleum Industry,* Technical Report; BP Amoco: U.K., 1999.
7. Hertwig, T.A.; Xu, A.; Ozyurt, D.B.; Indala, S.; Pike, R.W.; Knopf, F.C.; Hopper, J.R.; Yaws, C.L. Development of new processes for greenhouse gases management in multi-plant, chemical production complexes. Proceedings NATO CCMS Study Clean 2003 Annual Meeting, Cetraro, Italy, May 11–15, 2003.
8. Arakawa, H. Catalyst research of relevance to carbon management: progress, challenges and opportunities. Chem. Rev. **2001**, *101*, 953–996.
9. EIA. *International Energy Annual Report 2000*; Office of Energy Markets and End Use. U.S. Department of Energy: Washington, DC, 2002.
10. NRC. *Novel Approaches to Carbon Management: Separation, Capture, Sequestration and Conversion to Useful Products*, Workshop Report of the Committee on Novel Approaches to the Management of Greenhouse Gases from Energy Systems; National Research Council, National Academy Press: Washington, DC, 2003.
11. Kim, S.H.; Edmonds, J.A. *Potential for Advanced Carbon Capture and Sequestration Technologies in a Climate Constrained World*, PNNL-13095; Pacific Northwest National Laboratory, 2000.
12. Creutz, C.; Fujita, E. Carbon dioxide as a feedstock. In *Carbon Management: Implications for R&D in the Chemical Sciences and Technology*; Bell, A.T., Marks, T.J., Eds.; National Academy Press: Washington, DC, 2000.
13. Wells, G.M. *Petrochemicals and Processes*, 2nd Ed.; Ashgate Publishing Company: Brookfield, Vermont, 1999.
14. Moulijn, J.A.; Makkee, M.; Van Diepen, A. *Chemical Process Technology*; Wiley: New York, 2001.
15. Song, C.; Gafney, A.F.; Fujimoto, K.F. *CO_2 Conversion and Utilization*; ACS Symposium Series 809; American Chemical Society, Oxford University Press: Washington, DC, 2002; Chapter 1.
16. Peterson, R.W. *Giants on the River*; Homesite Company: Baton Rouge, LA, 1999.
17. Hertwig, T.A.; Xu, A.; Nagy, A.; Hopper, J.R.; Yaws, C.L. A prototype system for economic, environmental and sustainable optimization of a chemical complex. Clean Technol. Environ. Policy **2002**, *3* (4), 363–370.
18. Halmann, M.M.; Steinberg, M. *Greenhouse Gas Carbon Dioxide Mitigation: Science and Technology*; Lewis Publishers, CRC Press: Boca Raton, FL, 1999.
19. Inui, T. Advances in chemical conversions for mitigating carbon dioxide. In *Studies in Surface Science and Catalysis*, Proceedings of the Fourth International Conference on Carbon Dioxide Utilization; Elsevier Science Publishers: Amsterdam, 1998; Vol. 114.
20. Sullivan, B.P.; Krist, K.; Guard, H.E. *Electrochemical and Electrocatalytic Reactions of Carbon Dioxide*; Elsevier Science Publishers: New York, 1993.
21. Inoue, S.; Yamazaki, N. *Organic and Bio-organic Chemistry of Carbon Dioxide*; John Wiley & Sons: New York, 1982.
22. C&E News. Production: down but not out. C&E News **2002**, Jun 24, 60–65.
23. Constable, D. *Total Cost Assessment Methodology, Internal Managerial Decision Making Tool*; AIChE/CWRT: New York, 1999.
24. Smith, J.M.; Van Ness, H.C.; Abbott, M.M. *Introduction to Chemical Engineering Thermodynamics*, 6th Ed.; McGraw-Hill: New York, 2001; 475 pp.
25. Dodge, B.F. *Chemical Engineering Thermodynamics*; McGraw-Hill Book Company: New York, 1944.
26. Pelegrino, J.L. *Energy and Environmental Profile of the U.S. Chemical Industry*; U.S. DOE; Office of Industrial Technologies: Washington, DC, 2000.
27. Beaver, E. *A Pilot Study for Energy Performance Levels for the U.S. Chemical Industry*, Final Report on DOE Contract DE-AC05-00OR22725; AIChE/CWRT, 2001; www.aiche.org.
28. Xu, A.; Pike, R.W.; Hertwig, T.A.; Indala, S.; Knopf, F.C.; Hopper, J.R.; Yaws, C.L. Integrated chemical complex and cogeneration analysis system: energy conservation and greenhouse gas management solutions.Proceedings of the Sustainable Engineering Topical Conference, Sustainable Design Methodology, Paper No. 19f; AIChE Annual Meeting, Nov 3–8, 2002, Indianapolis, IN.

Heat Exchanger Operation and Troubleshooting

T. Reg. Bott
School of Engineering, Chemical Engineering, University of Birmingham, Birmingham, U.K.

INTRODUCTION

Heat exchangers are employed in many industrial processes, including petroleum refining, large-scale chemical and plastics manufacture, pharmaceutical production, power generation, food processing, desalination, metal production, etc. It is probably true that the manufacture of virtually all products available to modern society involves the transfer of heat. Apart from the need to control temperatures, a heat exchanger has a vital role in process economics and conserving energy, with the associated implications for the environment and global warming.

It is imperative that a heat exchanger is maintained and operated efficiently and effectively to meet all the criteria. As most operate on a continuous basis 24 hr a day, 365 days a year—the greatest hindrance to efficient operation is that the heat-transfer surfaces become dirty, usually referred to as ''fouling,'' with consequent reduction in heat-transfer capability. The choice of heat exchanger and its basic design are fundamental to satisfactory operation with the minimum of trouble. It sometimes represents optimization between operating and initial capital costs. Design of shell and tube heat exchangers must recognise the potential problem of tube failure from metal fatigue and fretting, due to tube oscillation in response to the fluid flow conditions.

It is, however, generally heat exchanger fouling that leads to difficulties in operation. This fouling leads to an increased pressure drop with the potential to produce inlet pressures greater than expected tolerance. As a result leakages may occur between streams, leading to contamination problems and even increased fouling. The increased pressure drop may also increase the maintenance requirements of the associated pumps. A further problem can be the development of corrosive conditions beneath the deposit.

Changes in operating conditions due to changed product demand or changed feed stock quality must be carefully considered to prevent the introduction of operating problems, particularly as these changes can affect the extent of surface fouling.

DESIGN, COMMISSIONING, AND OPERATION OF HEAT EXCHANGERS

Clearly the choice of heat exchanger and its basic design have an impact on its operation and ease of maintenance, particularly the extent of the fouling that accrues and its effect on efficiency. If severe fouling problems are anticipated, it may be necessary to specify a self-cleaning design such as a fluidized bed heat exchanger. It has to be recognized that most heat exchangers are unique in terms of process conditions and the quality of the material being processed, so that the extent of actual fouling is not accurately known at the design stage. Furthermore, during the life of the heat exchanger, the operating conditions may be modified in response to changes in the required product specification or in the origin of the raw material.

An allowance for anticipated fouling is incorporated into the design as an additional resistance to heat transfer. It could be argued that the choice of fouling allowance (often referred to as fouling factor) is little more than an inspired guess. The best guide for fouling allowance is experience, but for reasons already mentioned, experience may not be directly applicable to the design in hand.

To facilitate the design of shell and tube heat exchangers, the Tubular Exchanger Manufacturers Association (TEMA) in the U.S.A. has issued recommended fouling allowances for a variety of heat exchangers. In a survey[1] of the design and operation of tubular heat exchangers in the petroleum refining industry, six of nine heat exchangers were over-designed; one under-designed; one of a reasonable design; and one over-designed at the start of operations and under-designed at the end. An alternative to recommended values of fouling resistance is to use the company experience obtained from previous operations or to review published data. Careful scrutiny,[1] however, revealed many shortcomings in the use of published data relating to petroleum refining.

Allowance for fouling in design results in an increase of heat-transfer area and over-performance of the heat exchanger during early stages of operation. To provide the required temperature of, say, the cold stream, it might be necessary to reduce the velocity in the hot stream, but this could have a serious detrimental effect on the fouling of the heat-transfer surface on the hot side. It might be prudent, therefore, to recommend a ''pump around'' system so that velocity is maintained at the design level, but temperatures are controlled by suitable valve adjustment.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016837

During commissioning of the process plant, it is important to avoid subjecting the heat exchanger to conditions different from those taken into account during the design; high temperature and low velocity are to be avoided as this may create deposits that are difficult to remove.

Once the plant is accepted and fully operational, it becomes important to operate heat exchangers according to the conditions laid down in the design. Again attention to the temperature of operation and the fluid velocities is necessary to avoid accelerated fouling. Excursions of temperature particularly, even over short periods of time, may give rise to serious problems of deposits that are difficult to remove, affect performance, and are costly to remedy. Low velocity also may give rise to similar problems.

Slack management of operations and poor maintenance may give rise to accelerated and unexpected fouling. For instance, leaks across a heat exchanger, from one stream to the other, may introduce an unexpected and severe fouling problem, because of corrosion or chemical reactions. Again this can result in poor performance and may be expensive to rectify.

FOULING CONTROL

In general, the major problem in heat exchanger operation is the accumulation of deposits on surfaces. To avoid taking the exchanger out of service for cleaning, on-line fouling control is often employed. The choice of method will, of course, very much depend on the fouling problem encountered. There are two basic methods: control may be achieved by the use of chemical additives and by physical methods, or in some instances, with a combination of the two.

Every fouling problem is likely to be unique, so that it is impossible to cover all counter-fouling techniques as they apply to specific examples. The general approach, however, will indicate the technologies available and their potential application.

Chemical Additives

Chemical additives have been developed to tackle different fouling problems based on the mechanisms involved. In some instances, proprietary additives are designed to combat mixed fouling where two or more mechanisms are present. In the following, additives associated with individual mechanisms only are covered.

Particle deposition

The discussion will be centered on cooling water that may contain particulate matter, but the technology may be applied to other liquid systems.

The particles in the fluid passing through a heat exchanger are not always a problem, provided that settlement is avoided by imposing a velocity that ensures that they are kept in suspension. Avoidance of ''dead spots'' in the design is essential where particulate deposition is a possibility. Particle deposition often occurs in combination with other deposit-forming mechanisms, such as biofouling or scale formation. Particulate fouling is likely in cooling water circuits, the particles being introduced with the raw water or scrubbed from the atmosphere in a cooling tower. In aqueous systems, it is possible to use a flocculating agent and settle out the agglomerates formed. However, in a large ''once through'' system such as that in some power stations, this may not be feasible.

An alternative to coagulation and flocculation is to disperse fine particles by adding a dispersion agent; the particles remain in suspension and pass through the heat exchanger without deposition. The additive imparts similar charges to particles and surfaces so that settlement is prevented or reduced. Agents that are used for the purpose include surfactants and low molecular weight polymers. Dispersants may be used for other fouling problems such as biofouling in cooling water systems where the microorganisms are held away from the heat exchanger surface. Fouling from the corrosion of the heat-transfer surface may be restricted in this way, by preventing the laying down of deposits under which corrosion is promoted. Modification of the system pH may also reduce or eliminate particle deposition, but a change in pH may not be acceptable for the following reasons:

1. The process chemistry may not be compatible with the required pH change.
2. An aggressive environment may be created that replaces particle deposition with corrosion fouling.
3. There may be health and safety problems with handling large volumes of acid or alkali.
4. Additional costs may result if the pH of used cooling water must be neutralized before disposal.

Scale formation

The conventional chemicals used for the prevention or limitation of scale formation (the deposition of inverse solubility salts) include threshold agents, crystal modifiers, dispersants, and surfactants.

Threshold agents, in relatively small amounts, have the ability to inhibit or retard the precipitation of alkaline salts and scaling compounds. Even if the precipitation is not prevented, the action of the chemical additives often produces a ''softer'' sludge-like precipitate, which is more easily removed from the heat-transfer surface. Polyphosphate chain polymers such as hexametaphosphate have been used to control $CaCO_3$

scale formation in cooling water by threshold treatment. Inorganic polyphosphates are generally used for potable water applications, and organophosphates are used with cooling water. Sodium salts of ethylenediaminetetracetic acid (EDTA) and nitrilotriacetic acid (NTA) have been found to be effective for the control of scale formation in modern steam-raising equipment, where the temperatures of the superheater wall tend to be high. The action of the additive is to form chemical complexes with the calcium and magnesium salts. The higher cost of the EDTA compared with the cheaper phosphates usually limits the use of these chelating agents to waters with low hardness.[2]

To prevent or reduce the risk of robust crystalline structures forming on heat-transfer surfaces, it is helpful to employ a crystal modifier. The action of these chemical compounds is to interfere with the crystal structure, so that the scale is more susceptible to the removal forces created by the liquid flow across the deposit. Crystal modifiers including polymaleic acid are widely used in cooling water circuits.

Dispersants have been used successfully to reduce the potential for scaling of heat-transfer surfaces. The action is to impart a charge to crystallites to hold them apart, thereby preventing crystal growth. Organophosphorous compounds and polyelectrolytes are included in this group of chemicals. Surface-active agents may also be used to emulsify potential scale-forming contaminants to keep them in suspension or alternatively to establish a charge on the heat exchanger surface to repel depositing species.

Corrosion control

Control of fouling because of corrosion is possible by the employment of additives. For corrosion to occur, all the elements of the electrochemical circuit must be complete, so that an imposed electrical barrier in the circuit prevents the movement of ions and electrons, which is fundamental to the fouling mechanism. A thin layer of metal oxide can act as such a barrier, provided that the layer is continuous. The protective layer is itself a product of limited corrosion, its presence inhibiting further attack.

As described in the discussion of heat exchanger fouling elsewhere in this encyclopedia, anodic and cathodic reactions occur. Chemicals may be added to prevent these reactions; they are termed anodic and cathodic inhibitors. Cathodic inhibitors form a barrier at the cathode reducing or eliminating H^+ or O_2 transport to the cathode. They include nitrites, silicates, tannins, and orthophosphates. Anodic inhibitors prevent or restrict electron transfer and include polyphosphates, polyphosphonates, and molybdates. Some of these chemicals represent nutrients for aquatic life and may encourage the growth of microorganisms. Blends of anodic and cathodic inhibitors may be employed to maximize the effectiveness of the treatment.

Amines and other long-chain compounds may be added as alternatives to anodic and cathodic protection to form a protective film on metal surfaces. The group is generally known as ''filming amines.'' The long-chain molecules have ends that are hydrophilic and hydrophobic. Through the attachment of the hydrophilic group to the metal surface, the hydrophobic end repels water. The principle is that, through this arrangement, the surface becomes nonwettable and creates a barrier against the attack by water containing CO_2, NH_3, or O_2. The film-forming compounds are generally surfactants, so that when applied they tend to clean the surfaces with which they are in contact. The resulting particles, however, may represent a fouling potential downstream in the process plant. In addition, there is a tendency for the filming amines to react with the hardness salts and iron ions to produce a sludge that by itself represents a potential fouling problem. Before treatment, an estimate of the surface area to be protected is helpful to limit the extent of these unwanted chemical reactions.

Another method of introducing an electrical barrier is what is known as ''controlled scaling.'' It involves allowing limited scale deposits to form on the metal surfaces. A major drawback to the application of this technique is the lack of knowledge of the scaling potential in different parts of the system that allows adequate protection to be given, without excessive deposition. With inadequate information, the result could be scaling in one place and corrosion in another! An alternative is to raise the pH of the water by the use of neutralizing amines. These agents react with the H^+ ions released by the reaction of dissolved CO_2 with the water.

In steam condensers, such as may be found on power stations, the reduced pressure may induce air ingress into the condenser. The oxygen present in the air may represent a corrosion agent. Filming amines may be necessary to control corrosion under these circumstances. Alternatively hydrazine or another volatile oxygen scavenger may be used for oxygen and pH control.

The brief survey of chemical additives for the control of corrosion illustrates the many ways in which control can be achieved. The choice of method will depend on the effectiveness required, coupled with the costs involved.

Chemical reaction fouling

Fouling as a result of chemical reaction is generally associated, although not exclusively, with the processing of crude oil and petroleum fractions. The mechanisms involved include polymerization of components because of the presence of metal ions, and oxygen and coking (break-down reactions) reactions. Chemical

additives have been developed to deal with these problems:

1. Antioxidants scavenge free radicals which may be formed in the process stream and hence reduce the possibility of chain reactions. Suitable additives include phenols and organic compounds containing phosphorus and sulfur.[3]
2. Metal deactivators prevent the initiation of reactions catalyzed by transition metal ions, e.g. Ca^{2+}, Fe^{2+}, Zn^{2+}, and Mn^{2+}. Unless the activity of the metal ions is curtailed, they can give rise to free radicals for oxidation chain reactions to occur.
3. Dispersants and detergents may be added to impart an electric charge to particulate matter that may be the result of chemical reactions, so that the particles are prevented from agglomeration and deposition. The solubilizing property of detergents is the most important factor in reducing deposit formation. A wide range of complex organic chemicals are available that meet this requirement.

In a program to combat fouling from chemical reactions, it is usual to employ a blend of chemicals to offer wide-ranging protection. Many of the additives available are proprietary and it will be necessary to seek the advice of the suppliers before applying to the process stream.

Biofouling

The problem of biofouling, which involves microorganisms often occurs in cooling water systems. An effective way to control the deposition and growth of biofilms on heat exchangers is the use of chemical additives that act either as a biostat that limits the activity of the organisms or as a biocide that kills the organisms. The actions of these two additives are different. A biostat is considered to interfere with the metabolic processes within the organism, whereas a biocide, the use of which is more common, causes some irreversible damage to a vital cell component or function. Chemical dispersants may be employed to reduce the possibility of attachment to the heat exchanger surfaces. The discussion will concentrate on the application of biocides.

The use of chemicals for the control of biological growth has strong environmental connections, as the water is often taken from, and returned to, natural sources such as a river, lake, or sea. Unless proper steps are taken, the discharged water, containing the additive chemicals, could become an environmental hazard and a threat to the local ecology.

The final selection of biocide will require attention to the following points:

1. The water quality, which is likely to change with the season.
2. The concept of the system, whether recirculating or ''once through.''
3. Potential leaks from the process stream on the other side of the exchanger, which may provide nutrients for organisms, act as a biocide or react with the other added chemicals.

The activity of the biocide will very much depend on temperature and pH. Water velocity will affect the mass transfer of biocide to the biofilm and the shear forces acting to remove the biofilm.

Chlorine has been the preferred biocide for many years on account of its availability and relatively low cost. It has the drawback that it is stable and the products of its reaction with organic material are considered to be carcinogenic. It is possible to remove the chlorine from the water before discharge, and in some countries this is a legal obligation. Dechlorination techniques include air stripping, adsorption onto activated carbon, and the use of reducing chemicals. All these techniques impose a cost burden that might prove prohibitive. To meet the concern and accompanying legislation that is being imposed on the use of biocides, so-called ''environmentally biocides'' are being developed. In general, the cost of these biocides is relatively high, reflecting the manufacturing costs involved. There are, however, some biocides that are naturally environmentally friendly, such as hydrogen peroxide, which readily decomposes into water and oxygen. Ozone reverts to oxygen during use.

In view of the costs involved, it is necessary to develop biocide-dosing regimes that maximize biocide effectiveness for minimum cost.[4]

Freezing fouling

The usual method of controlling freezing fouling is by temperature adjustment, so that the heat-transfer surface is above the freezing point of the liquid in contact with it. Crystal modification is an alternative possibility for controlling freezing fouling, e.g., wax crystallization from waxy crude oils, but the high cost and possible contamination effects tend to preclude this technology.

Gas-side fouling

The use of chemicals to control fouling in gas systems, essentially flue gas, is less common and less researched than chemical control of fouling in liquid systems. The major difficulty with this fouling is that all the fouling mechanisms are likely to be present except biological growth, giving rise to a complex situation. Furthermore, the fuel quality is variable, depending on the origin of the fuel, and may even vary between individual

batches from the same source. Because of these difficulties, it is virtually impossible to estimate the required dose or to carry out laboratory tests to provide data. Effective control is achieved through plant trials that may cause severe operational problems in their own right, quite apart from the fouling problem in itself. The additives were originally developed to combat fouling from fuel oils, but the technology has been transferred to other systems such as coal and domestic refuse incineration. In general, the purpose of the additive is to modify the deposit so that it becomes easier to remove, either on-line or more likely on shut-down for cleaning. Relatively cheap additives are necessary, because of the large quantities required. The additive can be a solid as a powder or a liquid as a suspension. Effective mixing with the fuel can present problems, hence the additive is usually applied separately. Additives that are used include magnesium oxide (or hydroxide), limestone, silica, vermiculite, borax, aluminum oxide (or hydroxide), manganese compounds, and rare earth compounds. A large measure of success has been obtained by the use of copper oxychloride. It is usually applied as a suspension in oil, and small quantities can be effective.

The employment of additives for fouling control in combustion systems has a potentially excellent future, but a great deal more research work is required to put the technology on a sound footing.

Physical Methods of Control

Whereas the use of chemicals for fouling control is specific to a particular fouling mechanism, physical methods of control have a more general application. The simplest use of physical control is to increase the fluid velocity across the deposit to increase the shear forces acting on the deposit. The technique is only really feasible where the deposit is loosely bound, such as might be the case with magnetite in boiler water. It also has to be remembered that this technique would increase the pumping energy requirements and hence the operating cost.

A well-established technology that has been particularly successful in controlling fouling in the cooling water side of power station condensers is to circulate sponge rubber balls; the Taprogge system. The condensers contain a large number of tubes with the same internal diameter, which makes them ideal for the circulation of the balls. The balls are slightly larger than the internal diameter of the tubes and are therefore ''squeezed'' as they are forced through the tubes with the water flow. The wiping action of the balls pressed against the internal tube wall removes any deposit as it is formed. Problems that might accompany the application of the technique include potential erosion of the tube surface and the possibility that balls might become stuck at the tube entrance or inside the tube, causing water distribution problems and, if extensive, severe loss of efficiency.

The ''brush and cage'' system is similar in concept to the circulation of sponge rubber balls. It involves the use of a ''brush'' that fits the inside of each tube of a shell and tube heat exchanger. The brush is made of wires or polymer filaments and is made to oscillate between the ends of the tube by flow reversal. A cage is fitted at either end of the tube to ''catch'' the brush as it reaches the end of the tube. The cage and brush would offer a resistance to flow and hence increase the pressure drop through the exchanger. The movement of the brush against the surface of the tube tends to remove any deposit residing on the tube surface since the last sweep. The frequency of flow reversals will depend on the severity of the fouling problem. A major difficulty with this technology is the interruption in flow because of the reversal process, which may affect the stability of the process of which the heat exchanger is a part. The high cost of the brush and cage system may also be a drawback.

An alternative physical method for fouling control uses inserts within the tubes of a shell and tube exchanger. An example is the ''Spirelf'' system that consists of a spiral flexible metal device that is inserted and fixed into the tubes.[5] The action of the device is to vibrate in response to the flow of fluid through the tube in which it is situated, thereby keeping the tube surface clean. The vibration is controlled to avoid erosion of the tube surface. It is claimed[5] that the payback time is low; of the order of months.

An alternative insert for tubes is the spiral insert produced by Cal Gavin. The device consists of a matrix of wire loops on a twisted wire core. The diameter of the coils is such that it is a ''push fit.'' Laboratory studies using these wire matrix inserts with an Arabian crude oil demonstrate that the steady asymptotic value of the fouling resistance is reached in only 10 hr from start up and only of the order of 2–7% of the recommended TEMA value. The presence of the wires, particularly those in contact with the heat-transfer surface, creates turbulence that is largely responsible for the beneficial effects on the fouling.[6] The benefits are not without the potential penalty of increased pressure drop and hence increased pumping costs. By suitable design for the same duty, this penalty can be substantially reduced.

The potential for fouling on the shell side of shell and tube heat exchangers is well known. The problem arises largely because the fluid in the shell does not flow uniformly across the tubes. As a result, it is possible that deposits will form where the velocity is low. By the use of helical baffles, it is possible to redirect the fluid flow so that dead spaces are eliminated.[7]

The use of ultraviolet light

Ultraviolet light kills microorganisms and is, therefore, a candidate for the control of biofouling in cooling water and other water systems that may become contaminated. The technology involves exposing the organisms carried in the water stream to a beam of ultraviolet light, with the aim of killing them before they enter the heat exchanger. It will require a transparent section through which the light passes. The principal problem with the technique, apart from keeping the transparent section clean, is that although the majority of the microorganisms will be killed, some survivors can reach the heat-transfer surfaces and develop a fouling biofilm.

Magnetic devices

A recent publication[8] has several papers that expound the value of magnetic fields on the problem of fouling of heat-transfer surfaces. In general, these discussions cover applications in buildings, for hot water systems and tanks and the like. A successful application was recently reported[9] for control of precipitation of $CaCO_3$ in the hot water system in apartments by formation of soft aragonite, rather than hard scale of calcite. There appears to be little published about applications in the process industries. The considerable skepticism that surrounds the technology has probably directed engineers and technologists away from the use of magnetic fields. If a reliable theory concerning the operation of the technology is developed, it would certainly find a place in the armory of those whose work is to maintain the efficiency of heat exchangers.

The control of corrosion

There are two ways that are employed to control corrosion fouling: the so-called sacrificial anode and impressed current techniques. In the former technology, a sacrificial anode can be any metal that is anodic to the metal to be protected. When an electrical circuit is established, a galvanic cell is set up between the metal to be protected, and the sacrificial anode corrodes, providing a circuit to allow electrons to flow. The electrons could be provided by some direct current supply such as storage batteries, rectifiers, or DC generators, through an auxiliary anode. The electrochemical reactions at the cathode will be complementary to those at the sacrificial anode. The choice of anodic metal and the technology employed depends on a number of factors, but in most applications the major consideration is cost, related to which is the life of the anodic material. In the sacrificial anode technology, the presence of dissolved products from the corrosion of the anode may be unacceptable. Nonconsumable electrodes include such alloys as silicon–iron and platinum–titanium, and they are expensive. It is possible by the use of imposed currents to ''over-protect'' the system, but it may give rise to alternative problems comparable to corrosion.

An alternative is to employ anodic protection that relies on the formation and maintenance of protective films on metal surfaces by means of externally applied anodic currents.

Surface modification

It is possible to reduce corrosion by coating surfaces of heat exchangers. Furthermore, changes in the surface characteristics have the potential to reduce the incidence of fouling in general. In addition, the surfaces are generally smoother and therefore energy losses because of friction are lowered and the associated pumping costs are reduced. In general also coated surfaces facilitate off-line cleaning. Coating materials include vitreous enamels, plastics and polymers, and special paints. Recent technologies of surface modification, which have proved experimentally attractive, involve ion implantation.[10]

The major concern in respect of surface modification is the integrity and resistance to erosion. It could be a serious problem if cracks or local ''peeling'' occur that might facilitate localized corrosion or the buildup of deposits. It has to be remembered also that the presence of a film or coating on the heat exchanger surfaces could be regarded as a fouling layer that by itself imposes a resistance to heat transmission.

Nonmetallic heat exchangers and special designs

It may be feasible to fabricate a heat exchanger from a material that is not a metal or alloy, provided of course that it can withstand the temperatures and pressures of operation, and that it is durable. Flexible plastic tubes in a shell and tube exchanger allow the tubes to oscillate in response to the fluid turbulence in the shell, which helps to eliminate the problem of fouling inside the tubes. A limitation is that, in general, these alternative materials of construction have low thermal conductivities, so that a larger resistance to heat transfer is imposed compared with the traditional metal heat exchanger. Because carbon has inherent corrosion resistance, for potentially severe corrosion conditions it might be worthwhile to consider specially designed heat exchangers fabricated from graphitic carbon. It must be recognized however that these exchangers are less robust than those made conventionally from metal.

A special design of heat exchanger, which was originally introduced to handle viscous liquids, is the scraped surface heat exchanger. In the design, rotating

blades remove the viscous sublayer from the heat-transfer surface, and at the same time keep the surface clean.

Some physical methods of fouling control that are still in the development stage are the use of ultrasound and the circulation of polymer fibers.[11] The latter shows excellent promise for fouling control in cooling water.

The use of physical methods of control in cooling water systems has a major drawback in that it does not control the presence of *Legionella* that is demanded by legislation in many countries. It is necessary when using physical methods of fouling control to treat the water to kill bacteria before the water reaches the cooling tower or spray pond, where droplets of water are released into the atmosphere.

Gas-side fouling in combustion

A number of physical methods are available for handling fouling in combustion systems. So-called ''soot blowers'' that consist of directing a jet of air or steam at the heat-transfer surfaces, where fouling is likely to occur, have been employed for many years. The force created by the jet on the deposit weakens it and ''knocks'' it off the surface. Because of the need for the jets to be close to the surface, and the extent of the heat-transfer surfaces usually involved, it is necessary to use a number of soot blowers in a particular system. To be effective, the device has to be able to rotate or move relative to the heat-transfer surfaces so that they are fully covered; movement may be manual or motorized automatic.

Sonic energy is also used to dislodge deposits from the surfaces in combustion systems. Good results can be obtained in connection with loose friable deposits, but the efficiency is reduced if the deposits are sticky and tenacious.

The manual operation of water jets for on-line cleaning has been used for the removal of fouling deposits for many years, although modern soot blowing equipment is often preferred. Similarly shot cleaning, which involves the release of metal shot so that as it falls under the influence of gravity, it impacts the heat-transfer surface and removes any accumulated deposits. The technique is not widespread on account of the potential damage that could occur.

OFF-LINE CLEANING

Some heat exchangers operate continuously without fouling problems, but the majority, despite efforts to control the incidence of fouling, will require off-line cleaning at some stage. Often this can be accomplished at the annual (or periodic) shut-down, when major maintenance work is put in hand. But more frequent cleaning may be required, depending on the severity of the fouling problem and the cost of mitigation. In some processes such as milk sterilization, off-line cleaning may be required as frequently as twice a day! Under such circumstances, it is desirable to devise optimal cleaning schedules. The technique becomes more complex for the heat exchanger networks that may be found in crude oil preheat trains on refineries. It involves reliable simulation of the network, representative fouling data, and a robust optimization routine.[12] An alternative is to duplicate the heat-transfer system, so that one set is down for cleaning while the other is operating. The capital outlay involved could be considerable, possibly prohibitive for complex heat exchanger systems, but for individual heat exchangers it might be a reasonable solution. Clearly, the choice of heat exchanger type will have a bearing on the management of the cleaning cycle. In some instances, however, milk pasteurization for instance and food processing in general, it has to be recognized that in addition to fouling there is a question of hygiene that has to be met.

It is not possible to discuss the various methods of heat exchanger cleaning in detail, but the following gives a broad overview:

1. Manual cleaning may be carried out in a wet or dry condition. It may involve wiping, brushing, or scraping. It is generally a labor-intensive process and as a consequence it could be costly.
2. Water washing may be acceptable for many forms of deposit, but pressure steam or air lances may be found to be more effective for tenacious deposits. Detergents may be added to facilitate cleaning and it may be necessary to use high velocity blast cleaning with abrasive particles for effective deposit removal in some instances.
3. For the inside of tubes in shell and tube heat exchangers, where accessibility into the tubes is limited, it may be necessary to use drilling or rodding techniques. Lubrication may be afforded by water passing through the special drill.
4. Bullets, scrapers, or scrubbers may be ''fired'' into the tubes of shell and tube heat exchangers, propelled by high-pressure water or air. The projectiles dislodge deposits as they travel at high speed down the length of the tube.
5. Explosives have been used to clean combustion spaces in boiler plant.
6. Steam ''soaking'' involves allowing intimate contact between steam and the surface to be cleaned. The technology depends on the penetration of the deposit by steam and hot condensate, in addition to the effects of the differences in thermal expansion of metal and deposit, which produces ''spalling'' of the deposit.
7. Osmotic shock may be used for the destruction of biofilms, using the properties of the

semi-permeable membrane that microorganisms employ to imbibe nutrients. It is possible to subject the microorganisms to internal pressures that kill the cells by immersing the biofilm in a concentrated salt solution. A long time scale may be required for the full effects to be obtained.

8. Chemical cleaning is a method that employs the ''fill and soak'' principle. The chemical agents include acids, bases, and organic solvents. Circulation of the chemicals through the exchanger may be required to assist the cleaning process by imposing shear forces on the deposit and to carry away the debris.

It has to be emphasized that health and safety regulations will apply to many of these cleaning techniques. Of equal importance is the compatibility of the material of construction of the heat exchanger to be cleaned with the method of cleaning, to avoid physical and corrosive damage.

More details of cleaning techniques are to be found in books by Bott[13] and Muller-Steinhagen.[8]

CONCLUSIONS

The basic design of a heat exchanger, its installation, and commissioning lay the foundation for acceptable operation. The maintenance of efficient and reliable performance is likely to have a profound influence on the overall operating cost and hence on the profitability of the associated process. Recognition of likely operating problems, particularly the possible incidence of deposit formation on the heat-transfer surfaces, is essential. For many heat exchangers, fouling will occur and steps will be required to counter or reduce the adverse effects on heat-transfer efficiency. Most fouling problems can be tackled by on-line methods, but off-line cleaning is likely to be required as part of the overall maintenance program. The frequency of cleaning should ideally be determined by a logical examination of fouling experience and operating conditions, so that costly unscheduled ''shut-downs'' for restoration of efficient operation are avoided.

ARTICLE OF FURTHER INTEREST

Fouling of Heat Exchangers, p. 1043.

REFERENCES

1. Jones, G.M.; Campbell, J.; Bott, T.R. Fouling allowance in the design of refinery heat exchangers. Proceedings of the Second International Conference on Petroleum Phase Behaviour and Fouling, Copenhagen, Denmark, Aug 27—31, 2000; AIChE, Royal Danish Academy of Sciences and Letters, University of Denmark.
2. Kemmer, F.N. Boiler water treatment. In *The Nalco Water Handbook*, 2nd Ed.; McGraw Hill: New York, 1988; 39.1–39.66.
3. Watkinson, A.P. Critical review of organic fluid fouling. In *Final Report AWL/CNSV-TM-208*; Argonne National Laboratory: Illinois, 1988.
4. Grant, D.M.; Bott, T.R. Biocide dosing strategies for biofouling control. Heat Trans. Engineering **2005**, *24* (1), 44–50.
5. Baudelet, C.A. The Spirelf system fouling mitigation in refinery units. Second International Conference on Petroleum Phase Behaviour and Fouling, Copenhagen, Denmark, Aug 27–31, 2000; AIChE, Royal Danish Academy of Sciences and Letters, University of Denmark.
6. Crittenden, B.D.; Kolaczkowski, S.T.; Takemoto, T. Use of in tube inserts to reduce fouling from crude oils. In *Heat Transfer Atlanta 1993*; Volintine, B.G., Ed.; AIChE: New York, 1993.
7. Master, B.I.; Chunangad, K.S.; Pushpanathan, V. Fouling mitigation using Helixchanger heat exchangers. In *Heat Exchanger Fouling and Cleaning Fundamentals and Applications*; Santa Fe, U.S.A., May 18–22, 2003; Engineering Conferences International; New York, 2003.
8. Muller-Steinhagen, H. *Heat Exchanger Fouling Mitigation and Cleaning Technologies*; Publico Publications: Essen, Germany, 2000.
9. Kobe, S.; Drazic, G.; Novak, S.; McGuiness, P.J. The influence of impurity elements and magnetic fields on the crystalline form of calcium carbonate. In *Heat Exchanger Fouling and Cleaning Technologies*; Muller-Steinhagen, H., Ed.; Publico Publications: Essen, Germany, 2000.
10. Muller-Steinhagen, H.; Zhao, Q.; Reiss, M. Ion implantation—A new method of preparing low fouling metal surfaces. In *Understanding Heat Exchanger Fouling and Its Mitigation*; Bott, T.R., Ed.; Begell House: New York, U.S.A., 1999.
11. Bott, T.R. Potential physical methods for the control of biofouling in water systems. Trans. Inst. Chem. Eng. Part A **2001**, *79* (A4), 484–490.
12. Wilson, D.; Smaili, F.; Vassiliadis, V.S. Mitigation of fouling in refinery pre-heat streams by optional management of cleaning and antifouling treatment. Second International Conference on Petroleum Phase Behaviour and Fouling, Copenhagen, Denmark, Aug 27–31, 2000; AIChE, Royal Danish Academy of Sciences and Letters, University of Denmark.
13. Bott, T.R. *Fouling of Heat Exchangers*; Elsevier: Amsterdam, Holland, 1995.

Heat Transfer Fluids

Satish C. Mohapatra
Advanced Fluid Technologies, Inc., Dynalene Heat Transfer Fluids, Whitehall, Pennsylvania, U.S.A.

INTRODUCTION

Heat transfer fluid (HTF) is a very important component of many industrial systems. There are numerous processes in which thermal energy is transferred to or from an object through the physical contact with a heat transfer fluid, which is at a temperature either hotter or colder than the object. Some examples of such applications are in chemical manufacturing, breweries, industrial refrigeration/heating, ventilation, and air-conditioning (HVAC), environmental chambers, plastic manufacturing, supermarket refrigeration, pulp and paper industry, food and pharmaceuticals, oil and gas industry, ice rinks, and engine cooling.

Proper selection of an HTF is crucial to process engineers for a variety of reasons. Different properties of HTFs will impact the design, size, and, ultimately, the cost of the system. Following are some examples of how HTFs can influence a heat transfer process:

1. Lower viscosity and density of the fluid results in smaller pressure drop, therefore smaller pumps.
2. Higher heat transfer coefficient yields smaller heat exchanger, higher bulk temperature of the fluid, and lower flow rate.
3. Reactivity and instability of fluid can disrupt customer production by system fouling and corrosion.
4. Affinity for moisture can freeze system components in low-temperature applications.
5. Low surface tension of a fluid leads to leakage.

Because of the above considerations, selection of a low- or high-temperature HTF for a given application depends on a number of factors.[1] The choice of a particular fluid is invariably a compromise that best suits the specific application and economics. Following is a list of criteria that should be used while selecting a fluid:

1. Freezing point at least 10°C lower than the lowest operating temperature.
2. Low viscosity at low temperatures.
3. Good thermal properties (high specific heat and thermal conductivity).
4. Low flammability (high flash and fire points).
5. High boiling point (low vapor pressure at the highest operating temperature).
6. Noncorrosive toward the materials of construction.
7. Minimal environmental concerns (preferably nontoxic).
8. Good fluid service life (includes excellent thermal and oxidation stability).
9. Economical.

Based on the temperature range they use, the HTFs can be divided into four categories: 1) ultra-low-temperature fluids (lowest operating temperature of −50°C to −120°C or even lower); 2) low-temperature fluids (lowest operating temperature of 0°C to −50°C); 3) high-temperature fluids (highest operating temperature of 200–400°C); and 4) ultra-high-temperature fluids (highest operating temperature of >400°C).

ULTRA-LOW-TEMPERATURE FLUIDS

The current technologies to obtain lower than −50°C include direct use of a refrigerant or a liquefied gas and secondary cooling technology using various low-temperature fluids. In a secondary cooling process (Fig. 1), a primary refrigerant or a liquefied gas (liquid nitrogen or CO_2) is used for the cooling of the heat transfer fluid first, and then the cold heat transfer fluid is circulated through the customer process to enable uniform temperature distribution. This secondary cooling or refrigeration process uses a smaller amount of refrigerant to cool the HTF but may require additional capital expenditure because of the incorporation of another fluid loop. The general advantages associated with secondary cooling are:

1. Fewer refrigerant leaks because there is substantially less refrigerant piping (environmental benefit).
2. Significant reduction in the primary refrigerant charge.
3. Fewer service calls.
4. More stable and uniform process temperature.

Following are some fluid chemistries developed for ultra-low-temperature applications.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007798

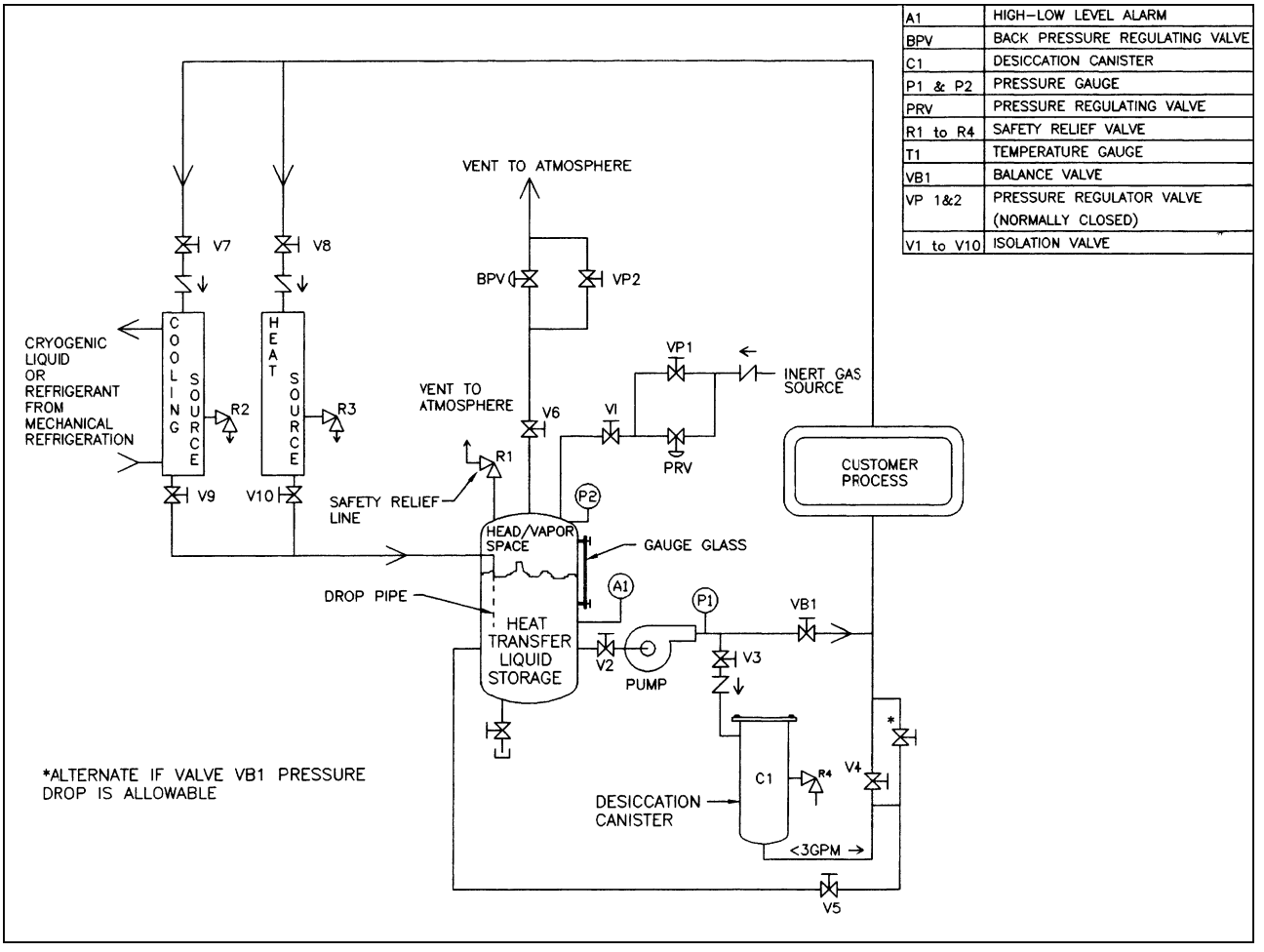


Fig. 1 Flow diagram of a typical secondary cooling/heating process.

Halocarbons

Certain halocarbons such as methylene chloride, trichloroethylene, (TCE) and fluorocarbons can be used as a secondary cooling medium at temperatures lower than −50°C. These fluids are nonflammable and noncorrosive under normal operating conditions. However, chlorinated compounds such as methylene chloride or TCE are very toxic and regulated by the Environmental Protection Agency (EPA).[2] So, these fluids will be removed from the existing systems in the next few years.

Fluorinated compounds such as hydrofluoroethers and perfluorocarbon ethers have certain unique properties that made these fluids compete in the low-temperature HTF market. First, these fluids are nonflammable and nontoxic. Some fluorinated compounds have zero ozone depleting potential and other environmental properties. Second, some of these fluids have a very low freezing point and low viscosity at low temperatures. However, these fluids are much more expensive compared to their counterparts. Also, because of the extremely low surface tension, leaks can develop around fittings causing an increase in the operating cost. Also, fluorocarbons with low freezing point posses a much lower boiling point than other HTFs. Therefore, these compounds are not suitable for applications where both low and high temperatures are desired. Typical applications of fluorocarbon-based fluids are in the pharmaceutical and semiconductor industries within a temperature range of −100°C to 150°C.

Hydrocarbons

Aromatics: Aromatic hydrocarbons such as diethyl benzene are common low-temperature HTFs in the temperature range of −70 to +260°C. The low-temperature heat transfer characteristics as well as the thermal stability of aromatic compounds are excellent. However, these alkylated benzene compounds cannot be classified as nontoxic. Also, these fluids have a strong odor, which can be irritating to the personnel handling it. Also, very few aromatic compounds have

a freezing point lower than −80°C. Hence, aromatic-based low-temperature HTFs are used above −70°C in closed, airtight systems and are typically found in chemical processing and industrial refrigeration.

Aliphatics: Aliphatic hydrocarbons of paraffinic and *iso*-paraffinic type are also used in some systems as low-temperature HTF. Many petroleum-based aliphatic compounds meet the Food and Drug Administration (FDA) and United States Department of Agriculture (USDA) criteria for ''incidental food contact.'' In addition, these petroleum-based fluids do not form hazardous degradation by-products. Most of these fluids have an nondiscernible odor and are nontoxic in case of contact with skin or ingestion. Even with all the above advantages, these fluids are not very common in low-temperature applications, the reason being the high viscosity at low temperatures, and the thermal stability of aliphatic compounds not being as good as that of aromatic compounds. Some of the *iso*-paraffinic based fluids (with 12–14 carbons) can be used from −60°C to 150°C. These fluids are preferred in food and pharmaceutical applications where toxicity is a major issue.

D-Limonene: Another class of low-temperature HTF is based on naturally derived terpenes such as D-limonene. U.S. Patent 3,597,355 describes D-limonene as being particularly preferred among all the monocycloterpenes because of its characteristic properties such as low viscosity at low temperatures.[3] D-Limonene is the major component in the oil of citrus fruit and is present in trace quantities in orange juice. It is recovered in commercial quantities by distilling orange oil obtained from citrus peels. Being derived from the citrus industry, D-limonene is considered a safe and environmentally friendly HTF, and hence it is preferred in many food and pharmaceutical processes. However, the melting point of D-limonene is about −78°C. Below this temperature, it becomes a thick white gel like substance that is impossible to pump. Therefore, the use of D-limonene is limited to only about −60°C. In addition, at temperatures above 50°C, D-limonene oxidizes rapidly in the presence of air. This oxidation triggers the acidification and polymerization of the molecules. Because of this effect, D-limonene is not recommended for use above 50°C.

Terpene Mixtures: Recently, U.S. Patent 5,847,246 has described the use of eutectic mixtures of monocycloterpenes for low-temperature applications.[4] These new environment friendly HTFs have a melting point lower than −110°C and thermophysical properties comparable to many other HTFs. Because the ingredients of these new HTFs are biobased and obtained from renewable sources, these fluids conserve energy for production as compared with the other synthetic HTFs. With the addition of proper antioxidants and using an inert gas such as N_2, these fluids can be used from −100°C to +150°C in a variety of applications.

Silicones

Another class of popular low-temperature HTF is dimethyl polysiloxane, commonly known as silicone oil. Because this is a synthetic polymeric compound, the molecular weight as well as the thermophysical properties can be adjusted by varying the chain length. Silicone fluids can be used at temperatures as low as −100°C and as high as +260°C. These fluids have excellent service life in closed systems in the absence of oxygen. Also, with essentially no odor and very low toxicity, silicone fluids are known to be workplace friendly. However, with low surface tension these fluids have the tendency to leak around pipe fittings, although this low surface tension improves the wetting property. Silicones are usually priced higher than the aromatic- and aliphatic-based HTFs. Because silicone fluids are virtually nontoxic, major applications are seen in the pharmaceutical industry.

More information on ultra-low-temperature HTFs can be obtained from the manufacturers.[5–10] Table 1

Table 1 Comparison of thermophysical properties of various low-temperature HTFs

	TCE	Aromatic	Aliphatic	Silicone	Terpene mixture
Freeze point (°C)	−86	<−84	<−101	<−110	<−110
Flash point (°C)	None	58	59−60	47	55
Viscosity (mPa sec)	1.4	3.9	15.0	6.9	4.3
Density (kg/m^3)	1560	920	815	927	885
Specific heat ($kJ/kg \cdot K$)	0.885	1.636	1.880	1.625	1.620
Thermal conductivity ($W/m \cdot K$)	0.140	0.143	0.115	0.125	0.144
RHTEF	1.0	0.54	0.23	0.37	0.51
Environmental and toxicity concerns	Yes	Yes	No	No	No

Viscosity, density, specific heat, and theermal conductivity of different fluids were obtained at –50°C. The RHTEF values are calculated with respect to TCE.

lists some of the properties and relative heat transfer efficiency factors (RHTEFs) with respect to TCE of commonly used low-temperature HTFs. Relative heat transfer efficiency factors was calculated based on Refs.[1,11], which shows the overall efficiency of an HTF. It is apparent that at −50°C, TCE has the highest RHTEF (equal to 1.0 because it is compared to itself) followed by aromatics and terpenes. However, it has already been discussed before that TCE is not more favorable because of environmental concerns. Therefore, aromatics, terpenes, silicones, and, in some cases, aliphatics and fluorocarbons are preferred as ultra-low-temperature HTFs.

LOW-TEMPERATURE FLUIDS

Low-temperature HTFs are routinely used in various process cooling applications. Examples of these HTFs can be found in the literature.[1,12] Some of these fluids are aqueous (or water based) and some are nonaqueous. In this category, the lowest operating temperature of the fluids is in the range of 0 to −50°C. Different chemistries that are utilized in this temperature range include mainly water-based fluids.

Water is the best HTF in the temperature range of 0–100°C because it has a high thermal conductivity, a high specific heat, and a low viscosity, which are all essential to provide excellent heat transport and transfer. In addition, water is the most inexpensive HTF, is abundant, nonflammable, and nontoxic. The problem arises only when temperatures lower than 0°C or higher than 100°C are encountered. For more than 100°C, the system needs to be pressurized to keep water in liquid form or superheated steam can be utilized. For temperatures lower than 0°C, a freeze point suppressant such as a water-soluble organic compound or a salt is added to water to lower the freezing point to the desired level. Examples of such freeze point suppressants are glycols such as ethylene and propylene glycol; alcohols such as methanol, ethanol, and isopropanol; brines such as chlorides, formates, and acetates of lithium, sodium, potassium, and cesium.

When a freeze point suppressant is added to water, it loses its heat transfer characteristics to some extent, which depends on the concentration of the suppressant. However, water-based fluids still exhibit a very high heat transfer coefficient compared to nonaqueous fluids. Nonaqueous fluids such as aromatics, aliphatics, and silicones are used in such applications when an aqueous fluid exhibits very high viscosity (>50 cP) at the lowest operating temperature or its freezing point is very close to the refrigerant temperature in the evaporator.

Some examples of low-temperature water-based fluids are discussed below.

Aqueous Solution of Organic Compounds

Ethylene Glycol: Commonly used as an antifreeze, ethylene glycol also has found use in refrigeration service. Common applications include process cooling at lower temperatures. Ethylene glycol is colorless and practically odorless and is completely miscible with water. When properly inhibited, it has a relatively low corrosivity, which is a major advantage when compared to salt-based brines. Ethylene glycol solutions can be used as a refrigeration system brine as low as −40°C. However, because of the high viscosity at low temperatures, this is used effectively at −10°C or above. Ethylene glycol is also toxic, hence it is not suitable for open baths or in the food and pharmaceutical industries.

The quality of water used to prepare a glycol solution is very important for the system. Typically, water with low chloride and sulfate ion concentration (<25 ppm) is recommended. Also, a monitoring schedule should be maintained to ensure that inhibitor depletion is avoided and pH of the solution is consistent. Once the inhibitor has been depleted, it is recommended that the old glycol be removed from the system and a new charge be installed.

Propylene Glycol: In its inhibited form, propylene glycol has the same advantages of low corrosivity shown by ethylene glycol. In addition, propylene glycol is considered nontoxic, hence it can be used in food applications. Other than lack of toxicity, it has no advantages over ethylene glycol, being higher in cost and more viscous. Because of the high viscosity at low temperatures (Fig. 2), it is used at −10°C or above.

Methanol/Water: This is a low-cost antifreeze solution, finding use in refrigeration services and ground source heat pumps. Similar to glycols, this can be inhibited to stop corrosion. This fluid can be used down to −40°C owing to its relatively high rate of heat transfer in this temperature range. Its main disadvantages as an HTF are its toxicological considerations. It is considered more harmful than ethylene glycol and consequently has found use only for process applications that are located outdoors. Also, methanol is a flammable liquid and, as such, introduces a potential fire hazard where it is stored, handled, or used.

Ethanol/Water: This is an aqueous solution of denatured grain alcohol. Its main advantage is that it is nontoxic. Therefore, it has found application in breweries, wineries, chemical plants, food freezing plants, and ground source heat pumps. As a flammable liquid, it requires certain precautions for handling and storage.

Salt Brines

Sodium Chloride: Water solutions of this salt are used for refrigeration services and other low-temperature

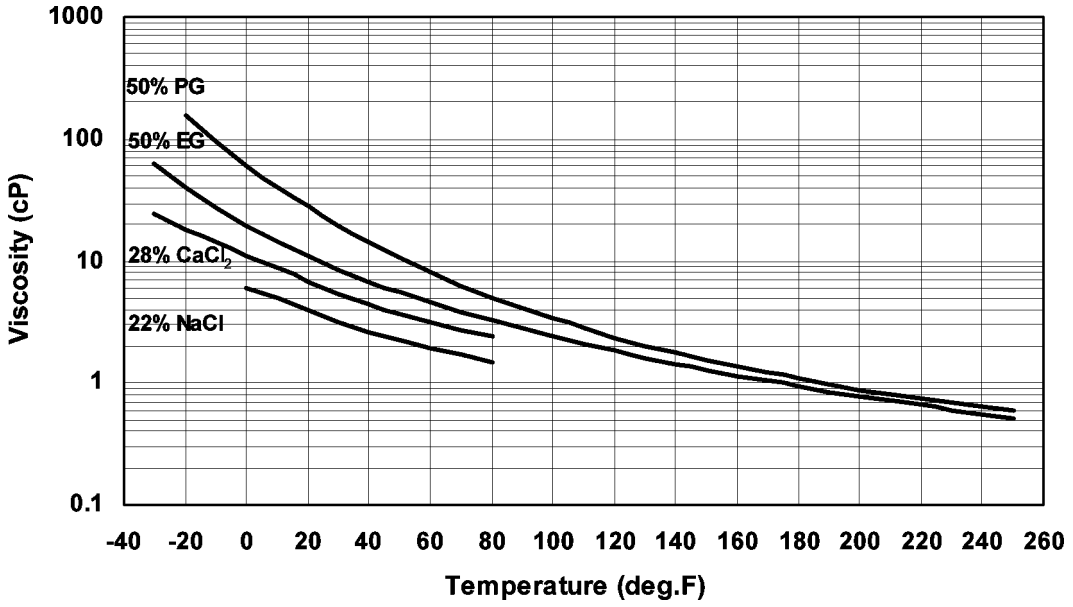


Fig. 2 Viscosity of various water-based low-temperature fluids.

applications. It can be used in applications involving contact with food and in open systems because of its low toxicity and nonflammable characteristics. A high heat transfer coefficient can be obtained with this fluid because of its good thermophysical properties. However, it has two main drawbacks: 1) it has a relatively high freezing point, which limits its application to about $-10^{\circ}C$ and 2) it is highly corrosive, requiring inhibitors that must be checked on a regular basis and replenished to prevent an acid condition from occurring in the system.

Calcium Chloride: Aqueous solutions of calcium chloride find wide use as a circulating brine. Similar to sodium chloride solution, it is nonflammable and nontoxic. It has a lower freezing point than sodium chloride solution and can be used down to $-37^{\circ}C$. The main disadvantages are that 1) it is highly corrosive; 2) it has a reduced heat transfer coefficient below $-20^{\circ}C$; and 3) it cannot be used in direct contact with foods.

Viscosity of calcium chloride and sodium chloride solutions is compared with ethylene and propylene glycol solutions in Fig. 2. It is apparent that there is a major difference in the viscosity of these fluids at low temperatures, which is very important for equipment design.

Dealing with Low-Temperature HTFs

As discussed before, low-temperature HTFs are used in several applications such as reactor cooling, industrial and supermarket refrigeration, lyophilization, and environmental chambers. Both aqueous and nonaqueous HTFs have been developed for these applications. Engineers routinely deal with various issues and challenges while designing, retrofitting, and using these fluids. Such issues are 1) flammability; 2) environmental concerns; 3) corrosion (material compatibility); 4) fluid contamination; 5) bioactivity, etc.

Flammability

A big advantage of water-based fluids over hydrocarbon or silicone fluids is that most of these are either nonflammable or have a relatively high flash point. Salt-based brines are nonflammable. Glycols are nonflammable only when there is at least 20% (v/v) water in the mixture. As discussed before, alcohol-based fluids are normally flammable, irrespective of the water concentration.

Environmental concerns

An important characteristic of a good HTF is that it should be nontoxic and environment friendly. The ideal candidate is the one that can be classified as food grade and also satisfy the USFDA and USDA criteria for ''incidental food contact.'' Additionally, its vapor should neither contribute to the greenhouse effect nor to the depletion of the ozone layer.

Ethylene glycol and methanol are toxic whereas propylene glycol and sodium chloride solutions can be used in direct contact with food. Other brines and alcohol-based fluids are nontoxic and can be used safely in the food and pharmaceutical industries. While many salt-based brines are safe to dispose off into the drains after sufficient dilution, local sewer authorities should be contacted to verify the local laws.

Corrosion (material compatibility)

Aqueous fluids are normally formulated with corrosion inhibitors. Because these fluids are ionic, improper selection of corrosion inhibitors can lead to severe corrosion. Glycols and brines are normally not recommended with galvanized steel and soft solder. Aluminum systems are not recommended above a certain temperature. Sodium and calcium chloride brines are very corrosive toward most of the metals even with the inhibitors.

Most plastics, rubbers, and elastomers are compatible with water-based fluids. However, they should be checked with the fluid manufacturer as well as the plastic/rubber/elastomer manufacturer to ensure compatibility.

Fluid contamination

Water-based fluids can be contaminated in various ways. In new systems, debris from welding, pipe joints, and other residues previously left over from the installers may be present. If not cleaned properly they can contaminate the water-based fluids. Fluxes used for soldering may contain chloride salts that are highly corrosive to metals including stainless steel.

Another way fluid contamination can occur is during fluid retrofit. If the original fluid is not completely removed from the system, it will contaminate the new fluid causing certain problems. In many cases a fluid can be contaminated by mistake. An operator may unintentionally add a wrong fluid into the tank to top off the original fluid. This can be a serious problem if the two fluids are incompatible. How to clean a system is discussed in the following section.

Bioactivity

Microbial activity is a major concern in systems in which water-based fluids are used. Particularly, glycol fluids provide a good source of nutrition to some types of biological species. In salt-based brines, however, microbes do not survive because of high osmotic pressure. When microbes start to grow inside a system, they create a layer known as ''biofilm'' on the walls of the pipes and heat exchangers. This reduces the heat transfer rate. Some microbes are capable of creating acids and hence cause a substantial amount of corrosion in the system.

The problem of bioactivity can be reduced or eliminated by the use of biocides. There are different types of biocides and one must be aware of the toxicity level of these to decide the most suitable biocide for that application. Already affected systems can be treated with biocides but it is very difficult to release the biofilm once it is formed. In that case a biocide treatment followed by a complete cleaning of the system with an acid (hydrochloric acid), alkali (trisodium phosphate), and deionized water is recommended.

HIGH-TEMPERATURE FLUIDS

At temperatures above 100°C, the saturation pressure of water may require the use of expensive equipment to contain the HTF. Glycol/water mixtures oxidize and degrade rapidly above 150°C. Therefore, an HTF other than water and glycol/water is normally utilized at temperatures higher than these. The HTFs used in these applications are normally referred to as high-temperature fluids. Table 2 provides a list of popular HTFs used in a variety of applications. This table also shows the manufacturers of these fluids along with some properties. It is seen that some fluids can be used at high as well as low temperatures because of their wide temperature range.

The most important property for a high-temperature HTF is its thermal stability at exceptionally high temperatures. These HTFs are normally used above their flash and fire points in closed and airtight systems. In many applications, an inert gas such as nitrogen is utilized in the head space of the holding/expansion tank.

There are two mechanisms by which the high-temperature fluids degrade, i.e., 1) thermal degradation and 2) oxidative degradation.

Thermal Degradation

A heating fluid reaches its highest temperature at the surface of the heating equipment such as electric heater, fired furnace, or vaporizer. At this surface, a stagnant film known as boundary layer develops creating an additional resistance for heat transfer. Temperature in this film region can be much higher compared to the bulk fluid temperature. This results in severe degradation of the fluid near the heat source. Fluid can also degrade outside the film depending on the bulk temperature.

There are two types of thermal degradation observed in high-temperature systems. In one type, the fluid molecules break down into low molecular weight products. Compounds with a large number of single bonds (i.e., aliphatics) are more susceptible to this type of degradation compared to compounds with double bonds (aromatics). A compound with both single and double bonds (alkylated aromatics) has better thermal stability than a pure aliphatic compound and worse than a pure aromatic compound.

The other important property of a high-temperature HTF is its boiling point. Because most of the HTFs are

Table 2 A list of widely used HTFs with some properties

Brand name	Manufacturer	Temperature range (°C)	Boiling point (°C)	Flash point (°C)	Autoignition temp (°C)	Freezing point (°C)
CALFLO LT[15]	Petro-Canada	−40 to 260	NA	171	323	
CALFLO AF[15]	Petro-Canada	Up to 316		221	343	
CALFLO HFT[15]	Petro-Canada	326		226	364	
DOWTHERM A[7]	Dow Chemical Co.	15 to 400	257	113	599	12
DOWTHERM G[7]	Dow Chemical Co	29 to 371	288	124	584	4
DOWTHERM J[7]	Dow Chemical Co.	−80 to 315	181	57	420	<−81
DOWTHERM Q[7]	Dow Chemical Co.	−35 to 330	267	120	412	
DOWTHERM MX[7]	Dow Chemical Co.	Maximum 330	328	165	420	
DYNALENE HC-50[5]	Dynalene HTF	−50 to 218	118	None	None	<−55°C
DYNALENE HT[5]	Dynalene HTF	20 to 343	385	200	450	
DYNALENE MV[5]	dynalene HTF	−112 to 163	176 to 179	53	388	<−129
FLUORINET[10]	3 M		30 to 215	NA	NA	−100 to −138
NOVEC[10]	3 M		34 to 130	NA	NA	
MARLOTHERM LH[9]	Sasol	Up to 360	278 to 282	130	450	−30
MARLOTHERM SH[9]	Sasol	60 to 350	385 to 395	200	450	
MULTITHERM 503[13]	Multitherm Co.	−12 to 260	324	154		−65
MULTITHERM IG-4[13]	Multitherm Co.	−6 to 316	411	227		
MULTITHERM PG-1[13]	Multitherm Co.	−29 to 316	349	171		−40
SYLTHERM 800[6]	Dow Chemical Co.	−40 to 400		160	385	−60
SYLTHEM XLT[6]	Dow Chemical Co.	−100 to 260		47	350	−111
THERMINOL 55[8]	Solutia	−25 to 290	351	177	343	
THERMINOL 59[8]	Solutia	−45 to 315	289	146	404	
THERMINOL 66[8]	Solutia	0 to 315	359	184	374	No Data
THERMINOL D-12[8]	Solutia	−85 to 230	192	59	247	No Data
PARATHERM HE[14]	Paratherm Co.	66 to 316	415	227	371	
PARATHERM MR[14]	Paratherm Co.	−1 to 288	301	149	327	
PARATHERM NF[14]	Paratherm Co.	49 to 343	343	174	366	
PARATHERM OR[14]	Paratherm Co.	79 to 288	333	177	332	
XCELTHERM LV1[14]	Radco Industries	7 to 371	258	122	604	7
XCELTHERM 445FP[14]	Radco Industries	229	357	229	379	
XCELTHERM 550[14]	Radco Industries	287	293	177	338	
XCELTHERM HT	Radco Industries	−37 to 348	390	214	450	

blends, the most volatile component boils first followed by the next and so on. So the boiling point of a blend is tricky and in many cases referred to as the initial boiling point, i.e., the temperature at which the first component boils. However, this does not give a true picture of the boiling point of the entire blend. So the HTF manufacturers provide a boiling point range (starting from when 10% of the fluid boiled to end at 90% of the fluid boiled).

High-Temperature Fluids

Normally, organic fluids (hydrocarbons and silicones) are used to attain a temperature between 200°C and 400°C. A great degree of information can be obtained from the fluid manufacturers as well as other published literature. Different chemistries that are utilized for these applications are discussed below.[5–10,13–20]

Petroleum Oils

Also known as mineral oils and hot oils, these petroleum-derived fluids are mostly aliphatic in nature. They are inexpensive, mostly nontoxic, and noncorrosive. Higher molecular weight oils have a high flash point but also have high viscosity. At higher temperatures, their decomposition rate is much more severe than that of the synthetic fluids. Exposure to air at high temperatures results in oxidation of the molecules. Subsequently, the oxidized fluid can react with moisture to form acids that can lead to corrosion of system components. The maximum operating temperature for petroleum oils is about 315°C. These fluids are used in applications where a nontoxic, less odor, economical fluid is desired.

Synthetic Fluids

Synthetic HTFs were developed to provide more thermal stability. These fluids are predominantly aromatic in nature, and can be used at lower and higher temperatures than their petroleum oil counterparts.

One popular synthetic fluid is a eutectic mixture containing 73.5% of diphenyl oxide and 26.5% of biphenyl. This is one of the oldest and most widely used synthetic HTFs. This fluid can be used in both liquid and vapor phase systems. Because of its high freezing point (12°C), it often requires protection against freezing. This fluid has exceptionally high thermal stability. The recommended maximum service temperature is 400°C at a pressure of 16 mPa (150 psia). However, because of the toxicity and odor problems, this fluid has lost popularity over the years.

Other popular synthetic fluids are alkylated aromatics, di- and triaryl compounds, mixture of diphenyl oxide and methylated biphenyl, partially hydrogenated terphenyls, mixture of isomeric dibenzyl toluene, polyalkylene glycol, synthetic paraffins, and silicone oils.

Dealing with High-Temperature HTFs

Similar to the low-temperature HTFs, engineers routinely deal with certain issues and problems with high-temperature fluids. They include 1) fluid expansion; 2) inert gas blanket; 3) draining and flushing; 4) venting; 5) heaters; 6) insulation; 7) filters; and 8) heat tracing.

Fluid expansion

High-temperature fluids usually go through a larger temperature change than the low-temperature fluids. These fluids can expand up to 30% when heated from ambient to operating temperatures. The reservoir or expansion tank must be sized to accommodate the increase in volume. The expansion tank should be at a higher level than and removed from the rest of the system so that the contained fluid exerts a positive pressure on the rest of the system. This arrangement also cools the vapors before venting.

Inert gas blanket

Organic HTFs oxidize when heated in the presence of air. This starts at about 150°C and the rate increases with temperature. Heat transfer fluid manufacturers usually recommend a blanket of an inert gas such as nitrogen or argon on the expansion tank to prevent oxidation. This inert gas can also be utilized to pressurize the system in case the fluid will be utilized above its boiling point or the net positive suction head (NPSH) of the pump needs to be elevated.

Draining and flushing

Draining and system flushing is required when the fluid needs to be changed. It is very difficult to effectively drain the entire system because of the presence of low-lying areas. Draining is enhanced by sloping the pipelines and including drain valves at all the low points during the initial installation. After draining is completed, a good flushing fluid can be used to clean the system of all the sludge deposits. Most HTF manufacturers carry a flushing fluid in their product line.

Venting

Flow of the HTF through the expansion tank during start-up is recommended because it facilitates venting

of vapors that result from residual water and any low boiling components in the fluid. The fluid, however, is seldom circulated through the expansion tank during normal operation. This allows the fluid in the tank to cool, which is necessary if there is no inert gas blanket.

At high temperatures, HTFs break down into low-boiling compounds. Vapor of these compounds needs to be vented from time to time to prevent pump cavitation, reduced fluid flow rate, lower fluid flash point, and reduction in safety because of continuously increasing pressure.

Heaters

As with all heat transfer equipment, there are trade-offs between the capital cost of equipment and the savings in operating cost. A good temperature control system is essential to obtain a long fluid life and limit unscheduled shutdowns. Gas-fired heating units usually have a single flow path heating coil. This eliminates zones with low fluid velocity and pockets of stagnant vapor. The heat flux stays around $20\,W/in.^2$ Hence, there is little danger of overheating the fluid. On the other hand, electric heaters consist of a bundle of elements surrounded by a cylinder filled with the fluid. Even with baffles to control the fluid flow pattern, it is impossible to completely eliminate zones with low fluid flow and collected vapor. The heat flux can reach $45\,W/in.^2$, which may cause degradation of the fluid at the heater surface.

Insulation

Insulation should be covered with an aluminum jacket, and the joints should be caulked. This protects the insulation from fluid leaks and accidental spills. It also prevents accumulation of the hot fluid in the porous insulation, which is very dangerous. Likely sources of leaks such as flanges, pump seals, and valve stems should not be insulated because exposure facilitates the early detection of leaks.

Filters

Most HTF systems do not have filters to remove coke, sludge, and other solids. One reason is that particulate filters are not very effective in removing the very fine solids formed when fluids degrade. Also, operators do not want to change filter cartridges containing hot fluids, especially when the fluids are toxic or have disagreeable odors. Filters containing high-temperature fiberglass and fine mesh stainless steel cartridges are very effective for removing particles. However, the filter should be installed in a sidestream where it does not interfere with the operation.

Heat tracing

Systems filled with biphenyl–diphenyl oxide mixtures or inorganic salt should be heat traced to avoid freezing during cool down to ambient temperature. This is very important because severe pump failure can occur and restarting of the system will be extremely difficult.

ULTRA-HIGH-TEMPERATURE FLUIDS

At temperatures above 400°C, organic compounds degrade so rapidly that their use becomes uneconomical. Under these conditions only inorganic salt mixtures and liquid metals have been employed.

Heat Transfer Salts

The most common salt mixture used is a eutectic blend of 40% sodium nitrite, 7% sodium nitrate, and 53% potassium nitrate.[21] Its temperature range is from 150°C to 540°C. The melting point of a fresh mixture of this salt composition is 142°C. However, sodium nitrite slowly undergoes endothermic breakdown to form sodium nitrate, sodium oxide, and nitrogen. Nitrite can also undergo oxidation to form nitrate. This process results in the increase of its melting point. It can react with carbon dioxide to form carbonates and with water to form hydroxides.

The salt is a strong oxidizing agent, hence should not be brought in contact with flammable or combustible materials. Even use of carbon steel is recommended up to 450°C. Above this temperature, low-alloy or austenitic stainless steel is recommended.

Liquid Metals

Liquid metals are used when temperature requirement is so high that even the nitrate/nitrite salt mixture becomes unsuitable. The most commonly used liquid metal is a eutectic mixture of sodium and potassium (44%). This has a very broad temperature range (40–760°C) and very high thermal conductivity. Lead and lead–bismuth eutectic can be used up to 900°C. There are several disadvantages with the use of liquid metals. Special precautions must be taken while using alkali metals because they react violently with water and burn in air. Mercury, lead, and bismuth-based mixtures are highly toxic, hence their applications are restricted. One common use of liquid metals is in the cooling of nuclear reactors.

CONCLUSIONS

In summary, several factors can affect the decision making process while choosing an HTF for a new system or for retrofitting a fluid in an existing system. Proper selection of the fluid can reduce the overall cost of a process in the long run.

Heat transfer fluids can be divided into various categories based on the lowest or highest temperature of operation. An operator or an engineer has to deal with different issues and problems with both low-temperature and high-temperature fluids.

The new frontiers in HTF research include nanofluids, phase change materials, fluids for specialty applications such as fuel cells, electronic cooling, thermal storage, etc., and environment friendly high-temperature fluids with excellent thermal stability.

REFERENCES

1. Mohapatra, S. Select heat transfer fluids for low-temperature applications. Chem. Eng. Prog. **2001**, Aug, 75–50.
2. EPA Health Effects Notebook for Hazardous Air Pollutants—Draft, Mar 1995; http://www.epa.gov/ttnuatw1/hlthef/tri-ethyl.html (accessed 1999).
3. Hsu, H. Heat Transfer Processes U.S. Patent 3,597,355, 1971.
4. Hsu, J.T.; Loikits, D.J. Fluid Heat Transfer Medium and Heat Transfer Process U.S. Patent 5,847,246, 1998.
5. Whitehall, P.A. Dynalene Heat Transfer Fluids. http://www.dynalene.com.
6. Syltherm Heat Transfer Fluids, Dow Chemical Co., Midland, MI.
7. Dowtherm Heat Transfer Fluids, Dow Chemical Co., Midland, MI.
8. Therminol Heat Transfer Fluids, Solutia Inc., St. Louis, MO.
9. Marlotherm Heat Transfer Fluids, Sasol, Inc., Houston, TX.
10. Hydrofluoroethers, 3M Specialty Chemicals Division. St. Paul, MN.
11. Ballard, D.; Manning, W.P. Boost heat-transfer system performance. Chem. Eng. Prog. **1990**, Nov, 51–59.
12. Mohapatra, S. Selecting a heat transfer fluid. Process Cooling **2000**, May–Jun, 39–42.
13. Multitherm Heat Transfer Fluids, Multitherm Corp., Colwyn, PA.
14. Paratherm Heat Transfer Fluids, Paratherm Corp., Conshohocken, PA.
15. Calflo Heat Transfer Fluids, Petro-Canada Products, Mississauga, Ontario, Canada.
16. Thermalane Heat Transfer Fluids, Coastal Chemicals Co., Houston, TX.
17. Xceltherm Heat Transfer Fluids, Radco Industries. LaFox, IL.
18. Minton, P.E.; Plants, C.A. Heat exchange technology: heat-transfer media other than water. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 3rd Ed.; Wiley: New York, 1980; Vol. 12, 171–191.
19. Minton, P.E. *Encyclopedia of Chemical Processing and Design*; McKetta, J., Ed.; Marcel Dekker, Inc.: New York, 1986; Vol. 25, 190–299.
20. Green, R.L.; Larsen, A.H.; Pauls, A.C. Get fluent about heat transfer fluids. Chem. Eng. **1989**, *96* (2), 216.
21. HITEC Heat Transfer Fluids, Coastal Chemicals Co., Houston, TX.

Heavy Water (Deuterium Oxide)

Sharad M. Dave
Bhabha Atomic Research Center, Mumbai, India

INTRODUCTION

Use of heavy water both as a neutron moderator and as a coolant in fission reactors has resulted in its worldwide demand and many heavy water plants (HWPs) have been constructed. This interest in heavy water, both scientific and technological, has been responsible for the generation of immense information regarding its various aspects. In this article, the nuclear and other properties of heavy water, its production from natural abundance, upgrading and detritiation of downgraded and irradiated water in the reactor are reviewed and an attempt is made to present an overview of the major aspects of heavy water management in nuclear power generation.

ISOTOPE OF HYDROGEN

Unambiguous identification of the isotope of hydrogen, having mass 2, was made by Urey, Brickwedde, and Murphy.[1] Urey received the Nobel prize for the discovery and named it deuterium and symbol D was used to designate it. Deuterium has a peculiar property of efficiently slowing down thermal neutrons, without absorbing them. Because of this very property, deuterium oxide, also known as heavy water, finds profound application in the nuclear industry. Tritium is another isotope of hydrogen having mass 3. It is radioactive, emits low-energy β-particles and has a half-life of 12.26 yr. Isotopes having mass 3 and 4 are unstable and have short half-lives.[2]

PROPERTIES OF HEAVY WATER

Although hydrogen and deuterium have similar physical properties, their nuclear properties are markedly different.

Neutron Moderation

D_2O is the best moderator of thermal neutrons. D_2O absorbs far less neutrons than H_2O and heavy water-moderated reactors can operate with less reactive fissile material. Different moderators are compared in Table 1. A very large moderating ratio for D_2O has led to the construction of reactors that can use a very low concentration of fissile material such as natural uranium (0.7% U^{235}).[2]

This neutron economy leads to low fuelling costs. Also, the production of about 3 g plutonium per kilogram of the fuel at the end of the cycle is an added advantage.[2]

D_2O is also an excellent coolant and as both moderator and coolant, it is not consumed, but does get downgraded and leaks out of the system. Fortunately, these losses are manageable. A few neutrons, captured by D_2O in the reactor, result in the formation of tritium, which exists in the form of DTO and presents hazards in the handling irradiated heavy water.[2]

Physical Properties of Heavy Water

The difference in molecular composition and weight of the isotopic water molecules gives rise to slight differences in their chemical and physical properties. These differences, even though very small, are not insignificant and provide the basis for their separation and analysis. They are also useful for investigation of chemical and biological isotope effects and the structure of molecules. Important properties of H_2O, D_2O, and T_2O are compiled in Table 2.

D_2O has higher density, higher freezing and boiling points, and lower vapor pressure and refractive index as compared to H_2O. The ionization constant for D_2O at 298°K is 1.95×10^{-15}, about one-fifth that for H_2O. Hence, the usual pH meter cannot be directly used as a pD meter. Neutrality is expected at 7.363.[1,2]

D_2O is a poorer solvent than H_2O. The difference in solubility decreases as temperature increases. Substances, in general, are about 16% more soluble in H_2O as compared to D_2O.[1,2]

Thermophysical Properties of Heavy Water

While standardizing the values for the thermophysical properties of water and steam, also known as water substance, the International Association for the Properties of Steam (IAPS) has established formulations for various properties.[3]

Ghosh and Dave have compiled the internationally accepted formulations for various thermophysical

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120023854

Table 1 Comparison of different moderators

Material	Slowing-down power	Absorption cross section, cm	Moderating ratio
Light water	1.28	2.2×10^{-2}	58
Heavy water	0.18	8.5×10^{-6}	21,000
Helium	1.0×10^{-5}	2.2×10^{-7}	45
Beryllium	0.16	1.2×10^{-3}	130
Graphite	6.5×10^{-2}	3.3×10^{-4}	200

(From Ref.[2].)

properties of heavy water substance and have calculated thermodynamic and transport properties of D_2O for the temperature and pressure ranges of 3.8–800°C and 0.01–100 MPa.[4]

NATURAL ABUNDANCE OF DEUTERIUM

The accepted abundances of hydrogen and deuterium are 99.9844% and 0.0156%. Because of the largest relative mass differences, the hydrogen isotopes show the largest variations in the abundance ratio. The major reasons for variations are differences in vapor pressures, equilibrium and kinetic isotope effects, hydration and ultrafiltration.

Deuterium contents of seawaters from major oceans, rivers, lakes, and different water sources in North America and the Indian subcontinent have been surveyed.[2] These studies show that water from the ocean depth has an average D-content of 156 ± 1 ppm, with little variation, whereas the same for surface water is lower. D-content of tropical seawaters is around 160 ± 1 ppm. Freshwater samples show large variations due to isotopic fractionation, which occur when water evaporates from land or sea and is condensed from air. Various factors such as local weather conditions, altitude, latitude, mean air temperature, drainage pattern, distance from the ocean, average precipitation, etc. are responsible for these variations.

Industrial Sources of Deuterium

To qualify for a feed for a HWP, the compound should be available in a very large quantity, with reasonable purity, and at a cheap rate.

A large quantity of feed is required to recover a rather small amount of heavy water. Hence, plants have to handle large flows. Only water, hydrogen, and natural gas fulfill the above conditions. Feed requirement for a 100 Mg D_2O plant are given in Table 3.

Both seawater and freshwater can be a source of deuterium. But freshwater is preferred because corrosion and other problems are less severe with it. Natural gas is available in large quantities, but there is no process available for extracting deuterium from hydrocarbons. Synthesis gas (abbreviated as syngas—a mixture of 75% hydrogen and 25% nitrogen by volume), used for producing ammonia, is a major source of hydrogen. Electrolytic hydrogen is very pure but is excessively expensive.

GENERAL PROCESS CHARACTERISTICS

Problems that are common to all heavy water processes are a very dilute feed, a large overall concentration ratio, and low recovery. Thus, irrespective of process, the ratio of feed to product is 8000 : 1 for 100% and 40,000 : 1 for 20% recovery.[5]

Rae compared different cascades for the H_2O–H_2S process.[6] For a two-stage cascade, the first stage gave a 10-fold enrichment, for a three-stage cascade, the first stage gave fourfold enrichment, whereas the second stage provided sixfold enrichment. Relative tower volumes showed a hefty 30% decrease for the two-stage arrangement and an additional 10% decrease for a three-stage cascade. For a single-stage cascade, the inventory of heavy water approaches the annual production capacity. For a two-stage case, the inventory is reduced by an order of magnitude. For a three-stage cascade, the inventory is further reduced by a factor of 2. Adding more stages increases plant complexity and capital cost. The first stage dominates the plant costs and hence, while comparing different multistage processes, only the first-stage figures are compared.

PARAMETERS FOR PROCESS EVALUATION

The separation factors and energy cost are the most important criteria for the commercial acceptance of any process.[7] The separation factor α denotes the degree of enrichment obtainable in a single equilibrium contact and is defined as

$$\alpha = [D/H + D]_A/[D/H + D]_B \quad (1)$$

where A and B are enriched and depleted streams in equilibrium, respectively. α Ranges from unity (no

Table 2 Physical properties of isotopic waters

Property	Units	$H_2{}^{16}O$	$D_2{}^{16}O$	$T_2{}^{16}O$
Natural abundance	%	99.9864	0.0146	
Moderating ratio		58	21,000	
Molecular weight	C-12 scale	18.015	20.028	22.028
Boiling point (760 torr)	°C	100.00	101.42	101.51
Melting point (MP) (760 torr)	°C	0	3.81	
Triple point	°C	0.01	3.82	4.4
Critical temperature	°C	374.15	370.09	
Critical pressure	bar	221.2	218.8	
Critical volume	cm^3/mol	55.3	55.0	
Density (25°C)	g/cm^3	0.99701	1.1044	1.2138
Maximum density	G/cm^3	0.99997	1.10600	1.21501
Temp. of maximum density	°C	3.984	11.185	13.403
ΔH_{fus} at MP	kcal/mol	1.436	1.515	
ΔH_{vap} at 25°C	kcal/mol	10.52	10.85	10.95
ΔH_f (liq.) at 25°C	kcal/mol	−68.32	−70.41	
ΔH_f (gas) at 25°C	kcal/mol	−57.80	−59.56	
ΔH_f (gas) at MP	kcal/mol	1.436	1.515	
ΔG_f (vap.) at 25°C	kcal/mol	2.06	2.14	
ΔG_f (liq.) at 25°C	kcal/mol	−56.69	−58.20	
ΔS_{fus} at MP	e.u.	5.26	5.42	
ΔS (vap.) at 25°C	e.u.	28.39	29.22	
ΔS of liq.	e.u.	16.75.	18.19	
ΔS of gas	e.u.	45.14	47.41	
$C_p$liq. at 25°C	cal/°/mol	17.99	20.16	
$C_v$liq. at 25°C	cal/°/mol	17.80	20.00	
$Comp_{isotherm}$ 25°C	$atm^{-1} \times 10^{-6}$	46.2	46.5	
$Comp_{adiabatic}$ 25°C	$atm^{-1} \times 10^{-6}$	46.2	46.5	
Vel_{sound} 4950 kHz, 25°C	M/sec	1946	1398	
Dipole moment, 25°C	debye	1.8479	1.8506	
Viscosity at 25°C	CP	0.8903	1.107	
Relative viscosity		1.0	1.25	
Surface tension, 25°C	dyne/cm	72.75	72.12	
Dielectric constant, 25°C		78.45	78.08	
Refractive index, 25°C (5893 Å)		1.33300	1.32830	
Molar refraction, 25°C	min/G/cm	3.712	3.679	
Verdet constant		0.013067	0.012556	
Ionization constant, 25°C		1×10^{-14}	1.95×10^{-15}	6×10^{-16}
Heat of ionization	kcal/mol	13.45	4.42	
Decomposition in molar FP	K/mol/kg	1.853	2.05	
Diamagnetic susceptibility, 20°C		-0.7200×10^{-6}	-0.6466×10^{-6}	
Molar susceptibility, 20°C		-12.97×10^{-6}	-12.76×10^{-6}	
Ionic conductivity, D_3O^+, 25°C	cm/mol	349.8	250.1	
Equivalent conductivity, 25°C	cm/mol	143.3	117.0	

(Continued)

Table 2 Physical properties of isotopic waters *(Continued)*

Property	Units	$H_2{}^{16}O$	$D_2{}^{16}O$	$T_2{}^{16}O$
Second virial coefficients				
at 200°C	cm/mol	−223.0	−226.0	
at 300°C	cm/mol	−117.2	−117.9	
at 450°C	cm/mol	−71.0	−71.1	

(From Ref.[2] and references therein.)

separation) to about 30. Parameters, such as flows, plant volume, and recovery depend directly on α, whereas the dependence of energy consumption on α is indirect. This interdependence of process parameters is schematically illustrated in Fig. 1. Certain parameters such as reaction rate, energy consumption, etc. often outweigh α.[2]

PROCESSES FOR HEAVY WATER ENRICHMENT

Separation processes can be broadly classified according to thermodynamic reversibility:

Irreversible processes: Diffusion, thermal diffusion, gravitational, crystallization, adsorption, biological, laser-based photochemical, electrolysis.

Reversible processes: Distillation, chemical exchange.

Irreversible processes require a continuous supply of work and use energy inefficiently. Among irreversible processes, electrolysis and laser-photochemical methods are important.

Electrolysis

During electrolysis, hydrogen is discharged more rapidly than deuterium, so that the electrolyte becomes enriched in D_2O.

Separation factor, α, is defined as the ratio of the relative abundance of deuterium in water to that in the evolved hydrogen gas.

In alkaline solutions, α ranging from 5.3 to 7.6 was observed for electrodes of iron, nickel, cobalt, silver, and lead, while values ranging from 3.4 to 6.5 were obtained for acid electrolytes. α values as high as 13.9 in alkali solutions and 14.7 in acid solutions were found for anode-polarized platinum cathodes. A similar gold electrode gave α as high as 17.6.[2] α increases with decreasing temperature. Current density is indirectly related to α. At higher current densities, the temperature of the cathode increases, which leads to lowering of α. The condition of the electrode surface and electrolytes has a marked effect on the separation.

The following relationship exists between the initial and final deuterium concentrations and their volumes:

$$[H_0/H][D/D_0]^0 = [V/V_0]^{\alpha-1} \quad (2)$$

Thus, the initial concentration is of vital importance for the economic performance of the electrolytic process.

For the electrolytic separation of heavy water, two different flow sheets are used, viz., electrolytic cascades with and without recycle. The former is used for primary while the latter is used for the final enrichment.[2]

Prior to 1943, all the heavy water was produced by electrolysis. Analysis of a typical electrolytic cascade (without recycle) showed that when the number of stages was 15, the maximum amount of D_2O produced in such a cascade was 0.224 g, with a by-product of about 10,000 g/mol of hydrogen. Consumption of electric power was 4836 kW hr/g D_2O.This enormous consumption of pure electrical energy makes this process unviable.[2,5]

However, for the final enrichment and upgrading of heavy water, electrolysis provides an efficient and economically viable route.

Table 3 Feed requirement of a 100 Mg D_2O/yr plant

Feed material	Deuterium concentration (ppm)	Deuterium recovery (%)	Feed rate (nm^3/day)
Water	145–155	20	8.85×10^3
Syngas	110–130	80	3.95×10^3
Hydrogen	90–145	80	2.95×10^6
Natural gas	110–134	50	2.30×10^6
Calculated values			

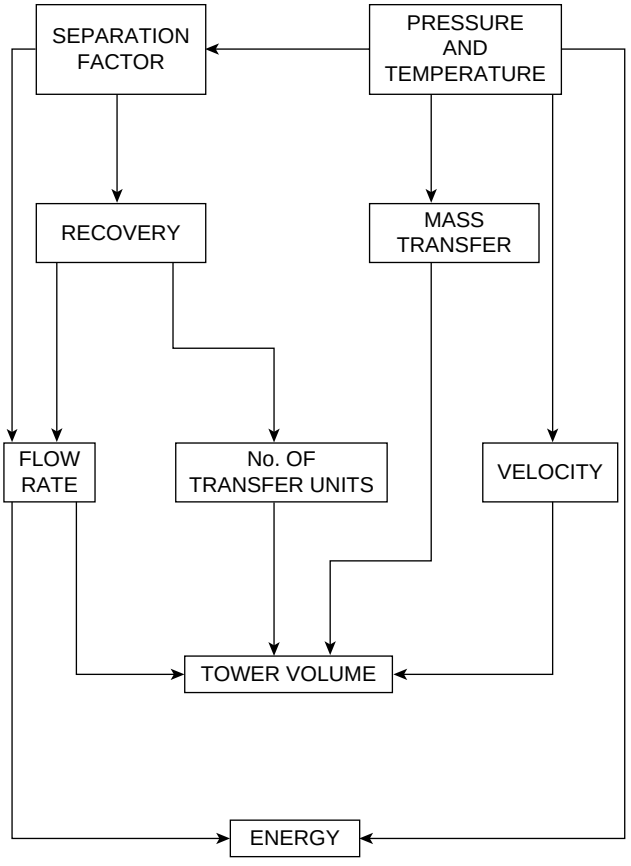


Fig. 1 Interdependence of process parameters.

Distillation

The distillation process utilizes the difference in vapor pressures of isotopic species. Because these differences are small, the process has to be repeated many times in a fractionating column, resulting in a cascade of several stages. Liquid is boiled at the bottom and vapor is condensed at the top, so that a continuous countercurrent contact is established.

Vapor pressures and separation factors

For a mixture of isotopic compounds such as XH_n, XH_{n-1}, $XH_{n-2}D_2$, XD_n, the separation factor, α, can be expressed as

$$\alpha_D = [p_{XH_n}/p_{XD_n}]^{1/n} \quad (3)$$

α is large for compounds having low molecular weight and increases as the temperature is lowered.

Salient features of distillation columns

Distillation columns consist of cylindrical vessels, arranged vertically, within which countercurrent contact between liquid and vapor is established. The less volatile isotope concentrates in liquid phase and is withdrawn from the bottom. A schematic representation of a distillation column is shown in Fig. 2.

Distillation columns can be broadly classified into two types, plate columns and packed columns. In plate columns, there is repeated contact between two phases, as in a bubble-cap column. Plate columns are generally used for primary enrichment. Random packings are used for small-diameter columns, whereas ordered packings are used in large-diameter towers. Packed columns are used for the final enrichment.

A process unit that separates an entering stream into two exit streams of different compositions is known as a stage. For a packed column, the height equivalent to a theoretical plate (HETP) signifies that height of a packed column that achieves the same separation as a theoretical plate. For a plate column, plate efficiency should be as high as possible, whereas in packed columns HETP should be as small as possible.[5] For such an arrangement, the plant volume is proportional to $1/(\alpha - 1)^2$. The effect of this term is overwhelming!

Distillation processes for separation of deuterium

Characteristics of distillation processes based on the above working materials are compiled in Table 4.

Three HWPs with a combined capacity of 1.15 Mg D_2O/mo were built in the United States in 1943 at Morgantown, Childsberg, and Dana but were shut down in the 1950s.[2]

Hydrogen distillation

Because of high separation factors, low-temperature distillation of hydrogen offers a viable route to D_2O production.

This process requires highly pure and a parasitic source of hydrogen. The bottom product gets enriched in deuterium, which is burnt in a stream of pure O_2 to give 99.8% D_2O.

Hydrogen distillation plants were built in France, Germany, Switzerland, Russia, and India. These plants were shut down in the late 1960s, with the exception of an Indian plant at Nangal, which has been operational for more than three decades, producing 14 Mg D_2O/yr.[2]

CHEMICAL EXCHANGE PROCESSES

Chemical exchange processes are characterized by large α, high recovery, and low energy requirement,

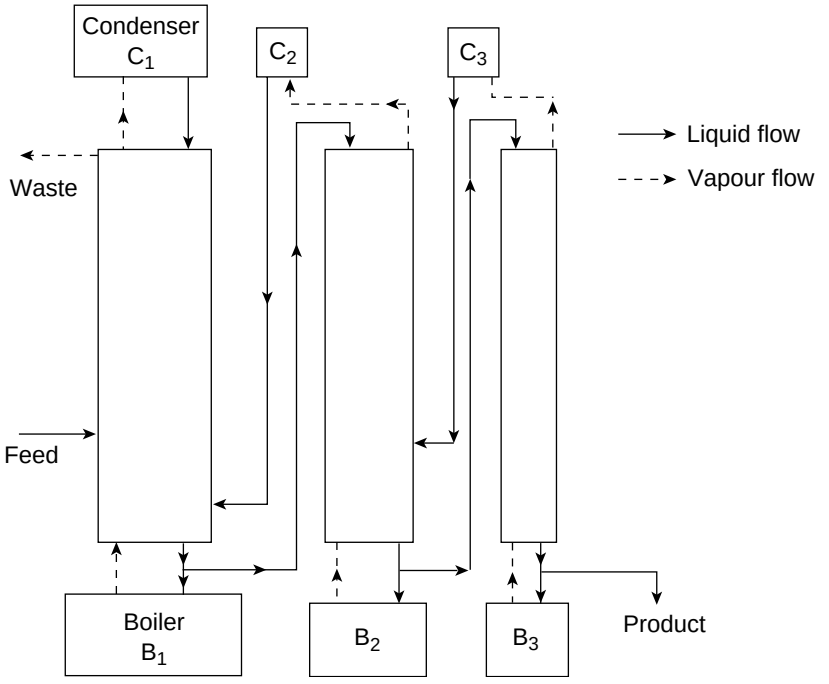


Fig. 2 Schematic diagram of a distillation column.

resulting in lower capital and operational costs. Commercially important exchange processes are:

- Water-fed hydrogen sulfide and hydrogen process.
- Hydrogen-fed ammonia and amine process.

The H_2O–H_2S process, being ionic, requires no catalyst. The remaining three require catalysts.

In large-scale separation, it is economical to combine chemical exchange for primary enrichment (10–25%) with either vacuum distillation or electrolysis for the final enrichment (99.8%).

There are two flow sheets available for achieving countercurrent gas–liquid exchange—monothermal and bithermal (dual temperature), as shown in Figs. 3A–C.

In a monothermal process the gas–liquid contact occurs at a single temperature and the reflux is provided by the conversion of liquid to gas by physical or chemical means in the equipment outside the exchange tower. This process is not suitable for H_2S–H_2O but is successfully used for NH_3–H_2 and H_2O–H_2 exchanges. In the former, ammonia is converted into hydrogen by cracking, whereas in the latter, water is electrolyzed to obtain hydrogen. This phase conversion is always expensive.

To circumvent the high cost of chemical reflux, Geib and Spevack suggested a dual-temperature process, which provides reflux by purely physical means.[5] The cold tower in bithermal operation performs the same function as monothermal tower and enriches deuterium by exchanging deuterium from the ascending gas stream to the descending liquid stream. The hot tower provides the reflux. Both flow sheets are used for producing heavy water.

Water–Hydrogen Sulfide Exchange Process

The first dual-temperature plant for the Manhattan District Heavy Water program was built at Dana and the second at Savannah River. The process was known as GS process (Girdler-Sulfide or Girdler-Spevack). A very simplified flow sheet of this process resembles Fig. 3C.

Table 4 Comparison of different distillation processes

Substance distilled	Temperature (K)	Pressure (kPa)	Separation factors, α	Number of stages, N_{min}	Boil-up rate, $(GlW)_{min}$
Hydrogen	24	250	1.5	37	2.08×10^4
Methane	112	100	0.9965	4373	1.99×10^6
Ammonia	235	100	1.036	431	1.99×10^5
Water	378	120	1.024	643	2.96×10^5

(From Ref.[6].)

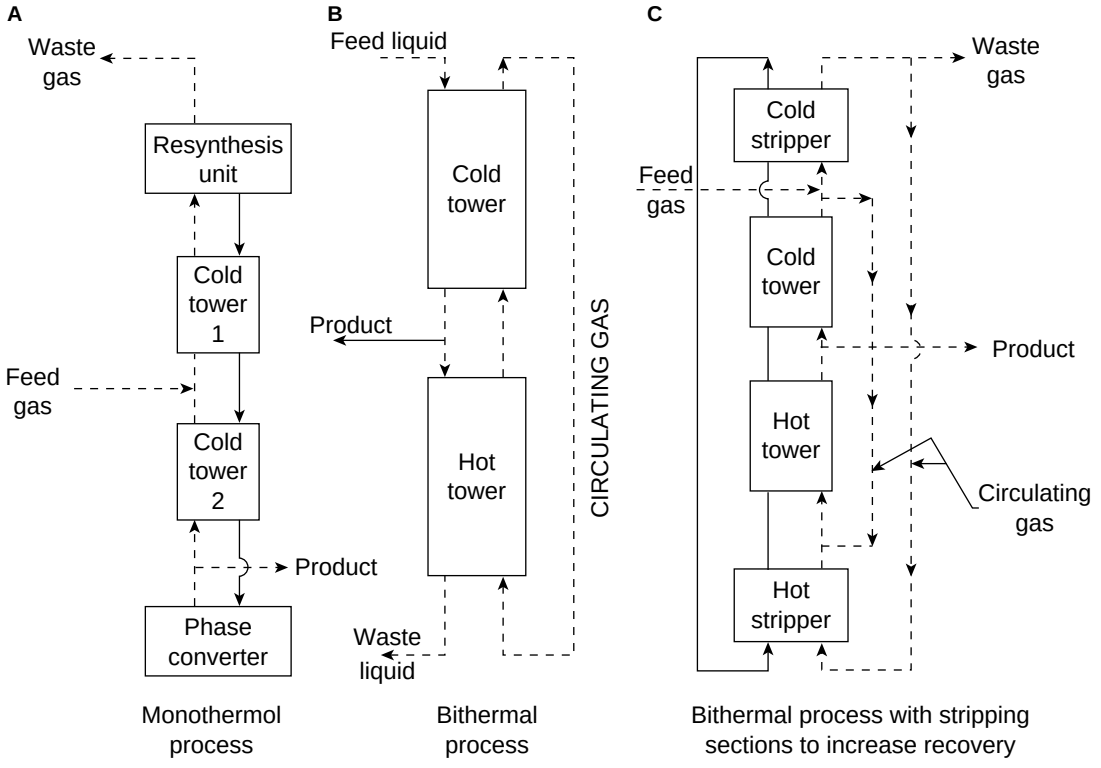


Fig. 3 Chemical exchange processes: (A) monothermal, (B) bithermal, and (C) bithermal process with stripping sections.

General process characteristics

The deuterium exchange reaction between water and hydrogen sulfide

$$H_2O(l) + HDS(g) = HDO(l) + H_2S(g) \quad (4)$$

proceeds rapidly and has an equilibrium constant of 2.32 at 32°C.

Temperature dependence of liquid–gas separation factors, β, is given by

$$\beta = -0.00736T + 4.534 \quad (5)$$

β decreases with increasing temperatures.[8] When a process is carried out at dual temperatures,

$$\beta_{\text{effective}} = \beta_{\text{cold}}/\beta_{\text{hot}} \quad (6)$$

Lower cold tower temperatures would be preferable. Unfortunately, below 28°C, H_2S forms solid hydrate.[5] This limits the cold tower temperature to 30°C. The higher hot tower temperatures are preferred for better recovery. But operating hot tower beyond 130°C is uneconomical, as this leads to increase in hot tower diameter and steam consumption, offsetting the marginal gain in recovery.

At 20 atm (2 MPa), the minimum safe cold tower temperature is around 30°C, below which hydrate formation occurs. The rapid increase in the condensation temperature above 20 atm is another reason for this being the optimum pressure.

Benedict, Pigford, and Levi have carried out mathematical analysis of the GS process.[5] An exhaustive treatment of the process, including calculations for flow rates, dependence of composition on number of stages, effect of solubility and humidity on process analysis, temperature profile in cold towers, simultaneous heat and mass transfer in heat transfer section, concentration reversal in heat transfer section, corrosion, materials of construction, feed purification, and safety, etc. have been reviewed by Dave, Sadhukhan and Novaro.[2]

The main drawback of the GS process is the highly corrosive nature of its aqueous solutions. A 400 Mg/yr GS plant requires an inventory of 800 Mg of H_2S, which is an extremely toxic, flammable, and corrosive gas with a distinct, disagreeable smell even at low concentrations. Hence, adequate measures must be taken for material selection, fabrication, feed purification, feed and waste discharges in water and the atmosphere, safety of staff, the surrounding population, and environment.[2]

Heavy water plants

Plants in the United States. Two plants, one at Dana, IN, and the other at Savannah River, SC, were built in 1952. The first plant at Dana was shut down in 1957

because of lack of demand. The Savannah plant was operating with one-third of its capacity, producing about 150 Mg D_2O/yr, and the plant began to be dismantled in 1994.

Canadian HWPs. The design concepts of Proctor and Thayer were incorporated in the CGE/Lummus process for large plants in Canada, viz., the Port Hawkesbury plant in Nova Scotia and the Bruce near Douglas Point in Ontario.[9]

In 1966, a plant with a capacity of 380 Mg D_2O/yr was commissioned at Port Hawkesbury, Nova Scotia, and production started in 1970. The plant at Glace Bay, Nova Scotia, built in 1963, faced numerous problems caused by seawater feed. The plant was rehabilitated in 1976. The Bruce HWP (BHW-A) at the Bruce site on Lake Huron was built in 1971. Two identical enriching units in parallel were used to double the plant capacity. The final enrichment was achieved by vacuum distillation. Three additional identical plants (BHWP-B, BHWP-C, and BHWP-D) were constructed at Bruce, Ontario, and at Le Prade in Quebec. BHWP-B was commissioned in 1979. Owing to lack of demand, the remaining plants were cancelled.[2,10]

Indian HWPs. Two plants, one at Kota, Rajasthan, and the other at Manuguru, Andhra Pradesh, are in operation. The Kota plant has a capacity of 85 Mg/yr of 99.8% D_2O. It uses GS process for primary enrichment (15%) and vacuum distillation for the final enrichment. The plant was commissioned in 1982.

The second plant, at Manuguru, was commissioned in 1991. The plant, with its own captive power station, has a capacity of 185 Mg/yr of 99.8% D_2O.[2]

The economics of the GS process

The construction and production costs of the Savannah River HWP, have been estimated as[5]

Investment: \$148/kg of D_2O/yr

Operating: \$17.60/kg D_2O/yr

Towers and piping account for almost 50% of equipment costs.

In 1972, it was estimated that a plant with a capacity of 400 Mg of D_2O/yr and a power plant would cost about \$115 million. This corresponds to a specific investment of \$290/kg of D_2O/yr. Thus, the H_2O–H_2S process is both capital and energy intensive.

Hydrogen–Water Exchange

Hydrogen–steam exchange was selected for the Manhattan District project and a 6 Mg/yr plant was built at Trail, Canada.[5]

This process has high separation factors, no corrosion, and large capacity, but all these advantages are nullified by the need for a catalyst. Separation factors have been calculated as function of temperature and deuterium concentrations in water.[11]

Catalysts and their performance

Carbon supported platinum catalysts catalyze this exchange in the gas phase, but they rapidly lose activity on contact with liquid water. For liquid–gas exchange, hydrophobic catalysts were developed by depositing platinum on porous polytetrafluoroethylene (PTFE, Teflon). Teflon provides a water-repellent environment to the platinum crystallite.[12] In a more successful approach, platinized carbon was prepared and bonded to a variety of column packings, using Teflon both as a bonding and as a waterproofing material. Butler has evaluated the performance of different types of Teflon-coated catalysts for water–hydrogen exchange.[12] He found that the reaction is fast and the catalyst stable. Impurities in feed, however, poisoned the catalysts. Many homogenous catalysts have been studied, but so far no suitable catalyst has been found for H_2O–H_2 exchange.

Monothermal hydrogen–water exchange

When hydrophobic catalysts were used for monothermal operation and reflux was provided by electrolysis, the process was termed combined electrolysis and catalytic exchange (CECE).

Because water can be used directly in the liquid form, there is a considerable saving in capital and operating costs. Because of combination with electrolysis, α is large, which bestows many advantages.[13]

But CECE involves production of a huge amount of electrolytic hydrogen and unless there is a direct market for this hydrogen, the process remains uneconomical for primary enrichment.

Dual-temperature H_2O–H_2 process

This process has been explored by many workers.[2] In most cases the cold and hot tower temperatures were at 100°C and 500–600°C. At high temperatures and high pressures, catalysts lose activity, making the process economically unviable.

Ammonia–Hydrogen Exchange Process

The ammonia–hydrogen process is represented by:

$$NH_3(l) + HD(g) = NH_2D(1) + H_2(g) \qquad (7)$$

Both monothermal and bithermal versions are feasible. For cold and hot tower temperatures of −40°C and +70°C, effective α is 2.03. For this exchange, α and its temperature coefficient are large, resulting in large recovery and small plant volume and low energy requirement. Corrosion and operational problems are manageable.

The ammonia–hydrogen plant requires a large source of ammonia or hydrogen. Synthesis gas, used for producing ammonia on an industrial scale, is an ideal source. But this limits the production capacity of HWP to about 100 Mg D_2O/yr corresponding to 1500 Mg NH_3/day in the ammonia plant.

Separation factors

The separation factors increase with decreasing temperature and follow the relationship

$$\ln \alpha = A/T + B \quad (8)$$

where A and B are constants. A corresponds to the heat of exchange reaction and was found to be 4.38 kJ/mol.[2] α increases with decreasing deuterium concentration of the system.[2]

Kinetics and the role of catalyst

Deuterium exchange between ammonia and hydrogen does not proceed spontaneously. Wilmarth, Dayton, and Flournoy found that potassium amide catalyzed this exchange.[14] Further studies showed that the rate of exchange is proportional to solubility of both hydrogen and catalyst in ammonia, which can be increased by intimate mixing of the liquid and gas phases.[2,14]

While high temperatures favor rapid mass transfer, low temperatures are desirable for higher α. NH_3 freezes at −78°C and slow kinetics at lower temperatures suggests a lower limit of temperature of −40°C for the cold tower.[2]

Properties of amide–ammonia solutions

The solubility of amides in liquid ammonia differs appreciably and increases from lithium to rubidium. Because of the high reactivity of amides and the excellent solvent properties of liquid ammonia, amide–ammonia solutions are highly reactive and react with most substances.[15]

The alkali amides react spontaneously with water and oxygen to form hydroxides; with CO they form formamide (HCONHK) and formate (HCOOK) and with CO_2 they form ammonium carbamate (NH_4CONH_2), which reacts with KNH_2 to yield stable potassium carbamate ($KCONH_2$). These reaction products are highly insoluble in liq. NH_3 and cause serious operational problems. Hence, it is imperative to have feed gas absolutely free from oxygenous impurities.[2]

Monothermal ammonia–hydrogen exchange process

The monothermal NH_3–H_2 process was developed in France. Deuterium exchange takes place between liquid NH_3 and gaseous H_2. KNH_2 is the catalyst, obtained by dissolving potassium metal in liquid ammonia. A schematic diagram of the process is given in Fig. 4.

The important steps in the process are:

Purification of feed gas: Stripping of syngas by molecular sieves and liq. NH_3 + KNH_2 wash to remove impurities.

The exchange process: To achieve reasonable mass transfer rates, an ejector contactor was developed by M/s Sulzer Brothers.[5] The plate consists of a ring of nozzles in which syngas breaks up and scatters the liquid in tiny droplets. At each stage, gas is expanded through nozzles and liquid is sucked into the high-velocity gas stream, giving a very intimate gas–liquid mixture.[2,5]

Catalyst recovery and recycle: Amide also gets enriched in its deuterium content. This deuterium and amide must be recovered by evaporating ammonia and stripping off deuterium by scrubbing.

Cracking of ammonia (ND_3): Complete dissociation of ammonia is achieved by heating it with catalyst at 550–580°C at 55–60 atm pressure. The product syngas can be burnt in air or oxygen to produce D_2O.

Integration of HWP with the ammonia synthesis plant: Purified synthesis gas after the pressure adjustment is sent to HWP, which acts as a by-pass to the ammonia synthesis unit.[2]

Plants based on the monothermal process

The monothermal ammonia–hydrogen exchange process was developed in France. The first plant was built at Mazingarbe and five plants have been built in India.

Mazingarbe HWP. The Mazingarbe plant was started in 1968. The basic design described above was followed. The final concentration of 99.8% was achieved by ammonia distillation. The Mazingarbe plant was shut down in 1972 because of an explosion in the ammonia synthesis plant.[2,5]

Indian HWPs. Four plants, based on the monothermal process, have been constructed at Baroda and Hazira in Gujarat, Tuticorin in Tamilnadu, and Thal in Maharashtra in India.[2]

The Baroda plant utilizes syngas from the Gujarat State Fertilizer Corporation plant at a pressure of

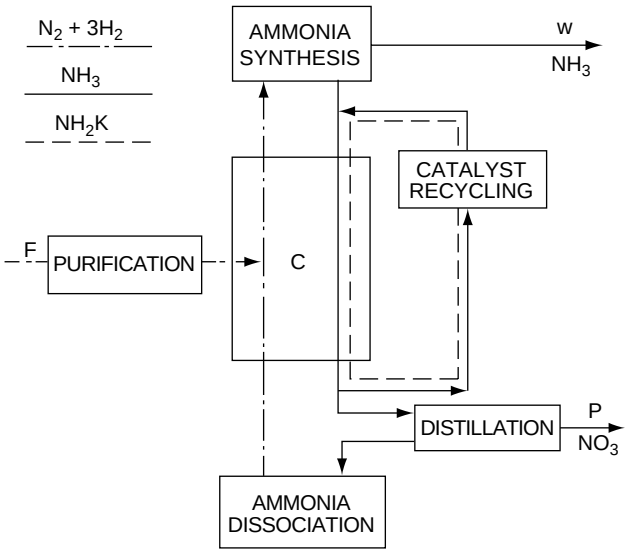


Fig. 4 Schematic monothermal ammonia–hydrogen process.

about 650 atm. It was started in 1983. The plants at Baroda and Tuticorin were built by Gelpara consortium for the Department of Atomic Energy, India. The final enrichment was carried out by distillation. Two more Indian HWPs, one at Thal and the other at Hazira, were designed and fabricated by the Department of Atomic Energy and were commissioned in 1986 and 1990, respectively. The Thal plant draws its requirement of syngas from two fertilizer plants, having a combined capacity of 1050 Mg of NH_3/day. Because of the larger capacity of the ammonia plant, two prototype HWPs, each having a capacity of 65 Mg D_2O/yr, have been set up. These plants are in operation.[2]

Despite the relatively low heat of decomposition of ammonia, 3×10^{-6} J/kg, cracking of ammonia represents the costliest operation.[2] Hence, bithermal process for the ammonia–hydrogen exchange was proposed.

Dual-temperature NH_3–H_2 process

A bithermal ammonia–hydrogen HWP has been constructed for the Government of India at Talcher by Uhde, GmBH of Germany.

The Talcher HWP is coupled with a 900 Mg/day ammonia synthesis plant of the Fertilizer Corporation of India to produce 62.5 Mg D_2O/yr. After purification, syngas enters the enrichment section, which consists of a three-stage cascade. The first stage is superimposed by two cold- and hot-stripping columns. Each stage consists of one pair of cold and hot columns. Deuterium concentration in ammonia is high at the base of the cold tower and in synthesis gas at the top of the hot column. Deuterium is transferred to water by ammonia–water exchange. The final enrichment is achieved by distillation.

A detailed analysis of this flow sheet is given by Benedict, Pigford, and Levi.[5] The Talcher plant, which was to be commissioned by the end of 1975, had a series of problems. The plant was commissioned in 1979 but it could never achieve sustained production and has become uneconomical to operate.

Large hot and cold tower volumes due to large flows and complex heat exchanger systems are the major drawbacks of the bithermal process.

Problems in the ammonia–hydrogen exchange process

1. The capacity of HWP is limited by the ammonia synthesis plant.
2. The deuterium content of syngas is low (110–120 ppm).
3. Interruptions in the ammonia plant lead to serious disruptions in the working of HWP.

These limitations can be removed by linking syngas feed of HWP with water. This can be achieved by:

1. Equilibration of depleted hydrogen with water.
2. Isotopic exchange of ammonia with water.

Amine–Hydrogen Exchange Process

Replacing ammonia by methylamine in the exchange

$$CH_3NH_2(l) + HD(g) = CH_3NHD(l) + H_2(g) \quad (9)$$

was first suggested by Bar-Eli and Klein.[16] Further studies have shown that monomethylamine (MMA) is a better substitute for ammonia and offers the following advantages:[17]

1. Larger separation factors, better kinetics, lower cold tower temperature.
2. Lower vapor pressure of MMA, low capital cost, low energy requirement.

Process characteristics

Separation factors for this process have been measured by Dave, Ghosh, and Sadhukhan and the temperature

dependence of the equilibrium constant can be expressed as[17]

$$\ln K = A/T + B \quad (10)$$

where A and B are constants. The gradient A corresponds to the enthalpy of reaction. ΔH was found to be 4.61 kJ/mol.[17]

Amide–amine solutions react vigorously with water and oxygenous impurities in syngas and form insoluble products. Hence, scrupulous purification of feed gas becomes inevitable.[2] At high pressures, hydrogen reacts with methylamide catalyst [potassium methylamide (PMA)] to form potassium hydride (KH), which is insoluble in amine. To circumvent this hydride precipitation and obtain a reasonable rate, an equimolar mixture of lithium methylamide (LMA) and PMA, called PLMA, was developed by Canadian workers.[6,18]

As conversion of methylamine to hydrogen is difficult, only bithermal operation is possible. The cold tower temperature is fixed at 223°K where the rate is fast and α is 7.9, the highest known for any practical exchange reaction. The optimum hot tower temperature is limited to 313 K, a limit set by the decomposition of the catalysis. Sulzer, GmBH and Atomic Energy Canada Limited have given a flow sheet for 65 tons/yr HWP.[6,18]

LASER SEPARATION OF ISOTOPES OF HYDROGEN

Laser isotope separation (LIS) utilizes small differences in the spectroscopic properties of isotopic substances. Each isotope-bearing substance absorbs a radiation of a particular wavelength. Separation of the excited species can be achieved by multiple-photon absorption or photopredissociation of molecules or chemical scavenging.

For laser separation of heavy water, the working material must be cheap, nontoxic, volatile, and easily redeuterable. It must have acceptable photon utilization and high optical selectivity for multiphoton absorption. Photoproducts must be stable and separable.

Concepts in Large-Scale Separation of Deuterium

A basic flow sheet containing common features of an LIS plant is schematically shown in Fig. 5. The two major components in a deuterium LIS plant are the laser isotopic separator and chemical exchange reactor.

The feed gas, entering the separator, is exposed to laser and is separated into enriched product stream and depleted stream.

The chemical exchange reactor replenishes the depleted stream by chemical exchange reaction with an abundant source of deuterium.[19]

Promising Deuterium Laser Separation Processes

Photopredissociation of formaldehyde and the consequent enrichment of deuterium were reported by Yeung and Moore.[20] When an equimolar mixture of H_2CO and D_2CO was irradiated with a ruby laser, the D_2CO molecules were excited and dissociated into $D_2 + CO$ and sixfold enrichment was obtained. Better enrichments were obtained by subsequent workers by using narrower wavelengths. Letokhov and Vanderleeden have given the mathematical analysis of the process.[21–26] Redeuteration of formaldehyde was studied by Vanderleeden and Dave and Ghosh.[2,25,26]

Multiphoton Dissociation

Marling and Herman investigated CF_3CDCl_2, known as Freon-123 (2,2-dichloro-1-1-1-trifluoroethane).[27] This compound absorbs selectively near 10.2 and 10.6 μm, accessible with CO_2 TEA laser, and undergoes dissociation to give stable CF=CFD as the main product. The selectivity for this process was as high as 1400. The depleted working material can be redeuterated by chemical exchange with water in the presence of NaOH catalyst.[27]

Another starting material, fluoroform-d, was reported by Tuccio and Harford.[28] This molecule undergoes multiphoton dissociation when irradiated with a pulsed CO_2 laser and dissociates to yield DF. An enrichment factor of greater than 5000 was obtained. Based on these studies, Marling, Herman, and Thomas have proposed a 100 Mg/yr deuterium separation plant using trifluoromethane as the working material.[29]

Formidable engineering and technological problems have to be solved before deuterium LIS can compete with other established processes.

FINAL ENRICHMENT AND UPGRADING OF HEAVY WATER

The primary enrichment processes discussed above give enrichment of 10–25% depending on the design. Further concentration of up to 99.8% is achieved by the final enrichment step. The final enrichment involves only a fraction of the total outlay and is a part of the heavy water program. Feed has more or less constant concentration.

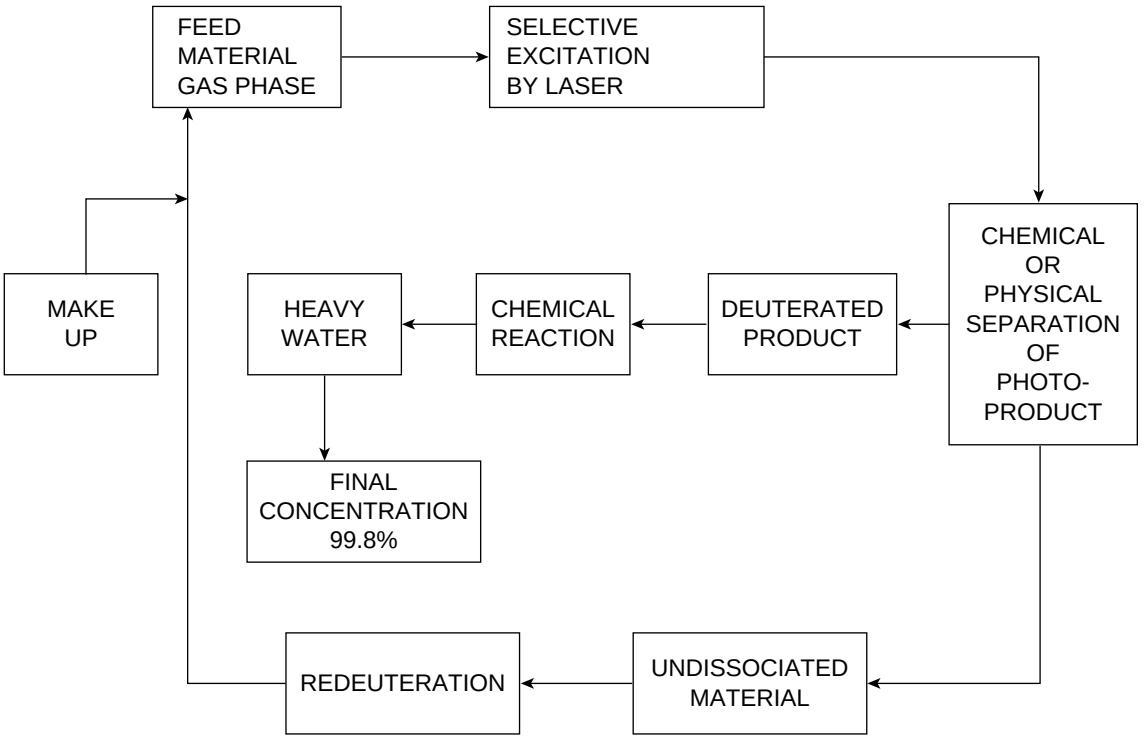


Fig. 5 Simplified flow sheet of an LIS plant for production of heavy water.

Heavy water reactors have a huge inventory of heavy water (0.9 tons/MWe), which leaks out. Yearly leakages amount to about 50%. Purity of recovered heavy water, which varies from 5% to 98% is radioactive. Hence, the upgrading facility should be near the reactor site.

Any process that is suitable for the final enrichment can be employed for upgrading also. The criteria for selecting a process are reliability, simplicity, and conservation of costly feed. Distillation and electrolysis fulfill these requirements.[2]

Distillation can be used for the final enrichment, or upgrading, or extracting of light water from the moderator heavy water.

As the capacity of the primary plant is large, the distillation columns for the final enrichment are large. For smaller-diameter columns, dixon rings are used. For large-diameter columns, ordered packings, developed by Sulzer Brothers, having a high surface area (therefore, low HETP), less pressure drop, and high wettability are used. The column and associated equipment should be leak-tight to prevent loss of costly D_2O.

For upgrading plants, the equipment must be leakproof and under vacuum to prevent tritium from escaping. There is a cleanup unit to avoid contamination. The electrolysis employed for the final enrichment or upgrading is a batch process and is carried out in a tank-type cell. The anode is usually made of nickel, while low-carbon steel is used as a cathode. The process temperature should be maintained as low as possible, for high separation factors. Basically, an electrolytic plant consists of an electrolytic section, a purification section, and a storage section. The separation factors for HTO/H_2O being much higher for electrolysis than for distillation, the tritium content in the reject in an electrolytic plant can be controlled effectively.

DETRITIATION OF HEAVY WATER

Tritium is produced by ternary fission and resides in the moderator. It is a soft β-emitter and is responsible for about 40% of man-rems exposure in the reactor.[2]

Processes for Detritiation

Distillation, electrolysis, chemical exchange, and laser-based photochemical methods are employed for separating tritium. In the detritiation process, tritium should concentrate in the gas phase as DT or T_2 and not in the aqueous phase. The permissible limit for tritium in air is 25,000 times larger than that for aqueous tritium.

Water distillation and electrolysis are not suitable because of high energy costs and tritium gets enriched

in the aqueous phase. Cryogenic distillation in combination with chemical exchange is used for detritiation as well as in tritium production. Tritium concentrates in the gas phase.

D_2O–DT exchange can be used for transferring tritium from heavy water to deuterium. Further enrichment is achieved by cryogenic distillation. Because of the similarity between deuterium and tritium, platinum on charcoal is the catalyst for vapor phase exchange, whereas hydrophobic catalyst is used for liquid–gas exchange.

Two basic schemes are available:

Vapor phase catalytic exchange (*VPCE*), *followed by cryogenic distillation*: Platinum on charcoal catalyst is used for the exchange. The heavy water has to be converted into steam. The exchange is, therefore, carried out at 200°C.

Liquid phase catalytic exchange (*LPCE*), *followed by cryogenic distillation*: Hydrophobic catalyst is used. Exchange is carried out between 50°C and 70°C. Commercial detritiation plants are based on VPCE.[2]

CONCLUSIONS

Heavy water is not only the most important component, but also the major cost contributor to the nuclear power program, which utilizes pressurized heavy water reactors.

The GS process has the advantage of unlimited feed but has serious limitations also. The prospects of sizeable cost improvement look bleak. Nevertheless, this proven process reigns supreme and is a major producer of heavy water. The ammonia–hydrogen process has emerged as a serious alternative. There is scope for improvement and its reliability is improving. The hydrogen–water process has the potential to become a major process, provided existing constraints are removed. In spite of many advantages, commercial exploitation of the amine–hydrogen process has to await solution of critical process chemistry problems. Laser separation of heavy water exhibits some potential, but it will take great effort and investment to remove existing limitations. Because of the dwindling demand for heavy water at present, serious R&D efforts to develop more efficient and cheaper processes are lacking and we will have to rely on the proven processes.

REFERENCES

1. Kirshenbaum, I.S. *Physical Properties and Analysis of Heavy Water*; Urey, H.C., Murphy, C.M., Eds.; McGraw-Hill Book Company, Inc.: New York, 1951; 1–79.
2. Dave, S.M.; Sadhukhan H.K. Novaro, O.A. *Heavy Water—Properties, Production and Analysis*; Quest Publications: Mumbai, 1997.
3. Hill, P.G.; MacMillan, R.D.C. Water and steam—their properties and current industrial applications. Proceedings of the 9th International Conference on the Properties of Steam. Straub, J., Scheffler, K., Eds.; Pergamon Press: New York, 1980; 105, 344.
4. Ghosh, S.K.; Dave, S.M.; Sadhukhan, H.K. *Thermodynamic and Transport Properties of Heavy Water Formulation and Results*; Report No. BARC-1354; Bhabha Atomic Research Centre: Trombay, 1987.
5. Benedict, M.; Pigford, T.H; Levi, W.F. *Nuclear Chemical Engineering*; 2nd Ed.; McGraw-Hill: New York, 1981.
6. Rae, H.K. *Separation of Hydrogen Isotopes*; ACS Symposium Series, 68; American Chemical Society: Washington, DC, 1978; 1–40.
7. Dave, S.M. Report No. BARC-1055; Bhabha Atomic Research Center: Trombay, 1980.
8. Dave, S.M.; Ghosh, S.K.; Sadhukhan, H.K. Indian J. Chem. **1981**, *20A*, 329–333.
9. Proctor, J.F.; Thayer, V.H. Chem. Eng. Prog. **1962**, *58*, 53–57.
10. Villani, S. Isotope separation. American Nuclear Society **1976**, 335–372.
11. Dave, S.M.; Ghosh, S.K.; Sadhukhan, H.K. *Deuterium Exchange Between, Liquid Water and Gaseous Hydrogen*; Report No. BARC-1164; Bhabha Atomic Research Centre: Trombay, 1982.
12. Butler, J.P. Separation. Sci. Technol. **1980**, *15*, 371.
13. Hammerli, M.; Butler, J.P.; Stevens, W.H. Peak power and heavy water production from nuclear electrolytic H_2 and O_2 in Canada. Int. J. Hydrogen Energy **1979**, *4*, 85 (Atomic Energy Canada Ltd., Report No. AECL-5512, 1976).
14. Wilmarth, W.K.; Dayton, J.C.; Flournoy, J.M. The mechanism of exchange of hydrogen gas and aqueous alkali. J. Am. Chem. Soc. **1953**, *75*, 4549.
15. Delmas, R. Report No. CEA-R-3377, 1966.
16. Bar-Eli, K.; Klein, F.S. J. Chem. Soc. **1962**, 3038.
17. Dave, S.M.; Ghosh, S.K.; Sadhukhan, H.K. *Methylamine Hydrogen Exchange*: *Part I and II*; Report No. 1116 and 1117; Bhabha Atomic Research Center: Trombay, 1981.
18. Bancroft, A.R.; Rae, H.K. *Heavy Water Production by Amine Hydrogen Exchange*, Report No. AECL-3684; Atomic Energy Canada Ltd., 1970.

19. Dave, S.M.; Ghosh, S.K.; Sadhukhan, H.K. Report No. BARC-1056; Bhabha Atomic Research Center: Trombay, 1980.
20. Yeung, E.S.; Moore, C.B. Isotopic separation by photopredissociation. Appl. Phys. Lett. **1972**, *21*, 109.
21. Letokhov, V.S. On selective laser photochemical reactions by means of photopredissociation of molecules. Chem. Phys. Lett. **1972**, *15*, 221.
22. Letokhov, V.S. Use of lasers to control selective chemical reactions. Science **1973**, 180.
23. Letokhov, V.S. Laser separation of isotopes. Annu. Rev. Phys. Chem. **1977**, *28*, 133.
24. Letokhov, V.S. Laser isotope separation. Nature **1979**, *277*, 605.
25. Vanderleeden, J.C. Laser Focus **1977**, *13*, 5.
26. Vanderleeden, J.C. Generalized concepts in large-scale laser isotope separation, with application to deuterium. J. Appl. Phys. **1980**, *51*, 1273.
27. Marling, J.B.; Herman, I.P. Deuterium separation with 1400-fold single-step isotopic enrichment and high yield by CO_2-laser multiple-photon dissociation of 2,2-dichloro-1,1,1-trifluoroethane. Appl. Phys. Lett. **1979**, *34*, 439.
28. Tuccio, S.A.; Harford, A., Jr. Deuterium enrichment via selective dissociation of fluoroform-d with a pulsed CO_2 laser. Chem. Phys. Lett. **1979**, *65*, 235.
29. Marling, J.B.; Herman, I.P.; Thomas, S.J. Deuterium separation at high pressure by nanosecond CO_2 laser multiple-photon dissociation. J. Chem. Phys. **1980**, *72*, 5603.

Heterogeneous Catalysis

Richard W. Rice
James G. Goodwin, Jr.
Department of Chemical Engineering, Clemson University, Clemson, South Carolina, U.S.A.

INTRODUCTION

Catalysts are materials that speed up a reaction by orders of magnitude without being consumed by the reaction. For example, iron (Fe) catalysts increase the rate of ammonia synthesis from nitrogen and hydrogen by a factor of roughly 10^{40}. Basically, catalysts provide an alternate mechanism for the reaction that involves at least one reactant chemically bonded to the catalyst. This alternative mechanism provides a lower energy barrier for reaction to occur, resulting in a lower ''apparent'' activation energy for reaction, but does not alter the thermodynamic features, i.e., heat of reaction or equilibrium constant. There are three classes of catalysts: heterogeneous (or solid inorganic catalysts), homogeneous (typically organometallic compounds, metal compounds, or liquid acids/bases), and enzymes (or biological catalysts). This entry focuses on only one of these classes, i.e., heterogeneous catalysts, but it is useful to briefly point out the main distinguishing features of all three.

As the name implies, a heterogeneous catalyst differs in phase from the reactants and products involved in the process. In general, heterogeneous catalysts are finely divided or porous solids that provide a large active surface area for reaction. They are the preferred catalysts for industrial use because they tend to be stable at higher temperatures, have high rates of reaction, and are easy to separate from the reaction media (gas or liquid) because of their phase difference. However, it is hard to achieve 100% selectivity to desired products. Homogeneous catalysts have higher selectivities, but are limited to lower temperatures because of their thermal stabilities. In addition, because they are in the same phase as the reaction mixture, expensive separation is usually required to recover the catalyst. In some cases, e.g., polymerizations, catalyst recovery is impracticable/impossible. Many homogeneously catalyzed reactions utilize liquid acids such as H_2SO_4 and HF. These strong liquid acids bring problems of process unit corrosion, catalyst separation, and, most importantly, disposal of the acid. With stricter environmental regulations being imposed on the chemical processing industry, processes using liquid acids have become less desirable.

Biocatalysts or enzymes are very interesting because of their extremely high selectivities (near 100%) at close to ambient temperature. For example, methane monooxygenase enzyme, which can be extracted from several bacteria, is able to oxidize methane directly at ambient temperature and pressure with a 100% yield of methanol. However, biocatalysts cannot be used much above 70°C and tend to prefer aqueous environments. Consequently, high rates cannot typically be obtained for industrial applications because of the temperature restriction. Immobilization of enzymes by tethering them to inert solids has been investigated with limited success. Such catalysts have features common to both heterogeneous and homogeneous catalysts.

HISTORY

Although scientific historians often cite Berzelius (1835) as the first to use the term ''catalysis,'' it is probably more meaningful to say that the science and technology of heterogeneous catalysis as we now know it began with Ostwald, who received the Nobel Prize in 1909 for his pioneering work. One of the earliest large-scale commercial uses of heterogeneous catalysis was in an ammonia synthesis plant commissioned by BASF in 1913. Industrial production of methanol from carbon monoxide and hydrogen via catalyzed processes began in the mid-1920s and the Fischer–Tropsch process for hydrocarbon synthesis from CO and H_2 was developed shortly thereafter. Conceptual/theoretical advances were also achieved in the 1920s with H.S. Taylor's hypothesis concerning active centers and Irving Langmuir's work on adsorption.[1] From these beginnings the development of both scientific and technological insight concerning catalysis has continually increased, thereby making possible enormous growth in many industries central to our economy. These include not only petroleum refining and the ''conventional'' chemical industry, but also pharmaceuticals, environmental protection, etc.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120018066

DEFINITION OF IMPORTANT TERMS

Active Site(s)

For most catalysts the accessible surface is nonuniform and typically only a small percentage of sites, i.e., individual atoms or small clusters of atoms, have the ability to cause rapid conversion of adsorbed reactant molecules to the desired product(s).[1] Although research has revealed the nature of such sites for some specific combinations of reaction and catalyst, there is no universally accepted description of what makes a site "active." One school of thought avers that an active site is a location where highly reactive intermediates (chemisorbed species) are stabilized long enough to undergo chemical transformation. For crystal surfaces it has been found for some reactions that only one of various exposed planes has the necessary features, i.e., atomic spacing and/or coordinative unsaturation, for causing a high rate of reaction.[2,3]

Activity

The rigorous definition of activity is the ratio of the rate of reaction (or rate constant) on a catalyst after some time, t, to the rate (or rate constant) on a fresh catalyst at the same conditions.[4] Thus, the loss of catalytic ability with time in use is called "deactivation." However, the term "activity" is also often used to merely refer to the rate constant, k, or to the intrinsic ability of a catalyst to carry out a reaction.

Reaction Rate

Because heterogeneously catalyzed reactions occur exclusively on the catalyst surface, the relevant normalizing quantity (basis) chosen for the rate is usually either the mass of the catalyst, the surface area, or something directly related to these. Thus, for a reaction such as $aA + bB \leftrightarrow mM + sS$, a commonly employed expression for calculating the rate based on mass (or weight) of catalyst is:

$$\begin{aligned}\text{Rate} &= \frac{\text{Moles A reacted}}{\text{Weight of catalyst} \times \text{Time}} \\ &= -\frac{1}{W}\frac{dN_A}{dt} \text{ for a batch reactor} \\ &= \frac{F_{A_0}X_A}{W} \text{ for a flow reactor}\end{aligned}$$

Here, N_A = moles of A, W = weight of catalyst, t = time, F_{A_0} = molar flow rate of A, and X_A = fractional conversion of A. For a given catalyst the rate is dependent on temperature and the concentration(s) of participating species. Often this relation is approximated as a simple power law expression, e.g., rate $\approx kC_A^{\alpha}C_B^{\beta}$, where C_A and C_B are the respective fluid phase concentrations of A and B, α and β are the respective orders of reaction, and k is referred to as the "rate constant" even though it generally is highly dependent on temperature. Typically, an Arrhenius expression is used for k, i.e., $k = k_0\, e^{-E/RT}$, where k_0 is a constant, E is the apparent activation energy, R is the gas constant, and T is the absolute temperature. The orders of reaction (α and β) are not necessarily equal to the reaction stoichiometric coefficients.

Turnover Frequency

The rate of reaction expressed as molecules reacted (or formed) per unit time per catalytic site (or per exposed atom of active metal for metal catalysts) is called the turnover frequency. For supported metal catalysts the calculation requires knowledge of the "dispersion," i.e., the fraction of the active metal available for adsorption of reactants. Boudart coined the term "demanding" (structure-sensitive) for catalyzed reactions for which the turnover frequency varies with the dispersion.[5] Related to this is the "ensemble effect," where the active site requires a specific multiatom grouping.[3]

Selectivity

For cases in which multiple reactions occur, e.g., parallel reactions such as $A + B \rightarrow M$ and $A + B \rightarrow P$, selectivity is the ratio of the rate of formation of one product relative to the other.

Carrier (or Support)

A solid, typically a highly porous metal oxide, which provides an appreciable surface area upon which the active component(s) is (are) dispersed.

Promoter

A substance that has little or no activity itself, but which imparts to the catalyst improved activity, selectivity, or stability.

CATALYST TYPES

The main types of heterogeneous catalysts are as follows.

Metal(s)

A wide variety of metals are used in the reduced (zero-valent) state as catalysts. Most often the metal is

dispersed within the pores of a support, but there are several bulk (unsupported) metal catalysts of commercial interest, e.g., bulk Fe ammonia synthesis catalysts, Pt/Rh wire gauzes, Raney Ni or Co.[6] The most commonly used supports are γ-Al_2O_3, SiO_2, and carbon, and the metal loading may vary from <1 wt% to, say, 40 wt%, depending on the application, support characteristics, and the relative cost of the metal.[7] Supported transition metals, particularly Group VIII metals such as Pt, Ni, and Fe, are frequently used as catalysts. Platinum is an excellent catalyst for many reactions, e.g., hydrogenation, dehydrogenation, oxidation, but is only used when its superior performance offsets its high cost. Often Pt is used in combination with other metals, e.g., Re for naphtha reforming and Pd and/or Rh for automotive catalytic converters.[8,9] Nickel is an active catalyst for hydrogenation and is often used for methanation, i.e., $CO + 3H_2 \rightarrow CH_4 + H_2O$.[6,10,11] Iron and cobalt have been used for Fischer–Tropsch synthesis of hydrocarbon oligomers.[10,12,13]

Metal Oxides and Sulfides

In addition to serving as supports, metal oxides are widely employed as catalysts themselves or as a ''co-catalyst'' with a metal in multifunctional catalysts for reactions involving two or more types of sites, e.g., dehydrogenation/hydrogenation on a metal and isomerization on an acid (oxide or hydroxyl) site. As discussed in the next section, metal oxides catalyze many reactions based on acid–base chemistry, utilizing either Lewis acid sites or Brönsted (proton donating) surface OH groups.[14] For example, γ-Al_2O_3 has been used industrially for alcohol dehydration.[3] The use of zeolites, i.e., crystalline aluminosilicates, as catalysts has continued to grow since their first major application in catalytic cracking of petroleum. The well-defined microporous structure of zeolites has been exploited to improve selectivity for several reactions by allowing only molecules of a restricted range of sizes/shapes to reach the active sites. In addition, by varying the Si/Al ratio and/or the exchangeable cation, the distribution and strength of acid sites can be varied considerably.[3] Metal oxides are also widely used as oxidation catalysts.[15] Usually, the catalyst and conditions are chosen to achieve selective partial oxidation, e.g., V_2O_5, Cr_2O_3, MoO_3, and WO_3 catalyze conversion of hydrocarbons to aldehydes, ketones, and acids. These metal oxides are often used in combination, e.g., as the catalyst used along with NH_3 in the selective catalytic reduction of NO and NO_2 to N_2, a redox reaction for pollution abatement.[9] Although often referred to as metal oxide catalysts, many petroleum hydrodesulfurization catalysts, e.g., alumina-supported cobalt-molybdate or nickel–molybdate, are actually closer to metal sulfide or hybrid oxide–sulfide catalysts because they are active only when properly sulfided.[11]

Acid–Base Catalysts

As mentioned in the preceding section, some zeolites exhibit appreciable surface acidity, which makes them useful for reactions involving carbonium (or carbenium) ions, e.g., cracking, isomerization, and alkylation. The widely used low Al content zeolite HZSM-5 has sites with very high proton donor strength at high temperature. Other solid very strong acid catalysts include sulfated zirconia and SbF_5 adsorbed on silica–alumina.[3] Chlorinated or fluorinated alumina, H_3PO_4-treated clays, and certain polymers, e.g., porous cross-linked poly(styrenesulfonic acid), are examples of more conventional solid acid catalysts.

The essential feature of base catalysis is formation of carbanions. Compared to acid catalysts, solid base catalysts are not as widely used commercially, but can be active for alkylation, condensation (oligomerization), isomerization, and dehydrohalogenation.[12] Common base catalysts are alkali metal oxides, carbonates, and hydroxides, as well as alkaline earth metal oxides.

CATALYSIS FUNDAMENTALS

Pore Structure and Surface Area

Although there are a few processes utilizing nonporous catalysts, e.g., ammonia oxidation over platinum wire gauze, most solid-catalyzed reactions are conducted over porous catalysts because such catalysts provide high surface area per unit mass (or volume). The latter usually consist of a porous support upon which is dispersed the catalytically active material. Such supported catalysts are especially used when the active material is very expensive, e.g., platinum, and a high dispersion is desired or when it would be hard to prepare a stable, high surface area unsupported version of the active material, i.e., most metals. Porous catalysts are often semiamorphous and have a rigid sponge-like structure with a complex network of interconnecting pores of various shapes and sizes. However, for simplicity, such pores are often modeled as being long cylinders of uniform diameter, $\bar{d}$, which can be roughly related to surface area, S_A, and pore volume, V_P, via $\bar{d} = 4V_P/S_A$. In reality, there is generally a distribution of pore sizes consisting of macropores ($\bar{d} > 50$ nm), mesopores (2–50 nm), and micropores (<2 nm). Some porous catalysts may be crystalline, e.g., zeolites, with well-defined micro- or mesopore structures.[16] Typical values for pore volume and surface area are 0.2–0.6 cm^3/g

and 25–300 m^2/g, respectively, but zeolites and activated carbons can have measured surface areas exceeding 1000 m^2/g.

Adsorption Isotherms

Unlike in the bulk of a solid, there is a lack of balance in the attractive forces at the surface. The unbalance can be relieved by attachment of molecules from a contacting fluid, i.e., by adsorption. ''Physical adsorption'' involves weak, e.g., van der Waals, interactions and resembles condensation. In contrast, ''chemisorption'' involves the formation of a chemical bond between the solid and the adsorbing molecule and generally occurs only at specific sites.[17] The equation relating the equilibrium extent of adsorption to the adsorbate partial pressure at constant temperature is called an ''adsorption isotherm.'' One of the simplest is the Langmuir isotherm, which attempts to describe single-layer physical adsorption on an assumed uniform surface.[3] It has the following general form: $\theta = K_A P_A/(1 + K_A P_A)$, where θ is the fractional surface coverage, P_A is the pressure of A, and K_A is the adsorption equilibrium constant. The latter is related to absolute temperature T via $K_A = K_0 e^{Q_{ads}/RT}$, where K_0 is a constant and Q_{ads} $(= -\Delta H_{ads})$ is the ''heat of adsorption.'' For chemisorption, θ is replaced by $C_{A,ads}/C_t$, where $C_{A,ads}$ is the surface concentration of adsorbed species A and C_t is the total concentration of potential sites.

The Coupling of Transport Phenomena and Surface Reaction

For a heterogeneously catalyzed reaction to take place the following sequence of events must occur:

1. Transport of reactant(s) through the bulk fluid phase to the exterior of the catalyst pellet.
2. Transport (diffusion) of reactants through the pores to an active site.
3. Adsorption of reactants at the active site.
4. Surface reaction at the active site (sometimes involving several steps and/or multiple sites).
5. Desorption of product(s) from the active site.
6. Transport (diffusion) of products through the pores to the catalyst's exterior.
7. Transport of product(s) into the bulk fluid phase from the exterior of the catalyst pellet.

This sequence is illustrated in Fig. 1. Steps 1 and 7 are typically complicated by the presence of a thin stagnant (laminar flow) fluid layer between the pellet exterior surface and the bulk fluid.

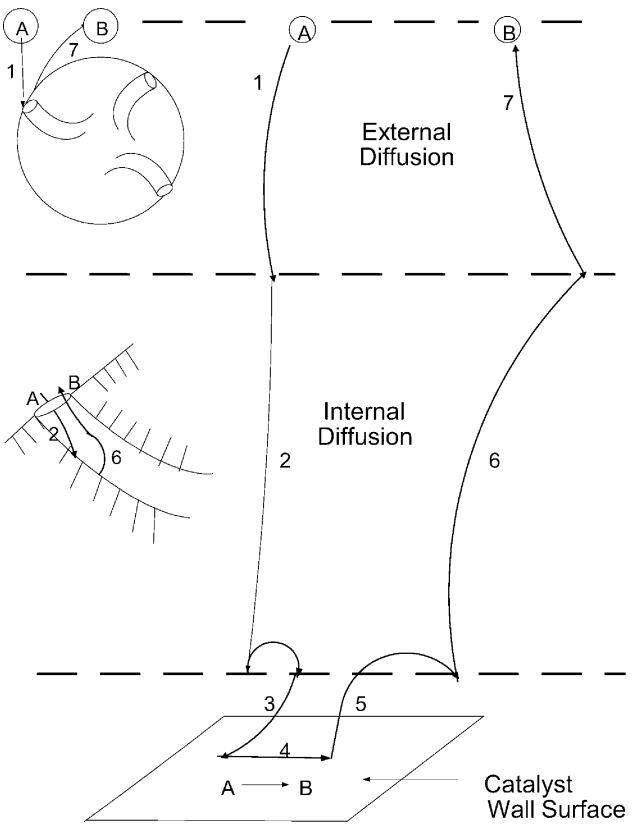


Fig. 1 Overall sequence of steps in a heterogeneously catalyzed reaction. (From Ref.[4].)

An extremely important concept in catalysis is that of the ''rate-limiting step,'' which is analogous to the idea that a chain is no stronger than its weakest link. At steady state the rate of each of the seven steps outlined above must be the same and will be equal to the rate of the slowest step. If external and internal (pore) mass transfer, as well as adsorption and desorption, are rapid enough that surface reaction is rate limiting, the observed rate will be the intrinsic rate, i.e., what the catalytic sites are capable of achieving under the existing conditions. However, if the sites are sufficiently active such that reactants cannot be supplied to them nearly as quickly as the sites could potentially process them, the system is said to be ''mass transfer-limited.'' Because the temperature dependence of surface reaction rate constants (exponential) is much greater than that of mass transfer coefficients or diffusivities (power law), the probability of mass transfer becoming rate limiting increases with increasing temperature. Similarly, it increases with decreasing average pore size and increasing particle size. As a system shifts from surface reaction limited to mass transfer limited, the apparent activation energy decreases appreciably and the apparent reaction order shifts toward first order.[4] When external mass transfer is limiting, the apparent activation energy can be very small, e.g., <8 kJ/mol.

Effectiveness Factor

A measure of the absence of internal (pore diffusion) mass transfer limitations is provided by the ''internal effectiveness factor,'' η, which is defined as the ratio of the actual overall rate of reaction to the rate that would be observed if the entire interior surface were exposed to the reactant concentration and temperature existing at the exterior of the catalyst pellet. A value of 1 for η implies that all of the sites are being utilized to their potential, while a value below, say, 0.5, signals that mass transfer is limiting performance. The value of η can be related to that of the Thiele modulus, ϕ, which is an important dimensionless parameter that roughly expresses a ratio of surface reaction rate to diffusion rate.[4] For the specific case of an nth order irreversible reaction occurring in a porous sphere, $\phi = \sqrt{k' R^2 \rho_c C_{AS}^{n-1} / D_e}$, where k' is the true reaction rate constant based on catalyst mass, R is the sphere radius, ρ_c is the sphere density, C_{AS} is the concentration of reactant A at the exterior, and D_e is the effective pore diffusivity. Figs. 2 and 3 show how ϕ influences the concentration profile within the pellet and η, respectively. Note that when pore diffusion is limiting, η is roughly inversely proportional to ϕ. An ''overall effectiveness factor,'' Ω, that also takes into account external mass transfer may also be calculated.[4]

Although this discussion has focused on mass transfer, heat transfer is also an important consideration, particularly for highly exothermic reactions. In such cases, significant intraparticle and interparticle temperature gradients can occur, with potentially strong effects on rate, selectivity, and stability.[10,18] Criteria have been published for estimating whether mass and/or heat transfer effects can be neglected.[4]

Catalytic Mechanisms and Kinetics

The mechanism of a catalyzed reaction such as A(g) ↔ B(g) + M(g) is the detailed set of elementary steps that show how the reaction is believed to occur on a molecular level, including specifying the rate-limiting step. Various mechanisms may be hypothesized, each generally leading to a different rate expression; thus, only those mechanisms that predict behavior that agrees with experimental results can be considered to be possible true explanations. It is virtually impossible to obtain absolute proof that a given mechanism uniquely describes what actually occurs at the molecular level. The procedure for obtaining a rate equation from a proposed mechanism was first proposed by Hinshelwood based on Langmuir's ideas concerning adsorption; thus, the resulting equation is referred to as reflecting Langmuir–Hinshelwood kinetics.[3] To implement this approach, one must write a rate equation for each of the elementary steps, then use the equations for the non-rate-limiting steps to express surface concentrations, e.g., C_{A^*S}, in terms of more readily measurable fluid phase concentrations, e.g., C_A. The results are then inserted into the rate equation for the rate-limiting step. In effect, the non-rate limiting steps are approximated as being at

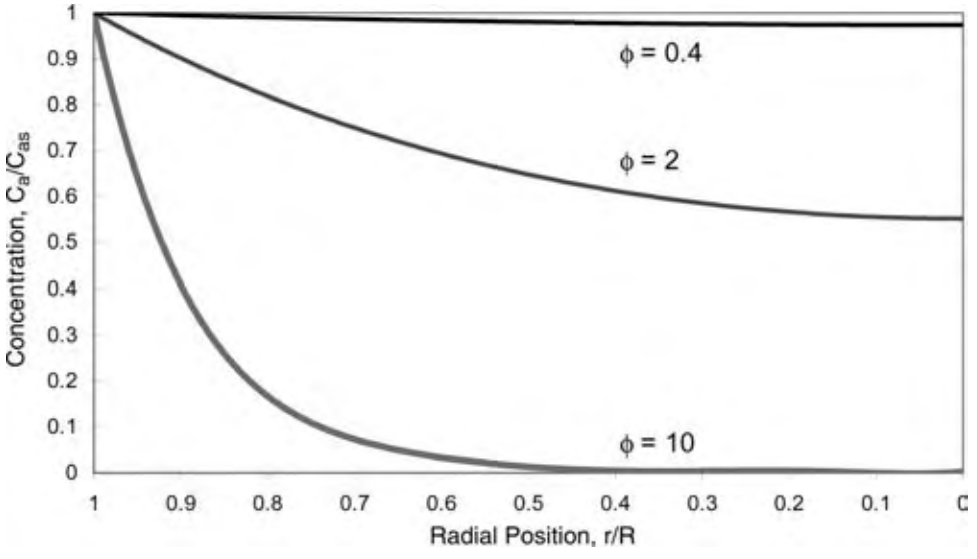


Fig. 2 Concentration profile within a porous catalyst for a first-order reaction. (Cas = concentration of reactant A at exterior surface, R = catalyst pellet radius, ϕ = Thiele modulus.)

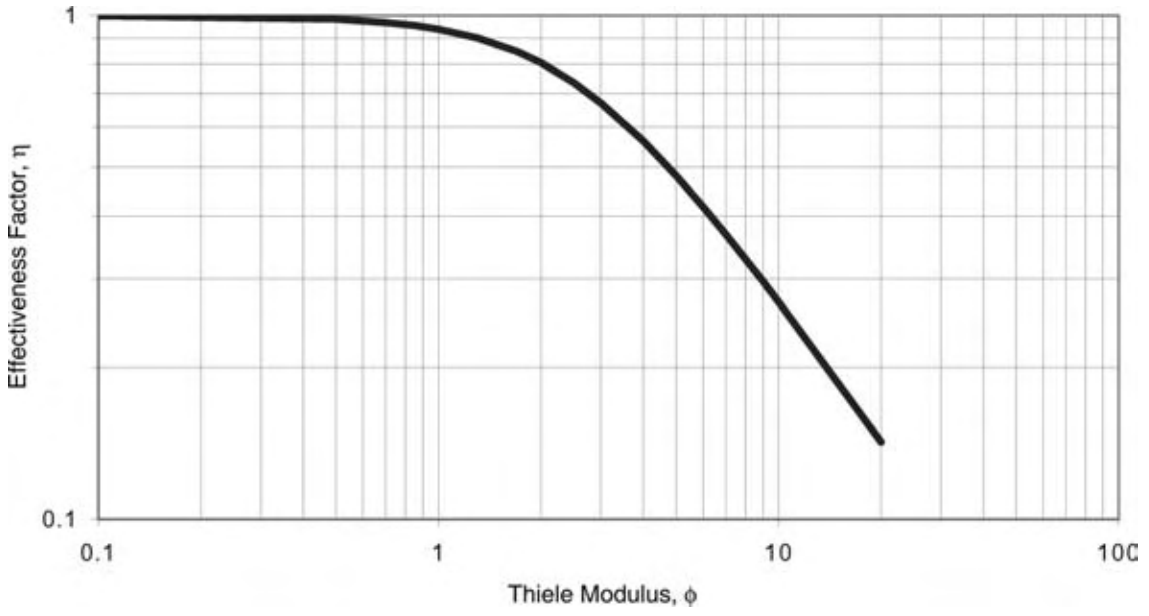


Fig. 3 Effectiveness factor vs. Thiele modulus for a first-order reaction in a spherical catalyst.

equilibrium. As an illustration, consider the following steps where A(g) = gaseous A, S = a vacant site, A*S = adsorbed A, etc.:

1. $A(g) + S \leftrightarrow A^*S$ Adsorption of reactant(s)
2. $A^*S \leftrightarrow B^*S + M(g)$ Surface reaction
3. $B^*S \leftrightarrow B(g) + S$ Desorption of product(s)

If surface reaction is assumed to be rate limiting and irreversible (and no adsorbed inerts are involved), the overall rate expression for consumption of A becomes $-r_A = kK_AC_A/(1 + K_AC_A + K_BC_B)$, where k is the surface reaction rate constant and K_A and K_B are adsorption equilibrium constants. If the surface is only sparsely covered, i.e., $K_AC_A + K_BC_B \ll 1$, this can be approximated as simply $-r_A \approx kK_AC_A = k'C_A$. This illustrates how a simple power law rate expression can apply, under some circumstances, for what is actually a relatively complex mechanism.

Deactivation

Ideally, a catalyst would remain unchanged regardless of time of use, but real catalysts undergo changes that lead to decreased activity and/or selectivity. There are three general categories (causes) of catalyst deactivation.

The first is ''poisoning'' in which small concentrations of impurities in the feed stream strongly adsorb on previously active sites, making these sites inactive. Usually, poisoning is irreversible, e.g., traces of heavy metals such as lead permanently deactivate supported noble metal catalysts used for many processes. In some cases a limited extent of poisoning is intentionally performed to optimize the selectivity of a catalyst for a particular product or reaction. One example of this is the presulfiding of petroleum reforming catalysts using H_2S to suppress undesired hydrogenolysis while only marginally affecting the desired dehydrodecyclization and isomerization reactions.

The second mode of deactivation is ''sintering,'' which denotes loss of surface area owing to crystallite growth. One type of sintering that often occurs for catalysts used at very high temperatures involves collapse of pores. Another type involves agglomeration of the original highly dispersed small clusters of active metal into a smaller number of larger clusters of lower dispersion. This can occur during regeneration of coke-deactivated catalysts if the burning of coke is not carefully controlled. In some cases, in situ methods have been developed to redisperse large clusters, but often extreme sintering requires replacement of the catalyst.

The third form of deactivation is called ''fouling'' or ''coking'' and refers to the formation of high molecular weight carbonaceous species (residues) that cover active sites. Depending on the hydrocarbon feedstock, catalyst, and conditions involved, the ''coke'' may range from graphitic to tar-like. The most common method for regeneration of coked catalysts is controlled combustion, but in some cases high-temperature steam or H_2 treatments have been used.

Various attempts have been made to mathematically model deactivation in a manner similar to that used for reaction kinetics, i.e., to express the rate of change in activity in terms of factors such as temperature, activity,

and concentration of relevant species.[4,19] These models are usually totally empirical, but are useful in predicting the decline in activity with time on stream. Such information can be used to guide the changes in operation, e.g., gradual raising of temperature, needed to compensate for deactivation if a fixed production rate is to be maintained.

CATALYST PREPARATION

Although both the number of applications for heterogeneous catalysis and the body of fundamental knowledge concerning catalysts have increased considerably over time, catalyst preparation still involves a major element of ''art'' (as opposed to ''science''). Because of the sensitivity of catalyst performance to subtle differences in preparation and/or pretreatment, achieving reproducible results for successive batches of catalyst, i.e., good quality control, can be a major challenge, requiring extreme attention to detail. Among the numerous techniques for making solid catalysts, some of the most common are briefly described below. Details concerning these and other methods can be found elsewhere.[6,10,16,20]

Impregnation

This involves first contacting the preformed support (carrier) with a solution, usually aqueous, containing the active metal(s) in the form of a readily decomposable salt, e.g., $Fe(NO_3)_3$, H_2PtCl_6, etc., then performing a series of additional steps. These include: 1) removing excess impregnating solution by filtration; 2) drying in air at, say, 120°C to remove solvent from the pores; 3) calcination (sometimes in an inert atmosphere, sometimes in air) at a sufficiently high temperature, e.g., 500°C, to complete removal of the solvent and to decompose the salt, and, in some cases; 4) pretreatment of the calcined catalyst, e.g., reduction using H_2 to ''activate'' the catalyst. ''Incipient wetness'' is a special form of impregnation in which only just enough solution is used to completely fill the pores. Usually, a nearly uniform distribution of the active component(s) within the porous support is desired and achieving this often requires not only allowing sufficient time for complete penetration by the solution, but also adjusting factors such as pH to avoid preferential deposition near the pellet exterior. Likewise, drying must be gradual to avoid unwanted redistribution (migration) of the active species. Because of the nature of certain reactions, some catalysts are prepared with the active phase intentionally nonuniformly distributed, e.g., ''egg-shell'' structured catalysts that have nearly all of the catalytic metal sites in the outer region of a porous support granule.[10]

Coprecipitation

In this method, the active component(s) and the support are made together by rapidly mixing concentrated solutions of the appropriate salts (plus an agent such as NH_4OH to induce precipitation), followed by filtration/centrifugation, drying, calcining, etc., as described above. Control of pH and agitation during precipitation is important for producing a relatively uniform particle size. This method can generally produce a catalyst with a higher ''loading,'' i.e., surface concentration, of active sites than can impregnation. Sol–gel processing is a technique that may be considered a special case of coprecipitation. Here, typically, a metal alkoxide undergoes hydrolysis/condensation to form a colloidal suspension (sol), which next undergoes further condensation to a wet gel, then conversion to a dry porous solid by extraction/evaporation of solvent.

Catalyst Forming

Most commercial catalysts are used in the form of either spheres or cylindrical pellets of 1–6 mm diameter, with the cylinders typically having a length to diameter ratio ≤ 3. There are three main methods that are used to form such shaped pellets from the irregularly shaped catalyst particles obtained via precipitation or from the powdered alumina, silica, etc. used in making the support for an impregnated catalyst.[6,12]

The first method is tableting (or pelletization), in which powder is mixed with plasticizing agents and die lubricants, e.g., stearic acid, and compressed in a die. A second method is extrusion, where a thick paste of powder and liquid is extruded through a die with many holes, yielding spaghetti-like strands that are subsequently cut into short cylinders, then dried, calcined, etc. Small amounts of clay, stearates, etc., are often added to facilitate extrusion and/or add strength to the formed pellet. The third method is granulation (or tumbling), which is usually used to obtain spheres. Most lubricating agents used in any of these methods are removed by calcination.

The conditions used for pellet forming can have a major influence on several important catalyst properties, including pore size distribution, pellet strength, and abrasion resistance. Both the size and shape of catalyst pellets affect the pressure drop across a packed bed reactor and also, as indicated earlier, affect the Thiele modulus and thus the effectiveness factor. Recently, monolith catalysts have begun to be used in circumstances where low-pressure drop and/or

enhanced heat transfer are important. One example is the use of monolith versions of the ''three-way'' catalyst used in automobile catalytic converters.[4,10] Monolith catalysts generally consist of an inert preformed ''honeycomb'' with many parallel channels that have had a washcoat of active powdered catalyst deposited via contacting with a slurry.

CATALYST CHARACTERIZATION

There are numerous properties that can affect the overall performance, i.e., activity, selectivity, deactivation, and durability, of a catalyst. As a result, full characterization of a catalyst requires using a wide variety of analytical/testing techniques.[21] These can be roughly sorted into three categories, corresponding to the general type of property studied. The following listing is not an exhaustive one, but includes most of the commonly used methods. More detailed information is available in the literature.[6,10,12,22]

Bulk (Nonsurface) Chemical Properties

1. Atomic absorption spectroscopy and inductively coupled plasma can be used for analysis of elemental composition.
2. Infrared, visible, and ultraviolet spectroscopy can determine bonding features, and thus structure.
3. X-ray diffraction can determine the crystal phases present and their average size.
4. Electron dispersive spectroscopic analysis can be used to determine the variation of composition over a cross section of a pellet.
5. X-ray fluorescence and electron diffraction (during electron microscopy) can determine particle composition and structure.
6. Nuclear magnetic resonance can determine the chemical environment of some elements.
7. Temperature-programmed reduction can determine the oxidation state of reducible species.
8. Extended x-ray fine structure analysis can determine the identity, coordination number, and distance of atoms surrounding a given element.

Surface Properties

1. X-ray photoelectron spectroscopy, Auger spectroscopy, and secondary ion mass spectroscopy are sophisticated techniques for determining the chemical nature of surface and ''near-surface'' layers.
3. Fourier transform infrared spectroscopy (FTIR) and attenuated total reflectance spectroscopy can give information on adsorbed species.
4. Physisorption of gases, e.g., N_2 or Ar, commonly referred to as the BET (Brunauer–Emmettt–Teller) method, can determine the total surface area.
5. Chemisorption of gases such as H_2 and CO can determine the surface area of specific reduced metal surface species (and thus the dispersion and average cluster size).
6. Temperature-programmed desorption (TPD) can determine the identity of adsorbed species.
7. Chemisorption of gaseous bases, e.g., ammonia or pyridine, followed by adsorption microcalorimetry, FTIR, and/or TPD, can determine the concentration, ''strength,'' and type of surface acid sites.
8. Steady-state isotopic transient kinetic analysis can determine the concentration and ''relative strength distribution'' of active sites.
9. Electron microscopy can reveal surface topography and structure.

Bulk Physical Properties

1. Mercury porosimetry (for macropores) and BET N_2 physisorption (for micro- and mesopores) can determine both the total pore volume and the pore size distribution.
2. Pellet crushing tests can determine compression strength.
3. Pellet tumbling tests can determine attrition loss, related to abrasion resistance.
4. ASTM fluidized bed and jet cup attrition tests can determine the attrition properties of powdered/pelleted catalysts.

IMPORTANT CATALYTIC PROCESSES

Because of page constraints, this section must be limited to only a handful of the many important processes in which heterogeneous catalysis plays an indispensable role. The coverage is merely intended as an introduction to the general features, e.g., feedstock(s), product(s), catalyst(s), and operating conditions. In addition to references cited under specific topics, Pines' book deals with a wide variety of reactions.[23]

Catalytic Cracking

In terms of throughput, catalytic cracking is the largest industrial catalytic process. Because the catalyst rapidly deactivates owing to coke formation, either

fluidized or moving bed reactors are used. These allow continuous regeneration by tightly controlled burning of coke off the catalyst (in a separate reactor). The feedstock is typically gas–oil (230–540°C boiling range) and the products are high-octane gasoline, middle distillate, and lower molecular weight olefins and paraffins. A typical catalyst used is a rare earth-stabilized Y zeolite imbedded in an amorphous silica–alumina binder. Operation is essentially at atmospheric pressure and the temperature range is normally 480–540°C.[3,10–12]

Reforming

As the name implies, the main purpose of reforming is to rearrange the molecular structure of naphtha to convert it to high-octane number gasoline. Secondarily, it also is a net producer of aromatics and hydrogen for use elsewhere in the refinery. The usual feed is heavy naphtha (95–205°C) consisting primarily of C_6–C_{11} paraffins and naphthenes. The increase in octane number results from isomerization of *n*-paraffins to branched paraffins, dehydrocyclization of paraffins to naphthenes and aromatics, and dehydrogenation of naphthenes to aromatics. The complex network of reactions involved requires that the catalyst be bifunctional, with the supported metal sites that are active for hydrogenation/dehydrogenation acting in concert with the acidic chlorinated alumina sites that are active for isomerization. The most commonly used catalyst is bimetallic Pt–Re/Al_2O_3 with 0.3–0.6 wt% of each metal and 0.7–1 wt% Cl.[8] Presulfiding fresh catalyst with H_2S is performed to add a trace surface concentration of S that suppresses undesired hydrogenolysis, but afterward the S content in the feed must be kept quite low, e.g., 1 ppm. Usually, several fixed bed adiabatic reactors are used in series, with interreactor heating to maintain temperature in the 480–520°C range. Pressure is typically 1–2 MPa and H_2 is introduced with the feed to minimize coking.[3,7,10–12]

Steam Reforming of Hydrocarbons

Steam reforming is a process in which either natural gas or naphtha is reacted with steam to produce H_2, CH_4, CO, and CO_2. After the initial formation of CO and H_2 from naphtha and steam, the primary reactions are the same as for natural gas feed, i.e., $CH_4 + H_2O \leftrightarrow CO + 3H_2$ and $CO + H_2O \leftrightarrow CO_2 + H_2$. The latter reaction is the water-gas shift. The conditions used depend on both the feed and the intended primary product, e.g., H_2 vs. "synthesis gas" (a mix of CO and H_2) vs. CH_4 (from naphtha).[6,11,12] For hydrogen production from methane the typical conditions are 800–850°C, 1–4 MPa, 5 mol H_2O/1 mol C in the feed. The catalyst used is normally Ni/NiO on a CaO–Al_2O_3 support. This is the main process for hydrogen production and will become even more important as H_2 usage grows for applications such as fuel cells and desulfurization.

Ammonia Synthesis

Because of the use of NH_3 in the production of fertilizer and explosives, ammonia synthesis is one of the most important large-scale catalytic processes. The reaction, $N_2 + 3H_2 \leftrightarrow 2NH_3$, is catalyzed using a magnetite, Fe_3O_4, catalyst with several promoters added, e.g., Al_2O_3, K_2O, CaO, MgO. Under the conditions normally used, i.e., 430–480°C, 15–30 MPa, conversion per pass through the catalyst bed is generally less than 80% of the equilibrium value, thus NH_3 is condensed downstream and unreacted H_2 and N_2 are recycled.[6,11,12]

Ethylene Epoxidation

Direct partial oxidation of ethylene over a supported silver catalyst accounts for essentially all of the very large quantity of ethylene oxide produced for subsequent use in making ethylene glycol (antifreeze), polyester fibers, polymers, and various chemicals. In addition to being industrially important, it is one of the most interesting reaction/catalyst systems because no other metal is even remotely comparable to Ag in terms of selectivity and because no other olefins appear to undergo similar reactions on any metal, including Ag.[3,10] The catalyst used is 10–15 wt% Ag on a low-area ($<1\,m^2/g$) α-Al_2O_3, the temperature is 230–280°C, and the pressure is 1–3 MPa. The reaction is believed to occur via a Rideal mechanism in which $C_2H_4(g)$ reacts with a previously adsorbed diatomic oxygen ion to form an activated complex, which subsequently splits at the O–O bond to give $C_2H_4O(g)$ and an oxygen atom adsorbed on Ag.

CONCLUSIONS

For at least the first half of the 21st century the world will continue to rely heavily on petroleum and coal as fuels and as hydrocarbon sources for use in making polymers, etc. Improved versions of existing catalysts, as well as new catalysts/processes, will be vital in making an orderly transition from reliance on nonrenewable resources. Included in this will be the continued development of practicable fuel cell technology and processes for synthesizing clean fuels from coal, tar sands, etc. Catalysis will play a role in the shift toward increased use of renewable/recycled materials and in efforts to minimize air pollution. Catalysts that mimic

biological systems will be developed for use in the pharmaceutical industry and elsewhere.[24] As in the past, improved instrumentation and fundamental understanding will undoubtedly evolve as catalysis meets the challenges of the future.

REFERENCES

1. Taylor, H.S. A theory of the catalytic surface. Proc. R. Soc. Lond. **1925**, *108A*, 105–111.
2. Thomas, J.M.; Thomas, W.J. *Introduction to the Principles of Heterogeneous Catalysis*; Academic Press: New York, 1967.
3. Gates, B.C. *Catalytic Chemistry*; John Wiley & Sons: New York, 1992.
4. Fogler, H.S. *Elements of Chemical Reaction Engineering*, 3rd Ed.; Prentice Hall: Upper Saddle River, NJ, 1999.
5. Boudart, M.; Djéda-Mariadassou, G. *Kinetics of Heterogeneous Catalytic Reactions*; Princeton University Press: Princeton, NJ, 1984.
6. Twigg, M.V. *Catalyst Handbook*, 2nd Ed.; Manson Publications: London, 1996.
7. Thomas, C.L. *Catalytic Processes and Proven Catalysts*; Academic Press: New York, 1970.
8. Sinfelt, J.H. *Bimetallic Catalysts*; John Wiley & Sons: New York, 1983.
9. Farrauto, R.J.; Bartholomew, C.H. *Fundamentals of Industrial Catalytic Processes*; Blackie Academic & Professional: New York, 1997.
10. Satterfield, C.N. *Heterogeneous Catalysis in Practice*; McGraw-Hill: New York, 1980.
11. Rase, H.F. *Handbook of Commercial Catalysts, Heterogeneous Catalysts*; CRC Press: Boca Raton, FL, 2000.
12. Oblad, A.G.; Goyal, S.K.; Ramakrishnan, R.; Sunder, S. Catalysis and catalysts. In *Encyclopedia of Chemical Processing and Design*; McKetta, J.J., Cunningham, W., Eds.; Marcel Dekker, Inc.: New York, 1976; Vol. 6, 420–490.
13. Anderson, R.B. *Fischer–Tropsch Synthesis*; Academic Press: Orlando, FL, 1984.
14. Krylov, O.V. *Catalysis by Nonmetals*; Academic Press: New York, 1970.
15. Hodnett, B.K. *Heterogeneous Catalytic Oxidation*; John Wiley & Sons: Chichester, New York, 2000.
16. Szostak, R. *Molecular Sieves, Principles of Synthesis and Identification*; Van Nostrand Reinhold: New York, 1989.
17. Hayward, D.O.; Trapnell, B.M.W. *Chemisorption*; Butterworths: London, 1964.
18. Petersen, E.E. *Chemical Reaction Analysis*; Prentice-Hall: Englewood Cliffs, NJ, 1965.
19. Butt, J.B.; Petersen, E.E. *Activation, Deactivation, and Poisoning of Catalysts*; Academic Press: San Diego, CA, 1988.
20. Richardson, J.T. *Principles of Catalyst Development*; Plenum Press: New York, 1989.
21. Anderson, J.R. *Introduction to Characterization and Testing of Catalysts*; Academic Press: Orlando, FL, 1985.
22. Somorjai, G.A. *Introduction to Surface Chemistry and Catalysis*; John Wiley & Sons: New York, 1994.
23. Pines, H. *The Chemistry of Catalytic Hydrocarbon Conversions*; Academic Press: New York, 1981.
24. Likhtenstein, G.I. *New Trends in Enzyme Catalysis and Biomimetic Chemical Reactions*; Kluwer Academic: Boston, 2003.

High-Pressure Reactor Design

Joseph M. Lambert, Jr.
Steven C. Hukvari
Parr Instrument Company, Moline, Illinois, U.S.A.

INTRODUCTION

Specific design criteria allow for the manufacturing of vessels and reactors that can be used safely at elevated temperature and pressure. The design criteria and testing of these vessels and reactors are governed by the American Society of Mechanical Engineers (ASME) code. Allowances need to be made for corrosion, welding efficiency, and connections. Complete designs also include provisions for agitation, heating, cooling, and control of the high-pressure reactor.

Many chemical processes are conducted at elevated pressures, particularly those processes that involve a reduction in the number of moles present as the reaction proceeds (e.g., hydrogenation) and those that benefit from increased solvent effects (e.g., supercritical processing). The safe application of these processes requires adequately designed pressure vessels. Vessels intended for use at elevated pressure need to have sufficient strength to safely contain the pressure. This is achieved by the careful selection of material and providing adequate wall thickness to support the intended pressure at the temperature of application. Numerous bodies of experts have developed such design criteria. The criteria discussed below are primarily that of the ASME.

To achieve the desired pressure containment at elevated temperatures, most pressure vessels are made of metal or metal alloys. Suitable materials must be ductile; that is, the material must expand (strain) with applied pressure (stress). It is important to select a material of construction that is in its ''proportional range'' such that the strain is linear, or proportional, to the stress. In this region, the proportionality is called Young's modulus. At sufficiently high stress (the proportional limit), the strain will be more than that predicted by the linear ratio. Up to this point, the material will return to its original dimensions when the stress is removed. Above the proportional limit, permanent deformation will occur and the material will not completely return to its original dimension when the stress is removed. The yield point of many materials is defined as the stress at which a permanent deformation of 0.2% is measured.

The classic approach to pressure vessel design differentiates between a thin-walled and a thick-walled vessel. The formulas for thin-walled pressure vessels are first-order equations and are easier to rearrange and solve for minimum thickness and maximum stress values. The thick-walled vessel formulas provide the most accurate value for the stresses in the pressure vessel wall, but solving the thin-walled equations provides comparatively accurate results and is, therefore, quite useful for preliminary design estimates.

THIN-WALLED PRESSURE VESSEL CALCULATIONS

The cylindrical shell is frequently used in pressure vessel design. For initial designs, it is useful to calculate the stresses in a thin-walled cylindrical shell that is uniformly loaded with internal pressure. For thin-walled pressure vessel calculations to be valid, the radial stresses in the shell need to be negligible. This is usually taken to be a valid assumption when the ratio of the vessel inner radius to the wall thickness (R/t) is greater than 10.[1]

The stresses in a thin-walled cylinder are divided into longitudinal and tangential components. The longitudinal stress that would tend to make a cylinder longer is generally expressed as one-half the tangential stress. The larger, and therefore more significant, is the tangential stress, the stress that attempts to widen or bulge the cylinder. As might be expected, the stress on a pressure vessel wall, expressed in pounds per square inch (psi), is proportional to the applied pressure and to the size of the vessel, while being inversely proportional to the wall thickness. The tangential stress is expressed as $\sigma_T = PR/t$, where P is the internal pressure (in psi); R is the inside radius of the cylinder in (in.), and t is the wall thickness (in.). For metric calculations, N/mm^2 and mm may be substituted for the English units.

A first level of improved results can be obtained by using the mean radius of the vessel ($R + 0.5t$) in these calculations.[1] Making this substitution leads to the following formula for the tangential stress: $\sigma_T = P(R + 0.5t)/t$. Values obtained from the use of this form of the equation provide stress values within 2% of the values calculated using the more exact thick-walled pressure vessel formulas.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007802

THICK-WALLED PRESSURE VESSEL CALCULATIONS

Generally referred to as the Lamé equations, the equations for a thick-walled cylinder are a continuation of the theory of thin-walled pressure vessels.[2] The thick-walled cylinder is viewed as a collection of thin laminar rings. A series of equations can be derived to find the stresses in a thick-walled cylinder, if the strain variation through the wall is such that all of the rings are in equilibrium and the stresses and deformations are consistent at the boundaries between the rings. This is likely a valid assumption for vessels that are not scaling apart or cracking from the applied stress.

For internally pressurized cylinders, the tangential stress on the outside cylinder wall is $2r_i^2p/(r_o + r_i)t$ and the tangential stress on the inside cylinder wall is $(r_o^2 + r_i^2)p/(r_o + r_i)t$, where r_i is the internal radius, r_o is the external radius, p is the internal pressure, and t is the wall thickness. Because r_o is always greater that r_i, the $(r_o^2 + r_i^2)$ term is always greater than $2r_i^2$. Therefore, the stress on the inside wall is always found to be greater than the stress on the outside wall. For this reason, it is the stress on the inside wall that is used for design calculations.

ASME CODE STANDARDS

A standard is a set of technical definitions and guidelines for designers and manufacturers. A code is a standard that has been adopted by one or more governmental bodies or has been incorporated into a business contract. ASME publishes its standards; accredits users of standards to ensure that they are capable of manufacturing products that meet those standards; and provides stamps that accredited manufacturers place on their products to indicate that a product was manufactured according to the ASME code.

A third-party inspector may be called upon to verify that a pressure vessel has been designed and manufactured according to the requirements set forth in a particular standard. Documentation from these inspections is often required for purchasing or installing a pressure vessel.

ASME PRESSURE VESSEL CALCULATIONS

The ASME code formula for the thickness of a cylindrical shell is listed in UG-27, as $t = PR/(SE - 0.6P)$.[3] In this formula, t is the minimum thickness of the shell (in.), P is the maximum allowable working pressure (MAWP) (psi), R is the internal radius of the vessel (in.), S is the allowable stress in the material listed in ASME Section II, and E is the weld joint efficiency. The allowable stress, S, includes a variety of factors including ultimate tensile strength of the material, the yield values of the material, and a safety factor to prohibit the possibility of underdesigning. ASME code permits the use of this formula until t exceeds $0.5R$.

The structure of this formula can quickly be related to the thin-walled pressure vessel cylinder equation. Using the equation that calculates the stress at the center of the vessel wall, $\sigma_T = P(R + 0.5t)/t$, and rearranging to solve for the thickness, results in $t = PR/\sigma_T - 0.5P$. The addition of the weld joint efficiency, E, and changing the coefficient before P to 0.6 results in the ASME code formula, $t = PR/(SE - 0.6P)$, which they feel best represents the minimum wall thickness required to contain an internal pressure, P, in a cylindrical vessel having a radius, R, and made of a material with an allowable stress, S.

A comparison of the equations developed above is shown in Fig. 1. Each of the four formulas, referred to as Thin-1, Thin-2, ASME, and Thick-walled, were used to evaluate the stress on the walls of the same hypothetical pressure vessel (with a radius of 2 in. and a pressure of 1000 psi) at various wall thicknesses between 1/8 and 7/8 in. The resultant stresses for the first three formulas were then normalized to the calculated thick-walled vessel stress and plotted to demonstrate the relative accuracy of the various methods.

ASME FORMULA FOR ELLIPSOIDAL HEADS

The ASME design formula from UG-32 for ellipsoidal heads having a major-to-minor axis ratio of 2 : 1 and being subjected to pressure on the concave side is $t = PD/(2SE - 0.2P)$, where t is the minimum thickness of the head (in.), P is the MAWP (psi), D is the internal diameter of the major diameter of the ellipsoid (in.) (and equal to the inside diameter of the vessel), S is the allowable stress of the material, as listed in ASME Section II, and E is the weld joint efficiency.[3]

ASME FORMULA FOR CIRCULAR FLAT HEADS

The ASME design formula from UG-34 for circular flat heads subjected to internal pressure is $t = d(CP/SE)^{1/2}$, where t is the minimum thickness of the head (in.), d is the internal diameter of the vessel (in.), P is the MAWP (psi), S is the allowable stress in the material listed in ASME Section II, and E is the weld joint efficiency.[3] The value of C ranges from 0.10 to 0.75 depending on the method of attachment of the head and the shell dimensions. For preliminary designs, a value of 0.33 for C will lead to a good approximation of the required head thickness.

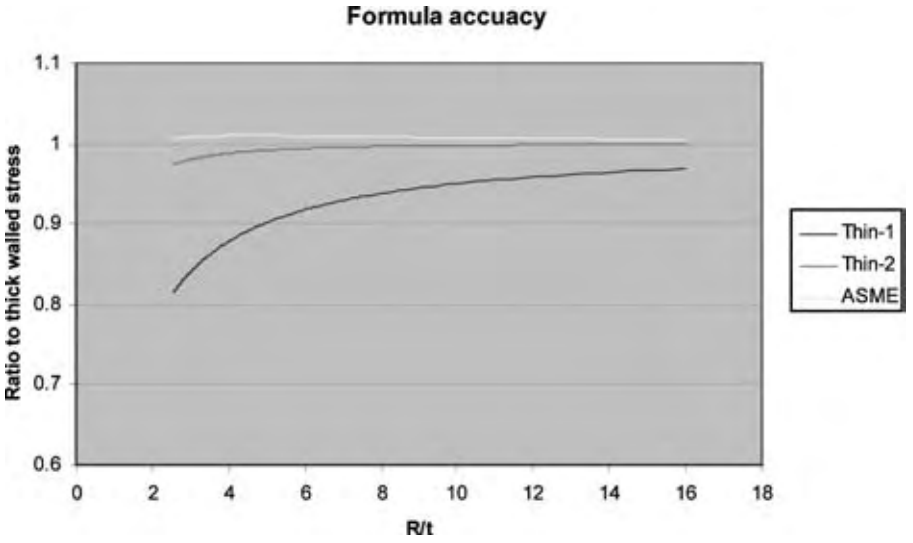


Fig. 1 Comparison of pressure vessel design formulas. (Equation Thin-1 is $\sigma_T = PR/t$; Equation Thin-2 is $\sigma_T = P(R + 0.5t)/t$; ASME equation is $\sigma_T = P(R + 0.6t)/t$, assuming a weld efficiency of 1). *(View this art in color at www.dekker.com.)*

Closures between heads and cylinders, often taking the form of clamps, split rings, or bolts, are also an important consideration in the proper design of the pressure vessel, for it is these components that must hold the vessel head to the cylinder. It is important to recognize the force that is exerted on the closure. For example, in a 1000 psi vessel, a circular head with a diameter of 12 in. will have over 110,000 lb of force pushing on it. This force must be evenly divided between bolts or spread out over the clamp or split ring closure.

ASME VALUES FOR ALLOWABLE STRESS

ASME publishes allowable stress values for materials in Section II, part D. Values for all ASME approved materials are listed in tabular form for operating temperatures ranging from −20°F to the maximum allowable working temperature of that material. These values are constantly under review and periodically change when better information is available. The basis for determining the tabulated stress values at a specific temperature is using the lesser of the tensile strength of the material divided by 3.5 or the yield strength divided by 1.5. Additional factors such as creep strength are considered by ASME materials committees and are factored into the final published values in the tables. The exact values of these derating factors used to calculate the final stress values are proprietary to ASME. Examples of allowable stress values are shown in Table 1.

Prior to the publication of the 1999 addenda to the ASME Boiler and Pressure Vessel Code, the criteria for determining the tabulated stress values at a specific temperature was based on the lesser of the tensile strength of the material divided by 4 or the yield strength divided by 1.5, commonly stated as the lesser of one-fourth the tensile or two-thirds the yield. The minimum value for the hydrostatic test pressure was 1.5 times the MAWP, corrected for temperature. In 1999, an increased confidence in material manufacturers and their testing programs prompted the ASME Materials Committee to rework all of the published allowable stress values based on the tensile strength of the material divided by 3.5 rather than 4. A consequence of this change was the reduction of the hydrostatic test multiplier from 1.5 to 1.3.

It is now common to hear it expressed that the "ASME code imparts a $3\frac{1}{2}$-to-1 safety factor." It is often common, in addition, that users of pressure reactors tend to impart their own safety factor of 2 or 3, so that the resultant vessel is designed with an overall safety factor of 10. Not only is such a vessel overdesigned, but it leads to operational difficulties such as excessive weight, high closure torque requirements, increased number of bolts, and poor heat transfer rates through the unnecessarily thick walls.

Table 1 Allowable stress values for selected materials

	Allowable stress (psi)		
Temperature (°C)	T316 SS	Alloy C-276	Titanium (Gr2)
0	20,000	27,300	14,300
100	20,000	27,300	12,136
200	19,412	27,300	8,912
300	17,280	25,676	6,780
400	16,092	23,484	316°C max.
500	15,472	22,472	

(From Ref.[3].)

STANDARD HYDROSTATIC PROOF TEST

The ASME code (UG 99) requires a hydrostatic test of each pressure vessel to validate the design, materials, and construction before that vessel is put into service. The minimum value for the test pressure is 1.3 times the MAWP, corrected for the temperature at which the test is conducted. This temperature correction is the ratio of the stress value at the hydrostatic test temperature (usually room temperature) to the stress value of the material used in the design formulas, usually at the maximum allowable working temperature of the pressure vessel. To conduct the test, water is pumped into the vessel to increase the internal pressure to the level of the test value and then released.

For example, a T316 stainless steel vessel designed for use at up to 1000 psi at 300°C needs to be hydrostatically tested at room temperature to a pressure of (1000) (1.3) (20000/17280) or 1505 psi. The room temperature and 300°C stress values of 20,000 and 17,280 psi, respectively, are found in Table 1.

ALTERNATIVE PROOF TESTS TO ESTABLISH MAWP

Recognizing that some reactor designs may be impractical or very difficult to completely validate with design formulas, ASME UG-101 offers alternative methods, such as using strain gages, deflection measurements, or pressurizing a vessel to failure, to determine the MAWP.

The strain gage and deflection methods require the vessel to be pressurized and then released at successively higher pressures until a permanent deflection of 2% is measured. The MAWP of the vessel is established by dividing this test pressure by 2.5.

In another type of deflection testing, the vessel is pressurized with water to the test pressure and then released. This procedure is repeated three times. The outside diameter of the vessel is measured and recorded at each successive increase and release of pressure. A vessel is considered to have passed the test if it expands to the same dimension each time it is pressurized and contracts to the same original dimension each time the pressure is released.

Another acceptable method of establishing the MAWP for a vessel is to pressurize it until it bursts. This burst pressure is divided by 5 to calculate the MAWP of the design.

The permanent deflection and burst methods of these alternative testing procedures are typically conducted at room temperature and, hence, provide maximum allowable pressure ratings for a vessel when used at room temperature. The corresponding pressure rating at an elevated temperature may be established by multiplying the calculated room temperature rating by the ratio of the stress value of the material at the maximum allowable working temperature of the pressure vessel divided by the stress value at room temperature. A vessel rating established by this method is only applicable for that specific design. Any alteration or subsequent changes to the design would require repeating the pressure tests to validate the new design.

WELD JOINT EFFICIENCY

Welding may be required on pressure vessels to construct seams or attach heads, nozzles, and flanges. External jackets for heating and cooling may also be welded to the main chamber of a pressure vessel. The value of the weld joint efficiency, E, is governed by the type of weld and the extent of radiographic examination of the weld. Radiography is a nondestructive examination of welds using x-rays to discover discontinuities or defects within the weld. Double-welded butt joints, for example, have an efficiency of 1.0 when they are fully radiographed but an efficiency of only 0.7 when no radiography is done.

Vessels can be designed with thinner walls when full radiography is performed to give a weld joint efficiency of 1.0. Conversely, the cost of radiographic examination can be traded for the increased cost and weight of a thicker vessel having a weld joint efficiency of 0.7. Such trade-offs need to be considered early in the design process. Welds that fail the examination must be reworked and retested.

The weld joint efficiency values of various types of welds are tabulated in ASME UW-12. The design of welded vessels should also take into account the capability of the prospective welders as well as the limitations of the existing welding procedures. New material combinations or variations in material thickness may require welding procedures to be revised and requalified. The person performing the welding needs to be certified for each material and procedure. Welding processes used for pressure vessel fabrication are usually gas tungsten arc welding (GTAW) and submerged metal arc welding (SMAW). Both of these processes are compatible with a wide variety of metals.

CRYOGENIC APPLICATIONS

Materials such as austenitic stainless steels, nickel-based alloys, and titanium alloys can be used as materials for pressure vessel components in cryogenic applications at temperatures as low as −200°C. Alloy steels have brittle transition points making their impact properties at low temperatures unsuitable for pressure applications. Closures and bolts must also be made of materials that remain ductile at low temperatures.

Gaskets made of polytetrafluoroethylene (PTFE) or silicone are commonly used in cryogenic applications.

TYPES OF LOADING TO CONSIDER, ASME UG-22

The design of laboratory reactors and similarly small vessels can be completed with consideration of only the effects of internal pressure. As vessels and tanks get larger for full-scale operations, the design must also consider the effects of additional external loads. Typically, these factors need to be included in vessels larger than 40 L. These loadings include:

1. The weight of the vessel and its contents.
2. Static loading from attached piping, insulation, internals, or external supports.
3. Dynamic loading due to pressure or thermal variations.
4. Wind, snow, and seismic loading.
5. Impact from water hammer shock during filling or draining.
6. Stresses induced by thermal gradients and differential thermal expansion.

SEALING

Three types of seals are commonly used to attach heads and wide ports to pressure vessels: flat gaskets, O-rings, and angled metal gasket seals. Flat gasket sealing surfaces are easy to machine and can be used with a variety of sealing materials. Lower-temperature flat gasket seals are often made of PTFE. Higher-temperature seals can be made of flexible graphite and can be used for temperatures in excess of 600°C. Both PTFE and flexible graphite have excellent chemical resistance properties. Soft metals, such as copper, nickel, or silver, can also be used in flat gasket designs.

Flat gasket seals must be preloaded with a force equivalent to the force that will be exerted by the pressure within the vessel. A flanged bolt closure or split ring bolt closure design can be used to achieve this preload. Closure bolts should be evenly tightened to a specified torque that will provide the proper closing force.

Grooves for O-ring seals can be machined into the head and cylinder walls of a pressure vessel to produce a gland type of seal arrangement. O-ring seals such as this do not require a preload on the closure parts, thus making assembly and disassembly of the vessel quicker. The elastomers used in O-ring seals introduce temperature and chemical compatibility restrictions. Practical temperature limits for O-ring seals are 150°C for NBR (Buna-N) and 225°C with FKM (fluorocarbon) material. Perfluoroelastomer (FFKM) material may not only increase the temperature limit to about 275°C, but also significantly increase the cost of the seal. The chemical compatibility of specific O-ring compounds should be checked against the reactants, intermediates, products, and by-products expected for each vessel application.

High-temperature reactions, incompatible with flexible graphite seals, may be sealed with angled metal gaskets. The preloading forces needed to deform and seal this type of gasket are significantly higher. These gaskets are best applied to smaller-diameter vessels where they are easier to install, tighten, and seal.

INTERNAL STIRRING

Internal stirring enhances many reactions.[4] Flat or pitched blade turbine impellers are typically used for good general mixing. Multiple turbine impellers are often employed, positioning one near the bottom of the vessel and the other at the vortex of the mixing liquid. Turbine impellers are typically sized to have a diameter of 30–60% of the diameter of the vessel in which they are used. Fluids with a high viscosity, above 50,000 cP, could benefit from anchor-type stirrers that slowly mix the vessel contents from the walls inward. Spiral stirrers are often used to keep heavy suspensions properly mixed. Gas entrainment impellers with hollow shafts are used in systems where high revolutions per minute can be obtained and there is a need for recirculating headspace gas into the liquid. Gas dispersion impellers are used to prevent gases introduced through a bottom sparge ring from traveling up the center shaft and avoiding contact with the bulk liquid.

Internal baffles are used to improve mixing in fluids that tend to rotate along with the impeller rather than being mixed. Baffles reduce the flow of fluids to speeds much lesser than that of the impeller. This increase in shear provides additional turbulence leading to improved mixing of the fluids. Baffles are typically designed to have a width of 1/12 to 1/10 of the vessel diameter and are located slightly off of the internal vessel wall to prevent accumulation of the material.

To prevent leakage and loss of pressure, a seal must be made between the vessel and the rotating agitator shaft. This seal can be made either with a magnetic coupling or with a mechanical seal.

Magnetically coupled drive mechanisms have a set of inner magnets located atop the impeller shaft that extends down into the pressure vessel. This inner magnet assembly is positioned so that it can be rotated within and be encapsulated by a small pressure vessel chamber. This chamber is made of nonmagnetic material and is attached and sealed to the head of the reactor vessel. This chamber is then surrounded by another set of matching magnets located in an

assembly that is supported in such a way as to permit its rotation with an attached motor. Rotation of the outer magnets will cause the inner magnets also to rotate, thereby providing a rotating shaft sealed to prevent leaking.

Mechanical seals are of two different types: single and double. Single mechanical seals are available in a variety of styles that provide a means of pressing a seal or set of seals against the shaft. Such a seal will ultimately wear off because of friction and high temperature. A means is usually provided to permit continuous or periodic tightening of the seal, such as spring loading or compression loading with a packing nut.

A double mechanical seal employs two seal arrangements, each similar to the single mechanical seal. A fluid, typically oil, is pumped through the region between the two seals. This oil serves the purpose of removing the heat of friction to add life to the seal ring and to create a liquid film to aid in sealing. The oil is pumped into the space between the two seal rings at a pressure slightly higher than the operating pressure of the vessel. This allows the lower seal to function as the primary pressure containment seal. The oil can be monitored for the presence of leaking material and provide early warning of failure. The upper seal prevents the escape of oil and vessel contents to the atmosphere.

ASPECT RATIO

The aspect ratio of a pressure vessel, defined as the ratio of the internal length to the internal diameter, is also important for improved mixing and heat transfer. As important as these factors may be, they must often be balanced with the cost of the pressure vessel. As the diameter of a vessel increases, the thickness of the wall must increase proportionally for safe operation at the same pressure. The heat transfer rate can become limited because of the thicker wall and because of the decreased surface to volume ratio of vessels with larger diameters.

SURFACE FINISH

The inside finish of a pressure reactor is an important physical attribute that can affect its functionality. A smoother surface finish is less conducive to corrosion and is easier to clean. The smoothness of a surface finish is expressed as the roughness average (R_a) of a surface. This is the average of the absolute values of the surface deviations from a centerline of these deviations. The values are typically stated in microinches or micrometers. The roughness is measured with instruments or prefinished sample blocks can be used for comparison. Typical machined finishes can range from 50 to 100 μin. R_a depending on the machining process and tooling used. Mechanical polishing can further smooth the finish to an R_a value of approximately 20 μin. Mechanical polishing may be followed by electropolishing to achieve extremely smooth ($R_a < 20$ μin.) finishes on stainless steel surfaces.[5]

Electropolishing cleans and polishes metals in an electrochemical bath. An electric current passing through the bath dissolves a small amount of metal from the surface of the part being polished, leaving a smooth, highly reflective surface. Electropolishing removes surface impurities and leaves the part passivated. Care should be taken to minimize mechanical scratches and polishing striations before a part is electropolished. These markings become accentuated rather than removed by electropolishing. Metals other than stainless steel may be electropolished, but the special techniques required are not commonly available with most vendors.

Abrasive blasting techniques can be used for large or irregular surfaces provided the abrasive medium is free of contaminants that could become embedded into the metal surface. Internally welded vessels should be finished according to prescribed welding procedures. Wire brushing or grinding should be done with materials compatible to the welded material, and at speeds that will not embed material into the weld.

CORROSION ALLOWANCE, ASME UG-25

Pressure vessels are subject to thinning by corrosion, erosion, or mechanical abrasion. To increase the desired useful life of the vessel, the design should include a suitable increase in wall thickness over the minimum design thickness required for safe pressure containment. In most cases, there is no specific code requirement for how much corrosion allowance a vessel requires. Vessels subject to corrosion should have provisions for complete draining as well as openings to allow for the inspection of internal surfaces.

Corrosion is generally considered to be in one of two forms: uniform or localized. Uniform corrosion occurs evenly on all surface areas of a pressure vessel. Such corrosion occurring at slow rates is quite typical and can be planned for by increasing the original wall thickness design. For vessel systems with a predictable corrosion rate, allowances can be calculated for vessel life spans of 2–20 yr.[6]

Localized corrosion occurs at a particular site, typically because of a breakdown of the passivation protection layer. This can often occur because of anionic attack from chlorides and similarly aggressive components in the vessel. Such localized corrosion is manifested as pitting at various sites across the surface or as crevice corrosion within the material extending beneath the surface.

Additionally, material can lose its alloying, such as when high temperature carbide formation leaches chromium out of stainless steel. This can lead to intergranular corrosion at grain boundaries and to stress corrosion cracking. One means of eliminating or greatly reducing this possibility is to use alloys that are low in carbon content as the material of construction.

Proper selection of the material of construction can avoid many catastrophic corrosion problems. Stainless steel is an excellent choice for many organic acids, but is rapidly attacked by halogen acids. Although readily attacked at elevated temperature, pressure, and concentration, stainless steel offers some protection against sulfuric, nitric, and phosphoric acid at low concentration and ambient temperature.

Sodium hydroxide, even at low concentrations, will cause stress corrosion cracking in stainless steel when subjected to temperatures above 100°C. Caustic solutions are much better handled in alloys with high concentrations of nickel, such as Alloy 400 or Alloy 600. Alloy 400 has also found success in applications involving fluoride, hydrogen fluoride, and hydrofluoric acid.[7]

Alloys rich in nickel, chromium, and molybdenum, such as C-22, C-276, and C-2000, offer the broadest range of corrosion inhibition. The high molybdenum content assists in diminishing corrosion by reducing acids, while the high chromium content assists in diminishing corrosion by oxidizing acids.[8]

Alloy B-3 can tolerate quite high concentrations of hydrochloric acid without severe corrosion, provided the solution is free of oxidizing ions such as Fe^{+3}, Cu^{+2}, Ni^{+2}, Mo^{+6}, and Ti^{+4}.[8] In contrast, hydrochloric acid solutions containing these ions can be easily handled in vessels made of titanium, which would quickly corrode in pure hydrochloric acid. Even low concentrations (<30 ppm) of ferric ion in hydrochloric acid have been demonstrated to greatly reduce the corrosion rate of titanium.[9]

PROCESS CONNECTIONS

Pressure vessels are usually designed with a variety of process connections. Ports for the attachment of valves are included for gas and liquid reagent addition, as well as product sampling and removal. Ports on top of a vessel can also be equipped with dip-tubes to permit access to the bottom of the vessel. Similarly, vessel bottoms can be ported for drain valves or equipped with inverted dip-tubes, usually to act as overflow weirs. Ports are also often provided for the installation of internal cooling coils.

Most pressure vessels are also equipped with a Bourdon-tube, analog pressure gage to indicate the internal pressure of the vessel at all times. This gage can also be used to verify any electronic readings taken by attached electronic pressure transducers. A proper pressure gage would have a usable range of about 1.5 to 2 times the maximum working pressure of the vessel.

Heated or chilled vessels require provisions for measuring the temperature of the vessel contents. Cooling coils, internal or external, can be used in conjunction with heaters to help control and stabilize process temperatures.

A relief valve or burst disk with a rating equivalent to or less than the rating of the vessel is not only a prudent accessory, but is also required by most safety codes. ASME recommends that the burst rating of a safety rupture should not exceed 110% of the MAWP. Rupture disks may be of a prebulged or scored type. Prebulged disks can be operated repeatedly to pressures up to 70% of their rating. Scored disks can be used repeatedly to pressures up to 90% of their rating.

In some cases, the pressure rating of various accessories that might be added to a pressure vessel may be lower than that of the vessel to which they are attached. In such cases, it is common to rate the overall system with a maximum operating pressure that is lower than the vessel's MAWP.

HEATERS

The contents of a pressure vessel are most commonly heated with external heaters. A variety of heater types, watt densities, and mounting techniques are available to the pressure vessel designer.

Clamp-on band heaters, for use up to 700°C, are specified when high-watt densities are required or the external vessel area is limited. Heating elements can be woven into a quartz fabric to permit flexibility and ease of removal. These heaters are typically specified for lower-temperature demands because of their lower watt density. More rigid heater designs employ fully exposed, radiant heating elements or heating elements embedded within a ceramic fiber matrix. The latter are the heater of choice for high-temperature applications.

Electric heating elements can be sealed within aluminum blocks to provide a more uniform heat distribution and a protective measure of safety in hazardous, explosive environments. Cooling channels can also be designed into the aluminum blocks to provide external cooling for the vessel.

External chambers, such as welded jackets, can provide a means for both heating and cooling from external sources such as steam generators or circulating baths. These jackets are themselves pressure vessels and have the same design considerations as the inner vessel.

CONTROL SYSTEMS

Control systems are used to monitor and regulate temperature, pressure, flow rates, stirring speeds, and compositional characteristics in a pressure reactor. Commercially available process controllers can sense and display temperatures and regulate heaters for temperature control. Controllers may also be specified to sense pressure and operate control valves for pressure control, adding reagents or relieving contents. Motor controllers are available for both AC and DC motors. They can vary from those with simple manual adjustment of speed settings to closed loop controllers that will maintain constant revolutions per minute, torque, or power to the motor.

Various types of probes are employed to monitor compositional changes. Among these are pH probes, dissolved oxygen probes, and high-pressure infrared probes. Such probes can be off-line, on-line, or in-line. Off-line probes require that samples be withdrawn from the reactor for analysis. In-line probes are typically installed on a slipstream that can be diverted from the reactor. In-line probes are installed directly into the reactor for real-time monitoring. Signals from the probes are interpreted by probe-specific electronics that, in turn, send a signal to the process controller for possible action.

Integrated microprocessor-based and PC-based control systems require correctly specified and programmed software, as well as an interface system between the process and the computer. Most commercially available control systems are user-configured with appropriate signal conditioning modules that convert the analog process signals into the digital signals acceptable to a computer. Most modules are multichannel boards designed to handle analog inputs, analog outputs, digital inputs, or digital outputs.

CONCLUSIONS

The safe design of a pressure vessel takes into account the strength of the material and the stresses that are imposed on it by internal pressure and exterior forces. The approach to satisfactory design can be best understood with an appreciation of the equations of a thin-walled vessel. Expansion of these concepts to the implications of a thick-walled vessel will lead to the code rules written by the ASME.

This governing body has published proper design equations and assigns allowable stress for each of a variety of materials of construction as a function of temperature. In addition to published formulas for common shapes of pressure vessels, the ASME code also provides methodologies to test irregular shapes by evaluating their point of permanent deformation or their bursting pressure.

The design of pressure vessels must also take into consideration weld efficiency, number and size of openings, and also add additional thickness to allow for controlled corrosion over time and use.

The complete design will also incorporate the aspect ratio needed for stirring or other desired flow paths, special finish as required by the process, and control of the vessel heating and cooling.

REFERENCES

1. Bednar, H.H. *Pressure Vessel Design Handbook*; Van Norstrand Reinhold Company: New York, 1981; 35–50.
2. Lindberg, M.R. *Mechanical Engineering Reference Manual*, 9th Ed.; Professional Publications, Inc.: Belmont, CA, 1995; 14–22.
3. ASME. In *ASME Boiler and Pressure Vessel Code*; The American Society of Mechanical Engineers: New York, 2001; Section VIII, Div. 1, 19–23, 33, 38, 79–87, Section II, Part D, 70–73, 234–237, 246–249.
4. Perry, R.H.; Green, D.W.; O'Hara, J. *Perry's Chemical Engineers' Handbook*, 7th Ed.; McGraw-Hill: New York, 1997.
5. Odberg, E.; Jones, F.; Horton, H. *Machinery's Handbook*, 23rd Ed.; Industrial Press, Inc.: New York, 1990; 700.
6. Megyesy, E.F. *Pressure Vessel Handbook*, 11th Ed.; Pressure Vessel Publishing: Tulsa, OK, 1998; 221.
7. Special Metals Corporation. In *High-Performance Alloys for Resistance to Aqueous Corrosion*; Special Metals Corporation: New Hartford, NY, 2000.
8. Haynes International. In *Corrosion Resistance of Hastelloy Alloys*; Haynes International: Kokomo, IN, 1984.
9. Titanium Metals Corporation. In *Corrosion Resistance of Titanium*; Titanium Metals Corporation: Denver, CO, 1993.

Hollow Fiber Technology

Vicki Chen
Pierre Le-Clech
UNESCO Centre for Membrane Science and Technology, School of Chemical Engineering, University of New South Wales, Sydney, New South Wales, Australia

INTRODUCTION

Hollow Fiber technology has transformed the membrane separation processes for over 40 years. By utilizing analogous fabrication techniques to textile fiber spinning, hollow fibers allows extremely large surface areas to be generated for heat and mass transfer within a compact module. The surface of the hollow fiber can be used as a contactor between different phases as well as a selective separation layer for diffusive, adsorptive, reactive, or convective processes. The commercial importance of hollow fiber technology is particularly evident in biomedical, filtration, and gas separation applications. This entry gives an overview of widely used hollow fiber materials, fabrication, modules, and applications. Operational and design issues inherent to hollow fiber modules are identified. New and potential hybrid processes and fiber materials/fabrication are also briefly discussed.

MATERIALS AND FABRICATION

Materials

As shown in Table 1, hollow fiber membranes can be synthetically manufactured from a large number of materials, which are generally classified into two groups, i.e., organic (polymeric) or inorganic (mainly glass or ceramic). The most important class of membrane material, especially for hollow fiber technology, remains the organic polymers. Because of their high impact on the market, polypropylene (PP), polyethylene (PE), polyethersulfone (PES), polyvinyl pyrrolidone (PVP), and polyvinylidene fluoride (PVDF) are the most common materials used for the fabrication of hollow fiber membranes. Supplied by leading membrane companies like Memcor, Mitsubishi Rayon, Asahi, Zenon, and X-Flow, these materials are easily spun and configured into dense and porous hollow fiber formats and are applied in a large range of applications (described in detail in ''Hollow Fiber Applications''). Other organic materials used in the fabrication of hollow fiber include compounds based on cellulose [cellulose acetate (-CA-) and triacetate (-CTA-), regenerated cellulose (-RC-)] and nitrogen [polyamide (-PA-), polyacrylonitrile (-PAN-)]. More recently, some hollow fibers based on material alloys (PES/PVP) and more complex organics (like polyetherimide, polybenzimidazole) are also commercially available or have been studied experimentally.

Although they provide higher chemical and temperature stability, inorganic-based hollow fibers are not as common as organics. Because of their higher production cost, lower permeability, and their tendency to break easily, only few companies, like CEPAration and Schott/Bioran, propose inorganic hollow fiber membrane systems, based on ceramic (aluminum oxide) and glass hollow fiber, respectively. Still, research for new inorganic materials is being carried out; Liu and coworkers[1] worked on the fabrication of orthorhombic perovskite hollow fiber, while Ismail and David[2] recently reviewed carbon-based hollow fiber technology.

Properties

Membrane morphology

Membrane could be configured into two types of geometry. While flat sheet membranes are generally assembled into plate and frame or spiral wound modules, tubular membranes can be externally or self-supported and are subdivided into different categories based on their bore diameter. Different terms and diameter thresholds could be found in the literature.[3,4] However, tubular membranes with a maximum external diameter of around 2 mm and that are self-supporting are generally considered as hollow fibers. Further discussion and comparison based on fouling, cleaning, packing, and capital investment is given for the different membrane configurations in Refs.[4,5].

Membrane morphology (or structure) is generally classified into two groups: porous and dense. Dense fibers utilize the chemical and the physical characteristics of their structure to provide separation depending on the diffusivity or solubility of the solute species. In the case of porous membranes (Fig. 1), internal structure could be symmetric, asymmetric, or composite. Asymmetric membranes feature a very thin active layer, responsible for the separation process, supported

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007805

Table 1 Commercially available hollow fiber membranes

Company (hollow fiber tradename)	Membrane material	ID, ED (mm)	Pore size (for porous membrane)	Applications/remarks
Air Liquide				
Medal	PI, polyaramide	0.1, 0.2		Gas separation, nitrogen generation, CO_2 removal
Air Products				
Permea (Prism)	PS			Nitrogen generation
Aquilo Gas Separation	Polyphenylene oxide	0.34, 0.54		Gas separation
	PS	0.25, 0.45		
Asahi Kasei				
Planova	RC	—	—	Hemofiltration, dialysis
APS series	PS-PVP	0.25, 0.32	12–72 nm	Hemofiltration, dialysis
PAN series	PAN-PVP	0.21, 0.27	10 nm	Hemofiltration, dialysis
Microza LGV	PVDF	0.7, 1.3	0.1 μm	MF-UF/distillation
Microza Aria	PAN	0.8, 1.4	13–80 kDa	Oil–water separator
Eutec	PP	—	—	
Berghof				
BMK series	PA, PS	ID: 0.6, 1.1, 1.5	2–100 kDa	Industrial/medical UF
Celgard LLC				
X series	PP	0.2, 0.3	0.04 μm	For oxygenation
Liqui-Cel		0.24, 0.3	0.03 μm	Membrane contactor (with characteristics)
		0.435, 0.575	nonporous	
CEPAration	Ceramic (aluminium oxide)	1–3, 2–4	0.3–40 nm, 0.1–1.4 μm	Food, pharmaceutical, biotechnological industries etc.
Daicen Membrane System				
Molsep	CA		150 kDa	
	PAN		30–150 kDa	
	PES		30–500 kDa	
Molpure	CTA	0.8	150 kDa	In-to-out filtration
Dainippon				
Sepharel/PF-D	Dense polyolefin			Degasing membrane, membrane contactor
DuPont				
BW-L	Aromatic PA			RO for brackish water 15 bar nominal pressure
Infilco Degremont				
Aquasource	CA derivative		0.01 μm	In-to-out filtration

Innovative Gas System				
Generon	Tetrabromo polycarbonate			Air separation
Fresenius	PS/PVP blend	0.1, 0.18		Hemofiltration
Kuraray	PVA			Blood purification
Hospal				
Diacepal	Cellulose diacetate			Hemodialysis
Acrylane	PES–PVP alloy			
Nephral	—			
Hydranautics				
HydraCap	PES	0.8,1.3	20 kDa	In-to-out filtration
HydraCapLD	PES	1.2–2	20 kDa	
Koch Membrane System				
Romicon	PAN	ID: From 0.5	50 kDa	Large range of applications, including membrane contractor
PMPW	PS	ID: 0.9	100 kDa	
TARGA	PS	ID: 1.5–1.1	10–100 kDa	
Membrana GmbH				
MicroPES	PES	0.3, 0.5	0.2 μm	Water, biotechnology (enzymes), food used for oxygenation
UltraPES	PES	ID: 0.7	UF range	
Accurel, Oxyphan	PP	0.28–1.8, 0.38–2.7	0.1–0.2 μm	
Memcor Ltd				
AXIA, PreMPT, CMF-S, MemJet	PP, PVDF	0.39, 0.65	0.2 μm	Submerged, wastewater
Microdyn Technologies	PP, PE, RC, Sulfonated PES	ID from 0.2	10 kDa to 0.4 μm	Membrane contactor
Millipore	PS	ID from 0.5	3 kDa to 0.1 μm	Membrane contactor
Minntech BV				
FilterFlo	PS (self-wetting Polyphen®)		0.05–0.45 μm	Medical, industrial and water
Mitsubishi Rayon				
Sterapore	Polypropylene, PE		0.4 μm	Submerged, wastewater
Nikkiso				
FLX-GW	Polyether polymer alloy			Dialysis
Parker-Finite	PI			Compressed air dryer
Pervasiv Ltd	Anion-exchange grafted in PE	0.4, 0.53		Ion exchange pervaporation (capacity 1.2 mmol/g dry membrane)

(Continued)

Table 1 Commercially available hollow fiber membranes *(Continued)*

Company (hollow fiber tradename)	Membrane material	ID, ED (mm)	Pore size (for porous membrane)	Applications/remarks
Praxair Inc	PI			Gas and hydrogen separation, nitrogen production, air dryer
Propylex	PP			Plasma separator
Schott/Bioran	Glass (>96% SiO_2)	0.3	10–90 nm	
Toyobo				
Hollosep	CTA, PA	0.1, 0.2		RO, artificial kidney
WK-Waterfilter				
ECO Series	PS	ID: 0.8	3.2–11 nm, 6–100 kDa	Food, pharmaceutical, chemical
X-Flow				
Capfil	PES/PVP blend	ID: 0.8	0.2–0.5 μm, and 150–200 kDa	NF also available Max pressure (300 kPa)
Zenon				
ZeeWeed 500	PVDF/PES		0.04 μm	Water and wastewater
ZeeWeed 1000			0.02 μm	

ID, internal diameter; ED, external diameter; CA, cellulose acetate; CTA, cellulose triacetate; PA, polyamide; PAN, polyacrylonitrile; PC, polycarbonate; PE, polyethylene; PES, polyethersulfone; PP, polypropylene; PS, polysulfone; PVDF, polyvinylidene fluoride; PVP, polyvinyl pyrrolidone; RC, regenerated cellulose.

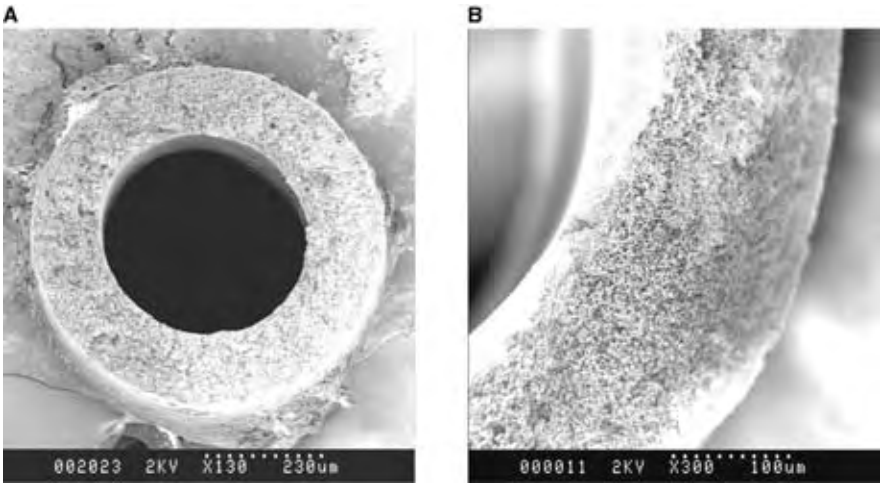


Fig. 1 Scanning electron microscopy of (A) 0.2 μm polypropylene and (B) 20 kDa polyethersulfone hollow fibers.

by a highly permeable, nonselective, and porous structure. In the case of composite membranes, those two layers are made from different materials (Fig. 2). For composite membrane, the active skin is usually a very thin (0.1–1 μm) dense layer, which could be located on either the inner or the outer fiber surface, or somewhere in between.[6] Dual layer spinneret could be used; but even if the compounds used in the active layer cannot be extruded into a hollow fiber shape, they are deposited (or chemically cross-linked) on the surface of the support fiber in a post-treatment stage. This process generally results in higher transport rates and lower hydraulic resistances, and therefore to lower operating costs, compared to those of symmetric membrane systems of similar nominal pore size. As a result, most of the academic and industrial research now concentrates on asymmetric or composite structures for hollow fiber materials. Based on a PES support matrix, dual layer asymmetric hollow fibers (Fig. 2) have been fabricated for gas separation with fluoropolyimide, matrimid, and polyimide.[1,7,8]

Thermal and chemical properties

For separation applications using porous membranes like microfiltration and ultrafiltration, the choice of polymeric material does not influence significantly the removal efficiency. However, because some membrane processes are required to operate under extreme conditions (temperature, pH, pressure), the chemical and thermal properties of the polymer must be adapted to a given application. Other important membrane properties like hydrophobicity, wettability, and adsorption are directly dependent on its material. While hydrophilic materials are used for aqueous feeds, hydrophobic polymers are required when gases and organic solvents are separated. For the treatment of aqueous feeds, naturally hydrophobic polymeric materials such as PE, PP, and PVDF are surface modified with some hydrophilic functional groups. The selection of the cleaning agent for a given hollow fiber membrane is also a delicate step in the process design. For example, polyamide (PA) membranes cannot be cleaned with chlorine-based agents, as chemical degradation would occur.

Mechanical properties

As a self-supported cylinder, hollow fiber membrane is required to withstand high transmembrane pressure without collapsing. Modulus of elasticity is a crucial parameter for the calculation of the collapse pressure of a given fiber. With a much more porous overall structure, asymmetric hollow fibers specifically require a high modulus of elasticity to avoid collapse of the

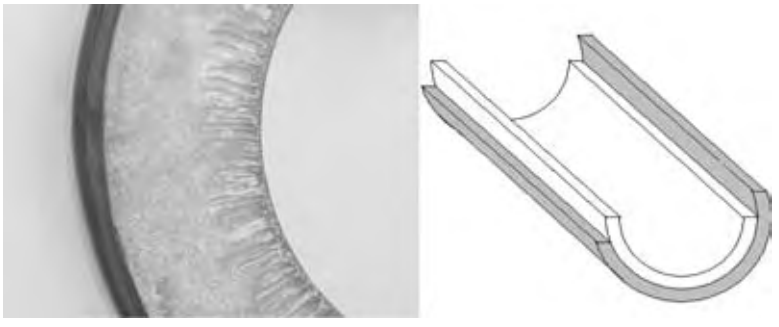


Fig. 2 Fluoropolyimide/polyethersulfone dual layer HF used for gas separation. (Courtesy of T.S. Chung, National University of Singapore.)

filament. For high pressure-driven processes like reverse osmosis, it is recommended to keep the internal diameter of the fibers at the order of magnitude of the fiber wall thickness. However, for microfiltration and uitrafiltration systems, where the transmembrane pressure is much less, the internal diameter can be significantly larger than the fiber wall. This allows for inside-out filtration (without the risk of lumen clogging) and generally reduces the hydraulic resistance offered by the membrane.

Fabrication

Hollow fibers membranes can be fabricated from most of the spinnable materials. Because of their self-supported properties, the dimensions and morphologies of hollow fiber are to be optimized and carefully assessed during the fabrication process. Three preparation methods based on spinning technology could be applied—wet (or dry–wet), melt, and dry spinning. While melt and wet spinning are generally used to obtain dense isotropic morphologies, the dry–wet method (Fig. 3) is the most widely used one as it leads to the formation of almost all membrane morphologies.

During the dry–wet spinning process (Fig. 3), a viscous, degassed, and filtered polymer solution (20–40% w/w polymer) is pumped through the outer bore of the spinneret. A nonsolvent fluid (also called bore injection fluid) is delivered through the inner part of the spinneret to maintain the annular structure and/or to control coagulation of the fiber bore. The newly formed thread spends a short time in the air before being immersed in a nonsolvent bath where coagulation occurs. The length of the air gap located between the spinneret and the coagulation bath is one of the main parameters determining the membrane dimension. Other important parameters affecting hollow fiber morphology are numerous and include dope composition and viscosity, spinning temperature, pumping rates, composition of coagulants, and the spinneret dimensions. Finally, residual liquid and solvent are removed during a washing step before the hollow fiber being collected on a spool. Depending on its desired functionality, other post-treatments could be applied to the hollow fiber such as (photo) chemical cross-linking, addition of antiplasticizers, and fluorination. For example, cellulose and polyamide-based hollow fiber often need to be kept wet (or plasticized) in order to maintain their morphology and properties. These fibers could be plasticised on-line or the hollow fiber collection spool could be directly immersed in a plasticizing medium. Final agents' addition to the storage solution could include formaldehyde (to avoid biological fouling) or magnesium ions (to increase fiber stability). Nonporous polymer hollow fibers have also been spun to provide polyester heat transfer surfaces for temperature in blood oxygenators, while others have been used for crystallization surfaces.[9]

HOLLOW FIBER MODULES

The small size of hollow fiber allows high surface area to volume ratios in hollow fiber modules ranging from 500 to 10,000 m^2/m^3 in comparison with 500 and 1000 m^2/m^3 for plate and frame and spiral wound modules, respectively. The majority of hollow fiber modules are assembled as shell and tube heat exchanger (Fig. 4). Flow can occur in the exterior (shell side) or the interior (lumen) of the fibers in a countercurrent or cocurrent manner. Shell-side flow can be introduced or collected using side ports in the external casing (Fig. 4A) or a central tube in the middle of the module to create radial flow across the fiber bundle

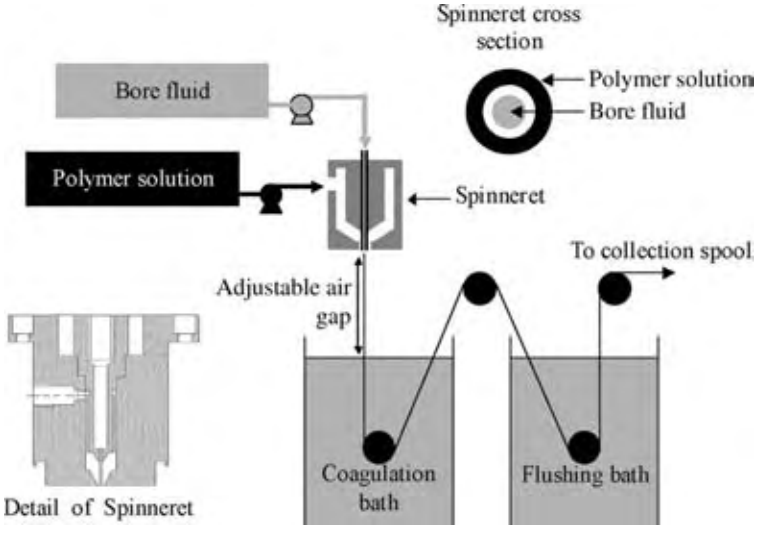


Fig. 3 Schematic of hollow fiber fabrication by dry-wet spinning process. (Spinneret details courtesy of T.S. Chung, National University of Singapore.)

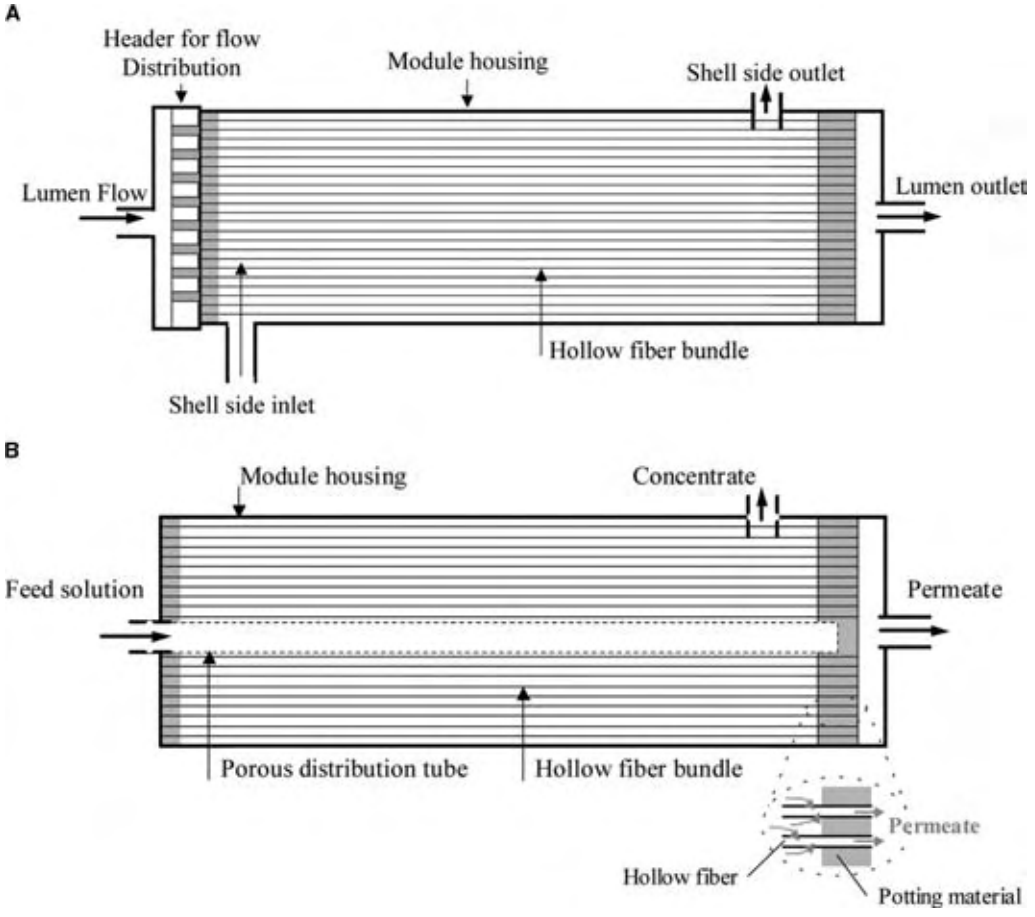


Fig. 4 Schematic of the axial view of a hollow fiber module featuring (A) shell-side ports and (B) operating from out-to-in the fiber with central feed tube. *(View this art in color at www.dekker.com.)*

(Fig. 4B). Typically for liquid–gas contactor applications, the gas flows in the lumen, while the liquid is fed in the shell to provide more mixing in the liquid side where mass transfer is more likely to be high.

Hollow fiber bundles, which can exceed 10,000 fibers, are rarely ordered arrays like heat exchangers. Most fiber modules consist of randomly packed fibers, closely packed (packing density of approximately 50%), with both ends of the fiber potted into a tubesheet with epoxy, silicone, or urethane resin (Fig. 5). The tubesheets are then cut to expose the fiber lumen. Headers on the potted ends are used to distribute or collect flow in the lumen while ports on the external casing perform the same purpose on the shell side. Metal casings are required for high pressure operations such as reverse osmosis and gas separation. The external casing is not required where the transmembrane pressure or the suction is applied solely via the lumen side. In these submerged membrane systems, fibers can be directly immersed in liquid to carry out filtration (Figs. 6 and 7) or serve as contactors. These modules can be, in the forms of bundles or sheets, potted into collection tubes. Fibers may also be potted only in one end and allowed to fluidize freely to reduce concentration polarization and improve mass transfer.[10] Unlike spiral wound modules that often exist in 4 and 8 in. diameter modules, hollow fiber modules are not as standardized and can vary significantly in length and diameter even for the same applications.

Flow maldistribution through these fiber bundles leading to low mass transfer and stagnant regions for fouling is a significant concern, particularly when liquids and shell-side feed are used. Woven or helically wound fiber packing has been used to overcome channeling of fluids between axial bundles of fibers and reduce concentration polarization in reverse osmosis and gas–liquid contactors and enhance heat transfer in oxygenators. Baffling and strategic location of the collection tube have been used in improving mass transfer for liquid contactors for the Liqui-Cel™ from Celgard shown in Fig. 8. Other strategies for generating transverse flow or turbulence in the module include crimping fibers to disrupt the axial flow, gas injection on shell or lumen side, and small amplitude oscillations.[11]

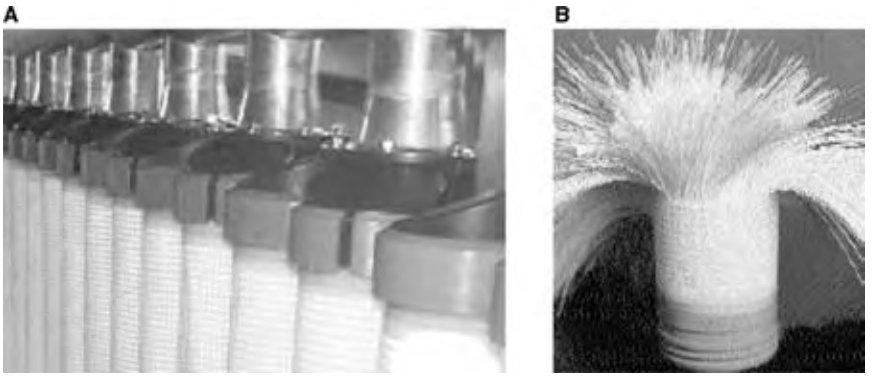


Fig. 5 Conventional (A) hollow fiber module assembly and (B) cutaway. (Courtesy of Memcor.) *(View this art in color at www.dekker.com.)*

Careful design of the headers and shell-side ports must be considered as they can be a source of significant flow maldistribution.[12]

To avoid major fouling and clogging problems, the nature of the feed to be treated has to be considered when hollow fiber modules are used. In the case of lumen to shell filtration, the inside diameter of the fiber is supposed to be at least 10 times the diameter of the largest species present in the feed.[13] However, when the permeate flows from shell to lumen, concentration and viscosity of the feed and the density of hollow fiber membrane per module may be critical parameters in the design process. Specific aeration or mixing requirements are necessary to keep the feed particles in suspension, and to avoid the clogging of the membrane module.

Where porous hollow fibers are used as contactors between phases, the pressure differential along the module needs to be lower than the critical breakthrough pressure determined by the surface tensions between the phases and the membrane pore size. High pressure differential between the entrance and exit of hollow fiber modules should be limited for filtration as it can promote excessive flux and fouling on one end of the module. Recirculation of permeate flow back into the feed size due to Starling flow can even occur during filtration if the pressure drop on the feed side drops below the permeate side as it flows through the module.[14] As lumen sizes decrease and hollow fiber lengthens, the pressure drop on both the lumen and the shell-side can become excessive.

A wide range of mass transfer correlations based on Reynold and Graetz numbers have been used to characterize the performance of hollow fiber module contactors. The variation of mass transfer correlations has been attributed to the nonideality in flow distribution, deviation from simple axial flow, and fiber inhomogeneity.[15,16] Modeling for concentration polarization build-up and two phase flow on the lumen side has also been developed for hollow fiber modules used in filtration.[17] Flow distribution in modules has also been characterized using residence time distribution

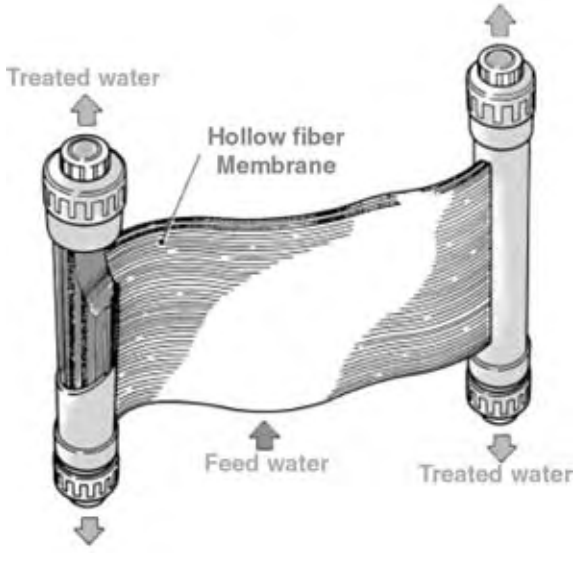


Fig. 6 Submerged, horizontal hollow fiber module. (Courtesy of Sterapore.) *(View this art in color at www.dekker.com.)*

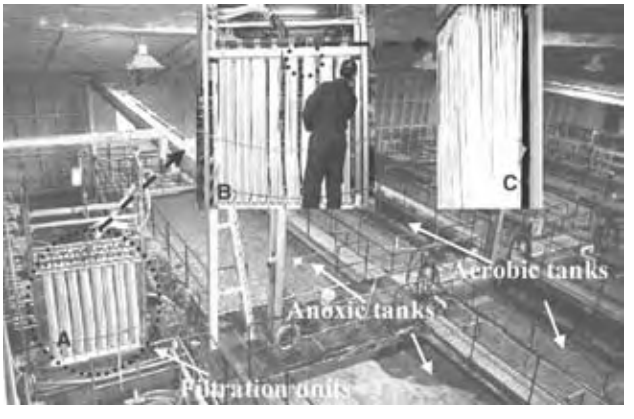


Fig. 7 (A) Overview of the module; (B) close-up of the module; and (C) close-up of the hollow fiber membrane used in a membrane bioreactor for wastewater treatment. (Courtesy of CH2M Hill.)

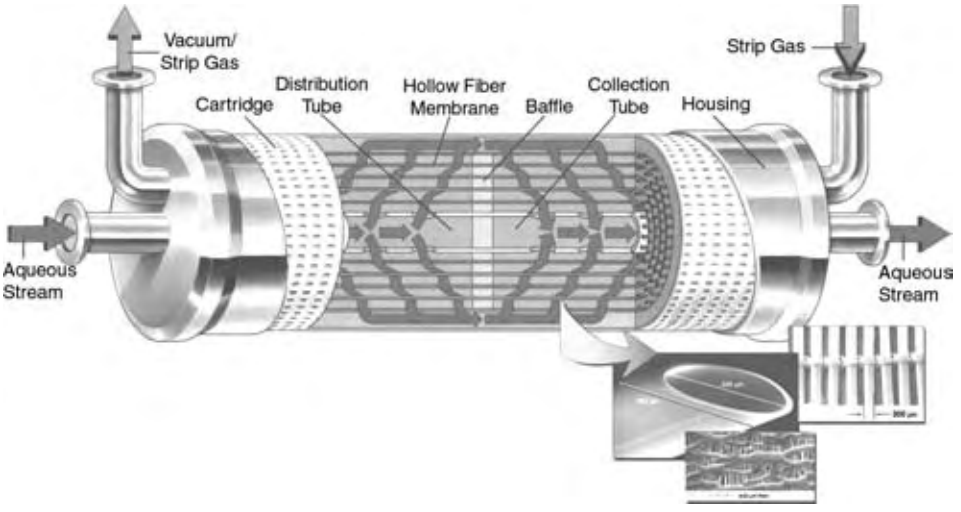


Fig. 8 Schematic of Liqui-Cel™ membrane contactor. (Courtesy of Celgard.) *(View this art in color at www.dekker.com.)*

measurements, nuclear magnetic resonance imaging, and x-ray computed tomography.[18] The hydrodynamics of shell-side flow and two phase flow around submerged hollow fiber systems however remains a challenge to fully characterize and model. Assembly of the modules into arrays allows easy scale-up and compact footprint of many hollow fiber processes. The modular nature of these arrays allows easy replacement and localized integrity testing when fiber breaks (Fig. 5).

HOLLOW FIBER APPLICATIONS

Filtration

With a nominal pore size ranging from 0.01 to few microns, microfiltration is generally used for the removal of colloidal, fine particles and bacteria, while ultrafiltration membranes and their molecular weight cutoff from 200 to 500,000 Da retain polymers, sugars, and viruses. Because of the relative ease of producing microporous, polymeric hollow fibers, the use of this membrane configuration for ultrafiltration and microfiltration treatments also varies widely. Only specific process requirements, like high chemical and thermal resistance of polymeric filters, have limited the application of hollow fiber microporous membranes; the development of ceramic or metal membranes (in a flat sheet or a tubular configuration) offers means to overcome these limitations. Hollow fiber microporous membranes could be used for concentration, purification, and fractionation in the following main industries:

- Food
- Medical, pharmaceutical, and biotechnological
- Water treatment and reuse
- Industrial wastewater treatment
- Air filtration

Clarification of rough beer, vinegar and pasteurization of clarified beer by cross-flow ultrafiltration are also very common processes utilizing hollow fiber ultrafiltration. As seen in Table 1, an important number of membrane manufacturers specialize in medical and pharmaceutical applications. In pharmaceutical and biotechnology industries, hollow fiber membranes are used for the concentration, separation, and purification of physiological activators such as antibiotics, vaccines, enzymes, proteins and peptides, as well as blood purification (hemofiltration). As a physical barrier for bacteria and viruses, membranes are also a popular option for the production of purified water for hospitals and pharmacies.

A growing application of hollow fiber technology is in the areas of water and wastewater treatment. Recent issues related to water shortage in many parts of the globe had forced the development of the number of potable water treatment .plants. Surface water sources can be very efficiently treated by hollow fiber membrane (sometimes coupled with coagulation) process. Based on the Asahi Kasei's Microza membrane, Pall Corporate commissioned 45 facilities, while US Filter/Memcor are responsible for 36 on-line plants and 11 under construction. Zenon Environment is probably the leader in the market with 68 plants on-line and 19 under construction. The biggest of these plants is located in Mery sur Oise (France) and reaches 44.4 mega-gallon per day (Pall Corporation). In many of these plants, hollow fiber microporous technology is used as a pretreatment, generally before nanofiltration or reverse osmosis membrane process. Because of their relatively small footprint and high reliability, these dual membrane systems are becoming more and more popular. The cost of treating sewage to indirect potable water reuse standards (i.e., by ultxafiltration followed by reverse osmosis processes) is only 39% of the cost of seawater desalination.[19] Hollow fiber membrane

technology is also increasingly used for industrial wastewater recovery and reuse.[20] Only limited usage of hollow fibers has occurred in reverse osmosis modules. The most common hollow fiber reverse osmosis membranes that attained the stage of economic application in desalination plants are generally made of polyamide (DuPont and Toyoto). Application of hollow fiber for nanofiltration has been explored by Akzo Nobel and others but are not yet commercially available.[21] By using 0.1 μm (or lower) pore size hollow fibers, compressed gases can also be sterilized as bacteria and viruses are removed.

Gas Separation and Pervaporation

Since its inception in 1979, gas separation using membranes has developed into an industry worth over $150 million worldwide, with the bulk being in hollow fiber modules.[22,23] The current major applications include air separation into nitrogen and oxygen, hydrogen separation from ammonia purge gas streams, and natural gas treatment.[24] The first two are relatively clean feed streams and thus are relatively straightforward processes, limited only by the maximum purity that can be produced commercially (approximately 99.9–99.99% nitrogen for air separation).[6] However natural gas treatment (CO_2, H_2S, and H_2O removal) and other potential petrochemical applications may have substantial foulants, including particulates, plasticizing, and condensable vapors. As transport of gas solutes through dense membrane requires high pressures (up to 70+ bars) and ultrathin selective layers, hollow fibers must be defect free, and the fibers and the potting material sufficiently robust to withstand solvent vapors commonly present in such feeds. The basis of the separation mechanisms is the ''solution–diffusion mechanisms'' of the solute through the free volume of the membrane material, typically polymeric thin films on a porous support layer. While the tradeoff between selectivity and permeability tradeoff with a common separation polymer is well known, current research has focused on a combination of short flexible molecules between bulky, rigid segments to enhance selectivity.[24,25] Other developments include using facilitated transport membranes where the complexing ligand is used to shuttle desired species across the membrane. Dense hollow fibers can also provide contact area for gas stripping or oxygenation using oxygen selective polymer such as polymethylpentene in low pressure applications, for instance blood oxygenators.

With respect to gas separation hollow fiber modules, the chemical compatibility and the mechanical stability of the fibers and potting are paramount. High pressure modules use shell-side feed while low pressure modules use lumen side feed when fouling is not an issue.[24] While flow maldistribution may not be as severe as liquid flow in hollow fiber modules, modeling suggests that at high purities or recoveries, it may become a significant limiting factor.[6]

In contrast to gas separation, pervaporation involves the solubilization and the diffusion of species from liquid feed and transport through membrane into a vapor phase, removed by vacuum or sweep gas. Instead of high pressure, temperature and the chemical potential of the species drive the separation, mainly dehydration of organic liquids and removal of organic species from water. Despite the larger surface area provided by hollow fiber modules, they are not as prevalent as plate and frame and spiral wound modules as the small channels contribute to resistance to permeate flow. This results in vapor pressure buildup in the permeate-side (typically the fiber lumen as the separation layer is often on the shell side), reducing driving force. Thus optimization of fiber length and fiber productivity for pervaporation are discussed in the literature.[26] Porous hollow fibers can be used for air purification, but only few companies (such as Parker-Finite and Praxair Inc.) have proposed using hollow fiber membranes permeable to water vapor to dry compressed air.

Hollow Fiber Contactors

Hollow fiber contactors use membranes to separate two phases and transport is due to diffusion, chemical reaction, or chemical potential rather than pressure. The main examples of hollow fiber contactors are found in dialysis, gas adsorption/deadsorption, and solvent extraction. Use of hydrophilic and hydrophobic fiber materials controls the wetting of the pores. Typically, the phase that has higher mass transfer is allowed to wet the pores in order to minimize overall mass transfer resistance.

In dialysis, size exclusion is the main separation mechanism, while osmotic pressure and concentration difference drive the transport across two typically aqueous phases. While dialysis is used in some analytical separations,[27] dialysis for the removal of toxins from blood (hemodialysis) is the most prominent application for hollow fiber technology in the biomedical field. The hemodialyzers are used to treat over one million people a year and have become a mass produced, disposable medical commodity.[28] While the first hemodialyzers were developed from cellulosic material (Cuprophane, RC, etc.), synthetic polymers such as polyacrylonitrile, poly(ether) sulfone, and polyvinyl pyrrolidone are increasingly used to improve blood compatibility and flux. Hemodialyzer modules consist of thousands of extremely fine hollow fibers

(approximately 250 μ with wall thicknesses of 5–10 μ) in a short casing (approximately 30 cm) to minimize fluid holdup. Dialysate flows in the shell side while blood flows in the lumen. As mass transfer can be severely compromised by flow maldistribution in the liquid phase, modules such as those by Membrana GmbH have incorporated wavelike undulations in the fiber and spacer fibers on the shell size to improve fluid distribution. Control of flux through the hollow fiber is also crucial as the pressure drop along hollow fiber modules can result in back-transport from the dialysate of endotoxins removed from the blood. Current research in hemodialysis is focused on better blood compatibility, low flux membranes, and adsorptive membranes that bind endotoxins.[28]

Hollow fibers can separate not just two liquid aqueous phases as in dialysis but also organic and aqueous phases. Liquid–liquid separation using porous hollow fibers to control the interfacial area has been proven successful for using organic solvent to strip organic solutes from the aqueous phase or aqueous phase to strip water soluble solutes from organic phase. So far, only small liquid–liquid extraction units have been used industrially.[29] Significant problems exist in controlling pressure fluctuations within the module and fouling of membrane surfaces, which cause fluids to breakthrough the pores of the fiber, resulting in emulsification or contamination of the feed or the stripping streams. The critical breakthrough pressure is governed by the Laplace equation relating minimum pressure to force liquid into a pore to surface tension of the liquid, surface energy of the material, and pores size. With hollow fibers, the pressure drop in lumen or shell can be significant thus making pressure control crucial to avoid undesired wetting and breakthrough of the fibers.

The application of hollow fibers for gas stripping and gas transfer into liquids has been commercialized successfully. As with liquid–liquid contactors, undesired fluid breakthrough can be reduced with careful pressure control to the shell and lumen side of the fiber throughout the module. Typically, the hydrophobic fibers are used to reduce mass transfer resistance from the stagnant liquid in the pores; thus the deposition of hydrophilic foulants on the membrane surface can provide an undesired conduit for liquid penetration into the hollow fiber. The use of nonporous, gas selective hollow fibers avoids some of these wetting problems.

Aside from blood oxygenators, the major applications include degassing (O_2 and CO_2 removal) for ultrapure water production (Membrana GmbH, W.L. Gore & Associates), bubble-less aeration (Membrana GmbH) and carbonation, and CO_2 removal from natural gas and exhaust gases (Kvaerner, W.L. Gore & Associates). The Kvaerner process involves using porous hollow fibers (polytetrafluoroethylene) to contact gases with an amine stripping solution to achieve CO_2 removal. Originally developed for off-shore natural gas processing, there are very few examples of large scale gas processing using hollow fiber membranes. However, as CO_2 removal has gained new importance, there are new efforts to expand hollow fiber contactors to flue gas and other large scale applications using cheaper polypropylene and alternative absorbants to the alkanolamines used in the original Kvaerner processes.

Other potential uses for hollow fiber contactors include membrane distillation where microporous membranes form a barrier to liquid water but allow vapor to pass. Studies of these modules have been so far limited to laboratory scale investigations. Extensive reviews of the mass transfer correlations and applications for hollow fiber contactors are given in the literature.[15]

Hybrid Processes

In hybrid processes, membranes are used to remove end-product from the reactor, allowing a chemical or a biochemical reaction to carry on. Hybrid membrane

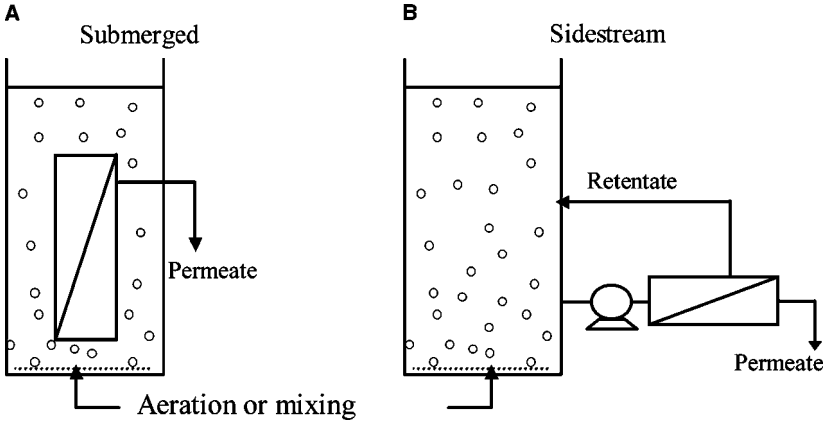


Fig. 9 Two operating configurations for hollow fiber membranes: (A) submerged (or immersed) and (B) sidestream (or cross-flow) *(View this art in color at www.dekker.com.)*

processes could be configured into two different designs. In a submerged configuration, the membrane is directly submerged in the reactor, while a sidestream membrane reactor features an external loop where the membrane separation unit is located outside the reactor (Fig. 9). Originally, most of research based on membrane reactors was carried out on gas phase reactions (catalysis-based dehydrogenation, hydrogenation, and oxidation), while studies on coagulant and adsorbent hybrid hollow fiber membrane processes for treatment of water and wastewater have been more recently reported.

The most remarkable development in membrane process is probably due to the exponentially increasing number and size of membrane bioreactor (MBR) treatment plants for municipal and industrial wastewaters.[30] Advantages of MBR over conventional process comprise higher process intensity and effluent quality, and reduction of footprint and sludge volumes. The process could be designed both in submerged or sidestream configurations (Fig. 9) with flat sheet or hollow fiber membranes. Although the relative merits of those designs are subject to discussion, a majority of contractors and users tend to choose the submerged hollow fiber technology, probably because of the backwashing option offered by the hollow fiber membrane.

Currently, three submerged hollow fiber MBR designs are commercially available. Mitsubishi Rayon, Zenon, and Memcor offer hollow fiber membrane specifically designed for MBR applications. Zenon, the market leader in terms of global capacity, has increased its MBR capacity from less than $1000\,m^3$/day in 1993 to almost $1.5\,Mm^3$/day currently. However, MBR could still be considered too costly, and issues related to membrane fouling and clogging are the main reasons limiting an even more widespread large-scale application. Membrane suppliers generally guarantee the integrity of the hollow fiber membrane for 3 to 5 years when they are used under normal conditions. A comparison between MBR and membrane used for tertiary filtration of effluents from conventional activated sludge process (both based on hollow fiber technology) showed that integrated MBR was the least expensive.[19] Proper hollow fiber module design is crucial in MBR where a high concentration (around 20 g/L) of active biomass is filtered. Aeration devices are generally located below the HF module to ensure good circulation of the biomass, promote membrane movement, both phenomena limiting clogging and fouling. In that same effort, recent optimization of the aeration of hollow fiber MBR has been recently developed by experimenting intermittent aeration and coarse bubbling. Fig. 7 shows one such wastewater treatment plant based on MBR—submerged hollow fiber technology in the Westview wastewater treatment plant in Canada.

CONCLUSIONS

Hollow fiber technology has allowed high surface area for mass and heat transfer to be generated relatively cheaply. However, the module design and the operational mode are crucial as the majority of capital or operating costs in chemical processing is often in achieving the pressure, flowrates, or turbulence in hollow fiber modules rather than the cost of the fibers itself. The advent of submerged hollow fiber systems has made a significant impact in the water and wastewater treatment areas, and new innovative module designs and operation modes are quickly developing. For gas separation and other dense membrane applications, sophisticated material science for developing more selective and robust fibers and potting materials will dictate the further penetration of hollow fiber modules into large scale petrochemical separations.

REFERENCES

1. Liu, Y.; Chung, T.S.; Wang, R.; Liu, D.F.; Chng, M.L. Chemical cross-linking modification of polyimide/poly(ethersulfone) dual layer hollow fiber membranes for gas separation. Industrial & Engineering Chemistry Research **2003**, *42* (6), 1190–1195.
2. Ismail, A.F.; David, L.I.B. A review on the latest development of carbon membranes for gas separation. Journal of Membrane Science **2001**, *193*, 1–18.
3. Ho, W.S.; Sirkar, K.K. *Membrane Handbook*; Van Nostrand Reinhold: New York, U.S.A., 1992.
4. Mulder, M. *Basic Principles of Membrane Technology*, 2nd Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1996.
5. Castro, B.N.; Gerla, P.E. Hollow fiber and spiral cheese whey ultrafiltration: Minimizing controlling resistances. Journal of Food Engineering **2005**, *69*, 495–502.
6. Lipscomb, G.G.; Sonalkar, S. Sources of non-ideal flow distribution and their effect on the performance of hollow fiber gas separation modules. Separation and Purification Technology **2004**, *33* (1), 1–36.
7. Li, D.F.; Chung, T.S.; Wang, R.; Liu, Y. Fabrication of fluoropolyimide polyethersulfone (PES) dual layer asymmetric hollow fiber membranes for gas separation. Journal of Membrane Science **2002**, *198*, 211–223.
8. Jiang, L.Y.; Chung, T.S.; Li, D.F.; Cao, C.; Kulprathipanja, A. Fabrication of matrimid/polyethersulfone dual layer hollow fiber

membranes for gas separation. Journal of Membrane Science **2004**, *240* (1–2), 91–103.
9. Zarkadas, D.M.; Sirkar, K.K. A new hollow fiber cooling crystallization technique. AICHE Annual Conference, Austin, Texas; American Institute of Chemical Engineers, 2004.
10. Ahmed, T.; Semmens, M.J. Use of sealed end hollow fibre for bubbleless membrane aeration: experimental studies. Journal of Membrane Science **1992**, *69*, 1–10.
11. Cui, Z.F.; Chang, S.; Fane, A.G. The use of gas bubbling to enhance membrane processes. Journal of Membrane Science **2003**, *221* (1–2), 1–35.
12. Bao, L.; Lipscomb, G.G. Mass transfer in axial flows through randomly packed fiber bundles. In *New Insights into Membrane Science and Technology: Polymeric and Biofunctional Membranes*; Bhattacharyya, D., Butterfield, D.A., Eds.; Elsevier, 2003; 5–26.
13. Porter, M.C. *Handbook of Industrial Membrane Technology*; Noyes Publications: Park Ridge, U.S.A., 1990.
14. Hammer, B.E.; Heath, C.A.; Belfort, G. Quantitative flow measurements in bioreactors by nuclear-magnetic resonance imaging. Bio-Technology, **1990**, *8* (4), 327–330.
15. Gabelman, A.; Hwang, S.-T. Hollow fiber membrane contactors. Journal of Membrane Science **1999**, *159*, 61–106.
16. Lipniski, F.; Field, R.W. Mass transfer performance for hollow fibre modules with shell-side axial feed flow: using an engineering approach to develop a framework. Journal of Membrane Science **2001**, *193* (2), 195–208.
17. Aimar, P.; Howell, J.A.; Clifton, M.J.; Sanchez, V. Concentration polarization buildup in hollow fibers — a method of measurement and its modeling in ultrafiltration. Journal of Membrane Science **1991**, *59* (1), 81–99.
18. Chen, V.; Li, H.; Fane, A.G. Non-invasive observation of synthetic membrane processes — a review of methods. Journal of Membrane Science **2004**, *241*, 23–44.
19. Cote, P.; Masini, M.; Mourato, D. Comparison of membrane options for water reuse and reclamation. Desalination **2004**, *167*, 1–11.
20. Judd, S.; Jefferson, B. Industrial waters. In *Membranes for Industrial Wastewater and Recovery and Re-use*; Elsevier Advance Technology: Oxford, 2003.
21. Sethi, S.; Wiesner, M.R. Simulated cost comparisons of hollow-fiber and integrated nanofiltration configurations. Water Research **2004**, *34* (9), 2589–2597.
22. Lonsdale, H.K. The growth of membrane technology. Journal of Membrane Science **1982**, *10* (2–3), 81–181.
23. Koros, W.J.; Fleming, G.K. Membrane-based gas separation. Journal of Membrane Science **1993**, *83*, 1–80.
24. Baker, R. Future directions of membrane gas separation. Industrial & Engineering Chemistry Research **2002**, *41*, 1393
25. Maier, G. Gas separation with polymer membranes. Angewandte Chemie **1998**, *37*, 2961–2974.
26. Feng, X.; Huang, R.Y.M. Liquid separation by membrane pervaporation. Industrial & Engineering Chemistry Research, **1997**, *36*, 1048–1066.
27. Moskvin, L.N.; Niskitina, T.G. Membrane methods of substance separation in analytical chemistry. Journal of Analytical Chemistry **2004**, *59* (1), 2–16.
28. Vienken, J.; Bowry, S. Quo vadis dialysis membrane. Artificial organs **2002**, *26* (2), 152–159.
29. Drioli, E.; Romano, M. Progress and new perspectives on integrated membrane operations for sustainable industrial growth. Industrial & Engineering Chemistry Research **2001**, *40*, 1277–1300.
30. Stephenson, T.; Judd, S.; Jefferson, B.; Brindle, K. *Membrane Bioreactors for Wastewater Treatment*; IWA Publishing: London, 2000.

Hybrid Materials (Organic–Inorganic)

C. Sanchez
Chimie de la Matière Condensée, Université Pierre et Marie Curie, Jussieu, Paris, France

G. J. A. A. Soler-Illia
Unidad de Actividad Química, CNEA, Centro Atómico Constituyentes, San Martín, Buenos Aires, Argentina

INTRODUCTION

Hybrid inorganic–organic materials can be broadly defined as synthetic materials with organic and inorganic components, intimately mixed, either homogenous systems derived from monomers and miscible organic and inorganic components or heterogenous and phase-separated systems where at least one of the components' domains has a dimension ranging from some angstroms to several nanometers. Hybrid nanocomposites had an explosive development since the 1980s, with the expansion of soft inorganic chemistry processes.[1–4] The mild synthetic conditions provided by the sol–gel process such as metallo-organic precursors, low processing temperatures and the versatility of the colloidal state allow for the mixing of the organic and inorganic components at the nanometer scale in virtually any ratio. These features, and the advancement of organometallic chemistry, and polymer and sol–gel processing, permit a high degree of control over both composition and structure (including nanostructure) of these materials, which present tunable structure–property relationships. This, in turn, permits tailoring and fine tuning of properties (mechanical, optical, electronic, thermal, chemical) in very broad ranges, and designing of specific systems for applications. Hybrid materials can be processed as gels, monoliths, thin films, fibers, particles, or powders. The seemingly unlimited variety, unique structure–property control, and the compositional and shaping flexibility give these materials a high potential for applications.

GENERAL REFERENCES AND APPLICATIONS

The subject of hybrid materials has been extensively reviewed in different fields, such as synthesis, properties and applications, textured hybrids, and biohybrids.[5–11] Therefore, in this entry we will mainly refer to those comprehensive reviews, to encourage the interested reader to explore into them for more specific sources.

Some hybrid materials have already entered the market. Commercial examples include materials from electronics to automotive coatings with varied mechanical and optical properties, adhesives, and composites, to cite a few.[12–14] Recent examples include the indigo dyes embedded in a silica/zirconia matrix (Toshiba TV screens), organically doped sol–gel glassware (Spiegelau), and sol–gel entrapped enzymes (Fluka).

CLASSIFICATION

A distinct characteristic of hybrid materials is that not only are their properties related to the chemical nature of the inorganic and organic components, but they also rely heavily on their synergy. Therefore, the interface between inorganic and organic domains is of paramount importance. Indeed, one key point is the control of this interface, the hybrid interface.

Hybrid materials can thus be broadly classified into two main classes, depending on the nature of the links and interactions existing at the hybrid interface (Fig. 1):[15] Class I hybrids include all systems where there are no covalent or iono-covalent bonds between the organic and inorganic components. Thus, only van der Waals, hydrogen bonding, or electrostatic forces are present. On the contrary, in Class II hybrids, at least parts of the inorganic and organic components are linked through strong covalent or iono-covalent bonds. Hybrids can also be characterized by the type and size of the organic or the inorganic precursors. Precursors can be two separate monomers or polymers, or they can be covalently linked. Generally, phase separation between the organic and the inorganic components will occur, owing to mutual insolubility. However, it is possible to obtain homogenous or single-phased hybrids by choosing bifunctional monomers presenting organic and inorganic components, or by combining both types of components in phases where one of them is in large excess.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120018068

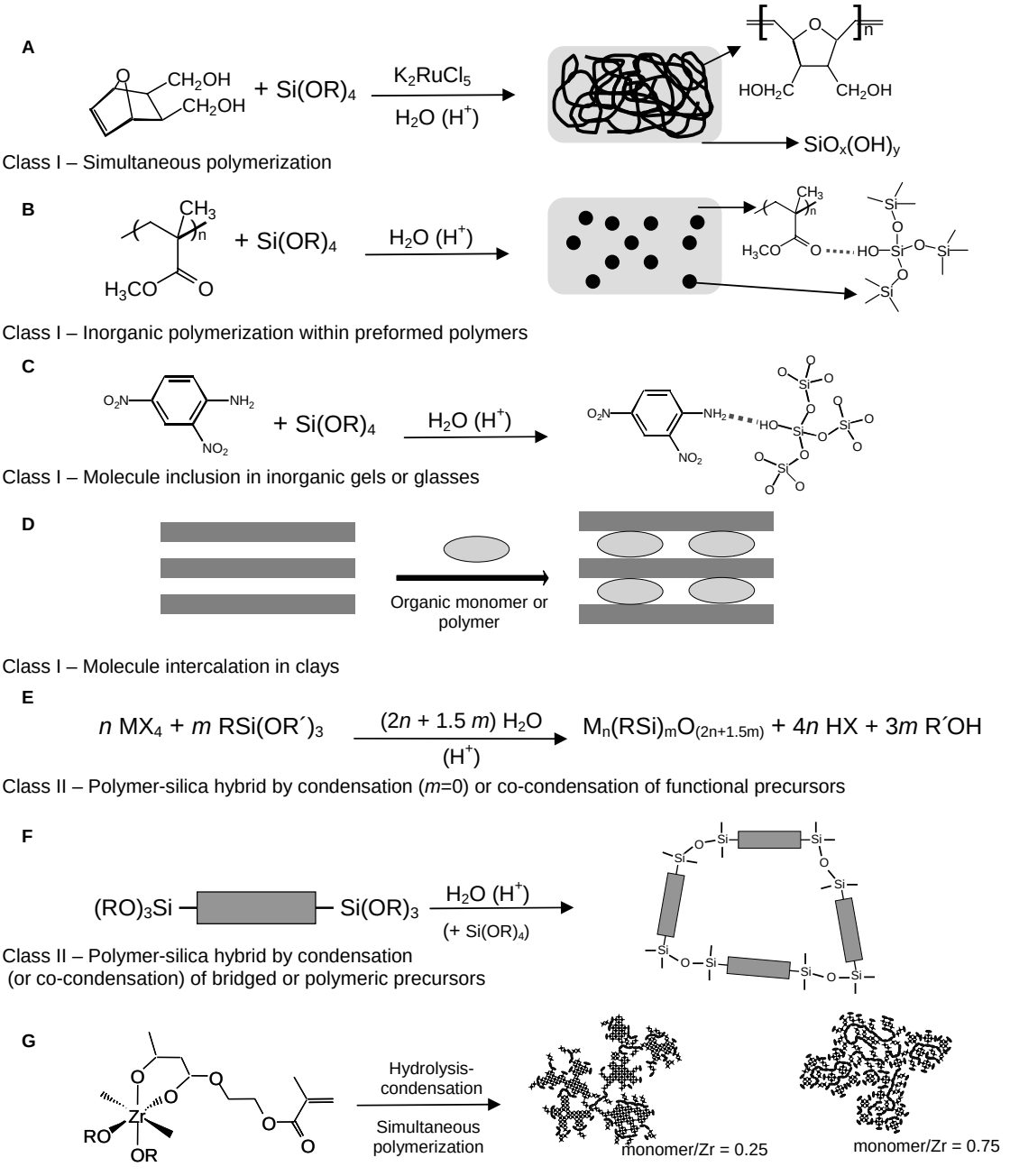


Fig. 1 Examples of hybrid organic–inorganic materials categorized by the nature of the hybrid interface (Class I or II), the type and the size of both precursors. (From Refs.[12,13,15].) *(View this art in color at www.dekker.com.)*

Brief Sol–Gel Concepts

Sol–gel processes, based on the controlled polymerization of inorganic molecular precursors in mild temperature conditions, organic solvents, and controlled water quantities are central in the development of hybrid materials. Typically, MX_n precursors are used, where M represents a metal center, n is its oxidation state, and X is a group such as alkoxy, RO–, which represents a deprotonated alcohol. X can also represent a halogen anion, as in metal chlorides. The sol–gel process implies connecting the metal centers with oxo- or hydroxo- bridges, therefore generating metal-oxo or metal-hydroxopolymers in solution. Hydrolysis and condensation reactions are activation and propagation steps. Hydrolysis of an alkoxy group

attached to a metal center leads to hydroxyl-metal species (Reaction 1).

$$M - OR + H_2O \longrightarrow M - OH + ROH \quad (1)$$

The hydroxylated metal species can react with other metal centers leading to condensation reactions, where an oligomer is formed by bridging two metal centers. In the case of oxolation, condensation leads to an oxo bridge, and water or alcohol is eliminated (Reaction 2).

$$M - OH + XO - M \longrightarrow M - O - M + X - OH$$
$$X = H \text{ or } R \quad (2)$$

In the case of olation, an addition reaction takes place, and a hydroxo bridge is formed. This reaction takes place when metal centers may have coordination higher than their valence, as in the case of Ti(IV), Zr(IV), or related cations (Reaction 3).

$$M - OH + M - OX \longrightarrow M - (OH) - M - OX$$
$$X = H \text{ or } R \quad (3)$$

In most sol–gel processing, the inorganic framework is built by successive hydrolysis and condensation reactions.[1] The structure, connectivity, and morphology of the final inorganic network depend strongly on the relative contribution of reactions 1–3.[16,17] Depending on intrinsic (i.e., metal center features such as coordination, acidity, lability, etc.) or extrinsic (i.e., tunable reaction conditions such as solvent, water contents, pH, catalyzers, reaction time, etc.) conditions, the hydroxo-oxo-polymers thus formed can exhibit a variety of structures, from branched arrangements to compact clusters. The growth of these structures can be arrested (for example, by poisoning or by limiting aggregations), and thus, sols made of suspended colloidal size entities are obtained. Some of these structures may grow extensively or aggregate until they reach macroscopic dimensions, trapping solvent and smaller monomers; a gel is thus obtained. Precipitates can also be obtained when extended growth of oxo-polymers leads to their insolubility. Drying of gels in ordinary or supercritical conditions leads to dried gels (xerogels) or high-surface aerogels, respectively. Solvent removal is a crucial step to control surface area, porosity, integrity, morphology, and eventual segregation.[16]

[1]While most sol–gel chemistry depends on the use of controlled water quantities to yield metal-oxo polymers, there is an alternative method to generate metal-oxo polymers in nonhydrolytic conditions, see, eg., Bourget, L.; Corriu, R.J.P.; Leclercq, D.; Mutin, P.H.; Vioux, A. J. Non-Cryst. Solids **1998**, *242*, 81 (and references therein).

Chemical control of activation/polymerization reactions [Eqs. (1)–(3)] permits tailoring of the size and shape of the inorganic polymers or colloids, as well as their and miscibility (i.e., the interactions, for example, hydrophobic/hydrophilic) with the organic counterparts. Functional precursors can also be used to cocondense with MX_4 precursors, or to modify the surface of inorganic entities, therefore enhancing the compatibility with organic monomers or polymers. The following sections will show general synthesis routes and some examples.

General Routes Toward Hybrid Materials

Fig. 2 shows the general chemical pathways used to obtain hybrid materials, from mutual dispersions of polymer domains attainable by the relatively simple combination of sol–gel techniques and polymerization (Route A), to the assembly of well-defined nanobuilding blocks (NBB), by using separators, or polymerization of functional NBB (Route B), to organized pore systems, obtained by the combination of sol–gel and self-assembly techniques (Route C). Materials with hierarchical structures and complex forms can also be obtained by ''integrative pathways,'' i.e., by combining controlled phase segregation, selective interactions, and morphosynthesis to the sol–gel/self-assembly tandem (Route D). Novel phases, composites, or dispersions bearing original properties (chemical storage, sensors, controlled delivery, etc.) can be created, which will have a deep impact on the technology of the 21st century. The chemical construction and patterning of materials with long-range order architectures (beyond nanometer size) remains an important challenge in the new field of ''organized matter chemistry.''[18]

Conventional Routes

Class I hybrid materials

The main strategies toward Class I hybrid materials encompass: 1) embedding small organic molecules in inorganic gels; 2) embedding preformed organic polymers in inorganic gels; 3) impregnation of previously formed porous inorganic matrices (glasses) by organic molecules; or 4) simultaneous formation of both polymer networks (organic and inorganic).

Organic molecules can be readily mixed with $M(OR)_4$ alkoxide precursors. Upon precursor hydrolysis, dye molecules, such as coumarines, rhodamines, pyranines, or others with nonlinear optical properties have been entrapped into silica, aluminosilica, or transition metal oxide matrices.[19] Alkoxides can also be mixed with polymers dissolved in alcohols or

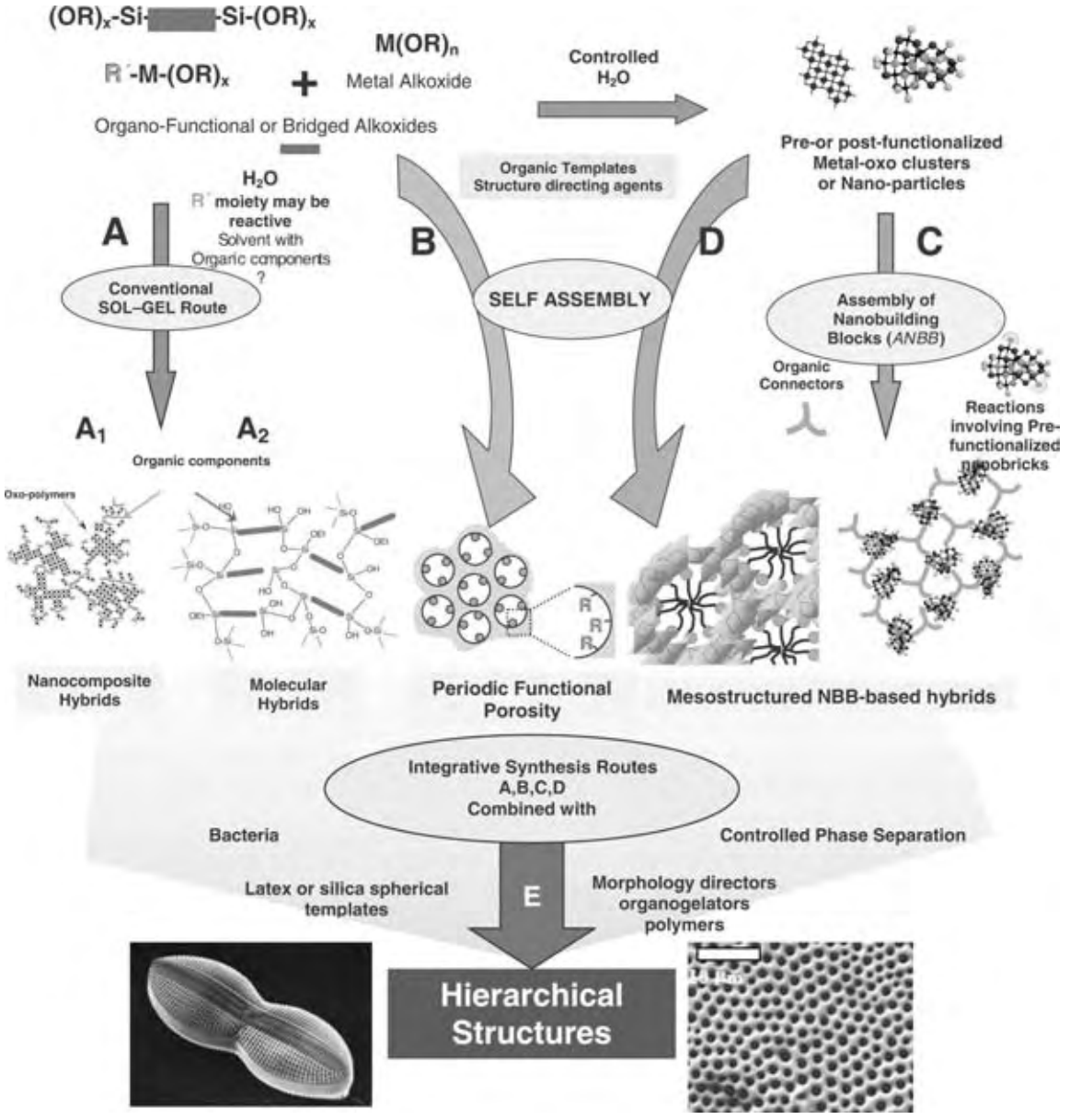


Fig. 2 Different paths to obtain hybrid materials from molecular sources. Path A: Sol–gel routes (A1: conventional route for hybrid nanocomposites, A2: molecularly homogenous hybrids). Path B: Assembly of nanobuilding blocks (ANBB), of prefunctionalized or postfunctionalized clusters or nanoparticles. Route C or D involve the use of templates capable of self-assembly, giving rise to organized phases. Path E involves integrative synthesis combining precedent paths from A to D and other processes, such as the use of lithography, casting, organogels or latex beads as templates, controlled phase separations, or external fields. (From Ref.[8].) *(View this art in color at www.dekker.com.)*

tetrahydrofuran (THF). This route results in transparent materials, where the inorganic particles are smaller and more homogenously dispersed than in a conventional blending process. Chemical control of the interactions between the inorganic (i.e., the –OH groups attached to the metal centers) and organic polymers at the hybrid interface permits avoiding or adjusting the phase separation.

A simultaneous polymerization process was developed by Novak, to generate hybrid materials in a single operation, to facilitate interpenetration of both domains (Fig. 1A).[20] Alkoxysilanes carrying polymerizable alkoxo groups (derived from methacrylate, methylmethacrylate, or norbornane) were used as precursors, to enhance the integration, and to minimize the shrinkage observed in pure silica systems. The control of polymerization kinetics of both inorganic and organic components results in large monolithic pieces with a wide polymer/silica composition domain.

Class II hybrid materials

Precursors. In these materials, strong links between the inorganic and organic domains are present. Heterofunctional inorganic precursors (RMX_{n-1}, where R is an organic group), where R is attached to the metal center, either covalently (C_{sp^3}–Si or C_{sp^3}–Sn) or by a coordination bond (usually a chelating group attached to transition metal centers such as Ti, Zr, etc.), can be used to build a covalent interface between the two types of components. R groups are not removed upon hydrolysis.

A large number of silicon precursors are commercially available, or can be prepared by well-established routes such as hydrosililation with $HSi(OR')_3$, or $HSiCl_3$.[21,22]

$$HSiX_3 + H_2C{=}CH{-}R'' \xrightarrow{cat} R''{-}CH_2{-}CH_2{-}SiX_3 \quad X = Cl \text{ or } -R \tag{4}$$

When the functional alkene is not available, other pathways are possible, such as substitution or addition reactions on $(R'O)_3Si(CH_2)_3Cl$, $(R'O)_3Si(CH_2)_3N{=}C{=}O$, or $(R'O)_3Si(CH_2)_3O_2CC(CH_3){=}CH_2$. For example,

$$R'O_3Si(CH_2)_3N{=}C{=}O + HOR \longrightarrow R'O_3Si(CH_2)_3NHCO_2R \tag{5}$$

Other possibilities are the reaction of $(R'O)_3Si(CH_2)_3Cl$ with a Grignard or organolithium reagents.[23,24] Currently, almost all organic moieties or organometallic fragments can be linked to silicon through a stable C–Si bond.[25]

Regarding nonsilicon precursors, C–Sn bonds are stable enough toward hydrolysis. Numerous precursors such as $R'SnX_3$ (X = Cl, –OR) and trichloro organostannanes are available. The trichloroorganotin precursors can also be transformed into trialkynylorganotins by reaction with alkynyllithium.

For the more electropositive metals or transition metals (Al, Ti, Zr, V, Nb, Ce, etc.), the C–M bond is easily cleaved by hydrolysis. Therefore, the links generally used to build Class II hybrids are formed by strong complexing organic ligands, obtained by deprotonation of HLZ (L being an anchoring function, Z, a general organic group), such as β-diketones, β-ketoesters, carboxylic acids, phosphates, or phosphonates. A proton exchange reaction leads to the modified alkoxide [Eq. (6), L is the complexing group].

$$M(OR)_n + HLZ \longrightarrow M(OR)_{n-1}(LZ) + HOR \tag{6}$$

Bidentate or cyclic ligands increase considerably the M–(LZ) bond strength toward hydrolysis, as they usually increase coordination around the metal center and lower the reactivity. This strategy has been extensively used to control hydrolysis–condensation processes of transition metal centers by decreasing their reactivity. Indeed, firmly attached complexing ligands reduce the functionality (and thus the connectivity) of the inorganic centers, act as a poison toward oxo-polymer growth, and provide an anchoring point to the organic component. The organic group can also give steric or solvophobic protection.

Examples. A huge quantity of Class II materials, mostly prepared from the functional precursors described above, has been so far described in the literature. Mixtures of functional precursors and ordinary alkoxides have also been employed, to increase the amount of inorganic component by coprocessing. Some of the hybrids thus obtained bear special names such as ORMOSIL (ORganically MOdified SILicates) and ORMOCERs (ORganically MOdified CERamics). The organic groups R play several roles:

1. Groups with simple functions, such as alkyl or aryl bound to Si or Sn, or complexing acetylacetone (acac) on transition metals. These groups act mainly as network modifiers of the inorganic framework. They are also used to control the kinetics of hydrolysis and condensation, and they can also lead to special properties such as hydrophobicity or plasticity to the final material. The resulting matrices are particularly interesting to host small optically active organic molecules.
2. The R groups can bear an active nonpolymerizable moiety. Thus, R groups are network modifiers, which give a special property to the material. Some of the numerous examples include chromophores for nonlinear optics, storage or lasing, crown ethers for transportation and ion scavenging, or organometallic fragments for catalysis.
3. The R groups can bear a polymerizable function such as methacrylate or epoxy acting as network formers. If the R–M bond is not cleaved, then R can also be considered a network modifier. Most of the works in this area are centered on silicon-based hybrids for protective coatings.

The systems derived from Zr-*n*-propoxide modified by acetoacetoxyethylmethacrylate (AAEM) constitute a complete example of a Class II hybrid (Fig. 1F). The heterofunctional precursor is $Zr(O{-}Pr^n)_{4-x}(AAEM)_x$. AAEM acts as a network former and modifier, by limiting the growth of the inorganic polymer, and allowing the formation of the organic polymer. Simultaneously, growth of the inorganic polymers also hinders AAEM polymerization. Thus, both polymerization processes are interdependent. The local and short-range structure of the polymers is dependent

on the complexation ratio, leading to dense Zr-oxo polymers associated with short poly(AAEM) chains (low AAEM/Zr ratio), or smaller and tenuous oxo-polymers cross-linked with longer poly(AAEM) chains (large AAEM/Zr ratio).[15]

Bridging heterofunctional precursors that present two or more trialkoxysilyl groups capping an organic molecular spacer (phenyl, biphenyl, alkene, or alkyne) have been synthesized.[7,23,24,26] The hydrolysis of these precursors leads to bridged polysilsesquioxanes (POSS). Precursors with varied functionality, rigidity, and morphology have been synthesized to prepare tailored pore size materials. The porous volume, shape, and surface area of the resulting hybrids depend on the precursor geometry and hydrolysis conditions (solvent, catalysis, etc.). Interestingly, $Si–C_{sp}$ bonds can be cleaved in soft conditions (methanol, 60°C, F^- catalysis), which permits removal of the alkyne spacers and generation of open porosity.[26]

Preformed polymers capped by trialkoxysilyl groups have also been used to generate hybrid materials, particularly to tune mechanical or optical properties. Polymers such as poly(dimethylsiloxane), poly(tetramethylene oxide), and poly(aryleneetherketone) undergo hydrolysis and condensation, and can cocondense with silica or transition metal oxo-polymers, yielding cross-linked strongly grafted hybrids, in which the inorganic oxo-polymers can be used as mineral fillers or matrices, depending on the relative proportions of both components.

Nanocomposite Hybrids: From Nanodispersions to Organization

The methods described above correspond to those attainable with conventional sol–gel chemistry. This strategy is simple, low cost, and yields amorphous hybrid materials, which can contain specific organic molecules, biocomponents or polyfunctional cross-linkable polymers (e.g., telechelic polymers). These materials exhibit an infinity of microstructures, can be transparent, and easily shaped as films or bulks (Fig. 2, Route A). However, they are generally polydisperse in size and locally heterogenous in chemical composition.

A better understanding and control of the local and semilocal structure of these materials and their degree of organization in different length scales are important issues, especially if tailored properties are sought for. For example, in nanocrystalline materials, electronic and optical properties can be tailored by varying particle size, and it is expected that ordered arrays can give rise to controlled magnetic or electronic coupling, triggering new properties. Inorganic precursors with preformed symmetry or anisotropies in their structure or morphology (e.g., lamellar materials, rods) can be combined with organic molecules or polymers to tailor mechanical properties. Ordered metallic nanoparticles give rise to interesting optical properties, derived from the coupling of plasmon resonances. Periodicity in the tens to the hundreds of nanometers (i.e., mesoscopic scale) can be attained by combining inorganic precursors and self-assembly or controlled phase segregation techniques.

Organization usually involves a controlled segregation at the nanoscopic level. Several approaches may be conceived to achieve such a control of the materials structure; they are schematized in Fig. 2:

- Self-assembling procedures (SA, Route B).
- The assembling of well-defined nanobuilding blocks (ANBB, Route C).
- The self-assembly of NBB, which combines the first two approaches (SANBB, Route D).
- Integrative synthesis, where SA, ANBB, and SANBB are combined with controlled phase segregation.

In the last 10 yr, the field of ''chemistry of organized matter'' has opened the path to create organized or textured inorganic or hybrid networks, templated by organic structure directing agents (Fig. 1, Routes B and D).[18,27–32] The success of this strategy is also clearly related to the ability to control and tune hybrid interfaces, which permits the versatile building of a continuous range of nanocomposites, from ordered dispersions of inorganic bricks in a hybrid matrix to highly controlled nanosegregation of organic polymers within inorganic matrices. In the latter case, one of the most striking examples is the synthesis of mesostructured hybrid networks (Routes B and D).

Nanocomposite Hybrids Derived from Lamellar[2]

Hybrid materials derived from the combination of organic molecules or polymers and delaminated clays, other lamellar oxides, and eventually sulfides or carbons are well known and present industrial interest (note that we will not refer to ordinary composites made up of modified silicates that are used as gelling agents, thickeners, or fillers). The lively color of the pigment Maya Blue developed in the eighth century by the Mayans is due to indigo dyes trapped in a lamellar mineral, palygorskite.[39] Oxide or silicate nanoplatelets finely dispersed in a polymeric matrix present interesting mechanical properties, owing to the ultralarge interfacial area between the organic and inorganic

[2]See the article ''Polymer/Clay Nanocomposites.''

components, which leads to dramatic changes in the conformation polymer molecules can adopt. This is exemplified by the nylon–clay hybrids (NCH) developed by Toyota, which show good mechanical properties and a high heat-distortion temperature in spite of the low content of the clay mineral, making NCH the first hybrid material applied to automotive engine parts.[40] In addition, the NCH film has low gas permeability. Additionally, the presence of nanodispersed matter can give rise to new electrical or optical properties. On the other hand, organic molecules or polymers embedded within clay or oxide matrices present interest in pigments, sorption, environmental issues, or drug delivery. Layered aluminosilicates (2 : 1 layered silicates, clay minerals, phylosilicates, smectites) are the most commonly used inorganic NBB. These silicates generally bear charge compensating alkaline cations or anions (in the case of layered double hydroxides, LDH) in the interlayer. Organic ions can be incorporated within the interlayer galleries either by coprecipitation or by ion exchange. Typically, cations such as organic ammonium and phosphonium are used to replace interlayer alkaline or alkaline earth cations present in clays. For LDH matrices, carboxylates, phosphates, sulfates, or phosphonic or sulfonic acids carrying organic residues are used to exchange anions. The trapped groups act as an immobilized phase, and present interesting sorption and phase behavior, depending on the degree of ion exchange. Monomers (typically acrylate) or even polymers can also be incorporated, leading to a wide range of compositions. Delamination of these precursors can take place in the adequate solvents, leading to highly disperse metal-oxo hybrid monolayers, which can be included within a matrix or eventually polymerized. When the inorganic and polymeric components are mutually compatible, the obtained systems are exfoliated and intercalated, and the final structure is a nanodispersion. In principle, a great variety of dispersion morphologies and secondary textures in the mesoscopic or macroscopic should be available. Several methods to prepare these nanocomposites include solution processing, mesophase mediated processing, in situ polymerization, or melt processing. A wide range of polymers has been used, from polystyrene to conductive polyaniline.[41]

Hybrid Nanocomposites from Functional Nanoentities: Assembly of NanoBuilding Blocks (ANBB)[8,15,19,42]

A further method to reach a better definition of the inorganic component consists in the use of perfectly calibrated nanosized objects, such as clusters or nanoparticles that keep their integrity in the final material. These NBB can be capped with functional (solvophylic/solvophobic, self-assembly active or polymerizable, Fig. 3) ligands, or connected through organic spacers, like telechelic molecules or polymers, or functional dendrimers (Fig. 1, Route C). The use of highly precondensed species presents several advantages:

- They exhibit a lower reactivity toward hydrolysis or attack of nucleophilic moieties than metal alkoxides.
- The NBB components are nanometric; nowadays, numerous synthesis procedures are available, which permit to obtain monodispersed nanoparticles with perfectly defined structures, which facilitates the characterization of the final materials.
- These NBB present per se interesting properties, which are indeed size tunable (luminescence, conductivity, and electronic and magnetic behaviors).

The variety found in the NBB (nature, structure, and functionality) and possible links allows one to build an amazing range of different architectures and organic–inorganic interfaces, associated to different assembling strategies. Moreover, the step-by-step preparation of these materials usually allows for high control over their semilocal structure.

Strategies combining the NBB approach with the use of organic templates that self-assemble and allow one to control the assembling step are also appearing (Fig. 1, Route D). This combination between the ''nanobuilding block approach'' and ''templated assembling'' will have a paramount importance in exploring the theme of ''synthesis with construction'' of hierarchically organized materials, in terms of structure and functions.

The NBB exhibit a large variety of composition, size, and interfaces between the organic and the inorganic components (covalent bonding, complexation, electrostatic interactions, etc.). Depending on the set of chosen experimental conditions, these NBB will keep or loose their integrity. They can be used as true building blocks that can be connected through organic spacers or condensation reactions, or as a reservoir of inorganic matter, which can be delivered at the hybrid interface to build an extended inorganic network. In the following section, several examples of hybrid materials and systems in which the core integrity of the NBB is preserved will be presented.

CLUSTER-BASED HYBRIDS[8,43]

Several well-documented routes exist for a variety of NBB such as oligosilsesquioxanes and derivatives, organotin-oxo clusters, organically functionalized heteropolyoxo-tungstates, transition metal oxo clusters,

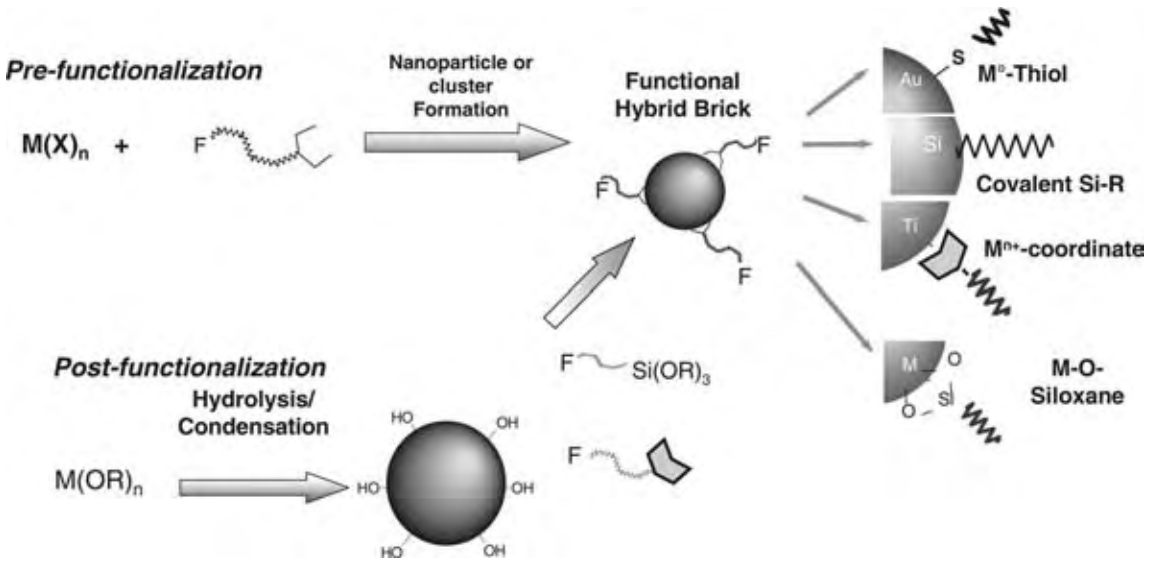


Fig. 3 Pre- and postfunctionalization routes leading to functional hybrid metal or oxide clusters and nanoparticles. Functional hybrids are the nanobuilding blocks (NBB) of complex hierarchical structures.

and finally with functionalized nanosized particles (metallic oxides, metals, chalcogenides).

A variety of Si-based building blocks have been used to elaborate hybrid materials, corresponding mostly to cage compounds of the general formula $[XSiO_{1.5}]_n$, where n is an even number ranging from 6 to 18.[44] The types of architectures that can be achieved depend on the nanobrick functionality. Nanobuilding blocks bearing a single double-bond yield upon polymerization (with or without a comonomer), linear polymeric backbones with pending silsesquioxane cubes. Such functional nanobricks allow a good dispersion and covalent anchoring of the nanoscale inorganic filler in the polymer, resulting in tuned changes in the thermal (i.e., glass transition temperatures) or mechanic (i.e., relaxation modulus) properties of the polymer with the inorganic loading.[8] Pendent POSS can also be introduced through polycondensation reactions as shown for segmented polyurethane-based elastomers. In cocondensed systems, the inorganic NBB can be selectively attached to one of the polymer segments, enhancing the tensile modulus and the strength of the elastomer.[45] All the examples presented above correspond to Class II hybrid materials, but POSS have also been simply blended with polymers to afford Class I hybrids. In such systems, the dispersability of the POSS is related to the organic side groups.

Tin–C_{sp^3} bonds are stable toward hydrolysis, and many organic functions can be added to Sn centers. Tin-6 and Tin-12 clusters have been used as NBB through different routes.[43] The Tin-6 cluster presents a variety of external functions (alkyl, vinyl, amine, etc.). Tin-12 or $\{(RSn)_{12}(\mu_3\text{-}O)_{14}(\mu_2\text{-}OH)_6\}^{2+}$ presents a versatile mixed interface. Assembly of these clusters within organic networks is possible by polymerization of the attached R groups containing double bonds. The positive charge of the cluster can be compensated with anionic monomers (acrylic acid, acrylamide), which can be subsequently polymerized. Compensating organic dianions or telechelic polymers can also be used to bridge the clusters and integrate them within polymeric matrices.

Several metal-oxo clusters (Ti^{IV}, Zr^{IV}, Ce^{IV}, Nb^{V}, and mixed composition) have been described in the literature.[8] Prefunctionalized clusters are interesting and controlled precursors for hybrids. Yet, because of their small size and high reactivity, functionalized clusters exhibit rather poor stability toward water or nucleophiles. Consequently, their assembly through radical initiated polymerization must be performed in nonprotic solvents such as toluene, benzene, or THF. Following this pathway, novel hybrid materials have been produced, in which the cluster size and identity is conserved. Clusters can also be connected by oxo bridges (in organic solvents, low water conditions) or organic spacers. The surface metallic atoms can be selectively complexed by groups such as carboxylate. Block copolymers or functional dendrimers with tuned functions, size, and symmetry have been used to create ordered NBB dispersions or mesoporous cluster-based solids. Mo-, V-, or W-based polyoxometallate (POM) NBB has also been used to generate hybrids with potential applications in electrochromism, photoelectrochemistry, sensors, catalysis, and light imaging.[46] Polyoxometallate-based Class I materials are based on weak interactions, as a result of the anionic charge carried by the POM. Polyoxometallates can be embedded in conducting polymers and used as

conductive electrodes. They have also been assembled in self-assembled monolayers or multilayers. The first example of Class II POM derivatives is the homopolymerization of difunctional POMs $[SiW_{11}O_{39}(RSi)_2O]^{4-}$ with unsaturated organic groups (vinyl, allyl, methacrylate, and styryl). These compounds present great potential in catalysis, chemotherapy, photo- and electrochromism.

NANOPARTICLE-BASED HYBRIDS[8]

Because of the stability problems that might be associated to small clusters with high reactivity, the devices sought nowadays are mainly based on nanoparticles (NP) ranging from 2 to 100 nm in diameter. Physical as well as wet chemical routes are converging toward a sound library of functionalized NP production procedures.[47–52] Functionalized NP already play an important role in pharmaceutical drug delivery systems, paint dispersion, tire reinforcement, catalyst optimization, and in many processes involving adhesives, biocements, varnishes, and lubricants. There are several ways to adapt the NBB approach to create nanoparticle-based hybrid materials. These include: 1) introducing NP into polymer or organic hosts; 2) synthesizing the nanoparticles inside such matrices; 3) connecting particles with adequate organic spacers; or 4) polymerizing functionalized metal-oxo nanoparticles. All these strategies have been shown above in the case of clusters. Building the hybrid structures requires a similar tuning at the hybrid interface level. Additionally, colloidal interactions must be taken into account. The resulting hybrid networks can be amorphous, nanostructured, or exhibit long-range ordering. The synthesis and nature of functionalized nanoparticles has been extensively reviewed.[50] (For a comprehensive review of synthesis methods and properties, see also Ref.[53].)

Organic functions can be added to the nanoparticle surface in a step subsequent to their synthesis (post-functionalization). This process strongly depends on the particle nature (Fig. 3).[54] Silica particles are functionalized by reactions in organic media with alcoxysilane [R′-Si(OR)$_3$] or chlorosilane (R′-SiCl$_3$) species; some organo-functionalized nanosilicas have already been commercialized by Degussa and Clariant. The hybridization strategy of nanosilicas has been used for the functionalization of alumina, zirconia, or titania nanoparticles with trimethoxysilylpropylmethacrylate, to design reactive ceramic fillers for PMMA-based composites. A variety of nanoparticles can be derivatized through the use of a functionalized complexing ligand (carboxylate, acetylacetone, thiols, silanes, etc.). Smart methods for polymerizing MMA at the surface of nanometric titania particles have already been developed in the paint and polymer industry.[55]

MULTISCALE ORDERING OF FUNCTIONAL COLLOIDAL NANOPARTICLES:[4]

Multiscale ordering of functional colloidal nanoparticles is a powerful technique for the creation of macroscopic devices. This can be performed via selective polymerization, self-assembly processes, or through controlled molecular recognition processes (Fig. 4). Further assembly between NBB is addressed following three main strategies, namely, electrostatic coupling, covalent, or self-assembly-based noncovalent binding.

Networks formed through electrostatic interactions: Mesoscopic structures of TiO_2 anatase nanoparticles have been formed through self-assembly processes involving Class I hybrid composites made of multiply charged polytitanate anions and tetramethylammonium cations. These anatase nanocrystals can self-assemble into highly ordered superlattices. The assembly of gold particles covered by quaternary ammonium bromide salts ($R_4N^+Br^-$) in 2D and 3D organized systems has been observed, the separation between the objects being controlled by the length of R. Nanosized silica NBB functionalized with primary amines can be coupled with a ''counter NBB'' displaying a complementary surface. The complementary NBB can be a gold nanoparticle, capped on its surface by thiol groups that carry alkylcarboxylic functions. Simple acid–base chemistry induces an immediate charge pairing, which results in the spontaneous formation of electrostatically bound mixed colloids.

Covalent networks: Covalent binding programmed assembly can be simply designed by mixing two sets of NBB presenting surfaces functionalized with complementary reactive functions (e.g., amine and aldehyde). A more sophisticated derivation is constituted by the so-called biomolecular route, which makes use of specific interactions, as those found in antigen–antibody pairs or streptavidin–biotin complexes, to induce assembling of NBB. A DNA-based method has also been presented, where noncomplementary DNA oligo-nucleotides attached on gold particles aggregate in the presence of an oligonucleotide duplex with ''sticky ends,'' which are complementary to the grafted sequences. The nanoparticles self-assemble into macroscopic materials, this assembly being reversible upon thermal denaturation.

Self-assembly-based networks: Ordered superlattices composed of nanosized semiconducting sulfides have been synthesized within lyotropic phases.[56] Hexagonal-packed arrays of nanocrystalline CdS (or similar structures such as ZnS, $Cd_{1-x}Zn_xS$, and CdSe) have been produced, a ''mineral copy'' of an (ethylene oxide)$_{10}$-oleyl/water mesophase presenting periodicities ranging between 7 and 10 nm.

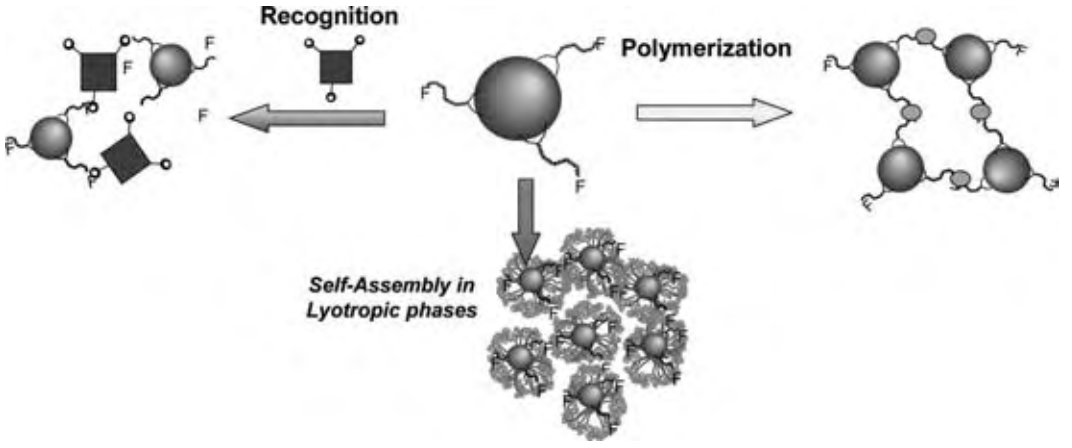


Fig. 4 Assembly strategies to organize NBB (functional clusters or nanoparticles) into hierarchical structures, by means other than colloidal interactions. *(View this art in color at www.dekker.com.)*

Textured Inorganic or Hybrid Materials with Tailored Porosity

Hybrid porous materials are interesting in that they combine inorganic–organic frameworks with tailored pores, leading to potential applications in catalysis, optics, sensing and separation, among others. Hybrid microporous materials have been mostly produced by hydrothermal synthesis, in a similar way to zeolites. The synthesis strategies and computational modeling leading to these solids have been extensively discussed in the literature, and will not be described in this entry.[57,58]

Sol–gel techniques can also be combined with the self-assembly of surfactants, leading to hybrid materials in which the function imparted by the organic component is to give order in the mesoscopic (2–50 nm) range. Inorganic condensation can proceed around micelles or micellar assemblies, leading to a hybrid composite in which the surfactant aggregates act as templates for cavities or pore arrays.[27] A variety of ionic and nonionic templates have been used to create porous matrices with organized monodisperse pores.[28] Template removal leads to mesoporous materials in the form of films, fibers, powders, monoliths, or macrospheres, which have promising applications in optics, electronics, chemical sensing, catalysis, separation, etc. Some templates can impart interesting properties to a hybrid nanostructured material; this is, for example, the case of sulfonated PS-based polymers, which can give rise to proton conducting robust hybrid membranes.[59] Apart from micellar templates, other methods to texture materials include organogel molecular assembly, latex or ceramic beads, or spinodal phase separation. The so-formed organic domains dimensionality and self-arrangement dictate the final porous structure of the inorganic network. A control at three levels is necessary to achieve such a control on the structure:

1. Tuning of the initial solution chemical composition allows selection of the desired stoichiometry, the reactivity, and adjustment of preferential interactions at the organic/inorganic interface when segregation and self-assembly take place.
2. Organized hybrid materials can be provoked through chemical (i.e., precipitation, gelation) or physical (i.e., evaporation, temperature) solicitations. This step is the most delicate one because phase segregation, self-assembly, and condensation must occur in this exact order to end up with the desire architecture.
3. Finally, the template has to be removed from the as-prepared hybrid materials to create monodisperse porosity.

So far, such nano-structuring approach has been well developed for pure or mixed metal oxide inorganic networks such as SiO_2, Al_2O_3, transition metal oxides, and nonoxide materials such as carbon, metals, or polymers.[60] Mesoporous materials have been prepared mostly by three different routes: precipitation methods, true liquid crystalline templating (TLCT), or evaporation-induced self-assembly (EISA). These alternative methods are complementary: while precipitation leads to high yields of a limited choice of inorganic networks, liquid TLCT or EISA permits a higher control of the inorganic condensation and processing as films, gels, powders, aerosols, or fibers.[27,28]

The use of functional precursors permits one to synthesize meso-organized materials with organic functions, either embedded within the walls or dangling in the pore interior. This can be performed by one-pot or postfunctionalization procedures

(Fig. 5).[61,62] A great number of organic functions have been included in mesoporous silica by cocondensation of silicon alkoxides and organosilane precursors [R′-$Si(OR)_3$, or bridging $(OR)_3Si$-R′-$Si(OR)_3$]. The connectivity, the hydrophilic–hydrophobic nature, and the specific interactions of the R′ group are essential for the placing of the organic functions either within or dangling from the walls. For example, polar R′ groups such as amino tend to place themselves at the organic/inorganic interface. Once the surfactant is removed by solvent extraction and the network stabilized by a chemical route, the R group remains covalently bonded to the framework and is in principle accessible through the porosity. Organic functions (pH probes, hydrophobic, electrochemical probes, etc.) can also be grafted onto silica or transition metal oxide walls by postfunctionalization by attaching organosilanes or complexing ligands (carboxylates, beta-diketonates, phosphonates, etc.) to the pore surface. This leads to hybrid mesostructured materials with tunable surfaces, opening a land of opportunities for designing new separation devices, catalysts, membranes, sensors, and nano-reactors. Biospecies such as active enzymes can be trapped within the pore network, by using adequate grafting groups and spacers.[63]

Morphology texturation by growth modifiers[27,64]

Organogelators are low-weight organic molecules that are able to form thermoreversible physical gels, exhibiting strongly anisotropic structures (i.e., fibers, ribbons, platelets), in a great variety of organic solvents. Combining this new family of texturing agents [with steroidal or diaminocyclohexane skeleton, 2,3-bis-*n*-decyloxyanthracene (DDOA) or 2,3-di-*N*-alkoxyphenazine (DAP)] with the sol–gel chemistry has recently permitted directing the condensation of silica into original shapes at the microscopic levels, and succeeded in transcripting the chiral information to a mineral network. DDOA has been successfully employed as template for fibrous silicas and aluminas with tuneable mesoporosity. Organically modified functionalized hybrid fibers with accessible and modifiable functionalities have also been synthesized from silicon alkoxides and organosilanes. Postfunctionalization can be performed in a second step with any other inorganic complex or even with fragile organic or biofunctions. Recently, novel organic–inorganic hybrids that present helical symmetries have been obtained through the hydrolysis of organosilica derivatives bearing an *R,R* or *S,S* chiral diureidocyclohexane spacer. Left- and right-handed helixes are self-generated depending of the configuration of the chosen organic subunit.

Texturation through phase separation

Phase separation in multicomponent systems can be used as a texturing tool.[65] In principle, three- or four-component systems (i.e., an inorganic alkoxide precursor, a texturing polymer—typically PEG, and a binary solvent) can lead to spinodal separation patterns upon solvent evaporation. Phase separation can be entropically or enthalpically driven, according to whether polymer or solvent droplets are segregated, respectively. The segregated droplets lead to pores in the micrometer range. Domain size, morphology, and dispersion depend on the rate and extent of species concentration and their diffusion in the medium. Mesoporosity can arise from template-inorganic residual interactions, leading to hierarchical porous systems. Original macrotextures shaped with coral-like, helical, or macroporous morphologies have been obtained; applications are sought in separation and chromatography. Functional nanoparticles can also be used as building blocks, in combination with biomimetic polymers such as poly-γ-benzyl-l-glutamate

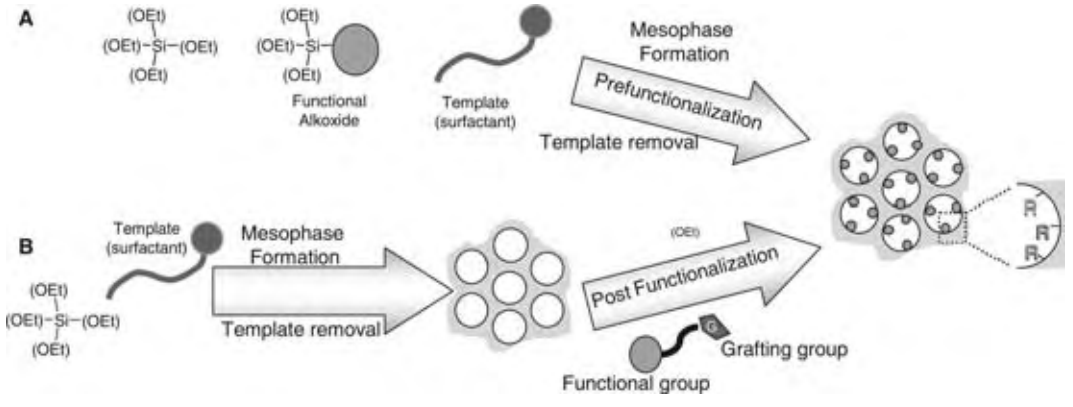


Fig. 5 Prefunctionalization (A) and postfunctionalization (B) routes toward hybrid materials organized in the mesoscopic scale. The meso-organized precursors are synthesized by reacting an inorganic precursor in the presence of a supramolecular template (surfactant). By Route A, organic functions can also be embedded within the walls. *(View this art in color at www.dekker.com.)*

(PBLG). By adjusting a single parameter, such as the template to inorganic ratio, a versatile tuning between templating effect and phase separation yields hierarchical porous materials presenting both micro and macro porosity with inorganic walls constituted of nano-crystalline cerium oxide particles.

BIOHYBRIDS[66–70]

In Nature, the combination of simple NBB such as silica, calcium carbonate, or calcium phosphate with organic biopolymers gives rise to biocomposites with well-defined mechanical properties and biological functions (protection, motion, sensing, storage, etc.). Natural materials such as teeth, bone, shells, etc., are a natural source of inspiration for the complex hybrid materials shown in the previous sections, where the structural role of the organic molecules was emphasized. But smaller biomolecules or molecule arrays (DNA, proteins, peptides, membranes, hormones) can perform highly selective and specific tasks such as molecular recognition, selective transport, or biocatalysis. Biomolecules are also fragile. Therefore, there is a great interest in immobilizing active species or even living organisms in suitable robust matrices.

Biohybrid materials are expected to find applications in smart biodevices such as sensors, biocatalysts, separation, and vectorization domains, etc. In the last few years, the encapsulation of biomolecules (proteins, enzymes, etc.) in silica gel has proven to be a promising alternative to biopolymer-based processes. Specific biosensors composed of enzymes immobilized in silica xerogels have recently been produced. Another developing domain concerns hybrids formed from inorganic nanoparticles or inorganic gels and biomolecules. Biotechnologies already use enzymes and bacteria as synthetic tools; their further encapsulation in solid matrices should bring modulated and enhanced biosynthetic properties. The exploitation of hybrid materials in domains including immunology tests, encapsulation, and/or vectorization is currently being tested. Good preservation of the enzyme activity can be tested by optical or electrochemical methods.

The possibility of using such inorganic hosts for whole-cell immobilization has been far less studied. In the case of bacteria, different approaches have been explored to maintain cell viability both during the encapsulation step and on a long-term scale. So far, the best conditions for bacteria survival involve an aqueous route using silicates and colloidal silica as inorganic precursors. Additives have been incorporated either during the gel formation or after bacteria encapsulation, leading to viability rates up to 85% after 1 mo of immobilization in silica.[71] The use of silica gels for the design of cell-based bioreactors or biosensors can now be envisioned.

CONCLUSIONS

Hybrid organic–inorganic materials are increasingly taking their position in the free spaces left between inorganic chemistry, polymer chemistry, organic chemistry, and biology. This land of research, initially worked out by the sol–gel community is at present thriving with the appearance of hybrid structures engineered from the molecular to the nanometric or micrometric scales, toward applications spanning biological and chemical sensing, catalysis, selective separation to optical communications. Nature is a source of inspiration for hybrid materials, in many properties such as sophistication, miniaturization, hierarchical organizations, hybridization, resistance, and adaptability. Hybrid materials that combine organic and inorganic components on a nanoscale with innovative controlled textures will allow designing of new or bioinspired materials reproducing principles or structures found in Nature.

Looking forward to the 21st century, nanosciences will be, as biology, one of the fields that will contribute to a high level of scientific and technological developments. Hybrid materials present the paramount advantage to both facilitate integration and miniaturization of the devices (nanomaterials, nanotechnologies) and afford a direct connection between inorganic, organic, and biological worlds. Functional precursors or functional NBB facilitate integrative synthesis pathways, where synergistic assembling and morphosynthesis can be strongly coupled. The chemical strategies offered by coupling soft chemistry processes with different macro templates (latex, bacteria, polymers, topological defects in LC, etc.), self-assembly processes, and external fields will allow, through an intelligent and tuned coding, to develop a new vectorial chemistry, able to direct the assembling of a large variety of structurally well-defined inorganic networks into complex architectures. The research of multiscale structured hybrids (from nanometer to millimeter) will open a land of opportunities for designing new materials. The future of this unifying field of research mainly depends on skills of all kind of chemists and is only limited by their imagination.

REFERENCES

1. Wilkes, G.L.; Orler, B.; Huang, H.H. Polymer Prep. **1985**, *26*, 300.
2. Sur, G.-S.; Mark, J.E. Eur. Polym. J. **1985**, *21*, 1051.

3. Wang, B.; Wilkes, G.L.; Smith, C.D.; McGrath, J.E. Polym. Commun. **1991**, *32*, 400.
4. Huang, H.H.; Orler, B.; Wilkes, G.L. Macromolecules **1987**, *20*, 1322.
5. Gómez Romero, P.; Sanchez C., Eds. *Functional Hybrid Materials*; Wiley: Chicester, 2004.
6. Sanchez, C., Ed. *Matériaux Hybrides*; Observatoire Français des Technique Avancées: Masson, Paris, France, 1996.
7. Loy, D.A.; Shea, K. Bridged polysilsesquioxanes. Highly porous hybrid organic–inorganic materials. Chem. Rev. **1995**, *95*, 1431.
8. Sanchez, C.; Soler-Illia, G.J.A.A.; Ribot, F.; Mayer, C.; Cabuil, V.; Lalot, T. Designed hybrid organic–inorganic nanocomposites from functional nanobuilding blocks. Chem. Mater. **2001**, *13*, 3061.
9. Sanchez, C.; Lebeau, B.; Chaput, F.; Boilot, J.-P. Optical properties of functional hybrid organic-inorganic nanocomposites. Adv. Mater. **2003**, *15*, 1969.
10. Sanchez, C.; Lebeau, B. Design and properties of hybrid organic–inorganic nanocomposites for photonics. MRS Bull. **2001**, May, 377.
11. Gómez-Romero, P. Hybrid organic–inorganic materials. In search of synergic activity. Adv. Mater. **2001**, *13*, 163.
12. Loy, D.A. MRS Bull. **2001**, May (special issue on Hybrid Materials), 364.
13. Eckert, H.; Ward, M., Eds. Chem. Mater. **2001**, (special issue on Nanostructured and Functional Hybrid Organic-Inorganic Materials), 13.
14. Arkles, B. Commercial applications of sol–gel-derived hybrid materials. MRS Bull. **2001**, May, 402.
15. Sanchez, C.; Ribot, F. Chemical design of hybrid organic-inorganic materials synthesized via sol–gel. New J. Chem. **1994**, *18*, 1007.
16. Brinker, C.J.; Scherrer, G.W. *Sol–Gel Science, The Physics and Chemistry of Sol-Gel Processing*; Academic Press: San-Diego, CA, 1990.
17. Livage, J.; Henry, M.; Sanchez, C. Sol–gel chemistry of transition metal oxides. Prog. Solid State Chem. **1988**, *18*, 259.
18. Mann, S.; Burkett, S.L.; Davis, S.A.; Fowler, C.E.; Mendelson, N.H.; Sims, S.D.; Walsh, D.; Whilton, N.T. Sol–gel synthesis of organized matter. Chem. Mater. **1997**, *9*, 2300.
19. Ribot, F.; Sanchez, C. Organically functionalized metallic oxo-clusters: structurally well-defined nanobuilding blocks for the design of hybrid organic-inorganic materials. Comments Inorg. Chem. **1999**, *4–6*, 327.
20. Novak, B.M. Hybrid nanocomposite materials-between inorganic glasses and organic polymers. Adv. Mater. **1993**, *5*, 422.
21. Noll, W. *Chemistry and Technology of Silicones*; Academic Press: New York, 1968.
22. Deschler, U.; Kleinschmit, P.; Panster, P. 3-Chloropropyltrialkoxysilanes—key intermediates for the commercial production of organofunctionalized silanes and polysiloxanes. Angew. Chem. Int. Ed. Eng. **1986**, *25*, 236.
23. Corriu, R.J.P.; Moreau, J.J.E.; Thepot, P.; Wong Chi Man, M. New mixed organic–inorganic polymers: hydrolysis and polycondensation of bis(trimethoxysilyl)organometallic precursors. Chem. Mater. **1992**, *4*, 1217.
24. Corriu, R.J.P. Ceramics and nanostructures from molecular precursors. Angew. Chem. Int. Ed. **2000**, *39*, 1376.
25. Schubert, U. Catalysts made of organic–inorganic hybrid materials. New J. Chem. **1994**, *18*, 1049.
26. Corriu, R.J.P.; Moreau, J.J.E.; Thepot, P.; Wong Chi Man, M. Hybrid silica gels containing 1,3-butadiyne bridging units. Thermal and chemical reactivity of the organic fragment. Chem. Mater. **1996**, *8*, 100.
27. Soler-Illia, G.J.A.A.; Sanchez, C.; Lebeau, B.; Patarin, J. Chemical strategies to design textured silica and metal oxide-based organised networks: from nanostructured networks to hierarchical structures. Chem. Rev. **2002**, *102*, 4093.
28. Soler-Illia, G.J.A.A.; Crepaldi, E.L.; Grosso, D.; Sanchez, C. Block copolymer-templated mesoporous materials. Curr. Opin. Colloid Surf. Sci. **2003**, *8*, 109.
29. Ying, J.Y.; Mehnert, C.; Wong, M.S. Synthesis and applications of supramolecular-templated mesoporous materials. Angew. Chem. Int. Engl. Ed. **1999**, *38*, 57.
30. Brinker, C.J.; Lu, Y.; Sellinger, A.; Fan, H. Evaporation-induced self assembly: nanostructures made easy. Adv. Mater. **1999**, *11*, 579.
31. Ozin, G.A. Panoscopic materials: synthesis over ''all'' length scales. Chem. Commun. **2000**, 419.
32. Stein, A.; Melde, B.J.; Schroden, R.C. Hybrid inorganic–organic mesoporous silicates—nanoscopic reactors coming of age. Adv. Mater. **2000**, *12*, 1403.
33. Leroux, F.; Besse, J.P. Polymer interleaved layered double hydroxide: a new emerging class of nanocomposites. Chem. Mater. **2001**, *13*, 3507.
34. Prevot, V.; Forano, C.; Besse, J.-P. Appl. Clay Sci. **2001**, *18*, 3.
35. Vaia, R.A.; Giannellis, E.P. Polymer nanocomposites: status and opportunities. MRS Bull. **2001**, May, 394.
36. Newman, S.P.; Jones, W. Synthesis, characterization and applications of layered double hydroxides containing organic guests. New J. Chem. **1998**, 105.

37. Vaia, R.A.; Krishnamoorti, R., Eds. *Polymer Nanocomposites*; American Chemical Society: Washington, DC, 2001.
38. *Nanocomposites 1999: Polymer Technology for the Next Century*; Principia Partners: Exton, PA, 1999.
39. Gómez-Romero, P.; Sanchez, C. Hybrid materials. Functional properties. From Maya Blue to 21st century materials. New J. Chem. **2005**, *29*, 57.
40. Usuki, A.; Kojima, Y.; Kawasumi, M.; Okada, A.; Fukushima, Y.; Kurauchi, T.; Kamigaito, O. J. Mater. Res. **1993**, *8*, 1179.
41. Ruiz-Hitzky, E. Conducting polymers intercalated in layered solids. Adv. Mater. **1993**, *5*, 334.
42. Bourgeat-Lami, E. Organic–inorganic nanostructured colloids. J. Nanosci. Nanotech. **2002**, *2*, 1 (and references therein).
43. Schubert, U. Polymers reinforced by covalently bonded inorganic clusters. Chem. Mater. **2001**, *13*, 3487.
44. Sellinger, A.; Laine, R. Silsesquioxanes as synthetic platforms. Thermally curable and photocurable inorganic/organic hybrids. Macromolecules **1996**, *29*, 2327.
45. Lichtenhan, J.D.; Vu, N.Q.; Carter, J.A.; Gilman, J.W.; Feher, F.J. Silsesquioxane-siloxane copolymers from polyhedral silsesquioxanes. Macromolecules **1993**, *26*, 2141.
46. Katsoulis, D.E. A survey of applications of polyoxometalates. Chem. Rev. **1998**, *98*, 359.
47. Bradley, J.S.; Chaudret, B., Eds. New J. Chem. **1998**, *22* (7 special issue on Nanoparticles, Part I).
48. Bradley, J.S.; Chaudret, B., Eds. New J. Chem. **1998**, *22* (11, special issue on Nanoparticles, Part II).
49. Schmid, G. Chem. Rev. **1992**, *92*, 1709.
50. Jolivet, J.-P. *Metal Oxide Chemistry and Synthesis: From Solution to Solid State*; Wiley–VCH: Chichester, 2000.
51. Interrante, L.V.; Hampden-Smith, M., Eds. *Chemistry of Advanced Materials: An Overview*; Wiley-VCH: New York, 1998.
52. Edelstein, A.S.; Cammarata, R.C., Eds. *Nanomaterials, Synthesis, Properties and Applications*; Institute of Physics Publishing: Philadelphia, PA, 1996.
53. Hadjipanayis, G.C.; Siegel, R.W., Eds. Nanophase Materials: Synthesis—Properties—Applications; Nato ASI Series, E. Appl. Sci. **1994**, 260.
54. Caruso, F. Nanoengineering of particle surfaces. Adv. Mater. **2001**, *13*, 11.
55. Hofman-Caris, C.H.M. Polymers at the surface of oxide nanoparticles. New J. Chem. **1994**, *18*, 1087.
56. Stupp, S.I.; Braun, P.V. Nanostructure templating in inorganic solids with organic lyotropic liquid crystals. Science **1999**, *277*, 1242.
57. Férey, G. Microporous solids: from organically templated inorganic skeletons to hybrid frameworks. Ecumenism in chemistry. Chem. Mater. **2001**, *13*, 3084.
58. Mellot-Draznieks, C.; Dutour, J.; Férey, G. Angew. Chem. Int. Ed. **2004**, *43*, 6290.
59. Mauritz, K.A.; Mountz, D.A.; Reuschle, D.A.; Blackwell, R.I. Self-assembled organic/inorganic hybrids as membrane materials. Electrochim. Acta **2004**, *50*, 565 (and references therein).
60. Schüth, F. Non-siliceous mesostructured and mesoporous materials. Chem. Mater. **2001**, *13*, 3184.
61. Shi, J.L.; Hua, Z.L.; Zhang, L.X. Nanocomposites from ordered mesoporous materials. J. Mater. Chem. **2004**, *14*, 795 (and references therein).
62. Sayari, A.; Hammoudi, S. Periodic mesoporous silica-based organic–inorganic nanocomposite materials. Chem. Mater. **2001**, *13*, 3151.
63. Chong, A.S.M.; Zhao, X.S. Design of large-pore mesoporous materials for immobilization of penicillin G acylase biocatalyst. Catal. Today **2004**, *93–95*, 293.
64. Van Bommel, K.J.C.; Friggeri, A.; Shinkai, S. Organic templates for the generation of inorganic materials. Angew. Chem. Int. Ed. **2003**, *42*, 980 (and references therein).
65. Nakanishi, K. Pore structure control of silica gels based on phase separation. J. Porous Mater. **1997**, *4*, 67.
66. Sanchez, C.; Arribart, H.; Giraud-Guille, M.M. Biomimetism and bioinspiration as tools for the design of innovative materials and systems. Nat. Mater. *in press* (and references therein).
67. Böttcher, H.; Soltmann, U.; Mertig, M.; Pompe, W. Biocers: ceramics with incorporated microorganisms for biocatalytic, biosorptive and functional materials development. J. Mater. Chem. **2004**, *14*, 2176.
68. Aizenberg, J.; Livage, J.; Mann, S., Eds. J. Mater. Chem. **2004**, *14* (special issue in New Developments in Bio-related Materials), 2059.
69. Mann, S. Biomineralization. In *Principles and Concepts in Bioinorganic Materials Chemistry*; Oxford University Press: Oxford, 2001.
70. Bäuerlein, E. Biomineralization of unicellular organisms: an unusual membrane biochemistry for the production of inorganic nano- and microstructures. Angew. Chem. Int. Ed. **2003**, *42*, 614 (and references therein).
71. Nassif, N.; Bouvet, O.; Räger, M.N.; Roux, C.; Coradin, T.; Livage, J. Living bacteria in silica gels. Nat. Mater. **2002**, *1*, 42.

Hydrocracking

James G. Speight
CD & W Inc., Laramie, Wyoming, U.S.A.

INTRODUCTION

The use of hydrogen in thermal processes is perhaps the single most significant advance in refining technology during the 20th century. The process uses the following principle—the presence of hydrogen during a thermal reaction of a petroleum feedstock terminates many of the coke-forming reactions and enhances the yields of the lower boiling components, such as gasoline, kerosene, and jet fuel.

Hydrogenation processes for the conversion of petroleum fractions and petroleum products may be classified as destructive and nondestructive. Destructive hydrogenation (*hydrogenolysis* or *hydrocracking*) is characterized by the cleavage of carbon–carbon linkages accompanied by hydrogen saturation of the fragments to produce lower boiling products. Such treatment requires severe processing conditions and the use of high hydrogen pressures to minimize phase separation followed by polymerization and condensation reactions that lead to coke formation. Many other reactions, such as isomerization, dehydrogenation, and cyclization, occur under the drastic conditions employed. Thus, *hydroprocessing* is a thermal conversion process in which hydrogen is used to accomplish the objectives of the refiner. Hydrotreating is a process in which hydrogen is used to convert heteroatom constituents into their heteroatom hydrogen analogs and hydrocarbons:

$$R\text{–}S\text{–}R1 + H^2 \rightarrow RH + R^1H + H_2S$$

On the other hand, *hydrocracking* is a process in which thermal decomposition is extensive and the hydrogen assists in the removal of the heteroatoms (nitrogen and sulfur) as well as mitigating the coke formation that usually accompanies thermal cracking of high molecular weight polar constituents. In addition to hydrodesulfurization and hydrodenitrogenation, the removal of aromatic constituents from some product streams has also become essential. The high aromatic content in diesel fuel has been recognized both to lower the fuel quality and to contribute significantly to the formation of undesired emissions in exhaust gases. Indeed, as a result of the stringent environmental regulations, processes for aromatic reduction in middle distillates have received considerable attention in recent years. Studies have shown that existing middle distillate hydrotreaters designed to reduce sulfur and nitrogen levels would lower the diesel aromatics only marginally. The major differences between *hydrotreating* and *hydrocracking* are the time at which the feedstock remains at reaction temperature and the extent of the decomposition of the nonheteroatom constituents and products. The upper limits of hydrotreating conditions may overlap with the lower limits of hydrocracking conditions. And where the reaction conditions overlap, feedstocks to be hydrotreated will generally be exposed to the reactor temperature for shorter periods; hence the reason why hydrotreating conditions may be referred to as mild.

The Hydrocracking Process

The purpose of hydrocracking is to convert high-boiling feedstocks to lower boiling products by cracking the hydrocarbons in the feed and hydrogenating the aromatic and unsaturated materials in the product streams.[1–6]

Thus, the hydrocracking process is petroleum refining that combines catalytic cracking and hydrogenation by which high molecular weight viscous (usually non- or low-volatile) feedstocks are cracked (thermally decomposed) in the presence of hydrogen to produce lower molecular weight more volatile products. The process employs high pressure, high temperature, a catalyst, and hydrogen. Hydrocracking is used for feedstocks that are difficult to process by either catalytic cracking or reforming, because these feedstocks are characterized usually by high polycyclic aromatic content and/or high concentrations of the two principal catalyst poisons, sulfur and nitrogen compounds.

The hydrocracking process largely depends on the nature of the feedstock and the relative rates of the two competing reactions, hydrogenation and cracking. Heavy aromatic feedstocks are converted into lower-boiling products under high temperatures (400–480°C, 750–900°F) and high hydrogen pressures (1000–2000 psi, often as high as 6000 psi for very heavy feedstocks), in the presence of specialty catalysts. When the feedstock has a high content of paraffinic constituents, the primary function of hydrogen is to prevent the

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008049

formation of polycyclic aromatic compounds that can result when paraffins are thermally decomposed at high temperatures. Another important role of hydrogen in the hydrocracking process is to reduce, or even prevent, the buildup of coke on the catalyst. Hydrogenation also serves to convert sulfur and nitrogen compounds present in the feedstock to hydrogen sulfide (H_2S) and ammonia (NH_3).

The hydrocracking process often produces relatively large amounts of iso-butane that can be used in alkylation units to prepare alkylate for gasoline blending. Hydrocracking, depending on the catalyst, can also cause isomerization of the paraffinic products that benefit liquid fuels in terms of pour point control and smoke point.

The hydrocracking process can either be a single-stage or a two-stage process.[4,6] In the single-stage process, the preheated feedstock and a hydrogen-rich gas (usually >60% v/v hydrogen) are allowed to contact the catalyst. However, the catalyst and process parameters must be versatile to allow hydrocracking, hydrodesulfurization, and hydrodenitrogenation to occur. This type of process may be preferred for gas oil feedstocks where contamination by high amounts of sulfur and nitrogen is not always a major issue. After the products leave the reactor, they are sent to a fractionator. Hydrogen and other hydrocarbon gases are separated for recycle to the front-end of the process. Hydrogen sulfide and ammonia are also removed from the product gases.

The two-stage process is more often applied to feedstocks that might also include heavy oil (either neat or blended with gas oil) where the gas oil might also act as a hydrogen donor in addition to feedstock. After the products leave the reactor, they are sent to a fractionator. Hydrogen and other hydrocarbon gases are separated for recycle. Hydrogen sulfide and ammonia are also removed from the product gases.

The single-stage reactor now becomes the first stage reactor in which the preheated feedstock is mixed with the recycled hydrogen-rich gas (make up hydrogen may also be introduced at this time) and sent to the first-stage reactor, where the catalysts convert sulfur and nitrogen compounds to hydrogen sulfide and ammonia. At this stage, hydrocracking may be limited through choice.

After the treated feedstock leaves the first stage, it is sent to a hydrocarbon separator where the treated feedstock is separated and the hydrogen is recycled to the first-stage feedstock. The treated feedstock is charged to a fractionator and depending on the products desired (gasoline components, jet fuel, and gas oil), the fractionator is operated to selectively separate part of the first stage reactor effluent.

The preheated fractionator bottoms are then mixed with a hydrogen-rich gas and charged to the second-stage reactor. As this material has already been subjected to some hydrogenation, cracking, and reforming in the first stage, the operations of the second stage are more severe (higher temperatures and pressures). Like the first-stage product, the second stage product is separated from the hydrogen and charged to the fractionator Fig. 1.

The single-stage process can be used to produce gasoline but is more often used to produce middle distillate from heavy vacuum gas oils (VGO).[4,6] The two-stage process was developed primarily to produce high yields of gasoline from straight-run gas oil, and the first stage may actually be a purification step to remove sulfur-containing (as well as nitrogen-containing) organic materials. In terms of sulfur removal, it appears that nonasphaltenes sulfur may be removed before the more refractory asphaltene sulfur, thereby requiring thorough desulfurization. This is a good reason for processes to use an extinction-recycling technique to maximize desulfurization and the yields of the desired product. Significant conversion of heavy feedstocks can be accomplished by hydrocracking at high severity. For some applications, the products boiling up to 340°C (650°F) can be blended to give the desired final product.

Because hydrocracking units operate at high temperatures and pressures, control of both hydrocarbon leaks and hydrogen releases is important to prevent fires. In addition, caution is required to ensure that explosive concentrations of catalytic dust do not form during recharging. Also, vigilance is necessary to protect against plugged reactor beds. Furthermore, unloading the spent catalyst (that may have coke deposited on it) requires additional precautions to prevent fires, some of which are the result of spontaneous ignition. The spent catalyst is often cooled before discharge or it should be protected from oxygen by a nitrogen blanket until it is cooled.

There is a potential for exposure to hydrocarbon gas and vapor emissions, hydrogen, and hydrogen sulfide gas because of high-pressure leaks. Large quantities of carbon monoxide may be released during catalyst regeneration and changeover. Catalyst steam stripping and regeneration create waste streams containing sour water and ammonia.

Process Design

As described earlier, the most common form of hydrocracking process is a *two-stage* operation.[4–6] This flow scheme has been very popular in the many refineries where it is necessary to maximize the yield of transportation fuels and has the flexibility to produce gasoline, naphtha, jet fuel, or diesel fuel to meet seasonal swings in product demand.

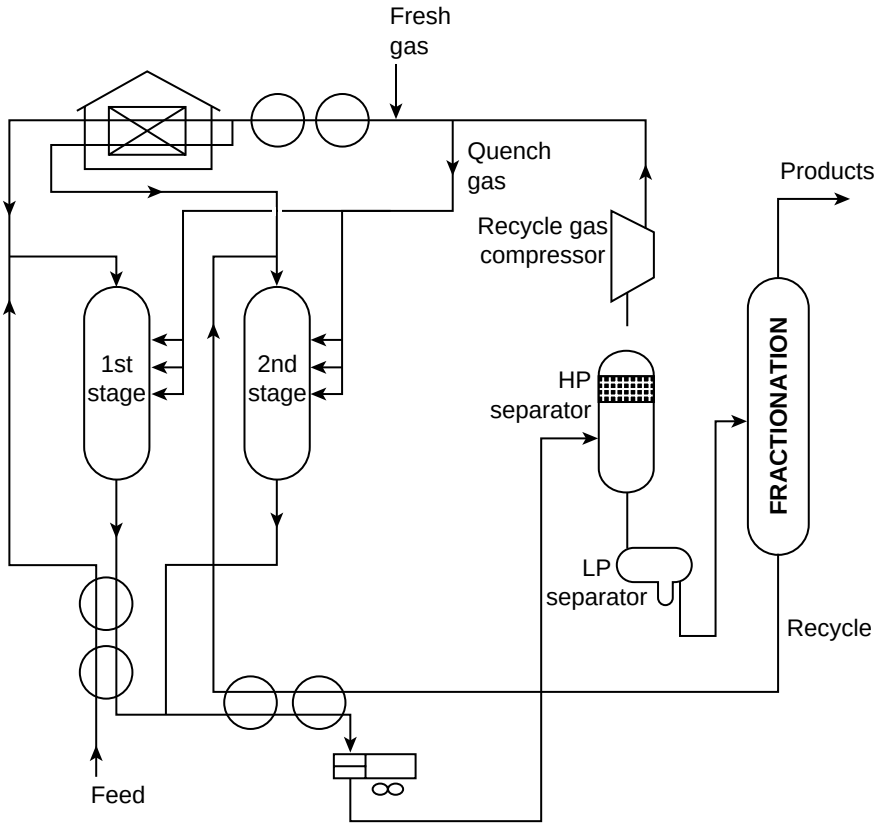


Fig. 1 A hydrocracking unit that can be operated as a single-stage or two-stage process.

This type of hydrocracker consists of two reactor stages together with a product distillation section. The choice of catalyst in each reaction stage depends on the product slate required and the character of the feedstock. In general, however, the first-stage catalyst is designed to remove nitrogen and heavy aromatics from raw petroleum stocks. The second-stage catalyst carries out a selective hydrocracking reaction on the cleaner oil produced in the first stage.

Both reactor stages have similar process flow schemes. The oil feed is combined with a preheated mixture of make up hydrogen and hydrogen-rich recycle gas and heated to reactor inlet temperature via a feed-effluent exchanger and a reactor charge heater. The reactor charge heater design philosophy is based on many years of safe operation with such two-phase furnaces. The feed-effluent exchangers take advantage of special high-pressure exchanger features that are designed to give leak-free end closures. From the charge heater, the partially vaporized feed enters the top of the reactor. The catalyst is loaded in separate beds in the reactor with facilities between the beds for quenching the reaction mix and ensuring good flow distribution through the catalyst.

The reactor effluent is cooled through a variety of heat exchangers including the feed-effluent exchanger and one or more air coolers. Deaerated condensate is injected into the first-stage reactor effluent before the final air cooler to remove ammonia and some of the hydrogen sulfide. This prevents solid ammonium bisulfide from depositing in the system. A body of expertise in the field of materials selection for hydrocracker cooling trains is quite important for proper design.

The reactor effluent leaving the air cooler is separated into hydrogen-rich recycle gas, a sour water stream, and a hydrocarbon liquid stream in the high-pressure separator. The sour water effluent stream is often sent to a plant for ammonia recovery and for purification so water can be recycled back to the hydrocracker. The hydrocarbon rich stream is pressure reduced and fed to the distillation section after light products are flashed off in a low-pressure separator.

The hydrogen-rich gas stream from the high-pressure separator is recycled back to the reactor feed by using a recycle compressor. Sometimes with sour feeds, the first-stage recycle gas is scrubbed with an amine system to remove hydrogen sulfide. If the feed sulfur level is high, this option can improve the performance of the catalyst and result in less costly materials of construction.

The distillation section consists of a hydrogen sulfide (H_2S) stripper and a recycle splitter. This latter column separates the product into the desired cuts. The column's bottom stream is recycled back to the second-stage feed. The recycle cut point is changed depending on the light products needed. It can be

as low as 160°C (320°F) if naphtha production is maximized (for aromatics) or as high as 380°C (720°F) if a low pour point diesel is needed. Between these two extremes, a recycle cut point of 260–285°C (500–550°F) results in high yields of high smoke point low freeze point jet fuel.

A *single-stage once-through* unit resembles the first stage of the two-stage plant. This type of hydrocracker usually requires the least capital investment. The feedstock is not completely converted to lighter products. For this application, the refiner must have a demand for highly refined heavy oil. In many refining situations, such an oil product can be used as lube oil plant feed or as fluid catalytic cracker feedstock or in low sulfur oil blends or as ethylene plant feed. It also lends itself to stepwise construction of a future two-stage hydrocracker for full feed conversion.

A *single-stage recycle* (SSREC) unit converts heavy oil completely into light products with a flow scheme resembling the second stage of the two-stage plant. Such a unit maximizes the yield of naphtha, jet fuel, or diesel depending on the recycle cut point used in the distillation section. This type of unit is more economical than the more complex two-stage unit when plant design capacity is less than about 10,000–15,000 bbl/day. Commercial SSREC units have operated to produce low pour point diesel fuel from waxy Middle Fast VGOs. Recent emphasis has been placed on upgrading lighter gas oils into jet fuels.

Building on the theme of *one-* or *two-stage* hydrocracking, the *once-through partial conversion* concept evolved. This concept offers the means to convert heavy VGO feed into high quality gasoline, jet fuel, and diesel products by a partial conversion operation. The advantage is lower initial capital investment and also lower utilities consumption than a plant designed for total conversion. Because total conversion of the higher molecular weight compounds in the feedstock is not required, once-through hydrocracking can be carried out at lower temperatures and in most cases at lower hydrogen partial pressures than in recycle hydrocracking, where total conversion of the feedstock is normally an objective.

Proper selection of the types of catalysts employed can even permit partial conversion of heavy gas oil feeds to diesel and lighter products at the low hydrogen partial pressures for which gas oil hydrotreaters are normally designed. This so-called *mild hydrocracking* has been attracting a great deal of interest from refiners who have existing hydrotreaters and wish to increase their refinery's conversion of fuel oil into lower boiling higher value products.

Recycle hydrocracking plants are designed to operate at hydrogen partial pressures from about 1200 psi to 2300 psi depending on the type of feed being processed. Hydrogen partial pressure is set in the design in part depending on required catalyst cycle length, and also to enable the catalyst to convert high molecular weight polynuclear aromatic and naphthene compounds that must be hydrogenated before they can be cracked. Hydrogen partial pressure also affects properties of the hydrocracked products that depend on hydrogen uptake, such as jet fuel aromatics content and smoke point and diesel cetane number. In general, the higher the feedstock endpoint, the higher the required hydrogen partial pressure necessary to achieve satisfactory performance of the plant.

Once-through, partial conversion hydrocracking of a given feedstock may be carried out at hydrogen partial pressures significantly lower than required for recycling and total conversion hydrocracking. The potential higher catalyst deactivation rates experienced at lower hydrogen partial pressures can be offset by using higher activity catalysts and designing the plant for lower catalyst space velocities. Catalyst deactivation is also reduced by the elimination of the recycle stream. The lower capital cost resulting from the reduction in plant operating pressure is much more significant than the increase resulting from the possible additional catalyst requirement and larger volume reactors.

Additional capital cost savings from once-through hydrocracking result from the reduced overall hydraulic capacity of the plant for a given fresh feed rate as a result of the elimination of a recycle oil stream. Hydraulic capacity at the same fresh feed rate is 30–40% lower for a once-through plant compared to the one designed for recycle.

Utilities savings for a once-through vs. recycle operation arise from lower pumping and compression costs as a result of the possible lower design pressure and also of lower hydrogen consumption. Additional savings are realized as a result of the lower oil and gas circulation rates required, because recycle of oil from the fractionator's bottom is not necessary.

Lower capital investment and operating costs are obvious advantages of once-through hydrocracking compared to a recycle design. This type of operation may be adaptable for use in an existing gas oil hydrotreater or atmospheric resid desulfurization plant. The change from hydrotreating to hydrocracking service will require some modifications and capital expenditure, but in most cases these changes will be minimal.

The fact that unconverted oil is produced by the plant is not necessarily a disadvantage. The unconverted oil produced by once-through hydrocracking is a high quality, low sulfur, and nitrogen material that is an excellent feed stock for an FCC unit or ethylene pyrolysis furnace or a source of high viscosity index lubricating oil base stock. The properties of the oil are a function of the degree of conversion and other plant operating conditions.

One disadvantage of once-through hydrocracking compared to a recycle operation is a somewhat reduced flexibility for varying the ratio of gasoline to middle distillate that is produced. A greater quantity of naphtha can be produced by increasing the conversion and jet fuel plus diesel yield can also be increased in this manner. But selectivity for higher boiling products is also a function of conversion. Selectivity decreases as once-through conversion increases. If conversion is increased too much, the yield of desired product will decrease, accompanied by an increase in light ends and gas production. Higher yields of gasoline or jet fuel plus diesel are possible from a recycle than from a once-through operation.

Middle distillate products made by once-through hydrocracking are generally higher in aromatic content of poorer burning quality than those produced by recycle hydrocracking. However, the quality is generally better when produced by catalytic cracking or from straight run. Middle distillate product quality improves as the degree of conversion increases and as hydrogen partial pressure is increased.

The hydrocracking process employs high-activity catalysts that produce a significant yield of light products. Catalyst selectivity for middle distillate is a function of both the conversion level and operating temperature, with values in excess of 90% being reported in commercial operation. In addition to the increased hydrocracking activity of the catalyst, percentage desulfurization and denitrogenation at start-of-run conditions are also substantially increased. End-of-cycle is reached when product sulfur has risen to the level achieved in conventional VGO hydrodesulfurization process.

An important consideration, however, is that commercial hydrocracking units are often limited by design constraints of existing VGO hydrotreating units. Thus, the proper choice of catalyst(s) is critical when searching for optimum performance. Typical commercial distillate hydrocracking catalysts contain both the hydrogenation (metal) and cracking (acid sites) functions required for service in existing desulfurization units.

Commercial Processes

The following processes are a sampling of those processes that are, or have been, in commercial practice.[4,6] This list is by no means exhaustive.

Gulf HDS Process

This is a regenerative fixed-bed process to upgrade residua by catalytic hydrogenation of refined heavy fuel oils or high quality catalytic charge stocks. Desulfurization and quality improvement are the primary purposes of the process, but if the operating conditions and catalysts are varied, light distillates can be produced and the viscosity of heavy material can be lowered. Long on-stream cycles are maintained by reducing random hydrocracking reactions to a minimum, and whole crude oils, virgin, or cracked residua may serve as feedstock. This process is suitable for the desulfurization of high-sulfur residua (atmospheric and vacuum), to produce low-sulfur fuel oils, or catalytic cracking feedstocks. In addition, the process can be used, through alternate design types, to upgrade high-sulfur crude oils or bitumen that are unsuited for the more conventional refining techniques.

H-Oil Process

The H-Oil process is a catalytic hydrogenation technique that uses a one-, two-, or three–stage ebullated-bed reactor in which considerable hydrocracking takes place during the reaction. The process is used to upgrade heavy sulfur-containing crude oils, residual stocks, and low-sulfur distillates, thereby reducing fuel oil yield.

IFP Hydrocracking Process

The process features a dual catalyst system: The first catalyst is a promoted nickel–molybdenum amorphous catalyst. It acts to remove sulfur and nitrogen and hydrogenate aromatic rings. The second catalyst is a zeolite that finishes the hydrogenation and promotes the hydrocracking reaction.

Isocracking Process

The process has been applied commercially in the full range of process flow schemes: single-stage, once-through liquid; single-stage, partial recycle of heavy oil; single-stage extinction recycle of oil (100% conversion); and two-stage extinction recycle of oil. The preferred flow scheme will depend on the feed properties, the processing objectives, and to some extent, the specified feed rate.

LC-Fining Process

The LC-Fining process is a hydrocracking process capable of desulfurizing, demetallizing, and upgrading a wide spectrum of heavy feedstocks by means of an expanded bed reactor. Operating with the expanded bed allows the processing of heavy feedstocks, such as atmospheric residua, vacuum residua, and oil sand bitumen.

Microcat-RC Process

The Microcat-RC process (also referred to as the M-Coke process) is a catalytic hydroconversion process operating at relatively moderate pressures and temperatures. The catalyst particles, containing a metal sulfide in a carbonaceous matrix formed within the process are uniformly dispersed throughout the feed. Because of their ultra small size (10^{-4} in. diameter), there are typically several more orders of magnitude of these microcatalyst particles per cubic centimeter of oil than is possible in other types of hydroconversion reactors using conventional catalyst particles. These results in smaller distances between particles and less time for a reactant molecule or intermediate to find an active catalyst site. Because of their physical structure, microcatalysts suffer none of the pore-plugging problems that plague conventional catalysts.

Mild Hydrocracking Process

The *mild hydrocracking process* uses operating conditions similar to those of a VGO desulfurizer to convert VGO to yield significant lighter products. Consequently, the flow scheme for a mild hydrocracking unit is virtually identical to that of a VGO desulfurizer.

Mild Resid Hydrocracking (MRH) Process

The mild resid hydrocracking process is a hydrocracking process designed to upgrade heavy feedstocks containing large amount of metals and asphaltene, such as vacuum residua and bitumen and to mainly produce middle distillates. The reactor is designed to maintain a mixed three-phase slurry of feedstock, fine powder catalyst and hydrogen, and to promote effective contact.

Residfining Process

Residfining is a catalytic fixed-bed process for the desulfurization and demetallization of residua. The process can also be used to pretreat residua to suitably low contaminant levels prior to catalytic cracking.

Unicracking Process

This is a fixed-bed catalytic process that employs a high-activity catalyst with a high tolerance for sulfur and nitrogen compounds and can be regenerated. The design is based upon a single-stage or a two-stage system with provisions to recycle up to extinction.

Veba Combi-Cracking (VCC) Process

The VCC process is a thermal hydrocracking/hydrogenation process for converting residua and other heavy feedstocks. The process is based on the Bergius–Pier technology that was used for coal hydrogenation in Germany until 1945. The heavy feedstock is hydrogenated (hydrocracked) using a commercial catalyst and liquid-phase hydrogenation reactor operating at 440–485°C (825–905°F) and 2175–4350 psi pressure. The product obtained from the reactor is fed into the hot separator operating at temperatures slightly below the reactor temperature. The liquid and solid materials are fed into a vacuum distillation column, the gaseous products are fed into gas-phase hydrogenation reactor operating at an identical pressure. This high temperature, high pressure coupling of the reactor products with further hydrogenation provides a specific process economics.

Catalysts

Hydrocracking catalysts typically contain separate hydrogenation and cracking functions. Palladium sulfide and promoted group VI sulfides (nickel molybdenum or nickel tungsten) provide the hydrogenation function. These active compositions saturate aromatics in the feed, saturate olefins formed in the cracking, and protect the catalysts from poisoning by coke. Zeolites or amorphous silica-alumina provide the cracking functions. The zeolites are usually type Y (faujasite), ion exchanged to replace sodium with hydrogen and make up 25–50% of the catalysts. Pentasils (silicalite or ZSM-5) may be included as dewaxing catalysts.

Hydrocracking catalysts, such as nickel (5% by weight) on silica-alumina, work best on feedstocks that have been hydrofined to low nitrogen and sulfur levels. The nickel catalyst then operates well at 350–370°C (660–700°F) and at a pressure of about 1500 psi to give good conversion of feed to lower boiling liquid fractions with minimum saturation of single-ring aromatics and a high iso-paraffin to *n*-paraffin ratio in the lower molecular weight paraffins.

Catalyst operating temperature can influence reaction selectivity because the activation energy for hydrotreating reactions is much lower than that of hydrocracking reactions. Therefore, raising the temperature in a residuum hydrotreater increases the extent of hydrocracking relative to hydrotreating, which also increases the hydrogen consumption.

Clays have been used as cracking catalysts particularly for heavy feedstocks and have also been explored in the demetallization and upgrading of heavy crude oil. The results indicated that the prepared catalyst

was mainly active toward demetallization and conversion of the heaviest fractions of crude oils.

Zeolite catalysts have also found use in the refining industry during the last two decades. Like the silica-alumina catalysts, zeolites also consist of a tetrahedral framework usually with a silicon or an aluminum atom at the center. The geometric characteristics of the zeolites are responsible for their special properties, which are particularly attractive to the refining industry. Specific zeolite catalysts have shown up to 10,000 times more activity than the so-called conventional catalysts in specific cracking tests. The mordenite-type catalysts are particularly worth mentioning because they have shown up to 200 times greater activity for hexane cracking in the temperature range 360–400°C (680–750°F).

Other zeolite catalysts have also shown remarkable adaptability to the refining industry. For example, the resistance to deactivation of the type Y zeolite catalysts containing either noble or nonnoble metals is remarkable, and catalyst life of up to 7 years has been obtained commercially in processing heavy gas oils in the Unicracking-JHC processes. Operating life depends on the nature of the feedstock, the severity of the operation, and the nature and extent of operational upsets. Gradual catalyst deactivation in commercial use is counteracted by incrementally raising the operating temperature to maintain the required conversion per pass. The more active a catalyst is, the lower the temperature required. When processing for gasoline, lower operating temperatures have the additional advantage—less of the feedstock is converted to iso-butane.

Zeolite catalysts provide the cracking function in many hydrocracking catalysts, as they do in fluid catalytic cracking catalysts. Zeolites are crystalline aluminosilicates, and among all commercial catalysts, the zeolite used at present is faujasite. Pentasil zeolites, including silicalite and ZSM-5 are also used for their ability to crack long chain paraffins selectively.

Typical levels are 25–50%, the weight of zeolite in the catalysts, with the remainder being the hydrogenation component and a silica (SiO_2) or alumina (Al_2O_3) binder. Exact recipes are guarded as trade secrets.

While zeolite catalysts provided a breakthrough that allowed catalytic hydrocracking to become commercially important, continued advances in the manufacture of amorphous silica alumina made these materials competitive in certain kinds of applications.

Typical catalysts of this type contain 60–80% of silica alumina, with the remainder being the hydrogenation component. The compositions of these catalysts are closely held secrets. Over the years, broad ranges of silica/alumina molar ratios have been used in various cracking applications, but silica is almost always in excess for high acidity and stability. A typical level might be 25% alumina (Al_2O_3).

Amorphous silica-alumina is made by a variety of precipitation techniques. The whole class of materials traces its beginning to silica gel technology, in which sodium silicate is acidified to precipitate the hydrous silica-alumina sulfate; sulfuric acid is used wholly or partly for this precipitation, and a mixed gel is formed. The properties of this gel, including acidity and porosity, can be varied by changing the recipe-concentrations, order of addition, pH, temperature, aging time, and the like. The gels are isolated by filtration and washed to remove sodium and other ions.

Careful control of the precipitation allows the pore size distributions of amorphous materials to be controlled rather well, but the distributions are still much broader than those in the zeolite catalysts. This limits the activity and selectivity. One effect of the reduced activity has been that these materials have been applied only in making middle distillates: diesel and turbine fuels. At higher process severities, the poor selectivity results in production of unacceptable amounts of methane (CH_4) to butane (C_4H_{10}) hydrocarbons.

Hydrocarbons, especially aromatic hydrocarbons, can react in the presence of strong acids to form coke. This coke is a complex polynuclear aromatic material that is low in hydrogen. Coke can deposit on the surface of a catalyst, blocking access to the active sites, and reducing the activity of the catalyst. Coke poisoning is a major problem in fluid catalytic cracking catalysts, where coked catalysts are circulated to a fluidized bed combustor to be regenerated. In hydrocracking, coke deposition is virtually eliminated by the catalyst's hydrogenation function.

In a well-designed hydrocracking system, the hydrogenation function adds hydrogen to the tarry deposits. This reduces the concentration of coke precursors on the surface. There is, however, a slow accumulation of coke that reduces activity over a 1–2 year period. Refiners respond to this slow reduction in activity by raising the average temperature of the catalyst bed to maintain conversions. Eventually, however, an upper limit to the allowable temperature is reached and the catalyst must be removed and regenerated.

Catalysts carrying coke deposits can be regenerated by burning off the accumulated coke. This is done by service in rotary or similar kilns rather than leaving catalysts in the hydrocracking reactor, where the reactions could damage the metals in the walls. Removing the catalysts also allows inspection and repair of the complex and expensive reactor internals. Regeneration of a large catalyst charge can take weeks or months, so refiners may own two catalyst loads, one in the reactor, one regenerated and ready for reload.

The thermal reactions also convert the metal sulfide hydrogenation functions to oxides and may result in agglomeration. Excellent progress has been made since the 1970s in regenerating hydrocracking

catalysts; similar regeneration of hydrotreating catalysts is widely practiced.

After combustion to remove the carbonaceous deposits, the catalysts are treated to disperse active metals. Vendor documents claim more than 95% recovery of activity and selectivity in these regenerations. Catalysts can undergo successive cycles of use and regeneration, providing long functional life with these expensive materials.

As illustrated earlier, for various forms of more conventional hydrocracking, the type of catalyst used can influence the product slate obtained. For example, for a mild hydrocracking operation at constant temperature, the selectivity of the catalyst varies from about 65% to about 90% by volume. Indeed, several catalytic systems have now been developed with a group of catalysts, specifically, for mild hydrocracking operations. Depending on the type of catalyst, they may be run as a single catalyst or in conjunction with a hydrotreating catalyst.

CONCLUSIONS

The purposes of hydroprocessing are (1) to improve existing petroleum products or develop new products or even new uses; (2) to convert inferior or low-grade materials into valuable products; and (3) to transform near-solid residua into liquid fuels. The distinguishing feature of the hydrogenating processes is that, although the composition of the feedstock is relatively unknown and a variety of reactions may occur simultaneously, the final product may actually meet all the required specifications for its particular use.

Hydrocracking is similar to catalytic cracking, with hydrogenation superimposed and with the reactions taking place either simultaneously or sequentially. Hydrocracking was initially used to upgrade low-value distillate feedstocks, such as cycle oils (highly aromatic products from a catalytic cracker that usually are not recycled to extinction for economic reasons), thermal and coker gas oils, and heavy-cracked and straight-run naphtha. These feedstocks are difficult to process either by catalytic cracking or reforming, because they are usually characterized by a high polycyclic aromatic content and/or by high concentrations of the two principal catalyst poisons, sulfur and nitrogen compounds.

Hydrocracking allows refiners the potential to balance fuel oil supply and demand by adding VGO cracking capacity. Situations where this is the case include: (1) refineries with no existing VGO cracking capacity; (2) refineries with more VGO available than VGO conversion capacity; (3) refineries where addition of vacuum residuum conversion capacity has resulted in production of additional cracking feedstocks boiling in the VGO range (e.g., coker gas oil); and (4) refineries that have one of the two types of VGO conversion units but could benefit from adding the second type. In some cases, a refiner might add both gas oil cracking and residuum conversion capacity simultaneously.

Those refiners who do choose gas oil cracking as part of their strategy for balancing residual fuel oil supply and demand must decide whether to select a hydrocracking unit or a fluid catalytic cracking unit. Although the two processes have been compared vigorously over the years, neither of the processes has evolved to be the universal choice for gas oil cracking. Both processes have their advantages and disadvantages, and process selection can be properly made only after careful consideration of many case-specific factors. Among the most important factors are: (1) product slate required; (2) amount of flexibility required to vary the product slate; (3) product quality (specifications) required; and (4) the need to integrate the new facilities in a logical and cost-effective way with any existing facilities.

REFERENCES

1. Gray, M.R. *Upgrading Petroleum Residues and Heavy Oils*; Marcel Dekker Inc.: New York, 1994.
2. Meyer, R.F., Ed.; *Heavy Crude and Tar Sands—Hydrocarbons for the 21st Century*; Petróleos de Venezuela S.A.: Caracas, Venezuela, 1991.
3. Speight, J.G. *Fuel Science and Technology Handbook*; Speight, J.G., Ed.; Marcel Dekker Inc.: New York, 1990; Part II, Chapters 12–16.
4. Speight, J.G. *The Chemistry and Technology of Petroleum*, 3rd Ed.; Marcel Dekker Inc.: New York, 1999.
5. Speight, J.G. *The Desulfurization of Heavy Oils and Residua*, 2nd Ed.; Marcel Dekker Inc.: New York, 1999.
6. Speight, J.G.; Ozum, B. *Petroleum Refining Processes*; Marcel Dekker Inc.: New York, 2002.

Hydrodesulfurization

James G. Speight
CD & W Inc., Laramie, Wyoming, U.S.A.

INTRODUCTION

The hydrodesulfurization process is a specific hydrogenation process and, as employed in petroleum refining, can be classified as nondestructive or destructive.

Nondestructive, or simple, hydrogenation is generally used for the purpose of improving product quality without an appreciable alteration of the boiling range. Mild processing conditions are employed so that only the more unstable materials are attacked and the sulfur, nitrogen, and oxygen compounds undergo hydrogenolysis to split out hydrogen sulfide (H_2S), ammonia (NH_3), and water (H_2O), respectively. An example is the hydrodesulfurization of naphtha with temperatures in the range of 330–370°C (625–695°F) and hydrogen partial pressures of 100–500 psi.

On the other hand, destructive hydrogenation (hydrocracking) is characterized by the cleavage of carbon–carbon bonds with concurrent addition of hydrogen to the fragments to produce saturated lower-boiling products. Such treatment requires severe reaction temperatures (usually of the order of 360–410°C, 680–770°F) as well as high hydrogen pressures to minimize reactions that lead to the formation of undesirable products such as coke.

HYDRODESULFURIZATION CHEMISTRY

Kinetic studies using individual sulfur compounds have usually indicated that simple first-order kinetics with respect to sulfur is the predominant mechanism by which sulfur is removed from the organic material as hydrogen sulfide. However, there is still much to be learned about the relative rates of reaction exhibited by the various compounds present in petroleum (see Ref.[1] and references cited therein).

The reactions involving the hydrogenolysis of sulfur compounds encountered in hydroprocessing are exothermic and thermodynamically complete under ordinary operating conditions. The various molecules have very different reactivity, with mercaptan sulfur much easier to eliminate than thiophene sulfur or dibenzothiophene sulfur.

The structural differences between the various sulfur-containing molecules make it impractical to have a single rate expression applicable to all reactions in hydrodesulfurization. Each sulfur-containing molecule has its own hydrogenolysis kinetics, which is usually complex because several successive equilibrium stages are involved and these are often controlled by internal diffusion limitations during refining.

The development of general kinetic data for the hydrodesulfurization of different feedstocks is complicated by the presence of a large number of sulfur compounds each of which may react at a different rate because of structural differences as well as differences in molecular weight. This may be reflected in the appearance of a complicated kinetic picture for hydrodesulfurization in which the kinetics is not, apparently, first order. The overall desulfurization reaction may be satisfied by a second-order kinetic expression, when it can, in fact, also be considered as two competing first-order reactions. These reactions are: 1) the removal of nonasphaltene sulfur and 2) the removal of asphaltene sulfur. It is the sum of these reactions that gives the second-order kinetic relationship.

In addition, the sulfur compounds in a feedstock may cause changes in the catalyst upon contact and, therefore, every effort should be made to ensure that the kinetic data from such investigations are derived under standard conditions. In this sense, several attempts have been made to accomplish standardization of the reaction conditions by presulfiding the catalyst, passage of the feedstock over the catalyst until the catalyst is stabilized, obtaining the data at various conditions, and then rechecking the initial data by repetition.

Thus, it has become possible to define certain general trends that occur in the hydrodesulfurization of petroleum feedstocks. One of the more noticeable facets of the hydrodesulfurization process is that the rate of reaction declines markedly with the molecular weight of the feedstock. For example, examination of the thiophene portion of a (narrow-boiling) feedstock and the resulting desulfurized product provides excellent evidence that benzothiophenes are removed in preference to the dibenzothiophenes and other condensed thiophenes. The sulfur compounds in heavy oils and residua are presumed to react (preferentially) in a similar manner.

It is also generally accepted that the simpler sulfur compounds (e.g., thiols, R-SH, and thioethers, R-S-R^1) are (unless steric influences offer resistance to the

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008050

hydrodesulfurization) easier to remove from petroleum feedstocks than the more complex cyclic sulfur compounds such as the benzothiophenes. It should be noted here that, because of the nature of the reaction, steric influences would be anticipated to play a lesser role in the hydrocracking process.

Residua hydrodesulfurization is considerably more complex than the hydrodesulfurization of model organic sulfur compounds or, for that matter, narrow-boiling petroleum fractions. In published studies of the kinetics of residua hydrodesulfurization, one of three approaches has generally been taken:

1. The reactions can be described in terms of simple first-order expressions.
2. The reactions can be described by the use of two simultaneous first-order expressions—one expression for easy-to-remove sulfur and a separate expression for difficult-to-remove sulfur.
3. The reactions can be described using a pseudo-second-order treatment.

Each of the three approaches has been used to describe hydrodesulfurization of residua under a variety of conditions with varying degrees of success, but it does appear that pseudo-second-order kinetics is favored. In this particular treatment, the rate of hydrodesulfurization is expressed by a simple second-order equation:

$$C/(1 - C) = k(1/\mathrm{LHSV})$$

where C is the wt% of sulfur in product/wt% of sulfur in the charge, k is the reaction rate constant, and LHSV is the liquid hourly space velocity (volume of liquid feed per hour per volume of catalyst).

Application of this model to a residuum desulfurization gave a linear relationship. However, it is difficult to accept the desulfurization reaction as a reaction that requires the interaction of two sulfur-containing molecules (as dictated by the second-order kinetics). To accommodate this anomaly, it has been suggested that, as there are many different types of sulfur compounds in residua and each may react at a different rate, the differences in reaction rates offered a reasonable explanation for the apparent second-order behavior. For example, an investigation of the hydrodesulfurization of an Arabian light-atmospheric residuum showed that the overall reaction could not be adequately represented by a first-order relationship. However, the reaction could be represented as the sum of two competing first-order reactions and the rates of desulfurization of the two fractions (the oil fraction and the asphaltene fraction) could be well represented as an overall second-order reaction.

Other kinetic work has shown that, for a fixed level of sulfur removal, the order of a reaction at constant temperature can be defined with respect to pressure:

$$k = 1/\mathrm{LHSV}(P_\mathrm{h})^n$$

where P_h is the hydrogen partial pressure, LHSV is the liquid volume hourly space velocity, k is a constant, and n is the order of the reaction. It has been concluded, on the basis of this equation, that the hydrodesulfurization of residuum is first order with respect to pressure over the range 800–2300 psi, although it does appear that the response to pressure diminishes markedly (and may even be minimal) above 1000 psi.

One marked effect of a hydrodesulfurization process is the buildup of hydrogen sulfide and the continued presence of this reaction product in the reactor reduces the rate of hydrodesulfurization. Thus, using the two first-order models, the effect of hydrogen sulfide on the process can be represented as:

$$k/k_0 = 1/(1 + k_1 P_{\mathrm{H_2S}})$$

where k is the rate constant with hydrogen sulfide present, k_0 is the rate constant in the absence of hydrogen sulfide, and k_1 is a constant.

Data obtained using this equation showed that a change in hydrogen sulfide concentration from 1% to 12% (by volume) could reduce by 50% the rate constants for the easy-to-desulfurize and the difficult-to-desulfurize reactions. On the basis of the data available from kinetic investigations, the kinetics of residuum hydrodesulfurization may be represented by the following general equation:

$$-\mathrm{d}s/\mathrm{d}t = \left[P_\mathrm{H}^n/(1 + k_\mathrm{a}A + k_\mathrm{s}P_{\mathrm{H_2S}})^m\right]k_i S_i$$

where s is the weight fraction of sulfur in the liquid phase, t is the residence time, P_H is the partial pressure of hydrogen, A is the weight fraction of asphaltenes in the liquid phase, $P_{\mathrm{H_2S}}$ is the partial pressure of hydrogen sulfide, S_i is the weight fraction of sulfur associated with component i in the range of i–j, k_a is the adsorption constant for asphaltenes, k_s is the adsorption constant for hydrogen sulfide, and k_i is the specific reaction rate constant for component i.

The latter constant, k_i, in the above relationship is a function of the chemistry of the component, the catalyst activity, and the reaction temperature. Therefore,

$$k_i = k_0(A/A_0)\mathrm{e}^{E/RT}$$

where k_0 is the reaction rate constant at standard catalyst activity, A_0 is the standard catalyst activity, A is

the catalyst activity, ΔE is the activation energy, R is the gas constant, and T is the absolute temperature.

This relationship gives activation energies for the hydrodesulfurization of various residua in the range 27–35 kcal/mol. In this context, it was interesting to note that the deasphalting of Khafji residuum had no effect on the activation energy of 30 kcal/mol and it was suggested that the activation energies of the various components in a particular residuum might be approximately the same.

THE HYDRODESULFURIZATION PROCESS

The hydrodesulfurization process operates using high hydrogen pressure, typically 1500–500 psi, and temperatures range from 290°C to 370°C (550–700°F). Several process configurations are used, depending on the feed and the design criteria.[1,2]

Process Description

The hydrodesulfurization process is essentially the reaction of hydrogen with a predominantly hydrocarbon feedstock to produce a desulfurized hydrocarbon product and hydrogen sulfide.[2,3] In the process, the feedstock is first pressurized to a pressure that is a little higher than that of the reactor section, mixed with hot recycle gas, and preheated to the temperature of the reactor inlet. The hot feedstock (and the recycle gas) is then introduced to the catalyst in the reactor where temperatures ranging from 290°C to 455°C (550–850°F) and pressures in the range 150–3000 psi prevail.

The use of a recycle gas technique in the hydrodesulfurization process minimizes losses of hydrogen—a very expensive commodity in petroleum refining. The hydrodesulfurization process requires high partial pressures of hydrogen to promote high desulfurization reaction rates and to diminish coke (carbon) deposition on the catalyst. To maintain high partial pressures of hydrogen, it is necessary to introduce hydrogen into the reactor at several times the rate of hydrogen consumption, and it is possible to recover most of the unused hydrogen in the separator after which it is then recycled for further use. Indeed, if it were not possible to recover most of the unused hydrogen, the process economics would probably be very questionable. However, tolerable amounts of hydrogen may be lost as a result of the solubility of the gas in the liquid hydrocarbon product or even during the removal of the hydrogen sulfide (and the light hydrocarbon gases) from the recycle gas.

Hydrogen requirements for the hydrodesulfurization process depend on the nature of the feedstock as well as on the extent of the desulfurization.[1,4] For example, the heavier feedstocks require substantially more hydrogen to produce a given product than the lighter feedstocks, or to yield a product with a predetermined amount of sulfur. The theoretical hydrogen requirements often differ markedly from the experimental values because of the nitrogen and oxygen contents of the feedstock. The metals content of the feedstock may also influence the hydrogen consumption by altering the characteristics of the catalyst. It is conceivable that metals deposition onto the catalyst may serve to increase the hydrocracking activity of the catalyst, thereby promoting higher consumption of hydrogen in the process.

Reactor Design

Reactor designs for hydrodesulfurization of various feedstocks vary in the way in which the feedstock is introduced into the reactor and in the arrangement, as well as the physical nature, of the catalyst bed. The conditions under which the hydrodesulfurization process operates (i.e., high temperatures and high pressures) dictate the required wall thickness (determined by the pressure/temperature/strength ratio). In addition, resistance of the reactor walls to corrosive attack by hydrogen sulfide and hydrogen (to name only two of the potential corrosive agents of all of the constituents in, or arising from, the feedstock) can be a problem. Precautions should be taken to ensure that wall thickness and composition yield maximum use and safety.

Reactors may vary from as little as 4 ft in diameter to as much as 20 ft in diameter and have a wall thickness anywhere from 4.5 to 10 in. or so.[3] These vessels may weigh from 150 tons to as much as 1000 tons. Obviously, before selecting a suitable reactor, shipping and handling requirements (in addition to the more conventional process economics) must be given serious consideration.

Once-through reactors, where the incompletely converted or unconverted fraction of the feedstock is separated from the lower-boiling products, are now being replaced by recycle reactors. In these reactors, any unconverted feedstock is sent back (recycled) to the reactor for further processing. In such a case, the volume flow of the combined (fresh and unconverted) feedstock is the sum of the inputs of the fresh feedstock and recycled feedstock:

$$F_T = F_F + F_R$$

where F_T is the total feedstock in the unit, F_F is the fresh feedstock, and F_R is the recycled feedstock. Thus, the recycle ratio, R, is the ratio of the recycled

feedstock to the fresh feedstock:

$$R = F_R/F_F$$

The recycle ratio may also be expressed as a percentage in some texts or references. However, the recycled feedstock, having been through the reactor at least once, will have a lower reactivity than the original feedstock, thereby reducing the reactivity of the feedstock in each recycle event.

The downflow fixed-bed reactor has been used widely for hydrodesulfurization processes and is so called because of the feedstock entry at the top of the reactor while the product stream is discharged from the base of the reactor. The catalyst is contained in the reactor as stationary beds with the feedstock and hydrogen passing through the bed in a downward direction. The exothermic nature of the reaction and the subsequent marked temperature rise from the inlet to the outlet of each catalyst bed require that the reaction mix be quenched by cold recycle gas at various points in the reactor, hence the incorporation of separate catalyst beds as part of the reactor design.

To combat the inevitable loss in desulfurizing activity of the catalyst that must be presumed to occur with time under any predetermined set of reaction conditions, the bed inlet temperature may be increased slowly, thereby increasing the overall temperature of the catalyst bed and so maintaining constant catalyst activity. Thus, depending on the nature of the feedstock, there may be a considerable difference between the start-of-run temperature and the end-of-run temperature.

In fixed-bed reactors, the catalyst may be poisoned (deactivated) progressively. For example, the first catalyst bed will most likely be poisoned by vanadium and nickel deposition initially at the inlet to the bed and then progressively through the bed as the active zone is gradually pushed downward into the bed. Once catalyst poisoning has progressed through the bed, the catalyst may have to be discarded or regenerated. However, catalyst regeneration may only suffice for a limited period of time and the catalyst may have to be completely replaced. This is especially true in the case of feedstocks such as residua and heavy oils.

Removal of the metal contaminants is not usually economical, or efficient, during rapid regeneration. In fact, the deposited metals are believed to form sulfates during removal of carbon and sulfur compounds by combustion, which produce a permanent poisoning effect. Thus, if fixed-bed reactors are to be used for residuum or heavy-oil hydrodesulfurization (in place of the more usual distillate hydrodesulfurization), it may be necessary to first process the heavier feedstocks to remove the metals (especially vanadium and nickel) and so decrease the extent of catalyst bed plugging. Precautions should also be taken to ensure that plugging of the bed does not lead to the formation of channels within the catalyst bed, which will also reduce the efficiency of the process and may even lead to pressure variances within the reactor because of the distorted flow patterns with eventual damage.

The radial-flow fixed-bed hydrodesulfurization reactor is a variation of the downflow fixed-bed reactor. Again, the feedstock enters the top of the reactor but instead of flowing downward through the catalyst bed, the feedstock is encouraged to flow through the bed in a radial direction and then out through the base of the reactor. There are certain advantages of this type of reactor, not the least of which is a low pressure drop through the catalyst bed; in addition, a radial bed reactor has a larger catalyst cross-sectional area as well as a shorter bed depth than the corresponding downflow reactor. It is this latter property that gives rise to the smaller pressure drop across the catalyst bed.

However, there are more chances of localized heating in the catalyst bed and (in addition to the more expensive reactor design per unit volume of catalyst bed) it may be more difficult to remove contaminants from the bed as part of the catalyst regeneration sequence. For this reason alone, it is preferable that this type of reactor is limited to hydrodesulfurization of low-boiling feedstocks such as naphtha and kerosene and application to the higher-boiling heavy oils and residua is usually not recommended.

In summary, fixed-bed processes have advantages in ease of scale-up and operation. The reactors operate in a downflow mode, with liquid feed trickling downward over the solid catalyst concurrent with the hydrogen gas. The usual catalyst is cobalt/molybdenum (Co/Mo) or nickel/molybdenum (Ni/Mo) on alumina (Al_2O_3) and contains 11–14% molybdenum and 2–3% of the promoter nickel or cobalt. The alumina typically has a pore volume of 0.5 ml/g. The catalyst is formed into pellets by extrusion, in shapes such as cylinders (ca. 2 mm in diameter), lobed cylinders, or rings.

These trickle bed reactors normally operate in the downflow configuration and have a number of operational problems, including poor distribution of liquid and pulsing operation at high liquid and gas loading. Scale-up of these liquid–gas–solid reactors is much more difficult than for a gas–solid or gas–liquid reactor. Nevertheless, the downflow system is convenient when the bed is filled with small catalyst particles. And, because the catalyst particles are free from diffusional limitations, these reactors are quite effective as filters of the incoming feed. Any suspended fine solids, such as fine clays from production operations, accumulate at the front end of the bed. Eventually, this will lead to high-pressure differentials between the inlet and outlet end of the reactor.

The main limitation of this type of reactor is the gradual accumulation of metals when heavy feedstocks are processed. The metals accumulate in the pores of the catalyst and gradually block access for hydrogenation and desulfurization. The length of operation is then dictated by the metal-holding capacity of the catalyst and the nickel and vanadium content of the feed. As the catalyst deactivates, the reactor feed temperature is gradually increased to maintain conversion. Toward the end of a run this mode of operation leads to accumulation of carbonaceous deposits on the catalyst, further reducing the activity.

The upflow expanded-bed reactor operates in such a way that the catalyst remains loosely packed and is less susceptible to plugging and therefore more suitable for the heavier feedstocks as well as for feedstocks that may contain considerable amounts of suspended solid material. Because of the nature of the catalyst bed, such suspended material will pass through the bed without causing frequent plugging problems. Furthermore, the expanded state of motion of the catalyst allows frequent withdrawal from, or addition to, the catalyst bed during operation of the reactor without the necessity of shutdown of the unit for catalyst replacement. This property alone makes the ebullated reactor ideally suited for the high-metal feedstocks (i.e., residua and heavy oils) that rapidly poison a catalyst with the ever-present catalyst replacement issues.

In the reactor, the feedstock and hydrogen are caused to flow upward through the catalyst bed in which each catalyst particle is reputed to have independent motion and can therefore (in theory) migrate throughout the entire catalyst bed. The heat of the reaction (which could be a problem in other reactors if it were not controlled) is dissipated in raising the temperature of fresh feedstock to that existing in the reactor. Spent catalyst can be withdrawn from the reactor and replaced with fresh catalyst on a daily basis if necessary and the need for time wasting and costly shutdowns can thus be eliminated.

Expanded-bed reactors normally operate with 1/32 in. extruded catalysts that require partial recycle of the liquid products to maintain the catalyst particles in an expanded and fluidized state.

The nearly completely random motion of the catalyst bed virtually ensures an isothermal operation, but the efficiency of the hydrodesulfurization reaction tends to suffer because of the back mixing of the product and feedstock. Hence, to effect sulfur removal at over 75% efficiency, it may be necessary to operate with two or more reactors in series. The need for two or more of these units to effectively desulfurize a feedstock may be cited as a disadvantage of the reactor, but the ability of the reactor to operate under isothermal conditions as well as the onstream catalyst addition–withdrawal system and the fact that the reactor size required for an expanded catalyst bed is often smaller than that required for a fixed bed can be cited in support of such a unit.

It is also possible to use finely divided catalyst in the expanded-bed reactor. If the catalyst is a suitable size (50–200 μm, i.e., 50×10^{-4} cm to 200×10^{-4} cm), it is possible to operate the expanded-bed reactor without recycling the liquid products to maintain the catalyst in the fluidized state. In addition, the finely divided catalyst has a relatively larger number of external pores on the surface than the extruded catalyst and is less likely to have metal contaminants plugging up these pores because of their size. The overall effect of a finely divided catalyst in this reactor is more efficient sulfur removal for a given set of process conditions.

CATALYSTS

Many of the catalysts for the hydrodesulfurization process are produced by composting a transition metal (or its salt) with a solid support.[4] The metal constituent is the active catalyst. The most commonly used materials for supports are alumina, silica, silica–alumina, kieselguhr, magnesia (and other metal oxides), as well as the zeolites. The support can be manufactured in a variety of shapes or may even be crushed to particles of the desired size. The metal constituent can then be added by contact of the support with an aqueous solution of the metal salt. The whole is then subjected to further treatment that will dictate the final form of the metal on the support (i.e., the metal oxide, the metal sulfide, or even the metal itself).

Molybdenum sulfide (MoS_2), usually supported on tailored alumina, is widely used in petroleum processes for hydrogenation reactions. It is a layered structure that can be made much more active by addition of cobalt or nickel. When promoted with cobalt sulfide (CoS), making what is called cobalt–molybdenum catalysts, it is widely used in hydrodesulfurization processes. The nickel sulfide (NiS) promoted version is used for hydrodenitrogenation as well as hydrodesulfurization. The closely related tungsten compound (WS_2) is used in commercial hydrocracking catalysts. Other sulfides [iron sulfide (FeS), chromium sulfide (Cr_2S_3), and vanadium sulfide (V_2S_5)] are also effective and are used in some catalysts. A valuable alternative to the base metal sulfides is palladium sulfide (PdS). Although it is expensive, palladium sulfide forms the basis for several very active catalysts. Zeolites loaded with transition metal sulfides have also been used as hydrodesulfurization catalysts.

The surface area of the catalyst is usually large (200–300 m^2/g), but almost this entire surface is contained within the pore space of the alumina. Cobalt and molybdenum are two of the more common metals

that are used as hydrodesulfurization catalysts and, as such, are dispersed in a thin layer within the pore system of the alumina. When these metals are used together as a hydrodesulfurization catalyst, the catalyst is more tolerant to poisoning agents, and is usually classed as being suitable for a wide variety of feedstocks, but more particularly to the heavy oils and residua. Other metals may be used such as a combination of nickel and molybdenum, but this catalyst is a more active hydrogenation catalyst than the cobalt–molybdenum catalyst and consumes more hydrogen per mole of sulfur removed. However, the nickel–molybdenum catalyst is useful for the hydrodesulfurization of catalytic cracking feedstocks (where maximum hydrogen consumption is desirable). The catalyst is more selective for nitrogen removal from feedstocks and presumably can tolerate higher nitrogen feedstocks without losing as much activity in a given time than the cobalt–molybdenum catalyst. Another catalyst that has received some attention is nickel–tungsten on alumina, which is a very active hydrogenation catalyst and displays a high activity in hydrocracking reactions.

Hydrodesulfurization catalysts are normally used as extrudates or as porous pellets, but the particle size and pore geometry have an important influence on process design, especially for the heavier feedstocks. The reaction rates of hydrodesulfurization catalysts are limited by the diffusion of the reactants into, and the products out of, the catalyst pore systems. Thus, as the catalyst particle size is decreased, the rate of desulfurization is increased but the pressure differential across the catalyst bed also diminishes and a balance must be reached between reaction rate and pressure drop across the bed.

Hydrodesulfurization catalysts are usually more active in the sulfide form with the external sulfur being applied either during the preparation from a sulfur-containing source or during the initial contact of the catalyst with the sulfur-containing feedstock. Chemically, the sulfiding process is a reduction in the oxide form of the metal (cobalt, nickel, etc.) on the alumina support by conversion to the metal sulfide. This can be achieved, in the absence of a sulfur-containing feedstock, by injecting hydrogen sulfide or any other low-boiling sulfide (carbon disulfide, dimethyl sulfide, and the like) directly into the recirculating hydrogen stream. If the feedstock is present and contains above 1% sulfur, addition of an external source is not necessary. Temperatures for the sulfiding process usually range from 290°C to 315°C (550–600°F), while minimum pressures of about 150 psi are also recommended.

CATALYST BED PLUGGING

Catalyst bed plugging can arise in a variety of ways, but the overall effect of bed plugging is always the same: expensive shutdowns and possibly the complete renewal of the expensive catalyst. Thus, the deposition of rust, coke, or metal salts (e.g., sodium chloride) from heavier and dirtier feedstock may contribute to the plugging of a catalyst bed. Vanadium and nickel may also be deposited onto the surface of the catalyst as well as into the pore system. Asphaltene deposition from residua and heavy oils is also a potential means of bed plugging; coagulation of the asphaltenes becomes appreciable at temperatures above 420°C (790°F) with the formation of hard, coke-like materials on the catalyst.

The exothermal nature of the reaction may also contribute indirectly to catalyst plugging. For example, lack of proper control over the heat liberated during the hydrodesulfurization process may lead to the formation of localized hot spots in the catalyst that can then initiate asphaltene coagulation at these points. In a similar manner, adsorption of the lighter constituents from a whole crude (heavy oil) or from, say, an atmospheric residuum could leave tarry materials within the void space of the catalyst, thereby leading to the eventual denaturing of the catalyst with a concurrent loss in catalyst efficiency. Finally, another cause of bed plugging may be the physical nature of the catalyst itself. Agglomeration of catalyst particles may lead to the formation of catalyst lumps within the bed, which again leads to catalyst inefficiency, presumably through formation of localized hot spots within the bed. The resulting end point of any of the above deficiencies will be an increased pressure drop through the catalyst bed that eventually requires process shutdown.

Catalyst bed plugging (and a subsequent expensive shutdown) may be minimized by using corrosion-resistant alloys throughout the entire system of feed lines, heaters, reactors, etc., thereby reducing the risk of rust accumulation in the feedstock. Removal of debris that has accumulated in the feedstock from the transportation and storage aspects may be achieved by mechanical filtering. Every attempt should be made to desalt the crude oil to remove any inorganic materials (sodium chloride, etc.) that may originate from the brines that are associated with crude oils as they are pumped to the surface of the earth. In the processing of the heavier feedstocks, it may be advisable to suspend the top layer of catalyst in basket-type devices so that this layer, when plugged, can be conveniently removed and replaced. It may be necessary to use two or more reactors in parallel rather than one large reactor—this would not only lead to more catalyst surface area for collecting debris, but if shutdown was inevitable, one of the reactors could be left onstream while the other reactor was being cleaned. Alternatively, upflow reactors would create the tendency for any debris in the feedstock to fall out of the catalyst,

and a circulating stream of liquid below the catalyst bed could be channeled through a filter to remove any solid materials. In a similar manner, prefiltering of the feedstock through a bed of bauxite or similar material will facilitate removal of debris and, in the presence of hydrogen, may also encourage removal of vanadium.

CATALYST POISONING

Hydrocarbons, especially aromatic hydrocarbons, can undergo multiple condensation reactions in the presence of catalysts to form coke. This coke is a complex polynuclear aromatic material that is low in hydrogen. Coke can deposit on the surface of a catalyst, blocking access to the active sites and reducing the activity of the catalyst. Poisoning by coke deposits is a major problem in fluid catalytic cracking catalysts where catalysts containing deposited coke are circulated to a fluidized bed combustor to be regenerated. In hydrocracking, coke deposition is virtually eliminated by the catalyst's hydrogenation function. However, the product referred to as ''coke'' is not a single material. The first products deposited are tarry deposits that can, with time and temperature, continue to condense to a solid deposit.

In a hydrodesulfurization system, the hydrogenation function adds hydrogen to the tarry deposits. This reduces the concentration of coke precursors on the surface. There is, however, a slow accumulation of coke that reduces activity over a 1–2 yr period. Refiners respond to this slow reduction in activity by raising the average temperature of the catalyst bed to maintain conversions. Eventually, however, an upper limit to the allowable temperature is reached and the catalyst must be removed and regenerated.

The metallurgy of the reactor, product quality considerations, or catalyst selectivity sets the upper temperature limits. Metallurgy considerations result from decreases in wall strength at higher temperatures. Product quality can be compromised at higher temperatures because thermal cracking reactions begin to compete with acid cracking as temperatures approach 400°C (750°F). Selectivity changes result from all the possible side reactions, which become more important as the temperatures rise.

Burning off the accumulated coke can regenerate catalysts containing deposited coke. This is achieved using rotary or similar kilns rather than leaving catalysts in the hydrocracking reactor, where the reactions could damage the metals in the walls. Removing the catalysts also allows inspection and repair of the complex and expensive reactor internals, discussed below. Regeneration of a large catalyst charge can take weeks or months, so refiners may own two catalyst loads, one in the reactor, one regenerated and ready for reload.

In addition, organometallic compounds produce metallic products that influence the reactivity of the catalyst by deposition on its surface. The transfer of metal from the feedstock to the catalyst constitutes an irreversible poisoning of the catalyst. After combustion to remove the carbonaceous deposits, the catalysts are treated to redisperse active metals.

PROCESS VARIABLES

The major process variables are 1) reactor temperature; 2) hydrogen pressure; 3) LHSV; and 4) hydrogen recycle rate.

Reactor Temperature

The temperature in the hydrodesulfurization reactor is often considered to be the primary means by which the process is controlled. For example, at stabilized reactor conditions, a rise of 10°C (18°F) in the reaction temperature will substantially increase, and may even double, the reaction rate. Generally, an increase in the temperature (from 360°C to 380°C, i.e., from 680°F to 715°F) will increase the conversion slightly or for a fixed conversion of about 90% enable the quantity of catalyst necessary for the process to be halved.

In the same manner as in hydrocracking, hydrogen is added at intermediate points in hydrodesulfurization reactors. This is important for control of reactor temperatures. The mechanical devices in the reactor, called reactor internals, that accomplish this step are very important to successful processes. If redistribution is not efficient, some areas of the catalyst bed will have more contact with the feedstock.

There are, however, limits to which the temperature can be increased without adversely affecting process efficiency; at temperatures above 410°C (770°F), thermal cracking of the hydrocarbon constituents becomes the predominant process that can lead to formation of considerable amounts of low-molecular-weight hydrocarbon liquids and gases. In addition, increasing the partial pressure of the hydrogen cannot diminish these high-temperature cracking reactions. In addition, excessively high temperatures (above 400°C or 750°F) lead to deactivation of the catalyst much more quickly than lower temperatures.

Hydrogen Pressure

The overall effect of increasing the partial pressure of hydrogen is to increase the extent of the conversion through an increase in catalyst activity. This is to be

expected because the essential function of the catalyst is to serve as a means by which the reactants are brought together, thereby promoting interaction between the feedstock constituents and the hydrogen. As with the temperature variable, there are also limitations to increasing the partial pressure of the hydrogen. Use of excessively high partial pressures (for conventional feedstocks—above 1000 psi; for heavy feedstocks this figure could be as high as 2000 psi) may only serve to saturate the catalyst, and any increase in the partial pressure of the hydrogen will affect the conversion only slightly.

Liquid Hourly Space Velocity

The LHSV is the ratio of the hourly volume flow of liquid in, say, barrels to the catalyst volume in barrels, and the reciprocal of the LHSV gives the contact time. Because the catalyst volume for the process will be constant, the space velocity will vary directly with the feed rate. A decrease in the LHSV (or, alternatively, an increase in the contact time) will usually bring about an increase in the efficiency (or extent) of the hydrodesulfurization process. To maintain a fixed rate of hydrodesulfurization when the feed rate is increased, it may be necessary to increase the temperature.

Hydrogen Recycle Rate

The optimum use of hydrodesulfurization catalysts requires a relatively high hydrogen partial pressure, and it is therefore necessary to introduce with the feedstock quantities of hydrogen that are considerably greater than required on the basis of stoichiometric chemical consumption. In all cases, process economics dictate that unused hydrogen should be recycled after it has been partially (or completely) freed from hydrogen sulfide that was produced in the previous pass.

The overall hydrogen consumption is a summation of several processes:

1. Removal of the sulfur, nitrogen, and oxygen in the feedstock as the hydrogenated analogs, i.e., hydrogen sulfide, ammonia, and water, respectively.
2. Addition of hydrogen to unsaturated (olefin) functions in the products as will occur in the prevailing conditions of the hydrodesulfurization process.
3. Destruction, by saturation (i.e., addition of hydrogen), of certain aromatic compound types.
4. Stabilization of unsaturated short-lived organic intermediates that exist during hydrocracking.

Thus, if the hydrogen is not recycled, the process economics is unfavorable and, in addition, the efficiency of the hydrodesulfurization reaction may be adversely affected because of the possible competing reactions outlined above.

The hydrodesulfurization process involves catalytic treatment with hydrogen to convert the various sulfur compounds present to hydrogen sulfide. The hydrogen sulfide is then separated and converted to elemental sulfur by the Claus process. From this point some of the hydrogen sulfide is oxidized to sulfur dioxide by air and sulfur is formed by the overall reaction.

CONCLUSIONS

Petroleum refining has entered a significant transition period as the industry moves into the 21st century. Refinery operations have evolved to include a range of next-generation processes (that include hydrotreating of heavy oils and residua) as the demand for transportation fuels and fuel oil has shown a steady growth.

Hydroprocessing petroleum fractions has long been an integral part of refining operations, and in one form or another, hydroprocesses are used in every modern refinery. The primary goal of the recent hydrodesulfurization processes is to convert heavy feedstocks to lower-boiling products and during the conversion there is a reduction in the sulfur content. The technology of hydroprocesses is well established for gas oil and lower-boiling products but processing heavy oils and residua presents several problems that are not found with distillate processing and that require process modifications to meet the special requirements that are necessary for heavy feedstock desulfurization.

REFERENCES

1. Speight, J.G. *The Desulfurization of Heavy Oils and Residua*, 2nd Ed.; Marcel Dekker Inc.: New York, 1999.
2. Speight, J.G. *The Chemistry and Technology of Petroleum*, 3rd Ed.; Marcel Dekker Inc.: New York, 1999.
3. Speight, J.G.; Ozum, B. *Petroleum Refining Processes*; Marcel Dekker Inc.: New York, 2002.
4. Topsøe, H.; Clausen, B.S.; Massoth, F.E. *Hydrotreating Catalysis*; Springer-Verlag: Berlin, 1996.

Hydrodynamics of Trickle-Bed Reactors

K. D. P. Nigam
Arunabha Kundu
Department of Chemical Engineering, Indian Institute of Technology, New Delhi, India

INTRODUCTION

Three-phase reactors consisting of flow of liquid and gas with a fixed bed of catalyst have various applications, particularly in the petroleum industry for hydroprocessing of oils. Mainly, trickle-bed reactors (TBRs) are used for this purpose. Trickle-bed reactors are catalytic randomly packed fixed-bed tubular devices traversed vertically downward by a gas–liquid stream (Fig. 1). The reactor is widely used in many gas–liquid–solid catalytic industrial applications like petroleum refining [hydrotreatment (hydrodesulfurization, hydrodenitrogenation, alkylation, etc.) and hydrocracking of petroleum fractions], petrochemical (hydrogenation of naphtha cracked main products for getting different usable products, hydrogenation of higher aldehydes, reactive distillation) and chemical industries (for oxidation and absorption purposes), waste treatment, and biochemical and electrochemical processing.[1] Typical process conditions and hydrogen consumption in TBR for various hydrotreating processes are given in Table 1.[2] The importance of hydroprocessing has been accelerated because of the imposition of stringent environmental regulations on petroleum products to achieve future goals of sustainability and a cleaner environment and a more demanding situation of middle distillates (mainly transportation fuels) against the availability of heavier and high-sulfur crudes.

The TBR reactor configuration ensures flexibility of operation and high throughputs of gas and liquid. Owing to a motionless catalyst bed, nearly plug flow is achieved in TBRs and in that respect, they are superior to other three-phase reactors where the catalyst is either slurried or fluidized.[3] Trickle-beds are usually operated at elevated pressures of about 2–30 MPa to slow down catalyst deactivation, increase the concentration of the gaseous component in the liquid phase, attain high conversion, achieve better heat transfer, and handle large gas volumes at less capital expense. High catalyst load, easy catalyst separation, and no catalyst attrition are other advantages for this type of reactor, whereas high pressure drop, flow maldistribution, and long diffusion distance are the areas concerned in the performance of TBR. Being a cocurrent mode of operation with respect to the flow of liquid and gas, it is also disadvantageous as conversion is suppressed by competitive adsorption because of different products (where most part of the bed is under NH_3 and H_2S rich condition for the case of hydroprocessing).

HYDRODYNAMIC PARAMETERS

The different hydrodynamic parameters that are important for the design and scale-up of TBR, are pressure drop, liquid holdup, flow regime, liquid distribution, and wetting efficiency. The interactions of different hydrodynamic parameters are shown in Fig. 2. It is established that pressure drop increases with the increase of operating pressure, gas mass flux, liquid mass flux, and viscosity of the liquid phase. Pressure drop is decreased with the increase of bed porosity and structure of the bed. The structure of the bed is dependent on the type of loading of catalyst (sock loading and dense loading). Though the dense loading gives higher pressure drop than that of sock loading, it improves the performance in TBR. The H_2/hydrocarbon ratio reduces up to 40% in dense loading for getting the same result even when keeping all other operating parameters the same with respect to sock loading.[4] Liquid holdup reduces with the raise of operating pressure. At relatively low gas mass flux, it decreases more rapidly than at higher gas mass flux. Though at higher liquid mass flux external wetting efficiency of the catalyst particles is not a crucial hydrodynamic parameter, it is important in TBR as it deals with very low liquid velocity. It has more importance in small-scale TBRs that are generally used for the development of better technology and catalyst evaluation because of lower cost and safer operation.[5] Wetting efficiency increases with a raise of gas mass flux and operating pressure as these process variables increase spreading of liquid over catalyst particles because of the increase in shear stress on the gas–liquid interface.[6] The increase in liquid mass flux increases the liquid holdup and wetting efficiency. Liquid distribution is an important parameter for proper utilization of the catalyst. Properly designed liquid distributors,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007996

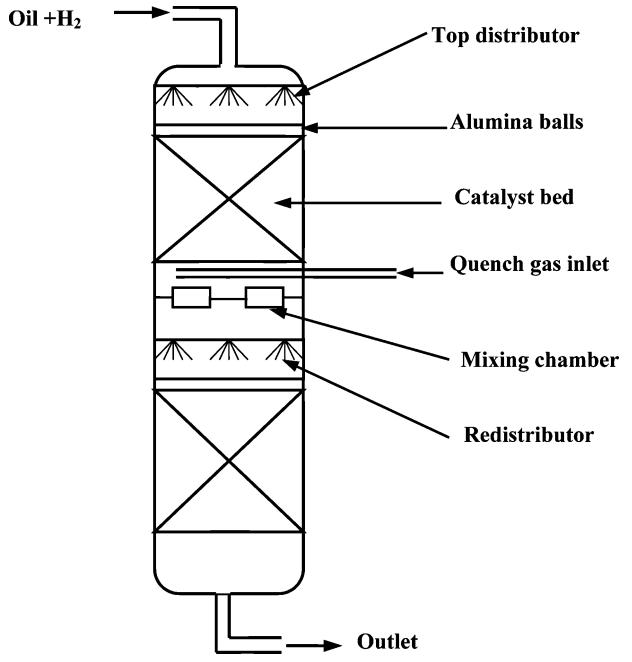


Fig. 1 Schematic diagram of one typical trickle-bed reactor.

properly designed structure of catalyst, and appropriate method of packing are necessary for getting uniform liquid distribution.

Pressure Gradient and Liquid Holdup

Accurate knowledge of two-phase frictional pressure drop and liquid holdup is essential for design, scale-up, and performance of TBR. The hydrodynamics is affected differently in each flow regime.[7–10] Of particular interest in the industry is the extensively used trickle flow encountered at low gas and liquid superficial velocities.

Two-phase pressure gradient is defined as the variation of the internal pressure per unit reactor length. The pressure gradient is related to the mechanical energy dissipation owing to the two-phase flow through the fixed bed of solid particles. It is needed in evaluating the mechanical energy losses, in sizing equipment for pumping and compression of the fluid, and most of the other relevant design variables like gas–liquid mass transfer, liquid–solid mass transfer, and heat transfer are often estimated using the knowledge of two-phase frictional pressure drop and liquid holdup.[11]

Liquid holdup (h_L) is defined as the volume of liquid contained in the bed per unit bed volume. It is a function of the physical properties of the fluid phases and the bed characteristics. It is a basic parameter for reactor design, because it is related to other important parameters, namely, pressure gradient, gas–liquid interfacial area, the mean residence time of the liquid phase, catalyst loading per unit volume, axial dispersion coefficient, mass transfer characteristics, and heat transfer coefficient at the wall, etc. The optimal value of liquid holdup is desirable for better performance of TBR as a high value of liquid holdup will increase mass transfer resistance while too low a value of liquid holdup will decrease the proper utilization of the catalyst bed. Sometimes, the term total liquid saturation (β_t) is used to describe the amount of liquid in the bed. It is defined as the volume of liquid present in a unit void volume of the reactor. Thus, the liquid holdup and total liquid saturation are related as:

$$h_L = \varepsilon\beta_t \tag{1}$$

As the catalyst particles are porous, the total liquid saturation (β_t) is the sum of internal or intraparticle liquid saturation and interparticle or external liquid saturation. The external liquid saturation is further divided into residual or static liquid saturation (β_s), and free draining or dynamic saturation (β_d). The static liquid saturation is a function of liquid physical properties and particle shape, size, and wettability. The dynamic liquid saturation is the fraction of volume of liquid collected after draining of the reactor on simultaneously closing the inlet and outlet streams.

Table 1 Typical operating conditions for different hydrotreating processes used in trickle-bed reactors

Hydrotreating process	Temperature (°C)	Hydrogen partial pressure (atm)	LHSV	Hydrogen consumption (Nm^3/m^3)
Naphtha	320	10–20	3–8	2–10
Kerosene	330	20–30	2–5	5–15
Atm. gas oil	340	25–40	1.5–4	20–40
VGO	360	50–90	1–2	50–80
ARDS	370–410	80–130	0.2–0.5	100–175
VGO HCR	380–430	90–200	0.5–1.5	150–300
Residue HCR	400–440	120–210	0.1–0.5	150–300

VGO, vacuum gas oil; ARDS, atm. residue desulfurization; VGO HCR, vacuum gas oil hydrocracking; Residue HCR, residue hydrocracking.

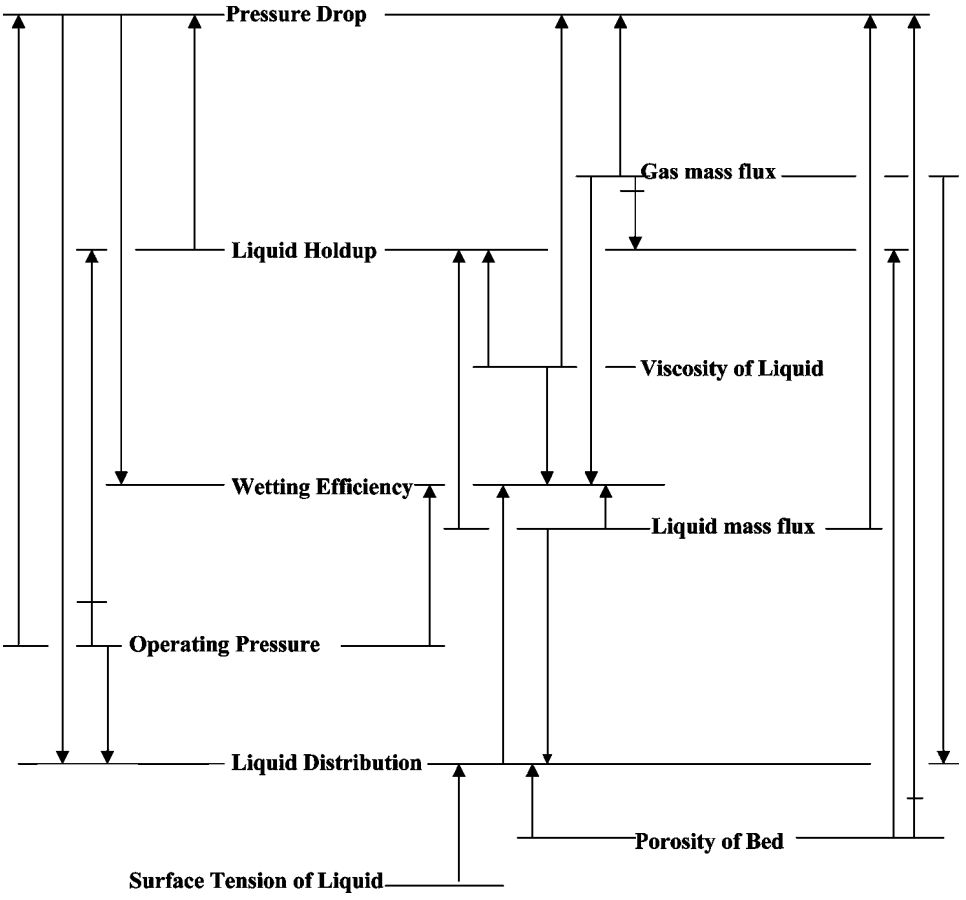


Fig. 2 Cause–effect diagram of different hydrodynamic parameters in TBR with the process variables and properties of gas and liquid (cause effect, increasing effect and decreasing effect).

There are three different techniques for measuring liquid holdup in a laboratory TBR operated under high pressure: 1) tracer method; 2) drainage method; and 3) gravimetric or weighing method. In the tracer method, the tracer is introduced in the liquid stream and the concentration of the tracer is analyzed at the outlet of the stream. Different types of common tracers are used. These are organic liquids (e.g., heptane and hexane), electrolytes (NaCl, KCl, NH_4Cl, etc.), and radioisotope tracers (bromine-82, molybdenum-99, and technetium-99m, etc.). For analyzing the tracer concentration at the outlet, gas chromatograph and refractometer are used for the case of organic liquid as a tracer. Conductivity measurements are taken when the tracer is an electrolyte.

For the case of radioisotope tracer experiments, nonintrusive methods are used to get the outlet concentration of tracer by utilizing collimated scintillation detectors. Radioisotope tracers have many advantages such as on-line detection, high detection sensibility, and availability in different compatible forms over conventional tracers.[12] This method can also help in troubleshooting and checking the performance of industrial TBR under operational conditions.

In the drainage method, the dynamic liquid holdup is measured in which the inlet and outlet streams are simultaneously shut off and the liquid drained from the reactor is measured. In the gravimetric method, the column is weighed during continuous operation. From the weight obtained, the weight of the dry column is subtracted so that the value of the liquid holdup can be obtained. The method permits the determination of both dynamic and static holdup. However, its application encounters two difficulties: one is linked with the effect of auxiliary equipment (connections with the fluid inlet and outlet lines) and the other with the inaccuracy of measuring differences between heavy weights. The tracer method and drainage method give comparable values of liquid holdup and are recommended for measuring liquid holdup in TBR at high pressure operation.[11,13–15] Nemec et al. measured liquid holdup in TBR at high pressure (up to 7 MPa) with the weighing method and recommended this procedure as it is less time-consuming compared to the other two methods and is less costly compared to the tracer method.[16] The volumetric liquid–liquid mass transfer coefficients between dynamic and static zones and volumetric liquid–solid mass transfer coefficients between static liquid and porous particles can also be evaluated from tracer studies.[17]

Recently, magnetic resonance imaging (MRI) technique has been applied for the measurement of liquid

holdup.[18] In this technique, high-resolution images are taken that are able to characterize liquid rivulets within the bed and to detect the presence of thin water films on the surfaces of the packing elements. Hence, magnetic resonance signal is detected only from the liquid.

The correlations and models that have been established on the basis of atmospheric pressure data are summarized in Refs.[1,19–22]. Investigations in pressurized TBR have been performed by Ellman et al., Wammes et al., Larachi et al., and Al-Dahhan et al.[20,21,23–28] These studies have shown that the hydrodynamic parameters are considerably affected by the gas density, and the correlations and models based on data at atmospheric pressure are not valid in the whole range of operating pressure.

The existing hydrodynamic models can be broadly classified into two different categories on the basis of empirical approach and theoretical approach. The empirical approach is based on dimensional analysis to produce explicit correlations for pressure drop and liquid holdup using flow variables and packing characteristics or using the Lockhart–Martinelli parameter, which was proposed for open horizontal tubes.[19–21,24–27,29–35] The theoretical approach involves balance equations deduced from the mass and momentum conservation laws. The different forces acting in the flow of gas and liquid over the catalyst particle are derived from different approaches. Sáez and Carbonell expressed the volume-averaged momentum equations for steady, one-dimensional incompressible flow of gas and liquid to derive the equations for liquid holdup and pressure drop expressed in terms of relative permeability.[36] The relative permeability of each phase has been correlated as a function of liquid saturation of each phase depending on the experimental results. In this concept, the phase interaction terms are considered to be negligible. The phenomenological slit model developed by Holub et al., Al-Dahhan and Dudukovic, and Iliuta et al. involves momentum and mass balances for the local flow in the assumed geometry and mapping the average bed properties to the assumed geometry.[22,28,37,38] It is a modified form of the Ergun equation. Initially, Holub et al. modeled the complex geometry of the actual void space in the catalyst bed at the pore level by the much simpler flow inside a rectangular slit.[22,38] In their model, the width of the slit is a function of bed porosity, the liquid was assumed to flow equally on the upper and lower surfaces, and the angle of inclination of the slit to the vertical axis is related to a tortuosity factor for the packed bed. The surface area per unit volume of solid in this rectangular slit was made equal to the surface area per unit volume of solid in the reactor. They introduced the concept of slip of the velocity and shear at the gas–liquid interface by introducing two slip parameters:

$$V_{i,\mathrm{G}} = f_{\mathrm{v}} V_{i,\mathrm{L}} \tag{2}$$

$$\tau_{i,\mathrm{L}} = f_{\mathrm{s}} \tau_{i,\mathrm{G}} \tag{3}$$

where $V_{i,\mathrm{L}}$ and $V_{i,\mathrm{G}}$ are interfacial velocity of gas and liquid respectively; $\tau_{i,\mathrm{L}}$ and $\tau_{i,\mathrm{G}}$ are interfacial shear stress of liquid and gas respectively; f_{s} and f_{v} are shear and velocity slip factors respectively.

The resulting pressure drop and liquid saturation model of Holub et al. indicates zero shear stress at the gas–liquid interface. Attou et al. considered a model based on macroscopic mass and momentum conservation laws in which the drag force has a contribution of both particle–liquid and gas–liquid interactions.[39] The liquid–solid and gas–liquid interaction forces are formulated on the basis of the Kozeny–Carman equation by taking into account the presence of liquid films and the gas–liquid slip motion. The gas–liquid interaction forces are generated because of the gas–liquid drag for the relative motion between the liquids and the force by which the gas pushes the liquid against the solid particles. The fundamental force balance model developed by Tung and Dhir involves force balance equations in liquid and gas phase in an elemental reactor volume.[40] Recently, Narasimhan et al. extended this approach with particle–gas drag, particle–liquid drag, and liquid–gas interfacial drag in addition to taking into consideration the excess drag for the tortuosity effect.[41] The tortuosity corrects the gravity term in the liquid phase and gas phase force balance equation, which has the expression

$$\tau = g\cos\theta - \mu g \sin\theta \tag{4}$$

where θ accounts for the distance traveled in the tortuous path and μ accounts for the excess loss in force caused by drags on the liquid and liquid following tortuous path. The comparison of different hydrodynamic models for TBR has been described by Kundu et al.[42]

Flow Regime

In TBR, basically four flow regimes have been observed depending on the gas and liquid and packing characteristics.[1]

Trickle flow regime

Trickle flow occurs at low liquid and gas flow rates. In trickle flow, the liquid flows down the reactor on the surface of the packing in the form of rivulets and films while the gas phase travels in the remaining void space. This regime is also termed gas continuous regime or

homogenous flow or low-interaction regime because very little interaction between gas and liquid exists in this regime.

Pulse flow regime

The pulse flow occurs at relatively high gas and liquid input flow rates. It refers to the information of slugs that have a higher liquid content than the remainder of the bed. The pulsing behavior refers to gas and liquid slugs traversing the reactor alternately.

Spray flow regime

Spray flow regime occurs at high gas flow rates and low liquid flow rates. The liquid phase moves down the reactor in the form of droplets entrained by the continuous gas phase.

Bubble/dispersed bubble flow regime

The bubble flow regime occurs at high liquid flow rates and low gas flow rates. The entire bed filled with the liquid and gas phase is in the form of slightly elongated bubbles. If the gas flow rate is increased, the bubbles become highly irregular in shape. For the foaming systems, an increase in gas and/or liquid flow rates leads to foaming, foaming pulsing, pulsing, and spray flow regime.

Flow regime transition between trickle flow and pulse flow

The knowledge of the hydrodynamics regime prevailing in the reactor is crucial in the design of TBR, as with the change in the hydrodynamic regime, all the hydrodynamic and kinetic parameters of the reactor also change considerably. The industrial TBRs are operated in the pulse flow regime. The transition between the two regimes is not very sharp. Operation in the trickle flow regime ensures a large liquid-residence time and provides a large single-pass conversion of the liquid reactant. This is therefore useful for reactions that are kinetic controlled. Operation in the pulsing regime provides an increase in the mass transfer coefficients and is therefore suitable for relatively fast mass transfer limited reactions. Heat transfer rate is also intensified because of the vigorous interactions between the phases, and this is of crucial importance in vessels packed with catalyst, in which highly exothermic reactions occur, e.g., in hydroprocessing. Catalyst utilization is also increased because of increased wetting of the catalyst particles. Peclet number (= LV/D, where L = catalyst bed length, V = superficial velocity, D = dispersion coefficient) also increases while changing the flow regime from trickle flow to pulse flow.[43,44] This is because the exchange rate between the two streams in the pulse flow regime is high and the pulses probably cause mixing between the different parallel flowing liquid streams in the column. Thus, if plug flow is required to attain a high yield, a pulse flow regime is preferred to a trickle one. Uniformity of flow through the bed is also obtained in pulse flow regime. Flexibility in changing the operating gas and liquid flow rates is also greater in the pulsing regime. Therefore, it is necessary to know the flow regime in which the reactor is operating for a given condition. It is also important to know whether the same flow regime will be maintained in scaling up from laboratory to pilot- or commercial-scale reactors.

Several investigators have studied the flow pattern in TBR at atmospheric pressure and presented their data in terms of flow maps. Various coordinates have been used by different investigators to present the flow maps to take into account the variation in density, viscosity, and surface tension of the fluids. The shift in the boundary between the trickle flow and the pulse flow regimes occurs at higher velocities of liquid and gas by increase in operating pressure.[24,45] This is because of a decrease in liquid holdup at higher operating pressure, in which the mean liquid film thickness becomes smaller and, consequently, the liquid film cannot collapse any more to initiate the pulses. The flow regime transition occurs at higher gas and liquid velocities as the packing hydrophobicity increases and the transition shifts to lower mass flow rates of liquid and gas as the concentration of surfactant increases.[46] The transition boundary shifts toward higher throughputs for small-size catalyst particles.[47]

The most commonly used method for detecting the flow regime transitions involves the visual observation of the flow pattern across the column walls. The subjective nature of this method results in the uncertainty of the flow regime boundaries. Another possible method is the measurement of conductivity traces. On the other hand, because of the various variables such as physical fluid properties, surface tension, bed porosity, size and shape of solid particles, which can significantly affect the flow regime, the existing flow maps are valid only in the particular range of conditions in which the experimental data are obtained. Other phenomena such as hysteresis and packing wettability have proven to be difficult in the use of these flow maps. Considering these limitations, theoretical or semiempirical models have been developed to predict the transition between the trickle and pulse regimes.

Liquid Distribution

Liquid distribution plays an important role in determining the reactor performance in TBR. To guarantee

uniform liquid distribution in a TBR has been a tough problem that has drawn much attention. The liquid tends to flow preferentially along the existing filaments where the porosity is high. The introduction of gas flow into the liquid–solid system smoothens the liquid distribution to some extent because of the competition between the liquid and gas phases for the interstitial pore space.[48]

Improper liquid distribution will result in nonoptimal use of catalyst, rapid deactivation of a part of the catalyst by creating hot spots or thermal instability, which means that improving flow distribution increases capacity and cycle length for existing plants. In highly exothermic reactions, the liquid is vaporized and the heat generated is not carried away by the liquid, leading to rapid deactivation of the catalyst.

The role played by the initial liquid distribution at the top of the bed is of decisive importance. A poorly designed liquid distributor can have a serious impact on the performance of TBR. The method of packing the catalyst particles, particularly the cylindrical extrudates, also affects the liquid distribution in TBR. If the cylindrical extrudates are packed with their major axis pointing downward toward one side of the reactor, the liquid will flow toward that side.[49] Al-Dahhan et al. reported that the essential requirement of the packed bed is to ensure reproducibility of the packing.[50] This is because even if a uniform liquid distribution is achieved at the distributor, significant bypassing (channeling) and/or segregation could occur because of the improper way in which the catalyst and fines are packed. The effect of prewetting of the catalyst bed is significant on the liquid distribution, as reported by Ravindra et al. and Kundu et al.[51,52] This prewetting effect can be explained in terms of liquid spreading over the dry catalyst particles (where pores are only responsible) and prewetted catalyst particles (where liquid film is only responsible).[53]

So, liquid distribution in TBR mainly depends on three factors:

1. Structure of the packed bed.
2. Gas–liquid–solid interaction.
3. Design of the distributor.

The above factors are well explained by Maiti et al.[54] Under the structure of the packed bed, the liquid distribution is dependent on: 1) porosity distribution; 2) D/d_p ratio (D is the diameter of reactor and d_p is the equivalent particle diameter); 3) H/d_p ratio (orientation problem) (H is the height of the catalyst bed); and 4) method of loading—convex, concave, and random (sock and dense loading).

Numerous studies on porosity distribution in randomly packed beds are available (Refs.[55,56] for experimental studies and Ref.[57] for computational study). These studies have shown that the porosity is higher near the vicinity of the wall and fluctuates significantly in the near-wall region (width of about four to five particle diameters). At a large ratio of D/d_p, there will also be lesser portions of wall and a large inner part of the catalyst bed, which will decrease the nonuniform distribution of liquid. Conversely, at a small ratio of D/d_p, the flow will be uneven because of the large portion of the wall. This effect is called the wall effect. Herskowitz and Smith suggested a value of 18 for the D/d_p ratio above which wall effect can be eliminated.[58] Al-Dahhan and Dudukovic proposed a value of 20, while Saroha et al. reported a value of 25 for D/d_p ratio.[28,48]

The dependence of gas–liquid–solid interaction on liquid distribution relies mainly on the gas and liquid mass fluxes and physical properties of liquids (surface tension, etc.). The design of liquid distributors in commercial hydroprocessing reactors is well explained by Jacobs and Milliken.[59] Three main factors are considered during the designing of the liquid distributor: 1) center-to-center spacing, which is directly proportional to particle size, is chosen as an optimum value so that radial mixing compensates for maldistribution; 2) liquid discharge pattern; and 3) wall coverage, in which the problem arises because of the absence of distributors located close to the reactor wall.

Different procedures for studies of liquid distribution in TBR are available in the literature. Though the liquid distribution studies using the collector at the outlet of the bed are mostly utilized because of its simplicity to carry out the experiments, there is a chance of flow redistribution at the exit region. So, recent studies emphasize the tomographic technique and the video imaging technique that provide the information on flow distribution more quantitatively and are useful for the validation of the computational fluid dynamics (CFD) model.[60] Marcandelli et al. suggested the residence time distribution (RTD) technique for the liquid distribution studies in a commercial reactor, which is also quantitative.[61] They also reported that liquid collectors and tomography are not good techniques for the liquid distribution studies in a commercial reactor because of unavailability of quantitative information and low spatial resolution with inapplicability in a porous medium, respectively.

CONCLUSIONS

In the past 50 yrs. much attention has been focused on pressure drop, liquid holdup, flow regime transition, and liquid distribution in TBR. Though significant progress in modeling the hydrodynamics of TBR theoretically has been made, the prediction of some

hydrodynamic parameters like pressure drop in TBR is still not so accurate because of the complexity of the gas–liquid flow pattern and their interaction. CFD simulation can be applied for the above purpose. The pulse flow regime is industrially favorable; more studies on flow behavior under this regime should be carried out. A good liquid distributor, appropriate structure of the bed, and starting with prewetted bed can improve liquid distribution in TBR. More emphasis is expected on studies of the effect of wettability on liquid distribution. To get a better estimation of structural information and flow field, MRI method can be applied because of its direct way of measurement. A high-quality distributor is to be developed based on cold flow modeling and supported by CFD calculations of commercial operating conditions with special emphasis on discharge pattern. The effect of flow dynamics on even and uneven irrigation of the particle may be explored.

The above aspects taken together will provide a holistic approach in the design of TBR, that will help to meet the EURO III and IV standards (5 and 10 ppmw) of the transportation fuels for a much better, cleaner, and greener environment.

REFERENCES

1. Saroha, A.K.; Nigam, K.D.P. Trickle bed reactors. Rev. Chem. Eng. **1996**, *12* (3–4), 207.
2. Nigam, K.D.P. Design of trickle-bed reactors. In *Hydroprocessing in Petroleum Refining Industry—A Compendium*; Verma, R.P., Bhatnagar, A.K., Eds.; Lovraj Memorial Trust: Delhi, 2000; 391.
3. Al-Dahhan, M.H.; Larachi, F.; Dudukovic, M.P.; Laurent, A. High-pressure trickle-bed reactors: a review. Ind. Eng. Chem. Res. **1997**, *36*, 3292.
4. Nooy, F.M. Dense loading of catalyst improves hydrotreater performance. Oil Gas J. **1984**, *12*, 152.
5. Chander, A.; Kundu, A.; Bej, S.K.; Dalai, A.K.; Vohra, D.K. Hydrodynamic characteristics of cocurrent upflow and downflow of gas and liquid in a fixed bed reactor. Fuel **2001**, *80*, 1043.
6. Al-Dahhan, M.H.; Dudukovic, M.P. Catalyst wetting efficiency in trickle-bed reactors at high pressure. Chem. Eng. Sci. **1995**, *50* (15), 2377.
7. Charpentier, J.C.; Favier, M. Some liquid holdup experimental data in trickle bed reactors for foaming and non-foaming hydrocarbons. Am. Inst. Chem. Eng. J. **1975**, *21*, 1213.
8. Specchia, V.; Baldi, G. Pressure drop and liquid holdup for two-phase concurrent flow in packed beds. Chem. Eng. Sci. **1977**, *32*, 515.
9. Rao, V.G.; Drinkenburg, A.A.H. A model for pressure drop in two-phase gas–liquid downflow through packed columns. Am. Inst. Chem. Eng. J. **1985**, *31*, 1010.
10. Xiao, Q.; Cheng, Z.M.; Jiang, X.; Anter, A.M.; Yuan, W.K. Hydrodynamics behavior of a trickle bed reactor under forced pulsing flow. Chem. Eng. Sci. **2001**, *56*, 1189.
11. Tosun, G. A study of co-current down-flow of non-foaming gas–liquid systems in a packed bed. 2. Pressure drop: search for a correlation. Ind. Eng. Chem. Process Des. Dev. **1984**, *23*, 35.
12. Pant, H.J.; Kundu, A.; Nigam, K.D.P. Radiotracer applications in chemical process industry. Rev. Chem. Eng. **2001**, *17*, 165.
13. Al-Dahhan, M.H.; Highfill, W. Liquid holdup measurement techniques in laboratory high pressure trickle bed reactors. Can. J. Chem. Eng. **1999**, *77*, 759.
14. Nigam, K.D.P.; Saroha, A.K.; Kundu, A.; Pant, H.J. Radioisotope tracer study in trickle bed reactors. Can. J. Chem. Eng. **2001**, *79*, 860.
15. Pant, H.J.; Saroha, A.K.; Nigam, K.D.P. Measurement of liquid holdup and axial dispersion in trickle bed reactors using radiotracer technique. Nukleonika **2000**, *45*, 235.
16. Nemec, D.; Bercic, G.; Levec, J. Gravimetric method for the determination of liquid holdup in pressurised trickle-bed reactors. Ind. Eng. Chem. Res. **2001**, *40*, 3418.
17. Nigam, K.D.P.; Iliuta, I.; Larachi, F. Liquid back-mixing and mass transfer effects in trickle-bed reactors filled with porous catalyst particles. Chem. Eng. Process. **2002**, *41*, 365.
18. Saderman, A.J.; Gladden, L.F. Magnetic resonance imaging as a quantitative probe of gas–liquid distribution and wetting efficiency in trickle-bed reactors. Chem. Eng. Sci. **2001**, *56*, 2615.
19. Sai, P.S.T.; Varma, Y.B.G. Pressure drop in gas–liquid downward flow through packed beds. Am. Inst. Chem. Eng. J. **1987**, *33*, 2027.
20. Ellman, M.J.; Midoux, N.; Laurent, A.; Charpentier, J.C. A New improved pressure drop correlation for trickle bed reactors. Chem. Eng. Sci. **1988**, *43*, 2201.
21. Ellman, M.J.; Midoux, N.; Wild, G.; Laurent, A.; Charpentier, J.C. A new improved liquid holdup correlation for trickle bed reactors. Chem. Eng. Sci. **1990**, *45*, 1677.
22. Holub, R.A.; Dudokovic, M.P.; Ramachandran, P.A. Pressure drop, liquid holdup, and flow regime transition in trickle flow. Am. Inst. Chem. Eng. J. **1993**, *39*, 302.
23. Wammes, W.J.A.; Westerterp, K.R. Hydrodynamics in a pressurised cocurrent gas–liquid trickle-bed reactor. Chem. Eng. Technol. **1991**, *14*, 406.

24. Wammes, W.J.A.; Mechielsen, S.J.; Westerterp, K.R. The influence of pressure on the liquid hold-up in a cocurrent gas–liquid trickle-bed reactor operating at low gas velocities. Chem. Eng. Sci. **1991**, *46*, 409.
25. Wammes, W.J.A.; Middelkamp, J.; Huisman, W.J.; deBaas, C.M.; Westerterp, K.R. Hydrodynamics in a cocurrent gas–liquid trickle bed at elevated pressures. Am. Inst. Chem. Eng. J. **1991**, *37*, 1849.
26. Larachi, F.; Laurent, A.; Midoux, N.; Wild, G. Experimental study of a trickle bed reactor operating at high pressure: two-phase pressure drop and liquid saturation. Chem. Eng. Sci. **1991**, *46*, 1233.
27. Larachi, F.; Laurent, A.; Midoux, N.; Wild, G. Some experimental liquid saturation results in fixed-bed reactors operated under elevated pressure in cocurrent upflow and downflow of the gas and the liquid. Ind. Eng. Chem. Res. **1991**, *30*, 2404.
28. Al-Dahhan, M.H.; Dudukovic, M.P. Pressure drop and liquid holdup in high-pressure trickle bed reactors. Chem. Eng. Sci. **1994**, *49*, 5681.
29. Larkins, R.P.; White, R.R.; Jeffery, D.W. Two-phase co-current flow in packed beds. Am. Inst. Chem. Eng. J. **1961**, *7*, 231.
30. Turpin, J.L.; Hungtinton, R.L. Prediction of pressure drop for two-phase, two component co-current flow in packed beds. Am. Inst. Chem. Eng. J. **1967**, *6*, 1196.
31. Sato, Y.; Hirose, T.; Takahashi, F.; Toda, M. Pressure loss and liquid holdup in packed bed reactor with co-current gas-liquid down-flow. J. Chem. Eng. Jpn. **1973**, *6*, 147.
32. Midoux, N.; Favier, M.; Charpentier, J.C. flow pattern, pressure loss, and liquid holdup data in gas–liquid down-flow packed beds with foaming and non-foaming hydrocarbons. J. Chem. Eng. Jpn. **1976**, *9*, 305.
33. Clements, L.D.; Schmidt, P.C. Two phase pressure drop in cocurrent downflow in packed beds: air–silicone oil systems. Am. Inst. Chem. Eng. J. **1980**, *26*, 314.
34. Xiao, Q.; Anter, A.M.; Cheng, Z.M.; Yuan, W.K. Correlations for dynamic liquid holdup under pulsing flow in a trickle-bed reactor. Chem. Eng. J. **2000**, *78*, 125.
35. Pinna, D.; Tronconi, E.; Tagliabue, L. High interaction regime Lockhart–Martinelli model for pressure drop in trickle-bed reactors. Am. Inst. Chem. Eng. J. **2001**, *47*, 19.
36. Saez, A.E.; Carbonell, R.G. Hydrodynamic parameters for gas–liquid cocurrent flow in packed beds. Am. Inst. Chem. Eng. J. **1985**, *31*, 52.
37. Holub, R.A.; Dudokovic, M.P.; Ramachandran, P.A. A phenomenological model for pressure drop, liquid holdup, and flow regime transition in gas–liquid trickle flow. Chem. Eng. Sci. **1992**, *47*, 2343.
38. Iliuta, I.; Larachi, F.; Al-Dahhan, M.H. Double-slit model for partially wetted trickle flow hydrodynamics. Am. Inst. Chem. Eng. J. **2000**, *46*, 597.
39. Attou, A.; Boyer, C.; Ferschneider, G. Modeling of the hydrodynamics of the cocurrent gas–liquid trickle flow through a trickle-bed reactor. Chem. Eng. Sci. **1999**, *54*, 785.
40. Tung, V.X.; Dhir, V.K.A Hydrodynamic model for two phase flow through porous media. Int. J. Multiphase Flow **1988**, *14*, 47.
41. Narasimhan, C.S.L.; Verma, R.P.; Kundu, A.; Nigam, K.D.P. Modeling hydrodynamics of trickle-bed reactors at high pressure. Am. Inst. Chem. Eng. J. **2002**, *48*, 2459.
42. Kundu, A.; Nigam, K.D.P.; Duquenne, A.-M.; Delmas, H. Recent developments on hydroprocessing reactors. Rev. Chem. Eng. **2003**, *19*, 531.
43. Iliuta, I.; Thyrion, F.C.; Bolle, L.; Giot, M. Comparison of hydrodynamic parameters for countercurrent and cocurrent flow through packed beds. Chem. Eng. Technol. **1997**, *20*, 171.
44. Boelhouwer, J.G.; Piepers, H.W.; Drinkenberg, A.A.H. Nature and characteristics of pulsing flow in trickle-bed reactors. Chem. Eng. Sci. **2002**, *57* (22–23), 4865.
45. Sie, S.T.; Krishna, R. Process development and scale up: III. Scale up and scale down of trickle bed processes. Rev. Chem. Eng. **1998**, *14*, 203.
46. Horowitz, G.I.; Cukierman, A.L.; Cassanello, M.C. Flow regime transition in trickle beds packed with particles of different wetting characteristics-check-up on new tools. Chem. Eng. Sci. **1997**, *52*, 3747.
47. Grosser, K.; Carbonell, R.G.; Sundaresan, S. Onset of pulsing in two-phase cocurrent downflow through a packed bed. Am. Inst. Chem. Eng. J. **1988**, *34*, 1850.
48. Saroha, A.K.; Nigam, K.D.P.; Saxena, A.K.; Kapoor, V.K. Liquid distribution in trickle bed reactors. Am. Inst. Chem. Eng. J. **1998**, *44*, 2044.
49. Ng, K.M.; Chu, C.F. Trickle-bed reactors. Chem. Eng. Prog. **1987**, *83*, 55.
50. Al-Dahhan, M.H.; Wu, Y.; Dudukovic, M.P. Reproducible technique for packing laboratory scale trickle-bed reactors with a mixture of catalyst and fines. Ind. Eng. Chem. Res. **1995**, *34*, 741.
51. Ravindra, P.V.; Rao, D.P.; Rao, M.S. Liquid flow texture in trickle-bed reactors: an experimental study. Ind. Eng. Chem. Res. **1997**, *36*, 5133.

52. Kundu, A.; Saroha, A.K.; Nigam, K.D.P. Liquid distribution studies in trickle-bed reactors. Chem. Eng. Sci. **2001**, *56*, 5963.
53. Khanna, R; Nigam, K.D.P. Partial wetting in porous catalysts: wettability and wetting efficiency. Chem. Eng. Sci. **2002**, *57*, 3401.
54. Maiti, R.N.; Sen, P.K.; Nigam, K.D.P. Trickle-bed reactors liquid distribution and flow texture. Rev. Chem. Eng. **2004**, *20*, 57.
55. Mantle, M.D.; Sederman, A.J.; Gladden, L.F. Single and two-phase flow in fixed-bed reactors: MRI flow visualization and Lattice-Boltzmann simulation. Chem. Eng. Sci. **2001**, *56*, 523.
56. Stephenson, J.L.; Stewart, W.E. Optical measurements of porosity and fluid motion in packed beds. Chem. Eng. Sci **1986**, *41*, 2161.
57. Spedding, P.L.; Spencer, R.M. Simulation of packing density and liquid flow in fixed beds. Comput. Chem. Eng **195**, *19*, 43.
58. Herskowitz, M.; Smith, J.M. Liquid distribution in trickle-bed reactors. Am. Inst. Chem. Eng. J. **1978**, *24*, 442.
59. Jacobs, G.E.; Milliken, A.S. Evaluating liquid distributors in hydroprocessing reactors. Hydrocarbon Processing, International Edition **2000**, *79*, 75.
60. Jiang, Y.; Khadilkar, M.R.; Al-Dahhan, M.H.; Dudukovic, M.P. Two-phase flow distribution in 2D trickle-bed reactors. Chem. Eng. Sci. **1999**, *54*, 2409.
61. Marcandelli, C.; Lamine, A.S.; Bernard, J.R.; Wild, G. Liquid distribution in trickle-bed reactor. Oil Gas Sci. Technol.—Rev. IFP **2000**, *55*, 407.

Hydrogels

Jae Hyung Park
Department of Advanced Polymer and Fiber Materials, Kyung Hee University, Gyeonggi-do, South Korea

Kang Moo Huh
Department of Polymer Science and Engineering, Chungnam National University, Daejeon, South Korea

Mingli Ye
Kinam Park
Departments of Pharmaceutics and Biomedical Engineering, Purdue University, West Lafayette, Indiana, U.S.A.

INTRODUCTION

A hydrogel is a three-dimensional polymer network made of a hydrophilic polymer or a mixture of polymers. In general, at least 10–20% of the total weight of a hydrogel is water. When a hydrogel is dried, it is called xerogel, or simply a dried hydrogel. When a dried hydrogel is placed in an aqueous environment, it can absorb a large amount of water and swell isotropically to maintain its original shape. Swollen hydrogels maintain their shape without dissolving even in abundant water because of the presence of chemical or physical cross-linking of polymer chains. The extent of swelling depends inversely on the cross-linking density. When more than 95% of the total weight is water, the hydrogel is also called superabsorbent. It is not unusual to see hydrogels with more than 99% of water.

Because of the presence of high water content and the rubbery property, hydrogels have been frequently compared with the natural tissues. The similarity to natural tissues renders hydrogels useful in biomedical and pharmaceutical applications. Further, depending on the chemical structures of the constituting polymers, hydrogels can be tailored to respond to external stimuli, such as temperature, pH, solvent composition, electric field, light, and specific biomolecules. Those hydrogels can undergo change in swelling/deswelling, shape change, and sol–gel transformation upon stimulation by external factors, and they are often called ''smart hydrogels.'' Such an interesting nature of the smart hydrogels has allowed their use in controlled drug delivery, biomechanical devices, and separation systems.

The advent of nanotechnology has provided new avenues for engineering materials in nano- and microscales. In recent years, there have been extensive studies on potential applications of hydrogels in nanotechnology. Most of the studies have tried to exploit the unique properties of hydrogels, such as the hydrophilic nature of the surface, soft physical properties, and environmental sensitivity. Of particular interest has been nanoscale fabrication and manipulation of hydrogel-based materials that may lead to scientific and technological advances. Here, we review the current technologies on the preparation and potential applications of hydrogel-based nanomaterials, including hydrogel nanoparticles, hydrogel-coated nano/micro devices, inorganic (or organic) nanoparticle-entrapped hydrogels, and molecularly imprinted hydrogels.

HYDROGELS IN NANOTECHNOLOGY

The hydrophilic polymer molecules of a hydrogel are interconnected by cross-linking, and this structure prevents dissolution of the polymer chains in an aqueous solution despite absorption of a large amount of water by the hydrogel. Hydrogels are generally classified into chemical and physical gels, according to the type of cross-links. Chemical gels are produced by cross-linking of hydrophilic polymers via covalent bonding. In an aqueous solution, they absorb water until they reach equilibrium swelling, which depends on the cross-linking density. On the other hand, physical gels are generated by noncovalent bonding, such as molecular entanglements, electrostatic interactions, hydrogen bonding, and hydrophobic interactions. These interactions, in contrast to covalent bonding, are reversible and can be disrupted by changes in physical conditions, such as temperature, pH, ionic strength, and stress.

Fig. 1 illustrates representative mechanisms of hydrogel formation. There are a number of different

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008107

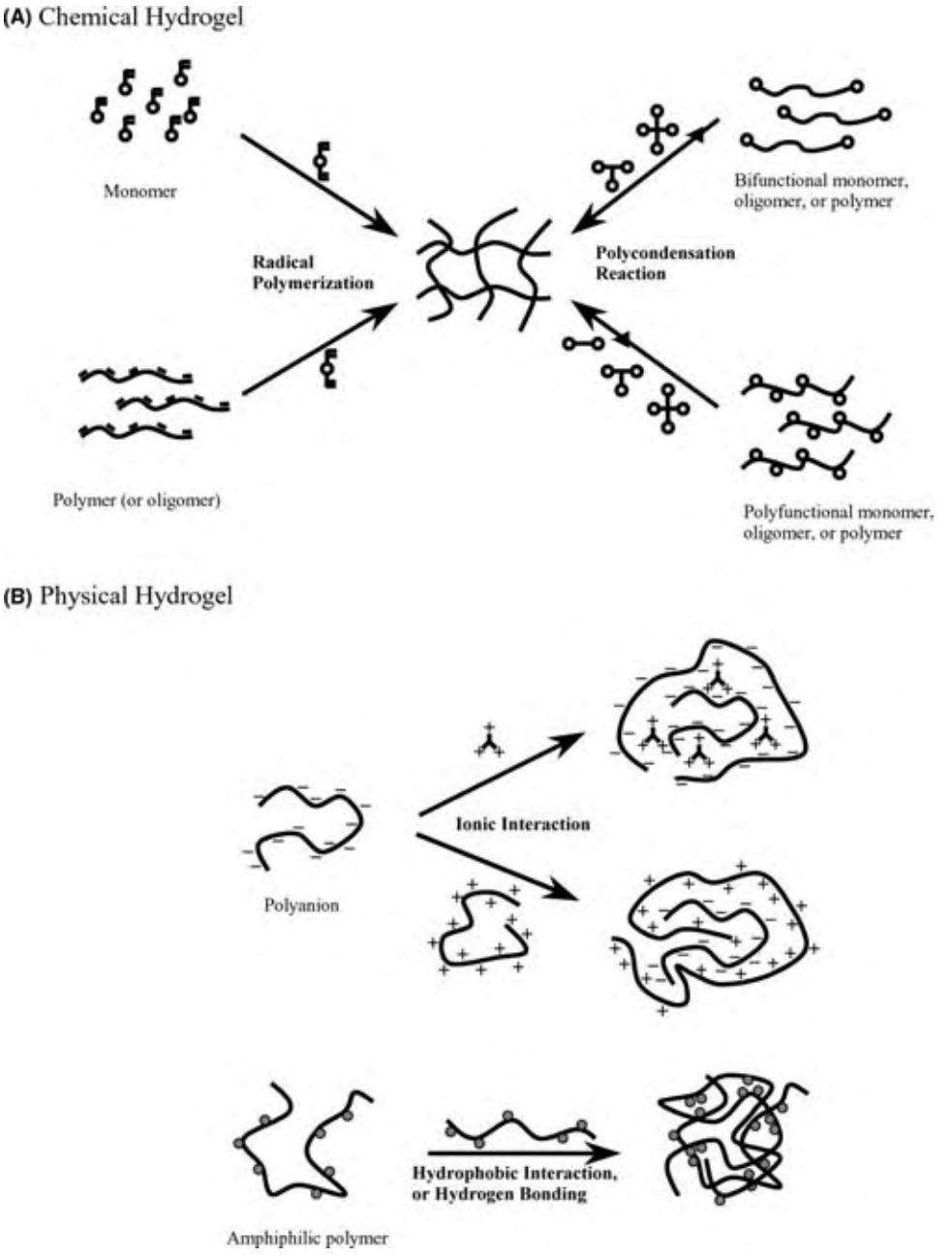


Fig. 1 Representative methods of hydrogel formation. (A) Chemically cross-linked hydrogels are prepared from monomers, oligomers, or polymers in the presence of cross-linking agents. The chemical cross-linking proceeds via radical polymerization or polycondensation reaction. (B) Physically cross-linked hydrogels can be formed by ionic interactions, hydrophobic interaction, or hydrogen bonding.

macromolecular structures that form physical or chemical hydrogels. Hydrogels can be designed to undergo biodegradation in a physiological solution,[1–3] exhibit rapid response to physical stimuli,[4–6] or reach the equilibrium swelling level within a few minutes.[7–9]

Furthermore, there are differences in preparation methods and properties between chemical and physical gels, and each gel type has its own advantages and disadvantages for the design of specific materials or devices involving nanotechnology. Physical gels have been primarily used for biomedical applications, especially in the form of polymer micelles or self-aggregates for controlled drug delivery to the specific sites of action. They are spherical in shape and have the mean diameter ranging from nanometers to micrometers. Because nanoparticles are frequently administrated via the systemic route, it is desirable to construct them with biodegradable polymers so that they can degrade

into low molecular weight entities eligible for renal excretion. Chemical gels that are covalently cross-linked provide a variety of potential applications because of their high stability in harsh environments, such as high temperature, acidic/basic solutions, and high stresses. In addition to drug delivery, chemical hydrogels have been considered as the constituents of diagnostic, electronic, and photonic devices.[10–12]

Nanoparticle-Bearing Hydrogels

Incorporation of functional nanoparticles into a hydrogel matrix produces unique properties, which cannot be found in other conventional organic/inorganic materials. Nanoparticles can be entrapped into a hydrogel matrix by chemical bonding or physical interactions with the polymer backbone of the hydrogel that have rubber elasticity and the stimuli-sensitive swelling/shrinking behavior. The entrapment of polymer nanoparticles, forming a crystalline colloidal array (CCA), into a hydrogel matrix has displayed different colors without adding coloring agents, responding to external stimuli.[10,13] By incorporating 10–12 nm magnetite (Fe_3O_4) or maghemite (γ-Fe_2O_3) into a hydrogel matrix, the shape of hydrogels can be modulated by applying magnetic field.[14,15]

Monodispersed colloidal particles are reported to form CCAs via slow particle sedimentation, centrifugation, and spin coating.[10,13,16] The CCA diffracts visible and near-infrared light at wavelengths dependent on the lattice spacing, which produces an intense color. By combining the characteristic of the CCA with stimuli-sensitive hydrogels, novel functional hydrogels have recently been prepared. Of various hydrophilic polymers, poly(*N*-isopropylacrylamide) (PNIPAM) and its copolymers have been widely used to prepare chemical gels containing CCAs, because they show fast stimuli-sensitive volume phase transition and are readily prepared by the free radical polymerization of monomers in the presence of a difunctional cross-linking agent.[10–12,14] Weissman et al.[10] prepared PNIPAM hydrogels by photopolymerization in the presence of polystyrene nanoparticles (99 nm in diameter) as a CCA component and *N*,*N*-methylene-*bis*-acrylamide as a cross-linking agent. This chemically cross-linked CCA hydrogels exhibited various colors according to the temperature that changes the hydrogel volume affecting the array lattice constant. Polyacrylamide-based hydrogels have also been investigated to entrap the CCA lattice.[13] Although polyacrylamide does not provide thermo-sensitive volume transition like PNIPAM, it allows incorporating molecular-recognition groups (e.g., crown ethers for metal ions and glucose oxidase for glucose) during its polymerization or by simple chemical modification. The recognition events may make the gel swell because of changes in environments within the hydrogel matrix, such as an osmotic pressure and pH, which increases the mean separation between the colloidal spheres, and thus changes the diffracted light to longer wavelengths. More recently, the use of monodispersed nanoparticles as a hydrogel matrix has been developed.[11,16,17] This technique involves preparation of monodispersed hydrogel nanoparticles by radial polymerization in the presence of the cross-linker and surfactant, and formation of nanoparticle networks (colloidal crystal gels) by centrifugation[16] or chemical cross-linking.[11,17] The resultant hydrogel matrices displayed a bright iridescence in the visible region of the spectrum and underwent a reversible color change in response to changes in the temperature or pH. For chemically cross-linked nanoparticle networks, the color was changed depending on the concentration and size of the nanoparticles.

Nanoparticles have also been incorporated into a hydrogel matrix to develop novel materials sensitive to stimuli, such as light[18] and magnetic fields.[14] The gold nanoshell, composed of a thin layer of gold surrounding a dielectric core (e.g., gold sulfide), is one of the representative nanoparticles that have been incorporated into chemical gels, resulting in unique optical properties.[19–21] The diameters of both the core and shell are known to be responsible for the optical properties of nanoshells.[19,20] As the core size and shell thickness can be readily manipulated during the fabrication process, the optical extinction profiles of the nanoshells can be adjusted to observe light at desired wavelengths. Recently, by using nanoshells which absorb near-infrared light, hydrogel/nanoshell composites have been prepared for photothermally modulated drug delivery.[22] Embedding the nanoshells in PNIPAM hydrogel produced unique properties by which the composite hydrogel exhibited volume transition upon the irradiation of near-infrared light, i.e., the nanoshells in the hydrogel matrix converted light to heat, raising the temperature of the composite above the low critical solution temperature of PNIPAM. Because near-infrared light is capable of being transmitted through tissue, such hydrogels bearing gold nanoshells may have promising potential as an injectable drug delivery system for low molecular weight drugs, peptides, proteins, and genes.[18,19,21,22]

Molecularly Imprinted Hydrogels

Design of a precise macromolecular architecture that can selectively recognize target molecules has gained significant attention because of its potential applications for separation processes, immunoassays,

biosensors, and catalysis.[23,24] Molecular imprinting technology has been developed as a response to the need to create such architectures. In general, molecular imprinting within polymers involves formation of prepolymerization complex between the template molecule and functional monomers, polymerization in the presence of a cross-linking agent and an appropriate solvent, and removal of the template.[25] Once the template is removed, the polymer network may have specific recognition elements for the target (or template) molecules. This nanoimprinting process is of great importance because it can create three-dimensional binding cavities for specific target molecules.

Because most recognition processes are associated with three-dimensional structure of the recognition site, it is preferable to limit the movement of the polymer chain that may affect affinity or selectivity to target molecules. Therefore, conventional methods to prepare molecularly imprinted polymers have used high ratios of the cross-linking agent to functional monomers, which leads to formation of rigid polymer matrix with low average molecular mass between cross-links.[26] On the contrary, imprinting within hydrogels requires different methods because they undergo changes in three-dimensional structure upon coming in contact with water. To maintain imprinting structure in an aqueous environment, hydrogels have been prepared by spatially varying cross-linking density.[25,27] As density fluctuations in the polymer network include microregions of localized higher cross-linking, hydrogels could retain an effective imprinting structure as well as proper rigidity to produce adequate specificity. Another promising strategy for imprinting within hydrogels is to match polymerization and rebinding solvents in terms of dielectric constant, polarity, and protic nature. This may reduce differences in swelling behavior, resulting in high binding affinity to target molecules. Also, in designing the network architecture for hydrogels, it is important to choose the length of the functional monomer and the molecular mass of the cross-linking monomer to endow with specificity to target molecules.[25,28]

Molecular imprinting provides shape-specific cavities (or nanovacuoles) that match the template molecule or chemical groups capable of specifically interacting with the template molecule. Of various polymers, ethylene glycol dimethyacrylate and methacrylic acid have been most widely used for the formation of imprinted polymers, where template molecules can be antibiotics,[29,30] carbohydrates,[19] peptides,[31] and enzymes.[32] Alvarez-Lorenzo et al.[33,34] and Hiratani et al.[35] have developed molecularly imprinted hydrogels by polymerizing in the presence of template molecule, functional monomer, and thermo-sensitive monomer. These hydrogels showed high affinity to target molecules as well as stimuli-sensitive recognition, by which the imprinted sites disappeared upon gel swelling and reformed upon shrinking.

The imprinted hydrogels, sensitive to analyte, have been the focus of many investigations for controlled drug release. A few examples are shown in Fig. 2. Fig. 2A shows that enzymes may be included in the hydrogel to invoke local pH changes by binding to analyte, and thus initiate the hydrogel swelling, modulating the drug release rate;[36,37] in Fig. 2B, cross-links can lose their function by free analytes and this leads to hydrogel swelling to change the drug release rate;[38,39] and in Fig. 2C, binding of analytes to specific sites or functional groups on the polymer backbone may change hydrophilicity (or hydrophobicity) of hydrogels, inducing swelling (or shrinking) in an aqueous environment.[40] It should be emphasized, however, that the currently available imprinting techniques involve nonspecific binding sites that decrease the specificity to target molecules. The precise control of the network structure of hydrogels via nanotechnology will contribute to a number of applications, including microfluidic devices, biomimetic sensors, drug delivery system, and membrane separation technology.

Hydrogel Nanoparticles

Because of their biocompatibility and soft rubbery nature, hydrogels have been extensively studied in biomedical and pharmaceutical fields. Macroscopic hydrogels have been studied for sustained drug delivery owing to their slow swelling kinetics. Microparticulate hydrogels for drug delivery have been examined more widely because of their unique properties resulting from small size. Initially, hydrophobic nano- and microparticles have been used extensively for systemic drug delivery, but it was soon found out that they were readily taken up by the reticuloendothelial system, and thus exhibited short residence time in blood. Thus, in an attempt to improve hydrophilicity for prolonged circulation time, the surfaces of the hydrophobic nanoparticles have been modified by conjugating, blending, and coating with hydrophilic polymers, such as polyethylene glycol (PEG) and PEG-containing block copolymers.[41,42]

Hydrogel nanoparticles have a special role in drug delivery in the sense that they have all the advantages of both nanoparticles and hydrogels with regard to the particle size and hydrophilicity. Hydrogel nanoparticles can swell rapidly because of large surface area and short diffusion path length for water. They can be modified to have reactive groups on the surface and are useful to introduce functional moieties, such as targeting and other bioactive moieties. Hydrogel nanoparticles consisting of stimuli-responsive polymers may exhibit corresponding responsive properties, which

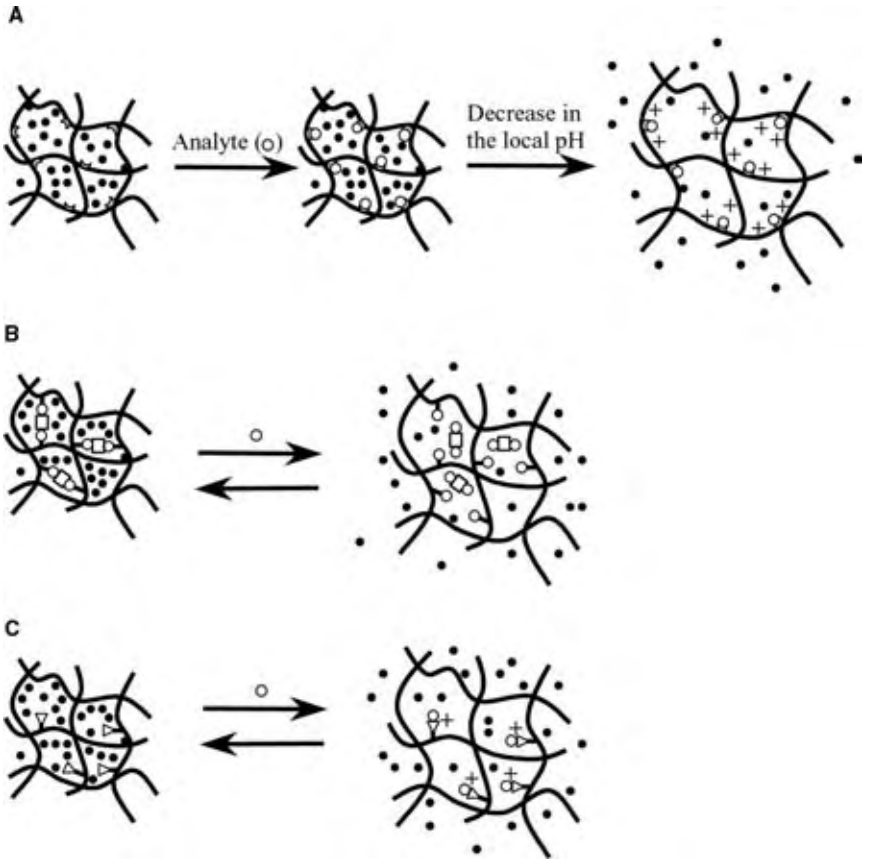


Fig. 2 Molecularly imprinted hydrogels for drug delivery. (A) Binding of analytes (○) to enzyme (ℭ) induces changes in the local pH. For cationic hydrogels, the acidic local pH results in ionization and swelling of hydrogel, resulting in faster release of drug (●). (B) Cross-linking agents (□) bind to analytes (○) anchored to the polymer backbone to maintain the hydrogel structure, which can swell and release the incorporated drugs as the concentration of free analyte increases, replacing the polymer-bound analytes. (C) Binding of analyte (○) of specific site or functional groups (▽) on the polymer backbone increases hydrophilicity of hydrogel, which induces swelling and drug release. (Modified from Ref.[26].)

are often found to become much faster than bulk hydrogels. Studies of hydrogel nanoparticle have intensified during the past decade because of enormous potentials in the development and implementation of new stimuli-responsive or smart materials, biomimetics, biosensors, artificial muscles, drug delivery systems, and chemical separation systems.

Chemically cross-linked hydrogel nanoparticles have been prepared in the presence of hydrophilic monomers, cross-linking agents, and emulsifiers. Advances in technology have enabled precise control of the core–shell structure of hydrogel nanoparticles.[43] The core–shell nanoparticles have been synthesized to modify surface properties of core particles or to provide stimuli-sensitivity for nonresponsive particles. Recently, multiresponsive core–shell hydrogel nanoparticles have also been developed by Jones and Lyon[43] and Gan and Lyon[44] They synthesized temperature- and pH-responsive hydrogel nanoparticles with core–shell morphologies, where core particles composed of PNIPAM were prepared via aqueous free radical polymerization, and then used as nuclei for subsequent polymerization of acrylic acid copolymers. Their swelling/deswelling thermodynamics were easily controlled by chemical manipulation of the core and shell structures, thus displaying both temperature and pH dependence. Without chemical cross-linking, core–shell hydrogel nanoparticles can also be prepared on the basis of the electrostatic interaction between water-soluble polymers.[45] Prokop et al.[46] have demonstrated that by atomizing the aqueous solution containing core polymer with negative charge, the nanosize droplets are encapsulated with cationic polymer solution by the electrostatic interaction (Fig. 3). This methodology showed promising potential as a protein delivery system.

Self-assembled hydrogel nanoparticles based on hydrophobically modified polysaccharides have been extensively studied as a drug carrier because of their excellent biocompatibility and ease of preparation. It is well known that polymeric amphiphiles, upon contact with an aqueous environment, spontaneously form micelles or micelle-like self-aggregates via undergoing intra- or intermolecular associations between hydrophobic moieties, primarily to minimize interfacial free energy. Hydrophobically modified polysaccharides are also known to self-assemble in aqueous media to form a unique core–shell structure that consists of hydrophobic segments and hydrophilic segments, respectively (Fig. 4). This type of hydrogels have multiple inner cores, which physically crosslink the hydrophilic polymer chains.[47] A number of polysaccharides have been investigated to create self-assembling systems, including dextran,[48] glycol chitosan,[49,50]

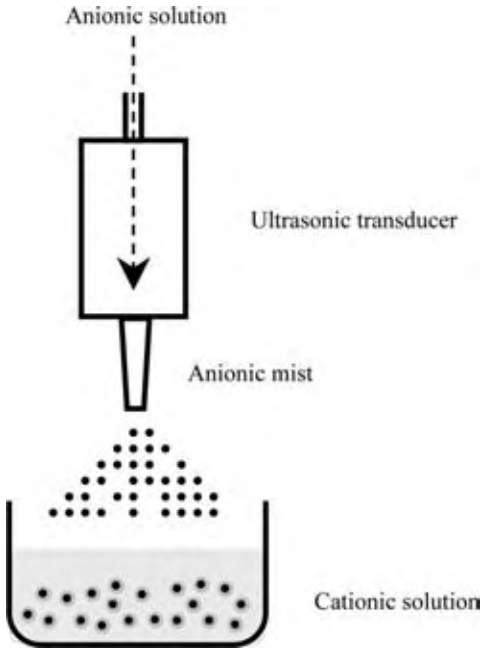


Fig. 3 Formation of the core–shell type hydrogel nanoparticles by electrostatic interactions. The anionic solution which contains core polymer is introduced as a mist into a cationic solution of shell polymer. (Modified from Ref.[45].)

pullulan,[51,52] and curdlan,[53] These polysaccharides are natural water-soluble polymers that are inherently biocompatible and biodegradable. The core–shell structure of self-assembled hydrogels can be employed as a potential delivery system that can effectively deliver hydrophobic drugs. It has been recently demonstrated that hydrophobically modified polysaccharides capable of forming nano-sized self-aggregates can imbibe hydrophobic drugs and release them in a sustained manner.[49] Hydrophobic moieties, conjugated to

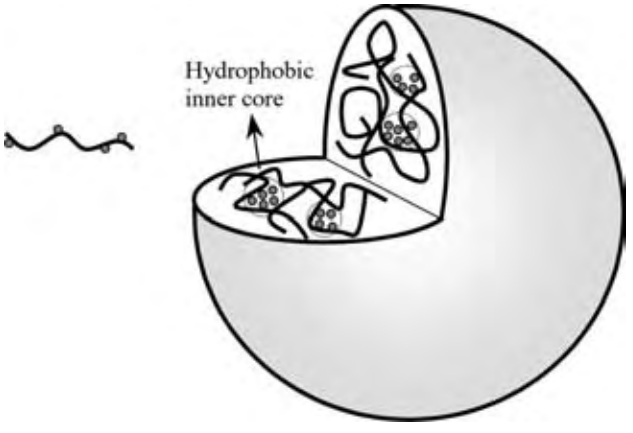


Fig. 4 Self-assembled hydrogel nanoparticles of hydrophobically modified polysaccharides. Note that nanoparticles have multiple inner cores which physically cross-link the hydrophilic polysaccharide chain. (Modified from Ref.[47].)

polysaccharides, can either be small molecules (e.g., cholesterol, alkyl chains, and bile acids)[47,51,53–55] or oligomers.[56] The conjugation of stimuli-sensitive hydrophobic moieties to polysaccharides may produce hydrogel nanoparticles, responsive to corresponding stimuli.[57] For example, Na et al.[53,56,57] have recently developed pH-sensitive hydrogel nanoparticles as an anticancer drug carrier. The extracellular pH of most solid tumors and inflammatory regions in the body is known to be lower than that in the normal tissues and blood (pH 7.4). To target the extracellular matrix of such disease sites, they prepared pullulan acetate-based nanoparticles bearing sulfonamide moieties, which show the hydrophobic nature at the low pH. The resulting nanoparticles rapidly released the anticancer drug (doxorubicin) at pH <7.0, whereas the drug release rate was substantially reduced at normal tissue pH (7.4).

Hydrogel Coating on the Surfaces

Surfaces of hydrophobic substrates have been frequently modified with hydrophilic polymers to achieve desirable properties for in vivo applications. Surface modification with hydrophilic polymers is known to minimize nonspecific interactions with blood proteins, cells, and tissues. The hydrophilic polymers commonly used for surface coating include PEG, polysaccharides, and poly(vinyl alcohol). To physically coat hydrophilic polymers on the nanoparticles, the solvent extraction/evaporation method has been used.[58,59] After the oil-in-water emulsion is prepared by adding the organic solution containing a hydrophobic polymer into the aqueous solution of a hydrophilic homopolymer or a block copolymer, the organic solvent is removed by evaporation or extraction, thus forming the hydrogel layer on the formed nanoparticles. The outer layer of a hydrophilic polymer in the nanoparticle is anchored by various interactions with the core polymer chains, such as physical entanglement, hydrophobic interaction, and hydrogen bonding. This approach has been frequently used for surface modification of biodegradable nanoparticles, such as poly(D,L-lactide) (PLA),[58] poly(lactide-co-glycolide),[59] and polyphosphazene.[60]

The hydrogel coating on the nanoparticles has also been achieved by radical polymerization. For poly (isobutyl cyanoacrylate) (PIBCA), the monomer was emulsified in an aqueous solution containing PEG that acts as a nucleophile initiator of polymerization through its hydroxyl terminal groups.[61] Once the aqueous pH is adjusted to 1, polymerization is initiated, thus forming PEG-coated PIBCA nanoparticles.[62,63] PNIPAM has been coated on the nanoparticles by radical polymerization in the presence of hydrophobic

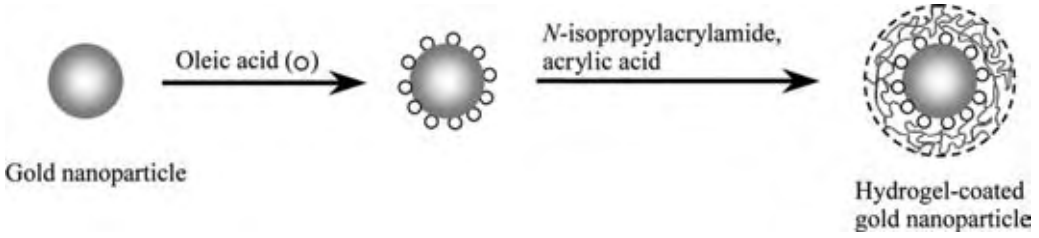


Fig. 5 Hydrogel-coated gold nanoparticles prepared by surfactant-free emulsion polymerization. After coating oleic acids on the gold nanoparticle, polymerization was carried out in the presence of *N*-isopropylacrylamide, acrylic acid, and ammonium persulfate (initiator). The size of resulting nanoparticles was in the range of 100–230 nm. (Modified from Ref.[69].)

nanoparticles to be coated, *N,N'*-methylenebisacrylamide (cross-linker), ammonium persulfate (initiator), and sodium dodecyl sulfate (emulsifier). The thickness of the outer hydrogel layer on the nanoparticles was readily controlled by varying the concentrations of the monomer and emulsifier.[64]

Hydrogels that coat metal and semiconductor nanoparticles are of considerable interest because of their unique size-dependent physicochemical properties.[65–67] Precise control of the structure and surface properties of nanoparticle would make them more attractive for use in biomedical applications. Inorganic nanoparticles have been conjugated to biomolecules such as sugars, peptides, proteins, and DNA. Such conjugates showed many advantages as fluorescent biological labels,[66–68] primarily appearing from inorganic nanoparticles, including high quantum efficiencies, optical activity over biocompatible wavelengths, and chemical or photochemical stability. It should be emphasized that in spite of numerous potential applications, inorganic

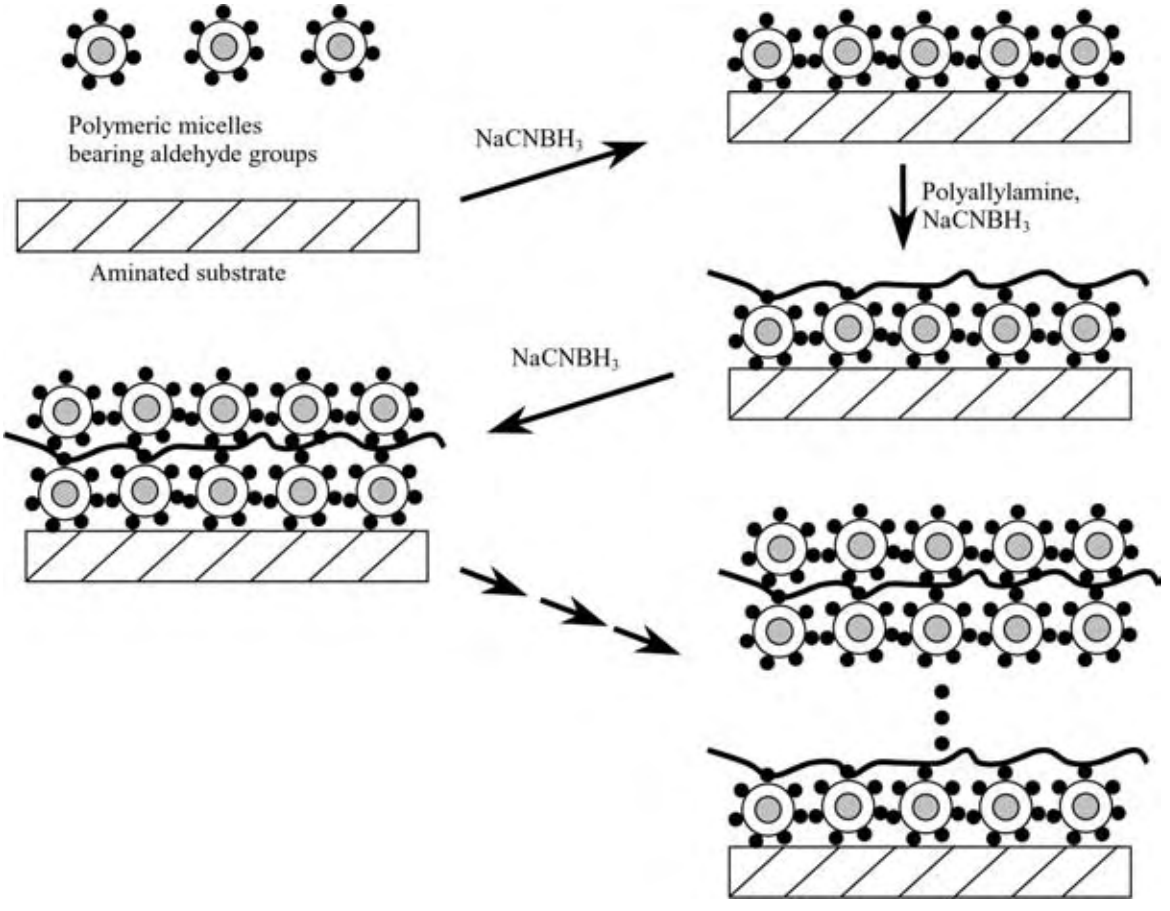


Fig. 6 Schematic illustration of the multilayered micellar coating on the surface. Polymeric micelles to be coated were first stabilized by polymerization of the hydrophobic inner core. The stable micelles were then immobilized on the aminated substrate by the reaction with aldehyde groups on the surface of polymeric micelle. (Modified from Ref.[74].)

nanoparticles have suffered from their aggregation and lack of biocompatibility. Hydrogel coating on such nanoparticles may not only prevent their aggregation by changing the surface hydrophilicity, but also improve their biocompatibility. Furthermore, use of stimuli-responsive hydrogels may provide unique properties for nanoparticles. Recent efforts have led to the development of hydrogel-coated inorganic nanoparticles that exhibit structural changes responsive to stimuli such as light. For example, hydrogel-coated gold nanoparticles have been prepared using surfactant-free emulsion polymerization method, as shown in Fig. 5.[69] The hydrogel layer was constructed with a mixture of *N*-isopropylacrylamide and acrylic acid, and its thickness could be varied by adjusting the amount of monomer and initiator, as well as the reaction time. The results revealed that the hydrogel can be thermally activated by exposure to light via the strong plasmon absorption of the gold nanoparticle core.

As mentioned earlier, the surface coating with hydrogel can improve the biocompatibility as well as provide specific functions. One of the promising strategies to improve the surface characteristics is to attach polymeric micelles onto the surfaces, thus forming the polymeric micelle-entrapped hydrogel layer. This approach is useful to maximize the number of tethered hydrophilic chains because the polymeric micelle has a high density of hydrophilic polymer on the surface, resulting in an effective nonfouling property. Further, as the polymeric micelles contain hydrophobic inner core as a reservoir of hydrophobic drugs, the surface coating with polymeric micelles may allow developing of biocompatible devices that can release the drug in a sustained manner. The structure of the polymeric micelles, however, is readily disrupted upon attachment to the surface, leading to the formation of a loosely packed layer structure.[70,71] In an attempt to stabilize the polymeric micelles, Ijima et al.[72] prepared heterobifunctional block copolymer of PEG–PLA, in which PEG had a reactive aldehyde group at the chain end, whereas PLA possessed a methacryloyl group that can be polymerized in the presence of the initiator. This amphiphilic copolymer was then exposed to an aqueous solution, which enabled it to form the polymeric micelle, followed by the polymerization of the hydrophobic inner core. The resulting micelles showed high stability in harsh environments.[72] The aldehyde groups at the end of PEG chain was used to chemically attach to the surfaces bearing the amino groups so that a single layer of polymer micelle is formed on the surface.[73] By introducing amino groups on the top of the micellar layer through tethering polyallylamine, multilayered highly organized micellar hydrogel can be coated on the surfaces.[74] Fig. 6 shows the schematic illustration of the formation of the multilayered micellar coating on the surface. The resulting surface with micellar hydrogel layers exhibited excellent resistance to protein adsorption. In addition, the incorporation of the hydrophobic drug into the micellar hydrophobic core (~10 nm in diameter) made it possible to release the drug in a controlled manner, depending on the number of coated layers.[75]

CONCLUSIONS

Recent advances in nanotechnology have enabled us to extend potential applications of micro- and nanoparticulate hydrogels. Combination of hydrogel and nanotechnology may afford a powerful means for manipulating the properties of surfaces and interfaces. Fabrication of nanostructures using hydrogels involves hydrophilic nanoparticles, molecular imprinting, nanoparticle-entrapped hydrogel, and nanoengineering for surface modification. These technologies will accelerate the development of various drug delivery and biomedical devices, as well as other electronic and photonic devices.

ACKNOWLEDGMENTS

This study was supported in part by the National Institute of Health through GM67044 and GM65284.

REFERENCES

1. Jeong, B.; Bae, Y.H.; Lee, D.S.; Kim, S.W. Biodegradable block copolymers as injectable drug-delivery systems. Nature **1997**, *388*, 860–862.
2. Kissel, T.; Li, Y.; Unger, F. ABA–triblock copolymers from biodegradable polyester A-blocks and hydrophilic poly(ethylene oxide) B-blocks as a candidate for in situ forming hydrogel delivery systems for proteins. Adv. Drug Deliv. Rev. **2002**, *54*, 99–134.
3. Holland, T.A.; Tessmar, J.K.; Tabata, Y.; Mikos, A.G. Transforming growth factor-beta 1 release from oligo(poly(ethylene glycol) fumarate) hydrogels in conditions that model the cartilage wound healing environment. J. Control. Release **2004**, *94*, 101–114.
4. Wang, C.; Stewart, R.J.; Kopecek, J. Hybrid hydrogels assembled from synthetic polymers and coiled-coil protein domains. Nature **1999**, *397*, 417–420.
5. Qiu, Y.; Park, K. Environment-sensitive hydrogels for drug delivery. Adv. Drug Deliv. Rev. **2001**, *53*, 321–339.
6. Kopecek, J. Smart and genetically engineered biomaterials and drug delivery systems. Eur. J. Pharm. Sci. **2003**, *20*, 1–16.

7. Chen, J.; Park, H.; Park, K. Synthesis of superporous hydrogels: hydrogels with fast swelling and superabsorbent properties. J. Biomed. Mater. Res. **1999**, *44*, 53–62.
8. Chen, J.; Park, K. Synthesis and characterization of superporous hydrogel composites. J. Control. Release **2000**, *65*, 73–82.
9. Gemeinhart, R.A.; Park, H.; Park, K. Effect of compression on fast swelling of poly(acrylamide-co-acrylic acid) superporous hydrogels. J. Biomed. Mater. Res. **2001**, *55*, 54–62.
10. Weissman, J.M.; Sunkara, H.B.; Tse, A.S.; Asher, S.A. Thermally switchable periodicities and diffraction from mesoscopically ordered materials. Science **1996**, *274*, 959–960.
11. Hu, Z.; Lu, X.; Gao, J.; Wang, C. Polymer gel nanoparticle networks. Adv. Mater. **2000**, *12*, 1173–1176.
12. Jones, C.D.; Lyon, L.A. Photothermal patterning of microgel/gold nanoparticle composite colloidal crystals. J. Am. Chem. Soc. **2003**, *125*, 460–465.
13. Holtz, J.H.; Asher, S.A. Polymerized colloidal crystal hydrogel films as intelligent chemical sensing materials. Nature **1997**, *389*, 829–832.
14. Xulu, P.M.; Filipcsei, G.; Zrinyi, M. Preparation and responsive properties of magnetically soft poly(*N*-isopropylacrylamide) gels. Macromolecules **2000**, *33*, 1716–1719.
15. Shipway, A.N.; Willner, I. Nanoparticles as structural and functional units in surface-confined architectures. Chem. Commun. (Camb.) **2001**, 2035–2045.
16. Debord, J.D.; Lyon, L.A. Thermoresponsive photonic crystals. J. Phys. Chem. B **2000**, *104*, 6327–6331.
17. Hu, Z.; Lu, X.; Gao, J. Hydrogel opals. Adv. Mater. **2001**, *13*, 1708–1712.
18. Sershen, S.; West, J. Implantable, polymeric systems for modulated drug delivery. Adv. Drug Deliv. Rev. **2002**, *54*, 1225–1235.
19. Averitt, R.D.; Sarkar, D.; Halas, N.J. Plasmon resonance shifts of Au coated Au_2S nanoshells: insight into multicomponent nanoparticle growth. Phys. Rev. Lett. **1997**, *78*, 4217–4220.
20. Averitt, R.D.; Westcott, S.L.; Halas, N.J. Linear optical properties of gold nanoshells. J. Opt. Soc. Am. B **1999**, *16*, 1824–1832.
21. Sershen, S.R.; Westcott, S.L.; Halas, N.J.; West, J.L. An optomechanical nanoshell-polymer composite. Appl. Phys. B **2001**, *73*, 1–3.
22. Sershen, S.R.; Westcott, S.L.; Halas, N.J.; West, J.L. Temperature-sensitive polymer-nanoshell composites for photothermally modulated drug delivery. J. Biomed. Mater. Res. **2000**, *51*, 293–298.
23. Takeuchi, T.; Haginaka, J. Separation and sensing based on molecular recognition using molecularly imprinted polymers. J. Chromatogr. B Biomed. Sci. Appl. **1999**, *728*, 1–20.
24. Piletsky, S.A.; Alcock, S.; Turner, A.P. Molecular imprinting: at the edge of the third millennium. Trends Biotechnol. **2001**, *19*, 9–12.
25. Langer, R.; Peppas, N.A. Advances in biomaterials, drug delivery, and bionanotechnolgy. AIChE J. **2003**, *49*, 2990–3006.
26. Byrne, M.E.; Park, K.; Peppas, N.A. Molecular imprinting within hydrogels. Adv. Drug Deliv. Rev. **2002**, *54*, 149–161.
27. Bures, P.; Huang, Y.; Oral, E.; Peppas, N.A. Surface modifications and molecular imprinting of polymers in medical and pharmaceutical applications. J. Control. Release **2001**, *72*, 25–33.
28. Hilt, J.Z.; Byrne, M.E. Configurational biomimesis in drug delivery: molecular imprinting of biologically significant molecules. Adv. Drug Deliv. Rev. **2004**, *56*, 1599–1620.
29. Lubke, C.; Lubke, M.; Whitcombe, M.J.; Vulfson, E.N. Imprinted polymers prepared with stoichiometric template-monomer complexes: efficient binding of ampicillin from aqueous solutions. Macromolecules **2000**, *33*, 5098–5105.
30. Lai, E.P.C.; Wu, S.G. Molecularly imprinted solid phase extraction for rapid screening of cephalexin in human plasma and serum. Anal. Chim. Acta **2003**, *481*, 165–174.
31. Kempe, M. Antibody-mimicking polymers as chiral stationary phases in HPLC. Anal. Chem. **1996**, *68*, 1948–1953.
32. Piletsky, S.A.; Piletsky, E.V.; Elgersma, A.V.; Yano, K.; Karube, I. Atrazine sensing by molecularly imprinted membranes. Biosens. Bioelectron. **1995**, *10*, 959–964.
33. Alvarez-Lorenzo, C.; Guney, O.; Oya, T.; Sakiyama, T.; Takeoka, Y.; Ito, K.; Wang, G.; Annaka, M.; Hara, K.; Du, R.; Chuang, J.; Wasserman, K.; Grosberg, A.Y.; Masamune, S.; Tanaka, T. Polymer gels that memorize elements of molecular conformation. Macromolecules **2000**, *33*, 8693–8697.
34. Alvarez–Lorenzo, C.; Hiratani, H.; Tanaka, K.; Stancil, K.; Grosberg, A.Y.; Tanaka, T. Simultaneous multiple-point adsorption of aluminum ions and charged molecules by a polyampholyte thermosensitive gel: controlling frustrations in a heteropolymer gel. Langmuir **2001**, *17*, 3616–3622.
35. Hiratani, H.; Alvarez–Lorenzo, C.; Chuang, J.; Guney, O.; Grosberg, A.Y.; Tanaka, K. Effect of reversible cross–linker, *N*,*N*-bis(acryloyl)cystamine, on calcium ion adsorption by imprinted gels. Langmuir **2001**, *17*, 4431–4436.

36. Podual, K.; Doyle, F.J., III; Peppas, N.A. Dynamic behavior of glucose oxidase-containing microparticles of poly(ethylene glycol)-grafted cationic hydrogels in an environment of changing pH. Biomaterials **2000**, *21*, 1439–1450.
37. Podual, K.; Doyle, F.J., III; Peppas, N.A. Glucose-sensitivity of glucose oxidase-containing cationic copolymer hydrogels having poly(ethylene glycol) grafts. J. Control. Release **2000**, *67*, 9–17.
38. Obaidat, A.A.; Park, K. Characterization of glucose dependent gel–sol phase transition of the polymeric glucose-concanavalin A hydrogel system. Pharm. Res. **1996**, *13*, 989–995.
39. Obaidat, A.A.; Park, K. Characterization of protein release through glucose-sensitive hydrogel membranes. Biomaterials **1997**, *18*, 801–806.
40. Kataoka, K.; Miyazaki, H.; Bunya, M.; Okano, T.; Sakurai, Y. Totally synthetic polymer gels responding to external glucose concentration: their preparation and application to on-off regulation of insulin release. J. Am. Chem. Soc. **1998**, *120*, 12,694–12,695.
41. Storm, G.; Belliot, S.O.; Daemen, T.; Lasic, D.D. Surface modification of nanoparticles to oppose uptake by the mononuclear phagocyte system. Adv. Drug Deliv. Rev. **1995**, *17*, 31–48.
42. Otsuka, H.; Nagasaki, Y.; Kataoka, K. PEGylated nanoparticles for biological and pharmaceutical applications. Adv. Drug Deliv. Rev. **2003**, *55*, 403–419.
43. Jones, C.D.; Lyon, L.A. Synthesis and characterization of multiresponsive core–shell microgels. Macromolecules **2000**, *33*, 8301–8306.
44. Gan, D.; Lyon, L.A. Tunable swelling kinetics in core–shell hydrogel nanoparticles. J. Am. Chem. Soc. **2001**, *123*, 7511–7517.
45. Prokop, A.; Holland, C.A.; Kozlov, E.; Moore, B.; Tanner, R.D. Water-based nanoparticulate polymeric system for protein delivery. Biotechnol. Bioeng. **2001**, *75*, 228–232.
46. Prokop, A.; Kozlov, E.; Newman, G.W.; Newman, M.J. Water-based nanoparticulate polymeric system for protein delivery: permeability control and vaccine application. Biotechnol. Bioeng. **2002**, *78*, 459–466.
47. Akiyoshi, K.; Deguchi, S.; Tajima, H.; Nishikawa, T.; Sunamoto, J. Microscopic structure and thermoresponsiveness of a hydrogel nanoparticle by self-assembly of a hydrophobized polysaccharide. Macromolecules **1997**, *30*, 857–861.
48. Kim, I.S.; Jeong, Y.I.; Kim, S.H. Self-assembled hydrogel nanoparticles composed of dextran and poly(ethylene glycol) macromer. Int. J. Pharm. **2000**, *205*, 109–116.
49. Son, Y.J.; Jang, J.-S.; Cho, Y.W.; Chung, H.; Park, R.-W.; Kwon, I.C.; Kim, I.-S.; Park, J.Y.; Seo, S.B.; Park, C.R.; Jeong, S.Y. Biodistribution and anti-tumor efficacy of doxorubicin loaded glycol-chitosan nanoaggregates by EPR effect. J. Control. Release **2003**, *91*, 135–145.
50. Park, J.H.; Kwon, S.; Nam, J.-O.; Park, R.-W.; Chung, H.; Seo, S.B.; Kim, I.-S.; Kwon, I.C.; Jeong, S.Y. Self-assembled nanoparticles based on glycol chitosan bearing 5beta-cholanic acid for RGD peptide delivery. J. Control. Release **2004**, *95*, 579–588.
51. Akiyoshi, K.; Kobayashi, S.; Shichibe, S.; Mix, D.; Baudys, M.; Wan Kim, S.; Sunamoto, J. Self–assembled hydrogel nanoparticle of cholesterol-bearing pullulan as a carrier of protein drugs: Complexation and stabilization of insulin. J. Control. Release **1998**, *54*, 313–320.
52. Gupta, M.; Gupta, A.K. Hydrogel pullulan nanoparticles encapsulating pBUDLacZ plasmid as an efficient gene delivery carrier. J. Control. Release **2004**, *99*, 157–166.
53. Na, K.; Park, K.-H.; Kim, S.W.; Bae, Y.H. Self-assembled hydrogel nanoparticles from curdlan derivatives: characterization, anti-cancer drug release and interaction with a hepatoma cell line (HepG2). J. Control. Release **2000**, *69*, 225–236.
54. Kwon, S.; Park, J.H.; Chung, H.; Kwon, I.C.; Jeong, S.Y.; Kim, I.S. Physicochemcial characteristics of self-assembled nanoparticles based on glycol chitosan bearing 5beta-cholanic acid. Langmuir **2003**, *19*, 10,188–10,193.
55. Park, J.H.; Kwon, S.; Nam, J.O.; Park, R.W.; Chung, H.; Seo, S.B.; Kim, I.S.; Kwon, I.C.; Jeong, S.Y. Self-assembled nanoparticles based on glycol chitosan bearing 5beta-cholanic acid for RGD peptide delivery. J. Control. Release **2004**, *95*, 579–588.
56. Na, K.; Lee, K.H.; Bae, Y.H. pH-sensitivity and pH-dependent interior structural change of self-assembled hydrogel nanoparticles of pullulan acetate/oligo-sulfonamide conjugate. J. Control. Release **2004**, *97*, 513–525.
57. Na, K.; Bae, Y.H. Self-assembled hydrogel nanoparticles responsive to tumor extracellular pH from pullulan derivative/sulfonamide conjugate: characterization, aggregation, and adriamycin release in vitro. Pharm. Res. **2002**, *19*, 681–688.
58. Landry, F.; Bazile, D.; Spenlehauer, G.; Veillard, M.; Kreuter, J. Release of the fluorescent marker prodan from poly(D,L-lactic acid) nanoparticles coated with albumin or polyvinyl alcohol in model digestive fluids (USP XXII). J. Control. Release **1997**, *44*, 227–236.
59. Win, K.Y.; Feng, S.S. Effects of particle size and surface coating on cellular uptake of polymeric nanoparticles for oral delivery of anticancer drugs. Biomaterials **2005**, *26*, 2713–2722.

60. Vandorpe, J.; Schacht, E.; Dunn, S.; Hawley, A.; Stolnik, S.; Davis, S.S.; Garnett, M.C.; Davies, M.C.; Illum, L. Long circulating biodegradable poly(phosphazene) nanoparticles surface modified with poly(phosphazene)–poly(ethylene oxide) copolymer. Biomaterials **1997**, *18*, 1147–1152.
61. Peracchia, M.T.; Vauthier, C.; Puisieux, F.; Couvreur, P. Development of sterically stabilized poly(isobutyl 2-cyanoacrylate) nanoparticles by chemical coupling of poly(ethylene glycol). J. Biomed. Mater. Res. **1997**, *34*, 317–326.
62. Peracchia, M.T.; Vauthier, C.; Passirani, C.; Couvreur, P.; Labarre, D. Complement consumption by poly(ethylene glycol) in different conformations chemically coupled to poly(isobutyl 2-cyanoacrylate) nanoparticles. Life Sci. **1997**, *61*, 749–761.
63. Peracchia, M.T.; Vauthier, C.; Desmaele, D.; Gulik, A.; Dedieu, J.C.; Demoy, M.; d'Angelo, J.; Couvreur, P. Pegylated nanoparticles from a novel methoxypolyethylene glycol cyanoacrylate–hexadecyl cyanoacrylate amphiphilic copolymer. Pharm. Res. **1998**, *15*, 550–556.
64. Gan, D.; Lyon, L.A. Fluorescence nonradiative energy transfer analysis of crosslinker heterogeneity in core–shell hydrogel nanoparticles. Anal. Chim. Acta **2003**, *496*, 53–63.
65. Schmid, G. Large clusters and colloids: metals in the embryonic state. Chem. Rev. **1992**, *92*, 1709–1727.
66. Bruchez, M., Jr.; Moronne, M.; Gin, P.; Weiss, S.; Alivisatos, A.P. Semiconductor nanocrystals as fluorescent biological labels. Science **1998**, *281*, 2013–2016.
67. Jaiswal, J.K.; Mattoussi, H.; Mauro, J.M.; Simon, S.M. Long-term multiple color imaging of live cells using quantum dot bioconjugates. Nat. Biotechnol. **2003**, *21*, 47–51.
68. Cao, Y.W.; Jin, R.; Mirkin, C.A. DNA-modified core–shell Ag/Au nanoparticles. J. Am. Chem. Soc. **2001**, *123*, 7961–7962.
69. Kim, J.H.; Lee, T.R. Thermo- and pH-responsive hydrogel-coated gold nanoparticles. Chem. Mater. **2004**, *16*, 3647–3651.
70. Frinha, J.P.S.; d'Oliveira, J.M.R.; Martinoho, J.M.; Xu, R.; Winnik, M.A. Structure in tethered chains: polymeric micelles and chains anchored on polystyrene latex spheres. Langmuir **1998**, *14*, 2291–2296.
71. Bijsterbosch, H.D.; Stuart, M.A.C.; Fleer, G.J. Adsorption kinetics of diblock copolymers from a micellar solution on silica and titania. Macromolecules **1998**, *31*, 9281–9294.
72. Iijima, M.; Nagasaki, Y.; Okada, T.; Kato, M.; Kataoka, K. Core-polymerized reactive micellesm from heterotelechelic amphiphilic block copolymers. Macromolecules **1999**, *32*, 1140–1146.
73. Emoto, K.; Nagasaki, Y.; Kataoka, K. A core–shell structured hydrogel thin layer on surfaces by lamination of a poly(ethylene glycol)-b-(D,L-lactic acid) micelle and polyallylamine. Langmuir **2000**, *16*, 5738–5742.
74. Emoto, K.; Iijima, M.; Nagasaki, Y.; Kataoka, K. Functionality of polymeric micelle hydrogels with organized three-dimensional architecture on surfaces. J. Am. Chem. Soc. **2000**, *122*, 2653–2654.
75. Otsuka, H.; Nagasaki, Y.; Kataoka, K. PEGylated nanoparticles for biological and pharmaceutical applications. Adv. Drug Deliv. Rev. **2003**, *55*, 403–419.

Hydrogen Bonding

J. Richard Elliott, Jr.
Department of Chemical Engineering, University of Akron, Akron, Ohio, U.S.A.

INTRODUCTION

Hydrogen bonding is well known as one of the elementary forces that bond atoms to form materials. Many of its properties are familiar, especially in biological systems, but its impact on chemical processing is also evident in the phase diagrams that dictate the nature of separation processes. By understanding these impacts and their explanation, it is possible to anticipate trends in behavior, which may suggest new process alternatives.

To review briefly, an excellent illustration of its role is in the special properties of water. For example, though similar in size and mass, the boiling temperature of methane is 111 K while that of water is 373 K, which is nearly a factor of 4. Furthermore, this strong attractive force causes liquid water at equilibrium to be relatively dense for the repulsive forces to provide balance. Based on molecular simulations parameterized to match vapor pressure and liquid density, we can estimate the effective diameter of methane to be 0.367 nm, while that of water is 0.306. This implies a packing fraction at the triple point of 0.44 for methane, compared to 0.50 for water. Note that hard spheres spontaneously crystallize to an expanded fcc structure with a packing fraction of 0.545 at liquid packing fractions of 0.495.[1] Hence, a liquid packing fraction of 0.50 is surprising. This effect is attributable to the tetrahedral coordination required for water to fully engage in a three-dimensional hydrogen bonding network. The tetrahedral hydrogen bonding disrupts simple crystallization and actually reduces the density on solidification, notably in floating ice. The tetrahedral structure of ice, along with similar structures in gas hydrates, is clear evidence of a form of bonding that is remarkably strong and not spherically symmetric.

A related anomaly is the hydrophobic effect. As hydrocarbon moieties cannot hydrogen bond, they disrupt water's hydrogen bonding network at substantial energetic penalty. This causes a large, positive Gibbs energy of mixing leading to liquid–liquid separation. Obviously, ''oil and water do not mix,'' but many subtle behaviors are possible when hydrocarbon and hydrogen bonding moieties are combined in a single molecule. Surfactants orient with their polar head groups toward water and their tails toward hydrocarbon regions. Similarly, polar moieties in proteins orient toward the outer regions, while hydrophobic moieties orient toward the center. Polar moieties may also hydrogen bond internally to form α-helices and β-sheets. The complex interaction between these hydrophobic and hydrophilic forces enables the delicate balance that creates unique protein structures, which is further clear evidence that these forces are irrefutable and nonnegligible.

The subject of hydrogen bonding may seem too abstract for a treatise on chemical processing. On the other hand, aspects of hydrogen bonding do contribute in a relatively simple manner to phase equilibria between homogenous solutions. Hydrogen bonding is significant in alcohols, aldehydes, amines, amides, and carboxylic acids as well as water. In mixtures, hydrogen bonding can occur between proton donors like alcohols and proton acceptors like imines, esters, ethers, and ketones. In most cases, the effects of hydrogen bonding are neglected in models of chemical processes or treated vaguely as giving rise to spherically symmetric local compositions. While it is possible to empirically correlate the impacts of these interactions with complex mixing rules, greater physical insight can be achieved by treating the interactions explicitly and systematically analyzing the impacts and trends.[2] In this way, an intuitive sense of the impact of hydrogen bonding on phase equilibria relevant to chemical processing may be developed.

In the presentation below, we discuss a single theoretical perspective on hydrogen bonding that should suffice to illustrate the phenomenology as concisely as possible. That theoretical perspective is Wertheim's.[3,4] The primary alternative is chemical theory, which treats hydrogen bonding as a series of chemical reactions. Elliott and Lira present an introduction to chemical theory and its generalization to Wertheim's theory.[5] Chemical theory is more general in its capacity to describe anomalous types of hydrogen bonding like enhanced formation of rings or specific oligomers like trimers or hexamers. But the application of chemical theory with that level of detail quickly becomes intractable for chemical processes that may include as many as 200 components. Wertheim's theory provides a reasonable balance between explicit treatment of hydrogen bonding and

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007809

its phenomenology on the one hand, and the general applicability on the other.

WERTHEIM'S THEORY

Wertheim's theory expresses the extent of hydrogen bonding in terms of the fraction of hydrogen bonding sites that are not bonded. The key equation can be recognized as a form of the law of mass action

$$1 - X_i^{A_k} = X_i^{A_k} \sum_j^{N_C} \sum_{l=1}^{N_j^D} x_j \alpha_{ij}^{A_k D l} X_j^{D l} \quad (1)$$

where $X_i^{A_k}$ is the fraction of the kth acceptors that are not bonded on the ith component, x_j is the mole fraction of the jth component, and $\alpha_{ij}^{A_k D l}$ is effectively a reaction equilibrium constant.

$$\alpha_{ij}^{A_k D l} = \rho K_{ij}^{A_k D l} \left[\exp\left(\beta \varepsilon_{ij}^{A_k D l} \right) - 1 \right] \quad (2)$$

where ρ is the molar density, $K_{ij}^{A_k D l}$ is the molar "bonding volume" and $\varepsilon_{ij}^{A_k D l}$ is the bonding energy between the kth acceptor on the ith component and the lth donor on the jth component, $\beta = 1/k_B T$ where k_B is Boltzmann's constant.

With these definitions, the left-hand side of Eq. (1) is clearly the fraction bonded, the sum of products of reactant combinations on the right-hand side. Note that an analogous set of equations is obtained for donor fractions by swapping the acceptor and donor labels. Together, these equations comprise a determinate system that can be solved for all bonding fractions of donors and acceptors. Keep in mind that X^A and X^D relate to fractional bonding at a given site, not for a molecule. Thus, the presence of donors and acceptors on the same molecule can lead to the formation of hydrogen bonded chains that strongly influence the true number of molecules in solution.

The bonding volume and bonding energy are, in principle, adjustable parameters that must be characterized for each type of acceptor–donor interaction. In practice, however, Elliott and Natarajan provided a general rule for estimating the bonding volume from the molecular volume and they showed how well-known guidelines for the bonding energy can be applied to reduce the number of parameters in the equation of state to the usual three.[6] An advantage of Wertheim's theory is that the bonding volume parameter has a very clear geometrical definition in terms of the assumed molecular potential function. This facilitates explicit recognition of hydrogen bonding effects through molecular simulation.

The thermodynamics of hydrogen bonding can be described in terms of the Helmholtz energy:

$$\frac{A^{HB}}{RT} = \sum_i^{N_C} \sum_k^{N_k^D} \ln\left(X_i^{A_k} \right) + \frac{1 - X_i^{A_k}}{2} + \ln\left(X_i^{D_k} \right) + \frac{1 - X_i^{D_k}}{2} \quad (3)$$

Differentiating the Helmholtz energy, one obtains other thermodynamic properties, like the compressibility factor, pressure, and internal energy. To clarify, consider the example of a single component with a single acceptor and a single donor. Noting that $X^A = X^D$, by symmetry in this instance, a simple quadratic equation results in

$$X^A = X^D = \frac{-1 + \sqrt{1 + 4\alpha}}{2\alpha} \quad (4)$$

Simplifying Eq. (3) leads to

$$\frac{A^{HB}}{RT} = 2\ln(X^A) + 1 - X^A \quad (5)$$

Thus, an equation of state is obtained that functions much like any other equation of state. That is, given a temperature and density, one obtains an estimate of α from Eq. (2), and consequently an estimate of X^A and A^{HB}. Contributions for disperse attraction and repulsion can be computed from density and temperature in the usual way. Combining the terms provides a complete estimate of the total Helmholtz energy, which can be differentiated to obtain the pressure. The only fundamental difference between this equation of state and those that treat hydrogen bonding implicitly is the intermediate step of computing X^A.

A popular equation for combining disperse attraction and repulsion with hydrogen bonding is the model of Statistical Associating Fluid Theory (SAFT).[7] A similar but simpler equation is the Elliott–Suresh–Donohue (ESD) model.[8] The ESD equation was used by Elliott and Natarajan in their generalized analysis and is more convenient for application in the present context, so the discussion here presents sample results based on the ESD equation.[6] An executable computer program can be obtained for performing the calculations described in Ref.[8a]. Results with the SAFT model are expected to be similar. The ESD equation is given for mixtures as

$$\frac{P\underline{V}}{n_T RT} = 1 + \frac{4\langle c\eta \rangle}{(1 - 1.9\eta)} - \frac{9.5\langle qY\eta \rangle}{1 + 1.7745\langle Y\eta \rangle} + Z^{HB} \quad (6)$$

where $\underline{V}$ = total volume (e.g., cm^3); $\eta = n/\underline{V}^*\Sigma x_i b_i$; b = the molecular volume (cm^3/mol); n = the superficial number of moles in the system (computing component mole numbers based on the molecular weight of a monomer); c = a ''shape factor'' representing the effect of nonsphericity on the repulsive term; $q = 1 + 1.90476(c - 1)$ is a shape factor that represents the effect of nonsphericity on the attractive term; $\varepsilon_{ij} = (1-k_{ij})^* (\varepsilon_i^* \varepsilon_j)^{0.5}$, the energy of disperse attraction (equivalent to well-depth of a square-well potential); $Y_{ij} = \exp(\varepsilon_{ij}/kT) - 1.06$; $\langle c\eta \rangle = n_T/\underline{V}^* \Sigma\Sigma x_i x_j (b_i c_j + b_j c_i)/2$; $\langle qY\eta \rangle = n_T/\underline{V}^* \Sigma\Sigma x_i x_j Y_{ij} (b_i q_j + b_j q_i)/2$; and $\langle Y\eta \rangle = n_T/\underline{V}^* \Sigma x_i b_i Y_{ii}$.

SAMPLE RESULTS OF WERTHEIM'S THEORY

We can illustrate the impact of hydrogen bonding in chemical processing with a few examples. The first example illustrates the impact on a system that solvates and associates, exhibiting only vapor–liquid equilibrium (VLE) behavior. Association is defined as hydrogen bonding of a component with itself. Solvation is hydrogen bonding between two different compounds. The second two examples illustrate the behavior of a system in which only one component associates. The lack of solvation gives rise to clear differences in the phase diagram at constant temperature. At constant pressure, the impact is to induce liquid instability at low temperatures. The role of hydrogen bonding is also evident in vapor pressure and density trends, as illustrated in the two final examples.

Fig. 1A illustrates the phase behavior of the system ethyl ether + methanol. Two characterizations of this system are presented. The solid line explicitly treats the solvation interaction ($k_{ij} = 0.03$; $\varepsilon_{ij}^{HB} = 17.5\,kJ/mol$) whereas the dashed line assumes zero solvation energy ($k_{ij} = -0.0393$). Note that the maximum pressure was constrained to match the experimental value for purposes of this comparison. A simple optimization would cause the dashed line to fit better at low ether compositions, but miss the azeotropic pressure and mask the distinctions between the two models. Several effects are clear. First, implicit treatment of the solvation energy results in a negative value for the disperse binary interaction parameter (k_{ij}). A strong solvation interaction should be suspected any time k_{ij} is found to be negative. Second, the shape of the phase diagram is altered by the presence of solvation interactions. The skewness shifts toward the associating component. The shift in skewness affects the composition of the azeotrope in a very sensitive manner. Third, the disperse binary interaction parameter, k_{ij}, is positive when hydrogen bonding is explicitly recognized, as would be expected from the differences in polarity between the two components. This recognition enhances the physical meaning of the parameters.

One advantage of enhancing the physical meaning of the parameters is that the strength of the solvation interactions can be anticipated to some extent through independent measurements, like those leading to the Kamlet–Taft parameters.[11] Kamlet et al. correlated an extensive collection of UV/visible, NMR, FTIR, and solubility data in terms of acidity (α) and basicity (β) parameters. To our knowledge, no attempt has been made to systematically translate their results in terms of Wertheim's theory, but it should be feasible. Such a translation would bring together a wealth of spectroscopic expertise with the rigorous insight afforded by present molecular simulations. Coleman et al. have performed similar measurements with FTIR relevant to polymer blending.[12]

Fig. 1B illustrates a somewhat simpler binary system than Fig. 1A in terms of its hydrogen bonding

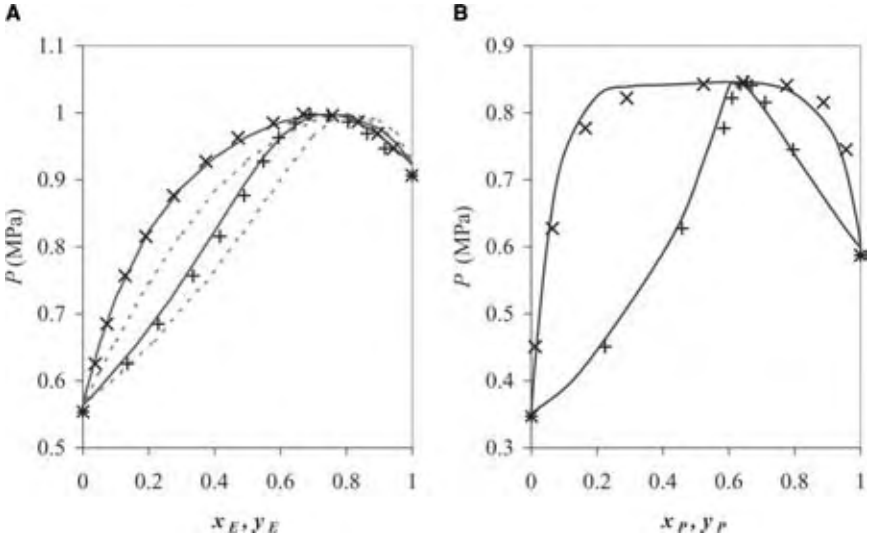


Fig. 1 Phase diagrams of methanol mixtures with (A) ethyl ether at 115°C and (B) *n*-pentane at 99.6°C. (From Refs.[9,10].) *(View this art in color at www.dekker.com.)*

behavior. The *n*-pentane + methanol system exhibits no solvation, but the pentane is very similar in nature to ethyl ether. For example, the boiling temperatures of pentane and ether are just 1.6°C apart, and they both have five backbone atoms making them similar in size and shape. Comparing Figs. 1A and B side by side shows the shift that occurs in the hydrogen bonding effect when one component is a hydrocarbon. The most striking feature is that the azeotropic pressure is 0.025 MPa higher than the highest vapor pressure for the pentane system, compared to just 0.008 for the ether. The corresponding maximum Gibbs excess energies are 1.7 and 0.8 kJ/mol. When the solvation effect is neglected in the ether + methanol system, a value of $k_{ij} = -0.0393$ is required to fit the azeotropic pressure, compared to a value of $k_{ij} = 0.05$ for the pentane + methanol system. These differences are various reflections of the moderation in nonideality that comes from the solvation effect. The $P-x$ projection appears much flatter for the pentane system because of the quick rise in the Gibbs excess energy. This flatness is usually a precursor to incipient liquid instability, but Thomas et al. have observed simple vapor–liquid behavior in the pentane + methanol system at 303 K.[13] One might expect liquid instability to occur at lower temperatures, however. The onset of liquid instability is readily observed in methanol + *n*-hexane, a system similar to the *n*-pentane system of Fig. 1B. A $T-x$ projection of the methanol + hexane system at 0.141 MPa is illustrated in Fig. 2A. Fig. 2A also illustrates a limitation of the hydrogen bonding theory in its present state. It is straightforward to correlate the behavior of either the VLE or the liquid-liquid equilibrium (LLE), but not both simultaneously. Two correlations are illustrated in Fig. 2A. The dashed line shows the result when the default characterization of methanol is applied and the liquid binodal is correlated ($c = 1.120$; $k_{ij} = 0.025$). In this case, the VLE region is misrepresented. The solid line shows the result when the characterization of methanol is customized to describe the VLE region by slightly adjusting the shape factor ($c = 1.095$; $k_{ij} = 0.030$). In this case, the LLE is misrepresented. The inaccuracy in these cases is an indication that there are other forces at work besides the hydrogen bonding and dispersion forces. The most likely explanation is the influence of polar moments and polarizability. The theory of polarity and polarizability in solution is not as thoroughly developed and tested as that of Wertheim and hydrogen bonding. Gray and Gubbins present a review of the current status.[15] Rigorous treatment of polarity and polarizability can be quite challenging, but local composition theories may offer effective means of accounting for these effects. Despite the inaccuracy involved in trying to represent VLE and LLE with a single set of parameters, most applications focus on one region or the other. In these cases, the key is to adapt the model to fit the desired region and the manner in which this may be achieved is illustrated by the example of methanol + *n*-hexane.

The excess volume is a very simple property that provides another indication of the impact of hydrogen bonding. Fig. 1B shows the excess volume for the methanol + *n*-hexane system at 45°C. Clearly, the excess volume is positive, but the mixture volume at 50 mol% methanol is roughly 90 L/mol, compared to 0.7 L/mol for the excess volume, a deviation of 0.8%. The magnitude of this deviation is similar to that of many mixtures that have difficulty packing, but in this case we can attribute the deviations to hydrogen bonding. Diluting the methanol with hexane breaks down the hydrogen bonding network. This causes the solution to expand. The ethanol + water system provides an interesting contrast. For ethanol + water, the

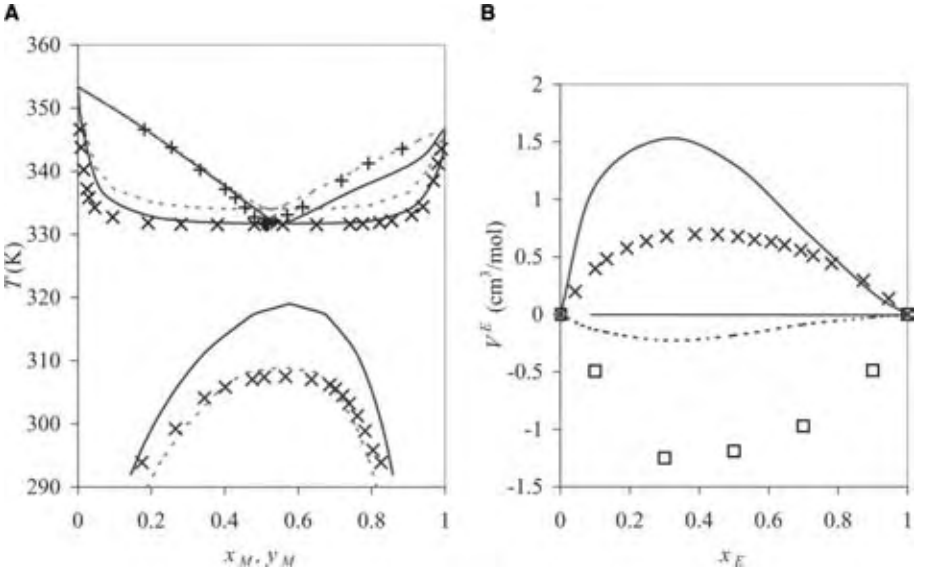


Fig. 2 Thermodynamics of the methanol + *n*-hexane system: (A) temperature-composition diagram and (B) excess volume. (From Ref.[14].) *(View this art in color at www.dekker.com.)*

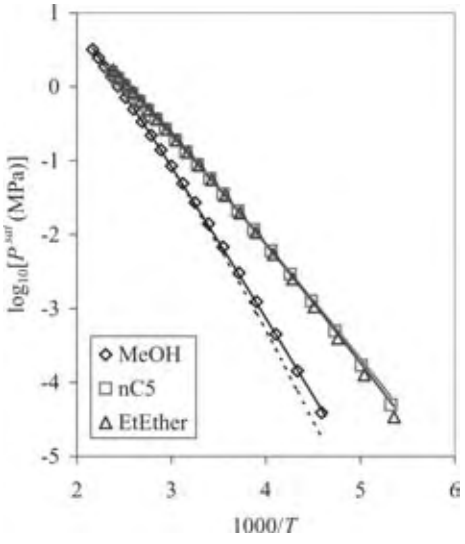


Fig. 3 The impact of hydrogen bonding on vapor pressure. *(View this art in color at www.dekker.com.)*

excess volume is negative, reflecting improved packing relative to the pure components. However, note that the excess Gibbs energy is positive for the ethanol + water system. Thus, the packing is enhanced, but the energy lost in breaking water's network overwhelms the packing efficiency. The hydrogen bonding theory fails to provide a quantitative description of this behavior, but the correct qualitative trend for such a complex phenomenon is noteworthy.

As a final example, we consider the impact of hydrogen bonding on the shape of the vapor pressure curve. Fig. 3 shows that modeling methanol as a hydrocarbon with the same critical properties and acentric factor gives a vapor pressure curve that diminishes too quickly at low temperatures. The hydrogen bonding contribution tends to flatten that curvature, raising the vapor pressure. Similar, but more pronounced behavior was observed for acetic acid.[8] Vapor pressure is an extremely important property when it comes to distillation modeling. Having a physical explanation for why the curve changes is thus quite valuable. With Wertheim's theory, this physical explanation comes complete with the ability to correlate the vapor pressure quantitatively.

HYDROGEN BONDING IN POLYMER BLENDS

A well-known problem in polymer processing is the incompatibility of polymer blends. Even small amounts of nylon in polyester, for example, can lead to reduced optical clarity and discoloration that undermines product marketability. Many other examples of polymer miscibility issues abound. According to Utracki, the majority of polymer blends being sold are actually immiscible blends that have been processed in such a way as to manage the compatibility problems.[16]

One way of managing compatibility problems is through an understanding of the hydrogen bonding in the blends. The dispersion energy is the root cause of the compatibility problem. Hydrogen bonding interactions can be specifically tailored in such a way as to counteract the disperse interactions. For example, polybutylmethacrylate (PBMA) and polyvinylphenol (PVPh) have solubility parameters that would normally make them highly incompatible (8.7 vs. 10.8), but the solvation interaction between the phenol and the carbonyl is strong enough to create a miscible blend below $\approx$160°C.[12,17] Blends that are compatibilized by hydrogen bonding tend to exhibit lower critical solution temperatures. This means that they go from single phase to two phase on heating. The reason is that hydrogen bonding is an exothermic reaction and exothermic reactions decrease in conversion as temperature is increased. A decrease in hydrogen bonding undermines the ability of the specific interactions to overcome the incompatible disperse interactions and the solution becomes unstable. A large number of examples have been reviewed by Coleman et al.[12] Furthermore, Coleman and Painter have developed a hydrogen bonding theory that describes many of the nuances that can occur in these kinds of blends.[18] Their theory is equivalent to Wertheim's theory except when dimer formation becomes peculiar. However, the dimer peculiarities can generally be neglected with little reduction in accuracy.

Polymer blending becomes even more complicated when three or more polymers are mixed. Such a scenario is suggested by adding a third component to enhance compatibility between two incompatible polymers. For example, PVPh could be added to a blend of PBMA and polymethylmethacrylate (PMMA) in the hope of bridging hydrogen bonds from the carbonyl through the PVPh to the pyridine. The third component would behave in the same was as a cosolvent or entrainer in VLE. Unfortunately, ternary polymer blends are subject to an additional, unfavorable contribution from the dispersion energy known as "$\Delta\chi$" effect.[19] Owing to the $\Delta\chi$ effect, islands of immiscibility can occur in a ternary phase diagram even when all binary combinations are completely miscible. Coupled to the $\Delta\chi$ effect, the third component has at least a small preference for one component or the other and then partitions with that component to such an extent that one of the components is left relatively isolated. As a specific example, Pomposo et al. analyzed the ternary system of PMMA + PVPh + polyethylmethacrylate (PEMA). Remarkably, PEMA and PMMA are immiscible, despite their minimal difference in structure. Pomposo et al. found that roughly 60 wt% PVPh is required at 165°C for the ternary blend to be miscible. At this composition, the blend should be

considered simply as "dirty" PVPh, rather than compatibilized PEMA + PMMA. The solution to this problem will likely involve some combination of block copolymer additive with custom tailored hydrogen bonding. Many problems remain to be solved before chemical processes based on this approach become viable.

CONCLUSIONS

We have outlined the impacts of hydrogen bonding on a number of key properties of interest in chemical processing. Most notably, the impact on phase diagrams promotes improved understanding of how the phenomena arise. Note that we have not attempted an exhaustive review of phase diagrams. An excellent overview of phase diagrams in chemical engineering is provided by Walas.[20] Furthermore, we have not attempted to cover aqueous or electrolyte systems as each of these comprise specialized topics in themselves.

Among the solutions considered, the hydrogen bonding is weak relative to aqueous systems, leading to more subtle variations in phase behavior as the disperse interactions compete with hydrogen bonding. The variations become evenmore subtle as one considers mixtures of amines or aldehydes with nonassociating species. Nevertheless, the trends are similar to the trends presented here even for these more weakly associating species. Wertheim's theory can achieve qualitative accuracy over a broad range of conditions. By narrowing down the range of conditions to a specific target, quantitative accuracy can usually be achieved while maintaining the clear physical picture afforded by the very general theory.

REFERENCES

1. Chang, J.; Sandler, S.I. Determination of liquid–solid transition using histogram reweighting method and expanded ensemble simulations. J. Chem. Phys. **2003**, *118*, 8930.
2. Huron, M.J.; Vidal, J. New mixing rules in simple equations of state for representing vapour–liquid equilibria of strongly non-ideal mixtures. Fluid Phase Equilib. **1979**, *3*, 255.
3. Wertheim, M.S. Fluid with highly directional attractive forces. I. Statistical thermodynamics. J. Stat. Phys. **1984**, *35*, 19.
4. Wertheim, M.S. Thermodynamic perturbation theory of polymerization. J. Chem. Phys. **1987**, *87*, 7323.
5. Elliott, J.R.; Lira, C.T. *Introductory Chemical Engineering Thermodynamics*; Prentice-Hall: Englewood Cliffs, NJ, 1999.
6. Elliott, J.R.; Natarajan, R.N. Extension of the ESD equation to polymer solutions. Ind. Chem. Eng. Res. **2002**, *41*, 1043.
7. Huang, S.; Radosz, M. Equation of state for small, large, polydisperse, and associating molecules. Ind. Eng. Chem. Res. **1990**, *29*, 2284.
8. Elliott, J.R.; Suresh, S.J.; Donohue, M.D. A simple equation of state for nonspherical and associating molecules. Ind. Eng. Chem. Res. **1990**, *29*, 1476.

8a. http://gozips.uakron.edu/~elliot1.

9. Srivastava, R.; Natarajan, G.; Smith, B.D. Total pressure vapor–liquid equilibrium data for binary systems of diethyl ether with acetone, acetonitrile, and methanol. J. Chem. Eng. Data **1986**, *31*, 89–93.
10. Wilsak, R.A.; Campbell, S.W.; Thodos, G. Vapor–liquid equilibrium measurements for the *n*-pentane-methanol system at 372.7, 397.7, and 422.6 K. Fluid Phase Equilib. **1987**, *33*, 157–171.
11. Kamlet, M.J.; Abboud, J.M.; Abraham, M.H.; Taft, R.W. Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromic parameters, pi*, alpha, and beta, and some methods for simplifying the generalized solvatochromic equation. J. Org. Chem. **1983**, *48*, 2877–2887.
12. Coleman, M.M.; Painter, G.J.F.; Painter, P.C. *Specific Interaction and the Miscibility of Polymer Blends*; Technomic Publishing: Lancaster, PA, 1991.
13. Thomas, S.; Bhethanabotla, V.R.; Campbell, S.W. Total pressure measurements for *n*-pentane-methanol-2-butanol at 303.15 K. J. Chem. Eng. Data **1991**, *36*, 374–378.
14. Blanco, A.M.; Ortega, J. Experimental study of miscibility, density, and isobaric vapor–liquid equilibrium values of mixtures of methanol in hydrocarbons (C5, C6). Fluid Phase Equilib. **1996**, *122*, 207–222.
15. Gray, C.G.; Gubbins, K.E. *Theory of Molecular Liquids*; Clarendon: Oxford, 1984.
16. Utracki, L.A. *Polymer Alloys and Blends: Thermodynamics and Rheology*; Hanser: Munich, 1989.
17. Patterson, D.; Robard, A. Thermodynamics of polymer compatibility. Macromolecules **1978**, *11*, 690–695.
18. Coleman, M.M.; Painter, P.C. Intramolecular screening and functional group accessibility effects in polymer blends: the prediction of phase behavior. Macromol. Chem. Phys. **1998**, *199*, 1307–1314.
19. Le Menestrel, C.; Bhagwagar, D.E.; Graf, J.F.; Painter, P.C.; Coleman, M.M. Hydrogen bonding in ternary polymer blend systems: determination of association parameters. Macromolecules **1992**, *25*, 7101.
20. Walas, S.M. *Phase Equilibria in Chemical Engineering*; Butterworth: Stoneham, MA, 1985.

Hydrogenation

Xiaobo Yang
Institute of Physical Chemistry and Electrochemistry, University of Hanover, Hanover, Germany

INTRODUCTION

Hydrogenation is applied in the manufacture of a wide variety of commercial products, including oleochemicals, such as edible fats and oils, fatty acids, fatty amines, etc.; specialty chemicals, e.g., pharmaceuticals, flavors and fragrances, herbicides, pesticides; and petrochemicals. While metal hydrides and other hydrogen transfer reagents are useful for hydrogenation experimentation in laboratory and small-scale processes, most hydrogenation processes are carried out through activation of molecular hydrogen in the presence of a catalyst, either in the heterogenous or in the homogenous mode. Mechanisms of hydrogen activation over catalysts and the way the activated hydrogen species performs the hydrogenation of unsaturated chemicals, as well as factors influencing and determining the hydrogenation selectivity are the topics being briefly introduced. Extensive examples of both heterogenously and homogenously catalyzed hydrogenation processes are discussed with respect to the choices of pertinent catalysts, suitable reactors, and conditional parameters, like the reaction temperature and H_2 pressure. Possibilities of combining a hydrogenation reactor to separation and recycling units are shown as well. Heterogenously catalyzed hydrogenation has been long established, and keeps defeating the homogenous processes in many aspects. Heterogenous catalysts, i.e., finely divided or porous metal powders or dispersed metallic nanoparticles on various solid supports, are usually highly stable. In the liquid phase operating mode, they are easily separated from products, regenerated, and recycled. Other advantages of heterogenous processes include a wide range of applicable reaction conditions, the possibility of hydrogenating barely reducible functional groups, e.g., aromatic rings, and the ability to combine with other catalytic functions to achieve, e.g., hydrogenolysis. High selectivity of a heterogenous hydrogenation process is realizable through catalyst modifications and fine-tuning of the conditional parameters. However, homogenous hydrogenation becomes a more feasible method, when mild conditions, high selectivity, in particular enantioselectivity, are required. The wide choice of ligands with variable structural features for metal–organic catalysts provides a precise control of the hydrogenation procedure in the homogenous mode. The future development of hydrogenation processes reveals, besides others, one trend, i.e., to heterogenize the homogenously catalyzed processes by entrapping or anchoring the metal organic active species on suitable solid materials, such as porous inorganic oxides and polymers.

GENERAL TERMS

Hydrogenation is the process of adding H_2 to multiple bonds, reducing them to lower bond orders. In principal all kinds of unsaturated multiple bonds such as C=C, C≡C, C=O, C=N, C≡N, N=O, N=N, N≡N in a chemical are potentially to be hydrogenized. However, the most simple and common type of hydrogenation is performed on C=C double bonds, producing saturated alkanes. Other examples are hydrogenation of nitrogen forming ammonia, hydrogenation of carbon monoxide forming methanol or hydrocarbons. When all the unsaturated bonds in a chemical are reduced during a hydrogenation process, the process is called total hydrogenation. When particular unsaturated bonds of a compound still remain after a hydrogenation process has been performed, it is called partial hydrogenation or selective hydrogenation. In other cases, a reaction of H_2 with unsaturated compounds results in breaking up (dissociation) of the molecule at the position of a multiple bond; this is called destructive hydrogenation or hydrogenolysis.

HYDROGENATION AGENT

To perform hydrogenation on an unsaturated compound one needs, besides others, active hydrogen species. That is, hydrogen has to be present, as an intermediate state, in the atomic form, either neutral or ionic, because molecular hydrogen with two covalently bonded atoms does not react with the unsaturated compound. There are various ways to achieve atomic hydrogen intermediates for a hydrogenation process: 1) using an ionic metal hydride as a hydrogenation agent; 2) using a catalyst to activate (break up) molecular hydrogen; or 3) transfer of hydrogen atoms in the presence of catalysts from one organic compound

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007811

to another, achieving effectively hydrogenation of the latter compound.

Hydride as Hydrogenation Agent

Metal hydrides can be roughly classified into three categories, i.e., ionic, covalent, and interstitial.[1] Ionic hydrides are those compounds of alkali or alkaline earth metals with hydrogen, in which hydrogen has obtained an electron from the very electropositive metal, and forms a hydride ion (H^-). A hydride can be binary, when it involves only two elements including hydrogen, e.g., LiH and MgH_2. It can as well be tertiary when one more metal element is included, e.g., $NaBH_4$ and $LiAlH_4$. Principally, all these ionic hydrides can be used as hydrogenation agents. However, binary hydrides are usually too strongly basic and too reactive. Besides hydrogenation they decompose through an acid–base pathway as shown in Scheme 1, which prevents their use for an effective hydrogenation. Tertiary hydrides are less reactive and are feasible hydrogenation agents. $NaBH_4$ can be used with H_2O as a solvent; $LiAlH_4$ is effective in diethylether, tetrahydrofuran, and liquid ammonia.

It is advantageous to use hydrides because hydrogenation can be carried out at mild conditions and selectively upon those functional groups of chemicals like carbonyl compounds, amides, sulfonates, halogenides, epoxides, and alkynes, etc., whereas C=C double bonds in the chemicals can remain unattacked. With carbonyl as an example it is shown in Scheme 2 that the hydrogenation using a hydride proceeds through a nucleophil mechanism, where the carbonyl becomes activated through coordination with the metal ion.

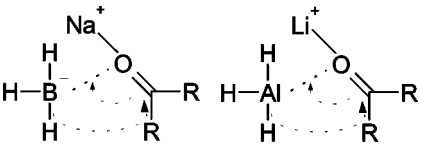
The special selectivity of hydrides toward functional groups other than C=C double bonds makes them useful in the synthesis of functionally structured molecules, e.g., reduction of amides leading to the synthesis of amines, as shown in Scheme 3.[2]

In case of an α,β-unsaturated carbonyl compound there are possibilities for variable regional selectivity: 1,2-reduction results in unsaturated alcohol; 1,4-reduction in saturated ketone, and in protonic solutions further to saturated alcohol (Scheme 4). In this case, Li^+ ion is a stronger Lewis acid than Na^+ ion, thus $LiAlH_4$ exhibits a higher selectivity toward 1,2-hydrogenation rather than 1,4-reduction. Similarly, $Ca(BH_4)_2$ can be employed, and behaves like $LiAlH_4$ owing to the stronger Lewis acidity of Ca^{2+} than Na^+ as well.[3]

Another important property of hydrides is the stereo-selectivity. This is determined by the way the

$$M^+ + H^- + H\text{-}R \longrightarrow H_2\,(gas) + M^+ + R^-$$

Scheme 1



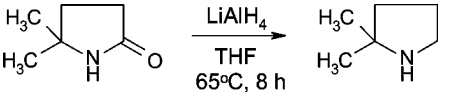
Scheme 2

bulky BH_4^- or AlH_4^- ions approaches the unsaturated bonds (Scheme 5). This kind of steric approach control becomes more propagated if the hydride ions (H^-) have been partly substituted by bulkier organic ions. For instance, $LiAlH_4$ can be mixed together with calculated amounts of proper alcohols, yielding $LiAlH_2(OR)_2$ or $LiAlH(OR)_3$. They approach the unsaturated functional groups only in particular configurations, thus elevating the stereo-selectivity. An example is shown as Scheme 6.[4]

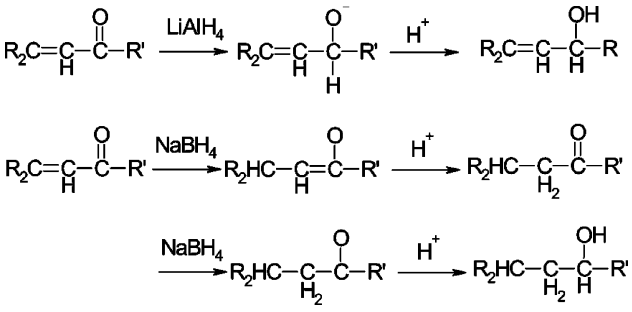
Besides the reduction of carbonyl groups, hydrides can furthermore be used to substitute halogenides, sulfonates, with hydrogen atoms, to selectively reduce epoxides to α-hydroxides, and alkynes to alkenes. The reactions are normally completed at mild conditions within a few hours, i.e., temperature ranges below 373 K and atmospheric pressure. Sometimes even lower temperatures (down to liquid nitrogen temperature) are applied to slow down the reaction rates and to suppress side reactions. Hydrides as hydrogenation agents are especially feasible for multistep synthesis of specialty chemicals such as medicines and pharmaceutically active components in laboratory or other small-scale experiments. Because the reactions are carried out homogenously in a liquid phase, all kinds of thermostatted and atmosphere-controllable batch and continuous reactor vessels are suitable. For transportation and handling, care has to be taken that contacts of hydrides with water and oxidative atmosphere, like ambient air, produce gaseous hydrogen and an enormous amount of heat, causing a disastrous explosion.

Activation of Molecular Hydrogen Using Catalysts

The history of hydrogen activation using a heterogenous catalyst (solid) dates back as early as 1823 when Dobereiner noted that hydrogen inflames spontaneously upon contact with finely divided platinum. Some 50 yr later chemists began to experiment with nickel, copper, iron, palladium, and other metals for



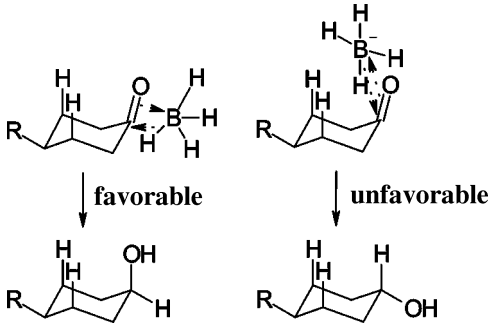
Scheme 3



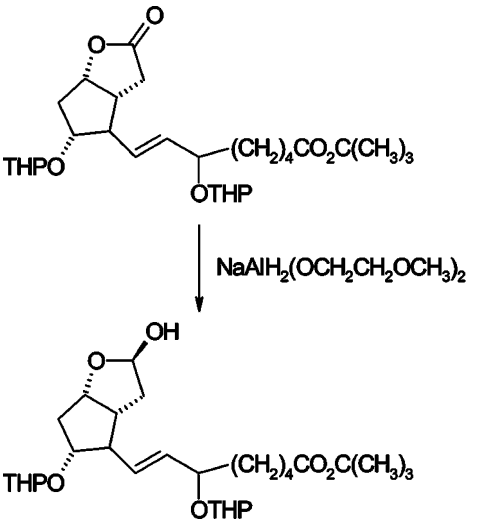
Scheme 4

hydrogenation and production of hydrogen from steam and hydrocarbons. From 1896 onward Sabatier and Senderens developed the process of hydrogenation of organic compounds in the vapor phase, including the hydrogenation of oleic to stearic acid, and nitrobenzene to aniline.[5] For years they erroneously recommended that no liquid should cover the catalyst surfaces. Nowadays by contrast most heterogenously catalyzed hydrogenation processes are performed in the liquid phase, with a solid catalyst, and in pressured H_2 atmosphere. The metal catalysts, noble metals like Pt, Pd, Ir or other metals such as Cr, Ni, Cu, and Fe are used as fine powder, porous particles, or highly dispersed nanoparticles supported on active carbons, aluminum oxides, or other oxide solids. This ensures high surface areas, thus a high effectiveness of the catalysts. Examples of common hydrogenation catalysts are the Raney Nickel, sponge-like nickel particles made by dissolution of aluminum from nickel–aluminum alloy, and Pd/C, supported palladium on high surface active carbon.

Scheme 7 sketches three possible intermediate states of the catalyst surface during a hydrogenation process. Using the concept of these intermediates the mechanism of hydrogenation can be explained: In all the three states, hydrogen molecules are dissociatively adsorbed on the metal surface as atoms. Then, the unsaturated bond of an organic compound can be either adsorbed



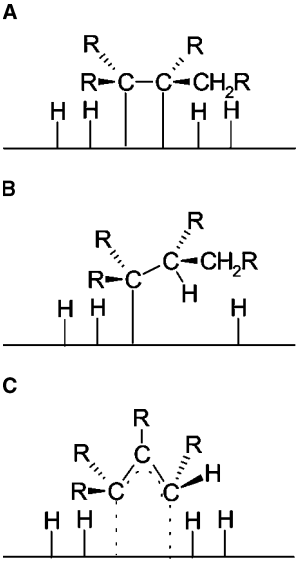
Scheme 5



Scheme 6

with its π and π^* orbitals on the metal surfaces, shown as case A, or inserted into the M–H bond, as in case B. Finally, a reductive elimination of the organics results in the hydrogenation products. Case C explains why the double-bond migration, *cis*- and *trans*-isomerization, and hydrogen exchange occur as side reactions of a hydrogenation process.

Different unsaturated bonds exhibit different reactivity toward hydrogenation. In general, alkynes are hydrogenated faster than alkenes under the same conditions. In turn, the hydrogenation of alkenes is faster than for other functional groups like carbonyls, nitrides, and nitrates. However, this chemo-selectivity is an issue to be tuned by altering catalysts. For instance, in the presence of Pd catalyst carbon–carbon



Scheme 7

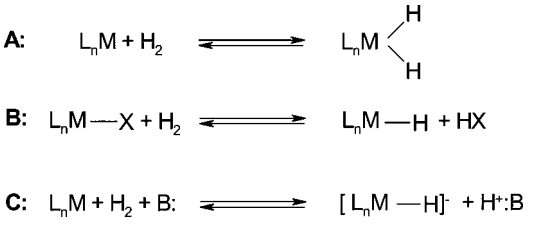
multiple bonds are reduced while C=O bonds remain. With Rh and Ir catalysts C=O bonds are more easily hydrogenated than C=C bonds.

Concerning the stereochemistry of a hydrogenation process, the adsorption of the unsaturated compound on the catalyst surface happens principally on the sterically less detained side, therefore hydrogen atoms are usually added on this side. However, this constraint is not that strict, and is easily amended through many conditional factors such as substituting groups on or close to the unsaturated bonds, solvents, and modifiers, i.e., a small quantity of a proper organic additive with particular structural features, etc. Details will be given with examples in the section Heterogenous Hydrogenation.

Beside heterogenous catalysts, there are various so-called homogenous catalysts exhibiting the function of activating molecular hydrogen, which can be employed in hydrogenation processes. These are transition metal–organic compounds, which according to the kind of hydrogen activation, can be divided into three types. The oxidative addition of hydrogen onto the metal site is the most common type of H_2 activation on a metal having d electrons (Scheme 8A). The organic complexes of early transition metals (d^0) activate hydrogen through their own hydrolysis, as shown in Scheme 8B. Lanthanides and actinides work similarly. The metal center has an orbital binding an anionic ligand, which is protonated off upon contact with hydrogen. During this process, the oxidation state of the center metal remains unchanged. In the heterolytic cleavage mode shown in Scheme 8C an external base is present to bind the proton that has been formed by the heterolytic cleavage of hydrogen on the metal center.

There are several advantages in using metal–organic compounds as hydrogenation catalysts. The catalytic activity can be varied by proper choice of the central metal species as well as the organic ligands. More significantly, the wide choice of organic ligands with a broad variety of chemical components and structural types allows very precise control of the way the unsaturated compound approaches the catalytic sites, therefore providing high chemical, regional, or stereo-selectivity. The section Homogenous Hydrogenation will be dedicated to these topics.

Hydrogen is a valuable chemical and expensive to manufacture. About 50% of commercial bulk hydrogen is currently produced by steam reforming of natural gas. Other hydrogen resources involve syn-gas, refinery side products, etc. At 973–1373 K and 0.3–2.5 MPa pressure, steam reacts with methane, yielding carbon monoxide and hydrogen over a catalyst: $CH_4 + H_2O = CO + 3H_2$. It is usually followed by the "water-gas shift" reaction, $CO + H_2O = CO_2 + H_2$, to recover additional hydrogen.[6] Electrolysis of water for hydrogen production has been discussed but no efficient method has been established yet. New methods investigated now are conversions of biomass derivatives, promising sustainable hydrogen production from renewable resources.[7]



Scheme 8

Other Reagents for Hydrogen Transfer

Both heterogenous and homogenous catalysts display the capability to activate hydrogen of a hydrocarbon or other organic compounds, and transfer it in the atomic form to another unsaturated compound by forming a more stable hydrogenated chemical. An example often used is diimine, or its derivatives, which, in the presence of Cu^{2+}, can hydrogenate alkenes, as shown in Scheme 9.

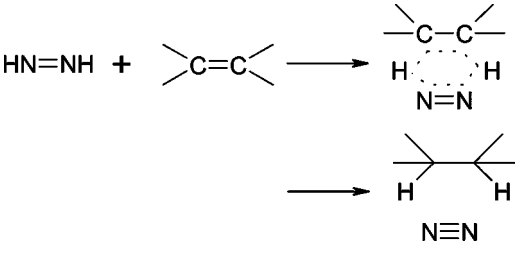
Another example is cyclohexane, which offers hydrogen in the presence of a noble metal catalyst by forming benzene.

Using these hydrogen transfer reagents to perform hydrogenation when it is not possible to employ a hydrogen atmosphere, or when high selectivity at mild reaction conditions is obliged, is a practical option. These are usually the cases of a laboratory hydrogenation experiment, or a synthesis of small-scale fine chemicals.

HETEROGENOUS HYDROGENATION

Ammonia Synthesis

The most important heterogenous hydrogenation process is probably the Haber process of ammonia synthesis, or more precisely, Haber–Bosch process.[8] It is a process of fixing air nitrogen, which is otherwise extremely unreactive. The ammonia produced is further



Scheme 9

processed into fertilizers, explosives, and polymers, on which the industrial civilization since the 20th century depends. In the production of fertilizer alone, the Haber process produces currently about 50% of all nitrogen compounds.

The Haber process is carried out at pressures of up to 20 MPa and temperatures from 623 to 723 K over a catalyst. It is an exothermic reaction ($\Delta H = -92\,\text{kJ/mol}$), thus an intermediate temperature favors the equilibrium on the product side, as does high pressure. The first Haber–Bosch reactor used a catalyst containing osmium and uranium. Current ammonia synthesis catalysts are commonly Fe catalysts with K promoters. The usual way to prepare the catalyst is to reduce magnetite, an iron oxide mineral, which is mixed with a minor amount (<1 wt%) of calcium, aluminum, and/or silicon oxides, in hot hydrogen. The resulting Fe catalyst is highly porous, whose large surface area is desirable for good activity. The potassium promoter is either added during the calcination procedure or introduced through a postsynthesis treatment with amounts around 0.5 wt%. It increases the electron density of the catalyst, and thus improves the activity. Furthermore, catalysts based on Co and some other transition metals are considered effective for ammonia synthesis as well. Corresponding processes are currently in a phase of development.

The working principal of a Haber–Bosch reactor is sketched in Fig. 1. Reactors actually used by the commercial processes, e.g., the Haber–Bosch reactor at BASF, the radial flow converter at Haldor–Topsoe bear the same working principal. This is a typical tube-bundle of fixed-bed reactors. The production volume is typically 1500 tons of ammonia per day. The feedstocks are air containing N_2 (and enriched O_2 to reach a high catalyst activity); H_2 is made from syn-gas. NH_3 leaves the reactor as liquid under the applied pressure. Unreacted N_2 and H_2 are separated from the product on-site and recycled. A new trend is to employ H_2 made from the natural gas because of its lower cost compared with the syn-gas made from coal.

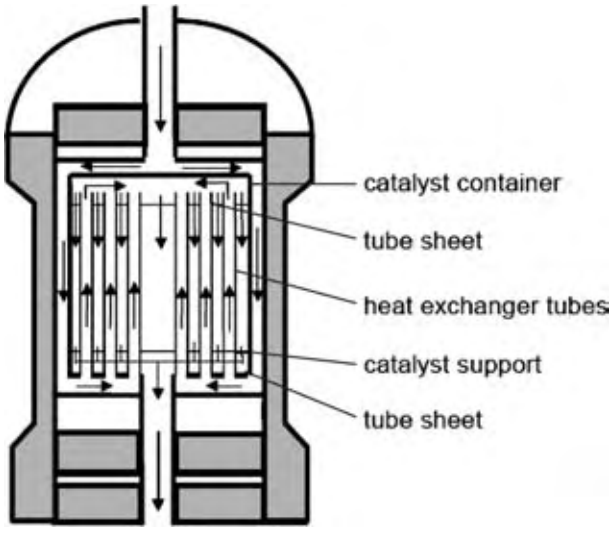


Fig. 1 Haber–Bosch reactor for ammonia synthesis.

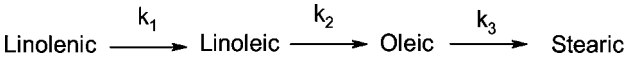
Vegetable Oil Hardening

Hardening the liquid vegetable oil to solid or semisolid, e.g., margarines, is demanded by the food industry, because the way the fat mixes with the flour produces a desired texture in baked products. The hardening is done by a partial hydrogenation process. The hardened oil is predominant in most commercial baked goods.[9]

Liquid vegetable oil is a mixture of linolenic, linoleic, oleic, and stearic esters. The hydrogenation process of these esters displays a chart of a reaction chain shown in Scheme 10. The processes of preferential hydrogenation of more unsaturated acids with minimum formation of completely saturated fatty acids are preferred by the food industry. The selectivity is expressed as the ratio $k_{\text{lino}}/k_{\text{ole}}$ or $k_{\text{lo}}/k_{\text{o}}$, where the variables k are the reaction rate constants, and the term ''lino'' represents the conversion of linolenic and linoleic compounds, while ''ole'' is for oleic. Thus, good processes have high $k_{\text{lo}}/k_{\text{o}}$ values.

Commercially available catalysts for oil hardening are mainly of two types: nickel catalysts and copper catalysts. A typical nickel catalyst contains 25 wt% Ni in hard fats as flakes. Copper catalysts for this purpose include the copper–chromium catalyst, which is a material containing 50 wt% CuO, 40 wt% Cr_2O_3, and 10 wt% BaO, and Cu dispersed on silicic acid supports with various loadings. Nickel catalysts cost less but are less selective than copper catalysts. The $k_{\text{lo}}/k_{\text{o}}$ value for a copper catalyst can be as high as 6–18. Hydrogenation can be performed on oils containing over 70% oleic acid, without increase in stearic acid.

The vegetable oil hardening processes are carried out typically in stirred tank vessels as schematically shown in Fig. 2, with units of catalyst separation and recycling. The conditional parameters are applicable at temperatures between 373 and 473 K, H_2 pressures of 1–5 MPa, and agitation rates up to 1000 rpm, typically around 750 rpm, within a time scale of a few hours, typically up to 5 hr. All these factors affect not only conversion but also selectivity, like all chain-type reactions. In terms of $k_{\text{lo}}/k_{\text{o}}$, higher temperatures have a positive effect, higher pressures and agitation rates have a negative effect, while they all increase the



Scheme 10

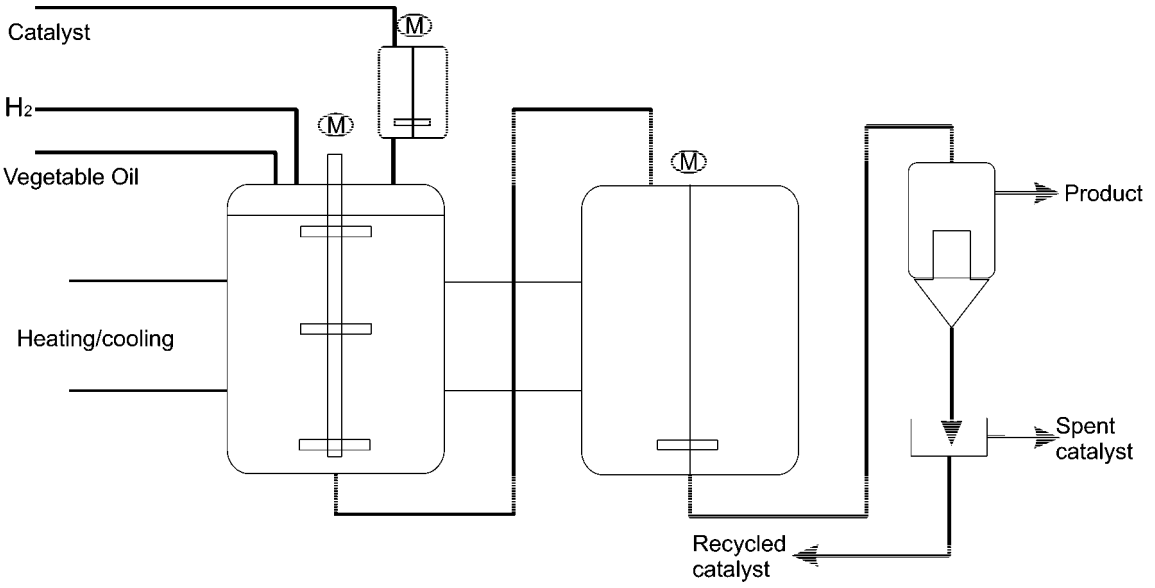


Fig. 2 Working principle of a stirred vessel usable for liquid-phase hydrogenation, e.g., vegetable oil hardening.

reaction rate. Thus, a compromise has to be found for these conditional factors.

Side reactions of a vegetable oil hydrogenation process are double-bond positional isomerization and *cis*- and *trans*-isomerization. The former type of double bond migration leads often to the formation of conjugated double bonds. The latter, because natural oil contains only the *cis*-isomers, results in the formation of an unnatural *trans*-fatty oil, with some of the remaining unsaturated bonds having transformed into the *trans*-double-bond form. While there is few concern about the double-bond migration, the formation of *trans*-fat stimulates heavy debates on its health implications. Manufacturers advertise that *trans*-fat is healthy because it is not digested and absorbed by the human digestion system. Thus, eating hydrogenated vegetable oil has the same effect as eating less fat. In contrast, other people insist that the useless *trans*-fat would deposit in blood vessels and increase the risk of heart attacks.[10]

Synthesis of γ-Butyrolactone by Hydrogenation of Maleic Anhydride

γ-Butyrolactone (GBL) is an important intermediate for the manufacture of butyric compounds and pyrrolidones used as plastilizers, polymer solvents, nylon precursors, and components of insecticides. γ-Butyrolactone is also used in photochemical etching, in vitamin and pharmaceutical preparations, and has many other applications. Synthesis of GBL by hydrogenation of maleic anhydride is a relatively new process providing a good example of studying the chemo-selectivity of hydrogenation.[11]

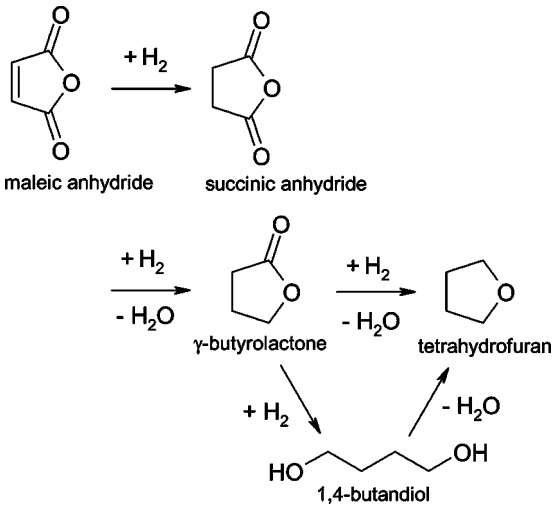
The reaction pathway, and therefore the possible side products, of maleic anhydride hydrogenation are given in Scheme 11. Because succinic anhydride is not desired, a catalyst with higher activity is needed to overcome the hurdle, leading H_2 to attack not only the C=C bond but also the carbonyl groups. Tetrahydrofuran (THF) and 1,4-butandiol (BDO) are other highly valued chemicals used as solvents and intermediate in pharmaceuticals. The production of GBL, THF, and BDO can be tuned as demanded.

The hydrogenation is carried out in a ''low-pressure vapor phase'' process, i.e., at around 473 K and under a few megapascals of H_2 pressure. Known effective catalysts include copper–chromium catalysts, and palladium–carbon catalysts, etc. The process can be coupled with maleic anhydride production units, using either benzene or butane feedstocks, and units to separate and refine GBL, THF, and BDO products. Fig. 3 shows a chart of such a process design. Other catalysts with higher activities allowing the process to be conducted at still lower temperatures in the liquid phase are under development.

γ-Butyrolactone is also the precursor to γ-hydroxybutyrate (GHB). γ-Hydroxybutyrate is a medicine with the function of a neurotransmitter. Misuse of the medical as a so-called ''date rape drug'' is dangerous. Illegal possession or sale of GHB, e.g., in the United States of America is to incur a penalty under the Texas law.

Heterogenous Asymmetric Hydrogenation

The possibility of performing enantio-selective hydrogenation over heterogenous catalysts is based on the fact that a chiral additive can govern the way a



Scheme 11

prochiral, unsaturated compound approaches the catalyst surface, and therefore the stereochemistry of its hydrogenation. Now there are two successful processes of enantio-selective hydrogenation. The first is the hydrogenation of β-ketonesters over the Raney Nickel catalyst with chiral tartaric acid and sodium bromide additives; the other is the hydrogenation of ethyl pyruvate over supported noble metals. Chiral promoters for this reaction are cinchona alkaloids. An impressive feature of these processes is that the enantio-selectivity can be switched by simple exchanges of the (*S*)-, and (*R*)-forms of the chiral additives.

Methyl or ethyl acetonacetate can be hydrogenated to (*R*)- or (*S*)-methyl- or ethyl-3-hydroxybutyrate on nickel catalysts. With (*R*)- or (*S*)-tartaric acid and sodium bromide as promoters, the hydogenation can reach enantiomeric excess (ee) values over 90%. The ee value is defined as the excess of the major enantiomer to the minor one over the total yields. The products are vitamin precursors. The function of tartaric acid is believed to form nickel(II) tartarate, which is adsorbed on the metal surface. The asymmetric site generated this way is responsible for the hydrogenation process. Naked nickel surface sites, without adsorbed tartarate, are poisoned by NaBr. The reaction is carried out at the boiling point of a solvent under a few megapascals H_2 pressure in stirred vessels.[12]

Cinchona promoted Pt or Pd (supported) catalysts are feasible for the hydrogenation of carbonyls or C=C bonds, both at the α-position of another carbonyl group. For example, ethyl pyruvate is hydrogenated on cinchonidine promoted Pt/C, producing ethyl lactate with ca. 95% ee.[13] The product is used in pharmaceuticals, agrochemicals, solvents used in the electronics industry, flavors and fragrances, health care foods, etc.

Two possible mechanisms have been proposed for this process. The first one is comparable with the Ni/tartaric acid system. Cinchonidine is adsorbed at the edges of Pt crystals or on ad-atoms on the Pt surfaces. These sites then allow the approach of ethyl pyruvate only with stereo-special configurations, resulting in enantio-selectivity. In the other mechanism a complex-like supramolecule of ethyl pyruvate and cinchonidine forms in the solution and approaches the metal surface afterward. The reaction is carried out as well in stirred vessels under a few megapascals.[14]

Asymmetric hydrogenation using modified solid metal catalysts is currently a topic under intensive research and development. In this respect, not only the reaction mechanism is studied, novel organic additives leading to more effective asymmetric hydrogenation are under consideration as well.[15]

Other Heterogenous Hydrogenation Processes

An important hydrogenation process, the Fischer–Tropsch process to hydrogenate carbon monoxide producing methanol and hydrocarbons using Fe and Co catalysts, is a stand-alone entry in this encyclopedia. Furthermore, hydrodesulfurization (HDS) has been drawing significant attention in recent years because of air pollution concerns and related legal regulations with ever more stringent limitations of

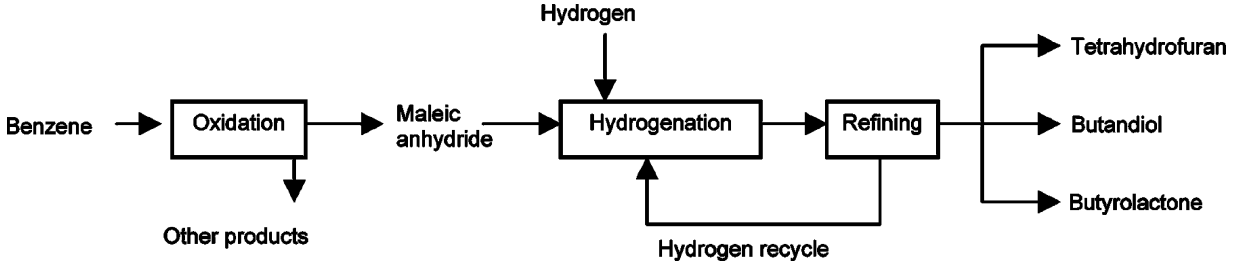


Fig. 3 Process chart for productions of tetrahydrofuran, butandiol, and γ-butyrolacton.

sulfur concentrations for transportation fuels.[16] An HDS process involves catalytic treatments with hydrogen to convert various sulfur compounds present in petroleum feedstocks. Sulfur compounds in the crude oil and various distillates include thiols, thioethers, thiophenes, polyaromatic thiophenes, and their derivatives. They are destroyed over HDS catalysts, producing hydrocarbons and hydrogen sulfide, which in turn is separated and converted to elemental sulfur via the Claus process. Currently employed catalysts are MoS_2 on γ-Al_2O_3 supports, with Co, and sometimes Ni, as promoters. Operation conditions are 673–773 K and 0.5–5 MPa hydrogen pressure, using multistage fixed-bed reactors. A key issue in the process design is the efficiency and recycling of hydrogen.[17]

Other noticeable hydrogenation processes are selective hydrogenation of alkynes, e.g., acetylene, to *cis*-alkenes, and similar partial hydrogenation processes. For partial hydrogenation, catalysts with reduced activity are required. Available catalysts include Lindlar's catalyst, a palladium catalyst partially poisoned with Pd and $CaCO_3$, and the nickel bromide catalyst, i.e., a $NaBH_4$ reduced $NiBr_2$.[18,19]

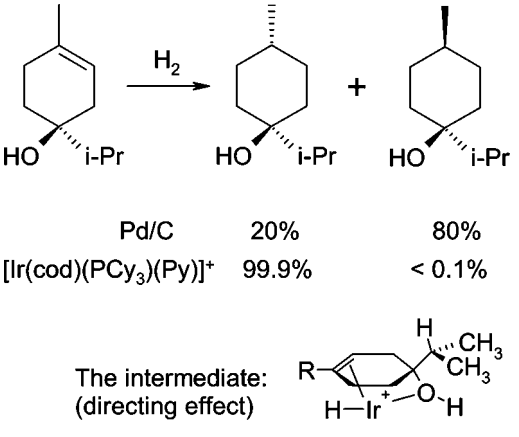
Hydrogenation using metal catalysts is a well-established method. Besides the hydrogenation of carbon–carbon unsaturated bonds, complete or partial saturations of aromatics, synthesis of fatty amines from nitriles, alcohols from ketones and aldehides can all be accomplished through heterogenous hydrogenation. Many catalysts such as Raney Nickel, copper–chromium, Pt/C, and Pd/C are commercially available in a large volume, with various modifications in metal loadings and surface areas. Chemical selectivity can usually be tuned by choosing a proper catalyst, e.g., Pd hydrogenates selectively C=C bonds, Rh and Ir hydrogenate preferentially carbonyls under suitable conditions. The choice of solvents when the process is carried out in the liquid-phase mode affects in certain circumstances the activity and selectivity because of their different polarity. EtOH, THF, dichloromethane are often used when a polar solvent is preferred, while cyclohexane is suitable as an apolar solvent. The processes can be performed either in solutions at rather low temperatures and under low H_2 pressures using a standard stirred vessel, or in the vapor phase with a fixed-bed flow-type reactor. In properly settled processes, the organics conversion can be completed without significant side products. H_2 is usually recycled.

Heterogenous hydrogenation catalysts can be combined with materials possessing other types of catalytic properties, resulting in bi- or multifunctional catalysts. For example, the catalyst used in the hydrocracking process is a combination of a noble metal, an acidic zeolite, and some other components. It leads, under the process conditions, to very effective hydrogenolysis of the long-chain components in the crude oil.[20]

HOMOGENOUS HYDROGENATION

Wilkinson's Catalysts

Wilkinson's catalyst, $RhCl(PPh_3)_3$, has been the first highly active homogenous hydrogenation catalyst. It was discovered by G. Wilkinson in 1964, and independently by R. Coffey almost at the same time.[21,22] The ligand helps the metal to be dissolvable in polar solvents and active at low temperatures. The mild condition then has an effect that the most reactive multiple bond in a molecule is hydrogenated with an extremely high selectivity. From mixtures of multiple unsaturated compounds, only the most reactive one is hydrogenated. The Wilkinson's catalyst can be simply prepared by reacting $RhCl_3{\cdot}3H_2O$ with PPh_3 in EtOH. The hydrogenation procedure over Wilkinson's catalyst follows three major steps. First, the unsaturated bond coordinates with a free orbital of the metal site, forming a π-complex. Then, this coordinated π-ligand interacts with a neighboring hydride, resulting in its insertion into the M–H bond. Finally, its reductive elimination gives a saturated product. Similar catalysts include the cationic $[Rh(cod)(PPh_3)_2]^+$, $[Ir(cod)(PCy_3)(Pyridine)]^+$, etc. They are even more active because the cationic metal sites are more electrophilic and favor the coordination of unsaturated compounds, which is often the rate determining reaction step. The stereoselectivity is determined by the way the multiple bond approaches the metal site. In certain cases, an inversed stereo-selectivity can be expected using the homogenous catalyst instead of the heterogenous catalyst, when other functional groups present in the compound offer a function to coordinate with the metal, as shown in Scheme 12 as an example. This kind of a contribution of the binding affinity of a neighboring functional group to an inversed stereo-selectivity is called "directing effect." Different functional groups in this respect



Scheme 12

display different strengths of directing effects. Generally, the order of the strength is as follows:

amide > OH > OR > ester ~ ketone

Amines and chelating ligands would bind too strongly to the catalyst, and therefore prevent hydrogenation.

The processes are carried out normally at 273 K with the Ir catalysts and 298 K with the Rh catalysts under 0.1 MPa H_2 pressure in stirred vessels. Feasible solvents are ethanol, acetone, dichloromethane, etc.

Lanthanide Hydrogenation Catalysts

$(Cp_2^*LuH)_2$ is present in solutions in its monometallic form, and is found to be the most active hydrogenation catalyst, reaching over 10 times the conversion rate as Wilkinson's catalysts.[23] Similar to Wilkinson's catalysts, the monometallic lanthanide has a free coordination site allowing the bonding of an unsaturated bond. The following hydride insertion and elimination then close up the catalytic circle. The processes are carried out at or below room temperature with 0.1 MPa H_2 pressure. This again enables the highly selective hydrogenation of the most reactive bond while other unsaturated sites remain intact.

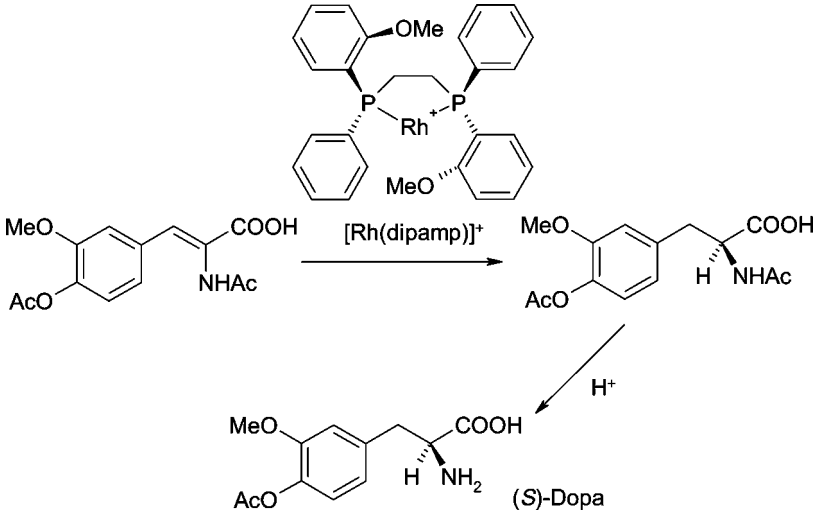
Homogenous Asymmetric Hydrogenation

About 95% of all hydrogenation processes are performed using Pd/C or Raney Nickel in the heterogenous mode.[24] The asymmetric hydrogenation, with high enantio-selectivities is the only area where the homogenous hydrogenation can predominate. The reason is that the high activity of the metal organics and the versatility of the available ligand structures provide precise controls of the coordination, H-insertion, and leaching of the unsaturated bond upon the catalytic site.

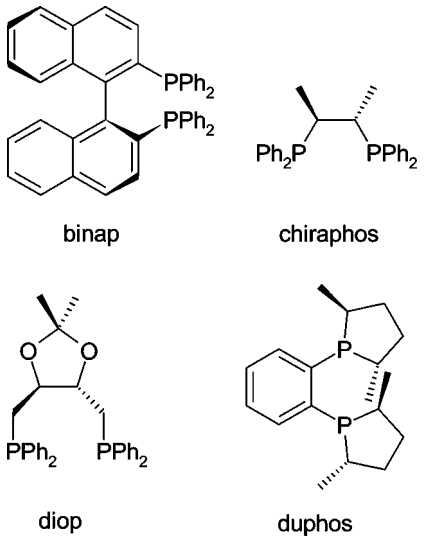
The first example has been the hydrogenation of α-acetamindocinnamatic acid producing (*S*)-Dopa, a medicine against Parkinson's disease.[25] As shown in Scheme 13, using the dipamp ligand the rhodium catalyst can achieve a 97% ee. The ligand, dipamp, is a chiral compound based on the antropisomery, i.e., the steric ligand is not freely rotational because of its bulkiness, therefore the molecule is rigid and asymmetric. The process producing (*S*) -Dopa has been commercialized at Monsanto, and won a Nobel prize in 2001 for the inventor, William Knowles, together with Ryoji Noyori for his contribution in heterogenous asymmetric hydrogenation, and K. Barry Sharpless, for chirally catalyzed oxidation.

Another often used chiral ligand is 1,1′-binaphthyl-2,2′-diyl-bis(phosphine), or binap.[26] It can be built into ruthenium catalysts, which display high activities and enantio-selectivities for the hydrogenation of alkenes, amines, allylalcohols, and unsaturated carbonyls. A directing effect is observed for an α-hydroxy group in these cases.

Besides dipamp and binap, many other effective chiral ligands, among which are mainly bi-phosphines, have been designed and are commercially available. Scheme 14 sketches a few examples in this category. There are many more chiral ligands that are obtained by connecting substituting groups to these basic structures. The chiral ligand enables a tailored design for hydrogenation of particular chemicals to the desired products. The example in Scheme 15 shows how complicated the structure of a hydrogenation catalyst can look.



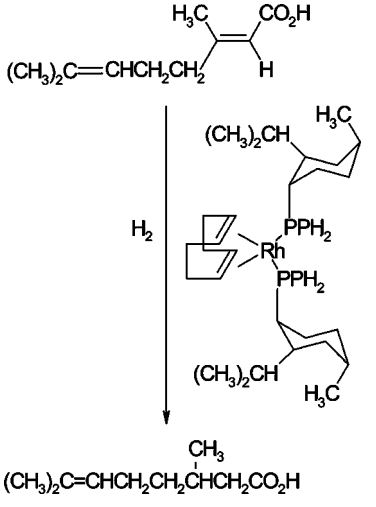
Scheme 13



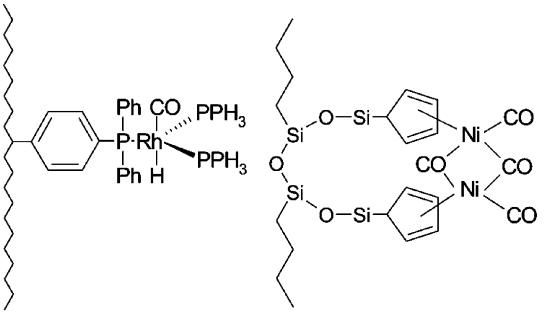
Scheme 14

Heterogenizing the Homogenous Hydrogenation

The homogenous hydrogenation processes are advantageous to the heterogenous processes in that they proceed faster and are highly selective under mild conditions. But there is a major drawback preventing their applications for large-volume productions, because the separation and recycling of catalysts from the products are tedious and cost-intensive. Therefore, methods to heterogenize the homogenous reactions are currently being developed. For the purpose of a heterogenization, the metal organic catalysts are entrapped or anchored to solid host materials, such as porous inorganic oxides, or polymers. Solid supramolecular materials are obtained through these procedures, which preserve the high activity and selectivity of the



Scheme 15



Scheme 16

metal organics. Further benefits of these materials are with respect to stability, ease of recovery, and further properties related to the porous structures such as high surface areas and the ability to diffusion control the reaction (e.g., shape-selectivity). Examples are the ''ship-in-the-bottle'' catalysts in zeolites, Wilkinson's catalysts anchored in the MCM-41 mesoporous silica through covalent bonding, and catalysts bonded to polymer backbones.[27–29]

Phosphorous is again the most common ligand element used to attach the metal complexes to inorganic or organic solid supports. Taking the organic polymers as examples, phosphorous can be bonded to the polymer chain as a dangling group in various ways. Then, these phosphorous groups can be used as ligands coordinating or chelating to the metallic catalysts. Other functionalized polymer chains or inorganic surface sites can be linked with metal sites by similar means. As shown in Scheme 16, many Ru, Rh, Ni, Ir catalysts have been heterogenized using this method. They have been proved to preserve the hydrogenation activity in various test reactions, such as saturation of acetylene, hexene hydrogenation, nitrobenzene hydrogenation, etc.[29]

Large-scale processes are yet to be developed based on these results. A major challenge is the catalyst leaching, as a heterogenized catalyst could undergo reversible or irreversible reactions forming species soluble in the reaction media, which may destroy the catalytically active species or even lead to undesired side reactions of the reactants and products. These problems have to be solved before successful processes are developed.

CONCLUSIONS

While hydrogenation using metal hydrides and other hydrogen transfer reagents offers a feasible method for laboratory experimentations, activation of molecular hydrogen with metal or metal organic catalysts is broadly employed in industrial hydrogenation processes.

Heterogenously catalyzed hydrogenation has long been established and has proved to be better than the homogenous processes in many aspects such as the stability of catalysts, the ease of separation of products from catalysts, a wide range of applicable reaction conditions, the possibility to hydrogenate hardly reducible functional groups, e.g., aromatic rings, and the ability to combine with other catalytic functions to achieve, e.g., hydrogenolysis. High selectivity of a heterogenous hydrogenation process is achieved through catalyst modifications and fine-tuning of the conditional parameters. However, homogenous hydrogenation becomes more widely applied when mild conditions, high selectivity, in particular enantio-selectivity, are required. The versatility of the structural features of usable ligands for metal–organic catalysts provides the possibility of a precise control of the hydrogenation procedure in the homogenous mode. The future development of hydrogenation processes exhibits, among all, one tendency to heterogenize the homogenously catalyzed processes by entrapping or anchoring the metal organic active species to suitable solid materials, such as porous inorganic oxides and polymers.

REFERENCES

1. Grochala, W.; Edwards, P.P. Thermal decomposition of the non-interstitial hydrides for the storage and production of hydrogen. Chem. Rev. **2004**, *104*, 1283–1315.
2. Moffet, R.B. 2,2-Dimethylpyrrolidine. Org. Synth. **1963**, *4*, 354–356.
3. Soai, K.; Ookawa, A. Mixed solvents containing methanol as useful reaction media for unique chemoselective reductions within lithium borohydride. J. Org. Chem. **1986**, *51*, 4000–4005.
4. Disselkoetter, H.; Liob, F.; Oedinger, H.; Wendish, D. Liebigs. Ann. Chem. **1982**, *150*.
5. Sabatier, P.; Senderens, J.B. Action du nickel sur l'éthylène synthèse de l'éthane. C. R. Acad. Sci. Paris **1897**, *124*, 1358–1360.
6. Xu, J.; Froment, G.F. Methane steam reforming, methanation and water-gas shift. AIChE J. **1989**, *35*, 88–96.
7. Czernik, S.; French, R.; Feik, C.; Chornet, E. Production of hydrogen from biomass by pyrolysis/steam reforming. In *Advances in Hydrogen Energy*; Gregoire-Padro, C., Lau, F., Eds.; Kluwer Academic/Plenum Publishers: Boston, 2000; 87–91.
8. Smil, V. *Enriching the Earth: Fritz Haber, Carl Bosch, and the Transformation of World Food Production*; MIT Press: Cambridge, MA, 2001.
9. Smidovnik, A.; Kobe, J.; Leskovsek, S.; Koloini, T. Kinetics of catalytic transfer hydrogenation of soybean oil. J. Am. Oil Chem. Soc. **1994**, *71*, 507–511.
10. Satchithanandam, S.; Oles, C.J.; Spease, C.J.; Brandt, M.M.; Yurawecz M.P. Rader, J.I. Trans, saturated, and unsaturated fat in foods in the United States prior to mandatory trans-fat labeling. Lipids **2004**, *39*, 11–18.
11. Castiglioni, G.L.; Fumagalli, C.; Guercio, A.; Lancia, R.; Messori, M.; Vaccari, A. Vapor-phase hydrogenation of maleic anhydride to γ-butyrolactone 2. Role of reaction parameters. Erdöl Kohle-Erdgas-Petrochem. **1994**, *47*, 337–341.
12. Izumi, Y.; Imaida, M.; Fukawa, H.; Akabori, S. Studies on modified hydrogenation catalyst 3. Asymmetric hydrogenation with modified Raney Nickel. Bull. Chem. Soc. Jpn. **1963**, *36*, 155–160.
13. Orino, Y.; Imai, S.; Niwa, S. Asymmetric hydrogenation of methyl pyruvate using Pt/C catalyst modified with cinchonidine. Nippon Kagaku Kaishi (J. Chem. Soc. Jpn.) **1979**, 1118–1120.
14. Margitfalvi, J.L.; Hegedüs, M.; Tfirst, E. Enantioselective hydrogenation of α-keto esters over cinchona-Pt/Al_2O_3 catalyst. Kinetic evidence for the substrate-modifier interaction in the liquid phase. Tetrahedron Asymm. **1996**, *7*, 571–580.
15. Blaser, H.-U.; Jalett, H.-P.; Müller, M.; Studer, M. Enantioselective hydrogenation of α-ketoesters using cinchona modified platinum catalysts and related systems: a review. Catal. Today **1997**, *37*, 441–463.
16. Topsoe, H.; Clausen, B.S.; Massoth, F.E. Hydrotreating catalysis. Catal. Sci. Tech. **1996**, *11*, 1–310.
17. Whitehurst, D.D.; Isoda, T.; Mochida, I. Present state of the art and future challenges in the hydrodesulfurization of polyaromatic sulfur compounds. Adv. Catal. **1998**, *4*, 345–471.
18. Lindlar, H.; Dubuis, R. Palladium catalyst for partial reduction of acetylenes. Org. Synth. **1973**, *5*, 880–882.
19. Brown, H.C.; Brown, C.A. The reaction of sodium borohydride with nickel acetate in ethanol solution—a highly selective nickel hydrogenation catalyst. J. Am. Chem. Soc. **1963**, *85*, 1005–1006.
20. Blauwhoff, P.M.M.; Gosselink, J.W.; Kieffer, E.P.; Sie, S.T.; Stork, W.H.J. Zeolites as catalysts in industrial processes. In *Catalysis and Zeolites*; Weitkamp, J., Puppe, L., Eds.; Springer: Berlin, 1999; 437–538.
21. Osborn, J.A.; Jardine, F.H.; Young, J.F; Wilkinson, G. Preparation and properties of tris(triphenylphosphine)halogenorhodium (1) and some reactions thereof including hydrogenation of

olefins and acetylenes and their derivatives. J. Chem. Soc. A **1966**, 1711–1718.
22. Chatt, J.; Coffey, R.S.; Shaw, B.L. Hydrocomplexes of iridium(III) stabilized by tertiary phosphines and arisines. J. Chem. Soc. **1965**, 7391–7395.
23. Giadello, M.A.; Conticello, V.P.; Brard, L.; Gagne, M.R.; Marks, T.J. Chiral organolanthanides designed for asymmetric catalysis—a kinetic and mechanistic study of enatioselective olefin hydrogenation/cyclization and hydrogenation by C-1-symmetrical $Me_2Si(Me_4C_5)(C_5H_3R^*)Ln$ complexes. J. Am. Chem. Soc. **1994**, *116*, 10241–10254.
24. Nishimura, S. *Handbook of Heterogenous Hydrogenation for Organic Synthesis*; John Wiley & Sons, Inc.: New York, 2001.
25. Knowles, W.S.; Christopfel, W.S.; Koenig, K.E.; Hobbs, C.F. Studies of asymmetric homogeneous catalysts. Adv. Chem. Ser. **1982**, *196*, 325–336.
26. Bosnich, B.; Roberts, N.K. Asymmetric catalytic-hydrogenation. Adv. Chem. Ser. **1982**, *196*, 337–354.
27. Herron, N. A cobalt oxygen carrier in zeolite Y: a molecular ''ship in a bottle''. Inorg. Chem. **1986**, *25*, 4714–4717.
28. Shyu, S.G.; Cheng, S.W.; Tzon, D.L. Immobilization of $Rh(PPh_3)_3Cl$ on phosphinated MCM-41 for catalytic hydrogenation of olefins. Chem. Commun. **1999**, 2337–2338.
29. Bailey, D.C.; Langer, S.H. Immobilized transition-metal carbonyls and related catalysts. Chem. Rev. **1981**, *81*, 109–148.

Hydrogenation Reactions in Dense Gas Systems

Gary Combes
Fariba Dehghani
Raffaella Mammucari
Neil R. Foster
School of Chemical Engineering and Industrial Chemistry, The University of New South Wales, Sydney, New South Wales, Australia

INTRODUCTION

Incremental improvements across all facets of the processing industries enable companies to maintain a competitive advantage. These improvements are often driven by the need to conform to more stringent environmental standards. The substantial cost of finding alternate reaction pathways can be avoided by optimizing current synthesis processes or the product recovery stages of a process. The potential of supercritical fluids (SCFs) to reduce processing waste has been demonstrated for various commercial separation processes. More recently, SCFs have been shown to provide a suitable solvent environment for synthesis, in which reaction rates and selectivities are enhanced. With the potentially high capital costs compensated by the superior downstream separation, processing, and minimization of process waste products, the use of SCFs for integrated synthesis processes has become an attractive area of research. In this entry, recent applications of dense gases for reactions, particularly hydrogenation reactions, have been reviewed.

BACKGROUND

The major applications of dense gases, fluids above or in the vicinity of the critical point, have been for the extraction of specific compounds from solid matrices or liquid mixtures. Well-known examples include the decaffeination of coffee, extraction of olfactory compounds for the perfume and food industries, the extraction of essential oils, and the use of SCFs for the extraction of heavy fractions from distillates. These processes benefit from the selectivity of dense gases. A variety of new applications has been identified for which the properties of dense gases, other than the solvent power, are utilized. One field is the utilization of dense gases as a reactant for reactions. The dense gas processes for the production of low density polyethylene (LDPE) by free radical polymerization, and more recently, the manufacture of fluoropolymers have annual outputs measured in bulk tonnages.

Generally though, the enhanced kinetics, selectivity, and/or separation provided by a SCF process must be traded off against the generally low solubility of components in SCFs. The result is that supercritical technology is generally applied to high value, low volume products that are typical of the fine chemical industry. Potential product examples include catalysts, pharmaceuticals, food additives, electronics, specialty polymers, graft copolymers, and fluoropolymers.

Much of the focus of early reaction research centered on reactions in supercritical water ($T_c = 647.2\,K$, $P_c = 220.5\,bar$), with some promising results arising from these studies in the area of waste destruction. The major disadvantage of using supercritical water as a reaction medium is that the reaction environment strongly promotes corrosion, which combined with high temperatures and pressures, leaves only the most technologically demanding processes economically viable. Innovations in reactor design are improving the economics of supercritical water oxidation technology.

The technological challenges posed by supercritical water do not exist for many other SCFs. The critical conditions required for other SCFs are less extreme. Supercritical methanol ($T_c = 512.6\,K$, $P_c = 80.8\,bar$) has been used to break down poly(ethylene terephthalate) (PET) waste matter to its basic synthesis components for reuse in the original process. Supercritical propane is the favored solvent for hydrogenations of long chain fatty acids and oils because of its superior solvation power for large molecular weight compounds compared to carbon dioxide. As solvents, SC-CO_2, ethane, and nitrous oxide have moderate critical temperatures (i.e., 31.4°C for CO_2). The most prevalent of these solvents, carbon dioxide, is considered environmentally benign, inexpensive, nonflammable, noncorrosive, and leaves no residue after use. These characteristics, combined with the attractive mass transfer properties of SC-CO_2, present a range of opportunities for the synthesis industry.

REACTIONS IN DENSE CO_2

The compilation of published material in Table 1 demonstrates the variety of reactions in which CO_2

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008077

Table 1 Reviews of dense gas synthesis chemistry

Reaction field	Review title	Reference
Uncatalyzed, homogeneous and heterogeneous metallic catalysis	Reactions in supercritical fluids—a review	[57]
	Reactions at supercritical conditions: applications and fundamentals	[58]
	Supercritical carbon dioxide as a substitute solvent for chemical synthesis and catalysis	[59]
	Pharmaceutical processing with supercritical carbon dioxide	[60]
	Recent advances in chemistry and chemical processing in dense phase carbon dioxide at Los Alamos	[61]
	Supercritical carbon dioxide as an innovative reaction medium for selective oxidation	[62]
	Supercritical chemistry: a review of chemical reactions in supercritical fluids solutions	[63]
Enzymatic catalysis	Enzymatic reactions and cell behavior in supercritical fluids	[64]
	Enzyme activity in supercritical fluids	[65]
	Enzymic synthesis in supercritical fluids	[66]
	Lipases in supercritical fluids	[67]
	Supercritical biocatalysis	[68]
Polymerization	Chain polymerization in inert near- and supercritical fluids	[69]
	Polymerizations in liquid and supercritical carbon dioxide	[70]
	Chain growth polymerizations in liquid and supercritical carbon dioxide	[71]

has been applied. Jessop and Leitner[1] and Noyori[2] provide a significant insight into dense gas synthesis chemistry. In some of the more recent works, the application of CO_2 for particle production is combined with CO_2 synthesis to obtain a process low on waste but of optimal efficiency.[3] As has been demonstrated, the unique properties of dense gases could enable the realization of a single process unit for both reaction and product separation.

Multiphase gaseous reaction systems are of prime importance in the pharmaceutical and specialty chemical industries, especially in the area of chiral synthesis. The importance of isomer-selective synthesis in the fine chemical industry is illustrated with the example of the aspartylphenylalanine methyl ester isomers. The (L,L) isomer is the common sweetening agent aspartame, whereas the (L,D) diastereoisomer has a bitter taste. In the pharmaceutical industry, thalidomide is a notable example. In the racemic or (S) isomeric form, the drug is a strong teratogen, whereas the (R) form has been claimed to possess no such side effects.[4] Enzymatic catalysis is also proving to be a significant growth area in supercritical reaction research, hence its justification in being included in the tabled survey of CO_2 reactions. A number of studies have already demonstrated that rates and selectivity for enzyme-catalyzed reactions can be enhanced in supercritical media.[5–7]

Supercritical CO_2 may not be the optimum solvent for a given reaction system. The low solubility of reaction components is perceived to be a major disadvantage of using SC-CO_2 for reactions. This limitation can be overcome through the combination of dense CO_2 and a conventional solvent. The term applied to this technique is gas antisolvent reactions (GASR) and has its origins in the gas antisolvent (GAS) particle process.[8] It is suggested that this area of reaction research offers potential for improvements in current multiphase processes without the need for considerable reengineering of the process. Evidence is provided to support this argument and its potential for future research. Finally, a consideration of some hybrid systems that incorporate not only reactions but also product separation/purification is presented to illustrate the potential for incorporating dense gases into existing processes.

Hydrogenations

Heterogeneous catalysis

The use of dense gases for heterogeneous hydrogenations is attractive given the ease of product purification and the significantly enhanced rates.[9–11] Other solvent

properties, such as viscosity, polarity, and local solvent density effects may also have an impact on kinetics. Heterogeneous hydrogenation reactions conducted in CO_2 offer the possibility of faster reaction rates through improved mass transfer and higher dissolved H_2 concentrations in reactions where the rate is greater than zeroth order in terms of hydrogen. The higher diffusivity and lower viscosity of SCFs may result in a faster removal of product from the catalyst pores, thus preventing catalyst fouling and reducing further side reactions. Improved mass transfer rates are manifest in the extended lifetime of catalysts, the enhanced product selectivity, and the reduction of residence times. The common hydrogenation catalysts of palladium and nickel exhibit very low solubility in SC-CO_2 and are suitable for continuous heterogeneous processes.

A number of heterogeneous hydrogenations have been carried out in CO_2.[12–15] Kröcher, Koppel, and Baiker[14] considered the heterogeneous catalytic synthesis of (*n,n*)-dimethylformamide (DMF) from SC-CO_2 using a variety of heterogeneous metal catalysts. From this study, silica matrix–stabilized ruthenium catalysts proved to be the most active and robust group 8 metal for the hydrogenation of CO_2 compared to metal catalysts such as iridium, palladium, rhodium, platinum, and cobalt. Turnover numbers [TON = (mole product)/(mole catalyst)] for the DMF synthesis were over three orders of magnitude greater than for the previously reported heterogeneous studies, although the TON were still less than those recorded in supercritical homogeneously catalyzed systems.

The relatively inert nature of CO_2 does not preclude it from deactivating catalysts. Minder et al. found that SC-CO_2 deactivated the platinum catalyst when used as a solvent for the hydrogenation of ethyl pyruvate.[16] Catalyst activity was maintained, however, when propane was used. The reaction rate in supercritical propane was significantly higher compared to that achieved in toluene. Tacke, Wieland, and Panster considered the hydrogenation of fatty acids in CO_2 with a variety of catalysts and reactor regimes using the commercial DELOXAN® supports.[15] The higher yield for the polysiloxane-supported 1% palladium catalyst bed compared to a 2% platinum catalyst on the same support in SC-CO_2 may have been because of catalyst deactivation by CO_2.[15] Van der Hark and Härröd observed significantly higher TON and much improved selectivities for *cis* product in the hydrogenation of fatty acid methyl esters in supercritical propane using conventional palladium or nickel catalysts.[17] These studies are evidence to show that catalyst technology plays an integral part in supercritical reactions and that CO_2 will not always be the best supercritical solvent.

The study by Hitzler et al. has illustrated the broad scope of potential hydrogenation reactions in SC-CO_2.[18] Some of the substrates investigated included *m*-cresol, benzaldehyde, acetophenone, 1-octene, and cyclohexene. Reactions were performed in 5 and 10 ml packed bed reactors and space times of up to $300\,hr^{-1}$ were achieved. Residence times of this magnitude render the issue of scale-up largely irrelevant for small (kilogram) quantities of product in a continuous process. This is advantageous for the commercialization of SCF reaction processes because it minimizes development and capital costs.

Hitzler et al. found that the product distribution could be varied by varying system parameters such as reactor temperature, hydrogen : substrate feed ratios and catalysts.[18] An example of this was the hydrogenation of acetophenone over a 5% Pd APII DELOXAN catalyst to ethylbenzene (EB), ethylcyclohexane (ECH), α-methyl benzenemethanol (MBM), and α-methyl-cyclohexanemethanol (MCH). Varying the temperature from 363 to 573 K at constant CO_2 flow (which altered the substrate : hydrogen ratio and residence time) resulted in a dramatic change in the selectivity toward the products EB, ECH, and MBM as shown in Table 2. The fact that the temperatures of these reactions were far higher than the critical temperature of CO_2 suggest that mass transfer characteristics (chiefly low viscosity) and chemical effects were significant factors.

There are few comprehensive kinetic studies included in the heterogeneous hydrogenation investigations conducted to date.[12,13] The published works by Bertucco et al.[12] and Devetta et al.[13] are notable because the investigations incorporate phase equilibria

Table 2 Conversion and selectivity for the hydrogenation of acetophenone in SC-CO_2

			Selectivity (%)			
Temperature (K)	**Sub : H_2**	**Conversion**	**MBM**	**EB**	**MCH**	**ECH**
363	1 : 2	97.5	90.0	7.5	0	0
453	1 : 5	100	14.0	41.0	28.0	17.0
473	1 : 3	91.5	0	73.0	0	18.5
573	1 : 6	94.0	0	4.5	0	89.5

(From Ref.[18].)

studies combined with equation of state (EOS) modeling to complete the kinetic studies for the system. In all cases, the reaction was the hydrogenation of an unsaturated ketone over an alumina-supported palladium catalyst. The ketone was a vitamin intermediate. The initial study was conducted in a modified internal recycle reactor to obtain perfectly mixed conditions.[12] The kinetic data were then used in subsequent modeling of a trickle-bed reactor in both two-and three-phase systems.[13] The aim of using SC-CO_2 in this instance was to enhance the solubility of hydrogen. As a consequence of the relatively extreme temperatures and pressures required to homogenize the fluids, three phases were present in the system. The results predicted by the kinetic models were in good agreement with the experimental data, largely because of the accurate quantitative evaluation of phase equilibria for the system.

Homogeneous catalysis

Supercritical media are well suited for conducting homogeneous catalytic reactions because of the absence of phase boundaries and the improved mass transfer rates. In the majority of the published kinetic works, CO_2 has been used as both the reactant and solvent to produce formic acid, methanol, formamide, and ethanol.[19–22] Jessop, Joo, and Tai[19] demonstrated the esterification of formic acid in a single stage process via the homogeneous catalytic hydrogenation of CO_2, in the presence of methanol and triethylamine using a variety of ruthenium-based catalysts. Not only was the methyl formate synthesis, a two-step reaction process, carried out concurrently in a single vessel, but the TONs were an order of magnitude higher and the equilibrium yields were greater than any reported for the ester synthesis. These results were achieved at 323 K, which is significantly lower than the 373–448 K used in prior studies.[23–25]

Novel organometallic compounds such as $Cr(CO)_5(H_2)$,[26] $(C_5R_5)Mn(CO)_2(\eta^2\text{-}H_2)$, and $(C_5H_5)Mn(CO)_2(\eta^2\text{-}H_2)$[27] have been synthesized and isolated in traditional flow reactors using SC-CO_2. Many other compounds have been identified through in situ infrared monitoring of the reactions. The work was notable because it combined synthesis and product purification in a very simple, though effective arrangement. Supercritical CO_2 was found to be a more versatile and stable environment for ligand manipulation than conventional solvents, which may allow the synthesis of further novel organometallics than predicted for conventional solvents. The reaction rate for the synthesis of $Cr(CO)_5(H_2)$ by Poliakoff, Howdle, and George was an order of magnitude faster in CO_2 than in *n*-heptane.[26] The rate enhancement was ascribed to the improved solubility of H_2 in SC-CO_2 compared to the liquid *n*-hexane and the weaker interactions between solute and reaction intermediates in the supercritical medium compared to the organic medium.

The improved reaction rates achieved in the SC-CO_2 medium combined with the variety of substrates considered so far illustrate the potential of this technology. As an indication of the economic viability of CO_2 processing in the fine chemical industry, a plant has been constructed at Thomas Swan and Co. in the UK.[28] The plant has been built for hydrogenation processes to produce fine chemicals and pharmaceutical intermediates to a rating of 1000 kg hr^{-1} of CO_2 at pressures up to 500 bar. However, the significant capital cost of supercritical processes, combined with the complex production and expensive nature of chiral products, demand that SC-CO_2 be considered for asymmetric hydrogenation.

Asymmetric homogeneous catalysis

Given the improved reaction rates of nonselective catalyzed reactions already achieved with SC-CO_2, it is not surprising that asymmetric hydrogenations have been a focal point of study. Economic and environmental regulatory guidelines are two of the major driving forces behind the efforts in the chemical industry to maximize efficiency and to minimize waste. The traditional methods for producing optically pure compounds involve preferential crystallization or diastereomeric crystallization, kinetic resolution, and catalytic asymmetric synthesis. Catalytic asymmetric synthesis has significant advantages over the other methods mentioned, including waste minimization and a reduction or elimination of separation requirements with associated savings in energy consumption. The global chiral pharmaceutical market was valued in excess of US $70 billion annually (as of 1997), making it an ideal candidate for SC-CO_2 reaction technology.[4]

Synthesis of optically pure compounds has been a long established challenge in chemistry. The focus has been on improving the catalysts involved in the reactions because of the difficult nature of separating chiral compounds.[29,30] A broad range of catalysts are available with very high selectivities in many organic solvents. The use of SCFs (in particular CO_2) as reaction media was investigated in 1995.[31] The small amount of available published data is indicative of the infancy of this research. However, a review of results to date highlights why carbon dioxide is a potential alternative to organic solvents for dissolved catalytic reactions or homogeneous catalysis.

The first published example of asymmetric hydrogenation was by Burk et al.[31] The enantioselective hydrogenation of α-enamides with cationic rhodium (R,R) Et-DuPHOS 1,2-bis(*trans*-2,5-diethylphospholano) benzene as a catalyst was investigated. To improve the

solubility of the catalyst, the very lipophilic counteranion, tetrakis(3,5-bis(trifluoromethyl)phenyl) borate (or BARF) was used. The enantiomeric excesses (e.e.) obtained were comparable to those achieved in conventional solvent systems of methanol and hexane. (Enantiomeric excess is a common measure of purity for one enantiomer over another, e.g., S over R is defined as $(S - R)/(S + R)$.) The results were catalyst and substrate selective, suggesting that steric and solvent effects were major factors.

The hydrogenation of β,β-disubstituted α-enamides by Burk et al. realized enantioselectivities that were significantly higher in SC-CO_2 than in conventional solvents.[31] The high enantioselectivities were shown not to be solely a function of the higher system pressures, but more a result of solvent effects and the improved hydrogen solubility in SC-CO_2 compared with organic solvents.[32] These observations were in agreement with the results obtained by Sun et al., who found that for H_2 limited reactions, the main kinetic factor controlling enantioselectivity in asymmetric catalytic hydrogenation was the H_2 concentration in solution.[33] Whilst the study by Sun et al. was applied to conventional solvents and not supercritical systems, it was a key factor for promoting the use of supercritical media such as CO_2.

A second investigation by Xiao et al. considered *trans*-2-methyl-2-butenoic acid as the substrate and a variety of ruthenium BINAP catalysts (BINAP is 2,2′-bis(diphenylphosphino)-1,1′-binaphtyl).[34] The results showed that the reaction in SC-CO_2 produced selectivities and yields that were comparable to those in methanol and hexane. The reaction took approximately twice as long in SC-CO_2 as in methanol, although no explanation was provided. Bonding of CO_2 to the catalyst may have inhibited, although not prohibited, the reaction. Equally, strong solvent–solute interactions may have inhibited the reaction. The presence of fluorinated alcohols enhanced the selectivity to levels above those achieved in neat methanol, possibly as a result of electronic effects and/or improved solubility. The lower pressures of hydrogen used in the SC-CO_2 trials still gave comparable yields, indicating that the miscibility of hydrogen in SC-CO_2 was not a significant factor. The limited results do not permit conclusions to be drawn on the relative rates or the true effect of hydrogen pressure.

In the work with asymmetric hydrogenation in SC-CO_2, Kainz et al. used modified iridium catalysts to hydrogenate imines.[35] The presence of BARF as a counterion provided the best results in CO_2. In comparison to dichloromethane, CO_2 gave up to double the turnover frequency (TOF = TON/time) but enantioselectivities were lower by ~10%. Again yield and selectivity were substrate specific. Lange et al. developed a chiral rhodium diphosphinite catalyst complex, which also used the BARF counterion and subjected it to a mechanistic study.[36] The hydrogenation of di-methylitaconate was conducted in both hexane and CO_2 and the results were compared. The enantioselectivity in CO_2 was comparable to that of hexane. However, the reaction was slower by a factor of five. Analysis by NMR indicated that CO_2 did not bond with the reaction intermediates. Thus, it was concluded that the solvent was either influencing the availability of the hydrogen at the reaction site or affecting specific steps in the catalytic cycle through changes in the stability of transition states.

There has been no comparative work on hydrogenation rates for solubilized compounds in CO_2 and conventional solvents to date. Most of the investigations mentioned above have shown technical feasibility, which for the synthesis of fine chemicals on the kilogram scale, is probably sufficient. An improved understanding of the reaction kinetics is necessary for the field to progress. In this respect, knowledge of catalyst solubility in CO_2 is required. An overview of work in conventional liquid solvents serves to highlight that ruthenium-, rhodium-, and iridium-based catalysts often provide improved selectivities and activities for hydrogenations. A comprehensive study along the lines of those reviewed by Jessop and Leitner;[1] Jessop, Joo, and Tai;[19] and Darr and Poliakoff[37] made on the solubility of various chelated transition metal catalysts in CO_2 and other supercritical mediums would be of immense use for future work in this area. Palo and Erkey determined the solubility of both dichloro bis(triphenylphosphine) nickel(II) and Wilkinson's catalyst in SC-CO_2 to be of the order of 0.01 mM.[38] The typical concentration used in homogeneous catalysis is about 1 mM, suggesting that catalyst development for CO_2 is essential to enhance the potential of CO_2 for selective catalytic reactions.

Hydroformylation

Hydroformylation reactions have been one of the most well researched areas of CO_2 reaction chemistry. Hydroformylation reactions are necessary for the formulation of complex chemicals. The first complete kinetic study of a hydroformylation reaction was in CO_2 and was first published in 1999.[39] Prior to this, most studies had considered the effect of dense CO_2 on linear : branch ratios or other forms of selectivity. Carbon dioxide has an effect on the selectivity of a variety of hydroformylation reactions and can enhance the rate of reaction[40] Hydroformylation is by its nature regioselective and typically the linear : branch or "*n : iso*" ratio is used as the measure of selectivity. The use of asymmetric catalysts to achieve chiral products has introduced a second degree of selectivity to catalyst design. Advancements in catalyst design, together with solvent selection, are expected to make

larger gains in selectivity, and thus simplify the separation and purification steps downstream.

Palo and Erkey determined a kinetic expression for the hydroformylation of 1-octene in SC-CO_2.[39] The ''CO_2'' phyllic catalyst developed was *trans*-RhCl(CO)$(P(p\text{-}CF_3C_6H_4)_3)_2$, which obtained an estimated solubility of at least 7.6 mM in SC-CO_2.[39] The observed solubility was a 100-fold improvement on that achieved with nonfluorinated catalysts. A half order dependence on the hydrogen concentration was observed, which contrasts markedly with results in conventional solvents, where the order with respect to hydrogen has always been one or greater.[41] The result suggests that either the improved hydrogen solubility or the suspected improved mass transfer rates had altered the rate-limiting step. Conversely, solvent effects such as higher local densities or the weaker solvent cage may have contributed to a change in the rate-limiting step in the mechanism. The evaluation and comparison of the Arrhenius equation for this reaction in CO_2 and conventional solvents would assist in determining the principal factor.

The work also revealed that unlike in conventional systems, there was no substrate inhibition and no critical catalyst concentration loading, which is often brought about by the formation of an inactive catalytic dimer species. The novel catalyst or the presence of CO_2 was suggested to be the cause of this effect. Carbon monoxide was found to have a strong effect on the selectivity of linear vs. branched product in the supercritical environment.

A comparison with other literature results highlighted the solubility issues that arise when comparing CO_2 with conventional solvents. As shown in Table 3, the concentrations of the gases are higher in CO_2 than in organic solvents. The substrate concentration in CO_2 is almost an order of magnitude lower than those in conventional solvents. The fine chemical and pharmaceutical industries, where this technology will be most applicable will obtain typically lower substrate solubility because the substrates will be far more complex CO_2-phobic molecules.

The improved catalytic efficiency of homogeneous hydroformylation reactions in CO_2 has been proven experimentally. The studies by Koch and Leitner of the homogeneous hydroformylation of 1-octene with the catalyst [(cod) Rh(hfacac)] (cod = cyclo-octadiene and hfacac = hexafluoroacetylacetonate) produced TON in CO_2 that were three times higher than those in toluene.[42] The addition of perfluoralkyl-substituted triarylphosphine and triarylphosphite ligands to the system produced improved regioselectivities of *n* : *iso* product compared to those achieved in conventional solvents. The side reaction of olefin isomerization that occurs in phosphite-modified systems in conventional solvents was totally suppressed in the dense gas environment. In addition, the work demonstrated catalyst activity for a number of substrates in liquid CO_2. The relative rates for the hydroformylation of the sterically hindered *trans*-3-hexene were at least five times greater in CO_2 than toluene under the same conditions. The results clearly indicate that the dense gas media can effect beneficial changes to the reaction chemistry.

Further catalyst development has occurred based on the improved regioselectivity of the perfluoroalkyl-substituted ligands for use in homogeneous hydroformylations in CO_2.[43,44] Kainz et al. developed a perfluoroalkyl-substituted arylphosphane ligand that achieved solubility improvements in CO_2 of several orders of magnitude.[43] A high degree of conversion and a high selectivity for *n* : *iso* product with no hydrogenation side reaction characterized the successful catalyst development. Franciò, Wittmann, and Leitner used the (R,S)-3-H_2F_6-BINAPHOS ligand complexed to [Rh(acac)] to perform the catalyzed hydroformylation of vinyl arenes in SC-CO_2.[44] Regioselectivity of 93% and enantioselectivity in excess of 90% were consistently attained at temperatures ranging from 308 to 333 K and pressures of 115–242 bar. The yields and rates were comparable to those in benzene and proved to be far superior than the results achieved in hexane. The yields in CO_2 systems were comparable with benzene at high substrate : catalyst loadings of 2000. The acceptance of CO_2 as a replacement solvent is more likely in this case given the environmental and health issues that benzene presents and the absence of solvent residue in the product. These studies indicate

Table 3 Comparison of concentrations used in hydroformylation experiments in CO_2 and conventional solvents

		Concentrations (10^{-3} mol dm^{-3})			
Solvent	**Substrate**	**H_2**	**CO**	**Catalyst**	**Substrate**
CO_2	1-Octene	130–2600	130–2200	0.63–0.25	0.04–0.96
Ethanol	1-Hexene	21.4–62.1	6.2–153	0.33–1.2	0.20–1.56
Toluene	1-Dodecene	8.9–30.8	8.4–121	1.0–8.0	0.18–2.2
Toluene	Ethylene	2.7–44.7	14.0–149	0.5–4.0	0.045–1.45

(From Ref.[72].)

that the considerable potential for homogeneous hydroformylation chemistry in CO_2 requires the concurrent development of catalyst technology.

Phase Transfer Reactions

Phase transfer catalyzed (PTC) reactions consist of two immiscible phases with the transfer of catalyst between the two phases such as liquid–liquid or solid–fluid phase. The reaction has been known to occur in a catalyst rich third phase. The PTC reactions are acknowledged as being a significant subset of reaction chemistry. The PTC industry is estimated to generate products worth billions of dollars annually in areas such as polymers, pharmaceuticals, and agrichemicals, although this value is not yet reflected in the degree of supercritical research into this area. The higher diffusion rates of solutes (both catalysts and reactants) through the dense CO_2 and the suspected improved mass transfer rates across the phase boundaries are features of PTC reaction that could benefit from the utilization of CO_2. The significantly lower solubility in CO_2 for both reactants and catalysts will impact on its overall economics.

The use of ionic liquids in combination with CO_2 has the potential to produce cleaner processes with improved selectivity.[45–48] The negligible miscibility of the ionic liquid in CO_2 compared with appreciable amounts of CO_2 that can be found in the liquid phase make the use of CO_2 as a ''green'' solvent attractive for continuous reaction processes. Sellin, Webb, and Cole-Hamilton[45] conducted a hydroformylation of hex-1-ene and 1-octene catalyzed by rhodium based catalyst in 1-butyl-3-methylimidazolium hexafluorophosphate (BMIMHF) in contact with CO_2. Improved *n* : *iso* product selectivity was obtained, compared with that using toluene with similar selectivity, but substantially lower yield (40% compared to >99%). Using 1-octene as a substrate and $[Rh_2(OAc)_4]$/[1-propyl-3-methylimidazolium] $[PhP(C_6H_4SO_3)_2]$ as catalyst, over 20 hr of continuous operation was achieved with minimal catalyst leaching at 373 K.

Brown et al. have conducted asymmetric hydrogenations of tiglic acid with $Ru(O_2CMe)_2$((R)-toluene BINAP) in BMIMHF with water using CO_2 to extract the product.[49] Catalyst activity and enantioselectivity were maintained through five sequential experiments. Liu et al. studied the hydrogenation of CO_2 with the catalyst $RuCl_2(Ph_2PCH_2CH_2PPh_2)_2$ in the presence of dialkylamines to produce *N,N*-dialkylformamides in the system of SC-CO_2/BMIMHF.[50] The results indicated that the selectivity for amines other than dimethylamine was higher in the SC-CO_2/BMIMHF system compared to conventional solvent systems. It was determined that CO_2 offered no advantage in terms of reaction rate for simple hydrogenations over hexane as a solvent. However, the separation step of the product from the solvent would be simpler and considerably less expensive using CO_2.

Reactions in a CO_2-Expanded Phase

The replacement of conventional solvents with CO_2 for hydrogenation reactions theoretically promises improved reaction rates with the improved gas miscibility. It is a different case when the optimum temperature lies below the critical temperature of CO_2. The solubility of gaseous hydrogen in a liquid phase increases with increasing temperature. If a reaction produces a higher selectivity for the desired product at lower temperatures, then the efficiency of the process becomes more complex. A conundrum exists whereby the reaction rates may be increased at the expense of selectivity, or the purification steps downstream are minimized but at the cost of a considerably slower reaction rate.

When phase equilibria are considered together with kinetic rate expressions, a dense gas expanded medium can be a viable alternative compared to SCF technology. There is a strong case that dense CO_2-expanded solvents may enhance the gas solubility in the liquid phase at temperatures below the critical point of CO_2. The issue that has been the commercial downfall of many potential CO_2 applications remains the poor solubility of most potential reactants in CO_2. It is possible that by using CO_2 in conjunction with conventional solvents, the solubility issue can be addressed. A significant advantage is also realized when it comes to product recovery and purification compared with the use of nonexpanded solvent. An added benefit is the reduction in organic solvent requirements that results through the use of CO_2-expanded media.

Given that the dense gas is miscible with the solvent, the degree of expansion has been found to be a function of the fraction of CO_2 in the liquid phase as shown in Fig. 1.[51] An arbitrary minimum for the amount of CO_2 required in a GASR system is defined by the fact that appreciable expansion is only gained at mole fractions beyond 60% of CO_2. The use of CO_2-expanded solvents theoretically provides the optimum tuning medium for reaction rates. The benefit of large scale operation is that it potentially affords the fine chemical industry the means to combine reaction chemistry together with simpler separation techniques. Significant improvements in processes, both current and new, can be achieved at significantly milder conditions than those required for homogeneous supercritical reactions.

The hydrogenation of CO_2 to form formic acid serves to illustrate the potential of expanded systems. Jessop et al. found that whilst the rate of hydrogenation

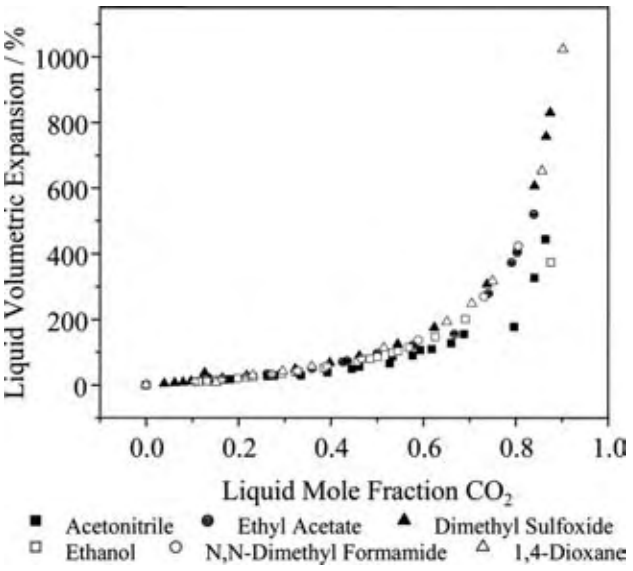


Fig. 1 Expansion curves of a variety of organic solvent/CO_2 systems at 298 K. (From Ref.[51].)

was fastest in the homogeneous phase, the TOF for the hydrogenation reaction in a two-phase mixture of methanol and CO_2 was still significantly faster than neat methanol.[19,20] The TON and TOF for the production of formic acid from CO_2 are shown in Table 4. Other solvents including dimethylsulfoxide (DMSO), acetonitrile, and tetrahydrofuran (THF), all of which expand with the addition of CO_2 were considered for the same reaction. The initial rate of hydrogenation in the two phase CO_2/DMSO system was comparable to that of the single CO_2/DMSO phase and CO_2/methanol system.

There have been few instances where a study has been focused on an expanded solvent as a reaction medium.[52–54] Kerler et al. reported that oxidations in a CO_2-expanded environment were more efficient than the corresponding reaction in SC-CO_2.[53] Kayaki, Naguchi, and Ikariya investigated the arylation of ethylene with iodobenzene using $PdCl_2\{P(OC_6H_5)_3\}_2$ and base.[55] The reaction was run over 18 hr at 403 K. The selectivity for styrene was increased from 57% in the absence of CO_2 to 83% in CO_2 at 101 bar. Selectivity increased at the expense of conversion as the CO_2 pressure was increased from 101 to 151 bar, which corresponded to the progression of the system from a binary mixture to a supercritical phase. Further studies with catalysts of higher solubility in the supercritical phase failed to provide higher TONs. The conclusion was drawn that the CO_2 was extracting the styrene from the liquid phase, thus preventing further reaction with the styrene. These studies were conducted at low concentrations and hence the true effect of the CO_2 on the reaction was not determined.

Combined Reaction and Separation Using Dense CO_2

The synthesis and purification processes are often numerous and complex. Dense gases present the possibility of combining the reaction and separation operations into a single integrated process and may provide a convenient way of removing the impurities through improved reaction control.

Darr and Poliakoff conducted one of the first studies where SC-CO_2 was used in a two-stage continuous reaction and purification process to form pure organometallic products.[55] The *cis*-$W(CO)_4$(pyridine)$_2$ or novel *cis*-$W(CO)_4$(3,5 lutidine)$_2$ were formed in a continuous flow reactor at 488 K and were then removed from solution via precipitation, with a second flow of CO_2 used to remove the excess solvent and reactant for reclamation and recycle. The benefits of this system, aside from the production of novel materials, include reduction of organic solvent consumption, decrease in number of stages and time, and production of 100% purity product.

The asymmetric catalytic hydrogenation of 2-(6′-methoxy-2′-naphthyl)propenoic acid to (S)-naproxen was monitored in both methanol- and CO_2-expanded methanol.[8] The catalyst used was [dichloro-(S)-(-)-2,2′-bis(diphenyl-phosphino)-1,1′-binapthyl]ruthenium(II). Addition of CO_2 to the methanol produced strong retardation of the reaction rate. An average reduction in enantioselectivity of 6% was observed compared to that in methanol. The enantioselectivity was found to increase as the temperature decreased in both neat methanol and CO_2-expanded methanol systems, indicating that the underlying mechanisms were similar. Insufficient oxygen removal may have prevented the

Table 4 Relative rates for the hydrogenation of CO_2

Catalyst precursor	Additive	No. of phases	TON	TOF (hr^{-1})
$RuH_2[P(C_6H_5)_3]_4$	H_2O	1	80	80
$RuH_2[P(C_6H_5)_3]_4$	Methanol	2	770	1500
$RuH_2[P(CH_3)_3]_4$	H_2O	1	1900	630
$RuH_2[P(CH_3)_3]_4$	Methanol	2	760	1500

(From Ref.[19].)

reaction from progressing to complete conversion and also retarded the reaction rate. Spectroscopic studies indicated that the effect of local density enhancements on the thermodynamic state of reaction components contributed to the retardation of the reaction in the expanded solvent media as well. While low enantioselectivity and reaction rate were achieved for this hydrogenation reaction in the CO_2-expanded methanol, the study does not close the avenue for further research in the area of hydrogenation reaction in gas expanded solvent systems.

Recent work by Warwick et al. for an alternate synthesis and purification process for a novel pharmaceutical have built upon the concept.[3] Warwick et al. investigated the effect of dense CO_2 as an antisolvent for improving yields of the synthesis of the nonsteroidal anti-inflammatory drug, copper indomethacin. The conventional synthesis is currently performed by dissolving reactant, copper acetate, and indomethacin in dimethylformamide (DMF) followed by using ethanol as an antisolvent to precipitate the product, including copper indomethacin and removing acetic acid as a by-product. It is important to remove the residual of indomethacin (reactant) from the product (i.e., copper indomethacin) because of the noted side effect of the former. The replacement of ethanol with CO_2 decreases dramatically the volumetric requirements of antisolvent and cuts processing time by 80%. The recovery of the antisolvent is considerably simpler using CO_2 compared to ethanol. By considering the solubilities of the individual components and utilizing the optimum reactant ratio, the yield can be improved to almost 100% whilst simultaneously achieving improved purity of product. The use of CO_2 also produced a significant reduction in residual solvent levels.

The combination of reaction and product separation into the desired micronized form is desirable from a processing view. Warwick[56] extended this work to using an aerosol solvent extraction system (ASES) arrangement. It was highlighted that for very fast reactions, ASES is a viable process for synthesizing and separating product in the single vessel. Using the component solubility of the system and a dense CO_2-based ASES arrangement, the product was synthesized and micronized in one vessel. Little or no solvent residue was realized by repeated washing of the micronized product.

CONCLUSIONS

It is clear that there is a wide scope for performing reactions in CO_2 and expanded solvents. An improved knowledge of solvent effects on reaction rates and selectivity at high pressure will enable targeted research to be applied to specific reactions to achieve improvements in synthesis processes.

The application of CO_2 to the considerably larger body of catalyzed reactions provides a greater challenge to our understanding of CO_2 chemistry. In many studies, technical feasibility of the reactions has been established and in several, complete investigations of the kinetics have been conducted. Utilizing dense gases for reactions often provides an advantage with the subsequent separation and purification processes being more efficient than the conventional solvent systems. In many cases, reaction rates and selectivity in a dense gas environment were improved compared to the conventional synthesis process conducted in an organic solvent. Most reaction investigations using CO_2 have typically been performed at dilute conditions to avoid solubility issues. The limits of a reaction system are rarely explored and hence the boundaries of a process (which include not only the reaction steps but the product recovery and purification) are never completely defined. Using dense gases for reactions should always be evaluated from the perspective of the complete process, rather than for just one stage.

REFERENCES

1. Jessop, P.g.; Leitner, W. *Chemical Synthesis Using Supercritical Fluids*; Jessop, P.g., Leitner, W., Eds.; 1999; 480 pp.
2. Noyori, R. In *Supercritical Fluids;* Chemical Reviews; Michl, J., Ed.; American Chemical Society: Washington, 1999; Vol. 99, 353–633.
3. Warwick, B.; Dehghani, F.; Foster, N.R.; Biffin, J.R.; Regtop, H.L. Synthesis, purification, and micronization of pharmaceuticals using the gas antisolvent technique. Ind. Eng. Chem. Res. **2000**, *39* (12), 4571–4579.
4. Stinson, S.C. Chiral drug market shows signs of maturity. Chem. Eng. News **1997**, *20* (10), 38–70.
5. Romero, M.D.; Calvo, L.; Alba, C.; Habulin, M.; Primozic, M.; Knez, Z. Enzymatic synthesis of isoamyl acetate with immobilized *Candida antarctica* lipase in supercritical carbon dioxide. J. Supercritic. Fluids **2005**, *33* (1), 77–84.
6. Nakamura, T.; Toshima, K.; Matsumura, S. One-step synthesis of n-octyl β-D-xylotrioside, xylobioside and xyloside from xylan and n-octanol using acetone powder of *Aureobasidium pullulans* in supercritical fluids. Biotechnol. Lett. **2000**, *22* (14), 1183–1189.
7. Celia, E.; Cernia, E.; Palocci, C.; Soro, S.; Turchet, T. Tuning *Pseudomonas cepacea* lipase (PCL) activity in supercritical fluids. J. Supercritic. Fluids **2005**, *33* (2), 193–199.

8. Combes, G.B.; Dehghani, F.; Lucien, F.P.; Dillow, A.K.; Foster, N.R. Asymmetric catalytic hydrogenation in CO_2 expanded methanol—an application of gas anti-solvent reactions (GASR). In *Reaction Engineering for Pollution Prevention*; Abraham, M.A., Hesketh, R.P., Eds.; Elsevier, 2000; 173–181.
9. Hitzler, M.G.; Smail, F.R.; Ross, S.K.; Poliakoff, M. Selective catalytic hydrogenation of organic compounds in supercritical fluids as a continuous process. Org. Process Res. Dev. **1998**, *2* (3), 137–146.
10. Ramirez, E.; Zgarni, S.; Larrayoz, M. A.; Recasens, F. Short compilation of published reaction rate data for catalytic hydrogenations in supercritical fluids. Eng. Life Sci. **2002**, *2* (9), 257–264.
11. Grunwaldt, J.-D.; Wandeler, R.; Baiker, A. Supercritical fluids in catalysis: opportunities of in situ spectroscopic studies and monitoring phase behavior. Catal. Rev. Sci. Eng. **2003**, *45* (1), 1–96.
12. Bertucco, A.; Canu, P.; Devetta, L; Zwahlen, A.G. Catalytic hydrogenation in supercritical CO_2: kinetic measurements in a gradientless internal-recycle reactor. Ind. Eng. Chem. Res. **1997**, *36* (7), 2626–2633.
13. Devetta, L.; Giovanzana, A.; Canu, P.; Bertucco, A.; Minder, B.J. Kinetic experiments and modeling of a three-phase catalytic hydrogenation reaction in supercritical CO_2. Catal. Today **1999**, *48* (1–4), 337–345.
14. Kröcher, O.; Koppel, R.A.; Baiker, A. Silica hybrid gel catalysts containing ruthenium complexes: influence of reaction parameters on the catalytic behavior in the synthesis of *N,N*-dimethylformamide from carbon dioxide. J. Mol. Catal. A Chem. **1999**, *140* (2), 185–193.
15. Tacke, T.; Wieland, S.; Panster, P. Selective and complete hydrogenation of vegetable oils and free fatty acids in supercritical fluids. In *Green Chemistry Using Liquid and Supercritical Carbon Dioxide*; Desimone, J.M., William, T., Eds.; Oxford University Press Inc.: New York, 2003; 228–240.
16. Minder, B.; Mallat, T.; Pickel, K.H.; Steiner, K.; Baiker, A. Enantioselective hydrogenation of ethyl pyruvate in supercritical fluids. Catal. Lett. **1995**, *34* (1,2), 1–9.
17. van den Hark, S.; Härröd, M. Hydrogenation of oleochemicals at supercritical single-phase conditions: influence of hydrogen and substrate concentrations on the process. Appl. Catal. A Gen. **2001**, *210* (1,2), 207–215.
18. Hitzler, M.G.; Smail, F.R.; Ross, S.K.; Poliakoff, M. Friedel-Crafts alkylation in supercritical fluids: continuous, selective and clean. Org. Process Res. Dev. **1998**, *2* (3), 137–146.
19. Jessop, P.G.; Joo, F.; Tai, C.-C. Recent advances in the homogeneous hydrogenation of carbon dioxide. Coord. Chem. Rev. **2004**, *248* (21–24), 2425–2442.
20. Jessop, P.G.; Hsiao, Y.; Ikariya, T.; Noyori, R. Homogeneous catalysis in supercritical fluids: hydrogenation of supercritical carbon dioxide to formic acid, alkyl formates, and formamides. J. Am. Chem. Soc. **1996**, *118* (2), 344–55.
21. Bianchini, C.; Glendenning, L. Catalytic production of dimethylformamide from supercritical carbon dioxide. Methyl formate synthesis by hydrogenation of supercritical carbon dioxide in the presence of methanol. Selectivity for hydrogenation or hydroformylation of olefins by hydridopentacarbonylmanganese(I) in supercritical carbon dioxide. Chemtracts: Org. Chem. **1996**, *9* (6), 318–321.
22. Jessop, P.G.; Hsiao, Y.; Ikariya, T.; Noyori, R. Methyl formate synthesis by hydrogenation of supercritical carbon dioxide in the presence of methanol. J. Chem. Soc., Chem. Commun. **1995**, *6*, 707–708.
23. Chen, W.; Liu, X.; Luo, S.; Liang, G.; Wu, Y.; Yu, Z.; Jia, Z. Methanol and methyl formate synthesis from synthesis gas over cuprous chloride catalyst in liquid phase. J. Nat. Gas Chem. **2000**, *9* (2), 139–146.
24. Parmaliana, A.; Frusteri, F.; Arena, F.; Mezzapica, A.; Sokolovskii, V. Synthesis of methyl formate via two-step methane partial oxidation. Catal. Today **1998**, *46* (2,3), 117–125.
25. Mueller, L.L.; Griffin, G.L. Formaldehyde conversion to methanol and methyl formate on copper/zinc oxide catalysts. J. Catal. **1987**, *105* (2), 352–358.
26. Poliakoff, M.; Howdle, S.M.; George, M.W. Clean chemistry in supercritical fluids. In *Process Technology Proceeding*; Rudolf von Rohr, C.T.Ph., Ed.; Elsevier Science B.V.: Amsterdam, Netherlands, 1996, 67–72.
27. Lee, P.; King, J.L.; Seebald S.; Poliakoff, M. Thermal exchange reactions of (n5-C5R5)Mn (CO)2(wL) compounds (wL = weakly-bound ligand) in supercritical fluid solution and the isolation of (n5-C5R5)Mn(n2-H2) using high-pressure flow and semiflow reactors. Organometallics **1998**, *17* (4), 524–533.
28. Parkinson, G. Supercritical CO_2 speeds up hydrogenation. Chem. Eng. **2000**, *103* (2), 21.
29. Brown, J.M. Catalysts for stereospecific synthesis. Insights Spec. Inorg. Chem. **1995**, 123–43.
30. O'Brien, M.K.; Vanasse, B. Asymmetric processes in the large-scale preparation of chiral drug

candidates. Curr. Opin. Drug Discov. Dev. **2000**, *3* (6), 793–806.
31. Burk, M.J.; Feng, S.; Gross, M.F.; Tumas, W. Asymmetric catalytic hydrogenation reactions in supercritical carbon dioxide. J. Am. Chem. Soc. **1995**, *117* (31), 8277–8278.
32. Wainwright, M.S.; Ahn, T.; Trimm, D.L. Solubility of hydrogen in alcohols and esters. J. Chem. Eng. Data **1987**, *32* (1), 22–24.
33. Sun, Y.; Landau, R.N.; Wang, J.; LeBlond, C.; Blackmond, D.G. A re-examination of pressure effects on enantioselectivity in asymmetric catalytic hydrogenation. J. Am. Chem. Soc. **1996**, *118* (6), 1348–1353.
34. Xiao, J.; Nefkens, S.C.A.; Jessop, P.G.; Ikariya, T.; Noyori, R. Asymmetric hydrogenation of α,β-unsaturated carboxylic acids in supercritical carbon dioxide. Tetrahedron Lett. **1996**, *37* (16), 2813–2816.
35. Kainz, S.; Brinkmann, A.; Leitner, W.; Pfaltz, A. Iridium-catalyzed enantioselective hydrogenation of imines in supercritical carbon dioxide. J. Am. Chem. Soc. **1999**, *121* (27), 6421–6429.
36. Lange, S.; Brinkmann, A.; Trautner, P.; Woelk, K.; Bargon, J.; Leitner, W. Mechanistic aspects of dihydrogen activation and transfer during asymmetric hydrogenation in supercritical carbon dioxide. Chirality **2000**, *12* (5/6), 450–457.
37. Darr, J.A.; Poliakoff, M. New directions in inorganic and metal-organic coordination chemistry in supercritical fluids. Chem. Rev. **1999**, *99*, 495–541.
38. Palo, D.R.; Erkey, C. Homogeneous catalytic hydroformylation of 1-octene in supercritical carbon dioxide using a novel rhodium catalyst with fluorinated arylphosphine ligands. Ind. Eng. Chem. Res. **1998**, *37* (10), 4203–4206.
39. Palo, D.R.; Erkey, C. Kinetics of the homogeneous catalytic hydroformylation of 1-octene in supercritical carbon dioxide with HRh(CO)$[P(p\text{-}CF_3C_6H_4)_3]_3$. Ind. Eng. Chem. Res. **1999**, *38* (5), 2163–2165.
40. Banet Osuna, A.M.; Chen, W.; Hope, E.G.; Kemmitt, R.D.W.; Paige, D.R.; Stuart, A.M.; Xiao, J.; Xu, L. Effects of the ponytails of arylphosphines on the hydroformylation of higher olefins in supercritical CO_2. Dalton **2000**, (22), 4052–4055.
41. Deshpande, R.M.; Bhanage, B.M.; Divekar, S.S.; Kanagasabapathy, S.; Chaudhari, R.V. Kinetics of hydroformylation of ethylene in a homogeneous medium: comparison in organic and aqueous systems. Ind. Eng. Chem. Res. **1998**, *37* (6), 2391–2396.
42. Koch, D.; Leitner, W. Rhodium-catalyzed hydroformylation in supercritical carbon dioxide. J. Am. Chem. Soc. **1998**, *120* (51), 13398–13404.
43. Kainz, S.; Koch, D.; Baumann, W.; Leitner, W. Perfluoroalkyl-substituted arylphosphines as ligands for homogeneous catalysis in supercritical carbon dioxide. Angew. Chem., Int. Ed. Engl. **1997**, *36* (15), 1628–1630.
44. Franciò, G.; Wittmann, K.; Leitner, W. Highly efficient enantioselective catalysis in supercritical carbon dioxide using the perfluoroalkyl-substituted ligand (R,S)-3-H2F6-BINAPHOS. J. Organomet. Chem. **2001**, *621* (1,2), 130–142.
45. Sellin, M.F.; Webb, P.B.; Cole-Hamilton, D.J. Continuous flow homogeneous catalysis: hydroformylation of alkenes in supercritical fluid-ionic liquid biphasic mixtures. Chem. Commun. (Camb.) **2001**, (8), 781–782.
46. Jessop, P.G.; Stanley, R.R.; Brown, R.A.; Eckert, C.A.; Liotta, C.L.; Ngo, T.T.; Pollet, P. Neoteric solvents for asymmetric hydrogenation: supercritical fluids, ionic liquids, and expanded ionic liquids. Green Chem. **2003**, *5* (2), 123–128.
47. Leitner, W. Recent advances in catalyst immobilization using supercritical carbon dioxide. Pure Appl. Chem. **2004**, *76* (3), 635–644.
48. Garcia, S.; Lourenco, N.M.T.; Lousa, D.; Sequeira, A.F.; Mimoso, P.; Cabral, J.M.S.; Afonso, C.A.M.; Barreiros, S. A comparative study of biocatalysis in non-conventional solvents: ionic liquids, supercritical fluids and organic media. Green Chem. **2004**, *6* (9), 466–470.
49. Brown, R.A.; Pollet, P.; McKoon, E.; Eckert, C.A.; Liotta, C.L.; Jessop, P.G. Asymmetric hydrogenation and catalyst recycling using ionic liquid and supercritical carbon dioxide, J. Am. Chem. Soc. **2001**, *123* (6), 1254–1255.
50. Liu, F.; Abrams, M.B.; Baker, R.T.; Tumas, W. Phase-separable catalysis using room temperature ionic liquids and supercritical carbon dioxide. Chem. Commun. (Camb.) **2001**, (5), 433–434.
51. Kordikowski, A.; Schenk, A.P.; Van Nielen, R.M.; Peters, C.J. Volume expansions and vapor–liquid equilibria of binary mixtures of a variety of polar solvents and certain near-critical solvents. J. Supercrit. Fluids. **1995**, *8* (3), 205–216.
52. Rajagopalan, B.; Wei, M.; Musie, C.,T.; Subramaniam, B.; Busch, D.H. Homogenous catalytic epoxidation of organic substrates in CO_2-expanded solvents in the presence of water-soluble oxidants and catalysts. Ind. Eng. Chem. Res. **2003**, *42*, 6505–6510.
53. Kerler, B.; Robinson, R.E.; Borovik, A.S.; Subramaniam, B. Application of CO_2-expanded solvents in heterogeneous catalysis: a case study. Appl. Catal. B Environ. **2004**, *49*, 91–98.
54. Kayaki, Y.; Noguchi, Y.; Ikariya, T. Enhanced product selectivity in the Mizoroki-Heck reaction

using a supercritical carbon dioxide-liquid biphasic system. Chem. Commun. (Camb.) **2000**, *22*, 2245–2246.
55. Darr, J.A.; Poliakoff, M. New directions in inorganic and metal-organic coordination chemistry in supercritical fluids. Chem. Rev. (Washington, D.C.). **1999**, *99* (2), 495–541.
56. Warwick, B.; The synthesis and purification of copper indomethacin. Ph.D. dissertation, The University of New South Wales, Sydney, Australia, 2001.
57. Subramaniam, B.; McHugh, M.A. Reactions in supercritical fluids—a review. Ind. Eng. Chem. Proc. Des. Dev. **1986**, (25), 1–12.
58. Savage, P.E.; Gopalan, S.; Mizan, T.I.; Martino, C.J.; Brock, E.E. Reactions at supercritical conditions: applications and fundamentals. AIChE J. **1995**, *41* (7), 1723–1778.
59. Morgenstern, D.A.; LeLacheur, R.M.; Morita, D.K.; Borkowsky, S.L.; Feng, S.; Brown, G.H.; Luan, L.; Gross, M.F.; Burk, M.J.; Tumas, W. Supercritical carbon dioxide as a substitute solvent for chemical synthesis and catalysis. ACS Symp. Ser., **1996**, *626* (Green Chemistry), 132–151.
60. Subramaniam, B.; Rajewski, R.A.; Snavely, K. Pharmaceutical processing with supercritical carbon dioxide. J. Pharm. Sci. **1997**, *86* (8), 885–890.
61. Buelow, S.; Dell'Orco, P.; Morita, D.; Pesiri, D.; Birnbaum, E.; Borkowsky, S.; Brown, G.; Feng, S.; Luan, L.; Morgenstern, D.; Tumas, W. Recent advances in chemistry and chemical processing in dense phase carbon dioxide at Los Alamos. Green Chem. **1998**, 265–285.
62. Loeker, F.; Leitner, W. Supercritical carbon dioxide as an innovative reaction medium for selective oxidation. In *DGMK Tagungsber*, Proceedings of the DGMK-Conference\Selective Oxidations in Petrochemistry, 1998; Vol. 9803, 209–215.
63. Poliakoff, M.; Howdle, S.M. Supercritical chemistry : a review of chemical reactions in supercritical fluids solutions. In *Third International Symposium on Supercritical Fluids*; International Society for the Advancement of Supercritical Fluids: Nottingham, UK, 1999.
64. Perrut, M. Enzymic reactions and cell behavior in supercritical fluids. Chem. Biochem. Eng. Q. **1994**, *8* (1), 25–30.
65. Kamat, S.V.; Beckman, E.J.; Russell, A.J. Enzyme activity in supercritical fluids. Crit. Rev. Biotechnol. **1995**, *15* (1), 41–71.
66. Nakamura, K. Enzymic synthesis in supercritical fluids. In *Supercritical Fluid Technology in Oil and Lipid Chemistry*; AOCS Press: Champaign, 1996; 306–320.
67. Cerina, E.; Palocci, C. Lipases in supercritical fluids. Methods Enzymol. **1997**, *286* (Lipases, Part B) 495–508.
68. Mesiano, A.J.; Beckman, E.J.; Russell, A.J. Supercritical biocatalysis. Chem. Rev. (Washington, D.C.). **1999**, *99* (2), 623–633.
69. Shaffer, K.A.; DeSimone, J.M. Chain polymerization in inert near- and supercritical fluids. Trends Polym. Sci. (Camb.). **1995**, *3* (5), 146–153.
70. Canelas, D.A.; Desimone, J.M. Polymerizations in liquid and supercritical carbon dioxide. Adv. Polym. Sci. **1997**, *133* (Metal Complex Catalysts, Supercritical Fluid Polymerization, Supramolecular Architecture), 103–140.
71. Quadir, M.A.; DeSimone, J.M. Chain growth polymerizations in liquid and supercritical carbon dioxide. ACS Symp. Ser. **1998**, *713* (Solvent-Free Polymerization and Processes), 156–180.
72. Palo, D.R.; Erkey, C. Kinetics of the homogeneous catalytic hydroformylation of 1-octene in supercritical carbon dioxide with HRh(CO)$[P(p\text{-}CF_3C_6H_4)_3]_3$. Ind. Eng. Chem. Res. **1997**, *36* (7), 2626–2633.

Hydrophilic Polymers for Biomedical Applications

Frank Davis
Séamus P. J. Higson
Cranfield University, Silsoe, U.K.

INTRODUCTION

The production of polymeric materials is one of the world's major industries. Polymers are utilized in many applications because of their processability, ease of manufacture, and diverse range of properties. Many of the commonest polymers such as polyethylene, polystyrene, and polytetrafluoroethylene (PTFE) are highly hydrophobic materials rendering them unsuitable for many biomedical applications. For applications that require contact with body fluids such as blood or urine, it is necessary for the materials to be hydrophilic and to be capable of maintaining intimate contact with the fluid in question for prolonged periods of time without significant loss of functional performance.

This entry begins by describing the nature of hydrophilic polymers and structures, together with properties of some of the most commonly encountered materials. Details are then given of some applications that require the use of polymers and how they can be applied. The entry finishes with a brief synopsis of the work in this field and possible future research and applications.

STRUCTURE AND PROPERTIES OF HYDROPHILIC POLYMERS

Polymeric materials made from simple hydrocarbon monomers such as polyethylene are extremely hydrophobic in nature. There is very little in the way of interaction between the polymer backbone and water, which means that these polymers will not adsorb water and their surfaces are not wettable. For a polymer to be hydrophilic, functional groups capable of undergoing interactions with water such as hydrogen bonding are usually required. Such groups include, but are not limited to, alcohols, ethers, esters, carboxylic acids, amines, and amides. These can interact with water via hydrogen bonding or dipolar interactions; such polymers can vary from just being wettable at their surfaces to being permeable to water, and in some cases are capable of adsorbing many times their own weight in water. Some of the most widely used polymers are discussed later. The properties and use of these materials in biomedical applications can only be summarized within the limits of this entry; however, several other works have been written on this subject.[1–3]

Hydrogels

Many of the polymers mentioned within this work exist as hydrogels. In essence, a hydrogel is a polymer that would normally be soluble in water but has been cross-linked to form a polymer network. The cross-linking process renders the polymers insoluble but does not remove their affinity for water. Therefore, the network can adsorb water with consequent swelling of the polymer. The nature of the polymer and the degree of cross-linking affect the swelling behavior; highly hydrophilic polymers containing few cross-links will tend to adsorb large amounts of water with a high degree of swelling. Less hydrophilic monomers, incorporation of hydrophobic comonomers, or a high degree of cross-linking, all act to reduce water adsorption, usually leading to a more rigid and firmer gel. Hydrogels often show high biocompatibility, usually because they have a high water content, within either a bioinert or a biodegradable polymer network.

Synthetic Hydrophilic Polymers

The structures of a number of synthetic hydrophilic polymers as a representative range of those commonly used are shown in Fig. 1. Many common hydrophilic polymers are based on methacrylate or acrylate backbones. One of the most common is polyacrylic acid, which forms the basis of many hydrogel materials (Fig. 1A). Although simple polyacrylic acid is water soluble, it can be easily cross-linked to form an insoluble network polymer that still, however, retains high affinity for water and is capable of adsorbing large amounts of water to form hydrogels; these are discussed later. Polymethacrylic acid-based polymers are similar in nature (Fig. 1B). One of the most commonly used polymers for contact lenses, polyhydroxyethyl methacrylate, is based on a similar backbone (Fig. 1C).

Polyvinyl alcohol is another widely used material, often as a surface coating because of its high biocompatibility (Fig. 1D). Usually, it can be obtained by the hydrolysis of polyvinyl acetate, with varying amounts

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120023855

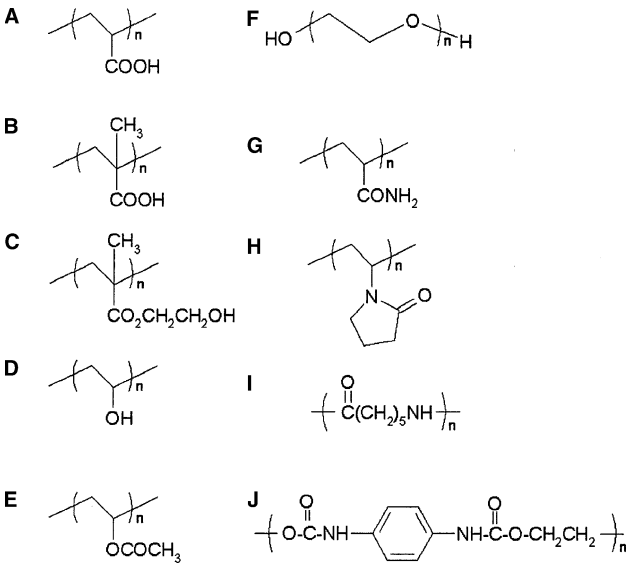


Fig. 1 Structures of some common synthetic hydrophilic polymers: (A) polyacrylic acid, (B) polymethacrylic acid, (C) polyhydroxyethyl methacrylate, (D) polyvinyl alcohol, (E) polyvinyl acetate, (F) PEG/PEO, (G) polyacrylamide, (H) polyvinylpyrrolidinone, (I) Nylon 6, and (J) a simple polyurethane.

of hydrolysis (Fig. 1E). Direct synthesis from the monomer is not possible because vinyl alcohol does not exist but tautomerizes to acetaldehyde. The solubility and physical properties of this polymer are highly dependent on its molecular weight and degree of hydrolysis. Cross-linking of the polymer leads to a variety of hydrogels.

Polyethylene glycol (PEG) and the very similar polyethylene oxide (PEO) are used as biocompatible coating agents and hydrogel forming materials, often as block or graft copolymers with other materials (Fig. 1F). They are often bound to polyurethanes to form hydrophilic foams such as Biopol® (Metabolix Inc.).

Polyacrylamides are also suitable for a variety of biomedical uses; the structure of polyacrylamide is shown in Fig. 1G, although the use of acrylamides in copolymers is much more common. Polyvinylpyrrolidinone has also found use as a biocompatible coating material (Fig. 1H). Polyacrylonitrile, though not suitable in itself, can be hydrolyzed to form some hydrophilic polymers such as the Hypan® (Hymedix Inc.) series of hydrogels.

Condensation polymers have been used although they are often not sufficiently hydrophilic enough to be considered within this entry. Typical polymers such as Nylon 6 and a typical polyurethane are shown (Figs. 1I and 1J). Some of these have been incorporated into composites with more hydrophilic materials; Biopol, for example, is formed from a copolymer of a polyurethane and PEG.

Natural Hydrophilic Polymers

Besides the synthetic materials, there are a variety of hydrophilic polymers obtained from natural sources. Some examples include collagen, alginate, and carrageen, all of which are obtained from natural sources.[1] Collagen is widely used in implants and many other possible uses and applications have been studied for this material.

A series of materials have been made by derivatizing cellulose (Fig. 2a). This naturally occurring polysaccharide contains reactive hydroxyl groups, which can be easily substituted, normally to form ethers. Typical materials include sodium carboxymethylcellulose, hydroxyethyl, and hydroxypropyl cellulose (Figs. 2b–2d). The properties of these polymers depend greatly on the molecular weight and degree of substitution; they have found uses both as gels and as coatings.

Hydrophilic Polymers: Use as Contact Lens Materials

The use of contact lenses, initially made from glass, to correct vision has been known since the 19th century. The first polymeric contact lens was a ''hard'' contact lens made in 1936 from polymethylmethacrylate (PMMA).[4] However, these lenses had to be taken out at night to prevent eye irritation and much research has been directed to making those that could be worn for much longer periods of time. This led to the development of the so-called ''soft'' lenses.

To be suitable for long-term wear, a contact lens material has to satisfy several criteria. It must be hydrophilic enough to maintain a stable, continuous tear film on its surface, resist fouling by tear components, not irritate the eye, and be comfortable to wear. It should also be realized that the metabolism of the cornea is highly dependent on dissolution and adsorption of atmospheric oxygen and therefore any lens material must have sufficient oxygen permeability to maintain this, else corneal anoxia will set in.

Hydrogels were the first polymers to be used for this application because of their favorable physical properties and compatibility with the ocular environment.

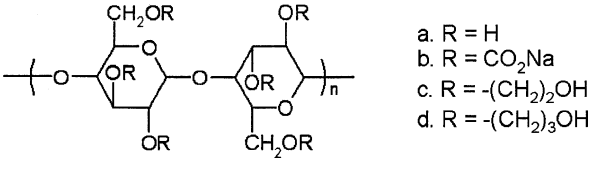


Fig. 2 Structures of some common cellulose-based hydrophilic polymers: (a) cellulose, (b) sodium carboxymethyl cellulose, (c) hydroxyethyl cellulose, and (d) hydroxypropyl cellulose.

The first commercial product, Polymacon® (Bausch & Lomb), was based on polyhydroxyethyl methacrylate (Fig. 1C).[5] The materials did suffer somewhat as they did not offer sufficient oxygen mass transport across the lens for prolonged use; hence, much further work has been devoted to developing materials with higher oxygen permeability. Organosilicon-based polymers such as polydimethylsiloxane have high oxygen permeability; however, most of these are very hydrophobic (Fig. 3A). They have, therefore, been incorporated into lenses along with hydrophilic materials in an attempt to make a material with the required properties. Early studies utilized copolymers containing dimethylsiloxane groups along with hydrophilic monomers such as hydroxyethyl methacrylate or *n*-vinyl pyrrolidinone.[6] Fluoroethers also have high oxygen permeability and a copolymer of a perfluoroether with methyl methacrylate was briefly marketed under the name Fluorofocon A® (3M); however, the relatively high cost of the fluorocarbon monomers eventually led to their withdrawal from the market (Fig. 3B).[7]

Most soft contact lenses commercially available are based on polyhydroxyethyl methacrylate; however, there are a few materials available, such as Lotrafilcon A® (Ciba Vision) and Balafilcon A® (Bausch & Lomb), based on siloxane or fluorosiloxane copolymers, that offer good oxygen permeability and so allow up to 30 days' continuous wear for contact lenses.[4] A survey of the patent literature shows that much interest is still being shown in the field of siloxane-based hydrogels in general and also in improving manufacturing processes for contact lenses formed from such hydrogels.

Besides external contact lenses, there are surgically implanted intraocular contact lenses that are usually used to replace the eye's natural lens after cataract surgery; these lenses are normally based on silicone or PMMA and lie outside the scope of this entry.

Implantable Membranes

Postsurgical adhesions are a common problem following major abdominal, gynaecological, and other forms of surgery. Adhesions are scars that form abnormal connections between tissue surfaces. Postsurgical adhesion formation is a natural consequence of surgery, resulting when tissue repairs itself following incision, cauterization, suturing, or other forms of trauma. This can lead to complications such as bowel obstruction and infertility. Implantable materials are in this case sometimes placed directly between tissue surfaces, for instance, between organs and the abdominal wall, to prevent adhesion formation.

A suitable membrane must be sterile, noninflammatory, and nontoxic, and because these cannot be removed from the body without further surgery, they are often also designed so as to be bioresorbable. A variety of polymers have been studied for this purpose. Typical materials available commercially include Seprafilm® (Genzyme Corp), which is based on a sodium hyaluronate/carboxymethyl cellulose composite, and Oxiplex® (Fziomed Inc.), which is based on a PEO/carboxymethyl cellulose composite.[8] Products designed for similar use but marketed in gel form include Oxiplex-SP®, Spraygel® (Confluent Surgical Inc.), and Resolve® (Life Medical Sciences Inc.). These can be applied directly to the site, rapidly form a solid hydrogel film to prevent adhesions, and are designed to be reabsorbed by the body after approximately 1 week.[9]

Sutures and Implants

Sutures are used to close wounds or incisions made during operations. If the material used for the suturing is bioresorbable, then this eliminates the need for removal of any stitches. Polymers such as polyglycolide [first marketed as Dexon® (Bayer Corp) in the 1960s] and polylactide are commonly used for this purpose, although a wider range of polymers are now available and have been recently reviewed elsewhere (Figs. 4A and 4B).[10]

Besides sutures, other medical devices such as orthopedic fixation devices are of interest because the use of biodegradable polymers means that not only does the device not require removal (a process that can cause refracture of the bone), but because if it degrades it will also slowly transfer stress over time to the damaged area, allowing healing of the tissues. At the time of writing, bioresorbable polymers that offer the necessary strength for use as bone plates for long bones are yet to be developed, although they have

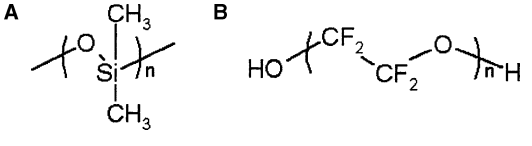


Fig. 3 Structures of some polymers used in contact lens formulations: (A) a typical polysiloxane and (B) a typical polyperfluoroether.

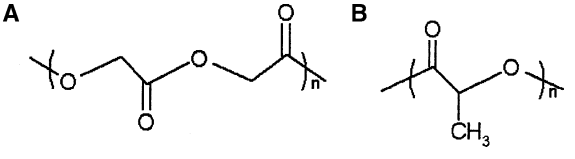


Fig. 4 Structures of some bioresorbable polymers: (A) polyglycolide acid and (B) polylactide.

found use in the manufacture of rods and pins for fracture fixation.[10]

Natural polymers have also been studied for cosmetic implants such as collagen. Bovine collagen implants have become increasingly popular for use in dermal augmentation of the skin since their introduction in the 1980s. They can be used either to remove unwanted wrinkles or to augment features such as collagen lip injections, which are gaining in popularity.

Burn and Wound Dressings

Many burn dressings are based on keeping a burn site moist while protecting it from the environment as well as helping in pain relief. Because hydrogels contain such a high water content, they have been investigated for use as burn and other wound dressings.[11] They have many desirable properties; for example, imposing the soft, moist texture can alleviate some pain; they also adhere to the wound but not so tightly that they cannot be easily detached from the wound for replacement; and finally, they are also transparent, allowing visual inspection of the healing process. Also, hydrogels prevent dehydration of the wound, have good oxygen permeability, and yet provide a barrier against bacteria. Finally, they can be easily manufactured with coadsorbed drugs for aiding in pain control, promotion of healing, and/or administration of antibiotics.

Materials are based on cross-linked hydrophilic polymers and are manufactured under a variety of trade names such as Spyroflex® (AGS Labs Inc.), Burnfree® (Burnfree Products), and Burnshield® (Levtrade International Ltd). Usually, they contain the hydrogel attached to an impermeable backing sheet and are currently available in up to blanket sizes. Typical polymers used include polyvinyl alcohol, PEG, and polyvinyl pyrrolidinone together with combinations of these materials. Irradiation is often used both to cross-link and to sterilize the hydrogel after it has been formed into the dressing. Collagen hydrogels are also being researched for their ability to allow cell migration and inhibit wound contraction because of their high tensile strength and low extensibility.[12]

Coatings

One major problem with inserting any type of biomedical device into the body that comes into contact with blood or other body fluids is that of biofouling. Both proteins and cells can adhere strongly to many foreign surfaces. In the case of devices such as contact lenses, for example, lipids and tear proteins can adsorb onto the surface of the lens. This can cause clouding of the lens, rendering it unsuitable for use, or lead to irritation. The irreversible adsorption of proteins and cells on synthetic surfaces has adverse effects on a variety of devices. Adverse biological reactions include fibrous encapsulation and blocking of small artificial blood vessels. Other devices, such as heart valves, can give rise to the formation of thromboses. This can cause either arterial blocking or the clot can detach from the surface, leading to the possibility of a stroke or heart attack. The application of hydrophilic polymer coatings has been widely studied to help render many surfaces biocompatible to minimize such risks.

Although a wide variety of polymeric materials have been utilized in the attempted prevention of biofouling as described in the extensive review by Kingschott and Grieser, by far the most popular materials have been ones based on PEG/PEO.[13]

The theoretical considerations of the causes of biofouling and why a PEG/PEO surface should resist the process so well is a topic for a complete review article in itself; however, it is widely reported that the lowest adsorption of proteins occurs for these materials.[13] One factor is thought to be the fact that the PEG/PEO chains are usually highly solvated, meaning that incoming protein molecules experience a surface that is largely composed of water. This surface mimics the typical conditions found within biological systems.

Several different approaches have been used to attach PEG/PEO coatings to a variety of surfaces. Among the simplest is the physical adsorption of Pluronic surfactants, which consist of PEO–polypropylene oxide (PPO)–PEO block terpolymers (Fig. 4A).[14] The more hydrophobic PPO section of the chain is absorbed onto the substrate being treated with the more hydrophilic PEO blocks stretching out into the aqueous phase. These materials display a variety of biocompatibilities depending on the makeup of the surfactant. Interestingly, increasing the length of the PPO section appeared to have a larger repulsive effect than doing so of the PEO section. This indicates that perhaps only short PEO chains are necessary for effective protein repulsion and the beneficial effect of increasing the PPO section is because of better anchoring of the surfactant and prevention of displacement by protein molecules.[14]

Other methods of attachment have been studied. Alkyl thiols are known to form well-packed chemically bound monolayers on noble metal surfaces and so this method has also been used to attach PEG chains to surfaces. A variety of thiol-substituted PEG oligomers were synthesized and spontaneously assembled onto gold and silver surfaces (Fig. 5B).[15] Quite short PEG chains were found to be sufficient to induce biocompatibility. It was found that the packing arrangement of the PEG derivatives was important in determining the surface properties with the resultant biocompatibility not simply being a direct function of packing density. Thiols pack more tightly on silver

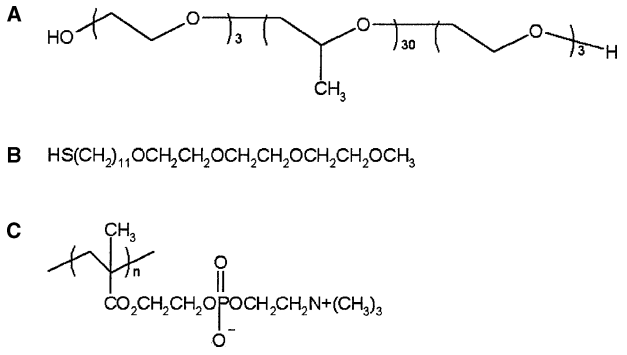


Fig. 5 Structures of some common synthetic polymers used in coatings: (A) a typical Pluronic surfactant, (B) a thiol-substituted PEG derivative, (C) a polymer containing a phospholipid head group analog.

than on gold; yet some treated silver surfaces showed higher levels of fibrinogen adsorption than the corresponding treated gold surface. This was attributed to the PEG chains being in different conformations and showed that simply increasing the surface packing density does not necessarily increase repulsion.

Plasma deposition has also been utilized to deposit PEO-like materials from volatile precursors onto a variety of subjects.[13] This technique involves generating a reactive plasma containing PEO-like monomers, which polymerize and deposit, often with chemical grafting, onto any surface within the plasma.[16] The availability of large-scale vacuum apparatus makes this technique feasible on an industrial scale. The materials deposited by this technique were often shown to contain only short PEO segments; yet greatly reduced protein deposition was observed and the small amounts (ng/cm) that did deposit were easily eluted.[16]

Other materials have also been studied for their ability to reduce protein adsorption onto surfaces. Because many cell membranes are based on phospholipids, polymers containing phospholipid-type head groups have been utilized for this purpose. Poly(2-methacroylethyl phosphoryl choline) could be plasma deposited onto silicone rubber and the adhesion of albumin reduced by factors of up to 80 (Fig. 5C).[17]

Polysaccharides have also been studied for this purpose, although generally they are not as effective as PEG/PEO systems.[13] Polymers based on dextran substituted with thiol groups could be adsorbed on gold surfaces and have been shown to significantly repel protein deposition, with the degree of repulsion being dependent on both the polymer molecular weight and the amount of thiol substitution.[18]

An additional benefit of coating samples with hydrophilic polymers is their ability to act as a lubricant. Lubrication of a device surface improves device insertion and manipulation; however, traditional treatments that include coatings with low-friction materials such as PTFE or silicone fluid, while improving slip in the body, are difficult for the physician to handle. Incorporation of hydrophilic polymers leads to the adsorption of water molecules in the presence of water or body fluids. This creates a sheath of water at the surface of the device. This watery interface causes the decrease in wet friction, achieving lower friction when wet, while still being easy to handle when dry.

Biosensors

Biosensors are devices that use the unique recognition qualities of biological molecules to selectively detect the presence of desired analytes within complex mixtures such as blood and serum. They generically offer simplified reagent less analyses for a range of biomedical and industrial applications, and for this reason the field has continued to develop into an ever-expanding and multidisciplinary one during the last couple of decades. The world market for biosensors is approximately $5 billion, and at the time of writing, approximately 85% of the world commercial market for biosensors is for blood glucose monitoring.[19] Biosensors can be used as in vivo devices, as in the simple glucose test where a drop of blood is extracted and tested; however, much research is now taking place on implantable, in vivo continuous reading biosensors.[19]

Sensors containing biomolecules such as glucose oxidase suffer from problems such as stability because the enzyme can denature upon storage or use. Also, the testing of complex mixtures infers the danger of unspecific adsorption of interferents or simple blocking of the sensor by protein or cell deposition. Hydrophilic polymers have been utilized in an attempt to solve some of these problems. The use of hydrophilic polymer membranes within biosensors has been recently reviewed.[20] The highly hydrated environment of a swollen hydrogel is a simple analog of conditions found within the biological environment. In this context, the trapping of active ingredients in a hydrogel has been utilized to improve the stability of the biomolecule.[20] Also, a thin hydrophilic polymer coating, such as PEO, can minimize biofouling while also helping to exclude common interferents such as ascorbate ions.[20]

Drug Delivery

The pharmaceutical industry has shown great interest in the development of controlled release systems based on hydrophilic polymers. Many drugs are most efficient when released into the body at a constant rate, rather than in individual doses. For example, a drug could be incorporated into a hydrogel either as it is synthesized or postsynthesis. This can then be ingested

by, implanted within, or be simply placed in contact with the patient, e.g., upon the skin. Once in vivo, the drug is released into the environment by diffusion or by the polymer being eroded or dissolved in some way, for example, by digestion. Ideally, this process will occur at a constant rate, thereby continually releasing controlled amounts of the drug. If the polymer is surgically implanted close to an affected site, it means the drug is delivered where it is most needed.

For a polymer to be suitable for this purpose, it must display several properties. First, it must be biocompatible and must biodegrade within a reasonable period of time. Both the polymer itself and its degradation products must be nontoxic and must create neither an allergic nor an inflammatory response. The nature of the drug itself can also affect the method of release; low molecular weight drugs are capable of diffusing out of the polymeric matrix, and if water soluble will be rapidly released. Larger molecules such as proteins tend to be trapped within the matrix until the polymer itself degrades and releases them. The rates of release are dependent on several variables, including the quantity/dosing of drugs within the composite, the rate of degradation of the polymer, the amount of water present (if it is a hydrogel type polymer), and the presence and degree of cross-linking. The application of a wide variety of hydrophilic biodegradable polymers to delivery of proteins has been extensively reviewed.[21] Typical polymers include hydrogels based on polyvinyl alcohol, polyvinylpyrrolidinone, or cellulose, and natural polymers such as alginase or collagen. Besides enabling the controlled release of the active material, the polymer can also be exploited as a stabilizing medium for what are often unstable biological agents.

The polymer-based delivery device can be used to apply the active agent in many ways, including ingestion, suppositories, skin patches, ocular and subcutaneous methods. The treatment of cancer is a field in which these methods are being widely researched, for example, ara-C, a treatment for acute leukemia has the least side effects when introduced subcutaneously. As an alternative to continuous infusion, the drug can be incorporated into cross-linked polyhydroxyethyl acrylate disks, which displayed steady controllable drug release.[22] A polycaprolactone/PEG composite was similarly used as a matrix for the anticonvulsant drug clonazepam and displayed long-term (>45 days) constant release properties.[23] Nitroglycerin is a problematic drug owing to its loss of tablet activity, often by volatilization of the active component.[24] This can be avoided by incorporation of the nitroglycerin into an acrylic-based hydrogel, which is then incorporated into a transdermal patch, as exemplified by products such as Deponit® (Schwarz Pharma), Minitran® (3M Pharma), and Nitrodisc® (G.D. Searle Company).

These products are usually based on simple polymeric systems; however, more complex architectures are known. Insulin can be contained within a ''smart'' porous membrane containing glucose oxidase encapsulated within a polymethacrylic acid/PEG copolymer. High levels of glucose interact with the glucose oxidase, causing a pH drop and shrinkage of the membrane, which leads to the release of insulin.[25]

Apart from drug delivery, recent work has focused on the use of hydrophilic polymers in gene therapy and has shown that encapsulation of the DNA inside a hydrophilic polymer can increase the transfection efficiency.[3]

Artificial Organs

In the human body, the kidneys remove toxic wastes such as urea from the blood for excretion via the bladder. When they fail to perform this function, it leads to the condition known as uremia, which if not addressed will ultimately lead to death. To prevent this, hemodialysis must be performed, usually by a kidney dialysis machine, which can be thought of as an artificial organ external to the body. A semipermeable membrane based on the hydrophilic polymer cellulose is an integral part of this process. Hemodialysis uses a cellulose membrane tube that is immersed in a large volume of fluid. Blood is first removed from the patient, pumped through this tubing, and then back into the patient intravenously. Cellulose membranes contain pores that allow passage of most solutes in the blood while retaining the proteins and cells.

As blood passes through the membrane, free exchange of the solutes occurs between blood and the external isotonic solution—a salt solution with ionic concentrations near or slightly lower than the desired concentrations in the blood. This means that the two solutions are in dynamic equilibrium, and so the concentration of these species within the blood does not change. Compounds such as urea that are present in the blood in excess pass, however, into the external solution and are thereby removed from the bloodstream.

Hydrophilic polymers are used as major components of artificial internal organs. The physical properties of these materials and, in particular, their high water content, soft texture, and consistency, give them a strong resemblance to actual living soft tissues. A possible application, for example, in the case of organ failure, would be to incorporate living cells from a donor into these hydrogels and then implant them in the patient. The high biocompatibility of the hydrogel would prevent rejection and help to keep the implanted cells viable. For example, hepatocytes could be encased inside highly permeable hydrogel (83% water) tubes

with a survival rate in vivo of 85% after 45 days.[26] The tubes maintained viability, prevented rejection, and allowed the passage of albumin from the encapsulated cells. A similar principle using membranes of calcium alginate hydrogel was used to encapsulate hepatocytes. Experiments showed that these units had the ability to replace liver function.[27]

Polymer-based heart valves are widely used as replacements for diseased or damaged human heart valves. Most mechanical heart valves are made from metals, silicone, or polyesters, although some work has gone into incorporating biocompatible coatings such as PEO into these systems.

Tissue Engineering

Biocompatible hydrophilic polymers, both natural and synthetic, can be used to promote tissue repair and regeneration. For example, gelatin scaffolds can be constructed by glutaraldehyde cross-linking, followed by freeze drying of the solution to give a porous matrix. This material can be specifically shaped to provide a support for tissue growth and organization over long periods.[28]

Coronary artery bypass grafting is commonly used to relieve conditions such as angina. Usually, the graft is taken from the patients themselves—a section of vein from the patient's leg being the most common. However, should there be a shortage of supply of this native material, the use of an artificial analog becomes necessary. Synthetic materials have been developed based on Dacron® (Dupont) or PTFE, although these are satisfactory only when used in large arteries; in smaller vessels they tend to cause thrombosis.[29] Current research includes the use of hydrophilic polymers as scaffold materials for the growth of cells with anticoagulatory properties, as recently reviewed elsewhere.[29,30] Both natural polymers such as collagen and synthetics such as biodegradable polyesters have been used.[29,30] The advantages of the polyesters are that not only are they stronger than many of the natural materials, but that over a period of time they are also reabsorbed by the body to be replaced with endogenous endothelial cells.[30]

CONCLUSIONS

The primary focus of this entry has been to provide a basic understanding of the field of hydrophilic polymers and their uses. These materials have a range of applications both in the bulk and as thin films. For example, hydrogel membranes are used as inhibitors of postsurgical adhesion formation and also as coatings. Dispensing these polymers onto the surface of medical devices can confer desirable surface properties that the substrate material lacks. Hydrogel coatings especially can result in vast improvements in biocompatibility or reduction in surface friction. In the bulk, their predictable and controllable diffusivity can lead to their use as drug delivery agents. Both natural and synthetic polymers have been utilized for these purposes.

The present polymer technology has resulted in improvements in several different medical fields, including drug therapy and surgical procedure. It is always difficult to predict future developments, but it appears likely that hydrophilic polymers will help further the development of more viable localized drug delivery, implantable sensors, and artificial organ substitutes than currently exist. Much research is being carried out worldwide in these fields, making this both an interesting and a fast moving field.

REFERENCES

1. LaPorte, R.J. *Hydrophilic Polymer Coatings for Biomedical Devices*; Technomic: Lancaster, 1997.
2. Ottenbrite, R.M. *Frontiers in Biomedical Polymer Applications*; Technomic: Lancaster, 1998; Vol. 1.
3. Ottenbrite, R.M. *Frontiers in Biomedical Polymer Applications*; Technomic: Lancaster, 1999; Vol. 2.
4. Nicolson, P.C.; Vogt, J. Soft contact lens polymers: an evolution. Biomaterials **2001**, *22*, 3273–3283.
5. Lui, Y.; Wilson, A.; Zantos, S. Contact lens. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th Ed.; Wiley: New York, 1993; Vol. 7, 191–228.
6. Mueller, K.; Kleiner, E Polysiloxane Hydrogels. US Patent 4,136,250, Jan 3, 1979.
7. Rice, D.; Ihlenfeld, J Contact lens containing a fluorinated telechelic polyether. US Patent 4,440,918, Apr 3, 1984.
8. Becker, J.M.; Dayton, M.T.; Fazio, V.M.; Beck, D.E.; Stryker, S.J.; Wexner, S.D.; Wolff, B.G.; Roberts, P.L.; Smith, L.E.; Sweeney, S.A.; Moore, M. Prevention of postoperative abdominal adhesions by a sodium hyaluronate-based bioresorbable membrane: a prospective, randomized, double-blind multicenter study. J. Am. Coll. Surg. **1996**, *183*, 297–306.
9. Mettler, L.; Audebert, A.; Lehmann-Willenbrock, E.; Jacobs, V.R.; Schive, K. Prospective clinical trial of spraygel as a barrier to adhesion formation: an interim analysis. J. Am. Assoc. Gynecol. Laparosc. **2003**, *10* (3), 339–344.
10. Middleton, J.C.; Tipton, A.J. Synthetic biodegradable polymers as medical devices. Med. Plast. Biomater. Mag. **1998**, *30*, 30–40.

11. Rosiak, J US Patent, 487,1,490, 1989.
12. Doillon, C.J.; Whyne, C.F.; Brandwein, S.; Silver, F.H. Collagen-based wound dressings—control of the pore structure and morphology. J. Biomed. Mater. Res. **1986**, *20*, 1219–1228.
13. Kingschott, P.; Grieser, H.J. Surfaces that resist bioadhesion. Curr. Opin. Solid State Mater. Sci. **1999**, *4*, 403–412.
14. Green, R.J.; Davies, M.C.; Roberts, C.J.; Tendler, S.J.B. A surface plasmon resonance study of albumin adsorption to PEO-PPO triblock copolymers. J. Biomed. Mater. Res. **1998**, *42*, 165–171.
15. Harder, P.; Grunze, M.; Dahint, R.; Whiteside, G.M.; Laibinis, P.E. Molecular conformation in oligo(ethylene glycol)-terminated self assembled monolayers on gold or silver surfaces determines their ability to resist protein adsorption. J. Phys. Chem. B **1998**, *102*, 426–436.
16. Shen, M.C.; Martinson, L.; Wagner, M.S.; Castner, D.G.; Ratner, B.D.; Horbett, T.A. PEO-like plasma polymerized tetraglyme surface interactions with leukocytes and proteins: in vitro and in vivo studies. J. Biomater. Sci. Polym. Ed. **2002**, *13* (4), 367–390.
17. Hsuie, G.H.; Lee, S.D.; Chang, P.C.; Kao, C.Y. Surface characterization and biological properties study of silicone rubber material grafted with phospholipid as biomaterial via plasma induced graft copolymerization. J. Biomed. Mater. Res. **1998**, *42*, 134–147.
18. Frazier, R.A.; Mattjis, G.; Davis, M.C.; Roberts, C.J.; Schacht, E.; Tendler, S.J.B. Characterization of protein-resistant dextran monolayers. Biomaterials **2000**, *21*, 957–966.
19. Newman, J.D.; Tigwell, L.J.; Turner, A.P.F.; Warner, P.J. *Biosensors: A Clearer View, Biosensors 2004—The 8th World Congress on Biosensors*; Elsevier: New York, 2004.
20. Nagels, L.J.; Staes, E. Polymer (bio)materials design for amperometric detection in LC and FIA. Trends Anal. Chem. **2001**, *20* (4), 178–185.
21. Gombotz, W.R.; Pettit, D. Biodegradable polymers for protein and peptide drug-delivery. Bioconjugate Chem. **1995**, *6*, 332–351.
22. Teijon, J.M.; Trigo, R.M.; Garcia, O.; Blanco, M.D. Cytarabine trapping in poly (2-hydroxyethyl methacrylate) hydrogels: drug delivery studies. Biomaterials **1997**, *18*, 383–388.
23. Cho, C.S.; Han, S.Y.; Ha, J.H.; Kim, S.H.; Lim, D.Y. Clonazepam release from bioerodible hydrogels based on semi-interpenetrating polymer networks composed of poly(epsilon-caprolactone) and poly(ethylene glycol) macromer. Int. J. Pharm. **1999**, *181*, 235–242.
24. Markovich, R.J.; Taylor, A.K.; Rosen, J. Drug migration from the adhesive matrix to the polymer film laminate facestock in a transdermal nitroglycerin system. J. Pharm. Biomed. Anal. **1997**, *16* (4), 651–660.
25. Gander, B.; Meinel, L.; Walter, E.; Merkle, H.P. Polymers as a platform for drug delivery: reviewing our current portfolio on poly(lactide-co-glycolide) (PLGA) microspheres. Chimia **2001**, *55*, 212–217.
26. Hongiger, H.; Balladur, P.; Mariani, P.; Calmus, Y.; Vaubourdolle, M.; Delelo, R.; Capeau, J.; Nordlinger, B. Permeability and biocompatibility of a new hydrogel used for encapsulation of hepatocytes. Biomaterials **1995**, *16*, 753–759.
27. Yanagi, K.; Ookawa, K.; Mizuno, S.; Ohshima, N. Performance of a new hybrid artificial liver support system using hepatocytes entrapped within a hydrogel. ASAIO Trans. **1989**, *35* (3), 570–572.
28. Kang, H.W.; Tabata, Y.; Ikada, Y. Fabrication of porous gelatin scaffolds for tissue engineering. Biomaterials **1999**, *20*, 1339–1344.
29. Thomas, C.T.; Campbell, G.R.; Campbell, J.H. Advances in vascular tissue engineering. Cardiovasc. Pathol. **2003**, *12*, 271–276.
30. Berglund, J.D.; Galis, Z.S. Designer blood vessels and therapeutic revascularisation. Brit. J. Pharmacol. **2003**, *140*, 627–636.

Hydrotreating Catalysts and Processes: Current Status and Path Forward

Arunabha Kundu
Nishith Dwivedi
Azad Singh
K. D. P. Nigam
Department of Chemical Engineering, Indian Institute of Technology, New Delhi, India

INTRODUCTION

The reaction of any petroleum fraction with hydrogen in the presence of a catalyst is commonly known as hydroprocessing. Depending on the purpose it serves, a hydroprocessing process can be broadly classified as hydrotreating, hydrorefining, and hydrocracking. Hydrotreating (HDT) processes remove the undesirable impurities such as sulfur, nitrogen, unsaturated molecules, and metals. It is a refining process for treating petroleum fractions from atmospheric or vacuum distillation units (e.g., naphthas, middle distillates, reformer feeds, residual fuel oil, and heavy gas oil) and other petroleum fractions (e.g., catalytic cracked naphtha, coker naphtha, gas oil, etc.) in the presence of catalysts and substantial quantities of hydrogen. Hydrotreating includes desulfurization, removal of substances (e.g., nitrogen compounds) that deactivate catalysts, conversion of olefins to paraffins to reduce gum formation in gasoline, and other processes to upgrade the quality of the fractions. Hydrorefining is a refining process similar to HDT that uses higher temperatures and pressures than HDT. The purpose of hydrorefining is to treat heavier molecular weight petroleum fractions. The fragmentation of heavy molecules into desirable ones is carried out during hydrocracking. All these processes have been widely used by petroleum refiners in the past and will also find continuous growth. The increasing importance of hydroprocessing has spurred research interest to develop more efficient catalysts and processes for commercial applications.

Hydrotreating reactions are generally carried out at high pressure (100–3000 psig) and temperature (270–350 °C) in trickle-bed reactors (TBRs) in which the liquid hydrocarbons and gaseous hydrogen are passed in a cocurrent downward flow through a fixed bed of solid catalysts. The following reactions take place during HDT:

- Hydrodesulfurization (HDS) (i.e., the conversion of organo-sulfur compounds to H_2S and similar molecular weight organic compounds).
- Hydrodenitrogenation (HDN) (i.e., the conversion of organo-nitrogen compounds to NH_3 and similar molecular weight organic compounds).
- Hydrodemetallization (HDM) (i.e., the precipitation of metal on catalyst in sulfide form).
- Hydrodeoxygenation (HDO) (i.e., the removal of –OH from molecule).
- Hydrogenation (HYD) (i.e., the hydrogenation of olefins to aliphatic compounds).

Environmental concerns are forcing drastic changes in motor fuel specifications. It is clear today that motor fuel quality will continue to be modified to improve both combustion quality and postcombustion gas clean-up performance. Regulations are becoming increasingly more stringent to minimize air pollution. For instance, 500 ppm S has been a common world specification; the Swedish Class 1 limit is 10 ppm S; in the United States, most refiners are designing for 5–8 ppm of the hydrotreaters; present EC regulations for the sulfur content of diesel fuels is 350 ppm and is expected to be below 50 ppm in 2005 and 10 ppm around 2008. In India, the sulfur specification will be 500 ppm by 2005 and 350 ppm by 2010 throughout the country. To eliminate sulfur compounds in fuels to get the very low value required by the new regulations, several proposals can be made. For example, to reduce the concentration from 500 to 10 ppm of sulfur in gas oil, it is necessary to do one or more of the following:

- To enhance the activity of the present catalysts considerably, by a factor of 4–5, which would correspond to very high conversion (higher than 99.9% for many gas oils).
- To increase the process severity, especially to increase the hydrogen pressure (from 1.5 to 2 times).
- By a better knowledge of the reactions in the process conditions, to find new catalyst combinations and synergism with noncatalytic processes by a better knowledge of the reactions involved in the process.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008052

In the second point, it means that refiners are faced, at a minimum, with revamps of their existing hydrotreaters, and possibly even reconfiguration of refineries. In some cases, especially where cracked stock feeds must be processed for on-road diesel, new grassroots, high-pressure hydrotreaters are being contemplated. These new units will be very expensive, with capital costs of the order of $1600–$2400 per barrel of the installed capacity, and they will be stretched near their technical limits to achieve ultralow sulfur levels. Hydrotreating process represents some of the most important catalytic processes and the annual sales of HDT catalysts represent close to 10% of the total world market for catalysts.[1] Cobalt–molybdenum and nickel–molybdenum are the most commonly used catalysts for this process. Both types of catalyst remove sulfur, nitrogen, and other contaminants from petroleumfeed. Cobalt–molybdenum catalysts, however, are selective for sulfur removal, while nickel–molybdenum catalysts are selective for nitrogen removal. CoMo catalysts showed a much better sulfur tolerance than NiMo catalysts in the HDS of dibenzothiophene.

HYDROTREATING REACTIONS

Hydrodesulfurization

Sulfur compounds are one of the most important heteroatoms or impurities present in petroleum crudes in varying amounts. Some crudes contain as low as 0.05 wt%, while some sour crudes have as high as 5 wt% sulfur. In addition to these organic sulfur compounds, some hydrogen sulfide can also be found. The distribution of sulfur compounds with respect to boiling range also changes from crude to crude, but generally the content of sulfur (in wt%) increases with the boiling range.

The types of sulfur compounds present in the petroleum fraction also change with the boiling range. Organic sulfides, disulfides, mercaptans, and thiophenes are found in the lighter boiling fractions. Different derivatives of thiophenes, e.g., benzothiophenes, dibenzothiophenes, are predominant in the heavier fractions.

Chemistry and Kinetics of Desulfurization

Hydrodesulfurization reactions occur when hydrogen reacts with the sulfur atom and forms hydrogen sulfide. The rate of the reaction depends on the type of sulfur compounds present, namely, aliphatic and aromatic thiols, sulfides, disulfides, various thiophenes and thiophene derivatives. The order of reactivity among these compounds decreases in the following order:

$$\text{RSH (thiols)} > \text{R–S–S–R}^1\text{(disulfides)} > \text{R–S–R}^1\text{(sulfides)} > \text{thiophenes}$$

Aliphatic and aromatic thiols have high reactivity and the sulfur atom can be removed easily. Aliphatic thiols may react through elimination and HYD:

$$\text{R–CH}_2\text{–CH}_2\text{–SH} \rightarrow \text{R–CH=CH}_2 \rightarrow \text{R–CH}_2\text{–CH}_3$$

or through hydrogenolysis, if a β-H atom is present

$$\text{R–CH}_2\text{–CH}_2\text{–SH} + \text{H}_2 \rightarrow \text{R–CH}_2\text{–CH}_3 + \text{H}_2\text{S}$$

The elimination takes place via acid–base catalysis and can be catalyzed by metal sulfide. Hydrogenation and hydrogenolysis take place on metal sulfide surface.

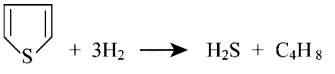
Sulfides and disulfides also react similarly to form hydrogen sulfide and hydrocarbons:

$$\text{RSSR}^1 + 3\text{H}_2 \rightarrow \text{RH} + \text{R}^1\text{H} + 2\text{H}_2\text{S}$$
$$\text{RSR}^1 + 2\text{H}_2 \rightarrow \text{RH} + \text{R}^1\text{H} + \text{H}_2\text{S}$$

Thiophene derivatives are slower to react and their reactivity decreases in the following order:

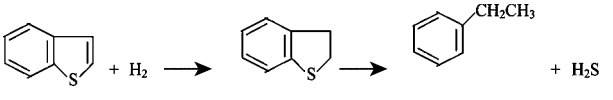
Thiophenes > benzothiophenes
> dibenzothiophenes
> alkyl substituted dibenzothiophenes.

Among the different isomers of alkyl substituted dibenzothiophenes, reactivity varies with the location of alkyl groups. The thiophenes react with hydrogen to form mixed isomers of C_4H_8 and H_2S.



The thiophene ring is not saturated prior to removal of sulfur atom. The first step may be involved with the simultaneous removal of a sulfur atom and the saturation of the hydrocarbon. Studies show that HDS and subsequent HYD reactions occur on separate sites.[2] In case of benzothiophene, however, the thiophene ring is hydrogenated to thiophene before sulfur atom is removed and this is in contrast with thiophene desulfurization.

The primary product from the desulfurization of dibenzothiophene is biphenyl with some amount of phenyl cyclohexane.



The actual mechanisms for all the above reactions are much more complex and involve a series of steps. The low reactivity of dibenzothiophenes is explained by their steric hindrance. When the sulfur atom is covered by two benzene rings on both sides, the benzene rings hinder the adsorption of the sulfur atom onto the catalyst surface. Transformation rates for the conversion of substituted dibenzothiophenes on a $NiMo/Al_2O_3$ catalyst are given in Table 1.[3]

Hydrodenitrogenation

In petroleum fractions, nitrogen is present as organonitrogen compounds. They are present in smaller concentrations than sulfur compounds. The ratio of nitrogen to sulfur varies from 1 : 2 to 1 : 10 depending on the origin of the crude. Normally, the nitrogen content in crude varies from 0.1 to 2 wt%. Nitrogen compounds found in petroleum are principally of the following two types:

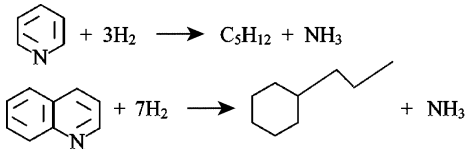
1. Basic nitrogen
2. Nonbasic nitrogen

The basic nitrogen compounds, which are mainly pyridine homologous, occur throughout the boiling ranges, but tend to concentrate in the higher boiling fractions. The nonbasic nitrogen compounds are usually of pyrole, indole, and carbazole types, and also occur in the heavier fractions.

Nitrogen compounds poison catalysts used in secondary processing. Basic nitrogen is a major source of coke formation in catalytic cracking and neutralizes the acid sites in the catalyst. They are also a source of instability in fuels and a source of NO_x emission on burning of heavy fuels. The only industrial method for removing nitrogen from petroleum fractions is HDN.

Chemistry of HDN

Nitrogen compounds are less reactive than sulfur compounds. These compounds are largely present in the form of heterocyclic compounds having five-membered pyrolic rings or six-membered pyridinic rings. The nonheterocyclic compounds, which include aniline, nitriles, and aliphatic amines are easily convertible and therefore are of little concern in industrial HDN. The reactions with pyridine and quinoline during HDN are given below:



In general, the hydrocyclic rings must first be saturated before ring breakup occurs at a C–N bond unlike HDS, which may not require complete HYD. Nitrogen is removed from the resulting amine or aniline as ammonia. Aliphatic amines react readily, but aromatic amines are more refractory. Hydrogenation of the aromatic ring is a prerequisite for the removal of nitrogen atom not only from pyridine, but also from aniline. Because of this, much higher H_2 pressures are required to obtain substantial conversions in HDN than in HDS.

Under the high pressure typical of industrial processes, HDN reactions are not limited by equilibrium. The reaction order for HDN of naphtha, kerosene, gas oil, and vacuum gas oil (VGO) streams is taken as 1 and activation energies are 10, 15, 23, and 25 kcal/mol, respectively.[2]

Hydrodemetallization

The occurrence of metallic constituent in crude oil is of great importance to the petroleum industry even though they are present in small amounts. Even very minute amounts of iron, copper, and particularly nickel and vanadium in the feedstock to the catalytic secondary processing units can affect the activity of the catalyst. These metals are present in the form of high molecular weight organometallic compounds. The concentration of nickel and vanadium can vary from 6 and 33 ppm, respectively, in light Arabian crude to 114 and 387 ppm, respectively, in Venezulean crude. A portion of the metal atom is surrounded by four pyrole-type rings. The nonporphyrinic structures comprise a wide variety of organo-metallic compounds. The principal method

Table 1 Transformation rates for the conversion of substituted dibenzo-thiophenes on a $NiMo/Al_2O_3$ catalyst

	Molecule			
	4,6-DMDBT[a]	**4,6-DEDBT**[b]	**4,6-DiBuDBT**[c]	**4,6-DiPrDBT**[d]
Transformation rate (relative to 4,6-DMDBT)	100	75	40	1.5

[a]4,6-Dimethyldibenzothiophene.
[b]4,6-Diethyldibenzothiophene.
[c]4,6-Dibutyldibenzothiophene.
[d]4,6-Dipropyldibenzothiophene.

of removing these metals is HDT in the presence of catalysts at elevated hydrogen pressure.

Chemistry of HDM

$$\text{M–P} + \text{H}_2 \leftrightarrow \text{M–PH}_2 \text{ reversible HYD}$$
$$\text{M–PH}_2 \rightarrow \text{breakup and deposition}$$
$$\text{irreversible hydrogenolysis}$$

Metal removal rates have generally been interpreted using a single kinetic rate expression of the form:

$$\text{Rate} = kP_{\text{H}_2}{}^m C^n$$

where C = concentration of organometallic compound, $m = 1$, and $n = 1$–2.

Unlike the other HDT reactions, in HDM reactions, metal sulfides are deposited on the catalyst causing permanent fouling leading to replacement of catalysts once their activity is exhausted. Consequently, the catalyst should be designed in such a way that a uniform deposit of metals is obtained throughout the catalyst and the catalyst has high metal retention capacity. Hydrodemetallization catalyst should also have large pore volume with macropores so that large reactant molecules can diffuse in and out of the pores. Ni–Mo catalysts are normally used for their high HYD activity. However, natural materials, e.g., manganese and bauxites have also been considered for HDM catalysts, especially as sacrificed catalyst to absorb metal deposition.

Hydrodeoxygenation

Hydrodeoxygenation involves the removal of oxygen from various oxygen containing compounds in petroleum feedstocks. The oxygen content in petroleum crudes is very low, typically of the order of less than 0.1 wt%. Carboxylic acids and to a lesser extent phenols are found in low- and medium-boiling fractions. However, fuels derived from tar sands, shale oils, and coal have a substantial amount of oxygen compounds. These may include ethers, furan, carboxylic acids, and phenols.

Hydrogenation

Hydrogenation processes are employed to remove unwanted olefins, diolefins, and polycyclic aromatics from various refinery streams. Hydrogenation of olefins takes place at atmospheric pressure, with the reactivity generally decreasing with increasing chain length and substituent groups adjacent to the double bond. Hydrogenation of aromatics requires high pressures of hydrogen for saturation of the double bond. This is partially due to low reactivity of resonance stabilized aromatic structures and partly to equilibrium constraints of pressures and temperatures employed. The relative activities of sulfided $NiMo/Al_2O_3$ for HYD of one ring of various multiring aromatic model compounds are found to be of the following order:[2]

$$\text{Benzene} < \text{phenanthrene} < \text{napthalene} < \text{anthracene}$$

CATALYSTS USED IN THE HYDROTREATING PROCESS AND ITS RECENT DEVELOPMENT

Composition of Catalysts

Supported metal sulfide constitutes the most important catalyst in the HDT process. Industrial HDT catalysts are composed of a molybdenum sulfide (or tungsten sulfide) phase promoted by cobalt or nickel and usually supported on alumina. This combination has the pivotal property in hydroprocessing to convert organosulfur compounds to clean hydrocarbons in the abundant presence of fouling molecules such as H_2S and NH_3. Hydrogen sulfide generally decreases reaction rates owing to competitive adsorption with sulfur containing molecules. However, the behavior is more complex as dissociative adsorption may alter the nature of the active sites. For instance, Leglise et al. pointed out that sulfhydryl species play a positive role in the catalytic mechanism of thiophene conversion at very low hydrogen sulfide partial pressures.[4] High conversion levels imply high hydrogen sulfide partial pressures, and hence the necessity for sulfur-tolerant catalysts. Generally, CoMo catalysts are preferred when a high HDS duty is required, whereas NiMo based materials, which have better HYD activity than CoMo catalysts, are used when nitrogen removal is the major goal. The HDN and HDS reactivity of pure sulfides and sulfide pairs can be ranked as below:

Pure sulfides:

$$\text{Mo} > \text{W} > \text{Ni} > \text{Co} \quad \text{for HDS and HDN}$$

Sulfide pairs:

$$\text{CoMo} > \text{NiMo} > \text{NiW} > \text{CoW} \quad \text{for HDS}$$
$$\text{NiMo} \cong \text{NiW} > \text{CoMo} > \text{CoW} \quad \text{for HDN}$$

Al_2O_3 is the only support that is being used in most commercial HDS catalysts. Before the 1980s, the alternative supports other than Al_2O_3 that were used in most of the investigations were SiO_2 and SiO_2–Al_2O_3. The use of carbon-supported catalysts

that showed outstanding activities for HDS reaction at the laboratory level stimulated some research interest in such systems because it allows easy metal recovery. A major drawback is its low density. Availability of methods to prepare oxides such as ZrO_2, TiO_2, MgO with high surface area generated more interest in these materials as supports. These materials when used as supports could impart four to five times higher activities in Mo and W catalysts. To take advantage of their high intrinsic activities, mixed oxides such as TiO_2–Al_2O_3, ZrO_2–TiO_2 supports figured in many investigations.[5,6] In recent years mesoporous materials such as MCM-41, zeolite supported catalysts, and hydrotalcite derived catalysts received considerable attention because of their potential for possible application in ultradeep desulfurization.

A large number of elements in the periodic table (I, Cl, F, B, P, As, Cs, Na, Ca, Mg, Ga, Li, K, Rb, Zn, Ti, Si, etc.) are used as additives for HDT catalysts. Among all these elements, phosphorous is mostly used as a third element in commercial catalysts. Although it is frequently used, its role is still unknown. Various researchers have reported the following explanations for its role:[7]

1. Phosphorous is generally used to enhance HDN activity by modifying Lewis acidic sites of alumina support.
2. The addition of P into $NiMo/Al_2O_3$ catalyst increases hydrocracking activity. It indicates that P has also an indirect effect on Bronsted acidity.
3. Phosphorous leads to increase in the size of MoS_2 slab and the number of MoS_2 stacking.
4. Phosphorous decreases the metal–support interaction.
5. Phosphorous inhibits the formation of $NiAl_2O_4$, to increase the amount of a P–Ni–Mo heteropoly compound or to form Mo–P tetrapoly and Co–phosphate-like compound.
6. The increase in the dispersion of Co and Mo species is also reported in the literature.
7. Phosphorus increases the sulfidability of molybdenum.
8. Phosphorous can increase the amount of octahedrally coordinated Ni in the polytungstate structure at the expense of Ni in tetrahedral sites.
9. Phosphorous has also been reported to affect the deactivation of HDT catalysts.

Structure of Chemical Components in a Catalyst

Structural information of HDT catalysts has been explained in terms of several models. These models are monolayer model, intercalation model, contact synergy model, and the Co–Mo–S model. In the monolayer model, the molybdenum species was assumed to be bonded to the surface of the alumina forming a monolayer. Interaction of the Mo^{+6} with the alumina was believed to occur via oxygen bridges resulting from reaction with surface OH groups. Cobalt (Co^{+2}) was assumed to be in tetrahedral positions in the surface of the alumina, replacing Al^{+3} ions, which is explained by the promotional effect of Co. In the intercalation model, it is assumed that MoS_2 (WS_2) is on the surface of the alumina carrier where a plane of Mo(W) atoms is sandwiched between two hexagonal, close-packed planes of sulfur atoms. The Co(Ni) ions are then believed to occupy octahedral intercalation positions in the van der Waals' gap between the layer. In the contact synergy model, the presence of Mo as MoS_2 and Co as Co_9S_8 has been proposed. In the Co–Mo–S model, the Co–Mo–S phase was shown to be MoS_2-like structures with the promoter atoms located at the edges in fivefold coordinated sites (tetragonal pyramidal geometry) at the (1010) edge planes of MoS_2.

Pretreatment of Catalysts

Hydrotreating catalysts are supplied in an oxidic form. During the operation, the oxidic form is converted to a sulfided form. In this case, the sulfur present in the feed is converted to H_2S, which subsequently reacts with the metal oxide to form sulfides. A significant operating time can be gained if the catalyst sulfiding is performed prior to HDT. This results from the decreased catalyst deactivation of the presulfided catalyst compared with that of the oxidic catalyst. Presulfiding of the catalyst can be performed either in situ or ex situ. It is evident that at least 40 hr of the operating time can be gained by using the ex situ presulfided catalyst.[8]

During the in situ presulfiding, the catalyst bed is contacted with a sulfur containing liquid or H_2S containing sour gas and H_2. This results in the conversion of metal oxides to the corresponding sulfides, e.g., MoO_3 is converted to MoS_2. Liquid phase sulfidings are preferred because of the better distribution of sulfur across the bed. Also, the liquid acts as a sink of heat generated by the exothermic sulfidation reactions. An example of the procedure used for the in situ presulfiding of the catalyst was published by Gorra et al.[9]. In this case, a hydrogenated diesel fraction was used, thus eliminating the risk of catalyst deactivation by aromatics in the untreated fraction. The fraction was spiked with dimethyldisulfide to give 1 wt% sulfur. An optimal presulfiding was achieved at about 350°C. Fundamental aspects of presulfiding were discussed by Zeuthen et al.[10].

The ex situ presulfiding, perhaps more accurately termed as presulfurizing, comprises wetting of the

oxidic catalyst with an alkylpolysulfide, e.g., ditertiononyl pentasulfide containing about 37 wt% sulfur.[11] Usually, a stoichiometric amount of sulfur is added, i.e., in the case of CoMo catalysts, the added amount should be sufficient to convert all MoO_3 and CoO to MoS_2 and Co_9S_8, respectively. After a thermal treatment, the active oxides of the catalyst are converted to the oxysulfides. In this form, the catalyst is ready to be loaded in the reactor. The final conversion of the oxysulfides to sulfides is completed in situ during the start-up in the presence of H_2 and the feed. The Sulficat and Acticat are perhaps the best-known ex situ presulfiding processes available commercially.[12,13]

DIFFERENT HYDROTREATING PROCESSES

Hydrotreating processes are carried out for different product streams, namely, naphtha, fluidized catalytic cracking gasoline, kerosene/aviation turbine fuel, diesel, and on feed streams to other secondary processes, namely, VGO, atmospheric resid, and vacuum resid. Different HDT processes and their characteristics with respect to operating pressure, temperature, liquid hourly space velocity, and hydrogen consumption are reported by Nigam.[14] The operating pressure and H_2 consumption in four main HDT processes are given in Fig. 1.

REACTORS USED IN HYDROTREATING AND COMPARISON OF THEIR PERFORMANCE

Fixed-Bed Reactor—Cocurrent and Countercurrent Operation

Two types of fixed-bed reactors are used in HDT processes. Fixed-bed reactors having randomly packed catalyst in tubular devices traversed vertically downward by a gas–liquid stream are mainly used and generally known as TBR. Depending on the capacity of the plant, the reactor may vary from 1 to 6 m in diameter. The TBR reactor configuration ensures flexibility of operation and high throughputs of gas and liquid. Owing to a motionless catalyst bed, nearly plug flow is achieved in TBRs and in that respect they are superior to other three-phase reactors where the catalyst is either slurried or fluidized.[15,16] Trickle-bed reactors are usually operated at elevated pressures of about 2–30 MPa to slow down catalyst deactivation, increase the concentration of the gaseous component in the liquid phase, attain high conversion, achieve better heat transfer, and handle large volumes of gas at less capital cost. The other fixed-bed reactor configuration with countercurrent flow of gas and liquid has been proposed by Kundu et al.[17]. Though it has advantages with respect to lower H_2S and NH_3 concentration in the major part of the bed and better axial temperature profile compared to conventional TBR, the countercurrent configuration is limited to low velocities far below those of industrial interest because of excessive pressure drop and flooding problem. This problem has been minimized by incorporating structured catalytic packing by sacrificing the volume of the reactor in an optimized way between the reactor volume and the rate of conversion (details are in Kundu et al.[18]).

Ebullated Bed Reactor

This reactor is used for feeds having high metal or when higher conversion is required. The catalysts are in suspension and the option for the addition and removal of catalysts is present. It is most applicable in highly exothermic reactions. The expansion rate of

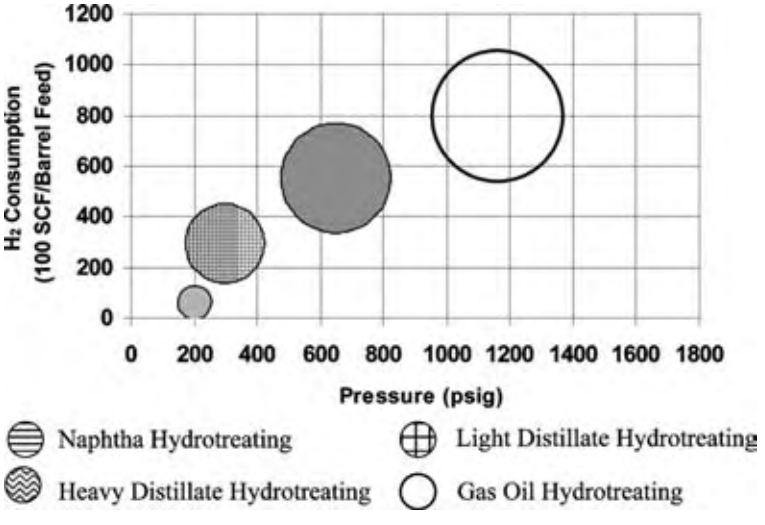


Fig. 1 Applications of hydrotreating processes and operating conditions.

the bed is kept at 1.3–1.5 under normal conditions. The advantages of the reactors are 1) high heat transfer rates; 2) uniform temperature; 3) catalyst replacement; and 4) much lower pressure drop. The disadvantages are 1) backmixed flow and the volume of the reactor not being fully utilized; 2) catalyst attrition due to motion; and 3) higher consumption of catalyst.

Moving-Bed Reactors

In moving-bed reactors, both the feed and the catalyst move in cocurrent downflow and the catalysts are continuously renewed. The metal content increases along the bed and metal-rich catalysts are withdrawn from the bottom. It can handle the feed having higher metal content. The moving-bed reactor can be used as a first reactor for demetallization and asphaltene disaggregation. Other conversions (HDN, HDS) can take place in a fixed bed downstream.

SELECTION OF A PROCESS FOR CATALYST DEVELOPMENT

There were basically two approaches, which were used in the past for HDT process development studies using catalyst in the commercially applied size and shape. The first one, which was followed 30–40 yr ago in various industrial research and development centers, was to test the commercial catalyst in large pilot plants. The second approach was to use a smaller pilot plant and simulate the data generated in these units, applying a suitable hydrodynamic model to predict the performance of a commercial unit. These are generally known as small-scale TBRs. Because of the presence of a liquid phase, the problems in these small TBRs are more complex as compared to those present in other small-scale fixed-bed catalytic reactors handling only vapor phases.

The catalysts are made in small quantities and to avoid the complexity of preparing shaped catalysts, it is common to test catalysts in the form of powder or fine particles. This also helps to compare the intrinsic reaction rates of various catalyst formulations. Because the number of catalyst samples to be tested are many, the testing method needs to be simple and fast. However, this should not compromise with the reliability and meaningfulness of the data generated. Because of the small amount of catalyst available, slurry reactors using batch autoclave or fixed-bed continuous microreactors are typically used for this stage of initial catalyst screening. A description of these two reactors for the purpose of initial screening of the catalyst is available in Bej.[19]

Reactors for Process Development and Optimization Studies

The best catalyst selected through previous screening tests is formed into a shape (cylindrical or multilobe) and undergoes further experimentation for process development and optimization. Some additional experiments are needed at this stage to select the proper size and shape of the catalyst and also to determine the effect of other ingredients such as binders and additives used in forming. Two points are very important in selecting an appropriate reactor at this stage of research. First, the catalysts need to be tested in their commercially applied size and shape. Second, the objective at this stage of research is to predict the full performance of the catalyst for a commercial operation rather than simple preliminary screening. The diameter of a small-scale (bench- or microscale) TBR is much smaller than those of a commercial reactor. Hence, the ratio of the reactor diameter to catalyst particle diameter is very low (<100). The ratio of the catalyst bed height to the particle diameter is also low in these reactors. Because of the smaller volume of catalyst used, the liquid flow rate is also lower. The values of these parameters along with the Reynolds number for various sizes of the reactors are given in Sie and Krishna.[20] Because of the low values of these factors, a number of problems such as incomplete wetting of catalyst and backmixing of liquid are observed in these small-scale reactors.[21]

Dilution Technique

The theories of dilution techniques have been explained in detail by various researchers.[20,22,23] The selection of the proper size of diluent is very important. Equal volumes of diluent and catalyst were used for this comparison. For the case of an undiluted bed, only commercial size catalyst is packed in a small-scale reactor. The wall effect is very significant and in this case causes channeling of liquid. Because of the high void space inside the catalyst bed, the liquid holdup in the catalyst bed is also very low. As a whole, there is incomplete wetting of catalyst and only partial utilization is achieved in this case. Besides this, an appreciable amount of axial backmixing is present in the undiluted catalyst bed. When a larger size of diluent is used, it cannot enter the void space between the catalyst particles. Thus, it does not increase liquid holdup and hence only partial utilization of catalyst is also obtained. However, the addition of diluent increases the bed height, which in turn reduces liquid axial dispersion to some extent. When the diluent size is smaller, it can enter the narrow void space between the catalyst particles and can increase the liquid holdup

of the bed. Thus, almost complete wetting of catalyst is achieved by using a fine size of diluent. The use of fine size of diluent also reduces the axial dispersion in a significant way.

In essence, diluting the catalyst bed with appropriate size of inert particles increases liquid holdup, improves catalyst wetting, and reduces liquid backmixing. The selection of diluent size depends on several factors such as the length and inner diameter of the reactor, the size, shape and amount of catalyst, and the flow rate of reactants. Therefore, there is a lot of research interest to determine the appropriate size of the diluent for various sizes of small-scale reactors to overcome their limitations.

Operating in Upflow Model

Sometimes, the researchers used upflow mode for the catalyst development in a bench-scale TBR for overcoming the partial wetting and wall flow. The upflow mode of operation suffers from the serious drawback of nonideal flow of liquid and the formation of stagnant zone inside the catalyst bed. A large difference in hydrodynamics for cocurrent downward flow and cocurrent upward flow was obtained when using a large size of diluents. This difference can be minimized by using higher liquid flow rates and smaller size diluents. The comparison of upflow and downflow operation with and without diluents was carried out by Chander et al., De wind et al., Khadilkar et al., and Dudukovic et al.[23–26].

COMMERCIAL HDT PROCESSES

Different commercial HDT processes using TBRs are available. For diesel hydrotreatment process, IFP's technology, and UOP's unionfining process is widely used in the petroleum refining industry. ABB Lummus in association with Criterion catalysts offers different HDT processes for upgradation of diesel quality under the family of technologies called SynSat Technology. The process uses intermediate by-product gas removal and optional countercurrent gas flow in addition to the conventional cocurrent downward mode of gas and liquid flow (Fig. 2). The HDT processes using licensed ebullating bed processes include:

- LC-Fining. Licensed by ABB Lummus Global Inc., Oxy Research and Development Co., and BP Amoco Corporation.
- H-Oil. Licensed by IFP North America and Texaco.
- T-Star. Licensed by IFP North America and Texaco.

These LC-Fining and H-Oil both use similar technologies but offer different mechanical designs.

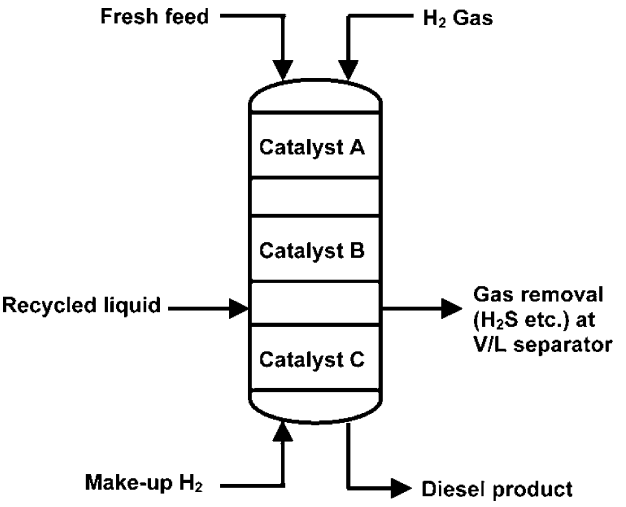


Fig. 2 SynSat process with Criterion/Lummus catalytic HDT reactor technology.

LC-Fining system consists of three sections, i.e., fresh catalyst handling, daily addition/withdrawal of catalyst to and from the reactors, and a spent catalyst handling system. The advanced design of the H-Oil reactor incorporates an improved internal recycle cup enabling a complete separation of the recycle liquid from the gas.[8] With this improvement the throughput of the feed was increased.

CONCLUSIONS

Hydrotreating is very important in petroleum refining industries and is an essential process for treatment of petroleum products owing to the stringent environmental regulations imposed on the world. The appropriate catalyst and reactors are necessary for getting deep conversion (e.g., deep HDS). Apart from the elementary composition of the catalyst, different types of catalyst supports and additives are used to get better performance of the catalyst. The development of the catalysts is still a result of trial-and-error experiments and empirical knowledge rather due to the desired fundamental understanding. Regarding the selection of reactors, countercurrent operation has been tried to commercialize for getting high conversion. There is a necessity to optimize the reactor volume utilization vs. the conversion for this reactor. Countercurrent TBR will also help to handle easily the heavy fraction of the oil.

REFERENCES

1. Topsoe, H.; Clausen, B.S.; Massoth, F.E. *Hydrotreating Catalysis*; Springer-Verlag: Berlin, 1996.

2. Manna, U. *Hydrotreatment*; Verma, R.P., Bhatnagar, A.K., Eds.; Lovraj Memorial Trust: Delhi, 2000; 103–168.
3. Breysse, M.; Djega-Mariadassou, G.; Pessayre, S.; Geantet, C.; Vrinat, M.; Pérot, G.; Lemaire, M. Deep desulfurization: reactions, catalysts and technological challenges. Catal. Today **2003**, *84* (3–4), 129–138.
4. Leglise, J.; Van Gestel, J.; Duchet, J.C. Promotion and inhibition by hydrogen sulfide of thiophene hydrodesulfurisation over a sulfide catalyst. J. Chem. Soc. Chem. Commun. **1994**, *5*, 611–612.
5. Segava, K.; Takahashi, K.; Satoh, S. Development of new catalysts for deep hydrodesulfurization of gas oil. Catal. Today **2000**, *63* (2–4), 123–131.
6. Maity, S.K.; Rana, M.S.; Srinivas, B.N.; Bej, S.K.; Murali Dhar, G.; Prasada Rao, T.S.R. Characterization and evaluation of ZrO_2 supported hydrotreating catalysts. J. Mol. Catal. A Chem. **2000**, *153* (1–2), 121–127.
7. Maity, S.K.; Ancheyta, J.; Soberanis, L.; Alonso, F. Catalysts for hydroprocessing of Maya heavy crude. Appl.Catal. A Gen. **2003**, *253* (1), 125–134.
8. Furimsky, E. Selection of catalysts and reactors for hydroprocessing. Appl. Catal. A Gen. **1998**, *171* (2), 177–206.
9. Gorra, F.; Scribano, G.; Christensen, P.; Andersen, K.V.; Corsaro, O.G. New catalyst, improved presulfiding result in 4+ year hydrotreater run. Oil Gas J. **1993**, *91* (34), 39–43.
10. Zeuthen, P.; Blom, P.; Muegge, B.; Massoth, F.E. Temperature-programmed sulfidation and oxidation of Ni-Mo/alumina catalysts and reaction with ammonia. Appl. Catal. **1991**, *68* (1), 117–130.
11. Dufresne, P.; Brahma, N.; Labruyere, F.; Lacroix, M.; Breysse, M. Activation of off site presulfided cobalt–molybdenum catalysts. Catal. Today **1996**, *29* (1–4), 251–254.
12. Eurecat Group. The SULFICAT Process, http://www.eurecat.fr/.
13. Catalyst Recovery Institute. actiCAT Presulfiding, http://www.catalystrecovery.com/.
14. Nigam, K.D.P. Design of trickle-bed reactors. In *Hydroprocessing in Petroleum Refining Industry—A Compendium*; Verma, R.P., Bhatnagar, A.K., Eds.; Lovraj Memorial Trust: Delhi, 2000; 391–424.
15. Saroha, A.K.; Nigam, K.D.P. Trickle bed reactors. Rev. Chem. Eng. **1996**, *12* (3–4), 207–347.
16. Al-Dahhan, M.H.; Larachi, F.; Dudukovic, M.P.; Laurent, A. High-pressure trickle-bed reactors: a review. Ind. Eng. Chem. Res. **1997**, *36* (8), 3292–3314.
17. Kundu, A.; Bej, S.K.; Nigam, K.D.P. A novel countercurrent fixed bed reactor. Can. J. Chem. Eng. **2003**, *81* (3–4), 831–837.
18. Kundu, A.; Nigam, K.D.P.; Duquenne, A.M.; Delmas, H. Recent developments in hydroprocessing reactors. Rev. Chem. Eng. **2003**, *19*, 531.
19. Bej, S.K. Performance evaluation of hydroprocessing catalysts. A. Rev. Exp. Tech. Energy Fuels **2002**, *16* (3), 774–784.
20. Sie, S.T.; Krishna, R. Process development and scale up: III. Scale up and scale down of trickle bed processes. Rev. Chem. Eng. **1998**, *14*, 203.
21. Kundu, A.; Nigam, K.D.P.; Verma, R.P. Catalyst wetting characteristics in trickle-bed reactors. AIChE J. **2003**, *49* (9), 2253–2263.
22. Sie, S.T. Scale effects in laboratory and pilot-plant reactors for trickle-flow processes. Rev. Inst. Franc. Pet. **1991**, *46*, 501.
23. Chander, A.; Kundu, A.; Bej, S.K.; Dalai, A.K.; Vohra, D.K. Hydrodynamic characteristics of cocurrent upflow and downflow of gas and liquid in a fixed bed reactor. Fuel **2001**, *80* (8), 1043–1053.
24. De Wind, M.; Plantenga, F.L.; Heinerman, J.J.L. Upflow versus downflow testing of hydrotreating catalysts. Appl. Catal. **1988**, *43* (2), 239–252.
25. Khadilkar, M.R.; Wu, X.Y.; Al-Dahhan, M.H.; Dudukovic, M.P. Comparison of trickle-bed and upflow reactor performance at high pressure: model predictions and experimental observations. Chem. Eng. Sci. **1996**, *51* (10), 2139–2148.
26. Dudukovic, M.P.; Larachi, F.; Mills, P.L. Multiphase reactors—revisited. Chem. Eng. Sci. **1999**, *54* (9), 1975–1995.

Immobilized Enzyme Technology

I

Charles G. Hill, Jr.
Department of Chemical and Biological Engineering, University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.

Cristina Otero
Departamento de Biocatálisis, Instituto de Catálisis y Petroleoquimica, CSIC, Campus Universidad Autonoma, Cantoblanco, Madrid, Spain

Hugo S. Garcia
UNIDA, Instituto Tecnologico de Veracruz, Veracruz, Mexico

INTRODUCTION

Enzymes are proteins employed by Mother Nature to catalyze the chemical reactions necessary to sustain life in plants and animals. As catalysts, enzymes may influence the rates and/or the directions of chemical reactions involving an enormous range of substrates (reactants). Enzymes function by combining with substrates to form enzyme–substrate complexes (reaction intermediates) that subsequently react further to yield products while regenerating the free enzyme.

The region of the enzyme that interacts with substrates is referred to as the active site. For reaction to occur there must be an appropriate fit between the three-dimensional structure of this site and the geometry of the reactant molecule so that an enzyme–substrate complex may form (Emil Fischer's ''lock and key'' hypothesis). Enzymes are relatively labile species and when subjected to unfavorable conditions of temperature, pH, pressure, chemical environment, etc., they can lose their catalytic activity. In these situations, deactivation of the enzyme can usually be attributed to changes in the geometric configuration of the active site.

Enzymes are characterized by unusual specific activities and remarkably high selectivities. They are effective catalysts at relatively low temperatures and ambient pressure. The primary driving force for efforts to develop immobilized forms of these biocatalysts is cost, especially when one is comparing process alternatives involving either conventional inorganic catalysts or soluble enzymes. Immobilization can permit conversion of labile enzymes into forms appropriate for use as catalysts in industrial processes—production of sweeteners, pharmaceutical intermediates, and fine chemicals—or as biosensors in analytical applications. Because of their high specificities, immobilized versions of enzymes are potentially useful in situations where it is necessary to obtain high yields of the desired product to minimize downstream processing costs and the environmental impact of a process.

Because the costs of isolation and purification of soluble enzymes are high and it is often both technically difficult and costly to recover an active form of the enzyme from product mixtures when the reaction of interest is completed, soluble enzymes are normally employed only in batch operations in which the enzymes are removed from the liquid product by precipitation. Thermal deactivation may be used instead to destroy the catalytic activity of the enzyme. Immobilization of the enzyme circumvents these difficulties because the solid phase containing the enzyme is easily recovered from the product mixture. Use of immobilized enzymes makes it possible to conduct the process in a continuous flow mode, thereby facilitating process control via manipulation of the flow rate of the process stream. One can offset losses in enzyme activity as time elapses by reducing the flow rate to maintain a constant product composition. Operation in this mode permits one to obtain more product per unit of enzyme employed.

TECHNIQUES FOR IMMOBILIZATION OF ENZYMES

A variety of physical and chemical methods have evolved for immobilizing enzymes on or within solid supports. Kennedy and Cabral employed a variation of the scheme in Fig. 1 to classify techniques for immobilization of enzymes.[1] Judicious choice of the support is essential not only for the stability of immobilized enzymes, but also for the operational characteristics of the device containing the immobilized enzyme and the economic viability of the intended application. The discussion below and the information in Table 1 indicate some of the criteria employed in selecting a mode of immobilization.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120018070

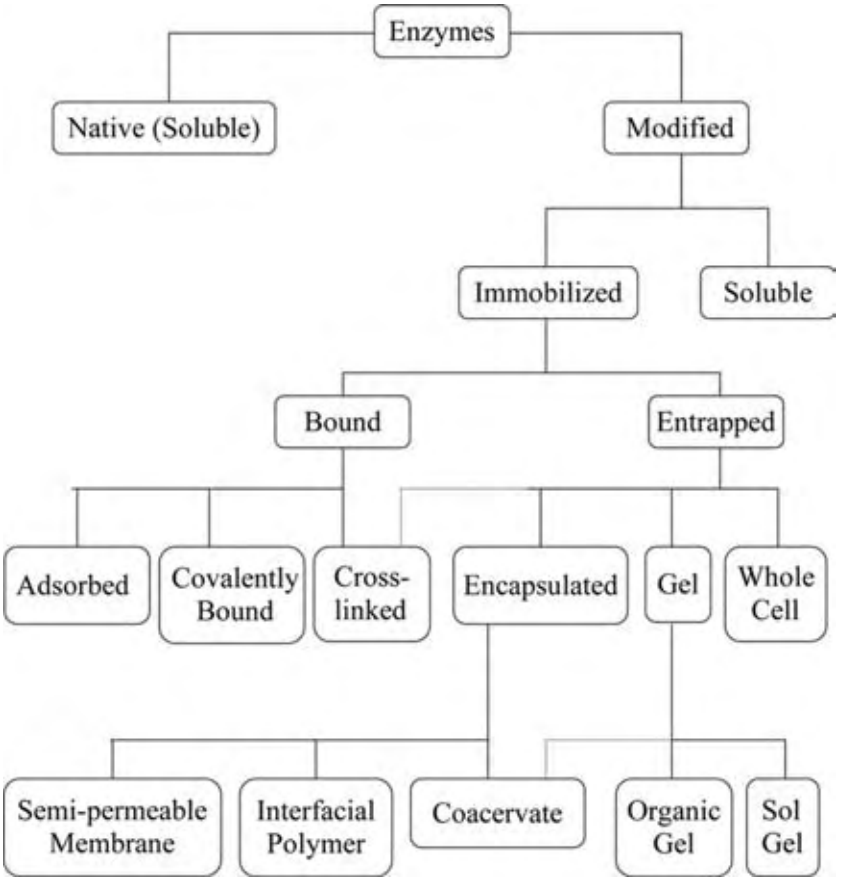


Fig. 1 Schematic representation of modes of immobilization of enzymes. (Dotted lines indicate potential alternative uses/classifications.)

Enzyme Carriers—Particulate Supports

Industrial scale processes involving immobilized enzymes are normally carried out in fixed-bed reactors. Hence, the desired characteristics of the catalyst support are closely akin to those for the heterogenous catalysts commonly employed in the chemical industry, namely:

- Chemical, mechanical, and thermal stability.
- Resistance to both microbial degradation and swelling or dissolution in the reaction medium.
- High permeability to reactant and product species (and to the enzyme during the immobilization process) (pore diameters in the 10–50 nm range and porosities of ~50% or better).
- The requisite hydrophobicity or hydrophilicity for the intended application.
- A moderately high specific surface area (~50 m^2/g) and a pore size distribution that provides adequate capacity for adsorption of the enzyme.
- A suitable shape and particle size (to minimize pressure drop), typically greater than 0.5 mm in diameter.
- Relatively low cost.
- Regenerability.
- Inertness with respect to both the enzyme-mediated reactions of interest and reactions leading to deactivation of the enzyme.

Because naturally occurring materials do not meet the morphological specifications, most carrier materials are synthesized via routes that produce the desired characteristics. Boller, Meier, and Menzler have indicated that although immobilization of enzymes on solid supports has been studied for half a century, there are no generally applicable rules for selecting the proper support for a specific application.[2] Nonetheless, they also indicate that microporous and mesoporous epoxy-activated acrylic beads with particle diameters in the 100–250 μm range (pore radii in the 10–100 nm range) are popular supports for the preparation of multiton quantities of catalysts for industrial biotransformations.

Methods of Immobilization

Several physical and chemical methods can be utilized to immobilize an enzyme on a solid support (Fig. 1).

Physical adsorption

The simplest method of immobilization is physical adsorption of the enzyme on the carrier. The procedure consists of contacting a solution of the enzyme with the support material under appropriate conditions, and

Table 1 Comparison of different modes of immobilization of enzymes

	Modes of immobilization							
Characteristic	**Physical adsorption**	**Ionic binding**	**Chelation**	**Covalent bonding**	**Cross-linking**	**Physical entrapment**	**Membrane entrapment**	**Whole cells**
Ease of preparation	Simple	Simple	Simple	Difficult	Intermediate	Difficult	Simple	Intermediate
Binding force	Weak	Intermediate	Intermediate	Strong	Strong	Intermediate	Weak	Intermediate
Fraction of original enzyme activity	Intermediate	High	High	High	Low	Small	High	High
Ease of regeneration	Possible	Possible	Possible	Rare	Impossible	Impossible	Possible	Impossible
Cost of immobilization	Low	Low	Intermediate	High	Intermediate	Intermediate	Intermediate	Intermediate
Stability	Low	Intermediate	Intermediate	High	High	High	Intermediate	Intermediate
General applicability	Yes	Yes	Yes	No	No	Yes	Yes	Some
Protection of enzyme from microbial attack	None	None	None	None	Good	Some	Some	Some

after allowing sufficient time to elapse, separating the solution from the now insoluble enzyme preparation by filtration, centrifugation, or other means. Because no chemical interactions are involved, there is little or no conformational change in the enzyme. Thus, the impact on the geometry of the site at which the biocatalyst interacts with substrates is minimal. The forces binding the enzyme to the carrier are relatively weak and may involve hydrogen bonding and hydrophobic interactions in addition to conventional van der Waals forces. The extents of adsorption and retention of activity are dependent on experimental parameters such as the pH of the solution/suspension, temperature, ionic strength, species concentrations, and the chemical nature of the solvent.

It is important to allow sufficient time for diffusion of the enzyme into the pore structure of the support to maximize the extent of physical adsorption of the enzyme. Surface coverages of the support by adsorbed enzymes may range from a fraction of a monolayer to multiple layers. Because the forces binding the enzyme to the support are relatively weak, enzymes immobilized by physical adsorption are susceptible to desorption during use. Shifts in microenvironmental conditions, such as changes in pH, ionic strength, temperature, composition of the solvent, etc., can lead to desorption with concomitant apparent loss of activity of the biocatalyst. On the other hand, these characteristics can sometimes be advantageous in protocols used to regenerate a fixed bed of immobilized enzyme, once it has lost a significant fraction of its original activity. In some cases, it may be appropriate to use a multifunctional cross-linking agent such as glutaraldehyde to chemically bind the adsorbed protein molecules to one another or to the underlying surface of the support. This approach minimizes the potential for loss of activity by desorption of the enzyme, but renders the task of regeneration of the activity of the biocatalyst much more difficult.

Ionic bonding and chelation

A useful variation of the physical adsorption method involves adsorption of the enzyme on carriers whose structures contain anion or cation exchange residues. Unless the pH of the system corresponds to the isoelectric point of the enzyme, the interactions of the net charge on the protein with opposite fixed charges on the solid support enhance the strength of adsorption of the enzyme on the support. In practice, both ionic bonding and physical adsorption occur simultaneously; the main difference is that when ionic forces are present, the strength of the interaction is greater. An alternative to employing charge bearing carriers is to enhance the strength of the adsorbate–adsorbent interaction by taking advantage of the ability of some transition metal compounds to form chelate structures with enzymes.

Covalent bonding

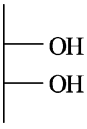
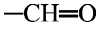
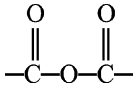
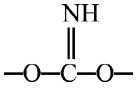
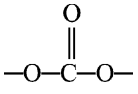
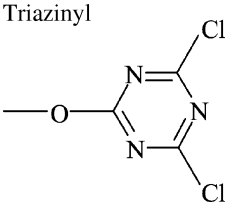
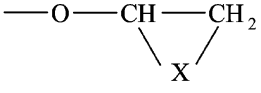
Stronger bonds between the enzyme and the carrier can be formed when covalent bonding is employed. Enzymes are copolymers (proteins) composed of a variety of amino acid monomers. They possess a number of reactive side chains that can be utilized for the purpose of forming covalent bonds with solid supports, whose surfaces contain appropriate functional groups. The functionalities present in the side chains of the protein typically include amino, carboxylic acid, sulfydryl, hydroxyl, imidazole, disulfide, indole, and phenol groups. The particular chemistry employed in forming covalent bonds between the enzyme and the carrier is selected on the basis of the chemical nature of the support and the intended application. The range of potential coupling processes is vast for both inorganic and organic supports, and methodologies based on several different types of chemical reactions are described in Refs.[1,3–5]. A sample of some of the most commonly employed types of reactions is presented in Table 2.

If the enzyme is to retain significant catalytic activity after being covalently bound to the support, immobilization should occur via functional groups that are not associated with the active site of the enzyme. This limitation may be difficult to surmount, and enzymes immobilized via covalent bonding may thus suffer significant losses of activity relative to the activity of the soluble precursor. Nonetheless, the wide variety of supports with functional groups capable of reacting (or susceptible to appropriate functionalization) via a range of chemistries with enzymes makes covalent bonding a generally applicable route for immobilization of enzymes. Enzymes bound to supports in this manner are not susceptible to desorption from the surface during use.

Cross-linking

Enzymes can be readily cross-linked using a bi- or multifunctional reagent, such as glutaraldehyde, bis-diazobenzidine-2,2-disulfonic acid, or toluene diisocyanate, which can react with free amino or carboxyl groups or with other functional groups that might be present in the enzyme. Cross-linking of proteins results in an insoluble polymer that may not possess appropriate physical properties for the intended application. Cross-linking is often used in combination with other techniques (especially physical entrapment and physical adsorption) to obtain a material that has enhanced stability.

Table 2 Examples of reactions commonly used for covalent bonding of enzymes to solid surfaces

Functional group of the support	Functional group of the enzyme	Coupling reagent
Amine, $-NH_2$	$-NH_2$	Glutaraldehyde
Hydroxyl, –OH	$-NH_2$	$(CH_3O)_3$ Si $(CH_2)_3$ NH_2 + glutaraldehyde
	$-NH_2$	Cyanogen bromide or tresyl chloride
Aldehyde, 	$-NH_2$	None
Acid anhydride, 	$-NH_2$	None
Imidocarbonate, 	$-NH_2$	None
Cyclic carbonate, 	$-NH_2$	None
Triazinyl 	$-NH_2$	None
 (X = NH, O, S)	$-NH_2$,-OH, -SH	None
$-COOH$	$-NH_2$	Carbodiimide (RN=C=NR)

Physical entrapment of enzymes

Enzymes can also be immobilized by physical entrapment, either in a solid matrix or encapsulated by a membrane that is permeable to low-molecular-weight species, but not to high-molecular-weight species. These membranes may be fabricated from polymers or formed by interfacial polymerization of appropriate monomers. The membranes may also be parts of living cells that have the capacity for bringing about the desired chemical transformations as part of the metabolic processes required for their existence. In addition, the membranes can consist of the walls of cells subjected to lysis.

Gel Formation. Entrapment of enzymes in solid matrices composed of synthetic or natural polymers or inorganic gels is a relatively simple process. The basic technique involves occlusion of the enzyme within the lattice of a solid matrix as the matrix is formed by polymerization, precipitation, or coacervation. Gels

formed from polysaccharides, especially those derived from cellulose and algae, have often been used in combination with cross-linking agents to immobilize enzymes. Both anionic polymers (e.g., carrageenan, carboxymethylcellulose, and sodium alginate) and cationic polymers (e.g., chitosan) have been used to form ionotropic gels within which enzymes or whole cells can be entrapped. This approach leads to a product that is not robust with respect to changes in ionic strength or pH, but it is usually the method of choice for immobilizing whole cells. Gels formed by polymerization of acrylic and methacrylic acids can be activated using a soluble carbodiimide as a precursor to production of covalently bound enzymes. In addition to the organic gels noted above, it is also possible to employ sol–gel techniques based on hydrolysis of metal alkoxides to produce inorganic polymeric structures within which enzymes can be physically entrapped.

A major disadvantage of the gel entrapment route to immobilization is the potential for physical loss of the enzyme as time elapses. To circumvent this problem, cross-linking agents, such as *N,N′*-methylene-bis-acrylamide or glutaraldehyde, may be used to more firmly immobilize the enzyme or to provide mechanical stability. However, the more rigid the matrix the greater is the possibility that diffusional resistance to transport of reactants (substrates) to the site of the enzyme and of products out of the gel will limit the reaction rate.

Encapsulation. Immobilization of enzymes by encapsulation within semipermeable structures dates back to the 1970s.[6] There are three fundamental variations of this approach. In coacervation, aqueous microdroplets containing the enzyme are suspended in a water-immiscible solvent containing a polymer, such as cellulose nitrate, polyvinylacetate, or polyethylene. A solid film of polymer can be induced to form at the interface between the two phases, thereby producing a microcapsule containing the enzyme. A second approach involves interfacial polymerization in which an aqueous solution of the enzyme and a monomer are dispersed in an immiscible solvent with the aid of a surfactant. A second (hydrophobic) monomer is then added to the solvent and condensation polymerization is allowed to proceed. This approach has been used extensively with nylons, but is also applicable to polyurethanes, other polyesters, and polyureas.

A third general approach to encapsulation involves the use of permselective membrane devices of the types employed in ultrafiltration and nanofiltration of aqueous solutions, especially those devices that employ the membrane in the form of hollow fibres. In effect, the enzymes are retained within a macrocapsule. An aqueous solution of the soluble enzyme or whole cells is contained on the retentate side of the membrane, while a solution containing the substrates is supplied to the permeate side of the membrane. The reactants are transported across the membrane to the retentate side where they undergo reaction. The products then diffuse out of the retentate zone across the membrane to the permeate side where they can be removed by convective transport. Throughout, the enzyme is retained on the retentate side provided that one employs a membrane whose molecular cutoff value is significantly below the molecular weight of the enzyme. An advantage of this approach is that this apparatus permits periodic replacement of the biocatalyst.

Reactions mediated by enzymes immobilized by coacervation, interfacial polymerization, retention by semipermeable membranes, or gelation are particularly susceptible to mass transfer/diffusional limitations on the rate.

Immobilized Cells. Instead of immobilizing individual enzymes in molecular form, one can elect to immobilize whole (living) cells containing the enzymes of interest. Many of the techniques employed for immobilization of enzymes are readily extended to immobilization of whole cells, especially those methods involving physical entrapment of the enzymes. In essence, immobilization of whole cells is just another means of physically encapsulating the enzyme(s) of interest. Furthermore, immobilization of whole cells circumvents the need for the multiplicity of processing steps involved in isolating and purifying intracellular enzymes with concomitant reductions in cost. The stability of the desired enzyme(s) is usually enhanced by virtue of the fact that its natural environment is maintained during both immobilization and use. This advantage is of particular benefit in the case of membrane-bound enzymes and for enzyme-mediated reactions involving the participation of either cofactors or multiple enzymes. The necessity for purification of multiple enzymes is avoided and the optimal spatial location of these enzymes within various compartments of the cell remains intact, as do the sites for regeneration of cofactors. Hence, the structural integrity of the catalytic complex is retained. Moreover, the enzymes present in immobilized cells are much more robust with respect to local perturbations in pH, temperature, ionic strength, and the presence of substances that cause deactivation of the enzyme (e.g., toxic metal ions).

On the other hand, use of whole cells as the vehicle for immobilization of enzymes is not without problems. These disadvantages include the susceptibility to mass transfer/diffusional limitations on reaction rates and possible losses in the yield of the desired product as a consequence of unwanted side reactions. In addition, there are potential problems associated with maintaining the integrity of the immobilized cells—supplying the nutrients, energy sources, or cofactors necessary to maintain the cells in a sufficiently viable condition to mediate the reaction(s) of interest.

Comparison of immobilization techniques

Table 1, an extension of the work of Kennedy and Cabral, is a concise summary of important characteristics of different immobilization techniques.[7]

APPLICATIONS OF IMMOBILIZED ENZYMES

Industrial Applications

In spite of the high expectations generated by immobilized enzyme technology in the last third of the 20th century, only a limited number of reductions to industrial practice have been accomplished. Very few large-scale immobilized enzyme processes can successfully compete with processes based on either free enzymes or more conventional catalysts. Some of these are indicated below.

Applications in the Food Industry

Immobilized glucose isomerase for the production of high-fructose corn syrup

A major shift in the technology employed for the production of sweeteners began in the early 1960s with the introduction of soluble enzymatic methods for the production of dextrose syrups from starch. (In commercial practice, the term dextrose is used instead of glucose.) This seminal change was followed about a decade later by utilization of immobilized glucose isomerase for the production of high-fructose corn syrups (HFCS) for use as sweeteners in the manufacture of soft drinks and other foods and beverages. Glucose and fructose have the same molecular formula ($C_6H_{12}O_6$), but differ in geometric configuration (Fig. 2) and sweetening power. Schenck has reviewed the technology for production of high-fructose syrups.[8]

Starch, a polymer formed from glucose monomers, is the principal storage carbohydrate of plants and commercially is obtained primarily from corn. The industrial process involves, first, acid or enzyme (soluble bacterial α-amylase) catalyzed hydrolysis of aqueous suspensions of gelatinized starch to obtain a product with a low dextrose equivalent (DE) of 5–12 and sugars such as glucose and maltose. (The DE is the percentage of the dry matter that consists of reducing sugars expressed as dextrose. This parameter indicates the percentage of the glycosidic linkages in the starch precursor that have been cleaved by hydrolysis.) This hydrolysis dissolves the starch. Further hydrolysis with soluble α-amylase yields a product with a DE in the range of 8–15. This product is then saccharified using one or more soluble enzymes. Fungal α-amylase and fungal glucoamylase (GA) are utilized either separately or in combination (and sometimes in combination with pullulanase) to produce a dextrose syrup with a typical DE of 42. Soluble enzyme preparations from *Aspergillus niger, A. oryzae*, or *Rhizopus oryzae* are then employed for additional saccharification to obtain the 95–98 DE feedstock necessary for the production of HFCS. The resulting products are sweeteners whose compositions and applications depend on the extent of hydrolysis mediated by the enzymes in question. Yields of glucose may be as high as 95–97%. However, the sweetening power of glucose suffers by comparison to that of fructose, and it is the subsequent conversion of glucose to fructose for which immobilized glucose isomerase (xylose isomerase) is an effective biocatalyst. This isomerization reaction constitutes the heart of the technology that brought about a revolution in the manufacture of sweeteners. The resulting syrups compete successfully with sucrose (cane sugar) in many food applications. Virtually all manufacturers of soft drinks use HFCSs in their formulations.

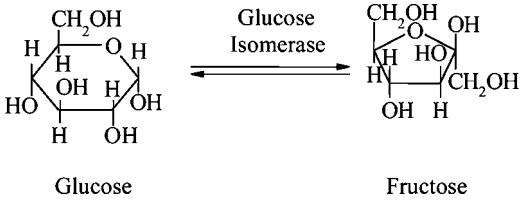


Fig. 2 Isomerization of glucose to fructose over an immobilized glucose isomerase.

Production of fructose from glucose became commercially viable only after adequate procedures for immobilization of glucose isomerase were developed, so that the same quantity of enzyme could isomerize large quantities of substrate in a packed-bed reactor fed continuously with a solution of maltodextrins. Process conditions (typically 55–65°C and pH 7.5–8.5) depend on the particular form of the immobilized enzyme. Reactor diameters are typically between 0.6 and 1.5 m, with corresponding heights of 2–5 m. Initial residence times are less than 1 hr, but to compensate for the loss of enzyme activity as time onstream elapses it is necessary to increase the residence time by reducing the flow rate of the feed stream so as to maintain the composition of the effluent constant. Manufacturers frequently employ large numbers of reactors in tandem to maintain constant production rates. After several months when the activity of a particular packed bed decreases to about 10% of its initial value, that bed can be removed from the reactor network and the biocatalyst replaced by a new charge of enzyme.

Because glucose isomerase is formed intracellularly in many bacterial strains of commercial interest, some industrial processes have utilized immobilized cells,

rather than isolated enzymes, in this application. In whole-cell processes the microbial cells are recovered from the fermentation broth and treated to maintain both enzymatic activity and particle integrity. For the biocatalysts derived from soluble enzymes, the enzyme is separated from the cells and purified prior to immobilization. Relatively few organisms (*Actinoplanes missouriensis, Bacillus coagulans, Streptomyces rubiginosus, S. olivochromogenes, S. murinus, and Microbacterium arborescens*) have been used to generate the glucose isomerase used in commercial operations. Although the equilibrium yield of fructose for the typical operating conditions cited above corresponds to 50–55% on a dry basis (db), the reactors are normally operated to obtain yields of 42–45% db to circumvent the limitations on the rate imposed by the approach to equilibrium and to obtain an economically viable reactor size. The effluent from the reactor is then polished to remove color and salts using a combination of activated carbon and ion exchange resins and then concentrated to ca. 71% (w/w) solids. This 42% HFCS product can be used directly in formulating some food products, but manufacturers of soft drinks who desire to effect complete replacement of sucrose in their formulations require that the 42% HFCS be fractionated (for example, by a continuous chromatographic technique) to obtain a product enriched in fructose and a raffinate enriched in glucose. The latter can then be recycled to the immobilized enzyme reactor (IMER) for further conversion to fructose. One possible mode of operation to produce the three fructose-rich syrups of commercial interest and crystalline fructose is shown in Fig. 3. Data provided by the USDA Economic Research Service indicates that in 2002 U.S. production of HFCS corresponded to 9.3 million short tons (db) of fructose.

Other food-related applications

Swaisgood has reviewed applications of immobilized enzymes in the food industry during the past 40 yr.[9] He discusses not only the HFCS application, but also a variety of others, some of which are no longer employed commercially. Table 3 contains a summary of these applications.

Applications Involving Fine Chemicals–Chiral Synthesis and Chiral Separations

In the mid-1980s researchers in the pharmaceutical industry demonstrated that single enantiomers of pharmaceutical compounds often functioned better as therapeutic agents than racemic mixtures. There are many examples for which one particular chiral form (enantiomer) of a compound demonstrated therapeutic efficacy, while the other chiral form was ineffective or produced deleterious effects. (For example, dextromethorpan is commonly used as a cough suppressant, while its enantiomer, levomethorphan, is a powerful narcotic.) By 2002 annual sales of the top 10 single-enantiomer drugs totaled $34.2 billion.[10]

In the pharmaceutical industry, immobilized enzymes (especially lipases) are used to mediate reactions of two general types: reactions involving prochiral substrates and kinetic resolution of racemates.

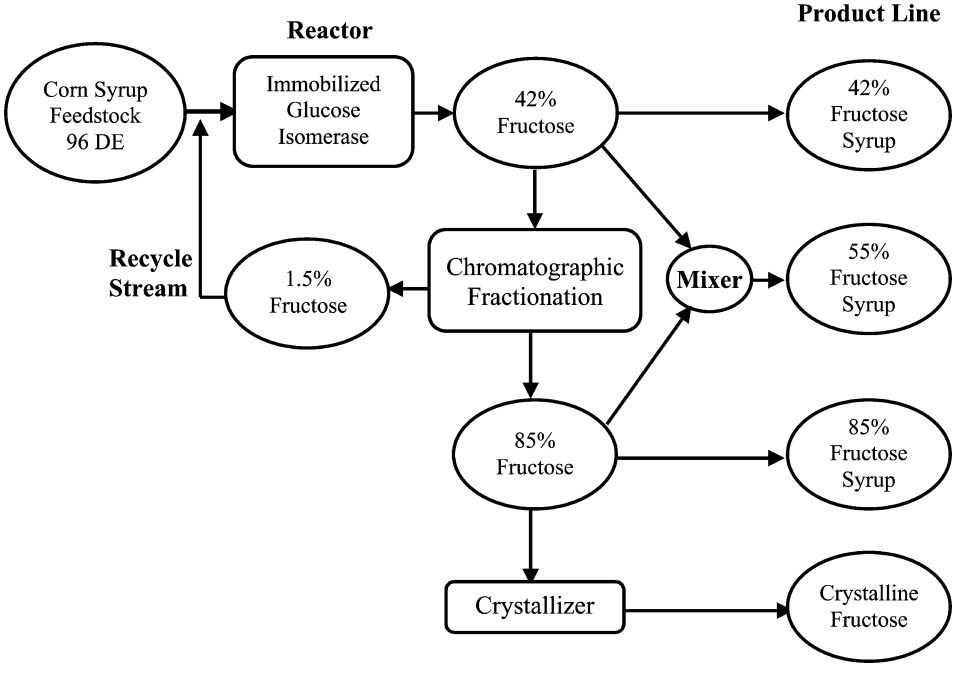


Fig. 3 Flow diagram for production of HFCS products.

Table 3 Applications of immobilized enzymes in the food industry

Product	Immobilized enzyme (application)	Countries	Half-life of enzyme	Productivity (wt/wt of enzyme)
Whey hydrolysate	β-Galactosidase (hydrolysis of lactose)	Finland, Norway, France, and U.K.	20 mo	2000
L-Amino acids	Aminoacylase (resolution of optical isomers)	Japan	65 days	n.a.
L-Phenylalanine	Transaminases (production of aspartame)	Japan	>8 mo	n.a.
5′-Ribonucleotides	5′-Phosphodiesterase (chemical synthesis)	Germany	>500 days	n.a.
Isomaltulose	Isomaltulose synthase (chemical conversion)	U.K., Germany, Japan	1 yr	1500
Invert sugar	Invertase (hydrolysis of sucrose)	Europe	n. a.	6000
Modified fats and oils	1,3-Specific lipases (migration of acyl groups)	Japan, U.K.	n. a.	1100

n.a., not available.

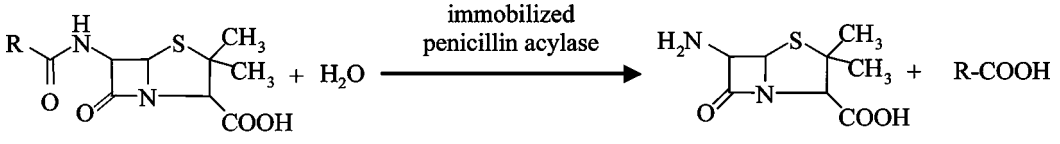
Although prochiral or chiral alcohols and carboxylic acid esters initially served as the primary classes of substrates, compounds susceptible to processing via these two routes now encompass diols, α- and β-hydroxy acids, cyanohydrins, chlorohydrins, diesters, lactones, amines, diamines, amino alcohols, and α-and β-amino acid derivatives. Gotor and Arroyo have reviewed the use of biocatalysts for the preparation of pharmaceutical intermediates and fine chemicals.[11,12] Some specific examples are indicated below.

Penicillins and cephalosporins are characterized by β-lactam structures and are the antibiotics that have traditionally been those most commonly used in the treatment of infections. Pharmaceutical companies have synthesized a variety of semisynthetic β-lactam compounds for use as oral antibiotics, for example, ampicillin and amoxicillin. These penicillin derivatives are prepared by acylation of 6-amino-penicillanic acid (6-APA) derived from penicillin G (benzyl penicillin) or penicillin V (phenoxymethyl penicillin). An immobilized penicillin amidase (penicillin acylase) from *Escherichia coli* or *Bacillus megaterium* is used to prepare the 6-APA in nearly quantitative yield (Fig. 4). This substance is used as the starting material for the production of a number of other penicillins. The immobilized enzyme can be reused more than 600 times in the batch reactor used to accomplish this transformation.

A similar reaction scheme can be used to produce derivatives of cephalosporin via acylation of 7-amino-cephalosporamic acid or 7-amino-desacetoxycephalosporamic acid. These compounds could be produced from a natural cephalosporin using an immobilized cephalosporin acylase, but alternative routes to these compounds are more cost-effective.

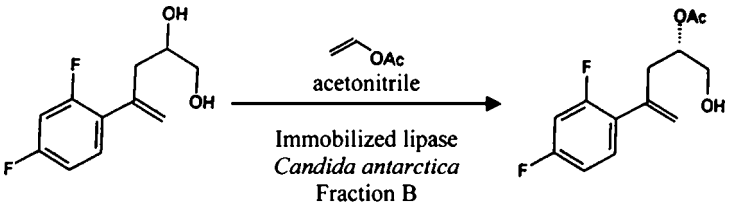
Immobilized forms or reticulated crystals of *Candida antarctica* lipase are effective biocatalysts for the synthesis of pure enantiomers utilized as anti-inflammatory agents. For example, one route for production of the *S*-isomers of 2-aryl propionic acids (ibuprofen, naproxen, ketoprofen, and flurbprofen) involves enantioselective hydrolysis of the corresponding racemic esters. Arroyo has indicated that an immobilized form of *C. antarctica* lipase (fraction B) is used to mediate the selective acetylation of a diol to form the *S*-enantiomer of a monoacetate (Fig. 5), which is further processed to obtain an antifungal agent.[12]

Analytical chromatographic separation of enantiomeric acids and N-substituted amino acids has been effected using immobilized α-chymotrypsin supported on activated silica.[13] Similarly, cellulase immobilized on silica gel can bring about chromatographic



natural penicillin 6-amino penicillanic acid

Fig. 4 Conversion of natural penicillin to 6-APA over an immobilized penicillin acylase.



diol S-monoacetate

Fig. 5 Generation of the *S*-enantiomer of the monoacetate precursor of an antifungal agent from a diol as mediated by an immobilized lipase from *Candida antarctica*, fraction B.

separation of enantiomers of propanolol.[14] This approach can be employed in industrial separation of racemic mixtures using simulated moving-bed chromatography as in the production of naproxen, warfarin, propanolol, and ephedrine.

Biosensors and IMERs

A biosensor is a sensing device consisting of a recognition element (enzyme or cell) in intimate contact with an appropriate transducer that is able to convert the concentration of a species involved in the recognition reaction into a measurable electronic signal. Biosensors based on immobilized enzymes have been employed for decades in analyses for solutes present in either aqueous solution or biological fluids. Fabrication of these biosensors frequently involves immobilization of an enzyme on a membrane electrode that is capable of donating electrons to (or accepting electrons from) species participating in an enzyme-mediated reaction. The substrate to be quantified diffuses to this surface where the biocatalytic reaction occurs. Electrochemical (potentiometric or amperometric) measurements are then employed to monitor changes in the concentration of the analyte of interest. In many applications, the transducer element is located downstream from an IMER that converts the substrate into a chemical form that stimulates the response of the transducer.

Analytical applications

Schuhmann has reviewed amperometric biosensors and indicated that these sensors can be categorized as devices employing: 1) direct electron transfer between redox proteins and electrodes modified with self-assembled monolayers; 2) anisotropic orientation of redox proteins at monolayer-modified electrodes; 3) electron transfer cascades via redox hydrogels; and 4) electron transfer via conducting polymers.[15] These biosensors are employed to quantify the concentrations of a wide variety of substrates. For these determinations, the choice of the support material and the method of immobilization are tailored to the particular application. Gupta and Mattiasson have described several unique applications of enzymes in bioanalytical systems.[16]

Physical entrapment of an enzyme in a gel or polymeric matrix is often employed in analytical applications of biosensors because much of the original activity of the free enzyme is retained. However, this approach is limited to relatively small analytes that can readily penetrate the solid matrix. For in vivo measurements, heparin may also be employed to coat the sensor to create a biocompatible nonthrombogenic surface. A technique for reversible binding of enzymes in biosensors involves use of antibodies raised against enzymes (antigens). For reversible immobilization of glycoenzymes one can utilize lectins (carbohydrate binding proteins). Novel protein structures (sequences) including one partner of an affinity pair can be constructed using recombinant DNA technology. These tagged enzymes can subsequently be immobilized to complementary ligands bound to the support for use in enzyme purification and in ELISA.

Analytical protocols based on the use of immobilized enzyme-mediated reactions that are highly specific to the analyte require little, if any, manipulation of the sample (extraction, addition of reagents, dialysis, filtration, etc.). The sensing element is normally reusable and has the great advantage that it does not require consumption of reagents during the reaction. Thus, a single biosensor may often be utilized to make hundreds or thousands of measurements so that the resulting cost per assay is low. In other cases, the biosensor involves a dipstick that undergoes a color change as the enzymatic reaction proceeds.

The most widely employed types of biosensors are those that employ an oxidase to generate hydrogen peroxide. A classic example is the electrode containing an immobilized glucose oxidase that generates an amperometric signal related to the concentration of glucose present in the sample. However, biosensors of these types are often susceptible to interference from other electrochemically active solutes in the sample. A wide variety of techniques have been developed to circumvent or minimize this problem, for example, application of a semipermeable membrane above the

enzyme matrix, use of enzyme field effect transistors, etc. For analytes that are not susceptible to quantitation by simple electrochemical methods, it may be necessary to employ coupled enzyme reactions or optical modes of detection, perhaps taking advantage of optical fibers. Other biosensors may utilize field effect transistors, thermistors, bioluminescence, or chemiluminescence for detection of the analyte.

The IMER approach does not require that the enzyme be placed in close proximity to the detector if the transducer signal is generated by a soluble product or cosubstrate of the enzymatic reaction. In the latter case, a variety of flow systems and postreactor detectors can be utilized to produce simultaneous determinations of the concentrations of several analytes. For example, an IMER can be combined with a high-performance liquid chromatography (HPLC) instrument (perhaps also in combination with mass spectroscopy) for purposes of both qualitative and quantitative analysis. The chemo-, stereo-, and regioselectivities of enzymes facilitate separation and/or identification of analytes that may be present as different isomers (e.g., in peptide analysis based on use of peptidase IMERs in combination with these techniques to obtain structural information about the sequence of amino acids in peptides).

In some cases, coimmobilization of multiple enzymes is required, especially when successive use of two or more enzymes is employed to convert the original substrate to a final product that is easier to detect or that permits one to drive an unfavorable equilibrium situation to completion by consumption of the initial product in a detection reaction yielding the product that serves as the actual analyte.

Two important general categories of biosensor applications are discussed below.

Detection and Analysis of Sugars. Because the refractive index detectors used in HPLC are not very selective and are characterized by low sensitivities for different sugars and their various isomers, reactors containing immobilized enzymes are often coupled to the corresponding chromatographic columns to obtain significant improvements in selectivity and sensitivity. Examples include:

1. Separation and determination of glucose and lactose in a penicillin fermentation broth using a glucose dehydrogenase (GDH) to catalyze oxidation of the β-anomeric form of aldoses to lactones in the presence of NAD^+. The analysis is based on amperometric detection of the NADH formed. The enzymatic reactor is packed with silanized porous glass beads activated with glutaraldehyde as a support for immobilization of GDH from *Bacillus megaterium*.[17]
2. Determination and quantification of sugars in the discharge from the sulfite pulping of lignocellulose using coimmobilization of a mutarotase (MT), from porcine kidney that converts all of the aldoses to their active β-anomers, with a xylose isomerase from *Streptomyces* sp. (that produces the detectable ketoses), a galactose dehydrogenase (from a recombinant *Escherichia coli*), and GDH from *Bacillus megaterium*.[18]
3. Detection of glucose and malto-oligomers (G2–G10) in corn syrup. This assay employs the selective hydrolysis of α-(1,4) and α-(1,6) linkages of malto-oligosaccharides to glucose by GA with amperometric detection at a gold electrode. Glucoamylase from *Aspergillus niger* is supported on Nucleosil 300 previously silanized with (γ-glycidoxypropyl) trimethoxysilane and activated with 1,1′-carbonyldiimidazole.[19]
4. Detection of oligosaccharides (e.g., stachyose, raffinose, sucrose, and fructose) in a soybean extract using invertase hydrolysis of β-D-fructofructoside to fructose, and further oxidation of this sugar by hexacyanoferrate (III) ion in the presence of fructose dehydrogenase (FDH). This analysis is based on a coimmobilization of invertase from *Candida utilis* and FDH from *Gluconobacter* on poly(vinyl alcohol) (PVA) beads and coulometric quantification of the hexacyanoferrate(II) ions formed.
5. Analysis of malto-oligosaccharides using an IMER containing GA, MT, and GDH coimmobilized on an aminated porous silica matrix. This analysis has been used for quantification of sugars in soft drinks, beer, and fermentation broth containing *Penicillium* and *Fusarium sysporum s.*

Analysis of Amino Acids. Differentiation of the L- and D-forms of amino acids is essential because they differ in their biological and physiological properties. Although chromatographic columns that effect separations of chiral compounds could be used for analysis of solutions of these acids, a combination of a reactor containing a stereoselective immobilized enzyme and a chromatographic system provides the necessary selectivity for such analyses.

1. Immobilized L-amino acid oxidase catalyzes the oxidation of L-amino acids to 2-oxo acids. Detection of the hydrogen peroxide product is accomplished using a fluorometric or chemiluminescent assay subsequent to the reaction in the packed bed. L-Amino acids can also be quantified amperometrically using a platinum electrode Ag/AgCl on which the enzyme is

immobilized. L-Lys, L-His, L-Cys, L-Arg, L-Met, L-Leu, L-Ile, L-Tyr, L-Phe, and L-Trp can be quantified by this procedure.

2. Branched-chain L-amino acids can be analyzed using leucine dehydrogenase immobilized on aminated PVA activated with glutaraldehyde. This enzyme catalyzes the deamination of L-Leu, L-Ile, and L-Val to 2-oxo acids in the presence of NAD^+.

There are many instances where it is helpful to use biosensors for the detection of contaminants and monitoring of air and water quality. In some cases, appropriate biosensors exist. Generally speaking, development of such biosensors is a demanding task because of the wide range of potential substrates (including such hazardous substances as chemical and biological warfare agents) and the associated problems of interference effects, the necessity for unattended operation, and the need for robust sensors in harsh environments. Analyses for aldehydes produced in industrial plants, incinerators, automobile exhausts, or foodstuffs (aroma and storage controls) are based on oxidation of aldehydes to carboxylic acids by immobilized aldehyde dehydrogenase in the presence of a thiol, K^+, and NAD^+. Determinations of some metals (e.g., zinc) via IMERs are sensitive, specific, fast (minutes), and do not require chromatographic separations.

Medical/Clinical Applications

Liang, Li, and Yang have reviewed biomedical applications of immobilized enzymes with emphasis on the use of biosensors for the diagnosis of disease states.[20] Electrodes containing immobilized enzymes constitute the primary technology used in this application. Table 4, adapted from Liang, Li, and Yang, contains a summary of applications of enzyme-based sensors in clinical diagnosis.[20]

Diagnosis of renal problems, xanthinuria, and toxemia of pregnancy via determination of the ratio of hypoxanthine to xanthine in plasma is facilitated by the use of biosensors. Xanthine oxidase immobilized on aminopropyl-CPG (controlled pore glass) activated with glutaraldehyde oxidizes hypoxanthine first to xanthine and then to uric acid. Use of an IMER with biosensors for hypoxanthine, xanthine, and uric acid provides the necessary data. Pre- or postcolumn enzymatic reactions catalyzed by creatinine deiminase, urease, alkaline phosphatase, ATPase, inorganic pyrophosphatase, or arylsufatase facilitate analysis of uremic toxins (simultaneous detection of electrolytes, serum urea, uric acid, creatinine, and methylguanidine).

Analysis of steroid hormones for control of endocrine functions by radioinmmunoassay fails for homologous steroids. Instead, one can employ highly specific steroid dehydrogenases to mediate oxidation of the hydroxy functions of hydroxysteroids. Then, highly sensitive fluorescence detection is used to determine specific positions of OH groups and α and β configurations with HPLC systems combined with an appropriate postcolumn enzymatic reactor. Similarly, one can determine serum bile acid concentrations as an indicator of liver disease by combining an IMER containing 3α-hydroxysteroid dehydrogenase with gas or liquid chromatography.

In addition to the steroid hormones, the steroids utilized for pharmaceutical applications can be selectively determined using appropriate immobilized enzymes in an IMER. Use of α-, β-, and/or stereoespecific dehydrogenases permits one to enhance the selective

Table 4 Applications of biosensors containing immobilized enzymes for clinical analyses

Substrate (analyte)	Immobilized enzyme(s)	Linear range (approximate)
Glucose	Glucose oxidase/GDH	50 mM
Lactate	Lactate oxidase	27 mM
Oxalate	Oxalate oxidase	1 mM
Urea	Urease	100 mM
Glutamate	Glutamate oxidase	200 μM
Carnitine	Carnitine dehydrogenase and diaphorase	1 nM
Theophylline	Theophylline oxidase	30 μM
Creatine and creatinine	Creatininase, creatinase, and sarcosine oxidase	30 mM
Cholesterol	Cholesterol oxidase	3 mM
Amino acids	Amino acid oxidase	10 mM
Acetylcholine and choline	Acetylcholine esterase and choline oxidase	100 μM
Bilirubin	Hemoglobin and glucose oxidase	
γ-Aminobutyric acid	Catalase and γ-glutamate oxidase	10 nM

detection and chiral resolution of steroids. Detection of anabolic steroids in urine by these methods is employed in sports medicine as a means of controlling illicit use of these drugs by athletes.

Flow injection analysis (FIA) for ethanol can be utilized for blood alcohol determination in drunken driving situations.

One can also envision applications involving the use of encapsulated or entrapped enzymes in bioreactors for therapeutic applications involving detoxification of deleterious substances or correction of metabolic deficiencies. In these applications, the enzymes could be contained within artificial cells [e.g., modified red blood cells (erythrocytes) or liposomes]. Liang, Li, and Yang have reviewed biomedical applications of immobilized enzyme bioreactors.[20]

SUMMARY

Techniques for immobilization of enzymes and applications of immobilized enzymes are discussed.

Both chemical and physical methods may be used to immobilize biocatalysts while retaining or modifying their activity, selectivity, or stability. Among the techniques used for immobilization of enzymes are physical adsorption, covalent bonding, ionic binding, chelation, cross-linking, physical entrapment, microencapsulation, and retention in permselective membrane reactors. The mode of immobilization employed for a particular application depends not only on the specific choice of enzyme and support, but also on the constraints imposed by the microenvironment associated with the application.

Commercial uses of immobilized enzyme technology are limited in scope, but encompass industrial production of HFCS, biosensors, clinical diagnostic procedures, chemical analyses, chiral syntheses, and therapeutic applications.

CONCLUSIONS

Immobilized enzyme technology is not a stagnant technology. It has evolved in recent decades to the point where it can be employed for select industrial processes and, more importantly, for rapid analyses of significant import in both clinical and analytical situations. The analytical applications also have important implications for monitoring air and water quality in support of environmental regulations, as well as in analyses of process effluents and industrial wastewaters. Biosensors based on immobilized enzyme technology offer significant commercial potential for detecting food spoilage and chemical and biological warfare agents, as well as for monitoring food storage conditions.

One of the areas in which immobilized enzymes are expected to have major impact is in the biocatalysis of reactions in organic media, for example, in the synthesis of chiral compounds as intermediates in the manufacture of pharmaceuticals and in the modification of naturally occurring fats and oils to produce value-added products targeted at the nutraceuticals market.

Use of recombinant DNA technology, other means of genetic engineering, and enzymes obtained from thermophilic and halophilic organisms can be expected to produce novel enzymes with enhanced selectivity, activity, or stability. These novel enzymes may be utilized to effect reactions at elevated temperature, in organic media, and in other harsh environments where stringent requirements must be met. In addition, further advances in permselective membrane technology and/or affinity separation media may facilitate improvements in the biosensors employed in analytical applications as a result of immobilization of both enzymes and cofactors in a manner that minimizes leakage problems.

Advances in the understanding of structure–activity/selectivity relations for enzymes evolving from the use of x-ray, NMR, and other instrumental methods for characterization of enzyme structures should contribute to the development of improved immobilized enzyme systems for both analytical and industrial applications. Immobilized enzyme technology has enormous potential, but significant advances on several fronts are necessary prior to widespread industrial use of this technology. Katchalski-Katzir has discussed this problem in a review of past successes and failures in efforts to employ immobilized enzymes in the food, pharmaceutical, and chemicals industries.[21]

From an industrial perspective, economic considerations are paramount when deciding whether or not to adopt a new technology. A crucial consideration for future development of large-scale industrial processes employing immobilized enzyme technology is that they be cost-effective relative to alternative technologies. To date, immobilized enzyme technology has, in most instances, failed to pass the test of economic viability for chemical, pharmaceutical, and food processing applications. On the other hand, this technology has found increasing numbers of applications in clinical and analytical applications, where it proves to be cost-effective. Indications are that in the future, novel and better commercial IMERs will be utilized more extensively and routinely in FIA systems in analytical applications.

ACKNOWLEDGMENTS

A sabbatical grant from the Spanish Ministerio de Educacion, Cultura, y Deportes and financial support

from the U.S. National Science Foundation (BES-00 77524) for Charles G. Hill, Jr. are gratefully acknowledged.

REFERENCES

1. Kennedy, J.F.; Cabral, J.M.S. Enzyme immobilization. In *Biotechnology, Vol. 7a, Enzyme Technology*; Rehm, H.J., Reed, G., Eds.; VCH: Weinheim, Germany, 1987; 347–404.
2. Boller, T.; Meier, C.; Menzler, S. EUPERGIT oxirane acrylic beads; how to make enzymes fit for biocatalysis. Org. Process Res. Dev. **2002**, *6*, 509–519.
3. Bickerstaff, G.F. Immobilization of enzymes and cells. In *Methods in Biotechnology*; Humana Press: Totowa, NJ, 1997; Vol. 1.
4. Scouten, W.H. A survey of enzyme coupling techniques. Methods Enzymol. **1987**, *135*, 30–65.
5. Woodward, J. Immobilized enzymes: adsorption and covalent coupling. In *Immobilized Cells and Enzymes: A Practical Approach*; Woodward, J., Ed.; JRL: Oxford, 1985; 3–17.
6. Chang, T.M.S. Microencapsulation of enzymes and biologicals. In *Methods in Enzymology*; Mosbach, K., Ed.; Academic Press: New York, 1976; Vol. 44, 201–218.
7. Kennedy, J.F.; Cabral, J.M.S. Enzyme immobilization. In *Biotechnology, Vol. 7a, Enzyme Technology*; Rehm, H.J., Reed, G., Eds.; VCH: Weinheim, Germany, 1987; 393.
8. Schenck, F.W. High fructose syrups—a review. Int. Sugar J. **2000**, *102*, 285–288.
9. Swaisgood, H.E. Use of immobilized enzymes in the food industry. In *Handbook of Food Enzymology*; Whitaker, J.R., Voragen, A.G.J., Wong, D.W.S., Eds.; Marcel Dekker, Inc.: New York, 2002; 359–366.
10. Rouhi, A.M. Chiral business. Chem. Eng. News **2003**, *81* (18), 45–55.
11. Gotor, V. Biocatalysis applied for the preparation of pharmaceuticals. Org. Proc. Res. Dev. **2002**, *6*, 420–426.
12. Arroyo, M. Empleo de biocatalizadores en la sintesis de compuestos de interes farmaceutico. Rev. R. Acad. Ciens. Exact. Fis. Nat. (Esp.) **2000**, *94*, 131–142.
13. Marle, I.; Karlsson, A.; Pettersson, C. Separation of enantiomers using α-chymotrypsin-silica as a chiral stationary phase. J. Chromatogr. **1992**, *604*, 185–196.
14. Fornstedt, T.; Sajonz, P.; Guiochon, G. Thermodynamic study of an unusual chiral separation. Propanolol enantiomers on an immobilized cellulase. J. Am. Chem. Soc. **1997**, *119*, 1254–1264.
15. Schuhmann, W. Amperometric enzyme biosensors based on optimized electron-transfer pathways and non-manual immobilization procedures. Rev. Mol. Biotechnol. **2002**, *82*, 425–441.
16. Gupta, M.N.; Mattiasson, B. Unique applications of immobilized proteins in bioanalytical systems. Bioanal. Appl. Enzymes **1992**, *36*, 1–34.
17. Marko-Varga, G. High-performance liquid chromatographic separation of some mono- and disaccharides with detection by a post-column enzyme reactor and a chemically modified electrode. J. Chromatogr. **1987**, *408*, 157–170.
18. Marko-Varga, G.; Dominguez, E.; Hahn-Hagerdahl, B.; Gorton, L. Selective post-column liquid chromatographic determination of sugars in spent sulfite liquor with two enzymic electrochemical detectors in parallel. J. Chromatogr. **1990**, *506*, 423–441.
19. Larew, L.A.; Johnson, D.C. Quantitation of chromatographically separated maltooligosaccharides with a single calibration curve using a postcolumn enzyme reactor and pulsed amperometric detection. Anal. Chem. **1988**, *60*, 1867–1872.
20. Liang, J.F.; Li, Y.T.; Yang, V.C. Biomedical application of immobilized enzymes. J. Pharm. Sci. **2000**, *89*, 979–990.
21. Katchalski-Katzir, E. Immobilized enzymes—learning from past successes and failures. Tibtech **1993**, *11*, 471–478.

BIBLIOGRAPHY

Godfrey, T., West, S., Eds.; *Industrial Enzymology*, 2nd Ed.; Macmillan Press: London, 1996.

Kress-Rogers, E., Ed.; *Handbook of Biosensors and Electronic Noses, Medicine, Food, and the Environment*; CRC Press: Boca Raton, FL, 1997.

Lam, S., Mallikin, G., Eds.; *Analytical Applications of Immobilized Enzyme Reactors*; Blackie Academic and Professional: Glasgow, U.K., 1994.

Liese, A.; Seelbach, K.; Wandrey, C. *Industrial Biotransformations*; John Wiley & Sons–VCH: Weinheim, Germany, 2000.

Rehm, H.-J.; Reed, G. *Biotechnology: A Comprehensive Treatise in 8 Volumes*; Kennedy, J.F., Ed.; VCH Verlagsgesellschaft mbH: Weinheim, Germany, 1987; Vol. 7a.

Schmid, A.; Dordick, J.S.; Hauer, B.; Kieners, A.; Wubbolts, M.; Witholt, B. Industrial biocatalysis today and tomorrow. Nature **2001**, *409*, 258–268.

Uhlig, H. *Industrial Enzymes and Their Applications*; Linsmaier-Bednar, E.M., Ed.; John Wiley & Sons: New York, 1998.

Incineration and Combustion

I

Selim M. Senkan
Department of Chemical Engineering, University of California, Los Angeles, California, U.S.A.

INTRODUCTION

Incineration is the thermal treatment of wastes generated as by-products of our technological society, and as such represents an important application of combustion.[1] Incineration not only reduces the volume of waste generated, but also results in the detoxication, decontamination, and sterilization of waste, while allowing the recovery of substantial amounts of energy and to a lesser extent materials. Although the prevention of waste generation in the first place is the most desirable option, our societal concerns for safety, security, and health, as well as simple processing inefficiencies inevitably results in the creation of some waste, which must be dealt with efficiently and effectively. Incineration provides an effective and efficient solution to these waste products.

The primary thermodynamic products of waste incineration, like any combustion process, are CO_2 and H_2O. When wastes contain chlorine and sulfur compounds, their combustion products include HCl and SO_x.[2] Incinerator exhausts may also contain unburned principle organic hazardous compounds (POHC) that was fed, particulate matter (PM), such as soot and fly ash; oxides of nitrogen (NO_x); products of incomplete combustion (PIC) that include carbon monoxide (CO), volatile organic compounds (VOC) such as methane, acetylene, 1,3-butadiene, and benzene; polycyclic aromatic hydrocarbons (PAH) such as anthracene, pyrene, and benzo(a)pyrene; chlorinated aromatics, including polychlorinated di-benzo dioxins (PCDD) and furans (PCDF); and metals, such as lead, arsenic, and mercury, and their oxides. Incineration also produces a residue (bottom ash) of incombustible and partially combusted waste that must be properly disposed in a secured landfill. As a result of several unfortunate and highly publicized incidents involving dioxin exposure in the past, little public support exists today for the broad utilization of incineration. This is regrettable because steadily increasing waste generation coupled with rapidly diminishing secured landfills call for the larger scale implementation of the versatile incineration technology.

One of the most extensive fundamental information sources on combustion and incineration is the *Proceedings of the International Symposium on Combustion*, which is regularly published by the Combustion Institute. In addition, selected books listed at the end of this entry provide information on combustion processes. In particular, the books by the National Academy of Engineering, *Waste Incineration and Public Health,*[3] *Pollutants from Combustion*: *Formation and Impact on Atmospheric Chemistry*,[4] and *Gas-Phase Combustion Chemistry*,[5] represent some of the more recent compilations. Information on federal government regulations pertaining incinerators and emissions from these devices can be found at the US Environmental Protection Agency (EPA) internet site http://www.epa.gov/epaoswer/hazwaste/combust/newmact/hazmact.htm. This web site provides information both on incinerator standards as well as the data EPA used to establish these standards. The industrial perspective on government regulations as well as technological issues facing incinerators can be obtained from the internet site of the Coalition for Responsible Waste Incineration (CRWI, www.crwi.org).

WASTE GENERATION

The waste generated can be categorized into three groups:

1. Municipal solid waste, defined to be nonhazardous, is widely generated by households, commercial establishments such as restaurants, public, private, and government facilities.
2. Hazardous waste, which is defined under the EPA's Resource Conservation and Recovery Act (RCRA) as potentially dangerous to human health or the environment on the basis of being flammable, toxic, corrosive, or reactive, are generated by manufacturing companies, universities, hospitals, government facilities, as well as to a smaller extent by households.
3. Medical waste is separately categorized because of its infectious or toxic characteristics, and is generated primarily by hospitals, medical laboratories, and offices.

As shown in Fig. 1, the amount of municipal solid waste generated in the United States steadily increased

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120037169

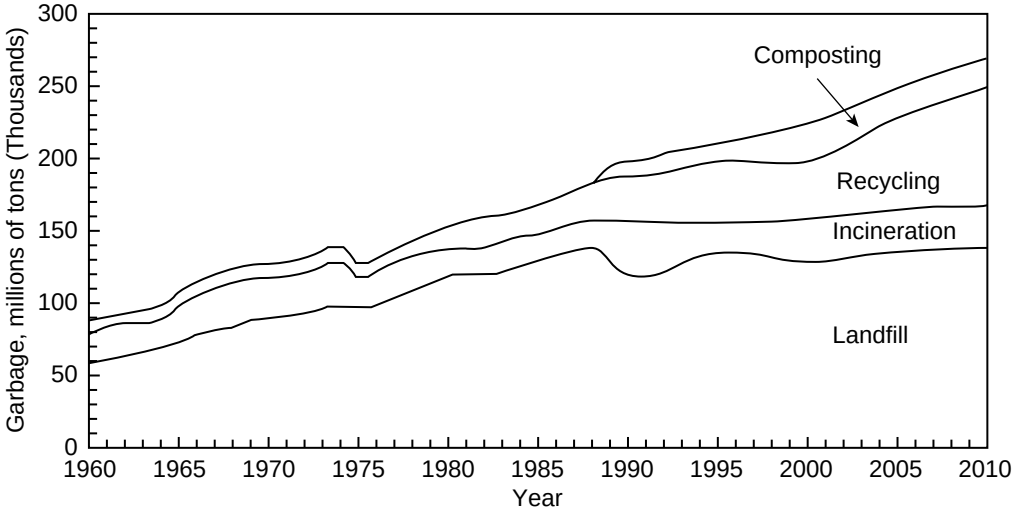


Fig. 1 Trends in municipal solid-waste generation and management in the United States, 1960–2010. (From Ref.[3].)

over the past four decades in spite of incentives for source reduction and recycling. Nearly, 80% of this waste is combustible (Fig. 2), thus can be disposed of by incineration. Yet, only 10–15% of municipal waste is incinerated (Fig. 1). This is because of: 1) the availability of lower-cost disposal alternatives, such as landfilling; 2) the opposition from local advocacy groups; and 3) the politically popular recycling programs, which decreased the public support for incineration. Clearly, as landfills become unavailable and exporting wastes are no longer feasible, we must rely on incineration technology to provide a more permanent solution to our growing waste disposal problems.

Municipal waste incinerators typically are large, state-of-the-art facilities where volume reduction and power and/or steam generation are often accomplished simultaneously. Owing to the presence of a large combustible fraction, municipal wastes can be incinerated directly with little use of an auxiliary hydrocarbon fuel.

Our estimates of the amount of hazardous wastes generated are less certain because of the inconsistent use of the definition of ''hazardous waste.'' Consequently, it has been difficult to obtain reliable historical data on generation and to determine the amount of waste reduction attained. For example, according

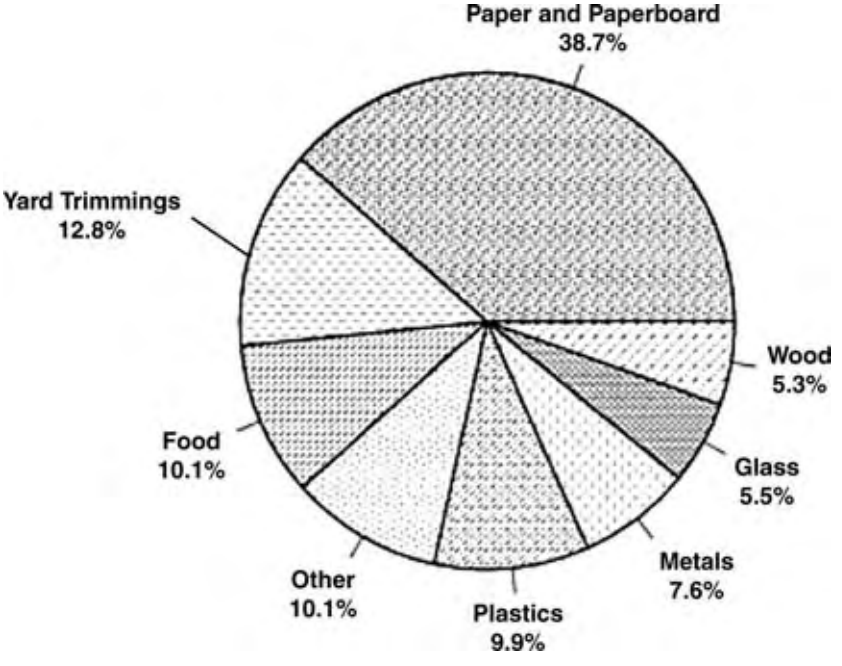


Fig. 2 Municipal solid-waste composition by weight. (From Ref.[3].)

to EPA estimates, over 264 million metric tons of hazardous wastes were managed in the early 1980s.[3] In contrast, the Chemical Manufacturers Association reported over 600 million metric tons of hazardous wastewater and about five million metric tons of hazardous solid waste during the same period.[3]

Rotary kilns (e.g., cement kilns), metal-recovery and smelting furnaces, mobile incinerators, and industrial boilers are primarily used to incinerate hazardous wastes. The obvious benefits of combustion of waste as fuel are the recovery of energy from the waste and the conservation of fossil fuels. Because the kiln and furnace operators are paid to take in the waste, rather than having to pay for fuel, also create economic incentives. Mobile incinerators are most commonly used for soil decontamination projects, and can be moved from site to site once the job is completed.

Medical wastes are generated in quantities significantly lower than the municipal and hazardous wastes, yet their infectious nature causes them to be the focus of significant attention. Typical composition of medical waste is shown in Fig. 3. As over 70% of the medical waste is combustible, incineration again represents an effective management tool. Because medical wastes consistently contain chlorinated plastics, e.g., polyvinyl chloride (PVC) gloves, as well as human tissue and organs, their incineration products can contain significant levels of chlorinated dioxins and furans.[6]

WASTE INCINERATORS

In Table 1, waste generation in the United States, the number of incineration facilities, and the amount of waste combusted are summarized. As seen from this table, the most prevalent incinerators are the medical waste incinerators. Industrial boilers and furnaces were the next most abundant type of incineration facilities, followed by municipal waste incinerators.

HEALTH EFFECTS OF INCINERATION

The adverse health effects of incineration by-products, or any combustion process, can be acute or long term. In Table 2, the acute toxicities of selected PIC are presented together with some other reference materials, where the toxicities are expressed in terms of LD_{50} (Lethal dose that kills 50% of laboratory animals).[7] As evident from this table, tetrachloro di-benzo dioxin (TCDD) has been shown to be extremely toxic with LD_{50} of 0.001 mg/kg, to guinea pigs. However, it is important to note that this extreme toxicity was not seen in other laboratory animals, as well as in humans. The latter was determined as a result of several tragic accidents in which large populations were exposed to dioxins (Seveso, Italy, Times Beach Missouri, U.S.A.). Unfortunately, LD_{50} values do not truly reflect the full adverse impact of a chemical or their mixtures.

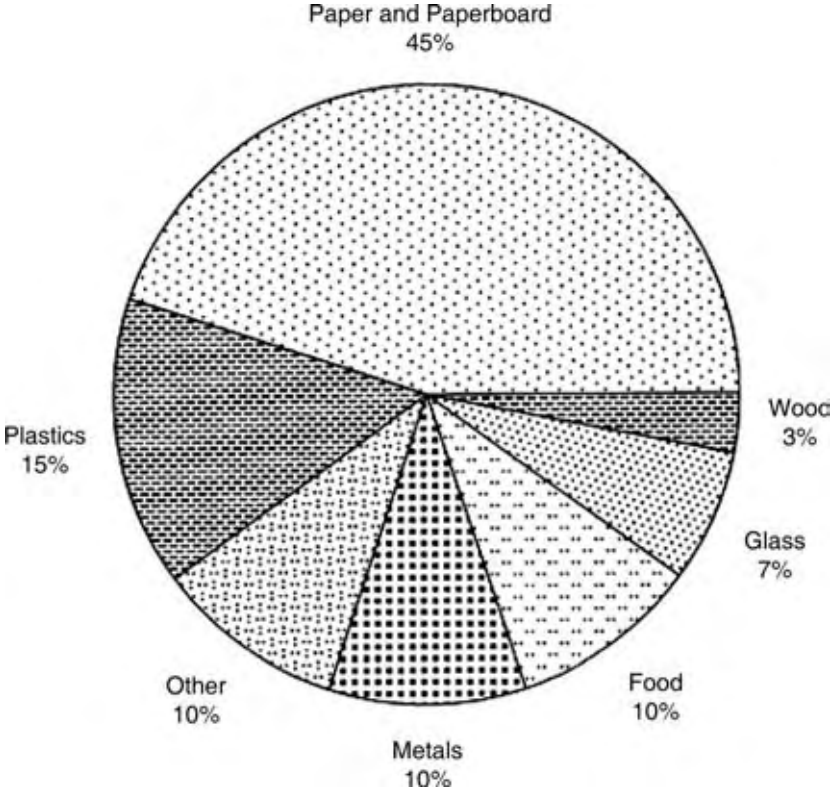


Fig. 3 Typical medical waste composition. (From Ref.[3].)

Table 1 Waste generation in the United States, number of incineration facilities, and amount of waste combusted

Type of waste	Amount generated (million tons/yr)	Number of incineration facilities	Amount of waste combusted (million tons/yr)
Municipal solid waste	209	122	36
Hazardous waste	276		3
On-site incinerators		129	
Commercial incinerators		20	
Industrial boilers and furnaces		950	
Cement kilns		18	
Light weight aggregate kilns		5	
Medical waste		1655	0.8

(From Ref.[3].)

Chronic exposure to some chemicals, such as some isomers of PAH, at doses that do not show any acute toxicities are well known to induce cancer, cause mutations, or are teratogens.[7] Complex mixtures present challenges even far greater than individual molecules.

In Fig. 4, the relative contribution of individual airborne hazardous pollutants to lung cancer rates are presented.[8] The results presented in Fig. 4 were obtained after removing cancer incidents that are directly attributable to tobacco smoke which primarily is an indoor pollutant. As seen from this figure the PIC (which includes PAH), 1,3-butadiene, benzene, and formaldehyde, which are produced by all combustion/incineration processes, were the largest contributors to lung cancer, representing in excess of 50% of the risk, far more than dioxins. However, because different chemicals target different parts of the body, it is important to consider other organs and types of cancer in assessing the full impact of chemicals on human health.[7]

Table 2 Approximate acute LD_{50} values for some chemicals for laboratory animals

Chemical	LD_{50} (mg/kg)
Ethanol	10,000
Sodium chloride	4000
Benzene	930
Phenobarbital sodium	150
Formaldehyde	100
Picrotoxin	5
Strychnine sulfate	2
Nicotine, arsenic	1
D-Tubocurarine	0.5
Hemicholinium-3	0.2
Tetrodotoxin, mercury	0.1
TCDD	0.001
Botulinum toxin	0.00001

INCINERATION TECHNOLOGY

Today waste incinerators can be designed and operated to achieve better than 99.999% destruction and removal efficiencies (DRE) of the waste compounds and to only emit extremely low concentrations of the pollutants of concern under normal operating conditions. In addition, operator training can help ensure that an incinerator facility is run at its peak combustion efficiency and the emission-control devices are operated optimally for maximum pollutant capture or neutralization. However, as with all types of facilities, there exist off-normal (upset) operating conditions that might result in the temporary increase in the emissions of pollutants. Such conditions occur during incinerator startup or shutdown or when the composition of the waste being burned changes sharply. Upset conditions can also be caused by malfunctioning equipment, operator error, or inadequate maintenance. Continuous monitoring of the combustion chamber and incinerator emissions can provide useful feedback and help control the incinerator operations.

Proper design and operation of an incinerator require attention to *temperature*, *turbulence* of the mixture being combusted, and residence *time* at the incineration temperature, generally referred to as the 3*T*s of combustion. To achieve efficient combustion, every part of the waste stream must reach an adequately high temperature for a sufficient period of time, and there must be adequate mixture of waste and oxygen. Cool spots can occur next to the furnace's walls where heat is extracted such as in boiler-type furnaces. Cold spots are less likely in refractory lined furnaces, such as in cement kilns.

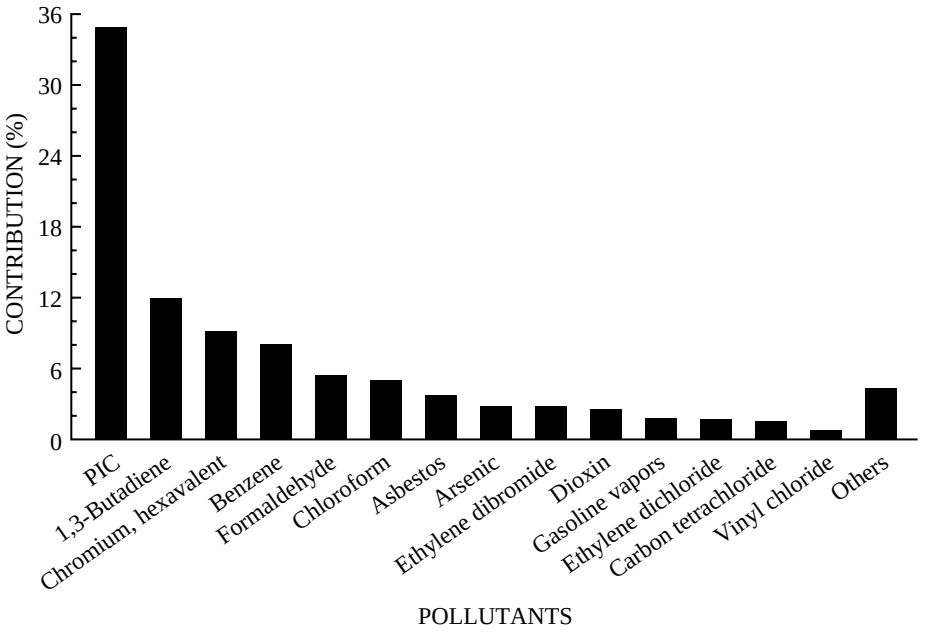


Fig. 4 Relative contribution of various air pollutants to lung cancer. (From Ref.[8].)

The temperature achieved in incineration is the result of heat released by the oxidation process, and has to be maintained high enough to ensure that combustion goes to completion, but not so high as to damage equipment or generate excessive thermal NO_x and metal emissions. Typically, temperatures are controlled by limiting the amount of material charged to the furnace to ensure that the heat-release rate is in the desired range, and then adjusting the resulting conditions by varying the amount of excess air.

Turbulence is needed to provide adequate contact between the waste and oxygen across the combustion chamber (macroscale mixing). Nevertheless, in the final stage, combustion at the molecular level occurs via diffusion or premixed flames, although the former is the dominant mode of waste destruction in practical systems.

Normally, incinerator exhausts are monitored for temperature, CO, NO_x, and O_2. Data covering the entire spectrum of emissions are often collected only for short periods of time. More emission information is needed, especially for dioxins and furans, heavy metals, and PM. Limited research is underway for the development of reliable and rugged technologies that will provide the real time, continuous monitoring of the entire spectrum of emissions from incinerators.[9,10] However, their routine use in incinerators as continuous emissions monitors is not practised yet.

A sketch of a modern waste-incineration facility is shown in Fig. 5. Such a facility includes the following:

1. Waste storage and feed preparation.
2. Incineration furnace producing hot gases and a bottom ash residue for disposal.
3. Exhaust gas cooling, frequently involving heat recovery via steam generation.
4. Further treatment of the cooled gases to remove air pollutants and disposal of residuals from this treatment process.
5. Dispersion of the treated exhaust gases (CO_2 and H_2O) to the atmosphere through an induced-draft fan and stack.

The types of incinerators used are listed in Table 3. Municipal solid-waste furnace designs have evolved over the years, with the newer waste incinerators being waste-to-energy plants that produce steam for electric power generation.

The predominant hazardous-waste incinerator designs are liquid-injection furnaces and rotary kilns. Hazardous wastes are also burned in cement kilns, light-weight aggregate kilns, industrial boilers, halogen–acid recovery furnaces, and sulfuric acid regeneration furnaces. Medical wastes are burned in fixed-hearth incinerators, with staged combustion. The smallest medical-waste incinerators are single-chamber, batch-operated devices, most of them employing afterburners.

Fig. 6 is a sketch of a rotary kiln incinerator, illustrating many of the essential phenomena associated with incineration. Rotary kilns are versatile incinerators because of their ability to treat solid, liquid, and gaseous wastes. Solid and liquid wastes are destroyed primarily through establishment of diffusion flames, i.e., flames that surround solid waste particles or droplets or form a flame sheet adjacent to surfaces.

To achieve highest levels of waste destruction, a secondary combustion chamber is often used in

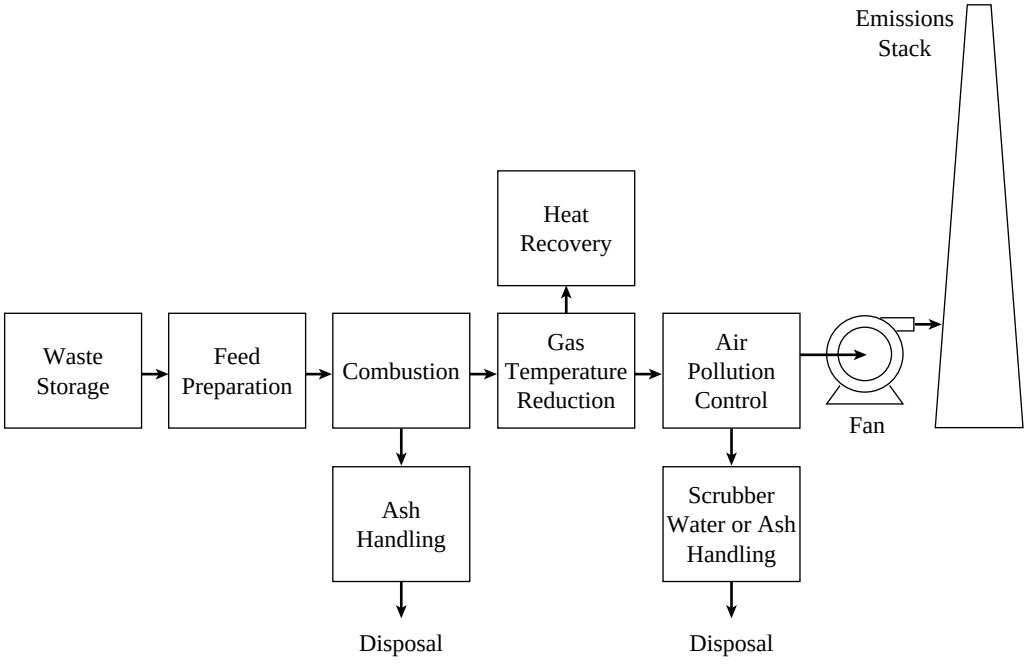


Fig. 5 A schematic of a typical waste incineration facility. Although incinerators have similar components indicated in the figure, combustion chamber designs vary significantly from application to application. (From Ref.[3].)

modern incinerators. The second stage ensures that wastes and by-products that may escape the first stage, for example, because of transient explosions of waste drums or ''puffs,'' are destroyed to achieve the required DRE. Two-stage combustion can also be used to simultaneously achieve waste destruction and NO_x reduction, although the former is the overriding factor in waste incineration. The product gases entraining both organic (i.e., soot) and inorganic (e.g., fly ash) particles are then subjected to a sequence of air pollution control devices (APCD), such as electrostatic precipitators, dust bags, etc. before being discharged

Table 3 Furnace designs used in waste inceneration

Waste type	Furnace design	Type application
Municipal solid waste	Mass burn Waterwall furnace Reciprocating or other continuous moving grate	Most newer municipal-scale facilities
	Mass burn Refractory furnace lining Various grate or stationary hearth designs	Old or small facilities
	Refuse-derived fuel Spreader-stoker/cyclone furnaces	Few facilities in the United States
	Fluidized bed	Foreign applications
Hazardous waste	Liquid injection	Common
	Rotary kiln with secondary combustion chamber	Common
	Fluidized bed	Few in the United States More common for biosludge incinerators
	Fixed hearth with secondary chamber	Mostly with plant trash co-feed
Medical waste	Multiple chamber	Old IIA design for older facilities
	Controlled-air primary chamber with afterburner	Predominant design in the United States since 1970s
	Rotary kiln with afterburner	Few in the United States

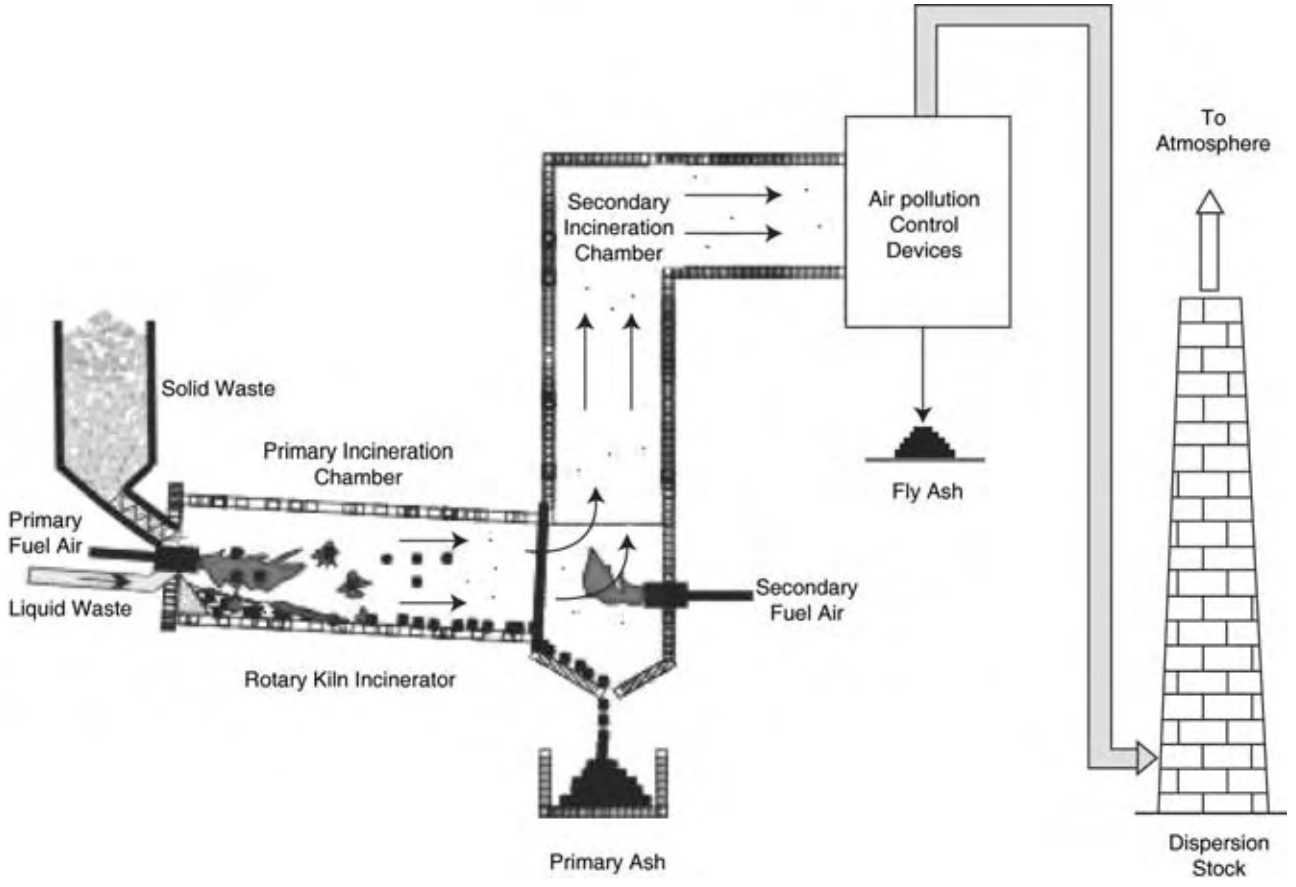


Fig. 6 Schematic of a versatile waste incinerator. Rotary kiln furnaces are preferred because of their applicability to treat different types of wastes, including solid wastes.

into the atmosphere through the stack. The primary ash produced in the main combustion chamber, together with the fly ash collected by the APCD must also be stabilized against leaching and stored in secure landfills.

INCINERATION SCIENCE

Incineration is a rapid, exothermic reaction between a fuel (waste) and oxygen (O_2). Incineration produces the same end products and by-products, whether the material burned is municipal solid waste, hazardous waste, or medical waste. This is because of the fact that complex fuel molecules first undergo thermal decompositions upon being preheated by the preceding flame, forming smaller molecules such as methane, acetylene, ethylene, carbon monoxide, hydrogen, and alike, and it is the combustion of these smaller molecules that primarily influence the nature of combustion products and pollutants formed. This aspect of combustion chemistry has significantly helped in the development of detailed kinetic mechanisms of combustion for all types of hydrocarbon fuels.[5,11,12] Solid decomposition products such as solid carbon (char) and ash also form when the materials being burned lack adequate hydrogen and oxygen or contain metals, respectively.

Premixed and Diffusion Flames

As noted earlier, waste destruction occur as a consequence of premixed and diffusion flames. In premixed flames, the fuel and oxidizer are mixed at the molecular level, and the relative amounts of the reactants are described by the equivalence ratio (ϕ), defined below:

$$\phi = (\text{Fuel amount}/\text{Oxidizer amount})_{\text{actual}}/(\text{Fuel amount}/\text{Oxidizer amount})_{\text{stoichiometric}}$$

Based on this definition, $\phi > 1$ represents fuel-rich and $\phi < 1$ represents fuel-lean conditions. Premixed systems generally are bounded by lower and upper flammability limits, i.e., ϕ, within which self-sustaining flames occur. Flame temperatures (T) first increase and then decrease as a function of increased equivalence

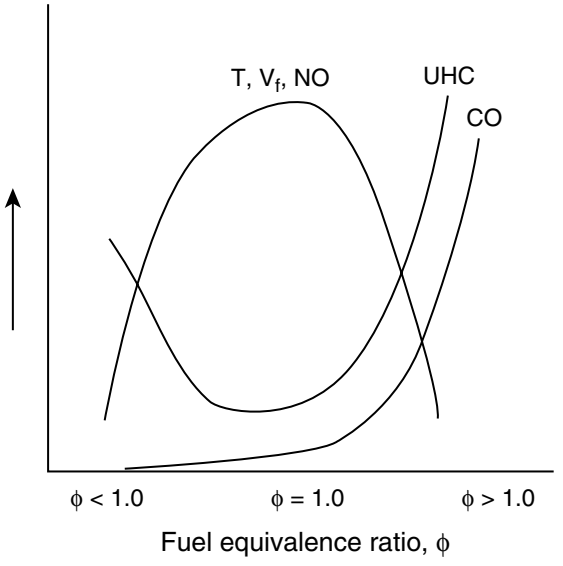


Fig. 7 Effects of equivalence ratio on flame temperature, flame velocity, NO, UHC, and CO.

ratio, exhibiting a maximum around the stoichiometric conditions as shown in Fig. 7. The presence of excess air or excess fuel decreases the maximum attainable flame temperatures. Premixed systems also exhibit a unique combustion wave propagation velocity (V_f), which vary with equivalence ratio similar to temperature. Because increased flame temperature and velocity correspond to increased intensity of combustion, the formation and emission of unburned hydrocarbons (UHC), which include both POHC and PIC, mirror these variables. The NO_x levels produced in flames also follow the temperature pattern because of combined thermodynamic and kinetic reasons as shall be discussed in what follows.

In diffusion flames, combustion reactions occur at an interface to which the fuel and oxygen diffuse from their sources in accordance with the overall combustion stoichiometry. Thus, there is no unique equivalence ratio that describes a diffusion flame. Because a flame is produced at the stoichiometric interface, the maximum flame temperatures reached in diffusion flames are high, e.g., close to adiabatic flame temperatures.

For stable combustion, flame temperatures generally must exceed 1400°C, and typically range approximately between 1500°C and 1900°C or more. The high temperatures preheat the incoming fuel and air through conduction, convection, and radiation. In premixed systems, the extent of air dilution can be used to control the maximum flame temperature. In diffusion flames, however, because stoichiometric concentrations prevail at the flame front, the resulting maximum temperatures are considerably higher, e.g., about 2000°C for natural gas and 2200°C for diesel fuel flames, respectively. Consequently, diffusion flames are easier to stabilize than premixed flames. On the other hand, the higher temperatures associated with diffusion flames promote the formation of certain pollutants, such as NO_x and soot.

In Fig. 8, actual laboratory pictures of premixed and diffusion flames of methane are presented. Flame luminosity increases with increasing equivalence ratio as a consequence of increased production of soot. Increasing luminosity also increases radiative heat transfer, which is desirable to extract as much thermal energy as possible from combustion. Also shown in Fig. 8 are probes to withdraw samples from within the flames, which subsequently can be analyzed by techniques such as gas chromatography, mass spectrometry, or both. Indeed, these types of measurements have been the primary source of information on flame chemistry, contributing significantly to our understanding of the incineration chemistry. For example, we now know that flames can contain hundreds of intermediate species, and detailed chemical kinetic mechanisms must be invoked to describe combustion processes.[13]

Fig. 9 shows concentration profiles of some selected species in a fuel rich, premixed methane flame.[14] Clearly, even the chemistry of combustion of a simple hydrocarbon fuel such as methane is considerably complex. The fuel molecule, before it ultimately produces CO_2 and H_2O, undergoes a series of intermediate

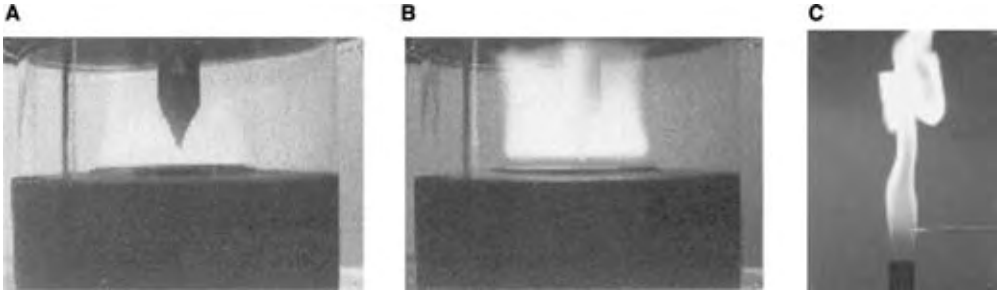


Fig. 8 Laminar premixed flames of methane. (A) Slightly fuel rich; (B) fuel-rich and sooting; and (C) diffusion flame. Note increased luminosity with increasing equivalence ratio.

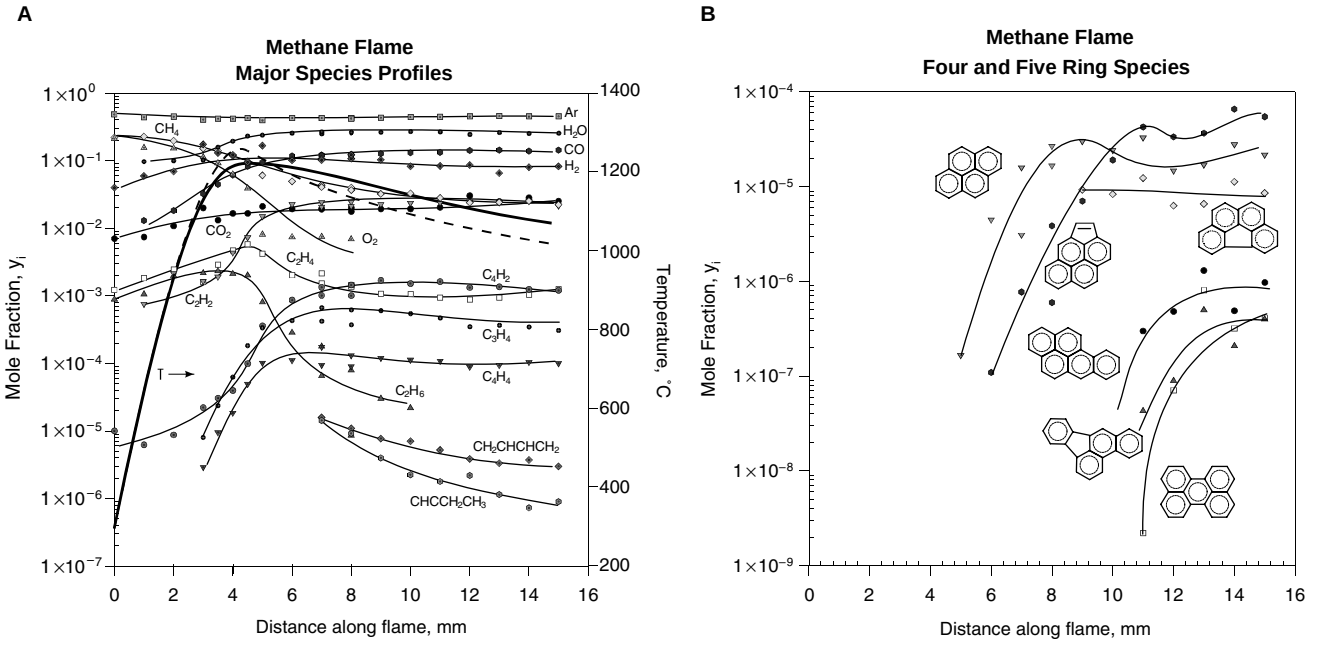


Fig. 9 (A) Major species concentration profiles and (B) Some of the polycyclic aromatic hydrocarbons (PAH) formed in a fuel-rich, premixed laminar methane flame; the formation of a large number of intermediates and by-products are evident. Highly toxic benzo-*a*-pyrene is the 3rd PAH from the bottom. (From Ref.[14].)

reaction steps, which can be described as:

$$CH_4 \rightarrow CH_3^{\bullet} \rightarrow CH_2O \rightarrow CO \rightarrow CO_2$$

(CH3$^{\bullet}$ represents a methyl radical)

where, CO oxidation represents the last sequence in CH_4 combustion, as well as in the combustion of all the hydrocarbons.[11] The main CO oxidation proceeds through its reaction with the OH$^{\bullet}$s:

$$CO + OH^{\bullet} \rightarrow CO_2$$

The OH$^{\bullet}$, together with H$^{\bullet}$ and O$^{\bullet}$, are some of the most important free radicals in flame combustion processes.[15,16] Collectively H$^{\bullet}$, O$^{\bullet}$, and OH$^{\bullet}$ help establish the *radical pool* in flames through the following chain branching reactions:

$$H^{\bullet} + O_2 \rightarrow OH^{\bullet} + O^{\bullet}$$

$$O^{\bullet} + H_2 \rightarrow OH^{\bullet} + H^{\bullet}$$

The interaction of H$^{\bullet}$ radicals with O_2 is particularly significant. At low temperatures, the H$^{\bullet}$ and O_2 combine to form relatively less reactive $HO_2^{\bullet}$. However, at higher temperatures associated with flames, the above chain branching reaction takes over and dominates the combustion process. Because H$^{\bullet}$ plays such a crucial role in flame propagation, its removal from the radical pool has been exploited as a strategy for the development of flame retardants.[17,18]

The above reactions also describe H_2 combustion, an essential submechanism in hydrocarbon combustion. Methane destruction in flames occurs through H$^{\bullet}$, OH$^{\bullet}$, and O$^{\bullet}$ radical attack, which can be shown as:

$$CH_4 + H^{\bullet} \rightarrow CH_3^{\bullet} + H_2$$

$$CH_4 + OH^{\bullet} \rightarrow CH_3 + H_2O$$

$$CH_4 + O^{\bullet} \rightarrow CH_3^{\bullet} + OH^{\bullet}$$

Once formed, $CH_3^{\bullet}$ radicals can then undergo recombination and dehydrogenation reactions forming a variety of C_2 species, the concentrations of which can reach substantial levels in fuel-rich flames (see Fig. 9):

$$CH_3^{\bullet} + CH_3^{\bullet} \rightarrow C_2H_6$$

$$C_2H_6 + H^{\bullet} \rightarrow C_2H_5^{\bullet} + H_2$$

$$C_2H_5^{\bullet} \rightarrow C_2H_4 + H^{\bullet}$$

$$C_2H_4 + H^{\bullet} \rightarrow C_2H_3^{\bullet} + H_2$$

$$C_2H_3^{\bullet} \rightarrow C_2H_2 + H^{\bullet}$$

Based on detailed chemical kinetic modeling studies, the vinyl radicals appear to play an important role

on flame chemistry.[13] In the presence of abundant oxygen, vinyl radicals undergo oxidation reactions producing CO and CO_2:

$$C_2H_3^\bullet + O_2 \rightarrow CHO^\bullet + CH_2O$$

$$CHO^\bullet \rightarrow CO + H$$

$$CO + OH^\bullet \rightarrow CO_2 + H^\bullet$$

However, under fuel-rich conditions, further molecular weight growth becomes likely by the competing polymerization reactions leading to C_3, C_4, C_5, C_6, aromatics, and polyaromatic species.[13,19,20] Examples of these reactions include:

$$C_2H_3^\bullet + CH_3 \rightarrow C_3H_6^\bullet \rightarrow C_3H_4 + H_2$$

$$C_2H_3^\bullet + C_2H_2 \rightarrow C_4H_5^\bullet \rightarrow C_4H_2 + H_2 + H^\bullet$$

$$C_4H_5^\bullet + C_2H_2 \rightarrow C_6H_7^\bullet \rightarrow C_6H_6(\text{benzene}) + H^\bullet$$

$$C_3H_4 + C_3H_3^\bullet \rightarrow C_6H_6(\text{benzene}) + H^\bullet$$

$$C_6H_6 + C_2H_3^\bullet \rightarrow C_6H_6C_2H_3^\bullet(+C_2H_2) \rightarrow C_{10}H_8(\text{naphthalene}) + H_2 + H^\bullet$$

Subsequent molecular weight growth reactions of C_6H_6 (benzene) and $C_{10}H_8$ (naphthalene) then produce even higher molecular weight PAH, some of which are presented in Fig. 8. Further growth in the molecular weights of hydrocarbon by-products results in the formation of species with extremely low vapor pressures. These larger species then condense and further dehydrogenate producing heterogeneous (i.e., liquid) nuclei for soot formation, which is an important universal characteristic of fuel rich flames (Fig. 10).

The nature of intermediates formed in diffusion flames is similar to the premixed ones, albeit differences in the contacting pattern. In Fig. 11, the species concentration profiles in a laminar ethylene diffusion flame front are presented.[21] The fuel and oxygen diffuse toward each other undergoing virtual annihilation within the flame zone concomitant with the establishment of a peak temperature of about 1600°C. Because premixed systems provide a better control of combustor temperature, and many practical combustion devices operate under diffusion limited conditions, considerable effort has been expended to ensure the rapid mixing of fuel and oxygen in combustion chambers and approach premixed conditions.

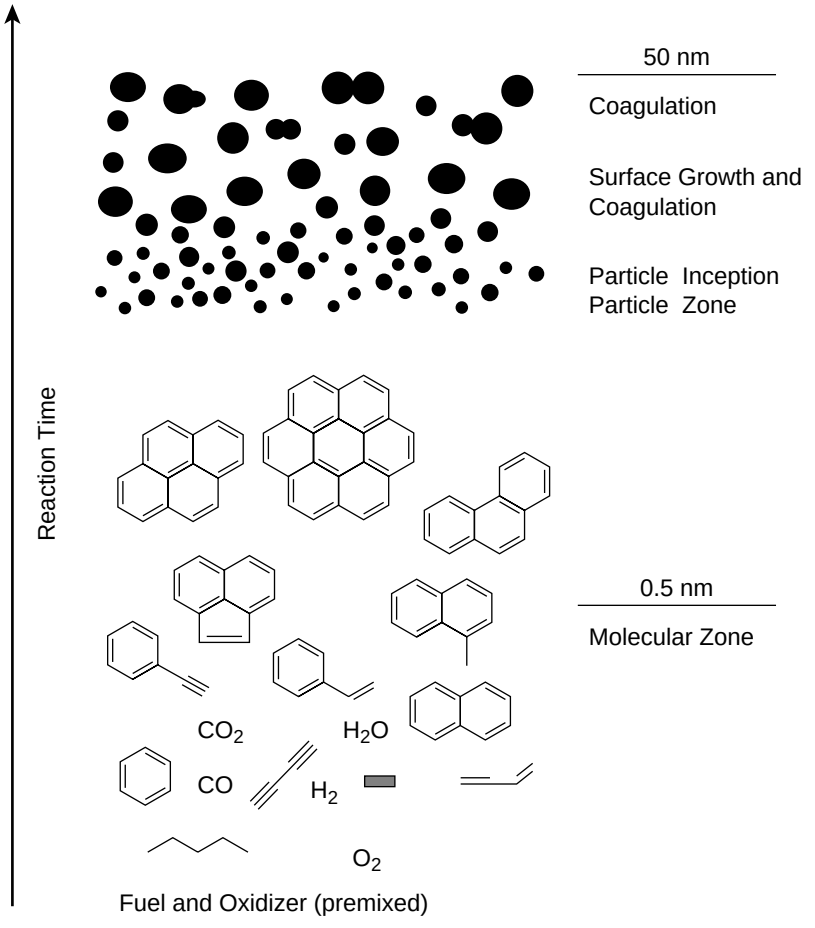


Fig. 10 Schematic of soot growth from fuel molecules through PAH. (From Bockhorn, 1994.)

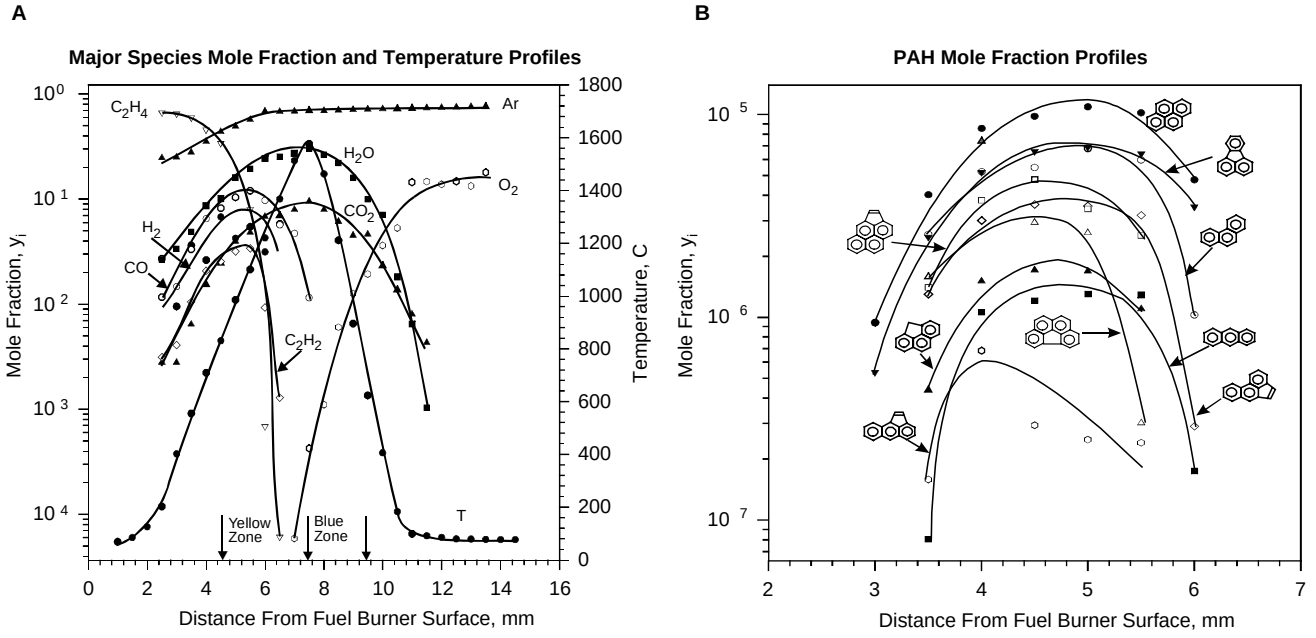


Fig. 11 Species concentration profiles vs. distance from fuel burner surface for (A) major species and (B) PAHs generated in an ethylene opposed flow diffusion flame. (From Ref.[21].)

NO_x, SO_x, and Other Pollutants

Although higher flame temperatures are desirable for improved waste/fuel destruction, there are also drawbacks. Most significantly, the reactions leading to NO_x formation increase with increasing temperature. This is shown in Fig. 12, where equilibrium NO levels are shown to increase with increasing temperature and decreasing fuel equivalence ratio in methane–air mixtures.[22] At high temperatures, molecular nitrogen and oxygen interact with one another in accordance with Zeldovich or ''thermal NO'' mechanism:[22]

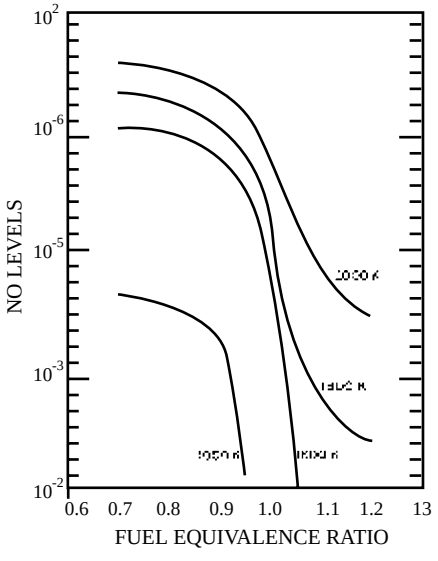


Fig. 12 Equilibrium NO concentration as a function of equivalence ratio for various flame temperatures. (From Ref.[22].)

$$O^{\bullet} + N_2 \rightarrow NO + N^{\bullet}$$

$$N^{\bullet} + O_2 \rightarrow NO + O^{\bullet}$$

The kinetics of these reactions are such that their role becomes significant only above 1500°C. Consequently, diffusion flames are particularly prone to higher levels of thermal NO production because of their higher peak flame temperatures. Besides the Zeldovich mechanism, NO formation can also occur via the ''prompt-NO'' mechanism[23] and from fuel nitrogen sources. In the prompt-NO mechanism, the reactions of CH radicals, produced by the sequential degradation of hydrocarbon fuels, with N_2 are responsible for NO production:

$$CH^{\bullet} + N_2 \rightarrow HCN + N^{\bullet}$$

$$HCN + O^{\bullet} \rightarrow NCO^{\bullet} + H^{\bullet} \quad \text{or} \quad NH^{\bullet} + CO$$

$$NCO + H^{\bullet} \rightarrow NH^{\bullet} + CO$$

$$NH^{\bullet} + O^{\bullet} \rightarrow NO + H^{\bullet}$$

$$NH^{\bullet} + H^{\bullet} \rightarrow N^{\bullet} + H_2$$

$$N^{\bullet} + O_2 \rightarrow NO + O^{\bullet}$$

These and related reactions that are believed to be involved in prompt-NO production are presented in

Fig. 13.[22] This same reaction mechanism also plays an important role in the thermal de-NO_x process to remove NO from combustion products using NH_3 and is discussed later.

When waste materials contain nitrogen bound in organic molecules, they can contribute to the formation of nitrogen oxides. The primary mechanism of NO_x formation in this case is the conversion of fuel-bound nitrogen to HCN and HNCO, followed by the conversion of these species into NH_x and then to NO, as described above. NO_2 and N_2O can also form in combustion processes through reactions such as:

$$NO + HO_2^{\bullet} \rightarrow NO_2 + OH^{\bullet}$$

$$O^{\bullet} + N_2 \rightarrow N_2O$$

There are two primary approaches for NO_x control: 1) combustion process modifications that minimize NO_x production in the first place and 2) postcombustion NO_x reduction methods. The nature of the fuel or combustible waste impacts the selection of the most suitable approach. For low-nitrogen feeds, thermal NO is the major contributor to NO, making temperature control the most important parameter. For incinerator feeds that contain significant levels of bound nitrogen, both the thermal- and fuel-NO formation processes become operative. In this case, the most effective strategy would be to implement staged combustion.[24,25]

Methods to control NO_x emissions in postcombustion gases can be categorized into noncatalytic and catalytic ones. Noncatalytic methods involve the injection of nitrogen-containing additives at concentrations comparable to the NO_x levels, to the combustion exhaust gases over a well-defined temperature window to induce NO_x reduction.[22] A large number of methods have also been developed for the selective catalytic reduction (SCR) of NO_x in combustion products.[26]

Some waste materials, coal, and some petroleum products also contain sulfur thereby creating SO_2 and SO_3 (SO_x) upon combustion. Together, NO_x and SO_x are the contributors to ''acid rain'' by their transformation into HNO_3 and H_2SO_4, respectively, in the atmosphere. These emissions, besides their obvious adverse environmental and health effects, also result in the erosion of structural and ornamental stonework and leads to corrosion of equipment. It is generally agreed that SO_2 formation occurs by the following sequential reactions of sulfur:[27]

$$\text{Fuel-S} \rightarrow RS \rightarrow SO \rightarrow SO_2 \rightarrow SO_3$$

The oxidation of SO_2 to SO_3 is thermodynamically favored at lower temperatures. However, as the homogeneous reaction kinetics are slow at low temperatures, SO_2 emissions dominate SO_x emissions from most combustion systems. On the other hand, significant SO_3 formation can occur through heterogeneously catalyzed reaction pathways, when the waste contains transition metals, especially vanadium. The most effective way to reduce SO_x emissions is to decrease the use of sulfur containing materials in the first place. In the absence of this, sulfur oxides can be readily removed from combustion flue gases by using sorbent adsorption.[28]

$$CaO + SO_2 + \tfrac{1}{2}O_2 \rightarrow CaSO_4$$

The $CaSO_4$ formed is stable, thus can safely be stored. However, the management of massive quantities of $CaSO_4$ that accumulate over a period of time can be a disposal problem. Other regenerative methods of SO_2 removal have also been developed.[28]

Combustion of Halogenated Hydrocarbons

Combustion of halogen containing wastes, in particular chlorine-containing compounds and chlorinated hydrocarbons (CHC) has been the focus of considerable controversy, and therefore, the research over the past several decades (see Refs.[29–32]). Technological and scientific issues faced in the combustion of CHC,

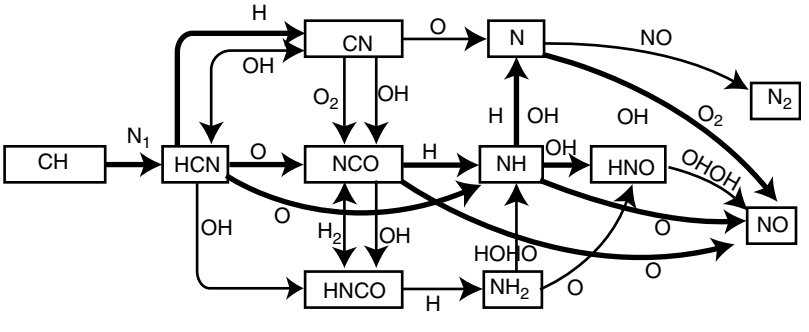


Fig. 13 NO production from fuel reactions. (From Ref.[22].)

in principle, are similar to those associated with the combustion of regular hydrocarbons as discussed in a recent review.[33] However, the differences in the reactivity of chlorine when compared to hydrogen, creates unique issues that must be addressed. For a general CHC designated by the formula $C_xH_yCl_z$, the overall stoichiometry of combustion can be described by the following equation:

$$C_xH_yCl_z + (x + (y - z)/4)O_2 \rightarrow xCO_2 + zHCl + ((y - z)/2)H_2O$$

As indicated above, the thermodynamically preferred CHC combustion product is HCl for $y > z$. However, for CHC compounds in which $y < z$, the formation of molecular chlorine must also be considered:

$$C_xH_yCl_z + xO_2 \rightarrow xCO_2 + yHCl + ((z - y)/2)Cl_2$$

In practical incinerators, the formation of Cl_2 is undesirable as it is difficult to remove from combustion effluents. Consequently, an auxiliary fuel with sufficient hydrogen content is used to increase H/Cl ratio and suppress Cl_2 formation. When methane is used, the overall combustion stoichiometry for the CHC/CH_4 mixture will be:

$$C_xH_yCl_z + (1/R)CH_4 + (2/R + x + (x - y)/4)O_2 \rightarrow (x + 1/R)CO_2 + zHCl + (2/R + (y - z)/2)H_2O$$

where R is the CHC/CH_4 ratio. Another reaction that is of significance in CHC combustion is the Deacon reaction:

$$2HCl + \tfrac{1}{2}O_2 = Cl_2 + H_2O$$

Excess air (fuel lean conditions) and high temperatures favor the conversion of HCl into Cl_2. The formation of Cl_2 can then lead to the production of chlorinated hydrocarbons, as seen in flame experiments.

In Fig. 14, atmospheric pressure premixed, flat flames of C_2H_3Cl are shown under both fuel-rich and fuel-lean conditions. Under fuel-rich conditions, CHC flames exhibit similar luminosity as hydrocarbon flames. However, under excess air conditions, the flames of chlorinated hydrocarbons exhibit white luminosity, likely because of the radiative recombination of chlorine atoms (Fig. 14A). Flames of highly chlorinated hydrocarbons also exhibit soot formation even at stoichiometric conditions (Fig. 14B) owing to the suppression of oxidation reactions.

The kinetics of chlorinated hydrocarbon reactions exhibit several distinguishing features when compared to the reactions of analogous hydrocarbons.[34] First, as a consequence of the weaker C–Cl bond dissociation energies, CHCs decompose at temperatures that are significantly lower than the analogous hydrocarbons. This leads to the early production of Cl radicals, thereby establishing the requisite free radical chain reactions, which destroy the parent CHC and lead to gas phase polymerization, ultimately forming soot.

In Fig. 15, the sooting limits of chlorinated methanes/methane/air mixtures are presented in terms of critical equivalence ratios and as a function of the chlorinated methane/methane molar ratio (R).[35] As the chlorine content of the mixture was increased, sooting occurred at lower equivalence ratios.

Second, chlorine and chlorinated compounds effectively compete with the important combustion chain branching reaction $H^\bullet + O_2 \rightarrow OH^\bullet + O^\bullet$ for the $H^\bullet$ radicals via the following reactions:

$$H^\bullet + CH_3Cl \rightarrow HCl + CH_3^\bullet$$

$$H^\bullet + CHC \rightarrow HCl + R^\bullet$$

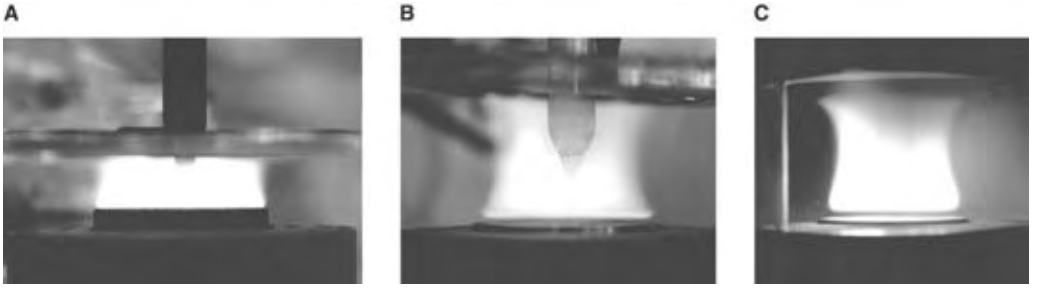


Fig. 14 (A) Premixed flames of trichloroethylene under fuel-rich, (B) stoichiometric, and (C) fuel-rich conditions. In contrast to regular hydrocarbons, soot formation can be seen in CHC flames under stoichiometric, and in some cases even under fuel lean conditions.

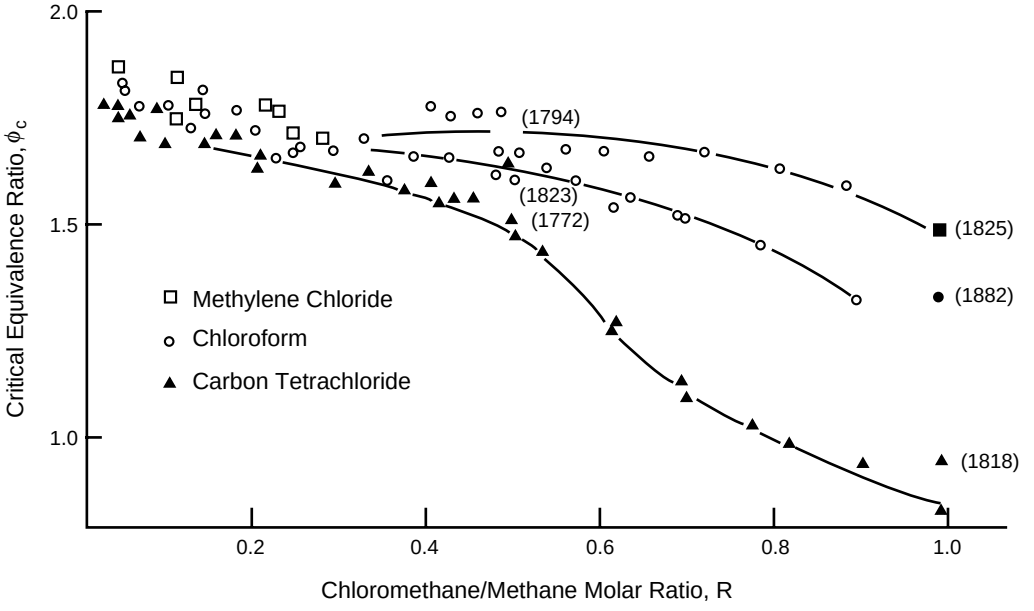
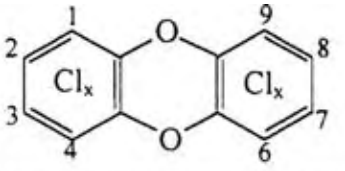


Fig. 15 Critical equivalence ratios (onset of soot formation) of premixed methane and chlorinated methane flames. (From Ref.[35].)

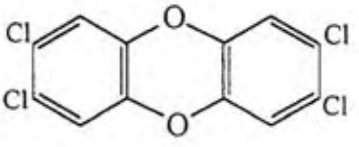
These reactions effectively remove the highly reactive $H^{\bullet}$ from radical pool, thereby inhibiting combustion processes.[18,36] This inhibition also manifests itself by the reduction in laminar burning velocities of hydrocarbon flames and by increased induction periods. In Fig. 16, the laminar burning velocities of chlorinated methanes are compared to methane.[37] Increased chlorination significantly reduces the burning speed of methane, rendering highly chlorinated CHC, such as CCl_4 nonflammable.

To better understand the fate of chlorine in combustion and to help in the design and operation of practical incinerators, the chemistry of flames of chlorinated hydrocarbons has been studied considerably. In Fig. 17, the detailed structures of CH_3Cl/CH_4 flame (more sooting) are compared to the CH_4 flame (less sooting).[38] As seen from this figure, in spite of differences in soot levels, the concentrations of CO, CO_2, and C_2H_2 in both flames were very close to one another, with the expected differences in HCl and H_2O. However, the levels of gas-phase aromatics and PAH in the more sooting CH_3Cl/CH_4 flame were significantly and consistently *lower* than the less sooting CH_4 flame. This is a surprising result with major practical and fundamental implications. Chlorinated compounds appear to assist in the rapid incorporation of PAH into soot, thereby reducing the levels of gas-phase PAH in the more sooting CH_3Cl/CH_4 flame.

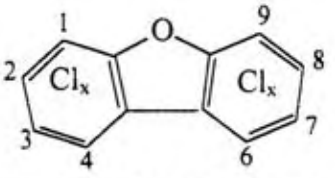
As noted earlier, polychlorinated dibenzo dioxins (PCDD) and polychlorinated dibenzo furans (PCDF) are two important classes of potentially toxic by-products that can form in trace levels (e.g., less than parts per billion, ppb) in waste incineration. The molecular structures of these compounds are:[29,33]



PolyChlorinated Dibenzo Dioxin (PCDD)



2,3,7,8 TetraChloro Dibenzo Dioxin (TCDD)
(most toxic isomer)

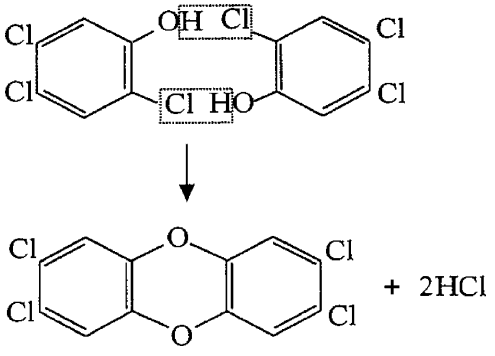


PolyChlorinated Dibenzo Furan (PCDF)

Although there are over 100 isomers of PCDD, the 2,3,7,8-tetrachlorodibenzodioxin has received the most attention because of its high toxicity in animal tests.[7] Three possible mechanisms have been put forward to account for dioxin and furan emissions

from incinerators: 1) unburned dioxins and furans in the original waste stream that survive the combustion chamber; 2) homogeneous gas-phase synthesis from aromatic precursors such as chlorinated benzenes and phenols;[39–41] and 3) heterogeneous and/or de novo synthesis from carbonaceous materials with particulate carbon and carbon species with functional groups believed to be the best starting materials.[29,39,42–44] The latter class also includes the surface assisted formation of PCDD/F from gaseous aromatic precursors.[39]

Conceptually, the simplest PCDD formation mechanism is the recombination of 2-chlorinated phenols via HCl elimination either homogeneously in the gas phase or on surfaces. For the case of 2,3,7,8 TCDD, this involves the combination of 2,4,5 trichlorophenol, as shown:



In homogeneous reaction mechanisms, PCDD and PCDF are considered to be formed from aromatic precursors at temperatures below 600°C with the active participation of OH radicals.[40]

The mechanisms of formation of PCDD and PCDF in heterogeneous systems have been reviewed by Addink and Olie,[29] and the following trends have been noted:

1. Carbon source: Virtually any carbon source seems to produce PCDD/F via the de novo route, including activated carbon, coal, charcoal, and soot.
2. Surface: In addition to fly ash, a large number of materials such as Al_2O_3, $Al_2O_3SiO_2$, carbon, firebrick, glass wool, MgSiO, $MgAlSiO_2$, SiO_2, SiO_2NaOH, and tenax were evaluated for dioxin and furan formation. The production of PCDD/F was possible on most of these surfaces, but the presence of transition metals appeared to be necessary.
3. Chlorine source: Both organic and inorganic, gaseous and solid chlorinated compounds appear to be capable of providing the necessary chlorine atoms for the formation of PCDD and PCDF.
4. Temperature: The lowest temperature reported for the formation of PCDD/F from activated carbon on fly ash has been 200 °C. The temperature range that maximizes PCDD/F formation varies: 300°C for charcoal/fly ash, 300–330°C for residual carbon on fly ash, and 350–375 °C for activated carbon/fly ash. At 470°C, a second maximum was observed with residual carbon on fly ash and detectable amounts of PCDD/F are formed even at 550 °C.
5. Catalysts and promoters: Transition metals such as Cu and Fe are well recognized to promote PCDD/F formation in laboratory experiments.
6. Oxygen source: Free oxygen appears to be essential for PCDD/F formation in most heterogeneous and homogeneous processes.

In spite of the above mentioned investigations that indicated several trends on PCDD/F formation, a study conducted for the American Society of Mechanical Engineers, ASME,[30] concluded that there was no statistically significant cross-incinerator correlation between chlorine content of the waste stream fed to

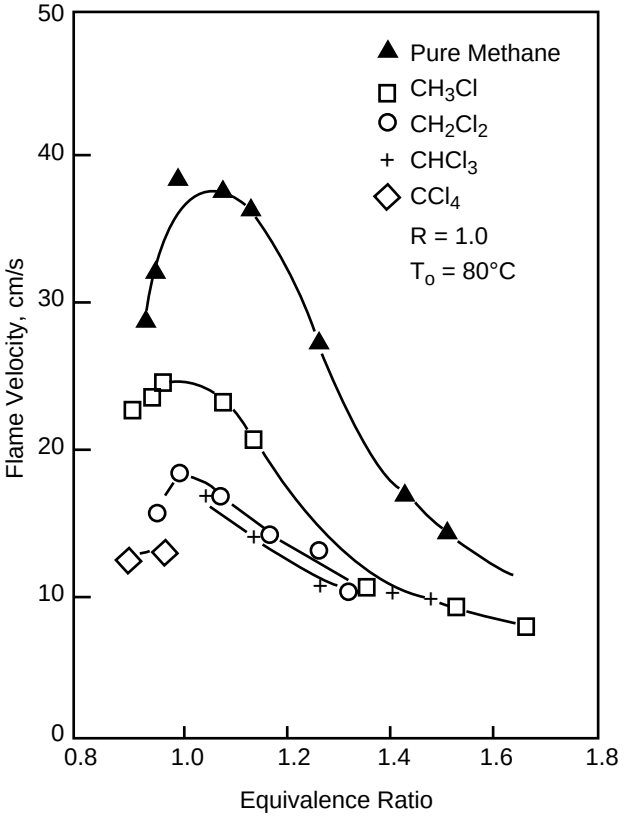


Fig. 16 Laminar burning velocities of chlorinated hydrocarbons as a function of equivalence ratio. Increased flame inhibition with increased chlorine content is evident.

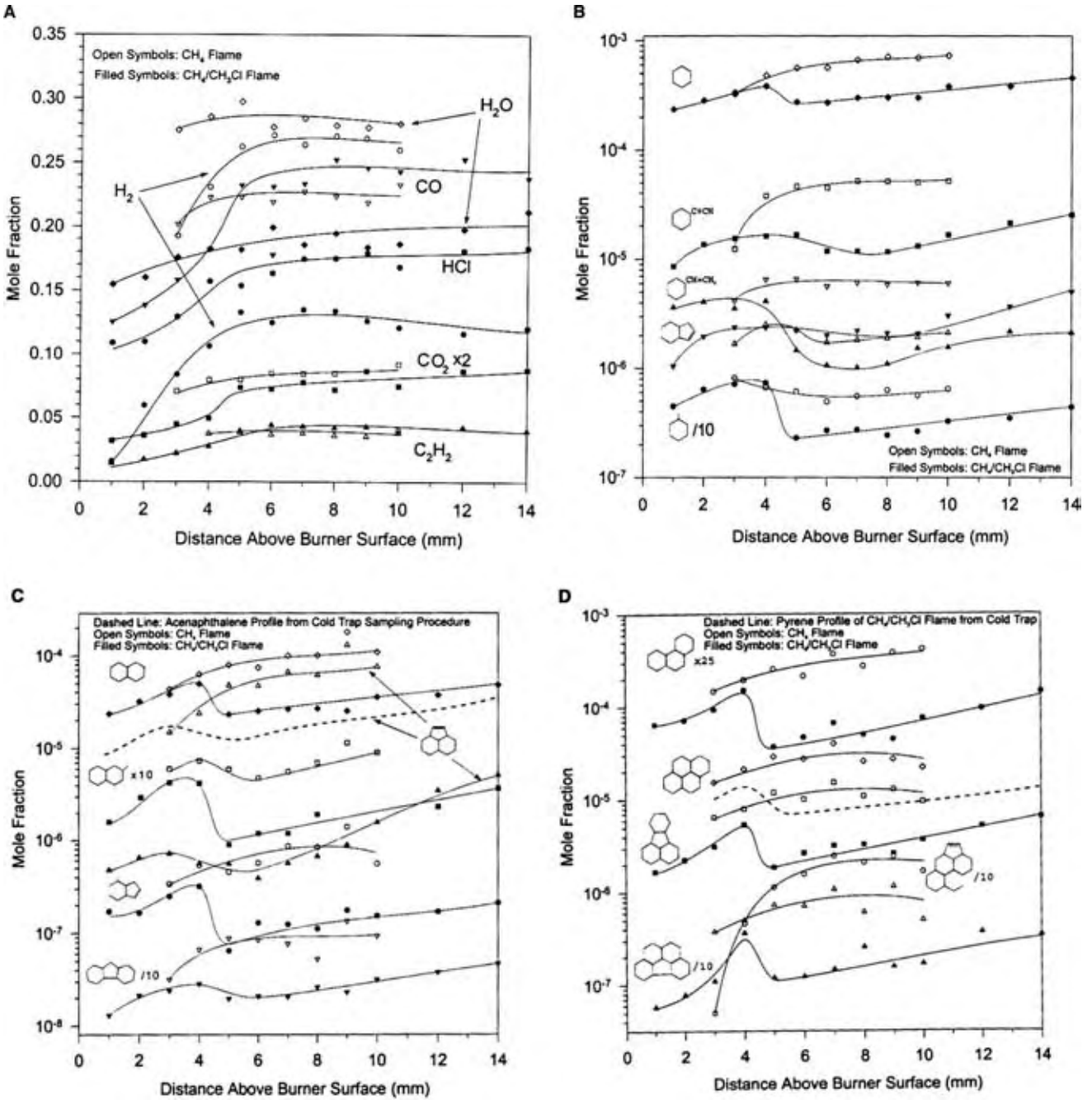


Fig. 17 Comparative species concentration profiles in the flames of CH_3Cl/CH_4 and CH_4 under similar equivalence ratio and carbon density. (A) Major species mole fraction profiles; (B) single ring aromatic species mole fraction profiles; (C) two ring aromatic species mole fraction profiles; (D) three and four ring aromatic species mole fraction profiles. (From Ref.[38].)

incinerators and the PCDD/F concentrations in the emissions from those incinerators. Although numerous factors have been identified above that lead to dioxin and furan formation, the only consensus at this point seems to be that good combustion efficiencies and low postcombustion temperatures reduce the secondary dioxin formation.

Particulate Matter and Heavy Metal Emissions from Incinerators

Particulate matter from waste combustors includes inorganic ash present in the waste and carbonaceous soot formed in the combustion process. The inorganic-ash fraction of the PM consists of mineral matter

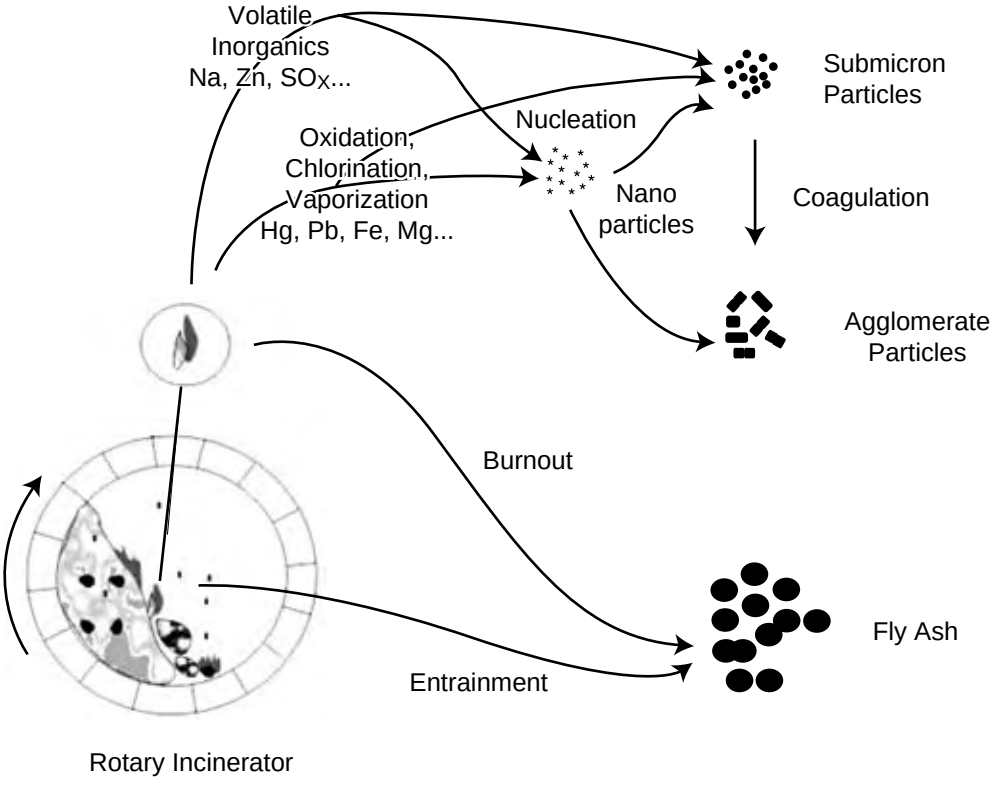


Fig. 18 The fate of metals and inorganics in the incineration of waste materials.

and metallic species. These materials are conserved in the combustion process and leave the combustion chamber as bottom ash or fly ash. Soot is a product of incomplete combustion that consists of unburned carbon in the form of fine particles or as deposits on inorganic particle.[11,45] The fate of metals in incineration is significantly complicated, but similar to those encountered in coal combustion.[46,47] Consequently, readers interested to learn more about metal emissions from incinerators can be well served by the extensive literature on coal combustion.[45–47] Metal emissions from incinerators can either be direct, i.e., through entrainment of small ash particles created during waste combustion or through a variety of physical and chemical processes (Fig. 18). The high temperatures associated with flames in the primary incineration chamber can gasify the more volatile metals such as Na, Hg, and Zn, while some are gasified through reactive processes involving oxidations [e.g., $Ni(CO)_4$], chlorinations (e.g., $PbCl_2$), etc. Subsequent decreases in the gas temperature then result in the nucleation of metal-containing gases, and can create nanoparticles, which can then either grow further by surface reactions, or agglomerate and/or sinter to form submicron size particles.

Particulate matter emissions can be controlled using a variety of APCD that include filtration collectors, including primary fabric filters (baghouses); electrostatic collectors, including dry and wet electrostatic precipitators and ionizing wet scrubbers; and wet inertial-impaction collectors, including venturi scrubbers and advanced designs that use flux-force condensation-enhancement techniques.

In addition to APCD, metal emissions from waste combustors can be minimized by: 1) limiting the metal content of the waste feed via source control; 2) designing and operating the combustion process to minimize metal vaporization; and 3) designing and operating the primary combustion chamber to minimize fly-ash carryover. From a practical standpoint, the second method is likely to be the most difficult to implement because the objective of the incineration process is to burn all the combustible waste completely and avoid PIC formation, both of which require the use of high temperatures. Therefore, the most-reliable methods of limiting metal emissions are source control and efficient use of APCD.

CONCLUSIONS

Today waste incinerators can be designed and operated to achieve better than 99.999% DRE of the waste compounds and to only emit extremely low concentrations of the pollutants of concern under normal

operating conditions. Consequently, incineration can be reliably used to reduce the volume and toxicity of a broad range of waste materials, while allowing the recovery of energy and materials. Like any other combustion process, incinerators can also produce potentially dangerous by-products. Nevertheless, the certain risks posed by specific incinerator emissions, e.g., dioxins, do not appear to be particularly higher than those posed by other combustion generated pollutants. It is unfortunate that the public image of incineration has been tarnished by several unfortunate accidents in the past. However, as we ran out of inexpensive waste management options, e.g., landfills, incineration is likely to become an essential tool in the management of wastes in the future.

REFERENCES

1. Lewis, B.; von Elbe, G. *Combustion, Flames and Explosions of Gases*; Academic Press: New York, 1961.
2. Yang, M.; Karra, S.; Senkan, S.M. Equilibrium-analysis of combustion incineration. Hazard Waste Hazard Mater. **1987**, *4*, 55–68.
3. NAE, National Academy of Engineering. In *Waste Incineration and Public Health*; National Academy Press: Washington DC, 2000.
4. Vovelle, C., Ed. *Pollutants from Combustion: Formation and Impact on Atmospheric Chemistry.* NATO Science Series C; Kluwer: Dordrecht, The Netherlands, 2000.
5. Gardiner W.C., Jr., Ed. *Gas-phase Combustion Chemistry*; Springer: New York, 2000.
6. Ferrza, M.C.M.A; Afonso, S.A.V. Dioxin emission factors for the incineration of different medical waste types. Arch. Environ. Contam. Toxicol. **2003**, *44*, 460–466.
7. Klaassen, C.D.; Amdur, A.O.; Doull, J., Eds. *Casarett and Doull's Toxicology*; Macmillan: New York, 1986.
8. Lewtas, J. Airborne carcinogens. Pharmacol. Toxicol. **1993**, *72*, S55–S63.
9. Gittins, C.; Castaldi, M.J.; Senkan, S. et al. Real-time quantitative analysis of combustion-generated polycyclic aromatic hydrocarbons by resonance-enhanced multiphotonionization time-of-flight mass spectrometry. Anal. Chem. **1997**, *69*, 286–293.
10. Cao, L.; Muhlberger, F.; Adam, T. et al. Resonance-enhanced multiphoton ionization and VUV-single photon ionization as soft and selective laser ionization methods for on-line time-of-flight mass spectrometry: investigation of the pyrolysis of typical organic contaminants in the steel recycling process. Anal. Chem. **2003**, *75*, 5639–5645.
11. Glassman, I. *Combustion*; Academic Press: San Diego, CA, 1996.
12. Westbrook, C.K.; Dryer, F.A. Chemical kinetic modeling of hydrocarbon combustion. Prog. Energy Combust. Sci. **1984**, *10*, 1–57.
13. Marinov, N.M.; Pitz, W.J.; Westbrook, C.K. et al. Aromatic and polycyclic aromatic hydrocarbon formation in a laminar premixed *n*-butane flame. Combust. Flame **1998**, *114*, 192–213.
14. Castaldi, M.J.; Vincitore, A.M.; Senkan, S. Micro-structures of premixed hydrocarbon flames: methane. Combust. Sci. Tech. **1995**, *107*, 1–19.
15. Fristrom, R.M.; Westenberg, A. A. *Flame Structure*; McGraw-Hill: New York, 1965.
16. Barnard, J.A.; Bradley, J.N. *Flame and Combustion*; Chapman and Hall: New York, 1985.
17. Pitts, W.M.; Nyden, M.R.; Gann, R.G. et al. *US National Institute of Standards and Technology*; NIST Technical Note, 1279; 1990.
18. Gann, R.G. Ed. *Halogenated Fire Suppressants*; American Chemical Society: Washington DC, 1975.
19. Frenklach, M.; Wang, H. Detailed mechanism and modeling of soot particle formation. In *Soot Formation in Combustion*; Bockhorn, H., Ed.; Springer: Berlin, 1994; 165.
20. Richter, H.; Howard, J.B. Formation of polycyclic aromatic hydrocarbons and their growth to soot—a review of chemical reaction pathways. Prog. Energy Combust. Sci. **2000**, *26*, 565–608.
21. Olten, N.; Senkan, S. On-line measurements of the polycyclic aromatic hydrocarbons (PAH) in counter-flow ethylene diffusion flame. Combust. Sci. Tech. **2000**, *159*, 1–15.
22. Bowman, C.T. Gas phase reaction mechanisms for nitrogen oxide formation and removal in combustion. In *Pollutants From Combustion: Formation and Impact on Atmospheric Chemistry*; Vovelle, C., Ed.; NATO Science Series C; Kluwer: Dordrecht, The Netherlands, 2000.
23. Fenimore, C.P. Formation of nitric oxide in premixed hydrocarbon flames. In 13th Symposium (Int'l) on Combustion. The Combustion Institute: Pittsburgh, 1971; 373–380.
24. Wendt, J.O.S.; Sterling, C.V.; Matovich, M.A. Reduction of sulfur trioxide and nitrogen oxides by secondary fuel injection. In 14th Symposium (Int'l) on Combustion. The Combustion Institute: Pittsburgh, 1973; 897.
25. Smoot, L.D.; Hill, S.C.; Xu, H. NO_x control through reburning. Prog. Energy Combust. Sci. **1998**, *24*, 385–408.
26. Parvulescu, V.I.; Grange, P.; Delmon, B. Catalytic removal of NO. Catal. Today **1998**, *46*, 233–316.

27. Hynes, A.J.; Wine, P.H. Kinetics and mechanisms of the oxidation of gaseous sulfur compounds. In *Combustion Chemistry*; Gardiner, W.C., Jr., Ed.; Springer: New York, 2000.
28. Flagan, R.C.; Seinfeld, J.H. *Fundamentals of Air Pollution Engineering*; Prentice Hall: Englewood Cliffs, New Jersey, 1988.
29. Addink, R.; Olie, K. Mechanisms of formation and destruction of polychlorinated dibenzo-*p*-dioxins and dibenzofurans in heterogeneous systems. Environ. Sci. Tech. **1995**, *29*, 1425–1435.
30. American Society of Mechanical Engineers. In *The Relationship between Chlorine in Waste Streams and Dioxin Emissions from Waste Incinerator Stacks*; ASME Press: Fairfield, New Jersey, 1995, CRTD. Vol. 36.
31. McKay, G. Dioxin characterization, formation and minimization during municipal solid waste (MSW) incineration. Chem. Eng. J. **2002**, *86*, 343–368.
32. Quaβ, U.; Fermann, M.; Broker, G. The European dioxin air emission inventory project—final results. Chemosphere **2004**, *54*, 1319–1327.
33. Senkan, S. Combustion of chlorinated hydrocarbons. In *Pollutants in Combustion*; Vovelle, C., Ed.; NATO Science Ser. C 547, 2000a; 303.
34. Senkan, S. Survey of rate coefficients in the C–H–Cl–O system. In *Gas-phase Combustion Chemistry*; Gardiner, W.C.Jr., Ed.; Springer: New York, 2000b; 389.
35. Senkan, S.; Gupta, A.; Robinson, J.M. Sooting limits of chlorinated-hydrocarbon methane air premixed flames. Combust. Flame **1983**, *49*, 305–314.
36. Westbrook, C.K. Inhibition of hydrocarbon oxidation in laminar flames and detonations by halogenated compounds. In 19th Symposium (Int'l) on Combustion. The Combustion Institute: Pittsburgh, 1982; 127–141.
37. Valeiras, H.; Gupta, A.K.; Senkan, S.M. Laminar burning velocities of chlorinated hydrocarbon–methane–air flames. Combust. Sci. Tech. **1984**, *36*, 123–133.
38. Huang, J.W.; Senkan, S.M. Polycyclic aromatic hydrocarbon and soot formation in premixed flames of CH_3Cl/CH_4 and CH_4. 26th Symposium (Int'l) on Combustion. The Combustion Institute: Pittsburgh, 1996; 2335–2341.
39. Schaub, W.M.; Tsang, W. Dioxin formation in incinerators. Environ. Sci. Tech. **1983**, *17*, 721–730.
40. Ritter, E.R.; Bozzelli, J.W. Pathways to chlorinated dibenzodioxins and dibenzofurans from partial oxidation of chlorinated aromatics by OH radical—thermodynamic and kinetic insights. Combust. Sci. Tech. **1994**, *101*, 153–169.
41. Lenoir, D.; Wehrmeier, A.; Schram K.W., et al. Thermal formation of polychlorinated dibenzo-*p*-dioxins and -furans: investigations on relevant pathways. Environ. Eng. Sci. **1998**, *15*, 37–47.
42. Vogg, H.; Stieglitz, L. Thermal-behavior of PCDD/PCDF in fly-ash from municipal incinerators. Chemosphere **1986**, *15*, 1373–1378.
43. Karasek, F.W.; Dickson, L.C. Model studies of polychlorinated dibenzo-*para*-dioxin formation during municipal refuse incineration. Science **1987**, *237*, 754–756.
44. Stieglitz, L. Selected topics on the de novo synthesis of PCDD/PCDF on fly ash. Environ. Eng. Sci. **1998**, *15*, 5–18.
45. Bartok, W.; Sarofim, A. Eds. *Fossil Fuel Combustion: A Source Book*; Wiley: New York, 1991.
46. Williams, A.; Pourkashanian, M.; Jones, J.M. Combustion of pulverised coal and biomass. Prog. Energy Combust. Sci. **2001**, *27*, 587–610.
47. Elliott, M.A. Ed. *Chemistry of Coal Utilization*; Wiley-Interscience: New York, 1981.

I

Injection Molding

I

David O. Kazmer
Department of Plastics Engineering, University of Massachusetts, Lowell, Massachusetts, U.S.A.

INTRODUCTION

Injection molding is a well-known process that is capable of economically producing very complex components with demanding specifications.[1] Many different kinds of products, from compact discs to audio and video-cassettes, to cutlery and glassware, to automotive parts, are molded using various types of injection molding. In the first half of this entry, a concise review of the process is provided. Then, prevalent guidelines for the design of plastic parts and injection molds are presented.

Continuing global competition is forcing producers of molded parts and suppliers of molding equipment to improve the molding process. Toward this end, an overview of the current state of the art in injection molding is provided in the second half of this entry. Modern machine and process control system designs are presented that provide improved control over the polymer melt. Then, variants of the injection molding process that have been developed to produce molded parts that are hollow, less dense, very thin, or composed of multiple materials are introduced. Finally, a summary of current technological and economic trends in the industry is provided.

PROCESS DESCRIPTION

Injection molding is a net shape manufacturing process in which a polymer melt is forced into an evacuated mold cavity that cools the polymer melt into a desired shape. As shown in Fig. 1, molding machines typically consist of two halves located on opposing sides of a stationary platen: 1) an injection unit that plasticizes the melt and transfers it into the mold, and 2) a clamping unit that closes the mold during the formation of the part and opens the mold for the ejection of the molded product.[2] The design and the operation of molding machines differ substantially, but, in general, adhere to underlying principles that are discussed next.

During the plastication stage, the polymer melt is typically plasticized from solid granules or pellets through the combined effect of heat conduction from the heated barrel and the internal shear heating caused by molecular deformation with the rotation of an internal screw. Screws in injection molding have many similarities with those in single screw extrusion, e.g., most of them have feed, transition, and metering zones. The primary difference, however, is that the screw reciprocates along its axis during the molding cycle. Specifically, the screw moves away from the mold as it rotates and forms a volume of melt at the front of the barrel. The screw then moves forward without turning to force the polymer melt into the injection mold. As such, reciprocating screws in injection molding usually include a nonreturn or check valve to prevent the polymer melt at the front of the barrel from flowing back into the screw during injection. While the dynamics and efficiency of plastication are well known to be a function of the screw design, material properties, and process conditions,[3] many, if not the majority of, molding processes rely on general purpose screw designs that provide reasonable performance for a wide variety of polymer resins.

During the injection stage, the polymer melt is forced from the barrel of the molding machine and enters the mold. Typically, the heated resin travels down a cooled and/or heated feed system, through one or more gates, and into one or more mold cavities where it will form one or more desired shapes. Control of the injection stage is well known to influence the properties of the molded products such as part weight, dimensions, esthetics, orientation, and others.[4–6] As such, modern molding machines allow velocity profiling to control the volumetric flow rate of the polymer melt into the mold during the injection stage. Molding processes are frequently set up so as to best provide a uniform melt front velocity by utilizing a low volumetric flow rate at the start of the injection stage when the polymer melt is just entering the cavity, then a higher volumetric flow rate as the polymer melt diverges and propagates throughout the mold, and then a lower volumetric flow rate as the polymer melt converges toward the end of the mold cavity.

The packing stage typically commences when the mold is filled or nearly filled with the polymer melt and provides additional resin into the mold cavity as the polymer melt cools and contracts. As the packing stage is also crucial to part quality, modern molding machines allow the melt pressure to be profiled as a function of time. The minimum packing pressure is usually set to avoid excessive volumetric shrinkage and

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007819

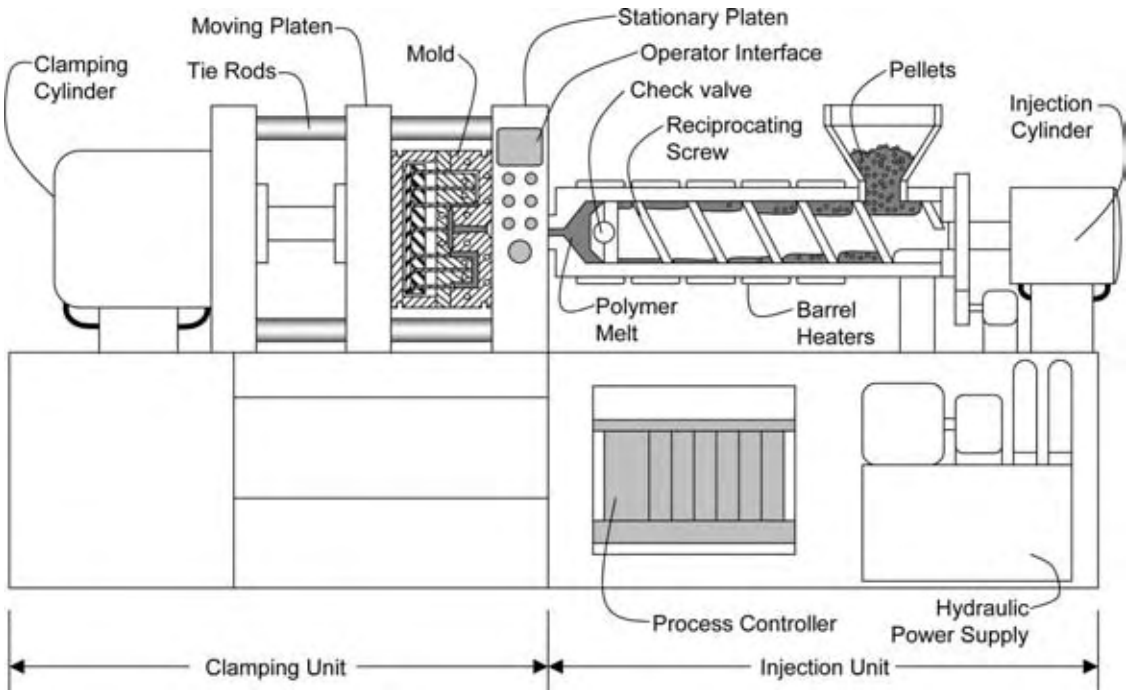


Fig. 1 Schematic of an injection molding machine.

related defects such as sink marks, internal voids, and dimensional instabilities.[7] The maximum packing pressure is usually set to avoid excessive pressure in the mold cavity and related defects such as flashing of the mold, internal stress, and dimensional instabilities.[8] The minimum and maximum packing times are determined by the volumetric shrinkage (required to achieve the desired part dimensions) and the solidification of the gate (which prevents additional polymer melt from entering the cavity at extended pack times), respectively. Accordingly, pack pressure profiling can be used to control the pressure, density, and stress distribution of the polymer melt as it solidifies in the cavity.

After the packing stage, molding processes typically require additional time to plasticate the polymer melt for the next molding cycle while the resin in the cavity solidifies and becomes rigid so that it may be ejected from the mold. Depending on the polymer properties and its inherent recrystallization rate, processing conditions, and the mold design, the cooling stage can range from seconds to minutes.[9] Excessively short cooling times can lead to difficulties in ejecting the part because of either internal compressive stresses that cause the part to remain in the mold cavity or excessive deformation of the part upon actuation of the ejection mechanism(s). Extended cooling times are sometimes used in an attempt to reduce part shrinkage, though this practice can lead to higher internal tensile stresses and difficulties in removing the molded part from the mold core(s) associated with increased ejection forces. Mold release is often utilized to facilitate part removal, though this practice introduces issues related to consistency and contamination. When the polymer melt is appropriately cooled, the molding machine typically actuates the necessary cores, slides, and pins to open the mold and remove the molded part(s).

PRODUCT AND MOLD DESIGN

Injection molding is commonly used because of its ability to form complex parts with features defined in three dimensions, and thereby consolidate the number of components in an assembly, and also reduce the number of fasteners required for the assembly (e.g., Fig. 2). Injection molded parts are usually designed to be thin relative to their length and width dimensions to reduce material and processing costs. As the wall thickness decreases, however, the part becomes very difficult to mold while losing stiffness and strength. As such, part design and mold design are of extreme importance, as a conservative design may result in excessive material and processing costs while an aggressive one may result in inferior performance and require costly design changes.

Computer-aided engineering applications have been developed to optimize the molded part design with respect to manufacturing and end use performance. Many, if not most, commercial molding applications utilize mold filling analyses to analyze the melt flow and pressure distribution in the injection mold before the mold is manufactured. Typically, these analyses

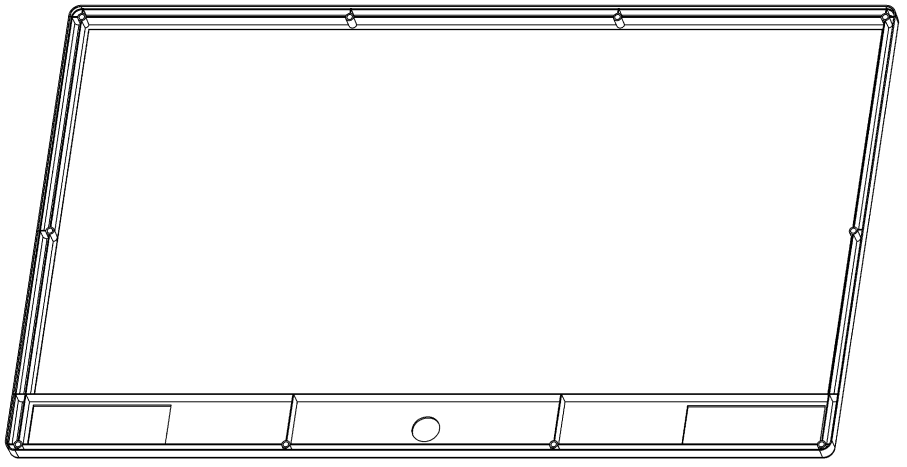


Fig. 2 Front bezel (face plate) for an HDTV with windows, ribs, and bosses.

(e.g., Moldflow, Moldex3D, and Fluent) include non-Newtonian modeling of the polymer viscosity using assumptions for viscous compressible flow. The implemented heat transfer analysis typically includes heat convection with the melt flow, heat conduction from the polymer melt to the metal mold, and internal heat generation from viscous dissipation. In Fig. 3, a typical result for a bezel design in which the flow is introduced at three locations indicated by cones is provided. The contour lines indicate the progressive location of the melt at different times during the mold filling process. For this design, the flow diverges radially from the gates before converging toward the end. This particular design is deficient in that there is a large area of the molded part that has not filled, which would result in localized areas with high melt pressure and nonuniform shrinkage in the molded parts. Most likely, this issue would be rectified by moving the two gates in the window closer together to reduce the flow length.

Computer-aided engineering applications (e.g., Ansys, Cosmos, and Abaqus) are also often utilized to analyze the stiffness and strength of candidate designs. Such concerns often arise in plastic part designs, as: 1) injection molded parts are thin relative to their length and width dimensions, and 2) polymeric materials have low elastic modulus and yield stress compared with metals. Because of this, injection molded parts are frequently designed with vertical ribs to ensure adequate stiffness and strength during end use.[10] As a result, molded plastic parts often have stiffness-to-weight ratios that are preferable to alternatively designed metal components.

Because of the widespread use of injection molded components, many guidelines have been developed with respect to the design of injection molding.[11] In general, it is desirable to use a uniform and optimal wall thickness throughout the part to reduce cooling times, and minimize residual stresses and part distortion upon ejection from the mold. Ribs should be used

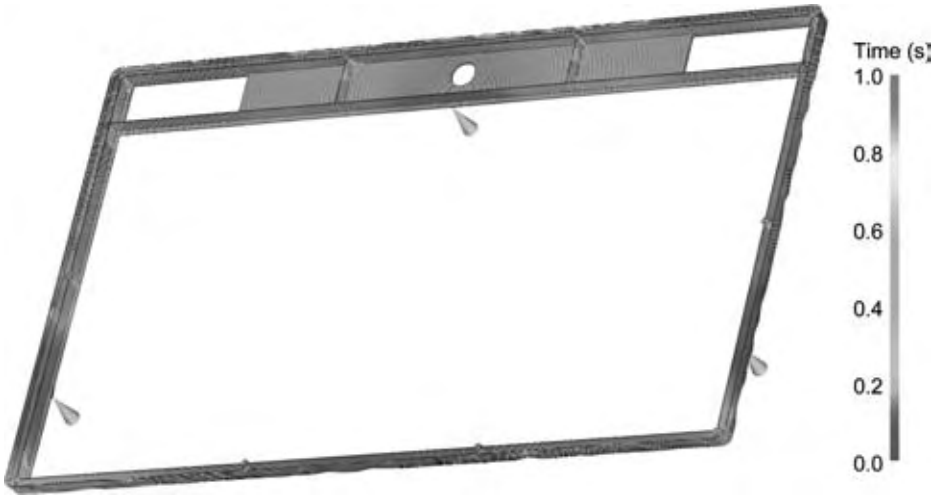


Fig. 3 Progression of polymer melt into mold cavity. *(View this art in color at www.dekker.com.)*

to improve the overall part stiffness, while gussets to secure bosses and other standing features. Generally, the thickness of the ribs and gussets, and other features should be no more than 70% of the part wall thickness to avoid sink and extended cycle times. It is also desirable to use a generous radius at all corners to avoid flow hesitation and mold hot spots during molding as well as stress concentrations in end use; an inside corner radius equal to the wall thickness is often used. Finally, it is important to design the molded parts to facilitate ejection from the mold by minimizing the number of undercuts that requires core pulls and also providing an acceptable amount of draft in the direction of the part ejection.

The design of injection molds[12] has improved with respect to the level of engineering analysis and manufacturing efficiency. Injection molds (e.g., Fig. 4) must be designed to ensure that the molded parts are economically and consistently produced. For a given molding application, the number of molds and mold cavities is first estimated based on the size and thickness of the part, the location of the gates, and the delivery of the polymer melt to the mold cavity(ies). When possible, the melt pressures are computed using mold filling analysis to estimate the forces likely to be exerted by the polymer melt on the mold components. Structural analysis, in the form of plate bending equations or finite element analysis, may be used to determine the number, location, and size of mold plates and support pillars. Cooling analysis is similarly conducted to determine the number, location, and size of cooling lines in the mold. Based on guidance from the material supplier and prior molding experience, the shrinkage of the plastic is estimated, which is used to determine steel-safe cavity and core designs that will form the molded part(s).

Because of the number of subsystems in an injection mold, mold design often requires multiple design iterations. For example, it is not uncommon for the design of the cooling lines to interfere with the placement of ejector pins. As such, the design must be continually refined until the designer is satisfied that the mold will function as desired. As mold making generally requires extensive machining and finishing time, it is quite common for mold makers to order the components and begin rough machining before the detailed mold design is completed. Owing to such concurrent engineering techniques and modern computer controlled machining centers, the cost and time required to produce a high quality mold have decreased significantly over time. After the mold is completed, the mold maker and molder typically perform molding trials to identify and correct any issues prior to the initiation of high volume production.

PROCESS CONTROL

Injection molding machines can control machine elements only with feedback from sensed process data. A fundamental difficulty in polymer processing is that very few of the final part properties can be ascertained within the molding cycle. Instrumentation does not yet exist, and may never exist, to yield information about

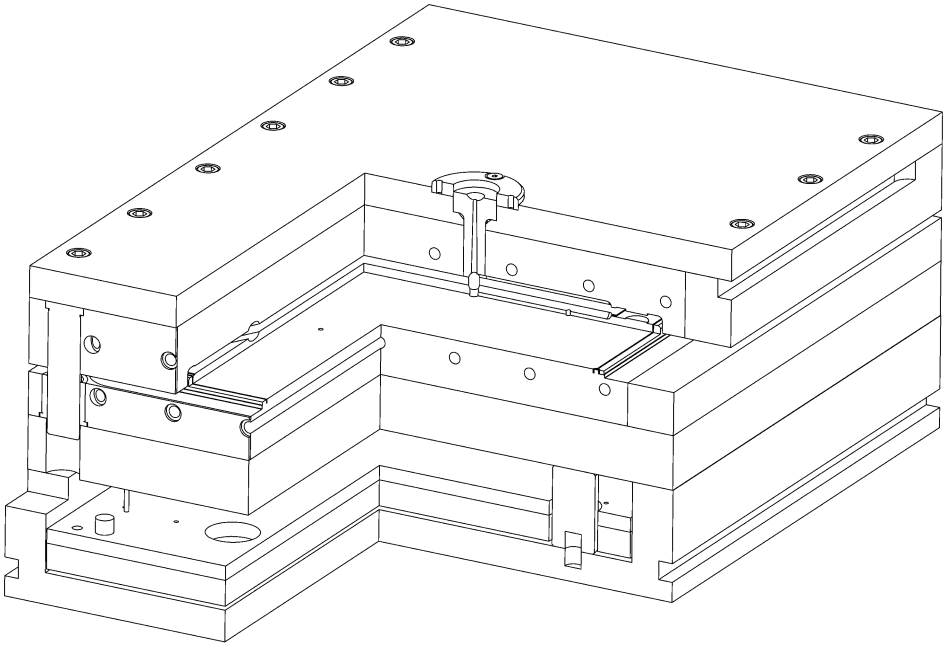


Fig. 4 Partial cutaway section of an injection mold.

esthetics, shrinkage, or structural integrity prior to removal of a plastic part from the mold. As the quality of the molded part is not available, in situ, molding machines operate in an open loop mode with respect to the quality of the molded products. Given the limited number of molding machine control settings, the process capability of injection molding will generally decline with an increasing number of molded parts and related specifications that are simultaneously demanded.[13]

Most machine manufacturers have adopted the use of standard programmable logic controllers in which multiple analog and digital input and output modules are used to acquire or transmit necessary feedback and control signals. The control signals may be generated through the CPU module in the programmable logic controller or alternatively through the use of dedicated axis controllers for faster response. Increasingly, industrial PCs are being used as an interface between the process controller and the operator, such that the PC provides enhanced ease of use, process diagnostics and maintenance, network connectivity, upgradability, and access to third party software.

The dynamics of the molding process are determined through control of different but related machine elements such as motors, heaters, servovalves, etc. These machine elements are typically controlled via a hierarchical closed loop control architecture as shown in Fig. 5.[14] At the innermost level, only the machine elements are regulated by real time comparison of the desired machine set points with the machine feedback, such that the difference (or error) is used to correct the process. At the second level, state variables such as melt temperature and melt pressure are controlled to track prespecified profiles and provide more precise control of the state of the melt. At the outermost level, the machine inputs are adjusted by the machine operator to improve the quality of the part through specification of better set points given feedback of part quality.

The performance of a closed loop controlled molding machine is dependent upon a number of system properties, such as the inertia and dynamic behavior of the process, availability and amount of control energy applied to the machine, the time response and resolution of the sensors providing feedback, and the validity of the control laws that convert perceived errors into corrective actions. Sustained advances in machine design and software algorithms have led to substantial gains in process control performance, such that the molding process control is not usually limited by the sweep time or algorithmic complexity of the controller. While there remains a significant variation in sweep time between molding machines, a control system sweep time (from input of feedback signal to output of control signal) of 2 msec is quite common, with much faster response times widely available. However, the process dynamics are not so dependent upon the response time of the controller, but rather on that of the integrated system. Consider the task of increasing the temperature of a 100 kg steel barrel (heat capacity of 473 J/kg °C) by 10°C, which theoretically requires 473,000 J. If four 1000 W heaters are utilized, then the minimum theoretical response time is about 120 sec or 2 min. Realistically, however, the response time will be much longer than the theoretical minimum, depending upon whether the system is over- or underdamped, and how much error is tolerable. As such, a reduction in the control system response time is likely to be less important than an improvement in the control law, controller tuning, heater design, or barrel design.

While machine control is important, it is the polymer state (pressure, temperature, and morphology) that directly determines the molded part quality. As such, recent technology developments have focused on closing the loop between the machine parameters and the polymer state. Conformal cooling and pulsed cooling are two molding technologies that have been

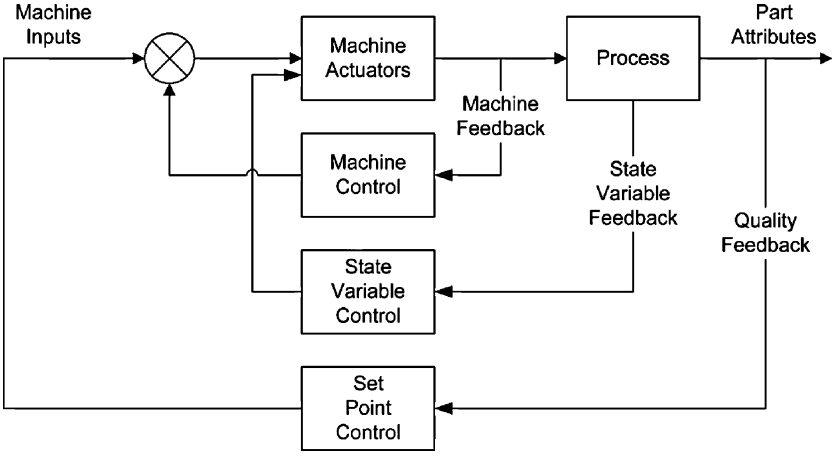


Fig. 5 A three-level hierarchical control architecture.

developed to control the state of the melt temperature in the mold cavity. In conformal cooling,[15] the injection mold is manufactured from powdered materials that are deposited in layers and selectively joined with binder from an ink-jet style printhead. Because of this manufacturing process, the mold tooling can be developed with cooling channels designed to be conformal with the molding cavity. Accordingly, such conformal cooling provides the ability to provide a uniform temperature across the molding cavity throughout the process cycle. Pulsed cooling[16] utilizes heaters to affect the surface temperature of the mold. By adding heat prior to and possibly during the molding process, it is possible in the pulsed cooling process to control boundary temperature condition at the polymer–mold interface during the subsequent injection molding cycle. However, cycle time and energy consumption may be increased because of the added heat.

Another fundamental state variable that can be regulated during the cycle is cavity pressure. Closed loop control of cavity pressure could automatically compensate for variations in melt viscosity and injection pressure to achieve a consistent process and consistency of molded products. Adaptive control methods have been developed to track the cavity pressure profile at one location in the mold.[17,18] In these earlier works, cavity pressure control was handicapped by the absence of actuators for distributed pressure control, as conventional molding machines are equipped with only one actuator (the screw), which prevents the simultaneous cavity pressure control at multiple points in the mold. This problem has been solved with the development of dynamic melt flow regulators that allow control of the flow and pressure of the polymer melt at multiple points in the mold.[19,20]

The traditional approach to machine setup in the plastics industry has been trial and error. For this, shots are taken during start-up, and part quality attributes are measured after each shot to evaluate the acceptability of produced parts. The process engineers then use their knowledge of the process to adjust the machine inputs in such a way as to improve the quality of the part from shot to shot. This tuning exercise is repeated until the specifications for part quality are satisfied. The main drawback of the traditional tuning approach is its inefficiency because of its ''ad hoc'' nature. An alternative to the traditional trial and error approach is the use of expert systems where corrective guidelines are presented in the form of if-then rules.[21] The main shortcoming of expert systems is that a generalized set of rules may not be applicable across a broad range of part geometries, material properties, and machine dynamics.

An analytical alternative for set point specification in more advanced molding operations is to develop an empirical model based on data obtained from a design of experiments (DOE).[22] Based on the developed empirical models, an objective function of an optimization problem is defined as a function of the part quality attributes. The set of inputs that produces the best set of quality attributes is obtained as the optimal point of this optimization problem. Alternatively, the Extensive Simplex Method has been developed to derive the global feasible process window using a constraint-based approach.[23] Such DOE-based methods offer a systematic approach for process setup,[24] but require significant investment in training and technology. The use of such methods, however, has increased with the increasing interest in Six Sigma and improved process capabilities.

ADVANCED MOLDING PROCESSES

While conventional injection molding is a very capable net shape manufacturing process, process variants are being continuously developed to efficiently produce molded parts with improved properties. In Fig. 6, a classification of several common variants of the injection molding process is provided. As suggested, these variants allow for the molding process to produce parts that are hollow, less dense, very thin, or composed of multiple materials. Other molding technologies have also been developed that are suitable for very high production quantities.

Several molding processes are available for the production of hollow parts. In gas assist[25,26] and water assist[27] molding, the injection mold is designed with thick flow channels. During the molding process, the polymer melt is injected and partially fills the mold. Gas or water is then injected at one or more points and forces the polymer melt to continue flowing down and out of the flow channel and into thinner adjacent sections while the gas or water cores out the molten center of the flow channel. Both processes thereby provide parts with defined external geometry and thicker sections that are coarsely hollowed out (e.g., door handles). Compared with gas assist, water assist provides faster heat transfer and improved internal surface finish, but requires handling of water from the molded part and around the molding machinery.

Injection blow molding and lost core molding are more advanced processes for producing parts with controlled external and internal geometry. In injection blow molding,[28] an injection mold is used to produce a preform with specific thickness and shape. The warm preform is then transferred to a subsequent blowing station where internal pressure is applied and forces the preform to the extremities of a differently shaped mold cavity to provide a hollow part with controlled wall thickness (e.g., soda bottles). In lost core molding,[29] a metal core with low melting temperature is cast with

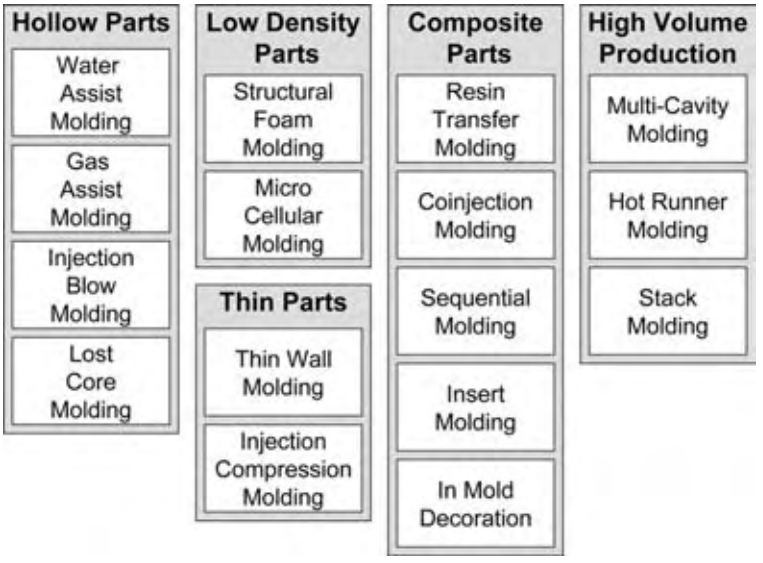


Fig. 6 Classification of various injection molding processes.

complex geometry and placed into an injection mold and then overmolded with the polymer melt per conventional injection molding. After the part is formed and cooled, the core is then melted out, thereby enabling complex internal and external part geometry with repeatable dimensional accuracy (e.g., automotive air intake manifolds) that could not otherwise be made.

Structural foam molding and microcellular molding are processes for creating molded parts with lower density. In structural foam molding[30] (also known as low pressure molding[31]), the polymer melt is used with a reactive blowing agent to create a foamed melt that is injected into the mold cavity. The process is frequently utilized with part designs that are thicker than those produced with conventional injection molding. The resulting molded parts usually exhibit a solid outer skin with an internal foamed core, thus achieving a very high stiffness-to-weight ratio. However, surface finish, foam consistency, and cycle economics are sometimes issues in application. More recently, microcellular molding[32] (also known as MuCell™) has been developed for application to thinner wall parts. In this process, a supercritical inert fluid is introduced into the polymer melt at a high pressure prior to the injection stage. During injection, the supercritical fluid reduces the melt viscosity, thereby reducing the injection pressure and clamp tonnage. With the subsequent pressure and temperature decay, the supercritical fluid rapidly changes to a gas with very small cell size and low cellular interconnectivity. With adequate process control, the resulting parts can thereby have a significant density reduction without a significant reduction in mechanical properties.

Thin wall molding and injection compression molding are processes utilized to obtain very thin molded parts. In thin wall molding, broadly defined as a mold design in which the length of flow is greater than one or more hundred times its wall thickness, resins with reduced melt viscosity are directly injected at a high pressure into a conventional injection mold.[33] Injection pressures in excess of 200 MPa and sometimes 400 MPa may be utilized to rapidly force the polymer melt throughout the mold cavity, thereby forming thin and complex parts (e.g., laptop housings). Because of the high pressures, thin wall molds are typically designed to be very stiff to avoid deflection during the molding process. In injection compression molding,[34] one side of the mold is displaced during the molding process so as to provide a larger wall thickness during injection of the polymer melt to allow for lower injection pressures. During or after the injection stage, the mold is closed to provide a reduced wall thickness while compensating for volumetric shrinkage. Accordingly, injection compression molding can be utilized to mold thin parts at lower pressures with uniform properties and reduced residual stress (e.g., compact discs[35]).

Given that assembly operations are expensive and can lead to failures, many molding processes have been developed to incorporate multiple materials and/or components into a single composite molding with improved properties and functionality. In resin transfer molding,[36] a composite mat or fabric is placed into a mold. The mold is then closed and the polymer injected into and around the composite fibers. The resulting parts (e.g., automotive body panels) usually have very good structural properties, though surface finish can be an issue. Coinjection molding[37] is another process that utilizes the serial injection of different materials into the mold cavity. Typically, a first

material with the desired esthetic and structural properties is injected to form the skin of the molding after which a second material is injected to form the core of the molding. One common application is the molding of fenders and door panels in the automotive industry with a high gloss exterior and internal recycled content.[38] To gain complete control of the distribution of multiple materials in a composite molding, sequential molding[39] (also known as two-shot or multishot molding) is commonly used. In this process, a molded part is produced in a first mold cavity with a first polymeric material. The part is then transferred to a second mold cavity into which a second polymeric material is injected, which fuses with the first molded part. The resulting parts may consist of multiple colors (e.g., automotive tail lights) or structural properties (e.g., soft grip razors).

Insert molding[40] can be considered a more generic version of sequential molding, as most components can be placed into an injection mold and overmolded with a polymeric material. Common applications of insert molding include the integration of electrical connectors in the housings of electrical devices as well as the incorporation of metal stiffeners and fasteners in structural components. Recently, in-mold decoration has become common.[41] In this process, a thin decorative film is inserted into the mold cavity and overmolded with the polymer melt. The resulting parts (e.g., cell phone faceplates) typically have complex printed patterns between the highly glossy surface and the polymer substrate.

While a single injection mold cavity may provide hundreds of thousands of moldings, applications requiring higher production volumes typically require multiple cavities to increase production rates and reduce production costs. In a multiple cavity mold, two or more cavities are connected via a runner system to the nozzle of the molding machine.[42] In a naturally balanced runner system, the mold geometry is symmetric such that the same melt flow and pressure are provided to each mold cavity. In this manner, a number of molded parts can be produced in the time normally taken to produce one molded part. Unfortunately, the runner system can utilize a significant amount of polymeric material and require extended cooling times to adequately solidify for ejection. To avoid excessive waste across many molding cycles and also to enable the operation of a higher number of mold cavities, the mold may be fitted with a hot runner system. In a hot runner system, the polymer melt is transmitted through a series of heated channels directly to the mold cavity. As a result, the polymer melt remains molten in the hot runner system and there is no wasted material or time. Typically, hot runner molding also improves molded part quality because of the improved melt transmission between the nozzle and the mold cavities.[43]

A difficulty with multicavity molding is that the force exerted by the melt onto the molding machine is proportional to the number of mold cavities. To reduce these clamping forces while operating with a higher number of cavities, stack molding was developed.[44] In stack molding, a second set of cavities is placed directly behind a first set of cavities in an injection mold. Both sets of cavities are molded at the same time, with the mold opening on two parting planes for ejection of the parts. As the molding pressures within the sets of mold cavities are opposing, two sets of mold cavities can be operated with the clamping forces required to mold only one set of cavities. As such, stack molds provide for very rapid and economical production of a large quantity of molded parts, given the additional upfront investment in molding technology.

CONCLUSIONS

While injection molding is a widely used and mature process, technological advances in computer-aided engineering and process control have brought new capabilities with respect to product design, mold making, and molding machine design. Product designs have become more complex and lightweight, with an ever-increasing number of designs incorporating composite material systems. Mold making has become extremely efficient in which the mold design utilizes current solid modeling techniques with libraries of mold components, and the mold manufacturing utilizes numerical controlled programming derived from the mold design on numerically controlled machining centers. Molding machines have become more capable and energy efficient with advances in machine components as well as easier to use with the incorporation of computer technology into user interfaces.

The selection adoption of advanced molding technologies is necessary to remain competitive. There has been a long term trend of consolidation by mergers and acquisition in the plastics industry. The resulting larger producers generally have a broader technology portfolio compared with the smaller producers. The smaller producers are responding by focusing on niche applications with a specialized technology strategy. These industry conditions are dynamic, but provide for very competitive services to customers.

ARTICLE OF FURTHER INTEREST

Process Optimization, p. 2439.

REFERENCES

1. Isayev, A.I. *Injection and Compression Molding Fundamentals;* Vol. 1, Marcel Dekker, 1987.
2. Johannaber, F. *Injection Molding Machines: A User's Guide*; Hanser Gardner Publications, 1994.
3. Rauwendaal, C. *Polymer Extrusion*, 4th Ed.; Hanser Gardner Publications, 2001.
4. Mann, J.W. Process parameter control: the key to optimization. Plast. Eng. **1974**, *30*, 25–27.
5. Lord, H.A.; Williams, G. Mold filling studies for the injection molding of thermoplastic materials: transient flow of plastic materials in the cavities of injection molding dies. Can. Controls Instrum. **1975**, 318–328.
6. Pandelidis, I.O.; Agrawal, A.R. Optimal anticipatory control of ram velocity in injection molding. Polym. Eng. Sci. **1988**, *28*, 147–156.
7. Chang, R.Y.; Tsaur, B.D. Experimental and theoretical studies of shrinkage, warpage, and sink marks of crystalline polymer injection molded parts. Polym. Eng. Sci. **1995**, *35*, 1222.
8. Zheng, R.; Kennedy, P.; Phan-Thien, N.; Fan, X.J. Thermoviscoelastic simulation of thermally and pressure induced stresses in injection molding. J. Non-Newtonian Fluid Mech. **1999**, *84*, 159–190.
9. Xu, H.; Kazmer, D.O. A stiffness criterion for cooling time estimation. Int. Polym. Process. **1999**, *13*, 249–255.
10. Trantina, G.; Nimmer, R. *Structural Analysis of Thermoplastic Components*; Hanster Gardner Publications, 1994.
11. Malloy, R. *Plastic Part Design for Injection Molding: An Introdution*; Hanser Gardner Publications, 1994.
12. Menges, G.; Michaeli, W.; Mohren, P. *How to Make Injection Molds*, 3rd Ed.; Hanser, 2001.
13. Kazmer, D.; Roser, C. Evaluation of product and process design robustness. Res. Eng. Design **1999**, *11*, 20–30.
14. Wang, K.K. Intelligent process control using on-line models. NSF Design and Manufacturing Grantees Conference, Los Angeles, CA: 1999.
15. Sachs, E.; Wylonis, E.; Cima, M.; Allen, S.; Michaels, S.; Sun, E.; Tang, H.; Guo, H. Injection molding tooling by three dimensional printing, a desktop manufacturing process. Proceedings of the society of plastics Engineers Annual Technical Conference, ANTEC, Society of Plastics Engineers: Brookfield, CT, 1995; Vol. 1, 997–1003.
16. Jansen, K.M.B.; Flaman, A.A.M. Influence of surface heating on the birefringence distribution in injection molded parts. Polym. Eng. Sci. **1994**, *34*, 898–904.
17. Gao, F.; Patterson, I.A.N.; Kamal, M.R. Self-tuning cavity pressure control of injection molding filling. Adv. Polym. Technol. **1994**, *13*, 111–120.
18. Nunn, R.E.; Grolman, C.P. Closed loop cavity pressure control in injection molding. J. Reinforced Plastics Composites **1991**, *9*, 2121.
19. Kazmer, D.O.; Barkan, P. Multi-cavity pressure control in the filling and packing stages of the injection molding process. Polym. Eng. Sci. **1997**, *37*, 1865–1879.
20. Kazmer, D.; Kudchakar, V.; Nageri, R. Performance of a self-regualating melt pressure valve. Proceedings of SPE Annual Technical Conference; 2005.
21. Shelesh-Nezhad, K.; Siores, E. Intelligent system for plastic injection molding process design. J. Mater. Process. Technol. **1997**, *63*, 458–462.
22. Myers, R.H.; Montgomery, D.C. Response surface methodology: process and product optimization using designed experiments. *Wiley Series in Probability and Statistics*; Wiley Interscience, 1995; 248 pp.
23. Zhu, L.; Kazmer, D. An extensive simplex method mapping the global feasibility. J. Eng. Optimization **2003**, *35*, 165–176.
24. Liu, C.; Manzione, L.T. Process studies in precision injection molding. I. Process parameters and precision. Polym. Eng. Sci. **1996**, *36*, 1–9.
25. Rothe, J. Improvements to injection moulding machinery. Kunststoffe German Plastics **1988**, *78*, 19–23.
26. Turng, L.S. Development and application of CAE technology for the gas-assisted injection molding process. Adv. Polym. Technol. **1995**, *14*, 1–13.
27. Liu, S.-J.; Chen, Y.-S. Water-assisted injection molding of thermoplastic materials: effects of processing parameters. Polym. Eng. Sci. **2003**, *43*, 1806–1817.
28. Abramo, R.J. Injection blow molding PVC containers. J. Vinyl Technol. **1992**, *14*, 87–92.
29. Michaeli, W.; Ziegmann, C. Micro assembly injection moulding for the generation of hybrid microstructures. Microsyst. Technol. **2003**, *9*, 427–430.
30. Wang, S.F.; Ogale, A.A. Structural foam molding. Modeling of bubble-growth process resulting from chemical blowing agents. Polym. Plast. Technol. Eng. **1990**, *29*, 355–369.
31. Shallenberg, D. Low pressure structural foam molding carry the load with style. Presented at Society of Manufacturing Engineers Annual Meeting, 1999; Vol. CM99–193, 1–13.
32. Yoon, J.D.; Hong, S.K.; Kim, J.H.; Cha, S.W. A mold surface treatment for improving surface

finish of injection molded microcellular parts. Cell. Polym. **2004**, *23*, 39–47.
33. Fassett, J. Thin wall molding: how its processing considerations differ from standard injection molding. Plast. Eng. **1995**, *51*, 35–37.
34. Yang, S.Y.; Nien, L. Experimental study on injection compression molding of cylindrical parts. Adv. Polym. Technol. **1996**, *15*, 205–213.
35. Fan, B.; Kazmer, D.O.; Bushko, W.C.; Theriault, R.P.; Poslinski, A.J. Simulation of injection-compression molding for optical media. Polym. Eng. Sci. **2003**, *43*, 596–606.
36. Lee, C.-L.; Wei, K.-H. Resin transfer molding (RTM) process of a high performance epoxy resin. Polym. Eng. Sci. **2000**, *40*, 935–943.
37. McRoskey, J. Coinjection advances open up new applications. Plast. World **1984**, *42* (9), 50–51.
38. Moss, M. Sequential coinjection hot runner. Proceedings of the society of Plastics Engineers Annual Technical Conference. ANTEC, Society of Plastics Engineers: Brookfield, CT, 1998; Vol. 1, 351–357.
39. Wilder, R.V. Multicomponent molding. Now a strategic capability Modern Plast. **1990**, *67*, 65–69.
40. Strasser, F. Molding metal inserts into plastic parts. Plast. Eng. **1980**, *36*, 17–22.
41. Enewoldsen, P.; Braun, H. In-mould decoration. Kunststoffe Plast Europe **1999**, *89*, 35–36.
42. Beaumont, J. *Runner and Gating Design Handbook: Tools for Successful Injection Molding*; Hanser, 2004.
43. Ohnuma, S. 64-Impression hotrunner mould for TPE cuff seals. Kunststoffe German Plastics **1989**, *79*, 17.
44. Strauch, R. Development and construction of a stack mold. Kunststoffe German Plastics **1988**, *78*, 15–16.

Ion Exchange

I

Sukalyan Sengupta
Civil and Environmental Engineering Department, University of Massachusetts, Dartmouth, Massachusetts, U.S.A.

Arup K. Sengupta
Department of Civil and Environmental Engineering, Lehigh University, Bethlehem, Pennsylvania, U.S.A.

INTRODUCTION

Ion exchange is a reversible chemical reaction where—in the most commonly used form—an ion from solution is exchanged with an equivalent amount of another ion of the same charge that is attached to an immobilized solid phase. However, presently the application of this technology has diversified to deal with gaseous streams as well as solids. The immobilized solid-phase particles are either naturally occurring inorganic zeolites or synthetic, organic ion-exchange resins. The primary morphology of ion exchangers is spheres of diameter 75–1000 μm, but membranes, composite sheets, and nanoparticles are increasingly being used for specialty applications. In most applications, the ion being removed from solution is a contaminant and is replaced by a benign ion from the solid phase, but this technology can easily be adapted to concentrate valuable ions also. The vast potential of this technology is derived from its ability to generate specific ion exchangers that have a high affinity for target ions. Thus selectivity of the ion exchanger plays a critical role and it allows engineers to design ion-exchange processes where the target ion is almost completely removed from the influent and the effluent contains ''nondetect'' amounts of it. Moreover, ion-exchange technology is almost always employed as a cyclical process. Therefore, the ability of a regenerating solution to desorb the target ion from the surface of the ion exchanger and reuse the ion exchanger for another exhaustion cycle is the critical factor in its success.

EVOLUTION OF ION EXCHANGE

The phenomenon of ion exchange was observed and scientifically documented for the first time by Way[1] and Thompson[2] in 1850. They confirmed the ion-exchange properties of soils and simulated the following naturally occurring cation-exchange reactions for calcium–potassium and calcium–ammonia systems:

$$2K^+(aq) + \overline{Ca^{2+}\text{–soil}} \Leftrightarrow \overline{(K^+)_2\text{–soil}} + Ca^{2+}(aq) \quad (1)$$

$$2NH_4^+(aq) + \overline{Ca^{2+}\text{–soil}} \Leftrightarrow \overline{(NH_4^+)_2\text{–soil}} + Ca^{2+}(aq) \quad (2)$$

Besides validating the occurrence of ion exchange, the experimental results of Waley and Thompson also established the preference of soil toward ammonium and potassium over calcium—a phenomenon that helped explain how ion exchange plays a key role in plants' uptake of nutrients from soils. In principle, the ion exchangers present in soil with varying selectivities toward different ions opened a new avenue to separate dissolved species by a simple filtration process.

The evolution of the ion-exchange process, however, happened rather slowly during the 19th and 20th centuries because of the difficulty associated with modifying or manipulating naturally occurring, inorganic clayey materials with low ion-exchange capacity. Also, the scientific understanding of ion-exchange fundamentals lagged well behind its application potential. It was gradually recognized that the primary factors that influence the relative selectivities of ions for such inorganic exchangers are:

1. Electrostatic (or coulombic) interaction between the ion-exchange sites and dissolved ions.
2. Hydrated ionic radii, which in turn are dependent on the solvation energy of the relevant ions.

Inorganic zeolites (synthetic or naturally occurring aluminosilicates) later found widespread application in softening hard water, i.e., removal of dissolved polyvalent cations, principally calcium and magnesium, through exchange with sodium.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030553

However, anion-exchange processes still remained unexplored and practically unobserved.

The immense potential of ion-exchange technology hit a new high when the first organic-base (polymeric) cation exchanger was synthesized by Adams and Holmes in 1953.[3] In less than 10 years, D'Alelio prepared the first polymeric anion exchanger.[4] Since then, synthesis of new ion exchangers has continued at an unceasing rate, and applications of ion exchangers in industries as diverse as power utilities, biotechnology, agriculture, pharmaceutics, chemicals, microelectronics, environmental treatment, etc. are growing. Ironically, the World War II and, more specifically, the race for nuclear technology helped catalyze the growth and maturity of the field of ion exchange at an accelerated pace. Ion exchange was found to be a viable process for separating some of the critical transuranium elements and, for understandable reasons, its application aroused lot of interest. In fact, some of the most fundamental studies on ion-exchange equilibria and kinetics were carried out during World War II and reported afterward in technical literature.[5,6] Since then, the understanding of ion-exchange fundamentals has kept good pace with new application frontiers. In this regard, the publication of the landmark *Ion Exchange* by Friedrich Helfferich in 1962 provided a strong theoretical framework, especially for scientists/engineers working in this field.[7]

ION-EXCHANGE THEORY

In an ion-exchange reaction between the aqueous phase and the polymeric ion exchanger, the latter can be viewed as a strongly ionized electrolyte with a relatively low dielectric constant. Consequently, attempts have been made to mimic specific aqueous-phase interactions in the solid phase. The following interactions are often operative in selective ion-exchange processes:

1. Electrostatic (or coulombic type).
2. Hydrophobic (or Van der Waals type).
3. Lewis acid–base (metal ligand type).
4. Ion–dipole.
5. Dipole–dipole.
6. Ion sieving.
7. Ion exclusion.
8. Steric effect.

Many of these interactions can be characterized and enhanced by modifying or tailoring the following composition variables of the ion exchangers:

- Polymer matrix.
- Covalently attached functional groups.
- Crosslinking.
- Porosity.

Also, for the ion-exchange processes to be viable, the ion exchangers should be amenable to regeneration or desorption so that they may be used for hundreds of cycles. In fact, the overall economy of an ion-exchange process is often dictated by the operating costs of the regenerant chemicals as opposed to the fixed cost of the ion exchangers. Thus, ideally, an ion-exchanger process should be reversible so that the target solutes can be desorbed efficiently, thus leading to energy-efficient separation. In reality, however, efficiency of desorption (or regeneration) tends to diminish for highly selective sorbents because of the strong bonding (or high free energy change) between the target ion and the sorbent.

To strike a balance between selectivity and regenerability, the intensity of solute–sorbent interaction has to lie within an envelope where ion exchange is selective but at the same time reversible. Fig. 1[8] helps quantify such a working regime for various types of interactions. The importance of reversibility is also well recognized for homogeneous separation processes using specialty complexing agents.[9] It is worthwhile to note that more than one interaction may be operative for the sorption (ion exchange) of a single solute. In such cases, individual free energy changes are additive and may enhance the overall sorption affinity. For example, perchlorate and chloride are identical electrostatically, i.e., both are monovalent anions. But perchlorate exhibits much higher affinity toward an anion exchanger with polystyrene matrix and tributyl functional group because of accompanying hydrophobic bonding between perchlorate and the resin matrix.

ION-EXCHANGE EQUILIBRIUM

The preference of the ion exchanger for the target ion is often expressed as a separation factor, α_{ij}, or a selectivity coefficient, K_{ij}, for binary exchange.[10] Consider, for example, the cation-exchange reaction of Na^+ for H^+ in a cation exchanger:

$$\overline{R^-H^+} + Na^+ \leftrightarrow \overline{R^-Na^+} + H^+ \tag{3}$$

$$K = \frac{\{\overline{R^-Na^+}\}\{H^+\}}{\{\overline{R^-H^+}\}\{Na^+\}} \tag{4}$$

In Eqs. (3) and (4), overbars denote the ion-exchanger phase. For a cation-exchange reaction as shown here, the exchangeable ions (Na^+ and H^+) are termed "counterions" and the negatively charged ions in solution are called "coions." K is the thermodynamic equilibrium constant, and {} denotes ionic activity. For convenience of measurement, concentrations are used in practice in place of activities. In this case, in Eq. (4), based on concentration, the selectivity

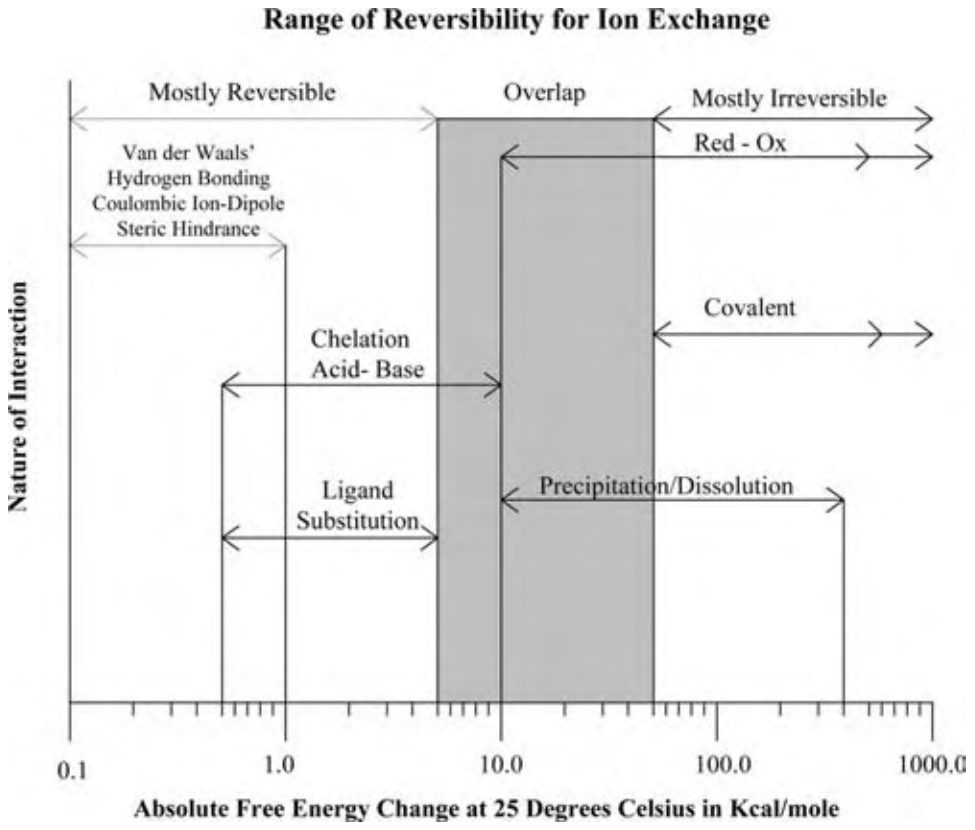


Fig. 1 A quantitative measure of various interactions in ion-exchange type sorption processes.

coefficient $K_{Na/H}$ describes the exchange. Note that $K_{Na/H}$ includes activity coefficient terms that are functions of ionic strength and thus is not a true constant, varying with different ionic strengths.

$$K_{Na/H} = \frac{[\overline{R^-Na^+}][H^+]}{[\overline{R^-H^+}][Na^+]} = \frac{q_{Na}C_H}{q_H C_{Na}} \quad (5)$$

where [ ] is the concentration (mole/L), q_{Na} the resin-phase concentration of sodium (equiv./L), and C_{Na} the aqueous-phase concentration of sodium (equiv./L).

The binary separation factor $\alpha_{Na/H}$ is a useful parameter that describes the ion-exchange equilibria because it is dimensionless:

$$\alpha_{ij} = \frac{\text{distribution of ion } i \text{ between phases}}{\text{distribution of ion } j \text{ between phases}} = \frac{y_i/x_i}{y_j/x_j} \quad (6)$$

Thus for the example above

$$\alpha_{Na/H} = \frac{y_{Na}/x_{Na}}{y_H/x_H} = \frac{y_{Na}x_H}{x_{Na}y_H} = \frac{(q_{Na}/q)(C_H/C_T)}{(C_{Na}/C_T)(q_H/q)} = \frac{q_{Na}C_H}{q_H C_{Na}} \quad (7)$$

where y_{Na} is the equivalent fraction of Na in resin (q_{Na}/q), y_H the equivalent fraction of H in resin (q_H/q), q_{Na} the concentration of Na on resin (equiv./L), q_H the concentration of H on resin (equiv./L), q the total exchange capacity of resin (equiv./L), x_{Na} the equivalent fraction of Na in water (C_{Na}/C_T), x_H the equivalent fraction of H in water (C_H/C_T), and C_T the total ionic concentration of water (equiv./L).

Eqs. (5) and (7) show that for homovalent exchange, the separation factor α_{ij} and the selectivity coefficient K_{ij} are equal. For our example, this translates as

$$K_{Na/H} = \frac{q_{Na}C_H}{q_H C_{Na}} = \alpha_{Na/H} \quad (8)$$

However, for heterovalent ion exchange, the separation factor is not equivalent to the selectivity coefficient. If we consider the case of calcium–sodium exchange,

$$\overline{2(R^-Na^+)} + Ca^{2+} \leftrightarrow \overline{(R^-)_2Ca^{2+}} + 2Na^+ \quad (9)$$

$$K_{Ca/Na} = \frac{q_{Ca}C_{Na}{}^2}{q_{Na}{}^2 C_{Ca}} \quad (10)$$

while

$$\alpha_{Ca/Na} = \frac{y_{Ca}/x_{Ca}}{y_{Na}/x_{Na}} \quad (11)$$

Combining Eqs. (10) and (11) we obtain,

$$\alpha_{Ca/Na} = K_{Ca/Na}\frac{q/C_T}{x_{Na}/y_{Na}} \quad (12)$$

Eq. (12) clearly shows that for heterovalent ion exchange, the separation factor can be manipulated through the total ionic concentration of the solution, C_T. The capacity of the resin, q, is a constant. The higher the C_T, the lower the $\alpha_{Ca/Na}$; i.e., selectivity tends to reverse in favor of the monovalent ion as the ionic strength increases. There is a general selectivity sequence for a typical cation or anion exchanger as shown in Table 1. Based on Table 1, it can be understood that a sodium-loaded cation exchanger can be used to remove a divalent cation such as Ca^{2+} from a dilute concentration of water—as is the case with hard water—as selectivity for $Ca^{2+} > Na^{+}$. When the resin is exhausted, a very high concentration Na^{+} solution (as NaCl) is passed through the resin bed.

Table 1 Relative selectivity of various counterions for a strong-acid cation exchanger

Counterion	Relative selectivity for AG 50W-X8 resin
H^{+}	1.0
Na^{+}	1.5
NH_4^{+}	1.95
K^{+}	2.5
Cu^{+}	5.3
Ag^{+}	7.6
Mn^{2+}	2.35
Mg^{2+}	2.5
Fe^{2+}	2.55
Zn^{2+}	2.7
Co^{2+}	2.8
Cu^{2+}	2.9
Cd^{2+}	2.95
Ni^{2+}	3.0
Ca^{2+}	3.9
Sr^{2+}	4.95
Hg^{2+}	7.2
Pb^{2+}	7.5
Ba^{2+}	8.7

AG 50W-X8 is a strong-acid cation exchanger with sulfonic acid functional group attached to a styrene-divinylbenzene copolymer matrix with 8% divinylbenzene and is available from Bio-rad Inc., California.

Based on Eq. (12), we see that as the C_T is increased, $\alpha_{Ca/Na}$ is decreased and for a certain C_T value, $\alpha_{Ca/Na} = 1.0$. If C_T is increased further, $\alpha_{Ca/Na} < 1.0$, i.e., selectivity reversal takes place and now the resin prefers Na^{+} to Ca^{2+}, which means that Ca^{2+} is easily eluted from the ion exchanger and it is back in Na^{+} form, where it can be used for another exhaustion cycle. This is best explained by the ion-exchange isotherm shown in Fig. 2. An ion-exchange isotherm is a constant temperature equilibrium plot of ion-exchanger-phase concentration vs. aqueous-phase concentration. A "favorable'' isotherm (convex to the x-axis) means that species i (which is plotted on each axis) is preferred to species j, the hidden or exchanging species. Thus if the isotherm is a 45° line, it means that the ion exchanger has no preference of i over j. And if the curve is concave to the x-axis, it indicates that j is preferred to i.

ION-EXCHANGE PROCESSES

Softening

The softening of hard waters (containing polyvalent cations) by ion exchange entails the substitution of polyvalent cations (principally Ca^{2+} and Mg^{2+}) by Na^{+}. This eliminates the undesirable characteristics of hard water, such as not lathering well, causing a scum in plumbing fixtures, and leaving hard, crusty scales on hot water heaters, coffee pots, etc. If the water has bicarbonate alkalinity, the ion-exchange reaction may be written as:

$$\overline{2R^{-}Na^{+}} + Ca(HCO_3)_2 \leftrightarrow \overline{(R^{-})_2Ca^{2+}} + 2NaHCO_3 \quad (13)$$

In a typical softening scenario, the combination of the facts that the influent is dilute (ionic strength of a few mequiv./L) and that divalent ions are more favored by the ion exchanger vis-à-vis monovalent ions results in almost complete migration of the target ion (Ca^{2+} in this case) from the solution to the ion-exchanger phase. When the ion-exchange capacity of the exchanger is reached, the ion-exchanger bed is backwashed with clean water at 6–8 gpm/ft^2 of bed area to expand the bed by about 50%. Then regeneration of the ion exchanger back to the sodium form is achieved by passing through the ion exchanger a strong brine solution (≈10%) to effect selectivity reversal. About 150–200% of theoretical amount of brine is needed to regenerate the exchanger. The regeneration cycle must be followed by a rinse cycle to remove the remaining brine from the bed. The total amount of rinse water needed is 20–35 gallons per cubic foot of

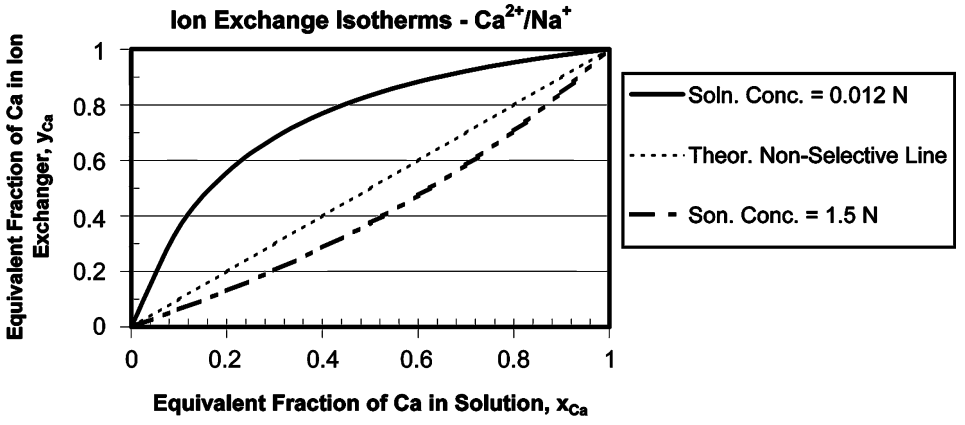


Fig. 2 Ion-exchange isotherm for Ca^{2+}/Na^{+} exchange showing selectivity reversal at high solution concentration.

resin. The total volume of the brine that is used during a regeneration cycle, together with the rinse water that follows, varies from 1.5% to 7% of the amount of water softened by the unit. The chloride concentration in this wastewater could be as high as 35,000–45,000 mg/L; therefore disposal of the spent brine becomes an important criterion. The steps in a conventional ion-exchange operation are shown in Fig. 3.

Demineralization

In many application scenarios such as high-pressure boilers, production of electronic components, synthesis of pharmaceuticals, etc. it is desirable to remove all the dissolved ions from the water. This process is termed demineralization. The primary use of ion exchangers is in demineralization because it can achieve the goal at low cost, and the quality is comparable to that by distillation. The process depends on two reaction stages. In the first, all cations are removed and replaced by H^{+}, using a hydrogen-loaded cation exchanger. In the second stage, acids produced are removed by a hydroxide-loaded anion exchanger. The cation exchanger is regenerated with an acid and the anion exchanger with an alkali. The reactions are

Cation-exchange step:

$$\overline{R^{-}H^{+}} + NaCl \leftrightarrow \overline{R^{-}Na^{+}} + HCl \quad \text{(service)} \tag{14}$$

$$\overline{2R^{-}Na^{+}} + H_2SO_4 \leftrightarrow \overline{2R^{-}H^{+}} + Na_2SO_4 \quad \text{(regeneration)} \tag{15}$$

Anion-exchange step:

$$\overline{R\text{–}CH_2N^{+}(CH_3)_3OH^{-}} + HCl \leftrightarrow \overline{R\text{–}CH_2N^{-}(CH_3)_3Cl^{-}} + H_2O \quad \text{(service)} \tag{16}$$

$$\overline{R\text{–}CH_2N^{+}(CH_3)_3Cl^{-}} + NaOH \leftrightarrow \overline{R\text{–}CH_2N^{-}(CH_3)_3OH^{-}} + NaCl \quad \text{(regeneration)} \tag{17}$$

Another major contaminant of pure water is silica. It is removed in the demineralization process by a strong-base anion exchanger in hydroxide mode.

Silica removal:

$$\overline{R\text{–}CH_2N^{+}(CH_3)_3OH^{-}} + H_4SiO_4 \leftrightarrow \overline{R\text{–}CH_2N^{+}(CH_3)_3H_3SiO_4^{-}} + H_2O \quad \text{(service)} \tag{18}$$

With a silica-loaded exchanger, a unique reaction of polymerization of silica inside and its migration into the interior of the exchanger also takes place that can

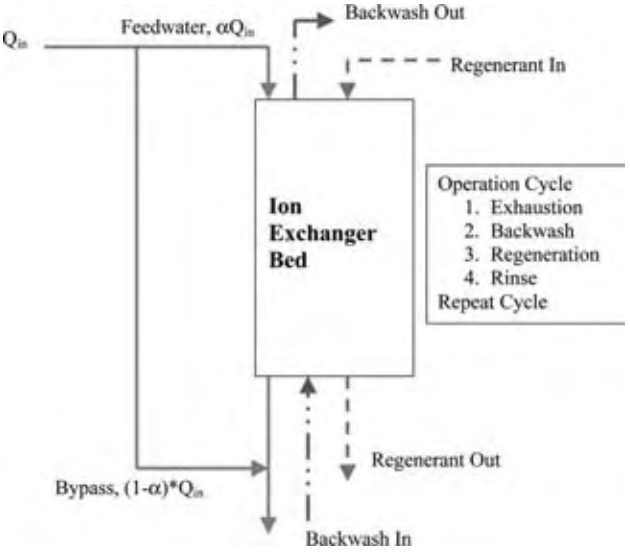


Fig. 3 Schematic of a conventional ion-exchange cycle. *(View this art in color at www.dekker.com.)*

be represented by the following reaction:

$$\overline{R\text{–}CH_2N^+(CH_3)_3H_3SiO_4^-} + xH_4SiO_4 \leftrightarrow \overline{R\text{–}CH_2N^+(CH_3)_3H_3SiO_4^- \cdot xH_4SiO_4} \quad \text{(polymerization)} \tag{19}$$

The exchanger is regenerated by a strong alkali:

$$\overline{R\text{–}CH_2N^+(CH_3)_3H_3SiO_4^- \cdot xH_4SiO_4} + NaOH \leftrightarrow \overline{R\text{–}CH_2N^+(CH_3)_3OH^-} + NaH_3SiO_4 + xH_4SiO_4 \tag{20}$$

ION-EXCHANGE MEMBRANE

Ion-exchanger membranes have been in use since the 1950s. A ''membrane,'' according to the usual definition, ''is a solid or liquid film or layer with a thickness that is small compared to its surface.''[7] In the case of ion-exchanger membranes, a broader definition has come into use. It includes any ion-exchange material, irrespective of its geometrical form, which can be used as a separation wall between two solutions. Two types of ion-exchanger membrane are in use: ''homogeneous'' and ''heterogeneous.'' ''Homogeneous'' membranes are coherent ion-exchanger gels in the shape of disks, ribbons, etc. Their structure is that of the usual ion-exchange resins. They are homogeneous only in dimensions that are large compared with the mesh width of the matrix. ''Heterogeneous'' membranes consist of colloidal ion-exchanger particles embedded in an inert binder (polystyrene, polyethylene, wax, etc.; see Fig. 4. Their mechanical stability is superior, but their electrochemical properties such as conductivity and barrier action are not as good as those of the homogeneous membranes.[11] However, at present, the combination of a strong fabric and a chemically stable binder has made heterogeneous membranes the material of choice in many industrial processes. Table 2 provides details of an anion-exchanger and a cation-exchanger membranes that are commercially available from Sybron Chemicals, Inc., a Bayer Company (no commercial endorsement implied).

The ability of ion-exchanger membranes to act as a separation wall, along with the chemical and electrochemical properties of the ion exchangers, provides manifold opportunities toward selective separation and movement of ions. The principal phenomenon in any ion-exchanger membrane is Donnan coion exclusion. Consider an ion-exchanger membrane that separates two solutions containing ions (both cation and anion) of different concentration. If the ions were without any charge, these concentration differences would be leveled out by diffusion. However, in the case of ions, such a process would disturb electroneutrality. For example, if a cation-exchanger membrane separates two solutions, any migration of anion from the solution to the membrane would result in the accumulation of negative charge in the membrane. The first few anions that diffuse into the membrane thus build up an electric potential between the two phases. This so-called ''Donnan potential'' pulls anions back into the solution. Thus, when in contact with electrolyte solutions of low or moderate concentrations, the membrane contains a large number of counterions but relatively few coions. Counterions are admitted to the membrane and have little difficulty in passing through from one solution to the other. Coions, on the other hand, are rather efficiently excluded from the membrane and thus find it difficult to pass through. In other words, the membrane is perm selective for counterions.

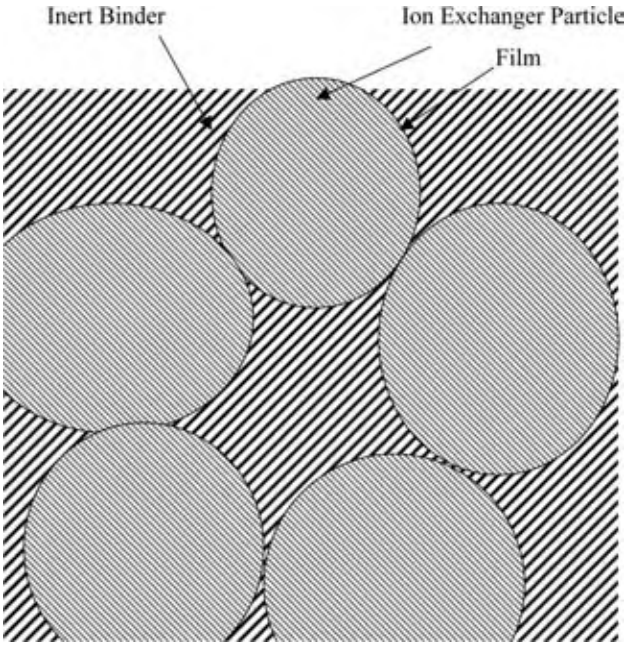


Fig. 4 Structure of a heterogeneous membrane (schematic). Colloidal ion-exchanger particles (1–10 μm in diameter) are embedded in an inert binder. (From Ref.[7].)

ELECTRODIALYSIS

Electrodialysis is an ion-exchanger membrane process used to treat brackish water. In this process, electric energy is used to transfer ionized salts from feedwater through membranes, leaving behind a purified product water.[12] In this process, cation- and anion-selective membranes are placed alternately between two electrodes (cathode and anode), as shown in

Table 2 Properties of two ion-exchanger membranes

Type	Cation	Anion
Brand name	MC-3470	MA-3475
Width (m)	1.09	1.22
Length (m)	3.1	3.1
Thickness (mm)	15	16
Exchange capacity (mequiv./g)	1.4	0.9
Mullen Burst test (bar)	10.3	10.3
Area resistance (Ω/cm)		
0.1 N NaCl	25	50
1.0 N NaCl	10	25
Permselectivity (0.5 N NaCl)	96	99
Water permeability (ml/hr/ft^2 at 5 psi)	25	25
Temperature stability (max.; °C)	80	80
Chemical stability (pH)	1–10	1–10
Current density (max.; A/ft^2)	50	50
Ionic form (as shipped)	Sodium	Chloride

Manufacturer: Sybron Chemicals, Inc., a Bayer Company, Birmingham, New Jersey.
Source: http://www.ion-exchange.com/products/membranes/index.html (accessed April 2004).

Fig. 5. These cation- and anion-selective membranes are separated by cut-out plastic spacers to form what is known as a membrane stack. A typical membrane stack has several hundred cell pairs, with each pair consisting of one ''dilute'' and one ''concentrated'' cells as shown in Fig. 6. The electrodes are charged by a direct current source. The anions and cations move through the respective selective membranes to form concentrated and dilute/pure solutions in the cell between alternating membranes. The concentrated and pure solutions are collected via separate piping manifolds for ultimate disposal or use. A patented process by Ionics of Watertown, Massachusetts, termed electro dialysis reversal (EDR) provides a

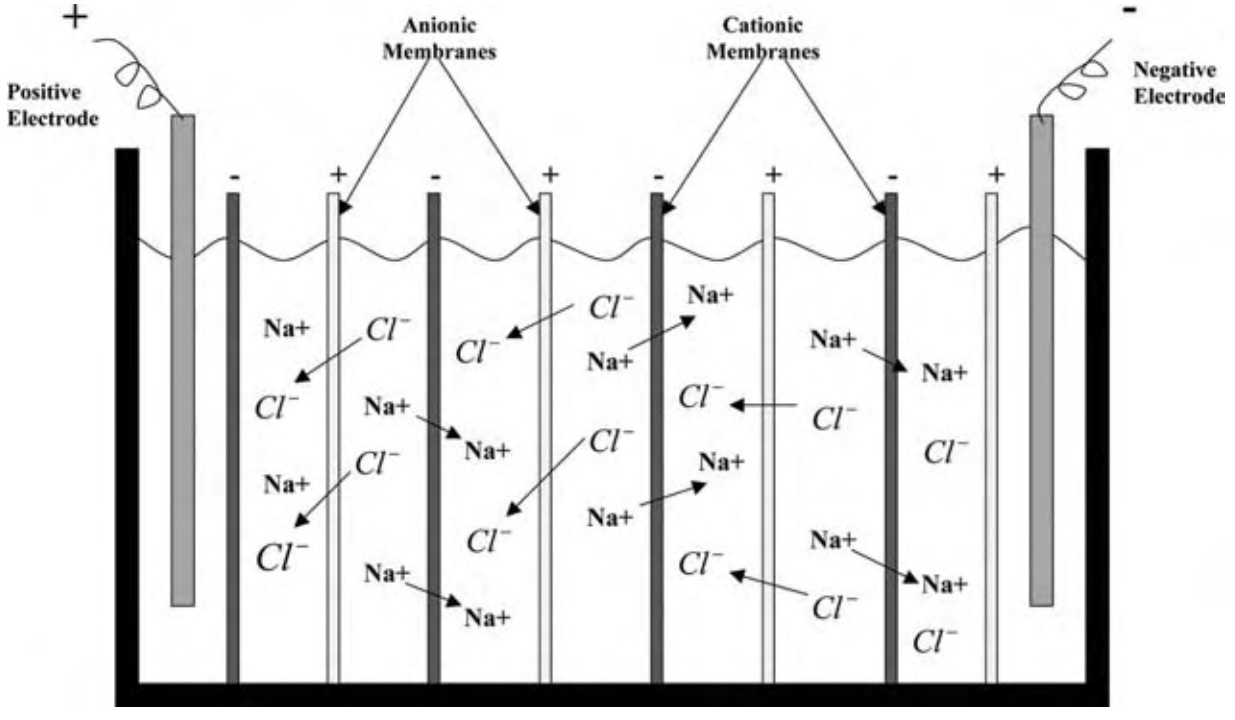


Fig. 5 Schematic of electrodialysis showing the electrodes and membrane stack. *(View this art in color at www.dekker.com.)*

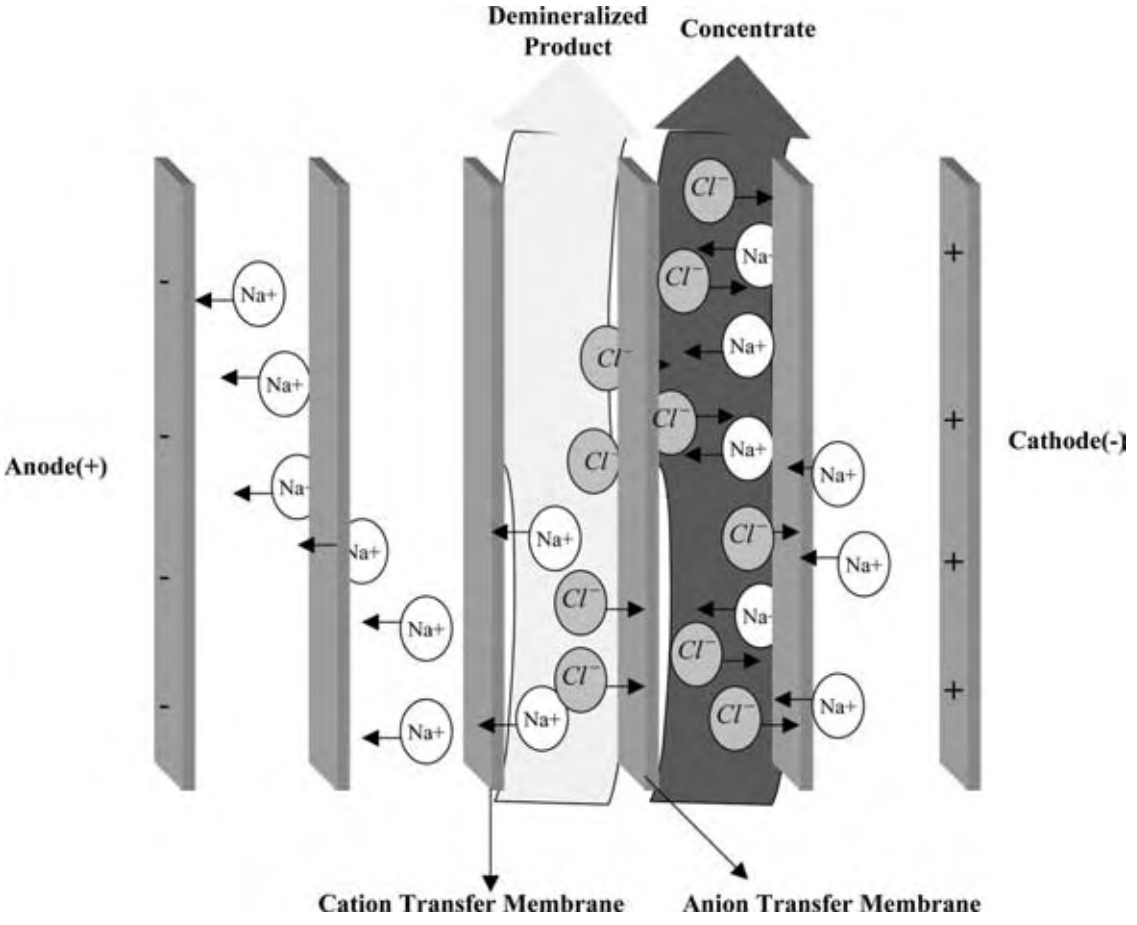


Fig. 6 Schematic of electrodialysis showing concentrate and product water. *(View this art in color at www.dekker.com.)*

self-cleaning capability that enables desalting of scaling or fouling waters and recovery of up to 94% of the feed as product water.

FLUXES IN MEMBRANE SYSTEMS

The overall flux J_i of an arbitrary species i is composed of three additive terms: the diffusion flux $(J_i)_{\text{diff}}$ caused by the chemical potential gradient of the species (Fig. 7), the electric transference $(J_i)_{\text{el}}$ caused by the electric potential gradient, and the transfer $(J_i)_{\text{con}}$ caused by convection. The diffusion flux is

$$\begin{aligned}(J_i)_{\text{diff}} &= (-)D_i C_i \operatorname{grad} \mu_i \\ &= (-)D_i(\operatorname{grad} C_i + C_i \operatorname{grad} \ln f_i)\end{aligned} \tag{21}$$

The electric transference is

$$\begin{aligned}(J_i)_{\text{el}} &= (-)u_i z_i C_i \operatorname{grad} \phi \\ &= (-)D_i z_i C_i \frac{\Gamma}{RT} \operatorname{grad} \phi\end{aligned} \tag{22}$$

and the transfer by convection is

$$(J_i)_{\text{con}} = C_i \beta \tag{23}$$

where β is the rate of motion of the center of gravity, T the absolute temperature, R the universal gas constant, Γ Faraday's constant, φ the electric potential, C_i the concentration of the ion, z_i the electrochemical valence of the ion, D_i the diffusion coefficient of the ion, u_i the electrochemical mobility of the ion, μ_i the chemical potential of the ion, and f_i the molar activity coefficient of the ion.

Thus the overall flux is given by

$$\begin{aligned}J_i &= (J_i)_{\text{diff}} + (J_i)_{\text{el}} + (J_i)_{\text{con}} \\ &= (-)D_{\text{i}}\left(\operatorname{grad} C_i + z_i C_i \frac{\Gamma}{RT} \operatorname{grad} \phi \right. \\ &\qquad \left. + C_i \operatorname{grad} \ln f_i\right) + C_i \beta\end{aligned} \tag{24}$$

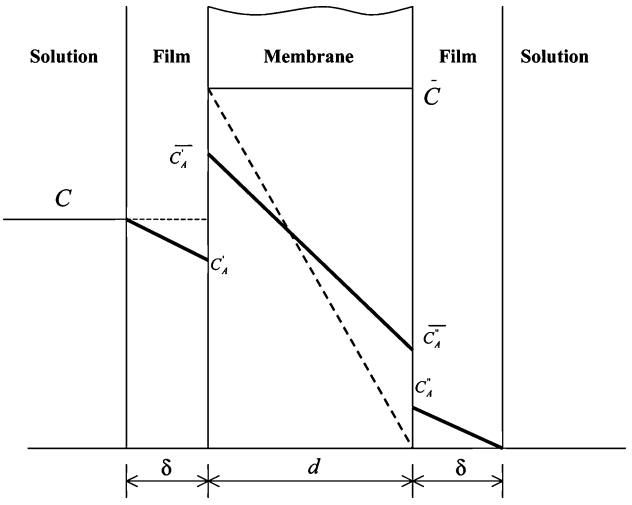


Fig. 7 Schematic of counterion diffusion across ion-exchange membrane.

DONNAN DIALYSIS

Donnan dialysis is based on the same principles as electrodialysis except that there is no external potential gradient applied; the only physical force causing movement of ions is diffusion. It has been used in wastewater and drinking water treatment, and in hydrometallurgical operations.[13–18] Its main advantage is that there is no need for an external electric field or a pressure gradient. Thus it is a cheap and a passive process, making it extremely easy to operate. It has been used mainly to concentrate (and recover) ions from a lower to a higher concentration, primarily by forcing the movement of a benign counterion in the opposite direction. For example, in Fig. 8, Donnan dialysis is used to concentrate NH_4^+ and PO_4^{3-} ions from sewage sludge. The sludge is placed in the central chamber. On one side of the sludge chamber is an acid chamber and on the opposite side is the alkali chamber. Both the side chambers are of much smaller volume than the sludge chamber. The acid chamber is separated from the sludge chamber by a selective cation-exchange membrane. This membrane only allows cations to pass through and is impermeable to anions. Similarly, the alkali chamber is separated from the sludge chamber by a selective anion-exchange membrane that only allows anions to pass through and is impervious to cations. A strong mineral acid is placed in the acid chamber. Because the acid chamber has a very high H^+ concentration and the sludge chamber is almost neutral, there is a high gradient of H^+ established and H^+ will tend to move from the acid chamber to the sludge chamber. At the same time, the sludge chamber has a very high concentration of NH_4^+, but the acid chamber has no NH_4^+ ion in the beginning. Therefore, to even out this high difference of concentration, NH_4^+ will move from the sludge chamber to the acid chamber. To maintain electroneutrality, each equivalent of H^+ that moves across the membrane in one direction has to be equal to an equivalent of NH_4^+ that moves in the opposite direction. Please note that NH_4^+ being a cation, it cannot move across the anion membrane. Similarly, the alkali chamber will contain a strong alkali in the beginning. This will set up a similar situation for the PO_4^{3-} to move from the sludge chamber to the alkali chamber and this will be equal to the equivalents of OH^- that will move from the alkali chamber to the sludge chamber. This process will go on until the concentrations of H^+ and NH_4^+ are partitioned to equilibrium in the sludge and the acid chambers, and the concentrations of OH^- and PO_4^{3-} are partitioned to equilibrium in the sludge and the alkali chambers. The net result is that NH_4^+ and PO_4^{3-} are concentrated in the acid and the alkali chambers, respectively. As the acid and alkali chambers have much smaller volume than the sludge chamber, the equilibrium concentrations of NH_4^+ and PO_4^{3-} in the acid and the alkali chambers are much higher than the final concentration in the sludge chamber. Please note that because of the low concentration of the heavy metals in the sludge and the selective nature of the cation-exchange

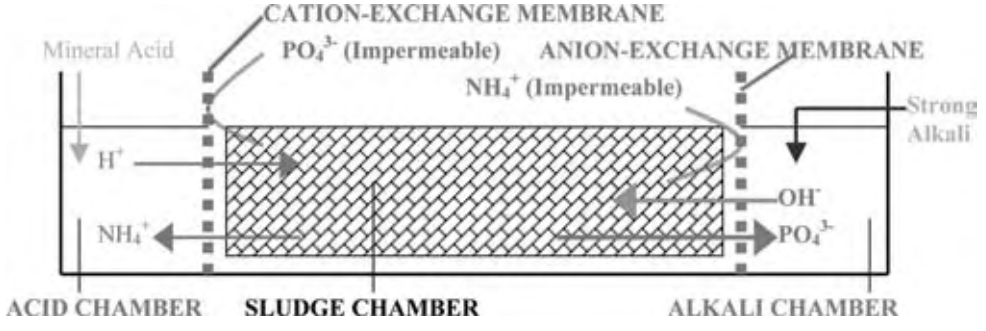


Fig. 8 Schematic diagram of Donnan dialysis process. *(View this art in color at www.dekker.com.)*

membrane, heavy metals remain in the sludge. The details of chemical equations are as follows:

When equilibrium is achieved, the redistributed NH_4^+ and H^+ will satisfy the following equation:

$$\left(\frac{C_{NH_4,AC}}{C_{NH_4,SC}}\right)^{z_H} = \left(\frac{C_{H,AC}}{C_{H,SC}}\right)^{z_{NH_4}} \quad (25)$$

The diffusion flux and the electric transference flux are provided in Eqs. (21) and (22). The convective flux is zero. The principle of electroneutrality mandates that the concentration of all counterions is equal to those of fixed ions, viz., the exchange capacity of the membrane (Q):

$$z_H q_H + z_{NH_4} q_{NH_4} = Q \quad (26)$$

Because of the condition of zero current, the sum of flux of feed ions J_H and that of driving ions J_{NH_4} is zero.

$$z_H J_H + z_{NH_4} J_{NH_4} = 0 \quad (27)$$

On solving the above equations simultaneously, the flux for ammonium transference can be determined. A similar set of equations can be generated for phosphate–hydroxide exchange. Table 3 shows data of preliminary, unpublished study conducted by the lead author in this area.

The advantages of this process are as follows:

1. It selectively extracts N and P from the sludge in a concentrated form that can be used as a mineral fertilizer.
2. The heavy metals and organic contaminants remain in the sludge.
3. It is a simple and passive process that does not require strict supervision.
4. The ion-exchange membranes (cation and anion) do not need any regeneration. Just a slight rinse after each cycle is adequate for its sustained performance.

COMPOSITE ION-EXCHANGE MATERIAL

Until now, our discussion was confined to using ion exchangers for removal of selective ions from a pure aqueous phase, i.e., there were no suspended solids in the wastestream, or using ion-exchanger membranes in Donnan dialysis. The ion-exchange resins involved were granular or spherical. With ion-exchanger membranes, they could be used with sludges and high-suspended solids only in Donnan dialysis, where the mixture is undisturbed. Lately, however, we have been encountering challenging environmental separation problems where the medium is essentially a slurry or sludge with high-suspended solid content. A vast new area of application would open up if the morphology of ion-exchanging polymers can be modified to remove selective ions from a solid-phase background. For example, a widespread environmental problem is the disposal of sludges or treatment of soil contaminated with minor fraction (often less than 5%) of heavy metals in the solid phase in an otherwise innocuous background of materials that are not important from a regulatory viewpoint. This problem stems from the fact that the heavy metals present can cause the sludge/soil to be designated as a ''hazardous'' waste, thus greatly increasing the cost of disposal. Selective and targeted removal of the heavy metals from the background solid phase would constitute an efficient treatment process as it would render the sludge non-hazardous, and may make it possible for the heavy metals to be concentrated and recycled/reused. The physical configuration of conventional ion exchangers (spherical or granular) makes their use inappropriate for such a case. A material identified in Refs.[19–26] shows potential of being appropriate under such conditions. It is a new class of sorptive/desorptive composite ion-exchange material (CIM) available commercially as thin sheets (approximately 0.5 mm thick) and is suitable for heavy-metal decontamination from sludges/slurries. The morphology of the material—along with its physical texture and tensile strength—makes it compatible for use with sludges/slurries. This property

Table 3 Results of preliminary, unfunded research

Time (days)	Nitrogen concentration in sludge chamber (mg/L)	Nitrogen concentration in acid chamber (mg/L)	Phosphorus concentration in sludge chamber (mg/L)	Phosphorus concentration in alkali chamber (mg/L)
0	2800	0	1000	0
1	2700	1200	900	800
2	2600	1800	850	1200
3	2550	2400	800	1900
5	2500	2800	760	2200
7	2450	3400	710	2800

of the material, when combined with the adapting of the chemistry of the reactor and/or creating an electric gradient induced ion-transport process, makes it possible to selectively remove heavy metals from a solid-phase background composed primarily of nonregulated elements. The CIM is a thin sheet prepared by comminuting a crosslinked polymeric ion exchanger to a fine powder and fabricating it mechanically into a microporous composite sheet consisting of ion-exchange particular matter enmeshed in polytetrafluoroethylene (PTFE). During this mechanical process, the PTFE microspheres are converted into microfibers that separate and enmesh the particles.[19] When dry, these composite sheets consist of >80 wt.% particles (polymeric ion exchanger) and <20 wt.% PTFE. They are porous (usually >40% voids, with pore size distribution that are uniformly below 0.5 μm). The ion-exchange microspheres are usually <100 μm in diameter and with a total thickness of ≈0.5 mm. As such, they are effective filters that remove suspended solids that are >0.5 μm from permeating fluids. Because of such sheetlike configuration, this material can be easily introduced into or withdrawn from a reactor with high concentration of suspended solids, with the target solutes being adsorbed onto or desorbed from the microadsorbents. Fig. 9 shows the electron microphotograph of the composite iminodiacetate (IDA) membrane (top) and a schematic depicting how the microbeads are trapped within the fibrous network of PTFE (bottom). Table 4 provides the salient properties of the composite IDA membrane used in the study. Note that the chelating microbeads constitute 90% of the composite membrane by mass. This feature allows the membrane to achieve the same level of performance as the parent chelating beads used in a fixed-bed operation. More details about characterization of the membrane are available in Ref.[24]. It may be noted that this material differs fundamentally from traditional ion-exchange membranes used in industrial process like Donnan dialysis (DD) and electrodialysis (ED) because of its high porosity. The membranes used in DD and ED have very low porosity and are strongly influenced by the "Donnan coion exclusion principle,"[27] which does not allow anions to pass through cation-exchange membrane and vice versa. However, in the case of a composite membrane, large gaps between ion exchangers allow anions to pass through freely even though it is a cation-exchange membrane. The suspended solids that

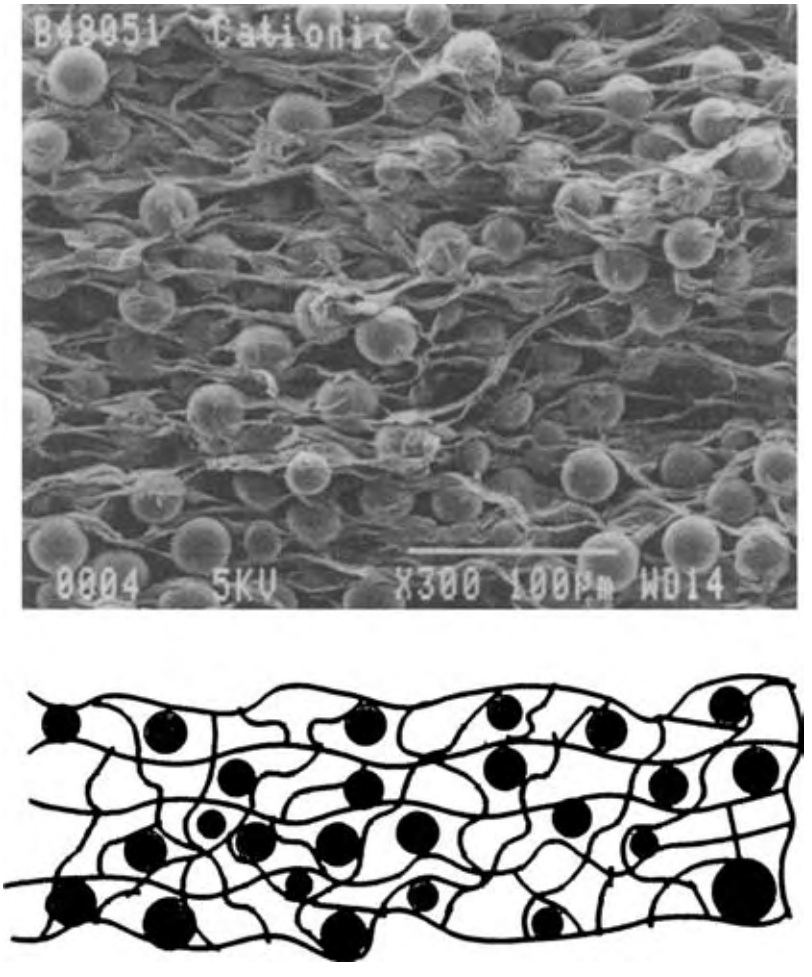


Fig. 9 Scanning electron microphotograph (300×) of the composite ion-exchange material (top) and schematic representation of the ion-exchanger beads present in a network of interlaced PTFE (bottom).

Table 4 Properties of the CIM

Composition	90% Chelex chelating resin, 10% Teflon
Pore size (nominal)	0.4 μm
Nominal capacity	3.2 mequiv./gm dry membrane
Membrane thickness	0.4–0.6 mm
Ionic form (as supplied)	Sodium
Resin matrix	Styrene-divinylbenzene
Functional group	Iminodiacetate
pH stability	1–14
Temperature operating range	0–75 °C
Chemical stability	Methanol; 1 N NaOH, 1 N H_2SO_4
Commercial availability	3M Corp., Minnesota, and Bio-Rad, California

are >0.5 μm are not able to penetrate across the skin of the membrane because of the pore size of the material, as explained above. However, water molecules and ions can easily move in and out of the thickness of the sheet, thus allowing for unimpeded ion-exchange reactions between target ions in solutions (heavy metals in this case) and the counterions of the membrane, as shown schematically in Fig. 10 and explained in detail by Sengupta.[19–26] After a design time interval, the membrane can be taken out and chemically regenerated with strong (3–5%) mineral acid solution.

PROCESS CONFIGURATION

Fig. 11 shows a conceptualized process schematic where a composite membrane strip is continuously run through the contaminated sludge (sorption step) and an acid bath (desorption step). Such a cyclic process configuration is relatively simple and can be implemented by using the composite membrane as a slow-moving belt. For a sludge containing heavy-metal hydroxide, say $Me(OH)_2$, the process works in two steps:

1. *Sorption:* The CIM, when in contact with the sludge, selectively removes dissolved heavy metals from the aqueous phase in preference to other nontoxic alkali and alkaline earth metal cations. Consequently, fresh heavy-metal hydroxide dissolves to maintain equilibrium, and the following reactions occur in series:

$$\text{Dissolution:}\quad Me(OH)_2\,(s) \Leftrightarrow Me^{2+} + 2\,OH^- \tag{28}$$

$$\text{CIM uptake:}\quad \overline{RN(CH_2\,COOH)_2} + Me^{2+} \Leftrightarrow \overline{RN(CH_2\,COO^-)_2\,Me^{2+}} + 2\,H^+ \tag{29}$$

3. *Desorption:* When the CIM is immersed in the acid chamber, the exchanger microbeads are efficiently regenerated according to the following reaction:

$$\overline{RN(CH_2COO^-)_2Me^{2+}} + 2H^+ \Leftrightarrow \overline{RN(CH_2COOH)_2} + Me^{2+} \tag{30}$$

The regenerated CIM is then ready for sorption again and the cycle is repeated. For such an arrangement to be practically feasible, the CIM sheet needs to be physically tough enough to withstand conveyor belt tension and resilient to the chemical forces that are created during sorption and regeneration. Previous studies conducted[21] in this area have confirmed that the CIM sheet is durable enough to withstand cyclical forces (physical and chemical) for more than 200 cycles.

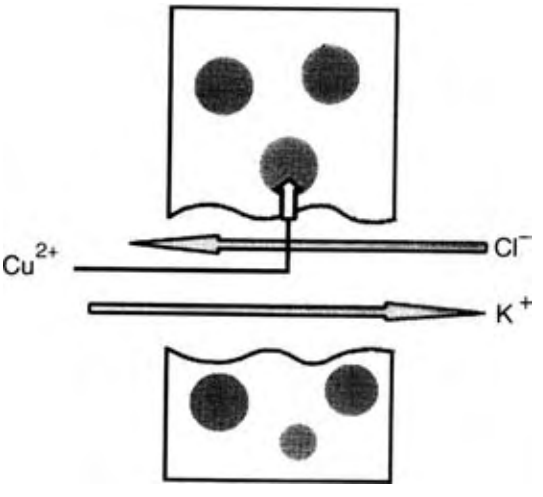


Fig. 10 Schematic diagram showing the porosity of the CIM to cations and anions and selective entrapment of heavy-metal cations.

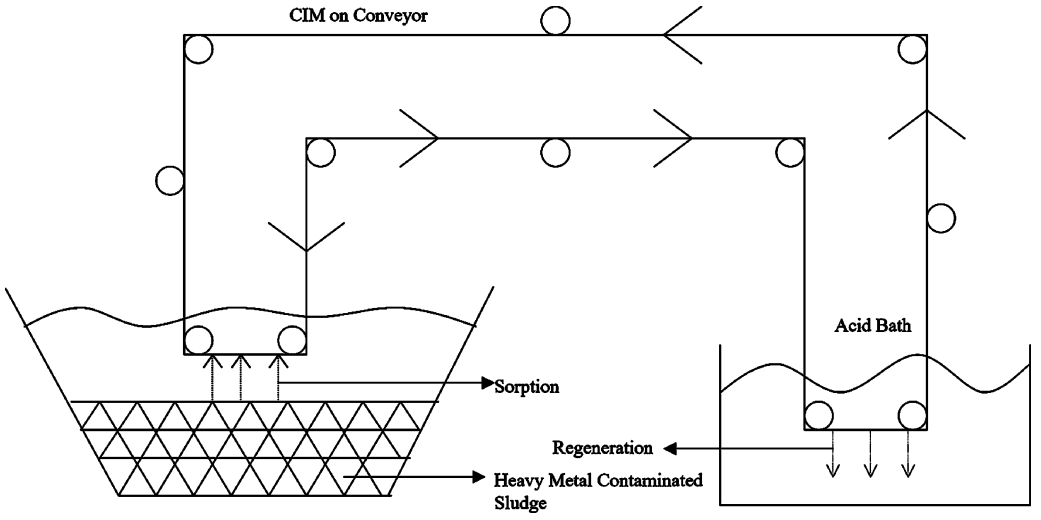


Fig. 11 A conceptualized continuous decontamination process for heavy-metal removal from a sludge reactor using CIM.

SORPTION KINETICS OF THE CIM

With chelating ion-exchanger beads that are used for heavy-metal removal, intraparticle diffusion is the most probable rate-limiting step. But for CIM composed of chelating resin, a significantly different physical configuration of the CIM may introduce additional diffusional resistance within it. As may be observed from the schematic diagram in Fig. 12, fairly stagnant pore liquid is present in the channels of the CIM between individual microbeads, and the solutes need to be transported through this pore liquid for sorption or desorption. This additional resistance to sorption/desorption is likely to retard kinetic rate.

Fig. 13 shows the results of a batch kinetic study (fractional metal uptake vs. time) comparing the parent chelating exchanger with the CIM under otherwise identical conditions. Fractional metal uptake, $F(t)$, is dimensionless and is defined as the ratio of the metal uptake $q(t)$ after time t and the metal uptake at equilibrium, q^0, i.e., $F(t) = q(t)/q^0$. Although the parent chelating microbeads were bigger in average diameter (50–100 mesh) than the microbeads within the CIM, the copper uptake rate, as speculated, was

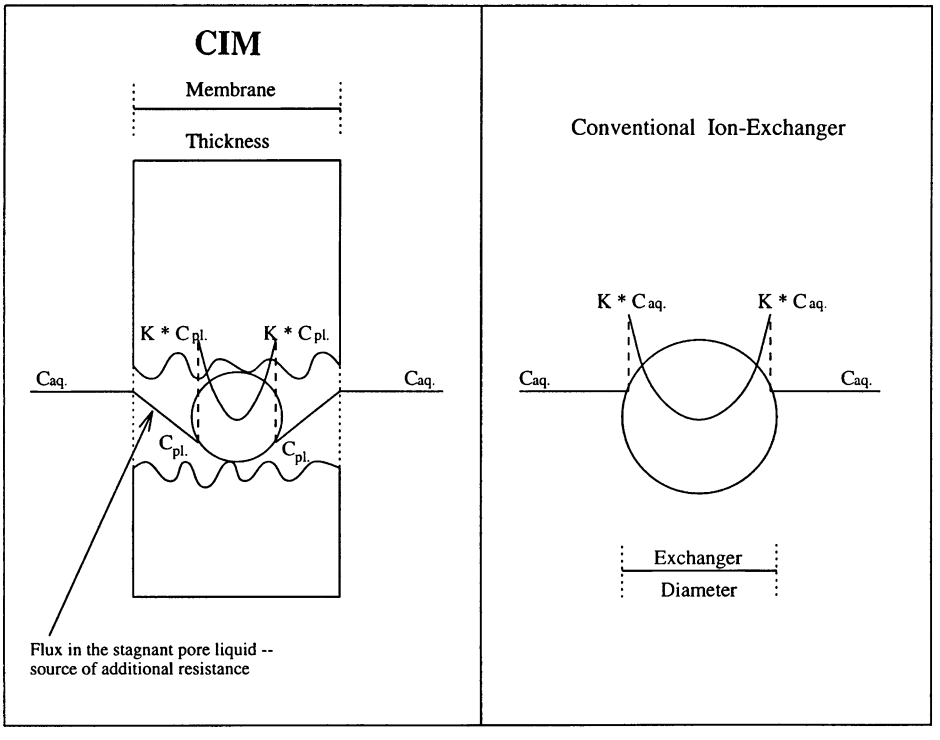


Fig. 12 Schematic showing the difference in flux curves between a conventional (spherical bead) ion exchanger and the CIM.

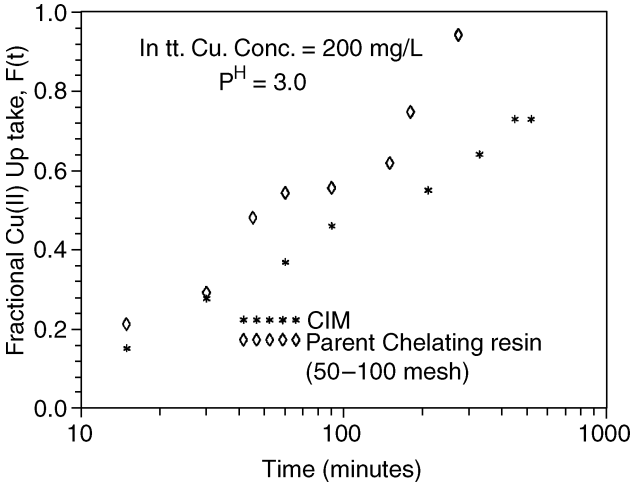


Fig. 13 Cooper uptake rates for the CIM vs. the parent chelating exchanger under identical solution condition.

slower with the CIM. To overcome the complexity arising because of the heterogeneity of the CIM (chelating microbeads randomly distributed in non-adsorbing Teflon fibers), a model was proposed,[20,21] in which the thin-sheetlike CIM may be viewed as a flat plate containing a pseudohomogeneous sorbent phase as shown in Fig. 14. Under the experimental conditions, it may be assumed that

1. Surface of the CIM is in equilibrium with the bulk of the liquid phase.
2. Total amount of solute in the solution and in the CIM sheet remains constant as the sorption process is carried out.
3. Solute (heavy metal) has high affinity toward the thin-sheet sorbent material.

This is a case of diffusion from a stirred solution with limited volume. The CIM is considered as a sheet of uniform material of thickness $2w$ placed in the solution containing the solute, which is allowed to diffuse into the sheet. The sheet occupies the space $-w \leq x \leq +w$, while the solution is of limited extent and occupies the space $-w - a \leq x \leq -w$, $w \leq x \leq +w + a$. The concentration of the solute in the solution is always uniform and is initially C_0 while initially the sheet is free from solute. Considering an apparent metal-ion diffusivity within the CIM phase, the PDE that must be solved in this case of metal uptake through a plane sheet (the thickness of the CIM) from a solution of limited volume is given as:

$$\frac{\partial C}{\partial t} = \overline{D}\frac{\partial^2 C}{\partial x^2} \tag{31}$$

where x is the axial space coordinate in the direction of CIM thickness and C the concentration of solute in the solution.

The IC of the above PDE is

$$C = 0, \quad -w < x < +w, \; t = 0 \tag{32}$$

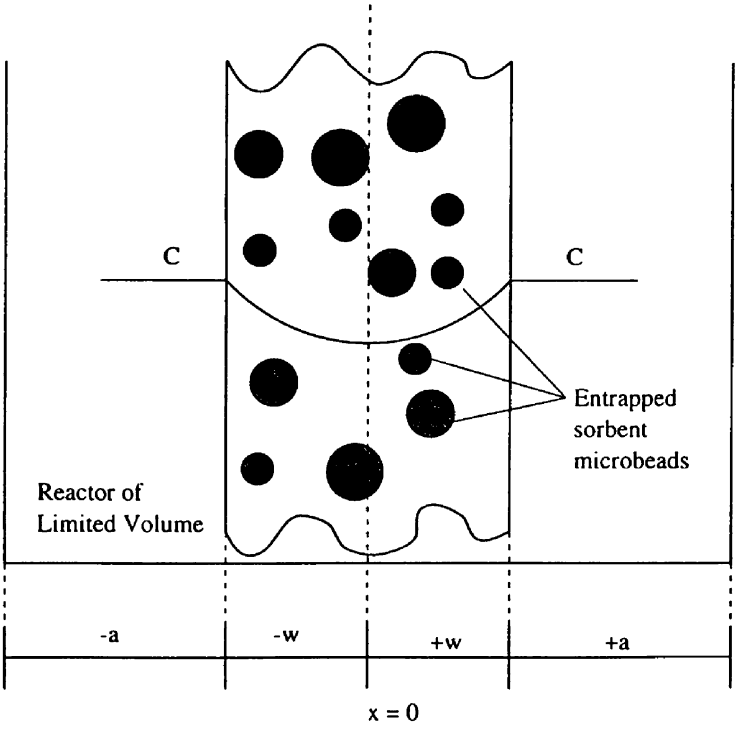


Fig. 14 Schematic showing sorption through an assumed flat flate with constant diffusivity from a reactor of limited volume.

and the BC expresses the fact that the rate at which the solute leaves the solution is always equal to that at which it enters the sheet over the surfaces, $x = \pm w$. This condition is mathematically expressed as

$$\frac{a}{K}\frac{\partial C}{\partial t} = \pm\overline{D}\frac{\partial C}{\partial x}, \quad x = \pm w,\ t > 0 \tag{33}$$

where K is the partition factor between the CIM and the solution, i.e., the concentration just within the sheet is K times that in the solution.

An analytical solution of this problem is given as,[28]

$$F(t) = \frac{q_t}{q^0} = 1 - \sum_{n=1}^{\infty}\frac{2\alpha(1+\alpha)}{1+\alpha+\alpha^2 q_n^2}\exp\left(\frac{-\overline{D}q_n^2 t}{w^2}\right) \tag{34}$$

where the q_n values are the nonzero positive roots of

$$\tan q_n = -\alpha q_n \tag{35}$$

and $\alpha = a/(Kw)$.

CONCLUSIONS

Ion-exchange technology has now matured enough to be the primary choice for a variety of applications spanning many industrial and municipal applications. Its primary advantage is that it can remove ions to almost nondetect levels from the matrix. Also, ion-exchange technology has the ability to selectively remove some ions from the background of other, nontarget ions. The primary area of research in this field is the synthesis of new ion exchangers (primarily organic) that can carry out selective removal of target ions that are determined to be toxic at very trace concentrations (at the level of μg/L or less), such as perchlorate, radium, uranium, cesium, heavy metals, oxyanions, etc. Other important areas of development in this field are the morphology of the ion exchanger and the nature of contact between the exchanger and the contaminated matrix, be it solid, liquid, or gas. Ion-exchanger membranes, composite ion-exchange material, nanofibers, microspheres, etc., have shown potential in many scenarios where the traditional ion-exchanger beads cannot be used. Lastly, the efficient disposal of wastestreams from regeneration of ion exchangers and the ability to recover/recycle valuable ions from this wastestream have received widespread attention of many researchers. It is expected that future developments in this area will result in ion-exchange technology making inroads into new application scenarios, either as a stand-alone operation, or in a treatment scheme in combination with other physical/chemical, biological, or thermal processes.

I

REFERENCES

1. Way, J.T. On the power of soils to absorb manure. J. Roy. Agric. Soc. Engl. **1850**, *11*, 313.
2. Thompson, H.S. On the absorbent power of soils. J. Roy. Soc. Engl. **1850**, *11*, 68.
3. Adams, B.A.; Holmes, E.L. Absorptive properties of synthetic resins—I. J. Soc. Chem. Ind. (Lond.) **1935**, *54*, 2T.
4. D'Alelio, G.F U.S. Patent 2,340,110, 1944.
5. Boyd, G.E.; Schubert, J.; Adamson, A.W. The exchange adsorption of ions from aqueous solutions by organic zeolites. I. Ion-exchange equilibria. J. Am. Chem. Soc. **1947**, *69*, 2818–2829.
6. Boyd, G.E.; Adamson, A.W.; Myers, L.S. The exchange adsorption of ions from solution from organic zeolites. II. Kinetics. J. Am. Chem. Soc. **1947**, *69*, 2836–2848.
7. Helfferich, F.G. *Ion Exchange*; McGraw-Hill: New York, 1962.
8. Sengupta, A.K. Introduction. In *Ion exchange Technolgy: Advances in Pollution Control*; Sengupta, A.K., Ed.; Technomic Publishing Company, Inc.: Lancaster, PA, 1995.
9. King, C.J. Chapter 15. In *Handbook of Separation Process Technology*; Rousseau, R.W., Ed.; John Wiley & Sons: New York, 1987.
10. Clifford, D.A. Ion exchange and inorganic adsorption. In *Water Quality and Treatment: A Handbook of Community Water Supplies*; Pontius, F.W., Ed.; McGraw-Hill Inc.: New York, 1990.
11. Hale, D.K.; McCuley, D.J. Structure and properties of heterogeneous cation-exchange membranes. Trans. Faraday Soc. **1961**, *57*, 135.
12. J.M. Montgomery, Consulting Engineers, Inc. Ion exchange and demineralization. In *Water Treatment Principles and Design*; John Wiley & Sons: New York, 1985.
13. Wallace, R.M. Concentration and separation of ions by Donnan membrane equilibrium. Ind. Eng. Chem. Prod. Res. Dev. **1967**, *6* (4), 423–431.
14. Cox, J.A.; Twardowski, Z. Tubular flow Donnan dialysis. Anal. Chem. **1980**, *52* (9), 1503–1505.
15. Kim, B.M. Donnan dialysis for the removal of chromates and cyanides. AIChE Symp. Ser. **1980**, *76* (197), 184–192.

16. Hichour, M.; Persin, F.; Molenat, J; Sandeaux, J.; Gavach, C. Defloururation des Eaux par Dialyse de Donnan et Electrodialyse. Desalination **1999**, *122*, 53–62.
17. Miyoshi, H.; Yamagami, M.; Katoka, T. Characteristic coefficients of nafion membranes. Chem. Exp. **1990**, *5* (10), 717–720.
18. Prakash, P.; Sengupta, A.K. Selective coagulant recovery from water treatment plant residuals using Donnan membrane process. Environ. Sci. Technol. **2003**, *37* (19), 4468–4474.
19. Errede, L.A.; Stoesz, J.D.; Sirvio, L.M. Swelling of particulate polymers enmeshed in poly(tetrafluoroethylene). J. Appl. Polym. Sci. **1986**, *31* (8), 2721–2737.
20. Sengupta, S.; Sengupta, A.K. Characterizing a new class of sorptive/desorptive ion exchange membrances for decontamination of heavy-metal-laten slugdes. Environ. Sci. Technol. **1993**, *27* (10), 2133–2140.
21. Sengupta, S. A New Separation and Decontamination Technique for Heavy-Metal-Laden-Sludges Using Sorptive/Desorptive Ion exchange Membrances. Ph.D. dissertation, Lehigh University, Bethlehem, PA, 1993.
22. Sengupta, A.K.; Shi, B. Selective alum recovery from clarifier sludge. J. Am. Water Works Assoc. **1992**, *84* (1), 96–103.
23. Sengupta, S.; Sengupta, A.K. Solid phase heavy metal separation using composite ion exchange membranes. Hazard. Waste Hazard. Mater. **1996**, *13* (2), 245.
24. Sengupta, S.; Sengupta, A.K. Heavy-metal separation from sludge using chelating ion exchangers with nontraditional morphology. React. Funct. Polym. **1997**, *35* (1–2), 111–134.
25. Sengupta, S. Electro-partitioning with composite ion exchange material: an innovative in-situ heavy-metal decontamination process. React. Func. Polym. **1999**, *40* (3), 263–273.
26. Sengupta, S.; Sengupta, A.K. Trace heavy metal separation by chelating ion exchangers. In *Environmental Separation of Heavy Metals: Engineering Processes*; Sengupta, A.K., Ed.; Lewis Publishers, CRC Press, 2002.
27. Burgess, J. *Ions in Solution: Basic Principles and Interactions*; Ellis Harwood Ltd., John Wiley & Sons: New York, NY, 1989.
28. Crank, J. *The Mathematics of Diffusion*; Oxford University Press: London, 1975.

Ion Exchange Resin

I

Sukalyan Sengupta
Civil and Environmental Engineering Department, University of Massachusetts, Dartmouth, Massachusetts, U.S.A.

Arup K. Sengupta
Department of Civil and Environmental Engineering, Lehigh University, Bethlehem, Pennsylvania, U.S.A.

INTRODUCTION

An ion exchange resin consists of a cross-linked framework—termed matrix—that consists of an irregular, macromolecular, three-dimensional network of hydrocarbon chains. Charged functional groups are attached to the matrix by covalent bonding. The predominant form of ion exchange resins is as spherical beads in the size range of 75–1500 μm. If the resin is capable of exchanging cations, it is termed as a cation exchange resin and the functional groups are negatively charged. Similarly, if the resin exchanges anions, it is termed as an anion exchange resin and the functional groups are positively charged. The fixed functional groups are balanced by counterions of opposite charge to maintain electroneutrality. The counterions are exchanged with other ions of the same charge in solution.

The matrix of a majority of ion exchange resins is made by the coplymerization of styrene and divinylbenzene (DVB). Styrene provides the general backbone, while DVB acts as the cross-linker, i.e., it interconnects the various styrene chains. This cross-linking agent makes the ion exchange resin insoluble, an extremely important characteristic of the material. However, the resin is elastic and can swell by taking up solvent. The chemical, thermal, and mechanical stability and the ion exchange behavior of the resins depend primarily on the structure and the degree of cross-linking and on the nature and number of the fixed ionic groups.[1] In general, the ability of a resin to absorb moisture and swell depends mainly on two factors: the functional group substitution and the degree of cross-linking. The influence of the functional group on swollen volume is described in the next section. The effect of cross-linking is inversely proportional to swollen volume. This can be easily understood by assuming the cross-linker to provide a higher spring constant. Because the moisture content of a resin varies widely, its density also varies; typically, the specific gravity of cationic resins is in the range of 1.10–1.35, while that of anionic resins varies from 1.05 to 1.15.[2] The bulk or apparent density of a resin will vary from 600 to 800 mg/L when wet.

The most important property of an ion exchange resin is its exchange capacity, i.e., the number of equivalents of counterion that can be exchanged. Please note that the unit of ion exchange is always equivalents (rather than moles) because ion exchange is an equivalent exchange process. As the volume of an ion exchange resin depends on the amount of moisture absorbed by the resin, the best way to report capacity is dry weight and it is usually reported in milliequivalents of counterion per dry gram of the resin. It is simply a measure of the extent of functional group substitution in the exchanger and is constant for a specific resin. It is generally in the 2–5 meq/g range. Please note that this number is the total or the theoretical exchange capacity: actual or operating capacity is always less than total and is dependent on operating conditions, such as superficial liquid velocity, regeneration scheme, column bed porosity, etc. Ion exchange resins are typically synthesized as gel type, with a pore size of 20–50 Å. This results in increased chemical and mechanical stability and lower water content. Therefore, more ionized groups are available per unit volume, resulting in higher capacity. Macroporous resins, on the other hand, have much wider pores (50–1,000,000 Å), with internal surface area ranging from 7 to 1500 m^2/g, much greater than the standard gel-type resins. The transport of reactants inside these resins is through internal pores or channels: this guarantees access to the interior of the beads even when nonpolar solvents are used.

GENERAL TYPES OF ION EXCHANGE RESIN

The ionizable group attached to the resin structure determines the functional capability of the ion exchange resin.[2] The four general types of ion exchange resin are as follows:

Strong-Acid Cation Exchange Resin: They contain the strongly acidic sulfonate (SO_3^-) functional group

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008053

and thus can operate over a very wide pH range. Two typical reactions of this type of resin are:

$$2\overline{RSO_3^-Na^+} + CaCl_2 \Leftrightarrow \overline{(RSO_3^-)_2Ca^{2+}} + 2NaCl \quad (1)$$

$$2\overline{RSO_3^-H^+} + CaCl_2 \Leftrightarrow \overline{(RSO_3^-)_2Ca^{2+}} + 2HCl \quad (2)$$

As may be noted from Eq. (2), strong-acid resins can convert neutral salts into their corresponding acids if operated in the hydrogen cycle. This ability is known as salt splitting, and distinguishes them from weak-acid resins. Also, note that Eq. (1) is a water-softening reaction. Details of softening are available in the entry ''Ion Exchange.'' Strong-acid resins have their largest swollen volume in the hydrogen form.

Weak-Acid Cation Exchange Resin: Weak-acid resins differ from strong-acid resins in that weak-acid resins function only in the neutral to alkaline pH range because the functional groups have a high affinity for H^+ and thus are not ionized at low pH.[3] The most common functional group is carboxylate (COO^-). Thus, weak-acid cation exchange resins require the presence of some alkali species to react with the more tightly held H^+ of the resin, for example:

$$\overline{2R^-H^+} + Ca(HCO_3)_2 \Leftrightarrow \overline{(R^-)_2Ca^{2+}} + 2H_2CO_3 \quad (3)$$

The exchange is, in effect, a neutralization with the alkalinity (HCO_3^-) neutralizing the H^+ of the resin. Because of the high affinity of weak-acid cation exchange resins for H^+, they show excellent regeneration efficiency with mineral acid. Weak-acid resins do not require the same sort of concentration driving force required to convert strong-acid resins to the hydrogen form. Weak-acid resins have their smallest swollen volume in the hydrogen form and the volume increases as the resin is converted to the salt form.

Strong-Base Anion Exchange Resin: They operate well throughout the entire pH range and will split neutral salts into their corresponding bases if operated on the hydroxide cycle. For example:

$$\overline{R^+OH^-} + NaCl \Leftrightarrow \overline{R^+Cl^-} + Na^+ + OH^- \quad (4)$$

The functional sites are derived from quaternary ammonium groups. Two types of strong-base anion exchange resins are available: 1) Type I has three methyl groups on the nitrogen atom, as shown in Figs. 1 and 2) Type II resins have an ethanol group that replaces one of the methyl groups, as shown in Fig. 2. Type I resins are more strongly basic, and therefore

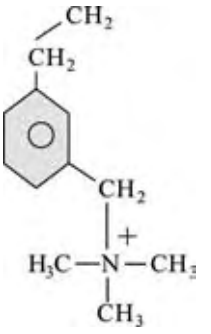


Fig. 1 Schematic presentation of a Type I strong-base anion exchange resin. *(View this art in color at www.dekker.com.)*

can provide better silica and carbonate removal vis-à-vis type II resins. Type II resins have greater regeneration efficiency and capacity. Strong-base resins have their largest swollen volume in the hydroxide form.

Weak-Base Anion Exchange Resin: They can be operated only in the acidic pH region where the primary, secondary, or tertiary amine functional groups are protonated, and thus can act as positively charged exchange sites for anions. The weak-base resins remove free mineral acidity, such as HCl or H_2SO_4, but cannot split salts. Consider the following reaction:

$$\overline{R_3N:} + HCl \Leftrightarrow \overline{R_3NH^+Cl^-} \quad (5)$$

$$\overline{R_3NH^+Cl^-} + HNO_3 \Leftrightarrow \overline{R_3NH^+NO_3^-} + HCl \quad (6)$$

In Eq. (5) the free form of the weak-base anion exchange resin becomes protonated on contact with a mineral acid (and low pH). To maintain electroneutrality in the resin, an equivalent of Cl^- enters the resin with every equivalent of H^+. Now this Cl^- loaded resin is ready for anion exchange as shown in Eq. (6), but this anion exchange will take place only at low pH. As with weak-acid resins, weak-base resins display excellent regenerability. They can be regenerated with

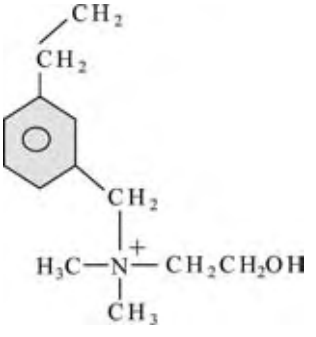


Fig. 2 Schematic presentation of a Type II strong-base anion exchange resin. *(View this art in color at www.dekker.com.)*

NaOH, NH_4OH, Na_2CO_3, or $Ca(OH)_2$. For complete regeneration, only slightly higher than the stoichiometric amount (<20%) is required. The swollen volume of a weak-base resin is the minimum when it is in the free form and increases as it is converted to salt forms.

ION EXCHANGE RESIN SELECTIVITY

Ion exchange resins display preference for certain ions over others and this characteristic is extremely critical during the choice of a resin for a particular application scenario. The fundamentals of ion exchange equilibria and selectivity are explained in the entry on ''Ion Exchange.'' This subject has been studied extensively by many researchers.[1,4] The following empirical rules can be used to approximate selectivities of resins.

The ion exchange resin tends to prefer:

1. The counterion of higher valence at low concentration and ordinary temperature. Thus,

   $$Th^{4+} > Al^{3+} > Ca^{2+} > Na^{+}$$

   and

   $$PO_4^{3-} > SO_4^{2-} > CI^{-}$$

2. For the case of constant valence, the counterion of higher atomic number (lower hydrated radius or smaller solvated volume) at low concentration and ordinary temperature. Thus,

   $$Cs^{+} > Rb^{+} > K^{+} > Na^{+} > Li^{+}$$

   $$Ba^{2+} > Sr^{2+} > Ca^{2+} > Mg^{2+} > Be^{2+}$$

3. For the case of constant valence, the counterion with greater polarizability at ordinary temperature and low concentration. A measure of polarizability is the Debye–Hückel parameter $\mathring{a}$, which gives the distance of closest approach between ions of opposite sign. Thus, highly polarized counterions, such as Ag^{+} and Tl^{+}, have abnormally low $\mathring{a}$ values and are preferred to alkali ions by the cation exchange resins with sulfonic acid groups.
4. The counterion that interacts more strongly with the fixed ionic groups or with the matrix. The former is the basis for chelating ion exchange resins and many other specialty resins and is explained in detail in the subsequent section. The interaction of the counterion with the matrix is the primary reason for a selectivity sequence for ClO_4^{-} shown below:

   Tributyl amine functional group
   > triethyl amine > trimethyl amine

5. The counterion that participates least in the complex formation with the coion. The explanation is straightforward: if the counterion forms a complex with the coion in solution, it will not be available for exchange with the loaded ions of the resin. Please see Refs.[5–9] for details on ion exchanger selectivity.

SPECIALTY ION EXCHANGE RESINS

In very few applications is ion exchange technology used to remove all cations/anions from a solvent. Perhaps, the only case of such an approach is in water demineralization, which is explained in detail in the entry ''Ion Exchange.'' The more common application scenario is that the solvent contains many counterions, most of which are benign but some counterions exist that are present in much lower concentration (than the benign counterions) and need to be removed from the solvent. Specialty ion exchange resins play an important role in such situations and offer a distinct advantage over other technologies. For example, in many waste streams, which are contaminated with heavy metals, the concentration of heavy metals is much smaller than benign cations, such as Ca^{2+}, Mg^{2+}, Na^{+}, etc. For example, let us consider a typical cation exchange reaction between Ni^{2+} and Ca^{2+}:

$$\overline{R_2Ca} + Ni^{2+} \leftrightarrow \overline{R_2Ni} + Ca^{2+} \tag{7}$$

Assuming ideality, the equilibrium constant or selectivity coefficient of Eq. (7) is:

$$K_1 = \frac{[\overline{R_2Ni}][Ca^{2+}]}{[Ni^{2+}][\overline{R_2Ca}]} \tag{8}$$

The ratio of Ni^{2+} to Ca^{2+} in the aqueous phase is, in general, much less than unity. Thus, for the Ni^{2+} removal process to be selective (i.e., high R_2Ni), the equilibrium constant, K_1, needs to be extremely high. Conventional cation exchangers with coulombic (electrostatic) type interaction are unable to attain such high selectivity (charge of calcium and nickel ion is the same).

Metal-Selective Chelating Resins

Most of the heavy-metal cations of interest, such as Cu^{2+}, Hg^{2+}, Pb^{2+}, Ni^{2+}, Cd^{2+}, Zn^{2+}, etc., are transition-metal cations and exhibit Lewis acid characteristics (electron acceptors). With organic and inorganic ligands (Lewis bases), all these heavy-metal cations form fairly strong complexes. Most of the complexes of these metal cations, depending on their coordination number, have regular or slightly distorted tetrahedral, octahedral, or square pyramid structures.[10] As Ca^{2+}, Mg^{2+}, and Na^{+}—the most commonly encountered competing nontoxic cations in water and wastewater—do not undergo such strong complexation, incorporating organic ligands as functional groups into the polymer matrix of the ion exchanger through covalent bonding was a natural progression of ideas to improve the exchanger's selectivity toward the toxic metal ions. These functionalized polymers are often referred to as chelating polymers, coordinating polymers, or metal selective ion exchange resins. Fig. 3 shows the chelating exchanger with imino-diacetate functionality, the workhorse of metal-selective exchangers and available from every major resin manufacturer around the world. Fig. 4 shows two commonly used routes for synthesizing this chelating polymer from styrene monomers (cross-linking not shown).[11] A typical ion exchange reaction between a metal ion, Me^{2+}, and Na^{+} for this resin may be presented in the following way:

$$\overline{RN(CH_2COONa)_2} + Me^{2+} \Leftrightarrow \overline{RN(CH_2COO)_2Me} + 2Na^{+} \quad (9)$$

Eq. (9), however, fails to reveal the Lewis acid–base (LAB) type interaction between the metal ion and the imino-diacetate functionality. Assuming that the water molecule has four coordinated water molecules in the aqueous phase, $[Me(H_2O)_4]^{2+}$, the overall exchange is shown in the below equation.

Note that three water molecules (ligands) from the coordination sphere of the metal ion are replaced by one nitrogen and two oxygen donor atoms in the imino-diacetate functionality. The arrows indicate the metal–ligand or LAB interaction and the high metal ion selectivity for this type of functional group is often attributed to the accompanying coordination

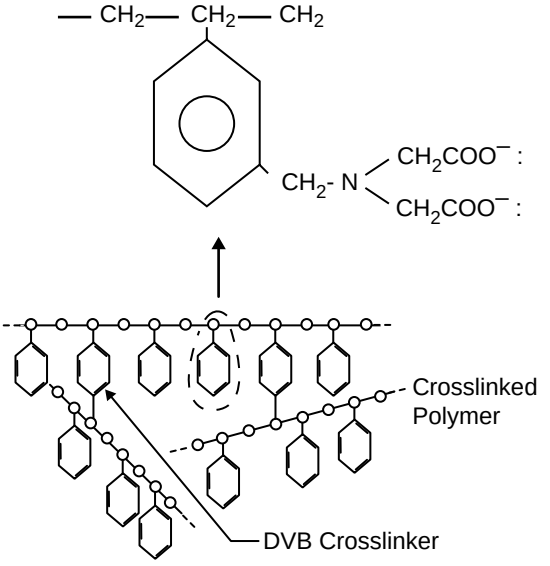


Fig. 3 Schematic presentation of three-dimensional cross-linked polystyrene with imino-diacetate functional group.

reaction in conjunction with exchange of ions. Fig. 5 provides experimentally determined Me^{2+}/Ca^{2+} separation factors for three commercial ion exchange resins (imino-diacetate functionality—IRC 718, thiol functionality—GT73, and picolylamine functionality—XFS4195) for various heavy metals at varying pH values, and the high selectivity of the metal ions can be readily noted.[12] Please see the entry "Ion Exchange" for details about separation factor. Like Ca^{2+}, these metal ions have a charge 2+ and, therefore, only coulombic/electrostatic interaction cannot be the reason for such a high Me^{2+}/Ca^{2+} separation factor. On the contrary, such a high selectivity is always attributed to a relatively strong Lewis acid characteristic of the toxic-metal cations favoring their selective uptake through coordination reactions.

To characterize high metal ion selectivity for chelating ion exchangers from a thermodynamic perspective, the metal ion uptake can be divided into two consecutive steps—ion exchange (IX) followed by LAB interaction, i.e.,

$$Me^{2+}(aq) \xrightarrow{IX(Step\ I)} \overline{RMe} \xrightarrow{LAB(Step\ II)} \overline{RMe}$$

$$\text{R}-\text{CH}_2\text{N}(\text{CH}_2\text{COO}^-\text{Na}^+)_2 + \text{Me}(\text{H}_2\text{O})_4^{2+} \longleftrightarrow \text{R}-\text{CH}_2\text{N}(\text{CH}_2\text{COO}^-)_2 \rightarrow \text{Me}(\text{H}_2\text{O})^2 + 2\text{Na}^+ + 3\text{H}_2\text{O} \quad (10)$$

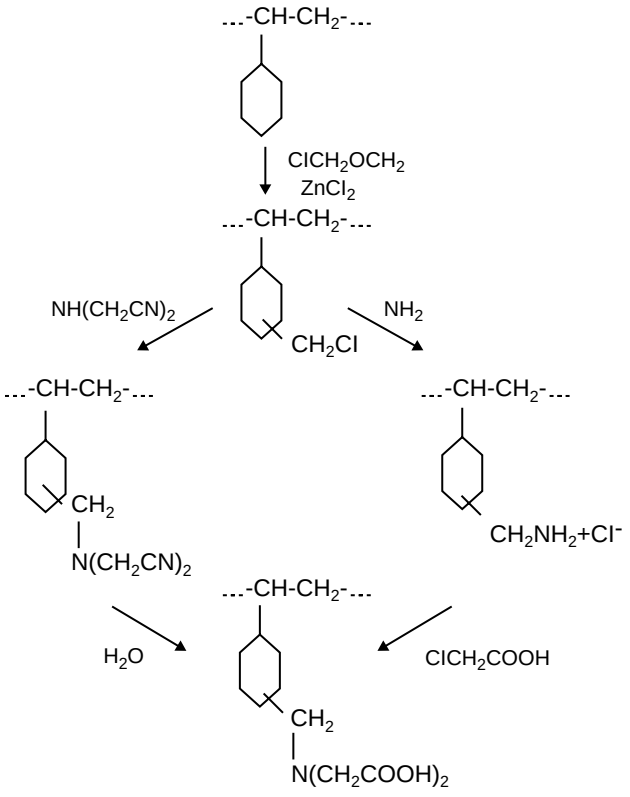


Fig. 4 Two parallel routes for synthesis of imino-diacetate resins from polystyrene (cross-linking not shown).

At the standard state, the overall free energy at equilibrium between Me^{2+}(aq) and RMe is given by:

$$\Delta G^0_{\text{overall}} = \Delta G^0_{\text{IX}} + \Delta G^0_{\text{LAB}} \tag{11}$$

or

$$-\text{RT ln } K_{\text{overall}} = -\text{RT ln } K_{\text{IX}} - \text{RT ln } K_{\text{LAB}}$$

or

$$K_{\text{overall}} = K_{\text{IX}}\, K_{\text{LAB}} \tag{12}$$

In general, for ion exchangers with chelating functionality, K_{LAB} is very high for most of the heavy-metal ions of interest because of their Lewis acid characteristics. Therefore, the overall equilibrium constants, according to Eq. (12), are also very high. For sodium, Na^+, LAB interaction (step II) is practically absent and hence,

$$K_{\text{overall}} = K_{\text{IX}} \tag{13}$$

For Ca^{2+}, however, LAB interaction is present but much weaker compared to most of the heavy-metal cations, and thus as a general rule, selectivity sequence for chelating exchangers may be written as follows:

$$K_{\text{overall}}\ (\text{heavy metal}) \gg K_{\text{overall}}\ (\text{calcium}) \gg K_{\text{overall}}\ (\text{sodium})$$

Such a high metal ion selectivity and more stringent environmental regulations have aroused high interest in the application of these chelating polymers for removal, separation, and purification of metal ions from heavy-metal contaminated water and wastewater streams.[13–21] Many chelating exchangers have been synthesized during the last 30 yr.[22–24] Table 1 provides composition of several commercially available chelating exchangers and other relevant information. These exchangers contain chelating functionalities with one or more donor atoms that can form coordinate bonds with metal ions. Depending on the number of donor atoms present in a repeating functionality, these exchangers are often referred to as mono-, bi-, or polydentate. Nitrogen, oxygen, and sulfur are the donor atoms in almost every chelating exchanger synthesized to date. Identifying the active donor atoms for a given application may provide useful clues to assess metal ion selectivity and other related properties for a given chelating polymer. These donor atoms form only a part of the complete chelating functionality, which is essentially either weak acid (say carboxylate, diacetate, thiol, etc.) or weak base (tertiary amine, pyridine, etc.). Because of the weak-acid or weak-base characteristic, these chelating functionalities exhibit high affinity toward the hydrogen ion. As a result, selective uptake of heavy-metal cations by chelating exchangers under highly acidic conditions (pH $<$ 2.0) is adversely affected because of strong competition from H^+. On the other hand, at neutral to alkaline pH, heavy-metal cations are quite insoluble because of low solubility product values for their hydroxides, carbonates, sulfides, etc. For effective heavy-metals removal, the optimum pH range for most of the chelating polymers is, thus, often limited to 2.0–7.0. Fig. 6 shows total copper uptake for different chelating exchangers as a function of pH.[25] Notice how metal removal capacity is essentially lost for IRC-718 (imino-diacetate functionality) and ES-467 (amino-phosphonate functionality) at pH $<$ 1.5. Note also the unusual behavior of XFS 4195; there is practically no reduction in copper uptake capacity even at a pH as low as 1.5. This is due to its unique metal ion binding mechanism and will be discussed later.

Chelating ion exchange resins can be titrated in a manner very similar to other aqueous-phase weak acids and weak bases. For example, strong-acid cation exchangers show an abrupt change in pH whereas weak-acid cation exchangers show gradual increase in pH—indicative of the latter's high affinity for H^+—as may be seen from Fig. 7. Titration curves are,

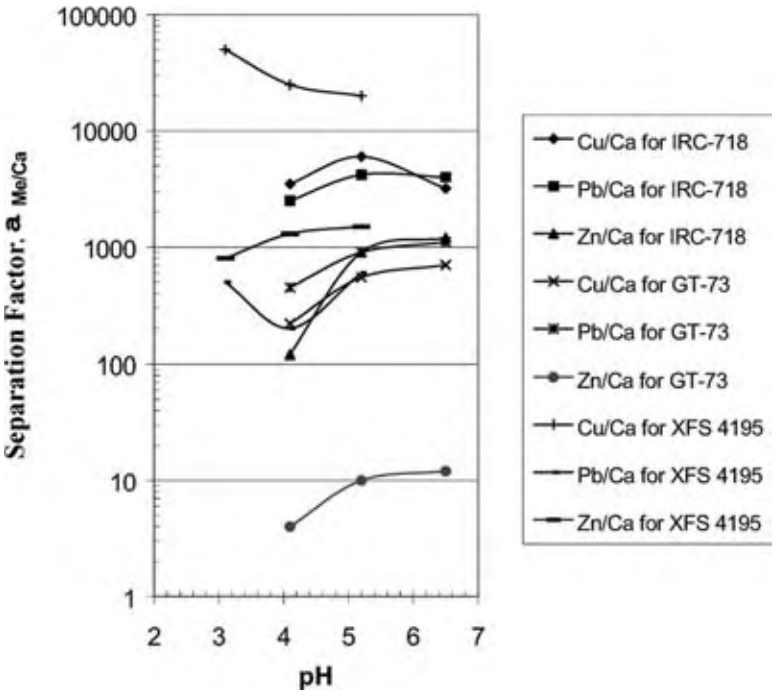


Fig. 5 Metal(II)/Ca separation factor for different commercial resins. *(View this art in color at www.dekker.com.)*

however, dependent on the aqueous-phase electrolyte concentration also; see excellent discussions of this subject by Marinsky and Helfferich.[1,26] A relationship to compute the apparent dissociation constant, K_a, of a weak-acid exchanger as a function of pH and aqueous-phase electrolyte concentration has been proposed by Helfferich and may be stated as follows:[1]

$$pK_a = pH - \log[A^-] + \log\left[\frac{\overline{X}}{2}\right] \quad (14)$$

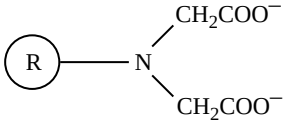
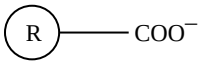
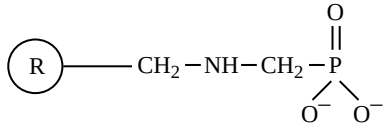
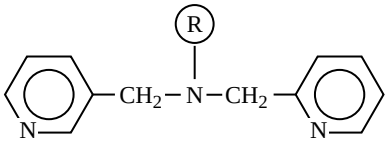
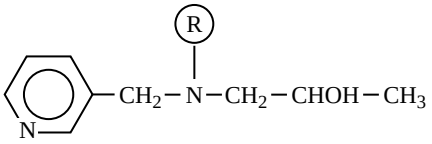
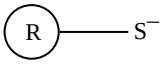
where $[A^-]$ is the aqueous-phase anion concentration and $[\overline{X}] = [RH] + [R]$ is the total concentration of (dissociated and undissociated) ionogenic groups in the exchanger. Note that the apparent pK_a decreases at a constant pH with an increase in electrolyte concentration (i.e., A^-), implying enhanced ionization of the exchanger.

Chelating exchangers' high preference for H^+ is often viewed as a shortcoming for heavy-metal removal under highly acidic conditions, but it offers an excellent regeneration of metal-loaded chelating polymers with moderately concentrated (2–10%) mineral acid. From a practical viewpoint, high regeneration efficiency of a chelating exchanger is just as desirable as the high metal ion affinity. Fig. 8 shows the regeneration of copper-loaded IRC-718 with 2% HCl.[27] As expected, copper desorption/elution was very sharp with copper concentration in the spent regenerant as high as 15,000 mg/L, indicating efficient regeneration in accordance with the principles of displacement chromatography.[1] Incidentally, several toxic metals, viz., Cu, Pb, Hg, Cd, Zn, and Ni, are included in EPA's list of priority pollutants and one of the major challenges is to minimize the volume of these metal-contaminated wastes.[27] High metal ion selectivity of the chelating exchangers accompanied by excellent acid-regeneration efficiency offers opportunities to concentrate and reduce the volume of metal-laden dilute wastewater streams, often by over 1000 times.[28]

KINETICS AND RATE PROCESSES

In spite of their high affinity toward most of the heavy-metal cations and favorable thermodynamics over a wide pH range, the metal-selective chelating cation exchange resins are often criticized because of their slow kinetics, which are primarily diffusion-controlled rate-limiting processes.[29–31] Several factors influence kinetics; they include degree of cross-linking and accompanying rigidity, steric property of the functional group, charge density, nature of polymer matrix, e.g., gel, macroporous, or pellicular, mechanism of metal ion binding, particle size, and other hydrodynamic conditions. Although no two cases are likely to be identical, ion exchange kinetics for chelating ion exchange resins is improved by decreasing the amount of cross-linking, by decreasing the hydrophobic nature of the polymer matrix, and, above all, by reducing the size of the polymer beads. Metal ion diffusivity in the exchanger is strongly dependent on the cross-linking because it influences the

Table 1 Salient information on some chelating ion exchange resins

Functionality	Electron donor atoms	Matrix, cross-linker	Nature of chelating group
	1 Nitrogen, 2 Oxygen	Polystyrene, divinylbenzene	Imino-diacetate
	1 Oxygen	Polymethacrylate	Carboxylate
	1 Nitrogen,2 Oxygen	Polystyrene, divinylbenzene	Aminophosphonate
	3 Nitrogen	Polystyrene, divinylbenzene	Picolyl-amine based
	2 Nitrogen	Polystyrene, divinylbenzene	Pyridine based
	1 Sulfur	Polystyrene, divinylbenzene	Thiol

swelling characteristic and water content of the polymer phase. Boyd and Soldanho determined the diffusivity of zinc ion for strong-acid cation exchange resins with varying degrees of DVB cross-linking, as reproduced in Table 2.[32] Note that effective zinc ion diffusivity drops by almost two orders of magnitude as the DVB increases from 5% to 23.8%.

In general, metal ion exchange by a chelating resin is intraparticle diffusion controlled (IPDC) accompanied by chemical reaction (chelation).[29–36] Both Fick's law and Nernst–Planck Eq.[30] have been used to describe ion exchange kinetics by intraparticle diffusion. For ion exchange resin beads with spherical geometry and constant exchanger-phase diffusion coefficient, the mathematical representation may be given as follows:

$$\frac{\partial \overline{C_i}}{\partial t} = \overline{D}\left(\frac{\partial^2 \overline{C_i}}{\partial r^2} + \frac{2}{r}\frac{\partial \overline{C_i}}{\partial r}\right) \tag{15}$$

where r = radial space coordinate, i.e., distance from bead center.

This equation is to be solved under the appropriate initial and boundary conditions. For the simple case of infinite solution volume, where the aqueous-phase concentration does not change because of sorption or desorption, fractional attainment of equilibrium, F, is given by:

$$F \approx \left[1 - \exp\left(-\frac{\overline{D}\, t\, \pi^2}{r_0^2}\right)\right]^{1/2} \tag{16}$$

where D is intraparticle self-diffusion coefficient of the exchanging ion, t is the time since the beginning of the run, and r_0 is the radius of the spherical ion exchange bead. The half-time, $t_{1/2}$, for attainment of 50% equilibrium capacity (i.e., $F = 0.5$) can be computed from Eq. (16) to give the following relationship:

$$t_{1/2} = \text{constant}\frac{r_0^2}{D} \tag{17}$$

The relative rate (inverse of $t_{1/2}$) is thus proportional to the diffusion coefficient in the exchanger

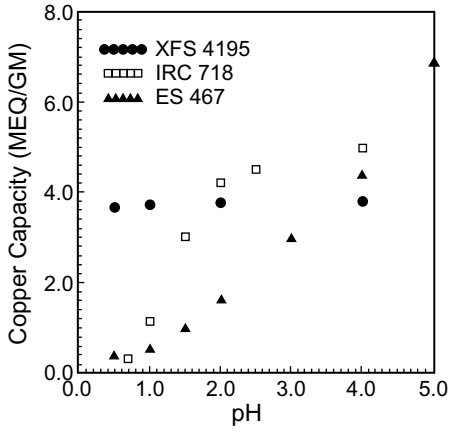


Fig. 6 Comparison of copper(II) uptake as a function of pH for three chelating exchangers in equilibrium with 300 mg/L Cu^{2+}. (From Ref.[25].)

and inversely proportional to the square of the bead radius. Note that $t_{1/2}$, according to IPDC is independent of any concentration term.

The above rate expression assumes homogenous exchanger phase where transport of ions takes place only in the solid phase. Such an assumption is valid for a gel-type resin but is truly debatable for macroporous exchangers. For a macroporous resin, solid phase and macropores (in the order of 500 Å) coexist in the interior of the resin and, therefore, parallel diffusion of counterions in the solid phase and also in the macropore is a strong possibility. Effective diffusivity in such cases tends to be concentration dependent.[35–38] Price, Helditch, and Streat studied this aspect quite rigorously with respect to Ni^{2+}–Na^{+} exchange for a macroporous chelating exchanger (ES 467, Duolite) with aminophosphonate functionality. Some of the salient findings of the study are:[37]

- Experiments carried out at different temperature (5–55°C) and different stirring speeds (200–3000 rpm) clearly indicated that the kinetics is

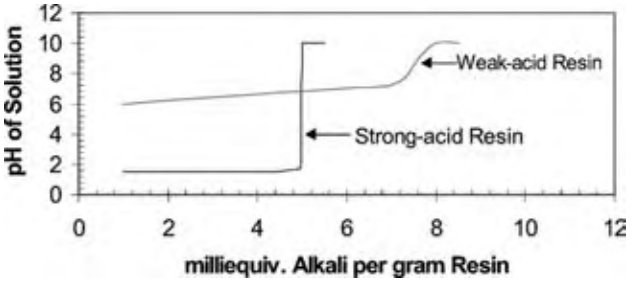


Fig. 7 pH titration curve of a strong-acid and a weak-acid cation exchange resin. *(View this art in color at www.dekker.com.)*

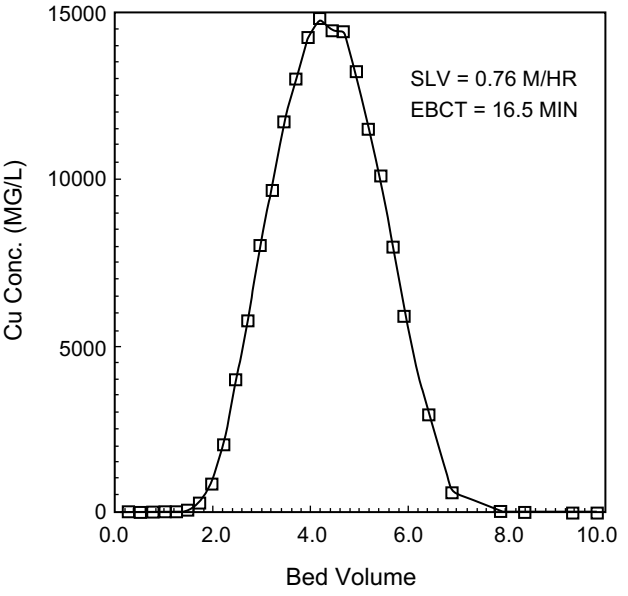


Fig. 8 Demonstration of high regeneration efficiency of copper(II)-loaded IRC-718 (imino-diacetate functionality) with 2% HCl.

IPDC. This was further confirmed by plotting $t_{1/2}$ vs. ion exchanger bead radius, r_0; $t_{1/2}$ varied with r_0^2 in agreement with Eq. (17).

- Aqueous-phase Ni concentration, however, was found to have a pronounced effect on ion exchange kinetics, which is not consistent with prediction from the IPDC model based on the Nernst–Planck equation. Experimental results could, however, be accounted for with the aid of a macroporous model, originally developed by Yoshida and Kataoka based on the assumption of parallel diffusion of counterions in the solid-gel phase and in the macropores.[38] This model predicts the concentration dependence of the effective self-diffusion coefficient, D_e, of counterion "i'' according to the following equation:

$$D_e = \frac{D_p\varepsilon_p + D_g(1-\varepsilon_p)Q/Z\,C_0}{\varepsilon_p + (1-\varepsilon_p)Q/ZC_0} \tag{18}$$

where D_p = diffusivity in the pore space, D_g = diffusivity in the gel phase, ε_p = void

Table 2 Effective diffusion coefficients for zinc ion in sulfonic acid cation resins of different cross-linking

Cross-linking (% DVB)	D_{Zn} (cm^2/sec)
5	3.7×10^{-7}
10	3.8×10^{-8}
16.2	8.1×10^{-9}
23.8	3.2×10^{-9}

(From Ref.[32].)

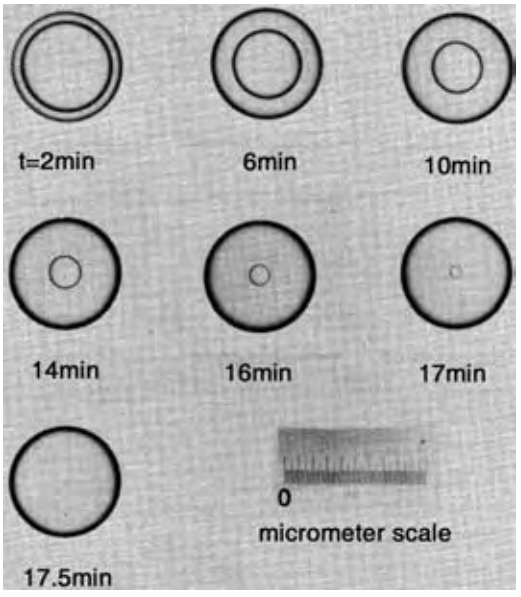


Fig. 9 Development of the Ca^{2+}–H^+ exchange for Amberlite IRC-84 (weak-acid cation exchange resin) in a 1.0 M HNO_3 solution.

fraction of pores, Q = total exchange capacity per unit volume, Z = charge of counterion i, and C_0 = concentration of i in the aqueous phase.

An increased counterion concentration in the aqueous phase has a favorable effect on kinetics for macroporous resins, i.e., $t_{1/2}$ decreases with an increase in aqueous-phase nickel concentration. This feature helps macroporous resins during regeneration also as the concentration of the regenerant (acid or alkali or salt) is usually very high (2–10%).

For gel-type resins, macropores are absent and intraparticle diffusion through the gel phase is the only mechanism for transport of the counterions. Hoell provided an optical verification to this effect for a gel resin with carboxylate functionality (IRC-84, Rohm & Haas Co.) using Ca^{2+}–H^+ exchange.[39] The exchanger was originally loaded with calcium ions and then placed in an acid solution; Ca^{2+} ions from the gel phase were slowly eluted out by hydrogen ions, progressing from the circumference toward the center. With time, the calcium-loaded core gradually shrank and eventually disappeared. Fig. 9 provides an excellent photographic testimony of this hypothesis.[31,39]

Helfferich discussed four scenarios of ion exchange kinetics and highlighted how the premises of some mathematical models are incompatible with facets of physical reality.[40] The specific cases considered were:

1. Nonlinear concentration gradients in Nernst–Planck film model.
2. Shell-core model in the absence of a mechanism that can produce shell-core behavior.
3. Reaction-controlled shell-core model.
4. Model for macroporous beads where diffusion in both the gel-phase and the macropores is comparable.

The guidelines proposed in this communication may serve as a basis to verify the applicability of any mathematical model under a given operating condition.[40]

MULTICOMPONENT ION EXCHANGE

In a typical ion exchange process the resin is loaded on a column and the influent containing the pollutant ions is passed through this bed during the exhaustion phase. It is common to find benign ions along with the pollutant ions in the influent. In fact, the concentration of benign ions is almost always greater than the pollutant ions. It is also expected that the ion exchange resin selected has maximum selectivity toward the most pollutant ion, i.e., the ion that needs to be removed to the maximum degree. Thus, we have a situation of an ion exchange resin in contact with n ions, each of different concentration and each having different selectivity toward the resin. In most applications, the ion that is preloaded on the resin (and the one that is exchanged for ions from the influent) has the least selectivity. When such a system is run, the effluent profile of each ion is different, and some ions undergo chromatographic elution, i.e., their effluent concentration may be higher than the influent concentration.

Table 3 Feedwater composition and binary separation factor ($\alpha_{Cu/ion}$) for the resin

		Feed water composition			
	Component	meq/L	Fraction	Presaturant composition (%)	α-value
1	Cu^{2+}	0.300	0.13043	0.00	1.00
2	Ni^{2+}	0.400	0.17391	0.00	3.00
3	Ca^{2+}	0.600	0.26087	0.00	5.00
4	Na^{2+}	1.000	0.43478	100.00	10.00
Total feed concentration		2.300			

Table 4 Equivalent fraction of mobile-phase composition in the four zones

	Component	1	2	3	4
1	Cu^{2+}	0.13043	0.00000	0.00000	0.00000
2	Ni^{2+}	0.17391	0.23698	0.00000	0.00000
3	Ca^{2+}	0.26087	0.30091	0.48226	0.00000
4	Na^{+}	0.43478	0.46211	0.51774	1.00000

Zone 1 = Na^{+}—presaturant, Zone 2 = Ca^{2+}—rich, Zone 3 = Ni^{2+}—rich, and Zone 4 = Cu^{2+}—rich according to Fig. 10.

As this can create a highly unfavorable situation, it is important to understand the following rules as they apply to multicomponent ion exchange:[41–43]

- The most preferred species will be the last to exceed the column and its effluent concentration will never exceed the influent concentration, i.e., it will never undergo chromatographic elution.
- The species exit the column in reverse preference order, with the least preferred species exiting first. Thus it is good practice to preload the ion exchange resin with the ion that is least preferred.
- The less preferred species will be concentrated in the ion exchange column and at some time will exit the column in concentrations exceeding their influent concentration. They will undergo chromatographic elution.
- When all the breakthrough fronts have exited the column, the entire bed is in equilibrium with the influent. At this time the effluent and the influent concentrations are equal.
- The effluent concentration of the presaturant ion decreases in steps as each new ion breaks through because the total equivalent concentration must be the same throughout the column run.

Let us visualize this through an example.

Let us assume a hypothetical influent with the following concentrations of cations:

Copper, Cu^{2+}	9.53 mg/L
Nickel, Ni^{2+}	11.74 mg/L
Calcium, Ca^{2+}	12.02 mg/L
Sodium, Na^{+}	23.0 mg/L

The equivalent concentration of each ion and the binary separation factors for each pair are provided in Tables 3–5. Please note that:

1. The ion exchange resin prefers copper the most.
2. The ion exchange resin prefers sodium the least and it is preloaded on the resin.

The ideal resin concentration profile inside the column is shown in Fig. 10. Based on the rules enumerated above, sodium exits at the beginning of the column run and for the first bed volumes, only sodium is present in the effluent at a concentration of 52.9 mg/L. The effluent breakthrough profile is shown in Fig. 11. Thus, sodium undergoes chromatographic elution. At bed volume 3520, calcium, the ion with the second-last preference for the resin, starts to exit from the resin bed. The emergence of calcium is simultaneous with a drop in sodium concentration, because the total equivalent concentration of the effluent has to be equal to its influent concentration, 2.3 meq/L. From 3520 to 4691 bed volumes, both sodium and calcium experience chromatographic elution. At bed volume, nickel appears in the effluent concomitant with a drop in sodium and calcium effluent concentration. From 4691 to 9184 bed volumes, sodium, calcium, and nickel experience chromatographic elution. At 9184 bed volume, copper appears in the effluent concomitant to a drop in the effluent concentration of sodium, calcium, and nickel. At this time the resin bed is exhausted and should be regenerated. Details of column cycle stages are explained in the entry ''Ion Exchange.'' From Fig. 11 it is clear that if both nickel and copper are to be removed from the influent, the column has to be stopped before nickel appears in the effluent (i.e., before 4691 bed volumes of operation)

Table 5 Equivalent fraction of solid-phase composition in the four zones

	Component	1	2	3	4
1	Cu^{2+}	0.45918	0.00000	0.00000	0.00000
2	Ni^{2+}	0.20408	0.42610	0.00000	0.00000
3	Ca^{2+}	0.18367	0.32463	0.65071	0.00000
4	Na^{+}	0.15306	0.24927	0.34929	1.00000

Zone 1 = Na^{+}—presaturant, Zone 2 = Ca^{2+}—rich, Zone 3 = Ni^{2+}—rich, Zone 4 = Cu^{2+}—rich according to Fig. 10.

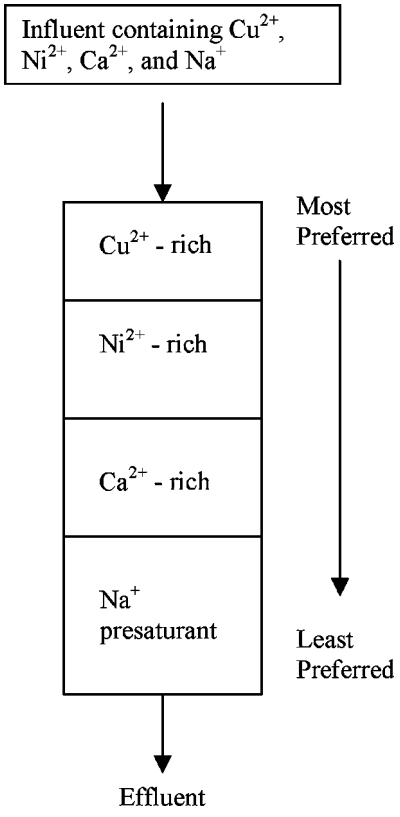


Fig. 10 Concentration profile of various ions inside the resin column.

because otherwise nickel would appear in the effluent at a concentration higher than the influent.

DEVELOPMENTS IN OVERCOMING SHORTCOMINGS OF CONVENTIONAL ION EXCHANGE PROCESSES

Particle Size Effects and Recoflo® Short-Bed Process

Like other heterogenous processes using small (<1 mm) particles, various reactor configurations have evolved for ion exchange, viz., packed bed, continuous countercurrent, fluidized bed, contact followed by sedimentation, etc. However, the packed-bed or fixed-bed process where the mobile liquid passes through stationary ion exchange beads in a column is by far the most popular unit operation because of its simplicity of construction and operation. This method is routinely used for metal ion removal, water softening, and water demineralization. The following steps are followed in a fixed-bed process:

- Exhaustion or removal of the contaminant
- Backdating
- In situ regeneration
- Rinsing
- Return to exhaustion

Timewise, exhaustion is much longer than all the other steps combined, and is practically equal to cycle time, t_{cycle}. Particle diameter in a packed bed influences both pressure drop and kinetics of ion exchange. For rigid, spherical particles like ion exchange beads, the pressure drop under laminar conditions can be estimated from the following equation:[44]

$$\Delta P = \frac{\mu v L}{K_p d_p^2} \tag{19}$$

where K_p is the permeability, μ is the viscosity of the liquid stream, L is the length of the packed bed, and v is the superficial liquid-phase velocity. Note that the pressure drop is inversely proportional to d_p^2. This is the reason why smaller diameter particles cannot be used in a packed bed. Assuming μ, v, and K_p to remain constant, the pressure drop remains the same if L/d_p^2 is kept constant. If, in addition, the L/t_{cycle} is kept constant, the capacity of the two columns of different heights will be the same even though the shorter column will exhaust faster.[45]

Poor kinetics is one of the major limitations of metal-selective ion exchange resins and as pointed out earlier, intraparticle diffusion is often the rate-limiting step. During the exhaustion cycle, there are three specific zones for a solute in a fixed bed, viz., saturated, unused, and mass-transfer zone (MTZ), as shown in Fig. 12. As the cycle advances, MTZ moves from the inlet toward the exit of the column. For a favorable isotherm, L_{MTZ} does not change during its passage inside the column but is dependent on the size of the adsorbent particle.[41,46] For intraparticle diffusion, L_{MTZ} is proportional to d_p^2. Improving kinetics in a fixed-bed process essentially means reducing L_{MTZ}. Thus, both hydrodynamic (high-pressure drop) and mass transfer (intraparticle diffusion) limitations in a fixed-bed process can be overcome by reducing the size of the polymer beads in accordance with the following guidelines:

1. If d_p is reduced in a way that L/d_p^2 remains constant, pressure drop across the bed does not change. In addition, if L/t_{cycle} is kept constant the column will have the same capacity as any larger column.
2. With a reduction in particle size, d_p, mass transfer rate by intraparticle diffusion will be enhanced and the length of the MTZ, L_{MTZ}, will be reduced in a way so that L_{MTZ}/L remains

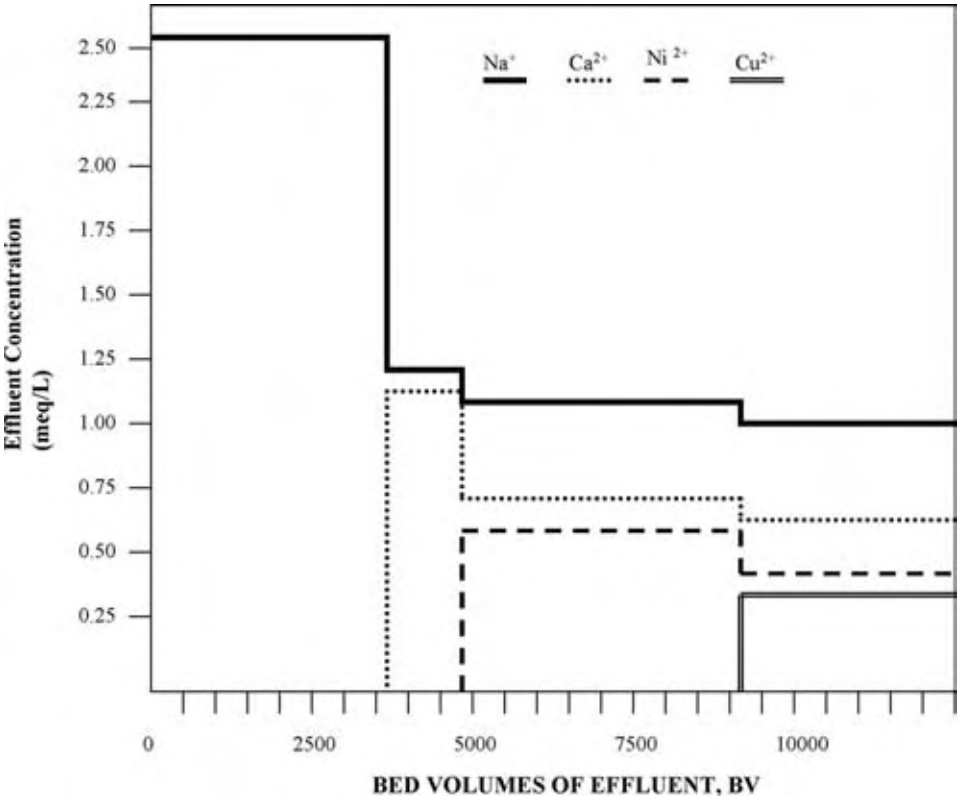


Fig. 11 Effluent history of the hypothetical influent solution based on the selectivity values provided in Table 3.

constant for a favorable isotherm, i.e., the fractional utilization of the column per cycle is unchanged.

The above theoretical principles form the basis of the ''Recoflo Short-Bed IX Process'' of ECO-TEC Limited, Ontario, Canada.[47,48] Recoflo uses much finer particle size (100–200 mesh) than is normally used in industrial ion exchange processes (20–40 mesh). Depending on the application, the total cycle time can vary from as low as 2 min to 1 hr. Consequently, the resin inventory for this process is low and that helps reduce the capital cost. On exhaustion, the short beds are regenerated in a countercurrent mode for maximum efficiency. The entire Recoflo system is assembled in a compact skid-mounted unit (Fig. 13). Poor kinetics and the relatively high price of chelating exchangers tend to limit their use for metals recovery. The improved kinetics (due to smaller particle size) and short-bed height characteristic of the Recoflo process greatly enhance the performance and economic viability of the chelating exchangers. They are now being used to recover metals from electroplating and hydrometallurgical wastestreams.[49] Grinstead has also used this principle for removing heavy-metal cations—particularly from dilute wastewater streams—and concentrating the heavy metals in the spent regenerant by over 1000 times.[28]

Continuous Annular Chromatography

We have discussed how metal ions can be removed from a dilute wastewater stream. Frequently, we come across a situation where the contaminated water contains several heavy metals. The methodologies outlined earlier can be effectively employed to remove all the heavy-metal cations, but the corollary is that the regenerant solution will have a mixture of all the heavy-metal cations. In other words, the heavy-metal cations are not separated in the regenerant stream. Such a mixture of salts of toxic metals has only a limited commercial value and there is a pressing need for separating the heavy-metal cations. Commonly encountered toxic metal cations, such as Cu^{2+}, Pb^{2+}, Ni^{2+}, Cd^{2+}, Zn^{2+}, etc., exhibit very similar physical and chemical characteristics and, therefore, separating one from another poses an obvious difficulty. A separation technique known as continuous annular chromatography (CAC) developed at Oak Ridge National Laboratory (ORNL) seems to show promise. In CAC, separation is achieved by using a slowly rotating annular bed of ion exchange resin, where feed is continuously introduced at a stationary point, while the eluent flows over the remainder of the annulus. The rotation of the sorbent bed coupled with the passage of the eluent causes the separated solutes of the feed to appear as helical bands, each of which has a

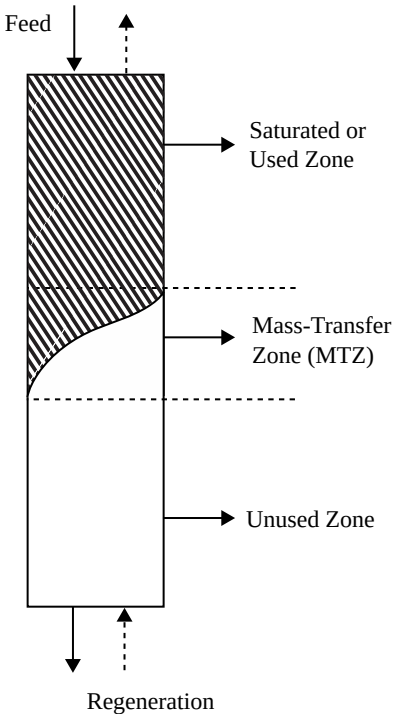


Fig. 12 Presence of three different zones in a fixed-bed column.

characteristic stationary exit point. Fig. 14 shows a conceptual diagram illustrating the key features of CAC, which is capable of achieving multicomponent chromatographic separation for a relatively concentrated mixture of metal salts. Various versions of a CAC apparatus have been constructed and tested experimentally at ORNL.[50–52] Multicomponent systems tried include the separation of Ni, Cu, and Co in ammoniacal solution, iron and ammonium in ammonium sulfate–sulfuric acid solution, and hafnium from zirconium in sulfuric acid solution. For the

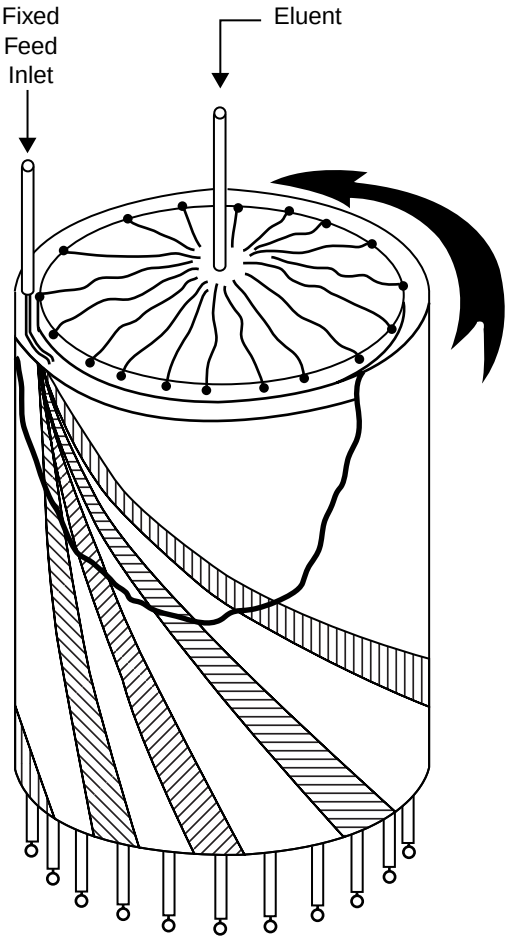


Fig. 14 Conceptual diagram illustrating the operation of the continuous annular chromatographic unit.

above-mentioned examples, Dowex 50W-X8, a polystyrene strong-acid (sulfonic group) cation exchanger resin with 8% DVB cross-linking from Dow Chemical, MI, was used as the stationary phase. Fig. 15 provides

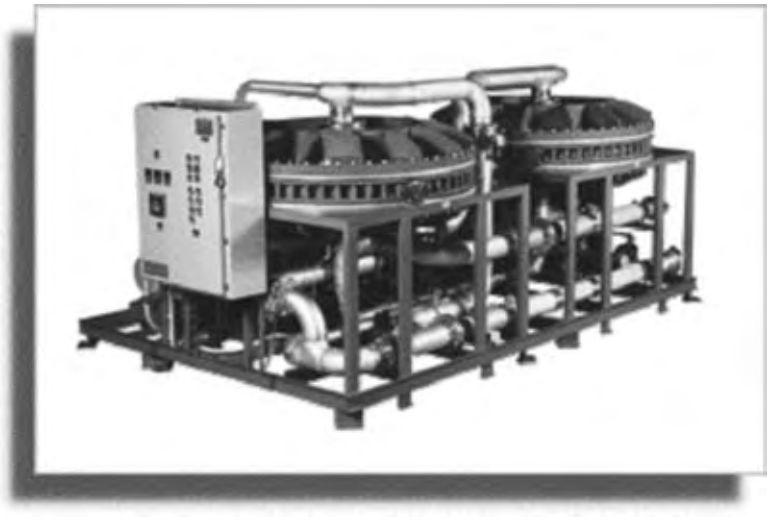


Fig. 13 A skid-mounted short-bed Recoflo unit. *(View this art in color at www.dekker.com.)*

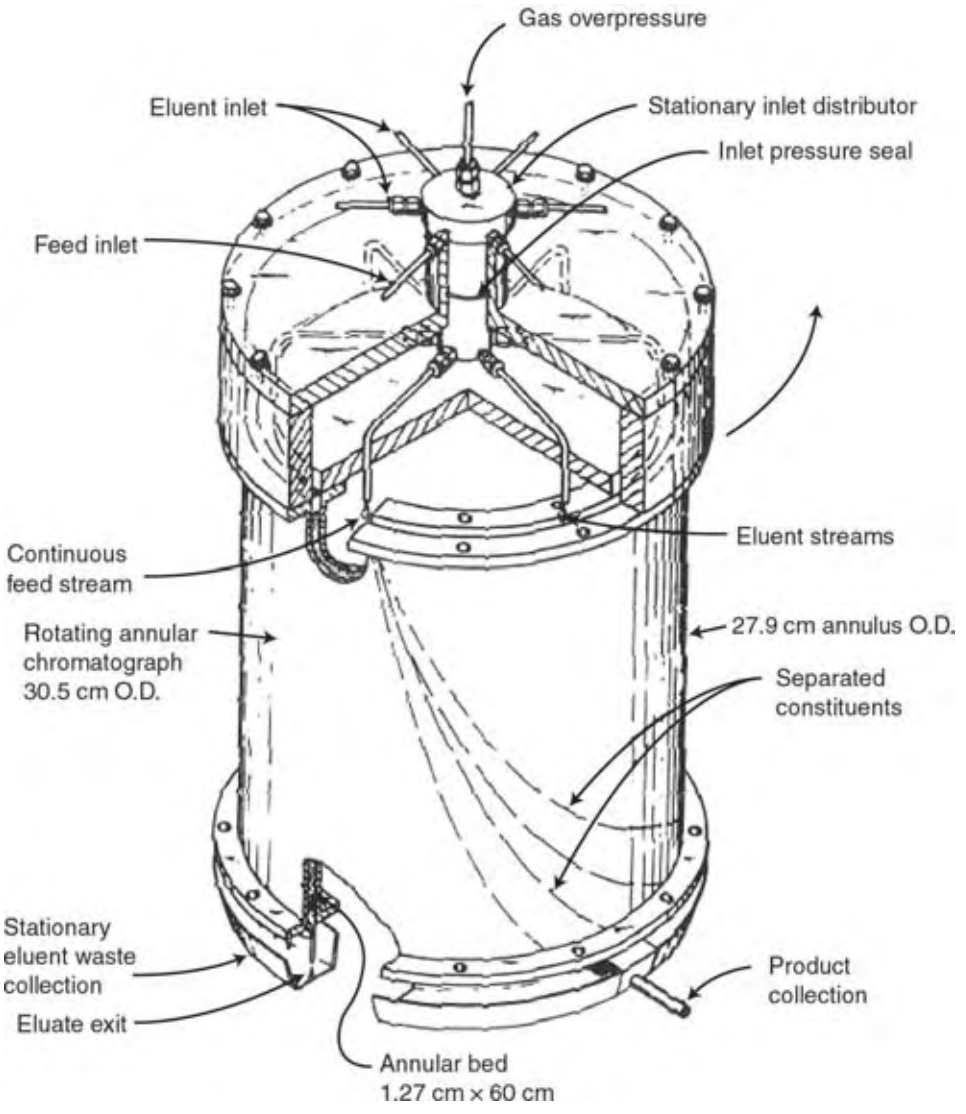


Fig. 15 Construction details of a continuous annular chromatographic unit used at ORNL.

some constructional details of CAC apparatus built for bench-scale testing at ORNL.

Fig. 16 shows separation of Cr(III) from Fe(III) with the CAC apparatus by Byers, Sisson, and Decarli II using ammonium as an isocratic eluent where the feed concentration was fairly high (5 gm/L).[52] Theoretically, elution and resolution of various components being separated can be predicted quite precisely using steady-state material balance and rate-limiting equations. From Fig. 16 one can note the excellent agreement of experimental results with the theoretical prediction (solid lines).

Using gradient elution, Byers, Sisson, and Decarli II have shown that the purity of Fe(III) and Cr(III) in the product can be increased by an order of magnitude.[52] However, all these studies were conducted with a strong-acid cation exchange resin, Dowex 50. Earlier discussion has indicated that chelating functionality may be used to improve the relative selectivity between the two metal ions, especially in the presence of ligands. Use of appropriate chelating exchangers may, therefore, broaden the application potential of CAC in separating heavy metals. Thus, CAC in conjunction with a proper choice of chelating ion exchange resins may be a useful tool in accomplishing this desired goal of heavy-metal–cation separation. Continuous annular chromatography apparatus is practically stationary and hence maintenance-free. Central treatment facilities, where wastewater streams are collected and treated for heavy-metal removal, are gaining momentum because of their overall economy and are being encouraged by the state and the federal regulatory authorities.[53] Using CAC in such facilities for separation and recovery of heavy metals may be a worthwhile venture in the coming years.

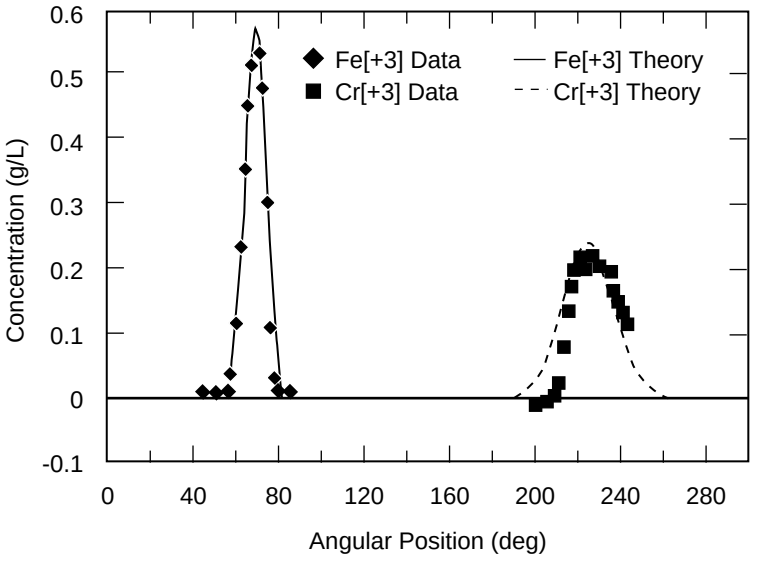


Fig. 16 Separation of Fe(III) and Cr(III) in a continuous annular chromatograph using 0.4 M $(NH_4)_2SO_4$ as the eluent.

CONCLUSIONS

Ion exchange resins have been used in a myriad of applications and based on trends, they will find newer application scenarios in the future. Present research in this field is primarily in the following areas:

1. *Size/morphology of the resin*: Pellicular resins that have diameter in the submicrometer range are being synthesized for applications where faster kinetics (of ion exchange during sorption and regeneration) is of prime importance, such as in chromatographic analysis. Nanoparticles of ion exchange resins are being developed by many researchers and because of their high surface area to volume ratio, their kinetics is faster than of conventional resins. Because of their nanosize, they can also be used in novel scenarios, such as in situ treatment of contaminated groundwater. Modifying the morphology has also been the focus of many researchers with the possibility of ion exchange membranes and composite sheets being used as a filtration material and under chemical/potential gradients.
2. Improving selectivity of the resin for target analytes. Many researchers have attempted to develop new resins that are targeted for certain analytes, such as ClO_4^-, As(V), Cr(VI), NOM, PCP, etc. The primary focus here is to improve the selectivity of the synthesized resin for targeted ion vis-à-vis the competing ions. This enhanced selectivity value is needed because the target ions have an extremely low concentration compared to the competing ions, but are needed to be removed to almost nondetectable levels (the guideline for ClO_4^- concentration in drinking water source aquifer is <1 ppb).
3. *Improving resin regeneration efficiency*: Since the late 1950s when ion exchange resins were used in industrial applications, researchers have tried to optimize regenerant solutions for specific resins operating under specific influent solutions, but presently many researchers have been attempting to increase manifold the regeneration efficiency of a certain resin by taking advantage of specific reactions between specific ions or by manipulating the regenerant dielectric constant.

REFERENCES

1. Helfferich, F.G. *Ion Exchange*; McGraw-Hill: New York, 1962.
2. James, M. Montgomery, Consulting Engineers, Inc. Ion exchange and demineralization. In *Water Treatment Principles and Design*; John Wiley & Sons: New York, 1985.
3. Clifford, D.A. Ion exchange and inorganic adsorption. In *Water Quality and Treatment*: *A Handbook of Community Water Supplies*; Pontius, F.W., Ed.; McGraw-Hill Inc.: New York, 1990.
4. Kunin, R. *Ion Exchnage Resins*; John Wiley & Sons: New York, 1958.
5. Reichenberg, D.; McCauley, D.J. Cation exchange equilibria on sulfonated polystyrene resins of varying degrees of crosslinking. J. Chem. Soc. **1955**, *2741*.
6. Clifford, D.; Weber, W. The determination of divalent/monovalent selectivity in anion exchangers. React. Polym. **1983**, *1*, 77.

7. Chu, B.; Whitney, D.C.; Diamond, R.M. Toward anion exchange resin selectivities. Inorg. Nucl. Chem. **1962**, *24*, 1405.
8. Kunin, R. Acrylic-based ion exchange resins and adsorbents. In *Amber Hi-Lites*; Rohm & Haas Co.: Philadelphia, PA, 1984; Vol. 173.
9. Sengupta, A.K.; Roy, T.; Jessen, D. Modified anion exchange resins for improved chromate selectivity. React. Polym. **1988**, *9*, 293.
10. Cotton, F.A.; Wilkinson, G. *Advanced Inorganic Chemistry*; Interscience Publishers: New York, 1972.
11. Pepper, K.W.; Hale, D.K. *Ion Exchange and Its Applications*; Society of Chemical Industry: London, 1955; 13.
12. Roy, T.K. Chelating Polymers: Their Properties and Applications in Relation to Removal, Recovery, and Separation of Toxic Metal Ions. M.S. thesis, Civil Engineering Department, Lehigh University, Bethlehem, PA, 1989.
13. Waitz, W.H. Ion exchange in heavy metals removal and recovery. In *Amber-Hi-Lites*; Rohm & Haas Co.: Philadelphia, PA, 1979; Vol. 162.
14. Bolto, B.A.; Pawlowski, L. *Wastewater Treatment by Ion Exchange*; E & F. N. Spon: London, 1987.
15. Janauer, G.E.; Gibbons, R.E.; Bernier, W.E. *Ion Exchange and Solvent Extraction*; Marinsky, J.A., Marcus, Y., Eds.; Marcel Dekker: New York, 1985; Vol. 9, Chapter 2.
16. Woelders, J.A.; Urlings, L.G.C.M.; Vanderpiji, P.P. In-situ remedial action of cadmium-polluted soil by ion exchange. *Proceedings of the International Symposium on Ion Exchange for Industry*; Streat, M., Ed.; Ellis Horwood: England, 1988; 169.
17. Sengupta, A.K.; Millan, E.; Roy, T. Potential of ion exchange resins and reactive polymers in eliminating/reducing hazardous contaminants. Proceedings of the 2nd International Conference on Physicochemical and Biological Detoxification of Hazardous Wastes, Atlantic City, NJ, May 3–5, 1988; 191.
18. Matejka, Z.; Zitkova, Z. The sorption of heavy-metal cations from EDTA complexes on acrylamide resins having oligo(ethyleneamine) moieties. React. Funct. Polym. **1997**, *35*, 81–88.
19. Courduvelis, C. New developments for the treatment of wastewater containing metal complexes. Proceedings AES 4th Conference on Advanced Pollution Control for the Metal Finishing Industry, Florida, 1982.
20. Marton-Schmidtt, E., et al. Separation of metal ions on ion exchange resin with ethylenediamine functional groups. J. Chromatogr. **1980**, *201*, 73–77.
21. Loureiro, J.M., et al. Recovery of copper, zinc, and lead from liquid streams by chelating ion exchange resins. Chem. Eng. Sci. **1988**, *43*, 1115–1124.
22. Warshawsky, A. Modern research in ion exchange. In *Ion Exchange*: *Science and Technology, NATO AISI Series*; Rodrigues, A.E., Ed.; Martinus Nijhoff Publishers: Boston, MA, 1986.
23. Warshawsky, A. Extraction of platinum group metals by ion exchange resins. In *Ion Exchange Processes in Hydrometallurgy*; Streat, M., Nadan, D., Eds.; Society of Chemical Industry, John Wiley and Sons, 1987; Vol. 19.
24. Hudson, M. Coordination chemistry of selective ion exchange resins. In *Ion Exchange*: *Science and Technology*; Rodrigues, A.E., Ed.; NATO AISI Series; Martinus Nijhoff Publishers: Boston, 1986.
25. Melling, J.; West, D.W. Comparative study of some chelating ion exchange resins for applications in hydrometallurgy. Proceedings of the 4th International Conference on Ion Exchange Technology at the University of Cambridge; England, Jul, 1984, 724–735.
26. Marinsky, J. Equations for the evaluation of formation constants of complexed ion species in crosslinked and linear polyelectrolyte systems. In *Ion Exchange and Solvent Extraction*; Marinsky, J., Marcus, Y., Eds.; Marcel Dekker Inc.: New York, 1973; Vol. 4.
27. Kokoszka, L.C.; Flood, J.W. *Environmental Management Handbook*: *Toxic Chemical Materials and Wastes*; Marcel Dekker, Inc.: New York, 1989; Chapter 2.
28. Grinstead, R.R.; Paalman, H.H. Metal ion scavenging from water with fine-mesh ion exchangers and microporous membranes. Environ. Prog. **1989**, *8* (1), 35.
29. Nativ, M.; Goldstein, S.; Schumuckler, G. Kinetics of ion exchange processes accompanied by chemical reactions. J. Inorg. Nucl. Chem. **1975**, *37*, 1951.
30. Helfferich, F.; Plesset, M.S. Ion exchange kinetics: a non-linear diffusion problem. J. Chem. Phys. **1958**, *28*, 411.
31. Hoell, W. Optical verification of ion exchange mechanisms in weak electrolyte resins. React. Polym. **1984**, *2*, 93–101.
32. Boyd, G.; Soldanho, B.A. Self diffusion of cations in and through sulfonated polystyrene cation exchange polymers. J. Am. Chem. Soc. **1953**, *75*, 6091.
33. Helfferich, F. Ion exchange kinetics, V: ion exchange accompanied by reaction. J. Phys. Chem. **1965**, *69*, 1178.
34. Schmuckler, G. Kinetics of moving-boundary ion exchange processes. React. Polym. **1984**, *2*, 103.

35. Yoshida, H.; Kataoka, T.; Ikeda, S. Intraparticle mass transfer in bidispersed porous ion exchanger, part I. Isotopic ion exchange. Can. J. Chem. Eng. **1985**, *62*, 422.
36. Yoshida, H.; Kataoka, T.; Fujikawa, S. Kinetics in a chelate exchanger I. Chem. Eng. Sci. **1986**, *41*, 2525.
37. Price, S.G.; Helditch, D.J.; Streat, M. Diffusion or chemical kinetics control in a chelating ion exchange resin system. In *Ion Exchange for Industry*; Streat, M., Ed.; Ellis Horwood Ltd.: Chichester, U.K., 1988.
38. Yoshida, H.; Kataoka, T. Intraparticle mass transfer in bidispersed porous ion exchanger, part II. Mutual ion exchange. Can. J. Chem. Eng. **1985**, *62*, 430.
39. Hoell, W.; Sontheimer, H. Ion exchange kinetics of the protonation of weak-acid ion exchange resins. Chem. Eng. Sci. **1977**, *32*, 755.
40. Helfferich, F. Models of physical reality in ion exchange kinetics. React. Polym. **1990**, *13*, 191–194.
41. Helfferich, F.; Klein, G. *Multicomponent Chromatography*: *Theory of Interference*; Marcel Dekker Inc.: New York, 1970.
42. Clifford, D.A. Multicomponent ion exchange calculations for selected ion separations. Ind. Eng. Chem. Fundam. **1982**, *21*, 141–153.
43. Snoeyink, V.L.; Cairns-Chambers, C.; Pfeffer, J.L. Strong-acid ion exchange for removing barium, radium, and hardenss. J. AWWA **1987**, *79* (8), 66–72.
44. Bird, R.B.; Stewart, W.E.; Lightfoot, E.N. *Transport Phenomena*; John Wiley & Sons: New York, 1960.
45. Wankat, P.C. Efficient fractionation by ion exchange. In *Ion Exchange Science & Technology*; Rodrigues, A.E., Ed.; NATO ASI Series; Martinus Nijhoff Publishers: Boston, MA, 1986.
46. Kovach, J.L. *Handbook of Separation Techniques for Chemical Engineers*; Schweitzer, P.A., Ed.; McGraw-Hill Inc.: New York, 1979; Chapter 3.
47. Brown, C.J. U.S. Patent 4,673,507, 1987.
48. Brown, C.J.; Fletcher, C.J. The recoflo short-bed ion exchange process. In *Ion Exchange for Industry*; Streat, M., Ed.; Ellis Horwood: Chichester, 1988.
49. Brown, C.J. Acid and metals recovery by Recoflo short-bed ion exchange. In *Separation Processes in Hydrometallurgy*; Davies, G.A., Ed.; Ellis Horwood: Chichester, U.K. 1987.
50. Scott, C.D.; Spence, R.D.; Sisson, W.G. Pressurized annular chromatograph for continuous separations. J. Chromatogr. **1976**, *126* (381), 400.
51. Begovich, J.M.; Byers, C.H.; Sisson, W.G. A high-capacity pressurized continuous chromatograph. Sep. Sci. Technol. **1983**, *18*, 1167.
52. Byers, C.H.; Sisson, W.G.; Decarli, II, J.P. The use of gradient elution in optimizing continuous annular ion exchange chromatography with applications to metal separations. In *Ion Exchange for Industry*; Streat, M., Ed.; Ellis Horwood: Chichester, U.K.
53. Chen, J.J. Metro recovery systems—a centralized metals recovery and treatment facility in twin cities, U.S.A. Proceedings of the Second International Symposium on Metals Speciation, Separation, and Recovery, Rome, Italy, May 1989.

Latex Processing

Alexander F. Routh
Department of Chemical and Process Engineering, University of Sheffield, Sheffield, U.K.

INTRODUCTION

Latex consists of tiny particles of polymer dispersed in liquid. It is naturally occurring and also made synthetically in large volumes. Latex is the major component of paints as well as being used in paper, textiles and adhesives, with an annual market for these polymer dispersions in excess of $20 billion.

Although available as a natural product from rubber trees, demand has necessitated synthetic production of latex. The major use of latex is in the formation of polymer films and a common problem with latex is ensuring stability prior to use.

This entry describes some common polymers used in latex formulations and standard synthesis methods. Particle interactions and reasons for the loss of stability will be reviewed and mechanisms of film formation will be outlined.

HISTORY OF LATEX

The origins of latex are from the rubber tree (Hevea Brazilliensis), which originated in the Amazon rain forest. From the second half of the 19th century, plantations were developed predominantly in South East Asia to harvest this natural product.[1,2]

Although Dupont had been making polychloroprene rubbers since 1920, synthetic manufacture of latex was extensively developed, during World War II, in the United States. The development was in response to the Japanese conquest of South East Asia and the subsequent shortage of rubber for tyres and inner tubes, with demand greatly exacerbated by the war. Rubber was made a ''strategic and critical material'' on June 28th 1940 and the Rubber Reserve Company was set up to produce raw materials, manufacture and sell synthetic rubbers.[3] By the late 20th century, industrial production of ''artificial'' latex was developed with 500 million kg of polymer produced as latex annually.[4]

USES

By far the largest use for latex is in emulsion paints with a common household paint, being around 50% polymer in aqueous solution. Latex is also used commonly in the manufacture of protective gloves, as an adhesive, as an additive in paper and textiles, and in the manufacture of foam rubber.[5]

DEFINITIONS

Latex is a dispersion of polymer particles in a liquid medium, where the particles will remain suspended indefinitely. This property means that latices are colloidal dispersions. By nature of its origin, latex is classified into *natural* latex for dispersions obtained from plants, and *synthetic* latex for dispersions that are man made, typically by a process called emulsion polymerization. Blackley[2] discusses a number of further classifications including *artificial* latex for dispersions in which the polymer is dispersed after synthesis, and *modified* latex where a chemical modification of existing latex is made.

EXAMPLES OF POLYMERS USED IN LATICES

The range of monomers available and copolymerization in varying compositional blends makes the number of possible polymer variants immense. Acrylic and styrene-acrylic copolymers are the most widely used polymers in paints and other coatings. According to Lesko[4] acrylic polymers ''account for 30% of commercial emulsion polymers.'' Other polymers commonly used in film forming latices are polychloroprene and poly(vinyl acetate). Polystyrene–cobutadiene latices are mixed with cement and sand to form rubberized cement, as well as being used as additives in textiles and paper industries, and in the manufacture of foam rubber.

The mechanical properties of polymers are very sensitive to temperature, hence an important design property is the glass transition temperature of the constituent particles and, subsequently, the final product. Styrene homopolymer has a glass transition temperature of around 100°C and butyl acrylate a glass transition temperature of around –50°C. Hence copolymers of these two monomers have a glass transition temperature in a range from –50°C to 100°C,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007833

Table 1 Glass transition temperatures of a range of acrylic homopolymers

Monomer	Acronym	Homopolymer T_g (°C)
Methyl acrylate	MA	+8
Ethyl acrylate	EA	−22
n-Butly acrylate	BA	−54
i-Butyl acrylate	IBA	−43
Methyl methacrylate	MMA	+105
Ethyl methacrylate	EMA	+65
n-Butyl methacrylate	BMA	+20
Styrene	S	+100

(From Ref.[4].)

depending on the composition. Table 1 shows the glass transition temperature of a range of acrylate polymers.

A further design consideration is the resistance of each polymer to different environments. For example poly(methyl methacrylate) is highly resistant to the action of sunlight, while polystyrene has low resistance. Conversely, polystyrene has excellent resistance to water where poly(methyl methacrylate) is only fair. (From Ref.[4].)

The chemical structure of a range of common polymers is given in Fig. 1.

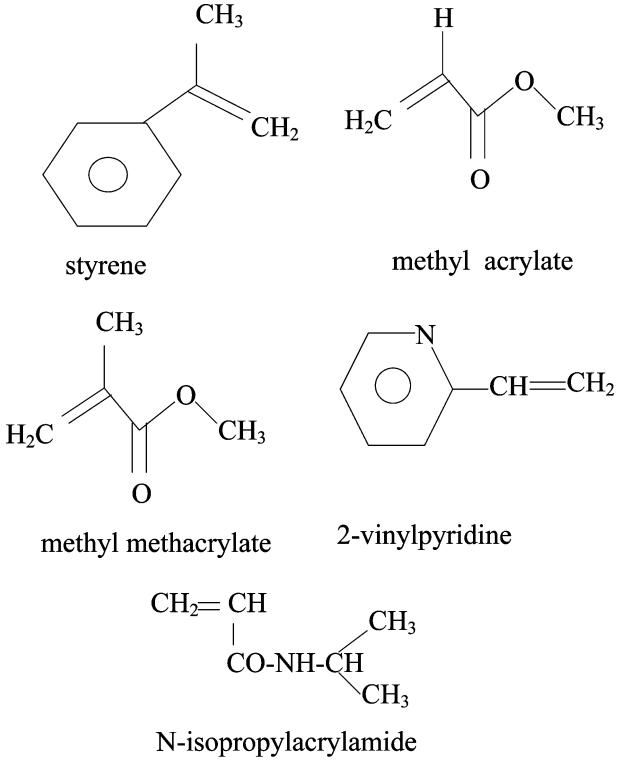


Fig. 1 Monomers commonly employed on latex production.

LATEX SYNTHESIS

The three main particle production methods are *emulsion*, *dispersion,* and *suspension* polymerization. In each case, the monomer is mixed with a continuous phase and an initiator. In addition a stabilizer in the form of surfactants may be required.

Emulsion Polymerization

In emulsion polymerization the monomer is barely soluble in the suspension medium, but the initiator is soluble. The monomer forms droplets of many microns in size that are surrounded by the stabilizing surfactant. Some monomer will dissolve in the bulk liquid and hence come into contact with the initiator, and thereby begin the polymerization. Further monomers will come into contact with the monomer radicals and polymer chains grow. These growing chains form the basis for particles and continue to grow by reacting with monomers in the bulk. The process continues until all the monomer is consumed. An example of emulsion polymerization is styrene in water with potassium persulfate as the initiator and sodium dodecylsulfonate as the stabilizer. For water-soluble monomers, the reaction can occur in organic media. Typically the monomer is present at volume fractions up to 50% and particles in the range of 50–300 nm are produced.[6]

It is not always necessary to use the stabilizing surfactant with surfactant free emulsion polymerization reported by Goodwin et al.[7]. The particles are stabilized by the surface charge on the particles resulting from fragments of the initiator used. In these cases, the monomer concentration needs to be drastically reduced to less than 5%, and particle sizes of between 100 and 1000 nm are reported.

The size of particles obtained from emulsion polymerization is determined by a balance between the rates of particle initiation and growth. The initial concentration of monomer and initiator, temperature, salt concentration, and the amount of emulsifier, affect the particle size. The particles produced are typically monodisperse with a polydispersity of less than 5%.

Dispersion Polymerization

In dispersion polymerization, the monomer and initiator are both soluble in the reaction medium, but the polymer chains are not. As the chains grow, they form micelles and continue growing. These micelles are primary particles and are typically very swollen with solvent and monomer, and hence further polymerization can occur inside the particles. The resulting particles are sterically stabilized by the addition of

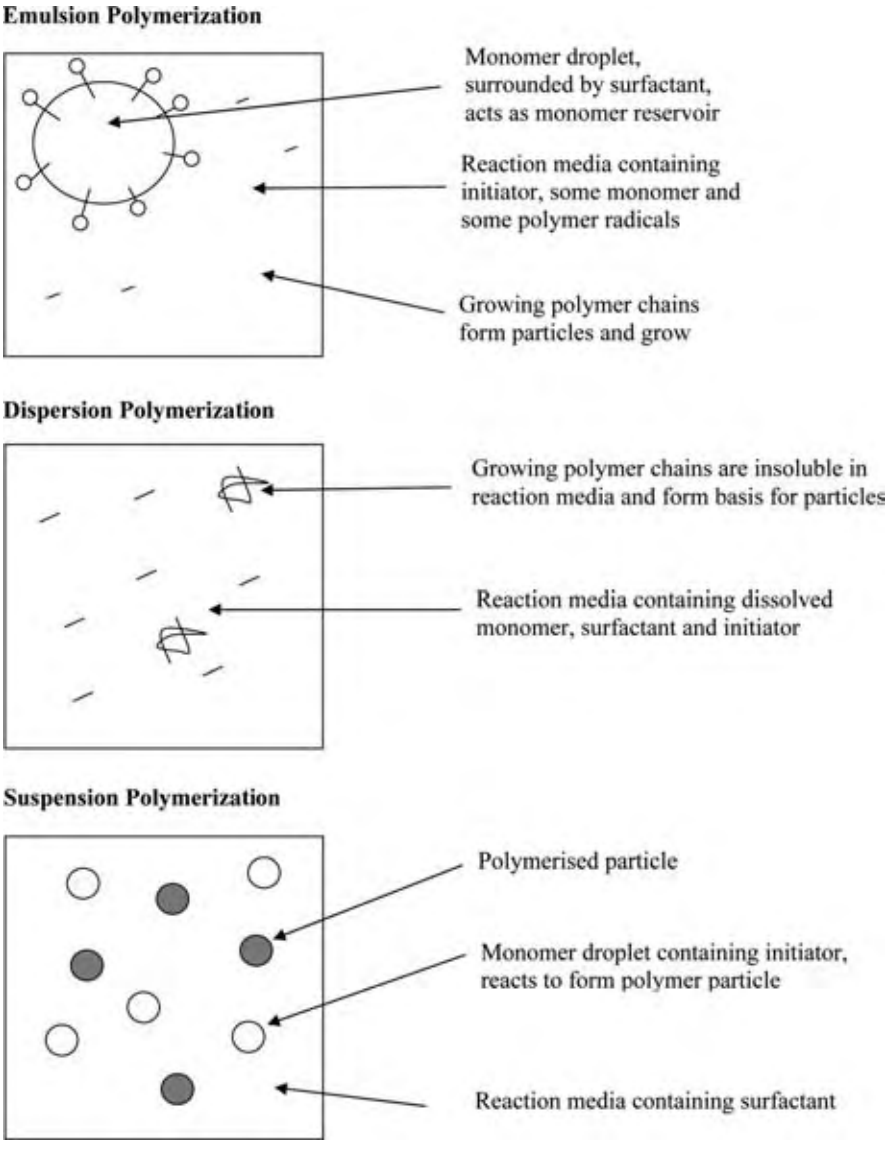


Fig. 2 Three methods to synthesize latex.

surfactants. Resulting particle sizes are in the range from 100 nm to 10 μm, the actual size being determined by the amounts of monomer, initiator and stabilizer, as well as the solvency of the reaction media. A typical dispersion polymerization reaction is styrene in ethanol using polyvinylpyrrolidone as a stabilizer and azo-bis-2-methylpropionitrile (AIBN) as the initiator.[6,8]

Suspension Polymerization

In suspension polymerization, both initiator and monomer are insoluble in the reaction medium, while the initiator is soluble in the monomer. Small droplets of the monomer are formed by stirring and the addition of surfactants whilst the reaction is usually initiated thermally. The polymerization of the monomer droplets leads to polymer particles of the same size as the original monomer droplet. A typical example is styrene and AIBN in water stabilized by polyvinylpyrrolidone and particles in the size range of 20 μm to 2 mm are routinely made. The volume fraction of the monomer is usually kept below 50%. Because of agitation during particle production, the resulting dispersions are generally polydisperse.[6–9]

The three synthesis methods are sketched in Fig. 2.

Pictures of Monodisperse Particles

Remarkable ordering is possible with highly monodisperse particles. This is demonstrated in Fig. 3, an SEM image of 200 nm radius polystyrene particles that were prepared by surfactant free emulsion polymerization.

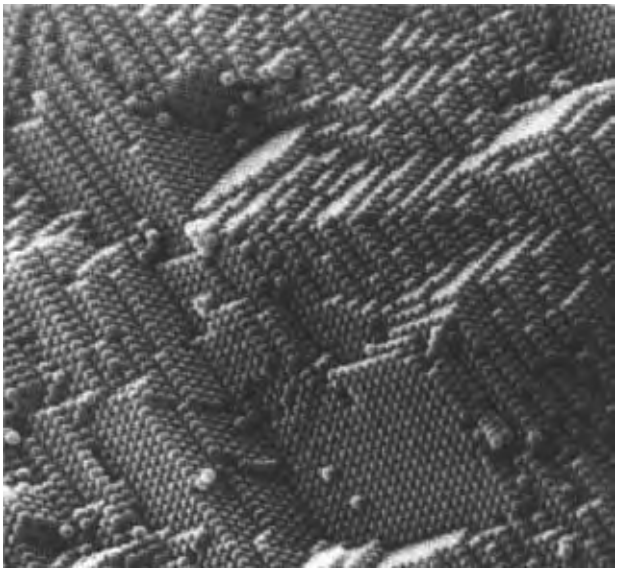


Fig. 3 Monodisperse polymer particles forming an ordered array. (From http://www.idelatex.com/body_pubsandservices.html, copyright Molecular Probes, Inc.).

RESPONSIVE POLYMERS—MICROGELS

A large body of work has recently investigated the properties and uses of responsive polymer latices. These particles, called microgels, swell in response to an external trigger, such as temperature or pH. A common microgel monomer is *N*-isopropylacrylamide, often referred to by its acronym of NIPAM. Particles composed of this monomer, cross-linked with methylene bisacrylamide, display a swelling transition at a temperature of 32°C, with particles being swollen at lower temperatures.[10] Other monomers commonly used to make microgels include styrene, divinylbenzene, acrylic acid, methacrylic acid and vinyl pyridine.[11]

Mechanism for Swelling

Poly NIPAM particles display a swelling below 32°C because water is a good solvent for the polymer at lower temperatures. The polymer chains try to dissolve, restricted only by the cross-linking holding the particles together. Vinyl pyridine based microgels swell below a pH of 3.5 because of protonation of the pyridine nitrogen that establishes a charge inside the microgel. Counter ions flood into the particle to balance the charge and the subsequent increase in osmotic pressure swells the particle. Again, a cross-linker (usually divinylbenzene) is required to stop the particle dissolving.[12] For acidic monomers swelling in the corresponding microgel occurs at high pH because of the osmotic pressure induced by dissociation.

Uses for Microgels

Microgels have found extensive use in the coatings industry as trigger dependent rheological modifiers. There are also reports of the use of PNIPAM particles for the uptake of heavy metals from wastewater and as enzyme supports, resulting in increased reaction rates.[13] The most exciting potential for microgels is as drug delivery agents. The challenges are numerous, with non-toxic responsive monomers required, but the promise of drug release in vivo on application of an external trigger is highly desirable.

STABILITY

A latex dispersion may remain invariant with time, in which case it is referred to as stable, or it may form a lump of polymeric material at the bottom of the container, in which case it is called unstable. It is obviously crucial to keep dispersions stable, at least over long enough time scales to allow their use. The sedimentation rate of a single latex particle is tiny, and Brownian motion is more than capable to keep particles dispersed. However, if particles stick together, or aggregate, they settle very quickly as sedimentation rate increases as the square of the floc size. This is the cause of dispersion instability: particles sticking together and then settling out of solution.

Interaction Potentials

Whether two isolated particles will come together or remain dispersed is determined by the energy of interaction between them. At close contact van der Waals interactions between individual molecules provide an infinite attraction meaning that all colloidal dispersion are, thermodynamically at least, unstable. Latex particles are charged because of electrical charges on their surfaces. This produces a repulsive interaction between particles that can provide a large energy barrier to aggregation and, thus, stabilize latex dispersions.

van der Waals Interactions

Two macroscopic bodies in solution have a mutual attraction. The basis for this is molecular: A dipole in body A will align in a certain orientation. This induces dipoles in body B to align in the opposite direction making charges of opposite sign closer to each other. The result is an attractive force between the two dipoles. When summed up over all possible dipoles in both macroscopic bodies, a net attractive energy is realized. This energy of interaction is dependent on the geometry and separation of the

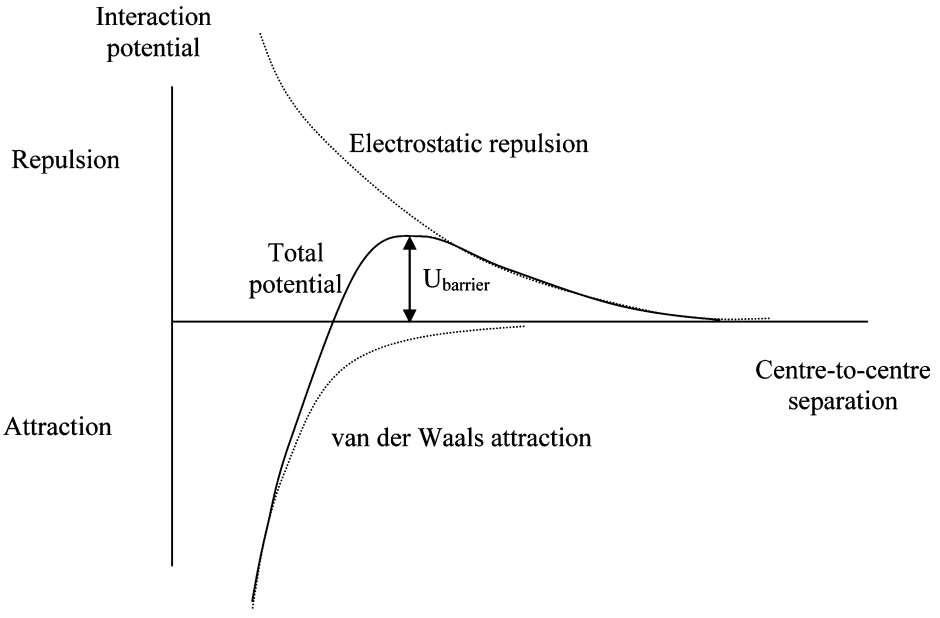


Fig. 4 Electrostatic and van der Waals potentials and a total interaction potential.

macroscopic bodies, as well as the medium they are dispersed in. For two spheres of radius R and center-to-center separation r, the interaction potential U_{vdw} is given by

$$U_{vdw} = -\frac{A}{6}\left(\frac{2R^2}{r^2-4R^2} + \frac{2R^2}{r^2} + \ln\frac{r^2-4R^2}{r^2}\right)$$

where A is called the Hamaker constant and depends on the polarizability of the particle material in the dispersed phase.[14]

Electrostatic Interactions

Two particles that have an electrostatic potential of the same sign will repel each other. The magnitude of this repulsion depends on the size of the potential and the properties of the surrounding fluid. If the surrounding solution contains a large number of ions the surface charge is screened and the inter-particle repulsion diminished. The range of electrostatic interaction between particles is called the Debye length and it is given the symbol κ^{-1}. As argued above, the Debye length is dependent on the local electrolyte concentration. For two spheres of surface potential ψ_s, radius R and center-to-center separation r the electrostatic potential U_{elec} is given by

$$U_{elec} = 4\pi\varepsilon\varepsilon_0\left(\frac{kT}{ze}\right)^2\frac{R^2}{r}\psi_s^2\exp[-\kappa(r-2R)]$$

where ε is the material permittivity, ε_0 the permittivity of free space, kT thermal energy, z the electrolyte valance and e the charge on an electron.[14] Addition of electrolyte reduces the Debye length (makes κ bigger) and reduces the effect of electrostatic repulsions.

DLVO Theory

The central theory for colloidal, and therefore latex, stability is because of the complimentary work of Derjaguin and Landau in Moscow and Verwey and Overbeek in Holland. This has become known as DLVO theory.[14] The idea is to represent a total energy of interaction as the sum of individual attractive and repulsive potentials. Fig. 4 sketches out the van der Waals and electrostatic potentials, as well as the total interaction for a particular particle size, surface potential, and electrolyte concentration.

The range of the electrostatic repulsion is determined by the external electrolyte concentration. If the electrostatics is over a longer range than the van der Waals attraction then the total interaction potential is repulsive at large separations. At short range the van der Waals attraction is always larger than the electrostatic repulsion and the total interaction potential is attractive. Therefore, a barrier to aggregation is present, marked as $U_{barrier}$ on Fig. 4. This is the energy individual particles require when they collide to be able to aggregate. It is analogous to the activation energy in chemical kinetics and depending on its magnitude provides a degree of kinetic stability to a latex dispersion. The addition of salt reduces the range of the electrostatic repulsion, allowing the van der Waals attraction to dominate. This diminishes the magnitude of the barrier to aggregation and destabilizes a latex dispersion.

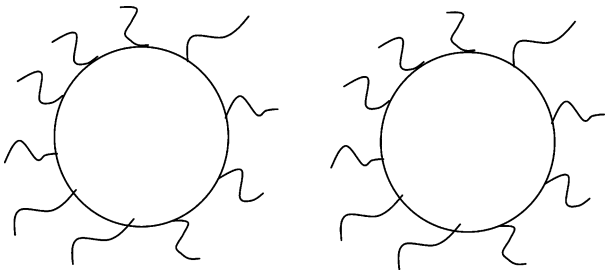


Fig. 5 Polymer chains grafted to particle surfaces provide a stabilizing repulsion: as particles approach, the chains overlap and repel.

Other Potentials

Many other interaction potentials are possible. The most common involve polymers added to systems.

Steric Stabilization

If polymer chains are attached onto the surface of particles they hinder aggregation and provide a degree of steric stability. The polymer chains must extend sufficiently into the solution, preventing approach of particles and hindering aggregation. This form of stability is especially useful in organic dispersions where electrostatic stabilization is not possible. An example is polystyrene particles with poly(ethylene oxide) chains terminally attached.[15] This is shown in Fig. 5.

Depletion Potential

A nonadsorbing polymer added to dispersions will often cause aggregation of particles. Asakura and Oosawa[16,17] were the first to describe the cause of this instability: as two particles approach the finite size of the polymer chains ensures their exclusion from the region between the two particles. The osmotic pressure is dependent on the concentration of macromolecule and hence it is diminished in this overlap region. The excess osmotic pressure in the main bulk fluid causes the two particles to be pushed together. This is called the depletion potential and has been extensively studied.[18–20] This is shown in Fig. 6.

AGGREGATION

If the interaction potential is such that aggregation does occur, then flocs of aggregated particles start to form. The two areas of interest are the rate of aggregation and the structures formed.

Aggregation Rate

The rate is determined by the diffusion of particles toward each other, and the likelihood of an individual collision achieving a stable bond. For strongly attractive potentials, such that every collision results in a stable bond, diffusion of particles towards each other is the slowest step in the aggregation process. In this case, the rate of aggregation is called the Smoluchowski rate[21] and the process called diffusion limited aggregation (DLA). The characteristic time for aggregation τ, scales inversely with the particle volume fraction ϕ as

$$\tau = \frac{\pi\eta R^3}{\phi kT}$$

where η is the viscosity of the continuous phase.

If a significant repulsion exists in the interaction potential then only a small percentage of particle–particle collisions will result in a stable bond. The characteristic time for aggregation is extended by a factor W, called the stability ratio, which takes into account this likelihood of ''reaction:''

$$\tau_{\text{slow}} = W\tau$$

The stability ratio is dependent on the total interaction potential U, such that

$$W = 2R\int_{2R}^{\infty}\frac{\exp(U/kT)}{r^2}dr$$

For very large stability ratios, where only a tiny fraction of collisions between particles result in a stable bond, the process is termed reaction limited aggregation (RLA).

Aggregate Structures

Beautiful fractal images are observed when latex particles aggregate. A fractal is a structure that replicates itself over a number of length scales and is defined by the relation

$$N = \left(\frac{a}{R}\right)^{d_f}$$

where N is the number of primary particles in the floc a is the size of the object and d_f is called the fractal dimension. The two limiting cases of aggregation (DLA and RLA) are found to have fractal dimensions of 1.75 and 2.10, respectively.[22–24]

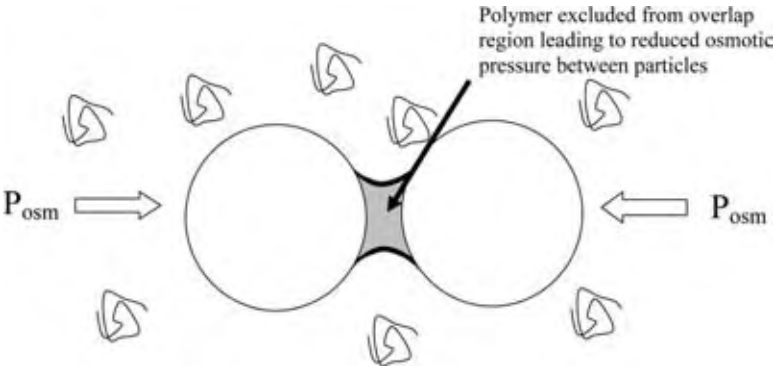


Fig. 6 Free polymer can cause destabilization: polymer, excluded from region between particles, exerts an osmotic pressure pushing particles together.

CHARACTERIZING PARTICLES

It is important to be able to characterize particles in a dispersion. Questions of interest include the particle size and charge. Ways to ascertain the glass transition temperature of the constituent polymer are discussed in the film formation section.

Dynamic Light Scattering

The easiest way to determine the size of colloidal particles is by dynamic light scattering. A laser is shone into a dilute dispersion and the intensity of light scattered at an angle, usually 90°, is followed. This scattered light is found to be self correlated for times shorter than the time it takes for the diffusion of a single particle out of the laser beam. From the self-correlation function the diffusion coefficient is obtained. The Stokes–Einstein relation relates the diffusion coefficient of particles D_0, to the thermal energy kT, fluid viscosity η and particle radius R through

$$D_0 = \frac{kT}{6\pi\eta R}$$

and hence the radius of colloidal particles is readily obtained.

Electrophoresis

Latex particles typically have a charge because of fragments of the initiator located on the surface. Additionally a bulk charge may result from ionizable moieties in the particle core.

In an electric field a charged particle will move towards the electrode of opposite charge. For a given electric field strength E, particles will acquire a velocity V, dependent on their mobility μ according to

$$V = \mu E$$

Moving particles will drag some fluid along with them. This is referred to as the added mass. There exists a radius at which the fluid does not move with the particle and this is called the slip plane. The electrical potential at this plane is lower in magnitude than the surface potential and is called the zeta potential, ζ. The magnitude of this potential determines the mobility. For very low electrolyte concentrations, such that $\kappa R \rightarrow \infty$ the mobility follows as

$$\mu = \frac{\varepsilon\varepsilon_0\zeta}{\eta}$$

Experimental Measurement

For particles large enough to be viewed with a microscope, but small enough that sedimentation is not relevant, an electric field is applied across a chamber and a very dilute sample of particles introduced. The particle motion is then followed visually and the mobility calculated.

For smaller particles, an indirect method is required to measure the particle velocity. The easiest is phase amplitude light scattering. An alternating electric field with a frequency of a few Hertz is applied. The phase shift of scattered light is dependant on the applied electric field and is proportional to the mobility of particles causing the scattering.

Iso-electric points

The source of electrostatic stabilization for latex particles is charged moieties on the surface of particles. At some pH value these groups protonate and the particles lose their surface charge. The pH at which particles have zero charge is referred to as the isoelectric point

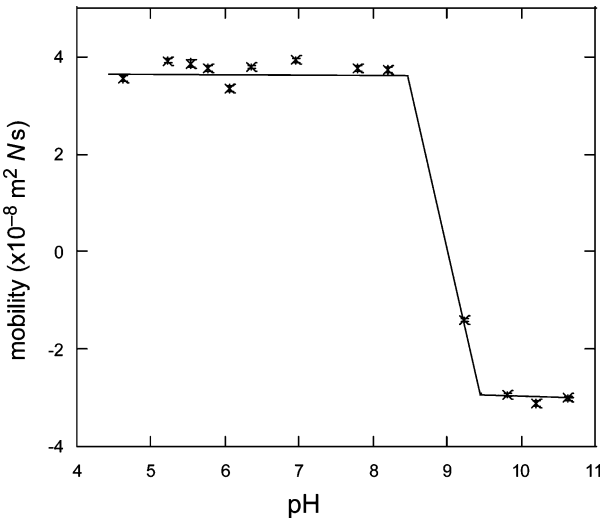


Fig. 7 Mobility versus pH data for polystyrene particles in 1 mM NaCl. The particles were made by surfactant free emulsion polymerization with 2,2-azo-bis(2-methylpropion amidine) dihydrochloride as the initiator.

(IEP) or the point of zero charge (PZC). Fig. 7 shows the mobility of polystyrene particles and an IEP of around pH 9 is seen. At the IEP the particles have very low electrostatic stabilization and readily aggregate.

RHEOLOGY

Latex dispersions display a rich range of rheological properties. The exact behavior of dispersions is dependent on both the volume fraction of particles and the interaction potential.

Shear Thinning

Shear thinning refers to the observation of a decrease in dispersion viscosity as the applied shear rate is increased. It is because of the breakdown of structure in dispersions as the shear rate is increased. For hard sphere latex dispersions, where the interaction potential is zero except at contact, there is found to be a shear thinning behavior for volume fractions above 50%.[25]

Low Shear Viscosity

The viscosity of a dispersion at very low shear rates increases as the volume fraction approaches close packing. For hard spheres a correlation that is found to cover all volume fractions is

$$\frac{\eta_0}{\mu} = \left(1 - \frac{\phi}{0.63}\right)^{-2}$$

where the critical volume fraction of 63% corresponds to random close packing.[25]

Geling

For charged latex particles in water above a critical volume fraction (often about 15%) the dispersion behaves as a viscoelastic solid, being able to support a shear stress. The transition from liquid to solid is found to correspond to the volume fraction of *equivalent hard spheres* being randomly close packed. The size of the equivalent sphere is the particle radius plus the effective size of the electrostatic interaction.[26]

FILM FORMATION

The major use of latex is in the formation of latex films, of which paint is a common example. It is common to split the film formation process up into the three transformations outlined below and the process is sketched in Fig. 8. Although any solvent is possible, water-borne coatings receive considerable attention in the literature, and hence aqueous based dispersions will be considered here. A stable dispersion of particles is concentrated by the evaporation of water. The particles pack and deform under a range of forces to give a structure without voids, although individual particles are still distinguishable. In the final step, polymer chains reptate across inter-particle boundaries, imparting mechanical strength to the film. There have been a number of excellent review articles written recently.[27–29]

Minimum Film Formation Temperature

Film formation requires deformation of polymer particles and reptation of polymer chains, and is strongly temperature dependent. The temperature at which a film will form is commonly measured on a minimum film formation temperature (MFFT) bar. Latex is applied to a metal bar with a pre-assigned temperature gradient. The coating is allowed to dry and a number of transitions are noted. Below a certain temperature the film displays cracks. This is called the crack point MFFT. At a lower temperature there is a transition from cloudy to clear, as the pores between particles become much smaller than the wavelength of light. This is called the cloudy-clear MFFT. A further transition is the temperature at which the film is able

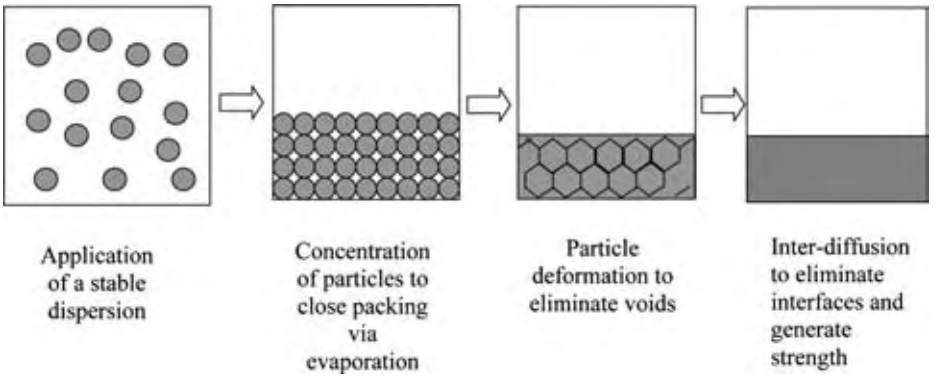


Fig. 8 Four stages of film formation.

to withstand externally applied mechanical shearing. This is called the knifepoint MFFT. The cloudy-clear and knifepoint MFFT are found to move to lower temperatures with time.[30] The MFFT are found to be within a few degrees of the polymer glass transition temperature.[31]

Use of Plasticizers

Particle deformation and polymer diffusion can only occur at temperatures above the glass transition temperature of the polymer. Final coatings, however, are required to be at temperatures considerably below the glass transition temperature. To get around this problem, it is common to add plasticizers to water borne latex dispersions to lower the glass transition temperature of the constituent polymer during the film formation process. Subsequent evaporation of the plasticizer results in a hard final coating. A common plasticizer is 2,2,4-trimethyl-1,3-pentanediol monoisobutyrate, usually referred to as Texanol Ester Alcohol.

Film Formation Mechanism

The final properties of latex coatings are dependent on the mechanism of film formation and how the film forms. For example, the development of mechanical strength is a direct consequence of polymer chain interdiffusion. Prediction of this strength is only possible from an understanding of the transformations occurring on the particle length scale. Here, the three major transformations in latex film formation are briefly outlined.

Transformation 1—Solvent evaporation

An aqueous latex sample initially loses water at a rate that is consistent with the evaporation of pure water. As the particles become more concentrated, two distinct cases have been observed: (1) If the particles consolidate at the top surface and deform significantly, restricting water flow to the top surface, evaporation is found to decrease, and the film becomes vertically inhomogeneous. (2) If the particles are sufficiently rigid, then evaporation is found to continue unhindered.[32]

In addition to the conceptually simple one-dimensional process, evaporating solvent from a latex dispersion is complicated by the presence of drying fronts. A wet film dries first at the thinnest part (normally an edge) and particles pack. Continued evaporation from the packed region causes a transverse flow of solvent in the film, carrying particles to the edge and resulting in a front of packed particles, passing laterally across the film.[33] This means that film formation is inherently two-dimensional and any control over film morphology must consider the lateral dimension.

Transformation 2—Particle deformation

Once particles come into contact, they will begin to deform. If deformation is incomplete then a cloudy film is seen and cracks may appear. The deformation step is, therefore. crucial to film formation, and has received a large amount of attention from academics. A number of driving forces for the deformation have been proposed. All the driving forces depend on the surface tension between the particles and the surrounding medium, or the water–air surface tension. These mechanisms are outlined below:

Wet Sintering. If the particles remain in an aqueous environment after coming into contact, the action of the polymer-water surface tension can cause the particles to consolidate. For this to occur, the deformation must be faster than the evaporation rate, and this was first postulated by Vanderhoff.[34]

Capillary Deformation. As water evaporates, curvature of the air-water interface causes a large negative capillary pressure in the water. Atmospheric pressure

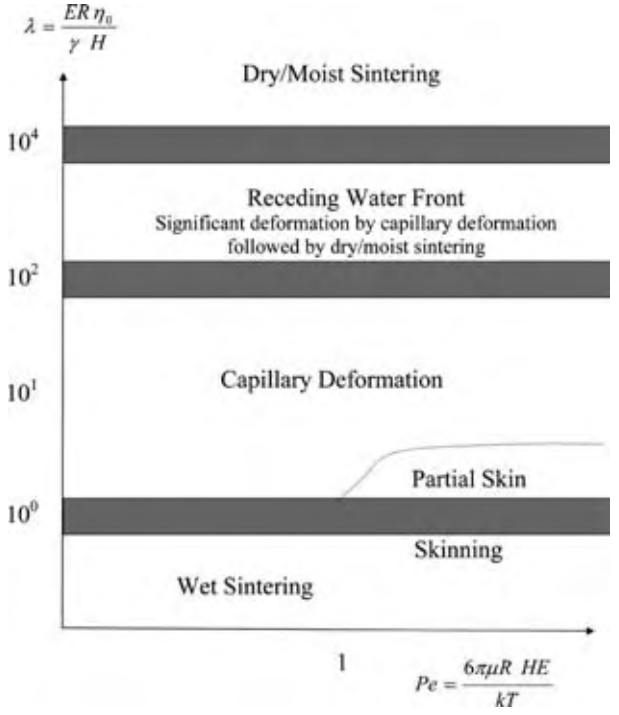


Fig. 9 Regimes of different deformation mechanisms. (From Ref.[32].)

pushes down on the film and can cause the deformation of particles. In this case, deformation is concurrent with evaporation and was first postulated by Brown.[35]

Dry Sintering. Once all the water has evaporated, the action of polymer-air surface tension can cause particles to deform. This is analogous to wet sintering, and was first postulated by Dillon, Matheson, and Bradford.[36] It has been argued that residual water must remain in ''dry'' films to balance residual humidity in the atmosphere, the water being present at the boundary between particles, as pendular rings. The capillary pressure in these rings is large and causes particles to deform. This deformation mechanism has been termed moist sintering.[37]

Receding Waterfront. If significant deformation occurs by capillary deformation, but the voids remain in the film as water recedes, an inhomogeneous film is observed. The deformation is completed by either the dry or moist sintering mechanisms referred to above. This was first identified by Keddie and termed stage II*.[38]

Sheetz Deformation. If the film is inhomogeneous, a continuous polymer layer or skin may form at the top surface. This slows evaporation considerably and allows more time for wet sintering to occur in the consolidated part of the film. This has been termed Sheetz deformation, although the argument about time scales is different to the original postulation by Sheetz.[39]

Predicted Deformation Mechanisms. Recent work has developed maps of the deformation mechanisms expected in films with different properties.[31,32] Two dimensionless groups were found to determine which of the deformation mechanism occurs. The first is the time for particle deformation compared to the time for evaporation, captured in $\lambda = ER\eta_0/\gamma H$, where E is the evaporation rate, η_0 is the polymer viscosity, and γ is the water-air surface tension. The second dimensionless group is the Peclet number, which determines the vertical homogeneity in the film, $Pe = 6\,\pi\,\eta\ R\ H\ E/kT$. The deformation regimes are shown in Fig. 9.

Transformation 3—Polymer reptation

Once the polymer particles have deformed, any free polymer chains will diffuse across particle boundaries. The strength of polymer films has been shown to increase with the depth of polymer interdiffusion, up to size of a single polymer chain.[40]

Design

When designing a film forming latex dispersion, the properties to consider are the final mechanical properties of the film, as well as the ease of film formation: The mechanical properties required, as well as the environment of operation will dictate the polymers suitable for the coating and may well dictate the glass transition temperature of the polymer. The crack points alluded to earlier correspond to the transition from capillary deformation to the receding water front regime. Therefore, a value of λ less than 100 will ensure a well-formed film.

CONCLUSIONS

This article has reviewed latex processing. The polymers used, synthesis of particles, major uses, and reasons for loss of dispersion stability have been outlined. The mechanism of latex film formation has been described, and the different properties resulting from different film forming conditions in latex explored.

ACKNOWLEDGMENTS

The author is very grateful to Dr. Peter Dowding for his invaluable advice and Ruth Dunleavey for her

countless helpful discussions and critical reading of the manuscript.

REFERENCES

1. Fitch, R.M. *Polymer Colloids: A Comprehensive Introduction*; Academic Press, 1997.
2. Blackley, D.C. *Polymer Latices Science and Technology: Fundamental Principles*; Chapman Hall, 1997; Vol. 1.
3. Blackley, D.C. *Polymer Latices Science and Technology: Types of Latices*; Chapman Hall, 1997; Vol. 2.
4. Lesko, P.M.; Sperry, P.R. Acrylic and styrene-acrylic polymers. In *Emulsion Polymerization and Emulsion Polymers*; Lovell, P., El-Aasser, M.S., Eds.; John Wiley and Sons Ltd., 1997; 620–651.
5. Blackley, D.C. *Polymer Latices Science and Technology: Applications of Latices*; Chapman Hall, 1997; Vol. 3.
6. Arshady, R. Suspension, emulsion and dispersion polymerization: a methodological survey. Colloid Polym. Sci. **1992**, *270*, 717–732.
7. Goodwin, J.W.; Hearn, J.; Ho, C.C.; Ottewill, R.H. The preparation and characterization of polymer latices formed in the absence of surface active agents. Br. Polym. J. **1973**, *5*, 347–362.
8. Adams, M.E.; Casey, B.S.; Mills, M.F.; Russell, G.T.; Napper, D.H.; Gilbert, R.G. High conversion emulsion, dispersion and suspension polymerization. Macromol. Chem., Macromol. Symp. **1990**, *35/36*, 1–12.
9. Horak, D. Uniform polymer beads of micrometer size. Acta Polym. **1996**, *47*, 20–28.
10. Pelton, R.H.; Chibante, P. Preparation of aqueous latices with N-isopropylacrylamide. Colloids Surf. **1986**, *20*, 247–256.
11. Saunders, B.R.; Vincent, B. Microgel particles as model colloids: theory, properties and applications. Adv. Colloid Interface Sci. **1999**, *80*, 1–25.
12. Loxley, A.; Vincent, B. Equilibrium and kinetic aspects of the pH-dependent swelling of poly (2-vinylpyridine-co-styrene) microgels. Colloid Polym. Sci. **1997**, *275* (12), 1108–1114.
13. Murray, M.J.; Snowden, M.J. The preparation, characterization and application of colloidal microgels. Adv. Colloid Interface Sci. **1995**, *54*, 73–91.
14. Russel, W.B.; Saville, D.A.; Schowalter, W.R. *Colloidal Dispersions*; Cambridge University Press, 1989.
15. Rawson, S.; Ryan, K.; Vincent, B. Depletion flocculation in sterically stabilized aqueous systems using polyelectrolytes. Colloids and Surfaces **1988**, *34*, 89–93.
16. Asakura, S.; Oosawa, F. On interaction between two bodies immersed in a solution of macromolecules. J. Chem. Phys. **1954**, *22* (7), 1255–1256.
17. Asakura, S.; Oosawa, F. Interaction between particles suspended in solutions of macromolecules. J. Polym. Sci. **1958**, *33*, 183–192.
18. Vincent, B.; Luckham, P.F.; Waite, F.A. The effect of free polymer on the stability of sterically stabilized dispersions. J. Colloid Interface Sci. **1980**, *73* (2), 508–521.
19. Vincent, B.; Edwards, J.; Emmett, S.; Croot, R. Phase separation in dispersions of weakly interacting particles in solutions of non-adsorbing polymer. Colloids Surf. **1988**, *31*, 267–298.
20. Anderson, V.J.; de Hoog, E.H.A.; Lekkerkerker, H.N.W. Mechanisms of phase separation and aggregation in colloid-polymer mixtures. Phys. Rev. E. **2001**, *65* (1), 11403.
21. Hogg, R.; Healy, T.W.; Fuertenau, D.W. Mutual coagulation of colloidal dispersions. Trans. Faraday Soc. **1965**, *62*, 1638–1651.
22. Carpineti, M.; Ferri, F.; Giglio, M.; Paganini, E.; Perini, U. Salt induced fast aggregation of polystyrene latex. Phys. Rev. A. **1990**, *42* (12), 7347–7354.
23. Lin, M.Y.; Lindsay, H.M.; Weitz, D.A.; Klein, R.; Ball, R.C.; Meakin, P. Universal diffusion-limited colloid aggregation. J. Phys.: Condens. Matter **1990**, *2* (13), 3093–3113.
24. Lin, M.Y.; Lindsay, H.M.; Weitz, D.A.; Klein, R.; Ball, R.C.; Meakin, P. Universal reaction-limited colloid aggregation. Phys. Rev. A. **1990**, *41* (4), 2005–2020.
25. De Kruif, C.G.; ven Iersel, E.M.F.; Vrij, A.; Russel, W.B. Hard sphere colloidal dispersions: viscosity as a function of shear rate and volume fraction. J. Chem. Phys. **1986**, *83*, 4717–4725.
26. Buscall, R.; Goodwin, J.W.; Hawkins, M.W.; Ottewill, R.H. Viscoelastic properties of concentrated latices: II Theoretical analysis. J. Chem. Soc., Faraday Trans. **1982**, *78*, 2889–2899.
27. Keddie, J. Film formation of latex. Mater. Sci. Eng. **1997**, *R21* (3), 101–169.
28. Winnik, M.A. The formation and properties of latex films. In *Emulsion polymerization and emulsion polymers*; Lovell, P., El-Aasser, M.S., Eds.; John Wiley and Sons Ltd, 1997; 468–515.
29. Steward, P.A.; Hearn, J.; Wilkinson, M.C. An overview of polymer latex film formation and properties. Adv. Colloid Interface Sci. **2000**, *86* (3), 195–267.

30. Sperry, P.R.; Snyder, B.S.; O'Dowd, M.L.; Lesko, P.M. Role of water in particle deformation and compaction in latex film formation. Langmuir **1994**, *10*, 2619–2628.
31. Routh, A.F.; Russel, W.B. A process model for latex film formation: limiting regimes for individual driving forces. Langmuir **1999**, *15* (22), 7762.
32. Routh, A.F.; Russel, W.B. Deformation mechanisms during latex film formation: experimental evidence. Ind. Eng. Chem. Res. **2001**, *40* (20), 4302–4308.
33. Routh, A.F.; Russel, W.B. Horizontal drying fronts during solvent evaporation from latex films. AIChE J. **1998**, *44* (9), 2088–2098.
34. Vanderhoff, J.W.; Tarkowski, H.L.; Jenkins, M.C.; Bradford, E.B. Theoretical consideration of the interfacial forces involved in the coalescence of latex particles. J. Macromol. Chem. **1966**, *1* (2), 361–397.
35. Brown, G.L. Formation of films from polymer dispersions. J. Polym. Sci. **1956**, *22*, 423–434.
36. Dillon, R.E.; Matheson, L.A.; Bradford, E.B. Sintering of synthetic latex particles. J. Colloid Sci. **1951**, *6* (2), 108–117.
37. Lin, F.; Meier, D.J. A study of latex film formation by atomic force microscopy: 1. A comparison of wet and dry conditions. Langmuir **1995**, *11*, 2726–2733.
38. Keddie, J.L.; Meredith, P.; Jones, R.A.L.; Donald, A.M. Kinetics of film formation in acrylic latices studied with multiple angle-of-incidence ellipsometry and environmental SEM. Macromolecules **1995**, *28*, 2673–2682.
39. Sheetz, D.P. Formation of films by drying of latex. J. Appl. Polym. Sci. **1965**, *9*, 3759–3773.
40. Prager, S.; Tirrell, M. The healing process at polymer-polymer interfaces. J. Chem. Phys. **1981**, *75* (10), 5194–5198.

Liquid–Liquid Mixing in Agitated Reactors

Richard V. Calabrese
Department of Chemical and Biomolecular Engineering, University of Maryland, College Park, Maryland, U.S.A.

Douglas E. Leng
Leng Associates, Midland, Michigan, U.S.A.

Piero M. Armenante
Otto H. York Department of Chemical Engineering, New Jersey Institute of Technology, Newark, New Jersey, U.S.A.

INTRODUCTION

This entry briefly describes the use of agitated vessels and reactors to create immiscible liquid–liquid dispersions. A more comprehensive treatment is given by Leng and Calabrese.[1] The term ''immiscible liquid–liquid system'' refers to two or more mutually insoluble liquids present as separate phases. These phases are referred to as the dispersed or drop phase, and the continuous or matrix phase, and their property variables are given subscripts of d and c, respectively. The dispersed phase is usually smaller in volume than the continuous phase, but under certain highly formulated conditions, it can represent up to 99% of the total volume of the system. Immiscible liquid–liquid systems are found extensively throughout the chemical, petroleum, and pharmaceutical industries. The rate of chemical reactions is often mass transfer controlled and is affected by the interfacial area per unit volume of dispersed phase. Examples include nitration, sulfonation, alkylation, hydrogenation, and halogenation. Failure to adequately suspend and disperse drops can lead to undesirable and sometimes catastrophic results.

Mechanical energy, typically supplied through a mechanical device such as an impeller, pump, or rotor–stator device, is required to disperse one phase into the other. Agitation plays a key role in liquid–liquid systems. It controls the break up of drops, referred to as *dispersion*; the combining of drops, known as *coalescence*; as well as the *suspension of drops* within the system. The magnitude and direction of convective flows produced by an agitator affect distribution and uniformity throughout the vessel, as well as the kinetics of dispersion. Agitation intensity is also important. Intense turbulence found near the impeller leads to drop dispersion, not coalescence. Lower turbulence or laminar/transitional conditions found beyond the impeller region, promote coalescence by enabling drops to remain in contact long enough for them to coalesce.

Coalescence, dispersion, and suspension phenomena are complex, inter-related, and scale dependent. For example, dispersion tends to dominate in small vessels and coalescence in large equipment. Nevertheless, industrial processes can be semi-quantitatively analyzed if they are either noncoalescing or slowly coalescing. This simplifies design and scale-up. Coalescence can usually be neglected, for practical purposes, in applications where the volume fraction of dispersed phase, $\phi \leq 0.1$. This is particularly true if surfactants and/or interfacial contaminants are present. However, scale-up is complicated by the time it takes to create dispersions. Small vessels reach a terminal state of dispersion much faster than large ones. If scale-up is done to form equal interfacial area, the large vessel will take much longer to achieve the task. Stirred vessels, rotor–stator mixers, static mixers, decanters, settlers, centrifuges, homogenizers, extraction columns, and electrostatic coalescers are examples of industrial process equipment used to contact liquid–liquid systems. Although this entry emphasizes stirred vessels, the fundamentals of phase behavior are applicable to a broad range of other equipment types. In the case of stirred vessels, for any given system, the mean drop size and drop size distribution depend on the selection, placement, and operational speed of the agitator. Excessive speed leads to hard-to-separate emulsions. Inadequate speed can cause phase separation. Both coalescence and dispersion are fluid motion dependent rate processes. Drop sizes depend on flow, shear, turbulence, and dispersion time, as well as physical and interfacial system properties.[1]

SELECTION AND CONFIGURATION OF LIQUID–LIQUID EQUIPMENT

Any impeller capable of pumping fluid and providing shear can produce liquid–liquid dispersions. Commonly

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120041244

used impellers include disk turbines, pitched-blade turbines, propellers, hydrofoils, paddles, retreat-curve impellers, and other proprietary designs. High-shear, high-speed devices such as rotor–stator mixers are also commonly used to produce emulsions containing finely dispersed, micron-size droplets. These devices are discussed in detail by Atiemo-Obeng and Calabrese.[2]

If the application requires high interfacial area (i.e., small drop diameters), a high-shear impeller, such as the Rushton or radial disk turbine (RDT), is a good choice (Fig. 1). Acceptable substitutes include the Scaba and Chemineer's BT-6 and CD-6 impellers, commonly used for gas–liquid dispersion. If moderate, yet gentle shear is required, such as for emulsion polymerization, the retreat-curve impeller is commonly chosen. When larger drops of a narrow size distribution are required, the loop impeller is a reasonable choice. Broad-blade paddles are also used.

Recommended impeller diameter-to-tank diameter (D/T) ratios, for liquid–liquid operations vary from 0.25 to 0.40 for radial disk turbines, from 0.4 to 0.6 for hydrofoils and propellers, and from 0.5 to 0.8 for retreat-curve, glassed-steel impellers. Vertical placement of the impeller depends on the vessel shape and the application. For dispersion by continuous addition of a dense phase fluid into a less dense fluid, the impeller should have a relatively small impeller clearance off the reactor bottom, C, with respect to the final height of the dispersion, H, i.e., the impeller should be placed low in the vessel ($C \approx H/4$ to $H/5$). For dispersion of light liquids, it is good practice to place a single impeller between $0.20 \leq C/H \leq 0.50$.

Multiple impellers are recommended if $H/T \gg 1.2$ or if $\Delta\rho > 0.15\,\mathrm{kg/m^3}$. Assuming a less dense dispersed phase, the second or top impeller often is a hydrofoil impeller placed midway between the radial disk turbine and the surface of the liquid. This impeller produces high flow at low power, provides effective circulation, and complements the flow pattern produced by the radial disk turbine. The diameter of the second impeller is usually greater than the radial disk turbine, typically $D/T \geq 0.45$. A good practice is to distribute the total power to ≈20% for the hydrofoil and ≈80% for the radial disk turbine. Table 1 lists equipment options for different drop sizing objectives.

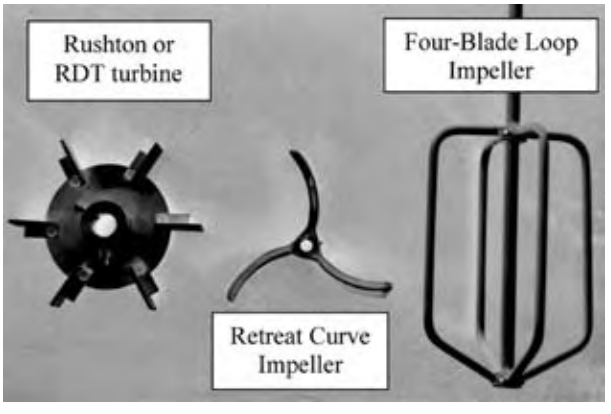


Fig. 1 Some impellers used for liquid–liquid dispersion. (From Ref.[1].)

The geometry for liquid–liquid processing is typically a cylindrical dish bottom vessel. It should be provided with four equally spaced baffles having a width between $T/12$ and $T/10$ and located at least $T/72$ off the wall. Other vessel and baffle arrangements are possible, as discussed by Hemrajani and Tatterson.[3] In all cases, it is essential to avoid stagnant regions in liquid–liquid operations, regardless of the process. This means that the use of flat and cone bottom tanks and tall slender vessels is not recommended. Placing baffles away from the wall, to permit flow between the wall and the baffle, prevents dispersed phase buildup in stagnant areas. Internal heating coils and ladders should also be avoided, if possible. Optimum flow patterns normally develop when the overall vessel shape is $1 < H/T < 1.2$.[1]

FORMING LIQUID–LIQUID DISPERSIONS

Both initial and final conditions affect dispersion formation (see Figs. 2 and 3, where oil is the lighter or the upper phase). If the lower phase is to be dispersed in the upper phase, the radial disk turbine should be placed in the upper phase and an upward-pumping axial flow turbine is placed in the lower phase, as shown in Fig. 2. When the upper oil layer is to be dispersed in the lower water layer, the arrangement shown in Fig. 3 is recommended. Here the axial flow turbine pumps downward. Both figures show the use of a radial disk turbine for dispersion and a propeller to improve circulation. Single impellers can also be used. Often, oil-in-water (O/W) and water-in-oil (W/O) regions initially coexist. The amount of each phase and the relative rates of coalescence (O/W vs. W/O) during transient conditions determine whether the final system is O/W or W/O. Fig. 4 shows the ideal location for a single turbine.

MECHANISMS OF DROP DISPERSION, DROP COALESCENCE, AND PHASE INVERSION

Drop deformation occurs when fluid dynamical forces, often referred to as "shear forces", in the surrounding fluid act on its surface. Surface and internal viscous forces resist it. Drop dispersion (breakage) occurs when the "shear forces" exceed the combined resistance force.

Table 1 Common types of equipment used for liquid–liquid dispersion

Description	Impeller types	Batch/continuous	Desired result	Comments
Stirred tanks; baffles	Flat, pitch, and disk type	Either	$30 \leq d_{32} \leq 300\,\mu m$	General; mass transfer operations
Stirred tanks; baffles	Retreat curve	Either	$30 \leq d_{32} \leq 300\,\mu m$	General; emulsion polymerization[a]
Stirred tanks; no baffles	Paddle, loop, special types	Batch	$100 \leq d_{32} \leq 1000\,\mu m$	Suspension polymerization;[a] suspending agent required
Static/in-line mixers	None	Continuous	$10 \leq d_{32} \leq 200\,\mu m$	Dispersant or protective colloid needed
Rotor–stator mixers	Slotted ring or impeller, along with slotted stator	Either, often continuous	$1 \leq d_{32} \leq 50\,\mu m$	Sparse data for scale-up; need extensive testing
Impingement Mixers	None	Continuous	$1 \leq d_{32} \leq 50\,\mu m$	Sparse data; work with vendors
Valve homogenizers; ultrasonic mixers	None	Usually continuous	$0.1 \leq d_{32} \leq 10\,\mu m$	Sparse data; work with vendors; feed is predispersed

[a]Drop size refers to monomer drops. Latex products are much smaller particles in the range from 0.1 μm to 0.5 μm.
(From Ref.[1].)

Coalescence is the combining of two or more drops, or a drop with a coalesced layer. The two-step process involves collision followed by film drainage. Collision frequency depends on both the agitation rate and the volume fraction of the dispersed phase. The drainage step depends on the magnitude and the duration of the force acting on the drop(s) to squeeze out the continuous phase film separating the drops to a critical thickness, believed to be in the range of ≈50 Å. The coalescence rate is the product of the collision rate and the coalescence efficiency (related to drainage rate). The mobility of the liquid–liquid interface also affects the film drainage rate. Clean, mobile interfaces promote efficient film drainage, and lead to higher coalescence probability. A simple static test for coalescibility is to agitate or shake a sample for 5 min and then watch it settle and coalesce. The results can be interpreted using Table 2. A noncoalescing or slowly coalescing system has a settling time that is greater than 5 min. A rapidly coalescing system has a settling time of less than 1 min.

Coalescence, dispersion, and settling are all affected by dispersed phase concentration or volume fraction, ϕ. Liquid–liquid systems can be categorized with respect to ϕ as follows:

- *Dilute Systems:* $\phi < 0.01$. Dispersion is affected only by hydrodynamics, and each drop is a single entity experiencing continuous phase fluid forces. Coalescence is neglected because few collisions occur.
- *Moderately Concentrated Systems:* $\phi < 0.2$. The behavior of systems in this concentration range depends on coalescence behavior. Ideal dilute dispersion theories may still apply, particularly if the system is noncoalescing. Even in the presence of coalescence, it is possible to roughly predict the droplet size for moderately concentrated systems.

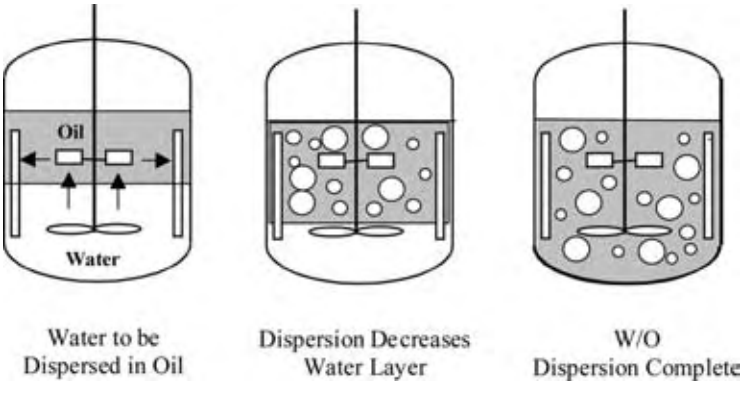


Fig. 2 Dual impeller arrangement for water-in-oil dispersion. Propeller is upward pumping. (From Ref.[1].)

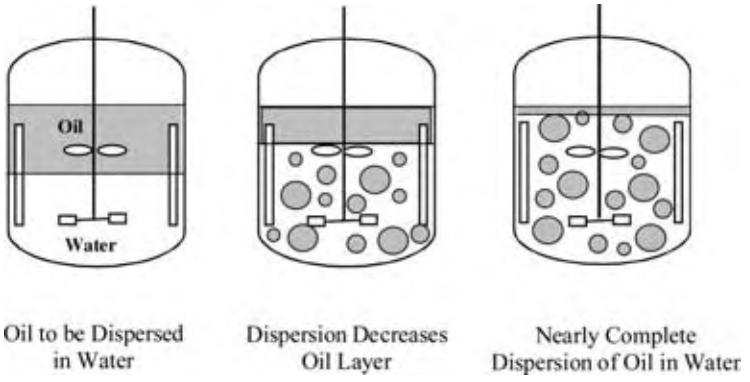


Fig. 3 Dual impeller arrangement for oil-in-water dispersion. Propeller is downward pumping. (From Ref.[1].)

- *More Concentrated Systems:* $\phi > 0.2$. This range is common in industry. Fast coalescence is probable for clean systems. Sprow[4] found that with coalescing systems, drop sizes were position dependent within the vessel. This behavior is very complex and extremely difficult to scale-up, as coalescence and dispersion dominate in different regions of the vessel.

Phase inversion is a commonly observed phenomenon in which the continuous phase abruptly becomes the dispersed phase and vice versa (see Pacek et al.[5,6] and Pacek, Nienow, and Moore [7]). Systematic studies on the effect of surfactant concentration and mixing on phase inversion and emulsion drop size have been carried out by Brooks and Richmond.[8–11] Fig. 5 shows schematically the steps occurring during phase inversion. Although conflicting information exists on the subject, the following conclusions can be made:

- Coalescence, not dispersion, dominates as the controlling mechanism in phase inversion. Factors affecting film drainage rates, such as agitation rate, interfacial tension, interface mobility, film viscosity and contact time, all apply;
- Inversion behavior is property dependent and therefore system specific.
- Surface-active agents play an important role, affecting film drainage rates.
- Every system has an operating region in which the oil phase is continuous, a region in which the aqueous phase is continuous, and an ambivalent region where either phase can be continuous, depending on initial conditions, flow circulation, and agitation intensity. This is shown in Fig. 6, which contains the data of Kinugasa et al.[12]
- The probability for phase inversion increases as drops get closer together.

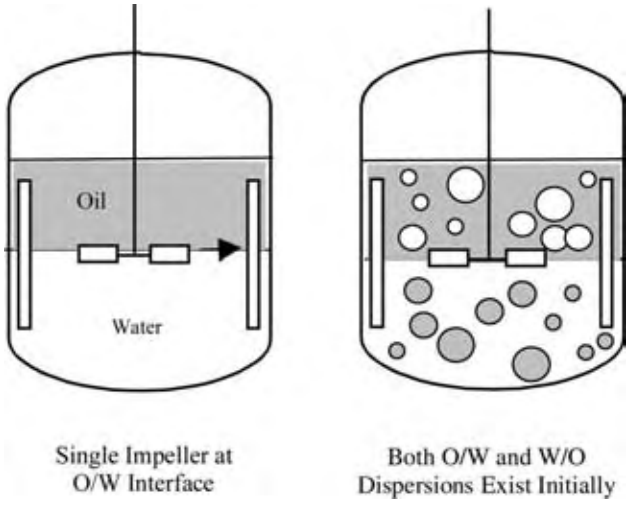


Fig. 4 Single radial disk turbine placed at oil-water interface. (From Ref.[1].)

PRACTICAL ASPECTS OF DISPERSION FORMATION

As previously discussed, placing a radial disk or Rushton turbine in the aqueous or lower phase, close to the interface, can be effective when making oil-in-water dispersions. A central interfacial vortex forms with the commencement of impeller motion. This directs a stream of the lighter oil phase to the impeller where it disperses. The volume of the oil layer decreases with continued dispersion until it is exhausted. Placing the turbine in the oil or upper phase, close to the interface, can result in water-in-oil dispersions, because a water-containing vortex forms, allowing water to be dispersed into the lighter oil phase.

Dispersions may also be formed by the continuous addition of one phase into another under agitated conditions. This method offers a safe procedure for handling exothermic reactions such as nitration and emulsion polymerization. The amount of phase addition will determine if phase inversion occurs.

Listed below are some general recommendations for immiscible liquid–liquid systems:

- Use multiple turbines if the system is rapidly coalescing, to provide additional dispersion capability.

Table 2 Characterization of the coalescibility of immiscible liquid–liquid systems

Time to separate	Characterization	Process implication
<10 sec	Very fast coalescence	Expect severe scale-up problems for agitated vessels; provide more dispersion opportunities. For example, use multiple impellers, provide for strong flow at the top and bottom of the vessel. Consider use of long static mixers
<1 min	Fast coalescence	Scale-up problems can be managed by careful selection of mixing equipment. Use multiple impellers, eliminate unnecessary internals, and provide for complete circulation
2–3 min	Moderate coalescence	Problems are less severe, design for coalescence. Use large impellers for dispersion and flow. Maintain ample flow at the top/bottom surfaces. Often can treat this case as noncoalescing
>5 min	Slow coalescence	Application can be treated as dispersion only

(From Ref.[1].)

Axial flow turbines can also be used to achieve better uniformity in circulation.

- Avoid excessive dispersion in noncoalescing systems. Creation of tiny, hard-to-coalesce drops can become a problem if phase separation is required later. Test the system using bench scale equipment to see if and at what speed undesirably small drops form.
- Use at least one axial flow hydrofoil-type impeller of high D/T (i.e., $0.4 \leq D/T \leq 0.6$) in addition to the Rushton or RDT turbine for systems having large phase density differences.
- Interfacial tension controls the ease of drop breakage. Low interfacial tension ($\sigma \leq 10$ dyne/cm or 0.01 N/m) systems require lower power for dispersion than high interfacial tension ($\sigma \geq 30$ dyne/cm or 0.03 N/m) systems.
- Use dish bottomed vessels with proportions $H/T < 1.2$.
- Baffling is always required for liquid–liquid dispersion, with the exception of suspension polymerization and certain highly shear sensitive emulsion polymerizations.

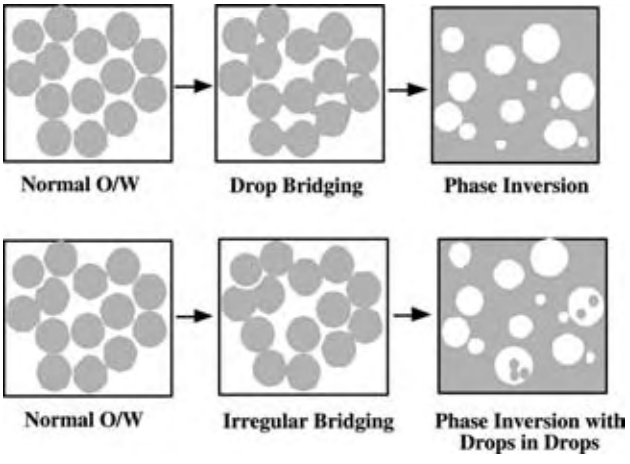


Fig. 5 Sequences in phase inversion. (From Ref.[1].)

Table 3 gives a summary of practical guidelines for scale-up of coalescing and noncoalescing systems.

CALCULATION OF MEAN DROP SIZE IN LIQUID–LIQUID DISPERSIONS

The ability to predict drop size is critical to determining both the interfacial area for mass transfer and the state of dispersion of the system. In dilute systems and in moderately concentrated systems where coalescence can be neglected, the following equation describes the maximum equilibrium (i.e., after a long time) drop diameter of an inviscid or low viscosity dispersed phase:

$$d_{\max} = C_1 \left(\frac{\sigma}{\rho_c}\right)^{3/5} \varepsilon_{\max}^{-2/5} \tag{1}$$

where the constant C_1 must be determined from experimental data. An extensive discussion of drop dispersion theory is given by Leng and Calabrese.[1] Davies[13] showed that values of $d_{\max}$ for a wide variety of dispersion devices could be correlated with maximum local power per mass, if a rough estimate of $\varepsilon_{\max}/\varepsilon_{\text{avg}}$ could be obtained. The results of the analysis for dilute inviscid dispersed phases, corrected for interfacial tension, are shown in Fig. 7. The slope of the line bounding the data is –2/5, as predicted by Eq. (1).

A more practically important drop size than $d_{\max}$ is the Sauter mean diameter d_{32}, defined as:

$$d_{32} = \frac{\sum_{i=1}^{i=m} n_i d_i^3}{\sum_{i=1}^{i=m} n_i d_i^2} \tag{2}$$

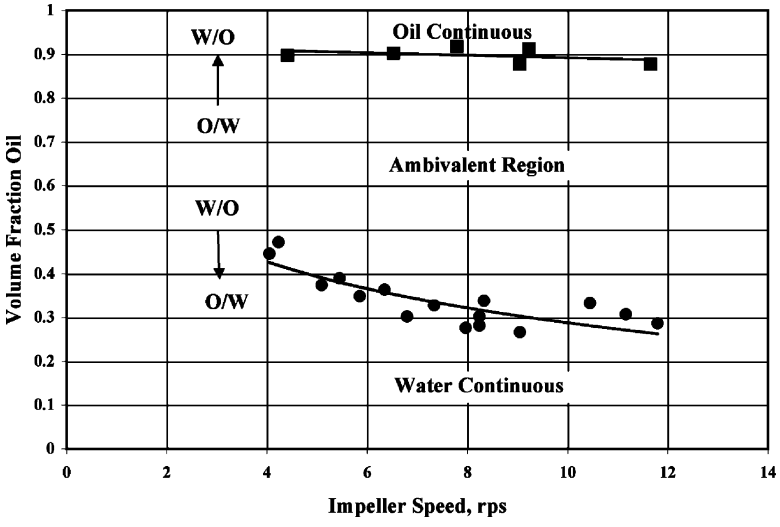


Fig. 6 Phase inversion boundaries for the kerosene-water system showing oil and water continuous regions, and an ambivalent region. (From Ref.[1,12].)

For geometrically similar turbulent systems, $\varepsilon_{max} \propto \varepsilon_{avg}$, which for constant power number gives $\varepsilon_{max} \propto N^3D^2$ (see Zhou and Kresta[14,15]). For a dilute system, the droplets will have sizes equal to d_{max} or smaller. There is considerable experimental evidence that $d_{max} \propto d_{32}$ (see Leng and Calabrese,[1] Wang and Calabrese[16]). Therefore, for geometrically similar systems, Eq. (1) is equivalent to:

$$\frac{d_{32}}{D} = C_2\, We^{-3/5} \tag{3}$$

where the Weber number $We = \rho_c N^2 D^3/\sigma$ is the ratio of inertial (disruptive) to surface (resistance) forces. For Rushton turbines, Chen and Middleman[17] found $C_2 = 0.053$ (Fig. 8) for a broad range of liquid–liquid pairs. Eqs. (1) and (3) show that, at equilibrium, dispersed phase systems created by turbulent flow scale-up by maintaining constant ε_{max}; or for practical industrial purposes, by constant ε_{avg}, which is equivalent to constant P/V_L. Large Weber numbers result in small drops and vice versa. These expressions are valid for dilute, noncoalescing systems of low μ_d, at equilibrium. Many stabilized or noncoalescing industrial systems with $\phi > 0.05$ can also be scaled by the constant P/V_L criterion.

If the drops are viscous, the internal viscous resistance to deformation must be considered. Furthermore,

Table 3 Guidelines for scale-up of liquid–liquid stirred vessels

Feature	Non/slowly coalescing system	Rapidly coalescing system
Scale-up criterion	P/V_L, constant	Circulation time, constant
Scale-up limitation, V_{Large}/V_{Small}	100/1	10/1 to 20/1
Baffles	Yes, but not for suspension polymerization	Yes
Impellers	Radial disk turbine and optional axial flow/ hydrofoil impeller	Multiple radial disk turbines and axial flow/hydrofoil impeller for better circulation
D/T	0.3–0.5	$\geq$0.5
Time to reach terminal drop size	Long times for large vessels	Short times under 30 min for most coalescing systems (all vessel sizes)
Geometric similarity	Maintain close similarity	Use more and larger turbines in larger vessel. Do not try to maintain geometric similarity
Speed/drives	Variable or fixed speed	Variable speed capability is essential. Consider over design to meet unpredicted performance
Risk	Low to moderate risk	High risk

(From Ref.[1].)

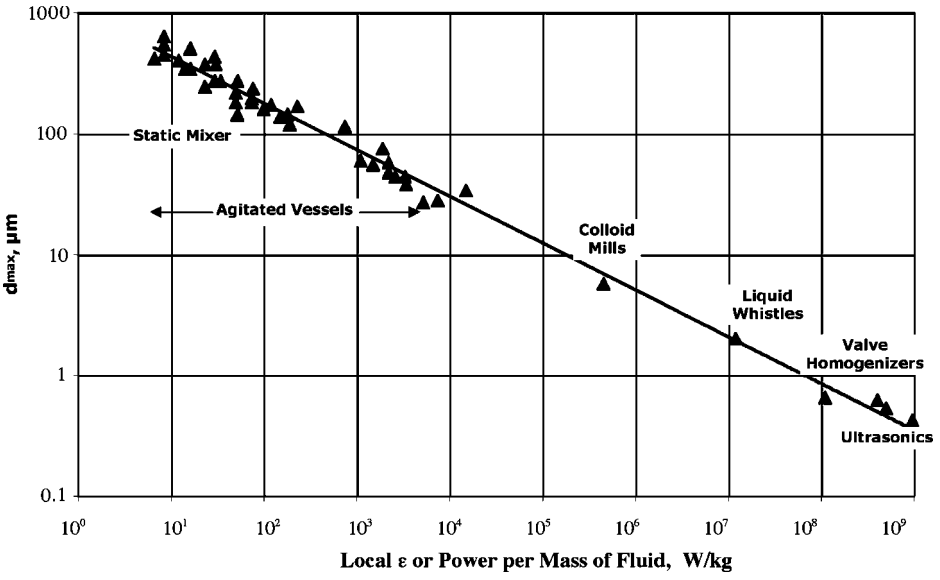


Fig. 7 Dependence of drop size on local power drawn by various dispersion devices. (From Ref.[1,13].)

the effect of ϕ on d_{32} should also be taken into account. In more concentrated systems ($0.01 \leq \phi \leq 0.3$), which can be described as noncoalescing, the relationships given above can be modified to account for both the viscosity effect (in $\mathrm{Vi} = (\rho_c/\rho_d)^{1/2}\mu_d ND/\sigma$) and ϕ to give the following equation for Rushton turbines, obtained by Calabrese, Wang, and Bryner:[18]

$$\frac{d_{32}}{D} = 0.054(1 + 3\phi)\mathrm{We}^{-3/5} \times \left[1 + 4.42(1 - 2.5\phi)\mathrm{Vi}\left(\frac{d_{32}}{D}\right)^{1/3}\right]^{3/5} \quad (4)$$

This equation should be used with some caution. It has only been experimentally validated for dilute dispersions ($\phi < 0.1$) and $D/T = 0.5$. Other equations for d_{32} are discussed by Calabrese and coworkers.[16,18,19]

CALCULATION OF THE MINIMUM IMPELLER SPEED FOR COMPLETE DISPERSION, N_{MIN}, IN LIQUID–LIQUID SYSTEMS

Settling and coalescence are common when the dispersed and continuous phases are of different density, and when agitation provides only minimal circulation throughout the vessel. It is, therefore, important to determine the minimum impeller speed, N_{min} to completely incorporate the dispersed phase, as droplets, into the continuous phase (i.e., to remove the initial stratification of the immiscible liquids in a vessel). Most reported work is semiempirical and follows the approach of the ''just suspended'' state of solids in liquids described in an earlier entry of this book, as well as by Atiemo-Obeng, Penney, and Armenante.[20]

There are analogies between N_{js}, the minimum agitation intensity to suspend solids, and the corresponding variable, N_{min}, for drop suspension. Both depend on density difference, continuous phase viscosity, and impeller diameter. However, N_{js} depends directly on particle size, while N_{min} depends instead on interfacial tension and the other physical properties that determine drop size.

Skelland and Seksaria[21] determined the minimum speed to form a liquid–liquid dispersion from two settled (separated) phases of different density, and included the sensitivity to impeller location. The vessels used were fully baffled. They determined N_{min} for systems of equal volumes of light and heavy phase. Studies included use of single impellers placed midway in the dense phase ($C = H/4$), at the O/W interface ($C = H/2$) and midway in the lighter phase ($C = 3H/4$). They also examined the use of dual impellers located midway in both phases. Several impeller types were tested. Their results are correlated by the following dimensionless equation:

$$\frac{N_{min}D^{0.5}}{g^{0.5}} = C_3\left(\frac{T}{D}\right)^{C_4}\left(\frac{\mu_c}{\mu_d}\right)^{\frac{1}{9}}\left(\frac{\Delta\rho}{\rho_c}\right)^{0.25}\left(\frac{\sigma}{D^2\rho_c g}\right)^{0.3} \quad (5)$$

The magnitude of the constants C_3 and C_4, given in Table 4, is a measure of the ease of suspension formation. Low C_3values indicate dispersions are formed at low speeds. Large C_3 values (single impellers) suggest that higher speeds are required for minimum suspension. Turbines at the O/W interface require lower speed than in other locations. Radial flat-blade turbines placed in the light phase appear to be inefficient.

Armenante and Huang[22] and Armenante, Huang, and Li.[23] found practically no advantage in using

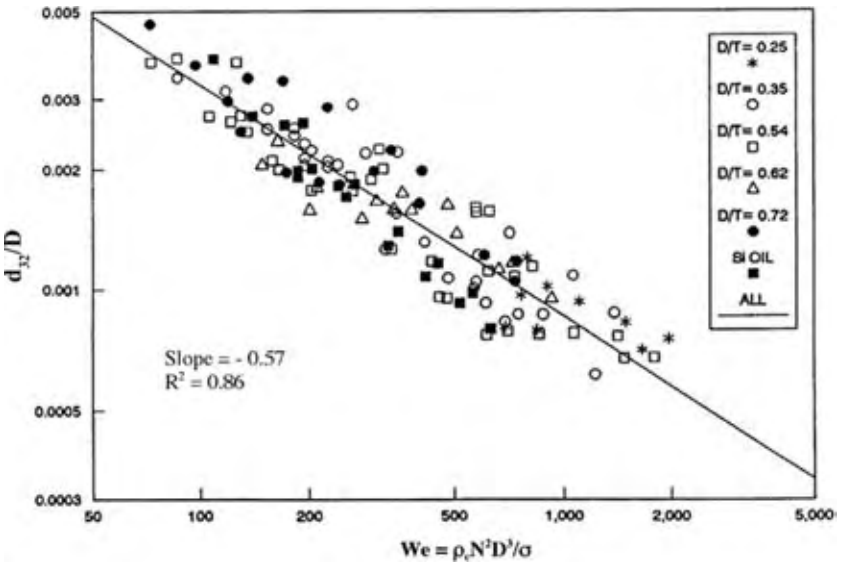


Fig. 8 Experimental data for 14 different liquid–liquid pairs. (From Ref.[1,17].)

multiple impellers for determining N_{min}. This is similar to the result for solid–liquid suspension. However, multiple impellers were useful in improving dispersed phase uniformity.

POWER DISSIPATION IN LIQUID–LIQUID STIRRED REACTORS

As discussed previously, the ultimate drop size is determined by ε_{max}, not ε_{avg}. However, most correlations for drop size use ε_{avg}, because data for ε_{max} are not readily available. The power dissipated by an impeller in liquid–liquid dispersions is:

$$P = N_p \bar{\rho} D^5 N^3 \tag{6}$$

where the volume average density of the liquid–liquid mixture is given by:

$$\bar{\rho} = \phi \rho_d + (1 - \phi)\rho_c \tag{7}$$

Other relevant equations as well as the diagrams to calculate the power number, N_p, are provided in an earlier entry, and by many investigators including Atiemo-Obeng, Penney, and Armenante;[20] Tatterson;[24] Bates, Fondy, and Corpstein;[25] Bates, Fondy, and Fenic;[26] and Armenante and coworkers.[23,27,28] The corresponding value of ε_{avg} for liquid–liquid dispersion is:

$$\varepsilon_{avg} = \frac{P}{\bar{\rho} V_L} \tag{8}$$

An estimate of ε_{max} in agitated vessels can be obtained by replacing V_L in Eq. (8) by the volume swept out by the impeller, as was used by Davies[13] in Fig. 7.

Table 4 Constants for use in Eq. 5

Impeller type	Clearance	C_3	C_4
Propeller	$H/4$	15.3	0.28
Propeller	$3H/4$	9.9	0.55
Propeller	$H/2$	15.3	0.39
Propeller	$H/4 + 3H/4$	5.2	0.92
Pitched-blade turbine	$H/4$	6.8	1.05
Pitched-blade turbine	$3H/4$	6.2	0.82
Pitched-blade turbine	$H/2$	3.0	1.59
Pitched-blade turbine	$H/4 + 3H/4$	3.4	0.87
Flat-blade turbine	$H/4$	3.2	1.62
Flat-blade turbine	$3H/4$	*	*
Flat-blade turbine	$H/2$	4.0	0.88
Flat-blade turbine	$H/4 + 3H/4$	*	*
Curved-blade turbine	$H/4$	3.6	1.46
Curved-blade turbine	$3H/4$	*	*
Curved-blade turbine	$H/2$	4.7	0.80
Curved-blade turbine	$H/4 + 3H/4$	4.3	0.54

*Insufficient data for correlation purposes. (From Ref.[1,20].)

INTERFACIAL AREA IN LIQUID–LIQUID DISPERSIONS

The liquid–liquid interfacial area per unit volume available for mass transfer can be obtained from:

$$a_V = \frac{6\phi}{d_{32}} \tag{9}$$

where d_{32} can be calculated from the equations given previously, at least for non- and weakly coalescing systems.

MASS TRANSFER COEFFICIENT IN LIQUID–LIQUID DISPERSIONS

In principle, the mass transfer coefficient for a *single* liquid spherical droplet in an immiscible liquid flowing with velocity ν_{LL} past the spherical droplet can be calculated from a Froessling-type equation, which was originally derived for a solid particle (see Atiemo-Obeng, Penney, and Armenante.[20]):

$$\mathrm{Sh_{LL}} = 2 + 0.6\mathrm{Re_{LL}^{0.5}Sc^{0.33}} \tag{10}$$

In this equation, the liquid–liquid Sherwood number for a single sphere, $\mathrm{Sh_{LL}} = k_{LL}d_{32}/D_L$, is a function of the liquid–liquid Reynolds number, $\mathrm{Re_{LL}} = \rho_L\nu_{LL}d_{32}/\mu$ and Schmidt number, $\mathrm{Sc} = \mu/(\rho_L D_L)$. This equation is of limited use in mixing vessels where the droplet-liquid slip velocity, ν_{LL}, cannot be determined. Instead, the liquid–liquid mass transfer coefficient can be obtained from empirical equations such as (see Skelland and Xien[29]; Skelland and Moeti[30]):

$$\frac{k_L\, d_{32}}{D_L} = 1.237 \times 10^{-5}\, \mathrm{Sc}_c^{1/3}\, \mathrm{Re}^{2/3} \times \mathrm{Fr}^{5/12}\left(\frac{D}{d_{32}}\right)^2\left(\frac{d_{32}}{T}\right)^{1/2}\left(\frac{\rho_d d_{32}^2 g}{\sigma}\right)^{5/4}\phi^{-1/2} \tag{11}$$

The generic equation for the mass transfer rate is:

$$\dot{m} = k_L a_V V_L \Delta C \tag{12}$$

where ΔC is the concentration driving force across the phases. This equation shows that knowledge of a_V [through Eq. (9)] and k_L [through Eq. (11)] is needed to calculate the mass transfer rate. In general, V_L can be taken as the entire volume of the dispersion, provided that the agitation intensity is well beyond N_{min}.

CONCLUSIONS

Apart from certain formulated products, such as emulsions, liquid–liquid dispersion is rarely carried out for its own sake. It is usually accompanied by heat/mass transfer and chemical reaction, thereby complicating scale-up. Mass transfer and heat transfer are significantly affected by mixing. Hence, mixing is a critical element in the analysis of any liquid–liquid reactor system. The material presented in this entry serves to demonstrate that fundamental knowledge must be coupled with practical insight and engineering judgment to solve problems associated with real industrial liquid–liquid applications.

NOMENCLATURE

A	Interfacial area between phases, m^2
a_V	Interfacial area per unit volume of liquid, m^2/m^3
B	Baffle width, m
C	Impeller clearance measured from the impeller centerline to the vessel bottom off the vessel bottom, m
C_1,C_2,C_3,C_4	Dimensionless empirical constants
ΔC	Concentration driving force for mass transfer, mol/m^3
d_{32}	Sauter mean drop diameter $\left(=\sum_{i=1}^{i=m} n_i d_i^3 / \sum_{i=1}^{i=m} n_i d_i^2\right)$, m
d_{max}	Maximum stable drop diameter, m
D	Impeller diameter, m
D_L	Diffusivity of dissolved component or reactant in liquid, m^2/s
g	Gravitational acceleration, m/s^2
H	Height of liquid in vessel, m
k_L	Mass transfer coefficient, m/s
k_{LL}	Mass transfer coefficient for a single spherical droplet immersed in a liquid flowing at constant velocity past the droplet, m/s
m_L	Mass of liquid, kg
$\dot{m}$	Rate of mass transfer of solute or reactant, kg/s
N	Impeller speed, rotations/s
N_{js}	Minimum speed to just suspend solid particles in vessel, rotations/s
N_{min}	Minimum impeller speed to completely incorporate dispersed phase into continuous phase in liquid–liquid systems, rotations/s
P	Power dissipation, W
t	Time, s
T	Vessel diameter, m
V_{LL}	Liquid velocity past an immiscible liquid droplet (slip velocity), m/s
V_L	Volume of liquid in vessel, m^3
W	Width of an impeller blade, m

Greek Symbols

ε_{avg}	Average energy dissipation rate (or power draw) per mass of mixture, W/kg
ε_{max}	Maximum energy dissipation rate (or power draw) per mass of mixture, W/kg
ϕ	Volume fraction of dispersed phase, i.e., ratio of the volume of dispersed phase to the volume of dispersed and continuous phases
μ	Viscosity, Pa s
μ_c	Viscosity of continuous phase, Pa s
μ_d	Viscosity of dispersed phase, Pa s

ν	Kinematic viscosity, m^2/s
ρ_L	Density of liquid, kg/m^3
ρ_c	Density of continuous phase, kg/m^3
ρ_d	Density of dispersed phase, kg/m^3
$\bar{\rho}$	Average density of mixture, kg/m^3
$\Delta\rho$	Density difference between phases, kg/m^3
σ	Interfacial tension, N/m

Dimensionless Groups

N_p	Power number, $P/(\bar{\rho}N^3D^5)$
Re	Impeller Reynolds number, $\bar{\rho}ND^2/\mu$
Re_{LL}	Liquid–liquid Reynolds number, $\rho_L V_{LL} d_{32}/\mu$
Sc	Schmidt number, $\mu/(\rho_L D_L)$
Sc_c	Schmidt number for the continuous phase, $\mu_c/(\rho_c D_L)$
Sh_{LL}	Sherwood number for a single droplet, $k_{LL}d_{32}/D_L$
Vi	Viscosity number, $(\rho_c/\rho_d)^{1/2}\mu_d ND/\sigma$
We	Weber number for droplets in stirred vessel, $\rho_c N^2 D^3/\sigma$

REFERENCES

1. Leng, D.E.; Calabrese, R.V. Immiscible liquid–liquid systems. In *Handbook of Industrial Mixing: Science and Practice*; Paul, E.L., Atiemo-Obeng, V.A., Kresta, S.M., Eds.; John Wiley & Sons: New York, 2004; 639–753.
2. Atiemo-Obeng, V.A.; Calabrese, R.V. Rotor–stator mixing devices. In *Handbook of Industrial Mixing: Science and Practice*; Paul, E.L., Atiemo-Obeng, V.A., Kresta, S.M., Eds.; John Wiley & Sons: New York, 2004; 479–505.
3. Hemrajani, R.R.; Tatterson, G.B. Mechanically stirred vessels. In *Handbook of Industrial Mixing: Science and Practice*; Paul, E.L., Atiemo-Obeng, V.A., Kresta, S.M., Eds.; John Wiley & Sons: New York, 2004; 345–390.
4. Sprow, F.B. Drop size distributions in strongly coalescing liquid–liquid systems. AIChE J. **1967**, *13*, 995–998.
5. Pacek, A.W.; Moore, I.P.T.; Calabrese, R.V.; Nienow, A.W. Evolution of drop size distributions and average drop diameters in liquid–liquid dispersions before and after phase inversion. Trans. Inst. Chem. Eng. **1993**, *71A*, 340–341.
6. Pacek, A.W.; Moore, I.P.T.; Nienow, A.W.; Calabrese, R.V. A video technique for the measurement of the dynamics of liquid–liquid dispersions during phase inversion. AIChE J. **1994**, *40*, 1940–1949.
7. Pacek, A.W.; Nienow, A.W.; Moore, I.P.T. On the structure of turbulent liquid–liquid dispersed flows in an agitated vessel. Chem. Eng. Sci. **1994**, *49*, 3485–3498.
8. Brooks, B.W.; Richmond, H.N. Dynamics of liquid–liquid phase inversion using non-ionic surfactants. Colloids Surf. **1991**, *58*, 131–148.
9. Brooks, B.W.; Richmond, H.N. Phase inversion in non-ionic surfactant-oil-water systems. I. The effect of transitional inversion on emulsion drop sizes. Chem. Eng. Sci. **1994**, *49*, 1053–1064.
10. Brooks, B.W.; Richmond, H.N. Phase inversion in non-ionic surfactant-oil-water systems. II. Drop size studies in catastrophic inversion with turbulent mixing. Chem. Eng. Sci. **1994**, *49*, 1065–1075.
11. Brooks, B.W.; Richmond, H.N. Phase inversion in non-ionic surfactant-oil-water systems. III. The effect of oil phase viscosity on catastrophic inversion and the relationship between the drop size present before and after catastrophic inversion. Chem. Eng. Sci. **1994**, *49*, 1843–1853.
12. Kinugasa, T.; Watanabe, K.; Sonove, T.; Takeuchi, H. Phase inversion of stirred liquid–liquid dispersions. Presented at the International Symposium on Liquid–Liquid Two Phase Flow and Transport Phenomena, Session 13, Antalya, Turkey, 1997.
13. Davies, J.T. A physical interpretation of drop sizes in homogenizers and agitated tanks, including the dispersion of viscous oils. Chem. Eng. Sci. **1987**, *42*, 1671–1676.
14. Zhou, G.; Kresta, S.M. Correlation of mean drop size with the turbulence energy dissipation and the flow in an agitated tank. Chem. Eng. Sci. **1998**, *53*, 2063–2079.
15. Zhou, G.; Kresta, S.M. Evolution of drop size distribution in liquid–liquid dispersions for various impellers. Chem. Eng. Sci. **1998**, *53*, 2099–2113.
16. Wang, C.Y.; Calabrese, R.V. Drop breakup in turbulent stirred-tank contactors. Part II. Relative influence of viscosity and interfacial tension. AIChE J. **1986**, *32*, 667–676.
17. Chen, H.T.; Middleman, S. Drop size distribution in agitated liquid–liquid systems. AIChE J. **1967**, *13*, 989–995.
18. Calabrese, R.V.; Wang, C.Y.; Bryner, N.P. Drop breakup in turbulent stirred-tank contactors. Part III. Correlations for mean size and drop size distribution. AIChE J. **1986**, *32*, 677–681.
19. Calabrese, R.V.; Chang, T.P.K.; Dang, P.T. Drop breakup in turbulent stirred-tank contactors. Part I. Effect of dispersed phase viscosity. AIChE J. **1986**, *32*, 657–666.
20. Atiemo-Obeng, V.A.; Penney, W.R.; Armenante, P.M. Solid–liquid mixing. In *Handbook of Industrial Mixing: Science and Practice*; Paul, E.L., Atiemo-Obeng, V.A., Kresta, S.M., Eds.; John Wiley & Sons: New York, 2004; 543–584.

21. Skelland, A.H.P.; Seksaria, R. Minimum impeller speeds for liquid–liquid dispersion in baffled vessels. Ind. Eng. Chem. Proc. Des. Dev. **1978**, *17*, 56–61.
22. Armenante, P.M.; Huang, Y.-T. Experimental determination of the minimum agitation speed for complete liquid–liquid dispersion in mechanically agitated vessels. Ind. Eng. Chem. Res. **1992**, *31*, 1398–1406.
23. Armenante, P.M.; Huang, Y.-T.; Li, T. Determination of the minimum agitation speed to attain the just dispersed state in solid–liquid and liquid–liquid reactors provided with multiple impellers. Chem. Eng. Sci. **1992**, *47*, 2865–2870.
24. Tatterson, G.B. *Fluid Mixing and Gas Dispersion in Agitated Tanks*; McGraw-Hill: New York, 1991.
25. Bates, R.L.; Fondy, P.L.; Corpstein, R.R. An examination of some geometric parameters of impeller power. Ind. Eng. Chem. Proc. Des. Dev. **1963**, *2*, 310–314.
26. Bates, R.L.; Fondy, P.L.; Fenic, J.C. Impeller characteristics and power. In *Mixing*; Uhl, V.W., Gray, J.B., Eds.; Academic Press: New York, 1966; Vol. I, 111–178.
27. Armenante, P.M.; Chang, G.-M. Power consumption in agitated vessels provided with multiple disk turbines. Ind. Eng. Chem. Res. **1998**, *37*, 284–291.
28. Armenante, P.M.; Mazzarotta, B.; Chang, G.-M. Power consumption in stirred tanks provided with multiple pitched-blade turbines. Ind. Eng. Chem. Res. **1999**, *38*, 2809–2816.
29. Skelland, A.H.P.; Xien, H. Dispersed-phase mass transfer in agitated liquid–liquid systems. Ind. Eng. Chem. Res. **1990**, *29*, 415–420.
30. Skelland, A.H.P.; Moeti, L.T. Mechanisms of continuous-phase mass transfer in agitated liquid–liquid systems. Ind. Eng. Chem. Res. **1990**, *29*, 2258–2267.

Lithium–Ion Battery

Chung-Chiun Liu
Xuekun Xing
Electronics Design Center and Department of Chemical Engineering, Case Western Reserve University, Cleveland, Ohio, U.S.A.

INTRODUCTION

The lithium-ion (Li-ion) battery is a large-scale-commercialized rechargeable battery system of the highest energy density to date. Sony introduced the first commercial Li-ion battery in the market in 1991. Following continuous improvement in battery performance, a rapid growth in production volume of this battery has continued unabated. The production volume of various Li-ion batteries approaches about 740 million units with a total revenue of about U.S. $2.4 billion worldwide in 2002.[1] In addition to its high energy density and specific energy, the Li-ion cell has a number of other advantages, such as long cycle life, wide operational temperatures, and reasonable rate performance. All these advantages make Li-ion batteries an attractive power source for many applications, especially for various portable electronic devices such as cellular phones, notebook PCs, PDAs, digital cameras, and audio–video systems. In some traditional applications, the Li-ion batteries will replace the nickel–cadmium (Ni–Cd) and nickel–metal hydride (Ni–MH) batteries, and this trend is expected to continue in the future.

The Li-ion cell has several marked characteristics that differentiate it from most of the other rechargeable cell systems. First of all, the Li-ion cell is a ''high-operation-voltage'' system with an average operation voltage of 3.6–3.7 V, which is about three times that of Ni–Cd or Ni–MH cells. The higher operation voltage becomes practically possible because of the employment of selected nonaqueous electrolytes, which have a much wider electrochemical window compared with its aqueous counterpart, for which the electrochemical window is governed by the water decomposition reaction. Second, there is no metallic lithium used in the Li-ion cell, which thus avoids the difficulties usually associated with the formation of the lithium dendrites on the metallic Li anode during charge–discharge cycles. Third, the cell reaction of the Li-ion cell is not a traditional redox reaction but rather an intercalation/deintercalation process of Li-ions in the anode- and the cathode-active materials, respectively.

Although the commercial Li-ion batteries have been marketed for just about 11 yr, the basic concept and research work related to the development of the Li-ion cell can be traced back to the early 1970s. Whittingham did pioneer work on the mechanism of the intercalation reaction, and various insertion compounds such as TiS_2, MoS_2, V_6O_{13}, etc. were studied and tested as cathode-active materials in nonaqueous solutions.[2–5] Steele suggested the use of graphite and TiS_2 as electrodes for battery applications in 1973.[6] Goodenough's group studied the properties of various lithiated transition metal oxides (Li_xMO_2, M = Ni, Co, Mn) and proposed to use these compounds as the cathode-active material in the early 1980s.[7–10] In the same period of time, the initial concept of the Li-ion cell with two intercalation electrodes in a nonaqueous solution was proposed and studied by Murphy and coworkers.[11] After intense closed-door research and development work, Sony introduced the first commercial Li-ion cell based on a carbon anode and a $LiCoO_2$ cathode in 1991.

CELL COMPONENTS AND CELL CHEMISTRY

Cathode Material and Cathode Reaction

Layered structural $LiCoO_2$ is the cathode-active material being used for most of the commercial Li-ion batteries today. In 1980 Goodenough et al. introduced this compound for the cathode-active material.[7–10] Fig. 1 shows the typical layered structure of $LiCoO_2$, in which the oxide ions form a cubic-close-packed array with close-packed (111) octahedral-site planes alternately filled with Li^+ and Co^{3+} ions. $LiCoO_2$ has good electronic conductivity mainly contributed by the availability of the holes of large O-2p character. On the other hand, Li^+ ions in the structure are in part movable, that is, Li^+ ions can be reversibly extracted from and reinserted to the $LiCoO_2$ structure, as shown in Eq. (1) below:

$$LiCoO_2 \underset{\text{discharge}}{\overset{\text{charge}}{\rightleftharpoons}} Li_{1-x}CoO_2 + xli^+ + xe^- \qquad (1)$$

It is worth noting that the reversibility of the reaction (1) is available only in a range of $0 \leq x \leq 0.5$.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007843

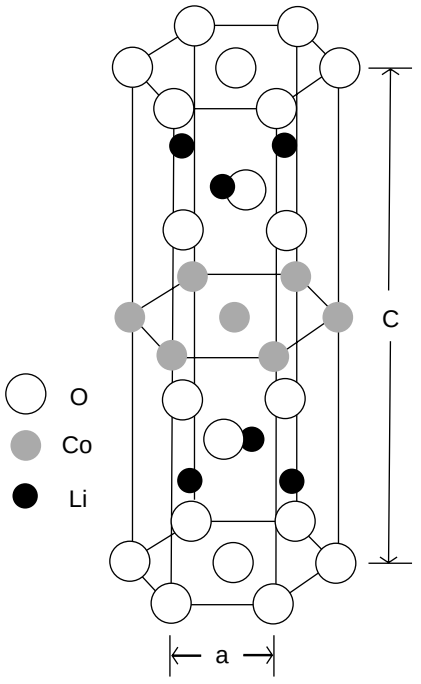


Fig. 1 Schematic diagram of the layered $LiCoO_2$ structure, showing the stacking of the O-Li-O-Co-O-Li-O layers. (From Ref.[10].)

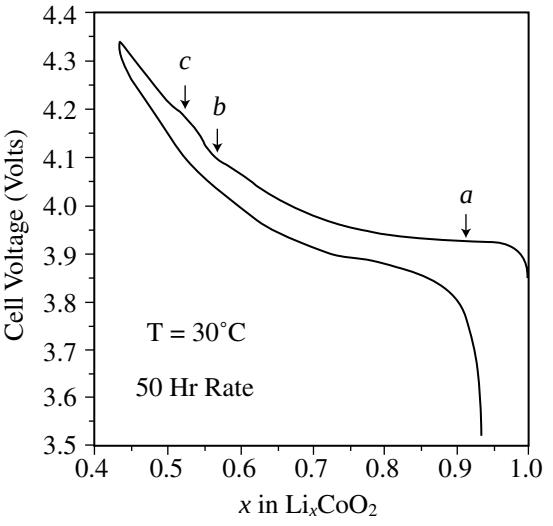


Fig. 2 Cell voltage vs. x in Li_xCoO_2 for the first charge and discharge. (From Ref.[12].)

In the case of $x > 0.5$, oxygen evaluation may take place and the crystal structure may be damaged accordingly.

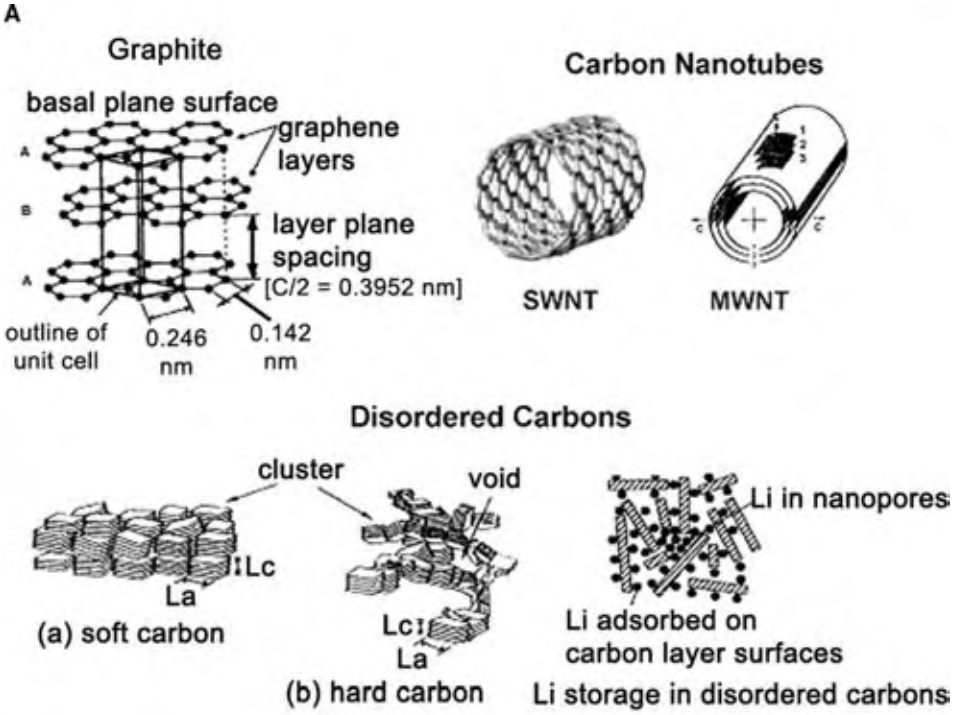
Fig. 2 shows the voltage-composition curve associated with Eq. (1). The three plateaus marked by arrows indicate the occurrence of three distinct phase transitions as x varies from 1 to 0.4 in Li_xCoO_2 during the deintercalation of lithium.[12] The limitation of $x = 0.5$ implies that the reversible specific capacity of $LiCoO_2$ is 137 mA hr/g. As shown in Fig. 2, within the limitation of the compositions, the voltage accompanied with the Li^+ intercalation/deintercalation process is around 4 V (vs. Li), which is much higher than what is shown by cathode materials in all aqueous batteries, and that leads to a much higher energy density of the Li-ion cell.

Anode Material and Anode Reaction

To date, all commercial Li-ion batteries use carbonaceous materials as the anode-active material. Fig. 3A shows the structural schematics of typical carbonaceous materials in the field. Among various carbonaceous materials, graphite and disordered carbons have been employed dominantly in Li-ion batteries.[13] Graphite has a typical layered structure that consists of stacked graphene sheets with an ABAB ... sequence along the C-axis as shown in Fig. 3A. The graphene sheets are bonded together by van der Waals forces with an intersheet gap of about 0.3354 nm, which provides a space and functions as a host for enabling the Li ion intercalation to take place. It has been found that the Li ion intercalation/deintercalation in/out graphite structure during charge/discharge processes is reversible, and the reaction scheme can be described as follows:

$$6C + xLi^+ + xe^- \underset{\text{discharge}}{\overset{\text{charge}}{\rightleftharpoons}} Li_xC_6 \qquad (2)$$

The relationship between the potential and the specific capacity (equivalently, the amount of Li-ions intercalated into the anode) of the graphite anode is shown in Fig. 3B. From Eq. (2), a theoretical specific capacity (372 mA hr/g at $x = 1$) of the graphite can be derived. One important characteristic of the Li-ion intercalation/deintercalation reaction in graphite is that the reaction takes place at potentials close to the metallic lithium potential in the given electrolytes (see Fig. 3B). This characteristic makes graphite act as an anode having an operational voltage similar to that of the metallic lithium while avoiding any cycling difficulty associated with the latter because of the formation of lithium dendrites in the cell charging process. The good reversibility and unique potential characteristics make graphite a good substitute for the metallic lithium and an ideal anode material for Li-ion cells that can provide both good charge/discharge cycle life and high energy density.



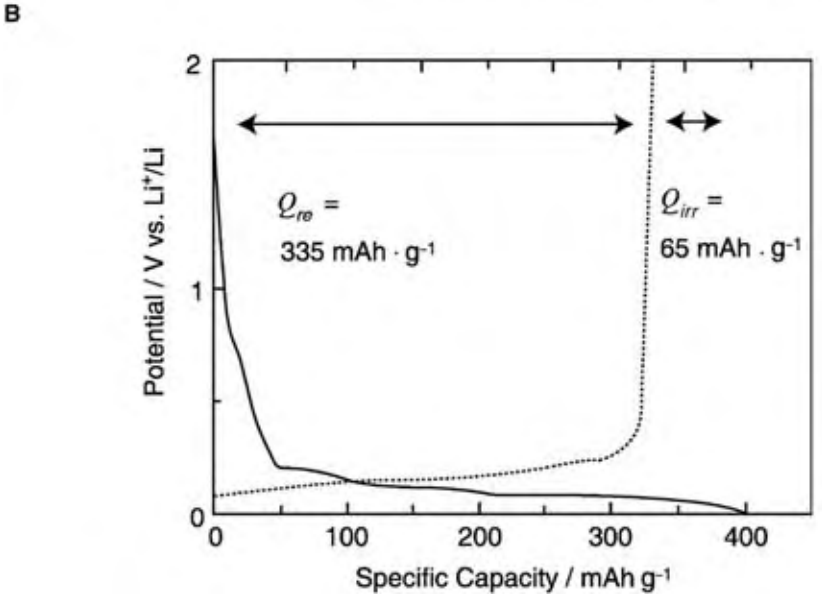


Fig. 3 (A) Schematic of structures of various types of carbons that can insert lithium reversibly. (B) Charge and discharge characteristics of natural graphite powder (NG-7) at the first cycle in the 1 M $LiClO_4$/EC + DEC (1 : 1 vol/vol). (From Ref.[17].)

The mechanism of lithium electrochemical intercalation into graphite has been substantially studied. The intercalation process is characterized by the staging phenomena and the lithium–graphite intercalation compounds have a staging structure. This means that the lithium intercalate layers undergo a successive distribution of stages with respect to graphite host layers following the progress of the intercalation, as schematically shown in Fig. 4. The stage structure changes from a higher to a lower stage during the charge (lithium intercalation), and an opposite order is followed during discharge (lithium deintercalation).

There is always some irreversible capacity loss during the first charge/discharge steps (Li intercalation/deintercalation) of a Li-ion cell. The irreversible capacity loss is attributed to the formation of a solid electrolyte interface (SEI) layer on the graphite surface at the initial stage of the lithium intercalation process.[14] The formation of the SEI layer is crucial to maintain the stability of a graphite anode for the long-term charge/discharge cycles.

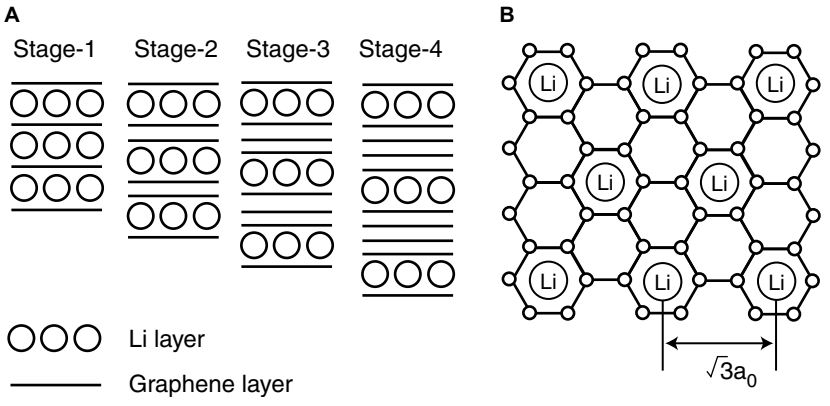


Fig. 4 Schematic of stage structures of lithium intercalation in graphite: (A) stage structures of Li-GICs and (B) in-phase structure of stage-1 LiGIC. (From Ref.[13].)

Overall Cell Reaction

The overall cell reaction of the Li-ion cell can thus be described as follows:

$$6C + LiCoO_2 \underset{\text{discharge}}{\overset{\text{charge}}{\rightleftharpoons}} Li_xC_6 + Li_{1-x}CoO_2 \qquad (3)$$

This cell reaction is the base of the high average operational voltage of 3.6–3.7 V for the Li-ion cell, which is substantially higher than any other available rechargeable battery systems.

Electrolyte and SEI Layer

The Li-ion cell uses a nonaqueous electrolyte as the ionic-conductive medium that consists of a Li salt and a combination of two or more organic solvents. The majority of commercially available Li-ion batteries use a mixture of solvents consisting of a cyclic alkylene carbonate [typically, ethylene carbonate (EC) and propylene carbonate (PC)]; and a linear alkyl carbonate [such as diethyl carbonate (DEC), dimethyl carbonate (DMC), and ethyl methyl carbonate (EMC)]. Two properties, among others, of the solvent are crucial to its application to a Li-ion cell. The first one is its electrochemical window, which should be sufficient to meet the requirement of the high operation voltage of a Li-ion cell. The second one is the ionic conductivity of the electrolyte, which should be large enough to meet the requirement of the rate capability of a Li-ion cell. The electrochemical window of a solvent can be determined from its oxidation potential and its reduction potential. Generally, organic solvents undergo oxidation at its oxidation potential with CO_2 and other small organic compounds as the oxidation products. On the other hand, organic solvents undergo reduction if the potential is low enough to approach its reduction potential. A comprehensive summary on the liquid electrolytes used in Li-ion cells has been presented recently.[15]

The solvent reduction takes place usually at the anode, and it is closely related to the formation of the SEI film on the surface of the anode. The SEI film is an electronic insulator but an ionic conductor, which covers the active surface of the anode and functions as a passivation layer to stabilize the anode-active material (e.g., blocking the solvent cointercalation into the anode-active material) for sustaining a long-term charge–discharge cycle life. The mechanism of the SEI formation at the anode has been studied in depth in recent years. It has been generally agreed that the SEI formation results from complex reactions involving solvents, salt, atmospheric impurities (e.g., O_2, N_2, CO_2, H_2O), and additives, if any.[16] It has been established that $(CH_2CO_2Li)_2$ is a major surface species formed on carbon electrodes for EC-based electrolytes. Usually, the SEI films involve other reaction products such as LiF, Li_2CO_3, Li_2O, LiOH, etc. Aurbach has provided a comprehensive summary for the formation and the composition of the SEI films in various electrolytes.[17] The SEI film formation and its properties (e.g., its adhesive property on the carbon surface, its forming potential and film smoothness, etc.) are crucial to maintain the long-term stability of graphite electrodes in Li insertion/deinsertion processes. The quality of the SEI film is dependent on many factors including solvents, salts, impurities, polarization current, and additives, and the latter has especially attracted substantial effort in recent years (see the following section).

Electrolyte Additives and Functional Electrolytes

The first major effort to bring additives into the electrolytes of a Li-ion cell is driven by overcoming the problems of PC-based electrolytes, which fail when used for the Li-ion cells with a graphite anode (large irreversible capacity loss at the first charge–discharge step, and poor charge–discharge cycle life).

One argument is to relate this failure to the co-intercalate of the PC solvent molecules into the graphite structure together with the solvated Li-ions.[18,19] This co-intercalation process may lead to an exfoliation of the graphite structure. In addition, co-intercalated PC molecules may undergo reduction within the graphite and its reduction products could block Li-ion to enter into the graphite lattice. As a result of these effects, the battery cycle life is adversely affected. Aurbach and coworkers have proposed a different mechanism, and they attribute the failure of the graphite electrode in the PC-based electrolytes to the failure in forming a smooth, cohesive SEI layer on the graphite surface mainly because of the structure hindrance caused by the methyl group on PC molecules.[17,20] Furthermore, the formation of propylene gas during PC reduction could cause higher internal pressure within graphite crevices and split the particles, which could result in the loss of electrical contact with the current collector because of isolation by surface films.

It has been found that the addition of some cyclic ethers, such as 12-crown-chromium-4 (12Cr4), to PC-based electrolytes can significantly improve the reversibility of a graphite electrode.[21,22] Crown ether is a strong complex agent for Li-ions, and the preferential formation of the Li^+–crown ether complexes shields PC out when the Li ions approach the electrode.

Separator

Similar to other battery systems, the Li-ion cell needs a separator to electronically separate its cathode and anode to avoid any electrical short circuit between them, while keeping a good ionic pathway through the electrolyte absorbed in the pore structure of the separator. An adequate separator material needs to meet the following requirements among others: 1) a good electronic insulator; 2) a high degree of porous structure that is able to absorb a sufficient quantity of the electrolyte ensuring a good ionic conductance; 3) a good chemical resistance to the electrolyte and good electrochemical stability in the cell operation conditions; 4) an adequate mechanical and physical strength; and 5) a good wettability to the electrolyte used. To date, the most commonly used separator materials for Li-ion cells are single-layer porous polyethylene (PE) and polypropylene/polyethylene/polypropylene multiple-layer (PP/PE/PP) films (Celgard type). These separators have a thermal shutdown function. The PE porous film will melt down partly when the temperature approaches its melting point (typically 130–140°C). This partial melt-down process leads to pore closing while maintaining the film integrity, which results in maximizing the internal resistance of the cell while minimizing the current and thus avoiding continuous increase in the cell temperature. This function is especially important to protect a Li-ion cell from overcharge and short circuit, because any abnormal current can be automatically cut off because of the rapid increase in the cell resistance.

Li-Ion Cell Construction

The first generation of the Li-ion battery manufactured by Sony is a cylindrical structure, and it is known as model 18650 (cell diameter: 18 mm, cell height: 65 mm). This cell has been manufactured on a very large scale and is widely employed in notebook PCs and other electronic devices. With the rapid growth of cellular phones, PDA, and other thin, light, portable electronic devices, a large portion of Li-ion batteries are being produced with a thin prismatic configuration to meet the market needs. A prismatic cell usually has a rectangular and/or square-shaped metal case (stainless steel or aluminum) and, to some extent, a porch envelope made of aluminum/polymer laminate. This laminate, which is initially employed for a Li-ion polymer (LiP) battery and has been extended to the so-called advanced Li-ion batteries, has a lighter weight and thus improves the cell's specific energy.

Both the cylindrical and the prismatic cells use a wound core structure that is usually referred to as a "jelly-roll." Fig. 5 shows the typical structure of a cylindrical Li-ion cell.[23] Prior to making the jelly-rolls, the anode and cathode films will first be prepared by coating the corresponding slurry consisting of selected active material, conductive material, binder, and solvent on the thin metal strips (copper foil for the anode, aluminum foil for the cathode, respectively), followed by drying, calendaring, and slitting operations. Then, an anode strip, a cathode strip, and the separator strips with adequate length, width, and thickness are combined together and wound into a tight coil, jelly-roll, via a winding machine. It is crucial to maintain close contact without a void or gap between electrodes and separators to ensure good performance of the Li-ion cells fabricated. The coil fabricated is then inserted into a metal casing of adequate dimension and shape, which functions not only as a container to house the electrodes and the electrolyte, but also to provide a hermetical seal from the atmosphere. This container also provides constant internal pressure to further closely hold the electrodes and the separators together. After a thorough vacuum-heat drying to remove any water inside the container, the casing with the jelly-roll is then filled with a selected electrolyte solution of adequate quantity. The electrolyte filling process is usually carried out with a precision pump under vacuum conditions, and it is operated in a dry-room or a dry-box to

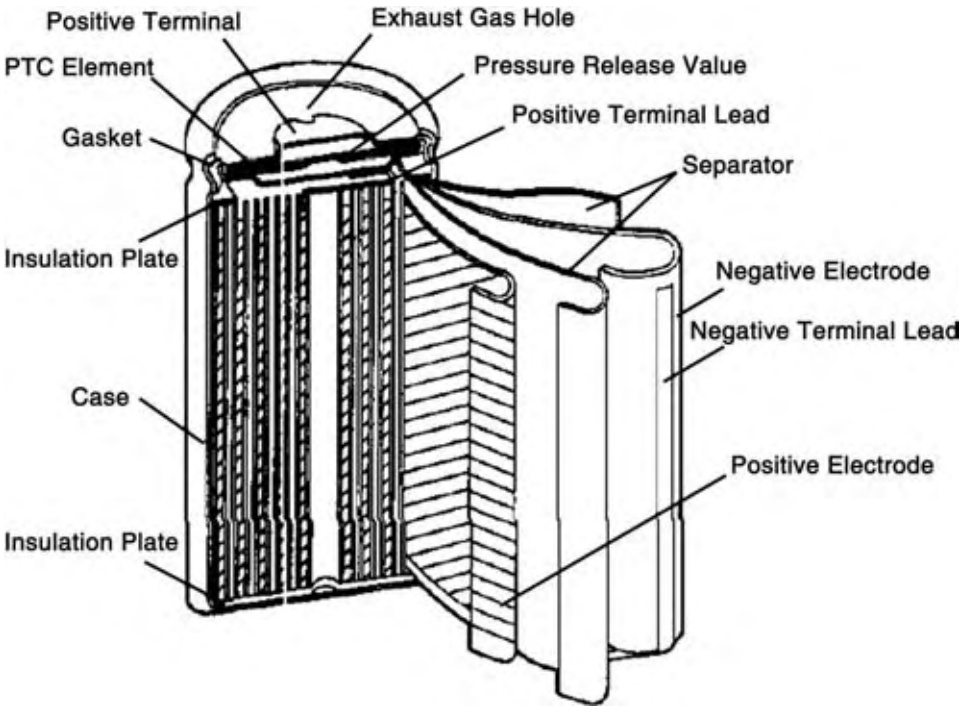


Fig. 5 Typical structure of the cylindrical Li-ion battery. (From Ref.[23].)

minimize moisture absorption. The final step in this assembly is to seal the cell casing by a controlled compression of a polymer gasket located between the cell casing and the casing cap. The sealed cell can then enter a first charging step, namely, the cell formation process.

LI-ION CELL PERFORMANCE CHARACTERISTICS

Charge Characteristics

The most common charging mode of Li-ion cells is the ''constant-current/constant-voltage'' (CCCV) method, by which the cell is charged first at a selected constant current (e.g., $c/2$, $1c$, or others); once the cell voltage approaches a predetermined value (typically, 4.1–4.2 V), the charging continues at this constant voltage until the charging current decreases to a predetermined cutoff value (e.g., $c/20$ or so). Fig. 6 shows a typical charge voltage–current-capacity profile of a cylindrical 18650 Li-ion cell. Unlike Ni–Cd cells, Li-ion cells usually have a near-100% charging efficiency at normal conditions; overcharging is not necessary and actually should be avoided. The charging voltage at the constant-voltage stage is crucial to ensure good performance of Li-ion cells. If the charging voltage is too low, the cell design capacity may not be fully used because of a lesser amount of Li extracted from the cathode at a lower cell-charging voltage. On the other hand, if the charging voltage is too high, it may cause structure damage of the cathode-active material because of overextracting Li. Besides, it may also cause metallic lithium deposition at the surface of the anode because of excess lithium over the intercalation capacity of the anode. All these may adversely impact the battery cycle life and may also induce safety concerns.

Discharge Characteristics

Li-ion cells can be discharged under different modes depending on the needs of the user. Constant current discharge is one of the most commonly used, by which the Li-ion cell undergoes discharge at a wide range of selected currents depending on the applications. The pulse current discharge [such as Global System for Mobile Communication (GSM) operation mode for cellular phones, etc.] is another mode widely employed in the mobile phone applications. The constant-power discharge is typically used in notebook PCs. Fig. 7 shows the typical voltage–current curves under two constant-power conditions for cylindrical 18650 cells. The cutoff voltage of discharge is usually in the range of 2.7–3.0 V. A lower cutoff voltage may shorten the cycle life of the cell because of the possible oxidation of the copper anode current collector.

Energy Density and Specific Energy

Li-ion batteries belong to the high-energy battery systems and have a substantially higher energy density and specific energy compared to other rechargeable

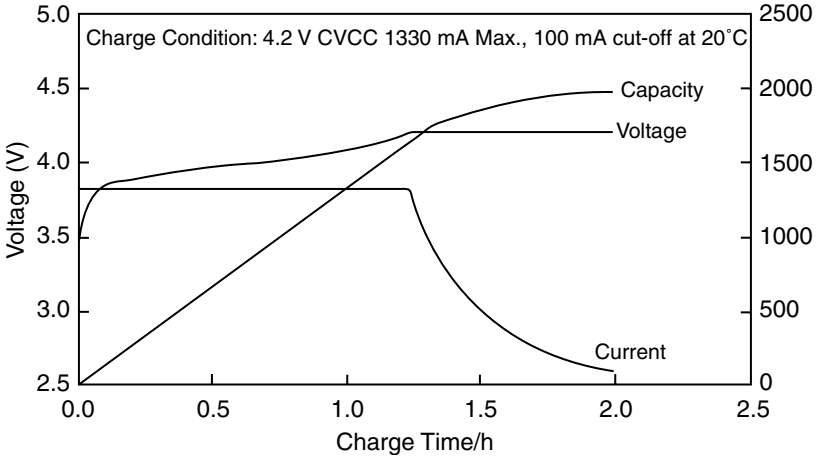


Fig. 6 Charge profile (current, voltage, and capacity) of Panasonic's CGR 18650A Li-ion battery (standard capacity: 2000 mA hr). (From Ref.[23].)

batteries such as lead–acid, Ni–Cd, Ni–MH, etc. The energy density of Li-ion batteries is related to the battery size, and it increases with increasing battery size. Typically, a single Li-ion cell of small to medium size used for normal cellular phone applications has an energy density approximately 350 W hr/L and a specific energy of about 170 W hr/kg, which are significantly higher than those of a lead–acid and a Ni–Cd battery.

Rate Capability

Li-ion batteries have a reasonable rate capability, which can meet the power requirement of various portable electronic devices. On the other hand, because Li-ion cells use nonaqueous electrolyte and the cell reaction involves a Li-ion solid diffusion process, its overall rate capability and power density/specific

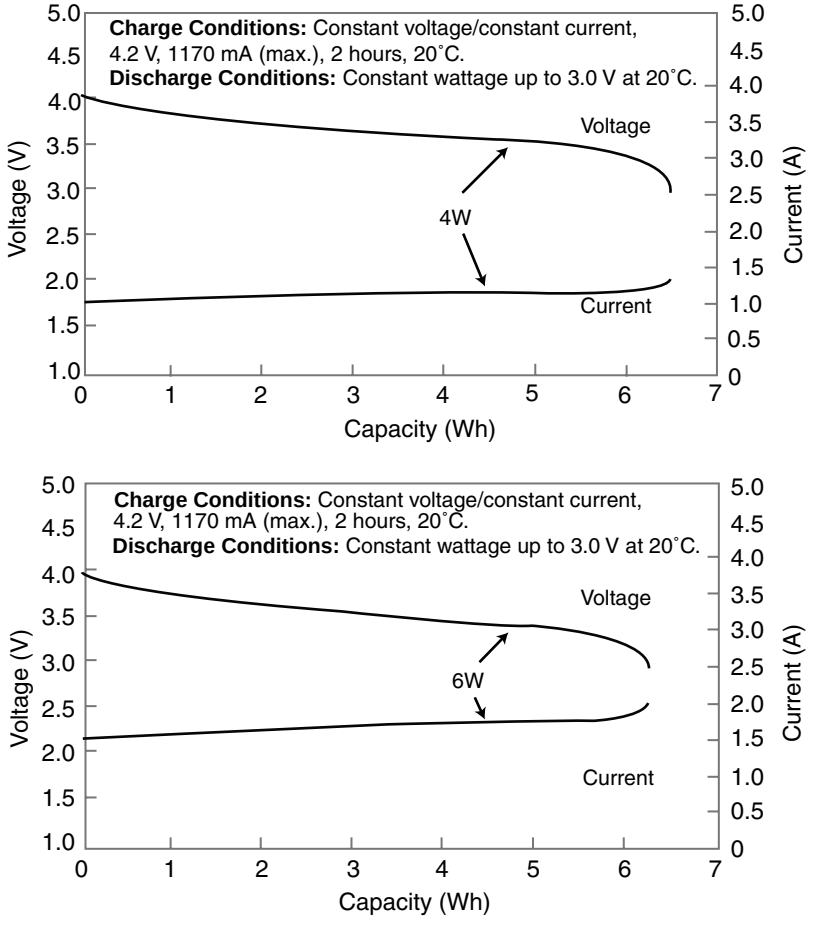


Fig. 7 Voltage–current profile at constant power discharge of Panasonic cylindrical CGR18650HG Li-ion batteries (standard capacity: 1800 mA hr). (From Ref.[23].) *(View this art in color at www.dekker.com.)*

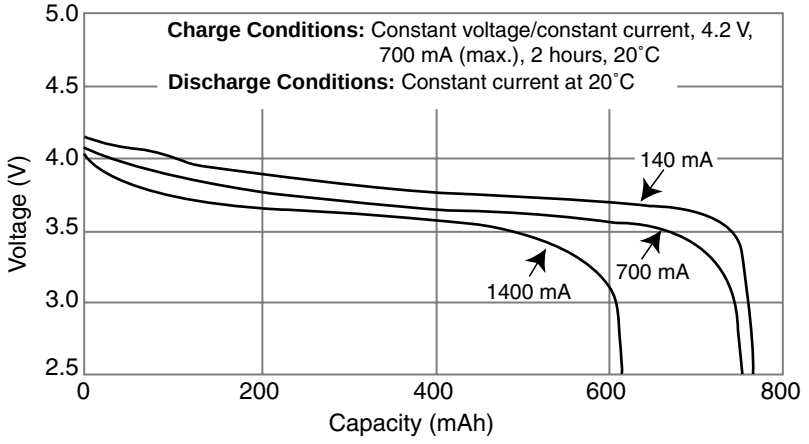


Fig. 8 Discharge characteristics at various drain rates of Panasonic CGA533048 Li-ion batteries (nominal capacity: 750 mA hr). (From Ref.[23].)

power are lower than for its aqueous counterparts such as lead–acid and Ni–Cd cells. Fig. 8 shows the typical discharge curves of a prismatic 533048 Li-ion cell at various drain rates at room temperature. The rate capability of Li-ion batteries can be improved through cell design such as reducing the electrode thickness, selecting adequate separators and electrolytes, and others.

Cycle Life

Li-ion batteries are known to have a long charge–discharge cycle life because the cell reactions do not involve large structural changes of active materials at both the anode and the cathode. At the normal charge/discharge conditions, a Li-ion battery can be cycled at 100% the depth of discharge (DOD) over 500 cycles, still remaining over 80% of the initial capacity. Fig. 9 shows the typical charging–discharging cycle performance of a prismatic 633450 Li-ion cell. Cell cycle life is strongly dependent on the cell design, the active materials, the electrolyte, and others. Manufacturing conditions, such as electrode coating quality, battery casing sealing quality, moisture control, etc., also play an important role in the impacting of the cycle life of a cell.

Storage Life and Self-Discharge

In addition to its long cycle life, Li-ion batteries have a low self-discharge rate and a long storage life as well. Figs. 10A and 10B show the typical storage performance of Li-ion cells. Another advantage of Li-ion batteries, compared to Ni–Cd batteries, is its lack of memory effect. Although a Li-ion cell can be stored at various states of charge (SOC), it is recommended to store it at a half-charged state, because a fully charged state may promote oxidative side reactions associated with the electrolyte, while a fully discharged state may lead to a possible copper oxidation. During long-term storage, Li-ion cells are expected to lose a portion of their original capacity because of the self-discharge. Although the major portion of this capacity loss can be restored after a few charging–discharging cycles (see Fig. 10B), there is always a portion of the capacity loss that cannot be totally recovered because

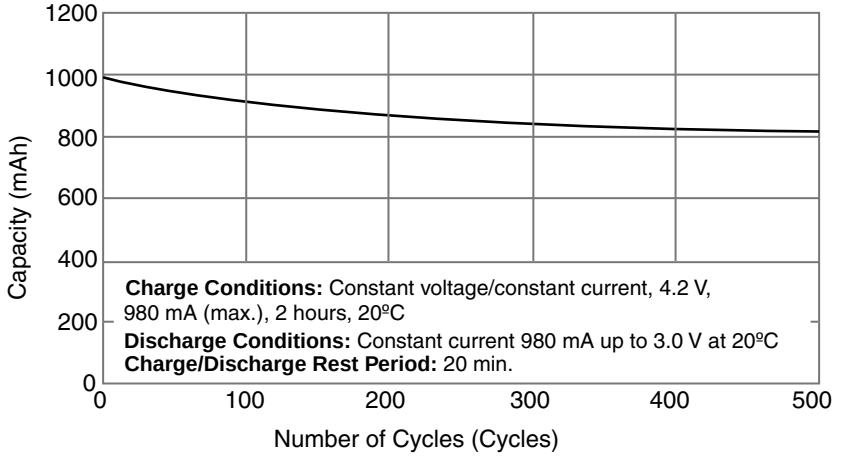


Fig. 9 Cycle life characteristics of Panasonic CGA633450A Li-ion batteries (standard capacity: 1035 mA hr). (From Ref.[23].)

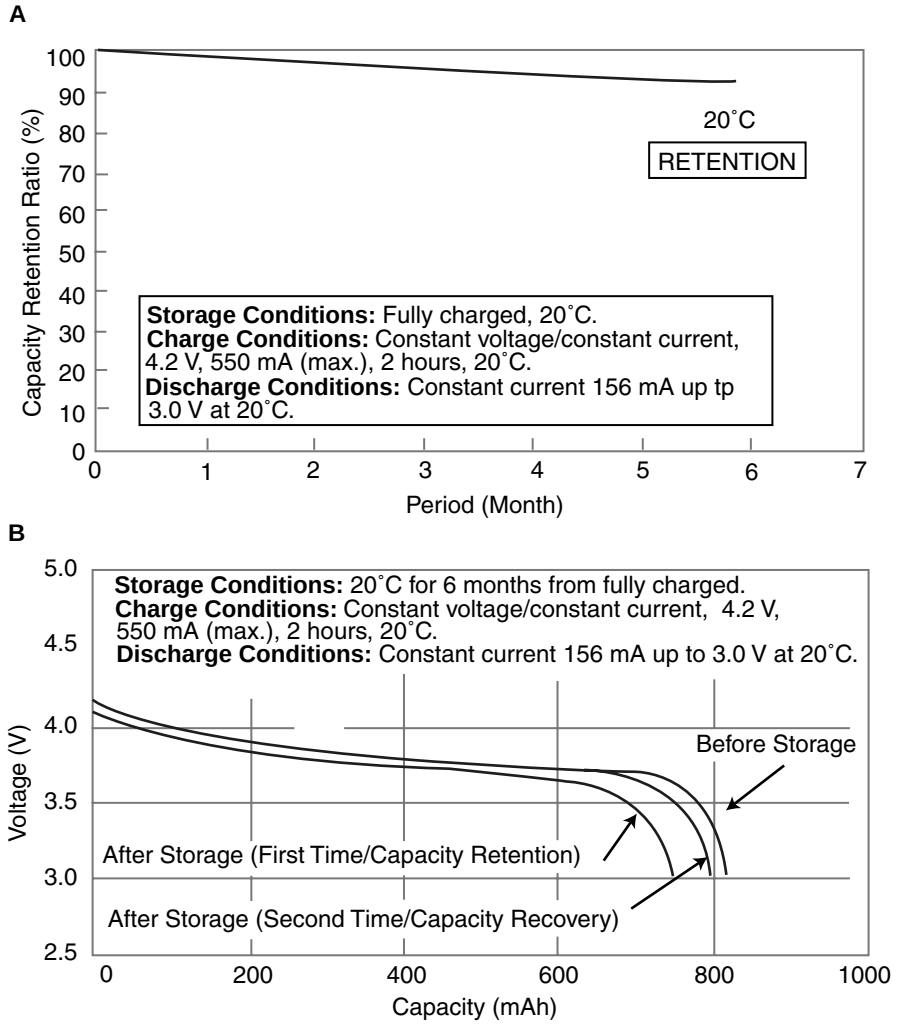


Fig. 10 Storage characteristics of Panasonic CGR17500 cylindrical batteries (standard capacity: 830 mA hr). (From Ref.[23].) *(View this art in color at www.dekker.com.)*

of the side reactions taking place during the storage. The time length of the storage and, more importantly, the temperature of the storage determine this ''irreversible capacity loss.''

Temperature Effect on Discharge and on Self-Discharge and Storage Life

Li-ion batteries can be discharged over a wide temperature range, typically −20°C to 60°C. The discharge capacity decreases with decreasing temperature, as shown in Fig. 11. The battery performance at low temperatures is to a large extent determined by the electrolyte employed. Some batteries are designed for applications at extremely low temperatures (e.g.,−40°C or lower), which require specially designed electrolytes to meet the requirements of the application.

The self-discharge rate of a Li-ion battery is usually lower compared to those of other rechargeable batteries. The capacity loss during storage is largely dependent on the storage temperature as well as the length of the storage time. A noticeably higher self-discharge rate is observed at an elevated temperature. It should be mentioned that the self-discharge rate is also impacted by the SOC of the battery, and the battery at the fully charged state usually shows the highest self-discharge rate.

Safety Issues Related to Li-Ion Batteries

Li-ion batteries use flammable organic electrolytes and work at a voltage range significantly higher than that of other aqueous rechargeable batteries, and safety becomes an important issue in their applications. Various inadequate operations such as overcharge, overdischarge, hard short circuit, abnormal high-temperature environment, impact, and others may cause the battery to catch fire or smoke. There are several industrial standards of safety tests for Li-ion batteries, such as Underwriter Laboratory (UL) 1640, International Electrical Committee (IEC), and others, which regulate Li-ion battery safety requirements.

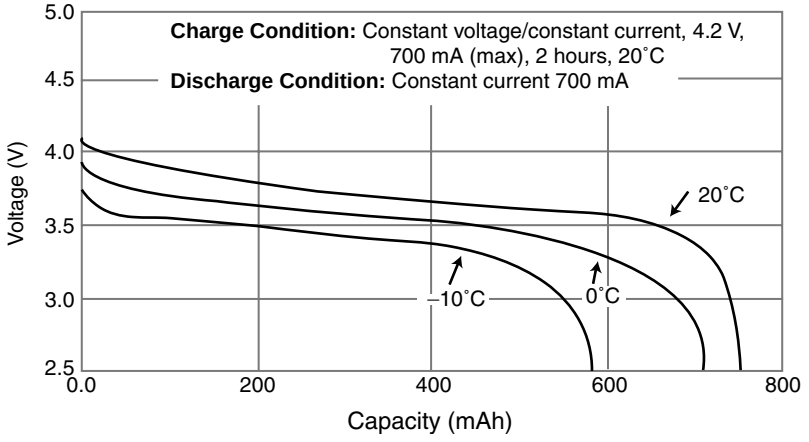


Fig. 11 Discharge characteristics at various temperatures of Panasonic CGA533048 batteries (nominal capacity: 750 mA hr). (From Ref.[23].)

These standards have been widely used by the Li-ion battery manufacturers and users to evaluate the battery safety characteristics to ensure battery safety in applications. Various safety devices including thermal control devices such as the positive temperature coefficient switches and electronic control devices such as various IC protection circuits have been successfully used for Li-ion batteries and battery packs. With an adequate combination of various safety devices, one can protect a Li-ion battery from overcharge, overdischarge, hard short circuit, impact, and other safety concerns throughout its applications.

Li-Ion Polymer Batteries

Commercial Li-ion batteries use liquid nonaqueous electrolytes. One of the new developments is to use a polymer electrolyte. The use of a polymer electrolyte provides several advantages over its liquid counterpart, such as avoiding any electrolyte leakage, making thin format batteries, implanting lightweight metal–plastic laminates as battery package material, etc. The most practical systems of LiP batteries to date are those with ''gelled polymer electrolyte.'' The gelled polymer electrolyte usually consists of a polymer matrix and a certain amount of liquid electrolyte absorbed in the polymer network. In general, there is no visible free liquid electrolyte in the gelled polymer electrolyte. Based on the type of polymers commonly used for the polymer matrix, the gelled polymer electrolytes can be categorized as polyethylene oxide (PEO) based, poly(vinylidene fluoride) /hexafluoropropylene (PVDF/HFP) based, poly(acrylonitrile) (PAN) based, and others. The PEO-based gelled polymer electrolyte received first attention in the development of the rechargeable batteries with metallic Li as the anode. This attention was then extended to other gelled polymers such as PAN, poly(methylmetacrylate), and PVDF. A common characteristic of these gelled polymer electrolytes is the direct trapping of the liquid electrolyte in the polymer network forming a gel-type membrane. In general, the gel-type membrane can be prepared by dissolving a selected quantity of a polymer and a Li salt in an adequate solvent or solvent-mixture, followed by homogenization, heat lamination, and other necessary processes. The obtained gel-electrolyte membrane has a conductivity around 10^{-3} S/cm or higher at room temperature.[24]

Recently, this type of gelled electrolyte has undergone various modifications in composition and processing, and has been successfully employed for manufacturing commercial LiP batteries.[25] One of the newer developments is to mix a known quantity of a monomer with a selected liquid electrolyte, which is then injected into a cell container with inserted jelly-roll, followed by polymerization under heating or radiation. The advantage of this type of LiP batteries is that it has basically the same manufacturing process as the liquid Li-ion batteries, in combination with an electrolyte gelling process. This process provides a practical means to manufacture LiP batteries on a large scale.

A different type of gel-electrolyte-based Li-ion batteries has been developed by Bellcore (now Telecordia Technologies).[26] The Bellcore plastic Li-ion battery technology has two major characteristics. First, this technology uses a special porous polymer separator that consists of a PVDF–HFP copolymer matrix mixed with a fine inorganic filler, such as the fumed SiO_2, and a liquid electrolyte (e.g., $LiPF_6$–EC–DMC) thoroughly absorbed in the polymer matrix. The PVDF–HFP copolymer matrix has a porous structure and its porosity is controlled by first adding a plasticizer (e.g., dibutyl phthalate) during the polymer/filler/solvent mixing process, and then extracting the plasticizer from the formed film to form a porous structure in a controlled way. The second major characteristic of the Bellcore technology is that the electrodes (anode and cathode, which both use PVDF–HFP as the binder) and separator are fused together by heat lamination or heat

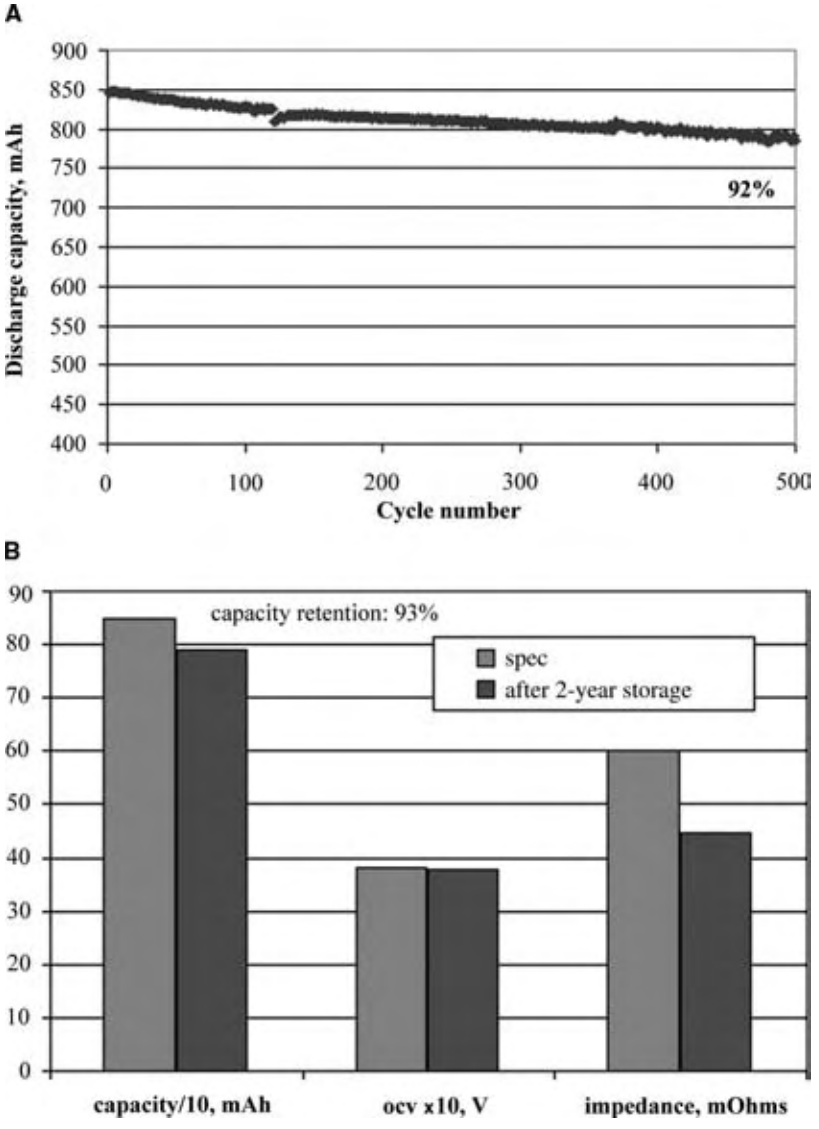


Fig. 12 (A) Cycle characteristics of NTK Powerdex pdx663448 LiP batteries (nominal capacity: 850 mA hr), charge: $c/2$, discharge: GSM, at room temperature. (B) Storage characteristics of NTK Powerdex pdx663448 LiP batteries (nominal capacity: 850 mA hr), storage conditions: half-charged state at room temperature for 2 yr. *(View this art in color at www.dekker.com.)*

pressing process. This special characteristic makes this technology especially advantageous to batteries with thin format and with stacking structure. Figs. 12A and 12B show the typical charge–discharge cycle performance and the long-term storage performance of the plastic Li-ion batteries based on the Bellcore technology (NTK Powerdex, Inc., unpublished data).

NEW DEVELOPMENTS IN THE LI-ION BATTERY FIELD

Development of New Cathode-Active Materials

Although a very good cathode-active material, $LiCoO_2$ has several disadvantages, such as high cost (due to the limited natural source of the cobalt element), toxicity of cobalt, and limited specific capacity, thermal stability at elevated temperatures, etc. Therefore, considerable efforts have been devoted to the development of new cathode-active materials other than $LiCoO_2$. Among several candidate cathode materials, the layered LiN_iO_2 has a higher specific capacity (e.g., $\sim$180 mA hr/g or higher) and a lower cost, but this material has some issues (such as lower thermal stability at elevated temperatures and the resulting safety concern, and ease of synthesis) that need to be resolved prior to its practical application.[27–29] To overcome these problems various modifications of the compound, such as $LiNi_{1-x}Co_xO_2$, $LiNi_{1-x-y}Co_xM_yO_2$ (M: other metals) and others, are being extensively studied.[30,31] Another promising candidate for the cathode-active material of a Li-ion cell is the Mn-based spinel ($LiMn_2O_4$) and doped-spinel ($Li_{1+x}Mn_{2-x-y}N_yO_4$) compounds.[9,32–36] The main advantages of the spinel compounds are its low cost, nontoxicity, and better safety performance at elevated temperatures. Two of the main problems to be

overcome for the spinel compounds are their relatively poorer cycle life at elevated temperatures and relatively lower specific capacity. Various doped spinel materials have been studied substantially.

Most recently, two new materials have attracted great interest in this development endeavor. One is the layered-structure $LiMn_{2-x}M_xO_2$ (M: other metals, e.g., Ni) compounds, the structure of which is stabilized with the dopant of other metals.[37,38] This group of materials has a higher specific capacity ($\sim$200 mA hr/g) and an operational voltage similar to that of $LiCoO_2$ and they appear to be promising alternates of the cathode-active material, although they are still at a developing stage at this juncture. Another material is the Olivien structured $LiFePO_4$ compound.[39] This material is relatively inexpensive, very stable at elevated temperatures, nontoxic, and has a specific capacity compatible to that of $LiCoO_2$ and an operational voltage about 200 mV lower than that of $LiCoO_2$. One of the major drawbacks of this material is its low electronic conductivity (10^{-8}–10^{-9} S/cm), which significantly impacts the rate capability for the Li-ion batteries with this compound as the cathode-active material. It has been reported that selected dopants can enhance its electronic conductivity by up to 10^8 order of magnitude.[40]

Development of New Anode-Active Material

Carbonaceous materials are widely used in commercial Li-ion batteries. To further improve the battery energy density and reduce the battery cost, the discovery of new anode-active materials with higher specific capacity and lower cost has become a goal in this field. Efforts in this area are mainly devoted to pursuing investigations in two directions: to find and improve the properties of new carbonaceous materials and to develop noncarbonaceous anode-active materials. For the former, several available carbonaceous materials, such as mesocarbon microbeads (MCMB), carbon fibers, are continuously improving in their specific capacity. New anodic materials, such as natural and synthetic graphite of lower cost, have been studied and special carbon materials such as carbon nanotubes are under study and development.[41–43] For the latter, the various alloys, such as silicon-based, tin-based, and intermetallic-based alloys, are of great interest to the field in terms of their feasibility as anode-active materials.[44–47] One of the advantages of the metal alloys as the anode-active material is their high specific capacity (e.g., 600 mA hr/g or more). However, two main drawbacks of these materials must be overcome before their practical usage. One is the large, irreversible capacity loss during the first charging–discharging step. Another is the large volume change induced by the lithium intercalation process, which may result in structure change and damage, and finally hurt the battery cycle life.

Development of Electrolytes

Research on electrolytes has focused on the following two areas: one is the development of new electrolyte additives and the other is the development of new Li salts.

The electrolyte additives can be categorized in three ways. The first additives are film-forming additives such as ethylene sulfite, vinylene carbonate, etc.[48,49] A common property of these additives in the electrolytes is that they are reduced at a relatively higher potential (compared with other alkyl carbonates solvents) on the carbon anode, which promotes formation of a more stable SEI film on the surface of the anode, and thus improves the tolerance of the graphite anode to the PC-based electrolytes and improves the battery performance, such as enhancing the cycle life, reducing the self-discharging rate especially at elevated temperatures. The second type of additives is used to reduce flammability of the electrolyte of Li-ion cells, or to provide nonflammable electrolytes for a Li-ion cell. Several inflammable additives such as trimethyl phosphate and other phosphorus (V) compounds have been studied in depth.[50–52] The third type of additives is used for a special purpose, e.g., as agents to protect battery overcharging.

The development of new Li salts to replace the commonly used $LiPF_6$, which is not stable with water and has low thermal stability at elevated temperatures, continues to draw attention. A newly developed Li salt, lithium bis(oxalato) borate[50,51] shows a number of advantages such as higher thermal stability at elevated temperatures, capability for improving compatibility between the graphite anode and PC-based electrolytes, and lower cost.[53,54]

CONCLUSIONS

As a newly commercialized high-energy battery system developed in the past decade, the Li-ion battery has various advantages over other commercial rechargeable batteries, which can be seen in its increasing number of applications. Continuous improvements in its performance and safety will extend its applications even wider. At present the majority of Li-ion batteries produced are relatively small in size and are focused largely on portable electronic devices applications. Substantial efforts have been made toward the development of Li-ion batteries, battery packs, and modules with larger size, higher power density, and safer performance to meet the requirement of other applications including electric vehicle power sources.

REFERENCES

1. Takeshita, H. Presentation on Rechargeable Batteries at Power 2002. Los Angeles, Sep 2002.
2. Whittingham, M.S. Insertion electrodes as SMART materials: the first 25 years and intercalation chemistry. Science **1976**, *192*, 1126.
3. Whittingham, M.S. The role of ternary phase in cathode reactions. J. Electrochem. Soc. **1976**, *123*, 315.
4. Whittingham, M.S. Chemistry of intercalation compounds: metal guests in chalcogenide hosts. Prog. Solid State Chem. **1978**, *12*, 41–99.
5. Whittingham, M.S. Abstract No. 112, extended abstract of 202nd ECS Meeting, Salt Lake City, Oct 2002.
6. Steele, B.C.H. *Fast Ion Transport in Solids: Solid-State Batteries and Devices*; Elsevier: New York, 1973; 103.
7. Mizushima, K.; Jones, P.C.; Wiseman, P.J.; Goodenough, J.B. Li_xCoO_2 ($0 < x < -1$): a new cathode material for batteries of high energy density. Mater. Res. Bull. **1980**, *15*, 783.
8. Mizushima, K.; Jones, P.C.; Wiseman, P.J.; Goodenough, J.B. Li_xCoO_2 ($0 < x \leq 1$): a new cathode material for batteries of high energy density. Solid State Ionics **1981**, *3–4*, 171.
9. Thackeray, M.M.; David, W.I.F.; Bruce, P.G.; Goodenough, J.B. Lithium insertion into manganese spinels. Mater. Res. Bull. **1983**, *18*, 461.
10. Goodenough, J.B. *Advances in Lithium-Ion Batteries*; van Sohalkwijk, W.A., Scrosati, B., Eds.; Kluwer Academic/Plenum Publishers: New York, 2002; 135(and references therein).
11. Murphy, D.W.; Disalvo, F.J.; Carides, J.N.; Waszczak, J.V. Topochemical reactions of rutile related structures with lithium. Mater. Res. Bull. **1978**, *13*, 1395.
12. Reimers, J.N.; Dahn, J.R. Electrochemical and in situ x-ray diffraction studies of lithium intercalation in Li_xCoO_2. J. Electrochem. Soc. **1992**, *139*, 2091.
13. Ogumi, Z.; Inada, M. *Advances in Lithium-Ion Batteries*; van Sohalkwijk, W.A., Scrosati, B., Eds.; Kluwer Academic/Plenum Publishers: New York, 2002; 79 (and references therein).
14. Peled, E. The electrochemical behavior of alkali and alkaline earth metals in nonaqueous battery systems—the solid electrolyte interphase model. J. Electrochem. Soc. **1979**, *126*, 2047.
15. Yamaki, J.-I. *Advances in Lithium-Ion Batteries*; van Sohalkwijk, W.A., Scrosati, B., Eds.; Kluwer Academic/Plenum Publishers: New York, 2002; 155 (and references therein).
16. Aurbach, D.; Markovisky, B.; Gamolsky, K.; Levi, E.; Ein-Eli, Y. On the correlation between surface chemistry and performance of graphite negative electrodes for Li-ion batteries. Electrochim. Acta **1999**, *45*, 67.
17. Aurbach, D. *Advances in Lithium-Ion Batteries*; van Sohalkwijk, W.A., Scrosati, B., Eds.; Kluwer Academic/Plenum Publishers: New York, 2002; 7 (and references therein).
18. Winter, M.; Wrodnigg, G.H.; Besenhard, J.O.; Biberacher, W.; Novak, P. Dilatometric investigations of graphite electrodes in nonaqueous lithium battery electrolytes. J. Electrochem. Soc. **2000**, *147*, 2427.
19. Chung, G.C.; Kim, H.J.; Jun, S.H.; Choi, J.W.; Kim, M.H. Origin of graphite exfoliation. An investigation of the important role of solvent cointercalation. J. Electrochem. Soc. **2000**, *147*, 4398.
20. Aurbach, D.; Levi, M.D.; Levi, E.; Schechter, A. Failure and stabilization mechanisms of graphite electrodes. J. Phys. Chem. B **1997**, *101*, 2195.
21. Dahn, J.R.; Fong, R.; Spoon, M. Suppression of staging in lithium-intercalated carbon by disorder in the host. J. Phys. Rev. B **1990**, *42*, 6424.
22. Shu, Z.X.; McMillan, R.S.; Murray, J.J. Electrochemical intercalation of lithium into graphite. J. Electrochem. Soc. **1993**, *140*, 922.
23. Panasonic Batteries, *Battery Technical Handbook*. CD-ROM; 2002.
24. Scrosati, B. *Advances in Lithium-Ion Batteries*; van Sohalkwijk, W.A., Scrosati, B., Eds.; Kluwer Academic/Plenum Publishers: New York, 2002; 251 (and references therein).
25. Nishi, Y. *Advances in Lithium-Ion Batteries*; van Sohalkwijk, W.A., Scrosati, B., Eds.; Kluwer Academic/Plenum Publishers: New York, 2002; 233 (and references therein).
26. Tarascon, J.M.; Gozdz, A.S.; Schmutz, C.; Schokpphi, F.K.; Warren, P.C. Performance of Bellcore's plastic rechargeable Li-ion batteries. Solid State Ionics **1996**, *49*, 86–88.
27. Thomas, M.G.S.R.; David, W.J.F.; Goodenough, J.B.; Groves, P. Synthesis and structural characterization of the normal spinel $Li[Ni_2]O_4$. Mater. Res. Bull. **1985**, *20*, 1137.
28. Dahn, J.R.; von Sacken, U.; Jurzow, M.W.; Al-Janaby, H. Rechargeable $LiNiO_2$/carbon cells. J. Electrochem. Soc. **1991**, *138*, 2207.
29. Bruce, P.G.; Lisowska-Oleksiak, A.; Saidi, M.Y.; Vincent, C.A. Vacancy diffusion in the intercalation electrode $Li_{1-x}NiO_2$. Solid State Ionics **1992**, *57*, 353.
30. Delmas, C.; Saadune, I.; Auradou, H.; Menetrier, M.; Hagenmuller, P. *Solid State Ionics: Materials and Applications*; Chowdari, B.V., et al., Eds.; World Scientific: Singapore, 1992; 255.
31. Sun, X.; Yang, X.Q.; McBreen, J.; Gao, Y.; Yakovleva, M.V.; Xing, X.K.; Darous, M.L.

New phases and phase transitions observed in over-charged states of $LiCoO_2$-based cathode materials. J. Power Sources **2001**, *274*, 97–98.
32. Hunter, J.C. Preparation of a new crystal form of manganese dioxide: λ-MnO_2. J. Solid State Chem. **1981**, *39*, 142.
33. Yang, X.Q.; Sun, X.; Lee, S.J.; McBreen, J.; Mukerjee, S.; Darous, M.L.; Xing, X.K. In situ synchrotron x-ray diffraction studies of the phase transitions in $Li_xMn_2O_4$ cathode materials. Electrochem. Solid-State Lett. **1999**, *2*, 157.
34. Mukerjee, S.; Thurston, T.R.; Jisrawi, N.M.; Yang, X.Q.; McBreen, J.; Darous, M.L.; Xing, X.K. Structural evolution of $Li_xMn_2O_4$ in lithium-ion battery cells measured in situ using synchrotron x-ray diffraction techniques. J. Electrochem. Soc. **1998**, *145*, 466.
35. Xia, Y.; Kumada, N.; Yoshio, M. Enhancing the elevated temperature performance of $Li/LiMn_2O_4$ cells by reducing $LiMn_2O_4$ surface area. J. Power Sources **2000**, *90*, 135.
36. Amatucci, G.G.; DePasquier, A.; Blyr, A.; Zhang, T.; Tarascon, J.M. The elevated temperature performance of the $LiMn_2O_4/C$ system: failure and solutions. Electrochim. Acta **1999**, *45*, 255.
37. Ohzuku, T.; Makimura, Y. Layered lithium insertion material of $LiNi_{0.5}Mn_{0.5}O_2$: a possible alterative to $LiCoO_2$ for advanced lithium-ion batteries. Chem. Lett. **2001**, *30*, 744.
38. Lu, Z.; MacNeil, D.D.; Dahn, J.R. Layered $Li[Ni_xCo_{1-2x}Mn_x]O_2$ cathode materials for lithium-ion batteries. Electrochem. Solid State Lett. **2001**, *4*, A200.
39. Padhi, A.K.; Nanjundaswamy, K.S.; Goodenough, J.B. Phospho-olivines as positive-electrode materials for rechargeable lithium batteries. J. Electrochem. Soc. **1997**, *144*, 1188.
40. Chung, S.-Y.; Bloking, J.T.; Chiang, Y.-M. Nature materials electronically conductive phospho-olivines as lithium storage electrodes. Nat. Mater. **2002**, *1*, 123.
41. Fujimoto, H.; Chinnasamy, N.; Mabuchi, A.; Kasuh, T. Abstract No. 5, 11th IMLB, Monterey, CA, Jun 2002.
42. Yang, Z.-H.; Wu, H.-Q. The electrochemical impedance measurements of carbon nanotubes. Chem. Phys. Lett. **2001**, *343*, 235.
43. Yang, Z.-H.; Wu, H.-Q. Electrochemical intercalation of lithium into carbon nanotubes. Solid State Ionics **2001**, *143/2*, 173.
44. Umeno, T.; Fukuda, K.; Wang, H.; Dimov, N.; Iwao, T.; Yoshio, M. Novel anode material for lithium-ion batteries: carbon-coated silicon prepared by thermal vapor decomposition. Chem. Lett. **2001**, *30*, 1186.
45. Dahn, J.R.; Beattie, S.D. Abstract No. 35, 202nd ECS Meeting, Salt Lake City, UT, Oct 2002.
46. Thackeray, M.M.; Vaughey, J.T.; Johnson, C.S.; Benedek, R.; Fransson, L.M.L.; Edstrom, K. Abstract No. 6, 202nd ECS Meeting, Salt Lake City, UT, Oct 2002.
47. Wachtler, M.; Besenhard, J.O.; Winter, M. Tin and tin-based intermetallics as new anode materials for lithium-ion cells. J. Power Sources **2001**, *94*, 189.
48. Wrodnigg, G.H.; Besenhard, J.O.; Winter, M. Ethylene sulfite as electrolyte additive for lithium-ion cells with graphitic anodes. J. Electrochem. Soc. **1999**, *146*, 470.
49. Simon, B.; Boeuve, J.P. Rechargeable lithium electrochemical cell. U.S. Patent 5,626,981, May 6,1997.
50. Wang, X.; Yasukawa, E.; Kasuya, S. Nonflammable trimethyl phosphate solvent-containing electrolytes for lithium-ion batteries: I. Fundamental properties. J. Electrochem. Soc. **2001**, *148*, A1058.
51. Lee, C.W.; Venkatachalapathy, R.; Prakash, J. A novel flame-retardant additive for lithium batteries. Electrochem. Solid State Lett. **2001**, *3*, 63.
52. Xu, K.; Ding, M.S.; Zhang, S.; Allen, J.; Jow, T.R. Abstract No. 204, 202nd ECS Meeting, Salt Lake City, UT, Oct 2002.
53. Lischka, U.; Wietelmann, U.; Wegner, M. Ger. DE. 19829030C1, 1999.
54. Xu, K.; Zhang, S.; Allen, J.; Jow, T.R. Abstract No. 202, 202nd ECS Meeting, Salt Lake City, UT, Oct 2002.

Loss Prevention in Chemical Processing

Joel M. Haight
Pennsylvania State University, University Park, Pennsylvania, U.S.A.

INTRODUCTION

As the chemical process industry evolved and advanced over the last 50 years, it could be said that probably more attention was paid to making the process work than to concerns that might be manifested if it failed. Accidents occur in nearly every facet of our lives. It is no surprise that the chemical process industry has also had its accident difficulties. Some of the watershed incidents that have been the impetus for much loss prevention improvement made over the last 15 years are discussed here. While many in the loss prevention field believe that we should not focus on incidents because they are in the past (and we need to be proactive), much can be learned from these incidents. In fact, it might be said that not only did we learn from these watershed incidents, but they also provided some of the motivation for change in our loss prevention systems. George Washington is credited with saying "We ought not to look back, unless it is to derive useful lessons from past errors, and for the purpose of profiting by dear-bought experience."[1]

BACKGROUND

Some of the historic process incidents are well documented in the literature; however, their impact on today's approaches to loss prevention is so profound that this entry would seem to be lacking if it did not, at least, summarize them. One of the earliest incidents occurred at Nypro Ltd., a caprolactam plant in Flixborough, England, in June 1974. It resulted in 28 fatalities and 89 injuries inside and outside of the plant, and damage to more than 1800 homes and 167 shops and businesses. A deflagration occurred after the failure of a temporary bellows joint and the line between two reactors in a six-reactor series of the operation's cyclohexane oxidation plant. It was installed when one of the reactors was removed for maintenance. When this temporary connection failed, about 35 tons of cyclohexane escaped and subsequently ignited. There were no calculations to determine if this temporary 20 in. line and bellows connection could safely handle the process conditions (120 psig and 145°C) and it failed. This incident highlights the need for a management of change process.[2]

A second incident of note occurred about two years later in Seveso, Italy. Approximately 2 kg of dioxin was released through a reactor's pressure relief valve. It was reported that the operating crew had gone home for the weekend without completing the last step in this batch process, which was to add quench water. Without the water, the unattended reaction continued and the vessel became overpressured. The pressure relief valve did its job and protected the reactor; however, 2 kg of 2,3,7,8-tetrachlorodibenzopara-dioxin was discharged to the atmosphere. Wind carried the contaminant over the countryside where people, animals, and plants were exposed to the dioxin. This incident highlighted the importance of standard operating procedures, training in unknown hazards, and emergency response.[2]

In 1984, one of the most notable process incidents occurred in Bhopal, India, at a plant owned by Union Carbide. On December 3, 1984, methyl isocyanate (MIC) was vented to the atmosphere from a vent gas scrubber after a run away reaction overwhelmed the scrubber. It is reported that more than 2500 people died and 20,000 people were injured as a result of exposure to MIC. The subsequent investigation showed that the run away reaction occurred in the MIC tanks after water was reportedly intentionally added to the tank. The safety systems, a pressure relief valve, a vent gas scrubber, and a flare were reportedly poorly maintained. The relief valve worked to vent the pressure, but the scrubber was overwhelmed, and the flare was down for maintenance. This incident highlighted several loss prevention areas that needed improvement. The first was mechanical integrity. No strong preventive maintenance effort appeared to be in place. The second was emergency preparedness and response. It appears that while the plant was built 1.5 miles away from the community, zoning problems allowed the community to expand to the plant limits. It does not appear that much effort was made to work with the community on communication/notification or evacuation needs.[2]

A fourth incident occurred in 1989 in Pasadena, Texas. An explosion occurred at a high density polyethylene plant, when 85,000 pounds of a flammable mixture of ethylene, isobutene, hexane, and hydrogen was released and ignited. The consequences included 23 fatalities, 314 injuries, and over $715 million in

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007844

capital losses. The investigation showed that a crew, attempting to unplug a settling leg in the polyethylene product take off system, did not follow proper procedures for carrying out the operation safely. This incident highlighted several deficiencies in the system's design, and given that there was no process hazards analysis, they were ignored. Also, the permit-to-work system was ineffective, there were no permanent combustible gas detection devices and alarms available, and building spacing was questionable in terms of both absolute distance and high occupancy (the control room).[3]

There have been several other incidents in the process industry such as an LPG explosion in Mexico City in 1984, a North Sea offshore oil platform explosion in 1988, and a fuel tank rupture in Pittsburgh in 1988. More recently, a 1999 naphtha explosion in California, a 2001 sulfuric acid tank explosion in Delaware, a 2002 hydrogen sulfide leak in Alabama, and a 2003 distillation tower explosion in Ohio all resulting in lower fatality numbers than the earlier incidents.[4]

These incidents have taught us much and, especially, the earlier ones seemed to have provided the motivation for the industry and government to begin to aggressively address the catastrophic loss potential associated with the chemical process industry. They helped spark development of several consensus and regulatory standards that have since become so much a way of life in the chemical process industry.

REGULATORY STANDARDS, CONSENSUS STANDARDS, AND VOLUNTARY PROGRAMS

Much of the impact on loss prevention in the last 12 years has come from the Occupational Safety and Health Administration's Process Safety Management of Highly Hazardous Chemicals regulation (PSM). This regulation, 29 CFR 1910.119, took effect in May 1992 and it was joined in 1996 by the Environmental Protection Agency's Risk Management Program (RMP) (40 CFR Subpart B). These performance standards are similar and had industry input through the former Chemical Manufacturers Association [now American Chemistry Council (ACC)] and the American Petroleum Institute (API), among others. These two standards assist the industry in preventing (and/or minimizing the consequences or likelihood of occurrence) catastrophic incidents like those that occurred in India, U.K., Mexico, Italy, U.S.A., and other countries.[5,6]

The Chemical Safety Board is a relatively recent addition to the loss prevention effort for the U.S. chemical process industry. While it was authorized by the Clean Air Act in 1990, it became operational in January 1998. It is an independent federal agency charged with the mission of preventing industrial chemical accidents. This is done through the accident investigation process: by identifying root causes and issuing recommendations to correct the causes. Their work is well publicized and when they issue an investigation report, it is made available to the public. Through this communication, they promote improvements in the loss prevention effort, and while they are not a regulatory body, their recommendations should be taken seriously.[4]

Several consensus standards and voluntary programs help to support process industry loss prevention efforts. The Responsible Care Program of the ACC helped get companies to address protecting not only their workers, but also their neighbors, their customers, and the rest of the community. This is promoted through practices such as Product Stewardship and Emergency Preparedness and Response. OSHA's Voluntary Protection Program (VPP) promotes the use of a safety management system and employee involvement in the management of safety and loss prevention.[5]

Contributions from consensus and industry organizations to the loss prevention effort are significant. The API developed Recommended Practice 750 Process Hazards Management and Recommended Practice 752 covering facility siting. The American Institute of Chemical Engineers' (AIChE) Center for Chemical Process Safety (CCPS) has developed a complete series of textbooks that address the same major elements of OSHA's PSM standard. They sponsor conferences, publish proceedings, develop and promote seminars, and provide industry with a process safety and loss prevention resource.

REGULATORY AND MANAGEMENT PRACTICES FOR PREVENTING LOSS

While the two main management systems that have driven loss prevention in the process industry are regulations, the way the industry has implemented them since 1992 represents an effective way to manage loss prevention and still comply with the regulations. The regulations are OSHA's PSM and EPA's RMP regulations. While they took effect at different times, they are similar in what they require of industry. Both address many aspects of how business operates. It appears to this author that over the last 12 years, industry representatives have made complying with these regulations more the way business gets done rather than an additional task for the sake of compliance.

OSHA's Process Safety Management of Highly Hazardous Chemicals

The OSHA regulation, 29 CFR 1910.119, addresses 14 elements. Through its list of highly hazardous

chemicals and threshold quantities, it first allows industry to determine if the regulation applies (if they handle a chemical from the list in quantities greater than the threshold quantity). This is determined by following the Applicability section [29 CFR 1910.119 (a)]. This covers many of the process industries as well as other industries, from food to power.[5]

The first major element is *Employee Participation* [29 CFR 1910.119 (c)]. This requires that affected employees (those who could be affected by an incident) not only have access to pertinent Process Safety Management information, but also be allowed to provide input to its development and use. It is expected that affected employees will be trained in the standard's content. Many companies not only welcome employees' input, but also involve them in the development and management of the individual elements. It is expected that all employee participation in PSM implementation will be documented.[5]

The second main element in OSHA's Process Safety Management standard is *Process Safety Information* (PSI) [29 CFR 1910.119 (d)]. This requires the covered operation to produce, maintain, and use information associated with the chemicals used, from raw materials to final products, process technology information such as the process design and batch recipes and covered equipment information. The types of information expected include material safety data sheets (MSDS), piping and instrument diagrams (P&ID), block flow diagrams, relief valve design and sizing information, ventilation system information, data sheets on pumps (pump curve data, etc.), compressors, pressure vessel ratings, and safe operating and design limits for equipment. While there is much information to be covered, space limitation of this entry does not allow complete coverage of PSI. However, the list above should give the reader a general idea of the required information.[5]

Not only must PSI be maintained up-to-date, but it must also be available to employees who work with or on the covered equipment. Many companies use electronic media. Only one copy has to be maintained, and all employees who have access to the organization's computer system should have access to the PSI. In some cases, a real-time electronic P&ID is available in which employees can access a particular piece of equipment on the P&ID and, from there, can access the details of that equipment, such as the MSDS, vessel thickness, the pump curve data, pressure relief valve sizing, design basis, etc. The regulation does not require electronic information; hence hard copies maintained in the control rooms or the supervisor's office are considered acceptable. Irrespective of the format, it still must be maintained up-to-date.[5]

The third major element is *Process Hazards Analysis* (PHA) [29 CFR 1910.119 (e)]. This element requires performing one of several types of analyses to identify process hazards that could lead to a catastrophic incident. The techniques one can use are What if studies, What if Checklists studies, Hazard and Operability (HAZOP) studies, Failure Mode and Effects Analysis (FMEA), and Fault Tree Analysis (FTA), among others. An initial PHA must be performed on the covered process, and it must be revalidated every five years after the initial PHA. In these studies, analysts are expected to identify failure potential, determine the potential consequences of that failure or hazard, identify existing safety systems, determine the risk associated with each hazard, and recommend corrective action if warranted. This corrective action is evaluated by the management team, resolved, and then implemented as deemed necessary. These studies are done by multidisciplinary teams. In each case, one of the representatives on the team must be an expert in the process itself and at least one member (preferably the leader) should be trained in the analytical method used.[5]

The next element is *Operating Procedures* [29 CFR 1910.119 (f)]. This requires that the operation of a covered process be directed by written operating procedures that are accessible to and used by the employees who operate that process. They must be up-to-date and must cover normal operations, start up, emergency shutdown, temporary operations, and start up after a maintenance turnaround. They must cover safe operating limits for the critical process parameters as well as how to prevent the process from operating outside those limits (and what to do if the limits are exceeded). These procedures must be certified annually as accurate and up-to-date, have the input of the operators, and they must be trained in their use.[5]

The *Training* element [29 CFR 1910.119 (g)] requires that anyone expected to operate a covered process must be trained in its safe operation. They must receive initial training when assigned and then receive refresher training at least every three years following initial training. The training must include the use of the Operating Procedure and trainees must show that they have understood the training. Any changes to the process covered by the Management of Change element (discussed later) must be incorporated into the training material, and operators must receive training in these changes.[5]

Contractor Safety programs have been in existence for many years, but the PSM Standard required that the program become formalized and more far-reaching than traditional contractor safety programs. The *Contractors* [29 CFR 1910.119 (h)] element requires that those considering the use of contractors must partially base their selection of a particular contractor on their safety performance. Covered process owners, once they select a contractor, must ensure that not only are the employees of the contractor qualified to do the

work for which they were hired, but also they understand the process and the hazards where they will be working. Owners are also required to regularly audit the performance of their contractors and keep incident records.[5]

If a covered process is started up for the first time or undergoes a maintenance turnaround, it must receive a *Pre-Start up Safety Review* (PSSR) [29 CFR 1910.119 (i)]. This element is most often integrated with the Management of Change element (discussed later) and is intended to help ensure that anytime there is a change to the process, it is reviewed and documented by representatives from appropriate disciplines. During this process, the people performing the PSSR check to ensure that necessary changes to operating procedures have been made and that operators have been trained as needed. The PSSR reviewer checks to ensure that required updates to the PSI have been made, that corrective action recommendations made by a PHA team have been resolved or completed, and that all required design specifications have been followed. These checks are to be made prior to start up, and any noted deficiencies that may jeopardize a safe start up are required to be followed up until resolution and/or implementation.[5]

The *Mechanical Integrity* [29 CFR 1910.119 (j)] element addresses requirements to maintain existing equipment in safe and working order. This requires identification of critical equipment and implementation of inspection, testing, and preventive maintenance programs to ensure that not only was the equipment built properly and to specifications, but also that it is maintained safe throughout its life. The type and frequency of inspections, testing, and preventive maintenance routines are not specified, but one can rely on manufacturer recommendations and internal company expert input to develop and establish these routines. The standard also addresses the need for all employees, whether contract or regular, to be trained to perform the intended maintenance, inspections, and testing. In addition to the maintenance aspect of mechanical integrity, the standard also requires that a quality assurance program be established to ensure that all new and existing equipment, spare parts, lubricants, and other consumable maintenance materials meet required specifications for dimensions, metallurgy or materials of construction, composition, strength, etc. Deficiencies found during inspection and testing routines must be managed.[5]

The *Hot Work Permit* [29 CFR 1910.119 (k)] element addresses a permit-to-work system for hot work in flammable areas and energy isolation for work on covered equipment. This element supports existing regulations, specifically 29 CFR 1910.252, on hot work. Hot work permits are important for any work that increases the risk of contact between a flammable vapor and an ignition source. It is a system in which the permits convey safety-related information (protective equipment requirements, special atmospheric testing, approval signatures, etc.) and involve transmission and posting that forces those involved in the task to, at least, review the permit if not approve and sign it.[5]

A number of incidents in the past occurred because a change made to a covered process was not adequately reviewed for unintended consequences. The case in Flixborough is an example. The element that governs this is *Management of Change* [29 CFR 1910.119 (l)]. This element requires evaluation of each change to be sure that the appropriate type and level of review has been done to ensure that it is safe and will not produce undesired consequences. It also requires that any information that is different after the change is revised and that employees are informed of the change or trained to use the new system. This involves input from many disciplines and extensive coordination by the change coordinator. It applies to changes in equipment, raw materials, intermediates and products, and methods or procedures and can even apply to changes in consumable supplies such as lubricants, gaskets, O-rings, etc.[5]

Even with all PSM elements in place, incidents still occur. When they do, the *Incident Investigation* [29 CFR 1910.119 (m)] element applies. An incident is a catastrophic loss event that occurred (or could have occurred) such as a fire, explosion, or toxic release in which people are injured or killed, and/or equipment, facilities, and/or the community are damaged. An investigation must begin within 48 hr of the incident's occurrence. The investigation must be done by a multidisciplinary team with the appropriate expertise to understand the process involved in the incident. This could involve operators, process and/or design engineers, instrumentation and/or electrical experts, materials experts, safety and/or fire experts, and those familiar with investigation techniques. The team determines the causes and develops recommended corrective actions to ensure that the incident does not recur. The team puts together a report of the findings, conclusions, and recommended corrective actions, along with the details of the incident. This report is to be made available to all employees who may be affected by this or a similar incident, or who may have a vested interest in the outcome of the corrective actions.[5]

Should an incident occur, there may be a need to interact with the outside community. If there is a potential for the incident to impact the community, the *Emergency Preparedness and Response* [29 CFR 1910.119 (n)] element will be important. This element relies on two other sections of the OSHA regulations to define compliance: 29 CFR 1910.38 (Emergency Action Plans) and 29 CFR 1910.120 (Hazardous Waste and Emergency Response). In addition, there

must be provisions in this element of the program for addressing small releases. The basic objective is to help ensure that everyone is prepared in the areas of alarms and alarm notification, actual response to the incident (fire fighters, toxic vapor monitoring specialists, etc.), notification of the community response organizations, evacuation plans and procedures, damage assessment, and clean up after the incident. The program must be written, and training and practice must be carried out for all employees involved in or affected by the emergency. Practice drills should be critiqued to ensure that weaknesses in the response system can be identified and corrected before a real incident occurs.[5]

Once all the elements of the Process Safety Management system are in place and are being implemented, the *Compliance Audit* [29 CFR 1910.119 (o)] element provides a means to evaluate the program's compliance level. Every three years, the program must be audited to ensure that it is being implemented as per the regulation. The audits can be done internally, but some companies hire outside consultants to get the viewpoint of an unbiased expert.[5]

There is also a *Trade Secret* [29 CFR 1910.119 (p)] element to help companies protect competitive advantage due to their chemical formulations. However, while this provision does offer some protection in that regard, the representatives of a covered process must still disclose their PSI for any hazards associated with that process or process materials.[5]

Environmental Protection Agency's Risk Management Program

The Environmental Protection Agency's Risk Management Program (RMP) Regulation (40 CFR Subpart B), which took effect in June 1996, is similar to OSHA's PSM standard. Its intent is also to prevent accidental releases of toxic or flammable materials. One difference is that it is directed at preventing off-site impact on the public, as opposed to OSHA's goal of protecting the on-site worker. The RMP regulation applies to companies using or handling materials from a list similar to OSHA's highly hazardous chemicals list above certain threshold quantities. The major elements of RMP include Hazard Assessment, Prevention Program, Emergency Response Program, and documentation maintained on-site and submitted to federal, state, and local environmental authorities.[3]

The *Hazard Assessment* element is a consequence analysis applied to a worst-case event (entire contents of a covered process released in 10 min. under the worst atmospheric conditions) as well as to an alternative case (generally more likely to occur, but usually with less severe consequences). The *Prevention Program* has 11 of the same elements as the PSM regulation, but differs in that it has no Employee Participation, Contractor Safety, Hot Work Permit, or Trade Secret elements, while the PSM standard does not have a Risk Assessment element found in the RMP regulation. The *Emergency Response Element* includes provisions for notifying, involving, and working with local emergency response organizations. While much of the PSM information can become public information if an incident occurs, it is not routinely and automatically made available to the public. In fact, since September 11, 2001, anyone requesting access to RMP-related information must apply in person and show just cause for accessing a company's RMP-related information. The results of the Hazard Assessments are available on EPA and state environmental agency websites.[3]

Implementation of these two government regulations accounts for much of the loss prevention effort in the chemical process industry. While these are regulatory, they form the basis of the management systems used by much of the industry to manage the catastrophic risk associated with process conditions that may increase the potential for loss of containment. While these management activities are thought to be effective, there are also a number of engineered approaches that are discussed.

ENGINEERING APPROACHES TO LOSS PREVENTION

There are many engineered systems that help to reduce the potential for catastrophic loss of containment. These engineered approaches are generic in nature in that one or more can be found in and applied to or on almost all process industry equipment. These systems or design specifications and practices include, but are not limited to, pressure relief systems and flares, automated control systems including detection and alarm systems, piping and equipment material specifications, purging and ventilation systems, fire suppression systems, explosion venting systems, positive pressure buildings, and equipment layout and siting practices.

Automated Control Systems (Including Detection and Alarms)

Many operations in the process industry are carried out by automated control systems. While technology is increasing and improving every day, automated control systems are not new to this industry. Computers have long been used to control pressures, temperatures, flows, levels, process stream concentrations, etc. and, to a great extent, these automated systems perform the control function more efficiently, effectively, and safely than humans. For this reason, automated

systems are relied upon extensively for their contribution to the loss prevention effort.[7]

These systems are designed to sense when process conditions are going outside safe limits. For example, if during a tank filling operation, the level goes above a specified limit, a level indicator/controller device (usually a float valve) generates a signal (based on the valve's position) that a transmitter then sends to an inlet control valve to close, to an outlet valve to open, to a discharge pump to start, to a charge pump to stop, etc. This prevents a spill. If a Temperature Indicator Controller senses an increased temperature, for example, a signal may be sent to cooling water or quench nitrogen valves prompting them to open and cool the process, thus preventing a runaway reaction. If there is a leak, a detection system measures the concentration of the material in the air and at a prescribed concentration, sends a signal to an alarm enunciator or control room screen alerting operators to the problem. It can also send a signal to a block valve that can isolate the process stream from the leak, minimizing the risk of overexposure to the leaking chemical.[7]

In large plant upsets, there is a possibility that the control operators are besieged with an alarm flood or shower. In one case in the author's experience, an upset was determined to have resulted in approximately 300 alarm signals in 5 min. Operators are unable to address all of these. A management strategy and technique that can help address this problem during an alarm flood and that seems to be gaining popularity is the practice called abnormal situation management (ASM). An ASM system can be built and integrated into a normal process control system. Control system engineers attempt to design these systems to take the actions that an experienced operator would to head off any equipment damage or incident. These systems contain an alarm priority system intended to suppress low priority alarms while the operators acknowledge and address the higher priority alarms. This is intended to reduce the information overload on the operator while still minimizing the risk of catastrophic incident because of an operator addressing a low priority alarm when a critical one gets lost in the flood.[8]

Automated system contribution to loss prevention is high; however, one must understand the limits of this equipment, its power source, its failure mode (open, closed, on, off, etc.), whether an alarm is registered at the point of the upset (possibly in a location where no one would see or hear it), or at a computer screen occupied 24 hr of the day.[7]

Pressure Relief Systems

Many vessels within the process industry operate at pressures above atmospheric pressure. Each vessel is rated for a certain range of pressures. If this pressure is exceeded, the vessel can potentially rupture, releasing its contents to the environment. A typical way to address this increased pressure potential, whether it is from an internal process-driven increase such as a runaway, exothermic reaction, or an external fire, is to install a pressure relief device. These devices are generally used to protect personnel from exposure to the contents of the vessel, minimize loss of product or other material, and/or prevent damage to the equipment.[3] This pressure relief device can be a spring-operated pressure relief valve or a rupture disk. The choice of which to use and when is beyond the scope of this entry; however, Crowl and Louvar[3] provide an excellent summary of the design issues and advantages and disadvantages.

Pressure relief devices must be properly sized (capacity), and discharge location is critical. These devices are generally sized for the most likely pressure increasing event, e.g., external fire or internal process upset such as an uncontrolled exothermic reaction. The engineer must calculate the temperature and pressure increases associated with the event as well as the expected release volume. He or she must also account for pressure drops across the relief device as well as for friction losses in the lines. A decision must also be made about whether to discharge to the atmosphere or to a closed system that includes a scrubber, a flare, or even as simple as a water tank, such as is sometimes used for venting anhydrous ammonia. Some of these design decision issues are addressed in more detail by Crowl and Louvar.[3]

Piping Specifications

Piping is similar to pressure vessels in that it must contain the pressure and the toxic or flammable material. A difference with piping is that the process stream is almost always flowing and this introduces a dynamic load on the system that must be considered when establishing piping specifications. Piping can rupture, corrode through, leak at a weld, leak at a flange or other connection point, and can even leak through its walls (if the molecular size of the process stream is small, e.g. hydrogen). For this reason, piping for each particular application must be designed to meet the stresses to which it will be exposed. This is most often done through the use of piping specifications. These specifications usually include an identifying number, a description of the safe range of process stream contents and conditions (maximum and minimum temperatures, maximum and minimum pressures, flow rates, etc.), and materials of construction (to minimize the risk of corrosion attack). It also contains information about required wall thickness, type, material and

size of gaskets, O-rings and fittings, and pipe joining technique, e.g., butt welded or flanged. The specifications also provide information about the types of valves to be used and any other types of connections or appurtenances (thermowells or injection points, etc.). All of these requirements, if adhered to, theoretically will help to prevent loss of containment of toxic and/or flammable materials.[9]

Many piping specifications are partially established by consensus standards such as the American National Standards Institute (ANSI), ANSI B31.3, Chemical Plant and Petroleum Piping Code, the American Petroleum Institute's (API) API 5L Specifications for Lined Pipe, or the National Fire Protection Association's (NFPA) NFPA 30 Flammable and Combustible Liquids Code. These can form the foundation for specifications; however, individual plant operating conditions and history should be considered when establishing piping specifications.[9]

Inerting, Purging, and Ventilation

Often, vessels and other process containers must be opened for inspection or maintenance. The container must be drained and cleaned, but often, residual toxic or flammable vapors can remain. Occasionally, work must be performed in this environment, while the flammable concentrations remain (e.g., some catalysts cannot be water washed or steamed). Because of this, protection methods must be applied to help minimize human exposure, and techniques such as inerting, purging, and ventilating are often used. If all the flammable material cannot be removed, a vessel is inerted with a material such as nitrogen or carbon dioxide to reduce oxygen concentrations to a point where ignition is not possible. If workers are expected to enter these vessels, a confined space entry permitting process must be used and additional attention must be paid to respiratory protection in this oxygen deficient environment. Hot nitrogen purges are sometimes applied to remove heavier hydrocarbons from catalysts or vessel internals. Other general types of purges such as vacuum and pressure purges are used to remove toxic or flammables from a work environment as well. When purging is used, one must decide what to do with the purged material, as it may not be safe to discharge it to the atmosphere. It may need to be routed through a collection medium such as a bag house, electrostatic precipitator, incinerator, or flare system.[3]

General dilution as well as local exhaust ventilation is also used to remove contaminants (toxic or flammable) from a work area. This is done more to protect individuals or small groups of employees than the public. Ventilation systems are sized based on space dimensions, expected contaminant concentration (emergency release or continuous emissions), physical and chemical properties of the contaminant, and extent and time of the occupancy. In each case, similar to purging, ventilation system discharge must be appropriately managed.[10,11]

Explosion Venting and Explosion Proof Equipment

Process operating areas are often designated as explosion proof (XP) or non-XP. An XP designation means that flammable vapors could be present under upset or other intermittent conditions. In these areas, many equipment-related measures must be taken, such as elimination or factory sealing of ignition sources, installing explosion proof housing around electrical equipment (housing can withstand an internal explosion), or installing explosion venting capability. Explosion venting does not prevent an explosion; it only minimizes potential damage by venting the overpressure. This could be in the form of a weak seam welded or riveted panel on a bag house filter or a weak seam weld on a tank roof-to-wall seam. In each case, the panel or roof plate opens to vent the overpressure and protects the rest of the equipment.[3]

When electrically driven equipment is to be operated in an XP and/or an electrically classified area (meaning flammable vapors are, or could be, present), it must be rated to operate in that environment. This means that the ignition source associated with the electricity used to power the equipment must be sealed to prevent the flammable vapors from infiltrating the housing (usually done by the manufacturer). The robustness and the integrity of these seals dictate whether the equipment can be operated in an environment where flammable vapors are always present [National Electric Code (NEC) Class 1 Division 1 area] or where flammable vapors could be present (NEC Class 1 Division 2).

Fire Suppression Systems

If a fire occurs in a process plant, the consequences can be devastating. Therefore, even though the main focus is on preventing fires, some attention must be paid to controlling and suppressing fires as well. There are many types of fire suppression systems in use in the chemical process industries, so decision making is necessary. One must decide between water-based and nonwater-based systems, between automatic and manual systems, or between wet pipe and dry pipe systems. How to make these decisions is beyond the scope of this entry; however, the *National Fire Protection Association's Fire Protection Handbook* and the *Society of Fire Protection Engineers Handbook of Fire*

Protection Engineering offer extensive additional reading in this area. The information available in these two sources is enough to rely upon for designing a fire suppression system.

In general, process plants are protected by the availability of fire hydrants, water monitors, and hose reels. In most cases, a fire will be fought manually by operators and other fire fighters. In some cases, with liquefied petroleum gas vessels, for example, a deluge system is installed over the vessel to help keep the vessel cool in case of a fire for preventing a Boiling Liquid Expanding Vapor Explosion (BLEVE). Sprinkler systems are sometimes found over critical motor-operated isolation valves, in compressor buildings or critical pump rooms. Nonwater-based systems are found in control rooms where computer equipment is located and in records retention areas. Automatic, water-based sprinkler systems are sometimes found in plant office areas.[12,13]

Facility Siting and Equipment Spacing

When conducting a process hazard analysis for an OSHA covered process, one must consider facility siting, also described in the American Petroleum Institute's Recommended Practice 752. It focuses on equipment and building spacing. Facility siting involves consequence analyses to determine the vulnerability zone associated with the extent of the potential damage from toxic vapor clouds, blast overpressure impact, or radiant heat effects. One determines which release scenarios should be considered, then decides on the extent of a vulnerability zone associated with a particular release scenario, and identifies buildings or facilities inside that zone. Once the buildings in the vulnerability zone are identified, one must determine if that building is occupied to the level (number of people hours) reaching unacceptable risk. Once the risk of impact is determined and it is considered unacceptable, a decision is made to remove the building activity to a location farther away from the release source, dismantle the building and build it in a location farther from the release source, pressurize the building, change the process to minimize the likelihood of the release scenario's occurrence, strengthen the blast resistance or vapor tightness of the building, or install a barrier between the source and the occupied building. This is a decision making process based on the extent of the risk and the feasibility of the countermeasures.[2]

Equipment spacing is a practice of determining safe distances between equipment in a process plant. The evaluation and input is most feasibly applied during the design phase and it is based on the expected vapor cloud travel distance, the presence of an ignition source in the equipment and the compatibility between materials handled in each piece of equipment or facility considered. The safest approach is to provide as much distance as possible between individual pieces of equipment; however, this uses more real estate and involves more cost associated with longer piping runs and process inefficiencies. There are trade offs and if one can install rated electrical equipment near a flammable vapor source, this may be less expensive than doubling the distance between them and be just as safe. Proper application of protective spacing principles can lower the risk of fire and explosion.

Other Protective Systems

Other loss prevention systems and devices that should be considered include:

- Static electricity controls—bonding and grounding, relaxation, and increasing process stream conductivity with additives.[3]
- Corrosion prevention—material selection, corrosion inhibition, cathodic protection, galvanic corrosion prevention, and extensive nondestructive and destructive inspection protocols.[9]
- Remote impounding flammable liquids—containment and drainage away from process equipment.

One can consult the references for this entry to learn more about these and other types of protection systems.

CONCLUSIONS

Loss prevention in the chemical process industries has come a long way in the last 15 years. It is now, thanks to regulatory and industry input, a much better managed function than it ever was. The OSHA PSM standard and the EPA RMP have become an integral part of the effort to reduce or prevent catastrophic loss incidents. We have learned much over the last 15 years, and the engineering and management systems approaches to loss prevention have contributed much to the loss prevention effort in the chemical process industries.

REFERENCES

1. Paradies, M.; Unger, L. *Taproot®, the System for Root Cause Analysis, Problem Investigation, and Proactive Improvement*, 1st Ed.; System Improvements: Knoxville, TN, 2000; *23*.
2. Flynn, A.M.; Theodore, L. *Health, Safety, and Accident Management in the Chemical Process Industries*, 1st Ed.; Marcel Dekker, Inc.:

New York, NY, and Basel, Switzerland, 2002; 1–77, 161–172.
3. Crowl, D.A.; Louvar, J.F. *Chemical Process Safety Fundamentals with Applications*, 2nd Ed.; Prentice Hall International Series in the Physical and Chemical Engineering Sciences, Prentice Hall PTR: Upper Saddle River, NJ, 2002; 27–29, 225–428.
4. Website of the Chemical Safety Board. www.csb.gov (accessed 1 September 2004).
5. Website of the Occupational Safety and Health Administration. www.osha.gov (accessed 1 August 2004).
6. Website of the U.S. Environmental Protection Agency. www.epa.gov (accessed 10 August 2004).
7. Haight, J.M.; Kecojevic, V. Process Safety Progress. *Automation vs. Human Intervention—What is the Best Fit for the Best Performance*? Vol. 24, No. 1, published online on January 2005.
8. Website for the Occupational Safety and Health Administration. www.instrumentation.co.za (accessed 7 February 2005).
9. O'Brien, T.; Luckiewicz, E. (chief authors). *Guidelines for Process Safety Fundamentals in General Plant Operations*, 1st Ed.; Center for Chemical Process Safety, American Institute of Chemical Engineers: New York, NY, 1995; 91–96, 103–114.
10. Committee on Industrial Ventilation. In *Industrial Ventilation, A Manual of Recommended Practices*, 23rd Ed.; American Conference of Governmental Industrial Hygienists: Cincinnati, OH, 1998; 2-2–2-7, 2–13.
11. Lipton, S; Lynch, J. *Health Hazard Control in the Chemical Process Industry*; John Wiley and Sons: New York, NY, 1987; 76–77.
12. DiNenno, P.J. (editor in chief). *SFPE Handbook of Fire Protection Engineering*, 3rd Ed.; National Fire Protection Association and Society of Fire Protection Engineers: Quincy, MA, Bethesda, MD, 2002; 4-1–4-72.
13. Cote, A.E. (editor in chief). *Fire Protection Handbook*, 18th Ed.; National Fire Protection Association: Quincy, MA, 1997; 6-1–6-136.

Low-Pressure Cascade Arc Torch

Hirotsugu Yasuda
University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

Arc discharge is a kind of DC discharge that is characterized by low voltage and high current, which is generated by a pointed small wire (cathode) and surrounding metal surface (anode).[1–7] The cathode and the anode are contained in a small nozzle (arc generator). The luminous gas phase created in the gap between the cathode and the anode is blown out of the arc generator by the flow of the carrier gas, which forms a jet of the luminous gas. The pressure of the carrier gas is generally superatmospheric at the entrance of the arc generator. When the luminous gas is blown into atmospheric pressure, the temperature of the flame is very hot, and the high-pressure arc torch is used as a heat source for cutting metals and welding processes. Cascade arc is the arc that is created by a special mode of arc generation, in which the cathode and the anode are separated by a series of conducting walls that are separated with each other by insulators.

CASCADE ARC GENERATOR AND LOW-PRESSURE CASCADE ARC TORCH REACTOR

Fig. 1 depicts the structure of a cascade arc generator. A disk of copper or brass, e.g., with diameter ~6 cm and thickness ~1 cm, which has a center hole, e.g., ~3 mm in diameter, is the major element of the cascade arc generator. Roughly, 7–10 rings are assembled with insulating gasket between rings so that 7–10 electrically floating conducting walls are aligned in the 2 mm diameter nozzle. Circulating temperature-controlled water cools each ring. A metal wire with low sputtering yield, typically tungsten wire, is used as the cathode, and the last ring is typically used as the anode.

When electrical voltage is applied between the cathode and the anode, the breakdown of the gas in the interelectrode space, which is surrounded by numbers of electrically floating conducting surfaces, occurs. The first arc discharge occurs between the cathode and the first ring, which establishes its electrical potential that is higher than the next ring. The discharge propagates from the first ring to the second ring, and so on to the anode surface in a cascading mode. Because of this mode of the propagation of arc discharge, this type of arc is termed as cascade arc or wall-stabilized arc. The advantage of cascade arc over other types of arc is in the ease of generating arc and the stability of the arc generation process.

A single monoatomic gas, e.g., argon, helium, etc., is used as the carrier gas of the cascade arc discharge. When the luminous gas is injected into an expansion chamber in low pressure, e.g., 1 torr or less, the flame extends a significant length (e.g., 1 m), which depends on the flow rate, input power, diameter of the nozzle, and the pressure of the expansion chamber. This mode of cascade arc torch is termed low-pressure cascade arc torch (LPCAT), which is useful in the surface modification by means of LPCAT treatment and LPCAT polymerization.

Low-pressure cascade arc torch can be used without employing the second gas that is injected into the expansion chamber, or with addition of a single or multiple gases or vapors. Because of a very high velocity of luminous gas jet stream, the location of the second gas injection is not a critical factor because the second gas is sucked into the jet stream. However, the second gas is generally introduced near the nozzle. Fig. 2 depicts the assembled cascade arc generator that can be attached to a vacuum system. In this configuration, the second gas injection port is integrated into the cascade arc generator. A pictorial view of an LPCAT reactor is shown in Fig. 3, which displays a stream of luminous gas injecting from the nozzle into the expansion chamber in vacuum.

As seen in Fig. 3, the LPCAT flame is relatively narrow implying that a uniform luminous gas phase is not created in the expansion chamber. Consequently, the treatment that can be achieved by an LPCAT is governed by the line of sight process, regardless of whether the substrate touches the luminous gas flame or not, and limited to a relatively small area that is exposed to the flame or near the tip of flame. When a substrate is placed along the line of the jet stream, the well-identifiable flame is destroyed, and gaseous species scatter in the downstream of the substrate. The scattered species could cause surface treatment effects; however, their extents are much smaller than that by the jet.

Plurals of integrated cascade arc generators, which are strategically placed on an expansion chamber, can be used to achieve a wider and uniform treatment.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008055

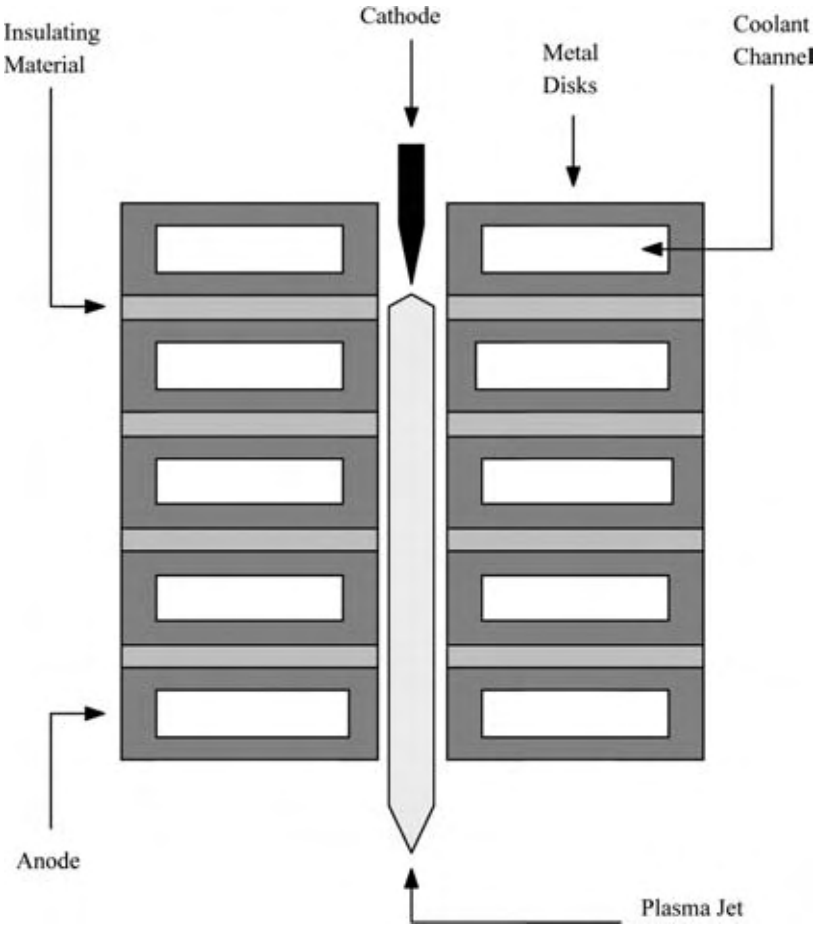


Fig. 1 General schematic of the cascade arc generator.

The relative motion of substrate with respect to the luminous gas jet is more or less mandatory for the uniform treatment. Fig. 4 depicts a reactor equipped with three cascade arc generators, of which two are used to treat substrates placed on a rotating plate.

Another mode of LPCAT processing is that the integrated cascade arc generator, such as the one shown in Fig. 2, is placed in a vacuum chamber held by a robot arm. In this mode, LPCAT jet could scan over a complex-shaped substrate by the robotic operation.

With low-pressure cascade arc, plasma formation (ionization/excitation of Ar) occurs in the cascade arc generator, and the major body of luminous gas is blown into an expansion chamber in vacuum. The majority of electrons and ions are captured by anode and cathode, respectively, of the cascade arc generator, and there is no external electrical field in the expanding plasma jet. Consequently, the photon-emitting excited neutrals of Ar cause the majority of chemical reactions that occur in the plasma jet. The luminous gas coming out of the nozzle interacts with gases existing in the space into which it is injected or the surface that is placed to intercept the jet.

Low-pressure cascade arc torch can be utilized in the following three modes in the surface modification of materials:

1. Without addition of the second gas or vapor, i.e., jet of excited argon neutrals.
2. With addition of the second gas that does not form the deposition of material (nonpolymer forming gas), i.e., jet of excited neutral species of the second gas.
3. With addition to the second gas that causes the deposition of material (polymer-forming gas and monomer), i.e., plasma polymerization by means of excited neutral species of argon.

The second gas is introduced in the expansion chamber. Because of an extremely high velocity of gas injecting from a small nozzle (e.g., 3 mm in diameter), the second gas injected into the expansion chamber in vacuum cannot migrate into the cascade arc generator, that is, the activation of Ar in the cascade arc generator and deactivation of the activated species of Ar in the expansion chamber, which activate the second gas introduced in the expansion chamber, are temporally

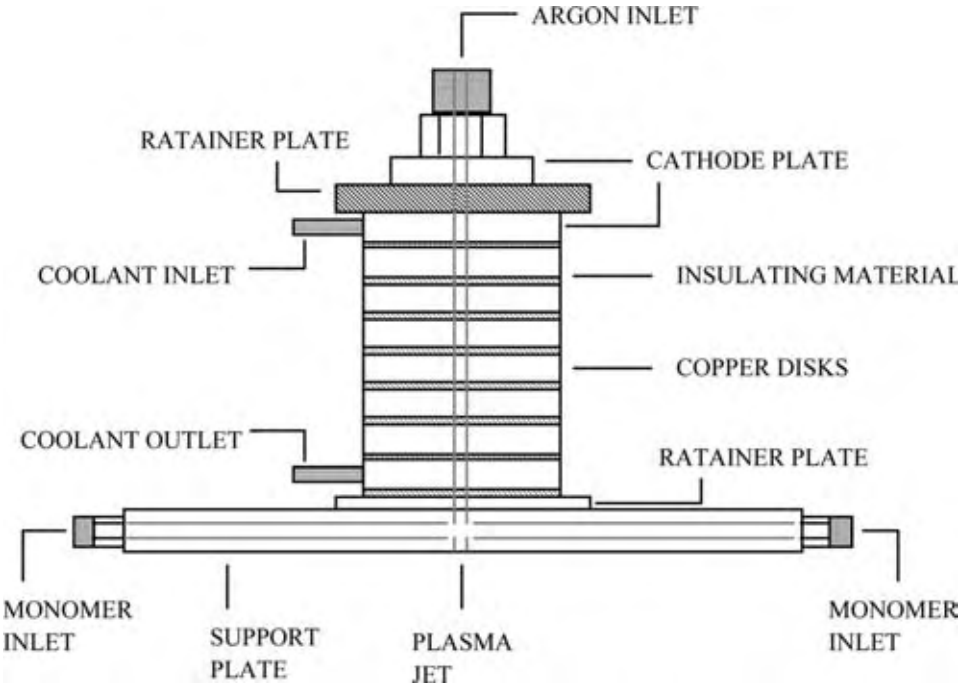


Fig. 2 Assembled cascade arc generator, which can be attached to a vacuum chamber.

and spatially separated. The LPCAT treatment and polymerization occur under such a totally (temporally, spatially, and kinetically) decoupled activation/deactivation system.

CREATION OF LUMINOUS GAS IN A CASCADE ARC GENERATOR

In LPCAT process, only an inert gas, such as Ar, exists in the cascade arc generator, and DC voltage is applied between the cathode and the anode. Therefore, it is a DC discharge of Ar, but it occurs under much higher pressure than most low-pressure DC discharges, and the gas travels very fast uniaxially in the generator. The basic process of ionization of Ar takes place in the cascade arc generator, which can be depicted as follows:

The ionization of Ar by high energy electron, e^*,

$$Ar + e^* \longrightarrow Ar^+ + 2e \quad (1)$$

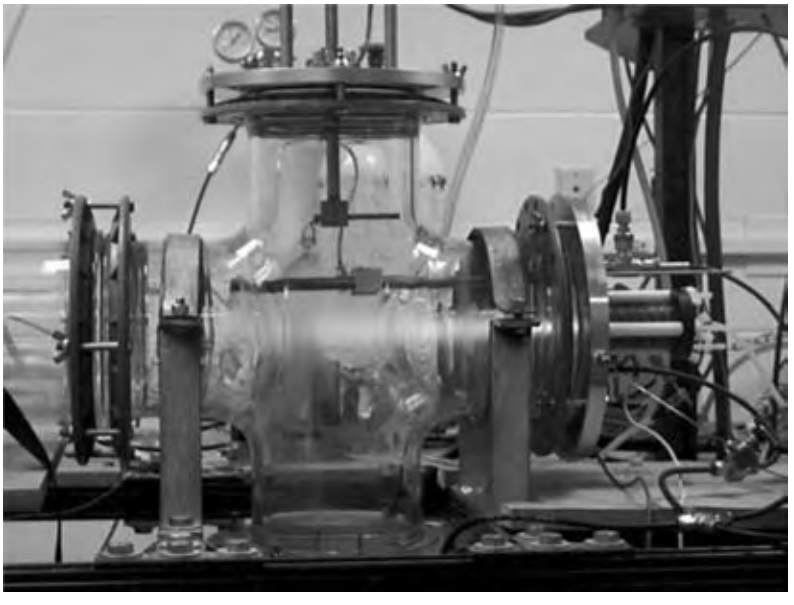


Fig. 3 Pictorial view of cascade arc torch reactor. *(View this art in color at www.dekker.com.)*

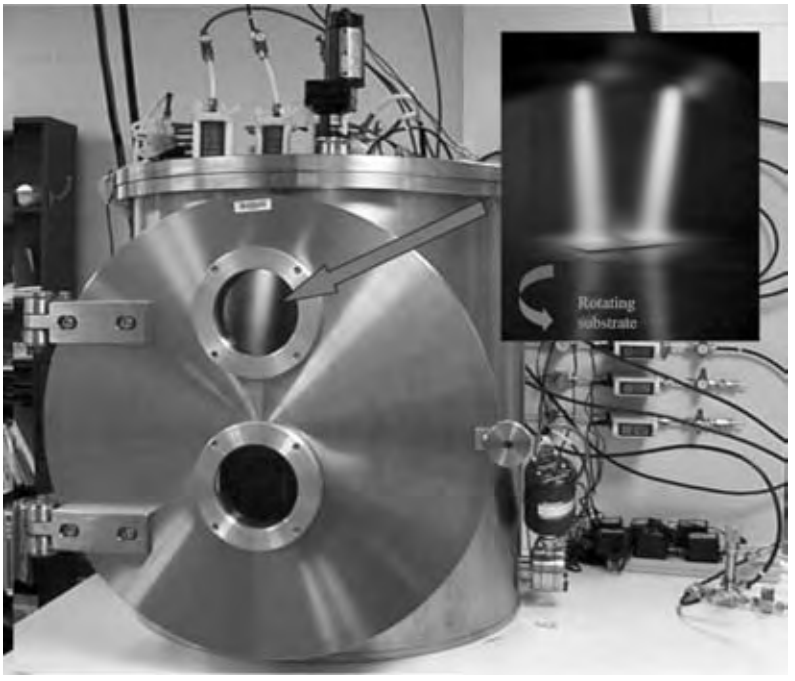


Fig. 4 The LTCAT model reactor with multiple cascade arc generators. *(View this art in color at www.dekker.com.)*

The creation of excited neutrals by an electron having the same level of energy also occurs simultaneously

$$Ar + e^* \longrightarrow Ar^* + e \quad (2)$$

The cathode captures most of Ar^+, and the anode captures the majority of electrons. When the luminous gas phase created in the cascade arc generator is blown out of the nozzle, the majority of species in the luminous gas jet are excited Ar neutrals and some strayed electrons.

In an expanding cascade arc plasma jet, the main characteristic is that electrons have very low kinetic energy (low electron temperature). The electron temperature estimates range from 0.3 to 1.5 eV in argon plasma jets by Langmuir double probe measurements.[5,8] As the double probes sample only the higher energy electrons, and not the bulk of electron distribution, the average electron temperature could be significantly lower than the above reported values. Because ions cannot be created at such low electron temperatures in the expansion chamber, only neutral lines are emitted in a pure argon or helium luminous gas jet.

In the expansion chamber, no electrical field exists, and no acceleration of electrons occurs. In such a passive environment, the number of electrons follows typical first-order decay as a function of the distance from the nozzle (Fig. 5).[8] Excited Ar neutrals outnumber electrons and dominate subsequent dissociation/excitation phenomena. The cascade arc luminous gas jet could be viewed as a jet stream of excited neutrals of the carrier gas.

Low-Pressure Cascade Arc Torch (Without Addition of the Second Gas)

In the expansion chamber without addition of the second gas, excited species Ar^* decay by emitting photons

$$Ar^* \longrightarrow Ar + h\nu_3 \quad (3)$$

The typical emission spectra of argon and helium are shown in Fig. 6 from 310 to 920 nm with few significant emissions outside of this region observed over the 200–1050 nm range. The emissions from the argon and helium plasma jets correspond exclusively to neutral argon or helium excited species, with no argon or helium ion lines.[8]

Some representative emission lines from the luminous gas jets are summarized in Table 1. The dominant features for argon plasma jet correspond to 4p $\rightarrow$ 4s transitions, with higher energy levels, such as 5p[2.5] $\rightarrow$ 4s[1.5], being quite clearly observed, as well. The excited atoms in a helium luminous gas jet have much higher energy levels than those in an argon luminous gas jet. With such a high energy over 23 eV, the excited helium neutrals can ionize nearly any molecular gases (excluding organic vapors, which mostly dissociate) injected into the expansion chamber.

The ratio of the concentration of excited species (energy carriers such as Ar) to that of the unexcited carrier gas is proportional to the power input

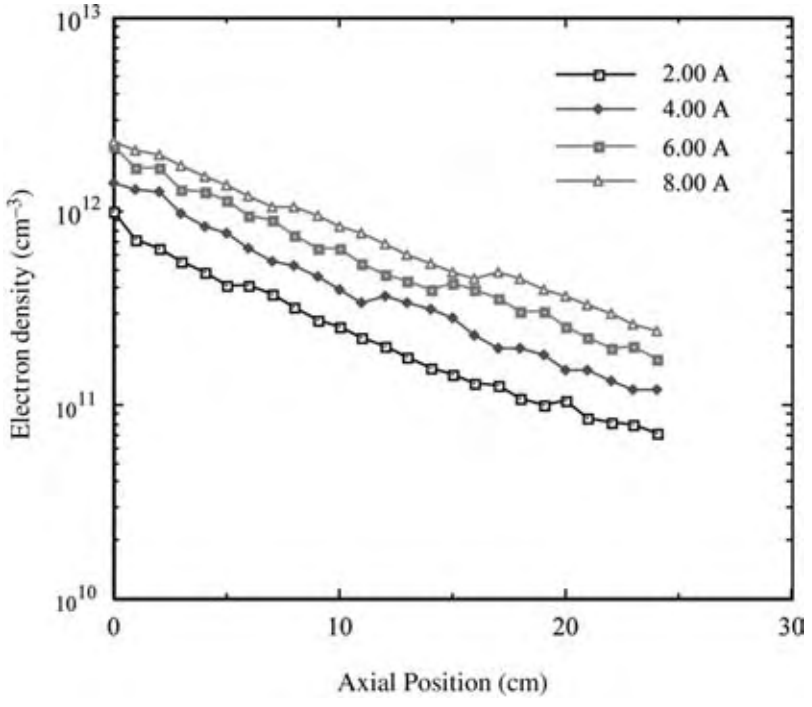


Fig. 5 Electron density (cm^{-3}) as a function of axial position and arc current. Other conditions are 2000 sccm argon and 560 mtorr (75 Pa). *(View this art in color at www.dekker.com.)*

given by *W*. This relationship can be expressed in Eqs. (4) and (5).[9]

$$\frac{[\text{Energy carrier}]}{[FM]_c} = \frac{[FM]_c^*}{[FM]_c} = \alpha_1 W \quad (4)$$

where α_1 is the proportionality constant (sec/J).

$$[\text{Energy carrier}] = [FM]_c^* = \alpha_1 W [FM]_c \quad (5)$$

In Eqs. (4) and (5), [Energy carrier] is the concentration (mass density) of excited species of the carrier gas in the jet stream blown out from the arc column into the expansion chamber (not the concentration in the chamber). The subscript "c" represents the carrier gas, and superscript "*" represents the excited carrier gas. *F* is molar flow rate, and *M* is the molecular weight of the gas. *FM* is the mass flow rate.

Fig. 7 shows the change of argon emission (763.5 nm) intensity against a combined experimental

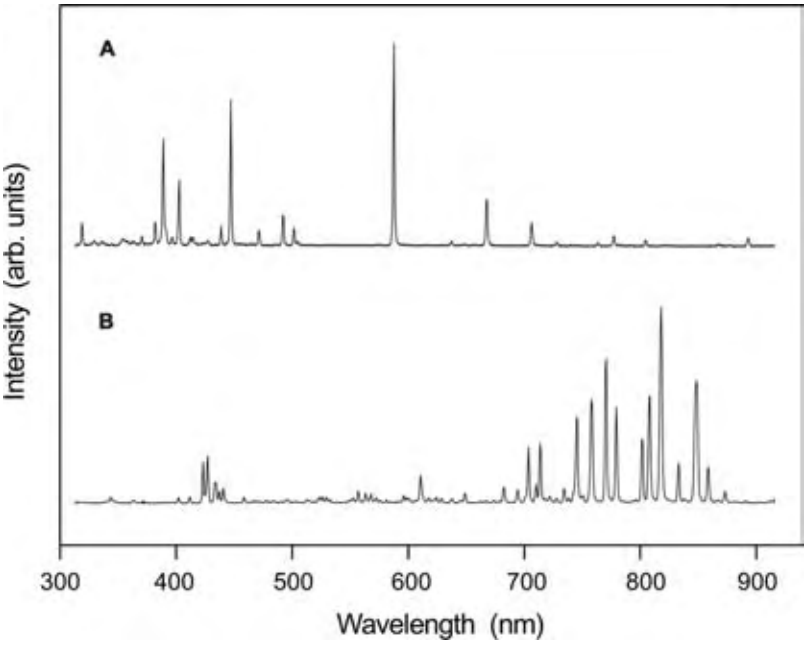


Fig. 6 Typical optical emission spectra of (A) helium plasma jet and (B) argon plasma jet; the spectra were obtained at an axial position 2.7 cm from the jet inlet; conditions are: (A) 3000 sccm helium, 1.35 kW, and 89 Pa; (B) 2000 sccm argon, 0.64 kW, and 75 Pa.

Table 1 Most intense emission lines observed in argon and helium plasma jets

Species	Emission wavelength (nm)	Transition	Energy of emitting state aboveground state, E_B (eV)
Ar	420.1	$5p[2\frac{1}{2}] \rightarrow 4s[1\frac{1}{2}]^0$	14.50
	696.5	$4p'[\frac{1}{2}] \rightarrow 4s[1\frac{1}{2}]^0$	13.33
	750.4	$4p'[\frac{1}{2}] \rightarrow 4s'[\frac{1}{2}]^0$	13.48
	763.5	$4p[1\frac{1}{2}] \rightarrow 4s[1\frac{1}{2}]^0$	13.17
	772.4	$4p'[\frac{1}{2}] \rightarrow 4s'[\frac{1}{2}]^0$	13.33
	794.8	$4p'[1\frac{1}{2}] \rightarrow 4s'[\frac{1}{2}]^0$	13.28
	811.5	$4p[2\frac{1}{2}] \rightarrow 4s[1\frac{1}{2}]^0$	13.08
	842.5	$4p[2\frac{1}{2}] \rightarrow 4s[1\frac{1}{2}]^0$	13.09
He	388.9	$3p^3P^0 \rightarrow 2s^3S$	23.01
	402.6	$5d^3D \rightarrow 2p^3P^0$	24.04
	447.2	$4d^3D \rightarrow 2p^3P^0$	23.73
	587.6	$3d^3D \rightarrow 2p^3P^0$	23.07
	667.8	$3d^1D \rightarrow 2p^1P^0$	23.07

parameter $W(FM)_c$, where W is the power input applied to the arc column and $(FM)_c$ is the carrier gas mass flow rate. The other argon emission lines, such as 706.7 and 604.3 nm, also showed similar trends. The argon emission intensity was proportional to the arc current at a fixed argon flow rate, and proportional to the argon flow rate at a fixed arc current. In other words, if the emission intensity was plotted against W or $(FM)_c$, the number of straight lines corresponding to values of $(FM)_c$ or W, respectively, were obtained, which are the similar dependences observed for the deposition in glow discharge on the discharge wattage and flow rate of monomer.[10]

The argon emission intensity showed a linear dependence on the combined parameter, $W(FM)_c$, i.e., the total energy applied to the monomers in the LPCAT process, as described by Eq. (5), can be expressed by this combined experimental parameter, $W(FM)_c$.

Addition of Nonpolymer-Forming Gases

The introduction of the second gas into the expansion chamber causes two significant visible changes: 1) the shrinking of the flame and 2) the change of color of the flame. These phenomena indicate that the excess energy carried by excited Ar neutrals are consumed

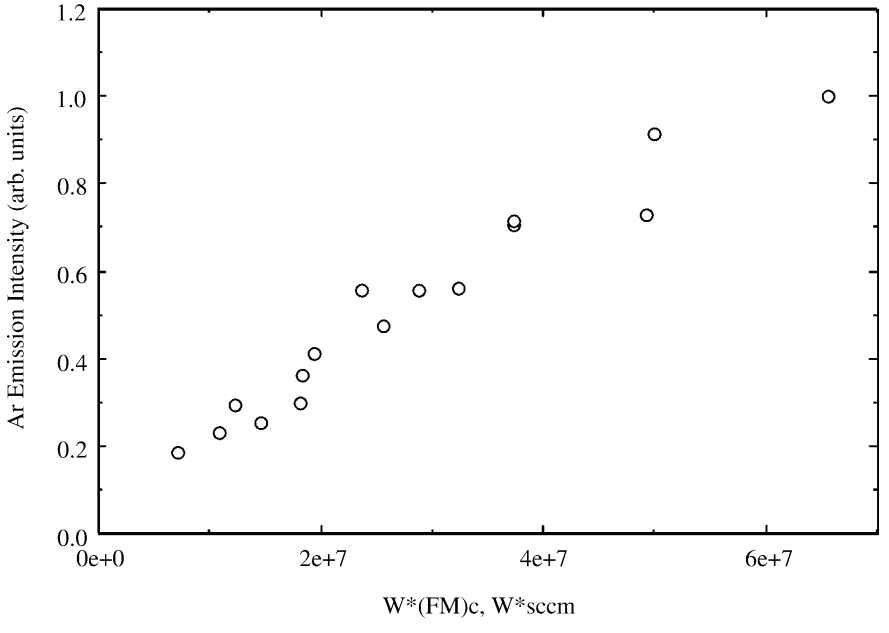


Fig. 7 The argon emission (763.5 nm) intensity as a function of the combined experimental parameter; W is the power applied to the arc column.

Table 2 Identified emission lines and bands

Species	λ(nm)	Transition	E_B(eV)	Remarks
N_2	337.1	$C^3\Pi_u \rightarrow B^3\Pi_g$	11.1	Second positive
	662.3	$B^3\Pi_g \rightarrow A^3\sum_a^+$	7.4	First positive
N_2^+	391.4	$B^2\Sigma_u^+ \rightarrow X^2\Sigma_g^+$	18.7	First negative
O	777.2	$3p^5P \rightarrow 3s^5S^0$	10.7	
	844.6	$3p^3P \rightarrow 3s^5S^0$	11.0	
O_2^+	525.1	$b^4\Sigma_g^- \rightarrow a^4\Pi_u$	18.2	First negative
H	656.2	$3d^2D \rightarrow 2p^2P^0$	12.09	H_α
	486.1	$4d^2D \rightarrow 2p^2P^0$	12.75	H_β
	434.0	$5d^2D \rightarrow 2p^2P^0$	13.06	H_γ
H_2		$a^3\Sigma_g^+ \rightarrow b^3\Sigma_u^+$	11.79	Continuum
NH	336.0	$A^3\Pi \rightarrow X^3\Sigma^-$	3.7	3360 Å system

in creating a new luminous gas phase, which consists of a mixture of the second gas and Ar (carrier gas), and the color depends on the nature of the second gas. The length of the mixed luminous gas flame also depends on the flow rate of the second gas (at a fixed flow rate of Ar), i.e., the higher the flow rate, the shorter the length of flame. Tables 2–4 depict characteristics of the luminous gas phase when a second gas is added to the expansion chamber.[11]

The electron-impact ionization does not occur in the cascade arc torch, and the energy transfer between excited neutrals of the carrier gas and the added gases becomes a dominant reaction, i.e., the Penning-type reactions or resonance reactions between gas molecules and excited neutrals of the carrier gas is the principal generation process for the reactive species.[12]

On addition of the second gases to the plasma jets, the emission of argon or helium plasma jets are highly quenched. The dominant features in the emission spectra are due to the excited species corresponding to the added gases, with only the strongest argon or helium lines remaining visible.[11]

Nitrogen

When nitrogen is added to plasma jets, all argon or helium emission lines are highly quenched, and a very strong pink flame for argon plasma jet and a green flame for helium plasma jet are formed. The dominant features in the spectra are due to nitrogen species. Fig. 8 depicts optical emission spectroscopy (OES) signals of luminous gas jet of N_2 added to: 1) Ar and 2) He LPCAT.

In radiofrequency (RF) plasmas (see Plasma Polymerization Coatings) that are characterized as low ionization degree and high electron temperature, the emission is mainly due to electron excitation and pooling reaction of the molecular metastable. Fig. 9 depicts various emissions observed in RF discharge of nitrogen. The main ion is N_2^+, which also can be excited by electrons to give the molecular ion emission. In expanding cascade arc helium plasma jet, only the spectrum of the first-negative system of N_2^+ was observed, as shown in Fig. 10, which depicts comparison of nitrogen luminous gas in RF and in LPCAT.

Table 3 Spectra emitted in argon plasma jet with addition of reactive gases

Reactive gas	Visual appearance	Observed spectra
None	Orange	Ar atom lines
N_2	Pink	N_2 second positive
O_2	Orange	O atom lines
Air	Pink	N_2 second positive O atom lines
H_2	Light blue	H_2 molecular continuum H_α, H_β lines
$N_2 + H_2$	Pink	N_2 first and second positive NH band H_α line
NH_3	Light blue	NH band H_α line

Table 4 Spectra emitted in helium plasma jet with addition of reactive gases

Reactive gas	Visual appearance	Observed spectra
None	White	He atom lines
N_2	Green	N_2^+ first negative
O_2	Light blue	O_2^+ first negative O atom lines
Air	Green	N_2^+ first negative O atom lines
H_2	White	NH band H_α, H_β lines

Argon-excited neutrals were consumed by the increasing nitrogen addition, and the emission intensity of the N_2 second-position emission band keeps increasing with increasing nitrogen flow rate from 0 to 60 sccm, which shows that more activated nitrogen species were produced (Fig. 11).

The electronic energy level of the excited argon or helium neutrals plays a dominant role in active nitrogen excitation processes in the plasma jets. As can be seen in Table 1, excited argon neutrals (E_B: ~13 eV) have more than enough energy to excite either state of N_2 first-position (E_B: ~7.4 eV) or state of N_2 second-position (E_B: ~11.1 eV), and excited helium neutrals (E_B: ~23 eV) have more than enough energy to excite any state of N_2 first position, N_2 second position, and N_2^+ first negative (E_B: ~18.7 eV). However, from the emission spectra of low-pressure cascade arc plasmas of nitrogen, a selective excitation of nitrogen species was observed. Only the activated species of N_2 second position (E_B: ~11.1 eV) in argon plasma jets and N_2^+ first negative (E_B: ~18.7 eV) in helium plasma jets were observed, which have the electronic energy levels close to those of excited argon atoms (E_B: ~13 eV) or helium atoms (E_B: ~23 eV), respectively. No excited species of nitrogen with higher or much lower energy levels than that of excited argon or helium neutrals was identified.

In the creation processes of reactive nitrogen species in the low-pressure cascade arc plasmas, only species that have close energy match to the excited species of carrier gas occur, e.g., N_2 second position and excited argon atoms or N_2^+ first negative and excited helium atoms. The selective excitation of nitrogen species was not observed in RF plasmas from either pure nitrogen or its mixture with argon or helium, probably because of the electron-impact ionization rather than the energy transfer type activation predominates in the RF glow discharge of nitrogen.

Oxygen

When oxygen was added into an argon plasma jet at 60 sccm, neither obvious quenching of argon emission

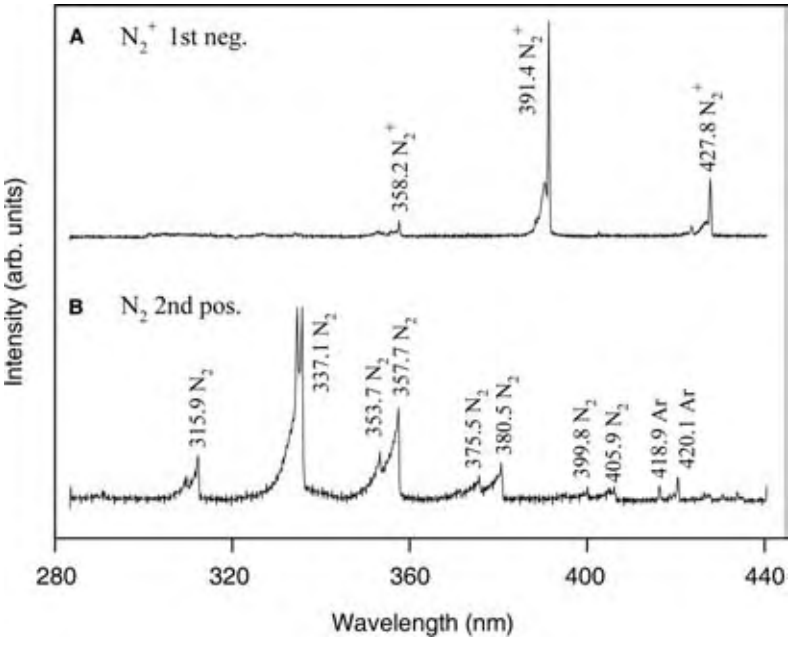


Fig. 8 Typical emission spectra of nitrogen in low-temperature cascade arc plasma jets: (A) N_2 in helium plasma jet and (B) N_2 in argon plasma jet.

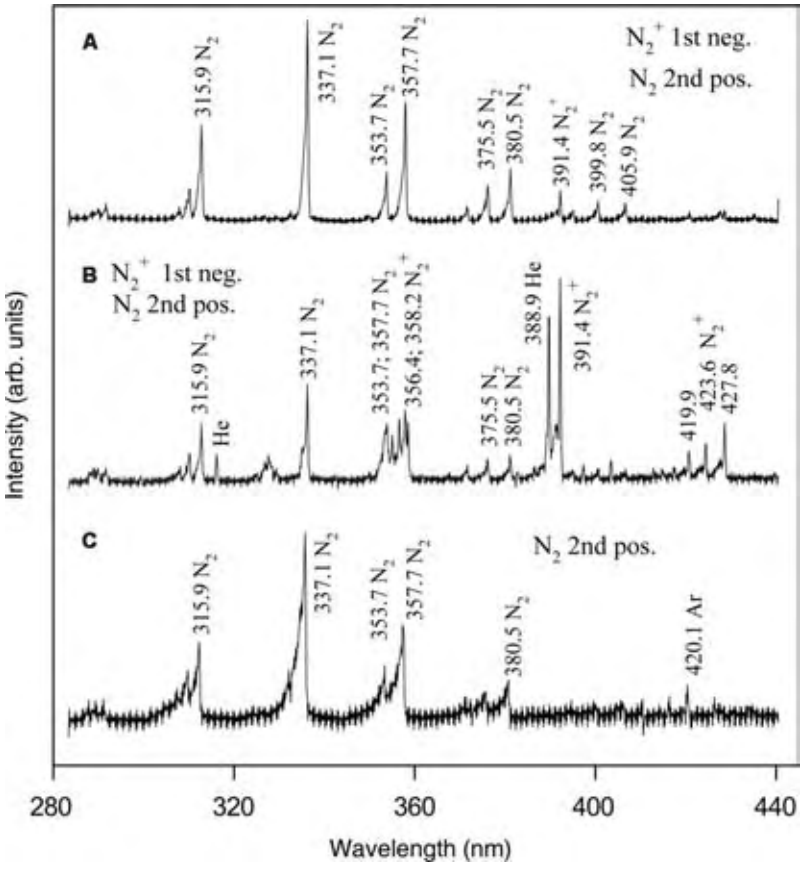


Fig. 9 Typical emission spectra of RF plasmas: (A) pure nitrogen, 60 sccm N_2, 10 Pa, RF power 40 W; (B) mixtures of 60 sccm nitrogen and 2000 sccm helium, 75 Pa, RF power 100 W; and (C) mixture of 60 sccm nitrogen and 2000 sccm argon, 75 Pa, RF power 250 W.

nor color change of the plasma jet was observed. The emission due to Ar atoms dominated the spectrum. However, the optical emissions due to excited oxygen atoms at 777.2 nm (E_B: ~10.7 eV) and 844.6 nm (E_B: ~11.0 eV) were clearly observed.

When oxygen was introduced into a helium plasma jet, rapid quenching of helium emission occurred and a very short light-blue flame was formed, and the emission due to N_2^+ first-negative system (E_B: ~18.2 eV) appeared in addition to the oxygen atom emission. Because of the obvious quenching, the emissions due to He atoms became very weak. This result can be clearly understood from the different energy levels between the excited argon and the helium neutrals, i.e., the excited helium atoms have more than enough energy to ionize oxygen molecules and excite them from the ground state to an excited state, but excited argon atoms do not have enough energy. The following reactions

$$He^* + O_2\left(X^3\sum_g^-\right) \longrightarrow O_2^+\left(b^4\sum_g^-\right) + e + He$$

$$O_2^+\left(b^4\sum_g^-\right) \rightarrow O_2^+\left(a^4\prod_u\right) + h\nu$$

are energetically possible for the production of O_2^+ first-negative system species. The generation process for excited oxygen atoms in argon plasma jet can also be clearly described by the following excitation transfer reaction:[12]

$$Ar^* + O_2\left(X^3\sum_g^-\right) \rightarrow O^* + O + Ar$$

Air

As the main components of air are nitrogen and oxygen, the main feature of the emission spectra of air in argon or in helium plasma jets is also a typical mixture of nitrogen and oxygen emission spectra in the plasma jets. When air was introduced into the argon plasma jet, a strong pink flame was formed and the main emissions were due to N_2 second-position molecules and excited oxygen atoms. When it was added to a helium plasma jet, the dominant emissions were due to N_2^+ first-negative molecular ions and excited oxygen atoms, but no O_2^+ first-negative emission was observed.

Hydrogen

When hydrogen gas was introduced into argon plasma jet, all the argon emissions were strongly quenched and a very short and brilliant flame with a light blue color was formed. The emissions from excited hydrogen atoms (E_B: ~12 eV) and molecules (E_B: ~11.8 eV) were

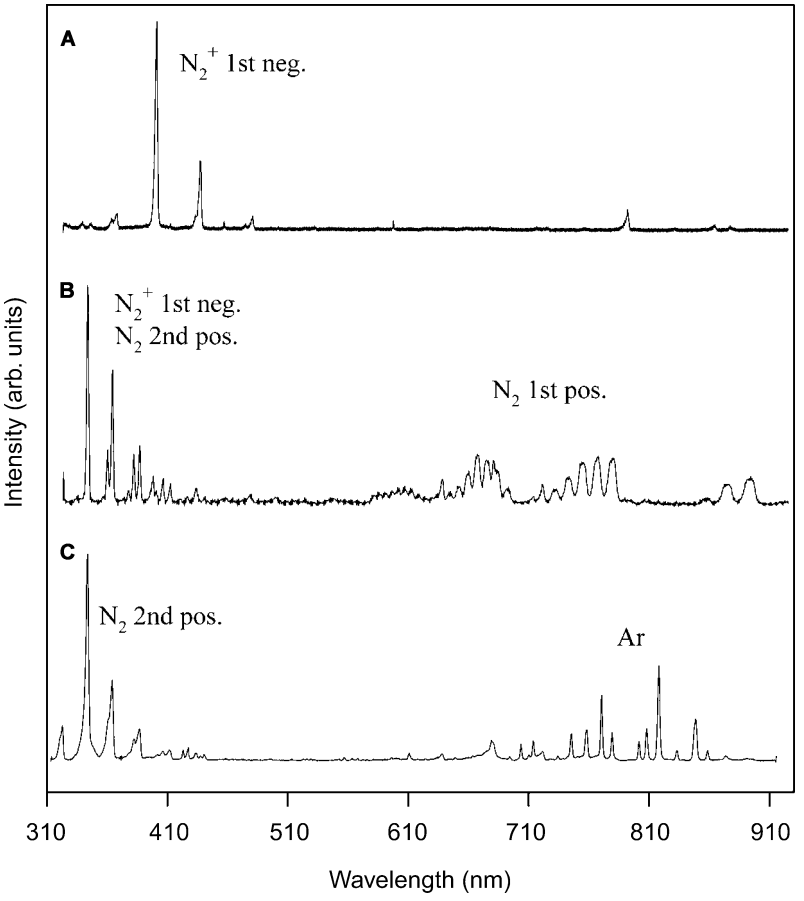


Fig. 10 Typical emission spectra of 60 sccm N_2 (A) in helium plasma jet, (B) RF plasma of pure N_2, 10 Pa, RF power 30 W, and (C) in argon plasma jets.

observed. However, when hydrogen gas was added to helium plasma jet, neither the strong quenching effect nor color change was observed. Only very weak emissions from excited hydrogen atoms were detected, and no emission from excited hydrogen molecules was observed, as the energy level of excited helium atoms exceeds the ionization energy for both H atoms (E_i: ~13.6 eV) and H_2 molecules (E_i: ~15.4 eV).

This system forms highly ionized so-called Penning mixtures.[13,14] The higher excited states of H_2^+ are partly stable and partly unstable, depending on the quantum numbers of the electron present. The stable excited states have, however, only very shallow minima of the potential curves.[15] That is the reason why no spectrum of H_2^+ is observed for the helium plasma jet. The argon excited neutrals, on the other hand, cannot ionize either hydrogen atoms or molecules, but could produce excited H_2 molecules, which can be detected by optical emission spectroscopy.

Ammonia

When ammonia gas was introduced into an argon plasma jet at a flow rate of 60 sccm, all argon emission lines disappeared, and a very short but brilliant light blue flame was formed. A very strong NH emission band was observed. In the ammonia RF plasma, some very weak N_2 emission bands due to N_2 second position appeared, but in the ammonia flame formed in an argon plasma jet, no emission related to N_2 species was observed.

Mixture of nitrogen and hydrogen

Figs. 12 and 13 show the emission spectra of an argon plasma jet with the addition of nitrogen and hydrogen at different compositions. The N_2 emission due to N_2 second-position intensities decreased greatly, even when a small amount of hydrogen existed. At a higher hydrogen composition in the mixture, very weak emission bands due to N_2 second position were observed. Quite strong emissions due to N_2 first position appeared with different levels of hydrogen in the gas mixture. Adding hydrogen to the system can limit the selective excitation of nitrogen species in low-pressure cascade arc plasmas, though the mechanism is not clear.

Addition of Material-Forming Gas

The creation of chemically reactive species from polymer-forming gas (monomer) follows the same

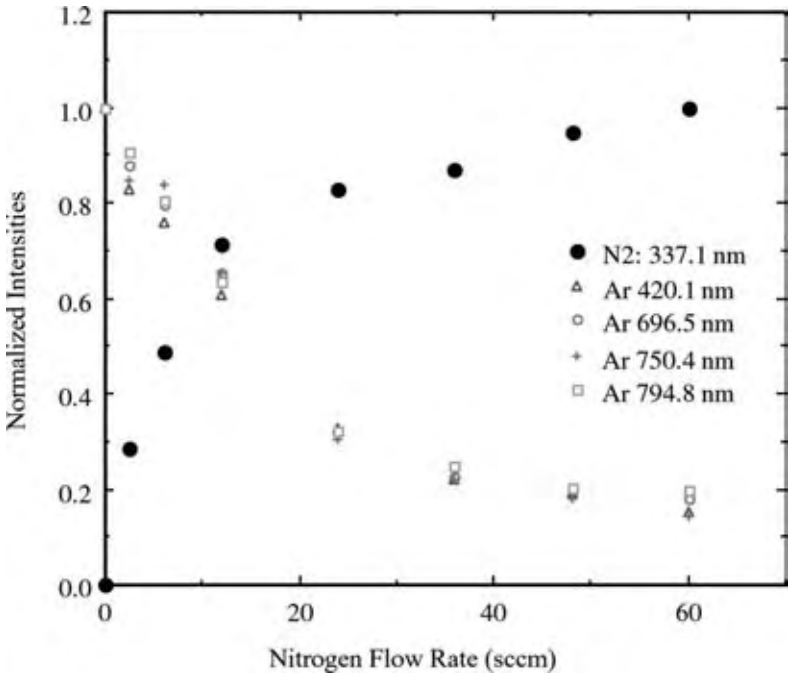


Fig. 11 The dependence of emission argon lines and nitrogen second-position band (337.1 nm) emission intensities at an axial position 2.7 cm from jet inlet on the flow rate of nitrogen added to argon plasma jet. Conditions are 2000 sccm argon, 60 sccm nitrogen, 75 Pa, 0.64 kW. *(View this art in color at www.dekker.com.)*

principle described for nonpolymer-forming gases, i.e., molecular dissociation by energy-transfer mechanism. On addition of monomers to the argon luminous gas jet, the emissions of argon luminous gas are highly quenched. The dominant features in the emission spectra are due to excited species corresponding to the relevant monomers with the strong argon lines remaining visible. Photon-emission aspects of some organic molecules are described below.

CH_4 or CH_3OH

The spectral features of OES spectra of low-temperature cascade arc plasmas of CH_4 and CH_3OH, are identical for both of these systems. Emission bands of CH radicals, emission lines of H atoms constitute the dominant emission characteristics of OES spectra for both the plasma systems. The CH emission ($A^2\Delta \rightarrow X^2\Pi$ at 431 nm) intensities are almost the same for both CH_4 and CH_3OH plasmas, and follow similar trends with the change of arc currents.

If one assumes that the densities of the electronically excited states, from which the observed emissions originate, are directly proportional to those of the ground state, the emission intensity profiles reflect approximately the species concentrations, especially at constant pressure. Therefore, the above results may indicate that, at the same operating conditions, almost the same quantity of CH radicals exist in CH_4 and CH_3OH cascade arc plasmas. From the OES spectra assignment, CH radicals are the only reactive species that could be attributed to the growth of plasma polymers in both CH_4 and CH_3OH cascade arc plasmas.

If one speculates only based on OES data, CH_4 and CH_3OH cascade arc plasmas should give similar deposition rates. However, the deposition rate did not follow the prediction. The deposition rates of CH_3OH in LPCAT are less than 1/10 of that for CH_4. The results of deposition rates indicated that there existed an inverse correlation between OES signals and the deposition rate, indicating that polymerizable species are not photon-emitting species.

CF_4 or C_2F_4

Clear-cut demonstrations for the case, in which polymerizable species that do not emit photons are created exclusively by the molecular dissociation by energy transfer principle, was found with addition of CF_4 or C_2F_4 to Ar LPCAT. The finding implies that the photon emission per se is not the essential indicator of chemical reaction pertinent to plasma polymerization. According to the literature, the CF_4 and C_2F_4 plasmas are the well-investigated fluorinated carbon plasma systems by OES diagnostics with conventional plasma sources. However, without influence of electron-impact ionization in LPCAT, these cases turned out to be the case for the formation of polymerizable species that do not emit photons, which occurs by the energy transfer mechanism.

The OES spectra of low-temperature cascade arc plasmas for carrier gas Ar only and with CF_4 and C_2F_4 addition were identical, although the quenching

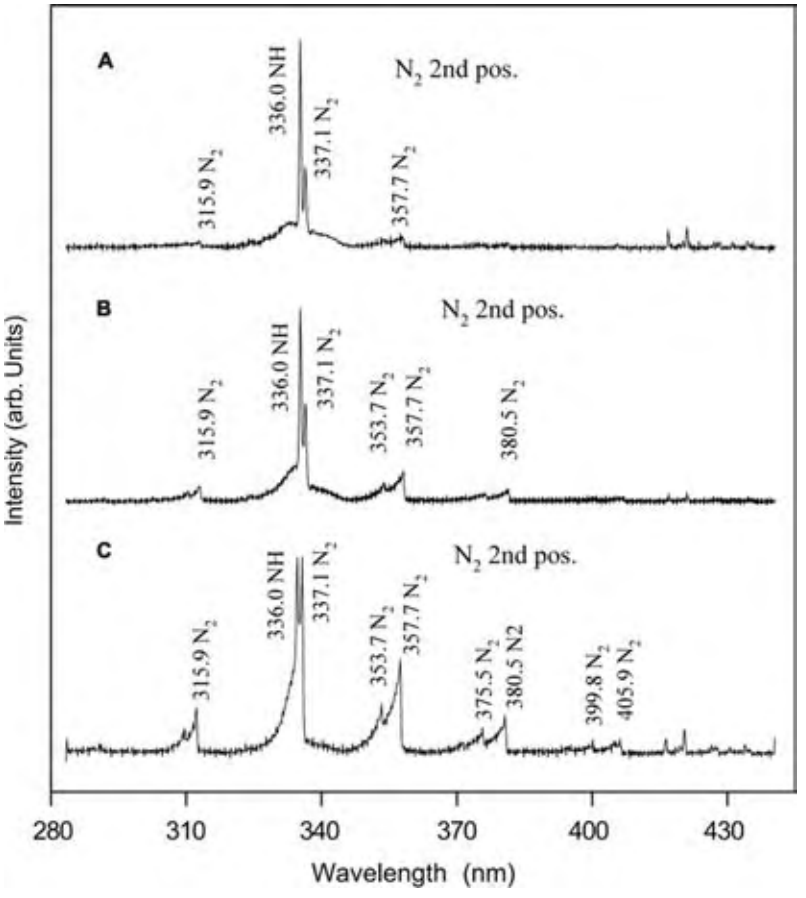


Fig. 12 The optical emission spectra of argon plasma jets (A) with addition of 10 sccm nitrogen and 10 sccm hydrogen, (B) with addition of 60 sccm nitrogen and 2.7 sccm hydrogen, and (C) with pure nitrogen addition, 60 sccm nitrogen. The cascade arc conditions are 2000 sccm argon, 0.64 kW, and 75 Pa.

of the flame occurred when the second gas is added. Low-temperature cascade arc plasmas of CF_4 and C_2F_4 did not show any additional peaks or continuums when compared to the Ar plasmas, like CF_2^+ or the CF lines that were found in the OES spectra for ($Ar + CF_4$) and ($Ar + C_2F_4$) RF plasmas. None of the species previously reported for CF_4 RF plasma, i.e., CF_2 and CF bands, as well as fluorine atom emissions, exists in the OES spectra of CF_4 or C_2F_4 LPCAT.[16] The addition of CF_4 or C_2F_4 quenched

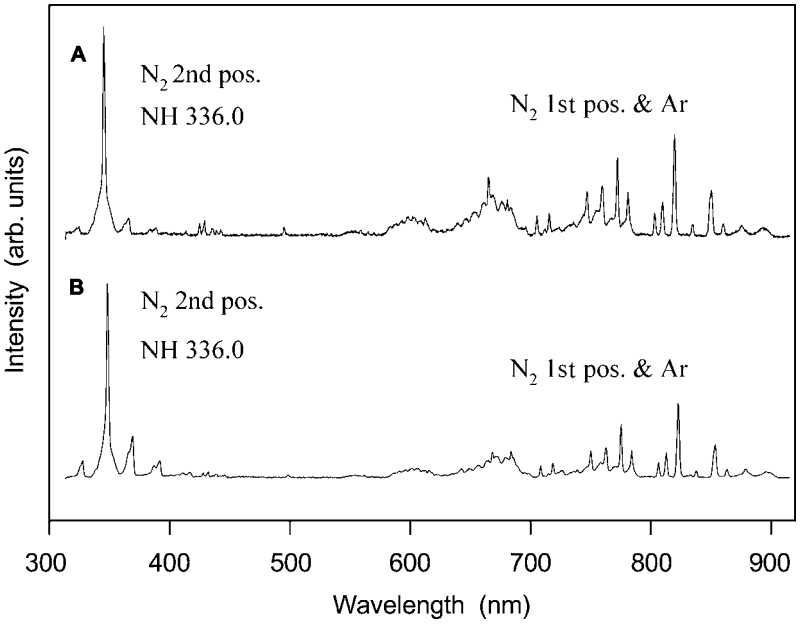


Fig. 13 The optical emission spectra of argon plasma jets with addition of nitrogen and hydrogen mixture: (A) 10 sccm nitrogen and 10 sccm hydrogen and (B) 60 sccm nitrogen and 2.7 sccm hydrogen. The cascade arc conditions are 2000 sccm argon, 0.64 kW, and 75 Pa.

the cascade arc torch flame (the excess energy is transferred to the second gas), but did not create reactive species of the added gas that emits photons, which can be identified by OES.

X-ray photoelectron spectroscopy data, on the other hand, showed that the low-temperature cascade arc torch treatment of (Ar + CF_4) and (Ar + C_2F_4) yielded just as good, if not better, fluorination of PET fibers as RF plasma treatment with these gases.[16,17] These examples clearly demonstrate that the main polymerizable species in plasma polymerization (free radicals) are not photon-emitting species in most cases. This is in accordance with the growth and deposition mechanism based on free radicals, which account for the presence of large amount of dangling bonds in most plasma polymers[10] (see ''Plasma Polymerization Coating'').

The formation and the dissipation of luminous gas phase of organic vapor by excited Ar neutrals in the cascade arc polymerization could be described as follows. A fraction of photon-emitting Ar neutrals (Ar^*), expressed by nA^*, interact with the monomer (M), and cause the dissociation of the monomer yielding fragmented photon-emitting species (A^* and B^*) and hydrogen atom (H) according to the scheme described by Eq. (1):

$$nAr^* + mM \longrightarrow m(A^* + B^*) + xH + (n - m)Ar^* \quad (6)$$

The remaining Ar^* emit photons according to Eq. (3). The photon emission by A^* and B^* can be expressed by the same equation, if these species do not chemically react. A fraction (β) of fragmented photon-emitting species reacts with the monomer or its fragmented moiety (represented by X) to yield polymerizable species (represented by $AX^\bullet$).

$$\beta A^* + X \longrightarrow \beta AX^\bullet \quad (7)$$

In the case in which no photon-emitting species are created, the formation of polymerizable species can be represented by a modified Eq. (6), as

$$nAr^* + mM \longrightarrow m(A^\bullet + B^\bullet) + xH + (n - m)Ar^* \quad (8)$$

The perfluorocarbons injected into Ar LPCAT represent the case represented by Eq. (8). On injection of a perfluorocarbon, such as CF_4, C_2F_4, etc., luminous gas jet shrinks. However, OES data indicate that there is no signal discernible to F-containing species. Radio-frequency discharge of mixtures of Ar and perfluorocarbon shows F-containing species, indicating the route to create polymerizable species in LPCAT is different from that in RF plasma because of the absence of electron or ion impact dissociation of monomer.

DEPENDENCE OF DEPOSITION RATE ON OPERATIONAL PARAMETERS

In the conventional glow discharge polymerization, the normalized deposition rate is given by

$$\text{Normalized deposition rate} = \alpha W/FM \quad (9)$$

where W is the electrical energy input to the discharge system, and FM (mass flow rate) represents the total mass of gases that adsorb the energy[10] (see ''Plasma Polymerization Coating''). W/FM represents the energy input per unit mass of gas involved, whereas W represents the energy input into the electrical circuit. As the carrier gas plasma is the direct energy source for the activation of monomers, the concentration of energy carrier [Energy carrier] represents the amount of energy available to activate LPCAT plasma polymerization in the cascade arc torch (in the expansion chamber).

The deposition rate in LPCAT polymerization can be described by the following expressions by replacing W in Eq. (9) (for general glow discharge polymerization) with [Energy carrier].[19]

$$\text{Normalized(DR)} = \alpha_2 \frac{[\text{Energy carrier}]}{[FM]_m} \quad (10)$$

By virtue of Eq. (5),

$$\text{Normalized(DR)} = \alpha_2 \frac{[FM]_c^*}{[FM]_m} = \alpha_2 \frac{\alpha_1 W[FM]_c}{[FM]_m} \quad (11)$$

where α_2 is a unit-less proportionality constant. Thus,

$$\text{Normalized(DR)} = \alpha_1\alpha_2 \frac{W[FM]_c}{[FM]_m} \quad (12)$$

In Eqs. (9)–(12), subscript ''m'' represents monomer, and the other experimental parameters are the same as described earlier. In essence, $W[FM]_c/[FM]_m$ is an energy-input factor (J/sec), the quantity of which can be considered as the energy applied to monomer per mass unit of monomers in unit time, transported by luminous carrier gas.

First, the energy transfer scheme based on $W[FM)_c/[FM]_m$ can be verified by examining OES spectrum of the second gas added into the expansion chamber of the cascade arc torch reactor as shown in Fig. 7.

The plotting deposition rates of different monomers with different feed rates in Normalized (DR) vs. $W[FM]_c/[FM]_m$ coordinates would give rise to the most objective comparison of their tendencies with regard to deposition. In LPCAT, a very important fact is that the polymerization (material formation) is ''atomic'' rather than ''molecular'' processes, implying that the depositing entities are fragmented species of the original monomer molecule.[10] Therefore, the deposition rate in LPCAT polymerization is determined largely by the type of atoms contained in the monomer structure, rather than by molecular structures.

Fig. 14 depicts the deposition characteristics of LPCAT polymerization. In this diagram, different monomers, different flow rates, and different electrical power inputs are all pooled together. The figure shows that: 1) the deposition rate in cascade arc torch can be well represented by Eq. (12), and 2) the deposition characteristics of silicon-containing monomers and hydrocarbon monomers are significantly different.

The silicon-containing monomers exhibited much higher deposition yield than hydrocarbon monomers. The wider scattering among various Si-containing organic molecules is due to the fact that each molecule contains different amount of C-based moieties. It can be estimated, from this figure, that Si-based moieties polymerize nearly seven times faster than C-based moieties. As there is a significant difference in the deposition behaviors of Si–C compounds and hydrocarbons, these results also indicate that the type of atoms contained in the monomer structure plays an important role in plasma polymerization in LPCAT.

Because of the relatively narrow beam of plasma jet, the deposition occurs in relatively narrow area. Deposition rate profiles as a function of axial position are shown in Fig. 15. Experimental conditions are 8.00 A arc current, 2000 sccm argon, 10.0 sccm methane, and 560 mtorr (75 Pa). The centerline deposition rate decreases sharply and the deposition rate at the shoulders of the profile increases with increased axial position, with profile broadening attributed to the diffusion of polymerizable species from the jet axis. The total mass deposition rate over a circle of 50 mm radius is given in Table 5, along with the conversion of methane to plasma polymer. Although the distribution profile broadens as the axial distance increases, the total deposition rate remains nearly constant within the distance examined. This implies that the distance from the nozzle, beyond a certain distance, is not a critical factor in practical deposition process, because the velocity of gas is so large that the variation of the distance is insignificant in the deposition process.

Low-pressure cascade arc torch polymerization or coating could be considered more or less the same as the plasma polymerization or coating by other conventional plasma processes. The ultrathin layers prepared by LPCAT polymerization have the general characteristics of plasma polymers, i.e., amorphous (noncrystalline), high concentration of the dangling bonds (free radicals trapped in immobile solid phase), and the high degree of the internal stress in the layer.

The internal stress follows the general pattern of its dependency on the energy input. Fig. 16 depicts the dependence of the internal stress on the power input

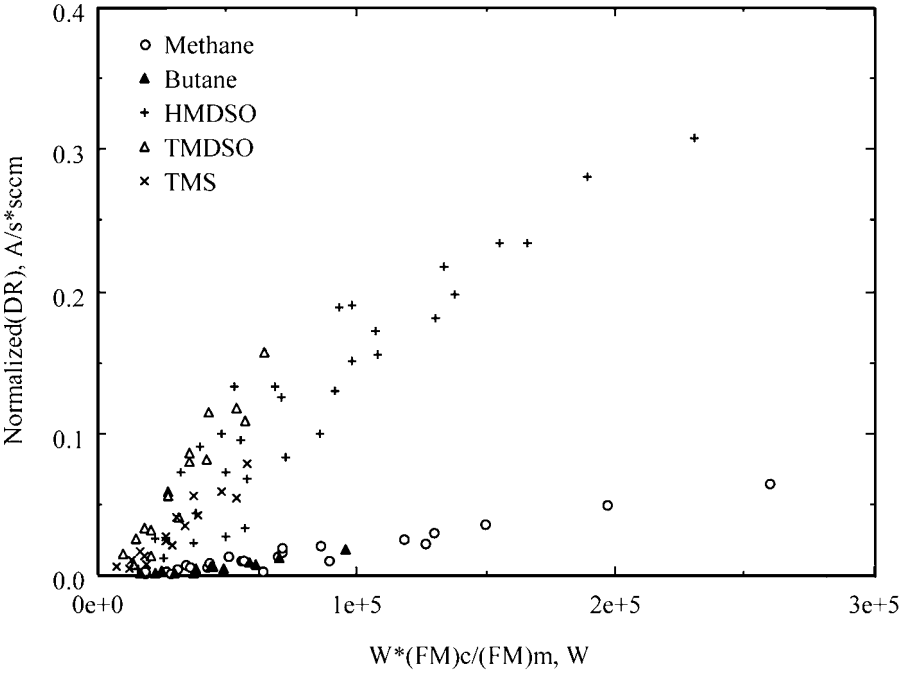


Fig. 14 Dependence of the normalized deposition rate of silicones and hydrocarbon monomers on the parameter $W^*(FM)_c/(FM)_m$ in cascade arc torch polymerization. The deposition rates were obtained at an axial position of 27.5 cm from the luminous gas jet inlet.

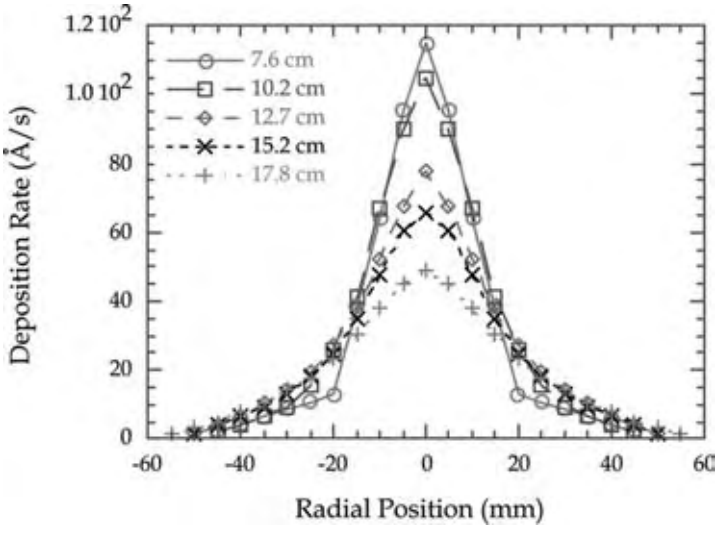


Fig. 15 Variation of the radial deposition rate distribution profile with axial position; conditions: 8.0 Å, 2000 sccm argon, 10.0 sccm methane, and 560 mtorr. *(View this art in color at www.dekker.com.)*

parameter. The larger monomers tend to create the higher internal stress, but in all cases, the internal stress increases with the power input. The internal stress is closely related to the refractive indices, as shown in Fig. 17. Si containing monomers have lesser internal stress and lower refractive indices compared to hydrocarbon (Nobonene), which showed conspicuously different lines in both the figures.

CHARACTERISTIC FEATURES OF LPCAT PROCESSING

The major effects of LPCAT treatment and LPCAT polymerization coating are, by and large, the same as conventional glow discharge treatment and plasma polymerization coating. However, the processing of LPCAT treatment or coating is significantly different in that it is a spray processing that requires the relative motion of the substrate with respect to the torch flame for the substrate larger than the diameter of LPCAT jet. Although the local treatment or deposition rate is high, it does not translate directly to the high overall processing rate depending on size and shape of the substrate.

Table 5 Total mass deposition rates and percent conversion of methane to plasma polymer at various axial positions. Conditions are 8.00 A, 2000 sccm argon, 10.0 sccm methane, and 560 mtorr

Axial position (cm)	Total rate ($\mu g/cm^2/sec$)	Conversion (%)
7.6	9.9	10.2
10.2	11.2	11.5
12.7	11.9	12.3
15.2	11.3	11.7
17.8	11.3	11.7

The process requires nearly three orders of magnitude higher flow rate of the carrier gas than typical flow rates of monomer or a mixture of monomer and a carrier gas in conventional glow discharge processes. The high consumption of carrier gas might necessitate the inclusion of carrier gas recovery system in an industrial scale operation.

The process, therefore, is not an alternative means to carry out conventional plasma processes, and the adaptation of the process should be done with careful identification of the specific goal that cannot be attained by other conventional plasma processes. The major features of the process that could distinguish LPCAT processes from other conventional plasma processes are:

1. The mode of creation of polymerizable or chemically reactive species, i.e., reactions caused by the interaction with excited neutrals of the carrier gas.
2. The short kinetic path length because of the high one-directional transport velocity of polymerizable species.

Whether or not these characteristic features are an advantage or a disadvantage entirely depends on the objectives to be accomplished by the processing. These two characteristic features, however, indicate that LPCAT process is better suited for the surface treatment rather than the surface coating.

The second feature seems to be responsible for the fact that the corrosion protection characteristics of

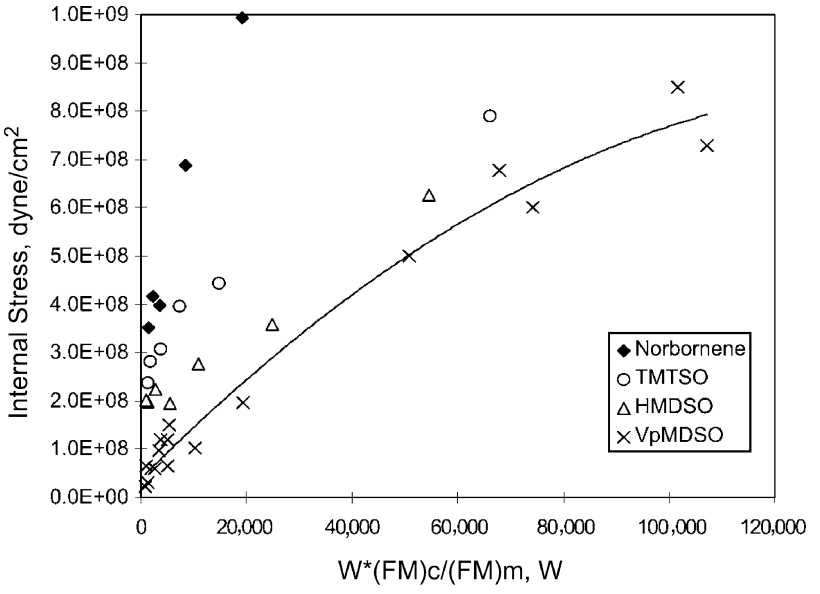


Fig. 16 Dependence of internal stress on the energy input parameter, $W^*(FM)_c/(FM)_m$, for cascade arc torch polymerization.

LPCAT coatings do not match that which can be attained by DC cathodic polymerization. The very high one-directional transport velocity hinders the interaction between polymerizable species existing in the luminous gas phase before they reach the target surface. The velocity of gaseous species leaving the nozzle of the cascade arc generator is estimated to be in the supersonic range, and LPCAT jet is close to the beam of excited species. The short kinetic path length means that smaller or oligomeric species deposit on the substrate. Consequently, the barrier characteristics and the adhesion of the coating become poorer than those that could be obtained by conventional plasma polymerization with a higher degree of kinetic path length. On the other hand, the same feature provides a great advantage when larger species should be avoided, e.g., in the case of deposition onto the surface of nanoparticles, which do not have large enough surfaces to accommodate larger species.

The first feature that chemically reactive species are created by the interaction of molecules (in gas phase or on the surface) with excited neutral species of Ar has a very significant influence in the surface modification of polymers. When a polymer surface is exposed to Ar or O_2 plasma, the energetic ions and electrons (at the level of ionization energy) bombard the surface. The influence on the surface is determined by the energy level. The chemical bonds involved in the molecules that

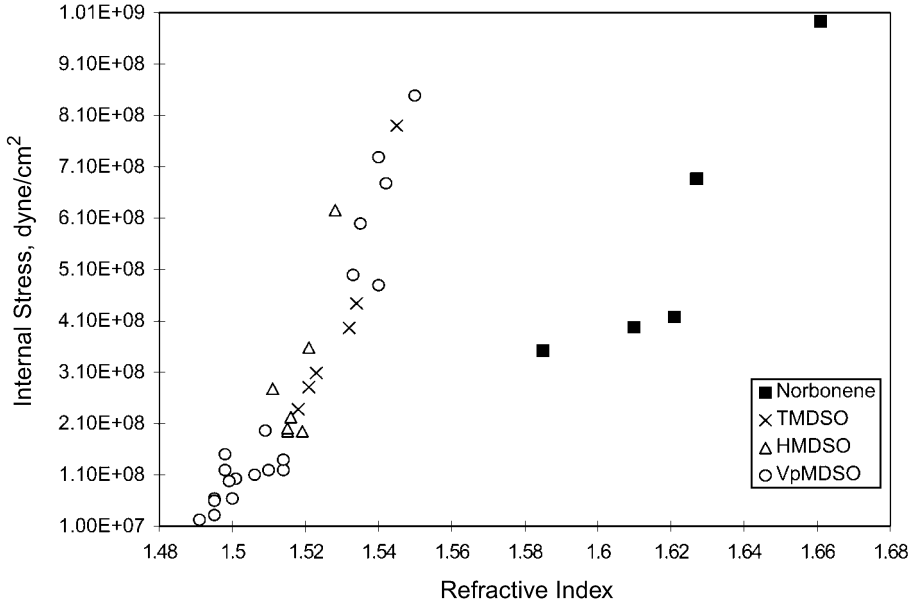


Fig. 17 Qualitative correlation of internal stress with refractive index of cascade arc torch plasma polymer films.

constitute the surface are relatively low (2–3 eV) compared to the ionization energy of the gas used in plasma (over 10 eV).

The cession of a σ-bond yields two free radicals. The free radical on the surface subsequently reacts with ambient oxygen when the treated substrate is exposed to air rendering the surface hydrophilic (in the case of Ar plasma). The high energy of impinging entities (electrons and/or ions) tends to yield excessive cession of bonds, which creates a weak boundary layer below the top surface. This situation could be visualized by the trends that plasma treatment of hydrophobic surface makes the surface paintable, but the paint does not adhere well because the paint could delaminate through the weak boundary sublayer.

In Ar LPCAT treatments, the excited neutrals of Ar, instead of Ar^+ and electrons with over 10 eV, interact with polymer molecules at the surface. The energy level of those excited species are nearly as high as those for ions and electrons; however, the interaction is through the energy transfer process in which the energy matching principle plays an important role. Consequently, the indiscriminate bombardment of highly energetic species does not take place in LPCAT treatment, which leads to a much less damaging surface treatment of polymers. Ar LPCAT treatment of thermoplastic olefins (used as the fascia of automobile bumper) yields much more durable and stable adhesion of paint applied on the treated surface than conventional plasma treatment processes could render.[18] Ar LPCAT treatment of fibrous polypropylene reinforcing materials for concrete also shows the much-improved energy absorbing characteristics of the fiber reinforced concrete by the same principle.[19] Trimethylsilane coating on zirconium oxides, which are used as the radio-opaque pigments in PMMA bone cement recipe, create bonding between the pigments and the polymer matrix and improves the fatigue life of the bone cement.[20]

CONCLUSIONS

Low-pressure cascade arc torch has the unique feature that excited neutral species of Ar that are created in the cascade arc generator are injected into the reaction chamber. The beams of Ar excited species can be used: 1) to cause chemical vapor deposition of the second gas that is injected in the expansion chamber or 2) to modify the surface of substrate, particularly polymers, by letting the excited Ar species react directly with the surface.

Because of the high rate of one-directional movement of the excited species of Ar, the chemical vapor deposition does not yield as well-developed a layer as that which can be obtained with plasma polymerization coating. On the other hand, the direct application of the excited species of Ar on polymeric surface yields excellent surface modifications without damaging the substrate polymers.

REFERENCES

1. Kroesen, G.M.W. Ph.D. thesis, Eindhoven University of Technology. Eindhoven, The Netherlands, 1988.
2. Beulens, J.J. Ph.D. thesis, Eindhoven University of Technology. Eindhoven, The Netherlands, 1992.
3. Beulens, J.J.; Kroesen, G.M.W.; Schram, D.C.; Timmermans, C.J.; Crouzen, P.C.N.; Vasmel, H.; Schuurmans, H.J.A.; Beijer, C.B.; Werner, J. J. Appl. Polym. Sci. Appl. Polym. Symp. **1990**, *46*, 527.
4. Kroesen, G.M.W.; Schram, D.C.; van de Sande, M.J.F. Fast deposition of amorphous hydrogenated carbon films using a supersonically expanding arc plasma. Plasma Chem. Plasma Process. **1990**, *10*, 49.
5. De Graaf, M.J.; Dahiya, R.P.; Jauberteau, J.L.; De Hoog, F.J.; van de Sande, M.J.F.; Schram, D.C. Colloq. Phys. **1990**, *51*, 5–387.
6. Buuron, A.J.M.; Otorbaev, D.K.; van de Sanden, M.C.M.; Schram, D.C. Absorption spectroscopy on the argon first excited state in an expanding thermal arc plasma. Phys. Rev. E **1994**, *50*, 1383.
7. Maecker, H. Naturforsh. Z. **1956**, *11a*, 457.
8. Fusselman, S.P; Yasuda, H. An overall mechanism for the deposition of plasma polymers from methane in a low-pressure argon plasma jet. Plasma Chem. Plasma Process. **1994**, *14*, 251.
9. Yu, Q.S.; Yasuda, H.K. Deposition behavior in a low-temperature cascade arc torch (CAT) polymerization process. J. Polym. Sci. A. Polym. Chem. Ed. **1999**, *37*, 967–982.
10. Yasuda, H. *Plasma Polymerization*; Academic Press: San Diego, CA, 1985.
11. Yu, Q.S.; Yasuda, H.K. An optical emission study on expanding low-temperature cascade arc plasmas. Plasma Chem. Plasma Process. **1998**, *18* (4), 461–485.
12. Ricard, A. *Plasma-Surface Interactions and Processing of Material*; Auciello, O., Ed.; Kluwer Academic Publisher: The Netherlands, 1990.
13. Acton, J.R.; Swift, J.D. *Cold Cathode Discharge Tubes*; Academic Press Inc.: New York, 1963.
14. Krogh, O.; Wicker, T.; Chapman, B. The role of gas phase reactions, electron impact, and collisional energy transfer processes relevant to plasma etching of polysilicon with H_2 and Cl_2. J. Vac. Sci. Technol. **1986**, *A4* (3), 1796.

15. Herzberg, G. *Molecular Spectra and Molecular Structure, I. Spectra of Diatomic Molecules*; Van Nostrand Reinhold Co.: New York, 1950.
16. Krentsel, E.; Fusselman, S.P.; Yasuda, H. Penetration of plasma surface modification, II. CF4 and C2F4 low-temperature cascade arc torch. J. Polym. Sci. A Polym. Chem. Ed. **1994**, *32* (10), 1839.
17. Yasuda, T.; Okuno, T.; Miyama, M.; Yasuda, H. Penetration of plasma surface modification. I. Cf4 and C2F4 glow discharge plasmas. J. Polym. Sci. A: Polym. Chem. **1994**, *32*, 1829.
18. Lin, Y.-S.; Yasuda, H.K. Low-temperature cascade arc torch treatments for enhanced adhesion of primer to thermoplastic olefins. J. Appl. Polym. Sci. **1998**, *67*, 855–863.
19. Zhang, C.; Gopalaratnam, V.S.; Yasuda, H.K. Plasma treatment of polymeric fibers for improved performance in cement matrices. J. Appl. Polym. Sci. **2000**, *76*, 1985–1996.
20. Kim, H.Y.; Yasuda, H.K. Improvement in fatigue properties of PMMA bone cement by means of plasma surface treatment of fillers. J. Biomed. Mater. Res. (Appl. Biomater.) **1999**, *48*, 135–142.

Lubrication Performance Factors for Chemical Process Plant Machinery

Jim C. Fitch
Drew D. Troyer
Noria Corporation, Tulsa, Oklahoma, U.S.A.

INTRODUCTION

The need for a maintenance organization to deliver machine reliability at low cost is in continuous demand today. With increasing frequency, lubrication programs are successfully accomplishing this feat by interlacing modern strategies and technologies with amazing skill and dexterity. Unfortunately, complacency is occurring among other organizations in bringing about necessary change to capitalize on opportunities in lubrication to reduce operating costs and production losses.

This entry addresses lubrication in the context of a modern maintenance organization in the chemical process industry. Less emphasis is placed on the selection of lubricants as this subject is expansive and exceeds the scope intended for this entry. Instead, emphasis will be on the activity of lubrication, which falls in the domain of those charged with the responsibility of maintenance and machine reliability. Lubrication deals with stability of the lubricant within the machine in controlling friction, wear, and heat generation, among other things.

LUBRICANT SELECTION

Many books have been published on the selection of lubricants for the numerous machine applications requiring their use. As stated above, it is not the objective of this entry to offer engineering advice on lubricant selection. However, owners and maintainers of equipment in process plants should have a basic understanding of the function of a lubricant. If a current lubricant fails to adequately serve these functions, changing the lubricant may be necessary in order to achieve the desired outcome.

The following are the ways that lubricant formulations may be changed to better serve a particular application:

Base Oil Chemistry: The base oil is the foundational building block of a lubricating oil or grease. The base oil can be derived from various sources including petroleum, oil seeds (vegetable), or synthetics. There are many different options to consider within each of these groups. These choices influence the physical and chemical properties of lubricants as well as the performance in a given condition.

Viscosity and Viscosity Index: Viscosity is the most important physical property of a lubricant. It provides strength for the base oil load bearing. The wrong viscosity, too high or too low, can have devastating consequences in terms of machine reliability. Viscosity index (VI) characterizes the degree to which viscosity changes in relation to temperature change.

Additives: The majority of lubricants on the market today employ the use of additives to enhance, suppress, or alter the properties of the base oil. Numerous additives serve varying formulation objectives, including:

- Oxidation inhibitors
- Pour point depressants
- Corrosion inhibitors
- Foam suppressants
- VI improvers
- Dispersants
- Detergents
- Metal deactivators

Grease Thickeners: Grease has nearly identical properties to lubricating oils, with the exception of the inclusion of a thickener, which gives it a solid or a semisolid consistency. Thickeners are often made of soaps of lithium, sodium, aluminum, or calcium. Like base oils and additives, thickeners possess a myriad of positive and detracting lubricating properties.

By matching base oil type, viscosity, VI, additives, and thickener type (for grease) to the lubrication objectives, the optimum composite performance can result. The primary functions of a lubricant in typical machinery applications include the following:

Friction Control: Friction is the enemy of machine reliability and efficient operation. The lubricant protects machine surfaces through either physical separation by viscous forces or by chemical action whereby additives form a ductile lubricating surface film.

Wear Control: Reducing friction and surface contact minimizes the rate of wear and distress to

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030626

load-bearing machine surfaces. The rate of wear and life expectancy of a machine and its components is largely influenced by the quality of lubrication.

Corrosion Control: Surfaces that become wetted by a lubricant and its additives are typically much less prone to corrosive damage from water, acids, bacteria, and other similar corrosion agents. Additives can neutralize acids as well as form a barrier film, which repels water and other chemically aggressive contaminants.

Temperature Control: In addition to reducing the generation of heat through friction reduction, the lubricant also serves to transport thermal energy from locations where it is generated to locations where it can dissipate by conduction or convection, typically through a heat exchanger.

Contamination Control: Lubricating oils collect contaminants, such as particles and moisture, and transport them to settling tanks, centrifugal separators, filters, etc. Some lubricants, such as crankcase oils, are formulated with additives to disperse contaminants that reduce the risk of deposit formation, corrosion, or premature filter plugging.

Power Transmission: Hydraulic systems utilize lubricating oils to serve all of the above listed functions plus the transfer of force and motion based on the fluids with general incompressibility.

In selecting a well-suited lubricant for a particular application, dozens of factors must be considered relating to load, speed, duty cycle, machine metallurgy, machine type, relubrication interval, lubricant application method, thermal range, contamination, reliability objectives, safety, and environmental issues.[1] Guidance from the original equipment manufacturer (OEM), industry design standards, lubricant suppliers, and specialists in tribology (study of friction and wear) and lubrication is generally well advised.

GREASE AND OIL APPLICATION METHODS

Once the lubricant type has been selected (base oil, viscosity, and additives), a decision must be made regarding the use of either oil or grease. Here, several more factors come into play. Questions that must be considered include:

- What are bearing speeds and loads?
- What is the operating temperature range for the lubricant?
- What type of bearing designs and sizes are used?
- Will the (lubricated) components operate for extended periods without interruption?
- Will the components be accessible during operation?
- Will the process allow for access during operation?
- Are the components exposed to the environment? Can the components be isolated from the environment?
- Are the components intended to be sacrificial (changed frequently)?
- Are the components highly specialized and sufficiently expensive that extra care must be provided to assure maximum protection available from the lubricant?
- Is there risk of the lubricant leaking onto or into the product being produced (or raw materials)?
- Will the lubricant need to be continuously filtered or dehydrated?
- Will the lubricant need to provide heat transfer?
- Will the lubricant need to seal out the possible ingress of contamination?

The reality for the practitioner and equipment maintainer is that the decision regarding the use of oil or grease and the application method is made well before the equipment reaches the plant floor. Still, for many reasons, lubricant selection and the method of application need to be changed or adjusted to best achieve reliability objectives.

Grease Application Methods

A high percentage of bearings and some gearboxes used in industry are only lubricated once, when the machine or component is built. These lubed-for-life machines are generally low duty with limited operational criticality. Service life varies, but 3–6 yr is typical. The cost of relubrication to extend the service life in such applications is found to exceed the value gained. For example, many companies do not relubricate small electric motors of less than 10–20 hp.[1] These motor bearings are typically sealed to control lubricant leakage and contaminant ingress.

Relubrication of larger bearings and gears can often extend service life by two to four times. However, with the practice of relubrication there are always risks or issues that must be considered. These include:

- Cost to store and handle lubricants
- Labor cost to relubricate machines
- Risk of introducing a wrong or an incompatible lubricant
- Risk of overlubricating or underlubricating the machine
- Safety risk associated with the lubrication technician working near operating equipment

Table 1 shows the advantages and disadvantages of common grease application methods.

Table 1 Advantages and disadvantages of common grease application methods

Grease application method	Advantages	Disadvantages
Grease gun (manual)	Simple to use, low front-end cost, technician can inspect machine	High labor cost, long intervals between relubrication, overgreasing a common problem, safety concerns
Grease cups and single-point lubricators—automatically apply grease slowly over a 3–12 month period	Enables lubrication in remote or restricted access areas, reduced labor cost, reduced grease consumption, increased machine reliability	Disadvantages depend on design and may include inconsistent grease delivery rate, temperature limitations, vibration influences, high costs, plumbing and installation requirements
Centralized automated multipoint lubrication system	Low labor cost, positive displacement delivery, continuous lubricant delivery	High hardware and installation costs, occasional reliability and maintenance problems

Oil Application Methods

Unlike grease, the options for introducing oil to a machine are much more varied and machine type dependent. In general, the larger and more critical the machine is to process operation, the more likely the use of oil, instead of grease, is the best option. Among other things, the use of oil facilitates the following capabilities:

- Ability to filter or separate dirt, water, sludge, and other contaminants
- Heat transfer through the use of heaters or coolers to stabilize lubricant and machine temperature
- Lubricant volume control through oil level gauges and constant level devices
- Simplified and more effective lubricant sampling for periodic analysis
- Ability to circulate the lubricant to insure homogeneous properties

While it is often believed that it is easier to grease a bearing than to reoil, the reliability of lubricated components is often enhanced when oil is the lubricant of choice. Table 2 describes the advantages and disadvantages of common oil application methods.

CONTAMINATION CONTROL

Contamination can be defined as any unwanted substance or energy that enters or contacts the oil. Contaminants come in many forms and may be highly destructive to the lubricant, its additives, and machine surfaces. It is often overlooked as a source of failure because its impact is usually slow and imperceptible. While it is impractical to attempt to completely eradicate contamination from in-service lubricants, control of contaminant levels within acceptable limits can be accomplished and is vitally important.

Particles, moisture, soot, heat, air, glycol, fuel, detergents, and process fluids are all contaminants commonly found in industrial lubricants and hydraulic fluids.[2] However, particle contamination is widely recognized as the most destructive contaminant to the oil and machine. This is the reason why the particle counter is the most widely used instrument in oil analysis today. The central strategy to its success in reducing maintenance costs and increasing machine reliability is proactive maintenance.

Proactive and Predictive Maintenance

While the benefits of detecting abnormal machine wear or an aging lubricant condition are important and frequently achieved with oil analysis programs, they are of low importance when compared to the more rewarding objective of failure avoidance. This is achieved by treating the causes of failure, not just the symptoms. Also, it is the foundation of the popular practice known as proactive maintenance.

Whenever a proactive maintenance strategy is applied, three steps are necessary to ensure that its benefits are achieved. Since proactive maintenance, by definition, involves continuous monitoring and controlling of machine failure root causes, the first step is simply to set a target, or a standard, associated with each root cause. In oil analysis, the most important root causes relate to fluid contamination (particles, moisture, heat, coolant, etc.).

However, the process of defining precise and challenging targets (for example, high cleanliness) is only the first step. The second step is achieving and sustaining control of the fluid's conditions within these targets. This often includes an audit of how fluids become contaminated and then systematically eliminating these entry points. Better filtration and the use of separators are often required.

The third step is the vital action element of providing the feedback loop of an oil analysis program.

Table 2 Advantages and disadvantages of common oil application methods

Application method	Advantages	Disadvantages
Manual methods using hand pad, brush or spray bottle	Low upfront cost, generally easy to apply, machine can be inspected by technicians	High labor costs, excess leakage, over supply immediate after application, difficult to perform on the run, safety risks, contamination risks
Drip feed devices (drop feed, wick feed, etc.)	Simple to use, no moving parts, continuous supply, variable feed rate control, easy to inspect and refill supply bowls	Contaminants can interfere for feed control, viscosity/temperature influences on feed rate, on control of contaminant ingression
Constant-level oilers	Better oil level management, low labor cost, better contamination control	Controlling contamination to the wrong oil level, defective level adjustment devices, tendency for feed supply tube to clog
Oil lifter devices, including oil rings, collars, slingers, paddle gears, and flingers	Machine self-lubricates without additional external hardware, low labor costs	Some devices are highly sensitive to oil level, oil viscosity/temperature, foaming conditions, wear, and operating speed. Also, dry start risks, sediment buildup risks
Bath and splash lubricated machines (typically gearing and bearings)	Machine self-lubricates without additional external hardware, low labor costs	Some are highly sensitive to oil level, oil viscosity/temperature, foaming conditions, and operating speed. Also, dry start risks, sediment buildup risks
Spray and mist lubrication systems–machines being lubricated by a directed spray of oil or oil mist	Low risk of contamination, no oil level control risks, improved machine reliability	High cost of hardware and installation, viscosity/temperature limitations, application limitations
Wet-sump and dry-sump circulating oil systems	Excellent cooling capabilities, enables use of filters and separators, multiple lube points can be supplied at once, low labor cost, the most representative oil sampling, limited oil volume control risk	Expensive hardware installation, leakage risks, large volume of oil required, possible aeration and foaming risks

When exceptions occur (for example, over target results), remedial actions can then be immediately commissioned. Using the proactive maintenance strategy, contamination control becomes a disciplined activity of monitoring and controlling high fluid cleanliness, not a basic activity of trending dirt levels.[3]

Finally, when the life extension benefits of proactive maintenance are flanked by the early warning benefits of predictive maintenance (PMs), a comprehensive condition-based maintenance program results. While proactive maintenance stresses root-cause control, PMs targets the detection of incipient failure of both the properties of the fluid and machine components such as bearings and gears. It is this unique, early detection of machine faults and abnormal wear, which is frequently referred to as the exclusive domain of oil analysis in the maintenance field (Fig. 1).[4,5]

Managing Particle Contamination

There is no single property of lubricating oil that challenges the reliability of machinery more than suspended particles. Very small particles can ride in oil almost indefinitely and, because they are not as friable (easily crumbled) or filterable as larger particles, the destruction can be continuous. Many studies have proven that greater damage is associated with small particles (Fig. 2).[5] Still, most maintenance professionals have misconceptions about the size of particles and the associated harm caused.

These misconceptions relate to the definitions applied to clean oil and dirty oil. It is this definition that influences the setting of appropriate target cleanliness levels for lubricating oils and hydraulic fluids. While there are numerous methods used to arrive at target cleanliness

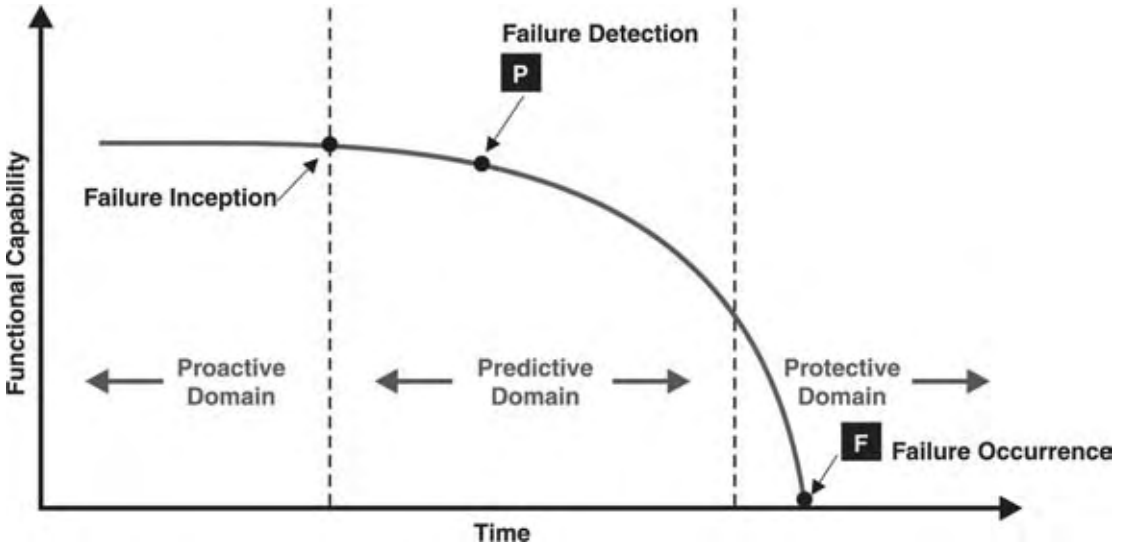


Fig. 1 Condition monitoring domains in the PF interval curve. *(View this art in color at www.dekker.com.)*

levels for oils in different applications, most combine the importance of machine reliability with the general contaminant sensitivity of the machine to set the target.

There are many expensive ways to achieve clean oil but experience has taught us the wisdom of contaminant exclusion—treating the cause not just the symptom. By effectively excluding the entry of contaminants and promptly removing contaminants if they do enter, the new cleanliness targets are frequently achieved. Concerns that filtration costs will increase are not often broached due to the greater overall control, especially from the standpoint of particle ingression.

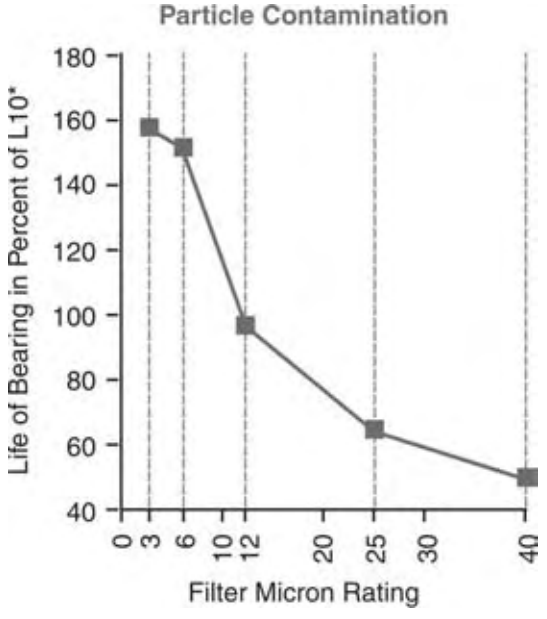


Fig. 2 Particle contamination vs. bearing life. *(View this art in color at www.dekker.com.)*

Managing Moisture Contamination

Moisture, when it contaminates hydraulic and lubricating oils, has a degrading effect on both the lubricant and the machine. For instance, when bearing lubricants are contaminated with less than 0.1% water, more than 75% of the bearings design life can be lost (Fig. 3).[6] Some additives adsorb to the water and are removed when the water separates from the oil while others are destroyed by water induced chemical reactions. Water also promotes oxidation of the oil's base stock, causes rust and corrosion of machine surfaces, and reduces critical, load-bearing film strength.[7] In summary, water represents a serious risk to equipment and should be aggressively controlled.

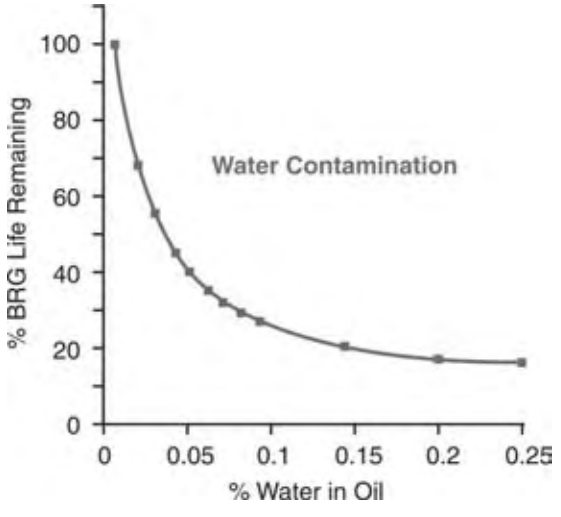


Fig. 3 Moisture contamination vs. bearing life. *(View this art in color at www.dekker.com.)*

Water coexists with oil in either a dissolved state or a free state. In a dissolved state, single water molecules are distributed throughout the oil due to the water's chemical attraction to the fluid. Numerous factors such as viscosity, base-stock type, base-stock condition, impurities, and additive package determine the volume of water that the oil will dissolve. The dissolved volume is a function of the oil's temperature; thus the ''humidity'' is reported as relative humidity, depending upon the temperature. If the oil has dissolved all the water it can at a given temperature, it is saturated. The dissolved water is very difficult to control and does only minimal harm to the machine and oil.[8] Free and emulsified water pose the greatest risk to the machine and they, besides the lubricant, should be placed under strict control.[9]

Along with particle contamination, it is important to monitor the presence of water in lubricating oils. Various methods are available to accomplish this including many simple field techniques. Once an abnormal level of water is observed, corrective action should be performed to dehydrate or change the oil.

Managing Air Contamination

Air may not immediately be thought of as a contaminant, but the presence of air in its various forms may have an impact on the ability of the lubricant to perform its design function. Almost all lubricating oil systems contain some air. Air is found in four phases: free air, dissolved air, entrained air, and foam. Free air is trapped in a system, such as an air pocket in a hydraulic line, and may have minimal contact with the fluid. It can be a contributing factor to other air problems when lines are not bled properly during equipment startup and free air is drawn into circulating oils.

Dissolved air is not readily drawn out of solution. It becomes a problem when temperatures rise rapidly or pressures drop. Petroleum oils contain as much as 12% dissolved air. When a system starts up or when it overheats, this air changes from a dissolved phase into small bubbles. If the bubbles are very small in diameter, they remain suspended in the liquid phase of the oil, particularly in high viscosity oils. This can cause air entrainment, which is characterized as a small amount of air in the form of extremely small bubbles dispersed throughout the bulk of the oil. Air entrainment is treated differently than foam and is typically a separate problem. Some of the potential effects of air entrainment include pump cavitation, spongy and erratic operation of hydraulics, loss of precision control, vibrations, oil oxidation, component wear due to reduced lubricant viscosity, equipment shutdown when low oil pressure switches trip, microdieseling due to the ignition of the bubble sheath at the high temperatures generated by compressed air bubbles, safety problems in turbines (if overspeed devices do not react quickly enough), and loss of head in centrifugal pumps.

Foam, on the other hand, is a collection of closely packed bubbles surrounded by thin films of oil that float on the surface of the oil. Generally, foam is cosmetic, but it must be treated if it makes oil level control impossible, if it spills onto the floor to create a safety or a housekeeping hazard, causes air locks at high points, or is so extreme that the equipment is lubricated with foam. Small amounts of foam do not necessarily need to be treated unless the system suffers from the air entrainment conditions listed here, although the presence of the foam may be symptomatic of a more serious problem.[10–12]

Managing Soft Contaminants (Oxides, Sludge, and Varnish)

The products formed from the oxidation of oil include weak carboxylic acids, sludge, and varnishes. The weak acids can be titrated by acid number (AN) testing and, when quantified, give an indication of the degree of oxidation of the oil. The approach of detecting and analyzing sludge and varnish problems in machinery is not the same as that used in oil analysis. In many instances this is because the evidence is not always in the oil. The sludge and varnish should be analyzed directly, using a different set of tests and evaluation parameters. Still, used oil analysis plays an important diagnostic role in helping to reveal candidate causes, as well as to rule out others.

The conditions that most commonly lead to sludge and varnish problems vary, which complicates the process of identifying the root cause analytically. There are at least 25 unique lubricant degradation mechanisms leading to sludge or varnish formation, including:

- Aeration of the fluid
- Sparking from static electricity
- Bulk thermal degradation
- Antifreeze contamination
- Soot coagulation
- Bulk oil oxidation
- Hydrolysis
- Prolonged cold storage
- Grease-contaminated oil
- Caustic detergent contamination
- Nitration
- Coking on hot surfaces
- Radiological contamination
- Poor engine combustion efficiency and blow-by
- Highly aromatic fuels

- Sulfation (fuel, H_2S, etc.)
- Lead corrosion reactions
- Reactive compressor gases
- Additive incompatibilities
- Base oil incompatibilities

Lubricants degrade in different ways and the products of this degradation are essentially referred to as sludge and varnish. These products are generally unstable in the oil and look for a place to deposit themselves. In certain instances, the deposits form on machine surfaces at the exact location where the oil has degraded, for example, hot surface coking. In other cases, the oil degrades in one location but deposits condense on a surface elsewhere.

The deposits that form on machine surfaces interfere with the fluid performance and the machine's mechanical movements. They can also contribute to wear and corrosion, or simply cling to surfaces. For example, deposits on the spool of a servo control valve can tighten the interference fit between the spool and the bore. Compounding this are the adherence properties of varnish, which can stick particles from the oil to silt lands, thus leading to common silt-lock valve failure. Other types of sludge and varnish-type failures include plugged orifices, damaged mechanical seals, plugged discharge ports on compressors, journal-bearing failure, premature plugging of oil filters, and diesel engine combustion-zone wear.[13–17]

Contaminant Exclusion and Removal

In many machines, the exclusion of contamination is the only way to control contamination. This is because these machines either do not have a filter or the filter in use is coarse, providing no practical protection in the particle size range of critical oil films. When contaminants are not removed by filtration or separators, a lubricant's contaminant level equals the machine's service hours multiplied by the contaminant amount ingressed per hour (ingression rate). For machines exposed to high ambient dust, particle counts can exceed recommended levels in just a few hours.

Even machines with good filters and separators are faced with ingression challenges. To maintain contaminant levels within targets, the filter must remove contamination at a rate equal to the ingression rate (mass balance). The lower the target cleanliness level, the more difficult this becomes. This is because fluids must stay within these high cleanliness targets, but by definition, particles are not densely packed in the oil but rather are sparsely distributed. This means that for every gallon of fluid that enters the filter, there are few particles from that gallon that are available to remove. Yet the filter must still remove particles at a rate equal to the ingression rate, otherwise the contaminant level will rise. This places increasing demand on the quality and capture efficiency of the filter (percent particles removed above a certain size).

The flow rate of the oil entering the filter is also a strong influence. The flow provides the necessary conveyance of particles to the filter. If flow rates are low, filters with even 100% capture efficiency cannot remove enough particles to keep up with ingression, causing contaminant levels to exceed targets. The higher the target cleanliness, the higher the minimum required flow rate for a given filter. After all, a filter can only remove particles it comes in contact with. Some by-pass and kidney loop (off-line) filters have flow rates that are

Table 3 Capabilities of filters and separators in removing contaminants of various types

	Air	Solids	Dissolved H_2O	Emulsified H_2O	Varnish and acids	Free H_2O
Absorption–polymeric adhesion				•		•
Air stripping gas injection	•		•	•		•
Mechanical						
Wire mesh		•				
Paper		•				
Microglass		•				
Cellulose		•				
Electrostatic–separation		•			•	
Centrifugal–radial acceleration		•		•		•
Coalescing separators				•		•
Distillation–thermal and vacuum	•		•	•		•
Settling–force of gravity	•	•				•
Magnetic separation		•				
Adsorptive media			•	•	•	

inadequate for stabilizing target cleanliness levels. Table 3 shows the capabilities of filters and separators in removing contaminants of various types.[18]

It is often said that the cost of excluding a gram of dirt is only about 10% of what it will cost once it enters the oil. Dirt places stress on additives, the base oil, and machine surfaces. Therefore, the cost to filter a gram of dirt from the oil is much higher than that of filtering a gram of dirt from the air intake/breather.

The word ingression refers to the introduction of particles into lubricants and hydraulic fluids regardless of the source (external or internal). Fig. 4 organizes common sources into three subcategories: built-in, ingested, and generated. Depending on the nature of the machine, the ingression rate and sources may vary considerably. For clean-environment indoor equipment, the primary sources can be from process fluids and internal generation (wear, corrosion, etc.).[19]

For many machines, reducing ingression means reducing top-end ingression of the particles entering through fill ports, vents, breathers, hatches, inspection ports, shaft seals, and other headspace openings. There are numerous ways to control top-end ingression in reservoirs, sumps, gearboxes, and bearing housings, as described below.

Purge methods

This involves the introduction of a clean gas or an aerosol into the headspace of the machine. A slight positive pressure is maintained to prevent the entry of ambient air. Examples include instrument air purge, oil mist purge, and nitrogen purge.

Isolation methods

Expansion chambers, piston/cylinder reservoirs, and bladders have been used to isolate headspace air from ambient air to prevent contamination. One disadvantage is that original moisture (humid air) is often unable to escape from the headspace, causing moisture to lock into the oil. In some cases, users have reported that this has led to heavy corrosion.

Filter breathers

If reservoirs and sumps can be tightly sealed so that all air exchanged between the atmosphere and the headspace can be directed through a single port, then high-quality filter breathers can be used to remove dust from incoming air at that port (vent). The quality of the filter (capture efficiency) should be no less than that of the oil filter in use.

LUBRICANT SAMPLING AND ANALYSIS

Lubricant sampling and analysis is a broad field comprising the activities of monitoring, reporting, and responding to information obtained from the analysis

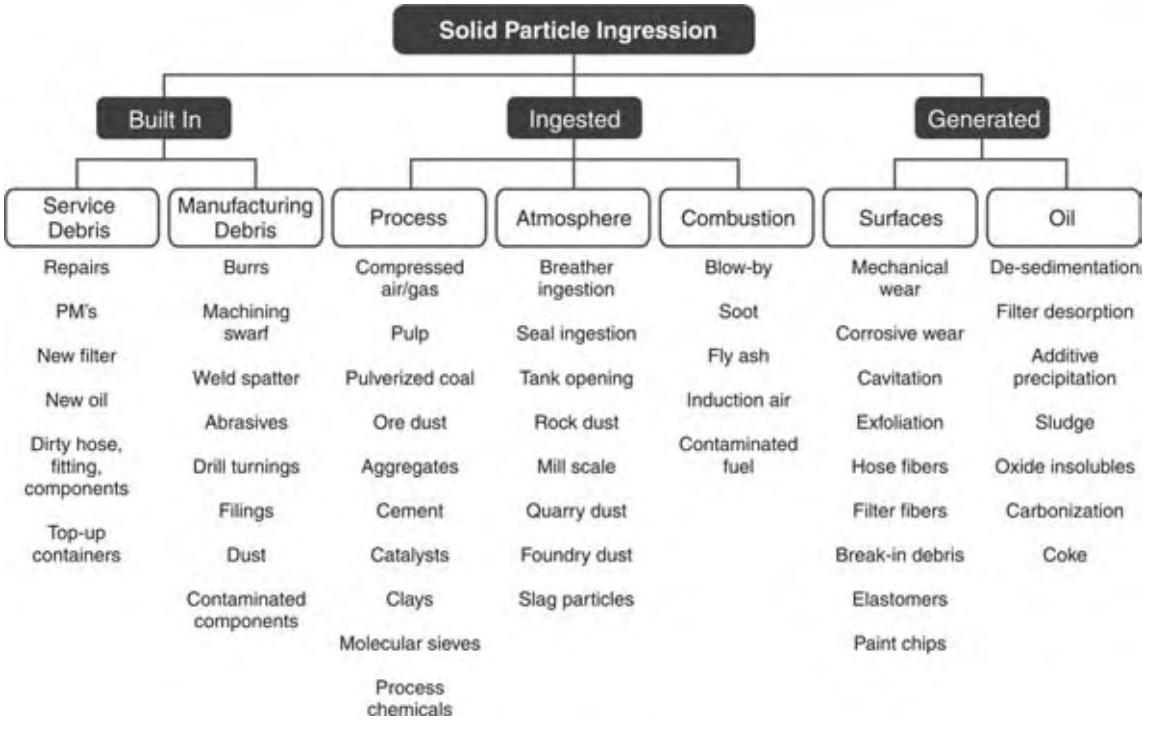


Fig. 4 Sources of particle contamination. *(View this art in color at www.dekker.com.)*

of grease, lubricating oils, and hydraulic fluids. Because the practice also encompasses the analysis of used grease and hydraulic fluids, this discussion will generically refer to the entire practice as ''lubricant analysis.'' Greases, hydraulic fluids, synthetic fluids, and lubricating oils are all lubricants and can be generally analyzed by similar methods.

Lubricant analysis involves numerous activities with the following objectives:

- Optimize machine reliability, productivity, and profitability.
- Reduce maintenance and operating costs.
- Minimize safety risks or related hazards.
- Reduce friction, heat, and energy consumption.
- Protect the environment, including air and water.
- Reduce the consumption of lubricants.

Many different types of conditions can be analyzed and reported through lubricant analysis. Commonly reported conditions include:

- Quality and condition of new oil deliveries.
- Distressed, degraded, or noncomplying in-service lubricant properties.
- Lubricant contamination.
- Abnormal root cause or stressing condition.
- Wear debris composition and physical characteristics.
- Leakage.
- State of lubricants in storage.

The methods of gathering and analyzing lubricant properties and conditions vary widely. Analytical instruments and procedures are often employed in a laboratory to determine important properties of routinely sampled lubricants. Similar instruments can be used remotely in the plant. This practice is often referred to as ''onsite analysis.'' In certain cases the instruments or sensors can be used in real time, dedicated to a specific machine and fluid. The management and reporting of lubricant analysis data is typically conducted with the aid of a computer and software developed for this purpose.

Objectives and Benefits of Lubricant Analysis

Lubricant analysis benefits the organization in several ways before, during, and after the occurrence of machine failure. Before the onset of failure, lubricant analysis confirms that the machine contains the correct lubricant and that the lubricant is physically and chemically fit for service. It also confirms that contamination levels are within tolerable limits. Lubricant analysis can be an important tool in the efforts to ensure the quality of new lubricant deliveries, storage effectiveness, and reclamation activities. Lubricant analysis helps to verify that lubricants are fit for service. These are proactive applications of lubricant analysis that have the potential to extend the life of the machinery.

When a machine fails, wear debris, which is detectable using lubricant analysis, is often produced in advance of any observable operational deterioration (Table 4). The early warning of lubricant analysis provides options for the maintenance professional. Also, lubricant analysis plays a key role in diagnosing problems once they are detected. When a problem is detected, operations personnel can inquire about the projected time interval to failure or the probability that the machine will last until a defined date. Failure prognostics, the process of estimating residual machine or lubricant life, is generally the least precise of the condition-monitoring activities, although it can still provide useful insight. Prognostic forecasts are improved when multiple data parameters are measured and defined.

When a machine fails, users analyze the oil from the failed machine. The oil often contains the remnants of the failure. This information can be used for the deployment of root-cause control measures to prevent similar failures in the future.

Strategic alignment of lubricant analysis to organizational objectives often begins with an audit of the program's component parts. The audit should identify the program's strengths and weaknesses and evaluate how resources are allocated. Often, the audit reveals that a machine was brought into a lubricant analysis program for the wrong reasons. The decision to include a machine for routine lubricant analysis must be based upon the effect a failure will have on production, safety or environmental compliance, the cost to repair or restore the equipment, and/or the likelihood that a failure will occur.

In some cases, the desire to avoid an unnecessary oil change or simply to extend oil drain intervals is an important objective, resulting in reduced lubricant consumption and associated costs. Mission criteria should also influence those machines that should be sampled, the frequency of lubricant analysis, and the types of tests that are conducted. Some machines are more operationally critical than others. They deserve more resources from maintenance technologies. The likelihood that failure will occur and specific challenges (high temperature, risk of contamination, etc.) faced by the lubricant are also factors that influence the type of lubricant analysis and the sampling frequency decision.

Well-implemented lubricant analysis can help optimize the use of maintenance and reliability resources to achieve improved machine reliability and reduced operating and maintenance costs. In certain hazardous

Table 4 Detection using lubricant analysis

	Root-cause detection	Incipient fault detection	Problem diagnosis	Failure prognosis	Post mortem
What oil analysis tells	When something occurs that can lead to failure—root cause conditions	When an early-stage fault exists that is otherwise going unnoticed; for example, abnormal wear	What is the nature of observed problem? Where does it come from? How severe is it? Can it be fixed?	That a machine is basically worn out and needs to be fixed or replaced	What caused the machine to fail? Could it have been "avoided''?
What you monitor	Particles, moisture, viscosity, temperature, additives, oxidation, AN/BN, soot, glycol, FTIR, RPVOT	Wear debris density, temperature, particle count, moisture, elemental analysis, viscosity, analytical ferrography	Wear debris, elemental analysis, moisture, particle count, temperature, viscosity, analytical ferrography, vibration analysis	Elemental analysis, analytical ferrography, vibration analysis, temperature	Analytical ferrography, ferrous density, elemental analysis
Maintenance mode	Proactive	Predictive	Predictive	Run-to-failure	Run-to-failure
Relative savings[a]	10	6	3	2	1

[a]10 = High, 1 = Low.

machinery applications, lubricant analysis can help ensure the safety of machinery. When combined with optimum lubrication practices, including the application and selection of lubricants, routine lubricant analysis can reduce energy consumption and waste stream emissions.

Oil Sampling

Oil sampling is a critical factor in successful lubricant analysis. Errors in obtaining a representative sample can impair all further analytical efforts. There are two primary goals in obtaining a representative oil sample. The first is to sample in a way that the information density in the oil is maximized, which is often referred to as maximizing data density. The data could be particles from dust, moisture, additive levels, and wear debris.

The second goal is to minimize data disturbance. Samples should be carefully extracted so that the concentration of information is uniform, consistent, and unaltered by the sampling process. It is important to make sure that the sample does not become contaminated during the sampling. This can distort the data, making it difficult to distinguish what was originally in the oil from what has come into the oil during the sampling process.

To ensure good data density and minimum data disturbance in oil sampling, the following factors should be considered:

Sampling Location: Not all locations in the machine will produce the same concentration of data. Some machines require multiple sampling locations to answer specific questions related to the machine's condition, usually on an exception basis.

Sampling Method: The procedure by which a sample is drawn is critical to the success of lubricant analysis. Sampling procedures should be documented and followed uniformly.

Sampling Hardware: The hardware used to extract the sample should not disturb sample quality. It should be easy to use, clean, rugged, and cost effective.

Sample Container: The type of bottle and cleanliness (ISO 3722) of bottle help ensure that a representative sample is achieved.

It is always advised to expend the necessary resources to equip machinery with proper sampling hardware (valves, access ports, etc.) to ensure that the above goals in oil sampling are achieved.

Testing Strategy

In its simplest and most basic form, lubricant analysis is performed to improve the quality of machine and lubrication maintenance decisions. When analysis is well designed and implemented, many necessary questions about the machine and lubricant can be answered without excessive expense or complexity. This is best accomplished by directing the testing program around three important categories of lubricant analysis:

Fluid Properties Analysis: This category of lubricant analysis deals with the assessment of the chemical, physical, and additive properties of the oil. The testing here confirms that the lubricant is in compliance with the originally specified performance properties.

Contamination Analysis: Contaminants of various types can enter the system and the lubricant from the environment during servicing or by internal generation. Contamination compromises machine reliability and promotes lubricant failure. Lubricant analysis targeting contamination can help ensure that goal-driven targets for contamination control are maintained.

Wear Debris Analysis: When machine surfaces wear, they generate metallic particles that enter the lubricant. Monitoring and analyzing the generated debris enables analysts to detect and evaluate abnormal conditions to assist in directing needed maintenance activities.

Tests selected for routine condition monitoring vary between lubricant types and machine application tests. However, numerous tests are almost always performed (viscometry and elemental spectroscopy for instance) while others are optional or exception tests. An example oil analysis test scheme for a particular lubricant is illustrated in Fig. 5. The following are the types of tests commonly performed in laboratories:

- Viscosity
- Particle counting
- Infrared spectroscopy
- Acid number or base number
- Oxidation stability
- Ferrous density
- Microscopic particle identification
- Elemental spectroscopy
- Water concentration tests
- Flash point

Test results are compared to historic trends and/or alarm levels set to alert equipment owners of nonconforming conditions. Practitioners of modern oil analysis programs receive data from laboratories electronically and only review data on an exception basis.[19,20]

PMS AND ROUTINE INSPECTIONS

Mechanics and/or operators regularly inspect machines. In most cases, the equipment is inspected every day and in some instances, every shift. Such

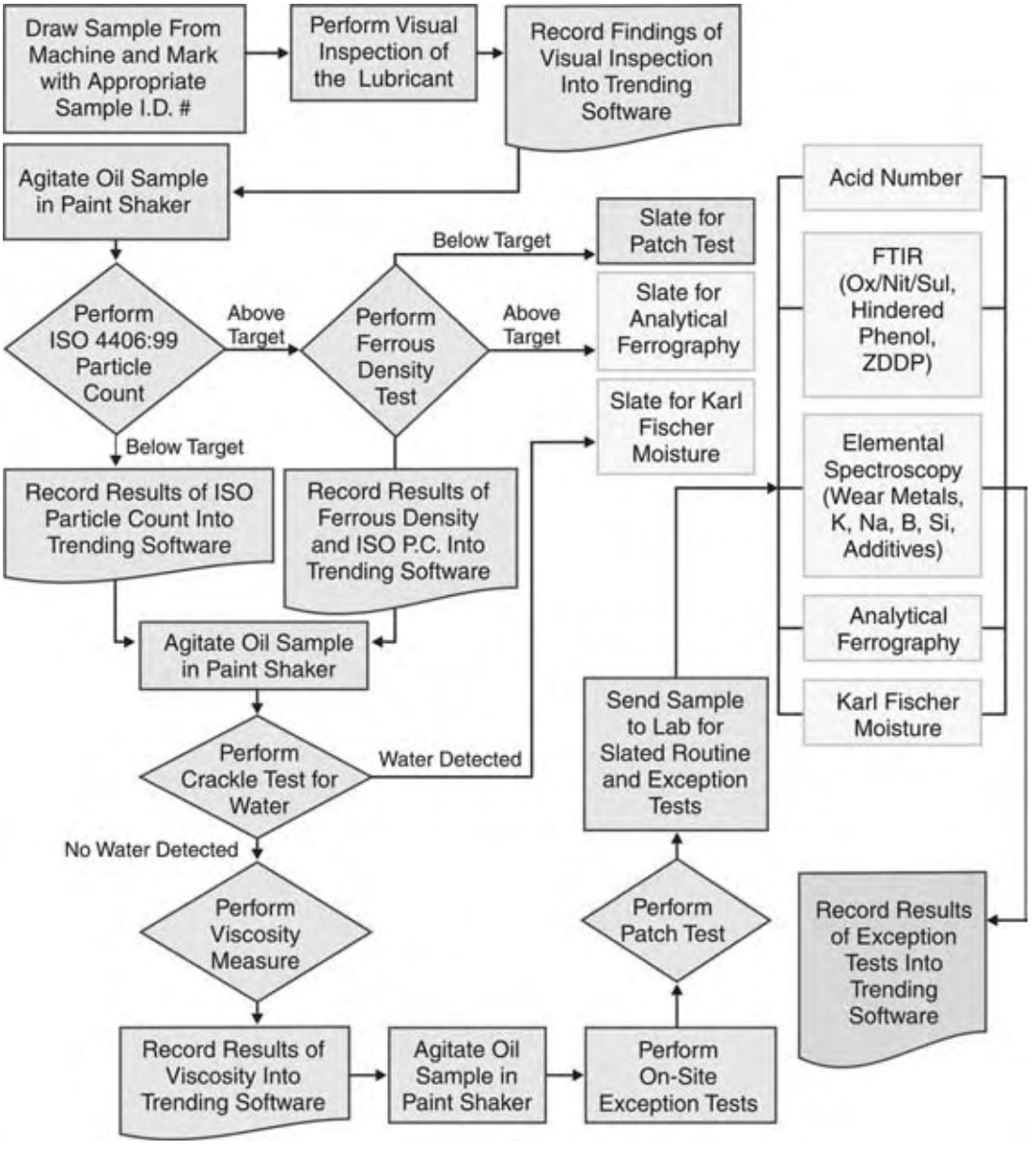


Fig. 5 Example routine and exception test slate. *(View this art in color at www.dekker.com.)*

inspections provide an excellent opportunity to verify that proper machinery lubrication is in order. Many variables can be quickly and easily monitored in a short amount of time. These inspection items can be revealing and will serve as the foundation of a precision lubrication management program.

Every plant, and for that matter every machine, will require a slightly different inspection routine and frequency. Below are six items to consider when setting up an inspection program.

Frequency: Not all machines in the plant are created equal. Some are highly critical and should be inspected on a regular basis, in some cases every shift or every day. Others require only weekly, biweekly, or monthly inspection. Because the lubrication inspection process is simple, it is always advised to err on the conservative side and inspect the machines frequently. The program can always be adjusted at a later time. When deciding on the frequency of inspections, consider the machine's propensity to fail, its impact on production, and the cost, time, and difficulty of restoring it upon failure. Too often, the sump volume is the primary determining factor. Sump volume is an easy way to group machines, but the cost of the lubricant is usually the least of a company's concerns.

Observed Parameters: For many machines, the number of observable parameters can reach into hundreds or even thousands. Obviously some priorities

must be established. It is best to select inspection parameters based upon their relative importance. How will an aberration affect the machine's performance? How frequently do detectable problems occur with a given inspection?

Training: Those performing the inspections must have some understanding about machinery lubrication, contamination control, etc. It is not necessary for them to be experts in the area as long as they have experts to turn to, but they do need a basic understanding. Inspectors also require task-based training on performing the inspection. For example, they must know that the desiccant is deemed failed when it turns from blue to pink. Modern machinery lubrication practices will be new to many of the people performing the inspections. Don't hold the team accountable unless they have been properly qualified to do the job.

Collection Mechanism: The trusty index card will work for performing inspections, but this requires the individual to remember the inspection criteria or refer to an inspection list and make notes about observations. A clipboard with a predefined set of questions (that can be answered "yes" or "no") for each machine is a definite improvement on the index card method. Be sure to leave a space for comments. Both the index card and the clipboard approach require manual entry of the data once they are collected. A more modern approach is to set up the inspection in an industrial-quality personal digital assistant. For each machine, the inspector can answer "yes" or "no," and as required, write short comments. These comments could be selected from a drop-down menu to minimize inconsistency. The data can then be uploaded to a host computer for storage. This speeds the process, ensures that the data will be captured, and reduces transposition errors.

Information Management: Be sure to track inspection findings carefully. Software is the best method because it can keep inspection data where other machine condition information is kept in order to facilitate diagnostics and failure root-cause analysis. Tracking the data helps identify frequently recurring problems, or lubrication "bad actors." It also helps optimize inspection frequency. It may be necessary to assign a number (typically 0 and 1) to the nominal ("yes" or "no") data collected to conform to the requirements of the software used. Some questions will be answered "yes" if things are OK, while others will be answered "no" if things are OK.

Remaining Steps: It seems that anytime a problem is encountered with any part of the lubrication system, an oil change is immediately scheduled with the assumption that the problem will be solved. This is referred to as lazy lubrication maintenance. Indeed, in certain cases the oil must be changed as a part of a complete and a cohesive corrective action. However, in many cases, an oil change is not required. Investigate the problem, identify its root cause, and implement a thoughtful corrective action that addresses the problems and eliminates any unnecessary steps.

Fig. 6 illustrates a list of recommended inspection items. Not all the items listed will be appropriate for all plants, and some additional items might need to be added to meet special needs. Also, some of the items listed are not routine, such as the evaluation of used filters and failed parts; hence, the work will need to be appropriately divided among operators, lubricant technicians, condition-monitoring technicians, and plant engineering.

Ineffective lubrication remains one of the leading causes for machinery failure. Routine inspection of the machine's lubrication system is among the easiest and most productive ways to avoid equipment failure and should play a pivotal role in the pursuit of precision lubrication. This program allows the operators and the mechanics who perform the routine inspections to have a better awareness of the importance of effective lubrication, resulting in the improvement of the quality of lubricant application.[21]

DEVELOPMENT OF PROPER LUBRICATION PROCEDURES

Despite overwhelming evidence suggesting that poor machinery lubrication spells trouble for the plant, most organizations lack clearly defined written procedures for performing basic lubrication tasks. The following outlines reasons why lubrication procedures are important elements of a proper lubrication procedure.[22]

The Importance of Lubrication Procedures

Work Scope: Procedures clearly scope the work an individual is expected to perform. They ensure work is executed the way management or engineering requires. If management wants 12 shots of grease pumped into the bearing, allowing 15 sec to elapse between shots, this desire can be clearly documented in the procedures.

Consistency: In the absence of procedures, five technicians are apt to perform the same task five different ways. In the absence of a procedure, each individual has the freedom to "personalize" the task at hand. Inconsistency produces undesirable results. Documented procedures bring uniformity into the lubrication task while keeping everyone on the same page.

Best Practices: A procedure creates the framework for standardizing best practice. It serves as the

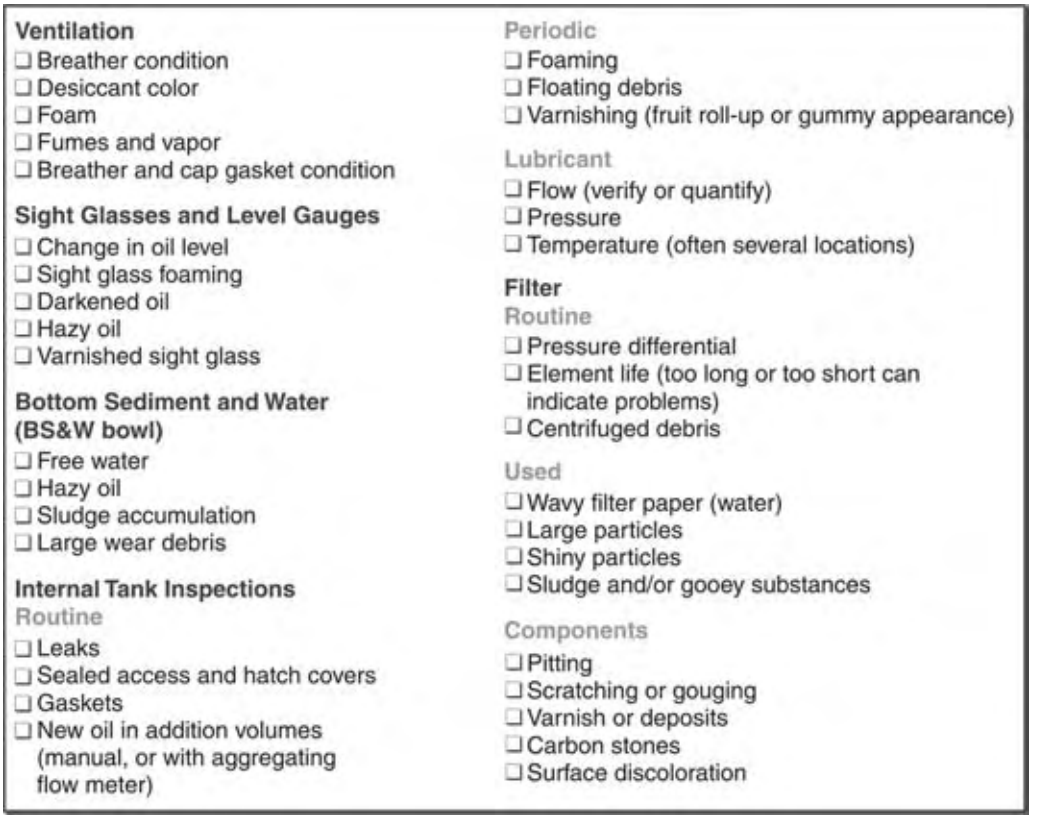


Fig. 6 Recommended inspection list. *(View this art in color at www.dekker.com.)*

container that gathers the experience and expertise of employees, outside consultants, vendors, and others into a single document. This convergence process also enables the team to align the procedure to the goals of the organization. Just enough ''best practice'' for one machine may be too much for another, depending upon the relative importance of the two machines to plant operations, even if they are identical in design.

Training: Lubrication procedures form the basis for training lube technicians. Basic training about lubrication, lubricants, oil analysis, etc. is designed to provide the foundation that enables the individual to think like a lube technician. Certification is another important part of the training process because it confirms that the individual possesses the skills to perform the job functions. This is called technology training. While it is important, technology training fails to convey specific task-based instructions for completing a lubrication-related work order. A set of procedures serves as a natural curriculum for task-based training. It also serves as the basis for evaluating an individual's ability to carry out the assigned tasks. Combining basic technology training and third-party certification with task-level training and skill verification creates a powerful combination and a valuable employee.

Elements of an Effective Lubrication Procedure

Emphasize Best Practice: As previously mentioned, procedures enable the incorporation of best practice. However, this is not automatic. A concerted effort must be made to build best practice into the procedure. Access the experience and knowledge of a maintenance team and bring in outside support as required to ensure that procedures are up-to-date and aligned with the business goals.

Communicate Clearly: Use clear, easy-to-understand language when creating procedures. Also, use digital photographs to reduce the procedure's dependence upon words. For intricate tasks, a digital video is an excellent way to communicate tasks that are difficult to describe orally. The best are procedures that include a top-view of the plant along with easy-to-spot landmarks that reveal the location of the machine. Getting to the right machine is the first step.

Electronic: Prepare lubrication procedures in an electronic form, preferably on a company-wide intranet, or onto an internet account for those who are moving toward web-based application support. When the procedures are electronic, they can be updated globally, attached to work orders, and linked to like

machines in a computerized maintenance management system system. Digital photographs and video images can be easily attached to a document. Documenting procedures electronically is more efficient and effective than the paper and three-ring binder method.

Continuous Improvement: Unfortunately, there is a downside to procedures. Without management, they can anchor the organization to the past, inhibiting the inclusion of new technology and best practices. Make sure the program includes a periodic review and improvement process to update and upgrade lubrication procedures. Keeping procedures in an electronic form simplifies continuous improvement because updates do not require tedious activities to physically replace pages in the lubrication manual. Changes can be documented and communicated in one memorandum, while updating the procedures themselves requires only the touch of a button.

There are a number of important areas that should be covered by lubrication procedures. These include the following:

- Cleaning and reconditioning of lubricant containers.
- Flushing machines and systems after overhaul and repair.
- Draining and refilling oil into a machine.
- Using a grease gun and other similar lubrication practices.
- Adding grease to single-point lubricators or centralized systems.
- Changing oil filters, separators, and breathers.
- Using a filter cart.
- Handling and disposing of waste oil/grease.

ACCESSORIZING EQUIPMENT FOR PROPER LUBRICATION

Many OEMs include only the minimum equipment for lubrication as a part of the original bill-of-material. Anything else is optional and therefore frequently left out. While these options add to the cost, many generate real value when selected. Following is a list of lubrication-related hardware and accessories that, where suitable for a particular machine and application, could and should be factory-installed. If not, they will need to be retrofitted onsite.[23,24]

Inspection

- Level gauges—located near fill ports and large enough to easily determine oil level.
- Expanded-metal guards on chains, couplings, belts, etc. Sheet-metal guards restrict easy inspection of lube points and moving seal parts.
- BS&W bowl—bottom sediment and water (BS&W) sight glasses enable quick inspection of low-lying contaminants and sludge.
- Mag plugs—magnetic drain plugs and other magnetic inspection devices enable wear metals to be examined and removed from lubricating oil.
- Large reservoirs and sumps should be equipped with inspection hatches and should have lips, gaskets, and compression clamps/bolts to control accidental dirt entry, ingression (dirt, wash-down sprays, etc.), and air movement.

Oil Sampling and Analysis

- Primary sampling ports/valves should be properly located and installed at the factory.
- Secondary sampling port/valves should be installed on various circulating systems.

Contamination Control

- All circulating systems should have quality beta-rated oil filters at a capture size and efficiency consistent with contamination control objectives.
- High-performance breathers should have capture size and efficiency consistent with contamination control objectives.
- Hydraulic cylinders should be equipped, where practical, with rod boots to control ingression.
- Reservoirs should have suitable baffling and be sized to enable contaminants to settle (dirt, water, sludge) and both air and gaseous contaminants to detrain.
- Dust protection covers should be installed where grease fittings are used.
- Off-line filters (kidney loop) should be installed on many bath/splash-lubricated machines.
- Headspace purge or other suitable headspace management equipment should be used with large reservoirs.

Instrumentation (Where Practical and Needed)

- Fluid pressure gauges.
- Flow meters.
- Temperature gauges.
- Free water alarm.
- Low oil-level alarms.
- Pressure differential gauges and filter bypass alarm.
- Air-intake vacuum gauges for diesel engines and breathers.

- Online oil properties' sensors: viscosity, ferrous density, particle count, moisture, etc.
- Headspace dew point meters.

Lubrication

- Optimum selection of lubricant delivery devices based on equipment performance needs drip, circulating, oil mist, splash oiled, constant-level oilers, single-point lubricator, grease fittings, centralized lube system, spray systems, etc.
- Optimum selection of seal type and quality for long-life service to control the ingress of contaminants and leakage.
- Use of temperature management systems, including heaters (start-up) and coolers as needed.
- Availability of grease purge ports where grease fittings are used.
- Proper selection of rolling-element bearing seals and shields.
- Proper use of lubricant return-line diffuser to control tank aeration.
- Installation of prelube system for engine cold-starts.
- Installation of power-flush quick- connects on tanks, reservoirs, and sumps.

LUBRICATION PROGRAM METRICS

Any pursuit intended to produce excellence needs a performance metric—machinery lubrication is no exception. If used properly, a performance metric works like a compass for the organization. It helps the users find their bearings and get back on track when performance is substandard. Once the organization performs on target, metrics help to keep it on track and facilitate continuous improvement. Likewise, it can serve as a common denominator for comparing plants within the organization, or even benchmarking entire industries. Performance metric also serves to push the team in the right direction. When the team is aware that something is being measured, trended, and managed, their behavior is affected.

Many organizations presently measure overall equipment effectiveness (OEE), which is the product of Availability $\times$ Quality rate $\times$ Yield.

Simply stated, OEE measures the plant's performance compared to the theoretical maximum production rate for the assets. It is a powerful metric that gives management an overall look at production effectiveness and eliminates the practice of hiding reliability shortcomings into scheduled downtime.

The metric overall lubrication effectiveness (OLE) (Fig. 7), employs a similar structure as the OEE, but focuses on variables that constitute effective machinery lubrication. The general equation for OLE is the product of the following three lubrication program-related factors:

- Percent compliance to scheduled lube PMs.
- Percent compliance to contamination control targets.
- Percent compliance to lubricant quality targets.

OLE measures an organization's performance in machinery lubrication, a critical input to machine reliability, and ultimately, profitability. Percent compliance to each of the target goals was selected because it is easy to calculate and understand. If OLE increases toward the target or holds steady above the target, it is heading in the right direction. If the OLE begins to trend downward, or holds steady below the target level, an intervention needs to occur in order to make corrective actions. The input variables are now discussed in more detail.

Calculating OLE

$$\mathrm{OLE} = \mathrm{PClpm} \times \mathrm{PCcct} \times \mathrm{PClqt}$$

where OLE, overall lubrication effectiveness; PClpm, percent conformance–lube PMs; PCcct, percent conformance–contamination control targets; and PClqt, percent conformance–lube quality targets.

Percent Compliance to Scheduled Lube PMs

Many lubrication tasks are scheduled activities related to performing inspections, sampling for oil analysis, regreasing, top-ups, and scheduled oil changes (where condition-based oil changes are not feasible). Several common problems exist with respect to completion of scheduled PMs. First, supervisors often pull designated technicians away from lube PMs to help with repair or to perform a number of other tasks within the plant. In some plants, lube technicians are called away from lubrication tasks 50% of the time or more. These PMs either pile up or get canceled, so they are not completed until the next time they come up on the schedule. Another problem, often referred to as ''pencil whipping,'' is the act of falsifying PM completion forms. This occurs when technicians have an overly heavy workload or they fail to see the relevance of a particular PM or group of PMs.

The keys to success in performing lubrication PMs include:

- Optimize PM tasks so that each one is relevant.
- Develop easy-to-follow written procedures for completing the tasks.

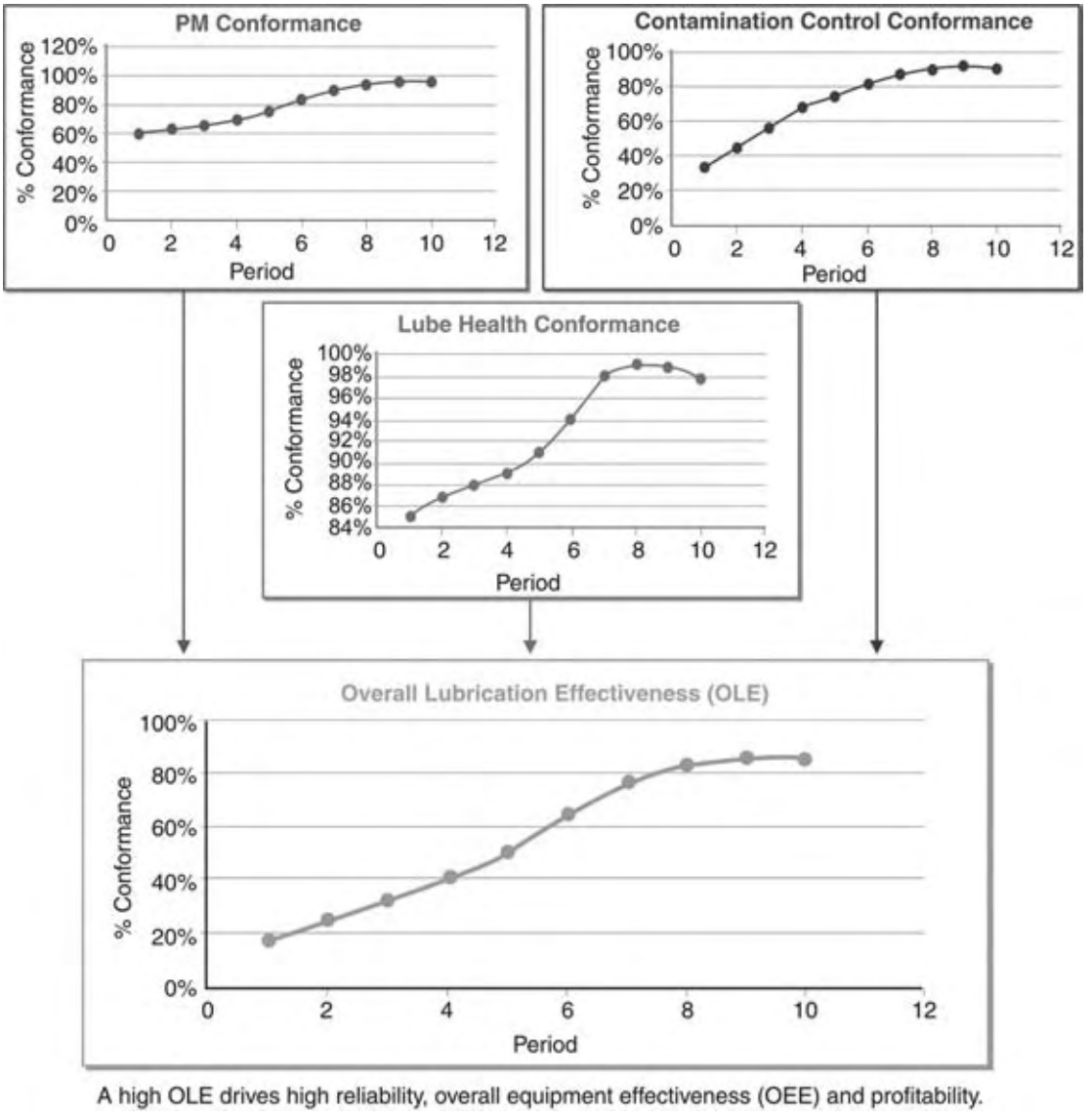


Fig. 7 Example overall lubrication effectiveness (OLE) performance metric. *(View this art in color at www.dekker.com.)*

- Train individuals to perform the tasks so they understand why those tasks are important.
- Periodically audit completion and quality.

Percent Compliance to Contamination Control Targets

The factor of percent compliance to contamination control targets aggregates the organization's performance in various aspects of contamination control. This includes new oil management and dispensing practices, filter selection and maintenance, breather selection and maintenance, seal selection and maintenance, proper combustion management, and other aspects of effective contamination control. If the machine is not in conformance, one or more of the above named (or other) sources of contamination requires attention.

Used oil analysis serves as the measurement apparatus for the contamination factor. Start by establishing meaningful target levels for each of the machines, taking into account mechanical sensitivity, application and/or environment severity and the machine's criticality to production, safety, environmental protection and/or other aspects of the mission. Compare the actual performance to targeted performance and report the percentage of the machines that comply with contamination targets (Fig. 8).

Percent Compliance to Lubricant Quality Targets

To operate reliably, the machine must contain the correct lubricant, the lubricant must be fit for service, and

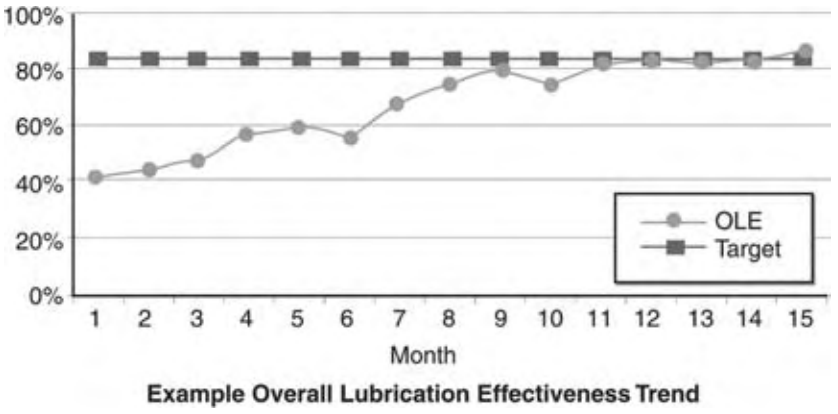


Fig. 8 Example overall lubrication effectiveness (OLE) trend. *(View this art in color at www.dekker.com.)*

it must be kept in the machine. Applying the incorrect lubricant or allowing the lubricant to become disabled due to chemical degradation and/or additive depletion can result in increased rates of mechanical wear and/or corrosion. Leakage of the lubricant can compromise the machine and create risk of injury and/or environmental mismanagement penalties.

The keys to success in ensuring lubricant quality are to select lubricants properly, store, and deliver them in a way that will minimize mixing and reduce chemical degradation, minimize aeration, run the machines as cool as the application will allow, and monitor regularly with oil analysis. Report the percentage of lubricants that conform to these lubricant quality targets.

OLE, like other global metrics, is designed to provide an overview for the management concerning how well the organization lubricates its equipment. The real work for lubrication and reliability engineers and technicians is in the factors themselves. As is the case with other global metrics, one can work to either improve the program, or work solely to improve the number. By including just a small percentage of the plant, excluding equipment that is hard to handle lubrication-wise, initiating flimsy contamination targets and lube quality alarms, arbitrarily cutting out PMs, or falsifying records can make the OLE look promising.

Trending an overall performance metric like the proposed OLE helps to drive improvement, keeps everyone focused, and serves as a rallying point to get and keep the team members working in the right direction and aligned to the goals of the organization.[25]

CONCLUSIONS

Modern lubrication programs in the chemical process industry are going through necessary transformations to modernize the selection and the use of lubricants to reduce operating costs and optimize machine reliability. The transformation has several unique elements, including:

Lubricant Selection: Lubricants are selected to best fit the changing demands from machinery and reliability expectations from users.

Application Methods: Applications methods of both oil and grease are changing to reduce labor content and improve the delivery to the needs of the machinery.

Contamination Control: Even the best lubricants and application methods can be severely compromised by invasion of harmful contaminants. Keeping contaminants in check using the proactive maintenance strategy is essential to effective lubrication.

Lubricant Sampling and Analysis: The routine analysis of lubricants is a vital condition-monitoring program element, providing information on the health of the lubricant, contaminant ingression, and machine condition.

PMs and Inspection Practices: Lubrication programs need to continually revisit scheduled PMs and inspection practices to rationalize their use for precision maintenance.

Lubrication Procedures and Documentation: Development of lubrication procedures consistent with industry best practices must include documentation and training.

Accessorizing Machinery for Precision Lubrication: Even the best lubrication procedures and PMs are of little value if the equipment is not properly equipped with hardware that enable their use.

Education and Skill Development: Today's lubrication technicians are knowledgeable workers and are well trained and possess skill competency certifications.

Program Metrics: OLE is an important program metric to trend continuous improvement and validate successes.

REFERENCES

1. Trujillo, G.; Troyer, D.D.; Fitch, J.C. *Best Practices for Machinery Lubrication*, Noria Corporation, 2004.
2. Fitch, J.C. What particles mean and why they need to be monitored and controlled. Practicing Oil Analysis Magazine. September 1998.
3. Troyer, D.D.; Fitch, J.C. An introduction to fluid contamination analysis. P/PM Technology. June 1995.
4. Moubray, J. *Reliability-Centered Maintenance*, 2nd Ed.; Industrial Press, Inc., 1997.
5. Contamination in lubrication systems for bearings in industrial gearboxes. SKF Ball Bearing Journal, **1993**, *242*.
6. Godfrey, D. Clean, dry, oil prolongs life of lubricated machines, Lubr. Eng. January 1989.
7. Troyer, D. The visual crackle—a new twist to an old technique. Practicing Oil Analysis Magazine. September 1998.
8. Barnes, M. Water–oil analysis 101. Practicing Oil Analysis Magazine. May 2002.
9. Fitch, J.C.; Jaggernauth, S. Moisture—the second most destructive lubricant contaminant, Part I—Its effects on bearing life. Proceedings of the JOAP Condition Monitoring International Conference, November, 1994.
10. Friesen, T.V. Transmission-hydraulic fluid foaming. *International Off-Highway and Powerplant Congress and Exposition*; Milwaukee, WI, 1987.
11. Duncanson, M. Sources of air-in-oil problems in circulating systems. Exxon Maintenance Conference, 1998; Exxon Company USA.
12. Fitch, J.C. Using oil analysis to monitor the depletion of defoamant additives. Practicing Oil Analysis Magazine. July 1998.
13. Fitch, J. Using oil analysis to control varnish and sludge. Practicing Oil Analysis Magazine. May 1999.
14. Wang, J.T. Dieseling in fluid power systems. Basic Fluid Power Res. J. **1982**, *16* (2), 305–311.
15. 'Hydraulic oil blackening' Fluid Power Symposium, April 8–10, 1981; BHRA Fluid Engineering, Cambridge, England.
16. Roberton, R.S.; Allen, J.M. Conserving hydraulic oil: new insight into oil degradation. Hydraulics and Pneumatics, WA761TP715.
17. Fitch, J.C. Demystifying sludge and varnish. Machinery Lubrication Magazine. January 2002.
18. Troyer, D.D. Removing water contamination, Machinery Lubrication Magazine. May–June 2001.
19. Fitch, J.C.; Troyer, D.; Barnes, M. *Oil Analysis for Maintenance Professionals*; Noria Corporation, 2004.
20. Troyer, D.D.; Fitch, J.C. *Oil Analysis Basics*; Noria Corporation, 2004.
21. Troyer, D. Boiling pot surveillance. Machinery Lubrication Magazine. January 2002.
22. Troyer, D. Everyone needs a playbook of lube procedures. Machinery Lubrication Magazine; November 2001.
23. Mackay, R. Current best reliability and maintenance practices of pumps and pump systems. Results Oriented Reliability Maintenance Conference, Raleigh, N.C., IDCON Inc.: November, 2002.
24. Fitch, J.C. Ask your OEM to accessorize new equipment for lubrication excellence. Machinery Lubrication Magazine. January 2003.
25. Troyer, D. Olé! rallying for a new lubrication performance metric. Machinery Lubrication Magazine. July 2002.

Mass Transfer Enhancement Because of Flow Instabilities

Vimal Kumar
K. D. P. Nigam
Department of Chemical Engineering, Indian Institute of Technology, New Delhi, India

INTRODUCTION

The transfer of mass within a fluid mixture or across a phase boundary is a process that plays a major role in various engineering and physiological applications. Typical operations where mass transfer is the dominant step are falling film evaporation and reaction, total and partial condensation, distillation and absorption in packed columns, liquid–liquid extraction, multiphase reactors, membrane separation, etc. The various mass transfer processes are classified according to equilibrium separation processes and rate-governed separation processes. Fig. 1 lists some of the prominent mass transfer operations showing the physical or chemical principle upon which the processes are based.

Several difficult but highly important approaches have come to the forefront in recent years to enhance the mass transfer operations, e.g., by increasing the surface area between the gas and the liquid, and/or by raising the intensity of mass transfer fluxes through increasing the intensity of turbulence. Nowadays, flow instabilities have been used to improve the mass transfer in chemical processes. In laminar flow regime, flow instability is often evidenced by the emergence of a structure that is periodic in space or time. More generally, it is associated with a loss of space–time symmetry by the initial flow considered. For instance, a small laminar stream of water from a faucet can give rise to regularly spaced drops because of an imbalance between gravity and surface tension.

Our main concern here is to present the mass transfer enhancement in several rate-controlled separation processes and how they are affected by the flow instabilities. These processes include membrane processes of reverse osmosis, ultra/microfiltration, gas permeation, and chromatography. In the following section, the different types of flow instabilities are classified and discussed. The axial dispersion in curved tubes is also discussed to understand the dispersion in the biological systems and radial mass transport in the chromatographic columns. Several experimental and theoretical studies have been reported on dispersion of solute in curved and coiled tubes under various laminar Newtonian and non-Newtonian flow conditions. The prior literature on dispersion in the laminar flow of Newtonian and non-Newtonian fluids through curved tubes has been compiled by Nigam and Saxena[1] and is analyzed in this entry. In addition, a new and more effective mass transfer device is also discussed based on the phenomenon of flow inversion. The effectiveness of the present device can be assessed by the fact that even at a Dean number of 3 a value of dispersion number as low as 0.0013 is obtained under the condition of significant molecular diffusion, and in the case of negligible molecular diffusion the value of dimensionless time at which the first element of tracer appears at the outlet is as high as 0.85. It is expected that this device can be used for chromatographic columns, cryogenic systems, and biochemical operations. The design also has potential as a membrane configuration for further improved operations.

TYPE OF FLOW INSTABILITIES

There are several direct and indirect approaches to enhance the mass transfer processes. In the membrane separation processes, concentration polarization and fouling are the two main limiting factors that affect the performance of any process. Therefore, for successful membrane process design, the flow configuration and module geometry must be considered along with membrane chemistry and structure. Fig. 2 shows a schematic analysis to overcome concentration polarization and fouling in membrane separation processes. A number of techniques have been used to achieve this goal by introducing various instabilities in the flow. Some of them are discussed here.

Roughness (placing protuberance or corrugation on the surface): These techniques are difficult to scale-up to intermediate or large size modules and are often limited by their high axial pressure drop. The approach, however, does demonstrate that shear stresses developed by flow instabilities enhance membrane permeation rates for difficult feeds.[2,3]

Pulsation: Superimposing an oscillating pressure gradient onto bulk axial flow in a tube produces a velocity profile with two equal maxima nearer the wall than the centerline. The net result of this is increase in the wall shear rate and the back migration of the solute in the bulk flow region, which results in improved performance.[4,5] Beside the improvement in performance,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007850

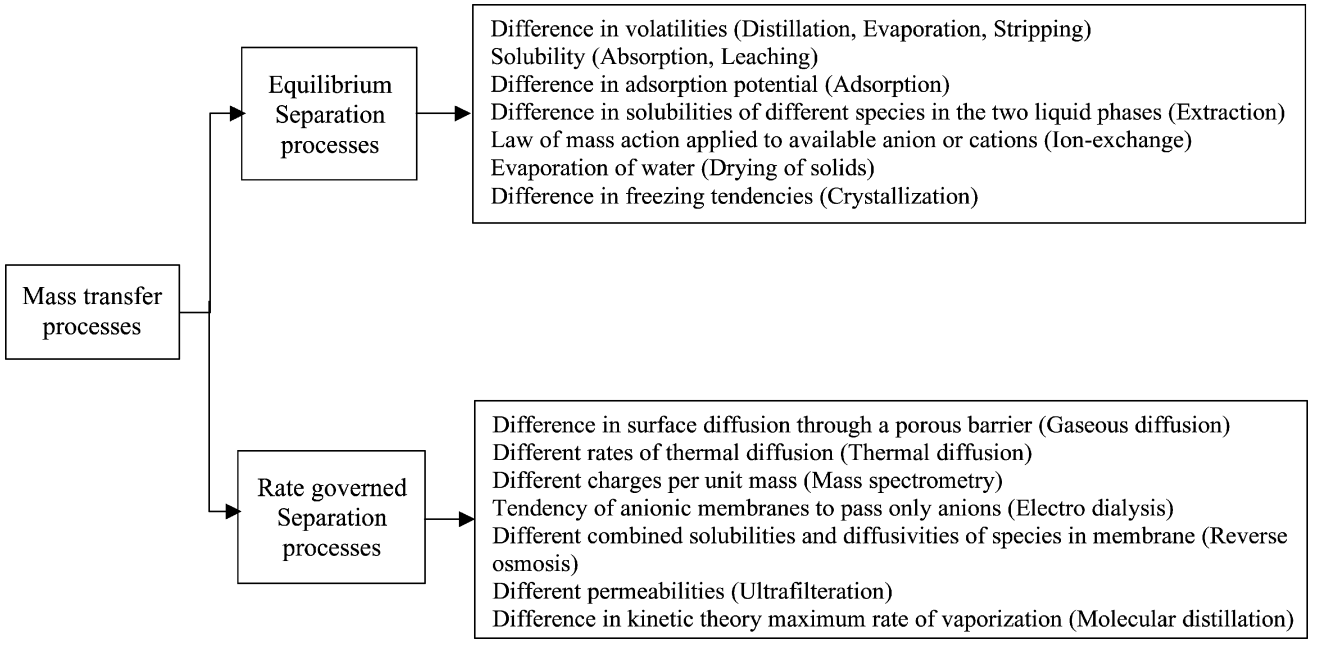


Fig. 1 Various mass transfer processes in the chemical industry.

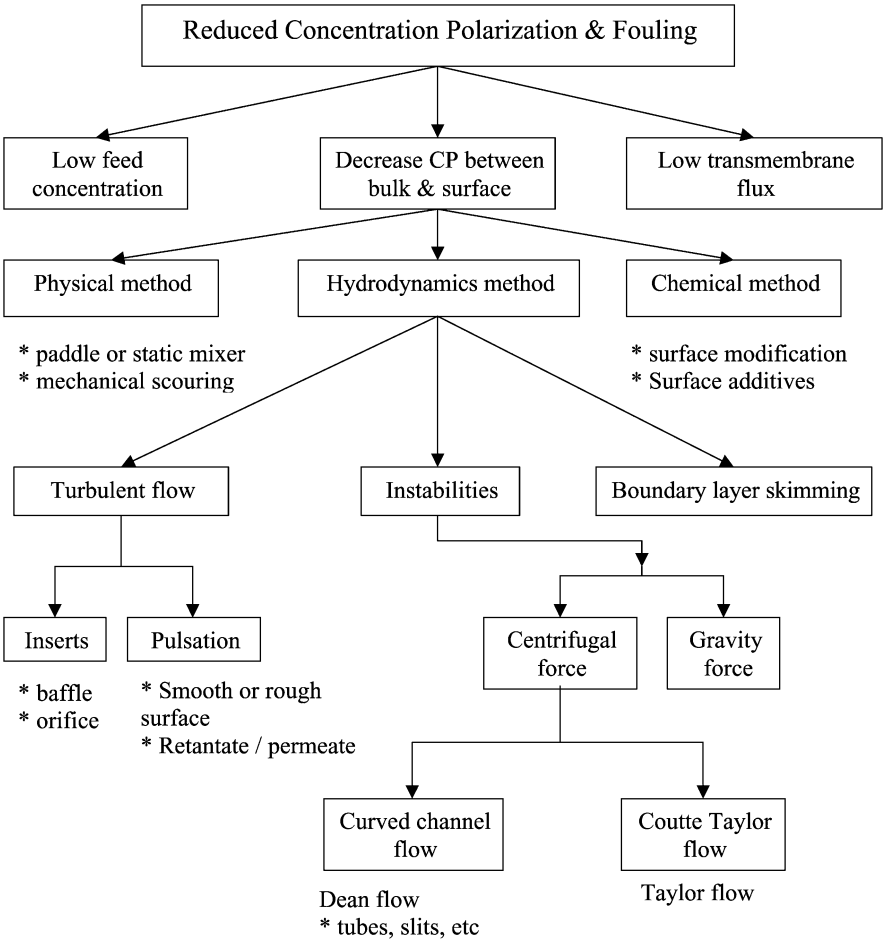


Fig. 2 Flow transfer instabilities in membrane separation processes. CP, concentration polarization.

pulsed axial flow decreases the penalty of high power consumption and loss in productivity. Detailed discussions on this aspect are given in Ref.[6].

Taylor Vortices: Taylor vorticesoccur in a rotating annular pipe and are very effective in reducing concentration polarization and particle deposition because of high shear rates. They are decoupled from the axial feed flow owing to the enhanced mass transfer between the boundary layer and the bulk phase.[7–12] The limitations are high-energy consumption, scale-up, and sealing problems.

Dean Vortices: When the fluid flows through a curved channel at Reynolds number above a critical value for incipient production of vortices, the flow changes from laminar to unstable laminar flow with counter rotating eddies or vortices.[13] These vortices are used to depolarize the solute buildup near the membrane–solution interface. These types of design are easy to scale-up and should not have exorbitant energy requirement. The only disadvantage of this type of equipment is the higher axial pressure drop as compared with conventional equipment.

Air Sparging: Gas sparging or injection of air bubbles has been effectively used to reduce concentration polarization and enhance mass transfer.[14–20] The secondary flows around bubbles promote mixing and reduce the thickness of the concentration polarization boundary layer. When the bubble diameter exceeds that of the membrane (tubular or hollow fiber), slugs are then formed; further increase in bubble diameter has no effect on flux improvement. Large slugs can displace most of the boundary layer and cause the pressure to pulsate. This results in enhancing the flux.

Backflushing Techniques: Periodic backflushing is widely used in membrane separation processes as a method to clean the membrane. By using a reverse transmembrane separation (TAMP) for very short periods of time, permeate is forced through the membrane in the reverse direction, and causes the filter cake or gel to expand, declog, and eventually be carried away. Periodic or intermittent backflushing is another form of instability that influences the concentration polarization boundary layer, particularly when high-frequency backflushing is used.[21,22] The efficiency of backflushing is dependent on optimizing the parameters (e.g., frequency, duration, pressure, single or multiple pulses, uniform or variable pressure, and duration) suitable for the specific application.

Further details of the mass transfer enhancement techniques in membrane separation processes are reported by Belfort[6] and Al-Bastaki and Abbas.[23] In this entry, the focus is on the flow instabilities produced by Dean vortices in curved and coiled tubes because of their advantages over the other techniques, viz., lower axial dispersion, better radial mixing, residence time distribution closer to plug flow, higher mass transfer, and critical Reynolds number for the transition from laminar to turbulent flow. Dean vortices arise from the centrifugal force and it may result in an enhanced wall shear stress. These vortices lower the concentration polarization, reduce the buildup of particle layers on the membrane, and improve the filtration performance.

BACKGROUND

The phenomenon of Dean vortices was first observed by Dean[13] in a curved tube. The fluid in the central part is driven toward the external wall by the centrifugal force, which gives rise to a secondary flow. This results in the inward movement of the fluid near the wall and the outward movement of the fluid near the center (Fig. 3A,B). Curved tubes may be classified as torus, bends, helical coils (with a constant curvature), and spirals (with a variable curvature). Fig. 4 shows the helical coiled, spiral, and bend tubes. The hydrodynamics in a coiled tube can be characterized by a dimensionless number named after Dean and is defined as:

$$N_{De} = N_{Re}\sqrt{\frac{d_i}{d_c}} \tag{1}$$

where N_{Re}—is the Reynolds number, d_i the internal diameter (m), and d_c—the coil diameter (m). In a helical tube, the so-called modified Dean number (N'_{De}) can be used to take into account the pitch effect, b:

$$N'_{De} = N_{Re}\sqrt{\frac{d_i}{d'_c}} \tag{2}$$

where d'_c denotes the effective coil diameter that varies as a function of the pitch of the coiled tube, i.e.,

$$d'_c = d_c\left[1 + \left(\frac{b}{\pi d_c}\right)^2\right] \tag{3}$$

The secondary flow in curved tube results in a pressure gradient between a maximum pressure at the outer wall and a minimum pressure at the inner wall. The frictional energy loss near the tube wall is increased, resulting in a higher-pressure drop as compared with that of a straight tube (Fig. 3C). The friction factor for laminar flow in the coiled tubes can be calculated from the correlation of Mishra and Gupta[24] giving the ratio of the friction factor in a coiled tube, f_c, to the friction factor in a straight tube, f:

$$\frac{f_c}{f} = [1 + 0.033(\log_{10} N_{De})^4] \tag{4}$$

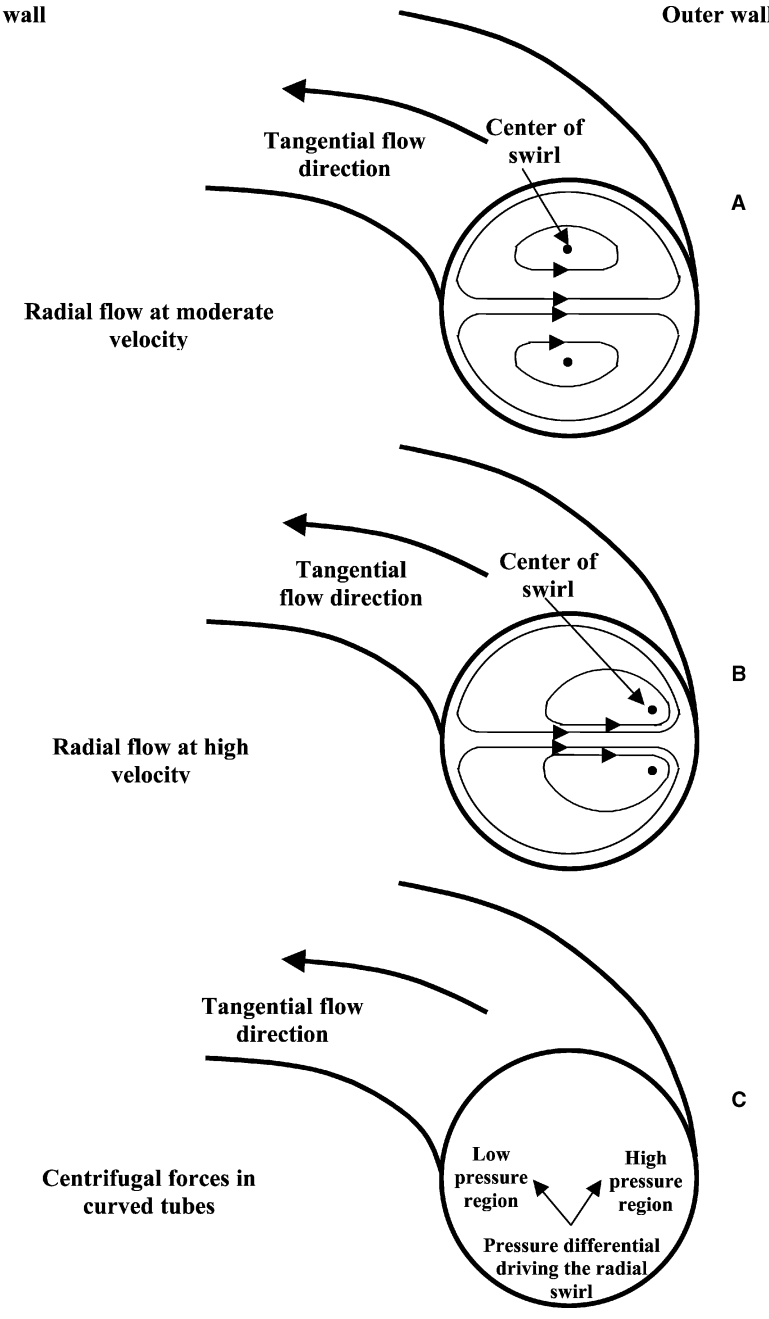


Fig. 3 Secondary flow in the cross-section of a curved tube: (A) radial flow at moderate velocity, (B) radial flow at high velocity, and (C) pressure gradient between a maximum pressure at the outer wall and a minimum pressure at the inner wall.

Experimental investigations showed that Eq. (4) is not only valid for helical tubes but also for sinusoidal curved tubes.

MASS TRANSFER IN MEMBRANES

In membrane separation processes, there is convective flow of the feed solution or suspension from the bulk toward the surface of the membrane, where separation and simultaneous slow diffusion of the rejected particles into the bulk occur because of a concentration difference. This slow back transport, coupled with membrane fouling, is a major limiting factor in the separation process. Over recent years, much attention has been devoted to the improvement of both the design and the operation of membrane separation modules. It has been reported that Dean vortices,[13] generated in curved geometries (which provide a high mixing potential by secondary flow) can be used to modify the hydrodynamics near the membrane surface.

The investigation of Dean vortices and their application to membrane separation processes has been the subject of several experimental and theoretical studies concerning the improvement of microfiltration (MF), ultrafiltration (UF), and nanofiltration (NF),

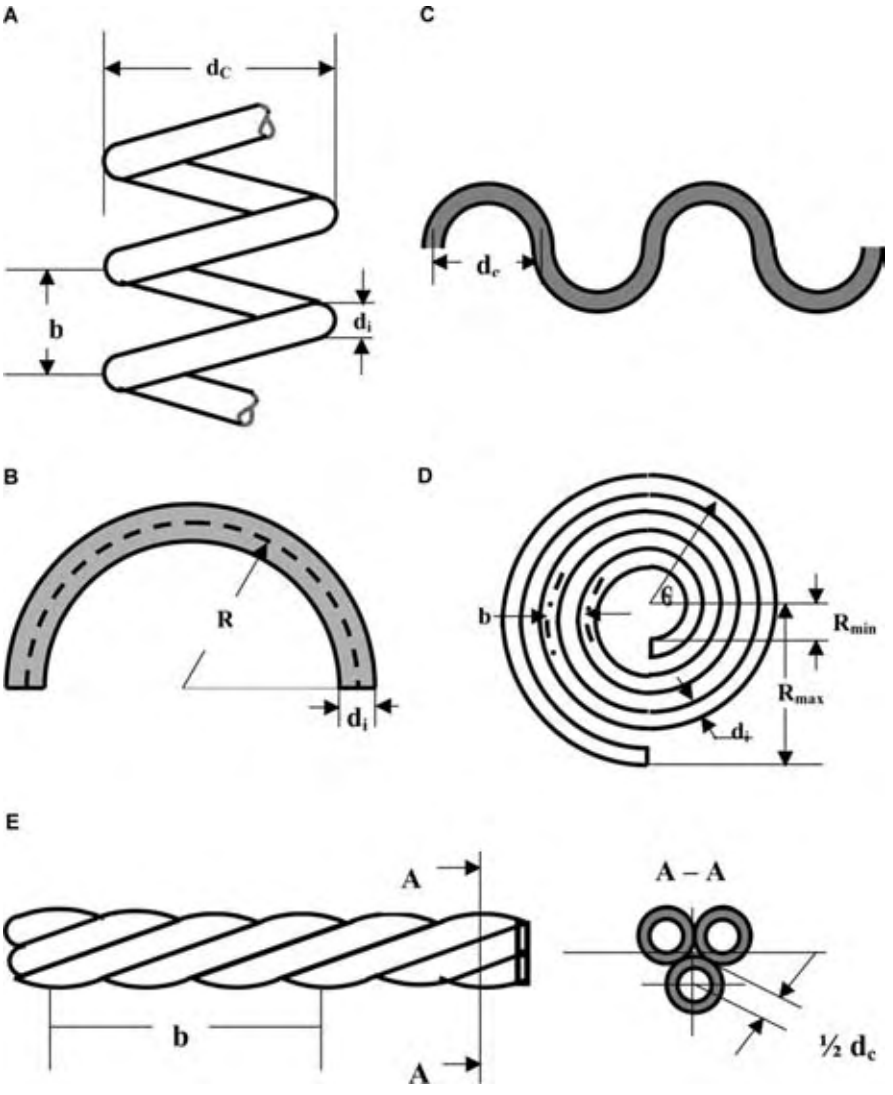


Fig. 4 Physical illustration of different types of curved tubes: (A) horizontal helical coil, (B) bend tube, (C) serpentine tube (meander-shaped tube), (D) spiral, where r is the tube radius, R the coiled tube curvature, b the coil pitch, and R_{min} and R_{max} are minimum and maximum radii of curvature of the beginning and end of the spiral, and (E) twisted tubes.

or pervaporation. Belfort and coworkers [25–27,31,38–40] experimentally and theoretically studied the Dean flow instabilities in curved channels to improve the performance of membrane processes for micro-, ultra- and nanofiltration in a helically coiled and curved slit membrane system because of Dean vortex flow. Moulin et al.[43] have shown an increased mass transfer coefficient for a helical hollow fiber module compared with a linear hollow fiber module and demonstrated the effectiveness of using Dean vortices for several applications, such as pervaporation and ultrafiltration of suspensions of yeast and bentonite.[51,52] Ophoff et al. proposed a design of helically twisted tubular membrane and reported significant flux improvement for the UF of a dextran solution.[41,42] Manno et al.[46] obtained up to 400% improvement of the limiting permeate flux for concentrated baker's yeast suspensions. Guigui et al.[47] experimentally demonstrated the efficiency of secondary flow for reducing membrane fouling in a coiled hollow fiber module. The shear stress is higher than in a straight module and is maximum near the external wall of the coiled tube. Gehlert et al.[48] used polysaccharide, protein, and yeast suspensions to compare UF and MF in the presence and absence of Dean vortices. In Fig. 5, a plot of enhancement in mass transfer vs. Reynolds number from different authors has been compiled.

Wille et al.[49] investigated the crossflow MF of polystyrene latex (0.005 wt.%) and baker's yeast suspension (0.3 wt.%) using tubular polypropylene membranes. Their investigation was focused on single straight and meander-shaped tubular membranes. Mallubhotla et al.[53,54] studied the helical nanofiltration (NF) modules and found improvement in the flux of aqueous solutions. Chung et al.[31] used a three-dimensional numerical model to quantify the concentration polarization for Dean vortex flow in a spiral reverse osmosis (RO) system. They showed that the presence of Dean vortices promotes mixing and inhibits the growth of the concentration polarization boundary layer. Kuakuvi et al.[58] employed a new

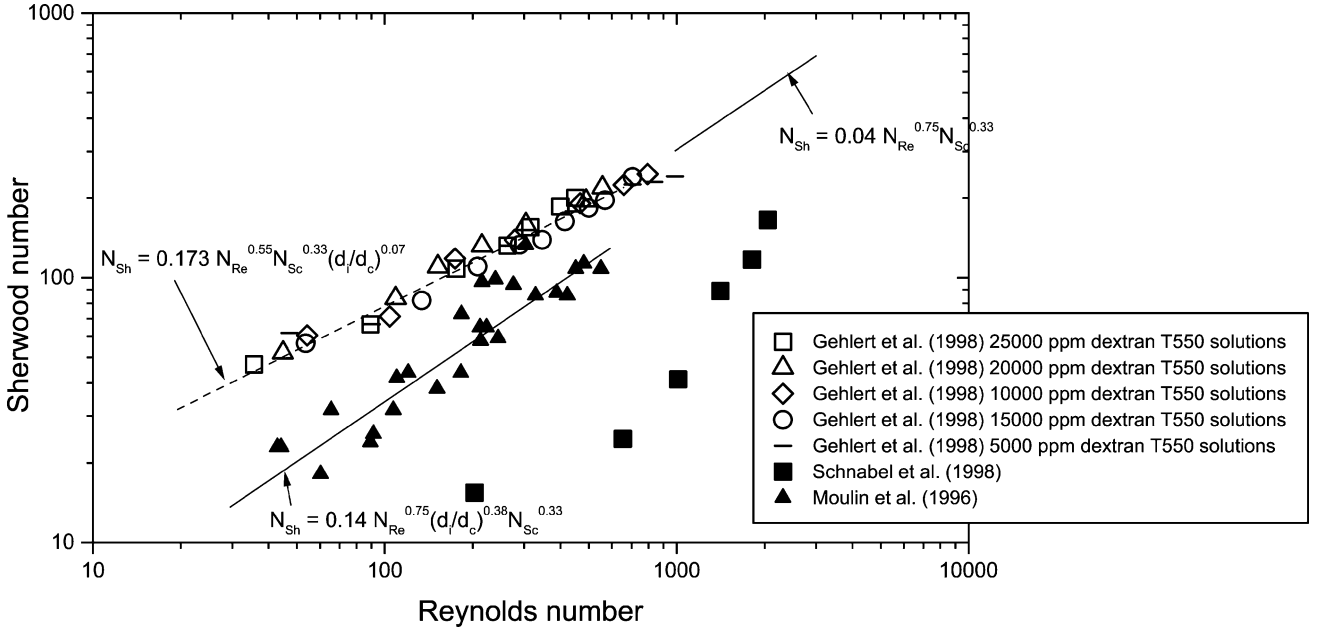


Fig. 5 Sherwood number vs. Reynolds number.

geometry, namely, woven hollow fiber UF membranes as a means of generating Dean vortices. They found that in the woven geometry, flux increases in a manner similar to the results observed for helical fibers but at lower energy consumption. Ghogomu et al.[59] experimentally compared the performance of several designs of curved membrane modules (straight, helically coiled, twisted and sinusoidal, or meander-shaped) with Dean vortices at steady state condition. They observed that the different curved geometry configurations gave the same limiting permeate flux at the same Dean number.

Al-akoum et al.[63]experimentally and theoretically reported the flux enhancement in three particular systems: shear-enhanced filtration with a vibrating membrane module, gas/liquid two-phase flows, and Dean vortices for yeast suspension system. They reported that the permeate flux was found to obey the empirical law: $J = K\tau_{wm}^{n}$ (where τ_{wm}^{n} is the mean wall shear stress, Pa) with $0.43 < n < 87$ and K depending on the membrane type and the yeast concentration for a particular system.

Several research groups [25–64] experimentally and theoretically reported the enhancement in mass transfer in membrane separation processes using Dean vortices, and their findings are summarized in Tables 1 and 2.

Helical Baffles and Stamps

A somewhat different methodology has also been used to enhance the mass transfer using Dean vortices to promote turbulence using helical baffles, and stamped and corrugated membranes. Gupta et al.[65] produced helical baffles by winding wires onto rod supports. These baffles were located axially inside a ceramic tubular membrane. They studied the separation of two different systems such as yeast suspensions and oil and water mixtures and observed up to 50% permeate flux enhancement as compared with linear modules at the same energy expenditure. Elmaleh and Ghaffor[66,67] used a UF membrane with helical baffles introduced in the filtration element and treated the suspensions of crude oil and biological cell. Significant flux improvements were reported. Broussous et al.[68] used a helical stamp (helical stamps are created on the internal surface of the ceramic macroporous support and shaped by a simple adaptation of the extrusion process) on the inside of a tubular ceramic MF membrane. They compared the performance of the proposed module with a smooth-surface membrane and reported that the permeate flux increased by a factor of 6. Scott and Lobato[69] studied the influence of crosscorrugated membranes and the orientation of flow to the angle of corrugation (90°, 45°, and 0°) on the mass transfer coefficients. They observed that the maximum improvement in mass transfer is reached when flow takes place at an angle of 90° to the corrugated membrane.

Energy Expenditure and Mass Transfer Enhancement

As the pressure drop is higher in coiled tube than in straight tube module, it is important to consider the performance of the coiled tube geometry in terms of

Table 1 Theoretical study for membrane separation processes in curved tubes

Author (Ref.)	Technique	Modules	Remarks
Brewster et al.[25]	Numerical solution (wide gap theory)	Spiral tube	Dean vortices in curved the tube were reported
Chung et al.[26]	Finite volume method and nuclear magnetic resonance flow imaging technique	Spiral tube	Velocity and pressure fields in the curved tube have been reported numerically, and it has been observed that Dean vortices can successfully depolarize solute buildup and fouling
Chung et al.[27]	Numerical solution	—	Flux improvement of 2–5 is obtained in the curved module over the straight module
Mallubhotla and Belfort[28]	MRI measurement and numerical technique	Helical coils	Velocity profiles and the dynamic behavior of Dean vortices in the curved tube flow were reported
Moulin et al.[29]	Finite volume method	Straight, torus, helical, and woven modules	Reported the influence of the geometrical parameter (d_s, d_i, b, N_{Re} and N_{De}) on the shear stress distribution and Dean vortices
Moll et al.[30]	Numerical solution and laser visualization	—	The numerical approach shows that the instantaneous velocity profile always exhibits a maximum value at the extrados of the coil, while in some cases, laser visualizations yield results that seem somewhat different and inconsistent
Chung et al.[31]	ADI scheme	Spiral tube	Dean vortex flow promotes rapid mixing and inhibits the growth of the solute concentration boundary layer

energy expenditure. Most of the studies reported so far are related to the constant energy dissipation in the straight and curved tube membranes. Moulin et al.[43] extended their work for the application in the gas–liquid membrane contactor for the case of water oxygenation. They reported that for a given energy dissipation, the permeate flux remains higher in the presence of secondary flows as shown in Fig. 6. Kuakuvi et al.[58] and Ghogomu et al.[59] presented an energy expenditure analysis in the curved membranes. According to their analyses, for a given energy consumption, the permeate flux is higher for a curved module than for a straight one. They have also reported a correlation for the energy consumption in curved tube modules:

$$\frac{W_S}{W_C} = \frac{\nu_S}{\nu_C} = 1 + 0.033[\log_{10}(gN_{Re})]^4 \quad (5)$$

Bubolz et al.[70] also reported a similar kind of analysis for laminar flow in coiled tubes.

AXIAL DISPERSION IN CURVED TUBES

Dispersion theory is concerned with the dispersal of a solute in a flowing liquid owing to the combined action of a nonuniform velocity profile and molecular diffusion. Numerous authors have discussed flow dispersion in a straight tube since the pioneering work of Taylor.[71] Additional complexities arise when flow instabilities are introduced using Dean vortices, where lateral mixing owing to molecular diffusion is augmented by the convective secondary motion. The dispersion in curved tubes is also of physiological interest because of the prevalence of the curved and bifurcational vessels in the circulatory and respiratory system.[72–74] Comprehensive reviews for the axial dispersion of Newtonian and non-Newtonian fluids in curved tube have been published.[75–99] Several experimental and theoretical studies have been reported on dispersion of the solute in curved tubes for Newtonian and non-Newtonian laminar flow conditions, and are listed in Tables 3 and 4. Fig. 7 shows the empirical correlation for dispersion in coiled tubes from various workers as a function of dimensionless number $N_{De}N_{Sc}^{1/2}$.

Axial Dispersion Owing to Tube Bending

It can be seen from Fig. 7 that in curved tubes, there is a drastic reduction in axial dispersion with an increase in Dean number. These studies reveal that very high Dean numbers are required to induce significant mixing in the cross-sectional plane. Saxena[100] proposed a simple and more effective alternative to helical coils. Helical coils are very efficient in inverting the flow

Table 2 Experimental study for membrane separation processes in curved tubes

Author (Ref.)	Membrane types and techniques	Solution	Remarks
Weissman and Mockros[32]	Helically coiled permeable walled tubes	Flow of gases into blood	Improvement in the mass transfer rates of gases into blood
Tanishita et al.[33,34]	Serpentine and hollow fiber membrane	Oxygen transfer into blood	Secondary flow enhances the mass transfer
Winzeler and Belfort[37]	Polypropylene MF membrane; polyethersulfone UF membrane	Baker's yeast; dairy whey	Dean vortices have five times higher performance than conventional devices
Chung et al.[37,38]	Acrylic transparent sheet	Water doped with $CuSO_4$	Dean vortices can depolarize the solute buildup near the membrane surface
Chung et al.[39]	Curved slit channel with 180° curve section	DI water, monodispersed polystyrene latex particle and yeast suspension	Dean vortices depolarize the solute buildup near the membrane–solution interface
Mallubhotla et al.[40]	Polypropylene microfiltration membrane	Cerevisiae broth suspension	Flux improvement of 60% was reported in curved module over the straight module
Ophoff et al.[41,42]	Twisted tubular membranes	Dextran T70 solution	Flux increase of 40% with respect to straight tube membranes
Moulin et al.[43]	Coiled silicone tubular membranes	Oxygenation operation with water	Dean vortices give better performance in terms of oxygen transfer, and the improvement was 2–4 times higher than in linear membranes
Mallubhotla and Belfort[44]	Polysulfone, polypropylene MF membrane	DI water with salt	Improvements in permeate flux of up to 150%, 80%, and 100% for the helically coiled modules over the linear modules for MF, UF, and NF at $N_{De} = 292, 450$
Crowder and Gooding[45]	Silicon rubber hollow fiber	O_2 removal from water and pervaporation of trichloroethylene	Mass transfer coefficient values observed for O_2 transfer were at least 20% higher than those achieved with traditional spiral wound modules
Manno et al.[46]	Cellulose acetate hollow fiber	Baker's yeast suspensions	Secondary flow enhances membrane permeation by a factor of up to 5
Guigui et al.[47]	Coiled hollow fiber module	Bentonite and yeast suspension	Concentration polarization and fouling are reduced and the limiting flux is improved by up to 5 times
Gehlert et al.[48]	Polyether sulfone	Polysaccharides, proteins, and beer yeast suspensions	A new empirical correlation of UF in dextran solution is proposed
Wille et al.[49]	Polypropylene membranes	Polystyrene latex and baker's yeast suspension	The efficiency for the curved modules is three times higher than for straight modules

Mallubhotla et al.[50]	Polyamide, poly(ether sulfone) helical membrane	Glutamic acid, glutamine, and lysine	Flux and rejection were higher for the helical module than for the linear module in the range of $375 \leq N_{Re} \leq 500$
Schnabel et al.[51]	Polydimethylsiloxane hollow fiber membranes	*n*-Butanol–water and chloroform–water	Improvement factor increases with Reynolds number and liquid mass transfer is twice in helical modules than that in straight modules
Moulin et al.[52]	Cellulose acetate hollow fibers	Colloidal bentonite solution and dextran solution	Secondary flow reduces the fouling phenomenon and the permeate flux is twice in the coiled membranes than in the linear membranes
Luque et al.[55]	Helical and linear modules	Cell suspensions (yeast, *Escherichia coli*, and mammalian cell cultures)	Flux and capacity improvements of up to 3.2-fold (constant TMP operation) and 3.9-fold (constant flux operation) with the helical module over the linear module
Kluge et al.[56]	Helically wound hollow fibers	Polyethylene glycol solutions and silica suspensions	Flux improvements of up to 45% in helical modules over straight modules
Kluge et al.[57]	Helical and linear modules	Suspended yeast cells (MF) and the solution of BSA and beer yeast (DF)	The capacity of helical module was 19 times that of the straight modules, and the flux improvement was 18–43%
Kuakuvi et al.[58]	Cellulose acetate hollow fiber membrane	Dextran solution	For a given energy consumption, the limiting flux observed in the woven hollow fiber module was higher than that in the helical and straight modules.
Ghogumu et al.[59]	Cellulose acetate hollow	Bentonite (montmorillonite) particle fiber membrane	A new empirical correlation was proposed and for a given permeate flow rate, the energy consumption is lower for a curved module than for a straight one
Schutyser et al.[60]	Polyethersulfone helical and linear modules	rBDNF inclusion bodies from an *E. coli* cell suspension	Dean vortex improves the filtration performance. A highly concentrated solution is obtained, and it is highly energy saving
Schutyser and Belfort[61]	Polyethersulfone helical and linear modules	Tetradecyltrimethyl-ammonium bromide and sodium salicylate with silica suspension	Proposed a new empirical correlation and reported that the permeate flux decreased with increasing TTASal concentration and increasing Reynolds number
Kaur and Agrawal[62]	Cellulose acetate membrane	Lysozome	Dean vortices enhance the mass transfer coefficient and an empirical correlation is proposed
Liu et al.[64]	Polysulfone hollow fiber module	30% TBP (in kerosene) + phenol + water	Coiled fibers provide better mass transfer performance than the straight fibers, and the improvement factor is in the range of 2–4

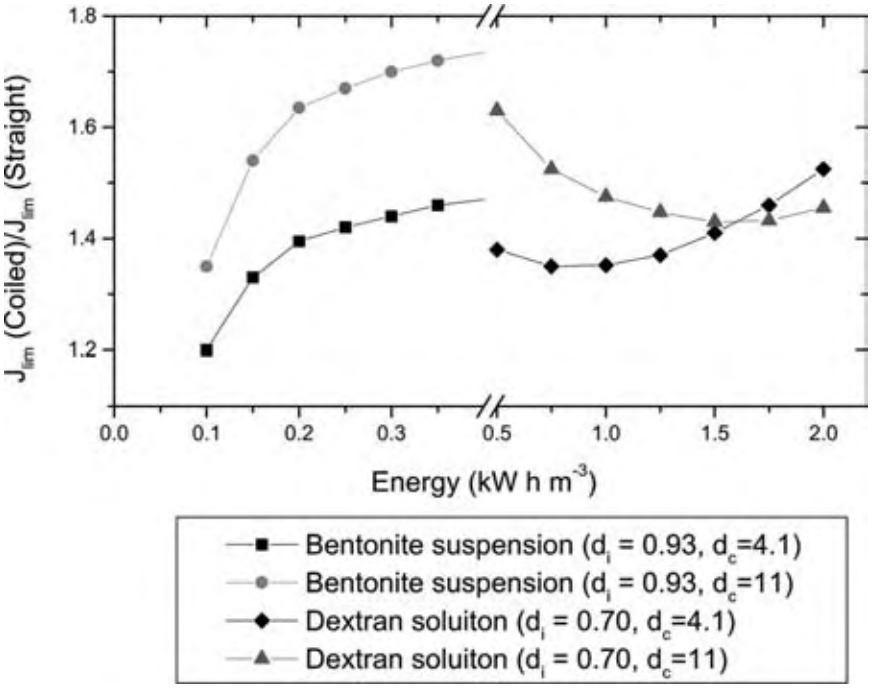


Fig. 6 Ratio of limiting permeate flux between a coiled and a straight module vs. energy consumption.[42] *(View this art in color at www.dekker.com.)*

and improving cross-sectional mixing in a coiled tube. Saxena[100] proposed that it is possible to obtain complete flow inversion by shifting the direction of centrifugal force. He conducted the residence time distribution experiments under the condition of significant molecular diffusion, using water as the flowing media. The Dean number was varied from 3 to 60. Taylor's dispersion model[101] was fitted to the experimentally obtained step input curve (F curve). Fig. 8 shows their best-fit computed values of dispersion number as a function of Dean number. It is interesting to mention that about a 20-fold reduction in dispersion number (D/uL, where D is diffusion coefficient, u the velocity in the tube, and L the length of tube) was obtained as compared with that of a helical coil in the new device.

AXIAL DISPERSION IN CHROMATOGRAPHIC COLUMNS

Ideally, a chromatographic column should yield the largest possible number of theoretical plates in the shortest possible time. While the use of microparticulate packings facilitates considerable increase in the efficiency, it has the disadvantages of poor reproducibility (owing to variations in mobile phase flow path length) and prohibitively high operating pressures. On the other hand, the use of open tubular columns is hindered by the relatively slow diffusion in liquids. Normally, radial mass transport is governed by molecular diffusion alone; in gas chromatography, where the diffusion coefficient is five times higher than in liquid chromatography, the radial diffusion is sufficiently fast to allow the use of capillary columns. In liquid chromatographic systems, however, undesirable band broadening occurs in columns with larger inner diameter. Such stringent inner diameter limitations result in detection and dead volume constraints.

Gidding[102] proposed a coupling concept analogy, in which the radial dispersion is represented by the combination of molecular diffusion and convection, thus providing a radial convection mechanism to facilitate the use of open tubular columns with larger internal diameters. Various techniques have been applied to enhance radial mass transport. Taylor[71] and Knox[103] worked in the turbulent regime while Desty and Douglas[104] utilized static mixers to enhance the radial mixing.

Effect of Dean Vortices

Hofmann and Halász[105] investigated the radial mixing without considering the axial dispersion in band broadening (band broadening is because of the multiple path of an analyte through the column packing, molecular diffusion, and effect of mass transfer between phases) in squeezed, twisted, and coiled tubes. Tijssen[106,107] predicted the heavy equivalent theoretical plate (HETP) in a chromatographic column using Dean's profile, based on dye tracer experiments in a transparent tube. He derived a correction factor to the Taylor dispersion model[71] using an analogy based

Table 3 Theoretical studies for the axial dispersion in curved tubes

Author (Ref.)	Technique	Parameters range	Remarks
Erdogan and Chatwin[75]	Perturbation analysis	$N_{De} \leq 17$	Followed Dean's solution; dispersion is reduced by curvature only if $N_{Sc} > 0.124$
Nunge et al.[76]	Perturbation analysis	$N_{De} \leq 17$	Followed Topakoglu's[77] solution; at low N_{Re} coiling enhances axial dispersion for liquids and at high values of N_{Re} coiling results in a reduced axial dispersion
Janssen[78]	Numerical integration	$N_{De}^2 N_{Sc} < 5000$	Introduced a new parameter $N_{De}^2 N_{Sc}$. For $N_{De}^2 N_{Sc} < 100$ no significant reduction, when compared with straight tube, was reported, while for $100 < N_{De}^2 N_{Sc} < 5000$ the ratio decreased more than threefold
Johnson and Kamm[79]	Monte Carlo and spectral model	$1 < N_{De}^2 N_{Sc} < 10^5$	Results were qualitatively similar to those of Ref.[77], the effective diffusion in coiled tubes fell to about 0.2 of its value in straight tube
Daskapoulos and Lenhoff[80]	Method of moments		Their results were very similar to those of Refs.[77,78] except for the fact that the results were applicable to higher values of Dean number
Eckman and Grotberg[81]	Perturbation technique	$N_{Sc} \approx 1$	An abrupt increase in transport rate was observed for the ratio of stroke length to tube diameter
Pedley and Kamm[82]	Asymptotic analysis and numerical solutions	$N_{Sc} \approx 1$	Effect of oscillatory flow on axial dispersion was reported; when the secondary flow time equals the oscillation period, a resonance is formed that gives rise to a prominent maximum in the transport rate
Sharp et al.[83]	Monte Carlo and spectral model	$0.01 < N_{De} < 1000$; $0 < \alpha < 100$; $N_{Sc} = 0.68, 10{,}000$	Effect of oscillatory flow on axial dispersion was reported; secondary flow decreases axial transport by a factor of 5 for low values of $\alpha^2 N_{Sc}$
Agarwal et al.[84]	Perturbation method	$0.2 < n < 1.5$; $1 < N_{De}^2 N_{Sc} < 700$; $\lambda = 30, 100$	Axial dispersion decreases with increase in the value of $N_{Re} N_{Sc}$ for a given coil and power law fluid, while it increases with increase in curvature ratio for a given $N_{Re} N_{Sc}$ and power law fluid
Agarwal and Jayaraman[85]	Spectral method with finite difference method	$100 < N_{De}^2 N_{Sc} < 10^4$	For the range of $N_{De}^2 N_{Sc}$ from 100 to 10^4 the axial dispersion in a circular curved tube is markedly less than that in a straight tube

Table 4 Experimental studies for the axial dispersion in curved tubes

Author (Ref.)	System (fluid)	Parameters range	Correlation proposed	Remarks
Koutsky and Adler[86]	Water–electrolyte	$300 < N_{Re} < 40{,}000$	$\log_{10}\left[N_{Di}\left(\frac{L}{D}\right)^{0.702}\left(\frac{R}{R_C}\right)^{-2.26}\right] = mZ + n$ where m, Z, and n are functions of ellipticity, curvature, and Reynolds number, respectively	For $N_{Re} < 300$, axial dispersion in helical coils was found to be of the same order as that in straight tube, and for equal power consumption helical coils facilitate less axial dispersion than do straight tubes
Van Andel et al.[87]	Air–butane, water–electrolyte	$100 < N_{Re} < 2000$; $\lambda = 30, 50$	—	For liquids, the Peclet number has no definite trends with the operating parameters, but for gases the axial dispersion approaches ideal plug flow
Trivedi and Vasudeva[88]	29% DEG solution in water	$10 < N_{Re} < 1600$; $1.5 \times 10^3 < N_{Sc} < 8.7 \times 10^3$; $10 < \lambda < 280$	—	1.5- to 500-fold reduction in values of K_c depending upon the values of N_{Re} and λ
Nigam and Vasudeva[89]	Water	$0.5 < N_{Re} < 10$; $10 < \lambda < 785$	—	For a nearly 80-fold variation in curvature ratio, the effect of coiling on K_c was negligible if $N_{De} < 0.15$
Mouljin et al.[90]	He–N_2, Kr–N_2, N_2–He, Pr–He, Fr–He	$1 < N_{Re} < 800$; $0.23 < N_{Sc} < 3.02$; $9 < \lambda < 47$	—	At low velocities, coiled tubes and straight tubes show identical dispersion behavior, and the Taylor–Aris model is applied up to $N_{De}N_{Sc}^{1/2} \approx 10$
Shetty and Vasudeva[91]	DEG solution	$7000 < N_{Sc}N_{De}^{1/2} < 55{,}600$; $10.3 < \lambda < 785$	$\frac{D_C - D_M}{D_S - D_M} = 1$ for $N_{De}N_{Sc}^{1/2} \le 7$; $\frac{D_C - D_M}{D_S - D_M} = 1.52 - 0.275 \ln(N_{De}N_{Sc})$ for $N_{De}N_{Sc}^{1/2} > 7$	Proposed an empirical correlation

Van Den Berg and Deelder[92]	Nitrobenzene–*iso*-octane, benzene–chloroform, benzene–hexane	$10 < N_{De}N_{Sc}{}^{1/2} < 200$	$\kappa = 1$ for $N_{De}N_{Sc}^{1/2} \leq 10$; $\kappa = 5.6\ (N_{De}N_{Sc}^{1/2})^{0.68}$ for $10 < N_{De}N_{Sc}^{1/2} < 200$	For increasing values of $N_{De}N_{Sc}{}^{1/2}$, the axial dispersion in the helical coils decreases as compared with that in straight tube
Saxena and Nigam[93]	Sodium carboxy methyl cellulose (CMC)	$N_{Re} < 140$; $0.6 < n < 1.0$; $10 < \lambda < 220$	—	Axial dispersion decreases with increase in the value of $N_{Re}N_{Sc}$, while it increases with increase in curvature ratio for flow of non-Newtonian fluids
Iyer and Vasudeva[96]	—	$N_{De}N_{Sc}^{1/2} \sim 7000$	$\frac{D_C - D_M}{D_S - D_M} = 1$ for $N_{De}N_{Sc}^{1/2} \leq 7$; $\frac{D_C - D_M}{D_S - D_M} = 1.65 - 0.863 \log(N_{De}N_{Sc}^{1/2})$ for $N_{De}N_{Sc}^{1/2} > 7$	Correlation is valid for higher values of $N_{De}N_{Sc}^{1/2}$
Deelder et al.[97]	Nitrobenzene–*iso*-octane, benzene–chloroform, benzene–hexane	$N_{De}N_{Sc}^{1/2} < 200$	$\kappa = 1$ for $N_{De}N_{Sc}^{1/2} \leq 12.5$; $\kappa = 1.52 - 0.275 \ln (N_{De}N_{Sc}^{1/2})$ for $12.5 < N_{De}N_{Sc}^{1/2} < 200$	Proposed an empirical correlation
Singh and Singh[98]	H_2–A, air–A, O_2–A, air–H_2	$10 < N_{Re} < 100$; $0.176 < N_{Sc} < 1.359$; $26.6 < \lambda < 98$	—	For $N_{De}N_{Sc}^{1/2} < 10$, the effect of coiling on axial dispersion is not significant
Kauffman and Kissinger[99]	Buffer solution of sodium phosphate monobasic	$N_{De}N_{Sc}^{1/2} > 4$	$\kappa = 1$ for $N_{De}N_{Sc}^{1/2} < 4$; $\kappa = 1.27 - 0.517l, \log(N_{De}N_{Sc}^{1/2}) + 0.0759 \log(N_{De}N_{Sc}^{1/2})^2$ for $N_{De}N_{Sc}^{1/2} > 4$	Serpentine geometry enhances the radial mixing better than the helically coiled reactors even at low flow rates

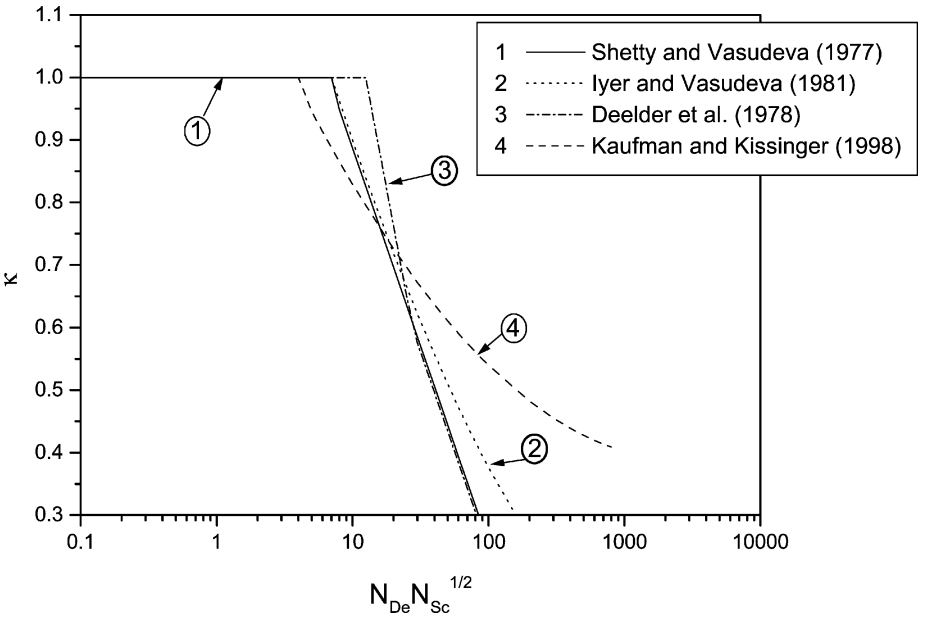
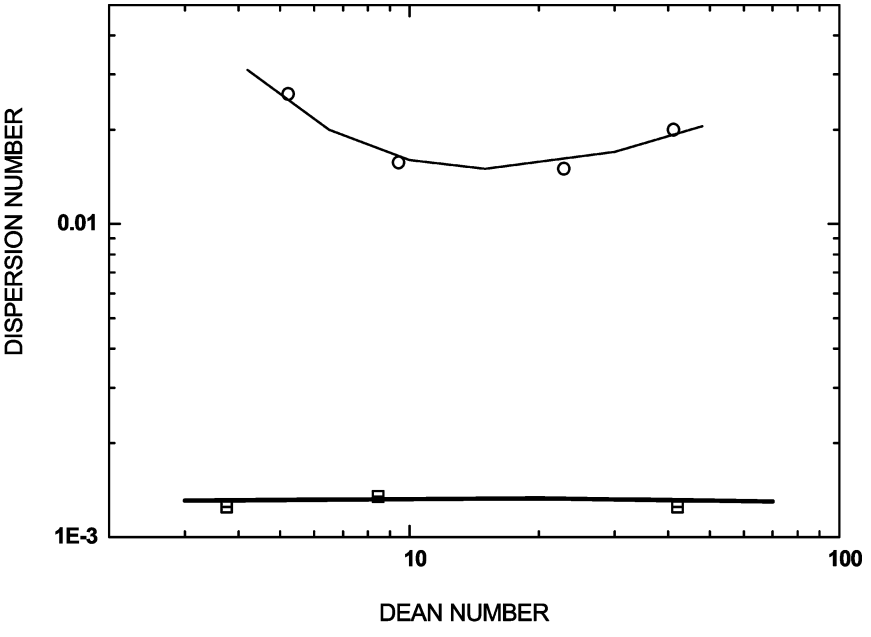


Fig. 7 Comparison of various axial dispersion correlations ($\kappa = D_c/D_s$, ratio of dispersion in coiled tube to the straight tube).

on heat transfer. Katz and Scott used serpentine tubes (Fig. 4C) as low dispersion connectors.[108] Tubes stitched through steel mesh framework and knitted onto themselves were used for postcolumn reactors.[110] According to Kauffman and Kissinger,[109] there is no distinct advantage in terms of band spreading between any of the serpentine geometries and the smallest diameter helical coil. However, using these geometries reduced the band spreading by up to 44% when compared with a linear tube of the same length.

Grochowicz et al.[110] modeled the secondary flow in serpentine and coiled tubes as the measure of the radial mixing using computational fluid dynamics techniques. They showed that flow characteristics in serpentine tubes result in considerably less band broadening per unit length than in linear tubes of the same inner



Coil		Symbol
Straight helix	-	○
Coil with Complete Flow Inversion	-	⊟

Fig. 8 Dispersion number vs. Dean number.

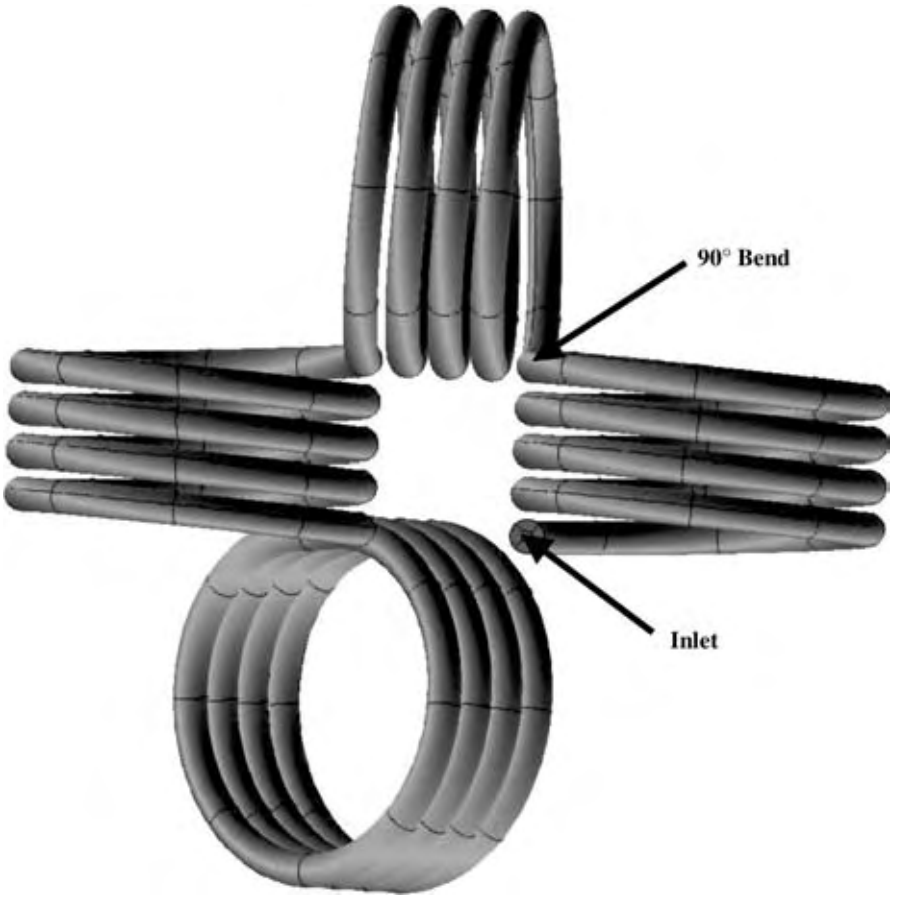


Fig. 9 Coiled flow inverter (the proposed device).

diameter. They[111] further reported the feasibility of rapid chromatographic separations in open tubular serpentine columns of 100–250 mm ID. They compared the serpentine and helical tubes from the point of view of band broadening. The use of comparatively large column diameters is permissible for the enhancement of native diffusion mass transfer with a secondary mechanism of radial mixing. The new device that has been proposed here (Fig. 9) can also be used as a chromatographic column because of its compactness, narrower RTD, and less axial dispersion as compared with the conventional, helical chromatographic columns.

CONCLUSIONS

Several novel techniques for mass transfer enhancement reported in the literature have been discussed in detail. The present study is focused on the mass transfer enhancement in the rate-controlled separation processes using flow instabilities. There are a large number of examples about the success of flow instabilities produced by Dean vortices in improving the performance by increasing flux and reducing fouling in membrane separation processes. Several curved modules (helically, coiled, twisted, spirally wound, and sinusoidal modules) with Dean vortices have been described, and their performances are analyzed in comparison with conventional linear modules. It has been reported that, for a given energy consumption, the permeate flux is higher for a curved module than for a straight one. Besides this, the influence of secondary flow on the extent of dispersion in curved and coiled tubes helps to enhance mass transfer. It has been reported that Dean vortices markedly reduced axial dispersion and, for equal power consumption, helical coils facilitate less axial dispersion than straight tubes. In the chromatographic systems undesirable band broadening occurs in columns, which results in detection and dead volume constraints. This can be avoided by using helical and serpentine columns. It has been observed that flow characteristics in serpentine tubes result in considerably less band broadening than in linear tubes of the same inner diameter. A further development in modeling of membrane filtration processes will provide a better understanding of the flow and mass transfer in curved membrane.

A new type of flow reactor is also presented, which has narrower RTD and higher heat and mass transfer as compared with the conventional reactors. The proposed reactor has a 20-fold reduction in axial

dispersion as compared with the helical coils. The ease of fabrication, compactness, and the lower axial dispersion found establish their superiority over other mechanical devices known in literature for inducing mixing in a cross-sectional plane. Because of the above advantages, the proposed device, which is based on the principle of flow inversion, will have a higher mass transfer coefficient and can be used in membrane separation (to improve permeate flux at low-energy consumption) and as a chromatographic column.

REFERENCES

1. Nigam, K.D.P.; Saxena, A.K. Residence time distribution in straight and cured tubes. In *Encyclopedia of Fluid Mechanics*; Cheremishinoff, N.P., Ed.; Gulf Publishing: U.S.A., 1986; Vol. 1, 675–762.
2. Van Der Wall, M.J.; Racz, G. Mass transfer in corrugated-plate membrane modules. I. Hyperfiltration experiments. J. Membr. Sci. **1989**, *40*, 243.
3. Thomas, D.G. Forced convection mass transfer in hyperfiltration at high fluxes. Ind. Eng. Chem. Fundam. **1973**, *12*, 396.
4. Kennedy, T.J.; Merson, R.L.; McCoy, B.J. Improving permeable flux by pulsed reverse osmosis. Chem. Eng. Sci. **1974**, *29*, 1927.
5. Bauser, H.; Chmiel, H.; Stroh, N.; Walitza, E. Interfacial effects with microfiltration membranes. J. Membr. Sci. **1982**, *11*, 321.
6. Belfort, G. Fluid mechanics in membrane filtration: recent developments. J. Membr. Sci. **1989**, *40*, 123–147.
7. Kroner, K.H.; Nissinen, V. Dynamic filtration of microbial suspensions using an axially rotating filter. J. Membr. Sci. **1985**, *36*, 85.
8. Kroner, K.H.; Nissinen, V.; Ziegler, H. Improved filtration and microbial suspension. Bio/Biotechnology **1987**, *5*, 921.
9. Holeschovsky, U.B.; Cooney, C.L. Quantitative description of ultrafiltration in a rotating filtration device. AIChE J. **1991**, *37*, 1919.
10. Murase, T.; Iritani, E.; Chidphong, P.; Kano, K.; Atsumi, K.; Shirato, M. High speed microfiltration using a rotating, cylindrical ceramic membrane. Int. Chem. Eng. **1991**, *31*, 370.
11. Belfort, G.; Mikulasek, P.; Pimbley, J.M.; Chung, K.-Y. Diagnosis of membrane fouling using a annular rotating filter. 2. Dilute particle suspensions of known particle size. J. Membr. Sci. **1993**, *77*, 23–39.
12. Pudjiono, P.I.; Tavare, N.S.; Garside, J.; Nigam, K.D.P. Residence time distribution from a continuous-couette flow device. Chem. Eng. J. **1992**, *48*, 101.
13. Dean, W.R. Notes on the motion of fluid in a curved pipe. Phil. Mag. **1927**, *4*, 208–233.
14. Li, Q.Y.; Cui, Z.F.; Pepper, D.S. Effect of bubble size and frequency on the permeate flux of gas sparged ultrafiltration with tubular membranes. Chem. Eng. J. **1997**, *67*, 71–75.
15. Ghosh, R.; Cui, Z.F. Mass transfer in gas-sparged ultrafiltration: upward slug flow in tubular membranes. J. Membr. Sci. **1999**, *162*, 91–102.
16. Cui, Z.F.; Wright, K.I.T. Flux enhancements with gas sparging in downward crossflow ultrafiltration: performance and mechanism. J. Membr. Sci. **1996**, *117*, 109–116.
17. Bellara, S.R.; Cui, Z.F.; Pepper, D.S. Gas sparging to enhance permeate flux in ultrafiltration using hollow fibre membranes. J. Membr. Sci. **1996**, *121*, 175–184.
18. Laborie, S.; Cabassud, C.; Durand-Bourlier, L.; Laine, J.M. Fouling control by air sparging inside hollow fibre membranes—effects on energy consumption. Desalination **1998**, *118*, 189–196.
19. Serra, C.; Durand-Bourlier, L.; Clifton, M.J.; Moulin, P.; Rouch, J.C.; Aptel, P. Use of air sparging to improve backwash efficiency in hollow-fiber modules. J. Membr. Sci. **1999**, *161*, 95–113.
20. Li, Q.Y.; Ghosh, R.; Bellara, S.R.; Cui, Z.F.; Pepper, S.D. Enhancement of ultrafiltration by gas sparging with flat sheet membrane modules. Sep. Purif. Tech. **1998**, *14*, 79–83.
21. van Hoof, S.C.J.M.; Hashim, A.; Kordes, A.J. Boron removal from seawater reverse osmosis permeate utilizing selective ion exchange resin. Desalination **1999**, *124*, 231.
22. Chellam, S.; Jacanangelo, J.S. Existence of critical recovery and impacts of operational mode on potable water microfiltration. J. Environ. Eng. **1998**, *124*, 1211.
23. Al-Bastaki, N.; Abbas, A. Use of fluid instabilities to enhance membrane performance: a review. Desalination **2001**, *136*, 255–262.
24. Mishra, P.; Gupta, S.N. Momentum transfer in curved pipes: Newtonian fluids. Ind. Eng. Chem. Proc. Des. Dev. **1979**, *18*, 130–137.
25. Brewster, M.E.; Chung, K.Y.; Belfort, G. Dean vortices with wall flux in a curved channel membrane system. 1. A new approach to membrane module design. J. Membr. Sci. **1993**, *81*, 127.

26. Chung, K.-Y.; Edelstein, W.A.; Li, X.; Belfort, G. Dean vortices in a curved channel membrane system. 5. Three dimensional magnetic resonance imaging and numerical analysis of the velocity field in a curved impermeable tube. AIChE J. **1993**, *39*, 1592–1602.
27. Chung, K.-Y.; Edelstein, W.A.; Belfort, G. Dean vortices with wall flux in a curved channel membrane system. 6. Two dimensional magnetic resonance imaging of the velocity field in a curved impermeable slit. J. Membr. Sci. **1993**, *81*, 151–162.
28. Mallubhotla, H.; Belfort, G. Dean vortices in curved tube flow: magnetic resonance imaging and microfiltration. Proceedings of the Eighth Annual Meeting of the North American Membrane Society (NAMS), Ottawa, Canada, 18–22 May 1996.
29. Moulin, Ph.; Veyretb, D.; Charbit, F. Dean vortices: comparison of numerical simulation of shear stress and improvement of mass transfer in membrane processes at low permeation fluxes. J. Membr. Sci. **2001**, *183*, 149–162.
30. Moll, R.; Moulin, Ph.; Veyret, D.; Charbit, F. Numerical simulations of Dean vortices: fluid trajectories. J. Membr. Sci. **2002**, *197*, 157–172.
31. Chung, K.-Y.; Brewser, M.E.; Belfort, G. Dean vortices with wall flux in a curved channel membrane system. 3. Concentration polarization in a spiral reverse osmosis slit. Chem. Eng. J. Japan **1998**, *31*, 683–693.
32. Weissman, M.H.; Mockros, L.F. Gas transfer to blood flowing in coiled tubes. J. Eng. Mech. Div., Proc. ASCE **1968**, *94*, 857–872.
33. Tanishita, K.; Richardson, P.D.; Galletti, P.M. Tightly wound coils of microporous tubing: progress with secondary flow blood oxygenator design. Trans. Am. Artif. Intern. Organs **1975**, *21*, 216.
34. Tanishita, K.; Ujihira, M.; Watabe, A.; Nakano, K.; Richardson, P.D.; Galletti, P.M. Gas transport in serpentine microporous tubes under steady and pulsatile blood flow conditions. ASME J. Biomech. Eng. **1991**, *113*, 223.
35. Dorson, W.; Baker, E.; Hull, H. A shell and tube oxygenator. Trans. Am. Soc. Artif. Intern. Organs **1968**, *14*, 242–249.
36. Srinivasan, S.; Tien, C. Reverse osmosis in a curved tubular membrane duct. Desalination **1971**, *9*, 127.
37. Winzeler, H.B.; Belfort, G. Enhanced performance for pressure-driven membrane processes: the argument for fluid instabilities. J. Membr. Sci. **1993**, *80*, 35.
38. Chung, K.-Y.; Bates, R.; Belfort, G. Dean vortices with wall flux in a curved channel membrane system. 4. Effect of vortices on permeation fluxes of suspensions in microporous membrane. J. Membr. Sci. **1993**, *81*, 139.
39. Chung, K.-Y.; Brewser, M.E.; Belfort, G. Dean vortices with wall flux in a curved channel membrane system. 2. The velocity field in a spiral channel. AIChE J. **1995**, *42*, 347–358.
40. Mallubhotla, H.; Nunes, E.; Belfort, C. Microfiltration of yeast suspensions with self-cleaning spiral vortices: possibilities for a new membrane module design. Biotechnol. Bioeng. **1995**, *48*, 375.
41. Ophoff, J.; Voss, G.S.; Racz, I.G.; Reith, T. Systematic approach of membrane module design based on hydrodynamics. The helically twisted tubular membrane module. Proceedings of ICOM'96, Yokohama, 1996.
42. Ophoff, J.; Voss, G.S.; Racz, I.G.; Reith, T. Flux enhancement by reduction of concentration polarisation due to secondary flow in twisted membrane tubes. Proceedings of Euromembrane University of Twente, The Netherlands, 23–27 June 1997; Kemperman, A.J.B., Koops, G.H., Eds.; 400–402.
43. Moulin, P.; Rouch, J.C.; Serra, C.; Clifton, M.J.; Aptel, P. Mass transfer improvement by secondary flows: Dean vortices in coiled tubular membranes. J. Membr. Sci. **1996**, *114*, 235–244.
44. Mallubhotla, H.; Belfort, G. Flux enhancement during Dean vortex microfiltration. 8. Further diagnostics. J. Membr. Sci. **1997**, *125*, 75–91.
45. Crowder, M.L.; Gooding, C.H. Spiral wound hollow fiber membrane wound: a new approach to higher mass transfer efficiency. J. Membr. Sci. **1997**, *137*, 17–29.
46. Manno, P.; Moulin, P.; Rouch, J.C.; Clifton, M.; Aptel, P. Mass transfer improvement in helically wound hollow-fibre ultrafiltration modules: bentonite and yeast suspensions. Sep. Purif. Tech. **1998**, *14*, 175.
47. Guigui, C.; Manno, C.; Moulin, P.; Clifton, M.J.; Rouch, J.C.; Aptel, P.; Laine, J.M. The use of Dean vortices in coiled hollow-fibre ultrafiltration membranes for water and wastewater treatment. Desalination **1998**, *118*, 73.
48. Gehlert, G.; Luque, S.; Belfort, G. Comparison of ultra- and microfiltration in the presence and absence of secondary flow with polysaccharides, proteins and yeast suspensions. Biotechnol. Prog. **1998**, *14*, 931.
49. Wille, M.; Bubolz, M.; Langer, G.; Werner, U. Enhanced efficiency of cross-flow microfiltration

in curved tubular membranes: experimental and theoretical studies. Proceedings of the 10th Annual Meeting of the North American Membrane Society, Cleveland, OH, U.S.A., 16–20 May 1998.

50. Mallubhotla, H.; Hoffmann, S.; Schmidt, M.; Vente, J.; Belfort, G. Flux enhancement during Dean vortex membrane nanofiltration. 10. Design, construction and system characterization. J. Membr. Sci. **1998**, *141*, 183–195.
51. Schnabel, S.; Moulin, P.; Nguyen, T.; Roizard, D.; Aptel, P. Removal of volatile organic components (VOC) from water by pervaporation with dense silicone hollow fibres. Separation improvement by Dean vortices. J. Membr. Sci. **1998**, *142*, 129.
52. Moulin, P.; Manno, P.; Rouch, J.C.; Serra, C.; Clifton, M.J.; Aptel, P. Flux improvement by Dean vortices: ultrafiltration of colloidal suspensions and macromolecular solutions. J. Membr. Sci. **1999**, *156*, 109.
53. Mallubhotla, H.; Schmidt, M.; Lee, K.H.; Belfort, G. Flux enhancement during Dean vortex tubular membrane nanofiltration. 13. Effects of concentration and solute type. J. Membr. Sci. **1999**, *153*, 259–269.
54. Mallubhotla, H.; Edelstein, W.A.; Earley, T.A.; Belfort, G. Magnetic resonance flow imaging and numerical analysis of curved tube flow. 16. Effect of curvature and flow rate on Dean vortex stability and bifurcation. AIChE J. **2001**, *47*, 1126–1140.
55. Luque, S.; Mallubhotla, H.; Gelhert, G.; Kuriyel, R.; Dzengeleski, S.; Pearl, S.; Belfort, G. A new coiled hollow fibre module design for enhanced microfiltration performance in Biotechnology. Biotechnol. Bioeng. **1999**, *65*, 247–257.
56. Kluge, T.; Kalra, A.; Belfort, G. Viscosity effects on Dean vortex membrane microfiltration. AIChE J. **1999**, *45*, 1913.
57. Kluge, T.; Rezende, C.; Wood, D.; Belfort, B. Protein transmission during Dean vortex microfiltration of yeast suspensions. Biotechnol. Bioeng. **1999**, *65*, 649–658.
58. Kuakuvi, D.N.; Moulin, P.; Charbit, F. Dean vortices: a comparison of woven versus helical and straight hollow fiber membrane modules. J. Membr. Sci. **2000**, *171*, 59–65.
59. Ghogomu, J.N.; Guigui, C.; Roucha, J.C.; Clifton, M.J.; Aptel, P. Hollow-fibre membrane module design: comparison of different curved geometries with Dean vortices. J. Membr. Sci. **2001**, *181*, 71–80.
60. Schutyser, M.; Rupp, R.; Wideman, J.; Belfort, G. Dean vortex membrane microfiltration and diafiltration of rBDNF *E.coli* inclusion bodies. Biotechnol. Prog. **2002**, *18*, 322–329.
61. Schutyser, M.; Belfort, G. Dean vortex membrane ultrafiltration non-Newtonian viscosity effects. Ind. Eng. Chem. Res. **2002**, *41*, 494–504.
62. Kaur, J.; Agrawal, G.P. Studies on protein transmission in thin channel flow module: the role of Dean vortices for improving mass transfer. J. Membr. Sci. **2002**, *196*, 1–11.
63. Al-akoum, O.; Mercier-Bonin, M.; Ding, L.; Fonade, C.; Aptel, P.; Jaffrin, M. Comparison of three different systems used for flux enhancement: application to cross flow filtration of yeast suspension. Desalination **2002**, *147*, 31–36.
64. Liu, S.H.; Luo, G.S.; Wang, Y.; Wang, Y.J. Preparation of the coiled hollow-fiber membrane and mass transfer performance in membrane extraction. J. Membr. Sci. **2003**, *215*, 203–211.
65. Gupta, B.B.; Howell, J.A.; Wu, D., et al. A helical baffle for cross-flow microfiltration. J. Membr. Sci. **1994**, *99*, 31–42.
66. Elmaleh, S.; Ghaffor, N. Cross-flow ultrafiltration of hydrocarbon and biological solid mixed suspensions. J. Membr. Sci. **1996**, *118*, 111–120.
67. Elmaleh, S.; Ghaffor, N. Upgrading oil refinery effluents by cross-flow ultrafiltration. Wat. Sci. Technol. **1996**, *34*, 231–238.
68. Broussous, L.; Ruiz, J.C.; Larbot, A.; Cot, L. Stamped ceramic porous tubes for tangential filtration. Sep. Purif. Tech. **1998**, *14*, 53–57.
69. Scott, K; Lobato, J. Mass transfer characteristics of cross-corrugated membranes. Desalination **2002**, *146*, 255–258.
70. Bubolz, M.; Wille, M.; Langer, G.; Werner, U. The use of Dean vortices for crossflow microfiltration: basic principles and further investigations. Sep. Purif. Tech. **2002**, *26*, 81–89.
71. Taylor, G.I. The criterion for turbulence in curved pipes. Proc. R. Soc. A **1929**, *124*, 243–249.
72. Chang, H.K.; Mockros, L.F. Convective dispersion of blood gases in curved channels. AIChE J. **1971**, *17*, 541–549.
73. Girloy, K.; Brighton, E.; Gaylor, J.D.S. Fluid cortices and mass transfer in a curved channel artificial membrane lung. AIChE J. **1977**, *23*, 106–115.
74. Patel, I.C.; Sirs, J.A. Indicator dilution measurement of flow parameters in curved tubes and

branching network. Phys. Med. Biol. hasis **1983**, *22*, 714–730.

75. Erdogan, M.E.; Chatwin, P.C. The effect of curvature and buoyancy on the laminar dispersion of solute in a horizontal tube. J. Fluid Mech. **1967**, *294*, 465–484.
76. Nunge, R.J.; Lin, T.S.; Gill, W.N. Laminar dispersion in curved tubes and Channels. J. Fluid Mech. **1972**, *52*, 363–383.
77. Topakoglu, H.C. Steady state laminar flow in incompressible viscous fluid in curved pipes. Math Mech. **1967**, *16*, 1321–1337.
78. Janssen, L.A.M. Axial dispersion in laminar flow through coiled tubes. Int. J. Heat Mass Transfer **1976**, *31*, 215–218.
79. Johnson, M.; Kamm, R.D. Numerical studies of steady flow dispersion at low Dean number in a gently curving tube. J. Fluid Mech. **1986**, *172*, 320–345.
80. Daskapoulos, P.; Lenhoff, A.M. Dispersion coefficient for laminar flow in curved tubes. AIChE J. **1988**, *34*, 2052–2058.
81. Eckman, D.M.; Grotberg, J.B. Oscillatory flow and mass transport in a curved tube. J. Fluid Mech. **1988**, *188*, 509–527.
82. Pedley, T.J.; Kamm, R.D. The effect of secondary motion on axial transport in oscillatory tube flow. **1988**, *193*, 347–367.
83. Sharp, M.K.; Kamm, R.D.; Shapiro, A.H.; Kimmel, E.; Karniadakis, G.E. Dispersion in a curved tube during oscillatory flow. J. Fluid Mech. **1991**, *223*, 537–563.
84. Agarwal, S.; Jayaraman, G.; Srivastava, V.K.; Nigam, K.D.P. Poer law fluids in a circular curved tube. I. Laminar flow. Polym. Plast. Technol. Eng. **1993**, *32*, 598–614; Agrawal, S.; Jayaraman, G.; Srivastava, V.K.; Nigam, K.D.P. Poer law fluids in a circular curved tube. II. Axial laminar dispersion. Polym.-Plast. Technol. Eng. **1993**, *32*, 615–634.
85. Agarwal, S.; Jayaraman, G. Numerical Simulation of dispersion in the flow of power law fluids in curved tubes. Polym. Plast. Technol. Eng. **1993**, *32*, 615–634.
86. Koutsky, J.A.; Adler, R.J. Minimization of axial dispersion by use of secondary flow in helical tubes. Can. J. Chem. Eng. **1964**, *42*, 239–246.
87. Van Andel, E.; Kramer, H.; Voogd, A. The residence time distribution of laminar flow in curved tubes. Chem. Eng. Sci. **1964**, *19*, 77–78.
88. Trivedi, R.N.; Vasudeva, K. Axial dispersion in laminar flow in helical coils. Chem. Eng. Sci. **1975**, *30*, 317–325.
89. Nigam, K.D.P.; Vasudeva, K. Influence of curvature and pulsation on laminar dispersion. Chem. Eng. Sci. **1976**, *31*, 835–857.
90. Mouljin, J.K.; Spijker, R.; Kolk, F.M. Axial dispersion of gases flowing through coiled columns. J. Chromatogr. **1977**, *142*, 155–166.
91. Shetty, V.D.; Vasudeva, K. Effect of Schmidt number in laminar dispersion in helical coils. Chem. Eng. Sci. **1977**, *32*, 782–783.
92. Van Den Berg, J.H.; Deelder, R.S. Measurement of axial dispersion in laminar flow through coiled capillary tubes. Chem. Eng. Sci. **1979**, *34*, 1345–1347.
93. Saxena, A.K.; Nigam, K.D.P. On RTD in coiled tube. Chem. Eng. Sci. **1979**, *34*, 425.
94. Singh, D.; Nigam, K.D.P. Laminar dispersion of polymer solutions in helical coils. J. Appl. Polym. Sci. **1979**, *23*, 3021–3025.
95. Saxena, A.K.; Nigam, K.D.P. Axial dispersion in laminar flow of polymer solution through coiled tubes. J. Appl. Polym. Sci. **1981**, *26*, 3475–3478.
96. Iyer, R.N.; Vasudeva, A.K. Laminar dispersion in helical coils. Chem. Eng. Sci. **1981**, *26*, 1104–1105.
97. Deelder, R.S.; Kuijpers, A.T.J.M.; van den Berg, J. Affinity chromatography study of the interaction of ribonucleotides with bovine pancreatic ribonuclease covalently bound to sepharose 4B. J. Chromatogr. **1983**, *255*, 545.
98. Singh, P.C.; Singh, S. Axial dispersion in gases flowing through small-bore helical column. Can. J. Chem. Eng. **1983**, *61*, 254–256.
99. Kauffman, A.D; Kissinger, P.T. Extra-column band spreading concerns in post-column photolysis reactors for microbore liquid chromatography. Curr. Sep. **1998**, *17*, 9–16.
100. Saxena, A.K. Laminar Dispersion in Helically Coiled Tubes. Ph.D. Thesis, Indian Institute of Technology, New Delhi, India, 1982.
101. Taylor, G.I. Dispersion of soluble matter in solvent flowing slowly through a tube. Proc. R. Soc. **1953**, *186*, 219.
102. Giddings, J.C. *Dynamics of Chromatography Principles and Theory Series*; Dekker: New York, 1965; Vol. 1.
103. Knox, J.H. Evidence for turbulence and coupling in chromatographic columns. Anal. Chem. **1966**, *38*, 253–261.
104. Desty, D.H.; Douglas, A.A. Study of new column forms in gas chromatography. J. Chromatogr. **1978**, *158*, 73.
105. Hofmann, K.; Halász, I. Mass transfer in ideal and geometrically deformed open tubes. 1. Ideal

and coiled tubes with circular cross-section. J. Chromatogr. **1979**, *173*, 211–228; Hofmann, K.; Halász, I. Mass transfer in ideal and geometrically deformed open tubes. 2. Potential application of ideal and coiled open tubes in liquid chromatography. J. Chromatogr. **1979**, *173*, 229–247; Hofmann, K.; Halász, I. Mass transfer in ideal and geometrically deformed open tubes. 3. Deformed tubes and plastic tubes. J. Chromatogr. **1980**, *199*, 3–22.

106. Tijssen, R. Liquid chromatography in helical coiled open tubular columns. Sep. Sci. Tech. **1978**, *13*, 681.
107. Tijssen, R. Axial dispersion and flow phenomena in helically coiled tubular reactors for flow analysis and chromatography. Anal. Chim. Acta **1980**, *114*, 71.
108. Katz, E.D.; Scott, R.P.W. Low-dispersion connecting tubes for liquid chromatography systems. J. Chromatogr. **1983**, *268*, 169.
109. Uihlein, M.; Schwab, E. A novel reactor for photochemical post-column derivatization in HPLC. Chromatographia **1982**, *15*, 140.
110. Grochowicz, P.R.; Simpson, C.F. Towards rapid chromatographic separations in open tubular serpentine columns; HPLC'96. Proceedings of the 20th International Symposium on High Performance Liquid Phase Separations, San Francisco, June 1996.
111. Grochowicz, P.R.; Simpson, C.F. Towards rapid open tubular chromatography; HPLC'97. Proceedings of the 20th International Symposium on High Performance Liquid Phase Separations, Birmingham, June 1997.

Materials Modeling

Sanat Mohanty
Gregg Caldwell
Manish Jain
Cristina U. Thomas
3M Company, St. Paul, Minnesota, U.S.A.

INTRODUCTION

Industrial scientists face the challenge of developing advanced materials by manipulating the relation between the chemical structure and the desired performance. Predicting the final performance property requires an integrated approach among the various length- and time scales of material behavior. Materials modeling must then answer questions at all these length scales integrating methods as shown schematically in Fig. 1. It must answer questions about bulk properties of a material and its behavior under external environments—thermal, mechanical, and electromagnetic. In designing a material with given bulk properties, one must understand how molecules or clusters of molecules can be engineered to attain these bulk properties. Thus, ab initio methods as well as atomistic and mesoscale methods become significant. This entry acts as a primer on various kinds of modeling techniques and points to references where any of these methods can be further explored.

CONTINUUM MODELS

Finite Element and Finite Difference

The behavior of materials is governed by the physical processes that act on those materials. Mathematical models of these physical processes are based on partial differential equations (PDEs). Most of the time, only materials undergoing simple processes can be treated with direct analytic solutions. Numerical methods then become the only alternative available for the solution of detailed and realistic models. Material developers call upon numerical methods to solve a PDE or a combination of PDEs on discrete set of points of the solution domain called ''discretization.'' Here, the solution domain is divided into subdomains having the discretization points as vertice; the distance between two adjacent vertices is the ''mesh size.'' Time is also subdivided into discrete intervals; ''timestep'' is the interval between two consecutive times at which the solution is obtained. The PDE is then approximated, or discretized, to obtain a system of algebraic equations, where the unknowns are the solution values at the discretization points. This system of algebraic equations can then be solved on a computer by ''direct'' or ''iterative'' techniques. The approximate solution of the original PDE has an error called ''discretization error.''

The finite difference approach is a widely used discretization technique because of its simplicity. Finite difference approximations of derivatives are obtained by using truncated Taylor series. The following Taylor expansions can be used:

$$u(x + \Delta x) = u(x) + \Delta x \frac{\partial u}{\partial x} + \frac{\Delta x^2}{2} \frac{\partial^2 u}{\partial x^2} + \frac{\Delta x^3}{6} \frac{\partial^3 u}{\partial x^3} + O(\Delta x^4) \quad (1)$$

$$u(x - \Delta x) = u(x) - \Delta x \frac{\partial u}{\partial x} + \frac{\Delta x^2}{2} \frac{\partial^2 u}{\partial x^2} - \frac{\Delta x^3}{6} \frac{\partial^3 u}{\partial x^3} + O(\Delta x^4) \quad (2)$$

The first order derivative is given by the following finite difference approximations: from Eq. (1): forward difference

$$\frac{\partial u}{\partial x} = \frac{u(x + \Delta x) - u(x)}{\Delta x} + O(\Delta x) \quad (3)$$

from Eq. (2): backward difference

$$\frac{\partial u}{\partial x} = \frac{u(x) - u(x - \Delta x)}{\Delta x} + O(\Delta x) \quad (4)$$

By subtracting Eq. (1) from Eq. (2): centered difference

$$\frac{\partial u}{\partial x} = \frac{u(x + \Delta x) - u(x - \Delta x)}{2\Delta x} + O(\Delta x^2) \quad (5)$$

An approximation for the second order derivative is obtained by adding Eq. (1) to Eq. (2):

$$\frac{\partial^2 u}{\partial x^2} = \frac{u(x + \Delta x) - 2u(x) + u(x - \Delta x)}{\Delta x^2} + O(\Delta x^2) \quad (6)$$

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120023856

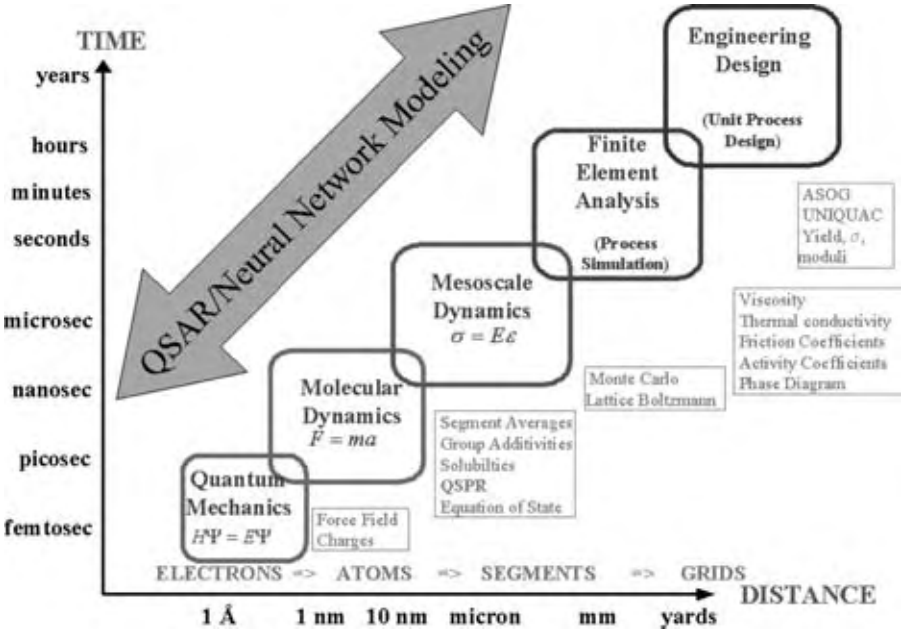


Fig. 1 Schematic of multiscale modeling approaches. *(View this art in color at www.dekker.com.)*

The terms $O(\Delta x)$ and $O(\Delta x^2)$ are the truncation errors. If a better approximation is needed, one could either reduce the size of the mesh or add more information by including higher order neighbors.

Finite element methods (FEM) are used to simulate physical systems governed by ordinary and partial differential equations. The basic concept is the separation of a complex materials system into simpler, disjoint entities called finite elements. A good analogy to illustrate this comes from structural mechanics where the mechanical response of an element is characterized by degrees of freedom. A set of node points (or simply nodes) represent values of the unknown functions for degrees of freedom. The element response is described by algebraic equations. The total response of the original system is approximated by that of the discrete model constructed by assembling all the elements. This separation process intuitively occurs when materials scientists or engineers analyze systems. Mathematically, FEM obtain numerical approximations to differential equations in a domain Ω, and this domain is replaced by the union of disjoint subdomains E, called finite elements. The unknown functions are locally approximated over each element by an interpolation expression (called shape functions) written in terms of values taken by the functions (and their derivatives in some cases) at a set of nodes located on the element boundaries. The union of the shape functions over adjacent elements forms an approximation, which is inserted into the governing equations to determine the unknown values by the Galerkin, least-squares, or Ritz methods.

USES OF FEM MODELS

Materials design via virtual prototyping is being increasingly incorporated into the product development process because of their criticality in minimizing costs and improving time-to-market. By including the effects of multiple physics when they are coupled together, FEM allow the researcher to address real-world structural, thermal, electromagnetic, and fluid-flow behaviors of 3-D product designs.

FEM have provided a wide range of solutions in many industries. For example, automotive engineers evaluate and optimize many performance and reliability aspects of their designs in sheet metal forming, gasket analysis, seal integrity, noise and vibration, crash simulations, etc. Aerospace/defense engineers routinely perform static, dynamic, and coupled acoustic–structural analysis of aircraft frames; simulations of space structures (solar sails, space radar, and antennas); wing panel buckling and crack propagation in the fuselage; bird strike simulations; the effects of underwater explosions on ships and submarines; structural analysis of submarine frames; and simulation of nozzles, piezoelectric motors, self-lubricating hinges, and various bearings. Consumer/electronics engineers utilize FEM to simulate the design, manufacturing, and performance of different types of consumer and electronic products and associated packaging including: drop testing of consumer electronics and fluid-filled containers; the manufacturing of new materials, such as paper and film; design and validation of acoustic properties of audio equipment; thermomechanical simulation of electronic products; thermal analysis of chip packaging; and the

design of consumer appliances such as washers and dryers. Manufacturing engineers design their equipment (pumps, motors, and compressors); perform fatigue analysis of power generation equipment, such as boilers, turbines, and heat exchangers; dynamic analysis of piping systems under various loads, such as thermal, pressure, seismic, etc. Biomedical engineers engage FEM in the design of mechanical connectors commonly used in all types of medical devices, simultaneously enhancing proprietary issues and regulatory approvals. They commonly use FEM to estimate the wear of materials used in dental implants, and hip, knee, and other prosthetic implants. In the rubber industry, developers seek the use of FEM to design bushings, seals, vibration damping systems, tire–road interactions, and tire wear.

SYSTEMS MODELING

Materials in the macroscopic sense follow laws of continuum models in which the nanoscale phenomenon is accounted for by statistical averages. Continuum models and analysis separate materials into solids (structures) and fluids. Computational solid mechanics and structural mechanics emphasize the analysis of solid materials and its structural design. Computational fluid mechanics treats material behaviors that involve the equilibrium and motion of liquid and gases. A relative new area, called multiphysics, includes materials systems that contain interacting fluids and structures such as phase changes (solidification, melting), or interaction of control, mechanical and electromagnetic (MEMS, sensors, etc.).

A more general concept is systems. Systems identify material objects that combine to perform a distinguishable or desired function. Examples of systems are airplanes, buildings, engines, electronic chips, trees, cats, human organs and cells, etc. There are no simple differential equations that can capture the diversity and the complexity of systems. It is important then to divide and conquer and finish with abstraction from these parts. Nowadays, materials designers define the system with the required functionality and separate it into subsystems, with each subsystem having its own design requirements. These subsystems can be subdivided again until components are specified. Components are sufficiently simple in geometry, connectivity, and raw materials so that they can be reasonably described by continuum mathematical models. For example, in the FEM, components are the finite elements and are obtained from FEM libraries based on mathematical models. The system model is then obtained (abstraction) by reversing the process, from component equations to subsystem equations, and from those equations to the complete system. In mechanical design, the system assembly process is governed by the classical principles of Newtonian mechanics. In molecular design, the system assembly process involves a more complex multiscale approach. In complex materials design, the processing parameters/environment play a key role in determining the final properties of the materials, and hence need to be part of the ''assembly process'' to properly describe the final system. When designing materials systems, one needs to address the issue of system optimization and realize that combinations of optimized subsystems do not necessarily yield an optimized system. One needs to involve the use of multidisciplinary optimization for a truly optimized system. Here, trade-offs are necessary as subsystems and the system do not share the same optimal point. Great advances in multidisciplinary optimization and computational power are enabling greater system optimization in which multidisciplinary teams work together rather than adopt a piecemeal approach of subsystem optimization.

ELECTRONIC AND ATOMISTIC MODELING

Macroscale models cannot answer questions regarding interaction between molecules, rates of change in these interactions, and the structures that evolve—questions that are important in catalysis, separation processes and media, surface engineering, electronics processing, among others. Models at the atomistic scale capture phenomena relevant to these questions. Significant advances have been made at the other extreme of the length- and time-scale spectrum for the development of methods that can accurately describe electronic structures of materials, specific interactions between atoms, as well as mobility of these atoms to engineer new materials.

Ab Initio Methods

Models that attempt to predict the behavior of materials using first principles quantum theory fall within this regime. These methods are applied to the development of ''traditional'' materials such as steels, refractory materials, ceramics, etc. as well as new materials such as those for microelectronics industries, catalysts of various kinds,[1,2] materials for fuel cell applications, to name a few. Some examples of such properties are electronic properties of solids such as conductivity, absorption spectra, etc.,[3] reactivity of molecules,[4] selective binding of molecules to specific sites on surfaces, catalytic reaction pathways, and active sites on molecules.

To gain insights into the above-mentioned problems, one needs to solve Schrodinger's equation:

$$H\Psi(\{\vec{r}\},\{\vec{R}\}) = E\Psi(\{\vec{r}\},\{\vec{R}\}) \quad (7)$$

$$H = T_{\text{electron}} + T_{\text{nuclei}} + U_{\text{electron-electron}} + U_{\text{electron-nuclei}} + U_{\text{nuclei-nuclei}} \quad (8)$$

where H is the Hamiltonian of the system, E the energy, Ψ the wave function (which is a function of electronic coordinates $\{\vec{r}\}$ and nuclear coordinates $\{\vec{R}\}$), T_{electron} the operator for kinetic energy of the electrons, T_{nuclei} for kinetic energy of the nuclei, $U_{\text{electron-electron}}$ for interaction between the electrons, $U_{\text{electron-nuclei}}$ for interaction between the nuclei and the electrons, and $U_{\text{nuclei-nuclei}}$ for interaction between the nuclei.

This equation has $3N(Z + 1)$ degrees of freedom, where N is the number of atoms and Z is the atomic number of the atoms. Apart from a few simple cases, an exact solution of this problem is far beyond the currently available computational power. A number of approximations have to be made to solve this many-body equation. The first and foremost is the Born–Oppenheimer approximation.[5] This is based on the fact that there is a huge difference in the mass of the electrons and the nuclei. It assumes that the motion of the electrons is determined by the instantaneous positions of the nuclei, and that of the nuclei by the average position of the electrons. The nuclear degrees of freedom thus only enter Schrodinger's equations as parameters. Density functional theory[6] is used to reduce the many-electron-problem to a series of coupled single particle equations. The central theorem for this theory implies that the charge density of the system uniquely determines the Hamiltonian of the system. Thus, all quantities that can be deduced a priori when the Hamiltonian is fixed can be written as functionals of the charge density. This theory (for which Walter Kohn got the Nobel Prize in 1998) is formally exact, but in practice some approximation has to be made for the exchange-correlation functional.[7] Besides this, often one also uses the pseudopotential approximation;[8] as only the valence electrons count in any chemical reaction, only they are considered explicitly. This approximation treats the remaining ''core'' electrons via an effective potential called pseudopotential. With these approximations one is left to solve:

$$\left(\frac{-\nabla^2}{2} + V_{\text{effective}}\right)\phi_i = \varepsilon_i\phi_i \quad (9)$$

where

$$V_{\text{effective}} = V_{\text{hartree}}[\rho] + V_{\text{ion}} + V_{\text{exchange-correlation}}[\rho] \quad (10)$$

$V_{\text{hartree}}[\rho]$ is the coulombic interaction between the electrons, V_{ion} the interaction between the nuclei and the electrons (either as a pseudopotential or just $1/r$ potential), and $V_{\text{exchange-correlation}}[\rho]$ the exchange-correlation potential from density functional theory. The charge density is given as

$$\rho(r) = \sum_{i=1,\text{occupied}} \phi_i^*\phi_i \quad (11)$$

The above three equations are solved self-consistently[9] to give a charge density and potential. Properties are then calculated based on this charge density and potential. This formalism based on density functional theory has become a very popular tool for investigating properties of materials.

Molecular Mechanics

It is not always necessary to detail the electronic behavior of materials; an accurate understanding of the atomic interactions is often sufficient to describe the phenomenon of interest with reasonable accuracy. In contrast to ab initio methods, molecular mechanics is used to compute molecular properties, which do not depend on electronic effects. These include geometry, rotational barriers, vibrational spectra, heats of formation, and the relative stability of conformers. As the calculations are fast and efficient, molecular mechanics can be used to examine systems containing thousands of atoms. However, unlike ab initio methods, molecular mechanics relies on experimentally derived parameters so that calculations on new molecular structures may be misleading.

The idea of building models to represent molecular structure is a fundamental tool that is used to understand chemistry and the structures that atoms adopt as they are found in molecules. In the not too distant past, molecular modeling was accomplished using plastic balls and straws. Watson and Crick discovered the structure of DNA using this type of model[10] in 1953. These plastic models were truly atomistic in nature in that they treat each atom as a discrete entity, with complex arrangement of electrons treated explicitly.

Today, computational techniques are used to model the conformational behavior and energetic properties of molecules. Molecular mechanics (MM) is a mathematical formalism, which uses a classical, physical

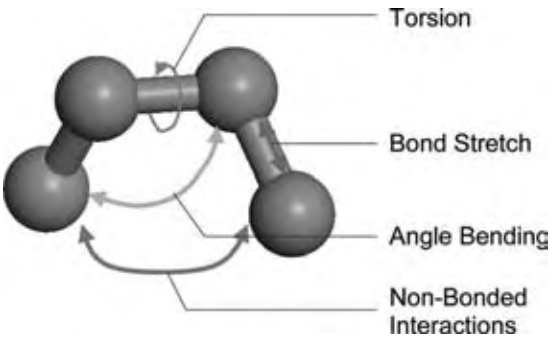


Fig. 2 The mechanical molecular model considers atoms as spheres and bonds as springs. The mathematics of spring deformation can be used to describe the ability of bonds to stretch, bend, and twist. *(View this art in color at www.dekker.com.)*

description of the molecule to calculate the energy of that particular conformation:

$$\text{energy} = f(\text{nuclear positions}) \tag{12}$$

Based on our history of physical and chemical experiences, we can think of molecules as mechanical assemblies made up of simple elements like balls (atoms), rods (bonds), and flexible joints (bond angles and torsional angles) —see Fig. 2. Eq. (12) can then be expressed as

$$\text{energy} = E_{\text{stretching}} + E_{\text{bending}} + E_{\text{torsion}} + E_{\text{nonbonded interactions}} \tag{13}$$

Eq. (13) makes the distinction between bonded (interactions between atoms that are connected by no more than three bonds) and nonbonded (interactions between atoms that are not connected to each other at all) interactions . The nonbonded energy accounts for repulsion, van der Waals attraction, and electrostatic interactions. van der Waals attraction occurs at short range and rapidly dies off as the interacting atoms move apart by a few angstroms. Repulsion occurs when the distance between interacting atoms becomes even slightly less than the sum of their contact radii.

The electrostatic interaction is modeled using a coulombic potential. The electrostatic energy is a function of the charge on the nonbonded atoms, their interatomic distance, and a molecular dielectric expression that accounts for the attenuation of electrostatic interaction by the environment (e.g., a solvent or the molecule itself).

The simplest calculation in molecular mechanics is the calculation of the potential energy of the system, which is performed by summing the numerous energy terms for the given conformation of the system using the given set of potential energy functions and parameters:

$$E_{\text{potential}} = \sum E_{\text{stretching}} + \sum E_{\text{bending}} + \sum E_{\text{torsion}} + \sum E_{\text{nonbonded interactions}} \tag{14}$$

Typically though, MM calculations are used to find the minimum energy conformations for a given molecule. There are numerous energy techniques used to find the minimum energy, from simple grid searching to more complex methods such as ''steepest descent,'' ''conjugate gradient,'' and ''Newton–Raphson.'' These methods tend to find the local minimum and not the global minimum, as the algorithms do not ''climb over'' energy barriers. To optimize the minimization procedure, it is usually best to combine several algorithms in a minimization scheme.

Some common applications of energy minimization in macromolecular systems are relieving strain in experimentally obtained structures, refinement of molecular models, or in search protocols aimed at finding the global energy minimum of the system.

Molecular Dynamics

All chemical process involves the motion of atoms within a molecule. Molecular dynamics (MD), in the broadest sense, is concerned with molecules in motion. It combines the energy calculations from molecular mechanics with equations of motion. Generally, an appropriate starting structure is selected (normally an energy minimized structure). Each atom in the system is then assigned a random velocity that is consistent with the Maxwell–Boltzmann distribution for the temperature of interest. The MM formalism is used to calculate the forces on all the atoms. Once the atom positions are known, the forces, velocities at time t, and the position of the atoms at some new time $t + \delta t$ can be predicted. More details about the method can be found in Ref.[11].

Molecular dynamics simulations have been used in a variety of ways. They can be used to compute mechanical moduli by studying the response of a model of the bulk polymer to a constant stress or strain, and to study the diffusion of molecules in membranes and polymers.[12] There are numerous biomolecular applications. Structural, dynamic, and thermodynamic data from molecular dynamics have provided insights into the structure–function relationships, binding affinities, mobility, and stability of proteins, nucleic acids, and other macromolecules that cannot be obtained from static models.

Monte Carlo Simulations

Often we are only interested in the equilibrium structure of a set of molecules. When the system is at high densities or includes a large number of conformations, other methods become computationally unfeasible, and one often uses Monte Carlo (MC) techniques. Monte Carlo methods use random number generators to integrate systems with very high degrees of freedom.[13] These integrals are then used in theories of statistical mechanics to evaluate thermodynamic properties of materials.[14]

Any macroscale thermodynamic state can be represented by a very large number of consistent microstates. The thermodynamic properties of the material are a result of contributions from these various microstates. Monte Carlo simulations ensure that these microstates are efficiently accounted for. Those that contribute significantly to the thermodynamic state are given greater importance through importance sampling algorithms.[13] Monte Carlo simulations further ensure that less accessible states are also accounted for[15] through various biasing methods.

STATISTICAL CORRELATIONS

While first principles modeling methods are most sought after—as they help one understand the mechanisms behind a phenomenon and hence engineer the phenomenon—it is not always possible owing to the complexity of certain systems or processes. In such situations, neural network models are often used to help understand the system or to optimize the system without necessarily understanding it completely.

Quantitative Structure–Property and Structure–Activity Relationships (QSPR/QSAR)

The idea that the behavior and properties of a molecule are concealed in the fundamental structural formulae has been around for a long time. Modern physical chemistry has become increasingly orientated toward understanding how these properties can be decoded from the structure. Ideally, all properties of a chemical compound would be calculated from first principles. This is, however, unlikely in the foreseeable future because of a number of reasons, including the lack of sufficient theory and limits of available computational power. An alternative approach to finding qualitative mathematical relationships between the intrinsic molecular structure and observable properties of a chemical compound will be extremely valuable to both industrial and academic chemists.

The concept of QSAR/QSPR is to transform searches for compounds with desired properties using chemical intuition and experience into a mathematically quantified and computerized form. Available programs enable scientists to easily generate and present molecular data including geometries, energies, and associated properties (electronic, spectroscopic, and bulk). The usual paradigm for displaying and manipulating these data is a table in which compounds are defined by individual rows, and molecular properties (or descriptors) are defined by the associated columns. A QSAR/QSPR attempts to find consistent relationships between the variations in the values of molecular properties and the biological activity or material property for a series of compounds so that these ''rules'' can be used to evaluate new chemical entities. A QSAR generally takes the form of a linear equation:

$$\text{Activity/property} = \text{Const.} + (C_1 \times P_1) + (C_2 \times P_2) + (C_3 \times P_3) + \cdots \quad (15)$$

where P_1 to P_n are descriptors computed for each molecule, and the coefficients C_1 to C_n are fitted so as to give the best correlation to the activity or property of interest.

Once a correlation between structure and activity/property is found, any number of compounds, including those not yet synthesized, can be readily screened on the computer to select structures with the properties desired. The most promising compounds may then be selected for synthesis and evaluation in the laboratory. In the pharmaceutical[16] and environmental areas,[17] QSPR has been applied to predict properties like the efficacy of drugs (with different substitutions) or behaviors like the accumulation of pollutants in fish, despite the fact that the exact mechanisms are unknown. The QSAR/QSPR approach conserves resources and accelerates the process of development of new molecules for use as drugs, materials, additives, or for any other purpose.

It is not easy to find successful structure–activity/property correlations, but the rapid growth of publications dealing with QSAR/QSPR studies clearly demonstrates the progress in this area. To obtain a significant correlation, it is crucial that appropriate descriptors be employed, whether they are theoretical, empirical, or derived from readily available experimental characteristics of the structures. Many descriptors reflect simple molecular properties and thus can provide insight into the physicochemical nature of the activity/property under consideration.

The QSPR/QSAR methods have many direct benefits like property prediction, target molecular design, and structural refinement, and indirectly it can help to

elucidate not-so-obvious physics, unknown mechanisms, or molecular behaviors.[18]

Neural Networks

Artificial neural network (ANN) is an information-processing paradigm inspired by the way the densely interconnected, parallel structure of the mammalian brain processes information. The processing ability of the network is stored in the interunit connection strengths, or "weights," obtained by a process of adaptation to, or "learning" from, a set of training patterns.[19] Neural networks can be used in the same general applications as QSPR/QSAR. While there is software available for generating a neural network, it is still very much an art to understand the model outputs, especially when trying to elucidate mechanistic information.

NANOSCALE AND MESOSCALE MODELING

Modeling at the continuum scale is well understood, and the accuracy, precision, and robustness of results using continuum methods have been analyzed. Modeling at the atomistic scale has also begun to be accepted as part of routine research protocol. Results from such methods are now accepted with a reasonable understanding of their accuracy and robustness. The weak link in the modeling spectrum is the area of mesoscopic and nanoscopic modeling. This is the scale where one cannot use continuum scale assumptions and yet the system one deals with is often larger than a few molecules, thus precluding atomistic modeling. How do we study systems that are made of clusters of thousands of molecules that interact in specific ways? How do we take results from atomistic simulations and correlate them to parameters we need for continuum studies?

Nanoscale is broadly defined as phenomena that occur in the range of 10–100 nm. Modeling nanoscale phenomena in industrial materials design becomes increasingly significant, as nanoscale engineering of materials becomes critical in developing superior materials. This includes engineering of molecular clusters or of functionalized nanoparticles, crystals, and liquid crystalline structures (such as clathrates and micelles) that are used as templates for porous membranes;[20] materials with specific enzymatic or catalytic properties;[21] control of flow behavior; nanocontainers with controlled release,[22] or materials with specific thermal, electromagnetic,[23] or mechanical properties.[24] As a significant fraction of these materials is designed by self-assembly of atoms or molecules to form nanosized clusters, modeling becomes a critical tool in engineering such materials. Modeling is also used to understand conditions or optimize processes to functionalize or engineer aspects of these structures—such as size of cluster, surface properties, or charges. Fig. 3 shows some examples of nanoscale clusters.

From a modeling perspective, there are a number of aspects to be addressed. Modeling helps to translate macroscopic observations based on nano- or mesoscale phenomena. Judicious use of simulations helps understand why certain behavior is seen—critical knowledge for engineering such behavior. For example, how can one predict the behavior of silicon nanoparticles whose surfaces have been modified with specific functional groups when they are in a polymer matrix[25]—or alternatively, how can one engineer the surfaces of nanoparticles so that polymer-particle composites may manifest a certain set of desired properties. Similarly, what kind of structures do micelles made up of specific amphiphiles form and how do they behave at different concentrations—or alternatively, what additives and amphiphiles could one use to engineer certain rheological properties in fluids, engineer specific surface properties, or design specific micellar structures,[26] how do molecules cluster in various environments,[27] and how does that affect their thermodynamic properties and their mobility. Another area with significant modeling effort is vis-à-vis biomolecules.[28] Simulations have been used to understand binding mechanisms of biomolecules on surfaces[29] and with each other, enzymatic pathways, mobility of molecules, design of encapsulants, etc.—all with the aim of engineering drugs for specific action as well as methods of targeted delivery. Nano- and mesoscale models can be used not only to look at the equilibrium behavior of materials but also to study mobility of particles and rupture behavior of adhesives—essentially nonequilibrium phenomena—as well.

Modeling mesosystems is difficult as the phenomenon of interest is neither atomistic, so that solutions can be grasped by understanding the behavior of a few atoms and computers are still not fast enough to deal with thousands of molecules rigorously, nor is it macroscopic so that continuum properties of the material can be assumed without losing events occurring in the smaller scale regime. Modeling phenomena that span different orders of length- and time scale require the use of multiscale models (schematic in Fig. 1).[30] Two kinds of multiscale methods—hierarchical and handshake—have provided feasible solutions to real problems.

Approximating a number of polymer molecules or solvent molecules as single beads with tunable interactions is an example of hierarchical models.[31] These models use information lumped from a smaller size scale or shorter time scales to predict phenomena at the scale of simulation and/or predict parameters or observations at larger scales. The parameters of the bead, in this case, are obtained from atomistic or quantum calculations on the molecules. The beads and

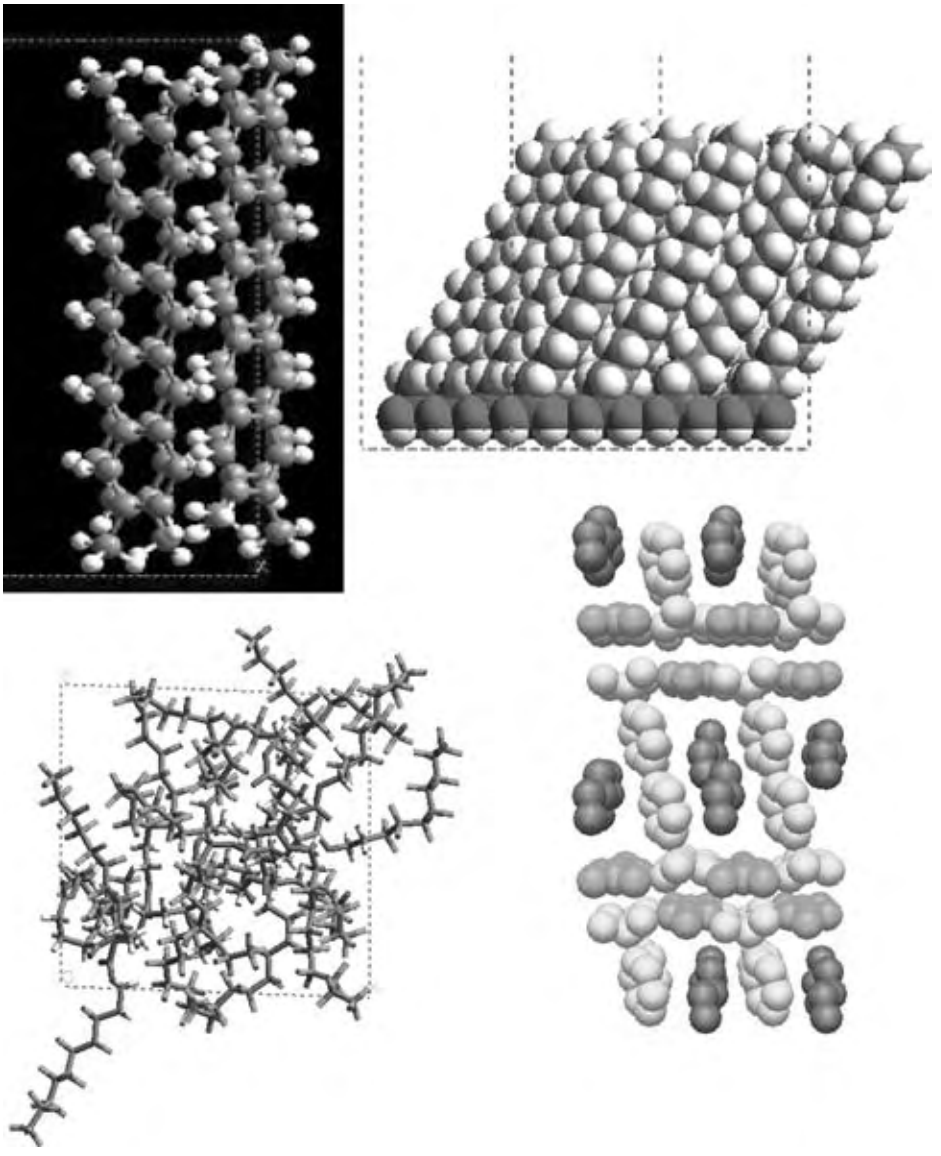


Fig. 3 Examples of nanoscale clusters whose properties may be controlled to engineer materials with specific thermal, mechanical, or electromagnetic properties. *(View this art in color at www.dekker.com.)*

chains, in turn, are used to predict viscosity and material properties, which can then be used in flow models or structure/stress models to predict macroscopic behavior of the material. Such methods are used to study equilibrium structure of polymers[32] and composites as well as nonequilibrium processes such as relaxation and/or fracture.[33] More complex versions of this algorithm have also been used in the study of nanoparticle interactions with polymers as well as a variety of solvents.[34]

Lattice models consist of spanning the system of interest by meshes or nodes. Molecules that are moving or reacting are lumped into coarser particles, and the nodes or the meshes keep a track of the extent of mobility or reaction. Examples include curing of polymers, diffusion in porous structures, and stress calculations in materials.

Handshake methods describe the problem by identifying one area where continuum assumptions hold while focusing on atomistic or mesoscale models to solve another aspect in a way that these two regions influence each other. Dynamic fracture is a very good example.[35] Energy from large-scale elastic fields is concentrated on the angstrom scale of the electrons that participate in atomic bonding. A simulation of this phenomenon requires an accurate description of atoms bonding at the crack tip, while at the same time including a proper description for very large volumes of strained material, with the resolution varying with distance from the crack tip. Far away, it is adequate to use the equations of motion for a macroscopic-averaged continuum field. This spatial decomposition makes it possible to combine different simulation methods describing the different physical regions into

a single, powerful simulation tool. Another example is the use of a handshake model to predict micelle free-energies: an analytical solution suffices in the continuum assumption of the hydrocarbon core, while molecular simulations are required to understand the head groups and their interactions with the solvent.[36]

Alternatively, different molecular models may be used in different regions. The combination of quantum mechanics (QM) (close to the docking site) and molecular mechanics (MM) (in the rest of the protein) methods allows one to predict the behavior of molecules docking on proteins. Particular attention must also be paid to an accurate joining of the two regimes at the boundary region.

Today, a significant fraction of related literature is on model systems, which explains trends and behavior of ''ball-spring models'' but is of little use in predicting the behavior of systems with specific chemistry. While there is a growing effort to solve real systems of industrial importance, this certainly is the area of weakness in the modeling toolbox today.

CONCLUSIONS

With the advent of powerful computational tools and algorithms, applications of modeling have become widespread. Modeling now not only offers a better understanding of materials phenomenon, but also has become a predictive tool useful in materials design. In this entry, we have described some modeling techniques that are useful for modeling phenomenon at various length- and time scales. These techniques allow industrial scientists to model real-world applications (and not just model systems). These techniques—at atomistic, mesoscale, and macroscale—when used either alone or in conjunction with one another, provide scientists with powerful tools to investigate properties of materials.

REFERENCES

1. Hammer, B.; Norskov, J.K. *Advances in Catalysis: Impact of Surface Science on Catalysis*; Elsevier: Vol. 45, 71 pp.
2. Group IV Zwitterion ANSA Metallocene (ZAM) Catalysis for α Olefin Polymerization. US Patent 5,939,503, 1999.
3. Chelikowsky, J.R.; Louie, S.G.; Martinez, G.; Shirley, E.L. The optical properties of materials (MRS 1999). In *Electronic Structure and Optical Properties of Semiconductors*, 2nd Ed.; Cohen, M.L., Chelikowsky, J.R., Eds.; Springer: Berlin, 1989.
4. Cramer, C.J. *Essentials of Computational Chemistry: Theories and Models*; John Wiley & Sons, 2002.
5. Born, M.; Oppenheimer, R. Ann. Phys. **1927**, *84*, 457.
6. Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. Phys. Rev. **1964**, *136*, B864; Kohn, W.; Sham, L. Self-consistent equations including exchange and correlation effects. Phys. Rev. **1965**, *140*, A1133.
7. Ceperley, D.M.; Alder, B.J. Ground state of the electron gas by a stochastic method. Phys. Rev. Lett. **1981**, *45*, 566; Perdew, J.; Zunger, A. Self-interaction corrections to density-functional approximations for many-electron systems. Phys. Rev. B **1981**, *23*, 5048.
8. Chelikowsky, J.R.; Cohen, M.L. Chapter 3. In *Handbook on Semiconductors*; Moss, T.S., Landsberg, P.T., Eds.; Elsevier: Amsterdam, 1992; Vol. 1.
9. Payne, M.C.; Teter, M.P.; Allan, D.C.; Arias, T.A.; Joannopoulos, J.D. Iterative minimization techniques for ab inito total-energy calculations; molecular dynamics and conjugate gradients. Rev. Mod. Phys. **1992**, *64*, 1045; Pickett, W. Pseudopotential methods in condensed matter applications, Comput. Phys. Rep. **1989**, *9*, 115; Beck, T.L. REal-space mesh techniques in density-functional theoryRev. Mod. Phys. **2000**, *72*, 1041–1080; Chelikowsky, J.R. The pseudopotential-density functional method applied to nanostructures. Physica D **2000**, *33*, R33.
10. Watson, J. *The Double Helix, A Personnel Account of the Discovery of the DNA*; Touchstone Publishers, 2001.
11. Rapaport, D.C. *The Art of Molecular Dynamics Simulation*; Cambridge University Press, 1995.
12. Jacobson, S.H. Molecular modeling studies of polymeric transdermal adhesives: structure and transport mechanisms. Pharmaceut. Technol. **1999**, September, 122–130.
13. Allen, M.P.; Tildesly, D.J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, 1987.
14. McQuarrie, D.A. *Statistical Mechanics*; Harper and Row: New York, 2000.
15. Frenkel, D.; Smit, B. *Understanding Molecular Simulations: From Algorithms to Applications*; Academic Press: San Diego, 1996.
16. Topliss, J.G. *Quantitative Structure–Activity Relationships of Drugs*; Academic Press: New York, 1983.
17. Carreira, L.A.; Hilal, S.; Karickhoff, S.W. Estimation of chemical reactivity parameters and physical properties of organic molecules using SPARC. In *Theoretical and Computational Chemistry, Quantitative Treatment of Solute/Solvent Interactions*; Politzer, P., Murray, P., Eds.; Elsevier, 1994.

18. Katritzky, A.R.; Lobanov, V.S.; Karelson, M. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. Chem. Soc. Rev. **1995**, *24*, 279–287.
19. Gurney, K. *An Introduction to Neural Networks*; UCL Press: London, 1997.
20. Yan, H.; Blanford, C.F.; Lytle, J.C.; Carter, C.B.; Smyrl, W.H.; Stein, A. Influence of processing conditions on structures of 3D ordered macroporous metals prepared by colloidal crystal templating. Chem. Mater. **2001**, *13*, 4314–4321.
21. Lim, M.H.; Blanford, C.F.; Stein, A. Synthesis of ordered microporous silicates with organosulfur surface groups and their applications as solid acid catalysts. Chem. Mater. **1998**, *10*, 467–470.
22. Kartik, A.; Radhakrishna, S.; Matthew, P.S.; Justin, L.P. Surfactant induced effects on turbulent swirling flows. Rheol. Acta **2002**, *41*, 25–34.
23. Kietzke, T.; Neher, D.; Landfester, K.; Montenegro, R.; Guntner, R.; Scherf, U. Novel approaches to polymer blends based on polymer nanoparticles. Nature Mater. **2003**, *2*, 408–412.
24. Hasegawa, N.; Okamoto, H.; Kawasumi, M.; Usuki, A. Preparation and mechanical properties of polystyrene-clay hybrids. J. Appl. Polym. Sci. **1999**, *74*, 3359.
25. Balazs, A.; Ginzburg, V.V.; Qiu, F.; Peng, G.; Jasnow, D. Multiscale model for binary mixtures containing nanoscopic particles. J. Phys. Chem. B **2000**, *104*, 3411–3422.
26. Bell, R.C.; Wang, H.; Iedema, M.J.; Cowin, J.P. Sculpting nano-scale liquid interfaces. In *Mesoscale Phenomena in Fluid Systems*; Case, F., Ed.; ACS Symposium Series, 2003.
27. Puvvada, S.; Blankschtein, D. Thermodynamic description of micellization, phase behavior and phase separation of aqueous solutions of surfactant mixtures. J. Phys. Chem. **1992**, *96*, 5567–5579.
28. http://kandinsky.chem.wisc.edu/~qiang/ (accessed July 2003).
29. Wang, P.; Vaidehi, N.; Tirrell, D.A.; Goddard III, W.A. Virtual screening for binding of phenylalanine analogues to phenylalanyl-tRNA synthetase. JACS **2002**, *124*, 14,442.
30. Thomas, C.U.; Caldwell, G.; Ross, R.B.; Mohanty, S.; Freedman, M. Importance of mesoscopic structures in the development of advanced materials. In *Mesoscale Phenomena in Fluid Systems*; Case, F., Ed.; ACS Symposium Series, 2003.
31. Clementi, E. Global scientific and engineering simulations on scalar, vector and parallel LCAP-type supercomputers. Phil. Trans. R. Soc. Lond. **1998**, *A326*, 445–470.
32. Theodorou, D.; Dodd, L.R. Atomistic Monte Carlo simulation and continuum mean field theory of the structure and equation of state properties of alkane and polymer melts. Adv. Polym. Sci. **1994**, *116*, 249.
33. Rottler, J.; Robbins, M.O. Yield conditions for deformation of amorphous polymer glasses. Phys. Rev. E **2001**, *64* (5), 051801.
34. Mei, W.; Mc Cormick, A.V.; Scriven, L.E. Kinetic gelation modeling: structural inhomogeneity during cross-linking polymerization. Macromolecules **2003**, *36*, 4140.
35. Bernstein, N. Multiscale modelling of materials. In *Multiscale Simulations of Brittle Fracture and the Quantum-Mechanical Nature of Bonding in Silicon,* MRS 2000 Fall Meeting Proceedings, Pittsburgh, 2001; Kubin, L.P., Bassani, J.L., Cho, K., Gao, H., Selinger, R.L.B., Eds.; Materials Research Society, 653.
36. Mohanty, S.; Davis, H.T.; Mc Cormick, A.V. Complementary use of simulations and free energy models for CTAB/NaSal systems. Langmuir **2001**, *17*, 7160–7171.

Measuring Experimental Quantities Using Simple Fluorescence

M

W. A. Hollerman
Department of Physics, University of Louisiana at Lafayette, Lafayette, Louisiana, U.S.A.

S. W. Allison
S. M. Goedeke
M. R. Cates
Engineering Science and Technology Division, Oak Ridge National Laboratory (ORNL), Oak Ridge, Tennessee, U.S.A.

INTRODUCTION

Phosphors are materials doped with impurities that give off cold light when excited. This fluorescence is caused by the ions in the lattice structure emitting a photon to de-excite, vs. nonluminescent phonon processes. Light emitted from a phosphor is not caused by thermal effects, and as such, is considered cold. Fluorescent materials are used in a variety of applications including television screens, lighting, photocopy lamps, scintillators, as x-ray conversion screens, and sensor technology. The materials used for these sensors are typically inorganics doped with impurities that provide characteristic fluorescence and are commonly referred to as phosphors. Sensor technologies based on these materials use characteristics of the light emission to determine various parameters, such as temperature, impact/pressure, and radiation dose. In the following sections, the fundamental principles of each of the three sensor types will be discussed and applications described.

TEMPERATURE EFFECTS

Phosphor-based temperature sensors are adaptable to the needs for a wide variety of situations and are based on fluorescence material properties. The thermal dependence of fluorescence may be exploited to provide a noncontact, emissivity-independent optical alternative to other more conventional techniques, such as those employing pyrometry, thermocouples, or thermistors. In fact, there are certain situations where fluorescence-based thermometry is the only useful approach, such as in devices like turbines or engines.

Background

Phosphors have many emission characteristics that can be temperature dependent. These characteristics include emission intensity, absorption wavelength shift, emission wavelength shift, and emission lifetime.

The emission spectra of a phosphor can have multiple distinct wavelength peaks. As stated earlier, the emission level of these peaks can change with temperature. For instance, Fig. 1 shows the emission spectra of $La_2O_2S:Eu$.[1,2] Each emission line is characterized by a wavelength for which the intensity is a maximum. Its value may change slightly with temperature, and this is termed a line shift. An emission line has a finite width called the line width, which is often referred to as the spectral width at half the maximum line intensity. Line width and line shift, which change as a function of temperature, are generally small and not often used in fluorescence thermometry. This spectrum at 36°C shows that the relative intensities of the lines near 465 and 512 nm are less than those at 4°C, and thus are temperature dependent. In fact, typically certain lines in these phosphors get weaker, i.e., become less bright, as the temperature of the material is increased. The intensity ratio of distinct emission peaks can also be used as a temperature sensor and has been suggested as a method to make two-dimensional measurements.

An absorption spectrum is obtained by measuring the amount of light transmitted through a specimen as a function of wavelength, whereas an excitation spectrum is determined by monitoring the intensity of an emission line while the excitation wavelength is varied. The absorption spectra of many phosphors consist of a relatively broad band at the blue or ultraviolet end of the spectrum, along with sharper absorption features in the visible and near infrared. The sharper features are often due to atomic transitions of the dopant atom and, as noted above, exhibit some temperature sensitivity. As the broad absorption band results from direct interaction with the host, it may show more temperature dependence. This characteristic can be utilized in two ways. First, using a set excitation wavelength no measurements can be made below the temperature where the excitation spectrum matches

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120041071

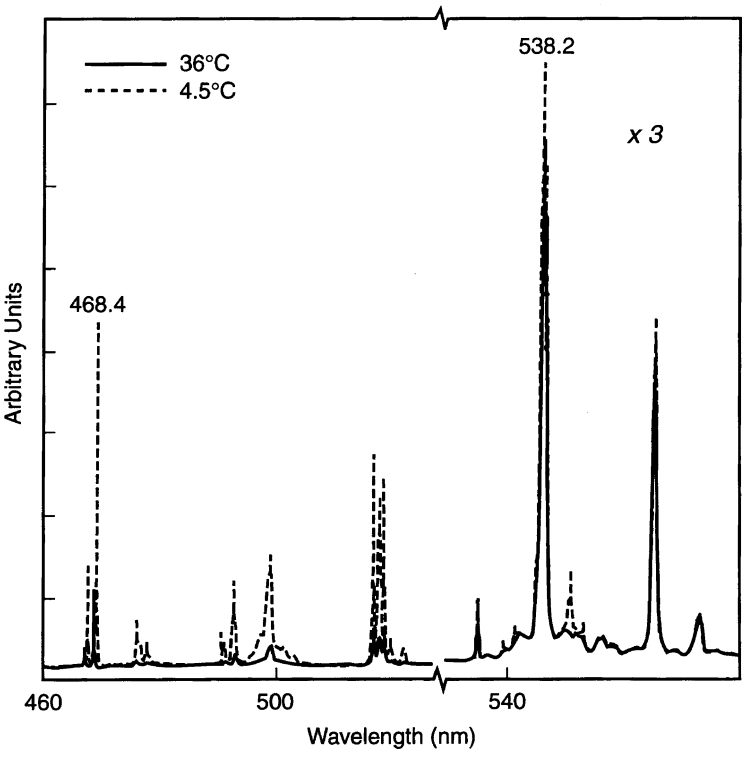


Fig. 1 Emission spectra for $La_2O_2S:Eu$ at two temperatures.

the excitation light. Second, this can be used as an indication of substrate heating. If the substrate is coated with a phosphor that is not excited at room temperature, as the light emission begins and gains intensity it becomes an indication of the substrate heating.

The emission lifetime is often used to determine surface temperature with the advantage that the technique is insensitive to blackbody background. This technique requires excitation by a pulsed source, the persistence of the resulting fluorescence can be observed providing that the length of the source pulse is much shorter than the persistence time of the phosphor's fluorescence. For certain phosphors, the prompt fluorescence decay time (τ) varies as a function of temperature and is defined by:

$$I = I_0 \exp\left\{-\frac{t}{\tau}\right\} \quad (1)$$

where I = fluorescence light intensity (arbitrary units), I_0 = initial fluorescence light intensity (arbitrary units), t = time from cessation of excitation source (sec), and τ = prompt fluorescence decay time (sec).

The time needed to reduce the light intensity to e^{-1} (36.8%) of its original value is defined as the prompt fluorescence decay time (τ). By measuring the decay time as a function of temperature, the phosphor can be calibrated.[1,2]

Calibration

A typical phosphor-based optical thermometry system will consist of at least the following components: 1) a source of excitation energy; 2) a means to deliver the energy to the target (typically beam-steering elements or a fiber-optic bundle if the energy is in the optical spectrum); 3) a fluorescing medium that is bonded to the target and illuminated by the incident flux; 4) an optical system to collect and transport the fluorescence that is subsequently generated; 5) a detector or an array of detectors to monitor the fluorescence signal; and 6) a data acquisition and analysis system that yields the target's temperature. Fig. 2 shows an elementary example of this arrangement. The particular system depicted here was designed to demonstrate the feasibility of making remote, noncontact measurements of the temperatures and strains in high-voltage devices in electrical substations in the Tennessee Valley authority's power grid.

By gradually raising the temperature of a phosphor sample and measuring the emission lifetime at selected temperatures, a calibration can be made. Once the lifetime values have been recorded over the expected temperature range, data are fit to produce a mathematical expression for the temperature as a function of the lifetime. Having calibrated the phosphor over the temperature range of interest, a small surface deposit of phosphor is excited with a pulsed laser and the fluorescent decay is measured (typically in less than 1 msec) to

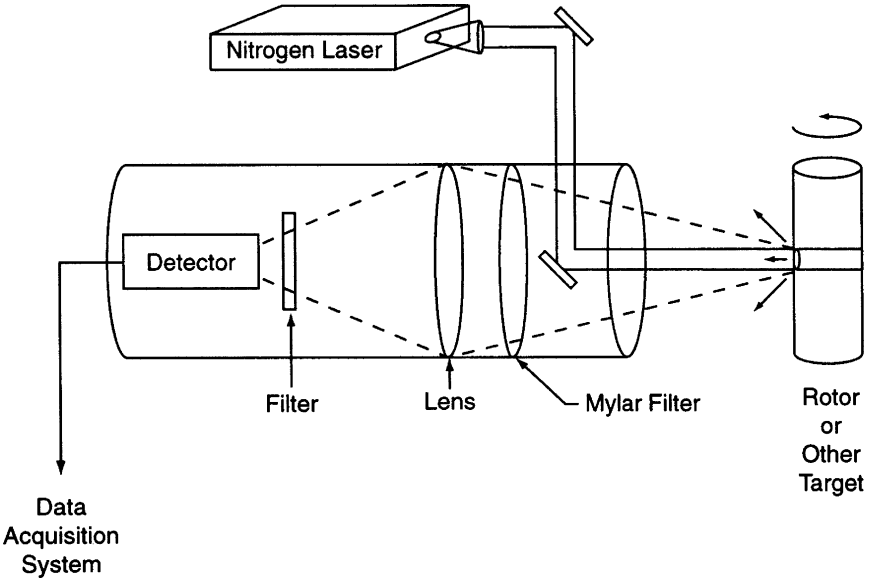


Fig. 2 Typical example of phosphor-based optical thermometry system.

calculate the temperature of the substrate. This technique has been evaluated for temperature ranging from cryogenic to upward of 2000 K. Fig. 3 shows the calibration curves for a variety of different phosphors.

Applications

Phosphor thermometry has been used in a variety of applications from aerospace to manufacturing. The technique becomes indispensable in situations that have large blackbody background, such as combustion, and when the surface is in motion, such as in jet turbines. The technique is also very useful when the surface emissivity varies with time. This technique can also be expanded to include measurements of heat flux by applying phosphor to two sides of an opaque insulator material.

Gas turbines

Previous successful applications led to efforts aimed at measuring the temperatures of surfaces and structures inside turbine engines, particularly those of the rotating blades and the stationary vanes. The history of that research is documented in a series of papers, as shown in Refs.[1,2] Several other groups have independently applied phosphor thermometry to turbomachinery, and in what follows we provide a synopsis of the work done in this field.

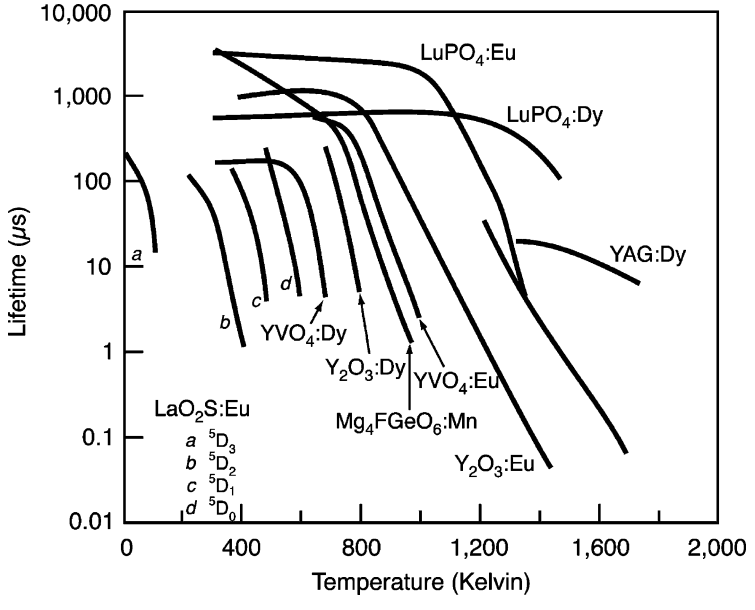


Fig. 3 Example calibration curves for a variety of phosphors. (From Refs.[1,2].)

One of the original intents of the Oak Ridge National Laboratory (ORNL) phosphor thermometry program was to provide a means of sensing the temperatures of objects within the combustion–flame environment. The first experimental tests of the technique were carried out inside a high-altitude simulation wind tunnel at the Arnold engineering development center (AEDC).[3] In them, the exhaust from the aft end of a Pratt & Whitney (PW) F100 engine impinged on a variable area extractor (VAE), a cone-shaped center body that aids in pressure recovery and the pumping process. Measurements were made of the VAE surface temperature, both with and without an afterburner plume intervening between the surface and detection optics. The VAE was water-cooled, consequently, its temperature was rather low, typically 150°C. The laser-induced fluorescence signals from the phosphor targets bonded on the VAE surface were successfully captured through the optical background created by the plume. This is shown in Fig. 4, which is a print made from a 16 mm film of the afterburner and the VAE. The white spot is the phosphor luminescence. The resulting data made it possible to track the temperature change of the VAE during the course of afterburner ignition transients.

Subsequent engine experiments were performed at a burner–rig installation. The usual arrangement involves mounting several blades on a mechanical carousel, which is then rotated at relatively low speeds through a jet flame. Phosphor thermometry tests by Tobin et al. in one such installation demonstrated the workability of the technique up to 1100°C.[4,5] Other studies have been carried out in spin pits, where the higher rotational speeds and the intermediate range temperatures encountered in operating turbine engines can be simulated without the exhaust flame present.

A number of in situ experiments have also been carried out, including measurements made on the first-stage stator vanes inside a PW 2037 engine on the vanes of a turbine disk running near full speed (14,600 rpm) in an advanced turbine–engine gas generator system, on the components of a PW joint technology developmental engine, and inside an experimental turbine engine at Virginia Polytechnic Institute and State University.[5–7] Collectively, these tests have demonstrated that phosphors based on materials such as yttrium oxide, yttrium vanadate, and yttrium aluminum garnet (YAG) can survive for extended periods even when bonded in place in the first stage of a turbine, where the blades are adjacent to the burner. Useful optical signals were obtained at temperatures well in excess of 1000°C. Separate calibration studies showed that the fluorescence decay parameters of some of the phosphors were still readily measurable at 1200°C.[8] The preferred design of the optical probe now consists of a two-fiber configuration with internal optical elements that are coated to minimize stray reflections, laser-induced damage to the input fiber, and the optical background due to blackbody radiation. On a related note, studies of the heat transfer through combustion engine components have also been undertaken with thermographic (temperature sensing) phosphors.[9]

Several others have also explored turbine engine and other aerospace applications of phosphor thermometry. For instance, Alaruri et al. exposed samples of Y_2O_3 : Eu to the combustion flows in a burner rig.[10] The gas velocities in the rig ranged from 245 to 407 m/sec at roughly atmospheric pressure. Data were taken over the range from 400°C to 1000°C with the overall accuracy of the measurements estimated to be ±3%. They noted that the phosphor's fluorescence

Fig. 4 Phosphor thermometry measured in an exhaust plume of jet engine afterburner.

was easily distinguishable from the luminescence because of the fuel pulses.

Phillips and Tilstra describe a thermographic phosphor temperature sensor based on the decay-time approach and designed using a 200 μm optical fiber to convey the fluorescence signals.[11] It has been tested over the range of 75–350°C, at altitude simulations of up to 21.3 km (70,000 ft), and under accelerative loadings equivalent to 20 g over the range of 5–2000 Hz. It can be used for measuring engine inlet total temperature, fuselage total temperature, and engine inlet immersion temperature, all without being subject to the effects of electromagnetic interference. Simons, McClean, and Stevens describe calibrations of additional phosphor materials of use for moderately high temperature measurements.[12]

Galvanneal steel

Galvanneal steel is an important material widely used in the production of automobiles. The accurate determination of temperature of galvanneal sheet during manufacture is a critical problem, which has recently been addressed with fluorescence thermometry.[13] In the galvanneal process, steel strip is dipped into a zinc bath. After exiting the bath, the sheet, with its molten-zinc coated surface, passes through an annealing furnace that promotes the diffusion of iron into the zinc coating. During this stage process, up to four metallurgical phases, or layers, are formed, each containing increasing amounts of zinc toward the surface. To control the final microstructure that determines the end product quality, precise control over the time–temperature relationship is crucial. The constantly changing emissivity of the strip surface and the strip motion rule out most optical pyrometers and thermocouples as practical measurement options. In addition, several parameters will vary during production, such as sheet thickness, sheet speed (which is up to 900 ft/min), and power to the induction furnaces. All of these affect the steel surface temperature and, so, must be controlled in real time to achieve a high-quality end product. Precision thermometry can provide for enhanced product quality, less product variability, and reduced process spoilage.

Clearly, the problem of surface emissivity is circumvented by fluorescence thermometry. A phosphor-deposition device applies the phosphor, under command from a control computer. As seen in Fig. 5, this deposition takes place several feet underneath the

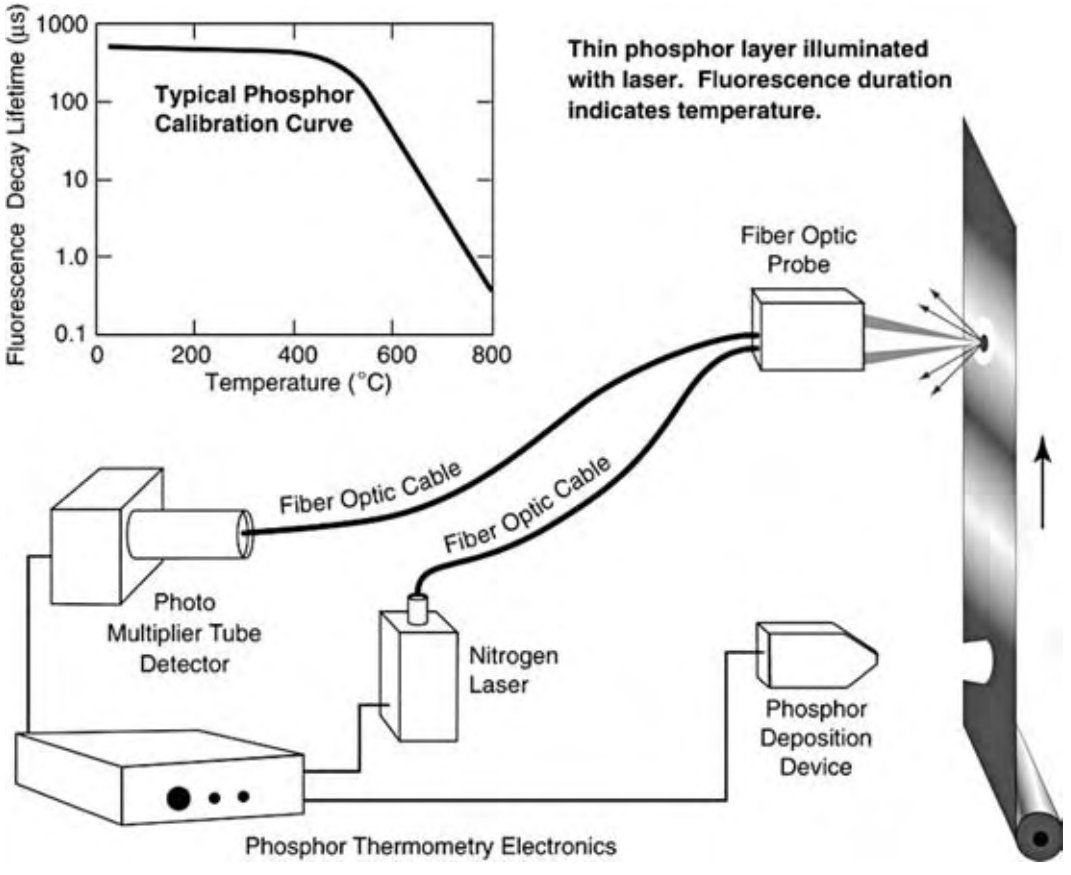


Fig. 5 Configuration of phosphor thermometry system for galvaneal steel.

temperature measurement equipment. This allows time for the thin phosphor layer to reach thermal equilibrium with the galvanneal surface. A laser with an optical fiber illuminates the phosphor. A given phosphor stripe is targeted several times by the laser so that the desired number of waveforms can be accumulated and averaged for good statistical sampling. Another optical fiber conveys the fluorescence photons to a remotely located photomultiplier tube (PMT) detector, which converts this optical signal to its electrical analog. The data analysis system is built around a personal computer (PC) using commercial software as well as hardware and software designed at ORNL. A custom interface board in the PC performs timing, control, and signal conditioning functions. The decay constant of the averaged signal is calculated and compared to an onboard calibration equation that determines the temperature reading. The most recent temperature is displayed on the computer screen. If the system is unable to obtain a stable measurement, no update occurs and the process is repeated. It should be pointed out that the introduction of microgram quantities of the phosphor material has been shown to have no detrimental impact on product quality.

An example of results, taken over a 100 min period, from a galvaneal line at National Steel is shown in Fig. 6. The temperature ranged between 460°C and 470°C. The apparatus can be used in other line-based industrial processes. For example, it was used at Bethlehem Steel for a much lower temperature. Different phosphor materials may be used depending on the temperature range of interest. This system received an R&D 100 Award for 1999.

In conclusion, the apparatus should be readily adaptable to other steel industry applications, such as slab heating, melt thermometry, roller thermometry, surfaces in reheat ovens, and galvalume processing. To extend the technology to these uses, issues like optical access, mechanical fixturing, phosphor adhesion, compatibility, and temperature range need to be addressed.

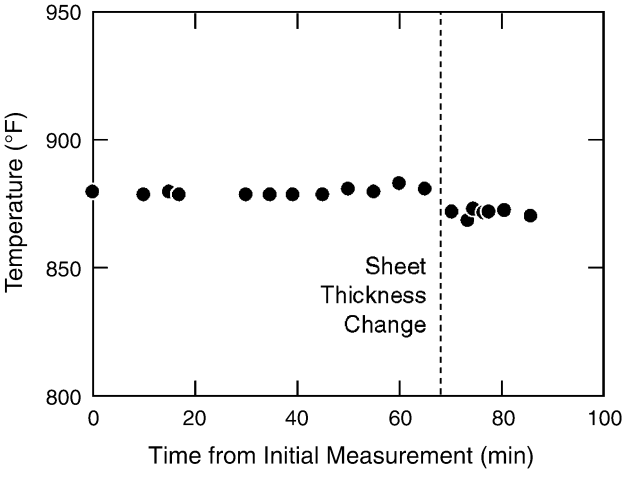


Fig. 6 Typical temperature data from galvaneal steel thermometry measurement.

Heat Flux

The determination of the heat flux through a surface is important in a variety of scientific and engineering applications. Noel, Turley, and Tobin, and Turley et al. pursued development of several thermal phosphor-based heat flux gauges.[14,15] A standard expression relates the heat flux, q, to the insulator thickness, d, thermal conductivity, k, and temperature difference, ΔT, across an insulating barrier:

$$q = \frac{k\Delta T}{d} \tag{2}$$

As originally conceived, a phosphor-based heat flux gauge would consist of a sandwich of two different types of phosphors separated by a UV-transparent insulator. The first realization of such a device incorporated $Gd_2O_2S:Tb$, in which the ratio of the D3 state at 415 nm to the D4 state at 490 nm served to provide the temperature of the top surface. The bottom surface was made of $La_2O_2S:Eu$, and its temperature was derived from the ratio of the emission of the D2 band at 511 nm to that of the D0 band at 614 nm. A demonstration arrangement, utilizing a hot-air stream, applied heat to one such gauge bonded to a water-cooled aluminum container. Levels of heat flux up to $40\,kW/m^2$ were measured in this way.

A second gauge consisted of an array of discrete circular and triangular spots of phosphor. This made it possible to obtain a two-dimensional determination of a heat flux through the surface of interest. Also, only one phosphor is required when, for instance, one-half of a semicircle of it is deposited on the top of the insulator with the matching half bonded onto the bottom. Fig. 7B is a close-up photograph of a fluorescing heat-flux gauge with matching triangular phosphor layers. Fig. 7A depicts the gauge attached to a turbine blade. Fig. 8 shows the experimental arrangement for this type of heat-flux gauge.

Baumann has also made heat flux measurements using thermographic phosphors.[16] Wind tunnel models made of three different materials (steel, Macor™ glass ceramic, and Norcoat™ 4000 silicone elastomer), all of which had thin coatings of $Y_2O_2S:Eu$ (0.15%) applied, were monitored during blow-down tests in a hypersonic flow facility. The response time of the measurement system was typically less than 200 msec, and the values of the heat flux measurements obtained with

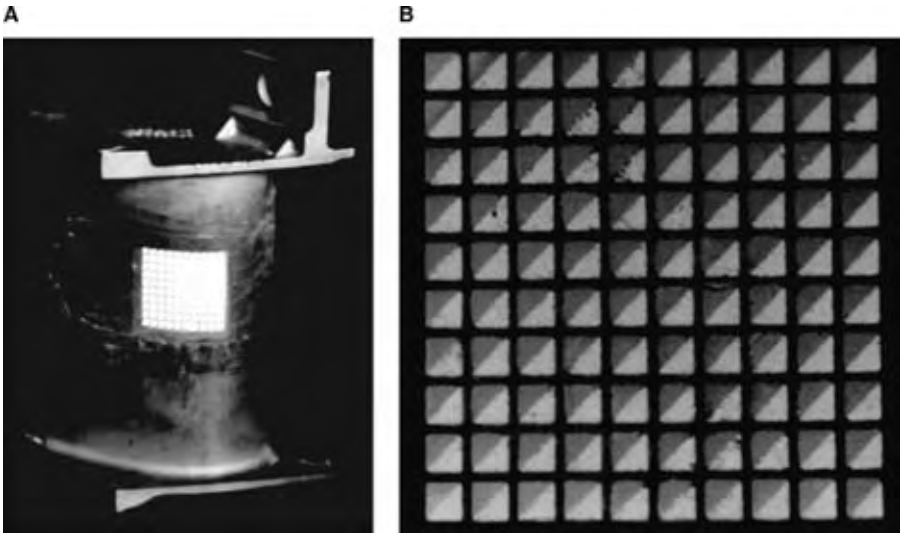
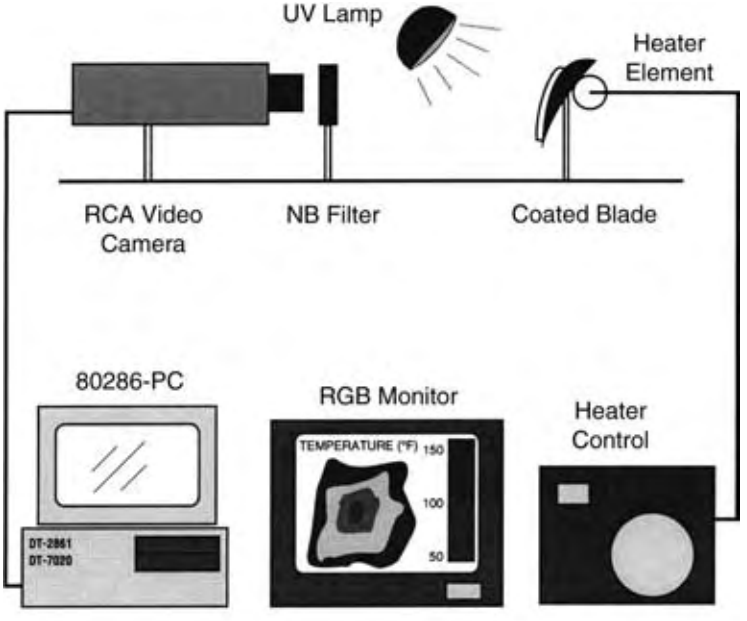


Fig. 7 (A) Phosphor-based heat flux gauge on turbine blade and (B) fluorescing gauge with matching triangular phosphor layers.

his technique compared well with the standard spot-gauge methods.

As a practical point, we note that the oxysulfide-based phosphors begin to degrade chemically in air near 450°C. However, other phosphors may be used for this application at higher temperatures. Another limiting factor encountered at high temperatures is the difficulty in finding a suitable insulator. To circumvent this, Noel has developed a theory and a resulting design that eliminates the need for an intervening insulator in the gauge.[17] As a related matter, we note that a detailed calculation of the heat flow through a thin



- RCA ISIT video camera.
- Narrow band filters: 410.5 and 490.5 nm
- Xenon (UV) light source
- 80286 computer
- DT-2861 frame grabber
- DE-7020 array processor

Fig. 8 Experimental arrangement phosphor-based heat flux gauge.

film heated by a laser has been carried out by Abraham and Halley.[18]

IMPACT EFFECTS

Triboluminescence (TL) is light produced while striking or rubbing two pieces of a material together. Sir Francis Bacon first studied it 400 yr ago and it is basically defined as the light from friction, as the term comes from the Greek tribein, meaning "to rub," and the Latin prefix lumin, meaning "light."[19,20]

Most crystals will emit light when fractured. Because most inorganic phosphors have a crystalline structure, they should emit TL. One phosphor, namely, zinc sulfide doped with manganese (ZnS : Mn), has often been noted for its TL. To develop an effective phosphor-based sensor, the relationship between TL intensity and impact velocity for ZnS : Mn or any other material must be quantified.

A specially designed drop tower for evaluating powder samples was constructed and is shown in Fig. 9.[21] It was made from 0.75 in. (19 mm) medium density fiberboard, 0.25 in. (6 mm) Plexiglas plate, and a common 1.25 in. (32 mm) piece of PVC pipe. The pipe has holes drilled in measured increments. These holes are used to place a pin that holds a 1.25 in. (32 mm) diameter steel ball (0.13 kg) bearing in place, and serves as the release mechanism.

The sample powder is placed on the Plexiglas plate as shown in Fig. 9B. The material is arranged so that it is aligned about the center of the tube. After several drop tests, the Plexiglas begins to show pits from the impact of the drop mass. These pits are used to align the sample so that the majority of the material is in the impact zone.

The ZnS : Mn powder can be widely displaced by the impact. After each test, the drop tube is removed, the drop mass is cleaned, and the powder is redistributed near the center of the target area. This redistribution is accomplished by placing a small piece of PVC in the tower base and moving the powder into an even layer. After each test, a small amount of the powder is removed when the ball bearing is cleaned. Apparently, this change in sample mass does not appear to affect the intensity or repeatability of the measurements.

To monitor TL intensity, a PMT was placed approximately 0.75 in. (9 mm) under the Plexiglas plate

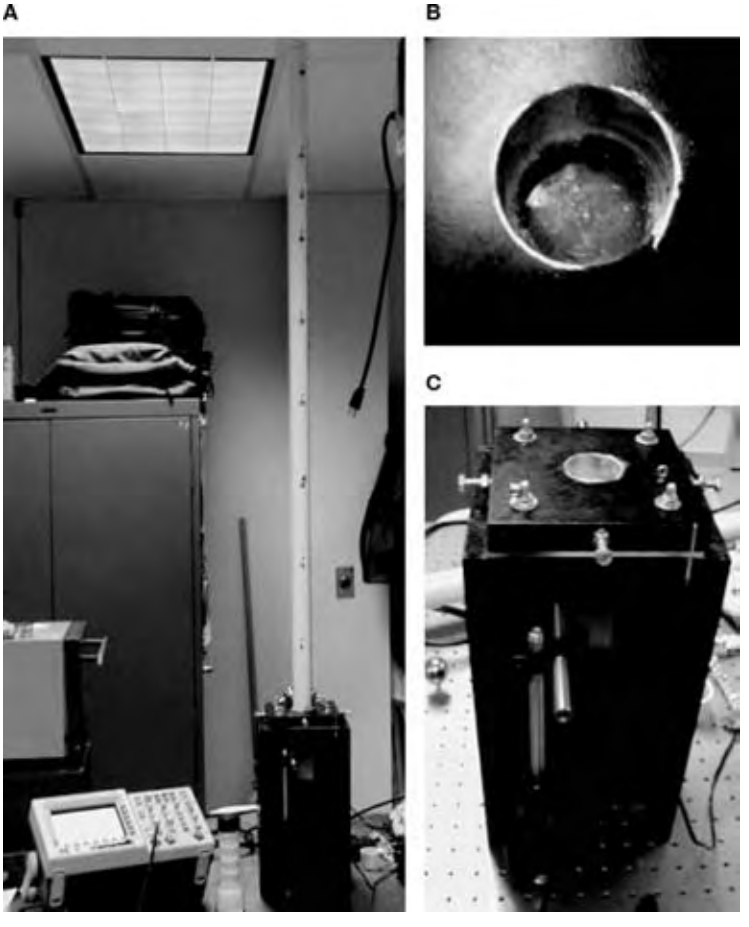


Fig. 9 Experimental arrangement for the drop tower: (A) drop tower with oscilloscope, (B) sample tray formed by tube support and Plexiglas bottom, and (C) drop tower base showing the mounting of the PMT.

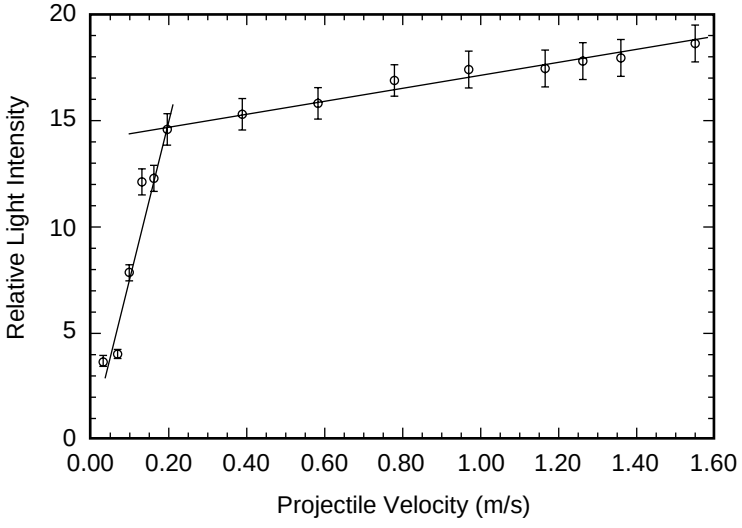


Fig. 10 Plot of the PMT output potential from ZnS : Mn and impact velocity.

as shown in Fig. 9C. To reduce effects because of stray light, a 589 nm filter is positioned in front of the PMT. Remember the maximum emission from ZnS : Mn is about 585 nm. The TDS 3052 oscilloscope captures the luminescence data in single sequence mode. The room light was turned off to further reduce background light noise in the PMT. The gain on the PMT was adjusted using a screw potentiometer until it was about half the total range.

Fig. 10 shows the variation in light emission characterized by the output potential from the PMT (V) vs. the drop velocity (m/sec). The impact velocity is defined based on Newton's equations for a free-falling object. These equations assume that air resistance is negligible and the ball does not make contact with the sides of the drop tube. The approximate measurement uncertainty is about ±5%, as shown by the error bars plotted in Fig. 10.[21]

The intensity of the triboluminesent response appears to be a function of impact velocity with two regions of interest as shown in Fig. 10. This is totally consistent with other measurements completed by the authors. The first region is in the threshold velocity ranges below 2 m/sec. The production of TL light appears to have a threshold at about 0.8 m/sec. Above the threshold, the projectile has sufficient velocity (or energy) to break ZnS : Mn crystals and to produce TL light. The TL luminescence increases rapidly until about 2 m/sec. The second region starts at about 2 m/sec and appears to be more like a saturation state, where the slope is much shallower and would indicate less sensitivity to impact velocity. Understanding these relationships will be very beneficial in the design of smart materials to detect damage caused by the impact of foreign objects. Once an impact is detected, the severity can be determined without the need for immediate inspection. These impact data could also be saved in a database and used to determine maintenance and inspection procedures. It was found that the reduction in TL material encountered in testing does not affect the repeatability of the measurements.

IONIZING RADIATION EFFECTS

For more than a century, phosphors that emit visible light when exposed to ionizing radiation have been used for a variety of scientific and engineering purposes. These materials exhibit a large fluorescence efficiency that makes them an effective accelerator beam detector or positioning system. Any phosphor used for this purpose will be exposed to large doses of beam radiation. This candidate material must be able to withstand such doses with a minimal reduction of fluorescence.

Half-Brightness Dose

The expression half-brightness dose ($N_{1/2}$) was coined as a consistent figure of merit to evaluate the effectiveness of a material in emitting fluorescence as a function of exposure. The $N_{1/2}$ was defined as the amount of exposure needed to reduce the fluorescence efficiency to one-half of its original value.

Birks and Black showed experimentally that the fluorescence efficiency of anthracene bombarded by alphas varies with total dose as

$$\frac{I}{I_0} = \frac{1}{1 + (N/N_{1/2})} \tag{3}$$

where I, I_0, N, and $N_{1/2}$ represent the fluorescence intensity, initial fluorescence intensity, total incident

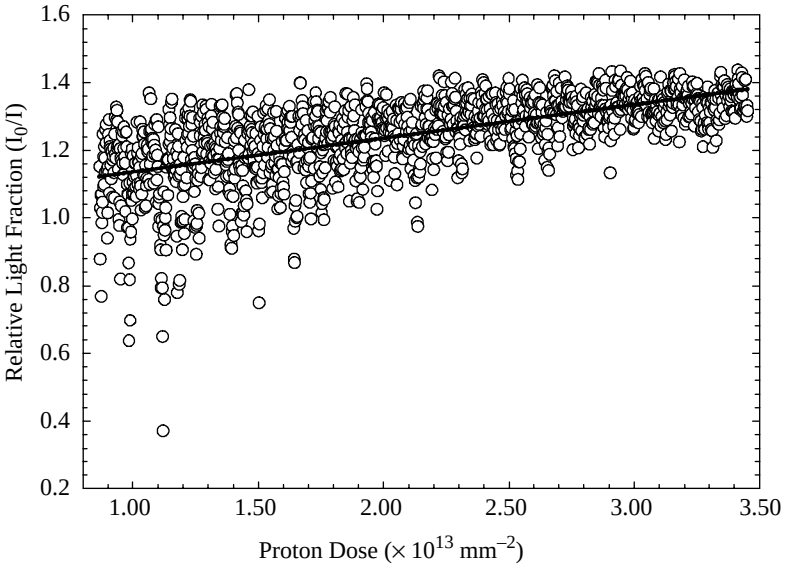


Fig. 11 Birks and Black plot for ZnS : Mn paint sample.

particle fluence, and the half-brightness dose, respectively.[22] The units of I and I_0 are related to the number of fluorescence photons interacting with the detector. When plotting the reciprocal of the light ratio (I_0/I) vs. proton dose, the resulting curve is linear with the slope equal to the inverse of $N_{1/2}$. The corresponding curve intercept was unity. Fig. 11 shows the relative light fraction as a function of proton dose for a ZnS : Mn paint sample. The line is determined by a linear, least-squares fit. Despite the low beam current, some charging/discharging of the sample was observed. The scattering of data points at low dose is probably due to this. As dose increases, the data are less scattered, suggesting that the protons were no longer building up on the surface. This could be the result of thinning of the layer due to beam sputtering or, more likely, a decrease in the proton cross section of the paint due to damage.

Schulman observed a similar effect to Eq. (3) when organic anthracene was exposed to γ-irradiation.[23] Black observed no efficiency degradation when the phosphor was exposed to 40 keV electrons, as they only cause ionization damage with no atomic displacements.[24] Northrop and Simpson found the fluorescence efficiency deteriorated in a similar fashion as was measured in previous measurements for organic phosphors.[25] Broser and Kallmann developed a similar relationship to Eq. (3) for inorganic phosphors irradiated using α-particles.[26] These results indicate that radiation produced quenching centers that compete with emission centers for absorbed energy.

3 MeV measurements

In the last decade, the authors have measured $N_{1/2}$ for several single-crystal and polycrystalline paint forms

Table 1 Selected 3 MeV proton $N_{1/2}$ data for several phosphor materials and forms

Phosphor				
Material	**Dopant**	**Sample crystal form**	**Cited reference**	$N_{1/2} \times (10^{14}\,\mathrm{mm}^{-2})$
YAG	Ce	PC paint	[27]	1.28 ± 0.21
		Single crystal		4.03 ± 0.65
Y_2O_2S	Eu	PC paint	[28]	0.60 ± 0.46
Gd_2O_2S	Pr	PC paint		0.16 ± 0.11
	Tb	PC paint		0.20 ± 0.13
Y_2SiO_5	Ce	Single crystal	[29]	0.28 ± 0.01
$Tb_3Ga_5O_{12}$	None	Single crystal	[30]	0.12 ± 0.01
ZnS	Mn	PC paint	[21]	9.02 ± 0.72

PC paint, PPMS paint with polycrystalline phosphor; single crystal, single slice of the given phosphor crystal.

Table 2 Half-brightness dose energy comparison for a selection of phosphors

Fluor		Ambient temperature $N_{1/2}(\times 10^{13}$p/mm$^3)$			150°C temperature $N_{1/2}(\times 10^{13}$p/mm$^2)$		
Material	**Dopant**	**3 Mev**	**45 Mev**	**Ratio (45/3)**	**3 Mev**	**45 Mev**	**Ratio (45/3)**
YAG : Ce	Ce	13 ± 11	28.0 ± 1.7	2.2 ± 1.9	1.4 ± 0.4	17.1 ± 6.6	12.2 ± 8.0
Y_2O_2S	Eu	6.0 ± 4.6	34.5 ± 29.5	5.7 ± 5.1	2.3 ± 0.4	9.24 ± 2.68	4.0 ± 1.3
Gd_2O_2S	Pr	1.6 ± 1.1	8.99 ± 2.43	5.6 ± 4.8	0.19 ± 0.06	15.9 ± 5.7	83.7 ± 40.0
	Tb	2.0 ± 1.3	14.7 ± 1.2	7.3 ± 4.8	1.0 ± 0.3	20.1 ± 1.2	20.1 ± 6.1

(From Ref.[28].)

for selected yttrium, gadolinium, and terbium phosphors prepared at ambient temperature. Table 1 shows the resulting $N_{1/2}$ values vary between $0.12 \times 10^{14}\,\text{mm}^{-2}$ and $4.03 \times 10^{14}\,\text{mm}^{-2}$.[21,27–30] These phosphors were excited using a 3 MeV proton beam from a small electrostatic accelerator.

The polycrystalline samples consisted of a paint containing approximately 70% poly(phenyl methyl) siloxane (PPMS) and 30% phosphor powder. This formulation was found to give the toughest and the most wear-resistant paint. The grain size for the powdered polycrystalline phosphors was measured to be less than 10 μm. Small phosphor crystal slices were mounted directly to the sample holder for measurement. Proton beam current was kept low to minimize electrical discharge during each irradiation sequence.

45 MeV measurements

The measured 45 MeV half-brightness dose for several fluor materials can be found in Table 2.[8] Note that the right column in both the ambient temperature and 150°C data sections of Table 1 are labeled ''Ratio (45/3).'' These columns contain entries for the half-brightness dose measured at 45°MeV divided by the equivalent value at 3 MeV. For example, the proton $N_{1/2}$ for YAG : Ce was measured to be 2.2 times larger at 45 MeV than it was at 3 MeV for the ambient temperature samples. The same ratio for the 150°C data indicates that YAG : Ce takes 12.2 times more proton irradiation to drop the light intensity to half of its original value. The half-brightness dose values for the ambient temperature samples were 2.2 to 7.3 times

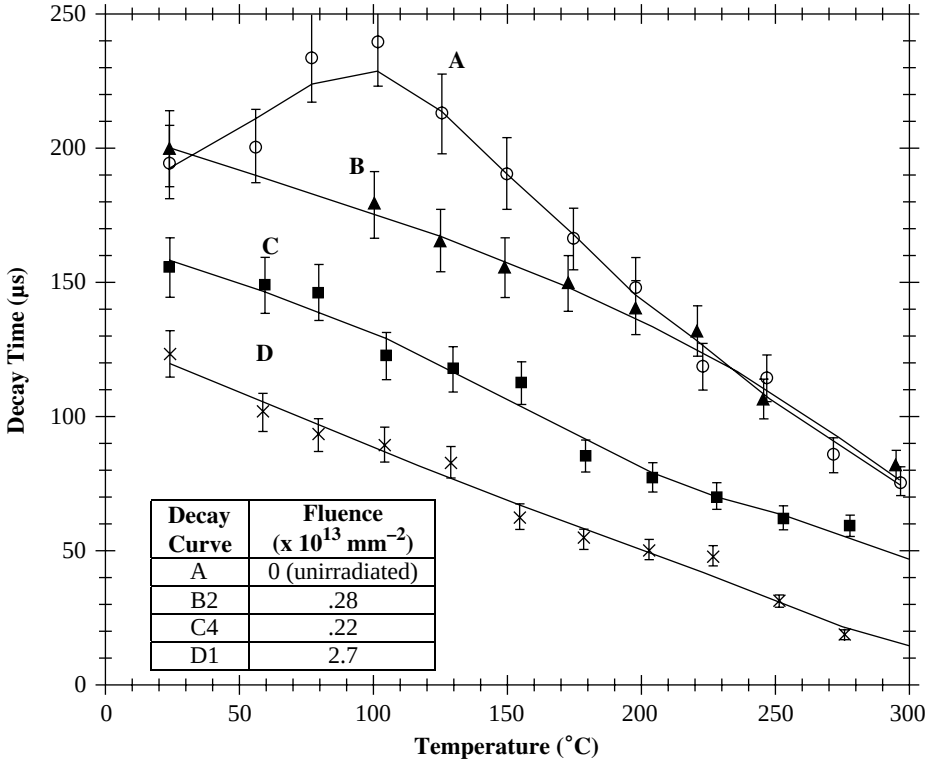


Fig. 12 Comparison of temperature-dependent decay times measured at the listed selection of 3 MeV proton fluences.

larger at 45 MeV than equivalent measurements at 3 MeV. The average ambient temperature ratio was approximately 5.2 for the set of four common fluors. The equivalent $N_{1/2}$ ratio at 150°C was 4.0–83.7 times larger at 45 MeV than the corresponding 3 MeV data. The average $N_{1/2}$ ratio at 150°C was 30 for the tested materials. The relationship between the half-brightness dose and the energy is unique at both ambient temperature and 150°C.

The proton interaction cross section is inversely proportional to energy, which causes the measured differences in half-brightness dose. Although fewer particles interact with the fluor at 45 MeV, each of the individual particles that do interact cause proportionally more damage to the scintillator than 3 MeV protons. A good analogy to this process is the use of a rifle bullet to knock down bowling pins. The probability that a rifle bullet (analogous to a 45 MeV proton) can hit a wooden pin is smaller than equivalent chance that a bowling ball (analogous to a 3 MeV proton) could do the same thing. Although the rifle bullet has a smaller interaction probability (cross section), the damage it causes when it hits a pin is much more extensive.

Fluorescence Decay Time

A plot of the temperature-dependent fluorescence decay time as a function of 3 MeV proton fluence is shown in Fig. 12.[31] The error bars shown in Fig. 12 for the measured decay times were calculated using standard uncertainty propagation techniques. Results indicate that a 3 MeV proton fluence as small as $2.28 \times 10^{13} \text{mm}^{-2}$ statistically reduces the ZnS : Mn decay time for measurement temperatures of less than 200°C. The unirradiated and $2.28 \times 10^{13} \text{mm}^{-2}$ decay time curves appear to be statistically identical after this temperature. A 3 MeV proton fluence of $2.28 \times 10^{13} \text{mm}^{-2}$ corresponds to one quarter of the half-brightness dose for ZnS : Mn.

Fig. 12 also shows that 3 MeV proton fluences greater than of $2.28 \times 10^{13} \text{mm}^{-2}$ uniformly reduce the ZnS : Mn decay time over all tested temperatures. It appears that reductions in temperature dependent decay time appear to be proportional to fluences larger than $2.28 \times 10^{13} \text{mm}^{-2}$. It is possible that radiation-induced dislocations in ZnS : Mn and the formation of sooty carbon in the polymer binder could account for these effects.

CONCLUSIONS

Fluorescent materials have been used in a wide variety of applications from lighting to x-ray detection. It has been well established that the light emission from many fluorescent materials are temperature dependent. This temperature dependence allows for the creation of noncontact and remote sensors that can be insensitive to blackbody backgrounds and radiofrequency interference. By observing the intensity of the light emitted after the material is struck, TL can be used to develop an impact sensor. Preliminary data indicate that there is a relationship between the light intensity and the impact energy for ZnS : Mn. Finally, it has been shown that the intensity of a fluorescent material is degraded by the bombardment of radiation. This degradation follows a linear pattern and has the potential to be used as an indicator of the total applied dose or fluence.

ACKNOWLEDGMENTS

We acknowledge the efforts of the many graduate and undergraduate students from the University of Virginia, University of Tennessee, Tennessee Technological University, University of Louisiana at Lafayette, and various other colleges and universities who have done much to establish the scientific and engineering foundations of phosphor measurement techniques over the past decade. We also thank the large number of colleagues at the Oak Ridge National Laboratory, the Los Alamos National Laboratory, Wright-Patterson Air Force Base, the University of Virginia, the University of Tennessee, EG&G Energy Measurements, Inc., and the International Bureau of Weights and Measures for their participation in this work. We give special thanks to Mr. W.A. (Bill) Stange of Wright Laboratory for his continued support of much of the effort described above for well over a decade. His influence has made possible the development of the method for important aerospace and turbine engine applications. Finally, we thank the workers at the Science and Engineering Library of the University of Virginia (Physics Branch Librarian Mr. James Shea in particular) and those at the Oak Ridge National Laboratory Library (Mr. Jon Arrowood in particular) for much valuable assistance in obtaining the various references cited here. The Oak Ridge National Laboratory is operated by UT-Battelle, LLC for the U.S. Department of Energy.

REFERENCES

1. Allison, S.W.; Gillies, G.T. Remote thermometry with thermographic phosphors instrumentation and applications. Rev. Sci. Instrum. **1997**, *68* (7), 2615–2650.

2. Allison, S.W.; Cates, M.R.; Goedeke, S.M.; Hollerman, W.A.; Womack, F.N.; Gillies, G.T. Remote thermometry with thermographic phosphors: instrumentation and applications. In *Handbook of Luminescence, Display Materials, and Devices, Inorganic Display Materials*; Nalwa, H.S., Rohwer, L.S., Eds.; American Scientific Publishers, 2003; Vol. 2, Chapter 4, 187–250.
3. Allison, S.W.; Cates, M.R.; Scudiere, M.B.; Bentley, H.T.; Borella, H.M.; Marshall, B.R. Remote thermometry in a combustion environment using the phosphor technique. Proc. SPIE **1987**, *788*, 90.
4. Tobin, K.W.; Allison, S.W.; Cates, M.R.; Capps, G.J.; Beshears, D.L.; Cyr, M.; Noel, B.W. High-temperature phosphor thermometry of rotating turbine blades. AIAA J. **1990**, *28*, 1485.
5. Tobin, K.W.; Cates, M.R.; Beshears, D.L.; Muhs, J.D.; Capps, G.J.; Smith, D.B.; Turley, W.D.; Borella, H.M.; O'Brian, W.F.; Roby, R.J.; Anderson, T.T. *Engine Testing of Thermographic Phosphors: Part 1. Pratt and Whitney Fixed-Blade Test and Part 2. Virginia Polytechnic Institute Turbine-Blade Test*; Technical Report No. ORNL/ATD-31; Martin Marietta Energy Systems, Inc., May 1990.
6. Noel, B.W.; Turley, W.D.; Lewis, W.; Tobin, K.W.; Beshears, D.L. Phosphor thermometry on turbine-engine blades and vanes. In *Temperature: Its Measurement and Control in Science and Industry*; Schooley, J.F., Ed.; American Institute of Physics: New York, 1992; Vol. 6, Part 2, 1249–1254.
7. Noel, B.W.; Turley, W.D.; Allison, S.W. Thermomographic phosphor temperature measurements: commerical and defense-related applications. In. In *Proceedings of the 40th International Instrumentation Symposium*; Instrument Society of America: Research Triangle Park, NC, 1994; 271–288.
8. Allison, S.W.; Cates, M.R.; Pogatshnik, G.J.; Bugos, A.R. *Solid-State Fluorescence Above 1000°C: Application to High-Temperature Laser Thermometry*; Technical Report No. ORNL/ATD-21; Martin Marietta Energy Systems, Inc., Jan 1990.
9. Domingo, N. *Heat Transfer Effectiveness of a Thermal Barrier Coating with Different Fuel Compositions*; Technical Report No. ORNL/TM-11816; Martin Marietta Energy Systems, Inc., Jun 1991.
10. Alaruri, S.; McFarland, D.; Brewington, A.; Thomas, M.; Sallee, N. Development of a fiber-optic probe for thermographic phosphor measurements in turbine engines. Opt. Lasers Eng. **1995**, *22*, 17.
11. Phillips, R.W.; Tilstra, S.D. Design of a fiber optic temperature sensor for aerospace applications. In *Temperature: Its Measurement and Control in Science and Industry*; Schooley, J.F., Ed.; American Institute of Physics: New York, 1992; Vol. 6, Part 2, 721–724.
12. Simons, A.J.; McClean, I.P.; Stevens, R. Phosphors for remote thermometry sensing in low temperature ranges. Electron. Lett. **1996**, *32* (3), 253.
13. Manges, W.W.; Allison, S.W.; Vehec, J.R. Galvanneal thermometry with a thermographic phosphor system. Iron Steel Engineer **1997**, *74* (12), 33–36.
14. Noel, B.W.; Turley, W.D.; Tobin, K.W. *Remote Temperature and Heat-Flux Measurement Using Thermographic Phosphors*; Los Alamos National Laboratory Technical Report No. LA-CP-91-0182, Apr 1991.
15. Turley, W.D.; Borella, H.M.; Noel, B.W.; Beasley, A.; Sartory, W.K.; Cates, M.R. *The Design and Characterization of a Prototype Optical Heat-Flux Gauge*; Los Alamos National Laboratory Technical Report No. LA-11408-MS, Jan 1989.
16. Baumann, P. Phosphor thermography technique in hypersonic blowdown wind tunnel: feasibility study with pinpoint measurement. Rech. Aérospace **1993**, *5*, 29.
17. Noel, B.W.; Beshears, D.L.; Borella, H.M.; Sartory, W.K.; Tobin, K.W.; Turley, W.D.; Williams, R.K. *A 2-D Imaging Heat-Flux Gauge*; Technical Report No. LA-12129-MS; Los Alamos National Laboratory, Jul 1991.
18. Abraham, E.; Halley, J.M. Some calculations of temperature profiles in thin-films with laser heating. Appl. Phys. A **1987**, *42*, 279.
19. Walton, A.J. Triboluminescence. Adv. Phys. **1977**, *26* (6), 887–948.
20. Sweeting, L. Triboluminescence with and without air. Chem. Mater. **2001**, *13* (3), 854–870.
21. Womack, F.N.; Bergeron, N.P.; Goedeke, S.M.; Hollerman, W.A.; Allison, S.W. Measurement of triboluminescence and proton half brightness dose for ZnS:Mn. IEEE Trans. Nucl. Sci. **2004**, *51* (4), 1737–1741.
22. Birks, J.B.; Black, F.A. Deterioration of anthracene under α particle bombardment. Proc. Phys. Soc. (Lond.) **1951**, *A64*, 874–875.
23. Schulman, J.H.; Etzel, H.W.; Allard, J.G. Application of luminescence changes in organic solids to dosimetry. J. Appl. Phys. **1957**, *28*, 1792.
24. Black, F.A. The decay in fluorescence efficiency of organic materials on irradiation by particle and photons. Philos. Mag. **1953**, *44*, 263.

25. Northrop, D.C.; Simpson, O. Electronic properties of aromatic hydrocarbon: II, fluorescence transfer in solid solutions. Proc. Phys. Soc. (Lond.) **1956**, *A234*, 815.
26. Broser, I.; Kallmann, H. On the elementary process of light excitation in phosphors by α particles, fast electrons, and gamma quanta. Z. Naturforsch. **1950**, *2A*, 439.
27. Hollerman, W.A.; Allison, S.W.; Goedeke, S.M.; Boudreaux, P.; Guidry, R.; Gates, E. Comparison of fluorescence properties for single crystal and polycrystalline YAG:Ce. IEEE Trans. Nucl. Sci. **2003**, *50* (4), 754–757.
28. Hollerman, W.A.; Glass, G.A.; Allison, S.W. Survey of recent research results for new fluor materials. Mater. Res. Soc. Symp. Proc. **1999**, *560*, 335–340.
29. Hollerman, W.A.; Fisher, J.H.; Holland, L.R.; Czirr, J.B. Spectroscopic analysis of proton induced fluorescence from yttrium orthosilicate. IEEE Trans. Nucl. Sci. **1993**, *40* (5), 1355–1358.
30. Hollerman, W.A.; Fisher, J.H.; Ila, D.; Jenkins, G.M.; Holland, L.R. Proton induced fluorescence properties of terbium gallium garnet. J. Mater. Res. **1995**, *10* (8), 1861–1863.
31. Hollerman, W.A.; Goedeke, S.M.; Bergeron, N.P.; Muntele, C.I.; Allison, S.W.; Ila, D. Effects of proton irradiation on triboluminescent materials such as ZnS:Mn. Nucl. Instrum. Methods Phys. Res. *in press*.

Membrane Reactors

Ivo F. J. Vankelecom
Centre for Surface Chemistry and Catalysis, Department of Interphase Chemistry, Faculty of Agricultural and Applied Biological Sciences, Katholieke Universiteit Leuven, Leuven, Belgium

INTRODUCTION

Although there is no commonly accepted definition of a membrane reactor (MR), the term is usually applied to operations where the unique abilities of membranes to organize, compartmentalize, and/or separate are exploited to perform a (bio)chemical conversion under conditions that are not feasible in the absence of a membrane. In every MR, the membrane separation and the (bio)catalytic conversion are thus combined in such a way that the synergies in the integrated setup entail enhanced processing and improved economics in terms of separation, selectivity, or yield, compared to a traditional configuration with reactor and separation separated in time and space. When the membrane itself carries the catalytic functions, it is mostly referred to as a ''reactive membrane.''

The combination of (bio)catalysis and membrane technology into an MR is truly multidisciplinary, involving aspects of material sciences, chemistry, biology, and biochemical engineering.

BASIC TERMINOLOGY OF MEMBRANE TECHNOLOGY

A membrane is defined as a semipermeable barrier between two phases that can either be liquid or gaseous.[1] Components permeate from one side of the membrane to the other under the influence of a driving force, being a gradient in concentration (or rather activity), (partial vapor) pressure, temperature, or electrical potential. Depending on this driving force, membrane processes can be further classified. Dialysis is the only concentration driven membrane process, while the pressure driven processes are subdivided into microfiltration (MF), ultrafiltration (UF), nanofiltration (NF), and reversed osmosis (RO), with the latter two often referred to together as hyperfiltration. The line between these nomenclatures is sometimes blurred, but generally pressure differences of 0.1–2 bar, 1–10 bar, 10–35 bar, and 15–100 bar refer to MF, UF, NF, and RO, respectively. The applied pressure increases as the membrane pore size decreases to end up as dense membranes for the finest separations. Pervaporation (PV) is the only partial pressure driven membrane process and the only one involving a phase transition at the membrane, through either a dense membrane or one with very small pores. Gas separations can be achieved using either dense or porous membranes, while electrodialysis is the only process involving an electrical potential, hence necessitating conductive membranes. Membrane distillation is nothing but a distillation taking place with a porous membrane separating the liquid and the vapor phases.

The membrane is characterized by a selectivity and a flux or permeance (flux divided by the thickness of the active layer of the membrane) or permeability (permeance divided by driving force). When used in an actual process, membranes are mounted in modules with a certain geometry: flat sheet membranes are turned into spiral wound or plate and frame modules, whereas cylindrical membranes with decreasing diameter are called tubular, capillary, or hollow fiber (Fig. 1). The use of inorganic membranes in the form of multihole elements or monoliths is becoming increasingly popular. The materials used to prepare membranes are very diverse, either organic or inorganic in origin, or a combination of both in hybrid membranes. The latter are also often referred to as composite, mixed matrix, or organo-mineral membranes. In general, inorganic membranes are chemically and thermally more stable, which can be beneficial during membrane cleaning procedures, but they are more difficult to prepare on a large scale and more expensive. They have a very high tensile modulus, but can be very brittle. In a module, the maximum possible membrane area per module volume is always lower than for polymeric membranes.

GENERAL ADVANTAGES AND DISADVANTAGES OF MRs

The general advantages of MRs clearly show how they fit in the scope of developing sustainable processes:[2,3]

- An integrated process involves a more compact process with lower investment costs and with substantial potential savings in processing costs, hence improving the overall economics.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030616

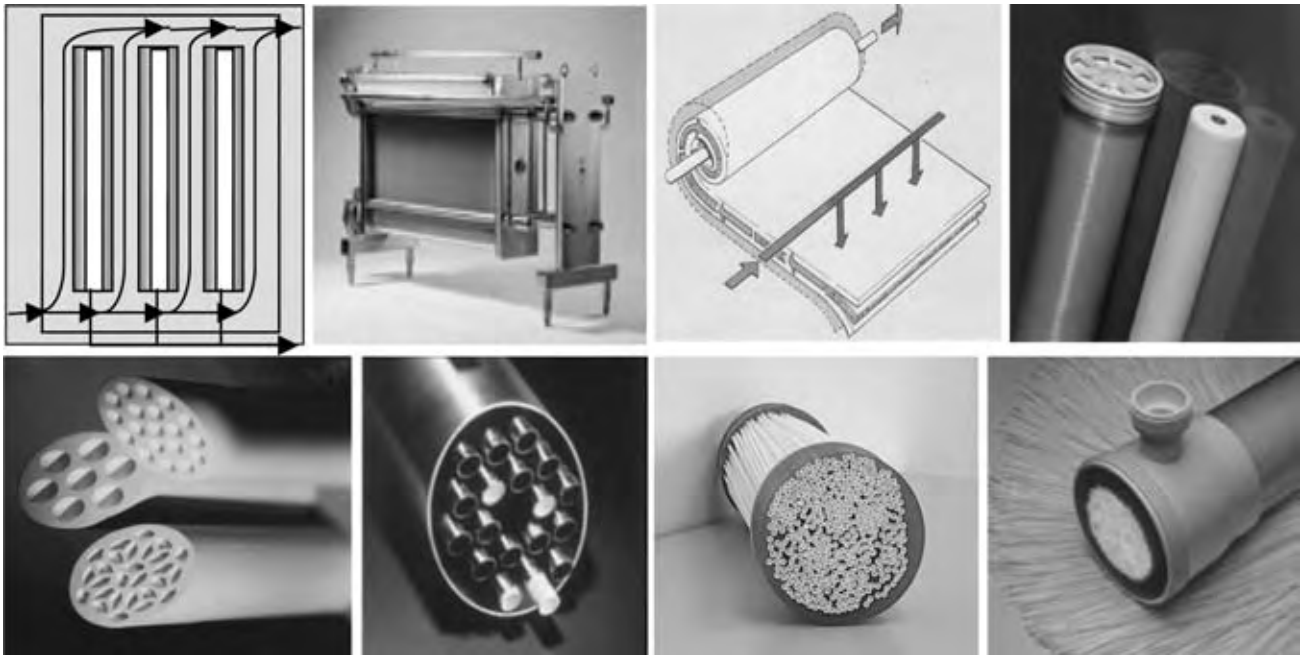


Fig. 1 Membrane modules (from left to right at top row: plate and frame module schematic, plate and frame module, spiral wound module schematic, spiral wound module; at bottom row: monoliths, tubular membranes, capillary membranes, hollow fibers. (Photos courtesy of http://www.niroinc.com/html/filtration/ftechsys.html.) *(View this art in color at www.dekker.com.)*

- Energy consumption can be saved, e.g., by less pumping.
- The removal of a product from a reaction mixture can shift the chemical equilibrium of the reaction and thus generate increased product yields or lead to similar yields, which are obtained at lower temperatures. The latter has the extra advantage of decreasing the extent of deleterious side reactions, such as coke formation.
- The continuous removal of product can decrease possible side reactions or product inhibition, thus increasing overall reaction rates.
- By selectively removing an intermediate product in consecutive reactions, enhanced selectivities for that intermediate can be realized.
- The downstream processing of the products can be substantially facilitated when removed from the reaction mixture by means of a membrane.
- The contact between two reactants can be mediated and controlled by a semipermeable membrane. In the field of selective oxidations, for instance, the oxygen feeding can thus be better controlled, or two hazardous reactants, like H_2 and O_2, can be reacted safely. Whereas the addition of gases to the reactor is often useful to give an extra stirring to the liquid mixture in the reactor, permeation through a dense, highly permeable membrane is sometimes preferred to integrate bubble-free aeration.
- When the membrane is placed as a contactor between two phases, solvents can be excluded and thus render the process environmentally and technically more attractive.
- Membrane separations often have the advantage of operating at much lower temperatures, especially when compared with, e.g., distillation, thus providing a solution for the limited thermal stability of either (bio)catalyst or product. Furthermore, membrane separations are not restricted to volatile components.
- Under specific conditions, the heat dissipated in an exothermic reaction can be used in an endothermic reaction taking place at the other side of the membrane, like, e.g., in hydrogenation/dehydrogenation.
- Coupling of sequential multistep reactions is possible.

In spite of these advantages, the combination and integration of two processes generally adds considerably to the technical complexity of the process, rendering modeling and prediction more difficult. Furthermore, MRs generally require more research, mainly related to optimization and design of a suitable reactor. Other possible drawbacks are the sealing of the reactor chambers, the manufacturing cost of the membrane and mainly the module, as well as the insufficient durability of the membranes and their fouling.

HIGH-TEMPERATURE CATALYTIC MRs

Most nonbiological MRs in the literature have been applied at high temperatures, with hydrogenations and dehydrogenations by far the most studied.[4–12] The reason was the existence of the remarkable Pd or Pd-alloy membranes showing an almost perfect hydrogen selectivity. In spite of all efforts, no industrial up-scaling has been realized yet, because of some important drawbacks. The cost of the membranes is very high, their commercial availability restricted, and their lifetime limited. Long-term operations may lead to surface inhomogeneity and competitive adsorption may lead to poisoning (e.g., with sulfides) or fouling (e.g., with carbon deposits). The intrinsically low permeability of such dense metal films is in fact the basic problem. To obtain sufficiently high permeabilities, their application is restricted to high temperatures. A reduction in membrane thickness is another possibility, but it creates an increased risk of defects. High-temperature applications entail difficulties when combining materials with different thermal expansion coefficients, leading to delamination of the membrane top layer from the support and bad sealing. Whereas Pd based membranes are strictly limited to H_2 permeations, highly selective membranes also exist for O_2 permeation based on Ag membranes or mixed inorganic oxides. They basically suffer from the same disadvantages as the Pd-membranes and even require temperatures above 500°C to show sufficient permeabilities. Nevertheless, because several important refinery and chemical feedstock reactions are excellent candidates to use MRs, many investigations have been performed already in this field of high-temperature applications.

Dehydrogenation Reactions

One of the most obvious choices for the application of MRs is the dehydrogenation of light alkanes to alkenes. Being endothermic, the reaction is performed at high temperatures. By removing selectively the hydrogen formed via a membrane, the equilibrium of these reactions can be driven to completion. At the same time, pure hydrogen is obtained at the permeate side, which is useful as a feed in other processes. In the pioneering studies by Gryaznov and coworkers, Pd or Pd-alloy dense membranes were used. Inorganic porous membranes have later been studied as alternatives, as well as zeolite and carbon molecular sieve membranes or proton-conducting solid state membranes.

Other dehydrogenations often investigated are the conversion of ethylbenzene to styrene, and cyclohexane to benzene, the latter being operated at such a low temperature that even polymer based membranes can become applicable. Also dehydrogenation of functionalized organics has been reported, as well as the nonoxidative coupling of methane to C_2-products. Fig. 2 illustrates how hydrogen removal through a selective Ag doped Pd-membrane increases the methylcyclohexane conversion to toluene and allows reaction at a lower temperature.

Hydrogenation Reactions

In liquid phase reactions, the use of hydrogen selective membranes as contactors between the gas and liquid phase to deliver the reactant at a controlled rate helps to avoid hot spots in the reactor or undesirable side reactions. Most interesting is the coupling of two reactions over one membrane with a dehydrogenation reaction at one side and a hydrogenation reaction at the other, with both hydrogen and energy flowing from one reactor to the other. Most importantly, hydrogen emerges at the other side of, e.g., a Pd-membrane in a highly reactive atomic form. Also, for this class of reactions, porous membranes and proton-conducting solid oxides have been tested. In the former case, the membrane can also function as a host for the catalyst, thus leading to a three-phase MR (gas–liquid–solid catalyst). Often studied reactions are the conversion of α-methyl styrene to cumene, cinnamaldehyde to

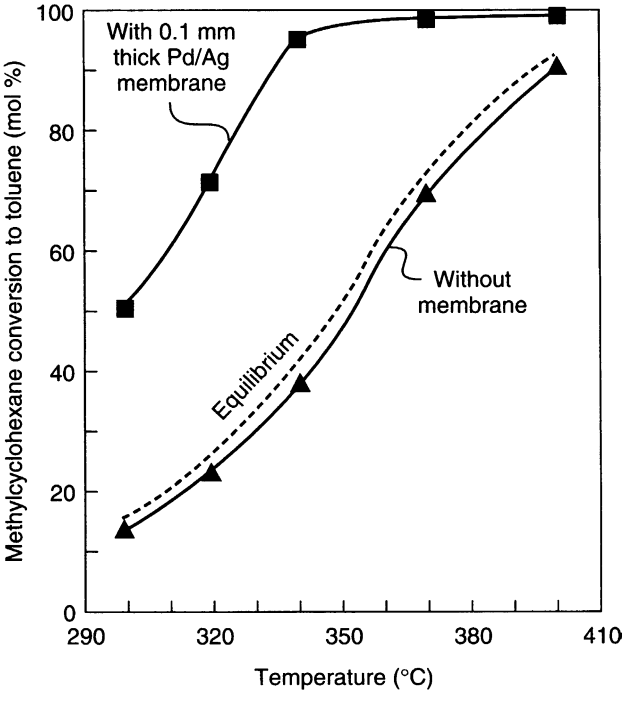


Fig. 2 Influence of temperature on the methylcyclohexane conversion to toluene in a membrane and a nonmembrane reactor. (From Ref.[8].)

hydrocinnamyl alcohol, and the hydrogenation of sunflower seed oil.

Selective Oxidations

The major advantage of using membranes in selective oxidations consists of the independent control of the concentration levels of each reactant in the reaction zone as oxygen and hydrocarbon (HC) can be fed in a controlled way from different sides of the membrane. In addition, the physical barrier between both reagents enhances safety aspects of the process. On the other hand, diminished reaction rates might appear because of the lowered oxygen partial pressure. The O_2/HC molar ratio might change along the length of the reactor owing to the depletion of a reagent during reaction. An elegant solution for the latter might be a ''chemical valve'' O_2 selective membrane showing an increased (HC + O_2) flux with decreasing O_2 concentration.[13]

Both dense and porous membranes, such as zeolite or alumina based membranes impregnated with catalytic species, have been applied. The dense membranes are mostly Ag or Ag-alloy membranes, solid oxides (Bi_2O_3, Mg stabilized zirconia, etc.), perovskites, and their oxygen deficient analogs brownmillerites. Brownmillerites and perovskites have increased thermal stability and better electronic conductivity. They refer to a general group of oxides with ABO_3 or $A_2B_2O_5$ structure, respectively.

Ion-conducting membranes could find application in two major industrial processes: the production of synthesis gas (syngas) by partial oxidation of methane and the oxidative coupling of methane to produce ethylene (Fig. 3). In the former, oxygen ions diffuse through the membrane to react with methane to form CO and H_2, a gas mixture that can be used for the formation of a variety of petrochemicals. In the latter, methane is converted to ethylene and hydrogen, permeating through the membrane to react with air oxygen to produce water. This second reaction produces the energy necessary to heat the process. Such reactions could allow the exploitation of natural gas fields for which utilization is currently not profitable because they are too far from the end users. The technical challenges on the membrane level are substantial: low-cost, defect-free anisotropic membranes operating at 800–1000°C without being poisoned or fouled.

Many other partial oxidations have been studied: methane conversion to methanol and formaldehyde,

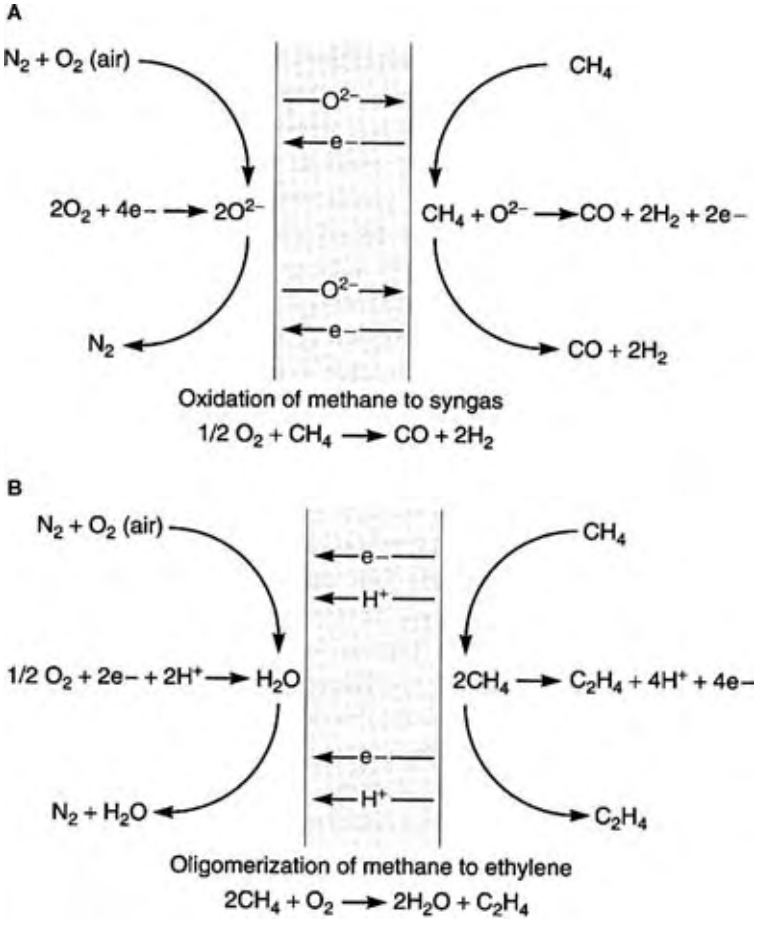


Fig. 3 A membrane reactor with ion-conducting membranes for the production of syngas (A) and ethylene (B). (From Ref.[8].)

benzyl alcohols to the corresponding aldehydes, ethane and propane to the respective oxides, butane to maleic anhydride, ethylene to acetaldehyde, propane to acrolein, oxidative dehydrogenation of alkanes and ethylbenzene, oxidative dimerization of propylene to benzene and hexadiene or of isobutene to branched octenes, oxidative dehydrogenation of methanol to CO_2 and H_2, and oxidative dehydrogenations of C_2–C_4 to their respective alkenes.

Other Reactions

Other major applications could lie in the water gas shift reaction, one of the oldest catalytic reactions to produce H_2, e.g., for the production of NH_3.

$$CO + H_2O = CO_2 + H_2$$

In this reaction, it would be extremely interesting to remove the CO_2 via selective membranes, but it is a challenge to find membranes stable at the harsh reaction conditions. Another important reaction with MR potential is the Fisher–Tropsch reaction to form higher alkanes:

$$CO + (1 - x)H_2 = CH_{2x} + H_2O$$

This reaction is exothermic, but for kinetic reasons it is most favorably performed at temperatures up to 700 K. The removal of water by, e.g., hydrophilic zeolite membranes would shift the reaction toward completion and reduce catalyst deactivation.

The decomposition of H_2S, NO, and NH_3; the breakdown of pollutants and volatile organics to CO_2 and H_2O; photocatalytic reactions involving TiO_2 as both catalyst and membrane; methanol reforming for the production of H_2, as well as direct H_2O decomposition for mobile fuel cell applications; CO_2 hydrogenation to produce methanol or CH_4 and H_2O; metathesis reactions of propene to ethylene and 2-butene; and the isomerization of *cis*-2-butene to *trans*-2-butene have all been reported.

LOW-TEMPERATURE CATALYTIC MRs[2]

In addition to the above-mentioned general advantages of combining reactions with membrane separations, the use of polymeric membranes in MRs entails some important extra possibilities, especially as reactive membranes (Fig. 4).[2] With a heterogeneous catalyst incorporated in a polymer matrix, a well-chosen polymeric environment can regulate the selective sorption of reagents and products with a beneficial effect on catalyst performance. In the case of an embedded homogeneous catalyst, the incorporation is a way to heterogenize and disperse the catalyst at the same time. Also, the coincorporation of additives in the membrane matrix can further increase the performance of such a heterogenized catalyst and substantially facilitate the downstream processing. Considering the high price of many homogeneous catalysts—especially the chiral transition metal complexes (TMCs)—the possibility of recycling them presents an important challenge, whereas good dispersions can generate higher stabilities and activities.

In selecting the most appropriate membrane, a much wider choice is available for polymeric than for either metallic or ceramic membranes. Polydimethylsiloxane (PDMS) is by far the most commonly used, thanks to its high permeability and stability. Polyvinylalcohol (PVA) and Nafion have also been described, especially for the more hydrophilic substrates. Operation of a polymer based MR at relatively

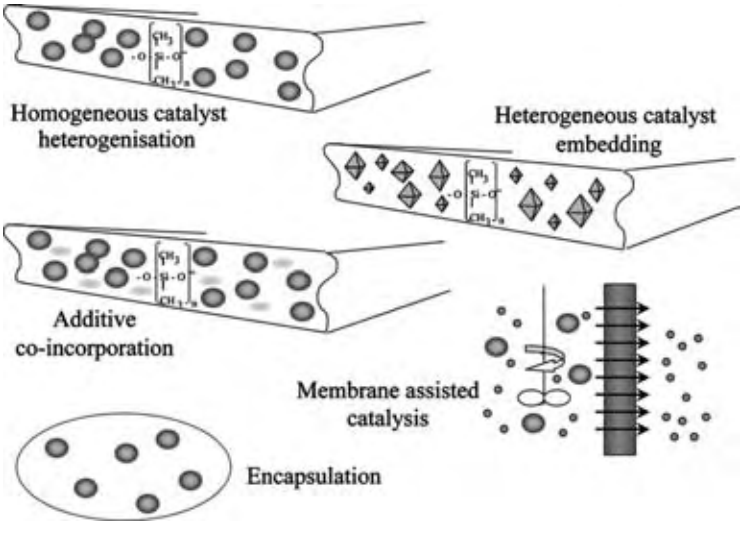


Fig. 4 Overview of the main applications of polymers in membrane reactors. *(View this art in color at www.dekker.com.)*

low temperatures is associated with less stringent demands for the materials in the module construction, thus leaving a wider choice to select the most optimal materials.

Encapsulation

Encapsulation refers to the confinement of a liquid solution within small capsules enclosed by a polymer or a surfactant. A potentially high interfacial area is thus created and the recovery of the catalyst is facilitated. The selective sorption through the membrane can further increase catalytic performances. Scaling-up is easy, but capsules should be as small as possible to prevent extra resistance to mass transfer in the non-agitated encapsulated volume. A problem associated with such capsules is the fact that there is no way to provide a fresh solution to the inner portion of the capsule or to continuously remove product from that phase. The capsules have to be either leached or broken at the end.

Phase transfer catalysts have been grafted onto the surface of porous capsules to facilitate product purification after reaction, and many types of immobilized cells, mycelia, enzymes, and catalysts have been encapsulated in polymers such as PDMS, PVA, or cellulose. In the specific case of PVA, they are named ''Lentikats,'' as commercialized by Genialab and used for nitrate and nitrite reduction and in the synthesis of fine chemicals. These beads show minimized diffusion limitations caused by the swelling of the polymeric environment under the reaction conditions. To avoid catalyst leaching, enlargement can be realized by linking them to, e.g., chitosan.

Incorporation of Heterogeneous Catalysts

Heterogeneous catalysts have been embedded in a membrane to improve their selectivity or activity thanks to a changed sorption in the catalyst, or to enable different reactor setups.[2] Activity can increase drastically when solvents become redundant with two immiscible reagent phases coming in contact via a membrane containing the embedded catalyst. It makes the system more environment friendly and facilitates product purification. The concept is claimed to be applicable to any catalytic reaction involving immiscible reagent phases, provided an appropriate choice is made of polymer type with respect to selective sorption and diffusion of reagents and products. Several oxidations, hydrogenations, isomerizations, dehydrations, esterifications, and epoxidations have been done this way already. The oxidation of propylene with hydrogen peroxide was described in a gas/liquid MR. Thanks to its transparent nature, PDMS has also been used in selective photo-oxidations. For photomineralization of organics in wastewaters or of volatile organics from air, acrylate based membranes were used containing a photocatalyst, mostly TiO_2, with a possible sensitizer. The systems were developed up to a preindustrial scale.

One of the most successful industrial applications of polymeric catalytic membranes is the Remedia Catalytic Filter System to destroy toxic gaseous dioxins and furans from stationary industrial combustion sources by converting them into water, CO_2, and HCl. The system consists of an expanded polytetrafluoroethylene (PTFE) microporous membrane, needle-punched into a scrim with a catalytically active PTFE felt. The catalyst is a V_2O_5 on a TiO_2 support. The microporous membrane captures the dust but allows gases to pass to the catalyst where they are converted at temperatures as high as 260°C.

Heterogenization of Homogeneous Catalysts

A whole set of TMCs, both chiral and achiral, have been already embedded for different hydrogenation, oxidation, and epoxidation reactions.[2] A specific and important problem occurring in liquid phase reactions with such membrane-occluded homogeneous catalysts is leaching of the complex and/or the coincorporated additive from the membrane. While the absence of strong interaction forces between the TMCs and the polymer is one of the strong points of this heterogenization with respect to catalyst reactivity, it renders them at the same time susceptible to leaching. This can be reduced or prevented by using solvents in which the complex does not dissolve, by establishing a chemical bond between the complex and the polymer, by enlarging the catalyst, or by selecting more appropriate reaction conditions, e.g., solvents that combine moderate membrane swelling with low solubility of the complex. On the other hand, adaptations on the level of the polymeric matrix can also restrict leaching: increasing the degree of cross-linking, decreasing the polymer chain length, or blending with other polymers to change the affinity.

TiO_2, Fe_2O_3, or Fe^{3+} have been embedded in Nafion, poly(ethylene) based polymers, and alginates for the photo-Fenton process to abate nonbiodegradable azo-dyes and in the selective oxidation of light alkanes. Similar or even higher activities were found owing to high polymer transparency, absence of aggregate formation, and a much better dispersion of the catalyst in the polymer.

On the basis of the metallic Pd-membranes, several authors incorporated Pd in polymeric membranes to create a higher active surface area with a smaller

amount of Pd. A variety of Pd-precursors and membrane polymers have been used to hydrogenate edible oils, isoprene, 1-octene, cyclopentadiene, propyne, propylene, ethylene, propadiene, and butadiene, and even in the controlled synthesis of hydrogen peroxide directly from molecular hydrogen and oxygen outside the explosion range.

Cation exchange membranes (CEMs) have been used as acid catalysts in the esterification of oleic acid with methanol. The reaction mixture was circulated between two CEMs. At the other side, these CEMs were in contact with a catalyzing acid, diluted in methanol. The proton thus appeared at the surface of the membrane, and was directed to the reaction mixture, while the anion associated with these protons remained at the other side preventing the proton from leaving the membrane.

Membrane-Assisted Catalysis

The key success factor of this concept is the choice of an appropriate membrane that allows the full retention of the catalyst and the complete passage of the products.[2] The aim is to achieve a high total turnover number, an easy recovery of the catalyst, or to operate in continuous or semibatch (also called ''repetitive batch'') modes. In contrast to a fixed bed reactor, membrane assisted catalysis has the advantage that fresh catalyst can be supplemented easily even in a continuous process. Because of the recovery of the catalyst, working at high catalyst concentrations becomes much more economical and competing noncatalyzed reactions can be controlled more easily. In a continuous mode especially, the educt concentration is kept low. Because the remaining free ligands are washed out during the initial phase of the membrane assisted process, the catalyst purification prior to the reaction can be shortened.

Mainly PV aided conversions have been studied and more in particular esterifications, a typical example of an equilibrium limited reaction with industrial relevance and well-known reaction mechanisms.[14,15] This hybrid process has already made it to several industrial applications. The thermodynamic equilibrium in such a reaction can be easily shifted and obtained in a shorter reaction time by removing one of the products. Pervaporation is especially interesting because it is not limited by relative volatility or azeotropes and energy consumption is generally low, because only the fraction that permeates undergoes the liquid/vapor phase change. It can also be operated at lower temperatures, which can better match the optimal conditions for reaction.

Because of the commercial availability of excellent permselective PVA, silica, or zeolitic membranes, it is especially the removal of water that became state-of-the-art technology over the past 15 yr. Modeling is important for these processes because, e.g., membrane permeability may change with temperature and composition of the mixture. A correct membrane surface area/volume ratio (A/V) is important because too low an A/V creates too slow a removal of the targeted compound, while too many reagents are removed with too high an A/V.

Apart from esterifications, dehydrations, condensations, oxidations, Diels–Alder alkylation, and hydrogenations also have been coupled to PV. In the production of alkyl coating resins at temperatures as high as 150–300°C, the use of microporous silica PV-membranes has been successfully investigated. A mixture of acids, acid anhydrides, and alcohols reacts to form a resin and water. In a kilogram-scale unit, energy savings turned out to be 40% and reactor efficiency increased by 30%.

When the membranes are too plasticized by the direct contact with the reaction mixture, vapor permeation can be a viable option. Gas separation has been described to enhance the conversion of the acid-catalyzed, equilibrium limited methyl-*tert*-butyl ether decomposition by the removal of methanol.

When rather small catalysts, like TMCs, have to be retained in pressure driven membrane processes, three approaches are possible: the catalyst can be enlarged to form dendrimers or hyperbranched polymers; it can be covalently bound to soluble polymers or insoluble supports; or membranes with a lower molecular weight cutoff (MWCO) can be selected, possibly at the expense of poor membrane flow.[16] A decrease in the MWCO of the membranes can be achieved by selecting tighter membranes or by changing the reaction conditions that influence this MWCO. Catalyst attachment or enlargement is often an elaborate task, creating structures that are nonuniform and uncontrolled. Mass transfer owing to hindered diffusion can lead to low catalytic activity. Another approach to enlarge TMCs is to add micelle forming amphiphiles to capture the catalyst. Emulsions can be easily retained by an appropriate UF membrane, often referred to as ''micellar enhanced UF.''

Whereas the pore-flow mechanism describes transport through porous UF, NF and RO membranes show a transient structure between porous and nonporous, with probably also sorption–diffusion as part of the transport mechanism. Nanofiltration is a relatively new membrane process with a nominal MWCO in the range of 200–1000 Da. Its application in water treatment has been growing rapidly, but the nonaqueous application is still an emerging field. All efforts to enlarge catalysts become superfluous when the membranes are capable of retaining off-the-shelf TMCs.

Aqueous NF played a crucial role in a continuous process that combines wet air oxidation, membrane separation, and biological treatment for the treatment of PEG containing wastewaters. In a first brief wet oxidation pretreatment, the polymers were degraded to a lower MW to increase the rate of the subsequent biodegradation of the permeate. The polymeric NF membrane proved resistant to the oxidative conditions created by excess oxidant still present in the oxidized process streams, even at 130°C. The integrated process thus showed a much higher treatment efficiency than any of the single optimized processes.

The idea of separating homogeneous TMCs with solvent-resistant NF(SRNF) membranes has only emerged over the last decade with the advent of commercial SRNF membranes. A nondestructive, energy efficient separation and concentration of reusable catalysts from reaction products can thus be realized. The SRNF-coupled catalysis can be run continuously, or alternatively in a semibatch operation mode to lower reactor occupancy or when reaction conditions are too harsh to be membrane compatible. Precipitating by-products could be simply removed from the reactor after filtration before refilling it.

For all catalysts (Jacobsen catalyst, Pd-BINAP, Wilkinson catalyst, Ru-BINAP, Rh-DUPHOS, phase-transfer catalysts) and solvents described, compatible commercial SRNF membranes were found that combined reasonable solvent fluxes with high catalyst rejection. However, some problems on the membrane level still need to be solved. Some membranes, marketed as SRNF membranes, swell to such an extent in certain solvents that they act like UF membranes with drastically increased MWCO. Fig. 5 shows the important consequences of tiny differences in rejections after many replacements of the reaction volume. Flux decline during operation can sometimes be observed owing to concentration polarization, increasing osmotic pressure, and membrane compaction. For some reactions, the membrane MWCO-curve is not steep enough or the cutoff is not low enough. Also, stability under certain ''aggressive conditions'' of solvent, temperature, pressure, and reagents can still be problematic. The very recent introduction of high-throughput membrane testing and the directed synthesis of membranes for a given purpose involving combinatorial synthesis techniques will undoubtedly further broaden the scope of available SRNF membranes and thus of catalytic applications.[17]

Also, dialysis was reported to remove TMCs from reaction mixtures. No mechanical pressure is needed here, clearly facilitating the mechanical and safety requirements. The first proof of concept was given with Ru-BINAP confined in a PDMS-''envelop'' simply submerged in the reactor. In a catalytic transfer hydrogenation, the PDMS-membrane retained the catalyst

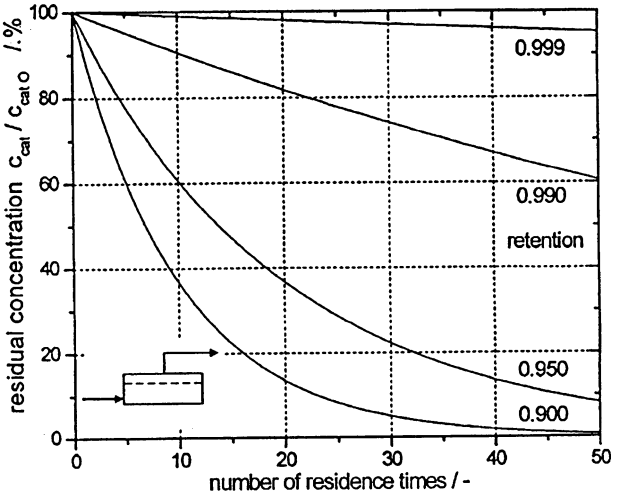


Fig. 5 Residual catalyst concentration as a function of the number of residence times. (From Brinkmann, N., Giebel, D., Lohmer, G.; Reetz, M.T.; Kragl, U. Allylic substitution with dendritic palladium catalysts in a continuously operating membrane reactor. J. Catal. **1999**, *183*, 163.)

insufficiently from the IPA-solution. Instead of enlarging the catalyst, the membrane composition and the composition of the solvent mixture were changed to improve the catalyst retention. Just like in SRNF, the use of solvents that induce strong swelling of the membranes makes laborious catalyst enlargement compulsory, as reported in SN_2 substitution reactions and Kharash addition.

MEMBRANE BIOREACTORS

Many industrial processes exist already in the production of biochemicals or for environmental applications.[18,19] One of the main reasons for this success is that the temperatures in these processes seldom exceed 60°C, thus enabling the use of inexpensive commercial polymeric membranes. Microfiltration and UF units are already proven and profitable technologies and well accepted by industry. Moreover, the fine chemical and biochemical MBR production leads to high value-added products, making the economics more favorable. The biocatalysts can be either enzymes or whole cells. The products range from food, liquid fuels (e.g., ethanol), and plant metabolites to more complex fine chemicals, such as pharmaceutical products or fragrances. Biological reactions typically generate a complex product mixture with some of these products often inhibitory or even toxic for the biocatalyst. Their continuous removal via a membrane separation process thus prolongs the biocatalyst's lifetime and increases the product turnover rate. But most often, the membrane's major role is in separating in situ the valuable

products from the unreacted raw materials and the often expensive biocatalysts, thus allowing continuous operation. Additionally, they increase the cellular concentration in the reactor, hence raising productivity and reactor throughput. The membrane offers the possibility of independently adjusting the residence times in the reactor of products, reactants, and biocatalysts to improve operational flexibility and provide effective process control, e.g., by controlled delivery of one of the reagents like in the bubble-free aeration. The membrane can act as the host for the biocatalyst when immobilized in the membrane pores.

In spite of the clear advantages, membrane-coupled processes often turn out to be still more expensive, because of the costs of membrane and all additional hardware associated with a membrane operation. Membrane separations tend to become more favorable for processes where the selectivity is more important than the conversion, because it replaces other purification steps that might lower such selectivities, like in the production of the chiral diltiazem intermediate. Membrane fouling, mass transfer limitations, biocatalyst activity loss, and biocatalyst denaturation are other potential disadvantages related to it.

Enzyme Membrane Bioreactors

The gap between enzyme catalysis and nonbiological catalysis will probably narrow in the future as adapted enzymes can work in organic solvents and catalysts that mimic enzymes are showing up. Biocatalysts are becoming commercially available to a greater extent, but often remain expensive. Enzymatic reactions are generally faster, but enzymes might need regular replacement because of denaturation. Recent efforts to increase the stability of enzymes were directed to either genetic manipulation or immobilization on a soluble polymer, on the membrane surface, or in the membrane pores. The amount of immobilized enzyme can be critical as blocking of the active site, multiple point attachment, or protein denaturation can occur above a certain loading. Furthermore, immobilization is often elaborate and may lead to only low immobilization yields and diffusional constraints in the supports.[8,20–22] Enzyme based MBRs (eMBRs) find application in the enzymatic hydrolysis of macromolecules, such as proteins, polysaccharides (cellulose, starch, etc.) or oligosaccharides (maltose, saccharose, lactose, etc.), and in the recycling of cofactors, such as NADH. An important application is the fruit juice clarification where the applied pectinases and cellulases are recovered via membrane filtration. Lipid hydrolyses or transesterifications catalyzed by lipases or phospholipases are complex micellar solutions of reactants and products. Enzyme based MBRs can be used here to selectively produce di- or monoglycerides by maintaining the system working at steady state at the desired conversion. Examples of polysaccharide production are also known, e.g., the synthesis of inulin from sucrose using a membrane immobilized fructosyltransferase.

A number of chiral drugs are currently being produced by means of eMBRs at scales up to 75 tons/yr. An optically active diltiazem intermediate is produced using an immobilized lipase. Also, for the production of naproxen from its ethyl ester, a lipase is immobilized in the wall of a membrane. The racemic ester is dissolved in methyl isobutyl ketone flowing through the shell side of the fiber, which is in contact with a buffered aqueous solution at the lumen side. The hydrolyzed substrate is transferred to the aqueous phase. The remaining enantiomer in the shell side is then racemized in a separate reactor and recirculated to the eMBR to achieve an almost 100% conversion.

An attractive MR concept is called ''cascade MR.'' Two reactions of which the optimal reaction conditions are significantly different can be carried out in series. Two different lyases have thus been used for the production of l-alanine from fumaric acid in two consecutive reactions. A UF membrane had to be used to retain both enzymes in their appropriate reactors. Each reactor could then be operated at the most optimal conditions of pH and temperature. In another example, GDP-mannose was prepared by linking two reactors through a UF membrane. The overall enzymatic consumption in the two-step eMR cascade was only 15% of that in the batch reactor and 45% of that in a single eMR.

Whole-Cell Membrane Bioreactors

Most of the work in this field involving plant or mammalian cells, bacteria, or yeast has been devoted to environmental applications treating wastewaters and to fermentation processes producing biochemicals.

With the earliest developments already in the 1960s, MBRs are state of the art in environmental applications and have been applied already worldwide in water treatment. They offer an important advantage compared to traditional physicochemical approaches: the complex organic contaminants present in wastewater streams are transformed into simple, harmless gaseous or water-soluble compounds, together with residual sludge, instead of just being transferred into a different medium. Also, they generally require a smaller volume and footprint than the more conventional biological treatment systems with biomass removal, e.g., via sedimentation. The membrane also offers an elegant way to control the process effectively via independent adjustment of wastewater residence

time in the bioreactor and the product withdrawal rate through the membrane. Moreover, MBRs generally need less pretreatment steps and the membranes offer an effective barrier for microbes and other particles.

Wastewater treatment can be achieved via either aerobic or anaerobic conditions. As shown in Fig. 6, the two processes differ in the rate of sludge production, hydraulic residence time, biomass concentration, and the type of pollutants remaining to be treated via other processes. Lower organics removal rates, the potential for odors, and a higher start-up time are among the main characteristics of the anaerobic processes. On the other hand, they require less energy to operate. The majority of the industrial wastewater treatment applications are aerobic. Still, the type of micro-organisms used and the process parameters vary depending on the wastewater's origin (domestic or industrial). Air bubbling is applied to supply oxygen for the biomass growth and to create turbulent flow conditions near the membrane surface to minimize fouling, the main problem of MBRs. The membranes are mostly immersed in the reactor to save on module costs and to combine aeration with fouling reduction (Fig. 7). In some cases, however, the membrane can be present in a module on a sidestream. The possibility of increasing the tangential flow may alleviate the fouling problem but increases the costs and can cause cell lysis via the high mechanical shear.

Because common MBRs are only able to remove the water-soluble compounds, poorly water-soluble compounds can accumulate to reach toxic levels. Organophilic membranes can be used then to extract these compounds, such as in the extractive MBR where the organics are removed via a hydrophobic membrane to an aqueous part where they are biotransformed. High concentrations of accumulating inorganics, acids,

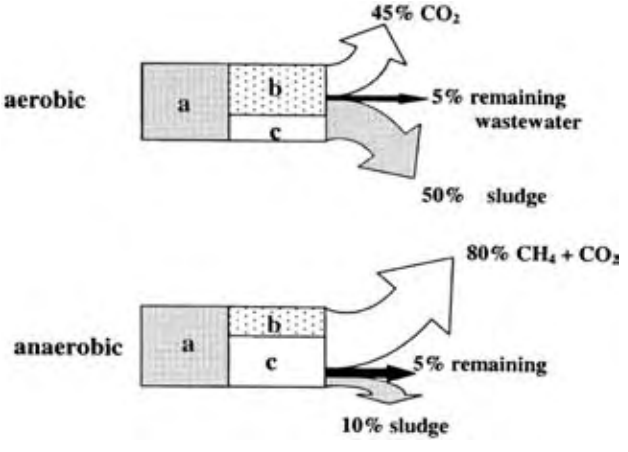


Fig. 6 Differences between aerobic and anaerobic biological wastewater treatment. a = wastewater; b = biomass content; c = hydraulic retention time. (From Ref.[9].)

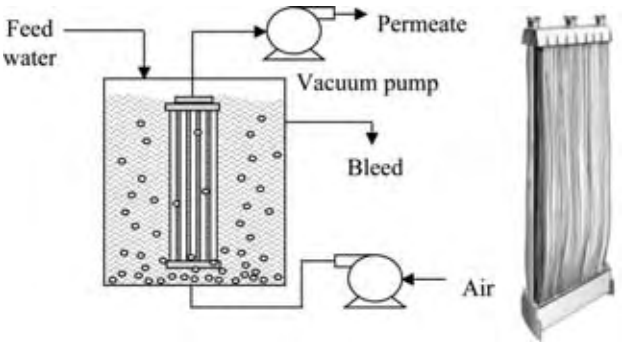


Fig. 7 Submerged membrane systems for wastewater treatment. *(View this art in color at www.dekker.com.)*

or biorefractory pollutants can prevent biological degradation.

Membrane bioreactors have finally also been applied to the treatment of gaseous phase wastes with a membrane used as a contactor, e.g., for the removal of organics (e.g., propene, chlorinated solvents, etc.) or inorganics (SO_2, NO_x, etc.). A biomass film is mostly grown on or within the membrane, or the bacteria can be homogeneously dispersed in the receiving liquid.

One of the most important processes in the production of biochemicals is the 40,000 tons/yr lactic acid production involving the *Lactobacillus* oxidation of lactose. The MBR productivity increased eightfold compared to a conventional batch reactor with a 19-fold increased biomass concentration. Even a 30-fold increased production of ethanol was found upon coupling the *Saccharomyces cerevisiae* fermentation to a membrane separation. Other successful industrial applications involve the pathogen-free production of growth hormones, the synthesis of homochiral cyanohydrins, the production of l-aspartic acid, phenylacetylcarbinol, vitamin B12, and the biotransformation of acrylonitrile to acrylamide.

In two-phase transformations, whole-cell MBRs use the membrane to separate the two different phases, thus avoiding phase mixing and emulsification. As a contactor, the membranes can also reduce mass transfer problems.

Perfusion MBRs have been introduced for the production of monoclonal antibodies. The mammalian cells that synthesize them are grown in the extracapillary space between the fibers in the module. Nutrients are supplied through the fibers, which also extract the metabolites continuously. The high cell concentrations between the fibers initiate high antibody harvests. These MBRs are also being investigated as an alternative concept for bioartificial organs such as liver and pancreas.

For the ABE (acetone–butanol–ethanol) fermentation, PV-coupled MRs have been described using a

variety of fermentation broths. The products formed inhibit the biocatalysis, making product removal from the fermentation broth via PV-coupled MRs an interesting option.

Enrichment of the impermeable components, i.e., salts and organic acids (mainly acetic and lactic acid) is an important challenge. The problem can be avoided by using a continuous ''bleed'' from the fermentor via an MF membrane. Another key challenge is membrane fouling. It can be minimized by selecting the right membrane material and conditions of operation, or by introducing an MF step before the PV unit to remove colloidal and macromolecular components. When the PV is applied only to the filtered cell-free broth, the PV efficiency can be increased by raising the temperature.

The technical challenges for MBRs are the high mechanical stress the biocatalysts experience at the elevated circulation rates required to maintain a good trans-membrane flux and biocatalyst deactivation because of limited mass transfer of nutrients or metabolites when immobilized in small pores or at high cell density. Another major problem is biofouling, caused by adsorption on the membrane of metabolites of coagulated proteins from lysed cells, and pore plugging owing to normal cell growth. Membrane cleaning can take place either by periodically shutting down the whole operation or continuously by employing two separate parallel units, one of which is in operation while cleaning the other. Membrane cleaning is obviously more difficult when whole-cell biocatalysts are immobilized in the small fibers. Immobilization of the biocatalysts in porous beads can facilitate this. Necessitating regular cleaning, the resistance to cleaning or sterilization cycles is a major issue in the selection of the membrane materials for MBRs.

CONCLUSIONS

In spite of many opportunities and lots of promises, multiyear large-scale industrial applications of MRs are still absent at high temperatures. The technical challenges to fit membranes leak-tight in high-temperature modules are substantial, and while membranes with extremely high hydrogen and oxygen selectivities are available already, their stability should be increased, and either their permeability should be increased or their operating temperature decreased.

Low and moderate temperature applications are more straightforward and small- to medium-scale niche applications are running already to produce fine chemicals. This is mainly the case in some specialized biotech applications where high-added-value products make the membrane process investments worthwhile. The largest current market is, however, in the treatment of low-added-value water, owing to the unique synergy between the biological treatment and the membrane process. The less severe requirements in terms of membrane housing and sealing, together with the possibility of using inexpensive, commercial polymeric membranes, have played a crucial role here, together with the fact that the MF and UF units that are coupled with the bioreactors are already proven and profitable technologies, well accepted by industry.

Polymer membranes have clearly matured much more than the metallic or ceramic membranes. For the former, fouling, long-term durability, module cost, and solvent resistance leave most room for further improvement. For the high-temperature applications with inorganic membranes, the membrane cost, the limited commercial availability, the low permeability, the module construction, and the operating costs probably remain the main factors hindering significant progress in, for instance, the petrochemical industry, where classical processes are mature and highly optimized, and where low returns on investments do not leave much room for implementing new technologies.

ACKNOWLEDGMENTS

An I.A.P.-P.A.I. grant on supramolecular catalysis sponsored by the Belgian Federal Government and a G.O.A. grant from the Flemish Government are gratefully acknowledged.

REFERENCES

1. Mulder, M. *Basic Principles of Membrane Technology*; Kluwer: Dordrecht, The Netherlands, 1991.
2. Vankelecom, I.F.J. Polymeric membranes in catalytic reactors. Chem. Rev. **2002**, *102* (10), 3779–3810.
3. Matson, S.L.; Quinn, J.A. Membrane reactors. In *Membrane Handbook*; Ho, W.S.W, Sirkar, K.K., Eds.; Van Nostrand Reinhold: New York, 1992; Chapter 43, 809–832.
4. Zaspalis, V.T.; Burggraaf, A.J. Inorganic membrane reactors to enhance the productivity of chemical processes. In *Inorganic Membranes, Synthesis, Characteristics and Applications*; Ramesh, R.R., Ed.; Van Nostrand Reinhold: New York, 1991; 177–207.
5. Armor, J.N. Overcoming equilibrium limitations in chemical processes. Appl. Catal. A: Gen. **2001**, *222*, 91–99.

6. Armor, J.N. Applications of catalytic inorganic membrane reactors to refinery products. J. Mem. Sci. **1998**, *147*, 217.
7. (accessed Mar 2005); http://www.eltronresearch.com/.
8. Baker, R.W. *Membrane Technology and Applications*; 2nd Ed.; Wiley-VCH: Weinheim, 2004; 509–517.
9. Marconi, J.G.S.; Tsotsis, T.T. *Catalytic Membranes and Membrane Reactors*; Wiley-VCH: Weinheim, 2002; 15–96.
10. Hsieh, H.P. *Inorganic Membranes for Separation and Reaction*; Membrane Science and Technology Series, 3; Elsevier: Amsterdam, 1996.
11. Falconer, J.F.; Noble, R.D.; Sperry, D.P. Catalytic membrane reactors. In *Membrane Separations Technology—Principles and Applications*; Membrane Science and Technology Series, 2; Noble, R.S., Stern, S.A., Eds.; Elsevier: Amsterdam, 1995; Chapter 14, 669–712.
12. Kikuchi, E.; Cattech; 1997, 67.
13. Julbe, A.; Farrusseng, D.; Guizard, C. Porous ceramic membranes for catalytic reactors—overview and new ideas. J. Membr. Sci. 181 (1), 3–20.
14. Marconi, J.G.S.; Tsotsis, T.T. Pervaporation membrane reactors. In *Catalytic Membranes and Membrane Reactors*; Wiley-VCH: Weinheim, 2002; Chapter 3, 15–96.
15. Kemmere, M.F.; Keurentjes, J.T.F. Industrial MRs. In *Membrane Technology in the Chemical Industry*; Nunes, S.P., Peinemann, K.-V., Eds.; Wiley-VCH: Weinheim, Germany, 2001; Chapter 5, 191.
16. Kragl, U.; Dreisbach, C. Membrane reactors in homogeneous catalysis. In *Applied Homogeneous Catalysis with Organometallic Compounds*, 2nd Ed.; Cornils, B., Hermann, W.A., Eds.; Wiley-VCH Weinheim; 2002; Vol. 2, 941–953.
17. Vandezande, P.; Gevers, L.E.M.; Paul, J.S.; Vankelecom, I.F.J.; Jacobs, P.A. High throughput screening for rapid development of membranes and membrane processes. J. Membr. Sci. **2005**, *250* (1–2), 305–310.
18. Marconi, J.G.S.; Tsotsis, T.T. Membrane bioreactors. In *Catalytic Membranes and Membrane Reactors*; Wiley-VCH: Weinheim, 2002; Chapter 4, 15–96.
19. Manam, J.; Sanderson, R. Membrane bioreactors. In *Water Treatment—Membrane Processes*; Mallevialle, J., Odendaal, P.E., Wiesner, M.R., Eds.; McGraw-Hill: New York, 1996; 17.1–17.31.
20. Paiva, A.L.; Malcata, F.X. Integration of reaction and separation with lipases: an overview. J. Mol. Catal. B: Enzym. **1997**, *3*, 99.
21. Alkorta, I.; Garbisu, C.; Llama, M.J.; Serra, J.L. Industrial applications of pectic enzymes: a review. Process Biochem. **1998**, *33* (1), 21.
22. Butterfield, D.A.; Bhattacharyya, D.; Daunert, S.; Bachas, L. Catalytic biofunctional membranes containing site-specifically immobilized enzyme arrays: a review. J. Membr. Sci. **2001**, *181*, 29–37.

Mesoporous Silica Films

Brian W. Eggiman
Hugh W. Hillhouse
Purdue University, West Lafayette, Indiana, U.S.A.

INTRODUCTION

Mesoporous materials are defined by IUPAC as porous materials having a pore diameter between 2 and 50 nm. In 1992, researchers at Mobil developed a technique[1] that relies on the self-assembly of surfactants to create mesoporous silica (MCM-41 and MCM-48) with well-defined pore diameter and pore geometry. Since then great interest has been generated by the possible applications of these materials. The powder form of MCM-41 has become commercially available, and doped or modified versions of the material are being employed as a catalyst in processes at ExxonMobil and Sumitomo Chemical. Techniques to synthesize thin films of these materials have been developed and applications are just now being commercialized. Air Products and Chemicals has recently developed a spin-on mesoporous thin film called MesoELK©, which is being marketed as a low-*k* dielectric thin film for the electronics industry. The potential applications of these films extend beyond low-*k* dielectrics to membrane separations, catalytic supports, and other novel optical, magnetic, and electronic applications. However, the commercialization of many of these ideas as well as the development of new applications is hindered by the nontrivial processing steps involved to reproducibly synthesize these films. Thus, this entry focuses on this issue by reviewing and discussing the steps required to reliably and reproducibly synthesize and process surfactant templated mesoporous silica thin films with controlled nanoscale architecture.

GENERAL METHODS FOR SYNTHESIS OF MESOPOROUS THIN FILMS

The synthesis of mesoporous silica films typically begins with the preparation of precursor solutions. These solutions contain a silica source (typically an alkoxide, although chloride and colloidal precursors can be used), a surfactant molecule used to template the mesostructure, an acid or base catalyst, and solvents. The nanoscale structure is then formed by a cooperative self-assembly of monomeric or partially condensed silica species and surfactant molecules into a liquid crystalline-like mesostructure, as shown in Fig. 1. After the deposition, postsynthesis treatments may be used to further condense the silica into a network and enhance the structural stability of the film. The surfactant template is then removed, leaving empty pores and thus creating an ordered mesoporous thin film where the pore system mimics the liquid crystalline order generated by the surfactant.

There has been extensive research on the synthesis of mesoporous silica thin films since the first reports in 1996.[2] However, most research has focused on two general synthesis techniques: evaporation induced self-assembly and spontaneous film growth from solution.

Evaporation Induced Self-Assembly (EISA) Method of Film Growth

In this technique, the self-assembly is driven by the evaporation of volatile solvents from a thin liquid coating of the precursor solution. As evaporation proceeds, the silica and surfactant species become more concentrated and begin to self-assemble into a nanostructured inorganic–organic hybrid structure. The self-assembly may occur over composition regions quite different from a solution of surfactant alone and may depend very sensitively on the state of the silica source, interfacial interactions, and environmental conditions. Upon self-assembly, both hydrophilic and hydrophobic domains are created with the silica segregating into the hydrophilic regions. Depending on the precursor solution, the silica may be monomeric, oligomeric, or form a condensed interconnected network. If the silica is sufficiently condensed, the template may be removed by calcination, extraction, or other techniques, creating a mesoporous silica thin film.

Both spin coating and dip coating have been used to create a uniform thin liquid coating of the precursor solution. In spin coating, a thin liquid film is formed on a suitable substrate by first dispensing a small amount of the liquid on the substrate followed by spinning (typically less than 6000 rpm). Alternately, the coating solution can be added dropwise to the already spinning substrate (Fig. 2). The solution then covers the substrate and spins off until interfacial and viscous

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120019073

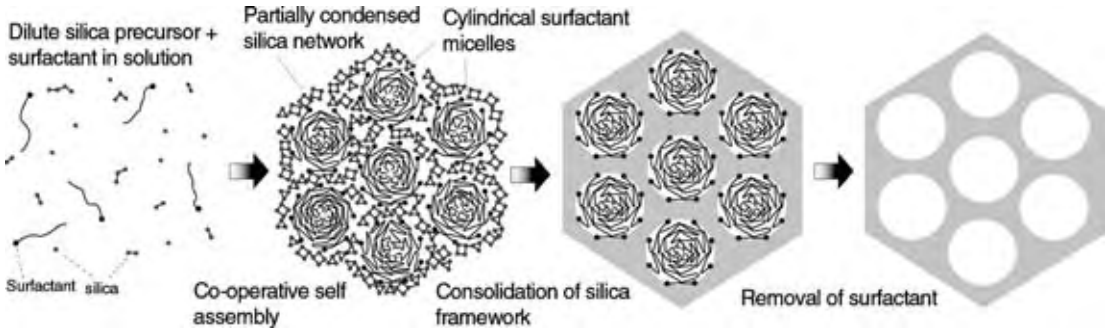


Fig. 1 A schematic for the formation of mesoporosity in a silica film shown for 2-D hexagonal (*P6mm*) phase. *(View this art in color at www.dekker.com.)*

effects balance the centrifugal forces, whereupon a thin liquid film remains.

For dip coating, a suitable substrate is immersed in the coating solution and withdrawn at a controlled rate, usually around 1 mm/sec.[3] During the withdrawal, the liquid film drains until interfacial, viscous, inertial, and disjoining effects balance gravitational forces. As the solvent in the coating solution evaporates and the solution becomes concentrated, nanostructured aggregates first assemble at the substrate–solution and air–solution interfaces and then grow into the liquid film.[4,5] After the solvent evaporates, the self-assembled thin film may be poorly condensed, and thus changes in environmental conditions (humidity, temperature, etc.) may alter the topology of the self-assembly (Fig. 3).

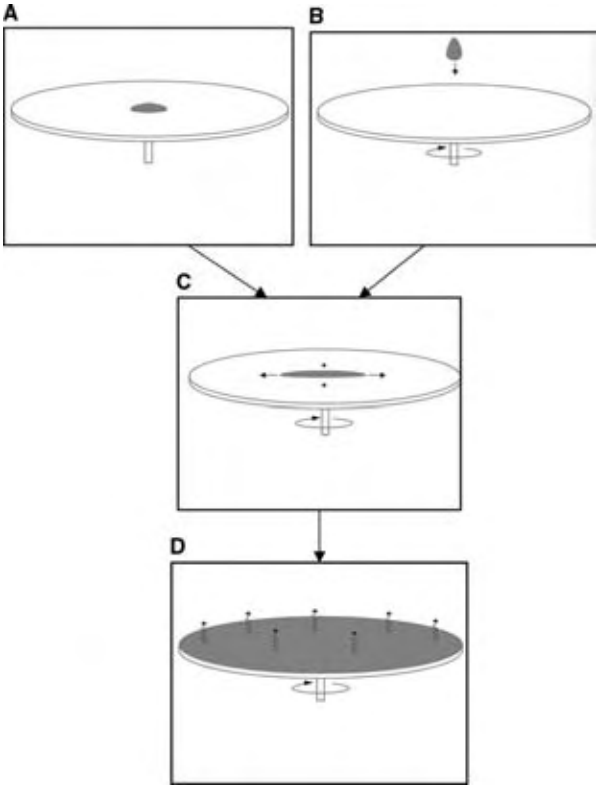


Fig. 2 A schematic for spin coating thin films. The coating solution is either (A) dripped on the substrate while it is at rest; or (B) dripped on while the substrate is spinning; (C) spin-off stage, forming a thin liquid coating on the substrate; (D) solvent evaporation results in self-assembly and thin film formation.

Spontaneous Growth from Solution

In this technique, favorable substrate–solution or air–solution interfacial interactions are used to drive the self-assembly and formation of a thin film of the mesostructured material. The precursor solutions used

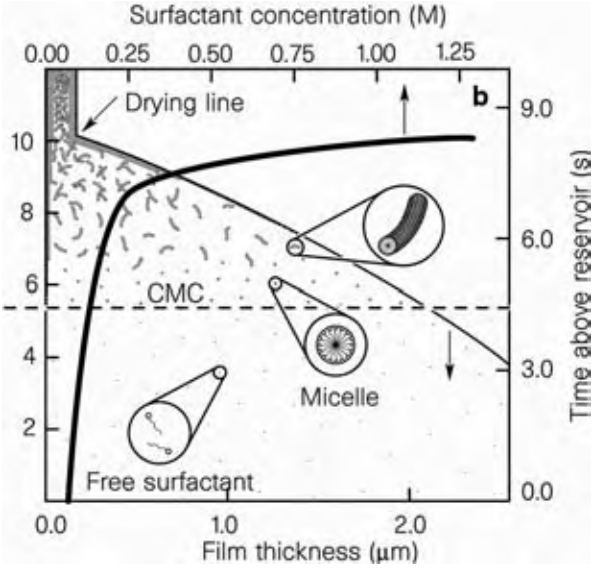


Fig. 3 Mechanism for the synthesis of a hexagonal silica thin film by dip coating. This is zoomed in at the substrate–bulk solution interface. The dotted line indicates the second critical micelle concentration, above which the micelles form cylindrical micellar rods. These rods then begin to assemble at the air–liquid interface and the liquid–substrate interface. (From Ref.[5].)

are typically (though not necessarily) more concentrated than those used for EISA but less concentrated than those used to form mesoporous powders. The key is to adjust the substrate and solution conditions to favor heterogeneous nucleation and growth. The first report of this method of synthesis of films on a substrate was from Yang et al.[2]. In this study, the surfactant and silica precursor heterogeneously nucleated a mesostructured phase on mica in acidic solution. Other reports have generalized the synthesis to include glass, graphite, mica,[7] and polyamide films.[8] In addition, with this technique, directional orientation can be imparted to the mesostructured film by the application of external fields[9] or epitaxial growth.[10]

There have been several other thin film synthesis techniques, such as the use of pulsed laser deposition.[11] However, the promising processing techniques for the economic mass production of mesoporous silica thin films are all based on EISA. Synthesis times can be as short as a few minutes and the processing equipment is readily available. Thus the remainder of this entry focuses on EISA techniques.

PROCESSING MESOPOROUS SILICA THIN FILMS BY SPIN AND DIP COATING

Chemistry and Reaction Mechanisms

The chemistry involved in the formation of mesoporous silica thin films is qualitatively well understood. However, specific reaction mechanisms of the individual steps are still debated. In addition, owing to the complexity of the sol–gel reaction pathways and cooperative self-assembly, full kinetic models have not been developed. From the time of mixing, hydrolysis reactions, condensation reactions, protonation and deprotonation, dynamic exchange with solution nucleophiles, complexation with solution ions and surfactants, and self-assembly, all occur in parallel and are discussed here. Although the sol–gel reactions involved may be acid or base catalyzed, mesoporous silica film formation is carried out under acidic conditions, as silica species are metastable and the relative rates of hydrolysis and condensation reactions lead to interconnected structures as opposed to the stable sols produced at higher pH. Silicon alkoxides are the primary silica source (tetramethyl orthosilicate, tetraethyl orthosilicate, tetrapropyl orthosilicate, etc.) and are abbreviated TMOS, TEOS, and TPOS, respectively. Starting from the alkoxide, $Si(OR)_4$, in ROH and H_2O solution, some of the general reactions are:

$$\equiv Si\text{–}OR + H_2O \underset{\text{esterification}}{\overset{\text{hydrolysis}}{\rightleftharpoons}} \equiv Si\text{–}OH + ROH$$

$$\equiv Si\text{–}OR + HO\text{–}Si\equiv \underset{\text{alcoholysis}}{\overset{\text{alcohol condensation}}{\rightleftharpoons}} \equiv Si\text{–}O\text{–}Si\equiv + ROH$$

$$\equiv Si\text{–}OH + HO\text{–}Si\equiv \underset{\text{hydrolysis}}{\overset{\text{watercondensation}}{\rightleftharpoons}} \equiv Si\text{–}O\text{–}Si\equiv + H_2O$$

$$\equiv Si\text{–}OH + H^+(aq) \underset{\text{deprotonation}}{\overset{\text{protonation}}{\rightleftharpoons}} \equiv Si\text{–}OH_2^+ \quad (\text{for pH} \leq 2)$$

Typically, condensation begins before hydrolysis is complete, and the reactions result in silica species of the form $Si(OR)_x(OH)_y(OH_2^+)_z$ where $x + y + z = 4$, in addition to further condensed oligomers. The isoelectric point of silica occurs at a pH of ~2, and thus below this pH some silica species will protonate to form stable cationic species. For the pH range of 0–2, for solutions prepared from only the alkoxide, HCl, alcohol, and water, the average z is less than 1. It is believed that both the hydrolysis and condensation reactions proceed via nucleophilic substitution (S_N2) reactions involving pentacoordinated reaction intermediates. This mechanism is commensurate with steric and inductive effects observed experimentally in silica solutions. Larger alkoxide groups result in steric hindrances of the pentacoordinated intermediate and increasing hydrolysis rates in the series TMOS > TEOS > TPOS. In addition, it is proposed that the first step in the condensation reactions is protonation of a Si–OH group, and it is observed experimentally that the condensation rate is proportional to the H^+ concentration.[12]

After hydrolysis begins, the neutral, charged, or partially condensed silica species may begin to associate and form complexes with counterions and surfactant molecules in solution. However, these associations evolve in time as the hydrolysis and condensation reactions proceed. Controlling this ''aging'' time is the key to reproducibly synthesize some mesophase structures (as discussed later). After forming a thin liquid coating, the ethanol and water begin to evaporate from the liquid film. The concentration of the solution increases, causing the silica complexes to cooperatively self-assemble with the surfactant into lyotropic liquid crystalline arrays.

Self-Assembly of Surfactant and Silica Species

The geometry of the assembled aggregates and the topology of the resulting mesophase depend strongly on the properties of the surfactant, the silica species, and solution conditions. Many of these aspects may

be understood qualitatively by considering a molecular packing parameter (sometimes referred to as the critical packing parameter, CPP) based on space filling arguments:

$$\text{CPP} = \frac{v}{al}$$

where v is the volume of the hydrophobic chain of the surfactant, a the effective cross-sectional area that one head group of the surfactant occupies on the micelle surface, and l the critical length of the tail. The volume may be estimated from the density of a liquid of pure hydrophobic chains, and the critical length will be less than the extended end-to-end tail length. The key parameter is a, which may be very sensitive to environmental or solution phase conditions. For example, adding a salt to solution will decrease the effective area of a charged head group by screening the repulsive electrostatic interactions from neighboring head groups. The CPP can be used to predict the geometry of the aggregate that is geometrically favored to form in solution, as shown in Fig. 4. For instance, a CPP below one-third will result in spherical micelles that may pack in a cubic structure. A CPP between one-third and one-half will result in cylindrical geometries that may assemble into a hexagonal array, and a CPP around 1 will yield phases with a low net curvature, such as lamellar phase or cubic bicontinuous phases.[13]

Self-assembly using cationic quaternary ammonium surfactants

The most prevalent cationic surfactant used for mesoporous thin film synthesis is cetyltrimethylammonium bromide (CTAB). It is a large head group cationic surfactant and has been shown to produce 3-D cubic ($Pm3n$), 3-D hexagonal ($P6_3/mmc$), and 2-D hexagonal mesostructures ($p6m$).[14] The resulting pore size of the mesoporous film may be tailored by using different alkyl tail lengths or adding swelling agents such as trimethyl benzene. The head group area may be controlled by adjusting salt concentrations or using larger alkyl groups on the ammonium head group. In addition, gemini surfactants (which are two covalently linked cationic quaternary ammonium surfactants) may be used to further tailor the geometry.[15] Also, it is important to note that the choice of surfactant counter anion and acid can significantly affect the self-assembly and are seen to follow the Hofmeister series.[16] These anions can affect the water structure in the solution, and structural differences between CTAB and cetyltrimethylammonium chloride (CTAC) are commonly observed.

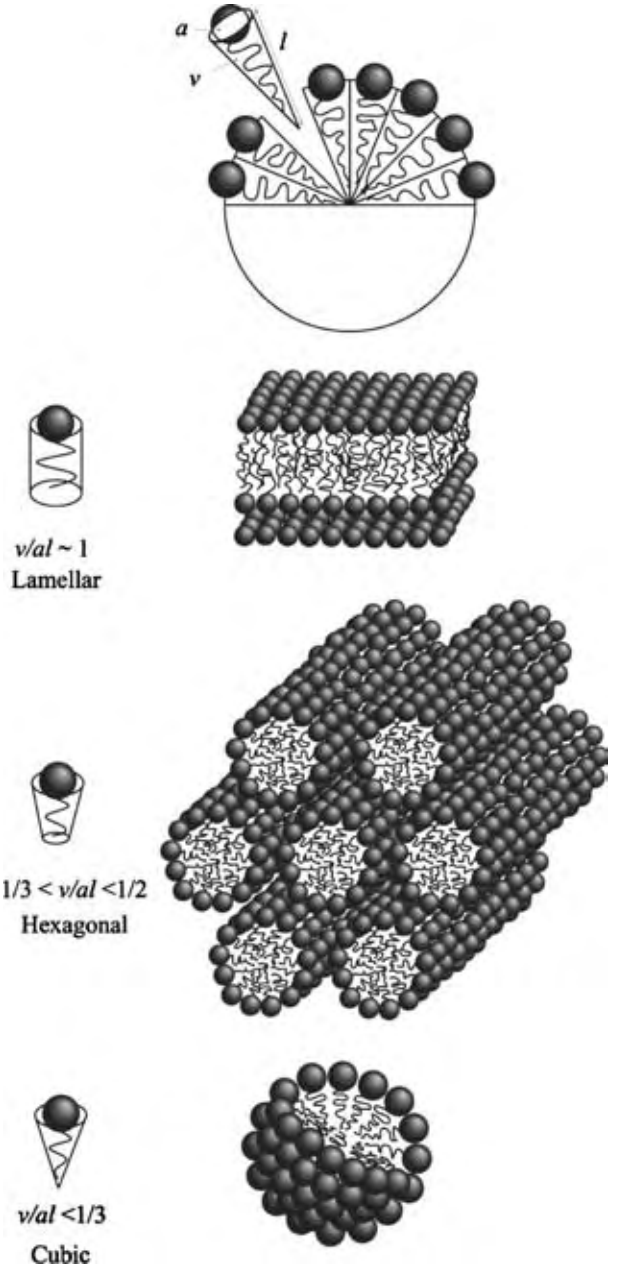


Fig. 4 Critical packing parameters and characteristic structure typical to mesoporous silica films. The top shows a cross-section of a micelle, displaying the parameters used to calculate the CPP. The spherical head group for the surfactant represents the effective cross-sectional area that the head group occupies. *(View this art in color at www.dekker.com.)*

Self-assembly using nonionic polymeric surfactants

Some of the most widely used surfactant templates are nonionic amphiphilic triblock copolymers composed of poly(ethylene oxide)–poly(propylene oxide)–poly (ethylene oxide) blocks.[17] These templates are nontoxic,

biodegradable, and commercially available. The two most prevalent block lengths used for templating mesoporous silica are Pluronic P123 ($EO_{20}PO_{70}EO_{20}$) and Pluronic F127 ($EO_{106}PO_{70}EO_{106}$) available from BASF. The ternary phase behavior of these amphiphiles with water and ethanol has been extensively studied by Holmqvist et al.[18,19] The external poly-(ethylene oxide) (PEO) blocks are the more hydrophilic portions and the internal poly-(propylene oxide) (PPO) blocks are more hydrophobic. The micelle is constructed with a hydrophobic PPO core surrounded by the hydrophilic PEO blocks. It is intuitive that the lengths of the blocks in the surfactant will have direct effects on the mesostructure of the film. The length of the PEO block mainly determines the phase mesostructure and wall thickness of the film. Typically, shorter PEO blocks (less than 10) will result in lamellar structures, medium length chains (between 17 and 37) favor two-dimensional hexagonal *p6mm* structures, and longer chains (in hundreds) favor cubic structures.[20] As the core of the micelle is constructed by the PPO block, the length of this block directly affects the pore size of the resulting mesoporous silica. The pore size can be tailored by varying the length and the hydrophobicity of this block, and these surfactants produce some of the largest pore diameters, as evident by the reported pore sizes of 120 Å for a cubic powder templated with a poly-(ethylene oxide)$_{39}$–poly-(butylene oxide)$_{47}$–poly-(ethylene oxide)$_{39}$ surfactant,[21] and 85 Å for a cubic film templated with F127. These diameters are much larger than those produced by low molecular weight alkyl(ethylene oxide) nonionic surfactants like CTAB, which has had a pore size of 34 Å reported for a cubic film.[17] In addition, these polymeric surfactants tend to produce mesoporous structures with significant microporosity between the mesopores.

While a surfactant can be chosen that preferentially forms one phase or another, any individual surfactant can form many different mesostructures. A recent study examined the effect of P123 concentration on the resulting mesophase.[3] Using a parameter Φ, defined as

$$\Phi = \frac{V_{pol}}{V_{pol} + V_{inorg}}$$

where V_{pol} is the volume of surfactant present in solution and V_{inorg} is the volume of nonvolatile components, they were able to predict what mesostructure would be produced at a certain surfactant concentration. It was found that for values of Φ between 30 and 36, a cubic structure was formed; for values between 40 and 55, a hexagonal phase is formed; and for larger values, 63–75, a lamellar phase is obtained. These data are consistent with the trends determined by the critical packing parameter. A low Φ indicates less volume of surfactant in solution than that of silica and hence fewer molecules of surfactant. This will increase the curvature of the micelles because of an increase of the effective area of one head group, and if the CPP is less than one-third, a cubic structure will form, which is what is observed for the low values of Φ.

Preparation of Dip and Spin Coating Solutions

Typically, silica precursor solutions and surfactant solutions are prepared separately where the alkoxide precursors in the silica solution are allowed to hydrolyze for some period before mixing with a surfactant solution to form the coating solution. Likewise, the surfactant solution is allowed to equilibrate before the mixing.

Preparation and prehydrolysis of the silica precursor solution

The silica precursor solution is typically prepared from a silicon alkoxide, water, and hydrochloric acid. The silica precursor solution is maintained at a pH of around 2, which coincides with the isoelectric point of silica. This is done to allow the hydrolysis of the silica while sustaining a low condensation rate. The first step in the acid catalyzed condensation is protonation of a Si–OH group, thus lower concentrations of the H^+ lower the overall condensation rate. A typical silica precursor preparation involves mixing 10.4 g of TEOS with 5.4 g water, 0.00197 g HCl, and 12 g of ethanol with stirring at room temperature for 20 min.[3] Once mixed, the reaction begins and solution age can greatly affect the self-assembly in the next stage. However, solutions may be prehydrolyzed and then chilled to retard the evolution of silica oligomers in solution.

Preparation and equilibration of the surfactant solution

A solution containing the surfactant and a portion of the solvent is often prepared separately from the silica solution and allowed to equilibrate. The larger molecular weight triblock copolymers may take some time to dissolve and should be allowed to equilibrate in solution to standardize the preparation of the coating solution. As there are no reactions taking place in these solutions, they may be stored and kept after preparation. A typical solution consists of 1.7 g Pluronic P123 ($EO_{20}PO_{70}EO_{20}$) and 8 g of ethanol.[3]

Preparation and aging of the coating solution

The coating solution refers to the final solution that is used for dipping or spinning and is prepared by mixing the silica solution and the surfactant solution. To control the self-assembly during solvent evaporation, the coating solutions are typically aged to allow the silica species and association complexes to evolve. The aging of the coating solution may differ widely depending on the components of the solution and the phase desired. It has been shown that for Pluronic P123 templates, a short coating solution aging time (10 min) is required to obtain a cubic mesostructure, while longer aging times (180 min) are required for hexagonal mesostructures. Also, trends show that lower temperatures, 5–15°C, favor a cubic mesostructure, while ambient temperatures are suitable for hexagonal films.[3] Keeping with our earlier example, the final composition of a typical coating solution would be 10.4 g TEOS, 5.4 g water, 0.00197 g HCL, 20 g ethanol, and 1.7 g Pluronic P123.

There are also many alternate routes to prepare the coating solution, and some syntheses involve no prehydrolysis of the silicon alkoxides.[22]

Coating Process and Self-Assembly of Mesostructured Films

A prerequisite for dip coating and spin coating is that the coating solution must wet the substrate. This is a necessary but not a sufficient criterion, as ultimately the disjoining pressure and liquid–solid interfacial energy must be able to resist droplet formation on the substrate. This is a function of the chemical interactions between the coating solution and the substrate in addition to surface roughness. Most research has employed borosilicate glass slides as substrates because of the low contact angle with water and low cost. Specific interactions with the substrate can be very important and may dramatically affect the self-assembly process.

In spin coating, after the initial stages where excess fluid spins off the substrate, viscous forces and interfacial interactions balance centrifugal forces, and a thin liquid film remains for the EISA. Thus, for a given spinning rate, higher concentrations of the surfactant and silica precursor will result in thicker films because of viscous effects, while faster spinning rates result in thinner films. Details of the fluid mechanics have been discussed by Bornside et al.[23] and in subsequent articles. In contrast, in dip coating, faster substrate withdrawal rates result in thicker coatings as more fluid is entrained on the substrate.[14] The final mesoporous film thickness is tunable, but typically preparations yield films around 200 nm thickness. In both techniques, x-ray diffraction data have shown decreased order in thicker films.

As the assembly of the mesostructure is driven by solvent evaporation and the ultimate equilibrium between the film and the vapor phase, external parameters such as relative humidity and the concentration of solvent in the vapor phase affect the formation of the mesostructured film. Recently, the effect of relative humidity on mesostructure has been studied.[6] Keeping the coating solution constant, films were dipped at three different relative humidities: 20%, 40%, and 70%. It was resolved that above 40% relative humidity, two regimes of solvent evaporation occur. During withdrawal, ethanol evaporates quickly, normally less than 20 sec, followed by the slower evaporation of water, which can take up to 1 min with a relative humidity of 70%. It was also determined that humidity can create a phase transformation within the films. If the relative humidity was too low, RH = 20%, the rate of water loss from the film occurred too quickly, which did not allow enough time for the CTAB to self-assemble, resulting in disordered films with no nanostructure. A relative humidity of 40% resulted in a film with *p6m* hexagonal structure, whereas at RH = 70%, a cubic *Pm3n* structure was preferentially formed. It was determined that the water content in the film at the high relative humidity caused an increase in the affective area of the head group of the surfactant, *a*, which promoted a cubic structure over the hexagonal structure that was formed at RH = 40%.

Postsynthesis Treatments of Mesostructured Films

Films are typically allowed to dry immediately after deposition to drive the silica condensation reaction toward completion. However, the film may not be fully condensed and stable after drying, and postsynthesis treatments are often employed to improve film stability, reduce shrinkage upon template removal, and drive the condensation reaction toward completion.

Drying and thermal treatment

After the film has been synthesized, the substrate and the film are normally allowed to dry for several hours. The details surrounding the drying of the film vary greatly depending on the coating solution and the desired film properties. However, it is important to note that immediately after dip or spin coating, the silica may be poorly condensed and capable of significant structural changes in response to the environment. Recently, this effect was dramatically demonstrated by Cagnol et al.[6] After coating, the mesophase structure was repeatedly modulated by varying the humidity.

When a film that was deposited at a relative humidity of 40% was exposed to an environment with a relative humidity of 85% 10 min after deposition, a phase change from 2-D hexagonal to cubic was observed. Other relative humidity changes led to a disordered to ordered transition in the film.

Thermal postsynthesis treatments have also been studied over a range of temperatures. These treatments are normally carried out well below the temperature required to remove the template and strive to strengthen the silica network while the template is still in place. In one report, the samples were treated at increasing temperatures, and it was shown that between the temperatures of 150°C and 175°C the order of the mesostructure increased.[24]

Chemical treatments

After drying, the films may be subjected to a variety of chemical postsynthesis treatments to increase the stability of the mesostructure. In one report, the as-deposited film was exposed to tetraethyl orthosilicate (TEOS) vapor in a closed container for 3 hr at 135°C.[22] This treatment, which was first carried out on hexagonal silica thin films templated with CTAB, adds silica to the structure, increases the density of the silica walls, and significantly improves the thermal stability of the films. Template removal of these treated materials yielded films with little to no peak shift in x-ray diffraction patterns and significantly reduced cracking. Ammonia atmospheres have also been reported to base catalyze the condensation of the silica and increase the thermal stability.[25] In the base treatment, the thin film was exposed to a saturated ammonia environment for 45 min. It was reported that upon template removal, the d spacing of the (200) plane shrunk by 15%. However, a prolonged exposure to ammonia creates cracks and delamination on the films. Hydrothermal treatments have also been reported to enhance the stability of the film. Heating in deionized water at 80°C overnight was reported to improve the polymerization of the silica and thermal stability of the film.[17]

Template Removal to Yield Mesoporous Films

Once the film is stable, the surfactant template must be removed to create the mesoporous film from the nonporous mesostructured film. However, upon removal of the template, films may contract and crack, and therefore care must be taken during removal to yield continuous and well-ordered films.

Calcination

Calcination of silica thin films involves heating the film to between 350°C and 500°C in the presence of oxygen to remove the templating surfactant. At these temperatures, the surfactants decompose and react with oxygen to form gaseous organic products (CO_2, H_2O, etc.), which diffuse out of the film. The temperature profile usually consists of a temperature ramp of 1°C/min to the desired calcination temperature, a holding period between 4 and 8 hr, and a decrease back to room temperature at a rate of 1°C/min. The calcination process and its effects on the film structure were studied in detail by Kleitz et al.[26]. They found significantly different decomposition pathways for different surfactants as characterized by in situ x-ray diffraction, thermogravimetric analysis, and mass spectrometry. For example, when a hexagonal film synthesized with the triblock copolymer Pluronic P123 was examined during calcination, one main step for mass loss was found. This loss occurred below 280°C and was accounted for by decomposition of the organic surfactant into smaller fragments with m/z ratios of 42, 58, and 74, which correspond to C_3H_6, C_3H_6O, and $C_3H_6O_2$, respectively. These components were removed over a relatively small range of temperatures, between 150 and 200°C. Between 200°C and 400°C, species were fully oxidized to carbon dioxide and water. Alternately, for a cubic mesostructured film templated by CTAB, between 150 and 250°C, the head group of CTAB surfactant was removed by way of Hofmann degradation in which a β-elimination reaction results in the loss of a β-proton and the cleavage of the trimethylamine head group from the alkyl chain. What is left in the pore is a mixture of different hydrocarbon species, which leave as C_2H_2, C_3H_6, C_4H_7, C_5H_9, CO_2, and water between 250°C and 350°C.

Solvent extraction

Solvent extraction is the process where the as-synthesized thin film containing the template is exposed to heated solvent to remove the template from the pores. The technique can involve maintaining the thin film in the solvent with stirring and heating, either in a closed container or in a reflux apparatus. One report indicates that the surfactant was removed by hot ethanol for 5 hr by soxhlet extraction.[25]

Ozone treatment

Surfactant removal in ozone has been reported in two different forms: one in which ozone was produced by UV treatment under ambient conditions[27] and the other in which ozone was bubbled through a

water solution that contained the bulk (powder) mesostructured silica.[28] In the UV treatment, bulk powder MCM-41 that was templated by CTAB was placed for 24 hr under a UV lamp, which produced wavelengths (254 and 180 nm) known to produce ozone from ambient oxygen. It was theorized that water and carbon dioxide were the major products of the reaction, which were eliminated from the pores by diffusion. Diffuse-reflectance infrared Fourier transform (DRIFT) spectra and nitrogen sorption isotherms were utilized to verify that the template was removed. In the second type of ozone treatment, the surfactant, CTAB, was removed from the powder MCM-41 by bubbling ozone for 30 min through a water solution, which contained the bulk mesostructured sample. This ozone treatment is advantageous when compared with the UV treatment because of the ability to control the amount of ozone that interacts with the surfactant and the amount of time that is required for the removal, 30 min for the bubbling compared with 24 hr for the UV treatment.

CHARACTERIZATION TECHNIQUES

To determine the success of synthesis efforts and to provide feedback on quality control, the mesoporous silica films are characterized by several techniques. The ability to determine the continuity and quality of the films must be quick and accurate. Owing to this fact, x-ray diffraction has proven the easiest way to resolve the film structure. However, phase identification by x-ray diffraction alone is not always conclusive (e.g., when only one peak is observed). Other techniques must be employed to fully determine the phase of the mesostructure.

X-Ray Diffraction (XRD)

X-ray diffraction is the most prevalent characterization technique to determine the phase of a film. Bragg diffraction is the result of the constructive interference of scattered x-rays that are scattered from regularly spaced variations in the electron density. The spacing of these variations is determined via Bragg's law:

$$d_{hkl} = \frac{\lambda}{2\sin\theta}$$

where λ is the wavelength of the incident radiation (typically 1.54 Å, for laboratory x-ray sources using Cu Kα radiation), d_{hkl} the periodic spacing indexed by the ($h\,k\,l$) Miller plane, and θ angle of incidence. Bragg's law is basically a scalar version of the von Laue equations. It encodes a spacing criterion necessary to observe a diffraction peak. However, there is an additional criterion for reflection geometry diffractometers that these variations be perpendicular to the plane of the film. Regular spacings along any other direction will not result in observed diffraction peaks. Details of the XRD pattern interpretation from mesoporous thin films are discussed by Hillhouse et al.[29]. The mesophase may then be identified from the peak positions. For the example of a lamellar phase thin film, where the sheets are parallel to the substrate, the observed peaks will be (1 0 0), (2 0 0), (3 0 0), and so on. These peaks will be equally spaced and diminishing in intensity. This diffraction pattern is also to be expected from an oriented 2-D hexagonal phase (with the 1 0 0 parallel to the substrate). However, upon removal of the template from the mesostructured silica in the lamellar phase, the film loses all structure and no x-ray pattern is expected. The hexagonal phase, on the other hand, will retain the diffraction peaks from (1 0 0) and, depending on the degree of orientation, the (2 0 0) and (3 0 0) peaks. The presence of these peaks after template is typically reported as evidence that a hexagonal structure was obtained, as shown in Fig. 5. However, many cubic phase films may show only one peak in the XRD pattern, and thus 2-D XRD, TEM, or polarized optical microscopy should be used to verify the mesophase structure.

Transmission Electron Microscopy (TEM)

Transmission electron microscopy is utilized regularly to characterize silica thin films. Although TEM is a more labor-intensive characterization technique than x-ray diffraction, it has several key advantages. High resolution transmission electron microscopy images from multiple orientations allow the mesostructure of a film to be determined conclusively. Pore size, wall thickness, orientation relative to the substrate, and long range order of the film can also be quickly determined. As the spacings in the mesostructure are much larger than that for traditional crystals (2–20 nm as compared with only a few Å), microscopes of moderate resolution may be used. Typically, clear images may be obtained on 100 keV microscopes. Sample preparation for TEM can be tedious and difficult. However, for mesophase identification, the film can be simply scraped off the substrate, dispersed in a solvent (ethanol), and coated onto TEM grids. Examples of TEM images can be seen in Figs. 6C–6F. These give clear insight into the structure of the films produced, as seen in Fig. 6D, where the hexagonal packing of the pores in a 2-D hexagonal film is apparent, and in Fig. 6E, where the 90° orientation of the pores with respect to each other elicits a cubic mesostructure.

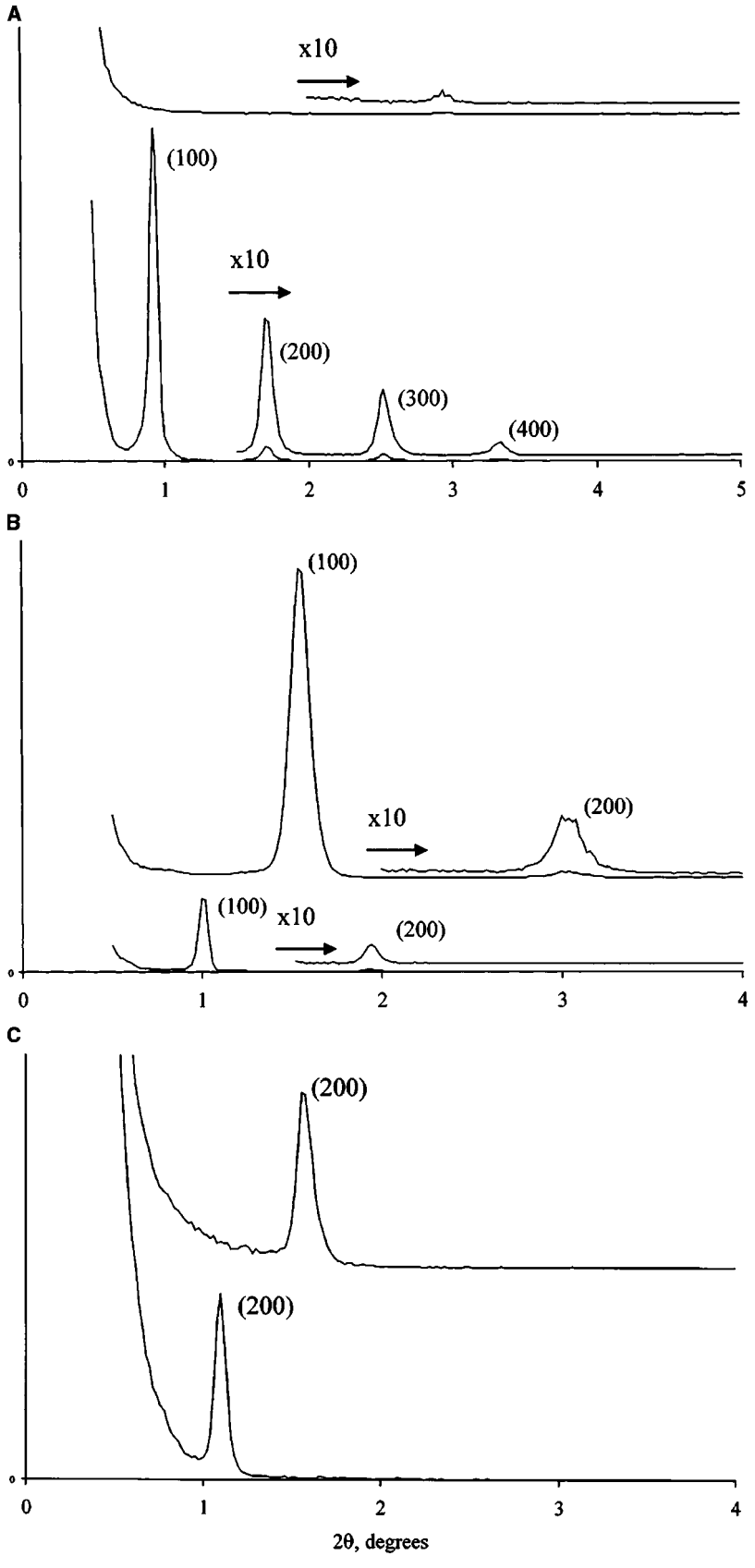


Fig. 5 X-ray diffraction patterns for silica films: (A) lamellar phase, (B) hexagonal, and (C) cubic. For each set of patterns, the bottom is as-synthesized, and the top is calcined. (From Ref.[3].)

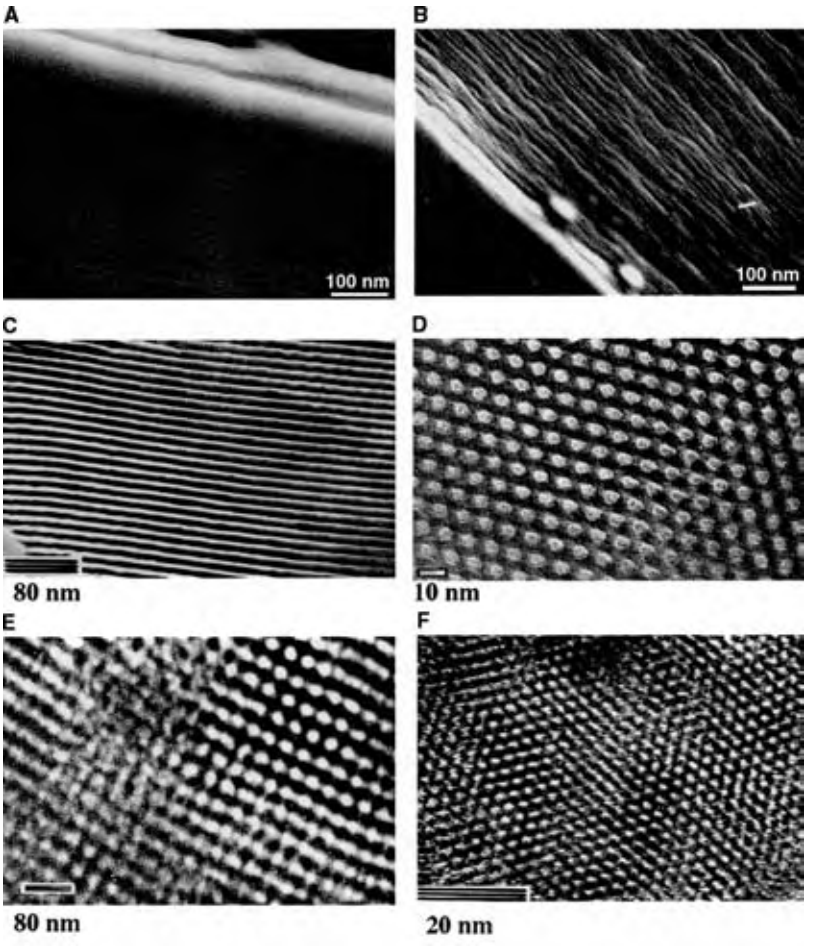


Fig. 6 (A,B) SEM images of an as-deposited and calcined hexagonal silica film, respectively, templated by Pluronic P123. (C) TEM image of the hexagonal film along the (1 1 0) zone axis. (D) TEM image of the hexagonal film along the (1 0 0) zone axis. (E,F) TEM images of a cubic silica film templated by Pluronic F127 along the (E) (1 1 0) zone axis and (F) (1 0 0) zone axis. (From Ref.[17].) *(View this art in color at www.dekker.com.)*

Scanning Electron Microscopy (SEM)

Typically, SEM is used to determine the continuity and the thickness of the film. Large-scale cracks and defects of the film can be determined by top view SEM images, and film thickness is easily obtained from cross-sections. It has been used extensively with powdered samples of mesoporous silica to determine morphology and thus assist in phase determination. However, thin films produced by dip and spin coating typically have a scarcity of features that may be resolved in SEM. This can be seen in Figs. 6A and 6B, SEM images of a mesoporous films templated with P123.

Other Characterization Techniques

There are other techniques that can be used to characterize the thin film. For powdered materials, nitrogen adsorption–desorption isotherms are typically used to determine the pore volume and surface area. However, this technique is rather difficult for films, as the physical amount of sample produced in one synthesis is not enough to perform an analysis. Attenuated total reflectance-infrared spectroscopy (ATR-IR) may be used to determine the chemical bonding characteristics of a thin film. For example, the extent of condensation of silica can be determined by monitoring the characteristic frequency of Si–OH.

CONCLUSIONS

Interest in the synthesis and processing of mesoporous silica materials has grown extensively since their discovery in 1992, and the exciting potential that these films hold in low-k dielectrics, sensors, nanowire fabrication, catalysis, membrane separations, and many other applications will continue to fuel academic and industrial interest in these films. While there are many new synthesis routes for processing mesoporous silica thin films, spin coating and dip coating remain the most facile methods available. These methods deliver high quality reproducible films that can be used for any of the variety of applications.

REFERENCES

1. Beck, J.S.; Vartuli, J.C.; Roth, W.J.; Leonowicz, M.E.; Kresge, C.T.; Schmitt, K.D.; Chu, C.T.-W.; Olson, D.H.; Sheppard, E.W.; McCullen, S.B.; Higgins, J.B.; Schlenker, J.L. A new family of mesoporous molecular sieves prepared with liquid crystal templates. J. Am. Chem. Soc. **1992**, *114* (27), 10,834–10,843.
2. Yang, H.; Kuperman, A.; Coombs, N.; Mamiche-Afara, S.; Ozin, G.A. Synthesis of oriented films of mesoporous silica on mica. Nature **1996**, *379* (6567), 703–705.
3. Alberius, P.C.A.; Frindell, K.L.; Hayward, R.C.; Kramer, E.J.; Stucky, G.D.; Chmelka, B.F. General predictive syntheses of cubic, hexagonal, and lamellar silica and titania mesostructured thin films. Chem. Mater. **2002**, *14* (8), 3284–3294.
4. Brinker, C.J.; Frye, G.C.; Hurd, A.J.; Ashley, C.S. Fundamentals of sol–gel dip coating. Thin Solid Films **1991**, *201* (1), 97–108.
5. Lu, Y.; Ganguli, R.; Drewien, C.A.; Anderson, M.T.; Brinker, C.J.; Gong, W.; Guo, Y.; Soyez, H.; Dunn, B.; Huang, M.; Zink, J.I. Continuous formation of supported cubic and hexagonal mesoporous films by sol–gel dip-coating. Nature **1997**, *389* (6649), 364–368.
6. Cagnol, F.; Grosso, D.; Soler-Illia, G.J.D.A.A.; Crepaldi, E.L.; Babonneau, F.; Amenitsch, H.; Sanchez, C. Humidity-controlled mesostructuration in CTAB-templated silica thin film processing. The existence of a modulable steady state. J. Mater. Chem. **2003**, *13* (1), 61–66.
7. Aksay, I.A.; Trau, M.; Manne, S.; Honma, I.; Yao, N.; Zhou, L.; Fenter, P.; Eisenberger, P.M.; Gruner, S.M. Biomimetic pathways for assembling inorganic thin films. Science **1996**, *273* (5277), 892–898.
8. Miyata, H.; Kuroda, K. Alignment of mesoporous silica on a glass substrate by a rubbing method. Chem. Mater. **1999**, *11* (6), 1609–1614.
9. Hillhouse, H.W.; Okubo, T.; van Egmond, J.W.; Tsapatsis, M. Preparation of supported mesoporous silica layers in a continuous flow cell. Chem. Mater. **1997**, *9* (7), 1505–1507.
10. Hillhouse, H.W.; van Egmond, J.W.; Tsapatsis, M.; Hanson, J.C.; Larese, J.C. Synthesis and characterization of ordered arrays of topological defects in mesoporous silica films. Chem. Mater. **2000**, *12* (10), 2888–2893.
11. Balkus, K.J.; Scott, A.S.; Gimon-Kinsel, M.E.; Blanco, J.H. Oriented films of mesoporous MCM-41 macroporous tubules via pulsed laser deposition. Microporous Mesoporous Mater. **2000**, *38* (1), 97–105.
12. Brinker, C.J.; Scherer, G.W. *Sol–Gel Science*; Academic Press: San Diego, 1990.
13. Israelachvili, J. *Intermolecular and Surface Forces*, 2nd Ed.; Academic Press: San Diego, 2003; 366–385.
14. Grosso, D.; Babonneau, F.; Albouy, P.-A.; Amenitsch, H.; Balkenenda, A.R.; Brunet-Bruneau, A.; Rivory, J. An in situ study of mesostructured CTAB-silica film formation during dip coating using time-resolved SAXS and interferometry measurements. Chem. Mater. **2002**, *14* (2), 931–939.
15. Huo, Q.; Margolese, D.I.; Stucky, G.D. Surfactant control of phases in the synthesis of mesoporous silica-based materials. Chem. Mater. **1996**, *8* (5), 1147–1160.
16. Leontidis, E. Hofmeister anion effects on surfactant self-assembly and the formation of mesoporous solids. Curr. Opin. Colloid Interface Sci. **2002**, *7* (1–2), 81–91.
17. Zhao, D.; Yang, P.; Melosh, N.; Feng, J.; Chmelka, B.F.; Stucky, G.D. Continuous mesoporous silica films with highly ordered large pore structure. Adv. Mater. **1998**, *10* (16), 1380–1385.
18. Holmqvist, P.; Alexandridis, P.; Lindman, B. Phase behavior and structure of ternary amphiphilic block copolymer–alkanol–water systems: comparison of poly(ethylene oxide) poly(propylene oxide) to poly(ethylene oxide) poly(tetrahydrofuran) copolymers. Langmuir **1997**, *13* (9), 2471–2479.
19. Holmqvist, P.; Alexandridis, P.; Lindman, B. Modification of the microstructure in block copolymer–water–''oil'' systems by varying the copolymer composition and the ''oil'' type: small-angle x-ray scattering and deuterium-NMR investigation. J. Phys. Chem. B **1998**, *102* (7), 1149–1158.
20. Kipkemboi, P.; Fogden, A.; Alfredsson, V.; Flodstrom, K. Triblock copolymers as templates in mesoporous silica formation: structural dependence on polymer chain length and synthesis temperature. Langmuir **2001**, *17* (17), 5398–5402.
21. Yu, C.; Yu, Y.; Zhao, D. Highly ordered large caged cubic mesoporous silica structures templated by triblock PEO–PBO–PEO copolymer. Chem. Commun. **2000**, (7), 575–576.
22. Nishiyama, N.; Tanaka, S.; Egashira, Y.; Oku, Y.; Ueyama, K. Enhancement of structural stability of mesoporous silica thin films prepared by spin-coating. Chem. Mater. **2002**, *14* (10), 4229–4234.
23. Bornside, D.E.; Macosko, C.W.; Scriven, L.E. Spin coating: one-dimensional model. J. Appl. Phys. **1989**, *66* (11), 5185–5193.

24. Klotz, M.; Ayral, A.; Guizard, C.; Cot, L. Synthesis conditions for hexagonal mesoporous silica layers. J. Mater. Chem. **2000**, *10* (3), 663–669.
25. Grosso, D.; Balkenenda, A.R.; Albouy, P.A.; Ayral, A.; Amenitsch, H.; Babonneau, F. Two-dimensional hexagonal mesoporous silica thin films prepared from block copolymers: detailed characterization and formation mechanism. Chem. Mater. **2001**, *13* (5), 1848–1856.
26. Kleitz, F.; Schmidt, W.; Schüth, F. Calcination behavior of different surfactant-templated mesostructured silica materials. Microporous Mesoporous Mater. **2003**, *65* (1), 1–29.
27. Keene, M.T.J.; Denoyel, R.; Llewellyn, P.L. Ozone treatment for the removal of surfactant to form MCM-41 type materials. Chem. Commun. **1998**, (20), 2203–2204.
28. Büchel, G.; Denoyel, R.; Llewellyn, P.L.; Rouquerol, J. In situ surfactant removal from MCM-type mesostructures by ozone treatment. J. Mater. Chem. **2001**, *11* (2), 589–593.
29. Hillhouse, H.W.; van Egmond, J.W.; Tsapatsis, M.; Hanson, J.C.; Larese, J.Z. The interpretation of X-ray diffraction data for the determination of channel orientation in mesoporous films. Microporous Mesoporous Mater. **2001**, *44*, 639–643.

Metallocene Catalysts for Olefin Polymerization

T. C. Chung
Department of Materials Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.

INTRODUCTION

Metallocene is a class of organometallic compounds that contains a transition metal sandwiched by one or two cyclopentadienyl ligands. Metallocene catalysts used in α-olefin polymerization are usually prepared by the combination of a metallocene (catalyst) from a group 4 metal (titanium, zirconium, and hafnium) and a cocatalyst to form an active metal cation. This cation contains both M-C species and an empty d-orbital suitable for α-olefin coordination/insertion reaction. The cocatalysts can be divided into two categories, including organoaluminum compounds (alkylaluminums and alkylaluminoxanes) and anionic counterions (borates or fluorinated borates), which form a weak coordination with the cationic active site. One unique feature of metallocene catalysts is the well-defined active site with a tunable ligand structure for manipulating olefin polymerization, and the resulting polymer structures. Although metallocene catalyst are soluble in common hydrocarbon solvents (used in α-olefin polymerization), the commercial catalysts are usually immobilized on supports for controlling polymer morphology and for reducing the requirement of a large excess of cocatalyst.

BACKGROUND INFORMATION

As early as the 1950s, the metallocene/alkylaluminum system was reported for mediating ethylene polymerization.[1–3] However, the polymer obtained was of low molecular weight and had extremely low catalytic activity, compared with those prepared by a heterogenous Ziegler–Natta catalyst system ($TiCl_4/AlEt_3$), discovered a few years earlier. Despite the ill-defined active sites, most of the research activities had focused on Ziegler–Natta catalysts and the resulting polyolefins in the following two decades. This includes polyethylene (PE), polypropylene (PP), poly(4-methyl-1-pentene), ethylene–propylene–elastomer, and ethylene–propylene–diene rubber. The commercial importance of polyolefin products and the furious competition among the major producers together led to technological advancement in developing Ziegler–Natta catalysts with superior activity and stereospecificity, and also to the development of economically viable production processes and products.

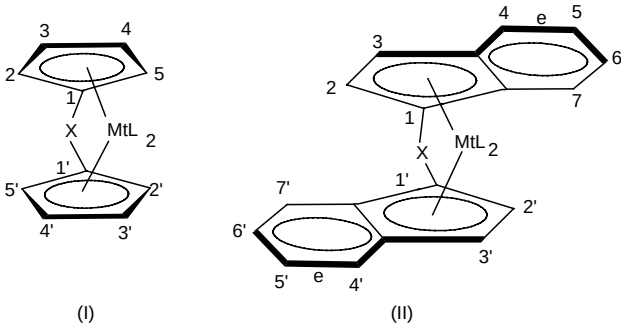
On the other hand, the research activity in a homogenous metallocene catalyst system was relatively dormant until the discovery of methylaluminoxane (MAO) cocatalyst by Sinn and Kaminsky in the late 1970s.[4] The combination of metallocene, such as Cp_2ZrCl_2, and MAO produces high-molecular weight PE with unprecedented catalytic activity (10^6 g PE/g Zr.hr.bar) that is much higher than that of the commercially used Ziegler–Natta catalysts. This historic discovery opened up another golden age in polyolefin research with intense competition across the world both in industrial and in academic laboratories.

In 1984, Ewen made the first stereospecific bridged metallocene catalyst (*rac*-ethylidenebis(indenyl)titanium dichloride/MAO) that produced isotactic PP (i-PP).[5] He also discovered *ansa*-ziroconium metallocene ($Me_2C(Cp\text{-}9\text{-}Flu)ZrCl_2$) for the production of highly syndiotactic polypropylene (s-PP) a couple of years later.[6] Around the same time, Kaminsky synthesized highly i-PP using a chiral *ansa*-zirconocene/MAO catalyst.[7] In the late 1980s, Dow discovered constrained geometry metallocene catalysts, which are titanium-based catalysts with monocyclopentadienyl and a donor ligand stabilizing the metal center. The resulting open active site allows the incorporation of higher α-olefins with comparative rates similar to that of ethylene. This brings in the significant advantage of metallocene catalysis to obtain polyolefin copolymers with a broad range of compositions and narrow composition and molecular weight distributions, which cannot be realized by heterogenous Ziegler–Natta catalysts. Since then, worldwide industrial and academic research in metallocene catalysis for olefin polymerization has been rapidly accelerating, and this advanced technology has been brought to commercial use.

METALLOCENE CATALYSTS AND POLYMERIZATION

Metallocene catalysts can be classified into several categories based on ligand symmetrical structures. This includes C_2, C_{2v}, C_s, C_1, oscillating structures, and bridged half-metallocenes. The main features of these

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007853



Scheme 1

catalysts and polymerization mechanisms, as well as their applications in olefin polymerization, will be discussed below.

C_2-Symmetric Metallocenes

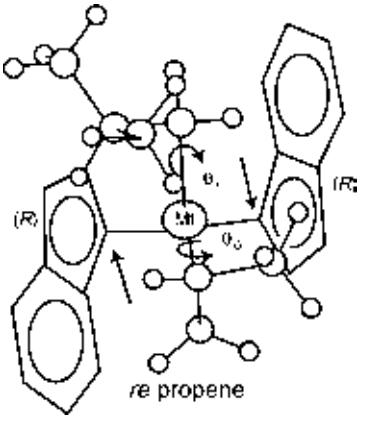
Chiral C_2-symmetric bridged metallocenes are the most successful highly enantioselective polymerization catalysts. The catalysts produce PPs with microstructures ranging from almost atactic to almost perfectly isotactic, and often contain a small amount of isolated regio-irregularities. The basic structures of these catalysts are shown in Scheme 1.

The bridge (X), such as $-Me_2Si-$, $-Me_2C-$, or $-CH_2-CH_2-$ connecting two cyclopentadienyl rings, prevents their rotation and locks them in a chiral configuration. In type (I), highly enantioselective catalysts can be obtained when preferably bulky substituents are present at position 3,3′ or 4,4′. In type (II), the role of such substituents is played by the phenyl of the 1-indenyl moieties and additional substitution. Particularly, position 4,4′ is highly beneficial. In both cases, ancillary substituents at position 2,2′ result in producing higher molecular weight polymers because of the reduction in β-H elimination of the propagating chain end.[8]

The stereoselectivity of catalysts during the α-olefin polymerization is well understood.[9] Scheme 2 shows the model for an active species representative of type (II), i.e., *rac*-$Me_2C(1\text{-Ind})_2ZrCl_2$. The catalytic complex is pseudotetrahedral and cationic, with an *iso*-butyl group simulating the growing polymer chain and a propylene molecule at the remaining coordination site. The aromatic ligand is in the (*R*,*R*) configuration. The growing polymer chain must adopt a conformation that minimizes the steric interaction with one of the two phenyl rings of the bis-indenyl moiety. The first C–C bond has to be bent to one side. This favors the 1,2-insertion of propylene with the enantioface that brings the methyl substituent *anti* to that C–C bond [*re*-face for the (*R*,*R*)-catalytic complex]. As a result, 2,1-insertion is always difficult owing to direct steric interactions of the CH_3 group with the aromatic rings. At the completion of each insertion step, the growing chain will always reside at the coordination site previously occupied by the monomer (chain-migratory insertion mechanism). However, this orderly coordination/insertion process may not always be the case, depending on the reaction condition (catalyst, temperature, etc.). There are some stereoerrors generated during the chain propagation. Generally, the C_2 symmetry ensures the equivalence of the two active sites; chain propagation is expected to be predominantly isotactic and site-controlled with occasional $\cdots$mmmmrrmmmm$\cdots$ stereodefects.

Table 1 shows the data of liquid propylene polymerization in the presence of some representative C_2-symmetric metallocene catalysts.[9] The flexibility of the bridge between two indenyl rings plays an important role in the microstructure and properties of the resulting polymer. Stereoselectivity and molecular weight increase in the order $Me_2C < Et < Me_2Si$. The introduction of an alkyl substituent of the 2 position in the *ansa*-bisindenyl zirconium complexes increases both the stereoregularity and the molecular weight of the produced PP. Combining substitutions in the 2- and 4- positions leads to some of the most successful isospecific zirconocene catalysts, which provide incredible catalyst activity, isotacticity, and molecular weight. These catalysts have also been successfully supported on a carrier, retaining their high activity at much lower Al/Zr ratios, and have met the requirements for industrial scale production of i-PP. In general, increasing temperature to a certain extent results in the decrease of stereoselectivity and molecular weight of the products, especially using *ansa*-metallocenes that contain flexible ethylene or silane bridge.

In general, C_2-symmetric metallocenes are not exceedingly regioselective in propylene polymerization.



Scheme 2

Table 1 Propylene polymerization with some representative *ansa*-zirconocene catalysts

Catalyst	Al/Zr	Temperature (°C)	Catalyst activity (kg PP/mol/hr)	mmmm (%)	2,1 (%)	T_m (°C)	M_v ($\times 10^3$)
rac-Et(1-Ind)$ZrCl_2$	8,000	50	140	87.4	0.6	134	33.6
rac-Et(1-Ind)$ZrCl_2$	8,000	70	252	83.4	0.7	125	19.6
rac-Me_2C(1-Ind)$ZrCl_2$	4,000	50	62	71.4	0.5	110	5.3
rac-Me_2C(1-Ind)$ZrCl_2$	8,000	70	145	76.7	0.6	124	5.6
rac-Me_2Si(1-Ind)$ZrCl_2$	3,000	50	17	90.3	0.5	144	56
rac-Me_2Si(1-Ind)$ZrCl_2$	nr	70	190	81.7	nr	137	36
rac-Me_2Si(2-Me-1-Ind)$ZrCl_2$	4,000	50	33	94.2			
rac-Me_2Si(2-Me-1-Ind)$ZrCl_2$	15,000	70	99	88.5	nr	145	195
rac-Me_2Si(2-Me-4-Ph-1-Ind)$ZrCl_2$	10,000	50	1,300	99.5	0.5		
rac-Me_2Si(2-Me-4-Ph-1-Ind)$ZrCl_2$	15,000	70	755	95.2	nr	157	729
rac-Me_2Si(2-Me-Benz[e]Ind)$ZrCl_2$	15,800	50	145	93	0.3	152	247.7
rac-Me_2Si(2-Me-Benz[e]Ind)$ZrCl_2$	15,000	70	403	88.7	nr	146	330

The resulting PP samples usually contain regioirregular enchainments in the range of 0.3–1.0 mol%. The end groups in the polymer chain are typical of saturated propyl groups and 2-methyl-prop-1-enyl(vinylidene) groups, owing to the chain transfer by intramolecular β-H elimination. Additional saturated end-groups, such as *iso*-butyl and unsaturated but-2-enyl, are also observed because of the chain transfer to $AlMe_3$ and β-H elimination, respectively, at a last inserted 2,1 unit.[10]

One major drawback of *rac*-C_2-symmetric *ansa*-metallocene catalysts is that either during the synthesis of the precursors or by subsequent epimerization the *meso*-form with C_s-symmetry can also be obtained. The active sites of catalytic species formed from *meso*-C_s-symmetric precursors are obviously nonchirotopic, and cannot exert any stereocontrol on the chain propagation. As a result, some atactic PP is invariably obtained.[11]

C_{2v}-Symmetric Metallocenes

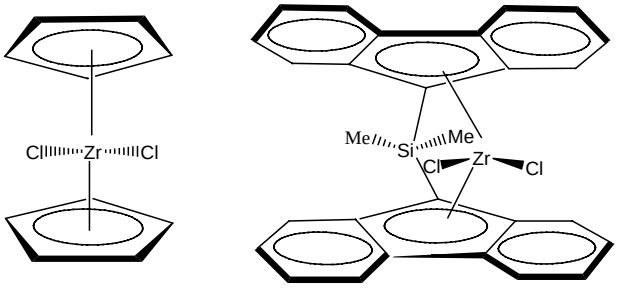
Two cyclopentadienyl-sandwiched Cp_2ZrX_2/MAO complexes (Cp = cyclopentadienyl; X = halogen or alkyl) with C_{2v}-symmetry form the earliest metallocene system that has been used for ethylene homopolymerization with remarkably high activity. The application in propylene polymerization is less useful. They produce PP with low activity and low molecular weight.[12] Unexpectedly, the Cp_2TiPh_2/MAO system can be used for the synthesis of predominately i-PP below room temperature. The stereochemical structure of the resulting polymer indicates a chain-end-controlled model in the polymerization. Bridged C_{2v}-symmetric metallocenes, such as $Me_2Si(9\text{-Flu})_2ZrCl_2$ (Flu = fluorenyl), provide convenient access to high molecular weight atactic PP, which performs like an elastomer.[13] The structure of two typical C_{2v}-symmetric metallocenes is shown in Scheme 3.

By means of ^{13}C and ^1H NMR, mainly propyl and 2-methyl-prop-1-enyl (vinylidene) end groups are detected in the polymers prepared by Cp_2MtX_2 (Mt = Zr, or Ti) catalyst. This indicates that the predominant monomer insertion is 1,2-insertion and that β-H elimination from the last inserted 1,2 unit is the major chain transfer pathway.[14] In PPs prepared with related catalyst systems $(Cp^*)_2ZrX_2$/MAO ($Cp^* = \eta^5$-pentamethylcyclopentadienyl), *iso*-butyl and prop-1-enyl (vinyl) groups are found to be the major end groups, which suggests that chain transfer consists of predominant β-methyl elimination.[15]

Bridged C_s-Symmetric Metallocenes

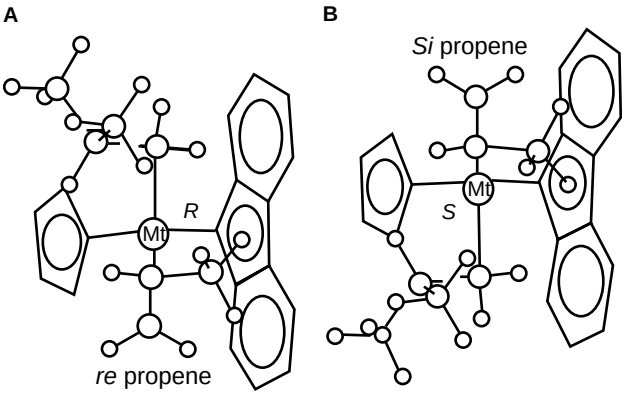
The first catalyst in producing highly s-PP is C_s-symmetric $Me_2C(Cp)(9\text{-Flu})ZrCl_2$ (Flu = fluorenyl), as illustrated in Scheme 4.[16] Although s-PP polymers in the experimental measurements always show lower melting points (T_m) than the corresponding i-PP polymers, the theoretical prediction of a fully s-PP will have T_m of 214°C, notably higher than that of i-PP (186°C).[17] It is a scientifically challenging and industrially important to develop new metallocene catalysts in order to prepare s-PP polymers with high syndiotacticity.

The invention of C_s-symmetric metallocenes has marked a major turning point in understanding the stereocontrol mechanism of metallocene catalysis. Scheme 5 shows the computer models of the cationic



Cp_2ZrCl_2 $Me_2Si(9\text{-}Flu)_2ZrCl_2$

Scheme 3



Scheme 5

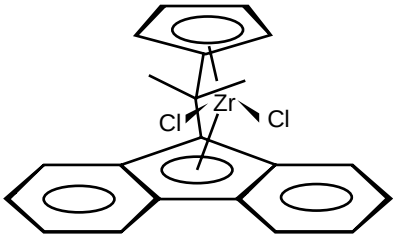
catalytic species $[Me_2C(Cp)(9\text{-}Flu)Mt(iso\text{-}butyl)]^+$ (Mt = Zr) with an η^2-coordinated propylene molecule.[18] The mutual arrangements of the first C–C bond of the *iso*-butyl group (simulating a growing PP chain) and of the monomer are the ones that minimize the nonbonded interactions at the 1,2-insertion step. According to the chain migratory polymerization mechanism, situations (1) and (2) (as given below) alternate regularly during chain propagation. In two successive insertion steps opposite propylene enantiofaces are inserted, therefore s-PP results. The presence of isolated insertion errors, showing rrrrmmrrrr stereodefects, indicates a site control mechanism.

At low polymerization temperatures with liquid propylene, $Me_2C(Cp)(9\text{-}Flu)ZrCl_2/MAO$ catalyst system can produce s-PP with a very high stereoregularity [(rrrr) > 0.95] and a melting point close to those of highly i-PP ($T_m > 150°C$).[19] Generally, C_s-symmetric metallocene catalysts are highly regioselective and no regioerrors are found in the typical s-PP.[20] The end groups are predominantly propyls and 2-methylprop-1-enyls (vinylidenes). At lower monomer concentration, small amounts of *iso*-butyls are found as a result of the chain transfer to $AlMe_3$.

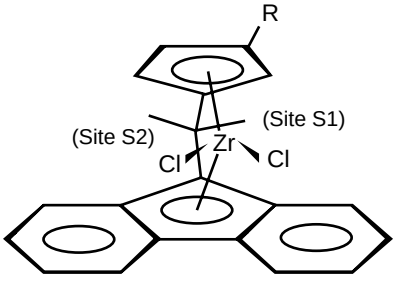
Bridged C_1-Symmetric Metallocenes

C_1-symmetric metallocenes are complexes lacking any symmetry, which can be divided into two categories: 1) those with one (substituted or nonsubstituted) cyclopentadienyl ligand having two homotopic faces (type I) and 2) those with two asymmetric cyclopentadienyls (type II), as illustrated in Scheme 6.

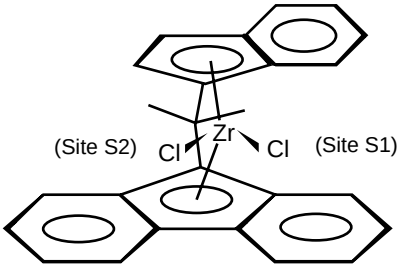
C_1-symmetric metallocenes present a synthetic advantage over C_2-symmetric ones in terms of their application for isospecific polymerization of propylene. As mentioned previously, a problem associated with the synthesis of *ansa*-C_2-symmetric metallocenes is that they are almost invariably generated along with their *meso*-isomers, which are difficult to remove from the catalyst mixture and often produce undesirable low molecular weight atactic PP with certain polymerization activity. The synthesis of pure C_2-symmetric catalysts usually requires multiple purification steps with low yields.[21,22] In contrast, a *meso*-form does not exist



Scheme 4



Type I



Type II

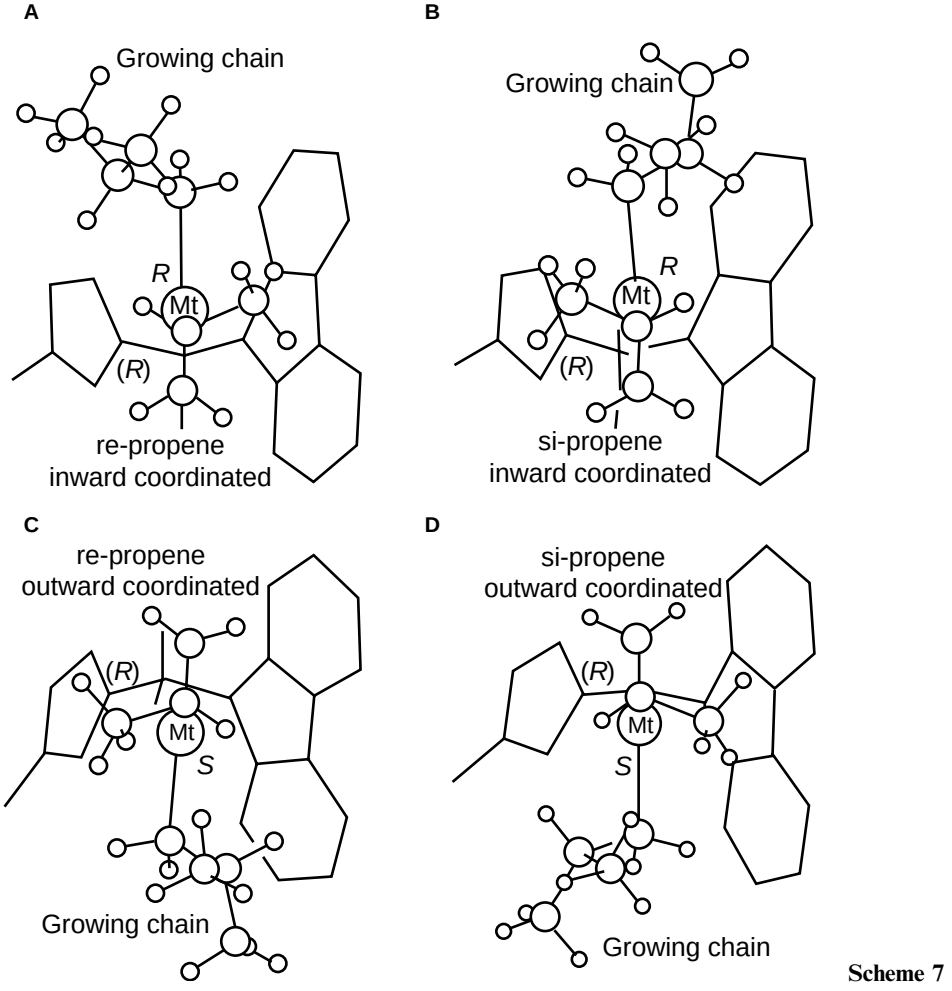
Scheme 6

in C_1-symmetric system, therefore the synthesis of ligand and catalyst is simple.

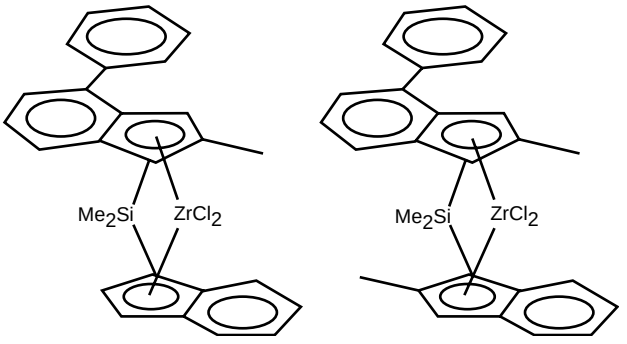
The common feature of C_1-symmetric metallocenes is that their two coordination sites are diastereotopic. On the size basis of the substitution on the cyclopentadienyl ligand, C_1-symmetric catalysts can vary in stereoselectivity to form PP that is hemi-isospecific (amorphous hi-PP) to partially isospecific (amorphous or low-crystallinity PP) to isospecific (i-PP with high T_m and high crystallinity).

The steroselectivity of C_1-symmetric metallocenes in propylene polymerization can be predicted based on rationalization of the ligand structure. In case of a typical type I metallocene, Me_2C(3-Me-Cp)(9-Flu)$ZrCl_2$, a computer model of the cationic catalytic species with growing polymer chain simulated by an *iso*-butyl group is shown in Scheme 7. Clearly, the model is enantioselective when the propylene molecule is inward coordinated, because structure A with a *re*-coordinated propylene is favored compared to structure B with a *si*-coordinated propylene. This is due to repulsive interactions between the growing chain and the fluorenyl ligand in structure B. On the contrary, the model is nonenantioselective when the growing chain is σ-bonded in the inward coordination position, because structures C and D are equally hampered by repulsive interactions of the growing chain, either with the 9-Flu ligand (structure C) or with the methyl group of the 3-MeCp ligand (structure D). According to a regular chain migration insertion mechanism, the insertions are alternately enantioselective and nanenantioselective. Therefore, the model is hemi-isospecific.[23] In fact, PPs with almost ideal hemi-isotactic structure can be obtained with Me_2C(3-Me-Cp)(9-Flu)$ZrCl_2$/MAO under conditions of chain migratory propagation (relatively low temperature and high propylene concentration).[24]

The stereoselectivity, however, strongly depends on the structure of the ligand. In particular, catalysts of type I with a bulky R substituent (e.g., R = *t*-butyl or trimethylsilyl), as well as those of type II, are predominantly isotactic selective.[25,26] The presence of *t*-butyl group or phenyl ring does not permit the growing chain to be located in the inward position. Particularly, in the absence of the monomer molecule, probably occurring at the end of each insertion step,



Scheme 7



Scheme 8

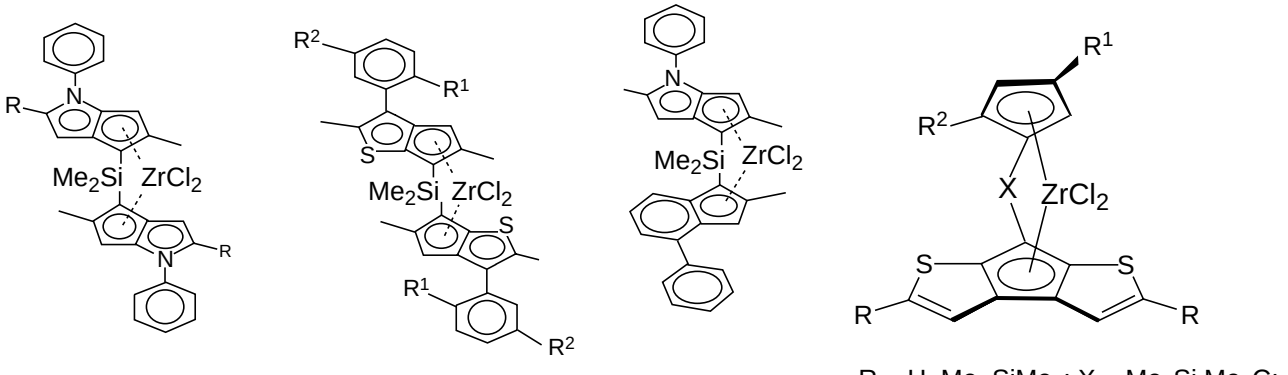
the steric stress of the ligand skeleton could force the growing chain to skip back to the less crowded outward position. Thus, insertion always occurs with the same relative disposition of the monomer and of the growing chain, and the model turns out to be isospecific. Experimental results show that Me_2C(3-*t*-Bu-Cp)(9-Flu)$ZrCl_2$/MAO produces i-PP with relatively high mmmm pentad contents from 77.5% at 60°C in liquid monomer to 87.8% at 50°C and pressure of 2 bar.[27,28] Spaleck reported the remarkable performance of C_1-symmetric systems, illustrated in Scheme 8, which combine the indenyl substitution of two different C_2-symmetric zirconocenes. These catalysts provide both high stereoregularity and high molecular weights. For example, Me_2Si(Ind)(2-Me-4-Ph-Ind)$ZrCl_2$ gives i-PP with mm = 96%, 2,1 = 0.4%, T_m = 155°C, and M_w = 530,000, at the relatively high polymerization temperature of 70°C.

Recently, Ewen et al. have designed C_2-, C_s-, and C_1-symmetric catalysts bearing heterocycle-condensed Cp ligands, as illustrated in Scheme 9, which add electronic effects to the stereocontrol.[29,30] The complexes, containing isopropylidene-bridged cyclopentadienyl and cyclopentyl thiophene ligands, show activity comparable to those of fluorenyl analogs, and obey the same symmetry rules for PP tacticity vs. catalyst symmetry. Such catalysts can produce i-PP with even higher melting temperature (161°C) and molecular weight ($M_w > 1$ million g/mol). However, when dithienocyclopentadiene is applied to substitute fluorenyl, the catalytic properties suffer greatly. Although low amounts of MAO are used to activate the metallocenes, only PPs with low isotacticity and melting point are produced.[31] Dialkoxyl substituted C_1-symmetric metallocene complexes, such as *rac*-[1-(5,6-dialkoxy-2-methyl-1-η^5-indenyl)-2-(9-η^5-fluorenyl)ethane]zirconium dichloride activated with triisobutylaluminum and $Ph_3C[B(C_6F_5)_4]$, also show few advantages compared with their counterparts without heteroatom substitutes in terms of catalytic activity, isotactity, molecular weight, and melting temperature of the products.

It should be noted that C_1-symmetric metallocenes show a stereoselectivity increase with elevated polymerization temperature and lowered monomer concentration, a behavior opposite to that displayed by the C_2-symmetric metallocenes.[32] The regioregularity of PP samples prepared with most C_1-symmetric metallocene catalysts is fairly high. The predominant monomer insertion mode is 1,2. Isolated 2,1 and/or 3,1 units can be observed with <0.5 mol%.

Metallocenes with Oscillating Structure

In unbridged metallocenes, like Cp_2MtX_2 or Ind_2MtX_2, the η^5-coordinated aromatic rings rotate freely even at very low temperatures. However, the rotation may become hindered in the presence of a bulky substituent, and the rate of rotation "oscillating" may be dependent on the temperature. Of special interest are (2-Ar-Ind)$_2$$MtX_2$ complexes (with Mt = Zr or Hf, and Ar = phenyl or a substituted



R = H or Me

$R^1 = R^2 = H$
R^1 = Me, R^2 = H
$R^1 = R^2$ = Me

R = H, Me, $SiMe_3$; X = Me_2Si,Me_2C;
R1 = Me, iPr, $SiMe_3$, Ph, tBu;
R2 = H, Me.

Scheme 9

aromatic moiety), which give rise to an equilibrium between two rotational isomers: one with quasi-C_2-symmetry (*rac*-like form, structure a) and the other with quasi-C_s-symmetry (*meso*-like form, structure b), as illustrated in Scheme 10.[33]

Polypropylenes prepared by (2-Ar-Ind)$_2$MtX$_2$/MAO catalysts contain both isotactic and atactic sequences, although with a complicated structure revealed by studies of solvent fractionation. These polymers perform as thermoplastic elastomers, involving atactic sequences imparting a moderately elastic response and isotactic sequences acting as physical cross-links.[34] The study on the regioregularity shows the predominant 1,2 propylene insertion with low amounts of detectable 2,1 insertion (0.2–0.3 mol%) only in a predominately isotactic environment.[35] This indicates that not only the stereoselectivity, but also the regioselectivity of the *rac*-like and *meso*-like forms of oscillating catalysts resemble that of the corresponding bridged bis-indenyl complexes.

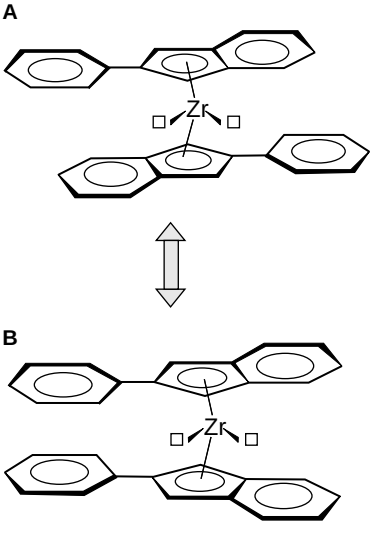
Bridged Half-Metallocenes

Many bridged and unbridged half-metallocenes have interesting catalytic behaviors. For example, CpTiCl$_3$/MAO is a highly active catalyst for syndiotactic polymerization of styrene.[36,37] However, such a catalyst cannot perform specific polymerization of propylene. Also, important half-metallocenes are constrained by the geometry of catalysts with the general formula Me$_2$Si(Me$_4$Cp)(NR)MtX$_2$(Mt = Ti or Zr, R = alkyl, X = halogen or alkyl), as illustrated in Scheme 11. This is well known as Dow ''Insite'' catalyst.[38] The absence of a second Cp ring and the short bridge results in a very open environment of the transition metal, allowing easier insertion of bulky monomers compared with bis-Cp systems. This special feature endows such catalysts with excellence performance in the copolymerizations of ethylene with higher α-olefins and even with styrene. In the case of propylene polymerization with such catalysts, the resulting atactic polymers are of poor regioregularity (with up to 5 mol% of 2,1 insertions).[39]

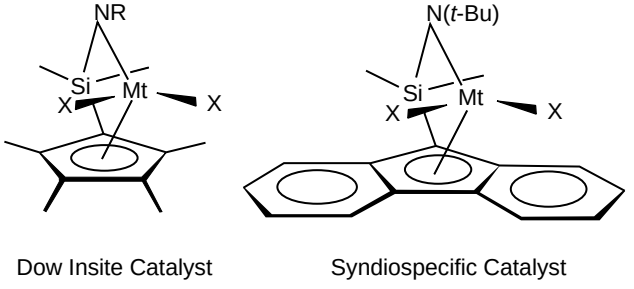
With the replacement of Cp by a 9-fluorenyl moiety, the metallocene promotes syndiotactic polymerization of propylene under site control. Syndiotactic PP with (rrrr) as high as 0.77 was obtained with Me$_2$Si(9-Flu)(N-*t*-Bu)ZrCl$_2$/MAO catalyst.[40] The stereoselectivity is due to the (pseudo)C_s-symmetry of the catalytic complex, and the stereocontrol mechanism is analog to that for the C_s-symmetric *ansa*-metallocenes.

Cocatalysts

Cocatalysts play an important role in the single-site polymerization of olefins. The relationship between the structure of cocatalysts and their activation in the metal-catalyzed olefin polymerization has been reviewed recently.[41] Methylaluminoxane is the most widely used cocatalyst, because it is able to activate a large number of metallocenes and other soluble transition complexes. It is usually synthesized by controlled hydrolysis of trimethylaluminum (TEA). The techniques for the preparation of MAO have been reviewed by Reddy and Sivaram.[42] It is generally believed that MAO consists of oligomers involving many [–O–Al(CH$_3$)–] units with the molecular weight of 1000–1500 g/mol. However, the precise chemical structure and composition of MAO are still not clear. A three-dimensional cage structure with four coordinate aluminum centers, resembling a half-open dodecahedron, was proposed by researchers.[43,44] The cage can seize an anion from the metallocene to form a stable AlX_4^- anion because of the delocalization of the electron over the whole cage. The utilities of MAO mainly include alkylation of halogenated metallocene complexes, serving as scavengers for moisture or other



Scheme 10 *(View this art in color at www.dekker.com.)*



Scheme 11

impurities, and stabilization of active species. Usually, a large excess of MAO (Al/Mt = 1000–50,000) is required to reach desirable activity. This is a major drawback of using MAO as a cocatalyst. In addition, conventional MAO has very low solubility in aliphatic solvents, as well as poor storage stability in solution. Therefore, commercial modified methylaluminoxane (MMAO) is prepared by controlled hydrolysis of a mixture consisting of TEA and triisobutylaluminum. This MMAO exhibits improved solution storage stability and solubility, and can be produced at a lower cost while providing good polymerization efficiency.

The other important cocatalyst is perfluoroaryl boranes, such as $B(C_6H_5)_3$, and highly Lewis acidic $B(C_6F_5)_3$, $NR_3H^+B(C_6F_5)_4^-$, and $Ph_3C^+B(C_6F_5)_4^-$, which can serve as substituents for MAO in combination with metallocene dialkyls.[45,46] A relatively small amount of boranes (borane/Mt = 1 : 1) is needed to form an effective catalyst for olefin polymerization. Unfortunately, this catalyst is rather sensitive to moisture and other impurities, and the stability of the resulting metallocene complexes is much lower than that formed with MAO. Usually, alklyaluminum (such as triisobutylaluminum) is added to the systems to scavenge the impurities and to stabilize the complexes.

Supported Metallocene Catalysts

As discussed, homogenous metallocene catalysts show special characteristics, such as a single active site, extremely high catalytic activity, capability of incorporating a high content of monomer, a narrow molecular weight distribution, and control of stereoregularity. However, the products prepared from homogenous metallocenes exhibit relatively poor apparent properties, compared with those from heterogenous Ziegler–Natta catalysts. They show irregular particle shape and low bulk density and contain a large amount of fine powder. Because such problems are the major concerns for the industrial application in slurry or gas-phase polymerization, immobilization of single-site metallocene on certain supports may improve the morphology of the final products and adaptation to the industrial process. Supported catalysts also show additional advantages in cost reduction due to lower amounts of MAO required and in giving products with higher molecular weight, higher melting temperature, and better stereoregularity. The carriers include inorganic supports, such as silica, Al_2O_3, $MgCl_2$, clay, zeolite, $CaCO_3$, and organic polymers, such as porous powdered high-density PE (HDPE), i-PP, nylon granules, and copolymer of styrene and divinylbenzene.[47–53] One of the great breakthroughs in heterogenous metallocene catalysts was made by investigators at Mitsubishi.[54] Acidic clays are formed by reacting $AlR_3$with clay, such as montmorillonite, hectorite, and mica, before contacting with metallocenes to form active metallocenes/AlR_3 complexes supported on clay. Overall, the AlR_3-treated clay minerals serve as both activators and carriers. The resulting catalysts show high activity comparable to those with MAO as cocatalyst, giving PE (co)polymers improved morphology and molecular weight distribution. This system has also been commercialized for propylene polymerization to form highly i-PP.

SYNTHESIS OF FUNCTIONAL POLYOLEFINS

The tunable metallocene catalyst with a well-defined polymerization mechanism provides distinctive advantages in the preparation of new polymers with well-controlled molecular structures, especially functional polyolefins that are very difficult to prepare by other methods.[55,56] Since the discovery of HDPE and i-PP about half a century ago, functionalization of polyolefin has been a scientifically challenging and industrially important area. The constant interest, despite lack of effective functionalization chemistry, is due to the strong desire to improve polyolefin's poor interactive properties. The hydrophobicity and low surface energy of polyolefin has limited its applications, especially in the areas of coating, blends, and composites, in which adhesion, comparability, dispersion, and paintability are paramount.

Polyolefin Containing Pendent Functional Groups

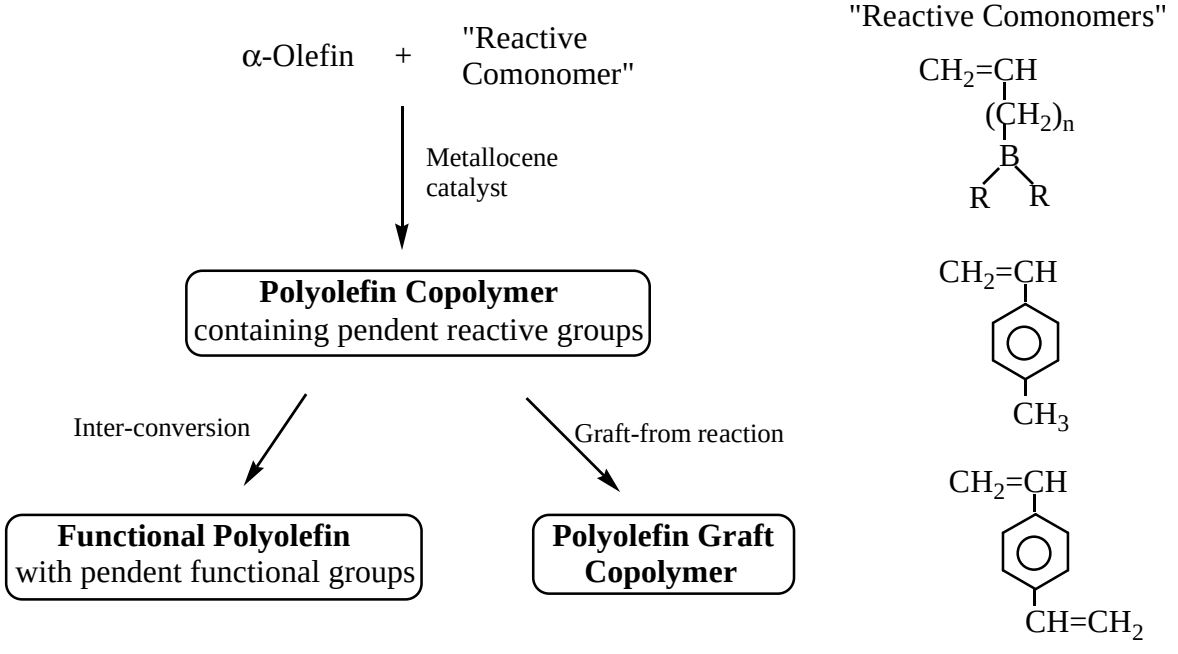
In the past decade, our group at Penn State has been focusing on a functionalization approach by the combination of metallocene catalysts and reactive comonomers. The chemistry takes the advantage of metallocene catalyst with a tunable single active site to prepare polyolefin copolymers with narrow molecular weight and composition distributions, high catalyst activities, and predictable tacticities and copolymer compositions.

As illustrated in Scheme 12, the metallocene-mediated copolymerization of α-olefin and reactive comonomer forms a copolymer containing several pendent reactive groups, and then serves as an ''intermediate'' for the transformation to functional polyolefins by various reaction mechanisms. In addition to the metallocene catalyst for effective copolymerization, the key factor in this approach is the design of a comonomer containing a reactive group that can simultaneously fulfill the following requirements. First, the reactive group must be stable to metallocene catalysts and soluble in hydrocarbon polymerization media.

Second, the reactive monomer should have good copolymerization reactivity with α-olefins. Third, the reactive group must be facile in the subsequent interconversion reaction to form polar groups under mild reaction conditions. It is particularly effective if the reactive group can serve as an initiator (with sufficient stability) of the polymerization of functional monomers. In other words, each incorporated reactive group can produce a functional polymer chain containing hundreds of functional groups. With the metallocene technology, the choice of reactive comonomers has dramatically increased.

Three reactive comonomers, including borane monomers, *p*-methylstyrene (*p*-MS), and divinylbenzene, have been incorporated in polyolefin copolymers with a broad compositional range and narrow molecular weight and composition distributions.[57–69] The resulting pendent reactive sites, i.e., borane, benzylic protons, and styrene units, located along the polyolefin backbone, are versatile in the subsequent transformation reactions, both in functionalization and graft reactions.[70–73] The transformation reactions occur selectively at the reactive sites. In other words, the concentration of functional groups is basically proportional to the concentration of the reactive sites, and the resulting functional polyolefin has a well-defined molecular structure.

One example is shown in Scheme 13, involving direct metallocene copolymerization of α-olefin and borane comonomer that is completely stable with the catalytic site, and can be effectively incorporated into polyolefin. The solubility of borane monomers and polymers in the reaction media (hexane and toluene) during metallocene polymerization is essential to ensure the optimum reaction condition for high polymer. In turn, the formed polyolefin contains reactive borane groups that can be selectively interconverted to the desirable functional groups (such as OH, NH_2, and halides) under mild reaction conditions.[55] This conversion can also be accomplished during the melt process (reactive extrusion). In addition, the borane side groups in polyolefin can then be selectively oxidized at the aliphatic C–B group for graft-form polymerization. The formed peroxyborane (B–O–O–C) initiates a controlled radical polymerization in the presence of free-radical-polymerizable monomers, such as methacrylates, vinyl acetate, acrylonitrile, etc., at ambient temperature. This controlled radical polymerization process minimizes the undesirable chain transfer reaction and termination (coupling and disproportional) reaction between the two growing chain ends, which result in the formation of homopolymers and cross-linked material. In most cases, the resulting graft copolymers are completely soluble and processible. The graft length (PMMA side chain) is controlled by the MMA concentration and reaction time. Some interesting graft polymers—including PE-*g*-PVA, PE-*g*-PMMA, PP-*g*-PMMA, PP-*g*-PMA, PP-*g*-PVA, and EP-*g*-PMMA—have been synthesized with controllable compositions and molecular microstructures.[56]



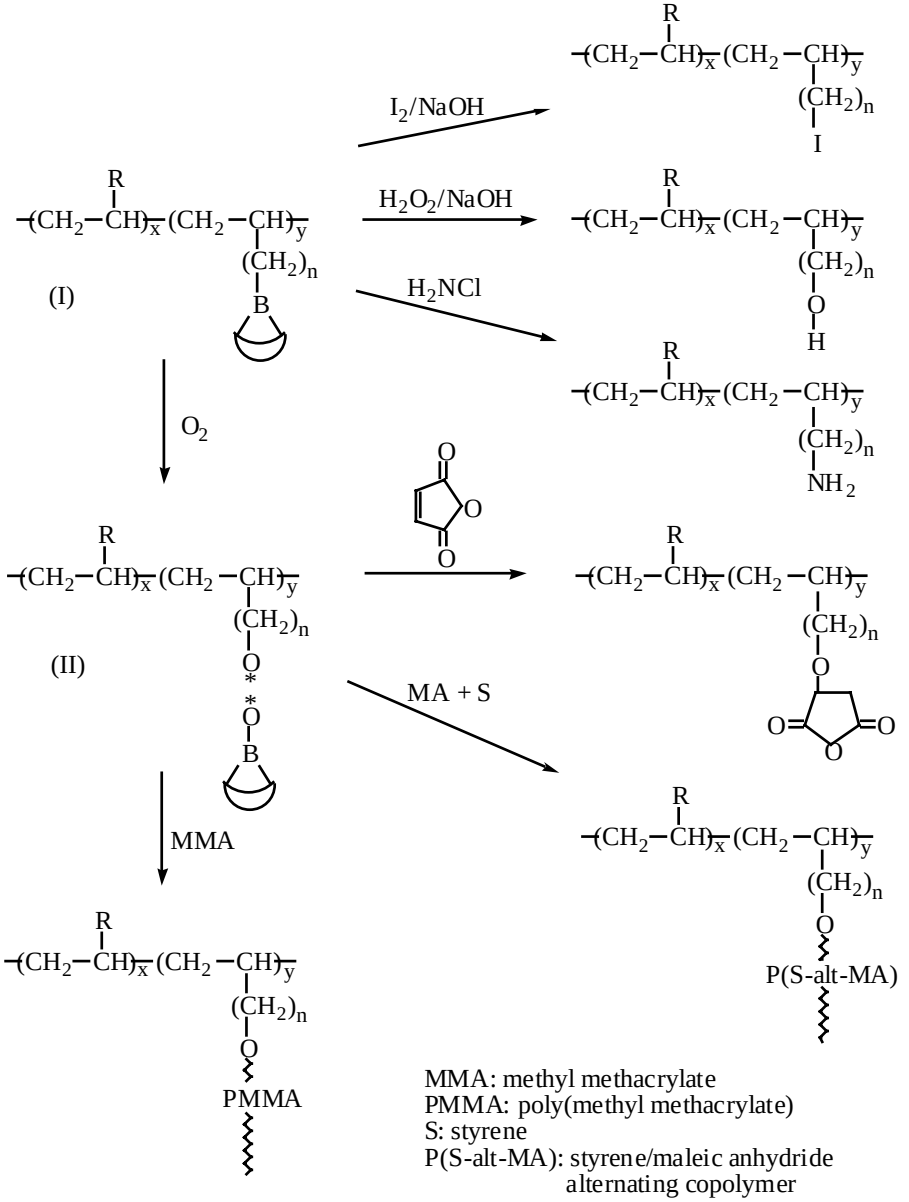
Scheme 12

Polyolefin Containing a Chain-End Functional Group

The chain-end functionalized polymer is a very attractive material that possesses an unperturbed polymer chain with desirable physical properties (such as melting temperature, crystallinity, glass transition temperature, melt flow, etc.) that are almost the same as those of the pure polymer.[74] Nevertheless, the terminal reactive group at the polymer chain end has good mobility and can provide a reactive site for many applications. This includes adhesion to the substrates, reactive blending, and formation of block copolymers.

The in situ chain transfer reaction by far is the most convenient and effective method for introducing a functional group to the polymer chain end.[56] With the combination of a well-defined metallocene reaction mechanism and suitable chain transfer agent (CT), it is possible to have a chain transfer reaction taking place with little change in the polymerization rate. Each polymer chain produced has a terminal functional group (the residue of the CT). Usually, the polymer's molecular weight is reversibly proportional to the CT/monomer ratio.

Several interesting CTs have been studied in the past decades with the major objective of introducing functional (polar) groups at the polyolefin chain end. In the past, aluminum-alkyl and zinc-alkyl were reported to act as coinitiator and CT in heterogenous Ziegler–Natta polymerization.[75] Recently, two hydride CTs, i.e., silane and borane containing Si–H and B–H moieties, respectively, have been successfully applied



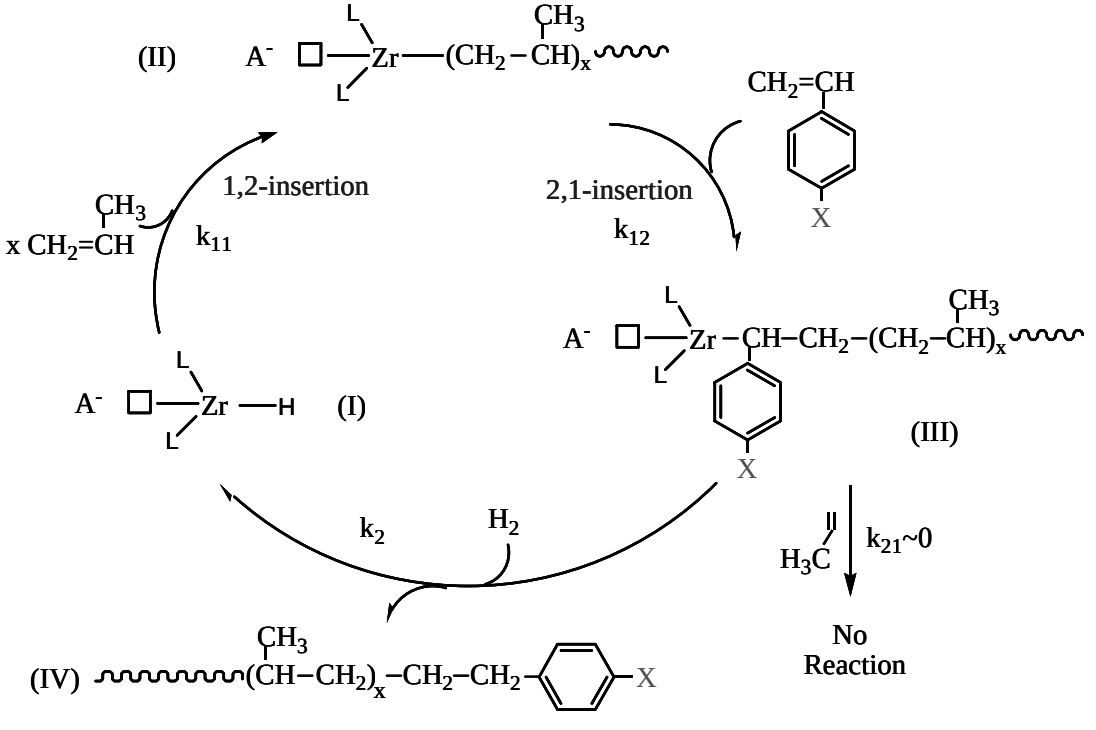
MMA: methyl methacrylate
PMMA: poly(methyl methacrylate)
S: styrene
P(S-alt-MA): styrene/maleic anhydride alternating copolymer

Scheme 13

in metallocene-mediated polymerization.[76–79] The chemistry is very general and applicable to most α-olefins and their mixtures. The resulting silane and borane terminated polyolefins, having relatively well-defined molecular structures, were used as intermediates in the preparation of more elaborate polymer structures. It is intriguing that the versatility of metallocene catalysis also allows the chain transfer reaction to extend to styrenic and allylic molecules, which usually serve as monomers.[80–84] The new route allows the preparation of chain end functional PP containing a terminal functional group, such as OH and NH_2, in one-pot polymerization process.

One example is illustrated in Scheme 14, which involves the combination of styrenic molecule (St-X) and hydrogen as the CT during metallocene-mediated propylene polymerization. Ironically, the research stemmed from a failure in the copolymerization of propylene and *p*-MS using *rac*-$SiMe_2$[2-Me-4-Ph(Ind)]$_2$ $ZrCl_2$/MAO complex. The catalyst was deactivated after 2,1-insertion of *p*-MS unit as a result of steric jamming, which forms a propagating site (III) with unfavorable 1,2-insertion of propylene (k_p) and lack of *p*-MS homopolymerization (k_{22} reaction). However, the sluggish propagating chain end (III) allows the reaction with hydrogen, which not only recovers the catalytic site (I) but also produces PP polymer with a terminal styrenic group (IV). The regenerated catalytic site (I) reinitiates the polymerization of propylene and continues the polymerization cycles. In other words, the ideal chain transfer reaction will not significantly affect the rate of polymerization, but will reduce the molecular weight of the resulting polymer. The molecular weight of chain-end functionalized PP will be linearly proportional to the molar ratio of [propylene]/[St-X], and independent of the [propylene]/[hydrogen] ration.

Table 2 summarizes a systematic study of propylene polymerization by using a *rac*-Me_2Si[2-Me-4-Ph(Ind)]$_2$ $ZrCl_2$/MAO catalyst in the presence of *p*-MS/hydrogen, which results in *p*-MS terminated PP (PP-t-*p*-MS).[81] The in situ chain transfer reaction is evidenced by its comparison with the control reactions that were carried out under similar reaction conditions—without hydrogen and/or *p*-MS. A small amount of *p*-MS (control 2) effectively stops the polymerization of propylene. The introduction of hydrogen restores the catalyst activity, as shown in run 4, which exhibits about 85% of the catalyst activity of control 1 (without CTs). Hydrogen is clearly needed to complete the chain transfer cycle during the polymerization. Comparing runs 4, 8, and 11 by altering the *p*-MS concentration, the higher the concentration of the *p*-MS the lower the molecular weight of the resulting



(X: H, CH_3, Cl, CH_2–OH, CH_2–CH_2–NH_2)

Scheme 14 *(View this art in color at www.dekker.com.)*

Table 2 A summary of PP-t-*p*-MS polymers prepared[a] by the combination of a *rac*-Me_2Si[2-Me-4-Ph(Ind)]$_2ZrCl_2$/MAO catalyst and *p*-MS/hydrogen CTs

Run	p-MS (M)	H_2 (psi)	Yield (g)	Catalyst activity (kg PP/mol catalyst/hr)	*p*-MS in PP (mol%)	*p*-MS, conversion (%)	M_n ($\times 10^{-3}$)	PDI (M_w/M_n)
Control 1	0	0	26.94	86,208	0	—	77,600	2.9
Control 2	0.0305	0	0.051	163	0.16	0.05	59,700	3.4
1	0.0305	2	3.80	12,160	0.14	8.30	55,500	2.4
2	0.0305	6	8.04	25,728	0.15	18.83	54,800	2.5
3	0.0305	12	12.04	38,528	0.15	28.19	55,400	2.3
4	0.0305	35	24.67	78,944	0.13	50.05	34,600	2.8
Control 3	0.076	0	~0	—	—	—	—	—
5	0.076	6	0.91	2,912	0.40	2.33	27,600	2.1
6	0.076	12	1.69	5,408	0.41	4.33	25,900	2.3
7	0.076	20	8.81	28,192	0.43	23.65	20,500	2.3
8	0.076	35	10.52	33,664	0.41	26.86	25,800	2.3
Control 4	0.153	0	~0	—	—	—	—	—
9	0.153	12	0.35	1,120	0.66	0.72	10,000	2.0
10	0.153	20	3.81	12,192	0.61	7.26	11,700	2.0
11	0.153	35	4.41	14,112	0.63	8.67	9,700	1.9

[a]Reaction conditions: 50 ml toluene, propylene (100 psi), [Zr] = 1.25×10^{-6} mol/L, [MAO]/[Zr] = 3000, temperature = 30°C, time = 15 min.

polymer. A polymer with very low molecular weight (couple of thousands) is obtained, and the molecular weight distribution is generally narrow, consistent with the well-defined single-site polymerization processes. The catalyst activity was also proportionally depressed with the concentration of *p*-MS, reflecting the competitive coordination at metallocene active sites between monomer and CTs. Three comparative reaction sets (runs 1–4, 5–8, and 9–11) were conducted under the same reaction conditions except for varying the hydrogen pressure. The change of hydrogen concentration does not significantly affect the polymer molecular weight and molecular weight distribution, but has a profound effect on the catalyst activity. Therefore, hydrogen does not engage in the initial chain transfer reaction, but rather assists in the completion of the reaction cycle. A sufficient quantity of hydrogen, proportional to the *p*-MS concentration, is needed to maintain high catalyst activity and *p*-MS conversion.

It is desirable to directly prepare PP with a desirable terminal functional group, such as Cl, OH, and NH_2. In other words, the ideal reaction is a one-pot in situ polymerization process, and no chain-end functionalization would be needed after the polymerization reaction. Three functional St-X molecules were investigated (83), including *p*-chlorostyrene (St-Cl), silane-protected *p*-vinylphenol (St-OSi), and silane-protected *p*-ethylaminostyrene (St-NSi_2), as illustrated in Scheme 14.[83] No external protection agent is needed for St-Cl in metallocene/MAO systems. However, both the OH and the NH_2 groups are very sensitive to the metallocene cationic site. The silane groups not only provide effective protection for both the OH and the NH_2 functional groups during the metallocene catalysis, but also can be completely deprotected by aqueous HCl solution during the sample workup procedure. The overall reaction benefits, especially, from the very small quantity of the CT needed in the preparation of high polymers. Therefore, the additional protection–deprotection step causes almost no change in the polymerization conditions and procedures. The polymer molecular weight is linearly proportional to the molar ratio of [propylene]/[St-X] in all three, St-Cl/H_2, St-OSi/H_2, and St-NSi_2/H_2, cases.[83] The degree of polymerization (X_n) follows a simple comparative equation $X_n = k_p[\text{olefin}]/k_{tr}[p\text{-MS}]$ with a chain transfer constant k_{tr}/k_p of 1/21 for St-Cl/H_2, 1/48 for St-OSi/H_2, and 1/34 for St-NSi_2/H_2, respectively.

CONCLUSIONS

This entry discusses metallocene catalysts and polymerization, which can be classified into several

categories based on ligand geometries, including C_{2v}-, C_2-, C_s-, and C_1-symmetry, oscillating structure, and bridged half-metallocenes, and the resulting polyolefins. The important features in metallocene-mediated α-olefin polymerization that have been stimulating the rapid development of this technology for commercialization include: 1) a tunable single catalytic site to produce a well-defined polymer with extremely high catalyst activity and predictable microstructure (i.e., tacticity, molecular weight, terminal group, etc.); 2) excellent ability of incorporating higher α-olefins and cycloolefins; 3) production of polyolefin copolymers with narrow molecular weight and composition distributions; and 4) a well-controlled polymerization mechanism allowing the incorporation of reactive (functional) groups in polyolefins. With the appropriate choice of metallocene catalysts and reactive comonomers and CTs, a broad range of functional polyolefins have been prepared with well-controlled molecular structures, containing functional groups (OH, NH_2, etc.) in the side chains and/or at the chain end, narrow molecular weight distribution ($M_w/M_n \sim 2$), and predictable composition and molecular weight.

ACKNOWLEDGMENT

The author would like to thank the Office of Naval Research for financial support.

REFERENCES

1. Wilkinson, G.; Briminham, I.M. Bis-cyclopentadienyl compounds of Ti, Zr, V, Nb and Ta. J. Am. Chem. Soc. **1954**, *76*, 4281.
2. Breslow, D.S.; Newburg, N.R. Bis-(cyclopentadienyl)-titanium dichloride–alkylaluminum complexes as catalysts for the polymerization of ethylene. J. Am. Chem. Soc. **1957**, *79*, 5072.
3. Natta, G.; Pino, P.; Corradini, P.; Danusso, F.; Namtica, E.; Mazzanti, G.; Moraglio, G. Crystalline high polymers of α-olefins. J. Am. Chem. Soc. **1955**, *77*, 1708.
4. Sinn, H.; Kaminsky, W. Ziegler-Natta catalysis. Adv. Organomet. Chem. **1980**, *18*, 99.
5. Ewen, J.A. Mechanisms of stereochemical control in propylene polymerizations with soluble Group 4B metallocene/methylalumoxane catalysts. J. Am. Chem. Soc. **1984**, *106*, 6355.
6. Ewen, J.A.; Jones, R.L.; Razavi, A.; Ferrara, J.D. Syndiospecific propylene polymerizations with Group IVB metallocenes. J. Am. Chem. Soc. **1988**, *110*, 6255.
7. Kaminsky, W.; Kulper, K.; Brintzinger, H.H.; Wild, F. Polymerization of propene and butane with a chiral zirconocene and MAO as co-catalyst. Angew. Chem. Int. Ed. Engl. **1985**, *24*, 507.
8. Busico, V.; Cipullo, R. Microstructure of polypropylene. Prog. Polym. Sci. **2001**, *26*, 443.
9. Resconi, L.; Cavallo, L.; Fait, A.; Piemontesi, F. Selectivity in propene polymerization with metallocene catalysts. Chem. Rev. **2000**, *100*, 1253.
10. Brintzinger, H.H.; Fischer, D.; Mulhaupt, R.; Rieger, B.; Waymouth, R.M. Stereospecific olefin polymerization with chiral metallocene catalysts. Angew. Chem. Int. Ed. Engl. **1995**, *34*, 1143.
11. Kaminsky, W. New polymers by metallocene catalysis. Macromol. Chem. Phys. **1996**, *197*, 3907.
12. Kulper, K.; Niedoba, S. Olefin polymerization with highly active soluble zirconinum compounds using MAO as co-catalyst. Makromol. Chem. Macromol. Symp. **1986**, *3*, 377.
13. Resconi, L.; Jones, R.L.; Rheingold, A.L.; Yap, G. High-molecular-weight atactic polypropylene from metallocene catalysts. 1. $Me_2Si(9\text{-}Flu)_2ZrX_2$ (X = Cl, Me). Organometallics **1996**, *15*, 998.
14. Tsuisui, T.; Mizuno, A.L.; Kashiwa, N. The Microstructure of propylene homo- and copolymers obtained with a Cp_2ZrCl_2 and MAO catalyst system. Polymer **1989**, *30*, 428.
15. Resconi, L.; Piemontesi, F.; Franciscono, G.; Abis, L.; Fiorani, T.J. Olefin polymerization at bis(pentamethylcyclopentadienyl)zirconium and -hafnium centers: chain-transfer mechanisms. J. Am. Chem. Soc. **1992**, *114*, 1025.
16. Ewen, J.A.; Jones, R.L.; Razavi, A.; Ferrara, J.D. Syndiospecific propylene polymerizations with Group IVB metallocenes. J. Am. Chem. Soc. **1988**, *110*, 6255.
17. Balbontin, G.; Dainelli, D.; Galimberti, M.; Paganetto, G. Thermal behavior of highly stereoregular syndiotactic polypropylene from homogeneous catalysts. Makromol. Chem. **1992**, *193*, 693.
18. Cavallo, L.; Guerra, G.; Vacatello, M.; Corradini, P. A possible model for the stereospecificity in the syndiospecific polymerization of propene with group 4a metallocenes. Macromolecules **1991**, *24*, 1784.
19. Longo, P.; Proto, A.; Grassi, A.; Ammendola, P. Stereospecific polymerization of propylene in the presence of homogeneous catalysts: ligand-monomer enantioselective interactions. Macromolecules **1991**, *24*, 4624.
20. Gureea, G.; Longo, P.; Cavallo, L.; Corradini, P.; Resconi, L. Relationship between regiospecificity and type of stereospecificity in propene polymerization with zirconocene-based catalysts. J. Am. Chem. Soc. **1997**, *119*, 4394.

21. Speleck, W.; Kuber, F.; Winter, A.; Rohrmann, J.; Bachmann, B.; Antberg, M.; Dolle, V.; Paulus, E. The influence of aromatic substituents on the polymerization behavior of bridged zirconocene catalysts. Organometallics **1994**, *13*, 954.
22. Stehling, U.; Diehold, J.; Kirsten, R.; Roll, W.; Brintzinger, H.H.; Jungling, S.; Mulhaupt, R.; Langhauser, F. *Ansa*-zirconocene polymerization catalysts with anelated ring ligands—effects on catalytic activity and polymer chain length. Organometallics **1994**, *13*, 964.
23. Guerra, G.; Cavallo, L.; Moscardi, G.; Vacatello, M.; Corradini, P. Back-skip of the growing chain at model complexes for the metallocene polymerization catalysis. Macromolecules **1996**, *29*, 4834.
24. Ewen, J.A.; Elder, M.J.; Jones, R.L.; Haspeslagh, L.; Atwood, J.L.; Bott, S.G.; Robinson, K. Metallocene/polypropylene structural relationships: implications on polymerization and stereochemical control mechanisms. Macromol. Symp. **1991**, *48*, 253.
25. Razavi, A.; Peter, L.; Nafpliotis, L.; Vereecke, D.; Den Dauw, K.; Atwood, J.L.; Thewald, U. The geometry of the site and its relevance for chain migration and stereospecificity. Makromol. Chem. Macromol. Symp. **1995**, *89*, 345.
26. Rieger, B.; Jany, G.; Fawzi, R.; Steimann, M. Unsymmetric *ansa*-zirconocene complexes with chiral ethylene bridges: influence of bridge conformation and monomer concentration on the stereoselectivity of the propene polymerization reaction. Organometallics **1994**, *13*, 647.
27. Razavi, A.; Vereecke, D.; Peters, L.; Den Dauw, K.; Nafpliotis, L.; Atwood, J.L. *Ziegler Catalysts*; Springer-Verlag: Berlin, 1995.
28. Kleinschmidt, R.; Reffke, M.; Fink, G. Investigation of the microstructure of polypropylene in dependence of the polymerization temperature for the systems i-Pr[3-RCpFlu]$ZrCl_2$/MAO and i-Pr{IndFlu}$ZrCl_2$/MAO. Macromol. Rapid Commun. **1999**, *20*, 284.
29. Ewen, J.A.; Jones, R.L.; Elder, M.J.; Rheingold, A.L.; Liable-Sands, L.M. Polymerization catalysts with cyclopentadienyl ligands ring-fused to pyrrole and thiophene heterocycles. J. Am. Chem. Soc. **1998**, *120*, 10,786.
30. Ewen, J.A.; Jones, R.L.; Elder, M.J.; Rheingold, A.L.; Liable-Sands, L.M.; Sommer, R.D. Chiral *ansa*-metallocenes with Cp ring-fused to thiophenes and pyrroles: syntheses, crystal structures, and isotactic polypropylene catalysts. J. Am. Chem. Soc. **2001**, *123*, 4763.
31. Nifant'ev, I.E.; Laishevtsev, I.; Ivchenko, P.V.; Kashulin, I.A.; Guidotti, S.; Piemontesi, F.; Camurati, I.; Resconi, L.; Klusener, P.A.; Rijsemus, J.H.; de Kloe, K.P.; Korndorffer, F.M. Macromol. Chem. Phys. **2004**, *205*, 2275.
32. Busico, V.; Cippullo, R.; Talarico, G. New evidence on the nature of the active sites in heterogeneous Ziegler-Natta catalysts for propene polymerization. Macromolecules **1997**, *30*, 4786.
33. Witte, P.; Lal, J.K.; Waymouth, R.M. Synthesis of unbridged bis(2-R-indenyl)zirconocenes containing functional groups and investigations in propylene polymerization. Organometallics **1999**, *18*, 4147.
34. Hu, Y.; Krejchi, M.T.; Shah, C.D.; Myers, C.L.; Waymouth, R.M. Elastomeric polypropylenes from unbridged (2-phenylindene)zirconocene catalysts: thermal characterization and mechanical properties. Macromolecules **1998**, *31*, 6908.
35. Lin, S.; Waymouth, R.M. Regioirregular propene insertion in polypropenes synthesized with unbridged bis(2-aryl)indenyl zirconium dichloride catalysts: implications on activity. Macromolecules **1999**, *32*, 8283.
36. Ishihara, N.; Kuramoto, M.; Uoi, M. Stereospecific polymerization of styrene giving the syndiotactic polymer. Macromolecules **1988**, *21*, 3356.
37. Yokata, K.; Inoue, T.; Naganuma, S.; Shozaki, H.; Tomotsu, N.; Kuramoto, M.; Ishihara, N. *In Metalorganic Catalysts for Synthesis and Polymerization*; Springer-Verlag: Berlin, 1999; 435–445.
38. Stevens, J.C.; Timmers, F.J.; Rosen, G.W.; Lai, S.Y. Constrained Geometry Addition Polymerization Catalysts, Processes for Their Preparation, Precursors Therefore, Methods of Use, and Novel Polymers Formed Therewith. European Patent Application EP 0416815 A2, Mar 13, 1991.
39. Mcknight, A.L.; Masood, M.; Waymouth, R.M.; Straus, D.A. Selectivity in propylene polymerization with group 4 Cp-amido catalysts. Organometallics **1997**, *16*, 2879.
40. Shiomura, T.; Asakura, T.; Inoue, N. Inversion of stereoselectivity in a metallocene catalyst. Macromol. Rapid Commun. **1996**, *17*, 9.
41. Chen, E.Y.; Marks, T.J. Cocatalysts for metal-catalyzed olefin polymerization: activators, activation processes, and structure-activity relationships. Chem. Rev. **2000**, *100*, 1391.
42. Reddy, S.S.; Sivaram, S. Homogeneous metallocene-MAO catalyst systems for ethyleme polymerization. Prog. Polym. Sci. **1995**, *20*, 309.
43. Sinn, H. Proposals for structure and effect of MAO based on mass balances and phase separation experiments. Macromol. Symp. **1995**, *97*, 27.
44. Koide, Y.; Bott, S.G.; Barron, A.R. Alumoxanes as cocatalysts in the palladium-catalyzed copolymerization of carbon monoxide and ethylene: genesis

of a structure-activity relationship. Organometallics **1996**, *15*, 2213.
45. Yang, X.; Stern, C.L.; Marks, T.J. Cationic zirconocene olefin polymerization catalysts based on the organo-Lewis acid tris(pentafluorophenyl)-borane. A synthetic, structural, solution dynamic, and polymerization catalytic study. J. Am. Chem. Soc. **1994**, *116*, 10,015.
46. Bochmann, M.; Lancaster, S.J. Base-free cationic zirconium benzyl complexes as highly active polymerization catalysts. Organometallics **1992**, *12*, 633.
47. Takahashi, T.; Yamamoto, K.; Hirakawa, K. Powder Catalyst Composition and Process for Polymerizing Olefins with the Use Thereof. U.S. Patent 5,474,962, Dec 12, 1995.
48. Kaminaka, M.; Soga, K. Polymerization of propene with catalyst systems composed of Al_2O_3 or $MgCl_2$-supported zirconocene and $Al(CH_3)_3$. Polymer **1992**, *33*, 1105.
49. Soga, K.; Arai, T.; Nozawa, H.; Uozumi, T. Recent development in heterogeneous metallocene catalysts. Macromol. Symp. **1995**, *97*, 53.
50. Woo, S.; Ko, Y.; Han, T. Polymerization of ethylene over metallocene confined inside the supercage of a NaY zeolite. Macromol. Rapid Commun. **1995**, *16*, 489.
51. Lee, D.; Yoon, K. Metallocene/MAO polymerization catalysts supported on cyclodextrin. Macromol. Symp. **1995**, *97*, 185.
52. Herrmann, H.F.; Bachmann, B.; Hierholzer, B.; Spaleck, W. Catalyst for the Polymerization of Olefins, Process for Its Preparation and Its Use. U.S. Patent 5,942,586, Aug 24, 1999.
53. Sangokoya, S.A. Heterogeneous Methylaluminoxane Catalyst System. U.S. Patent 5,308,815, May 3, 1994.
54. Sugano, T.; Iwama, N.; Isobe, E.; Suzuki, T.; Maruyama, Y. Catalysts for Polymerization of Alpha-Olefins, Process for Producing Alpha-Olefin Polymers, Novel Transition Metal Compounds and Catalyst Components for Polymerization of Alpha-Olefin. U.S. Patent Application 20020045535, Apr 18, 2002.
55. Chung, T.C. *Functionalization of Polyolefins*; Academic Press: London, 2002.
56. Chung, T.C. Synthesis of functional polyolefin copolymers with graft and block structures. Prog. Polym. Sci. **2002**, *27*, 39.
57. Chung, T.C. Method for Preparing Functional α-Olefin Polymers and Copolymers. U.S. Patent 4,734,472, Mar 29, 1988.
58. Chung, T.C. Synthesis of polyalcohols via Ziegler-Natta polymerization. Macromolecules **1988**, *21*, 865.
59. Chung, T.C.; Rhubright, D. Kinetic aspects of the copolymerization between α-olefins and borane monomers in Ziegler-Natta catalyst. Macromolecules **1993**, *26*, 3019.
60. Chung, T.C.; Lu, H.L.; Li, C.L. Synthesis and functionalization of unsaturated polyethylene; poly(ethylene-co-1,4-hexadiene). Macromolecules **1994**, *27*, 7533.
61. Chung, T.C.; Lu, H.L.; Li, C.L. Functionalization of polyethylene using borane reagents and metallocene catalysts. Polym. Int. **1995**, *37*, 197.
62. Chung, T.C.; Lu, H.L. Alpha-Olefin/Para-Alkylstyrene Copolymers and Functionalized Copolymers Thereof. U.S. Patent 5,543,484, Aug 6, 1996.
63. Chung, T.C.; Lu, H.L. Synthesis of poly(ethylene-co-p-methylstyrene) copolymers by metallocene catalysts with constrained ligand geometry. J. Polym. Sci. Pt. A: Polym. Chem. **1997**, *35*, 575.
64. Chung, T.C.; Lu, H.L. Kinetic and microstructure studies of poly(ethylene-co-p-methylstyrene) copolymers prepared by metallocene catalysts with constrained ligand geometry. J. Polym. Sci. Pt. A: Polym. Chem. **1998**, *36*, 1017.
65. Lu, H.L.; Hong, S.; Chung, T.C. Synthesis of new polyolefin elastomers, poly(ethylene-ter-propylene-ter-p-methylstyrene) and poly(ethylene-ter-1-octene-ter-p-methylstyrene). Macromolecules **1998**, *31*, 2028.
66. Lu, H.L.; Hong, S.; Chung, T.C. Synthesis of poly(propylene-co-p-methylstyrene) copolymers and functionalization. J. Polym. Sci. Pt. A: Polym. Chem. **1999**, *37*, 2795.
67. Chung, T.C.; Dong, J.Y. A Process of Preparing Linear Copolymers of Alpha-Olefins and Divinylbenzene with Narrow Molecular Weight and Composition Distributions. U.S. Patent 6,096,849, Aug 1, 2000.
68. Chung, T.C.; Dong, J.Y. Synthesis of linear ethylene/divinylbenzene copolymers by metallocene catalysts. Macromolecules **2002**, *35*, 2868.
69. Dong, J.Y.; Hong, H.; Chung, T.C.; Wang, H.C.; Datta, S. Synthesis of linear polyolefin elastomers containing divinylbenzene units and applications in crosslinking, functionalization, and graft reactions. Macromolecules **2003**, *36*, 6000.
70. Chung, T.C.; Janvikul, W.; Bernard, R.; Jiang, G.J. Synthesis of ethylene-propylene rubber graft copolymers by borane approach. Macromolecules **1994**, *27*, 26.
71. Lu, B.; Chung, T.C. Maleic anhydride modified PP with controllable molecular structure; new synthetic route via borane teriminated PP. Macromolecules **1998**, *31*, 5943.
72. Chung, T.C.; Lu, H.L.; Ding, R.D. Synthesis of polyethylene-g-polystyrene and polyethylene-g-poly(p-methylstyrene) copolymers. Macromolecules **1997**, *30*, 1272.

73. Lu, B.; Chung, T.C. New maleic anhydride modified polypropylene copolymers with block structure; synthesis and application in PP/nylon blends. Macromolecules **1999**, *32*, 2525.
74. Wang, Z.M.; Nakajima, H.; Manias, E.; Chung, T.C. Exfoliated PP/clay nanocomposites using ammonium-terminated PP as the organic modification montmorillonite. Macromolecules **2003**, *36*, 8919.
75. Shiono, T.; Soga, K. Synthesis of terminally aluminum-functionalized polypropylene. Macromolecules **1992**, *25*, 3356.
76. Fu, P.F.; Marks, T.J. Silanes as chain transfer agents in metallocene-mediated olefin polymerization. Facile in situ catalytic synthesis of silyl-terminated polyolefins. J. Am. Chem. Soc. **1995**, *117*, 10,747.
77. Koo, K.; Marks, T.J. Silanolytic chain transfer in Ziegler-Natta catalysis. Organotitanium-mediated formation of new silapolyolefins and polyolefin architectures. J. Am. Chem. Soc. **1998**, *120*, 4019.
78. Xu, G.; Chung, T.C. Borane chain transfer agent in metallocene-mediated olefin polymerization; synthesis of borane-terminated polyethylene and diblock copolymers containing polyethylene and polar polymer. J. Am. Chem. Soc. **1999**, *121*, 6763.
79. Xu, G.; Chung, T.C. Synthesis of syndiotactic polystyrene (s-PS) containing a terminal polar group and diblock copolymers containing s-PS and polar polymers. Macromolecules **1999**, *32*, 8689.
80. Chung, T.C.; Dong, J.Y. Polyolefin Containing a Terminal Styrene or Styrene Derivatives Unit. U.S. Patent 6,479,600, Nov 12, 2002.
81. Chung, T.C.; Dong, J.Y. A novel consecutive chain transfer reaction to p-methylstyrene and hydrogen during metallocene-mediated olefin polymerization. J. Am. Chem. Soc. **2001**, *123*, 4871.
82. Dong, J.Y.; Chung, T.C. Synthesis of polyethylene containing a terminal p-MS group; metallocene-mediated ethylene polymerization with a consecutive chain transfer reaction to P-MS and hydrogen. Macromolecules **2002**, *35*, 1622.
83. Dong, J.Y.; Wang, Z.M.; Han, H.; Chung, T.C. Synthesis of isotactic polypropylene containing a terminal Cl, OH, and NH_2 group via metallocene-mediated polymerization/chain transfer reaction. Macromolecules **2002**, *35*, 9352.
84. Hagihara, H.; Tsuchihara, K.; Sugiyama, J.; Takeuchi, K.; Shiono, T. Copolymerization of propylene and polar allyl monomer with zirconocene/methylaluminoxane catalyst: catalytic synthesis of amino-terminated isotactic polypropylene. Macromolecules **2004**, *37*, 5145.

Microelectronics Fabrication

Edmund G. Seebauer
Charlotte T. M. Kwok
Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.

INTRODUCTION

Chemical engineering can be defined as the study and practice of transforming substances at a large scale for the tangible improvement of the human condition. Such transformations are executed to produce other useful substances or energy, and lie at the heart of vast segments of the chemical, petroleum, and pharmaceutical industries. It is often less well recognized that chemical engineering principles form the core of most aspects of microelectronic fabrication. The microchip that emerges from a fabrication plant bears little resemblance to a tank car full of a petrochemical, yet the fabrication plant and the refinery–factory complex that generate these two kinds of products are in fact related to each other closely. Both facilities represent large-capitalization enterprises incorporating large sets of unit operations, with optimal production depending upon an understanding of phenomena on length scales ranging from the molecular to the factory level. This entry seeks to highlight how chemical engineering principles play a central role in microelectronic device fabrication.

HISTORICAL PERSPECTIVE

From modest beginnings in the middle of the last century (Table 1), microelectronic devices have come in just a few decades, to pervade nearly all aspects of modern life. This enormous change stems not only from the invention of new kinds of devices as outlined in Table 1 and materials listed in Table 2, but also from the steep power-law increase in speed and performance of integrated circuits (ICs). The simplest and most widely known metric for this improvement is commonly referred to as ''Moore's Law.''[1] This ''law'' represents a correlation originally articulated by Intel's former chief executive Gordon Moore in 1965, and states that device performance doubles roughly every 18 months. Fig. 1 shows this correlation graphically; it has been obeyed for approximately four decades. Some social commentators have suggested that this vast improvement in technology has led to improved macroeconomic performance and that the corresponding ''information revolution'' has manifestly advanced the human prospect.[2] Whether these claims are true should be the subject of another discussion. It is clear, however, that the giddy whirl of better hardware and frillier software has required consumers and businesses to upgrade their computer systems regularly, with substantial attendant costs.

Chemical engineers have played a central role in this rapid advance, particularly during the past 15 years or so. Employment statistics drive home this point. A typical new fab constructed by Intel employs about 30% chemical engineers, while the corresponding number for IBM is about 25%. These percentages equal or exceed those for any other discipline. However, their contributions remian sometimes hidden. This fact illustrates the point. Armytage had made:[3] ''The artistry of a bridge-builder is obvious to the naked eye, but the activities of the chemical engineer are not, until the products are bottled, batched or baled. Both profoundly affect the progress of mankind.''

CHEMICAL ENGINEERING AT MANY LENGTH SCALES

In a commodity chemical plant, chemical engineers have recognized for nearly a century how issues of chemistry and transport arise at length scales ranging from the factory level down to the molecular level. Such issues also pervade microelectronic fabrication as described in the following sections.

The Factory Level

A fabrication plant (or ''fab'') typically operates at 20,000 wafer starts per month or more, with each wafer taking several weeks to wend its way through the entire process sequence of over 400 individual steps. A typical fab now costs roughly US$2–3 billion to build and requires 2–3 years to complete. Fab-siting decisions rely on sophisticated and quantitative models[a] that

[a]An example is the Jupiter software from Abbie Gregg, Inc., a start-up consulting firm. See www.abbiegregg.com.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007859

Table 1 Major events in microelectronic device development

Year	Event
~1925	Vacuum tubes dominate complex circuits
~1940	First commercial sources developed for high-purity Si and Ge
1947	First transistor built (point-contact form) by Bardeen, Brattain, and Shockley of Bell Telephone Labs
1952	Junction transistor developed to avoid reliability problems with point-contact transistor
1954	First Si-based solar cell
1958	IMPATT diode invented for microwave generation; forerunner of wireless communication phones
1959	Jack Kilby files patent for first integrated circuit
1960	Invention of ruby laser, the forerunner of all solid-state lasers
1960s–1970s	MIS (metal–insulator–semiconductor) development leads to greatly reduced power consumption and miniaturization
1980s	1 μm design rule reached. GaAs circuits developed
1990s	0.18 μm design rule reached. GaAs and Si–Ge wireless devices become widespread

systematically consider factors such as government incentives, cost of land and electrical power, and proximity to universities and business partners. Crucial factors of specifically chemical nature include environmental regulations regarding gas emissions and access to water supplies for both intake and discharge.

Design of the main building itself depends upon issues of both chemistry and transport. For example, piping and equipment for chemical handling typically occupy one- quarter of the entire floor space, and need constrained optimization to minimize the required footprint while keeping incompatible chemicals stowed apart from each other. Poisonous gases such as arsine (AsH_3) and pyrophoric gases such as silane (SiH_4) have to be stored for easy access to contractors who specialize in changing large cylinders of these materials. The gases must then be piped to their place of use while keeping added contamination below parts-per-billion. Transport issues involved in such distribution are nontrivial. Many pieces of fabrication equipment (typically called ''tools'') require environment-controlled airspace, especially to minimize contamination from particulates. Particulate removal and clean air handling is clearly a problem of separations and transport. Elimination of mechanical vibration is also crucial for some tools, particularly in lithography.

Fab operation resembles a commodity chemical factory regarding the need for careful scheduling and debottlenecking. Yield levels typically exceed 90% for virtually all tools; hence major productivity and profitability improvements must originate more from optimization of the overall process flow rather than of individual unit operations.[4] In fact, a typical tool is often in production only about 30% of the time, with the balance falling prey to testing and tuning, set up, scheduled and unscheduled maintenance, and unscheduled shutdowns. The shutdowns include cancelled orders, late delivery of materials, turnover of trained operators, and software failures. Manufacturing execution systems that coordinate data management at the factory level are used to aid with these tasks.[b] Advanced systems include scheduling use of consumable items like wafers and chemicals to avoid shortages while minimizing inventory to avoid space and shelf life problems.

The complexity and expense of fab construction and operation have led to the development of detailed models for the cost-of-ownership of tools and usage of consumables. These models often call for the subcontracting of services to specialized firms. Such services include front-end processes such as silicidation and ion implantation, and back-end processes such as assembly, electrical test, and packaging. Foundry manufacturers such as Taiwan Semiconductor Manufacturing Corporation (TSMC) and Singapore's Chartered Semiconductor Manufacturing represent successful examples of this trend.

The Unit Operation Level

An overall process flow for making an integrated circuit typically consists of several hundred individual steps. A greatly simplified process flow for making an integrated circuit is seen in Fig. 2. The flow was drawn with Si-based IC fabrication in mind, but most of the major steps apply to nearly any microelectronic process. Steps before the creation of metal interconnect structures are commonly termed ''front-end,'' while subsequent ones are called ''back-end.'' Most process

[b]An example is the FAB300 software from Consilium. See www.consilium.com.

Table 2 Common materials for semiconductor devices

Material	Advantages	Uses
Si	Easily forms a high-quality oxide and simple resistive metal contact	Most widely used semiconductor for ICs.
GaAs	Better optical properties than Si Higher electron mobility for fast devices Ga can be partially replaced with Al ($Al_xGa_{1-x}As$) for tunable light emission	Dominates high-performance optical and wireless devices
Ge	Can be alloyed with Si (Si_xGe_{1-x}) for specialized devices	Si_xGe_{1-x} quantum wells, heterojunction bipolar transistors, microwave devices
InP	Can be alloyed with GaAs ($In_xGa_{1-x}As_yP_{1-y}$) to emit light with wavelength that depends upon stoichiometry	Widespread use in light sources for fiber-optic communications at 1.3 μm Some microwave devices, radiation-hard solar cells
GaN	Emits blue light	Blue lasers and light emitting diodes
InSb, $Hg_xCd_{1-x}Te$	Emits and absorbs light far into the infrared	Infrared detectors, night vision

steps fall into just a few categories that chemical engineers would consider to be "unit operations."[5] Tools for accomplishing these unit operations can be staged sequentially in a process flow, or sometimes in a "cluster" form as shown in Fig. 3, wherein several separate unit operations are carried out using a single piece of apparatus. Irrespective of whether tools are arranged individually or in clusters, chemical engineering treatments focus on classical unit-operation concerns of optimizing and controlling temperatures, pressures, concentrations, reaction times, and the like.

Various schemes can be used for classifying the unit operations. Initial wafer fabrication falls into a class by itself. However, once device creation begins, the operations can be classified into five general categories: deposition, lithography, etching, doping, and cleaning. Deposition, etching, and cleaning share a common feature of reactions occuring at a solid interface, often with an intervening solid or fluid boundary layer that inhibits diffusion of reactants and products. Such effects are shown schematically in Fig. 4. In deposition, the net effect of the diffusion–reaction is to leave material behind on the substrate. In etching, portions of the substrate itself are removed, but in cleaning only contaminants are removed leaving the substrate intact.

Wafer fabrication

Electronic device fabrication requires a substrate, which is typically a highly purified slice of single-crystal semiconductor. For optoelectronic applications, the substrate could be GaAs, InP, or some other semiconductor; but for most integrated circuits and many Si–Ge wireless applications, the wafer is made of Si.

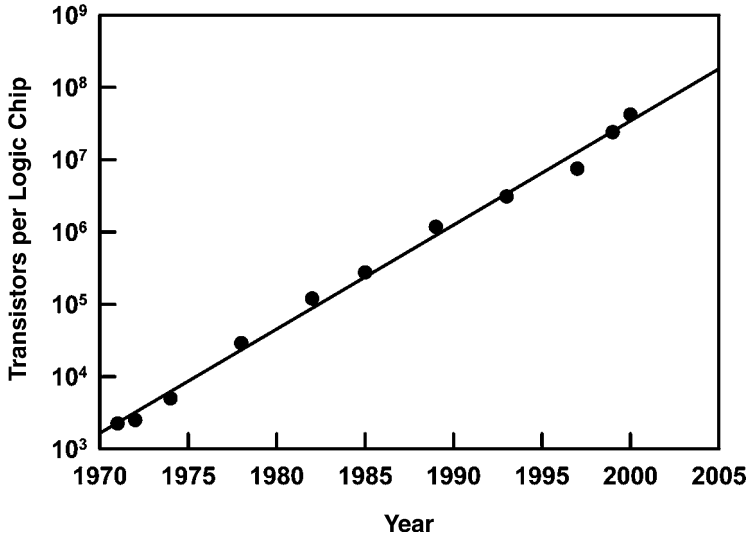


Fig. 1 A graphical representation of Moore's law for the past several decades.

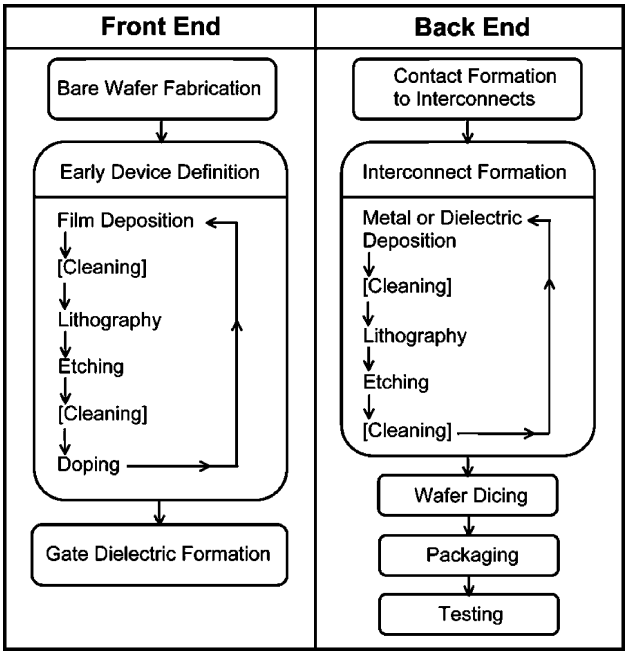


Fig. 2 A simplified process flow for making an integrated circuit. Most unit operations fall into one of the five categories: deposition, lithography, etching, doping and cleaning.

Fig. 5 shows a process sequence for the initial stages of making electronic-grade Si. The unit operations in this sequence resemble those typical of conventional chemical processing, and include furnaces, fluidized bed reactors, and distillation columns. The product semiconductor emerging from this process must however be purified further than what distillation allows, and must also be converted into single-crystal form. Invariably, the latter process involves melting the semiconductor again and crystallizing it slowly under carefully controlled conditions. Various methods exist for this purpose, including Bridgeman, floating zone, and Czochralski techniques. The Czochralski technique is favored for Si and GaAs, because crystallization occurs with no contact between the growing solid and the walls of the melt container, thereby resulting in very little defect production in the solid. The melt solidifies into a boule that is slowly pulled from the liquid as heat is continually withdrawn. Boule diameters capable of yielding Si wafers up to 300 mm across are now a routine practice. Modeling the heat transfer, impurity segregation into the liquid, and defect formation in the solid has now reached a high degree of sophistication.[6] After crystal pulling from the melt is complete, the boule is sliced into wafers that are polished and chemically cleaned in readiness for device fabrication.

Deposition

Deposition represents the primary means by which new layers are introduced onto the wafer substrate for subsequent feature delineation by lithographic patterning and etching.

Physical vapor deposition (PVD) is used most often for metals, especially where ultrahigh purity is needed in the grown film and perfect crystallinity is not crucially important. The method employs either evaporation or ion sputtering to vaporize a source material in high vacuum. The vaporized gas travels directly to the substrate without gas collisions and deposits there. Chemical reactions are minimal (leading to a material of very high purity), and transport involves line-of-sight ballistic motion. This motion leads to shadowing effects that yield poor conformality in the coverage of

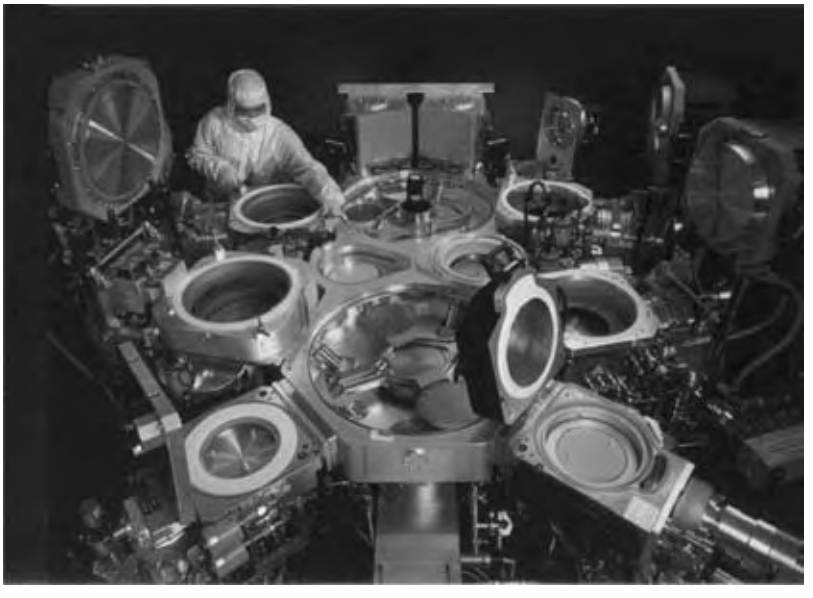


Fig. 3 Cluster tool (''Endura Classic'' from Applied Materials, Inc.) for accomplishing metal deposition on 200 mm diameter wafers. The various individual tools for unit operations surround a central handler (lower center) that transfers wafers among them. (Photo courtesy of Applied Materials Inc.) *(View this art in color at www.dekker.com.)*

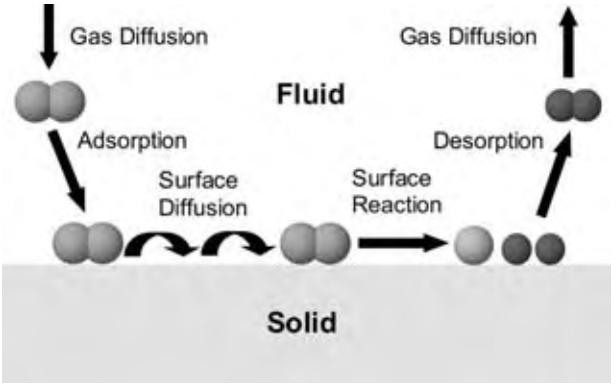


Fig. 4 Elementary molecular steps common to many deposition, etching, and cleaning processes.

recessed regions of the surface (e.g., trenches). Rotating the wafer during deposition mitigates some of these effects, as does heating the wafer to promote lateral transport by surface diffusion.

Spin casting is a liquid-based form of PVD that works at atmospheric pressure, finding use for polymers and other nonvolatile materials. These applications appear mainly in lithography and is discussed in the next section.

Chemical vapor deposition (CVD) finds its primary use in semiconductors and insulators, especially where epitaxial growth is required. (Epitaxy is the growth of a single-crystal film with a fixed orientation with respect to the crystal structure of the underlying substrate.) It employs reactive gas-phase precursors that decompose over the hot substrate, leaving solid behind. Fig. 6 shows an example of a tool commonly used for epitaxial deposition. Because of the surface reactions involved, films can sometimes be grown selectively on some regions of the substrate versus others. An example is the epitaxial growth of Si on Si using disilane, in which growth on wafer regions covered with SiO_2 is completely suppressed.[7] PVD is incapable of such selectivity. Moreover, the chemical reactions permit the growth of compounds such as GaAs with exact stoichiometry—a feat PVD cannot match. CVD often affords greater conformality as well. However, in typical applications where pressures range from millitorr to atmospheric, reactions and diffusion in the gaseous overlayer complicate the predictive modeling of CVD growth rate and thickness uniformity. Other potential CVD reactions suffer from a lack of reactive precursor gases. Sparking a plasma during growth can help to solve this problem by introducing additional reaction pathways. However, selectivity is often lost, the films suffer more contamination, and there is a possibility of ion-induced damage.

Atomic layer deposition (ALD) is a variant of CVD that relies upon the alternate exposure of a substrate to different precursor gases, with each gas being adsorbed to saturation. Between each adsorption step, the reactor is either purged with an inert gas or pumped down, so that several surface reactions take place during different stages of the cycle. Most ALD processes employed today include reactions that liberate gaseous products upon exposure to each of the precursor gases. This incarnation has been termed ''reaction sequence ALD,''[8] and should be distinguished from ''chemisorption saturated ALD,'' in which exposure to one of the precursors leads merely to chemisorbed layer, with no gaseous product produced. The reaction sequence version of ALD has clear advantages over CVD and PVD as films become increasingly thin. These advantages include exceptional thickness control, conformality, interface sharpness.

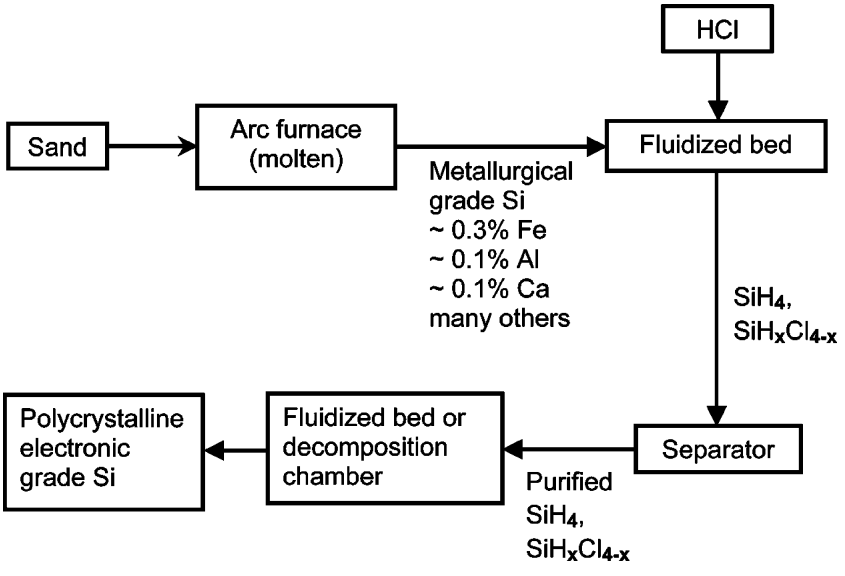


Fig. 5 Schematic process flow for making electronic-grade silicon.

Fig. 6 Deposition system used for epitaxial CVD. Heating in this case is accomplished by banks of lamps (e.g., upper right). (Photo courtesy of Applied Materials Inc.) *(View this art in color at www.dekker.com.)*

Electrodeposition is a liquid-phase analog of plasma-enhanced CVD, used especially in the formation of copper interconnects. This method employs electrochemical voltages instead of surface temperature to control the growth characteristics. It is however more difficult to model and predict than CVD, because of solvation effects and the relative dearth of experimental methods to examine individual elementary steps in the key reactions.[9]

The deposition processes discussed so far typically operate such that all the material required for the growing film comes from the overlying gas or liquid phase. Other deposition reactions involve reaction (and therefore consumption) of the underlying substrate itself. Examples of such deposition processes include thermal oxidation, nitridation, or silicidation of silicon, which can be accomplished by exposing a silicon wafer at high temperature to oxygen, ammonia, or titanium tetrachloride, respectively, to form silicon dioxide, silicon nitride, or titanium disilicide. Solid-phase diffusion and reaction processes are involved in each case.

Lithography

Individual microelectronic devices consist of numerous layers of carefully patterned thin films. Patterning depends upon the ability to define and ''draw'' a two-dimensional pattern on a wafer substrate. Currently, this ability depends on photolithography by optical projection, in which a polymer ''resist'' film is deposited on the wafer and subsequently exposed to ultraviolet light through a mask to render the illuminated areas either more or less susceptible to the action of a solvent. Use of this solvent selectively removes either the illuminated or unilluminated polymer (depending upon whether the resist is ''positive'' or ''negative,'' respectively), leaving regions of bare wafer available for further processing.

Limitations on photolithography originate primarily from image aberrations and diffraction effects introduced by the optical projection system. Image aberrations can be reduced to a considerable extent by using ''stepper'' technology, in which successive small portions of the whole substrate wafer are exposed in sequence using a small-field lens system together with a mechanical stepper motor. However, diffraction problems can be ultimately solved only by using shorter wavelengths of light to expose the resist. Pulsed excimer lasers that emit deep in the ultraviolet region have found extensive use for this purpose: for example, KrF lasers with output at 248 nm.

The deposition, exposure, and removal of resists involve quite a few issues of chemistry and transport[10] that benefit from the chemical engineering perspective. For example, resists are typically deposited by a spin-on process, in which the polymer is dissolved in a solvent and poured onto a rapidly spinning wafer. With the addition of suitable viscosity thinners and careful choice of spinning speed, this mixture spreads uniformly over the wafer with simultaneous evaporation of the solvent. Detailed modeling of the simultaneous flow and evaporation presents a nontrivial transport problem. Chemical issues make their appearance during optical exposure. Positive resists contain

roughly 20% of a photosensitizer such as naphthoquinone diazide, which absorbs light and decomposes to render the resist more soluble. By contrast, negative resists contain only 2–3% of a photosensitive agent such as a quinone or an azo compound that crosslinks to the resin matrix and renders the resist less soluble. The ever-shorter UV wavelengths needed to achieve high spatial resolution have an increasing tendency to pass through resists without stimulating such chemistries, so that the differential solubility between the exposed and the unexposed regions decreases. Chemical issues also appear during resist removal, both during pattern formation and subsequent stripping of all resists after the pattern has been defined. The issues typically revolve around reagent disposal; processes based upon organic solvents or oxidizing acids present more problems in this regard than dry plasma processes based upon oxygen, for example. However, dry processing with energetic ions typically induces undesired wafer damage in the surface region that needs to be avoided.

Etching

After lithography is used to deposit resist in a patterned way, the wafer must be exposed to some sort of etching to emblazon the pattern into the underlying substrate. Other forms of etching seek to remove gross debris left over from previous processing steps. Such debris could include photoresist that needs to be stripped off after a pattern-defining etch, or metal left over from a solid-phase silicidation reaction with silicon.

Wet etching appears mostly in applications where high spatial resolution is not critical. The process employs an aqueous acid or some other chemically reactive solution. The chemical specificity of liquid reagents is often exquisite, permitting the etching of one solid to the complete exclusion of others. However, in most cases the etching reaction is isotropic, meaning that the sidewalls of a trench disappear at the same rate as the bottom as shown in Fig. 7. Also, the etching reaction cannot be stopped instantaneously; an ill-defined period of time is needed to rinse the solution from the surface (including within trenches). Thus, controllability is too poor for fine features.

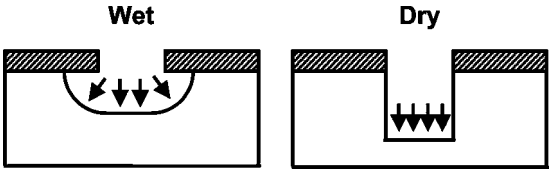


Fig. 7 Isotropic etching characteristic of wet etching (left) yields undercutting, leading to poor aspect ratios. Anisotropic etching characteristic of reactive ion etching (right) gives higher aspect ratios.

Etching of SiO_2 is almost always done with aqueous HF, because the solution is very selective against etching any underlying Si. However addition of HNO_3 to aqueous HF makes a fine etchant for Si; the nitric acid oxidizes the Si to SiO_2, which is immediately removed by HF. For elemental metals, some form of electrochemical oxidation usually operates to dissolve the metal. Both aqueous acids and bases can serve this function, although acids find greater use because they do not attack SiO_2 elsewhere on the wafer the way bases do.

"Dry" or plasma etching operates via the action of ions and/or free radicals on the surface. Plasma etching can be run in several modes, depending upon the relative balance that is sought between the effects of ions and free radicals. At one extreme lies ion milling, in which an inert-gas plasma is run at low pressures and high bias voltages (>300 V) with respect to the wafer. In this case, material removal takes place by physical erosion only. Although both the rate and the degree of anisotropy are high, chemical specificity is low and the substrate usually suffers considerable ion-induced damage. At the other extreme lies a free radical plasma, which employs halogen-containing gases at very low bias voltages (<70 V). In this case, material removal takes place almost entirely by chemical reactions. The chemical specificity is high and substrate damage is avoided, but the rate and anisotropy are low. Between these two extremes is reactive ion etching (RIE), wherein both ion sputtering and chemical reaction play a role. The plasma runs with a gas capable of producing reactive radicals of Cl, F, or O, and operates at intermediate bias voltages typically between 7 V and 300 V. The combination of moderate rate, anisotropy, chemical selectivity, and damage levels makes this mode the most common one for dry etching. Careful tuning of the bias voltage, pressure, and type of source gas offers extensive opportunities for optimizing these metrics of process performance.[11] Process controllability and reproducibility is good because the reaction stops as soon as the power to the plasma is shut off. This controllability, together with the considerable anisotropy (Fig. 7) offered by RIE, gives this method overwhelming advantages over wet etching for fine-feature patterning.

Chemical mechanical polishing (CMP) represents a specialized type of all-wafer etching meant to clean, smoothen, and planarize the entire wafer in preparation for subsequent lithographic steps. Modeling of CMP is complex[12] because both mechanical abrasion and chemical reaction are involved.

Doping

Doping intentionally introduces impurities into the semiconductor substrate to control its electronic

properties such as conductivity. Atoms of the dopant often have a valence similar to but not the same as those in the host lattice. For example, Si and Ge have valences of four and are typically doped with elements having valences of three (boron) or five (arsenic). GaAs comprises elements of valence three and five, and can be doped with Si having valence four.[13]

When deposited-thin films must be doped uniformly (as in the fabrication of many optoelectronic devices), it is common practice to use a CVD process wherein small amounts of a precursor gas containing the dopant are added to the reacting mixture. For example, Si-doped GaAs can be made using $(C_2H_5)_3Ga$ and AsH_3 with a little SiH_4 added.

However, often the substrate itself must be doped. If the doped region needs to be fairly thick and device fabrication has not reached stages where extended annealing at high temperatures could be damaging, then thermal solid-state diffusion processes can be employed. Several such methods exist, but they all involve forming a layer of dopant oxide (e.g., B_2O_3 or As_2O_3 on Si) on top of the semiconductor and then heating to temperatures that are appreciable fractions of the wafer melting temperature. The oxide then releases dopant atoms into the substrate, which diffuse in.

In more critical applications involving pn junction formation, the diffusion profiles afforded by thermal diffusion are too broad and deep to be useful. It is then preferable to introduce dopants into the substrate bulk by implantation of energetic ions. With ions having energies of only a few hundred electron volts, pn junctions can be made that are as shallow as 15–20 nm. A postimplant annealing step is then required to remove crystal defects and move the dopant onto lattice sites where it can become electrically active. However, the residual crystal defects mediate extraordinarily fast diffusion of dopants, particularly in the case of boron,[14] which tends to make the pn junction significantly deeper than originally intended. The precise temperature trajectory of the annealing exerts an enormous influence on this ''transient enhanced diffusion'' and therefore on the shape of the diffused profile. Current experimental and modeling efforts focus on the optimization of this trajectory through lamp-based rapid thermal processing (RTP), in which heating rates of 400°C/sec or more can be reached.[15]

Cleaning

Wafer cleaning can take several forms;[16] but most commonly it involves using aqueous liquids to remove small quantities of chemical residues and particulates left over from previous processing steps. For example, care must be taken to avoid leaving unreacted metal atoms on a silicon surface, because these atoms can diffuse into the silicon and adversely affect its electronic characteristics. Halogens such as chlorine must be scrupulously removed from the vicinity of chip-level interconnects or wire connections within a device package to inhibit corrosion of the elemental metals located there. Residual native oxides and organic species must be removed before epitaxial growth to ensure a high-quality interface. In all these cases, particulates that adhere to the wafer surface must be removed to avoid interfering with subsequent lithographic steps.

In some cases, simple rinsing with deionized, particulate-free water suffices. More commonly, however, removal of residues requires some form of chemical action. The reagents are then acids such as HCl or HNO_3 (for transition metals), HF (for SiO_2), or oxidants based on H_2O_2 (for organics). Regardless of the cleaning solution used, exposure to the liquid phase can be performed in either semibatch mode via spraying or batch mode by dipping in successive baths.

The Molecular Level

Good modeling of the chemistry and transport governing tool behavior almost always involves the use of reaction and diffusion rate expressions with appropriate parameters. As the forms of these expressions together with their parameters are manifestations of molecular phenomena, even phenomenological modeling often demands at least some recourse to a molecular view. As device dimensions have decreased progressively, such a view has become correspondingly more important.

As an example, Fig. 4 shows elementary steps common to most deposition, etching, and cleaning processes. First, a reactive molecule must approach the interface of interest, often diffusing through a fluid boundary layer. Sometimes a solid diffusion layer is also present. Upon reaching the interface, the molecule may then pass through some or all the following elementary chemical steps: adsorption, surface diffusion, surface reaction, and desorption. Product molecules also usually need to desorb, diffusing back through any solid or fluid boundary layers back into the bulk fluid.

As another example, atomic clustering and dopant motion during annealing after ion implantation can be modeled by a complex set of reaction–diffusion equations.[17] Determination of appropriate rate parameters requires an assessment of the importance of surface space charge on the motion of charged bulk interstitials and of transport aided by photogenerated carrier recombination. Such an assessment plays a key role in determining optimal temperature trajectories for RTP in postimplant annealing.

Environmental, Safety, and Health Issues

Semiconductor processing has witnessed ever-increasing attention to environmental, safety, and health (ES&H) issues. Chemical engineers play a central role in solving these problems.

Environmental issues show up throughout a processing sequence and usually involve the need to dispose of large quantities of hazardous reagents cost-effectively. For example, photoresist removal has often been accomplished with strong liquid oxidants such as H_2SO_4–H_2O_2 mixtures. However, new methods based on aqueous ozone are being developed,[18] as disposal of large quantities of acid can be avoided. Wafer cleaning employs very large quantities of water, and the effluent typically contains trace amounts of dissolved metals and other undesirable species. Separation processes are required before the effluent can be released back into the environment.

Safety concerns are near the top of the list of concerns in both fab and tool design. The gases typically used as reagents or produced as products in chemical vapor deposition, etching, and doping are exceptionally poisonous (e.g., AsH_3 and B_2H_6), corrosive (e.g., HCl and HF), flammable (e.g., organic solvents), or pyrophoric (e.g., SiH_4). Fab designs must ensure that liquid acids and bases are stored well apart from each other and that cylinders of poisonous gases be changed well away from where primary fab activity is located. Tool designs require provision for automatic shutdown and proper venting in case of power loss, operator error, or other malfunction.

Health issues arise, especially with respect to long-term exposure to low levels of these chemicals. The need for employees to work in a clean-room environment while wearing unwieldy ''bunny suits'' to reduce particulate contamination and protect against chemical exposure can also present problems with physical and mental stress, especially given the intense time pressures associated with fabs that run around the clock.

EMERGING PROCESSING CHALLENGES

Forecasting the future is always uncertain business, but the Si-based fabrication industry may be unique in the depth, breadth, and sophistication it uses to forecast developments and plan research and investments accordingly. Such forecasting germinated in the United States during the mid-1980s in response to intense competitive pressures from overseas, especially from Japan. In the face of relentlessly shrinking market share, especially in the equipment manufacturing business, the Semiconductor Industry Association (SIA), comprising virtually all major US integrated circuit manufacturers, began to recognize clearly the benefits of joint research and development in a technology base that was ''precompetitive,'' meaning common and fundamental to the engineering that member companies needed to do. Among the creations of the SIA was the National Technology Roadmap for Semiconductors (NTRS), which was produced by a broad-based committee of industry experts to define precompetitive areas of technology and to project how that technology would need to develop to keep the industry on its profitable Moore's Law trajectory.

During the 1990s, the NTRS received updates at roughly three-year intervals and became an exquisitely detailed document that defined the needs and pace of development for everything—from lithographic techniques to shallow junction design. While the industry largely managed to keep up with the aggressive milestones set down by the various roadmaps, it became clear by the end of the decade that certain technical problems would likely become so large and difficult that more than a national effort would be needed. Thus, representatives from major corporations in the Far East and Europe were consulted for the 1999 version of the roadmap, which accordingly became the ITRS, with ''I'' standing for ''international.''[19]

The ''Wall''

Even though the 2002 ITRS is still quite aggressive in projecting Moore's law trajectory, many experts believe that this trajectory is approaching the end of its nearly long run. Cost and technical limitations imposed by device and material physics will end rapid power-law device scaling. Many experts believe these problems will converge concurrently quite soon, erecting what many call simply ''the Wall.'' Although some industry observers have been predicting the end of Moore's law for decades, there are several reasons to believe that forecast may soon prove accurate.

Staggering costs of new fabs

The staggering capital costs and substantial lead times for building new fabs are causing the microelectronics industry to resemble the commodity chemical industry, with boom–bust cycles of production capacity and profitability. This behavior has been particularly evident in the production of dynamic random access memory (DRAM) devices. This segment of the industry was plagued in the late 1990s by overcapacity and marginal profitability, leading to extensive consolidation. Early in that decade, there were four significant DRAM manufacturers in the United States. By the end of the decade there remained only one (Micron Technologies). Economic turmoil in Pacific Rim countries exacerbated the problem in the late 1990s, but the ingredients for

the transformation of a specialty-product business into a commodity business were already in place. Consolidation has been steadily taking place in the microprocessor unit business as well. The 1998 sale of Digital Equipment's microprocessor manufacturing operations represents one such example.

Finding a cost-effective lithography

Optical lithography has provided the mainstay of pattern delineation on wafers for many years. In the early days of fabrication when device dimensions were much larger than a wavelength of UV light, lithography was relatively straightforward to accomplish. However, with current critical dimensions for commodity devices dropping below 0.1 μm, which is below the wavelength of light sources suitable for production, much more advanced projection methods like phase shifting are needed.[20] Reducing the wavelength also helps. The industry has recently relied upon XeCl excimer lasers with $\lambda = 308$ nm. Transition is taking place to KrF ($\lambda = 248$ nm) and may eventually move to ArF[21] ($\lambda = 193$ nm) or F_2[22] ($\lambda = 157$ nm). However, resists with suitable chemistry and photoactivity become progressively more difficult to make at shorter wavelengths.[23] This problem, coupled with the lack of practical light sources below 157 nm, will probably terminate the continued scaling of conventional UV-based lithography. Over the years, the industry has worked extensively on alternative techniques, such as electron-based SCALPEL[24] and X-ray lithography.[25] However, such techniques have generally foundered against the shoals of cost-effectiveness in a large-scale production environment.[26]

Interconnect delays

As individual devices get closer together on a chip, the spacing between the metal interconnects linking them together decreases correspondingly. Elementary electromagnetic principles then dictate that the capacitance C between the wires, which varies inversely with separation, must also increase. Thinner interconnects also lead to an increase in resistance R. Thus, even though the interconnects get shorter, the propagation delay (governed by $R \times C$) stops decreasing when device scaling reaches a certain point, and actually increases beyond that point. Fig. 8 shows this effect graphically.[27] Even with clever device layout on the chip, such physics eventually halts progress in improving clock speeds. This problem can be avoided to some extent by using metals with lower R. Indeed, for this reason the industry is moving from aluminum-based interconnects to copper.[28] Moreover, the capacitance can be reduced by using insulators with low dielectric constant k to separate the interconnects. The industry currently employs dielectrics

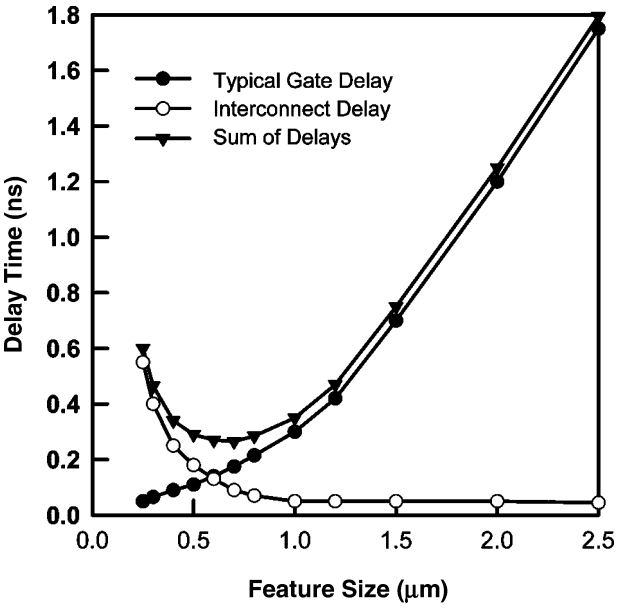


Fig. 8 Gate and interconnect delay versus feature size. Interconnect delay is shown for repeater spacings of 3000 μm. (Some data adapted from Ref.[29].)

based on SiO_2 ($k = 4$) and is actively investigating fluorocarbon polymers, aerogels, and other "low-k dielectrics" as replacements.[29] However, these measures are expensive to implement and represent stopgaps at best. Copper has about the lowest resistance available for a practical interconnect material, and k lower than unity is impossible. At this point entirely new paradigms like optical interconnects, are needed.

Problems with gate/capacitor dielectric materials

As device dimensions scale down, additional physical limitations arise because of fundamental aspects of the properties of the constituent materials. For example, central to both the field-effect transistors and capacitors that form the backbone of most complementary metal oxide semiconductor (CMOS) logic and dynamic memory devices is a thin film of dielectric material to prevent the flow of direct current between key components. In a transistor, this material lies sandwiched between the gate and the underlying channel between the source and drain. In a capacitor, this material is between the two electrodes. In both cases, nitrogen-doped amorphous SiO_2 is typically used, with thicknesses down to 2 nm.[30] While field effects in transistors and capacitance in capacitors rise as this distance increases, ways are needed to suppress current leakage to provide the requisite dc electrical isolation. Much work has focused on leakage mechanisms involving defects caused by strain, contamination, and injected charge.[31] However, this problem has been largely solved, so that below about 4 nm, the significant electrical conductivity because of

quantum mechanical tunneling poses an absolute limit on the ability of SiO_2 to provide isolation. A solution must be found to the tunneling problem—one that is universally recognized to require an insulator with a higher dielectric constant k. The search for ''high-k dielectrics'' (as opposed to the low-k dielectrics of the previous section) has proceeded for about the last decade. Much research has focused on materials such as Ta_2O_5 ($k = 20$–50), various ferroelectrics such as $Ba_xSr_{1-x}TiO_3$ ($k \approx 100$–400),[32] and more recently, oxides and silicates of metals from Group IVA of the periodic table. Examples include ZrO_2, HfO_2, $ZrSiO_4$, $HfSiO_4$, and $HfSi_xO_y$,[33] with the first two oxides receiving most attention in recent years. Ta_2O_5 has been developed sufficiently to find current use in production (for memory devices, but not transistor gates). $Ba_xSr_{1-x}TiO_3$ suffers from severe reliability problems. The Group IVA materials show more promise, but still require development to solve difficulties not only with current leakage and interfacial properties but also with process integration issues such as patterning and etching. Though are some reasons to believe that a suitable replacement for SiO_2 does not exist,[34] there is very recent evidence that the key problems can be solved.[35]

CONCLUSIONS

Current microelectronic processing involves working around fundamental constraints imposed by cost and physics. However, working in the presence of constraints is fundamental to chemical engineering: improved devices and processes are developed through optimization of all kinds. In other words, the industry must continue to learn how to squeeze most out of what Mother Nature permits. In device fabrication, where rates of transport and physical/chemical transformation govern, chemical engineering principles are needed more than ever.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation (CTS 02-03237). Opinions expressed here are those of the authors and not necessarily those of NSF. The authors thank M. Isenburger and N. Henshaw of Intel and W. Davies of IBM for these statistics.

REFERENCES

1. Moore, G.E. Cramming more components onto integrated circuits. Electronics **1965**, *38*, 114–117.
2. Gordon, R.J. Does the ''new economy'' measure up to the great inventions of the past? J. Econ. Pers. **2000**, *14*, 49–74.
3. Armytage, W.H.G. *A Social History of Engineering*; MIT Press: Cambridge, 1961.
4. Sack, E.A. Global yield engineering for IC production. Solid State Technol. **1998**, *41*, 81–85.
5. Hess, D.W., Jensen, K.F., Eds.; *Microelectronics Processing: Chemical Engineering Aspects*; Advances in Chemistry Series 221; American Chemical Society: Washington, DC, 1989.
6. Sinno, T.; Brown, R.A.; Ammon, W. von; Dornberger, E. Point defect dynamics and the oxidation-induced stacking-fault ring in Czochralski-grown silicon crystals. J. Electrochem. Soc. **1998**, *145* (1), 302–318.
7. Violette, K.E.; O′Neil, P.A.; Öztürk, M.C.; Christensen, K.; Maher, D.M. Low temperature selective silicon epitaxy by ultra high vacuum rapid thermal chemical vapor deposition using Si_2H_6, H_2 and Cl_2. Appl. Phys. Lett. **1996**, *68* (1), 66–68.
8. Sneh, O.; Clark-Phelps, R.B.; Londergan, A.R.; Winkler, J.; Seidel, T.E. Thin film atomic layer deposition equipment for semiconductor processing. Thin Solid Films **2002**, *402* (1–2), 248–261.
9. Drews, T.O.; Ganley, J.C.; Alkire, R.C. Evolution of surface roughness during copper electrodeposition in the presence of additives. J. Electrochem. Soc. **2003**, *150*, C325.
10. Wolf, S; Tauber, R.N. *Silicon Processing for the VLSI Era: Process Technology*, 2nd Ed.; Lattice Press: Sunset Beach, CA, 1999; Vol. 1.
11. Zhang, D.; Kushner, M.J. Investigations of surface reactions during C_2F_6 plasma etching of SiO_2 with equipment and feature scale models. J. Vac. Sci. Technol. **2001**, *A19*, 524.
12. Singh, R.K.; Bajaj, R. Advances in chemical-mechanical planarization. MRS Bulletin **2002**, *10*, 743–747.
13. Masi, M.; Simka, H.; Jensen, K.F.; Kuech, T.F.; Potemski, R. Simulation of carbon doping of GaAs during MOVPE. J. Crystal Growth **1999**, *124*, 483–492.
14. Downey, D.F.; Falk, S.W.; Bertuch, A.F.; Marcus, S.D. Effects of 'fast' rapid thermal anneals on sub-keV boron and BF_2 ion implants. J. Electronic Mater. **1999**, *28*, 1340.
15. Thakur, R.P.S.; Timans, P.J.; Tay, S.P. RTP technology for tomorrow. Solid State Technol. **1998**, *41*, 171–175.
16. Campbell, S.A. *The Science and Engineering of Microelectronic Fabrication*, 2nd Ed.; Oxford: New York, 2001.
17. Jung, M.Y.L.; Gunawan, R.; Braatz, R.D.; Seebauer, E.G. New physics for modeling transient

enhanced diffusion in RTP. Electrochemical Society **2000**, *2000-9*, 15–20.
18. Nelson, S. Ozonated water for photoresist removal. Solid State Technol. **1999**, *42*, 107–112.
19. The 2002 ITRS can be viewed at http://public.itrs.net/Files/2002Update/2002Update.htm.
20. Shibuya, M. Resolution enhancement techniques for optical lithography and optical imaging theory. Optical Review **1997**, *4*, 151–160.
21. Rothschild, M.; Forte, A.R.; Kunz, R.R.; Palmateer, S.C.; Sedlacek, J.H.C. Lithography at a wavelength of 193 nm. IBM J. Res. Dev. **1997**, *41*, 49–55.
22. Bloomstein, T.M.; Horn, M.W.; Rothschild, M.; Kunz, R.R.; Palmacci, S.T.; Goodman, R.B. Lithography with 157 nm lasers. J. Vac. Sci. Technol. **1997**, *B15*, 2112–2116.
23. Reichmanis, E.; Nalamasu, O.; Houlihan, F.M. Organic materials challenges for 193 nm imaging. Acc. Chem. Res. **1999**, *32*, 659–667.
24. Harriott, L.; Waskiewicz, W.; Novembre, A.; Liddle, J.A. Favored SCALPEL's continued progress. Solid State Technol. **1999**, *42*, 73–78.
25. Silverman, J.P. Challenges and progress in X-ray lithography. J. Vac. Sci. Technol. **1998**, *B16*, 3137–3141.
26. Ito, T.; Okazaki, S. Pushing the limits of lithography. Nature **2000**, *406*, 1027–1031.
27. Havemann, R.H.; Hutchby, J. A. High-performance interconnects: an integration overview. Proceedings of the IEEE **2001**, *89*, 586.
28. Murarka, S.P. Multilevel interconnections for ULSI and GSI era. Mater. Sci. & Eng. Rept. **1997**, *19*, 87–151.
29. Peters, L. Low-k dielectrics: will spin-on or CVD prevail? Semiconductor International. **1999**, *23*, 108–124.
30. Maiti, B.; Tobin, P.J.; Misra, V.; Hegde, R.I.; Reid, K.G.; Gelatos, C. High performance 20 angstrom NO oxynitride for gate dielectric in deep subquarter micron CMOS technology. In *IEDM Technical Digest*; International Electron Devices Meeting, Washington, DC, Dec 7–10, 1997; Institute of Electrical and Electronics Engineers: New York, 1997; 651–654.
31. Balk, P. Dielectrics in microelectronics: problems and perspectives. J. Non-Cryst. Solids. **1995**, *187*, 1–9.
32. Wessels, B.W. Metalorganic chemical vapor deposition of ferroelectric oxide thin films for electronic and optical applications. Ann. Rev. Mater. Sci. **1995**, *25*, 525–546.
33. Wilk, G.D.; Wallace, R.M. Electrical properties of hafnium silicate gate dielectrics deposited directly on silicon. Appl. Phys. Lett. **1999**, *74*, 2854–2856.
34. Lucovsky, G.; Yang, H.; Niimi, H.; Thorpe, M.F.; Phillips, J.C. Intrinsic limitations on ultimate device performance and reliability at (I) semiconductor-dielectric interfaces and (II) internal interfaces in stacked dielectrics. J. Vac. Sci. Technol. **2000**, *B18*, 2179–2186.
35. Lammers, D. Companies put high-k on fast track for 45 nm. EE Times **2003**, Dec. 15, 2003, 18310280.

Microfabrication

Chung-Chiun Liu
Electronics Design Center and Department of Chemical Engineering, Case Western Reserve University, Cleveland, Ohio, U.S.A.

INTRODUCTION

Microfabrication is a mature silicon-based manufacturing processing. It is effective to produce miniature size, geometrically well-defined, highly reproducible structures. While microfabrication processes are used extensively in microelectronic industries, it is interesting to know that these processes are intimately related to chemical engineering sciences. For instance, most of these microfabrication processes are batch processing, and the principles of transport phenomena and reaction kinetics are applied broadly in these microfabrication processes. There are excellent books describing the microfabrication processing in detail.[1,2]

Microfabrication processes consist of many different procedures, and four major processes will be discussed here:

- Lithographic reduction
- Oxidation
- Thick and thin film metallization
- Photoresist patterning and chemical wet etching

LITHOGRAPHIC REDUCTION

Historically, lithography is a stone-writing or stone-rubbing technique. In this process, an image or picture is first carved onto a flat-surface stone. The stone is then properly treated with chemicals resulting in the image and nonimage areas of the stone. Subsequently, ink is applied onto the carved stone surface, and an image receptive material, such as paper, is placed over the stone. By rubbing the paper over the stone, the image on the carved stone will then be transferred onto the paper. In microelectronic processing, lithographic technique is used to translate a designed pattern onto a substrate through the use of a mask. This translation is often a reduction of the original sizes.

The most commonly used lithographic technique is photolithography. It combines the photographic reduction capability and the lithographic pattern translation through a mask. Photolithography is relatively simple, but it is limited to minimum size reduction. Nevertheless, it can be used to produce geometrically well-defined microstructures with a high degree of reproducibility. Photolithography involves three major tasks: 1) photographic reduction; 2) mask fabrication; and 3) optical printing.

Photographic reduction starts with a design, which is generally produced using a computer-aided design (CAD) software package. The design can be any type of device that can be an integrated circuit, a sensor, or others. Consequently, the structure of this device may be multilayer. This will require more than a single mask, and in this case, proper alignment for the layers is critical. The original design of the device can be transferred onto a rubylith, then the design should be photographically reduced onto a special plate for the mask.

The mask is an integral part of the lithographic technology as well as for the patterning process. In IC processing, it is further defined as reticles and masks. A reticle is considered to contain a designed pattern that can be stepped and repeated over an entire substrate. A mask is viewed to be able to expose the designed pattern over the entire substrate at once. Obviously, the resolution and definition of a reticle or a mask are critical to the final definition of a microstructure. The polarity of a reticle or a mask can be either a dark field or a light field. Fig. 1 shows an example of the light field and a dark field of a mask. Obviously, the light parts of a mask permit the light, ultraviolet light, to penetrate through whereas the dark portions would not allow any light to go through. This will directly influence the results of photoresist patterning.

Using the patterned mask, the transformation of the design of the device to the substrate can be accomplished by optical printing techniques. There are three general optical photographic techniques, and they are contact printing, proximity printing, and projection printing. Each of these printing techniques has its own merits and limitations. In contact printing, the mask is in direct contact to the surface of the substrate or the surface of the material layer, such as photoresist or others, which covers the surface of the substrate. In this configuration, the image of the device can be translated directly to the surface of the substrate without any optical distortion. Because the mask is directly touching the surface layer of the substrate, damage, such as scratching on the mask, often occurs after a limited number of usages. This may not be acceptable economically.

In the proximity printing, a gap space is placed between the mask and the surface layer of the substrate.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007861

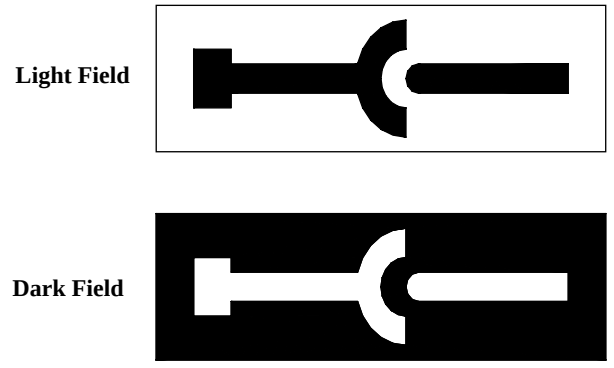


Fig. 1 The light field and dark field polarities of a typical mask.

There is no direct contact between the mask and the surface layer, minimizing the damage due to direct contacting. Because there is a gap space between the mask and the surface layer of the substrate, there is a distortion between the pattern on the mask and its translation onto the surface layer of the substrate. As expected, an increase in the gap space will degrade the resolution of the patterned image significantly.

Projection printing involves a much more complicated optical system. Optical lenses and mirrors are used to focus the pattern on the mask onto the substrate surface. There are several types of projection printing, including projection scanners, reduction step-and-repeat projection aligners, and nonreduction step-and-repeat aligners. Each type of the projection printing has its own merits and limitations. In general, the feature of step-and-repeat in projection printing is very useful to reproduce the image of a device accurately on the substrate surface, resulting in producing highly reproducible geometric structures.

The merits, advantages, and limitations of each optical printing technique have been discussed elsewhere.[3] For the microelectronic industry, as the demand of smaller device size increases, advancement in lithographic techniques is needed. This advancement includes the x-ray, electron-beam, and ion-beam lithographics. As the names of these lithographic techniques imply, various high-energy sources for the optical printing are used to enhance the resolution of the device structure. There is continuous development in lithographic technology including scanning, probe, and atomic force microscopic lithography, three-dimensional lithography and the so-called ''soft'' lithography. The description of these techniques is beyond the scope of this article.

OXIDATION

Silicon is a semiconductive material. When silicon is used as the substrate for the fabrication of IC and devices, the surface of the silicon substrate needs to form an insulation layer on which the IC and devices can be fabricated. This insulation layer can be silicon dioxide, silicon nitride, or other materials, and silicon dioxide is, however, the most commonly used insulation layer. Silicon dioxide is stable and tenacious and is well suited as the foundation layer for the construction of IC and devices. The formation of the SiO_2 layer on the silicon substrate can be accomplished by various means, and thermal oxidation is probably the preferred approach. The thickness of the oxide layer varies depending on the application, and a thickness range of 50–10,000 Å SiO_2 formed by thermal oxidation is commonly used.

The basic mechanism for the formation of SiO_2 on the silicon surface is well understood. The seminal work by Deal and Grove provided a mathematical model that describes the growth kinetics of the SiO_2 layer.[4] In this model, the oxidation process progresses as the diffusion of an oxidant takes place at the interface of Si/SiO_2. The oxidant can be oxygen in a dry oxidation process and H_2O molecule in a wet oxidation process. Because the oxidation reaction takes place at the Si/SiO_2 interface, the continuous growth of the oxide film is a moving boundary transport phenomena. Also, as the oxide layer grows, silicon will be consumed by the oxidation process. Based on the relative density and molecular weights of Si and SiO_2, the amount of silicon consumed is 44% of the final oxide thickness.[2,4] This implies that for an oxide layer of 10,000 Å, 4400 Å of the Si will be consumed.[2,4]

Thermal oxidation of silicon can be achieved under dry or wet oxidation condition. In dry oxidation, oxygen is the oxidant whereas in wet oxidation, water molecule is the oxidant. Thermal oxidation is often carried out at elevated temperatures, such as 600–1250°C. The oxide growth rate is generally faster in wet oxidation compared to dry oxidation. Also, the growth rate of the oxide depends on the crystallographic orientation of the silicon, that is, the linear oxidation rate of silicon qualitatively is [110] > [111] > z[100]. However, crossover in growth rate is possible; for instance, the oxide growth rate at 700–1000°C, [111] > [110], but not at 1100°C.

As in all microfabrication processes, the cleanliness of the substrate is very critical for the silicon oxidation process. As discussed above, the crystalline orientation will influence the oxidation rate of silicon dioxide. There are other operational parameters that will affect the oxidation rate of silicon. This includes the dopants in silicon, the trace amount of water, the concentration of Cl-bearing species, the temperature control, and its profile. Extensive assessment of each parameter on the growth rate and the quality of the SiO_2 layer on silicon can be found in microfabrication related literature elsewhere.[1,2]

There are other insulating materials that can be used instead of silicon dioxide. Silicon nitride, alumina, and aluminum nitride are a few that are often used. The selection of a proper insulation layer is based on the specific needs and the properties of selected materials.

THICK AND THIN FILM METALLIZATION

Deposition or formation of a metallic film is an important aspect in microfabrication. For traditional IC and electronic devices, metal films such as aluminum or copper film provide the electrical contact and connection from the device to external circuitry or equipment. In this circumstance, the line width and its electrical continuity are of interest. With the advancement of using microfabricated devices in chemical and biological application, formation of metallic films, such as gold, platinum, and others, becomes necessary. The formation of this metal film can be accomplished by thick or thin film techniques.

Thick film technology has been used extensively in the microelectronic industry, particularly in the manufacturing of hybrid circuit boards. Basically, thick film technology is a silk-screening printing process.[5] In general, the process involves a screen with the designed pattern, an ink or paste containing the desired component, and a mechanical squeegee. The screen has different mesh sizes and its selection is based on the properties of the ink or paste, such as particle size, viscosity, and others. The thick film ink or paste usually contains three components, the desired film material, such as gold, platinum, or others, a binding compound, and a solvent. For the formation of a metallic film, such as gold, platinum, and others, particles of the metal are used. These particles are then mixed with a binding compound, or the binder, into a solvent. The binder is often an inorganic compound, such as glass, and it serves to bind the metallic particles onto the surface of the substrate. However, organic binders, such as polyethylene oxide, are also used. The solvent is the vehicle to produce a homogenous mixture for the silk-screen printing. It is essential that the prepared ink or paste is homogenous in nature to obtain a uniform thick film. The printed thick films are not limited to metallic film; insulation materials, carbon, and other nonmetallic films can also be silk-screened using appropriate inks and pastes.

The thick film printing sequence involves first transferring the design onto a screen, then placing the thick film ink over the screen, and then using a squeegee to spread the ink over the screen. In this process, the ink or paste will go through the opening of the mask, forming the planned structure onto the substrate. Fig. 2 shows a typical sequence of the thick film processing. In this process, the thickness of the thick film in a single passage of the squeegee is approximately 5–30 μm. This thickness is controlled by the pressure applied by the squeegee, the gap space set between the screen and the substrate, and the properties of the ink or paste, such as viscosity and composition. After printing, the substrate is placed in an oven at a relatively low temperature, i.e., approximately 70–80°C, to remove the solvent. Then, the substrate is put inside a high-temperature furnace to carry out the binding process. For commercially available thick film inks or pastes, information for the heating temperature and the heating profile is readily accessible.

One of the attractive features of the thick film process is that the capital investment for thick film printing is relatively low. Basically, it involves a thick film printer, an oven, and a furnace. One of the notable commercial successes of using thick film techniques is the manufacture of glucose strips for diabetic patient management. The strip involves thick film printing and is cost-effective, making the disposable strip a reality.

However, one must also recognize the limitation of thick film printing techniques. For example, the minimum line width in thick film printing is in the order of 20–50 μm. Also, the reproducibility of the thick film printing is about ±10%. Therefore, the user of thick film printing technique needs to recognize its limitation.

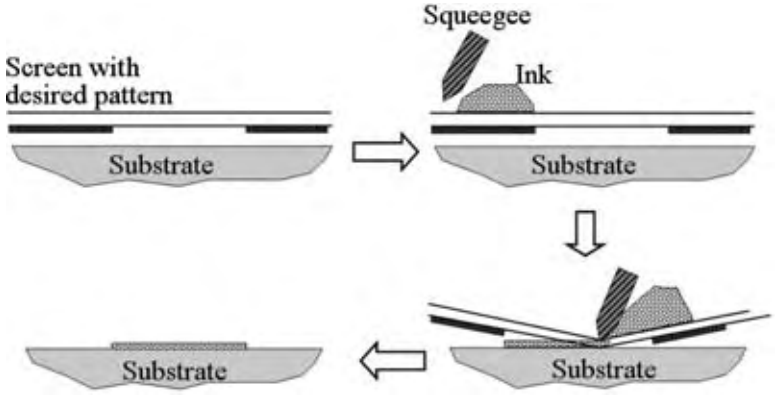


Fig. 2 Typical sequence of a thick film printing processing.

For the formation of a metallic film in addition to thick film silk-screen technique, thin film metallization is another means for the film deposition. Deposition of thin film can be accomplished by either physical or chemical means, and thin film technology has been extensively used in the microelectronics industry. Physical means is basically a vapor deposition, and there are various methods to carry out physical vapor deposition. In general, the process involves the following: 1) the planned deposited metal is physically converted into vapor phase and 2) the metallic vapor is transported at reduced pressure and condensed onto the surface of the substrate. Physical vapor deposition includes thermal evaporation, electronic beam assisted evaporation, ion-beam and plasma sputtering method, and others. The physical depositions follow the steps described above. In essence, the metal is converted into molecules in the vapor phase and then condensed onto the substrate. Consequently, the deposition is based on molecules and is uniform and very smooth.

Thermal evaporation involves an application of heat to evaporate the film material. This heating and thin film deposition process is carried out in a high-vacuum environment. In the vacuum environment, the vaporization temperature of the material becomes lower, and the vaporized molecules will strike on the substrate surface instead of collision with other molecules according to Knudsen's mean free path. Thermal evaporation has been extensively used in the microelectronics industry. The advantages of thermal evaporation include a high deposition rate and relatively low damage of the substrate because of the low energy of the impinging metal molecules. However, thermal evaporation also has its limitation. For instance, it is very difficult to deposit an alloy film with predetermined composition. In situ cleaning is important for thin film deposition, because a cleaned surface enhances the quality of the film deposited.

Electron beam assisted deposition involves the application of a stream of electrons to evaporate the source material of the deposited film. The electron beam power can be established at a very high value. This tends to enhance the deposition rate. However, the high-energy beam may cause damage on the surface of the substrate.

Deposition of thin film by ion beam and plasma sputtering are very similar, in principle, to electron beam assisted evaporation, with the difference in the energy source applied. Details of each of these physical vapor deposition methods can be found elsewhere.[6]

In addition to physical vapor deposition, thin film can also be formed using chemical vapor deposition (CVD). In general, a CVD involves a gaseous reactant containing the deposited component, and another reactant, often a reducing reagent, in the gaseous phase. Inside the deposition chamber, a reaction involving the gaseous reactants takes place leading to the formation of a nonvolatile solid film on the surface of the substrate.

In CVD, the process involves steps in sequence. This includes the transportation of gaseous molecules from the bulk to the reaction sites, a surface reaction, and the transportation of the by-products into the bulk. Similar to a heterogenous chemical reaction, the slowest step in the process is the rate-limiting step.

Chemical vapor deposition includes various systems, and they are low-pressure CVD (LPCVD), atmospheric pressure CVD (APCVD), plasma enhanced CVD (PECVD), and others. Each type of CVD system has its own advantages and limitations. For instance, in LPCVD, the reactor is usually operated at $\sim$1 torr. Under this condition, the diffusivity of the gaseous species increases significantly compared to that under atmospheric pressure. Consequently, this increase in transport of the gaseous species to the reaction sites and the by-products from the reaction sites in LPCVD will not become the rate-limiting steps. This leads to the surface reaction step to be the rate limiting one.

Atmospheric pressure CVD was first used for CVD in the microelectronics industry. The reactor for APCVD is, in general, relatively simple. However, as compared to LPCVD, APCVD can be mass-transport rate limited. Most of the APCVD used are for low-temperature oxide deposition and epitaxy.

Both LPCVD and APCVD are defined by the pressure regime. There is another type of CVD that is defined by the energy input to heat the reactive surface of the substrate. This heating can be accomplished by resistant, radiofrequency induction, glow discharge (plasma), and photon heating. Each one has its merits and limitations. When radiant heating is involved, both the substrate and the reactor wall will become hot, and the reactor is defined as hot-wall reactor. On the other hand, when the heating is directly on the substrate, the reactor wall does not become hot, and the reactor is defined as cold-wall reactor. It can be appreciated that the geometry of the reactor of CVD is horizontal or vertical or in another form and the control of gaseous flow rate will affect the performance of the CVD process directly.

Deposition of a metallic film onto the silicon dioxide surface requires a good adhesion between the two layers. An enhancement of the adhesion can be accomplished by first depositing a very thin layer of titanium or chromium, 50–100 Å in thickness, then the desired metallic film.

PHOTORESIST PATTERNING AND CHEMICAL WET ETCHING

In microfabrication, photoresist patterning is an important aspect of the definition of microstructures. In this processing, a layer of the photoresist material

is spin-coated onto the surface of the substrate. The coated photoresist layer is then exposed to radiation, which can be ultraviolet light, electrons, or x-rays. This exposure is usually through a mask that defines the pattern and the area to be radiated. After the radiation and in the subsequent development step, the photoresist layer yields the pattern based on the area where the photoresist remains or is removed. Photoresist generally consists of a base resin, a sensitizer, and a solvent, and the solvent serves as the vehicle to form a homogenous mixture for uniform coating. The basic resin is normally a polymer. There are two groups of photoresist, the positive and the negative photoresists. When radiated, the photochemical reaction of a positive photoresist will weaken the exposed polymer, resulting in rupturing or scissoring of the main and side polymer chains. Consequently, the exposed polymer becomes soluble in the developing solution and can be removed from the surface of the substrate. On the other hand, under radiation, the exposed part of a negative photoresist tends to cross-link forming a polymer layer insoluble in the developing solution.

Poly(methylmethacrylate) is a well-known single component positive photoresist, whereas a diazoquinone ester and a phenolic novolak resin is a two component positive photoresist. As anticipated, radiation will render the scission or rupture the main and side chain of the polymer making it soluble in an alkaline ''developing'' solution. Hydroxide solutions such as KOH and tetramethylammonium hydroxide are commonly used as the developer for positive photoresist.

Bis(aryl)azide rubber resist is a common negative photoresist. In a photoreaction, the evolution of nitrogen from the acylazide produces a highly reactive intermediate, nitrene. Through a series of reactions the nitrene intermediate leads to the formation of polymer by cross-linking the resin. The unreacted negative photoresist can then be dissolved in organic solvent, and xylene is commonly used.

Positive and negative photoresists have their own advantages and disadvantages. General consideration of the photoresists includes its adhesion on the substrate, the minimum feature size permitted, the resistance to chemicals, the cost, and others. Most of the commercially available photoresists provide information about the properties and the suitable developers of the photoresists.

After the photoresist layer is spin-coated onto a substrate, before it is exposed to radiation and further development, a soft-bake step often takes place. In this soft-bake step, the photoresist-coated substrate is placed inside a low-temperature oven to be baked at a temperature range approximately 90–100°C for a few minutes to about 30 min. This soft-bake procedure intends to remove the solvent from the photoresist, enhance the adhesion to the substrate, and release the internal stress of the photoresist film. In essence, the soft-bake step intends to provide a photoresist film that will perform well in the following radiation exposure and development process.

It should be recognized now that the mask making, the selection of the polarity, and the relation of the photoresist are intimately related.

The development of the photoresist on a substrate leads to the defining of the actual pattern on the substrate. The material for the actual pattern can be a thin metal film, silicon dioxide, or others. To release this pattern wet chemical etching is used. Obviously, the chemical used in the etching will attack neither the photoresist nor the material that is covered by the photoresist. Chemical etching, also referred to as wet etching, is generally isotropic. Thus, it has limitations in defining features of size less than 3 μm in width. However, wet etching remains an important processing step when the feature size of similar width is more than 3 μm. Wet etching for a specific material requires a specific chemical solvent. This solvent, the etchant, is capable of dissolving and removing the film layer, but not the masking or the substrate material. Wet etching is also a chemical reaction in series. It involves the transport of the etchant to the reacting surface, a surface reaction, and the transport of the products from the etching process to the bulk solution. As anticipated, the slowest step in the process is the rate limiting one.

Silicon, either single crystal or polycrystalline, can be wet etched in a mixture of nitric acid and hydrofluoric acid. This etching reaction is based on the nitric acid reaction with silicon forming silicon dioxide that is then dissolved by hydrofluoric acid. Different compositions of the HNO_3 and HF mixtures yield different etching rates. Both water and acetic acid have been used for the silicon etching mixture. The etching rate of silicon is also related to the crystalline structure of the silicon plane. This orientation etching is termed anisotropic etching. It is well recognized that etching at a <111> plane will create a V-shaped profile with an angle of 54.7°. For etching processes that are independent of the crystalline orientation, bulk micromachining techniques, such as reactive ion etching and plasma etching, can be used. These processes are well established and are described extensively in three-dimensional bulk micromachining literature.

Wet chemical etching is also a commonly used process to pattern the metallic film structure. In these cases, a suitable solvent for the metallic film is used to dissolve or etch away the exposed area of the metallic film. This wet etching technique has been used for various metallic films, such as gold, copper, aluminum, and others. However, there are cases where chemical etching of the metal can be difficult. For instance, the etching of the platinum film is not easy because of the required etchant for the platinum film. One of the patterning processes to overcome the potential

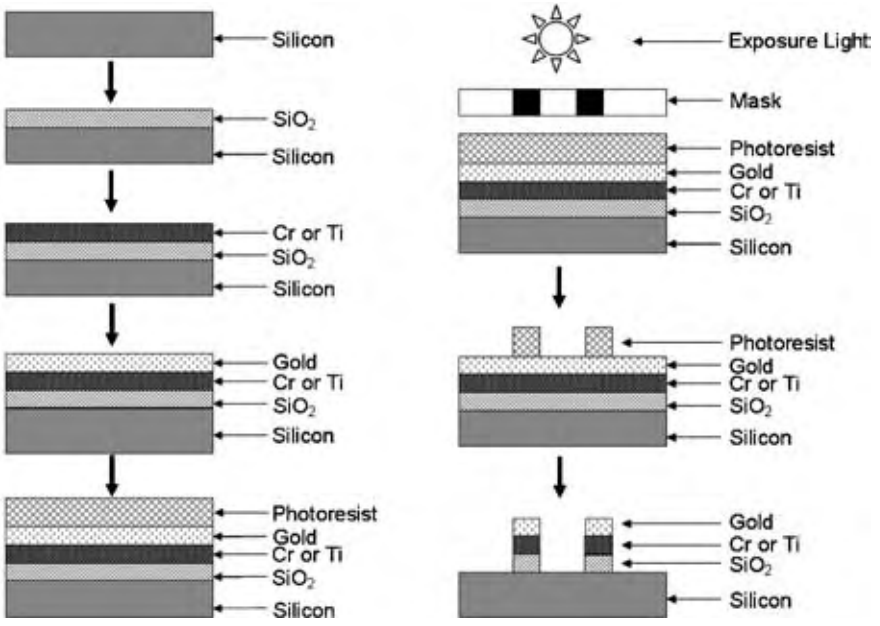


Fig. 3 Processing steps for the deposition and patterning of a thin film metal structure. (The thicknesses of the metal layers are not proportionally scaled.)

problems in wet etching is the lift-off technique. Lift-off is considered to be a technique for forming a pattern on a substrate surface by an additive process as opposed to the removal (or subtractive) process in the etching process.[1–3] In the lift-off process, the photoresist layer is first spin-coated and patterned on the surface of the substrate using lithographic and mask-based optical printing techniques. However, the photoresist is inverse-patterned in this lift-off process. The film, metallic or otherwise, to be patterned is then deposited over the inverse-patterned photoresist layer as well as the open area in the pattern. The substrate is then placed into a solvent that will dissolve or remove the photoresist. Consequently, any film material deposited on top of the photoresist will be lifted off leaving the patterned portion where the film material is deposited, on the surface of the substrate. As anticipated, there are advantages and disadvantages in the lift-off process in patterning. The choice of a proper process in patterning, as in other microfabrication processes, depends on the needs, the materials selected, and the required dimensions, and available facilities. Fig. 3 shows a typical thin film metallization processing sequence. This processing sequence summarizes the various steps in the thin film microfabrication.

CONCLUSIONS

In conclusion, silicon-based microfabrication processes prove to be effective in producing miniature size, highly reproducible microstructures at modest manufacturing cost. This provides opportunities for the development of microsensors for chemical processing control, microreactors, and microfluidics, and many other unique applications. The combination of the chemical engineering fundamentals, transport phenomenon and reaction kinetics, and the microfabrication technology provides an excellent opportunity and technical challenge in numerous research and development endeavors in the years to come.

REFERENCES

1. Madou, M.J. *Fundamentals of Microfabrication*, 2nd Ed.; CRC Press: Boca Raton, FL, 2002.
2. Wolf, S.; Tauber, R.N. *Silicon Processing for the VLSI Era, Volume 1—Process Technology*; Lattice Press: Sunset Beach, CA, 1986.
3. Elliot, D.J. *Integrated Circuit Fabrication Technology*; McGraw Hill: New York, 1982.
4. Deal, B.E.; Grove, A.S. General relationship for the thermal oxidation of silicon. J. Appl. Phys. **1965**, *36*, 3770.
5. Lambrechts, M.; Sansen, W. *Biosensors: Microelectro-Chemical Devices*; The Institute of Physics Publishing: Philadelphia, PA, 1992.
6. Hill, R.J. *Physical Vapor Deposition*; Temescal: Berkeley, CA, 1976.

Microgravity Processing of Materials

Robert Naumann
University of Alabama, Huntsville, Alabama, U.S.A.

INTRODUCTION

Why study materials science in space? Given the high cost of getting to and from low Earth orbit, the actual manufacture of most materials in space on a commercial basis seems unlikely. A principal justification for carrying out materials science experiments in space is the use of the virtual absence of the effects of gravity as a tool to obtain a more fundamental understanding of the processing of materials with the objective of being able to develop improved processing strategies for use on Earth.

Much of what goes on in a multicomponent melt is affected by gravity, but some of the more subtle interfacial effects are not. Yet these interfacial effects, which are poorly understood because they are masked by gravitational effects, can play important roles in many terrestrial processes. Using microgravity as a tool to sort out the nongravitational from the gravitational effects may add to our understanding of how these interfacial processes operate, which is needed to control their effects.

With the computational capability that exists today, it is possible in theory to model a complete solidification process in a complex mold, taking into account the heat and mass flow, culminating in the prediction of the microstructure of the final casting at critical places. However, to do this, the basic laws or theories governing the development of the microstructure must be known along with accurate values for the thermophysical properties of the constituents. Establishing the physical basis and workable models for the solidification process has, over the last half-century, transformed metallurgy from an industrial art, based on empiricism, to a more exact science. However, because of the complicating effects of convection, many of these theories or models in use today are based on the assumption of no convective flows. It is understood that many of these models do not apply exactly, but they are assumed to be basically correct and attempts are then made to "fix them up" by adding the effects of convection. However, most of these theories and models have never been tested in the absence of convection, so various subtleties may have been overlooked. The ability to experiment in microgravity provides opportunities to test some of these theories and models to make sure they are valid or at least to define the limits of their validity. This entry reviews some of the significant experiments that were performed during the Spacelab series of flights involving NASA, the European Space Agency (ESA), and the National Space Development Agency of Japan (NASDA). The types of experiments can be logically categorized under the topics of Alloy Solidification, Measurement of Thermophysical Properties, Undercooling Experiments, and Crystal Growth Experiments. Readers are encouraged to refer to the citations for more details about these experiments and their findings.

ALLOY SOLIDIFICATION

When one attempts to solidify a multicomponent system from the melt, difficulties arise. The solute atoms generally do not fit into the lattice as easily as the host atoms forming the matrix, and segregation results. The melt containing the rejected atoms will have a different density from the bulk liquid, resulting in solutally driven convection that causes the final solid to have a nonuniform composition on a macroscopic scale (macrosegregation). If a single crystal with uniform composition is required, a thermal gradient of sufficient magnitude must be applied to stabilize the solidification front or it will break down from plane front to cellular or dendritic growth. In most industrial processes that do not require single crystals, no attempt is made to stabilize the interface against breakdown, so the resulting dendrites form the microstructure of the casting. Therefore, it is important to understand how the structure of these dendrites depends on the processing conditions.

The microstructure will coarsen with time at high temperatures, meaning the grains or other microstructural features will grow larger, which alters the mechanical properties. Any particles intentionally or unintentionally dispersed in the melt will generally have a different density and will tend to either sink or float. Particle behavior is also affected by the interfacial energy between their surface and the melt, which determines if they are engulfed by an advancing solidification front or pushed ahead of it.

The alloy solidification experiments that have been performed on Spacelab missions generally fall into four categories: 1) interfacial stability; 2) evolution

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039235

of the microstructure during and after solidification; 3) Liquid phase sintering; and 4) particle/solidification interactions.

Interfacial Stability

As solidification progresses in a binary or multicomponent system, the rejected solute builds up in front of the solidification interface, which has the effect of lowering the freezing temperature at the interface because some of the component with the higher freezing temperature has already been removed. However, the bulk melt away from the solidification interface has the original composition, which has a higher freezing point. Therefore, the freezing point of the melt rises from the lower value at the growth front to the higher value of the bulk melt. Unless the imposed thermal field in front of the solidification interface is everywhere higher than the local freezing point of this melt, the melt is said to be constitutionally undercooled (or supercooled), which can lead to an interfacial instability. If a small fluctuation displaces a portion of the growth interface ahead and it finds the local freezing point to be lower than the local temperature, it will continue to advance. Thus the interface will break down, first into a cellular pattern if the constitutional undercooling is small, or into long finger-like projections if the undercooling is larger. The sides of these projections will also break down to form secondary arms, which in turn can break down to from tertiary arms. The resulting structure resembles a fir tree, hence the term ''dendrite.'' Where single crystals of uniform composition are required, the interface can be stabilized by applying a sufficient thermal gradient at the growth interface to keep the local temperature above the local freezing point and prevent interfacial breakdown.

A simple constitutional supercooling criterion (known as the CS criterion), which predicts the ratio of the gradient required to stabilize the interface moving at a particular growth velocity for a given solidification system, was developed by Rutter and Chalmers in 1953.[1] In 1964, Mullins and Sekerka[2] developed a more rigorous theory based on a stability analysis that included the liquid–solid interfacial energy, which can provide a stabilizing effect on the interface. Like most theories concerning solidification phenomena, it was necessary to assume no convection to simplify these analyses.

A unique and highly sophisticated apparatus for studying details of the solidification process was developed by Favier and co-workers at the French Centre d'Etude Nucléaire, Grenoble (CENG), under a cooperative program between NASA and the French Centre National d'Etudes Spatiales (CNES) or National Space Agency and the French Commissariat á l'Energie Atomique (CEA) or Atomic Energy Commission. The official name is Materials for the Study of Interesting Phenomena of Solidification on Earth and in Orbit or MESPHISTO. The middle of a sample is melted using two furnaces. One solidification front is kept stationary while the other is moved back and forth to create a solidification front that can be moved at different velocities. By measuring the differential Seebeck voltage between the stationary and moving interfaces, the kinetic undercooling can be determined as a function of growth velocity. At the freezing point, a solid will remain in equilibrium with its melt indefinitely. This kinetic undercooling is the driving force for continued solidification. The kinetic undercooling will be small for plane front solidification, but since additional interfacial energy must be provided as the plane front begins to break down, the kinetic undercooling must get larger. Thus, the transition from plane front solidification to cellular growth can be observed as a change in the Seebeck voltage, and the critical growth velocity where the plane front interface begins to break down can be determined accurately. The MEPHISTO instrument, flown on four Spacelab missions, provided the first opportunities to perform a critical test of the Mullins–Sekerka theory as well as to explore other important phenomena involved in the solidification process.

Favier used the USMP-1 opportunity to explore the interfacial breakdown in Bi-doped Sn, which, like most metals, solidifies as a plane front with little kinetic undercooling. His U.S. co-investigator, Abbaschian, investigated interfacial stability on the other side of the phase diagram; i.e., Sn-doped Bi, which solidifies with a faceted interface. The purpose was to test the extension of the Mullins–Sekerka stability criterion to include the effects of anisotropy, which acts to stabilize the interface against breakdown into cellular and dendritic growth.[3,4]

Microstructure Evolution

The strength and other properties of an alloy depend on its microstructure, which is characterized by the size, orientation, and composition of the grains and other features that make up the solid. One of the main tasks of a materials scientist is to design solidification processes to produce the microstructure that will give the material the desired properties. To accomplish this it is necessary to understand how this microstructure develops as a function of processing parameters.

Dendrite formation

Since the microstructure and hence the mechanical properties of the casting are largely controlled by the

structure of the dendrites, it becomes important to know how dendrite growth depends on processing parameters so that the desired microstructure can be engineered.

Glicksman and co-workers carried out a series of precisely controlled dendritic growth experiments on three Spacelab missions using succinonitrile, which solidifies body-centered cubic, and pivalic acid, which solidifies face-centered cubic. Both systems have unusually low entropies of fusion, more typical of metals than of organics and are widely used as transparent metal analogs. One of the governing factors in the growth of these thermal dendrites is the heat flow from the dendrite to the surrounding melt. Ivantsov[5] obtained an exact solution to the conductive heat flow problem for a parabolic shaped dendrite, which relates the product of growth rate and tip radius to the undercooling. However, from this simple analysis, there seems to be no fundamental relationship between the tip radius R and the growth velocity V.

The question becomes, how does nature select a unique operating state? Experimental observations of dendrite solidification in pure systems suggest that VR^2 is either a constant for a specific material, or a weakly varying function of the undercooling. A large body of terrestrial data has been taken on several systems, but convection effects, especially in the crucial region of low undercoolings where the growth rate is comparable to the convective flow velocities, have not been able to provide an adequate test of the selection rules governing this process (Fig. 1). The establishment of such relationships was the motivation behind this set of flight experiments.[6,7]

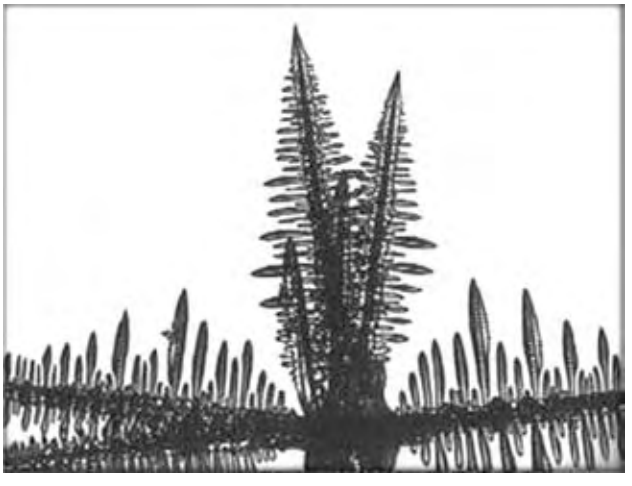


Fig. 1 Thermal dendrites grown in a microgravity experiment as part of the Isothermal Dendrite Growth Experiments (IDGE) carried out by the Renasslear Polytechnic Institute team headed by Glicksman. (Taken from the IDGE home page on http://www.rpi.edu/locker/56/000756/.)

Coarsening

Since the yield strength of a metal is inversely proportional to the square root of the average grain size, coarsening is of major importance in the evolution of the microstructure of alloys. This is particularly true in precipitation or dispersion hardened alloys because the added strength provided by the dispersed phase declines rapidly if the particles grow past a critical size. Coarsening is driven by the excess interfacial energy in a finely dispersed material, which could be reduced if fewer larger particles were present. The melting point of a small particle is lower than that of a larger particle of the same composition because of its excess surface energy (Gibbs-Thompson effect), so the smaller particles dissolve to feed the growth of larger particles in a melt. Coarsening also occurs in fully solidified systems through solid state diffusion.

This process was first recognized by Ostwald[8] and is known as Ostwald ripening. The mathematical details were worked out independently by Lifshitz and Slyozov[9] and by Wagner[10] and is known as the LSW theory. However, this theory is based on a mean field approximation and is restricted to low volume fraction systems. Voorhees and coworkers extended the LSW theory to finite volume fraction systems and conducted a series of flight experiments designed to test this and similar theories.[11–13]

Liquid Phase Sintering

Liquid phase sintering (LPS) is a widely used process for forming composites containing refractory particles such as tungsten, rhenium, or various carbides in a metal matrix. Sintered products include cutting tools, bearings, contact points, and other irregularly shaped parts where it is desirable to combine extreme hardness with the toughness and thermal or electrical conduction of the metal matrix. The refractory particles are combined with the metal matrix powder and isostatically pressed and heated to above the melting point of the matrix phase. If proper attention is paid to the wettability of the refractory particles, the molten host metal will infiltrate between the grains of the solid particles and envelop them. There are some obvious gravity effects because of the large difference in densities often encountered between particle and host phase. Consequently, these effects restrict the process to large volume fractions of the solid phase since the solid particles will essentially have to support themselves during the process. Even under these circumstances, there are differences in the particle size and morphology between the top and bottom of the specimen due to the gravity-imposed hydrostatic pressure.

German and co-workers conducted liquid phase sintering experiments during three Spacelab missions using W particles in a Fe–Ni matrix with the amount of W ranging from 78–98 wt% W in 5 wt% increments. The major results from this series of experiments are universal models for coarsening, slumping and distortion, and grain agglomeration.[14,15]

Particle/Solidification Front Interactions

Small ceramic particles are sometimes added to metals to block the motion of dislocations (dispersion hardening) or for flux pinning in type II superconductors. In the preparation of composite materials, it is important to know how such particles interact with the solidification front. If the particle is not wetted by the melt, intermolecular forces at the advancing solidification will tend to repel the particle. These forces are pitted against inertia and drag forces that tend to engulf the particle. There have been a number of attempts to model this process. It is generally accepted that for a particular system there is a critical velocity below which the particle will be pushed ahead of the solidification front, and above which it will be engulfed by the advancing solid. Buoyancy and convective flows complicate the picture in normal gravity and it is important to be able to separate these effects from the more fundamental interactions that take place at the solidification front. As a result, a variety of experiments were conducted on Spacelab missions to develop a better understanding of particle-solidification front interactions.

A theory developed by Shangguan, Ahuja, and Stefanescu to predict the critical velocity was validated for the case of zirconia particles being pushed by Al on a later Spacelab mission.[16,17]

MEASUREMENT OF THERMOPHYSICAL PROPERTIES

Not much is known about the thermophysical properties of liquid metals, especially the transport properties such as chemical and thermal diffusivities. The existing data are sparse and the scatter makes it difficult to make an accurate determination of the temperature dependency of these properties. This situation was the motivation for Froberg's experiment on Spacelab-1 in which he measured the temperature dependence of the self-diffusion of Sn from 240°C to 1250°C. He found that the diffusion coefficients were 30–50% lower than the accepted values and seemed to follow a T^2 dependence as opposed to the Arrhenius behavior observed in solid state diffusion.[18] These results prompted a number of experiments in subsequent missions designed to measure diffusion coefficients of liquid metals. Some of these experiments found a power law dependence on temperature, with the exponent varying around 2 for different materials, while others found a better fit with an Arrhenius model. However, all of the microgravity experiments consistently measured diffusion coefficients that were 30–50% lower than the accepted values, suggesting that all existing transport data for liquid metals may be contaminated by unwanted convection.[19–21] A recent analysis demonstrated the difficulty of eliminating convective transport in such measurements in normal gravity.[22]

UNDERCOOLING EXPERIMENTS

A space electromagnetic levitation facility designated as TEMPUS (Tiegelfreies Elektromagnetisches Prozessieren Unter Schwerelosigkeit) was developed by the Institute for Space Simulation, Cologne, Germany. It uses a quadrapole coil for sample positioning and a dipole coil for sample heating. Since only very small positioning forces are required in a microgravity environment, electromagnetically driven flows from positioning can be minimized. The sample chamber can be evacuated, or backfilled with an inert gas to reduce evaporation of samples with high vapor pressure. A team of U.S. and German investigators, using some innovative noncontact instrumentation was able, for the first time, to measure a variety of thermophysical properties including surface tension, viscosity, thermal conductivity, and enthalpy in the undercooled state of a number of alloys that are candidates for metallic glasses.[23–26] Other phenomena such as quasicrystal formation and dynamic nucleation were also explored and ferromagnetic magnetic ordering was in the liquid state of $Co_{80}Pd_{20}$ when it was cooled below its Curie temperature.[27,28]

CRYSTAL GROWTH EXPERIMENTS

Growth from the Melt

Macrosegregation has always been a problem in the growth of alloy-type multicomponent single crystals from the melt. Even in systems that are not subject to thermosolutal convection and can be stabilized by vertical Bridgman growth by placing hot over cold, radial thermal gradients can drive convective flows that result in radial as well as axial compositional inhomogeneities. A number of early flight experiments attempted to avoid this problem by growing such

crystals in microgravity under the assumption that by reducing the effects of gravity by some six orders of magnitude, the resulting residual convective flows would produce negligible compositional redistribution.[29] After these early experiments failed to produce the desired results, a serious analysis was undertaken of the effects of small convective flows. It was shown that a transverse quasi-steady acceleration of less than 1μg was sufficient to cause significant macrosegregation and that it would be necessary to align the thermal gradient along the residual acceleration vector to provide a small stabilizing force if unwanted solute redistribution is to be avoided.[30] The residual acceleration in the Spacelab was primarily due to gravity gradient effects, which occur when the experiment is some distance from the center of mass of the spacecraft. Beginning with the first United States Material Laboratory (USML-1) mission in 1992, all subsequent low-gravity emphasis Spacelab missions were flown with the orientation of the Space Shuttle controlled to provide a definable nearly constant residual acceleration vector during at least a portion of the mission.[31,32]

Larson designed an experiment to investigate how gravity might influence the formation of defects during the growth of $Cd_{0.95}Zn_{0.05}Te$ on the USML-1 mission. A headspace was left in the growth ampoule to accommodate thermal expansion. There was an unanticipated 0.5 μg lateral acceleration from the venting on the Space Shuttle that had the fortuitous effect of nudging the melt against one wall of the growth ampoule and leaving the opposite side of the melt free of wall contact. The space-grown crystal was found to have a dislocation density $\sim 100\times$ lower than the control sample grown in the same furnace under normal gravity. The X-ray rocking curve at full-width half-maximum, (an indication of structural perfection) was reduced by a factor of 2.[32] Larson attributed these improvements in crystal quality due to lack of strain from the virtual absence of hydrostatic pressure and minimum wall contact and adhesion, resulting in very low stress being exerted on the crystal during growth and postsolidification cooling.

Larson repeated this experiment on USML-2 using a novel ampoule design that would minimize wall contact with the sample. A second sample had a spring-plunger system that forced the sample to fill the ampoule, thereby assuring wall contact. Postflight analysis confirmed the lack of strain effects seen previously and found that twin formation was virtually absent in the region grown without wall contact; whereas, the sample in the spring-loaded ampoule was highly strained at the exterior and heavily twinned.[33,34] This set of experiments dramatically showed how wall effects can influence defect formation in crystals grown from the melt.

A number of crystals were grown in microgravity using the floating zone process. The low Bond number (ratio of hydrostatic forces to capillary forces) allows much larger zones than are possible in normal gravity. Surface tension-drive convection (Marangoni flows) was found to produce time-dependent flows resulting in dopant striations.[35] It was found that coating the sample with thin oxide films could eliminate such flows[36] and silicon, gallium arsenide, and other compound semiconductors were grown with low dislocation densities.[37,38]

Solution Growth

Growth of small molecule crystals

Lal developed a novel method for growing crystals from aqueous solution by extracting heat through the crystal using a cooled sting.[39] This method was used on the International Microgravity Laboratory (IML) missions to grow triglycene sulfate (TGS), a pyroelectric material used for detection of far infrared radiation. A smooth transition from seed to new growth was seen without the veil of dislocations (''ghost of the seed'') that usually surround the seed crystal during growth on the ground. The TGS crystal grown on the IML-2 mission was examined with high resolution monochromatic synchrotron X-radiation diffraction imaging using the National Synchrotron Light source at the Brookhaven National Laboratory. The X-ray topographic images indicated an extraordinary crystal quality.[40] The only inclusions were due to the incorporation of polystyrene particles intentionally inserted in the growth solution to study the fluid motion in low-g. The detectivity (D^*) of the space-grown crystal was found to be significantly higher than the seed crystal, and the loss tangent was reduced from 0.12–0.18 for the seed to 0.007 for the space-grown material.[41]

Growth of biomolecular crystals

A large number of proteins and other biomolecular crystals have been grown in space. The first Shuttle flight after the Challenger accident (STS-26) yielded crystals of 4 different proteins that were shown to have better diffraction resolution than the best crystals of these proteins that had ever been grown on Earth.[42] This feat was even more remarkable, considering that the crystals produced in only a handful of space experiments were compared with the best crystals of these particular proteins that had been grown in

thousands of experiments by the world's most qualified researchers whose professional success depends heavily on obtaining the molecular structure from the X-ray diffraction data from such crystals. These results formed the impetus for the major effort in the growth of biomolecular crystals for structural analysis by X-ray crystallography that was sponsored by NASA and ESA.

Not all space-grown biomolecular crystals were superior to those grown on Earth, but there were a number of cases of larger crystals, improved diffraction resolution, and lower mosasicity to merit a serious study among theorists to investigate how gravity effects influence the perfection of such crystals.[43] When a crystal grows on Earth, the nutrient next to the growth interface is depleted and the less dense solvent rises, bringing more solute to the growth site. Since growth kinetics are slow for most biomolecular crystals, the convective flows always bring nutrient to the crystal faster than it can be incorporated into the lattice. Thus, the growth in normal gravity is generally kinetics limited. In space, the solute must diffuse in from the surrounding region, which greatly reduces the transport to the growing crystal. One of the arguments for why protein crystals grow better in microgravity is that diffusion-limited growth slows the rate

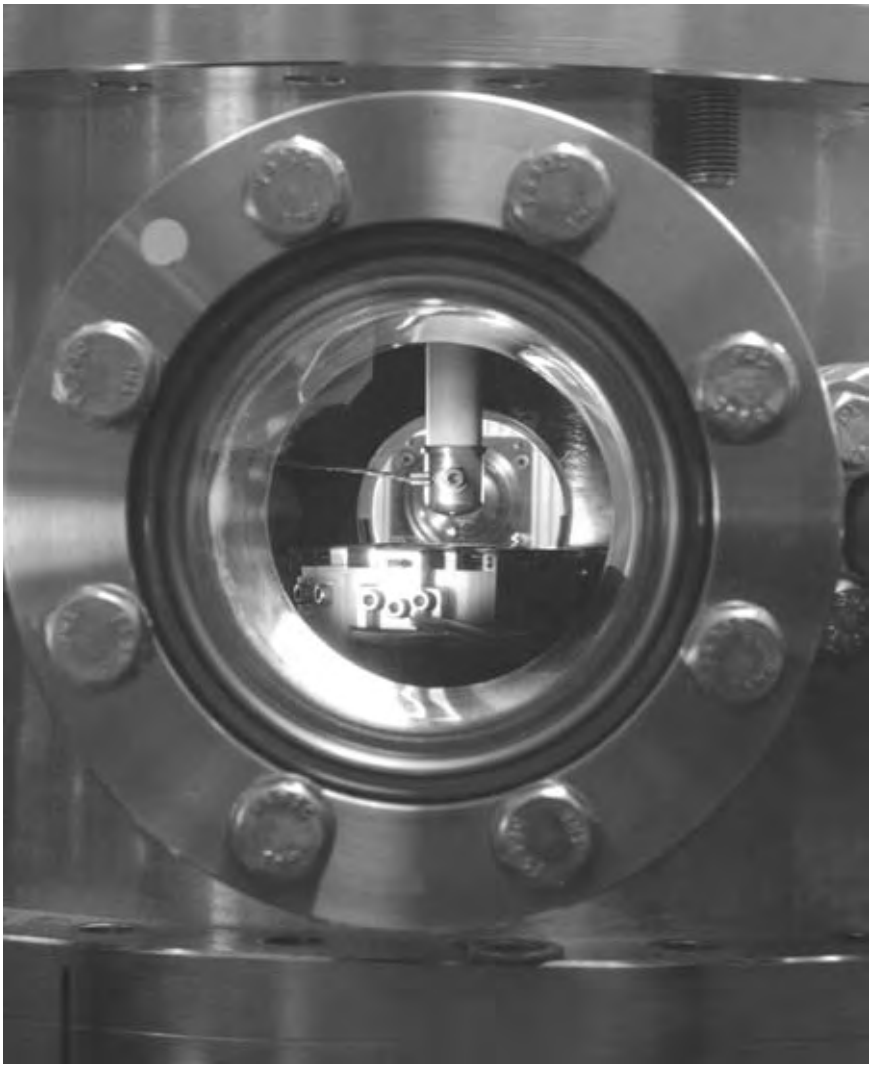


Fig. 2 A specimen levitated in the MSFC electrostatic levitator facility. A version of this device has recently been installed in the MUCAT sector of the beamline at the Advanced Photon Source for realtime studies of solidification of high temperature materials. Thermophysical properties measurements made on $Zr_{57}Nb_5Ni1_{2.6}Al_{10}Cu_{15.4}$ and $Ti_{34}Zr_{11}Al_{7.5}Cu_{47}Ni_8$ in the undercooled state helped William Johnson at Liquidmetal Technology design metal glass-forming systems with cooling rates slow enough to form in bulk quantities. Dr. Rick Weber, Containerless Processing, Inc., also used the MSFC electrostatic levitator to develop a new class of laser host glasses consisting of rare earth oxides and alumina. (Photo courtesy of NASA.) *(View this art in color at www.dekker.com.)*

at which nutrient is transported to the growing crystal so that growth becomes transport limited rather than kinetics limited. This gives surface kinetics a chance to go to equilibrium, allowing the admolecules to reach a configuration that minimizes the surface free energy and thus produce greater order in the lattice. Also, the diffusion field surrounding the growing crystal can act as a molecular weight filter that tends to exclude foreign impurities with higher molecular weights or oligomers of the growth molecules.[44]

APPLICATIONS

Space experiments have validated many of the theories and models that form the basis of our knowledge of alloy solidification and have begun to demonstrate the range of their applicability. Several lessons learned from materials experiments in space have been put to practice.

One of the more important ''real world'' problems has to do with attempts to dispersion harden superalloy single crystal gas turbine blades by incorporating very small (submicron) oxide particles during the growth process to increase their creep resistance. A uniform dispersion can be achieved by powder metallurgical techniques, which can then be densified by hot isostatic pressing (HIP). However, when the blade is melted so that it can be directionally solidified into a single crystal, the particles tend to agglomerate and are not uniformly incorporated into the superalloy matrix. The solidification velocities required to achieve plane front solidification are generally below the critical velocity for engulfment of such small particles. At higher solidification rates, the particles tend to be pushed laterally by the dendrites and wind up being clumped together, trapped in the last-to-freeze interdendritic fluid. This problem prompted several flight experiments by industrial firms trying to sort out gravitational effects from nongravitational effects that remain as barriers to developing this process.[45,46]

Much of the computational fluid dynamics modeling that was developed to analyze the effects of residual accelerations on solidification have been adopted along with magnetic damping to help control unwanted solute redistribution from convective flows in processes used on Earth. Researchers at NASA/MSFC picked up on the idea of using detached growth to minimize crystal defects. By carefully balancing the pressure in the growth ampoule against capillarity forces, they have been able to grow $Ge_{1-x}Si_x$ with minimal wall contact in the laboratory.[47]

Structural biologists, who must crystallize complex biomacromolecules to obtain structural information from X-ray diffraction, have learned to mimic growth in microgravity by growing such crystals in X-ray capillaries. Capillaries with inside diameters of 200–300 μm restrict flows to the same order of those in microgravity and crystals can be grown under diffusion-limited conditions.[48]

Given the realization that the few existing data on transport properties of liquid metals may be in serious error, care must be taken in using such data to design processes where these properties become critical and efforts must be made to refine the available data. The pioneering work that was done using the TEMPUS electromagnetic levitator on the 1997 Material Science Laboratory (MSL-1R) mission demonstrated the usefulness of being able to levitate a small sample quiescently. The desire to continue such experiments on the ground led to the development of the electrostatic levitator (ESL), now in operation at the NASA Marshall Space Flight Center (Fig. 2). The ESL is restricted to much smaller particles than can be levitated in microgravity and the types of data that can be extracted are somewhat limited, but data obtained from the ELS combined with the TEMPUS have been used to develop a new class of laser host glasses,[49] a commercial line of bulk metallic glasses, (Fig. 3)[50] and have provided new insight into the nucleation of metals from the melt.[51]

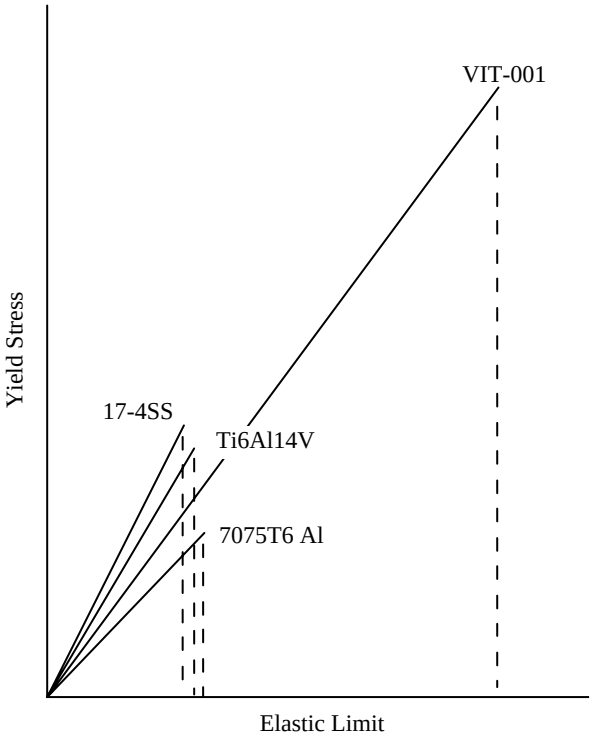


Fig. 3 Elastic performance of bulk metallic glass VIT-001 compared with stainless steel, aluminum, and titanium alloys.

CONCLUSIONS

The rational behind some of the more significant materials processing experiments that have been carried out in space and their results have been presented along with a few key references that describe the experiments and their results in much more detail than could be presented here. The list of experiments and the publications that have resulted from them are by no means exhaustive. More detailed results from the U.S. experiments can be obtained from the Mission Science Review reports available from the Marshall Space Flight Center, Huntsville, AL. The European results are available in the EEA (Erasmus Experiment Archive) http://spaceflight.esa.int/eea/.

REFERENCES

1. Rutter, J.W.; Chalmers, B. A mathematical analysis of solute redistribution during solidification. Can. J. Phys. **1953**, *31*, 15.
2. Mullins, W.W.; Sekerka, R.F. Morphological stability of a particle growing by diffusion or heat flow. J. Appl. Phys. **1963**, *34*, 323–329.
3. Favier, J.J.; Garandet, J.P.; Rouzaud, A.; Camel, D. Mass transport phenomena during solidification in microgravity; Preliminary results of the first Mephisto flight experiment. J. Cryst. Growth. **1994**, *140* (1–2), 237–243.
4. Abbaschian, R.; Gokhale, A.B.; Allen, D.B. A study of directional solidification of faceted Bi-Sn alloys in microgravity. In *Solidification Science and Processing*; Ohnaka, I., Stefanescu, D.M., Eds.; TMS (The Minerals, Metals and Materials Society): Warrendale, PA, 1996; 73–84 pp.
5. Ivantsov, G.P. Temperature field around spherical, cylindrical and needle-shaped crystals which grow in supercooled melt. Dokl. Akad. Nauk. SSSR. **1947**, *58*, 567 (in Russian).
6. Glicksman, M.E.; Koss, M.B.; Winsa, E.A. Dendritic growth velocities in microgravity. Phys. Rev. Lett. **1994**, *73*, 573–576.
7. Glicksman, M.E.; Koss, M.B.; Bushnell, L.T.; LaCombe, J.C.; Winsa, E.A. Dendritic growth of succinonitrile in terrestrial and microgravity conditions as a test of theory. ISIJ Int. **1995**, *35* (6), 604.
8. Ostwald, W. Studien uber die Bildung und Umwandlung fester Korper. Z. Phys. Chem. **1897**, *22*, 289.
9. Lifshitz, I.M.; Slyozov, V.V. The kinetics of precipitation from supersaturated solid solutions. J. Phys. Chem. Solids **1961**, *19*, 35.
10. Wagner, C.Z. Theorie der Alterung von Niederschlägen durch Umlösen, Z. Elektrochem. **1961**, *65*, 581.
11. Ratke, L.; Uffelmann, D.; Bender, W.; Voorhees, P.W. Theory of Ostwald ripening due to second-order reaction. Scripta Metall. et Mater. **1995**, *33*, 363.
12. Kuehmann, C.J.; Voorhees, P.W. Ostwald ripening in ternary alloys. Metall. Mater. Trans. A. **1996**, *27A*, 937–943.
13. Seyhan, I.; Ratke, L.; Bender, W.; Voorhees, P.W. Ostwald ripening of solid-liquid Pb-Sn dispersions. Metall. Mater. Trans. A. **1996**, *27A*, 2470–2478.
14. Raman, R.; German, R.M. A mathematical model for gravity-induced distortion during liquid phase sintering. Metall. Mater. Trans. A. **1995**, *26A*, 653–659.
15. German, R.M.; Liu, Y.; Griffo, A.A. Gravitational effects on grain coarsening during liquid-phase sintering. Metall. Mater. Trans. A. **1997**, *28A*, 215–221.
16. Shangguan, D.; Ahuja, S.; Stefanescu, D. An analytical model for the interaction between an insoluble particle and an advancing solidification front. Metall. Mater. Trans. A. **1992**, *23A*, 669–680.
17. Stefanescu, D.M.; Juretzko, F.R.; Dhindaw, B.K.; Catalina, A.; Sen, S.; Curreri, P.A. Particle engulfment and pushing by solidifying interfaces: Part II. Microgravity experiments and theoretical analysis. Metall. Mater. Trans. A. **1998**, *29A*, 1697–1706.
18. Frohberg, G.; Kraatz, K.H.; Wever, H. Self-diffusion of Sn-112 and Sn-124 in liquid Sn. Proceedings of the 5th European Symposium, Materials Science under Microgravity, Results of Spacelab-1, ESA SP-222, 1984; 201–206.
19. Mathiak, G.; Griesche, A.; Kraatz, K.H.; Frohberg, G. Diffusion in liquid metals. J. Non-Cryst. Solids **1996**, *205–207* (1), 412–416.
20. Yoda, S. Measurement of diffusion coefficient by shear cell method. In *Microgravity Science Laboratory (MSL-1) Final Report*; Robinson, M.B., Ed.; NASA/CP-1998-208868, 1998; 86–92.
21. Itami, T.; Aoki, H.; Kaneko, M.; Uchida, M.; Sisa, A.; Amano, S.; Odawara, O.; Masaki, M.; Ooida, T.; Oda, H.; Yoda, S. Diffusion of liquid metals and alloys: The study of self-diffusion under microgravity in liquid Sn in the wide temperature range. In *Microgravity Science Laboratory (MSL-1) Final Report*; Robinson, M.B., Ed.; NASA /CP-1998-208868, 1998; 60–69.
22. Alexander, J.I.D.; Ramus, J.F.; Rosenberger, F. Numerical simulations of the convective contamination of diffusivity measurements in

liquids. Microgravity Sci. Technol. **1966**, *IX/3*, 158–162.

23. Schaefers, K.; Kuppermann, G.; Thiedemann, U.; Qin, J.; Frohberg, M.G. New variant for measuring the surface tension of liquid metals and alloys by the oscillating drop method. Int. J. Thermophysics **1966**, *175*, 1173–1179.
24. Schaefers, K.; Qin, J.; Rosner-Kuhn, M.; Frohberg, M.G. Mixing enthalpies of liquid Ni-V, Ni-Nb and Ni-Ta alloys measured by levitation alloying calorimetry. Can. Metall. Quart. **1996**, *35* (1), 47.
25. Wunderlich, R.K.; Lee, D.S.; Johnson, W.L.; Fecht, H.-J. Noncontact modulation calorimetry of metallic liquids in low Earth orbit. Phys. Rev. B. **1997**, *55* (1), 26–29.
26. Schroers, J.; Holland-Moritz, D.; Herlach, D.M.; Grushko, B.; Urban, K. Undercooling and solidification behaviour of a metastable decagonal quasicrystalline phase and crystalline phases in Al-Co. Mater. Sci. Eng. A. **1997**, *226* (22), 990–994.
27. Busch, R.; Bakke, E.; Johnson, W.L. Viscosity of the supercooled liquid and relaxation at the glass transition of the $Zr_{46.75}Ti_{8.25}Cu_{7.5}Ni_{10}Be_{27.5}$ bulk metallic glass forming alloy. Acta Materialia. **1998**, *46/13*, 4725–4732.
28. Holland-Moritz, D.; Schroers, J.; Herlach, D.M.; Grushko, B.; Urban, K. Undercooling and solidification behaviour of melts of the quasicrystal-forming alloys Al-Cu-Fe and Al-Cu-Co. Acta Materialia **1998**, *46* (5), 1601–1615.
29. Hofmeister, W.H.; Bayuzick, R.J.; Hyers, R.; Trapaga, G. Cavitation-induced nucleation of zirconium in low earth orbit. Appl. Phys. Lett. **1999**, *74/48*, 2711.
30. Naumann, R.J. An analytical model for transport from quasi-steady and periodic accelerations on spacecraft. Int. J. Heat Mass Transfer **2000**, *43*, 2917–2930.
31. Gillies, D.C.; Lehoczky, S.L.; Szofran, F.R.; Watring, D.A.; Alexander, H.A.; Jerman, G.A. Effect of residual accelerations during microgravity directional solidification of mercury cadmium telluride on the USMP-2 mission. J. Cryst. Growth **1997**, *174*, 101–107.
32. Gillies, D.C.; Lehoczky, S.L.; Szofran, F.R.; Larson, D.J., Jr.; Su, C.-H.; Sha, Y.-G.; Alexander, H.A. Bulk growth of II-VI crystals in the microgravity environment of USML-1. Proc. of SPIE — Int. Soc. Opt. Eng. **1993**, *2021*, 10–21.
33. Chung, H.; Raghothamachar, B.; Dudley, M.; Larson, D.J., Jr. Synchrotron white beam x-ray topography characterization of structural defects in microgravity and ground-based CdZnTe crystals. Proc. of SPIE — Int. Soc. Opt. Eng. **1996**, *2809*, 45–56.
34. Larson, D.J., Jr.; Dudley, M.; Chung, H.; Raghothamachar, B. Characterization of Zn-alloyed CdTe compound semiconductors processed in microgravity on USML-1 and USML-2. Adv. Space Res. **1998**, *22*, 1179–1188.
35. Croell, A.; Müeller-Sebert, W.; Nitsche, R. Critical Marangoni number for the onset of time-dependent convection in silicon. Mater. Res. Bull. **1989**, *24*, 995–1004.
36. Croell, A.; Müller, W.; Nitsche, R. Floating zone growth of surface coated silicon under microgravity. J. Cryst. Growth. **1985**, *79*, 65–70.
37. Herrmann, F.M.; Mueller, G. Growth of GaAs crystals by the floating-zone technique under microgravity. J. Cryst. Growth. **1995**, *156*, 350–360.
38. Croell, A.; Kaiser, T.; Schweizer, M.; Danilewsky, A.N.; Lauer, S.; Tegetmeier, A.; Benz, K.W. Floating-zone and floating-solution-zone growth of GaSb under microgravity. J. Cryst. Growth. **1998**, *191*, 365–376.
39. Lal, R.B.; Aggarwal, M.D.; Kroes, R.L.; Wilcox, W.R. A new technique of solution crystal growth. Phys. Status Solidi (a) **1983**, *80*, 547.
40. Steiner, B.; Dobbyn, R.C.; Black, D.; Burdette, H.; Kuriyama, M.; Spal, R.; van den Berg, L.; Fripp, A.; Simcheck, R.; Lal, R.B.; Batra, A.K.; Matthiesen, D.; Ditchek, B. High resolution synchrotron X-radiation diffraction imaging of crystals grown in microgravity and closely related terrestrial crystals. J. Res. National Inst. Standards Technol. **1991**, *96*, 305.
41. Batra, A.K.; Lal, R.B.; Aggarwal, M.D. Electrical properties of TGS crystals grown by new technique. J. Mater. Sci. Lett. **1985**, *4*, 1415.
42. DeLucas, L.J.; Smith, C.D.; Smith, W.H.; Kumar, S.V.; Senadhi, S.E.; Ealick, S.E.; Carter, D.C.; Snyder, R.S.; Weber, P.C.; Salemme, R.; Ohlendorf, D.H.; Navia, M.A.; McKeefer, B.M.; Nagabhushan, T.L.; Nelson, G.; Bugg, C.E. Protein crystal growth in microgravity. Science **1989**, *246*, 651–654.
43. Chernov, A.A. Protein versus conventional crystals; creation of defects. J. Cryst. Growth **1997**, *174*, 354–361.
44. McPherson, A. Virus and protein crystal growth on earth and in microgravity. Appl. Phys. D **1993**, *26*, 104.
45. Amende, W.; Holl, S. Remelting and solidification of turbine blade samples within te scope of the D-2 mission. Adv. Space Res. **1994**, *16* (7), 191–194.

M

46. Busse, P.; Deuerler, F.; Poetschke, J. Stability of the ODS alloy CMSX6-Al_2O_3 during melting and solidification under low gravity. J. Cryst. Growth. **1998**, *193*, 413–425.
47. Voltz, M.P.; Schweizer, M.; Kaiser, N.; Cobb, S.; Vujisic, L.; Motakef, S.; Szofran, F.R. Bridgman growth of detached GeSi crystals. J. Cryst. Growth. **2002**, *237–239*, 1844–1848.
48. Lopez-Jaramillo, F.J.; Garcia-Ruiz, J.M.; Gavira, J.A.; Otalora, F. Crystallization and cryocrystallography inside X-ray capillaries. J. Appl. Cryst. **2001**, *34*, 365–370.
49. www.containerless.com/realglass.htm. See Containerless Research, Inc., Glass Products Division homepage See also NSF press release, NSF PR 03-119. October 7, 2003, ''New glass can replace expensive crystals in some lasers and bring high power to small packages.''
50. See Liquidmetal Technologies home page, www.liquidmetal.com/index/
51. Kelton, K.F.; Lee, G.W.; Gangopadhyay, A.K.; Hyers, R.W.; Rathz, T.J.; Rogers, J.R.; Robinson, M.B.; Robinson, D.S. First X-ray scattering studies on electrostatically levitated metallic liquids: demonstrated influence of local icosahedral order on the nucleation barrier. Phys. Rev. Lett. **2003**, *90* (19), 195504:1–4.

Microreactors and Microreaction Engineering

M

Richard I. Masel
Scott Gold
Zheng Ni
Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.

INTRODUCTION

Microreactors are chemical reactors with characteristic length scales ranging from the order of one to hundreds of micrometers. Such small length scales give microreactors many unique and potentially very useful properties. Fluid flow is nearly always laminar in microreactors, and timescales for heat and mass transfer become very short. Because smaller amounts of chemicals, typically on the order of microliters, are required for microreactors, hazardous and even explosive materials may be used at much less risk, opening up new reaction pathways. The small size of microreactors also makes them ideal for portable applications such as in micropower generation. The nature of microfabrication technology also is such that microreactors can be cheaply mass produced. Presently, microreactors are used mainly in research laboratories. However, projections suggest that the microreactors will soon be used in consumer products and in devices for medical laboratories.

WHAT ARE MICROREACTORS AND WHY ARE THEY USEFUL?

Microreactors are chemical reactors that are much smaller than typical industrial scale reactors, on the order of 1000 times smaller than the smallest reactors that were used for large-scale chemical production in the 1990s. Generally, they have a volume of $10\,cm^3$ or less and characteristic length scales on the order of 1 mm or less. Presently, microreactors are used mainly in research laboratories. However, projections suggest that the microreactors will soon be used in consumer products and in devices for medical laboratories.

Historically, the use of microreactors dates back to the 1940s when they were developed to measure kinetics of catalytic reactions.[1–5] One of the key early findings was Denbigh's 1965 observation that if a reactor were made small enough, temperature and concentration gradients with the reactor would be negligible, so that "differential" (i.e., gradientless) behavior would be observed.[6] This allowed much more accurate kinetic measurement to be made. Indeed, small differential reactors for research purposes are still sold today.

More recently, interest has been growing in the use of microreactors for consumer products, devices for medical laboratories, and for small-scale chemical production. For example, Fig. 1 shows a picture of Casio's proposed microreactor to power a laptop computer. This microreactor would produce hydrogen from methanol, then feed the hydrogen into a fuel cell to power the laptop.[7] The device is expected to have a lifetime 10 times that of lithium batteries. Another application for which microreactors have shown great promise is in amplification of DNA molecules by the polymerase chain reaction (PCR) method. The process requires three steps, each done at a different temperature, which are repeated to produce the desired amount of DNA. The speed of the process is generally limited by the speed of the thermal cycles. The rapid heat transfer that can be achieved in microreactors makes them ideal for this process, dramatically increasing the speed of the thermal cycles relative to conventional-scale devices. A schematic of a PCR device developed by Manz et al. is shown in Fig. 2.[8] These are just two of the many case studies demonstrating applications where microreactors, provide significant advantages of conventional systems.

To date, microreactors have not yet appeared in consumer products or for other large-scale uses. Thus, while microreactors have significant potential, it is unclear when, or if, that potential will be realized.

Generally, the main advantages of microreactors are

- Small size
- Rapid heat and mass transfer
- Ability to be mass produced

The size is critical to the use of microreactors in portable devices, such as many electronics goods. Microreactors could also be made small enough and be easy enough to mass produce so that they could be used in consumer products. Another advantage of the small size is that lesser amounts of chemicals must be used. This is highly desirable in dealing with chemicals that present significant environmental,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008058

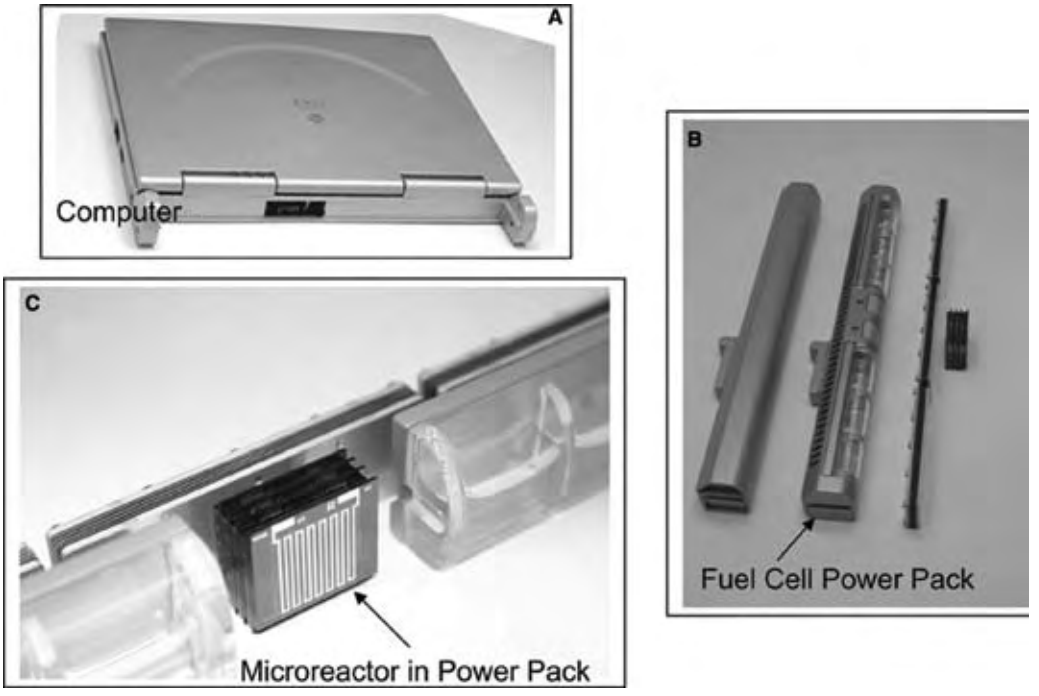


Fig. 1 A commercial version of the Casio methanol reformer and fuel cell for a laptop computer [above and a front and back view of the microreactor for methanol reforming (below)]. (From Ref.[7].) *(View this art in color at www.dekker.com.)*

health, and/or safety risks or with materials that have a high cost or are difficult to obtain, as is the case with many biological materials or some pharmaceuticals, for example.[74,75] One particular application calls for the production and use on the spot of metastable compounds. For example, phosgene is an intermediate in the production of isocyanides, but it is very toxic. Jensen et al. found that they could make the phosgene in a microreactor and use it immediately in a subsequent reaction.[9] This approach avoids the need for large-scale storage and/or transport of phosgene, thus substantially increasing the safety of the process.

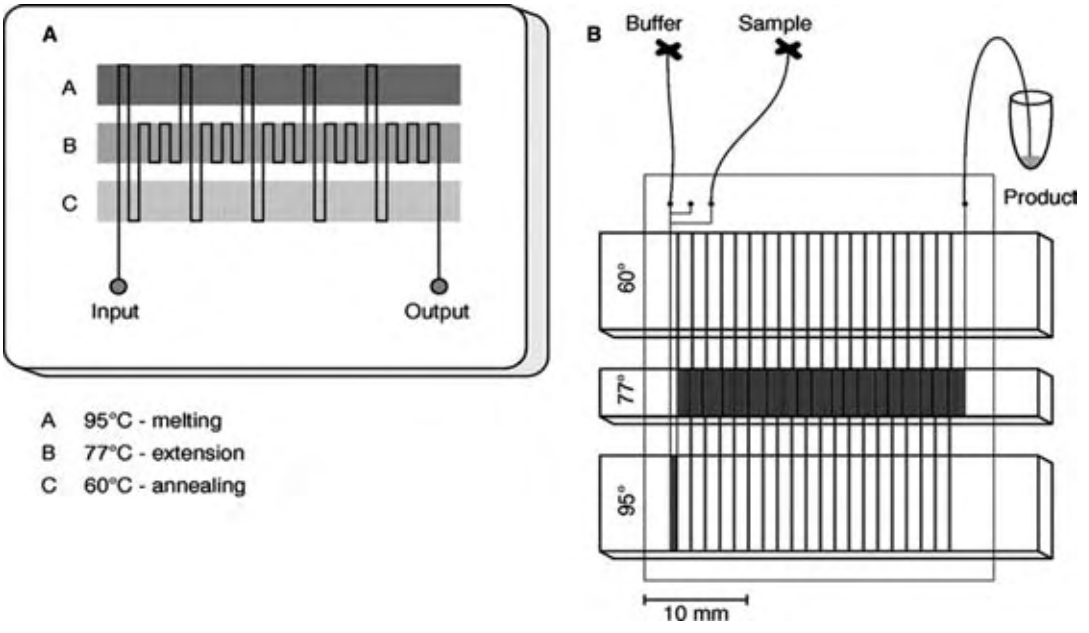


Fig. 2 Layout of a PCR chip reported by Manz et al. (A) Schematic of a chip for flow-through PCR. (B) Layout of the device used in the Manz study. (From Ref.[8].)

Another advantage of microreactors is that the heat and mass transfer rates are enhanced owing to short characteristic length scales. With regard to thermal management, this allows for more uniform heating throughout a reaction volume, more efficient heat removal, and better temperature control than is typically possible in macroscale systems. Heat transfer coefficients up to 25,000 $W/m^2/K$ have been reported in microreactors as compared to 2500 $W/m^2/K$ in conventional scale reactors. Extremely endothermic, exothermic, or even potentially explosive reactions can thus be safely and efficiently controlled in microreactors, opening up opportunities for new chemistries not feasible in conventional reactors. Short diffusion times allow for highly efficient mixing even under the laminar flow conditions that are characteristic of most microreactors.

High surface area to volume ratios are another distinctive feature of microscale reaction systems. Interfacial areas per unit volume in falling film microreactors have been reported to be as high as 25000 m^2/m^3. By comparison, interfacial areas in bubble columns are typically of the order of 1–200 m^2/m^3. As a consequence, surface effects that are often neglected in the macroscale become dominant. This makes a tremendous difference in gas–liquid reactions, where mass transfer from the gas to the liquid often limits the rate.

Microfabrication technology used to manufacture microreactors also introduces many advantages, most notably the ability to rapidly and cheaply mass-produce devices. The low cost of microfabricated devices makes it possible for these devices to be disposable, a characteristic desirable for many medical applications. Rapid scale-up of production by operating many microreactors in parallel can also be accomplished. Microfabrication also presents the opportunity for complete systems in a single monolithic device or systems on a chip as microreactors are incorporated with chemical sensors and analysis devices, microseparation systems, microfluidic components, and/or microelectronics.

CHARACTERISTICS OF MICROREACTORS

Scale

In discussing microreaction systems, it is helpful to first distinguish the characteristic size of a microreactor in comparison with conventional scale reactors. Two different definitions of the term microreactor are commonly used in the literature. The first defines any reactor that is an order of magnitude or more smaller than its conventional scale counterpart as a microreactor. For example, an industrial reforming reactor might be 10,000–100,000 L in volume. Using this first definition, a 1 L reforming reactor could be considered a microreactor. This definition is not very useful as it includes reactors of virtually any scale. A second, more precise and useful definition is that a microreactor is a reaction system with microscale characteristic length scales, on the order of 1–100 μm, or at most on the order of millimeters. This length scale typically refers not to the system as a whole, but rather to the length scale of the reaction vessel or fluid flow channels, or to the volume scale of materials handled by the microreaction system.[10] The volumes of reactants and products dealt with are typically on the order of microliters or less. We will adopt the latter definition here and call a microreactor, any reactor where the critical dimensions for fluid flow are about a millimeter or less. This definition is more useful than the first as transport processes are quite different at these critical device dimensions. Generally, the flow is laminar, heat and mass transfer is rapid, and wall interactions become more important when device sizes are millimeters or less. Thus, millimeter and microscale devices are fundamentally different from devices with a larger scale.

Fluid Mechanics in Microreactors

The small characteristic length scales found in microreactors significantly affect flow behavior in these systems and, consequently, their design and performance. In the section that follows, we will give an overview of some of the important characteristics of fluid mechanics at the microscale. The reader is referred to one of the many texts or review articles on microfluidics for a more thorough discussion of the topic.[10–14,73]

Flow regime

As a general rule, continuum mechanical descriptions of fluid flow still apply in devices with microscale critical dimensions.[10] Also, owing to the small characteristic length scales involved, the Reynolds number ($Re = uL/v$, where u is the characteristic flow velocity, L is the characteristic length scale, and v is the kinematic viscosity of the fluid) is very small, in the order of 10 or less, resulting in laminar flow.[10] Exceptions to both of these general rules, though not commonly encountered, can and do occur. For example, continuum mechanics do not apply to very diffuse gas flows where the mean free path becomes large when compared to the characteristic length scale for the flow ($Kn > 10^{-1}$, where Kn is the Knudsen number) or for very high shear rates in liquid flows. For length scales at the higher end of the microrange (on the order of hundreds of micrometers to millimeters), especially for very high flow rates, turbulent flow can be achieved.[10]

Fig. 3 illustrates the flow of two streams of dye inserted into a microfluidic channel.[15] Note that the

Fig. 3 Illustration of laminar flow in a microfluidic channel. (From Ref.[15].) *(View this art in color at www.dekker.com.)*

two dyes stay on their own side of the channel as mixing occurs only due to diffusion. Such a flow pattern is characteristic of laminar flow, and significantly impacts the design of microreactors when mixing of two or more materials is required, as discussed in the section on ''Mixing.''

Gas flows

Gas flows in microreactors are well described using the kinetic theory of gases. Two dimensionless groups of particular interest for gas flows are the Mach number and the Knudsen number. The Mach number, $Ma = u/c_s$, where u is the characteristic velocity of the flow and c_s is the speed of sound, gives a measure of the compressibility of the flow. If $Ma < 0.3$, which is common in microscale systems, the flow is considered to be incompressible. Exceptions to this condition occur when the walls of the system are heated leading to large changes in the density of the gas or when self-heating of the gas occurs due to viscous dissipation as might happen in very long, thin channels.[10] The Knudsen number provides a measure of how rarified the flow is. As previously mentioned, at high Kn ($Kn > 0.1$), continuum mechanics break down as the flow transitions into the free molecular regime. For smaller Knudsen numbers, as typically observed in microscale systems, the magnitude of the Kn determines the boundary conditions applied to the Navier–Stokes equations. For $Kn < 10^{-3}$, a no-slip boundary condition applies, while for $10^{-3} < Kn < 10^{-1}$, a slip boundary condition applies.[10]

Liquid flows

Liquid flow behavior is far more complex and not nearly as well understood as gas flow, especially in microscale. Generally, when the dimensions of a flow channel are 10 μm or more, the fluid behavior is simple and laminar. However, as the channel dimensions shrink below 1 μm new processes start to occur.[10] The presence of the wall can affect the phase behavior of the fluid causing colloidal behavior. Numerous and thus far unexplained deviations from classical laminar flow behavior have been observed. At very high strain rates, deviations from Newtonian behavior are observed in fluids that are Newtonian in macroscale systems.[16] Liquid viscosities have been observed to increase, decrease, and remain constant in microfluidic devices as compared to viscosities in larger systems.[10] Deviations from the no-slip boundary condition have been observed to occur at high shear rates.[17] One important deviation from no-slip conditions occurs at moving contact lines, such as when capillary forces pull a liquid into a hydrophilic channel. The point at which the gas, liquid, and solid phases move along the channel wall is in violation of the no-slip boundary condition. Ho and Tai review discrepancies between classical Stokes flow theory and observations of flow in microchannels.[11] No adequate theory is yet available to explain these deviations from classical behavior.[18]

Two-phase flow

Two-phase flow can also be utilized in microreactors. Fig. 4 shows a flow map for two-phase flow on the microscale.[19] Note that two phase flow is much simpler in a microreactor than is typical in a macroscale reactor. Only two-flow regimes are important: slug flow and annular-dry flow; complex turbulent conditions have not been observed. At low gas and liquid flow rates, the flow regime is slug–annular, in which small slugs of gas are present in the centers of the channels. With increasing gas flow, the flow regime changes to annular, in which the gas pockets become continuous streams in the centers of the channels, while the liquid streams are confined to the edges. At high liquid flow rates, dispersed flow is attained.

Mass Transport

Reduced length scales also have a major impact on mass transfer in microreactors. The timescale for diffusion, τ, scales as the square of the length scale and can be expressed as $\tau = L^2/D$ where L is the characteristic length and D the diffusion coefficient. Typical diffusion coefficients are 10^0–10^{-2} cm^2/sec in gases and 10^{-4} to 10^{-6} cm^2/sec in liquids. As length scales decrease in microreaction systems, very short timescales for diffusion can be achieved leading to substantially enhanced mass transfer rates. This has particular importance with regard to mixing in microreactors. Mixing can be crucial to reactor performance, and is

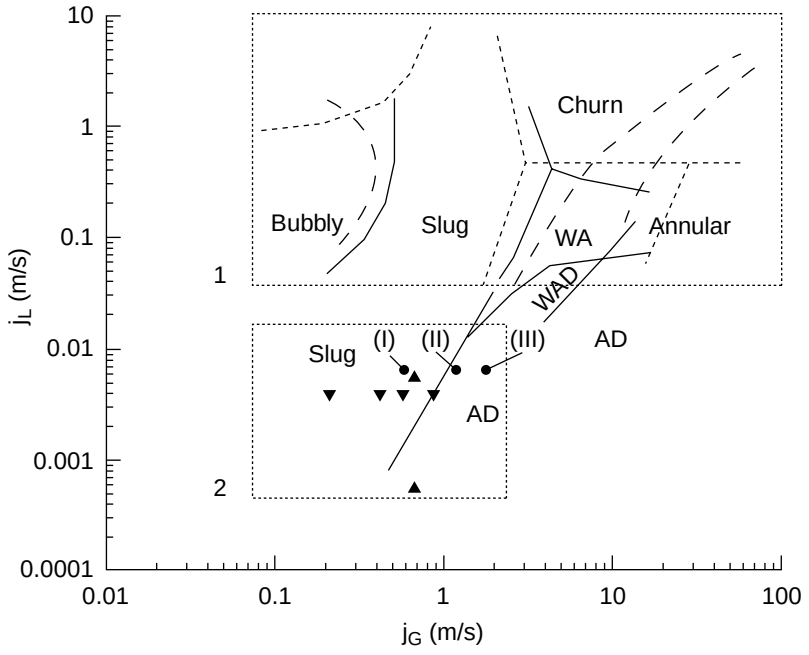


Fig. 4 Gas–liquid flow regime map in microchannels. Flow regimes are annular, churn, wavy annular (WA), wavy annular-dry (WAD), slug, bubbly, and annular-dry (AD) flows. (From Ref.[19].)

often the limiting factor in determining conversion. In conventional macroscale systems, mixing is usually achieved by the introduction of turbulence to the system. Because flow in microchemical systems is nearly always laminar, microscale mixing technologies must rely on the diffusion to achieve mixing. The typical approach taken in mixing in microreactors and other microfluidic systems is to reduce the diffusion length scale. Various micromixer designs will be discussed in the section on ''Mixing.''

Significant mass transfer advantages can be achieved in two-phase flows in microreactors.[20] Slug–annular flows provide contact areas of $\sim 1500\,m^2/m^3$ while dispersed flows provide contact areas of $\sim 16{,}000\,m^2/m^3$ in microreactors.[20] By way of comparison, bubble tower contact areas are ~ 1–$200\,m^2/m^3$. Thus, there is much more area for mass transfer in a two-phase microreactor than in a bubble column, or other two-phase mixer in addition to having a shorter characteristic length for diffusion. This effect is particularly important in gas–liquid reactions where the rate of mass transfer from the gas to the liquid limits the reaction rate. For the hydrogenation of cyclohexene to cyclohexane in this type of reactor, the mass transfer rate constant, $K_L a$, was found to be in the range 3–7 sec^{-1}, which is two orders of magnitude higher than that of conventional reactors.[20]

Despite the small dimensions, conventional models for dispersion have been applied to microreactors. Since flow is laminar, a Taylor–Aris or shear-induced dispersion model generally describes dispersion.[21–25] Beard has applied the Taylor–Aris dispersion model to correlate diffusion rates in a T-sensor with rectangular microchannels.[23–25]

Beard's derivation was for a rectangular channel of height W (y dimension) much less than the width S (x dimension), flow in the z direction, and a nonreactive solute of concentration c. The z component of laminar fluid velocity is approximated to be parabolic and is given by

$$\nu_z(y) = \frac{6\overline{V}}{W^2}\left(y^2 - Wy\right)$$

where $\overline{V}$ is the mean velocity. Owing to the high aspect ratio of the channel, variations of velocity in the x direction were neglected. The two dimensional advection–diffusion equation derived by Beard is given by

$$\frac{\partial c}{\partial t} = D_0\frac{\partial^2 c}{\partial x^2} + \left(D_0 + \frac{33}{560}\frac{\overline{V}^2 W^2}{D_0}\right)\frac{\partial^2 c}{\partial z^2} - \overline{V}\frac{\partial c}{\partial z}$$

where D_0 is the molecular diffusion coefficient, and the effective dispersion coefficient is given by

$$\widetilde{D} = \frac{33}{560}\frac{\overline{V}^2 W^2}{D_0}$$

Ghosal has also examined electrophoretic flow in microchannels using a Taylor–Aris dispersion model.[26] In both of the above cases Taylor–Aris models were reported to work well. On the other hand,[67] Ni, Seebauer and Masel found that at high flow rates over

porous reactor surface the Taylor–Aris model did not work as well, presumably because the wall roughness induced secondary flows.

Heat Transfer

Analogous to the enhancement of diffusion in microreactors is the enhancement of heat transfer at small length scales and high relative surface areas. Heat transfer coefficients as high as $25{,}000\,W/m^2/K$ have been reported in micro heat exchangers. This compares to $2500\,W/m^2/K$ in a conventional scale heat exchanger. Surface areas per unit volume are typically on the order of $10{,}000\,m^2/m^3$ in microheat exchangers, compared to $1–500\,m^2/m^3$ in conventional scale reactors. The net effect is that heat transfer is much more rapid on the microscale than on the conventional scale. Heat generated in a microreactor can be rapidly dispersed. The small scale also makes it much easier to control reaction temperatures uniformly throughout the reaction volume using external heating or cooling systems. As a result, microreactors are very attractive for both highly exothermic and highly endothermic reactions, where hot or cold spots can create problems, such as thermal quenching, thermal activation of undesired side reactions, or explosions.

Convective heat loss to the environment can be much more significant in microreactors than in conventional systems because of their high surface to volume ratios. A consequence of this is that insulation generally enhances heat loss rather than mitigating it. The critical radius of a tube below which the use of an outer insulation layer will increase heat losses is given by the ratio of the thermal conductivity of the insulation layer (k) and the heat transfer coefficient at the interface between the insulation and the ambient (h_0), $r_{critical} = k/h_0$. For most commonly used materials and ambient air, this critical radius is on the order of 1 mm.[27]

There are several examples where rapid heat transfer in microreactors has been utilized to perform highly exothermic reactions that would not be viable in conventional reactors. Localized hot spots occurring in conventional reactors to directly fluorinate aromatics can lead to undesired side reactions, including the formation of fluorine radicals and nonselective free-radical side reactions. Jenson et al. have run direct fluorination reactions of aromatics in microreactors, taking advantage of the rapid heat transport in the microreaction system to operate under very aggressive reaction conditions (up to 10 vol.% fluorine).[19] Highly exothermic nitration reactions of aromatics have also been widely studied in microreactors.[28,29] At high temperatures, hydrogen cyanide can hydrolyze to ammonia. Hessel et al. were able to produce hydrogen cyanide isothermally in a microreactor via the Andrussow route and avoid this undesired side reaction.[30] Another example is the production of the antibiotic ciprofloxacin which requires a complex 20–30-step synthesis to avoid explosions in conventional reactors. CPC Systems have recently patented a synthesis process utilizing a microreactor to produce ciprofloxacin via a five-step synthesis, where the first step is a highly exothermic and potentially explosive cycloaddition.[31]

Surface Reactions

As length scales are reduced, the surface area to volume ratio increases dramatically. For example, consider a cylinder of 1 cm in length. For a radius of 1 cm, the surface area to volume ratio is $2\,cm^{-1}$. For a radius of 100 μm, the surface area to volume ratio is $200\,cm^{-1}$. This high surface area to volume ratio inherent in microscale systems has many consequences. Among these are the increased importance of surface reactions and surface tension.

The relatively high surface of microreactors can be taken advantage of to enhance surface-catalyzed reactions. For example, Langer et al. have utilized reactive polymer coatings in microreactors to produce various biomolecules and to perform bioassays, as illustrated in Fig. 5.[32,33] The addition of monoliths or posts to the microreactor design can be utilized to further increase surface area for catalytic reactions. Sadykov et al. report the use of Pt-based catalyst wash coated onto monolithic supports in a microreactor for the selective production of propylene by propane oxidative dehydrogenation.[34] As shown in Fig. 6, Masel et al. utilize alumina surfaces that have been anodized to further enhance surface areas of catalytic supports in microreactors allowing very high conversions and production rates to be achieved in the reforming of ammonia to hydrogen.[35,36]

Surface Tension

Surface tension forces are proportional to the contact area between two immiscible phases and decrease linearly with the length scale. Other forces, however, tend to decrease with higher powers of the length scale. For example, inertial forces decrease proportionally to the cube of the length scale. Consequently, the relative importance of surface tension is magnified at small length scales. One result of surface tension is that it can be extremely difficult to force liquids into channels made of materials not wetted by that liquid. Similarly, it can be a challenge to get liquids out of small channels that are wetted by the liquid. While surface tension can create some engineering difficulties in microreactors, it can also be harnessed. In continuous-flow systems, the use of capillaries is an effective way

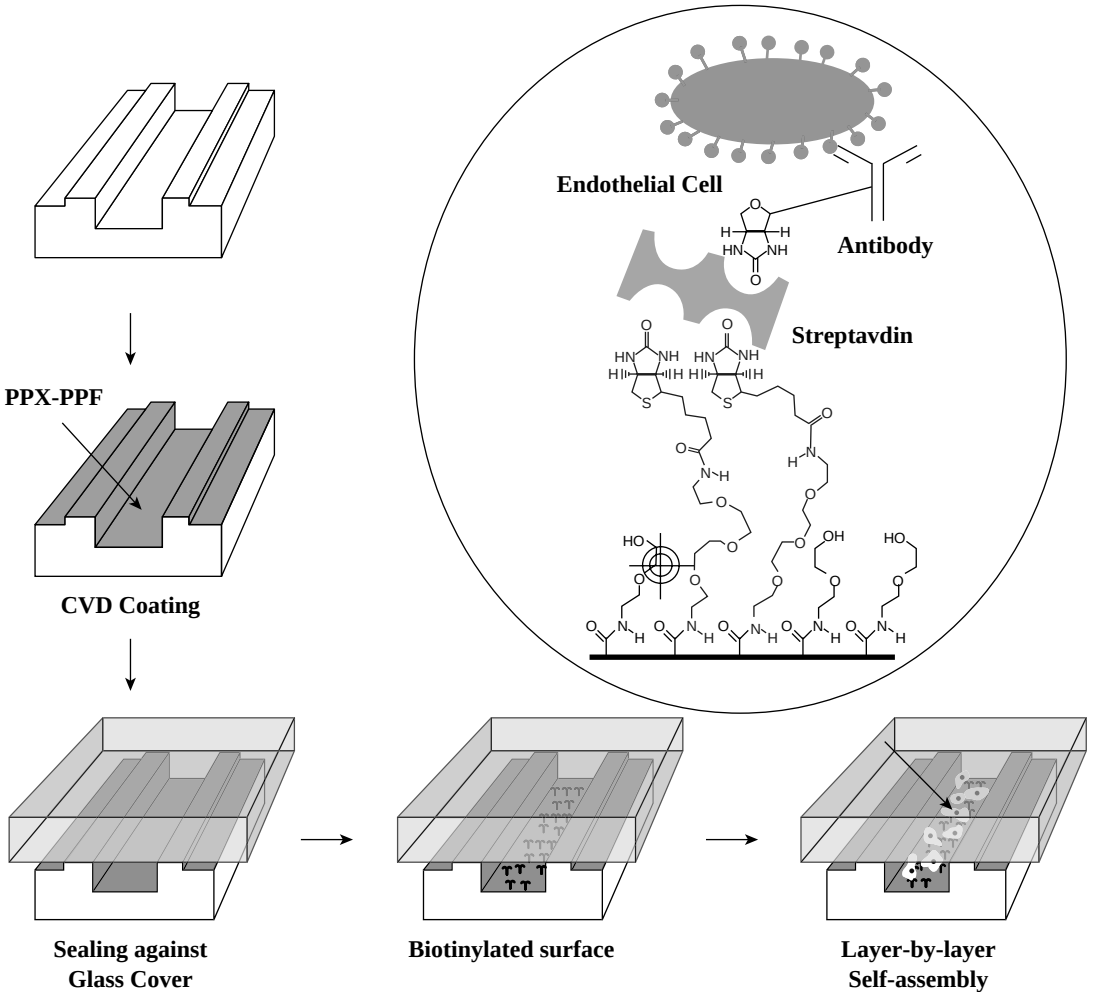


Fig. 5 Schematic representation of the surface modification steps used by Langer et al. to produce a biologically active surface in a microreactor. Biotin-labeled human anti-α_5-integrin (HAI) was bound to modified PDMS and used to study cell surface receptor activity. (From Ref.[33].) *(View this art in color at www.dekker.com.)*

of separating liquids from gases.[37] By controlling surface tension gradients, movement and mixing of liquid droplets can be achieved without the use of pumps and valves. Several approaches have been used to control surface tension gradients, including electrowetting or electrocapillarity, thermocapillarity, electrochemical processes, and light.[38–41]

Short Residence Time

Another consequence of the small length scales in microreactors is that very short residence times are commonly achieved creating engineering challenges to be overcome or major advantages depending on the application. If the residence time approaches the kinetic reaction timescale, conversions can become unacceptably low necessitating device designs that maximize residence times or enhance reaction rates. For example, in microcombustion systems, the kinetic reaction timescale is of the order of 0.01–0.1 msec for hydrocarbon–air reactions. In this case, the short residence time can lead to unacceptably low conversions. One obvious approach to overcoming this problem is to increase the reactor size and, consequently, the residence time. Other approaches to enhancing conversion such as premixing of the reactants, incorporation of catalytic materials into the reactor, or external heating of the reactor can be used to mitigate the problems posed by the short residence times and allow device size to remain small.[42,43] Short residence times achievable in microreactors can in some case be an advantage rather than a problem. For example, unstable or metastable intermediates can be produced and transferred quickly to a subsequent process.[10] Similarly, hazardous materials can be produced at the point of use.

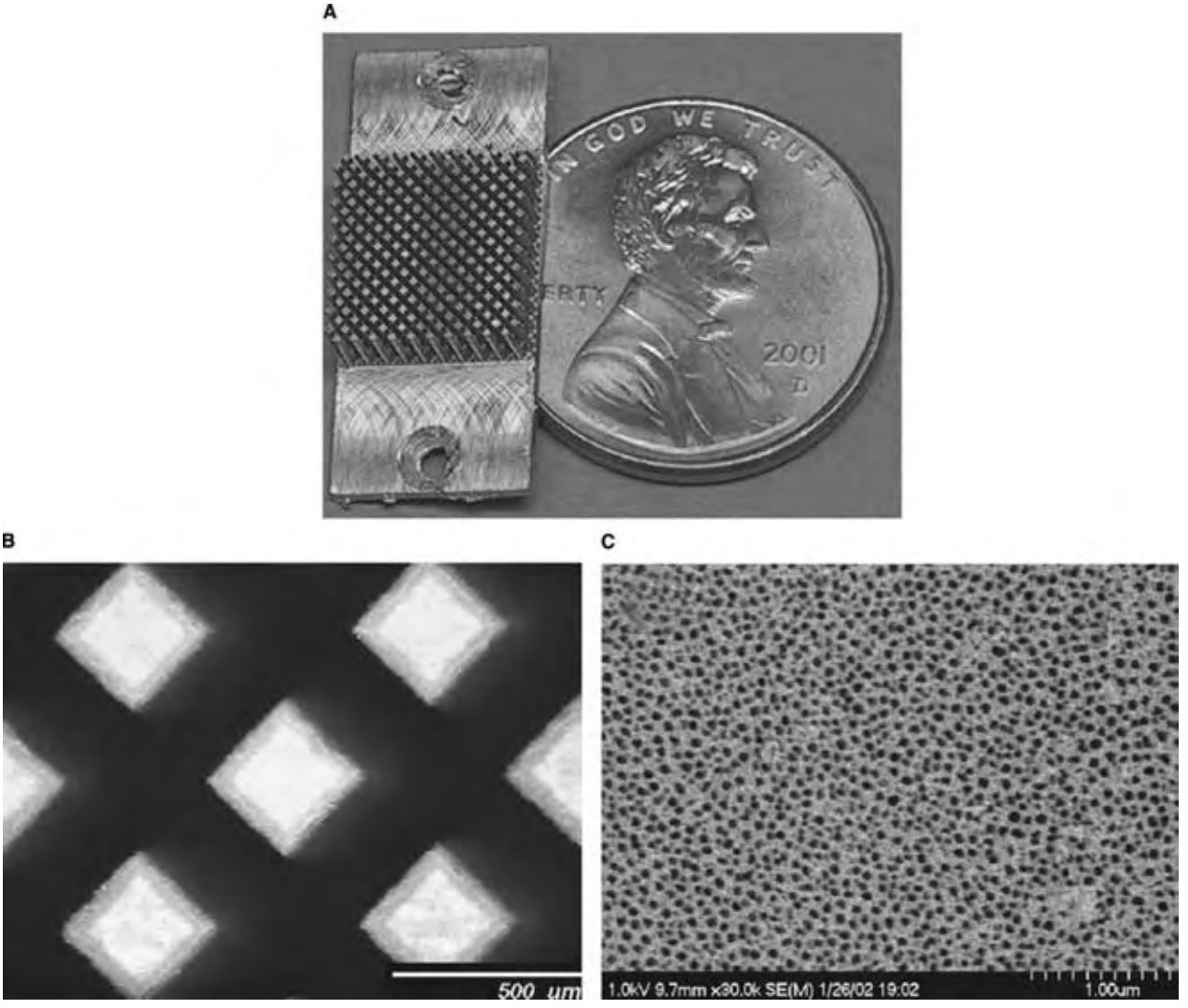


Fig. 6 (A) Anodized alumina, posted microreactor used for reforming of ammonia to hydrogen. (B) Optical microscope image of the microreactor posts. (C) Scanning electron microscopy image of a microreactor postsurface illustrating the porosity of the anodized surface. *(View this art in color at www.dekker.com.)*

This generates opportunities to use new reaction pathways not feasible in conventional macroscale reactors. For example, Sadykov et al.[34] utilize rapid thermal quenching in conjunction with short residence times in a microreactor to suppress undesired side reactions in propylene production by the oxidative dehydrogenation of propane.

PRINCIPLES OF MICROREACTION ENGINEERING

Design of Microreactors

Fluid movement

For fluid movement to occur, there must be a driving force for that movement. In many cases, flow in the microreactor system is driven by external, macroscale pumps, flow controllers, etc. Often, however, it is desirable to control fluid flow within the microscale system. A wide array of designs for micromechanical pumps and valves have been described in the microfluidics literature for just that purpose.[10]

In addition to conventional pressure driven flow, electrokinetic flow is also a commonly used means of transporting liquids in microfluidic devices. One type of electrokinetic flow, electroosmotic flow, relies on the presence of an electrical double layer at the solid–liquid interface. A negatively charged surface in a flow channel will attract cationic species from the fluid to form an electrical double layer at the surface. Application of an external voltage can pull those cationic species through the flow channel inducing bulk flow. The electroosmotic flow velocity can be described

by the Helmholtz–Smoluchowski equation:

$$u_{EOF} = \mu_{EOF}E = \frac{\varepsilon\zeta E}{\eta}$$

where u_{EOF} is the electroosmotic flow velocity, μ_{EOF} is the electrophoretic mobility, E is the electric field strength, ε is the dielectric constant of the fluid, and ζ is the zeta potential of the flow channel wall.[10] Velocity profiles in electroosmotic flow tend to be very flat. Significant flow velocities, as high as 5 cm/min, can be achieved with electroosmotic flow.[10]

Mixing

In conventional reactors and fluid systems, mixing is typically achieved by introducing eddies to the system. Eddies can be created in the wake of a ball moving through the fluid, by introducing fluid to the reactor through a jet, with a stirrer, or with any of a number of other approaches. These approaches are not viable in the microscale because of the large fluid volumes required and the fact that Reynolds numbers in the microscale are so low. Micromixer designs seek to enhance diffusional mixing by decreasing the characteristic diffusion length. In addition to achieving fast mixing times, it is also desirable to minimize the device area of the micromixer and to design it such that it can be integrated easily into the larger microchemical system. Several designs for microscale mixers have been developed, which can be divided into two classes, passive and active. They will be briefly described in the next two sections. More extensive discussion of micromixer design are given in the Refs.[10,44,68–72].

Passive Mixing. Passive or static mixing technologies are typically simple in that they involve no moving parts. Probably, the most common types of micromixing devices are lamination micromixers, of which there are two main types, parallel and sequential. In parallel micromixers, inlet streams are split into substreams, which are then joined into a mixed stream as laminae, in the simplest case, two fluids to be mixed are joined in a ''T'' or ''Y'' pattern. An example of a ''T'' micromixer is shown in Fig. 7. Splitting into multiple fluid streams can further enhance mixing. Another approach to enhanced mixing is to introduce a throttle design by reducing the size of the inlet, illustrated in a ''Y'' mixer in Fig. 8.[10,45] This reduces the timescale for diffusion at the throttled inlet. Johnson, Ross, and Locascio reported enhanced mixing when small wells were etched into the combined flow channel just past the confluence of two streams in a T mixer.[46] Another strategy used to enhance mixing in lamination micromixers is sequential switching. The flows of the fluids are alternately turned on and off. Sequential lamination micromixers, illustrated in Fig. 9, provide more rapid mixing by resplitting and then recombining the mixed fluid streams several times in sequence. One downside to these micromixers is that a more complicated structure is required. Injection or microplume mixers operate in a similar manner. In these micromixers, one fluid stream is split into many, which are then injected into a second fluid stream. By increasing the contact surface between the two fluids, the mixing path and therefore the mixing time is decreased. Application of these mixing technologies can be used for single-phase or multiphase mixing.

Another mixing strategy involves utilizing posts within a reaction chamber in a checkerboard pattern, as shown in Fig. 6.[47] In such an arrangement, characteristic diffusion lengths are decreased in the areas between the posts and between the post and the wall.[47] When this strategy is employed, the posts are commonly used as catalyst supports. The posts mix the streams further and provide a high surface area for gas–liquid contacting, while also having a lower pressure drop than a packed-bed reactor with a

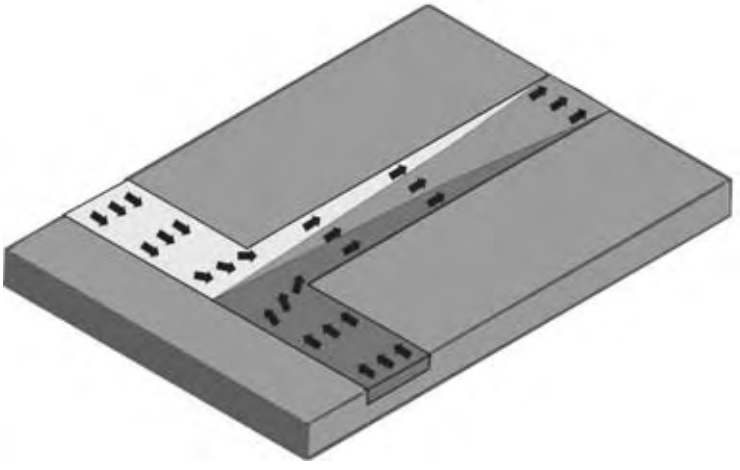


Fig. 7 A ''T'' micromixer. *(View this art in color at www.dekker.com.)*

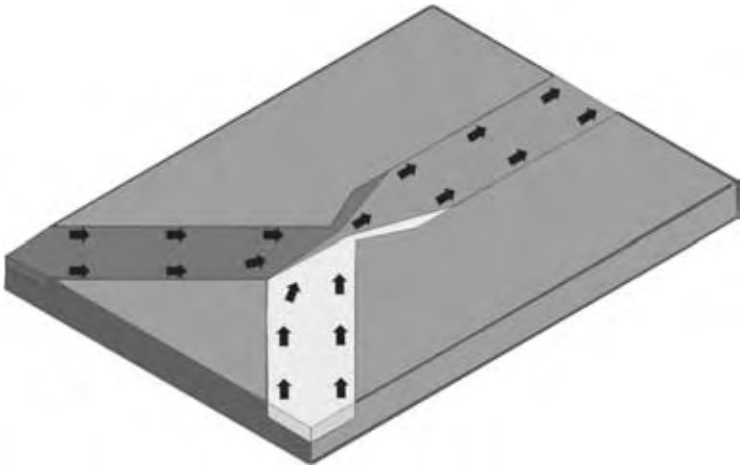


Fig. 8 A "Y" micromixer with throttling at the confluence of the impinging streams to enhance mixing. *(View this art in color at www.dekker.com.)*

comparable catalyst surface area. The uniformity of arrangement also allows for precise control of flow characteristics and minimization of channeling.

Active Mixing. Because they employ moving parts, active micromixers are inherently more complicated than their passive counterparts. Active or dynamic mixing technologies utilize actuators to agitate a fluid and thus cause mixing. A variety of different types of actuators have been employed to induce fluid mixing. Ultrasonic mixers, for example, utilize acoustic actuators.[48,49] Micromechanical pumps have also been utilized to drive fluid into a mixing chamber from the opposite direction in chaotic advection mixers.[50,51] Other active mixing technologies include magneto hydrodynamic and electrokinetic micromixers.[52,53]

Thermal management

One of the chief advantages of microreactors is the rapid heat transfer achievable at such small length scales. Consequently, microreactors are used in many applications where thermal management is critical. Many approaches for thermal management have been used in conjunction with microreactors. One widely examined method is the use of microheat exchangers. In design, microheat exchangers resemble their conventional counterparts. As in conventional heat exchangers, the thermal conductivity of the heat transfer fluid and the heat exchanger material as well as the flow configuration (cross-flow, countercurrent flow, etc.) are important design parameters.[54] Decreased length scales magnify the importance of material selection in microheat exchangers, as axial heat conduction in the walls of the heat exchanger is enhanced.[54–56] Contrary to intuition, materials with high heat conductivities reduce heat transfer efficiency when axial heat conduction becomes important. As thermal conductivities approach zero, the insulating effect of the material reduces heat transfer efficiency. Optimal heat conductivities are of the same order of magnitude as glass and other ceramics.[55]

Another approach to thermal management is ohmic heating.[57] This can be done by incorporating

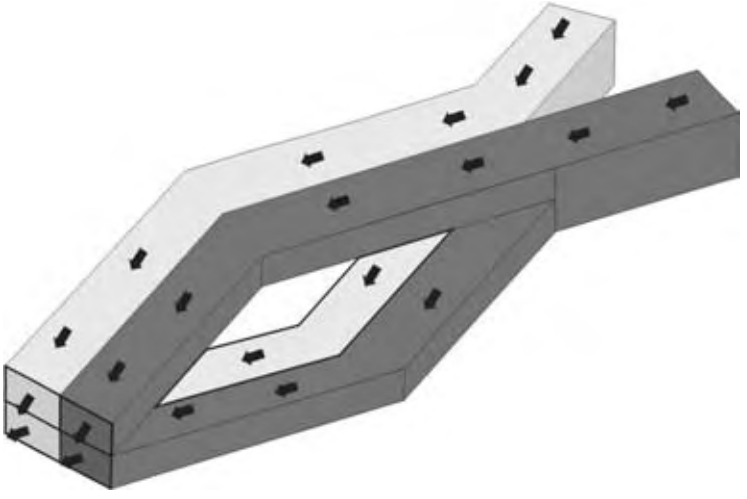


Fig. 9 A sequential lamination micromixer. *(View this art in color at www.dekker.com.)*

opposing electrodes to resistively dissipate electrical power as heat through a conductive medium.[57] Incorporation of a simple resistor into the reactor design can also be used.[27]

For high-temperature systems, radiative heat loss can be very significant. In conventional systems, reflective shields or radiation mirrors and low-emissivity material are commonly used to mitigate radiative heat loss. It is expected that similar approaches would be effective in microreactors as well.[27]

Fired heaters are often used to generate high temperatures in large-scale reactors. However, combustion is much more difficult on the microscale than on the conventional scale. Under normal conditions, flames will not propagate in small, confined spaces. The two dominant mechanisms of flame quenching, thermal and radical quenching, are surface phenomena, which are magnified in small devices. Flame arrestors employ a wire mesh to capture free radicals to terminate combustion reactions and quench flames, taking advantage of the high surface area and small gaps in the mesh. Masel et al. have demonstrated microburners with stable flames in combustion channels as small as 100 μm as illustrated in Fig. 10.

There are two primary challenges in fabricating microburners, surface free-radical quenching of the combustion reaction and thermal quenching due to the rapid heat transfer that occurs in small devices. Radical quenching is a heterogenous kinetic process whereby highly reactive radical intermediates undergo termination reactions at the wall. The removal of these radicals, crucial to the propagation of homogenous catalytic cycles in the combustion mechanism, quenches the flame. In microburners, thermal quenching of confined flames occurs when heat generated by the combustion process fails to keep pace with the heat loss to the walls. The walls effectively act as an enthalpy sink for the homogenous-combustion zone, lowering the near-wall temperature and retarding the kinetic mechanisms, further reducing heat generation and leading to flame extinction. Thermal quenching is less dominant when the surfaces are at higher temperatures but does act to reduce the reaction rates of the homogenous reactions. Both quenching mechanisms act to reduce the reaction rates of the homogenous reactions. To overcome both thermal and radical quenching, materials of construction with low thermal conductivities and relatively inert surfaces are generally chosen. This usually restricts material choices to insulative oxides, such as silicon dioxide and alumina. Thermal management schemes, such as the use of excess enthalpy (via a Swiss roll burner design) and cutting edge aerogel insulation, are also employed to maintain high device temperatures. While the ionic surface electronic structure of ideal insulating oxides

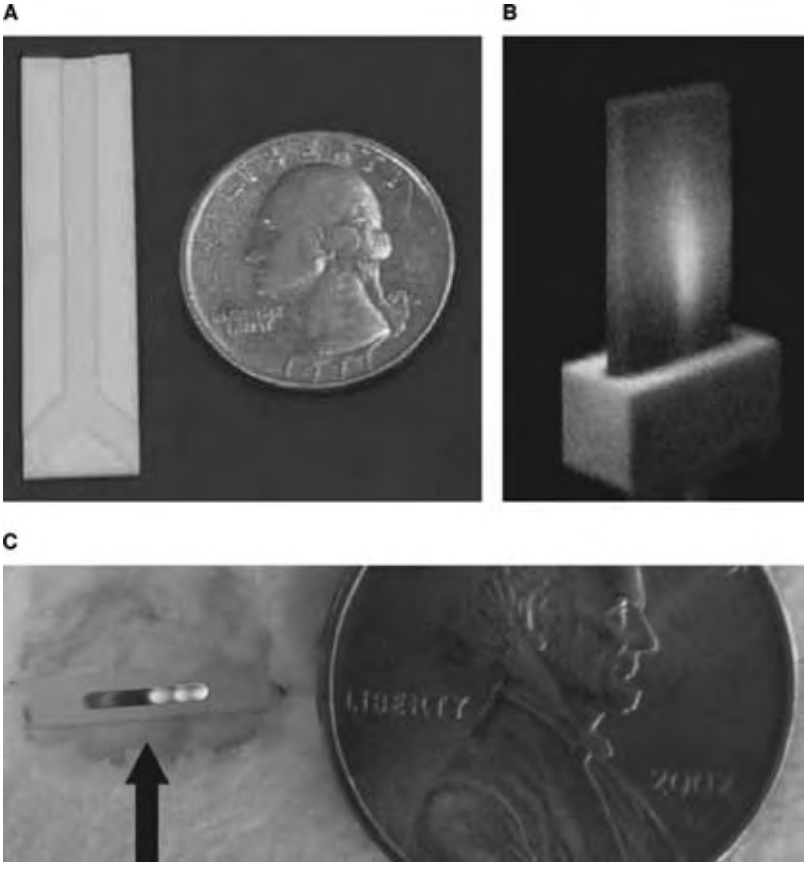


Fig. 10 Illustration of a microburner developed by Masel et al.: (A) cross section of alumina burner; (B) microburner operation; and (C) flame cells observed from top of microreactor. *(View this art in color at www.dekker.com.)*

generally renders them inert to heterogenous reactions, defects, impurities, and grain boundaries are present in reality. These surface features are active sites for radical quenching via radical recombination. Surface preparation to reduce the reactivity of the oxide surfaces, such as chemical cleaning and etching and thermal annealing, can be performed to reduce the occurrences and effects of these surface features. Another technique used to maintain high-temperature walls and to prevent radical quenching is using catalytic materials of construction, such as platinum. The heterogenous combustion, although not as reactive and exothermic as homogenous combustion can stabilize small gap combustion processes.

CASE STUDIES

Microreactors have been applied to many diverse applications and will doubtless be applied to many new reaction engineering problems in the future. Below are some specific examples of the uses of microreactors that illustrate some of the advantages of these systems. The applications described below clearly indicate the broad array of reaction engineering problems to which microreactors can be applied.

Hydrogen Production for Portable Electronics

One application of microreactors that is likely to appear in commercial devices soon is in the generation of hydrogen for portable-power generation. The power needs of portable electronic devices increase every year. Advances in battery technology have not kept pace with the growing needs. Thus, a power gap is developing, where portable devices can be built, but are not powered. One alternative is to use a fuel cell to generate the power. There are several designs including a direct fuel cell, and indirect fuel cell systems, where a hydrocarbon fuel is first reformed to form hydrogen, then the hydrogen is oxidized in a fuel cell to produce electricity. Such a design has been proposed by Casio to power laptop computers, as shown in Fig. 1. In response to this, many research efforts have focused on the use of microreactors to reform other fuels into hydrogen for use in microfuel cells. The small size and weight of microreactors makes them ideal for this application where portability is critical.

Masel et al. describe the synthesis and properties of a ruthenium-impregnated anodic aluminum catalyst in microreactors for the production of hydrogen from an ammonia feed, as illustrated in Fig. 6.[35,36] The catalyst structure was synthesized using microelectric discharge machining to create a series of posts on an aluminum substrate.[35,36]

The posts were anodized to yield a 60 μm covering of anodic alumina with an average surface area of $16\,m^2/g$ and an average pore size of 50 nm.[35,36] Ruthenium metal was dispersed on the alumina using conventional wet impregnation.[35,36] The enhanced surface area for the microreactor allows very high conversions and hydrogen production rates to be achieved relative to the reactor size.[35,36] A $0.9\,cm \times 0.9\,cm$ reactor containing 250 posts decomposed anhydrous ammonia at 650°C to yield 10 sccm of hydrogen at 95% conversion, while a similar channel design produced 200 sccm of hydrogen in one volume less than $1\,cm^3$.[35,36] A somewhat different approach is reported by Jensen et al. utilizing two suspended, thin-walled

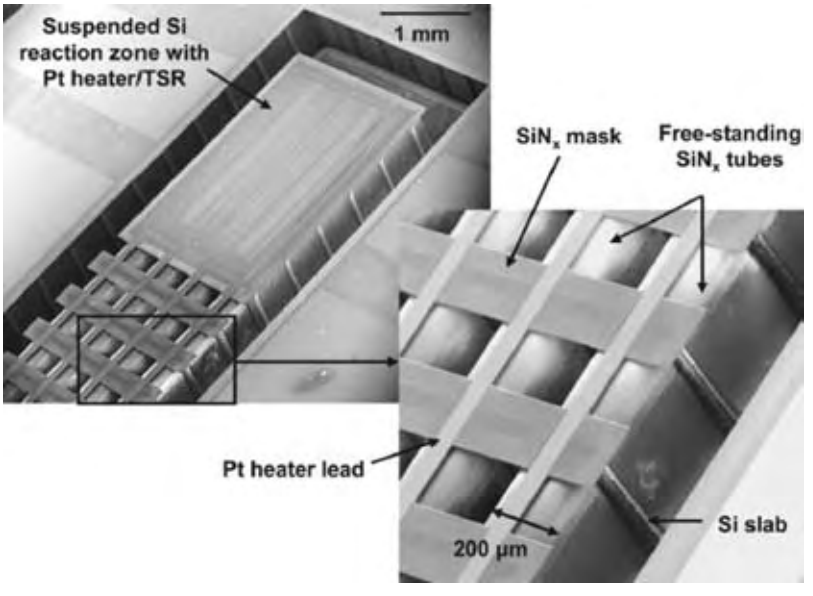


Fig. 11 Scanning electron microscopy photos of microreactor produced by Jensen et al. for reforming of ammonia showing four free-standing SiN_x tubes, a suspended Si reaction zone with integrated thin-film platinum heater and temperature sensing resistor (TSR), and Si slabs wrapped around the four tubes. (From Ref.[27].)

(2 μm) silicon nitride tubes in a ''U'' shape, as illustrated in Fig. 11.[27] The inlet and outlet ends of the U-shaped tubes are-fixed in a silicon substrate, while the remainder of the tube is kept floating to help thermally isolate it from the rest of the system and the floating end of the tubes are encased in silicon to enhance heat transfer between the tubes. The U shape allows flexibility for thermal expansion. Wash-coating was utilized to coat the channels with catalyst. Butane combustion was performed in one of the two tubes. Heat produced from this reaction was utilized to enhance ammonia reforming in the second tube. Yet another approach is described by Holladay et al. and is shown in Fig. 12.[58,59,76]A proprietary catalyst on a monolithic support was utilized to perform steam microreforming in an integrated system that included an external combustion heater, fuel vaporizer/preheater, reheater, heat exchanger, and CO filtering unit.[58,59]

Polymerase Chain Reaction

The PCR is another case with a huge potential market. The PCR method is commonly used to create copies of specific fragments of DNA. First, the individual strands of DNA are unwound and separated by heating to 90–96°C. Then, the DNA is cooled to 50–60°C so that primers can bind to the DNA. The primers allow the duplication process to start at a specific point on the DNA. Finally, the temperature is raised to 72°C and *Taq*-polymerase is used to grow two strands of DNA for the original single strand. Generally, the cycle is repeated perhaps 30 times to produce a factor of 2^{30} increase in DNA concentration.

In a conventional scale device, each cycle takes 5–10 min, for a total time of 150–300 min. However, it is possible to do more rapid thermal cycling in a microscale device owing due to the short length scales for heat transfer, thus reducing the analysis time. A schematic of such a device reported by Manz et al. is shown in Fig. 2. It utilizes flow channel repeatedly passing through three different temperature controlled zones on the chip to perform the required thermal cycling. Their device was able to perform 20 cycles in times ranging from 90 sec to 18.7 min depending on the flow rate through the system.[8] A different design is reported by Ramsey et al. (Fig. 13) which utilized an on-chip Peltier heater/cooler to perform the thermal cycling.[60] Electrophoretic sizing and detection is also performed on the same chip. Total analysis time, including 10 thermal cycles, was less than 20 min.[60]

Phosgene Production

Phosgene production in a packed-bed microreactor is a case study that illustrates the potential of the safety advantages of microreactors.[9] Commonly used as an intermediate in the production of isocyanates, which in turn are used in the manufacturing of pharmaceuticals, pesticides, and polyurethanes, phosgene is also extremely hazardous requiring specialized facilities for storage, and handling, and is subject to a variety of transportation restrictions. Microreactors present a convenient method for the production of phosgene at the point of use, minimizing the risks associated with large-scale storage and/or transport of phosgene. Operation of an array

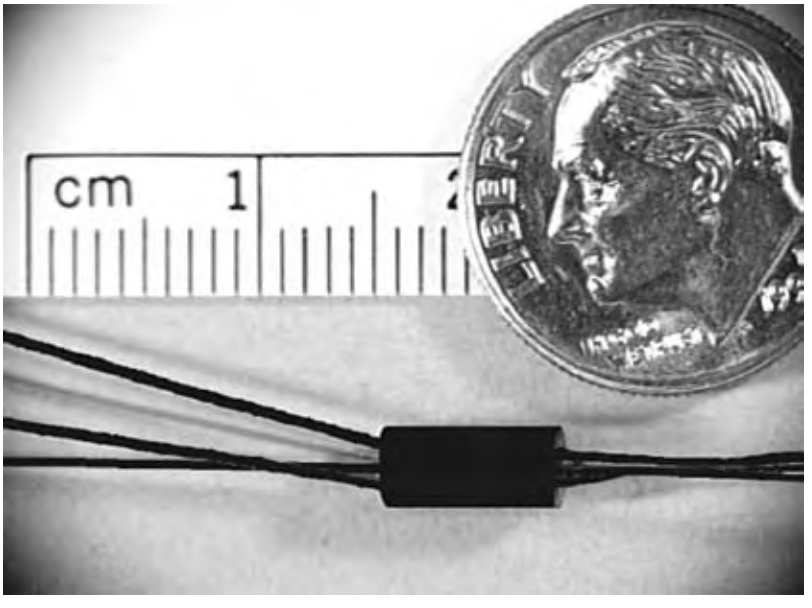


Fig. 12 An integrated fuel processing system from Holladay et al. (From Ref.[58].)

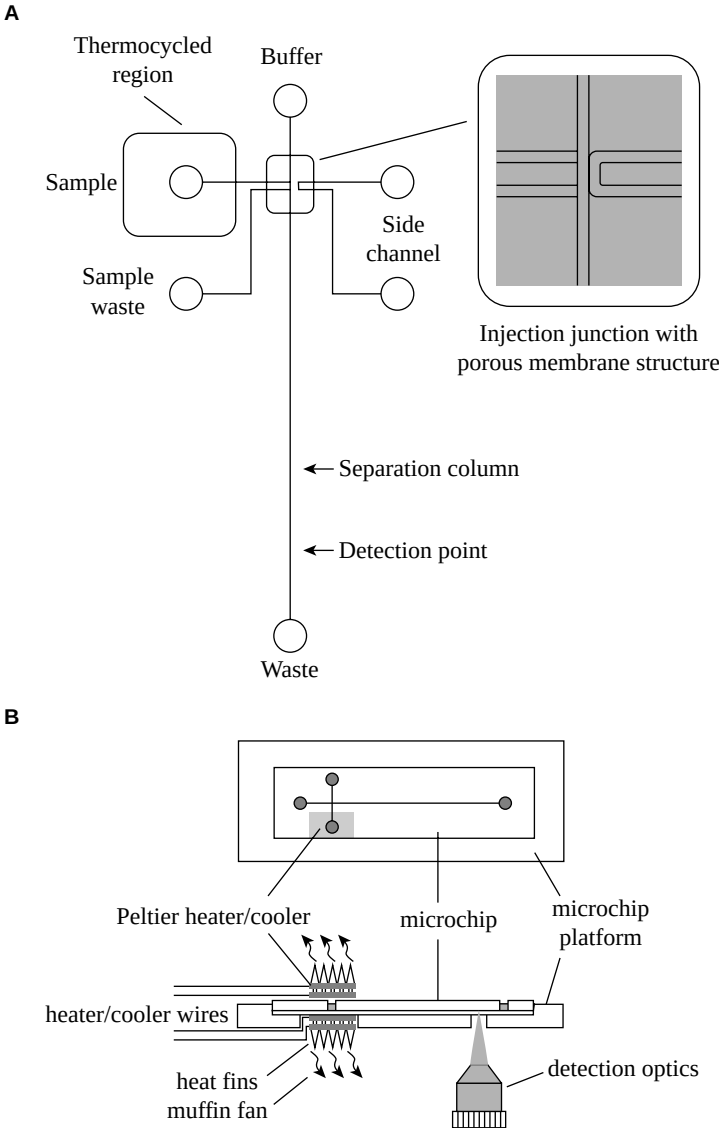


Fig. 13 Ploymerare chain reactor system on a chip reported by Ramsey et al. (A) Layout of chip. (B) On-chip Peltier heater/cooler for thermal cycling of PCR process. (From Ref.[60].)

of such reactors can allow significant production rates to be achieved. The failure of any one of these reactors, however, would only result in a minute chemical release, which mitigated the safety hazard.

The reactor described by Jensen et al. (Fig. 14) consists of a reaction channel 20 mm long, 300 μm deep, and 625 μm wide etched in single-crystal silicon and capped with Pyrex. A thick silicon dioxide coating protects the channels from being etched by chlorine used to produce phosgene. The reactor is loaded with catalyst by pulling a vacuum from the outlet of the reactor and feeding catalyst particles through feed channels perpendicular to the main reaction channel. A series of posts with 25 mm gaps at the outlet is used to keep the catalyst particles from escaping the reactor. Inlet flow is split among several streams to form a parallel micromixer at the inlet. Activated carbon particles between 53 and 73 μm in diameter with a surface area of $850\,m^2/g$ were used as catalyst.

The phosgene formation reaction is moderately exothermic ($\Delta H = -26$ kcal/mol). Despite this, Jensen et al. observed no temperature increase in the reactor when flows were switched from bypassing the reactor to going through the reactor owing to the high thermal conductivity of the silicon reactor and the short length scale for thermal conduction in the microreactor. No undesired side-product formation was observed either, presumably because hot spots, common in conventional reactors, were suppressed as a result of the rapid heat transport. Complete conversion was achieved in this system and significant production rates demonstrated. A 10-channel microreaction system like the one described by Jensen et al. could produce 11 g/hr. Greater production could of course be achieved by adding more reaction channels.

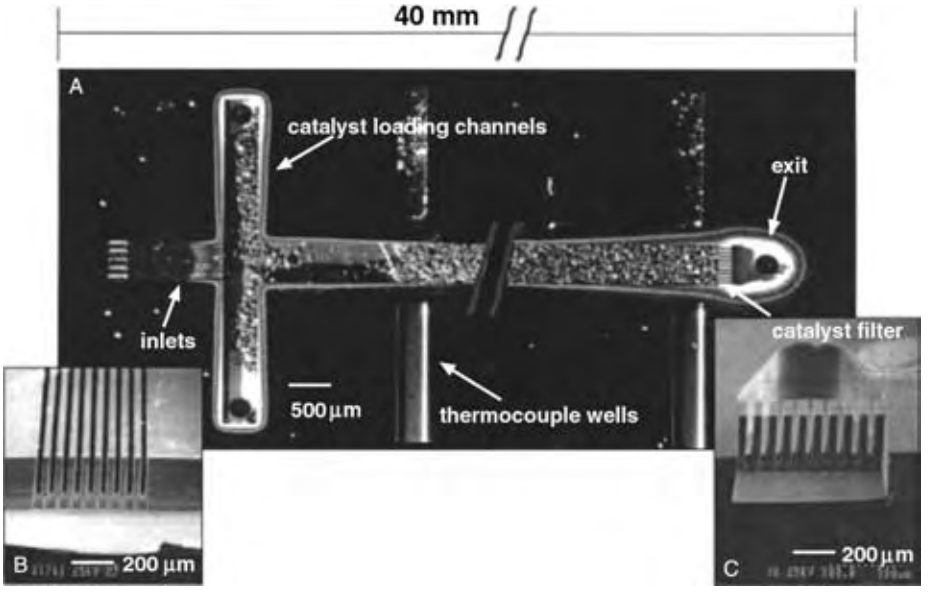


Fig. 14 A packed-bed microreactor used for phosgene production by Jensen et al.: (A) top view; (B) interleaved inlets; (C) catalyst filter. (From Ref.[9].)

Production of *N*-methoxycarbonyl-L-*tert*-leucine

Researchers at Johnson and Johnson have reported the use of microreactors in the drug development process.[61] They utilized a commercially available CYTOS benchtop system, shown in Fig. 15, to examine several reactions. One such reaction was the highly exothermic reaction to form *N*-methoxycarbonyl-L-*tert*-leucine by the addition of methylchloroformate to L-*tert* leucine. Such highly exothermic reactions present safety hazards in large-scale systems. Utilizing the CYTOS system, they were able to perform this reaction in the laboratory and achieved 91% yield.[61]

Photochemical Oxidation of Terpene

Short length scales combined with the use of optically transparent materials allow photoactivated reactions to be performed with much greater ease in microreactors than in conventional reactors. One illustration of these advantages is seen in terpene oxidation by singlet oxygen.[62] To efficiently carry out this reaction, illumination of the entire volume of the reactor with uniform light intensity and wavelength is desired. Complications in conventional reactors include the high absorptivity of the photosensitizers often used and excessive heating of the reactor by the light source. With microreactors,

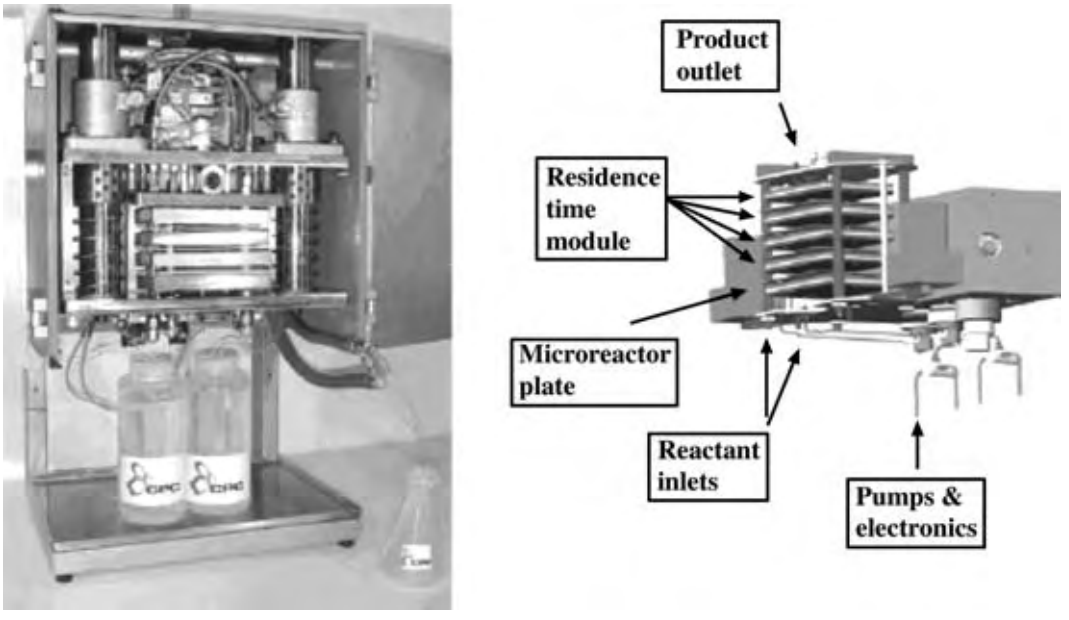


Fig. 15 CYTOS benchtop continuous microreactor system. (From Ref.[61].) *(View this art in color at www.dekker.com.)*

the pathlength for the light is greatly reduced, introducing several advantages. A lower-intensity light source can be used to fully and more uniformly illuminate the reagents, reducing not only the energy used for illumination, but also the undesired excess heating that creates problems with larger light sources.

MATERIALS AND FABRICATION

Nearly any material that can be used in conventional reactors can also be used to fabricate microreactors. Silicon is one of the most widely used materials for microreactors, primarily owing to the fact that microfabrication technology for silicon, borrowed largely from the microelectronics industry, is well developed and there is a large established knowledge base for these processes. Silicon micromachining can be accomplished through a wet etching procedure, such as in aqueous KOH, ethylene diamine pyrocatechol, tetramethyl ammonium hydroxide, or N_2H_4, or through a dry etching procedure, such as deep reactive ion etch, most commonly performed in a SF_6 plasma. A wide array of processes are available for creating structures and coatings from other materials on silicon substrates such as chemical vapor deposition, physical vapor deposition, and electrodeposition.

An inherent advantage of silicon microfabrication techniques is that numerous devices can be rapidly and cheaply produced. This presents the opportunity to develop disposable systems which would be valuable for many chemical analysis and medical applications or to operate many devices in parallel to scale-up production from microreaction systems. The large knowledge base for these processes due to their long history of use in the microelectronics industry also presents a major advantage. A major drawback to silicon microfabrication is the high cost of much of the associated equipment.

Microreactor material choices are by no means limited to silicon and other electronic materials as a wide array of microfabrication technologies for other materials have also been developed. Processes such as soft lithography (63), micromolding (64), and microstereolithography (10) have been utilized to fabricate microreactors from polymeric materials.[10,63,64] Electrodischarge machining has also been used to manufacture microreaction systems from metals and other conductive materials.[65] As can be seen, the choice of materials is generally not by the availability of microfabrication technologies for the material.

Often complex microchemical and microelectromechanical systems have components made in more than one substrate that must be bonded together to form the complete system. These systems also must be packaged to be used commercially. Bonding processes between silicon, glasses, polymers, ceramics, and metals are, hence, of great importance for producing practical microchemical systems. One of the most widely used bonding technologies in microfabrication is anodic bonding in which glass and silicon wafers are held together at temperatures around 400°C under an electric field. Many materials, including glasses and some polymers, can be bonded together thermally. Adhesives and eutectics are also approaches commonly used in bonding materials in microfabrication.

As can be seen, there are a broad array of microfabrication techniques available for many diverse materials. A thorough treatment of microfabrication technology is well beyond the scope of this work. For a more detailed treatment, the reader is referred to Madou's text on the subject.[66]

CONCLUSIONS

Microreaction engineering is a rapidly expanding field. As shown here, microreactors have significant inherent advantages for many applications. Because of their small size, they offer great opportunities for portable applications. Additionally, microreactors provide environmental, health, and safety advantages owing to the small amount of chemicals used and the ability to produce chemicals at the point of use, among other things. The greatest benefits of microreactors result from the enhanced heat and mass transport at short length scales.

The uses to which microreactors have been applied are diverse and cover many fields. Examples cited here include organic synthesis, combustion, reforming, and immunoassay applications. These are only a small sample of the applications to which microreactors have been applied and without doubt an even smaller sample of the future applications of these unique reaction systems.

ACKNOWLEDGMENTS

This work was supported by the Department of Defense Multidisciplinary University Research Initiative (MURI) program administered by the Army Research Office under contract DAAD19-01-1-0582. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Defense or the Army Research Office.

REFERENCES

1. Korbach, P.F.; Stewart, W.E. Kinetic and equilibrium studies of benzene hydrogenation in batch

recycle reactor. Ind. Eng. Chem. Fundam. **1964**, *3* (1), 24–27.
2. Pansing, W.F.; Malloy, J.B. A differential reactor system. Chem. Eng. Prog. **1962**, *58* (12), 53–54.
3. Hougen, O.A.; Watson, K.M. Solid catalysts and reaction rates-general principles. J. Ind. Eng. Chem. **1943**, *35*, 539–541.
4. Carberry, J.J. Heat and mass diffusional intrusions in catalytic reactor behavior. Catal. Rev. **1969**, *3* (1), 61–91.
5. Vern, J.; Weekman, W. Laboratory reactors and their limitations. AIChE J. **1974**, *20* (5), 833–840.
6. Denbigh, K. *Chemical Reactor Theory: An Introduction*; University Press: Cambridge, U.K., 1965.
7. http://world.casio.com/info/2002/fuelcell.html (accessed Aug 2004).
8. Kopp, M.U.; Mello, A.J.d.; Manz, A. Chemical amplification: continuous-flow PCR on a chip. Science **1998**, *280* (5366), 1046–1048.
9. Ajmera, S.K.; Losey, M.W.; Jensen, K.F.; Schmidt, M.A. Microfabricated packed-bed reactor for phosgene synthesis. AIChE J. **2001**, *47* (7), 1639–1647.
10. Nguyen, N.-T.; Wereley, S.T. *Fundamentals and Applications of Microfluidics*; Artech House, Inc.: Boston, MA, 2002.
11. Ho, C.; Tai, Y. Micro-electro-mechanical systems (MEMS) and fluid flows. Annu. Rev. Fluid Mech. **1998**, *30*, 579–612.
12. Gad-El-Hak, M. Flow physics. In *MEMS Handbook*; Gad-El-Hak, M., Ed.; CRC Press: Boca Raton, FC, 2002.
13. Koch, M.; Evans, A.; Brunnschweiler, A. *Microfluidic Technology and Applications*; Research Studies Press, Ltd.: Hertfordshire, U.K., 2000.
14. Gad-el-Hak, M. The fluid mechanics of microdevices—the Freeman Scholar Lecture. J. Fluids Eng. **1999**, *121* (1), 5–33.
15. Kenis, P.J.A.; Ismagilov, R.F.; Whitesides, G.M. Microfabrication inside capillaries using multiphase laminar flow patterning. Science **1999**, *285* (5424), 83–85.
16. Loose, W.; Hess, S. Rheology of dense fluids via nonequilibrium molecular hydrodynamics: shear thinning and ordering transition. Rheol. Acta **1989**, *28*, 91–101.
17. Thompson, P.A.; Troian, S.M. A general boundary condition for liquid flow at solid surfaces. Nature **1997**, *389*, 360–362.
18. Judy, J.; Maynes, D.; Webb, B.W. Characterization of frictional pressure drop for liquid flows through microchannels. Int. J. Heat Mass Transfer **2002**, *45*, 3477–3489.
19. Mas, N.d.; Gunther, A.; Schmidt, M.A.; Jenson, K.F. Microfabricated multiphase reactors for the selective direct fluorination of aromatics. Ind. Eng. Chem. Res. **2003**, *42*, 698–710.
20. Losey, M.W.; Jackman, R.J.; Firebaugh, S.L.; Schmidt, M.A.; Jensen, K.F. Design and fabrication of microfluidic devices for multiphase mixing and reaction. J. Microelectromech. Syst. **2002**, *11* (6), 709–717.
21. Taylor, G. Dispersion of soluble matter in solvent flowing slowly through a tube. Proc. R. Soc. A (Lond.) **1953**, *A219* (1137), 186–203.
22. Aris, R. On the dispersion of a solute in a fluid flowing through a tube. Proc. R. Soc. A (Lond.) **1956**, *235* (1200), 67–77.
23. Beard, D.A. Taylor dispersion of a solute in a microfluidic channel. J. Appl. Phys. **2001**, *89* (8), 4667–4669.
24. Dorfman, K.D.; Brenner, H. Comment on ''Taylor dispersion of a solute in a microfluidic channel.'' J. Appl. Phys. **2001**, *90* (12), 6553–6554.
25. Beard, D.A. Response to ''Comment on Taylor dispersion of a solute in a microfluidic channel.'' J. Appl. Phys. **2001**, *90* (12), 6555–6556.
26. Ghosal, S. Band broadening in a microcapillary with a stepwise change in the z-potential. Anal. Chem. **2002**, *74*, 4198–4203.
27. Arana, L.R.; Schaevitz, S.B.; Franz, A.J.; Schmidt, M.A.; Jensen, K.F. A microfabricated suspended-tube chemical reactor for thermally efficient fuel processing. J. Microelectromech. Syst. **2003**, *12* (5), 600–612.
28. Dummann, G.; Quittmann, U.; Groschel, L.; Agar, D.W.; Worz, O.; Morgenschweis, K. The capillary-microreactor: a new reactor concept for the intensification of heat and mass transfer in liquid–liquid reactions. Cataly. Today **2003**, *79–80*, 433–439.
29. Burns, J.R.; Ramshaw, C. Development of a microreactor for chemical production. Chem. Eng. Res. Des. **1999**, *77* (3), 206–211.
30. Hessel, V.; Ehrfeld, W.; Golbig, K.; Hofmann, C.; Jungwirth, S.; Lowe, H.; Richter, T.; Storz, M.; Wolf, A.; Worz, O.; Breysse, J. High temperature HCN generation in an integrated microreaction system. 3rd International Conference on Microreaction Technology, **2001**; 151–164.
31. Taghavi-Moghadam, S.; Kleemann, A.; Golbig, K.G. Microreaction technology as a novel approach to drag design, process development and reliability. Org. Process Res. Dev. **2001**, *5* (6), 652–658.
32. Lahann, J.; Balcells, M.; Rodon, T.; Lee, J.; Choi, I.S.; Jensen, K.F.; Langer, R. Reactive polymer coatings: a platform for patterning proteins and mammalian cells onto a broad range of materials. Langmuir **2002**, *18*, 3632–3638.

33. Lahann, J.; Balcells, M.; Lu, H.; Rodon, T.; Jensen, K.F.; Langer, R. Reactive polymer coatings: a first step toward surface engineering of microfluidic devices. Anal. Chem. **2003**, *75*, 2117–2122.
34. Sadykof, V.A.; Pavlova, S.N.; Saputina, N.F.; Zolotarskii, I.A.; Pakhomov, N.A.; Moroz, E.M.; Kuzmin, V.A.; Kalinkin, A.V. Oxidative dehydrogenation of propane over monoliths at short contact times. Catal. Today **2000**, *61*, 93–99.
35. Ganley, J.C.; Seebauer, E.G.; Masel, R.I. Microreactors for fuel conversion. Proceedings of the 40th Power Sources Conference, 2002; 367–371.
36. Ganley, J.C.; Seebauer, E.G.; Masel, R.I. Porous anodic alumina posts as a catalyst support in microreactors for production of hydrogen from ammonia. AIChE J **2004**, *50*, 829–835.
37. Jensen, K.F. Towards integrated microsystems for chemical synthesis. Micro Total Anal. Syst. **2002**, Proceedings of the 6th μTAS 2002 Symposium, Nara, Japan, 2002, 642–645.
38. Lee, J.; Kim, C.-J. Surface-tension-driven microactuation based on continuous electrowetting. J. Microelectromech. Syst. **2000**, *9* (2), 171–180.
39. Sammarco, T.S.; Burns, M.A. Thermocapillary pumping of discrete drops in microfabricated analysis devices. AIChE J. **1999**, *45* (2), 350–366.
40. Gallardo, B.S.; Gupta, V.K.; Eagerton, F.D.; Jong, L.I.; Craig, V.S.; Shah, R.R.; Abbott, N.L. Electrochemical principles for active control of liquids on submillimeter scales. Science **1999**, *283* (5398), 57–60.
41. Ichimura, K.; Oh, S.-K.; Nakagawa, M. Light-driven motion of liquids on a photoresponsive surface. Science **2000**, *288* (5471), 1624–1626.
42. Spadaccini, C.M.; Zhang, X.; Cadou, C.P.; Miki, N.; Waitz, I.A. Preliminary development of a hydrocarbon-fueled catalytic micro-combustor. Sensors Acutators A. Phys. **2003**, *103*, 219–224.
43. Waitz, I.A.; Gauba, G.; Tzeng, Y.-S. Combustors for micro-gas turbine engines. J. Fluids Eng. **1998**, *120* (1), 109–117.
44. Ehrfeld, W.; Hessel, V.; Lowe, H. *Microreactors: New Technology for Modern Chemistry*; Wiley–VCH: Weinheim, Germany, 2001.
45. Knight, J.B.; Vishwanath, A.; Brody, J.P.; Austin, R.H. Hydrodynamic focusing on a silicon chip: mixing nanoliters in microseconds. Phys. Rev. Lett. **1998**, *80* (17), 3863–3866.
46. Johnson, T.J.; Ross, D.; Locascio, L.E. Rapid microfluidic mixing. Anal. Chem. **2002**, *74* (1), 45–51.
47. Lin, Y.; Gerfen, G.J.; Rousseau, D.L.; Yeh, S.-R. Ultrafast microfluidic mixer and freeze-quenching device. Anal. Chem. **2003**, *75* (20), 5381–5386.
48. Yang, Z.; Matsumoto, S.; Goto, H.; Matsumoto, M.; Maeda, R. Ultrasonic micromixer for microfluidic systems. Sens. Actuators A. **2001**, *93* (3), 266–272.
49. Rife, J.C.; Bell, M.I.; Horowitz, J.S.; Kabler, M.N.; Auyeung, R.C.Y.; Kim, W.J. Miniature valveless ultrasonic pumps and mixers. Sens. Actuators A. **2000**, *86* (1–2), 135–140.
50. Tsai, J.-H.; Lin, L. Active microfluidic mixer and gas bubble filter driven by thermal bubble micropump. Sens. Actuators A. **2002**, *97–98*, 665–671.
51. Jones, S.W.; Young, W.R. Shear dispersion and anomalous diffusion by chaotic advection. J. Fluid Mech. **1994**, *280*, 149–172.
52. Bau, H.H.; Zhong, J.; Yi, M. A minute magneto hydro dynamic (MHD) mixer. Sens. Actuators B. **2001**, *79* (2–3), 207–215.
53. Oddy, M.H.; Santiago, J.G.; Mikkelsen, J.C. Electrokinetic instability micromixing. Anal. Chem. **2001**, *73* (24), 5822–5832.
54. Bier, W.; Keller, W.; Linder, G.; Seidel, D.; Schubert, K. Gas to gas heat transfer in micro heat exchangers. Chem. Eng. Process. **1993**, *32*, 33–43.
55. Stief, T.; Langer, O.-U.; Schubert, K. Numerical investigations of optimal heat conductivity in micro heat exchangers. Chem. Eng. Technol. **1999**, *21* (4), 297–303.
56. Peterson, R.B. Numerical modeling of conduction effects in microscale counterflow heat exchangers. Microscale Thermophys. Eng. **1999**, *3*, 17–30.
57. Cordero, N.; West, J.; Berney, H. Thermal modeling of ohmic heating microreactors. Microelectron. **2003**, *34*, 1137–1142.
58. Holladay, J.D.; Jones, E.O.; Phelps, M.; Hu, J. Microfuel processor for use in a miniature power supply. J. Power Sources **2002**, *108*, 21–27.
59. Palo, D.R.; Holladay, J.D.; Rozmiarek, R.T.; Guzman-Leong, C.E.; Wang, Y.; Hu, J.; Chin, Y.-H.; Dagle, R.A.; Baker, E.G. Development of a soldier-portable fuel cell power system part I: a bread-board methanol fuel processor. J. Power Sources **2002**, *108*, 28–34.
60. Khandurina, J.; McKnight, T.E.; Jacobson, S.C.; Waters, L.C.; Foote, R.S.; Ramsey, J.M. Integrated system for rapid PCR-based DNA analysis in microfluidic devices. Anal. Chem. **2000**, *72* (13), 2995–3000.
61. Zhang, X.; Stefanick, S.; Villani, F.J. Application of microreactor technology in process development. Org. Process Res. Dev. **2004**, *8*, 455–460.
62. Wooton, R.C.R.; Fortt, R.; Mello, A.J.d. A microfabricated nanoreactor for safe, continuous generation and use of singlet oxygen. Org. Process Res. Dev. **2002**, *6* (2), 187–189.

63. Xia, Y.; Whitesides, G.M. Soft lithography. Ann. Rev. Mater. Sci. **1998**, *28*, 153–184.
64. Becker, H.; Heim, U. Hot embossing as a method for the fabrication of polymer high aspect ratio structures. Sens. Actuators A. **2000**, *83* (1–3), 130–135.
65. Stampfl, J.; Leitgeb, R.; Cheng, Y.-L.; Prinz, F.B. Electro-discharge machining of mesoscopic parts with electroplated copper and hot-pressed silver tungsten electrodes. J. Micromech. Microeng. **2000**, *10*, 1–6.
66. Madou, M.J. *Fundamentals of Microfabrication: The Science of Miniaturization*, 2nd Ed.; CRC Press: Boca Raton, FL, 2002.
67. Ni, Z.; Seebauer, E.G.; Masel, R.I. ''Effects of microreactor geometry on performance: differences between posted reactors and channel reactors.'' Ind. Eng. Chem. Res. **2005**, *44*, 4267–4271.
68. Baier, T.; Drese, K.S.; Schonfeld, F.; Schwab, U. A microfluidic mixing network. Chem. Eng. Tech. **2005**, *2*, 362–366.
69. Nguyen, N.T.; Wu, Z.G. Micromixers—a review J. Micromech. Microeng. **2005**, *15* (2), R1–R16.
70. Vuppu, A.K.; Garcia, A.A.; Saha, S.K.; Phelan, P.E.; Hayes, M.A.; Calhoun, R. Modeling microflow and stirring around a microrotor in creeping flow using a quasi-steady-state analysis. Lab Chip. **2004**, *4* (3), 201–208.
71. Bringer, M.R.; Gerdts, C.J.; Song, H.; Tice, J.D.; Ismagilov, R.F. Microfluidic systems for chemical kinetics that rely on chaotic mixing in droplets. Philos. Trans. R. Soc. Lond. Ser. A—Math. Phys. Eng. Sci. **2004**, *362*, 1087–1104.
72. Floyd, T.M.; Schmidt, M.A.; Jensen, K.F. Silicon micromixers with infrared detection for studies of liquid-phase reactions. Ind. Eng. Chem. Res. **2005**, *44* (8), 2351–2358.
73. Stone, H.A.; Stroock, A.D.; Ajdari, A. Engineering flows in small devices: Microfluidics toward a lab-on-a-chip. Annu. Rev. Fluid Mech. **2004**, *36*, 381–411.
74. Watts, P.; Haswell, S.J. The application of micro reactors for organic synthesis. Chem. Soc. Rev. **2005**, *34* (3), 235–246.
75. Doku, G.N.; Verboom, W.; Reinhoudt, D.N.; van den Berg, A. On-microchip multiphase chemistry—a review of microreactor design principles and reagent contacting modes. Tetrahedron. **2004**, *61* (11), 2733–2742.
76. Holladay, J.D.; Wang, Y.; Jones, E. Review of developments in portable hydrogen production using microreactor technology. Chem. Rev. **2004**, *104*, 4767–4789.

Microscale Fuel Cells

J. D. Holladay
V. V. Viswanathan
Pacific Northwest National Laboratory, Richland, Washington, U.S.A.

INTRODUCTION

Fuel cell research has intensified in the past few years. In stationary and mobile applications, fuel cells are attractive for their potential high efficiency and low or no emissions.[1] With increasing number of people using laptop computers and other handheld electronic devices, the desire for longer operating time than what current battery technology provides has sparked interest in 10–100 W_e fuel cells.[2] But perhaps some of the most innovative works on fuel cells have been research dedicated to applying silicon fabrication techniques to fuel cell technology, and creating low power microscale fuel cells applicable to microelectromechanical systems (MEMS), microsensors, cell phones, PDAs, and other low power (0.001–5 W_e) applications. In this small power range, fuel cells offer the decoupling of the energy converter from the energy storage, which may enable longer operating times and instant or near-instant charging.[3] To date, most of the microscale fuel cells being developed are based on proton exchange membrane fuel cell (PEMFC) or direct methanol fuel cell (DMFC) technology. The following section discusses the requirements and considerations that need to be addressed in the development of microscale fuel cells, as well as some proposed designs and fabrication strategies.

FUEL CELL SYSTEM CONSIDERATIONS

There are many issues that need to be addressed for the making of a successful system: fuel cell design, oxidant and fuel delivery, thermal management, and power conditioning. These issues are not independent, because the performance, or lack thereof, of one affects the others. To be successful, all of these components need to be integrated and optimized to work together.

Fuel Cell Design

Fuel cell design needs to include strategies to enable high electrical conductivity of the current collectors and electrodes, high ionic conductivity in the electrolyte and electrodes, appropriate catalysts, and high fuel utilization. High electrical conductivity is required between the catalyst and the current collectors and between the current collectors and external loads. However, the cathode and the anode of a particular cell should be electrically insulated from each other. This is done by the electrolyte membrane separator.[4]

On the other hand, there must be good ionic conductivity between the electrodes and the catalyst surface. For most polymeric membranes, ionic conductivity is strongly dependent on the state of hydration of the membrane.[5] For example, the anode can dry out, thereby seriously hindering the performance. As the anode dries out, localized hot spots can form, which can irreversibly damage the fuel cell. At the other extreme, during high current operation, water produced at the cathode can condense and form a thin layer that blocks oxygen from the catalyst surface.[6] Thus, excess water must be removed from the cathode, while keeping the anode appropriately humidified. There are four modes of water transport in the stack: water produced electrochemically through applied load, that produced because of crossover, water crossover because of proton conduction, and that through simple permeation. Water loss and management in fuel cells are of greater importance as the size decreases. Various water management approaches have been proposed.[7,8] Strategies to manage water include humidifying the fuel and/or oxidant, using an active air blower to provide the oxidant, and removing any excess water.[6]

Water clogging issues in DMFC are similar to those in PEMFC. At high current density, clogging is more severe, especially at low stoichiometric ratio of oxygen, with the problem becoming more exacerbated for cells downstream from the air inlet. Controlling the pressure drop in the cathode flow field was found to be critical for water management. Higher temperatures also increase water transport across the membranes, thus creating water management issues. Higher airflows were needed at high temperature to support higher current density. For DMFCs, this was probably because of high methanol crossover at low current density. Hence, for low power density operation, it appears that unless crossover can be minimized, it is preferable to operate at a low temperature (20°C) in a DMFC system.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120019081

Electrocatalyst selection and design are the key aspects of PEM fuel cells. The most popular catalyst is platinum for the anode and the cathode in pure hydrogen cells. For direct methanol fuel cells and for hydrogen cells with carbon monoxide present, a platinum/ruthenium alloy is used.

Fuel utilization plays an important role in determining system efficiency. Obviously, it is desirable to use as much fuel as possible; however, when nearly all of it is used, fuel delivery to the electrode becomes increasingly difficult. This is because of the diffusional mass-transport limitations restricting the rate of transport of the unused fuel to the active catalyst sites. A significant drop in the current density occurs in the depleted region. This is more significant in direct methanol fuel cells than in hydrogen dead ended fuel cells because of the rapid diffusion of hydrogen. Therefore, in most cases, when operating on methanol the drop in the cell potential prohibits operating the cell to full fuel utilization, and the methanol fuel must either be recycled or expelled.[6] As methanol has a high toxicity, expelling it to the environment is generally not desired, leaving recycling as the preferred option.[6]

Fuel and Oxidant Delivery

The fuel cell does not contain its own fuel; therefore fuel storage and delivery system are very important to a successful system. One of the first questions that needs to be addressed is whether the fuel cell will operate on hydrogen, reformed fuel, or direct methanol. Another important question is whether the system will be active or passive with regard to the oxidant.

Hydrogen can be stored as a gas or a liquid, and in metal hydrides, chemical hydrides, and carbon storage material.[6] If hydrogen is stored as a gas, it requires high pressures and often a large volume, which negates one of the key advantages of a hydrogen fuel cell—high power density per unit volume.[6] Metal hydrides can safely store hydrogen and, depending on the media, it has an acceptable energy density.[9] However, the specific energy is very poor and the materials tend to be expensive.[9] Because hydrogen in all these cases is stored at pressure, this simplifies the fuel supply issue for the fuel cell. It can be supplied passively, requiring no pumps, with only a latching valve to turn the supply on and off.

Using a direct methanol system may improve the energy density of the power supply and make fuel storage easier, though at the cost of increased complexity within the fuel cell. The open circuit of the DMFC decreases with the concentration of methanol feed to the anode owing to methanol crossover and leads to lower performance at low current density because of methanol oxidation at the cathode. However, at a high current density, higher methanol concentration gives better performance owing to lower concentration polarization. Hence, a methanol tolerant cathode would significantly improve the fuel cell performance. Membranes that minimize methanol crossover would also be beneficial. Therefore, to successfully use a DMFC, methanol concentration must be tightly controlled. This requires a pump and a mixer, which add to the complexity to the system, and using passive (wicking) techniques to provide methanol at a low rate so that it has a chance of reacting before the concentrations get too high or by decreasing methanol crossover by tailoring the membrane properties.

Several methods have been used to determine methanol transport through perfluorosulfonic acid membranes. Verdrugge[10] measured methanol diffusivity of Nafion equilibrated with sulfuric acid at room temperature using radioactive tracers and estimated the effective diffusion coefficient to be $1.15 \times 10^{-5}\,cm^2/sec$ at 25°C for Nafion 117. Kauranen and Skou[11] measured time responses of anodic peak currents on two working electrodes and estimated that the permeability of methanol for Nafion 117 at 60°C was $4.9 \times 10^{-6}\,cm^2/s$, and the activation energy was 12 kJ/mol. Narayanan et al.[12] measured the methanol crossover rates by estimating the carbon dioxide content of the cathode exit stream. The crossover rate decreased with increasing current density because of an increased utilization of methanol at high current densities. Increasing the membrane thickness from 5 to 14 mils caused the crossover to reduce by 40–50%. Heinzel et al.[13] described a technique to measure methanol transport through Nafion membranes. Methanol was completely electro-oxidized on the opposite electrode at high electrode potential. Both the diffusion coefficient and the methanol concentration were determined from the measured transient current density following a potential step. Corrections for electro-osmotic drag were found necessary even for low methanol concentrations. The results agreed well with NMR measurements.

A number of organizations are involved in developing membranes with lower methanol permeability. At Giner Inc., Nafion membrane was impregnated with cross-linked polystyrene.[14] Other approaches include electrochemically polymerizing a cation edge film on the cathode to limit methanol diffusion to the cathode,[15] imbibing Nafion 117 with concentrated H_3PO_4,[16] and using membranes based on tin mordenite.[17] Ma, Cheng, and Zhao sputtered a thin layer of Pt/Pd–Ag/Pt on Nafion membrane to reduce methanol crossover.[18] This membrane reduced methanol crossover and gave higher performance than a Nafion membrane cell, with performance increasing with coating thickness up to 1 μm.

The most complicated way to provide fuel to a microscale fuel cell is the use of reforming technology to process a hydrocarbon into a hydrogen rich stream,

which is fed to the fuel cell. This process typically requires the use of multiple reactors and operates at higher temperatures (200–300°C for methanol processing[19] and >600°C for other hydrocarbon fuels[20]). Reforming of hydrocarbons theoretically enables higher energy densities to be achieved. The fuel cell operating on reformate hydrogen is often smaller, requires lower platinum loadings, and is more efficient than a direct methanol fuel cell. The disadvantages include increased thermal management requirements, more unit operations to control, and the need for liquid and air pumps.

Thermal Management

At maximum power, a typical fuel cell operating on hydrogen is only 50% efficient. This translates into 1 W of thermal power that needs to be dissipated for every watt of electricity produced.[21,22] A battery typically operates at an efficiency of 80% or more. Therefore, the use of a fuel cell can increase the thermal load in an electronic device by 60–100%.[21,22] The fuel cell efficiency can be improved by operating the stack at a higher voltage. However, this requires oversizing the stack, which increases the mass and the volume of the power supply, and also, perhaps more significantly, increased cost.

Thermal management has been achieved by passive convection cooling, active air cooling, or liquid cooling. The simplest cooling technique is to design the stack that is able to dissipate the generated heat using natural convection. This design eliminates the need for temperature monitoring, active cooling elements that require parasitic power to operate, and also increased controls.[21] The disadvantages of this approach are increased size because of large cooling plates and heat sinks, and also thermal gradients that may develop in the fuel cell, limiting the size and power of the fuel cell.[21] In addition, space limitations and high ambient temperatures may limit the effectiveness of natural convection cooling.[23] The active air cooling technique uses a blower or a fan to cool the stack. Special cooling plates are again used. These plates are often smaller than those used for natural convection, but are typically larger than those used in liquid cooled systems. The final approach is liquid cooling. This is generally used in larger applications (>1 kW) where large amounts of heat are produced.[23] Liquid cooling stacks are typically more compact and have smaller thermal gradients, enabling high power densities. The disadvantages are the need for a radiator to cool the liquid and the increased complexity of the system.[23] Depending on the power output of the microscale fuel cell, natural convection cooling and active air cooling are typically used.[21,23]

Power Conditioning and Load Handling

Microscale fuel cells are being developed to power a wide range of portable electronic goods. The input voltages and duty cycles will vary from application to application. Fuel cell based power supply must be able to provide the appropriate voltage (1V, 3V, 12V, etc. depending on the device) and also respond to sudden changes in power demand even from a ''cold start.'' The fuel cell voltage is typically achieved by increasing the number of cells in the stack until the desired voltage is attained or by using a DC to DC converter to step up or step down the fuel cell stack voltage. Load handling is a more difficult challenge. Under mild conditions, this may not be a problem; however, under extreme cold startup, this may be a significant challenge.[6,22] In addition, membrane hydration may also be required if the system is dry.[6,22] The change in demand is especially significant for direct methanol systems where control of the methanol concentration is required. For example, a sudden decrease in power demand may flood the system with methanol, thus significantly increasing crossover and its problems, whereas a sudden increase in power demand will starve the system. In either case, the fuel cell may not be able to respond to the demand.[6,22]

The two most common strategies to handle load changes are to oversize the fuel cell and to include a battery in the system to handle peak loads.[6,22] There has been significant development in microscale batteries that can be integrated into the smaller systems (see for example Refs.[24–26]), while more conventional batteries can be used in the ''larger'' systems.[6,22] The most extreme example of this hybrid technology is the battery handling all loads with the fuel cell trickle charging the battery. This setup lengthens the battery's life and enables the load handling during transients; however, it can only be used when the duty cycle has periods of low power that enable the fuel cell to completely or significantly charge the battery.[6,22]

FABRICATION

To build an efficient, high-quality microscale fuel cell, microfabrication techniques need to be combined with appropriate materials such as Nafion based membrane electrode assemblies (MEAs). These techniques must be able to produce three-dimensional structures, allow reactant and product flow into and out of the device, process appropriate materials, and should be of low cost. Fortunately, traditional thin film techniques can be modified for microscale fuel cell fabrication, while maintaining their advantages of surface preparation, sensor integration, and finishing or packaging. In addition, other techniques are also available and are discussed in the following sections.

Membrane Electrode Assembly Fabrication

Some common ways of fabricating the MEA for microscale fuel cells include hot pressing the membrane onto preformed electrodes or hot pressing a complete MEA onto a substrate, screen printing a membrane onto a substrate and depositing the electrodes on either side, and using spin deposition to deposit the membrane followed by electrode deposition.

In the hot pressing technique, the membrane is often pretreated prior to use. The exact pretreatment depends upon the membrane used, but the pretreatment of Nafion 112 is used as an example here. The pretreatment of Nafion 112 often consists of a four-step sequential boiling process.[27,28] The treatment was as follows: 3% H_2O_2, deionized water, 0.5 M H_2SO_4, and final deionized water. The membrane was placed in each boiling solution for 1 hr.[27,28] The membrane can then be hot pressed onto a substrate or onto cathode and anode electrodes (such as catalysts loaded onto a supported carbon paper).[27,28] Typical hot press conditions are a temperature of 135°C and a pressure of 30 MPa.[27,28]

Screen printing membranes onto a substrate is another way of forming the fuel cell. Wainright et al.[29] have developed a procedure, which begins with a typical 5% Nafion solution. This is modified to increase its viscosity to print a well-defined film.[29] The modification consists of adding higher boiling solvents, such as dimethyl sulfoxide (DMSO) or ethylene glycol, and then concentrating the solution. Using this solution, the films are cast and the solvent is then removed under vacuum at approximately 80°C.[29] They were able to cast 25 μm thick free standing films from 10 to 15 wt.% solutions.[29] Once cast, the electrodes are deposited on the films by printing a slurry of C/Pt particles (e.g., 40 wt.% Pt on XC-72 from Etek Inc.).[29] Other ways of forming the electrodes include deposition of the films onto preformed electrode substrate, vapor or sputter deposition techniques, and spray coating with an air brush.[30]

A third common way of depositing ion conducting membranes is to use a spin coating technique.[31] The specifics of this technique will depend on the viscosity of the solution to be deposited and the thickness of the desired layer. Morse et al.[31] used the following process to form their electrolyte membranes. They first formed a catalyst layer by electron beam deposition of 5 nm of Pt onto a prepared substrate. They used a perfluorinated sulfonic acid polymer in a 5 wt.% of a 950 weight equivalent solution. It was deposited over the catalyst using a pipette and spun at 500 rpm for 30 sec.[31] After baking at 120°C for 30 min, another electrode layer was deposited by electron beam deposition of 5 nm of Pt.[31]

Microfabrication

This section briefly reviews the material selection and some common microfabrication techniques.

Materials

The materials used in fuel cell construction must be chemically compatible, have appropriate thermal properties, and be structurally sound. In Table 1 are listed the advantages and disadvantages of some common materials used in microfabrication. There are three general categories of materials: metals, silicon (which includes materials containing silicon or those that are processed with semiconductor fabrication techniques), and low-temperature cofired ceramics (LTCC).[32,33] The material selection process is complicated by the fabrication process, as it may alter the material properties. Although microscale devices have the potential to be manufactured using high-volume techniques, the design, material, and fabrication selections must take cost into account to avoid developing a laboratory curiosity rather than a commercially viable device.[34]

Fabrication techniques

Popular microfabrication methods include lithographie, galvanoformung, abformtechnik (LIGA), wet and dry etching processes, micromachining, lamination, and soft lithography. An overview of these techniques is given here and specifics can be found in the references.

Table 1 Microreactor material and benefits and drawbacks

Substrate	Benefits	Drawbacks
Metal	Standard fabrication techniques; durable; low to modest costs; no clean room required	Poor compatibility with ceramics and glass
Silicon	Well-characterized silicon fabrication techniques; high precision; low cost	Fragile; requires a clean room
LTCC	Flexible fabrication; refractory and durable; low cost; no clean room required	Nonstandard fabrication; sealing

(From Refs.[32,33].)

Lithographie, Galvanoformung, Abformtechnik (LIGA). The LIGA technique, developed by researchers in Germany, with Forschnungszentrum Karlsruhe and the Institute for Microtechnology Mainz (IMM) being the primary drivers, combines deep lithography, electroplating, and molding[34,35] using three general steps. Pattern transfer using a serial beam writing process or mask into a photoresist or a special photosensitive epoxy, such as SU8, deposited on an electrically conductive substrate is the first step.[34] After removal of the undesired material (e.g., the developed mask), electroplating is used to form a relief structure on the exposed substrate. To complete the second step, the resist or epoxy is removed leaving the metallic structure. In some applications, this structure may be the final product; however, usually it is used as a master tool for a replication process (such as injection molding, casting, or embossing), which is the final step.[34,35]

Wet and Dry Etching Processes. The most common processes in microfabrication are the wet and dry etching processes. Liquid etch solutions (such as potassium hydroxide for silicon) are used to remove unwanted material from the substrate anisotropically in a wet etch. Dry etching uses plasmas or reactive plasmas and may be anisotropic or isotropic, depending on the source.[34] The procedures, fluids, temperatures, and times are material- and pattern dependent. One form of dry etching that is increasingly being used is deep reactive ion etching (DRIE).[35] Because wet etching is an anisotropic process, it has strong restrictions on the geometries that can be made. Isotropic dry etching enables a wider variety of geometries but is more limited in the materials that can be etched. The initial step in the process is pattern transfer into a photoresist or other protective layer. The unprotected substrate areas are etched using either technique. Next, the protective layer is removed.[34]

Micromachining. There are three main techniques in micromachining: milling, laser radiation, and electrodischarge machining. These processes typically do not need the initial deposition of a resist, an epoxy, or other protective layers. The equipment tends to be less expensive and has lower maintenance costs compared with that used in LIGA and wet/dry etching. Though not required, computer control is usually used. The milling technique includes traditional milling, turning, and grinding, but with ultraprecision machines to produce small features.[34,35] The feature sizes are not as small as of those constructed from LIGA and other fabrication processes; however, almost any material can be used. As the name implies, the laser radiation uses lasers to remove material or, in some situations, to build up material.[34] This technique removes the material by melting, evaporation, decomposition, photoablation, or a combination of these. Some development involves building structures by photochemically crosslinking inorganic compounds or powder solidification by laser sintering. Both processes can be used to make devices to critical dimensions of approximately 10 μm; however, the surfaces tend to be rough, which may cause problems in some cases.[34] Electrodischarge machining (EDM), also called microelectrodischarge machining, removes material by small sparks in a dielectric fluid, such as deionized water or oil, between the work piece and an electrode. This technique is limited to electrically conductive work pieces. The main disadvantages of this type of machining are the relatively rough surfaces, limitations in miniaturization to the size of the electrodes and spark, and relatively long machining times, which may limit the technology to prototyping and manufacture of mold inserts.[34]

Lamination. Lamination processes are typically done by stacking several sheets with different patterns and then braze or bond them together. In typical MEMS-based approaches, DRIE is used to pattern silicon or other substrates. The substrates are then stacked and bonded together using silicon bonding methods. Another approach, used by researchers at the Pacific Northwest National Laboratory (PNNL), is the fabrication of thin metal laminates using stamping, embossing, or processes described previously.[35] After stacking, these laminates are brazed or diffusion bonded into a single block.[35–37] Lamination is also particularly well suited for fabricating ceramic devices. Ceramic tapes in the ''green state'' are cut, molded, laser cut, etc. to the desired pattern. The layers are stacked and then cured in high-temperature furnaces, bonding them into a single structure.[35]

Soft Lithography. Soft lithography refers to a collection of techniques originally developed by Xia and Whitesides.[37] They typically involve pattern transfer, often by stamping, using an elastomer such as poly (dimethylsiloxane). This technique has been combined with other polymers, electroplating techniques, or molding of ceramics to make low-cost fast prototyping of devices.[35,37]

MICROSCALE FUEL CELL DESIGNS

In designing a microscale fuel cell, there are several considerations that need to be accounted for. These include the following: is the fuel cell to be completely active or passive? will it operate at room temperature or elevated temperatures? will the fuel be at atmospheric pressure or elevated pressure? will external humidification be required? and finally, will the fabrication techniques

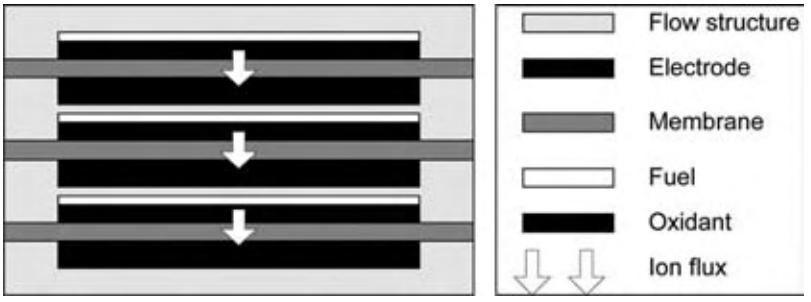


Fig. 1 Conventional bipolar plate configuration. (From Ref.[3].)

be compatible with the materials selected?[29] Several designs are reviewed here: a conventional bipolar plate configuration, a banded configuration, a ''flip-flop'' design, a monolithic configuration, and a mixed reactant configuration.

Conventional Bipolar Design

One way to make a microfuel cell is to size down conventional stacks. In this design (Fig. 1), the MEA, gas diffusion layer (GDL), if used, and bipolar plates are made thinner and smaller in area to make the microfuel cell. This design has the advantages of being able to achieve high voltages and one that most conventional fuel cell manufacturers prefer. For microscale fuel cells, this is a difficult design to manufacture, as it requires a large number of components and is very difficult to manifold.[27,28]

Banded Fuel Cell Configuration

In the banded fuel cell design, the cells are in a side-by-side planar array (Fig. 2). Higher voltages are achieved by overlapping adjacent cell electrodes. For example, the anode of one cell crosses the electrolyte to become the cathode of the adjacent cell.[27,28,38,39] These crossmembrane connections can be made by the use of edge-tabs or by routing breaches through the membrane.[3] This configuration enables very compact designs that show promise of improved volumetric packaging compared with conventional bipolar stacks in microscale fuel cells or other low power designs.[38] Other advantages of this design include relatively easy packaging and simpler manifolding of the fuel and oxidant. Major limitations are the crossmembrane interconnections. Interconnections by edge-tabs limit design flexibility and may require longer conductive lengths, which could increase the resistive losses in the cell. However, edge-tabs also offer the possibility of connecting the cells in parallel and/or serial configurations that would enable a wide range of voltages and currents to be fabricated.[29] For microscale cells, these losses may be minimized, but could still affect the cell performance.[3] If the interconnections are made by breaching the membrane, then the problem is sealing, especially when a polymer membrane that swells is employed.[3]

''Flip-Flop'' Fuel Cell Configuration

The ''flip-flop'' fuel cell design was developed to eliminate the need for crossmembrane interconnection of the electrodes that is apparent in the banded design. In this configuration, the anode and the cathode of adjacent cells are still electrically connected, but the connection no longer needs to cross the electrolyte membrane, as the anode and the cathode of adjacent cells are on the same side of the electrolyte membrane (Fig. 3).[3,27,28] The benefits of such a design include a fully continuous electrolyte membrane, all interconnects being on the same level so that all the electrodes are on the same side, and its conductor paths may have a lower resistance because this design does not require electronically conductive paths to the perimeter.[3,27,28] The disadvantage is the increased difficulty in sealing and in gas manifolding

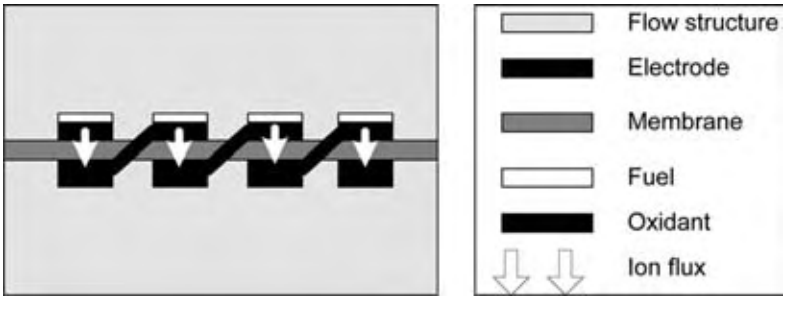


Fig. 2 Banded fuel cell configuration. (From Ref.[3].)

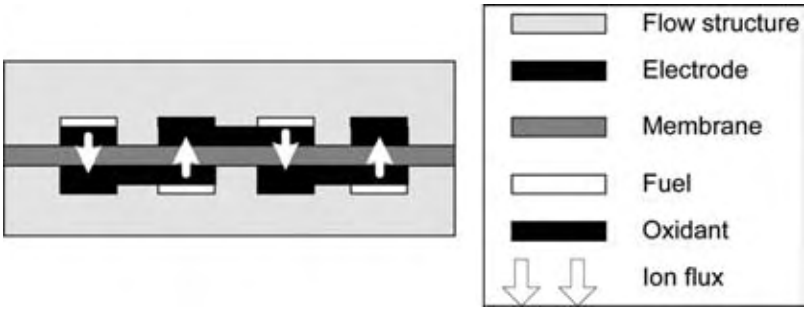


Fig. 3 ''Flip-flop'' fuel cell configuration. (From Ref.[3].)

because of the requirement of the reactant chambers to alternate between fuel and oxidant.[3,27,28]

Lee et al.[3] fabricated a four-cell assembly banded fuel cell as follows. On a 100 mm diameter silicon substrate (500 μm thick), the flow chambers were etched 200 μm deep and square flow distribution pillars (100 μm × 100 μm) were arranged in a rectangular array 100 μm apart. Cross channels were etched on different planar levels of the wafer to provide appropriate gas routing for the fuel and oxidant. These routes were in a checkerboard design. The etchant gas was SF_6 (130 sccm, 36 mT) and intermittent passivation with C_4F_8 (85 sccm, 18 mT) in an STS Multiplex ICP Deep Reactive Ion Etcher with an inductively coupled plasma at 13.56 MHz, coil rf power 600 W, and electrode rf power 120 W. After the etch was complete, a thermal oxide was grown on the wafer in a tube furnace. Electrically conductive regions were made by first depositing 15 nm of chromium (as adhesion promoter) followed by 100 nm of gold by electron beam evaporation. The prepared flow channels were then pressed onto MEAs by supplementary backing plates to interface gas fittings. Silicone gaskets were used to seal the substrates and backing plates together. Hand-tightened fasteners were used to supply the pressure. The cells were tested with compressed hydrogen and oxygen and were able to provide over 40 mW/cm^2 of power.[3]

Monolithic Fuel Cell Configuration

The monolithic design has been called an ''unfolded fuel cell'' with the anode and cathode on the same substrate.[6] The cathode and the anode of a particular cell are adjacent with each other in a coplanar configuration. Multiple cells are fabricated next to each other such that the cathode and the anode of adjacent cells are adjoining (Fig. 4). Advantages of this design include the ability to fabricate all the fuel cell components on the same structure, improved humidification control because the top structure is the moisture sensitive electrolyte membrane, very small separations between cathode and anode are now possible, and the cathode and anode electrodes can be sized differently enhancing the surface area of the limiting electrode to improve the cell's performance.[6] The main disadvantages of this design are the lower power density because the area is increased by 50% compared with a stacked design, the current distribution is not uniform, and also large metal lines are required to pull out the current.[6]

Mixed Fuel and Oxidant Fuel Cell Configuration

The mixed fuel and oxidant fuel cell design is similar to the monolithic fuel cell configuration except that there is only one reactant flow channel for both the fuel and the oxidant (Fig. 5).[6,40,41] Two approaches to this fuel cell configuration include the use of very selective catalysts for the electrodes so that the catalyst that reacts the fuel is significantly less active in the reaction of the oxidant and vice versa.[6,40] A second strategy is to have the component that is more active be consumed on the near side of the electrolyte membrane separator.

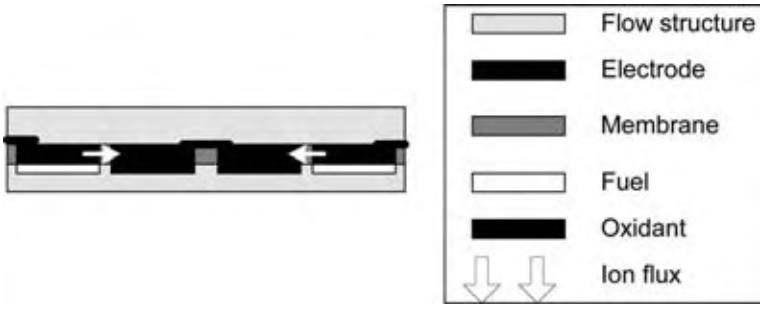


Fig. 4 Monolithic fuel cell configuration. Note: the two end electrodes are electronically connected.[3,6]

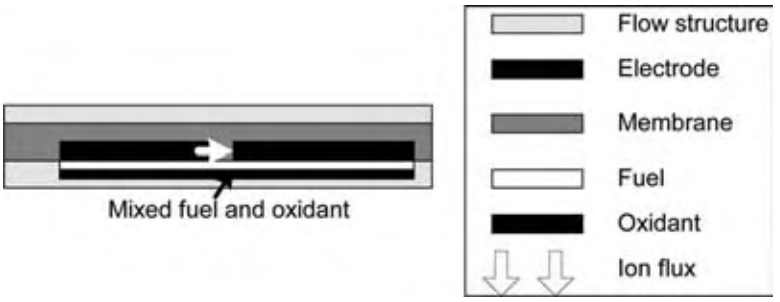


Fig. 5 Mixed fuel and oxidant fuel cell configuration. (From Ref.[3].)

The less active reactant is then free to diffuse to the other electrode and react.[6,41] The advantage of these designs is simple design. The main disadvantages are poor kinetic performance of the electrodes in the first strategy and operating the fuel cell with a mixed potential for the second.[6]

MINIATURE DMFC

There have been significant works in the development of direct methanol fuel cells at JPL,[12,42,43] University of Newcastle,[44] Korea Institute of Science and Technology,[45] Los Alamos National Laboratory,[13] and other institutions. The performance of the DMFC is highly dependent on methanol crossover, catalyst loading, and methanol concentration. At high temperature and high methanol concentration, the kinetics for methanol oxidation is highly enhanced and methanol crossover is also increased under these conditions. Methanol crossover has been extensively studied.[13,46,47] It has been found that high temperature increases methanol crossover, while high current density decreases crossover. Hence, at low current density and high temperature, the crossover current density was high, thus lowering the cathode potential. High methanol concentration increased crossover, while it improved kinetics at the anode. The effect of crossover on the cathode was shown to be $\sim$70 mV. Hence, the use of a membrane that would minimize crossover, along with the use of high catalytic activity at the anode, was identified as an effective means for preventing crossover and increasing cell performance.

The minimum goal for commercialization of DMFC is 250–300 mW/cm^2 for transportation and 30–40 mW/cm^2 for portable applications operating on methanol and air.[48] Siemens obtained 20 mW/cm^2 with pressurized methanol and oxygen, but the single cell was not durable. The University of Newcastle used liquid and vapor feed systems with low loading electrodes containing 2.5 mg/cm^2. A power density of 0.2 W/cm^2 was achieved at 98°C with pressurized oxygen and 2 M methanol solution, while pressurized air gave a power density of 60 mW/cm^2. JPL/Giner measured a cell voltage of 0.47 V at 400 mA/cm^2 in oxygen, while the cell voltage was 0.38 V in air. Electrodes with low Pt loading of 0.5 mg/cm^2 gave 150 mW/cm^2 at 95°C. Using a thinner Nafion 112 membrane, LANL developed DMFCs that had a cell voltage of 0.57 V at 400 mW/cm^2. High temperature and pressure were used to enhance electrode kinetics to counter methanol crossover.

Kho et al.[45] developed miniature DMFCs that operate under passive feed conditions. The effect of methanol concentration, membrane thickness, and amount of catalyst loading on the performance of air-breathing DMFCs was determined. Maximum power density was obtained for methanol concentration in the 4–5 M range. This was much higher than optimum methanol concentration in the active feed regime. At $>$6 M methanol, there was a sharp drop in performance. Higher concentration is needed to increase the transport rate of methanol. Thinner membranes gave better performance than thicker membranes, thus showing that proton conductance is more important than methanol crossover for the investigated range of membrane thickness. The cell performance increased with catalyst loading up to 6 mg/cm^2. This indicated that excess catalyst could result in providing a mass-transport barrier to methanol and oxygen. Under active methanol feed conditions, 2 M methanol concentration gave better results than 4 M because of lower crossover. The fuel cell was monopolar with six cells with an active area of 4.5 cm^2 per cell, with the electrodes of same polarity on the same surface of the polymer membrane. Other companies that have developed miniature DMFCs are Poylfuel, Samsung Advanced Institute of Technology, Casio, MTI, Toshiba, and NEC.

CONCLUSIONS

Typically, power source design has lagged behind the requirements of the portable electronic goods industry. As the demand for portable electronic devices increases and as those devices increase their power and energy

requirements, there will be a need for advanced power supplies with large power and energy density. There is significant developmental effort in the area of microscale fuel cells. Development has been focused primarily on hydrogen and more recently on direct methanol fuel cells. For low power applications, DMFCs appear to be very attractive, if methanol crossover through the membrane is reduced or cathode catalyst tolerance to methanol oxidation is increased. If a cost-effective and efficient hydrogen storage design can be obtained, PEM fuel cells show great promise. Many researchers are developing designs that enable the use of high-volume fabrication similar to that of the semiconductor industry in the hope of developing an inexpensive device. Various stack designs are available for the optimization of power or energy density.

REFERENCES

1. Hoogers, G. Introduction. In *Fuel Cell Technology Handbook*; Hoogers, G., Ed.; CRC Press: Boca Raton, 2003; 1-1–1-5.
2. Hoogers, G. Portable applications. In *Fuel Cell Technology Handbook*; Hoogers, G., Ed.; CRC Press: Boca Raton, 2003; 9-1–9-14.
3. Lee, S.-J.; Chang-Chien, A.; Cha, S.W.; O'Hayre, R.; Park, Y.I.; Saito, Y.; Prinz, F.B. Design and fabrication of a micro fuel cell array with "flip-flop'' interconnection. J. Power Sources **2002**, *112*, 410–418.
4. Srinivasan, S.; Davé, B.B.; Murugesamoorthi, K.A.; Parthasarathy, A.; Appleby, A.J. Overview of fuel cell technology. In *Fuel Cell Systems*; Blomen, L.J.M.J., Mugerwa, M.N., Eds.; Plenum Press: New York, 1993; 37–72.
5. Zawodzinski, T.A.; Springer, T.E.; Davey, J.; Jestel, R.; Lopez, C.; Valerio, J.; Gottesfeld, S. A comparative study of water uptake by and transport through ionomeric fuel cell membranes. J. Electrochem. Soc. **1993**, *140*, 1981–1985.
6. Meyers, J.P.; Maynard, H.L. Design considerations for miniaturized PEM fuel cells. J. Power Sources **2002**, *109*, 76–88.
7. Blum, A.; Duvdevani, T.; Philosoph, M.; Rudoy, N.; Peled, E. Water-neutral micro direct-methanol fuel cell (DMFC) for portable applications. J. Power Sources **2003**, *117*, 22–25.
8. Peled, E.; Duvdevani, T.; Aharon, A.; Melman, A.A. Direct methanol fuel cell based on a novel low-cost nanoporous proton-conducting membrane. Electrochem. Solid-State Lett. **2000**, *3*, 525–528.
9. Hoogers, G. The fueling problem: fuel cell systems. In *Fuel Cell Technology Handbook*; Hoogers, G., Ed.; CRC Press: Boca Raton, 2003; 5-1–5-23.
10. Verdrugge, M.W. Methanol diffusion in perfluorinated ion-conducting membranes. J. Electrochem. Soc. **1989**, *136*, 417–423.
11. Kauranen, P.S.; Skou, E. Methanol permeability in perfluorosulfonate proton exchange membranes at elevated temperatures. J. Appl. Electrochem. **1996**, *26*, 909–915.
12. Narayanan, S.R.; Halpert, G.; Chun, W.; Jeffries-Nakamura, B.; Valdez, T.I.; Frank, H.; Surampudi, S. The status of direct methanol fuel cell technology at JPL. The Eleventh Annual Battery Conference on Applications and Advances, Long Beach, California, Seo, E.T., Ed.; IEEE, 1996; 96–99.
13. Heinzel, A.; Hebling, C.; Muller, M.; Zedda, M.; Muller, C. Fuel cells for low power applications. Proceedings of the Seventh Ulmer Elektrochemische Tage, 26–27 June 2000; Elsevier Science B.V.: Ulm, 2002; Vol. 105, 250–255.
14. Kosek, J.; Hamdan, M.; LaConti, A.B. *Reducing Methanol Crossover in CH_3OH-Fuel Cell Membranes*; NASA Tech Briefs, Glenn Research Center: Cleveland, OH, 2000; Vol. LEW-16669, 1–11.
15. Potje-Kamloth, K.; Josowicz, M.; Vielstich, W. Polymer coated oxygen cathode for methanol fuel cell application. In *Fall Meeting of the Electrochemical Society*; Electrochemical Society: Toronto, 1992.
16. Savinell, R.E.; Yeager, E.; Tryk, D.; Landau, U.; Wainwright, J.; Wong, D.; Lux, K.; Litt, M.; Rogers, C. A polymer electrolyte for operation at temperatures up to 200°C. J. Electrochem. Soc. **1994**, *141*, L46–L48.
17. Kjaer, J.; Yde-Anderson, S. Solid state direct methanol fuel cells. Proceedings of the 26th Intersociety Energy Conversion Engineering Conference, Atlanta, GA, 1991; 542.
18. Ma, Z.Q.; Cheng, P.; Zhao, T.S. A palladium-alloy deposited Nafion membrane for direct methanol fuel cells. J. Membr. Sci. **2003**, *215*, 327–336.
19. Holladay, J.D.; Wainright, J.S.; Jones, E.O.; Gano, S.R. Microscale power generation using a mesoscale fuel cell integrated with a microscale fuel processor. J. Power Sources **2004**, *130*, 111–118.
20. Song, C.S. Fuel processing for low-temperature and high-temperature fuel cells—challenges, and opportunities for sustainable development in the 21st century. Catal. Today **2002**, *77*, 17–49.
21. Hahn, R.; Krumm, M.; Reichl, H. Thermal management of portable micro fuel cell stacks. Proceedings of the 19th Annual IEEE Semiconductor Thermal Measurement And Management Symposium, San Jose, CA, 11–13 March 2003; Institute of Electrical and Electronics Engineers Inc., 2003; 202–209.
22. Maynard, H.L.; Meyers, J.P. Miniature fuel cells for portable power: design considerations and

challenges. Proceedings of the 29th Conference on the Physics and Chemistry of Semiconductor Interfaces, Santa Fe, NM, 6–10 January 2002; American Institute of Physics Inc., 2002; Vol. 20, 1287–1297.
23. Hoogers, G. Fuel cell components and their impact on performance. In *Fuel Cell Technology Handbook*; Hoogers, G., Ed.; CRC Press: Boca Raton, 2003; 4-1–4-26.
24. Holladay, J.D.; Humble, P.; Harb, J.; LaFollette, R.; Salmon, L.; Barksdale, R.; Anderson, B. Ni/Zn and Li-ion rechargeable microbatteries for use with MEMS devices. The 195th Meeting of the Electrochemical Society Inc., Seattle, WA, May 1998, ECS, 1998.
25. Humble, P.H.; Harb, J.N. Optimization of nickel–zinc microbatteries for hybrid powered microsensor systems. J. Electrochem. Soc. **2003**, *150*, A1182–A1187.
26. Bates, J.B.; Dudney, N.J.; Neudecker, B.; Ueda, A.; Evans, C.D. Thin-film lithium and lithium-ion batteries. Solid State Ionics **2000**, *134*, 33–45.
27. Yu, J.; Cheng, P.; Ma, Z.; Yi, B. Fabrication of miniature silicon wafer fuel cells with improved performance. J. Power Sources **2003**, *124*, 40–46.
28. Yu, J.; Cheng, P.; Ma, Z.; Yi, B. Fabrication of a miniature twin-fuel-cell on silicon wafer. Electrochim. Acta **2003**, *48*, 1537–1541.
29. Wainright, J.S.; Savinell, R.F.; Liu, C.C.; Litt, M. Microfabricated fuel cells. Electrochim. Acta **2003**, *48*, 2869–2877.
30. Kelley, S.C.; Deluga, G.A.; Smyrl, W.H. Miniature fuel cells fabricated on silicon substrates. J. AIChE **2002**, *48*, 1071–1082.
31. Morse, J.D.; Jankowski, A.F.; Graff, R.T.; Hayes, J.P. Novel proton exchange membrane thin-film fuel cell for microscale energy conversion. Proceedings of the 46th National Symposium of the American Vacuum Society, 25 October–29 October 1999; J. Vac. Sci. Technol. A: Vacuum, Surfaces Films **2000**, *18*, 2003–2005.
32. Johnson, W.L.; Phillips, C.B.; Chen, Z.; Ransom, T.S.; Thompson, L.T. Novel, heat integrated, low temperature co-fired, ceramic (LTCC) methanol reformer for hydrogen generation. Proceedings of the Microreaction Technology and Process Intensification Symposium at the 226th American Chemical Society National Meeting, New York, 7–11 September 2003; ACS, Abstract number 41.
33. Thompson, L.T. Novel fuel processing catalysts and reactors. In *Microreaction Technology and Process Intensification*, Symposium at the 226th American Chemical Society National Meeting, New York, 7–11 September 2003; ACS, Abstract Number 35.
34. Ehrfeld, W.; Ehrfeld, U. Micro fabrication for process intensification. Proceedings of the Fifth International Conference on Microreaction Technology, Strausburg, France, 27–30 May 2001; Matlosz, M., Ehrfeld, W., Baselt, J.P., Eds.; Springer-Verlag: New York, 2001; 3–12.
35. Tonkovich, A.Y.; Zilka, J.L.; LaMont, M.J.; Wang, Y.; Wegeng, R.S. Microchannel reactors for fuel processing applications. I. Water gas shift reactor. Chem. Eng. Sci. **1999**, *54*, 2947–2951.
36. Ameel, T.A.; Papautsky, I.; Warrington, R.O.; Wegeng, R.S.; Drost, M.K. Miniaturization technologies for advanced energy conversion and transfer systems. J. Propulsion Power **2000**, *16*, 577–582.
37. Xia, Y.N.; Whitesides, G.M. Soft lithography. Angew. Chem. Int. Ed. **1998**, *37*, 550–575.
38. Heinzel, A.; Nolte, R.; Ledjeff-Hey, K.; Zedda, M. Membrane fuel cells—concepts and system design. Electrochim. Acta **1998**, *43*, 3817–3820.
39. Ledjeff, K. Polymer Electrolyte Membrane Fuel Cell U.S. Patent 5,863,672, 1999.
40. Calabrese Barton, S.; Fuller, T.F.; West, A.; Electrode design for a strip-cell direct methanol fuel cell. ECS Meeting Abstracts, 1998; ECS, MA 98-2, 1089.
41. Dyer, C.K. Replacing the battery in portable electronics. Sci. Am. **1998**, *281*, 88
42. Narayanan, S.R.; Valdez, T.I.; Kindler, A.; Witham, C. Advances in direct methanol fuel cells for mobile and portable applications. Proceedings of the 2002 Fuel Cell Seminar, Palm Springs, 2002; Binder, M., Ed.; 2002; 1000–1003.
43. Valdez, T.I.; Narayan, S.R.; Frank, H.; Chun, W. Direct methanol fuel cell for portable applications. Proceedings of the 12th Annual Battery Conference on Applications and Advances, Long Beach, CA, 1997; Frank, H.A., Seo, E.T., Eds.; IEEE, 1997.
44. Argyropoulos, P.; Scott, K.; Taama, W.M. Dynamic response of the direct methanol fuel cell under variable load conditions. J. Power Sources **2000**, *87*, 153–161.
45. Kho, B.K.; Cho, E.-A.; Oh, I.-H.; Hong, S.-A.; Ha, H.Y. The effects of operating conditions on the performance of air-breathing direct methanol fuel cells. Proceedings of the 2002 Fuel Cell Seminar, Palm Springs, CA., 2002; Binder, M., Ed.; 263–265.
46. Heinzel, A.; Barragán, V.M. A review of the state-of-the-art of the methanol crossover in direct methanol fuel cells. J. Power Sources **1999**, *84*, 70–74.
47. Scott, K.; Taama, W.M.; Argyropoulos, P.; Sundmacher, K. The impact of mass transport and methanol crossover on the direct methanol fuel cell. J. Power Sources **1999**, *83*, 204–216.
48. Hogarth, M. Prospects of the direct methanol fuel cell. In *Fuel Cell Technology Handbook*; Hoogers, G., Ed.; CRC Press: Boca Raton, 2003; 7–9.

Microscopy of Ionomers

Andreas Taubert
Department of Chemistry, University of Basel, Basel, Switzerland

Brian P. Kirkmeyer
International Flavors and Fragrances, Union Beach, New Jersey, U.S.A.

INTRODUCTION

DuPont's Surlyns, which are semicrystalline poly(ethylene-*ran*-methacrylic acid) (EMAA) copolymers neutralized with various cations, were developed in the 1960s. They have ever since been used as coatings and adhesives, in food packaging, and other fields. The Surlyns are the first example of a then-new class of polymeric materials, so-called ionomers, that is, uncharged polymers with a small fraction of ionic groups. These can be neutralized with a variety of counterions, and the presence of these neutralized acid groups significantly influences the physical properties of ionomers. For example, the impact strength and melt viscosity of ionomers depend not only on the type of neutralizing cation, but also on the neutralization level. This is advantageous because by simply changing a cation or the neutralization level of an ionomer, one can tune the properties for a desired application. It is now generally accepted that the neutralized acid groups and the cations form larger structures, so-called ionic or ion-rich aggregates, within a predominantly hydrophobic matrix. These ionic structures act as physical cross-links between polymer chains and largely control the ionomer properties, particularly if there is no crystallization of the main polymer chain.

Despite the relative simplicity of most ionomers, questions about them remain. One of the key questions is how structure and dynamics on different length scales connect. Specifically, how does the metal coordination to the neutralized acid groups (which is on an angstrom level) correlate with the size, shape, and distribution of the ion-rich aggregates in the hydrophobic matrix (which is on a nanometer length scale)? Furthermore, how does the microscopic structure control the macroscopic properties like melt viscosity or elastic modulus?

Because the micro- and mesoscopic behavior of an ionomer governs its macroscopic properties, understanding these phenomena is the key to new and improved ionomers. Much research has thus been directed at investigating ionomers at different length scales. We will here present approaches to study ionomers on the intermediate mesoscopic (i.e., nanometer) length scale and show how one can obtain information about processing effects on ionomer morphology. The focus of this entry is on ionomer microscopy, which has recently been established as a complementary technique to, for example, x-ray scattering or IR spectroscopy for the investigation of ionomers. For a comprehensive treatment of ionomers, in particular their micro- and macroscopic properties and their chemistry, the reader should refer to the current literature.

IONOMER CHEMISTRY

We will briefly introduce some important ionomers (see Fig. 1), but for a thorough treatment of ionomer chemistry, see, e.g., Refs.[1,2]. The simplest classification of ionomers is semicrystalline vs. amorphous ionomers. The prototypical semicrystalline ionomer is EMAA (Surlyn, DuPont) neutralized with various cations. Also from DuPont, Nafion is a perfluorinated polyethylene with sulfonic acid or sulfonate groups on short side chains. Other commercial ionomers like Aciplex (Asahi Chemical Company), Flemion (Asahi Glass Company), and Neosepta (Tokuyama) are structurally similar to Nafion. For a recent review on Nafion see Mauritz and Moore.[3]

Arguably the most important amorphous ionomer is sulfonated polystyrene (SPS). Other ionomers include poly(styrene-*ran*-methacrylic acid) (SMAA), polyurethanes, siloxanes, butadiene-based elastomers, ethylene-propylene-diene terpolymers, acrylates and methacrylates, polyphosphoesters, polyimides, and many others.[1,2]

The most important ionic groups are carboxylates and sulfonates, but other groups like ammonium have also been reported.[4] Special classes of ionomers are zwitterionic ionomers, which contain both cationic and anionic groups, as well as ionomers with regular architectures like block ionomers, mono- or telechelic ionomers, star ionomers, or ionenes, which have regularly spaced ionic groups.[1,2]

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007825

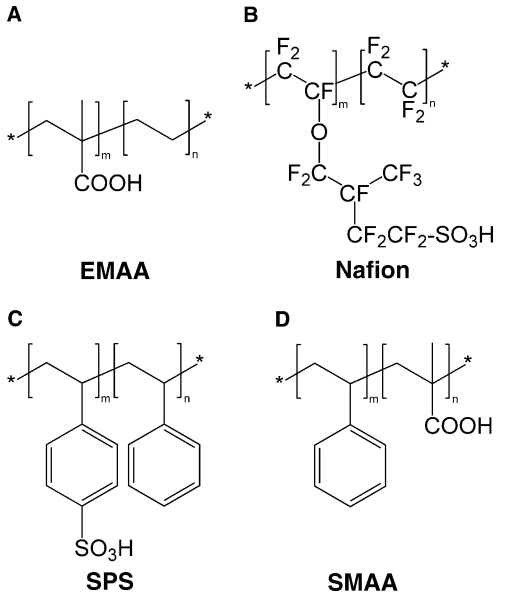


Fig. 1 Important ionomers: (A) poly(ethylene-*co*-methacrylic acid), the free acid form of Surlyn, (B) Nafion, (C) SPS, and (D) poly(styrene-*co*-methacrylic acid). For the sake of simplicity, all polymers are shown in the acid form. In the ionomers, the acids are partly or fully neutralized with cations such as Na^+, NH_4^+, or Zn^{2+}.

LOCAL ENVIRONMENT OF THE NEUTRALIZING ION

As stated above, the ionic groups act as physical cross-links between polymer chains; cross-linking is thus a phenomenon that occurs on an angstrom length scale. Ionomer structure and processing effects at this level have mainly been probed with solid-state nuclear magnetic resonance (SSNMR), Fourier transform infrared spectroscopy (FTIR), and extended x-ray absorption fine structure spectroscopy (EXAFS). Most of these experiments yield information about the local structure around the neutralizing cation (generally a metal). For example, ^{23}Na-SSNMR experiments by O'Connell et al. found a narrow peak at 7 ppm and a broad peak at ca. 20 ppm in the spectra of Na-SPS.[5] The authors interpret these peaks as being due to isolated and aggregated Na^+, respectively, where the breadth of the latter signal is caused by an overlap of the signals from dimers, trimers, etc. coexisting in the aggregates. Similar observations have recently been made by Moore and coworkers in Na-neutralized poly(butylene terephthalate) ionomers.[6]

Solid-state NMR also allows for the study of sample preparation effects on the local cation environment. Na-SPS with 1.7% styrene sulfonate groups cast from dimethyl formamide or tetrahydrofuran (THF)/water mixtures shows more aggregated ions, i.e., a more intense peak at ~20 ppm, than ionomers cast from less polar THF or THF/EtOH mixtures.[7] Humidification and subsequent drying also result in more Na^+ in the aggregates than before; however, annealing Na-SPS above the PS glass transition temperature leads to more ions in the polymer matrix. This process is partly reversible and can be controlled by adjusting the processing conditions.[8] In contrast to Na-SPS, ^{23}Na-SSNMR spectra of Na-EMAA only exhibit one broad signal at ~12 ppm, which is independent of the thermal history of the sample.[9] This implies that processing affects Na-SPS much stronger than Na-EMAA.

Extended x-ray absorption fine structure spectroscopy shows that in Zn-EMAA the oxygen atoms around Zn^{2+} form well-defined oxygen shells.[10] In Na-EMAA the environment around Na^+ is much less ordered, in particular at high neutralization levels.[11] This is in principle supported by the results of Kuwabara and Horii,[5] because the broad peak in ^{23}Na-SSNMR spectra of Na-EMAA can be attributed to dimers, trimers, etc. in the aggregates, similar to the suggestion by O'Connell et al. for Na-SPS.[9] Extended x-ray absorption fine structure spectroscopy has also shown that ionic aggregates may contain a significant amount of nonionic material (see, e.g., a new study by Grady and coworkers).[12]

Of particular interest with regard to processing is a study investigating the local cation environment of Zn-EMAA at high pressure.[13] The coordination geometry around Zn^{2+} depends on the applied pressure P: under vacuum (at $P = 0$), IR detects bands at 1624 and 1538 cm^{-1} that were ascribed to hexacoordinated Zn-carboxylates. Increasing P intensifies a band at 1585 cm^{-1}, which was assigned to a tetracoordinated Zn-carboxylate. A remarkable feature of this pressure-induced coordination change is that even atmospheric pressure significantly changes the coordination. Furthermore, if the pressure is applied below the melting temperature of the crystalline polyethylene backbone, the coordination change depends on the temperature. These results clearly show that temperature and pressure will have a drastic impact on the ionic structure on an angstrom length scale, which will in turn influence the macroscopic properties of the material.

X-RAY SCATTERING OF IONOMERS

Rather than probing very short-ranged features and interactions, many authors describe the structure and processing effects in ionomers as a function of the entire ionic aggregate, i.e., on a mesoscopic (nanometer length) scale. The morphology of the ionic aggregates has mainly been investigated with small-angle x-ray scattering (SAXS). Typical ionomer SAXS patterns show a broad scattering peak in the region from ca. 0.5 to 4 nm^{-1} and a low-angle upturn near the

beam stop. Different models have been proposed as to how the scattering peak shape and position relate to the size, shape, and spatial distribution of the ionic aggregates.[14–19] The low-angle upturn was explained as the result of parasitic scattering, remaining neutralizing agent, or other impurities.[17,20,21] However, already in 1987, Galambos et al. have suggested that the upturn is due to the presence of metal ions in the neutralized acid groups.[22] Eisenberg et al. have also suggested an inhomogenous distribution of so-called multiplets, i.e., very small, ion-containing entities, as the source for the upturn.[19]

Anomalous SAXS and ultra-SAXS further support the claim that both the scattering peak and the low-angle upturn are due to the neutralizing cations in the ionomers and not to impurities or artifacts.[23–27] Wu et al., Ding et al., and Li et al. suggest various kinds of long-range heterogeneities, for example, a nonrandom distribution of unaggregated ionic groups in the matrix, as the source of the low-angle upturn.[25–27] They also suggest sample preparation by compression molding, nonequilibrium structures arising from short annealing time, or compositional variations in the SPS as possible sources of the upturn. Register and Cooper assign the upturn to the microphase-separated structure of ion-rich aggregates, amorphous matrix, and crystalline matrix, where applicable.[28,29]

This chapter clearly shows that although scattering and spectroscopy are powerful methods, they have limitations in terms of the selection and applicability of the models used for data interpretation. For example, many of the suggested causes for the low-angle upturn in ionomer SAXS patterns are plausible, but they cannot be verified independently. Microscopy, however, can in certain cases provide a model-free assessment of ionomer morphology, which in turn will confirm or rule out the applicability of a certain model. In the remainder of the text we will thus focus on microscopy of ionomers.

MICROSCOPY OF IONOMERS

Imaging methods can yield valuable information about ionomer morphology, because imaging provides a

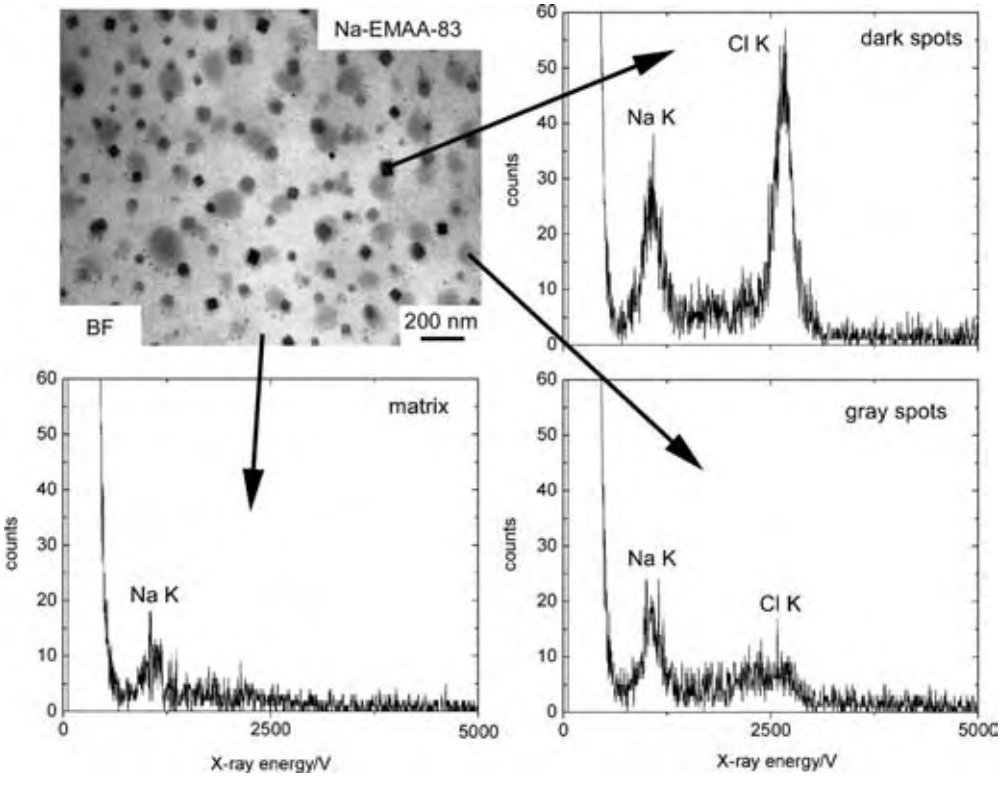


Fig. 2 Bright field STEM image of Na-EMAA neutralized to 83% and exposed to a warm and humid environment for about 5 days. The image shows a light background (the presumed matrix), gray spots, and dark rectangular features. By placing the electron beam on either one of these features it is possible to acquire energy dispersive x-ray spectra, which carry information about the chemical composition of the respective feature. The x-ray spectra acquired on a light gray, a dark gray, and a black feature show that the presumed matrix does contain some Na. However, this does have to be treated with care as stray electrons may also cause signals from adjacent areas. In any case, the dark gray and, in particular, also the black features do contain Na and Cl. These results thus indicate that aging of ionomers can include the formation of metal chloride nanoparticles on exposure to a certain environment. Unfortunately, the mechanism of the precipitation is not clear to date.

model-independent determination of the presence, size, size distribution, shape, shape variation, and spatial distribution of the ionic aggregates. Early transmission electron microscopy (TEM) experiments have, however, been inconclusive: stained solvent-cast ionomer films produce artifacts in the TEM but no reliable data.[30] Stained microtomed samples do not show features that can be assigned to either an ionic aggregate or a region without aggregates (matrix). It was thus concluded that either the ionic aggregates must be smaller than 2–3 nm in diameter, because otherwise they should be visible in the TEM, or that they are much more diffuse than expected. This would lead to a rather nondescript contrast variation in the images.

The same paper also reports the formation of MCl (M = Na, K, Cs) nanocrystals upon exposure of M-EMAA ionomers to chloroform. Interestingly, using scanning TEM (STEM) and energy dispersive x-ray spectroscopy, Winey and coworkers have observed the same phenomenon on thin sections of Na-EMAA exposed to the humid conditions of a Philadelphia summer for a few days (see Fig. 2). These results clearly show that aging effects in ionomers can be observed and identified on a nanometer scale with the appropriate electron microscopy technique.

Porat et al. have used sulfur mapping and zero loss imaging to investigate Nafion.[31] They report S-rich (and hence $-SO_3^-$ rich) domains of about 5 nm in diameter. Some of the S-rich domains appear to have a ring-like shape. This is intriguing because recently, Winey and coworkers observed similar, so-called vesicular, aggregates in Zn-SPS.[32] The shape observed by Porat et al. could thus be interpreted as evidence for vesicular aggregates in Nafion.[31] Electron diffraction also shows that Nafion crystallizes in a structure very similar to polyethylene, which is unlike the crystal structure of polytetrafluoroethylene, the Nafion base polymer. Unfortunately, this interesting publication has never been followed up.

The authors are only aware of one other study where conventional TEM was used to image ion-rich domains in ionomers. Huang et al. observe features interpreted as ion-rich domains in negatively stained poly(dimethyl siloxane) ionomers with pendant quaternary ammonium groups neutralized with chloride anions (Fig. 3).[4] Here, one does, however, have to

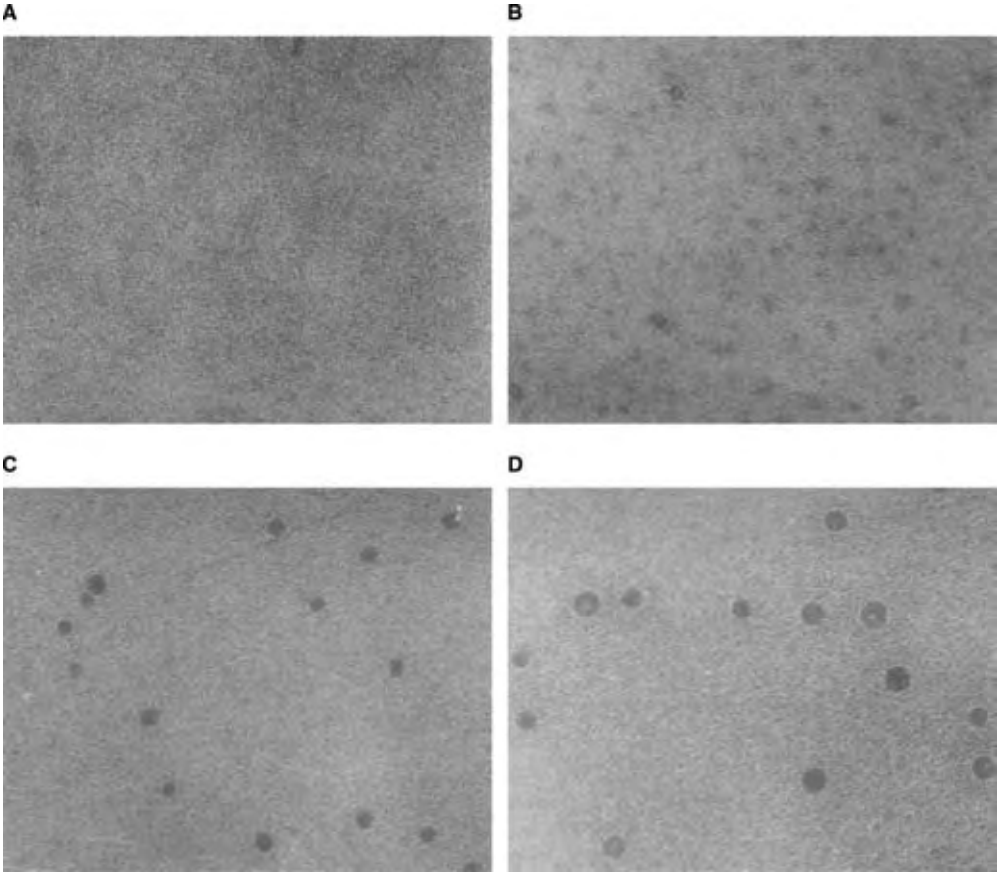


Fig. 3 Transmission electron micrographs of poly(dimethyl siloxane) ionomers with different contents of the ionic γ-(benzyl dimethylamino)propylmethylsiloxy unit: (A) 1.3 mol%, (B) 2.3 mol%, (C) 4.1 mol%, and (D) 10.0 mol%. Some aggregates may have a vesicular structure. (Reprinted with permission from Ref.[4]. Copyright 2004 Wiley–VCH.)

ask how staining affects the morphology of the polymer; also the authors do not elaborate about the shape of the claimed ionic aggregates. This is unfortunate because some of them appear to be vesicular rather than spherical. They thus represent an interesting deviation from models of ionic aggregation, possibly similar to the TEM results by Porat et al. or more recent studies by Winey and coworkers (see below).[31]

Using STEM rather than conventional TEM, Winey and coworkers have successfully imaged ionic aggregates in various ionomers. Scanning TEM has the advantage over TEM that here contrast scales approximately with the atomic number squared, Z^2.[33] This means that STEM provides a powerful tool for imaging high-Z elements like Zn in a low-Z matrix like C. This is exactly what is required for the investigation of ionomer morphology if one is interested in the size, shape, and distribution of ion-rich domains in ionomers.

Scanning TEM shows that semicrystalline Zn-EMAA neutralized between 17% and 78% contains nearly monodisperse spherical aggregates randomly distributed in the polymer matrix (Fig. 4).[34,35] The aggregates in both as-extruded and recrystallized Zn-EMAA are spherical. The diameter of the aggregates is ca. 2 nm, independent of the neutralization level; it increases slightly upon recrystallization. This is more pronounced at higher neutralization levels (see Fig. 5). Finally, at 17% neutralization the ionic aggregates are widely separated relative to the polymer chains, which is not the case at higher neutralization levels. Similar to Grady et al.[12] the authors suggest that the ionic aggregates include segments from the nonionic polymer backbone and that the aggregate morphology is primarily the result of ionic interactions in the system.[34,35]

These results thus support the Yarusso-Cooper model, where it was suggested that in any ionomer,

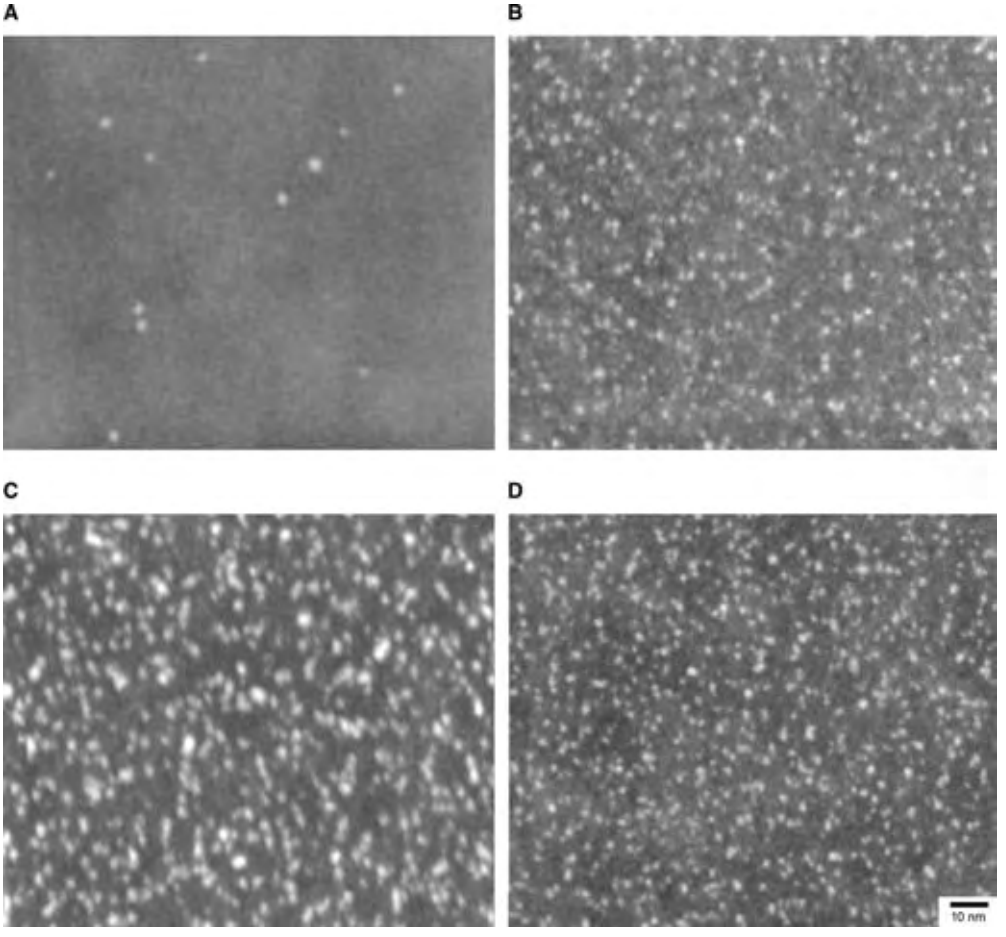


Fig. 4 Annular dark field STEM images of as-extruded EMAA neutralized to various extents with Zn^{2+}: (A) 17%, (B) 29%, (C) 55%, and (D) 78%. Here, bright regions correspond to higher average atomic number and hence to Zn-rich ionic aggregates. Nearly spherical ionic aggregates of ca. 2.1 nm are present at all neutralization levels, though the areal number density of aggregates increases substantially between 17% and 29% neutralization. (Reprinted with permission from Ref.[34]. Copyright 2004 American Chemical Society.)

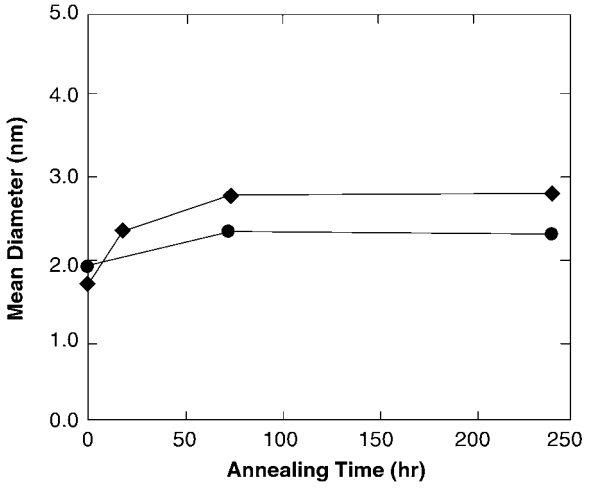


Fig. 5 Effect of thermal history on aggregate size in Zn-EMAA: the mean diameter for Zn-EMAA neutralized to 29% (circles) and 78% (diamonds) with Zn^{2+} increases slightly with annealing time. (Reprinted with permission from Ref.[34]. Copyright 2004 American Chemical Society.)

the ion-rich domains are only a few nanometers in diameter and are randomly dispersed in a hydrophobic matrix.[17] It is furthermore similar to the Eisenberg–Hird–Moore model, where spherical aggregates of ionic groups were postulated.[19] However, a significant deviation from the models was found at low neutralization levels, where only a few aggregates were observed.

It came as a major surprise when STEM images of amorphous Zn-SPS revealed a macrophase-separated morphology at the micrometer size scale (Fig. 6). In lightly neutralized ionomers, only some regions contain detectable aggregates.[32,36] Furthermore, a new aggregate morphology was observed: vesicular aggregates coexist with the expected spherical domains (Fig. 7). The spherical aggregates have diameters from 4 to 10 nm, whereas vesicular aggregates have diameters from 9 to 55 nm and an aggregate wall thickness of about 3 nm. Lightly neutralized ionomers exhibit only spherical aggregates, whereas higher neutralization levels exhibit both aggregate types. These ionic aggregates thus contradict previous interpretations of SAXS data with respect to size, size dispersity, shape, and spatial distribution, but they clearly demonstrate the need for a technique that is able to assess ionomer morphologies without the need to infer structures from a model. These results further raise the question whether the features observed by Porat et al. in Nafion are not actually vesicular aggregates.[31] If indeed they are, one may have to reassess existing Nafion morphological models.

Similar to Zn-SPS, Cs-neutralized poly(styrene-*ran*-methacrylic acid) (Cs-SMAA) ionomers exhibit Cs-rich vesicular aggregates that are randomly distributed in a

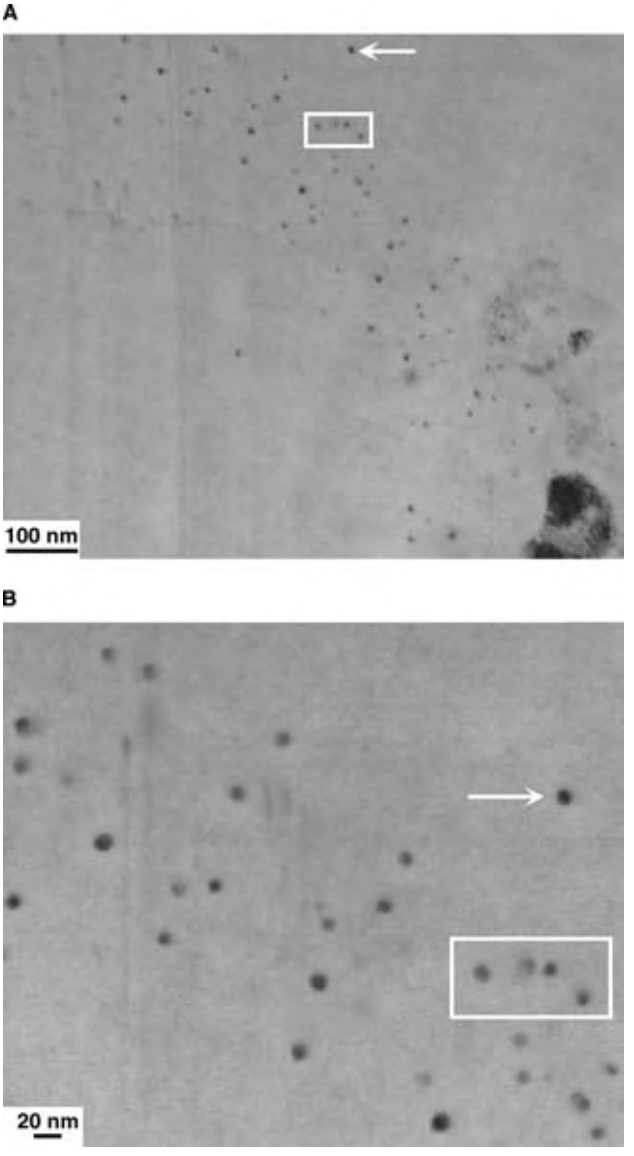


Fig. 6 Bright field STEM images showing the shape homogeneity and spatial heterogeneity in Zn-SPS neutralized to 25%. (A) Macrophase separation of ionic aggregates. Aggregates are only observed in a ca. 200 nm band across the field of view. (B) The Zn-rich aggregates indicated by the box and arrow in (A) are shown at higher magnification. Only uniform and spherical aggregates exist at 25% neutralization. (Reprinted with permission from Ref.[32]. Copyright 2004 Wiley–VCH.)

polystyrene-rich matrix.[37] Here, the aggregates are typically 5–20 nm in diameter and have a shell thickness of ca. 3 nm. The vesicular aggregates are invariant to annealing and their dimensions and number density remain unchanged. The aggregate diameters and shell thickness are independent of the copolymer concentration and the rate of neutralizing agent addition during solution neutralization. Like Zn-SPS, Cs-SMAA is thus quite resistant to various treatments.

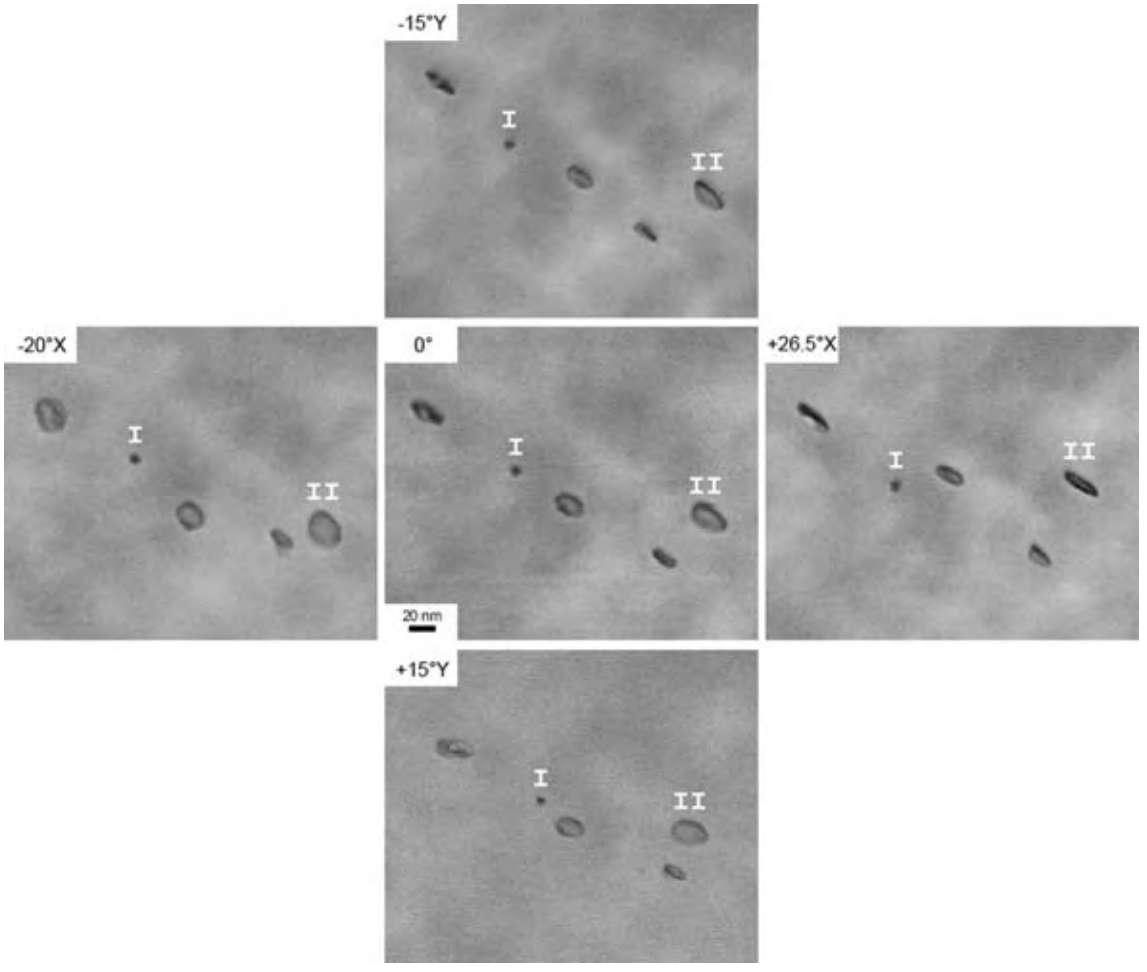


Fig. 7 Bright field STEM images showing the shape heterogeneity in SPS neutralized to 100% with Zn. The double-tilt series shows multiple projections of the Zn-rich aggregates to determine the aggregate shape. The horizontal and vertical sets of images were collected after rotating the sample about the *x* and *y* axes, respectively, by the amounts indicated in the figure relative to the untilted central image. The index I refers to a spherical aggregate; II refers to vesicular aggregates with a wall thickness of about 3 nm. (Reprinted with permission from Ref.[32]. Copyright 2004 Wiley–VCH.)

Also similar to Zn-SPS, Na-EMAA does exhibit macrophase separation after annealing at 60°C (Fig. 8).[38] Recrystallized Na-EMAA has a macrophase-separated structure where three phases with distinct boundaries coexist: no aggregates are observed in the so-called phase I, phase II shows rather small spherical aggregates with diameters of ca. 2–15 nm, and phase III exhibits large, approximately spherical aggregates with diameters from ca. 20 to 160 nm and rather diffuse boundaries. In contrast, the original, as-extruded Na-EMAA does not show any features on the STEM length scale. Like the findings discussed above, this study shows that thermal history has a significant impact on ionomer morphology on a nanometer length scale. Also, it is interesting to compare the results from ^{23}Na-SSNMR with the STEM results: NMR suggests that thermal treatments do not have a measurable impact on the aggregation in Na-EMAA, but STEM shows that on a nanometer length scale, processing drastically affects the morphology.[9,38]

Na- and Zn-EMAA blown films exhibit the same characteristics as the bulk material, that is, the blowup process does nothing to the morphology of Zn-EMAA but Na-EMAA does change with blow conditions similar to bulk Na-EMAA (N. M. Benetatos and K. I. Winey, in preparation).

Finally, two Al-neutralized copolyimides from (4,4'-hexafluoroisopropylidene) diphthalic anhydride, (4,4'-hexafluoroisopropylidene) dianiline, and 3,5-diaminobenzoic acid exhibit a variety of aggregate shapes and sizes that have not been previously observed or even postulated.[39] Scanning TEM shows Al-rich aggregates in both copolymers, but the aggregate size and shape distributions in the ionomer with a high ionic fraction are much more heterogenous than in the ionomer with a lower ionic fraction. This is despite

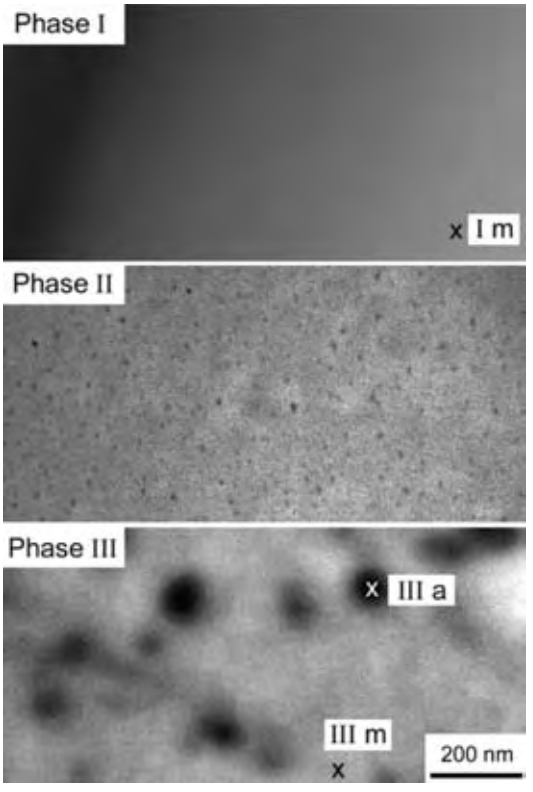


Fig. 8 Bright field STEM images of Na-EMAA showing three distinct phases. Phase I is featureless on the STEM length scale, phase II contains small spherical aggregates, and phase III contains large spherical aggregates. The designations ''a'' and ''m'' denote the locations of energy dispersive x-ray spectra taken of the aggregates and the matrix, respectively (data not shown, see Ref.[38]). (Reprinted with permission from Ref.[38]. Copyright 2004 American Chemical Society.)

the fact that the SAXS patterns of the two samples are identical. The former ionomer contains spherical aggregates ranging from ca. 5 to 25 nm in diameter, chain-like features, vesicular aggregates, and long, stripe-like aggregates with a length of up to 1000 nm. The latter ionomer contains only a few roughly spherical aggregates. Large fractions of both ionomers do not exhibit visible aggregates on the STEM length scale; this is again similar to several of the polymers described above.

At this point, it has to be noted that the analysis of STEM data is not straightforward. Similar to AFM, where the images are a convolution of the tip shape and size and the actual feature size and shape, the shape and size of the beam in the STEM will lead to convoluted images. Kirkmeyer et al.[40] have recently shown that STEM image deconvolution provides additional useful information about the ionic aggregates (Fig. 9). The deconvoluted images are sharper than the original and exhibit decreased noise levels. This simplifies detection of the aggregates and measurement of their size and shape in comparison with the original image. Both raw and deconvoluted images of spherical ionic aggregates in Zn-SMAA suggest that the electron density varies smoothly from the edge to the center of the aggregates. Line scans across single aggregates in deconvoluted STEM images suggest that the three-dimensional density distribution of these aggregates resembles a radially symmetric Gaussian distribution and not a uniformly dense sphere. This issue is currently addressed by N. M. Benetatos and K. I. Winey (personal communication).

What is not known, however, is whether image deconvolution can eliminate beam spreading as a contributor to the line scan shape. One would expect that beam spreading from a dense array of sources like the aggregates would obscure features beyond the resolution capabilities of the microscope. In this instance, deconvolution would provide only a modicum of aid in image interpretation. The authors of this entry however believe that beam spreading is a minimal contributor to the image.

Another significant issue is how STEM-observed ionic aggregates and SAXS-proposed multiplets are related, if at all. One possibility is that the aggregates and multiplets are actually the same features, but the measurement techniques skew the data. An example of this is related to the line scans detailed above. If the aggregates/multiplets are in fact Gaussian spheres, then a SAXS hard sphere data fit would interpret the features as smaller than they are, while a STEM direct measurement at the periphery of the features would interpret them as larger. Another possibility is a two-tiered morphology, each one accessible by only one technique. The 2–3 nm features observed in STEM would scatter near the beamstop in SAXS, while the <1 nm features from the SAXS data fits would not generate enough contrast to be observed in STEM. In this instance, the aggregates could be responsible for the low-angle upturn in SAXS. However, to the authors' knowledge, this issue is to date unresolved.

Besides (S)TEM, atomic force microscopy (AFM) has also been used to image ionomers. Chomakova-Haefke et al. report that dry Nafion has a disordered network structure, which reorganizes into parallel fibrils upon swelling.[41] No substantial changes in the fibril dimensions are found in the swollen state. This was interpreted as an indication that swelling occurs at the supramolecular level but this interpretation was not elaborated further. Also in Nafion, Lehmani et al. found spherical grains of ca. 11 nm in diameter surrounded by ca. 5 nm thick regions.[42] The authors speculate that the 5 nm features may correspond to some type of ionic aggregate, but both studies have remained inconclusive as to what the features really represent.

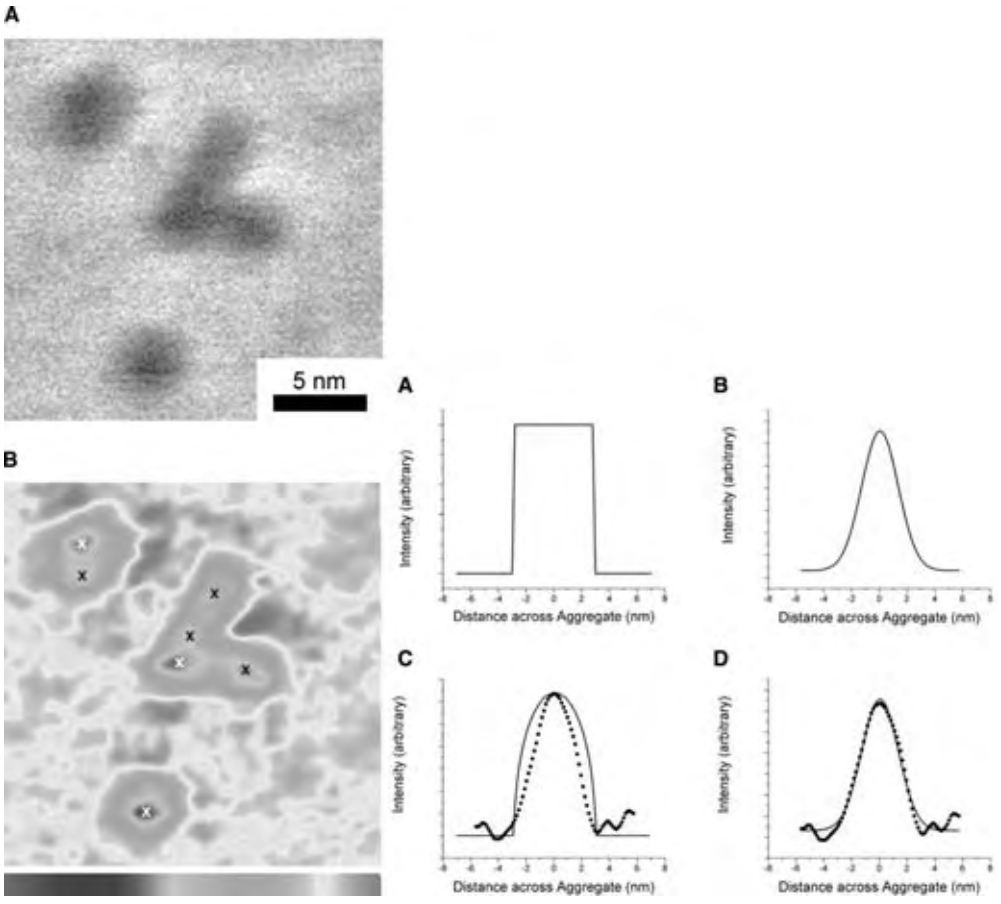


Fig. 9 Left: High-magnification images of the Zn-SMAA ionomer morphology: (A) grayscale raw STEM image and (B) grayscale image model. The images show both the decrease in the noise level and the multiple brightness maxima in various ionic aggregates, indicating an overlap of ionic aggregates in the projection. Each ''x'' indicates a brightness maximum. Right: (A) Line scan through the center of a three-dimensional solid sphere, (B) line scan through the center of a three-dimensional Gaussian sphere, (C) intensity profile across a projected ionic aggregate in Zn-SMAA (∗) and the best-fit curve of the intensity profile across a projected three-dimensional solid sphere (–), and (D) intensity profile across a projected ionic aggregate in Zn-SMAA (∗) and the best-fit curve of the intensity profile across a projected three-dimensional Gaussian sphere (–). (Reprinted with permission from Ref.[40]. Copyright 2004 Wiley–VCH.) *(View this art in color at www.dekker.com.)*

Furthermore, using tapping mode AFM, Hill et al. have found features called ''bundles of micelles'' in Nafion thin films on silicon.[43] The films show no dewetting even above the glass transition temperature, but high-ionic-strength solutions like 20 wt% $NaHSO_4$ destroy the aggregates and lead to dewetting.

Scanning tunnelling microscopy, scanning electrochemical microscopy, and AFM-surface potential measurements have also been used to investigate Nafion films. Scanning electrochemical microscopy reveals a domain-like structure containing circular features ca. 1-2 nm in diameter made up of a conductive center (presumed to be ion-rich regions) surrounded by a much less conductive zone.[44] Atomic force microscopy surface potential measurements detect features that were interpreted as ion channels in Nafion membranes. The size of the claimed ion channels was 40–100 nm and several tens of the presumed channels were observed in a 10 × 10 μm region.[45]

James et al. have studied Nafion membranes in different states of hydration using an AFM with an environmental chamber to control the humidity (Fig. 10).[46,47] Tapping mode was used to identify the hydrophobic and hydrophilic regions of Nafion. As a main finding the authors report that the number of the supposed ionic domains decreases while the average size increases with increasing humidity. Similarly, Umeda and coworkers have shown that as-prepared Nafion films have an irregular surface morphology with features of ca. 10–30 μm. Upon exposure to methanol vapor the surface became relatively flat.[49,50] These studies provide an important piece of information for the fuel cell community, because the results are obtained from ''active'' material and not, like in

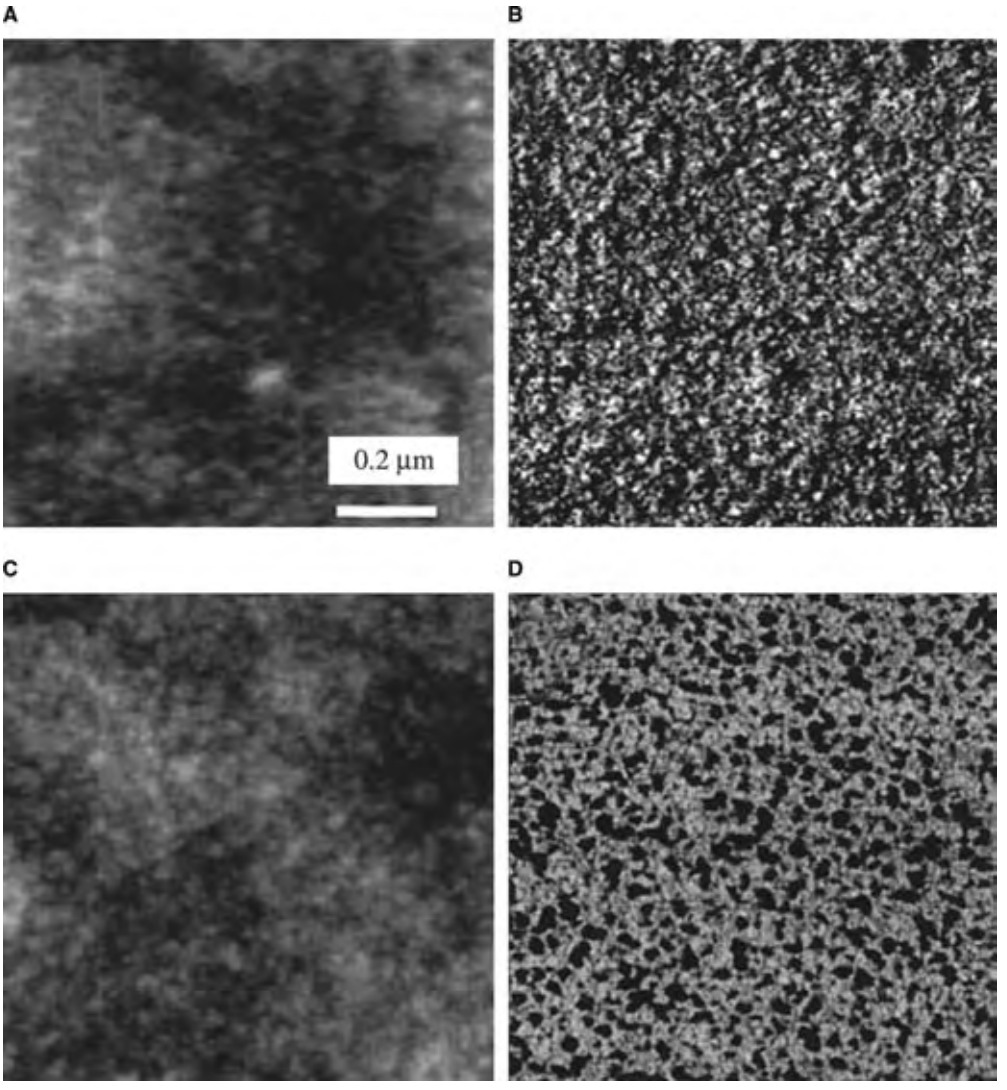


Fig. 10 Tapping mode AFM images of Nafion 115. (A) One-micrometer topography image of Nafion 115 free acid (protonated polymer) imaged under ambient conditions. (B) The corresponding phase image. (C) One-micrometer image of Nafion 115 neutralized with Cs imaged under ambient conditions. (D) The corresponding phase image. The images exhibit a marked difference in the phase contrast between the two ion forms: the phase range is significantly larger in the Cs^+ ion, 60° as opposed to 10° for the H^+ form.[48] There is no significant difference in the topography images; therefore topographic coupling is not responsible for the change in phase contrast. These images thus demonstrate that phase imaging can indeed yield some information about aggregation in ionomers. (Reprinted with permission from Ref.[48]. Copyright 2004 American Chemical Society.)

(S)TEM, from dry, thin sections of an ionomer. They also demonstrate, however, that ionomer physics is far from understood. For example, it is not clear yet how the different features and dimensions observed in AFM connect to one another or to the results obtained from other studies.

Much like James et al., Sauer and coworkers use tapping mode AFM technique to detect ionic aggregates in Zn-EMAA and Nafion.[46,47,51,52] Zn-EMAA exhibits crystalline polyethylene lamellae and noncrystalline regions. Operating the AFM under special low oscillation amplitude gave images that were interpreted by the authors as showing ion-rich regions and individual ionic domains. The diameter of the claimed individual domains was ca. 2 nm, similar to results obtained from STEM.[34] Further experiments showed that the lamellar morphology and perfection are controlled by chain mobility in the melt, which depends on acid level, neutralization level, and counterion. Nafion contains individual fluorocarbon crystallites separated by a matrix containing features interpreted as the ionic domains. While no particularly surprising structures were thus revealed in these studies of either Zn-EMAA or Nafion, it would be interesting to investigate the effects of thermal treatment on Na-EMAA or polymers where unexpected morphologies have

been observed in STEM to determine whether or not the same features are observed with the AFM.

Using scanning electron microscopy (SEM) with x-ray microanalysis rather than (S)TEM or AFM, Schulze et al. and Tierney and Register have investigated the elemental distribution of Nafion and EMAA ionomers, respectively, on a micrometer length scale.[53,54] Schulze et al. showed that the Na counterion in Nafion/polytetrafluoroethylene (PTFE) membranes in fuel cells are closely associated with the sulfonate anion and that PTFE does not take up Na during the solution neutralization of Nafion.[53] Tierney and Register have measured ion-hopping times by allowing a finite slab of one ionomer to diffuse into a matrix of a second ionomer. The distribution of metal cations was measured via x-ray line profiling in the SEM and it was shown that in Mg/Ca, Mg/Li, Na/Ca, and Na/Li diffusion couples the diffusion coefficient follows inversely with melt viscosity.[54]

Before concluding it must be clearly stated that when using a microscope it is of utmost importance to realize its limitations. One of the major issuess in AFM and STEM (at the high resolutions and magnifications required for ionomer work) is drift: if a feature is only a few nanometers in diameter, great care has to be taken to ensure that the sample does not move during the measurement. Even drift of 1 or 2 nm can dramatically affect the outcome. Furthermore, James et al. have pointed out that tapping-mode phase imaging is a very useful tool for identifying and mapping regions of different properties regardless of their topographical nature but that the interpretation of the AFM data is not trivial.[48] Both STEM and AFM do of course also have limitations as far as the atomic resolution of the ion-rich aggregates is concerned. They are thus not suited for the elucidation of the atomic arrangements in, e.g., ion-rich domains. Their power lies in the capability to provide an independent view of ionomers, which can validate existing models of ionomer morphology and thus contribute to a better understanding of the physics of ionomers.

CONCLUSIONS

This article shows that microscopy of ionomers is a powerful complementary technique to various spectroscopic and macroscopic techniques like EXAFS and rheology for the investigation of ionomers. If precautions are taken and data are carefully analyzed, microscopy can thus add valuable insights to results from other studies. Because of this, microscopy of ionomers is not only of interest for the scientist trying to understand the physics behind ionomer properties and behavior, but also for the engineer attempting to improve or adapt an existing or new ionomer for a desired application.

ACKNOWLEDGMENTS

The authors thank Prof. K. I. Winey for her support and many discussions during their tenure in the Winey group at the University of Pennsylvania. We also thank R. M. Walters, K. E. Sohn, N. Zhou, and in particular, J. Tam, N. M. Benetatos, Prof. R. J. Composto, and Prof. Y. E. Elabd for help in the laboratory and many useful discussions. Prof. J.-S. Kim is acknowledged for useful suggestions during manuscript preparation. Dr. D. M. Yates is acknowledged for help with the STEM and Prof. D. B. Williams for very insightful discussions on the technology of STEM. The writing of this entry was supported by the Swiss National Science Foundation and International Flavors and Fragrances (IFF). While at UPenn, A. T. was funded by the Petroleum Research Fund of the American Chemical Society (ACS).

REFERENCES

1. Eisenberg, A.; Kim, J.-S. *Introduction to Ionomers*; John Wiley & Sons: New York, 1998; 352.
2. Tant, M.R., Mauritz, K.A., Wilkes, G.A. Eds.; *Ionomers–Synthesis, Structure, Properties, and Applications*; Kluwer Academic Publishers: Dordrecht, 1997.
3. Mauritz, K.A.; Moore, R.B. State of understanding of Nafion. Chem. Rev. **2004**, *104*, 4535.
4. Huang, Z.; Yu, Y.; Huang, Y. Ion aggregation in the polysiloxane ionomers bearing pendant quaternary ammonium groups. J. Appl. Polym. Sci. **2002**, *83*, 3099.
5. O'Connell, E.M.; Root, T.W.; Cooper, S.L. Morphological studies of lightly sulfonated polystyrene using ^{23}Na NMR. 1. Effects of sample composition. Macromolecules **1994**, *27*, 5803.
6. Page, K.A.; Schilling, G.D.; Moore, R.B. Influence of ionic aggregation on the surface energies of crystallites in poly(butylene terephthalate) ionomers. Polymer **2004**, *45*, 8425.
7. O'Connell, E.M.; Root, T.W.; Cooper, S.L. Morphological studies of lightly sulfonated polystyrene using ^{23}Na NMR. 2. Effects of solution casting. Macromolecules **1995**, *28*, 3995.
8. O'Connell, E.M.; Root, T.W.; Cooper, S.L. Morphological studies of lightly sulfonated polystyrene using ^{23}Na NMR. 3. Effects of humidification and annealing. Macromolecules **1995**, *28*, 4000.
9. Kuwabara, K.; Horii, F. Solid-state NMR analyses of the crystalline-noncrystalline structure and its thermal changes for ethylene ionomers. J. Polym. Sci. B Polym. Phys. **2002**, *40*, 1142.
10. Grady, B.P.; Floyd, J.A.; Gennetti, W.B.; Vanhoorne, P.; Register, R.A. X-ray absorption

studies of zinc neutralized ethylene methacrylic acid ionomers. Polymer **1999**, *40*, 283.
11. Farrell, K.V.; Grady, B.P. EXAFS spectroscopy studies of cation local environment in sodium neutralized ethylene copolymer ionomers. Macromolecules **2001**, *34*, 7108.
12. Grady, B.P.; Goossens, J.G.P.; Wouters, M.E.L. Morphology of zinc-neutralized maleated ethylene-propylene copolymer ionomers: structure of ionic aggregates as studied by x-ray absorption spectroscopy. Macromolecules **2004**, *37*, 8585.
13. Kutsumizu, S.; Nakumura, Y.; Yano, S. Pressure-induced coordination-structural change around zinc(II) in zinc(II)-neutralized ethylene-methacrylic acid ionomers. 1. Infrared spectroscopic studies. Macromolecules **2001**, *34*, 3033.
14. Eisenberg, A. Clustering of ions in organic polymers. A theoretical approach. Macromolecules **1970**, *3*, 147.
15. Marx, C.L.; Caulfield, D.L.; Cooper, S.L. Morphology of ionomers. Macromolecules **1973**, *6*, 344.
16. MacKnight, W.J.; Taggart, W.P.; Stein, R.S. A model for the structure of ionomers. J. Polym. Sci. Symp. Ser. **1974**, *45*, 113.
17. Yarusso, D.; Cooper, S.L. Microstructure of ionomers: interpretation of small-angle x-ray scattering data. Macromolecules **1983**, *16*, 1871.
18. Roche, E.J.; Stein, R.S.; Russell, T.P.; MacKnight, W.J. Small-angle x-ray scattering study of ionomer deformation. J. Polym. Sci. Polym. Phys. **1980**, *18*, 1497.
19. Eisenberg, A.; Hird, B.; Moore, R.B. A new multiplet-cluster model for the morphology of random ionomers. Macromolecules **1990**, *23*, 4098.
20. Yarusso, D.; Cooper, S.L. Analysis of SAXS data from ionomer systems. Polymer **1985**, *26*, 371.
21. Williams, C.E.; Russell, T.P.; Jerome, R.; Horrion, J. Ionic aggregation in model ionomers. Macromolecules **1986**, *19*, 2877.
22. Galambos, A.F.; Stockton, W.B.; Koberstein, J.T.; Sen, A.; Weiss, R.A.; Russell, T.P. Observation of cluster formation in an ionomer. Macromolecules **1987**, *20*, 3091.
23. Chu, B.; Wu, D.Q.; Lundberg, R.D.; MacKnight, W.J. Small-angle x-ray scattering (SAXS) studies of sulfonated polystyrene ionomers. 1. Anomalous SAXS. Macromolecules **1993**, *26*, 994.
24. Wang, J.; Li, Y.; Peiffer, D.G.; Chu, B. Small-angle x-ray scattering investigation of temperature influences on microstructures of an ionomer. Macromolecules **1993**, *26*, 2633–2635.
25. Wu, D.Q.; Chu, B.; Lundberg, R.D.; MacKnight, W.J. Small-angle X-ray scattering (SAXS) studies of sulfonated polystyrene ionomers. 2. Correlation function analysis. Macromolecules **1993**, *26*, 1000.
26. Ding, Y.S.; Hubbard, S.R.; Hodgson, K.O.; Register, R.A.; Cooper, S.L. Anomalous small-angle x-ray scattering from a sulfonated polystyrene ionomer. Macromolecules **1988**, *21*, 1698.
27. Li, Y.; Peiffer, D.G.; Chu, B. Long-range inhomogeneities in sulfonated polystyrene ionomers. Macromolecules **1993**, *26*, 4006.
28. Register, R.A.; Cooper, S.L. Anomalous small-angle x-ray scattering from nickel-neutralized ionomers. 1. Amorphous polymer matrices. Macromolecules **1990**, *23*, 310.
29. Register, R.A.; Cooper, S.L. Anomalous small-angle x-ray scattering from nickel-neutralized ionomers. 2. Semicrystalline polymer matrices. Macromolecules **1990**, *23*, 318.
30. Handlin, D.L.; MacKnight, W.J.; Thomas, E.L. Critical evaluation of electron microscopy of ionomers. Macromolecules **1981**, *14*, 795.
31. Porat, Z.; Fryer, J.R.; Huxham, M.; Rubinstein, I. Electron microscopy investigation of the microstructure of Nafion films. J. Phys. Chem. **1995**, *99*, 4667.
32. Kirkmeyer, B.P.; Weiss, R.A.; Winey, K.I. Spherical and vesicular ionic aggregates in Zn-neutralized sulfonated polystyrene ionomers. J. Polym. Sci. B Polym. Phys. **2001**, *39*, 477.
33. Williams, D.B.; Carter, C.E. *Transmission Electron Microscopy—A Textbook for Materials Science*; Plenum Press: New York, London, 1996.
34. Winey, K.I.; Laurer, J.H.; Kirkmeyer, B.P. Ionic aggregates in partially Zn-neutralized poly-(ethylene-ran-methacrylic acid) ionomers: shape, size and size distribution. Macromolecules **2000**, *33*, 507.
35. Laurer, J.H.; Winey, K.I. Direct imaging of ionic aggregates in Zn-neutralized poly(ethylene-ran-methacrylic acid) copolymers. Macromolecules **1998**, *31*, 9106.
36. Kirkmeyer, B.P.; Winey, K.I.; Weiss, R.A. Imaging ionic aggregates in Zn-neutralized sulfonated polystyrene ionomers: shape and spatial heterogeneity. Microsc. Microanal. **2000**, *6*, 1112.
37. Kirkmeyer, B.P.; Taubert, A.; Kim, J.-S.; Winey, K.I. Vesicular ionic aggregates in poly(styrene-ran-methacrylic acid) ionomers neutralized with Cs. Macromolecules **2002**, *35*, 2648.
38. Taubert, A.; Winey, K.I. Imaging and x-ray microanalysis of a poly(ethylene-ran-methacrylic acid) ionomer melt-neutralized with sodium. Macromolecules **2002**, *35*, 7419.
39. Taubert, A.; Wind, J.D.; Paul, D.A.; Koros, W.J.; Winey, K.I. Novel polyimide ionomers: CO_2 plasticization, morphology, and ion distribution. Polymer **2003**, *44*, 1881.
40. Kirkmeyer, B.P.; Winey, K.I.; Puetter, R.C.; Yahil, A. Deconvolution of scanning transmission

electron microscopy images of ionomers. J. Polym. Sci. B Polym. Phys. **2004**, *41*, 319.
41. Chomakova-Haefke, M.; Nyffenegger, R.; Schmidt, E. Structure reorganization in polymer films of Nafion due to swelling studied by scanning force microscopy. Appl. Phys. A **1994**, *59*, 151.
42. Lehmani, A.; Durand-Vidal, S.; Turq, P. Surface morphology of Nafion 117 membrane by tapping mode atomic force microscopy. J. Appl. Polym. Sci. **1998**, *68*, 503.
43. Hill, T.A.; Carroll, D.L.; Czerw, R.; Martin, C.W.; Perahia, D. Atomic force microscopy studies on the dewetting of perfluorinated ionomer thin films. J. Polym. Sci. B Polym. Phys. **2002**, *41*, 149.
44. Fan, F.-R.F.; Bard, A.J.; Guckenberger, R.; Heim, M. STM on wet insulators: electrochemistry or tunneling? Science **1995**, *270*, 1849.
45. Kanamura, K.; Morikawa, H.; Umegaki, T. Observation of interface between Pt electrode and Nafion membrane. J. Electrochem. Soc. **2003**, *150*, A193.
46. James, P.J.; Elliott, J.A.; McMaster, T.J.; Newton, J.M.; Elliott, A.M.S.; Hanna, S.; Miles, M.J. Hydration of Nafion studied by AFM and x-ray scattering. J. Mater. Sci. **2000**, *35*, 5111.
47. James, P.J.; McMaster, T.J.; Newton, J.M.; Miles, M.J. In situ rehydration of perfluorosulfonated ion-exchange membrane studies by AFM. Polymer **2000**, *41*, 4223–4231.
48. James, P.J.; Antognozzi, M.; Tamayo, T.J.; Newton, J.M.; Miles, M.J. Interpretation of contrast in tapping mode AFM and shear force microscopy. A study of Nafion. Langmuir **2001**, *17*, 349.
49. Affoune, A.M.; Yamada, A.; Umeda, M. Surface observation of solvent-impregnated Nafion membrane with atomic force microscopy. Langmuir **2004**, *20*, 6965.
50. Umeda, M.; Ojima, H.; Mohamedi, M.; Uchida, I. Methanol vapor-induced morphology and current-voltage characteristic changes in a cast-coated Nafion film/interdigitated microarray electrode. J. Polym. Sci. B Polym. Phys. **2002**, *40*, 1103.
51. McLean, R.S.; Doyle, M.; Sauer, B.B. High resolution imaging of ionic domains and crystal morphology in ionomers using AFM techniques. Macromolecules **2000**, *33*, 6541.
52. Sauer, B.B.; McLean, R.S. AFM and x-ray studies of crystal and ionic domain morphology in poly(ethylene-co-methacrylic acid) ionomers. Macromolecules **2000**, *33*, 7939.
53. Schulze, M.; v.Bradke, M.; Reissner, R.; Lorenz, M.; Gülzow, E. Characterization of polymers in PEFC electrodes with EDX and XPS. Fresenius J. Anal. Chem. **1999**, *365*, 123.
54. Tierney, N.K.; Register, R.A. Ion hopping in ethylene-methacrylic acid ionomer melts as probed by rheometry and cation diffusion measurements. Macromolecules **2002**, *35*, 2358.

Microwave Processing of Ceramics

Eldon D. Case
Chemical Engineering and Materials Science Department, Michigan State University, East Lansing, Michigan, U.S.A.

INTRODUCTION

Microwave processing of ceramics may be categorized in many ways, but two helpful schemes are: i) high-temperature processing and ii) low- or intermediate-temperature processing. For ceramics, high-temperature processing includes sintering and joining of ceramics.

Of course, microwaves are used for a variety of low-temperature processing applications. For materials other than ceramics, low-temperature applications of microwave heating include drying wood, preparing food, and synthesizing polymeric materials. For ceramics, low-temperature microwave applications include drying, binder burnout, and the synthesis of oxide and nonoxide ceramic powders.[1] In addition, microwave heating can be used to fabricate ceramic coatings or perform rapid thermal annealing over a range of temperatures.

In general, ceramic microstructure (grain size and porosity) is of central importance in processing, as many of the mechanical, thermal, electronic, and optical properties of ceramics are closely linked with their microstructure.[2] Microwave heating allows one to engineer microstructures and/or optimize processing times in ways that are not available via conventional radiant energy processing. This applies not only to microwave sintering and joining but also to microwave-assisted powder synthesis, drying, and fabrication of coatings. The apparent advantages of microwave processing stem from the volumetric heating and the inverse temperature profiles available from microwave heating. In addition, microwave heating (especially for high-temperature processing) can result in energy savings compared with conventional radiant heating.

MICROWAVE HEATING OF CERAMICS

In the 1950s, engineers working at Raytheon discovered that microwave radiation could be used to heat materials.[3] By the 1960s, scientists and engineers were performing the first microwave sintering studies on ceramic powders. While a number of the earlier experiments were done with 2.45 GHz radiation with magnetron tube power supplies, high power, high frequency gyrotron microwave sources were developed as part of the nuclear weapons research programs in a number of countries.[3] These gyrotron millimeter wave sources have subsequently been used to sinter and to join a variety of ceramic materials. Since the 1990s, the use of microwave heating to dry ceramics, synthesize ceramic powders, and in the fabrication of ceramic coatings has increased steadily.

POWER DISSIPATED BY THE MICROWAVE FIELD

Parameters Governing Microwave Power Absorption

The mechanisms by which microwave energy is absorbed by the entire processing system (the microwave applicator, the specimens, and the thermal insulation) are crucial to understand the microwave processing of ceramic materials. The microwave energy absorbed per unit time per unit volume, p_{abs} (the powder density), can be related to dielectric properties of the material and to the applied electromagnetic field:[4,5]

$$p_{abs} = 2\pi f \varepsilon_r' \tan\delta \langle E^2 \rangle = 2\pi f \varepsilon_r'' \langle E^2 \rangle \quad (1a)$$

where f is the frequency of the incident microwave field, ε_r' the relative dielectric constant of the material, $\tan\delta$ the loss tangent of the dielectric material, $\langle E^2 \rangle$ the time-average of the square of the local electric field, E, ε_r'' the dielectric loss constant ($= \varepsilon_r' \tan\delta$).

The dissipation of microwave power by ceramics is a function of the material properties ε_r' and ε_r'', which are the real (storage) and imaginary (loss) parts of the dielectric constant, respectively. In addition to the frequency dependence of p_{abs} [Eq. (1a)], ε_r' and ε_r'' are themselves functions of the microwave frequency, the ambient temperature, T, as well as material properties, including a number of microstructural and chemical variables. Thus, to emphasize the range of parameters that impact the local power dissipation, we define the following symbols: M denotes the microstructural variables, especially volume fraction porosity (VFP), as well as the size, shape, and the distribution of size

Encyclopedia of Chemical Processing DOI: 10.1081/-E-ECHP-120030627

and shape of pores; C_D denotes defect chemistry, chemistry of atomic-scale defects, especially in terms of concentration and chemical species, the effects of point defects such as vacancies and interstitial atoms.

Using the position vector $\vec{r}$ to indicate the spatial dependence of the material properties, the dependence of ε_r', ε_r'', and tan δ on microwave frequency and temperature as well as on the material's microstructure and chemistry can be denoted symbolically as $\varepsilon_r' \{f, T, M(\vec{r}), C_D(\vec{r})\}$, $\varepsilon_r'' \{f, T, M(\vec{r}), C_D(\vec{r})\}$, tan $\delta\{f, T, M(\vec{r}), C_D(\vec{r})\}$, respectively. To further emphasize the dependence of p_{abs} on a potpourri of physical variables, Eq. (1a) can be rewritten as

$$p_{abs} = 2\pi f \varepsilon_r'\{f, T, M(\vec{r}), C_D(\vec{r})\} \tan \delta\{f, T, M(\vec{r}), \times C_D(\vec{r})\}\langle E^2 \rangle \quad (1b)$$

Examples of the dependence of ε_r', ε_r'', and tan δ on temperature, frequency, and impurity level are given in Figs. 1–3. For a microwave frequency of about 3.5–4.0 GHz, the temperature dependence of ε_r' for three aluminas of differing purity (the purities are 99.5%, 99%, and 90–95% alumina for the topmost, middle, and bottom curves, respectively) is shown in Fig. 1.[6] Fig. 2 shows the dramatic temperature dependence of tan δ for the same three aluminas, with the onset of the rapid rise in tan δ occurring at increasingly lower temperatures as the concentration of impurities increases. Fig. 3 demonstrates the sensitivity of ε_r'' to changes in frequency for magnesium oxide at seven different temperatures from 25°C to 1200°C.[6]

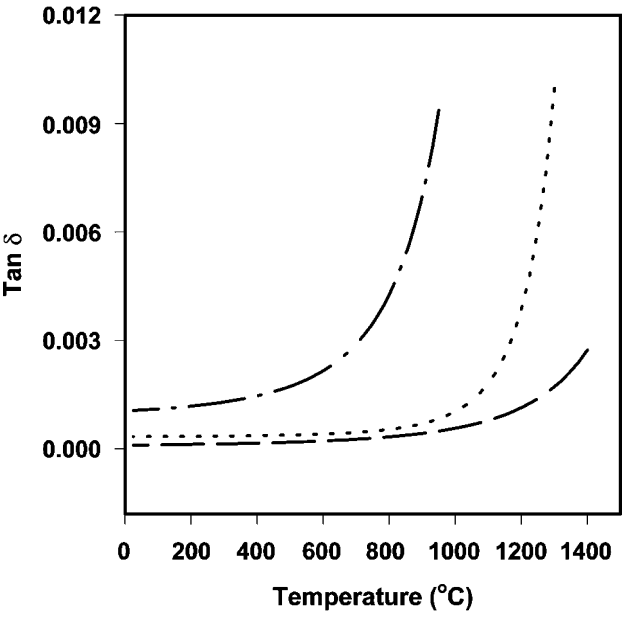


Fig. 2 The temperature dependence of tan δ for polycrystalline alumina with the following purities: 99.5% (bottom curve), 99% (middle curve), and 90–95% (topmost curve). (From Ref.[6].)

As can be seen from Eq. (1b), the magnitude of ε_r'' or tan δ indicates how easily a material absorbs the microwave energy that is subsequently converted into heat energy. At room temperature, most ceramics are "low loss" and thus not heated readily by

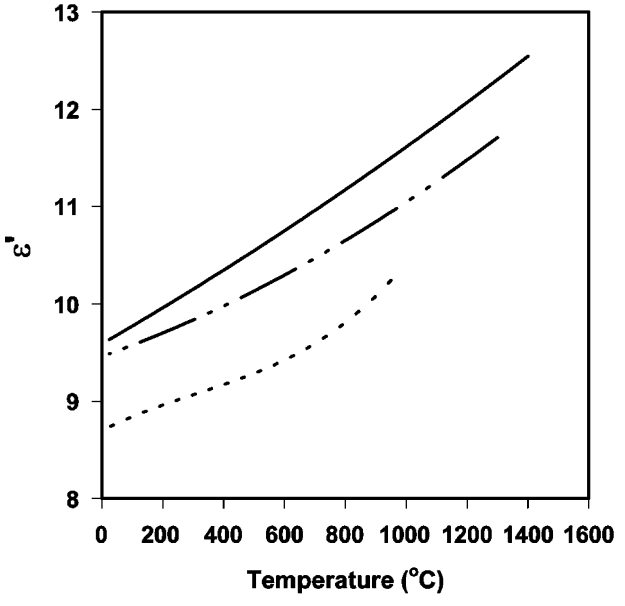


Fig. 1 The temperature dependence of ε_r' for polycrystalline alumina with the following purities: 99.5% (topmost curve), 99% (middle curve), and 90–95% (bottom curve). (From Ref.[6].)

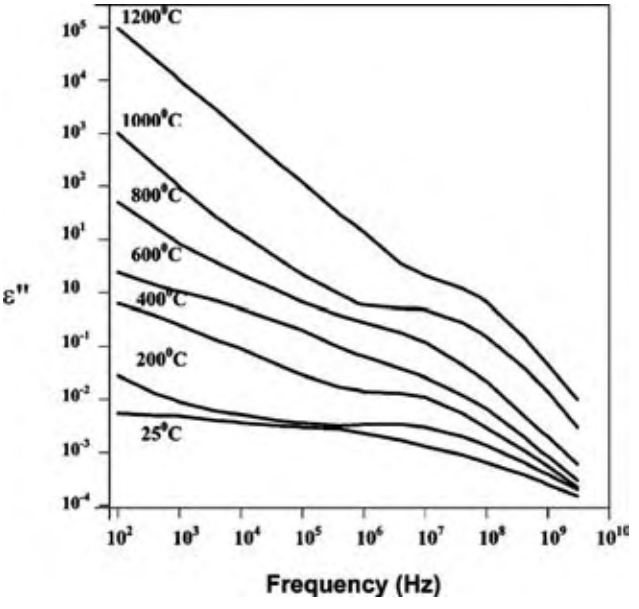


Fig. 3 The frequency and temperature dependence of ε_r'' for polycrystalline magnesium oxide. (From Ref.[6].)

microwaves, with tan δ values of roughly 10^{-3} to 10^{-4} for microwave frequencies. In contrast, ceramic susceptor materials (such as SiC) are lossy dielectrics at room temperature with tan δ values of about 10^{-1} to 1 at microwave frequencies. However, for most ceramics, tan δ and ε_r'' rise rapidly for temperatures above roughly 600–1000°C; hence, most ceramics readily couple with microwaves at elevated temperature.[1,6]

For high-temperature microwave processing of ceramics, the use of refractory specimen enclosures (caskets) is a natural consequence of the temperature dependence of ε_r'' or tan δ. Refractory caskets typically: 1) are fabricated from susceptor materials or 2) include susceptor materials that are placed inside the refractory casket, near the dielectric specimens. At low temperatures, the susceptor materials within the casket absorb microwave energy, heating the casket. The heated casket then radiatively heats the low loss ceramic specimens contained within it. However, as the processing temperature increases, the specimen's dielectric loss also increases (as in Fig. 2) until the specimen also couples directly with the microwave energy. This combination of radiative and direct microwave heating is termed "hybrid heating" and is often used in sintering, joining, and etching of ceramics.

In addition to the conversion of microwave power into heat, the dielectric constants also determine the manner in which microwave radiation penetrates a ceramic. Low loss ceramics are relatively transparent to microwave radiation, while electrically conductive materials reflect microwave radiation (Fig. 4).

Partitioning of Power Within the Cavity

Several power dissipation mechanisms are involved in heating a dielectric using microwave radiation. For an extended body such as a specimen or the refractory casket, p_{abs} is the local microwave power absorption. P_C and P_S give more global measure of the power dissipation in a dielectric, the spatial averages of the power density in the dielectric body. For the casket, p_{absc} is integrated over the casket volume, V_C, to give[4,5]

$$P_C = \int_{V_C} p_{absc}\, dV \tag{2a}$$

Likewise, the power dissipated in the specimen, P_S, is given by p_{abss} integrated over the specimen volume, V_S, to give[4,5]

$$P_S = \int_{V_S} p_{abss}\, dV \tag{2b}$$

The total microwave power, P_T, absorbed by the entire processing system is[4,5]

$$P_T = P_I - P_R = P_W + P_C + P_S \tag{3}$$

where P_I is the microwave input power, P_R the reflected microwave power, P_W the power dissipated by the cavity wall, P_C the power absorbed by the casket, and P_S the power absorbed by the specimen.

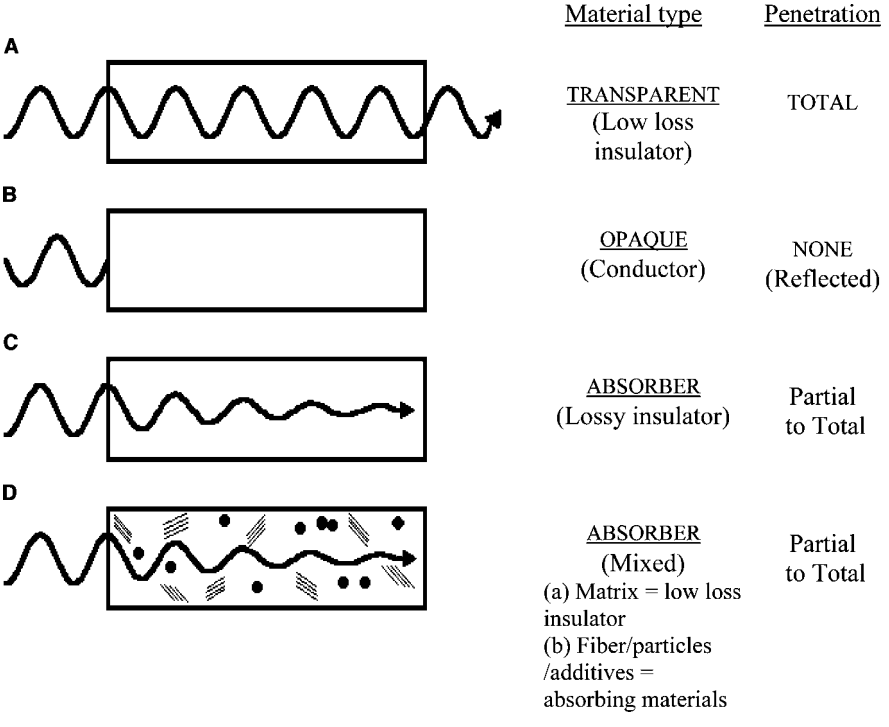


Fig. 4 Schematic diagram showing the nature of reflection and absorption from (A) low loss dielectric, (B) metallic, (C) monolithic high loss dielectric, and (D) a composite or particulate ceramic with high dielectric loss. (From Sutton, W.H. Microwave processing of ceramic materials. Am. Ceram. Soc. Bull. **1989**, *68*, 376–386, Fig. 2, page 377. Reprinted with permission of The American Ceramic Society, www. ceramics.org. Copyright 1989. All rights reserved.)

In a microwave cavity loaded with a lossy dielectric material (such as a ceramic), the power is absorbed mainly by the casket (specimen enclosure) and the specimen. The losses in the cavity wall can typically be ignored[7] such that Eq. (3) can be written as

$$P_T = P_I - P_R = P_C + P_S \quad (4)$$

The discussion above has several implications. During high-temperature heating, the power is dissipated mainly by the refractory casket and the specimens. Also, refractory specimen enclosures (caskets) can allow the high-temperature microwave heating of ceramics that have low losses at room temperature. In addition, for microwave heating of caskets and specimens alike, the dependence of the power dissipation upon material properties [Eq. (1b)] complicates the modeling of the microwave heating process.

Analytical and Numerical Modeling of Microwave Heating

The temperature fields induced by microwave heating can be modeled via the simultaneous solution of Maxwell's equations (for the electromagnetic component of the problem) and the heat equation. The modeling is very challenging, in part because the dielectric constants ε_r' and ε_r'' are functions of temperature and the microwave frequency as well as of the microstructural and chemical details of the ceramic [Eq. (1b)]. Also, the thermal conductivity, k (which is needed for the heat transfer calculations), typically is a function of temperature as well as of microstructural variables such as porosity.[2]

A consequence of the complex interplay of the dielectric and thermal properties with the imposed microwave field is that both Maxwell's equations and the Fourier heat equation are mathematically nonlinear (i.e., they are in general nonlinear partial differential equations). Although analytical solutions have been proposed under particular assumptions,[8] most often microwave heating is modeled numerically via methods such as finite difference time domain (FDTD) techniques.[9] Both the analytical and the numerical solutions presume that the numerical values of the dielectric constants and the thermal conductivity are known over the temperature, microstructural, and chemical composition range of interest, but it is rare in practice to have such complete databases on the pertinent material properties.

As an example of the property changes induced by microwave processing, in the sintering process a porous ceramic powder compact evolves toward a dense body during heat treatment at elevated temperatures. During densification, changes in ε_r', ε_r'', and k can be large in amplitude and complex in nature. The lack of data on ε_r', ε_r'', and k can then result in large discrepancies between the predicted and the actual temperature fields during microwave processing. Moreover, even if the temperature field within the ceramic body is fully specified, it still can be extremely difficult to map the temperature–time values into the microstructural evolution (densification and/or grain growth, for example) experienced by a particular ceramic powder compact.

MICROWAVE CASKETS (SUSCEPTORS AND THERMAL INSULATORS)

For high-temperature microwave ceramic processing, the refractory specimen enclosure (the casket) serves as both i) a thermal insulator and ii) a microwave susceptor. The susceptor allows the microwave energy to couple with the material at high temperature. The thermal insulation diminishes the dissipation of thermal energy. On occasion, the caskets are referred to as "susceptors," which does not fully depict the dual role of the specimen enclosure or "casket" as both a thermal insulator and a microwave susceptor (absorber).

In general, there are two basic strategies in casket construction: one is to incorporate separate thermally insulating and susceptor materials and the other is to include materials that can serve both functions. The casket may consist of a number of substructures, reflecting its dual role. The outer part of the casket often is either a prismatic box or a cylinder composed of ceramic insulating material. Within the outer insulating shell of the casket are the specimens and a "setter" material that supports the specimens during densification or joining. Some casket designs position microwave susceptor materials and/or additional thermal insulation between the outer shell and the specimens (Fig. 5). Prior to discussing particular casket configurations, it is of interest to consider particular examples of insulating and susceptor materials used in microwave processing.

The susceptor materials used in high-temperature processing include zirconia, boron nitride, graphite, carbon black, sodium-beta alumina, zinc oxide, and silicon carbide. While each of these susceptor materials has relatively high dielectric losses at room temperature, silicon carbide is also refractory with a relatively good resistance to oxidation at temperatures up to roughly 1500°C.[10]

Ceramics used for thermal insulation of microwave caskets include alumina, aluminosilicates, mullite, and fused silica, each of which is relatively "transparent" to microwave radiation and thus have low dielectric loss at room temperature. The insulating materials included in the casket are typically in the form of ceramic fibers, fiberboard, or a granular bed of ceramic

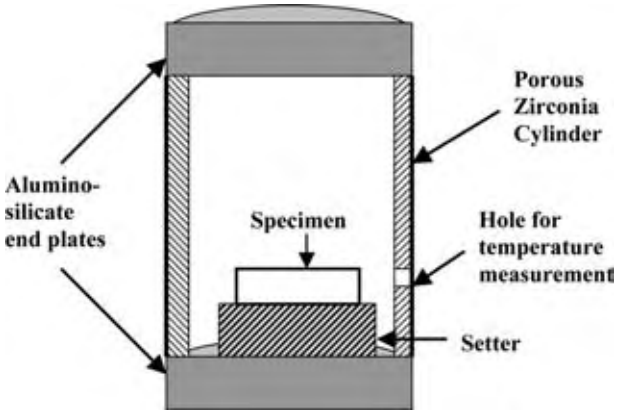


Fig. 5 A cross-sectional schematic view of the type of casket shown in Fig. 6. The casket is sectioned along a plane coincident with the casket's cylinder axis and shows the specimen setter and the specimen.

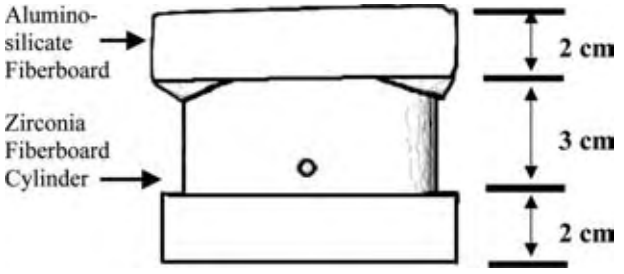


Fig. 6 A schematic drawing of microwave casket (refractory specimen enclosure) composed of a hollow zirconia fiberboard material with disk-shaped aluminosilicate end caps. *(View this art in color at www.dekker.com.)*

particles where the material's high porosity not only boosts the material's insulating capabilities but also tends to inhibit thermal shock damage in the insulator. (Rapid heating and cooling induces thermal gradients that can in turn generate mechanical stress and cracking. Such damage is referred as thermal shock damage.) In particular, ceramic fiberboard material offers some structural stability compared with powders or entirely fibrous materials. Fiberboard also is commercially available in sheets or hollow cylinders that are easily cut to the needed dimensions for caskets. For example, zirconia and/or aluminosilicate fiberboard has been used widely as microwave casket materials during joining and sintering operations at temperatures as high as 1600–1650°C.[4]

Although zirconia and aluminosilicate fiberboard insulation materials are quite versatile, hot spots tend to form in them during microwave processing, with the hot spots becoming increasingly prevalent in the temperature range of 1500–1650°C.[4,11] Such local hotspots are a manifestation of the "thermal runaway" phenomenon that may occur in both specimens and casket materials. As a material's temperature is raised, its dielectric loss, ε_r'', typically increases. Increasing ε_r'' then induces an increase in the local power absorption per unit volume (p_{abs}), which leads to further increases in the material's local temperature. Hot spots can result from inhomogeneities in either the material's dielectric loss and/or the electromagnetic field distribution and can lead to local melting.

In Fig. 6, the refractory casket is composed of a hollow, porous fiberboard zirconia cylinder. The zirconia cylinder is capped on both ends with porous endplates cut from a flat aluminosilicate fiberboard. In this case, the zirconia cylinder serves as both a susceptor and a thermal insulator. The processing temperature is measured via a circular hole in the cylindrical wall of the casket, which allows one to site an optical pyrometer on the specimen during heating (Figs. 5 and 6).

As an alternative, the casket may include separate insulator and susceptor materials. For example, in some caskets, a circular array of solid cylindrical SiC susceptor elements is embedded in porous ceramic insulation surrounding the specimen.[10] Instead of an array of rods, susceptors consisting of open tubes or crucibles also have been used.[10] Extremely high microwave sintering temperatures (>2100°C) have been achieved using a multistage casket consisting of i) a BN crucible enclosing the specimen, ii) a granular bed of BN particle/zirconia fibers outside the BN crucible, and iii) an outer layer of an alumina fiberboard.[12]

In addition to thermal runaway, microwave processing encounters the typical problems associated with high-temperature processing materials. For example, at elevated temperatures, local melting can occur because of eutectic reactions between materials. Solid-state diffusion or vapor phase transport between the specimen and the setter can contaminate the specimen. Another type of problem encountered during high-temperature processing is that fibrous and fiberboard materials tend to densify (and hence shrink and crack) if held at elevated temperatures for extended lengths of time. This may limit the useful lifetime of ceramic fiberboard casket elements to perhaps between 5 and 20 heating cycles.

HIGH-TEMPERATURE PROCESSING OF CERAMICS

Energy Savings

Microwave heating can offer significant energy savings compared with conventional heating,[1] especially for the high-temperature densification or joining of ceramics. For conventional radiant heating, significant thermal energy losses occur from the furnace walls. In contrast, for microwave heating the cavity walls

are relatively cool, as little microwave energy is absorbed directly by the cavity walls when lossy dielectrics such as ceramics are heated. A useful gauge of the energy cost is the energy consumed per unit time per unit mass of a processed ceramic. For example, the microwave sintering of alumina at 1600°C consumes about 4 kWh/kg, while conventional sintering consumes roughly 60 kWh/kg. Similarly, microwave sintering consumes roughly 3 kWh/kg during the 2 hr needed to densify a silicon nitride specimen, while for the same specimen conventional heating would consume roughly 20 kWh/kg for each hour of the 12 hr processing cycle.[1]

Microwave Sintering

During the 1980s and 1990s, much of the effort in microwave processing of ceramics focused on sintering. During that time, there also was a growing interest in ceramic/ceramic joining using microwave energy.

Microwave sintering has been successful with many different ceramics, including oxides such as alumina, zirconia, and hydroxyapatite (HAP) as well as nonoxides such as silicon nitride, silicon carbide, and titanium diboride.[10] The time–temperature history of the processing cycle (Fig. 7) varies with the details of the material and the furnace. For conventional heating, the sintering hold time, T_{hold}, is often several hours or more (Fig. 7). For microwave sintering, T_{hold} is typically 10–20 min or less and T_{hold} may approach zero to minimize grain growth.[3]

The primary interest in microwave sintering of ceramics has been in the promise of ''microstructural engineering'' of particular materials. For example, microwave sintering of SrO- and MgO-doped $LaGaO_3$ rapidly produces (within 10 min) a dense polycrystalline body with a more uniform microstructure than is obtained via conventional sintering,[13] with microwave heated specimens having a 4 μm average grain size. In contrast, the smallest mean grain size that the same researchers were able to achieve via conventional sintering was 10 μm.[13]

In general, covalently bonded materials are difficult to sinter because of their inherently low value of self-diffusivity. However, using microwave heating, silicon nitride with 20 wt.% yttria-doped zirconia has been sintered at 1400°C in a nitrogen atmosphere at a pressure 0.1 MPa. In contrast, via a conventionally heated hot isostatic press, a sintering temperature of 1850°C and a nitrogen pressure of about 180 MPa were required to densify the same silicon nitride/zirconia composition as completely as was achieved by microwave heating at 1400°C.[3]

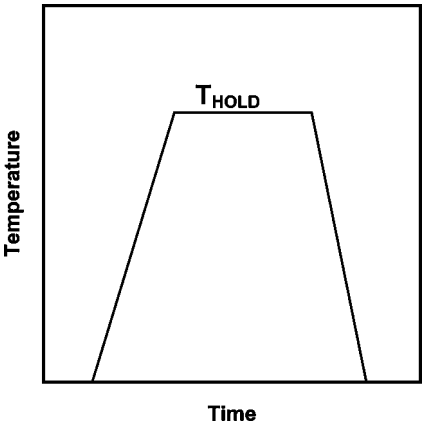


Fig. 7 A schematic of the time–temperature history of the processing cycle. Typically for conventional heating, the time at T_{hold}, the sintering or joining temperature, is in the order of 1 hr to several hours. For microwave processing, the time at temperature T_{hold} is 1020 min or less.

Ceramic/Ceramic Joining

Joining of dissimilar ceramics provides the synergistic advantages of combining materials with differing mechanical, thermal, electrical, or magnetic properties. In addition, ceramic/ceramic joining can create internal channels in ceramics by joining subcomponents with surface channels,[14] thus transforming the surface channels into bulk-penetrating conduits for cooling fluids, fuels, biochemical fluids, or medicine. Ceramics have been joined using both conventional and microwave processing,[15] but as is the case for sintering of ceramics, microwave joining may provide a means of joining at lower temperatures and in shorter times than is possible for conventional processing.

Successful ceramic/ceramic joints between both similar and dissimilar ceramic materials have been achieved both i) by layering materials in the green state, then firing,[16] and ii) by joining fully dense materials (Fig. 8).[17–19] In some instances, microwave

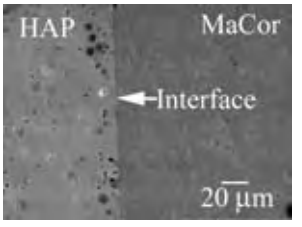


Fig. 8 A scanning electron microscope micrograph of microwave-joined MaCor™ and hydroxyapatite, joined at 1020°C for 20 min in a single-mode 2.45 GHz microwave cavity. (MaCor is a mica-platelet reinforced glass ceramic and HAP is a bioceramic material.) (From Ref.[19]. Reprinted with permission of The American Ceramic Society, www.ceramics.org. Copyright 2003. All rights reserved.)

joining can enhance the performance of the joint, compared with the joining by conventional heating. For example, for ceramic/ceramic joints formed in polycrystalline MgF_2 (an infrared window material), joined specimens exhibited a loss of transmittance because of a conventionally formed joint, while joints formed at the same temperature by microwave heating showed essentially no loss in transmittance because of the presence of the joint.[20]

Thermal Etching

Grain boundary grooving (also called thermal etching) occurs during the high-temperature treatment of polycrystalline ceramics by either conventional or microwave heating.[21] The grooves, which represent a balance between grain boundary and surface energies, form along grain boundaries via mass diffusion. Grain boundary etching is of considerable practical importance in terms of grain size measurement. (The surfaces of sintered specimens are polished and then etched to reveal grain boundaries, which allows one to determine both grain size and grain shape.) In the last 10–15 years, thermal grooving has supplanted chemical etching as perhaps the most commonly used method to facilitate grain size measurement in both ceramics and metals, because thermal etching is typically easier to control and yields more uniform results than chemical etching.

An accelerated rate of grain boundary grooving (thermal etching) has been observed in alumina ceramics (Fig. 9), where the etching rate was especially enhanced as the level of impurities increased.[21] The microwave enhancement of the grain boundary etching

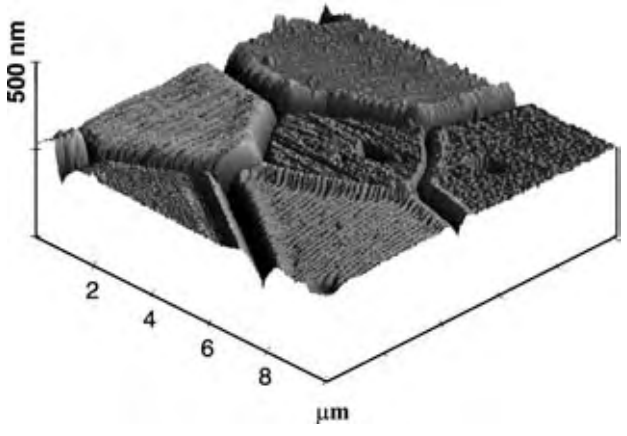


Fig. 9 An atomic force microscope (AFM) image of a polycrystalline alumina specimen that has been polished and then thermally etched by heating for 1 hr at 1527°C in a 2.45 GHz resonant cavity microwave. (From Ref.[21]. Reprinted with permission of EDP Sciences, P.A. de Courtaboeuf, Les Ulis Cedex, France.)

of polycrystalline ceramics is of practical interest, as grain boundary etching is a typical and reliable method of determining the grain size (GS). For both research and development of ceramic materials and quality control for production of ceramic materials, grain size measurements are crucial because a number of mechanical, thermal, and electrical properties of ceramics are functions of grain size.[2]

LOW AND INTERMEDIATE TEMPERATURE PROCESSING OF CERAMICS

Microwave-Assisted Synthesis of Ceramic Powders

While microwave heating has been used to accelerate chemical synthesis in a variety of organic systems, microwave-assisted ceramic powder production also has been done successfully for a number of years.[22] Since the early 1990s, a wide range of ceramic oxides and ferrites have been synthesized using microwave energy,[22,23] along with selected carbides, nitrides, and chalcogenides.[24]

Microwave heating can accelerate chemical reactions in liquids via dielectric heating, reducing the time required to fabricate a given ceramic powder by a factor of 10 or more.[22] It also offers the significant benefits of reducing the particle size and enhancing powder uniformity compared with powder synthesis via conventional heating. In addition, microwave-assisted synthesis of ceramic powders can typically be done at relatively low temperatures with low microwave power. As a consequence, ceramic powder synthesis often can be performed using domestic 2.45 GHz multimode microwave ovens, which tend to be relatively simple, inexpensive, robust, and easily adapted to a variety of uses.

A key aspect of the uniformity of the temperature field in both low- and high-temperature processing is the nature of the thermal gradients within the material. Consider the temperature distributions within a flat ceramic slab of thickness L (Fig. 10). For microwave heating (top curve in Fig. 10), the temperature is relatively uniform within the bulk, with a drop in temperature near the specimen surface owing to heat losses. In contrast, for conventional heating from the specimen surfaces (bottom curve in Fig. 10), the temperature is highest at the surface and lowest near the specimen's midplane.

Microwave-assisted synthesis using a copper acetate and sodium hydroxide solution with an ethanol solvent has produced quasispherical CuO nanoparticles with most particles ranging from about 3 to 5 nm and a mean grain size of roughly 4 nm.[25] In general, it is not easy to fabricate ceramic nanoparticles, in part because it is difficult to achieve the uniform thermal

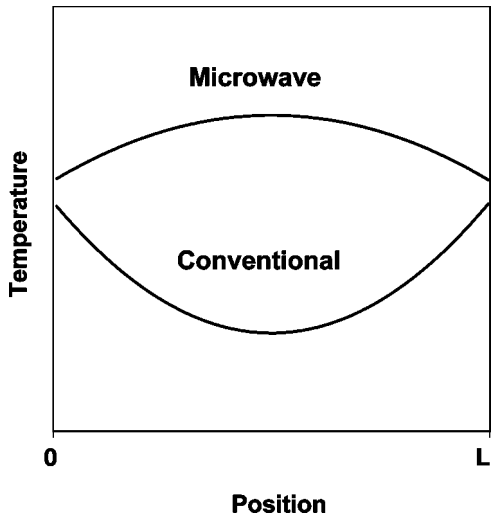


Fig. 10 Schematic of temperature distributions in a flat ceramic slab of thickness L for volumetric microwave heating (top curve) and conventional heating from the slab surfaces (bottom curve). For conventional heating, the finite value of thermal conductivity, k, gives the highest temperatures near the specimen surface and the lowest temperature along the specimen's midplane. Conversely, for microwave heating the heating is more uniform, with decreasing temperature near the slab surface because of heat losses from the surfaces.

environment needed for nanoparticles to nucleate and grow. Fortunately, microwave energy couples well with polar solvents resulting in uniform heating of the solution that in turn promotes the uniform reactant concentration and temperature fields required to synthesize ceramic nanoparticles.[25]

At a temperature of about 80°C, heating using a 2.45 GHz 650 W kitchen microwave oven produced spherical, submicron zirconia powders from zirconyl chloride precursor solutions.[26] A significant link between stirring, microwave heating, and the consequent powder size distributions was noted. For conventional heating, stirring affects the particle size distribution. Without stirring the particle size distribution is broad, while stirring during heating increases the uniformity of both the reactant concentration and the temperature field within the solution, and thus stirring gives a narrower particle size distribution. However, a disadvantage of stirring is that it leads to a shear-induced particle agglomeration.[26] For microwave heating, stirring is not necessary to obtain a uniform temperature field. Microwave heating produces simultaneous nucleation throughout the solution, yielding uniform size distribution without the stirring-induced particle agglomeration.[26]

Suboxide SnO is a candidate anode material for lithium-ion secondary batteries.[26] For the battery to function properly, it is critical that the material remains as the suboxide SnO rather than being oxidized to SnO_2. During conventional SnO processing, the solution is heated (aged) for more than 1 wk to reduce the oxidation rate, then crystallization is induced by further heat treatment (at roughly 75–95°C). In contrast, microwave synthesis of SnO can be completed in roughly 4 min using intermittent or pulsed heating consisting of 1 min long heating intervals interleaved with zero power intervals when the material is allowed to cool.[27] After the powder synthesis is complete, the powders can also be dried by microwave heating, enabling the entire process to be completed in about 15 min, thus avoiding the weeklong aging time.

In addition to the reduced processing time, the physical properties of microwave-fabricated SnO powders can be significantly different than conventionally prepared SnO powders. While microwave heating gives crystalline SnO, conventional heating yields amorphous SnO. In addition, microwave heating enhances the crystallization rate but does not that of oxidation. The microwave acceleration of SnO crystallization rates may be an athermal effect that may be linked to the coupling of microwave energy with atomic reactive sites.[27]

Ceramic Coatings on Powders, Fibers, and Substrates

Ceramic coatings on fibers and powders have a variety of uses. For example, porous ceramic coatings on nanoscale metallic or ceramic particles can improve the catalytic properties of a powder.[28,29] Also, the carbon fibers used as reinforcement in metallic matrices can be coated with a thin ceramic film (such as SiC or TiN) to reduce the rate of interdiffusion that may occur between the matrix materials and the fibers, and enhance the wetting of the fiber surface by metals.[29]

As is the case with ceramic sintering, joining, and powder synthesis, microwave-assisted fabrication of ceramic coatings can offer unique benefits. To expose the material surfaces for reaction with a gas phase, fibers[29] or powder particles[28] may be suspended by the flow of gas in the fluidized bed. In addition to the flow of the fluidizing gas, low frequency mechanical vibrations (for example, 10–15 Hz) may be applied to the container to reduce the tendency for the powders or fibers to sinter together to form clumps.[28] As the powder size decreases, the powder's sinterability increases. While vibration may not be necessary for particles as large as, say, 50 μm, it may be beneficial for powder particles that are a few microns in diameter.[28]

Careful control of the microwave-assisted reactions can produce nanometer-thick coatings.[28,29] For example, during the microwave-assisted fabrication of ceramic coatings on metallic powders, the microwave power may be switched on and off to produce a pulsed

microwave input signal with a pulse duration in the order of a few seconds and a pulse repetition rate of roughly 0.1–0.5 Hz. Pulsing the input power helps to thermally quench the outer surface of the particle to avoid the initiation of a self-propagating high-temperature synthesis (SHS) reaction that (for susceptible materials) would otherwise consume the entire particle rather than limiting the reaction to a thin outer film.[28] For metals, the skin depth for microwave radiation is in the order of a fraction of a micron, and this limited volume of microwave-induced heating also assists in limiting the formation of the ceramic coating to a thin film on the metal particle surface.

Microwave-assisted fabrication of ceramic coatings on fibers and powders can be done at intermediate temperatures. For example, at temperatures of about 800–900°C, carbon fibers have been coated with thin TiN layers using a microwave plasma-assisted fluidized bed with the precursor vapor ($TiCl_4$ in this case) being introduced into the reactor by the fluidizing gas.[29]

In addition to depositing thin coatings on the surfaces of powders and fibers, microwave energy can aid in processing thin, protective films on planar electronic substrates. Submicron thick boron carbon nitride (BCN) thin films have been successfully deposited on Si(1 0 0) substrates using plasma-assisted pulsed laser processing at temperatures lower than 80°C.[30] A pulsed Nd/YAG laser first ablated the surface of a boron carbide target in a nitrogen environment, then a microwave-induced low-energy, low-temperature nitrogen plasma was used to ''pin'' the BCN film onto the silicon substrates. The quality of the plasma deposited films was enhanced compared with that of boron carbide films deposited in vacuum, namely the plasma-assisted deposition process gave a lower oxygen concentration than films grown by conventional heating in vacuum.[30]

Microwave Drying of Granular Ceramics, Sol–Gel Precursors, and Cement

Microwave energy already has proven to be very advantageous for the rapid drying of granular food materials, grains, and pharmaceuticals, where the microwave-assisted drying rates are up to six times faster than those for conventional heating.[31] While conventional drying of particles in a fluidized bed relies on heat transfer via convection from the circulating drying gas to the wet surfaces of the particles, microwave heating rapidly achieves relatively uniform temperatures throughout the fluidized bed volume by the direct coupling of microwave energy to either water or other microwave-absorbing (polar) liquids.[31]

Fluidized bed drying is potentially important in the immobilization of high-level liquid radioactive waste (HLLW) materials. For example, sol–gel derived ceramic microspheres with diameters of 20–50 μm and porosities of about 40–50% are currently used to immobilize HLLW.[31] Efficient drying of the HLLW is related to the good coupling between microwave radiation and water in the porous ceramic microspheres. In particular, for a pilot-scale drying operation of 10 kg batches of ceramic microspheres, the microwave energy consumed per kilogram of moisture removed was about 6.6 J/kg for a moisture loading of 0.10 kg moisture/kg powder, while for a moisture loading of 0.20, the microwave energy consumed per kilogram of moisture decreased to about 5.5 J/kg.[31] Thus, the microwave energy consumed per kilogram of water removed decreased as the moisture content increased, such that microwave drying can become even more efficient as the moisture content increases.

Microwave radiation has also been used to accelerate the curing of cement paste and mortar. For Portland cement mortar, microwave heating at temperatures between 40°C and 60°C can shorten the curing time by up to a factor of 3 while maintaining the long-term strength of the mortar.[32] However, microwave curing at 80°C and above can decrease the strength of the mortar. In fact, if the curing temperatures extend above ≈90°C, then steam generated in the specimen bulk can generate cracks in the cement.[32] The results for the acceleration of curing were roughly similar for Portland cement paste and for mortars blended with pozzolanic materials including silica fume, blast furnace slag, and fly ash.

Another ceramic processing technique that involves drying as a key step is that of the sol–gel production of ceramic powders. Sol–gel precursors consist of small particles weakly bound together, permeated with networks of fine, interconnected pores.[33] These pores typically are filled with either an alcohol-based or a water-based liquid. Microwave heating has been successfully applied to dry sol–gel precursors.[34]

ATHERMAL EFFECTS

Athermal (or alternatively, nonthermal) effects in microwave processing have been a subject of long-standing controversy. With respect to microwave processing, ''thermal'' effects involve the direct conversion of microwave energy to heat, yielding a particular temperature field in the body. The thermal energy in turn stimulates the standard thermally activated processes associated with chemical reactions and mass diffusion. If only thermal effects are present, then given a particular temperature field, the effects on sintering rate or microstructural development are identical, regardless of whether the agent producing the given thermal field

is a microwave or conventional heating. A crucial point is that the temperature distribution produced in the microwave can be fundamentally different than the temperature distribution created by thermal heating, as microwave heating of ceramics can lead to the volumetric, selective, and/or rapid heating effects. Thus if solely thermal effects are present, then these effects only modify the temperature distribution within the body, based on the details of how the microwave energy is absorbed within the body according to its dielectric properties [Eq. (1b)].

However, there have been a number of reports of athermal effects in processing ceramic materials, where the sintering rate or the microstructure evolution (grain size and/or porosity) resulting from microwave heating differed from that obtained by a conventional heat treatment at the same temperature. Thus, athermal effects refer to mechanisms that operate in addition to the conventional thermal effects and may be a function of, for example, the electric field intensity or the frequency.

Modeling of possible microwave-induced athermal effects has been done in terms of a ''ponderomotive effect,'' which is enhanced by the tangential component of the electric field at grain boundaries.[1] The ponderomotive force is likely enhanced during sintering by the geometry of the small interparticle neck regions between the particles.[1] Nevertheless, sintering theory and practice specify that the interparticle necks are only small during the initial stages of sintering,[2] implying that the ponderomotive force may be of lesser importance during the intermediate and final stages of sintering.

CONCLUSIONS

The rapid, volumetric heating provided by microwave energy has a number of potential benefits. The resulting inverse temperature distribution (Fig. 10) promotes rapid and efficient drying of granular ceramics, sol–gels, and concrete. The uniform heating of polar liquids means that microwave heating can produce ceramic nanopowders with narrower grain size distributions and smaller mean particle sizes than are achievable by conventional heating. Moreover, the ability to precisely control the frequency and repetition rate of the microwave energy aids in depositing nanometer-thick ceramic coatings on particles, fibers, and planar substrates.

Microwave sintering, joining, and thermal etching are three high-temperature processes for which there have been claims that microwave energy can densify, join, or etch at lower temperatures and/or shorter times than are required to produce the same results by conventional heating. Regardless of the presence of athermal microwave effects, the temperature distributions achieved via microwave heating are different than those resulting from conventional heating, so that differences are to be expected between the results obtained from conventional and microwave heating. Also, especially with respect to high-temperature processes, microwave heating provides energy savings, in part because microwave heating avoids the substantial furnace-wall losses that are omnipresent in conventional radiant heating.

Microwave processing of ceramics will likely expand in scope in the future. The processes of slip casting and tape casting are additional examples that involve controlled drying of granular ceramics and are thus areas of potential application for microwave drying. In addition, large microwave ovens of the type now used to dry lumber may be used to speed up the drying of preformed concrete structures. The rapid development of microwave-assisted ceramic powder synthesis will likely continue in the future, encompassing an even greater scope of oxide and nonoxide powder production. The deposition of nanometer-scale ceramic coatings on fibers, powder particles, and substrates will also expand and diversify in the future.

REFERENCES

1. Bykov, Y.V.; Rybakov, K.I.; Semenov, V.E. High-temperature microwave processing of materials. J. Phys. D.: Appl. Phys. **2001**, *34*, R55–R75.
2. Barsoum, M.W. *Fundamentals of Ceramics*; McGraw-Hill: New York, 1997.
3. Link, G.; Feher, L.; Thumm, M.; Ritzhaupt-Kleissl, H.J.; Bohme, R.; Weisengurger, A. Sintering of advanced ceramics using a 30-GHz, 10-kW, CW industrial gyrotron. IEEE Trans. Plasma Sci. **1999**, *27* (2), 547–554.
4. Lee, K.Y.; Case, E.D.; Asmussen J., Jr. The steady-state temperature as a function of casket geometry for microwave-heated refractory caskets. Mater. Res. Innovations **1997**, *1* (2), 101–116.
5. Lee, K.Y.; Dearhouse, P.H.; Case, E.D. Microwave sintering of alumina using four different single-cavity modes. J. Mater. Syn. Process **1999**, *7* (3), 159–166.
6. Westphal, W.B.; Sils, A. Dielectric constant and loss data. In *Technical Report AFML-TR-72-39*; Air Force Materials Laboratory: Wright-Patterson Air Force Base, OH, 1972; Vol. 70, 19–21.
7. Mak, P.; Asumssen, J. Experimental investigation of the matching and impressed electric field of a multipolar electron cyclotron resonance discharge. J. Vac. Sci. Technol. A. **1997**, *15* (1), 154–168.

8. Hill, J.M.; Marchant, T.R. Modeling of microwave heating. Appl. Math. Model. **1996**, *20*, 3–15.
9. Jackson, H.W.; Barmatz, M.; Wagner, P. Transient temperature behavior of a sphere heated by microwaves. In *Microwaves, Theory And Applications In Materials Processing II; Ceramic Transactions*; Clark, D.E., Tinga, T.R., Laia, J.R., Jr., Eds.; Am. Cer. Soc.: Westerville, OH, 1993; Vol. 36, 189–199.
10. Zhao, C.; Vleugels, J.; Groffils, C.; Luypaert, P.J.; Van der Biest, O. Hybrid sintering with a tubular susceptor in a cylindrical single-mode microwave furnace. Acta Mater. **2000**, *48*, 3795–3801.
11. Lee, K.Y.; Case, E.D. Steady-state temperature of microwave-heated refractories as a function of microwave power and refractory geometry. Mater. Sci. Eng. **1999**, *A269*, 8–20.
12. Xu, G.F.; Olorunyolemi, T.; Carmel, Y.; Lloyd, I.K.; Wilson O.C., Jr. Design and construction of insulation configuration for ultra-high-temperature microwave processing of ceramics. J. Am. Ceram. Soc. **2003**, *86* (12), 2082–2086.
13. Subasri, R.; Mathews, T.; Sreedharan, O.M. Microwave assisted synthesis and sintering of $La_{0.8}Sr_{0.2}Ga_{0.83}Mg_{0.17}O_{2.815}$. Mater. Lett. **2003**, *57*, 1792–1797.
14. Case, E.D.; Ren, F.; Kwon, P.; Kok, C.K.; Rachedi, R.; Klenow, B. Machining and ceramic/ceramic joining to form internal meso-scale channels. Int. J. Appl. Ceram. Technol. **2004**, *1* (1), 95–103.
15. Case, E.D.; Crimp, M.A. Joining of ceramic materials using spin on interlayers. Adv. Eng. Mater. **2001**, *3*, 395–399.
16. Zheng, J.; Akinc, M. Green state joining of SiC without applied pressure. J. Am. Ceram. Soc. **2001**, *84* (11), 2479–2483.
17. Binner, J.G.P.; Fernie, J.A.; Whitaker, P.A. An investigation into microwave bonding mechanisms via a study of silicon carbide and zirconia. J. Mater. Sci. **1998**, *33* (12), 3009–3015.
18. Binner, J.G.P.; Fernie, J.A.; Whitaker, P.A.; Cross, T.E. The effect of composition on the bonding of alumina ceramics. J. Mater. Sci. **1998**, *33* (12), 3017–3029.
19. Case, E.D. Selection and function of interlayer materials in ceramic/ceramic joining. In *Ceramics Transactions*; Lewinsohn, C.A., Singh, M., Loehman, R., Eds.; American Ceramic Society: Westerville, OH, 2003; Vol. 138, 3–28.
20. Case, E.D.; Lee, J.G.; Lee, K.Y. Joining of optical and infrared materials using spin-on layers. In *Joining of Advanced and Specialty Materials*; Singh, M., Indacochea, J.E., Hauser, D., Eds.; ASM International: Materials Park, OH, 1998; 17–26.
21. Lee, K.Y.; Case, E.D. A comparison of theoretical and experimental profiles for thermally-induced grain-boundary grooving. Eur. Phys. J., Appl. Phys. **1999**, *8* (3), 197–214.
22. Komarneni, S.; Li, Q.H.; Roy, R. Microwave-hydrothermal processing of layered anion exchangers. J. Mater. Res. **1996**, *11* (8), 1866–1869.
23. Costa, A.C.F.M.; Fagury-Neto, E.; Morelli, M.R.; Kiminami, R.H.G.A. Microwave synthesis of Ni–Zn ferrite powders. Mater. Sci. Forum **2003**, *16* (4), 705–710.
24. Peelamedu, R.D.; Fleming, M.; Agrawal, D.K.; Roy, R. Preparation of titanium nitride: microwave-induced carbothermal reaction of titanium dioxide. J. Am. Ceram. Soc. **2002**, *85* (1), 117–122.
25. Wang, H.; Xu, J.Z.; Zhu, J.J.; Chen, H.Y. Preparation of CuO nanoparticles by microwave irradiation. J. Crystal Growth **2002**, *244*, 88–94.
26. Moon, Y.T.; Kim, D.K.; Kim, C.H. Preparation of mono-dispersed ZrO_2 by the microwave heating of zirconyl chloride solutions. J. Am. Ceram. Soc. **1995**, *78* (4), 1103–1106.
27. Wu, D.S.; Han, C.Y.; Wang, S.Y.; Wu, N.L.; Rusakova, I.A. Microwave-assisted solution synthesis of SnO nano-crystallites. Mater. Lett. **2002**, *53*, 155–159.
28. Jain, A.; Brezinsky, K. Microwave-assisted combustion synthesis of tantalum nitride in a fluidized bed. J. Am. Ceram. Soc. **2003**, *86* (2), 222–226.
29. Tap, R.; Willert-Porada, M. Synthesis of composite powders and coating of fibers by combined CVD-PECVD in a microwave heated fluidized bed reactor. In *Microwave and Radio Frequency Applications*; Folz, D.C., Booske, J.H., Clark, D.E., Gerling, J.F., Eds.; The American Ceramic Society: Westville, OH, 2003; 89–97.
30. Ling, H.; Wu, J.D.; Sun, J.; Shi, W.; Ying, Z.F.; Li, F.M. Electron cyclotron resonance plasma-assisted pulsed laser deposition of boron carbon nitride films. Diamond Relat. Mater. **2002**, *11*, 1623–1628.
31. Sizgek, E.; Sizgek, G.D. Drying characteristic of porous ceramic microspheres in microwave heated fluidized bed. Chem. Eng. Technol. **2002**, *25* (3), 287–292.
32. Sohn, D.; Johnson, D.L. Microwave curing effects on the 28-day strength of cementitious materials. Cement Concrete Res. **1999**, *29*, 241–247.
33. Rahaman, M.N. *Ceramic Processing and Sintering*; 2nd Ed.; Marcel Dekker: New York, 2003.
34. Ananthakumar, S.; Hareesh, U.S. Microwave drying of sol–gel alumina–30 vol.%SiC precursor for particulate ceramic matrix composites. Brit. Ceram. Trans. **1998**, *97* (5), 236–239.

Mixing and Chemical Reactions

M

Edward L. Paul
Merck and Co., Inc., Sea Girt, New Jersey, U.S.A.

Suzanne M. Kresta
Department of Chemical and Materials Engineering, University of Alberta, Edmonton, Alberta, Canada

Arthur W. Etchells
DuPont Fellow, Philadelphia, Pennsylvania, U.S.A.

INTRODUCTION

In developing a complex reaction for scale-up, or to improve yield, the mixing issues that must be considered are: 1) will mixing affect the reaction beyond the requirement to blend the reagents; 2) how are these mixing requirements evaluated; and 3) what mixing equipment design will optimize yield and selectivity? This chapter provides guidelines for addressing these questions. The same concepts can be applied to several other processes involving rapid mixing controlled steps that are not usually thought of as kinetically controlled, e.g., crystallization, precipitation, and physicochemical steps, such as surface and dispersion stability. Understanding the impact of mixing on these processes can improve yields by as much as 20% and reduce by-product formation, and can mean the difference between a successful product and one that fails on scale-up.

MIXING-SENSITIVE REACTIONS

Mixing can be a rate limiting mechanism in chemical reactions, reducing the rate of fast reactions and changing the rate, yield, and product distribution for complex reactions. Both homogeneous and heterogeneous reaction systems can be affected. This issue was first identified by Danckwerts and Levenspiel.[1,2] Since that time, the field has advanced significantly, moving well beyond the concept of residence time distribution to the modeling of local mixing rates and time varying concentration fields. Recent comprehensive treatments are given in Refs.[3,4].

Mixing effects in chemical reactions must be formulated in terms of local mixing rates or local mixing times. The easily formulated global blend time seldom has an effect, while the time constants based on local conditions in the reactor, such as local mixing time or local mass transfer rate can be very important. When the rates of reaction, local mixing, and mass transfer approach one another, mixing will affect the outcome of the process. Reactions cannot take place until the reagents are mixed locally to the molecular level. Thus, the processes of mixing/blending and chemical reaction operate in series initially, and then in parallel. When the chemical reaction is slow with a half-life of several minutes, and the mixing takes place quickly, say within seconds, then the mixing is essentially completed before a significant chemical reaction takes place. There is no effect of mixing on the slow chemical reaction and ideal reactor analysis can be used. The mixing requirement is satisfied by the blending of the reagents. When the chemical reaction is fast, however, the local mixing rate can be the rate-determining step.

Mixing effects can change the apparent kinetics of the reaction so that the measured kinetics is limited by the rate of mixing rather than by the rate of reaction. This problem is so pervasive that the rate of mixing for fast reactions is often mistaken for the rate of a chemical reaction. For a very fast chemical reaction, for example, an acid–base neutralization with a half-life of 0.001 sec, the rate of the chemical reaction depends on the rate of the mixing, which is much slower. If the reaction rate was measured, the result would be the mixing rate and not the molecular chemical reaction rate. The result is the ''apparent'' reaction rate.

When reactions are fast relative to the mixing rate, not only are the apparent reaction rates affected but the whole time and temperature history of the reaction mechanism is also affected, yielding different selectivities and yields, depending on the intensity of the mixing. This often leads to a scale-up/scale-down problem, where yields of the desirable products in a plant-scale reactor are not as good as those in a small-scale reactor in the laboratory or the pilot plant. If the yield drops from the pilot-scale to the plant-scale reactor when all other important variables (temperature, pressure, and composition) have been held constant, then there is a mixing problem. Fast

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030914

chemical reactions where the mixing rate and the reaction rates are intertwined may be either homogeneous (single phase) or heterogeneous (multiphase). Typical examples of such reactions can be azo-couplings, esterification, nitrations, sulfonations, and aminations.[5] In either case, local mixing rates provide important information for successful scale-up.

Time Scales and Local Mixing Rates

The competition between reaction and mixing is given by the Damkoehler number (*Da*), which is the ratio between the reaction rate and the local mixing rate, or conversely, the ratio of the characteristic local mixing time τ_M and the reaction time τ_R:

$$Da = \tau_M/\tau_R \quad (1)$$

A smaller *Da* indicates less effect of mixing. Larger *Da* indicates that mixing will be a concern. Local mixing rates and mass transfer rates can be accurately estimated for many common reactor configurations of both homogeneous and heterogeneous reactions. These estimates are combined with an estimate of the reaction rate to identify the conditions under which mixing effects may be critical to the course of a reaction. Care should be taken in estimating the characteristic time constant for a reaction. Many reactions consist of a series of steps and some of the initial ones may be mixing-controlled even though the overall reaction time is long. Classical examples of mixing-controlled reactions are the competitive-consecutive and competitive-parallel schemes, where the desired reaction is fast and an undesired side reaction is much slower, but fast enough to be of the order of the local mixing time.

Reaction Schemes of the Greatest Interest

Mixing effects on product distribution are of importance in multiple reactions because the impact of product distribution on design and economics can be profound. In such reactions, the desired product is one of two or more possible products. Economics is directly affected by the yield of the desired product and both design and economics are affected by downstream separation requirements.

Mixing effects may be significant for two major types of complex reaction:

1. Competitive-consecutive
   a. $A + B \rightarrow R$
   b. $R + B \rightarrow S$
2. Competitive-parallel
   a. $A + B \rightarrow R$
   b. $A + C \rightarrow U$

Fig. 1 illustrates the mixing problem for the competitive-consecutive case. R is the desired product and S is the undesired overreaction product. During the time from when the reactants are first contacted to when they are completely mixed on the molecular scale, reaction of A with B to form the desired product R occurs along with the undesired reaction of R with B to form S. When A and B are well mixed at the molecular scale, mainly R is formed, but when there is a boundary between A and B, a significant amount of undesired S appears. Competitive-parallel reactions can be subject to similar mixing effects where the first reaction is the desired one and the second is a simultaneous decomposition of A to form the undesired U. While these two reaction systems have received the most attention, the course of any reaction that is

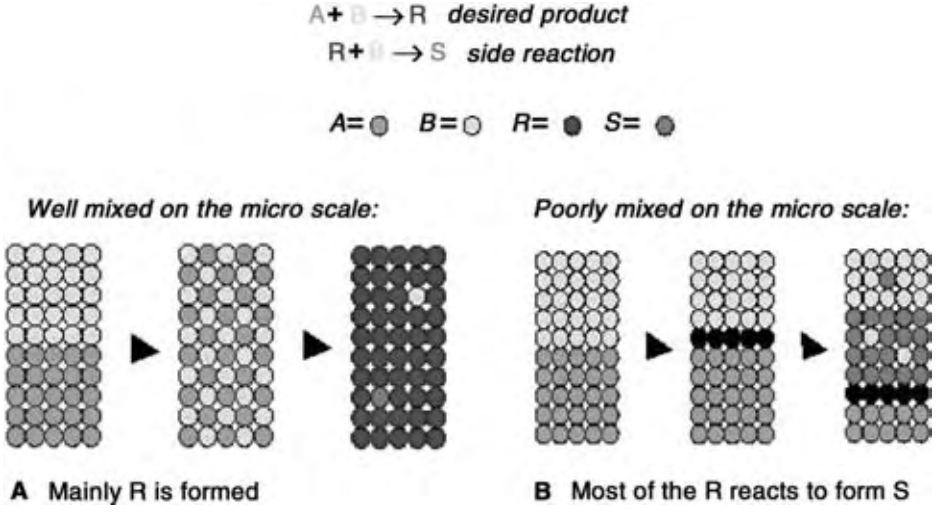


Fig. 1 Diffusion and chemical reaction at an A–B mixing surface. In this competitive-consecutive reaction the first reaction, which forms the desired product (R), is fast and the consecutive reaction step forming the undesired by-product (S) is slower. Local mixing conditions at the molecular scale determine the amount of undesired by-product (S) formed. *(View this art in color at www.dekker.com.)*

influenced by concentration has the potential to be influenced by mixing. The effect can be on the reaction rate and the product distribution or both.

In the ideal semibatch, consecutive-competitive case, reagent B is added to a tank containing A and is instantaneously mixed to the molecular level with the vessel contents. The objective is to determine how mixing conditions can affect the yield of the desired product. The maximum selectivity for the desired product R is a function of the rate constants k_{R1} and k_{R2}, the overall molar charge ratio of A to B, and the degree of conversion of A. The degree of conversion of A can depend on the charge ratio and the residence time. The discussion that follows is limited to the case of an equimolar charge ratio of A to B, and sufficient residence time for all of the B to react. The maximum selectivity for R in the absence of mixing effects then becomes a function only of k_{R1}/k_{R2} and the molar charge ratio. If we now fix k_{R1}/k_{R2} and the molar charge ratio, the selectivity is fixed, as are the yield and the degree of conversion of A.

The term "expected (ideal) yield," Y_{exp}, is used to denote the yield that would be obtained for a competitive-consecutive reaction under conditions of perfect mixing and complete conversion of the limiting reactant, as presented in Eq. (2):

$$Y_{exp} = R/A_0 = 1/(1 - \kappa)[(A/A_0)^{\kappa} - (A/A_0)] \quad (2)$$

where $\kappa = k_{R2}/k_{R1}$ and R and A denote molar concentrations. This equation applies to both batch and semibatch operations provided both the reaction rates depend on B in the same way (e.g., second order), B is added to A in the semibatch case, and B is completely consumed. The relationship between the magnitude of the secondary reaction rate constant and the yield is illustrated in Fig. 2, where the mixing *Da*, based on k_{R2}, is plotted against X_S, the yield of unwanted product based on the consumption of B. This is one minus the yield. This relationship was developed by Bourne using X_S to represent the amount of S formed where $X_S = 2S/(2S + R)$ and k_{R2} represents the undesired reaction kinetics.[6] The figure shows that mixing effects of homogenous reactions can only reduce yield below the expected (ideal) level as calculated by Eq. (2). The primary concern is the magnitude of the yield reduction attributable to deviation from instantaneous, perfect mixing to the molecular scale.

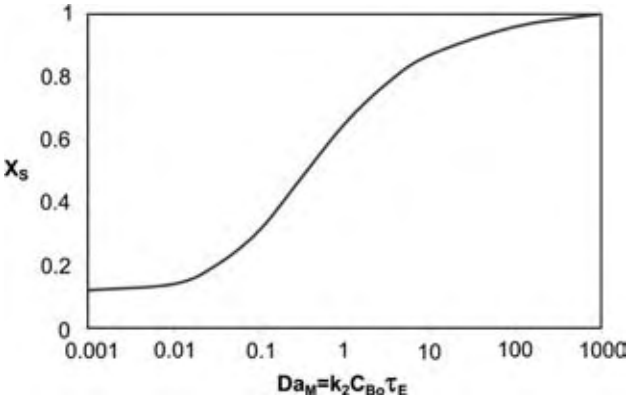


Fig. 2 By-product selectivity, X_s, as a function of *Da* based on k_2. This data from Bourne show the increased by-product formation with increasing mixing time based on the engulfment model, τ_E. As the reaction rate for the second reaction, k_2C_B, increases, the mixing time must decrease to maintain yield. (From Ref.[6].) *(View this art in color at www.dekker.com.)*

Relating Mixing and Reaction Time Scales: The Mixing *Da*

Here, we consider the characteristic local mixing time. The final phase of mixing during which chemical reactions can occur and before complete molecular homogeneity is achieved is the molecular diffusion controlled mixing of the smallest eddies in the turbulence energy dissipation spectrum. The smallest eddy size can vary over several orders of magnitude from $\sim$1 μm in intense jet mixing to $>$ 100 μm in stirred tanks with low-shear impellers. The time and length scales of turbulence and small-scale diffusion are discussed by Brodkey and Kresta.[4]

The magnitude of yield reduction because of imperfect mixing is determined by the following factors: 1) Local mixing time, a measure of the time from initial contact of the reactants to final homogeneity on a molecular scale at a given point. Any overreaction of R to S must occur during this time. 2) Chemical kinetics as given by the absolute values of the rate constants, k_{R1} and k_{R2}. 3) The *Da*, the ratio of rates of the first or the second reaction and the local mixing rate. The characteristic reaction time of interest is the one that produces the undesirable product. Other faster reactions are blending controlled. The reaction time constant consists of a reaction rate and a concentration. The concentration to be used is that of the feed, not the well-mixed concentration in the reactor. The reaction of interest takes place before dilution.

There are several mixing or blending times that can be measured and observed in an agitated vessel. The bulk blend time is the time it takes to reach a uniform concentration throughout the vessel. The local mixing time is a measure of how fast the material at a given point loses its identity. The local mixing time varies with position, while the bulk blending time is a characteristic of the whole tank. Bulk blending time is limited by the slowest rate of mixing in the vessel and varies

with the tank geometry. Local mixing times depend on the local turbulence.

Blend time

Although local mixing time is more relevant to yield effects of mixing-controlled reactions, blend times provide a rapid initial comparison of mixing and reaction time constants. If the reaction is slower than the blend time, mixing will not be an issue. The blend time is defined as the time required for the spatial variation of average concentration to drop to 5% of the original variation after a pulse disturbance to the vessel. Multiple probes must be used to obtain an accurate result. Typically, conductivity probes and salt-water pulses are used to make the measurements of blend time.[7,8] The Grenville correlations for blend times are extensively used for design and scale-up. They are dependent on the Reynolds number range as follows:

$$N\tau_B = 5.2(T/D)^2 \Big/ N_p^{1/3} \quad \text{for } N_p^{1/3}\,\mathrm{Re} > 6400 \tag{3}$$

$$N\tau_B = 3.35 \times 10^4 (T/D)^2 \Big/ \left(\mathrm{Re}N_p^{2/3}\right) \quad \text{for } 500 < N_p^{1/3}\mathrm{Re} < 6400 \tag{4}$$

Vessel blend times are typically 2 sec in a 1 L vessel and 20 sec in a 20,000 L vessel for low-viscosity liquids. If the blend time is small compared to the time constants of all key reactions, then there will be no mixing effects.

Local mixing time scales

In dealing with mixing effects on a reaction, two topics are of interest. The first is the rate of addition of the feed stream. The second is its rate of disappearance or the inverse local mixing time. For low-viscosity liquids, very rapid mixing times—with local mixing time constants, τ_M, as short as 0.01 sec—are easily obtained in liter-sized reactors, but because of mechanical limitations local mixing times of the order of 0.1 sec or longer often occur in reactors of 10,000 L or more. The rate of addition with the local mixing time determines the amount of the undesirable product that can be formed. There are many formulations for these two effects. The discussion of local mixing timescales must begin with a definition of the turbulent scales that underlie many mixing-time formulations. These scales are developed, as previously referenced, by Brodkey and Kresta.

Micromixing analysis considers the smallest scales of mixing, where the striations in the smallest turbulent eddies are dissipated by molecular diffusion. It is assumed that the amount added and the rates of addition are very small and that the scale of interest is set by the local turbulence. The earliest approach was to assume that the added material did not mix at the molecular scale until the Kolmogoroff scale was reached and the subsequent mixing took place by molecular diffusion. This is a good concept for moderate-viscosity turbulent systems and, with striation thickness replacing eddy size, can be used in fully laminar mixing systems. This molecular diffusion concept has been replaced by the engulfment model, τ_E, which is a more realistic way of treating the breakup of the added reactant as its scale is reduced by turbulence in low-viscosity systems:

$$\tau_E = 17\left(\frac{\nu}{\varepsilon}\right)^{1/2} \tag{5}$$

The differences in the approaches are small. Both include local energy dissipation and the viscosity. The molecular diffusivity is only important when the viscosity is high. In mechanically agitated tanks the local energy dissipation can vary by two to three orders of magnitude so that local mixing rates also vary dramatically. This provides a strong incentive to feed in the region of highest turbulence, close to the impeller. Local values of the dissipation, ε, are given by Patterson et al.[4] Alternate reactor configurations, where the dissipation is more uniform and can often be more tightly controlled, are increasingly in use in the chemical process industries. These include static mixers, impinging jet mixheads, and, in the more specialized field of reactive extrusion, twin screw extruders with kneading blocks.[9–11]

Mesomixing is used to describe a set of phenomena between macromixing, or blend time, which involves the whole vessel, and local micromixing at the smallest eddy scales. Thoma looked at the time of addition in semibatch operation and observed that when the addition time was very short there was an additional undesirable selectivity change.[12–14] High rates of addition overwhelm the local ability to take the material away by micromixing. The scales of mixing then penetrate into the larger eddy sizes in the inertial convective range of turbulence. This delays the molecular scale mixing. They called this phenomenon mesomixing. Mesomixing may be visualized as the mixing rate at the scale of the feed plume. Several models for mesomixing are discussed in Ref.[3]. This is an area of rapid development and change.

Summary: *Da*

Given the reaction time and the local mixing time one can calculate *Da* and use Fig. 2 for an estimate of yield and selectivity. Fig. 2, however, is based on the

assumption that the local mixing time is constant as the fed reactant moves away from the inlet position. In fact, there is a wide distribution of values of local mixing time constants in any mixing equipment and the entering material moves through many different zones of varying energy dissipation. The distribution of energy dissipation in a reactor is thus very important. This explains the interest in laser Doppler anemometry and computational fluid dynamics for the modeling of stirred tanks, and in static mixers and impinging jet mixheads as alternate reactor geometries.

This distribution of energy dissipation complicates any mathematical analysis immensely. It also explains why many modelers have gone to zone model analyses to more accurately predict the path of the reactants. In such models the vessel is divided into a number of zones of different energy intensities where the local mixing is allowed to vary. An example of such a model developed is given in Ref.[15]. Other examples of contributions in this area are discussed in Refs.[16,17].

In the absence of a full mathematical model of mixing and reaction kinetics, but given good estimates of the reaction and mixing rates, reactor design focuses on the following:

1. Determining and/or predicting which reactions are mixing sensitive.
2. Guidelines to minimize yield loss in mixing-sensitive reactions on scale-up from bench-scale to industrial reactors.

The homogeneous reaction case is considered first. All of the mixing concepts for homogeneous reactions will also apply to heterogeneous reactions, but with the added complication of multiphase mass transfer.

REACTOR DESIGN FOR MIXING-SENSITIVE HOMOGENOUS REACTIONS

The *Da* is extremely useful for initial evaluation of reaction conditions under which mixing effects must be considered. The reaction rate constant of the undesired reaction, k_{R2}, can vary over several orders of magnitude. When experimental data are not available but the class of reaction is known, the magnitude of k_{R2} can be estimated within two orders of magnitude or less. The mixing rate in vessels should not vary by more than two orders of magnitude. With these bracketed values, upper and lower limits on *Da* can be readily estimated and used as a first measure of mixing sensitivity as follows:[6]

$Da < 0.001$—Reaction is much slower than mixing and the selectivity is determined by chemical kinetics alone.

$Da > 1000$—Reaction is much faster than mixing and the selectivity can approach the asymptotic limit for minimum yield.

$0.001 < Da < 1000$—Reaction and mixing rates compete and both micromixing and chemical kinetics must be considered.

These concepts can be used to implement a developmental protocol for a new chemical reaction.[4]

Once the reaction is established, three key factors must be addressed for successful development and scale-up of homogenous reactions: 1) the effect of concentration on yield for competitive-consecutive reactions; 2) the effect of feed rate or addition time on yield; and 3) feed pipe backmixing.

Effect of Concentration

The rates of a reaction are determined by rate constants, concentrations of reactants, and temperature. At a given temperature, the rate of a second-order reaction depends on the product of the rate constant and the concentrations of the reactants. The concentration at which the chemical reaction becomes faster than the mixing is the critical concentration at which conversion and yield will be affected by mixing. It must be emphasized that the relative reaction rate at a point in the reactor is proportional to the product of the rate constant and the local concentration. It is recommended that bench and pilot data for mixing-sensitive reactions can be obtained at the same concentrations as they are to be used in the commercial plant. This eliminates concentration as a concern in scale-up.

Effect of Feed Point

One of the most important concepts from micromixing theory is the importance of addition position for selectivity in competitive-consecutive homogeneous reactions. It has been shown that there is a wide range of turbulent energy dissipation rates in a stirred tank.[4] The effect of position on mixing selectivity has been shown by a number of researchers and several methods have been used to demonstrate this effect. Nienow and Inoue give a set of examples using a small tank and the semibatch barium sulfate method of Villermaux to demonstrate the importance of feed position, as shown in Fig. 3.[18] The conversion to undesired by-product varies from 83% to 0.6%, depending on the feed location. Many other experimental measurements of yield and selectivity as a function of feed location have borne out this result.[16,19–23]

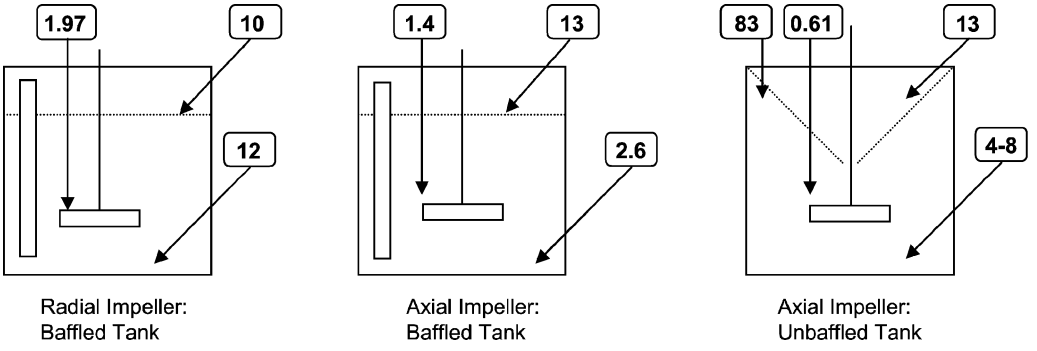


Fig. 3 Impact of different feed positions on the precipitation of barium sulfate. The selectivity to by-product as a percent of reactant is shown for feed into zones of high and low turbulent energy dissipation. The impeller speed and reactant addition time were held constant. More by-product is formed at feed points where the local mixing is slow.

Effect of Feed Injection Velocity

The importance of feed rate on yield for a mixing-sensitive reaction was demonstrated in Ref.[24]. The addition time in a semibatch reaction is often increased on scale-up because of heat transfer limitations. In the case of a mixing-sensitive reaction, the time of addition is increased on scale-up to compensate for the increase in blend time and to maintain the expected molar ratio at the feed point. The minimum feed time to achieve the expected yield is, therefore, scale dependent. Feed times that are too short will result in mesomixing conditions and reduced yield.

At feed times greater than the minimum, the yield becomes independent of feed time. The minimum feed time is a function primarily of the rate constant of the primary reaction and the mixing intensity at the point of feed introduction.

When feed velocities are low, backmixing into reactor feed pipes can lower yield by causing a slower overall mixing rate of the reactants. Jo et al. have shown that feed pipe backmixing can have a significant effect on yield, and they developed recommendations for V_f/V_{tip}, the feed pipe exit velocity divided by the impeller blade tip velocity.[25] Their results for turbulent flow in the feed pipe are summarized in Table 1. For laminar flow in the feed pipe, $V_f/V_{tip,min}$ was always lower than for turbulent flow. A good rule of thumb for turbulent flow is to design for $V_f/V_{tip} > 0.5$ except for Case 1, where $V_f/V_{tip} > 2$ is necessary.

Simple Guidelines for Scale-Up of Mixing-Sensitive Homogenous Reactions

Laboratory studies of mixing-sensitive homogenous reactions have been done by many investigators using the four reactions that are known as the ''Bourne Reactions.''[3] These studies and others fleshed out the previous theoretical predictions and established the mixing characteristics of stirred vessels in various configurations, as well as several types of in-line mixers. For fast reactions having timescales of seconds and tens of seconds, the concepts of micromixing and mesomixing can be reduced to a set of the following simple guidelines:

- Always add the ingredients to the point of highest turbulence. Avoid adding to the surface, a point of low turbulence.
- Scale up and scale down based on constant power per unit volume or mass. Even then there can be a loss of yield or by-product selectivity on scale-up.
- Consider using smaller reactors with higher energy dissipation rates such as in-line mixers in recirculation loops.
- Consider diluting the incoming reagents.
- Question the size of the reactor.
- If experiments show a possibility of mixing-sensitive reactions and the rate of addition is important, consider multiple point injections. The feed time will have to be increased in large-scale equipment.

MIXING AND TRANSPORT EFFECTS IN HETEROGENOUS CHEMICAL REACTORS

When chemical reactors have more than one phase, the problem increases in complexity because the reaction and mass transfer processes interact and another time constant is introduced. The interaction is governed by the relative rates of the reaction and mass transfer. In some cases, chemical reactions are mass transfer rate controlled (very fast chemical reactions) and in others, they are reaction kinetics controlled (very slow chemical reactions); however, in reality very few reactions strictly fit this classification. Thorough discussions of this problem are given in Refs.[26–28]. Classifications of the relative contributions of mass transfer and reaction kinetics in heterogeneous systems

Table 1 Recommended minimum V_f/V_{tip} for selected geometries for turbulent feed pipe conditions

Case	Impeller	Feed position	D/T	G/D	$V_f/V_{tip,min}$
1	6BD[a]	Radial/midplane[b]	0.53	0.1	1.9
2	6BD[a]	Above/near shaft[c]	0.53	0.55	0.25
3	HE-3[d]	Radial/midplane[b]	0.53	0.1	0.1
4	HE-3[d]	Above/near shaft[c]	0.53	0.55	0.15

[a]Six-blade disk turbine.
[b]Injection radially inward toward the impeller at its midplane at a distance G/D.
[c]Injection downward into the impeller at about $D/4$ from the centerline of the impeller shaft and G/D above the impeller midplane.
[d]High-efficiency three-blade down pumping turbine.

are outlined in Refs.[2,29]. These can be useful in an initial evaluation of reaction systems.

Homogenous reactions are more amenable to quantitative analysis and can, therefore, be developed more completely, as discussed earlier. The same local considerations apply for heterogeneous reactions, where the expected molar ratios between reactants cannot be maintained. In heterogeneous reactions, mass transfer limitations at phase boundaries, as well as local mixing limitations may affect the local concentration and reaction.

Homogeneous vs. Heterogeneous Selectivity

For simple reactions overall reaction rates may be affected and usually decrease, but yield is unaffected given equal degrees of conversion. For complex reactions, the selectivity may decrease but may also increase under certain circumstances. Significant selectivity effects can occur in heterogeneous systems at far lower absolute reaction rates because the mass transfer limitations can be very severe. These effects can be subjected to considerable magnification on scale-up to plant operations. This can be thought of as changing the k_{R2}/k_{R1} ratio, as measured by independent determination of the rate constants, to an apparent value caused by mass transfer limitations. Heterogeneous systems can also sometimes be manipulated to achieve improved yields, when compared to a homogenous system with the same reactions. There can be a great advantage to running under heterogeneous conditions or in some cases to deliberately creating a heterogeneous system for the purpose of improving selectivity.

The course of reactions is determined by events at the molecular scale, whether or not the reactive molecules are in the liquid, solid, or gas phase when they enter the reaction zone. As in the case of homogeneous reactions, the course of a complex reaction will be determined by local molar ratios and chemical kinetics. The degree of deviation from expected kinetic behavior is determined by the reaction rate relative to the rates of mass transfer and mixing. Possible mixing interactions in the films around heterogeneous phases, including dissolving solids, gasses, and liquid–liquid dispersions, are shown for illustration in Fig. 4. Differences in selectivity between the same reaction run under homogeneous and heterogeneous conditions are illustrated in the example below.

Competitive-Consecutive Reaction Example: Solid–Liquid Compared with Homogenous[4,30]

This example compares a reaction run using reagent addition as a dissolving solid and the same reagent added in solution. The two reactions were run in the

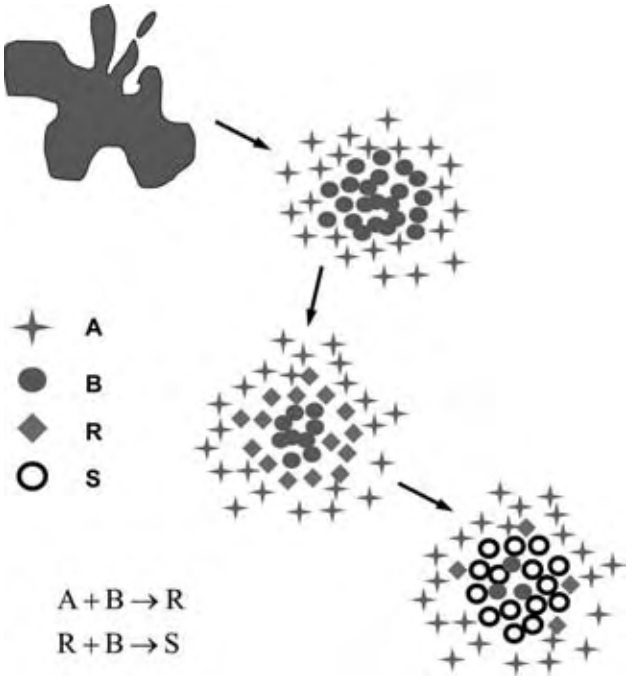


Fig. 4 Simultaneous mass transfer and reaction in the films around solid particles, gas bubbles, and liquid drops. For a heterogeneous competitive-consecutive reaction, mass transfer rates, reaction rates, and mixing rates can play a role. *(View this art in color at www.dekker.com.)*

same pilot-plant equipment with the same mixing conditions. The data for the product distribution in this consecutive–competitive reaction system allow direct comparison of product distributions obtained under homogenous and heterogeneous conditions. The reaction is a classical competitive-consecutive bromination to mono- and dibromo-substituted products, where the mono-substituted product is the desired product. Dibromo formation represents a yield loss both in this step and in the reaction steps to follow. The reaction is run in the semibatch mode in all cases. The dissolving reagent is *N*-bromo succinimide (NBS) and the reaction solvent is acetone. The pilot-plant conditions are shown in Table 2. The NBS is added over a 6-hr period because the reaction is highly exothermic. The actual reaction rate is not known but the addition requires 6 hr for heat removal. The impeller is a six-blade Rushton turbine.

Results from powder addition of NBS were satisfactory in the laboratory but in the pilot-plant, an increase in dibromo from 2% (Laboratory) to 8% (pilot-plant) was observed. This is an unacceptable increase in overreaction to dibromo. The apparent rate constant ratio, k_{R2}/k_{R1} for the two scales can be calculated from the product distributions [Eq. (1)]. This apparent ratio increased from 0.02 to 0.08 results in a decrease in selectivity.

Mixing effects in the film around the dissolving NBS are the obvious reason—the reaction rate is fast enough to allow significant reaction in the film before the dissolved NBS can be mixed to the molecular level. This indicates that the mass transfer rate is slower than the reaction rate. Possible solutions include: 1) reduce the particle size of the NBS by milling to reduce dissolution time and 2) eliminate mass transfer limitations by predissolving NBS in the reaction solvent and running as a homogeneous reaction. The latter option was run in the laboratory and was shown to reduce dibromo below that obtained with a powder addition ($< 1\%$). The same reduction was achieved in the pilot plant. In this example, the homogeneous reaction environment is more selective than the film around a dissolving reagent for a consecutive reaction.

Table 2 Pilot-plant conditions for bromination with NBS

Variable	Pilot-plant condition
Vessel volume (m^3)	0.75
Vessel diameter, T (m)	1
Impeller diameter, D (m)	0.4
Impeller speed, N (rpm)	175
Reaction volume H/T	0.5
Power/volume (W/m^3)	117
Local power (W/m^3) at point of solution addition	1170

Heterogeneous Reactions with Parallel Homogenous Reactions

The yield and selectivity of heterogeneous reactions can also be affected by mass transfer in extending the time for completion of a reaction during which a parallel reaction—possibly decomposition of A, B, or R—can be occurring in the bulk phase as well as in the films around the dispersed phase (or in the dispersed phase for liquid–liquid reactions). This problem can develop when the desired reaction rate can only be achieved at a temperature at which the starting materials, or any intermediate, or the product can react or decompose during the reaction time. This reaction time can be longer than expected on scale-up if the mass transfer rates do not duplicate those in the laboratory or in piloting. Shorter overall reaction times can also be realized on scale-up when mass transfer rates are increased by improved mixing, e.g., for liquid–liquid or gas–liquid dispersion. Reasons for slower mass transfer and extended reaction time for each type of contact are as follows:

Gas–liquid: Lower overall mass transfer rate because of insufficient gas dispersion–holdup and surface area.

Solid–liquid: Slower dissolution time because of variation in reagent particle size and mass transfer.

Liquid–liquid: Larger dispersed phase drop size and higher coalescence rates than expected.

All these factors are mixing dependent and can contribute to scale-up problems if mass transfer rates are not successfully reproduced.

Simple Guidelines for Scale-Up of Mixing-Sensitive Heterogeneous Reactions

The complex interaction between mass transfer and reaction kinetics requires determination of mixing sensitivity for virtually all heterogeneous reactions for which competitive and/or consecutive reactions are possible to be sure of successful scale-up. This requirement is prompted by issues of both: 1) overall reaction completion time and 2) undesired reactions in the films around the discontinuous phase(s). The following set of guidelines may be useful in evaluating mixing sensitivity and for scale-up.

Determine reaction by-product formation and true kinetics in the laboratory by running under dilute homogeneous conditions if possible.

Extend time of reaction beyond that required for completion of the primary reaction to determine whether or not the mass transfer rate must be duplicated on scale-up to prevent simultaneous homogeneous by-product formation and/or continuing consecutive reactions.

For solid–liquid reactions, determine whether the particle size or addition time affects yield. On scale-up provide sufficient power for off-bottom suspension.

For gas–liquid reactions, determine whether the gas–liquid mass transfer rate (a function of sparging and mixing) affects addition time and yield. On scale-up provide sufficient power and gas sparging to maintain interfacial surface area.

For liquid–liquid reactions, determine whether the drop-size distribution and/or additional time affects yield. On scale-up provide sufficient power and shear to generate the required drop-size distribution.

CONCLUSIONS

When local mixing rates are slower than the rate of reaction, but it is not feasible to perform complete reactor simulations, scale-up based on local mixing conditions is essential. For stirred reactors with multiple reactions where mixing affects the yield, the simplest approach is to hold the power per unit volume constant. This will only work if the feed locations are in the most turbulent location and strict geometric similarity is maintained. A more precise scale-up criterion is to hold the rate of turbulent energy dissipation per unit mass in the region of most intense mixing constant, and to feed at this location. This is particularly useful while feeding into the impeller stream of a stirred vessel. For geometrically similar mixing vessels, the local turbulence energy dissipation rate per unit mass is generally proportional to the overall power per unit volume, so the two criteria are essentially the same. In some cases, equal selectivity cannot be achieved on scale-up, even to a pilot-scale vessel. In this case, an in-line mixer may be considered. Results such as this are controlled by the local intensity of turbulence and the associated mixing field.

REFERENCES

1. Danckwerts, P.V. The effect of incomplete mixing on homogenous reactions. Chem. Eng. Sci. **1958**, *8*, 93–99.
2. Levenspiel, O. *Chemical Reaction Engineering*, 3rd Ed.; John Wiley and Sons: New York, 1998.
3. Baldyga, J.; Bourne, J.R. *Turbulent Mixing and Chemical Reactions*; John Wiley and Sons: Chichester, 1999.
4. Baldyga, J.; Bourne, J.R.; Paul, E.L.; Atiemo-Obeng, V.; Kresta, S., Eds. *Handbook of Industrial Mixing*; John Wiley & Sons: New York, 2004.
5. Muller, F.L.; Carpenter, K.J. On the role of dispersion and fast reactions in the development of processes. Recents Prog. Genie Precedes **1997**, *11* (51), 333–340.
6. Sharratt, P.N. *Handbook of Batch Process Design*; Blackie Academic & Professional: London, 1997.
7. Grenville, R.K.; Blending of Viscous Newtonian and Pseudo-Plastic Fluids. Ph.D. Thesis, Cranfield Institute of Technology, U.K., 1992.
8. Nienow, A.W. On impeller circulation and mixing effectiveness in the turbulent flow regime. Chem. Eng. Sci. **1997**, *52*, 2557–2565.
9. Fang, J.Z.; Lee, D.J. Micromixing efficiency in static mixer. Chem. Eng. Sci. **2001**, *56*, 3797–3802.
10. Johnson, B.K.; Prud'home, R.K. Chemical processing and micromixing in confined impinging jets. AIChE J. **2003**, *49* (9), 2264–2282.
11. Shearer, G.; Tzoganakis, C. Distributive mixing profiles for co-rotating twin screw extruders. Adv. Polym. Techol. **2001**, *20* (3), 169–190.
12. Thoma, S. Interactions Between Macro- and Micro-Mixing in Stirred Tank Reactors. Dissertation No. 9012, ETH, Zurich, 1989.
13. Bourne, J.R.; Thoma, S. Some factors determining the critical feed time of a semi-batch reactor. Chem. Eng. Res. Des. **1991**, *69*, 321–323.
14. Thoma, S.; Ranade, V.V.; Bourne, J.R. Interaction between macro- and micro mixing during reactions in stirred tanks. Can. J. Chem. Eng. **1991**, *69*, 1135–1141.
15. Patterson, G.K. Simulating turbulent-field mixers and reactors. In *Application of Turbulence Theory to Mixing Operations*; Brodkey, R.S., Ed.; Academic Press: New York, 1975; 223–275.
16. Bourne, J.R.; Yu, S. Investigation of micromixing in stirred tank reactors using parallel reactions. Ind. Eng. Chem. Res. **1994**, *33*, 41–55.
17. Baldyga, J.; Podgorska, W.; Pohorecki, R. Mixing-precipitation model with application to double feed semibatch precipitation. Chem. Eng. Sci. **1995**, *50*, 1281–1300.
18. Nienow, A.W.; Inoue, K. A study of precipitation micromixing, macromixing, size distribution, and morphology. Paper No. 9.4; CHISA: Prague, Czech Republic, 1993.
19. Paul, E.L.; Treybal, R.E. Mixing and product distribution for a liquid-phase, second-order, competitive-consecutive reaction. AIChE J. **1971**, *17*, 718–731.

20. Bourne, J.R.; Kozicki, F.; Moergeli, U.; Rys, P. Mixing and fast chemical reaction iii: model–experiment comparisons. Chem. Eng. Sci. **1981**, *36*, 1655–1663.
21. Bourne, J.R.; Del'Ava, P. Micro- and macromixing in stirred tank reactors of different sizes. Chem. Eng. Res. Des. **1987**, *65*, 180–186.
22. Baldyga, J.; Bourne, J.R. The effect of micromixing on parallel reactions. Chem. Eng. Sci. **1990**, *45*, 907–916.
23. Baldyga, J.; Bourne, J.R.; Hearn, S.J. Interaction between chemical reactions and mixing on various scales. Chem. Eng. Sci. **1997**, *52*, 457–466.
24. Baldyga, J.; Bourne, J.R. Interactions between mixing on various scales in stirred tank reactors. Chem. Eng. Sci. **1992**, *47*, 1839–1848.
25. Jo, M.C.; Penney, W.R.; Fasano, J.B. Backmixing into reactor feedpipes caused by turbulence in an agitated vessel. AIChE Sym, Ser. No. 299, **1994**, *90*, 41–49.
26. Cichy, P.T.; Russell, T.W.F. Two-phase reactor design tubular reactors—reactor model parameters. Ind. Eng. Chem. Res. **1969**, *61*, 15.
27. Schaftlein, R.W.; Russell, T.W.F. Two-phase reactor design, tank-type reactors. Ind. Eng. Chem. **1968**, *60*, 12–20.
28. Astarita, G. *Mass Transfer with Chemical Reaction*; Elsevier, 1967.
29. Doraiswamy, L.K.; Sharma, M.M. *Heterogeneous Reactions: Analysis, Examples, and Reactor Design*; John Wiley and Sons: New York, 1984.
30. Homsi, K.L.; Thompson, A.; Thien, M.P. A facile and selective large-scale bromination of aminotoluidine using N-bromosuccinimide/acetone solution addition. AIChE National Meeting, St. Louis, Nov 1993.

Molecular Bioengineering

Sundararajan V. Madihally
School of Chemical Engineering, Oklahoma State University, Stillwater, Oklahoma, U.S.A.

INTRODUCTION

Molecular bioengineering encompasses a broad area in which engineering principles are used to address problems at a molecular level, i.e., at the fundamental level of building blocks of life such as amino acids, proteins, carbohydrates, fatty acids, RNA, DNA, and their combinations. Although molecular bioengineering is still in its infancy, it is grounded on a solid scientific foundation attributed to the efforts of many disciplines. For example, the field of molecular biology provides ample information on the molecular events and their implications in health and disease. Significant advances in deciphering the human genome, culturing human embryonic stem cells (ESCs), and discovering the ''plastic'' nature of adult stem cells have opened a new window of opportunity for developing cellular therapies and also for generating functionally replaceable devices and/or tissue parts. Furthermore, advances in proteomic and genomic technologies have generated a driving force in high throughput screening and discovery of regulatory factors operating at the level of building blocks. Biochemistry provides a better understanding of the receptor–ligand interactions and enzyme–inhibitor interactions that have been studied using various experiments and modeling techniques. However, these concepts are not completely utilized to engineer efficient products and do not have related function and pathology to genomic sequence in biology and medicine.

The focus of molecular bioengineering is to utilize these interdisciplinary technologies to develop novel tools such as biosensors, efficient therapeutic molecules, and genetically modified products; design and develop efficient data analysis and data sharing techniques for a large pool of information typically obtained by proteomic and genomic technologies. Evaluating regulatory events that lead to a disease state in conjunction with biochemical alterations can lead to development of a highly efficient therapeutic molecule or development of a tool that can be used to diagnose a disease with high precision at the initial stages of the disease. Significant improvements in computational tools and information technologies have already played a greater role with the emergence of bioinformatics. To understand the output results, one has to develop sophisticated statistical tools that can help establish the network of information involved in the regulatory pathways controlling the activity of living organisms. In this entry, an overview of molecular bioengineering areas where significant activity and advancement have occurred is described by grouping them into four major categories as shown in Fig. 1. The first two categories are less defined and both engineering and fundamental sciences have to make simultaneous progress. On the other hand, the last two categories have a better understanding from the fundamental science perspective, but significant progress in engineering is still required.

TISSUE ENGINEERING

Tissue engineering has given promise for generating functionally replaceable tissue parts. The technology is based on using biodegradable porous scaffolds to guide and support the ingrowth of cells during tissue regeneration either at the site of grafting or in vitro. Here, the importance of biomimicry using molecular bioengineering is discussed.

Scaffolds are used with and without prior cell-seeded configurations based on the tissue to be replaced. In many applications where functionality of a tissue is critical to the survival of a patient, developing tissues outside the body or colonization of cells prior to transplantation is the preferred modus operandi. For example, the cardiovascular system poses unique challenges, as porous leaky materials cannot be grafted without cell colonization. Also, the absence of cellular components typically leads to blood clotting and overgrowth of cells that clog blood flow and lead to the subsequent failure of the graft. Another example is the liver cells, which are the metabolic hub of the body. If the liver has to be replaced, the replaced tissue should have the ability to provide the critical functions. To colonize the scaffolds, various molecular events have to be carefully programmed to regulate cellular bioactivity such as adhesion, spreading, migration, proliferation, and functionality. In addition, one has to identify an immunocompatible cell source in a quantity sufficient to support the functional load of the system. One practice is to retrieve autologous cells from a patient, grow them in vitro using tissue culture techniques, and transplant the cell-seeded scaffold back

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120027001

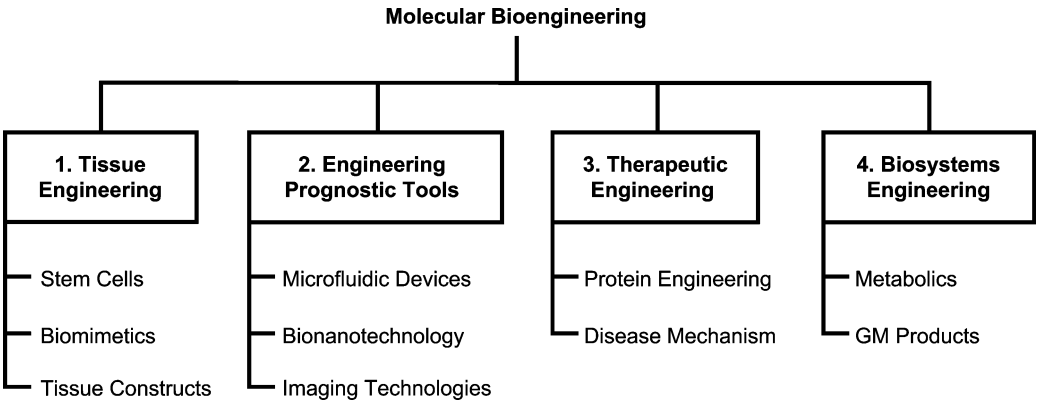


Fig. 1 Categories in molecular bioengineering.

to the patient. This procedure is restricted by the time taken in developing a graft and complications associated with the pathogenesis of a disease state. If allogeneic cells are isolated from a donor, the donor should be immunocompatible so that the graft rejection problem is abated. Hence, restrictions that limit the availability of transplantable donor tissue/organs also limit cell sources.

Stem cells

An alternative approach to broaden the availability of immunocompatible cells is to use stem cells that have the potential to differentiate into various cell types and also can be genetically manipulated because of their inherent self-renewal property. Thus, one could hope for the availability of functionally replaceable tissues ''off the shelf.''[1] Efficacy of using stem cells has been clinically proven beneficial in treating a variety of blood-related disorders. Furthermore, ''plasticity'' of stem cells has also been demonstrated in various animal models, suggesting that stem cells from one tissue type can be converted into cells of other tissues as well; for example, when bone marrow cells were transplanted into liver-injured animals, these cells restored the liver by differentiating into liver cells.[2] However, there is a need to engineer in vitro systems that can provide various microenvironments conducive to drive stem cells to defined cell types and then colonize them to generate needed tissue parts.

Exploiting the plasticity of stem cells in tissue regeneration will be the focus of many researchers in the coming years, as more concepts on the required elements and signaling mechanism(s) that regulate differentiation and proliferation of a variety of cell types are being understood. One of the activities that a number of investigators have focused on is evaluating the role of various soluble factors on the regulation of lineage development in culture systems. However, in addition to numerous soluble factors, dynamic interaction between stem cells and other cell types present in the microenvironment, membrane-bound factors, and polymeric matrix components that surround the stem cells directly regulate the process of differentiation and proliferation. Thus to recreate the microenvironment, coculturing stem cells with other cell types or incorporating molecules that significantly influence the cellular events into an artificial matrix has been the strategy. For example, an approach to incorporate bioregulatory activity is to use matrix molecules such as collagen (a protein) and glycosaminoglycans (GAGs, a polysaccharide) in scaffolds.[3] The latter are involved in diverse regulatory functions that include altering the bioactivity of mitogenic factor, colocalizing factors within a microenvironment (e.g., compartmentalization of the bone marrow microenvironment[4]) and directly regulating cell proliferation and differentiation (e.g., CD34 + cells).[5] Also, collagen/GAG-based skin equivalents are already in clinical use[6,7] and are under investigation in various tissue engineering applications. In addition, many other matrix elements such as fibronectin, vitronectin, laminin, fibrinogen, and denatured collagen are involved in orchestrating the dynamic events of cellular differentiation and proliferation and have to be carefully integrated into the matrix to program the differentiation and proliferation of stem cells so that transplantable tissues can be regenerated.

Biomimetics

While using individual components such as GAGs or collagen, a bigger engineering challenge is to process these materials into the required architecture without compromising their bioregulatory function. For example, an approach in using GAGs is to form an ionic complex with positively charged polymers, as ionically immobilized GAGs retain their bioactivity.[5,8] However, the ionic interactions are unstable and pH dependent, and the initial high content of GAGs can dissociate rapidly from the surface. To incorporate

bioregulation of other matrix elements, grafting a small peptide arginine–glycine–aspartic (RGD) acid onto polymers is an approach taken by many investigators.[9] Using RGD is based on the understanding that the majority of communication across the cell wall takes place via integrins, a family of transmembrane receptors, which communicate with many matrix elements through the RGD binding domain.[10] Nevertheless, availability of new molecules with specific biofunctions and ability to tailor their bioresponses to specific medical or biotechnological needs via protein engineering have added another driving force to biomaterials research.[11] The wide variety of ways by which such biomolecules and cells can be combined with polymeric biomaterials provides tremendous opportunities for tissue regeneration. Development of polymers that possess the capabilities of regulating the differentiation of stem cells to required mature cell types, new artificial biomimetic systems by exact placement of functional groups on polymer backbones, crosslinked structures, or macromolecular assemblies are the things to come.

Tissue constructs

Apart from replacing tissues in the body, developing tissue constructs outside the body will significantly help understand the progression of a pathological condition that is not clearly understood thus far. For example, development of solid tumors is aided by the growth of new blood vessels in a given location in a human body.[12] Although the signaling mechanism regulating the morphogenesis of endothelial cells that line the inner wall of blood vessels is an intensely investigated research area, the reason for the growth of new vessels is yet to be clearly understood. This is fundamentally limited by the lack of sensitive diagnostic tools that can detect the sprouting of new vessels in the body during the onset of a pathological condition in addition to the absence of reliable in vitro model systems. For example, routinely used magnetic resonance imaging and computer tomography techniques have 100–500 μm resolution, which is insufficient to carefully detect the initial stages of microvessels inside the body. On the contrary, confocal microscopy has a 0.1 μm resolution and can only be used in few-millimeters-thick samples. However, the sample size and the sample preparatory steps limit the usage of confocal microscopy in clinical applications. Needless to say, developing a reliable in vitro model that mimics various microenvironments will significantly help understand the molecular events that lead to sprouting of new vessels and the formation of solid tumors. Apart from solid tumors, understanding new blood vessel formation will also help comprehend its unequivocal role in a variety of patho/physiological processes including embryogenesis,[13] normal tissue development and function, atherosclerosis, diabetes, wound healing, and tissue engineering. Furthermore, developed tissue constructs can be used for toxicology studies and to test the effect of pathogens, i.e., as real time sensors for detecting biological agents. This would be highly effective in preventing infection altogether or alerting persons at risk about the need for immediate medical attention during bioterrorist attack and allergy seasons.

Engineering Prognostic Tools

Microfluidic devices

The microfabrication technology developed to create microelectronic devices has been adapted to create sophisticated, miniaturized analytical devices. Development of capabilities to miniaturize analytical devices and components offers a number of potential benefits including the ability to reduce sample sizes, development of low cost devices disposable after single-use devices, and improved device portability. Further, careful engineering of these devices will allow a wide variety of separations, chemical reactions, and detection steps that can also be carried out in an integrated fashion. Extensive work has been done on producing such micro-analytical systems on silicon or glass substrates using processes commonly employed in electronic chip manufacturing. A number of devices are currently under development for use in various laboratory techniques including bioseparations, microdialysis, and mass spectroscopy. Advances in the development of lab-on-chip devices shrink, and potentially simplify, laboratory tests like DNA analysis,[14] antigen detection and high-throughput genotyping.

Apart from developing miniature analytical tools, microfabrication technology can be utilized to develop portable biosensors or programmed drug delivery devices and also to understand the cell–cell or cell material interactions, which are useful for tissue engineering applications. While developing biosensors, one could utilize the cell-specific responses to obtain information for both pharmaceutical and chemical safety, and drug efficacy profiles in vitro as a screening tool.[15] Throughout their life cycle within specific tissues and organs, various cell types respond to a variety of chemical and physical signals by several means, i.e., either producing specific proteins or specific free radicals. Some of these events have been recreated using in vitro cell culture technique and molecular events are extensively cataloged for a few cell types. Knowing the desired output, a specific cell type is selected and cultured on microfabricated substrate to a desired pattern. In the case of electroanalytical

measurements and biosensing, electrode and biosensor devices can be miniaturized down to a cell-size scale and can be positioned directly at the vicinity of the cell surface, where cellular signaling substances are captured before they diffuse. Although quantitatively the signal is very small, it is enhanced through either circuit amplifiers or catalytic reactions on the biosensor.

Microfabrication technique can also be utilized, in conjunction with tissue engineering, to spatially control cell distribution in a cell culture, which can be used to understand the cell shape dynamics, cell division, and cell–cell interactions.[16] This information will be of significant use while engineering new materials for tissue regeneration and to develop biosensors of high sensitivity.

Bionanotechnology

The possibility of synthesizing nanoparticles, i.e., particles in the size range of 2–6 nm, is of considerable interest because of their unique size-dependent properties. Also, the particle size is dimensionally similar to that of biological macromolecules (e.g., nucleic acids and proteins).[17] These similarities could allow major advances in medical diagnostics, targeted therapeutics, molecular biology, and cell biology. Nanoparticles are among the first nanoscale materials to be directly useful in biology. Colloidal semiconductor nanocrystals called ''quantum dots'' (QDs) are the most attractive particles in biomedicine. They are generally composed of atoms of elements from groups II to VI or III to V of the periodic table. A normally used QD is zinc sulfide-capped cadmium selenide (CdSe-ZnS). Their size range of 2–10 nm or 10–50 atoms, often referred to as a size less than the Bohr radius in physics, leads to a quantum confinement effect dictated by the rules of quantum mechanics. This effect endows QDs with unique optical and electronic properties such as: a) exceptional photochemical stability; b) high photobleaching threshold; c) continuous absorption profiles; and d) readily tunable emission properties (from the UV to the IR), which allows simultaneous excitation of several particle sizes at a single wavelength; size-tunable narrow spectral line widths. For example, in comparison with commonly used organic dye rhodamine, the QD luminescent label is 20 times brighter, 100 times stable against photobleaching, and one-third as wide in spectral line width. In addition, the large surface-area-to-volume ratios of QDs make them appealing for the design of targeted molecular probes.

Imaging technologies

Targeted molecular probes are very useful for diagnostic purposes or as drug delivery systems and antigen/antibody immunoassays. Present imaging technologies rely mostly on nonspecific macroscopic physiological or metabolic changes that differentiate pathological tissues from normal tissue rather than identifying specific molecular events responsible for disease.

Specificity in targeting could be obtained by conjugating the surface with small molecules, such as receptor ligands or enzyme substrates, or higher-molecular-weight affinity ligands, such as monoclonal antibodies or recombinant proteins. While using these particles, the main concern is the fate of these particles when administered into the body. All foreign particles once injected into the bloodstream are detected by the immune surveillance system, which leads to adsorption of biological elements, especially circulating plasma proteins. This process is known as opsonization and is critical in dictating the fate of the injected particles. Normally opsonization renders the particles recognizable to the body′s major defense system, i.e., recognizable to the immune cells as a foreign material, and eliminates the particle. However, if the particles are large in size, then they are typically removed by the liver. Knowing immune surveillance mechanism(s), suitable surface modifications can be attempted to render the particles to remain in the blood circulation for a longer time or be directed to sites of interest. For example, uncoated poly(lactide-co-glycolide) (PLGA) particles are cleared rapidly, while particles coated with a block copolymer of polylactide-poly(ethylene glycol) (PLA-PEG) remained in the blood stream for a longer period of time.[18] This is attributed to rendered surface hydrophilicity and sterical stabilization of the particle. Recently, targeting specific tissues have been accomplished by tagging specific peptides to QDs and imaging the tissue;[19] because QDs are endowed with optical properties, the targeted tissue can be probed with increased sensitivity and for an extended period of time. Although producing stable QD-biomolecule complexes with clearly defined characteristics has been a challenge, advances in molecular engineering will allow the development of these targeted QDs of various diseases, and will potentially revolutionize medical treatment.

Therapeutic Engineering

Protein engineering

Evaluating molecular mechanisms is critical to understand the onset of a disease state. However, developing an effective and efficient therapy with minimum side effects using these concepts is more important. A classic example for this scenario is the understanding of diabetes type I (insulin-dependent diabetes). The fact that it is caused by the abnormal secretion of

insulin has been known for many years. Although significant improvement in insulin therapy has been made, it is still far from mimicking the physiological secretory profile of pancreatic β-cells. Many years of knowledge on the fate of insulin in the body, such as the need for dissociation of the insulin hexamer for insulin absorption and interaction with its receptor, has resulted in biosynthetic insulin (rapid-acting, intermediate-acting, and long-acting), produced by recombinant DNA techniques. However, the pharmacokinetic profiles of biosynthetic insulin do not adequately reproduce the physiological postprandial insulin response. This has led to the development of molecular analogs with slight modifications that prevent the spontaneous polymerization underlying delayed absorption.[20] Lispro is one such rapidly absorbed and short-acting insulin analog in which the amino acids at positions 28 (lysine) and 29 (proline) of the B chain are interchanged and has recently gained more attention. In the years to come, one could imagine having a controlled delivery device that could regulate the insulin dosage by sensing the level of glucose in the blood stream.[21] This concept not only decreases the number of doses a patient has to take but also improves the efficacy and alleviates the problem of overdosing. Also, administering insulin in other pathological states could prove beneficial because of its involvement in diverse biological processes such as fatty acid metabolism, protein synthesis, and protein degradation.[22]

Apart from insulin, engineering proteins has also advanced the antibody-based therapeutic approach, based on the cascade of interactions that lead to a disease state. Antibodies bind to antigens such as proteins, glycoproteins, phospholipids, and glycolipids with highest specificity, and one has to consider the half-life of circulation of antibodies. The factors affecting the antibody stability have to be carefully assessed so that novel molecules can be engineered without altering their target-recognition specificity. In this regard, a number of antibodies have been conjugated with high molecular weight PEG, which result in antibodies that have significant improvement in their half-life while still retaining their molecular recognition. Recent developments in engineering chimeric (containing components from two species) and humanized antibodies as well as antibody fragments and bispecific (two specificities that can direct the effector cells to a required target site) antibodies have rejuvenated the antibody-based technologies. Even more excitingly, approaches to utilize bioreactors and genetically modified lower eukaryotic organisms, plants, and milk from animals have provided alternative strategies for large-scale production of antibodies, overcoming many earlier clinical concerns. These developments have given new impetus in developing targeted immunotherapies for the treatment of asthma, autoimmune diseases, infection and inflammation, cancer, transplantation, poisoning, substance abuse, and other diseases. Antibody engineering has begun to show concrete success in its long-term efforts to develop targeted immunotherapies. Advances in phage display antibody libraries can now help to define novel gene function and the measurement of abnormal protein expression in pathological states.

Disease mechanism

Molecular events leading to many other pathological states are yet to be clearly understood. For example, more than seven decades ago, Cuthbertson[23] reported that severe trauma results in a generalized hypercatabolic state characteristically marked by severe skeletal muscle loss; recently, similar muscle losses have been observed in cancer and AIDS patients. Protein loss could not be explained by disuse alone; moreover, effects were observed at sites distal from the locus of injury. The net catabolic effect is attributed to imbalanced protein synthesis vs. degradation. Sustained loss of proteins is termed as ''muscle wasting'' and results in diminished muscle strength, with secondary complications including prolonged mechanical respiration, increased risk of pneumonia, poor wound healing, long-term hospitalization, and extended rehabilitation. Apart from muscle wasting, severe trauma also leads to depressed immune responses and multiorgan dysfunction syndrome, which together constitute the major cause of morbidity and mortality in post-traumatic patients. Based on our current understanding of mechanisms underlying muscle wasting and immune function, therapies to block proteolysis and/or augment immunity (e.g., cytokines) have been attempted. However, these have been largely ineffective, suggesting a need for better understanding of the pathways mediating the dysfunctional hypercatabolic state.[24] The underlying mechanisms also remain incompletely understood because of technical problems associated with whole muscle preparations and the lack of clinical data to support the mechanistic hypotheses. Thus, there is a need for models or in vitro tissue engineered constructs that mimic the pathological states. Recent advances in proteomics and genomics will significantly help in the understanding of the cascade of molecular events leading to a pathological state and subsequent development of target-based therapies.

Biosystems Engineering

Using the concepts of microbiology and plant biology, engineering micro-organisms and food crops has made substantial progress.

Metabolic engineering

Utilizing the biochemical understanding of metabolic pathways of prokaryotic and eukaryotic cells, pharmaceutical molecules such as aromatic amino acids that are difficult to process through the traditional chemical processing techniques can be synthesized.[25] For centuries, this concept has been utilized in the fermentation process to produce simpler ethanol from the various less-expensive, renewable carbon sources. Recent developments in metabolic engineering have shown that other molecules can be produced by diverting the competing reactions at the genetic level, i.e., by engineering the genetic sequence to increase selectivity of a biochemical reaction based on the required product.[26] In addition, producing secondary metabolites or small molecules important for pharmaceutical and materials applications is also possible.

For commercial feasibility, factors other than the metabolic pathways affecting the yield of the product synthesized by micro-organisms have to be analyzed. The higher sensitivity of the engineered cells to changes in environmental factors can cause cell death and decrease productivity. Thus, cells have to be made more robust to perturbations in environmental factors. Also, supplements used to grow the cells can become very expensive and hence the process has to be extended to include cost-effective substrates. Thus, apart from extending the process to cost-effective substrates, cells have to be made more robust to perturbations in environmental factors. Furthermore, efforts focusing on engineering cells by random selection methods have failed where genes are introduced based on conclusions derived in the absence of quantitative analysis of pathways or enzyme-dependent reactions carrying out various functions related to cell growth and product formation. Thus, to make a systematic approach to target selection for the improvement of productivity, methods of modeling the metabolic pathways using reaction kinetics and material balances are gaining more attention; if a genetic manipulation is made, how does it affect substrate utilization and, importantly, the selectivity of the required product. However, in any given scenario, completing the material balance is a challenging task because of the possibility of multiple product formation. With the emergence of proteomics and genomics, metabolic engineering is poised to succeed in the coming years and will play a significant role in the bioprocess industry.

Bioconversion can be thought of as an upstream integration of metabolic engineering, i.e., use micro-organisms to degrade renewable resources and then use the formed glucose or other substrates to synthesize required pharmaceutical products. Cellulose is a major building block of various crops. The potential for using cellulosic materials in bioconversions is well recognized, as the enzymatic hydrolysis provides an environmentally friendly means of depolymerizing cellulose. However, the high cost associated with the enzyme and also the very low yield obtained remains a major problem. A number of pretreatment strategies such as acid treatment, steaming, or peroxide treatment are being investigated to reduce the cost and also to increase the yield and is an upcoming area.

Genetically modified (GM) products

Another area in biosystems engineering focuses on developing crops, particularly fruit and vegetables, which are more resistant to handling or to infections. Pectin, the components constituting the skin of many fruits and vegetables, play a key role in plant physiology and plant pathology. The degree of methylation of these polysaccharides has been correlated with the extensibility of the cell walls.[27] Nevertheless, pectins are the first line of defense against infectious agents and also a target for genetic engineering to improve the shelf life of the crop. Thus, significant research has been focused on understanding how plant pathogens invade a fruit or a vegetable such as melons, apples, tomatoes, soybean, and maize so that one could develop a strategy that can prevent the invasion, thus leading to increased productivity. In the realm of plant–pathogen interactions, the interaction between plant polygalacturonase-inhibiting proteins (PGIPs) and pathogene-derived polygalacturonases (PG) is of significant interest as part of the plant's defense system, as it plays an important role in the prevention of the pathogenic micro-organism penetration in addition to pectic substance metabolism. Because the resistance of plant tissues to infection frequently correlates with PGIP expression and its inhibitory action on fungal PG, PGIP is a candidate for genetic engineering to obtain transgenic plants resistant to fungal infection or to impart extended shelf life to fruits. Further, PGIP is responsible for the specificity of cell–cell interactions during pollination or inoculation by fungi nonpathogenic for a particular plant. It is localized in plant cell walls and is encoded by a gene family whose expression is stimulated by injury or fungal infection.

To better understand the interactions, models describing the PG–PGIP interaction have to be developed that can be used to engineer better products.[28] However, the presence of a number of isoforms of the two interacting species and the alterations in the growth and development of fruits and vegetables have to be included in the models. Furthermore, with several classes of hydrolytic enzymes, the interaction of the enzyme with a specific inhibitor-protein results in a complex of two proteins that is inactive or of greatly depressed activity. Models can provide a better quantitative method to estimate the inhibitor-protein

activities and also generate the information on the important stage-specific alterations in the expression of inhibitor-proteins during growth and development. This could lead to minimal genetic manipulation of only the dominant inhibitory factors while engineering new products or to a better understanding of the process of evolution.

CONCLUSIONS

In summary, the emerging field of molecular bioengineering spans a wide spectrum of applications and will greatly expand the scope to tackle challenges in molecular medicine. A number of concepts will show significant progress in the coming years. Further advances in materials processing and micromachining will accelerate the development of more sophisticated diagnostic tools and therapies, such as imaging and antibody targeted chemotherapy.

ARTICLES OF FURTHER INTEREST

Tissue Engineering, p. 3115.

REFERENCES

1. Marshall, E. The business of stem cells. Science **2000**, *287* (5457), 1419–1421.
2. Petersen, B.E.; Bowen, W.C.; Patrene, K.D.; Mars, W.M.; Sullivan, A.K.; Murase, N.; Boggs, S.S.; Greenberger, J.S.; Goff, J.P. Bone marrow as a potential source of hepatic oval cells. Science **1999**, *284* (5417), 1168–1170.
3. Lindahl, U.; Kusche-Gullberg, M.; Kjellen, L. Regulated diversity of heparan sulfate. J. Biol. Chem. **1998**, *273* (39), 24979–24982.
4. Gupta, P.; McCarthy, J.B.; Verfaillie, C.M. Stromal fibroblast heparan sulfate is required for cytokine-mediated ex vivo maintenance of human long-term culture-initiating cells. Blood **1996**, *87* (8), 3229–3236.
5. Madihally, S.V.; Flake, A.W.; Matthew, H.W. Maintenance of CD34 expression during proliferation of CD34 + cord blood cells on glycosaminoglycan surfaces. Stem Cells **1999**, *17* (5), 295–305.
6. Orgill, D.P.; Straus, F.H., 2nd; Lee, R.C. The use of collagen-GAG membranes in reconstructive surgery. Ann. NY. Acad. Sci. **1999**, *888*, 233–248.
7. Yannas, I.V. Models of organ regeneration processes induced by templates. Ann. NY. Acad. Sci. **1997**, *831*, 280–293.
8. Chupa, J.M.; Foster, A.M.; Sumner, S.R.; Madihally, S.V.; Matthew, H.W. Vascular cell responses to polysaccharide materials: in vitro and in vivo evaluations. Biomaterials **2000**, *21* (22), 2315–2322.
9. Massia, S.P.; Hubbell, J.A. Covalently attached GRGD on polymer surfaces promotes biospecific adhesion of mammalian cells. Ann. NY. Acad. Sci. **1990**, *589*, 261–270.
10. Ruoslahti, E.; Engvall, E. Integrins and vascular extracellular matrix assembly. J. Clin. Invest. **1997**, *99* (6), 1149–1152.
11. Seliktar, D.; Zisch, A.H.; Lutolf, M.P.; Wrana, J.L.; Hubbell, J.A. MMP-2 sensitive, VEGF-bearing bioactive hydrogels for promotion of vascular healing. J. Biomed. Mater. Res. **2004**, *68A* (4), 704–716.
12. Folkman, J.; Shing, Y. Angiogenesis. J. Biol. Chem. **1992**, *267* (16), 10931–10934.
13. D'Amore, P.A.; Thompson, R.W. Mechanisms of angiogenesis. Annu. Rev. Physiol. **1987**, *49*, 453–464.
14. Hong, J.W.; Quake, S.R. Integrated nanoliter systems. Nat. Biotechnol. **2003**, *21* (10), 1179–1183.
15. Kamei, K.; Haruyama, T.; Mie, M.; Yanagida, Y.; Aizawa, M.; Kobatake, E. The construction of endothelial cellular biosensing system for the control of blood pressure drugs. Biosens. Bioelectron. **2004**, *19* (9), 1121–1124.
16. Folch, A.; Toner, M. Microengineering of cellular interactions. Annu. Rev. Biomed. Eng. **2000**, *2*, 227–256.
17. Whitesides, G.M. The ''right'' size in nanobiotechnology. Nat. Biotechnol. **2003**, *21* (10), 1161–1165.
18. Davis, S.S. Biomedical applications of nanotechnology—implications for drug targeting and gene therapy. Trends Biotechnol. **1997**, *15* (6), 217–224.
19. Akerman, M.E.; Chan, W.C.; Laakkonen, P.; Bhatia, S.N.; Ruoslahti, E. Nanocrystal targeting in vivo. Proc. Natl. Acad. Sci. U.S.A. **2002**, *99* (20), 12617–12621.
20. Holleman, F.; Hoekstra, J.B. Insulin lispro. N. Engl. J. Med. **1997**, *337* (3), 176–183.
21. Shichiri, M.; Sakakida, M.; Nishida, K.; Shimoda, S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial

endocrine pancreas. Artif. Organs **1998**, *22* (1), 32–42.
22. Solomon, V.; Madihally, S.; Mitchell, R.N.; Yarmush, M.; Toner, M. Antiproteolytic action of insulin in burn-injured rats. J. Surg. Res. **2002**, *105* (2), 234–242.
23. Cuthbertson, D.P. The disturbance of metabolism produced by bony and non-bony injury, with notes on certain abnormal conditions of bone. Biochem. J. **1930**, *24*, 1244–1263.
24. Madihally, S.V.; Toner, M.; Yarmush, M.L.; Mitchell, R.N. Interferon gamma modulates trauma-induced muscle wasting and immune dysfunction. Ann. Surg. **2002**, *236* (5), 649–657.
25. Buckland, B.C.; Robinson, D.K.; Chartrain, M. Biocatalysis for pharmaceuticals—status and prospects for a key technology. Metab. Eng. **2000**, *2* (1), 42–48.
26. Koffas, M.A.; Jung, G.Y.; Stephanopoulos, G. Engineering metabolism and product formation in *Corynebacterium glutamicum* by coordinated gene overexpression. Metab. Eng. **2003**, *5* (1), 32–41.
27. Darley, C.P.; Forrester, A.M.; McQueen-Mason, S.J. The molecular basis of plant cell wall extension. Plant Mol. Biol. **2001**, *47* (1–2), 179–195.
28. Fish, W.W.; Madihally, S.V. Modeling the inhibitor activity and relative binding affinities in enzyme-inhibitor-protein systems: application to developmental regulation in a PG–PGIP system. Biotechnol. Prog. **2004**, *20* (3), 721–727.

Molecular Modeling for Nonequilibrium Chemical Processes

M

Dionisios G. Vlachos
Department of Chemical Engineering and Center for Catalytic Science and Technology, University of Delaware, Newark, Delaware, U.S.A.

INTRODUCTION

Applied computational quantum chemistry, such as density functional theory (DFT), and molecular modeling, including molecular dynamics (MD) and Monte Carlo (MC) techniques, have become essential tools for science and engineering research. Molecular modeling has its roots in Newton's laws of motion and the Metropolis stochastic evolution of phase space. Yet, it is only the advances in computational power coupled with novel algorithmic and theoretical developments (e.g., the Gibbs ensemble, the histogram reweighting MC with finite-size scaling for critical exponent determination, the Swendsen–Wang algorithm, etc.) since the early 1980s that have rendered molecular modeling an indispensable tool in computing ''equilibrium structure'' of bulk matter, small clusters, and interfaces. Examples of its applications include liquid and solid reconstruction, protein folding, crystallization, adhesion of macromolecules (e.g., biopolymers) on surfaces, phase diagrams of adsorbed layers, and segregation in alloys.[1–4] The introduction of molecular modeling for nonequilibrium systems entailing transport and/or chemistry has naturally been slower (selected application examples in chemical engineering are given in Ref.[5]). For example, the first report of MC simulation of surface kinetics appeared in Ref.[6], but it was the work of Ziff et al.[7] a few years later that actually ''ignited'' the scientific community. Monte Carlo methods were introduced earlier to study epitaxial growth of films, such as the growth modes and the transition to rough crystals (an overview of earlier work is given in Refs.[8,9]).

The delay in using molecular modeling in chemical process industry stems from many challenges, including the inherent complexity of nonequilibrium processes, the lack of a rigorous nonequilibrium statistical mechanics theory, the lack of experimental techniques with nanometer spatial and short temporal resolution that can be confronted with molecular models, and the inherent multiscale complexity of industrial processes. Despite these obstacles, molecular modeling is nowadays clearly an integral part of multiscale and nanoscience research,[10–12] as it can handle complex many-body systems and provide the ''exact'' numerical solution for a given input. Molecular models provide unprecedented insights into complex problems, such as reaction mechanisms[13] and diffusion paths.[14] In addition, they are already assisting design of new materials and catalysts, e.g., Refs.[15,16]. Their integration into macroscopic scale models creates a new multiscale-modeling paradigm by enabling for the first time design of materials and processes with atomic resolution.[17] This article describes major uses of molecular modeling, focusing briefly on MD and mainly on kinetic MC (KMC) methods, and presents examples from the author's group on heterogeneous catalysis and microporous materials.

MOLECULAR DYNAMICS SIMULATION

Molecular dynamics is a true first principles dynamic molecular model. It simply solves the equations of motion. Given an intermolecular potential, MD provides the exact spatial and temporal evolution of the system. The stiffness caused by fast vibrations compared with slow molecular relaxations demands relatively small time steps and challenges current simulations.[3,4] As an example, the time scale associated with vibrations is a fraction of a picosecond, whereas those associated with diffusion or reaction may easily be in the seconds to hours range depending on the activation energy. Consequently, MD on a single processor is usually limited to short time and length scales (e.g., pico- to nanoseconds and 1–2 nm).

Molecular dynamics has been successful in revealing preferred adsorption sites within microporous materials; diffusion paths in microporous materials and on single crystals, e.g., Refs.[14,18]; calculation of sticking coefficients of small adsorbates, e.g., Ref.[19], etc. It is in the calculation of dynamics though, such as sticking coefficients and diffusion paths, where the real merit of MD lies. It works well when there is no large activation barrier compared to the thermal energy, i.e., a fluid-like behavior. The small reachable scales have limited direct comparison of MD simulation results with data from most-far-from-equilibrium systems,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120026492

such as heterogeneous catalysts and transport through microporous membranes. Use of MD data into the phenomenological Stefan–Maxwell equations has been one approach in reaching larger length and time scales for microporous materials.[20,21] For large activation energies, alternative molecular modeling tools, including transition state theory and reaction coordinate search along with statistical mechanics, have been invoked.[17,18,22] The spatiotemporal dynamics of largely activated processes can be modeled using KMC, which is discussed next. The hierarchy of multiscale tools is large to fully describe here, though a recent review attempts to do this.[23]

MONTE CARLO METHODS FOR PROCESSES WITH TRANSPORT OR CHEMISTRY

Monte Carlo is a stochastic method that in many cases circumvents, at least in part, the time/length problem of MD. The introduction of the Metropolis MC algorithm in 1953[24] established a computational statistical mechanics framework for computing equilibrium solutions. For equilibrium structure determination, MC is often preferred to MD. On the other hand, MD is preferred for dynamics and when collective behavior dominates. Starting from a description of the physical system in terms of a Hamiltonian, MC solves stochastically an underlying master equation using pseudorandom numbers by constructing the probability with which the various states of the system have to be weighted according to a Markov process.

It is often stated that MC methods lack real time and results are usually reported in MC events or steps. While this is immaterial as far as equilibrium is concerned, following real dynamics is essential for comparison to solutions of partial differential equations and/or experimental data. It turns out that MC simulations follow the stochastic dynamics of a master equation, and with appropriate parameterization of the transition probabilities per unit time, they provide continuous time information as well. For example, Gillespie has laid down the time foundations of MC for chemical reactions in a spatially homogeneous system.[25,26] His approach is easily extendable to arbitrarily complex computational systems when individual events have a prescribed transition probability per unit time,[27] and is often referred to as the *kinetic Monte Carlo or dynamic Monte Carlo* (*DMC*) method. The microscopic processes along with their corresponding transition probabilities per unit time can be obtained via either experiments such as field emission or fast scanning tunneling microscopy[28,29] or shorter time scale DFT/MD simulations discussed earlier. The creation of a database/lookup table of transition probabilities emerges as a powerful modeling approach in computational materials science and reaction arenas.[30,31]

The extension of Gillespie's algorithm to spatially distributed systems is straightforward. A lattice is used to represent binding sites of adsorbates, which correspond to local minima of the potential energy surface. The discrete nature of KMC coupled with possible separation of time scales of various processes could render KMC inefficient. The work of Bortz et al. on the *n-fold or continuous time MC* (*CTMC*) method[32] can lead to computational speedup of the KMC method, which, however, has been underutilized most probably because of its difficult implementation. This method classifies all atoms in a finite number of classes according to their transition probability. Probabilities are computed a priori and each event is successful, in contrast to the Metropolis method (and other null event algorithms) whose fraction of unsuccessful (null) events increases drastically at low temperatures and for stiff problems.[27,33] In conjunction with efficient search within a class and dynamic variation of atom coordinates,[34] the CPU time can be practically independent of lattice size.[27] After each event, the time is incremented by a continuous amount.

PARAMETERIZATION OF MOLECULAR MODELS

The input to a microscopic MD or MC simulation is an intermolecular potential, whereas that to a KMC simulation is the microscopic mechanisms along with their corresponding transition probabilities per unit time. This input is obviously critical for the success of molecular simulation. Trial and error is often involved when determining the parameters of this input. As an example, in the case of MD, a functional form of the intermolecular potential is initially assumed; MD simulations are carried out with a guessed set of parameters of the potential to compute an ensemble average property that is subsequently compared with that of a corresponding experiment. An iterative approach must then be employed whereby a different set of parameters is assumed, and MD simulations are repeated to improve the agreement with the data. This process is tedious and time consuming, and may lead to a poor parameterization because of multiple local minima. Determining the parameters of molecular models is actually an optimization problem. It differs somewhat from reverse engineering, discussed next, in that forward molecular simulations are performed, and the potential form in MD or the paths in KMC are assumed to be correct (only the parameters are unknown).

A goal in parameterization of any model is to carry out a minimum number of simulations while maximizing the information content of these simulations. Extensive literature exists on optimum design of experiments that meets this objective.[35,36] The overall idea is based on a surface response methodology where one carries out a limited number of appropriately chosen molecular simulations to develop an empirical, low degree polynomial as a function of important parameters; this polynomial is subsequently used to determine the parameters so that simulation results are in close agreement with experimental data. Rigorous determination of parameters follows by minimizing an objective (cost) function that describes the difference between experimental data and simulation predictions. The approach can be further accelerated via hierarchical multiscale modeling: the corresponding mean field model is first parameterized to zoom in the correct range of parameter space (a mean field model assumes a spatially homogeneous system and ignores correlations and thermal fluctuations[37]). With the parameters of the mean field model as an initial guess, training of the parameters of the molecular model follows. The hierarchical approach is systematic, and has recently been demonstrated in the case of surface reactions (CO oxidation on Pt) and zeolites for KMC simulation.[38,39] Next a brief introduction to zeolites is given followed by an example of such parameterization.

Zeolites

Zeolites are crystalline aluminosilicate solid materials consisting of networks with pores of molecular dimensions. Because of the periodicity of the solid network and the nanometer size of the pores, zeolites have an unprecedented resolution (atomic size and shape selectivity) in a number of industrial applications, including catalysis, adsorption, and ion exchange. Currently, zeolites are also being explored as hosts for growth of nanowires, quantum dots, and sensors.

An example of molecular model parameterization: Benzene in NaX zeolite

NaX is a Faujasite type zeolite used industrially as a cracking catalyst. Its structure is depicted in the inset of Fig. 1. Benzene in NaX zeolite has been modeled extensively[18] and found to bind strongly in a facially coordinated orientation above four Na cations (S_{II} sites) and weakly in four sites near Na(III′) cations (W sites).

Figs. 1 and 2 depict the result of hierarchical training of a KMC model of benzene in the NaX lattice by simultaneously fitting self-diffusivity data at different loadings θ (number of benzene molecules per cage) and adsorption isotherms, respectively.[39] The optimized

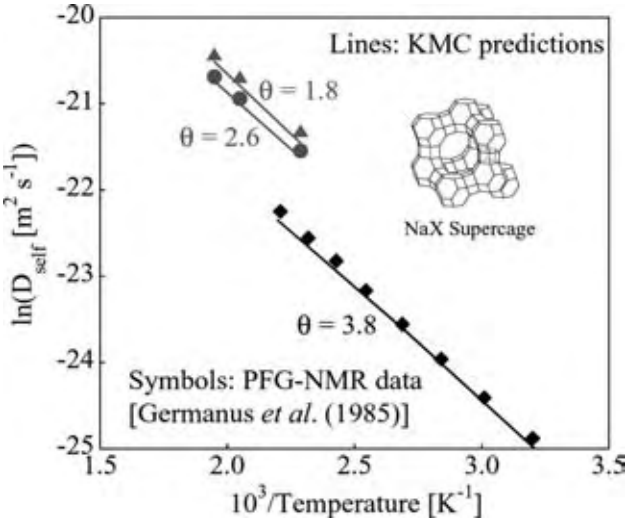


Fig. 1 Comparison of experimental and simulated self-diffusivities of benzene in NaX zeolite (supercage depicted in inset) vs. inverse temperature for three loadings, θ, indicated. *(View this art in color at www.dekker.com.)*

model behaves well even under extrapolation. Finally, the optimized parameters of the mean field and of KMC model differ by less than a factor of 2, underscoring the possibility for hierarchical multiscale modeling.

MOLECULAR MODELING AS A PREDICTIVE TOOL

Molecular modeling provides an unprecedented level of understanding. For most engineering applications,

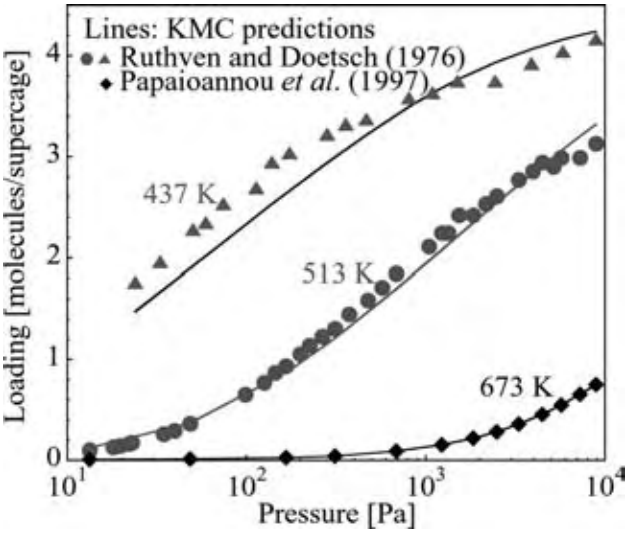


Fig. 2 Comparison of experimental and simulated isotherms of benzene in NaX zeolite vs. pressure for three temperatures indicated. *(View this art in color at www.dekker.com.)*

a spatially averaged quantity, e.g., a reaction rate or a flux, suffices, and this can easily be computed using molecular simulation output. However, this is not always a strict means for validating a molecular model (spatiotemporal data are a much more stringent requirement). Next we present two illustrative examples where KMC is used to make predictions and further discuss the validation issue.

Thin Zeolitic Microporous Membranes

Most technologies of microporous materials involve many simultaneous microscopic mechanisms (e.g., adsorption, desorption, migration, reaction, etc.) of molecular species. In particular, zeolites grown as oriented thin films or as membranes have received attention for their potential applications in high-resolution gas separations. Such systems are characterized by adsorption and desorption at the membrane boundaries and diffusion through the membrane interior. Armed with a KMC model trained with adsorption and diffusion data under equilibrium for the NaX zeolite/benzene system, the permeation for such membranes can be predicted (see inset in Fig. 3 for a schematic of the membrane). Fig. 3 shows an example of benzene loading profiles through a relatively thin membrane predicted via direct KMC. As expected, the strong adsorption sites (S_{II}) are nearly saturated whereas the weak adsorption sites (W) are nearly empty. Interesting nonmonotonic loading profile on W sites, shown in Fig. 3, arises mainly because of the difference in the density of benzene inside and outside the membrane.

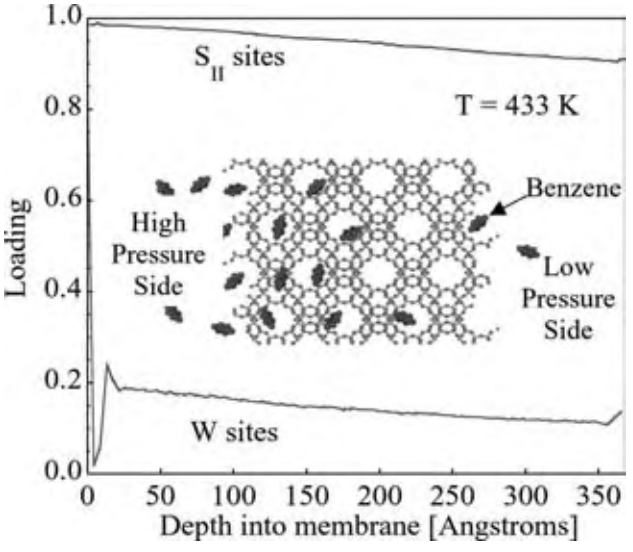


Fig. 3 One-dimensional loading profiles of benzene across a NaX zeolitic, single crystal membrane. The loading is the spatial average over planes perpendicular to the main diffusion direction (three-dimensional simulations are conducted; periodic boundary conditions are employed in the transverse direction and Robin at the membrane interfaces exposed to the high- and low-pressure sides). The inset shows a schematic of the membrane. *(View this art in color at www.dekker.com.)*

Simulations, such as the ones shown here, provide insights into loading profiles, which currently do not exist experimentally. At the same time, concentration profiles are a fingerprint of the diffusion mechanism[40] and could be used as an out-of-equilibrium test for stricter validation of a molecular model. Furthermore, molecular simulations enable one to predict flux and selectivity through membranes. These are spatiotemporal average quantities that can be compared to permeation experiments (such experiments are routinely conducted in several laboratories). More importantly, such simulations could be employed to understand how the microstructure and orientation of zeolite films affect separation properties and how important defects, e.g., grain boundaries, are for membrane performance. Such questions are presently unanswered. Yet, substantial improvements are possible if one were able to know what microstructures are best and subsequently how to make them.[41] Answers to such questions could provide insights into fabrication of systems with improved performance. Currently, real zeolitic membranes are polycrystalline and much thicker (1–30 μm) than the state-of-the-art simulated ones, and thus, comparison of KMC simulation to experiments is still infeasible. The above issues highlight the lack of strict out-of-equilibrium validation of molecular models and the limitation of KMC to reach realistic scales of experimental relevance.

Spatiotemporal Patterns on Catalyst Surfaces: The H_2/O_2 Reaction on Pt

Agreement of a spatially averaged quantity, such as a reaction rate and a temperature-programmed desorption spectrum, is often insufficient to test whether the underlying mechanisms and input in a molecular model are correct. Ideally, validation of molecular simulation predictions demands spatiotemporal data over a multitude of length and time scales. Unfortunately, such data are rarely available (see the membrane application example earlier). However, advances in scanning probe techniques start rendering such comparisons feasible.

We illustrate the above point with an example from the hydrogen oxidation on a single Pt(1 1 1) crystal. This system was recently studied experimentally by Ertl, Wintterlin, and coworkers, and motivated simulation because the corresponding continuum diffusion-reaction model was unable to capture some experimental features. Their experiments entailed predosing the

catalyst with oxygen and subsequently exposing the crystal to a hydrogen background.[42] Under certain conditions, a circular-looking wave propagates across the surface that is believed to emanate from an initial water cluster. The wave has a ring consisting of hydroxyl (OH), and as it travels outward into the pre-adsorbed oxygen pool, leaves behind mainly water.

Results from KMC simulations carried out with a detailed chemistry model were reported in Ref.[43]. The KMC simulation is able to capture qualitative experimental features without parameter adjustment. However, simulated length scales are slightly short, and finite-size effects probably affect the longer time evolution. Furthermore, the time scales predicted were longer than the experimental ones (dispersive NEXAFS data show that the difference in time scales is probably smaller[44]).

Sensitivity analysis (SA) could be an invaluable tool for understanding microscopic mechanisms, which unfortunately has been underutilized in molecular simulation in part because of its cost and the lack of system approach of the molecular simulation community.[23] We provide an SA illustration for the wave propagation problem. We have found that an increase in the diffusion coefficient of water leads to a reduction of predicted time scales, bringing simulation and experimental time scales closer. However, when one does not account for adsorbate–adsorbate interactions in diffusion, the increase of water diffusivity destroys the pattern, as shown in Fig. 4 (left panel). It is found that water–water and water–oxygen intermolecular forces via hydrogen bonding play a significant role in stabilizing the front and controlling time scales (Fig. 4, right panel). These simulations appear to be the first comparison of KMC with spatially distributed experimental data. More importantly, it appears possible to reveal the role of various molecular scale interactions in mesoscopic pattern formation.

REVERSE ENGINEERING USING MONTE CARLO METHODS

Recent advances in experimental techniques at the nanoscale and in high throughput experimentation (HTE) provide either information with unprecedented spatiotemporal resolution or massive data. Extraction of information from such data is also an optimization problem that is often referred to as ''reverse engineering.'' Here one bypasses the question of which are the actual kinetics giving rise to the experimental structures, i.e., how do these structures form and whether they are metastable or not, and simply focuses on determining these structures from data. Therefore, forward nonequilibrium KMC simulation is not carried out. Instead, MC is merely used as an efficient global minimum search (optimization) method [note that the term reverse MC (RMC) is also often employed] to determine structures consistent with experimental data.

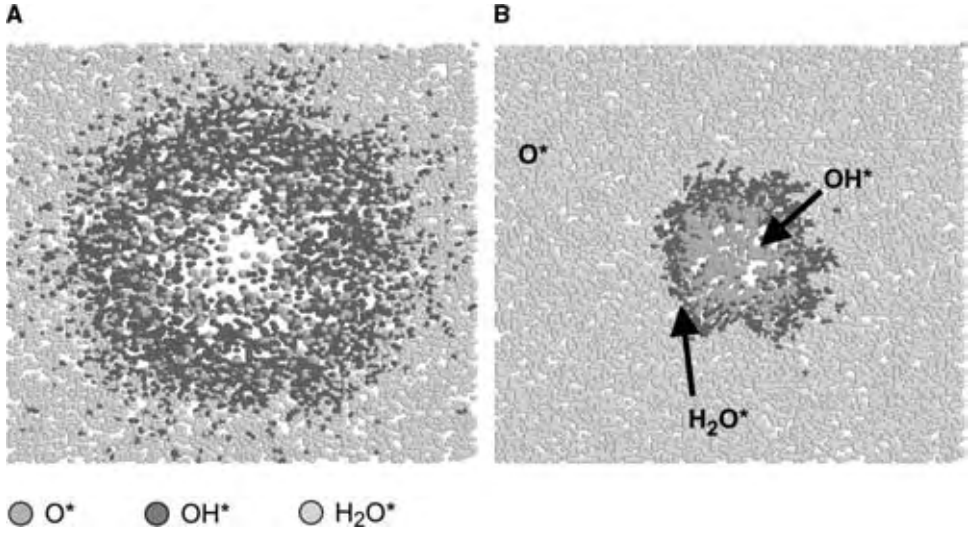


Fig. 4 Effect of intermolecular forces on wave propagation in the H_2/O_2 reaction on a single Pt crystal. The red dots represent OH, the gray adsorbed O, and the yellow water. For Fickian diffusion of water, i.e., in the absence of intermolecular forces, the wave propagates faster as the water diffusivity increases. However, the wave becomes ''diffuse'' (left graph) and eventually is lost (not shown) as realistic time scales are reached by increasing the diffusivity (A). Even weak attractive interactions between water molecules and between water and oxygen (e.g., 0.05 kcal/mol), owing to H bonding, are sufficient to keep the front sharp (right graph), even for the fast diffusivity used in the left simulation, so that the front is reminiscent of experimental data (B). *(View this art in color at www.dekker.com.)*

The overall idea of reverse engineering is simple. A ''cost'' function is formulated that represents the distance between experimental data and simulated results (this task parallels the molecular model parameterization discussed above). Multiple experimental points and multiple types of data can be added to the cost function. This is an important, highly recommended strategy because different types of experiments are usually sensitive to different parameters. Simultaneous use of multiple types of data constrains the system, resulting in a smaller subset of possible solutions. Given a cost function, one attempts random displacement of atoms or molecules in the system, which are accepted or rejected according to the Metropolis criterion.[24] According to the Metropolis criterion, the system can climb uphill in energy (in this case the cost function mimics the potential energy, and temperature has really no physical meaning). However, large barriers separating minima may prohibit the reach of global minimum in reasonable computational times.

The introduction of simulated annealing[45] has provided one of the best global optimization techniques and has become a widespread research tool. By slowly lowering the temperature from an initially high value, the system passes through quasiequilibrium states, and in principle the global minimum is obtained at low temperatures. Although this is undoubtedly a better approach compared with a single temperature MC simulation, possible large barriers still yield no guarantee that the global minimum is obtained. When the global minimum is desired, it is advisable that multiple runs starting from different initial configurations be exploited. Recent advances in tempering MC techniques provide an alternative approach.[46] A comparison of the efficiency of different MC search algorithms for HTE is given in Ref.[47]. Reverse engineering examples embody reconstruction of microporous carbon materials, nanoparticles of zeolites, porous media used as adsorption columns and fixed bed reactors, and design of libraries of materials, among many others.[47,48] Next we provide an example of reconstruction.

Structure Determination of Zeolite Nanoparticles in Hydrothermal Synthesis of Silicalite-1

One of the best-known zeolites is ZSM-5 whose lattice (MFI) is depicted in Fig. 5A. Its purely siliceous form is known as silicalite-1. Reverse engineering is used to determine the structure of silica nanoparticles formed during hydrothermal synthesis of silicalite-1 prior to any detectable crystallization of MFI zeolite particles. These nanoparticles have been detected prior to crystallization with scattering techniques and are believed

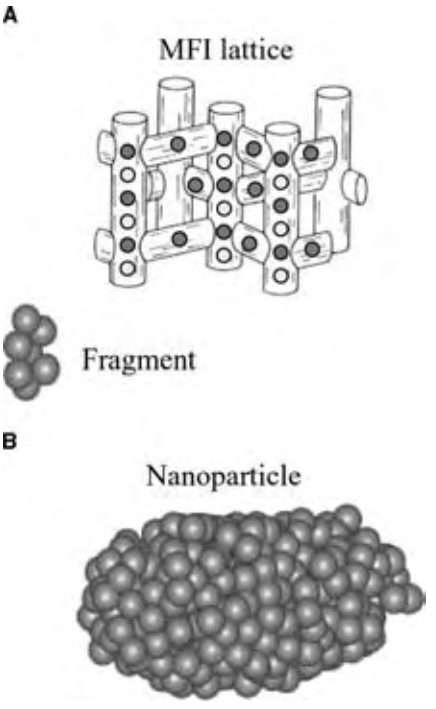


Fig. 5 Schematic of MFI lattice (A) and structure of siliceous nanoparticle determined by simultaneously fitting NMR and SANS data using simulated annealing (B). Small clusters (fragments) are often found to better fit the data. Each sphere in B represents a SiO unit. *(View this art in color at www.dekker.com.)*

to be the building growth units of zeolites.[49] Their size is usually in the order of 3–4 nm. Because of their apparent role in crystal growth of zeolites, an understanding of their structure is important.[50]

To determine the shape and structure of these nanoparticles, in situ ^{29}Si-NMR data have been obtained that reveal the local Si connectivity. Here Q^n ($n = 0, \ldots, 4$) denotes the coordination of a Si atom obtained from NMR, i.e., n is the number of Si atoms connected via an O bridge to a central Si atom. Knowledge of Q^n alone is insufficient to probe the long-range order, such as the size and shape of nanoparticles. In situ SANS data can provide information about the particle size, through the pair distance distribution function (PDDF). Using both NMR and SANS data within simulated annealing, possible structures have been computed (in preparation). Fig. 6 depicts how the simulated annealing works in this example, starting from a completely random distribution of ~400 SiO units. At high temperatures short-range order is built, manifested by the formation of small clusters (not shown), and the objective function decreases because the NMR data are fitted. Once a lower temperature is reached, longer-range order forms via an apparent Ostwald ripening type of mechanism, and the SANS

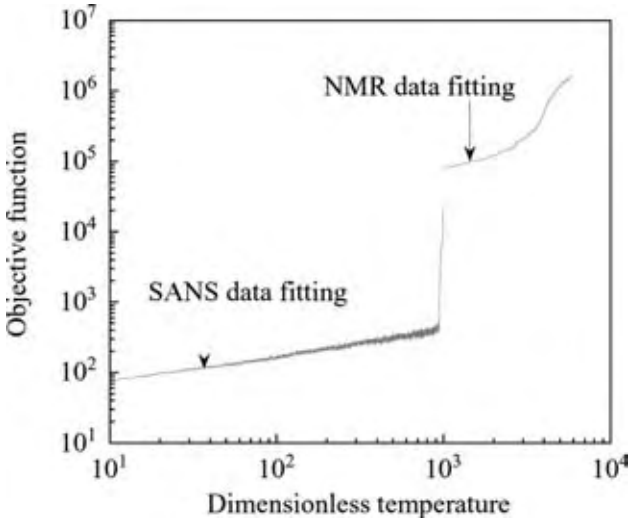


Fig. 6 Reduction of cost function with reducing temperature in simulated annealing calculations when determining the structure of siliceous nanoparticles by fitting NMR and SANS data. Short-range order occurs at higher temperatures and long-range order at lower temperatures. *(View this art in color at www.dekker.com.)*

data are fitted. An example of a possible structure is given in Fig. 5B (note that many structures can fit the experimental data). Fig. 7 and Table 1 show how well the PDDF/SANS and NMR data are captured, respectively. An important conclusion is that these nanoparticles are amorphous, elongated ellipsoids at

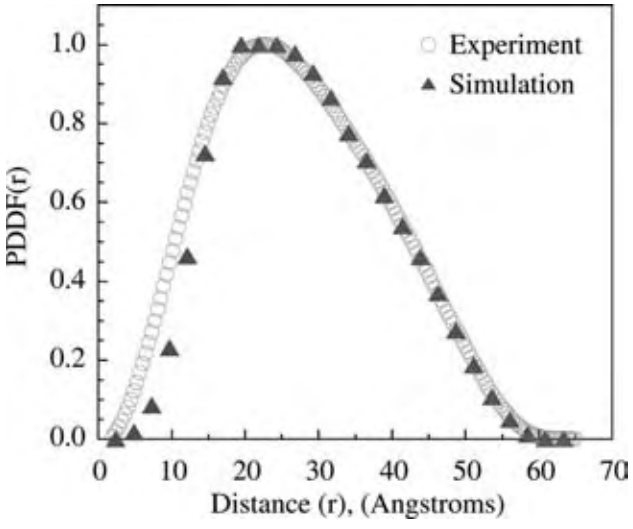


Fig. 7 Experimental and simulated PDDF of siliceous nanoparticles. The experimental PDDF is obtained via Fourier transform of SANS data and the simulated one from determining the nanoparticle structure (reverse engineering). The PDDF is fitted simultaneously with NMR data (see Table 1). *(View this art in color at www.dekker.com.)*

Table 1 Comparison of experimental NMR data and modeling results from determined structure of nanoparticles (see Fig. 5B for estimated structure)

	Q^0	Q^1	Q^2	Q^3	Q^4
Experimental	0	3	88	306	3
Model	0	3	87	307	3

The NMR data are fitted simultaneously with SANS data depicted in Fig. 7 according to the simulated annealing calculation shown in Fig. 6.

room temperature that do not possess three-member rings. This information is important in building growth models.

EXTENSIONS OF MOLECULAR MODELING TO LARGER SCALES

Molecular models can considerably impact the chemical process industry. Obviously, numerous problems fall beyond the realm of conventional molecular simulation (see the example above on zeolitic membranes). Examples include dynamics of protein folding, diffusion through microporous membranes and human cells, formation of quantum dots in heteroepitaxial growth of semiconductors, and pattern formation on catalyst surfaces.

Currently, coarse graining of conventional molecular models, including MD, Brownian dynamics, and MC methods, to larger scales (often termed mesoscales) is pursued very actively. Most attempts of coarse graining are empirical in nature. One approach entails coarse graining in time of MD and KMC[51,52] to partially overcome the time scale gap. In the case of spatially varying systems, the system, e.g., a peptide, a polymer molecule, or a catalyst surface, coarse graining in space is carried out by discretization into coarse cells (also termed as interacting sites or lumps), each containing a number of actual atoms or microscopic sites (see Fig. 8). The united atom approach, where the atomistic view of carbon and hydrogen in a methyl or a methylene is replaced with a single sphere, is a widespread example of such coarse graining. Recent models though go beyond the united atom approach and combine a number of methylene (and possibly other) groups into a single lump. Lumping alone is obviously not sufficient for coarse-graining. Coarse-grained rules, such as potentials and transition probabilities, must also be derived. Currently these are fitted or reverse engineered so that the coarse simulation provides an accurate description of some microscopic simulation property, such as the pair distribution function (see for example Ref. [53]). Such spatial coarse graining can result in tremendous CPU savings because of removing degrees of freedom, i.e., fast relaxations, and

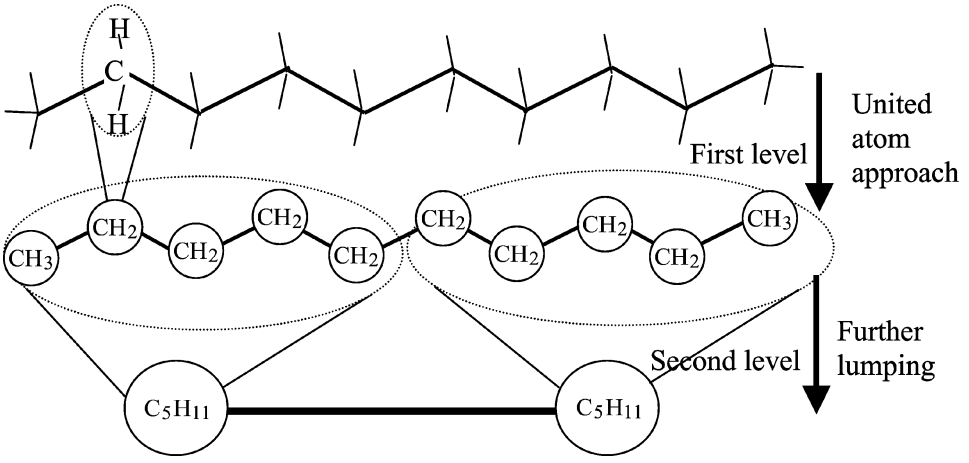


Fig. 8 Schematic of coarse graining of an *n*-decane molecule. In the first level, the united atom approach lumps hydrogen and carbons into methyl and methylene groups (spheres). In the second level, lumping of five groups into two supergroups (lumps or coarse cells) is depicted.

by reducing potential function evaluations (see grains in CPU in MD[53] and KMC[54,55]).

Several problems need to be overcome to render mesoscopic modeling an integral part of chemical process industry. One is the selection of the best size of coarse cells. Another is that the coarse rules are condition specific, i.e., they depend on density, temperature, etc. As a result, the coarse parameters need to be re-estimated for every new set of conditions. This is time consuming and tedious. These problems stem from the fact that first attempts of coarse graining are empirical in nature and lack a mathematical foundation that determines the factors controlling errors in coarse graining. Recent theoretical work on coarse graining of KMC[54–56] evidenced great promise to overcome the aforementioned problems. Here one derives from the microscopic mechanisms and rules, through nonequilibrium statistical mechanics, the coarse rules without fitting them to microscopic simulation results. As a result, reparameterization of the coarse rules is unnecessary. Furthermore, error estimates can be developed to guide optimum selection of coarse cell size. Also, microscopic cells can be used wherever there is a ''high activity,'' such as a chemical reaction, whereas coarse cells could be used far from such activity. These developments are at research level stages and it will take some time before they become an integral part of the chemical process industry.

CONCLUSIONS

Molecular modeling for nonequilibrium processes, such as heterogeneous catalysis and growth of advanced materials, is a relatively new tool in the chemical process industry. The advent of algorithms and advances in computer power on the one hand, and the development of coarse graining and multiscale techniques more generally on the other offer great promise for overcoming current limitations of molecular simulation and transform it into a design, predictive tool.

REFERENCES

1. Binder, K., Ed.; *Monte Carlo Methods in Statistical Physics*; Springer-Verlag: Berlin, 1986; Vol. 7.
2. Binder, K. Atomistic modeling of materials properties by Monte-Carlo simulation. Adv. Mater **1992**, *4*, 540–547.
3. Allen, M.P.; Tildesley, D.J. *Computer Simulation of Liquids*; Oxford Science Publications: Oxford, 1989.
4. Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*; Academic Press: New York, 1996.
5. Chakraborty, A.K., Ed.; *Molecular Modeling and Theory in Chemical Engineering*; Academic Press: New York, 2001; Vol. 28.
6. Wicke, E.; Kunmann, P.; Keil, W.; Schiefler, J. Unstable and oscillatory behavior in heterogeneous catalysis. Berichte der Bunsen-Gesellschaft-Phys. Chem. Chem. Phys. **1980**, *84*, 315–323.
7. Ziff, R.M.; Gulari, E.; Barshad, Y. Kinetic phase transitions in an irreversible surface-reaction model. Phys. Rev. Lett. **1986**, *56*, 2553–2556.
8. Gilmer, G. Computer models of crystal growth. Science **1980**, *208*, 355–363.
9. Muller-Krumbhaar, H. Kinetics of crystal growth. In *Current Topics in Materials Science*; Kaldis, E., Ed.; North-Holland, 1978; 1–46.
10. Thompson, T.B. Chemical industry of the future: technology roadmap for computational chemistry; DOE Workshop Roadmap for Computational Chemistry, 1999; http://itri.loyola.edu/molmodel.

11. Cummings, P.T.; Peterson, B.K.; Hall, C.K.; Neurock, M.; Panagiotopoulos, A.Z.; Westmoreland, P.R. Future directions in molecular modeling and simulation: fundamentals and applications. NSF Workshop Report1997; http://flory.engr.utk.edu/nsf.
12. Siegel, R.W.; Hu, E.; Roco, M.C. *Nanostructure Science and Technology: A Worldwide Study*; Kluwer Academic Publishers, 1999. Prepared by National Science and Technology Council (NSTC), Committee on Technology, The Interagency Working Group on NanoScience, Engineering and Technology (IWGN): WTEC, Loyola College in Maryland (http://itri.loyola.edu/nano/final/).
13. Norskov, J.K.; Bligard, T.; Logadottir, A.; Bahn, S.; Hansen, L.B.; Bollinger, M.; Bengaard, H.; Hammer, B.; Sljivancanin, Z.; Mavrikakis, M.; Xu, Y.; Dahl, S.; Jacobsen, C.J.H. Universality in heterogeneous catalysis. J. Catal. **2002**, *209*, 275–278.
14. Huang, D.Y.; Chen, Y.; Fichthorn, K.A. A molecular-dynamics simulations study of the adsorption and diffusion dynamics of short *n*-alkanes on Pt(1 1 1). Chem. Phys. **1994**, *101*, 11021–11030.
15. Jacobsen, C.J.H.; Dahl, S.; Clausen, B.S.; Bahn, S.; Logadottir, A.; Nørskov, J.K. Catalyst design by interpolation in the periodic table: bimetallic ammonia synthesis catalysts. J. Am. Chem. Soc. **2001**, *123*, 8404–8405.
16. Jóhannesson, G.H.; Bligaard, T.; Ruban, A.V.; Skriver, H.L.; Jacobsen, K.W.; Nørskov, J.K. Combined electronic structure and evolutionary search approach to materials design. Phys. Rev. Lett. **2002**, *88*, 255501–255506.
17. Raimondeau, S.; Vlachos, D.G. Recent developments on multiscale, hierarchical modeling of chemical reactors. Chem. Eng. J **2002**, *90*, 3–23.
18. Auerbach, S.M. Theory and simulation of jump dynamics, diffusion and phase equilibrium in nanopores. Int. Rev. Phys. Chem. **2000**, *19*, 155–198.
19. McCormack, D.A.; Kroes, G.-J. Accuracy of trajectory methods for activated adsorption of H_2 on Cu(1 0 0). Chem. Phys. Lett. **1998**, *296*, 515–520.
20. Skoulidas, A.I.; Bowen, T.C.; Doelling, C.M.; Falconer, J.L.; Noble, R.D.; Sholl, D.S. Comparing atomistic simulations and experimental measurements for CH_4/CF_4 mixture permeation through silicalite membranes. J. Membrane Sci. **2003**, *227*, 123–136.
21. Chempath, S.; Krishna, R.; Snurr, R.Q. Nonequilibrium molecular dynamics simulations of diffusion of binary mixtures containing short *n*-alkanes in faujasite. J. Phys. Chem **2004**, *108*, 13481–13491.
22. Hammer, B.; Nielsen, O.H.; Mortensen, J.J.; Bengtsson, L.; Hansen, L.B.; Madsen, A.C.E.; Morikawa, Y.; Bligaard, T.; Christensen, A. DACAPO Version 2.7; CAMP, Technical University, Denmark, 2003.
23. Vlachos, D.G. A review of multiscale analysis: examples from systems biology, materials engineering, and other fluid-surface interacting systems. Adv. Chem. Eng. **2005**, *in press*.
24. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of state calculations by fast computing machines. J. Chem. Phys. **1953**, *21*, 1087–1092.
25. Gillespie, D.T. A general method for numerically simulating the stochastic evolution of coupled chemical reactions. J. Comp. Phys. **1976**, *22*, 403–434.
26. Gillespie, D.T. Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. **1977**, *81*, 2340–2361.
27. Reese, J.S.; Raimondeau, S.; Vlachos, D.G. Monte Carlo algorithms for complex surface reaction mechanisms: efficiency and accuracy. J. Comp. Phys. **2001**, *173*, 302–321.
28. Gomer, R. Diffusion of adsorbates on metal surfaces. Rep. Prog. Phys. **1990**, *53*, 917–1002.
29. Renisch, S.; Schuster, R.; Wintterlin, J.; Ertl, G. Dynamics of adatom motion under the influence of mutual interactions: O/Ru(0 0 0 1). Phys. Rev. Lett. **1999**, *82*, 3839–3842.
30. Maroudas, D. Modeling of radical-surface interactions in the plasma-enhanced chemical vapor deposition of silicon thin films. In *Molecular Modeling and Theory in Chemical Engineering*; Chakraborty, A.K., Ed.; Academic Press: New York, 2001; 252–296.
31. Raimondeau, S.; Aghalayam, P.; Vlachos, D.G.; Katsoulakis, M. Bridging the gap of multiple scales: from microscopic to mesoscopic to macroscopic models. In *Foundation of molecular Modeling and Simulation*; AIChE Symposium Series No. 325; 2001; Vol. 97, 155–158.
32. Bortz, A.B.; Kalos, M.H.; Lebowitz, J.L. A new algorithm for Monte Carlo simulations of Ising spin systems. J. Comp. Phys. **1975**, *17*, 10–18.
33. Vlachos, D.G.; Schmidt, L.D.; Aris, R. Kinetics of faceting of crystals in growth, etching, and equilibrium. Phys. Rev. B **1993**, *47*, 4896–4909.
34. Nicholson, D. Grand ensemble Monte Carlo. CCP5 Quart. **1984**, *11*, 19–24.
35. Box, G.E.P.; Hunter, W.G.; Hunter, J.S. *Statistics for Experiments. An Introduction to Design, Data Analysis, and Model Building*; Wiley: New York, 1978.
36. Box, G.E.P.; Draper, N.R. *Empirical Model-Building and Response Surfaces*; Wiley: New York, 1987.

37. Hill, T.L. *An Introduction to Statistical Thermodynamics*; Dover: New York, 1986.
38. Raimondeau, S.; Aghalayam, P.; Mhadeshwar, A.B.; Vlachos, D.G. Parameter optimization in molecular models: application to surface kinetics. Ind. Eng. Chem. Res. **2003**, *42*, 1174–1183.
39. Snyder, M.A.; Vlachos, D.G. Rational, hierarchical parameterization of complex zeolite-guest molecular models. Mol.Sim. **2004**, *30*, 561–577.
40. Vlachos, D.G.; Katsoulakis, M.A. Derivation and validation of mesoscopic theories for diffusion of interacting molecules. Phys. Rev. Lett. **2000**, *85*, 3898–3901.
41. Lai, Z.; Bonilla, G.; Diaz-Carretero, I.; Sujaoti, K.; Amat, M.A.; Kokkoli, E.; Terasaki, O.; Thompson, R.W.; Tsapatsis, M.; Vlachos, D.G. Microstructural optimization of a zeolite membrane for organic vapor separation. Science **2003**, *300*, 456–460.
42. Sachs, C.; Hildebrand, M.; Volkening, S.; Wintterlin, J.; Ertl, G. Spatiotemporal self-organization in a surface reaction: from the atomic to the mesoscopic scale. Science **2001**, *293*, 1635–1638.
43. Raimondeau, S.; Vlachos, D.G. Front propagation at low temperatures and multiscale modeling for the catalytic combustion of H_2 on Pt. Chem. Eng. Sci. **2003**, *58*, 657–663.
44. Nagasaka, M.; Kondoh, H.; Amemiya, K.; Nambu, A.; Nakai, I.; Shimada, T.; Ohta, T. Water formation reaction on Pt(1 1 1): near edge x-ray absorption fine structure experiments and kinetic Monte Carlo simulations. J. Chem. Phys. **2003**, *119*, 9233–9241.
45. Kirkpatrick, S.; Gelatt, C.D.J.; Vecchi, M.P. Optimization by simulated annealing. Science, **1983**, *220*, 671–680.
46. Yan, Q.; De Pablo, J.J. Hyperparallel tempering Monte Carlo and its applications. In *Molecular Modeling and Theory in Chemical Engineering*; Chakraborty, A.K., Ed.; Academic Press: New York, 2001; 1–20.
47. Deem, M.W. A statistical mechanical approach to combinatorial chemistry. In *Molecular Modeling and Theory in Chemical Engineering*; Chakraborty, A.K., Ed.; Academic Press: New York, 2001; 81–121.
48. Lastoskie, C.M.; Gubbins, K.E. Characterization of porous materials using molecular theory and simulation. In *Molecular Modeling and Theory in Chemical Engineering*; Chakraborty, A.K., Ed.; Academic Press: New York, 2001; 149–201.
49. Nikolakis, V.; Kokkoli, E.; Tirrell, M.; Tsapatsis, M.; Vlachos, D.G. Zeolite growth by addition of subcolloidal particles: modeling and experimental validation. Chem. Mater. **2000**, *12*, 845–853.
50. Kragten, D.D.; Fedeyko, J.M.; Sawant, K.R.; Rimer, J.D.; Vlachos, D.G.; Lobo, R.F.; Tsapatsis, M. The structure of the silica phase extracted from silica/TPAOH solutions containing nanoparticles. J. Phys. Chem. B **2003**, *107*, 10006–10016.
51. Voter, A.F. Hyperdynamics: accelerated molecular dynamics of infrequent events. Phys. Rev. Lett. **1997**, *78*, 3908–3911.
52. Gillespie, D.T. Approximate accelerated stochastic simulation of chemically reacting systems. J. Chem. Phys. **2001**, *115*, 1716–1733.
53. Marrink, S.J.; de Vries, A.H.; Mark, A.E. Coarse-grained model for semiquantitative lipid simulations. J. Phys. Chem. B **2004**, *108*, 750–760.
54. Katsoulakis, M.A.; Majda, A.J.; Vlachos, D.G. Coarse-grained stochastic processes and Monte Carlo simulations in lattice systems. J. Comp. Phys. **2003**, *186*, 250–278.
55. Katsoulakis, M.A.; Vlachos, D.G. Coarse-grained stochastic processes and kinetic Monte Carlo simulators for the diffusion of interacting particles. J. Chem. Phys. **2003**, *119*, 9412–9428.
56. Katsoulakis, M.; Majda, A.J.; Vlachos, D.G. Coarse-grained stochastic processes for microscopic lattice systems. Proc. Natl. Acad. Sci. **2003**, *100*, 782–787.

Molecular Self-Assembly

Yoon Seob Lee
James F. Rathman
Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio, U.S.A.

INTRODUCTION

Molecular self-assembly is the process in which molecules or molecular aggregates spontaneously organize into two- or three-dimensional supramolecular architectural objects. The formations of micelles and liquid crystalline phases in solutions containing surfactant molecules are common examples; however, in general, self-assembly phenomena are not limited to molecular-scale building blocks. Self-assembly is observed in processes on diverse length scales, from the atomic to the macroscale. Examples include amphiphilic polymers, such as block copolymers and dendrimers, and colloidal spheres and nanoparticles. The forces involved in self-assembly are weaker than the much stronger forces in covalent bonds. The organization of a self-assembled structure is maintained by a balance of weaker intermolecular forces. These forces include hydrogen bonding, coulombic interaction, coordination bonding, and van der Waals interactions. For example, self-assembly of amphiphilic molecules in water involves both hydrophobic forces, which tend to promote aggregation, and hydrophilic interactions, which favor the dispersion of molecules. Thermodynamically, self-assembly is in many cases similar to a phase transition with an important difference: the equilibrium size of the condensed phase may be finite. As the majority of self-assembly processes involve nanometer-sized entities, external effects such as capillary force, electric field, magnetic field, and flow can be strong and effective factors for controlling the self-organization.

Biological systems provide numerous examples of self-assembled objects. Owing to the relatively weak interactions involved, a self-assembled structure is much more sensitive and responsive to its environment than a more rigid structure held together by covalent bonds. Unlike processes involving simple surfactants, polymers, and nanoparticles, self-assembly processes in biological systems are usually directional and functional and often lead to the formation of extremely complex structures. For example, the three-dimensional structure adopted by a protein in solution is critical to the protein's function, and this structure is determined by both strong (covalent) and weak interactions. Thanks to the weaker interactions, the protein can respond dynamically to changes in its environment.

The wide spectrum of self-assembly phenomena can be categorized in various ways. In this entry, we discuss the similarities and the differences between two- and three-dimensional systems. The last section of this entry describes recent and possible future applications of self-assembly processes, mainly related to advanced materials, environmental issues, biotechnology, and nanotechnology. Emulsions, microemulsions, and foams are examples of important and common applications in which self-assembly plays a key role. These have a wide variety of industry applications from cosmetics, foods, detergents, oil recovery, drug formulation/delivery, petroleum refining, and mining. As these are the subjects of other topics in this encyclopedia, they are not covered here.

SELF-ASSEMBLY IN BULK (THREE-DIMENSIONAL SELF-ASSEMBLY)

Surfactant Self-Assembly and Micelles

Surfactants (short for surface active agents) are molecules that contain hydrophobic and hydrophilic groups within the same molecule.[1] A typical surfactant contains a hydrophilic ''head'' group and a hydrophobic ''tail'' group. Head groups can be cationic, anionic, nonionic, and zwitterionic.[2] Hydrocarbon tail groups can consist of one, two, or three saturated or nonsaturated chains. Surfactants with two head groups that are connected covalently through a linker (usually another hydrocarbon chain) are called Gemini surfactants.[3] Cationic and anionic surfactants are salts such as quaternary ammoniums and sulfates or sulfonates, respectively. Common nonionic surfactants include long chain derivatives of ethylene glycol head groups. Most zwitterionic surfactants have biological origins such as phospholipids.

In Fig. 1, the typical self-assembly process of surfactant molecules in aqueous solution to form aggregates called micelles is illustrated. At very low concentrations, the solution is a simple dispersion of individual

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007870

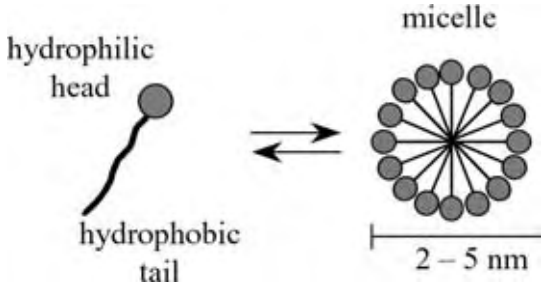


Fig. 1 Schematic of molecular self-assembly process of surfactant molecules. *(View this art in color at www.dekker.com.)*

surfactant molecules. Micellization begins at a concentration called the critical micelle concentration (cmc). Below the cmc, the hydrocarbon chains of surfactant monomers are encapsulated by the clusters of water molecules; the diffusion rate of molecules in this ''iceberg structure'' surrounding the hydrocarbon tails is approximately 1000 times slower than that of free water molecules.[1] Aggregation of surfactant monomers into micelles results in a release of the bound water because of favorable attractive interactions between surfactant tails. This phenomenon is called the ''hydrophobic effect'' and is the main driving force for micellization and many other self-assembly processes. Note that in a micelle the hydrophilic head groups remain exposed to the aqueous solution. When the head groups are ionic or dipoles, there is a net electrostatic repulsion between head groups. The equilibrium size and the shape of a micelle depend on the balance between the attractive and the repulsive forces. Any internal or external factor that affects this force balance can consequently change the micelle sizes and shapes. Such effects include the molecular structures of surfactants, surfactant concentration, solution pH, temperature, pressure, and the types and concentrations of counterions. Micellar diameters are typically in the range of 2–5 nm, primarily depending on the length of hydrocarbon chains. Micelles are dynamic objects—covalent bonding plays no role in this process. There is a continuous exchange of surfactant molecules between the monomer and micelle ''phases'' on a time scale of $\sim 10^{-5}$ sec. The micelle as a whole also may break up and reform, a relatively slow process that occurs on a time scale of 10^{-3} sec.

In Fig. 2, a variety of micelle structures are shown. Typical shapes of micelles are spherical, rod-like, and worm-like. At high concentrations of surfactant or at high concentrations of counterions, liquid crystals are usually formed. Hexagonal, cubic, and lamellar are common liquid crystal phases that occupy much of a

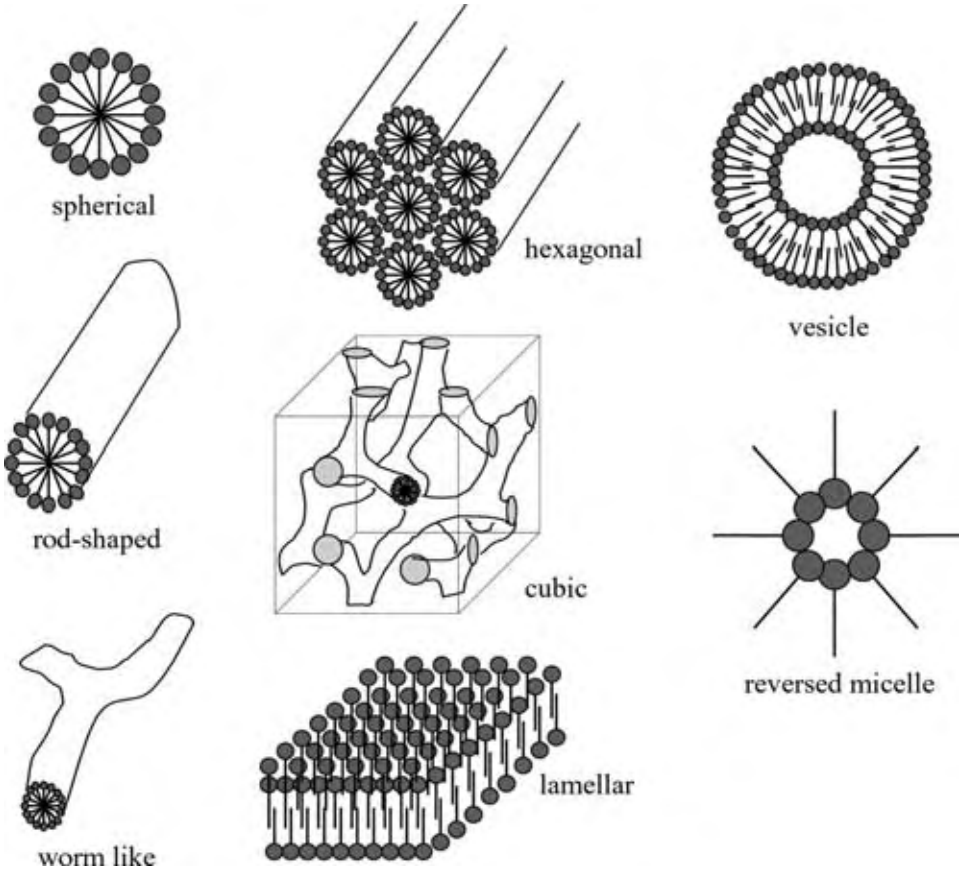


Fig. 2 Schematics of a variety of micelle structures. *(View this art in color at www.dekker.com.)*

typical phase diagram. Only three space groups are known for cubic structures of surfactant liquid crystals: *Ia*3*d*, *Pm*3*n*, and *Im*3*m*. Depending on the components and conditions, there can also be intermediate phases of liquid crystals, but usually only in a very narrow range of concentrations. These include orthorhombic, rhombohedral, and tetragonal phases. Because their structures and properties are strikingly similar to those of the biological membranes, vesicles are widely being applied as models for the study of biological cells.[4] Lipids having two hydrocarbon chains are typical amphiphiles that form vesicles (liposomes). Mixed systems of anionic and cationic surfactants at relatively low concentration may form vesicles too. When the medium consists of a binary or ternary aqueous mixture with organic solvents, their dielectric constant strongly affects the micellization, and reverse micelles can be formed at low water content. In a reverse micelle, the hydrophilic head groups are oriented inward, surrounding a small core of water molecules, while the hydrocarbon chains are oriented outward toward the oil-like solvent. Hexagonal and cubic liquid crystals have their reverse micelle-like counterparts in a media rich in the oil-like component.[5] Phase transitions between all these micelle structures including liquid crystals are possible through first order transitions.

The molecular packing parameter is a useful concept for understanding how the structure of an individual surfactant molecule influences the type of aggregate that will form by self-assembly. As illustrated in Fig. 3, the packing parameter (N_s) is simply the ratio of the volume of the hydrophobic tail to the product of the tail length and the effective head group area. Owing to the many factors that influence self-assembly, the packing parameter is an overly simplistic concept; for example, the effective head group area is often difficult to estimate because of hydration and electrostatic interactions. Despite these limitations, this dimensionless parameter does in many cases provide at least a qualitative understanding. Low values of N_s favor structures with positive curvature such as normal micelles; intermediate values favor surfaces with no net curvature, such as bilayers and cubic structures; and high values result in negative curvature.

Self-Assembly of Amphiphilic Polymers

The basic principles and driving forces for the self-assembly of polymers (polymer micellization) are the same as for the micellization of surfactant molecules. Dendrimers, and di- and tri-block copolymers are the typical amphiphilic polymers that form polymer micelles. For block copolymers, polypropylene or polybutylene segments compose the micelle cores, while polyethylene segments are the micelle corona that is exposed to water. Depending on the length of the polymer chains, micellization can occur either by intramolecular folding or by intermolecular aggregation. Polymer micelles have shapes analogous to surfactant micelles including liquid crystals, but the sizes are usually larger than their surfactant counterparts because of their molecular characteristics.[6] There are also intermicellar and intramicellar exchange processes in polymer micelles, but the diffusion rate is much slower than that of the surfactant micelles. Neutron scattering is an excellent tool to study polymer micelles, especially through contrast matching techniques.

Biological Self-Assembly

Many essential life activities are strongly controlled and regulated by self-assembly processes.[7] In addition

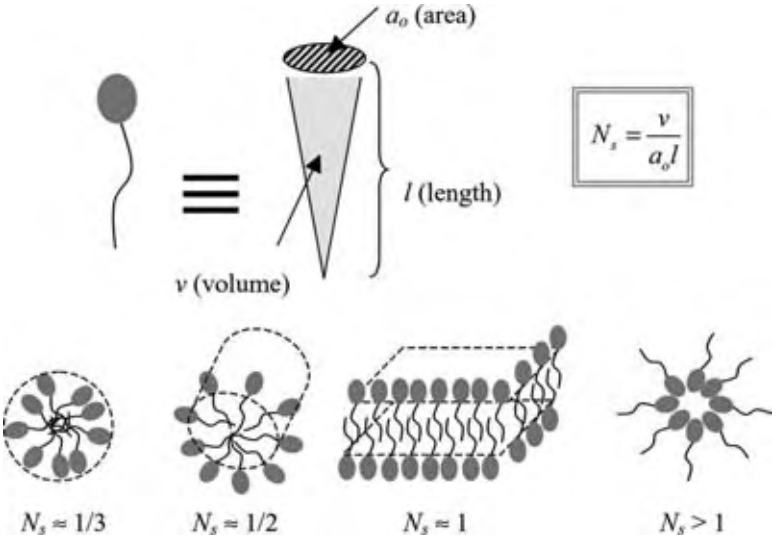


Fig. 3 Molecular packing parameter. *(View this art in color at www.dekker.com.)*

to the hydrophobic and electrostatic forces that are major players in the self-assembly of surfactants and polymers, the formation of self-assembled biological structures also often involves hydrogen bonding, π–π stacking, and coordination bonding. Covalent bonding, often involving disulfide bonds, may also be an important part of the self-assembly process.[8] These rather strong forces not only provide additional driving forces for biological self-assembly, but in many cases act as ''functional forces'' as well (Fig. 4).[9,10] As a result, biological self-assembly processes are directional (DNA, RNA, tobacco mosaic virus), functional (proteins, molecular motor), nearly irreversible (streptavidin–avidin complexation), molecular recognizing (antibody–antigen), and highly selective (enzymes and membranes). A variety of vital life activities such as cell–cell interaction, intercellular aggregation, activation of the cytoskeleton, transport through plasma membranes, cell fusion and lysis, focal adhesion, roles of collagen and fibronectin, and movement of certain cells on solid surface are largely because of biological self-assembly, which can be understood as the hierarchical evolution of self-assembly in nature.[11] Owing to its diversity and complexity, biological self-assembly is more difficult to characterize than surfactant and polymer self-assemblies.

Meso-/Macroscale Self-Assembly

Self-assembly is not limited to molecular-scale building blocks. It is possible in much larger entities, usually aided by external forces such as capillary effects, electric and magnetic fields, and shear and elongational flows. These approaches rely on the careful design and modification of surface chemistry of the building block objects. For example, assembly of solid objects on the millimeter-to-centimeter size into an amazing level of hierarchy and architectural variety using lateral capillary forces has been demonstrated.[12]

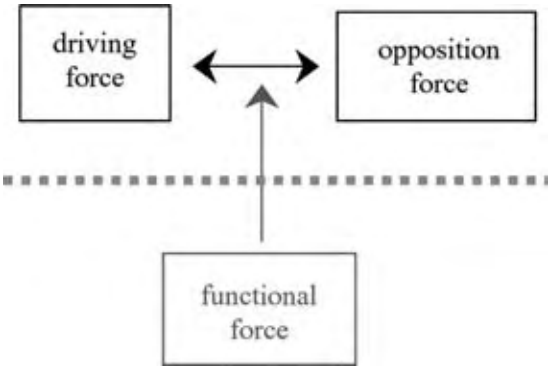


Fig. 4 Process of molecular self-assembly in biological systems. *(View this art in color at www.dekker.com.)*

Assembly and directional orientation of liquid crystals of 5CB (4-cyano-4′-pentylbiphenyl) under magnetic field provide another example of this category of self-assembly.[13]

SELF-ASSEMBLY AT INTERFACE (TWO-DIMENSIONAL SELF-ASSEMBLY)

Some of the most interesting examples of self-assembly arise at the interfacial region between two bulk phases. Surfactant molecules, for example, tend to concentrate at interfaces, and self-assembly can give rise to intricate patterns and structures.

Surface Micelles at Liquid–Solid Interface

Just as surfactants in bulk solution aggregate to form three-dimensional structures such as micelles, those adsorbed at an interface may self-assemble to form two-dimensional structures. In the past decade, the long-debated existence of ''surface micelles'' at liquid–solid interfaces has been conclusively proven, in large part thanks to the development of new imaging techniques such as high-resolution atomic force microscopy (AFM).[14] In Fig. 5, side views of two types of possible surface micelles at solid surfaces formed in solution are presented: hemimicelle and semimicelle. Hydrophobic and electrostatic forces largely govern the formation of these surface micelles and determine their sizes (diameter). As this micellization is believed to proceed after adsorption (physisorption) of amphiphilic monomers onto the solid surfaces from solution, solid surface properties such as hydrophobicity, lateral defects, and epitaxy strongly affect the final geometry of surface micelles. Hemimicelles are usually formed on hydrophobic solid surfaces, while semimicelles are formed on hydrophilic solid surfaces. Lateral dimension and length are in many cases affected by surface epitaxy. Typical factors governing bulk micellization such as monomer concentration, solution pH, type and concentration of counterions, temperature, and pressure are still important during this two-dimensional micellization. Bulk micelles have the capacity to solubilize nonpolar oil-like solutes inside

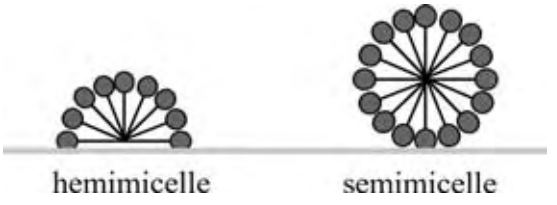


Fig. 5 Schematic side views of two typical admicelles at solid surfaces. *(View this art in color at www.dekker.com.)*

the micelles in a process called ''micellar solubilization.'' Analogous to this bulk micellar solubilization, both hemimicelle and semimicelle are capable of solubilizing nonpolar solutes inside their core region (coadsorption, adsolubilization, or surface solubilization).[15] Lateral dimensions of both hemimicelle and semimicelle are either a rod-like linear or a worm-like curved geometry. Both geometries are always highly parallel to each other and are believed to be governed by surface epitaxy. As the formation of surface micelles begins with the physisorption of monomers, there is always a monomer exchange process between surface micelles and monomer and micelles in bulk solution.

Bilayer structures may also form at solid surfaces instead of surface micelles. Whether or not they are intermediate phases for surface micelles or true stable thermodynamic phases is currently not known. As in bulk micelles, the phase transition that proceeds gradually by the change of micelle surface curvature seems to be possible in surface micelles also.

Surface Micelles at Gas–Liquid Interface

In Fig. 6, the top and side views of surface micelles at an air–liquid interface are illustrated. First proposed by Langmuir in the 1930s and debated very extensively in the literature, the existence of this equilibrium surface micelle has only recently been proven by both experimental data and theory.[16] Thermodynamic energetics and molecular packing for three-dimensional self-assembly in bulk can be directly extended to this two-dimensional system. For surface micelles, the critical micellar area (cma) is now analogous to the cmc for characterizing the onset of micelle. Both surfactants and amphiphilic polymers can form these surface micelles, with the number of aggregation molecules ranging from 10 to 1000. Structural transitions, as in bulk micelles, are possible, such as from circular-to-rod and liquid crystal-like structures, and even reverse micelles, as the surface curvature (or molecular packing geometry) is increased. The exchanging process between monomer and micelle, and intermicellar diffusion is also important, but there is no experimental report in this area to date.

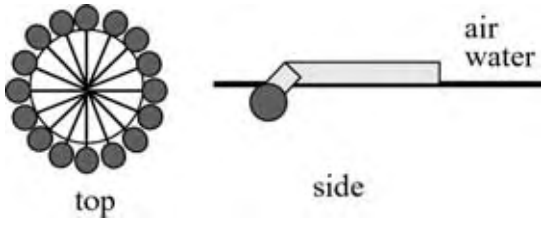


Fig. 6 Illustration of top and side views of surface micelle at air–liquid interface. *(View this art in color at www.dekker.com.)*

Langmuir Monolayers and Langmuir–Blodgett (LB) Films

In Fig. 7, the typical process for the formation of Langmuir monolayers and LB films is presented. When a Langmuir monolayer is transferred to a solid substrate at an angle other than vertical it is called Langmuir–Schafer deposition, and the deposited film is an LS film. There are no experimental and theoretical reports so far as to whether the concept of molecular packing (that was successfully applied to hemimicelles, semimicelles, and surface micelles) can be extended to these systems. The surface pressure–area isotherm is an excellent tool to follow this self-assembly event at a molecular level. At low surface pressure, the gas-like phase or ''liquid extended'' phase of amphiphilic monomers is dominant. As the surface pressure is increased, the gas-like phase is equilibrated with microsized domains called ''liquid condensed'' phase. The formation of this domain is caused by the competition between dipole moment and line tension. At high surface pressure, a tightly packed ''solid'' phase is formed. Technically, LB or LS films can be obtained at any stage of this phase transition of Langmuir monolayer, but the act of the meniscus at the point of solid–liquid–gas interface and the interaction between Langmuir monolayer and substrate often create defects such as stripe patterns, holes, and deformations of domains. This is indeed a mechanical assembly of amphiphiles (mainly insoluble or nearly insoluble) at air–liquid interface rather than a purely spontaneous thermodynamic self-assembly such as surfactant or polymer micelles. Hydrophobic and electrostatic interactions are important for the formation of domains or phases, but they are not major driving forces for the assembly. Molecular geometry of amphiphiles highly affects the size and shape of the domains, the stability of monolayers and films, and overall phase transitions. Usual parameters such as solution (subphase in this case) pH, temperature, and counterions in subphase are also strong factors for the phase formation and transitions. Surface tension and spreading are also important colloidal factors for the assembly of Langmuir and LB films.[17] Dynamic processes, such as self-diffusion or tumbling of monomers in the monolayer or film, and molecular orientation such as vertical molecular tilting or lateral positioning, are currently under intensive study.

Self-Assembled Monolayers (SAMs)

In Fig. 8, the typical structure of a SAM on a solid surface is shown. The driving force for the formation of SAMs is either coordination bonding or covalent bonding. Amphiphiles or even short chain hydrocarbons

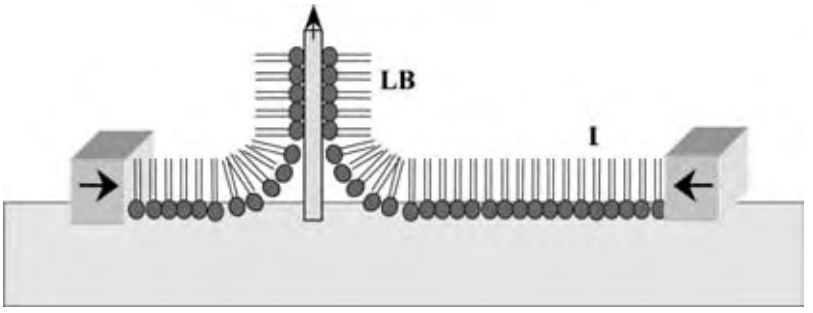


Fig. 7 Schematic for the formation of Langmuir monolayer (L) and LB film. *(View this art in color at www.dekker.com.)*

with functional (usually terminal) groups such as thiol, phosphonoic, or silane can form SAMs on a variety of solid surfaces such as gold, silver, platinum, copper, mica, silica, etc. Molecular packing is governed by hydrophobic effects, molecular structures, and surface epitaxy. Electrostatic interaction is usually not a major factor in this assembly. In a variety of media such as aqueous solution, gas, or supercritical fluid, SAMs are formed spontaneously. They show an excellent stability upon the exposure to air owing to the strong bonding between organics and substrates.[18] Domain formation and phase transitions observed in Langmuir monolayers and LB films are rarely observed in SAMs for this reason. Even though the bonding is rather strong between the monomer and solid substrate, there is an exchange process between the monomer and SAM upon the immersion of the SAM in solution.

Layer-by-Layer Self-Assembly

Alternating deposition of organic polyelectrolytes with multiple cationic and anionic charges can create well-defined nanometer-sized multilayers on solid surfaces. Electrostatic interaction is the major player for this type of self-assembly. Hydrophobic forces and molecular structure largely determine the film thickness and physical properties, but are not the major driving forces. Unlike SAMs, which often cannot be successfully formed on curved surfaces owing to the packing constraint (too loose packing on the end region of the chain that is exposed to air), assembly through layer-by-layer approach can be successfully performed on highly curved surfaces, even on surfaces of colloidal particles.[19] The flexibility of polymer chains should be carefully considered in this process. This assembly process is somewhat mechanical, in a sense that it requires constant repetition of immersion and drying. The result is highly organized well-defined self-assembled multilayers of polymeric polyelectrolytes through a very simple process. This cannot be obtained through LB or SAM techniques. Detailed structural study, diffusion issues, and exchange process are under investigation.

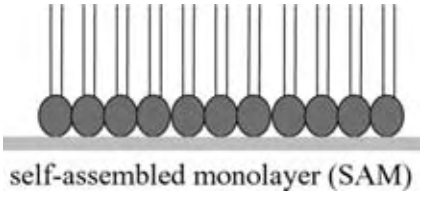


Fig. 8 Typical structure of self-assembled monolayer on solid surface. *(View this art in color at www.dekker.com.)*

CURRENT AND FUTURE APPLICATIONS

Biomineralization and Biomimetic Materials

The structural variety and the morphological variety of biominerals such as aragonite and calcite are used in nature to achieve diverse functionality and amazing mechanical properties. Understanding the process by which these materials are formed will therefore provide excellent insight into the development of new synthetic materials.[20] The key to biomineralization, the process in which these materials are produced in nature, is the cooperative self-assembly between inorganic constituents and self-assembled bio-organics such as proteins, enzymes, and membrane lipids. These self-assembled structures act as templates on which the polymerization of reactive inorganics takes place to form the final structured solid material. Biominerals can be formed through nonselective random polymerization or selective growth on specific sites on bio-organics. An important characteristic of this process is that the self-assembly usually proceeds multistepwise, making the final structures highly hierarchical.[21] Mimicking the biomineralization process in the laboratory is believed to be a very promising and efficient route toward the discovery of new materials with excellent mechanical (high volume-to-weight ratio), chemical (functional sites), and magnetic properties. Diatoms, animal bone, abalone shell, spider silk, and eggshell are among the many examples of complex materials produced in nature by biomineralization.

Advanced Materials

Examples of advanced materials include highly porous materials, such as zeolites and mesoporous materials, and multifunctional materials such as magnetorheological fluids and optomagnetic multilayered films. Methods in which organic self-assembled aggregates served as the guide or removable template for the synthesis of meso- and macroporous materials are prime examples of the utilization of self-assembly in this field of research.[22] This process can be considered as the analogy of biomineralization, in the sense that the cooperative self-assembly process is strongly involved between the inorganic constituents and the organic self-aggregates, and the entire process proceeds in a multistepwise manner. There is no selective assembly involved. The mesoporous materials have well-controlled pore sizes and shapes between two and a couple of hundred nanometers and high inner surface areas of $1000\,m^2/g$ or even higher. These remarkable physical properties are tunable through the manipulation of the self-assembly processes. Pore sizes are directly determined by the size of micelles, which is the function of the length of the monomers. Three-dimensional pore arrangement in hexagonal, cubic, or lamellar is the result of respective formation of liquid crystals of organic–inorganic composites. Often phase transitions between mesophases are observed, which is the result of the change of charge matching and packing geometry of micelles. Modification of the inner surface of mesopores for the secondary functionality such as selective binding and separation of heavy metals from wasted water can be accomplished using SAMs. Multilayering is another key in the field of advanced materials. Examples include the preparation of multifunctional materials, nanosized devices, biological sensors, drug delivery systems, and biomimetic systems for the study of biological systems. For this task, LB film techniques and layer-by-layer self-assembly provide facile and economic routes with molecular level precision. In combination with molecular recognition and directional properties of biological self-assembly, this approach can provide powerful structural diversity and control.

Nanoscience and Nanotechnology

The core concept of this fast growing new field is the ability to manipulate/assemble nanometer-scale building blocks into integrated systems.[23] Self-organizing phenomena and nanoscale events share the same working length scale of 1 nm–1 μm and are driven by the same major forces of formation, primarily weak intermolecular forces. This provides the foundation of the so-called ''bottom-up'' construction of nanoscale supramolecular architecture. The variety of above-mentioned classes of self-assemblies can provide a rich spectrum of tools to achieve this goal. Understanding and manipulation of self-assembly processes and self-aggregates in general represent the ability to assemble/link molecular or colloidal level entities together with high precision. Biological self-assembly can bring direction, selectivity, and hierarchy to this. Compared with ''top-down'' approaches that rely on creating features by ''chiseling'' bulk raw materials into nanosized architectures, the bottom-up approach can be used to create features with dimensions smaller than what can be attained by top-down methods such as lithography or microprinting. Another suggested advantage of the bottom-up approach is that it can possibly be designed to have built-in error-checking process.[24] Some of the details include synthesis and nanofabrication of nanosized material such as nanoparticles, nanocrystals, and colloidal particles, nanoelectronics, bioelectronics, molecular switches, hybridization of noncomparable materials such as bioinorganic, and biometal for multifunctional materials, and processing of functional organics such as conducting polymers or conjugate oligomers.[25] The key point of course is how to improve upon the precision that is currently available by top-down techniques. This requires a thorough understanding of the interplay between various types of colloidal forces, the exact role of each force, the interaction with the surrounding environment, and, most importantly, a facile way to control them in a mass scale.

Biological Issues and Biotechnology

In mature, self-assembly plays a key role in a huge array of nanoscale events such as protein folding/defolding, gene expression, RNA transfer, and functions of cell membrane. Understanding the delicate balance between driving and opposition forces, role of functional forces, molecular and structural origin of these forces, and changes of these forces by the surrounding factors are vital to understand biological systems. Self-assembled monolayers such as SAMs and Langmuir monolayers provide useful model systems for the study of biological membranes.[26] Similarly, liposomes and amphiphilic vesicles provide highly simplistic but useful models of biological cells. These systems provide a wide variety of site-specific functionality and domain- or phase-orientated issues that can be easily reproduced and tested in vitro.

Biotechnology deals with the engineering of biological systems. Development of target drug delivery vehicles, functional biomaterials that can be comparable with both biosystems and solid materials, such as inorganics or metals, and cell cultivation for the large-scale

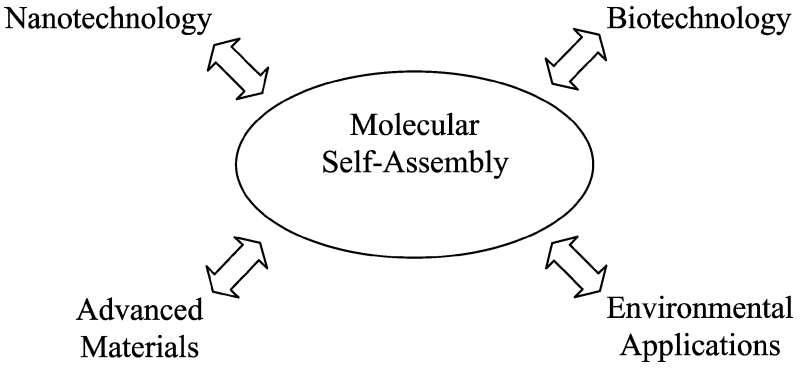


Fig. 9 Diagram of bottom-up self-assembly approach to nanotechnology, biotechnology, advanced materials, and environmental issues.

and effective production of molecular drugs are among many possible excellent examples.[27] Issues related to molecular self-assembly again are inevitably involved with these subjects, often because of effects resulting from weak intermolecular forces and molecular packing.

Environmental Issues

Two basic approaches for a better environment are minimal use of potential pollutants and effective removal of pollutants. Surfactant or polymer micelles have been demonstrated to effectively catalyze many types of reactions in aqueous solution involving organic reactants that are only sparingly soluble. Environmentally toxic materials that are insoluble or have limited solubility in water are solubilized and reacted inside micelles. Systems such as this that are macroscopically homogeneous but microscopically heterogeneous can be used as alternate solvents to minimize the use of organic solvent in industry processes. High loading, subsequent changes of micellar phases, and separation of desired product after the reaction are the issues for major research in this regard. The second approach can include microfiltration using micellar membrane systems, soil remediation using colloidal materials such as mesoporous materials and nanoparticles (for example, bimetallic nanoparticles), and sonochemical removal of pollutants. Micellar solubilization, counterion binding, and interfacial self-assembly are among the key concepts in addition to the self-assembly process itself.[28]

CONCLUSIONS

The spontaneous formation of structures by self-assembly relies on a delicate balance of relatively weak intermolecular forces such as the hydrophobic force, electrostatic and van der Waals interactions, and hydrogen bonding. Building blocks for self-assembly can range from small atoms to much larger macroscale objects. A wide variety of two- and three-dimensional architectures are possible; common examples include micelles, liquid crystals, tertiary protein structure, cell membranes, monolayers, multilayers, vesicles, and liposomes. Self-assembled structures are responsive to their environment; hence, factors such as temperature, pressure, solution pH, and additives strongly affect the self-assembly process. External factors such as capillary force, flow, magnetic field, and electric field can also be used to control the self-assembly process.

Self-assembly plays a key role in many areas (Fig. 9). Using self-aggregates as templates has proven to be an extremely powerful approach for developing novel materials, especially mesoporous particles and films for use as catalysts, catalyst supports, sorbents, and molecular sieves. Biological systems offer many examples of highly complex process in which intricate structures formed by self-assembly play a key role. Understanding these systems will help drive the development of next-generation materials.

REFERENCES

1. Evans, D.F.; Wennerström, H. *The Colloidal Domain: Where Physics, Chemistry, and Biology Meet*, 2nd Ed.; Wiley-VCH, 1999.
2. Hiemenz, P.C.; Rajagopalan, R. *Principles of Colloid and Surface Chemistry*; Marcel Dekker, 1997.
3. Hait, S.K.; Moulik, S.P. Gemini surfactants: a distinct class of self-assembling molecules. Curr. Sci. **2002**, *82* (9), 1101–1111.
4. Lasic, D.D. *Liposomes: From Physics to Applications*; Elsevier, 1993.
5. Israelachvili, J. *Intermolecular and Surface Forces*, 2nd Ed.; Academic Press, 1992.
6. Mortensen, K. Small-angle scattering studies of block copolymer micelles, micellar mesophases and networks. In *Amphiphilic Block Copolymers*; Alexandridis, P., Lindman, B., Eds.; Elsevier: Amsterdam, 2000; 191–220.
7. Philp, D.; Stoddart, J.F. Self-assembly in natural and unnatural systems. Angew. Chem. Int. Ed. Engl. **1996**, *35*, 1154–1196.

8. Lehn, J.-M.; Ball, P. Supramolecular chemistry. In *The New Chemistry*; Hall, N., Ed.; Cambridge University Press, 2000; 300–351.
9. Antonietti, M.; Göltner, C. Superstructures of functional colloids: chemistry on the nanometer scale. Angew. Chem. Int. Ed. Engl. **1997**, *36*, 910–928.
10. Lindsey, J.S. Self-assembly in synthetic routes to molecular devices. Biological principles and chemical perspectives: a review. New J. Chem. **1991**, *15*, 153–180.
11. Seckler, R. Assembly of multi-subunit structures. Front. Molec. Biol. **2000**, *32*, 279–308.
12. Choi, I.S.; Bowden, N.; Whitesides, G.M. Macroscopic hierarchical, two-dimensional self-assembly. Angew. Chem. Int. Ed. **1999**, *38* (20), 3078–3081.
13. Gupta, V.K.; Abbott, N.L. Design of surfaces for patterned alignment of liquid crystals on planar and curved substrates. Science **1997**, *276*, 1533–1536.
14. Manne, S.; Gaub, H.E. Molecular organization of surfactants at solid–liquid interfaces. Science **1995**, *270*, 1480–1482.
15. Dickson, J.; O'Haver, J. Adsolubilization of naphthalene and α-naphthol in C_nTAB admicelles. Langmuir **2002**, *18* (24), 9171–9176.
16. Israelachvili, J. Self-assembly in two dimensions: surface micelles and domain formation in monolayers. Langmuir **1994**, *10*, 3774–3781.
17. Basu, J.K.; Sanyal, M.K. Ordering and growth of Langmuir–Blodgett films: x-ray scattering studies. Phys. Rep. **2002**, *363*, 1–84.
18. Schwartz, D.K. Mechanisms and kinetics of self-assembled monolayer formation. Annu. Rev. Phys. Chem. **2001**, *52*, 107–137.
19. Salditt, T.; Schubert, U.S. Layer-by-layer self-assembly of supramolecular and biomolecular films. Rev. Mol. Biotechnol. **2002**, *90* (1), 55–70.
20. Ball, P. *Made to Measure: New Materials for the 21st Century*; Princeton University Press, 1998.
21. Tirrell, M.; Kokkoli, E.; Biesalski, M. The role of surface science in bioengineered materials. Surf. Sci. **2002**, *500* (1–3), 61–83.
22. Soler-Illia, G.J.A.A.; Sanchez, C.; Lebeau, B.; Patarin, J. Chemical strategies to design textured materials: from microporous and mesoporous oxides to nanonetworks and hierarchical structures. Chem. Rev. **2002**, *102*, 4093–4138.
23. Committee for the Review of the National Nanotechnology Initiative, National Research Council. In *Small Wonders, Endless Frontiers: A Review of the National Nanotechnology Initiative*; National Academies Press: Washington, DC, 2002.
24. Whitesides, G.M. Self-assembling materials. Sci. Am. Sep. **1995**, 146–149.
25. Bashir, R. DNA-mediated artificial nanobiostructures: state of the art and future directions. Superlattices Microstruct. **2001**, *29*, 1–16.
26. Scott, H.L. Modeling the lipid component of membrane. Curr. Opin. Struct. Biol. **2002**, *12* (4), 495–502.
27. Torchilin, V.P. Drug targeting. Eur. J. Pharm. Sci. **2000**, *11*, S81–S91.
28. Deitsch, J.J.; Rockaway, E.J. Surfactant-enhanced desorption of organic pollutants from natural soil. Physicochem. Groundwater Remed. **2001**, 217–243.

Molecularly Imprinted Polymers

Gregory T. Rushton
Department of Chemistry and Biochemistry, Kennesaw State University, Kennesaw, Georgia, U.S.A.

Ken D. Shimizu
Department of Chemistry and Biochemistry, University of South Carolina, Columbia, South Carolina, U.S.A.

INTRODUCTION

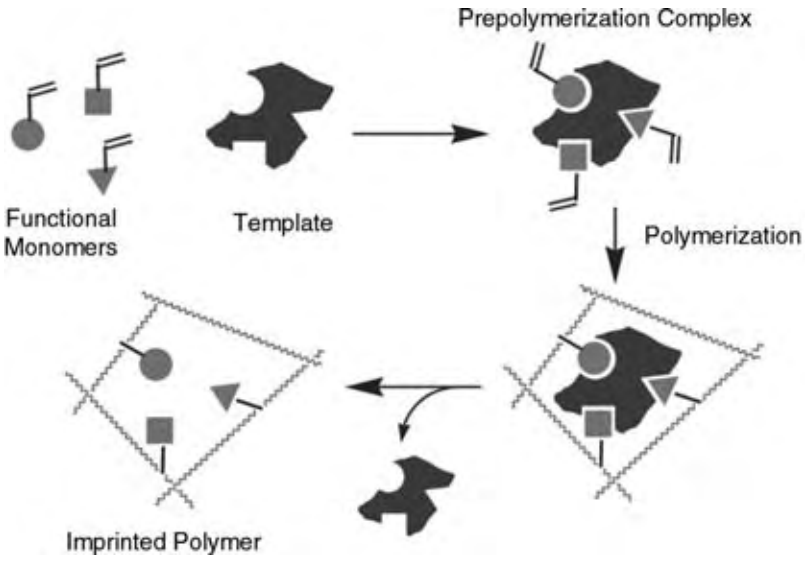
Molecularly imprinted polymers (MIPs) are polymers formed in the presence of a template molecule. Removal of the template from the polymer matrix creates complementary binding sites with affinity and selectivity for the template molecule. Molecularly imprinted polymers are attractive materials capable of molecular recognition owing to their versatility, ease of preparation, and robust physical and chemical properties. Over the past 30 years, MIPs have been developed as stationary phases in chromatography, heterogeneous catalysts in organic synthesis, and sensors for a wide array of biologically relevant compounds. This review seeks to highlight the accomplishments of these materials over this period of time, which has led to their rise in popularity and their potential in commercial applications. We begin by describing the primary approaches to preparing MIPs, and then discuss their success in various applications. We conclude with an eye to the future and suggest where the imprinting field may be heading.

BACKGROUND

Molecular recognition is an important feature of all essential biological processes. For example, enzymes and monoclonal antibodies are able to identify a specific substrate in the presence of many other structurally similar entities and then carry out their catalytic and immune response functions. The study of the recognition properties of these systems has led to insight into the mechanisms of binding and the subsequent rational design of many pharmaceuticals and chemosensors. Both biological and synthetic receptors have been intensely studied as the focus of research groups but each presents a unique set of challenges.[1–4] Biological recognition systems, such as enzymes and antibodies, have a fairly narrow range of chemical, thermal, and physical stability, which limits their ''shelf life'' and ultimately, their utility in commercial applications. They also typically, require time-consuming preparation from biological sources and the necessity to sacrifice animals. Synthetic molecular receptors, on the other hand, are more chemically and thermally stable and can, in principle, be prepared in large quantities. However, synthetic molecular receptors typically have much lower selectivity and binding affinities when compared to their biological counterparts, and they also require lengthy multistep syntheses.

Since the early 1970s, an alternative route to easily-tailored recognition materials has attracted considerable interest in the form of MIPs.[5,6] Molecularly imprinted polymers are synthetic polymers that contain shape-specific binding cavities that are lined with complementary functional groups, mimicking the recognition properties of biological systems.[7,8] These materials are versatile and easy to prepare. A generalized imprinting scheme is shown below (Scheme 1). First, a template or ''guest'' molecule is chosen and is linked to functional monomer(s) via either covalent or noncovalent bonds to form a prepolymerization complex. The prepolymerization complex is then preserved by polymerization into a highly cross-linked polymer matrix, when rigid cavities are formed that are complementary in size and shape to the template. Upon removal of the template molecule, the polymer is ready to be used as a recognition material for the host molecule. Imprinted polymers have many attractive characteristics, such as the ability to be prepared and utilized in organic solvents. They retain their recognition properties when stored dry for long periods of time and are thermally stable up to 125°C. Molecularly imprinted polymers are readily accessible in large quantities as they can be formed in a single step from commercially available starting materials. Finally, the recognition properties of imprinted polymers can be easily tailored to new substrates simply by carrying out the polymerization using the appropriate template molecule. This combination of attributes is complementary to the molecular receptors produced by immunological and organic synthesis. For example, MIPs can be made using highly toxic or hydrolytically

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120041040

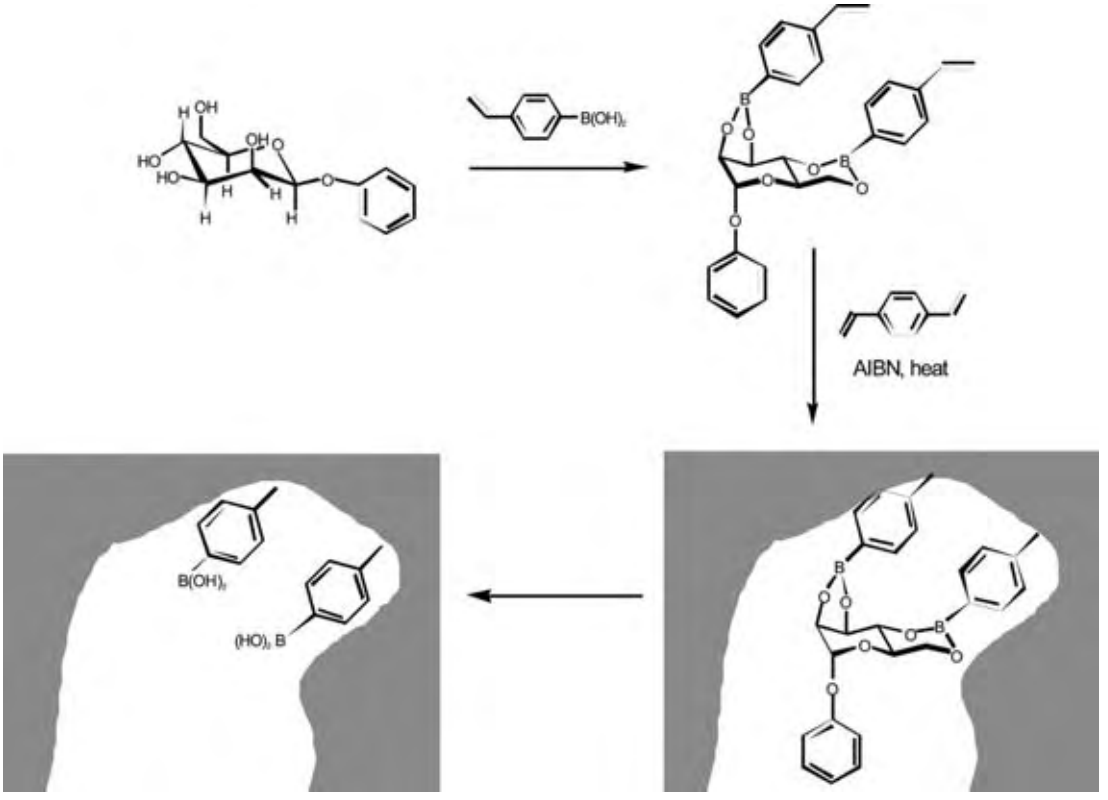


Scheme 1 The imprinting process begins with the selection of a template molecule and complementary functional monomer(s). *(View this art in color at www.dekker.com.)*

unstable templates, such as nerve toxins, pesticides, or short peptides for which antibodies cannot be directly elicited. The molecular imprinting process is also more efficient than the preparation of most synthetic or biological receptors.

The first forays into molecular imprinting took place in the 1940s when Frank Dickey, a student of Linus Pauling, prepared silica gels in the presence of organic dyes, such as methyl orange and found that these gels showed improved adsorption of the dyes as compared to a control gel.[9] Although these results were promising, it was not until the early 1970s that Gunter Wulff began conducting research that served as the impetus for the current generation of molecularly imprinted materials. Initial efforts by Wulff et al. focused on the difficult problem of carbohydrate recognition using boronic acid monomers that form covalent boronate linkages with 1,2-diols. Although the boronate ester is a covalent linkage, it is reversible and can be cleaved under hydrolytic conditions (Scheme 2). This covalent imprinting approach yields a homogeneous population of binding sites with a high affinity for the template, which could even resolve enantiomers when the polymer was used as the stationary phase in liquid chromatography. For example, when phenyl-α-mannopyranoside was selected as the template and (4-vinyl)-phenylboronic acid was used as the monomer, the resulting MIP could discriminate D- and L-enantiomers with baseline separation and separation factors up to 4.56.[10] The enantioselectivity of the imprinted polymer provided verification that the imprinting process produced template selective binding cavities. The constituent monomers are achiral and, therefore, the enantioselectivity of the imprinted polymer can only arise from the imprinting process. Polymers made without template did not show any enantioselectivity when measured under similar conditions.

While the covalent imprinting approach proved successful, it was limited to templates that could form labile covalent bonds to a functional monomer, such as boronic esters, imines, ketals, and disulfides. In the 1980s, a new noncovalent imprinting approach was introduced by Klaus Mosbach. The noncovalent imprinting approach quickly became the preferred approach owing to the ease of synthesis and the ability to tailor the materials to accommodate a wide range of template molecules. The key feature of the approach was the in situ formation of the prepolymerization complexes via self-assembly of the template and functional monomers using reversible noncovalent interactions, such as hydrogen bonding, electrostatic interactions, solvaphobic, and π–π interactions. This simplifies the imprinting procedure into a more versatile one-pot process (Scheme 3). The self-assembled prepolymerization complex is formed by simply combining the component template and functional monomer(s) and then is polymerized into a highly cross-linked polymer matrix. The reversibility of the template–functional monomer interactions in the prepolymerization complex allowed a common set of functional monomers and cross-linkers to be used with many different template molecules. The most common pair of functional monomer and cross-linking agent is methacrylic acid (MAA) and ethylene glycol dimethacrylate (EDMA). This copolymer system has been used to imprint amino acids, pharmaceuticals, and herbicides. An additional advantage of the noncovalent imprinting approach is that interactions of the imprinted binding cavity with the template molecule are also noncovalent interactions. This allows for much faster binding kinetics, which is important for

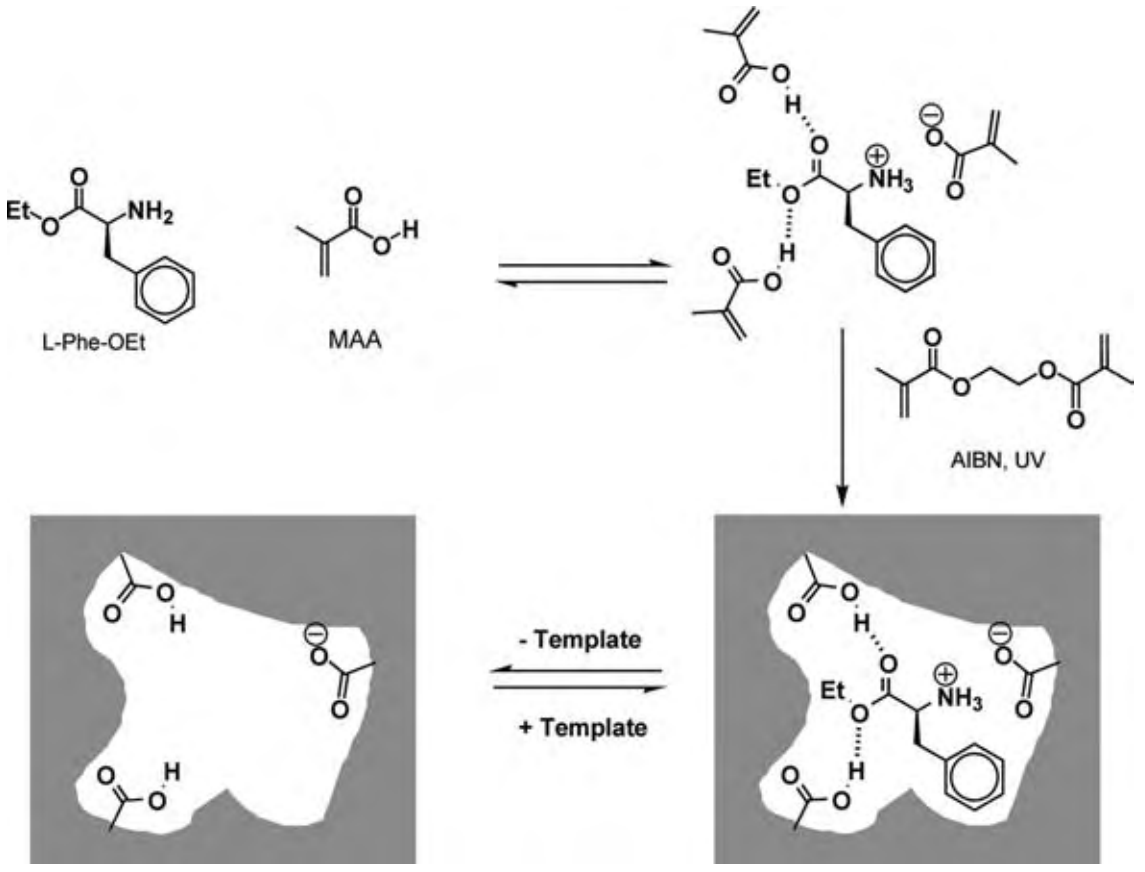


Scheme 2 Covalent imprinting of α-phenyl-D-mannopyranoside in a DVB/4-vinylphenylboronic acid matrix, via the formation of covalent boronic ester linkages between the 4-vinylphenylboronic acid and the carbohydrate. *(View this art in color at www.dekker.com.)*

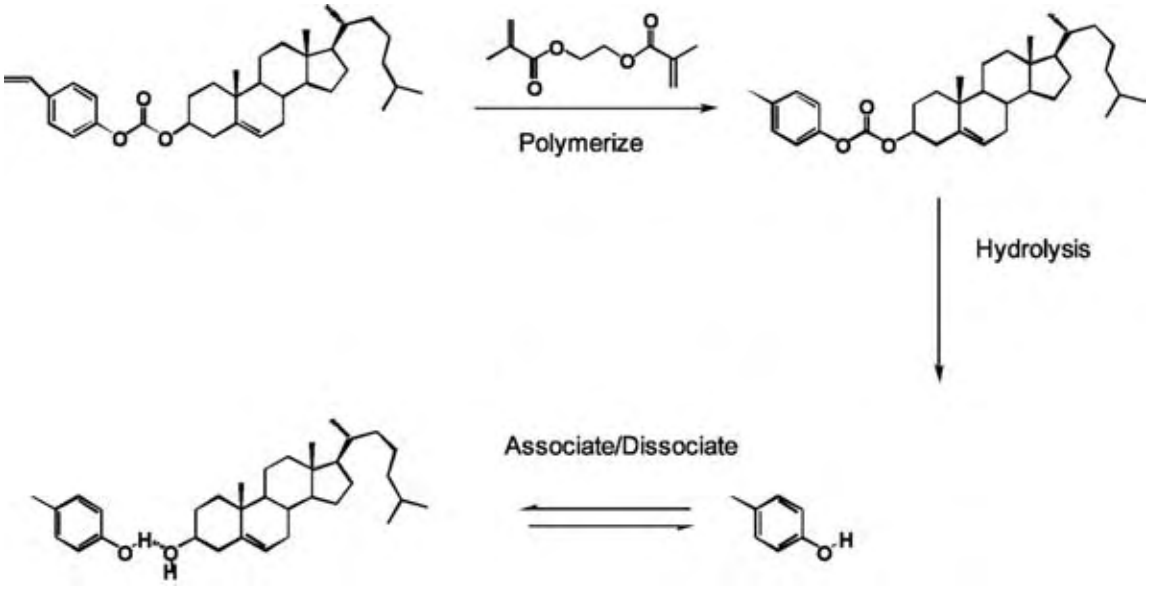
chromatographic and sensing applications. One disadvantage of the noncovalent imprinting approach is binding site heterogeneity arising from the dynamic nature of the noncovalent prepolymerization complex. Noncovalently imprinted polymers contain a range of binding sites that vary from low affinity to high affinity.[11,12] Unfortunately, the binding site distribution is heavily weighted toward the low-affinity selectivity binding sites. The MIPs prepared by the covalent approach tend to have a more homogeneous binding distribution that does not extend out as far into the high-affinity region.

More recently, hybrid-imprinting strategies have been developed that seek to combine the advantages of covalent and noncovalent imprinting methodologies. The hybrid imprinting methods attempt to form the prepolymerization complex in a stoichiometric fashion like the covalent imprinting approach, while still producing an imprinted polymer that forms noncovalent interactions with the template molecule. For example, the sacrificial spacer approach developed by Whitcombe et al. uses covalent bonds to connect the monomers and template like the covalent approach. However, hydrolysis of the template unmasks hydrogen-bonding functionality that forms noncovalent interactions to rebind the guest. A specific example of Whitcombe's sacrificial spacer imprinting method is outlined in Scheme 4. A cholesterol containing prepolymerization complex was prepared using (4-vinyl)-phenyl carbonate ester. Polymerization with EDMA followed by hydrolysis of the template from the polymer matrix yielded a phenol moiety positioned in a complementarily shaped binding cavity that could bind cholesterol through noncovalent hydrogen bonding interactions.[13] This polymer was able to discriminate among cholesterol analogs, such as epicholesterol and cholesterol acetate.

Another hybrid imprinting method is to use functional monomers that form particularly strong noncovalent interactions with the template. The ''stoichiometric'' noncovalent imprinting approach retains the synthetic efficiency ''one pot'' imprinting procedure of the noncovalent approach but also produces MIPs with a more homogeneous distribution of high-affinity, high-selectivity binding sites like the covalent approach. One moiety that has been used is the amidine group, which can form both strong hydrogen bonds and electrostatic interactions with carboxylic acids, phosphonates, and phosphoric esters. Sellergren used MAA to bind the pneumonia drug pentamidine



Scheme 3 Noncovalent imprinting of L-phenylalanine methyl ester in a methacrylic acid (MAA)/ethyleneglycol dimethacrylate (EDMA) polymer matrix. *(View this art in color at www.dekker.com.)*



Scheme 4 A hybrid method of imprinting by Whitcombe et al. uses stoichiometric amounts of monomer and template during polymerization and then relies on noncovalent interactions during the rebinding phase.

and Wulff used the amidine moiety in a monomer to resolve two enantiomers of *N*-(4-carboxybenzoyl)-phenylglycine with separation factors up to 2.8.[14]

COMPOSITION AND PREPARATION OF MIPs

One of the most attractive features of the imprinting process is the ease with which new MIPs can be prepared for and tailored to specific applications. This is particularly the case for the noncovalent imprinting approach. In most cases, MIPs can be synthesized and processed within a few days using readily available starting materials. First, a functional monomer must be selected that forms reversible interactions with the chosen template molecule. The most common functional monomers, such as acrylic acids and vinylpyridines, are inexpensive, commercially available, and readily polymerized by free radical polymerization.[15] Although the imprinting mechanism is not restricted to free radical polymerization of vinyl monomers, radical polymerizations are the most common method of preparation of MIPs because of their high yields in a range of solvents and tolerance to many acidic and basic functional groups that are used in the imprinting process.

The second variable is the cross-linking agent, which serves as a solid support for the functional monomer, ensuring the proper distance and orientation of the functional groups around the template molecule. The cross-linking agent also provides shape selectivity by forming a rigid cavity around the template. These cavities must remain intact during polymerization and subsequent extraction of the template from the matrix for selective binding properties to be observed. Common cross-linking agents are EDMA, divinylbenzene (DVB), and the bisacrylamides (Fig. 1). Wulff systematically examined the influence of the cross-linker on the selectivity of the resulting imprinted polymer. A high percentage of cross-linking agent, typically 80%, was required for the matrix to maintain the integrity of the binding cavity after cleavage of the template.

The final component in the polymerization mixture is the solvent. Although the choice of solvent is often overlooked, it can have dramatic consequences on the surface area, properties of the materials, and binding affinity and selectivity of the resulting MIP. The solvent must dissolve the template, monomers, and cross-linker without disrupting the stability of the prepolymerization complex. Common solvents include acetonitrile, chloroform, and toluene. The solvent also acts as a porogen, creating ''macropores'' within the imprinted polymer so that the template can be efficiently removed from the cross-linked matrix and can also access the binding sites during the rebinding process. Imprinted polymers are typically macroporous monoliths with high surface areas of 100–600 m^2/g. Near theta solvents in which short polymer chains are soluble but longer polymer chains are insoluble are particularly effective in yielding highly macroporous morphology because of phase separation during polymerization.

Other variables in the imprinting process are stoichiometry and temperature, which are particularly important in the cases of noncovalent imprinted polymers. Lower temperatures and higher functional monomer to template ratios stabilize noncovalent prepolymerization complexes, resulting in noncovalent MIPs with higher capacities and selectivities. The imprinting process can be carried out at lower temperatures by either UV initiated radical polymerizations or by using azo-based radical initiators.

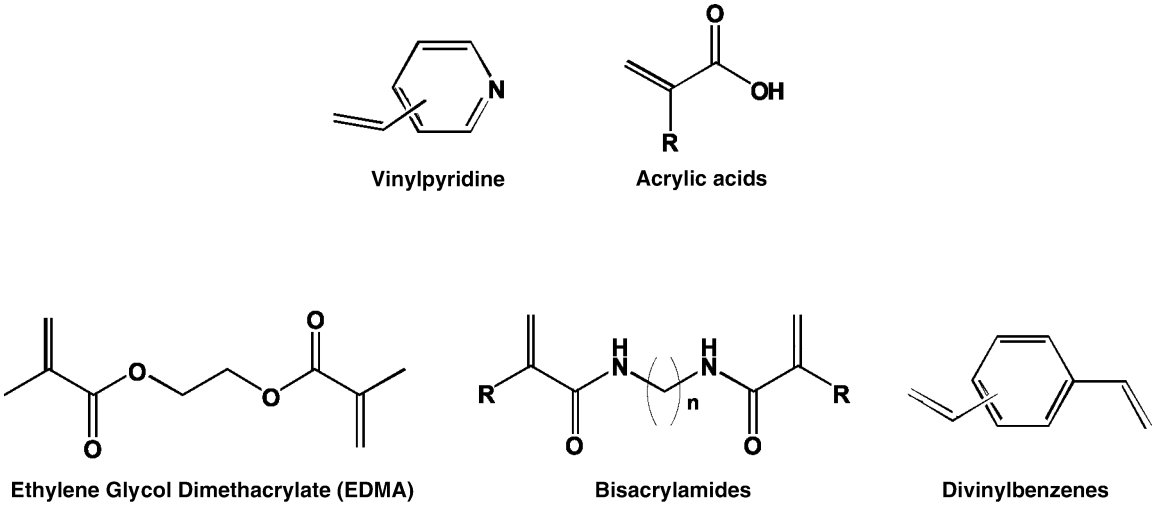


Fig. 1 Common functional monomers (top) and cross-linking agents utilized in noncovalent imprinting.

Once the imprinted polymer has been synthesized, the template must be washed out of the matrix. In most cases, greater than 80% of the template can be effectively removed by Soxhlet extraction. This is sufficient for most applications, and the recovered template can be reused, which is important if the template molecule is particularly valuable. However, a small amount of template usually remains in the highest affinity sites and slowly leaches out over time, which can interfere with sensing applications at very low concentrations (nM). These issues can be circumvented by selecting applications for MIPs that are carried out at micromolar (μM) or higher concentrations or by indirectly monitoring binding of the analyte by radioligand or fluorescently labeled analyte assays.

POLYMER MORPHOLOGIES

As the imprinting field moves toward commercial applications, the need to conveniently and efficiently produce the polymers in different formats is becoming increasingly important. In this section, the primary methods utilized for the preparation of MIPs and their relative merits for various applications will be presented.

By far the most popular approach for preparing MIPs is bulk polymerization. In this strategy, the template, monomer, cross-linker, and initiator are combined in a suitable porogenic solvent and irradiated or heated until a sufficient degree of polymerization has taken place. The macroporous monolith is then ground and sieved to achieve a narrow size distribution of particles. Polymers synthesized in this way are used in chromatography, solid phase extractions, and binding assays.[16] Grinding the polymer does not destroy or alter the binding sites. However, it does enhance the accessibility of the binding sites, which is important for chromatographic applications. The polymer monoliths can also be used directly, which eliminates the grinding and sieving steps that contribute to lower yields. Molecularly imprinted polymer monoliths have been synthesized in chromatography columns and have shown similar levels of affinity and selectivity.

For chromatographic applications, monodisperse spherical MIP particles are preferable to the irregularly shaped MIP particles formed by grinding and sieving the polymer monoliths. Consequently, some researchers have developed methods that produce spherical MIP particles through precipitation polymerization and emulsion polymerization.[17] Monodisperse MIP microspheres were prepared by Mosbach et al. under similar conditions to the MIP monolith preparation but under significantly more dilute precipitation polymerization conditions.[18] The MIP microspheres grew to 1–3 μm diameter with a narrow size distribution. These imprinted microspheres can then be packed more efficiently into chromatography columns or into solid-phase extraction (SPE) cartridges than the particles prepared by bulk polymerization techniques. Larger spherical imprinted polymer particles can be prepared by modification of preformed latex particles either by reswelling with a secondary polymerization mixture or by coating a spherical core particle with an imprinted polymer shell.

Imprinted thin films or membranes have also been prepared for applications in chromatography, sensing, and SPE and can be either freestanding or supported on a solid substrate. The MIP films have been synthesized by bulk thin film polymerization, surface initiated polymerization, and phase inversion precipitation of a preformed linear polymer. For example, Rotello and Penelle et al. fabricated a polychlorinated aromatic quartz crystal microbalance (QCM) sensor by bulk thin polymerization with a hexachlorobenzene imprinted thin film directly on to gold surface. The issues of adherence of the polymer film to the gold substrate and the film integrity on drying were resolved by the choice of a more flexible 1,5-bis(2-acetylaminoacryloyloxy)pentane cross-linker. Sellegren et al. have synthesized MIP thin film stability by preparation of MIPs via surface initiated polymerization. Azo-based radical initiators were anchored to silica surfaces to develop new high surface area MIP stationary phases.[19] Kobayashi et al. have developed a unique preparation of MIP thin films by precipitation of linear polymers in the presence of templating agents. For example, Nylon-6 and L-glutamine dissolved in formic acid were cast onto a solid substrate to a thickness of 100 μm.[20] Rinsing the film with water leads to a phase inversion, yielding a macroporous film with enantioselective recognition properties. The imprinted Nylon-6 membranes formed by this inversion polymerization have been used as the recognition element in a QCM sensor and as a filtration membrane.

APPLICATIONS OF MIPs

The most common application for MIPs has been as stationary phases in chromatographic or SPE formats. The stability of MIPs to a wide range of solvents, temperatures, and pressures makes them particularly well suited to be stationary phases in high-performance liquid chromatography (HPLC) and capillary electrophoresis. Some of the most impressive studies using MIPs are reports of MIP systems that preferentially bind a single enantiomer over its antipode, such as amino acids and drugs. One of the key advantages of MIP-based stationary phases, in comparison to general enantioselective stationary phases, is that there is a logical order of elution of enantiomers. The imprinted

enantiomer is more strongly retained and typically elutes as a broader peak after the nonimprinted enantiomer. The known order of elution is useful in identifying the respective enantiomers in analytical applications. Mosbach et al., for example, successfully separated a racemic mixture ($\alpha = 2.9$) of the β-blocker timolol, using an HPLC column packed with an *s*-timolol imprinted polymer (Fig. 2).[21] Mosbach et al. also developed stationary phases based on MIP imprinted with a short peptide and observed separation factors as high as $\alpha = 17.8$, which exceeds that of commercial chiral stationary phases.[22]

Solid-phase extraction is a widely used method for cleanup and preconcentration of a solution containing a mixture of analytes. The SPE sorbents preferentially bind a specific analyte while allowing the other analytes to pass through. A solvent capable of disrupting the polymer–analyte interactions is then introduced to wash off the desired analyte. The concentrated and purified analyte is then quantified by a secondary method, such as gas or liquid chromatography. An attractive feature of this imprinting application is that the polymer need only work in ''on'' or ''off'' mode, either strongly adsorbing the template or releasing the analyte in contrast to the chromatographic mode where the MIP column must also exhibit good resolution to be effective. Theodoridis, for example, recently demonstrated the selective extraction of caffeine from beverages and spiked plasma using an SPE cartridge loaded with an MIP.[23] Zhu et al. recovered 96% of the herbicide metsulfuron-methyl in spiked river water using an MIP-SPE cartridge. The stability of MIP-based SPEs was also demonstrated as the SPE was shown to be reusable for up to 200 sorption studies.[24] In another study, an MIP imprinted with 4-nonylphenol showed selectivity for the template over 11 related phenolic pollutants.[25] These successes of MIPs in SPE technology have led to the establishment of MIP Technologies, Lund, Sweden (www.miptechnologies.se). MIP Technologies' MIP4SPE-Triazine10 cartridge was tailored, using the imprinting process, to selectively adsorb triazine and triazine metabolites from environmental samples. MIP Technologies have also developed other MIP-based sorbents for clenbuterol, nicotine metabolites, and riboflavin and are working toward others for steroids, peptides, and nicotine.

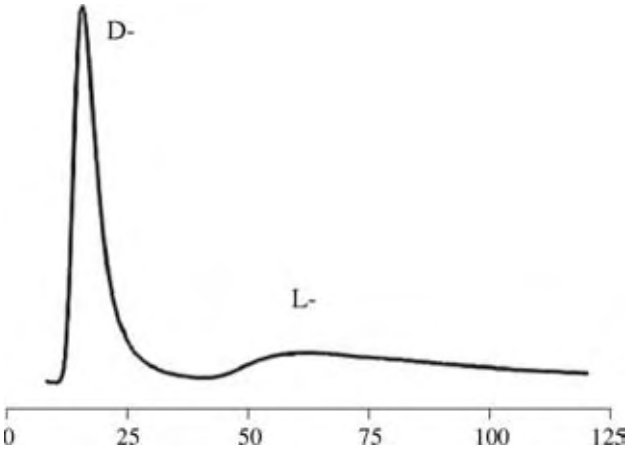


Fig. 2 Elution profile of a racemic mixture of *t*-BOC-D-tyrosine (first peak) and *t*-BOC-L-tyrosine (second peak) on a molecularly imprinted polymer made using *t*-BOC-L-tyrosine as template and *N,O*-bismethacryloyl ethanolamine as the only monomer. HPLC conditions: particle size 45–63 μm; mobile-phase: 99/1, acetonitrile/acetic acid; flow rate 0.1 mL/min; injected volume 5 μl of a 2.0 mM racemic solution; detection at $\lambda = 270$ nm. (Courtesy of David Spivak, Lousiana State University.) *(View this art in color at www.dekker.com.)*

The MIPs have also been utilized as heterogeneous catalysts.[26] The strategy is analogous to work on catalytic antibodies. The MIPs are imprinted with transition state analogs, and then the imprinted polymers tested for their ability to effect rate and selectivity enhancements. Beach and Shea, for example, prepared an MIP capable of catalyzing the dehydrohalogenation of 4-fluoro-4-(*p*-nitrophenyl)butan-2-one.[27] Sellergren and Shea have developed an MIP that mimics enzyme action by hydrolyzing a protected phenylalanine ester.[28] To date, MIP catalysts have produced only modest rate enhancements. The future of catalytic applications for MIPs will depend on the ability to form binding sites that bind the transition state of the reaction with higher affinity or include catalytic functionality. One interesting development along these lines is transition metal containing MIP catalysts.[29] For example, Severin has prepared MIPs containing ruthenium arene complexes that showed enhanced regioselectivity in the catalytic reduction of 4-acetylbenzophenone.

The MIPs have also been utilized as the recognition elements in pseudoimmunoassays.[30–32] In this approach, MIPs are substituted for antibodies to quantify the amount of analyte in a biological sample, such as blood plasma. Most MIP immunoassays are competitive binding studies in which a radio- or fluorescent-labeled analyte is added to a mixture of the MIP and unlabeled analyte. After equilibrium is reached, some fraction of the labeled species is bound to the polymer surface and thus can be separated from the supernatant. The supernatant is then analyzed via scintillation or fluorescence techniques to determine the concentration of the original unlabeled analyte. Mosbach et al. have demonstrated that MIP-based immunoassays can rival the selectivity of antibody-based assays.[33] Imprinted polymers for the opioid receptor ligands enkephalin and morphine were prepared and showed submicromolar (μM) level selectivity in a radioligand competition assay in aqueous buffers. The analysis

by Andersson et al. of *S*-propranolol in human plasma and urine using an MIP assay demonstrated remarkable accuracy and low cross-reactivity with structurally similar analogs.[34] With propranolol, nanomolar selectivity has been demonstrated using an MIP assay. The MIP assays have also been developed for the demanding application of measuring enantiomeric excess (ee).[35] An L-phenylalanine anilide (L-PAA) imprinted polymer was equilibrated with PAA solutions of varying ee. After equilibration, the concentrations of PAA remaining in solution could be correlated to the ee of the original solutions with a standard error of only $\pm 5\%$ ee. Although MIPs have shown comparable selectivity to antibodies at low concentrations, MIPs more typically show high degrees of cross-reactivity and low selectivity. Greene et al. have addressed this issue by grouping a number of different MIPs together into an MIP sensor array.[36] Six aryl amine analytes, which included structurally similar analytes, such as ephedrine and psuedoephedrine, were classified with 94% accuracy, using the MIP array assay.

Sensors based on MIPs have also been developed in which the MIP is the recognition element of the sensor, often replacing a polymer thin film or antibody. The MIPs lack signaling functionality and thus the primary difficulty is coupling a singling element to the MIP binding site. Molecularly imprinted polymers containing fluorophores and electrochemically active monomers have been prepared. For example, MIPs synthesized from conductive polymers, such as polypyrrole and polyanaline by electrochemical polymerization have been tested and shown to be able to sense the presence of the template molecule.[37] Alternatively, MIPs can be prepared on signaling surfaces, such as an electrode or a gold QCM or surface plasmon resonance surface. For example, Shoji et al. prepared an MIP-based sensor for atrazine by polymerization on the surface of a gold electrode and measuring the reduction potential of atrazine vs. a Ag/AgCl electrode.[38] The sensor showed good sensitivity for atrazine over other triazines. Kroger et al. developed a sensor for 2,4 dichlorophenoxyacetic acid, and Marx et al. coated glassy carbon electrodes with sol–gel MIP matrices to sense parathion in aqueous solutions.[39]

FUTURE APPLICATIONS FOR MIPs

The MIPs are particularly versatile materials, especially as the number and types of templates that have been successfully imprinted increase. For example, imprinted polymers have been prepared using metal ions as templates.[40] Potential applications for these polymeric ionophores include the remediation of toxic metals from the environment, recognition elements for ion sensors for medical diagnostics, catalysis, and chromatography.[29] Other potential applications for MIPs have been as drugs. Researchers from Geltex Pharmaceuticals reported the use of MIPs as cholesterol lowering drugs. An imprinted version of the commercial cholesterol sequestering polymeric drug, colesevelam–HCl, was prepared and shown to have higher capacity and affinity for bile acids than the non-imprinted polymer. The imprinted colesevelam–HCl was synthesized by cross-linking poly(allylammoniumchloride) with epichlorohydrin in the presence of sodium cholate as a templating agent. The non-imprinted polymer had a capacity for bile acids in deionized water of 1.30 mmol/g, and the imprinted polymer had a capacity of 1.97 mmol/g. Other medical applications for MIPs include MIP hydrogels for timed and triggered drug release.

Another limitation of imprinted polymers is the size of the template. Most template molecules are small molecules of less than 20–30 Å in length.[41] Larger templates become physically entrapped in the highly cross-linked matrix and cannot be removed from their binding sites. Thus, important biological macromolecules and structures, such as DNA, proteins, viruses, and cells cannot be imprinted by traditional methods. However, new surface imprinting methods are being developed, which should greatly increase the scope and utility of imprinted materials. For example, Ratner et al. have imprinted proteins using cellulose films.[42] The proteins are first coated with disaccharides. Then, a glow-discharge plasma deposition covalently links the disaccharides into a polymer film around the proteins. A number of different proteins were imprinted by this method including immunoglobulin G, ribonuclease, and streptavidin. The protein-imprinted films could be patterned into a microarray using microcontact printing, opening up the possibility of biomedical applications. Nusslein et al. have imprinted even larger biological structures, specifically bacteria, using a polymer film on the surface of a QCM. The imprinted polymer film displayed recognition abilities for the shape and surface functionality of the imprinted bacteria. The cell-selective QCM was able to differentiate gram-positive and gram-negative cells, cell aggregates, and cell shapes.

CONCLUSIONS

The MIPs have already shown great potential in a wide range of applications requiring molecular recognition elements. The MIPs can replace synthetic molecular receptors or antibodies in many applications and can function with similar levels of affinity and selectivity. The MIPs have the advantage that they can be tailored to recognize an ever-increasing pool of templates including chiral amines, carbohydrates, proteins, and

bacteria. With improvements in the imprinting process and the development of new imprinted polymer morphologies and formats, MIPs will find new applications in catalysis, separation materials, pharmaceuticals, medical devices, and sensors.

ACKNOWLEDGMENT

The authors are grateful to the National Institutes of Health (GM062593) for support of research on MIPs.

REFERENCES

1. Breining, S.R. Recent developments in the synthesis of nicotinic acetylcholine receptor ligands. Curr. Top. Med. Chem. **2004**, *4* (6), 609–629.
2. Hiley, C.R.; Ford, W.R. Cannabinoid pharmacology in the cardiovascular system: potential protective mechanisms through lipid signaling. Biol. Rev. Camb. Philos. Soc. **2004**, *79* (1), 187–205.
3. Abe, H.; Mawatari, Y.; Teraoka, H.; Fujimoto, K.; Inouye, M. Synthesis and molecular recognition of pyrenophanes with polycationic or amphiphilic functionalities: artificial plate-shaped cavitant incorporating arenes and nucleotides in water. J. Org. Chem. **2004**, *69* (2), 495–504.
4. Lehn, J.-M. *Supramolecular Chemistry: Concepts and Perspectives*; VCH: New York, 1995.
5. Sellergren, B., Ed. *Molecularly Imprinted Polymers. Man Made Mimics of Antibodies and Their Applications in Analytical Chemistry*; Elsevier: Amsterdam, 2001.
6. Molecularly imprinted materials. Sci. Technol. **2005**, 734.
7. Alexander, C.; Davidson, L.; Hayes, W. Imprinted polymers: artificial molecular recognition materials with applications in synthesis and catalysis. Tetrahedron **2003**, *59* (12), 2025–2027.
8. Haupt, K. Molecularly imprinted polymers in analytical chemistry. Analyst **2001**, *126* (5), 747–756.
9. Dickey, F.H. Specific adsorption. J. Phys. Chem. **1955**, *59*, 695–707.
10. Wulff, G.; Vesper, W.; Grobe-Einsler, R.; Sarhan, A. Enzyme-analog built polymers, 4. The synthesis of polymers containing chiral cavities and their use for the resolution of racemates. Makromol. Chem. **1977**, *178* (10), 2799–2816.
11. Umpleby, R.J., II; Bode, M.; Shimizu, K.D. Measurement of the continuous distribution of binding sites in molecularly imprinted polymers. Analyst. **2000**, *125* (7), 1261–1265.
12. Umpleby, R.J., II; Baxter, S.C.; Chen, Y.; Shah, R.N.; Shimizu, K.D. Characterization of molecularly imprinted polymers with the Langmuir–Freundlich isotherm. Anal. Chem. **2001**, *73* (19), 4584–4591.
13. Whitcombe, M.J.; Rodriguez, M.E.; Villar, P.; Vulfson, E.N. A new method for the introduction of recognition site functionality into polymers prepared by molecular imprinting: synthesis and characterization of polymeric receptors for cholesterol. J. Am. Chem. Soc. **1995**, *117*, 7105–7111.
14. Wulff, G.; Schönfeld, R. Polymerizable amidines—adhesion mediators and binding sites for molecular imprinting. Adv. Mater. **1998**, *10*, 957–959.
15. Sellergren, B., Ed. *Molecularly Imprinted Polymers. Man Made Mimics of Antibodies and Their Applications in Analytical Chemistry*; Elsevier: Amsterdam, 2001.
16. Sellergren, B. Imprinted chiral stationary phases in high-performance liquid chromatography. J. Chromatogr. A. **2001**, *906* (1–2), 227–252.
17. Tovar, G.E.M.; Krauter, I.; Gruber, C. Molecularly imprinted polymer nanospheres as fully synthetic affinity receptors. Top. Curr. Chem. **2003**, *227*, 125–144.
18. Ye, L.; Weiss, R.; Mosbach, K. Synthesis and characterization of molecularly imprinted microspheres. Macromolecules **2000**, *33* (22), 8239–8245.
19. Sulitzky, C.; Ruckert, B.; Hall, A.J.; Lanza, F.; Unger, K.; Sellergren, B. Grafting of molecularly imprinted polymer films on silica supports containing surface-bound free radical initiators. Macromolecules **2002**, *35* (1), 79–91.
20. Reddy, P.S.; Kobayashi, T.; Abe, M.; Fujii, N. Molecular imprinted Nylon-6 as a recognition material of amino acids. Eur. Polym. J. **2002**, *38* (3), 521–529.
21. Fischer, L.; Müller, R.; Ekberg, B.; Mosbach, K. Direct enantioseparation of ß-adrenergic blockers using a chiral stationary phase prepared by molecular imprinting. J. Am. Chem. Soc. **1991**, *113*, 9358–9360.
22. Ramstrom, O.; Nicholls, I.A.; Mosbach, K. Synthetic peptide receptor mimics—highly stereoselective recognition in noncovalent molecularly imprinted polymers. Tetrahedron Asymmetry **1994**, *5* (4), 649–656.
23. Theodoridis, G.; Manesiotis, P. Selective solid-phase extraction sorbent for caffeine made by molecular imprinting. J. Chromatogr. A. **2002**, *948* (1–2), 163–169.
24. Zhu, Q.Z.; DeGelmann, P.; Niessner, R.; Knopp, D. Selective trace analysis of sulfonylurea herbicides in water and soil samples based on solid-phase extraction using a molecularly imprinted polymer. Environ. Sci. Technol. **2002**, *365*, 411–5420.

25. Masque, N.; Marce, R.M.; Borrull, F.; Cormack, P.A.G.; Sherrington, D.C. Synthesis and evaluation of a molecularly imprinted polymer for selective on-line solid-phase extraction of 4-nitrophenol from environmental water. Anal. Chem. **2000**, *72* (17), 4122–4126.
26. Wulff, G. Enzyme-like catalysis by molecularly imprinted polymers. Chem. Rev. **2002**, *102* (1), 1–27.
27. Beach, J.V.; Shea, K.J. Designed catalysts. A synthetic network polymer that catalyzes the dehydrofluorination of 4-fluoro-(*p*-nitrophenyl)-butan-2-one. J. Am. Chem. Soc. **1994**, *116*, 379–380.
28. Sellergren, B.; Shea, K.J. Enantioselective ester hydrolysis catalysed by imprinted polymers. Tetrahedron Asymmetry **1994**, *5*, 1403–1406.
29. Severin, K. Imprinted polymers with transition metal catalysts. Curr. Opin. Chem Biol. **2000**, *4* (6), 710–714.
30. Ansell, R.J. Molecularly imprinted polymers in pseudoimmunoassay. J. Chromatogr. B. **2004**, *804*, 151–165.
31. Ansell, R.J.; Ramstrom, O.; Mosbach, K. Towards artificial antibodies by the technique of molecular imprinting. Clin. Chem. **1996**, *42*, 1506–1512.
32. Ansell, R.J. MIP-ligand binding assays (pseudo-immunoassays). Bioseparation **2002**, *10*, 365–377.
33. Andersson, L.I.; Muller, R.; Vlatakis, G.; Mosbach, K. Mimics of the binding-sites of opioid receptors obtained by molecular imprinting of enkephalin and morphine. Proc. Natl. Acad. Sci. USA. **1995**, *92* (11), 4788–4792.
34. Bengtsson, H.; Roos, U.; Andersson, L.I. Molecular imprint based radioassay for direct determination of *S*-propranolol in human plasma. Anal. Commun. **1997**, *34*, 233
35. Chen, Y.; Shimizu, K. Measurement of enantiomeric excess using molecularly imprinted polymers. Org. Lett. **2002**, *4* (17), 2937–2940.
36. Greene, N.T.; Morgan, S.L.; Shimizu, K.D. Molecularly imprinted polymer sensor arrays. J. Chem. Soc. Chem. Commun. **2004**, *10*, 1172–1173.
37. Piletsky, S.A.; Turner, A.P.F. Electrochemical sensors based on molecularly imprinted polymers. Electroanalysis **2002**, *14* (5), 317–323.
38. Shoji, R.; Takeuchi, T.; Kubo, I. Atrazine sensor based on molecularly imprinted polymer-modified gold electrode. Anal. Chem. **2003**, *75*, 4882–4886.
39. Marx, S.; Zaltsman, A.; Turyan, I.; Mandler, D. Parathion sensor based on molecularly imprinted sol–gel films. Anal. Chem. **2004**, *76*, 120–126.
40. Striegler, S. Designing selective sites in templated polymers utilizing coordinative bonds. J. Chromatogr. B. **2004**, *1804*, 183–195.
41. Spivak, D.A.; Shea, K.J. Investigation into the scope and limitations of molecular imprinting with DNA molecules. Anal. Chim. Acta **2001**, *435* (1), 65–74.
42. Shi, H.Q.; Tsai, W.B.; Garrison, M.D.; Ferrari, S.; Ratner, B.D. Template-imprinted nanostructured surfaces for protein recognition. Nature **1999**, *398* (6728), 593–597.

Molten Carbonate Fuel Cells

Prabhu Ganesan
Branko N. Popov
Department of Chemical Engineering, University of South Carolina, Columbia, South Carolina, U.S.A.

Rajam Pattabiraman
Central Electrochemical Research Institute, Karaikudi, India

INTRODUCTION

Fuel cells are electrochemical power generators that convert chemical energy of a fuel such as hydrogen directly into electricity. Power generation through fuel cells is considered to be a new option for future energy production. Fuel cells have several advantages relative to other power generation technologies, such as: 1) very high efficiency levels for conversion of fuels into electricity; 2) low emissions of NO_x, CO_x, and other airborne pollutants, which contribute to air quality problems; 3) an inherently modular construction, minimum siting requirements, flexibility in fuel sources, and stable response to load changes; and 4) the power plant efficiencies are independent of power plant size. Unlike the other conventional energy sources, the high-temperature combustion processes are absent in fuel cells. The fuel cells are highly efficient and convert the fuel into electricity without flame or smoke and are more ecofriendly. Fuel cells are commonly classified according to the operating temperature and electrolyte employed. There are five different types of fuel cells under development, namely, alkaline fuel cell (AFC), polymer electrolyte membrane fuel cell (PEMFC), phosphoric acid fuel cell (PAFC), molten carbonate fuel cell (MCFC), and solid oxide fuel cell (SOFC). The AFC, PEMFC, and PAFC are categorized under low-temperature fuel cells, which operate in the temperature range between 50°C and 200°C. The MCFC and SOFC are placed under the high-temperature fuel cells category, which operate at 650°C and 700–1000°C, respectively. Only recently have efforts been focused on the development of direct methanol fuel cell (DMFC) as centralized power packs for stationary applications after the success of the PEMFC technology.

The MCFC technology is under consideration for on-site power generation source because of its high efficiency and ability to utilize a variety of fuel sources such as hydrogen, natural gas, and other hydrogen-rich hydrocarbons. Theoretically, a single cell can deliver an open circuit voltage of around 1.02 V. An optimized cell can deliver a current density of 150 mA/cm^2 at 0.70 V. The state-of-the-art MCFC utilizes Ni-10 wt% Cr anode, lithiated nickel oxide cathode, and $(Li_2 + K_2)CO_3$ embedded in $LiAlO_2$ is served as electrolyte. Normally, the cell is operated at 650°C at which the electrolytes are in molten stage favoring the hydrogen oxidation and oxygen reduction reactions at the anode and cathode, respectively. The efficiency of MCFC is as high as 50% because of its high operating temperature. The electrodes and the electrolyte structures are prepared by different methods and tape casting process is the most widely accepted method for preparing these components. The typical thickness for the anode is in the range of 0.5–1.0 mm while the same for the cathode is in the range between 0.5 and 1.5 mm. The thickness of the $LiAlO_2$ electrolyte retention matrix ranges between 0.75 and 1.0 mm. Pure nickel was the choice of anode material during the early stages of MCFC development but it has now been replaced by Ni–Cr alloy. Lithiated nickel oxide has been the choice of cathode material. Alternate materials such as $LiCoO_2$, $LiFeO_2$, and Co-coated nickel are being considered as potential cathodes because they show good stability at the cathode operating conditions. Lithium aluminate has been continuously used as an electrolyte retention material because of its high stability in molten carbonate and ability to transport carbonate ions from cathode to anode. This material also has minimum resistance and effectively separates the electrodes with minimum gas crossover. Because of its high corrosion resistance stainless steel is served as both cathode current collector and bipolar plate. Different coatings have been applied on these materials to improve its corrosion resistance especially at the cathode side. The ideal MCFC technology would employ either direct internal reforming (DIR) or indirect internal reforming (IIR) concept, wherein the fuel source can be processed to produce the hydrogen fuel.

MOLTEN CARBONATE FUEL CELL

Molten carbonate fuel cells have been termed as the "second-generation fuel cell" as they have lagged

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039322

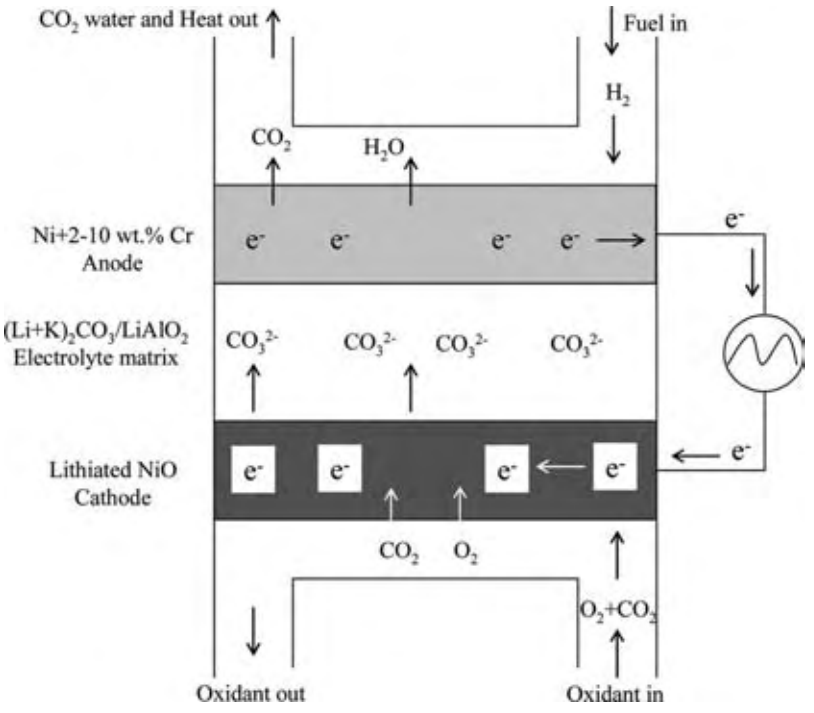


Fig. 1 Schematic representation of molten carbonate fuel cell operation.

behind in developmental efforts when compared to other types of fuel cells.[1–5] Molten carbonate fuel cells consist of nickel anode, nickel oxide cathode, and lithium aluminate as an electrolyte retention matrix. The electrolyte is a mixture of potassium and lithium carbonates. The cell operates at a nominal temperature of 650°C, where the carbonates are in the liquid phase. The molten carbonate is immobilized within a porous ceramic matrix, which takes the form of a planar tile. Electrodes are located above and below the tile to make up a single fuel cell, which can easily be assembled into a stack to produce a useful output voltage.[6]

The electrodes are flat. The anode is composed of porous sintered nickel along with additives, which inhibit the loss of surface area during operation. The anode is in direct contact with the electrolyte matrix. The cathode is a porous nickel oxide, which is initially fabricated in the form of a porous sintered nickel and is subsequently oxidized during the cell operation.

The electrolyte tile is formed from $LiAlO_2$ either by hot pressing or by tape casting methods and filled with $Li_2CO_3 + K_2CO_3$ (62/38 mol% eutectic). A schematic of MCFC is shown in Fig. 1. During the cell operation, the following electrochemical reactions take place at the electrodes.

Anode reaction:

$$H_2 + CO_3^{2-} \rightarrow H_2O + CO_2 + 2e^- \quad (1)$$

The CO_2 produced is transferred to the cathode chamber, where it is reduced along with O_2 to form CO_3^{2-} ion.

Cathode reaction:

$$\tfrac{1}{2}O_2 + CO_2 + 2e^- \rightarrow CO_3^{2-} \quad (2)$$

The carbonate ion takes part in the electrode reactions and the melt remains constant in composition by a continuous transfer of CO_3^{2-} ion from the cathode to the anode.

The overall reaction is:

$$H_2 + \tfrac{1}{2}O_2 \rightarrow H_2O \quad (3)$$

If carbonaceous fuels are used, the CO present in the anode gas chamber oxidizes to CO_2:

$$CO + CO_3^{2-} \rightarrow 2CO_2 + 2e^- \quad (4)$$

Thus, the presence of CO_2 or CO in the fuel stream is not detrimental to the fuel cell performance. The materials currently used for each of the cell components are summarized in Table 1. These components are typically fabricated individually and assembled into alternating layers to form a stack.

THERMODYNAMICS OF MCFC REACTIONS

The overall cell reaction occurring in MCFC can be written as:

$$H_2(g) + \tfrac{1}{2}O_2(g) + CO_{2(\text{cathode})} \rightarrow H_2O + CO_{2(\text{anode})} \quad (5)$$

Table 1 Typical characteristics of the electrode and matrix materials in MCFC

Characteristics	Anode	Cathode	Matrix
Material	Nickel	Lithiated nickel oxide (5 wt%)	$LiAlO_2$
Stabilizer	Chromium	—	—
Reinforcement material	Nickel	Stainless steel mesh	Al_2O_3 and ZrO_2 fibers
Porosity (%)	50–70	70–80	70–75
Mean pore diameter (μm)	3–4	10–15	7–10
BET surface area (m^2/g)	0.1–1.0	0.1–0.5	1.0–20
Thickness (mm)	0.5–1.0	0.5–1.5	0.75–1.0
Other additives	Al_2O_3, MgO, $LiAlO_2$, $LiMoO_2$, $LiCrO_2$, Li_2ZrO_3	Perovskites, Ag, $LaSrNiO_3$	$LiAlO_2$ fiber
Alternate materials	Ni–Al alloy, Ni–Cr alloy, Li_2TiO_3, Cu–Ni alloy	$LiCoO_2$, $Li_2Mn_2O_3$, $LaSrGaO_3$	$SrTiO_3$
Fabrication technique	Compaction, sintering, hot pressing, slurry casting, tape casting	Compaction, sintering, hot pressing, slurry casting, tape casting	Hot pressing, hot rolling, plasma spraying, paper making, tape casting

The reversible voltage for this reaction is given by the equation:

$$E = E^0 + \frac{RT}{2F} \ln \frac{[pH_2][pO_2]^{1/2}}{[pH_2O]} + \frac{RT}{nF} \times \ln \frac{[pCO_2]_{(cathode)}}{[pCO_2]_{(anode)}} \quad (6)$$

When the partial pressure of CO_2 is identical in both compartments and the electrolyte is invariant, the cell voltage depends only on the partial pressures of H_2, O_2, and H_2O. The theoretical reversible voltage E^0 at 650°C calculated by using Eq. (6) is 1.019 V with a theoretical efficiency of $\Delta G/\Delta H$ equal to 79.45%.[7,8] The CO_2 produced at the anode compartment is recycled to the cathode for the continuous formation of CO_3^{2-} ions.

ADVANTAGES OF MCFC

The MCFC because of its high operating temperature has higher efficiency (>50%) and faster electrode kinetics than any other fuel cell system.[9,10] At 650°C, almost a theoretical reversible potential is established at the interface with low electrode overpotentials, which does not require any noble metal catalysts. The CO does not poison the anode, because in the MCFC it is oxidized at the anode interface.

The MCFC operating temperature is also suitable for internal reforming of hydrocarbon fuels such as methane. The heat required for the reforming reaction can be supplied directly from the heat generated in the fuel cell, which minimizes the need for stack cooling devices. These advantages resulted in the development of MCFC stack technology with integrated reformers. The high-grade waste heat can also be used to generate more power by utilizing a steam turbine (bottoming cycle). The MCFC has an efficiency of 50% or above. However, with the addition of the bottoming cycle, the efficiency increases up to 60%. The overall combined heat and power (CHP) efficiency of an MCFC system will probably exceed 60%. This high efficiency and low heat/power ratio make the MCFC ideal for CHP or co-generation.

HISTORY OF MCFC

Bauer (1921) developed the first high-temperature MCFC based on a molten $(Na/K)_2CO_3$ electrolyte, immobilized in a MgO matrix.[4] The MCFC in its present form was developed by G. H. J. Broers in 1951. Small laboratory cells were constructed using a non-sintered MgO + molten carbonate "paste electrolyte." The durability of these cells was tested up to 6 mo. Common features of today's cells and the early Broers' cells are the nickel-based electrodes with planar and bipolar construction and the alkali carbonate electrolyte in inert matrix filler. The basic operating principles are still the same and the cell is represented below.

$$\underset{\text{(Nickel oxide)}}{Air(O_2) + CO_2, Cathode(+)/\gamma} - \underset{\text{(Matrix/electrolyte)}}{LiAlO_2} + \underset{\text{(Ni-based alloy)}}{M_xCO_3/anode(-), fuel(H_2)}$$

Major technological breakthroughs have been made in the electrode materials, electrolyte composition, and fabrication technologies. Argonne National Laboratory (ANL), U.S.A., and Institute of Gas Technology (IGT), U.S.A., are the pioneers in optimizing electrolyte, engineering, and component fabrication. The IGT optimized the parameters for the cell stack and has done modeling optimization studies in collaboration with Physical Sciences Inc. Several theoretical models have been derived for the molten carbonate fuel cathode.[11–21] First principles based theoretical models for MCFC cathode can be divided into the thin film model and the agglomerate model.[11,12] Wilemski assumed that the MCFC cathode could be described as a cylindrical pore covered with a thin film of electrolyte.[11] Gases flowing through the pore dissolve at the surface of the film and diffuse to the surface of the pore and react there. While the model gives good agreement with experimental data, it requires knowledge of the pore diameter, length, and film thickness. Further, the entire description of the electrode is limited and cannot be used for cathode design analysis or two-dimensional simulations. The more common and popular approach for describing the MCFC cathode is the agglomerate model proposed by Yuh and Selman.[12] In this approach, the electrode is assumed to consist of cylindrical agglomerates completely flooded with electrolyte. Gaseous species move through straight cylindrical channels of macropores. A three-phase homogenous model and a full cell model have been developed at the University of South Carolina, U.S.A.[22,23]

Government Industries Research Institute, Osaka (GIRIO), Japan, conducted basic research on MCFC components during the past two decades. Currently, Fuel Cell Energy (FCE), U.S.A., MTU Friedrichshafen, Ansaldo, Italy, and Ishikawajima-Harima Heavy Industries, Japan, are actively pursuing the commercialization of MCFCs.[24]

CELL COMPONENTS

Table 2 provides a chronology of MCFC component technology development.[25] In the early 1960s precious metals such as Pt and Ag were used as anode and cathode, respectively. Continued research in this field resulted in the use of nickel and lithiated nickel as anode and cathode. Until the mid-1970s, there was no significant change in the development of electrodes and the electrolyte matrix. A major development in the 1980s was the evolution in the fabrication of electrolyte structures. The developments in cell components for MCFCs have been reviewed by Maru and co-workers, Petri and Benjamin, and Selman.[26–29] Over the past 30 yr, the performance of single cells has improved from about 10 to $>150\,mW/cm^2$. During the 1980s, major improvements were made in both the performance and the endurance of MCFC stacks. Several MCFC stack developers have produced cell stacks with cell areas up to $1\,m^2$. Recently, FCE, U.S.A., has tested a stack consisting of nearly 300 cells with effective area of $\sim 9000\,cm^2$ delivering >250 kW power. ERC, U.S.A., also tested a stack with $246 \times 5600\,cm^2$ cells and $253 \times 7800\,cm^2$ cells producing 125 and 253 kW, respectively. M-C Power, U.S.A., has tested a stack with $250 \times 1\,m^2$ cells producing 250 kW.[25]

OPTIMIZATION OF MOLTEN CARBONATE FUEL CELL TECHNOLOGY

Attempts were made in recent years related to the development of the component for MCFC. These

Table 2 Evolution of MCFC component technology

Component	ca. 1965	ca. 1975	Current status
Anode	Pt, Pd, or Ni	Ni-10 Cr	Ni–Cr/Ni–Al/Ni–Al–Cr 3–6 μm pore size 45–70% initial porosity 0.20–1.5 mm thickness 0.1–1 m^2/g
Cathode	Ag_2O or lithiated NiO	Lithiated NiO	Lithiated NiO–MgO 7–15 μm pore size 70–80% initial porosity 60–65% after lithiation and oxidation 0.5–1 mm thickness 0.5 m^2/g
Electrolyte support	MgO	Mixture of α-, β-, and γ-LiAlO2 10–20 m^2/g 1.8 mm thickness	γ-LiAlO2, α-LiAlO2 0.1–12 m^2/g 0.5–1 mm thickness
Carbonate electrolyte (mol%)	52 Li-48 Na 43.5 Li-31.5 Na-25 K"paste"	62 Li-38 K hot press "tile" 1.8 mm thickness	62 Li-38 K 60 Li-40 Na 51 Li-48 Na tape casting 0.5–1 mm thickness

Table 3 Issues in MCFC technology

Component	Issues	Results
Electrolyte	Loss due to volatilization	Decrease in the conductivity
	Loss due to filling inside the electrodes	Increase in the resistivity of the layer
	Loss due to reaction with the separator plate	Decrease in the cell voltage Increase in the polarization of electrodes
Matrix	Particle growth	Changes in the microstructure
	Stability and crack formation	Increase in the ionic resistivity
	Dissolution of γ-$LiAlO_2$ in the electrolyte	Decrease in the cell voltage
	Phase transformation from γ to α variety	Decrease in the cell life
Anode	Anode sintering	Change in the pore structure, gas leak through the edges, poor strength
	Anode deformation and creep	Forms a BBP layer
	Extent of electrolyte filling	Increase in the polarization due to increase in the area of the electrode
Cathode	Cathode stability	Cathode shrinkage
	Electrocatalytic activity	Changes in the microstructure
	Cathode dissolution	Deformations at the separator ribs
	Extent of electrolyte filling	Changes in the electrocatalytic activity Increase in the polarization due to increase in the electronic resistivity of the electrode
Separator plates	Corrosion due to reaction with the electrolyte	Decrease in the cell voltage Increase in the resistivity between cells due to increase in polarization

studies are summarized in Table 3.[30] Besides the fact, that the present approaches function properly in full-size cells at atmospheric pressure, research was carried out in addressing alternate cathode materials and electrolytes, performance improvement, life extension beyond the commercialization goal of 5 yr, and cost reduction.[31] These studies provide updated information on the promising materials for electrodes, the electrolyte matrix, and the capability of the cell to tolerate trace contaminants in the fuel supply.

ANODE

The following anode properties were recognized to be of fundamental importance with respect to cell performance: 1) electronic conductivity; 2) electrocatalytic activity; 3) thermodynamic stability; 4) chemical inertness; 5) moderate cost; 6) interconnected porosity; and 7) ease of fabrication.

ANODE PERFORMANCE

Table 1 presents the anode performance characteristics. Efficient fuel cell operation requires that the electrodes must have sufficient electrolytes to maintain the electrode/electrolyte reactant gas interface necessary for the electrochemical reaction to occur. The electrode performance of MCFC depends on the number of pores occupied by the electrolyte. This requires careful control of the overall porosity and the pore size distribution within the electrode structure.[32–36]

The MCFC anodes are made from a porous sintered nickel with a thickness of 0.8–1.0 mm and a porosity of 55–70% with a mean pore diameter of 5 μm.[1] This porosity range provides adequate interconnected pores for mass transport of gaseous reactants and adequate surface area for the anodic electrocatalytic reactions. Because the anode kinetics is faster than that of the cathode, less active surface area is sufficient for the anodic process. Partial flooding of the comparatively thick anode is therefore acceptable at the anode interface.

In MCFC, the three-phase contact, i.e., the contact between the catalyst, electrolyte, and the reacting gas is accomplished by balancing the capillary pressures, which establishes the electrolyte interfacial boundaries in the porous electrodes.[25] Theoretically, the relation between the pores in the cell components is given by the equation:

$$\frac{\gamma_c \cos\theta_c}{D_c} = \frac{\gamma_e \cos\theta_e}{D_e} = \frac{\gamma_a \cos\theta_a}{D_a} \quad (7)$$

where γ is the interfacial surface tension (dynes/cm), θ is the contact angle of the electrolyte, D is the pore diameter (μm), and the subscripts a, c, and e refer to the anode, cathode, and electrolyte matrix, respectively. An optimized electrolyte distribution could be achieved by appropriate coordination of the pore diameters at the anode and cathode with that of the electrolyte retention matrix. This pore size arrangement allows the electrolyte retention matrix to remain completely filled with molten carbonate electrolyte, while the porous electrodes are partially filled, depending on their pore size distribution. Trials have been made to control the electrolyte occupation at the anode.[35] The MCFC anode pores are smaller than the gas-filled (large) pores of the cathode. Ideally, a dual pore structure electrode can be employed, the microporous side containing the molten liquid electrolyte and the macroporous side containing the fuel gas. The matrix should have a pore size that is smaller than that of both the electrodes, so that it is always filled with the electrolyte.

BUBBLE PRESSURE BARRIER

The ideal function of the porous structure in the electrode is to maintain the equilibrium between the gas and electrolyte within these pores of the anode. The fine pore layer of the anode can also be used as an electrolyte reservoir, which prevents gas leakage through the electrolyte.

Alternatively, an additional layer constructed by using fine nickel powder, $LiAlO_2$, and NiO is positioned between the anode and the electrolyte and filled with molten carbonate electrolyte. The purpose of this additional layer is to prevent gas crossover from one electrode to the other if cracks develop in the electrolyte structure. This bubble barrier layer serves as a reinforcement of the electrolyte matrix. This bubble pressure barrier (BPB) can be fabricated as an integral part of the anode structure. Typically, the pores of this barrier layer are smaller than the anode pores and provide ionic transport through the cell.[25]

Various materials have been evaluated for use as anodes in MCFC.[26,37] There are many methods reported in the literature for the preparation of porous sintered nickel structures with adequate porosity and pore size distribution. The oldest method is the gravity sintering method.[38,39] Improvements have been made by incorporating a BPB layer to control the wetting properties of the anode.[40] Organic binders were used during the cold rolling stage, which was removed during the sintering process.[41] Composite coating method was employed to deposit Raney alloys into the nickel matrix.[42]

From the point of view of larger-scale fabrication of MCFCs, these processes for preparing the electrodes may not be efficient. To prepare larger-area electrodes, slurry/slip casting and tape casting methods were adopted as an improvement on the aqueous slurry process.[26,34,43] Aqueous tape casting method employs polyvinyl alcohol or methylcellulose as binder. To obtain fast curing green electrode sheets, a nonaqueous process has also been reported.[44,45] To provide better contact with the electrolyte substrates, preparation of combined tapes by a nonaqueous process and in-cell sintering procedures were adopted.[46,47]

Pure nickel tends to sinter at the fuel cell operating conditions, resulting in loss of surface area and pore growth. Additives have been used to control the sintering. For example, the addition of Cr_2O_3 to nickel has been shown to effectively prevent anode sintering, by the formation of submicrometer $LiCrO_2$ on the nickel surface. Additions of Co, Cr, and Cu metals to nickel powder have also been tried and a Ni-10% Cr alloy has become the standard anode.[25]

The anode structure is also susceptible to deformation (creep) under compressive load during normal stack operation (i.e., shrinkage of anode thickness under load). Anode creep results in decreased porosity, increased contact resistance, and leaks. To stabilize the anode structure, addition of oxides such as Al_2O_3, $LiAlO_2$ and ZrO_2 has been reported in the literature. The oxide dispersion strengthened anodes have demonstrated encouraging results toward minimizing the creep. The kinetics of the anode reaction was not affected. Recent reports indicated that addition of Al, Cu, and Ti also shows encouraging results in retarding the anode creep.[30]

The anode is stabilized by using mixed powders of Ni–Cr, Ni–Al, or refractory oxides and sintered at 900–1100°C under a reducing atmosphere to provide a creep resistant anode structure.[43,48] The function of the additives is to reduce the loss of porosity during sintering and develop creep resistant materials. The creep is referred to as the shrinkage in thickness and change in shape. The sintering resistance is increased by the additives, which are usually metals or oxides of metals.

The sintering of the anodes can be minimized by stabilizing the anode with Al/Cr.[49] The addition of aluminum and chromium inhibits porous Cu–Ni anodes from being sintered because of the formation of aluminum oxide and chromium oxide on their surfaces.[50] Addition of Cr or Cr_2O_3 prevents the anode sintering by formation of submicrometer size $LiCrO_2$ on the nickel surface. Ni alloy with 10 wt% Cr has become the standard anode.[51] Heat treatment after impregnation caused a decomposition of the metal salts and resulted in finely dispersed oxide particles on the surface of the sintered electrodes. This stabilization method effected lower porosity changes of the anode in MCFC operating condition. Also, addition of ceramic oxides such as $LiAlO_2$ stabilized the anode structure. Ni powder with 20% $LiAlO_2$ resulted in stabilization of the electrode

exposed to further compaction by sintering and retained the initial porosity.

CATHODE

The primary challenge in commercialization of MCFC remains in the proper selection of materials for the cathode. The life expectancy of the electrode structure is aimed toward 40,000 hr for successful commercialization of MCFC.[52] The following cathode properties were recognized as of fundamental importance with respect to the cell performance: 1) high electronic conductivity at 650°C ($\sigma > 1$ S/cm); 2) low chemical reactivity and solubility in the electrolyte; 3) thermodynamic stability at 650°C in carbonate electrolyte at different partial pressures of O_2/CO_2 mixtures; 4) high electrocatalytic activity for the oxygen reduction reaction; and 5) suitability for the fabrication of porous electrodes.[26]

The most widely used cathode material is nickel oxide doped with 3 atomic percent (at%) of Li to provide the required electrical conductivity and electrocatalytic activity. The cathode structures were fabricated by sintering either metallic nickel or nickel oxide into 0.4 mm-thick porous plaque. In situ oxidation results in electrochemically conducting ceramic $Li_xNi_{1-x}O$, where x is in the range of 0.022–0.04. At 650°C, the resistivity of the lithiated nickel oxide is about 0.20 Ω cm. Typical cathodes have a porosity of 55% and a mean pore size of about 10 μm. They usually develop a bimodal pore distribution. The small pores are flooded with electrolyte, thereby providing an extended reaction surface and ionic conduction. The larger pores facilitate mass transport in the gaseous phase. Because the lithiated nickel oxide is completely wetted with CO_3^{2-} electrolyte, the agglomerates are covered with a thin film of electrolyte. Gases diffuse through the film to react at the electrode. Thus, it is important to avoid flooding of the entire electrode. The cathode performance is sensitive to the thickness of the electrode and the degree of filling of pores with the electrolyte. Ohmic losses in both the liquid and solid phases increase as the cathode thickness increases. Diffusion losses in the gas phase also increase with cathode thickness. On the other hand, activation and liquid phase diffusion losses decrease as the cathode thickness increases. The optimum value was identified at a thickness of 0.5 mm, when the total losses are at a minimum.[25]

ISSUES IN CATHODE DEVELOPMENT—NiO DISSOLUTION

The major problem identified during testing of MCFC for long durations exceeding 1000 hr of operation is the dissolution of NiO cathodes and accumulation of Ni nodules in the matrix toward the anode. These phenomena result in a cell shorting. There is an excellent review in the literature that extensively covers this problem.[53] NiO dissolves in molten carbonate electrolyte (62–75 mol% Li^+) slowly to about 10–15 ppm according to:

$$NiO + CO_2 \longrightarrow Ni^{2+} + CO_3^{2-} \quad (8)$$

The NiO solubility increases with increasing temperature, pCO_2, and cation fraction Li^+.[54] The dissolved Ni^{2+} ions diffuse through the electrolyte toward the anode, where it is reduced and precipitated as metallic nickel. This causes a sink for the Ni^{2+} ions, which facilitates further NiO dissolution. The lifetime of MCFC is significantly reduced by the slow dissolution of nickel oxide cathode and precipitation of Ni ions in the matrix.

ALTERNATE CATHODE MATERIALS

To solve the problems associated with the instability of state-of-the-art NiO cathodes, research has been focused on the development of new stable cathode materials to replace NiO. Currently, perovskite types of compounds and mixed metal oxides such as $LiFeO_2$ and $LiCoO_2$ have been evaluated as cathode materials.[55–67]

Preliminary work on alternate $LiFeO_2$ cathodes showed better chemical stability under the cathode environment and a lower dissolution rate.[25] However, because of slower kinetics, the catalytic activity of this material for oxygen reduction is very poor when compared to the state-of-the-art NiO cathode. Because $LiFeO_2$ shows better performance than NiO under pressurized operation it is still under consideration to be used as a cathode material.[25]

NiO doped with 5 mol% Li showed only 43 mV overpotential and higher performance at 160 mA/cm^2 compared to the state-of-the-art NiO cathode.[25] Performance improvements were suggested with Co-doped $LiFeO_2$ and Ni–Fe–Co ternary alloys.[26,68] NiO coated with CoO or $LiCoO_2$ or mixed oxides such as $LiFe_{0.5}Co_{0.5}O_2$ and mixture of $LiFeCoNiO_2$ electrodes have also been suggested as cathodes for MCFC.[69–71] Also, lithium (Li) content was found to affect the solubility of either prelithiated NiO or prelithiated cobaltite.[72]

Other materials such as Ni–Ti alloys, cerium incorporated NiO, lanthanum impregnated NiO, electroless cobalt encapsulated NiO, cobalt-doped lithium nickel oxides, and $La_{0.8}Sr_{0.2}CoO_3$ coated NiO were also suggested as alternate cathode materials.[73–78] The electrochemical cell shown in Fig. 2 made of

alumina tubes was used for the evaluation of different cathode materials under cathode operating conditions.

ELECTROLYTE

The MCFC electrolyte structure is referred to as a ''tile'' or ''matrix.'' It consists of a mixture of alkali carbonates supported within a porous support (matrix) made from finely divided $LiAlO_2$ or ceramic particles, by capillary action.[79] It contains about 45% by volume of the molten carbonate electrolyte, in most cases an eutectic mixture of 62 mol% Li_2CO_3 + 38 mol% K_2CO_3 (m.p. 487°C). Binary and ternary electrolyte compositions have also been investigated.[80] Typical compositions are 52 mol% Li_2CO_3 + 48 mol% Na_2CO_3 (m.p. 500°C) and 43.5 mol% Li_2CO_3 + 31.5 mol% Na_2CO_3 + 25 mol% K_2CO_3 (m.p. 397°C). Electrolyte compositions may be varied within the ternary system to tailor properties such as ionic conductivity, melting point, vapor pressure, gas solubility, surface tension, coefficient of thermal expansion, and corrosivity.

Electrolyte optimization is a key to MCFC lifetime.[30] The electrolyte composition affects the cell performance via: 1) tile resistance, which depends on the ionic conductivity, and 2) the polarization of the electrodes. The latter depends primarily on the electrode kinetics and gas solubility in the electrolyte.

High Li^+ and Na^+ ion contents in the melt result in higher electrical conductance of the electrolyte, while high K^+ content promotes the gas solubility.[79] The conductance measurements with various binary and ternary systems indicate that the ternaries of Li, Na, and K carbonates can also be successfully used as alternate electrolytes.[81] Typical conductance values of various electrolytes are as follows:[82]

62 mol% Li_2CO_3 + 38 mol% K_2CO_3	2.513 S/cm
52 mol% Li_2CO_3 + 48 mol% K_2CO_3	3.53 S/cm
60 mol% Na_2CO_3 + 40 mol% K_2CO_3	2.351 S/cm

Addition of Na^+ lowers the melting point of the eutectic, which is advantageous because of the fact that it lowers the vapor pressure and the expansion coefficient. Li_2CO_3/Na_2CO_3 eutectic electrolyte is expected as a new candidate electrolyte for MCFC, because it offers higher ionic conductivity and lower NiO cathode dissolution rate than the Li_2CO_3/K_2CO_3 eutectic electrolyte.[83] It was found that Li/Na electrolyte has the following superiority over Li/K:[84–88] 1) high ionic conductivity; 2) low volatility; 3) low solubility of the electrode materials; 4) no electrolyte segregation as observed with Li/K electrolyte; 5) no phase change and no particle growth of $LiAlO_2$; 6) no degradation of the cell components; 7) low kinetics at low pO_2; 8) small power density enhancement; and 9) good performance at high pressures. Addition of Ba/Ca salts lowers the melting point and increases the life stability. However, the O_2 solubility and the electrode kinetics is lower in Li/Na when compared with that in Li/K.[89]

The cell containing $(Li_{0.75}Cs_{0.25})CO_3$ electrolyte has higher cell voltage and lower cathode and anode overpotentials and it shows smaller dependence on operating temperature.[90] Li/Na/Ba/Ca electrolytes ranging from 3% to 5% Ba/Ca have shown stable and long-term performance.[91] Addition of alkaline

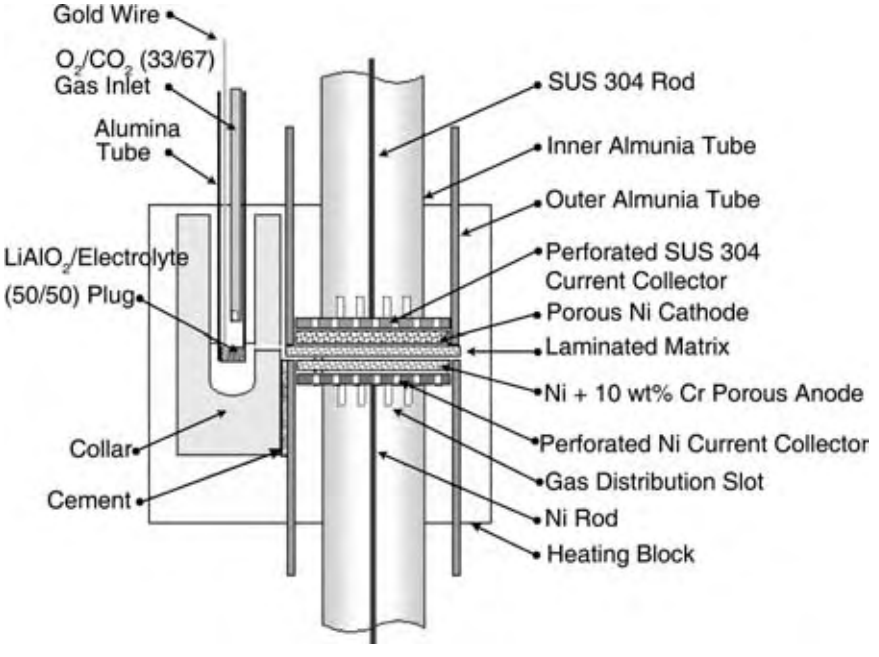


Fig. 2 Electrochemical cell used for the evaluation of MCFC cathode materials. *(View this art in color at www.dekker.com.)*

earth carbonates such as $SrCO_3$, $BaCO_3$, and $CaCO_3$ to the MCFC electrolyte increases the cathode stability and the cell performance.[92]

ELECTROLYTE RETENTION MATRIX

The electrolyte matrix provides: 1) transport of CO_3^{2-} ions from the cathode to the anode within its pore structure with a minimum resistance, and 2) separation between the fuel and oxidant gases with a minimum crossover through the voids. The matrix should satisfy the following requirements: 1) the matrix material must be chemically inert against molten carbonates, adjacent cell components, and reactant gases, and must be morphologically stable; 2) the matrix material must be of minimum crystallite size, consistent with stability; and 3) the matrix must be porous having appropriate pore size distribution. Typically, the pore size distribution should be in the range of 0.1–0.5 μm to control the distribution of CO_3^{2-} ion. The electrolyte retention matrix material should be electronically insulating and should not dissolve in the molten alkali carbonate melt.[93] It should possess porosity of more than 40%, which will enable complete flooding by the electrolyte.

The strength of the matrix and the electrolyte structure depends on the relative amount of carbonate and $LiAlO_2$. At low carbonate contents, the structure is rigid. Currently, 40 wt% of $LiAlO_2$ and 60 wt% carbonate mixtures are used to form the matrix. At the fuel cell operating temperature, the electrolyte structure is a thick paste, which provides gas seals (called the wet seal) at the edges of the cell.

To satisfy these requirements, the matrices are made of MgO, Al_2O_3, and $LiAlO_2$. The allotropic form γ-$LiAlO_2$ was found to be thermally stable and hence widely used. The status of MCFC matrix materials and their characteristics are summarized in Table 1.

Three allotropic forms of $LiAlO_2$ (α, β, and γ) have been reported.[94] Both α and β-$LiAlO_2$ undergo an irreversible transformation to γ-$LiAlO_2$ in the presence of Li_2CO_3/K_2CO_3 (62/38 mol%) at high temperatures. The phase changes are accompanied by a change in particle morphology and a decrease in surface area. Therefore, γ-$LiAlO_2$ appears to be the most desirable crystalline form for MCFC applications. The size, the shape, and the distribution of particles of $LiAlO_2$ control the carbonate retention, mechanical properties, and effective ionic conductivity of the electrolyte structures. A matrix with a narrow pore size distribution with a small mean pore size (<0.2 μm) can be attained from this powder and is considered quite acceptable. The most effective shape for $LiAlO_2$ particles appears to be fibers of submicrometer diameter.

$LiAlO_2$ powders have been prepared by using three different methods. The solid-state high-temperature reaction was developed at ANL, U.S.A.[95,96] This method involves the reaction between γ-Al_2O_3 and Li_2CO_3 at 600°C for 10 hr, leading to the formation of α-$LiAlO_2$, which was homogenized and subjected to further high-temperature treatment at 1000°C when it is converted into γ-$LiAlO_2$. The following reaction sequences are reported.[95,96]

$$\gamma\text{-}Al_2O_3 + Li_2CO_3 \xrightarrow{10\,hr,\ 600^\circ C} \alpha\text{-}LiAlO_2 \quad (9)$$

$$\alpha\text{-}LiAlO_2 \xrightarrow{24\,hr,\ 1000^\circ C} \gamma\text{-}LiAlO_2 \quad (10)$$

The α-$LiAlO_2$ thus obtained was generally of very high surface area (57–69 m^2/g), which reduces to a low-surface-area (0.4–10 m^2/g) γ-$LiAlO_2$ after the heat treatment. General Electric (GE) reported two procedures, namely aqueous slurry process and chloride synthesis routes.[4]

ALTERNATE MATERIALS

Strontium titanate ($SrTiO_3$) is recommended as an alternate material for MCFC matrix and GE has carried out extensive studies with $SrTiO_3$ tiles.[4,97] It was reported that $SrTiO_3$ reacted with Li_2CO_3 to form Li_2TiO_3 and $SrCO_3$. This reaction is inhibited by the addition of $SrCO_3$ to the carbonate melt, resulting in lowering the melting point of the melt. Unfortunately, no further studies were done either to characterize this material or to identify a more suitable matrix material than $LiAlO_2$ for MCFC.[98]

Other alternate materials such as Ga_2O_3, In_2O_3, SnO_2, PbO, Bi_2O_3, Ti_2O_3, and lithium double oxides such as $LiGaO_2$, $LiSnO_2$, and Li_2SnO_3 and zirconia were also reported in the literature.[99,100] All these materials showed a very poor performance. Thus, currently γ-$LiAlO_2$ is continuously used as matrix material for MCFC application.

ELECTROLYTE RETENTION MATRIX PREPARATION

Several researchers have reviewed the state-of-the-art and advances in MCFC matrix technology from time to time.[83,96,101] Simple, cost-effective, and scalable techniques were demonstrated for the preparation of γ-$LiAlO_2$ and to fabricate electrolyte retention matrices.[102,103] Hot pressing of the support material and electrolyte mixture at a temperature 5–10°C below the melting point of the carbonate phase was the standard method of fabrication of a carbonate fuel cell matrix (tile) in the 1980s.[104–106] Typically, tiles were hot pressed at 5000 psi and contained more than

50 vol% of the electrolyte. They were relatively thick with thickness varying from 1.8 to 2.5 mm.

Because of the inherent functional and manufacturing problems with the hot pressed tiles, alternate fabrication procedures like hot rolling, cold pressing, and sintering followed by impregnation were adopted.[107–109] Other methods such as cold pressing followed by sintering and low-pressure compression molding were also reported to produce crack-free electrolyte matrix support.[108,109] Tiles of large sizes are very difficult for production by these techniques. The major drawback of this process is that the tiles were characterized by 1) poor uniformity of microstructure; 2) low porosity; and 3) high IR drop.

Alternate methods such as vacuum casting, cold and hot roll milling flow cast technique, and paper making were adopted for the fabrication of electrolyte tile.[110–115] Electrophoretic deposition offers fabrication of very thin electrolyte layers (0.25–0.50 mm).[116] The process involved dispersion of very finely divided ceramic powder in an organic liquid and deposition onto an electrode under a d.c. electric field. This technique has also been proved to be less promising because of poor adhesion to the electrodes, matrix cracking, and lack of reproducibility. Although both the technique and the deposited product were initially well suited for MCFC applications, structural fracture resulted during impregnation with the alkali carbonate electrolyte.

TAPE CASTING TECHNIQUE

The tape casting process, which is utilized widely in the electronics industry, has been adopted by many MCFC developers.[117,118] The tape casting process was originally introduced for matrix fabrication, because of its suitability for low-cost processing of very thin structures with large areas.[101] The tape casting process was demonstrated to be advantageous by IGT, Energy Research Corporation (ERC), and United Technologies Corporation (UTC).[38,119] This process involves dispersing the ceramic powder in an organic vehicle (solvent) along with binders, plasticizers, and additives to yield proper slip rheology. As presented in Fig. 3, a typical tape casting process begins by dispersing the dried ceramic powders in solvents containing a dissolved organic deflocculant.

Although the basic formulation of the tape casting slurry and the tape casting process variables is known, optimization of these variables is desired. The current activities are focused to improve the performance of the carbonate fuel cell matrix manufactured by the tape casting method. The crack propagation was arrested by the use of Al_2O_3 or $LiAlO_2$ fibers. Various additives such as submicrometer size $LiAlO_2$ or ceramic fiber have been reported for reinforcement to avoid matrix cracking.[120,121]

MATRIX STABILITY

It is important to improve the endurance and the reliability of the electrolyte plate for the commercialization of MCFC. The electrolyte loss from the matrix plate increases the cell resistance and deteriorates the cell voltage. The formation of cracks in the electrolyte plate causes a gas cross-leakage between the fuel gas and the oxidizer gas. The pore structure of the matrix plate must be stable and fine to support liquid electrolyte under MCFC operation. It is necessary to prevent the formation of cracks in electrolyte plates during thermal cycling. Because of the immediate and large

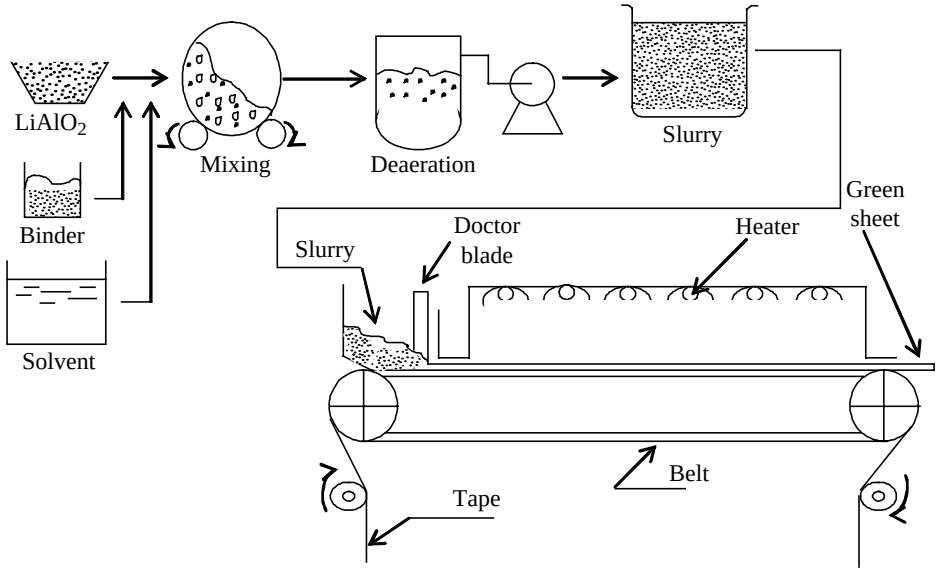


Fig. 3 Tape casting process—a schematic.

performance loss associated with gas crossover, resistance to thermal cycling is considered a major critical issue for stack lifetime. The stability of electrolyte plate has been improved by using advanced $LiAlO_2$ powder and its durability by the addition of the ceramic fiber.[122,123]

ELECTROLYTE RETENTION CAPACITY

To prolong the life of MCFC, the amount of electrolyte in the matrix must be maintained at an appropriate level over long-term operation. The growth of particles of $LiAlO_2$ as an electrolyte retention material in molten carbonates leads to a decrease in the electrolyte retention ability. These phenomena result in a decrease of the fuel cell performance. It was found that zirconia powder added to lithium aluminate keeps the electrolyte retention ability constant for over 7000 hr in Li/Na carbonates and $pCO_2 = 0.1$.[124]

ELECTROLYTE LOSS AND MANAGEMENT

The causes that trigger the electrolyte loss from the cell packages are vaporization and reaction of the electrolyte with the metallic components.[25] Corrosion of cell metallic parts consumes Li^+ ions. Li^+-rich compositions tend to have a very good performance. Carbonate creepage, i.e., migration of electrolyte from one end of the stack to the other through the external manifolds, gaskets, etc. is another cause. Engineering solutions are aimed to solve this with a proper stack design and by manifold concept.

Electrolyte loss by volatilization is a continuous phenomenon. This can be overcome by using Na^+ containing electrolyte. Also, use of thicker electrodes can store more electrolytes in it. The effect of matrix thickness and cathode pCO_2 on cell shorting has also been discussed.[125] The relationship between cell life and matrix thickness is not linear. Cells employing thick matrix have longer life. On the other hand, the resistance of the thick matrix is high.[26] It is important to determine the matrix thickness in consideration of the cell life and performance. Also, the use of corrosion resistant materials in cathode development will reduce the carbonate loss. There is a tendency for the electrolyte to migrate from the positive end of the stack to the negative end of the stack. This may cause the end cells to perform poorly compared to the central cells. The electrolyte loss is through the gasket used to couple the external manifolds to the cell stack. The standard gasket material is porous and provides a conduit for electrolyte transfer. Stacks with internal manifolding do not require a gasket and may not experience this problem.[25]

CURRENT COLLECTORS AND BIPOLAR PLATES

Austenitic stainless steels like 310S, 316, or 316L are typically used for the construction of cathode and anode current collectors and bipolar separator plates.[126–128] Corrosion of these steel components is a major lifetime-limiting factor in MCFC. The corrosion behavior of stainless steel components in molten carbonate conditions has been studied extensively during the past decade. Research is being aimed at increasing the corrosion resistance of these components by altering the alloy composition or by surface modification techniques.[126,129–132]

The corrosion resistance of stainless steels and nickel-based alloys in aqueous solutions can often be increased by addition of chromium or aluminum.[133–135] Chromium protects the base metal from corrosion by forming an oxide layer at the surface. Chromium is also considered to be an important alloying metal for steels in MCFC applications. Chromium containing stainless steel, however, leads to the induced loss of electrolyte. Previous studies done to characterize the corrosion behavior of chromium in MCFC conditions have shown the formation of several lithium chromium oxides by reaction with the electrolyte.[133] This corrosion process also results in increased ohmic loss because of the formation of scales on the steel. Aluminum additions similarly have a positive effect on corrosion resistance.[134,135] However, corrosion scales formed in aluminum containing alloys show low conductivity leading to a significant ohmic polarization loss.

Nickel electrocladding is popularly used as a surface modification technique to reduce corrosion on the anode side of the bipolar plate. However, nickel electroplated parts do undergo corrosion. Also, nickel cladding does not offer good barrier protection to the diffusion of oxygen. Alloying, or coating with aluminum or alumina, is often reported to improve corrosion resistance. Aluminum coating typically gives lifetime protection in the wet-seal area.[135] However, the high electrical resistance of an Al_2O_3 or $LiAlO_2$ scale is not acceptable in the active area.

The nature of the corrosion scale is very important in deciding the corrosion resistance of the component. Surface modification or alloying alters the composition and complexity of the corrosion scales, which prevents further corrosion and outward diffusion of alloy components. Vossen et al. have shown the complex scale formation in the case of stainless steel and nickel-based alloys in comparison to pure metals.[133,136–142] These scales are understood to cause transport problems to the outward diffusion of alloy components and thereby decrease the dissolution of the steel components. Therefore, any improvement to the corrosion resistant properties of the current collector has to come by way

of novel coatings, which possess the ability to alter the corrosion scale structure of the alloy component. By using this approach, the barrier properties of the corrosion scale to the oxygen diffusion can be improved. Also, the coating should restrict the outward diffusion of steel components, a common problem associated with Ni coating.

Nickel containing scales exhibit higher conductivity because of the presence of trivalent nickel ions, which introduce vacancies in the lattice of the scale.[136] Therefore, nickel-based coating can lead to superior conductivity and good protection provided that it is alloyed properly with corrosion resistant elements. Cobalt has a lower solubility in molten carbonate and electroless Co has been successfully used for a variety of corrosion-resistant applications.[82,143] Electroless plating of Ni–Co gives rise to deposition of uniform layers of nanostructured material, which would result in better protection of the substrate.

INTERNAL REFORMING

The MCFC operates at sufficient high temperatures to allow natural gas to be reformed within the fuel cell stack itself. This involves addition of a reforming catalyst in the fuel gas passages adjacent to the anode, where the reforming reaction takes place.[144] This concept is called DIR step.[25] A direct fuel cell (DFC) is a unique type that employs the DIR concept for MCFC (IRMCFC), eliminating the need for an external reformer and the associated heat transfer loop. Heat and steam are supplied directly by the electrochemical reaction and this reaction drives the reforming reaction to completion by removing the hydrogen as it is formed. The water gas shift reaction:

$$CO + H_2O \longrightarrow CO_2 + H_2 \qquad (11)$$

also occurs, i.e., reaction of the carbon monoxide with steam to produce additional hydrogen. In this way an external reformer is not needed. Other advantages claimed for the IRMCFC are greater system simplicity, multifuel capacity, increased reliability, which can lead to a 30% cost reduction. Fuel to electricity conversion efficiencies greater than 60% can be achieved. But the major disadvantage is that the catalyst can become deactivated in the presence of carbonate, which migrates out of the cell.[25]

In another alternative, in IIR mode, the reforming reaction takes place in a separate catalyst compartment located in close thermal contact with the cell. This method allows the heat from the fuel cell to be used for the reforming reaction, but there is no hydrogen removal or steam production to help move the reaction forward. Indirect reforming therefore is not as efficient as direct reforming but it does have the advantage that the catalyst is less likely to become deactivated.[25] A combination of both methods of internal reforming may give optimum cell characteristics for the longest cell life.

In the direct internal reforming MCFC (DIR-MCFC) system, the direct reformation of fuel at the anode can give fuel saving of 20%, resulting in 12% improvement in fuel cell electrical efficiency. The schematic representation of the IIR and DIR concepts is shown in Fig. 4.

Simple molecules such as methanol can be directly fed to a carbonate fuel cell anode. Hydrocarbon molecules, such as methane, can be reformed in situ with the help of a catalyst.

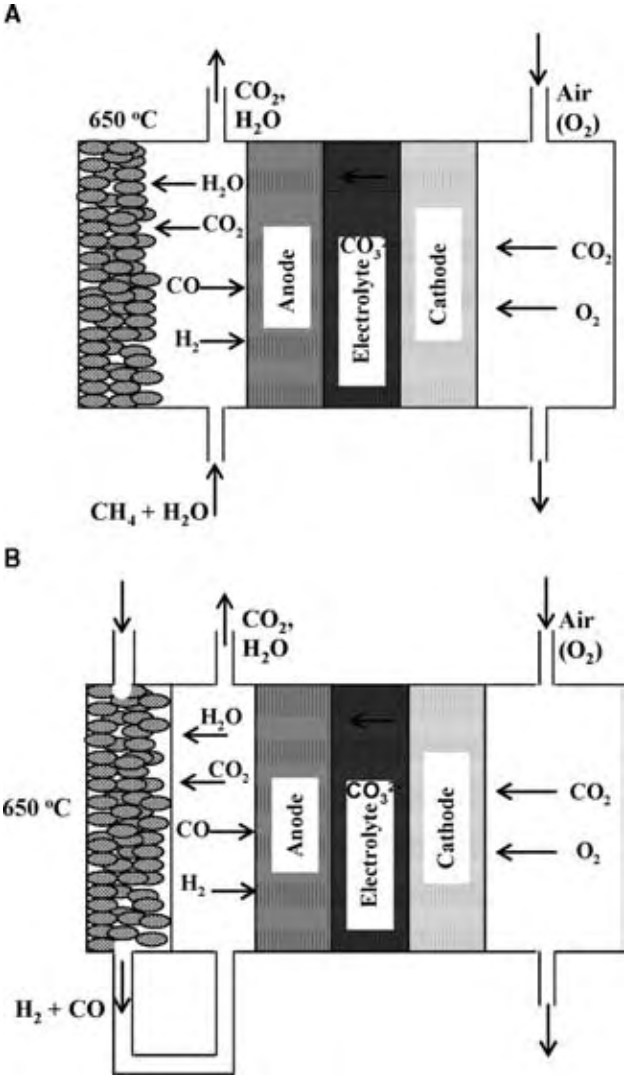


Fig. 4 Schematic representation of (A) direct internal reforming (DIR) and (B) indirect internal reforming (IIR) MCFC concepts.

The reformation reaction is:

$$CH_4 + H_2O \longrightarrow CO + 3H_2$$
$$H_{650^\circ C} = +53.87\,\text{kcal/mol} \quad (12)$$

The H_2 thus formed then reacts in the cell as:

$$H_2 + \frac{1}{2}O_2 \longrightarrow H_2O(g)$$
$$H_{690^\circ C} = -59.26\,\text{kcal/mol} \quad (13)$$

The reforming reaction is endothermic, whereas the overall cell reaction is exothermic. It is estimated that for typical operating parameters, approximately 60% of the heat produced by the fuel cell can be consumed for the reforming reaction.[144]

Different nickel-based catalysts on ceramic supports have been extensively investigated as internal reforming catalysts. Ni/MgO or Ni supported or Al_2O_3 provide sufficient catalytic activity for the steam reformation reaction.[145] It possesses the stability and converts about 85% CH_4 to H_2 at 650°C. But it is deactivated by the alkali metal cations of the electrolyte. No thorough study has been made on the suitability of available catalysts besides nickel. Ni–cermet anodes are suggested as alternatives. Nickel catalyst supported with γ-$LiAlO_2$ was also used to study the steam reforming of methane in internally reformed MCFC.[146,147]

Catalyst support such as Co/Al_2O_3 and Rh/Al_2O_3 were used for the production of hydrogen from ethanol for internal reforming MCFC application.[148] Research is also focused on the use of nickel catalyst supported on $MgO–TiO_2$ composite oxide with varying ratio of Mg to Ti.[149]

CONCLUSIONS

The MCFC is a promising power generating source because of its unique characteristics such as high fuel efficiency and ability to use various carbonaceous fuels. Although Ni-10 wt% Cr is used in the state-of-the-art MCFC as anode, it needs to be improved in terms of better creep and sintering resistance. In spite of the development in the alternate cathode material research, lithiated nickel oxide has been the choice of cathode material in the kilowatt-level MCFC stacks developed by many companies. Continuous research in the development of stable electrolyte retention matrix, identification of suitable molten carbonate electrolyte composition, and additives to the electrolyte will be a significant milestone. Also, research in the area of current collector/bipolar plate to overcome the corrosion problem will lead to the MCFC technology hitting the commercial market in the near future.

M

REFERENCES

1. Penner, S.S. *Assessment of Research Needs for Advanced Fuel Cells*; DOE/ER/30060-T; U.S. Department of Energy: Washington, DC, 1985.
2. Department of Energy. Fuel Cells Technology Status Report, DOE/METC/87-0257; U.S. Department of Energy: Washington, DC, 1986.
3. Institute of Gas Technology. Symposium Papers on Fuel Cells—Status and Outlook, IGT: U.S.A, 1982.
4. General Electric Company. *Development of Molten Carbonate Fuel Cell Power Plant*, Final Report for U.S. DOE/Contr. DE-A02 80ET/7019, 1985.
5. Blomen, L.J.M.J.; Mugerwa, M.N. *Fuel Cell Systems*; Plenum Press: New York, 1993; 345.
6. Kordesch, K.; Simader, G. *Fuel Cells and Their Applications*; Weinheim: New York, 1996; 111.
7. Benjamin, T.G.; Camara, E.H.; Marianowski, L.G. *Hand Book of Fuel Cell Performance*; Institute of Gas Technology: Chicago, IL, 1980.
8. Kinoshita, K.; Mcharnow, R.F.; Cairns, E.J. *Fuel Cells—A Hand Book*; U.S. DOE/METC-88/6096/ 1988.
9. Schora, F.C.; Camara, E.H. Fuel Cells Power for the Future—Energy and Environment for the 21st Century Proceedings, Cambridge, MA, Mar 26–28, 1990; Tester, J.W., Wood, D.O., Ferrari, N.A., Eds.; M.I.T. Press: Cambridge MA, 1991; 959–971.
10. Masters, J.; Dicks, A.L. The Potential and Market Penetration of MCFC. 18th World Gas Conference, Berlin, Jul 9, 1991 (published 1992).
11. Wilemski, G. Simple porous electrode models for molten carbonate fuel cells. J. Electrochem. Soc. **1983**, *130* (1), 117–120.
12. Yuh, C.Y.; Selman, J.R. Polarization of the molten carbonate fuel cell anode and cathode. J. Electrochem. Soc. **1984**, *131* (9), 2062–2068.
13. Prins-Jansen, J.A.; Hemmes, K.; de Wit, J.H.W. An extensive treatment of the agglomerate model for porous electrodes in molten carbonate fuel cells—I. Qualitative analysis of the steady-state model. Electrochim. Acta **1997**, *42* (23–24), 3585–3600.
14. Kunz, H.R.; Murphy, L.A. The effect of oxidant composition on the performance of molten carbonate fuel cells. J. Electrochem. Soc. **1988**, *135* (5), 1124–1131.

15. Fontes, E.; Lagergren, C.; Simonsson, D. Mathematical modelling of the MCFC cathode. Electrochim. Acta **1993**, *38* (18), 2669–2682.
16. Christensen, P.S.; Livbjerg, H. A new model for gas diffusion electrodes. Application to molten carbonate fuel cells. Chem. Eng. Sci. **1992**, *47* (9–11), 2933–2938.
17. Fontes, E.; Fontes, M.; Simonsson, D. Effects of different design parameters on the performance of MCFC cathodes. Electrochim. Acta **1996**, *41* (1), 1–13.
18. Prins-Jansen, J.A.; Fehribach, J.D.; Hemmes, K.; de Wit, J.H.W. A three-phase homogeneous model for porous electrodes in molten carbonate fuel cells. J. Electrochem. Soc. **1996**, *143* (5), 1617–1628.
19. Lee, G.L.; Selman, J.R.; Pomp, L. Comparison of MCFC cathode materials by porous electrode performance modeling. J. Electrochem. Soc. **1993**, *140* (2), 390–396.
20. Fontes, E.; Lagergren, C.; Simonsson, D. Mathematical modelling of the MCFC cathode on the linear polarisation of the NiO cathode. J. Electroanal. Chem. **1997**, *432* (1–2), 121–128.
21. Fontes, E.; Fontes, M.; Lindbergh, G.; Simonsson, D. Influence of gas phase mass transfer limitations on molten carbonate fuel cell cathodes. J. Appl. Electrochem. **1997**, *27* (10), 1149–1156.
22. Subramanian, N.; Haran, B.S.; Ganesan, P.; White, R.E.; Popov, B.N. Analysis of molten carbonate fuel cell performance using a three-phase homogeneous model. J. Electrochem. Soc. **2003**, *150* (1), A46–A56.
23. Subramanian, N.; Haran, B.S.; White, R.E.; Popov, B.N. Full cell mathematical model of a MCFC. J. Electrochem. Soc. **2003**, *150* (10), A1360–A1367.
24. http://www.dt.navy.mil/ip/mfp/Acrobat%20Files/Technical%20Papers/References/newref06.pdf. (accessed Dec 18 2004).
25. EG&G Services Parsons, Inc. *Fuel Cell Handbook*, 6th Ed.; Science Applications International Corporation, Under Contract No. DE-AM26-99FT40575, U.S. Department of Energy, Office of Fossil Energy, National Energy Technology, Laboratory: Morgantown, West Virginia, Nov 2002.
26. Pigeaud, A.; Maru, H.C.; Paetsch, L.; Doyon, J.; Bernard, R. Recent developments in porous electrodes for molten carbonate fuel cells. Proceedings of the Symposium on Porous Electrodes: Theory and Practices, Maru, H.C., Katan, T., Klein, M.G., Eds.; The Electrochemical Society, Inc.: Pennington, NJ, 1984; 234–259.
27. Maru, H.C.; Paetsch, L.; Pigeaud, A. Review of molten carbonate fuel cell matrix technology. Proceedings of the Symposium on Molten Carbonate Fuel Cell Technology, Selman, J.R., Claar, T.D., Eds.; The Electrochemical Society, Inc.: Pennington, NJ, 1984; 20–53.
28. Petri, R.J.; Benjamin, T.G. Molten carbonate fuel cell component design requirements. Proceedings of the 21st Intersociety Energy Conversion Engineering Conference, American Chemical Society: Washington, DC, 1986; Vol. 2, 1156–1162.
29. Selman, J.R. Molten carbonate fuel cells (MCFCs). Energy **1986**, *11* (1–2), 153–208.
30. Pattabiraman, R. Molten carbonate fuel cells. Proceedings Two Day Conference on Fuel Cell Technology, Applications and Opportunities, Engineers India Ltd.: Gurgaon, New Delhi, Nov 26, 27, 2001.
31. Farooque, M.; Maru, H.C. Carbonate Fuel Cells Overview. Preprints of Symposia, American Chemical Society, Division of Fuel Chemistry, 2000; Vol. 46 (2), 440–442.
32. Iacovangelo, C.D. Electrolyte loss and performance decay of molten carbonate fuel cells. J. Electrochem. Soc. **1986**, *133* (2), 280–289.
33. Lanzi, O.; Landau, U. Effect of pore structure on current and potential distributions in a porous electrode. J. Electrochem. Soc. **1990**, *137* (2), 585–592.
34. Baumgartner, C.E. Controlled-pore-size composite nickel oxide structures for carbonate fuel cell cathodes. J. Am. Ceram. Soc. **1990**, *73* (3), 516–520.
35. Mitsushima, S.; Okada, H.; Takeuchi, M.; Nishimura, N. Lifetime modeling for molten carbonate fuel cells. Denki Kagaku **1992**, *60* (10), 906–911.
36. Iwase, Y.; Okada, H.; Kuroe, S.; Mitsuishima, S.; Takeuchi, M. Enhancement and stabilization of the molten carbonate fuel cell performance by optimization of electrode pore distributions. Denki Kagaku **1994**, *62* (2), 152–157.
37. Yabe, H.; Ito, Y.; Ema, K.; Oishi, J. Surface modification of nickel electrodes for molten carbonate fuel cells. J. Power Sources **1988**, *24* (3), 207–214.
38. United Technologies Corporation. Quarterly Report No. 4 for DOE DE-AC-01-79, ET-15440; United Technologies Corporation: South Windsor, CT, Jul–Sep, 1980.
39. General Electric Co. *Development of Molten Carbonate Fuel Cells for Power Generation*, Report No. SRD 80-053; General Electric Co.: Apr 1980.
40. Iacovangelo, C.D.; Karas, B.R. Bubble pressure barriers for molten carbonate fuel cells: materials, properties and in-cell testing. J. Electrochem. Soc. **1986**, *133* (8), 1595–1600.

41. Bychin, V.P.; Zvezdkin, V.A.; Samatov, O.M. Porous nickel anodes for molten-carbonate fuel cells. Russ. J. Electrochem. **1993**, *29* (11), 1346–1349.
42. Kudo, T.; Nishina, T.; Uchida, I. Fabrication of porous gas-diffusion anode for molten-carbonate fuel cell by composite coating method. Denki Kagaku **1990**, *58* (4), 354–359.
43. Pigeaud, A. Energy Research Corporation, Final Report; DOE-ET-11304-19, Feb–Dec, 1980.
44. Antolini, E.; Ferretti, M.; Gemme, S. Preparation of porous nickel electrodes for molten carbonate fuel cells by non-aqueous tape casting. J. Mater. Sci. **1996**, *31* (8), 2187–2192.
45. Li, F.; Wang, C.-M.; Hu, K.-A. Optimization of non-aqueous nickel slips for manufacture of molten carbonate fuel cell electrodes by tape casting method. Mater. Res. Bull. **2002**, *37* (12), 1907–1921.
46. Niikura, J.; Hatoh, K.; Taniguchi, N.; Gamo, T.; Iwaki, T. Fabrication and properties of combined electrode/electrolyte tape for molten carbonate fuel cells. J. Appl. Electrochem. **1990**, *20* (4), 606–610.
47. Niikura, J.; Hatoh, K.; Iwaki, T. In-cell sintering anodes for molten carbonate fuel cell. J. Chem. Soc. Jpn. **1989**, *1*, 20–25.
48. Energy Research Corporation, Report No. DOE-ET-11304-29, Sep 1983.
49. Wendt, H.; Boehme, O.; Leidich, F.U.; Brenscheidt, T. Materials and production technologies of molten carbonate fuel cells. In *Proceedings—Electrochemical Society*, Proceedings of the Third International Symposium on Carbonate Fuel Cell Technology, 1993; Vol. 93-3, 485–495.
50. Oh, I.-H.; Yoon, S.P.; Lim, T.H.; Nam, S.W.; Hong, S.-A.; Lim, H.C. Effect of the structural changes of the Ni-Cr anode on the molten carbonate fuel cell performance. Denki Kagaku **1996**, *64* (6), 497–507.
51. Hwang, E.R.; Park, J.W.; Kim, Y.D.; Kim, S.J.; Kang, S.G. Effect of alloying elements on the copper-base anode for molten carbonate fuel cells. J. Power Sources **1997**, *69* (1–2), 55–60.
52. Pierce, R.D.; Smith, J.L.; Poeppel, R.B. A review of cathode development for molten carbonate fuel cells. In *Molten Carbonate Fuel Cell Technology*, Proceedings Electrochemical Society, 1984; Vol. 84–13, 147–174.
53. Freni, S.; Barone, F.; Puglisi, M. The dissolution process of the NiO cathodes for molten carbonate fuel cells: state-of-the-art. Int. J. Energy Res. **1998**, *22* (1), 17–31.
54. Baumgartner, C.E. Solubility and transport of nickel (II) oxide cathodes in molten carbonate fuel cells. J. Am. Ceram. Soc. **1986**, *69* (2), 162–168.
55. Joon, K. Fuel cells—a 21st century power system. J. Power Sources **1998**, *71* (1–2), 12–18.
56. Watanabe, M.; Takao, M.; Namikawa, Y. Control of micropore structures of porous oxide tiles prepared by a slurry sintering method for molten carbonate fuel cells. Denki Kagaku **1989**, *57* (9), 898–901.
57. Giorgi, L.; Moreno, A.; Pozio, A.; Simonetti, E. Cathode materials for molten carbonate fuel cells. In *Carbonate Fuel Cell Technology*, Proceedings Electrochemical Society, 1999; Vol. 99-20, 265–286.
58. Veldhuis, J.B.J.; Eckes, F.C.; Plomp, L. The dissolution properties of $LiCoO_2$ in molten 62 : 38 mole percent Li:K carbonate. J. Electrochem. Soc. **1992**, *139* (1), L6–L8.
59. Marini, A.; Berbenni, V.; Massarotti, V.; Capsoni, D. Phase composition, microstructure, and sintering in the system Co-Li_2CO_3. J. Solid State Chem. **1995**, *116* (1), 15–27.
60. Legergren, C.; Lundblad, A.; Bergman, B. Synthesis and performance of $LiCoO_2$ cathodes for the molten carbonate fuel cell (MCFC). J. Electrochem. Soc. **1994**, *141* (11), 2959–2966.
61. Giorgi, L.; Carewska, M.; Scaccia, S.; Simonetti, E.; Giacometti, E.; Tulli, R. Development of molten carbonate fuel cell using novel cathode material. Int. J. Hydrogen Energy **1996**, *21* (6), 491–496.
62. Giorgi, L.; Carewska, M.; Scaccia, S.; Simonetti, E.; Zarzana, F. Investigation on $LiCoO_2$ materials for MCFC alternative cathodes. Denki Kagaku **1996**, *64* (6), 482–485.
63. Makkus, R.C.; Hemmes, K.; de Wit, J.H.W. A comparative study of NiO(Li), $LiFeO_2$, and $LiCoO_2$ porous cathodes for molten carbonate fuel cells. J. Electrochem. Soc. **1994**, *141* (12), 3429–3438.
64. Lagergren, C.; Lindbergh, G. Experimental determination of effective conductivities in porous molten carbonate fuel cell electrodes. Electrochim. Acta **1998**, *44* (2–3), 503–511.
65. Lagergren, C.; Simonsson, D. The effects of oxidant gas composition on the polarization of porous $LiCoO_2$ electrodes for the molten carbonate fuel cell. J. Electrochem. Soc. **1997**, *144* (11), 3813–3817.
66. Bernd, R.; Uwe, J. $LiCoO_2$ cathodes with optimized microstructure. One of the keys for high performance and long life MCFC. Denki Kagaku **1996**, *64* (6), 519–525.
67. Prins-Jansen, J.A.; Plevier, G.A.J.M.; Hemmes, K.; De Wit, J.H.W. An *ac*-impedance study of dense and porous electrodes in molten-carbonate

fuel cells. Electrochim. Acta **1996**, *41* (7–8), 1323–1329.
68. Kudo, T.; Kihara, K.; Hisamitsu, Y.; Yu, Q.; Mohamedi, M.; Uchida, I. Electrochemical stability and solubility of Ni-Fe-Co ternary alloy oxides in Li/Na carbonate eutectic as an alternative material for MCFC cathodes. J. Mater. Chem. **2002**, *12* (8), 2496–2500.
69. Fukui, T.; Ohara, S.; Hotta, T.; Okawa, H.; Naito, M. Properties of NiO cathode coated with lithiated Co and Ni solid solution oxide for MCFCs. J. Power Sources **2000**, *86* (1–2), 340–346.
70. Fukui, T.; Hotta, T.; Okawa, H.; Ohara, S.; Naito, M.; Nogi, K. Performance and morphology of a new cathode made from CoO/Ni composite particles for MCFCs. Electrochemistry **2001**, *69* (5), 335–339.
71. Takehisa, F.; Hajime, O.; Tadashi, H.; Makio, N. Synthesis of CoO/Ni composite powders for molten carbonate fuel cells. J. Am. Ceram. Soc. **2001**, *84* (1), 233–235.
72. Zhang, X.G.; Capobianco, P.; Torazza, A.; Passalacqua, B. A solubility study of prelithiated NiO and prelithiated cobaltite as cathode materials for MCFC in 62 Li-38 K eutectic carbonate melts. Electrochemistry **1999**, *67* (6), 603–607.
73. Gourba, E.; Cassir, M.; Tessier, C. Chemical and electrochemical behaviour of Ni–Ti in the cathodic conditions used in molten carbonate fuel cells. J. Electroanal. Chem. **2001**, *503* (1–2), 69–77.
74. Soler, J.; Gonzalez, T.; Escudero, M.J.; Rodrigo, T.; Daza, L. Endurance test on a single cell of a novel cathode material for MCFC. J. Power Sources **2002**, *106* (1–2), 189–195.
75. Escudero, M.J.; Novoa, X.R.; Rodrigo, T.; Daza, L. Influence of lanthanum oxide as quality promoter on cathodes for MCFC. J. Power Sources **2002**, *106* (1–2), 196–205.
76. Durairajan, A.; Colon-Mercado, H.; Haran, B.; White, R.; Popov, B. Electrochemical characterization of cobalt-encapsulated nickel as cathodes for MCFC. J. Power Sources **2002**, *104* (2), 157–168.
77. Ganesan, P.; Colon, H.; Haran, B.; White, R.; Popov, N.B. Study of cobalt-doped lithium–nickel oxides as cathodes for MCFC. J. Power Sources **2002**, *111* (1), 109–120.
78. Ganesan, P.; Colon, H.; Haran, B.; Popov, N.B. Performance of $La_{0.8}Sr_{0.2}CoO_3$ coated NiO as cathodes for molten carbonate fuel cells. J. Power Sources **2003**, *115* (1), 12–18.
79. Selman, J.R.; Maru, H.C. Physical chemistry and electrochemistry of alkali carbonate melts with special reference to the molten-carbonate fuel cell. Adv. Molten Salt Chem. **1981**, *4*, 159–390.
80. Kunz, H.R. Binary electrolyte for molten carbonate fuel cells. U.S. Patent, 4,591,538, May 27, 1986.
81. Tanase, S.; Miyazaki, Y.; Yanagida, M.; Tanimoto, K.; Kodama, T. Numerical formulation of electrical conductance data of molten alkali carbonates. Prog. Batteries Solar Cells **1988**, *7*, 396–402.
82. Maru, H.C.; Paetsch, L.; Pigeaud, A. Review of molten carbonate fuel cell matrix technology. Proceedings of the 1st Symposium on Molten Carbonate Fuel Cell Technology, PV. 84-13; Electrochemical Society Inc.: Pennington, NJ, 1984; 20–53.
83. Yoshikawa, M.; Mugikura, Y.; Izaki, Y.; Watanabe, T. The relationship between the internal resistance and electrolyte loss in molten carbonate fuel cell. The cell life using Li_2CO_3/Na_2CO_3 electrolyte. Denki Kagaku **1998**, *66* (3), 279–285.
84. Wendt, H.; Brenscheidt, Th.; Kah, M. Different molten alkali carbonate eutectics as fuel cell electrolytes for MCFCs. Is lithium/potassium or lithium/sodium the appropriate choice? A critical survey. High Temp. Mater. Processes **1998**, *2* (4), 597–612.
85. Yoshioka, S.; Urushibata, H. Superiority of Li_2CO_3/Na_2CO_3 electrolyte in molten carbonate fuel cells. I. Influence of operation conditions on polarization behavior. Denki Kagaku **1996**, *64* (8), 909–914.
86. Yoshioka, S.; Urushibata, H. Superiority of Li_2CO_3/Na_2CO_3 electrolyte in molten carbonate fuel cell. II. Effect of the basicity of the electrolyte on the life performance. Denki Kagaku **1996**, *64* (10), 1074–1079.
87. Fujita, Y.; Nishimura, T.; Urushibata, H.; Sasaki, A. Degradation of the components in molten carbonate fuel cells with Li/Na electrolyte. Proceedings of the 4th Symposium on Molten Carbonate Fuel Cell Technology, PV. 97-4, Selman, J.R., Ed.; Electrochemical Society Inc.: Pennington, NJ, 1997; 191–202.
88. Yoshikawa, M.; Mugikura, Y.; Watanabe, T.; Ota, T.; Suzuki, A. The behavior of MCFCs using Li/K and Li/Na carbonates as the electrolyte at high pressure. J. Electrochem. Soc. **1999**, *146* (8), 2834–2840.
89. Janowitz, K.; Kah, M.; Wendt, H. Molten carbonate fuel cell research: part I. Comparing cathodic oxygen reduction in lithium/potassium and lithium/sodium carbonate melts. Electrochim. Acta **1999**, *45* (7), 1025–1037.

90. Smith, D.S.; Winnick, J. Cesium-containing electrolyte for the molten carbonate fuel cell. Electrochem. Solid State Lett. **1999**, *2* (5), 207–209.
91. Kaun, T.D.; Schoeler, A.; Centeno, C.-J.; Krumpelt, M. Improved MCFC performance with Li/Na/Ba/Ca carbonate electrolyte. Proceedings of the 5th Symposium on Molten Carbonate Fuel Cell Technology, PV. 99-20, Uchida, I., Ed.; Electrochemical Society Inc.; Pennington, NJ, 1999; 219–227.
92. Kojima, T.; Miyazaki, Y.; Nomura, K.; Tanimoto, K. Density, molar volume, and surface tension of molten Li_2CO_3-Na_2CO_3 and Li_2CO_3-K_2CO_3 containing alkaline earth (Ca, Sr, and Ba) carbonates. J. Electrochem. Soc. **2003**, *150* (11), E535–E542.
93. Bohme, O.; Leidich, F.U.; Salge, H.J.; Wendt, H. Development of materials and production technologies for molten carbonate fuel cells. Int. J. Hydrogen Energy **1994**, *19* (4), 349–355.
94. Maru, H.C.; Farooque, M.; Pigeaud, A. Review of carbonate fuel cell matrix and electrolyte. Proceedings of the 2nd Symposium on Molten Carbonate Fuel Cell Technology, PV. 90-16, Electrochemical Society Inc.: Pennington, NJ, 1990; 121–136.
95. Kinoshita, K.; Sim, J.W.; Ackerman, J.P. Preparation and characterization of lithium aluminate. Mater. Res. Bull. **1978**, *13* (5), 445–455.
96. Kinoshita, K.; Sim, K.W.; Kucera, H. Synthesis of fine particle size lithium aluminate for application in molten carbonate fuel cells. Mater. Res. Bull. **1979**, *14* (10), 1357–1368.
97. Arendt, R.H.; Curran, M.J. Alternate synthesis of electrolyte matrix for molten carbonate fuel cells. J. Electrochem. Soc. **1980**, *127* (8), 1660–1663.
98. Arendt, R.H. Alternate matrix materials for molten carbonate fuel cell electrolyte structures. J. Electrochem. Soc. **1982**, *129* (5), 979–983.
99. Tanimoto, K.; Miyazaki, Y.; Yanagida, M.; Tanase, S.; Kojima, T.; Okuyama, H.; Kodama, T. Alternative matrix materials for molten carbonate fuel cell. Denki Kagaku **1990**, *41* (2), 51–55.
100. Nakagawa, K.; Ohashi, T. A reversible change between lithium zirconate and zirconia in molten carbonate. Electrochemistry **1999**, *67* (6), 618–621.
101. Paetsch, L.M.; Doyon, J.D.; Farooque, M. Review of carbonate fuel cell matrix and electrolyte developments. Proceedings of the 3rd Symposium on Molten Carbonate Fuel Cell Technology, PV. 93-3, Electrochemical Society Inc.: Pennington, NJ, 1993; 89–105.
102. Yuh, C.Y.; Huang, C.M.; Farooque, M. Advances in carbonate fuel cell matrix and electrolyte. Proceedings of the 4th Symposium on Molten Carbonate Fuel Cell Technology, PV. 97-4, Electrochemical Society Inc.: Pennington, NJ, 1997; 66–78.
103. Yuh, C.Y.; Farooque, M.; Maru, H. Advances in carbonate fuel cell matrix and electrolyte. Proceedings of the Symposium on Molten Carbonate Fuel Cell Technology, PV. 99-20; Electrochemical Society Inc.: Pennington, NJ, 1999; 189–201.
104. Singh, R.N. Fracture strength of a porous lithium aluminate structure for application in molten carbonate fuel cells. Proc. Ceram. Sci. Eng. **1981**, *1* (7–8)(B), 500–507.
105. Singh, R.N.; Dusek, J.T.; Sim, J.W. Fabrication and properties of a porous lithium aluminate electrolyte retainer for molten carbonate fuel cells. Am. Ceram. Soc. Bull. **1981**, *60* (6), 629–635.
106. Sim, J.W.; Singh, R.N.; Kinoshita, K. Testing of sintered lithium aluminate structures in molten carbonate fuel cells. J. Electrochem. Soc. **1980**, *127* (8), 1766–1768.
107. Arendt, R.H.; Curran, M.J. Alternate fabrication process for molten carbonate fuel cell electrolyte structures. J. Electrochem. Soc. **1980**, *127* (8), 1663–1666.
108. Alvani, C.; Roncari, E. Preparation and characterization of γ-lithium aluminate tiles for molten carbonate fuel cells. High Temp.–High Pressures **1988**, *20* (3), 247–250.
109. Birk, S.; Ibeh, C.C. Plastics in fuel cell applications: an in-lab developed and fabricated molten carbonate fuel cell (MCFC) electrolyte matrix support with polyolefin-based binders. 57th Annual Technical Conference of the Society of Plastic Engineers, 1999; Vol. 2, 2629–2633.
110. Pasco, W.D.; Arendt, R.H. Some alternate fabrication processes for molten carbonate fuel cell electrolyte structures. J. Electrochem. Soc. **1986**, *133* (12), 2498–2502.
111. Iacovangelo, C.D.; Pasco, W.D. Hot-roll-milled electrolyte structures for molten carbonate fuel cells. J. Electrochem. Soc. **1988**, *135* (1), 221–224.
112. Li, N.; Yi, B.; Kong, L.; Zhang, E.; Ling, H.; Qu, T.; Cheng, Y. Cold-roll-milled electrolyte membrane for molten carbonate fuel cells. Mo Kexue Yu Jishu **1997**, *17* (2), 24–29.
113. Li, N.; Yi, B.; Lin, H.K.; Lianying, Z.; Enjun, Q.; Tianxi, C.Y. Preparation of $LiAlO_2$ used as matrix materials in molten carbonate fuel cells. Wuji Cailiao Xuebao **1997**, *12* (2), 211–217.
114. Lin, H.; Yi, B.; Li, N.; Qu, T.; Zhang, E.; Kong, L.; Chen, Y. Study on the performance of MCFC matrix prepared by flow cast. Dianhuaxue **1998**, *4* (4), 406–413.
115. Koseki, K.; Nishihara, H.; Shundo, H.; Nakanishi, T. Molten carbonate fuel cells with

electrolyte plates prepared with paper-making method. Proceedings of the 24th Intersociety Energy Conservation Engineering Conference, 1989; Vol. 3, 1529–1534.
116. Baumgartner, C.E.; DeCarlo, V.J.; Glugla, P.G.; Grimaldi, J.J. Molten carbonate fuel cell electrolyte structure fabrication using electrophoretic deposition. J. Electrochem. Soc **1985**, *132* (1), 57–63.
117. Hyatt, E.P. Making thin, flat ceramics—a review. Am. Ceram. Soc. Bull. **1986**, *65* (4), 637–638.
118. Williams, J.C. Doctor-blade process. In *Treatise on Materials Science and Technology*; Wang, F.F.Y., Ed.; Academic Press: New York, 1976; Vol. 9, 173–198.
119. Paetsch, L.; Pigeaud, A.; Chamberlin, R.; Maru, H. *Development of Molten Carbonate Fuel Cell Components*, Final Report to EPRI, Report No. AP 5789, Jul 1989.
120. Bushnell, C.L.; Bregoli, L.J.; Schroll, C.R. Electrolyte matrix for molten carbonate fuel cells. U.S. Patent 4,322,482, Mar 30, 1982.
121. Fairchild, G.H.; Brown, P.M. The fabrication of lithium aluminate/lithium silicate-based ceramics. Adv. Ceram. **1989**, *25*, 123–139, (Fabr. Prop. Lithium Ceram.).
122. Nirasawa, H.; Kawachi, T.; Ogawa, T.; Hori, M.; Tomimatsu, N.; Nakagawa, K.; Ohzu, H.; Yamazaki, Y. Stabilized matrix for molten carbonate fuel cell. 31st Intersociety Energy Conversion Engineering Conference, Washington, DC, Aug 11–16, 1996; 1132–1137.
123. Shoji, C.; Matsuo, T.; Suzuki, A.; Yamamasu, Y. Development of electrolyte plate for molten carbonate fuel cell. In *Materials for Electrochemical Energy Storage and Conversion II—Batteries, Capacitors and Fuel Cells*, Materials Research Society Symposium Proceedings, 1998; Vol. 496, 211–216.
124. Yasumoto, E.; Hatoh, K.; Gamo, T. Particle growth behaviour of $LiAlO_2$ containing ZrO_2 in Li/Na carbonate electrolytes. J. Power Sources **1998**, *71* (1–2), 159–163.
125. Mugikura, Y.; Abe, T.; Yoshioka, S.; Urushibata, H. Nio dissolution in molten carbonate fuel cells: effect on performance and life. J. Electrochem. Soc. **1995**, *142* (9), 2971–2976.
126. Donado, R.A.; Marionowski, L.G.; Maru, H.C.; Selman, J.R. Corrosion of the wet-seal area in molten carbonate fuel cells II. Experimental results. J. Electrochem. Soc. **1984**, *131* (11), 2541–2543.
127. Yuh, C.; Johnson, R.; Farooque, M.; Maru, H. Status of carbonate fuel cell materials. J. Power Sources **1985**, *56* (1), 1–10.
128. Tanimoto, K.; Miyazaki, Y.; Yanagida, M.; Tanase, S.; Kojima, T.; Ohtori, N.; Kodama, T. Corrosivity of molten alkali and alkaline earth carbonate to SUS 310 S under a simulated MCFC cathode environment. Denki Kagaku **1994**, *62* (5), 445–450.
129. Shimada, T.; Ariga, N.; Sakai, J.; Masamura, K. Corrosion resistant alloy for molten carbonate fuel cell. Trans. Mater. Res. Soc. Jpn. **1994**, *18A*, 569–571, (Ecomaterials).
130. Matsuyama, H.; Nishina, T.; Uchida, I. Corrosion studies of Fe-based Cr alloys containing Al in molten carbonates. In *Molten Salt Chemistry and Technology*, Proceedings of the Electrochemical Society, Saboungi, M.-L., Kojima, H., Duruz, J., Shores, D., Eds.; The Electrochemical Society Proceeding Series; Electrochemical Society Inc.: Pennington, NJ, 1993; Vol. 93-9, 436–445.
131. Keijze, M.; Hemmes, K.; Van der Put, P.J.J.M.; De Wit, J.H.W.; Schoonman, J. A search for suitable coating materials on separator plates for molten carbonate fuel cells. Corros. Sci. **1997**, *39* (3), 483–494.
132. Okuyama, M.; Ushioda, M.; Itoi, Y. Formation of graded metal-ceramic coatings on iron by plasma spray and their corrosion resistance to molten alkali carbonates. Denki Kagaku **1992**, *60* (6), 508–514.
133. Vossen, J.P.T.; Makkus, R.C.; De Wit, J.H.W. Corrosion behavior of chromium in molten carbonate. J. Electrochem. Soc. **1996**, *143* (1), 66–73.
134. Davis, H.J.; Kinnibrugh, D.R. Passivation phenomena and potentiostatic corrosion in molten alkali metal carbonates. J. Electrochem. Soc. **1970**, *117* (3), 392–396.
135. Swaroop, R.B.; Sim, J.W.; Kinoshita, K. Corrosion protection of molten carbonate fuel cell gas seals. J. Electrochem. Soc. **1978**, *125* (11), 1799–1800.
136. Vossen, J.P.T.; Plomp, L.; De Wit, J.H.W.; Rietveld, G. Corrosion behavior of stainless steel and nickel-base alloys in molten carbonate. J. Electrochem. Soc. **1995**, *142* (10), 3327–3335.
137. Vossen, J.P.T.; Makkus, R.C.; De Wit, J.H.W. Corrosion behavior of nickel-chromium alloys in molten carbonate. Mater. Corros. **1997**, *48* (3), 157–164.
138. Vossen, J.P.T.; Janssen, A.H.H.; De Wit, J.H.W. Corrosion behavior of nickel-iron alloys in molten carbonate. J. Electrochem. Soc. **1996**, *143* (1), 58–65.
139. Vossen, J.P.T.; Plomp, L.; De Wit, J.H.W. Corrosion of nickel in molten carbonate. J. Electrochem. Soc. **1994**, *141* (11), 3040–3048.

140. Vossen, J.P.T.; Ament, P.C.H.; De Wit, J.H.W. Mechanisms for oxidation and passive behavior of nickel in molten carbonate. J. Electrochem. Soc. **1996**, *143* (7), 2272–2279.
141. Vossen, J.P.T.; Janssen, A.H.H.; De Witt, J.H.W. The corrosion behavior of NiAl in molten carbonate at 650°C. Mater. Corros. **1996**, *47* (12), 703–708.
142. Vossen, J.P.T.; Makkus, R.C.; Janssen, A.H.H.; De Wit, J.H.W. Corrosion behavior of nickel-aluminum alloys in molten carbonate. Mater. Corros. **1997**, *48* (4), 228–236.
143. Durairajan, A.; Haran, B.; Popov, N.B.; White, R.E. Cycle life and utilization studies on cobalt microencapsulated AB_5 type metal hydride. J. Power Sources **1999**, *83* (1–2), 114–120.
144. Maru, H.C.; Baker, B.S. Applications of internal reforming molten carbonate fuel cells. Energy Technol. **1986**, *13*, 1551–1559.
145. Gonjyo, Y.; Matsumura, M.; Tanaka, T. Performance of direct internal reforming molten carbonate fuel cells. Proceedings of the 26th Intersociety Energy Conversion Engineering Conference, American Nuclear Society: LaGrange Park, IL, 1991; 600–605.
146. Parmaliana, A.; Frusteri, F.; Tsiakaras, P.; Giordano, N. Out of the cell performance of reforming catalysts for direct molten carbonate fuel cells (DMCFC). In *Advances in Hydrogen Energy 5*, 6th World Hydrogen Energy Conference, 1986; Vol. 3, 1252–1258.
147. Mori, T.; Higashiyama, K.; Yoshioka, S.; Kobayashi, T.; Itoh, S. Steam reforming reaction of methane in directly-fueled molten carbonate fuel cell. J. Electrochem. Soc. **1989**, *136* (8), 2230–2234.
148. Cavallaro, S.; Mondello, N.; Freni, S. Hydrogen produced from ethanol for internal reforming molten carbonate fuel cell. J. Power Sources **2001**, *102* (1–2), 198–204.
149. Choi, J.-S.; Kwon, H.-H.; Lim, T.-H.; Hong, S.-A.; Lee, H.-I. Development of nickel catalyst supported on $MgO–TiO_2$ composite oxide for DIR-MCFC. Catal. Today **2004**, *93–95*, 553–560.

Multiphase Mixing and Solid–Liquid Mixing in Agitated Reactors

Piero M. Armenante
Otto H. York Department of Chemical Engineering, New Jersey Institute of Technology, Newark, New Jersey, U.S.A.

Victor Atiemo-Obeng
Engineering Science and Market Development, The Dow Chemical Company, Midland, Michigan, U.S.A.

W. Roy Penney
Department of Chemical Engineering, University of Arkansas, Fayetteville, Arkansas, U.S.A.

INTRODUCTION

Multiphase reactions can be significantly affected by how well mixed the system is and how intimately dispersed the phases are. The reason for this is easy to explain, but more difficult to quantify: although the course of any reaction is determined exclusively by the *local* concentrations of the reactants and the intrinsic reaction kinetic rates, in any real reactive system, the local reactant concentrations depend not only on how fast the reactants are depleted by the reaction, but also on how fast they are locally replenished from the bulk of the phases in which they initially reside. The latter phenomenon is directly related to the existence of a mass transfer step (in series with the reaction step), which determines the rate at which the reactants in different phases are brought in contact with each other. In many cases, especially if the rate of reaction is "fast" with respect to the mass transfer rate, the latter mechanism can become controlling over the former, and the overall reaction process is dominated by mass transfer and, hence, multiphase mixing.

BACKGROUND

Mixing has no impact on the intrinsic reaction kinetics, but it has a controlling effect on the temporal variation in species concentration and the mass transfer rate. This can be shown by examining the typical convective mass transfer rate equation:

$$\dot{m} = k_L A(C_{\text{interface}} - C_{\text{bulk}}) = k_L a_v V_L \Delta C \qquad (1)$$

All the variables on the right-hand side of this equation are directly affected by mixing. More specifically, the mixing effects generated by the mechanical energy introduced into the reactor impact the overall mass transfer rate $\dot{m}$ by affecting the following variables:

1. State of dispersion or suspension of the dispersed phase, i.e., *degree of macroscopic homogeneity* of the dispersed phase throughout the continuous phase. This variable affects mainly the volume of the liquid that is *effectively* available for the reaction. Depending on the intensity of mixing, such a volume can be equal to, or smaller than, the liquid volume in the reactor, V_L. The local concentration gradient ΔC is also affected by the degree of homogeneity, in that poor mixing can lead to variations in the bulk concentration, C_{bulk}, from location to location within the reactor.
2. Specific interfacial area, a_v, and overall interfacial area, A, available for mass transfer.
3. Mass transfer coefficient at the interface (k_L).

The first variable refers to how well interdispersed the phases are in different portions of the reactors. In most cases, mixing affects this variable by determining the major flow patterns existing in the reactor under a specific operating regime. For example, the loading regime in a stirred liquid sparged with a gas is defined as the condition at which the impeller can disperse the gas throughout the upper part of a vessel only (but not necessarily in the portion below the impeller). Obviously, if one phase, i.e., the gas in this example, is dispersed only in a portion of the liquid in the reactor, the liquid volume available for mass transfer will be reduced and C_{bulk} will vary from location to location within the vessel at any given time. Similar remarks can be made for liquid–liquid dispersions, and especially solids–liquid suspensions, where it is critical to ensure that the solids are in the "just-suspended" regime, thus fully exposing their entire

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030574

surface to the liquid, if mass transfer effects are to be minimized.

The second variable, the interfacial area, is also greatly affected by mixing, especially in those systems, such as liquid–liquid and gas–liquid dispersions, where mixing effects produce a breakup of the dispersed phase into smaller ''elements'' (bubbles or droplets): the smaller these elements, the larger the interfacial area. Finally, the turbulent mixing produced by agitation has a well-known effect on the third variable, the interfacial mass transfer coefficient, by reducing the thickness of the boundary layer around a dispersed phase: the greater the local turbulent intensity, the thinner the boundary layer and the higher the mass transfer coefficient.

In summary, mixing controls every aspect of mass transfer. Hence, mixing, through mass transfer, is a key factor in the mass balances for the reactants, strongly affecting the local concentration of the reactants and, ultimately, the overall reaction process, especially if the reaction kinetic rates are comparable to, or faster than, the mass transfer rates.

It should also be stressed that the mass transfer is affected by scale, whereas the intrinsic reaction kinetics is not. To make matters worse, small reactors are typically much better mixers than large reactors, implying that many mixing phenomena are negatively impacted by scale and may manifest themselves only at large scales. Therefore, failure to analyze the multiphase mixing aspects of a reaction process may result in undesired results on scale-up, such as reduced yield of a desired product, or partial dispersion, and, hence, reduced utilization of a suspended solid catalyst.

Computational fluid dynamics (CFD) has been used in recent years to predict the flow, mixing, and reaction characteristics of single and multiphase systems, including a variety of complex mixing-reactive systems.[1–5] Although the majority of these applications have described processes occurring in single-phase systems, CFD is being currently used to study more complex multiphase systems, and it is expected to be even more extensively used in the future for such applications.

The remainder of this entry as well as the other two entries in this encyclopedia entitled ''Liquid–Liquid Mixing in Agitated Reactors'' and ''Gas–Liquid Mixing in Agitated Reactors'' deal exclusively with systems in which a liquid is the continuous phase and a second dispersed system is introduced as a finely divided solid (solid–liquid mixing), an immiscible liquid (liquid–liquid mixing), or a sparged gas (gas–liquid mixing). Most of the discussion is focused on multiphase reactors whose content is mechanically stirred by an impeller operating under turbulent regime. Each system is analyzed independently, but the impact of mixing on the three main variables listed above, i.e., degree of dispersion, a_v (and A), and k_L, is examined in a similar fashion in each case.

POWER DISSIPATED BY IMPELLERS IN AGITATED VESSELS

In general, increasing the mixedness of a system (e.g., by increasing the agitation speed) has a beneficial effect on all the above-mentioned variables responsible for mass transfer, and, hence, on the overall mass transfer rate. However, a higher rate of mechanical energy dissipation (power) is typically required to achieve such an increased level of mixedness. Power dissipation has a significant impact on all the variables controlling mass transfer. Therefore, before proceeding any further, it is important to distinguish among different types or impellers and determine their mixing or hydrodynamic characteristics.

Impellers come in a variety of shapes and forms. They are commonly classified as either radial impellers, producing a main radial flow perpendicular to the shaft (as shown in Fig. 1), or axial impellers, generating a flow mainly oriented parallel to the shaft (as shown in Fig. 2).[6] More recently, hydrofoil (or fluidfoil) impellers have been developed, typically producing strong axial flow and low shear (as shown in Fig. 3).[6] Radial impellers typically dissipate more power as turbulent shear than axial and hydrofoil impellers, because of the significant shear which they generate. Consequently, they are used in those applications in which new surface area must be created, and high mass transfer rates are required (e.g., gas–liquid and liquid–liquid applications). Axial and hydrofoil impellers are more common when high flow and low shear are demanded (e.g., solids suspension), although they may have other applications as well (e.g., fermentation). Close-clearance impellers (such as rotor–stator homogenizers) are a separate class of impellers finding applications in higher-viscosity fluids, or where very high shear is required (e.g., liquid–liquid emulsification).[6]

The power dissipated by an impeller rotating in a homogenous, single-phase liquid in a baffled vessel is a function of the type of impeller, the flow regime in which the impeller operates (laminar vs. turbulent), which is, in turn, a function of the impeller Reynolds number, $Re = \rho_L N D^2/\mu$, and a number of geometric ratios. For an agitated vessel, the impeller power dissipation, P, and the power dissipation per unit liquid mass, ε, can be calculated, respectively, from:

$$P = N_p \rho_L N^3 D^5 \tag{2}$$

and

$$\varepsilon = \varepsilon_{avg} = \frac{P}{\rho_L V_L} \tag{3}$$

where the impeller power number, N_p, is available, for selected impellers in selected system geometries (from Fig. 4).[6] For sufficiently high Re, the flow is turbulent

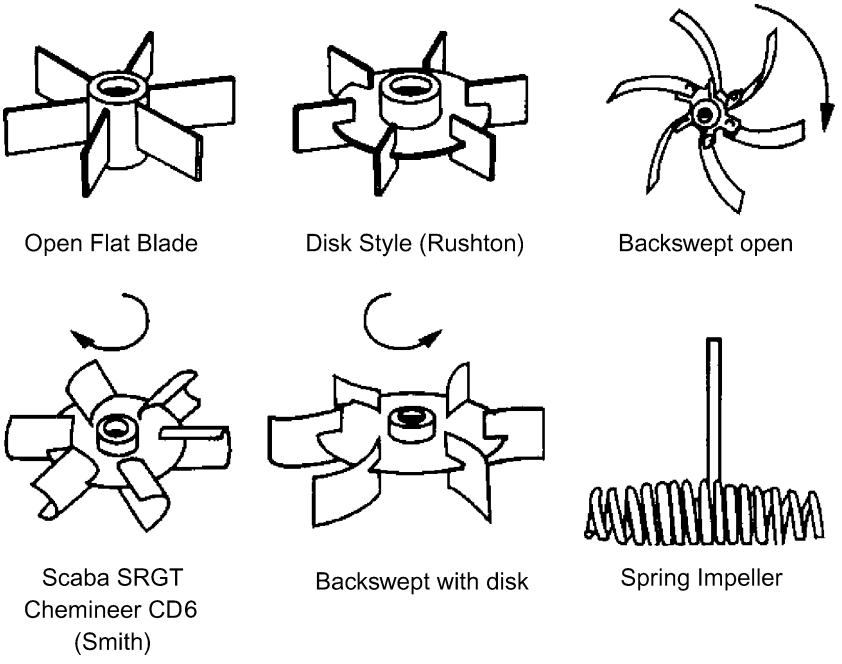


Fig. 1 Radial flow impellers. (From Ref.[6].)

and the power number becomes a constant, which depends only on the type and geometry of the impeller–vessel configuration. For example, for a standard Rushton turbine in a baffled standard impeller–vessel configuration ($H/T = 1$; $C/T = 1/3$; and $D/T = 1/3$) $N_p = 5$. Additional N_p values for different types of impellers are provided below, in the next two entries, and elsewhere.[7–13] The power numbers for multiple impellers mounted on the same shaft are also available.[11,14–16]

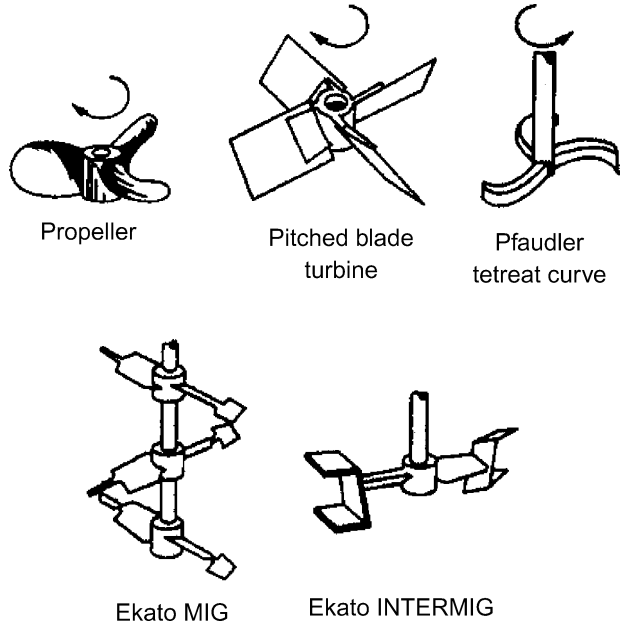


Fig. 2 Axial flow impellers. (From Ref.[6].)

MIXING IN SOLID–LIQUID AGITATED REACTORS

Solid–liquid reactors contain a two-phase mixture of the continuous liquid phase and finely divided, small, solid particles resulting from the addition or formation of solids. Typical operations involving solid–liquid mixing include dissolution and leaching, crystallization, precipitation, adsorption, desorption, ion exchange, solid catalyzed reaction, and suspension polymerization. All these unit operations involve solid–liquid mass transfer. The objectives of solid–liquid mixing operations are to promote suspension of solids, resuspension of settled solids, incorporation of floating solids, dispersion of solid aggregates, or control of particle size from the action of fluid shear as well as any abrasion due to particle–particle, particle–wall, and particle–impeller impacts, and mass transfer at the solid–liquid interface.[7]

Selection and Configuration of Solid–Liquid Equipment

Solids suspension is usually carried out in mechanically agitated vessels with or without draft tubes, as shown in Figs. 5 and 6.[7] Dished heads (ASME dished, elliptical, or torispherical heads) are preferred. A single impeller is usually sufficient for off-bottom solids suspension in vessels with dished heads and $H/T < 1.3$. Dual impellers are recommended for vessels with $1.3 < H/T < 2.5$, which are used for uniform suspension of fast-settling solids. Three or more impellers

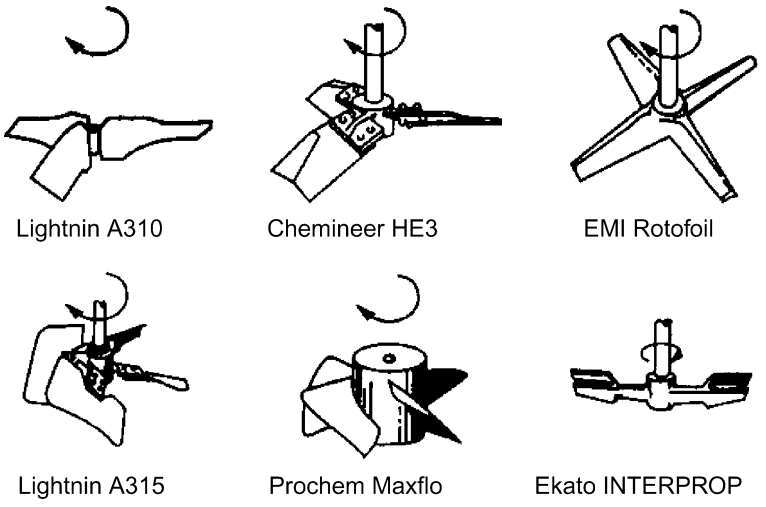


Fig. 3 Hydrofoil impellers. (From Ref.[6].)

may be required if $H/T > 2.5$. A vessel with such a high aspect ratio is a poor choice for solid suspension.

Baffles are essential for solids suspension operations involving solids that are heavier than the liquid. They convert the swirling motion into top-down or axial fluid motion, which helps to lift and suspend the solids.

In steel or alloy vessels, the recommended baffle design for suspension of settling solids is four, flat-blade baffles, each baffle with a width, B, equal to $T/12$, and a wall clearance of at least $T/72$. The baffles should extend to the lower end of the straight section of the cylindrical vessel, at the intersection of the cylindrical section with the bottom dish head. In glass-lined equipment, the recommended baffles are either fin or beaver tail types.[7] A minimum of two baffles is recommended. These baffles are generally less effective than the standard four flat-blade baffle design.

Solids suspension and solids distribution are primarily governed by the bulk or convective flows in the vessel. Axial impellers and hydrofoil impellers (e.g., Lightnin A-310, A-320; Chemineer HE-3, APV LE-20, Ekato Viscoprop) whose discharge is axially directed are more efficient than others in achieving solids suspension. The recommended impeller speed should be higher than N_{js} (defined below). For multi-process batch reactors, mixers equipped with variable speed drives permit the mixer to be operated at different impeller speeds to accommodate the different mixing needs of the various steps in the process.

Typical values for the impeller clearance off the vessel bottom are $T/4$ for hydrofoil impellers and $T/3$ for pitched-blade turbines. Hydrofoil impellers may be a poor choice when solid suspension is accompanied by other mixing duties, such as liquid–liquid dispersion or gas dispersion. For such cases, a multiple impeller system consisting of a high-efficiency impeller in combination with a 45° pitched-blade impeller should be evaluated. Small pitched blade impellers with a

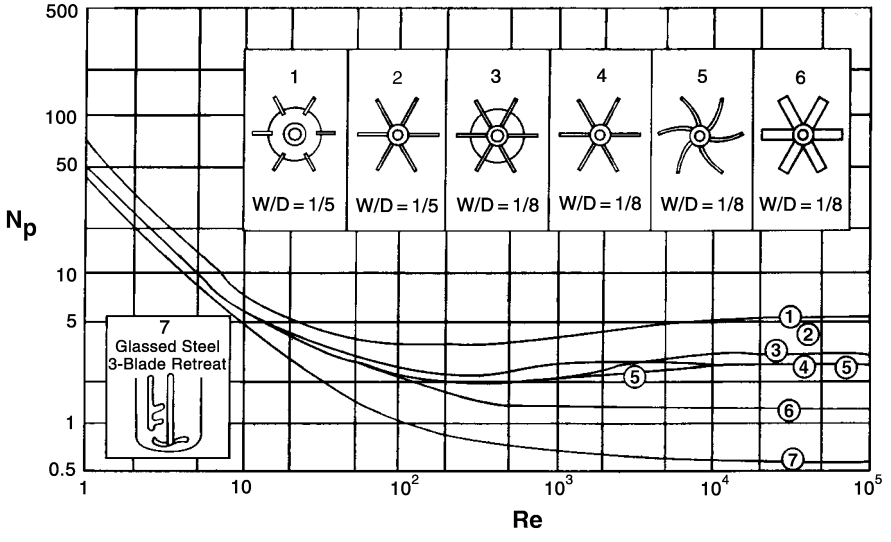


Fig. 4 Power number, N_p, vs. impeller Reynolds number, Re, for seven different impellers. (From Ref.[6].)

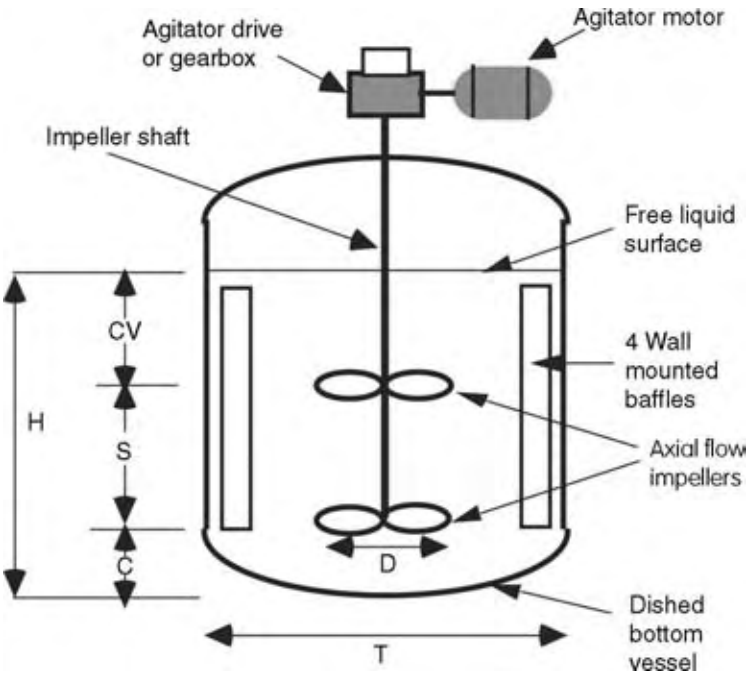


Fig. 5 Schematic representation of a typical mechanically agitated vessel. (From Ref.[7].)

diameter $D < T/2.5$, located near the vessel base ($C < T/4$) are appropriate for solid suspension, and also aid in the discharge of the solids during slurry transfer. If the impeller is located very close to the tank bottom, a four-bladed flat-blade impeller is more appropriate than a pitched-blade turbine.

Nozzles for solids addition should be provided with gratings or screens to keep large solid chunks or foreign matter from entering the vessels. Bottom nozzles should be as short as practical and be installed with flush-bottom valves to prevent the solids from collecting and plugging the nozzle.[7]

Settling Velocity and Solid Suspension

A dense solid particle placed in a quiescent fluid will accelerate to a steady-state settling velocity (free- or still-fluid settling velocity). Equations to calculate the settling velocity are given elsewhere.[17] In an agitated

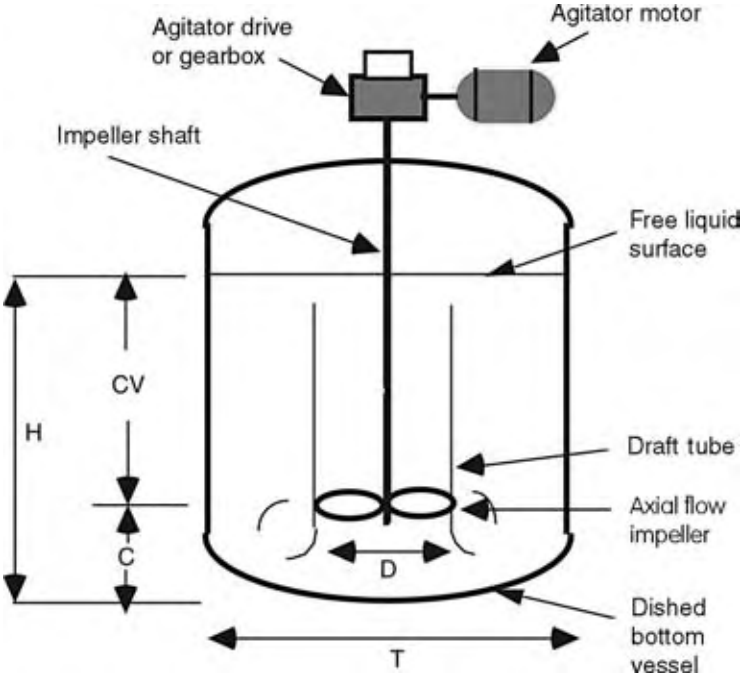


Fig. 6 Schematic representation of a mechanically agitated vessel with a draft tube. (From Ref.[7].)

Table 1 Impact of desired result on mixer design (power and speed depend on mixing criteria and settling velocity)

		Power ratio at settling velocity (ft/min)		
Suspension criteria	**Speed ratio**	**16–60 (Difficult)**	**4–8 (Moderate)**	**0.1–0.16 (Easy)**
On-bottom motion	1	1	1	1
Complete off-bottom suspension	1.7	5	3	2
Total uniformity	2.9	25	9	4

(From Ref.[19].)

solid suspension, the particle settling velocity is always less than the free-settling velocities because of the complex turbulent hydrodynamic field and solid–solid interactions.[18]

The magnitude of the free-settling velocity has proven useful in characterizing solid suspension problems into easy, moderate, or difficult categories, as reported in Table 1.[19]

Hydrodynamics of Solid Suspension and Distribution

Solid suspension requires the input of mechanical energy into the fluid–solid system by some mode of agitation. The input energy creates a turbulent flow field in which solid particles are lifted from the vessel base and subsequently dispersed and distributed throughout the liquid. Solids lifting from the vessel base is achieved by a combination of the drag and the lift forces of the moving fluid on the solid particles and the bursts of turbulent eddies originating from the bulk flow in the vessel (Fig. 7).[7,20,21]

The distribution and magnitude of the mean fluid velocities and large anisotropic turbulent eddies generated by a given agitator determine to what degree the solid suspension may be achieved. Thus, different agitator designs may achieve different degrees of suspensions at similar energy input. For any given impeller, the degree of suspension will vary with D/T as well as C/T, at constant power input.

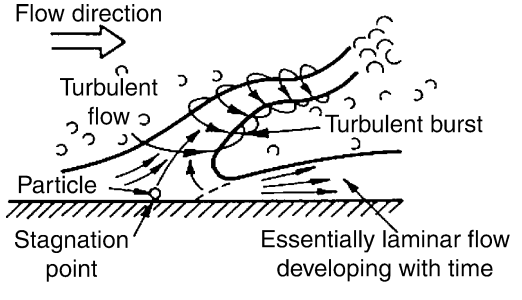


Fig. 7 Sudden pickup of solids by turbulent burst. (From Refs.[7,20].)

Solids Suspension and Distribution Regimes

In agitated vessels, the degree of solid suspension, or solid suspension regime, is generally classified into the following three levels (Fig. 8).[7]

- *On-Bottom Motion or Partial Suspension Regime*: This state is characterized by the complete motion of all particles around the bottom of the vessel. It excludes the formation of fillets, i.e., loose aggregations of particles in corners or other parts of the vessel bottom. As the particles are in constant contact with the base of the vessel and with one another, not all the particle surface area is available for chemical reaction, mass, or heat transfer.
- *Off-Bottom or Complete Suspension* (*Just-Suspended*) *Regime*: This state is characterized by the complete off-bottom motion of all particles with no particle remaining on the base of the vessel for more than 1–2 sec (Zwietering criterion).[22] Under this condition, the total surface area of the particles is exposed to the fluid for chemical reaction, mass or heat transfer. The just-suspended regime refers to the minimum agitation conditions at which all particles attain complete suspension.
- *Nearly Uniform Suspension Regime*: This is the state of suspension at which particle concentration and particle size distribution are roughly uniform throughout the vessel. Any further increase in agitation speed or power does not appreciably enhance the solids distribution in the fluid. A coefficient of variation of the solid concentration of about 0.05 (i.e., a uniformity of 95%) is often considered adequate for most process applications. In practice, a concentration gradient as a function of vertical position will always exist.[23] Complete uniformity is theoretically unattainable and impractical to achieve because a thin clear fluid layer always exists at the air–liquid interface, as the axial lift velocity, and, hence, the particle concentration approaches zero near the liquid surface. Nearly uniform suspension is often the desired process result for operations where a representative sample of solids is required or a uniform concentration of solids must

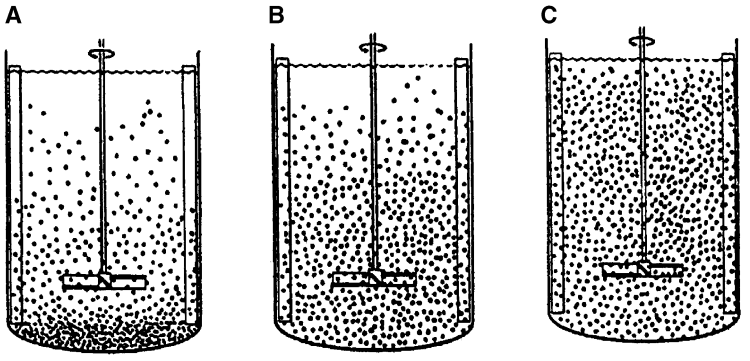


Fig. 8 Degrees of suspension. (A) Partial suspension: some solids rest on the bottom of the tank for short periods, a useful condition only for dissolution of very soluble solids. (B) Complete suspension: all solids are off the bottom of the vessel, minimum desired condition for most solid–liquid systems. (C) Nearly uniform suspension: solids suspended uniformly throughout the vessel, required condition for crystallization, solid-catalyzed reaction. (From Ref.[7].)

be achieved. For example, in crystallization, nonuniform solids concentration may lead to unacceptably high local supersaturation levels and subsequent nonuniformity in crystal growth. Similarly, feeding of filters and reactors from batch feed tanks requires nominal complete uniformity.

Calculation of the Just-Suspended Impeller Speed, N_{JS}, in Solid–Liquid Suspensions

The attainment of the solids just-suspended regime is essential to ensure that all solids are suspended off the tank bottom, thus ensuring that their surface area is fully exposed to the fluid. Therefore, the ability to determine or predict the minimum impeller speed, N_{js}, for the just-suspended state is a critical step in any solid–liquid mixing operation. Techniques for measuring N_{js} are discussed elsewhere.[24,25] In addition, N_{js} has been the subject of significant studies and correlations are available to predict it.[21] The most widely used equation for the determination of N_{js} is the Zwietering equation:[22]

$$Re_{js}^{0.1} Fr_{js}^{0.45} \left(\frac{D}{d_p}\right)^{0.2} X^{-0.13} = S \tag{4}$$

which is often rewritten as

$$N_{js} = S\nu^{0.1} \left[\frac{g(\rho_S - \rho_L)}{\rho_L}\right]^{0.45} \frac{d_p^{0.2}}{D^{0.85}} X^{0.13} \tag{5}$$

With the exception of the density difference, the influence of fluid and particle properties on N_{js} is not significant, as indicated by the small exponents on the kinematic viscosity, ν, the particle diameter, d_p, and the percent solids-to-liquid mass ratio, X ($100 \times m_S/m_L$). Obviously, this correlation fails as $X \rightarrow 0$, because N_{js} can never be zero. This error is handled practically by using $X = 5$ for solids concentrations smaller that 5%. The density difference is the property having the largest influence on N_{js}. Its exponent reflects the effect of the terminal settling velocity of the particles. The exponent on the impeller diameter, D, represents the effect of scale. Note that an exponent of -0.67 on D would imply a scaling rule based on power per unit volume [as it can be derived from Eqs. (2) and (3)].

The effects of the geometry of the impeller, vessel, and its internals are incorporated into the S parameter. Values of S are listed in Table 2[26] for a variety of impellers. The S value varies with D/T, and C/T. Zwietering provided plots of S as a function of D/T and C/T.[22] More recently, Armenante and Uehara Nagamine and Armenante, Uehara Nagamine, and Susanto have sought simple mathematical expressions to describe the effects of geometry (D/T and C/T) on S.[25,27] Their results, reported in Table 3, are yet to be validated with data from large-scale tests and for vessels with dished bottoms. N_{js} is 10–20% lower in dished-bottomed vessels than in flat-bottomed vessels.[28,26] Solids suspension is impractical with conical bottoms.

When the power number of the impellers is taken into account, it becomes clear that axial-flow impellers (e.g., Lightnin A-310, Chemineer HE-3) are able to achieve the just-suspended state at a lower power input than other impellers, such as pitched-blade or disk turbines. The resulting axial flow developed by high-efficiency impellers is higher at the vessel base, per unit of power expended, than that for radial flow impellers. These impellers are also more effective at higher clearances from the vessel base, i.e., larger C/T values.

The use of multiple impellers usually has a negligible impact on N_{js}, as solid suspension is dominated by the hydrodynamics at the bottom of the vessel, which is primarily affected by the lowest impeller.[29]

Chowdhury has pointed out regions of interest where Zwietering's correlation is not as reliable.[30] They include: 1) solids loading below 2% (v/v); 2) high d_p/T values; and 3) high solids loading (>15%, v/v).

Table 2 Value of the S parameter for solid suspension in dished-bottomed vessels

Impeller geometry and location	S value
A-310 ($T/2.4$)	
C = $T/4$	6.9
A-310 ($T/2$)	
C = $T/4$	7.1
30°PBT ($T/3$, $D/2.5$)	
$C = T/4$	6.4
$C = T/6$	7.1
$C = T/8$	7.2
45°PBT ($T/3.3$, $D/2.1$)	
$C = T/4$	4.5
$C = T/8$	4.3
45°PBT ($T/3$, $D/3.5$)	
$C = T/4$	4.8
$C = T/6$	4.6
$C = T/8$	4.2
45°PBT ($T/2.5$, $D/2.8$)	
$C = T/4$	4.7
$C = T/8$	3.4
45°PBT ($T/2$, $D/3.5$)	
$C = T/4$	5.2
$C = T/6$	4.2
$C = T/8$	3.7
45°PBT ($T/2$, $D/6$)	
$C = T/4$	5.5
$C = T/8$	—
45°PBT ($T/1.7$, $D/3.5$)	
$C = T/4$	6.7
$C = T/6$	5.1
$C = T/8$	4.4
45°PBT ($T/1.7$, $D/4.3$)	
$C = T/4$	6.8
$C = T/8$	3.8
45°PBT ($T/1.4$, $D/5.0$)	
$C = T/4$	5.4
$C = T/8$	4.5
45°PBT ($T/3$, $D/4$)	
$C = T/4$	4.4
$C = T/6$	4.1
$C = T/8$	3.7
90°PBT ($T/3$, $D/5$)	
$C = T/4$	4.4
$C = T/6$	4.1
$C = T/8$	4.1

The first term in parentheses refers to the value of the impeller diameter, D; the second is the impeller blade width, B.
(From Ref.[26].)

Cloud Height and Solid Distribution in Stirred Tanks

In solid suspensions, there is a distinct level to which most of the solids are lifted within the fluid even at speeds above N_{js}. The distance from the bottom of the vessel to this level is called the cloud height (H_{cloud}). The liquid below this height is solid rich, while above it there is only an occasional visit by a few small solids.

Recently, Bittorf and Kresta have successfully correlated the cloud height data of Bujalski et al. and Hicks, Myers, and Bakker with the following equation for purely axial impellers:[31–33]

$$H_{cloud} = \frac{N}{N_{js}}\left[0.84 - 1.05\frac{C}{T} + 0.7\frac{(D/T)^2}{1 - (D/T)^2}\right] \tag{6}$$

This equation is appropriate for $0.154 < D/T < 0.52$ and for solids with a terminal settling velocity less than 0.143 m/sec. The correlations produced by Magelli et al. can be used to predict the solid distribution as a function of the position along the vertical axis.[23]

Power Dissipation in Solid–Liquid Stirred Reactors

The power dissipated by the impeller in solid–liquid dispersions can be calculated from an equation similar to Eq. (2), i.e.,

$$P = N_p\bar{\rho}D^5N^3 \tag{7}$$

where the average density of the solid–liquid mixture is now given by

$$\bar{\rho} = \phi\rho_S + (1 - \phi)\rho_L \tag{8}$$

Such a value for the average density is only approximately correct for power calculations (although it is appropriate in most cases), as the solids fraction in the impeller region may differ from ϕ depending on N, i.e., on the solids suspension state.

The volumetric solids faction ϕ is related to X as follows

$$\phi = \frac{(X/100)\rho_L}{\rho_S + (X/100)\rho_L} \tag{9}$$

Interfacial Area in Solid–Liquid Suspensions

The surface area of a given mass of solids is strictly determined by the size and geometry of the particles,

Table 3 Correlations for the S parameter for solid suspension in flat-bottomed vessels

Impeller type	Correlation for S	Source
Dis turbine ($C/T < 1/6$)	$S = \left[0.99(T/D)^{1.40}\exp(2.18(C_b/T))\right]$	Armenante and Uehara Nagamine[25]
Dis turbine ($C/T > 1/5$)	$S = \left[2.10(T/D)^{1.18}\exp(0.24(C_b/T))\right]$	Armenante, Uehara Nagamine and Susanto[27]
6-Blade flat-blade turbine ($C/T < 1/5$)	$S = \left[1.43(T/D)^{1.20}\exp(1.95(C_b/T))\right]$	Armenante and Uehara Nagamine[25]
6-Blade flat-blade turbine ($C/T > 1/4$)	$S = \left[1.78(T/D)^{1.16}\exp(1.77(C_b/T))\right]$	Armenante, Uehara Nagamine and Susanto[27]
6-Blade 45°C pitched-blade turbine ($C/T < 1/4$)	$S = \left[2.28(T/D)^{0.83}\exp(0.65(C_b/T))\right]$	Armenante and Uehara Nagamine[25]
Chemineer HE-3 ($C/T < 1/4$)	$S = \left[3.49(T/D)^{0.79}\exp(0.66(C_b/T))\right]$	Armenante and Uehara Nagamine[25]

and cannot be altered by mixing. However, the solids interfacial area available for effective mass transfer with the liquid depends strongly on the solids suspension state. If the solids are accumulated at the vessel bottom their surface will not be fully exposed to the liquid, and most of the liquid in the vessel will not be brought in effective contact with the solids. This is clearly shown in Fig. 9.[7,34–36] Increasing N when $N < N_{js}$ produces a dramatic increase in the mass transfer rate primarily because the particles' exposed area in Eq. (1) increases with N as more particles become fully suspended. Increasing N beyond N_{js} produces only a moderate increase in the mass transfer rate, as the interfacial area, A, now corresponds to the effective solids surface area, and the mass transfer rate increases solely because k_L increases with P (and hence N), as quantified in the next section. This constitutes an additional reason why solid–liquid operations should always be conducted at an agitation speed $\geq N_{js}$.

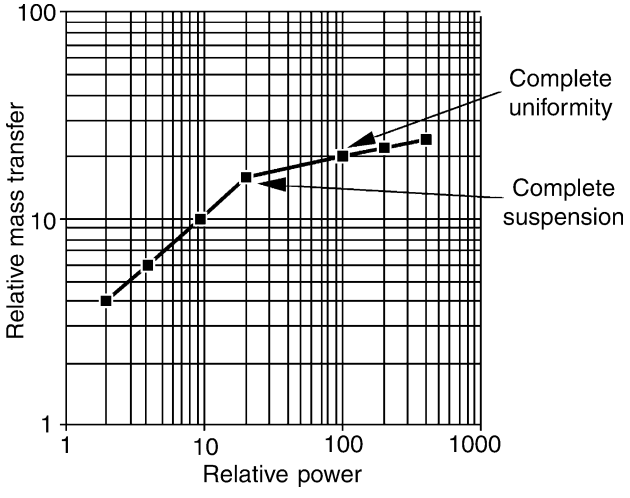


Fig. 9 Relative mass transfer as a function of impeller power. The mass transfer increases sharply up to the point of complete suspension, and at a much lower rate to nearly complete uniformity. (From Ref.[7].)

When $N > N_{js}$, the surface area of the particles, now fully exposed to the liquid, is given by

$$A = a_v V_L = \frac{6\phi}{\psi d_{32pe}} V_L \tag{10}$$

implying that

$$a_v = \frac{6\phi}{\psi d_{32pe}} \tag{11}$$

where ψ is the particle sphericity factor (i.e., the ratio of surface area of a sphere having the same volume as the particle to the surface area of the particle), and d_{32pe} is the Sauter mean equivalent diameter[17]

$$d_{32pe} = \frac{\sum_{i=1}^{i=m} n_i d_{pe\,i}^3}{\sum_{i=1}^{i=m} n_i d_{pe\,i}^2} \tag{12}$$

The equivalent particle diameter, d_{pe} is the diameter of a sphere of equal volume as the particle.

Mass Transfer Coefficient in Solid–Liquid Suspensions

In reactive solid–liquid systems, the process can be dominated by one or more of the following mechanisms: 1) external (extraparticle) mass transfer; 2) internal (intraparticle) mass transfer; or 3) chemical reaction (at the surface, below the surface, or in the liquid). The agitation speed only affects the interfacial area available for mass transfer (if $N < N_{js}$, as described above), and the external mass transfer coefficient

k_L (in all cases). Only processes in which the external mass transfer is partially or completely rate controlling are affected by the agitation speed. Experimentally, this phenomenon can be observed by varying the agitator speed and determining the effect on the overall reaction process rate. If the reaction conversion is affected by N, mass transfer plays a role.

The mass transfer coefficient to a single solid spherical particle immersed in a liquid flowing with velocity v_{SL} past the particle can be calculated from:

$$Sh_{SL} = 2 + 0.6Re_{SL}^{0.5}Sc^{0.33} \tag{13}$$

where the solid–liquid Sherwood number for a single sphere, $Sh_{SL} = k_{SL}(d_p/D_L)$ is a function of the solid–liquid Re, $Re_{SL} = \rho_L \nu_{SL}(d_p/\mu)$.[37] Eq. (13) is of limited use in mixing vessels, where the particle–liquid slip velocity v_{SL} cannot be easily calculated. Instead, the following approach can be used. In any turbulent system, the Kolmogoroff's microscale, λ_K, defines the size of the smallest eddy in the stirred vessel:

$$\lambda_K = \left(\frac{\nu^3}{\varepsilon}\right)^{1/4} \tag{14}$$

Armenante and Kirwan have defined as macroparticles and microparticles solid particles having diameters larger or smaller than λ_K, respectively.[38] Then, the overall particle–liquid mass transfer coefficient for fully suspended macroparticles is given by:[39,40]

$$Sh = 2 + 0.47Re_p^{0.62}Sc^{0.36}(D/T)^{0.17} \quad (\text{for } d_p > \lambda_K) \tag{15}$$

while the following equation applies to microparticles:[38]

$$Sh = 2 + 0.52Re_p^{0.52}Sc^{1/3} \quad (\text{for } d_p < \lambda_K) \tag{16}$$

In these equations, $Sh = k_L d_p/D_L$, and the particle Re for the turbulent system, Re_p, is defined as

$$Re_p = \left(\frac{d_p}{\lambda_K}\right)^{4/3} = \frac{d_p^{4/3}\varepsilon^{1/3}}{\nu} \tag{17}$$

Although Eqs. (15) and (16) were obtained for single-size particles, it is reasonable to assume that they apply to multisized particle dispersions also, in which case d_p should be replaced with d_{32pe}. The effect of agitation on solid–liquid mass transfer can be predicted from these equations. For example, for macroparticles it is:

$$Sh = \frac{k_L d_p}{D_L} \propto Re_p^{0.62} = \left(\frac{d_p^{4/3}\varepsilon^{1/3}}{\nu}\right)^{0.62}$$
$$\Rightarrow k_L \propto \varepsilon^{0.21} = \left(\frac{P}{\rho_L V_L}\right)^{0.21} \propto P^{0.21} \propto N^{0.62} \tag{18}$$

These results show that k_L in solid–liquid systems increases only modestly with power dissipation. For example, a 20% increase in N increases P by 73%, but k_L by only 12%. In conclusion, at agitation speeds below N_{js}, $k_L a_v$ increases significantly with N (as shown in Fig. 9), primarily because the solids' exposed surface area increases with N, as more particles become suspended.[7] Above N_{js}, the value of $a_v = A/V_L$ is constant and k_L increases only modestly with N [Eqs. (15), (16), and (18)]. In many cases, operating slightly above N_{js} is economically optimal.

CONCLUSIONS

Multiphase mixing is a critical element in the analysis of any multiphase reactor system. Multiphase mixing phenomena involving solid–liquid, liquid–liquid, and gas–liquid present not only significant differences, especially in terms of equipment and operation, but also many similarities. Each of these multiphase systems should be analyzed to determine the optimal equipment configuration and operating conditions required to generate an adequate dispersion of the dispersed phase in the continuous phase, a high interfacial area, and a high mass transfer coefficient. Quantitative information of this kind combined with knowledge on the effect of operating variables will result in the ability to predict, estimate, or conduct precisely aimed experiments to determine the rate of multiphase mass transfer for the reactants. Such knowledge will in turn be essential to determine the relative importance of mass transfer vis-à-vis the reaction kinetics in the overall reaction process, thus enabling the designer to predict the behavior of the system, especially during scale-up.

NOMENCLATURE

A	Interfacial area between phases (m^2)
a_v	Interfacial area per unit volume of liquid (m^2/m^3)
B	Baffle width (m)
C	Impeller clearance off the vessel bottom, measured from impeller centerline to vessel bottom (m)

C_b Impeller clearance measured from bottom of the impeller to vessel bottom (m)
ΔC Concentration driving force for mass transfer (mol/m^3)
C_{bulk} Concentration of dissolved component or reactant in bulk liquid (mol/m^3)
$C_{interface}$ Concentration of dissolved component or reactant in liquid at interface between dispersed and continuous phases (mol/m^3)
d_p Particle diameter (m)
d_{pe} Particle equivalent diameter, i.e., diameter of a sphere of equal volume as the particle (m)
d_{32} Sauter mean diameter $= \sum_{i=1}^{i=m} n_i d_i^3 / \sum_{i=1}^{i=m} n_i d_i^2$ (m)
d_{32pe} Sauter mean equivalent diameter $= \sum_{i=1}^{i=m} n_i d_{pe\,i}^3 / \sum_{i=1}^{i=m} n_i d_{pe\,i}^2$ (m)
D Impeller diameter (m)
D_L Diffusivity of dissolved component or reactant in liquid (m^2/sec)
g Gravitational acceleration (m/sec^2)
H Height of liquid in vessel (m)
H_{cloud} Cloud height, i.e., distance from bottom of vessel to height where most of the solids are lifted in solid-suspensions (m)
k_L Mass transfer coefficient (m/sec)
k_{SL} Mass transfer coefficient for a single sphere immersed in a liquid flowing at constant velocity past the sphere (m/sec)
m_L Mass of liquid (kg)
m_S Mass of solids (kg)
$\dot{m}$ Rate of mass transfer of solute or reactant (kg/sec)
N Impeller speed (revolutions/sec)
N_{js} Minimum speed to just suspend solid particles in vessel (revolutions/sec)
P Power dissipation (W)
S Nondimensional parameter in solid–liquid suspension
t Time (sec)
T Vessel diameter (m)
ν_{SL} Liquid velocity past a solid particle (slip velocity) (m/sec)
V_L Volume of liquid in vessel (m^3)
W Width of an impeller blade (m)
X Mass ratio of suspended solids to liquid

Greek Symbols

ε Energy dissipation rate (or power draw) per mass of mixture (W/kg)
ε_{avg} Average energy dissipation rate (or power draw) per mass of mixture (W/kg)
ϕ Volume fraction of dispersed phase, i.e., ratio of volume of dispersed phase total volume of dispersion (–)
λ_K Kolmogoroff microscale of turbulence (m)
ψ Particle sphericity factor, i.e., ratio of surface area of a sphere of same volume as the particle to surface area of particle (–)
μ Viscosity (Pa sec)
ν Kinematic viscosity (m^2/sec)
ρ_L Density of liquid (kg/m^3)
ρ_S Density of solid phase (kg/m^3)
$\bar{\rho}$ Average density of mixture (kg/m^3)

Dimensionless Groups

Fr Impeller Froude number $= N^2 D/g$
Fr_{js} Modified impeller Froude number at N_{js} for solid–liquid suspensions $[\rho_L/(\rho_S - \rho_L)]\, N_{js}^2 D/g$
N_p Power number $= P/(\rho_L N^3 D^5)$
Re Impeller Reynolds number $= \rho N D^2/\mu$
Re_{js} Impeller Reynolds number at N_{js} for solid–liquid suspensions $= N_{js} D^2/\nu$
Re_p Particle Reynolds number in a turbulent system $= \left(d_p^{4/3}\varepsilon^{1/3}\right)\Big/\nu$
Sc Schmidt number $= \mu/(\rho_L D_L)$
Sh Sherwood number $= k_L d_p/D_L$
Sh_{SL} Sherwood number for a single sphere $= k_{SL} d_p/D_L$

REFERENCES

1. Marshall, E.M.; Bakker, A. Computational fluid mixing. In *Handbook of Industrial Mixing: Science and Practice*; Paul, E.L., Atiemo-Obeng, V.A., Kresta, S.M., Eds.; John Wiley & Sons: New York, 2004; 257–343.
2. Akiti, O.; Armenante, P.M. Experimentally-validated micromixing-based CFD model for fed-batch stirred-tank reactors. AIChE J. **2004**, *50* (3), 566–577.
3. Akiti, O.; Yeboah, A.; Bai, G.; Armenante, P.M. Hydrodynamic effects on mixing and competitive reactions in laboratory reactors. Chem. Eng. Sci. **2005**, *60*, 2341–2354.
4. Marchisio, D.L.; Barresi, A.A. CFD simulation of mixing and reaction: the relevance of the micro-mixing model. Chem. Eng. Sci. **2003**, *58*, 3579–3587.
5. Aubin, J.; Fletcher, D.F.; Xuereb, C. Design of micromixers using CFD modelling. Chem. Eng. Sci. **2005**, *60*, 2503–2516.

6. Hemrajani, R.R.; Tatterson, G.B. Mechanically stirred vessel. In *Handbook of Industrial Mixing: Science and Practice*; Paul, E.L., Atiemo-Obeng, V.A., Kresta, S.M., Eds.; John Wiley & Sons: New York, 2004; 345–390.
7. Atiemo-Obeng, V.A.; Penney, W.R.; Armenante, P.M. Solid–liquid mixing. In *Handbook of Industrial Mixing: Science and Practice*; Paul, E.L., Atiemo-Obeng, V.A., Kresta, S.M., Eds.; John Wiley & Sons: New York, 2004; 543–584.
8. Tatterson, G.B. *Fluid Mixing and Gas Dispersion in Agitated Tanks*; McGraw-Hill: New York, 1991.
9. Chudacek, M.W. Impeller power numbers and impeller flow numbers in profiled bottom tanks. Ind. Eng. Chem. Des. Dev. **1985**, *24*, 858–867.
10. Oldshue, J.Y. *Fluid Mixing Technology*; McGraw Hill: New York, 1983.
11. Bates, R.L.; Fondy, P.L.; Corpstein, R.R. An examination of some geometric parameters of impeller power. Ind. Eng. Chem. Proc. Des. Dev. **1963**, *2*, 310–314.
12. Bates, R.L.; Fondy, P.L.; Fenic, J.C. Impeller characteristics and power. In *Mixing*; Uhl, V.W., Gray, J.B., Eds.; Academic Press: New York, 1966; Vol. 1, 111–178.
13. Rushton, J.H.; Costich, E.W.; Everett, H.J. Power characteristics of mixing impellers. Chem. Eng. Progr. Pt. I **1950**, *46*, 395–402; Pt. II **1950**, *46*, 467–476.
14. Armenante, P.M.; Mazzarotta, B.; Chang, G.-M. Power consumption in stirred tanks provided with multiple pitched-blade turbines. Ind. Eng. Chem. Res. **1999**, *38*, 2809–2816.
15. Armenante, P.M.; Chang, G.-M. Power consumption in agitated vessels provided with multiple disk turbines. Ind. Eng. Chem. Res. **1998**, *37*, 284–291.
16. Hudcova, V.; Machon, V.; Nienow, A.W. Gas–liquid dispersion with dual Rushton turbine impellers. Biotechnol. Bioeng. **1989**, *34*, 617–628.
17. Geankoplis, C.J. *Transport Processes and Separation Process Principles*, 4th Ed.; Prentice Hall: Upper Saddle River, NJ, 2003.
18. Guiraud, P.; Costes, J.; Bertrand, J. Local measurements of fluid and particle velocities in a stirred suspension. Chem. Eng. J. **1997**, *68*, 75–86.
19. Oldshue, J.Y. Fluid mixing technology and practice. Chem. Eng. **1983b**, *90* (13), 83–108.
20. Cleaver, J.W.; Yates, B. Mechanism of detachment of colloidal particles from a flat substrate in turbulent flow. J. Colloid Interphase Sci. **1973**, *44*, 464–474.
21. Nienow, A.W. The dispersion of solids in liquids. In *Mixing of Liquids by Mechanical Agitation*; Ulbrecht, J.J., Patterson, G.K., Eds.; Gordon and Breach Science Publishers, 1985; 273–307.
22. Zwietering, T.N. Suspending of solid particles in liquid by agitators. Chem. Eng. Sci. **1958**, *8*, 244.
23. Magelli, F.; Fajner, D.; Nocentini, M.; Pasquali, G. Solid distribution in vessels stirred with multiple impellers. Chem. Eng. Sci. **1990**, *45*, 615–625.
24. Brown, D.A.R.; Jones, P.N.; Middleton, J.C.; Papadopoulos, G.; Arik, E.B. Experimental methods. In *Handbook of Industrial Mixing: Science and Practice*; Paul, E.L., Atiemo-Obeng, V.A., Kresta, S.M., Eds.; John Wiley & Sons: New York; 145–256.
25. Armenante, P.M.; Uehara Nagamine, E. Effect of low off-bottom impeller clearance on the minimum agitation speed for complete suspension of solids in stirred tanks. Chem. Eng. Sci. **1998**, *53*, 1757–1775.
26. Mak, A.T.C. Solid–Liquid Mixing in Mechanically Agitated Vessel. Ph.D. thesis, University College London, 1992.
27. Armenante, P.M.; Uehara Nagamine, E.; Susanto, J. Determination of correlations to predict the minimum agitation speed for complete solid suspension in agitated vessels. Can. J. Chem. Eng. **1998**, *76* (3), 413–419.
28. Guerci, D.; Conti, R.; Sicardi, S. Proceedings of the International College on Mechanical Agitation, ENSIGC, 1986; Toulouse, 3-8–3-24.
29. Armenante, P.M.; Huang, Y.-T.; Li, T. Determination of the minimum agitation speed to attain the just dispersed state in solid–liquid and liquid–liquid reactors provided with multiple impellers. Chem. Eng. Sci. **1992**, *47*, 2865–2870.
30. Chowdhury, N.H. Improved Predictive Methods for Solids Suspension in Agitated Vessels at High Solids Loadings. Ph.D. thesis, University of Arkansas, Fayetteville, AK, 1997.
31. Bittorf, K.J.; Kresta, S.M. Prediction of cloud height for solid suspension in stirred tanks. CHISA Conference Proceedings, Prague, Aug 25–29, 2002.
32. Bujalski, W.; Takenaka, K.; Paolini, S.; Jahoda, M.; Paglianti, A.; Takahashi, K.; Nienow, A.W.; Etchells, A.W.B. Suspension and liquid

homogenization in high solids concentration stirred chemical reactors. Chem. Eng. Res. Des. **1999**, *77*, 241–247.

33. Hicks, M.T.; Myers, K.J.; Bakker, A. Cloud height in solids suspension agitation. Chem. Eng. Commun. **1997**, *160*, 137–155.
34. Nienow, A.W.; Miles, D. The effect of impeller/tank configurations on fluid-particle mass transfer. Chem. Eng. J. **1978**, *15*, 13–24.
35. Chaudhari, R.V.; Ramachandran, P.A. Three phase slurry reactors. AIChE J. **1980**, *26*, 177–201.
36. Conti, R.; Sicardi, S. Mass transfer from freely-suspended particles in stirred tanks. Chem. Eng. Commun. **1982**, *14*, 91.
37. Ranz, W.E.; Marshall, W.R. Evaporation from drops. Chem. Eng. Progr. Pt. 1 **1952**, *48*, 141–146; Pt. 2 **1952**, *48*, 173–180.
38. Armenante, P.M.; Kirwan, D.J. Mass transfer to microparticles in agitated systems. Chem. Eng. Sci. **1989**, *12*, 2781–2796.
39. Levins, D.M.; Glastonbury, J. Application of Kolmogoroff's theory to particle–liquid mass transfer in agitated vessels. Chem. Eng. Sci. **1972**, *27*, 537–542.
40. Levins, D.M.; Glastonbury, J. Particle–liquid hydrodynamics and mass transfer in a stirred vessel. Trans. Inst. Chem. Eng. **1972**, *50*, 132–146.

Multiphase Reactors

Stanley M. Barnett
University of Rhode Island, Kingston, Rhode Island, U.S.A.

INTRODUCTION

Multiphase reactors are very common in the chemical, petroleum, food, mining, pharmaceutical, and semiconductor industries. They are also prominent in many civilian processes. The production of drinking water can consist of several multiphase reactors and separators, removing soluble salts, organic matter, live organisms, and insoluble material. Our waste treatment systems usually have an activated sludge reactor in which air is forced into the solid–liquid waste stream. In the past, when untreated wastes fouled potential drinking water, people produced beer as a method of purification using a multiphase reactor in a procedure called fermentation.

Antibiotics are produced by a three-phase process: Air is sparged into a vessel or tank to supply oxygen, keep the producing cells suspended in a liquid broth, and mix the contents of the vessel. Coal liquefaction is another three-phase process with hydrogen as the gas phase, coal as the solid phase. Fuel cells, a key component of distributed energy systems, are three- or more-phase reactors depending on the type of fuel cell: solid oxide, molten carbonate, phosphoric acid, or proton exchange membrane. They can be catalytic electrochemical reactors with membranes and multiple solid phases. Genetic engineering requires many multiphase reactor systems. Enzymatic peptide synthesis is a multiphase reaction. Sample applications are summarized in Table 1. With such diversity, this entry will focus on a few key types with some new applications. The reactors can be divided into just a few broad categories. These are column contactors with one phase fixed, fluidized or transported, stirred-tank reactors with at least one phase kept distributed by agitation, film reactors, membrane reactors, and large diameter tank reactors, including open top tanks and ponds. The goals for each reactor design are usually identical, maximize conversion of the reactants and yield of product. Cost minimization, energy efficiency, and conforming to space or weight constraints are additional requirements. In some situations, extra care is needed to avoid contamination of the environment or the product. The differing reactor designs optimize heat and mass transfer, usually limiting factors controlling conversion and yield. For example, by passing or channeling of a reactant must be minimized. Flow patterns can be used to generate sufficient turbulence to ensure intimate contact of a reactant or catalyst. Alternately, agitation might be minimized to ensure that shear sensitive organisms live long enough to produce the desired product(s).

Governing equations are the continuity equation, the chemical reactions and their thermodynamic relationships, and the heat, mass, and momentum equations. Elastic behavior of an expanding bed of particles sometimes must be included. These equations can be many and complex because we are dealing with both multiphase and multicomponent systems. Correlations are often in terms of phase-based dimensionless groups such as Reynolds numbers, Froude numbers, and Weber numbers.

MULTIPHASE BIOREACTORS FOR BIOTECHNOLOGY AND ENVIRONMENTAL APPLICATIONS

Activated Sludge Process

The objective of an activated sludge process is to remove soluble and insoluble organics from municipal or industrial wastewaters and convert the reactants into a microbial suspension. Traditionally, the tank is narrow and long enough to provide 6–8 hr of contact between the fluid and air, usually blown in through diffusers (see Fig. 1A). The narrow flow paths are the first step above a simple pond. On a more sophisticated level, the long, narrow tank with diffusers is replaced by a cylindrical tank with an agitator, designed to beat air into the fluid as well as mix the tank contents (see Fig. 1B). Recycle streams allow for stabilization of the reactor by returning live microorganisms to the process. An important design parameter is power input per unit volume, which controls the aeration efficiency.

For certain industrial applications, a biofilm reactor is preferred. In one type a rotating drum becomes coated with the waste stream as it moves through a tank of the wastes. As the drum surface moves from below the liquid level to above, the attached wet film becomes exposed to air (see Fig. 1C). Alternately, the waste liquid stream can be poured onto a bed of

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039327

Table 1 Multiphase reactors: some examples

Reactor types	Applications
Gas–liquid	Absorbers
	Gas phase olefin polymerization
Gas–solid	Fluidized beds
	Trash incinerators
	Catalytic crackers
	Coal gasification
	Reformers
	Ore processing
Liquid–liquid	Enzyme reactions
	Bulk polymerization-casting
	Solution polymerization
Liquid–solid	Ion exchange
	Electrode reactions (including plating)
	Enzyme reactions peptide synthesis
	Liquid phase olefin polymerization
	Cheese making
	High-fructose corn sweeteners
Three-phase:	Hydrogenation
gas–liquid–solid	Fermenters
	Airlift
	Stirred tank
	Enzyme production
	Alcohol
	Fuel cells
	Electrolysis of water
	Trickle bed filters
	Coal liquefaction
	Wastewater treatment
Liquid–liquid–solid	Emulsion polymerization
	Suspension polymerization
	Enzyme reactions
	Detergents
	Proteolysis

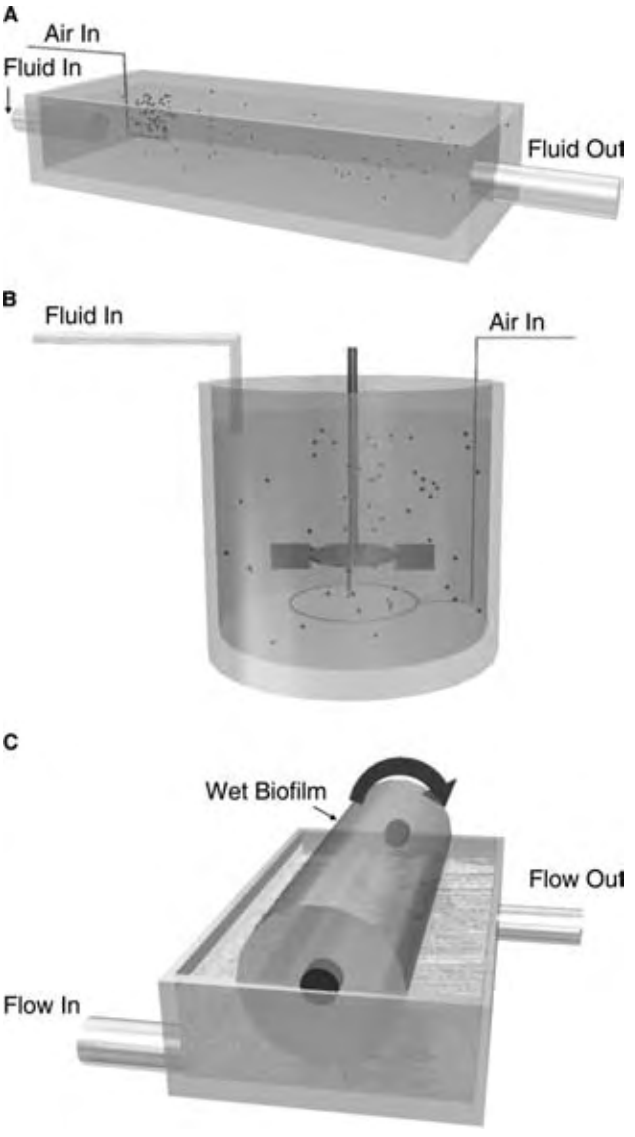


Fig. 1 Bioreactors. (A) activated sludge reactor; (B) stirred tank reactor; and (C) rotating biofilm reactor. *(View this art in color at www.dekker.com.)*

particles—stones, crushed glass, or commercial packing to allow intimate contact between the liquid and upflowing air stream. The key is the improved mass transfer of oxygen from the air stream into a thin film of the wastewater because of the turbulence created by liquid falling from one packing particle or tray to another, without the energy input required for the aerator–agitator of the tanks. There can be an environmental problem if volatile organic chemicals are released into the atmosphere, instead of being converted by the treatment system.

Stirred Tank Reactors and Fermenters

The waste treatment systems described above can also be used to grow algae as a protein-rich product found in foods as nutritional supplements. A pond, or shallow tank similar to that in Fig. 1(A), provides the large surface area/volume ratio required for photosynthetic processes.

In the pharmaceutical industry, the reactors are closed to maintain aseptic conditions, and agitation and aeration are provided as needed. The classic example is the Monod chemostat, often found in laboratories. A representative chemostat, fermentation vessel, or bioreactor, is shown in Fig. 2 with an air sparger, agitator, and feed or exit streams. The continuous flow system is usually replaced by a batch process for production in the pharmaceutical and biotechnology industries. The general relationships for a constant volume system are developed in terms of cell concentration (X), often the component of interest, and substrate concentration (S), such as a sugar or oxygen,

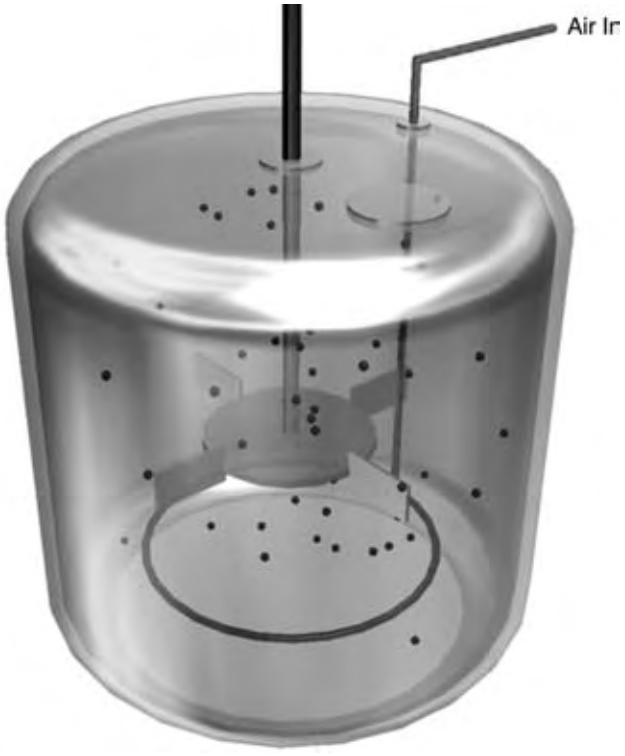


Fig. 2 Stirred tank reactor. *(View this art in color at www.dekker.com.)*

the contributing component.[1] Simple rate equations for cells and substrate are:

$$dX/Xdt = \mu = \mu_{max}S/(K_s + S)$$

and

$$dS/dt = -\mu_{max}SX/Y_{X/S}(K_s + S)$$

Here, μ is the specific growth rate, K_s the Monod coefficient, and $Y_{X/S}$ the yield of cells per unit of substrate. After the integration and use of initial substrate (S_0) and cell (X_0) concentrations the following results were obtained:

$$(X_0 + Y_{X/S}(S_0 + K_s))\ln[(X_0 + Y_{X/S}(S_0 - S))/X_0] - K_sY_{X/S}\ln(S/S_0) = \mu_{max}(X_0 + Y_{X/S}S_0)t$$

These equations can be easily adjusted to reaction conditions, for example, replacing S with light intensity for photosynthetic reactions.[2] A hybrid reactor, consisting of suspended biomass, as above, plus particles coated with biofilm, can be evaluated by segregating the two systems for separate analysis and then adding the two contributions.

S, substrate or reactant concentration, is often controlled by factors such as agitation. Mass transfer of a reactant such as oxygen is correlated with power input per unit volume. Power input to a multiphase stirred vessel is complex. Most commonly used correlations in the literature are valid over a narrow range of conditions. Power must be total power, which includes mechanical power input plus that from bubble expansion or other multiphase issues. The total power number correlates well over a broad operating region with impeller liquid pumping and gas Froude numbers plus various geometric ratios, each to the 0.25 power.[3]

Stirred tank reactors are also common in the chemical and plastics industry. Multiphase processes include emulsion and suspension polymerization plus bulk polymerization of high-impact polystyrene.

Oscillatory Flow Reactors

These are tubular reactors with intensified mixing due to built-in eddy creating and shedding devices. These devices might be baffles coupled with positive displacement devices capable of creating oscillatory flow patterns. Both macro- and micromixing are enhanced, although Reynolds numbers are typically about 50. Because mixing control can be independently achieved by means of amplitude and frequency of oscillation, the mixing effect is decoupled from the flow rate.[4,5] This reactor is an active area of research.

Membrane Reactors

Nano- and ultrafiltration membranes will retain biological catalysts, such as enzymes, which are in the 10,000–100,00 molecular weight range. Larger particles, such as whole cells, can be retained by microfilters. Membrane materials can be in the form of flat sheets as in a filter press type operation or in the spiral wound format. They can also be formed as tubes for a shell and tube arrangement. Or they can be set up as rotating cylinders. The feed stream usually flows tangentially across the surface, helping to minimize cake buildup on the membrane surface. The permeate exits perpendicular to the membrane surface. Membrane pore size or molecular weight cutoff is chosen on the basis of whether we want the substrate and product to be retained or passed through the membrane. Membranes can also be hydrophilic or hydrophobic.

An example is the production of glycerol and fatty acids from plant oils, such as olive oil, catalyzed by the enzyme lipase. The oil flows on the feed side, exiting with the fatty acid product in the concentrate stream shown to the right of the membrane in Fig. 3.

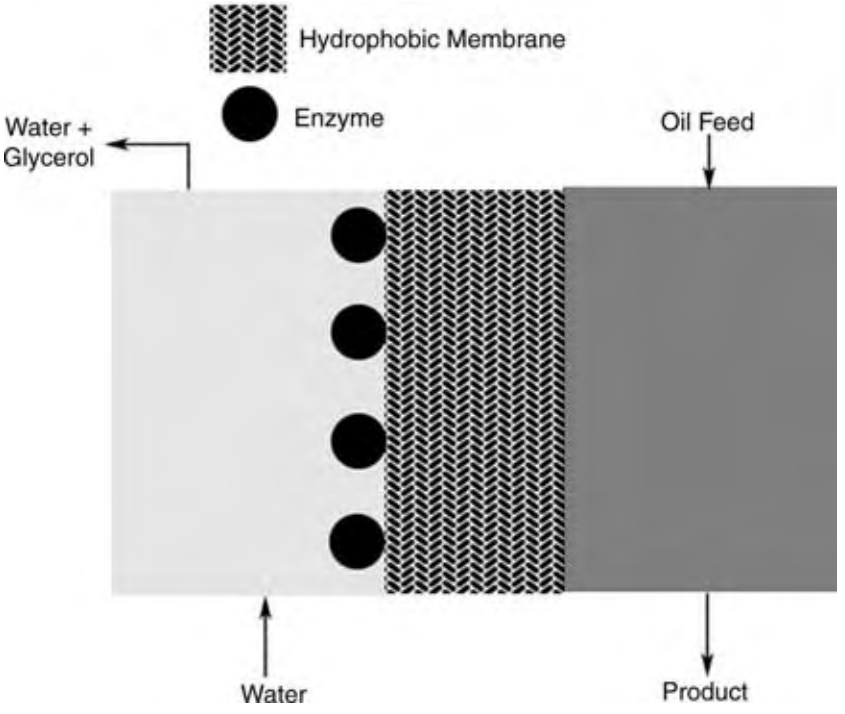


Fig. 3 Multiphase membrane reactor.

The lipase, shown as black circles in Fig. 3, can be attached to the hydrophobic membrane or allowed to wet the membrane surface, taking advantage of the surface-active properties of the enzymes. The hydrophobic oil can also wet the membrane pores and contact the lipase. The glycerol produced by the reaction is water-soluble and moves into the aqueous permeate stream on the left side of the membrane. The undissociated acid returns to the oil stream.[6]

MULTIPHASE REACTORS IN THE ENERGY AND CHEMICAL INDUSTRIES

Packed Bed Reactors

The standard, when one considers multiphase reactors, has become more complex over the years. Most can be classified as reactions over heterogenous catalysts. The catalytic activity occurs in one phase, the solid phase, while transport of the reactants occurs in a gas or liquid phase, or both. A common example is the catalytic converter for automobile exhaust gas. The key steps for a packed bed reactor are:

1. Bulk movement of reactants to the catalyst particle.
2. Mass transfer from the gas or liquid phase to the surface of the catalyst.
3. Diffusion of the reactants from the catalyst surface to the internal pore surface.
4. Adsorption of key molecules onto the catalyst site.
5. Reaction proceeds.
6. Desorption of product from the catalyst site.
7. Pore diffusion to the outer surface of the particle.
8. Interphase transport into the liquid and gaseous phases.

Equations can be written for each step, but it is easier to determine the rate controlling step(s) and focus on just that section of a complex process. In Fig. 4, the cartoon on the right provides a view of the reactant entering and product leaving a catalyst particle.

For industrial reactors, the effectiveness factor (η) is used to provide a measure of the actual reaction rate, as affected by operating conditions, in comparison to the intrinsic reaction kinetics.[7] Assuming that the trickle bed reactor shown in Fig. 4 is operated so that interphase transport of one of the reactants, steps 2 or 8 above, is controlling,

$$\eta = \frac{k_l a \tanh\left(L\sqrt{k/D_a}\right)}{k_l a L\sqrt{k/D_a} + k \tanh\left(L\sqrt{k/D_a}\right)}$$

The rate for this first-order reaction is $-\eta k c_a$, where, k is the reaction rate constant, c_a the concentration of the limiting reactant, L the pore length, k_l the mass

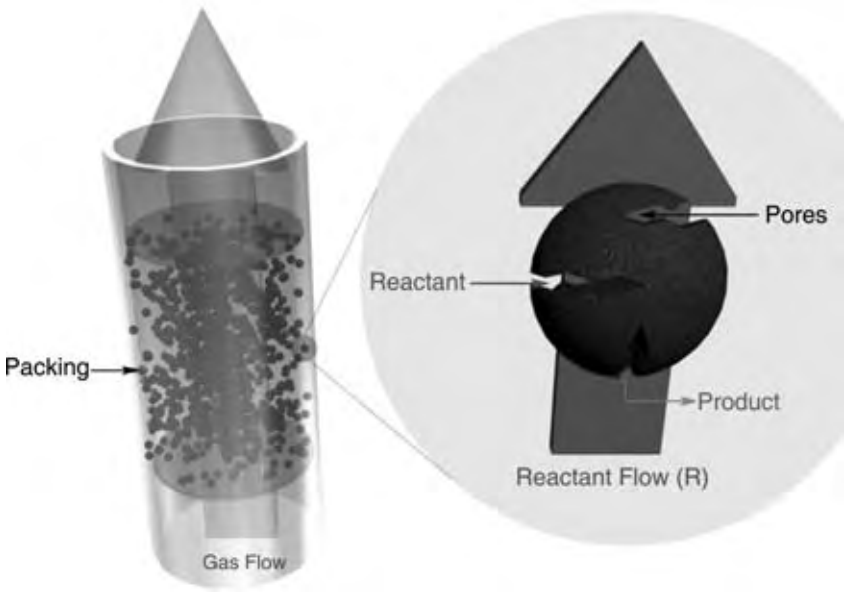


Fig. 4 Packed bed reactor regions. *(View this art in color at www.dekker.com.)*

transfer coefficient, a the specific surface area, and D_a the diffusivity of the limiting substance.

This rate can be added to the global rate equation. Here, ε is the porosity and u_s the superficial velocity.

$$u_s \frac{\partial c_a}{\partial z} = D_a \left(\frac{1}{r} \frac{\partial c_a}{\partial r} + \frac{\partial^2 c_a}{\partial r^2} \right) + \varepsilon \eta c_a$$

A related energy–balance relationship will provide temperature profiles.

Pressure drop is determined using the Ergun equation:

$$\frac{-\Delta P}{L} + \rho_f g = \frac{E_1 \mu_f (1 - \varepsilon)^2 U}{D_p^2 \varepsilon^3} + \frac{E_2 \rho_f (1 - \varepsilon) U^2}{D_p \varepsilon^3}$$

The Es are constants best determined by experiment, ε porosity, ρ_f fluid density, U superficial velocity, and D_p particle diameter.[8]

Two-phase flow through a packed bed, gas–solid flow, for example, requires modification of the Ergun equation. The change includes accounting for mixture density and viscosity. Particle-bed interaction factors must also be considered.[9]

Reforming

To provide hydrogen for fuel cells, various fuels can be used as a feedstock. Currently, methane is the source of choice for hydrogen. Fuel cells are accompanied by a reactor called a reformer, in which the methane is reacted with steam over a supported nickel catalyst at temperatures above 500°C. This is a gas–solid packed bed type reactor.

Hydrocarbon Hydrocracking

A hydrocracker is a three-phase operation. The gas phase supplies hydrogen, the liquid phase supplies the heavy hydrocarbons, and the catalyst is the solid phase. This unit can be operated as a trickle bed reactor, with gas and liquid phases fed in at the top. Products, removed from the bottom, are in both the gas and the liquid phases. The key steps and analysis are similar to the packed bed reactor above. Pressure drop and holdup can be determined from the Ergun equation and gas and liquid phase Reynolds numbers. Relationships for transitions to pulsating and other flows can also be developed.[8]

A hydrocracker can also be run with both liquid and gas fed from the bottom and the solids kept in suspension by drag forces of the fluids. This reactor is a fluidized bed reactor explained below.

Coal Liquefaction

A related reactor is that for coal liquefaction, which can be carried out in a three-phase slurry bubble column (see Fig. 5). Hydrogen can be supplied at the bottom of a column of downcoming product—oil. The solid coal reactant is blended with the product or carrier oil and fed at the top. The generic process depicted in Fig. 5 is a generalization of the liquefaction reactor in the Exxon Donor Solvent Process. As the gas flow rate increases, the bubbles change from uniformly small to chaotic. In the H–coal process, both the gas and a coal–oil slurry are fed from the bottom in an ebullating-bed reactor. Catalyst solids are fed from the top. This reactor operates as an expanded

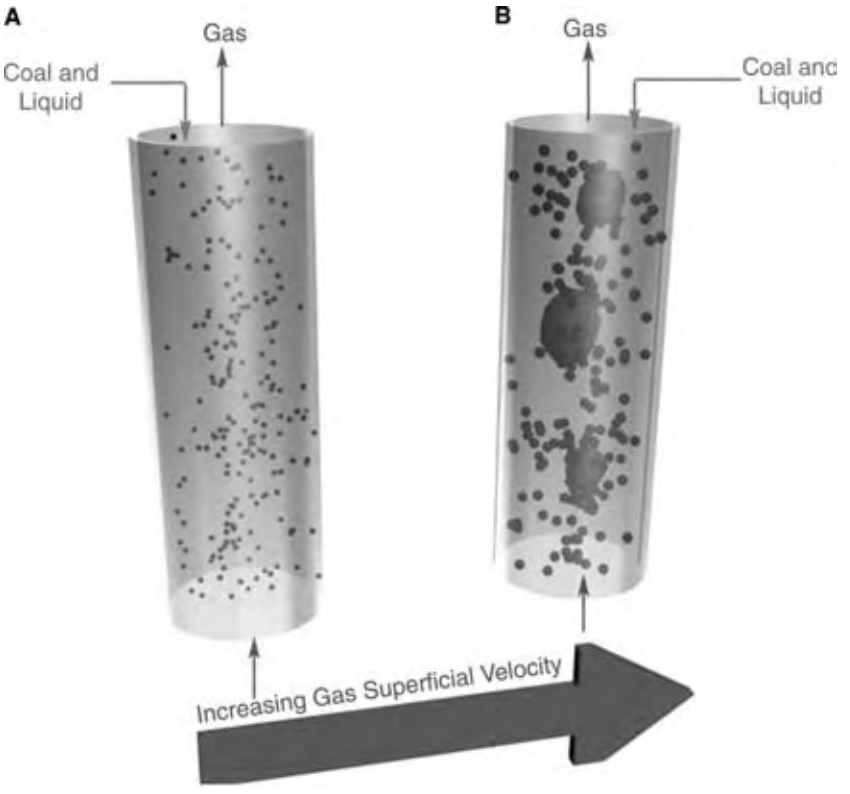


Fig. 5 Coal liquefaction column. *(View this art in color at www.dekker.com.)*

bed type fluidized bed reactor. The products of these reactors are further processed using other multiphase reactors. The solvent-refined coal process (SRC) is related.[10] Coal gasification, Fisher–Tropsch syntheses, and methods of impurity removal, such as desulfurization, are also in three-phase columns.

Understanding three-phase flow dynamics is necessary for design of these reactors. A major aid is the classification scheme of Fan, depending on stream direction and degree of fluidization, expanded bed or transported bed. For example, a continuous liquid phase reactor, with gas and liquid inlets at the bottom and solids at the bottom, is classified as Mode E-I-a-1 if it operates as an expanded bed and Mode T-I-a-1 if all phases are flowing together, called transport.

Slurry bubble columns, with liquid and solid countercurrent to the gas stream, Mode E-1-a-2 or liquid batch Mode E-111-a, demonstrate three flow patterns: the homogenous bubble regime (see Fig. 5A), the larger coalesced bubble or churn flow regime (Fig. 5B), and the huge bubble or slugging regime. Gas bubble size depends on gas and liquid velocities and holdups. Predicting gas holdup is a complex function of flow velocities, solids loading, and Newtonian or non-Newtonian fluid behavior.[11,12]

For gas holdup, bypassing and the effect of solids loading are important issues. Gas holdup increases with solids loading at low solids level, peaks and then decreases with solids content. Gas holdup also increases with superficial gas velocity. Liquid holdup is a function of superficial flow rate and solid particle size. Important dimensionless groups are the liquid and gas Froude numbers, the liquid Reynolds number, and the Weber number. The transitions from channel flow to slugging flow, or to plug flow or homogenous distribution flow are needed design factors.[13]

Dispersed phase holdups, usually gas and solids, have traditionally been determined by differential pressure tranducers, conductivity measurements, photography, particle counters, and volume changes. Ultrasonic techniques have also been used.[14] Liquid flow is complex, with up and down patterns. Local liquid velocities tend to be higher than predicted by energy balances. Circulation cells occur, increasing residence time distributions. Solids suspension depends on passing a critical fluid velocity, about two to four times the particle terminal velocity. Solids loading and equipment design are important variables.[10,15]

Three-Phase Air Lift Reactors

This annular reactor operates with bubbles of gas causing the liquid phase to rise in the central section of the reactor (see Fig. 6). As the gas leaves in the separator or reversal zone at the top of the center column, the heavier liquid falls in the annular downcomer, and is lifted again in the bottom flow-reversing

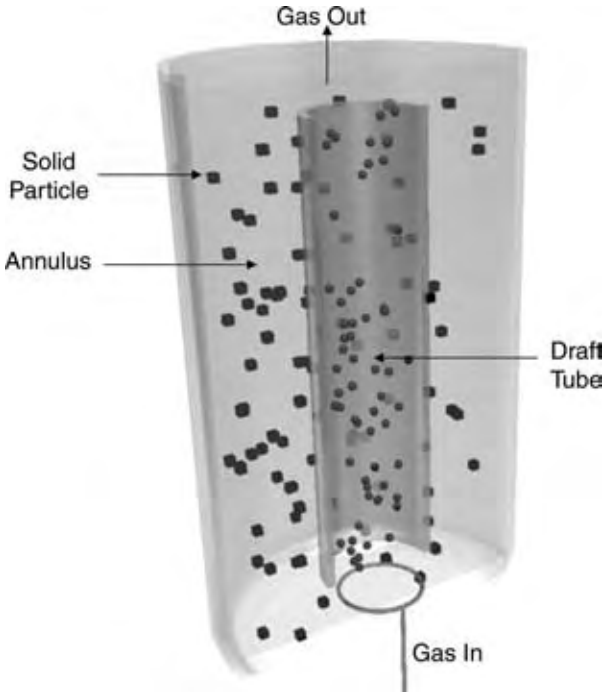


Fig. 6 Annular reactor. *(View this art in color at www.dekker.com.)*

zone by the gas distribution system. The solids circulate with the liquid. This reactor and variations are also used as fermenters, for immobilized cell reactors and for catalytic hydrogenation. Modeling is accomplished by separately evaluating the riser, downcomer, and two transition zones and then combining the modules.[16] This type of reactor has also been used with microbubbles to gain interfacial area.

Fluidized Beds

Instead of a fixed bed reactor, the solid particles, catalyst or reactant, can be moving or fluidized. Normally, the entire solid particle bed is lifted by the upflowing gas stream (see Fig. 7). The minimum superficial fluidizing velocity (U_{mf}), considered a property of the solid particles, can be estimated from:

$$U_{mf} = 0.0009(g\Delta\rho)^{0.934} d_w^{1.8}/\mu^{0.87}\rho_f 0.006$$

Here, g is the gravitational acceleration, ρ_f the density of the gas, ρ_s the density of a particle, $\Delta\rho = \rho_s - \rho_f$, and μ is the gas viscosity. The d_w is calculated as an inverse weighted average from a sieve analysis. Entrainment increases with gas velocity and viscosity, vessel diameter, and particle attrition. It decreases with increasing freeboard height and particle density.[17] Pressure drop is nearly constant across the fluidization region. Clusters of particles can be a problem, but an issue appears to be defining a cluster, by size or effect on reactor hydrodynamics and erosion.[18]

Once U_{mf} is reached, various operating regions are encountered. At first, the bed expands, and then a bubbling/slugging region ensues. The desired turbulent region follows and as gas velocity increases beyond the particle terminal settling velocity, a circulating bed is feasible. At a still higher gas velocity, the solids can be conveyed.[19]

High temperatures have been found to effect particle–particle interaction, beyond the changes in fluid properties such as density and viscosity. Particle classification can change.[20]

Applications include coal and trash combustion, coal and biomass gasification, and chemical, including polymer, synthesis. Both particles and the fluidizing gas stream change properties during the process. Coal is converted to ash, carbon dioxide, and carbon monoxide. The coal particle can also become fragmented. Coal and steam can yield hydrogen and carbon monoxide.

In a spouted bed version of a fluidized reactor, gas enters from a nozzle located at the bottom of the reactor, solids flow in from the sides, or annulus, and are carried up again forming a spout. At the top of the spout, the solids move to the annulus and return to the bottom of the spout.

Using a hybrid arrangement, a fermentation vessel can also be based on fluidized bed technology. The solid phase consists of particles surrounded by an organism-rich biofilm. It is kept in suspension by a flowing liquid medium saturated with air or, both the gas and liquid enter the bottom of the column as separate streams and both help suspend the solids.[21]

Fuel Cells

A growing application of multiphase reactors is for fuel cells. There are many types but they have certain characteristics in common. The basic construction of the fuel cell, as shown in Fig. 8, is that of cathode–electrolyte–anode connected to a load, such as an electric motor. Oxygen is fed to the cathode and hydrogen to the anode. The electrode structure is porous so that gas can enter from one side and the electrolyte from the other. In an acid electrolyte, as in the phosphoric acid fuel cell, the hydrogen ionizes, releasing electrons. The hydrogen ions move across the electrolyte from anode to cathode. At the cathode, they react with oxygen to form water. In an alkaline electrolyte, as is found in the potassium hydroxide fuel cell, the oxygen at the cathode reacts with water in the electrolyte and electrons returning from the load circuit, to form hydroxyl ions. In this situation, it is

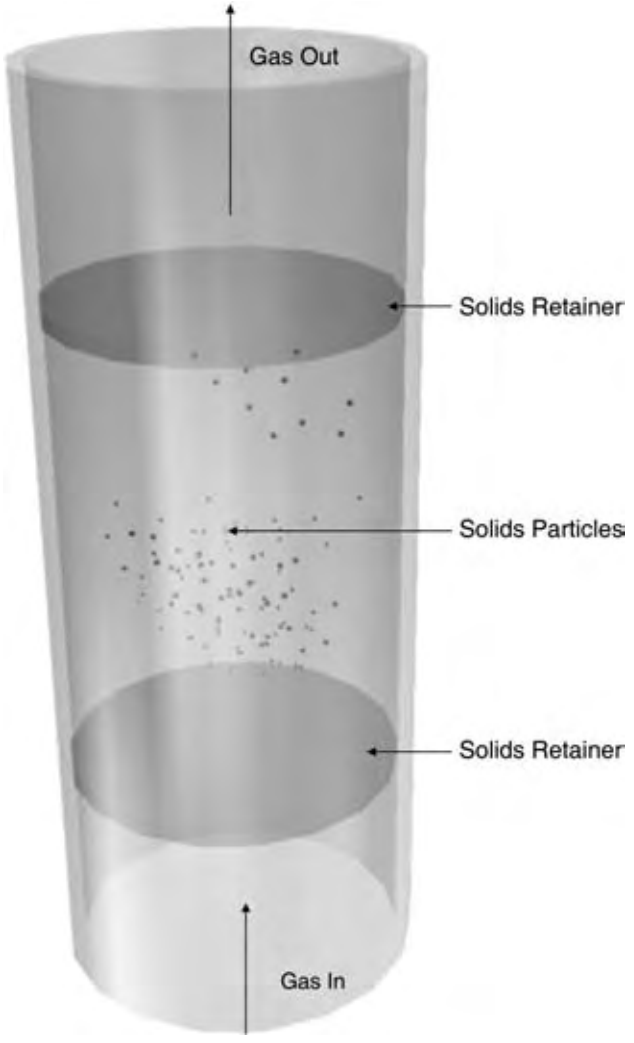


Fig. 7 Fluidized bed reactor. *(View this art in color at www.dekker.com.)*

the hydroxyl ions that move across the electrolyte to the anode to react with hydrogen to form water and release electrons to drive the load. For both types of cells, the anodes and cathodes are usually carbon with a dispersed platinum catalyst. The fuel cell is thus a gas, liquid, multisolid reactor producing energy. Key variables used to compare systems are power density (power produced/unit volume) and specific power (power/mass). To replace the internal combustion engine in automobiles, the goal for power density is bettering 1 kW/L.

For transportation applications, such as automobiles and buses, the liquid electrolyte is replaced with a solid electrolyte-polytetrafluoroethylene containing pendent sulfonated fluoroethylene groups. Water is absorbed onto the sulfonated side chains, hydrogen ions are weakly attracted and are able to move from site to site, creating a weak acid. With an electrode structure consisting of a carbon-supported platinum catalyst, the result is a gas–solid–solid–solid–solid system that can operate at no more than 80°C.

A high-temperature fuel cell capable of operating in the 800–1100°C range, is the solid oxide fuel cell. The high temperatures allow for increased energy efficiencies due to easily recoverable heat. The electrolyte is zirconia doped with small amounts of yttria and other components. The anode can be a zirconia cermet, the metal is nickel. Instead of hydroxyl ions moving from cathode to anode as above, oxygen ions are transported across the electrolyte moving from one vacant oxygen site to another across the crystal structure. This cell is another example of a gas–multisolid phase system.[22]

Hydrogen by Electrolysis

One source of hydrogen for the hydrogen economy is electrolysis of water. Hydrogen is evolved at the cathode. Focusing on the hydrogen generating region, the cathode and the electrolyte between the membrane separator and cathode, the cathode reaction that can occur is:[23]

$$2H_2O + 2e^- \rightarrow H_2 + 2OH^-$$

The key equations are the conservation and momentum equations with a buoyancy term and an added term for interphase friction (F). The latter is given by:

$$F = 0.75 C_d \rho_L \alpha_L \alpha_g |u_L - u_G| d_b$$

C_d, the drag coefficient, is a function of the Reynolds number (Re) and the Weber number (We), both based on the bubble diameter (d_b) and slip velocity ($|u_L - u_G|$).[24,25] An equation can be added for transport across the gas–liquid interface. Ionic species diffusion must be accounted for by the Nernst equation with the current density from Faraday's law, equations for conservation of current and ionic species, and the diffusivities modified by the gas fraction. Solving for all equations with appropriate boundary conditions indicates electrolyte velocity to be the controlling variable for hydrogen production.[23]

Microfluidic Channel Reactors

Recently, researchers have been carrying out multiphase reactions in nanometer to micrometer sized channels in terms of diameter or height and width. In one example, hydrogenation of a liquid flowing over a solid catalyst immobilized on the wall is carried out

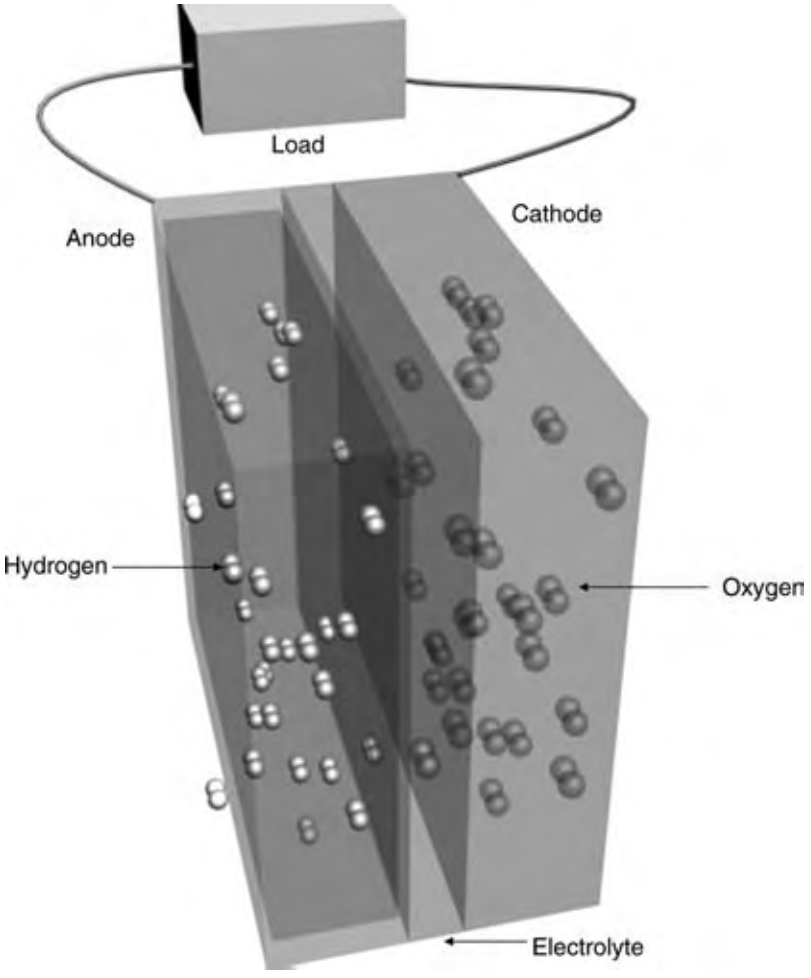


Fig. 8 Fuel cell. *(View this art in color at www.dekker.com.)*

by gaseous hydrogen flowing through the channel. The specific surface area (a) used to calculate the effectiveness factor for the packed bed reactor above is 100 times or more greater for a microreactor than for conventional reactors because of the large number of microchannels that can be placed in a given volume.[26] Wall effects, normally neglected in reactor flow calculations, must be considered in calculations for nanometer range channels. Current practical applications are focused on analysis: the lab-on-a-chip. The oscillatory flow reactor introduced above can utilize these narrow channels.

An application of microfluidic reactors is the development of a membraneless fuel cell. Two streams, one containing a fuel such as methanol, the other an oxygen-saturated acid or alkaline stream, are merged without mixing. The laminar flow pattern in the narrow channel helps to maintain separate streams without the use of membrane separators. Opposite walls function as the electrodes and are doped with catalyst. Ion exchange, protons for the acid system, takes place through the liquid–liquid interface. This is an example of a solid–liquid–liquid–solid multiphase reactor.[27]

CONCLUSIONS

Multiphase reactor types are highly varied. The simplest approach to analyzing and predicting their behavior is to focus on the rate limiting steps or segment the reactor and model each segment and its contributions separately. Correlations are invariably a function of phase-based Reynolds and Froude numbers. Fractional volumes and properties of the solids are factors. Where interfacial tension is an important factor, the Weber number can be added.

Because of energy, sustainability, and security issues, multiphase reactors will have an ever-increasing role in all the countries. Considerable effort is needed to better predict the very complex phenomena occurring in the many types of multiphase reactors.

REFERENCES

1. Blanch, H.W.; Clark, D.S. *Biochemical Engineering*; Marcel Dekker: New York, 1996.

2. You, T.; Barnett, S.M. Effect of light quality on production of extracellular polysaccharides and growth rate of Porphyridium cruentum. Biochem. Eng. J. **2004**, *19*, 251–258.
3. Gray, D.J.; Treybal, R.E.; Barnett, S.M. Mixing of single and two-phase systems. AIChE J. **1982**, *28*, 195–199.
4. Stonestreet, P.; Van Der Veeken, P.M.J. The effect of oscillatory flow and bulk flow components on residence time distribution in baffled tube reactors. Chem. Eng. Res. Des. **1999**, *77* (A8), 671–684.
5. Stonestreet, P.; Harvey, A.P. A mixing methodology for continuous oscillatory flow reactors. Chem. Eng. Res. Des. **2002**, *80* (A1), 31–44.
6. Prazeres, D.M.F.; Cabral, J.M.S. Enzymatic membrane reactors. In *Multiphase Bioreactor Design*; Cabral, J.M.S. et al., Ed.; Taylor & Francis: New York, 2001.
7. Nauman, E.B. *Chemical Reactor Design*; Wiley: New York, 1987.
8. Dudukovic, M.P.; Devanathan, N.; Holub, R. Multiphase reactors: models and experimental verification. Rev. Inst. Francais Pet. **1991**, *46*, 439–466.
9. Wang, Z.L.; Ding, Y.L.; Ghadiri, M. Flow of a gas–solid two-phase mixture through a packed bed. Chem. Eng. Sci. **2004**, *59*, 3071–3079.
10. Dadyburjor, D.B.; Liu, Z. Coal liquefaction. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 5th Ed.; John Wiley & Sons: Hoboken, NJ, 2004; vol. 6, 832–869.
11. Fan, L.-S. *Gas–Liquid Solid Fluidization Engineering*; Butterworths: Boston, MA, 1989.
12. Fan, L.-S. Bubble dynamics in liquid–solid suspensions. In *Developments in Fluidization and Fluid Particle Systems*; Chen, J.C., Ed., AIChE Symp. Ser. 91; 1995, 1–44.
13. Mikkilineni, S. Hydrodynamic Design of Multiphase Bubble Columns. Ph.D. thesis, University of Rhode Island, Rhode Island, 1988.
14. Zheng, Y.; Zhang, Q. Simultaneous measurement of gas and solid holdups in multiphase systems using ultrasonic technique. Chem. Eng. Sci. **2004**, *59*, 3505–3514.
15. Walas, M. *Chemical Process Equipment*; Butterworth-Heinemann: Stoneham, MA, 1990.
16. Livingston, A.G.; Zhang, S.F. Hydrodynamic behavior of three-phase (gas–solid–liquid) airlift reactors. Chem. Eng. Sci. **1993**, *48* (9), 1641–1654.
17. Porter, et al. Solids drying and gas solid systems. In *Perry's Chemical Engineers' Handbook*, 6th Ed.; Perry, R.H., Green, D., Eds.; McGraw-Hill: New York, 1984.
18. Chen, J.C. Clusters. In *Progress in Fluidization and Fluid Particle Systems*; King, D., Ed.; AIChE Symp. Ser. *92*, 1997, 1–5.
19. Grace, J.R. Agricola aground: characterization and interpretation of fluidization phenomena. In *Fluidized Processes*; Weimer, A.W., Ed.; AIChE Symp. Ser. *88*, 1992.
20. Cui, H.; Chaouki, J. Effects of temperature on local two-phase flow structure in bubbling and turbulent beds of FCC particles. Chem. Eng. Sci. **2004**, *59*, 3413–3422.
21. Lee, H.; Barnett, S.M. A predictive model for the allowable operating velocities and the biomass concentration in a three phase fluidized biofilm reactor. J. Ind. Eng. Chem. **2003**, *9*, 202–211.
22. Blomen, L.J.M.J.; Mugerwa, M.N. *Fuel Cell Systems*; Plenum Press: New York, 1993.
23. Mat, M.D.; Aldas, K.; Ilegbusi, O.J. A two-phase flow model for hydrogen evolution in an electrochemical cell. Int. J. Hyd. Energy **2004**, *29*, 1015–1023.
24. Clift, R.; Grace, J.R.; Weber, M.E. *Bubbles, Drops and Particles*; Academic Press: New York, 1978.
25. Ilegbusi, O.J.; Iguchi, M.; Nakajima, K.; Sano, M.; Sakamoto, M. Modeling mean flow and turbulence characteristics in gas agitated bath with top layer. Metall. Mater. Trans. B **1998**, *29B*, 211–222.
26. Kobayashi E.J., et al. Hydrodynamic design of multiphase bubble columns. Science **2004**, *304*, 1305–1308.
27. Choban, E.R.; Markoski, L.J.; Wieckowski, A.; Kenis, P.J.A. Microfluidic fuel cell based on laminar flow. J. Power Sources **2004**, *128*, 54–60.

Nanoimprint Technology and Its Applications

L. Jay Guo
Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan, U.S.A.

INTRODUCTION

Nanoimprint is an emerging lithographic technology that promises high-throughput patterning of nanostructures. Based on the mechanical embossing principle, nanoimprint technique can achieve pattern resolutions beyond the limitations set by the light diffractions or beam scatterings that exist in other conventional techniques. The basic principles of nanoimprint and some of the recent progress in this field are reviewed in this entry. Nanoimprint not only can create resist patterns as in lithography, but can also imprint functional device structures in polymers. This property can be exploited in several potential applications in the area of electronics, photonics, microelectrical-mechanical system (MEMS)/BioMEMS, and biotechnology.

PRINCIPLES OF NANOIMPRINT

The ability to replicate patterns from micro- to nanoscale is of importance to the advance of micro- and nanotechnologies and the study of nanosciences. Critical issues that need to be considered in developing new lithography methodologies include resolution, reliability, speed, and overlay accuracy. Considerable industrial effort has been devoted to the leading-edge UV optical lithography, and the so-called next generation lithography (NGL) techniques by exposing resist material with energetic beams from Extreme UV, electron-beam, ion-beam, or x-ray sources. But a lot of issues are yet to be solved in deep ultra-violet (DUV) photolithography and various NGL methods, and the high costs of such equipments are prohibitive for many users.

In recent years, researchers have investigated a number of alternative and potentially low-cost techniques; most notably, microcontact printing (μCP), nanoimprint, atomic force microscope (AFM) lithography, and dip-pen lithography. Since mid-1990s, nanoimprint lithography (NIL)[1,2] has emerged as one of the most promising technologies for high-throughput nanoscale patterning. In this method and its successful variants such as the step-and-flash imprint lithography (S-FIL), pattern replication is done nontraditionally by mechanical deforming of the resist materials, which completely free itself from the resolution-limiting factors such as light diffraction and beam scattering that are often inherent with the more traditional approaches. Investigations by several researchers in the sub-50 nm regime indicate that imprint lithography resolution is only limited by the resolution of the mold or template fabrication process.

Imprinting-based lithography also presents many new challenges to researchers. Many critical issues need to be addressed for the further progress of this technology, such as its limitations in handling complex patterns with varied feature density, patterning over topographies, and pattern alignment, and defect and critical dimension (CD) control, to name an important few. There is also great demand in new material systems with properties more suitable for the imprint application, where off-the-shelf commercial polymers have become inadequate to satisfy the special requirement in nanoimprinting. In addition, new techniques are needed for many nontraditional microelectronics applications. The basic principle and component requirements for nanoimprinting and some of the recent development in the area of imprint technology and its various applications are reviewed in this entry. The readers are also referred to a recent book, *Alternative Lithography*,[3] which includes chapters dedicated to the discussion of various aspects of nanoimprint lithography written by several experts in the fields.

The principle of nanoimprint is quite simple. As shown in Fig. 1A, NIL uses a hard mold that contains nanoscale features defined on its surface to emboss into polymer material cast on the wafer substrate under controlled temperature and pressure, thereby creating thickness contrast in the polymer material, which can be further transferred through the resist layer via an O_2 plasma-based anisotropic etching process. Nanoimprint has the capability of patterning sub-10 nm features,[4] yet only entails simple equipment setup and easy processing. This is the key reason that NIL attracted wide attention within only a few years after its inception.

Most of the device fabrications require several lithography steps and precise overlay. A derivative of imprint lithography, S-FIL, provides an effective solution to address registration issues. It uses a transparent fused silica template, facilitating the viewing of alignment marks on the template and wafer simultaneously.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120027069

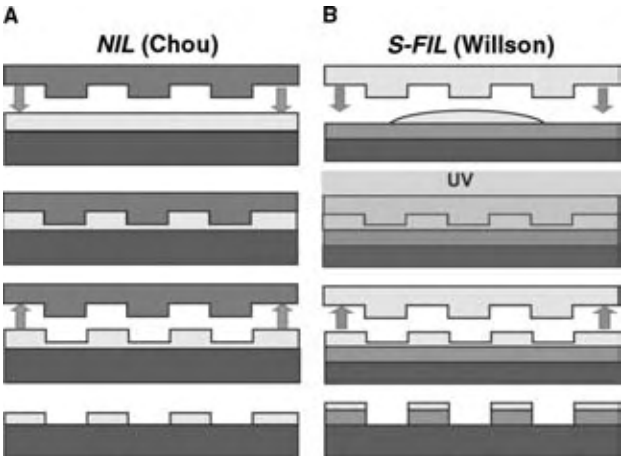


Fig. 1 Schematic of (A) nanoimprint lithography and (B) step-and-flash imprint lithography. *(View this art in color at www.dekker.com.)*

In addition, the imprint process is performed at low pressure and room temperature, minimizing magnification and distortion errors. In the S-FIL process, as illustrated in Fig. 1B, the substrate is first coated with an organic transfer layer; then a surface-treated, transparent template with surface relief patterns is brought close and aligned to the coated substrate. Once in proximity, a drop of low viscosity, photopolymerizable organosilicon solution is introduced into the gap between the template and the substrate. The organosilicon fluid spreads out and fills the gap by capillary action. Next the template is pressed against the substrate to close the gap, and the assembly is irradiated with UV light, which cures the photopolymer to make it a solidified and silicon-rich replica of the template. Significant efforts from both academia and industry have been put in S-FIL research and development,[5] template fabrication methods,[6,7] and defect analysis.[7,8] Because of its ability to pattern at room temperature and at low pressure, the template can be stepped to pattern the whole wafer area similar to that of a stepper lithography tool. These imprinting-based lithography technologies, especially S-FIL, are entering the competition of NGL methodologies for future semiconductor IC manufacturing, and are emerging as strong contenders. Recently the International Technology Roadmap for Semiconductors (ITRS) also announced the inclusion of NIL onto their roadmap as a candidate technology to begin production in 2013.

NANOIMPRINT MOLD

Material Selection and Mold Fabrication

Mold in nanoimprint technique plays the same role as photomask in photolithography. Both NIL and S-FIL utilize hard molds to replicate patterns, which distinguish themselves from the widely studied μCP technique that uses soft and flexible poly(dimethylsiloxane) (PDMS) stamp. The hard features on NIL molds and S-FIL templates can allow imprinting of features in polymer materials with resolutions below 10 nm, which is not possible to achieve by μCP using soft stamp. Hard molds have mostly been made in Si or SiO_2 for nanoimprint, while it is also possible to use metal as the mold material.

Considerations for selecting mold materials include hardness, compatibility with traditional microfabrication processing, and thermal expansion coefficient, to name an important few. Li has examined the relative hardness of various materials suitable for imprinting mold, including Si, SiO_2, SiC, silicon nitride, and sapphire. Taniguchi et al.[9,10] have investigated diamond as a potential mold material for NIL. On the other hand, works from many groups have shown that Si and SiO_2 have sufficient hardness and durability for the nanoimprint application. Thermal expansion coefficient is especially important in thermal NIL where a temperature over 100°C is typically required in the imprinting step. Thermal mismatch between the mold and the substrate could result in pattern distortions or stress build-up during the cooling cycle, which would affect the pattern fidelity. In this regard, Si mold and Si substrate make a very good pair for the NIL process.

Nanoscale protrusion features on the mold or template are typically fabricated by a combination of electron-beam lithography (EBL) and anisotropic reactive ion etching (RIE). When silica or glass wafer is used as the substrate, especially for S-FIL template, special considerations are required to eliminate or minimize the electron charging effect because of the insulator nature of the substrate. Two effective methods have been developed to fabricate S-FIL templates on standard photomask plate.[6,7] The first method uses a very thin (15 nm) layer of Cr as a hard mask, on top of which electron beam resist is coated and patterned by direct EBL writing. The thin Cr layer suppresses charging during electron-beam exposure of the template and minimizes the CD losses during the pattern transfer through the Cr by wet etching. Because the high etch selectivity of glass to Cr in a fluorine-based RIE process, a sub-20 nm Cr layer is also sufficient as a hard mask during the etching of the glass substrate. The second scheme incorporates a conductive and transparent layer of indium tin oxide (ITO) on the glass substrate, and charging is suppressed not only during EBL process of making the template, but also during final inspection stage, which greatly facilitates the template manufacturing.

Low Surface Energy Coating on Mold

One would normally imagine that the nanoscale protrusion features on the mold are a weak point in NIL.

But this is usually not the case. The reason is that in imprint lithography, mold durability is maintained by choosing the proper polymer resist layer thickness such that the mold protrusion will not contact the hard substrate directly. In other words, a thin viscous polymer fluid exists between the mold protrusions and the substrate surface and acts as a ''soft cushion,'' which effectively protects the nanofeatures on the mold. The trade-off of this is the existence of a residual resist layer that is present in the recessed regions, which has to be removed by a separate plasma etching step before pattern definition can be completed.

Mold used for imprint lithography typically has high density of nanoscale protrusion features on its surface, which effectively increases the total surface area that contacts the imprinted polymer, and therefore leading to a strong tendency of adhesion of imprinted polymer to the mold. This effect can be easily seen by the sticking of resist material on a mold without any special treatment. The solution to this problem is to apply a low surface tension coating to the mold to reduce its surface energy. This can be done by plasma deposition of fluoropolymers or by using a monolayer of perfluorosilane surfactant molecules. The latter technique is the most widely adopted method. Surfactant molecules, such as 1H,1H,2H,2H-perfluorodecyl-trichlorosilane (FDTS), can be deposited on mold surface either by liquid phase or by vapor phase through a silanization process (Fig. 2). Silanization of oxidized silicon with an $RSiCl_3$ precursor begins with the hydrolysis of the polar headgroup, which converts the Si–Cl bonds to Si–OH (silanol) groups. The generated silanol groups, which are strongly attracted to the hydrophilic surface of oxided silicon, condense and react with the hydroxyl group on the surface as well as with other monomer silanol groups to form networks of covalent siloxane bonds, Si–O–Si. Such covalent bonding makes the surfactant coating layer chemically and thermally stable. Vapor phase coating has shown to provide better imprint results for nanoscale features, because it avoids the difficulty of liquid wetting of the nanoscale trench features on the mold.[11]

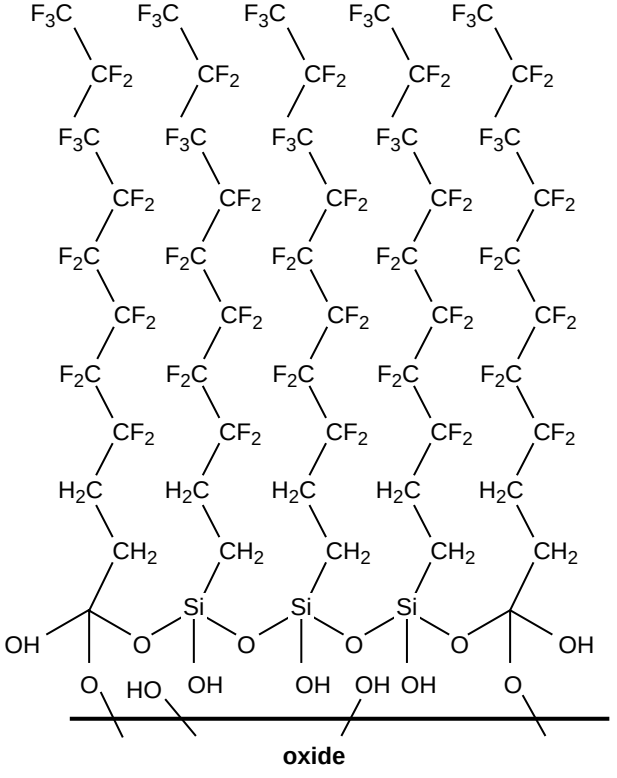


Fig. 2 Formation of a monolayer of FDTS molecules on SiO_2 to create a low-energy surface.

A more important issue in terms of mold durability is actually related to the wearing and stability of this surfactant coating layer. This could be a critical issue for the acceptance of imprint-based technology in industrial applications. A good coating could allow hundreds to thousands of imprints before it loses its effectiveness. We should point out that the durability issue of the surface coating on molds could be alleviated if the mold itself is made of a material that has low surface energy. For example, it has been demonstrated that amorphous fluoropolymers, such as Teflon AF ($T_g = 240°C$), can be used as an imprinting mold without any surface treatment. Mold fabrication can also be simplified by casting of the fluoropolymer solutions over a prefabricated template and drying off the solvent[12] or by direct molding or imprinting of this fluoropolymer at 350°C under a high pressure. A flexible film mold can be obtained using these methods, which can provide better conformal contact with the substrate to be patterned and reduce the pressure needed during the imprinting step.

NANOIMPRINT RESIST

Requirement for Nanoimprint

Because imprint lithography makes a conformal replica of the surface relief patterns by mechanical embossing, the resist materials used in imprinting should be deformable easily under an applied pressure. Step-and-flash imprint lithography uses a very effective solution by utilizing a liquid resist that can be cured and solidified by UV exposure. The viscosity of the initial monomer liquid is chosen to be lower than 4 cps to allow a very low-pressure printing. In NIL, typically a thermal plastic material is used as imprint resist, and a suitable imprint temperature is chosen to be 70–80°C above the material's glass transition temperature (T_g). This choice can be explained by considering the typical deformation behavior of thermal plastic polymer as a function of the temperature (Fig. 3). At a temperature

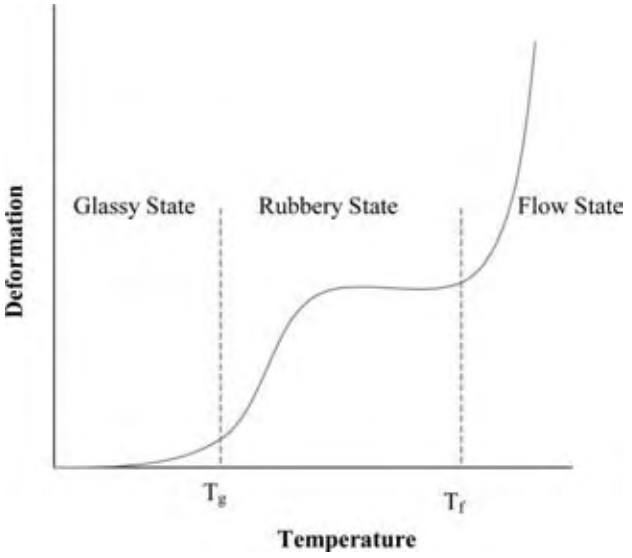


Fig. 3 Typical deformation behavior of thermal plastic polymers as a function of temperature. *(View this art in color at www.dekker.com.)*

below T_g, the major contribution to the deformation comes from the elongation of the atomic distance, and the deformation is ideal elastic. The Young's modulus for glassy polymers just below T_g is approximately constant over a wide range of polymers (3×10^9 Pa) and the magnitude of deformation is very small. Above T_g, the local motions of chain segments take place, and the modulus of the material drops by several orders of magnitude. However, the entire chains are still fixed by the temporary network of entanglements. A rubber–elastic plateau region exists beyond T_g, where relative large deformation may occur because of extension of chain segments fixed between entanglement points. The modulus stays relatively the same in the rubbery state and the deformation will recover after the release of force. Next is the rubbery flow region for linear amorphous polymers; but it does not occur for crosslinked polymers. Finally, with further increase in temperature, viscous liquid flow state is reached. In this regime the motion of entire chains takes place and the polymer flows by chain sliding. The modulus and viscosity are further reduced in this region and the deformation is irreversible, which makes it the right temperature range for NIL patterning. It is perceived that a good imprinting result can be acquired when imprinting temperature is set to be higher than the flow temperature (T_f) of the polymer. Empirically the optimal imprinting temperature is found 70–80°C above the T_g of the material used. In this temperature range, the viscosity of the polymer can be described by the following equation by Williams, Landel, and Ferry:

$$\log \eta(T) = \log \eta(T_g) - \frac{C_1(T - T_g)}{C_2 + (T - T_g)} \quad (1)$$

where the values for C_1 and C_2 are 17.44 and 51.6, respectively. This equation describes the effect of temperature on viscosity for a large number of polymers. Because T_g is the onset temperature for molecular motion in polymers, factors that increase the energy required for molecular motion also increase T_g; those that decrease the energy requirement lower T_g. These considerations can be exploited in choosing the desired T_g for the imprint resist. Therefore factors such as intermolecular forces, intrachain steric hindrance (e.g., branching or crosslinking) and bulky and stiff side groups can be used to increase T_g. While flexible bonds and flexible side groups can decrease T_g.

Often, it is desirable to do patterning at lower temperature for higher throughput and reducing the thermal expansion mismatch. From the above analysis, one can use materials that have a lower T_g as NIL resist. In Fig. 4A, the imprinted patterns in poly(benzyl methacrylate) ($M_w \sim 70{,}000$, $T_g = 54$°C) at temperature of 134°C are shown; and in Fig. 4B, the patterns imprinted in poly(cyclohexyl acrylate) ($M_w \sim 150{,}000$; $T_g = 19$°C) at 99°C are shown. In both cases, the imprinting temperatures were chosen to be $T_g + 80$°C °C and the applied pressure was 50 kg/cm^2, and very good pattern definition was obtained.

For typical thermal plastic, not only the viscosity has temperature dependence, but it also strongly depends on the polymer's molecular weight M_w, relative to the so-called critical molecular weight of a given polymer M_c. M_c can be interpreted as the molecular weight at which a temporary network of entanglements

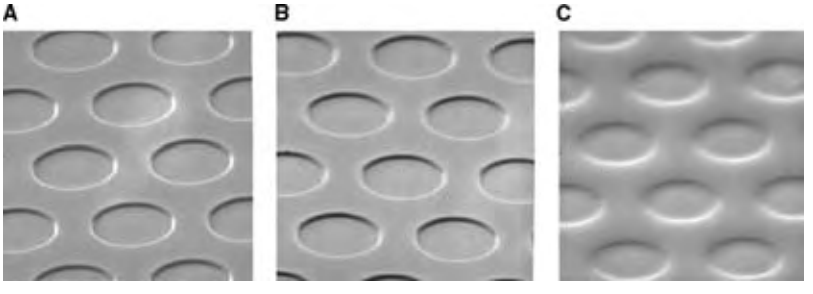


Fig. 4 SEM micrograph of patterns imprinted in (A) poly(benzyl methacrylate); (B) poly(cyclohexyl acrylate); and (C) relaxation of poly(cyclohexyl acrylate) patterns 10 days after the imprinting. *(View this art in color at www.dekker.com.)*

spans over macroscopic dimensions. Below M_c entanglements may be present, but their number is too small to lead to a sufficiently connected temporary network. Accordingly, the polymer melt resembles the flow properties of a low molecular weight compound, namely, the viscosity increases linearly with the molecular weight. Above M_c, the segment length between entanglements appears to be constant according to the experimental observation, and the viscosity shows a stronger dependence on M, as the number of entanglements per chain has to increase with M for a constant segment length.[3] If the Newtonian viscosity at critical molecular weight is denoted by η_c, the viscosity of a linear polymer in the flow state can be described by the following equations:

$$M_w < M_c, \text{ unentangled molecules}$$
$$M_w > M_c, \text{ entangled molecules}$$

In practice, low molecular weight polymers with $M < M_c$ can be imprinted at lower temperatures, lower pressures, or within shorter times. However, the absence of a network of entanglements may lead to more brittle behavior and could result in the fracture of the imprinted polymer features during the mold separation step. Therefore, the choice of both T_g and M_w is important in pattern structural stability. On the other hand, this property can also be exploited in deciding the structures to be obtained in using the reverse imprinting technique to pattern over topographies (see the section titled ''Nanoimprint Mold''). In addition, the stress build-up during the thermal cycling of the NIL process also affects the pattern integrity during mold separation, and Hirai et al. have investigated this problem in detail.[13] For more discussions on the material rheology issues related to nanoimprint, the readers are referred to chapters 3 and 4 in Ref.[3].

Another important criterion for a good NIL resist is that the imprinted pattern should maintain its mechanical integrity during mold–substrate separation as well any subsequent pattern transfer steps. Although low T_g material can be used for NIL for the sake of reducing processing temperature, the imprinted patterns are also unstable and tend to deform at temperatures close to the imprinting temperature. In Fig. 4C, the pattern relaxation at room temperature is shown, which was observed 10 days after the structures shown in Fig. 4B were imprinted (in this case the resist has a T_g of 19°C, less than room temperature). From this demonstration, it can be inferred that thermal curable or thermosetting polymers could be excellent resist systems for NIL because of the possibility of low-pressure imprinting and good mechanical integrity after crosslinking by thermal treatment. We recently developed a new, fast thermal curable liquid resist that can be imprinted under a low pressure with a very high precision and throughput. This system is based on the hydrosilylation chemistry of siloxane polymers shown in the following scheme. It consists of four basic chemical components: a vinyl terminated polydimethylsiloxane polymer, a silyl-hydride (Si-H) -based dimethylsiloxane crosslinker, a platinum catalyst, and an inhibitor. The high Si content in this polymer system guarantees that the resist has high etching resistance in RIE processes. (See below.)

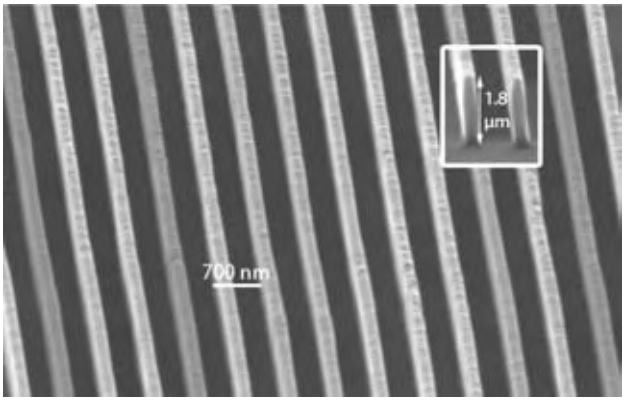
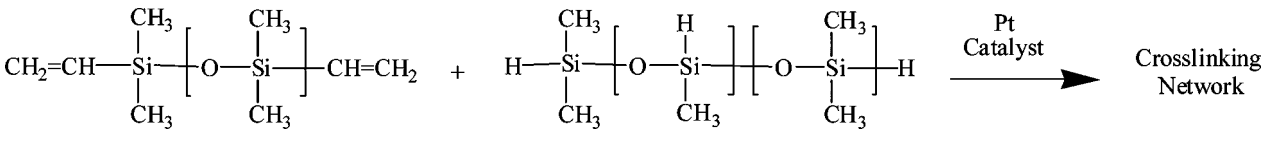


Fig. 5 SEM micrograph of an imprinted high-aspect ratio grating pattern with a 250 nm line width, and a 1.8 μm height using PDMS-based thermal curable liquid resist.

In nanoimprinting, the liquid resist is first spin-coated on a silicon wafer, followed by imprinting with pressures in the range 30–100 psi at room temperature, and crosslinking after being heated above 80°C within a minute. Because of the low surface tension of the patterned film ($\gamma = 19.8$ dyne/cm at 20°C), the mold separation is quite easy even for high aspect ratio features and the imprinted resist surface is very smooth (Fig. 5). Importantly, the mold separation does not require cooling to room temperature, thereby increasing the process throughput.



DEVELOPING BETTER RESIST MATERIALS FOR NANOIMPRINT LITHOGRAPHY

The rapid development of this technique in recent years has also stimulated the research for new materials that are better suited as nanoimprint resists. Because imprint lithography makes a conformal replica of the surface relief patterns by mechanical embossing, the resist materials used in imprinting need to be deformable easily under an applied pressure. The most commonly used materials in the original NIL scheme are thermal plastic polymers, which soften significantly when heated above their glass transition temperatures. Many commercial thermal plastic materials, e.g., poly(methyl methacrylate) (PMMA) and polystyrene (PS), have been used as NIL resists.[2,4,14] Obviously, these materials are not optimized at all for the special requirement in NIL process. One critical requirement for candidate polymers used in imprinting-based lithography is that it should provide excellent mold releasing properties during the demolding process, while at the same should not compromise its adhesion to the substrate. Commercially available polymer materials can hardly satisfy these seemingly contradictory requirements. Although the mold surface is normally treated with low surface energy surfactant, when imprinting high-density or high aspect ratio patterns, the imprinted polymer tends to adhere to the mold, creating pattern defects that are intolerable for many device applications. Block copolymer materials that have dual surface properties by microphase segregation can be exploited to solve this problem. In addition, higher dry etching resistance is highly desirable if the imprinted polymer pattern is to be used as a dry etching mask for further pattern transfer. Recently we exploited the use of siloxane copolymers as NIL resist, such as polystyrene-*b*-poly(dimethyl siloxane) (PS-PDMS) block copolymer. The presence of PDMS imparts the copolymers with many properties that are favorable for NIL, e.g., low surface energy for easy mold release and high Si content for chemical etch resistances—in particular, extremely low etch rates (comparable to SiO_2) in oxygen plasma, to which organic polymers are quite susceptible. In addition, the copolymers are thermoplastic and can be thermally deformed reversibly. An imprinted grating pattern with 250 nm line width is shown in Fig. 6.

On the other hand, the high temperature and pressure required for the nanoimprinting of thermal plastic materials limit the throughput and the application scope of the NIL technique. In addition, the thermal expansion mismatch between the mold and the substrate frequently incurred often presents an obstacle for pattern alignment over large substrates. Although various means have been attempted to make the thermoplastics imprintable at a temperature close to room temperature, including dissolving the polymer into its monomer or solvent[15] using PDMS stamp for solvent evaporation,[14] the success has been very limited.

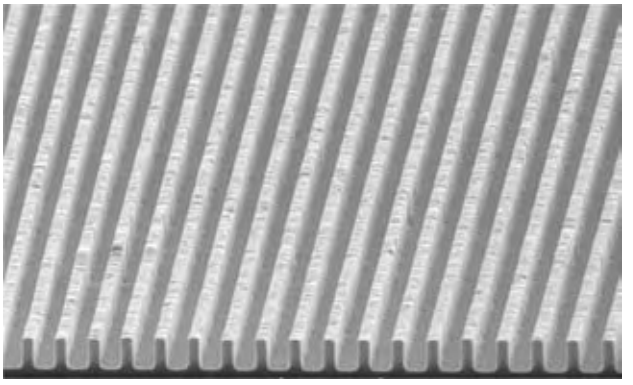


Fig. 6 SEM micrograph of 250 nm line width patterns imprinted in PS-PDMS diblock copolymer.

A very attractive method for achieving room-temperature and low-pressure imprinting is to utilize a liquid precursor that can be cured and solidified by heating or UV light. Comparably, UV-curable material system is preferred to thermal curing system because the heating and cooling cycle of the latter can slow down the process throughput. The most widely used UV-curing formulations are based on free radical polymerization of acrylic and methacrylic monomers because of their high reactivity, as in S-FIL.[4] However, such material systems suffer from oxygen sensitivity issue: oxygen scavenges free radical species and thus inhibits polymerization process at the resist surface, making the process prone to defect generation. Besides, the acrylate-based UV-imprint resist has a large shrinkage upon curing ($\sim$10%), which not only affects the faithful pattern definition but also tends to cause delamination of the resist film from certain substrates, especially metals and plastics.

As a result, a number of critical issues have to be resolved before this potentially powerful technique can be exploited for any industrial-scale nanopatterning. First, it is preferable that the imprinting resist precursor is in liquid state and can be spin-coated on the substrate to allow low-pressure and high-throughput processing. Second, the shrinkage of the resist after crosslinking must be very small to ensure the fidelity of pattern transfer and to prevent possible film delamination. Third, a resist with a good dry etching resistance is highly desirable as it allows the resist to be used directly as a hard mask in RIE process to transfer the patterns into the underlying substrate (Fig. 7). We developed a new UV-curable epoxysilicone material based on cationic crosslinking of cycloaliphatic epoxies. This resist combines a number of aforementioned desired features for nanoimprinting.[16] Because cationic polymerization is not prone to oxygen

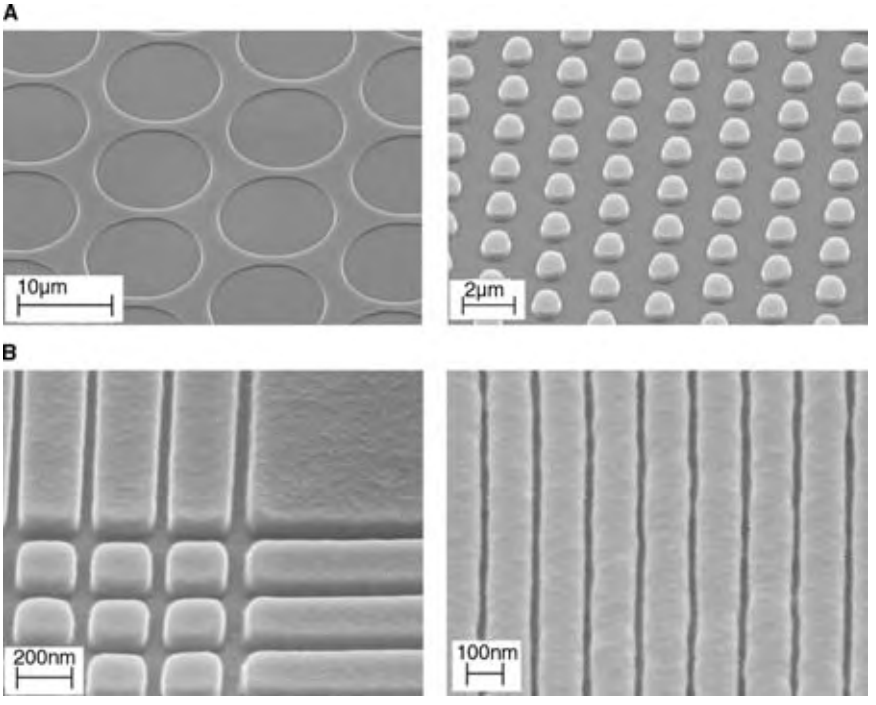


Fig. 7 SEM micrographs of imprinted and UV-cured resist patterns: (A) 20 μm diameter recessed patterns (left) and 1 μm diameter protruded patterns (right) and (B) sub-100 nm trench patterns (left) and 20 nm trenches (right).

inhibition as compared to the free radical polymerization of the acrylate monomers, fewer defects are expected. The resist exhibits very good dry etching resistance because of high silicon content. Furthermore, its very low shrinkage after curing, only a fraction of that of the acrylate system, allows a reliable patterning. In addition, with a suitable undercoating polymer, a very uniform liquid precursor can be formed simply by spin-coating, which also allows other processes such as lift-off to be easily performed.

The novel liquid nanoimprint resist consists of a silicone-diepoxy monomer, a silicone crosslinking agent, and a photoacid generator. The epoxysilicone materials were prepared via hydrosilylation of 4-vinyl-1-cyclohexane 1,2-epoxide with the corresponding SiH-functional silicone intermediates in the presence of a platinum catalyst and purified by vacuum distillation. A typical liquid resist formulation comprises 94 wt.% diepoxy monomer, 5 wt.% crosslinkers, and 1 wt.% PAG. Organic solvents, such as propylene glycol monomethyl ether acetate (PGMEA), can be used to adjust the viscosity of the resist so that film thickness from 1 μm down to sub-50 nm can be readily obtained. With a suitable under-coating layer, such as baked PMMA or SU-8 resist, stable and uniform liquid thin films can be formed on Si or other substrate.

With such a UV-curable nanoimprint resist, both micro- and nanoscale patterns can be easily achieved at room temperature and at a pressure of less than 0.1 MPa by using a conventional contact exposure tool. In Fig. 7A, recessed hole pattern of 20 μm in diameter and protruded pillar pattern of 1 μm in diameter after imprinting and curing is shown, and in Fig. 7B, various nanoscale patterns are shown. These results demonstrate that patterns as small as 20 nm (limited by the feature resolution on the current mold) can be easily achieved by conventional contact aligner using this new resist material.

OTHER IMPRINT-BASED PATTERNING TECHNIQUES

Combined Nanoimprint and Photolithography

Although NIL has proved to be very successful in nanopatterning, especially in replicating periodic nanoscale features with uniform sizes, it still has several limitations as a flexible lithographic technique. A general lithographic technique should be capable of producing both large and small features in various combinations and distributions—a typical requirement in micro- and nanofabrication processes. In NIL, protrusion features on the mold physically deform and displace the polymer. Larger features on the mold must displace more polymer material over larger distances. Thus, patterns with large features are much more difficult to imprint than nanopatterns. Moreover, NIL pattern defects or even failures in the form of incomplete pattern transfer can occur because of the high viscosity of the polymer melt and the mold pattern complexity.

The aforementioned defect generations are related to the mold pattern itself and are inherent in NIL process. To solve those problems, we developed a technique

that integrates photolithography into NIL to take advantages of both techniques while compensating each other's limitations.[17] The schematic of such a combined-nanoimprint-and-photolithography (CNP) technique is shown in Fig. 8. A hybrid mask concept is introduced in this process (Fig. 8A), which is made of UV transparent material, and acts both as a NIL mold and as a photolithography mask. Protrusions are made on the mold for imprinting nanoscale features, while metal pads are embedded into the mold serving as light mask for photolithography to replicate large patterns. Fabrication details of such hybrid mold can be found in Ref.[17]. The resist used in this method can be either negative tone photoresists or UV-curable materials.

The processing steps of CNP are very simple. The hybrid mold is first imprinted into the resist layer by pressure, then the whole mold–substrate assembly is exposed by UV radiation (Fig. 8B). After the hybrid mold and the substrate are separated (Fig. 8C), the substrate is immersed in a developer solution to remove the unexposed resist that are blocked by the metal pads (Fig. 8D). After developing, both large-scale and nanoscale patterns are created in the polymer resist in one step. The effectiveness of this new technique was demonstrated by using a negative tone photoresist SU-8.

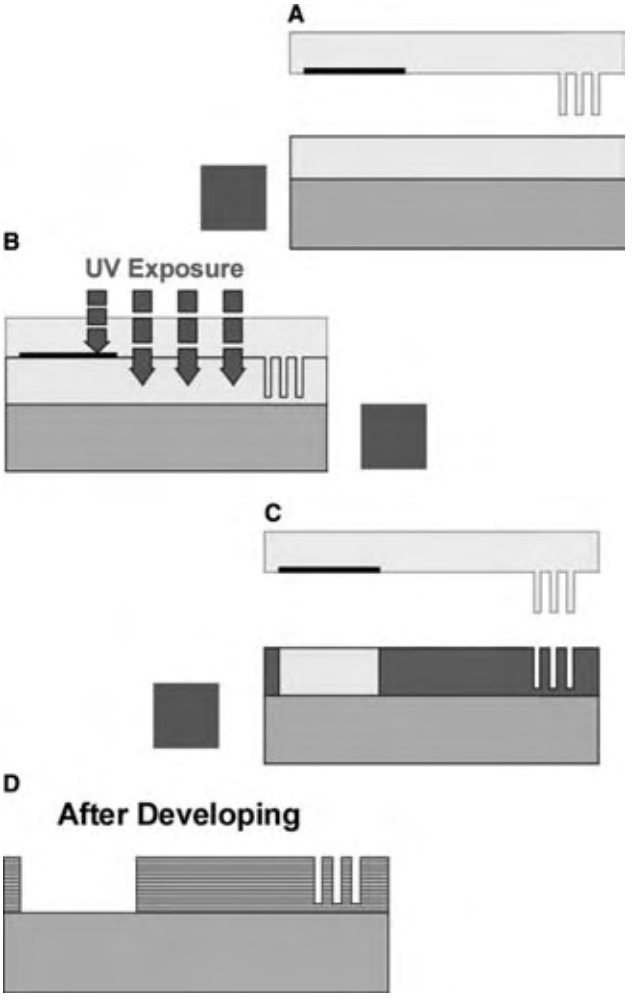


Fig. 8 Schematic of the CNP technique using a hybrid mold. *(View this art in color at www.dekker.com.)*

There are many advantages of the CNP method by using the new hybrid mask concept. First, it enables one-step lithography of arbitrary patterns containing both large-scale and nanoscale structures, which is often required in functional device fabrications. Second, because there are only nanoscale mold protrusion features on the hybrid mold, it allows low imprinting pressure to be used as only very small amount of polymer needs to be displaced. Third, by forming the large patterns as a photomask (i.e., making them as metal pads), it reduces the complexity of the relief pattern on the hybrid mask-mold. This simplifies residual layer thickness distribution, which can ease the step for residue removal significantly. Finally, the throughput could also improve because, in NIL, imprinting nanoscale features can be a very fast process, while establishing a quasi-equilibrium distribution of residual polymers when imprinting large patterns could take significant time. The last point can be quantified using a simple model by considering the squeezed flow of a Newtonian fluid between plates that have a radius R and a gap distance of d. Such a structure can approximate the process where a mold protrusion with size R is imprinted into a resist of thickness d on a substrate. One can obtain a simple solution that expresses the pressure (P) in terms of these dimensions and to the speed of imprinting ($\mathrm{d}h/\mathrm{d}t$) and fluid viscosity h. Numerical simulation based on finite difference method using nonsteady-state Navier–Stokes equation has also given such relationship for simple periodic mold features.[18] Integrating this equation can give the time required to reduce the thickness of the fluidic layer by half (i.e., to imprint half way through the feature height). This simple model says that under the sample pressure the imprinting time scales as quadratic function of the pattern size. Therefore, the time required for imprinting large size pattern will be significantly longer than that for nanopatterns.

Reverse-Nanoimprint Technique

This technique was developed to address the issue of patterning on topographies and on flexible substrate (Fig. 9). Nanoimprint lithography has a great advantage in that it can directly imprint device structures in polymer materials[19] (sometimes in single step) and create patterned features using polymers with certain desirable physical or chemical functionalities.[20–22] But NIL patterning is typically carried out at a temperature 70–80°C above the glass transition

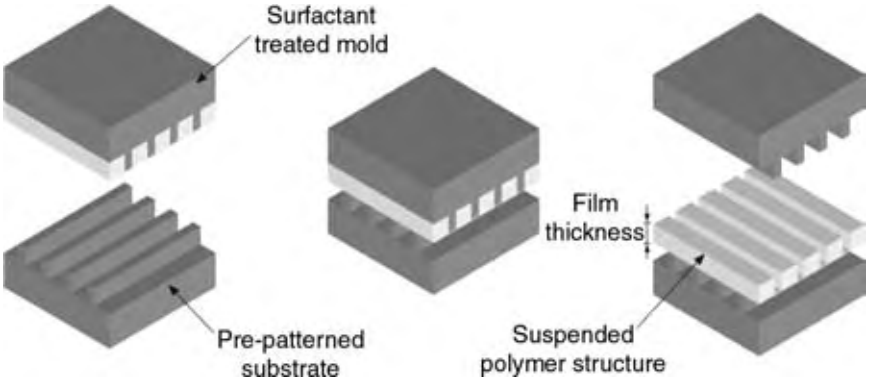


Fig. 9 Schematics of reverse-imprinting polymer nanostructures over topography. *(View this art in color at www.dekker.com.)*

temperature (T_g) of the polymer material to ensure that the polymer has a sufficiently reduced viscosity that can be imprinted at reasonable pressure. For micro- and nanoscale patterning, especially when directly patterning functional polymers, it is often desirable to operate at lower temperature and with reduced pressure. This is required not only in certain applications or limited by certain substrate characteristics, but also if the mold and substrate are made from different materials with mismatch in thermal expansion coefficients, higher temperature will create more stress during the NIL thermal cycle, and could present problem for mask–sample alignment. To target these problems, room temperature NIL has been previously demonstrated by using spin-on-glass or hydrogen silsequioxane (HSQ) as imprint resist material,[23,24] but the pressure needed was very high because of the very high modulus and viscosity of such inorganic material.

It was developed based on the following consideration. If a polymer film is spin-coated onto a mold, the polymer will fill up the trench regions of the surface relief patterns. This means that a replica of the mold pattern is formed in the polymer film simply by spin-coating. Now if this film can be transferred from the mold to a substrate, patterned structures are obtained (Fig. 10A). In Fig. 10B, the imprinted 350 nm line-spacing PMMA grating obtained by reverse imprint process that was done at a temperature of 105°C, i.e., at the T_g of PMMA, is shown. The key to the successful film transfer lies in the fact that the mold has a lower surface energy than that of the substrate, so the polymer film has better adhesion to the substrate and therefore can be detached from the mold.

Because in reverse imprinting coating of polymer resist on the substrate is not required, it is therefore possible to use this technique to transfer patterns onto substrates that are not suitable for spin-coating or have surface topographies. We have demonstrated that this technique can be applied to flexible substrates and substrates that already have other fabricated features (e.g., nonflat surface or prepatterned surface) to create nanoscale patterns (Figs. 11A and 11B).[25] The ability to pattern over topography has solved a long-standing problem in imprint-based lithography. Previous efforts to solve this problem often involve multilayer resist approaches with a thick polymer planarization layer on top of the nonflat substrate.[26] Those approaches not only require complex processes with multiple steps, but also entail deep etching steps to etch through the thick planarization layer, which often degrades the resolution and fidelity of the pattern. But reverse-imprint solved this problem very effectively. In Fig. 12A, the reverse-imprinted polycarbonate grating

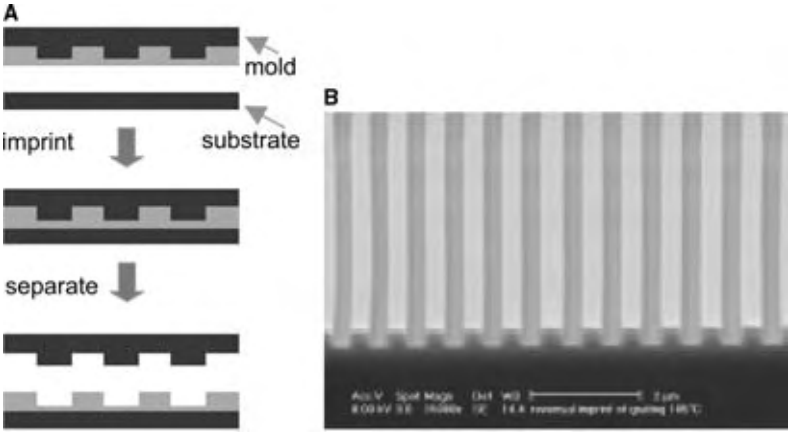


Fig. 10 (A) Schematic of reverse nanoimprint, and (B) SEM of reverse imprinted PMMA gratings with 350 nm line width/spacing. *(View this art in color at www.dekker.com.)*

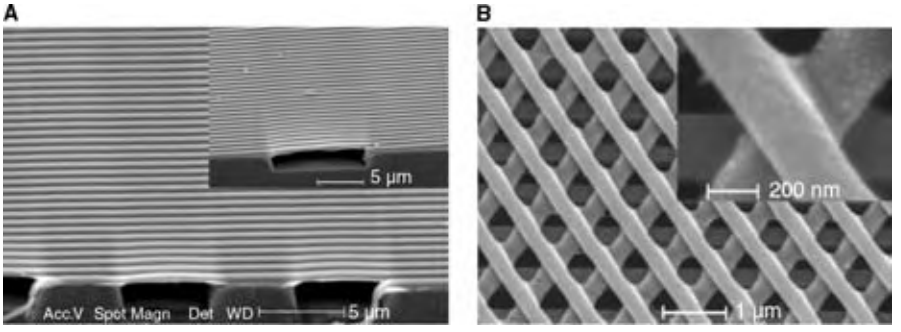


Fig. 11 Imprinting of 700 nm period grating on topographies: (A) PC pattern imprinted across 5 μm gaps and 10 μm gaps as shown in the inset. (B) PMMA pattern imprinted on the protruded surfaces for line spacing of greater than 3 μm.[25]

structure is shown that is suspended over etched features on Si substrate.

The reverse-imprinting technique can potentially offer a simple method to fabricate three-dimensional (3D) polymer nanostructures. Such 3D structure can be achieved by simply repeating the process and building up the structure in a layer-by-layer fashion. An example of imprinted three-layer nanostructure, using three different polymers with progressively lower T_g, is shown in Fig. 12B.[24] Note that this approach can also allow a single polymer material to be used for all layers if the material can be thermal or UV cured.

There are many potential applications of such 3D polymer nanostructures. Multilayered structure with varied grating periods can be used as size-controlled filters in microfluidics to select and separate particles of different sizes. Another potential application is to fabricate periodic 3D polymer structures and infiltrate the polymer template with high refractive index inorganic materials to create 3D photonic crystal materials (Fig. 12). Not only the dimensions of such structure can be designed and controlled but also it is straightforward to form defect waveguide and resonator structures that are useful for many photonic device applications, by simply removing rows on the mold. This example only serves as an illustration of a potential application of 3D printing method—we should not undermine the challenge of stringent pattern alignment required in this particular application.

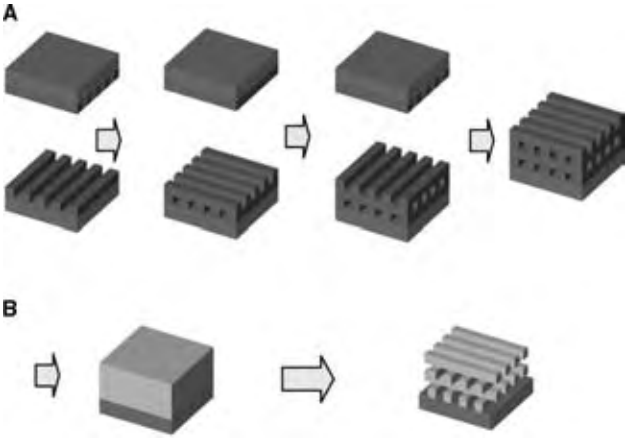


Fig. 12 Schematic of (A) building 3D polymer nanostructures by using reverse-imprint; thermal plastic or photosensitive material is spin-coated on the mold for pattern transfer and (B) infiltrate the 3D periodic structure with other materials, such as inorganic materials that have high refractive index; then remove the polymer template layer to create a 3D pattern of the infiltrated material that is complementary to the original polymer resist pattern. *(View this art in color at www.dekker.com.)*

CONCLUSIONS

Nanoimprint technique has, for the first time, enabled parallel nanoscale processing capability with simple equipment setup. The simplicity of this method has made it appealing to researchers in various fields. New techniques based on the concept of mechanical printing or embossing are appearing very rapidly. New variations of the technique aimed at different applications have also been developed at a fast pace. More and more researchers are attracted by these nonconventional patterning technologies because of their scalability, low-cost and potential new applications. It is certain that advances in new materials for imprinting will fuel the development in this field.

The simplicity and high resolution provided by this technique have also found numerous applications in electronics such as hybrid plastic electronics,[27] organic TFTs and electronics,[28,29] and nanoelectronic devices in Si;[30,31] in photonics such as organic lasers, high-resolution OLED pixels,[32] diffractive optical elements,[33] waveguide polarizers,[34] and active[20] and nonlinear optical polymer nanostructures;[21] in magnetic devices like high-density quantized magnetic disks[35] and patterned magnetic media;[36] as well as in biological applications such as manipulating DNA in nanofluidic channels[37,38] and nanoscale protein patterning.[39] Because of its fast development in the past decade and its potential for sub-100 nm lithography, MIT's Technology Review listed NIL as one of 10 emerging technologies that are likely to change the world.[40]

ACKNOWLEDGMENTS

The author would like to thank all the graduate students, postdoctoral fellows, and other colleagues at the University of Michigan who have contributed to the work described here: Dr. Xing Cheng, Dr. Li-Rong Bao, Dr. Xu-Dong Huang, Dr. Tan Li, Philip Choi, Wayne Fung, Professor Stella W. Pang, and Professor Albert F. Yee; and special thanks to Dr. Peng-Fei Fu at Dow Corning Corporation for his collaboration and contribution to some of the work described in this entry.

REFERENCES

1. Chou, S.Y.; Krauss, P.R.; Renstrom, P.J. Imprint of sub-25 nm vias and trenches in polymers. Appl. Phys. Lett. **1995**, *67* (21), 3114–3116.
2. Chou, S.Y.; Krauss, P.R.; Renstrom, P.J. Imprint lithography with 25-nanometer resolution. Science **1996**, *272* (5258), 85–87.
3. Sotomayor Torres, C.M., Ed. *Alternative Lithography*; Kluwer: Boston, 2003; Chapter 3.
4. Chou, S.Y.; Krauss, P.R. Imprint lithography with sub-10 nm feature size and high throughput. Microelectr. Eng. **1997**, *35* (1–4), 237–240.
5. Ruchhoeft, P.; Colburn, M.; Choi, B.; Nounu, H.; Johnson, S.; Bailey, T.; Damle, S.; Stewart, M.; Ekerdt, J.; Sreenivasan, S.V.; Wolfe, J.C.; Willson, C.G. Patterning curved surfaces: template generation by ion beam proximity lithography and relief transfer by step and flash imprint lithography. J. Vacuum Sci. Technol. B. **1999**, *17* (6), 2965–2969.
6. Bailey, T.C.; Resnick, D.J.; Mancini, D.; Nordquist, K.J.; Dauksher, W.J.; Ainley, E.; Talin, A.; Gehoski, K.; Baker, J.H.; Choi, B.J.; Johnson, S.; Colburn, M.; Meissl, M.; Sreenivasan, S.V.; Ekerdt, J.G.; Willson, C.G. Template fabrication schemes for step and flash imprint lithography. Microelectr. Eng. **2002**, *61–62*, 461–467.
7. Dauksher, W.J.; Nordquist, K.J.; Mancini, D.P.; Resnick, D.J.; Baker, J.H.; Hooper, A.E.; Talin, A.A.; Bailey, T.C.; Lemonds, A.M.; Sreenivasan, S.V.; Ekerdt, J.G.; Willson, C.G. Characterization of and imprint results using indium tin oxide-based step and flash imprint lithography templates. J. Vacuum Sci. Technol. B. **2002**, *20* (6), 2857–2861.
8. Bailey, T.; Choi, B.J.; Colburn, M.; Meissl, M.; Shaya, S.; Ekerdt, J.G.; Sreenivasan, S.V.; Willson, C.G.; Step and flash imprint lithography: template surface treatment and defect analysis. J. Vacuum Sci. Technol. B. **2000**, *18* (6), 3572–3577.
9. Taniguchi, J.; Tokano, Y.; Miyamoto, I.; Komuro, M.; Hiroshima, H.; Kobayashi, K.; Miyazaki, T.; Ohyi, H. Preparation of diamond mold using electron beam lithography for application to nanoimprint lithography. Jpn. J. Appl. Phys. 1 (Regular Papers Short Notes Review Papers) **2000**, *39* (12B), 7070–7074.
10. Taniguchi, J.; Tokano, Y.; Miyamoto, I.; Komuro, M.; Hiroshima, H. Diamond nanoimprint lithography. Nanotechnology **2002**, *13* (5), 592–596.
11. Beck, M.; Graczyk, M.; Maximov, I.; Sarwe, E.L.; Ling, T.G.I.; Keil, M.; Montelius, L. Improving stamps for 10 nm level wafer scale nanoimprint lithography. Microelectr. Eng. **2002**, *61–62*, 441–448.
12. Khang, D.Y.; Lee, H.H. Sub-100 nm patterning with an amorphous fluoropolymer mold. Langmuir **2004**, *20* (6), 2445–2448.
13. Hirai, Y.; Fujiwara, M.; Okuno, T.; Tanaka, Y.; Endo, M.; Irie, S.; Nakagawa, K.; Sasago, M. Study of the resist deformation in nanoimprint lithography. J. Vacuum Sci. Technol. B **2001**, *19* (6), 2811–2815.
14. Khang, D.Y.; Yoon, H.; Lee, H.H. Room-temperature imprint lithography. Adv. Mater. **2001**, *13* (10), 749–752.
15. Jung, G.Y.; Ganapathiappan, S.; Li, X.; Ohlberg, D.A.A.; Olynick, D.L.; Chen, Y.; Tong, W.M.; Williams, R.S. Fabrication of molecular-electronic circuits by nanoimprint lithography at low temperatures and pressures. Appl. Phys. A: Mater. Sci. Process. **2004**, *78* (8), 1169–1173.
16. Cheng, X.; Guo, L.J.; Fu, P.-F. Room temperature and low pressure nanoimprinting based on cationic photopolymerization of novel epoxysilicone monomers. Adv. Mater. **2005**, *17*, 1419–1424.
17. Cheng, X.; Guo, L.J. One-step lithography for various size patterns with a hybrid mask-mold. Microelectr. Eng. **2004**, *71* (3–4), 288–293.
18. Wu, L.; Chou, S.Y. 47th EIPBN Technical Digest. 2003.
19. Chao, C.Y.; Guo, L.J. Polymer microring resonators fabricated by nanoimprint technique. J. Vacuum Sci. Technol. B. **2002**, *20* (6), 2862–2866.
20. Wang, J.; Sun, X.Y.; Chen, L.; Chou, S.Y. Direct nanoimprint of submicron organic light-emitting structures. Appl. Phys. Lett. **1999**, *75* (18), 2767–2769.
21. Guo, L.J.; Cheng, X.; Chao, C.Y. Fabrication of photonic nanostructures in nonlinear optical polymers. J. Modern Optics **2002**, *49* (3–4), 663–673.

22. Pisignano, D.; Persano, L.; Visconti, P.; Cingolani, R.; Gigli, G.; Barbarella, G.; Favaretto, L. Oligomer-based organic distributed feedback lasers by room-temperature nanoimprint lithography. Appl. Phys. Lett. **2003**, *83* (13), 2545–2547.
23. Matsui, S.; Igaku, Y.; Ishigaki, H.; Fujita, J.; Ishida, M.; Ochiai, Y.; Komuro, M.; Hiroshima, H. Room temperature replication in spin on glass by nanoimprint technology. J. Vacuum Sci. Technol. B. **2001**, *19* (6), 2801–2805.
24. Matsui, S.; Igaku, Y.; Ishigaki, H.; Fujita, J.; Ishida, M.; Ochiai, Y.; Namatsu, H.; Komuro, M. Room-temperature nanoimprint and nanotransfer printing using hydrogen silsequioxane. J. Vacuum Sci. Technol. B. **2003**, *21* (2), 688–692.
25. Bao, L.R.; Tan, L.; Huang, X.D.; Kong, Y.P.; Guo, L.J.; Pang, S.W.; Yee, A.F. Polymer inking as a micro- and nanopatterning technique. J. Vacuum Sci. Technol. B. **2003**, *21* (6), 2749–2754.
26. Sun, X.Y.; Zhuang, L.; Zhang, W.; Chou, S.Y. Multilayer resist methods for nanoimprint lithography on nonflat surfaces. J. Vacuum Sci. Technol. B. **1998**, *16* (6), 3922–3925.
27. McAlpine, M.C.; Friedman, R.S.; Lieber, D.M. Nanoimprint lithography for hybrid plastic electronics. Nano Lett. **2003**, *3* (4), 443–445.
28. Austin, M.D.; Chou, S.Y. Fabrication of 70 nm channel length polymer organic thin-film transistors using nanoimprint lithography. Appl. Phys. Lett. **2002**, *81* (23), 4431–4433.
29. Cedeno, C.C.; Seekamp, J.; Kam, A.P.; Hoffmann, T.; Zankovych, S.; Torres, C.M.S.; Menozzi, C.; Cavallini, M.; Murgia, M.; Ruani, G.; Biscarini, F.; Behl, M.; Zentel, R.; Ahopelto, J. Nanoimprint lithography for organic electronics. Microelectr. Eng. **2002**, *61–62*, 25–31.
30. Guo, L.J.; Krauss, P.R.; Chou, S.Y. Nanoscale silicon field effect transistors fabricated using imprint lithography. Appl. Phys. Lett. **1997**, *71* (13), 1881–1883.
31. Zhang, W.; Chou, S.Y. Fabrication of 60-nm transistors on 4-in. wafer using nanoimprint at all lithography levels. Appl. Phys. Lett. **2003**, *83* (8), 1632–1634.
32. Cheng, X.; Hong, Y.T.; Kanicki, J.; Guo, L.J. High-resolution organic polymer light-emitting pixels fabricated by imprinting technique. J. Vacuum Sci. Technol. B. **2002**, *20* (6), 2877–2880.
33. Hirai, Y.; Harada, S.; Kikuta, H.; Tanaka, Y.; Okano, M.; Isaka, S.; Kobayasi, M. Imprint lithography for curved cross-sectional structure using replicated Ni mold. J. Vacuum Sci. Technol. B. **2002**, *20* (6), 2867–2871.
34. Wang, J.; Schablitsky, S.; Yu, Z.N.; Wu, W.; Chou, S.Y. Fabrication of a new broadband waveguide polarizer with a double-layer 190 nm period metal-gratings using nanoimprint lithography. J. Vacuum Sci. Technol. B. **1999**, *17* (6), 2957–2960.
35. Wu, W.; Cui, B.; Sun, X.Y.; Zhang, W.; Zhuang, L.; Kong, L.S.; Chou, S.Y. Large area high density quantized magnetic disks fabricated using nanoimprint lithography. J. Vacuum Sci. Technol. B. **1998**, *16* (6), 3825–3829.
36. McClelland, G.M.; Hart, M.W.; Rettner, C.T.; Best, M.E.; Carter, K.R.; Terris, B.D. Nanoscale patterning of magnetic islands by imprint lithography using a flexible mold. Appl. Phys. Lett. **2002**, *81* (8), 1483–1485.
37. Cao, H.; Yu, Z.N.; Wang, J.; Tegenfeldt, J.O.; Austin, R.H.; Chen, E.; Wu, W.; Chou, S.Y. Fabrication of 10 nm enclosed nanofluidic channels. Appl. Phys. Lett. **2002**, *81* (1), 174–176.
38. Guo, L.J.; Cheng, X.; Chou, C.F. Fabrication of size-controllable nanofluidic channels by nanoimprinting and its application for DNA stretching. Nano Lett. **2004**, *4* (1), 69–73.
39. Hoff, J.D.; Cheng, L.J.; Meyhofer, E.; Guo, L.J.; Hunt, A.J. Nanoscale protein patterning by imprint lithography. Nano Lett. **2004**, *4* (5), 953–957.
40. Special Report: 10 emerging technologies that will change the world. MIT Technol. Rev. **2003**, *106*, 36.

Nanomaterials

David S. J. Arney
Jimmie R. Baran
Allen R. Siedle
Matthew H. Frey
3M Company, St. Paul, Minnesota, U.S.A.

INTRODUCTION

Industrial interest in nanomaterials derives from the novel properties they exhibit. These are defined for this entry as materials having engineered discrete particulate domains with diameters in the range of 1 nm to a few hundred nanometers. These domains may appear in many forms, such as dispersions of nanoparticles in a liquid, on surfaces, or embedded in a continuous matrix. The unique properties of nanomaterials are a consequence of the small size and extremely large interfacial areas. In this regime, dramatic variations in the chemical and physical properties of a material may be effected. Representative examples of size-critical properties, enabling new industrial applications, reviewed in this entry include surface and interfacial, catalytic, optical, and mechanical.

NANOPARTICLES IN SURFACE AND INTERFACIAL PHENOMENA

Small particles have been known for years to stabilize dispersions, foams, and emulsions.[1] Two prominent theories have been put forth to explain this phenomenon. One is based on the ''surface activity'' of the particles, while the other is based on viscosity control of the continuous phase.

The first theory claims that the particles are ''surface active but, unlike surfactants, are not amphiphilic.''[2] Partially modified particles preferentially partition the interface between the two phases, with the key physical characteristic for stabilization being the contact angle of the individual particle with the interface. Particles having a 90° contact angle (approximately equal amounts of hydrophobic and hydrophilic character) with the interface will create the most stable emulsions. The particles used in these studies are nanometer sized (~10–30 nm), but this is the primary particle size. The particles are actually agglomerated and are typically several hundred nanometers in size. This distinction is important for Eq. (1), which is used to support the proposed mechanism:

$$E = \pi r^2 \gamma_{\alpha,\beta}(1 \pm \cos\theta)^2 \quad (1)$$

where E is the energy required to remove a particle from the interface between the two liquid phases, r the radius (in meters) of the particle, $\gamma_{\alpha,\beta}$ the interfacial tension between the two phases, and θ the contact angle the particle makes with the interface.

This implies that the larger the particle (up to where gravity becomes a factor), the more stable the emulsion. It is also implicit that a particle with a too high (or too low) contact angle or of too small size should not be held at the interface and would be located in the bulk. This theory concludes that particles bound to the interface will stabilize systems, while those that are not bound to the interface cannot be used for stabilization.

Proponents readily admit that there is no significant reduction in surface tension when the particles are incorporated into the fluids. No arguments are proposed as to how the particles are bound to the surface. Their presence cannot be detected by surface tension measurements.

There is little applicability of this mechanism to stabilization by small particles. For instance, using the values exemplified earlier, the energy required to remove a particle with a diameter of 200 nm (approximate actual size of the particles in the above study) and a contact angle of 150° from a water/toluene interface (interfacial tension = 0.036 N/m) is 4927 kT, while a 5 nm particle in the same system has a binding energy of 3 kT. Therefore, a 200 nm particle will be ''irreversibly'' bound to the interface, while a 5 nm particle should not be held at the interface and if stabilization occurs, it must take place by a different mechanism.

An alternative theory based on results using unmodified, nanometer-sized, individual silica particles in an aqueous system was proposed almost 15 years ago.[3] Using reflected light microinterferometry, it was determined that the nanoparticles formed ordered structures inside the thin liquid films located between the dispersed phase regimes. This mechanism relies on

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008059

purely physical effects for stabilization, not on thermodynamic/surface tension phenomena. To briefly summarize this work, the diagram in Fig. 1 will be useful.

The large circles represent a dispersed phase (bubbles, droplets, or solids), while the small ones represent surface modified nanoparticles (SMNs) dispersed in the continuous phase that can be found between dispersed phase domains. The diagram on the left illustrates that the SMNs must be dispersed in the continuous phase. If they are not well dispersed or are agglomerated, they will not work as efficiently.

In the next diagram, the thin liquid film located between the dispersed phase regimes begins to drain and the discontinuous phase regimes begin to approach each other. When this happens, the SMNs are forced into the restricted volume of the thinning film, resulting in an ordered structure. The SMNs begin acting as steric barriers to coalescence or flocculation. As this ordering takes place, a localized concentration of the SMNs occurs—the liquid drains faster than the SMNs. This results in a very concentrated dispersion in the localized environment between domains. As with any concentrated dispersion, the viscosity in this region (only) increases. With the viscosity increasing in the region between domains, drainage is slowed and the system is further stabilized.

In the third diagram, the local concentration increases and the SMNs actually create well-ordered layers between the dispersed phase regimes while the liquid continues to drain. The SMNs cannot get out of the way fast enough, which is more commonly known as ''depletion flocculation.''[4]

Depletion flocculation occurs when two particles approach each other to within a distance that is smaller than the particle size so that no other particle can fit into the space between them. The osmotic pressures between the particles and in the rest of the dispersion are not balanced, and this pressure difference pushes the two particles toward each other. This leads to a further increase in viscosity and stability.

Finally, in the last diagram, drainage cannot be completely halted and layers of SMNs will eventually drain away. This stepwise draining is attributed to the existence of an osmotic pressure gradient—SMNs leave the film and ''vacancies'' appear in their place. The higher chemical potential of the SMNs causes the SMNs to slowly diffuse to the meniscus, leaving behind vacancies. The vacancies reach an equilibrium concentration that can only be increased by the further removal of SMNs from the film. This is known as the diffusive osmotic model.[5]

These diagrams demonstrate that the SMNs should be well dispersed in the continuous phase, monodisperse, and small in size. This can be visualized if one assumes that the SMNs in the diagrams are 5 nm particles. Thus, there would be four layers of particles between the dispersed phase regions. If one replaces the 5 nm particles with 20 nm particles, then a single 20 nm particle will essentially take up the same space (linearly) as four 5 nm particles. Replacing the 5 nm particles with 20 nm particles makes the path to coalesce or flocculation less tortuous. Also, at equal weights of particles, there are fewer 20 nm particles than 5 nm particles. It is well known that the viscosity

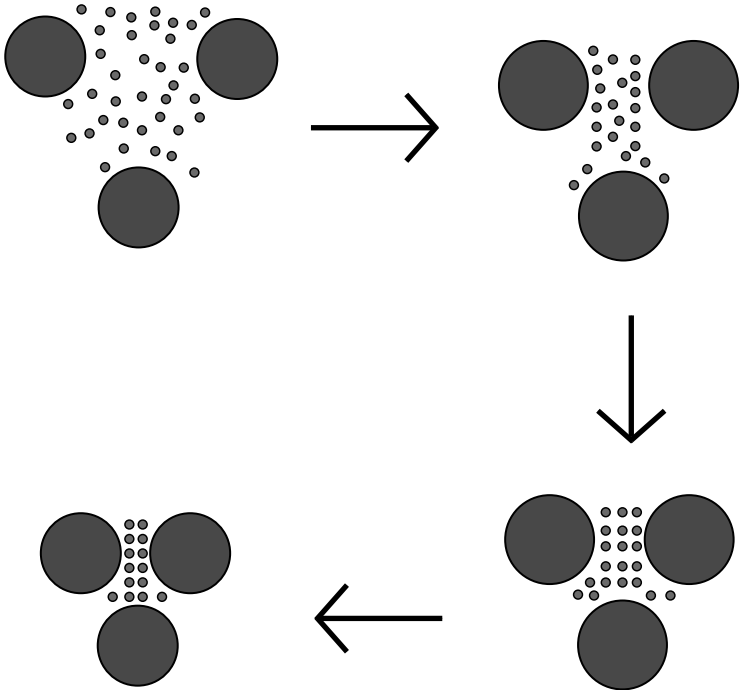


Fig. 1 Proposed stabilization mechanism. *(View this art in color at www.dekker.com.)*

of a dispersion of smaller particles will increase faster with an increase in solid content than that of a dispersion of larger particles, leading to a more stabilized system.[6] Therefore, with everything else being equal, smaller particles should stabilize better.

There are limitations on the size. As has been estimated by theoretical calculations, for spherical particles, ~4 nm will be the approximate minimum size that should stabilize an emulsion or foam. This represents the smallest particle that will not be enveloped by the amplitude of surface corrugations.[3]

Work was conducted in emulsion stabilization using only nanoparticles. It was surprising to find that these emulsions were stable for months (sometimes years) without high shear mixing. Particles with the same surface modification would stabilize a water-in-oil system and a fluorocarbon-in-oil system. This is nearly unheard of in emulsion science. Typically for a surfactant to be useful, it must be located at the interface between two phases, meaning that it must have a limited amount of solubility in both phases. Because water and fluorochemicals are vastly different in their chemical potentials, it is unlikely that a single surfactant would stabilize both types of systems. Fig. 2 shows optical micrographs of these multiple emulsions. On the left is one of the more interesting ''micelle'' shapes. On the right is the actual breaking of one of the emulsion droplets.

This led to the inclusion of two types of particles into the emulsions and the formation of multiple emulsions. For example, hydrophobically modified particles were dispersed in toluene, while hydrophilically modified particles were dispersed in water. Emulsification of this system produced water-in-toluene-in-water or toluene-in-water-in-toluene multiple emulsions. Formulations of one type over the other were achieved by changing the water/toluene ratio or hydrophobic/hydrophilic particle ratio. Fluorocarbon-in-toluene-in-water multiple emulsions were made by adding a fluorochemical fluid. Fig. 3 is an optical micrograph of a multiple emulsion containing Fluorinert® FC75 in toluene in water.

These multiple emulsions could be made by simple mixing, without any special order of addition of the ingredients. In fact, all three liquid components (the particles were previously dispersed in the respective phases) in the FC-in-toluene-in-water system were combined and emulsified simply by mixing. It is well understood that multiple emulsions are extremely process dependent and are nearly a black art. The ability to create these emulsions with minimal processing is a vast improvement over the current art.[7]

Dispersion stabilization with nanoparticles is also known. A recent example of a dispersion stabilized by nanoparticles was published by Tohver et al.[8] This group used zirconia particles to stabilize an aqueous colloidal system of larger silica particles. The dispersion was stabilized by electrostatic stabilization and thus is essentially applicable only to aqueous systems. Surface modification of the particles changes the stabilization mechanism to steric stabilization, and dispersions in both aqueous and nonaqueous systems have been demonstrated.

Foam, emulsion, and dispersion stabilization can be accomplished with organic molecules. With the proper chemistry, organic molecules also exhibit stabilization capabilities. Because many of these molecules are hydrophobic to begin with, further hydrophobic modification only makes them more compatible with the solvent, resulting in very few, if any, lyophobic/lyophilic regions that would cause them to partition the interface.

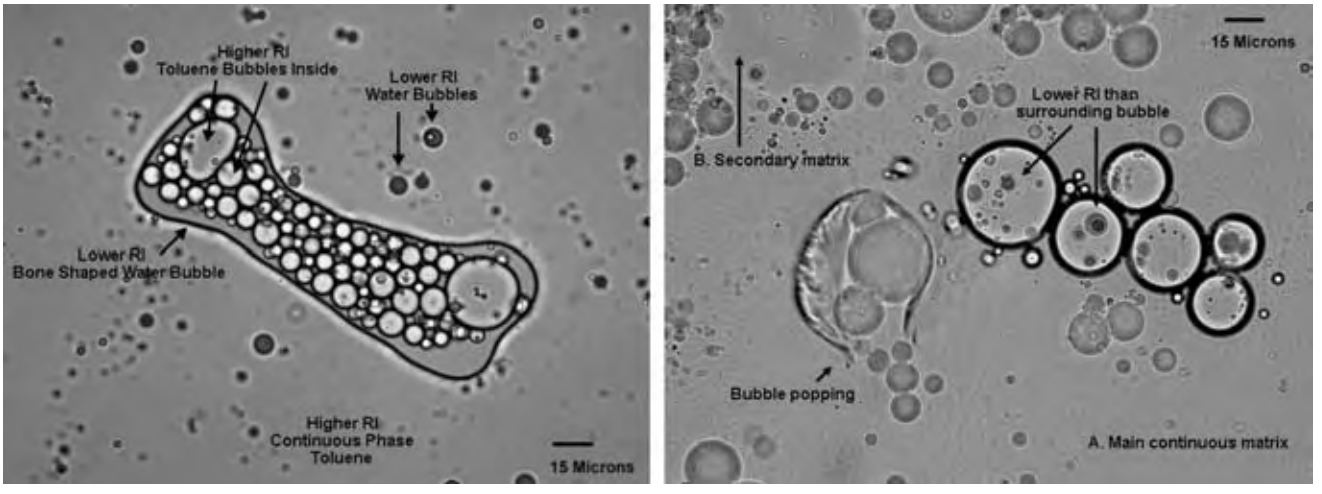


Fig. 2 Optical micrographs of multiple emulsions stabilized with nanoparticles.

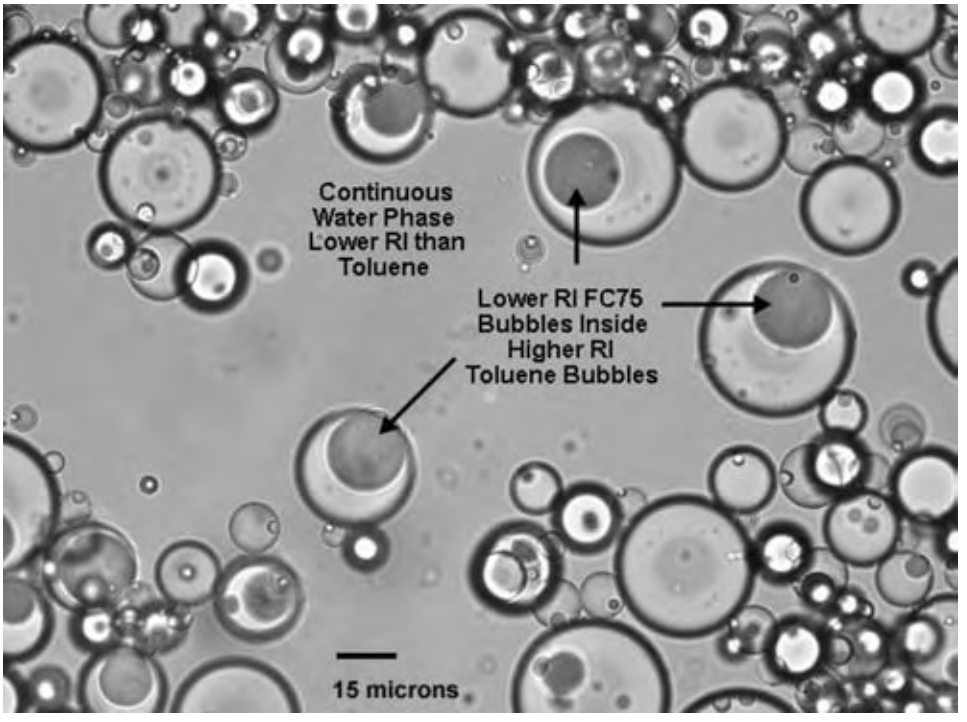


Fig. 3 An optical micrograph of fluorocarbon-in-toluene-in-water emulsion stabilized with nanoparticles.

GOLD NANOPARTICLE CATALYSTS

The science of metal nanoparticles has had a significant impact on commerce and industry, most prominently through the chemical transformations of hydrocarbons that lead to fuels and petrochemicals. Additionally, the physics of small metal particles has been fascinating for a long time. A problem considered in latter times has centered on quantum size effects: how electronic, and perhaps chemical, properties change as the particle size is reduced in three (nanoclusters), two (nanowires), or one (rafts or islands) dimensions. Gold nanoparticles are singled out in this entry because clusters and particles formed from main group and d-block elements comprise too vast a field. The majority of work on gold nanoparticles has been done in the past two decades. This smaller subfield replicates the puzzles and surprises presented by metal nanoparticles in general. Because it is so new, there are many technological opportunities to be explored and many fundamental questions remain to be answered. Gold nanoparticles are at a frontier in catalytic chemistry today.

The surface chemistry of bulk gold is limited in scope compared with its lighter congeners copper and silver, and to elements of the transition series. Oxygen absorption is strong but occurs rapidly only at $>500^{\circ}C$.[9,10] Chemisorptive properties are weak although of some importance in electrochemistry. Gold has long been known to catalyze the oxidation of CO to CO_2; the discovery that this reaction proceeds rapidly at ambient temperature on small gold particles was a breakthrough. Catalysis by or on gold nanoparticles has since developed rapidly, and this[11] as well as CO oxidation[12] has been recently reviewed. Broad topics discussed here include a) preparation and characterization of gold nanoparticle catalysts; b) the nature of the active site(s) and the mechanism of CO oxidation; and c) other processes catalyzed by gold nanoparticles.

Catalyst Syntheses and Characterization

Gold catalysts are made through two general methods that differ in terms of the principles involved, the materials produced, and the level of their understanding that has been achieved. In both, a solid support is used to stabilize the particles against sintering and to provide them in a form that is convenient to use.

Chemical methods

The first method involves chemical reduction of Au(III), usually as the hydrated H_3O^+ salt of $AuCl_4^-$. A preformed support may be impregnated with this "$HAuCl_4$" and then treated with a reducing agent. Alternatively, when the pH of a mixture of $HAuCl_4$ and various metal ions is raised, gold hydroxide and metal hydroxide coprecipitate. Drying, calcination, and reduction of this mixture provide Au(0) supported upon a metal oxide. Suitable reducing agents include H_2, BH_4^-, citrate, and ascorbate. Permutations of

support materials and reaction conditions are myriad. Some systematic studies are available,[13–18] but a real understanding of the role of processing variables has not been achieved. Supports can be carbon, metal oxides, or mesoporous oxides such as zeolites. Oxides can be categorized as reducible (TiO_2, Fe_2O_3, CeO_2) or not (ZrO_2, SiO_2, Al_2O_3).

Chemical vapor deposition (CVD) has been used to prepare nanoparticle catalysts, although the technology is quite expensive. Dimethylgold acetylacetonate absorbs on supports such as MgO. Decomposition at $>100°C$ produces an efficient CO oxidation catalyst. X-ray absorption near edge (XANES) and extended X-ray absorption fine structure spectroscopies (EXAFS) showed that, under steady-state CO oxidation conditions, the catalyst contained Au(0) in the form of (on average) Au_6 clusters and also additional gold as Au(I).[19–21]

Preformed gold clusters, obtained by the reduction of $(Ph_3P)AuCl$ followed by treatment with O_2 to remove capping ligands and then with H_2, have been used as catalysts.[22]

Usually, these catalysts are characterized in terms of their efficiency in some test reaction, usually CO oxidation, under arbitrary conditions (O_2 and CO concentrations, humidity, flow rates, and catalyst time on stream) that would be difficult to reproduce in another laboratory. Characterization on any length scale, from molecular to morphological, is extremely difficult. In many cases, it is uncertain whether all the reactants and by-products have been removed from the catalyst surface. It is reported that chlorides poison the active sites.[23] This raises the question of a role for gold ions, and these, rather than Au(0), are reported to be the active sites in the water-gas shift reaction.[24]

Transmission electron microscopy (TEM) and scanning tunneling microscopy (STM) have been used to visualize small particles.[25–27] Catalysts can contain large amounts of large particles and crystalline bulk gold. One study of Au/Al_2O_3, prepared by sputtering, found that some gold was unaccounted for in the mass balance and in size regime below TEM resolution limits ($\sim$1 nm).[28] It is not clear whether all the catalytically active gold has (or can be) viewed by microscopy. Gold particles <2 nm in size have structureless absorption, and larger ones show a surface plasmon peak in the visible.[29,30] Colors of gold catalysts thus can provide a crude, qualitative measure of gold aggregation. On MgO, they vary from light blue to blue, purple, and then pink with increasing particle size.[11] Individual characterization techniques have different limitations. For example, in x-ray photoelectron spectroscopy (ESCA) and TEM, a sample prepared (and used) under ambient conditions is studied under high vacuum conditions. In ESCA, photoelectrons can cause spurious reduction of cationic gold.

Incisive characterization of gold nanoparticle catalysts, however prepared, faces formidable obstacles. The root cause is this: virtually all experimental probes average over or ''see'' all (or almost all) of the gold in the sample. But it is not clear whether all the gold is responsible for all the catalytic activity. The effect of a small fraction of the total gold can be greatly amplified in catalysis. The same problem exists with particles and clusters of other metals. However, for gold, the effect on activity of particle size is extremely large, going from high at <10 nm diameter to zero for the bulk-like metal. In contrast, Ru and Pt are active under all degrees of dispersion (allowing for scaling for surface area).

Sputtering methods

This process is conducted under vacuum and deposition is line-of-sight. It is expensive and best suited to planar surfaces having relatively small areas. Impediments to commercial applications are aids to fundamental research. Catalysts can be prepared, studied, and tested under UHV conditions and in the absence of chemical contaminants. Most of our detailed understanding of gold nanoparticle catalysts comes from work on sputter-coated model catalysts.

Basic studies

It is difficult to deduce what gold particle morphologies arise from heterogeneous chemical reduction of $HAuCl_4$. Understanding of the model catalysts is much easier. In brief, a) nucleation of gold clusters occurs at surface defects that act as traps; b) on Al_2O_3, there are two kinds of traps at <0.8 and >1.6 eV; c) the defect density is ca. 3×10^{12} sites per cm^2 (10^{-3} monolayer); and d) when the clusters grow to >600 atoms, they leave the traps. This can explain the bimodal size distribution of the clusters.[31] Atomistic definition of these traps is needed.

Fundamental studies of metal nanoparticle growth on oxide surfaces indicate that, on a larger length scale, particles tend to collect at dislocations and grain boundaries,[32–34] and mobility on rough surfaces is lower than on smooth ones.[35] There is a significant size dependence,[36] and oxygen can enhance lateral diffusion.[37] Many theoretical studies have explored stability–size–geometry relationships in small gold clusters.[38–40] There are ''magic numbers'': Au_{20} in T_d form is particularly stable.[41] Metal–support interactions (which need to be explored further) and relativistic effects[42] tend to stabilize flat 2-D structures over polyhedral 3-D ones.

Ionized gold clusters generated in the gas phase can be mass sorted, deposited with low kinetic energy onto oxide surfaces, and then charge neutralized. For Au_n

($n > 2$) on MgO, Au_8 was found to be the smallest cluster that catalyzed CO oxidation and then only on defect-rich surfaces. Adsorption is accompanied by partial electron transfer into the cluster from surface defects (possibly oxygen vacancies). This may play a key role in enhancing catalytic activity. On Au_8, CO is a weakly bound η^1 ligand.[43,44] Oxidation by small gold clusters has been studied experimentally[45] and theoretically.[46,47] Interestingly, in the gas phase, only Au_n^- ($n \leq 20$, $\neq 16$ and even) reacts with O_2 to form $Au_n(O_2)^-$.

Scanning tunneling microscopy studies of gold nanoparticles on TiO_2 suggested that clusters may have unusual catalytic properties as one cluster dimension becomes smaller than three atomic spacings. Two-atom thick clusters show maximum activity for catalytic CO oxidation at diameters 2.5–3.0 nm. A metal-to-nonmetal transition occurs as cluster size decreases below 3.5 nm in diameter and 1 nm in height (ca. 300 atoms).[48,49] These results raise fundamental questions about the electronic nature of gold nanoparticles, particularly when they are small and thin. Also, O_2 is reported to bind more strongly to ultrathin Au islands than to thick ones.[50] Two-dimensional Au islands are thermodynamically less stable than three-dimensional aggregates and sintering occurs on annealing or in the presence of O_2.

A different view of CO oxidation comes from computational studies of extended Au(1 1 1). Steps and kinks play a critical role in reducing the activation energy for O_2 dissociation.[51] The enhanced activity of thin islands may be related to step density (geometric effects) rather than to quantum size effects.[52] The literature suggests a key role for such defects. Gold is an effective catalyst because it can bind CO and O_2 but weakly enough so that subsequent processes have achievable activation barriers.

Hydrogenation

Use of gold nanoparticles as hydrogenation catalysts is a recent development.[53] In hydrogenation of acrolein using Au/ZnO, edge sites effected reduction of carbonyl groups, whereas reduction of the C=C bonds occurred on facial sites.[54] Thiophene doping of supported gold increased the rate of carbonyl hydrogenation of crotonaldehyde,[55] a remarkable example of promotion of gold by sulfur.

Silane chemistry

Alkylsilanes, $RSiH_3$, are absorbed on gold with loss of hydrogen to form monolayers of RSi≡moieties.[56] Reaction of $RSiH_3$ with water is catalyzed by gold nanoparticles. Products include siloxane nanowires, filaments, and tubes.[57] These results indicate that gold nanoparticles have considerable potential in the synthesis of novel materials and in developing new chemistry of p-block hydrides.

Other applications

Gold nanoparticles have been used to catalyze the liquid-phase oxidation of *vic*-diols, e.g., ethylene glycol to glycolic acid.[58] This may presage an extensive liquid-phase catalytic chemistry. In the gas phase, $Au_n(OH)^-$ clusters form adducts with O_2.[59] Propylene oxide is formed over Au/TiO_2 from a mixture of propylene, H_2, and O_2. Very small gold particles tend instead to form propane.[60–63] Numerous other applications will surely soon be published.

NANOPHASE GLASS-CERAMICS

Nanophase glass-ceramics are an important class of materials that incorporate a nanoscale, particulate phase in a continuous matrix. These nanoscale grains are formed by controlled nucleation and growth from a precursor glass during a separate heat-treatment step that follows higher-temperature forming. A nanophase microstructure can impart unique optical and mechanical properties to glass-ceramic materials, as demonstrated through some commercially important examples.

Glass-ceramic materials were invented in the 1950s at Corning Glass Works by S. Donald Stookey[64] who discovered that when small amounts (e.g., <5 wt.%) of nucleating agents (e.g., TiO_2, ZrO_2) were added to certain glass-forming compositions, a separate heat-treatment step after forming could generate highly crystalline bodies. The devitrified materials displayed superior mechanical and thermal properties, including exceptional thermal shock resistance and strength. Glass-ceramic materials with many-fold increases in strength, as well as elimination of thermal expansion, as compared with their parent glasses, can be engineered. Through development efforts at Corning, Schott, Nippon Electric Glass, and IBM, among others, glass-ceramic materials have found useful application in cookware, cooktops, dental restorations, electronic packaging substrates, building materials, and machined structural components. Commercial glass-ceramics are sold under such trade names as MACOR®, DICOR®, Vitronit™, Corning Ware®, Vision®, Neoceram®, CERAN®, EMPRESS®, and others.[65]

Although a complete treatment is beyond the scope of this entry, it is worth considering the source of improved mechanical properties of glass-ceramic materials over simple glass materials. There are many aspects of mechanical performance that combine to define the utility of a material, including numerous

modes of strength and fracture toughness, as well as surface hardness and thermomechanical attributes. At the same time, the variety of glass-ceramic compositions has become enormous in the past half century. Accordingly, there is no way to efficiently capture the full breadth of what is known about the development of mechanical properties of glass-ceramic materials in general. However, there are some dominant themes that have emerged. Glass-ceramic materials, as brittle solids, display superior strength mainly by virtue of improved abrasion resistance and flaw tolerance, which develop when crystalline grains form regions with some inherently superior properties. However, those grains also frequently serve to create unique localized stress conditions for crack suppression and barriers for crack deflection in the microstructure. More detailed analyses of specific materials systems can follow numerous lines of explanation that include considerations of such factors as differences in thermal expansion coefficient, anisotropy in thermal expansion, and chemical substitution effects. Glass-ceramic materials, as composite materials, develop properties that depend on their different constituent phases as well as interactions between them. An example is the elimination or reversal of thermal expansion of some glass-ceramics that results from the crystallization of low or net-negative thermal expansion phases in a higher net-positive thermal expansion residual glass matrix. Finally, effects of grain size on strength of ceramics, mainly the effect that decreasing grain size almost always leads to higher strength, are manifest in glass-ceramics as well. Thus, it is not uncommon to observe eventual reductions in strength when glass-ceramics are heat treated to increase the grain size.

Recently, glass-ceramic materials with exceptionally fine-scale microstructures have been reported at an increasing rate. In most of the original work by Stookey, crystal sizes in the range of 0.1–20 μm were reported. Although such materials have great utility in a wide range of applications, certain advantages can be realized when the crystal size is further reduced to the nanoscale (<100 nm). Primarily, with a nanoscale crystal size, one can access advantageous combinations of high mechanical performance and improved optical transparency for many compositions. As mentioned, glass-ceramic materials in general are characterized by much higher strength, fracture toughness, and thermal shock resistance than their glass counterparts. However, glass-ceramics take on even more value when their optical properties can be tailored for certain applications. The rapid growth in the use of glass ceramic materials in cooktops, following a migration from opaque β-spodumene to transparent β-quartz glass-ceramics, is an example. In the latter case, optical transparency is achieved mainly through index of refraction matching and low birefringence for the crystalline phase, although a small crystal size of around 100 nm is achieved.

A good example where a unique combination of mechanical and optical properties has been achieved in a nanophase glass-ceramic material relates to high index of refraction beads or microspheres. 3M has recently patented transparent nanophase glass-ceramic microspheres that include comparatively large amounts of ZrO_2 and TiO_2, with index of refraction values that can exceed 2.0.[66–68] The materials are particularly useful as exposed lens elements in durable retroreflective pavement markings. Without index matching between the crystal grains and the surrounding glass phase, within the microspheres, these materials rely on ultrafine crystal size for transparency. Useful beads comprise crystal grains, with a size usually smaller than 20 nm. Other new nanophase glass-ceramic materials from 3M have been reported with ternary rare-earth-oxide-Al_2O_3–ZrO_2 compositions.[69] The glasses are formed with near-eutectic compositions, and can be devitrified with nanoscale grain structure and exceptional mechanical properties, including very high hardness. At the same time, optical transparency is preserved.

There are other examples of glass-ceramic materials that are crystallized to develop a particular property, but crystallized specifically with nanoscale grain structure to preserve optical transparency. Transparent oxyfluoride glass-ceramics with exceptional optical frequency upconversion were first reported by researchers at Sumita Optical Glass and NTT Opto-Electronics Laboratories in 1993.[70] Since that time, research at Corning and elsewhere has further proven the potential utility of the materials, which comprise a nanoscale dispersion of doped lead or rare earth fluoride crystals in an aluminosilicate matrix.[71] The nanophase glass-ceramics are preferred to competing materials, like simple glasses and single crystals. Nanophase glass-ceramics offer the outstanding fluorescence properties of an active rare earth doped fluoride, with their low phonon coupling losses, together with mechanical and chemical durability of an aluminosilicate glass, which derives from the residual matrix after devitrification.

Nanophase glass-ceramics, being nanocomposite materials, offer the potential to achieve unique combinations of properties that have not been demonstrated in other materials systems. Most significantly, small feature sizes in the multiphase microstructures can lead to optical transparency. At the same time, an interaction between phases, or a functional phase itself, can impart additional properties, such as high strength. Just as glass-ceramics in general have experienced extensive technical development and commercial application since their invention half a century ago, the advancement and the use of nanophase glass-ceramics

in particular can be expected to grow substantially over the coming years.

CONCLUSIONS

The capability to controllably engineer discrete nanoscale particulate domains in a material is a powerful tool for accessing new and unique material properties that enable innovative industrial applications. It is expected that research and development work on nanomaterials will continue to increase and numerous new applications will be identified.

REFERENCES

1. Scarlett, A.J.; Morgan, W.L.; Hilderbrand, J.H. Emulsification by solid powder. J. Phys. Chem. **1927**, *31*, 1566–1571.
2. Binks, B.P. Particles as surfactants—similarities and differences. Curr. Opin. Coll. Surf. Sci. **2002**, *7* (1–2), 21–41.
3. Nikolov, A.D.; Wasan, D.T. Ordered micelle structuring in thin films formed from anionic surfactant solutions. J. Coll. Interface Sci. **1989**, *133* (1), 1–12.
4. Asakura, S.; Oosawa, F. Interaction between particles suspended in solutions of macromolecules. J. Polym. Sci. **1958**, *33* (126), 183–192.
5. Kralchevsky, P.A.; Nikolov, A.D.; Wasan, D.T.; Ivanov, I.B. Formation and expansion of dark spots in stratifying foam films. Langmuir **1990**, *6* (6), 1180–1189.
6. Hug, J.E.; von Swol, F.; Zukowski, C.F. The freezing of colloidal suspensions in confined spaces. Langmuir **1995**, *11* (1), 111–118.
7. Silva Chunha, A.; Grossiord, J.L.; Seiller, M. The formulations and industrial applications of multiple emulsions: an area of fast development. In *New Products and Applications in Surfactant Technology*; Karsa, D.R., Ed.; CRC Press LLC: Boca Raton, FL, 1998; Vol. 1, 205–226.
8. Tohver, V.; Chan, A.; Sakurada, O.; Lewis, J.A. Nanoparticle engineering of complex fluid behavior. Langmuir **2001**, *17* (12), 8414–8421.
9. Chesters, M.A.; Somorjai, G.A. Chemisorption of oxygen on stepped surfaces. Surf. Sci. **1975**, *52*, 21–28.
10. Jugnet, Y.; SantosAires, F.J.; Deranlot, C.V.; Piccolo, L.; Bertolini, J.C. Chemisorption on Au(1 1 0). Surf. Sci. **2002**, *521* (1,2), L639–L634.
11. Bond, G.C.; Thompson, D.T. Catalysis by gold. Catal. Rev. Sci. Eng. **1999**, *41* (3,5), 319–388.
12. Bond, G.C.; Thompson, D.T. Gold catalyzed oxidation of carbon monoxide. Gold Bull. **2000**, *33* (2), 41–51.
13. Schumacher, B.; Plzak, V.; Kinne, M.; Behm, R.J. Highly active Au/TiO_2 catalysts. Catal. Lett. **2003**, *89* (1,2), 109–113.
14. Kung, H.H.; Kung, M.C.; Costello, C.K. Supported catalysts for CO oxidation. J. Catal. **2003**, *126* (1,2), 425–432.
15. Wolf, A.; Schüth, F. Synthesis of highly active gold catalysts. Appl. Catal. A. **2002**, *226*, 1–13.
16. Grisel, R.J.; Kooyman, P.J.; Nieuenhuys, B.E. Influence of the preparation of Au/Al_2O_3 on oxidation activity. J. Catal. **2000**, *191* (2), 430–437.
17. Schubert, M.; Hackenberg, S.; van Veen, A.C.; Muhler, M.; Plzak, V.; Behm, R.J. CO oxidation over gold catalysts. J. Catal. **2001**, *197* (1), 113–122.
18. Zanella, R.; Giorgio, S.; Henry, C.R.; Louis, C. Methods for preparation of gold nanoparticles. J. Phys. Chem. B. **2002**, *106* (31), 7634–7642.
19. Guzman, J.; Gates, B.C. Cationic and reduced gold in oxidation catalysts. J. Phys. Chem. B. **2002**, *106* (31), 7659–7665.
20. Guzman, J.; Gates, B.C. Mononuclear gold-complex catalyst. Angew. Chem. Int. Ed. **2002**, *42* (6), 690–693.
21. Okumura, M.; Tsubota, S.; Haruta, M. Preparation of supported gold catalysts. J. Mol. Catal. A. **2003**, *199* (1,2), 73–84.
22. Martra, G.; Prati, L.; Manfredi, C.; Biella, R.M.; Coluccia, S. Deposition of gold sols from $(Ph_3P)_3AuCl$. J. Phys. Chem. B. **2003**, *107* (23), 5343–5459.
23. Ohm, H.S.; Yang, J.H.; Costello, C.K.; Wang, Y.M.; Bare, S.R.; Kung, H.H.; Kung, M.C. Effect of chloride on supported catalysts. J. Catal. **2002**, *210* (2), 375–386.
24. Fu, Q.; Saltsburg, H.; Flytzani-Stephanopoulos, M. Ceria-based water gas shift catalysts. Science **2003**, *301* (5635), 935–938.
25. Boccuzzi, F.; Cerrato, G.; Pinna, F.; Strukul, G. Reduced reactivity on used gold samples. J. Phys. Chem. B. **1998**, *102* (30), 5733–5736.
26. Guczi, L.; Horváth, D.; Pászti, Z.; Toth, L.; Horváth, A.; Karacs, A.; Petõ, G.J. Modeling gold nanoparticles. J. Phys. Chem. B. **2000**, *104* (14), 3183–3193.
27. Akita, T.; Lu, P.; Ichikawa, S.; Tanaka, K.; Haruta, M. Dispersion of nanoparticles in Au/TiO_2. Surf. Interface Anal. **2001**, *31* (2), 73–78.
28. Carrey, J.; Maurice, J.-L.; Petroff, F.; Vaurès, A. Evidence of cluster mobility. Surf. Sci. **2002**, *504*, 75–82.
29. Alvarez, M.M.; Khoury, J.T.; Schaaf, T.G.; Shafigullin, M.N. Spectra of nanocrystal gold

molecules. J. Phys. Chem. **1997**, *101* (19), 3706–3712.
30. Link, S.; El-Sayed, M.A. Surface plasmon electronic oscillations. J. Phys. Chem. B. **1999**, *103* (40), 8410–8426.
31. Carrey, J.; Maurice, J.-L.; Varuès, A. Evidence of carrier mobility. Phys. Rev. Lett. **2001**, *86* (20), 4600–4603.
32. Frank, M.; Bäumer, M. From atoms to crystallites. Phys. Chem. Chem. Phys. **2000**, *2* (17), 3723–3738.
33. Freund, H.-J.; Baumer, M.; Kuhlenback, H.K. Supported cluster model systems. Adv. Catal. **2000**, *45*, 333–378.
34. Henry, C.R. Studies of supported catalysts. Surf. Sci. Reports. **1998**, *31* (7,8), 235–325.
35. Yoon, B.; Luedtke, W.D.; Gao, J.; Landman, U. Gold on defective graphite clusters. J. Phys. Chem. B. **2003**, *107* (24), 5882–5891.
36. Campbell, S.C.; Parker, S.C.; Starr, D.E. Size-dependent nanoparticle sintering. Science **2002**, *298* (5594), 811–814.
37. Sykes, E.H.C.; Williams, F.J.; Tikhov, M.; Lambert, R.M. Nucleation, growth mobility and absorption. J. Phys. Chem. B. **2002**, *106* (21), 5390–5394.
38. Wang, J.; Wang, G.; Zhuo, J. Density functional study of Au_n clusters. Phys. Rev. B **2002**, *66* (3), 035418–035424.
39. Häkkinen, H.; Landman, U. Clusters and their anions. Phys. Rev. B. **2000**, *62* (7), R2287–R2290.
40. Häberlen, O.; Chung, S.-C.; Stener, M.; Rösch, N. From clusters to bulk. J. Chem. Phys. **1997**, *106* (12), 5189–5201.
41. Li, J.; Li, X.; Zhai, H.-J.; Wang, L.-S. Au_{20}: tetrahedral cluster. Science **2003**, *299* (5608), 864–867.
42. Häkkinen, H.; Moesler, M.; Landman, U. Bonding in gold: relativistic effects. Phys. Rev. Lett. **2002**, *89* (3), 033401.
43. Sanchez, A.; Abbet, S.; Heiz, U.; Schneider, W.-D.; Häkkinen, H.; Barnett, R.N.; Landman, U. Why gold is not noble. J. Phys. Chem. B. **1999**, *103* (48), 9573–9578.
44. Heiz, U.; Sanchez, A.; Abbet, A.; Schneider, W.-D. Using nanoassembled model catalysts. Chem. Phys. **2000**, *262* (1), 189–200.
45. Stolcic, D.; Fischer, M.; Ganteför, G.; Kim, Y.D.; Sun, Q.; Jena, P. Observation of key reaction intermediates. J. Am. Chem. Soc. **2003**, *125* (10), 2848–2849.
46. Lopez, N.; Nørskov, J.K. Catalytic oxidation by gold nanoparticles. J. Am. Chem. Soc. **2002**, *124* (38), 11262–11263.
47. Hammer, B.; Nørskov, J.K. Theoretical surface science and catalysis. Adv. Catal. **2000**, *45*, 71–129.
48. Valden, M.; Lai, X.; Goodman, D.W. Onset of activity of gold on titania. Science **1998**, *281* (5383), 1647–1650.
49. Lai, X.; St. Clair, T.P.; Valden, M.; Goodman, D.W. Scanning tunneling microscopy of clusters. Prog. Surf. Sci. **1998**, *59* (1–4), 25–52.
50. Bondzie, V.A.; Parker, S.C.; Campbell, C.T. Kinetics of CO oxidation. Catal. Lett. **1999**, *63* (3,4), 143–151.
51. Liu, Z.-P.; Hu, P.; Alavi, A. DFT study of CO oxidation. J. Am. Chem. Soc. **2002**, *124* (49), 14770–14779.
52. Mavrikitis, M.; Stoltze, P.; Nørskov, J.K. Making gold less noble. Catal. Lett. **2000**, *64* (2–4), 101–106.
53. Mohr, C.; Claus, P. Hydrogenation properties of gold. Sci. Prog. **2001**, *84* (4), 311–334.
54. Mohr, C.; Hofmeister, H.; Radnik, J.; Claus, P. Gold-catalyzed hydrogenation of acrolein. J. Am. Chem. Soc. **2003**, *125* (7), 1905–1911.
55. Bailie, J.E.; Hutchings, G.J. Promotion by sulfur of gold catalysts. Chem. Commun. **1999**, *21*, 2151–2152.
56. Owens, T.M.; Nicholson, K.T.; Banaszak Holl, M.; Süzer, S. Alkylsilane-based monolayers on gold. J. Am. Chem. Soc. **2002**, *124* (24), 6800–6801.
57. Prasad, B.L.V.; Stoeva, S.; Sorensen, C.; Zaikovski, V.; Klabunde, K.J. Polymerization of alkylsilanes to nanowires. J. Am. Chem. Soc. **2003**, *125* (35), 10,488–10,489.
58. Prati, L.; Rossi, M.J. Selective liquid phase oxidation. J. Catal. **1998**, *176* (2), 552–560.
59. Wallace, W.T.; Wyrwas, R.B.; Whetten, R.L.; Mitrić, R.; Bonačyić-Koutecký, V. Absorption on gold cluster anions. J. Am. Chem. Soc. **2003**, *125* (27), 8408–8414.
60. Hayashi, T.; Tanaka, K.; Haruta, M.J. Catalysis **1998**, *178* (2), 566–575.
61. Ajo, H.M.; Bondzie, V.A.; Campbell, C.T. Propene absorption on gold. Catal. Lett. **2002**, *78* (1–4), 359–368.
62. Zwijnenburg, A.; Goosens, A.; Sloof, W.G.; Crajé, W.J.; van der Kraan, A.M.; de Jongh, L.J.; Makkee, M.; Moulijn, J. Mossbauer characterization of Au/TiO_2. J. Phys. Chem. B. **2002**, *106* (38), 9853–9862.
63. Kapoor, M.P.; Sinha, A.K.; Seelan, S.; Inagaki, S.; Tsubota, S.; Yoshida, H.; Haruta, M. Hydrophobicity induced vapor phase oxidation. Chem. Commun. **2002**, *23*, 2902–2903.
64. Stookey, S.D. Ceramic Body and Method of Making US Patent 2,971,853, 14 February 1961.

65. Höland, W.; Beall, G. *Glass-Ceramic Technology*; The American Ceramic Society: Westerville, OH, 2002.
66. Kasai, T.; Budd, K.D.; Lieder, S.L.; Laird, J.A.; Yokoyama, C.; Naruse, T.; Matsumoto, K.; Ono, H. Transparent beads and their production . US Patent 6,335,083, 1 January 2002.
67. Budd, K.D.; Kasai, T.; Roscoe, S.B.; Yokoyama, C.; Bailey, J.E. Transparent Microspheres US Patent 6,245,700, 12 June 2001.
68. Frey, M.H.; Studiner, C.J.; Budd, K.D.; Kasai, T.; Roscoe, S.B.; Yokoyama, C.; Bailey, J.E. Glass-ceramic Microspheres Impart Yellow Color to Retroreflected Light US Patent 6,479,417, 12 November 2002.
69. Rosenflanz, A. Al_2O_3-Y_2O_3-ZrO_2/HfO_2 Materials, and Methods of Making and Using U.S. Patent Application 20,030,110,708, 19 June 2003.
70. Wang, Y; Ohwaki, J. New transparent vitroceramics codoped with Er^{3+} and Yb^{3+} for efficient frequency upconversion. Appl. Phys. Lett. **1993**, *63* (24), 3268–3270.
71. Beall, G.H.; Pinckney, L.R. Nanophase glass-ceramics. J. Am. Ceram. Soc. **1999**, *82* (1), 5–16.

Nanoporous Dielectric Materials

Jorge A. Lubguban
Shubhra Gangopadhyay
University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

A solid, liquid, or gas that exhibits poor conductivity of electricity is generally considered as dielectric material. Examples of such material include porcelain, mica, glass, plastics, and metal oxides, which have a wide range of dielectric constants (k). The dielectric constant relates to the capacitance enhancement over that of a vacuum. In microelectronics, dielectric materials are simply called dielectrics. They are used as insulation between metal interconnect and as passivation layers. The insulation capability of a material is determined by its breakdown strength. Low dielectric constant materials have a high breakdown strength and the ability to withstand intense electrostatic field. Nanoporous dielectric materials are related to microelectronics and are material candidates for future dielectrics. The semiconductor industry needs to continuously improve performance and speed in ultra-large-scale integration devices has motivated researchers to look into nanoporous dielectrics as a replacement of the standard SiO_2 material.

HISTORY OF DIELECTRIC MATERIALS

Since the invention of integrated circuits (ICs) in 1959, the progress of this technology has been extremely fast paced.[1] Advances in device fabrication, such as lithography, etching, and film deposition, in over four decades have brought extraordinary increases in IC complexity together with tremendous improvement in chip performance. The main factor driving this complexity and improved performance is the ability to scale down device dimensions into the microscopic regime. This can be illustrated by the continued shrinking of the transistor gate lengths from 7 μm to submicrometer levels today, facilitating the decrease in gate delays to negligible levels as well as the increase in transistor density. The number of transistors in a single microprocessor chip if plotted as a function of time from 1970 to 2000 traces a semilog behavior and follows a trend that roughly doubles the number of transistors every 18 mo.[1] The regular doubling of transistors is known as Moore's law. Fig. 1 illustrates Moore's law for IC.[1] It shows an exponential increase in the transistors per chip as a function of time for different generations of microprocessors. Advanced semiconductor chips are estimated to have more than 0.5 billion transistors and require over 10,000 m of wiring distributed over six to eight levels.[2]

As devices are continually scaled down to satisfy the ever-increasing demand for speed, it is recognized that device physics will not be the first limiting factor for performance improvement.[3] The main hindrance lies in the back end of the line (BEOL), which consists of metal interconnect and dielectric materials. The metal lines are responsible for transmitting electrical current and distributing clock signals that control the timing and synchronize the operation, whereas the dielectric materials are used as insulation. As feature size (generally the metal oxide semiconductor field effect transistor channel length) decreases below 0.1 μm, the BEOL interconnect resistance–capacitance (RC) delay, power dissipation, and cross-talk noise increase significantly.[4] The RC delay is given by the following equation:

$$\mathrm{RC}_{\mathrm{delay}} = 2\rho\varepsilon_0 k\left[(4L^2/P_{\mathrm{m}}^2) + (L^2/T^2)\right] \quad (1)$$

where ρ = interconnect resistivity, ε = permittivity free space, k = dielectric constant of insulator, L = total line length, P_{m} = metal pitch, and T = metal thickness. The power consumption, P, is given by:[3]

$$P = \alpha C f V^2 \quad (2)$$

where α = wire activity, C = capacitance, f = frequency, and V = voltage. With the increase in L and f and decrease in T and P_{m}, both $\mathrm{RC}_{\mathrm{delay}}$ and P increase. To address these issues, alternative materials must be used for BEOL. The standard for BEOL materials (Al and SiO_2) needs to be replaced with low-resistivity metal and low-dielectric constant materials. The successful implementation of Cu as replacement for Al has resulted in a decrease in $\mathrm{RC}_{\mathrm{delay}}$ by as much as 37% because of the lower resistivity of Cu. Fluorinated silicon dioxide films with k value from 3.2 to 3.6 have been successfully integrated at the 180 nm node replacing SiO_2 dielectric with $k = 4.0$.[5] Organosilicates with dielectric constant of 2.6–2.8 deposited by current chemical vapor deposition (CVD) tool have also been introduced and developed with Cu technology.[6]

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120026941

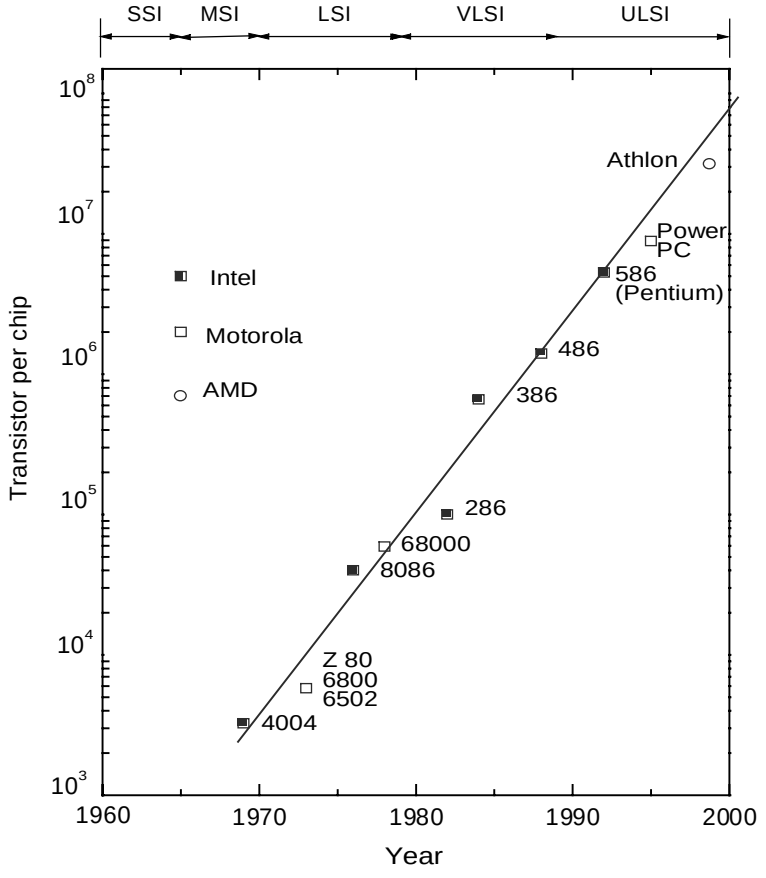


Fig. 1 Exponential increase of transistor count as a function of time for different generations of microprocessors. (From Ref.[1].) *(View this art in color at www.dekker.com.)*

Dielectric materials with even lower k values are needed for future generation IC devices. The International Technology Roadmap for Semiconductors projecting the overall technology requirements since 1994 suggested that for technology nodes of less than 65 nm, materials with $k < 2.1$ are needed. As a limited number of fully dense materials satisfy $k < 2.5$, researchers have focused on the preparation of porous films. The ability to tailor the degree of porosity, size, and shape of pores offers versatility and extendability from one generation to the next, which makes porous films attractive candidates. Fig. 2 shows the dielectric constant for methylsilsesquioxane (MSSQ) (discussed later) film vs. the pore volume fraction, V, as obtained from Bruggeman-effective-medium approximation. The figure shows that we can decrease the dielectric constant by tailoring V.

NANOPOROUS DIELECTRICS DEFINITIONS

The most common characteristic of a porous material is its gas–solid interphases.[7] Historically speaking, porous materials are defined in terms of their adsorption properties.[8] The pore sizes of the materials are frequently characterized from their gas adsorption isotherm, i.e., the amount of gas adsorbed by the material at a fixed temperature as a function of pressure.[9] The International Union for Pure and Applied Chemistry (IUPAC) has proposed to classify porous materials in terms of these isotherms. Nanoporous materials

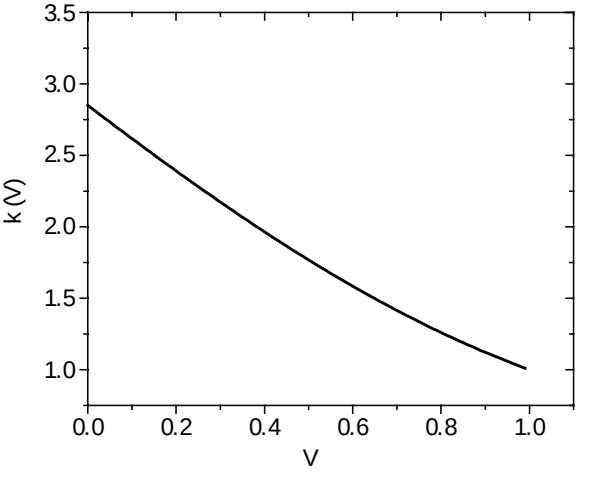


Fig. 2 Bruggeman effective medium approximation of a two-component film of a PMSSQ ($k = 2.85$) and air ($k = 1$).

pertain to those with pore sizes in the nanometer range of 1–1000 nm. However, for practical semiconductor applications pore sizes of interest must be significantly smaller than current feature sizes. The IUPAC defines pore diameters as follows: micropores as having 0–2 nm-diameter pores; mesopores as having 2–50 nm-diameter pores; and macropores as having greater than 50 nm-diameter pores.[10] Pores can be classified as closed or interconnected. Further classifications are in terms of network material, which can be inorganic or organic, and also in terms of being ordered or disordered.[7] Nanoporous materials can be characterized by gas adsorption methods, mercury porosimetry, small-angle x-ray scattering (SAXS), small-angle neutron scattering, positron annihilation lifetime spectroscopy, and electron microscopy methods like scanning electron microscopy, and transmission electron microscopy (TEM).[3,8,11]

PREPARATION OF NANOPOROUS DIELECTRIC MATERIALS

Thin film dielectrics are usually deposited using chemical vapor deposition (CVD). A variation of CVD utilizing a plasma discharge is called plasma-enhanced CVD (PECVD) and is the standard in IC fabrication for the deposition of dielectric films. Plasma-enhanced CVD involves the formation of a solid film in a substrate surface from volatile precursors (vapor or gas) in a plasma discharge. The precursors are chosen to contain the constituent elements of the final film and chemical reactions in the gas phase are encouraged. They are condensed in a substrate that is heated or cooled. It will be shown later that porosity can be introduced in the PECVD films. Spin coating is another preparation technique and a popular choice in future nanoporous dielectric materials. In spin coating, a solution (in soluble form or ''sol'') is prepared containing the material constituents of the final film mixed with a compatible solvent. The succeeding important steps are spinning, baking, and curing. Typical rotational speeds are 2000–6000 rpm. The bake step, typically below 200°C, expels the solvents and the final step called cure, done typically above 200°C, hardens the film via cross-linking or polymerization. The final thickness of the film depends on the fluid viscosity and density, rotation speed, rate of evaporation, and other factors. The initial amount of solution has little effect on the final thickness. Both CVD and spin-coating methods produced films with good planarization and gap fill properties.

REQUIREMENTS FOR DIELECTRIC MATERIALS

The most important requirement for future dielectric material is low dielectric constant. However, this is hardly the main requirement. Researchers have routinely shown films with k-values lower than 2.0 but these are not feasible from the point of view of manufacturing. The International Technology Roadmap for Semiconductors, which assessed the overall technology requirements for future microelectronic devices, has projected that no manufacturable dielectric with $k < 2.4$ will be found until 2007.[12] The main challenge lies in the integration of these films, which undergo harsh conditions in IC fabrication. Table 1 enumerates important properties of candidate materials.[13] The candidate material should not only satisfy these requirements, but also must have versatility and extendability for a future generation of nodes to be cost-effective.[4]

Table 1 Property requirements of low-dielectric-constant interlayer-dielectric materials

Electrical	Chemical	Mechanical	Thermal
k, Anisotropy	Chemical resistance	Thickness uniformity	High thermal stability
Low dissipation	Etch selectivity	Good adhesion	Low coefficient of thermal expansion
Low leakage current	Low moisture	Low stress	
Low charge trapping	Absorption	High hardness	Low thermal shrinkage
High electric-field strength	Low solubility in H_2O Low gas permeability	Low shrinkage Crack resistance	Low thermal weight loss High thermal conductivity
High reliability	High purity No metal corrosion Long storage life Environmentally safe	High tensile modulus	

(From Ref.[13].)

FACTORS AFFECTING THE DIELECTRIC CONSTANT

The dielectric constant of a material depends primarily on its polarizability and, hence, strongly depends on density. The three types of polarization that contribute to the dielectric constant are electronic, ionic, and orientational polarizations and are given by the Debye equation[3]

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{N}{3\varepsilon_0}\left(\alpha_e + \alpha_d + \frac{\mu^2}{3kT}\right) \quad (3)$$

where α_e is the electronic polarization, α_d is the ionic polarization, and $\mu^2/3kT$ is the permanent dipole contribution.

To obtain low k, these three polarizations are kept as low as possible. An obvious way to have low polarizations is reduction of the density of the material (N). The lower density will decrease the number of polarizable species in the films and thus results in a lower dielectric constant. This is done by incorporating low molecular weight molecules, space occupying molecules, inherently open structures, and, more significantly, introducing porosity. This entry presents an overview of nanoporous dielectric materials. This overview is not exhaustive because new materials are being developed as this entry is written.

NANOPOROUS DIELECTRIC MATERIALS AND PORE GENERATION METHODS

Nanoporous dielectric materials in microelectronics are inorganic, organic, or hybrids. Silsesquioxane (SSQ)-based, silica-based, and organic polymers are well studied. There are many dense materials that exhibit molecular porosity because of their inherent open structure and also because of incorporation of space occupying structures like methyl groups that provide steric hindrance. These materials will be mentioned here as they are a popular matrix host for the introduction of porosity. Maex et al. have classified two routes for fabrication of nanoporous materials as constitutive and subtractive.[3] In subtractive route, less stable materials from a stable matrix are removed selectively by different processes to produce a porous material. This is an effective method in producing porous materials with tunable porosities. Examples of constitutive porous materials are CVD and spin-coating organosilicate films. In these, methyl structures provide steric hindrance and open structures creating molecular-level porosities. The addition of sacrificial material or porogen (pore generator) in a film matrix is the most widely used technique in subtractive pore generation. Basically, it relies on incorporation of a thermally degradable material within a thermally stable matrix. The sacrificial material is removed usually by thermal degradation and volatilization to produce a porous structure. There are other techniques that will be shown later to remove the sacrificial material. One can use inorganic, organic, and hybrid materials in this approach. The availability of pore generators with different morphologies and the ability to vary the precursor ratios provide versatility and control not only the pore size and pore distribution but also the pore shape of the film.

BLOCK COPOLYMERS

A block copolymer is composed of chemically distinct and mutually incompatible homopolymer segments, and is covalently joined at the ends. It can be synthesized through different chemical pathways, typically, anionic polymerization. This type of polymerization is initiated by nucleophilic addition to two double bonds of the monomer. In the case of linear AB block copolymer synthesis, a nucleophile or a carbanion can originate from an organometallic species such as alkyl lithium or Grignard reagents (initiators). The carbanion now reacts with another monomer molecule on sequential monomer addition and at this point an appropriate terminating agent is added. To fabricate a nanoporous material, a thermally degradable polymer block is used with another block that is thermally stable to act as the matrix. The polymers are mixed with a casting solvent and spin coated to a substrate. The precursors form block copolymers and undergo self-assembly, where the thermally degradable block is surrounded by the thermally stable block by manipulating precursor ratios. The film is then annealed at higher temperatures, but below the glass transition temperature (T_g) of the matrix, to volatilize the thermally degradable block producing a nanoporous dielectric. Matrix polymers must have good dielectric properties and a higher T_g when compared with the decomposition temperature of the sacrificial polymer. Suitable matrix polymers enumerated by Hedrick et al. include polyimides, poly(phenylguinoxalines), poly(benzoxazoles), poly(benzimidazoles), poly(triazoles), poly(oxadiazoles), and poly(benzothiazoles).[14] For the thermally decomposable oligomers, poly(propylene oxide), poly(methylmetharcrylate), and aliphatic polycarbonates like poly(propylene carbonate) and poly(ethylene carbonate) can be used.[14] Common organic casting solvents include tetrachloroethane, *N*-methylpyrrolidone or γ-butyrolactone. An illustration scheme for producing nanopores via block copolymers is shown in Fig. 3 where poly(propyleneglycol) (PPG)–polyimide–PPG triblocks are used as building blocks

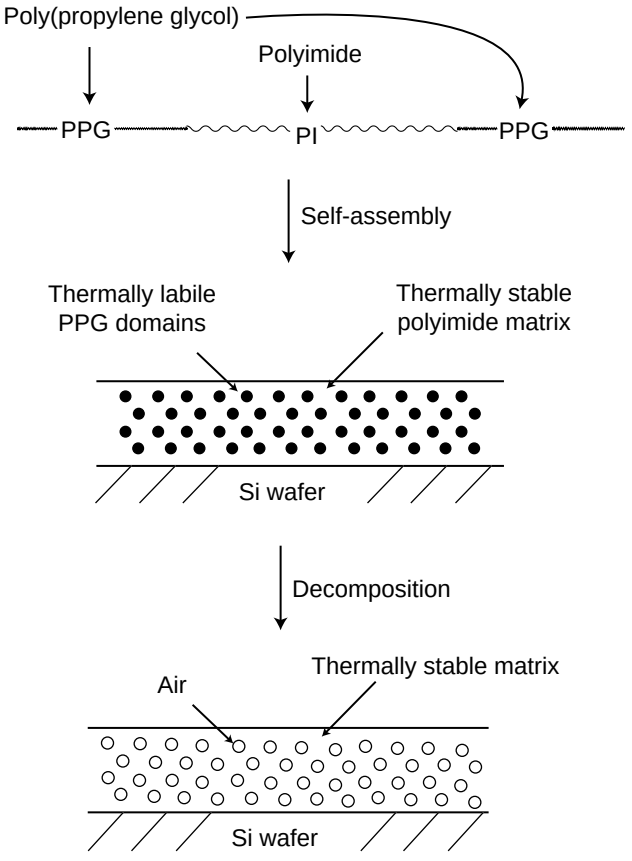


Fig. 3 Schematic representation of the self-assembly of block copolymers to give porous dielectric materials. (From Ref.[15].)

for self-assembly.[15] The weight percent of the polyimide matrix is increased relative to PPG to form nanoscopic domains of PPG surrounded by polyimides. The PPG is then removed by thermal decomposition and volatilization at a temperature below the glass transition (T_g) of the polyimide to prevent collapse of matrix. This process leaves behind pores (typically 5–10 nm), which take the morphology of the sacrificial block.[16] This successful approach has drawbacks because of the low inherent T_g of the matrix polymer, low hardness, and non-cross-linking of the organic matrix, which is prone to collapse. This approach has been extended to inorganic–organic hybrid materials.

SOLVENT AS POROGEN APPROACH

Another variation of sacrificial material technique to produce porous materials with less than 30 nm pore diameter and less than 10 μm thickness is using a high-boiling point solvent as the templating material.[17,18] This solvent is mixed with a low-boiling point solvent and then used as solvents for the polymer material that acts as the host matrix in the final film. Typically, the low-boiling point solvent has a volatilization temperature of less than 100°C and the high-boiling point solvent, a temperature greater than 150°C. After spin casting, the film is partially dried to expel a large amount of low-boiling point solvent. The film is then put into contact with a chemical, which is miscible to the high-boiling point solvent but immiscible to the polymer to induce nonsolvent phase separation in a process called phase inversion. The film is then annealed to completely remove the solvents and toughen the matrix producing a porous material. Cross-linking of the polymer matrix before or after phase inversion may be done by ultra violet or chemical processes. O'Neill et al. provided polymer precursors, such as poly(arylene ethers), polyimides, poly(phenyl quinoxalines), substituted poly(*p*-phenylenes), poly(benzobisoxazoles) polybenzimidazoles, polytriazoles, and mixtures thereof.[18] The low-boiling point solvents includes tetrahydrofuran, acetone, 1,4-dioxane, 1,3-dioxolane, ethyl acetate, methyl ethyl ketone, cyclohexanone, cyclopentanone, and mixtures thereof. The high-boiling point solvent can be selected from dimethylformamide, dimethylacetamide, *N*-methyl pyrrolidone, ethylene carbonate, propylene carbonate, glycerol and derivatives, naphthalene and substituted versions, acetic acid anyhydride, propionic acid and propionic acid anyhydride, dimethyl sulfone, benzophenone, diphenyl sulfone, sulfolane, phenol, *m*-cresol, dimethyl sulfoxide, diphenyl ether, terphenyl, cyclohexanone, cyclopentanone, and mixtures thereof.[2] The nonsolvent (immiscible to the polymer matrix) for the polymers can be selected from the group consisting of water, methanol, ethanol, isopropanol, toluene, hexane, xylene, cyclohexane, butanol, cyclopentane, octane, and mixtures of these solvents.[18]

NANOPOROUS SILICA VIA SOL–GEL TECHNIQUE

Silica in the form of thin films as well as oxide monoliths, fibers, and powders can be prepared from sol–gel method. In contrast with the fabrication of conventional inorganic glasses at much higher melting temperature, sol–gel processing is performed at low temperatures to produce oxide materials with desirable hardness, optical transparency, chemical durability, tailored porosity, and thermal resistance.[9] The sol–gel method involves formation of a colloidal suspension (sol) and gelation to form a network in a continuous liquid phase (gel). One starts with an aqueous solution containing oxides or alkoxides, mutual solvent, and catalyst. Usually an external catalyst is added like mineral acids and ammonia as well as acetic acid, KOH, amines, KF, and HF for rapid and

complete hydrolysis. The most widely used alkoxides are tetramethoxysilane (TMOS) and tetraethoxysilane (TEOS), because they readily react with water. However, other alkoxides like aluminates, titanates, and borates are also used. Alkoxides and water are not miscible, hence we use a mutual solvent like alcohol. The general reactions involved in sol–gel are hydrolysis and condensation. If an alkoxide precursor is used, a general scheme to produce silica can be written as:[7]

$$Si(OR)_4 \xrightarrow{H_2O} [Si(OH)_4] \longrightarrow SiO_2 \quad (4)$$

The hydrolysis reaction replaces the alkoxide groups (OR) with hydroxyl groups, after which condensation reactions involving the silanol groups (Si–OH) occur to produce siloxane bonds (Si–O–Si). The condensation reactions form a soluble high molecular weight polysilicate (a sol) and these polysilicates link together to form a three-dimensional network, whose pores are filled with solvent molecules (a gel).[9] If the liquid in the gel is removed without collapsing the structure, a highly porous and extremely low-density material can be obtained. Thin films can be produced by spin-coating or dip-coating the sol in a substrate. The spin coating is divided into four stages: deposition, spin-up, spin-off, and evaporation.[19] Gel aging and drying after the spin-coating step are the two most critical steps in sol–gel thin film processing. Drying usually involves shrinkage, implying reduction in pore volume and pore sizes and possible pore collapse. Shrinkage during drying must not occur to prevent cracking, pullout, adhesive loss, and delamination in the patterned substrate. Generally, the film must gel first before drying to prevent collapse. There are two routes of drying: if the film is dried by solvent evaporation, it is called xerogel and if dried supercritically, it is called aerogel. The final stage of sol–gel processing is heat treatment. The organics and adsorbed solvents burn off at temperatures of about 400°C. Fig. 4 is a simple illustration of sol–gel silica formation in thin films.

The porosity of the film can be tailored during initial stages and other processing conditions. The initial factors include pH (acid- or base-catalyzed), temperature of reaction, reagent concentrations, and H_2O/Si molar ratio. The processing conditions that affect the porosity are aging, temperature, and time of drying.[20] For example, silica xerogels processed under acidic conditions display characteristics of microporous materials, while those processed under basic conditions display mesoporous characteristics. Silica materials with porosity of 2–80 nm with a variety of pore shapes can be obtained by sol–gel method.[21] The acceptance of this technology is hampered by elaborate processing conditions and the hydrophilic nature of the pores, which leads to increased dielectric constants.[21]

NANOPOROUS SILICA VIA SURFACTANT-TEMPLATING APPROACH

The semiconductor industry has recognized the potential of silica mesophases for use as dielectrics, because it is built on the idea of true templating.[22] Here, the final pore size and geometry of the material are a direct resemblance of the initial surfactant array size and shape. In the fabrication of these bulk materials, the pore geometry can be controlled primarily by the concentration of surfactant in the solution as shown in Fig. 5 and also by the processing conditions.[22] The pore size is controlled by the length of surfactant, addition of auxiliary organics, choice of solvent, template removal method, aging conditions, processing conditions, and postfunctionalization of the isolated porous material.[22] However, there is difficulty in controlling the film formation process of this material. The basic process that will be presented here is based on fabrication of bulk material. It is recognized that film formation is different and the mechanism is complex and not yet well understood.

This technique uses surfactant as organic templates for the inorganic network to build on. The surfactants are molecules with a hydrophilic head group and a hydrophobic tail. They can aggregate into micelles with different supramolecular structures like spherical, cylindrical, hexagonal closed-packed liquid crystal, and cubic bicontinuous shapes depending on its concentration in the solution. The fabrication of mesoporous silica is a three-step process that includes synthesis, drying, and template removal. In film formation, a solution is made, spin-coated into a substrate, and then dried after which the template is removed. The synthesis of mesoporous silica requires a silica source, surfactant, water, and catalyst. The silica source acts as the inorganic network and conforms to the template, because it is precondensed initially. The reaction

Fig. 4 Nanoporous silica thin films by sol–gel process. *(View this art in color at www.dekker.com.)*

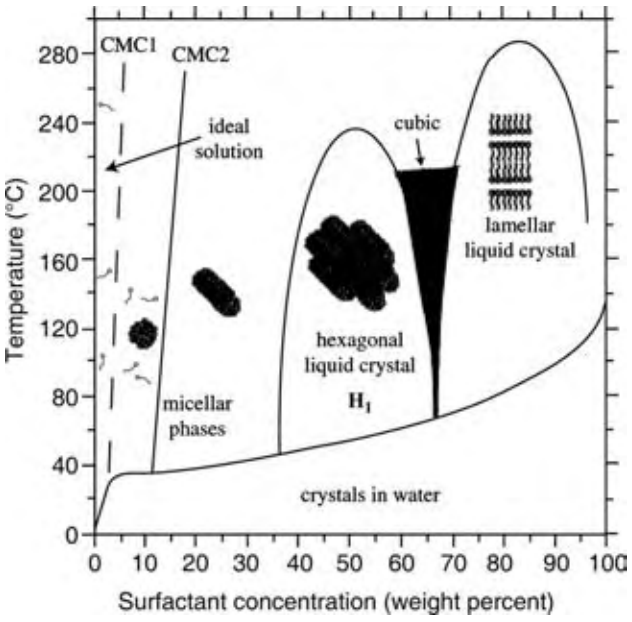


Fig. 5 Schematic phase diagram for C_{16} TMABr in water. CMC1 is increased to a higher concentration. (From Ref.[22].)

between the network and templating materials is via electrostatic, van der Waals, and hydrogen bonding interactions. Examples of nonmolecular silica sources are fumed silicas, precipitated silicas, or water glass, whereas molecular silica sources are alkoxysilanes, such as TEOS or TMOS. There are varieties of surfactants with different sizes, shapes, functionalities, and charges. Examples include sulfates, sulfonates, phosphates, carbolyxic acid, alkyllammonium salts, gemini surfactant, cetylethypiperidinium, and bichain salts, primary amines, poly(oxyethylene oxides), and octaethylene glycol monohexadecyl ether.[23–27] Surfactant concentrations can be as low as 0.5 wt% and typically 15–30 wt%. After synthesis, the films are dried to polymerize the inorganic silica and remove the solvents. There are several methods to remove the template and these are solvent extraction, calcination, oxygen plasma treatment, and supercritical extraction.[22] Acidic ethanol has been used to extract the organic template. Chen, Li, and Davis also reportedly removed 100% of the organic material in MCM-41 using HCl at 70°C for about 30 hr.[28] Pure ethanol has also been used, but extraction is only 85%.[29] Calcination of MCM-41 has been done in flowing N_2, O_2, and air at 540°C.[30,31] In spin coating, a preferential evaporation of the solvent increases the surfactant and silica concentrations in the film and the surfactant start to self-assemble into micelles, while the silica begins to vitrify around the supramolecular structures.[32] The self-assembly during the formation of the film is a complex process governed by the interactions at the interphases between organic and inorganic phases in the films.[33] The self-assembly and silica formation mechanisms are said to occur simultaneously during the removal of alcohol from the thermodynamic equilibrium.[34]

ORGANOSILICATES, SSQs-BASED MATERIALS, AND THEIR INORGANIC–ORGANIC HYBRIDS

One type of material that is being developed for low-k applications is organosilicate film.[35–37] Basically, the structure of this material is composed of a silicon–oxygen backbone with methyl (CH_3) incorporation. The methyl groups make a robust structure leading to a less dense film and may result in a lower k. For PECVD organosilicate films, deposition is done using either SiH_4/hydrocarbon gas mixtures or liquid sources. Published reports using liquid sources as deposition precursors show films with low k, good gap-filling capability, and high thermal stability. In the study conducted by A. Nara and H. Itoh, tetramethylsilane [$Si(CH_3)_4$ (TMS)] was used as a liquid precursor with O_2 gas and the result is a film having a value of k below 3.0 after annealing at 500°C.[36] Uchida et al. produced a OH-free organic silica using tetra-isocyanate-silane, $Si(NCO)_4$; di-methyl silyl di-*iso*-cyanate silane, $(CH_3)_2Si(NCO)_4$; tri-methyl amine, $N(CH_3)_3$; and di-methyl ethyl amine, $N(CH_3)_2C_2H_5$.[38] Siloxane-based materials also discussed below are chemically stable against thermal decomposition and moisture absorption. One such material utilized as a precursor for PECVD is hexamethyldisiloxane [$(CH_3)_3SiOSi(CH_3)_3$], which can be mixed with O_2, CF_4, and toluene ($C_6H_5CH_3$).[39,40] The result is a film with a low dielectric constant and the value decreased with annealing and/or addition of flourine. Precursors like tetravinyltetramethylcyclotetrasiloxane [TVTMCTS (ATMI)] have been investigated by this group. The structure of the liquid is a cyclic Si–O with methyl and vinyl groups attached to Si. Some organosilicates (OSG) precursors enumerated by Laxman include methylsilane, $(CH_3)SiH_3$; dimethylsilane, $(CH_3)_2SiH_2$; trimethylsilane, $(CH_3)_3SiH$; TMS, $(CH_3)_4Si$; dimethyldimethoxysilane; and 1,3,5,7,-tetramethyltetrasiloxane.[39]

Some OSG films by PECVD available from several manufacturers are as follows:[41] 1) Applied Materials (Santa Clara, CA) —Black diamond I and II, $k = 2.4$–3.1; 2) Novellus Systems (San Jose, CA)—CORAL, $k = 2.2$–2.8; 3) Trikon Technologies (Newport, U.K.)—Flowfill CVD, $k = 2.8$; 4) ISMT PECVD solutions, $k = 2.2$; and 5) ASM International (ASMI, Bilthoven, Netherlands), $k = 2.7$.

For spin coating films:[41] 1) Dow Chemicals (Midland, MI)—SiLK, $k = 2.65$, Dow Corning (Midland, MI)—Fox HSQ, $k = 2.9$, XLK porous HSQ,

$k = 2.0$; 2) JSR—MSQ, $k = 2.7$, porous film, $k = 2.0$–2.2; 3) Air Products—MesoELK, $k = 2.2$; and 4) Honeywell Electronics materials (Sunnyvale, CA)—Nanoglass, $k = 2.0$.

Silsesquioxane refers to structures with an empirical formula of $RSiO_{3/2}$ where R is H or any alkyl, aryl, arylene, or organofunctional derivatives of alkyl, aryl, or arylene groups. There are different structures of SSQ and these are random, ladder, and cage structures, as shown in Fig. 6.[42] The ladder and cage structures are preferable because of the inherently open structure. This material is hydrophobic because of the methyl end groups, which are nonpolar. The fully cured SSQs have many good qualities for application to interlayer dielectric (ILD)-like excellent thermal stability greater than 400°C, k value equal to 2.85, low dielectric loss, good electrical properties, low moisture absorption, etc. Two SSQs that are used in microelectronics are hydrogen silsesquioxane (HSQ) and methylsilsesquioxane (MSSQ). Hydrogen silsesquioxane has H as the termination element with a dielectric constant of 3.0–3.2 and MSSQ has CH_3 end group with $k = 2.8$.[3]

To obtain an even lower k, porosity is introduced to MSSQ by incorporating preformed organic materials by "reactive blending."[43,44] This also utilizes the concept of sacrificial-porogen technique. It is an improvement of block copolymer, because silica-based matrix materials are used. Its advantages are compatibility with current fabrication processes and the ability to withstand rigorous requirements of the semiconductor industry. The process is to develop an inorganic–organic, nanophase-separated, hybrid material and subsequent removal of the organic material by thermolysis to produce nanoporous film. To form nanoscopic domains, preformed macromolecular porogens with controlled structure and architecture are used. The organic macromolecule must be tailored to have strong reactive sites capable of bonding to the silicate matrix, and hence, the concept of reactive blending. This chemical bonding and other matrix–template interactions like hydrogen bonding will significantly prevent the mobility of the organic phase until the polymer mobility is restricted by matrix vitrification. It also prevents macromolecule aggregation and hence, avoids macrophase separation. Thus, the sacrificial polymer must be chosen to have several end groups. The matrix must have low molecular weight and a lot of functional groups to have a low dielectric constant. The inorganic–organic precursors are dissolved in a common solvent, spin-coated, and then prebaked to remove the solvent. The inorganic–organic hybrid undergoes another annealing stage, where the matrix vitrifies without decomposition of the organic phase. Then, the sacrificial organic material is removed by thermal decomposition and volatilization, whereby the matrix further cross-links. Fig. 7 shows a schematic diagram of this technique.

IBM has prepared a number of architecturally defined star-like, dendrimeric, and hyperbranched polycaprolactones from functionalized core molecules as porogen macromolecules.[44] Short linear chains of polycaprolactone are attached to the chain ends of the starting dendrimers like dendritic polyesters derived from 2,2-bis(hydroxymethyl)propionate. The degree of polymerization of the linear blocks could be accurately controlled simply by varying the ratio of ε-caprolactone to the dendrimer. The number of branches could be varied from 6 to 48 by controlling the size or generation number of the dendrimer. The precondensed SSQ and the porogen are usually dissolved in a common solvent, such as propylene glycol monomethylether acetate (PM-acetate) and then spin coated. The matrix resin vitrifies at temperatures above 150°C, whereas the porogen start to decompose from 300°C. As stated above, it is important that the matrix stiffens first before decomposition and volatilization of

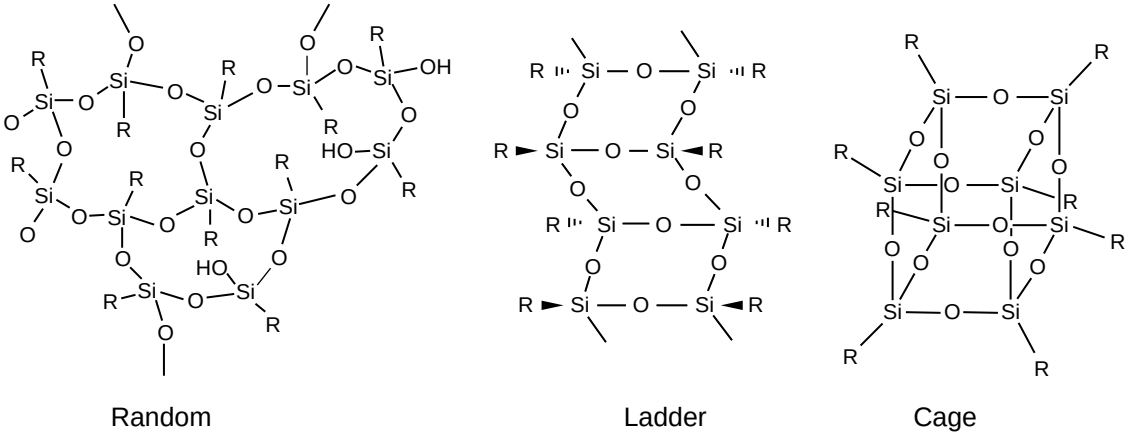


Fig. 6 Different structures of silsesquioxane.

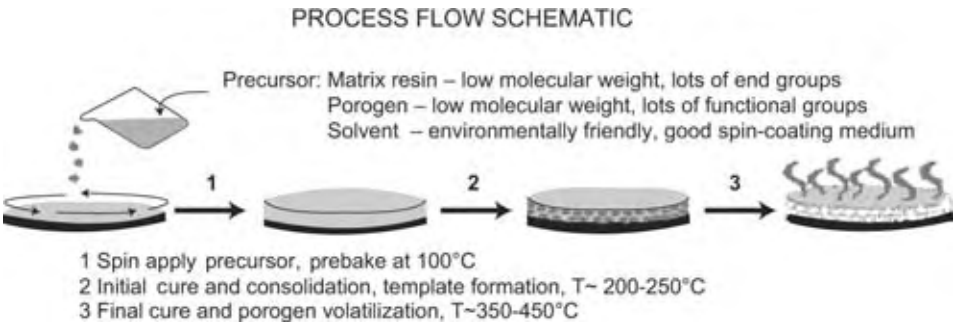


Fig. 7 Schematic diagram of pore generation in organic–inorganic hybrid. *(View this art in color at www.dekker.com.)*

the porogen. The amount of various components was varied to generate different porosities with decreasing k from 2.7 to 1.65, as the amount of dendrimers was increased from 0 to 40 wt%.[15] Rajagopalan et al. have shown removal of PPG pore generator in an MSSQ matrix via an alternative route to annealing, which is supercritical CO_2 ($SCCO_2$).[45] Here, the porogens are extracted from the matrix by the ability of $SCCO_2$ to penetrate the packed matrix structure, dissolve the porogens, and remove it from the matrix via depressurization. The extraction was found to be also effective in the production of inorganic–organic hybrid nanoporous material. Fig. 8 shows TEM micrographs of two porous films with open- and closed-pore morphologies processed by $SCCO_2$.[45] The pore distribution of these $SCCO_2$ processed films was shown in Fig. 9 and compared with the thermal removal of the porogens.

NANOPOROUS PLASMA-ENHANCED CVD FILMS VIA ANNEALING AND SUPERCRITICAL CO_2 TREATMENTS

Plasma-enhanced CVD has been the standard for dielectric depositions. Extending its use provides obvious advantage in terms of manufacturing costs. Researchers have shown that nanoporous PECVD films can be obtained by postdeposition treatments. In this method, a composite/dual-phased film consisting of a thermally stable matrix like silicon dioxide or silicon carbide and a less stable material like CH_x is prepared by PECVD.[46–48] In some cases, a non-cross-linked low -molecular weight CF species is incorporated in low-density organosilicate films. Postdeposition treatments of the films like annealing and alternatively applying supercritical CO_2 remove the less-stable materials (CH_x and CF) leaving behind a film with voids/pores and hence, a nanoporous structure.[46–49] The porosity can be tailored by the amount of CH_x or CF incorporation by adjusting precursor ratios and deposition conditions. Grill and Patel have prepared a dual-phased material of SiCOH by PECVD of tetramethylcyclotetrasiloxane (TMCTS) mixed with an organic precursor for the source of CH_x.[47] The TMCTS acts as the host matrix, whereas the organic precursor is the material to be removed. The films were annealed after deposition at 400°C to remove the sacrificial organic material. The results show that the dielectric constant decreases from 2.8 for the nonporous material to 2.05 for the nanoporous material. Measurements of positron annihilation

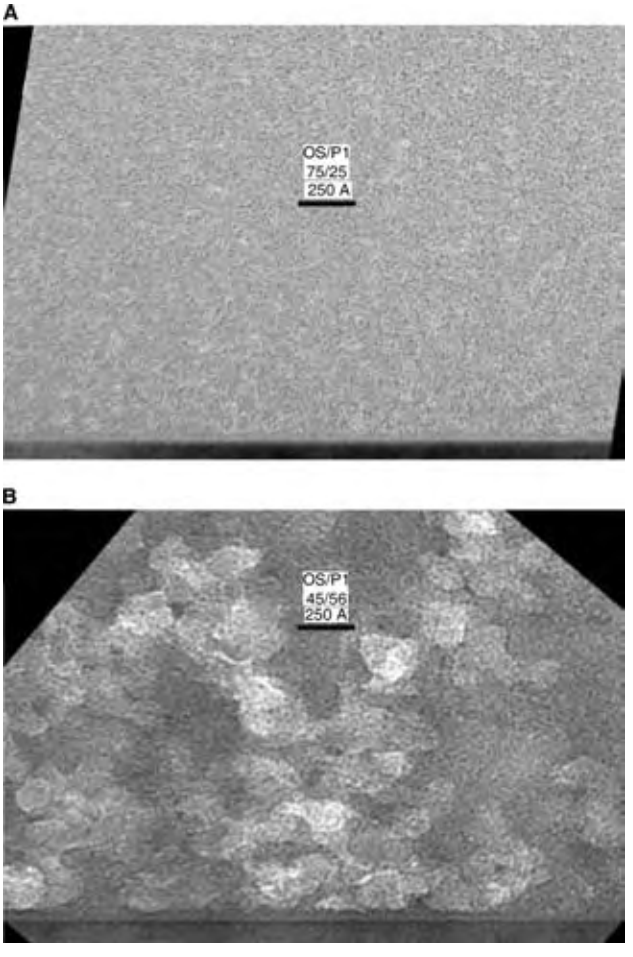


Fig. 8 Cross-sectional TEM images of nanoporous MSSQ films prepared by $SCCO_2$ treatment at 200°C for the closed-pore (A) and at 160°C for the open-pore (B) morphologies. (From Ref.[45].)

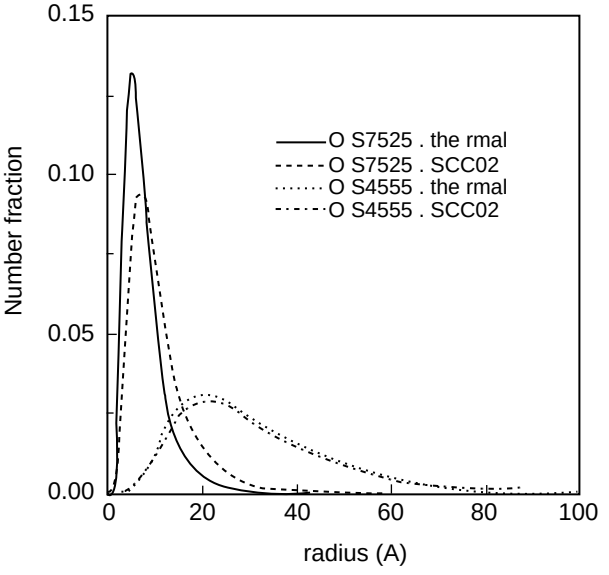


Fig. 9 Pore size distribution obtained from the best fits of SAXS measurements. (From Ref.[45].)

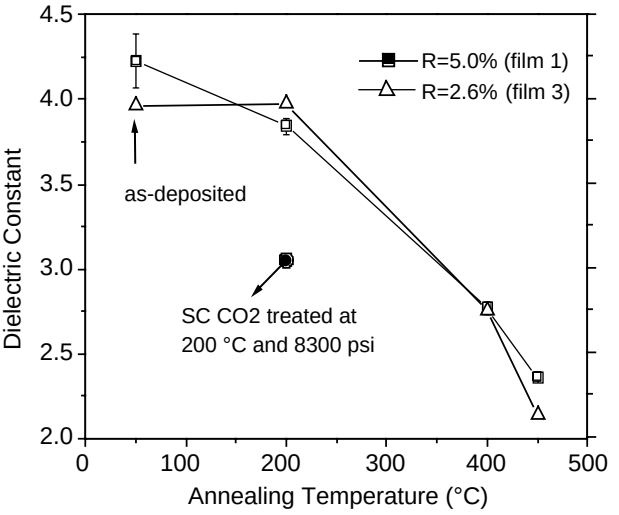


Fig. 11 Effect annealing temperature and $SCCO_2$ treatments on the dielectric constant of Films 1 and 3. *R* stands for gas flow ratio given by $\{[C_4H_{12}Si]/[(C_4H_{12}Si) + (CH_4)]\} \times 100\%$.

spectroscopy and ellipsometric porosimetry show that the annealed films show porosities up to 30–39% for films with *k*-values of 2.05. The pore sizes were measured to be less than 5 nm. A similar strategy is developed by Lahlouh et al.[48] where a-SiC : H films with lots of less stable CH_x and other species were deposited from a diethylsilane liquid precursor and CH_4. The films were deposited at the low radiofrequency power and deposition temperature to deposit a less stable material. The films were then annealed from 200°C to 450°C and also treated with $SCCO_2$ at 200°C. Results show that CH_x, species like the OH group, SiH_n, and C=O were removed by the treatments, possibly leaving voids/pores as shown in Fig. 10. The dielectric constant of the films decreased from 4.2 to 2.1 after 450°C annealing. $SCCO_2$ treatments at 200°C show a more effective extraction of the less stable species. The results are shown in Fig. 11. Lubguban et al. have deposited a nanocomposite film from TVTMCTS and C_4F_8 via PECVD.[49] It was shown that a pure CF film can be removed totally by $SCCO_2$ and pure TVTMCTS film cannot be removed. These facts are used to create a nanocomposite film with stable and less stable parts. The result is a film with a open and stable structure of TVTMCTS with a C : F species that can be removed via supercritical CO_2. Measurements of dielectric constants show decreases to about 10–14%.

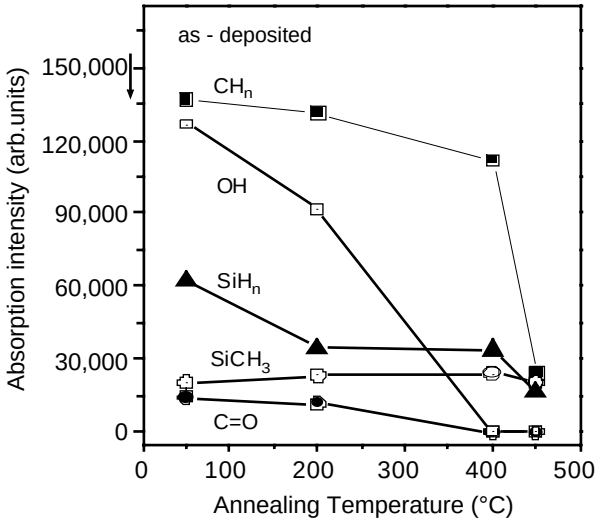


Fig. 10 Effect of the different annealing temperatures on the concentration of C–H_n, –OH, Si–H, Si–CH_3, and C=O bonds in the films.

CONCLUSIONS

In this review, nanoporous dielectric materials have been presented. The preparation techniques, methods of pore generation, and materials classification were discussed. Inorganic, organic, and hybrid films were also discussed. The nanoporous dielectric films were classified by preparation method, which includes block copolymer, solvent as porogen approach, surfactant templating, and sol–gel approach. This is still a very active field of research because no one film has satisfied all the stringent requirements in semiconductor device processing.

REFERENCES

1. Streetman, B.G.; Banerjee, S. *Solid State Electronics*, 5th Ed.; Prentice Hall: New Jersey, 2000; 422, Copyright Pearson Education, Inc.; Upper Saddle River, NJ.
2. Miller, R.D.; Volksen, W.; Hedrick, J.L.; Hawker, C.J.; Remenar, J.F.; Furuta, P.; Nguyen, C.V.; Yoon, D.; Toney, M.; Rice, D.P.; Hay, J. Conference Proceedings ULSI XV, Materials Research Society, c. 2000; 327–333.
3. Maex, K.; Baklanov, M.R.; Shamiryan, D.; Iacopi, F.; Brongersma, S.H.; Yanovitskaya, Z.S.J. Low dielectric constant materials for microelectronics. Appl. Phys. **2003**, *93*, 8793–8841.
4. Lee, W.W.; Ho, P.S. Low-dielectric constant materials for ULSI interlayer-dielectric applications. MRS Bull. **1997**, *22*, (Oct), 19–23.
5. *International Techology Roadmap for Semiconductors*; 2001 Ed.
6. Golden, J.H.; Hawker, C.J.; Ho, P.S. Designing porous low-k dielectrics. Semiconductor Int. **2001**, *24*, (May), 79.
7. Polarz, S; Smarsely, B. Nanoporous materials J. Nanosci. Nanotech. **2002**, *2* (6), 581.
8. Gregg, S.J.; Sing, K.S. *Adsorption, Surface Area, and Porosity*, 2nd Ed.; Academic Press: London, 1982.
9. Barton, T.J.; Bull, L.M.; Klemperer, W.G.; Loy, D.A.; McEnaney, B.; Misono, M.; Monson, P.A.; Pez, G.; Scherer, G.W.; Vertuli, J.C.; Yaghi, O.M. Tailored porous materials. Chem. Mater. **1999**, *11*, 2633–2656.
10. Sing, K.S.W.; Everett, D.H.; Haul, R.A.W.; Moscou, L.; Pierotti, R.A.; Rouquerol, J.; Siemieniewski, T. Pure Appl. Chem. **1985**, *57*, 603.
11. McEnaney, B.; Mays, T.J. *Porosity in Carbons: Characterization and Applications*; Patrick, J.W., Ed.; Edward Arnold, Inc: London, 1995; 93 pp.
12. *International Technology Roadmap for Semiconductors*; 2003 Ed.
13. Ting, C.H.; Seidel, T.E. *Low-Dielectric Constant Materials-Synthesis and Applications in Microelectronics*; Lu, T.M., Murarka, S.P., Kuan, T.S., Ting, C.H., Eds.; Materials Research Society Symposium Proceedings 381, Pittsburgh, PA, 1995; 3 pp, Copyright the Materials Research Society.
14. Hedrick, J.L.; Hofer, D.C.; Labadie, J.W.; Prime, R.B.; Paul, T.P. U.S. patent 5,776,990, 1998.
15. Hawker, C.J.; Hedrick, J.L.; Miller, R.D.; Volksen, W. Supramolecular approaches to nanoscale dielectric foams for advanced microelectronic devices. MRS Bull. **2000**, *25*, (Apr), 54, Copyright the MRS Bulletin.
16. Hedrick, J.L.; Labadie, J.W.; Russell, T.P.; Hofer, D.; Wakharher, V. High temperature polymer foams. Polymer **1993**, *34*, 4717.
17. Jin, C.; Wetzel, J. Characterization and integration of porous extra low-k (XLK) dielectrics. Proc. IEEE **2000**, 99.
18. O'Neill, M.L.; Robeson, L.M.; Burgoyne, W.F.; Langsam, M. Nanoporous Polymer Films for Extreme Low and Interlayer Dielectrics. . U.S. Patent 6,187,248, 2001.
19. Brinker, C.J.; Hurd, A.J.; Schunk, P.R.; Fyre, G.C.; Ashley, C.S.J. Non-cryst. Solids **1992**, *147&148*, 424–436.
20. http://www.psrc.usm.edu/mauritz/solgel.html.
21. Jin, C.; Luttmer, J.D.; Smith, D.M.; Ramos, T.A. Nanoporous silica as ultralow-*k* dielectric. MRS Bull. **1997**, *22*, (Oct), 39–42.
22. Raman, N.K.; Anderson, M.T.; Brinker, C.J. Template-based approaches to the preparation of amorphous, nanoporous silicas. Chem. Mater. **1996**, *8*, 1682–1701, Copyright the American Chemical Society.
23. Huo, Q.; Margolese, D.I.; Ciesla, U.; Demuth, D.G.; Feng, P.; Gier, T.E.; Sieger, P.; Firouzi, A.; Chlemka, B.F.; Schuth, F.; Stucky, G.D. Organization of organic molecules with inorganic molecular species into nanocomposite biphase arrays. Chem. Mater. **1994**, *6*, 1176–1191.
24. Antonelli, D.M.; Ying, J.Y. Synthesis of hexagonally packed mesoporous TiO_2 by a modified sol-gel method. Angew. Chem. Int. Ed. Engl. **1995**, *34*, 2014–2017.
25. Tanev, P.T.; Pinnavaia, T.J. A neutral templating route to mesoporous molecular-sieves. Science **1995**, *267*, 865–867.
26. Bagshaw, S.A.; Prouzet, E.; Pinavaia, T.J. Templating of mesoporous molecular-sieves by nonionic polyethylene oxide surfactants. Science **1995**, *269*, 1242–1244.
27. Attard, G.S.; Glyde, J.C.; Goltner, C.G. Liquid crystalline phases as templates for the synthesis of mesoporous silica. Nature **1995**, *378*, 366–368.
28. Chen, C.Y.; Li, H.X.; Davis, M.E. Studies on mesoporous materials I. Synthesis and characterization of MCM-41. Micropor. Mater. **1993**, *2*, 17–26.
29. Luca, V.; Maclachlan, D.J.; Hook, J.M.; Withers, R. Synthesis and characterization of mesostructured vanadium oxide. Chem. Mater. **1995**, *7*, 2220–2223.
30. Kresge, C.T.; Leonowicz, M.E.; Roth, W.J.; Vartulli, J.C.; Beck, J.S. Ordered mesoporous molecular sieves synthesized by a liquid-crystal template mechanism. Nature **1992**, *359*, 710.
31. Beck, J.S.; Vartulli, J.C.; Roth, W.J.; Leonowicz, M.E.; Kresge, C.T.; Schmitt, K.D.; Chu, C.T.-W.; Olson, K.H.; Sheppard, E.W.; McCullen,

S.B.; Higgins, J.B.; Schlenker, J.L. A new family of mesoporous molecular sieves prepared with liquid crystal templates. J. Am. Chem. Soc. **1992**, *114*, 10834–10843.
32. Brinker, C.J.; Anderson, M.T.; Ganguli, R.; Lu, Y. Process to Form Mesostructured Films. U.S. Patent 5,858,457, 1999.
33. Grosso, D.; Babonneau, F.; Sanchez, C.; Soler-Illia, G.J.; de, A.A.; Crepalde, E.L.; Albouy, P.A.; Amenitsch, H.; Balkenende, A.R.; Brunet-Bruneau, A. Self assembly of 2D-hexagonal templated TiO2 mesostructured films during dip coating. In *Materials and Surfaces*: *Structural Studies*; Research Highlights, 2003; 36-39.
34. Ogawa, M.; Kikuchi, T. Preparation of self-standing transparent films of silica-surfactant mesostructured materials and the conversion to porous silica films. Adv. Mater. **1998**, *10* (14), 1077.
35. Nakano, T.; Tokunaga, K.; Ohta, T.J. Effects of Si-C bond content on film properties of organic spin-on glass. Electrochem. Soc. **1995**, *142*, 1303.
36. Nara, A.; Itoh, H. Low dielectric constant insulator formed by downstream plasma CVD at room temperature using TMS/O_2. Jpn. J. Appl. Phys. Pt. 1 **1997**, *36*, 1477.
37. Uchida, Y.; Taguchi, K.; Nagai, T.; Sugahara, S.; Matsumura, M. Chemical-vapor deposition of OH-free and low-k organic-silica films. Jpn. J. Appl. Phys. Pt. 1 **1998**, *37*, 6369.
38. Uchida, Y.; Taguchi, K.; Nagai, T.; Sugahara, S.; Matsumura, M. A fluorinated organic-silica film with extremely low dielectric constant. Jpn. J. Appl. Phys. Pt. 1 **1999**, *38*, 2368.
39. Laxman, R.K.; Arkles, B.; Tabler, T.A. Synthesizing low-k CVD materials for fab use. Semi conductor Int. **2000**, Nov.
40. Fujii, T.; Hiramatsu, M.; Nawata, M. Thin Solid Films **1999**, *457*, 343–344.
41. Peters, L. Industry divides on low-*k* dielectrics choices. Semiconductor Int. **2001**, May.
42. Baney, R.H.; Itoh, M.; Sakakibara, A.; Suzuki, T. Silsesquioxane. Chem. Rev. **1995**, *95* (5), 1426.
43. Miller, R.D.; Hedrick, J.L.; Yoon, D.Y.; Cook, R.F.; Hummel, J.P. Phase-separated inorganic-organic hybrid for microelectronic applications. MRS Bull. **1997**, *22*, (Oct), 41.
44. Miller, R.D.; Volksen, W.; Hedrick, J.L.; Hawker, C.J.; Remenar, J.F.; Furuta, P.; Nguyen, C.V.; Yoon, D.; Toney, M.; Rice, D.P.; Hay, J. Sacrificial macromolecular porogens: a route to porous organosilicates for on-chip insulator applications. Conference Proceedings ULSI XV, Materials Research Society **2000**, 327–332.
45. Rajagopalan, T.; Lahlouh, B.; Lubguban, J.A.; Biswas, N.; Gangopadhyay, S.; Sun, J.; Huang, D.H.; Simon, S.L.; Mallikarjunan, A.; Kim, H.-C.; Volksen, W.; Toney, M.F.; Huang, E.; Rice, P.M.; Delenia, E.; Miller, R.D. Supercritical carbon dioxide extraction of porogens for the preparation of ultralow-dielectric-constant films. Appl. Phys. Lett. **2003**, *82*, 4328, Copyright the American Institute of Physics.
46. Grill, A.; Patel, V. Ultralow-*k* dielectrics prepared by plasma-enhanced chemical vapor deposition. Appl. Phys. Lett. **2001**, *79* (6), 803.
47. Grill, A.; Patel, V.; Rodbell, K.P.; Huang, E.; Baklanov, M.R.; Mogilnikov, K.P.; Toney, M.; Kim, H.-C. Porosity in plasma enhanced chemical vapor deposited SiCOH dielectrics: a comparative study. J. Appl. Phys. **2003**, *94* (5), 3427.
48. Lahlouh, B.; Lubguban, J.A.; Biswas, N.; Rajagopalan, T.; Mehta, N.; Sun, J.; Huang, D.; Simon, S.L.; Gangopadhyay, S. Post treatments of plasma-enhanced chemical vapor deposited a-SIC:H films for low-k dielectrics, accepted for publication in Thin Solid Films.
49. Lubguban, J.A.; Rajagopalan, T.; Lahlouh, B.; Simon, S.L.; Gangopadhyay, S. Supercritical carbon dioxide extraction to produce low-*k* plasma enhanced chemical vapor deposited dielectric films. Appl. Phys. Lett. **2002**, *81* (23), 4407.

Nanostructured Materials

Vikrant N. Urade
Hugh W. Hillhouse
Purdue University, West Lafayette, Indiana, U.S.A.

INTRODUCTION

Profound changes may occur in the electronic properties of materials as their characteristic length scale is reduced to the nanoscale. These changes affect the optical properties, mechanical response, adsorption behavior, and catalytic properties of the material. The possibility of manipulating the structure of materials at the nanometer length scale to alter these properties has opened up an array of opportunities for the materials science, chemistry, and the engineering communities. Synergistic efforts by researchers in these areas have led to the discovery of a large number of new nanostructured materials. Many of these materials have promising applications in heterogeneous catalysts, molecular-sieve adsorbents, sensors, hydrogen and methane storage materials, solar energy conversion, thermoelectric devices, and magnetic data storage. However, in many cases, these properties are sensitively dependent on the processing conditions.

Because of the explosion in new materials, we introduce a nomenclature scheme to help clarify the field. In general, we define a nanomaterial as any material with a characteristic length scale that is less than 100 nm. This characteristic length scale may be the overall size of the sample (such as a 10 nm diameter particle), or it may be an internal length scale over which the structure varies. The realm of nanomaterials may then be divided into four subcategories depending on the dimensionality of the material (Fig. 1). Thus we reserve the term ''nanostructured materials'' for the materials that are macroscopic in three dimensions but contain internal structure, on the nanometer length scale, of less than 100 nm but greater than 1 nm (this excludes normal crystal structures such as fcc, rocksalt, zinc blende, etc.). This classification thus encompasses an array of materials including zeolites, self-assembled inorganic materials such as mesoporous metal oxides, and self-assembled block-copolymers. If the material is porous and has pores of diameter less than 100 nm, then we classify the material as nanoporous. Historically, porous materials have been classified by the IUPAC as being microporous if the pore dimensions are below 2 nm, mesoporous if they are between 2 and 50 nm, and macroporous if the pore dimensions exceed 50 nm. While the choice of the term ''microporous'' for a material with pores less than 2 nm is unfortunate, this terminology is still heavily used.

There have been many breakthroughs in the synthesis of nanostructured materials in the last 10–15 years, including the synthesis of large pore zeolites with pores larger than 1 nm,[1,2] the discovery of surfactant templated mesoporous materials by the scientists at Mobil,[3] the extension of this templating method to nonsilica systems including transition metal oxides, metals, and carbons,[4–6] and the synthesis of nanoporous metal organic frameworks (MOFs).[7] This entry focuses on the synthesis strategies and the mechanisms involved in the synthesis of nanostructured materials and the processing parameters that are vital for the successful synthesis of these nanostructures. Special emphasis is placed on the synthesis of materials that are closer to applications and are therefore of more immediate technological importance to the process engineer. Further, within the branch of nanostructured materials, materials may be further categorized depending on the synthesis procedures, as illustrated in Fig. 1. This entry also focuses on templated nanostructured materials assembled by the bottom-up approach.

GENERAL STRATEGIES TO SYNTHESIZE NANOSTRUCTURED MATERIALS FROM THE BOTTOM-UP APPROACH

Nanostructured materials can be synthesized from the so-called ''top down'' or ''bottom-up'' approach. In the first approach, features at the micron (or submicron) length scale are created on a substrate by masking and exposing selected regions of a radiation sensitive layer (typically a polymeric photoresist) to a UV source. This exposure is followed by various chemical treatments and mechanical steps to obtain the desired spatial pattern on a substrate. However, the feature sizes that can be obtained with this approach are limited to the length scale of the wavelength of the radiation employed. If features at the nanometer scale are desired, one must start from the bottom (i.e., use individual molecules or clusters) and assemble templates that will impart the nanostructure to the desired material.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120018074

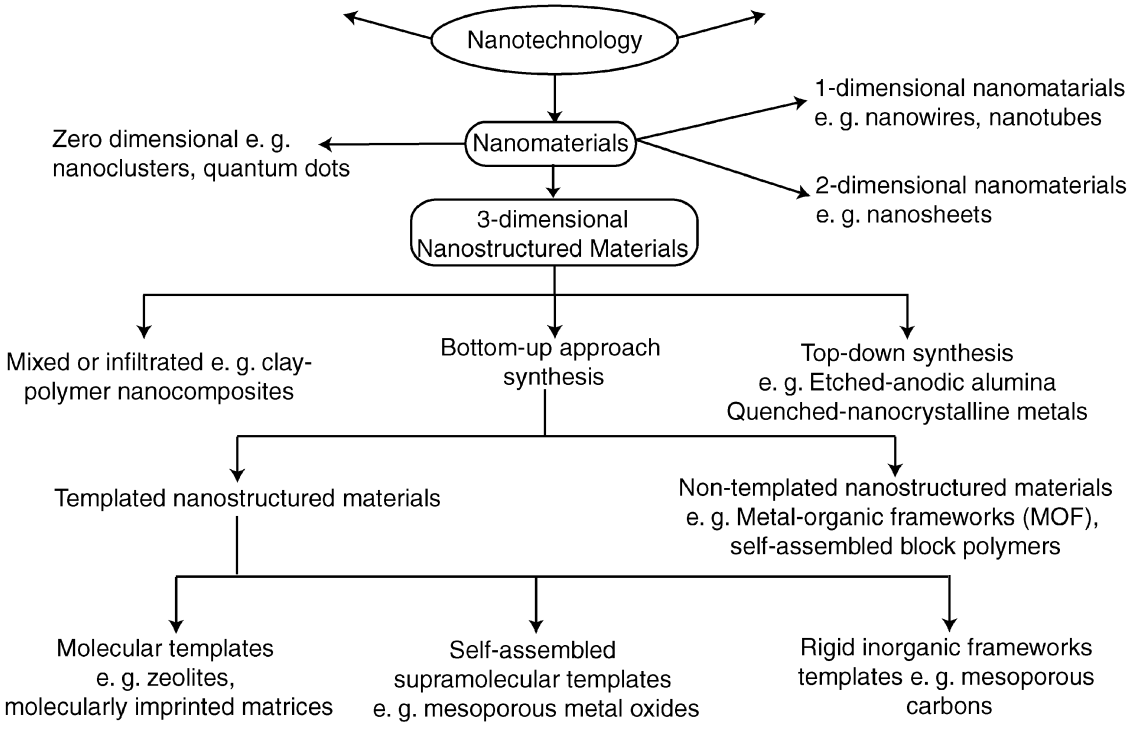


Fig. 1 A schematic illustrating various classes of nanostructured materials based on the synthesis strategies.

Templating Route to Synthesize Nanostructured Materials

The general strategy for the synthesis of nanostructured materials involves the use of templates, which can be ionic, molecular, or supramolecular structures and act as molds. These molds guide the formation of the structure and leave behind porosity when removed by suitable means (if desired). The templating approach allows control over the size and the shape of the resulting pores. In a typical synthesis of a nanoporous metal oxide material, a suitable template molecule (typically an amphiphile) is added to the synthesis mixture. The synthesis mixture contains inorganic precursors that are capable of interacting with the template molecule either via electrostatic or entropic pathways to create an ordered nanostructured assembly of the template and the inorganic material. The composition of the mixture is adjusted such that the desired material self-assembles and phase separates from the mixture. At this stage, the template molecule is occluded in the inorganic framework, directing the nanoscale structure. The removal of the template creates ordered voids in the inorganic framework, whose size and topology are determined by the template used. Template removal is essential for applications of these materials as catalyst supports, adsorbents, or molecular sieves. The inorganic framework must be structurally rigid and capable of supporting itself in the absence of the template.

Types of Templates

Depending on the size and the shape of the pore system desired, the nature of template molecules also varies. Microporous materials are typically synthesized using individual molecules or ions as templates. In this case, these molecules may not constitute the entire template, but may be combined with solvent or water molecules to form the templating species. When larger pore sizes are desired, the templates are supramolecular assemblies of surfactants arranged in various configurations depending on the synthesis parameters. It is also possible to use other porous inorganic materials as templates and create new porous materials inside them. These templates are termed as ''rigid templates,'' as opposed to ''soft'' surfactant templates, which are labile and undergo phase transition in the solution. The words ''endotemplating'' and ''exotemplating'' are also used to indicate the use of surfactant micelles and inorganic porous solids as templates, respectively.

USE OF SINGLE MOLECULES AS TEMPLATES TO SYNTHESIZE MICROPOROUS MATERIALS

The use of single molecules or ions as templates results in materials with the smallest pore sizes (typically less than 1 nm), and hence they are called microporous according to the IUPAC definitions. Most microporous materials synthesized to date are zeolites.

Zeolites are a subclass of microporous materials in which the crystalline inorganic framework is composed of four-coordinated species interconnected by two-coordinated species. Traditionally these materials are aluminosilicates; however, many different compositions have been synthesized.[8] The templates used in the synthesis of microporous materials are typically small ionic or neutral molecular species. The function of the template in the synthesis of microporous materials is little understood, and there are at least four different modes by which an additive can operate in a zeolite synthesis:[9] a) It may act as a space filler occupying the voids in the structure, thereby energetically stabilizing less dense inorganic framework; b) the additive may control the equilibria in the synthesis mixture, such as solution pH or complexation equilibria; c) it may preorganize the solution species to favor the nucleation of a specific structure; d) it may act as a true template determining the size and the shape of the voids in the structure.

Noncrystalline microporous materials may also be synthesized by a technique called molecular imprinting. In this process, a molecular template, called a ''print molecule,'' is used to direct the arrangement of functional monomers around the template, which can then be fixed by chemical polymerization with a crosslinking monomer. This results in the formation of a rigid matrix with the template embedded in it. The removal of the template exposes the functional sites, which can specifically recognize the print molecule or molecules similar to it. This technique can be effectively used to synthesize catalytic and enzymatic hosts having specific interactions with a particular kind of molecule and can be used for separations as well. For a review of molecular imprinting, the reader is referred to Refs.[10–12].

USE OF SELF-ASSEMBLED ARRAYS OF MOLECULES AS TEMPLATES FOR MESOPOROUS MATERIALS

The pore sizes obtained using molecular or ionic templates are restricted because of the small sizes of the templates used (typically less than 15 Å), restricting the accessibility of the internal surface to small molecules. Hence, a great deal of effort has been devoted to the synthesis of materials having pore sizes larger than those in zeolites. In the last decade, progress toward this goal has been made following the discovery of surfactant templated silica by the scientists at Mobil.[3] A variety of materials (silica and nonsilica oxides, carbons, metals, chalcogenides, phosphates, borates, and sulfides) have been synthesized in highly ordered forms having extremely high surface areas ($>1000\,m^2/g$) and narrow pore size distributions, with the pore size tunable all the way from 2 to 30 nm by varying the size of the surfactant and the synthesis conditions. These materials, with pore diameters between 2 and 50 nm, are called mesoporous or mesostructured materials, depending on whether the template has been removed or not. The most well known of these materials is the Mobil family of mesoporous materials (M41S), including a two-dimensional hexagonal mesoporous silica (MCM-41) and a cubic phase mesoporous silica (MCM-48).

The templates used in the synthesis of mesostructured and mesoporous materials can be classified into two categories: The first class of templates includes ''soft'' templates, which are ordered arrays of self-assembled surfactant micelles, similar to the ones used by the researchers at Mobil. Alternately, mesoporous materials can themselves be used as templates to synthesize new mesostructured materials, and such templates can be termed as ''rigid'' templates. In the following sections, we focus on the use of supramolecular assemblies of surfactants as well as on the use of rigid templates as the templates for the synthesis of mesostructured materials.

Self-Assembled Surfactant Templates

Surfactants are amphiphiles and as such have both hydrophilic and hydrophobic portions. If the contrast in the hydrophobicity of the hydrophilic and hydrophobic moieties is strong, then above a certain concentration (critical micelle concentration, cmc) these molecules may self-assemble in solution to form aggregates called micelles. If the concentration is increased further, the micelles may themselves self-assemble into a tertiary structure and form a lyotropic liquid crystalline phase to minimize the free energy of the system. There are three important parameters that characterize the self-assembly process of the surfactant: cmc, aggregation number, and molecular packing factor. Below the cmc, surfactant molecules predominantly exist as monomers in solution. The cmc is experimentally identified as the concentration at which added surfactant starts to enter into an aggregate.[13] The aggregation number is the dominant number of monomeric units in an aggregate. The geometry and the topology of the assembly are affected by the aggregation number, the geometry of the surfactant molecule, concentration of the surfactant, temperature, and the presence of other species in solution. For the effect of the aggregation number on the microstructure of some block copolymers that are commonly used as templating agents, the reader is referred to Ref.[14]. The effect of surfactant geometry can be qualitatively, and in many cases quantitatively, described by making use of the concept of molecular packing factor, g.

Surfactant molecules aggregate into different shapes in solution (spherical, cylindrical, lamellar, spherical bilayer, vesicular, etc.), which result in minimum interaction between the hydrophobic moieties of the surfactant and the polar solvent. The concept of molecular packing factor is useful in visualizing and qualitatively predicting the geometry of the micelles from the known geometry of individual surfactant molecules. Israelachvili et al.[15] proposed the concept of molecular packing parameter and demonstrated how the size and the shape of the aggregate at equilibrium can be predicted from a combination of molecular packing considerations and general thermodynamic principles. The molecular packing factor is defined as:

$$g = \frac{V_0}{a l_0}$$

where V_0 is the volume of the surfactant tail, l_0 the critical length of the tail (not the actual length of a fully extended tail), and a the effective area of the head group at the surface of the micelle. Thus, low values of g result in structures of high curvature, while a value of $g = 1$ has no net curvature. Table 1 summarizes the range of values of the packing factor for which various structures are obtained. The key point to keep in mind is that solution conditions can drastically affect the effective area of the head group. For instance, highly charged metal oxide ions can effectively screen the charge of adjacent head groups, reducing their effective area, and thus increase the packing parameter.

Use of Functional Dendrimers for the Self-Assembly of Nanobuilding Blocks

Chemically well-defined inorganic clusters (often termed as ''nanobuilding blocks'') can be assembled into an ordered array using dendrimers capped with functional groups that can chemically bond with the inorganic clusters. Various metal oxo-based hybrid materials, including Ti- and Ce-based gels, have been synthesized using this synthetic strategy.[16,17]

Table 1 Molecular packing factor values corresponding to various micellar structures

Packing factor, g	Structure obtained
1/3	Spherical
1/2	Infinite cylinders
1	Planar bilayers

Between $g = 1/3$ and $g = 1/2$ or between $g = 1/2$ and $g = 1$, complex phenomena such as formation of aggregates of lower or higher symmetry or phase separation may occur.

Synthesis Strategies

Self-assembled mesostructured materials are typically synthesized from dilute solutions of inorganic precursors and surfactant molecules wherein the self-assembly process occurs through electrostatic and entropic interactions and is a cooperative one. However, in some cases, mesostructured materials are synthesized at much higher surfactant concentrations such that a liquid crystalline phase actually pre-exists in solution, and the inorganic precursor is preferentially accommodated in the hydrophilic domains. This approach, called ''true liquid crystal templating'' (TLCT, route II in Fig. 2), was used by Attard et al. for the synthesis of mesoporous silica, metals, and alloys.[18,19] In addition, there have been several novel synthesis routes developed that utilize inorganic precursors with covalently bonded hydrophobic ligands or take advantage of specific solution interactions such as hydrogen bonding.

Cooperative self-assembly

After the addition of the inorganic precursor, cooperative assembly between the surfactant and the inorganic precursor takes place, leading to the formation of the mesostructure. This approach is similar to the one used by the Mobil scientists and is also the most commonly used one for the synthesis of nonsilica systems, where the control of precursor reactivity demands lower concentrations of the inorganic precursors and hence of the surfactant. In Fig. 2, the formation of a mesostructure by this approach is illustrated (route I).

To help explain this mechanism of the self-assembly process, Stucky and coworkers[20,21] proposed a very useful concept of charge density matching. This concept refers simply to the fact that electroneutrality is maintained locally at the interfaces. At the isoelectric point (IEP) of the inorganic material, the inorganic framework is charge neutral. When the pH is below the IEP, the framework carries a positive charge, and when it is greater, it is negatively charged. The electrostatic interactions between the charged inorganic moieties (I^- or I^+ depending on pH) and the surfactant head groups (S^+ and S^- depending on the type of surfactant, cationic or anionic) lead to the formation of a hybrid interface. The interaction can be direct (S^+I^- or S^-I^+) if the surfactant head groups and inorganic moieties carry opposite charges or they can be mediated by counterions associated with the surfactant ($S^+H^-I^+$ or $S^-H^+I^-$). The shape of the resultant hybrid structure is determined by the balance of charge density between the inorganic moieties and the surfactant head groups. If the charge density of the inorganic framework is reduced, the average head group area of the surfactant assembly increases and the packing

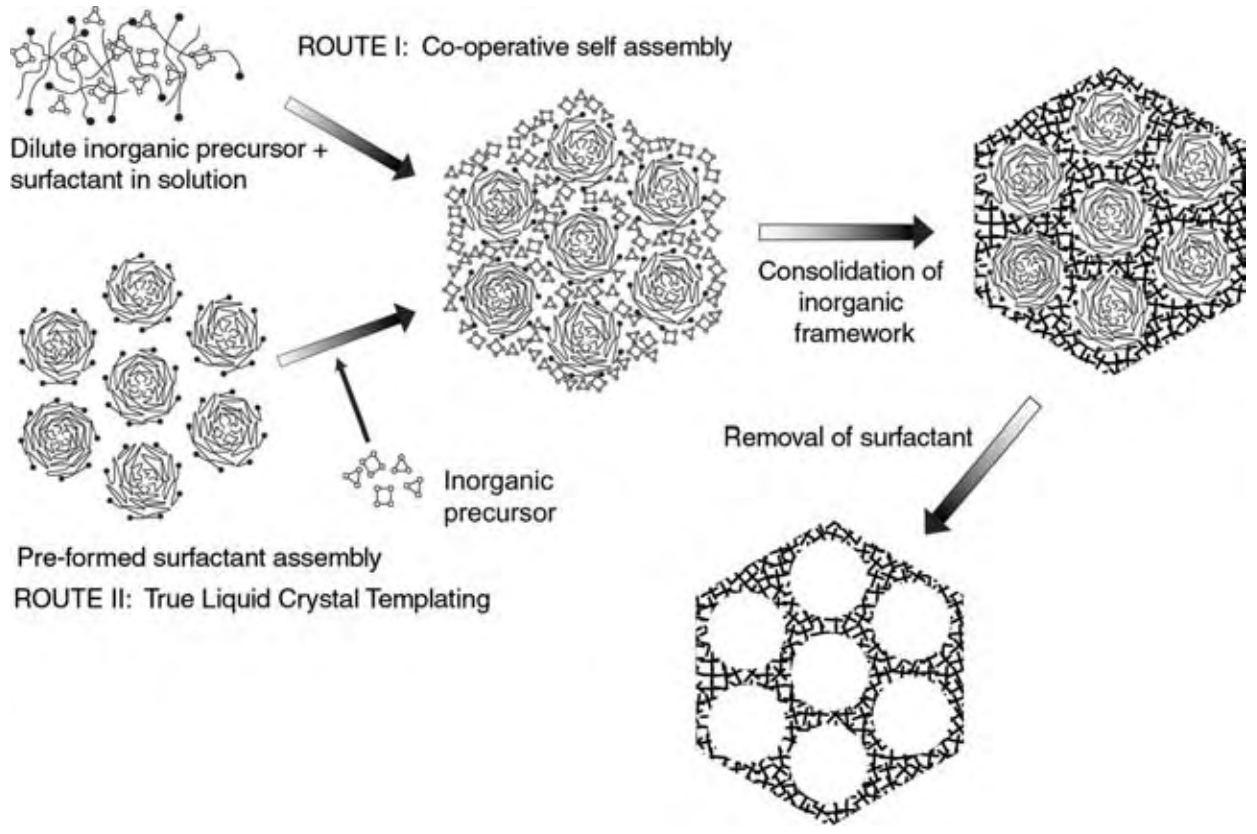


Fig. 2 A schematic illustrating various routes used to synthesize mesostructured materials. Route I: A cooperative self-assembly route relying on the interaction between the surfactant molecules and the inorganic precursors. Route II: TLCT route. The inorganic framework condenses around preformed surfactant micelles in this case.

factor g decreases, resulting in the formation of more curved structures.

In addition to using charged surfactants, Tanev and Pinnavaia[22] synthesized mesoporous silica by utilizing the hydrogen bonding interactions between the head group of an alkyl amine surfactant (S^0) and the hydroxylated silica precursor, tetra ethyl ortho silicate (I^0). The bonding between the amine head group and the hydroxylated inorganic precursor occurs through the exchange of the lone pair of electrons on nitrogen (Fig. 3). This work led to a very useful development reported by Zhao et al.[23] in which nonionic amphiphilic triblock-copolymers were used as templates. These commercially available templates are quite robust in the synthesis of mesoporous materials and produce thicker walled materials than the cationic surfactants. One unique feature of the triblock-copolymer templated materials is that many structures have micropores in the walls that interconnect the mesoporosity.

Ligand-assisted templating

As opposed to the liquid crystal or electrostatic templating mechanisms, which rely on coulombic, van der Waals, or hydrogen bonding interactions between the surfactant molecules and the inorganic species, Antonelli et al.[24] proposed a ligand assisted templating mechanism, which relies on the formation of a covalent bond between the surfactant head group and the inorganic moieties. The amine surfactants were pretreated with metal alkoxides in the absence of water to form metal-ligated surfactants. Addition of water, which acts as both a solvent and a reactant, to the alkoxide–surfactant solution initiates the self-assembly and alkoxide hydrolysis–condensation. The reaction scheme following this route is given in Fig. 4.

Processing of Mesostructured Materials: Effect of Synthesis Parameters

The formation of self-assembled materials is governed by a number of experimental parameters: the choice of inorganic precursors and surfactants, the inorganic to surfactant ratio, amount of water and other solvents used, pH during synthesis, additives, reaction time and temperature, treatments used to stiffen the inorganic framework, and treatments to remove surfactant and obtain porosity, all decide the final ordering, porosity and surface area. It is well known that successful

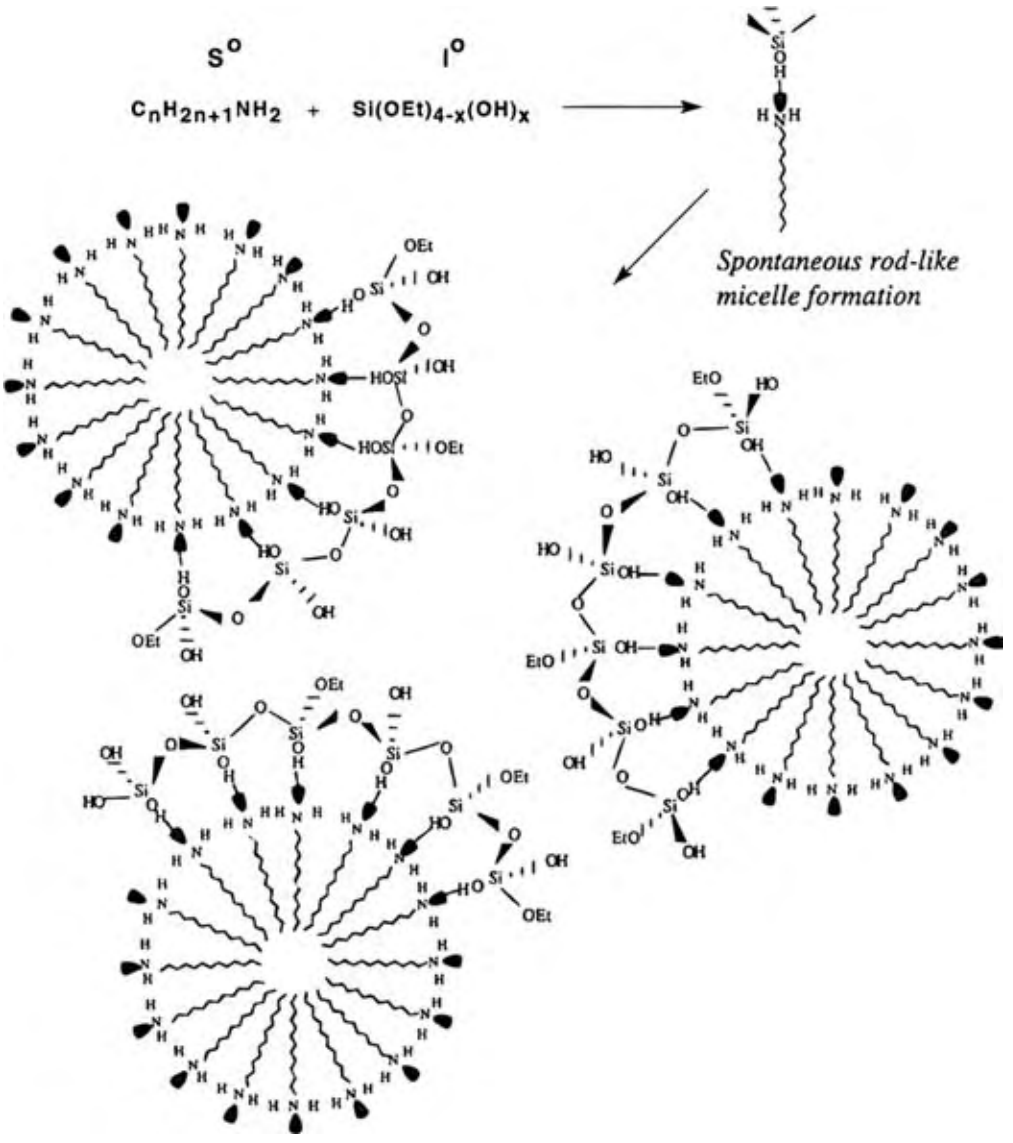


Fig. 3 The neutral templating route proposed by Pinnavaia et al. The electron lone pairs indicated by the shaded lobes on the surfactant head groups participate in the hydrogen bonding process with the framework silanol (Si–OH) groups. (From Ref.[22].)

synthesis depends on tuning the rates of the reactions that lead to the formation of inorganic framework with that of the self-assembly of the surfactant micelles.

Control of reactivity of inorganic precursor

Silica is the most extensively studied system among the self-assembled materials. Commonly used precursors in the synthesis of silica are tetraalkyl orthosilicates ($Si(OR)_4$). However, in nonsilica systems, including transition metal oxides or chalcogenides, the precursors are typically very reactive, and hence some means must be devised to control the reactivity of these precursors to avoid outright precipitation of solids having no mesostructure. The ways to control the reactivity of the inorganic precursors are summarized here. These are specific mostly to metal oxide systems, as they are the most extensively studied single class of mesostructured materials and are well characterized.

Addition of Complexing Agents. Complexation of metallic centers in the precursors by various chelating agents such as acetylacetone, ethylene glycol, and triethanolamine allows the control of the hydrolysis and condensation reactions of the inorganic precursors in the synthesis of metal oxides. Alternately, surfactants with a complexing ability can also be used.

Synthesis Under Low pH Conditions. The reactivity of transition metal alkoxides $M(OR)_n$ can be controlled

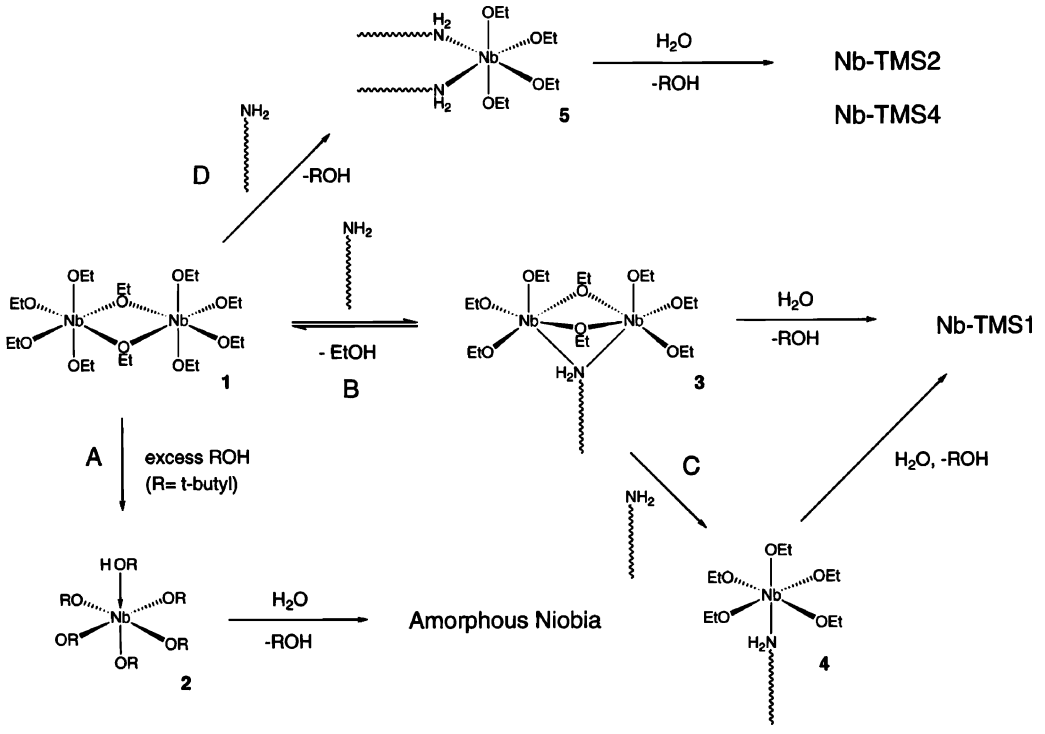


Fig. 4 Ligand assisted templating mechanism for the synthesis of niobium oxide. The amine surfactant forms covalent bonds with the niobium (V) ethoxide, resulting in the formation of a hybrid interface. Upon exposure to water, the system organizes into various structures such as bidimensional hexagonal (Nb-TMS1), three-dimensional hexagonal (Nb-TMS2), lamellar (Nb-TMS4), or a cubic phase. (Nb-TMS3 is not shown in the figure.) (From Ref.[24].)

by carrying out syntheses at very low pH. The protonation of M–OH groups by the excess H^+ ions in the solution, as well as increased rates of the depolymerization reaction, leads to the formation of small hydrophilic inorganic species almost all uncondensed, which can interact better with the polar head group of the surfactant.

Use of Mixed Precursors. When the framework of mesostructured materials consists of mixed materials (mixed metal oxides) or materials that require more than one inorganic precursors (e.g., metal phosphates require metal and phosphorous precursors), it is necessary that the inorganic precursors interact with each other more strongly than with molecules of the same species (strength of interactions should vary as $I_1I_2 > I_1I_1, I_2I_2$, where I_1 and I_2 represent two different inorganic precursors). Hence, the difference in the acidity/basicity of the precursors should be maximum to maximize the interaction between them. Tian et al.[25] reported a variety of metal phosphates, borates, and mixed metal oxides using this approach.

Use of Nonaqueous Solvents and Addition of Limited Quantities of Water. Use of limited quantities of water in nonaqueous solvents allows the control of the degree of hydrolysis of the transition metal oxide precursors. The resultant hydrolyzed, polar species can have a better interaction with the polar head group of the surfactant, resulting in the formation of ordered structures. Soler-illia et al. reported a ''modulation of hybrid interface'' approach,[26] which relies on the addition of controlled quantities of water to the solution of the inorganic precursor and a nonionic surfactant in an organic solvent to obtain ordered mesostructures.

Use of Dilute Solutions. The reactivity of the inorganic precursors can also be controlled by using low concentrations to lower the reaction rates. Consequent evaporation of solvent results in the formation of organized structure. This approach, introduced by Brinker et al.[27] and termed as ''evaporation induced self-assembly.'' (EISA), allows the formation of powders, films, gels, and monoliths.

Use of Nanoparticles. Instead of using reactive precursors that undergo reactions converting them to the final inorganic material, nanometer sized particles of final inorganic materials can themselves be used as the precursors. These particles will be inert, in that they will not interact with each other, but their interaction with the surfactants will allow formation of ordered mesostructure.

Surfactant to inorganic precursor ratio

The type of mesostructure obtained depends strongly on the surfactant to inorganic ratio. In fact, there is a close correlation between the surfactant to solvent ratio in the phase diagram of a surfactant and the surfactant to inorganic ratio in the mesostructured materials obtained. Alberius et al. demonstrated this correlation by the so-called ''general predictive synthesis'' approach.[28] They used the phase diagrams of the water-surfactant system to guide the synthesis of mesoporous silica and titania films. There was a very close correlation between the values of the volume fraction of the surfactant over which different phases are obtained in the water-surfactant system and in the silica-surfactant and titania-surfactant systems.

Effect of additives

Additives can have multifold effects on the structure of mesoporous materials: they can affect the structure at the nanometer level, i.e., change the phase obtained, or they can act at the micrometer level, changing the morphology of the resultant materials. The morphology control is discussed separately. When they are acting at the nanometer level, the additives can modulate the rate of the hydrolysis and condensation reactions associated with the inorganic precursors and hence change the range of variables over which ordered materials can be obtained. For example, Kim et al. obtained highly ordered silica mesostructures over a diverse range of pH by controlling the relative rates of hydrolysis and condensation of the silica species through the use of fluoride ions and tetramethyl orthosilicate (TMOS) as the inorganic precursor.[29]

Ionic additives can accumulate at the surfactant–inorganic interface and modulate the interfacial curvature, thereby changing the value of ''a,'' the effective head group area per surfactant molecule. This can change the structure of surfactant micelles and hence the structure of the mesophase obtained. The addition of salts to solution can also lower the critical micelle concentration, thus broadening the synthesis domain in surfactant concentration, as well as increase the range of surfactants that can be used in producing ordered mesostructures. The salt addition can also affect the radius of the micelles and hence the unit cell parameter and porosity of the inorganic material obtained after surfactant removal.[30]

The dramatic effect of salt addition on the mesostructure was demonstrated by Che et al.[31]. Starting with solutions with the same surfactant: inorganic ratio and working under the same synthesis condition, and by varying only the type of acid used (H_2SO_4, HCl, HBr, and HNO_3), the authors obtained a three-dimensional hexagonal structure with H_2SO_4 ($P6_3/mmc$ space group), a two-dimensional hexagonal with HBr ($p6m$), a bicontinuous cubic with HNO_3 ($Ia3d$), and a cubic ($Pm3n$) with HCl. (For an introduction to the space group nomenclature, which is quite common among the zeolite and mesoporous materials community to describe the crystal structures, the reader is referred to Ref.[32]. Various ''International Tables for Crystallography'' also constitute a useful and handy reference.) This modulation of structure was explained in terms of the adsorption strength of the acid anions on the head groups of the surfactant micelles. The acid anions are more or less hydrated in the surfactant solution. Less strongly hydrated ions have, in general, smaller ionic radii and bind more closely and strongly on the head group of the surfactant. The small anions contribute to the partial reduction in the electrostatic repulsion between the charged surfactant head groups and the decrease in the effective area of surfactant, ''a,'' thereby resulting in a significant increase in the g value. Ionic radii decrease in the following order: $\frac{1}{2}SO_4^{2-} > Cl^- > Br^- > NO_3^-$. Hence, H_2SO_4 leads to the formation of the 3D hexagonal $P6_3/mmc$ mesophase with a smaller g parameter and HNO_3 favors the formation of the $Ia3d$ mesophase with a larger g parameter. Liu et al.[33] found that the addition of organosiloxane or small organic molecules like benzene and its alkyl derivatives to the synthesis mixture containing TEOS, as silica source, and nonionic block-polymeric surfactants results in the formation of cubic bicontinuous $Ia3d$ phase at room temperature under acidic conditions. Previous reports of the synthesis of this phase were limited to alkaline media and high temperature synthesis.

Effect of environmental variables

Relative Humidity (RH) of Atmosphere. Relative humidity is an extremely important parameter for the synthesis of thin films from solution by dip or spin coating. The dynamics of evaporation is highly dependent on the RH of the atmosphere. The importance of controlling the RH was demonstrated by Cagnol et al.[34]. Starting with a solution with a fixed composition, different mesostructures were obtained just by changing the RH of the atmosphere. Structures with higher curvature and lower values of the packing factor g, such as cubic structures, require higher values of RH. More water is retained in the films at higher RH and hence the flexibility of inorganic framework, where the water is mainly located, is enhanced. The inorganic framework, therefore, can conform to the more curved interface resulting in the formation of cubic structure. At lower values of RH, on the other hand, structures requiring less curvature (such as

2D hexagonal mesostructure) are obtained. It was also demonstrated that the final structure is formed through a series of intermediate structures, the sequence of which depends on the relative humidity. It is also possible to change the structure after the dip coating by changing the relative humidity.

Organic Solvent Environment. The partial pressure of the organic solvent in the atmosphere has effects similar to that of water. The organic solvent generally has high volatility and evaporates faster. At high concentrations of the solvent, though, a high quantity of solvent is maintained in the films and the films remain low in viscosity, paving way for phase transitions.[35] Gallis and Landry studied the transformation processes leading to the formation of MCM-48 from MCM-41 and found that the presence of ethanol at the interface was necessary for the phase transition.[36] The authors theorized that hydrolysis of TEOS leads to the production of ethanol in the vicinity of the organic–inorganic interface. Upon heating, ethanol was driven further into the organic region, increasing the surfactant packing parameter and causing the system to transform from MCM-41 into MCM-48.

Strategies for Stabilization of the Inorganic Framework and Removal of the Template

Most of the applications of surfactant templated materials depend on the porosity of the inorganic framework; hence removal of the template is necessary. However, the inorganic framework must be strong enough to retain its fidelity even in the absence of the template. This is particularly important for the synthesis of materials with reactive precursors, as the strategies to control the precursor reactivity result in a framework that has a lower degree of polymerization and, hence, lower structural stability. This necessitates post-treatment of the mesostructured materials to enhance the polymerization of the inorganic framework and removal of template without destroying the mesostructure. The strategies employed to achieve this end are summarized here.

Thermal Treatments to Consolidate the Inorganic Framework. The surfactants are usually removed by heating the organic–inorganic hybrids to temperatures high enough to oxidize the template. Thermal treatment of the hybrids at a temperature that is lower than that required for template removal but high enough to enhance the rate of condensation will lead to consolidation of the inorganic framework in the presence of the template; this will enhance the stability of the inorganic framework for subsequent template removal at higher temperatures.

Mild Template Removal. Instead of using strong thermal treatments to remove the templates, mild template removal techniques, e.g., by refluxing in a solvent that can dissolve the template, use of UV irradiation, or use of mild oxidants, can also be used for template removal without structural collapse.

Doping with Foreign Atoms to Reduce the Size of Crystallites. In the case of transition metal oxides, calcination of the mesostructures is accompanied by crystallization of the inorganic framework. The size of the crystallites formed in the inorganic framework walls grows as the temperature increases, and it has been shown that when the size of these crystallites becomes comparable to the wall thickness, the structure collapses. Doping of the inorganic walls by heteroatoms prevents the growth of the crystallites, and, therefore, retains the structural integrity of the inorganic framework after template removal.

Aging Under Mild Conditions. Longer duration of aging, if it is permitted by the synthesis parameters, leads to a higher degree of condensation, and, therefore, stronger inorganic walls.

Treatment with Vapors of the Inorganic Precursor. The density of the inorganic walls can be increased post-synthesis by subjecting the material to an environment saturated with the vapors of the inorganic precursor. The vapors infiltrate into the wall of the structure, and fill any defects and microporous voids in the framework, which would otherwise lead to the collapse of the structure on template removal. The denser walls that result from the vapor treatment have much higher structural strength than those of untreated materials.

Morphology Control of Mesostructured Materials

The control of morphology of the materials at a macroscopic level is essential from an application point of view. Various morphologies of mesostructured materials, especially silica and nonsilica mesoporous oxides, have been obtained, such as free standing and supported thin films, spheres, fibers, and monoliths, using a combination of sol–gel chemistry and emulsion chemistry. Syntheses in acidic solutions afford better control over the morphology of the materials. Addition of a cosolvent immiscible with the solvent used for synthesis also leads to the formation of different morphologies of the resultant materials. Completely different morphologies are observed on the oil side and the water side of the resultant emulsion.[37] Morphology can also be affected by the synthesis temperature; polymerization at low temperatures leads to the formation of faceted single

crystals under thermodynamically controlled conditions. For a complete discussion of the control of morphology of mesostructured materials, the reader is referred to a review by Zhao et al.[38].

USE OF RIGID TEMPLATES

Rigid vs. Nonrigid Templates

Although very versatile and well researched, the surfactant templating route has some disadvantages. Surfactants are labile molecules and undergo thermal degradation typically at temperatures lower than ~350 °C. Most transition metal oxides undergo amorphous to crystalline transition at a temperature higher than this, which is responsible for the destruction of mesostructure, as discussed earlier. Most applications of the metal oxides, on the other hand, depend on the crystallinity of the wall. This necessitates the use of rigid templates. It has also been demonstrated that the use of mesoporous carbons as rigid templates for the synthesis of mesoporous metal oxides results in the formation of mesostructures that are otherwise not obtained with surfactant templates. A complete infiltration of inorganic precursors into the mesopores of the template is necessary to obtain dense wall with as little volume shrinkage as possible.

Materials Synthesized

The first ordered mesoporous materials synthesized using rigid templates were carbons. Ryoo et al. reported the synthesis of highly ordered cubic mesoporous carbons (termed CMK-1) using MCM-48 silica as the template[6] (and references therein). As a general route for the preparation of mesoporous carbons, the silica template is infiltrated with a solution of carbon precursor or an organic compound. This is followed by pyrolysis, in which the carbon precursor is converted to carbon. The silica template is then removed by treating with HF or NaOH. Mesoporous carbon materials typically have very high surface areas (1500–1800 m^2/g). But the carbon prepared using MCM-41 as a template (which has a 2D hexagonal structure) collapses upon the template removal to yield disordered microporous structure with high surface area. The reason behind this is explained in Fig. 5. It is necessary that when the silica template is 2D hexagonal, there should be some micropores connecting the mesopores to each other. Otherwise, the resultant carbon nanorods are disconnected and they collapse upon one another after template removal. When large pore 2D hexagonal silica, SBA-15, having micropores in the wall was used as template, the synthesized mesoporous carbon, termed CMK-3, was a true replica of the silica structure.

Tian et al. demonstrated the use of microwave digested silica as a template for the synthesis of a variety of metal oxides having crystalline structures in the form of nanowires and nanospheres.[39] The use of microwaves instead of the usual high temperature calcination to remove the surfactant from the silica during its synthesis leaves behind a large number of surface silanol groups (Si–OH). This leads to better hydrophilic affinity interactions with the inorganic precursors resulting in higher loading of the pores.

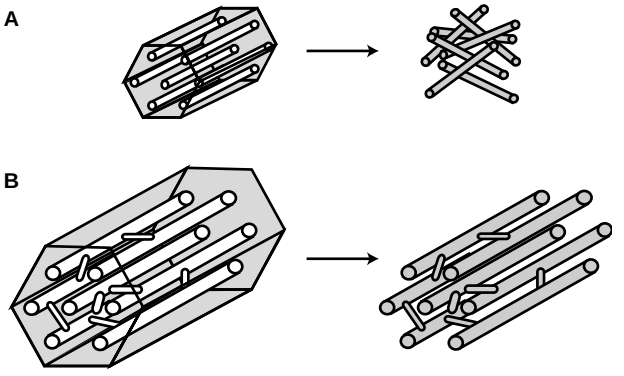


Fig. 5 A schematic illustrating the necessity of an interconnected pore system for the formation of mesoporous carbon. When MCM-41 (A), which does not have interconnections, was used as template, the structure collapsed upon calcination. The use of large pore interconnected SBA-15 (B) led to a stable, highly ordered material. (Adapted from Ref.[6].)

CONCLUSIONS

Phenomenal progress has been made in the synthesis of nanostructured materials in the last decade. A deeper understanding of the formation mechanisms has been established and it is now possible to synthesize these materials in a reproducible way. Modification of the properties of these materials should now pave the way for the use of these materials for conventional applications (catalysis, separation, adsorption, etc.) and for novel applications in the fields of solar energy conversion, electronics, hydrogen generation, etc.

REFERENCES

1. Davis, M.E.; Saldarriaga, C.; Montes, C.; Garces, J.; Crowder, C. A molecular-sieve with 18-membered rings. Nature **1988**, *331* (6158), 698–699.
2. Dessau, R.M.; Schlenker, J.L.; Higgins, J.B. Framework topology of Alpo4-8—the 1st 14-ring molecular-sieve. Zeolites **1990**, *10* (6), 522–524.

3. Kresge, C.T.; Leonowicz, M.E.; Roth, W.J.; Vartuli, J.C.; Beck, J.S. Ordered mesoporous molecular-sieves synthesized by a liquid-crystal template mechanism. Nature **1992**, *359* (6397), 710–712.
4. Yang, P.D.; Zhao, D.Y.; Margolese, D.I.; Chmelka, B.F.; Stucky, G.D. Generalized syntheses of large-pore mesoporous metal oxides with semicrystalline frameworks. Nature **1998**, *396* (6707), 152–155.
5. Attard, G.S.; Coleman, N.R.B.; Elliott, J.M. The preparation of mesoporous metals from performed surfactant assemblies. In *Mesoporous Molecular Sieves*; 1998; Vol. 117, 89–94.
6. Ryoo, R.; Joo, S.H.; Kruk, M.; Jaroniec, M. Ordered mesoporous carbons. Adv. Mater. **2001**, *13* (9), 677–681.
7. Chen, B.L.; Eddaoudi, M.; Hyde, S.T.; O'Keeffe, M.; Yaghi, O.M. Interwoven metal–organic framework on a periodic minimal surface with extra-large pores. Science **2001**, *291* (5506), 1021–1023.
8. Breck, D. *Zeolite Molecular Sieves: Structure, Chemistry, and Use*; Krieger Publishing Company, 1984.
9. Schuth, F. Endo- and exotemplating to create high-surface-area inorganic materials. Angew. Chem. Int. Ed. **2003**, *42* (31), 3604–3622.
10. Ekberg, B.; Mosbach, K. Molecular imprinting—a technique for producing specific separation materials. Trends Biotechnol. **1989**, *7* (4), 92–96.
11. Collinson, M.M. Sol–gel strategies for the preparation of selective materials for chemical analysis. Crit. Rev. Anal. Chem. **1999**, *29* (4), 289–311.
12. Wulff, G. Molecular imprinting in cross-linked materials with the aid of molecular templates—a way towards artificial antibodies. Angew. Chem. Int. Ed. (English) **1995**, *34* (17), 1812–1832.
13. Evans, D.F.; Wennerström, H. *The Colloidal Domain: Where Physics, Chemistry, Biology, And Technology Meet*, 2nd Ed.; Wiley-VCH: New York, 1999.
14. Vasilescu, M.; Caragheorgheopol, A.; Caldararu, H. Aggregation numbers and microstructure characterization of self-assembled aggregates of poly(ethylene oxide) surfactants and related block-copolymers, studied by spectroscopic methods. Adv. Colloid Interf. Sci. **2001**, *89*, 169–194.
15. Israelachvili, J.; Mitchell, D.J.; Ninham, B.W. Theory of self-assembly of hydrocarbon amphiphiles into micelles and bilayers. J. Chem. Soc. Faraday Trans. **1976**, *2* (72), 1525.
16. Soler-Illia, G.; Rozes, L.; Boggiano, M.K.; Sanchez, C.; Turrin, C.O.; Caminade, A.M.; Majoral, J.P. New mesotextured hybrid materials made from assemblies of dendrimers and titanium(IV)-oxo-organo clusters. Angew. Chem. Int. Ed. **2000**, *39* (23), 4249–4254.
17. Bouchara, A.; Rozes, L.; Soler-Illia, G.J.D.; Sanchez, C.; Turrin, C.O.; Caminade, A.M.; Majoral, J.P. Use of functional dendritic macromolecules for the design of metal oxo based hybrid materials. J. Sol–Gel Sci. Technol. **2003**, *26* (1–3), 629–633.
18. Attard, G.S.; Glyde, J.C.; Goltner, C.G. Liquid-crystalline phases as templates for the synthesis of mesoporous silica. Nature **1995**, *378* (6555), 366–368.
19. Attard, G.S.; Leclerc, S.A.A.; Maniguet, S.; Russell, A.E.; Nandhakumar, I.; Bartlett, P.N. Mesoporous Pt/Ru alloy from the hexagonal lyotropic liquid crystalline phase of a nonionic surfactant. Chem. Mater. **2001**, *13* (5), 1444–1446.
20. Huo, Q.S.; Margolese, D.I.; Ciesla, U.; Demuth, D.G.; Feng, P.Y.; Gier, T.E.; Sieger, P.; Firouzi, A.; Chmelka, B.F.; Schuth, F.; Stucky, G.D. Organization of organic-molecules with inorganic molecular-species into nanocomposite biphase arrays. Chem. Mater. **1994**, *6* (8), 1176–1191.
21. Monnier, A.; Schuth, F.; Huo, Q.; Kumar, D.; Margolese, D.; Maxwell, R.S.; Stucky, G.D.; Krishnamurty, M.; Petroff, P.; Firouzi, A.; Janicke, M.; Chmelka, B.F. Cooperative formation of inorganic-organic interfaces in the synthesis of silicate mesostructures. Science **1993**, *261* (5126), 1299–1303.
22. Tanev, P.T.; Pinnavaia, T.J. Mesoporous silica molecular sieves prepared by ionic and neutral surfactant templating: a comparison of physical properties. Chem. Mater. **1996**, *8* (8), 2068–2079.
23. Zhao, D.Y.; Feng, J.L.; Huo, Q.S.; Melosh, N.; Fredrickson, G.H.; Chmelka, B.F.; Stucky, G.D. Triblock copolymer syntheses of mesoporous silica with periodic 50 to 300 angstrom pores. Science **1998**, *279* (5350), 548–552.
24. Antonelli, D.M.; Nakahira, A.; Ying, J.Y. Ligand-assisted liquid crystal templating in mesoporous niobium oxide molecular sieves. Inorg. Chem. **1996**, *35* (11), 3126–3136.
25. Tian, B.Z.; Liu, X.Y.; Tu, B.; Yu, C.Z.; Fan, J.; Wang, L.M.; Xie, S.H.; Stucky, G.D.; Zhao, D.Y. Self-adjusted synthesis of ordered stable mesoporous minerals by acid–base pairs. Nat. Mater. **2003**, *2* (3), 159–163.
26. Soler-illia, G.J.D.; Sanchez, C.; Lebeau, B.; Patarin, J. Chemical strategies to design textured materials: from microporous and mesoporous oxides to nanonetworks and hierarchical structures. Chem. Rev. **2002**, *102* (11), 4093–4138.
27. Brinker, C.J.; Lu, Y.F.; Sellinger, A.; Fan, H.Y. Evaporation-induced self-assembly: nanostructures made easy. Adv. Mater. **1999**, *11* (7), 579–585.

28. Alberius, P.C.A.; Frindell, K.L.; Hayward, R.C.; Kramer, E.J.; Stucky, G.D.; Chmelka, B.F. General predictive syntheses of cubic, hexagonal, and lamellar silica and titania mesostructured thin films. Chem. Mater. **2002**, *14* (8), 3284–3294.
29. Kim, J.M.; Han, Y.J.; Chmelka, B.F.; Stucky, G.D. One-step synthesis of ordered mesocomposites with non-ionic amphiphilic block copolymers: implications of isoelectric point, hydrolysis rate and fluoride. Chem. Commun. **2000**, *24*, 2437–2438.
30. Yu, C.Z.; Tian, B.Z.; Fan, B.; Stucky, G.D.; Zhao, D.Y. Salt effect in the synthesis of mesoporous silica templated by non-ionic block copolymers. Chem. Commun. **2001**, *24*, 2726–2727.
31. Che, S.N.; Lim, S.Y.; Kaneda, M.; Yoshitake, H.; Terasaki, O.; Tatsumi, T. The effect of the counteranion on the formation of mesoporous materials under the acidic synthesis process. J. Am. Chem. Soc. **2002**, *124* (47), 13962–13963.
32. Sands, D.E. *Introduction to Crystallography*; Dover Publications: New York, 1994.
33. Liu, X.Y.; Tian, B.Z.; Yu, C.Z.; Gao, F.; Xie, S.H.; Tu, B.; Che, R.C.; Peng, L.M.; Zhao, D.Y. Room-temperature synthesis in acidic media of large-pore three-dimensional bicontinuous mesoporous silica with *Ia*3*d* symmetry. Angew. Chem. Int. Ed. **2002**, *41* (20), 3876–3878.
34. Cagnol, F.; Grosso, D.; Soler-Illia, G.; Crepaldi, E.L.; Babonneau, F.; Amenitsch, H.; Sanchez, C. Humidity-controlled mesostructuration in CTAB-templated silica thin film processing. The existence of a modulable steady state. J. Mater. Chem. **2003**, *13* (1), 61–66.
35. Alonso, B.; Albouy, P.A.; Durand, D.; Babonneau, F. Directing role of pH and ethanol vapour on the formation of 2D or 3D mesostructured silica and hybrid organo-silica thin films. N.J. Chem. **2002**, *26* (10), 1270–1272.
36. Gallis, K.W.; Landry, C.C. Synthesis of MCM-48 by a phase transformation process. Chem. Mater. **1997**, *9* (10), 2035–2038.
37. Schacht, S.; Huo, Q.; VoigtMartin, I.G.; Stucky, G.D.; Schuth, F. Oil–water interface templating of mesoporous macroscale structures. Science **1996**, *273* (5276), 768–771.
38. Zhao, D.Y.; Yang, P.D.; Huo, Q.S.; Chmelka, B.F.; Stucky, G.D. Topological construction of mesoporous materials. Curr. Opin. Solid-State Mater. Sci. **1998**, *3* (1), 111–121.
39. Tian, B.Z.; Liu, X.Y.; Yang, H.F.; Xie, S.H.; Yu, C.Z.; Tu, B.; Zhao, D.Y. General synthesis of ordered crystallized metal oxide nanoarrays replicated by microwave-digested mesoporous silica. Adv. Mater. **2003**, *15* (16), 1370–1374.

Nanotribology

Jonathan W. Bender
Department of Chemical Engineering, University of South Carolina, Columbia, South Carolina, U.S.A.

Xiaodong Li
Department of Mechanical Engineering, University of South Carolina, Columbia, South Carolina, U.S.A.

INTRODUCTION

Nanotribology is the study of friction, lubrication, and wear at nanometer scales. Trends in device miniaturization, coatings for protection from wear in extreme environments, and lubricant technology motivate the development of improved materials and structure–property relations. Because engineering surfaces are innately rough and contact is made via a multiplicity of micro- and nano-sized contacts, measuring material properties and lubricant performance at this scale is critical, especially as classical descriptions of material behavior fail. This entry discusses efforts to measure and predict friction and material deformation at the nanoscale, leading to the development of improved structural materials for tribological applications.

MOTIVATION FOR NANOTRIBOLOGY

Tribology, the study of friction and wear, is of immense importance to the transportation and process industries. Over $100 billion in capital is lost each year because of wear-related reduced functionality of machines and devices.[1] Except in cases where transmitting mechanical forces between surfaces is desirable (such as traction in sneakers, automotive brakes, and fingertips), tribology is substantially focused on reducing friction as a means of improving mechanical longevity. However, mathematical relationships between friction and wear remain elusive, even for the simplest surfaces, because of the complex interplay of chemical reactions, mechanical deformation, heat conduction, and lubrication phenomena that occur during sliding. Thus, the key need of tribology is the development of constitutive relationships that link surface physicochemical and mechanical characteristics to energy dissipation and material deformation.

The development of instrumentation capable of studying and manipulating matter on a molecular scale has ushered in new opportunities to explore surfaces in relative motion (Fig. 1).[2] The field of nanotribology has emerged from the belief that a more fundamental understanding of the physics of interacting surfaces on the nanometer scale will yield new design methodologies of larger (micrometers and up) systems. Technologically, nanotribology is driven by trends in device miniaturization, composite coating development for extreme wear environments, nanofluidics and nanorheology for bioseparations, and lubrication research, and data storage technology, and by the need to meet increasingly stringent environmental and technological challenges in the ''conventional'' lubricant industry. Fundamental scientific issues need to be addressed as well, including the origins of friction and modes of energy dissipation, and how materials well described at the macro level respond differently when probed at nanometer scales. For example, fluids such as vegetable oil that exhibit only viscous behavior under bulk conditions behave viscoelastically when confined to molecular dimensions.

Earlier books and reviews have provided considerable detail on specific aspects of nanotribology.[3–11] Some are written from the lay person's perspective, and provide a historical account of tribology leading to specific research efforts in the field.[3–5] Others are compilations of the state-of-the-art contributions in various fundamental and applied aspects of nanotribology.[6–11] The intention here is to provide a brief overview of the purpose and progress of nanotribology in developing a more complete understanding of the behavior of sliding interfaces. The first two sections explore the atomic and molecular origins of static and kinetic friction, the third focuses on differences of the response of fluids to molecular confinement, the fourth focuses on new nanoindentation methods and insight gained on the deformation of materials at the nanoscale. The relevance of the research from a practical viewpoint of materials and systems development is highlighted.

MOLECULAR-LEVEL EXPLANATION OF MACROSCOPIC FRICTION ''LAWS''

Despite the ubiquitous presence of friction, most people study friction for no more than one or two

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039301

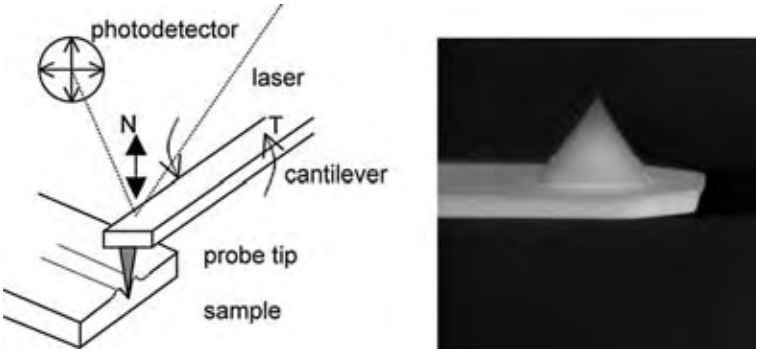


Fig. 1 Schematic of the AFM cantilever and cantilever tip micrograph. Cantilever bending normal to the surface is used to calculate a normal force (N) and torsional motion is used to calculate the frictional force (T). Cantilever deflections are detected by a photodetector.

lectures in a basic physics class. Classical ''rules'' are taught that are based on everyday observations and a host of experimental data. To summarize:

1. Frictional forces do not depend on an ''apparent'' area of contact, but rather the ''true'' area of contact between the opposing asperities (Fig. 2).
2. Frictional forces are proportional to normal load (known as Amontons' laws). This has been explained by the increase in true contact area between the surfaces as the load increases.
3. The force required to begin sliding an object from rest (static friction) is generally greater than the force required to sustain motion (kinetic friction).
4. The kinetic friction force is independent of sliding velocity (Coulomb's law).

Surprisingly, these rules apply to a wide range of paired surfaces, irrespective of roughness, hardness, or the presence of a lubricant.

Today, many undergraduate physics textbooks describe friction in terms of the work required to separate adhesive microcontacts formed during sliding. As the load increases, the true contact area increases, resulting in a larger resistance to motion. This intuitively simple picture explains the load dependence of friction (1 and 2 above), but does not provide insight into the origins of friction, mechanisms of friction energy dissipation, and the differences between static and kinetic friction. Despite embellishment, models based on a purely classical description of surfaces cannot explain the commonly observed friction rules. Thus, considerable theoretical models, molecular dynamics (MD), computer simulations, and nanoscale experimentation are converging on the mechanisms of energy dissipation, origins of static friction, and the velocity independence of friction. These efforts are briefly described.

Proportionality of Friction and Load

For surfaces that deform plastically, the contact area A is proportional to the applied load. A single elastic contact deforms as $A^{2/3}$ as load increases, and would not be expected to follow this rule.[12] However, when considering an exponential surface height distribution, which leads to a multiplicity of elastic asperity contacts, a linearity between load and contact area is recovered. Using instrumentation developed in the last 15 yr, notably the atomic force microscope (AFM) and surface forces apparatus (SFA), researchers have explored the universality of friction–load proportionality over a much wider range of dimensions and surface characteristics. Indeed, SFA experiments have shown friction–load proportionality between atomically smooth mica surfaces in dry air over square micrometers of contact area. A contact mechanics expression for elastic contacts that incorporates the effects of adhesion was used. Similarly, AFM experiments of

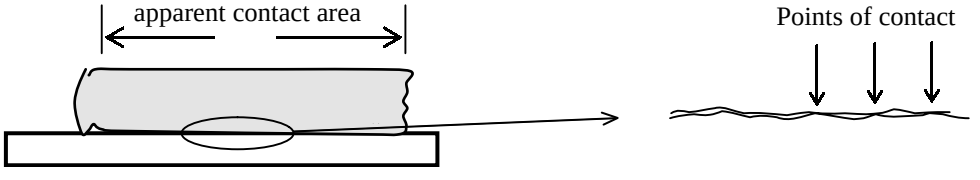


Fig. 2 Comparison of the apparent contact area to the real contact area between individual asperities. Differences between these areas can exceed 1000 : 1. *(View this art in color at www.dekker.com.)*

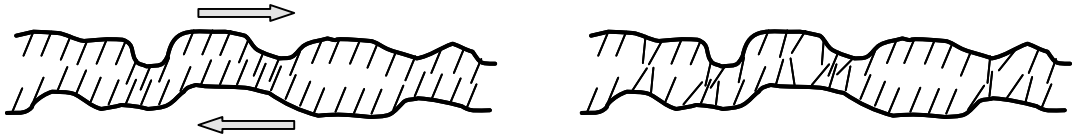


Fig. 3 Representation of the relaxation mechanism. The figure to the left has surface molecules that are responding simply to the shear, whereas the figure to the right shows the surfaces at rest. The molecules move to their locally lower energy state. *(View this art in color at www.dekker.com.)*

a single-asperity contact between two highly stiff materials showed a friction–load proportionality. Surprisingly, these nanotribological investigations have validated the applicability of Amontons' laws to nanometer dimensions.[13] Further, they clearly demonstrate the presence of friction in the absence of wear through studies of well-characterized surfaces, and demonstrate the significant role that adhesion plays in the generation of friction.

The Origin of Static Friction

Atomically flat, weakly interacting surfaces are predicted to exhibit no friction. In this limit, no asperities block motion, and the atoms of one surface cannot incite movement of atoms on the opposing surface. Thus, no relaxation of the surfaces to local energy minima occurs, and no energy is required to initiate sliding (Fig. 3). In practice, these conditions are difficult to achieve, so simulations have explored the effect of atomic impurities on static friction. In MD simulations, both the motions of the surface impurity molecules and the atoms that compose the surface are considered explicitly. The friction is obtained from an atomistically applied damping constant required to keep the system isothermal. Through these computations, a static co-efficient of friction has been predicted in the case of deformable surfaces under compression.[14] Upon cessation of sliding, the atoms at the interface move swiftly to a low energy state and require a finite force to initiate motion. During sliding, the atoms are not given an opportunity to reach a low energy state. Thus, the force required to initiate motion is larger than that required to continue motion. In the case of atomic impurities, the relaxation of the interface generally occurs on time scales too rapid for experimental verification. However, using an SFA to confine liquid films of oligomers to molecular thicknesses, an increase in the stress required to initiate motion as a function of elapsed time was demonstrated.[15] This provides strong evidence that static friction arises from interfacial relaxation processes.

Velocity-Independent Friction

Of all the classical friction rules, this is perhaps the most complex one to elucidate. In contrast to most everyday observations, many theories and simulations generally predict a viscous response where the friction is proportional to, rather than independent of, sliding velocity.[9] In fact, quartz crystal microgravimetry (QCM) experiments also show a velocity-dependent friction.[5] Quartz crystal microgravimetry is able to measure the degree of slippage of condensed gas films on a rapidly oscillating and resonating metal surface (Fig. 4). Through resonant frequency and amplitude changes, the slippage of a layer of condensed krypton was found to be dependent on QCM oscillation amplitude. (That is, the friction seen in this system is velocity dependent, in contrast with rule 4.) In lubricated systems, velocity independence may be partially explained by the shear thinning of the lubricant, in which the viscosity of the fluid decreases as shear rate is increased. Further, wall slip can occur, resulting in a momentary reduction in friction. However, the occurrence of velocity independence in dry systems has not been explained, including those conducted at a nanometer scale using AFM. Thus, a more universal explanation for this friction behavior is needed, and provides impetus for future research in this area.

NATURE OF FRICTION

The focus of this discussion so far has been on research that compares macroscopically observed friction

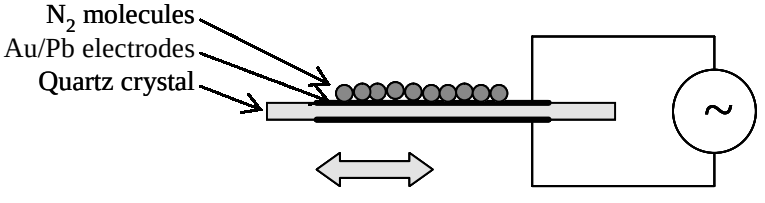


Fig. 4 Schematic of the QCM experiment with Pb-coated electrodes. *(View this art in color at www.dekker.com.)*

phenomena to that found at nanometer scales. No wear has been allowed in most simulations, or is desired in nanoscale experiments, and all frictional energy was dissipated in the form of heat. Because wear is a primary concern of tribological research, the modes by which energy is dissipated are of significant interest. From a materials design perspective, it is desirable to have the energy dissipated in the form of heat, rather than in the form of viscoelastic or plastic deformation, which results in permanent damage. The focus of this entry is on exploring the mechanisms by which energy is dissipated and the connections between friction, atomic, and molecular structure.

Phononic and Electronic Energy Dissipation

Energy transfer away from a sliding contact occurs through a number of means. Lattice vibrations can be excited through localized variations in relative velocity caused by asperity contact.[5,9] A ''wave'' of atoms vibrating in concert, termed a phonon, can be transmitted through the solid. These atomic vibrations may be absorbed at grain boundaries, resulting in atomic dislocations at the boundary. With sufficient accumulation of energy, the material may fracture and be removed from the surface. Energy may also be dissipated via surface exchange of electric charge. A classic experiment is to charge a balloon by rubbing it against hair. This electronic friction can be modeled as charges and dipoles moving through a ''viscous'' medium. Coulombic interactions between dissimilar surfaces can also generate resistance to motion even with noncontacting surfaces, although these forces are generally small. To experimentally distinguish between phononic and electronic contributions to the friction, Dayo, Alnasrallah, and Krim performed QCM experiments of monolayers of nitrogen on lead under ultrahigh vacuum (UHV) conditions (Fig. 4).[16] As the system temperature cools to below the superconducting temperature of lead, the electronic contribution to friction will disappear, allowing a calculation of the relative contributions of phononic and electronic friction to be measured. While other experimental evidence suggested that both phononic and electronic contributions to friction existed, these marked the first nanotribological studies that could determine the relative magnitude of these effects.

Commensurate and Incommensurate Surfaces

The term ''commensurate'' is applied to opposing surfaces that possess the same lattice spacing and orientation, as though the surfaces had been generated by cleaving a single crystal (Fig. 5). While commensurate surfaces do not typically appear in tribology, they assist in elucidating the origins of friction and modes of energy dissipation: Given the above discussion on phonon transport and slip at grain boundaries, it follows that friction would be sensitive to lattice orientation. Indeed, theory and simulations predict a significant dependence of friction on the commensurate nature of the interface.[5]

Numerous tribological studies have measured the friction anisotropy that results as a function of changing lattice orientation on both elastic and plastic contacts.[17] Under elastic deformation conditions, both adhesion and friction have been shown to be highly dependent on slight changes in orientation in SFA studies of atomically flat mica surfaces. The friction and adhesion anisotropies persisted even when the surfaces were separated by four molecular layers of liquid. Using AFM, friction anisotropy and lattice-directed sliding have been measured between diamond surfaces and between MoO_3 nanocrystals and single-crystal MoS_2, respectively.

In contrast, for plastically deformed surfaces, friction anisotropy appears to be primarily attributable to the movement of atomic slip planes within the bulk of the metal, and not to commensurability at the sliding interface. Early experiments using diamond surfaces showed that friction anisotropy disappeared at low loads where no plastic deformation was evidenced.

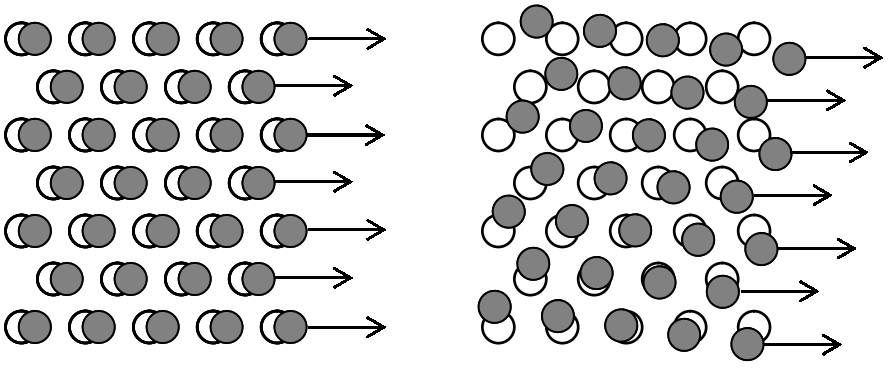


Fig. 5 Atomic configurations representing a commensurate surface (on the left) and incommensurate surface (on the right). Sliding friction is predicted and experimentally shown to be different between these cases.

Friction anisotropy has also been seen between clean and monolayer-coated metal (100) single crystals in UHV, and persists when the surface is covered with as much as 20–40 monolayers of octane, beyond which no anisotropy is detectable.[17]

Energy Dissipation in Polymers

Many of the above nanotribology theories, computer simulations, and experiments have focused on surfaces composed of covalently or metallically bonded atoms. Polymers are distinguished on the basis of their covalently bonded backbone, but entangled macromolecular structure. Relaxation times for polymers can extend from seconds to days, and lend themselves nicely to nanotribology studies of time-dependent behavior. Current nanotribology experiments on polymers using the AFM have found that the dependence of friction co-efficients on sliding speed can be traced to polymer segment motion.[18] When the sliding time is comparable to the relaxation times associated with the movement of polymer functional groups, the friction is maximized. This behavior has been seen in several polymers, including poly(vinyl alcohol), poly (vinyl acetate), poly(ethylene terephthalate), and polystyrene thin films. Further, because of the inherently fast response time of the cantilever compared to macroscopic load cells, the AFM can acquire ''instantaneous'' frictional forces on a nanometer scale. In a separate study, it has been suggested that the distribution of friction forces is reflective of glass-like and rubber-like relaxations, again tying friction to the relaxation behavior of the polymer.

NANORHEOLOGY

One central issue in the development of lubricants for tribological purposes is the shear behavior of molecules confined to small, thin layers. Conventional tribological experiments often use ill-defined surfaces, and more than one parameter change during measurement (viscous heating, surface roughness changes), making the interpretation of friction behavior on the basis of the shape of the molecule, functionality, and size difficult. One of the most remarkable nanotribological experiments performed using the SFA was the finding that simple molecules form a layered structure upon angstrom-level confinement between atomically smooth mica surfaces.[19] Simulations and models have followed that predict this behavior.[9] Alternating compressive and tensile forces were created as the mica surfaces were brought into close proximity, with a period that corresponded to the segment diameter of a methyl group. The oscillations in force decreased in magnitude using a linear chain having a pendant methyl group, which interrupts the layered structure. This provided an important clue as to the importance of molecular structure in the load carrying capacity of lubricant films. Subsequent SFA experimentation with perfluoroethers demonstrated that branched molecules produce thinning, more compression resistant films than linear chains, and that as molecular weight increases, the lubricant exhibited solid-like to liquid-like transitions with increasing strain.[20] Since then, the SFA has proven to be indispensable in investigating the effects of lubricant physicochemical properties such as friction, stick-slip, and nanorheology.

Stick-Slip Behavior

Under certain conditions, systems can exhibit an unsteady stick-slip motion. Technologically, stick-slip behavior is usually undesirable, causing mechanical shudder in power transmission equipment, for example, and it is difficult to predict and control. In essence, as the velocity decreases, an increasing friction force slows the system further, potentially bringing the system to rest. At this point, the applied tangential forces build up, ultimately reinitiating motion. From the preceding discussion of the origins of static friction, relaxations of lubricant and surface molecules can explain this velocity-dependent friction. However, theory, nanoscale simulations, and experimentation have shown that confined lubricant films can also undergo a solid-like–liquid-like phase transition.[21] This more abrupt behavior is temperature sensitive, but surprisingly load and speed independent. These findings suggest that friction may not always be ''smoothly'' dependent on the viscoelastic nature of a lubricant film, but may undergo sharp transitions attributable to confinement-dependent phase behavior.

End-Chain Functionality

Because of the importance that friction modifiers play in lubricant technology, the effect of varying chain-end chemical functionality on nanorheology of alkane chains has been studied.[22] Measurements of the dynamic shear stress revealed differences in solid-like and liquid-like behavior of the thin films depending on the strength of hydrogen bonding between end groups. Carboxylic acid groups have strong bonding and exhibit a more solid-like behavior than corresponding measurements with hydroxyl-terminated alkanes. Further, self-association between end groups affects film thickness. These results clearly establish links between nanorheological behavior and the behavior of lubricants in larger-scale applications.

Lubricity of Aqueous Systems

Water is a reasonable lubricant under low load conditions, but a poor one under higher loads. For this reason, considerable attention has recently been focused on aqueous systems in an effort to understand the mechanisms by which diarthrodial joints and other biological systems maintain such low-friction co-efficients. The large number of variables associated with the choices of surfactant or polymer type and conformation (whether native or synthetic, hydrated or polyelectrolytic, homopolymer, or diblock, tethered, or untethered) coupled with the variability in testing conditions creates an inherently complex experimental space. Currently, there is insufficient information to draw significant generalizations. However, experiments have shown that hydrated and electrolytic polymer-coated surfaces remain more fluid-like compared to nonaqueous systems, and that their ability to lubricate directly depends on hydration.[23] Clearly, nanotribological investigations of biolubrication will continue to be fruitful areas of research.

Effects of Density Fluctuations

The nature of thin films may also be influenced by rapid, thermally driven fluctuations in the local density of liquids. Heuberger, Zäch and Spencer performed force and refractive index measurements of hexane confined to nanometer thickness using fast spectral correlation spectroscopy, and observed the fluctuations in density.[24] This finding may help explain observations of high surface mobility, anomalously fast self-diffusion, and apparent solidification in liquid films.

NANOINDENTATION

In addition to the behavior of intervening lubricants, mechanical properties of material surfaces and solid thin films clearly affect tribological performance.[25] Recent advances have shown the capability to fabricate and control materials and structures on the scale of nanometers, which has spurred research both in tribological films for extreme environments and in the design of durable micro/nanodevices. Measuring mechanical properties on the nanoscale is essential, considering the strong size and structure dependence of both thick and ultrathin films, nanocomposites, nanowires, nanotubes, nanobeams, and nanoparticles.[26] Direct nanoindentation shows continuing promise for measuring hardness, elastic modulus, scratch resistance, film adhesion, fracture toughness, creep resistance, and fatigue of materials.[27] Further, nanoindention mimics micro- and nanoscale physical contact at the tribological interface of magnetic storage and MEMS/NEMS devices. Thus, the discussion here will focus on the recent developments in nanoindentation techniques, and on the use of nanoindentation to elucidate the mechanisms of solid deformation.

Nanoindentation Techniques

Recently developed nanoindention instruments can accurately measure loads as small as a nano-Newton and displacements of about an angstrom. Clearly, this depends on the indenting tip structure. The Berkovich triangular pyramidal diamond indenter is widely used for measuring hardness and elastic modulus at nanometer scales. Ideally, the indenter tip radius should be as small as possible, and has been fabricated to 20 nm dimensions using focus ion beam techniques. At a small indentation depth, friction between the indenter and the material being indented, which has been long ignored, needs to be considered in analyzing the load–displacement curve. Surface roughness greatly affects the hardness and elastic modulus measurement.

Hardness and elastic modulus measurements

A depth-sensing nanoindenter is capable of continuously measuring load and displacement throughout nanoindentation (Fig. 6). Experimentation has been focused on improving the accuracy of hardness and modulus measurements through concurrent measurements of contact area. Material pile-up around the indentation, unaccounted for in earlier analysis procedures, can now be measured via subsequent topographical imaging using an AFM (Fig. 7). Numerical and finite element analyses that consider pile-up effects have been used to extract more accurate mechanical property measurements from load–displacement curves. Versatile methodologies are still under development.

To measure the hardness and elastic modulus of thin films while avoiding the influence of the substrate, peak indentation depth cannot exceed about 30% of the film thickness.[25] Because commercial nanoindenters can make a minimum penetration depth of 10–15 nm, hardness and elastic modulus of films thinner than 30 nm cannot be measured. Clearly, new techniques for fabricating sharper indenters and new nanoindentation theories are needed to extend this technique. For film thicknesses less than 30 nm, nanoscratch tests are widely accepted to evaluate the mechanical properties (discussed later). Alternatively, assuming the hardness and elastic modulus of a film do not change with thickness, thicker films can be used.

A significant improvement in nanoindentation testing is continuous stiffness measurement (CSM).[27]

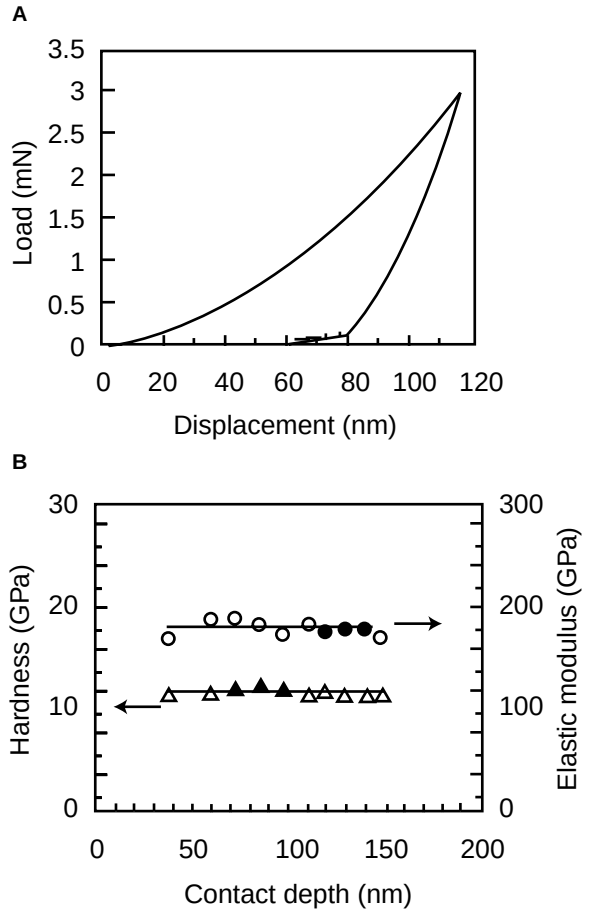


Fig. 6 (A) A representative load–displacement curve of an indentation made at 3 mN peak indentation load and (B) the hardness and elastic modulus as a function of indentation contact depth for the $Al_{74.6}Co_{16.9}Ni_{8.4}$ quasicrystal. (Li, X.; Zhang, L.; Gao, H. Micro/nanomechanical characterization of a single decagonal AlCoNi quasicrystal. J. Phys. D: Appl. Phys. **2004**, *37,* 753–757.)

During CSM, a smaller, oscillatory force is superimposed on a steady larger force. This enables CSM during loading. Without CSM, stiffness can only be measured from the unloading cycle, which requires a larger indentation depth to be obtained. Further, CSM allows mechanical properties to be obtained without the need for multiple, discrete unloading cycles, and avoids creep effects. Finally, the CSM can detect microstructural and mechanical property changes with indentation depth in nonuniform, multilayered materials.

For viscoelastic materials such as polymers and many biomaterials, recently developed nanoindentation dynamic mechanical analysis (DMA) techniques enable the study of dynamic response as a function of load amplitude and frequency. The DMA stiffness mapping technique offers the unique capability of mapping storage and loss moduli of the sample surface. This technique has proven to be very powerful in characterizing copolymers and composites.

Fracture toughness measurements

In general, a material with high strength and fracture toughness exhibits high wear resistance. Fracture toughness of a material is a measure of its resistance to the propagation of cracks, and can be measured via nanoindentation testing. Fracture toughness can be obtained by measuring the length of cracks generated radially from the corners of the indentation. However, several micrometers of indentation are needed for many brittle materials, which exceeds the thickness of many tribologically important thin films. Thus, the fracture toughness measurement of ultrathin films depends on the development of new indenter geometries.

Indenters that displace more volume with shallower indentations, such as the cube-corner indenter, generate stresses and strain concentrations sufficient to create radial cracks in thin films. Further, it has been experimentally proven that the fracture toughness measurement equation is applicable to the radial cracks by the cube-corner indenter.[25] Improvements to toughness measurements from nanoindentation beyond simple models could clearly be made by atomistic computer simulation of crack initiation and propagation, which are currently lacking. Indentation experiments capable of investigating fracture of nanometerthick films are ongoing.

As an alternative to indenter development, new nanoindentation methodologies for measuring the fracture toughness of ultrathin films (of the order of 100 nm thickness) have been recently developed. By measuring crack length via SEM/AFM and calculating the energy release in indentation-induced through-thickness cracking from steps in the loading curve, the fracture toughness can be obtained. Fracture toughness of thin films with thicknesses less than 100 nm has not yet been achieved.

Fatigue measurements

Recently, nanoscale fatigue damage has received much attention because frequent, high-magnitude impacts that occur at the interfaces of magnetic storage and MEMS/NEMS devices lead to fatal flaws.[28] Nanoindenters outfitted with CSM are ideally suited to simulate nanoscale fatigue damage through high-frequency repetitive surface impact. During fatigue tests, sharp decreases in contact stiffness or a sudden increase in indentation depth indicates specimen damage. Fatigue resistance as a function of both load and amplitude frequency has been studied for these thin films and nanostructures. The number of cycles to failure as a function of load, commonly obtained

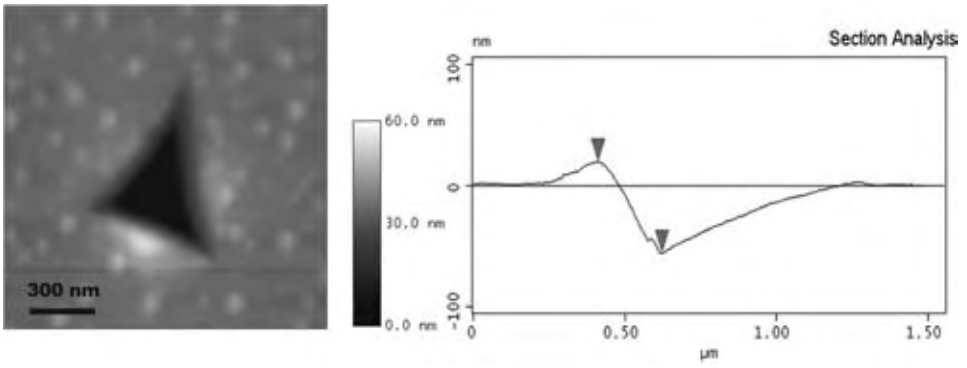


Fig. 7 Atomic force microscopy image and the cross section of the residual impression made at 3 mN. (Li, X.; Zhang, L.; Gao, H. Micro/nanomechanical characterization of a single decagonal AlCoNi quasicrystal. J. Phys. D: Appl. Phys. **2004**, *37,* 753–757).

in macroscale tribological durability testing, can be used to gage device or surface durability in a specific service.

Fatigue tests on nanoscale silicon beams used in constructing MEMS/NEMS devices, such as radio frequency switches, are of critical importance. Fatigue damage is the main failure mechanism of these devices. Silicon itself is a brittle material. No fatigue damage has been found in bulk silicon. Nanoscale fatigue tests on silicon nanobeams in ambient air have suggested that the failure of the beams arises from sequential fracturing of the oxide surface layer and oxidation of the underlying exposed silicon. More specifically, many film/substrate systems exhibit: 1) indentation-induced compression; followed by 2) delamination and buckling of the film from the substrate; and finally 3) a ring-like crack formation of the delaminated film.

Scratch and wear resistance measurements

Nanoscratch tests have been used to simulate the effect of third-body particulate wear debris on component surface scratching during use. The load at which the co-efficient of friction or friction force suddenly increases is identified as the critical load, and is used to evaluate scratch resistance and adhesion strength. The depth-sensing nanoindenter, usually equipped with a conical indenter, can elucidate the mode of failure, whether elastic/plastic deformation, cracking, or delamination.

Finally, wear tests can be realized by using a nanoindenter tip sliding against the sample surface at a constant load. Cyclically sliding against the samples surface is analogous to ball-on-flat wear tests. The indenter can detect the co-efficient of friction and wear depth in situ, offering a real-time study of wear mechanisms. The wear scar can then be imaged by the same tip. Debris particles, delamination, and buckling can be monitored. In summary, continuing advances in nanoindentation and related methods will accelerate materials development and durability testing for ultrathin films and microdevices.

Material Deformation Mechanisms

In the preceding section on energy dissipation mechanisms in solids, the origins of friction in the absence of material deformation have been studied. The focus here is to review the role nanoindentation experimentation has played in elucidating material deformations. Briefly, in the initial stages of deformation, many surfaces deform elastically. Surprisingly, this elastic deformation has been successfully modeled using classical theories of contact mechanics even at nanometer scales.[13] If the load is applied for sufficient durations, materials such as metals and polymers will creep, where the individual or collective diffusion of atoms or molecules will relax the stress.[29] Beyond a critical stress, the material will deform via crack propagation (in brittle materials) or via the nucleation and migration of atomic lattice dislocations (Fig. 8).[30,31] The brittle–ductile transition has been described through a balance between the energy associated with creating additional free surface and the energy required to create and move a dislocation.

"Pop-in" discontinuity

Upon exceeding a critical stress, ductile surfaces deform plastically. Often in nanoindentation experiments under controlled load, the indenter moves in a

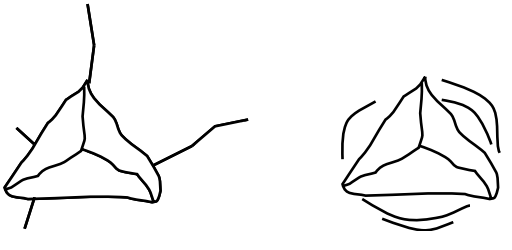


Fig. 8 Differences in response to indentation. The colored area is the indentation. In a brittle material, cracks form, whereas in a more ductile material, dislocation motion accommodates plastic deformation.

discontinuous, step-like fashion, analogous to an avalanche (Fig. 9). This ''pop-in'' effect has been shown to arise from one of the two mechanisms, breakthrough of the harder oxide overlayer that covers most metals exposed to air and nucleation and migration of crystal dislocations. To correlate nanoindentation pop-in events with atomic motion, recent MD simulations have shown a correlation between indenter motion and the formation of dislocation loops and cross-slip events. A central finding of these simulations is the importance of elucidating the relative roles of dislocation nucleation vs. dislocation migration. These simulations were performed on an oxide-free surface. Experimentally, the presence of an oxide has a dramatic effect on the critical load behavior. Electropolished tungsten exhibited an initial elastic deformation regime, followed by a staircase yielding phenomenon indicative of multiple pop-in events, whereas mechanically polished tungsten exhibited a steady plastic deformation with no initial elastic regime. Because both have equal oxide layers, the pop-in effect was attributed to dislocation nucleation. More recent nanoindentation studies have focused on silicon because of its use in MEMS fabrication, its crystallinity (which allows comparisons to simulations), and its unusual transition to a conductive phase under high loads.

Amorphous alloys

Because of their high strength and large elastic deformation, bulk metallic glasses are emerging as structural materials for MEMS and other devices. As the name suggests, these metal alloys have no inherent crystalline structure, and exhibit decidedly different indentation behavior from their crystalline counterparts. Plastic deformation is highly localized to shear bands, which propagate rapidly across a macroscopic specimen.[32] Recent nanoindentation studies show small, incremental discontinuities in response to load. Further, nanocrystallites form in shear bands during indentation attributable to flow dilation inside the bands and the radially enhanced atomic mobility within deforming shear bands.

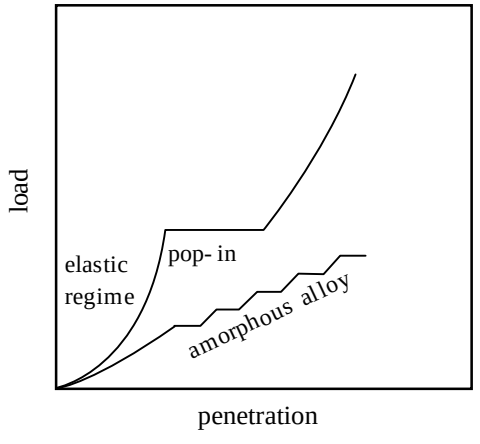


Fig. 9 Graphical representation of results from a nanoindentation experiment. For many solids, an elastic regime at low loads is followed by a discontinuous excursion. For amorphous alloys, a series of ''staircase'' deformations occur in response to shear banding.

Hard thin films

To increase the wear resistance of surfaces, silicon and metals are often coated with a hard nitride, carbide, boride, or oxide film.[33] Nanoindentation and fracture simulations have been used extensively to elucidate failure mechanisms of these typically more brittle surfaces, which include crack propagation and film delamination. Considerable attention has also focused on nanocomposite materials, which possess nanocrystalline inclusions in an otherwise amorphous matrix. The nanocrystalline component is sufficiently small to preclude the formation of stable dislocations, and thus provide a higher hardness.

CONCLUSIONS

Clearly, nanotribology has made significant progress in our understanding of the nature of friction, how friction is influenced by surface and lubricant structure, and how materials deform under specific loads. This has led to more complete explanations of macroscale behavior, such as the classical ''rules'' of friction and the nanoscale deformation mechanisms of metals, ceramics, thin films, and nanocomposites. Further, with the progress in synthesizing nanoscale materials, such as carbon and boron nitride nanotubes, the need to develop tools to probe mechanical properties at the nanoscale has never been greater. Despite these successes, several challenges remain:

1. Mathematical structure–property relations have yet to be developed that tie molecular level information to a prediction of friction and wear behavior. Current theories focus on idealized structures and interfacial interactions that are rarely realized in practice. Thus, statistical representations of the surface topography and interfacial forces need to be integrated with nanotribological measurements to develop more widely applicable and predictive tools.
2. Many of the instruments used in nanotribology experiments are limited in terms of the surfaces and materials used. The surface forces apparatus accommodates ''soft'' materials such as

lubricants, polymers, and surfactants, but the ''solid'' surfaces are limited primarily to mica. Atomic force microscopy studies typically use a silicon or silicon nitride probe tip. Replacing the sharp tip with a colloidal particle of arbitrary composition presents a promising new avenue for exploring specific interactions of a wider range of materials. Further, by replacing a sharp tip having a radius of curvature of 15–30 nm with a 3–10 μm spherical particle, the probe will more readily glide over surfaces rather than cut through them. This allows the study of a wider range of lubrication phenomena. An example is the recent study of the friction of cellulosic fibers, which show stick-slip behavior, dependence on load, and dependence on the presence of small amounts of high molecular weight polyelectrolytes.[34]

3. Fatiguing processes at a nanoscale have not been extensively studied. Processes leading to wear (material removal) are often initiated only after repeated loading and unloading. Because of the difficulties associated with characterizing subsurface material deformations or slight changes in atomic structure, current nanotribology research focuses primarily on either scratch testing, in which the surface is scored during first contact, or friction measurements made under no-wear conditions. An exception is Ref.[35] where fatigue-induced wear of mica was studied using AFM and modeled based on the friction energy dissipated and the number of sliding events. Their results represent one of the first efforts in relating friction to wear at the nanoscale. Nevertheless, long-term materials testing (hours to days) are impractical for these instruments; yet, the majority of macroscale tribological testing is of this nature.
4. The atomic and molecular structure of tribological surfaces and lubricants under shear plays a crucial role in tribology. Unfortunately, techniques to probe nanostructure spectroscopically or optically simultaneously with tribological probing are limited. This limits the ability to draw conclusive structure–property relations for all but the most idealized macroscopic surfaces, such as those described here.

These challenges present a host of opportunities in nanotribology and allied fields. Most of the advances discussed above were made within the last 15–20 yr, with a focus on the last ten. In this respect, nanotribology is by no means a mature field. With continued progress, further impact on our scientific and technological understanding of the nature of sliding surfaces and nanomechanics is expected.

REFERENCES

1. Oak Ridge National Laboratory; http://www.ornl.gov/css/Tribology/Tribology.htm.
2. Carpick, R.W.; Salmeron, M. Scratching the surface: fundamental investigations of tribology with atomic force microscopy. Chem. Rev. **1997**, *97*, 1163–1194.
3. Krim, J. Friction at the atomic scale. Sci. Am. **1996**, *275* (4), 74–80.
4. Granick, S. Soft matter in a tight spot. Phys. Today **1999**, *57*, 26–31.
5. Krim, J. Surface science and the atomic-scale origins of friction: what once was old is new again. Surf. Sci. **2002**, *500*, 741–758.
6. Handbook of Micro/NanoTribology, 2nd Ed.; Bhushan, B., Ed.; CRC Press: Boca Raton, 1999.
7. Hsu, S.M., Ying, Z.C., Ed. *Nanotribology, Critical Assessment and Research Needs*; Kluwer Academic Publishers: Dordrecht, 2003.
8. Gao, J.; Luedtke, W.D.; Gourdon, D.; Ruths, M.; Israelachvili, J.N.; Landman, U. Frictional forces and Amontons' law: from the molecular to the macroscopic scale. J. Phys. Chem. B **2004**, *108*, 3410–3425.
9. Persson, B.N.J. Sliding friction. Surf Sci. Rep. **1999**, *33*, 83–119.
10. Bhushan, B.; Israelachvili, J.N.; Landman, U. Nanotribology: friction, wear and lubrication at the atomic scale. Nature **1995**, *374*, 607–616.
11. Bhushan, B. Nanoscale tribophysics and tribomechanics. Wear **1999**, *225–229*, 465–492.
12. Johnson, K.L. *Contact Mechanics*; Cambridge University Press: Cambridge, 1985.
13. Enachescu, M.; van den Oetelaar, R.J.A.; Carpick, R.W.; Ogletree, D.F.; Flipse, C.F.J.; Salmeron, M. Observation of proportionality between friction and contact area at the nanometer scale. Tribol. Lett. **1999**, *7*, 73–78.
14. He, G.; Müser, M.H.; Robbins, M.O. Adsorbed layers and the origin of static friction. Science **1999**, *284* (5420), 1650–1652.
15. Van Alsten, J.; Granick, S. The origin of static friction in ultrathin liquid films. Langmuir **1990**, *6*, 876–880.
16. Dayo, A.; Alnasrallah, W.; Krim, J. Superconductivity-dependent sliding friction. Phys. Rev. Lett. **1998**, *80*, 1690–1693.
17. Mancinelli, C.M.; Gellman, A.J. Friction anisotropy at Pd(100)/Pd(100) interfaces. Langmuir **2004**, *20*, 1680–1687.
18. Hammerschmidt, J.A.; Gladfelter, W.L.; Haugstad, G. Probing polymer viscoelastic relaxations with temperature-controlled friction force microscopy. Macromolecules **1999**, *32*, 3360–3367.

19. Christenson, H.K.; Gruen, D.W.R.; Horn, R.G.; Israelachvili, J.N. Structuring in liquid alkanes between solid surfaces: force measurements and mean-field theory. J. Chem. Phys. **1987**, *83*, 1834–1841.
20. Reiter, G.; Demirel, A.L.; Granick, S. From static to kinetic friction in confined liquid films. Science **1994**, *263*, 1741–1744.
21. Yoshizawa, H.; Israelachvili, J. Fundamental mechanisms of interfacial friction. 2. Stick-slip friction of spherical and chain molecules. J. Chem. Phys. **1993**, *97*, 11300–11313.
22. Ruths, M.; Granick, S. Tribology of confined Fomblin-Z perfluoropolyalkyl ethers: role of chain-end chemical functionality. J. Phys. Chem. B **1999**, *103*, 8711–8721.
23. Raviv, U.; Glasson, S.; Kampf, N.; Gohy, J.-F.; Jérôme, R.; Klein, J. Lubrication by charged polymers. Nature **2003**, *425*, 163–165.
24. Heuberger, M.; Zäch, M.; Spencer, N.D. Density fluctuations under confinement: when is a fluid not a fluid? Science **2001**, *292*, 905–908.
25. Bhushan, B.; Li, X. Nanomechanical characterization of solid surfaces and thin films (invited). Int. Mater. Rev. **2003**, *48*, 125–164.
26. Elmustafa, A.A.; Stone, D.S. Indentation size effect in polycrystalline fcc metals. Acta Mater. **2002**, *50*, 3641–3650.
27. Li, X.; Bhushan, B. A review of nanoindentation continuous stiffness measurement technique and its applications. Mater. Charact. **2002**, *48*, 11–36.
28. Li, X.; Bhushan, B. Development of a nanoscale fatigue measurement technique and its application to ultrathin amorphous carbon coatings. Scripta Mater. **2002**, *47*, 473–479.
29. Klinger, L.; Rabkin, E. Theory of nanoindentation creep controlled by interfacial diffusion. Scripta Mater. **2003**, *48*, 1475–1481.
30. Kysar, J.W. Energy dissipation mechanisms in ductile fracture. J. Mech. Phys. Solids **2003**, *51*, 795–824.
31. Smith, R.; Christopher, D.; Kenny, S.D.; Richter, A.; Wolf, B. Defect generation and pileup during nanoindentation of Fe single crystals. Phys. Rev. B **2003**, *67*, 245405.
32. Schuh, C.A.; Nieh, T.G. A nanoindentation study of serrated flow in bulk metallic glasses. Acta Mater. **2003**, *51*, 87–99.
33. Patscheider, J. Nanocomposite hard coatings for wear protection. Mater. Res. Bull. **2003**, *28*, 180–183.
34. Zauscher, S.; Klingenberg, D.J. Friction between cellulose surfaces measured with colloidal probe microscopy. Colloids Surf. **2001**, *A178*, 213–229.
35. Kopta, S.; Salmeron, M. The atomic scale origin of wear on mica and its contribution to friction. J. Chem. Phys. **2000**, *113* (18), 8249–8252.

Natural Gas Hydrates

P. R. Bishnoi
Matthew A. Clarke
Department of Chemical and Petroleum Engineering, University of Calgary, Calgary, Alberta, Canada

INTRODUCTION

Gas hydrates are nonstoichiometric crystalline compounds that belong to a group of solids called clathrate. They are formed from mixtures of water and low molar mass gases at high pressures and low temperatures. Through hydrogen bonding, water molecules form a framework containing relatively large cavities, which can be occupied by certain gas molecules that get linked to the framework by van der Waals forces. The hydrate forming gases include light alkanes (methane to *n*-butane), carbon dioxide, hydrogen sulfide, nitrogen, and oxygen.

The first recorded observation of gas hydrates was made in 1823 when Humphrey Davy and Michael Faraday[1] bubbled chlorine gas through water. As the mixtures cooled, they noticed a solid material forming at temperatures above the normal freezing point of water. Throughout the remainder of the 19th century, much work was undertaken cataloging the various molecules that could form hydrates and the conditions at which each hydrate was stable.

The second phase of gas hydrates research began in the 1930s when Hammerschmidt[2] observed that gas hydrates were responsible for plugging natural gas pipelines. Following the discovery by Hammerschmidt,[2] Katz and coworkers[3,4] pioneered experimental investigations of the thermodynamics of gas hydrate formation for several natural gas systems. Also in this period, van der Waals and Platteeuw[5] developed the first thermodynamic model for the solid hydrate phase. This enabled development of computer-based predictive methods[6,7] to determine the incipient hydrate conditions. The 1960s saw another shift in gas hydrate research when natural gas hydrates (NGHs) (primarily composed of methane) were observed as a naturally occurring constituent of subsurface sediments in the giant gas fields of the Western Siberia basin.[8] In 1974, marine NGHs were directly observed when Soviet scientists discovered large hydrate nodules from the floor of the Black Sea. In recent years, there has also been speculation[9] that gas hydrates are present at the polar regions of Mars and on Europa. Interest in gas hydrates received additional impetus during the 1970s when rising energy costs and depletion of conventional hydrocarbon reserves pushed the search for new oil and gas into increasingly deeper offshore waters and arctic permafrost regions. Environmental concerns due to oil and gas activities in the cold waters of the arctic or deep offshore waters generated interest in understanding the impact of gas hydrate formation on the dynamics of gas/oil plumes. Exploitation of oil and gas from deep offshore waters also necessitated the development of kinetic inhibitors and flow modifiers for multiphase flow assurance in the offshore pipelines.

THE STRUCTURE AND PROPERTIES OF GAS HYDRATES

Gas hydrates constitute a class of solids in which small molecules occupy almost spherical holes in icelike lattices made up of hydrogen-bonded water molecules. This class of solids is known as clathrates. Technically speaking, clathrate compounds are characterized by the structural combinations of two substances that remain associated not through strong attractive forces, but because strong mutual binding of the molecules of one sort makes possible the firm enclosure of the other.

The Crystal Structure of Gas Hydrates

Gas hydrates can exist in one of four structures: structure I (sI),[10] structure II (sII),[11,12] structure H,[13] and a new, as of yet unnamed structure.[14]

All four structures contain the basic building block water cavity referred to as the 5^{12} cage. The notation 5^{12} indicates a crystal structure that contains 12 pentagonal faces. The 5^{12} cage can accommodate small guest molecules such as H_2S and CH_4.

In addition to the 5^{12} cage, there are two other commonly encountered cages: the $5^{12}6^2$ cage that can accommodate slightly larger molecules, such as CO_2 and C_2H_6 and the $5^{12}6^4$ cage that can accommodate even larger molecules such as propane and *n*-butane. The 5^{12} cage occurs in both sI and sII hydrates, the $5^{12}6^2$ cage is the large cage in the sI hydrate and the $5^{12}6^4$ cage is the large cage for the sII hydrate.

The sH hydrate is composed of the basic 5^{12} cage, as well as a $4^3 5^6 6^3$ cage and a $5^{12}6^8$ cage. The $4^3 5^6 6^3$ cage

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008060

Table 1 Geometry and occupying species of hydrate cavities

	Structure I	Structure II	Structure H
Cavity types	5^{12}, $5^{12}6^2$	5^{12}, $5^{12}6^4$	5^{12}, $4^35^66^3$, $5^{12}6^8$
Radius (Å)	3.91, 4.33	3.902, 4.683	3.91, 4.06, 5.71
Cages/unit cell	2, 6	16, 8	3, 2, 1
Co-ordination number	20, 24	20, 28	20, 20, 36
Crystal type	Cubic	Cubic	Hexagonal
Guest species	CH_4, C_2H_6, Xe, O_2, H_2S, CO_2	Ar, Kr, N_2, C_3H_8, *i*-C_4H_{10}, $(CH_2)_3O$	Butane[a], neopentane[a], benzene[a], cyclopentane[a], cyclohexane[a], cycloheptane[a], cyclo-octane[a], adamantane[a]

[a]Requires the presence of a ''helper'' molecule to stabilize the structure.
(From Refs.[10–13].)

contains three fairly strained square faces, six pentagonal faces, and three hexagonal faces. The $5^{12}6^8$ cage contains 12 pentagonal and 8 hexagonal faces. Structure H hydrates are double hydrates. Small guest molecules, such as xenon, methane, or hydrogen sulfide, occupy the two small cages of the sH hydrate while a larger molecule occupies the large sH cage. Structure I is a body centered cubic crystal, structure II is a diamond cubic crystal and structure H is a hexagonal crystal. Diagrams of the cages and the unit cell for structures I, II, and H are readily available.[10–13]

A comparison of the geometrical properties of structures I, II, and H is given in Table 1. The list of guest species contains some of the most common hydrate formers for each structure. Although both pure methane and pure ethane form structure I hydrates, it has recently been observed[15] that hydrates formed from mixtures of methane and ethane can actually form structure II.

The new, and currently unnamed, structure consists of alternating stacks of structure II and structure H hydrates. It has been synthesized by crystallizing choline hydroxide with tetra-*n*-propylammonium fluoride from aqueous solution.

Physical Properties of Gas Hydrates

In many respects, the acoustic, strength, thermal, and rheological properties of gas hydrates are similar to those of ice. However, there are a few properties, including the dielectric constant and the thermal conductivity, which differ significantly from that of ice. The vast majority of data available on the physical properties of gas hydrates is only for methane hydrates. Davidson[16] presented a comparison of the physical properties of hydrates with those of ice. Table 2 summarizes some of these properties.

A handful of researchers[17,18] have measured the hydration numbers (the ratio of the number of water molecules per molecule of hydrate former in a unit cell) for natural gas and methane and found that they lie between 5.9 and 8.2.

THERMODYNAMIC ASPECTS OF NGH FORMATION AND DECOMPOSITION

Phase Behavior of Hydrate Forming Systems

The phase behavior of NGHs has been extensively researched. The temperature, pressure, the composition of the gas, and the state of the aqueous solution determine the incipient conditions of gas hydrate formation. The most common method for representing phase behavior for gas hydrates is by a temperature versus pressure

Table 2 Comparison of ice and methane hydrate properties

Property	Ice	Methane hydrate
Dielectric constant at 273 K	94	58
Isothermal Young's modulus at 268 K (10^9 Pa)	9.5	8.4
Speed of longitudinal sound at 273 K		
Velocity (km/sec)	3.8	3.3
Transit time (μsec/ft)	80	92
Velocity ratio V_p/V_s at 272 K	1.88	1.95
Poisson's ratio	0.33	0.33
Bulk modulus (272 K)	8.8	5.6
Shear modulus (272 K)	3.9	2.4
Bulk density (g/cm^3)	0.916	0.912
Adiabatic bulk compressibility at 273 K 10^{-11} Pa	12	14
Thermal conductivity at 263 K (W/m-K)	2.23	0.49 ± 0.02

(From Ref.[16].)

phase diagram. Makogan,[8] Berecz and Balla-Achs,[19] and Sloan[20] have many examples of phase diagrams for various hydrate forming systems. For a single gas, the Gibbs phase rule suggests that for the binary system of water and a hydrate former gas, there is only one degree of freedom for three phases to coexist in equilibrium. In other words, for a given temperature there is only one pressure for the equilibrium of three phases of solid hydrate, liquid water, and gas phases. For the single gas, the quadruple point has no degree of freedom.

For gases, such as methane, which are supercritical at hydrate forming temperatures, there is one quadruple point, as indicated by point Q1 in Fig. 1. At this point, ice, liquid water, gas and hydrate are in equilibrium. For gases that are subcritical at hydrate forming temperatures, such as ethane,[20,21] there are two quadruple points (Q1 and Q2 in Fig. 2). While Q1 lies at approximately the freezing point of water, Q2 is at approximately the intersection of the hydrate–water–gas three-phase equilibrium curve with the vapor pressure curve. At this latter point, liquid water, gas, hydrate, and liquid hydrate former are all in equilibrium. As seen in Fig. 2, the hydrate formation pressure increases sharply above this quadruple point.

Of particular interest to those in the natural gas industry is the phase diagram of hydrate systems in the presence of inhibitors. Fig. 3 shows the phase diagram for methane hydrates in the presence of methanol[22] and a NaCl and KCl mixture.[23] The solid line is the three-phase equilibrium curve for methane in pure water. As seen from Fig. 3, forming hydrates in the presence of either an alcohol[22] or salt[23] increases the pressure required for gas hydrate formation, at a given temperature.

The equilibrium behavior of structure H hydrates has only recently been studied.[24–27] Fig. 4 shows the incipient hydrate formation conditions of a mixture of methane, carbon dioxide, and neo-hexane, as measured by Servio et al.[27]

Formation of gas hydrates in porous media

Experimental data on gas hydrate formation in porous medium are very sparse. Ershov and Yakushev[28]

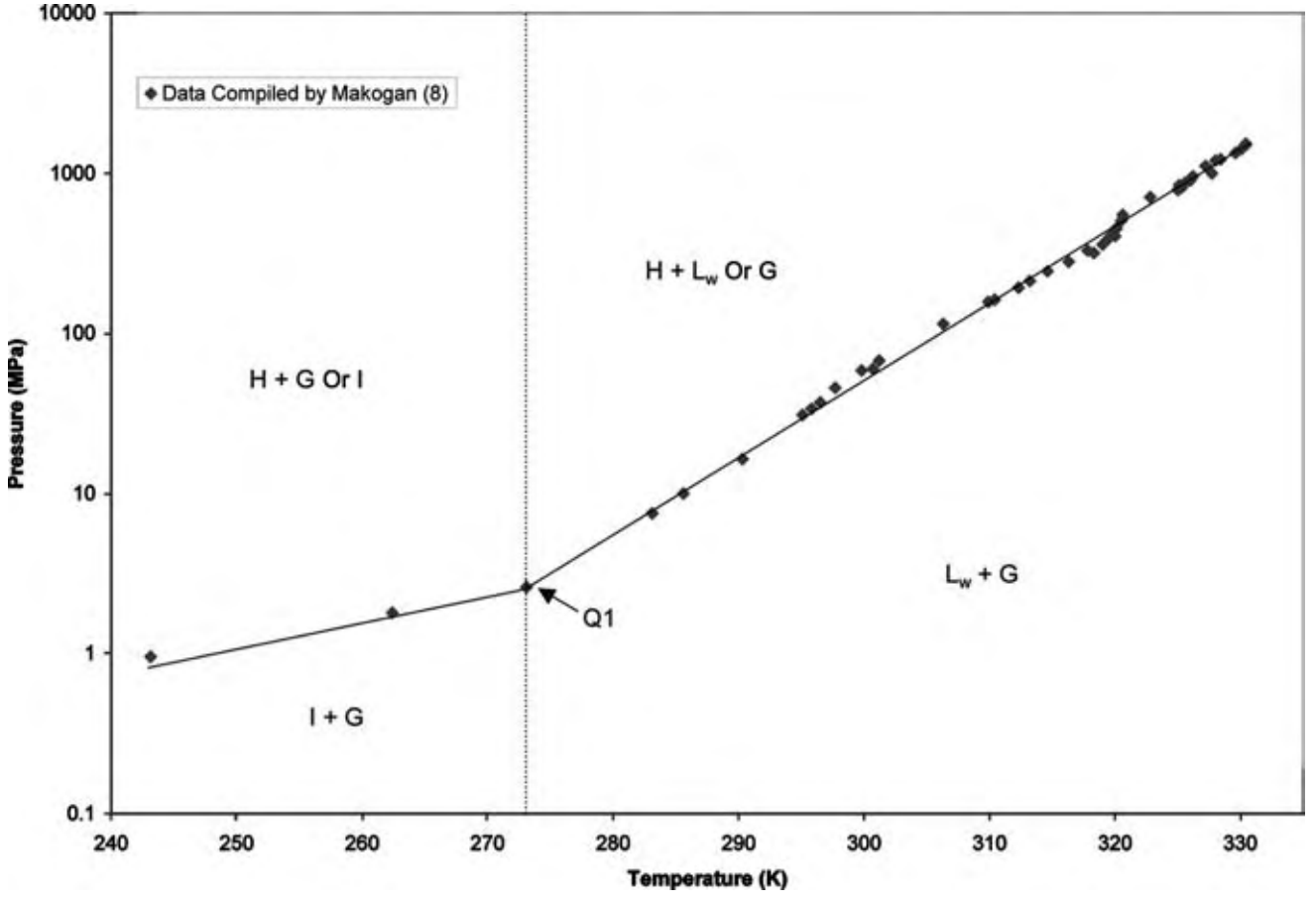


Fig. 1 Incipient hydrate forming conditions for pure methane. The solid line represents the three-phase equilibrium for either ice–hydrate–gas ($T < 273.15$ K), or liquid water–hydrate–gas ($T > 273.15$ K). *(View this art in color at www.dekker.com.)*

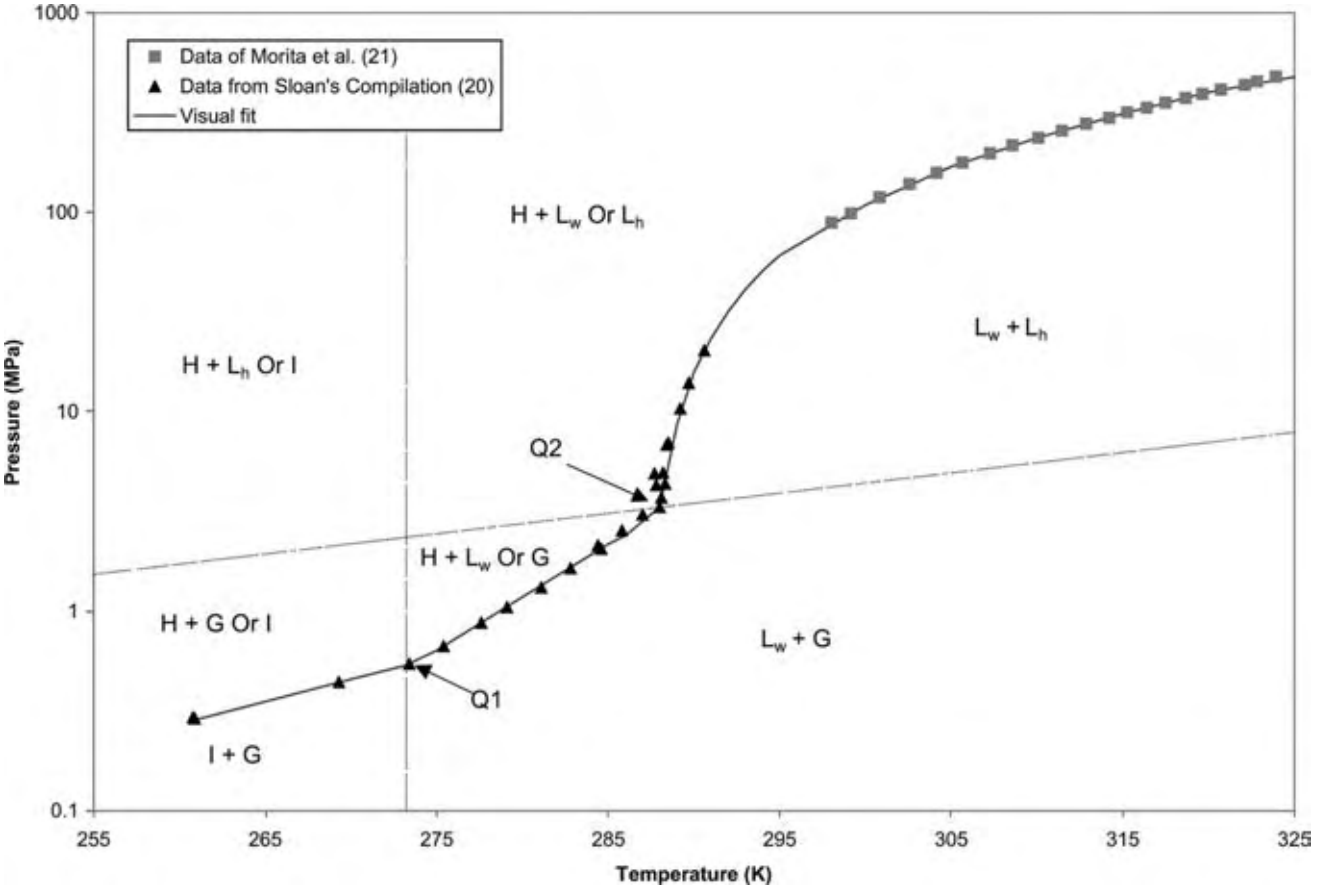


Fig. 2 Incipient hydrate forming conditions for pure ethane. The solid line represents the three-phase equilibrium for either ice-hydrate-gas ($T < 273.15\,\text{K}$), or liquid water–hydrate–gas ($273.15\,\text{K} < T < 288.1\,\text{K}$) or liquid water–hydrate–liquid ethane ($T > 288.1\,\text{K}$). *(View this art in color at www.dekker.com.)*

conducted experiments on rocks freezing by decomposing gas hydrates. Handa and Stupin[29] measured the dissociation characteristics of methane and propane in 70-Å-radius silica gel pores. In both cases, the equilibrium pressures were between 20 and 100% higher than those for hydrates in free water, as seen in Fig. 5. Ahmadi, Chuang, Smith[30] measured equilibrium pressures over a range of temperatures for dissociation to free gas and liquid water of various sI (methane, ethane, and carbon dioxide) hydrates and one sII (propane) hydrate confined in silica gel pores of nominal radii 3–7 nm. Each of these porous media contained a broad distribution of pore radii. The observed pressures again were higher than those in free water.

General Computational Procedures

Incipient point computations

At the three-phase equilibrium condition, the chemical potential (or fugacity) of water in the hydrate phase will be equal to those in the liquid solution (or ice phase, depending on the temperature) and in the gas phase. For computation, it is necessary to have a model to compute the chemical potential in each phase. For the vapor phase, an equation of state is most commonly used. The model of van der Waals and Platteeuw[5] is generally used to calculate the chemical potential (or fugacity) of water in the hydrate phase. The model is given by

$$\frac{\mu_{\text{w}}^{\text{MT}} - \mu_{\text{w}}^{\text{H}}}{RT_{\text{f}}} = \sum_{m=1}^{2} \nu_m \ln\left(1 + \sum_{j=1}^{\text{NH}} C_{mj} f_j\right) \qquad (1)$$

In Eq. (1), the Langmuir constant, C_{mj}, is normally computed from a Kihara potential function.[6]

For calculating the chemical potential of water in the liquid solution, or ice phase, Holder, Corbin, Papadopoulos[7] generated chemical potential, enthalpy, and heat capacity functions for gas hydrates at temperatures between 150 and 300 K and derived

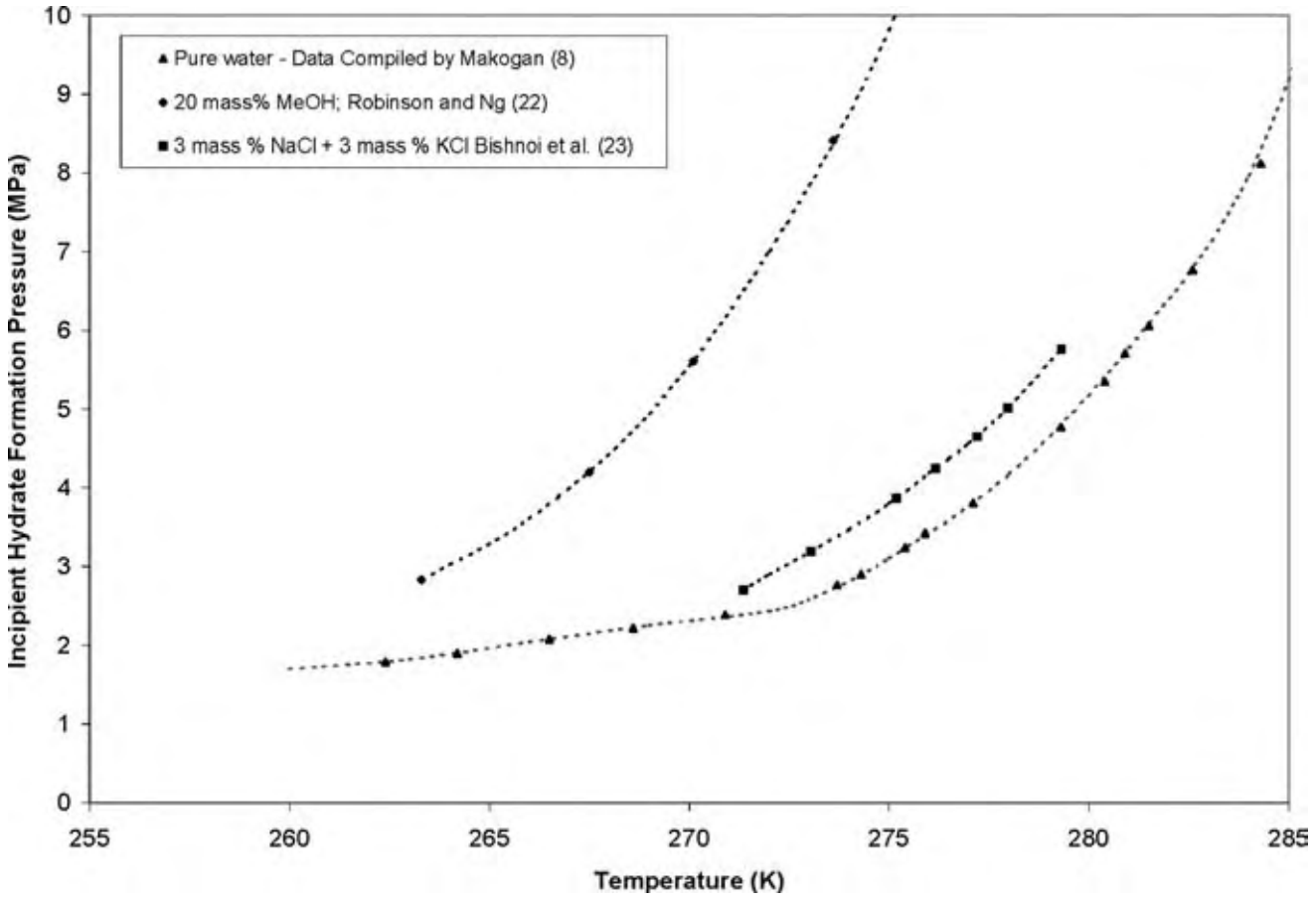


Fig. 3 Three-phase equilibrium curves showing incipient hydrate forming conditions for methane in the presence of various inhibitors. The dashed lines are visual fits of the experimental data. *(View this art in color at www.dekker.com.)*

the following relationship:

$$\frac{\mu_w^{MT} - \mu_w^{L}}{RT_f} = \frac{\Delta\mu_w^{o}}{RT_o} - \int_{T_o}^{T_f} \frac{\Delta h_w^{MT-L_o}}{RT^2} dT + \int_{P_o}^{P} \frac{\Delta v_w^{MT-L_o}}{RT_f} dP - \ln a_w \qquad (2)$$

At equilibrium, the right hand sides of Eqs (1) and (2) are equal. In Eq. (2), the activity of water may be calculated from either an equation of state or from a liquid solution model.

Englezos and Bishnoi[31] developed a model to predict the formation conditions of gas hydrates for sparingly soluble gases in solutions of aqueous electrolytes. More recently, Clarke and Bishnoi[32,33] proposed an equation of state for high-pressure aqueous systems, which may contain soluble gases, single or mixed electrolytes, and/or a second nonaqueous solvent. The equation of state was successfully used in conjunction with the model of van der Waals and Platteeuw[5] to calculate the formation conditions of hydrates formed from pure and mixed gases, in the presence of single salts, mixed salts, and salt/alcohol systems. An alternative approach to modeling hydrate formation in the presence of salts and alcohols was undertaken by Jager, Ballard, Sloan[34] who used the Bromley activity model to compute the activity of water.

Activity models have also been used to compute incipient hydrate conditions in the presence of polymers. Englezos and Hall[35] presented a predictive model for the calculation of the incipient hydrate formation conditions in a hydrocarbon–water–polymer system. The activity of water in the polymer–water solution is computed using the UNIFAC model.

In addition to salts, alcohols, and polymers, the presence of a porous medium will also have an inhibiting effect on the formation of gas hydrates. Clarke, Pooladi-Darvish, Bishnoi[36] derived the following expression for the activity of water in a pore of radius r, and used it along with the model of van der Waals and Platteeuw[5] to model the incipient conditions for

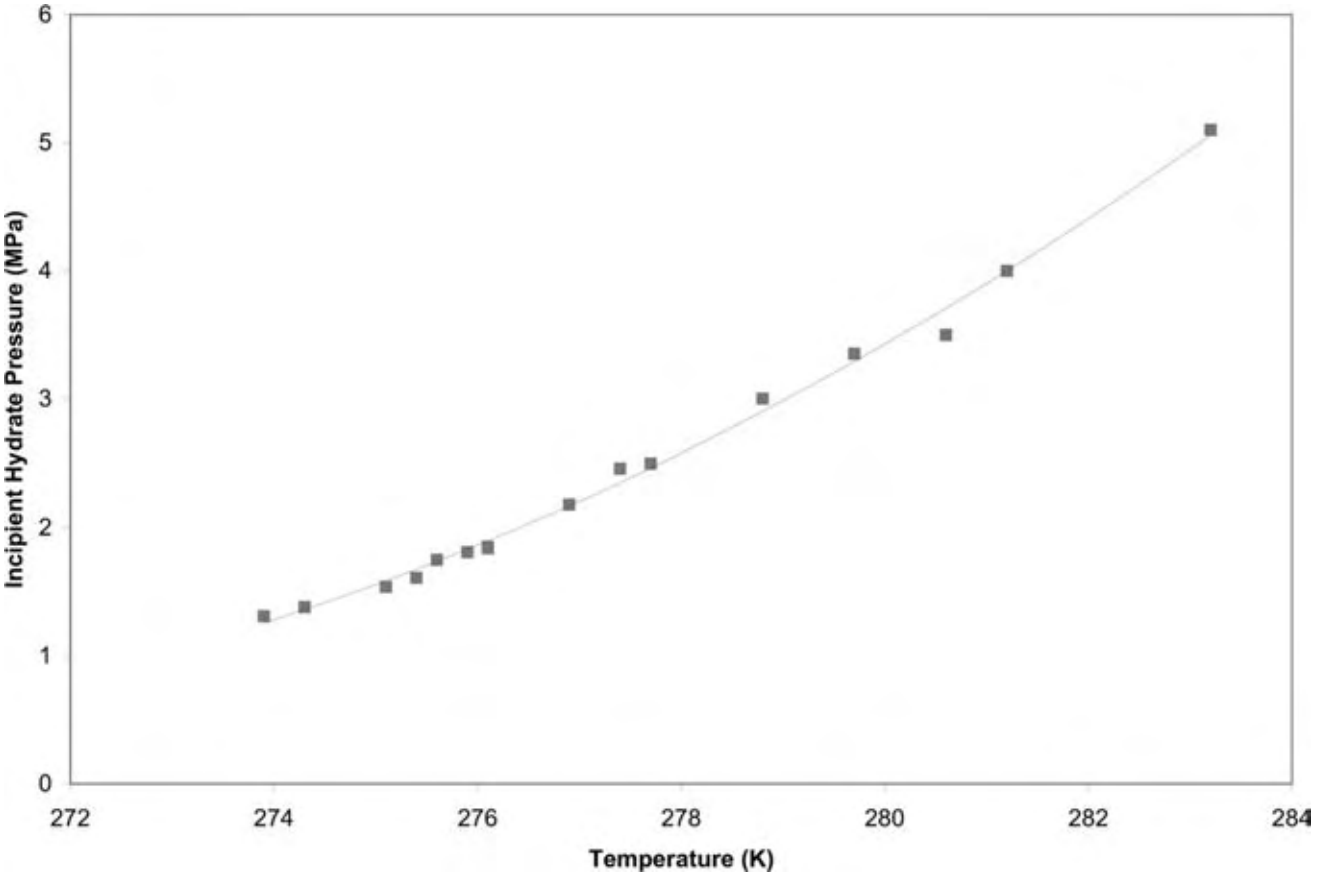


Fig. 4 Three-phase equilibrium curves showing incipient hydrate forming conditions for sH hydrates formed from a mixture of CH_4/CO_2/neo-hexane. Points are the experimental data of Servio and Englezos and the solid line is a visual fit. (From Ref.[27].) *(View this art in color at www.dekker.com.)*

methane and propane hydrate formation in 70-Å silica gel pores.

$$\ln a_w = \frac{-2\sigma v_l}{rRT_f}\cos\theta \quad (3)$$

This equation was used with Eqs. (1) and (2) to model the formation of methane and propane hydrates in 70-Å silica gel pores.

Phase fraction calculations

The calculations described in the earlier section gave the incipient point only. At the incipient point, the amount of hydrate formed is infinitesimally small. At conditions that depart from the incipient pressure and temperature, it is possible to estimate the phase fractions (or amounts) of hydrates that have formed. The pioneering work in this field was that of Bishnoi et al.,[37] in which solid phases of structure I and II gas hydrates were included in multiphase flash calculations. The multiphase equilibrium calculations are based on a Gibbs free energy minimization methodology developed by Gupta, Bishnoi, Kalogerakis[38] Recently, Ballard and Sloan[39,40] made minor modifications to the methodology of Gupta, Bishnoi, Kalogeratis[38] and reported calculations on the hydrate phase fractions for some systems.

KINETICS OF GAS HYDRATE FORMATION AND DECOMPOSITION

Hydrate Formation

The process of hydrate formation is a heterogeneous process having similarities with crystallization processes. The difference in the two processes is that in the hydrate formation the solute (hydrate former) is supplied from another fluid phase (gas or liquid) to the aqueous liquid phase where it combines with water and crystallizes as solid hydrate. Also, the process is generally conducted at high pressures. It is because of these factors that the kinetics of hydrate formation and decomposition has been studied by only a few researchers. The similarities lie in the fact that the

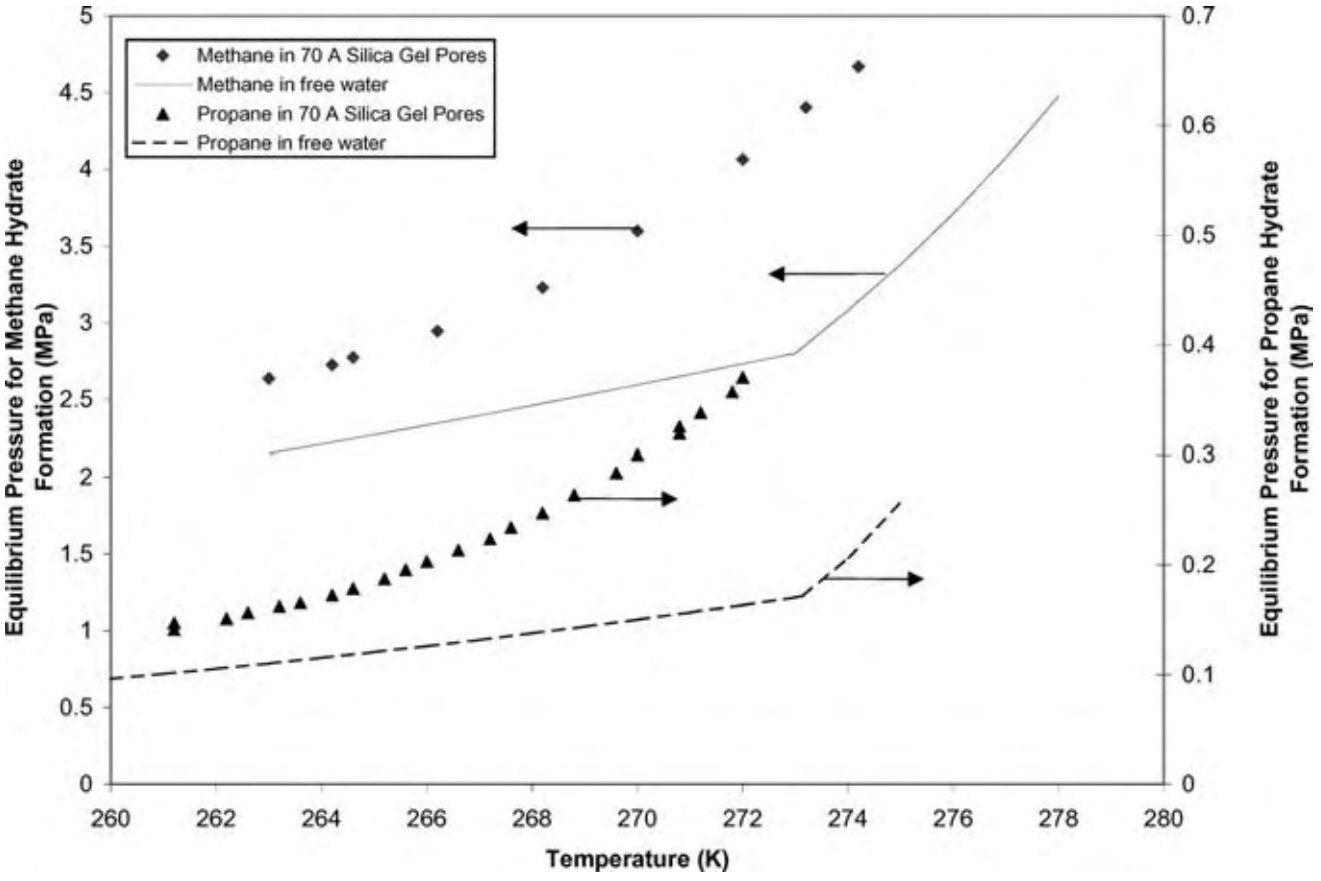


Fig. 5 Three-phase equilibrium curves showing incipient hydrate forming conditions for methane and propane hydrates formed in 70-Å silica gel pores. Points are the data from Handa and Stupin. (From Ref.[29].) *(View this art in color at www.dekker.com.)*

hydrate formation, like crystallization, consists of two different phenomena—nucleation and growth. A crystal nucleus that reaches the thermodynamically stable critical size grows to form crystals. The kinetics of hydrate formation thus requires studying these two different phenomena.

Nucleation

Experimental information on gas hydrate nucleation at a microscopic level is almost nonexistent or at best very limited. Most of the studies on the hydrate nucleation are based on a macroscopic approach. Although there are some differences, gas hydrate nucleation has similarities with salt crystal nucleation.[41] For the nucleation to occur, supersaturation of the aqueous solution with the hydrate former gas is required. The supersaturation is necessary to overcome the free energy barrier for creating a new surface of a solid hydrate nucleus. The degree of supersaturation or the driving force for nucleation may be defined in terms of difference in the chemical potential or the fugacity of a hydrate former in the solution and that at the equilibrium state. The state of supersaturation is a metastable state in which the nucleation processes consisting of the formation and breakage of gas–water clusters occur. These processes continue to eventually form critical size particles that are thermodynamically stable. Formation of the critical size particles leads to nucleation of the solid hydrate phase. The hydrate nucleation may occur in the liquid solution where a sufficient degree of supersaturation exists. For a non-stirred solution, the probability of reaching the desired supersaturation is highest at the gas (or liquid hydrate former) liquid (water) interface. For a well-stirred system the desired supersaturation may reach anywhere in the liquid solution. Englezos et al.[42] estimated the thermodynamic critical size of the hydrate nucleus, assuming primary homogeneous nucleation, to be of the order of 10^{-8} m.

Hydrate nucleation, like salt crystallization, is intrinsically a stochastic process.[43] For a deterministic process, a set of consistent input variables will produce a consistent reproducible result. Hydrate nucleation being a stochastic process will not produce a single set of outputs, but a broad statistical distribution of results owing to the presence of a probabilistic element

in the nucleation mechanism. Because the stochastic characteristics of a homogeneous nucleation event may be masked by heterogeneous influences, the intrinsic kinetics of hydrate nucleation processes is difficult, if not impossible, to measure. The heterogeneous influences may be due to active foreign bodies, surface effects, mechanical perturbations and agitation, thermal history of water, and thermal shocks. Experimentally, the random nature of the process may be dominated by these influences, producing a seemingly deterministic behavior that leads to repeatable experimental results. In addition to the heterogeneous influences, the degree of supersaturation, i.e., driving force and local supersaturation affect the rate of hydrate nucleation. Hydrate nucleation kinetics, in the macroscopic approach, has generally been studied experimentally by observing induction time, defined as the duration of the metastable (supersaturation) state. Modeling of experimental data on induction time is reported in the literature.[41,44] The parameters in such models are equipment dependent and hence cannot be used for other equipment and industrial applications.

Gas Hydrate Crystal Growth

Hydrate crystal growth, unlike hydrate nucleation, is not a stochastic process. Hence, the intrinsic kinetics of hydrate growth can be studied experimentally and modeled. The intrinsic rate constant obtained for this process can be independent of the equipment used for its determination, if determined carefully with proper experimental design, and thus can be used for industrial or other applications.

Experimental information on the kinetics of gas hydrate growth on molecular level is either nonexistent or at best extremely limited. Most of the studies on the hydrate formation kinetics are based on a macroscopic approach. Although hydrate growth is a heterogeneous process, a few researchers have studied and modeled hydrate growth by considering it a homogeneous process.[45,46] The first work to recognize and quantitatively account for the fact that the growth of gas hydrate particles is a heterogeneous crystallization process was that of Bishnoi and coworkers.[42] They formulated a macroscopic mechanistic model based upon crystallization and mass transfer theories. They experimentally studied the hydrate formation in bulk water using a semi-batch high-pressure stirred reactor by monitoring gas consumption during its dissolution in water and consumption for hydrate formation. The experiments were conducted so that the heat and mass transfer effects around a hydrate particle are eliminated. Thus, the measured rate constant is the intrinsic rate constant, which is not system dependent. The data, obtained after the appearance of hydrate nuclei, were analyzed to formulate the model. The studies were undertaken for hydrates of methane, ethane, and their mixtures. Recently, they completed a study on carbon dioxide hydrate formation by installing a particle size analyzer in the high-pressure reactor.[47] The rate model is formulated by considering the intrinsic step of hydrate growth as the ''reaction'' at the hydrate solid–liquid interface. The reaction is a combination of steps for the creation of water structure near the interface, the incorporation of guest (hydrate former) by adsorption in the lattice, and subsequent stabilization of the hydrate structure framework by hydrogen bonding and van der Waals forces. The intrinsic model for the rate of hydrate growth per particle is given by:

$$\left(\frac{dn_H}{dt}\right)_p = K^* A_p (f - f_{eq}) \tag{4}$$

where

$$\frac{1}{K^*} = \frac{1}{k_r} + \frac{1}{k_d} \tag{4a}$$

In the above model, the driving force is given by $(f - f_{eq})$, which is the difference between the fugacity of the dissolved gas and its three-phase equilibrium value. In Eq. (4a), k_r is the intrinsic rate constant for the hydrate particle growth reaction and k_d is the mass transfer coefficient around the particle. If the experiments are carried out under conditions such that heat and mass transfer resistances around the particle are eliminated, then $k_d \gg k_r$ and $K^* \approx k_r$. The intrinsic rate constants K^*, for methane, ethane, and carbon dioxide are reported in Table 3.

The heterogeneous nature of gas hydrate formation was also acknowledged by Herri, Gruy, Cournil,[48] who developed a new model of gas hydrate crystallization and an experimental technique that uses in situ turbidimetry measurements to study the crystallization of methane hydrate.[49]

A handful of studies have investigated the effects of additives, such as salts and surfactants, on the kinetics of hydrate formation.[50–52]

Gas Hydrate Decomposition

Hydrate crystal decomposition, like the hydrate growth, is a deterministic process. Hence, the process is amenable to study experimentally and modeling. Although some studies have been undertaken at the molecular level, most of them on the hydrate decomposition kinetics are based on a macroscopic approach. The hydrate decomposition is a heterogeneous process where liquid water and gas are released as the solid

Table 3 Intrinsic rate constants for hydrate formation for CH_4, C_2H_6, and CO_2 hydrates

Temperature (°C)	K^a for CH_4 (mol/m^2 s MPa)	K^a for C_2H_6 (mol/m^2 s MPa)	K^a for CO_2 (mol/m^2 s MPa)
1.0	0.65×10^{-5}	0.12×10^{-5}	6.42×10^{-3}
3.0	0.55×10^{-5}	0.11×10^{-5}	4.82×10^{-3}
4.0	—	—	3.21×10^{-3}
5.0	—	—	—
6.0	0.57×10^{-5}	0.13×10^{-5}	—
6.5	—	—	5.27×10^{-3}
9.0	0.58×10^{-5}	0.14×10^{-5}	—

[a]Rate constants for CH_4 and C_2H_6 are currently being re-evaluated using an in situ particle size analyzer.
(From Refs.[42,47].)

shrinks due to its decomposition. Bishnoi and coworkers[53–57] experimentally studied the hydrate decomposition in bulk water using a semibatch high-pressure stirred reactor by monitoring gas released during the decomposition. The experiments are conducted so that the heat and mass transfer effects around a hydrate particle are negligible. Thus, the measured rate constant is the intrinsic rate constant, which is not system dependent and thus, can be used for industrial or other applications.

In their earlier work,[53] the experiments were conducted without determining the particle size analysis in the water slurry in the reactor. Later, they installed a particle size analyzer externally to the reactor to obtain the data with the particle size analysis.[54–56] Recently, they installed a particle size analyzer in the high-pressure reactor for their experiments.[57]

To obtain the intrinsic rate constant from the data, Kim et al.[53] developed a mechanistic model given below:

$$\left(\frac{dn_H}{dt}\right)_p = -K_d A_p \left(f_{eq} - f_g^V\right) \quad (5)$$

The model is formulated on the premise that the decomposing hydrate particle is surrounded by a cloud of the product gas; hence the driving force for the decomposition process is expressed in terms of the fugacity difference given in Eq. (1). The process of decomposition possibly involves (1) destruction of the clathrate host (water) lattice at the surface of the particle, and (2) and desorption of the guest (hydrate former) molecules from the surface. The particle size distribution was incorporated in the calculations for the determination of the intrinsic rate constants.[54–57] The following Arrhenius type equation is used to represent the effect of temperature on the intrinsic rate constant:

$$K_d = K_d^o \exp\left(-\frac{\Delta E}{RT}\right) \quad (6)$$

The intrinsic rate constant, K_d^o, and activation energies, ΔE, for CH_4, C_2H_6, and CO_2 hydrate decomposition are given in Table 4. For the decomposition of hydrates formed from mixtures of methane and ethane, Clarke and Bishnoi[56] showed that the total rate of decomposition is equal to the sum of the decomposition rates of the individual species.

PRACTICAL ISSUES AND APPLICATIONS

Industrial Issues

It is of interest to avoid hydrate formation or modify its flow characteristics to circumvent the problem of plugging natural gas pipelines or process equipment. Interest in gas hydrate research also stems from the planned activities to exploit the huge natural deposits of gas hydrates as an energy resource.

Flow assurance

In many pipelines, the temperature and the pressure conditions that are encountered place the flowing fluid well within the hydrate stability envelope. Experts estimate that controlling and preventing hydrate formation, or ''flow assurance,'' costs industry more than $100 million per year.[58] The problem is extremely severe in off-shore pipelines. In fact, the replacement of hydrate plugged flow lines in deep water is estimated to cost roughly $1

Table 4 Intrinsic rate constants, K_d^o, and activation energies, ΔE, for methane, ethane, and CO_2 hydrate decomposition

Gas	K_d^o (mol m^{-2} Pa^{-1} sec^{-1})	ΔE (kJ/mol)
CH_4	3.6×10^4	81
C_2H_6	2.56×10^8	104
CO_2	1.83×10^8	103

(From Refs.[54,55,57].)

million per mile.[59] Conventional methods of preventing hydrate formation in pipelines are to process the petroleum fluids, typically by heating the fluid, water dew point control through moisture removal, or to inject thermodynamic inhibitors so that the operating conditions of the pipelines lie outside the hydrate stability envelopes. These may be called the ''hydrate avoidance'' methods. More recently, kinetic methods of delaying hydrate formation and hydrate flow modifiers have been developed. The methods, based on modifying the flow characteristics of hydrates seem to be gaining popularity with industry, especially for off-shore applications.

Of the above-mentioned techniques, thermodynamic inhibitors, which include alcohols, salts, and glycols, are by far the most prevalent. For example, adding methanol to a natural gas will shift the equilibrium conditions so that a higher pressure is required to form hydrates, at a given temperature, as illustrated for methane in Fig. 3. Methods for estimating the saturation water content of natural gases and amounts of methanol or glycol required to suppress hydrate formation are discussed by Katz,[3] Sloan,[20] and Campbell.[60] Current practice for the estimations is to use computer software based on phase equilibrium calculations.[61]

Kinetic inhibitors are typically water-soluble polymers or copolymers that delay hydrate nucleation and/or growth. An inhibitor molecule slows crystal growth by either adsorbing on to the growth sites on the crystal surface, or by fitting into the crystal lattice. Antiagglomerants are designed to specifically interact with the growing hydrate crystal surface. These inhibitors permit hydrates to form but inhibit agglomeration, deposition, and plugging.[62]

Natural gas recovery from hydrate deposits

Fig. 6 shows the worldwide distribution of the currently known hydrate deposits.[63] It can be seen that the majority of the deposits are located offshore, either in subduction zones, passive margins, or offshore hydrocarbon deposits. Methane hydrates can exist in oceanic sediment where the ocean is at least 300 m deep. In the permafrost regions, methane hydrate is present at depths between 150 and 2000 m below the surface. It is interesting to note from Fig. 6 that NGH deposits are distributed worldwide quite evenly unlike conventional oil and gas deposits.

On the Earth, the total amount of carbon sequestered in offshore and onshore hydrate deposits is believed to be larger than all other carbon sources combined. Fig. 7 shows the relative sizes of organic carbon reserves. Of the estimated 18,777 gigatons of carbon, 10,000 gigatons are believed to be in the hydrate form.[64]

Currently, the techniques to recover natural gas from in situ hydrate deposits are in their infancy. Possible techniques include dissociating the in situ hydrates by pressure reduction, heating, or solvent injection. Two test wells were drilled for characterization and production testing in the Mackenzie Delta region of Canada, by international efforts in 1998 and 2002. The results of the first test well have been reported.[65] The release of production test results from the second well is expected. On 25 August 2005 from the Geological Survey of Canada.[66]

Environmental Issues

Seafloor stability and global warming

As oil and gas exploration extends into progressively deeper waters, the potential hazard posed by gas hydrates to operations is gaining increasing recognition. Hazards can be considered as arising from two possible events: (1) the release of high-pressure gas trapped below the hydrate stability zone, or (2) the destabilization of in situ hydrates. A major issue is how gas hydrates alter the physical properties of sediment. The link between seafloor failure and gas

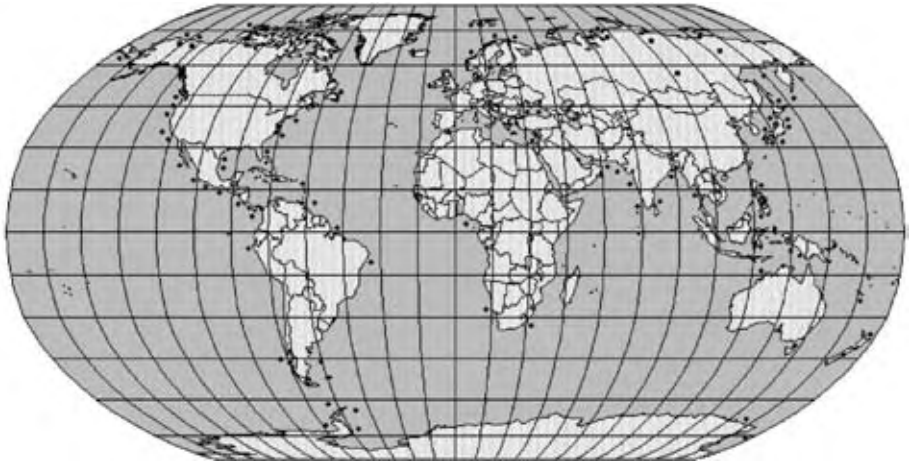


Fig. 6 Worldwide deposits of gas hydrates. Circles are offshore deposits and squares are onshore deposits. (From Ref.[63].) *(View this art in color at www.dekker.com.)*

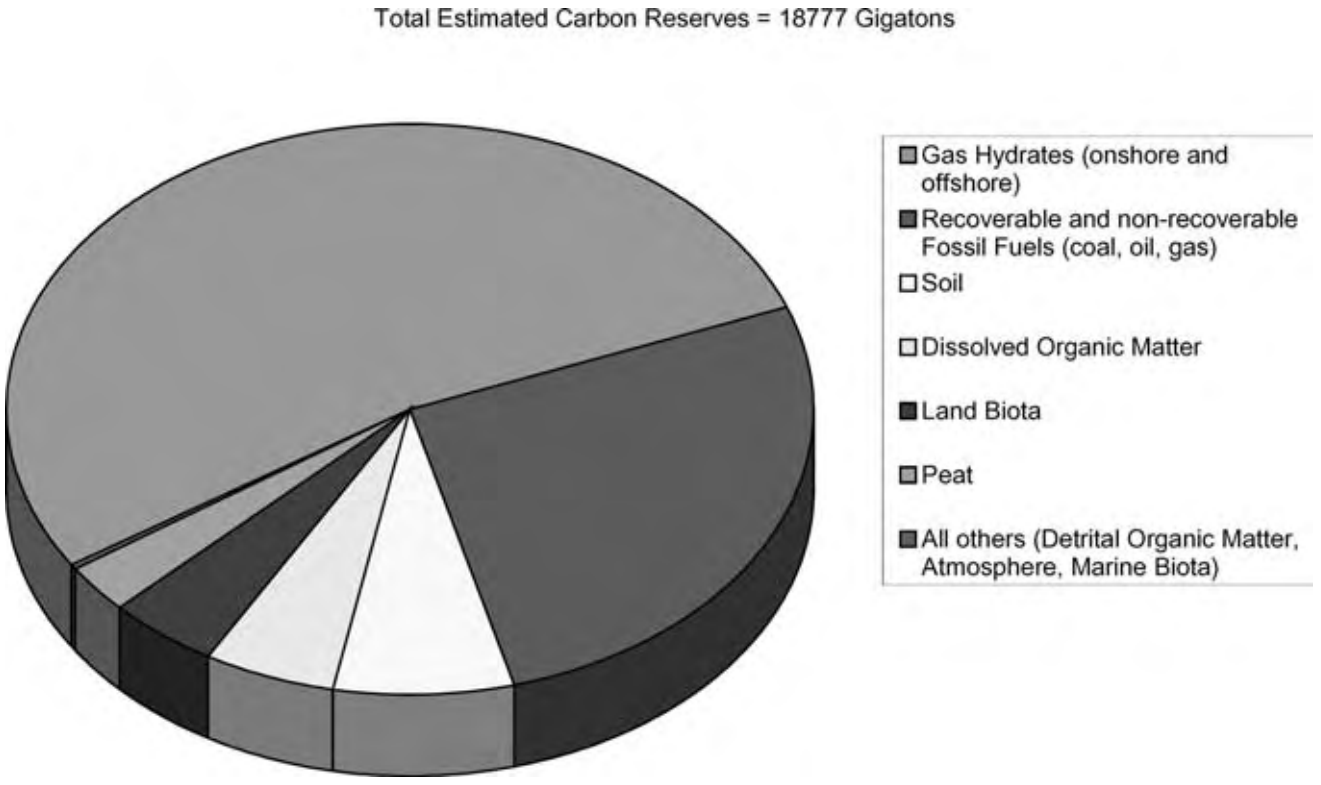


Fig. 7 Relative sizes of organic carbon reserves. (From Ref.[64].) *(View this art in color at www.dekker.com.)*

hydrate destabilization has been well established, especially with respect to the previous glacial–interglacial eustatic sea-level changes. Sea slope failure, as a result of gas hydrate decomposition, is considered to pose a significant hazard to underwater installations, pipelines, and cables, and, in extreme cases, to coastal populations through the generation of tsunamis.

As a greenhouse gas, methane is roughly 10 times more potent than CO_2. Over geologic time scales, there is evidence pointing to periodic, large releases of methane into the atmosphere. During formation of large polar ice sheets the sea level falls, thereby reducing the pressure on the ocean margin gas hydrates. It has been suggested that the shallower gas hydrate deposits became unstable and released methane into the atmosphere, which contributed to warming and the ending of the Ice Age.[67] On the other hand, the presence of increased amounts of CO_2 in the atmosphere due to human activities is likely to lead to global warming, which could increase surface temperatures. This, in turn, could release methane from hydrate deposits in ocean sediments or permafrost regions leading to further global warming.

Deep-water well blowouts

In the event of a deep-water well blowout, one of the environmental concerns is whether oil will surface and if so, where, when, and what will be the thickness of the oil slick. In the high-pressure and low-temperature conditions encountered in deep water, the gases are likely to form hydrates. As the density of hydrates is similar to that of oil, the conversion of gas into hydrates has a significant impact on the behavior of the jet/plume due to the alteration of the buoyancy. This phenomenon was first studied experimentally by Maini and Bishnoi,[68] later modeled by Topham,[69] and more recently by Chen and Yapa,[70] who also provide a summary of other important experimental work in this area. Chen and Yapa incorporated kinetic models of hydrate growth[42] and hydrate decomposition[53] in their plume model.

Applications of gas hydrates

Gas hydrates for natural gas storage

Natural gas could possibly be stored at low pressures and temperatures in the form of a hydrate. The relative density of gas in the hydrate lattice exceeds its liquid density. According to the calculations of Parent,[71] a natural gas with a volume of $4.42\,m^3$ at 15°C and atmospheric pressure only needs $0.028\,m^3$ volume for storage in the hydrate state. Thus, only 1/156 of the volume in the free state is needed. Rogers and Zhong[52] conducted a study on storing natural gas

in hydrate form in which they found that adding surfactants greatly increased the rate of hydrate formation in storage vessels.

Transportation of natural gas

Gudmundsson and Borrehaug[72] have presented results from a study to determine the feasibility of transporting natural gas in the frozen hydrate form. A NGH transportation chain comprises hydrate production, marine transport, and regasification elements. Even after a conservative economic analysis, it was found that the capital cost associated with an NGH chain will be 28% lower than the equivalent LNG chain. Takaoki et al.[73] found that the most suitable way to store hydrate for transportation was as small pellets.

Separation processes involving gas hydrates

Knox et al.[74] investigated using gas hydrates for the desalination of seawater. In this process, pure water is separated from its solution in a solid form by a reaction with certain agents to form a gas hydrate. The hydrate crystals are then separated from the brine, washed, and decomposed into propane gas and pure water.

Ngan and Englezos[75] formed propane hydrates in pulp mill effluents to recover water. Another application of gas hydrates in separations was seen in the study of Purwanto et al.,[76] which sought to investigate the concentration of coffee solutions through gas hydrate formation.

CONCLUSIONS

Since their discovery in the early 19th century, gas hydrates have gone from being merely a laboratory curiosity to a serious problem for the natural gas industry to potentially becoming the largest source of methane. The emerging gas hydrate technologies have the potential not only to provide a huge source of methane, but may also one day be a means for natural gas storage and transportation and for various separations. However, in order to shift to these processes from the conceptual stage to becoming commercially feasible, it is still necessary to further enhance current understandings of hydrate science and engineering.

NOMENCLATURE

a	Activity
A	Surface area
C	Langmuir constant
f	Fugacity
h	Enthalpy
K^*	Intrinsic rate constant for hydrate formation
K_d	Intrinsic rate constant for hydrate decomposition
n	Moles
NH	Number of hydrate formers
P	Pressure
r	Pore radius
R	Universal gas constant
T	Temperature
v	Molar volume

Greek Letters

μ	Chemical potential
ν	Number of cavities per water molecule
σ	Interfacial energy
θ	Wetting angle

Subscripts and Superscripts

eq	Three-phase equilibrium conditions
f	Hydrate formation conditions
g	Gas phase
H	Hydrate or in hydrate
L	Liquid solution or phase
MT	Empty lattice
o	Standard state
p	Particle
V	Vapor phase
w	Water

REFERENCES

1. Faraday, M.; Davy, H. On fluid chlorine. Phil. Trans. R. Soc. Lond **1823**, *113*, 160–165.
2. Hammerschmidt, E.G. Formation of gas hydrates in natural gas transmission lines. Ind. Eng. Chem. Res. **1934**, *26*, 851–857.
3. Katz, D.L. *Handbook of Natural Gas Engineering*; McGraw Hill: New York, 1959.
4. Katz, D.L.; Lee, R.L. *Handbook of Natural Gas Engineering*, 2nd Ed.; McGraw Hill: New York, 1991.
5. van der Waals, J.H.; Platteeuw, J.C. Clathrate solutions. Adv. Chem. Phys. **1959**, *2* (1), 1–45.
6. Parrish, W.R.; Prausnitz, J.M. Dissociation pressures of gas hydrates formed by gas mixtures. Ind. Eng. Chem. Proc. Des. **1972**, *11*, 26–36.
7. Holder, G.D.; Corbin, G.; Papadopoulos, K.D. Thermodynamic and molecular properties of gas hydrates containing methane, argon and krypton. Ind. Eng. Chem. Fund. **1980**, *19*, 282–289.

8. Makogan, Y.F. *Hydrates of Hydrocarbons*; Penn Well Books: Tulsa, 1997.
9. Kargel, J.S.; Tanaka, K.L.; Baker, G.; Komatsu, G.; MacAyeal, D.R. Formation and dissociation of clathrate hydrates on mars: polar caps, northern plains and highlands. Lunar Planet. Sci. **2002**, *XXXI*. www.lpi.usra.edu/meetings/lpsc2000/pdf/1981.pdf.
10. Claussen, W.F. A second water structure for inert gas hydrates. J. Chem. Phys. **1951**, *19*, 1425–1429.
11. von Stackelberg, M.; Müller, H.R. Feste Gas Hydrate II. Zeit. Elektrochem. **1954**, *58*, 25–36.
12. Jeffery, G.A.; McMullan, R.K. The clathrate hydrates. Prog. Inorg. Chem. **1967**, *8*, 45–53.
13. Ripmeester, J.A.; Ratcliffe, C.I. Low-temperature cross-polarization/magic angle spinning ^{13}C NMR of solid methane hydrates: structure, cage occupancy, hydration number. J. Phys. Chem. **1988**, *92*, 337–339.
14. Konstantin, A.; Uchadin, K.A.; Ripmeester, J.A. A complex clathrate hydrate structure showing bimodal guest hydration. Nature **1999**, *397*, 420–423.
15. Subramanian, S.; Kini, R.A.; Dec, S.F.; Sloan, E.D. Methane and ethane can form structure II hydrates. Chem. Eng. Sci. **2000**, *55*, 1981–1999.
16. Davidson, D. Gas hydrates as clathrate ices. In *Natural Gas Hydrates—Properties, Occurrence and Recovery*; Cox, J., Ed.; Butterworth: Woburn, MA, 1983; 1–16.
17. Handa, Y.P. Calorimetric studies of laboratory synthesized and naturally occurring gas hydrates. Ind. Eng. Chem. Res. **1988**, *27*, 872–874.
18. Uchida, T.; Hirano, T.; Ebinuma, T.; Narita, H.; Gohara, K.; Mae, S.; Matsumoto, R. Raman spectroscopic determination of hydration number of methane hydrates. AIChE J. **1999**, *45*, 2641–2645.
19. Berecz, E.; Balla-Achs, M. *Gas Hydrates*; Elsevier: Amsterdam, 1983.
20. Sloan, E.D. *Clathrate Hydrates of Natural Gases*, 2nd Ed.; Marcel Dekker Inc.: New York, 1998.
21. Morita, K.; Nakano, S.; Ohgaki, K. Structure and stability of ethane hydrate crystal. Fluid Phase Equilibria **2000**, *169*, 167–175.
22. Ng, H.J.; Robinson, D.B. Hydrate formation in systems containing methane, ethane, propane, carbon dioxide or hydrogen sulfide in the presence of methanol. Fluid Phase Equilibria **1985**, *21*, 145–155.
23. Bishnoi, P.R.; Dholabhai, P.D.; Mahadev, K.N. Solid deposition in hydrocarbon systems kinetics and thermodynamics of gas hydrate formation. Task #2, Gas hydrate equilibrium studies. Report for GRI, Contract Number 5091-260-2138. Report numbers GRI-96/1043 and GRI-96/1044, 1996.
24. Mehta, A.P.; Sloan, E.D. Structure H hydrates: the state of the art, Preprint, 75th Annual GPA Convention, March 11–13, 1996; Denver: Colorado.
25. Thomas, M.; Behar, E. Modelling of structure H hydrate equilibria for methane, intermediate hydrocarbon molecules and water systems. Proceedings of the 75th GPA Annual Convention, March 11–13, 1996; Denver: Colorado.
26. Hütz, U.; Englezos, P. Measurement of structure H hydrate phase equilibrium and the effect of electrolytes. Fluid Phase Equilibrium **1996**, *117*, 178–185.
27. Servio, P.; Lagers, F.; Peters, C.; Englezos, P. Gas hydrate phase equilibrium in the system methane-carbon dioxide-neohexane and water. Fluid Phase Equilibria **1999**, *158-160*, 795–800.
28. Ershov, E.D.; Yakushev, V.S. Experimental research on gas hydrate decomposition in frozen rocks. Cold Regions Sci. Technol. **1992**, *20*, 147–156.
29. Handa, Y.P.; Stupin, D. Thermodynamic properties and dissociation characteristics of methane and propane hydrates in 70-Å-radius silica gel pores. J. Phys. Chem. **1992**, *96*, 8599–8606.
30. Ahmadi, G.; Chuang, J.; Smith, D. A simple model for natural gas production from hydrate decomposition. Proceedings of the Third International Conference on Gas Hydrates, Salt Lake City, Utah, July 21–24; Holder, G., Bishnoi, P.R., Eds.; Annals of the New York Academy of Science, 1999; 420–427.
31. Englezos, P.; Bishnoi, P.R. Prediction of gas hydrate formation conditions in aqueous electrolyte solutions. AIChE J. **1988**, *34*, 1718–1721.
32. Clarke, M.A.; Bishnoi, P.R. Prediction of hydrate formation conditions in aqueous electrolyte solutions in the presence of methanol. Proceedings of the Fourth International Conference on Gas Hydrates, Yokohama, Japan, May 17–20, 2002; 406.
33. Clarke, M.A.; Bishnoi, P.R. Development of a new equation of state for mixed salt and mixed solvent systems, and application to vapour liquid equilibrium and solid (hydrate) vapour liquid equilibrium calculations. Fluid Phase Equilibria **2004**, *220*, 21–35.
34. Jager, M.D.; Ballard, A.L.; Sloan, E.D. The next generation of hydrate prediction: II. Dedicated aqueous phase fugacity model for hydrate prediction. Fluid Phase Equilibria **2003**, *211*, 85–107.
35. Englezos, P.; Hall, S. Phase equilibrium data on carbon dioxide hydrate in the presence of electrolytes, water soluble polymers and montmorillonite. Can. J. Chem. Eng. **1994**, *72*, 887–893.

36. Clarke, M.A.; Pooladi-Darvish, M.; Bishnoi, P.R. A method to predict equilibrium conditions of gas hydrate formation in porous media. Ind. Eng. Chem. Res. **1999**, *38*, 2485–2493.
37. Bishnoi, P.R.; Gupta, A.K.; Englezos, P.; Kalogerakis, N. Multiphase equilibrium flash calculations for systems containing gas hydrates. Fluid Phase Equilibria **1989**, *53*, 97–104.
38. Gupta, A.K.; Bishnoi, P.R.; Kalogerakis, N. A method for the simultaneous phase equilibria and stability calculations for multiphase reacting and nonreacting systems. Fluid Phase Equilibria **1991**, *63*, 65–89.
39. Ballard, A.L.; Sloan, E.D. The next generation of hydrate prediction: Part III. Gibbs energy minimization formalism. Fluid Phase Equilibria **2002**, *218*, 15–31.
40. Ballard, A.L.; Sloan, E.D. The next generation of hydrate prediction: I. Hydrate standard states and incorporation of spectroscopy. Fluid Phase Equilibria **2003**, *194-197*, 371–383.
41. Natrajan, V.; Bishnoi, P.R.; Kalogerakis, N. Induction phenomena in gas hydrate nucleation. Chem. Eng. Sci. **1994**, *49*, 2075–2087.
42. Englezos, P.; Dholabhai, P.; Kalogerakis, N.; Bishnoi, P.R. Kinetics of methane and ethane hydrate formation. Chem. Eng. Sci. **1987**, *42*, 2659–2666.
43. Parent, J.S.; Bishnoi, P.R. Investigations into the nucleation behaviour of methane gas hydrates. Chem. Eng. Comm. **1996**, *144*, 51–64.
44. Natarajan, V.; Bishnoi, P.R. Formation and decomposition of gas hydrates. Fluid Phase Equilibria **1996**, *117*, 168–177.
45. Vysniauskus, A.; Bishnoi, P.R. A kinetic study of methane hydrate formation. Chem. Eng. Sci. **1983**, *38*, 1061–1069.
46. Shindo, Y.; Lund, P.C.; Fujioka, Y.; Komiyama, H. Kinetics and mechanism of the formation of CO_2 hydrate. Int. J. Chem. Kinet. **1993**, *25*, 777–782.
47. Clarke, M.A.; Bishnoi, P.R. Determination of the intrinsic rate of CO_2 hydrate formation using in situ particle size analysis. Chem. Eng. Sci. **2004**, *60*, 695–709.
48. Herri, J.; Gruy, F.; Cournil, M.; Kinetics of methane hydrate formation Proceedings of the Second International Conference on Natural Gas Hydrates, Toulouse, France, June 2–6, Guillon, O., Ed.; INP ENGISC, 1996; 243–250.
49. Herri, J.; Gruy, F.; Pic, J.S.; Cournil, M. Methane hydrate crystallization mechanism from in situ particle sizing. AIChE J. **1999**, *45*, 590–602.
50. Dholabhai, P.; Kalogerakis, N.; Bishnoi, P.R. Kinetics of methane hydrate formation in aqueous electrolyte solutions. Can. J. Chem. Eng. **1993**, *71*, 68–74.
51. Kalogerakis, N.; Jamaluddin, A.K.M.; Dholabhai, P.D. Effect of surfactants on hydrate formation kinetics. SPE 25188 1993.
52. Rogers, R.E.; Zhong, Y. Feasibility of storing natural gas in hydrates commercially. Proceedings of the Third International Conference on Gas Hydrates, Salt Lake City, Utah, July 21–24, 1999; 843–850.
53. Kim, H.C.; Bishnoi, P.R.; Heidemann, R.A.; Rizvi, S.S.H. Kinetics of methane hydrate decomposition. Chem. Eng. Sci. **1987**, *42*, 1645–1653.
54. Clarke, M.A.; Bishnoi, P.R. Determination of the intrinsic rate of ethane gas hydrate decomposition. Chem. Eng. Sci. **2000**, *55*, 4869–4883.
55. Clarke, M.A.; Bishnoi, P.R. Determination of the activation energy and intrinsic rate constant of methane gas hydrate decomposition. Can. J. Chem. Eng. **2001**, *79*, 143–147.
56. Clarke, M.A.; Bishnoi, P.R. Measuring and modelling the rate of decomposition of gas hydrates formed from mixtures of methane and ethane. Chem. Eng. Sci. **2001**, *56*, 4715–4724.
57. Clarke, M.A.; Bishnoi, P.R. Determination of the intrinsic rate of CO_2 hydrate decomposition using in situ particle size analysis. Chem. Eng. Sci. **2004**, *59*, 2983–2993.
58. Notz, P.K.; Bumgardner, S.B.; Schaneman, B.D.; Todd, J.L. Application of kinetic inhibitors to gas hydrate problems. SPE Prod. Facil. **1996**, *11* (4), 256–260.
59. Hansen, A.B.; Clasen, T.L.; Bass, R.M.; Direct impedance heating of deepwater flowlines. Proceedings of the Annual Offshore Technology Conference, SPE, 1999 Vol. 3, 691–701.
60. Campbell, J. *Gas Conditioning and processing*; Vol. II, Section 17; Campbell Petroleum Series, Inc., 1981.
61. Englezos, P.; Huang, Z.; Bishnoi, P.R. Prediction of natural gas hydrate formation conditions in the presence of methanol using the Trebble-Bishnoi equation of state. J. Can. Pet. Tech. **1991**, *30*, 148–154.
62. Mehta, A.P.; Hebert, P.B.; Cadena, E.R.; Weatherman, J.P. Fulfilling the promise of low dosage hydrate inhibitors: journey from academic curiosity to successful field implementation. Proceedings of the Annual Offshore Technology Conference, SPE, 2002; 565–571.
63. Kvenvolden, K. A gas hydrate and humans. Proceedings of the Third International Conference on Gas Hydrates, Salt Lake City, Utah, July 21–24, Holder, G., Bishnoi, P.R., Eds.; Annals of the New York Academy of Science, 1999; 17–22.

64. Lee, S.Y.; Holder, G.D. Methane hydrates potential as a future energy source. Fuel Proc. Technol. **2001**, *71*, 181–186.
65. Dallimore, S.R.; Uchida, T.; Collett, T.S. Scientific results from Japex/JNOC/GSC Maillik 2L-38 gas hydrate research well, MacKenzie Delta, Northwest Territories, Canada, Geological Survey of Canada, Bulletin 544, 1999.
66. Dallimore, S.R., Collett, T.S., Eds.; Scientific results from the Malik 2002 gas hydrate production research well program, MacKenzie Delta, Northwest Territories, Canada, Geological Survey of Canada Bulletin 585, 2005.
67. Spence, G.D.; Hyndman, R.D. The challenge of deep ocean drilling for natural gas hydrate. Geosci. Can. **2001**, *28*, 170–186.
68. Maini, B.B.; Bishnoi, P.R. Experimental investigation of hydrate formation behaviour of a natural gas bubble in a simulated deep sea environment. Chem. Eng. Sci. **1981**, *36*, 183–189.
69. Topham, D.R. Modelling of hydrocarbon bubble plumes to include gas hydrate formation. Chem. Eng. Sci. **1984**, *39*, 1613–1622.
70. Chen, F.; Yapa, P.D. Estimating hydrate formation and decomposition of gases released in a deepwater ocean plume. J. Mar. Syst. **2001**, *30*, 21–32.
71. Parent, J.D. Equilibrium compositions and enthalpy changes for the reactions of carbon, oxygen and steam. Inst. Gas Technol. Res. Bull. **1948**, *1*.
72. Gudmundsson, J.; Borrehaug, A. Frozen hydrate for transport of natural gas. Proceedings of the Second International Conference on Natural Gas Hydrates, Toulouse, France, June 2–6, Guillon, O., Ed.; INP ENGISC, 1996; 983–986.
73. Takaoki, T.; Iwasaki, T.; Katoh, T.; Takashi, A.; Kiyoshi, H. Use of hydrate pellets for natural gas transportation. I. Advantage of pellet form of natural gas hydrate in sea transportation. Proceedings of the Fourth International Conference on Gas Hydrates, Yokohama, Japan, May 17–20, 2002; 406.
74. Knox, W.G.; Hess, M.; Jones, G.E. Smith, H.B. The hydrate process. Chem. Eng. Prog. **1961**, *57* (2), 66–76.
75. Ngan, Y.T.; Englezos, P. Concentration of mechanical pulp mill effluents and NaCl solutions through propane hydrate formation. Ind. Eng. Chem. Res. **1996**, *35*, 1894–1900.
76. Purwanto, Y.A.; Oshita, S.; Seo, Y.; Kawagoe, Y. Concentration of liquid foods by the use of gas hydrate. J. Food Eng. **2001**, *47*, 133–138.

Natural Gas Utilization

Peter R. Pujadó
UOP LLC, Des Plaines, Illinois, U.S.A.

INTRODUCTION

Natural gas is a complex mixture of light hydrocarbons and some nonhydrocarbon materials. Methane (CH_4) is its major constituent, but natural gas may contain much heavier hydrocarbons, including pentanes or hexanes. This gives rise to terminology like ''wet gas'' for a natural gas that contains condensable hydrocarbons and ''dry gas'' for a natural gas that is essentially free of recoverable liquid hydrocarbons. Other components of natural gas include carbon dioxide (CO_2), nitrogen, hydrogen sulfide (H_2S), mercaptans (HS-R), helium, water, and other minor contaminants. The composition of the natural gas is often related to its provenance. ''Associated gas'' is natural gas that is recovered from an oil reservoir and, having been in contact with oil, tends to contain a larger proportion of hydrocarbons other than methane. On the other hand, natural gas that has not been in contact with oil, sometimes called ''nonassociated gas,'' tends to be leaner on hydrocarbons other than methane and often contains only small amounts of ethane and propane. ''Natural gas liquids'' are light hydrocarbons like ethane, propane, and butanes that can be recovered from natural gas by condensation.

BACKGROUND

Although natural gas has been known for thousands of years, sometimes as a subject of awe and worship, its first documented ''industrial'' use appears to have been by the Chinese who used it to evaporate salt brine in around the eighth century. It appears in the environment either as a result of gas seeps from underground coal mines or shallow oil reservoirs, or as swamp gas from the decomposition of organic compounds, and not insignificantly from the flatulence of animals, in particular cattle. All natural gas appears to be of organic origin, although some theories postulate the existence of abiotic gas reservoirs deep in the Earth's crust. Regardless of its source, the release of natural gas into the environment is of growing concern, as, because of the relative quantities, methane is potentially a more significant contributor to the ''greenhouse effect'' than CO_2.

In recent times, natural gas was first used domestically for lighting in Fredonia, New York, in around 1821. Later, in 1865, the Fredonia Gas, Light, and Waterworks Company became the first natural gas company in the U.S.A. The high calorific value of natural gas soon prompted its use as an industrial fuel for the steel and glass industry in Pittsburgh in the late nineteenth century, and its use has continued to expand ever since. The availability of natural gas in Europe has always been very limited and it has been only in relatively recent years that its use has expanded as a result of the discovery of oil and gas in the North Sea and the import of natural gas from Russia by pipeline and from North Africa and the Middle East by ship as liquefied natural gas (LNG). Eventually though, natural gas and liquefied petroleum gases (LPG; propane, butanes, or mixtures thereof) completely replaced the old ''town gas'' or ''water gas,'' a highly toxic mixture of carbon monoxide (CO), hydrogen, methane, and other hydrocarbons, which is obtained by blowing air and steam over red-hot coal, and which, despite its very low calorific value, for many years was the mainstay for domestic illumination in many European countries.

Today, natural gas is the third most important fuel used worldwide, after oil and coal. Like all fossil fuels, the reserves of natural gas are far from being inexhaustible but, fortunately so far, newer fields and reserves have been identified at a faster rate than its consumption so that, overall, the proven reserves have been increasing. If the recovery of deep-ocean natural gas hydrates were to become feasible in the future, much larger reserves would be available. The uses of natural gas today can be summarized as:

1. Residential fuel.
2. Commercial fuel.
3. Industrial fuel.
4. Power generation fuel.
5. Transportation fuel.
6. Chemical feedstock.
   a. Direct chemical applications.
   b. Indirect chemical applications (through synthesis gas, or ''syn gas'').
7. Feedstock for liquid fuels (also through syn gas).

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120019083

NATURAL GAS RESERVES

Because of their nature, encyclopedias are never comprehensive nor up to date, and this one is no exception. For the best overview of the reserves and production of natural gas and almost all other fuels it is best to consult the specialized literature. Two particularly valuable resources are the *Oil & Gas Journal* and the annual *BP Statistical Review of World Energy*, currently available in the internet at http://www.bp.com and also available in hard copy as far back as 1951. These and other publications from the International Energy Agency (IEA), the European Union, and other sources, provide a detailed, up-to-date picture of the energy balance for all fuels, on a country-by-country, and year-by-year basis.

At the end of 2004, the proved reserves of natural gas were 179.53 trillion (10^{12}) cubic meters (m^3). [Please refer to the appendix on units and unit conversions—in this writing we use the American and British definitions of billion (10^9) and trillion (10^{12}), which should not be confused with the European definitions, 10^{12} and 10^{18}, respectively. Also, unless noted otherwise, the volumetric amounts are expressed in terms of standard m^3, or scm, measured at 15°C and 1 atm or 1013 mbar]. In Table 1 the change in reserves over the past 20 years is illustrated:[1]

The increase in reserves is short of spectacular, but conservation has had nothing to do with it. Worldwide consumption increased from 2.04 trillion m^3 in 1992 to 2.54 trillion m^3 in 2002 and about 2.7 trillion m^3 in 2004. The main contributor to the growth in reserves has been the discovery of massive fields in the Russian Federation, from western Siberia all the way to Sakhalin island in the Far East, and the potential for future discoveries in Siberia, Kazakhstan, and surrounding areas. Iran, Qatar, Nigeria, the United Arab Emirates, and, to a lesser extent, Trinidad, Bolivia, Venezuela, Egypt, Australia, etc. have also been significant contributors to the growth in reserves. In contrast, the proved reserves in the U.S.A., Canada, and Mexico have decreased. The reserves/production (R/P) ratio for a given year provides an indication of the length of time that the remaining reserves would last if production were to continue at the same level as for that year. Typically, the R/P ratios are low for industrialized countries with a large appetite for fuel—for example, 9.6 for the U.S.A. at the end of 2002—and high for less developed areas. Not surprisingly, we are seeing a shift in the location of large gas-consuming industries from high-cost industrialized countries where natural gas may cost as much as $6.00 per million Btu to less developed areas where gas is abundant at the source and its cost may be as low as $0.50 per million Btu.

Figs. 1–7 are reproduced here from the *BP Statistical Review of World Energy 2005*, by permission from BP p.l.c.

Table 1 Worldwide natural gas reserves

Year end	Trillion (m^3)
1982	85.9
1992	138.3
2002	155.8
2004	179.5

FUEL PROPERTIES OF NATURAL GAS

Natural gas is a mixture of variable composition. Its properties, therefore, cannot be defined precisely. A typical range of compositions could be as given in Table 2.

Because natural gas may contain significant amounts of nonhydrocarbons, its calorific value will be a strong function of its composition. In fact, for transportation and fuel utilization, it is necessary to remove these nonhydrocarbons to the highest extent possible. Typically, natural gas is treated at the source, either for a single well or, more commonly, at a central point for several wells. The gas is desulfurized or sweetened to bring the hydrogen sulfide content down to a few ppm. Usually amines or other solvents are used for the removal of H_2S by absorption and stripping; elemental sulfur can then be recovered by using a Claus unit or similar. Depending on the system, CO_2 can also be removed or depleted by absorption/stripping; alternatively, it can be separated from natural gas using membranes (membranes have also been used for H_2S but to a lesser extent).

Typically, natural gas will also be dried to remove water vapor using a glycol system, a cryogenic unit, or molecular sieve adsorbents, and it may be treated for the removal of mercury, arsenic (often as arsine, AsH_3), or phosphorous (often as phosphine, PH_3). Various adsorption units are available for the removal of traces of these metals.

At the end what matters is to have a natural gas with a high calorific value. In Table 3, heating values for typical components in clean natural gas are provided:[2]

Although natural gas is usually priced in terms of its high heating value (HHV), most users only benefit from its low heating value (LHV), as the latent heat of water vapor condensation is seldom recovered in smaller applications. In Table 4, the heating value calculation for a hypothetical gas composition is illustrated.

Nitrogen and CO_2 dilute natural gas and lower its heating value. A ''good'' natural gas typically will have high heating values of about 1000 Btu/scf (pure methane has an HHV of 1009 Btu/scf). Commercially,

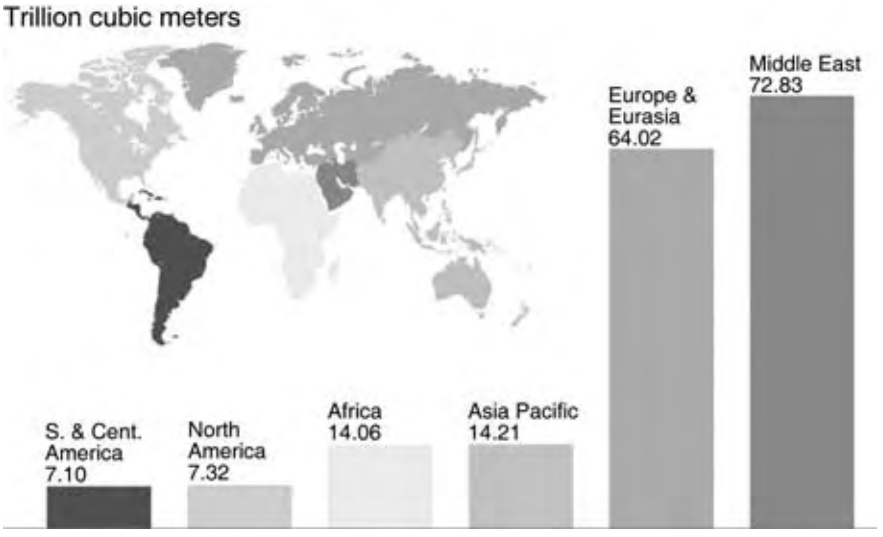


Fig. 1 Distribution of proved natural gas reserves (2004). (From Ref.[1].) *(View this art in color at www.dekker.com.)*

heating requirements are usually expressed in terms of millions of Btu (mm Btu). One mm Btu = 1.055 giga-Joules (GJ). Domestic gas consumption in the U.S.A. is often expressed in terms of ''therms.'' One therm = 10^5 Btu = 0.1 mm Btu.

The success of natural gas as a clean fuel has greatly increased its consumption in power plant applications. Natural gas fueled power plants are significantly less costly to build than coal fired or nuclear power plants. They are also easier to operate so that they lend themselves to ''peaking'' operations in which they are only brought on stream when required and may sit idle the rest of the time. In principle, natural gas fired power plants are also cheaper to operate, but the increased demand for natural gas has rapidly increased feedstock prices in the U.S.A. in recent years.

When facilities for shipping natural gas by pressurized pipelines are not available, the only feasible solution is to liquefy it in cryogenic installations and to ship it as LNG. Most LNG landing facilities are located in Japan, Korea, and in southern Europe. Although the facilities to land natural gas as LNG in the U.S.A. are currently very limited (one in the Gulf Coast and one in New England), it can be expected that others will be built as a result of the growing demand and diminishing domestic supply. While this entry does not address LNG or the shipping of natural gas in general, it must be realized that the natural gas requirements for LNG facilities that are commercially and economically viable can be formidable. A typical facility to handle 5 mm t/yr of LNG requires 22.2 mm scmd of natural gas, or about 222 billion scm

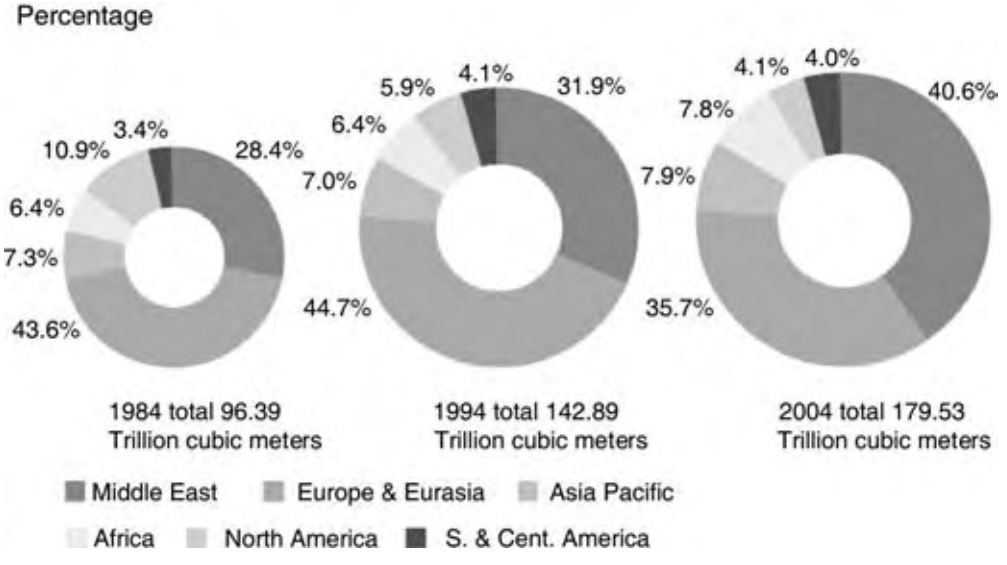


Fig. 2 Distribution of proved natural gas reserves (2004). (From Ref.[1].) *(View this art in color at www.dekker.com.)*

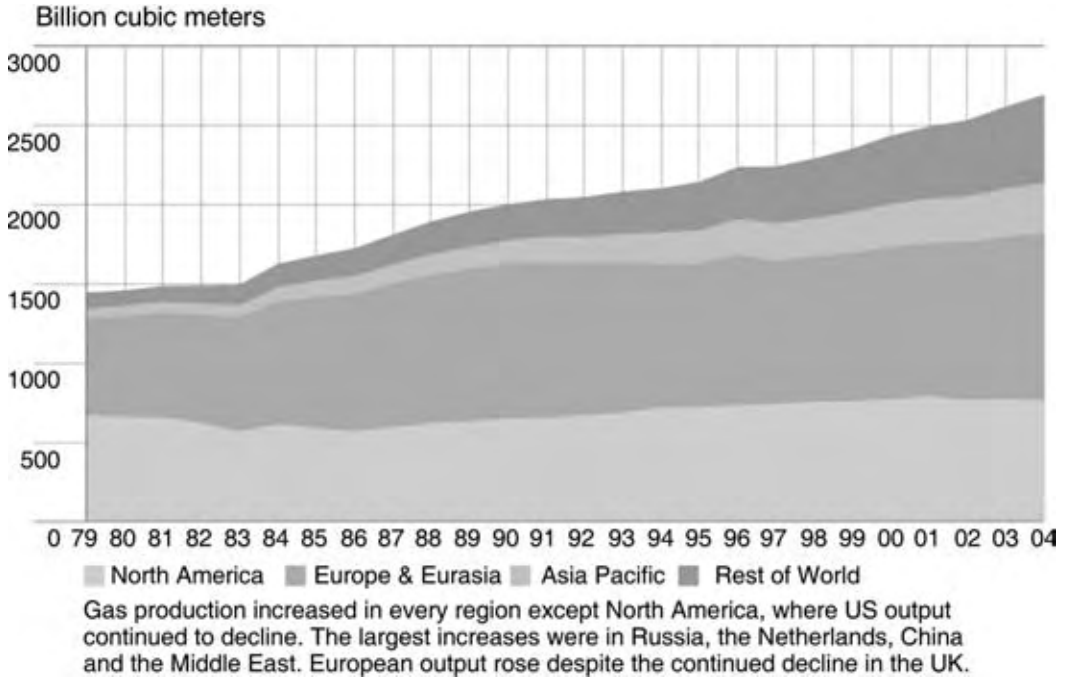


Fig. 3 Geographic distribution of natural gas production (2004). (From Ref.[1].) *(View this art in color at www.dekker.com.)*

over a 30-year time span. Liquefaction, shipping, and revaporization costs are significant. Even with natural gas priced at source at not more than \$0.75 per mm Btu, the landed price is likely to exceed \$3.50 per mm Btu, before local distribution costs and profit margins.

NATURAL GAS AS TRANSPORTATION FUEL

In addition to its use as domestic and industrial fuel, natural gas is extensively used as a transportation fuel. In transportation applications, natural gas offers the advantages of being a clean burning fuel that can

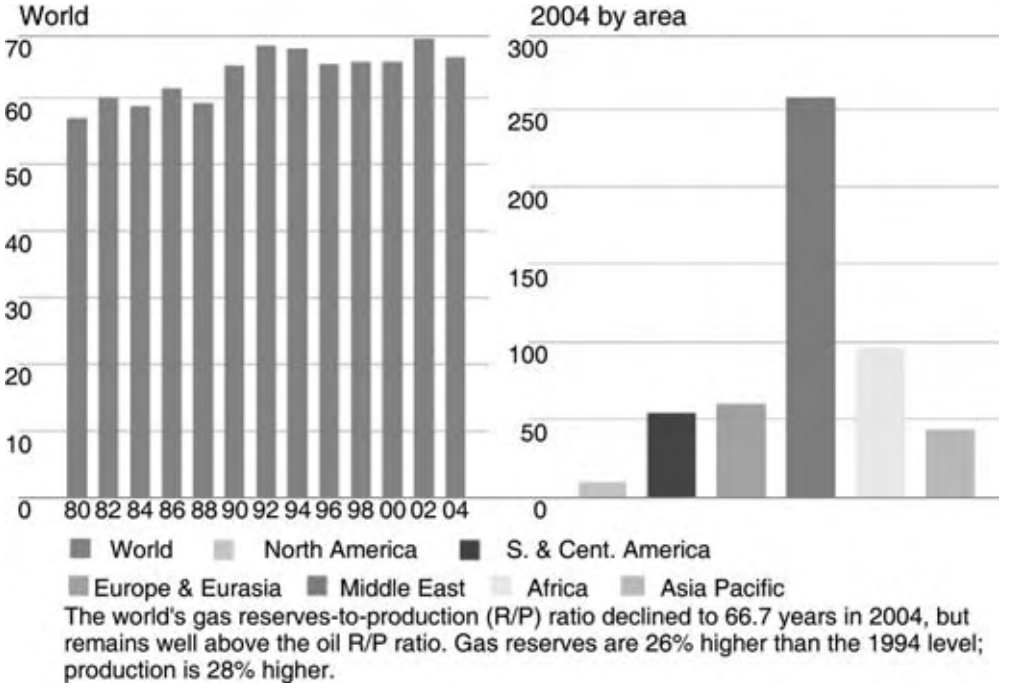


Fig. 4 Geographic reserve/production ratios of natural gas (2004). (From Ref.[1].) *(View this art in color at www.dekker.com.)*

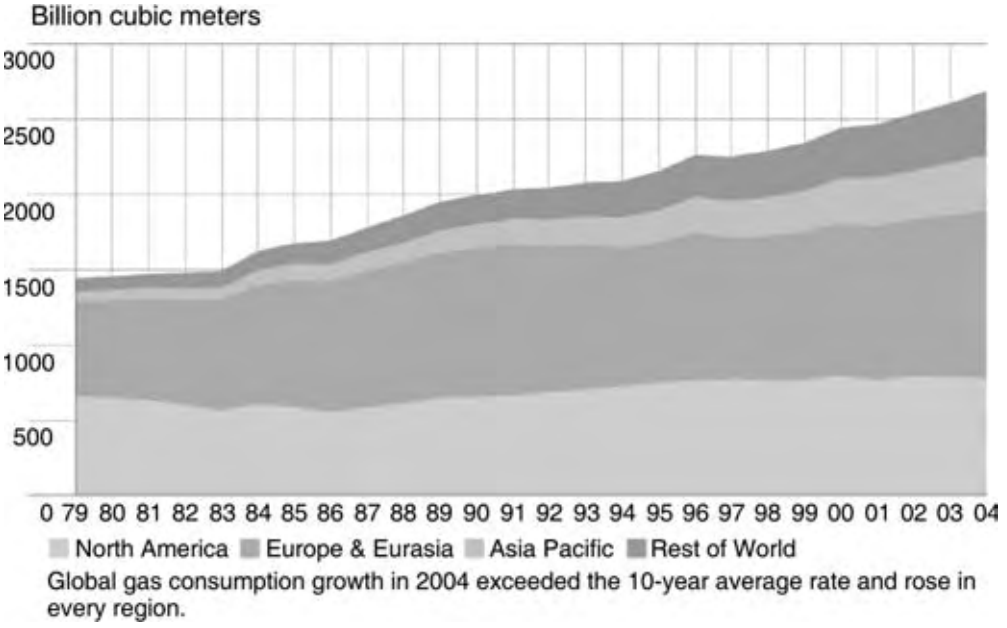


Fig. 5 Geographic consumption of natural gas (2004). (From Ref.[1].) *(View this art in color at www.dekker.com.)*

generate significantly reduced exhaust and greenhouse gas emissions than in comparable gasoline- or diesel-powered vehicles. In addition, because methane, the main ingredient of natural gas, has an exceptionally high octane number with $(R + M)/2 > 120$, it can be readily used in conventional gasoline-powered engines.

The main difficulty with natural gas is its storage requirements and limitations. As a gas, its energy capacity per unit volume is exceedingly low. Therefore, the conventional solution has been to raise its pressure to about 2400 psi or higher. Compressed natural gas (CNG) has a much higher energy content per unit volume and is of practical use in vehicular applications. While in the U.S.A. uses of CNG have been largely limited to municipal transit buses and fleet applications, in other countries, in particular New Zealand, CNG has gained wide acceptance as an automobile fuel. In Table 5, comparative values of the energy content of various fuels are provided.

Higher energy content of CNG can be obtained by storing it at even higher pressures. Pressures as high as 3500–5000 psig have been used, although there may be

Fig. 6 Per capita consumption of natural gas (2004). (From Ref.[1].) *(View this art in color at www.dekker.com.)*

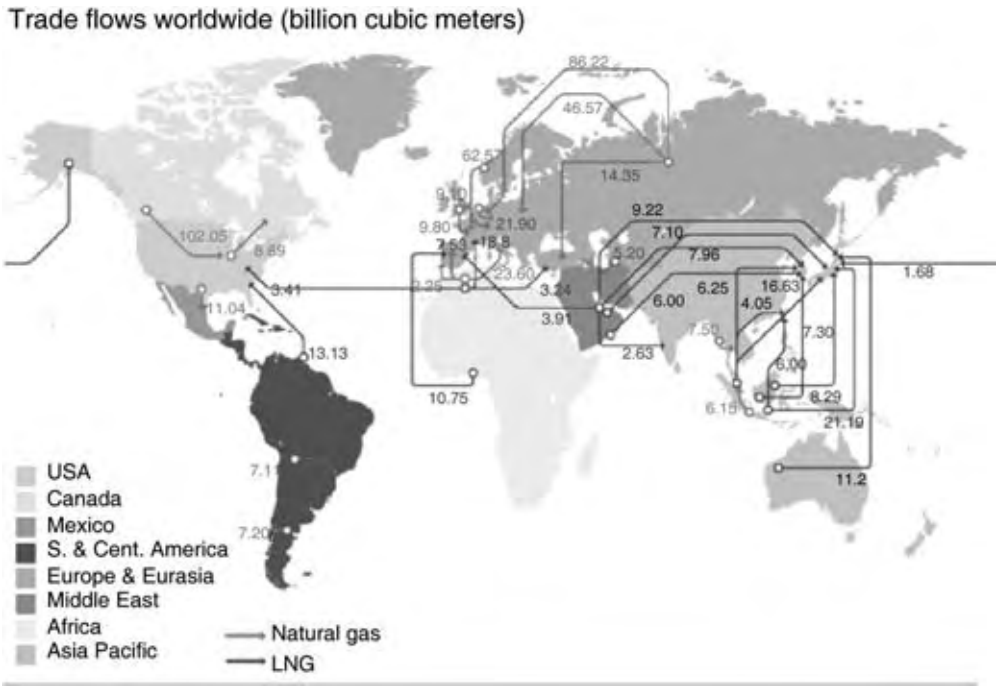


Fig. 7 Natural gas shipping and trade movements (2004). (From Ref.[1].) *(View this art in color at www.dekker.com.)*

some concerns as to the integrity of the storage tanks, principally in the event of an accident.

Clearly, CNG has a significant energy content, but still falls short of the energy contained in an equivalent volume of gasoline or diesel fuel. Typically, even with an expanded fuel tank volume made possible by its lower weight, the driving range of a CNG-powered vehicle is only about one-third to one-half that of the same vehicle with a gasoline-powered engine.

An alternative that has been proposed is to take advantage of the solubility of methane in hydrocarbons under medium to elevated pressures. Thus, by operating at pressures comparable to or lower than those used for CNG, it is possible to greatly increment the volumetric energy content by coupling the energy content of the solvent with that of the solute.

Data for methane/hydrocarbon mixtures are not well established. Some results can be generated through commercially available simulators like Aspen or Hysys but are somewhat questionable because of the nonidealities of these mixtures at elevated pressures. Also, the vapor–liquid equilibrium data currently available are very limited and mostly concern blends of methane with *n*-paraffins, which have very poor ignition characteristics in internal combustion engines.

With the above-mentioned uncertainties and limitations concerning the accuracy of commercial simulators, results obtained with Aspen, which can be used to at least directionally, illustrate the merits of using methane/diesel blends for a particular set of conditions with a diesel fuel at room temperature (70°F), are illustrated in Table 6.

Table 2 Typical composition ranges for natural gas

	Mol%
Hydrocarbons	
Methane	75–99
Ethane	1–15
Propane	1–10
n-Butane	<2
iso-Butane	<1
n-Pentane	<1
iso-Pentane	<1
Hexane and heavier	<1
Nonhydrocarbons	
Nitrogen	0–15
Carbon dioxide	0–30
Hydrogen sulfide	0–30
Helium	0–5

Table 3 Heating values of individual natural gas components

	MW	HHV (Btu/lb)	LHV (Btu/lb)
Nitrogen	28.014	0	0
CO_2	44.011	0	0
Hydrogen	2.016	60,958	51,571
Methane	16.043	23,861	21,502
Ethane	30.070	22,304	20,416
Propane	44.097	21,646	19,929

Table 4 Heating values calculation for a given natural gas composition

	Mol%	Wt.%	HHV (Btu/scf)	LHV (Btu/scf)
Nitrogen	2.00	3.17	0.00	0.00
CO_2	3.00	7.47	0.00	0.00
Hydrogen	1.00	0.11	3.24	2.74
Methane	90.00	81.66	907.87	818.11
Ethane	3.00	5.10	53.02	48.53
Propane	1.00	2.49	25.15	23.16
Average MW = 17.683			989.28	892.54

Thus, by using mixtures of natural gas with hydrocarbon fuels, we can maintain the volumetric energy density of the mixture at values close to those of the hydrocarbon while adding significant amounts of natural gas to the hydrocarbon fuel pool, thereby incrementing the total energy content by 10% or more, depending on the pressure.

While the above results reflect mixtures of methane with a standard diesel fuel, similar or even better effects can be observed with blends of methane with gasoline components. In Table 7, as an example, literature data on the solubility of methane in toluene at 300°F (solubilities are much higher at lower temperatures) are provided. This also illustrates the need to generate additional data for different hydrocarbons over a broad range of operating conditions, i.e., pressure and temperature.

Methane forms azeotropes with many hydrocarbons. For example, methane and *n*-butane at 32°F and 2660 psig form an azeotrope with 92.9 mol% methane. Azeotropes are ideal for fuel blend applications because the composition remains essentially constant as the fuel is being used and the tank is being emptied. Little has been done to date to identify the optimal temperature and pressure conditions for the formation of azeotropes for the various hydrocarbons and for commercial gasoline and diesel compositions.

DIRECT USES OF NATURAL GAS AS A CHEMICAL FEEDSTOCK

For the most part we refer to the direct use of methane (100% basis) as a chemical feedstock. The methane molecule is very inert in relation to other molecules, except when used as fuel with oxygen (air); however, there are a few applications involving the direct reaction of methane to chemical derivatives. The more characteristic applications are listed here:[4,5]

- Production of acetylene.
- Production of hydrogen cyanide (HCN).
- Production of chloromethanes.

Other applications have been proposed but as yet without success, like the direct coupling of methane to ethylene or the direct conversion of methane to methanol. Recently, some disclosures were made on the reaction of methane with alkanes to yield augmented alkanes.[6]

Production of Acetylene

Most acetylene is produced nowadays from methane and other light hydrocarbons. Methane is usually the preferred raw material. Although ethane, propane, and butane can also be used, the yield of acetylene declines with increasing molecular weight of the feedstock and,

Table 5 Typical heating values of various fuels

LHV	Btu/U.S. gallon
Gasoline	115,000+
Methanol	~60,000
Diesel	128,000+
Methane (atmospheric pressure)	120
CNG at 2400 psig	23,000

Table 6 Estimated energy of methane/diesel fuel blends

Pressure (psi)	1,450	2,900
Diesel mass (lb)	100.0	100.0
Diesel density (kg/m^3)	847	852
Diesel density (lb/gal)	7.07	7.11
Diesel volume (gal)	14.2	14.1
Mixture mass (lb)	105.4	113.0
Mixture density (kg/m^3)	696	571
Mixture density (lb/gal)	5.81	4.76
Mixture volume (gal)	18.2	21.0
Diesel energy (mm Btu)	1.81	1.81
Mixture energy (mm Btu)	1.93	2.10
Volume energy density (Btu/gal)		
Diesel	128,000	128,000
CNG (at 2400 psig)	23,000	23,000
Mixture	110,000	100,000

Table 7 Solubility of methane in toluene at 300°F

Pressure (psig)	Mol% methane
278.5	3.53
424.4	5.45
731.4	9.54
1441.4	19.49
2199.4	28.79
2924.4	38.58
3614.4	48.97

in any case, ethane, propane, and butane are better used in the production of olefins by steam cracking.

Production of acetylene from methane requires very high temperatures in the order of 1000–1500°C, very short residence times, and a very fast quench to prevent decomposition of the product.[5] Typically, methane can be heated by: a) contact with a very hot solid; b) in situ partial oxidation of the feedstock; or c) electric discharge or plasma processes. In a typical electric arc process (e.g., the Hüls process), the furnace may be a vertical tube with an injection chamber in which methane is fed tangentially to generate a rapidly rotating vortex that flows through an arc created between electrodes at some 7000 V. This has the advantage of making the temperature of the reactor more uniform and of assisting in the separation of soot particles. The reactor is cooled by water circulation. Because of the reduced dimensions of such reactors, high productivities can be obtained with a residence time of about 2 msec. The hot effluent gas is rapidly quenched by water or by injecting a hydrocarbon that can be thermally cracked to olefins, thus allowing for the dual production of acetylene and olefins.

Production of Hydrogen Cyanide

Although substantial amounts of hydrogen cyanide are produced and recovered as a byproduct from the manufacture of acrylonitrile by the ammoxidation of propylene, some hydrogen cyanide is also made on purpose from methane. Basically, two main approaches are available.[5]

The Andrussov process involves the ammoxidation of methane:

$$CH_4 + NH_3 + \frac{3}{2}O_2 \rightarrow HCN + 3H_2O$$
$$\Delta H^0_{298} \sim -480 \text{ kJ/mol}$$

The reaction is highly exothermic and uses volumetric flow ratios, air/CH_4/NH_3, of about 5/1/1. Whereas propylene ammoxidation proceeds at about 440°C, the ammoxidation of methane requires temperatures of about 1100–1200°C. Although HCN yields are high, in the order of about 80 mol%, the gaseous effluent is very dilute with only about 6–8 vol.% HCN, and it has to be rapidly quenched to about 400°C to avoid undesirable side reactions.

The Blausäure Methan Ammoniak (BMA) process of Degussa uses an endothermic amination reaction in the absence of oxygen:

$$CH_4 + NH_3 \rightarrow HCN + 3H_2$$
$$\Delta H^0_{298} \sim +250 \text{ kJ/mol}$$

The reaction takes place in sintered alumina tubes covered with platinum and heated externally to about 1200–1300°C. The HCN in the effluent may be as high as 20 vol.% and the molar yields may approach 85%.

Production of Chloromethanes

Because of its relatively poor selectivity to individual components, methane is not the preferred feedstock for the production of chloromethanes.[4] Methyl chloride, for example, is best obtained by the reaction of methanol with hydrogen chloride, and methylene dichloride can be obtained by the chlorination of methyl chloride. Methane is difficult to chlorinate. Because methyl chloride is easier to chlorinate than methane, a large excess of methane is required if a reasonable yield of methyl chloride is desired. For example, a mol ratio of 0.1/1 for chlorine to methane at about 450°C over a copper chloride catalyst will yield mostly methyl chloride:

$$CH_4 + Cl_2 \rightarrow CH_3Cl + HCl$$

As the chlorine-to-methane ratio increases, the higher chlorinated methanes are favored. At a chlorine-to-methane ratio of about 1.4/1, methylene dichloride is favored; at about 2.6/1, chloroform is mostly obtained; and at 3.4/1, the end-product is mostly carbon tetrachloride.

Methane chlorination reactions do not require the use of a catalyst, but UV light may be used to activate the molecules. Typical reaction temperatures are in the 350–400°C range.

Other Direct Applications of Natural Gas as a Chemical Feedstock

Oxidative methane coupling at high temperatures and very short residence times would hold great promise in the production of ethylene. As noted throughout, methane (natural gas) is abundantly available at relatively low cost, albeit in remote locations, and its

conversion to ethylene could be of significant commercial interest if and when such an approach can be fully developed. A similar direct approach has been proposed for the conversion of methane to methanol. However, all the results to date show very poor selectivities at moderate per-pass conversions or reasonable selectivities at very low methane conversions. At present, these direct applications can only be viewed as long-term exploratory research projects.

INDIRECT USES OF NATURAL GAS AS A CHEMICAL FEEDSTOCK

In reality, the largest direct use of natural gas as a chemical feedstock is in the production of syn gas, a mixture of hydrogen and carbon monoxide:

By steam reforming (endothermic)

$$CH_4 + H_2O \rightarrow CO + 3H_2$$
$$\Delta H_{298} \sim +49.27 \text{ kcal/mol}$$

where 1 kcal = 4.184 kJ.

By partial oxidation (exothermic)

$$CH_4 + \frac{1}{2}O_2 \rightarrow CO + 2H_2$$
$$\Delta H^0_{298} \sim -8.53 \text{ kcal/mol}$$

or by various combinations of the above (i.e., autothermal reactors, gas-heated reactors, etc.), as might be required to balance the endothermic reaction requirements of steam reforming with the exothermic heat generated by partial oxidation. It should be noted though that catalytic processes for the production of syn gas require deep desulfurization of the natural gas. Noncatalytic processes, like partial oxidation, do not require feed desulfurization and, thus, can easily accommodate heavier feedstocks all the way to refinery residues, but the product syn gas has to be desulfurized for most downstream applications.

Syn gas by itself has no direct commercial use. However, Syn gas, apart from its use in hydroformylation reactions, is the main intermediate in the production of:

- Hydrogen.
- Ammonia and urea.
- Carbon monoxide for carbonylation reactions.
- Methanol.
- Liquid fuels by the Fischer–Tropsch reaction mechanism.

Steam reforming provides the highest hydrogen yield and is preferred when hydrogen is the desired product, as for the production of ammonia or hydrogen for refinery uses (hydrotreating, hydrocracking, etc.). In these cases, the product syn gas invariably is made to undergo the shift reaction:

$$CO + H_2O \rightarrow CO_2 + H_2$$
$$\Delta H^0_{298} \sim -9.83 \text{ kcal/mol}$$

Thus, with steam reforming, it is possible to obtain an aggregate of 4 mol of hydrogen per mol of methane. This ratio is always lower when some form of partial oxidation is used, but that is less of a problem when methanol is desired as end-product (theoretical $H_2/CO = 2.0$) or for gas to liquids (GTL) in the production of hydrocarbon fuels ($H_2/CO \sim 2.0$).

$$2H_2 + CO \rightarrow -CH_2- + H_2O$$

Traditional steam reformers are limited to processing about 100 mm scfd of natural gas in a single train, equivalent to the production of about 3000 t/day methanol (1 scf = 0.02679 Nm^3, inclusive of the temperature correction; see the appendix). While this was regarded as a world-scale unit until recently, current methanol projects under construction are for 5000 t/day and projects are being announced for methanol capacities in excess of 15,000 t/day. Likewise, GTL projects have been announced for capacities well in excess of 100,000 barrels/day (bpsd) of liquids (1 oil barrel = 42 U.S. gallons = 159 L) or a feed stock requirement of almost 1000 mm scfd of natural gas. These applications require huge units for the generation of syn gas and, as a result, partial oxidation or autothermal schemes are often preferred.

METHANOL UTILIZATION

Traditional demand for methanol is fairly limited. Apart from its use as a solvent and as a feedstock for the production of formaldehyde and acetic acid, the use of methanol is not growing in demand significantly. Because of unrelated concerns with ground water contamination in the U.S.A. and a few other countries, MTBE (methyl *tert*-butyl ether), one of the fastest growing outlets for methanol in recent years, has slowed down to almost a halt. Currently, the principal momentum behind the renewed interest in methanol production resides in its application in nontraditional areas. These can be summarized as follows:

- Methanol to gasoline (MTG).
- Methanol to olefins (MTO), especially ethylene and propylene.
- Methanol to propylene (MTP).
- Methanol to dimethyl ether.

Dimethyl ether is being contemplated as a high-cetane, clean burning diesel fuel alternative, or as a substitute for methanol in the MTO or MTP reactions mentioned above.

In all these projects, economics plays a leading role. It is always necessary to secure the feedstock at a very low price (typically natural gas at not more than about \$0.75 per mscf), which is only available in remote locations (e.g., ''stranded'' natural gas) with low local demand. This can be translated into a cost of methanol production of about \$80 per tonne, or perhaps a maximum of about \$90 per tonne if allowances are made for shipping costs, etc. Considering that methanol has a fairly low heating value of only about 60,000 Btu per U.S. gallon (LHV) or about 21.1 GJ/t, these ''low'' methanol prices are not insignificant, as \$90 per tonne translates into about \$4.50 per mm Btu (~\$4.25 per giga-Joule). Not surprisingly, although an MTG unit was built and operated by Mobil near New Plymouth in the Northern Island of New Zealand, the production of gasoline (at fuel value) from methanol proved uneconomical and the unit was shut down.

Utilization of methanol or DME for petrochemical applications has a much better chance of succeeding, as there is a much higher value added in going from methanol to light olefins. Even when the markets are soft, ethylene and propylene command prices of at least about \$400 per tonne, and usually substantially more. Even on a pure fuel value basis, \$400 per tonne for ethylene would be equivalent to about \$8.00 per giga-Joule, or a very significant value-added markup.

A typical MTO unit for the production of about 1 milliont/yr of ethylene and propylene (say, about 500,000 t/yr each) requires about 2.7 million t/yr methanol or about 259 mm scfd (6.9 mm Nm^3/day) of natural gas, which, at \$0.75 per mm Btu, corresponds to a relatively modest annual feedstock expense of about \$65 million. This compares favorably with \$400 million/yr value for the olefin product (if priced at \$400 per tonne) or about \$750 million/yr for the end-product polyolefins (if priced at about \$750 per tonne).

Although the attractiveness of the projects is case specific as it depends on the plant capacity, unit location and local investment requirements, cost of the feed natural gas, scope of the project (whether only olefins or all the way to polyolefins), etc. on the average MTO or GTP (gas-to-polymers) projects provide returns on investment in the order of 20–25%.

Main licensors in this area are UOP LLC in the U.S.A. (together with Norsk Hydro in Norway) with MTO technology for the simultaneous production of ethylene and propylene at adjustable $C_2=/C_3=$ ratios between 0.6 and about 1.3 and Lurgi GmbH in Germany with MTP technology for the production of propylene. The former uses a microspheroidal catalyst based on a SAPO-34 molecular sieve catalyst contained in fluidized-bed reactors scalable up to about 1.5 mm t/yr of light olefins in a single train, depending on the product slate and the operating conditions. Lurgi appears to favor a modular approach with individual trains of three fixed-bed reactors in series using a modified HZSM-5 molecular sieve catalyst supplied by Süd-Chemie;[7] it is believed that each train can produce up to about 200,000 or 250,000 t/yr of propylene. Both processes use proprietary schemes to recover and reprocess heavier olefin byproducts in the $C_4=$ to $C_6=$ range.

At the time of writing this entry, one MTO unit for 800,000+ t/yr of ethylene-plus-propylene had been designed by UOP for Axinova (Eurochem Technologies Corporation Pte Ltd) to be located at the Lekki peninsula near Lagos, Nigeria.[8] ExxonMobil are also showing a high level of patent activity in this area and it is not unreasonable to expect that they may develop one or more commercial MTO projects too.

GAS TO LIQUIDS

Production of liquid hydrocarbons from syn gas makes use of the well-established Fischer–Tropsch technology using either the older Fe catalysts or the more modern Co catalysts. Originally commercialized in Germany toward the early part of the twentieth century and widely used during the World War II, it was further developed by Sasol in South Africa. All these earlier applications used syn gas derived from coal and made use of a variety of reactor types like the multitubular ARGE low-temperature (250°C) reactors first used by Sasol based on the German design or the Synthol high-temperature (350°C) fluidized-bed reactors for a lighter product slate or, more recently, ebullated bed or slurry phase reactors also at the lower-temperature conditions.[9] One of the goals in Fischer–Tropsch technologies is to control the product slate by controlling the degree of polymerization. In a very simplified form, Fischer–Tropsch reactions follow a Schulz–Flory equal-growth-probability polymerization mechanism of the form:

$$P_n = n(1 - \alpha)^2\alpha^{(n-1)}$$

where P_n is the probability (fractional molar concentration) of having a hydrocarbon with an n-long carbon chain in the product, and α is the degree of polymerization, and such that

$$\Sigma P_n = 1$$

In Figs. 8–10, the chain growth patterns that develop as a function of the degree of polymerization[10] are illustrated on the basis of the simplified mechanism shown above. Lower values of α yield lower molecular weight

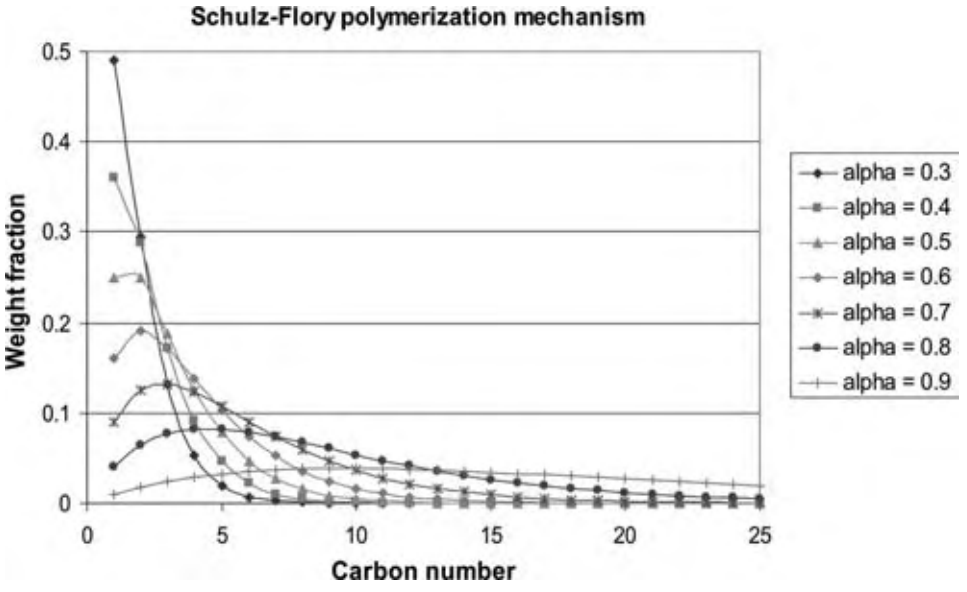


Fig. 8 Fischer–Tropsch chain growth mechanism (Schulz–Flory polymerization mechanism). *(View this art in color at www.dekker.com.)*

products, but unfortunately large amounts of methane also. Practical applications of Fischer–Tropsch usually require relatively high values for $\alpha(>0.8)$ that lead to the maximization of products in the diesel range (say, C_{12} to C_{18}) and also to the production of waxy materials. Often, a mild hydrocracking step is advisable to crack the heavier materials to more useful products, but Fischer–Tropsch can also be used advantageously for the production of waxes and lube oils as more valuable byproducts. Typically, about 80% of the product is in the diesel range while the remaining 20% is a paraffinic naphtha, which is ideal as steam cracker feedstock.

At present, the only operating Fischer–Tropsch-based units in the world are Sasol I, II, and III, all based on coal feedstocks, with an aggregate capacity of about 180,000 bpsd, the natural-gas-based PetroSA (Mossgas) unit, also in South Africa, with close to 35,000 bpsd capacity, and the Shell unit at Bintulu in Malaysia originally designed for 12,500 bpsd, but since revamped to about 15,000–16,000 bpsd. The PetroSA unit uses Synthol high-temperature reactors licensed from Sasol, while the Shell unit uses fixed-bed tubular reactors (Fig. 11).[11]

Many GTL projects have been announced around the world, some of the more credible ones include:

Sasol-Chevron	Escravos, Nigeria	34,000 bpsd
Shell	Ras Laffan, Qatar	140,000 bpsd
Conoco-Qatar Petroleum	Qatar	120,000 bpsd
ExxonMobil	Qatar	100,000 bpsd
Sasol-Qatar Petroleum	Qatar	34,000 bpsd

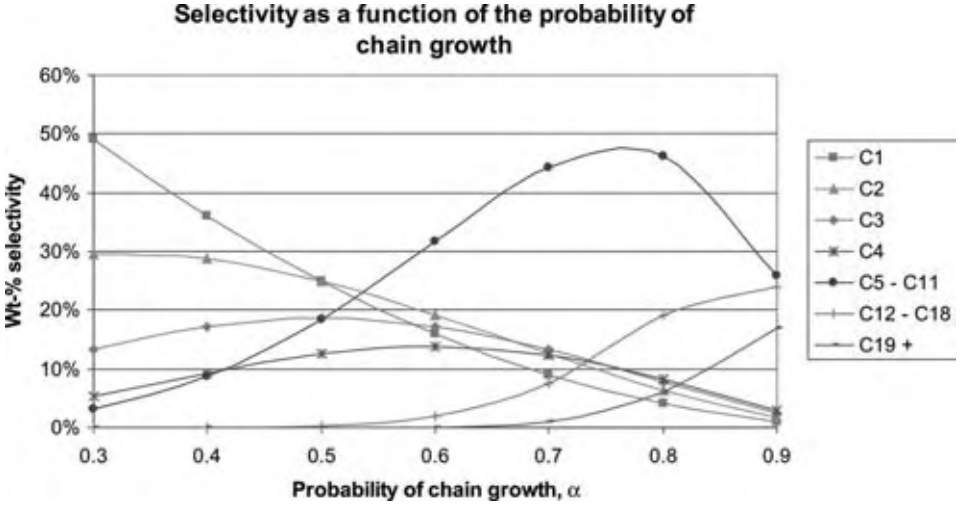


Fig. 9 Fischer–Tropsch chain growth mechanism (selectivity as a function of the probability of chain growth). *(View this art in color at www.dekker.com.)*

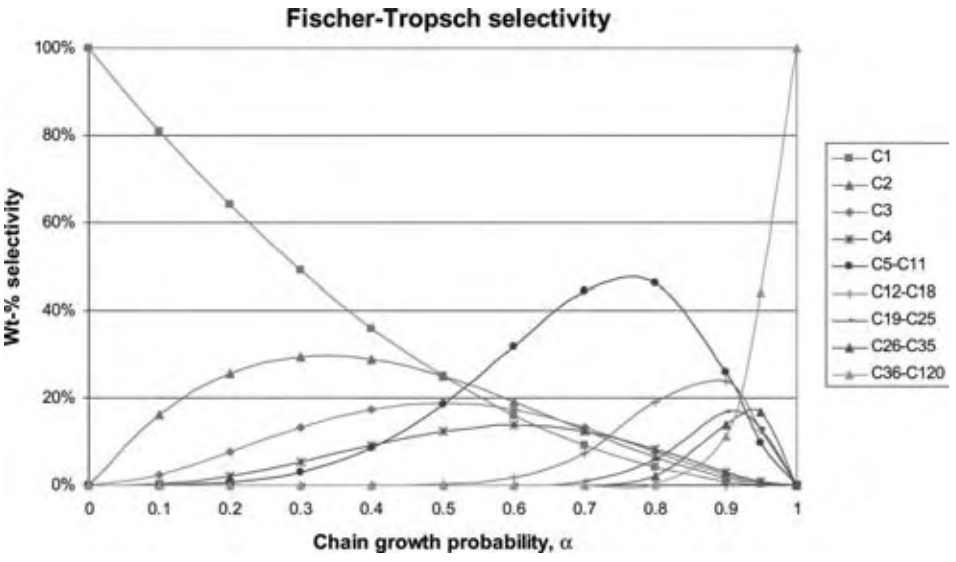


Fig. 10 Fischer–Tropsch chain growth mechanism (Fischer–Tropsch selectivity). *(View this art in color at www.dekker.com.)*

Other projects have also been announced in Iran, Australia, South America, etc. Iran and Bolivia, in particular, appear to be likely candidates for GTL facilities in the future.

Gas-to-liquid plants consume in the order of 9500 scf per barrel of product. Thus, a 100,000 bpsd unit will consume about 950 mm scfd (26.9 mm scmd) of natural gas. Therefore, over a typical project life of about 30 years (10,000 days), the cumulative consumption of natural gas will be 260 mm scm. Only very large natural gas fields can support such production rates. Obviously, Qatar, Iran, and some other countries with very large gas reserves are in a good position to satisfy the needs for such projects.

Although GTL can be advantageously used to add value to natural gas reserves in remote or less industrialized locations, the fact is that the return from GTL facilities is fairly modest, as basically they make

Fig. 11 Reactors at Shell's MDS unit, Bintulu, Malaysia. *(View this art in color at www.dekker.com.)*

fuels (naphtha, diesel, and middle distillates) out of another fuel (natural gas). Even when natural gas is only priced at low values of less than about \$0.75 per mm Btu, the typical rates or return for these facilities are in the order of 10–12% and depend greatly on the investment costs. Better returns can be expected for GTL units that recover more valuable fractions like waxes, synthetic lube oils, *alpha*-olefins, etc. Although older and smaller gas-based units like the Shell unit at Bintulu, Malaysia, were reported to require investments in excess of \$60,000 per daily barrel of installed capacity, more recent reports indicate that, at least for larger world-scale units, the investment costs may be in the order of \$23,000 per daily barrel of installed capacity. These numbers, however, may be taken with a certain degree of skepticism, as they often ignore the large investments for services and infrastructure at remote construction sites. The investment requirements for the 140,000 bpsd GTL unit that Shell is planning to build in Qatar have been disclosed[12] as being in the order of \$5 billion, or about \$36,000 per daily barrel of installed capacity.

CONCLUSIONS

The appetite for natural gas as an industrial fuel continues unabated and is only limited by the supplies available in the major industrialized countries, in particular in the U.S.A., which has led to significant increases in its pricing. Large reserves of natural gas at low cost exist in remote of less industrialized areas. This has led to the increase in the number of LNG facilities and also to the construction of facilities at such locations where natural gas can be advantageously exploited in new ventures for the manufacture of methanol and light olefins and for the conversion of natural gas into liquid fuels for transportation applications.

APPENDIX

Units and Equivalences

Volumetric gas rates are defined either at ''normal'' conditions (0°C and 1 atm or 1013 mbar of absolute pressure) or at ''standard'' conditions; in English units ''standard'' conditions are defined at 60°F (15.56°C) and 1013 mbar, while in metric units ''standard'' conditions are defined at 15°C and 1013 mbar. There is practically no difference between the two ''standard'' conditions, but care should be used not to confuse them with ''normal'' conditions. The abbreviations cf and cm are commonly used to denote ft^3 and m^3, respectively. Volumes defined at ''standard'' conditions are about 1.055 (=288.15/273.15) or 1.057 (=288.71/273.15) times the same volumes defined at ''normal'' conditions.

1 kmol of ideal gas at ''normal'' conditions occupies 22.414 N m^3 or 791.5 N ft^3.

1 kmol of ideal gas at ''standard'' (60°F) conditions occupies 23.690 scm or 836.6 scf.

1 lb mol = 0.4536 kmol.

ACKNOWLEDGMENTS

The author wishes to thank BP p.l.c. for granting permission to reproduce figures and data from the BP Statistical Review of World Energy, and Shell for granting permission to reproduce a photographic view of the reactor section of their GTL unit in Bintulu, Malaysia.

REFERENCES

1. BP Statistical Review of World Energy 2005; http://www.bp.com.
2. Table 3-207. *Perry's Chemical Engineers' Handbook*; 6th. Ed.; McGraw-Hill Book Company, 1984; 3-155–3-157.
3. Ali Mansoori, G.; University of Illinois, Chicago; Private communication.
4. Waddams, A.L.; *Chemicals from Petroleum*, 4th Ed.; Gulf Publishing Company, 1980.
5. Chauvel, A.; Lefebvre, G. *Petrochemical Pricess. Vol. 1—Synthesis-Gas Derivatives and Major Hydrocarbons*, 2nd Ed.; Éditions Technip: Paris, 1985 (in French), 1989 (in English).
6. Process for Manufacturing Alkanes by Reacting Other Alkanes with Methane. BP Chemicals LtdInternational Patent Publication WO03/06655214 August 2003.
7. Chemical Week, 14 February 2001.
8. Chemical Week, 2 October 2002.
9. Hill, C. *Sasol's Experience in the Conversion of Gas into Liquids*. Persian Gulf Gas Resources, Tehran, 7–8 November 1999.
10. Pujadó, P.R. Bent Lecture, University of Missouri, Columbia, 28 February 2002.
11. de Graaf, W.; Schrauwen, F. World scale GTL, Hydrocarbon Engineering 2002 May. See also; http://www.shellglobalsolutions.com/gasevents/gas_liquids/World_Scale_Gas_to_Liquid_Engineering_Choices.pdf.
12. Watts, P.; Fabricius, N. The Qatar Shell Gas-to-Liquids Project, Doha, 20 October 2003. See also; http://www.shell.com/qatar.

New Flame Retardant Materials: Nonhalogenated Additives from Brominated Starting Materials and Inherently Low-Flammability Polymers

Alexander B. Morgan
Dow Chemical Company, Core R&D, Midland, Michigan, U.S.A.

Joshua Jurs
Jason Stephenson
James M. Tour
Departments of Chemistry, Mechanical Engineering and Materials Science, and Center for Nanoscale Science and Technology, Rice University, Houston, Texas, U.S.A.

INTRODUCTION

Flame retardancy of polymeric materials is an important field of science because of its beneficial societal impacts in preventing fire losses. It is also a broad and complex field of science, with many different routes available to provide fire safety with polymeric materials. In this entry, we discuss two classes of materials that impart fire safety. The first is nonhalogenated flame retardant additives derived from halogenated flame retardant feedstocks. The second is inherently low-flammability polymers. We discuss the advantages and disadvantages that these materials bring to flame retardancy in relation to current technology, as well as their flammability performance. Also presented are the chemical syntheses and processing of these materials. These materials are still new, and the larger use and application may not occur until new advances are made or new regulations mandate their use.

BACKGROUND

Fire safety is an all-encompassing need in modern society as the damage caused by fire results in significant loss. According to the National Fire Prevention Association, in 2002 there were 1,687,500 fires, resulting in $10.3 billion of property loss and 3,380 fire deaths in the U.S.A.[1] If one were able to collect data on fires worldwide, the degree of damage and number of fire deaths would be significantly larger, especially in developing or fast-growth countries. Worldwide, plastics and polymers make up a large amount of materials used in everyday life, and in many cases these plastics contribute significantly to fires when ignition sources are present. The plastics and polymers contribute to fires because many of them are inherently flammable owing to their chemical structure being based upon petrochemical feedstocks. Therefore, as plastics contribute to fire risk scenarios, efforts have been taken to provide flame retardancy to these materials. Flame retardancy can be provided to polymeric materials in a variety of ways, including flame retardant additives, the use of non- or low-flammable polymers, and engineering solutions that seek to keep flammable materials away from ignition or fire sources. The use of flame retardant additives is the most commonly used approach, followed by engineering solutions and then the use of inherently fire-safe polymers. The primary reason for this is economics, but other reasons include regulations and market demands. There are advantages and disadvantages to each flame retardant approach. Flame retardant additives have economical advantages and can be highly effective in multiple polymers. However, they can give an undesirable balance of properties to the final polymer formulation, as well as introducing recycling and environmental issues when the product containing that plastic has reached its end of use. In addition, additives can fail to provide fire performance beyond the product requirement test, should that product encounter a fire risk scenario outside of its original design. Engineering solutions can have some economic advantages and they can be easily implemented, but they also can be easily defeated, leading to catastrophic fire loss. Inherently low flammability or nonflammable polymers can withstand a variety of fire risks and provide some of the highest levels of fire safety, but they can be very expensive and/or fail to have the proper balance of properties because of their chemical structures. For the purposes of this entry, we focus on the use of flame retardant additives and non- or low-flammability polymers.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007768

The flame retardant additives market worldwide was $2 billion with a 3–3.5% growth rate per year as of 2002.[2] Despite the healthy growth of this additive industry, there are significant concerns that are causing changes in the flame retardant additive area. The main concern is focused on the perceived persistence, bioaccumulation, and toxicity (PBT) of these additives. As mentioned earlier, one of the drawbacks of additives is that they can leach out of the plastic, either during the lifetime of the final article or during end-of-use which includes landfill, recycling, or incineration. While there are no disagreements with the fact that flame retardants provide fire safety during the use of the product, and this safety performance has a clear environmental benefit in the prevention of fire,[3] the impact of the additive at the end the product's lifetime is causing concern with government regulators.

Halogenated flame retardants are the main class of flame retardants coming under regulatory scrutiny because of their perceived PBT. In some cases, extensive scientific studies have shown that some of the halogenated flame retardants have definite PBT characteristics, and these materials have been banned from further manufacture and use.[4–6] It should be noted, however, that other studies are inconclusive and the use/lifetime of some halogenated flame retardants is still in question. From a product lifetime point of view, many of the halogenated flame retardant additives have been in use for decades, some for over 50 years, and their effective service may be coming to an end simply because they are being replaced by less expensive or newer products perceived or shown to be more environmentally friendly in their life cycle analysis. Whether a product is environmentally friendly or not may be subjective. Careful data analysis is required to determine if a material is compatible with the environment.

The next major class of flame retardant additives that are nonhalogenated is the phosphorus-based flame retardants, but even these materials have some regulatory environmental concerns.[7,8] Other nonhalogenated flame retardants that are not phosphorus-based exist, including mineral fillers (i.e., $Al(OH)_3$, $Mg(OH)_2$),[9–11] expandable graphite,[12–15] melamine,[16,17] and polymer nanocomposites combined with other flame retardants.[18–26] Each of these materials has its own advantages and disadvantages, and effectiveness in one polymer system often does not translate into another system.

Inherently fire-safe polymers do not have any of the regulatory issues associated with additives because they can be recycled/incinerated with minimal issues. Today, some fire-safe polymers are in use where the product requirements demand them, and cost is not as large an issue as it is in the additive applications. These applications typically include mass transport (i.e., aircraft, subways, trains, buses, and maritime shipping) where fire risk must be addressed to prevent significant human or material losses. Currently, market forces and local regulations that perceive additive FR solutions as negative are beginning to embrace inherently fire-safe materials as the replacement technology.

To summarize, the field of flame retardancy for polymers is in a state of flux, and multiple new technologies and approaches are expected to arise in the coming decade. In this entry, we focus on some new technology from our laboratory, including nonhalogenated flame retardants synthesized from brominated starting materials, and inherently fire-safe and low-flammability polymers.

NEW NONHALOGENATED FR ADDITIVES

Before describing the new nonhalogenated FR additives, we must first explain how polymer combustion occurs and how various flame retardants work. Polymer combustion occurs when the polymer has undergone enough thermal degradation to volatize into flammable decomposition species, such as monomer and polymer fragments, which then combine with oxygen for combustion upon the application of an ignition source. Once combustion begins, the oxygen is consumed at the flame front, and heat from the flame radiates back down to the polymer, assisting in additional polymer thermal degradation, melting, and fuel pyrolysis. The process continues until the flame is extinguished or the fuel is completely consumed (Fig. 1).[27–29]

To provide flame retardancy to the polymer, there are three main mechanisms that a flame retardant can use, and the most effective flame retardants use two or all three mechanisms in concert. The first is by altering the gas phase or combustion chemistry. Additives that work by this mechanism are referred to as vapor phase flame retardants (see Fig. 1) during combustion. Halogenated and some phosphorus flame

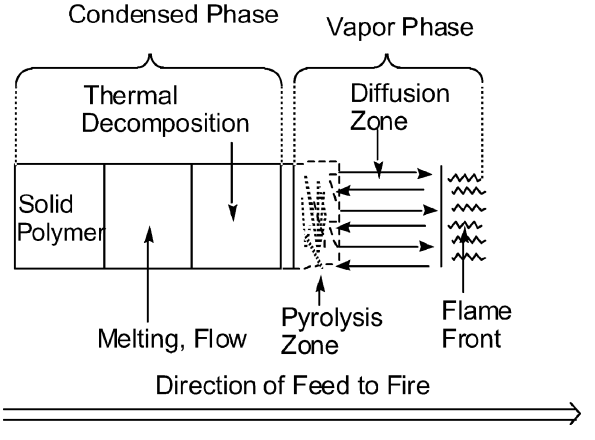


Fig. 1 The polymer combustion process in fires.

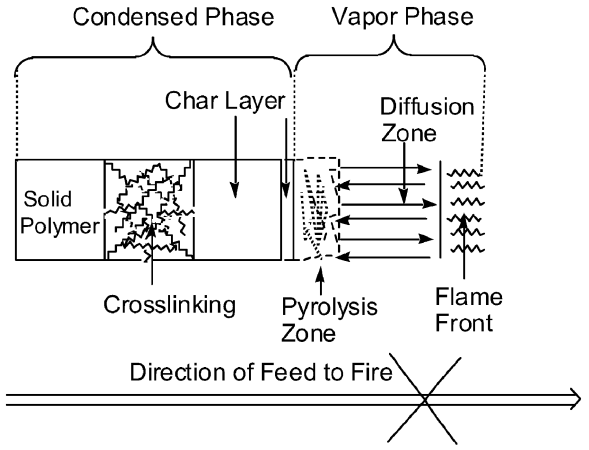


Fig. 2 Formation of a char layer extinguishes polymer fires.

retardants work by this route by reducing heat from combustion through the scavenging of reactive free radicals, in either a catalytic or a stoichiometric manner. The second is by cooling the system through endothermic decomposition. These materials are usually metal hydroxides or carbonates that release nonflammable gases, such as water or CO_2, which cool the polymer and dilute the total amount of fuel going to the flame front. The third mechanism by which flame retardancy is achieved is through crosslinking or char formation. There are a variety of flame retardants that rely on this mechanism through numerous chemical pathways. Char formation is a strictly condensed phase phenomenon. It works by preventing fuel release through the crosslinking of the polymer matrix or by inducing the formation of high thermal stability chemical bonds. The prevention of fuel release retards combustion, which in turn slows down the rate of flame growth and further polymer decomposition. Additionally, as char forms, it may become thermally stable enough to act as an insulator to prevent heat from being transmitted to nondecomposed polymer, thus extinguishing the flame and protecting the non-heat-exposed polymer (Fig. 2). There is an additional benefit to char formation in that it will keep the heat release rate of the polymer low during burning. This will lower the fire risk to other items near the burning polymer, as low heat release rates have much lower chances of propagating fire to nonflame retardant objects (Fig. 3).[30,31]

As mentioned earlier, brominated flame retardants are under regulatory scrutiny that may lead to their eventual abandonment as chemicals used in industry. The chemistry of bromination is well understood, and it can yield a variety of different products and structures. In light of this, brominated starting materials, in particular brominated flame retardants, could be a very useful route into new nonhalogenated flame retardants, and the C–Br bond is usually quite reactive to a variety of organic chemical reactions. Work in our laboratory focused on the use of brominated flame retardants as starting materials in flame retardant additive syntheses and their conversion to nonhalogenated FR derivatives.[32–34]

The first class of compounds, which were nonhalogenated and nonphosphorus containing in structure, was alkyne-functionalized materials made from brominated aromatic compounds via a Pd/Cu mediated crosscoupling reaction.[35,36] The brominated aromatics were all commercial flame retardants, some of which presented synthetic challenges because of their insolubility when highly brominated. For example, decabromodiphenyl ether (**1** where X = Br) is a commonly used flame retardant in a variety of applications.[37] However, it is highly insoluble in most organic solvents. It was found that as the reaction proceeded over time, the addition of each phenylalkyne group to the structure improved its solubility, leading to additional reaction and eventual conversion of a majority of the aromatic bromides (only 1–2 wt.% Br remained unreacted) to alkynes. Other materials, such as the tetrabromobisphenol A carbonate oligomers (**2** where X = Br), reacted with 100% Br conversion in less time, because of a combination of better solubility and carbonate functionality assisting the Pd/Cu reaction. The choice of amine base, palladium catalyst, and ratio of alkyne to bromide under reaction conditions does need to be optimized with each aryl halide

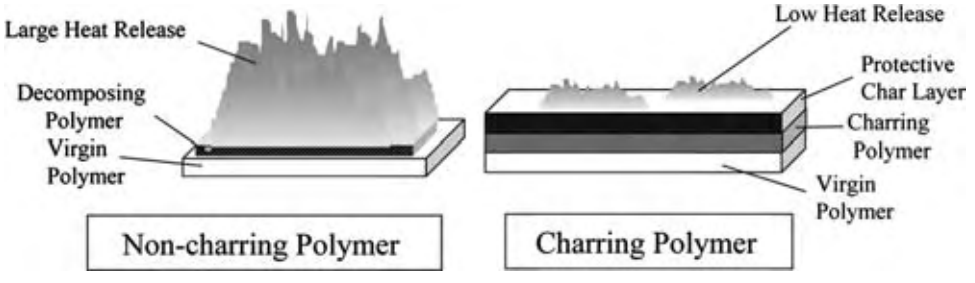


Fig. 3 Large heat release (left) leads to larger fires while low heat release because of char formation (right) retards fire propagation. (From Ref.[30].) *(View this art in color at www.dekker.com.)*

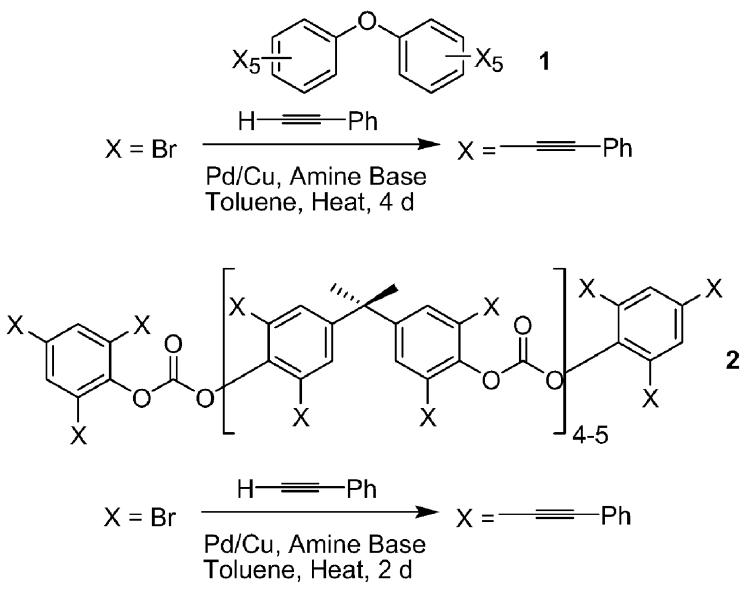


Fig. 4 Synthesis of alkyne-based flame retardants.

one considers for this type of reaction. A schematic of the alkyne-based flame retardants is shown in Fig. 4.

As mentioned earlier, phosphorus flame retardants are another class of commonly used materials, and they come in both halogenated and nonhalogenated versions. Just as alkyne-based flame retardants were made from halogenated materials, attempts were made to synthesize alkyne/phosphorus materials via the same Pd/Cu chemistry with phenylacetylene and an aryl bromide phosphate. However, it was found that the phosphate prevented the reaction from occurring; thus a different synthetic route was chosen that is far more versatile in producing derivative chemical structures. For the synthesis of existing phosphate flame

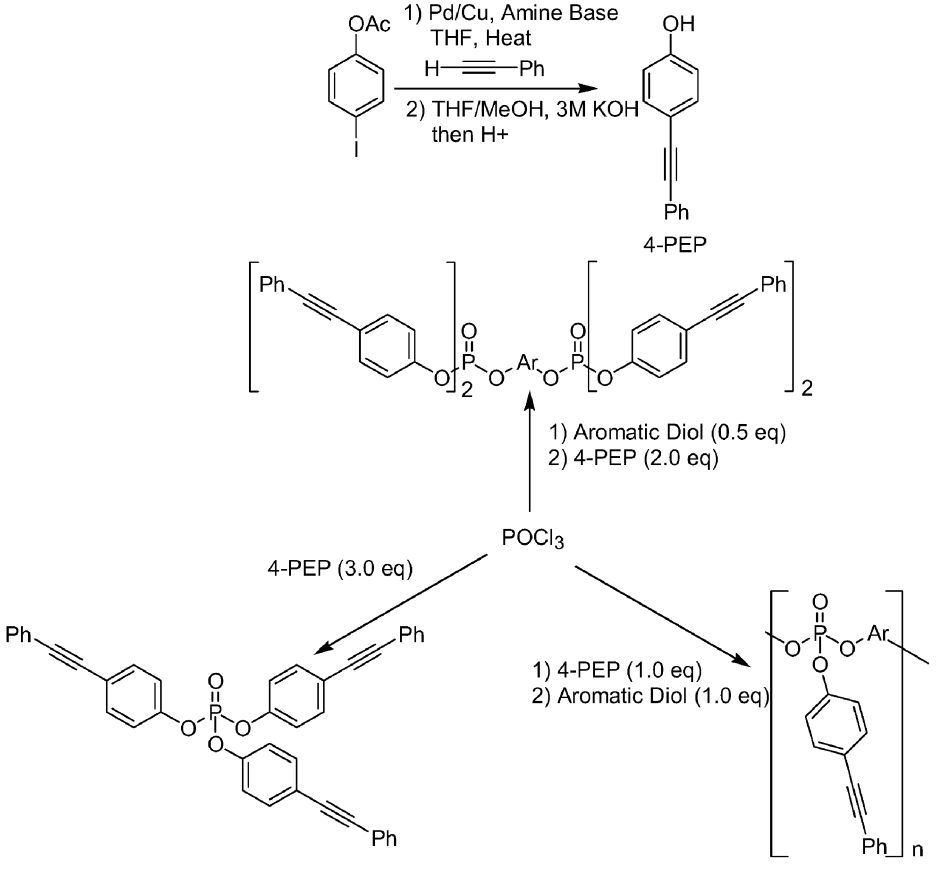


Fig. 5 Synthesis of phosphorus-based flame retardants.

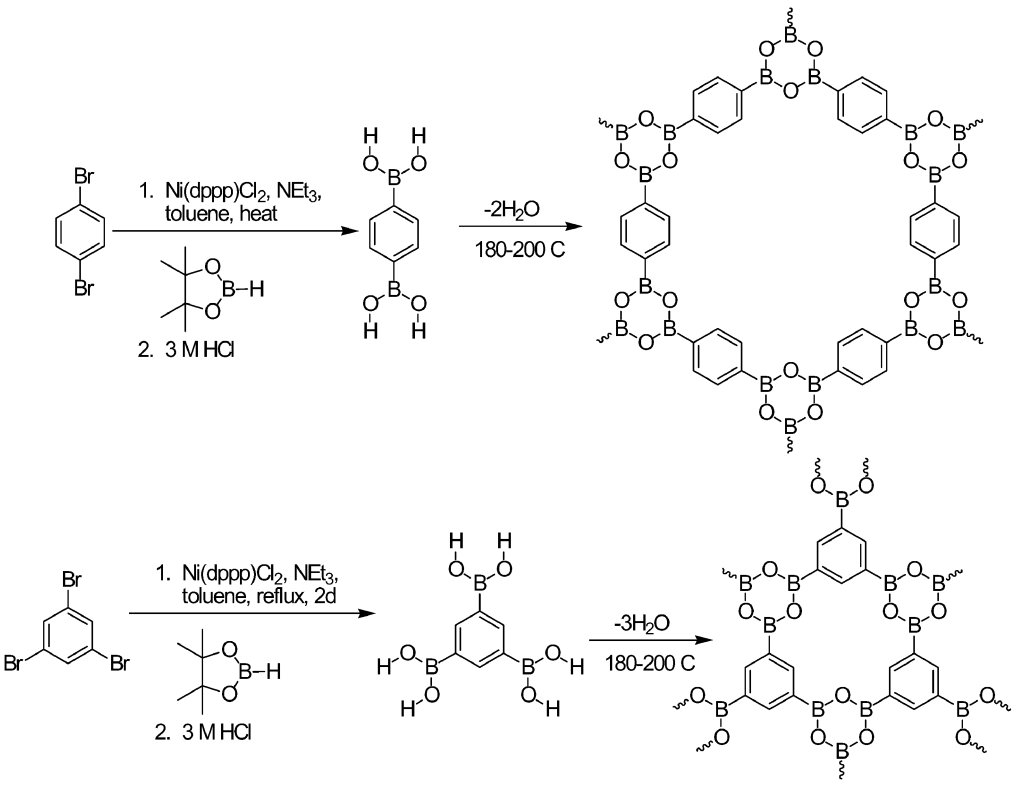


Fig. 6 Synthesis of aromatic boronic acid flame retardants, showing the dehydration process that can take place in formulated products when subjected to heat.

retardants, the primary starting material is phosphorus oxychloride ($POCl_3$), whereupon various alcohols and phenols are reacted to give the resulting phosphate. To synthesize alkyne/phosphorus materials, it is necessary to put the alkyne functional group on phenol before adding the phenol to $POCl_3$. The phenol (4-phenylethynyl phenol, or **4-PEP**) was synthesized from 4-iodophenyl acetate. The acetate group is important, as the phenol can inhibit the Pd/Cu reaction by preventing oxidative addition between the C–I bond. The iodide, rather than bromide, is essential for reaction, as the reactivity of 4-haloaryl acetates is quite low to Pd/Cu conditions, with chlorides being unreactive and bromides reacting only in low yields over long reaction times. There are some new generation catalysts and ligand combinations that permit these types of reactions to proceed with aryl chlorides;[38–43] however, at the time of our research, they were not available. Once the Pd/Cu-catalyzed reaction is complete and the phenylethynyl group has been added, the acetate can be removed with the use of base, followed by acid quench to give the phenol. This phenol can then be esterified with $POCl_3$, and by controlling the ratio of $POCl_3$ to phenol and 1,3-/1,4-aromatic diols (such as bisphenol A, hydroquinone, and resorcinol) along with the order of addition, a variety of diphosphate or polyphosphate structures can be obtained (Fig. 5).

While synthesis of these alkyne-based or phosphorus/alkyne materials can be accomplished, melt compounding the materials into a polymer to impart flame retardancy is not straightforward. Compound **1** (Fig. 4) and other 1,2-dialkynyl materials synthesized in our laboratory were found to have poor thermal stability above 200°C. More specifically, the alkyne groups began to react via Bergman cyclizations[44–47] to generate a variety of crosslinked highly aromatic species. Therefore, these materials would only be suitable in polymers that can be processed below 200°C, otherwise the alkyne groups would be nonfunctional under fire conditions and the alkyne crosslinking during polymer processing would result in polymer degradation. 1,3- and 1,4-Dialkyne materials have much higher thermal stability in that the alkyne groups do not begin crosslinking until 350–400°C.[48] These materials were compounded into a wide range of polymers, but the

$$H\bullet + HX \longrightarrow H_2 + X\bullet$$

$$HO\bullet + HX \longrightarrow X\bullet + H_2O$$

$$X\bullet + RH \longrightarrow HX + R\bullet$$

$$X = Cl \text{ or } Br$$

Fig. 7 Capping of radicals in the vapor phase with halogens.

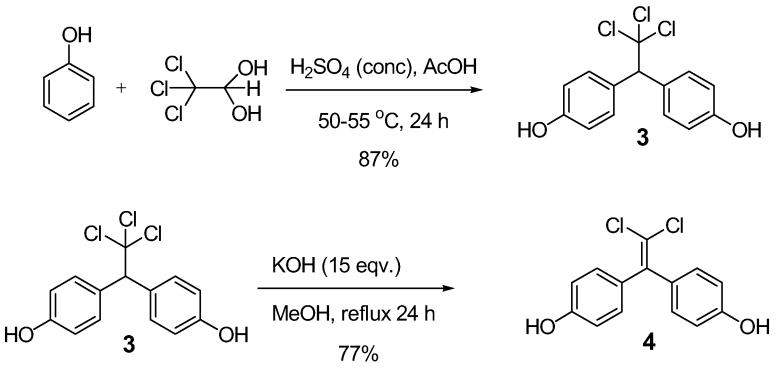


Fig. 8 The classical synthesis of bisphenol C (BPC).

flame retardancy they provided was limited. Although the presence of the alkyne provided a significant amount of char formation, thus preventing fuel from being volatilized and combusted, the heat generated during the crosslinking/char formation process overwhelmed any benefits because of charring. The phosphorus/alkyne materials were more effective because the phosphate structure is capable of crosslinking chemistry in some polymers and, when combined with the alkynes, gave good flame retardancy in polycarbonate.

The final category of flame retardants synthesized from brominated starting materials is the aromatic boronic acids. There are several ways to make aromatic boronic acids from brominated aromatics, including Grignard,[49–51] lithium–halogen exchange,[52] and catalytic routes.[53–57] Grignard and lithium–halogen exchange routes gave complex mixtures with impurities that could not easily be removed, and in some cases not all of the halogens were exchanged when more than one halogen per aromatic ring was present. We found that aromatic monoboronic acids were ineffective as flame retardants, but di- and triboronic acids were very effective at low loadings in polycarbonate and poly(acrylonitrile–butadiene–styrene) (ABS).[58] Catalytic routes to making di- and triboronic acids were the most useful, but optimization of the chemistry is still needed. The catalytic route found to be effective in producing boronic acids was the use of pinacol borane and a nickel(II) catalyst. The reaction works through an oxidative addition between the Ar–Br bond, followed by transmetallation with the pinacol borane, and finally reductive elimination to yield the aryl pinacol borate. Acid hydrolysis of the pinacol ester gave the final boronic acid (Fig. 6). Specifically, we synthesized two aromatic boronic acids: 1,4-benzene diboronic acid, and 1,3,5-benzene triboronic acid.

As mentioned with the alkyne additive above, melt compounding can affect the structure of the additive. If this material is compounded below 180°C, the boronic acid structure remains intact. However, if compounded above 180–200°C, the boronic acid loses water and condenses to form a boroxine network structure. Depending upon the boronic acid substitution, a variety of boroxine networks can be formed. We found that the boronic acids were ineffective as flame retardants in high-impact polystyrene, but were effective in ABS and polycarbonate. The latter two are compounded at temperatures above 200°C; therefore the active flame retardant species is the boroxine network. However, there is a limit to the effectiveness of the boroxine network as a flame retardant, as antagonistic flame performance is found at high loadings of boroxine in polycarbonate. Furthermore, the loss of water during melt compounding results in polycarbonate

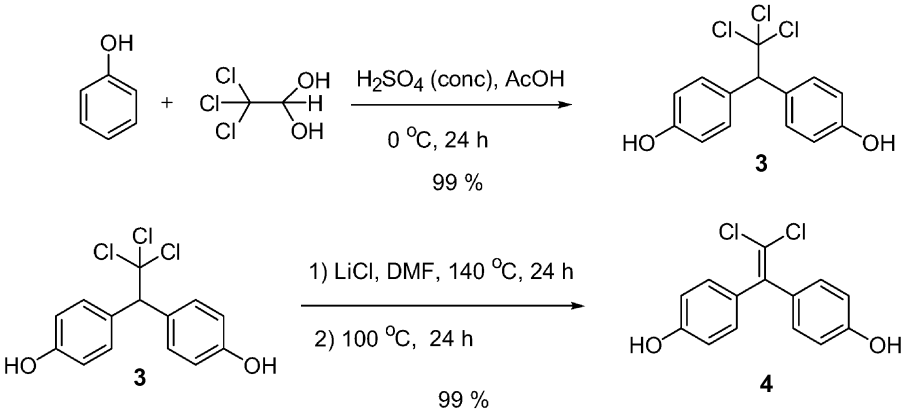


Fig. 9 The improved synthesis of BPC.

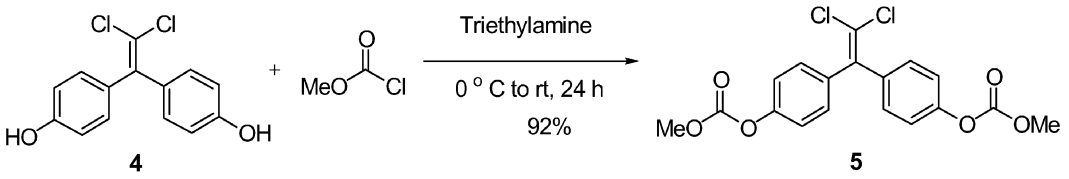


Fig. 10 Synthesis of the BPC bis(methylcarbonate) monomer from BPC.

molecular weight degradation by the hydrolysis of the carbonate groups. The loss in polycarbonate molecular weight results in higher flammability, rendering the flame retardant less effective during a fire. As removing all of the water from the boronic acid/boroxine network prior to compounding is difficult, there is a limit to the effective loading for this flame retardant additive. At 5 wt% loading in PC, UL-94 5V and V-0 ratings were achieved at the 3.2 mm sample thickness. In ABS, the limit does not exist, as the release of water will not affect ABS molecular weight during melt compounding. At high loadings of boronic acid/boroxine network, UL-94 V ratings were not achieved in ABS; however long burn times with no dripping (5–6 min to consume an entire burn bar) were observed. Presumably with an additional synergist this material would be more effective in ABS. The mechanism of flame retardancy for these materials is presumably a condensed phase mechanism, with the boroxine network assisting in char formation in PC and complexing with the nitrile groups in ABS to assist in char formation and antidrip behavior. The boroxine systems act antagonistically with additional halogenated flame retardants; as it appears, the formation of hydrogen halide, either Cl or Br, results in degradation of the char that the boroxine forms during a fire. There is still more to be learned about this class of materials, and additional studies should help to improve their performance and utility in a wider range of polymers.

NEW INHERENTLY NONFLAMMABLE POLYMERS

The chemical structure of a polymer determines its flammability, or to be more specific, the chemical structure of the polymer degradation product dictates the base flammability of the polymer. When a polymer breaks down there are several pathways by which the polymer can decompose or "unzip" to form high energy free radicals. In Fig 1, it is illustrated that a burning piece of plastic has two distinct phases: a vapor phase and a condensed phase. In the vapor phase, the combustion cycle can be stopped by physically diluting the flame with noncombustible gases and chemically reacting the high energy free radicals with halogens.[37] Obviously, materials that decompose to give flammable small molecules, such as ethylene from polyethylene, are highly flammable materials, whereas materials with nonflammable species (such as halogen) are inherently nonflammable, as they release species that inhibit the flame. Materials such as polyvinyl chloride (PVC) and polytetrafluoroethylene (PTFE) have very low flammability and nonflammability, respectively, because of their chemical structure and thermal degradation products.[59]

Halogenated flame retardants inhibit the flame efficiently by acting as a radical trap. The halogen radicals, chlorine and bromine in particular, catalytically stop the radical chain by reacting with the high energy free radicals, $HO^{\bullet}$ and $H^{\bullet}$, in the vapor phase to produce lower energy radical species that eventually lower the energy of the entire combustion cycle, thus extinguishing the flame once the free-radical pathway can no longer be sustained (Fig. 7).[37] The generation of H_2O from $HO^{\bullet}$ is a highly exothermic reaction in the chain branching step that should be mitigated to ensure efficient flame retardancy.

Bromine and chlorine are two reactive halogens that have worked well for flame retardant purposes. Organofluorine compounds have not been as effective as flame retardant additives because the carbon–fluorine bond energy is so high that other events dominate at temperatures where halogenated flame retardants operate.[60]

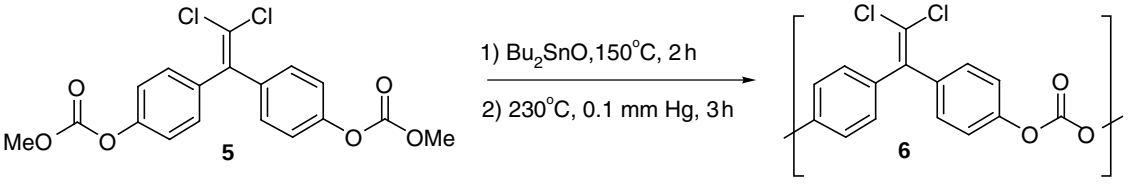


Fig. 11 Polymerization of BPC bis(methylcarbonate) **5** to polymer **6**.

Table 1 Pulsed combustion flow calorimetry (PCFC) data for the polycarbonates: Comparison of PCFC data for polymer **6** to that for BPA polycarbonate

Polymer	HR capacity	Total heat (kJ/g)	Char yields (%)
BPC**6** polycarbonate	29	3	50.1
BPA polycarbonate	359	16.3	21.7

Organofluorine compounds are used as extinguishing agents to put out high-intensity fires, such as aircraft fuel-pool fires, whereas polymer fires tend to have lower temperatures, and therefore materials with C–F bonds tend not to be effective flame retardant additives. Conversely, the iodine–carbon bond is so weak that it is easily cleaved by light and the released HI or I_2 is leached out of the plastic, making it unsuitable for many applications.[60] Halogenated flame retardants can work effectively by themselves, but there are cases where their action must be enhanced by the addition of a synergist to achieve the right balance of properties in the final product. The synergist acts to improve the activity of the additive in the polymer, thus lowering the amount of halogenated additive needed. One synergist, antimony oxide (Sb_2O_3), acts as a halogen shuttle bringing SbX_3 into the vapor phase.[37,60]

The goal of the polymer industry is the design of a low-cost flame retardant polymer that does not require additives yet still has favorable physical and mechanical properties. New flame-resistant engineering plastics that have been commercialized are used in specialty applications. These include polyaryl sulfone and polyetheretherketone (PEEK). These polymers work as heat-/flame-resistant polymers, but because of their high cost of production and poor processibility, their commercial use has been limited.

The monomer that we have used as a backbone for our work toward flame retardant polymers is commonly called bisphenol C (BPC) or 1,1-dichloro-2,2-bis(4-hydroxyphenyl) ethylene. As has been shown by many research groups, BPC can be used as a blendable additive in a commercial plastic or as part of a polymer backbone to effectively impart flame resistance to certain polymeric materials.[61–69] When thermally decomposed, BPC exothermically produces volatile products such as HCl and CO_2, and the unique structure formed upon thermal degradation leads to a very stable carbon structure (char).[70–72] It is this pyrolysis byproduct and the high char forming nature of BPC that give inherently low flammability and flame retardancy (Fig. 3) in these polymers and blends.

The thrust of our research has been to incorporate the BPC moiety into a polymer backbone that can impart flame retardancy without additives. The incorporation of this monomer into a thermoplastic has been approached in several ways including the following: nucleophilic aromatic polymerizations,[73] nucleophilic displacement under phase transfer conditions (PTC),[74] diene metathesis,[75–77] and vinyl addition polymerization.[78]

While halogenated flame retardants have PBT perceptions, halogenated polymers do not have this problem, provided that a polymer with halogen in the backbone does not degrade to give structures with PBT concerns. By incorporating the halogen into the polymer backbone, it is not readily available for bioaccumulation or environmental degradation. Also, these halogenated materials are more easily recycled,[79] as they do not degrade upon multiple regrind/remelts during plastic recycling, and do not lose flame retardant effectiveness, thus further decreasing the environmental impact of inherently nonflammable, halogenated polymers.

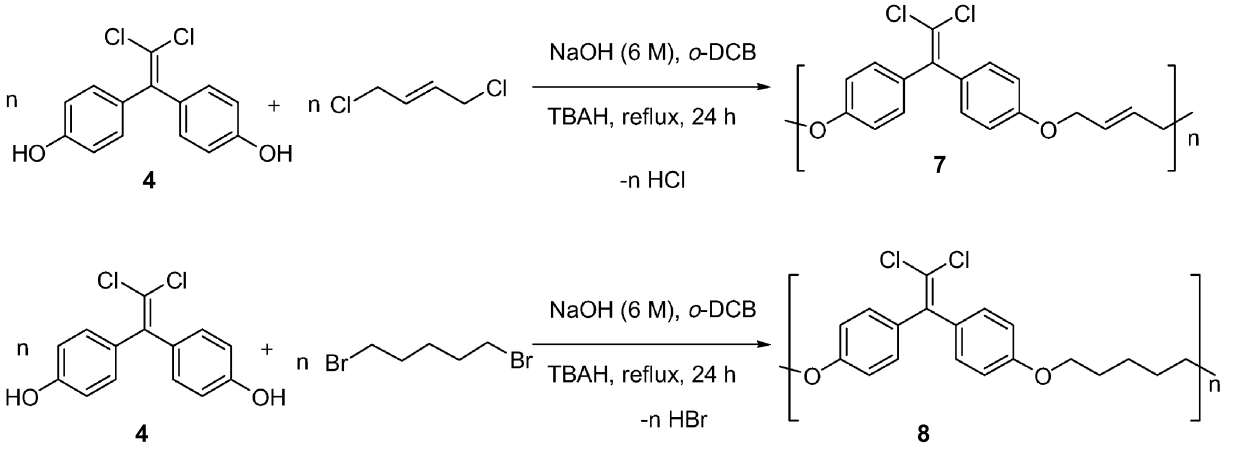


Fig. 12 Synthesis of polyethers **7** and **8** via phase transfer catalysis.

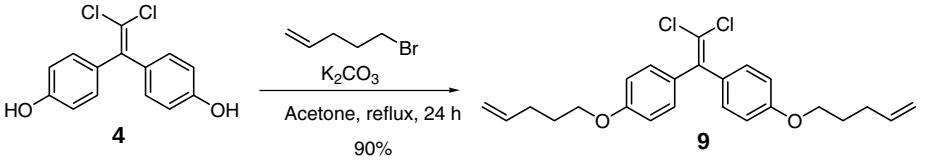


Fig. 13 Preparation of diether monomer **9** via alkylation of BPC **4**.

BISPHENOL C AND BPC POLYCARBONATE

As it was believed that bisphenolic compounds would have laxative properties, the original research on that class of compounds was done by Ex-Lax Inc.,[80] who pursued the development of BPC.[81] The synthesis of BPC was carried out by condensing phenol with chloral to produce 1,1,1-trichloro-2,2-bis(4-hydroxyphenyl) ethane (**3**) followed by a dehydrohalogenation to give the corresponding dehydrohalogenated compound BPC (**4**) (Fig. 8).[80]

A newer process for synthesizing BPC results in near quantitative yields (Fig. 9).[60] Several decolorizing and recrystallization steps are needed to obtain material that is suitable for condensation polymerizations, but overall this procedure requires less workup time and produces higher yields than the original protocols.[61–69] Compound **4** is now commercially available from Sigma–Aldrich.

The use of BPC as a flame retardant additive or polymer was not explored until General Electric (GE) examined its use in the late 1970s.[59,82] The flame-resistant properties of polycarbonate were improved by copolymerizing BPC with bisphenol A (BPA).[59,82] Phosgene gas is used in GE's process but because of the high toxicity of phosgene, we opted to use a non-phosgene-containing polymerization procedure for our small scale investigations.[53–55] In this new process to make carbonate monomers, BPC was transformed into the dicarbonate by condensation with methyl chloroformate to give BPC bis(methylcarbonate) (**5**) (Fig. 10).[83] Compound **5** was then polymerized into **6** using dibutyltin oxide as a catalyst, with high heat and vacuum to drive off CO_2 and MeOH (Fig. 11).[83]

Polymer **6** was tested in a pulsed combustion flow calorimetry (PCFC) apparatus (Table 1). The flame retardant properties of this polycarbonate were previously tested by GE, so no full scale flammability testing was pursued in our laboratory.[59] The PCFC is a small scale flammability testing instrument developed by the FAA for its fire-safe materials program.[84–87] This instrument measures the oxygen consumed by a sample exposed to rapid pyrolysis temperature profiles, thus measuring the heat release (HR) capacity as well as the total heat evolved from the sample. The instrument measures the inherent flammability of a polymeric material; the lower the HR capacity and total heat, the less flammable the material. The PCFC results obtained from the BPC polycarbonate show a much lower heat release capacity and also a much higher char yield, compared with traditional BPA polycarbonate, thereby underscoring the effectiveness of the BPC system.

BPC Polyarylethers

The inflexibility of most aromatic polymers is a problem because they do not melt and flow, making processing the polymers difficult. To make flexible polymers with reasonable melting points and glass transition temperatures, aliphatic and olefinic groups were added to the polymeric backbone. The problem in using these types of functional groups is that they

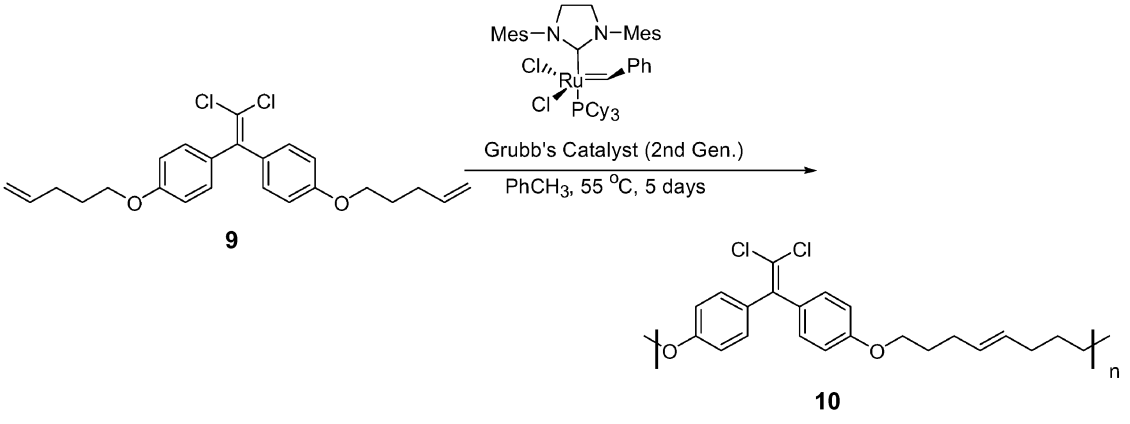


Fig. 14 Acyclic diene metathesis polymerization of **9** to give polymer **10**.

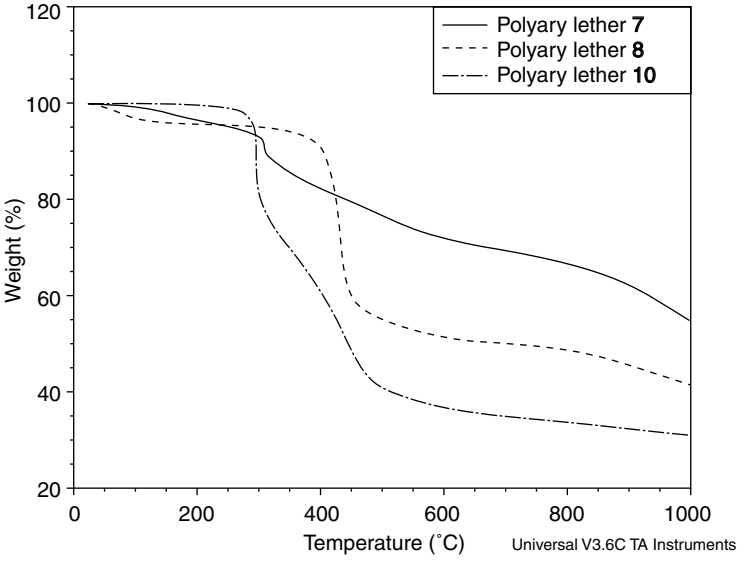


Fig. 15 Thermogravimetric (TG) analyses of polymers **7**, **8**, and **10** to determine weight loss on heating. Weight retention indicates some flame resistance.

might increase the amount of fuel for the fire, and therefore the flammability of the polymer. On the other hand, it is known in flame retardant chemistry that olefins can crosslink during the burning process and increase char formation.[69] In addition, Wagener describes some of his unsaturated polyethers as having a high degree of thermal stability as measured by thermogravimetric analyses.[88]

The first polymer with a flexible backbone was synthesized by a process known in the literature for making bisphenol A polyaryl ether under phase transfer catalysis conditions.[83] We used this procedure to couple BPC with 1,4-dichloro-2-butene (DCB) or 1,5-dibromopentane in *ortho*-dichlorobenzene (*o*-DCB) to yield polyethers **7** and **8** (Fig. 12).

The next approach for obtaining higher molecular weight polymers was to explore acyclic diene metathesis (ADMET) polymerizations.[75–77] The aim was to achieve higher molecular weight flame-resistant polymers. We modeled the reaction using aliphatic diene monomers such as 1,5-hexadiene and 1,9-decadiene under test conditions to optimize conditions before making BPC-derived products. At this point, we decided to functionalize the BPC with an olefin. Wagener has done considerable work on the influence of the neighboring group effect, and how it affects the ADMET reaction. These studies have concluded that atoms such as oxygen and sulfur act as Lewis bases, complexing with the metal (Ru) and hindering the metathesis reaction.[88] To circumvent this problem, three methylene spacer groups were incorporated into the monomer. This was achieved by condensing BPC with 5-bromo-1-pentene under basic conditions to yield monomer **9** (Fig. 13). After purification of this monomer by multiple recrystallizations from methanol, we polymerized **9** using ''second generation'' Grubb's catalyst[88] in toluene at 55°C with a positive flow of argon for 5 d to yield polyether **10** (Fig. 14).

The incorporation of the aliphatic (pentyl) and olefinic (butenyl and octenyl) functionalities in the present work has yielded processible polymers with a defined T_g. In addition, these materials are flame retardant with little to no additives being required (1 wt% PTFE was used as an antidrip additive) to pass vertical flammability tests, such as UL-94 V.

Polyarylether **7** was the first to be synthesized using PTC conditions. It is the most thermally stable of the three polymers. Comparisons of the TG analysis

Table 2 PCFC results for the polyarylethers (comparison of PCFC HR capacity, total heat, and char data for polymers **7**, **8**, and **10** to polyethylene and polystyrene)

Polymer	HR capacity (J/g K)	Total heat (kJ/g)	Char yields (%)
Polyarylether **7**	37	8	58
Polyarylether **8**	205	13	37
Polyarylether **10**	146	18	20
Polyethylene	1676	42	0
Polystyrene	927	39	0

Table 3 UL-94 test results for the polyarylethers

Additive	First ignition[a] (s)	Observed dripping[b]	Second ignition[a] (s)	Observed dripping[b]	UL-94 rating
Polyarylether **7**, 1 wt.% PTFE	3, 0	No, no	0, 0	No, no	V-0
Polyarylether **8**, 1 wt.% PTFE	1, 6	No, no	3, 0	No, no	V-0
Polyarylether **10**, 1 wt.% PTFE	6, 0	No, no	0, 2	No, no	V-0

[a]Time to self-extinguishing after the first and second 10 s ignition in two sample runs.
[b]Molten sample did (yes) or did not (no) drip onto cotton patch underneath ignited bar during UL-94 test.

(Fig. 15) and PCFC data from the three polyarylethers show that **7** gives the highest char yields (58%) followed by **8** and **10** (Table 2). The thermal stability of the polymers is reduced with the incorporation of more aliphatic/olefinic groups, which results in a lower percentage of the char yields, which is apparent when you compare heat release capacity and char yields of polyarylethers **7** and **10.**

The PCFC total heat release data correlate well with the TG analysis char yields. Polyarylether **10** has a lower heat release capacity than polyarylether **8**, but also has a lower char yield. The thermal stability of these polymers can be attributed to the BPC structure; in addition, the olefin functionality in polyarylethers **7** and **10** might also give rise to its thermal resistance.[89] Almost half the polymer weight in **7, 8**, and **10** is retained; this indicates some flame resistance because we would have otherwise expected complete incineration (Fig. 15).

The flammability of the three polyarylethers **7, 8,** and **10** was measured using the UL-94 V test, and all three samples obtained a V-0 rating when mixed with 1 wt.% PTFE as an antidrip additive. It is important to note that the samples were tested without antidrip additive, and did not burn but dripped excessively, making characterization impossible. It was observed that while these polymers where subjected to prolonged exposure to the Bunsen burner flame, the fire never consumed them. The UL-94 sample results are shown in Table 3.

Asymmetric BPC Polymers

The incorporation of flexible linkers for making processible polymers, which are found in vinyl addition polymers, is not usually considered flame resistant and therefore is not generally a suitable option for

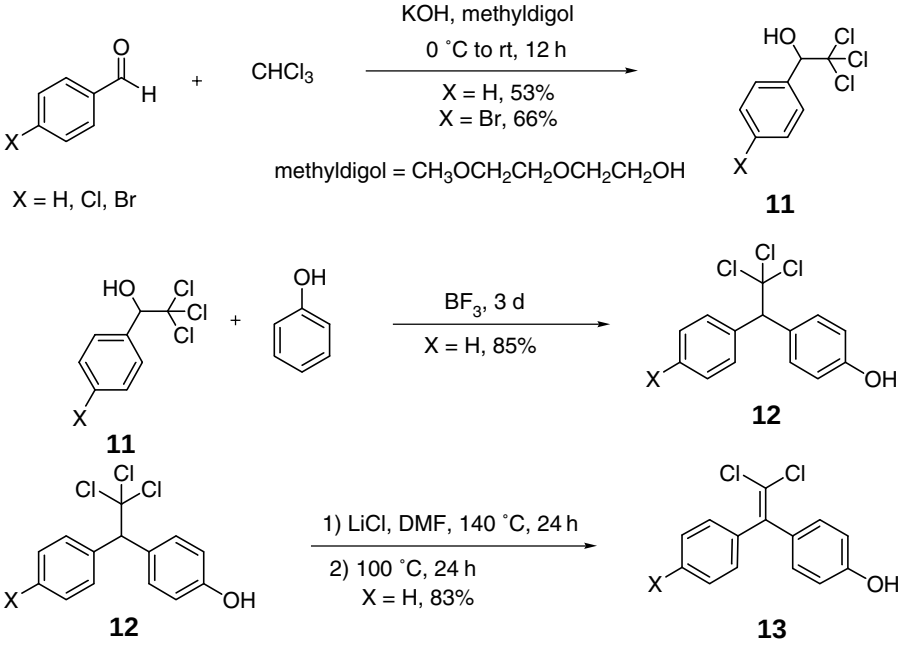


Fig. 16 Synthesis of asymmetric BPC **13**.

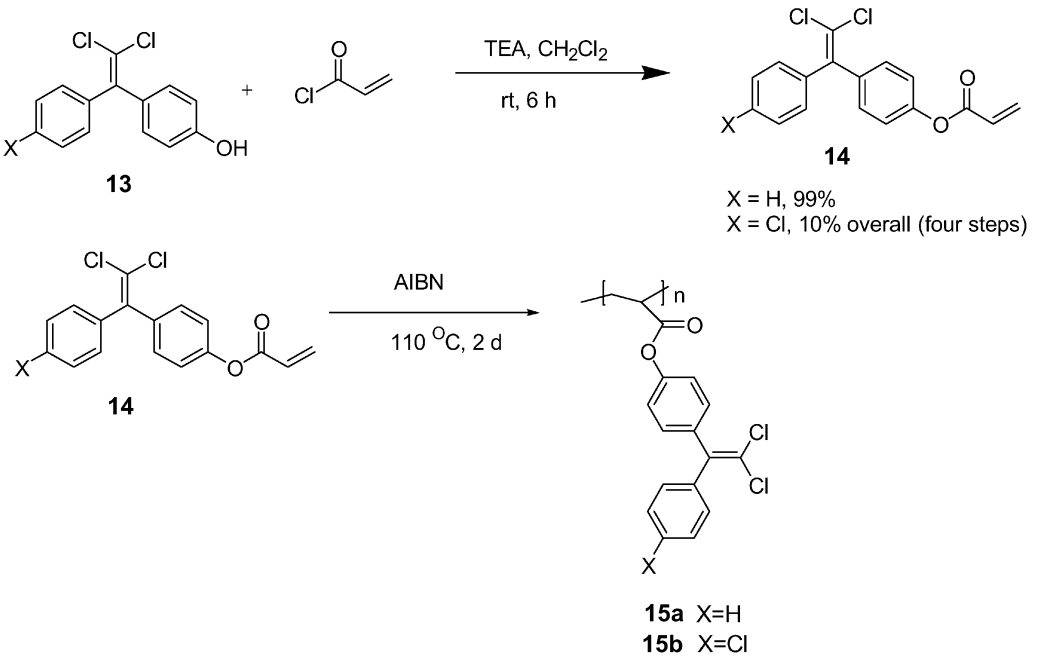


Fig. 17 Synthesis of asymmetric acrylate polymers **15a** and **15b**.

flame retardant polymers. However, because of the excellent flame retardant characteristics of BPC, we have made new asymmetric BPC (ABPC) monomers with an easily polymerized acrylate or styrene functional group. Acrylate polymers are known for their resistance to heat, sunlight, and weathering, which makes them excellent candidates for flame retardant polymers.[90] Styrene polymers are one of the most widely used plastics in the world having applications in markets such as those for building materials, packaging,

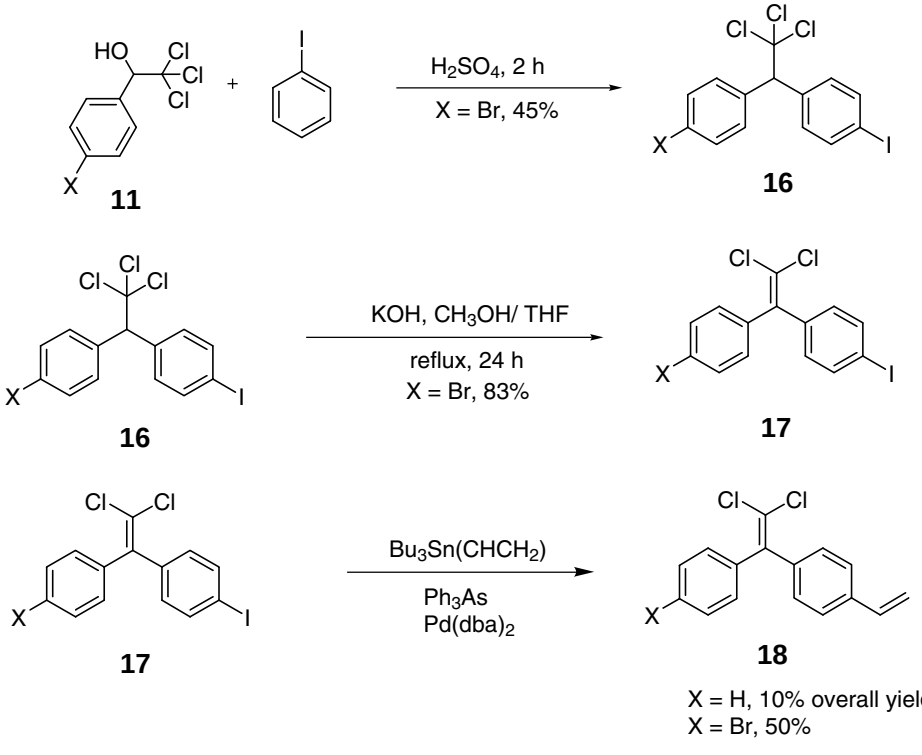


Fig. 18 Synthesis of asymmetric BPC styrene monomer **18**.

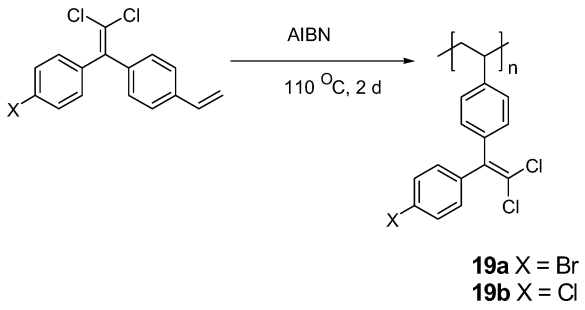


Fig. 19 Polymerization of BPC styrene monomers to give polymers **19a** and **19b**.

appliances, and automobiles, among others. By utilizing these two very important classes of polymers, it is possible to incorporate flame retardant properties into the most common plastics used today. Our approach to increasing the flame retardancy of vinyl polymers was to incorporate BPC as a pendant group of the vinyl backbone. As we will show, this work has produced a flame retardant polymer that can be easily melted and molded during processing. Such properties make the commercialization of the polymer more likely.

Acrylate BPC structures are made from inexpensive and readily available starting materials. The first step is to produce the trichloroethanol carbinol **11**. The carbinol is synthesized by slowly adding a solution of potassium hydroxide and methyldigol (diethylene glycol monomethylether) to a stirring solution of aromatic aldehyde (X can be H or Cl) and chloroform chilled at 0°C (Fig. 16).[91] The resulting carbinol **11** was purified by vacuum distillation and subsequently coupled with phenol in the presence of BF_3 gas over several days to yield **12**.[91] This process required several daily additions of BF_3 gas and mechanical stirring to mix the dark viscous product that was then purified by column chromatography. The dehydrohalogenation of **12** was completed using lithium chloride in DMF at 140°C for 24 hr, then slowly decreasing the heat to 100°C, to yield the asymmetric BPC analog **13** (Fig. 16).[92]

The final step to afford the acrylate monomer is the addition of acryloyl chloride to a solution of **13** and triethylamine in methylene chloride to produce the acrylate monomer **14** (Fig. 17). The acrylate polymer **15** was made by a bulk polymerization catalyzed by 2,2′-azobisisobutyronitrile (AIBN) at 110°C for 2 d (Fig. 17).

Styrene BPC polymers are made in much the same way; however, different attachment groups must be utilized to incorporate the styrene group onto the BPC molecule. Continuing on with the synthesis of **16**, sulfuric acid (H_2SO_4) was used to couple iodobenzene to make the asymmetric BPC analog as seen in Fig. 18.[93] The iodo-functionalized BPC compound was then dehydrohalogenated using potassium hydroxide and a methanol/THF mixture under reflux conditions to yield the BPC **17** (Fig. 18). The final step to afford the styrene BPC monomer (**18**) is to couple a vinyl group to the aryl iodide using Stille coupling procedures (Fig. 18).[94] The vinyl polymer **19** was polymerized under the same conditions as for **15** (Fig. 19).

The burn results for polymers **15** and **19** show that the polymers are inherently flame retardant with low base flammability (Table 4). The polymer dripped but did not ignite the cotton when it was subjected to the UL-94 flame test, and with the addition of 1 wt% PTFE, it did not drip. The PCFC results show that these polymers have a high heat release capacity when compared with the BPC carbonates and aryl ethers, but it is still significantly less than that of the base commodity polymers, such as polyethylene or polystyrene (Table 5).

CONCLUSIONS

In this entry, we have summarized recent work in the field of nonhalogenated flame retardant additives and

Table 4 UL-94 test for flammability—polyacrylates and polystyrenes

Compound tested	First ignition[a] (s)	Observed dripping[b]	Second ignition[a] (s)	Observed dripping[b]	UL-94 rating
Polyacrylate **15a**	0, 0	No, no	1, 1	No, yes[d]	V-0
Polyacrylate **15a**, 1 wt.% PTFE	0, 0	No, no	0, 0	No, no	V-0
Polyacrylate **15b**	0, 0	No, no	1, 1	No, no	V-0
Polystyrene **19a**	0[c]	No[c]	0[c]	No[c]	V-0[c]

[a]Time to self-extinguishing after the first or second 10 s ignition. The two numbers are for two separate tests on two separate samples.
[b]Molten polymer did (yes) or did not (no) drip onto cotton patch underneath ignited bar during UL-94 test.
[c]Preliminary tests were performed on smaller samples than dictated by UL-94 standards.
[d]The dripping did not ignite the cotton.

Table 5 PCFC Results for the polyacrylates and polystyrenes

Polymer	Heat capacity (J/g K)	Total heat (kJ/g)	CHAR yields (%)
Polyacrylate **15a**	185	12	17.1
Polyacrylate **15b**	224	10	11
Polystyrene **19a**	263	8	26
Polystyrene **19b**	320	16	20

inherently nonflammable polymers. Currently, none of the additives we developed are in commercial use because of cost issues associated with their production. However, process chemistry and economy of scale would greatly lower the cost for these additives if new regulations pushing to higher levels of fire safety, or nonhalogenated flame retardants, mandated their use. The inherently nonflammable polymers could be commercialized if there were sufficient market incentive. Because the BPC materials have very low flammability, their use may be mandated in high-performance applications, such as aircraft interiors and military/aerospace applications. The synthetic routes into these materials are not difficult, which indicates that the potential cost for these materials should be more than that of the existing commodity plastics, but not prohibitively expensive. The only potential concern with the synthetic approaches into BPC would be the use of chloral, which is an FDA regulated substance, and requires special licenses for its large scale commercial use.

To summarize the key findings of our work:

1. Alkyne, phosphorus/alkyne, and boronic acid flame retardant additives can be synthesized from brominated feedstocks via catalysis reactions.
   a. These additives are effective in providing flame retardancy in polycarbonate and require an additional synergist (currently unknown) to provide flame retardancy in other polymers.
2. BPC is a very versatile molecule that can be used to produce a variety of polymeric materials with inherently low levels of flammability, allowing them to be used with no additional flame retardant in some applications.
   a. Polycarbonates via transesterification with a tin catalyst.
   b. Polyarylethers via condensation or ADMET polymerizations.
   c. Vinyl polymers (acrylate or styrenic-like) via free-radical polymerizations.
3. The increasing market demand for higher fire safety and lower environmental impact will likely stimulate the commercial development of new flame retardant systems. These new systems may be additive or inherently polymer based, as described in this entry, and it may be that these materials are early analogs of future systems that someday will be in commercial use.

ACKNOWLEDGMENT

We thank the FAA, grant number 02-G-023, for funding this research.

REFERENCES

1. Karter, M.J. *Fire Loss in the US, 2002*; National Fire Prevention Association: Quincy, MA, 2003.
2. Davenport, R.E.; Fink, U.; Sasano, T. Flame Retardants SRI International Report, November 2002.
3. Simonson, M.; Blomqvist, P.; Boldizar, A.; Moller, K.; Rossell, L.; Tullin, C.; Stripple, H.; Sundqvist, J.O. Fire-LCA model: TV case study. SP Report 2000, 13.
4. http://www.lga.de/en/news/news_chemical_analysis_040323.shtml (accessed October 2004).
5. http://wwwdb.europarl.eu.int/oeil/oeil_viewdnl.ProcedureView?lang=2&procid=1632 (accessed October 2004).
6. http://www.epa.gov/waterscience/fish/forum/2004/presentations/wednesday/mcdonald.pdf (accessed October 2004).
7. http://www.inchem.org/documents/ehc/ehc/ehc209.htm (accessed October 2004).
8. http://www.mindfully.org/Plastic/Computer-Triphenyl-Phosphate.htm (accessed October 2004).
9. Camino, G.; Maffezzoli, A.; Braglia, M.; De Lazzaro, M.; Zammarano, M. Effect of hydroxides and hydroxycarbonate structure on fire retardant effectiveness and mechanical properties in ethylene-vinyl acetate copolymer. Polym. Degrad. Stab. **2001**, *74*, 457–464.
10. Wang, Z.; Hu, K.; Gui, Z. Thermal degradation of flame-retarded polyethylene/magnesium hydroxide/poly(ethylene-co-propylene) elastomer composites. Polym. Int. **2003**, *52*, 1016–1020.
11. Grand, A.F.; Wilkie, C.A. Inorganic hydroxides and hydroxycarbonates: their Function and use as flame-retardant additives. In *Fire Retardancy of Polymeric Materials*; Marcel Dekker, Inc: New York, 2000; 285–352.
12. Schartel, B.; Braun, U.; Schwarz, U.; Reinemann, S. Fire retardancy of polypropylene/flax blends. Polymer. **2003**, *44*, 6241–2350.

13. Le Bras, M.; Bourbigot, S.; Delobel, R.; Camino, G.; Eling, B.; Lindsay, C.; Roels, T. Thermal degradation of polyurethane and polyurethane/expandable graphite coatings. Polym. Degrad. Stab. **2001**, *74*, 493–499.
14. Duquesne, S.; Delobel, R.; Le Bras, M.; Camino, G. A comparative study of the mechanism of action of ammonium polyphosphate and expandable graphite in polyurethane. Polym. Degrad. Stab. **2002**, *77*, 333–344.
15. Modesti, M.; Lorenzetti, A. Flame retardancy of polyisocyanurate–polyurethane foams: use of different charring agents. Polym. Degrad. Stab. **2002**, *78*, 341–347.
16. Weil, E.D.; McSwigan, B. Melamine phosphate flame retardants. Plastics Compound. **1994**, *17*, 31–39.
17. Gijsman, P.; Steenbakkers, R.; Furst, C.; Kersjes, J. Differences in the flame retardant mechanism of melamine cyanurate in polyamide 6 and polyamide 66. Polym. Degrad. Stab. **2002**, *78*, 219–224.
18. Ray, S.S.; Okamoto, M. Polymer/layered silicate nanocomposites: a review from preparation to processing. Prog. Polym. Sci. **2003**, *28*, 1539.
19. Gilman, J.W.; Jackson, C.L.; Morgan, A.B.; Harris, R.; Manias, E.; Giannelis, E.P.; Wuthenow, M.; Hilton, D.; Phillips, S.H. Flammability properties of polymer-layered-silicate nanocomposites. Polypropylene and polystyrene nanocomposites. Chem. Mater. **2000**, *12*, 1866–1873.
20. Chigwada, G.; Wilkie, C.A. Synergy between conventional phosphorus fire retardants and organically-modified clays can lead to fire retardancy of styrenics. Polym. Degrad. Stab. **2003**, *80*, 551–557.
21. Bartholmai, M.; Schartel, B. Layered silicate polymer nanocomposites: new approach or illusion for fire retardancy? Investigations of the potentials and the tasks using a model system. Polym. Adv. Technol. **2004**, *15*, 355–364.
22. Morgan, A.B.; Chu, L.L.; Harris, J.D. A flammability performance comparison between synthetic and natural clays in polystyrene nanocomposites. Proceedings of Flame Retardants, 27–28, January 2004; Interscience Communications: London, U.K., 2004; 85–96.
23. Zanetti, M.; Camino, G.; Canavese, D.; Morgan, A.B.; Lamelas, F.J.; Wilkie, C.A. Fire retardant halogen-antimony-clay synergism in polypropylene layered silicate nanocomposites. Chem. Mater. **2002**, *14*, 189–193.
24. Dabrowski, F.; Le Bras, M.; Cartier, L.; Bourbigot, S. The use of clay in an EVA-based intumescent formulation. Comparison with the intumescent formulation using polyamide-6 clay nanocomposite as carbonisation agent. J. Fire Sci. **2001**, *19*, 219–241.
25. Beyer, G. Flame retardant properties of EVA-nanocomposites and improvements by combination of nanofillers with aluminum trihydrate. Fire Mater. **2001**, *25*, 193–197.
26. Gilman, J.W.; Kashiwagi, T. Polymer-layered silicate nanocomposites with conventional flame retardants. In *Polymer-Clay Nanocomposites*; Pinnavaia, T.J., Beall, G.W., Eds.; John Wiley & Sons, 2000; 193–206.
27. Stevens, M.P. *Polymer Chemistry: An Introduction*; 3rd Ed.; Oxford University Press: New York, 1999; 110–112.
28. Nelson, G.L. Fire and polymers. Chemistry **1978**, *51*, 22
29. Tewarson, A. Generation of heat and chemical compounds in fires. In *SFPE Handbook of Fire Protection Engineering*, 2nd Ed.; Society of Fire Protection Engineering: Boston, 1995; 53–124.
30. Jeff Gilman, NIST-BFRL, U.S. Dept. of Commerce, Gaithersburg, MD, 20899.
31. Babrauskas, V.; Peacock, R.D. Heat release rate: the single most important variable in fire hazard. Fire Safety J. **1992**, *18*, 255–272.
32. Morgan, A.B.; Tour, J.M. Synthesis and testing of nonhalogenated alkyne-containing flame-retarding polymer additives. Macromolecules. **1998**, *31*, 2857–2865.
33. Morgan, A.B.; Tour, J.M. Synthesis and testing of nonhalogenated alkyne/phosphorus-containing polymer additives: potent condensed-phase flame retardants. J. Appl. Polym. Sci. **1999**, *73*, 707–718.
34. Morgan, A.B.; Tour, J.M. Synthesis, flame-retardancy testing, and preliminary mechanism studies of nonhalogenated aromatic boronic acids: a new class of condensed-phase polymer flame-retardant additives for acrylonitrile–butadiene–styrene and polycarbonate. J. Appl. Polym. Sci. **2000**, *76*, 1257–1268.
35. Sonogashira, K.; Tohda, Y.; Hagihara, N. A convenient synthesis of acetylenes: catalytic substitutions of acetylenic hydrogen with bromoalkenes, iodoarenes and bromopyridines. Tetrahedron Lett. **1975**, 259.
36. Sonogashira, K. Development of Pd–Cu catalyzed cross-coupling of terminal acetylenes with sp^2-carbon halides. J. Organomet. Chem. **2002**, *653*, 46–49.
37. Georlette, P.; Simons, J.; Costa, L. Halogen-containing fire-retardant compounds. In *Fire Retardancy of Polymeric Materials*; Grand, A.F., Wilkie, C.A., Eds.; Marcel Dekker: New York, 2000.

38. Littke, A.F.; Fu, G.C. Palladium-catalyzed coupling reactions of aryl chlorides. Angew. Chem., Int. Ed. **2002**, *41*, 4176.
39. Wolfe, J.P.; Tomori, J.; Sadighi, J.P.; Yin, J.; Buchwald, S.L. Simple, efficient catalyst system for the palladium-catalyzed amination of aryl chlorides, bromides, and triflates. J. Org. Chem. **2000**, *65*, 1158.
40. Littke, A.F.; Dai, C.; Fu, G.C. Versatile catalysts for the Suzuki cross-coupling of arylboronic acids with aryl and vinyl halides and triflates under mild conditions. J. Am. Chem. Soc. **2000**, *122*, 4020.
41. Old, D.W.; Wolfe, J.P.; Buchwald, S.L. Highly active catalyst for palladium-catalyzed cross-coupling reactions: room-temperature Suzuki couplings and amination of unactivated aryl chlorides. J. Am. Chem. Soc. **1998**, *120*, 9722.
42. Hartwig, J.F.; Kawatsura, M.; Hauck, S.I.; Shaughnessy, K.H.; Alcazar-Roman, L.M. Room-temperature palladium-catalyzed amination of aryl bromides and chlorides and extended scope of aromatic C–N bond formation with a commercial ligand. J. Org. Chem. **1999**, *64*, 5575.
43. Gelman, D.; Buchwald, S. Efficient palladium-catalyzed coupling of aryl chlorides and tosylates with terminal alkynes: use of a copper cocatalyst inhibits the reaction. Angew. Chem., Int. Ed. **2003**, *42*, 5993–5996.
44. Bergman, R.G. Reactive 1,4-dehydroaromatics. Acc. Chem. Res. **1973**, *6*, 25.
45. Bergman, R.G.; Lockhart, T.P.; Comita, P.B. Kinetic evidence for the formation of discrete 1,4-dehydrobenzene intermediates. Trapping by inter- and intramolecular hydrogen atom transfer and observation of high-temperature CIDNP. J. Am. Chem. Soc. **1981**, *103*, 4082.
46. Bergman, R.G.; Lockhart, T.P. Evidence for the reactive spin state of 1,4-dehydrobenzenes. J. Am. Chem. Soc. **1981**, *103*, 4091.
47. Bergman, R.G.; Bharucha, K.N.; Marsh, R.M.; Minto, R.E. Double cycloaromatization of (Z,Z)-deca-3,7-diene-1,5,9-triyne: evidence for the intermediacy and diradical character of 2,6-didehydronaphthalene. J. Am. Chem. Soc. **1992**, *114*, 3120.
48. Sastri, S.B.; Keller, T.M.; Jones, K.M.; Armistead, J.P. Studies on cure chemistry of new acetylenic resins. Macromolecules. **1993**, *26*, 6171.
49. Nielsen, D.R.; McEwen, W.E. Benzenediboronic acids. J. Am. Chem. Soc. **1957**, *79*, 3081.
50. Hoffman, A.K.; Thomas, W.M. The synthesis of p-vinylphenylboronic acid and some of its derivatives. J. Am. Chem. Soc. **1959**, *81*, 580.
51. Coutts, I.C.G.; Goldschmid, H.R.; Musgrave, O.C. Organoboron compounds Part VII. Aliphatic and aromatic diboronic acids. J. Chem. Soc. C **1970**, 488.
52. Todd, M.H.; Balasubramanian, S.; Abell, C. Studies on the synthesis, characterisation and reactivity of aromatic diboronic acids. Tetrahedron Lett. **1997**, *38*, 6781.
53. Ishiyama, T.; Murata, M.; Miyaura, N. Palladium(0)-catalyzed cross-coupling reaction of alkoxydiboron with haloarenes: a direct procedure for arylboronic esters. J. Org. Chem. **1995**, *60*, 7508.
54. Ishiyama, T.; Miyaura, N. Synthesis of aryl boronates via palladium-catalyzed cross-coupling reactions of alkoxydiboron with aryl halides or triflates. In *Advances in Boron Chemistry*; Sieber, W., Ed.; The Royal Society of Chemistry: Cambridge, U.K., 1997; 92.
55. Murata, M.; Watanabe, S.; Masuda, Y. Novel palladium(0)-catalyzed coupling reaction of dialkoxyborane with aryl halides: convenient synthetic route to arylboronates. J. Org. Chem. **1997**, *62*, 6458.
56. Murata, M.; Oyama, T.; Watanabe, S.; Masuda, Y. Palladium-catalyzed borylation of aryl halides or triflates with dialkoxyborane: a novel and facile synthetic route to arylboronates. J. Org. Chem. **2000**, *65*, 164.
57. Cho, J.Y.; Iverson, C.N.; Smith, M.R. Steric and chelate directing effects in aromatic borylation. J. Am. Chem. Soc. **2000**, *122*, 12868.
58. Morgan, A.B.; Tour, J.M. Synthesis and Use of Non-Halogenated Aromatic Compounds as Flame Retardants for Polymer-Containing Materials. U.S. Patent 6,566,429 B2, 20 May, 2003.
59. Hirschler, M.M. Chemical aspects of thermal decomposition of polymeric materials. In *Fire Retardancy of Polymeric Materials*; Grand, A.F., Wilkie, C.A., Eds.; Marcel Dekker: New York, 2000.
60. Lyons, J.W. *The Chemistry and Uses of Fire Retardants*; John Wiley and Sons Inc.: New York, 1970.
61. Sobiczewski, Z.; Wielgosz, Z. Stability of polycarbonates prepared from chlorinated bisphenols. Plaste Kautschuk **1968**, *15*, 176–179.
62. Sobiczewski, Z.; Wielgosz, Z.; Janicka, K. Stability and stabilization of polycarbonates prepared from chlorobisphenols. Plaste Kautschuk **1969**, *16*, 99–102.
63. Stewart, J.R. Synthesis and characterization of chlorinated bisphenol-based polymers and polycarbodiimides as inherently fire-safe polymers. Ph.D. Thesis, University of Massachusetts: Amherst, MA, 2000.
64. Factor, A.; Orlando, C.M. Polycarbonates from 1,1-dichloro-2,2-bis(4-hydroxyphenyl)ethylene and

bisphenol A: a highly flame-resistant family of engineering thermoplastics. J. Polym. Sci. Polym. Chem. Ed. **1980**, *18*, 579–592.
65. Rusanov, A.L.; Korshak, V.V. Heat- and fire-resistant polymers based on chloral and its derivatives. Uspekhi Khimii **1989**, *58*, 1006–1031.
66. Rusanov, A.L. Condensation polymers based on chloral and its derivatives. Prog. Polym. Sci. **1994**, *19*, 589–662.
67. Jurs, J.L.; Mickelson, E.T.; Abramowitz, D.B.; Tour, J.M. Novel flame retardant polymer blends. Proceedings of the Society for the Advancement of Material and Process Engineering, Long Beach, CA, 2000.
68. Jurs, J.L. Novel flame retardant polymers blends. In International Aircraft Fire and Cabin Safety Research Conference. Atlantic City, NJ, 2001.
69. Jurs, J.L.; Tour, J.M. Novel flame retardant polyarylethers: synthesis and testing. Polymer **2003**, 44.
70. Lyon, R. Fire resistant polymers based on bisphenol-C. International Aircraft Fire and Cabin Safety Research. Atlantic City, NJ, 2001.
71. Zhang, H.; Westmoreland, P.R.; Farris, R.J.; Coughlin, E.B.; Plichata, A. Thermal decomposition and flammability of fire-resistant, UV/visible-sensitive polyarylates, copolymers and blends. Polymer **2002**, *43*, 5463–5472.
72. Stoliarov, S.I.; Westmoreland, P.R. Mechanism of thermal decomposition of bisphenol-C polycarbonate: nature of its fire resistance. International Aircraft Fire & Cabin Safety Research Conference. Atlantic City, NJ, 2001.
73. Riley, D.J.; Gungor, A.; Srinivasan, A.; Sankarapandian, M.; Tchatchoua, C.; Muggli, M.W.; Ward, T.C.; McGrath, J.E. Synthesis and characterization of flame resistant poly(arylene ether)s. Polym. Eng. Sci. **1997**, *37*, 1501–1511.
74. Boileau, S. *New Methods for Polymer Synthesis*; Plenum Press: New York, 1992.
75. Wagener, K.B.; Boncella, J.M.; Nel, J.G. Acyclic diene metathesis (ADMET) polymerization. Macromolecules **1991**, *24*, 2649–2657.
76. Wagener, K.B.; Brzezinska, K. Acyclic diene metathesis (ADMET) polymerization: Synthesis of unsaturated polyethers. Macromolecules **1991**, *24*, 5273–5277.
77. Wagener, K.B.; Lehman, S.E. Comparison of the kinetics of acyclic diene metathesis promoted by Grubbs ruthenium olefin metathesis catalysts. Macromolecules **2002**, *35*, 48–53.
78. Stephenson, J.J.; Jurs, J.L.; Tour, J.M. Vinyl bisphenol C for flame retardant polymers. Proceedings of the Society for the Advancement of Material and Process Engineering, Long Beach, 2004; Vol. 49.
79. Almeras, X.; Le Bras, M.; Hornsby, P.; Bourbigot, S.; Marosi, Gy.; Anna, P.; Delobel, R. Artificial weathering and recycling effect on intumescent polypropylene-based blends. J. Fire Sci. **2004**, *22*, 143–161.
80. Hubacher, M.H. Bis(p-hydroxyphenyl)acetic acid. J. Org. Chem. **1959**, *24*, 1949–1951.
81. Hubacher, M.H.; Doernberg, S.; Horner, A. Laxatives: chemical structure and potency of phthaleins and hydroxyanthraquinones. J. Am. Pharm. Assoc. **1953**, *42*, 23.
82. Mark, V.; Hedges, C.V. Polycarbonate compositions having improved flame retardance and improved water vapor transmission U.S. Patent 4,182,838, 8 January, 1980.
83. Haba, O.; Ueda, M.; Kuze, S. Synthesis of polycarbonate from dimethyl carbonate and bisphenol-A through a non-phosgene process. J. Polym. Sci. A, Polym. Chem. **1999**, *37*, 2087–2093.
84. Lyon, R.E.; Walters, R.N. Microscale combustion calorimeter U.S Patent 5,981,290, 9 November, 1999.
85. Lyon, R.E.; Walters, R.N.; Gandhi, S. A pyrolysis-combustion flow calorimeter study of polymer heat release rate. Recent Adv. Flame Retard. Polym. Mater. **1998**, *9*, 334–353.
86. Walters, R.N.; Lyon, R.E. Microscale combustion calorimeter for determining flammability. Recent Adv. Flame Retard. Polym. Mater. **1997**, *8*, 298–308.
87. Walters, R.N.; Lyon, R.E. Molar group contributions to polymer flammability. J. Appl. Polym. Sci. **2002**, *87*, 548–563.
88. Wagener, K.B.; Brzezinska, K.; Anderson, J.D.; Younkin, T.R.; Steppe, K.; DeBoer, W. Kinetics of Acyclic Diene Metathesis (ADMET) Polymerization Influence of the Negative Neighboring Group Effect. Macromolecules **1997**, *30*, 7363–7369.
89. Lenz, R.W. *Organic Chemistry of Synthetic High Polymers*; Interscience Publishers: New York, 1967.
90. Horn, M.B. *Acrylic Resins*, 2nd Ed; Reinhold Publishing Corporation: New York, 1962.
91. Goodford, P.J.; Hudson, A.T.; Sheppey, G.C.; Wootton, R. Physicochemical-activity relationships in asymmetrical analogues of methoxychlor. J. Med. Chem. **1976**, *19*, 1239–1247.
92. Sundstrom, G. Metabolic hydroxylation of the aromatic rings of 1,1-dichloro-2,2-bis(4-chlorophenyl)ethylene (*p*,*p*′-DDE) by the rat. J. Agric. Food Chem. **1977**, *25*, 18–21.
93. Cristol, S. Some positional isomers of DDT analogs. J. Am. Chem. Soc. **1949**, *71*, 2875–2876.
94. Stille, J. Palladium-catalyzed coupling of vinyl triflates with organostannanes A short synthesis of pleraplysillin-1. J. Am. Chem. Soc. **1984**, *106*, 4630–4632.

Nickel–Cadmium Battery

Chung-Chiun Liu
Electronics Design Center and Department of Chemical Engineering, Case Western Reserve University, Cleveland, Ohio, U.S.A.

Xuekun Xing
NTK Powerdex, Inc., Wixom, Michigan, U.S.A., and Department of Chemical Engineering, Case Western Reserve University, Cleveland, Ohio, U.S.A.

INTRODUCTION

Nickel–cadmium (Ni–Cd) battery is one of the rechargeable battery systems produced on a large scale industrially. It is also one of the most mature and the widest employed rechargeable power sources in various industrial and consumer markets. At its peak, the annual production worldwide of the small-scale Ni–Cd batteries (cylindrical and prismatic) approached about 1.6 billion cells and over U.S.$2 billion in 1994 (R. Brodd, private communication).[1] This is mainly attributed to the various advantages of the Ni–Cd battery, such as its high discharge rate and high power density, long cycle life and long storage life, wide operational temperature range, etc. Ni–Cd batteries have been widely used in the transportation industry (power sources for trains, aircraft, and others), in the aerospace industry (power sources for rockets, satellites, and others), in the consumer market (portable power for cordless phone, cordless tools, toys and hobbies, portable electronic devices, portable lighting, and others), for standby power (uninterrupted power sources or UPS, emergency lighting, electronic backup, etc.), and in various military applications.[2–9] However, Ni–Cd batteries have some drawbacks: they are environmentally unfriendly (owing to the toxicity of cadmium), and have a relatively high self-discharge, memory effect in some cases, and relatively low energy density. Because of these disadvantages, the Ni–Cd battery has undergone a continuous decrease in its annual production volume in recent years, and in applications in selected fields it has been replaced by other more advanced battery systems such as nickel–metal hydride and Li-ion batteries. However, despite these drawbacks, the Ni–Cd battery remains a popular choice for various applications, and it still has a large production volume (about 1.16 billion cylindrical and prismatic cells and about U.S.$ 1.1 billion in 2001 and still holds a unique position in the field of rechargeable batteries (R. Brodd, private communication).[1]

The developmental history of the Ni–Cd battery can be traced back to the late 19th and early 20th centuries. The initial work was conducted by Jungner and others, who brought the ''alkaline electrolytes'' concept into the rechargeable battery field and introduced the ''pocket-plate'' Ni–Cd battery and its modifications.[2–8] The second milestone in the history of the Ni–Cd battery was the development of ''sintered-plate'' batteries, which was initially proposed by Pflider in 1928.[8] The application of sintered-plate electrodes resulted in a significantly improved battery energy density and a better rate capability. The disadvantages of the sintered-plate electrodes were the relatively higher fabrication cost and a possible memory effect during battery storage. A series of developments in the electrode structure of the Ni–Cd battery followed, to further improve the battery performance, which included the ''fiber-plate'' electrodes in the late 1970s and early 1980s, the ''pressed-plate'' or ''pasted'' electrodes, and more recently, the ''foam nickel'' electrodes, etc.[4–8] On the other hand, the cell structure of the Ni–Cd batteries underwent a development from the original vented cell structure to the sealed cell structures. Both structures of Ni–Cd batteries are in current production depending on the nature of the applications, and they represent two different concepts to overcome the problem of cell internal pressure buildup due to gas formation during the battery charge–discharge cycle. With the vented cell structure, the gas formed in the cell is removed via a ''physical-release'' mechanism by a specially designed gas valve on the cell container, while in the case of the sealed cell structure, the gas formed is removed via a ''chemical-combination'' mechanism by a special chemical reaction and special cell design. The sealed cell is a maintenance-free system and it can improve the battery cycle and the storage life. Therefore, it is very attractive for many applications, especially in the consumer market and for portable devices.

CELL COMPONENTS AND CELL CHEMISTRY

Cathode/Anode/Cell Reactions

The Ni–Cd cell consists of four base elements, that is, cathode, anode, separator, and electrolyte. Nickel

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007878

hydroxide ($Ni(OH)_2$) is the active material of the cathode at its discharge state. Among its various forms, beta-$Ni(OH)_2$ is believed to be the actual form undergoing a charge transfer reaction and it can be reversibly converted to nickel oxyhydroxide (NiOOH) as shown in Eq. (1) below:

$$Ni(OH)_2 + OH^- \underset{\text{discharge}}{\overset{\text{charge}}{=\!=\!=\!=\!=\!=\!=\!=}} NiOOH + H_2O + e^- \quad (1)$$

with an equilibrium potential $E^\circ = 0.49\,V$ vs. standard hydrogen electrode (SHE).

It has been established that the above cathode reaction is actually a proton transfer (or proton insertion/deinsertion) reaction between $Ni(OH)_2$ and NiOOH lattices. Eq. (1) can thus be rewritten as:

$$Ni(OH)_2 \underset{\text{discharge}}{\overset{\text{charge}}{=\!=\!=\!=\!=\!=\!=\!=}} NiOOH + H^+ + e^- \quad (2)$$

$Ni(OH)_2$ and NiOOH coexist in a single phase of a homogenous solid-state solution, and their relative concentration ratio in the solid solution varies with the state of charge (SOC) of the cell. Therefore, the cathode potential varies with the battery SOC governed by the Nernst equation. The temperature coefficient of the cathode reaction is about $-0.5\,mV/^\circ C$.

For the anode, cadmium hydroxide ($Cd(OH)_2$) is the active material at its discharge state. $Cd(OH)_2$ is reversibly converted to metallic Cd during the charging process, as shown in Eq. (3) below:

$$Cd(OH)_2 + 2e^- \underset{\text{discharge}}{\overset{\text{charge}}{=\!=\!=\!=\!=\!=\!=\!=}} Cd + 2OH^- \quad (3)$$

with an equilibrium potential, $E^\circ = -0.81\,V$ vs. SHE.

It should be noted that Eq. (3) is only a simplified reaction scheme for the anodic reaction. Actual anodic discharge reaction may take place via a ''dissolution–precipitation'' mechanism with soluble complex ions such as $Cd(OH)_3^-$, $Cd(OH)_4^{2-}$ as intermediates, as shown below:

$$Cd + 3OH^- \xrightarrow{\text{dissociation}} Cd(OH)_3^- + 2e^- \quad (4)$$

$$Cd(OH)_3^- \xrightarrow{\text{precipitation}} Cd(OH)_2 + OH^- \quad (5)$$

Unlike the cathode reaction, the anode reaction is not a homogenous single-phase reaction, but rather a two-phase [Cd and $Cd(OH)_2$] charge transfer reaction. Therefore, the anode potential is independent of its SOC. The anode reaction mechanism plays a significant role in impacting the reduction in battery capacity during its charge/discharge cycling. The temperature coefficient of the anode reaction is $-1.01\,mV/^\circ C$.

The overall cell reaction is thus shown as:

$$2Ni(OH)_2 + Cd(OH)_2 \underset{\text{discharge}}{\overset{\text{charge}}{=\!=\!=\!=\!=\!=\!=\!=}} 2NiOOH + Cd + 2H_2O \quad (6)$$

with cell voltage at the standard state $E_{cell} = 1.30\,V$.

Reactions in Overcharge and Overdischarge

In overcharge, the following reactions may take place at the cathode:

$$4OH^- \rightarrow 2H_2O + O_2 + 4e^- \quad (7)$$

at the anode:

$$2H_2O + 2e^- \rightarrow H_2 + 2OH^- \quad (8)$$

$$O_2 + 2H_2O + 4e^- \rightarrow 4OH^- \quad (9)$$

In overdischarge, the following reaction may occur at the cathode:

$$2H_2O + 2e^- \rightarrow H_2 + 2OH^- \quad (10)$$

The charging efficiency in the Ni–Cd battery is generally only about 80%. Part of the electrical energy during charging is consumed by side reactions, such as the water decomposition reaction. Although cadmium electrode has a high overpotential to hydrogen evolution reaction [Eq. (8)], which can be further suppressed by using excess anode, the O_2 evaluation at the cathode during charging is generally inevitable. Because the charging efficiency is lower than 100%, the Ni–Cd cell needs to be overcharged to a certain degree to attain the desired discharge capacity of the given cell.

Additives and Their Functions

Various additives are used for the Ni–Cd batteries to improve the battery performance. Additives are selected based on their special functions to improve the electrode structure and/or electrode chemical and electrochemical properties. For example, cadmium hydroxide $Cd(OH)_2$ is added to the cathode to prevent phase segregation and to help maintain a single phase of the solid solution during the transfer between $Ni(OH)_2$ and NiOOH in charge and discharge processes. Because $Cd(OH)_2$ is isomorphous with both $Ni(OH)_2$ and NiOOH, this structural functionality can improve the cycle life of the battery. Cd or CdO can increase the overpotential of oxygen evolution

reaction at the cathode and thus can suppress the gas evaluation and reduce the self-discharge rate because oxygen evaluation at the cathode is one of the major factors that causes cell self-discharge. Similar functions have been found with a number of metals, such as Zn, Li, Co, etc., which can also be used as cathode additives. On the other hand, metallic cadmium grain growth and the formation of large-size crystals may cause loss of battery capacity during the charge–discharge cycle; some materials having antiagglomorant function are used as anode additives. These include selected iron compounds, nickel, and selected organic polymers (such as PVA, cellulose, etc.), which can prevent Cd crystal growth, stabilize Cd structures, and thus minimize the loss of capacity during the charge–discharge cycle. Indium is a different type of anode additive and its basic function is to enhance the electrochemical activity of the anode-active material, $Cd(OH)_2$. Other selected compounds can be used as electrolyte additives; for example, a small amount of lithium hydroxide, LiOH, can be added into the electrolyte to improve the battery cycle life at elevated temperatures.

Electrolyte

The typical electrolyte of the Ni–Cd battery is the concentrated potassium hydroxide (KOH) aqueous solution (~30% by weight). This electrolyte solution provides high ionic conductivity and a wide temperature window of a liquid phase. In addition, both the cathodic and the anodic active materials have low solubility in this electrolyte. These properties make a substantial contribution to the performance of the Ni–Cd, especially the high rate, wide operational temperature range, long cycle life, and long shelf life.

Separator

The main function of the separator in a cell is to prevent direct contact between the cathode and the anode and thus avoid an internal electrical short circuit of the cell. In addition, a separator needs to provide a good pathway for the ion transport in the cell. Therefore, in addition to its good chemical and electrochemical stability in the cell's environment, a good separator should have a high degree of porosity and a good wettability with regard to the electrolyte used. In the case of the vented cell type of Ni–Cd batteries, a separator also functions as a gas barrier to prevent gases formed in the cell from passing from the cathode area to the anode area, or vice versa. The commonly employed separators in Ni–Cd batteries are nonwoven nylon and polyimide. A newly developed separator in recent years is polymer impregnated Zircar material, which shows an extremely high chemical stability in concentrated caustic solutions, and can thus further reduce the self-discharge rate and improve battery life. Zircar is a high-temperature fibrous ceramic material containing zirconia as a major component.

BATTERY CONSTRUCTION

Ni–Cd cells can be categorized into two major groups of batteries, vented and sealed, based on the cell construction.

Vented Ni–Cd Batteries

As mentioned, gas (oxygen) evaluation at the cathode takes place in the overcharge process, which results in a continuous accumulation of gases and a buildup of the internal pressure of the cell during its charge–discharge cycle. To resolve this problem, the vented cell structural concept was introduced into the Ni–Cd battery in its original form during its early development stage, and it has been widely employed in manufacturing Ni–Cd batteries since then. For the vented cell construction, the cell container is fabricated with an open space that is covered by a removable cap (vent cap). The main function of the cap is twofold: first, it serves as a check valve to release gases generated during cell overcharge (such as from water decomposition) and prevent the increase of the cell internal pressure and second, it serves as an inlet for replenishment of water to the electrolyte when water is consumed during cell usage.

For the vented Ni–Cd batteries, two types of electrode structures and configurations are used. One is the so-called pocket-plate electrode and the other is the sintered-plate electrode. In the case of the pocket plate, electrodes are made by pressing or feeding electrode composite materials (cathode: nickel hydroxide mixed with conductive materials, such as graphite, binder, and other selected additives; anode: cadmium hydroxide mixed with graphite, binder, and other selected additives) into flat pockets of preshaped perforated steel strips. The steel strips are plated with nickel to avoid iron contamination in the system. The pocket functions as a current collector and also provides mechanical support. The thickness of the pocket-plate electrode can be controlled according to the requirement of the battery discharge rate, and it is generally in the range of 1.5–5.0 mm. The pocket-plate Ni–Cd batteries have a number of advantages, such as the simplicity of the electrode fabrication process, its low cost, the good physical stability and reliability, the long storage life, etc. One of its main disadvantages is its relatively low energy density, which results largely

from the loose structure of the electrodes and the inefficient volume utilization of the cell.

The sintered-plate electrode was developed to successfully change the electrode structure from loose to compact. By this technology, electrodes can be made denser and thus thinner compared with pocket-plate electrodes. Accordingly, the energy density of the sintered-plate Ni–Cd battery is significantly higher than its pocket-plate counterpart (up to 50% higher). The sintered-plate electrode consists of two main components, that is, a sintered porous metal substrate (usually called ''plaque'') and the active material dispersively impregnated into the pores of the plaque. The plaque is usually made with a perforated (or woven screens) nickel or nickel-plated steel strip (or wire) substrate. This substrate is then coated with carbonyl nickel powder plus binder and pore former via either the ''wet coating'' (slurry coating) or the ''dry powder'' process. The coated substrate is then sintered in a reducing atmosphere at 800–1000°C for a controlled time duration to become a compact, highly porous plaque. The plaque obtained by this process usually has porosity up to 85% and its thickness can be controlled based on the requirements. The plaque has the dual function of both host of the active materials and current collector for the given electrode. The active materials (nickel hydroxide for the cathode and cadmium hydroxide for the anode) are dispersed and held in the pores of the plaque. This is usually accomplished by using an impregnation process. There are several methods of impregnation, including mechanical, chemical, and electrochemical. Chemical impregnation is the most typical and widely employed method in the industry. In this process, the plaque is first filled with aqueous nickel nitrate $Ni(NO_3)_2$ solution [or cadmium nitrate solution $Cd(NO_3)_2$ for anode], and then the $Ni(NO_3)_2$-filled plaque is soaked and polarized cathodically in a caustic NaOH solution, in which the corresponding metal hydroxide is precipitated in the pores of the plaque. This impregnation process can be repeated several times until the desired loading level is approached. For the electrochemical impregnation process, the plaque is soaked in an acidic aqueous or alcoholic solution of nickel nitrate, and it is then cathodically polarized in this solution, and nickel hydroxide is then formed on the plaque.

Unlike the pocket-plate electrodes, the sintered-plate electrodes do not need the addition of any conductive material (such as graphite) to create enough electronic conductivity of the electrode, because plaque itself provides this function. On the other hand, the sintered-plate electrodes do not need the addition of the binder material to the active material to create the structural integrity of the electrode, because the plaque itself provides this integrity in the first place. Having these structural advantages, the sintered-plate Ni–Cd batteries show substantially higher material utilization and a higher energy density than the pocket-plate Ni–Cd batteries. The sintered-plate Ni–Cd batteries have other advantages such as high physical reliability, high performance rate, good storage life, etc. One of the disadvantages of the sintered-plate Ni–Cd battery is its relatively higher manufacturing cost, which is associated with costly processes such as sintering and impregnating. Another major disadvantage of the sintered-plate Ni–Cd battery is its memory effect.

The memory effect is generally referred to as an apparent decrease in the discharge capacity and power with respect to a given discharge cutoff voltage (VOC) after repeated shallow discharge–charge cycles. The apparent discharge capacity and power decrease results from depression or decrease of the discharge voltage; in other words, the discharge capacity could remain the same if the discharge VOC decreases accordingly. Therefore, the memory effect can be more correctly described as the depression of discharge voltage. The memory effect usually is not observed in the pocket-plate Ni–Cd batteries but it occurs in the sintered-plate Ni–Cd batteries. Although the mechanism of the memory effect is not fully clarified, it is believed to be associated with the change in the anode structure. The morphological changes of the unused metallic cadmium in a shallow discharge may lead to an increase in cell resistance, R, and thus an increase in IR drop during discharge. This is one of the proposed mechanisms for the voltage depression (IR mechanism). A different mechanism for the memory effect is the formation of intermetallic alloys such as Ni_5Cd_2 in the anode during the overcharge. Intermetallic alloys may have a higher discharge potential compared with metallic cadmium, which leads to a decreased cell discharge voltage. The main source of nickel in the anode to form Ni–Cd alloys is the residual nickel hydroxide that is formed from Ni-plaque during the impregnation process. Thus, it is important to control the impregnation process in order to reduce the formation of nickel hydroxide. Or, one can either completely remove Ni substrates or eliminate the impregnation process in the anode fabrication. The employment of the ''plastic-bonded'' or pressed-plate electrodes is one of the new developments following the above idea. It is noticed that the memory effect is actually a reversible effect, which can be recovered totally by an electrical reconditioning process called the ''maintenance cycle.'' In this process, the cell is forced to undergo a complete discharge (e.g., short circuiting the cell) to use up all unused active materials of the charged state, and this is then followed by a complete charge/overcharge step. The capacity and power of the cell can usually be restored by this treatment.

Vented sintered-plate Ni–Cd batteries are usually designed with capacity-matched cathode and anode.

Under this condition, the oxygen evaluation on the cathode and the hydrogen evaluation on the anode may occur during overcharge. To minimize and avoid oxygen transferring from the cathode area to the anode area and recombining with hydrogen, the separator used should be O_2 impermeable. In addition, vented Ni–Cd batteries usually use excess amount of the electrolyte to form the flood cell.

As discussed above, the electrodes of Ni–Cd cells are categorized in pocket-plate electrodes and sintered-plate electrodes, both of which are widely employed in the Ni–Cd battery industry. On the other hand, several new structures of the electrode have been developed in recent decades. One is the so-called fiber-plate electrode developed in the late 1970s, in which a mat of nickel fibers or nickel-plated plastic fibers is used instead of the pocket of perforated nickel-plated steel strips in the pocket-plate structure. Compared with the pocket-plate electrode, the fiber-plate electrode has higher utilization of the electrode-active materials and higher energy density. The next development in this field was the introduction of plastic-bonded or pressed-plate or pasted electrodes, in which the electrode composite (active material, conductive material, and plastic binder) is mixed with a solvent to become a paste first and then the paste is extruded or rolled on a current collector of nickel-plated perforated steel. The advantages of the plastic-bonded electrode are its cost-effective processing, higher structural integrity, and compactness of the electrode, which improves the energy density. The plastic-bonded electrode is widely used for the fabrication of the anode of Ni–Cd batteries. A further new development in the mid-1980s was the introduction of the foam-nickel electrodes. A porous Ni-foam substrate is fabricated first by plating the nickel on a porous plastic material (such as polyurethane, etc.), which is made electrically conductive by a pretreatment step consisting of an electrode-less deposition, and then removing the plastic material by a pyrolysis treatment after Ni plating. The Ni-foam framework obtained possesses a high degree of porosity (e.g., 90–95% free volume) and a high surface area, which creates a large capacity for holding active material and provides an effective conductive pathway. This results in enhanced cell energy density. The Ni-foam electrode structure has been widely employed for the cathode of Ni–Cd batteries in both industrial and consumer products.

Sealed Ni–Cd Batteries

The use of the vented cell structure can prevent the buildup of internal pressure in the cell resulting from gas generation during overcharge. One of the disadvantages of the vented structure, however, is that the batteries need to be routinely maintained, e.g., by replenishing water in the electrolyte to compensate for water consumed during overcharge. The sealed Ni–Cd battery, a new generation of the cell construction, was developed to further resolve the gas generation problem. The mechanism of the sealed Ni–Cd cell is based on the following chemical reaction:

$$\tfrac{1}{2}O_2 + Cd + H_2O \rightarrow Cd(OH)_2 \qquad (11)$$

where the oxygen, generated on the cathode during overcharge, transfers to the anode area and reacts with cadmium to produce cadmium hydroxide, which is continuously converted to cadmium during charge. Through this recombination route, the oxygen formed can be ''consumed'' or ''eliminated'' continuously, thus avoiding the buildup of internal pressure in the cell. The cell can be completely sealed and becomes maintenance free. The sealed Ni–Cd batteries have the same cell chemistry as that of the vented ones, but have the following differences in terms of cell design:

1. The cell is designed with excess anode-active materials, so that the overcharge takes place only on the cathode, and oxygen gas will be generated on the cathode, but no hydrogen gas is formed on the anode.
2. The separator employed in the sealed cells is not a gas barrier, but rather an oxygen permeable membrane (such as nonwoven nylon, polypropylene, etc.).
3. The amount of electrolyte is limited, that is, the cell is of ''starved'' type rather than the ''flooded'' type for vented batteries. The starved electrolyte can facilitate the oxygen transfer process.

In addition to being maintenance free, the sealed Ni–Cd batteries provide many other advantages, such as long shelf life, long cycle life, good performance rate, good low-temperature performance, etc. The sealed Ni–Cd batteries have been widely employed in various industrial and consumer markets. In the field of portable power sources, the sealed Ni–Cd batteries are especially attractive for various portable electronic devices such as wireless phones, camcorders, PCs, etc., and for various consumer products such as toys, wireless tools, photography, etc. The sealed Ni–Cd batteries can be manufactured with different constructions. The cylindrical cell is the most common one and Fig. 1 shows its general structure.[9]

In this structure, one of the core elements is a round jellyroll, which consists of a cathode strip, an anode strip, and a separator strip between them. The jellyroll is inserted into a cylindrical-shaped Ni-plated steel can with adequate electrical connections. The controlled

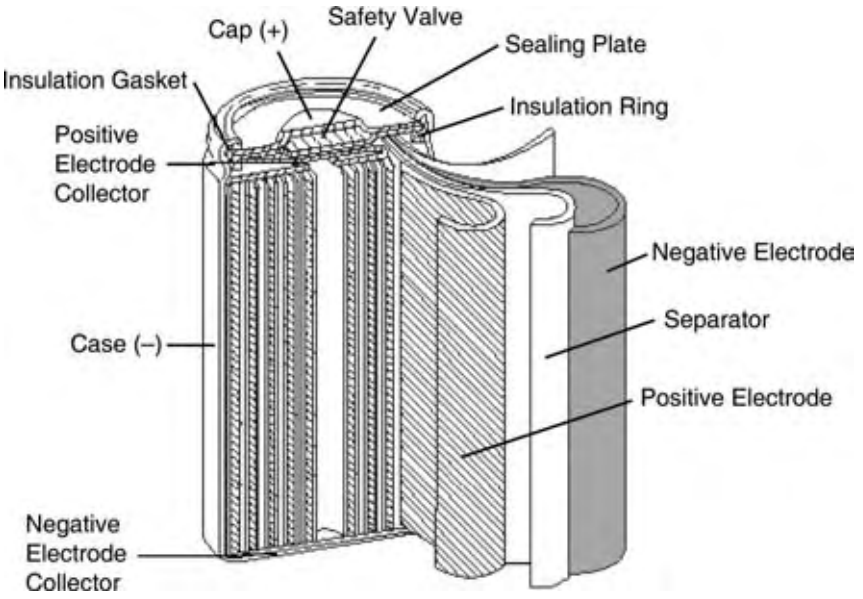


Fig. 1 Construction of the sealed Ni–Cd battery. (From Ref.[9].) *(View this art in color at www.dekker.com.)*

amount of electrolyte is then added into the can and the electrolyte is absorbed by the electrodes and the separator. The sealed Ni–Cd batteries can also be manufactured with other configurations, such as button cells, prismatic cells, and rectangular cells. Among them, button cells are usually used for applications of low capacity and low drain rate, while prismatic cells have more applications in portable and compact devices because of their structural advantage of enhancing volume efficiency by 20% with respect to the cylindrical structure.

OPERATION MODES

Charge Mode and Its Characteristics

The constant-current charge is the most commonly used charge mode for Ni–Cd batteries. Usually, a low or medium charge rate is recommended (e.g., a rate of a few tenths of coulombs rate) to avoid a fast increase in cell temperature during the charge. It is usually necessary to have a charge control circuit on the cell to control charge depth (ending charge), for which various parameters, such as charge time, cell temperature, and voltage change rate, can be selected as the charge controlling factors. Trickle and float charge are also used for Ni–Cd batteries in which the batteries are kept in a continuous charge step at a very low charge current (such as 0.02–0.05 C) to maintain them always in a fully charged state. Another charge mode is the constant voltage charge. In general, the charge efficiency in Ni–Cd batteries is significantly lower than 100% (typically about 70% for vented cells and about 80% for sealed cells) because of the side reactions occurring during the charge. Therefore, the overcharge is generally needed to obtain a full charge state. The charge efficiency is impacted by the charge rate and the charge temperature. For the sealed Ni–Cd batteries, the charge efficiency decreases with increasing charge temperature, which can be attributed to the decrease in potential for oxygen generation on the cathode. Although a higher charge rate leads to a higher charge efficiency, it is not recommended to use an overhigh charge rate because it could cause the buildup of internal pressure in the cell once the oxygen generation rate on the cathode is higher than the oxygen recombination rate on the anode. On the other hand, a high charge rate may cause an increase in cell temperature due to heat generated in the oxygen recombination reaction. Fig. 2 shows the typical charge characteristics of Ni–Cd batteries.

It is important to control the end point of the charge process for the constant-current charge mode. One method is to set up a predetermined VCO at which the charge process of constant current will be automatically terminated. Although this method is simple and effective, the selection of the VCO and the impact of temperature on the cell charge voltage must be considered. It has been found that the temperature effect of the charge voltage at constant-current conditions is $-4\,\mathrm{mV}/^{\circ}\mathrm{C}$. Owing to this effect, a VCO that is adequate at temperature T_1 may become too high if the cell temperature changes to T_2 ($T_2 > T_1$), which in turn may hurt the battery. To prevent this adverse effect, a temperature compensation element should be used. By using a temperature sensitive control unit to monitor cell temperature in situ, and with the function of the temperature compensation element installed in the battery case, charge process can be controlled

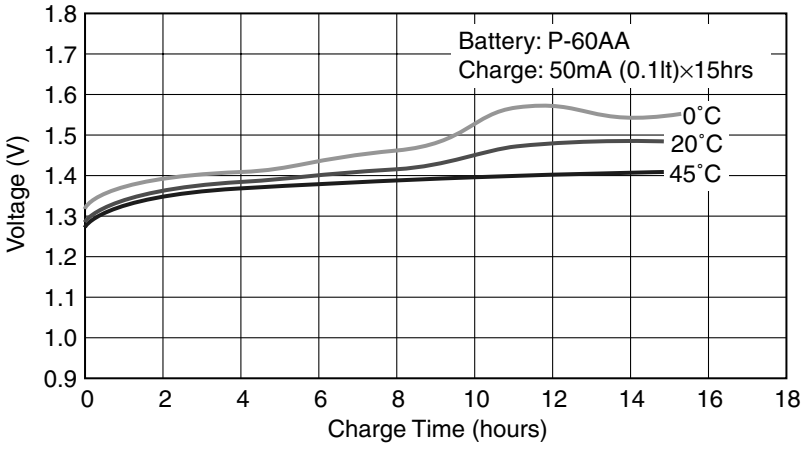


Fig. 2 The charge characteristics of Ni–Cd batteries. (From Ref.[9].) *(View this art in color at www.dekker.com.)*

adequately. Another commonly used charge control method is to use voltage change rate as a signal to terminate the charge, because the cell voltage undergoes a sharp rise when the charge approaches the end. This characteristic property can be used for controlling the charging process. It is noted that the ''end-of-charge'' voltage is higher for the vented Ni–Cd batteries than for the sealed Ni–Cd batteries; this is attributed to a higher overpotential for the gas evaluation reactions in the former, and a decreased overpotential for the latter due to depolarization of the anode once the oxygen recombination reaction takes place on it. Table 1 summarizes various charge systems of Ni–Cd batteries.[7]

Discharge Mode and Its Characteristics

The discharge mode of Ni–Cd batteries is dependent on their applications. The most common one is the constant-current discharge, in which the current can be changed from very low (e.g., in the case of button-batteries) to very large (e.g., for batteries for wireless tools). There are also other discharge modes, such as constant-resistance discharge, content-power discharge, etc.

BATTERY PERFORMANCE CHARACTERISTICS

Energy Density and Power Density

The energy density of Ni–Cd batteries is higher than for lead-acid batteries. Generally, the energy density of vented sintered-plate Ni–Cd cells ranges from 58 to 96 W hr/L and the specific energy ranges from 30 to 37 W hr/kg.[4] The value of the energy density varies with the cell structure and battery size. Ni–Cd batteries can be discharged at high rates and have a high power density. The typical power density of vented sintered-plate Ni–Cd cells ranges from 730 to 1250 W/L, and the specific power ranges from 330 to 460 W/kg.[4] In general, vented Ni–Cd batteries have higher power capability than the sealed ones.

Effect of Drain Rates on Discharge

Ni–Cd batteries have an open circuit voltage of about 1.35 V and an average close circuit voltage of about 1.20 V at normal discharge rates. The discharge capacity and the average voltage change depending on the discharge rate. Fig. 3A shows a family of typical discharge curves at various drain rates for sealed Ni–Cd batteries. As shown in the plot, the battery capacity decreases with increasing drain rates.

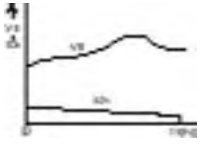
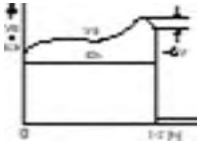
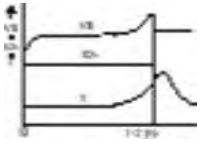
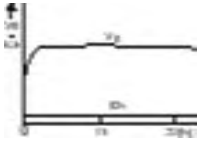
Effect of Temperature on Discharge

Ni–Cd batteries can be used at a wide range of temperatures (-40° to $+50^{\circ}$C). Fig. 3B shows a family of discharge curves at various temperatures for sealed Ni–Cd batteries. As shown in the plot, both capacity and average voltage decrease with decreasing temperatures. At –20°C, about 80% of the rated capacity can be obtained at 0.2 C rate.

Cycling Life

For Ni–Cd batteries, the cycling life is usually measured to the stage when the discharge capacity decreases to 80% of the rated capacity. Ni–Cd batteries generally have a cycling life of 500 cycles or more. The cycling life is even longer if the battery undergoes shallow discharge–charge cycling (low depth of discharge). The cycling life of the battery is also impacted by other factors such as charge mode and charge rate, discharge mode and discharge rate, battery operation temperature,

Table 1 General comparison of the various charge systems of Ni–Cd batteries

	Cycle (repeated) use				Standby use
Charge system	**Semiconstant current charge**	**Timer controlled charge**	**$-\Delta V$ cutoff charge**	**dT/dt cutoff charge**	**Trickle charge**
Operation VB: battery voltage Ich: charge current T: battery temperature					
Features	Most typical charge system Simple and economical	More reliable than semiconstant current charge system Relatively simple and economical	Most popular	Charging circuit costs more than the others but overcharge can be avoided enabling longer life cycle than $-\Delta V$ charge method	Simple and economical Applicable to the equipment for continuous long charge
No. of output terminals	2	2	2	3	2
Charge time (hr)	15	6–8	1–2	1–2	30 or longer
Charge current (C mA)	0.1	0.2	0.5–1	0.5–1	Frequent charge: 0.05–0.033
Trickle current (C mA)	—	0.05–0.033	0.05–0.033	0.05–0.033	Less frequent charge: 0.033–0.02
Charge level at charge control (%)	—	Approx. 120	Approx. 110–120	Approx. 100–110	—
Application examples	Shavers Digital cordless phones Toys	Cordless phones Shavers	Data terminals Camcorders Wireless equipment Cellular phones	Power tools Electric tools Notebook PCs Cellular phones	Emergency lights Guide lights Memory backup

(From Ref.[9].)

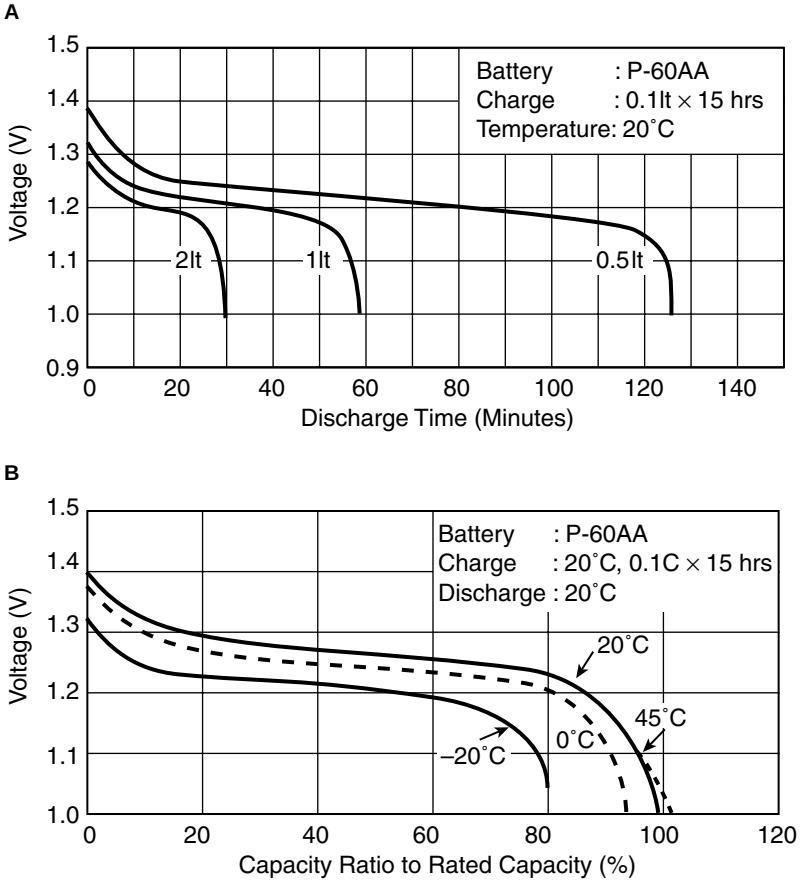


Fig. 3 Discharge characteristics of Ni–Cd batteries: (A) the effect of rates and (B) the effect of temperature. (From Ref.[9].)

and others. Fig. 4 shows a typical cycling performance for a sealed Ni–Cd cell.

Self-Discharge and Shelf Life

The shelf life of Ni–Cd batteries is largely determined by the cell self-discharge rate. The mechanism of the self-discharge for Ni–Cd batteries includes the following reactions:

1. The decomposition of the charged cathode-active material NiOOH:

$$2\mathrm{NiOOH} + \mathrm{H_2O} \rightarrow 2\mathrm{Ni(OH)_2} + \tfrac{1}{2}\mathrm{O_2} \quad (12)$$

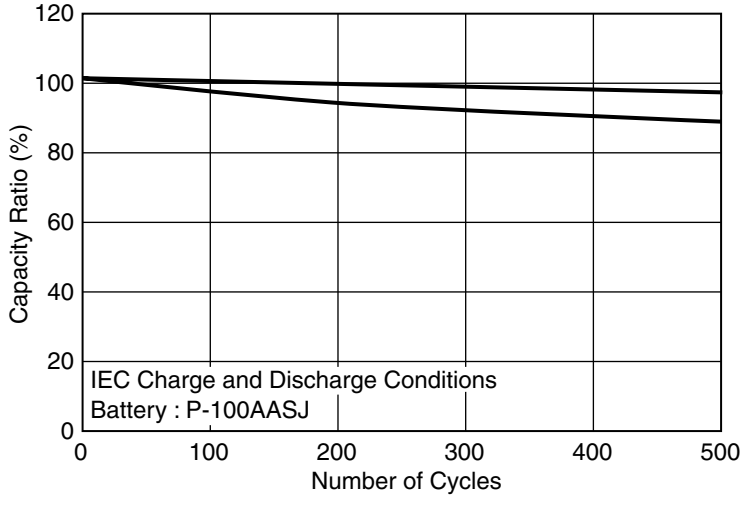


Fig. 4 Typical cycle life characteristics of Ni–Cd batteries. (From Ref.[9].)

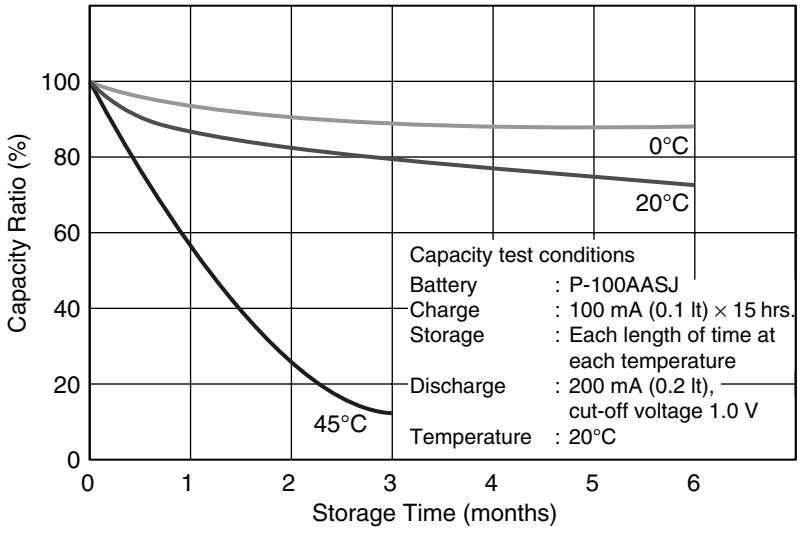


Fig. 5 Typical self-discharge characteristics of Ni–Cd batteries. (From Ref.[9].) *(View this art in color at www.dekker.com.)*

This decomposition reaction can be further enhanced by the O_2 recombination reaction at Cd anode, which in effect consumes part of the charged anode-active material (Cd).

2. Consumption of NiOOH and Cd (charged active materials) by residual inorganic impurities, mainly nitrate ions. The residual NO_3^- ions come from the electrode impregnation process; NO_3^- ions can react with Cd to form NO_2^- and NH_3. Both NO_2^- and NH_3 can, in turn, react with NiOOH. Therefore, both charged anode- and cathode-active compounds (Cd and NiOOH) can be gradually consumed by $NO_3^-/NO_2^-/NH_3$ shuttle reactions and this leads to capacity loss during the battery storage.
3. Consumption of NiOOH by organic impurities. Organic impurities are formed as a result of the decomposition of the separator materials in concentrated alkaline solutions. The organic impurities formed can react with NiOOH to cause capacity loss.

In general, the self-discharge rate increases with increasing temperature; accordingly, the shelf life of Ni–Cd batteries becomes shorter at high temperatures. Besides, several other factors such as electrode structure, process control during cell manufacturing, separator materials, etc. also affect the cell self-discharge rate and the shelf life. Fig. 5 shows the typical shelf life of sealed Ni–Cd cells at various temperatures.

CONCLUSIONS

The Ni–Cd battery is one of the important rechargeable batteries in both industrial and consumer markets. This battery system has special advantages in its high rate capability, high power density, and long cycle life, among others. Toxicity of cadmium is one of its major drawbacks, which has resulted in a gradual decrease in its production volume since the mid-1990s.

REFERENCES

1. Takeshita, H. Presentation in the 18th International Seminar & Exhibit on Primary & Secondary Batteries, March 2001.
2. Falk, S.U.; Salkind, A.J. *Alkaline Storage Batteries*; Electrochemical Society Series; John Wiley & Sons: New York, 1969.
3. Barak, M. *Electrochemical Power Sources*; Peter Peregrinus: London, 1980.
4. Linden, D. *Handbook of Batteries*, 2nd Ed.; McGraw Hill: New York, 1995.
5. Gates Energy Products. *Rechargeable Batteries Applications Handbook*; Butterworth-Heinemann: Boston, 1992.
6. Buchmann, I. *Batteries in a Portable World*, 2nd Ed.; Cadex Electronics Inc., 2001.
7. Vincent, C.A.; Scrosati, B. *Modern Batteries*, 2nd Ed.; John Wiley & Sons: New York, 1997.
8. Shukla, A.K.; Venugopalan, S.; Hariprakash, B.J. Nickel-based rechargeable batteries. Power Sources **2001**, *100,* 125 (and references therein).
9. Panasonic Industrial Company. *Battery Technical Handbook*; CD-ROM, 2002.

NMR in Chemical Processing

Sangrama K. Sahoo
Peter L. Rinaldi
Department of Chemistry, The University of Akron, Akron, Ohio, U.S.A.

INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy since its discovery in the 1940s has become one of the most valuable characterization tools for chemical and biological substances. Purcell et al.[1] and Bloch et al.[2] independently discovered the NMR phenomenon in 1946 and shared a Nobel Prize in 1952 for their work. In the first two decades following its discovery, NMR was used for the measurement of magnetic shielding, electric field gradients, and coupling tensors in molecular physics and for the chemical structure determination of small molecules. However, with the implementation of Fourier transform methods[3] and diffusion phenomena,[4] NMR has become a popular characterization tool in chemistry, physics, biochemistry, materials science, and medicine.[5] Another breakthrough in the 1970s used the deliberate introduction of field gradients along the sample to image nuclear spin density, a technique now commonly known as magnetic resonance imaging (MRI).[6] These developments led to many important applications of NMR as evident from the award of Nobel Prizes in Chemistry to Professor Richard Ernst in 1991 and Professor Kurt Wüthrich in 2002 for their work in the development of NMR methodology and their biological applications.

Nuclear magnetic resonance spectroscopy offers engineers the opportunity to study the influence of process parameters on heterogeneities in the physical and chemical structure of materials. It was first used in chemical engineering as a noninvasive probe of structure and flow.[7] One advantage of NMR is its utility to probe all three forms of matter: solid, liquid, and gas. Although solution NMR was first used in chemical processing, developments in solid-state NMR and imaging in the 1980s provide additional analytical tools. It is possible to obtain qualitative and quantitative information on adsorption, transport, and reaction processes in situ by NMR spectroscopic and imaging techniques.[7] These data are utilized in modeling studies to develop theories for predicting the behavior of relevant processes. The temperature dependence of the NMR spectrum makes it a unique tool for analyzing molecular dynamics behavior and reaction kinetics.

Conventional one-dimensional NMR involves the collection of a spectrum with signal intensity plotted as a function of resonance frequency (in ppm). However, recent developments in NMR led to the use of multidimensional (2D- and 3D-) NMR. These nD spectra are usually displayed as contour maps where each of the axes of the spectrum contains information relating to spectral parameters such as chemical shift (δ), scalar coupling (J), and diffusion coefficient (D); correlations between these spectral parameters provide additional structural and dynamics information.[8] These studies can be performed on solid powders, gels, liquid crystalline solutions, supercritical fluids, liquids, and solutions. Before discussing the application of NMR in chemical processing, basic principles of NMR spectroscopy are discussed, followed by applications categorized by studies in solution, solid state, and imaging.

BASIC PRINCIPLES OF NMR

Nuclear magnetic resonance spectroscopy is a technique to study the perturbation of atomic nuclei in a static magnetic field (B_0) in the presence of a second oscillating magnetic field (in the radiofrequency range).[8–11] All atomic nuclei possess a characteristic known as nuclear spin (I). However, only those nuclei with nonzero spin are NMR active, undergoing precession when placed in a static magnetic field. The unique precession frequency, known as the ''Larmor frequency,'' for each nucleus is dependent on B_0 and is unique for each atom. The irradiation frequency ω_0 applied to induce the resonance condition must match the magnetic moment's precession and is related to B_0 and magnetogyric ratio, γ, of the nucleus by

$$\omega_0 = \gamma B_0$$

Nuclei with $I = 1/2$ have two energy states, $m = \pm 1/2$, with the nuclear moment, μ, aligned with (lower energy) or against the field (higher energy) (Fig. 1). Generally, nuclei with spin I populate $2I + 1$ energy levels. Every element in the periodic table has at least one NMR-active isotope; some of the common NMR-active nuclei and their properties

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120018073

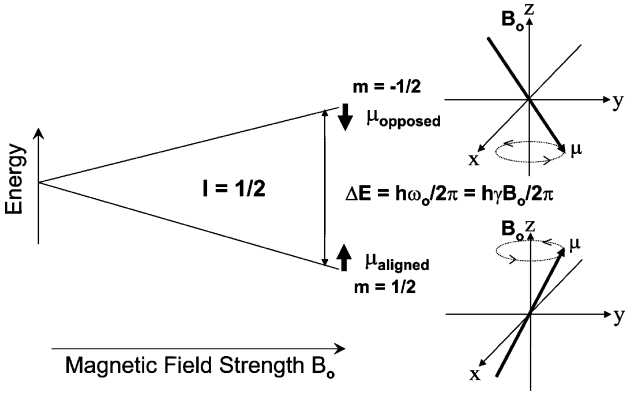


Fig. 1 Energy separation between nuclear spin states.

are given in Table 1. In a stationary (laboratory) reference frame at equilibrium, the spins are randomly oriented and precess on the surface of a cone (Fig. 2a). However, as there is a small excess of nuclei in the lower energy level, the average or net orientation of these magnetic moments is aligned along the magnetic fields and is represented by vector M_0. The intensity of the NMR signal is proportional to M_0. In an rf field, B_1, precessing at the Larmour frequency, this magnetization will rotate around an axis, which is the vector sum of $B_0 + B_1$ and will tip into the $x'y'$-plane. In a reference frame rotating about B_0 at a frequency ω_0, the nuclear moments B_1 appear stationary, and M_0 will precess about B_1 toward the y-axis (Fig. 2c). The external field (B_0) effectively disappears in the rotating frame and there is no apparent precession.

In pulsed NMR, a short rf pulse is applied to all spins at the Larmor frequency, creating a small B_1 field along the x-axis in the rotating frame and causing the nuclear spins to rotate toward the xy-plane as a coherent packet (Fig. 3). Typically, the pulse duration is such that M_0 flips into the y-axis where the decaying xy component of the magnetization can be detected as a time domain signal, commonly known as a free induction decay (FID). Each unique atom gives rise to a decaying cosine signal with its own characteristic frequency. The Fourier transformation of the time domain signal gives the characteristic frequency domain NMR spectrum.

After the perturbation, the net magnetization returns to equilibrium by two relaxation processes. The recovery of z-magnetization M_z to its equilibrium value, $M_z = M_0$, which depends on the rate of exchange of energy between the nuclei and the lattice, is called spin-lattice relaxation or longitudinal relaxation and is governed by the time constant T_1. The common method of determining T_1 is by inversion recovery.[10] Similarly, the recovery of M_{xy} to its equilibrium value $M_{xy} = 0$ is known as spin–spin or transverse relaxation, T_2. Spin-echo experiments are used for measuring T_2.[10] The line width of the resonance at half height is inversely proportional to T_2. The resonance line shape of dilute solutions is approximately Lorentzian, with a width at half-height, $\Delta\nu_{1/2}$, given by

$$\Delta\nu_{1/2} = \frac{1}{\pi T_2^*} = \frac{1}{\pi T_2} + \Delta\nu_{B_0\text{inhomo}}$$

For dilute solutions, T_2^* is large and the line width is narrow (on the order of Hz). As the solution viscosity increases, T_2 decreases and line width broadens (in solids, line widths are 10^2–10^4 Hz). For dilute solutions of small rapidly tumbling molecules $T_1 = T_2$; however, as the solution viscosity or the molecular weight increases, T_2 decreases whereas T_1 increases.

The magnetogyric ratio is different for each isotope; thus at a constant B_0, different isotopes resonate at different frequencies. For example, if $B_0 = 14\,\text{T}$, ^{1}H atoms resonate at 600 MHz and ^{13}C atoms resonate at 150 MHz. When atoms are placed in a static magnetic field, their electronic environments create small perturbations in the effective magnetic field sensed by nuclei in slightly different local environments. The

Table 1 The properties of some common NMR-active nuclei

Nucleus	I	$\gamma/10^7$ rad T^{-1} s^{-1})	Natural abundance (%)	Resonance frequency[a]	Relative receptivity[b]
^{1}H	1/2	26.8	99.99	100.00	62.90
^{2}H	1	4.1	0.02	15.35	0.23
^{13}C	1/2	6.7	1.10	25.15	1.00
^{15}N	1/2	−2.7	0.37	10.14	0.07
^{19}F	1/2	25.2	100.00	94.09	52.40
^{23}Na	3/2	7.1	100.00	26.45	1.16
^{27}A1	5/2	7.0	100.00	26.06	1.11
^{29}Si	1/2	−5.3	4.67	19.87	0.49
^{31}P	1/2	10.8	100.00	40.48	4.17

[a]Resonance frequency of nuclei in MHz at B_0 field of 2.35 T.
[b]Relative receptivity of various nuclei with respect to ^{13}C.

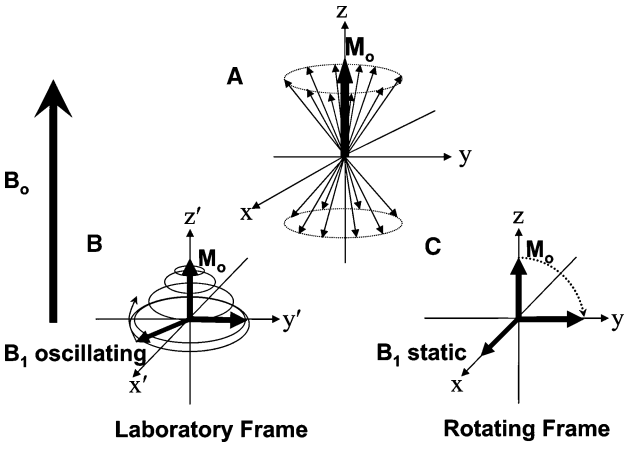


Fig. 2 (A) Macroscopic equilibrium magnetization M_0 from the vector sum of the individual nuclear moments μ; (B) magnetization interacting with B_1 rf field in the laboratory reference frame; and (C) magnetization interacting with B_1 field in the rotating reference frame.

magnetic field at the nucleus (B_{eff}) is therefore different from B_0 by a fraction σ,

$$B_{\text{eff}} = B_0(1 - \sigma)$$

where σ is known as the shielding constant. Depending on the direction of effective field, the shift in the resonance position might be deshielded (higher frequency) or shielded (lower frequency). This spread of resonance frequencies, which is proportional to B_0, is known as the chemical shift. To facilitate comparison of results from instruments with different B_0s, the chemical shifts are reported relative to the resonance frequency of a standard in field-independent units. The ^{1}H, ^{13}C, and ^{29}Si resonances of tetramethylsilane (TMS) are generally used for ^{1}H, ^{13}C, and ^{29}Si chemical shift reference standards, respectively. This chemical shift is reported in ppm and given the symbol δ, where

$$\delta = \frac{(\nu - \nu_{\text{ref}}) \times 10^6}{\nu_{\text{ref}}}$$

Structural information can be obtained from through-bond, scalar, or J coupling and through-space (nuclear Overhauser effect, NOE) magnetization transfer NMR experiments between pairs of protons.[11] Spin–spin (J) coupling is the through-bond interaction between the moments of two neighboring nonequivalent nuclei. If a nucleus is coupled to n equivalent nuclei, a multiplet is observed with approximate intensity ratios defined by polynomial coefficients of $(x + 1)^n$. The ^{1}H spectrum of 2-propanol in Fig. 4 serves as an example. Three groups of signals (multiplets) are present with relative areas 1 : 1 : 6 from the methine, hydroxyl, and methyl protons. The hydroxyl protons rapidly exchange between molecules, so coupling to them is not detected. The methyl proton is a 1 : 1 doublet from coupling to the methine proton; and the methine proton is a 1 : 6 : 15 : 20 : 15 : 6 : 1 septet from coupling to six equivalent methyl protons. The separation between peaks in both multiplets is 6 Hz because it is caused by the same interaction. An enormous amount of structural information can be obtained from such simple interpretation of ^{1}H NMR spectra.

Commonly ^{1}H NMR is used, as its high γ and essentially 100% natural abundance, and its presence in the vast majority of chemical and biological molecules ensures broad applicability. However, the narrow ^{1}H chemical shift dispersion (12 ppm) is a drawback, because ^{1}H NMR resonances in complex structures often overlap (Fig. 5). The acquisition of a ^{1}H NMR spectrum is fairly simple and takes less than a minute; however, this is not true for other nuclei that have low sensitivity because of their low natural abundance and

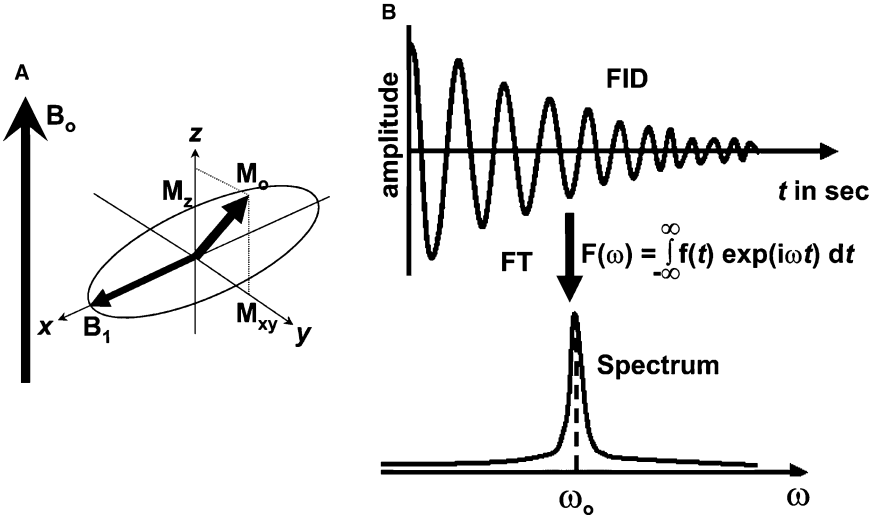


Fig. 3 (A) Interaction of M_0 with B_0 and B_1 fields; (B) time domain FID; and (C) frequency domain spectrum obtained from Fourier transformation of the FID.

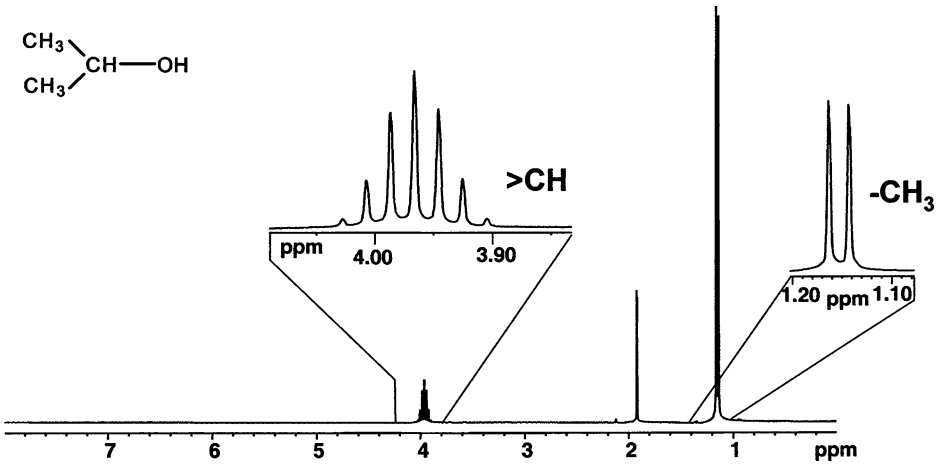


Fig. 4 ^{1}H NMR spectrum of 2-propanol showing splittings because of spin–spin coupling.

low γ. For example, the ^{13}C NMR signal is 100 times weaker than the ^{1}H NMR signal from the same number of atoms, as 99% of the carbons are ^{12}C with $I = 0$, and only 1% of the atoms are ^{13}C with $I = 1/2$. Furthermore, γ_C is only $1/4 \times \gamma_H$, resulting in an additional $(\gamma_H/\gamma_C)^{5/2}$ times, i.e., 32-fold, reduction in the ^{13}C NMR signal relative to that of ^{1}H (relative signal strengths for commonly observed isotopes are summarized in the last two columns of Table 1).

With the development of Fourier transform NMR, ^{13}C NMR spectroscopy has become a routine sensitive tool for characterizing organic and biological

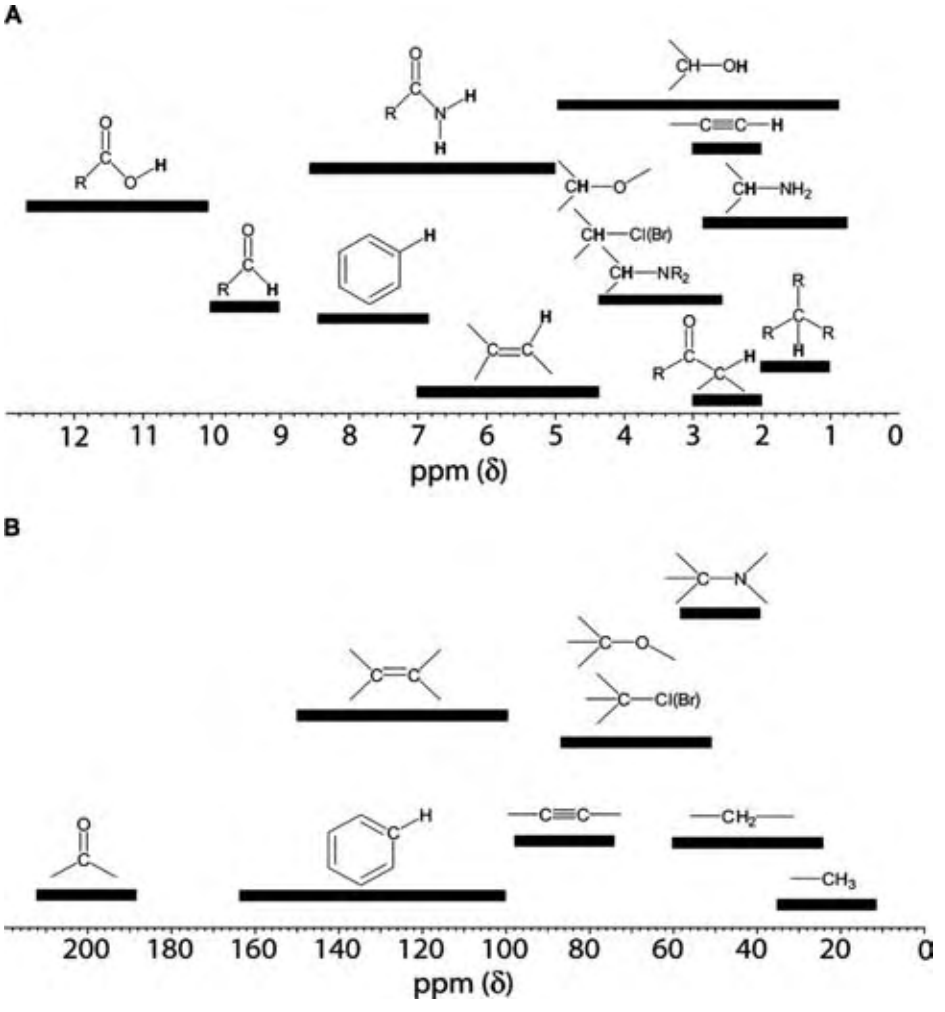


Fig. 5 Typical chemical shifts of (A) proton and (B) carbon resonances in organic compounds.

molecules. Advantages of ^{13}C NMR are its wide chemical shift range (250 ppm; Fig. 5), proton decoupling is used to remove J coupling so that each carbon produces a single resonance, the sensitivity of ^{13}C chemical shift to its chemical environment is such that carbons that differ by substituent four or five bonds away can be distinguished from each other.

Other commonly used nuclei are ^{15}N, ^{19}F, ^{29}Si, and ^{31}P. However, their use is limited in scope, as they are not always present. ^{19}F NMR is a very sensitive technique when fluorine is present, as ^{19}F has 100% natural abundance, $I = 1/2$ and γ_F is as large as γ_H providing 83% of the receptivity of ^{1}H NMR for the same number of atoms. ^{19}F NMR exhibits a large chemical shift range (several thousand ppm) and is highly sensitive to minute environmental changes.

The role of quantitative NMR in chemical processing is significant and has been studied in detail. The ratio of the areas under the NMR resonances is directly related to the relative number of atoms contributing to the resonances. This is a unique feature of NMR; in other forms of spectroscopy, absorption coefficients differ for different groups, and extensive calibrations are needed. Two factors complicate the efforts to perform quantitative analysis such as compositional analysis of mixtures NOE and different T_1s.

The NOE results in an enhancement of the ^{13}C signals when simultaneous ^{1}H decoupling is performed throughout the experiment. Under optimal conditions, signals are enhanced by threefold compared to the intensities obtained if the NOE is not present. Because the NOEs are not always full for all ^{13}C atoms, the signal intensity ratios might not accurately reflect the ratios of ^{13}C atoms. To eliminate this problem, spectra are obtained with a gated decoupling to suppress the NOE.[11] Also, for better quantitative results, it is beneficial to reduce the T_1s by introducing a paramagnetic relaxation reagent, such as Cr(acetylacetonate)$_3$. The loss of NOE can result in as much as a threefold loss in signal-to-noise ratio, requiring a ninefold increase in signal averaging time. However, when paramagnetic relaxation reagents are employed, it is common to observe a 10 to 100-fold reduction in the relaxation delay needed for quantitative analysis, more than compensating for the loss in sensitivity from the lack of NOE.

SOLUTION NMR

The majority of industrial applications of solution NMR spectroscopy are for quality control and routine analysis. Production line samples are usually taken and analyzed off-line. Many industrial NMR units are equipped with sample changers and probes capable of detecting more than one nucleus with automatic control of routine experiments, for high throughput operation. ^{1}H and ^{13}C NMR are common. For unambiguous resonance assignments, more complicated methods such as DEPT and 2D-NMR are performed. It is also possible to detect insensitive nuclei like ^{13}C through ^{1}H by 2D-correlation techniques like HMQC and HMBC. The development of pulsed field gradient (PFG) methods makes it possible to obtain artifact-free 2D and 3D spectra in a short time.

Common Industrial Applications

A common industrial application of NMR is the determination of functional groups in various natural products, synthetic chemicals, and drug molecules.[12] For example, the estimation of hydroxyl groups is performed by comparing ^{1}H NMR spectra with and without deuterium exchange of hydroxyl protons. This method is also useful in determining the contributions from configurational and conformational isomers as well as in the test of optical and enantiomeric purity, which has applications in pharmaceutical industries. Other functional groups such as aldehydes, ketones, and carboxylic acids can be easily identified by their unique ^{1}H and/or ^{13}C NMR spectra. Also, NMR is used in the analysis of vegetable oils and fats (which are predominantly triglycerides of C_{18} fatty acids) and in food processing industries. Another useful application of ^{1}H NMR is the determination of the degree of unsaturation in oils and fats, commonly known as iodine number.

The use of low-field NMR in quality control applications has increased because of the ease of operation and reliability of the data.[13] Most of the applications involve the measurement of T_2 and/or T_1 relaxation times, to yield quantitative information relating to the concentrations of individual components. The applications include the determination of viscosity and the measurement of moisture, fat, hydrogen, fluorine, and carbon contents.[14] One of the most common uses of this technique is the quantitative determination of the state of water in soils, foodstuffs, coals, and catalysts.[14]

Applications of NMR in process control, although not as popular as FT-IR and NIR, provide an alternative when the latter techniques fail. The advantages of low-field (10–60 MHz) low-resolution NMR in process control applications are its noninvasive nature and the lack of need for calibration procedures with feed or product, as it is tolerant to changes in process environment and feed quality. Process/quality control NMR applications are based on separation of the FID into separate responses from components with different T_2s (Fig. 6). Longer in-stream analysis times, high

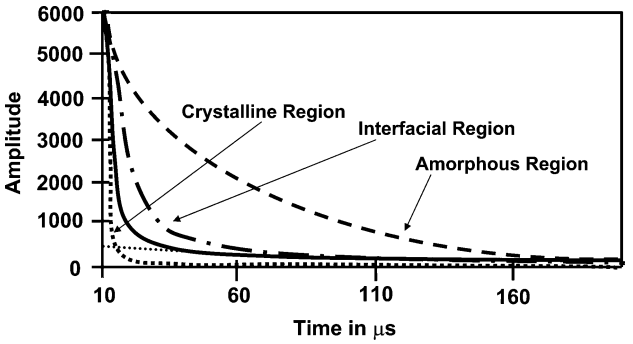


Fig. 6 A typical semicrystalline polypropylene FID (—) and its three distinct components from fast-decaying (crystalline ·····), medium-decaying (interfacial — . —), and slower-decaying (amorphous or atactic - - - - -) polymer domains.

accuracy, rapid measurement capabilities, and low instrument maintenance make process NMR useful for acid alkylation control, blending of petroleum fractions and crude oils, refinery feed optimization, base oil manufacturing, styrene–butadiene rubber, and ethylene–propylene copolymers.[14]

Although NMR is versatile in characterization of molecules and mixtures, often in industrial applications, it is utilized in combination with chromatographic techniques, such as LC-NMR in process control applications to improve efficiency.[15] Recent improvements in sensitivity, the design of flow cells, and rapid data acquisition make NMR a suitable detection system for HPLC.[15]

Polymer Processing

Solution NMR is widely used in polymer processing for the qualitative and quantitative analyses of tacticity, end-groups, degradation products, chain defects, and monomer sequence distribution.[16] A typical application is in the characterization of monomer sequence distribution by quantitative ^{13}C NMR spectroscopy.[17] For example, Fig. 7 shows a typical ^{13}C NMR spectrum of ethylene-co-1-butene.[18] From the relative peak areas, it is possible to determine the fractions of the two monomers, their reactivity ratios, the triad distribution, and the ''blockiness'' or ''randomness'' of the monomer distributions.[17] All of these structure factors play an important role in the polymer's physical and mechanical properties.

Nuclear magnetic resonance is also used to determine the tacticity of polypropylene and vinyl polymers.[19] For example, high-resolution ^{13}C NMR spectra show the variations in spectral patterns with varying tacticity of polypropylene (Fig. 8). Not only is it possible to distinguish between various stereo-regular polymers, but the higher resolution of ^{13}C NMR also makes it sensitive for measuring stereo-defects, regio-defects, and end-groups. The greatly increased spectral dispersion of solution state 2D- and 3D-NMR is increasingly being used to study complex polymer architectures.[20]

The use of NMR relaxation studies in the polymer solutions is well known for the study of polymer chain dynamics, polymer/polymer, polymer/solvent,

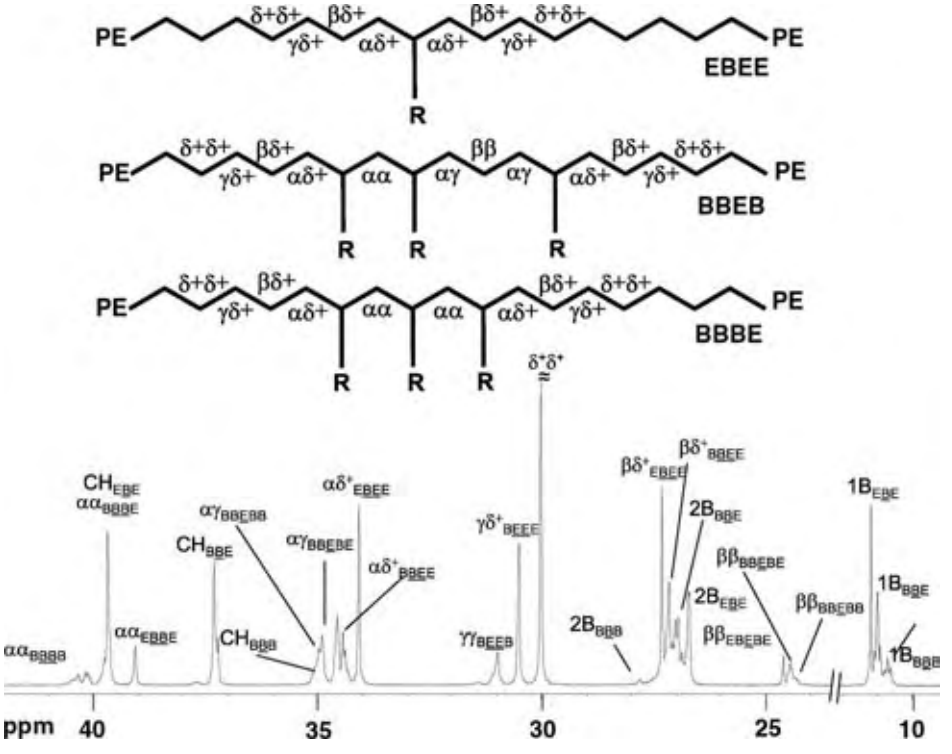


Fig. 7 Aliphatic region from the 188.6 MHz 1D ^{13}C NMR spectra of poly(ethylene-co-1-butene) containing 41% 1-butene. (From Ref.[18].)

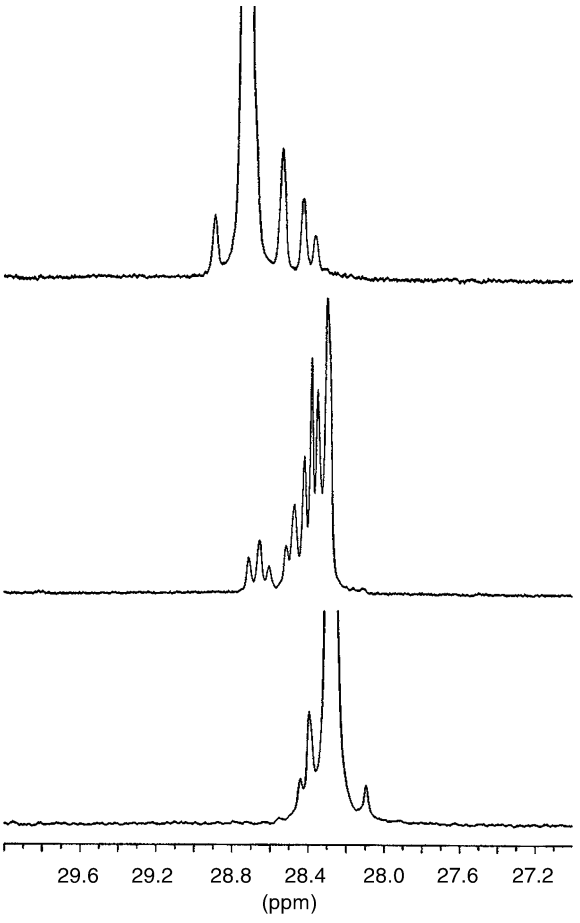


Fig. 8 Methine pentad region from the 100 MHz ^{13}C NMR spectra of isotactic (top), atactic (middle), and syndiotactic (bottom) polypropenes. (From Ref.[19].)

polymer additive interactions, and phase transitions. T_1 and T_2 relaxation measurements of dilute solutions provide valuable information on energy barriers of intramolecular motional processes linked to bulk properties.[21]

Fuels and Lubricants

The use of NMR for the compositional analysis of crude oils and fractionated products is routine in industrial production. Common analyses include the determination of saturated and aromatic hydrocarbon content, average structural parameters such as the percentages of *n*-paraffins, *iso*-paraffins, cyclo-paraffins, mono-, di-, and polyaromatics. These data are used for the development of correlations between the compositions and their characteristics.[22,23] Spectral editing such as DEPT is routinely used for the unambiguous assignments of resonances in complex mixtures, and recent trends indicate the utility of 2D-correlation techniques for such purpose.[24] In addition, NMR is used to determine additive constituents in formulated oils and lubricants. For example, ^{31}P NMR is used to analyze the structure and decomposition pathways of additives like dialkyldiothiophosphates that are found in used crankcase and gear oils. Alcohols and oxygenates are added to gasoline as fuel extenders and to improve the octane rating. ^{1}H NMR of 60 MHz is used to monitor alcohols in fuels by quantifying the CH_3 singlet at 3.4 ppm, which is correlated to the methanol concentrations in gasoline. Similarly, the concentrations of oxygenates such as methyl *t*-butyl ether (MTBE) can be determined from the signal near 3 ppm (Fig. 9).[25]

Pharmaceutical Applications

Modern approaches in drug discovery rely on NMR-based screening techniques to investigate protein–ligand interactions. Parameters of NMR such as chemical shifts, relaxation, and diffusion are used to identify small ligand molecules and macromolecular targets.[26] The development of structure–activity relationships (SAR) using 2D-NMR of proteins has opened new avenues for the discovery and analysis of drugs.[27] This technique is based on the use of chemical shift changes to screen for low-affinity ligands, in combination with structural information to direct a

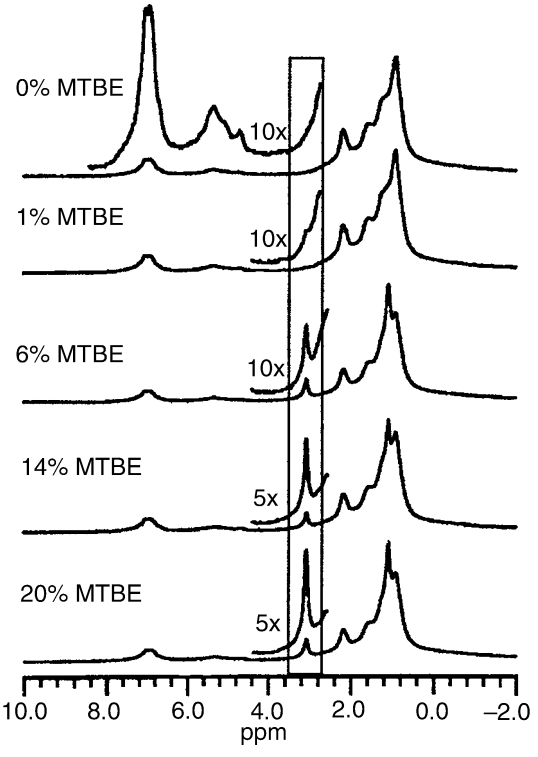


Fig. 9 ^{1}H NMR spectra at 42 MHz of various MTBE blends with regular and unleaded gasoline. The highlighted area shows signals from MTBE. (From Ref.[25].)

linked fragment approach for enhancing binding affinity. Another exciting area is high throughput screening for potential drug molecules or fragments using combinatorial techniques. The recent availability of high sensitivity cryogenic probes and flow methods makes it feasible for NMR to keep up with the requirements to screen more than 10,000 compounds per day. It is possible to prepare, characterize, and screen libraries of more than 200,000 compounds in less than a month.[26]

SOLID-STATE NMR

Solid-state NMR (SSNMR) is becoming an indispensable tool in the chemical and material processing industries because of its utility for studying structure characteristics such as internuclear distances, bond torsion angles, atomic orientations, spin diffusion, molecular dynamics, and exchange processes.[28] The most common SSNMR protocol is the use of magic-angle spinning (i.e., fast sample rotation about an axis oriented at 54.7° relative to the static magnetic field) to average chemical shift anisotropy, high power decoupling (to remove homonuclear and heteronuclear dipolar couplings), and crosspolarization (magnetization transfer from high γ nuclei, usually ^{1}H, to the observed low γ nuclei) to increase the sensitivity for detection of the low γ nuclei; the combined method is commonly referred to as CP/MAS NMR. For maximum enhancement by polarization transfer from nucleus I to nucleus S, the Hartmann–Hahn condition $\gamma_I B_{1I} = \gamma_S B_{1S}$ (where γ is magnetogyric ratio and B_1 is the rf field strength) must be satisfied. However, one of the limitations of the crosspolarization method is that it discriminates between protonated and nonprotonated nuclei, yielding nonquantitative information. For quantitative measurement, spectra obtained by CP method are compared with data from the time consuming Bloch Decay or single pulse excitation (SPE) NMR. To study the structure and dynamics of a variety of solids, SPE and CP/MAS NMR along with relaxation measurements such as T_1, T_2, and $T_{1\rho}$ are being used.

Although ^{1}H is the most commonly detected nucleus in solution NMR, this is not true in SSNMR. The strong homonuclear dipole–dipole interactions broaden the line widths of ^{1}H resonances, and fast magic-angle spinning (>40 kHz) is needed to produce spectra with the line narrowing needed to resolve signals of chemically distinct protons. Another method of line narrowing in solid-state ^{1}H NMR is the use of combined rotation and multipulse spectroscopy (CRAMPS) NMR, which is most often used for the analysis of solid substances in which MAS rate is not sufficient to narrow the lines of the proton resonances.

Polymer Processing

As it is difficult to find good solvents for many commercial polymers and there is a need to relate NMR properties directly to macroscopic properties such as crystallinity, adsorption, transport phenomena, blending, and determination of domain sizes, SSNMR is important in polymer processing.[29] The characterization of polymer morphology is important as the majority of the polymers are semicrystalline; thus SSNMR can be used for the analysis of both crystalline and amorphous phases, and exchange process between the two.[30] For example, ^{13}C CP/MAS NMR data of polypropylene can be correlated with its crystalline, interfacial/entangled, and amorphous components. It is also used for the analysis of crosslinked polymers that are generally not soluble. Another engineering-based application of SSNMR is ^{1}H and ^{13}C relaxation measurements of stimuli-responsive polymeric gels, which include superabsorbing polymers and hydrogels.[7]

Catalysis

Applications of SSNMR in catalysis include the study of zeolites and supported-metal catalysts. It has provided information on the structural changes of catalysts with changes in environment, and on interactions and reactions that occur on catalyst surfaces and in zeolite pores.[31] Most studies on zeolite structures are based on the application of SPE and CP/MAS NMR of ^{29}Si and ^{27}Al. The collection of NMR data with and without CP enhancement helps to assign the peaks in ^{29}Si NMR, as only peaks from ^{29}Si atoms having hydrogen neighbors are observed in the CP experiment, whereas the SPE experiment produces spectra from all ^{29}Si atoms (Fig. 10). ^{29}Si CP/MAS NMR was also used to study hydroxyl groups on catalyst surfaces and spillover phenomena. Multinuclear NMR can be used for the structural characterization of the coke species that deactivate catalysts. Multiple quantum NMR was used to study the distributions of adsorbed species within porous catalysts, and in situ NMR of adsorbed species is used to study mechanisms, acid site distribution, and porosity. Most of the studies focus on the state of sorbed reactants such as chloroform using ^{13}C NMR, or the state of catalyst architecture using ^{29}Si and ^{27}Al NMR.[31,32] In general, glass ampoules containing adsorbed catalyst are charged with reactants on a vacuum line, flame sealed, heated to high temperature for variable reaction times, and then quench cooled. This method has been used to identify the reactants, intermediates, and products at various stages of the methanol to gasoline (MTG) transformation (Fig. 11).[32]

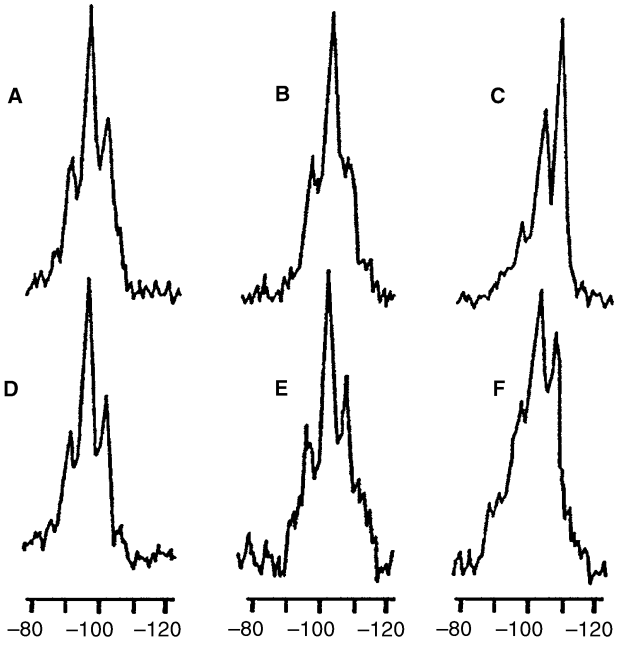


Fig. 10 ^{29}Si MAS NMR spectra of Y zeolites as synthesized (A and D) and dealuminated to varying degrees (B–F). Spectra were obtained without (A–C) and with (D–F) cross-polarization. (From Ref.[31].)

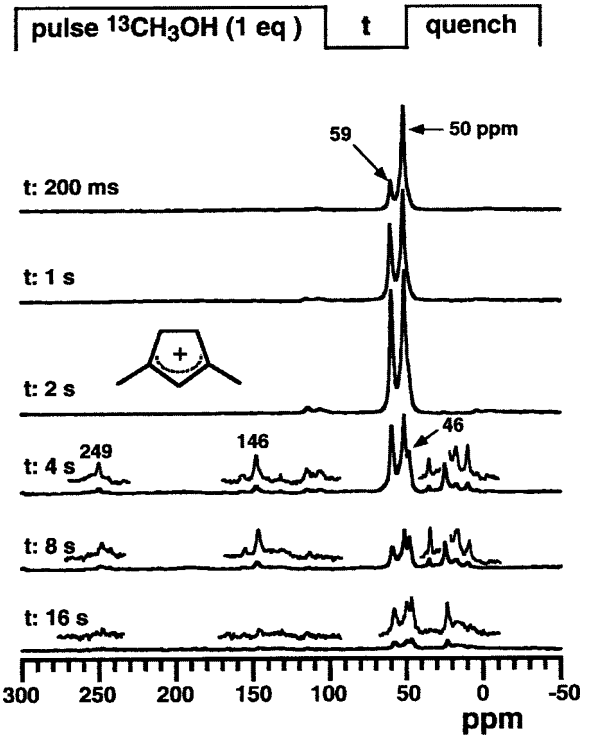


Fig. 11 ^{13}C MAS NMR spectra at 90.5 MHz from pulse-quench in situ studies of MTG chemistry on zeolite HZSM-5 at 643 K with He carrier gas flow and reaction times from 200 ms to 16 s are shown. ^{13}C signals of methanol (50 ppm), dimethyl ether (59 ppm), and cations (249, 146, and 46 ppm) are indicated. Spectra were measured at 298 K after a rapid thermal quench. (From Ref.[32].)

Pharmaceutical Applications

In the pharmaceutical industry, SSNMR is commonly used to study polymorphism, hydrogen bonding, crystal packing, and solid–solid interactions.[33] Shelf life or activity decay is often determined by the bioavailability of different polymorphs. Fig. 12 shows very different ^{13}C CP/MAS spectra from two polymorphs of an analgesic, flufenamic acid. It is also used for the study of inclusion complexes, drug–excipient interactions, or the effect of moisture on drug substances or formulations.

NMR IMAGING

Nuclear magnetic resonance imaging, commonly known as MRI, involves using magnetic field gradients to image nuclear spin density in one, two, and three dimensions with submillimeter resolution.[6] A noninvasive 3D volume image can be obtained by applying linear gradients that are varied in each of the *x*-, *y*-, and *z*-directions. The resultant spin–density image provides information regarding the concentration of given molecular species, relaxation characteristics of the sample, and local transport properties. The image construction is usually carried out by measuring the strength of ^{1}H NMR signal of mobile molecules (e.g., free H_2O in living organisms). Although proton imaging is most common, ^{19}F, hyperpolarized ^{129}Xe, ^{31}P,

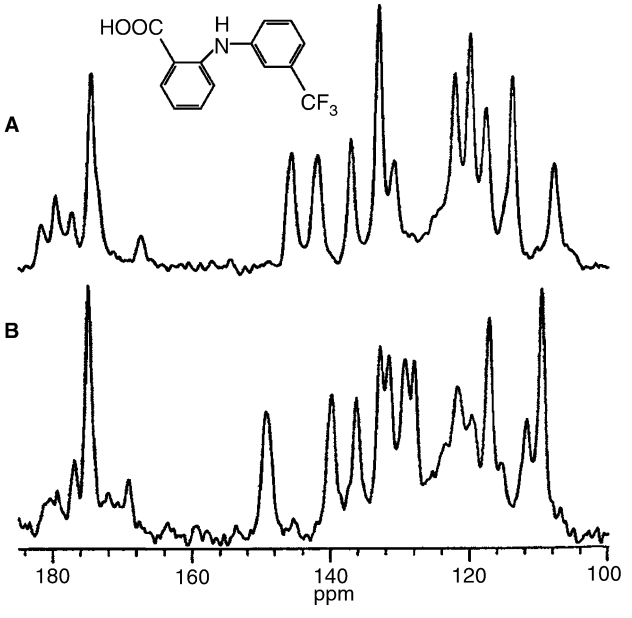


Fig. 12 ^{13}C CP/MAS NMR spectra of flufenamic acid polymorphs: A) Form III and B) Form I. (From Ref.[33].)

and other abundant nuclei can be used for constructing images of materials. These noninvasive imaging methods are used to study the nonequilibrium distribution of heat or strain, transport phenomena, and the progress of chemical reactions.[7]

Although clinical MRI has been ubiquitous for many years, only in recent years has the use of MRI grown in applications such as study of ceramic processing, catalysis, food processing, polymeric materials, and transport in reactors.[7,34] Normally, the size of objects studied is limited by the size of the magnet dimensions. One of the newer developments in this field is a mobile imaging instrument, the NMR MOUSE.[35] A mouse sized probe is moved along an object's surface, and spin density is measured in a stray field created by the mouse, making it possible to image large objects like tires.

One of the major focuses of the engineering research is the understanding of the spatial distribution of flow and diffusion. Nuclear magnetic resonance imaging has enormous untapped potential for noninvasively studying materials transport in reactions and in flowing systems. This method is used to study diffusion of fluids into the solid materials such as water migration into controlled drug release polymers, hydration of materials, and material porosity. It is also possible to investigate the pore and void structure of catalyst support pellets to evaluate the efficiency of liquid transport through these porous media. It has been used to study the influence of reactor design on catalytic processes and to study flow phenomena such as gas and liquid distributions during the trickle flow operation of fixed bed reactors.

Ceramic Processing

The nondestructive nature of MRI makes it ideal to study the microstructure at the various stages of ceramics processing. For example, T_2 contrast measurement is used to discriminate between crystalline, amorphous, and highly plasticized binder materials to measure binder distribution, damage during binder burnout, and slip-casting processing of ceramics. Unlike medical imaging, materials imaging is complicated by short T_2 relaxation times of binder materials, which are polymers such as poly(ethylene glycol), poly(vinyl alcohol), or crosslinked elastomers. These difficulties are overcome by using techniques such as homonuclear line narrowing, back-projection imaging, multiple quantum coherence, and magic-angle spinning.[36]

Polymer Processing

Nuclear magnetic resonance imaging is used extensively to characterize the structure and swelling properties of polymers.[37] It is ideal for studying material inhomogeneities such as filler clustering and air pockets, composite structures such as fiber reinforcements and cords in tires, the structure of vulcanized rubber, concentration and structural gradients as a function of crosslink density, swelling and diffusion characteristics, thermal and mechanical aging, and local stress and strain distributions. For example, Fig. 13a shows the MRI image of four layers of an elastomer differing only in their filler concentrations. The intensity of the image is readily quantified as shown by the scale on the side of the figure. The second layer from the right shows a defect that is approximately 0.2 mm in diameter. Fig. 13b shows another MRI image of a rigid polymer with a defect of a few tenths of a millimeter in diameter. Other novel applications include the study of polymerization kinetics[38] in which T_2 weighted images show the transition from ''liquid'' to ''solid'' during PMMA synthesis.

Transport in Reactors

The applications of NMR to study the transport processes within the reactors, filler cakes, and packed columns and for imaging of packing distribution and

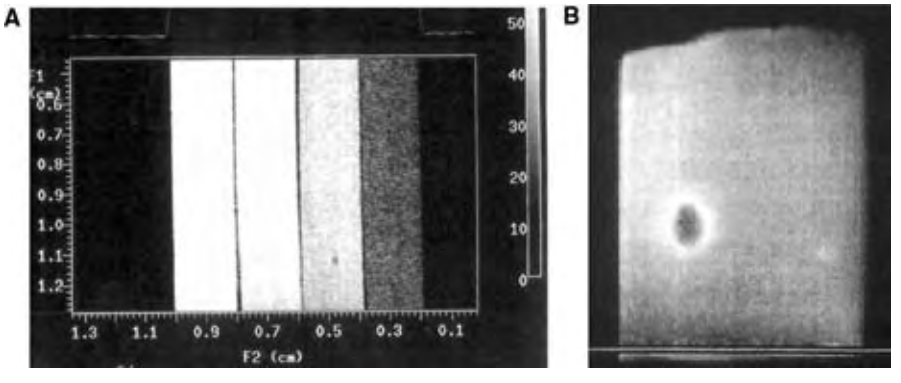


Fig. 13 (A) Cross section of MRI image (obtained at 9.4 T) of four layers of filled elastomer composite materials. Total data accumulation time was ca. 2 hr. (B) MRI image of defect in composite layer.

the fluid flow processes are of considerable interest in engineering research. Nuclear magnetic resonance imaging can be used to optimize the fabrication of preparative chromatographic columns, as relaxation measurement provides information on the mobility of the solvent in the stationary phase. ^{23}Na imaging was used to study real-time ion-exchange processes in a column.

CONCLUSIONS

Nuclear magnetic resonance spectroscopy has found application in all branches of science. Modern solution, solid-state, and imaging NMR techniques provide unique and detailed information on molecular structure, dynamics, morphology, and other interesting physical characteristics of various materials including polymers, pharmaceuticals, chemicals, and fossil fuels. Despite the expense, NMR is routinely used to monitor the characteristics of various commercial materials. When rapid, low-cost analyses of many samples are required, NMR techniques are often used in method development and as a primary standard for calibrating other low-cost spectroscopic instruments that might be used inline. With the availability of low-field, low-cost NMR instruments, the routine use of NMR is becoming popular in process/quality control applications. Expensive research grade instruments will continue to play an important role in designing production processes, in understanding the molecular basis for production problems so that they can be corrected, and in developing new process monitoring methodology.

REFERENCES

1. Purcell, E.M.; Torrey, H.C.; Pound, R.V. Resonance absorption by nuclear magnetic moments in a solid. Phys. Rev. **1946**, *69*, 37–38.
2. Bloch, F.; Hansen, W.W.; Packard, M. The nuclear induction experiment. Phys. Rev. **1946**, *70*, 474–485.
3. Ernst, R.R.; Anderson, W.A. Application of Fourier transform spectroscopy to magnetic resonance. Rev. Sci. Instrum. **1966**, *37* (1), 93–102.
4. Stejskal, E.O.; Tanner, J.E. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. J. Chem. Phys. **1965**, *42*, 288–292.
5. Grant, D.M., Harris, R.K. Eds.; *Encyclopedia of NMR*; John Wiley: New York, 1996.
6. Lauterbur, P.C. Image formations by induced local interactions. Examples employing nuclear magnetic resonance. Nature **1973**, *242*, 190–191.
7. Gladden, L.F. Nuclear magnetic resonance in chemical engineering: principles and applications. Chem. Eng. Sci. **1994**, *49*, 3339–3408.
8. Ernst, R.R.; Bodenhausen, G.; Wokaun, A. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*; Oxford: New York, 1987.
9. Braun, S.; Kalinowski, H.-O.; Berger, S. *150 and More Basic NMR Experiment. A Practical Course*; Wiley-VCH: Weinheim, Germany, 1998.
10. Farrar, T.C.; Becker, E.D. *Pulse and Fourier Transform NMR. Introduction to Theory and Method*; Academic Press: New York, 1971.
11. Gunther, H. *NMR Spectroscopy: Basic Principles, Concepts and Applications in Chemistry*; John Wiley & Sons: Chichester, 1995.
12. Wendisch, D.A.W. Nuclear magnetic resonance in industry. Appl. Spectrsc. Rev. **1993**, *28* (3), 165–230.
13. Maciel, G.E. NMR in industrial process control and quality control. In *Nuclear Magnetic Resonance in Modern Technology*; Maciel, G.E., Ed.; Kluwer Academic: Netherlands, 1994.
14. Nordon, A.; McGill, C.A.; Littlejohn, D. Process NMR spectroscopy. Analyst **2001**, *126*, 260–272.
15. Albert, K. LC NMR theory and experiment. In *On-Line LC NMR and Related Techniques*; Albert, K., Ed.; John Wiley & Sons: England, 2002.
16. Koenig, J.L. *Spectroscopy of Polymers*; 2nd Ed.; Elsevier, 1999.
17. Randall, J.C. A review of high-resolution liquid carbon-13 nuclear magnetic resonance characterizations of ethylene-based polymers. J. Mac. Sci. Rev. Macromol. Chem. Phys. **1989**, *C29*, 201–317.
18. Sahoo, S.K.; Zhang, T.; Reddy, D.V.; Rinaldi, P.L.; McIntosh, L.H.; Quirk, R.P. Multidimensional NMR studies of poly(ethylene-co-butene) microstructure. Macromolecules **2003**, *36*, 4017–4028.
19. Resconi, L.; Cavallo, L.; Fait, A.; Piemontesi, F. Selectivity in propene polymerization with metallocene catalysts. Chem. Rev. **2000**, *100*, 1253–1345.
20. Rinaldi, P.L. Polymer characterization by 3D solution NMR. ACS Symp. Ser. **2002**, *834*, 94–122.
21. Dais, P.; Spyros, A. ^{13}C nuclear magnetic relaxation and local dynamics of synthetic polymers in dilute solution and in the bulk state. Prog. Nucl. Magn. Reson. Spectrsc. **1995**, *27*, 555–633.
22. Alam, T.M.; Alam, M.K. Chemometric analysis of nuclear magnetic resonance spectroscopy data. Spectroscopy **2001**, *16* (1), 18, 20–24, 26–27.
23. Singh, I.D.; Sahoo, S.K.; Adhvaryu, A. Characterization of new generation base fluids. Tribo Test **2001**, *8* (2), 123–129.

24. Kapur, G.S.; Berger, S. Simplification and assignment of proton and two-dimensional hetero-correlated NMR spectra of petroleum fractions using gradient selected editing pulse sequences. Fuel **2002**, *81*, 883–892.
25. Skloss, T.W.; Kim, A.J.; Haw, J.F. High-resolution NMR process analyzer for oxygenates in gasoline. Anal. Chem. **1994**, *66*, 536–542.
26. Stockman, B.J.; Dalvit, C. NMR screening techniques in drug discovery and drug design. Prog. Nucl. Magn. Reson. Spectrsc. **2002**, *41*, 187–231.
27. Shuker, S.B.; Hajduk, P.J.; Meadows, R.P.; Fesik, S.W. Discovery of high affinity ligands for proteins: SAR by NMR. Science **1996**, *274*, 1531–1534.
28. Laws, D.D.; Bitter, H.-M.L.; Jerschow, A. Solid-state NMR spectroscopic methods in chemistry. Angew. Chem. Int. Ed. **2002**, *41*, 3096–3129.
29. Mirau, P.A. NMR characterization of polymers. In *Applied Polymer Science, 21st Century*; Craver, C.D., Carraher, Jr. C.E., Eds.; Elsevier, 2000.
30. Schmidt-Rohr, K.; Speiss, H.W. *Multidimensional Solid-State NMR and Polymer*; Academic Press: New York, 1994.
31. Bell, A.T.; Pines, A. *NMR Techniques in Catalysis*; Marcel Dekker Inc.: New York, 1994.
32. Haw, J.F. In-situ NMR of heterogeneous catalysis: new methods and opportunities. Top. Catal. **1999**, *8* (1, 2), 81–86.
33. Tishmack, P.A.; Bugay, D.E.; Byrn, S.R. Solid-state nuclear magnetic resonance spectroscopy—pharmaceutical applications. J. Pharm. Sci. **2003**, *92*, 441–474.
34. Miller, J.B. NMR imaging of materials. Prog. Nucl. Magn. Reson. Spectrsc. **1998**, *33*, 273–308.
35. Eidmann, G.; Savelsberg, R.; Blümler, P.; Blümich, B. The NMR MOUSE, a mobile universal surface explorer. J. Magn. Reson. Ser. A. **1996**, *122*, 104–109.
36. Brown, S.P.; Emsley, L. Solid state NMR. In *Handbook of Spectroscopy*; Vo-Dinh, T., Gauglitz, G., Eds.; Wiley-VCH: Weinheim, Germany, 2003; 269–326.
37. Blümich, B.; Blümler, B. NMR imaging of polymeric materials. Mackromol. Chem. **1993**, *194*, 2133–2161.
38. Jackson, P.; Clayden, N.J.; Walton, N.J.; Carpenter, T.A.; Hall, L.D.; Jezzard, P. Magnetic resonance imaging studies of the polymerization of methyl methacrylate. Polym. Int. **1991**, *24*, 139–143.

NMR Spectroscopy of Polymers in Solution

Sangrama K. Sahoo
Peter L. Rinaldi
Department of Chemistry, The University of Akron, Akron, Ohio, U.S.A.

INTRODUCTION

In the past decades, nuclear magnetic resonance (NMR) spectroscopy has been used extensively to study various aspects of polymer chemistry and engineering. Fig. 1 shows the relationship among polymerization conditions, polymer structure, and the material's physical structure and end uses. Solution, solid state, and imaging NMR techniques contribute to understanding the physical and chemical aspects of the route from raw materials to final product.[1,2] Solution NMR provides information about all aspects of the polymerization reactions and the final structure of the synthesized polymer. This information can be correlated with the material's final properties and provide feedback to control the initial polymerization process so that the fraction of structures responsible for desirable properties can be controlled in a systematic way.

High-resolution ^{1}H and ^{13}C NMR, and recently, multidimensional methods have revealed the microstructures of complex polymers.[3] In particular, multidimensional (2D- and 3D-) NMR have proven to be useful techniques to identify small amounts of irregular structures in synthetic polymers. In this entry, specific topics to be covered include the use of solution NMR methods to study polymer stereochemistry/tacticity, monomer composition and sequence distribution, short-chain branches, and chain-end structure, as these parameters influence the material's mechanical, thermal, optical, and electrical properties.

NMR METHODOLOGY

One of the major differences between NMR characterization of synthetic polymers and other organic structures is that polymers contain the same repeat units, differing only by their sequence, stereochemistry, branching, blocking, or termination by end-groups. As a consequence, polymer NMR spectra contain huge signals from the main chain repeat units in the presence of much smaller signals, often representing the most interesting polymer structural features. Many synthetic polymers are soluble in organic solvents and can be made in large quantities; however, detection of signals from small numbers of these unique structures presents a challenge as the dynamic range of the NMR instrument places a limit on the detection range. However, modern pulsed Fourier transform NMR with high-field magnets, advanced electronics, fast computers, advanced experimental methodology, and new data processing methods has provided researchers with the capability of detecting minor structural components in polymers below 1 unit per 100,000 atoms.[3] Nuclear magnetic resonance has the advantage over other analytical techniques in that quantitation of structural components can be accomplished without using external calibrations. Reviews of polymer NMR characterization techniques are discussed in several classical monographs.[4–6]

Observable NMR spectral parameters such as chemical shift, spin–spin coupling, and peak intensities in simple 1D spectra allow one to obtain polymer composition, tacticity, sequence distribution, and mechanism of polymerization. However, the poor mobility of polymers often leads to high viscosity solutions and rapid T_2 relaxation, resulting in the poor spectral resolution. The viscosity can be decreased by dilution and high measurement temperature. Fortunately, many polymers possess segmental mobility, which facilitates the observation of high-resolution spectra with the line widths in the range of 1–10 Hz.

Most polymer characterization is accomplished with ^{1}H and ^{13}C NMR although ^{15}N, ^{19}F, ^{29}Si, and ^{31}P NMR are also routinely used. ^{1}H NMR is most commonly used because it is present in most chemical structures, has 100% natural abundance and a high gyromagnetic ratio, making it the most sensitive of all the NMR observable nuclei to detect. Moreover, because of its fast relaxation and the ability to extract information from spin–spin coupling constants, ^{1}H NMR is one of the most useful techniques for determining chemical structure. However, because of its small chemical shift range (only 10 ppm) and the complexity of each proton resonance, severe peak overlap occurs for ^{1}H atoms situated in similar chemical environments.

On the other hand, the ^{13}C nucleus has low natural abundance of only 1.1% and a gyromagnetic ratio one-fourth that of ^{1}H, making ^{13}C detection almost 6000 times more difficult than ^{1}H detection. The advantages

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120024576

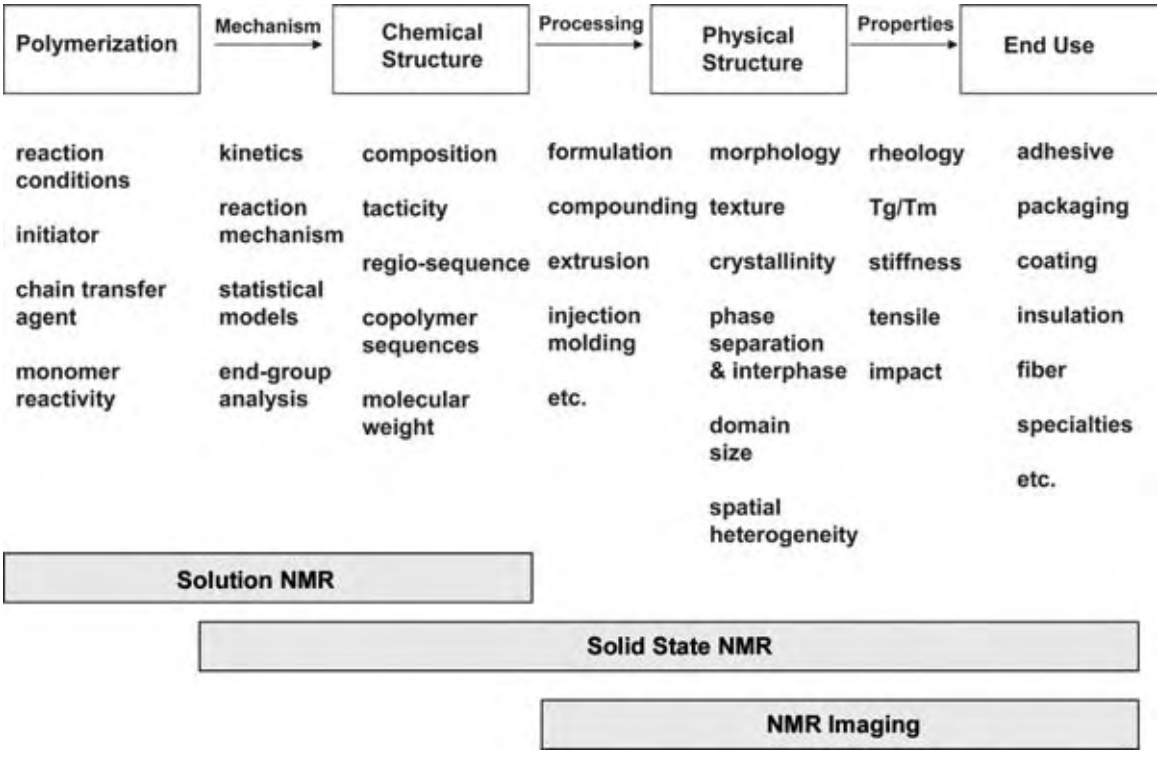


Fig. 1 Summary of polymer NMR studies. (From Ref.[1].) *(View this art in color at www.dekker.com.)*

of ^{13}C NMR are its large chemical shift range of 250 ppm, the sensitivity of its chemical shift to the small changes in its chemical environment, and that ^{13}C signals appear as single sharp lines. ^{13}C NMR is also insensitive to long-range effects, such as solvent effects and diamagnetic anisotropy of neighboring groups. Modern, high-field NMR instruments make ^{13}C NMR routine, despite its insensitivity.

The most fundamental step in NMR studies of polymer microstructures is the assignment of resonances to structural features of the polymer. This is accomplished by a variety of methods, including the use of low molecular weight model compounds, estimation of chemical shift by using empirical relationship, synthesis of polymers having known structures, isotopic enrichment, spectral simulation based on statistical models, NMR spectral editing, and multidimensional NMR.[4] The following section presents a brief overview of these methods of resonance assignments with some specific examples pertaining to ethylene-based polymers, as these are one of the most important groups of polymers that have been extensively characterized by NMR.

Model Compound Analysis

One of the first methods of analyzing polymer microstructure was to systematically study model compounds and to compare their spectra with those of the polymer to provide resonance assignments. The major drawbacks of this characterization method are the difficulty in obtaining suitable model compounds and the fact that the methodology is time consuming, and that prediction of chemical shifts can be ambiguous in the spectra of complex structures where many peaks fall in a narrow region.

When the polymerization mechanism is known, peak intensities in the spectra of polymers with different monomer compositions can be compared with calculated intensities based on various polymerization models [e.g., first order Bernoullian (B) or first order Markovian model]. For example, if polymerization mechanism of a vinyl monomer fits a Bernoullian model, then the fractions of triads will be given by:

$$P_{mm} = P_m^2 = (m)^2 \quad (1)$$

$$P_{rm/mr} = 2P_m(1 - P_m) = 2(r)(m) \quad (2)$$

$$P_{rr} = P_r^2 = (1 - P_m)^2 = (r)^2 \quad (3)$$

where P_m is the probability that a monomer will add in a meso (m) fashion to the end of the growing chain and P_{xx} is the final fractional probability of forming xx diad units in the polymer chain. Even when the statistical model for polymerization is not known, most resolvable

Table 1 A_l and corrective term S_l used in chemical shift calculation

Carbon position, l	A_l (ppm)	Carbon types	S_l (ppm)
α	9.1 ± 0.10	1° (3°)	−1.10 ± 0.20
β	9.4 ± 0.10	1° (4°)	−3.35 ± 0.35
γ	−2.5 ± 0.10	2° (3°)	−2.50 ± 0.25
δ	0.3 ± 0.10	2° (4°)	−7.5
ε	0.1 ± 0.10	3° (2°)	−3.65 ± 0.15
		3° (3°)	−9.45
		4° (1°)	−1.50 ± 0.10
		4° (2°)	−8.35

(From Ref.[7].)

resonances can be assigned based on qualitative comparisons of the signal intensity variations with varying polymer compositions.

The estimation of chemical shifts by examining the spectra of model compounds is not always feasible, and the prediction models fail to distinguish between two or more stereosequences as they cannot always be distinguished on the basis of intensity alone. To overcome these limitations, large numbers of organic compounds have been analyzed by NMR and their chemical shifts have been used to determine a set of empirical correlations that are used to determine the structure based on the polymer's NMR spectrum. The carbon chemical shifts of hydrocarbon-based polymers such as polyethylenes can be predicted by empirical additivity rules such as:

$$\delta_C = B + \sum_l A_l n_l + \sum_l S_l \qquad (4)$$

where δ_C is the calculated chemical shift, B the chemical shift of methane (−2.3 ppm), n_l the number of carbons at position l bonds away from the carbon of interest, A_l the additive shift because of carbon l, and S_l a term used to account for branching; the values of A_l and S_l are given in Table 1.[7] These empirical correlations are very useful for main chain resonance assignment and in many cases are valid for assignments of resonance of end-groups, branches, crosslinks, and regioisomers, which show distinctly different chemical shifts and *γ-gauche* effects. The *γ-gauche* effect along with the rotational isomeric state (RIS) model[8] has also helped in characterizing polyolefin tacticity.[9]

Isotopic Enrichment

Selective isotopic labeling overcomes ^{13}C NMRs insensitivity and permits reliable peak intensity measurements.[10] The enriched sample allows NMR to be performed on dilute solutions, reducing sample viscosity, with concomitant line narrowing. Selective spectral windows can be used to record the NMR spectrum without worrying about the folding of weak signals from unlabeled sites; hence, better resolutions can be achieved especially in 2D-NMR experiments. Selective labeling also facilitates the detection of weak resonances from chain-end structures, which give information on the polymerization initiation and termination mechanisms. The sensitivity enhancement achieved ameliorates the dynamic range limitations of NMR.

Statistical Modeling

Statistical models for the analysis of NMR data are used in two complementary approaches (Fig. 2): an analytical (model fitting) approach and a synthetic (computer simulation) approach.[11] In the analytical approach, assigned NMR resonance intensities are fit to expected intensities based on statistical models. In the synthetic approach, spectral intensities are first calculated using reaction probabilities predicted by theoretical models; these theoretical intensities are matched with those observed in the NMR spectrum. The calculation is based on theoretical probability expressions or Monte Carlo simulation. In an integrated approach, both methods are used for more complex systems.[12]

Simple propagation models discussed earlier fail to provide good fits when there is compositional heterogeneity in the polymer structure because of different comonomer reactivity ratios or deviations from the statistical combinations of comonomer placements on polymer chains. To overcome these drawbacks,

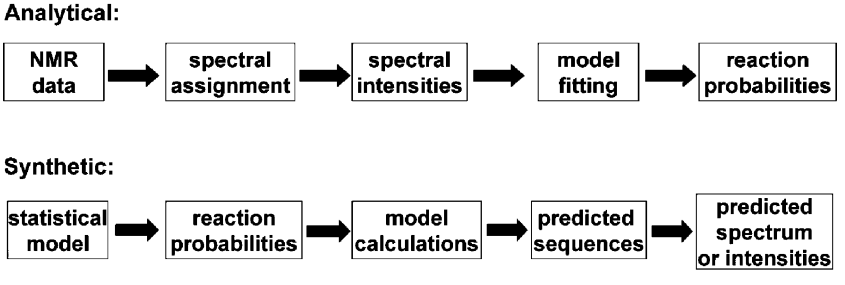


Fig. 2 Statistical models to analyze NMR data. (From Ref.[11].)

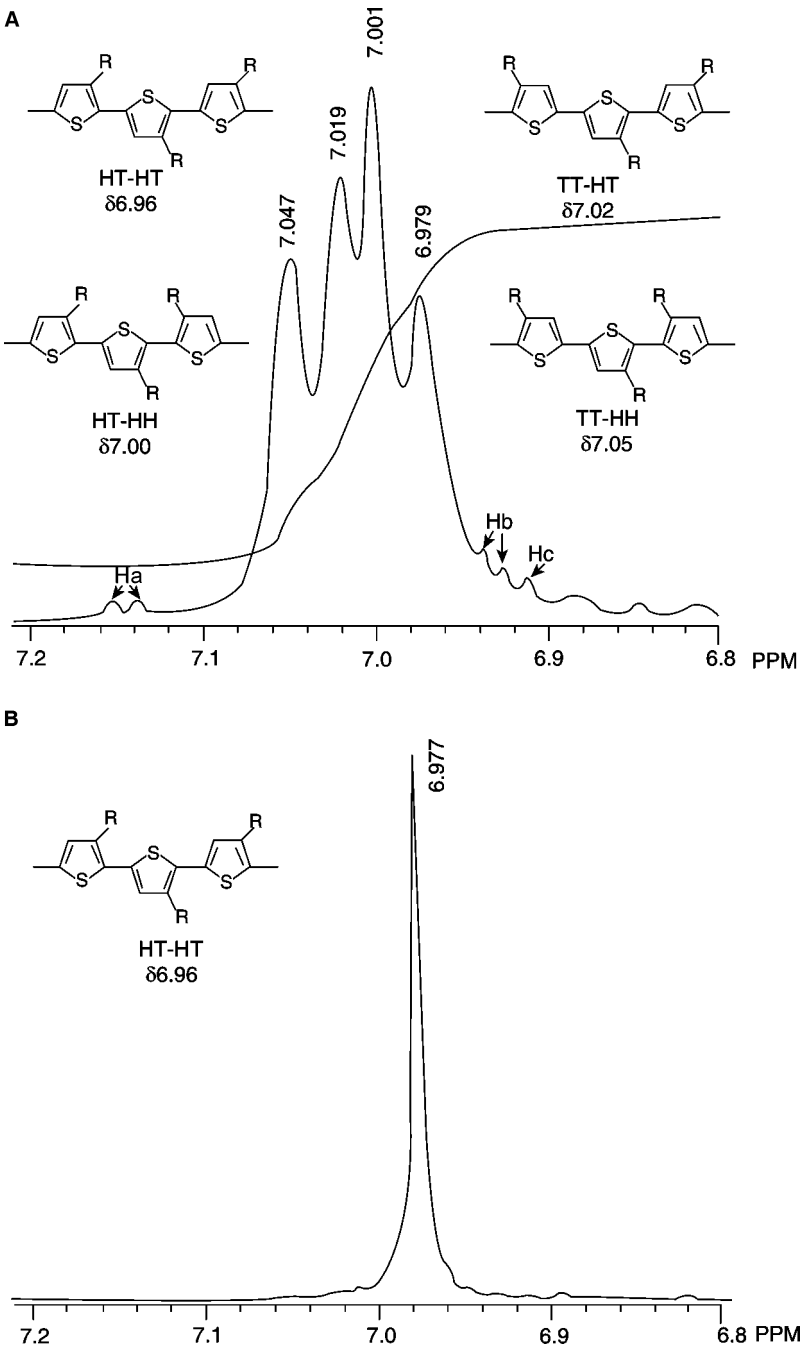


Fig. 3 Proton NMR spectra of (A) regioirregular and (B) regioregular poly(3-hexyl thiophene). (From Ref.[23].)

various models were proposed to specifically include the effects of compositional heterogeneity.[6] Models that are suitable for statistical, conversion, and process heterogeneities can be generalized as the perturbed Markovian model, perturbed enantiomorphic-site (E) model, perturbed model with exponentially modified Gaussian, and perturbed model applicable to any arbitrary chemical composition distribution.[12] For multicomponent heterogeneity, the bicatalytic-site model (involving a B component and an E component) is routinely used.

Multidimensional NMR

Recently, multidimensional NMR has become a standard technique for the assignment of polymer resonances and for characterization of polymer microstructure. Two-dimensional and three-dimensional NMR have been useful for identification of resonances from trace structures such as block junctions, chain-ends, or chain branches that are usually obscured in 1D-NMR spectra because of overlap with polymer backbone resonances. Complex spectral features are simplified

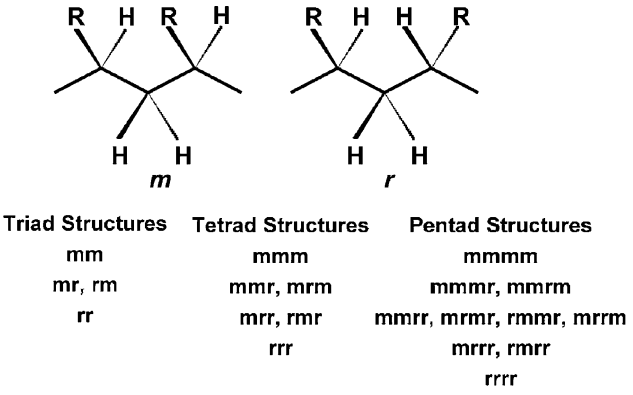


Fig. 4 Stereochemical arrangement of groups in a vinyl polymer.

by dispersion of the overlapping 1D-NMR resonances into additional frequency dimensions, and the correlations observed in multidimensional spectra provide atomic connectivity information, helping in identification of structures in complex polymers. There are many types of 2D-NMR experiments; however, the most commonly used methodologies for the polymer analysis depend upon combinations of homonuclear correlation experiments such as ^{1}H–^{1}H COSY and NOESY;[13] or heteronuclear ^{13}C–^{1}H HETCOR and COLOC, ^{1}H–^{13}C HMQC, HSQC, and HMBC;[14] or ^{13}C–^{13}C INADEQUATE experiments.[15]

The methodologies based on ^{1}H–^{1}H COSY and NOESY experiments are the most sensitive, as they rely on detection of high-sensitivity, abundant ^{1}H nuclei. COSY provides correlations among ^{1}H atoms that are *J*-coupled, while NOESY provides those among protons based on their separation in space. The drawback of using these homonuclear 2D methods is their complexity, as two coupled protons can produce many peaks in the 2D spectrum.

Two classes of heteronuclear correlation experiments are used to provide information about one-bond and two/three-bond carbon–hydrogen attachments. These experiments are 100-fold less sensitive, as they require detection of signals from 1% of the molecules containing ^{13}C. The earlier experiments such as ^{13}C–^{1}H HETCOR and COLOC involve ^{13}C detection; these have been superseded by more sensitive ^{1}H-detected HMQC, HSQC, and HMBC experiments, which provide ca. 30-fold sensitivity improvement over ^{13}C detection. Heteronuclear correlation experiments provide simpler spectra (a single peak is observed for each C–H attachment) and they take advantage of the much greater ^{13}C spectral dispersion.

Homonuclear ^{13}C–^{13}C INADEQUATE experiments are the most difficult and least frequently used. They require detection of ^{13}C–^{13}C fragments that are present in only $1/10^{4}$ molecules. When sufficient sample and instrument time is available, they provide unambiguous identification of a molecule's entire carbon skeleton and utilize the enormous ^{13}C chemical shift dispersion in both dimensions of the 2D spectrum.

Triple resonance 3D-NMR has been used to study polymers containing three NMR active nuclei,[16] and parallel methods for studying hydrocarbon-based polymers are now being used.[17] ^{1}H–^{13}C–X (X = ^{19}F, ^{29}Si, and ^{31}P) NMR correlation experiments help to simplify and unambiguously assign both main chain and chain-end resonances in a variety of fluoropolymers, polyorganosilanes, and phosphorus-containing polystyrenes.

The availability of pulsed field gradient (PFG)[18] techniques has had the most significant impact in terms of making it possible to perform 2D- and 3D-NMR experiments on polymer samples. These methods have taken the place of traditional radiofrequency phase cycling methods for coherence selection. By optimizing the use of the spectrometer's dynamic range, PFG techniques not only save time but also drastically reduce artifact noise.

Often it is impossible to resolve the signals from different molecular species in complex polymer mixtures by simple 1D- and 2D-NMR. However, development of diffusion-ordered spectroscopy (DOSY),[19] which employs PFGs, provides another method of separating the resonances of different polymer species. In DOSY experiments, the normal 1D spectrum is obtained in one dimension of the 2D spectrum and these spectra are dispersed in a second dimension based on the structure's diffusion coefficient. It is possible to resolve separate spectra of small monomer and large polymer molecules;

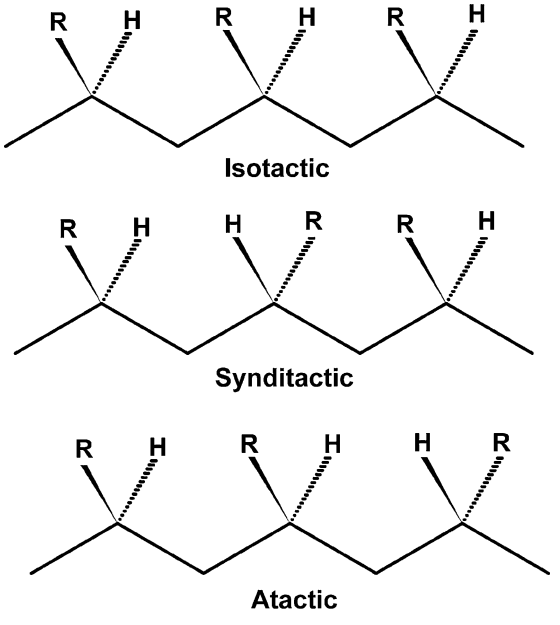


Fig. 5 Regular stereochemical arrangements found in vinyl polymers.

to produce a spectrum with dispersion based on molecular weight distribution of polymer molecules; and to distinguish between rapidly moving chain-ends and main-chain groups in the same polymer molecule.

Hyphenated NMR Techniques

Recent trends indicate an increased use of hyphenated NMR techniques, especially those involving NMR in combination with chromatographic methods such as HPLC, GPC, or SEC.[20] Online SEC- and GPC-NMR have found useful applications in the direct determination of molecular weight and molecular weight distribution.[21,22]

APPLICATIONS

Structural Isomerism

Three kinds of structural isomerism are observed in polymers: regioisomerism, stereoisomerism, and geometrical isomerism, depending on the polymerization mechanism. For example, with modern catalysts, structures contained in polyolefins can be controlled to determine stereoregularity, monomer sequence distribution, and branching. The relative amounts of these structures are directly related to the polymer's preparation conditions and end properties. Nuclear magnetic resonance provides the key link between the polymer's production/processing conditions and its characteristics.

Regioisomerism arises from head-to-tail (H-T), head-to-head (H-H) vs. tail-to-tail (T-T) addition of asymmetric monomer units. Regiodefects produce large changes in chemical shifts compared with the shifts of atoms in regioregular parts of the molecule. For example, in poly(2-hexylthiophene-2,5-diyl) distinct resonances from regioregular (HT-HT) and regioirregular (HT-HT, HT-HH, TT-HT, and HH-TT) polymers are observed in the aromatic region of the ^{1}H NMR spectrum shown in Fig. 3.[23] Moreover, 2D INADEQUATE NMR has been used to characterize regioregular and regioirregular polypropylenes[15] and to identify the resonances from H-H and/or T-T units. Once the resonances from these defect structures are reliably identified, their relative intensities can be used to determine the relative abundances of these structures.

Nuclear magnetic resonance has been extensively used to investigate stereoisomerism in polymers such as polypropylene, polyacrylates, vinyl polymers,

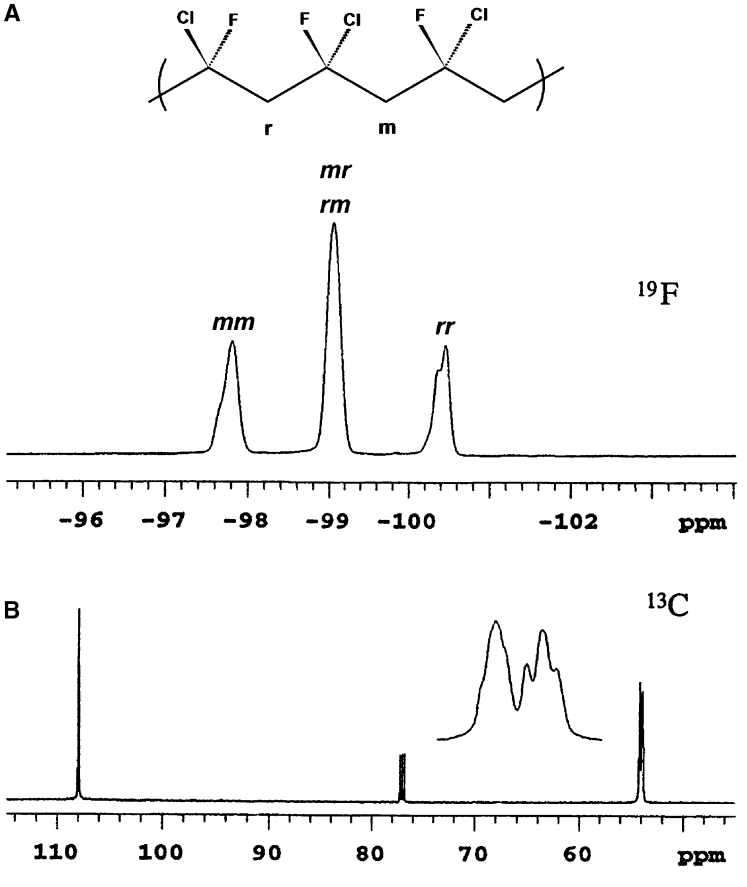


Fig. 6 NMR spectra of poly(1-chloro-1-fluoroethylene): (A) ^{19}F spectrum, (B) ^{13}C spectrum obtained with ^{1}H and ^{19}F decoupling. (From Ref.[24].)

ethylene/α-olefin copolymers, and a variety of other copolymers. Monosubstituted olefins when polymerized introduce stereogenic centers along the polymer's backbone. As most polymers are prepared in the racemic form, absolute stereochemistry is unimportant. However, the relative configurations of sites in neighboring monomer units are important in determining and understanding polymer structure–property relationships. The two possible relative configurations of stereogenic centers in consecutive monomer repeat units (diads) are meso (*m*) and racemic (*r*) (Fig. 4). If the polymer backbone is drawn in an all trans-planar conformation, as illustrated in top portion of Fig. 4, the *m* diad has R-substituents in adjoining monomer units on the same side, and the *r* structure has the R groups on opposite sides. The configurations of longer monomer stereosequences are defined by the relative configurations of their component diads.

A monosubstituted vinyl monomer yields polymers having three regioregular arrangements of configuration (Fig. 5), described by triad stereosequences. The isotactic structure has all R groups on the same side of the backbone (*mm*) the syndiotactic structure has R groups on alternate sides of the polymer's backbone (*rr*) and the heterotactic or atactic structure has R groups randomly oriented on either side of the polymer's backbone.

In the NMR spectra of polymers of monosubstituted ethylenes, resonance patterns from the atoms at branch sites and substituents bound to it are determined by the configuration at the α-carbon relative to those of the stereogenic centers on the adjacent monomer units.[5] The observed patterns will be those arising from the four possible triad sequences: *mm*, *mr*, *rm*, and *rr*. If the polymerization is random, four stereosequences are produced in equal amounts, *mr* and *rm* triad sequences are usually indistinguishable, and three resonances in a ratio of 1 : 2 : 1 are observed for *mm*, *mr*/*rm*, and *rr* triad stereosequences. The ^{19}F NMR spectrum of poly(1-chloro-1-fluoroethylene) (Fig. 6A) serves as an excellent illustration.[24] When the substituents have a large influence on shielding, or when the spectra are obtained at very high magnetic field, the chemical shift of a nucleus can be sensitive to the stereochemistry more than one monomer away. In these cases, the resonance patterns observed will be for higher *n*-ad (where n = odd integer) stereosequences. The unresolved fine structure in Fig. 6A arises from these remote stereosequence effects. When observing the CH_2 resonances of methylenes between stereogenic centers, chemical shifts are influenced by their relative configurations, resulting in two resonances from only two possible diad sequences, *m* and *r*. This pattern is clear in the ^{13}C NMR spectrum of poly (1-chloro-1-fluoroethylene) obtained with both ^{1}H and ^{19}F decoupling (Fig. 6B). Two groups of methylene ^{13}C resonances are observed near 54 ppm; the incompletely resolved splitting at 54 ppm arises because the

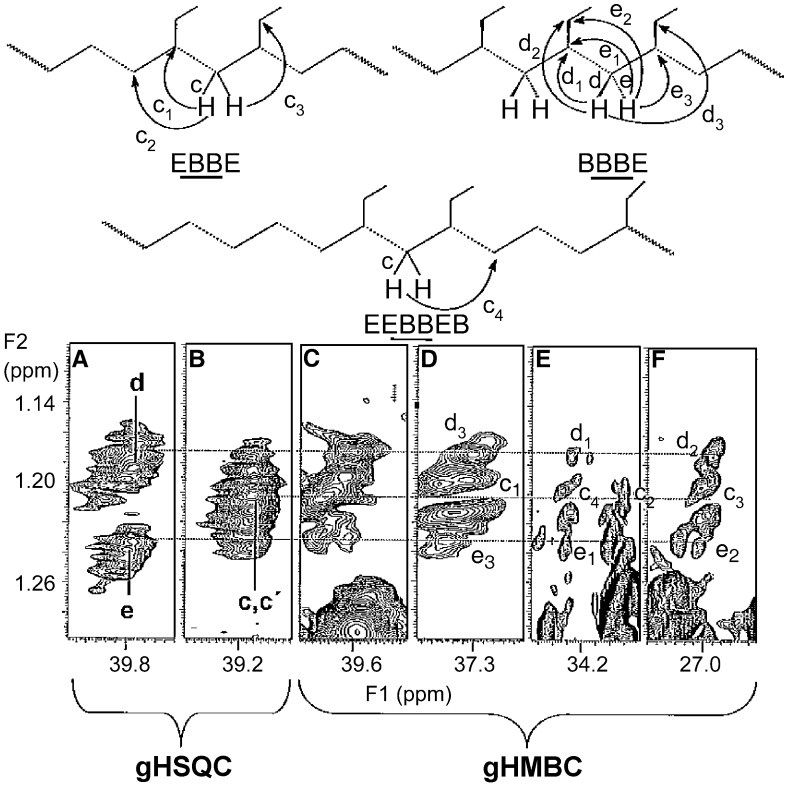


Fig. 7 Expansions from the gHSQC and gHMBC spectra of poly(ethylene-co-1-butene) in the αα proton chemical shift region: (A, B) gHSQC spectrum showing only the carbon chemical shift regions containing crosspeaks, (C–F) gHMBC spectrum showing only the carbon chemical shift regions containing crosspeaks. (From Ref.[14].)

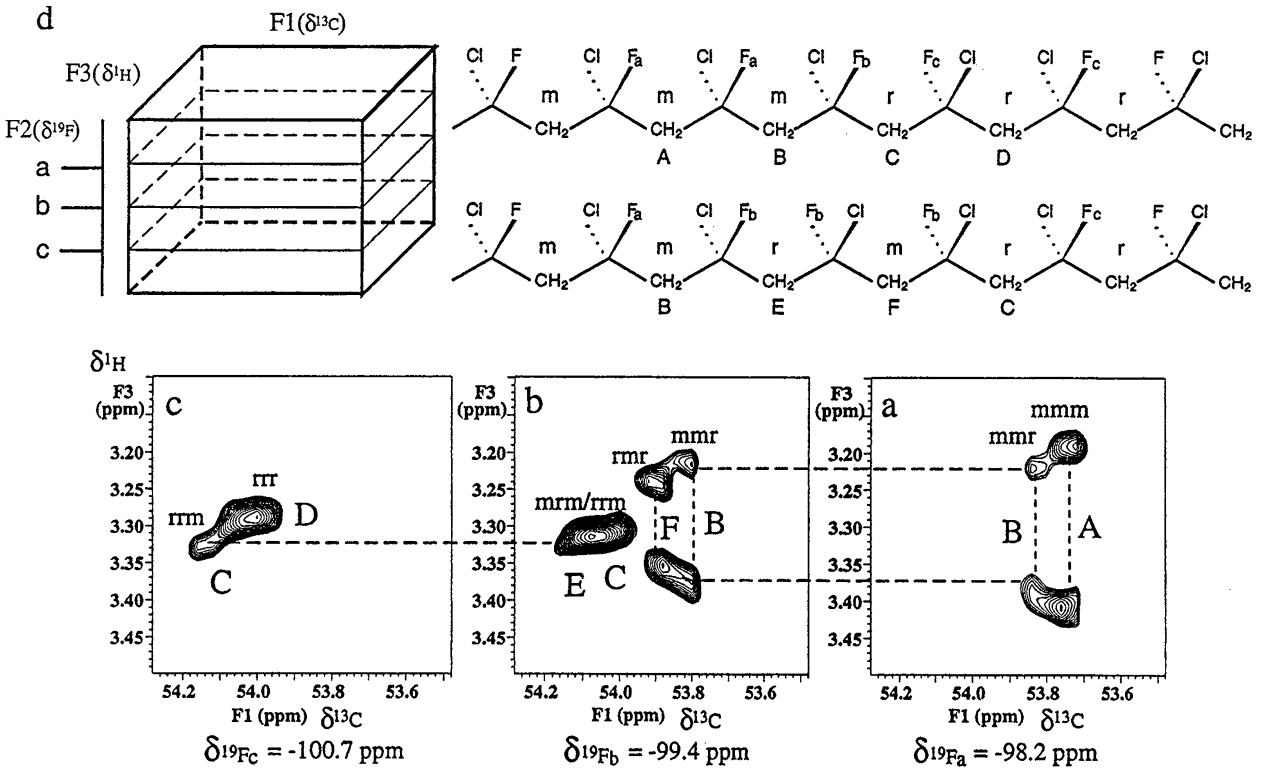


Fig. 8. 3D-HCF NMR spectrum of PCFE with $f1f3$ slices at different ^{19}F frequencies: (a) $f1f3$ slice at $\delta^{19}F = -98.2$ ppm; (b) $f1f3$ slice at $\delta^{19}F = -99.4$ ppm; (c) $f1f3$ slice at $\delta^{19}F = -100.7$ ppm; (d) schematic illustration of the 3D spectrum showing the relative positions of the slices. (From Ref.[24].)

methylene chemical shifts are sensitive to structure differences two monomer units away. Generally, the resonance patterns of symmetric groups between stereogenic centers are determined by the relative distributions of diad or higher n-ad (where n = even integer) stereosequences.

Although the patterns in Fig. 6 indicate the influence of remote stereosequence effects on the spectrum,

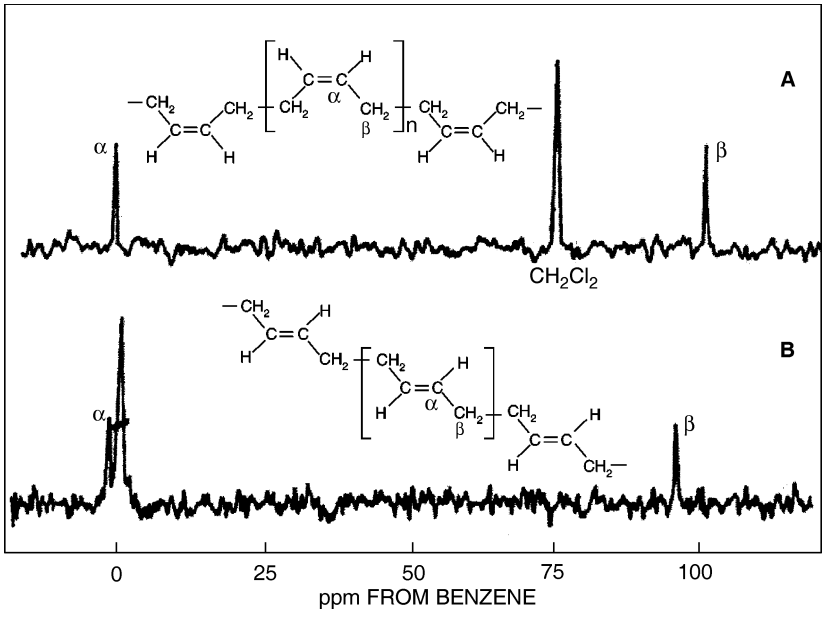


Fig. 9 ^{13}C NMR spectra of 15.08 MHz of (A) *cis*-1,4-polybutadiene and (B) *trans*-1,4-polybutadiene. (From Ref.[25].)

it is difficult to resolve resonances from higher *n*-ad sequences. By dispersing the resonances into two frequency dimensions using multidimensional NMR, it is routinely possible to resolve and assign resonances from stereosequence effects in polymers. One such example is seen in the HSQC and HMBC NMR spectra of poly(ethylene-co-1-butene), which shows the presence of *m* and *r* diads in sequences with two butene units (Fig. 7).[14] Additional detail can be resolved with the addition of more frequency dimensions. For example, from the $^{1}H/^{13}C/^{19}F$-triple resonance 3D-NMR spectrum of poly(1-chloro-1-fluoroethylene), six methylene and ten ^{19}F resonances are resolved from the possible tetrad and pentad stereo sequences, respectively (Fig. 8).[24]

Lastly, NMR is useful in identifying geometrical isomers when *cis*- and *trans*-double bonds are formed from polymerization of dienes such as butadiene. ^{1}H and ^{13}C 1D-NMR is commonly used to distinguish between these isomeric forms, as they have distinct chemical shift patterns as shown by the spectra of *cis*- and *trans*-polybutadiene in Fig. 9.[25]

Monomer Sequence Distribution

The examination of monomer sequence distributions by NMR is one of the most extensively used applications in materials science. When two (or more) dissimilar monomers A and B are copolymerized, a polymer is obtained with varying placements of A and B units along the backbone as shown in Fig. 10. It is important to know the relative distribution of monomer sequences, as these have an influence on the polymer's properties, and information about the distributions is valuable for studies of copolymerization mechanism. Initially, NMR was the only technique available to determine monomer sequences.

The monomer sequence is governed by the relative preference of monomer A to add to a growing polymer chain-end containing the same monomer P-A (where P is the rest of the polymer chain) or the comonomer B (in P-B) as illustrated in Fig. 11. This preference is described by reactivity ratios r_1 and r_2, the ratios of the rate constants for the addition of similar and dissimilar monomer units at the growing chain-end.

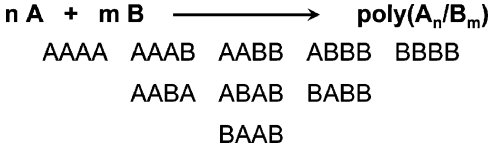


Fig. 10 Reaction of monomers A and B to form a copolymer, and the 10 possible tetrad monomer sequences.

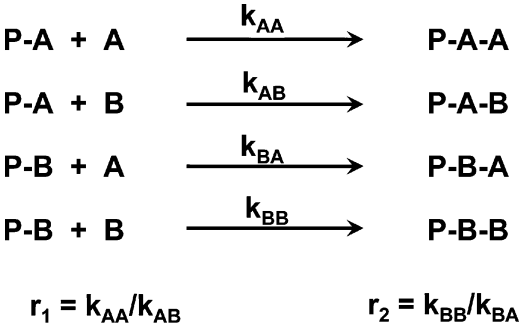


Fig. 11 Reactivity ratios (r_1 and r_2) for addition of monomer to a polymer chain-end containing similar or dissimilar monomer units.

The reactivity ratios can be calculated from the resonance intensities of different monomer sequences in the spectra of polymers obtained from termination of the polymerization at low monomer conversion (<10–20% to avoid complications from changes in monomer concentrations during the reaction).

Generally, the NMR analysis begins with comparison of the spectra of the corresponding homopolymers and a series of copolymers prepared by systematically varying the monomer compositions. Fig. 12C shows the C–F resonances from the ^{13}C NMR spectra

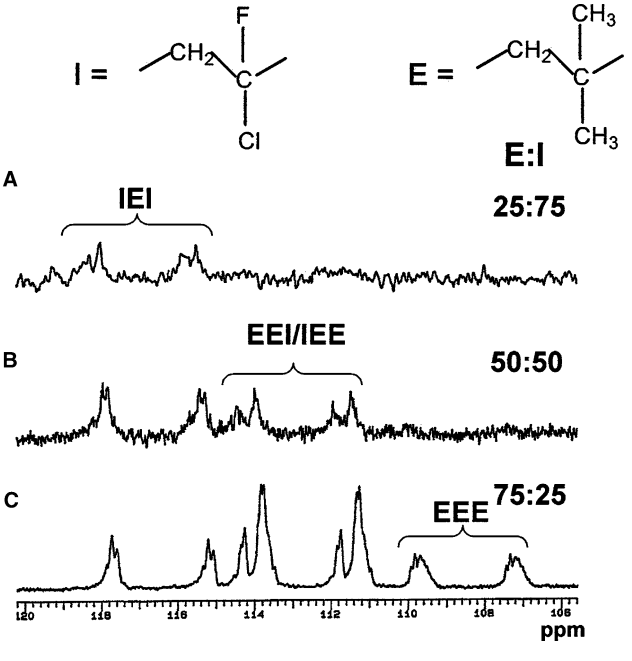


Fig. 12 Expansion of the C–F region from the ^{13}C NMR spectra of poly(1-choloro-1-fluoroethylene-co-isobutylene) obtained from different monomer feed ratios of 1-choloro-1-fluoroethylene (E) and isobutylene (I).

of copolymers prepared from different ratios of 1-chloro-1-fluoroethylene (E) and isobutylene (I). The spectra contain three groups of doublets with shift ranges of 107–110, 111–115, and 115–118 ppm. These groups of peaks arise from C–F carbons of E units centered in EEE, EEI/IEE, and IEI triads. The large doublet splitting results from one-bond C–F couplings. The additional fine structure is a result of stereosequence and longer-range monomer sequence effects.

Most copolymer analyses are carried out by fitting the calculated spectra to experimental ^{1}H and/or ^{13}C NMR spectra based on various propagation models.[26] More detailed information on the compositions of higher order (greater than tetrads) monomer sequences can be obtained only by combining the results from 1D- and 2D-NMR. Examples of these applications can be found in studies on poly(ethylene-co-1-alkene), poly(ethylene-co-alkyl acrylate), and poly(ethylene-co-vinyl monomer).[6,26]

Chain-End Analysis

Nuclear magnetic resonance is useful for determining the number and the structure of chain-ends. The structures of polymer chain-ends are usually relatively complex, producing many unique ^{13}C resonances. Identification of these resonances provides information about the regiospecificity, stereospecificity, the relative rates, and the mechanisms of initiation and termination reactions. In many cases, the unique chain-end structures are the primary reaction sites for degradation and further reactions in otherwise inert materials. Identification of only a single chain-end resonance permits calculation of the polymer's number average molecular weight (M_n). Most of the NMR analyses of chain-end structures rely upon chemical shift calculation and use of model compounds. A good example is presented in Fig. 13 showing 750 MHz ^{1}H NMR spectrum chain-end structures arising from initiation, chain transfer, and termination reactions of poly(methyl methacrylate) (PMMA) prepared with benzoylperoxide in toluene at 100°C;[27] resonances are observed corresponding to *m* and *r* chain-end diad sequences.

Multidimensional NMR combined with multiple resonance techniques can be used to remove most of the resonances from the spectrum and to selectively detect unique structures such as those found at chain-ends.[28] For example, the unique occurrence of

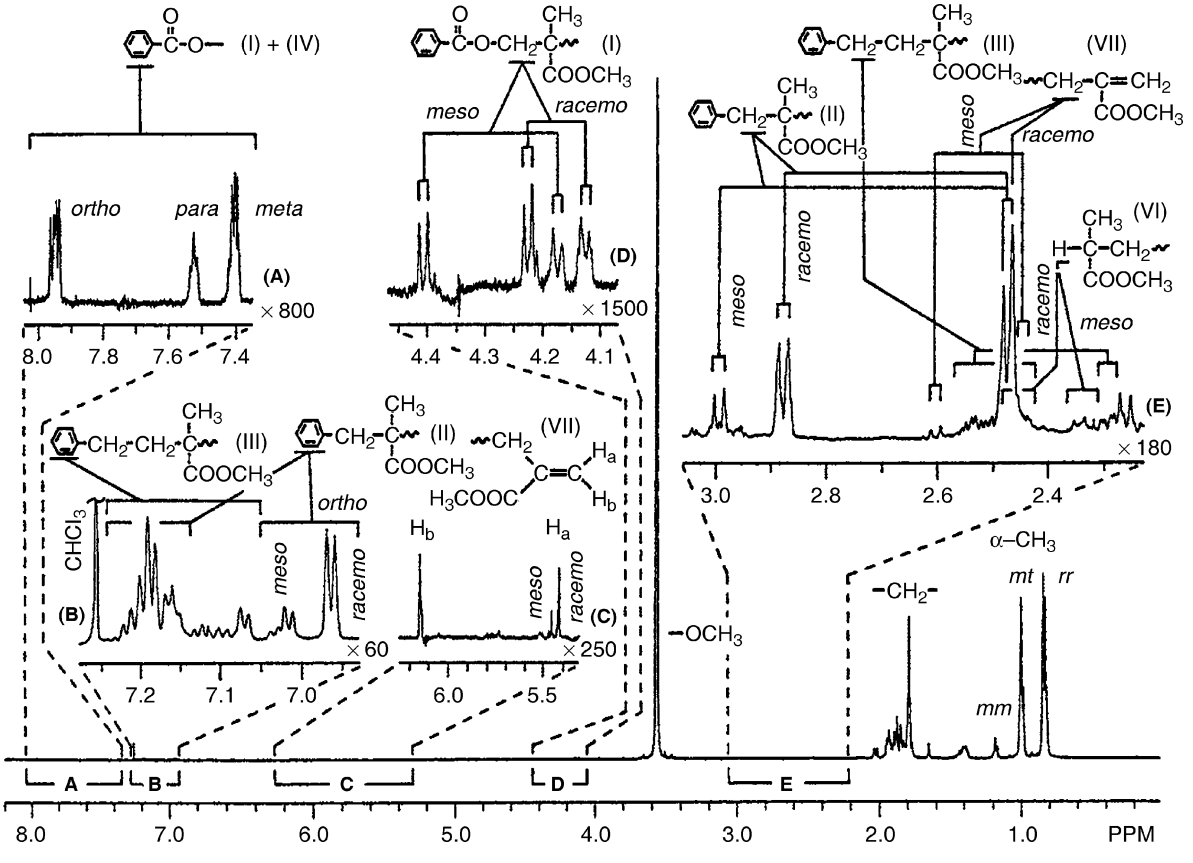


Fig. 13 ^{1}H NMR spectrum of 750 MHz of PMMA prepared with benzoyl peroxide in toluene at 100°C. (From Ref.[27].)

phosphorus (from initiator) at the end of polystyrene chains was exploited by using $^{1}H/^{13}C/^{31}P$ triple resonance multidimensional NMR to produce spectra containing only those resonances from CH_n groups directly bound to ^{31}P at the chain-ends.[28] In Fig. 14A, an HMQC spectrum shows the correlations from both main chain and chain-end structures, whereas a $^{1}H/^{13}C/^{31}P$ triple resonance 2D-NMR spectrum (Fig. 14B) shows only correlations from chain-end structures, simplifying the spectral assignment. The full 3D-NMR spectrum shows resolved resonances from eight different chain-end structures.

A quantitative comparison of end-group resonances with those of the rest of the polymer provides its number average molecular weight (M_n). The relationships between M_n and peak area are generally defined by

$$M_n = \frac{\sum_i (I_i \times m_i)}{I_{CE}} \quad (5)$$

where I_i is the intensity of a unique resonance from each repeat unit in the polymer, m_i the mass of repeat unit i, and I_{CE} the intensity of a chain-end resonance. When possible, the average of several resonances for each monomer unit and for the chain-end structure can provide a more reliable measure of M_n than that obtained from the integration of one peak for each structure fragment. The mobility of the chain-end atoms is usually significantly greater than the mobilities of the atoms in the polymer backbone, leading to

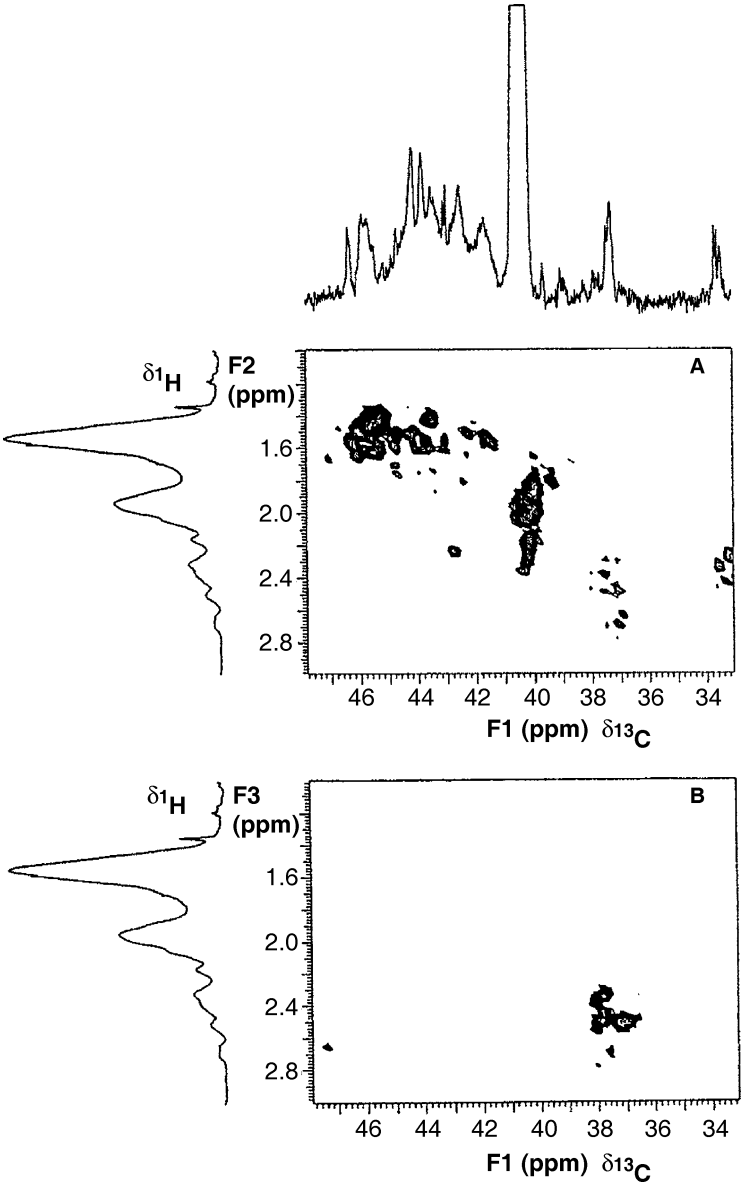


Fig. 14 (A) Aliphatic region from the 2D HMQC NMR spectrum of polystyrene (*Mn* Å 3000, %P Å 1.55) containing diphenylphosphinyl chain-ends, showing the 1D ^{1}H and ^{13}C spectra along the axes; (B) corresponding region from the projection of *f2f3* planes from the 3D PFG-$^{1}H^{13}C^{31}P$-NMR spectrum. (From Ref.[28].)

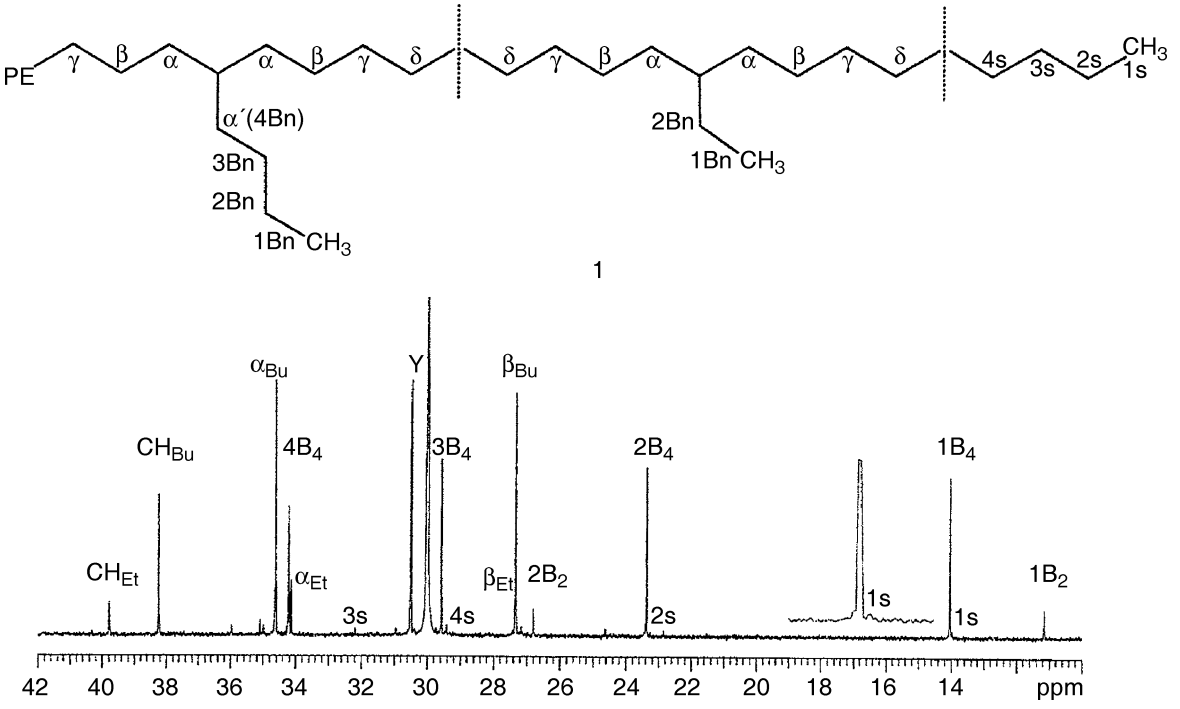


Fig. 15 ^{13}C NMR spectrum of 188.6 MHz of polyethylene containing ethyl and butyl branches. (From Ref.[31].)

significantly different T_1 relaxation times for the nuclei in the former. Consequently, errors from inadequate relaxation delays are greater than those introduced in determining monomer and stereosequence distribution, where the atoms in the main chain repeat units are likely to have similar T_1s. This is particularly true when the areas of methyl resonances are compared with polymer backbone resonances; relaxation delays governed by the longest T_1s must be used.

Branching Structures

The presence of a small number of defects in a polymer chain can have a large influence on a polymer's properties. Two important aspects of polymer chemistry involve detection and structure identification of chain branching and junctions between segments in block copolymers, as low levels of these structures can have a large effect on a polymer's crystallinity,

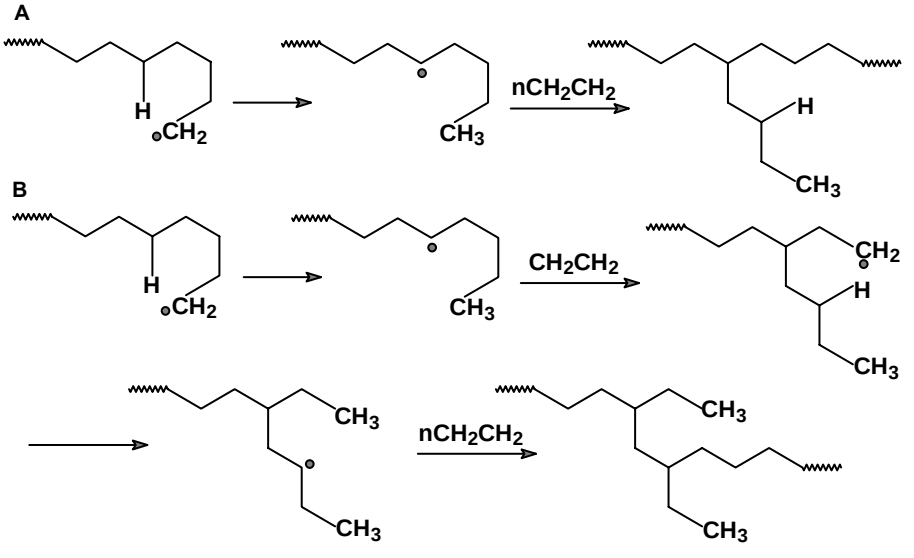


Fig. 16 Typical short-chain branching mechanisms for ethylene homopolymerization: (A) formation of a butyl branch; (B) formation of paired ethyl branches. (From Ref.[29].)

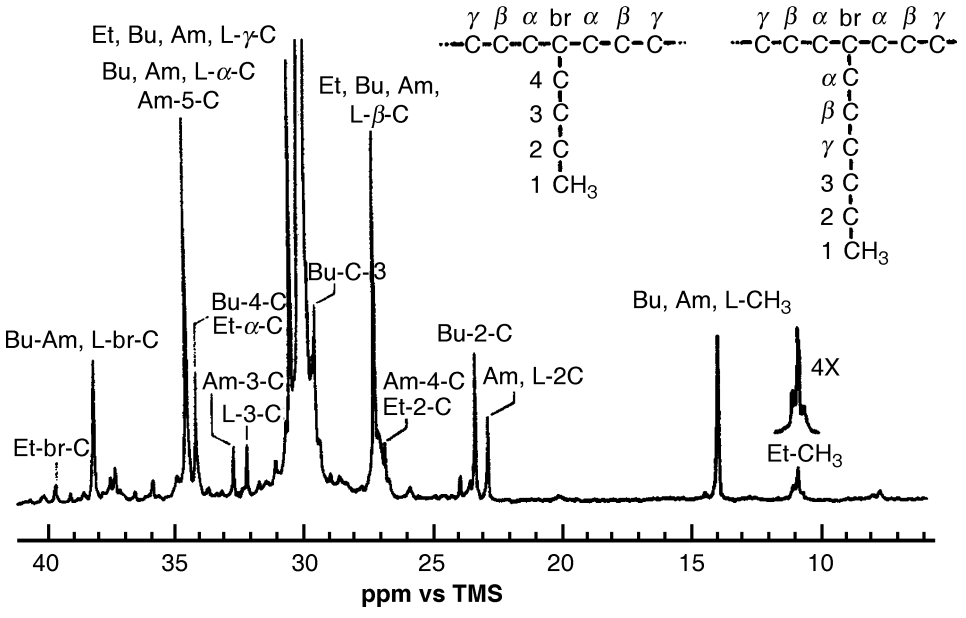


Fig. 17 ^{13}C NMR spectrum of 25 MHz of low-density polyethylene. (From Ref.[4].)

density, and thermal, rheological, and mechanical properties. As in the study of chain-end structures, branch points can be present in very low concentration and can be hard to detect. An excellent example is in the spectrum of polyethylene (Fig. 15), where branching arises from rearrangements and/or hydrogen migration during the polymerization[29,30] or by deliberate introduction of α-olefin comonomers.[26,31]

Fig. 15 shows a ^{13}C NMR spectrum of a polyethylene containing ethyl and butyl branches.[31] The peaks are labeled with the resonance assignments referring to the carbon labels on the structure. The intense peak at 30.0 ppm is from the major component, long sequences of methylene groups ($\delta^+\delta^+$ methylenes) and is plotted far off scale. Three additional sets of resonances are observed for the structures formed at chain-ends (ns), ethyl branches (CH_{Et}, nB_2, α_{Et}, β_{Et}), and butyl branches (CH_{Bu}, nB_4, α_{Bu}, β_{Bu}). In this case, the concentrations of branch structures are low, so that these structures are isolated from one another. Quantitative results are typically reported as the number of branches per 1000 carbons (or number of branches per 10,000 carbons) in the polymer. This can be calculated from the ratio of the integral from one peak of the branch

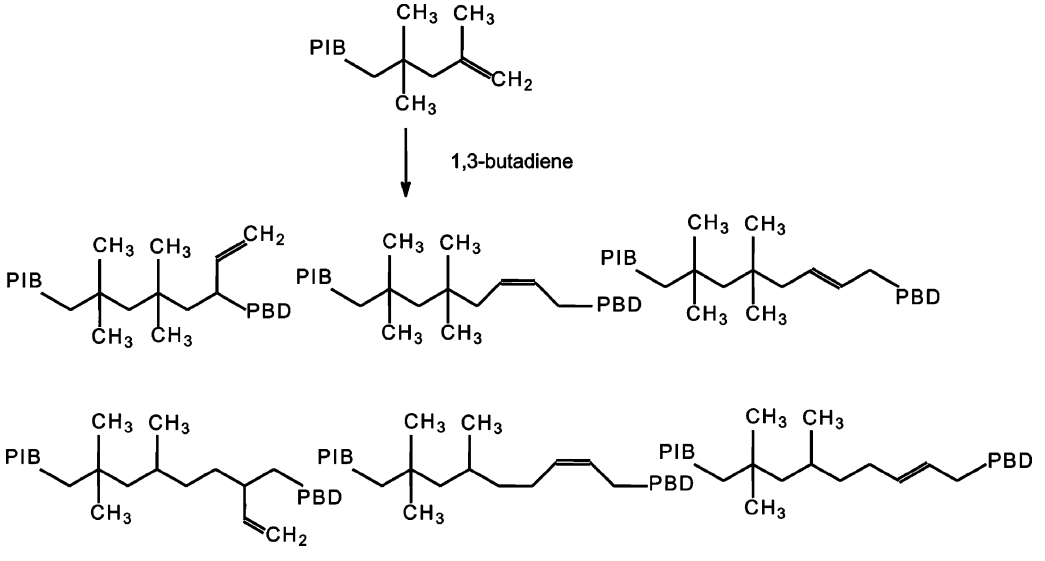


Fig. 18 Reaction of PIB chain-end to form PIB-*b*-PBD copolymer.

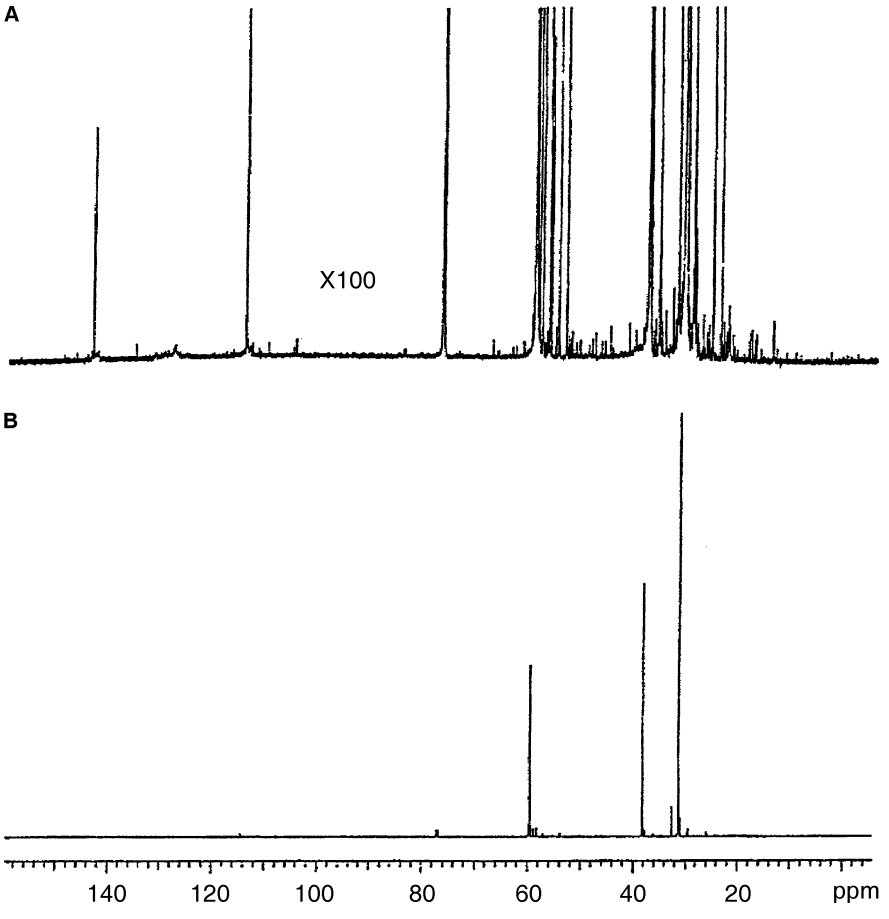


Fig. 19 ^{13}C NMR spectrum of 150 MHz of poly(isobutylene-*b*-butadiene): (A) normal plot and (B) plot of the same region with 100 × vertical amplification.

structure divided by the total integrated area of peaks from the polymer. For example, the number of ethyl branches (N_{Et}) per 1000 carbons is

$$N_{Et} \text{ per 1000 carbons} = \frac{1000 \times I_{CH_{Et}}}{\sum_i I_i} \quad (6)$$

where $I_{CH_{Et}}$ is the area under the signal from CH carbons of ethyl branches and the denominator indicates a summation of the areas of all the polymer's peaks.

McCord et al.[29,30] in a series of papers presented ^{13}C and 2D-NMR evidence for short-chain branch (SCB) forming reactions in the case of ethylene and α-olefin homo- and copolymerization. Backbiting (SCB formation) is a type of intramolecular chain transfer in which the growing radical chain-end curls back on itself to form a five-, six- or seven-membered ring intermediates, transferring the radical back along the chain by hydrogen atom abstraction to form short-chain (propyl, butyl, or amyl) branches (Fig. 16). Fig. 17 shows the ^{13}C NMR spectrum of polyethylene synthesized at 270°C indicating the presence of butyl and amyl branches. It was shown that with increasing reaction temperature, the concentration of branches increases.

Block Copolymers

Block copolymers made from monomers A and B contain long sequences of A units connected to long sequences of B units. To understand the block polymerization mechanism, it is useful to measure the number of structures at block junctions. The issues associated with the study of block copolymers are similar to those discussed earlier in relation to chain-end and branching analysis; the resonances from block junctions are small compared with those of the main chain peaks.

Data from poly(isobutylene-*b*-butadiene) (PIB-*b*-PBD) are an excellent illustration of the characterization of block junction structures.[32] This polymer is more complicated than most because the 1,3-butadiene monomers add to PIB at the PIB–PBD junction in one of three ways (illustrated in Fig. 18): by 1,2-addition to form vinyl (V) units, and by 1,4-addition to form *cis*-1,4-units (C) or *trans*-1,4-units (T). Six structures are possible for the block junction if one considers that addition of V, C, or T units occurs at either of two sites on the terminal isobutylene unit. Because additional butadiene units can add to form C, T, or V units in the second position from the block junction, the number of possible structures increases dramatically.

The ^{13}C NMR spectrum of PIB–PBD, shown in Fig. 19A, is dominated by the resonances of the PIB portion of the polymer. In the spectrum plotted with 100 × vertical expansion (Fig. 19B), a large number of weak signals are detected from the numerous chain-end and block junction structures. In the aliphatic region (10.0–60.0 ppm), the signals are so numerous that they appear to be noise. Resonance assignments obtained by chemical shift predictions or by comparison with the shifts of model compounds are ambiguous in this case, as many signals appear in a small region. Multidimensional NMR experiments were used to obtain unequivocal assignments before these peaks were used to obtain useful information about polymer's structure.

CONCLUSIONS

Nuclear magnetic resonance has become the premier technique for characterizing polymer structure dynamics and interactions in solution. Advances in multidimensional NMR and gradient spectroscopy along with labeling methods continue to enhance applications in the characterizing complex polymers. It is now routinely used in all aspects of fundamental and applied polymer work. Although the instrumentation is somewhat expensive, it is often the only way to obtain necessary analyses. In process monitoring and quality control environments requiring high sample volume and rapid throughput, NMR is used to obtain primary calibration standards for use with other less expensive and more rapid instrumental analysis techniques.

REFERENCES

1. Cheng, H.N.; Early, T.A. NMR studies of polymeric materials—an overview. Macromol. Symp. **1994**, *86*, 1–14.
2. Cheng, H.N.; English, A.D. Advances in the NMR spectroscopy of polymers: an overview. ACS Symp. Ser. **2003**, *834*, 3–20.
3. Grant, D.M.; Harris, R.K. *Encyclopedia of NMR*; John Wiley: New York, 1996.
4. Bovey, F.A.; Mirau, P.A. *NMR of Polymers*; Academic Press: New York, 1996.
5. Koenig, J.L. *Spectroscopy of Polymers*; 2nd Ed.; Elsevier, 1999.
6. Cheng, H.N. Structural studies of polymers by solution NMR. Rapra Rev. Rep. **2001**, *11* (5), 1–156.
7. Breitmaier, E.; Voelter, W. *Carbon-13 NMR Spectroscopy*, 3rd Ed.; VCH: Weinheim, 1987.
8. Suter, U.W.; Flory, P.J. Conformational energy and configurational statistics of polypropylene. Macromolecules **1975**, *8*, 765–777.
9. Busico, V.; Cipullo, R. Microstructure of polypropylene. Prog. Polym. Sci. **2001**, *26*, 443–533.
10. Moad, G. Applications of labeling and multidimensional NMR in the characterization of synthetic polymers. Annu. Rep. NMR Spectrosc. **1992**, *29*, 287–323.
11. Cheng, H.N. ^{13}C NMR sequence determination and modeling of polypropylene oils. Macromol. Symp. **1994**, *86*, 77–102.
12. Cheng, H.N. Computer-aided NMR methods for polymers. Polym. News **2000**, *25*, 114–122.
13. Bovey, F.A.; Mirau, P.A. The two-dimensional NMR spectroscopy of macromolecules. Makromol. Chem. Macromol. Symp. **1990**, *34*, 1–16.
14. Sahoo, S.K.; Zhang, T.; Reddy, D.V.; Rinaldi, P.L.; McIntosh, L.H.; Quirk, R.P. Multidimensional NMR studies of poly(ethylene-co-butene) microstructure. Macromolecules **2003**, *36*, 4017–4028.
15. Asakura, T.; Nakayama, N.; Demura, M.; Asano, A. Carbon-13 NMR spectral assignments of regioirregular polypropylene determined from two-dimensional INADEQUATE spectra and chemical shift calculations. Macromolecules **1992**, *25*, 4876–4881.
16. Rinaldi, P.L. Polymer characterization by 3D NMR. ACS Symp. Ser. **2003**, *834*, 94–122.
17. Monwar, M.; Sahoo, S.K.; Rinaldi, P.L.; McCord, E.F.; Marshall, D.R.; Buback, M.; Latz, H. A 3D-NMR method for studying hydrocarbon-based polymer structures. Macromolecules **2003**, *36* (18), 6695–6697.
18. Hurd, R.E. Gradient-enhanced spectroscopy. J. Magn. Reson. **1990**, *87* (2), 422–428.
19. Morris, G.A. Diffusion-ordered spectroscopy (DOSY). Encl. Nucl. Magn. Reson. **2002**, *9*, 35–44.
20. Klaus, A. *On-line LC-NMR and Related Techniques*; John Wiley: New York, 2002.
21. Ute, K.; Niimi, R.; Matsunaga, M.; Hatada, K.; Kitayama, T. On-line SEC-NMR analysis of the stereocomplex of uniform isotactic and uniform syndiotactic poly(methyl methacrylate)s. Macromol. Chem. Phys. **2001**, *202*, 3081–3086.
22. Hatada, K.; Ute, K.; Yamamoto, M.; Nishimura, T.; Kashiyama, M. On-line GPC/NMR analyses of block and random copolymers of methyl and butyl methacrylates prepared with tert-butylmagnesium bromide. Polym. Bull. **1989**, *21*, 489–495.
23. Chen, T.-A.; Rieke, R.D. The first regioregular head-to-tail poly(3-hexylthiophene-2,5-diyl) and a regiorandom isopolymer: Ni vs Pd catalysis of 2(5)-bromo-5(2)-(bromozincio)-3-hexylthiophene polymerization. J. Am. Chem. Soc. **1992**, *114*, 10087–10088.
24. Li, L.; Rinaldi, P.L. Characterization of poly (1-chloro-1-fluoroethylene) fluoropolymer using

$^{1}H/^{13}C/^{19}F$ triple resonance 3D-NMR. Macromolecules **1997**, *30*, 520–525.
25. Dutch, W.M.; Grant, D.M. Carbon-13 chemical shift studies of the 1,4-polybutadienes and the 1,4-polyisoprenes. Macromolecules **1970**, *3*, 165–174.
26. Randall, J.C. A review of high-resolution liquid carbon-13 nuclear magnetic resonance characterizations of ethylene-based polymers. J. Mac. Sci., Rev. Macromol. Chem. Phys. **1989**, *C29* (2, 3), 201–317.
27. Hatada, K.; Kitayama, T.; Ute, K.; Terawaki, Y.; Yanagida, T. End-group analysis of poly(methyl methacrylate) prepared with benzoyl peroxide by 750 MHz high-resolution ^{1}H NMR spectroscopy. Macromolecules **1997**, *30*, 6754–6759.
28. Saito, T.; Rinaldi, P.L. Evaluation of 3D NMR experiments for the characterization of polymer structure. J. Magn. Reson. **1998**, *132*, 41–53.
29. McCord, E.F.; Shaw, W.H., Jr.; Hutchinson, R.A. Short-chain branching structures in ethylene copolymers prepared by high-pressure free-radical polymerization: an NMR analysis. Macromolecules **1997**, *30*, 246–256.
30. McCord, E.F.; McLain, S.J.; Nelson, L.T.J.; Arthur, S.D.; Coughlin, E.B.; Ittel, S.D.; Johnson, L.K.; Tempel, D.; Killian, C.M.; Brookhart, M. ^{13}C and 2D NMR analysis of propylene polymers made with α-diimine late metal catalysts. Macromolecules **2001**, *34*, 362–371.
31. Liu, W.; Ray, D.G., III; Rinaldi, P.L.; Zens, T. High-temperature pulsed-field-gradient multidimensional NMR of polymers. J. Magn. Reson. **1999**, *140*, 482–486.
32. Tokles, M.; Keifer, P.A.; Rinaldi, P.L. Characterization of polyisobutylene–polybutadiene diblocks by HMQC and HMBC. Macromolecules **1995**, *28*, 3944–3952.

NO_x Removal

Mike Bradford
Rajiv Grover
Jacobs Engineering Group Inc., Houston, Texas, U.S.A.

INTRODUCTION

Nitrogen oxides (NO_x) are a combination of nitrogen oxide (NO) and nitrogen dioxide (NO_2). Nitrogen oxides can combine with volatile organic compounds (VOCs) to form ozone. For this reason, NO_x control has become a target of the environmental regulatory authorities. The minimization or reduction of NO_x is required because of ever-tightening regulations and permitting requirements.

The emission control requirements of the existing as well as future regulations require an understanding of the NO_x-emitting sources, pollution prevention opportunities, emission reduction technologies, costs of controls, and impacts on existing operations and plant safety. This entry discusses a number of aspects of NO_x control:

- NO_x reductions through process changes and/or pollution prevention activities
- ultra low NO_x burners (ULNB)
- burner-related NO_x reduction, such as flue gas recirculation (FGR)
- postcombustion NO_x control options, such as selective noncatalytic reduction (SNCR), selective catalytic reduction (SCR), and catalytic oxidation/scrubbing.

When evaluating these technologies, it should be kept in mind that in many cases, retrofitting the NO_x controls can have positive side benefits. In many instances, the installation of NO_x controls can result in better energy efficiency and in lower maintenance requirements.

TECHNOLOGY OVERVIEW

Table 1[1,2] lists various technologies available for NO_x reduction. It also lists the estimated percentage of reduction using each technology and the approximate ppmvd NO_x that can be achieved. The table is divided into two sections: one for fired heaters and the other for boilers, because NO_x generation is very different for the two types of systems.

FIRED HEATER NO_x REDUCTION

Process Modifications and/or Pollution Prevention NO_x Reduction

A number of process modifications and/or energy efficiency improvements can be made to reduce NO_x emissions.[1,2] Also, some of the standard NO_x reduction technologies have the additional benefit of improving energy efficiency. Possible process modifications include:

- Replacement of boilers with a cogeneration system.
- Replacement of inefficient heaters with more efficient heaters.
- Provision of a new convection section or upgradation of the existing one. These modifications can be used to preheat the process fluid, thereby improving the heater efficiency and decreasing the firing rate. The lower firing rate results in less NO_x generation, and reduction in operating cost could pay for the modification.

  It must be noted, however, that modification to the convection section is often required to provide the correct temperature for installation of an SCR. An increase in energy efficiency may be an additional benefit of installing an SCR.
- Upgradation of the heater controls to allow tighter control of the excess air. Tighter controls could be used to reduce the oxygen concentration in the flue gas. A reduction in the O_2 concentration in the flue gas from 6% to 2% reduces the firing rate by only few percentages. But this still has two advantages.

  i. The lower firing rate decreases the NO_x generation a little
  ii. Less NO_x is generated if the oxygen concentration is low. In general, a decrease in O_2 concentration from 6% to 2% reduces the NO_x generation by approximately 30%.

  Combining the above two benefits results in a NO_x reduction of a little over 30%.
- Sealing the heater can also reduce the O_2 concentration in the firebox and improve the heater

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007882

Table 1 Potential NO_x reduction technologies

Technology	Approximate reduction (%)	Approximate lbs/MM Btu	Approximate ppmv at 3% O_2
Fired heaters			
Process modifications/pollution prevention			
Standard burners	Base case	0.14	120
Low NO_x burners	60%	0.06	50
ULNB–first generation	80%	0.025–0.06	20–50
ULNB – current best demonstrated	90%	0.012–0.05	10–40
FGR	30%	(1)	(1)
Boilers			
Standard burners	Base case	0.25	200
ULNB–first generation (w/o FGR)	70%	0.08	65
ULNB–first generation (w/o FGR)	84%	0.04	35
ULNB – current best demonstrated (w FGR)	96%	0.01	8
FGR	55–90% (2)	(1)	(1)
Miscellaneous			
SNCR	40–90%	(1)	(1)
Catalytic scrubbing	70–90%	(1)	(1)
SCR	90–97% (3)	(1)	$\geq$5 ppmvd

Notes: (1) Final NO_x level depends on the starting point; (2) Depends on % of FGR and level of air preheat; (3) >90% reduction requires special attention to NH_3 distribution.

efficiency. Like the upgradation in controls, this can be an additional advantage of the installation of ULNB.

Ultra Low NO_x Burners

Burners have been undergoing a rapid state of development based on pressures to reduce NO_x formation and stages of development include:

- Standard burners that were designed more for fuel economy and maintenance ease than for NO_x control.
- Low NO_x burners (LNB) specifically designed for lower NO_x emissions. These burners typically utilized air staging to reduce NO_x.
- Ultra low NO_x burners that typically had a combination of air staging and fuel staging. The latest version of ULNB includes new, more experimental improvements.

Impacts on Standard Burner NO_x Generation

Nitrogen oxides are formed in fired systems in three ways:

- Thermal NO_x, formed because of local hot spots in the flame.
- Prompt NO_x, formed because of reactions at the fuel–oxygen interface.
- Fuel NO_x, resulting from the burning of fuels containing organic nitrogen. Some of the organic nitrogen is oxidized into NO_x. This type of NO_x formation tends to be negligible with natural gas and fuel gas firing.

Standard burners typically result in NO_x formation in the 0.12–0.14 lb/MM Btu (100–120 ppmv) for fired heaters and the 0.2–0.3 range for boilers. Factors that influence this number are:

- Excess air: the lower the excess air (%O_2), the lower the NO_x (Fig. 1).[3]
- Firebox temperature: high firebox temperatures result in higher NO_x formation (Fig. 2).[3]
- Air preheat: higher air preheat results in higher NO_x (Fig. 3).[3] Note: This figure only applies to standard burners. Air preheat does not have as great an effect on ULNB.

Ultra Low NO_x Burners for Fired Heaters

While this section discusses the ULNB used for fired heaters only, boiler burners are discussed in the subsequent section. These two types of fired units are discussed separately because of their different firing characteristics.

Currently, there are two recognized generations of ULNB. The first generation, under ideal conditions, has demonstrated NO_x levels as low as 0.025 lb/MM Btu.

Burners that have achieved NO_x levels as low as 0.012 lb/MM Btu are currently under development, but their performance is highly variable because it is

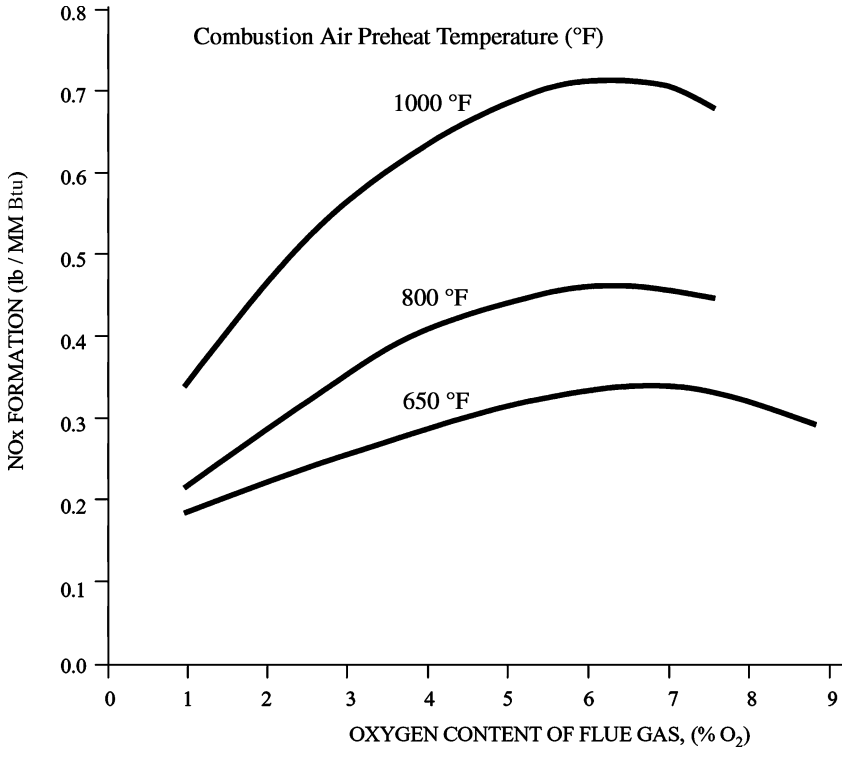


Fig. 1 Stack gas O_2 concentration vs. NO_x. (From Ref.[3].)

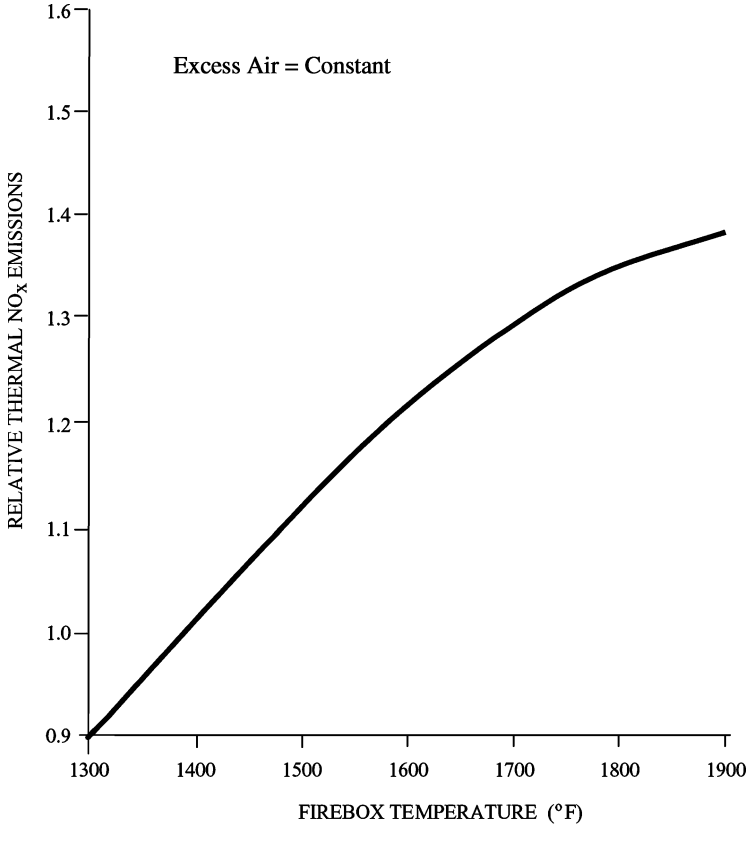


Fig. 2 Firebox temperature vs. NO_x formation. (From Ref.[3].)

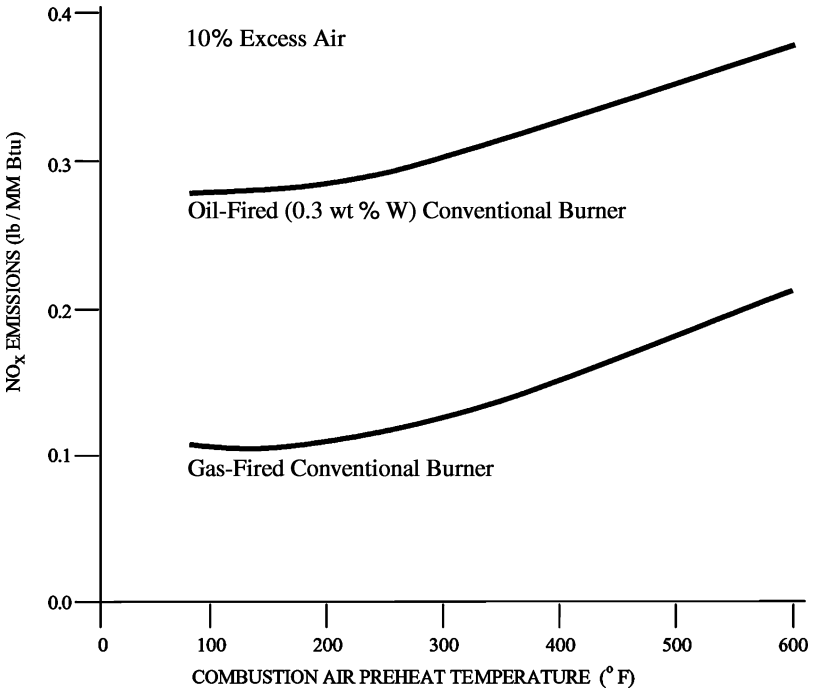


Fig. 3 Impact of air preheat on NO_x formation. (From Ref.[3].)

highly dependent on flux levels and flow patterns within the heaters.

Ultra low NO_x burner's performance

Burner performances are dependent on the following:

- No air leaks: Most of the ULNB use internal FGR to dilute the flame. Air leaks result in a high oxygen level in the flue gas that is recirculated and result in higher NO_x levels. Leaks into the furnace may need to be welded shut, and gaskets may need to be installed to stop air leaks around view ports.
- Firebox temperatures: Low NO_x guarantees are easier to obtain if firebox temperatures are low.
- Fuel gas: Fuel gas systems containing hydrogen cause problems with some burners, because hydrogen burns hot and also owing to swings in the hydrogen concentration. Some vendors derate their burners in fuel gas systems.

More advance control systems may be required to ensure the performance of the burners. Control modifications that may be required include:

- Installation of continuous emission monitoring systems (CEMS).
- Use of oxygen analyzers near the bridgewall and/or in the floor of the heater.
- Automatic controls on stack dampers. The stack damper is used to control the heater draft. This in turn approximately controls the O_2 concentration in the firebox.
- Automatic modulation of the burner inlet air registers. This is used in conjunction with the stack damper control. The stack damper controls the draft and the O_2 concentration in the firebox controls the inlet air registers.
- CO trim control. This has proven to be the optimum control scheme. An ''optical through-the-stack'' light beam is used to measure the CO. This CO control is used in place of oxygen control.
- Upgrades to the burner management system (BMS). This may be triggered by the above changes to the control systems.
- The installation of coalescers/filters to remove liquid and particulates from the fuel gas. The ULNB that depend on inducing internal FGR have small fuel tips. These fuel tips are very susceptible to plugging if the fuel gas is not clean.

Additional potential restrictions on the use of ULNB are:

- Size: The latest generation of ULNB is significantly larger than standard burners.
- Minimum spacing: ULNB depend on FGR, inside the firebox, quenching the flame temperature. Recent tests have demonstrated that if burner spacing is too tight, the flue gas cannot recirculate

inside the burner circle, and NO_x generation goes up.
- Computational fluid dynamics (CFD) computer modeling can be used to flag problem heaters.

Vendors are doing a tremendous amount of research to develop burners that will achieve very low NO_x levels. As this is an evolving technology, the odds are that the burners we install next year will be different from those we install this year.

Ultra Low NO_x Boiler Burners

Boilers are higher NO_x producers than fired heaters because the firebox temperatures are higher and they typically have air preheat systems. Firebox temperatures typically run in the 2100°F range, at full loads.

Historically, FGR has been used extensively to minimize boiler NO_x. The reasons that FGR have been used so extensively are:

- Historically, there has been no other way to achieve low NO_x levels because of the high firebox fluxes/firebox temperatures.
- Boilers normally are forced draft (FD) versus most fired heaters that are natural draft. It is cheaper to combine FGR with an FD system because the FD fan can be used to induce the flue gas and the plenums already exist to distribute the air.
- Boilers frequently have air preheat systems. The induced draft (ID) fans used for air preheat can be used to recirculate the flue gas.

In addition to FGR, LNB boiler burners have used air staging to minimize NO_x formation. ULNB boiler burners also use fuel staging and/or a combination of air and fuel staging.

A number of approaches to achieve the quenching effect of fuel gas recirculation can be used:

- The conventional FGR from the stack to the air-side of the burner
- Steam or water injection into the air-side of the burner
- Induced FGR from the stack into the fuel gas
- Steam injection into the fuel gas

Fig. 4 provides information on the impact of FGR on NO_x reduction. Boiler NO_x levels that have been achieved are:

- First generation ULNB

Without FGR	0.08 lb/MM Btu
With FGR	0.04
With FGR and water injection (air-side)	0.033

- Latest generation burners

Without FGR (natural gas)	0.04 lb/MM Btu
Without FGR (fuel gas)	0.03 (one installation)
With FGR (natural gas)	0.01
With induced FGR into the fuel gas (IFGR) and steam (fuel gas)	0.01 (low flux boiler)

Coalescers/filters may or may not be required to clean up the fuel gas. Burners that utilize FGR as the fuel gas have larger than normal fuel tips, and thus are less susceptible to plugging than standard burner

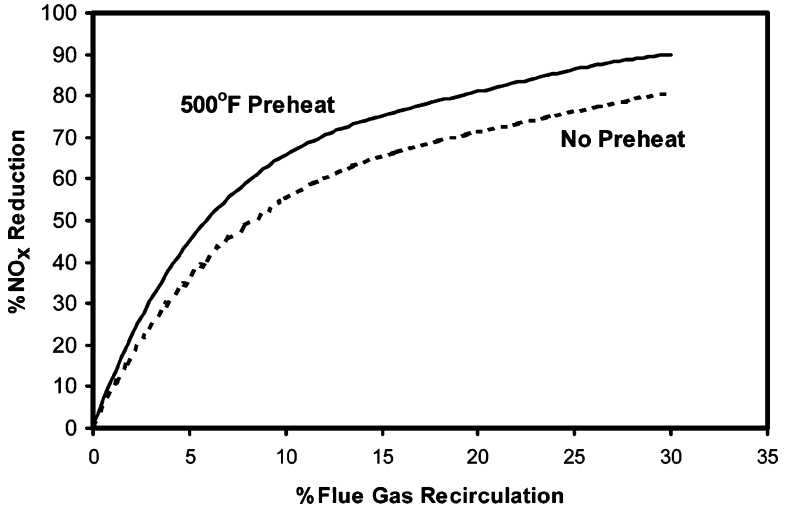


Fig. 4 Impact of FGR on NO_x reduction.

tips. In general, boiler burners do not depend on inducing FGR within the firebox, they do not have the small diameter burner tips that fired heater ULNB use. So, they also are no more susceptible to fouling than the standard burners.

Burner-Related NO_x Reduction

Conventional FGR

Historically, recirculation rates in the 15–20% neighborhood have been used, with NO_x reductions in the 40–55% range. The maximum stoichiometric recirculation possible is 40%. The latest regulations have been forcing recirculation close to 30% for NO_x reductions as high as 70%, but these systems require very tight control systems to maintain flame stability.

Fig. 4 shows the NO_x reduction that can be achieved for various FGR rates. It must be noted that

- FGR has a greater impact if the system has an air preheater.
- The greatest benefits are achieved with the first 5–10% FGR. Beyond 10%, the returns diminish. However, to achieve very low NO_x numbers, the vendors are pushing the systems all the way out to the 30% FGR range.
- The curves are for boilers only. Because of their high firebox temperatures, boilers tend to have high NO_x and benefit more from FGR. Estimates are that fired heaters can achieve a maximum of about 30% NO_x reduction from FGR.

FGR design details include:

- An FD system is required, which is one reason this technology has been more popular with boilers than with fired heaters. Boilers are typically FD while fired heaters are typically natural draft.
- An eductor in the FD air can be used on new installations to educt in the recirculated air. For retrofits, the FD fan may not have enough capacity for an eductor, so a recirculating fan may be required.
- The burners may need to be retrofitted with ''bussels'' for introduction of the recirculated air.

Alternates to the conventional FGR are the educting of flue gas into the fuel gas and/or steam injection into the fuel gas.

Advantages of FGR include:

- Being less expensive than an SCR.
- Perhaps the only way to achieve low NO_x on boilers, because of the high firebox temperatures and the potential impact from air preheat.
- Use in combination with ULNB to achieve very low NO_x concentrations, especially in boilers.
- Use in recirculating flue gas in the fuel gas to reduce the heat of combustion of the fuel. The reduced heat of combustion reduces the NO_x generation.

Disadvantages of FGR include:

- Higher investment and higher operating costs than the installation of ULNB.
- Need to increase the capacity of the FD fan.

Steam or Water Injection (Air-Side of Burner)

Steam or water injection into the air-side of the burner can be used to quench the flame. Unless free steam is available, water injection is more economical.

The disadvantage of water injection is that it introduces particulates into the flue gas from the dissolved salts in the water. These particulates can deposit on the tubes or result in an increase in particulate emissions from the boiler. Boiler feedwater has been used to minimize this, although it is not required by the burner vendors.

Fig. 5 provides information on NO_x reduction versus steam injection rates. Water injection can achieve the same NO_x reductions, but with lower injection rates.

IFGR

This technology was developed by one of the vendors based on observations that low-Btu fuel gases produce low NO_x. Inducing the flue gas into the fuel gas dilutes the fuel gas. This dilution has a greater impact than FGR into the air-side of the burner because it reduces both the flame temperature and localized hot spots. A general rule of thumb is that 15% IFGR will produce the same NO_x reduction as 30% FGR.

A fan is not required. Instead, an eductor in the fuel gas is used to educt the recirculated flue gas. Because of the higher fuel gas flow, larger fuel tips are required in the burner.

This can be a very economical technology, as long as extensive internal modifications are not required inside the boiler and the existing FD fan can handle the extra pressure drop.

Steam Injection into the Fuel Gas

This is the same principle as IFGR—the steam dilutes the fuel gas thus reducing both the flame temperature and the localized hot spots. Systems have used up to 8% steam injection (the steam injection rate is 8% of the total steam production from the boiler).

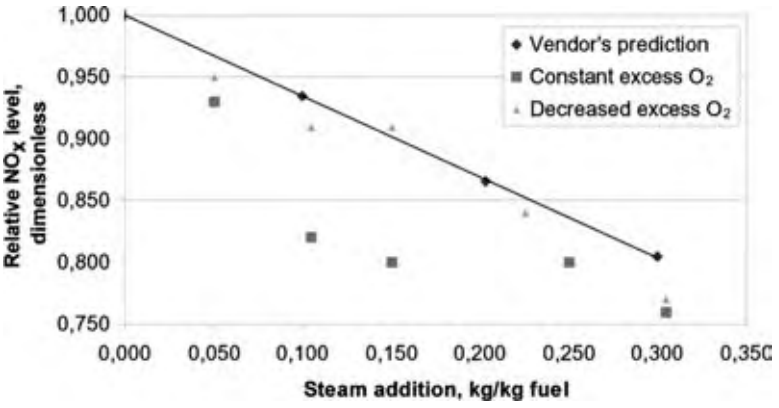


Fig. 5 NO_x reduction vs. steam injection to air-side of burner. *(View this art in color at www.dekker.com.)*

The main drawback of steam injection is the potentially high operating cost. If 8% steam injection is required, in general it would be more economical to install an ID fan and use FGR into the air-side—the higher investment would offset the operating cost.

An ideal application of steam injection would be in cases where the steam is normally not needed. For example, one plant normally does not require steam, but adds it when needed to offset NO_x production in other parts of the refinery. As the addition frequency is low, operating costs are low and there was no requirement for FGR fans.

SNCR

SNCR units can achieve approximately a 40% NO_x reduction without requiring the installation of an SCR. Combined with other technologies, SNCR can potentially produce low NO_x concentrations. A number of vendors license SNCR technologies.

SNCRs utilize the reducing capability of ammonia and urea to reduce NO_x to nitrogen. The requirements for good conversion include:

- Proper temperature. The reactions occur over a very narrow temperature range; however, additives can be used to extend the range.

Ammonia reaction	1600–2200°F (with versions down to 1400°F)
Urea reaction	1600–2000°F (with versions down to 1200°F)

- Good mixing. The ammonia is typically injected into holes drilled into the firebox. These injectors can be installed on-line.
- Adequate residence time.
- No impingement of the injected chemical against tubes.

The equations for the reaction of ammonia with NO_x are:

$$4NO + 4NH_3 + O_2 = 4N_2 + 6H_2O \quad (NH_3/NO = 1/1)$$

$$2NO + 4NH_3 + 2O_2 = 3N_2 + 6H_2O \quad (NH_3/NO = 2/1)$$

$$6NO_2 + 8NH_3 = 7N_2 + 12H_2O$$

For temperatures between 1600°F and 2200°F, the top and bottom equations dominate.

Most of the NO_x is NO (90–93% by volume); hence, the theoretical NH_3/NO_x ratio is about 1.03, plus the NH_3 slip.

The equation for the reaction of urea is:

$$CO(NH_2)_2 + 2NO + 1/2O_2 = 2N_2 + CO_2 + 2H_2O$$

Advantages of SNCR include:

- Lower capital investment than SCRs.
- Can be used in dirty/fouling services (particulates and/or high sulfur). However, any SO_3 can react with the NH_3 and form ammonium bisulfate precipitates (see the section on SCR technology).
- Can be combined with other technologies, such as catalytic scrubbing or partial burn regenerators, to achieve greater NO_x reduction.
- SNCR may be economical in conjunction with an SCR. The SNCR can be used to reduce the required size of the SCR by reducing the inlet NO_x concentration to the SCR.

Disadvantages of SNCR include:

- Operates only in a narrow temperature range.
- The residence time may not be long enough for the SNCR reaction to occur.

- High ammonia consumption/ammonia slip. An excess of 20–50% is required to achieve a 40% NO_x reduction. By comparison, SCRs require < 5% excess ammonia. Because of the higher ammonia slip, operating costs could be higher than SCRs.

SCRs

SCRs utilize the following equations for the reaction between ammonia and NO_x at high temperatures:

$$4NO + 4NH_3 + O_2 = 4N_2 + 6H_2O$$
$$2NO + 4NH_3 + 2O_2 = 3N_2 + 6H_2O$$
(minor at standard temperatures)
$$6NO_2 + 8NH_3 = 7N_2 + 12H_2O$$

This is basically the same reaction that occurs with SNCR. The difference is that the catalyst allows the reaction to proceed at lower temperatures.

The NO_x is primarily NO (90–93% by volume). So, the theoretical NH_3 required is approximately 1.03 NH_3/NO_x plus the slip. The slip is the unreacted NH_3 that exits the stack.

There are three basic types of SCR technology:

- Low temperature catalysts: These catalysts have the advantage of operating at lower temperatures, but are susceptible to sulfur and particulates.
- Medium temperature catalysts.
- High temperature catalysts: These tend to use zeolites.

In addition to the choices between the types of catalysts, the following are also required in designing SCR systems:

- Catalyst volume (space velocity) versus percent conversion. Modules can be mounted in series for greater conversion.
- Provisions for excess catalyst to allow long runs between unit turnarounds.
- Where does one locate the SCR in the flue gas train? There is some flexibility in operating temperature, but in general the higher the temperature, the lower the catalyst volume.
- Horizontal or vertical flow. Vertical flow is most common because it occupies less space. Vertical down flow is preferred in many applications because it allows particulates to drop through the catalyst.
- The design of any downstream heat recovery equipment. Any SO_3 in the flue gas will react with the NH_3 added in the SCR to form ammonium sulfate salts (ammonium sulfate and ammonium bisulfate). These salts form a solid at high temperatures and tend to precipitate on heat transfer surfaces. Note: phase diagrams are available that predict the solid dew points. A more detailed discussion of ammonium salt deposit concerns is included later in this section.

Advantages of SCRs include:

- A high percentage of NO_x removal. Recent systems achieve 80–97% reductions, depending on the design space velocity. The volume of catalyst used can be varied depending on the required percentage of reduction.
- Only approach technically feasible in applications that are difficult for burners.
- Less ammonia slip and a greater NO_x reduction than SNCRs.

Disadvantages of SCRs include:

- The temperature profile in the heat recovery train may not be optimal for the performance of the SCR system. This may require modifications to the heat recovery train and the installation of some type of temperature control.
- Sulfur can precipitate as ammonium bisulfate, thus fouling the catalyst.
- Sulfur can also precipitate as ammonium sulfate on downstream heat transfer surfaces. Any new heat recovery system must be designed to deal with this precipitation. Existing systems must be checked to make sure they can be cleaned without disrupting the process.
- Safety concerns for the storage and handling of ammonia. This can be minimized by using aqueous ammonia, but with an increase in operating cost.
- High capital investment.
- Higher operating costs than some of the other options.
- The increase in pressure drop could require a conversion to FD or the installation of ID fans.
- Low temperature units are particularly susceptible to fouling in high sulfur applications. Plus, some low temperature catalysts are susceptible to particulate plugging.

Low Temperature SCRs

A low temperature installation (GE Frame 3 Gas Turbine) is seen in Fig. 6.[4] A catalyst being loaded is shown in Fig. 7.[4]

Advantages of low temperature SCRs include:

- Can be operated at a lower temperature because it either has a higher surface area or a higher vanadium content for a higher activity

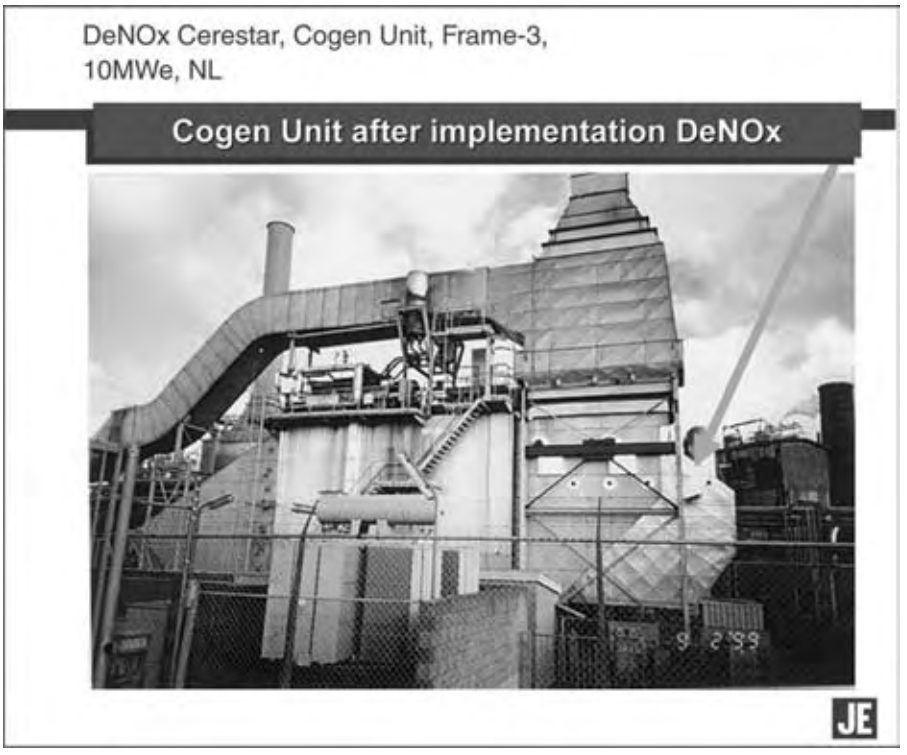


Fig. 6 A low temperature installation. (Courtesy of Cerestar.) *(View this art in color at www.dekker.com.)*

- Can be located at the end of the stack gas train, low temperature area. This minimizes the need to run ducting from a high temperature area and then return the flue gas to the stack gas train. This is its biggest advantage.
- Operating temperature range = 320–680°F (but is not likely to be economical above 500°F).
- Can be used in installations of firing natural gas and many installations of firing plant/refinery fuel gas.

Disadvantages of low temperature SCRs include:

- Susceptible to sulfur concentrations.
- Cost: higher catalyst volumes are required at the lower temperatures.

Low temperature sensitivity to sulfur include:

- The problem is operation below the freezing point of ammonium bisulfate. This results in deposition on the catalyst. At the end of this section is a phase-chart showing the deposition temperature of both ammonium sulfate and ammonium bisulfate, as a function of SO_3 and ammonia concentrations.
- The allowable sulfur level depends on the length of runs between turnarounds. SO_2 levels as high as 100 mg/m^3 can be tolerated with frequent turnarounds and the installation of excess catalyst. However, lower sulfur levels are required for turnaround frequencies of four or five years.

Approaches to minimizing the impacts of sulfur include:

- Increasing the temperature to stay above the dew point.
- Increasing catalyst volume to provide spare surface area for deposition. This has been used successfully in a refinery with an H_2S level in the fuel gas of 50 ppmv.
- Periodic operation at higher temperatures to desublime the deposited ammonium bisulfate. If possible, flow can be stopped to the convection section to raise temperatures. Alternately, duct burners could be installed for periodic operation.

Medium Temperature SCRs

Medium temperature catalysts operate in the 500–725°F temperature range. Three basic types of medium temperature SCRs are:

- Honeycomb units
- Corrugated designs
- Plate

A honeycomb catalyst is shown in Fig. 8. A corrugated unit is similar in design except that the walls are thinner, thus allowing more open area. A plate-type catalyst has long, rectangular openings between the plates.

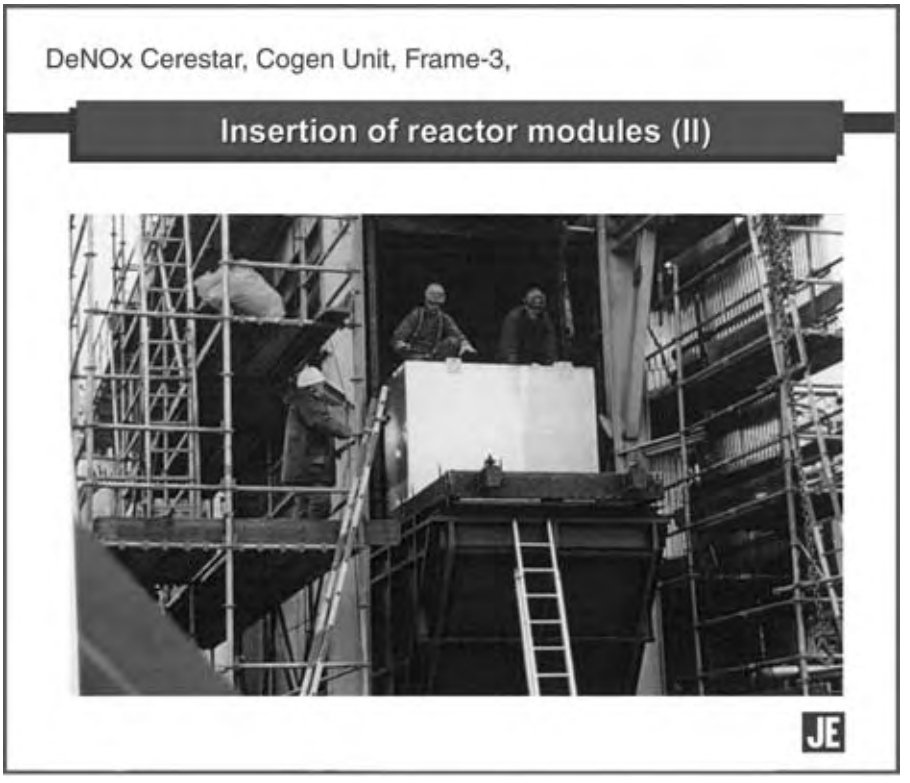


Fig. 7 A catalyst being loaded in a low temperature installation. (Courtesy of Cerestar.) *(View this art in color at www.dekker.com.)*

Advantages of medium temperature catalysts are:

- Honeycomb catalysts have been in use for a number of years.
- Can stand high sulfur and high particulate loadings. A soot blower can be provided for particulate control (the particulates are stirred up and flushed through the honeycomb).

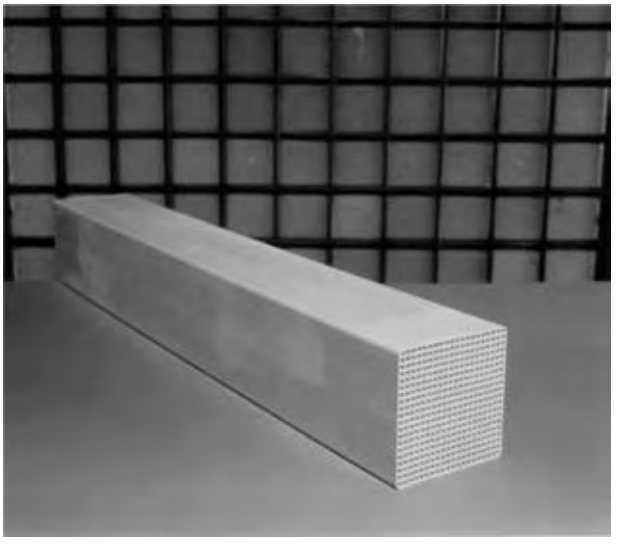


Fig. 8 Honey comb catalyst in front of module. *(View this art in color at www.dekker.com.)*

- Low-vanadium catalysts can be used that minimize the conversion of SO_2 into SO_3. This minimizes the precipitation of ammonium bisulfate on downstream heat recovery units. However, in general these low-vanadium catalysts are not economical because the catalyst volumes are larger (low-vanadium catalysts are less active).

Disadvantages of medium temperature catalysts are:

- The ideal temperature may not exist in the flue gas system. A typical operating temperature range is 550–725°F. This temperature range may not exist in all facilities, and may require modification of the heat recovery train.
- Increased pressure drop. Could require conversion to FD burners, or installation of an ID fan.
- Precipitation can still occur in the catalyst bed because of capillary action that raises the dew point of the ammonium salts. This is negligible for natural gas and refinery fuel gas fired systems but can be significant for applications with high sulfur in the fuel.

High-Temperature SCRs

High-temperature catalysts operate in the 650–1100°F temperature range. These are primarily zeolite based catalysts.

The advantages of high temperature catalysts are:

- Can be used in applications that have little or no heat recovery.
- Can stand periodic high-temperature upsets.
- Zeolite catalysts can be used for high sulfur applications. The reaction with zeolite catalysts occurs inside the molecular sieve body, rather than on the surface of a metal catalyst. This eliminates the sulfur poisoning of metallic catalysts and reduces the conversion of SO_2 into SO_3.

The main disadvantage of the high temperature catalyst is cost.

SCR Support Facilities

SCR systems require the following support facilities:

- Ammonia storage and vaporization facilities.
- An air compressor provides dilution air for the ammonia to aid injection and distribution. Air to ammonia ratio of approximately 20/1 is used to stay safely below the 15% LEL, for NH_3, and to provide enough volume for good NH_3 distribution.
- Control of the ammonia injection rate. As a minimum, the ammonia feed rate should be controlled by the exit NO_x concentration, possibly adjusted for the firing rate. For larger installations, those with varying NO_x concentrations or for those with a tight ammonia slip specification, an inlet NO_x analyzer or an ammonia analyzer could be used to reset the ammonia/feed ratio.
- Soot blowers may be required for services with high particulate loadings. Down flow designs can also help in flushing any solids through the system.
- The ammonia injection is typically through a distribution grid (AIG) located upstream of the reactor. This distribution is critical to the performance of the system and to minimize ammonia slip. For a high percent NO_x reduction, this distributor should be located as far upstream as possible to get very good mixing. The grid should be removable for on-stream cleaning if there are high concentrations of SO_3 in the flue gas.
- Turning vanes may be required to ensure distribution of the airflow across the catalyst. Computer modeling (CFD) or "cold flow" modeling may be required to ensure the proper air distribution and the proper design of the straightening vanes.

Precipitation of Ammonium Sulfate Salts

The ammonia added to the SCR can react with any SO_3 to form ammonium sulfate salts. These salts can precipitate at high temperatures to deactivate the SCR catalyst or to foul downstream heat transfer surfaces.

A portion of the SO_2 present in the flue gas is actually present as SO_3. The SO_2 is frequently either measured or calculated from sulfur in the fuel. The SO_3, as a percentage of the total SO_x, can vary depending on the source, but tends to range between 3% and 10% of the total SO_x.

- In addition to the SO_3 in the incoming flue gas, part of the SO_2 is converted into SO_3 by the SCR catalyst. This percent conversion can vary from less than 1–5%, depending on the operating temperature and the type of catalyst.

The temperature at which the ammonium sulfate salts start precipitating (dew point) is a function of both the ammonia and the SO_3 concentration.

Fig. 9 is taken from API 536 (as seen in Fig. 6). This figure can be used to determine the dew point of the mixtures in and out of the SCR.

Because of capillary action, the liquid in micropores of a catalyst has a lower vapor pressure than the free liquid. This means that ammonium salts have a higher

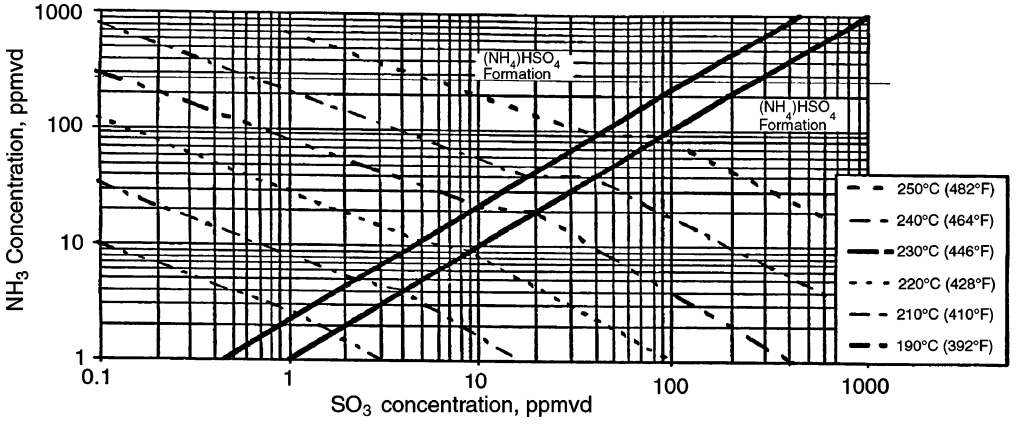


Fig. 9 Ammonium salt dew points 10% water. (Courtesy of API 536.)

dew point in the pores of the SCR catalyst than would be indicated by the curves above.

CATALYTIC SCRUBBING NO_X REDUCTION

NO_2 can easily be scrubbed with caustic solutions, but NO cannot be scrubbed. Most thermal NO_x is NO; so the NO_x in the vent from fired units cannot be scrubbed unless the NO is converted into NO_2. However, many nonfired processes generate NO_2 that can be scrubbed. A number of noncatalytic scrubbing systems have been installed on nitric acid plants and plants generating NO_2 in the process.

There are a number of methods to convert NO into NO_2 so that the NO_x can be scrubbed:

- Adding oxidizing additives to the scrubbing solution.
- Oxidizing the NO in the vapor stream upstream of the scrubber.

Additives to Scrubbing Solution

A number of vendors have proprietary technology for the catalytic oxidation of NO in scrubbing systems. These technologies use liquid-phase oxidizing agents to convert the NO into NO_2, or chelating agents to tie-up the NO in solution. However, the NO must first be dissolved in the scrubbing solution before it can react. So, a large number of theoretical stages are required to dissolve and then react with the NO.

Vapor Phase Oxidation Upstream of Scrubber

An alternate approach to catalytic scrubbing is the injection of ozone, chlorine, or hydrogen peroxide into the flue gas upstream of the scrubber. Ozone is the more powerful oxidizing agent, with chlorine just behind. Hydrogen peroxide is not always practical because of the long reaction times required.

Ozone injection can reportedly achieve up to a 95% NO_x reduction because the ozone reportedly converts both the NO and NO_2 into N_2O_5. N_2O_5 is more soluble in caustic solutions than NO or NO_2. Considerations for ozone injection include:

- The temperature must be low, below 300°F and ideally below 200°F, to minimize decomposition of the ozone.
- The large oxygen generation requirement. Ozone is generated from oxygen. For large systems, the most likely approach is either the installation of a pressure-swing absorption plant to generate the oxygen or pipeline oxygen.
- The low oxygen to ozone conversion: electric arcs convert around 10% of the oxygen to ozone. Only one-third of the ozone reacts; so effectively only 3% of the oxygen ends up as reactive ozone.
- High power requirements—approximately 5.5kW power is required/pound of ozone generated.
- A long residence time is required upstream of the scrubber for complete oxidation.
- Excess ozone slip must be avoided, as the purpose of the NO_x removal is ozone reduction. Ozone is only slightly soluble in water, but will react with any organic compounds in the water to be removed. Ozone will also react with the sulfites present in scrubbers used to remove SO_2, and will thus be destroyed in the scrubber.

GAS TURBINE NO_X REDUCTION[5]

The vent from gas turbines has a much higher percentage of O_2 than heater and boiler vents. The gas turbine vents have around 15% O_2 unless duct burners are installed. Duct burners can reduce the oxygen to about 8–10%. The NH_3 slip for gas turbines is typically specified at 15% O_2.

Common gas turbine NO_x reduction technologies include:

- Steam or water injection: This has been a standard NO_x control approach. This technology works by cooling the flame. Water is more effective than steam, but high purity water must be used to prevent the deposition of solids. Water or steam is added at a ratio to the fuel (water-to-fuel ratio or WFR) in ranges from 0.28 to 2.48 (natural gas fuel).
- Steam or water injection combined with LNB: This can provide a slight improvement over steam or water injection alone.
- Lean combustion and lean premix systems: This technology has been acceptable on new installations in many parts of the country but is not proven on retrofit applications.
- SCR: This technology has been applied in California for a number of years. Nitrogen oxide levels as low as 4 ppmvd, with 5 ppmvd NH_3 slip, have been demonstrated. This is the most expensive NO_x control option, but it can achieve much lower NO_x numbers than the other technologies.

In addition to the technologies listed here, there are a number of technologies under development:

- Catalytic combustion: This technology uses a catalyst to combust the air/fuel mixture at a lower temperature. One test has demonstrated a NO_x level of 5 ppmvd.
- The SCONOXTMGT catalytic reduction process developed by Goal Line Environmental Technologies, LLC:[6] This technology uses a catalyst for the simultaneous reduction in NO_x, VOCs, and CO.

Typical NO_x levels than can be achieved by the different technologies are:

Uncontrolled gas turbine NO_x	150 ppmvd
Water or steam injection	25–90
Lean combustion/lean premix	9–65 (new installations only)
Catalytic	3–5 (but is still experimental)
SCR	3

THERMAL OXIDIZER NO_X REDUCTION

Thermal oxidizers are historically high NO_x generators because the available pressure drop is too low for good mixing between the vent gas and any air and/or fuel. Thermal oxidizer NO_x levels have traditionally run in the 0.2–0.3 lb/MM Btu range. Lower numbers can be experienced if the organics are diluted in an inert, such as nitrogen, to form a low Btu fuel. Numbers can be much higher if the vent contains NH_3 or organic nitrogen, both of which tend to oxidize to NO_x.

The vent stream can be the fuel source or the air source. If the vent stream is dilute organics in nitrogen, then both fuel and air will need to be added.

Ways to control NO_x from thermal oxidizers include:

- Staged combustion: The first stage operates in a reducing mode. Additional air is added to complete the combustion in the second stage. This is the standard approach for vents containing NH_3 or organic nitrogen, because it prevents these compounds from forming NO_x. But it is more expensive because it requires two fireboxes and separate control systems.
- Staged introduction of waste gases: Gases are added at multiple points in the burner section. Staging can reduce peak flame temperatures and localized hot spots.
- Increase ID fan pressure drop and provide better mixing of the air and fuel. A better burner can be used if sufficient pressure drop is available.
- If both air and fuel must be added, it can be done so in a low NO_x burner. The vent stream could be added in a way that dilutes the flame and thus reduces NO_x.
- Water or steam injection into the burner: This has been used inadvertently when the thermal oxidizer also is used to destroy waste water streams.

CONCLUSIONS

Nitrogen oxides, or NO_x, are a combination of nitrogen oxide (NO) and nitrogen dioxide (NO_2). Nitrogen oxides can combine with VOCs to form ozone. For this reason, NO_x control has become the target of the environmental regulatory authorities. The minimization or reduction of NO_x may be required because of ever-tightening regulations and permitting requirements.

The emission control requirements of the existing as well as future regulations requires an understanding of the NO_x emitting sources, pollution prevention opportunities, emission reduction technologies, costs of controls, and impacts on existing operations and plant safety. This article discusses a number of aspects of NO_x control:

- Nitrogen oxides' reductions through process changes and/or pollution prevention activities
- Ultra low NO_x burners
- Burner-related NO_x reduction, such as FGR.
- Postcombustion NO_x control options such as: selective Noncatalytic reduction (SNCR), SCR, and catalytic oxidation/scrubbing.

When evaluating these technologies, one must remember that in many cases retrofitting the NO_x controls can have positive additional benefits. In many instances, the installation of NO_x controls can result in better energy efficiency and lower maintenance requirements.

ACKNOWLEDGMENT

Portions of this paper are reproduced with permission from Chemical Engineering Progress (March 2002 and April 2002). Copyright © 2002 AIChE. All rights reserved.

REFERENCES

1. Bradford, M.; Rajiv, G. Controlling NO_x emissions Part 1. CEP Magazine **2002**, *98* (3), 42–46.
2. Bradford, M.; Rajiv, G. Controlling NO_x emissions Part 2. CEP Magazine **2002**, *98* (4), 38–42.
3. Alternative Control Techniques Document – NO_x Emissions from Process Heaters (Revised); U. S. Environmental Protection Agency Bulletin EPA-453/R-93-034, September 1993.
4. Paul, P.; Maaskant, O. Catalytic NO_x emission reduction at a cogeneration plant in the food industry, NOXCONF, Paris, France, March 2001.
5. Holland, C. D. Special Report – A Summary of NO_x Reduction TechNOlogies; The Texas Institute for Advancement of Chemical Technology, Special Report, 2000.
6. Goal Line Environmental Technologies, LLC, personal correspondence.

Numerical Computations for Chemical Process Analysis and Design

David G. Retzloff
Department of Chemical Engineering, University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

Chemical industries can be traced back to the Middle East as early as 700 B.C. The Phoenicians in the sixth century B.C. were already making soap, and by the tenth century A.D. the Chinese were producing gunpowder. The early origins of present day chemical manufacturing began in the 17th century A.D. and were dominated by the British in the early 1800s and later by the Germans in the late 1880s. However, it was not until the 19th century that large-scale chemical industries made their appearance. Chemistry as a science originated in the 18th century. Up until the 19th century, chemical plants consisted of pots and kettles in which the chemicals of the times were made. They were built using experience gained from a trial and error approach to chemical processing. A chemical process design based on numerical computations could not be made during this period as the concepts of stoichiometry, Avogadro's number, and the periodic table are required for mass balance. The periodic table first appeared in printed form in 1869. Avogadro's number was proposed in 1811 and the first experimental measurements of this number were done in 1827. Jeremias Richter enunciated the principles of stoichiometry in 1792. These principles were not in widespread use in the 1880s. In addition, one also needed the laws of thermodynamics to write the energy balance equation. To put this into historical perspective, Mikhail Vasilievitch Lomonossov first formulated the laws of conservation of energy and mass in 1748 in a letter to Euler. Antoine Lavoisier published a book entitled Elementary Treatise of Chemistry in 1789, which contained the first concise statement of the Law of Conservation of Mass. It would take the work of Thompson (1798), Carnot (1824), Kelvin (1824), Clausius (1857), Maxwell (1859), Boltzmann (1872), Gibbs (1873), and others to establish the foundations of thermodynamics that are required to formulate the energy balance on a chemical process. The science of chemistry began in the 1830s in Germany and England. It was not until 1888 that chemical engineering formally began as a profession when Professor Lewis Norton introduced *Course X* at the Massachusetts Institute of Technology, thus creating a formal degree in chemical engineering. Hence, it is very unlikely that process design calculations including mass and energy balances, as we currently understand them, were used before 1900. One of the first textbooks on plant design was by Vilbrandt in 1942.[1] The initial design of a chemical plant consisted of bringing together the concepts of the chemistry and mechanical equipment principles to design and build the early chemical plants. The chemical plant designer relied heavily on experience gained from building previous chemical plants and elementary calculations to account for the chemicals used and the chemistry that occurred. The empirical relationships between the equipment size and the quantity of material processed per unit time were being developed in this period. This was followed by the development of the concept of unit operations[2] that established the basics of a chemical plant design, as we know it today. The basic calculations that were done were algebraic in nature and consisted of the steady-state equations for mass and energy for the process equipment. The calculations were performed using a pencil, paper, and a slide rule. The formulation of the design equations frequently resulted in nonlinear algebraic equations, transcendental equations, or differential equations. However, to obtain a numerical solution, these equations are reduced to a system of linear equations of the form $Ax = B$ by appropriate linearization of the original design equations. Thus, the numerical solution of the design equations requires that one establish that the linearized equations converge in a well-defined mathematical sense to the original equations and then solve the linearized equations by using well-established numerical techniques. It is also common to encounter a ''trial-and-error'' approach to solving the design equations that can be reduced to solving a system of linear equations to obtain updated values of those quantities that were determined initially by guessing, combined with a criterion for terminating the calculations when the unknowns are determined with sufficient accuracy. The computations involved are ideally suited to modern computers. As a result, a number of chemical process simulators have become available and simulators are now routinely used to perform chemical process design calculations and evaluate design alternatives. For this reason, numerical methods in

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120026494

chemical process design are also discussed in the context of a process simulator. The process simulator is first discussed, followed by a review of the numerical methods employed and the assimilation of numerical algorithms into process simulators to represent process units that are not already available as library routines.

CHEMICAL PROCESS SIMULATORS

Chemical process design consists of five components. The first identifies the equipment components that will be used in each block of the generic process flow diagram shown in Fig. 1.

The second consists of writing the material, energy, momentum, and empirical equations to determine the composition, flow rates, and size of each piece of equipment needed to process the materials. The third involves solving the resulting equations to determine the specifications for each piece of equipment in the plant design. In the fourth, an economic analysis is made of the proposed plant. The fifth is the optimization of the chemical process design both economically and functionally. In the plant design, several alternative processing strategies are usually considered. The economic analysis normally determines which, if any, of the process alternatives will be implemented. From the point of view of numerical complexity, economic analysis is a straightforward process that involves applying reasonable cost-estimating procedures. However, it is not the focus of this article; emphasis is on the numerical methods used in chemical process design.

Process design has matured to the point that there currently exist a number of commercially available process simulators for chemical process design. For instance, AspenPlus, ChemCad, HySys, ICAS, and Sim42 are among the well-known simulator packages. These simulators have a number of similarities with regard to defining the design to be simulated. They begin with a simulation flow sheet somewhat similar to Fig. 1 in which specific process units are depicted for each of the generic blocks. Streams representing the flow of material from one unit to another connect the process units. The simulation flow sheet consists of a collection of icons to represent the mathematical models and subroutines that describe the entire chemical process. Each process unit is associated with a mathematical model that describes its physical behavior. This is in contrast to the process flow sheet that depicts the process units and their interconnection, and represents the material and energy flows in the chemical process. This simulation flow sheet also depicts the flow of information during the numerical calculations. Current simulators contain thermodynamic models and empirical equations that enable process units to be appropriately sized. Once the simulation flow sheet has been constructed, information on the required thermo-physical properties are supplied to the computational program usually via a graphical user interface (GUI) that allows the user to identify the relevant chemicals that occur so that the necessary thermo-physical data are available from the database for the simulation run. There are two types of simulations that are normally run in process design. The first is the steady-state simulation, which identifies the behavior of the process for a large portion of the plant operation. In this simulation, the flow rates, the composition of flow streams, the thermodynamic state of the material, and the size of each piece of process equipment is determined. The second is dynamic simulation. The object of dynamic simulation is to evaluate the response of process units and the overall process design to disturbances. The viability of control schemes can be examined in dynamic simulation. A simulation flow sheet is required to establish the order in which the equations or modules will be solved in the simulation process. Although the simulation flow sheet is not a numerical method, it is a part of the overall strategy for performing chemical process design

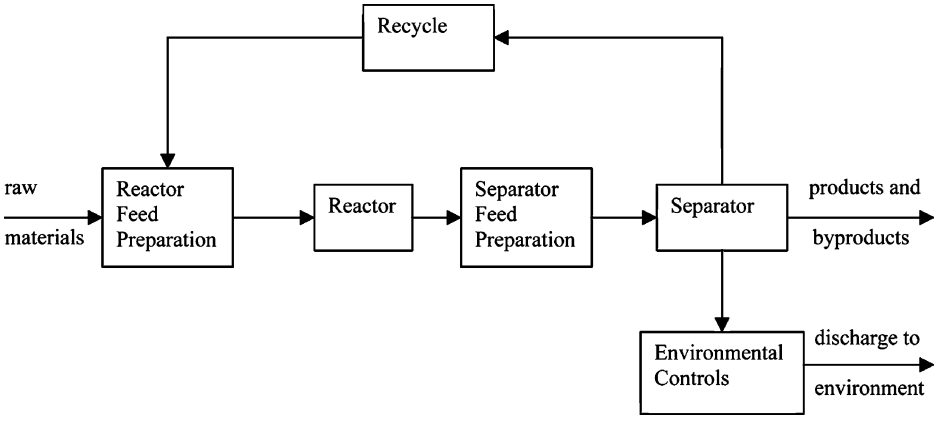


Fig. 1 A generic process flow diagram for chemical plant design.

evaluations. The procedure associated with determining a simulation flow sheet is presented next, followed by a discussion of the numerical methods employed in chemical process plant simulations.

GENERATION OF SIMULATION FLOW SHEETS

In chemical process plant simulations, one of the main considerations is the identification of all the feasible process flowsheets and the elimination of the nonoptimal ones. When only one reaction occurs to produce the required products and only one set of raw materials can be used, the problem of identifying the optimal flowsheets is relatively straightforward and can be accomplished by the application of common sense, heuristics, and analysis. However, when there are multiple reaction pathways to the products and multiple choices for the starting raw materials, the identification of the optimal flowsheets becomes a daunting task. Recently, two algorithmic approaches have been developed to determine the optimal flowsheets. The first approach developed by Friedler et al.[3] is based on graph theory and uses P-graphs to identify feasible flowsheets. Such flowsheets are optimized with respect to an objective function usually based on economic considerations. Optimal flowsheets are used to perform the process simulation. The second approach stems from the ideas of natural selection and genetics and is called a genetic algorithm. Laquerbe et al.[4] discuss the challenges in establishing a natural selection-like criteria for determining the possible flowsheets. The resulting flowsheets are then run on a process simulator to determine equipment sizes, processing conditions, and flow stream information. This information is obtained from computer programs that utilize the numerical methods to solve the mathematical equations that describe the process units that appear in the flowsheet. These numerical methods are the subject of the following section.

NUMERICAL METHODS FOR PROCESS SIMULATORS AND PROCESS DESIGN

Steady-State Simulations

For steady-state simulations, several solution methods have been used. The one most frequently used is the sequential approach. In this method, numerical modules are used that calculate the output stream of a process unit from the input streams coupled with any additional information that is required to uniquely define the performance of the process unit. The simulation treats one unit model at a time. For a simulation flow sheet consisting of interconnected units, the order in which the numerical modules are solved is determined by flow sheet decomposition methods based on process topology. When recycle is present, a tear-stream (guesses for the variables not uniquely determined by the equations describing the process unit) is used to provide an iterative calculation process to determine the tear variables. Convergence is obtained by using fixed-point iteration, Wegstein acceleration, or a quasi-Newton method.

A second solution method is based on an equation-oriented approach, which represents each process unit with a set of equations that are either part of the simulator model library or provided by the user. This method frequently uses an ordered set of equations. Procedures have been developed to automatically determine this information after the simulation flow sheet topology and the complete set of equations have been defined for the process simulation. The equations for all of the process units are solved simultaneously. The equations describing every process in the simulation flow sheet are solved by standard numerical methods that are discussed later. The methods for simulation flow sheet decomposition and equation ordering are covered in the work of Cameron and Hangos[5] and Westerberg et al.[6]

The procedures employed by a simulation package come directly from the approach that would be taken if the chemical process design were to be done by using a handheld calculator and process topology. This should not be surprising as the primary function of the computer is to perform calculations faster than is possible by an individual. The procedure is depicted in Fig. 2. The steady-state behavior of a process unit can be treated as a lumped parameter system. As such, it can be described mathematically in terms of its inputs and outputs. The distillation column and the constant flow stirred tank reactor are examples. In these cases, the equations describing the process equipment are algebraic in nature. If $\{x_1, \ldots, x_n\}$ represent the dependent variables of the problem such as concentration, temperature, and so on, then the equations representing the mass, energy, and momentum equations can be written as

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= 0 \\ f_2(x_1, \ldots, x_n) &= 0 \\ f_3(x_1, \ldots, x_n) &= 0 \\ &\cdots\cdots\cdots\cdots \\ f_n(x_1, \ldots, x_n) &= 0 \end{aligned} \tag{1}$$

This set of equations must be augmented with additional equations to determine the size of the process equipment from empirical relationships. The

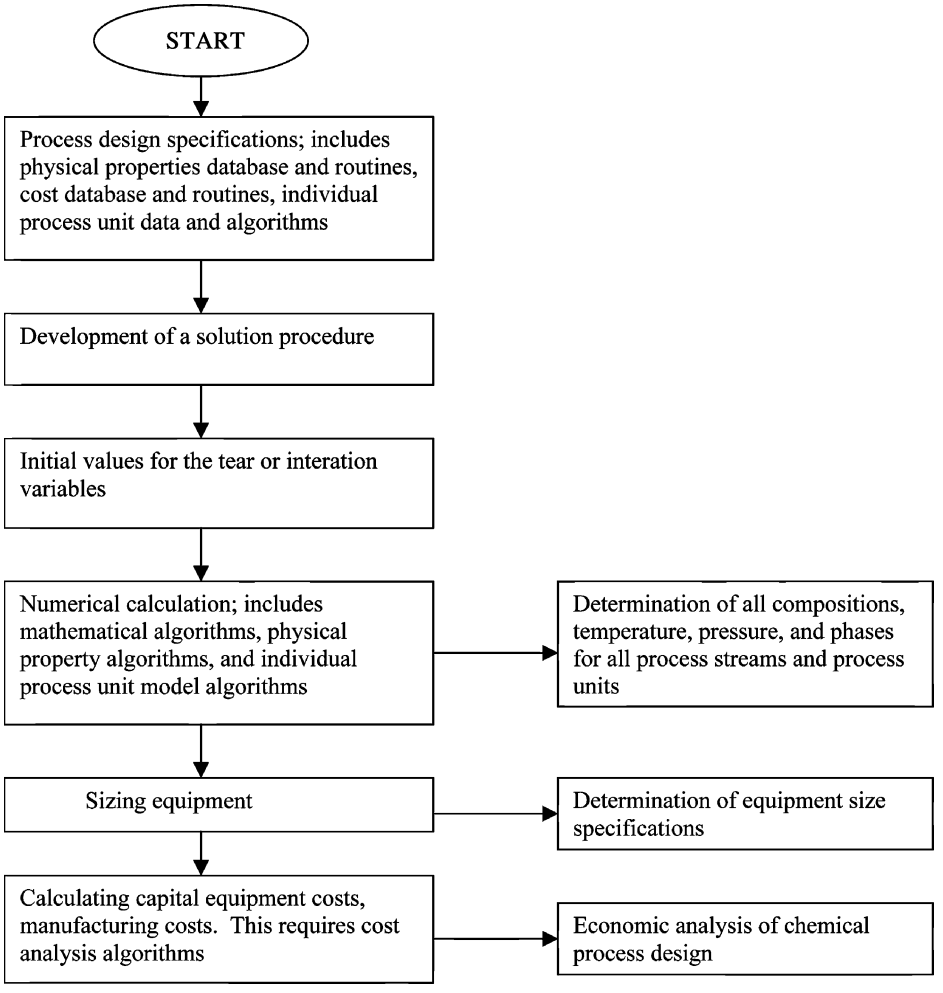


Fig. 2 General solution procedure for process design. This is implemented in process simulator packages.

conceptually simplest method to solve these equations is to rewrite them as

$$\begin{aligned} x_1 &= g_1(x_1, \ldots, x_n) \\ x_2 &= g_2(x_1, \ldots, x_n) \\ x_3 &= g_3(x_1, \ldots, x_n) \\ &\ldots\ldots\ldots\ldots\ldots \\ x_n &= g_n(x_1, \ldots, x_n) \end{aligned} \tag{2}$$

and use a fixed point iteration scheme. Essentially starting with an initial guess for $\{x_1, \ldots, x_n\}$ and substituting into the right-hand-side results in an updated estimate for $\{x_1, \ldots, x_n\}$. Intuitively what is required is that at each step of the iteration, the latest estimates for $\{x_1, \ldots, x_n\}$ are closer in norm to the solution than the previous values. This is called a contraction mapping and a sufficient condition for this to occur is that the equations satisfy a Lipshitz condition.

The conceptually simplest numerical approach to solving the equations is to linearize Eq. (1) about the unknown solution using a Taylor series expansion and rewrite the results to solve for the unknown solution. This yields the vector Newton's iterative method for solving the equations, which takes the form

$$\mathbf{J}^{(k)}\delta^{(k)} = -\mathbf{F}^{(k)} \tag{3}$$

where $\mathbf{J}^{(k)}$ is the Jacobian of Eq. (1) and the terms in Eq. (3) are defined as

$$\delta^{(k)} = \begin{vmatrix} x_1^{(k+1)} - x_1^{(k)} \\ x_2^{(k+1)} - x_2^{(k)} \\ \ldots\ldots\ldots\ldots \\ x_{n-1}^{(k+1)} - x_{n-1}^{(k)} \\ x_n^{(k+1)} - x_n^{(k)} \end{vmatrix}$$

$$\mathbf{J}^{(k)} = \left| \frac{\partial f_i^k}{\partial x_j} \right|$$

$$\mathbf{F}^{(k)} = \begin{vmatrix} f_1(x_1^{(k)}, \ldots, x_n^{(k)}) \\ f_2(x_1^{(k)}, \ldots, x_n^{(k)}) \\ \ldots\ldots\ldots\ldots\ldots \\ f_{n-1}(x_1^{(k)}, \ldots, x_n^{(k)}) \\ f_n(x_1^{(k)}, \ldots, x_n^{(k)}) \end{vmatrix} \tag{4}$$

This is a system of equations of the form $Ax = B$. There are several numeral algorithms to solve this equation including Gauss elimination, Gauss–Jacobi method, Cholesky method, and the LU decomposition method, which are direct methods to solve equations of this type. For a general matrix A, with no special properties such as symmetric, band diagonal, and the like, the LU decomposition is a well-established and frequently used algorithm.

The LU decomposition to solve $Ax = B$ given A and B, is developed in two steps. The two matrices L and U are determined from the equation $LU = A$ and the requirement that U be upper triangular and L be lower triangular with unit diagonal elements. Thus for A an $n \times n$ matrix, A, L, and U have the form

$$A = \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \cdots & a_{1N} \\ a_{21} & a_{22} & a_{23} & a_{24} & \cdots & a_{2N} \\ a_{31} & a_{32} & a_{33} & a_{34} & \cdots & a_{3N} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{N1} & a_{N2} & a_{N13} & a_{N4} & \cdots & a_{NN} \end{vmatrix}$$

$$L = \begin{vmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & 0 & \cdots & 0 \\ l_{31} & l_{32} & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{N1} & l_{N2} & l_{N3} & l_{N4} & \cdots & 1 \end{vmatrix}$$

$$U = \begin{vmatrix} u_{11} & u_{12} & u_{13} & u_{14} & \cdots & u_{1N} \\ 0 & u_{22} & u_{23} & u_{24} & \cdots & u_{2N} \\ 0 & 0 & u_{33} & u_{34} & \cdots & u_{3N} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & u_{NN} \end{vmatrix} \tag{5}$$

The matrices L and U are found by solving $LU = A$. Once L and U are known, the problem $Ax = B$ is solved for x by first solving for y in the equation $Ly = B$ and then solving for x in the equation $Ux = y$. Solving the equation $Ly = B$ for y is known as forward substitution and solving the equation $Ux = y$ for x is called back substitution.

Then, the algorithm becomes

$$\begin{aligned} u_{1j} &= a_{1j} \quad j = 1, 2, \ldots, N; \\ l_{i1} &= a_{i1}/u_{11} \quad i = 2, \ldots, N \\ u_{ij} &= a_{ij} - \sum l_{ik}\, u_{kj}; \quad i = 2, \ldots, N; \\ &\qquad j = i + 1,\ i + 2, \ldots, N \\ l_{ij} &= (a_{ij} - \sum l_{ik}\, u_{kj})/u_{jj}; \\ &\qquad j = 2, 3, \ldots, N - 1; \\ &\qquad i = j + 1,\ j + 2, \ldots, N \end{aligned} \tag{6}$$

$$y_1 = b_1, \quad y_i = b_i - \sum l_{ij}\, y_j; \ i = 2, 3, 4, \ldots, N$$

and

$$x_N = y_N/u_{NN}, \quad x_i = (1/u_{ii})\left(y_i - \sum u_{ij}\, x_j\right); \quad i = N - 1, N - 2, \ldots, 1$$

To improve the accuracy of the calculation, partial pivoting obtained in the form of exchanging rows so that the largest diagonal element is in the row on which one is currently calculating is employed. The column containing the diagonal element of the row that is to be calculated is searched, starting below the diagonal element, for an element larger in magnitude. If such an element is found, the two rows are interchanged so that the largest possible element in magnitude will appear at the diagonal position. This is required to ensure stablity of the algorithm. If zero or a very small element appears at the diagonal position during this computation, then an error message is generated and the calculation terminated to prevent division by zero. To efficiently use the dynamic memory of the computer, the elements of the L and U matrices replace those of the A matrix as they are calculated. However, the diagonal elements of the L matrix that are 1s are not stored. Furthermore, the elements of the b vector are stored as an extra column of the A matrix, which is then called the augmented A matrix. As the elements of the solution vector y are determined, they are stored in the row of the augmented matrix A that originally held the b vector. The x vector is calculated from the elements of the final augmented A matrix. The x vector can be stored as the last column of the augmented A matrix if desired or can be calculated and written to a file. This algorithm is easily written as a parallel program that can be executed in a distributed computer cluster or a Beowulf cluster.

The main drawback of Newton's method is the evaluation of the derivatives of the functions in the expression for $\mathbf{J}^{(k)}$ in Eq. (6). This requirement involves either writing explicit equations for the derivatives and encoding these expressions into the simulator program or numerically computing the derivative. This is computationally intensive and must be iterated until $\|\delta^{(k)}\| < \epsilon$ at which point the solution is $\mathbf{x}^{(k+1)} = \{x_1^{(k+1)} x_2^{(k+1)} \cdots x_n^{(k+1)}\}^T$. Alternatively, one can replace $(\mathbf{J}^{(k)})^{-1}$ in Eq. (3) with an approximation $\mathbf{H}^{(k)}$ that is easy to compute and use iteration to solve the resulting equation

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{H}^{(k)}\mathbf{F}^{(k)} \tag{7}$$

Broyden's method and Wegstein's method are two examples of this approach that use different approximations for $(\mathbf{J}^{(k)})^{-1}$. In the former,[7] which is based on the Householder, theorem $\mathbf{H}^{(k)}$ is calculated recursively using the equation

$$\mathbf{H}^{(k+1)} = \mathbf{H}^{(k)} + [(\delta^{(k)} - \mathbf{H}^{(k)}\mathbf{F}^{(k)})(\delta^{(k)})^T\mathbf{H}^{(k)}]/ [(\delta^{(k)})^T\mathbf{H}^{(k)}\mathbf{F}^{(k)}] \tag{8}$$

The latter uses a secant approximation to the derivatives in $\mathbf{J}^{(k)}$. In this method, the functions in Eq. (1) are written as $f_i(\mathbf{x}^{(k)}) = x_i^{(k)} - g_i(\mathbf{x}^{(k)})$ and the method can be written for each component in $\mathbf{x}^{(k+1)}$ as

$$\begin{aligned} x_i^{(k+1)} &= (1 - \omega_i)x_i^{(k)} + \omega_i\, g_i(\mathbf{x}^{(k)}), \\ \omega_i &= 1/[1 - q_i], \\ q_i &= [g_i(\mathbf{x}^{(k)}) - g_i(\mathbf{x}^{(k-1)})]/[x_i^{(k)} - x_i^{(k-1)}] \end{aligned} \tag{9}$$

The Wegstein acceleration is a popular algorithm in chemical process simulators as it is easy to implement in a computer and well adapted to handle recycle calculation when recycle streams are present in the process flow diagram and tear variables are used to solve the recycle loops. However, the convergence of these numerical schemes can be a problem when recycle streams are present. The difficulty arises from the method employed to obtain new estimates of the assumed tear variables in the iteration procedure. If the method for determining successive estimates of the tear variables is equivalent to a fixed-point iteration scheme, then convergence is assured. However, establishing that this is the case is often difficult. These numerical methods are currently included in commercially available chemical process simulators. Thus to take advantage of these numerical procedures, a user has to supply only a mathematical description of any chemical process unit that is not already available in the process simulator. Most simulators are written using FORTRAN as the computer language and have incorporated a procedure for writing user-defined models of chemical processing units and producing object module files for these process units that can be used by the process simulator in connection with its simulation flow sheet and process topology sequencing of the order of solution of the process units.

We next consider the simulation of the unsteady-state behavior of chemical process units.

Dynamic Simulations

Dynamic simulations represent the temporal and the spatial behavior of a chemical process unit in the presence of perturbations or at process startup. There is a natural division in the types of numerical methods used to solve the equations describing the dynamic behavior of the process. In lumped parameter descriptions of the process units, the resulting equations are ordinary time evolution differential equations, whereas for distributed parameter descriptions of process units the resulting equations are parabolic partial differential equations. The numerical methods used to solve these equations are very different and necessitate a separate discussion. Numerical methods used to solve ordinary differential equations describing the dynamics are considered first followed by a discussion of the methods employed to solve evolution equations of the parabolic type.

Dynamics Described by Ordinary Differential Equations

Examples of chemical process units in this category include constant flow stirred tank reactors, separation processes such as distillation columns, liquid–liquid extractors, and other units that are described in terms of the thermodynamic equilibrium stage concept. To develop a numerical algorithm, time derivatives are replaced by finite differences. There are a number of numerical algorithms to solve the resulting equations. For differential equations that are not stiff (the dynamics do not exhibit multiple time scales, i.e., the dynamical behavior does not exhibit time intervals for which the dependent variables change very slowly and other time intervals in which the dependent variables change rapidly with time), the fourth-order Runge–Kutta algorithm is a viable algorithm. For stiff differential equations, either Gear's method or the implicit form of the modified mid-point method discussed by Bader and Deuflhard[8] provide efficient algorithms. The modified mid-point method is presented because the author's experience is that it works well with both stiff and nonstiff differential equations and systems of differential equations even though it

is weakly unstable and the truncation error is of order h^3. The presentation focus is on systems of differential equations, as there is negligible additional effort to formulate the method for more than one evolution differential equation.

Let $\mathbf{Y}$ denote the column vector of physical dependent variables (e.g., mass or mole concentration or temperature) and $\mathbf{Y}_n$ the value of $\mathbf{Y}$ at time $t_n = n\Delta t$, then the system of differential equations under consideration is written as

$$\frac{d\mathbf{Y}}{dt} = \mathbf{f}(\mathbf{Y}) \tag{10}$$

The finite difference algorithm is obtained by replacing the time derivatives by a forward difference, using an implicit rule to evaluate $\mathbf{F(Y)}$ at time t_n, and setting $h = \Delta t$. The result is

$$\mathbf{Y}_{n+1} - \mathbf{Y}_n = 2h\mathbf{f}([\mathbf{Y}_{n+1} + \mathbf{Y}_{n-1}]/2) \tag{11}$$

The numerical algorithm is obtained by linearizing $\mathbf{f}$ about $\mathbf{Y}_n$ and rewriting Eq. (11) as

$$\mathbf{Y}_{n+1} - \mathbf{Y}_n = \mathbf{Y}_n - \mathbf{Y}_{n-1} + 2[\mathbf{I} + \partial\mathbf{f}/\partial\mathbf{Y}]^{-1} \times [h\mathbf{f}(\mathbf{Y}_n) - (\mathbf{Y}_n - \mathbf{Y}_{n-1})] \tag{12}$$

By defining $\Delta_p = \mathbf{Y}_{p+1} - \mathbf{Y}_p$ the computational procedure can be written in the standard form

$$\begin{aligned} \Delta_p &= \Delta_{p-1} + 2[\mathbf{I} + \partial\mathbf{f}/\partial\mathbf{Y}]^{-1}[h\mathbf{f}(\mathbf{Y}_p) - \Delta_p]; \\ & \quad p = 1, \ldots, q-1 \\ \mathbf{Y}_{p+1} &= \mathbf{Y}_p + \Delta_p \end{aligned} \tag{13}$$

with t_q as the final time of the calculation. This is not a self-starting algorithm. To start the calculations, the semi-implicit Euler algorithm is used in the following form:

$$\begin{aligned} \Delta_0 &= [\mathbf{I} + \partial\mathbf{f}/\partial\mathbf{Y}]^{-1} h\mathbf{f}(\mathbf{Y}_0) \\ \mathbf{Y}_1 &= \mathbf{Y}_0 + \Delta_0 \end{aligned} \tag{14}$$

The matrix inversion $[\mathbf{I} + \partial\mathbf{f}/\partial\mathbf{Y}]^{-1}$ can be accomplished by using the LU-decomposition method discussed previously. The final value for $\mathbf{Y}$, i.e., $\mathbf{Y}_q$, can be calculated from Eq. (13), but it is frequently determined by the smoothing operation given by

$$\begin{aligned} \Delta_q &= [\mathbf{I} + \partial\mathbf{f}/\partial\mathbf{Y}]^{-1}[h\mathbf{f}(\mathbf{Y}_q) - \Delta_{q-1}] \\ \langle\mathbf{Y}_q\rangle &= (\mathbf{Y}_{q+1} + \mathbf{Y}_{q-1})/2 = \mathbf{Y}_q + \Delta_q \end{aligned} \tag{15}$$

The algorithm is applied by choosing an interval size that integer divides t_q and is suitable for representing the solution graphically. Then, Eqs. (13–15) are applied across the interval with an appropriately chosen step size, h, so that the desired accuracy is obtained. The process is repeated for each succeeding interval until $\mathbf{Y}_q$ is determined. In general, reducing the step size lessens the error in the approximation of the derivative by a finite difference. Step size halving, recomputing to the same end time, and comparing the values of $\mathbf{Y}$ for two successive time steps provide a measure of the accuracy of the calculation. This procedure for ascertaining the accuracy of the calculation can be computationally optimized by incorporating adaptive step size control and extrapolation to zero step size in the algorithm so that the largest step size possible is used to compute $\mathbf{Y}$ across the interval consistent with the specified error tolerance. Extrapolation to zero step size as the name implies takes the values of $\mathbf{Y}$ for all available step sizes across the interval and extrapolates to the value of $\mathbf{Y}$ at the end of the interval that corresponds to choosing a step size of zero. The extrapolation is done using polynomial functions or rational functions. A complete implementation of this algorithm in C with adaptive step size control and extrapolation to zero step size can be found in Ref.[9].

The presence of recycle streams increases the complexity. A recycle stream in this context is a stream that is returned to a unit. It appears earlier in the simulation flow diagram with no change in physical properties and chemical composition from its process unit of origin. The recycle streams introduce a time delay in the mathematical equations and change the mathematical description from a finite dimensional system of equations to an infinite dimensional system of equations. Discussion of these issues is beyond the scope of this entry.

One method to solve partial differential equations using the numerical schemes developed for solving time dependent ordinary differential methods is the method of lines.[10] In this method, the spatial derivatives at time t are replaced by discrete approximations such as finite differences or finite element methods such as collocation or Galerkin. The reason for this approach is the advanced stage of development of schemes to solve ordinary differential equations. The resulting numerical schemes are frequently similar to those developed directly for partial differential equations.

A frequent application of dynamic simulation is to investigate a proposed control strategy for a chemical plant. It is usually assumed that the plant is operating at steady-state initially. The objective is to investigate the response of the system including controllers to process disturbances and evaluate the performance obtained for different types of control action. What

is required is a map that transforms input (both control actions and upsets) changes into output responses. The method of lines is a good candidate for a numerical scheme to provide the required map.

Dynamics Described by Partial Differential Equations

Examples of chemical process units in this category include plug flow reactors, laminar flow reactors, turbulent flow reactors, plasma reactors, and separation units that are described in terms of the mass transfer concept. To develop a numerical algorithm, the time and spatial derivatives are replaced by finite difference approximations. In general, the time derivative is represented by a forward difference, whereas the second order spatial derivatives are approximated by central differences as follows for the dependent variable Y in Cartesian coordinates:

$$\begin{aligned}
\frac{\partial Y}{\partial t} &= \frac{Y^{n+1}_{jkl} - Y^{n}_{jkl}}{\Delta t},\\
\frac{\partial^2 Y}{\partial x^2} &= \frac{Y^{n}_{j+1kl} - 2Y^{n}_{jkl} + Y^{n}_{j=1kl}}{(\Delta x)^2},\\
\frac{\partial^2 Y}{\partial y^2} &= \frac{Y^{n}_{jk+1l} - 2Y^{n}_{jkl} + Y^{n}_{jk-1l}}{(\Delta y)^2}\\
\frac{\partial^2 Y}{\partial z^2} &= \frac{Y^{n}_{jkl+1} - 2Y^{n}_{jkl} + Y^{n}_{jkl-1}}{(\Delta z)^2},\\
Y^{n}_{jkl} &= Y(t = n\Delta t,\ x = j\Delta x, y = k\Delta y, z = l\Delta z)
\end{aligned} \tag{16}$$

There are a number of numerical algorithms to solve the difference equation representation of the partial differential equations. Implicit algorithms such as Crank–Nicolson scheme where the finite difference representations for the spatial derivatives are averaged over two successive times, $t = n\Delta t$ and $t = (n + 1)\Delta t$, are frequently used because they are usually unconditionally stable algorithms. Most conservation laws lead to equations of the form

$$\frac{\partial \mathbf{Y}}{\partial t} = \nu\Delta^2\mathbf{Y} + \mathbf{F}(t, \mathbf{x}, \mathbf{Y}) \tag{17}$$

which are parabolic equations.

The resulting numerical scheme for solving this system of differential equations can be written in the following general form:

$$\mathbf{Q}_1\mathbf{Y}^{n+1} = \mathbf{Q}_2\mathbf{Y}^{n} + \Delta t\mathbf{G} \tag{18}$$

The ordering for the dependent variables Y^{n}_{jkl} is

$$\begin{aligned}
\mathbf{Y}^{n} = [&Y^{n}{}_{111}, \ldots, Y^{n}{}_{\text{jmax}\,11}, Y^{n}{}_{121}, \ldots,\\
&Y^{n}{}_{\text{jmax}\,21}, \ldots Y^{n}{}_{\text{jmax kmax}\,11}, \ldots Y^{n}{}_{\text{jmax kmax lmax}}]^{\mathrm{T}}
\end{aligned} \tag{19}$$

The exact form of the matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ depends on the type of partial differential equations that make up the system of equations describing the process units, i.e., parabolic, elliptic, or hyperbolic, as well as the type of applicable boundary conditions, i.e., Dirichlet, Neuman, or Robin boundary conditions. The matrix $\mathbf{G}$ contains the source terms as well as any nonlinear terms present in $\mathbf{F}$. It may or may not be averaged over two successive times corresponding to the indices n and $n + 1$. The numerical scheme solves for the unknown dependent variables at time $t = (n + 1)\Delta t$ and all spatial positions on the grid in terms of the values of the dependent variables at time $t = n\Delta t$ and all spatial positions. Boundary conditions of the Neuman or Robin type, which involve evaluation of the flux at the boundary, require additional consideration. The approximation of the derivative at the boundary by a finite difference introduces an error into the calculation at the boundary that propagates inward from the boundary as the computation steps forward in time. This requires a modification of the algorithm to compensate for this effect.

From an elementary point of view, Eq. (18) can be solved for $\mathbf{Y}^{n+1}$ by multiplying Eq. (18) by the matrix $\mathbf{Q}_1^{-1}$. From the computational point of view, this is not necessarily the most efficient approach. If $\mathbf{F}$ consists of only nonhomogenous terms, then the nonhomogenous alternating direction implicit schemes such as the Douglas–Gunn scheme[11] are computationally attractive. When $\mathbf{F}$ contains nonlinear terms in the dependent variables, a Newton-like scheme is employed. Similar methods are used when the partial differential equations are of the hyperbolic or the elliptic type. The key issues in developing a numerical scheme for solving these balance equations are stability, convergence, and the order of accuracy of the numerical scheme in the variables Δt, and the spatial step sizes. The numerical scheme must be both stable and convergent. In addition, all terms in the partial differential equations should be approximated consistently in the spatial and time step sizes. Multigrid methods that employ two distinct mesh sizes with the operations of prolongation and retraction as well as finite element methods have also been successfully used to obtain numerical solutions for the balance equations.

Recycle streams introduce time delays that appear in the boundary conditions. A discussion of the effects of

recycle streams on the stability and the convergence of the numerical scheme is beyond the scope of this entry.

COMPUTATIONAL CONSIDERATIONS

Solving chemical plant design problems involves developing a numerical scheme that is stable and convergent as well as implementing the resulting algorithm on a computer. The basic consideration is the time required for execution of the algorithm and printing out the results in appropriate format. A large number of design simulations and computations of moderate complexity can be handled by modern personal computers running at speeds in the gigahertz range. Beginning 1997, distributed computing clusters (frequently called Beowulf clusters) composed of off-the-shelf components capable of supercomputer speeds have made their debut and are now commonplace and commercially available. The underlying communication programs necessary for distributed computing are at a mature state of development and freely available as packages under the name of MPI (message passing interface) and PVM (parallel virtual machine). To realize the power of cluster computing, parallel programming is required. To date, there are a few commercial packages available that take a program written in serial format and translate it to a parallel program. However, the most efficient parallel programs are still written by skilled programmers. Most computational programs are not what is termed massively parallel but are rather composed of segments that can be run in parallel separated by sequential segments. At the moment, this requires the skills of an experienced programmer to write a time efficient parallel code. As a result, chemical process simulators have not yet been written to run on Beowulf-type clusters. This will undoubtedly change in the future as translators from serial to parallel computer code become more efficient and cluster computing becomes commonplace. These computer clusters can handle computationally intensive design simulations and have the advantage that they can be constructed using standard personal computers or standard personal computer parts making them economically attractive.

CONCLUSIONS

Models of the steady-state behavior of a chemical plant consist of a system of algebraic equations that can be solved by well-established algebraic numerical methods. These can be implemented on computer clusters, which can achieve supercomputing execution times. Recycle streams require the use of tear variables and an iterative numerical scheme to obtain a solution. The convergence of the numerical algorithm is an important consideration. Several numerical schemes can be used to obtain convergence including Newton's, Broyden's, and Wegstein's methods. Dynamic simulations are required to obtain the time-dependent behavior of the chemical plant. One of the primary objectives of dynamic simulations is the evaluation of proposed control systems and strategies. Chemical process units whose descriptions do not involve spatial dependence give rise to ordinary differential equations that can be numerically solved by finite difference methods such as the modified midpoint method, Gear's method, or the Runga–Kutta method. Gear's method and the modified midpoint method are required for units described by stiff differential equations. Recycle streams introduce time delays into the resulting differential equations, thereby increasing the complexity of the description. When the spatial dependence of the dynamic behavior of a chemical process unit must be considered, partial differential equations arise. Numerical schemes based on implicit finite difference approximations are used to obtain convergent stable algorithms. Convergence, stability, and consistency are important considerations when choosing a particular numerical scheme. The consistency requirement allows one to estimate the time- and spatial-step sizes needed to meet accuracy considerations. Numerical schemes for partial differential equations inevitably lead to systems of algebraic equations for the unknown process variables that are solved numerically by Newton-like methods or explicit matrix methods. When a computer is used in generating the numerical solution, a major consideration is the computational time required. Usually, iterative methods are faster, i.e., require less computational time, than explicit methods for a required numerical accuracy. Currently, computer clusters employing message passing can be used to reduce the computational time. To utilize the supercomputer power of a cluster, the numerical scheme must be written in parallel computer code for execution on the cluster. Computer software is available for converting serial computer code to parallel code. This reduces the effort required to obtain a parallel program. However, experienced programmers currently produce the most efficient (shortest execution time) parallel programs. This is because of the requirement that writing efficient parallel code requires that all parts of the program that can be executed in parallel by multiple CPU (central processing units) must be identified and the associated communication overhead among the CPUs be taken into account. As numerical libraries of efficient parallel code for specific numerical tasks such as matrix multiplication, matrix inversion, and the solution to differential equations become available, producing parallel

code will become a routine task and the use of computer clusters will be a standard practice.

REFERENCES

1. Vilbrandt, F.C. *Chemical Engineering Plant Design*; McGraw-Hill: New York, 1934.
2. Badger, W.L.; McCabe, W.L. *Elements of Chemical Engineering*; McGraw-Hill: New York, 1936.
3. Friedler, F.; Varga, J.B.; Fan, L.T. Decision-mapping: A tool for consistent and complete decisions in process synthesis. Chem. Eng. Sci. **1995**, *50*, 1755–1768.
4. Laquerbe, C.; Laborde, J.C.; Soares, S.; Floquet, P.; Pibouleau L. Domenech, S. Synthesis of RTD models via stochastic procedures: simulated annealing and genetic algorithms. Comput. Chem. Eng. **2001**, *25*, 1169–1183.
5. Cameron, I.; Hangos, K. *Process Modelling and Model Analysis*; Academic Press: New York, 2001.
6. Westerberg, A.W.; Hutchison, H.P.; Motard, R.L.; Winter, P. *Process Flowsheeting*; Cambridge University Press: New York, 1979.
7. Broyden, C.G. A class of methods for solving nonlinear simultaneous equations. Math. Comput. **1965**, *19*, 577–593.
8. Bader, G.; Deuflhard, P. A semi-implicit mid-point rule for stiff systems of ordinary differential equations. Numer. Math. **1983**, *41*, 373–398.
9. Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. *Numerical Recipes in C*; Cambridge University Press: New York, 1992.
10. Liskovets, O.A. The method of lines. J. Differ. Equ. **1965**, *1*, 1308.
11. Douglas, J., Jr.; Gunn, J. A General formulation of alternating direction methods. Part I. Parabolic and hyperbolic problems. Numer. Math. **1964**, *6*, 428–453.

Onsite and Offsite Emergency Preparedness for Chemical Facilities and Chemical Transportation

O

Michael K. Lindell
Hazard Reduction and Recovery Center, Texas A&M University, College Station, Texas, U.S.A.

Ronald W. Perry
School of Public Affairs, Arizona State University, Tempe, Arizona, U.S.A.

INTRODUCTION

Chemicals provide enormous benefits to American society but impose well-recognized risks to the health and safety of plant workers. In addition, there are also risks to adjacent neighborhoods and those living and working near transportation routes linking chemical production, storage, and distribution facilities. All these risks can be managed through effective programs for onsite and offsite emergency preparedness (EP). This entry begins by addressing the demands chemical emergencies place on facility personnel and community emergency responders. It then turns to the processes of developing an emergency preparedness program, formulating an emergency operations plan, and establishing coordination among facility departments and with public sector organizations. This is followed by a discussion of methods for conducting hazard/vulnerability/risk analyses, emergency assessment analyses, expedient hazard mitigation analyses, personnel and population protection analyses, and incident management analyses. The entry concludes with a discussion of methods for developing plans and procedures; establishing emergency operation centers (EOCs); conducting and evaluating training, drills, and exercises; and disseminating risk information.

UNDERSTANDING EMERGENCY DEMANDS

Chemical incidents impose specific agent-generated and response-generated demands on emergency response organizations.[1] Agent-generated demands are caused by the hazard agent itself and threaten human health and safety, property, and the environment. By contrast, response-generated demands are caused by the need to coordinate the activities of the many individuals and organizations seeking to respond to the incident. Major disasters elicit a significant outpouring of assistance, frequently creating convergence at an incident scene. Moreover, when incident demands deviate from planners' expectations, emergency response organizations generally improvise adaptive responses within existing organizations and—in some cases—new organizations emerge to meet unforeseen needs.

Contrary to their expectations, emergency responders typically find households responding to disasters with rational information seeking and adaptive action. Indeed, panic, shock, passivity, disorientation, and antisocial activities such as looting are *extremely* rare.[2] Not only is household emergency response generally rational, it also is frequently prosocial. That is, many individuals, households, and larger organizations volunteer to help during emergencies. Unfortunately, the convergence of unexpected volunteers, the emergence of new organizations, and the improvisation of unplanned responses can hinder an orderly emergency response. Consequently, facility EP coordinators need to develop emergency programs in coordination with community emergency managers and also participate in community-wide emergency response planning, training, and exercising.

To address the demands of major incidents, emergency response organizations must perform four basic types of functions—emergency assessment, expedient hazard mitigation, personnel/population protection, and incident management.[3,4] Emergency assessment actions characterize the nature and the magnitude of an incident by evaluating conditions in the physical and social environment. Expedient hazard mitigation involves *preventive* and *corrective actions* that limit the magnitude of hazard impacts by controlling the source (e.g., plugging leaks in hazmat containers) or the spread (e.g., diking around spills) of hazards. Personnel population protection actions preserve safety and health; for example, by requiring emergency responders to wear respirators and implementing evacuations of community residents. Finally, incident management actions ensure that the emergency assessment, expedient hazard mitigation, and population protection actions are performed promptly and effectively. In addition, incident management actions ensure responders are

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120040387

supported by adequate emergency resources such as support staff, facilities, equipment, and materials.

DEVELOPING AN EP PROGRAM

A capability for prompt and effective emergency response is based on the quality of the facility's EP program, which must have clear support from the plant manager and any levels of management above that. In many instances, the responsibility for facility EP is assigned to the manager for health, safety, and environment. Whoever is assigned this responsibility, the EP coordinator must understand the duties of his or her position: specifically, who does the EP coordinator report to, who reports to him/her, what duties for the position are specifically listed in the job description, and what are the specific qualifications (education, training, and experience) that are listed in the job description.

Once the basic expectations of the job have been established, EP coordinators should develop EP program plans that systematically direct their efforts over the course of each year. The National Fire Protection Association's standard for emergency management programs, NFPA 1600, lists a set of criteria that can be used to guide this development effort.[5] In the public sector, the Federal Emergency Management Agency[6] has advised local emergency managers to set annual goals in each of the major programmatic areas for which they are responsible. The next step is to conduct organizational capability analyses that assign the tasks needed to perform the emergency response functions to organizations having the resources needed to perform those tasks. There is relatively little guidance on performing organizational capability analyses, but the best procedure is to contact subject matter experts (SMEs) for information about each emergency response function.[7] In the case of emergency public information, for example, this would include public safety and health officials and members of the local news media. Once the SMEs have been identified, they can be asked what specific tasks are needed to perform each function under different emergency scenarios (e.g., fire, explosion, toxic chemical release). For each task, they can then be asked to identify the personnel, facilities, equipment, and materials needed. Emergency planners should exercise caution when asking SMEs about performance of tasks under conditions they have not previously experienced. Fire brigade members, for example, might assume their experience in handling small-scale fires and chemical spills will generalize to large-scale incidents involving major conflagrations threatening to cause catastrophic failures of toxic chemical tanks. Thus, when the hazard/vulnerability/risk analysis identifies disaster conditions that might be significantly different from the facility's previous experience, the EP coordinator should seek outside expertise.

This capability assessment is likely to identify satisfactory levels of capability in some areas but not in others. Thus, the EP coordinator should document the capability shortfall and devise a multiyear development plan to reduce it. Emergency preparedness is often a low priority for most organizations; hence, the limited funds available for this activity make it likely that a multiyear (typically five year) development plan will be needed that sets specific annual milestones. Once the annual milestones have been set, facility EP coordinators should monitor their programs' achievement of these objectives.

ESTABLISHING COORDINATION WITH FACILITY DEPARTMENTS AND PUBLIC SECTOR ORGANIZATIONS

A facility EP coordinator can facilitate the development of EP program but cannot implement it alone. Ultimately, EP is the responsibility of the facility's operational departments; hence, the EP coordinator must staff the emergency response organization from the employees in these departments. In addition, the EP coordinator should work with the local fire department, local emergency management agency (LEMA), and local emergency planning committee (LEPC). Obviously, coordination with the local fire department is essential to ensure the adequacy of offsite support (e.g., number of personnel and apparatus) and compatibility of equipment (e.g., hose fittings and radio frequencies). However, there also is a need to ensure the compatibility in the structure of the onsite and offsite emergency response organizations. Moreover, coordination with the LEMA and LEPC is also important in achieving a realistic appraisal of the geographic areas at risk if there are incidents at the facility, the nature of the special facilities and populations that could be affected by these incidents, and the ability of community agencies to provide the emergency response resources needed to respond to these incidents.

CONDUCTING HAZARD/VULNERABILITY/ RISK ANALYSES

Incidents involving fires, explosions, or chemical releases can be initiated by internal (accident or sabotage) or external (geophysical, meteorological, or hydrological events; or terrorist attacks) causes. The types of hazards that can occur at a chemical facility, their initiating events, their consequences, and their likelihoods of occurrence are assessed using hazard analysis. This process begins by identifying dangerous

chemicals (i.e., those that are threats because of their flammability, reactivity, or toxicity), their locations, and the quantities at those locations. Once the chemical inventory has been developed, this information can be used to assess the threats these chemicals pose to the facility, its workers, its neighbors, and the environment. In the case of Extremely Hazardous Substances defined under SARA Title III, Vulnerable Zones (VZs) can be computed using data on the chemical's toxicity, its quantity available for release, the type of spill (liquid or gaseous), the postulated release duration (e.g., 10 min), assumed meteorological conditions (wind speed and atmospheric stability), and terrain (urban or rural). Available methods include manual computations,[8] *ALOHA*,[9] see information about *CAMEO* at www.epa.gov/ceppo/cameo], or *RMP*Comp* at ⟨yosemite.epa.gov/oswer/ceppoweb.nsf/content/rmp-comp.htm⟩.

Once the radii of the VZs for the different chemicals have been computed, these can be overlaid onto a map with the release point in the center of the circle and the radius drawn around it (see Fig. 1). Moreover, there will also be a rectangular VZ surrounding the transportation route to the facility. The facility and the transportation route VZs can then be examined to identify areas of residential, commercial, and industrial land use (see Ref.[10] for an analysis of hazardous waste transportation to an incinerator). In particular, special attention should be given to identifying the locations of special facilities whose inhabitants have limited ability to receive warnings or to take protective action (Table 1).

CONDUCTING EMERGENCY ASSESSMENT ANALYSES

The facility emergency response organization must be prepared to promptly and accurately assess the nature and magnitude of an emergency. As Table 2 indicates, this includes detecting a threatening situation promptly, safely conducting a reconnaissance to obtain critical information about the product and container, and monitoring environmental conditions that affect the direction and extent of any fires, explosions, or releases to air, water, soil, or groundwater.[4,12,13] Data from the reconnaissance and environmental monitoring should be integrated into an emergency classification system (e.g., Level I: Threat to a single building; Level II: Sitewide threat; Level III: Minor offsite threat; Level IV: Major offsite threat) that allows offsite emergency response organizations to anticipate the need for providing technical assistance to the facility or implementing population protective actions. An emergency classification system can be constructed using the information from the hazard/vulnerability/risk analysis by overlaying the radius of the VZ onto a site map. For example, a chemical whose VZ lies entirely within a single building on the facility site can produce only a Level I incident.

Emergency assessment also includes assessing any damage that might increase the severity of the threat, monitoring any fires or releases that occur, and projecting the potential impacts of these events to offsite locations.[4,12,13] These damage assessments should be communicated to offsite agencies so that they can be aware of the potential for any minor incidents (e.g., Level I or II) escalating to Level III or IV. The EP coordinator should ensure adequate personnel, facilities, and equipment to support the emergency assessment function—especially equipment for fire and chemical release detection and meteorological monitoring.

CONDUCTING EXPEDIENT HAZARD MITIGATION ANALYSES

The facility emergency response organization must be prepared to assess the need for, and implement, actions

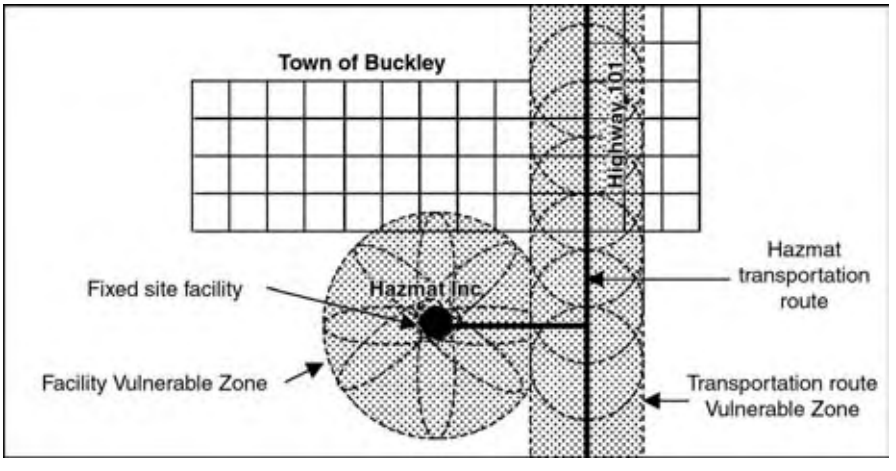


Fig. 1 Vulnerable zones around fixed-site facility and transportation route. (Adapted from Ref.[11].)

Table 1 Reference list of special facilities

Health related
Hospitals
Nursing homes
Halfway houses (drug, alcohol, mental retardation)
Mental institutions
Penal
Jails
Prisons
Detention camps
Reformatories
Assembly and athletic
Auditoriums
Theaters
Exhibition halls
Gymnasiums
Athletic stadiums or fields
Amusement and recreation
Beaches
Camp/conference centers
Amusement parks/fairgrounds/race courses
Campgrounds/RV parks
Parks/lakes/rivers
Golf courses
Ski resorts
Community recreation centers
Religious
Churches/synagogues/mosques
Evangelical group centers
High density residential
Hotels/motels
Apartment/condominium complexes
Mobile home parks
Dormitories
College
Military
Convents/monasteries
Transportation
Rivers/lakes
Dam locks/toll booths
Ferry/railroad/bus terminals
Commercial
Shopping centers
Central business districts
Commercial/industrial parks
Educational
Day care centers
Preschools/kindergartens
Elementary/secondary schools
Vocational/business/specialty schools
Colleges/universities

(From Ref.[3].)

to prevent fires, explosions, or releases. It must also be able to limit the severity of such events when they do occur and terminate them as soon as possible. As Table 2 indicates, there are four principal methods of expedient hazard mitigation—leak control, spill control, fire control, and container stabilization.[12] *Leak control* limits the rate at which chemical product escapes from containment to the environment. In turn, there are two types of leak controls: direct controls and indirect controls. Direct controls restore the integrity of a compromised container by patching, plugging, overpacking, or crimping; indirect controls include product shut-off, product transfer, and product displacement.

Spill control limits the rate at which a chemical disperses through the environment. Gaseous releases to the atmosphere can be controlled by ventilation, dissolution, dispersion, and diversion. Liquid releases to ground can be controlled by diking, retention, adsorption, and neutralization. Liquid releases to water can be controlled by damming, diverting, booming, and absorption. Solid releases to ground can be controlled by blanketing. By contrast, *fire control* involves extinguishment, exposure (adjacent structures) protection, and controlled burn.

Finally, *container stabilization* attempts to restore an unstable container to a stable physical location or orientation. Container stabilization is principally used in transportation incidents, but natural hazards, such as severe inland flooding and earthquakes, can displace stationary tanks from their foundations.[13] As in the case of emergency assessment analyses, the EP coordinator must ensure the availability of adequate personnel, facilities, equipment, and materials to support expedient hazard mitigation.

CONDUCTING PERSONNEL AND POPULATION PROTECTION ANALYSES

As Table 2 indicates, the facility emergency response organization and the offsite emergency response organization have many personnel and population protection tasks to perform in a chemical emergency. Onsite actions focus on protecting members of the onsite emergency response organization, whereas offsite actions focus on prompt initiation of protective actions for the population at risk. Protection of the onsite emergency response organization requires the use of personal protective equipment (PPE) and the implementation of procedures to ensure the safety of those attempting to assess the incident, fight fires, prevent explosions, and control releases. In addition, it is important for facility EP coordinators to recognize that many major emergencies require prompt and effective protective action by local residents and special facility populations near the facility. Indeed, facility

Table 2 Generic emergency response functions

Response function	Onsite actions	Offsite actions
Emergency assessment	Threat detection, reconnaissance, and emergency classification Product monitoring Container monitoring Environmental monitoring Release monitoring Impact projection Damage assessment	Incident monitoring Environmental monitoring Population monitoring Impact projection Damage assessment
Expedient hazard mitigation	Leak control (patching, plugging, overpacking, crimping, product shut-off/transfer/displacement) Spill control (air ventilation, dissolution, dispersion, diversion; water damming, diverting, booming, absorption, diking, retention, adsorption, neutralization, surface blanketing) Fire control (extinguishment, controlled burn, exposure protection, withdrawal) Container stabilization	
Personnel and population protection	Impact (''hot/warm/cold'') zone access control and security Personal protective equipment Medical monitoring Hazard exposure control Search and rescue Decontamination First aid and transport of victims Emergency medical care and morgues	Protective action selection (evacuation, sheltering in-place) Population warning Protective action implementation (transportation support traffic management) Evacuation zone access control and security Search and rescue Reception and care of victims Emergency medical care and morgues Evacuation zone re-entry
Incident management	Agency notification and mobilization Mobilization of facilities and equipment Internal direction and control External coordination Public information Administrative and logistical support Documentation (incident data collection and after-action analysis) Incident recovery (resource assessment and replacement) Demobilization	Agency notification and mobilization Mobilization of facilities and equipment Internal direction and control External coordination Public information Administrative and logistical support Documentation (incident data collection and after-action analysis) Incident recovery (resource assessment and replacement) Demobilization

personnel might sometimes need to warn offsite populations at the same time as they notify local officials of an emergency. Consequently, EP coordinators must understand the process for offsite population protection. First, as Table 2 indicates, one or more protective action recommendations must be selected that generally involves a choice between evacuation and sheltering in-place. The most common protective action for environmental hazards, evacuation, appears deceptively simple: just warn everyone in the risk area to leave. Indeed, a rapid evacuation is relatively easy to achieve when the risk area population is small, all evacuating households are united and have their own vehicles, and the capacity of the evacuation route system is high.[3,14,15] However, evacuation can take many hours to clear the risk area when the population is large in relation to the capacity of the evacuation route system. Indeed, evacuation time estimates for some major urban areas around commercial nuclear power plants, where the VZs are 10 miles in radius, have been

estimated to exceed 30 hr.[16] By contrast, sheltering in-place is the most common protective action recommendation for some hazards (e.g., tornadoes), but the criteria for choosing between evacuation and sheltering in-place can be complex for chemical emergencies. For further discussion, see Ref.[3].

In addition, the risk area population must be warned about the hazard. The seven primary warning mechanisms are face-to-face warnings, mobile loudspeakers, sirens, commercial radio and television, tone alert radio, newspapers, and telephones.[3] These warning mechanisms differ in their precision of dissemination, penetration of normal activities, specificity of the message, susceptibility to message distortion, rate of dissemination over time, receiver requirements, sender requirements, and feedback (verification of receipt). EP coordinators should work with local emergency managers to select the most appropriate warning mechanisms based on the characteristics of the jurisdiction (e.g., population density and wealth) and the hazards to which it is exposed (especially speed of onset and scope of impact).

Unlike the other emergency response functions, population protection does not require the EP coordinator to ensure the availability of adequate personnel, facilities, equipment, and materials because this is the responsibility of the offsite emergency response organization. Nonetheless, facilities sometimes provide technical or financial support for the development of warning systems and implement these warning systems in rapidly escalating incidents.

CONDUCTING INCIDENT MANAGEMENT ANALYSES

As Table 2 indicates, incident management involves a series of tasks that are virtually identical for both the onsite and the offsite emergency organizations. The similarity in tasks for the two otherwise quite different organizations arises from the fact that incident management is the function that copes with the response-generated demands of the emergency. Although the requirements of most tasks listed in Table 2 are self-evident, it is important to distinguish internal direction and control from external coordination because the former refers to authority relations with teams inside the facility emergency response organization, whereas the latter refers to collaborative (nonhierarchical) relationships with offsite organizations. Similarly, it is important to understand that population warning disseminates information about the hazard, protective action recommendations, and sources of further information to those at risk. By contrast, public information is directed toward those segments of the population that are *not* at risk. As is the case with the other emergency response functions, the EP coordinator must ensure the availability of adequate personnel, facilities, equipment, and materials to support incident management.

DEVELOPING PLANS AND PROCEDURES

For many years, the chemical industry has provided technical guidance to support the development of facility emergency response plans. For example, the Chemical Manufacturers Association *CAER Program Handbook*[17] defined the planning standards for an emergency response plan as including: 1) assignment of organizational responsibilities; 2) risk evaluation; 3) notification procedures and communication systems; 4) emergency equipment and facilities; 5) assessment capabilities; 6) protective action procedures; 7) public education and information; 8) postemergency procedures; 9) training and drills; and 10) program maintenance. In this regard, the *CAER Program Handbook* complements federal government guidance that is oriented primarily toward local government agencies for the development of radiological emergency response plans,[18] chemical incident response plans,[7] or all-hazards emergency operations plans.[19]

More recent guidance has promoted the adoption of the incident command system/incident management system (ICS/IMS), which seeks to establish a universal command structure for emergency response under the authority of a single incident commander.[20] The incident commander (IC) can be supported by a unified command consisting of representatives from other response organizations (e.g., local fire chief, state environmental protection representative, and federal coordinating officer). The IC directs an emergency response organization consisting of five sections. Command consists of the IC and the information, safety, and liaison functions. The operations section, which is responsible for tactical operations, operates a staging area for mobilizing personnel reporting to the scene and can comprise divisions (defined geographical areas), functional groups (assigned to specific tasks such as leak control, evacuation management, and emergency medical services), or combinations of units in task forces and strike teams. The Planning section has a resources unit, situation unit, demobilization unit, and documentation unit. The Logistics section has a service branch and a support branch. The service branch contains a communications unit, a medical unit, and a food unit, whereas the support branch contains a supply unit, a facilities unit, and a ground support unit. Finally, there is a finance and administration section that has a time unit, a procurement unit, a compensation and claims unit, and a cost unit.

Incident command system/incident management system provides two important advantages—unified command and standardization of roles. Facility EP Coordinators should understand how their emergency response organizations will link with the local version of ICS/IMS adopted by local government agencies so that the onsite and the offsite organizations can work together effectively. For further discussion of ICS/IMS, see Refs.[21–23].

ESTABLISHING EOCs

Emergency Operations Centers are facilities that provide technical assistance to emergency responders at the scene of an incident. Emergency operations centers, which are permanently located in areas expected to be safe from hazard exposures, are important because the resources needed to respond to a major incident can be widely dispersed; hence, the EOC relieves the IC and his/her command staff of the need to locate them. Moreover, many organizations participate in the incident response and each organization must have a capability for obtaining and processing timely information about onsite and offsite aspects of the incident. This capability is established by collocation of senior decision makers with telecommunications and information processing equipment in the EOC. An EOC should be designed to provide enough space to house the emergency response functions taking place within it and allocate this space so that its staff are located close to the equipment, information, and materials they need. For further discussion of EOC design, see Ref.[24].

CONDUCTING AND EVALUATING TRAINING, DRILLS, AND EXERCISES

Training provides the means by which emergency responder skills are upgraded; hence, the facility emergency response plan should describe the content and the frequency of the emergency training to be provided for all response personnel—including members of the senior management team. Classroom instruction should include discussion of each individual's assigned tasks and a review of the procedures for which each individual is responsible. To promote flexibility in organizational response, all emergency personnel should receive an explanation of the nature of the hazard, measures they should take to ensure their own personal protection, and an overview of the emergency response plan and the rationale for its components.[25]

Once a training program has been implemented, it must be evaluated using drills, exercises, and (occasionally) responses to actual incidents. Drills and exercises involve the measurement of performance over varying degrees of scope. Drills usually involve the performance of a single individual or a small team over a period of minutes to hours, whereas exercises and incidents involve the performance larger organizations and multiorganizational networks over a period of hours to days.

In preparing for drills and exercises, the first task is to specify clearly what is to be tested.[26] Typically, drills and exercises are used to test people, facilities, and equipment on tasks that are difficult, critical, and are performed infrequently. Exercises are more comprehensive than drills because they are used to test people's ability to perform both taskwork and teamwork. While the former is obviously the ability to competently perform each separate aspect of the emergency response, the latter is the ability to allocate resources and schedule tasks to achieve a coordinated performance that is efficient, effective, and timely.[27]

Walk-through drills should be conducted to ensure procedures are current (e.g., testing notification procedures to verify telephone numbers are current), or determine the capabilities of individuals or small teams to perform their designated emergency duties. In tabletop exercises, agency heads and assistants gather informally to discuss solutions to very general emergency situations presented by the local emergency manager. Response actions are described but not actually implemented.

Functional exercises are conducted using the staff from a single department to perform a limited number of specific tasks in response to a more detailed scenario than is used in drills. By contrast, full-scale exercises involve staff from many departments performing a wide range emergency response functions. The scenario for a full-scale exercise might address initial detection, emergency classification, activation of emergency response organizations, continuing emergency assessment, communication and coordination with other responding organizations, protective response (including protective action decision making and implementation, and medical support), and termination of the emergency. The scenario for a full scale exercise is often very detailed and is usually constructed by a training committee or an external consultant. Functional exercises and full scale exercises can be announced in advance, but should be based on emergency scenarios whose contents are withheld from all responding personnel.

All three forms of exercises (tabletop, functional, and full-scale) and incident responses will benefit from an immediate oral and later written critique by the players, controllers, and evaluators.[26] The discussions should address whether the response was consistent with the emergency plan and procedures and, if it is an exercise critique, whether the exercise objectives

were met. If there were deviations from the emergency plan and procedures, the participants should discuss why this occurred. In some cases, the conclusion will be that the emergency plan or procedures need to be revised. In other cases, the solution will be to reassign personnel, improve personnel training, or upgrade facilities and equipment. The results of the critique should be documented in a written report that contains specific recommendations for action, assignment of responsibility for implementation, and a schedule for completion. More detailed guidance on this process can be found in Refs.[26,28].

DISSEMINATING RISK INFORMATION

Risk communication should be a process by which stakeholders share information about hazards affecting a community.[29] Stakeholders include businesses and households that are vulnerable to a specific hazard, as well as community and industry personnel who are responsible for managing a hazard in ways that reduce the risk to an acceptable level. Communication among stakeholders can be analyzed in terms of who (source) says what (message), via what medium (channel), to whom (receiver), and directed at what kind of change (effect). Sources (authorities, news media, and peers) are perceived primarily in terms of expertise and trustworthiness. Messages vary in their content—especially their information about a hazard, its characteristics (e.g., magnitude, location, and time of impact), alternative protective actions and their attributes (efficacy; cost; safety; and requirements for time and effort, knowledge and skill, tools and equipment; and cooperation from others).

The information channels available for use include print media such as newspapers, magazines, and brochures; electronic media such as television, radio, telephone, and the internet; and face-to-face interaction through personal conversations and public meetings. The distinctions among these information channels are important because they differ in the ways in which they accommodate the information processing activities of the receivers. For example, orally presented information is ephemeral and is easily lost unless otherwise recorded, whereas written information inherently provides a record that can be referred to at a later time.

Receivers differ in many respects, but the most important of these are psychological characteristics that have direct effects on the communication process. For example, receivers differ in their perceptions of source credibility, access to communication channels, prior beliefs about hazards and protective actions, ability to understand and remember message content, and access to resources needed to implement protective action. The effects of a message on a receiver include attention, comprehension, acceptance, retention, and action.

Finally, feedback is an important component of the communication model because some attempts are unidirectional, whereas others are interactive. Unidirectional communications appeal to many EP coordinators because they appear to be less time consuming and sometimes this actually is the case. Frequently, however, interactive communication is needed for receivers to indicate that they have not comprehended the message that was sent or that the message sent by the source did not satisfy their information needs.

It is important to recognize that risk information is transmitted through social networks.[2] Thus, a warning message is likely to be relayed from the original source (e.g., an authority) through an intermediary (e.g., a risk area resident) to an ultimate receiver (e.g., a peer such as a friend, relative, neighbor, or coworker) in addition to—or even instead of—being transmitted directly from the original source to the ultimate receiver. In addition, warning messages are often transmitted through peer networks and requests for confirmation or further information transmitted back to authorities and the news media.

There are six basic functions that should be addressed in a community risk communication program.[29] These are strategic analysis, operational analysis, resource mobilization, program development, program implementation for the continuing hazard phase, and program implementation for the escalating crisis and emergency response phases.[29] The first function, strategic analysis, involves four tasks. The first task is for facility EP coordinators to conduct hazard/vulnerability/risk analyses to identify the facility's hazards and the neighborhoods nearby that are at greatest risk. During the second task, EP coordinators should become aware of the community context to identify emergency response relevant variables such as the ethnic composition, communication channels, perceptions of authorities, and level of education in nearby neighborhoods. The third task is for EP coordinators to work with the LEPC to identify the community's prevailing perceptions of environmental hazards and hazard adjustments. During the fourth task, EP coordinators should set appropriate goals for the risk communication program, beginning with establishing a basic awareness of the hazards to which the facility exposes the community, continuing through explanations of alternative hazard adjustments, and concluding with development of active citizen support for the LEPC.

Operational analysis comprises five tasks, the first of which is to identify and assess feasible hazard adjustments for the community and its households/businesses. To accomplish this objective, EP coordinators

can access resources such as the American Red Cross web site at http://www.redcross.org/services/disaster/beprepared, where they can find information about recommended household adjustments for a wide range of hazards. During the second task, EP coordinators should work with the LEPC to identify ways to provide incentives, sanctions, and technological innovations for these hazard adjustments. During the third task, EP coordinators should work with the LEPC to identify the available risk communication sources in the community. In choosing among authorities (local, state, and federal government agencies; facility operators; and scientists), news media (especially newspapers, television, and radio), and peers (friends, relatives, neighbors, and coworkers), EP coordinators should recognize that official sources are generally the most credible and message recipients infer credibility from the source's credentials (e.g., job title and educational degrees), acceptance by other sources of established credibility, or previous history of job performance.[3] However, the degree of expertise attributed to different sources varies from one hazard to another and by gender, ethnicity, and other demographic characteristics.[30–32] During the fourth task, EP coordinators should work with the LEPC to identify the available risk communication channels in the community from electronic (radio, television, and, increasingly, web sites) and print (newspapers, magazines, brochures, posters, telephone book inserts, comic/coloring books, and reports) media. Additional communication channels include informal face-to-face conversations (drop-in hours at local libraries, newsletters, information booths at local events, and shopping malls) and formal meetings with or without audiovisual presentations such as computer simulations, slide shows, and films. During the fifth task, EP coordinators should work with the LEPC to identify the differences among audience segments in terms of their access to and preferences for different types of media (e.g., radio) and specific channels (e.g., specific radio stations) within each medium, and if there are any audience segments with language barriers.

There are five steps in *resource mobilization*. During the first task, EP coordinators should obtain the support of senior management by "selling" the importance of community hazard vulnerability and identifying hazard mitigation, emergency preparedness, emergency response, and disaster recovery as effective solutions. In the second task, EP coordinators should enlist the participation of other departments in their facilities and work with the LEPC to do the same with government agencies. EP coordinators should adopt a collaborative strategy to ensure all chemical facilities within a community are aware of any risk communication programs being planned and implemented by other organizations and to develop a coalition that pools their resources.[33–35] During the third task, EP Coordinators should work with the LEPC to enlist the participation of nongovernmental (nonprofit) and private sector organizations—many of which can identify geographic areas in which there is a high concentration of vulnerable households, and which can also assist these households to prepare for emergencies. The fourth task is for EP coordinators to work through the LEPC to develop active contact with reporters and editors who can provide access to channels citizens routinely use. The fifth task involves having EP coordinators work through the LEPC with neighborhood associations and service organizations to facilitate their organizational effectiveness and to provide them with opportunities to learn about environmental hazards and feasible adjustments to those hazards.

Program development for all phases involves five tasks. During the first task, EP Coordinators should staff, train, and exercise a crisis communications team. This team forms a critical link between technical experts and the population at risk; hence, its members should have skills in communicating with both groups. In addition, the crisis communications team should be represented by a spokesperson who is technically competent to explain emergency conditions clearly and will be perceived as credible because of relevant credentials (e.g., job title and educational degrees), acceptance by other sources of known credibility, or a demonstrated history of job performance that has enhanced credibility.[3,36] It will also be helpful if this spokesperson receives training from public relations experts.[37] The second task is for EP coordinators to establish procedures for maintaining an effective communication flow in an escalating crisis and in emergency response to ensure each organization receives all the information it needs as promptly as possible. In the third task, EP coordinators work with the LEPC to develop a comprehensive risk communication program that presents information about hazards and hazard adjustments in a form that attracts attention and is easily understood and retained. In addition, this information should be repeated over time and emphasize the personal consequences of environmental hazards, the need to accept personal responsibility for protective action, the efficacy of alternative hazard adjustments, and information about hazard adjustments' resource requirements. During the fourth task, EP coordinators should encourage the LEPC to use informal communication networks in the community. The fifth task is for EP coordinators to establish procedures for obtaining feedback from the news media and the public. Feedback is usually limited in public hearings; hence, many scholars recommend informal channels of communication such as advisory panels and meetings with neighborhood associations and civic organizations.[37–39]

During emerging crises, feedback can be obtained by reading copies of local newspapers, and monitoring radio and television broadcasts. In addition, EP coordinators can work with the LEPC to obtain feedback from citizens via rumor control centers with a telephone number or a web site address that has been publicized in advance.

There are five steps to *program implementation for the continuing hazard phase*. During the first task, EP coordinators should build source credibility by increasing perceptions of expertise and trustworthiness. Thus, members of the crisis communication team should ensure its procedures are coordinated with all relevant agencies' emergency operation plans, and emergency personnel from each facility or agency should develop a demonstrated history of effective job performance during minor incidents such as severe storms and minor floods. To enhance trustworthiness, messages must be perceived as accurate, objective, and complete[40,41] and the source is perceived as competent, open and honest, caring and concerned, and sympathetic.[42] These qualities can be enhanced when EP coordinators promote meaningful public involvement by working with the LEPC to involve the community in the continuing hazard phase, avoid secret meetings, explain constraints in risk communication procedures (especially what constraints on public participation are imposed by law and agency policy), and provide accurate information that is responsive to people's requests. The second task is for EP coordinators to use a variety of channels to disseminate hazard information because each channel can affect a different stage of information processing. For example, radio or television ''spots'' might have their greatest impact in establishing initial hazard awareness (i.e., attracting attention to the hazard) and maintaining its intrusiveness by means of frequent thought and discussion. By contrast, printed materials are most effective in providing the detailed information needed to establish a perception of threat and identifying suitable hazard adjustments. In the third task, EP coordinators describe community or facility hazard adjustments being planned or implemented. In particular, local residents should be informed of any hazard mitigation actions being taken so they will understand their risk is being reduced. In the fourth task, EP coordinators should work with the LEPC to describe feasible household hazard adjustments risk area residents can take to protect themselves. For example, households can prepare for airborne releases of toxic chemicals by reducing air infiltration in their homes,[3] or drinking bottled or boiled water if local wells are contaminated. The fifth task is for EP coordinators to evaluate program effectiveness by measuring the degree to which it has achieved its objectives.[43] Thus, they should determine how to measure the goals they have set, how to collect the data needed to measure performance, and how to analyze these data. This comparison process can then serve as the basis for determining whether changes need to be made in the risk communication program.

Program implementation for the escalating crisis and emergency response phases involves six activities. The first task is for EP Coordinators to work with their plant managers and public information managers to classify the situation in terms of its severity. To a great extent, ''perception is reality,'' so a crisis exists if facility operators, civil authorities, the news media, or a significant proportion of those in the community believe a situation is dangerous. Nonetheless, facility personnel can exert some control over other people's definition of the situation by establishing specific criteria in advance of an incident that systematically define elevated conditions of threat. To accomplish the second task, EP coordinators should activate the crisis communication team promptly so that its members can make all appropriate contacts and open all necessary communication links. This makes it possible for all organizations to be aware of the information being disseminated by other organizations, identify any disagreements, and prepare appropriate explanations before they are contacted by the news media. During the third task, facility personnel should determine the appropriate time to release sensitive information by developing procedures that define when information is to be released. This is a challenging task because even experts disagree on the rules for determining when to release information.[44] On the one hand, early releases of information are often characterized by a significant degree of uncertainty; hence, there is a possibility that crisis conditions might never materialize or that they will be less severe than initially expected. However, withholding information to avoid unnecessary disruption can be misinterpreted as a ''cover-up'' if the data are leaked.[37] The fourth task involves having facility personnel select the communication channels appropriate to the situation. The newsworthiness of an escalating crisis ensures there will be little difficulty in obtaining news media coverage, but facility personnel need to promote dialogue through two-way communication in small groups as well as in press conferences.[37] The fifth task is for EP coordinators to maintain source credibility with the news media and the public. If the available data are incomplete, facility personnel should be honest about what is and is not known. In addition, they should recognize that the news media have many other sources of information; hence, it is important to respond promptly to reporters facing imminent deadlines.[45] To accomplish the sixth task, facility personnel should provide timely and accurate information to the news media and the public. In particular, news releases should be no longer than two

pages with simple short sentences in plain English. They should contain a dateline (date and location of release), the organizational source (including point of contact) for the information, a summary lead that provides a one sentence abstract of the press release, the text of the press release, and a brief description of any attachments.[45] These should be supplemented by fact sheets containing basic background information that is appropriate to any incident. Facility personnel should state that they do not know the answer to a question when this is the case, but should make a commitment to answer the question at a later time. Last, they should evaluate performance through postincident critiques that permit all members of the crisis communication team to review the goals of the risk communication program, the event logs kept during the incident, and other available documentation to identify deficiencies in organizational performance. Each participant should be encouraged to address any deficiencies by recommending improvements in plans, procedures, training, facilities, equipment, or materials and supplies.

CONCLUSIONS

Chemical risks can be managed through effective programs for onsite and offsite emergency preparedness. Such programs begin by anticipating incident demands and developing an emergency preparedness program to meet these demands. Facility emergency coordinators should conduct hazard/vulnerability/risk analyses, emergency assessment analyses, expedient hazard mitigation analyses, personnel and population protection analyses, and incident management analyses to guide the development of their emergency response plans. In addition, they should use established methods for developing plans and procedures; establishing EOCs; conducting and evaluating training, drills, and exercises; and disseminating risk information.

ACKNOWLEDGMENT

This entry is based upon work supported by the National Science Foundation under Grant CMS-0219155. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Quarantelli, E.L. Disaster planning: small and large—past, present, and future. Proceedings: American Red Cross EFO Division Disaster Conference; American Red Cross Eastern Field Office: Alexandria, VA, 1981.
2. Drabek, T.E. *Human System Responses to Disaster: A Sociological Perspective*; Springer-Verlag: New York, 1986.
3. Lindell, M.K.; Perry, R.W. *Behavioral Foundations of Community Emergency Planning*; Hemisphere Press: Washington, DC, 1992.
4. Lindell, M.K.; Perry, R.W. Identifying and managing conjoint threats: earthquake-induced hazardous materials releases in the U.S. J. Hazard. Mater. **1996**, *50*, 31–46.
5. National Fire Protection Association. *NFPA 1600 Standard on Disaster/Emergency Management and Business Continuity Programs, 2004 Edition*; National Fire Protection Association: Quincy, MA, 2004.
6. Federal Emergency Management Agency. *The Emergency Program Manager*; Federal Emergency Management Agency: Washington, DC, 1993.
7. National Response Team. *Hazardous Materials Emergency Planning Guide, NRT-1*; National Response Team: Washington, DC, 1987; 28–34.
8. U.S. Environmental Protection Agency, Federal Emergency Management Agency, and U.S. Department of Transportation. *Technical Guidance for Hazards Analysis*; U.S. Environmental Protection Agency, Federal Emergency Management Agency, and U.S. Department of Transportation: Washington, DC, 1987; 28–34.
9. Federal Emergency Management Agency, U.S. Department of Transportation, and U.S. Environmental Protection Agency. *Handbook of Chemical Hazard Analysis Procedures*; Federal Emergency Management Agency, U.S. Department of Transportation, and U.S. Environmental Protection Agency: Washington, DC.
10. Lindell, M.K. Assessing emergency preparedness in support of hazardous facility risk analyses: an application at a U.S. hazardous waste incinerator. J. Hazard. Mater. **1995**, *40*, 297–319.
11. Lindell, M.K. Hazardous materials. In *Planning and Urban Design Standards*; American Planning Association, Ed.; John Wiley and Sons: New York, 1995.
12. Lesak, D.M. *Hazardous Materials: Strategies and Tactics*; Prentice-Hall: Upper Saddle River, NJ, 1999.
13. Lindell, M.K.; Perry, R.W. Hazardous materials releases in the Northridge earthquake: implications for seismic risk assessment. Risk Anal. **1997**, *17*, 147–156.
14. Lindell, M.K.; Prater, C.S.; Perry, R.W.; Wu, J.Y. *EMBLEM: An Empirically-Based Large Scale Evacuation Time Estimate Model*; Texas A&M

University Hazard Reduction & Recovery Center: College Station, TX, 2002, http://www.txdps.state.tx.us/dem/documents.htm.
15. Urbanik, T. Evacuation time estimates for nuclear power plants. J. Hazard. Mater. **2000**, *75*, 165–180.
16. Urbanik, T. *An Analysis of Evacuation Time Estimates Around 52 Nuclear Power Plant Sites: Analysis and Evaluation*; NUREG/CR-1856; U.S. Nuclear Regulatory Commission: Washington, DC, 1981.
17. Chemical Manufacturers Association. *Community Awareness & Emergency Response: Program Handbook*; Chemical Manufacturers Association: Washington, DC, 1985.
18. U.S. Nuclear Regulatory Commission. *Criteria for Preparation and Evaluation of Radiological Emergency Response Plans and Preparedness in Support of Nuclear Power Plants*; NUREG-0654, FEMA-REP-1; U.S. Nuclear Regulatory Commission: Washington, DC, 1980.
19. Federal Emergency Management Agency. *Guide for All-Hazard Emergency Operations Planning*; SLG-101; Federal Emergency Management Agency: Washington, DC, 1996.
20. Perry, R.W.; Lindell, M.K. Responding to disasters. In *Emergency Management: Principles and Practice for Local Government*, 2nd Ed.; Tierney, K., Waugh, W., Eds.; International City Management Association: Washington, DC, 1996.
21. Lindell, M.K.; Prater, C.S.; Perry, R.W. *Emergency Management Principles and Practices*; John Wiley and Sons: Hobeken, NJ, 2006.
22. Brunacini, A.V. *Fire Command: The Essentials of IMS*; National Fire Protection Association: Quincy, MA, 2002.
23. Erickson, P.A. *Emergency Response Planning for Corporate and Municipal Managers*; Academic Press: San Diego, CA, 1999.
24. Lindell, M.K.; Perry, R.W. Emergency planning and preparedness. In *Emergency Management: Principles and Practice for Local Government*, 2nd Ed.; Tierney, K., Waugh, W., Eds.; International City Management Association: Washington, DC, 1996.
25. Ford, J.K.; Schmidt, A. Emergency preparedness training: strategies for enhancing real-world performance. J. Hazard. Mater. **2000**, *75*, 195–215.
26. National Response Team. *Developing a Hazardous Materials Exercise Program: A Handbook for State and Local Officials, NRT-2*; National Response Team: Washington, DC, 1990.
27. McIntyre, R.M.; Salas, E. Measuring and managing for team performance: lessons from complex environments. In *Team Effectiveness and Decision Making in Organizations*; Guzzo, R.A., Salas, E., Eds.; Josey-Bass: San Francisco, 1995; 9–45.
28. Emergency Management Institute. In *Exercise Design*, IS-139; Federal Emergency Management Agency Emergency Management Institute: Emmitsburg, MD, 2003.
29. Lindell, M.K.; Perry, R.W. *Communicating Environmental Risk in Multiethnic Communities*; Sage: Thousand Oaks, CA, 2004.
30. Nigg, J.M. Awareness and behavior: public response to prediction awareness. In *Perspectives on Increasing Hazard Awareness*; Saarinen, T.F., Ed.; University of Colorado Institute of Behavioral Science: Boulder, CO, 1982; 36–51.
31. Perry, R.W. Racial and ethnic minority citizens in disasters. In *The Sociology of Disasters*; Dynes, R., Pelanda, C., Eds.; Franco Angelli: Gorizia, Italy, 1987; 87–99.
32. Perry, R.W.; Nelson, L. Ethnicity and hazard information dissemination. Environ. Manag. **1991**, *15*, 581–587.
33. Drabek, T.E. *Emergency Management: Strategies for Maintaining Organizational Integrity*; Springer-Verlag: New York, 1990.
34. Gillespie, D.F.; Colignon, R.A.; Banerjee, M.M.; Murty, S.A.; Rogge, M. *Partnerships for Community Preparedness*; University of Colorado Natural Hazards Research and Applications Information Center: Boulder, CO, 1993.
35. Lindell, M.K.; Whitney, D.J.; Futch, C.J.; Clause, C.S. The Local Emergency Planning Committee: a better way to coordinate disaster planning. In *Disaster Management in the U.S. and Canada: The Politics, Policymaking, Administration and Analysis of Emergency Management*; Sylves, R.T., Waugh, W.L., Jr., Eds.; Charles C. Thomas Publishers: Springfield, IL, 1996.
36. Perry, R.W.; Lindell, M.K. *Living with Mt. St. Helens: Human Adjustment to Volcano Hazards*; Washington State University Press: Pullman, WA, 1990.
37. Hance, B.; Chess, C.; Sandman, P. *Improving Dialogue with Communities*; New Jersey Department of Environmental Protection: New Brunswick, NJ, 1988.
38. Committee on Risk Perception and Communication. *Improving Risk Communication*; National Academy of Sciences: Washington, DC, 1989.
39. Covello, V.S. Case studies of risk communication: introduction. In *Risk Communication*; Davies, J.C., Covello, V.T., Allen, F.W., Eds.; The Conservation Foundation: Washington, DC, 1987; 63–65.
40. Meyer, P. Defining and measuring credibility of newspapers: developing an index. Journalism Q **1988**, *65*, 567–574, 588.

41. Trumbo, C.W.; McComas, K.A. The function of credibility in information processing for risk perception. Risk Anal. **2003**, *23*, 343–353.
42. Maeda, Y.; Miyahara, M. Determinants of trust in industry, government and citizen's groups in Japan. Risk Anal. **2003**, *23*, 303–310.
43. Stallen, P.J.M. Developing communications about risks of major industrial accidents in the Netherlands. In *Communicating Risks to The Public: International Perspectives*; Kasperson, R.E., Stallen, P.J.M., Eds.; Kluwer Academic Publishers: London, 1991; 55–66.
44. Kasperson, R. Panel discussion on "trust and credibility: the central issue?" In *Risk Communication*; Davies, J.C., Covello, V.T., Allen, F.W., Eds.; The Conservation Foundation: Washington, DC, 1987; 43–62.
45. Churchill, R.E. Effective media relations. In *The Public Health Consequences of Disasters*; Noji, E.K., Ed.; Oxford University Press: New York, 1997; 122–132.

Oriented Morphologies: Development in Polymer Processing

Mario A. Perez
3M Company, St. Paul, Minnesota, U.S.A.

INTRODUCTION

Most of the key findings in polymer morphology development and its connection to property enhancement took place during the latter half of the 20th century. Polymer morphology is still a very active area of research as there is a great need for understanding macrostructure obtained with new materials and processes. This entry is an overview of morphologies that can be obtained in polymeric materials and their connection to useful commercial properties. It is limited to semicrystalline homopolymers. Blends and copolymers were omitted, as they should be treated separately. Development of methods for property improvement through morphology creation or modification has focused on linear chain polymers or thermoplastic semicrystalline materials. This is because of the nature of amorphous and thermosetting polymers. The study of extended amorphous regions is already encompassed within the framework of semicrystalline thermoplastics.

Thermosetting materials, as the name implies, cannot be easily deformed to attain an oriented morphology once crosslinking has taken place. The notable and well documented exception being the deformation of crosslinked rubbers. Furthermore, in uncrosslinked thermosets, molecular length is often insufficient to form metastable oriented morphologies. Within linear polymers or semicrystalline thermoplastics, one can find those that form liquid crystalline phases, spherulitic structures, self-assembled layers, hard segmental sections, functional bonding, and strong intermolecular secondary forces.

Polymers are commercially available in various forms, such as pellets, powders, solutions, and dispersions. In the undissolved state, most thermoplastic semicrystalline polymers are shaped by processing beyond their melting point, followed by forming and cooling. Several choices are available for drawing, but two are of great practical importance: a) extension in the molten state and rapid quenching and b) cooling to or below crystallization followed by deformation, with the two techniques applicable being melt drawing and solid state drawing, respectively. An array of macrostructures can be obtained by these techniques.

MORPHOLOGY DEVELOPMENT

The deformation of long chain polymer molecules has always been of great industrial interest as more value can be placed on a material that has improved properties. Molecular extension, or alternatively molecular orientation, is of particular interest as it can enhance mechanical properties of an otherwise weak polymer. Oriented materials are inherently anisotropic. These anisotropic regions can be found directly in semicrystalline polymers where chains organize themselves into crystalline domains.

Quiescent Systems

Polymer chains aggregate and fold spontaneously upon cooling from a molten or a diluted state. Evidence of folding in polymer single crystals was first presented by Keller.[1] Molecules aggregate side by side to form lamellar crystallites where the chain backbone is preferentially aligned in a given direction. It is customary to label the chain axis direction within the crystallite, the *c*-axis direction. This is the direction perpendicular to the plane of the crystallite. Typical lamellar thicknesses ranges from 100 to 200 Å. Chain folds are present on two sides of this lamellar crystallite. Fig. 1 shows a sketch and a micrograph of a polymer single crystal grown from a dilute solution. The lozenge shape is characteristic of low concentration solutions and lower crystallization temperatures. More dendritic structures are observed as the concentration increases and the growth temperature is higher.

Intermolecular forces hold the crystallite together. These forces are overcome by increased molecular vibration when the melting point is approached. Alternatively, these crystallites can be cleaved, deformed, aligned, or rotated upon mechanical loading. The maximum property attainable with any polymer system is that of a fully aligned chain where covalent linkages bear the load. Even though the properties of perfect polymer crystals have been measured, they cannot be fully realized in practical commercial processes. Among the factors that preclude attainment of maximum properties through molecular extension are chain

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120023858

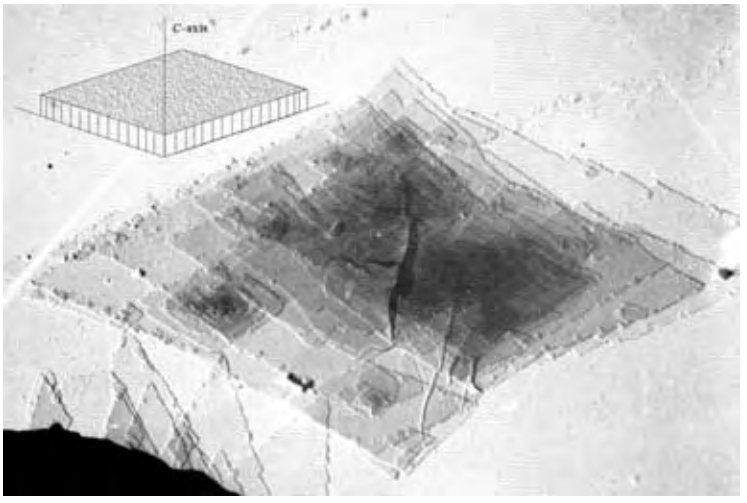


Fig. 1 Sketch of a polymer single crystal and electron micrograph of polyethylene single crystals (Marlex 50) grown from hot xylene. Chain backbones in the crystals are mostly aligned with the *c*-crystallographic axis. (From Ref.[1].)

entanglements (and other topological constraints), chain size distributions, chain slippage, tacticity defects, and kinetic constraints. The chemistry of linear polymers can also be tailored to provide inherent stiffness and strength.

Besides single crystals grown from solution, saturated polymer solutions and melts yield larger scale macrostructures under quiescent conditions. A very common macrostructure is spherulitic morphology. Spherulites are structures that grow radially from a single crystal or heterogeneity until their boundaries meet. Spherulite sizes vary a great deal; however, for most polymers typical sizes range from 1 to 20 μ. Semicrystalline thermoplastics like polyethylene and polypropylene readily form spherulitic structure upon cooling. More exotic polymers also form these structures. For example, spherulites of a crystalline diepoxide can be observed in Fig. 2. A maltese cross pattern, as shown in Fig. 2, is typical when spherulites are examined with a visible light microscope under cross polarization. A clover leaf pattern is formed instead with small angle light scattering.[2] The periodic nature of tangential polarizability from the center to the boundary of a spherulite is attributed to the helicoidal positioning of the crystalline lamellae around a radial axis, where the higher refractive index observed is the *c*-direction. In spherulites, the *c*-axis of lamellar regions is aligned tangentially to the radial direction in a spiral screw pitch staircase-like structure.

Besides those macrostructures obtained in the bulk, one can find oriented structures that originate at the boundaries from contact with other materials. This can happen during crystallization of a polymer against a steel surface, as in the case with the polymer processing practice of injection molding. It can also take place in composites where crystallization occurs near the fiber's boundary, or on a regular and crystalline face of another material. Such material may be a

Fig. 2 Visible light micrograph of spherulites from a crystalline diepoxide taken with a 5X lens and cross polarization. The observed maltese cross pattern arises from the spiral positioning of lamella along the radial growth direction. The high refractive index *c*-axis is tangential to the spherulite's radius.

mineral, an organic crystal, or a polymer. These morphologies originate from the difference in nucleation potential between the interface and the bulk. So, growth starts first at the interface and propagates into the bulk. In the case of contact with heterogeneous irregular or nonperiodic surfaces, the phenomenon is called transcrystallization. Transcrystalline behavior is widely known and has been reported for a variety of polymers and surfaces.[3,4] This phenomenon is of particular importance in fiber reinforced composites where the matrix polymer is semicrystalline. Spherulitic branches radiating away from the interface, coupled with an abundance of nucleation sites and tight packing, give rise to local oriented structures.

In cases where substrate surface periodicity influences the nature of the polymer crystalline structure obtained, the growth is called epitaxial. Polymer molecules are housed along the crystallographic plane of a given substrate providing an oriented morphology. For example, polyethylene can be crystallized in this manner using oriented polypropylene as the substrate. Polyethylene (1 0 0) planes align parallel to the (0 1 0) lattice plane of the oriented polypropylene monoclinic α-crystalline form.[5] The resulting structure of a high-density polyethylene crystallized on uniaxially oriented polypropylene is shown in Fig. 3. In this figure, the orientation direction of polypropylene is along the horizontal—polyethylene lamellae align at a 40° angle with the horizontal.[6] A related phenomenon, cross-hatching, is one where a bimodal lamellar orientation is observed with the same polymer and within the same spherulite.[7,8] Cross-hatching has been observed in the spherulitic structure of isotactic polypropylene that is generated by high temperature crystallization. In this case, orthogonal lathlike growths stemming from the radial structure are observed. It is implied from this discussion that a temperature gradient can either influence, or be the origin of an oriented morphology at an interface.

Within the boundaries of the quiescent methods discussed here, these oriented morphologies do not propagate far into the bulk, but typically only a few tenths of microns. So it has been of interest to develop methods to impart an oriented structure throughout the bulk as well. Usually, these methods involve polymer deformation or flow. The following sections cover a discussion of the most common situations where deformation provides oriented morphologies.

Deformation-Induced Morphologies

In semicrystalline polymers, crystallites and the macrostructures that they contribute to form, provide an added dimension to the consequences of imposed deformation.

Two events need to be managed during this deformation: chain extension (or unfolding) and chain relaxation. The ability to immobilize a chain after extension depends on several factors, such as drawing temperature in melts and polymer concentration in solutions. As mentioned earlier, random chains from melts and solutions minimize their energy by organizing themselves into chain-folded lamellae when

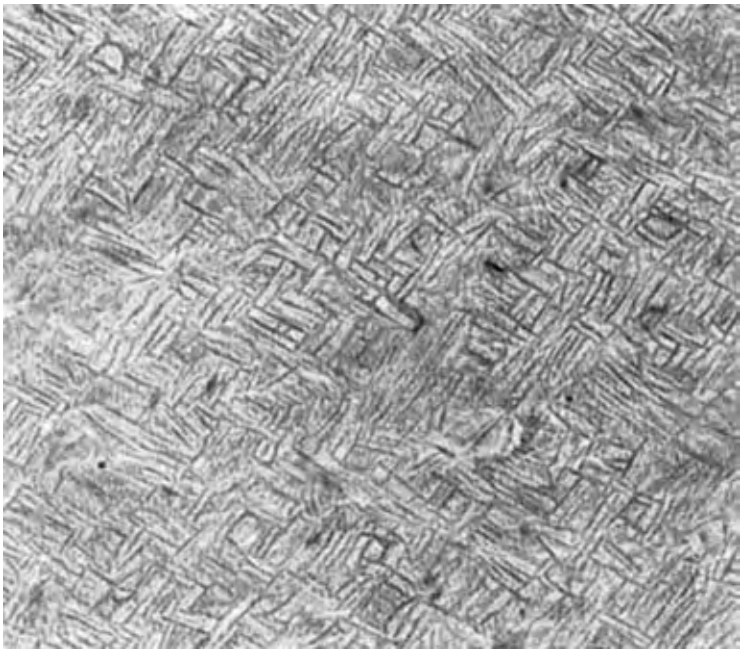


Fig. 3 Epitaxial growth of high-density polyethylene on oriented polypropylene. Polyethylene lamellae aligns at a 40° angle with the direction of orientation of the polypropylene. The orientation direction of polypropylene in the micrograph is along the horizontal. (From Ref.[6].)

conditions allow for good mobility. Chain-folded lamellae are observed in diluted polymer systems and molten polymers. The onset of lamellae formation depends on the concentration of the solution, molecular length, and cooling rate.

Flowing solutions

Fibrous crystals or crystalline bundles can be formed by the increasing alignment and proximity of chains undergoing macroscopic extension in solvent and melt systems. Studies of flow-induced morphologies in solvent systems date back to the mid-1960s and early 1970s. Beyond plain fibrillar structures, the first flow-induced polymeric structures with lamellar overgrowth resulted from experiments on cooling stirred dilute solutions.[9] These crystallized polymer structures obtained with nonquiescent solutions were termed ''shish-kebabs.'' In this composite structure, the ''shish'' are core fibril bundles and the ''kebabs'' are chain-folded lamellae that are formed periodically around the fibril core. An example of shish-kebab structure is shown in Fig. 4. The fibrillar core has been recognized to be composed of the longest molecules that precipitate from the solution first in this case. This fibrillar core then comprises elongated crystals that resemble a tight bundle. These bundles are connected with regions having a lower degree of order. The origin of the periodicity of the spacing between the lamellar kebabs is not clear, but it is probably dependent on the following factors: Strength of the flow field, frequency of the disordered tie regions in the fibrillar core, crystallization rate at the imposed temperature gradient, remaining unprecipitated molecular lengths, and chain stiffness. In some instances, the kebabs may not be connected to the fibrillar core. In addition, a distinction between regions within the main kebabs has been made, and macro- and microdomains have been reported.[10]

Melts

Studies of flow-induced morphologies in polymer solutions were crucial to the understanding of more complicated structures crystallized from flowing melts or deformed solids. Shish-kebab structures have been found to be present in melt drawn films.[11] Values of 20–35 nm were observed for the core fibrillar structures. Lamellar thicknesses in the range of 40–60 nm were obtained with scanning force microscopy. In addition to shish-kebabs, other structures can be prevalent in deformed melts such as needle crystal morphologies, oriented fringed micelles, and parallel stacked lamellae. Lamellar stacked and needle crystal morphologies are shown in Figs. 5[12] and 6.[13,14] In addition, another structure prevalent in polymer processing is the fringed micelle morphology.

A stacked lamellar morphology can be obtained by solid-state deformation of a crystallized polymer and also by freezing or capturing molecular orientation from the melt. In general, oriented fringed micelle structures can be obtained at deformation temperatures below or around the glass transition temperature and needle crystal morphologies through deformation of the polymer near its melting temperature.[13,14] As the chains between crystallites become taut and aligned, other phenomena can take place. For example, an apparent increase in overall crystallinity because of increased chain proximity and registry, and a departure from initial crystallite dimensions. Order is

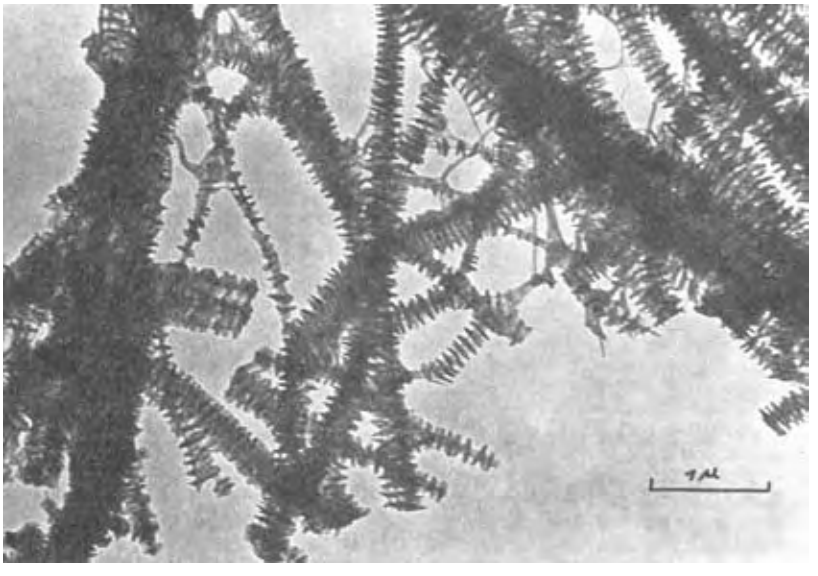


Fig. 4 Electron diffraction micrograph of polyethylene ''shish-kebab'' structure. (From Ref.[9].)

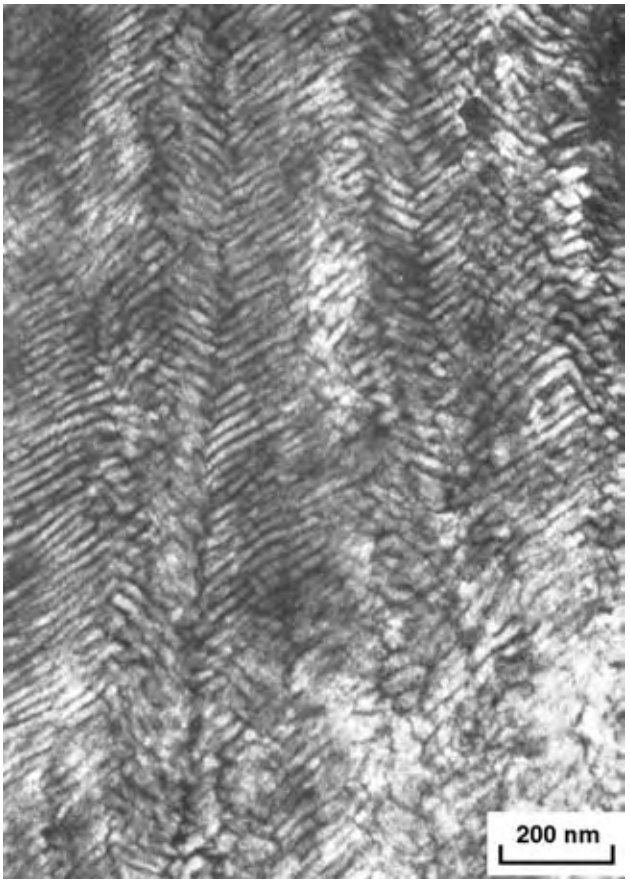


Fig. 5 Lamellar structure of polyethylene. (From Ref.[5].)

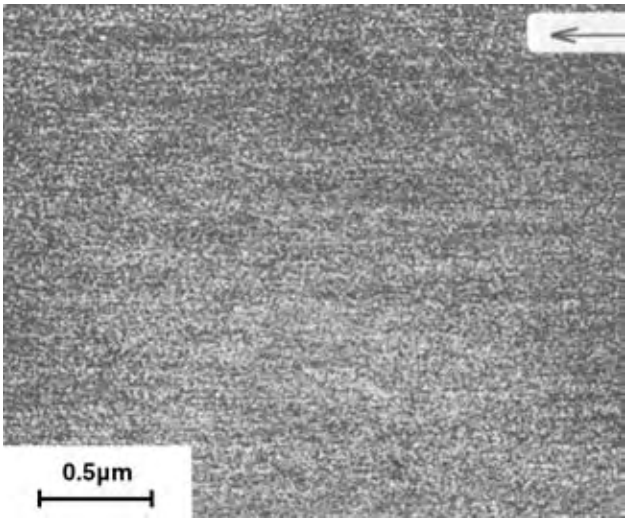


Fig. 6 Needle crystal morphology of polybutene-1. (From Ref.[13].)

increased further if these processes are followed by an annealing step.

Overall crystallinity increases observed in oriented morphologies are accompanied by melting point elevations. Elevated melting is detected for semicrystalline polymers that undergo processes that extend chains beyond their kinetically restricted coiled conformations. In thermodynamic terms, entropy is lowered upon chain extension. Chain extension leads to a shift in melting to higher temperatures and larger endotherm area. A larger endotherm area implies a higher degree of apparent crystallinity. This effect is shown in Fig. 7, which shows differential scanning calorimetry traces for polypropylene. Smaller crystallites melt first. Crystallites with higher dimensions along the chain direction are more resistant to melting, such as in the case of needle crystals.

METHODS

Uniaxial Extension

A common way to produce oriented morphologies or attain enhanced properties is to mechanically deform the polymer in one direction. This is uniaxial extension or draw. To attain molecular extension, imposed stress needs to transfer from the outer layers of the polymer sample or melt to the individual chains. This can happen through frictional contact among chains or between a chain and its surrounding medium. In the solid state, the contact is influenced by chain stiffness, functional groups, molecular length, charges, and aggregates (crystallites or heterogeneous particles). In the most basic of cases such as parallel plate flow, deformation of the fluid can come from simple shear by setting one of the plates in motion. Momentum transfers to the polymer's layers that are in contact with the moving plate as the fluid velocity at the wall is zero. If the channel is large enough, shear flow evolves in a flow with a strong rotational component. Therefore, polymer orientation through shear flow is not necessarily effective. An elongational flow field is the best way to extend a polymer molecule. Because of the nature of this elongational flow field, there is an inherent material acceleration, i.e., a chain can be extended more effectively as it proceeds downstream. In these flows, higher molecular weight chains—there is an upper limit to this—extend at lower strain rates, while lower molecular weight entities need to accelerate downstream in the flow field further to extend under high strain rates.

In uniaxial macroscopic deformation, the term draw ratio (DR) is customarily used to denote nominal draw (not molecular draw), or the change from the sample's initial dimension to its deformed state at each drawing step.[15] Even though this is by no means a rigorous definition, it applies to deformations induced during simple sample extension or during stretching by differential roll speed. DR assumes full extension efficiency.

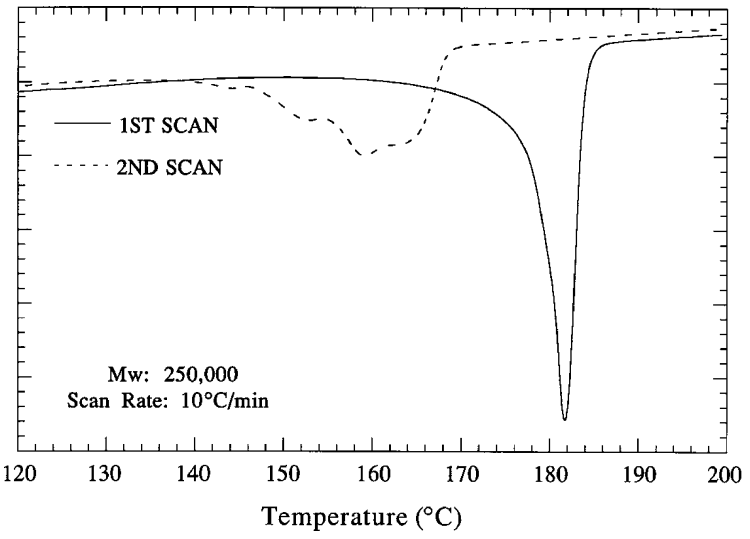


Fig. 7 Differential scanning calorimetry traces of an isotactic polypropylene. The first heating trace exhibits a sharp high melting transition with larger endotherm area and higher apparent crystallinity. Sample: Uniaxially oriented rod with high aspect ratio crystallites. The second heating trace shows the unioriented specimen's melting endotherm.

This is not necessarily the case at the molecular level where the efficiency of permanent chain extension can be influenced by factors like chain relaxation, topological restraints, chain size and distribution, dissipation of frictional heat, chain repeating structure, and chain slippage. An expression that accounts for chain slippage and provides a measure of true molecular extension has been proposed by Porter et al.:[16]

$$\text{MDR} = [(L_t - L_s)/L_o)] + 1$$

where L_t is the total length after full deformation; L_s the length of sample after shrinkage on heating beyond melting or glass transition; and L_o the length before deformation.

Affine deformation takes place when molecular draw ratio (MDR) = DR.[15] This is true in the absence of chain slippage. In a typical sample with a wide distribution of molecular lengths, this would be an average molecular DR. In practice, macroscopically uniform orientation properties are hard to obtain, because only sample surfaces and ends are in contact with rolls or clamping devices to effect extension. For example, during roll drawing, differences in oriented morphologies are produced when one side of the film is in contact with a heated roll and the other side is exposed to air or heat transfer medium. Nonuniformities in roll heating or surface finish differences may also affect final structure and property development.

Common uniaxial deformation methodologies

The most common type of uniaxial extension can take place during simple tensile drawing of cast polymeric sheets. In this case, the two ends are gripped and the material is pulled to extend, and then chain extension and crystallite rearrangement takes place.

Neck formation (sudden loss of nominal cross-sectional area) after a yield stress maximum is typical for room temperature deformation in several thermoplastic materials. Free radical generation has been detected during tensile deformation long before specimen fracture occurs.[17] This is because of the rupture of highly strained load bearing intercrystalline tie molecules. Lamellar planes do move and rotate toward the direction of tensile draw as well as suffer some shear-induced size modification. This size modification may produce populations of various lamellar aspect ratios that will in turn influence the breadth of the melting endotherm or may altogether form another melting maxima. Lamellar dimensions have a large effect on melting temperature. The Gibbs–Thompson equation[18] can be utilized to calculate this effect. This is shown in Fig. 8.

Another common processing operation that involves uniaxial orientation is fiber spinning. It is customary to extrude using a multiple orifice die (spinneret) and subsequently impose a draw step while quenching in a tank. In general, higher spinning speeds lead to higher levels of induced orientation and better load bearing properties. Because of geometry, important uniaxial texture information in samples made by fiber spinning can be extracted unequivocally with a few measurements. The same average structure should be observed as the sample is rotated to spin around the fiber axis. Wide-angle X-ray patterns can be utilized to determine the degree of orientation of crystallographic planes when the beam is perpendicular to the fiber axis. The span of the diffraction spot in the azimuthal direction provides a measurement of the degree of orientation of a particular diffraction plane with respect to the fiber axis. The breadth of the diffraction spots in the radial direction can be utilized to compute the average dimension of a crystallite in the direction perpendicular to the plane.

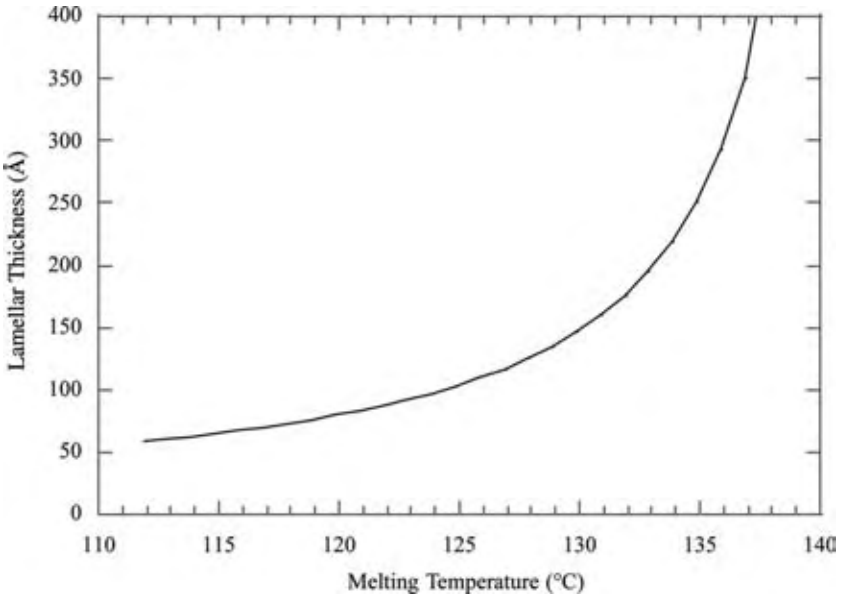


Fig. 8 Theoretical calculation of the effect of lamellar thickness on the melting point of a polymer crystallite using the Thompson–Gibbs equation. (From Ref.[18].) *(View this art in color at www.dekker.com.)*

Less information can be obtained if the beam is parallel to the fiber axis because crystalline plane orientation should have no bias. Fig. 9 shows a wide angle X-ray diffraction pattern of a uniaxially oriented polypropylene when the beam is perpendicular and parallel to the sample's fiber axis. No preferred orientation direction is seen when the beam is parallel to the fiber axis. Besides wide angle X-ray analysis, other popular techniques utilized for the characterization of orientation include small angle X-ray diffraction, optical microscopy, birefringence, infrared and Raman spectroscopy, nuclear magnetic resonance spectroscopy, and neutron scattering. Owing to increased availability of beam time in national laboratories, synchrotron high energy X-ray diffraction is being utilized more and has led to new findings on topics that have generated a great deal of debate in the past. Real time studies of rapidly occurring events and new collimation techniques to resolve submicron length scales have been reported with synchrotron X-ray sources.[19]

Ultraoriented structures

Beyond regular processing operations such as fiber spinning, film stretching, and molding, more exotic methods for obtaining oriented morphologies and ultimate properties have been proposed by both industry and academia. However, most of these methods suffer from line speeds that only amount to a few centimeters per minute. These methods have been reviewed in the literature.[20] A summary of some of these methodologies is provided in Fig. 10. Melt spinning is included for comparison. Most methods devised for obtaining ultraoriented structures are solid-state deformation methods. The only method, as listed in Fig. 10, capable of providing a high orientation directly from the melt is melt transformation extrusion and coextrusion.[21,22] In this process, a flowing melt is forced to flow through a converging cavity at very high pressures, and the orientation attained is frozen at a land section where a crystallization front is formed. Coextrusion is implemented to lubricate the converging and land section walls with a lower viscosity polymer. This allows for higher drawing speeds.

In solid-state extrusion, a solid material is processed without melting. A heated solid plug of a thermoplastic material below its crystallization temperature, or glass transition in amorphous materials, is forced through a conical cavity. There are several versions of this methodology: hydrostatic, where a solid billet forced through the orifice via pressurized fluid; coextrusion,[23] where the solid billet is housed between fluoropolymer sleeves; and ram or plug extrusion, where the billet is pushed by a piston. Successful hydrostatic extrusion trials in uniaxial and biaxial deformation modes have been reported.[24,25] Very high crystallinities and orientation functions can also be obtained by these methods.[26]

Solid-state extrusion has also yielded some of the highest properties for uniaxially oriented morphologies. Tensile moduli (210 GPa) nearing the theoretical value of a polyethylene single crystal have been attained.[27]

A similar technique, die drawing, is a solid-state extrusion method that is assisted by a tensile force to help overcome the high constraints and polymer strain rates at the die exit. Several methods were devised to address the problem of low lineal speeds because of the aforementioned constraints. These methods include solid state rolling,[28] roll drawing,[29,30] and roll trusion.[31]

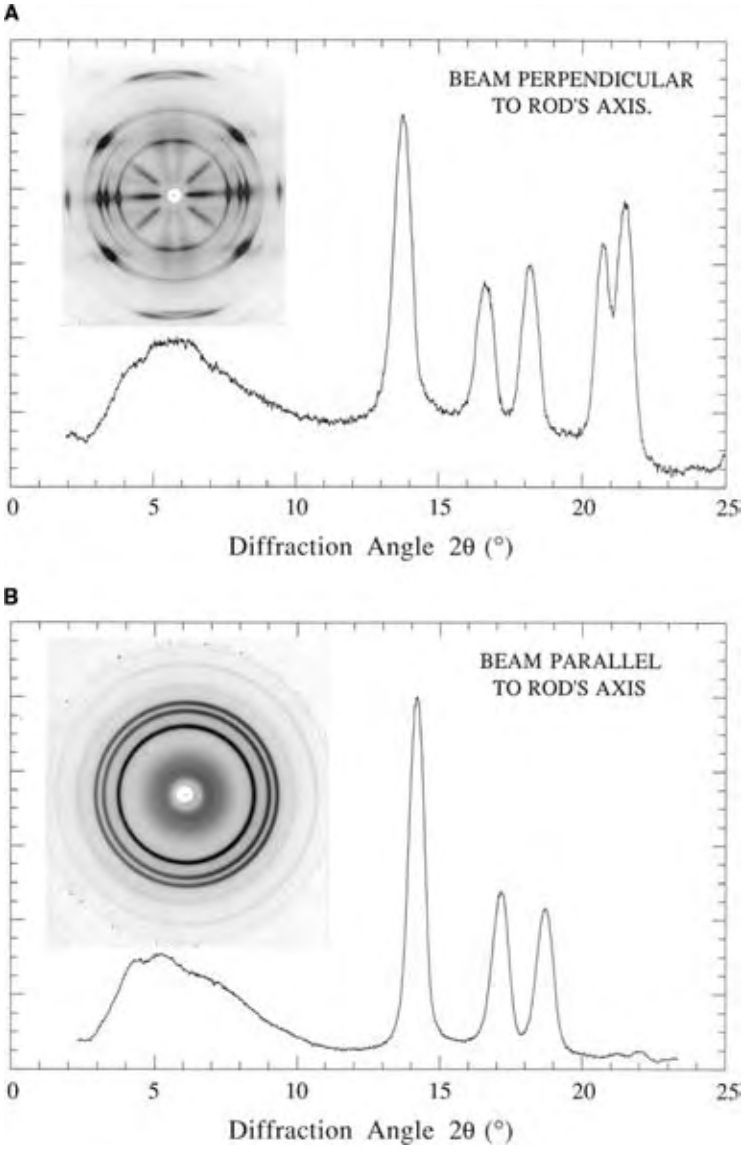


Fig. 9 Wide angle X-ray diffraction of uniaxially oriented isotactic polypropylene rod. (A) Beam perpendicular to the fiber axis. (B) Beam parallel to the fiber axis. Fiber patterns and full scans.

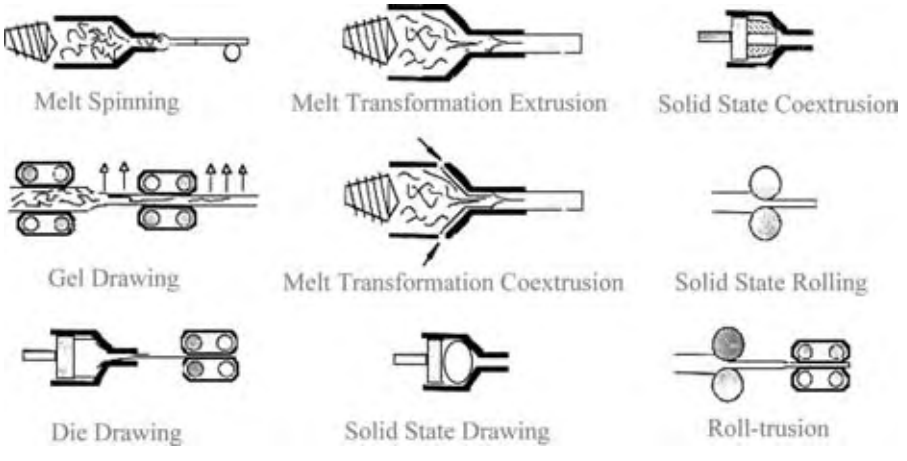


Fig. 10 Deformation methods to produce ultrahigh modulus materials. *(View this art in color at www.dekker.com.)*

Roll trusion suggests the delivery of a system where conditions can be manipulated to provide control of structure formation in two directions.

As done in some of the rolling processes mentioned earlier, it is often desirable to combine stretching with a pressure-induced orientation step during solid-state drawing. Such combinations may involve a hydrostatic extrusion or calendering steps. A higher degree of molecular orientation can be attained with processes that use a compression deformation step than tensile draw only at a given extension ratio.[32] The compressive deformation step induces crystallite cleaving. Crystalline planes shear and slip, thus changing the apparent crystallite size and getting the structure ready for the tensile drawing steps.[33] Unless these compressive steps are done using narrow channels, the nature of the deformation may no longer be uniaxial. Such is the case in calendering where a biaxial character can be obtained.

A tensile modulus comparison for various materials made using some of these ordering processes are shown in Table 1.[34] The reported moduli in this figure were for the direction of maximum level of extension or order. If density is factored into mechanical property comparisons, superior properties are obtained with these ordered polymers when compared to metals and glass.

From the processes shown in Fig. 10, only one is solvent based, i.e., gel spinning.[35] Here ultrahigh molecular weight polyethylene may be diluted to 95% in mixtures of decalin, paraffin oil, or xylene to enable processing and avoid topological restrictions. The solvent is evaporated upon drawing. Surface porosity may arise depending on solvent evaporation conditions. The manufacturing feasibility attained by this process results from a combination of high DR (70 : 1) and line speed (app. 25 m/min). Fibers made by this process are commercially available.

Table 1 Tensile modulus values of highly ordered polymers and reference materials

Material	Tensile modulus (Gpa)
Unoriented thermoplastics	0.1–4
Wood (fiber direction)	15
Glass	70
Highly oriented polyethylene fiber	100–150
Polyparaphenylene-terephthalamide fibers	80–130
Steel	200
Polyphenylenebenzobisoxazole fibers	180–280
Polydiimidazopyridynylene-dihydroxyphenylene fibers	300
Diamond	1200

(From Ref.[34].)

Order in Rodlike Polymers

Most rodlike high polymers are materials with rigid aromatic backbone structures that have very high or intractable melting points. Some of these are liquid crystalline materials. In lyotropic liquid crystalline systems, ordering takes places in solution, while in thermotropic, the oriented domains form in the molten state. Some of these highly ordered systems are made commercially by spinning from solutions. Structure development is dependent on how the solvent is evaporated effectively to yield the best properties. Most liquid crystalline polymers can be made to attain very high properties in tension. However, their compressive properties are very poor because of low shear properties where crystalline planes can be slip sheared. Among these aromatic rod polymers are polyphenyleneterephthalamide, polyphenylenebenzobisoxazole, polydiimidazopyridynylene-dihydroxyphenylene, and a series of thermotropic polyesters. A model of fibrillar hierarchy for highly oriented liquid crystalline polymers has been proposed to describe microfibril size, shape, and organization.[36] This model elucidates morphological causes for poor compression performance and high tensile properties.

Biaxial Orientation

In a biaxial texture, polymer chains are oriented with high bias in the deformation plane. Most biaxial operations with high DRs are done in sequence. After casting, the film is subjected to a machine direction or uniaxial deformation step. This step is followed by a tenter or transverse direction draw. In this latter step, the edges of the film are clamped to effect stretching. The film is still under tension in the forward or machine direction and moves at normal line speeds through the tenter oven. As in fiber spinning, the film may undergo an annealing or a tensilizing step to provide a stable temperature set or a desirable final property. The deformation mechanism during this tranverse orientation step can be dominant. The morphology that is formed during the initial uniaxial or machine direction orientation is intensely modified or destroyed during transverse drawing. Previously obtained crystallite sizes are reduced by this deformation perpendicular to the original draw direction.

Because of recent technological advances, high DR simultaneous biaxial orientation can be achieved continuously at industrial speeds. Ideally for the simultaneous case—at an equal stretch ratio in each direction—in-plane periodicity or other morphological

features should be homogeneous. In isotactic polypropylene, the order attained in the film plane is dependent on the deformation method utilized, sequential or simultaneous biaxial. Molecular orientation developed normal to the plane is independent of either method.[37] In the most industrially common biaxial orientation method, sequential drawing, the sample initial morphology is of extreme importance to the development of morphology during the deformation steps. In this process, a cast sheet is stretched uniaxially, followed by transverse stretching and annealing steps. Crystalline plane orientation is strongly dependent on the degree of crystallinity and crystallite dimensions obtained during previous steps. The temperature at which this transverse or restretching step takes place dictates the type of morphology that will be ultimately attained. For example, the morphology of polypropylene can be substantially changed if the restretching step is done above 130°C. Studies on biaxial orientation of polypropylene conclude that high temperature restretching (around 155°C) yields a smaller size crystallite population through lamellar breakage.[38] Formation of intermolecular crystallites in the direction of stretching was observed upon cooling.

In terms of macrostructure, studies on the deformation of spherulites during biaxial orientation elucidated that spherulitic break-up occurs not only at their boundaries, but also at their centers.[39] In either case and preferentially at higher temperatures, a fibrillar morphology was observed to develop at higher levels of stretching. Less data are available on direct methods to quantify the behavior of amorphous segments as they take part in these deformations.

Besides film tentering, hydrostatic extrusion offers methodologies for biaxially orienting thick sheets.[40] Good mechanical properties have been obtained in the plane of the film with this method, but void formation was reported to be extensive. Biaxial orientation of polyethylenes has also been attained through full compressive deformation.[41] Using this latter method, a modulus of 10 Giga Pascals was obtained in the plane of the sheet with ultrahigh molecular weight polyethylene. Dies with converging-diverging geometries in solid-state extrusion have been shown to enhance mechanical properties along the extrudate's width.[42]

Cyclic Processes

The most common cyclic process for thermoplastics is injection molding. Modern machinery is of the reciprocating screw type. In this process, the screw rotates and loads a metered dosage of polymer toward the tip. A ring valve at the tip of the screw prevents backflow during forward movement of the screw during the injection part of the cycle. The molten polymer flows into mold, which is held at lower temperatures. Depending on the complexity of the mold, different morphologies may be encountered. For example, those originating from the effects of flows near walls and pins, merging flows that form weld lines, and flows at gates and vents. Core or bulk morphologies differ from wall morphologies and depend a lot on the rate of cooling during the injection molding cycle. A strong shearing deformation is imposed on a polymer flowing near walls in a mold cavity. In general, skin layers tend to have a higher level of molecular orientation than the bulk. This is coupled with molecular extension caused by fountain flow or advancing flow fronts. Weld lines yield weak morphological boundaries as melt fronts flow around cold pins or other inserts.

Molecular relaxation in adjacent sections of a part with varying degrees of orientation can induce differential shrinkage that leads to warpage. Rapid cooling of these macrostructures, before and after part ejection, can give rise to a complex state of stress. Residual stress formation was found, at least for amorphous polymers, to be caused by contributions from frozen-in orientation as well as nonequilibrium shrinkage and viscoelastic behavior changes.[43] Residual stress commonly develops as the material undergoes inhomogeneous cooling through its glass transition temperature.

Studies of the orientation phenomena in injection molding have shown that irrespective of the molding conditions and type of thermoplastic, molecular orientation is higher near the walls and decreases toward the interior.[44] Low crystallinity and flow-induced structures have been observed near the walls for injection molded isotactic polypropylene parts.[45] High modulus materials have also been obtained with some injection molding techniques.[46]

CONCLUSIONS

A broad view of principal findings and processes utilized for the development of oriented polymer morphologies has been presented. New trends toward the advancement of this topic are being developed within the realm of multidisciplinary research. Studies of order development in polymers have—for a few years already—transcended beyond traditional disciplines in chemistry and engineering. Genetically engineered polymers, nanoparticles, self-assembled molecules, supercritical fluids, and hybrids are some of the few areas that are now an integral part of macromolecular structural property relationship studies.

ACKNOWLEDGMENT

My sincere appreciation goes to the 3M Company for kindly donating equipment time to develop a deeper

understanding of morphology development in oriented polymer systems.

REFERENCES

1. Keller, A. A note on single crystals in polymers: evidence for a folded chain configuration. Phil. Mag. **1957**, *2*, 1171.
2. Stein, R.S.; Rodhes, M. Photographic light scattering by polyethylene films. J. Appl. Phys. **1960**, *31* (11), 1873.
3. Lovering, E.G. Transcrystallinity and X-ray diffraction in *trans*-1,4-polyisoprene. J. Polym. Sci. Part A-2 **1970**, *8*, 1697.
4. Waddon, A.J.; Hill, M.J.; Keller, A.; Blundell, D.J. On the crystal texture of linear polyaryls (PEEK, PEK, PPS). J. Mater. Sci. **1987**, *22*, 1773.
5. Lotz, B.; Wittman, J.C. Epitaxy of helical polyolefins: polymer blends and polymer-nucleating agent systems. Makromol. Chem. **1984**, *185*, 2043.
6. Peterman, J.; Broza, G.; Riek, U.; Kawaguchi, A. Epitaxial interfaces in semicrystalline polymer and their applications. J. Mater. Sci. **1987**, *22*, 1477.
7. Khoury, F. The spherulitic crystallization of isotactic polypropylene from solution: on the evolution of monoclinic spherulites from dendritic chain-folded crystal precursors. J. Res. Nat. Bur. Stand. **1966**, *A-70*, 29.
8. Norton, D.R.; Keller, A. The spherulitic and lamellar morphology of melt crystallized isotactic polypropylene. Polymer **1985**, *26*, 704.
9. Pennings, A.J.; Kiel, A.M. Fractionation of polymers grown from solution III. On the morphology of fibrillar polyethylene crystals grown in solution. Koll. Zeist. fur Polym. **1965**, Band 205, Heft 2, 160.
10. Keller, A.; Wilmouth, F.M. Some macroscopic properties of stirring induced crystals of polyethylene. J. Macromol. Sci. Phys. **1972**, *B6-3*, 493.
11. Jandt, K.D.; Buhk, M.; Miles, M.J.; Petermann, J. Shish-kebab crystals in polyethylene investigated by scanning force microscopy. Polymer **1994**, *35*, 2458.
12. Grubb, D.T.; Dlugosz, J.; Keller, A. Direct observation of lamellar morphology in polyethylene. J. Mater. Sci. Lett. **1975**, *10*, 1826.
13. Petermann, J.; Gohil, R.M. A new method for preparation of high modulus films. J. Mater. Sci. Lett. **1979**, *14*, 2260.
14. Petermann, J.; Riek, U. Morphologies and mechanical properties of PET films crystallized under high strain rates. J. Polym. Sci. Part B: Polym. Phys. **1987**, *25*, 279.
15. Porter, R.S.; Wang, L.H. Uniaxial extension and order development in flexible chain polymers. J.M.S.—Rev. Macromol. Chem. Phys. **1995**, *C35* (1), 63.
16. Porter, R.S.; Daniels, M.; Watts, M.P.; Pereira, J.R; DeTeresa, S.; Zachariades, A. Shrinkage as a measure of the deformation efficiency of ultra-high density polyethylene. J. Mater. Sci. Lett. **1981**, *16*, 1134.
17. Peterlin, A. Radical formation and fracture of highly drawn crystalline polymers. J. Macromol. Sci. Phys. **1972**, *B6* (4), 583.
18. Bodor, G. *Structural Investigation of Polymers*, 1st Ed.; Ellis Horwood: New York, 1991; 227 pp.
19. Chu, B.; Hsiao, S. Small angle X-ray scattering of polymers. Chem. Rev. **2001**, *101*, 1727.
20. Bigg, D.M. Mechanical property enhancement of semicrystalline polymers—a review. Polym. Eng. Sci. **1988**, *28*, 830.
21. Collier, J.R.; Tam, T.Y.; Newcome, J.; Dinos, N. Extrusion of highly oriented polyolefin fibers. Polym. Eng. Sci. **1976**, *16* (3), 204.
22. Perez, M.A.; Collier, J.R. Melt transformation coextrusion I. Polym. Eng. Sci. **1989**, *29* (15), 1004.
23. Griswold, P.D.; Zachariades, A.E.; Porter, R.S. Solid state coextrusion: a new technique for ultra-drawing thermoplastics illustrated with high density polyethylene. Polym. Eng. Sci. **1978**, *18* (11), 861.
24. Yoon, H.N.; Pae, K.D.; Sauer, J.A. Hydrostatic extrusion of polypropylene and properties of extrudates. Polym. Eng. Sci. **1976**, *16* (8), 567.
25. Pan, S.J.; Tang, H.I.; Hiltner, A.; Baer, E. Biaxial orientation of polypropylene by hydrostatic solid state extrusion. Part II. Morphology Properties **1987**, *27* (12), 869.
26. Chuah, H.H.; Porter, R.S. Solid-state extrusion of chain-extended polyethylene. J. Polym. Sci. Phys. **1984**, *22*, 1353.
27. Porter, R.S.; Kanamoto, T.; Zachariades, A.E. Property opportunities with polyolefins: a review. Preparations and applications of high stiffness and strength by uniaxial draw. Polymer **1994**, *35*, 4979.
28. Bigg, D.M.; Smith, E.G.; Epstein, M.M.; Fiorentino, R.J. High modulus semicrystalline polymers by solid state rolling. Polym. Eng. Sci. **1982**, *22* (1), 27.
29. Burke, P.E.; Weatherly, G.C.; Woodhams, R.T. Uniaxial roll-drawing of isotactic polypropylene sheet. Polym. Eng. Sci. **1987**, *27* (7), 518.
30. Tate, K.R.; Perrin, A.R.; Woodhams, R.T. Molecular orientation of polypropylene by rolling-drawing. Polym. Eng. Sci. **1988**, *28*, 1264.
31. McGill, J.H.; Sun, D.C. *Rolltrusion: Double-Orientation of Polymers using a Single Step Process. Into Integration of Fundamental Polymer Science and Technology—2*; Lemstra, P.J.; Kleintjens, L.A., Eds.; Elsevier: New York, 1988; 490.

32. Ward, I.M.; Coates, P.D.; Dumolin, M.M., Eds.; *Solid Phase Processing of Polymers*; Hanser Gardner Publications, Inc.: Cincinnati, Ohio, 2000; 288 pp.
33. Saraf, R.F.; Porter, R.S. A deformation induced order–disorder transition in isotactic polypropylene. Polym. Eng. Sci. **1988**, *28* (13), 842.
34. Lemstra, P.J.; Bastiaansen, C.W.M.; Peijs, T.; Jacobs, M.J.N. Fibres based on ultra-high molecular weight polyethylene—processing and applications. In *Solid Phase Processing of Polymers*; Ward, I.M., Coates, P.D., Dumolin, M.M., Eds.; Hanser Gardner Publications, Inc.: Cincinnati, Ohio, 2000; Section 5.3, 173 pp.
35. Barham, P.J.; Keller, A. Review: high strength polyethylene fibers from solution and gel spinning. J. Mater. Sci. **1985**, *20*, 2281.
36. Sawyer, L.C.; Chen, R.T.; Jaimeson, M.G.; Musselman, I.H.; Russell, P.E. The Fibrillar hierarchy in liquid crystalline polymers. J. Mater. Sci. **1993**, *28*, 225.
37. Hsu, S.L.; Wang, Y. Personal communication. Polym. Sci. & Eng. Univ. of Mass. At Amherst. 1998.
38. Iwato, N.I.; Tanaka, H.; Okajima, S. Studies on biaxial stretching of polypropylene film. X. High temperature X-ray and optical study of type III orientation. J. Appl. Polym. Sci. **1973**, *17*, 2533.
39. Tanaka, H.; Masuko, T.; Homma, K.; Okajima, S. Studies on the biaxial stretching of polypropylene film. III. Electron microscopy observation of the one-step biaxial stretching of isotactic polypropylene spherulites. J. Polym. Sci. **1969**, *7* (Part A-1), 1997.
40. Tang, H.I.; Hiltner, A.; Baer, E. Biaxial orientation of polypropylene by hydrostatic solid state extrusion. Part III. Mechanical properties and deformation mechanisms. Polym. Eng. Sci. **1987**, *27* (12), 876.
41. Prins, A.J.; Kortschot, M.T.; Woodhams, R.T. Biaxial orientation of linear polyethylenes using the compressive deformation process. Polym. Sci. Eng. **1977**, *37* (2), 261.
42. Mascia, L.; Zhao, J. Transverse orientation dies for solid state extrusion of polymers. Part II. Orientation and properties of extrudates. Polym. Eng. Sci. **1991**, *31* (9), 677.
43. Isayev, A.I. Orientation development in the injection molding of amorphous polymers. Polym. Eng. Sci. **1983**, *23* (5), 271.
44. Fujiyama, M.; Wakino, T. Molecular orientation in injection-molded polypropylene copolymers with ethylene. Inter. Polym. Proc. **1992**, *VII* (2), 159.
45. Saiu, M.; Brucatto, V.; Piccarolo, S.; Titomanlio, G. Injection molding of iPP: an integrated experimental investigation. Inter. Polym. Proc. **1992**, *VII* (3), 267.
46. Kubat, J.; Manson, J.A. High Modulus/high strength polyethylene obtained by high pressure injection molding. Polym. Eng. Sci. **1983**, *23* (16), 869.

Osmotic Distillation

Bob Johnson
School of Physical and Chemical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia

INTRODUCTION

Osmotic distillation (OD) is a membrane separation process that is used primarily for the dewatering of liquid food products at mild (usually ambient) operating temperatures. It is similar to membrane distillation (MD) in its mode of operation. That is, water evaporates from the feed stream (solution being concentrated) and is then transferred in the vapor phase through the pores of a hydrophobic (liquid water repelling) membrane to the strip stream, where it recondenses. The only driving force for mass transfer is a water vapor pressure gradient. However, OD and MD differ in the method by which this gradient is maintained and it is this difference, which determines the applicability of each technique to particular processing applications. In MD, the water vapor pressure of the feed stream is increased, relative to that of the strip stream, by heating the feed stream. In OD, the water vapor pressure of the strip stream is lowered, relative to that of the feed stream, by using a concentrated aqueous solution of an osmotic agent (inorganic salt) as the strip solution.[1,2]

The mild operating temperature of OD is conductive to the concentration of solutions containing thermally labile components and those containing volatile solutes. Thermal degradation or a substantial loss of volatile solutes during concentrate production often detracts from the quality of the concentrate. The most thoroughly investigated applications of OD so far have been the concentration of fruit juices, in particular grape, orange, tomato, and passion fruit juices.[3–10] These solutions have traditionally been concentrated to their required levels (usually greater than 60 Brix) by multistage vacuum evaporation, with obvious quality deterioration. The same levels of concentration have been achieved using OD, with no detectable degradation and with minimal loss of volatile flavor and fragrance components.

This treatment of OD covers the fundamental principles and various practical aspects of the process. Since the first patent on the subject appeared in 1988,[1] OD has been transformed from a laboratory novelty into a commercially viable process in a range of niche applications. However, the wider commercialization of OD has been hampered by the lack of a suitable membrane–membrane module (housing) combination. The interest in OD has been driven by recognition of the numerous potential applications afforded by a process, which facilitates the concentration of aqueous solutions under conditions that maintain a high degree of product integrity.

PRINCIPLES OF OSMOTIC DISTILLATION

OD Membranes

Like MD, OD utilizes a microporous hydrophobic membrane (pore diameter 0.01–1.0 μ) to separate the feed stream from the strip stream. Both streams are pumped across their respective membrane surfaces in cross-flow mode (parallel to membrane surface), to minimize the effects of concentration polarization (see below). Microporous hydrophobic membranes allow the passage of water vapor through the pores, but aqueous solutions are excluded from entering under normal operating pressures. Feed and strip pressures are typically less than 150 kPa, with the feed pressure maintained at 20–40 kPa above that of the strip stream. The use of a higher feed pressure is a precaution against contamination of the feed by the strip solution in the event of membrane leakage. In most cases, this pressure gradient is well below the critical penetration pressure of the membrane, ΔP_c.

ΔP_c depends on the liquid surface tension, γ_l, the liquid–solid (membrane) contact angle, θ (greater than 90° for hydrophobic membranes), the membrane pore radius, r, and the pore geometry co-efficient, B, in accordance with the Laplace equation [Eq. (1)]. B has a value of 1 for cylindrical pores. The higher the liquid surface tension, the larger the contact angle, and the smaller the pore radius, the greater the pressure required for pore intrusion by the liquid.

$$\Delta P_c = \frac{-2B\gamma_l \cos\theta}{r} \quad (1)$$

Young's equation [Eq. (2)] shows that high contact angles are achieved when the solid (membrane) surface tension, γ_s, is low, and the interfacial solid–liquid tension, γ_{sl}, and liquid surface tension are both high. Accordingly, the most suitable membranes for use in

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016181

OD applications are those fabricated from nonpolar polymers with low surface free energies. The most commonly used OD membrane materials are polyolefins, such as polyethylene and polypropylene, and fluoropolymers, such as polytetrafluoroethylene (PTFE) and polyvinylidine difluoride (PVDF).[11]

$$\cos\,\theta\;=\;\frac{\gamma_s\;-\;\gamma_{\mathrm{sl}}}{\gamma_l} \tag{2}$$

The surface tension of most concentrated aqueous solutions of inorganic salts, such as those employed in OD as strip solutions, is considerably greater than that of pure water. Intrusion of these solutions into microporous, hydrophobic membranes of the types used in OD is, therefore, unlikely under moderate operating pressures. However, some aqueous feeds contain amphiphilic components that may depress the liquid surface tension, and thereby reduce the critical penetration pressure. In such cases it may be necessary to use a membrane with a pore diameter of less than 0.1 μ, to prevent liquid intrusion. For most applications however, membranes with a nominal pore diameter of 0.2 μ have been found to be suitable.

Mass Transfer

The water flux, J, which is normally expressed as kg (or L) $m^{-2}\,h^{-1}$, is proportional to the water vapor pressure gradient, Δp_m, between the feed–membrane and strip–membrane interfaces, and the membrane mass transfer co-efficient K_m, [Eq. (3)]. The vapor pressure gradient between the two interfaces depends on the water activity, a_w, in the bulk feed and strip streams, and the extent to which concentration polarization reduces that activity at each interface. Whilst K_m can be estimated using established diffusional transport equations, it is more difficult to estimate values for the water vapor pressure at the membrane wall for use in Eq. (3). However, an overall approach using the vapor pressures of the bulk solutions and semi-empirical correlations that take account of the different conditions near the membrane wall can be used to estimate J.

$$J\;=\;K_m\Delta p_m \tag{3}$$

Maintenance of water vapor pressure gradient

The water vapor pressure of the strip stream is rendered lower than that of the feed stream by the use of a concentrated solution of a solute with high osmotic activity (high water solubility and low equivalent weight) as the strip stream. The strip solution requires periodic reconcentration to maintain the vapor pressure gradient at an effective level. Depending on the size of the plant, an evaporator that utilizes factory waste heat is usually adequate for this operation because of the relatively small fluxes of OD systems (1–3 $kg\,m^{-2}\,h^{-1}$). Typically, the evaporator is placed in a closed loop with the strip tank. Osmotic distillation feed streams contain solutes with low osmotic activity (sugars, polysaccharides, protein), and hence their water vapor pressure is relatively insensitive to concentration changes as OD proceeds.[11] For example, at 25°C a 10 wt% sucrose solution has a vapor pressure of about 25 mm Hg (3.3 kPa), whereas that of a 50 wt% solution is only slightly lower at about 23 mm Hg (3.1 kPa).

In addition to the requirement for high osmotic activity, there are several other factors to be considered when selecting an osmotic agent for use in OD. The agent should be thermally stable and nonvolatile, to withstand reconcentration for reuse. Also, it should preferably be nontoxic, noncorrosive, and of low cost. The agents most frequently used are the alkali and alkaline earth metal halides, particularly NaCl and $CaCl_2$. However, halide salts are quite corrosive to ferrous alloys and are, therefore, unsuitable for prolonged use in most applications. Also, NaCl has relatively low water solubility and $CaCl_2$ is sensitive to precipitation as $CaCO_3$ in the presence of carbon dioxide. Agents found to be most suitable, particularly for the concentration of liquid foods and pharmaceutical products, are the potassium salts of *ortho-* and pyrophosphoric acid. For example, a 50 wt% solution of K_2HPO_4 with a water vapor pressure of about 17 mm Hg (2.2 kPa) at 25°C has been found to be an effective strip solution. Phosphate salts have the added advantage of being present in many biological fluids and are, therefore, considered safe in the event of leakage through a faulty membrane.[11]

Membrane mass transfer

Three mechanisms have been invoked to describe the transfer of water vapor from the feed–membrane interface to the strip–membrane interface, namely Poiseuille flow, Knudsen diffusion, and Fickian (molecular) diffusion.[12] The operative mechanism in a particular system depends on the pore diameter and on whether or not the pores are filled with stationary air. The membrane mass transfer co-efficient (as $kg\,m^{-2}\,h^{-1}\,Pa^{-1}$) applicable to each of these mechanisms can be estimated using Eqs. (4), (5), and (6), respectively. Here, r is the pore radius, ε is the membrane porosity, M_w is the molar mass of water, p_m is the mean water vapor pressure in the pores, δ is the membrane thickness, χ is the pore tortuosity, η is the viscosity of water vapor, R is the gas constant, T is the absolute temperature, D_w is the diffusion co-efficient of water vapor in air, P_a

is the pressure due to stagnant gas (air) in the pores, and P is the total gas pressure in the pores.

Poiseuille flow:

$$K_m = r^2 \varepsilon M_w p_m / 8\delta\chi\eta RT \quad (4)$$

Knudsen diffusion:

$$K_m = (2/3)(r\varepsilon/\delta\chi)(8M_w/\pi RT)^{0.5} \quad (5)$$

Fickian diffusion:

$$K_m = (1/P_a)(\varepsilon/\delta\chi)(D_w P M_w/RT) \quad (6)$$

When air is absent in the pores and the pore diameter is substantially greater than the mean free path, λ, of the diffusing water molecules (distance traveled between collisions with other molecules), water molecules collide more frequently with each other than with the pore walls and the Poiseuille flow relationship applies. The mean free path can be calculated using Eq. (7), where k_b is the Boltzmann constant and σ is the collision diameter of the water molecule. However, for pore diameters in the range of membranes that are suitable for OD applications and for a gas-phase pressure attributable to water vapor alone at ambient temperature (approximately 20 mm Hg, 2.7 kPa), the mean free path is significantly greater than the pore diameter. This results in more frequent collisions of the water molecules with the pore walls than with each other, and Knudsen diffusion predominates.

$$\lambda = k_b T / P(2\pi\sigma^2)^{0.5} \quad (7)$$

In practice, both the feed and strip solutions are saturated with air at ambient temperature, and consequently, the membrane pores contain air at about atmospheric pressure. In some applications, the air may have a higher carbon dioxide or nitrogen content than usual, depending on whether the feed has been sparged with one of these gases to remove oxygen, and thereby prevent oxidative deterioration or microbial attack. Knudsen diffusion may occur in the presence of air when the pore diameter is sufficiently small, and in constricted regions within otherwise larger pores. However, in the presence of air, water vapor–air collisions usually predominate over water vapor–pore wall collisions, and mass transfer through the membrane is by simple Fickian diffusion. In this case, the rate of mass transfer is independent of pore diameter.

The most appropriate model for mass transfer through air-filled pores can be determined by reference to the Knudsen number, K_n, which is the ratio of the mean free path of the water molecule and the pore diameter [Eq. (8)]. A Knudsen number of 10 or greater is indicative of Knudsen diffusion, whereas a value of 0.01 or less is indicative of Fickian diffusion. Intermediate Knudsen numbers signify the coexistence of both mechanisms. Fickian diffusion is significantly slower than Knudsen diffusion at the same vapor pressure gradient. Experimental measurements of water flux in the presence and absence of air in the pores have shown a two to three-fold increase on degassing. Lowering the total gas pressure in the pores decreases the frequency of water vapor–air collisions, and thereby increases the mean free path of the water molecules. However, feed pretreatment by heating or exposure to a vacuum to remove air is often counterproductive to the preservation of product quality.

$$K_n = \frac{\lambda}{2r} \quad (8)$$

Feed and strip mass transfer

Mass transfer in the feed and strip solutions is limited by the extent of concentration polarization. On the feed side of the membrane, concentration polarization refers to an increase in the concentration of solutes at and near the feed–membrane interface because of evaporation of water into the membrane pores (Fig. 1). The resulting solute concentration gradient between the membrane–feed interface, where the concentration is greatest, and the bulk solution induces diffusive transport of rejected solutes back through the concentration polarization boundary layer into the bulk stream. Bulk solution is simultaneously transported to the membrane wall by convection. When equilibrium has been established under a given set of operating conditions (stream flow rate, temperature, fluid dynamics imposed by membrane module design), the rate of back diffusion is equal to the rate at which the solutes are carried to the membrane surface by convective flow.[13]

On the strip side of the membrane, concentration polarization refers to an increase in the water concentration at and near the strip–membrane interface because of condensation of permeate into the strip solution. At equilibrium, the solutes (osmotic agent) diffuse from the bulk stream towards the membrane wall at the same rate as their concentration is reduced by permeate condensation.[12] Water is transported away from the membrane by convection.

Feed-side and strip-side concentration polarization result in a reduction in the driving force for mass transfer. There is a decrease in water activity at the feed–membrane interface and an increase at the strip–membrane interface. This results in a reduction in the water vapor pressure gradient across the membrane. The feed side and strip side mass transfer co-efficients, K_f and K_s, respectively, can be expressed in terms of the solute diffusion co-efficient in the boundary layer, D_s,

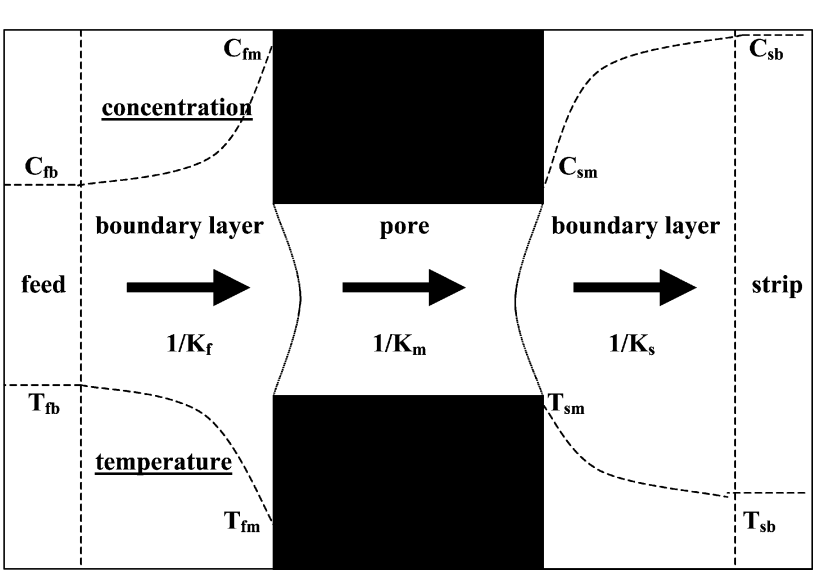


Fig. 1 Solute concentration profile, temperature profile, and mass transfer resistances in OD. The subscripts f, b, m, and s refer to the feed, bulk, membrane, and strip, respectively.

and the boundary layer thickness, z [Eq. (9)]. The magnitude of each mass transfer co-efficient is determined by the viscosity and density of the solution, on which D_s is dependent, and on the hydrodynamic properties of the system, on which z is dependent. However, it is difficult to estimate values for D_s and z and hence an overall approach that takes both the physical properties and hydrodynamics of the solutions into account is used in flux estimations.

$$K_f \text{ or } K_s = \frac{D_s}{z} \tag{9}$$

Overall mass transfer

The water flux achieved in OD can be described in terms of an overall mass transfer co-efficient, K, and the water vapor pressure gradient between the bulk feed and strip streams [Eq. (10)]. The total resistance to mass transfer, given by $1/K$, is the sum of three separate resistances in series [Eq. (11)]. Here, $1/K_f$, $1/K_m$ and $1/K_s$ are the resistances imposed by the feed-side boundary layer, the membrane, and the strip-side boundary layer respectively.

$$J = K\Delta p \tag{10}$$

$$\frac{1}{K} = \frac{1}{K_f} + \frac{1}{K_m} + \frac{1}{K_s} \tag{11}$$

K_m may be estimated using the equation applicable to either Knudsen or Fickian diffusion. K_f and K_s are estimated using empirical correlations of the dimensionless Sherwood, Sh, Reynolds, Re, and Schmidt, Sc, numbers.[11] Several such correlations have been proposed to describe flow through membrane-bound channels of various configurations. These include rectangular channels, such as those present in flat-sheet and spiral-wound modules, and hollow fibers. As an example, Eqs. (12) and (13) show the correlations for the calculation of the Sherwood number applicable to laminar (Re less than 2100) and turbulent flow (Re greater than 4000) through the lumen of a hollow fiber membrane. Having estimated the Sherwood number, the liquid (feed or strip solution) mass transfer co-efficient, K_l, can be calculated using Eq. (14). The Reynolds number and the Schmidt number are calculated using Eqs. (15) and (16), respectively. In these equations, d_h is the hydraulic diameter of the fiber, L is the fiber length, D_s is the diffusion co-efficient of the solute in water, v is the mean solution velocity, ρ is the solution density, and η is the dynamic viscosity of the solution. This approach allows the flux to be estimated by using the vapor pressures of the bulk feed and strip solutions [Eq. (10)].

$$\text{Sh} = 1.86, \text{Re}^{0.33}\ \text{Sc}^{0.33}\ (d_h/L)^{0.33} \tag{12}$$

$$\text{Sh} = 0.023\ \text{Re}^{0.8}\ \text{Sc}^{0.33} \tag{13}$$

$$\text{Sh} = K_l d_h / D_s \tag{14}$$

$$\text{Re} = v d_h \rho / \eta \tag{15}$$

$$\text{Sc} = \eta / \rho D_s \tag{16}$$

Viscous fingering is a complicating factor in the prediction of water flux using the above correlations.[11] This phenomenon, which is rarely encountered in other membrane concentration processes, can be quite

detrimental to process performance. Viscous fingering results from the anomalously high viscosities of feed solutions containing hydrophilic solutes, such as sugars, polysaccharides, and proteins, when concentrated to high levels at ambient temperature. As the feed solution passes through the membrane-bound channel, the solution near the membrane wall becomes increasingly concentrated. Upon reaching a critical concentration, the viscosity rises rapidly with further dehydration. Viscous transport of this layer through the channel becomes progressively slower because it is bounded by flowing liquid of lower viscosity and density. Stagnation of the boundary layer results and the incoming solution "fingers" through the center of the channel with increasing velocity.

The decrease in residence time in the channel, and the blockage of access of the incoming feed solution to the membrane, result in large decreases in the process performance. An example of the rapid increase in viscosity beyond a critical solute concentration is that associated with the concentration of grape juice. The production of grape juice concentrates has been the major commercial application of OD up to the present time.[3] The viscosity of Gordo grape juice increases slowly from about 2×10^{-3} Pa s at a solutes concentration of 18 Brix (single strength) to about 50×10^{-3} Pa s at 60 Brix. On further concentration to the standard concentrate level of 68 Brix, the viscosity increases rapidly to about 220×10^{-3} Pa s, with an accompanying rapid flux decline.[11]

Another complicating factor in the prediction of OD flux using the above correlations is temperature polarization.[13] This is an internally generated phenomenon resulting from mass transfer through the membrane. There is a reduction in temperature at the feed–membrane interface because of the latent heat requirement associated with water evaporation. Condensation of the vapor at the strip–membrane interface releases this latent heat with a consequent increase in temperature. The resulting temperature gradient across the membrane acts in opposition to the vapor pressure gradient generated by the osmotic agent (Fig. 1). The latent heat of evaporation is supplied by convection in the bulk feed stream, conduction across the feed-side boundary layer, and conduction of the heat released on permeate condensation back across the membrane. The net result is a reduction in feed temperature. Heat that is not returned to the feed side from the strip–membrane interface is lost to the strip stream by conduction across the boundary layer and convection in the bulk strip stream. This results in an increase in strip temperature.

To maximize the amount of heat that is returned to the feed side by conduction across the membrane, it is desirable to select a membrane with good thermal conductivity and minimum practical thickness. Membranes used for OD have generally been found to allow recovery of most of this energy. Consequently, the resulting temperature difference between the feed and strip streams seldom exceed 2°C.[9] Frictional heating in the process equipment usually compensates for the slight cooling of the feed stream. The temperature of the strip solution is usually maintained constant by refrigeration. Whilst most authors have considered the effect of temperature polarization on OD flux to be negligible,[14,15] others have suggested that neglecting the driving force decay can lead to over-predictions of the water flux by more than 20%.[13,16] The requirement for a membrane with good thermal conductivity is the opposite of that for MD. The latter process relies on the maintenance of a temperature gradient across the membrane to provide the water vapor pressure driving force, and hence a membrane of low thermal conductivity is preferred.

Effect of temperature on mass transfer

The water vapor flux increases exponentially as the mean temperature of the system increases in accordance with the Antoine equation [Eq. (17)]. Here, T is the absolute temperature and A, B, and C are constants. Temperature also affects water flux through the thermal sensitivity of solution viscosity, solute diffusivity, and the diffusion co-efficient of water vapor in air-filled membrane pores. Elevated temperatures tend to lower feed-side, membrane, and strip-side resistances to mass transfer. However, operation at such temperatures may lead to a loss of product integrity through thermal degradation or volatiles loss.

$$\mathrm{Ln}\, p = \frac{A - B}{T + C} \tag{17}$$

PRACTICAL ASPECTS OF OSMOTIC DISTILLATION

Product Integrity

Osmotic distillation is unique among concentration processes in that feeds containing thermally labile or shear-sensitive components, and those with appreciable volatility, can be concentrated with little or no loss of product integrity. This characteristic is particularly important for the concentration of fruit juices and other liquid foodstuffs, such as vegetable juices, and tea and coffee extracts. These feeds contain a complex mixture of essential volatile hydrophilic flavor, and fragrance components in very low concentrations. Substantial depletion of these components by direct destruction or evaporation may render the product

unacceptable to consumers. The evaporative loss of volatile solutes during multistage vacuum evaporation, which is the conventional method of concentrate production, can be offset to some degree by condensation of the vapor mixture, followed by rectification to recover the volatiles for reblending. However, this process increases the extent of thermal degradation and results in a significant incremental processing cost. Osmotic distillation concentrates that have been reconstituted to single-strength are organoleptically very similar to that of the fresh feeds.

Reverse osmosis (RO) is able to concentrate these feeds at ambient temperature but may damage the product as the result of shear stresses at the high operating pressures required. Also, RO is limited to concentration to 25–30 Brix in most applications because of the high osmotic pressure of the feed. The high feed-side pressure required in overcoming this osmotic pressure to effect further concentration renders RO both uneconomical and impracticable for concentration to high levels. However, RO is among the least costly means for water removal from dilute solutions and may, therefore, be used successfully as a preconcentration step in a hybrid process with OD. Although RO always results in a small leakage of low molar mass solutes through the membrane, concentrates of a quality comparable to that achieved by OD alone have been produced in this way, with a significant reduction in processing costs.[11] For example, grape juice has been concentrated from 18 Brix (single strength) to 30 Brix by RO, followed by concentration to 68 Brix by OD. In this case, RO removed one-half of the water originally present in the juice.

There are several factors that allow OD to effect the concentration of feeds with a minimal reduction of volatile flavor and fragrance components. In general, low temperature operation substantially depresses the vapor pressure of these components relative to that of water, thereby reducing the driving force for transmembrane transport of these solutes. Also, the solubilities of these lipophilic components are substantially lower in the concentrated strip solution than in the feed solution. Consequently, the vapor pressures of these components over the strip solution are higher than over the feed solution at the same concentration. Thus, the vapor pressure driving force for transfer of these solutes across the membrane is considerably lower than that in simple evaporation. Additionally, the molar masses of these components are considerably greater than that of water and consequently their diffusive permeabilities through the membrane are much lower.

Notwithstanding these factors, there is always a small loss of volatile solutes across the OD membrane. However, there is evidence to suggest that this loss can be controlled by the choice of membrane.[8] Membranes having relatively large pore diameters at the surface have been shown to provide better retention of volatile solutes per unit water removal than those with smaller pore openings. It has been suggested that pores with large diameters at the membrane surface allow greater intrusion of the meniscus of both the feed and strip streams into the pore openings. This provides an effective increase in the thickness of the boundary layers at the pore entrances with a resulting increase in the resistance to mass transfer through the aqueous medium. While larger pores are preferred on the basis of reduced volatiles loss, the choice of membrane is dictated by the commercial availability of modules that are most suitable for OD, in particular the four-port hollow fiber modules (see the following sections).

Membrane Modules

As in the case of all membrane separation processes, the choice of an appropriate module type is based on feed type (viscosity, suspended solids content, and particle size), required membrane packing density (based on flux, total throughput, and available floor space), good flow hydrodynamics (for minimization of concentration polarization, effective cleaning and sanitation), and module cost. The various types of membrane modules and their fabrication have been reviewed by Strathmann.[17] Hollow fiber modules, which have the highest membrane packing density of all module types, are the most suitable for use in OD because of the inherently low flux of this process. However, the membranes that have provided the best fluxes and volatiles retention because of their relatively large pore diameters and porosities, that is those fabricated from PTFE, have not yet become available in hollow fibers with an acceptably low wall thickness.

CONCLUSIONS

Osmotic distillation is a relatively new membrane separation process, which is used primarily for the dewatering of liquid food products. The majority of reported applications have involved the production of fruit juice concentrates. The main advantage of OD over vacuum distillation, which is the conventional method for concentrate production, is the preservation of product integrity. Osmotic distillation is operated at mild temperatures (typically ambient), and thereby produces concentrates that are free from the effects of thermal degradation. Also, the delicate volatile flavor and fragrance components that are essential to consumer acceptance of fruit juice concentrates are largely retained.

Osmotic distillation also has advantages over MD with respect to concentrate quality. Osmotic and

membrane distillation are similar in that both the processes utilize a water vapor pressure gradient across a microporous hydrophobic membrane as the driving force for water removal. However, the advantage of OD lies in the use of a concentrated osmotic agent as the strip solution to maintain the vapor pressure gradient, rather than subject the feed stream to heating with a consequent reduction in concentrate quality. Membrane distillation has the advantage of higher fluxes than OD. Osmotic distillation also has a major advantage over RO, a pressure-driven process, despite the high quality of RO concentrates. The high osmotic pressure of the feed at concentrations of 25–30 Brix prevents further concentration from occurring, whereas OD is unlimited by osmotic pressure in the degree of concentration achieved. However, there is potential for the use of the cost effective RO process as a preliminary concentration step in a hybrid process with OD.

While several niche applications for OD have been identified, the commercial acceptance of the technology has been hampered by the nonavailability of a suitable membrane–membrane module combination. Fluoropolymer membranes, such as PTFE and PVDF, have been shown to provide superior flux performance, but are still unavailable in hollow fiber form with a suitable thickness for use in OD applications. The inherently low flux of OD requires that membrane-packing density be maximized for effective operation, and hence the available flat-sheet form of perfluorocarbon membranes is unsuitable for commercial use. Four-port hollow fiber modules that provide excellent fluid dynamics are currently available, but only low-flux polypropylene membranes are utilized.

NOMENCLATURE

a_w	Water activity
B	Pore geometry co-efficient
Brix	Approximate measurement of % by weight of solutes, strictly applying to aqueous sucrose solutions only (% w/w)
C	Concentration ($mol\,m^{-3}$ or $mol\,L^{-1}$, subscripts: f, feed; b, bulk; s, strip; and m, membrane)
D_s	Diffusion co-efficient of solute in water ($m^2\,s^{-1}$)
D_w	Diffusion co-efficient of water vapor in air ($m^2\,s^{-1}$)
d_h	Hydraulic diameter (m)
J	Water flux ($kg\,m^{-2}\,s^{-1}$ or $kg\,m^{-2}\,h^{-1}$)
K	Overall mass transfer co-efficient ($kg\,m^{-2}\,s^{-1}\,Pa^{-1}$)
K_l	Liquid mass transfer co-efficient ($kg\,m^{-2}\,s^{-1}\,Pa^{-1}$)
K_f	Feed side mass transfer co-efficient ($kg\,m^{-2}\,s^{-1}\,Pa^{-1}$)
K_s	Strip side mass transfer co-efficient ($kg\,m^{-2}\,s^{-1}\,Pa^{-1}$)
K_m	Membrane mass transfer co-efficient ($kg\,m^{-2}\,s^{-1}\,Pa^{-1}$)
K_n	Knudsen number
k_b	Boltzmann co-efficient ($1.381 \times 10^{-23}\,J\,K^{-1}$)
M_w	Molar mass of water ($18.0\,g\,mol^{-1}$)
P	Total gas pressure in membrane pores (Pa)
P_a	Pressure because of stagnant gas (air) in membrane pores (Pa)
ΔP_c	Critical penetration pressure (Pa)
Δp_m	Water vapor pressure gradient across membrane (Pa)
p_m	Mean water vapor pressure in membrane pores (Pa)
r	Membrane pore radius (m)
R	Gas constant ($8.314\,J\,mol^{-1}\,K^{-1}$)
Re	Reynolds number
Sc	Schmidt number
Sh	Sherwood number
T	Absolute temperature (K, subscripts: f, feed; b, bulk; s, strip; and m, membrane).
v	Mean solution velocity ($m\,s^{-1}$)
z	Boundary layer thickness (m)
ρ	Density ($kg\,m^{-3}$)
θ	Contact angle (°)
η	Viscosity (Pa s)
χ	Pore tortuosity
ε	Membrane porosity
δ	Membrane thickness (m)
σ	Collision diameter (m)
λ	Mean free path of diffusing species (m)
γ_l	Liquid surface tension ($N\,m^{-1}$)
γ_s	Solid (membrane) surface tension ($N\,m^{-1}$)
γ_{sl}	Interfacial solid–liquid tension ($N\,m^{-1}$)

REFERENCES

1. Lefebvre, M.S.M. Method of Performing Osmotic Distillation. US Patent 4781837, November 1, 1988.
2. Johnson, R.A.; Valks, R.H.; Lefebvre, M.S. Osmotic distillation—a low temperature concentration technique. Austral. J. Biotechnol. **1989**, *3*, 206–207.
3. Thompson, D. The application of osmotic distillation for the wine industry. The Australian Grapegrower & Winemaker **1991**, *328* (April), 12–14.
4. Sheng, J.; Johnson, R.A.; Lefebvre, M.S. Mass and heat transfer mechanisms in the osmotic distillation process. Desalination **1991**, *80*, 113–121.

5. Durham, R.J.; Nguyen, M.H. Hydrophobic membrane evaluation and cleaning for osmotic distillation of tomato puree. J. Membr. Sci. **1994**, *87*, 181–189.
6. Deblay, P. Un nouveau procede de concentration de solutions aqueuses: l'evaporation osmotique. Bios **1994**, *250*, 56–57.
7. Jerial, O.; Reynes, M.; Courel, M.; Durand, N.; Dornier, M.; Deblay, P. Comparison de quelques techniques de concentration des jus de fruits. Fruits **1996**, *51*, 437–450.
8. Barbe, A.M.; Bartley, J.P.; Jacobs, A.L.; Johnson, R.A. Retention of volatile flavor/fragrance components in the concentration of liquid foods by osmotic distillation. J. Membr. Sci. **1998**, *145*, 67–75.
9. Bailey, A.F.G.; Barbe, A.M.; Hogan, P.A.; Johnson, R.A.; Sheng, J. The effect of ultrafiltration on the subsequent concentration of grape juice by osmotic distillation. J. Membr. Sci. **2000**, *164*, 195–204.
10. Vaillant, F.; Jeanton, E.; Dornier, M.; O'Brien, G.M.; Reynes, M.; Decloux, M. Concentration of passion fruit juice on an industrial pilot scale using osmotic evaporation. J. Food Eng. **2001**, *47*, 195–202.
11. Hogan, P.A.; Canning, R.P.; Petersen, P.A.; Johnson, R.A.; Michaels, A.S. A new option: osmotic distillation. Chem. Eng. Prog. **1998**, *94* (7), 49–61.
12. Kunz, W.; Benhabiles, A.; Ben-Aim, R. Osmotic evaporation through macroporous hydrophobic membranes: a survey of current research and applications. J. Membr. Sci. **1996**, *121*, 25–36.
13. Courel, M.; Dornier, M.; Rios, G.M.; Reynes, M. Modeling of water transport in osmotic distillation using asymmetric membrane. J. Membr. Sci. **2000**, *173*, 107–122.
14. Mengual, J.I.; Ortiz De Zarate, J.M.; Pena, L.; Velazquez, A. Osmotic distillation through porous hydrophobic membranes. J. Membr. Sci. **1993**, *82*, 129–140.
15. Bandini, S.; Gostoli, C. Concentrating aqueous solutions by gas membrane extraction. Proceedings of Euromembrane'95 Conference, European Society for Membrane Science and Technology; Bath, UK, 1995; Vol. 2, 19–24.
16. Cervellati, A.; Zardi, G.; Gostoli, C. Osmotic distillation: developments in technology and modeling. Proceedings of Conference on Membrane Distillation, Osmotic Distillation and Membrane Contactors; Cetraro, Italy, 1994; 39–42.
17. Strathmann, H. Synthetic membranes and their preparation. In *Handbook of Industrial Membrane Technology*; Porter, M.C., Ed.; Noyes Publications: Westwood, New Jersey, 1990; 1 pp.

Ozone Treatment

Jiangning Wu
Department of Chemical Engineering, Ryerson University, Toronto, Ontario, Canada

INTRODUCTION

This entry introduces applications of ozone technology in various areas: water and wastewater treatment; control of the microbial safety of food; decontamination of soils; polymer surface modification; and bleaching paper pulps. For water and wastewater treatment, in addition to being used alone, ozone is increasingly used in combination with heterogenous catalysts, UV/H_2O_2 (advanced oxidation process), and biological treatment to enhance ozonation efficiency. The discussion that follows mainly introduces the applications of ozone in water and wastewater treatment because ozone has been both extensively and intensively used in this area; however, it does briefly describe other applications.

PROPERTIES OF OZONE

Ozone (O_3) is a very powerful oxidant (oxidation potential, $E^\circ = 2.07\,V$).[1] It can oxidize a wide spectrum of chemicals and oxidatively inactivate diverse microorganisms. Ozone's self-decomposition to oxygen makes it a residue-free technology. Mainly because of the strong oxidative capability and the unique residue-free characteristic, ozone has been used extensively in water and wastewater treatment.[2] Other applications of ozone include control of the microbial safety of food, decontamination of soils contaminated by polycyclic aromatic hydrocarbons (PAHs), modification of polymer surfaces to increase their biocompatibility, and bleaching paper pulps.[3–8] In addition to being used alone for water and wastewater treatment, ozone is increasingly used in combination with heterogenous catalysts, UV/H_2O_2 (advanced oxidation process), and biological treatment to enhance ozonation efficiency.[5,6,9]

Ozone is a metastable molecule produced from elemental oxygen.[2] It has a pungent, characteristic odor described as similar to "fresh air after a thunderstorm."[10] The solubility of ozone is 13 times that of oxygen at 0–30°C and ozone has a longer half-life in the gaseous state than in aqueous solution.[11] The major physical and chemical properties of ozone are listed in Table 1, and the oxidation potential of ozone together with several commonly used oxidizing agents is shown in Table 2.

GENERATION OF OZONE

Because of ozone's instability, it has to be generated on-site of applications. Ozone can be generated by electrical (corona discharge), photochemical, electrochemical, thermal, and chemonuclear methods.[3,13] Among these methods, corona discharge (CD) is the most widely used to produce large amounts of ozone.[2,3] The first industrial CD ozone generator was developed by von Siemens in 1857, and the CD ozone is generated by passing dried oxygen or air through a high-voltage electrical field.[2] The photochemical ozone generation, on the other hand, involves the formation of ozone from oxygen exposed to UV light below 200 nm. UV-generated ozone is mostly used for residential pool and spa treatment.[5]

OZONE TREATMENT OF WATER AND WASTEWATER

Ozone is residue-free in water due to its self-decomposition into oxygen. As described above, ozone has been extensively used in water and wastewater treatment since its first full-scale application in drinking water treatment at Oudshoorn, Netherlands, in 1893.[2] To date, ozone has been used full scale for disinfection and purification of ground and surface waters, treatment of municipal and industrial wastewater, and treatment of swimming pool and cooling tower waters.[2,4,5,14]

In water and wastewater treatment ozone is used for the following purposes: 1) disinfection; 2) oxidation of organic compounds, including removal of taste, odor, and color; and 3) oxidation of inorganic compounds.[2,13,15] Ozone is usually used as pre-, intermediate-, or posttreatment with other processes such as sedimentation, adsorption, filtration, etc.

Ozone treatment of water is relatively expensive when compared with chlorine. Regardless of the relatively higher cost, however, considering the overall health and environmental problems caused by chlorination, ozone sometimes cannot be replaced by chlorine. Moreover, continuing efforts on lowering the energy consumption of ozone generation and optimizing ozonation processes would result in considerable cost reduction and thus would enable

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007888

Table 1 Physicochemical properties of ozone

Property	Value
Melting point (°C)	−251
Boiling point (°C)	−112
Critical pressure (atm)	54.62
Critical temperature (°C)	−12.1
Specific gravity	1.658 higher than air, 1.71 g/cm^3 at −183°C
Critical density (kg/m^3)	436
Heat of vaporization (cal/mol)[a]	2,980
Heat of formation (cal/mol)[b]	33,880
Free energy of formation (cal/mol)[b]	38,860
Oxidation potential (V)[c]	2.07

[a]At the boiling point temperature.
[b]At 1 atm and 25°C.
[c]At pH 0.
(From Ref.[1].)

ozone treatment to be a more competitively economic technology.

Reaction Kinetics of Ozone with Solutes in Water

When ozone is dissolved in aqueous medium, it reacts with solutes in two mechanisms: the direct reaction and the indirect reaction.[16] The two reactions may occur simultaneously. The direct reaction involves molecular ozone reaction with solutes while the indirect reaction involves the reaction between the solute and the hydroxyl free radical ($OH^{\bullet}$), which is produced from the decomposition of ozone in water. Hydroxyl radical is an even stronger oxidant than molecular ozone (oxidation potential, $E^{\circ} = 3.06$ V). When in reaction with solutes, molecular ozone is selective whereas the hydroxyl free radical is nonselective.[16] The efficacy of the hydroxyl radical for water treatment mainly depends on the quality of the water because of its nonselective nature. It may be less effective for waters containing a high concentration of radical scavengers such as *tert*-butanol, bicarbonate, carbonate, and natural organic matter.[17] Fig. 1 shows the different pathways of ozone reacting with solutes.

Table 2 Oxidizing agents and their oxidation potential

Oxidizing agent	Oxidation potential (mV)
Fluorine	3.06
Ozone	2.07
Permanganate	1.67
Chlorine dioxide	1.50
Hypochlorus acid	1.49
Chlorine gas	1.36

(From Ref.[12].)

For direct reactions between ozone and the solute, the reaction is first order with respect to ozone and the solute, respectively. Overall, the reaction is second order:[2]

$$-\frac{d[O_3]}{dt} = k[O_3][M] \quad (1)$$

where k is the reaction rate constant and $[O_3]$ and [M] are concentrations of ozone and the solute, respectively.

For indirect reactions between the hydroxyl radical and the solute, the rate of reaction mainly depends on the rate of ozone decomposition, which in turn, significantly depends on the nature of solutes in water. Because the solutes vary from water to water, it is difficult to express the indirect reactions in a general rate law. Numerous researchers have conducted investigations on rate equations of ozone decomposition, and have derived the equations with different complexity and different orders (half-, first-, second-, or third-order reactions with respect to the ozone).[15] More details can be found in Ref.[15]. It has also been found that pH is a very important parameter for ozone decomposition. Generally, when pH <7, it has a small effect on the rate of ozone decomposition. However, at higher pH, the rate of ozone decomposition increases significantly with pH.[15,16]

Drinking Water Treatment

The ability of ozone to disinfect polluted water was recognized by de Meritens in 1886 and the number of ozone treatment facilities for drinking water grew rapidly prior to 1914, mainly located in Europe.[2] Later, the use of ozone was significantly slowed down because of the development of an inexpensive chlorine treatment method. However, in the late 1970s, it was discovered that some chlorine disinfection by-products (DBP), trihalomethanes (THMs), were carcinogens. In addition, chlorine was incapable of inactivating some parasitic organisms. Consequently, there was a renewed interest in the applications of ozone in drinking water treatment.[2,13] Till 1997, there were more than 1000 drinking water treatment plants in Europe and over 200 in the United States.[5,18] In the United States, ozone is applied to nearly all bottled waters to provide final disinfection.[5] In general, ozone is used in drinking water treatment for disinfection, removal of taste, odor, and color, and minimization of the formation of chlorination by-products (THMs).

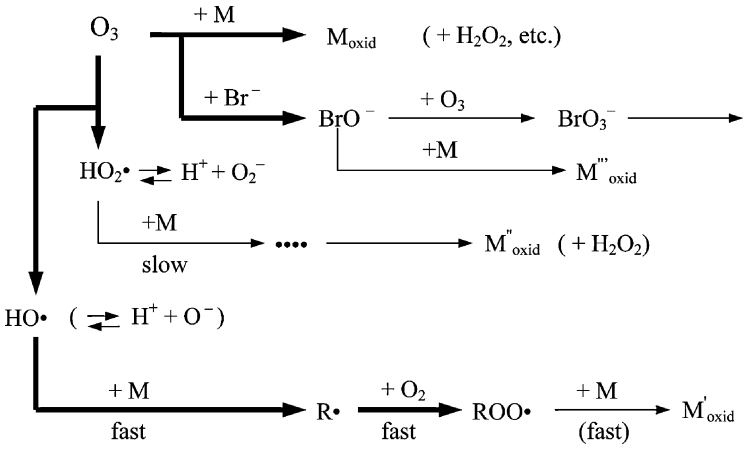


Fig. 1 This scheme shows the different pathways of the reactions of ozone with solutes (M) (including Br^-), and the formation of secondary oxidants, which also react with solutes M, but which produce different products from those formed on direct reaction with ozone. (From Ref.[16].)

Disinfection

Ozone has been found to be very effective on inactivating a wide range of microorganisms. It is generally accepted that molecular ozone is a more effective biocide than hydroxyl radicals, because the latter are very short-lived and nonselective. The resistance of microorganisms follows the increasing order: bacteria, viruses, and parasite cysts.[13,19]

The disinfection kinetics can be expressed by the law of Chick–Watson:[2]

$$\ln\left(\frac{N}{N_0}\right) = -AC^n t \qquad (2)$$

where N is the number of microorganisms surviving at time t; N_0 the initial number of microorganisms; C the concentration of disinfectant; t the contact time; A the specific coefficient of lethality, and n the coefficient of dilution.

Based on the Chick–Watson law, the concept of Ct has been adopted, where C is the dissolved disinfectant concentration (in mg/L) and t is the time (in min) the water is in contact with the disinfectant.[2,5,19] Because of ozone's powerful disinfecting capabilities, it has the minimum value of Ct compared with chlorine and other commonly used disinfectants as shown in Table 3.

Oxidation of organic compounds

Natural Organic Matter. Natural organic matter (NOM), the major component of total organic carbon (TOC) in raw drinking water, has been recognized as the principal precursor of chlorine DBPs.[20] The well-known DBPs are THMs that have already been identified as human carcinogens. The presence of NOM also causes other water quality problems such as color and odor. In addition, a water distribution system containing NOMs favors the bacterial regrowth.[13,19]

The main constituents of NOM in natural water are humic substances consisting of humic and fulvic acids.[2] Humic substances are the precursors of THMs. These must be degraded before chlorine disinfection because once THMs are formed they cannot be removed by chemical oxidation. Ozonation can degrade humic substances into low molecular weight compounds that are less reactive toward chlorine, and hence can minimize the formation of THMs.[19,21]

As mentioned above, humic substances may not be completely mineralized by ozonation alone. Instead,

Table 3 Ct values (mg · min/L) for 99% inactivation of microorganisms with disinfectants at 5°C

	Disinfectant			
Microorganism	**Free chlorine (pH 6–7)**	**Preformed chloramine (pH 8–9)**	**Chloride dioxide (pH 6–7)**	**Ozone (pH 6–7)**
E. *coli*	0.034–0.05	95–180	0.4–0.75	0.02
Polio I	1.1–2.5	770–3740	0.2–6.7	0.1–0.2
Rotavirus	0.01–0.05	3810–6480	0.2–2.1	0.006–0.06
Phage f2	0.08–0.18	—	—	—
G. lamblia cysts	47–>150	—	—	0.5–0.6
G. muris cysts	30–630	1400	7.2–18.5	1.8–2.0

(From Ref.[2].)

ozonation of humic substances usually leads to the formation of small molecules, mainly aldehydes and carboxylic acids. Therefore, the TOC is either reduced or kept unchanged. However, the low molecular weight aldehydes and carboxylic acids could be easily removed using a medium such as granular activated carbon (GAC) that allows the development of biological activity. In fact, combined ozone/GAC filtration proved to be efficient in reducing humic substances and dissolved organic carbon (DOC).[19,21]

Ozone is particularly effective in terms of color reduction. The removal of color and UV-absorbance is usually conducted in preozonation steps. The color removal results from the depolymerization and loss of aromaticity of humic substances.[14,19]

Ozone can sometimes solve taste and odor problems when the taste and odor are caused by unsaturated organic compounds. For example, unsaturated aldehydes, which are important taste and odor by-products of *Synura* species, are readily oxidizable by ozone.[2] Also, ozone is quite effective in removing 2-methylisoborneol (MIB), which causes the musty–earthy taste and odor and is rarely removed by conventional water treatment methods, from the natural water. However, if the taste- and odor-causing compounds are saturated, ozone treatment may have little effect.[2,22]

Synthetic Organic Micropollutants. Micropollutants that frequently presented in water include chlorobenzenes, PAHs, polychlorinated biphenyls (PCBs), and pesticides. Polycyclic aromatic hydrocarbons can be degraded effectively by ozone. But oxidation of chlorobenzenes and PCBs is slow by molecular ozone. However, they are more reactive with hydroxyl radicals. Therefore, it is recommended that either ozonation be conducted at high pH or advanced oxidation processes (AOP) be used. As for the pesticides, their removal by ozonation may be related to their water solubility. However, faster degradations are usually obtained with AOP.[19]

Oxidation of inorganic compounds

Inorganic compounds including manganese and iron usually are oxidized by preozonation to form insoluble species. Preozonation must be followed by a filtration or a coagulation–flocculation–decantation step to finally remove those inorganics.[19]

It should be noted that for water containing bromide, ozonation results in the formation of bromate, a potential carcinogen. Therefore, the presence of bromate has been regulated in water intended for human consumption. In Europe, bromate concentration cannot exceed 25 μg/L from December 25, 2003, and 10 μg/L after December 25, 2008.[18]

Wastewater Treatment

Ozone treatment of wastewater includes municipal and industrial wastewater treatment.

Municipal wastewater treatment

In municipal wastewater treatment, ozone has been used primarily for disinfection following primary or secondary treatment to meet microbiological standards of discharge. Although chlorine or chlorine compounds are most often used for disinfection, because of the potential DBP production as discussed in the drinking water treatment section, ozone is used to avoid those by-products. In the United States, ozone treatment of municipal wastewater for the purpose of disinfection is declining since 1976. The reason is that in 1976 the U.S. Environmental Protection Agency (USEPA) changed its disinfection policy. The USEPA required the sewage treatment plants to disinfect their effluents only when the receiving waters are to be used for 1) potable water intake; 2) agriculture; or 3) human contact with the waters.[5,14] In 1994, there were more than 40 municipal wastewater treatment plants in the United States that had ozonation facilities.[17]

In addition to disinfection, ozone has been used in municipal wastewater treatment to enhance biodegradability of recalcitrant organic chemicals, control odor, improve suspended solids removal, improve the performance of granular activated carbon, and condition sludge.[5,17]

Industrial wastewater treatment

The full-scale applications of ozone have been installed to treat wastewaters from marine aquaria, electronic chip manufacture, textile industry, petroleum refinery, and electroplating. Ozone has also been used to treat landfill leachates and meat processing wastewater. These applications of ozone are described below.

Treatment of Marine Aquaria Wastewaters. Wastewater from marine aquaria contains high levels of BOD_5 (5-day biological oxygen demand), NH_3, and microorganisms. Owing to restrictions against discharge of aquaria wastewater, the wastewater must be recycled for reuse. Ozone has been used to oxidize excess BOD_5, COD (chemical oxygen demand), and to disinfect pathogenic microorganisms. Sea World of Florida (Orlando) and at least 35 U.S. marine aquaria have installed ozonation facilities followed by biofiltration to treat their recirculating artificial brines. During biofiltration, additional BOD_5 is removed along with NH_3. These techniques allow 100% recycling of wastewater and permit the animals to thrive in a chlorine-free environment. Many U.S. zoos also have installed

ozonation facilities in various animal exhibits. In fact, the acceptance of ozone in aquaria and zoos is growing steadily.[4,5]

Treatment of Electronic Chip Manufacture Wastewater. Electronic chip manufactures need a large quantity of very high-purity water per day, because impurities cause lower electrical yields or even failures of the integrated circuits. Ozone is a recognized oxidizing/disinfecting agent for water purification purpose. It has been used for treatment of influent water to the plant and rinse wastewaters for recycling, and for the storage of high-purity process waters.[4,5] Integrated circuit manufacturers that use ozone technique for water treatment include Bell Telephone Laboratories and IBM Corporation. Bell uses ozone mainly for bacteria control. After ozonation 5–20 CFU (colony forming units) per milliliter can be achieved. On the other hand, IBM uses the combined ozone/H_2O_2 technique for the recycling of spent deionized water. This technique can oxidize trace organic materials (isopropyl alcohol, acetone, photographic stabilizers, and chlorinated organic solvents) and ammonia. When ozone treatment is followed by reverse osmosis and ion exchange, the quality of the treated water meets the standards of reuse.

Treatment of Textile Wastewater. Textile wastewater is characterized by heavy color and a large number of diversified auxiliary agents. Conventional treatment processes do not readily degrade dyes in textile wastewater because dyes are stable to light, oxidizing agents, and aerobic digestion.[23] The oxidation potential of ozone is 1.52 times that of chlorine, as shown in Table 2; ozone decolorization of textile effluents is, therefore, quite successful.[4,24,25] Ozone decolorization of the dyes can be explained by the oxidative cleavage of conjugated chains that shifts the absorption of light from the visible to the invisible portion of the spectrum.[26] During the decolorization processes, ozone may also oxidize many organic compounds to low molecular weight substances, resulting in an increase in the ratio of BOD_5 to COD for the wastewater, indicating an increase in biodegradability.[24] Because raw textile wastewater usually contains a large amount of oxidizable substances, which are ozone consuming, it is advisable to use ozone as a final treatment or at least following chemical coagulation.[27]

Full-scale ozone applications of textile wastewater treatment have been installed in Japan, the United Kingdom, Italy, Germany, etc.[4,13]

Treatment of Petroleum Refinery Wastewater. Ozone treatment of petroleum refinery wastewater has been used mainly to remove organic contaminants, especially phenols.[4] The combination of ozonation followed by GAC adsorption could lower the hydrocarbons and other organic contaminant levels to 0.1 mg/L. In addition, lifetime of the GAC adsorbent is extended 10–50%.

Treatment of Electroplating Wastewater. Toxic cyanide ions are the major pollutant in electroplating wastewater. They must be removed before discharge of the wastewater. As shown in Eqs. (3) and (4), ozone can oxidize free cyanide ion rapidly to less toxic cyanate ion, which then slowly hydrolyzes to nitrogen and ammonia. The reaction equations are as follows:[4]

$$CN^- + O_3 \Rightarrow CNO^- + O_2 \quad (3)$$

$$2CNO^- + 2H_2O \Rightarrow 2CO_2 + N_2 + 4H^+ \quad (4)$$

Boeing Aircraft plant in Wichita, KS, installed an ozone treatment system for its electroplating wastewaters in the 1950s. Later in the 1970s, to remove the ozone-stable iron–cyanide complex, full-scale ozone/UV systems were installed at the U.S. Tinker Air Force Base (Oklahoma). The hydroxyl free radicals generated can readily destroy the complexes. The Cadillac Motor Division at Detroit, MI, also installed a full-scale ozone cyanide remediation unit in the 1980s. So far, ozone application in the treatment of electroplating wastewater is not extensive. The main reason is that the less expensive alkaline chlorination can also oxidize cyanide ions to satisfy the wastewater discharge standards. However, alkaline chlorination cannot convert all of the toxic cyanide ions to cyanate ions. Therefore, the advantage of ozone application is to completely eliminate the cyanide ions so that the discharge of the wastewater and the subsequent sludge are cyanide-free.[4,5]

Treatment of Landfill Leachates. Landfill leachate contains large quantities of nonbiodegradable and toxic constituents.[28] Some heterogenous catalytic ozonation processes (Ecoclear or Catazone systems) can successfully treat landfill leachates.[4,13,28] In addition, ozone-based AOPs have also been successfully used to remove COD or TOC either as pretreatment or at the final stage of the treatment. Both heterogenous catalytic ozonation and the AOP processes are introduced below.

Treatment of Meat Processing Wastewater. Meat processors consume a large quantity of water and produce a significant amount of wastewater. Because of increasing water costs and environmental constraints, reuse of the processing water has become an attractive option for meat processors. Recycling of poultry chill water is permitted by the U.S. Department of Agriculture (USDA) regulations, provided that certain reconditioning criteria (microbial load, COD, light transmission, etc.)

can be satisfied.[29] Ozonation of poultry chiller water could decrease microbial load >2 logs and COD by ca. 33%. Ozonation could also increase light transmission (at 500 nm) of the poultry chiller water.[3,30] Also, ozone is effective on disinfection of red meat processing water.[31]

Treatment of Other Waters

Treatment of cooling tower water

Ozone has been used to replace chromium-containing chemicals to control biofouling for the cooling tower water. The U.S. cooling water market for ozone is in steady growth. Approximately 500 cooling towers have installed ozone.[5]

Treatment of swimming pool water

The major concern for the swimming pool water is the control of microorganisms. In the United States, more than 500,000 pools and spas have installed UV-generated ozone units. Only a few dozen use CD-generated ozone. Consequently, the objectionable odors of chlorine can be avoided.[5]

Some Typical Combined Use of Ozone with Catalysts or with Other Processes in Water Treatment

Heterogenous catalytic ozonation

Heterogenous catalytic ozonation used in water treatment usually implies the use of insoluble solid catalysts to enhance ozonation efficiency. The advantages of catalytic ozonation over regular ozonation are as follows: 1) accelerated degradation of pollutants; 2) considerably reduced ozone consumption and reaction time; and 3) higher degree of mineralization of the pollutants. The use of heterogenous catalysts also allows easy separation of the catalysts from the water for reuse or regeneration.[9,32]

Commonly used solid catalysts are metal oxides such as TiO_2, MnO_2, etc. Activated carbon is also a member of the heterogenous catalyst family.[17,32] Examples of industrial applications of heterogenous catalytic ozonation include CATAZONE and ECOCLEAR processes. CATAZONE process uses TiO_2 as catalyst and this process could almost mineralize oxalic acid completely. CATAZONE can also effectively remove halogenated pesticides such as aldrin, dieldrin, and hexachlorobenzene.[17] Unlike CATAZONE process, in which TiO_2 is used as catalyst, in ECOCLEAR process, ozone gas and wastewater flow cocurrently in a fixed-bed reactor in the presence of the activated carbons, which is used as the catalyst. The ECOCLEAR process has been successfully applied in full scale for the treatment of biologically pretreated landfill leachates since 1992.[13]

Ozone/biological treatment

Biological treatment of wastewater is usually the most cost-effective method. However, recalcitrant contaminants in wastewaters are the challenge for biological treatment because they are resistant to biodegradation. Because many recalcitrant contaminants can be oxidized by ozone to produce biodegradable derivatives, the combination of ozone and biological treatment (usually activated sludge treatment) is very useful for wastewaters containing such contaminants.[13,17]

Advanced oxidation processes

Advanced oxidation processes (AOP) have been defined as processes that involve the generation of hydroxyl radicals in sufficient quantity for effective water purification.[13] Ozone-based AOPs generally include the activation of free radicals by 1) hydrogen peroxide (O_3/H_2O_2); 2) UV (O_3/UV); and 3) hydrogen peroxide and UV ($O_3/H_2O_2/UV$).[17] Advanced oxidation processes improve ozonation efficacy, hence can achieve higher-degree mineralization of organic contaminants.[32]

Ozone-based AOPs are being used increasingly to treat landfill leachates.[4] They are also used for groundwater treatment to destroy trichloroethylene (TCE), tetrachloroethylene, and pentachlorophenol. In addition, they are used for groundwater remediation at Superfund sites in the United States to destroy volatile organic compounds and benzidines. Another application of ozone-based AOPs involves their use at U.S. ammunition plants to destroy explosives.[5]

In fact, heterogenous catalytic ozonation introduced above is a novel type of AOP, because a large amount of radicals are produced during the processes.[32]

OZONE APPLICATIONS IN THE FOOD INDUSTRY

Ozone's strong antimicrobial capability and its self-decomposition into nontoxic oxygen make it very useful in enhancing the microbiological safety and quality of food. However, in contrast to the fact that ozone has been used extensively for water treatment since the beginning of the 20th century, its applications in food industries have been limited in the United States. Only in 1982, the United States Food and Drug Administration (USFDA) granted generally recognized as safe (GRAS) status for use of ozone as a disinfectant

in bottled water.[3,10] In 1997, ozone was granted GRAS status by FDA for use as a disinfectant in foods and food processing. Finally, in June 2001, the FDA formally approved the use of ozone as an antimicrobial agent in the gas or liquid phases on food.[33] This approval enhanced ozone's applications in the food industry as an alternative sanitizer/disinfectant for chlorine in North America.

Ozone inactivates numerous bacteria as vegetative cells and as spores. In general, inactivation of spores is more difficult than that of vegetative cells. Ozone inactivates such gram-positive bacteria as *Listeria monocytogenes*, *Staphylococcus aureus*, *Bacillus cereus, Enterococcus faecalis*, and such gram-negative bacteria as *Pseudomonas aeruginosa, Salmonella typhimurium*, and *Yersinia enterocolitica*. Ozone is also an effective fungicidal agent. It inactivates such fungi as *Aspergillus* and *Penicillium*. As for virus inactivation, ozone is effective for a wide range of viruses including Venezuelan equine encephalomyelitis virus, hepatitis A, influenza A, vesicular stomatitis virus, and infectious bovine rhinotracheitis virus as well as several bacteriophage strains. Ozone can also inactivate such protozoa as *Cryptosporidium parvum, Naegleria gruberi, and Giardia lamblia*. Efficacy of ozone inactivation of microflora on food depends on the nature and composition of the food surface, the type of microbial contaminants, and the degree of attachment or association of microorganisms with food.[3,10]

When ozone is used for food protection, it can be used either in gas phase or after dissolving in water. Gaseous ozone is usually used for surface microflora control of beef; disinfection of hatchery, hatching eggs, recycled poultry chiller water, poultry carcass, and contaminated eggs; inactivation of bacteria and fungi associated with cereal grains, peas, beans, spices; mold control of cheese; removal of odor and color from fish flesh; increase of shelf life of fruits and vegetables; disinfection of slaughterhouse effluent.[3,31,34] On the other hand, water containing dissolved ozone is used for reduction of bacterial contamination of beef carcasses and fish, and for sanitation of meat transport vehicles and food plant equipment.

When ozone is used for the treatment of food, the applied dosage should be carefully optimized. Otherwise, at the same time pathogenic microorganisms are inactivated, the quality of food such as color and taste could deteriorate when the applied ozone dosage is over certain thresholds.

OZONE APPLICATIONS IN SOIL DECONTAMINATION

Soils contaminated by PAHs are widespread, mainly due to fuel oil spills. Biodegradation has been successfully used for soils contaminated with low molecular weight PAHs. However, the higher molecular weight PAHs are resistant to bioremediation.[6,35,36] Similarly, ozone alone has been found to be very efficient for the destruction of PAHs containing two and threebenzene rings in soils, but was not effective for those PAHs containing four and five rings.[35] However, the integrated ozonation and biodegradation processes are found to be more effective than either of them alone. The preozonation generates products that have a better solubility in water and thus a better bioavailability, which is of great importance for a successful bioremediation of the PAHs.[36] Therefore, the combination of ozonation and subsequent biological degradation may be a successful technology in the remediation of soils contaminated with PAHs.

The use of gaseous ozone for in situ chemical oxidation is more effective than aqueous-based systems for the treatment of contaminated soils, because the diffusivity of ozone gas is much greater than that for aqueous species. In addition, the concentration of ozone in the gas phase can be orders of magnitude higher than that obtainable in aqueous solutions, and ozone is more stable in the gas phase than in water.[6]

OZONE APPLICATIONS IN POLYMER SURFACE MODIFICATION

Polymers have been extensively used in the medical field for a large number of important implants and devices since the early 1950s. In spite of their extensive applications, to date the general biocompatibility problem of the polymer materials has not been completely solved. The natural inert surfaces of polymers hinder their applications as biomaterials, which require excellent wettablity. One approach to improve the biocompatibility of some polymers is the ozone-induced graft polymerization. In this approach, the polymer is oxidized by gaseous ozone to generate peroxide groups on the surface, and the peroxide groups are capable of initiating free radical graft copolymerization of various monomers with hydrophilic groups, resulting in modified surfaces with increased hydrophilicity and reduced protein adsorption.[7]

Surface ozone oxidation is applied in polymers because it has the following advantages: it introduces peroxides uniformly on the polymer surfaces; it is applicable to complex geometries; it can be handled easily; and it is relatively economical as compared with other techniques such as corona discharge, plasma treatment, flame treatment, irradiation with gamma rays.[7,8,37]

The common base polymers whose surfaces would be oxidatively modified by ozone include polyurethanes (PU), segmented polyether urethane (SPEU),

polypropylene (PP), polyethylene (PE), polymethylmethacrylate (PMMA). The common monomers used for graft include acrylamide (AAm), 2-hydroxyethyl methacrylate (HEMA), polyethylene oxide (PEO), *N, N′*-dimethyl-*N*-methacryloyloxyethyl-*N*-(3-sulfopropyl ammonium (DMMSA). The ozone-induced graft polymerization has been successfully applied for grafting HEMA monomer onto PP, grafting AAm onto PU film, grafting PEG onto PU, PE, and PMMA, and grafting DMMSA on to SPEU.[7,8,37,38]

OZONE APPLICATIONS IN BLEACHING PAPER PULPS

Ozone has been increasingly used to replace chlorine for bleaching paper pulps. Changing the bleaching agent from chlorine to ozone eliminates the production of AOX (adsorbable organic halides) compounds, which are very difficult to destroy by chemical oxidation once formed, hence they will remain in the paper mill effluents. Moreover, the chlorine-free bleaching could result in the recycle/reuse of a significant amount of process waters, or could even allow the paper mills to be effluent-free.[4,5]

The world's largest industrial ozonation system of 420 kg O_3/hr production capacity is in operation in a pulp bleaching process in Finland, and at least 10 commercial ozone pulp bleaching mills are in operation in various countries throughout the world.[5,13]

CONCLUSIONS

Ozone has been applied successfully and extensively for water and wastewater treatment. Ozone also has been used as a safe and effective antimicrobial agent in many food applications. Other applications of ozone include soil decontamination, polymer surface modification, and bleaching paper pulps. It is recognized that for water treatment, the combined use of ozone with either biological treatment, or heterogenous catalysts, or UV and/or H_2O_2 makes the whole process more efficient.

REFERENCES

1. Perry, R.H.; Green, D.W. *Perry's Chemical Engineers' Handbook*, 7th Ed.; McGraw-Hill: New York, 1997.
2. Langlais, B.; Reckhow, D.A.; Brink, D.R. *Ozonation in Water Treatment: Application and Engineering*; Lewis Publishers: New York, 1991.
3. Kim, J.; Yousef, A.E.; Dave, S. Application of ozone for enhancing the microbiological safety and quality of foods: a review. J. Food Prot. **1999**, *62* (9), 1071–1087.
4. Rice, R.G. Applications of ozone for industrial wastewater treatment—a review. Ozone Sci. Eng. **1997**, *18*, 477–515.
5. Rice, R.G. Ozone in the United States of America—state-of-the-art. Ozone Sci. Eng. **1999**, *21*, 99–118.
6. Masten, S.J.; Davies, S.H.R. Efficacy of in situ ozonation for the remediation of PAH contaminated soils. J. Contam. Hydrol. **1997**, *28*, 327–335.
7. Fujimoto, K.; Takebayashi, Y.; Inoue, H.; Ikada, Y. Ozone-induced graft polymerization onto polymer surface. J. Polym. Sci. A. Polym. Chem. **1993**, *31*, 1035–1043.
8. Yuan, Y.; Zhang, J.; Ai, F.; Yuan, J.; Zhou, J.; Shen, J. Surface modification of SPEU films by ozone induced graft copolymerization to improve hemocompatibility. Colloid. Surf. B. Biointerfaces **2003**, *29*, 247–256.
9. Yong, K.; Wu, J.; Andrews, S. Heterogeneous catalytic ozonation of aqueous reactive dye. Ozone Sci. Eng. *27* (4) *in press*.
10. Guzel-Seydim, Z.B.; Greene, A.K.; Seydim, A.C. Use of ozone in the food industry. LWT—Food Sci.Technol. **2004**, *37*, 453–460.
11. Rice, R.G. Application of ozone in water and waste water treatment. In *Analytical Aspects of Ozone Treatment of Water and Waste Water*; Rice, R.G., Browning, M.J., Eds.; The Institute: New York, 1986; 7–26.
12. Manley, T.C.; Niegowski, S.J. Ozone. In *Encyclopedia of Chemical Technology*, 2nd Ed.; Wiley: New York, 1967; 410–432.
13. Gottschalk, C.; Libra, J.A.; Saupe, A. *Ozonation of Water and Waste Water: A Practical Guide to Understanding Ozone and Its Application*; Willey–VCH: New York, 2000.
14. Paraskeva, P.; Graham, J.D. Ozonation of municipal wastewater effluents. Water Environ. Res. **2002**, *74* (6), 569–580.
15. Beltran, F.J. *Ozone Reaction Kinetics for Water and Wastewater Systems*; Lewis Publishers: New York, 2004.
16. Hoigne, J. The chemistry of ozone in water. In *Process Technologies for Water Treatment*; Stucki, S., Ed.; Plenum Press: New York, 1988; 121–143.
17. Masten, S.J.; Davies, S.H.R. The use of ozoneation to degrade organic contaminants in wastewaters. Environ. Sci. Technol. **1994**, *28* (4), 181A–185A.
18. Bouland, S.; Duguet, J.; Monties, A. Minimizing bromate concentration by controlling the ozone reaction time in a full-scale plant. Ozone Sci. Eng. **2004**, *26*, 381–388.

19. Camel, V.; Bermond, A. The use of ozone and associated oxidation processes in drinking water treatment. Water Res. **1998**, *32* (11), 3208–3222.
20. Chaiket, T.; Singer, P.C.; Miles, A.; Moran, M.; Pallotta, C. Effectiveness of coagulation, ozonation, and biofiltration in controlling DBPs. J. AWWA **2002**, *94* (12), 81–95.
21. Yu, M.J.; Kim, Y.H.; Han, I.; Kim, H.C. Ozonation of Han River humic substances. Water Sci. Technol. **2002**, *46*, 21–26.
22. Jung, S.W.; Baek, K.H.; Yu, M.J. Treatment of taste and odor materials by oxidation and adsorption. Water Sci. Technol. **2004**, *49*, 289–295.
23. Mishra, G.; Tripathy, M. A critical review of the treatments for decolorization of textile effluent. Colourage **1993**, *10*, 35–38.
24. Wu, J.; Wang, T. Effects of some water-quality and operating parameters on the decolorization of reactive dye solutions by ozone. J. Environ. Sci. Health **2001**, *A36* (7), 1335–1347.
25. Wu, J.; Wang, T. Ozonation of aqueous azo dye in a semi-batch reactor. Water Res. **2001**, *35*, 1093–1099.
26. Carriere, J.; Jones, J.P.; Broadbent, A.D. Decolorization of textile dye solutions. Ozone Sci. Eng. **1993**, *15*, 189–200.
27. Vandevivere, P.C.; Bianchi, R.; Verstraete, W. Treatment and reuse of wastewater from the textile wet-processing industry: review of emerging technologies. J. Chem. Technol. Biotechnol. **1998**, *72*, 289–302.
28. Wu, J.J.; Wu, C.; Ma, H.; Chang, C. Treatment of landfill leachate by ozone-based advanced oxidation processes. Chemosphere **2004**, *54*, 997–1003.
29. USDA. *Code of Federal Regulations—Poultry Products: Temperatures and Chilling and Freezing Procedures—Title 9, Part 381.66*; Office of the Federal Register, National Archives and Records Administration: Washington, DC, 1997.
30. Diaz, M.E.; Law, S.E.; Frank, J. Control of pathogenic microorganisms and turbidity in poultry-processing chiller water using UV-enhanced ozonation. Ozone Sci. Eng. **2001**, *23*, 53–64.
31. Wu, J.; Doan, H. Disinfection of recycled red-meat-processing wastewater by ozone. J. Chem. Technol. Biotechnol. **2005**, *80*, 828–833.
32. Legube, B.; Leitner, N.K.V. Catalytic ozonation: a promising advanced oxidation technology for water treatment. Catal. Today **1999**, *53*, 61–72.
33. Rice, R.G.; Graham, D.M. U.S. FDA Regulatory approval of ozone as an antimicrobial agent—what is allowed and what needs to be understood. Ozone News **2001**, *29*, 22–31.
34. Allen, B.; Wu, J.; Doan, H. Inactivation of fungi associated with barley grain by gaseous ozone. J. Environ. Sci. Health **2003**, B*38* (5), 617–630.
35. Nam, K.; Kukor, J.J. Combined ozonation and biodegradation for remediation of mixtures of polycyclic aromatic hydrocarbons in soil. Biodegradation **2000**, *11*, 1–9.
36. Stehr, J.; Muller, T.; Svensson, K.; Kamnerdpetch, C.; Scheper, T. Basic examinations on chemical pre-oxidation by ozone for enhancing bioremediation of phenanthrene contaminated soils. Appl. Microbiol. Biotechnol. **2001**, *57*, 803–809.
37. Ko, Y.G.; Kim, Y.H.; Park, K.D.; Lee, H.J.; Lee, W.K.; Park, H.D.; Kim, S.H.; Lee, G.S.; Ahn, D.J. Immobilization of poly(ethylene glycol) or its sulfonate onto polymer surfaces by ozone oxidation. Biomaterials **2001**, *22*, 2115–2123.
38. Wang, Y.; Kim, J.; Choo, K.; Lee, Y.; Lee, C. Hydrophilic modification of polypropylene microfiltration membranes by ozone-induced graft polymerization. J. Membr. Sci. **2000**, *169*, 269–276.

Packed Absorption Column Design

Karl B. Schnelle, Jr.
Department of Chemical Engineering, Vanderbilt University, Nashville, Tennessee, U.S.A.

Partha Dey
P.A. Consulting, Nashville, Tennessee, U.S.A.

INTRODUCTION

Absorption columns may be packed towers or plate towers. As the current process usage is leaning more toward packed towers and especially structured packing, this entry will concentrate on vertical packed tower design. There are two directions of flow in a vertical tower, countercurrent and cocurrent. This entry presents the methodology for vertical countercurrent and cocurrent packed absorption column design. In the basic sense, the design consists of determining the tower diameter and packing height. The tower diameter depends on the fluid dynamics of the flow. The tower height depends on the vapor–liquid equilibrium and mass transfer characteristics, which are also functions of the fluid dynamics. Cross-flow towers are also used in absorption. Here, the gas flows horizontally and the liquid is sprayed down on the gas flow. Design is more complicated for the cross-flow tower and will not be considered in this entry.

In countercurrent flow, the tower must be designed to prevent flooding, which occurs when the upward flow of the gas is of sufficient velocity, to prevent the liquid flow from coming down the column. The advantage of a cocurrent tower is that the fluids flow both in the same direction, usually downward, and flooding will not occur. High flow rates of both the liquid and gas may be used, which results in smaller column diameters. The disadvantage of cocurrent flow absorbers is that only one equilibrium stage is available. Therefore, the mass transfer may not be sufficient.

Tower height depends on vapor–liquid equilibrium and mass transfer. The driving force for mass transfer in absorption is the concentration difference of the solute between the gaseous and liquid phases. The driving force can be interpreted as the difference between the partial pressure of the soluble gas in the gas mixture and the vapor pressure of the solute gas in the liquid film in contact with the gas. If the driving is negative, desorption or stripping will occur, and the concentration of the gas being treated will increase.

THERMODYNAMIC EQUILIBRIUM RELATIONSHIPS FOR ABSORPTION COLUMNS

For an extensive discussion of phase equilibria, consult Smith, van Ness, and Abbott[1] for an introductory treatise, and Prausnitz, Lichtenthaler, and Gomes de Azevado[2] for a more advanced text.

Conditions for Equilibrium

The concept of fugacity was developed by Lewis.[3] Fugacity for a vapor can be defined by Eq. (1).

$$\frac{f_i^v}{y_i P} \rightarrow 1 \quad \text{as } P \rightarrow 0 \tag{1}$$

Here, y_i is the mole fraction of component i in the vapor. As $P \rightarrow 0$, the vapor approaches an ideal gas. Therefore, for an ideal gas, the fugacity is the partial pressure $y_i P$ as seen in Eq. (2).

$$f_i^v = y_i P \tag{2}$$

Equilibrium for a vapor and a liquid (VLE) is defined in Eq. (3).

$$f_i^v = f_i^l \tag{3}$$

Raoult's Law

Raoult's law is the simplest quantitative expression for vapor–liquid equilibrium. This law is based on the vapor phase being an ideal gas and the liquid phase being an ideal solution. Therefore, the vapor-phase fugacity is given by Eq. (2). The effect of pressure on the liquid phase is very small, and since the vapor is ideal, the liquid fugacity may be written as in Eq. (4). Here, P_i^{sat} is the vapor pressure of component i at the temperature of the solution.

$$f_i^l = x_i P^{sat} \tag{4}$$

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120041351

Substituting Eqs. (2) and (4) into Eq. (3) results in Eq. (5) known as Raoult's law.

$$y_i P = x_i P_i^{sat} \quad (5)$$

Note that this law was developed essentially for an ideal solution where the vapor is an ideal gas. Therefore, Raoult's law is very limited in its usefulness. Essentially, it is valid for solutions of chemically similar compounds at a pressure of up to 2 or 3 bars. Raoult's law can be applied only to systems where the component's vapor pressure is known. If the temperature of the system is above the critical temperature for any component, there is no vapor pressure for that component, and other means must be employed to determine compositions. This situation is described by Henry's law in the following section.

Henry's Law

In the case where the solute is at a temperature above its critical temperature, the liquid mixture cannot exist over the entire composition range. Assume that we are dealing with an ideal dilute solution where the solvent does not dissolve in the solute. We can write the fugacity for an ideal solution as in Eq. (6) where f_i is the fugacity of pure component i.

$$f_i^{id} = x_i f_i \quad (6)$$

As we have a dilute solution, we can assume that no solvent dissolves in the solute. Let subscript 1 refer to the solvent and subscript 2 to the solute. Then, as $x_2 \rightarrow 0$, Eq. (7) results in

$$\lim_{x_2 \rightarrow 0} \frac{f_1^{id}}{x_1} = f_1 \quad (7)$$

For component 2, the solute, we can define Henry's law constant by Eq. (8).

$$\lim_{x_2 \rightarrow 0} \frac{f_2^{id}}{x_2} = H_{2,1} \quad (8)$$

Henry's law for a dilute solution and for the situation where the pressure is low enough that the ideal gas can apply is given by Eq. (9) and this is very similar to Eq. (5),

$$y_2 P = x_2 H_{2,1} \quad (9)$$

Raoult's law, but now it is applied to dilute solutions. It is also apparent that Eq. (9) can be applied to the case where gases are only slightly soluble in a solvent. A classic case is to determine the saturation concentration of oxygen in natural waters and in waste treatment plant effluents.

The Case for Nonideal Liquid Solutions and Ideal Vapor Solutions

In the case of the ideal gas solution at a moderate pressure, Eq. (2) defines the fugacity for the vapor. The activity coefficient can be defined by Eq. (10), where f_i^* is the fugacity of

$$\gamma_i^l = \frac{f_i^l}{x_i f_i^*} \quad (10)$$

component i in the standard state. If we now assume that the pure component at the pressure and temperature of the solution becomes the standard state of the liquid, f_i^* becomes the vapor pressure of component i. The vapor–liquid equilibrium can now be written as Eq. (11).

$$y_i P = x_i \gamma_i^l P_i^{sat} \quad (11)$$

This formulation is Raoult's law extended by the inclusion of the activity coefficient and is not much more difficult to use and it has much greater utility and accuracy at moderate pressures than Raoult's law. It is especially useful when correlations are available for liquid-phase activity coefficients γ_i^l.

Correlations Representing VLE Data—Margules, van Laar, Wilson, and UNIQUAC

Wohl[4] proposed a general method for expressing excess Gibbs energies as a function of effective volume fractions. Liquid mole fractions can be related to the excess volume fractions. The advantage of Wohl's method is that a rough physical significance could be attached to the parameters of the equations. Prausnitz, Lichtenthaler, and Gomes de Azevado[2] give a detailed description of Wohl's method and the relation to equations representing the excess Gibbs energy. The Margules and van Laar equations are common relationships that can be derived from Wohl's work. Margules equations are best for molecules of similar size, and van Laar's equations are best for molecules of different sizes.

There are many other equations, which have been proposed, that do not result from Wohl's method. Two of the most popular equations are the Wilson[6] and the universal quasi-chemical theory (UNIQUAC) by Abrams and Prausnitz.[5] These equations are based on the concept of local composition models, which was proposed by Wilson[6] in his paper. It is presumed in a solution that there are local compositions that differ

from that of overall mixture. These local compositions result from differences in molecular size and intermolecular forces. The local compositions account for the short-range order and nonrandom molecular orientations.

Multicomponent Solutions

It is difficult to fit data with the local composition models due to their complex logarithmic forms. However, they are readily generalizable to multicomponent systems. Smith, van Ness, and Abbott[1] and Prausnitz, Lichtenthaler, and Gomes de Azevedo[2] present the Wilson and UNIQUAC models extended for multicomponent solutions. They both employ the constants from binary data. However, the constants are not unique in the sense that valid, but different constants may be obtained from different sets of data. Wilson's equations cannot be employed for immiscible solutions, but the UNIQUAC model may be used to describe such solutions. However, Wilson's equations are good for polar or associating compounds. A compilation of Wilson parameters can be found by Hirata, Ohe, and Nagahama.[7]

The UNIQUAC method of Abrams and Prausnitz[5] divides the excess Gibbs free energy into two parts, the combinatorial part and a part describing the intermolecular forces. The sizes and shapes of the molecule determine the combinatorial part and are thus dependent on the compositions and require only pure component data. As the residual part depends on the intermolecular forces, two adjustable binary parameters are used to better describe the intermolecular forces. The UNIQUAC equations are about as simple for multicomponent solutions as for binary solutions. Parameters for the UNIQUAC equations can be found by Gmehling, Onken, and Arlt.[8]

UNIFAC—A Group Contribution Method

When it is necessary to estimate activity coefficients where no data or very limited data are available, estimates may be made by using a group contribution method. In this case, a molecule is divided into functional groups, or subgroups of the molecule. These subgroups are assumed to act independently of the molecule in which they appear. Molecular interactions are accounted for by properly weighted sums of group interactions. Fredenslund, Jones, and Prausnitz[9] developed the method for UNIQUAC and named it as universal functional activity coefficient (UNIFAC). Smith, van Ness, and Abbott[1] report the equations for the activity coefficients of multicomponent solutions and their parameters. These equations are very similar to the equations for UNIQUAC. Recent sources for the group interaction parameters may be found by Hansen et al.[10] Gmehling, Li, and Schiller,[8] and Fredenslund and Sørensen.[9]

Dealing with the Nonideal Case

The fugacity of the vapor phase can be accurately represented in terms of a fugacity coefficient defined by Eq. (12). For an ideal gas, $\phi_i^v = 1.00$, and the vapor-phase fugacity can be

$$\phi_i^v = \frac{f_i^v}{y_i P} \tag{12}$$

represented by Eq. (2). A most exact representation for vapor–liquid equilibrium can then be written as Eq. (13). A liquid-phase fugacity can be calculated from Eq. (14).

$$y_i^v \phi_i^v P = x_i \gamma_i^l f_i^l \tag{13}$$

In Eq. (13), ϕ_i^v becomes the fugacity coefficient at saturation conditions.

$$f_i^c = \phi_i^{sat} P_i^{sat} \exp\left[\frac{v_i^c (P - P_i^{sat})}{RT}\right] \tag{14}$$

If it is assumed that the specific volume, v_i^c, of the phase is very nearly constant over a large pressure increase, then when the pressure difference in the exponential is small enough, the exponential will be nearly one. Thus, at moderate pressures the fugacity of a condensed phase is nearly equal to the vapor pressure. This approximation was used in Eq. (11) to write Raoult's Law. When the pressure is low enough that the fugacity coefficient is nearly 1.00, Eq. (13) reduces to Raoult's Law extended with the activity coefficient included or Eq. (11). The fugacity coefficient can be calculated from equations of state, and the activity coefficient can be found from various correlations as discussed earlier.

The Equilibrium *K* Value

A useful formulation for VLE in separation calculations is the equilibrium K value defined in Eq. (15).

$$K_i \equiv \frac{y_i}{x_i} = \frac{\gamma_i^l f_i^l}{\phi_i^v P} \tag{15}$$

Under conditions of low pressure when the fugacity coefficient for the vapor is nearly 1.00, then K becomes Eq. (16), which for the case of an ideal solution can

then be simplified to Eq. (17).

$$K_i = \frac{\gamma_i^l P_i^{sat}}{P} \tag{16}$$

$$K_i = \frac{P_i^{sat}}{P} \tag{17}$$

Originally there were charts prepared to make use of these definitions. The DePriester Charts[11] for the homologous series of light hydrocarbons can serve as a basis with approximate validity for simple calculations of VLE situations.

MASS TRANSFER

The Whitman Two-Film Theory

Mass transfer in real absorption equipment resembles a molecular diffusion process only in the basic idea of a concentration difference driving force. However, the two-film theory of Whitman[12] can be used to construct a model similar in many respects to molecular diffusion equations. Fig. 1 is a schematic representing the Whitman two-film theory:

The theory may be summarized as follows:

1. Visualize two films, the gas and liquid films, on either side of the interface.
2. Material is transferred in the bulk of the phases by convection currents. Concentration differences are negligible except in the vicinity of the interface.
3. On each side of the interface, convection currents die out and the two thin films form.
4. Both films offer resistance to the mass transfer. Transfer takes place through these films by a mechanism similar to molecular diffusion.
5. The interface is at equilibrium and offers no resistance to mass transfer.
6. There is uniform composition in the main stream.
7. At the interface, y_{Ai} and x_{Ai} are considered to be in equilibrium described by Henry's law, $y_{Ai} = m_i x_{Ai}$, where m_i is the Henry's law constant.

In a packed absorption column, the fluid is in turbulent motion. Mass transfer through the films is defined by k_y and k_x, which are now turbulent mass transfer coefficients. An equation similar to that for molecular diffusion can be used to describe the mass transfer. However, in this case, the concentration difference is expressed in terms of mole fractions at the interface. The molar mass transferred $\overline{N}_A$ can be found from Eq. (18).

$$\overline{N}_A = k_y(y_A - y_{Ai}) = k_x(x_{Ai} - x_A) \tag{18}$$

Overall Mass Transfer Coefficients

Mass transfer coefficients defined on the basis of interfacial mole fractions have little practical value because

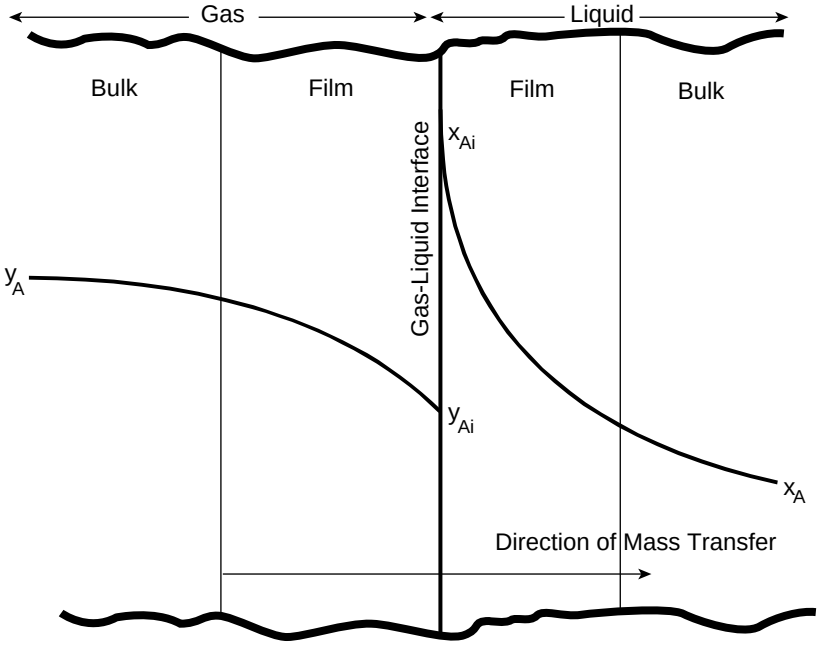


Fig. 1 Whitman two-film theory.

these mole fractions cannot be measured. A new mass transfer coefficient can be defined, which can be determined from measured data. Two new pseudo-mole fractions are defined as illustrated in the Fig. 2. Overall mass transfer coefficients K_y and K_x are defined based on these new pseudo-mole fractions defined below:

y_A^*, equilibrium mole fraction of the solute in the vapor corresponding to the mole fraction x_A in the liquid.

x_A^*, equilibrium mole fraction of the solute in the liquid corresponding to the mole fraction y_A in the vapor.

The mass transfer may now be written as

$$\overline{N}_A = K_y(y_A - y_A^*) = K_x(x_A^* - x_A) \tag{19}$$

The mass balance line or the operating line and the equilibrium curve are both plotted in Fig. 2. The construction in the figure illustrates how to determine y_A^* for any y_A, and x_A^* for any x_A. The overall coefficients can be related as follows to the individual phase coefficients making use of Henry's law constant m.

$$\frac{1}{K_y} = \frac{1}{k_y} + \frac{m}{k_x} \tag{20}$$

Volume-Based Mass Transfer Coefficients

Absorption towers are packed with plastic, metal, or ceramic pieces to provide a greater area for contact between the vapor and the liquid phases. Each of these packings has a characteristic area per unit volume of packing, which can be denoted by the quantity "a," where

$$a = \frac{\text{Interfacial area}}{\text{Unit volume}}$$

The flow rate of both phases, viscosity, density, surface tension, and size and shape of the packing determine the value of "a". These same factors affect the value of the mass transfer coefficients K_y and K_x. Therefore, it is expedient to include "a" in the mass transfer equation and define two new quantities $K_y a$ and $K_x a$. These quantities would then be correlated with the solution parameters as functions of various chemical systems. If A is the absorption tower cross-sectional area, and z the packing height, then Az is the tower packing volume. Defining A_i as the total interfacial area:

$$A_i = aAz \tag{21}$$

We can now rewrite Eq. (19) in terms of a differential rate of mass transfer dN_A, with units (moles/unit unit time) = (moles/unit time – interfacial area) × (interfacial area),

$$\begin{aligned} dN_A &= \overline{N}_A dA_i = K_y a(y_A - y_A^*) A dz \\ &= K_x a(x_A^* - x_A) A dz \end{aligned} \tag{22}$$

Eq. (22) defines the volumetric mass transfer coefficients $K_y a$ and $K_x a$, which have the typical unit moles/(hr m^3 mole fraction). Note also that the term Adz represents the differential packed tower volume. Therefore, Eq. (22) can serve as a model to determine packing height in a packed absorption tower.

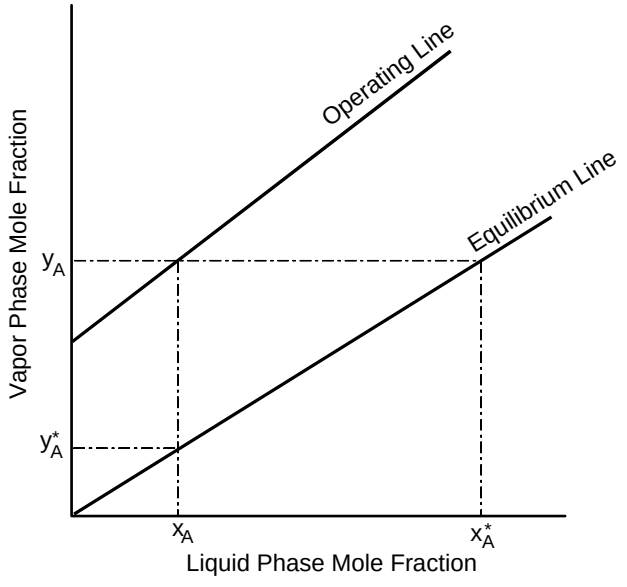


Fig. 2 Pseudo-mole fractions defined.

PACKED TOWER DESIGN

In a packed tower operating in countercurrent flow at a constant liquid rate, the pressure drop varies with the gas mass velocity. With liquid holdup in the packing, there is smaller void space in the column, and a higher pressure drop exists than with dry packing. When plotting gas flow rate vs. pressure drop at a constant liquid flow rate and lower gas flow rates, the liquid flow rate lines are nearly parallel to the dry line up to high liquid flow rates. However, there is a continually increasing pressure drop with gas flow at any liquid flow rate until the slope of the curves become infinite. The point at which the rate of change in pressure drop increases more rapidly than a constant value has been called the upper loading point. The liquid holdup now increases with the increasing gas rate.

Above the loading region, the column will approach the maximum hydraulic capacity at the point where the slope becomes infinite. This is the upper limit to the gas rate. The velocity corresponding to this upper limit is known as the flooding velocity. Flooding can be observed as a crowning of the packing that begins at the bottom and progresses to the top of the packing where a layer of liquid is visible above the packing. At gas rates greater than the flooding velocity, the column will act as a gas bubbler. The greater the liquid rate, the lower the gas velocity at which flooding occurs. Larger more open packings flood at high velocities than smaller, denser packings.

Liquid holdup is a function of liquid flow rate and column pressure drop. Two types of holdup have been defined. Static holdup is the volume of liquid per volume of packing that remains after gas and liquid flows are stopped and bed has drained. Static holdup depends on packing surface characteristics. The second type is operating holdup that is the volume of liquid per volume of packing that drains out of the bed after gas and liquid flows have been stopped. The gas flow rate has little effect on holdup below loading.

Packed towers may be operated above 90% of the flooding velocity when pressure controls are provided to maintain a fixed maximum pressure drop not to exceed the desired flooding velocity. Strigle[13] suggests the following operational characteristics for countercurrent towers.

- Pressure drop between 0.25 and 0.60 in H_2O per foot of packing.
- Air velocity between 5.0 and 8.0 ft/sec with modern high-capacity plastic packing.
- Liquid irrigation rates typically between 2 and 8 gpm/ft^2 of column cross-sectional area.

Choosing a Liquid–Gas Flow Ratio

Fig. 3 illustrates a packed absorption tower with countercurrent flow where flow rates are constant. Making a differential mass balance around the differential cross section in molar units results in Eq. (23). This equation can now be integrated from the top of the column down

$$\mathrm{d}(Gy_\mathrm{A}) = \mathrm{d}(Lx_\mathrm{A}) \tag{23}$$

to the tower cross section, producing Eq. (24), defined as the operating line. A common

$$y_\mathrm{A}G - y_{\mathrm{A}1}G_1 = x_\mathrm{A}L - x_{\mathrm{A}1}L_1 \tag{24}$$

equation for general use, especially when dilute solutions are not expected, is to rewrite the flow rates

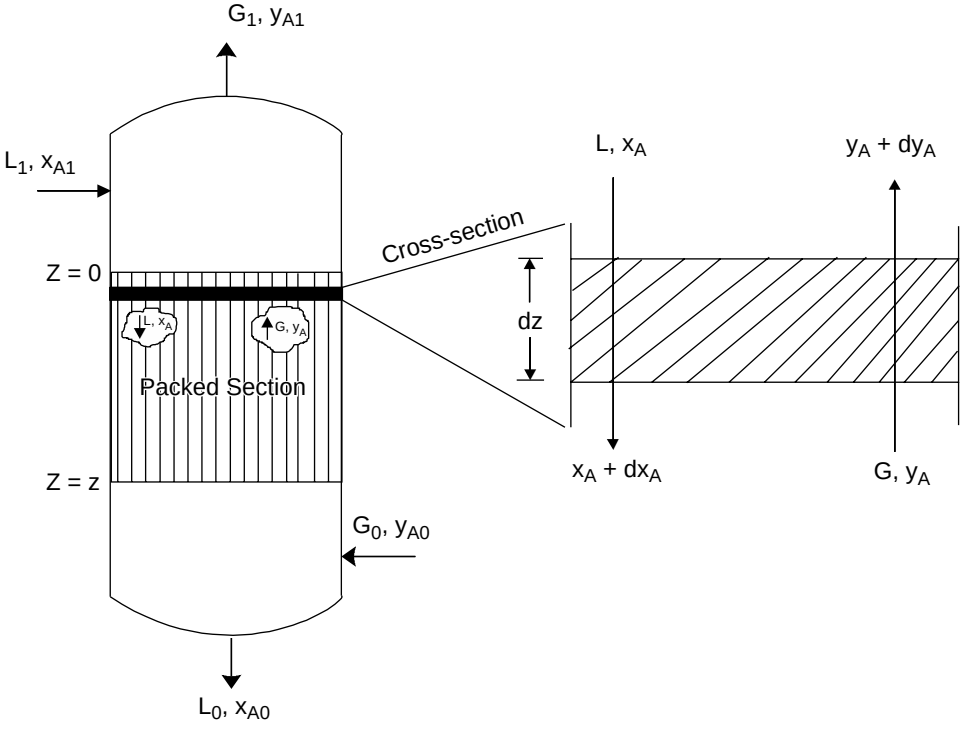


Fig. 3 Countercurrent flow-packed absorption tower.

on a solute-free basis. Eq. (25) uses these solute-free flow rates designated by the subscript B.

$$\left(\frac{y_A}{1 - y_A}\right)G_B + \left(\frac{x_{A1}}{1 - x_{A1}}\right)L_B = \left(\frac{x_A}{1 - x_A}\right)L_B + \left(\frac{y_{A1}}{1 - y_{A1}}\right)G_B \quad (25)$$

For dilute solutions, which might be found in most air pollution control conditions. Eq. (25) reduces to Eq. (24)—a general equation that can be used along with a general equilibrium relationship such as Eqs. (13) and (15). For this case of a dilute solution, Eq. (24) will be used along with Eq. (9), Henry's law. Then, both the equilibrium line and the operating line will plot as a straight line on an x–y plot.

Choosing a Liquid–Gas Flow Rate

When designing an absorption tower for removal of a component, usually the following variables will be known:

G, molar gas flow rate.

L, molar liquid flow rate.

y_{A0}, mole fraction of A, gas at inlet coming from process.

y_{A1}, mole fraction of A, gas at outlet (specified by control agency regulation if discharged to the atmosphere).

x_{A1}, mole fraction of A, liquid into tower, quite frequently is zero if solvent is used only once.

Eq. (24), the material balance up to the tower cross section, can be rearranged as follows.

$$y_{A0} = \left(\frac{L}{G}\right)(x_{A0} - x_{A1}) + y_{A1} \quad (26)$$

Thus, if L or (L/G) can be found, the entire material balance will be solved.

In absorption, on a plot of y_A vs. x_A, the equilibrium curve can be concave up or can be nearly a straight line, or the curve can be "S" shaped, having a concave downward portion. These situations are illustrated in the two equilibrium curves of Fig. 4 on which the operating line as specified by Eq. (26), the material balance, is also plotted. In the case of Fig. 4A with a concave equilibrium line, three possible operating lines are shown. Note that on these plots (L/G) is the slope of the operating line. As can be seen, the line furthest to the right with the smallest slope represents the limit to which the slope of the operating line can be drawn. Thus, this slope represents a minimum (L/G) ratio in which x_{A0} and y_{A0} are actually in equilibrium, and the line could not be drawn with a smaller slope because it would go through equilibrium. As true thermodynamic equilibrium can never be reached practically, this point represents a theoretical minimum (L/G) that never can be attained practically.

In practice, (L/G) ratios are usually set at 1.1–1.7 times this minimum rate. Thus, if G is known, (L/G) is determined and x_{A0} can be found. In the case of Fig. 4B, with an "S"-shaped equilibrium curve, the minimum value of (L/G) is reached at some point within the column where the operating and equilibrium lines are tangent and thermodynamic equilibrium would be "exceeded," if the (L/G) ratio were any smaller.

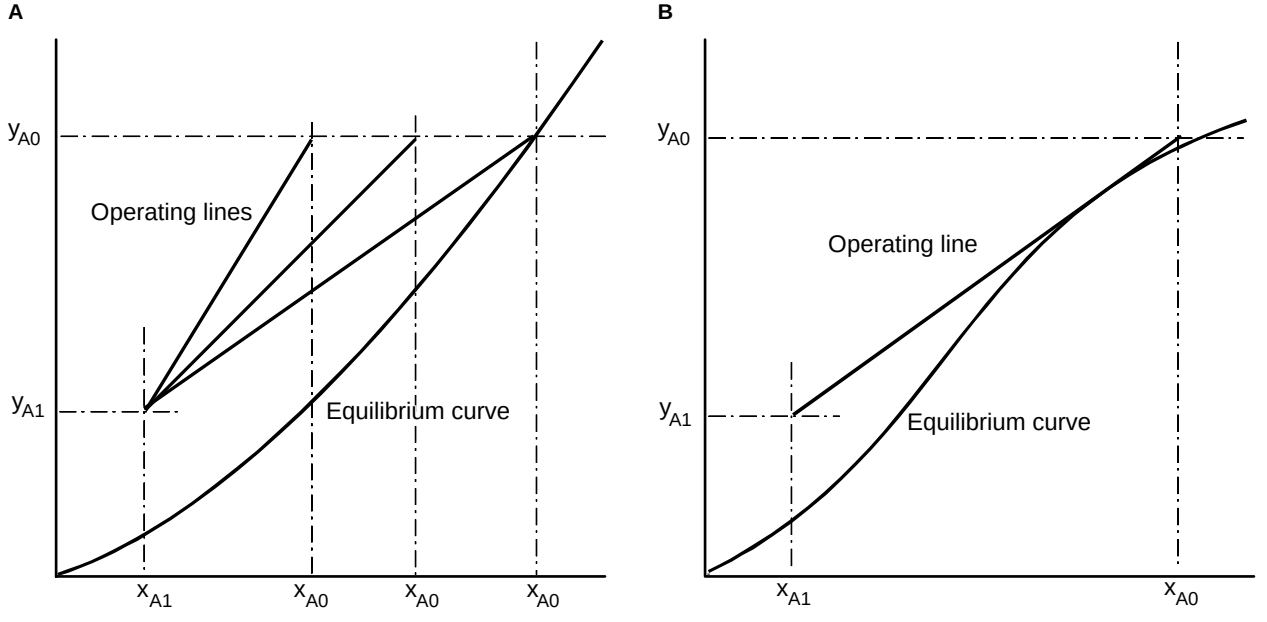


Fig. 4 Operating lines and equilibrium curve relationships.

Determining Tower Diameter—Random-Dumped Packing

From the (L/G) ratio and the overall material balance, having selected the type of packing and the pressure drop per foot of packing, the tower diameter can now be determined. Current accepted practice is to use a modified Sherwood, Shipley, and Holloway[14] flooding correlation to determine tower diameter. Strigle[13] presents such a correlation in Fig. 5 modified with the addition of Sherwood's flooding curve. This correlation uses a linear scale for the ordinate, which is expressed in terms of a capacity factor; C_S. Table 1 presents the definition of terms used in Fig. 5. It should be noted that the pressure-drop correlation ordinate contains a term defined as the packing factor. The packing factor is a characteristic of a particular packing size and shape. Originally it was defined as a^3/ε where a is the interfacial area per packed volume and ε is the void fraction. However, this definition for F did not adequately predict the packing hydraulic performance. Therefore, Table 2 from Strigle[13] presents packing factors that have been calculated from experimentally determined pressure drop.

Since (L/G) is known, we can calculate the value of the abscissa in Fig. 5, if the gas density ρ_G and the liquid density ρ_L are known. Although these densities may vary through the column, the variations will be small, especially in the dilute solution case, and an average of the top and bottom values from the column should be sufficiently accurate. If the entering gas is very dilute, the entering values of density could be used.

From the abscissa, the packing factor, and the required pressure drop per foot of packing, the ordinate of the curve can be read and the value of G^* in $\text{lbs}/\text{ft}^2\,\text{sec}$ calculated. The column cross-sectional area A_s is then:

$$A_s(\text{ft}^2) = \frac{G(\text{lbs/sec})}{G^*\left(\text{lbs}/\text{ft}^2\,\text{sec}\right)} \tag{27}$$

The column diameter can be found from:

$$D = \sqrt{\frac{4G}{\pi G^*}} \tag{28}$$

Determining Tower Diameter—Structured Packing

Parkinson and Ondrey[15] report that structured packing is becoming more frequently used in air separations. They also report that structured packings are being favored because they have a higher capacity than dumped packing, and the scale up is more predicable.

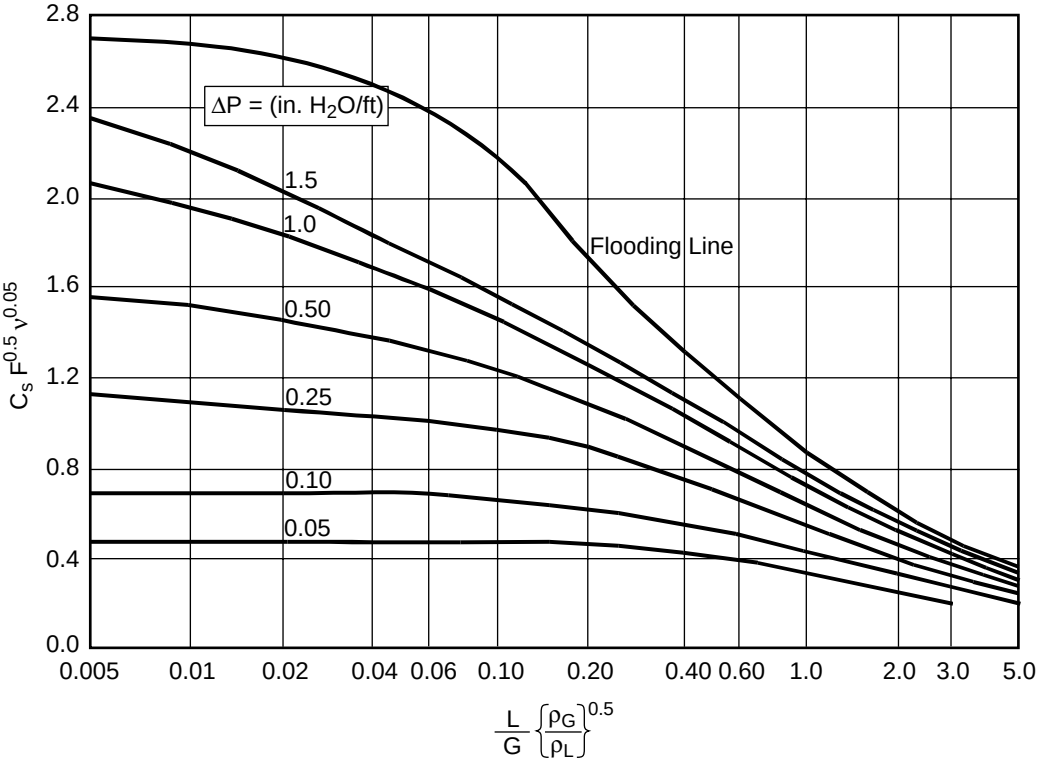


Fig. 5 Generalized pressure-drop correlation. (See Table 1.)

Table 1 Variable definitions and units for Fig. 5

$\overline{L}$	Superficial liquid mass velocity in lbs/ft^2 hr
L	Liquid mass flow rate in lbs/hr
$\overline{G}$	Superficial gas mass velocity in lbs/ft^2 hr
G	Gas mass flow rate in lbs/hr
G^*	Superficial gas mass velocity in lbs/ft^2 sec
ρ_G	Gas density in lb/ft^3
ρ_L	Liquid density in lb/ft^3
V	Kinematic liquid viscosity in cSt
F	Packing factor
$V = G^*/\rho_G$	Superficial velocity in ft/sec
$C_S = V\sqrt{\frac{\rho_G}{\rho_L - \rho_G}}$	Capacity factor in ft/sec

Correlations similar to Fig. 5 have been used to predict the pressure drop in structured packing. Bravo, Rocha, and Fair[16] developed a generalized method of predicting the pressure drop in structured packing. A review of the methods for prediction of the pressure drop in structured packing is presented by Fair and Bravo.[17] Bravo, Rocha, and Fair[16] present a new more comprehensive pressure drop model developed from distillation data. Because the pressure drop curves for random and structured packings are similar, Strigle[13] reports that the pressure drop for structured packing can be estimated from the generalized correlation of Fig. 5. However, in this case the operating pressure affects the packing factor.

Determining Height of Packing in the Tower: The HTU Method

The differential rate of mass transfer given by Eq. (22) is equal to the differential rate of change in mass within a phase, Eq. (23). For the gas phase, the differential rate of mass transfer of component A is equal to the differential rate of change in the mass of A in the incoming gas stream. It can be shown that the differential tower height, dz, is given by Eq. (29). A similar equation based on the liquid phase can be written, which would be useful in

$$\frac{G\mathrm{d}(y_A)}{(1 - y_A)} = K_y a\left(y_A - y_A^*\right) A\mathrm{d}(z) \tag{29}$$

stripping calculations. We now define a mass transfer coefficient, Eq. (30), based on the logarithmic mean concentration, Eq. (31). This mass transfer coefficient is independent of

$$K_y = \frac{K_y^0}{(1 - y_A)_{LM}} \tag{30}$$

$$(1 - y_A)_{LM} = \frac{(1 - y_A) - \left(1 - y_A^*\right)}{\ln\left[\frac{(1 - y_A)}{(1 - y_A^*)}\right]} \tag{31}$$

Table 2 Packing factors, F, for random-dumped packings

Normal packing size (in.)		1/2	5/8	3/4	1	1/4	1 1/2	2	3 or 3 1/2
IMTP packing (metal)		51		41		24	18	12	
Hy-pak packing (metal)					45		29	26	16
Super intalox saddles (ceramic)					60			30	
Super intalox saddles (plastic)					40			28	18
Pall rings (plastic)			95		55		40	26	17
Pall rings (metal)			81		56		40	27	18
Intalox saddles (ceramic)		200		145	92		52	40	22
Raschig rings (ceramic)	580	380	255	179	125	93	65	37	
Raschig rings (1/32-in. metal)		300	170	155	115				
Raschig rings (1/16-in. metal)		410	300	220	144	110	83	57	32
Berl saddles (ceramic)		240		170	110		65	45	

(From Ref.[13].)

flow rate through the column. Define the set of flow rates based on the tower cross section. This

$$\overline{G} = \frac{G}{A} \quad \text{and} \quad \overline{L} = \frac{L}{A} \tag{32}$$

equation can now be integrated down the column from $z = 0$ to $z = z$ to determine the tower height shown in Eq. (33). The ratio of flow rate to mass transfer coefficient has been designated

$$z = \left(\frac{\overline{G}}{K_y^0 a}\right) \int_{y_{A1}}^{y_{A0}} \frac{(1 - y_A)_{LM}}{(1 - y_A)(y_A - y_A^*)} dy_A \tag{33}$$

as the height of a transfer unit (HTU) or for the gas phase, H_{OG}. Therefore, H_{OG} has been

$$H_{OG} = \left(\frac{\overline{G}}{K_y^0 a}\right) \tag{34}$$

defined in such a way that it remains constant through the absorption column. The integral portion of Eq. (33) is designated as the number of overall mass transfer units (NTU) or for the gas phase N_{OG}. Thus

$$N_{OG} = \int_{y_{A1}}^{y_{A0}} \frac{(1 - y_A)_{LM}}{(1 - y_A)(y_A - y_A^*)} dy_A \tag{35}$$

The integral may be evaluated by graphical or numerical techniques. The height of the column may now be calculated from

$$z = H_{OG} N_{OG} \tag{36}$$

Dilute solution case

For dilute solutions, Henry's law is usually a good choice for an equilibrium relationship. In this case, $y_A^* = mx$, which is defined in relation to the overall mass transfer by Eq. (19). For a dilute solution

$$(1 - y_A) \approx (1 - y_A)_{LM} \approx 1.0 \tag{37}$$

and Eq. (35) reduces to

$$N_{OG} = \int_{y_{A1}}^{y_{A0}} \frac{dy_A}{(y_A - y_A^*)} \tag{38}$$

The operating line, Eq. (24), may now be rewritten as

$$y_A = \left(\frac{L}{G}\right)(x_A - x_{A1}) + y_{A1} \tag{39}$$

Note that the ratio $\frac{L}{G} = \frac{\overline{L}}{\overline{G}}$ and introducing Henry's law constant m, the absorption factor A can be defined as

$$A = \frac{L}{mG} = \frac{\overline{L}}{m\overline{G}} \tag{40}$$

Recall that $y_A^* = mx_A$ and solve Eq. (39) for y_A^* after multiplying and dividing by Henry's law constant m, and Eq. (41) results.

$$y_A^* = \left(\frac{y_A - y_{A1}}{A}\right) + mx_{A1} \tag{41}$$

Substituting Eq. (41) into Eq. (38) and integrating, we get:

$$N_{OG} = \frac{\ln\left[\left(\frac{y_{A0} - mx_{A1}}{y_{A1} - mx_{A1}}\right)\left(1 - \frac{1}{A}\right) + \left(\frac{1}{A}\right)\right]}{\left(1 - \frac{1}{A}\right)} \tag{42}$$

When a pure solvent such as water is used, $x_{A1} = 0.0$, and Eq. (42) reduces to

$$N_{OG} = \left(\frac{1}{1 - \frac{1}{A}}\right) \ln\left[\left(\frac{y_{A0}}{y_{A1}}\right)\left(1 - \frac{1}{A}\right) + \frac{1}{A}\right] \tag{43}$$

Colburn[18] put Eq. (43) into graphical form, which makes it easier to use. (See Ref.[19].)

Correlation for the effect of the L/G ratio on the NTU

Eq. (43) is based on Henry's law, which is a straight line on a y_A vs. x_A plot. On the same plot Eq. (39), the operating line, will intersect Henry's law line at the point where $y_{A0} = mx_{A0}$, which can be seen in Fig. 4. This represents the minimum L/G ratio, $(L/G)_{min}$. Eq. (43) can then be redefined as follows. Recall that Eq. (43) is based on $x_{A1} = 0.0$. Solve Eq. (39) for $(L/G)_{min}$, substitute the equilibrium value for x_{A0}, and rearrange the equation with the following result.

$$\left(\frac{L}{G}\right)_{min} = \left(1 - \frac{y_{A1}}{y_{A0}}\right) m \tag{44}$$

Define b as a multiple of $(L/G)_{min}$ to set the actual (L/G), then

$$\left(\frac{L}{G}\right) = b\left(\frac{L}{G}\right)_{min} \tag{45}$$

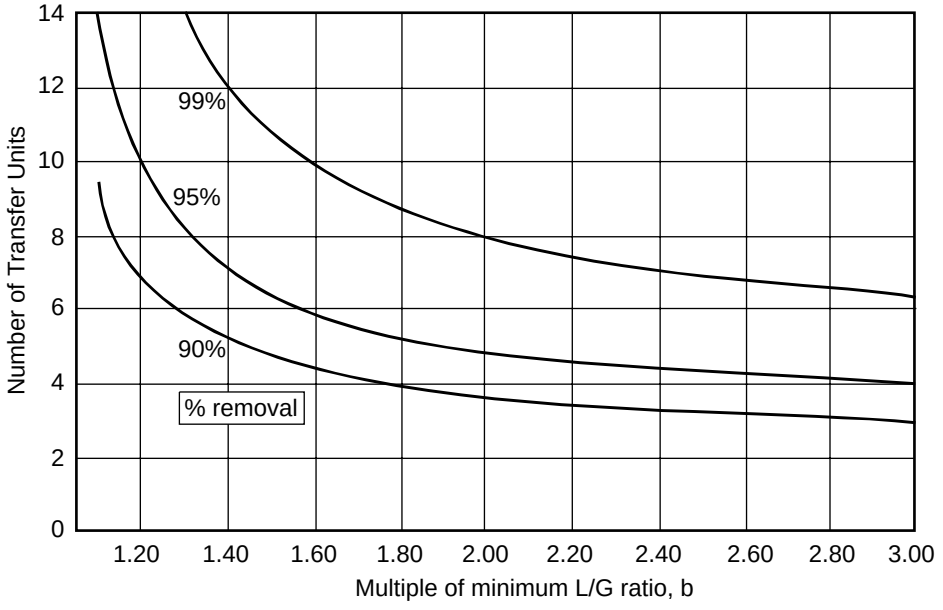


Fig. 6 Effect of the multiple of minimum L/G ratio on the number of transfer units.

and define Ab in the following equation

$$Ab = \frac{b(L/G)_{\min}}{m} = b(1 - y_{A1}/y_{A0}) \tag{46}$$

Under these conditions, the operating line may now be written as

$$y_A = \frac{y_A^*}{m}\left(\frac{L}{G}\right) + y_{A1} \tag{47}$$

which can be substituted into Eq. (38) and integrated with the following result.

$$N_{OG} = \left(\frac{1}{1 - \frac{1}{Ab}}\right) \ln\left[\left(\frac{y_{A0}}{y_{A1}}\right)\left(1 - \frac{1}{Ab}\right) + \frac{1}{Ab}\right] \tag{48}$$

This equation looks the same as Eq. (43) with Ab substituted for A. As Ab is a function of b, which is a multiple of $(L/G)_{\min}$, Fig. 6 is a plot of Eq. (48), N_{OG} vs. b, with (y_{A1}/y_{A0}) as the parameter fixed at 0.90, 0.95, and 0.99 and $x_{A0} = 0.0$. Note that this parameter is the fraction of material A removed from the inlet stream.

Knowing values for $\overline{G}$ and $K_y^0 a$, H_{OG} can be calculated from Eq. (34). For any given value of b, Fig. 6 can be used to determine N_{OG}, if the removal fraction required is between 0.95 and 0.99. Fig. 6 shows that, for dilute solutions, the number of transfer units cannot be materially reduced by increasing b, the multiple of $(L/G)_{\min}$ above 1.70.

COCURRENT TOWER DESIGN

Wankat[20] has a description of the design method for cocurrent packed towers. The procedure to determine the tower height using the HTU method follows that for countercurrent flow.

However, now both streams enter the column at the top. H_{OG} is based on Eq. (39), and N_{OG} is now written as follows.

$$N_{OG} = \int_{y_{A0}}^{y_{A1}} \frac{dy_A}{(y_A - y_A^*)} \tag{49}$$

Eq. (39), the operating line, now becomes

$$y_A = \left(\frac{L}{G}\right)(x_{A1} - x_A) + y_{A1} \tag{50}$$

With a straight operating line and a straight equilibrium line (Henry's Law for example), where the entering absorption liquid is pure, i.e., $x_{A0} = 1.0$, N_{OG} as shown in Eq. (48), it now becomes

$$N_{OG} = \left(\frac{1}{1 + \frac{1}{Ab}}\right) \ln\left[\left(\frac{y_{A1}}{y_{A0}}\right)\left(1 + \frac{1}{Ab}\right) + \frac{1}{Ab}\right] \tag{51}$$

HENRY'S LAW CONSTANTS AND MASS TRANSFER INFORMATION

Limited Henry's Law constants and mass transfer information can be found by Perry and Green.[19]

Yaws, Yang, and Pan[21] have published a listing of Henry's Law constants for 362 organic compounds in water. Strigle[13] also lists mass transfer coefficients for dumped packing. Mass transfer coefficients for structured packing are usually greater than in dumped packing. Manufacturers of structured packing should be consulted for available mass transfer data.

CONCLUSIONS

The methodology and models for packed column design were developed in this entry. The methods are valid for isothermal columns where heat generation from reactions would be negligible. Vapor–liquid equilibrium data are necessary to carry out the design; these data are frequently available in the literature. Mass transfer information is necessary as well to carry out the design. However, mass transfer data are much more difficult to come by. General information required for determining pressure drop is available in the literature. However, companies producing packing can supply data more accurately for their particular packings. Although random or dumped packing is still used, stacked packing is gaining much favor for use in columns. Stacked packing can usually be handled much easier and most usually is more efficient for mass transfer and provides lower pressure drop.

Plate Columns: Plate columns or equilibrium stage columns remain available for absorption operations. They are not as frequently used as the packed column. The design proceeds similar to that of a distillation plate column. As the distillation stage or plate column design began to evolve, it was recognized that the system of equations was very difficult to model. Sorel[22] developed the equilibrium model to solve this problem. In this model, it is assumed that the vapor and the liquid leaving stage are in equilibrium. It is then possible to use thermodynamic calculations to determine the phase equilibria and the energy balance. In carrying out design, it is assumed that the column is composed of equilibrium stages, and the number of equilibrium stages is calculated. These ideal stages are converted to actual stages by means of tray efficiencies. Then, the column diameter can be determined. A graphical solution to Sorel's method developed independently by Ponchon[23] and Savarit[24] solves both the mass and energy balance and phase equilibrium relationships simultaneously stage by stage. This same concept can be applied to absorption column design. Consult Wankat[20] for a discussion of the methodology.

Multicomponent Systems: According to Sherwood and Pigford,[25] in the recovery of natural gasoline and the treatment of refinery gases the problem of multicomponent solutions was of enormous importance in the petroleum industry. The gas to be treated usually consisted of mixtures of methane and several other lower hydrocarbons. The solvent used was usually a nonvolatile light hydrocarbon oil in which the hydrocarbons to be absorbed were highly soluble. In the petroleum industry, plate columns were used and the theoretical-plate concept was employed in the design calculations. Strigle[13] reports on a design technique that could be used for packed absorption tower design. A key component is selected and the absorption efficiency for that component is set. The depth of packing and solvent rate is determined for that component. Then, the effect on the other components is considered. The mass transfer coefficient and driving force will differ for the other components of the solution. Therefore, the economic effects of the removal of the other components need to be considered and the design changed, if necessary, to meet those requirements. Strigle[13] discusses several examples in detail.

REFERENCES

1. Smith, J.M.; van Ness, H.C.; Abbott, M.M. *Introduction to Chemical Engineering Thermodynamics*, 7th Ed.; The McGraw-Hill Companies, Inc., 2005.
2. Prausnitz, J.M.; Lichtenthaler, R.N.; Gomes de Azevado, E. *Molecular Thermodynamics of Fluid-Phase Equilibria*, 3rd Ed.; Prentice Hall: Upper Saddle River, New Jersey, 1999.
3. Lewis, G.N.; Randall, M. *Thermodynamics.* McGraw-Hill book Co., 1923.
4. Wohl, K. Thermodynamic evaluation of binary and ternary liquids. Trans. AIChE **1946**, *42*, 215–249.
5. Abrams, D.S.; Prausnitz, J.M. Statistical thermodynamics of liquid mixtures: a new expression for the excess Gibbs energy of partly or completely miscible systems. AIChE J. **1975**, *21* (1), 116–128.
6. Wilson, G.M. Vapor–liquid equilibrium. XI. A new expression for the excess free energy of mixing. J. Am. Chem. Soc. **1964**, *86* (2), 127–130.
7. Hirata, M.; Ohe, S.; Nagahama, K. *Computer Aided Data Book of Vapor–Liquid Equilibria*; Elsevier Scientific Publishing Company: New York, 1975.
8. Gmehling, J.; Onken, U.; Arlt, W. *Vapor–Liquid Equilibrium Data Collection;* Chemistry Data Series; DECHEMA: Frankfurt am Main, 1974–1990; Vol. 1, Parts 1–8.
9. Fredenslund, A.; Jones, R.L.; Prausnitz, J.M. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. AIChE J. **1975**, *21* (6), 1086–1099.
10. Hansen, H.K.; Rasmussen, P.; Fredenslund, A.; Schiller, M.; Gmehling, J. Vapor–liquid equilibria

by UNIFAC group contribution. 5. Revision and extension. Ind. Eng. Chem. Res. **1991**, *30* (10), 2352–2355.
11. DePriester, C.L. The *K* values for systems of light hydrocarbons. Chem. Eng. Proc. Symp. Ser. **1953**, *49* (7), 41–42.
12. Whitman, W.G. The two film theory of gas absorption. Chem. Metal. **1923**, *2*, E146–E148.
13. Strigle, R.F. Packed Tower Design and Applications. Random and Structured Packings, 2nd Ed.; Gulf Publishing Co.: Houston, 1994.
14. Sherwood, T.K.; Shipley, G.H.; Holloway, F.A.L. Flooding velocities in packed columns. Ind. Eng. Chem. **1938**, *30* (7), 765–769.
15. Parkinson, G.; Ondrey, G. Packing Towers. Chem. Eng. Newsfront, **1999**, (December), 39.
16. Bravo, J.L.; Rocha, J.A.; Fair, J.R. Pressure drop in structured packings. Hydrocarbon Process. **1986**, *65* (3), 45–49.
17. Fair, J.R.; Bravo, J.L. Distillation columns containing structured packing. Chem. Eng. Prog. **1990**, *86* (1), 19–29.
18. Colburn, A.P. Simplified calculation of diffusional processes. Ind. Eng. Chem. **1961**, *33*, 459–467.
19. Perry, R.H.; Green, D.W. *Chemical Engineer's Handbook*, 7th Ed.; McGraw-Hill Book Company: New York, 1997.
20. Wankat, P.C. *Equilibrium Staged Separations*; Prentice Hall: Upper Saddle river, New Jersey, 1988; 369–379.
21. Yaws, C.; Yang, H.; Pan, X. Henry's Law constant for 362 organic compounds in water. Chem. Eng. **1991**, *97* (November), 179–185.
22. Sorel, E. *La Rectification de l'Alcool*; Gauthier-Villiers: Paris, 1893.
23. Ponchon, M. Etude graphique de la distillation fractionée industrielle. Tech. Moderne **1921**, *13* (20), 53–55.
24. Savarit, R. Eléments de distillation. Arts and Mètiers **1922**, *65*, 142, 178, 241, 266, 307.
25. Sherwood, T.K.; Pigford, R.L. *Absorption and Extraction*, 2nd Ed.; McGraw-Hill Book Company, Inc., 1952.

Particle–Particle Interaction: Improvements in the Prediction of DLVO Forces

Anh V. Nguyen
Linh T. T. Tran
Discipline of Chemical Engineering, The University of Newcastle, Callaghan, New South Wales, Australia

Jan D. Miller
Department of Metallurgical Engineering, University of Utah, Salt Lake City, Utah, U.S.A.

INTRODUCTION

Particle–particle interaction is central to a wide range of engineering applications and processing industries. Examples include coagulation, flocculation, dispersion, emulsification, and froth flotation. In these applications, the particle size is small, and the overall particulate behavior is determined by forces associated with the surface properties rather than those related to mass or volume. The surface properties of a particle in a liquid medium are the result of a complex interaction between molecules, atoms, and ions at the particle surface and in the surrounding liquid. If a number of particles are present, interactions also take place between particles at short separation distances, and it is this interaction that is of most interest as it can determine the overall stability or instability of dispersions and/or suspensions.

It is customarily assumed that the overall particle–particle interaction can be quantified by a net surface force, which is the sum of a number of independent forces. The most often considered force components are those due to the electrodynamic or van der Waals interactions, the electrostatic double-layer interaction, and other non-DLVO interactions. The first two interactions form the basis of the celebrated Derjaguin–Landau–Verwey–Overbeek (DLVO) theory on colloid stability and coagulation. The non-DLVO forces are usually determined by subtracting the DLVO forces from the experimental data. Therefore, precise prediction of DLVO forces is also critical to the determination of the non-DLVO forces. The surface force apparatus and atomic force microscopy (AFM) have been used to successfully quantify these interaction forces and have revealed important information about the surface force components. This chapter focuses on improved predictions for DLVO forces between colloid and nano-sized particles. The force data obtained with AFM tips are used to illustrate limits of the renowned Derjaguin approximation when applied to surfaces with nano-sized radii of curvature.

This entry is organized in the following paragraphs: First, the advanced determination of van der Waals interaction between spherical particles is described. Second, the relevant approximate expressions and direct numerical solutions for the double-layer interaction between spherical surfaces are reviewed. Third, the experimental data obtained for AFM tips having nano-sized radii of curvature and the DLVO forces predicted by the Derjaguin approximation and improved predictions are compared. Finally, a summary of the review and recommended equations for determining the DLVO interaction force and energy between colloid and nano-sized particles is included.

VAN DER WAALS INTERACTIONS BETWEEN COLLOIDAL PARTICLES

There are essentially two approaches for the determination of the van der Waals interaction energy and forces. In the first, due largely to Hamaker,[1] the interaction between macroscopic bodies is calculated by a pairwise summation of all the relevant microscopic interactions, which are assumed to be nonretarded and additive. In reality, the microscopic interactions are retarded by the nearby atoms and molecules, which are included in the classical theories on retardation via the finite speed of light propagating in dispersed media.[2] The second, more rigorous, approach developed by Lifshitz[3] is based on quantum-mechanics theory and depends entirely on the macroscopic electrodynamic properties of the interacting particles and the media such as dielectric constants and refractive indices. The electric fields established by the fluctuating dipoles are considered to interact both constructively and destructively. The result of these fluctuating and many-body interactions is the formation of a standing wave between the bodies whereby only certain modes, or frequencies of electromagnetic radiation may pass through. The summation of all fluctuation modes in

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030608

the electromagnetic field gives the van der Waals interaction energy.[4] The full theory is complicated, and this is probably the reason that the additive approach of Hamaker is still used by many scientists. In some cases, the results of the Hamaker theory are accurate when an experimentally determined Hamaker constant is used. For many practical systems of colloidal particles, the Hamaker constant is not known but the Hamaker prediction can be useful if the (effective) Hamaker constant (function) is determined from the Lifshitz continuum approach.

Hamaker Microscopic Approach

In the Hamaker approach, the energies of all the atoms in one body with all the atoms in the other one are summed, leading to an integral expression for the interaction energy, E, between the two macroscopic bodies: $E_{vdW} = -\oint\oint (C\rho_1\rho_2/r^6)\mathrm{d}v_1\mathrm{d}v_2$, where C is the coefficient of the appropriate interaction between elementary particles separated by distance r and $\mathrm{d}v_i$ ($i = 1, 2$) the volume elements of bodies 1 and 2 at distance r, and with densities of atoms or molecules of ρ_i. The problem appears simple since it only requires the calculation of a closed integral. However, the results in closed analytical forms are available only for some simple systems.[4] The Hamaker prediction for interaction energy splits into two terms: the Hamaker constant described by $A = \pi^2 C\rho_1\rho_2$ and the second term that represents the geometrical dependence of the van der Waals energy.

For the van der Waals energy per unit area, E^*_{vdW}, between two half-spaces separated by a distance h, the Hamaker theory gives

$$E^*_{vdW} = -\frac{A}{12\pi h^2} \tag{1}$$

The interaction force, F, is then determined by the first derivative of the energy with respect to the distance by $F(h) = -\mathrm{d}E/\mathrm{d}h$.

The Hamaker interaction energy for two spheres separated by the inter-surface shortest distance, h, is described by

$$E_{vdW} = -\frac{A}{6}\left\{\frac{2R_1R_2}{r^2 - (R_1+R_2)^2} + \frac{2R_1R_2}{r^2 - (R_1 - R_2)^2} + \ln\frac{r^2 - (R_1 + R_2)^2}{r^2 - (R_1 - R_2)^2}\right\} \tag{2}$$

where R_1 and R_2 are the radii of two spheres and r the inter-center distance, $r = R_1 + R_2 + h$. We will see later that Eqs. (1) and (2) are useful for combining the Hamaker microscopic and Lifshitz macroscopic theories.

Lifshitz Macroscopic Approach

The description of the van der Waals interaction based on the Lifshitz[3,5] approach is now sufficiently advanced to provide accurate predictions for the complete interaction energy. For the geometry of two half-spaces, the exact theory is available in a formulation suited for computational purposes.[4] In parallel with work on planar systems, there has been a focus on the interaction between spheres.[4,6–8] These developed theories have been used as the exact solutions in the validation of the approximate predictions using the Hamaker approach. The significant contribution of the continuum approach to our understanding of the van der Waals interaction lies in the reliable prediction of the Hamaker constant. The interaction energy for two half-spaces and two spheres is summarized below.

The van der Waals interaction energy, E^*_{132}, per unit area between two half-spaces 1, 2 immersed in a medium 3 as a function of the separation, h, is described by the Lifshitz macroscopic theory as follows:[4]

$$E^*_{132}(h) = \frac{k_BT}{8\pi h^2}{\sum_{N=0}^{\infty}}' \int_{x_N}^{\infty} x\ln[(1 - y_{13}y_{23}e^{-x}) \times (1 - z_{13}z_{23}e^{-x})]\mathrm{d}x \tag{3}$$

where $y_{a3} = \frac{x\varepsilon_a - s_a\varepsilon_3}{x\varepsilon_a + s_a\varepsilon_3}$, $z_{a3} = \frac{x - s_a}{x + s_a}$, $\xi_N = 2N\pi k_BT/\hbar$, $x_N = 2h\xi_N\sqrt{\varepsilon_3}/c$, $\varepsilon_a = \varepsilon_a(i\xi_N)$, and $s_a^2 = x^2 + x_N^2\{\varepsilon_a/\varepsilon_3 - 1\}$. In these equations, the subscript $a = 1$ to 3, $i = \sqrt{-1}$, c is the speed of light, k_B Boltzmann's constant, T the absolute temperature, $\hbar$ the Planck constant divided by 2π, and $i\xi_N$ are the discrete equally spaced imaginary frequencies. The prime on the summation symbol indicates that the zero-frequency term, which accounts for the contributions due to the orientation and induction interactions, is divided by 2. Eq. (3) shows that the interaction energy depends on electromagnetic fluctuations, via the relative permittivity, $\varepsilon(i\xi_N)$, of all three materials.

For two spheres separated by the (shortest) distance, h, the interaction energy is given by the Lifshitz macroscopic theory:[6,9]

$$E_{132}(h) = k_BT\sum_{N=0}^{\infty}\sum_{m,n=1}^{\infty}(2m+1)(2n+1)\Delta_m\Delta_n \times \sum_{\mu=-\mu_m}^{\mu_m} V^{\mu}_{mn}(K_3r)V^{-\mu}_{nm}(K_3r) \tag{4}$$

where

$$\Delta_m = \frac{\begin{array}{r}(x_3^2 - x_1^2)mi_m(x_3)i_m(x_1) + x_3x_1[x_1i_m(x_1)i_{m-1}(x_3) \\ - x_3i_m(x_3)i_{m-1}(x_1)]\end{array}}{\begin{array}{r}(x_3^2 - x_1^2)mk_m(x_3)i_m(x_1) - x_3x_1[x_1i_m(x_1)k_{m-1}(x_3) \\ + x_3k_m(x_3)i_{m-1}(x_1)]\end{array}}$$

$$V_{mn}^{\mu}(x) = U_{mn}^{\mu}(x) + \frac{n - \mu + 1}{(n + 1)(2n + 1)} xU_{mn+1}^{\mu}(x) - \frac{n + \mu}{n(2n + 1)} xU_{mn-1}^{\mu}(x),$$

$$U_{mn}^{\mu}(x) = \left(\frac{2}{x}\right)^{\mu} \sum_{v=0}^{v_m} S_{mn}^{\mu v} k_{m+n-\mu-2v}(x),$$

$$V_{mn}^{-\mu}(x) = \frac{(n - \mu)!(m - \mu)!}{(n + \mu)!(m + \mu)!} V_{mn+1}^{\mu}(x),$$

and

$$S_{mn}^{\mu v} = \frac{\begin{array}{r}\Gamma(m - v + 1/2)\Gamma(n - v + 1/2)\Gamma(m + v + 1/2) \\ \times (m + n - v)!(m + n - \mu - 2v + 1/2)\end{array}}{\begin{array}{r}\Gamma(m + n - \mu - v + 3/2)\Gamma(\mu + 1/2)\Gamma(1/2) \\ \times (m - \mu - v)!(n - \mu - v)!v!\end{array}}.$$

In these equations, $K_a = (\xi_N/c)\sqrt{\varepsilon_a(i\xi_N)}$ ($a = 1, 2, 3$), $x_a = K_aR, \varepsilon_a$ is the permittivity, $\mu_m = \min(m, n)$, $v_m = \min(m - \mu, n - \mu)$, r the inter-center distance, $r = R_1 + R_2 + h$, Γ the gamma function, and i_m and k_m are the modified spherical Bessel functions of the order m and of the first and second kind, respectively. The susceptibility function Δ_m is purposely used for particle 1 with radius $R = R_1$. The susceptibility function Δ_n is used for particle 2 and can be described similarly with x_1 replaced by x_2 and $R = R_2$. There is a typographic error in Δ_m given in:[9] the term in the brackets must be multiplied by the particle radius. The modified spherical Bessel functions are not uniquely defined in the literature. In this chapter, the functions are defined by the respective modified Bessel functions: $i_m(x) = \sqrt{\pi/(2x)}I_{m+1/2}(x)$ and $k_m(x) = \sqrt{2/(\pi x)}K_{m+1/2}(x)$. Because the definition of modified spherical Bessel functions of the third kind is not unique in the literature, the Bessel functions of the third kind are not used in the above equations.

The convergence of Langbein's solution described by Eq. (4) is slow. Pailthorpe and Russel[9] evaluated Eq. (4) by computing the difference between the retarded and nonretarded interactions, which converges much faster. They then evaluated the nonretarded interaction by using Love's prediction,[7] which is faster than evaluating Langbein's nonretarded expression, and added this to the difference. Fig. 1 shows the results of the calculation and comparison with different models.

The full spectra of permittivities are required in the calculation of the van der Waals interaction energy using Eqs. (3) and (4). For water and a few materials, the dependence of the permittivity, ε_a, on the sampling frequency, ξ_N, is available (see Fig. 2).[10–12] For highly polar liquids such as water, the relaxation in the microwave and infrared appears significant and the oscillator model for the spectrum of water permittivity has a number of terms as described by Eq. (5), and the model parameters are given in Table 1.

$$\varepsilon(i\xi) = 1 + \frac{d_m}{1 + \xi\tau_m} + \sum_j \frac{f_j}{\omega_j^2 + g_j\xi + \xi^2} \quad (5)$$

For other materials, the permittivity spectra can be approximately determined using the refractive index, $\sqrt{B}$, and the static dielectric constant, $\varepsilon(0)$, as follows:

$$\varepsilon(i\xi_N) = \begin{cases} \dfrac{B + (\xi_N/\omega)^2}{1 + (\xi_N/\omega)^2} & \text{for } N \geq 1 \\ \varepsilon(0) & \text{for } N = 0 \end{cases} \quad (6)$$

where ω is the characteristic relaxation frequency of the UV region, which is $\sim 2 \times 10^{16}$ rad/s.[14] The nonzero sampling frequency ξ_N ($N \geq 1$) begins at $\xi_1 = 2.4 \times 10^{14}$ rad/s and is closely spaced in the UV region; the relaxation in the UV dominates the portion of the dielectric spectra most important for the van der Waals interaction. For most nonmetallic materials with simple spectra, a single UV relaxation may be sufficient for the Hamaker constant calculation.[12] Thus, the simple oscillator model described by Eq. (6) is sufficient for determining the van der Waals energy. In this simplification, water presents an important exception: the permanent dipole contribution of water is only taken into account in the static dielectric constant but is ignored in calculating $\varepsilon(i\xi)$ at nonzero frequencies. The best fit with the available permittivity for water gives $B = 1.887$, which is slightly higher than the square of the refractive index of water. Parameters of Eq. (6) for some materials are given in Table 2. The simplified model given by Eq. (6) also allows simple equations for van der Waals energy to be developed.

Approximate and Simplified Equations for van der Waals Interaction Between Spheres

A number of approximate models have been developed for the calculation of the van der Waals interaction energy between two spherical particles. The

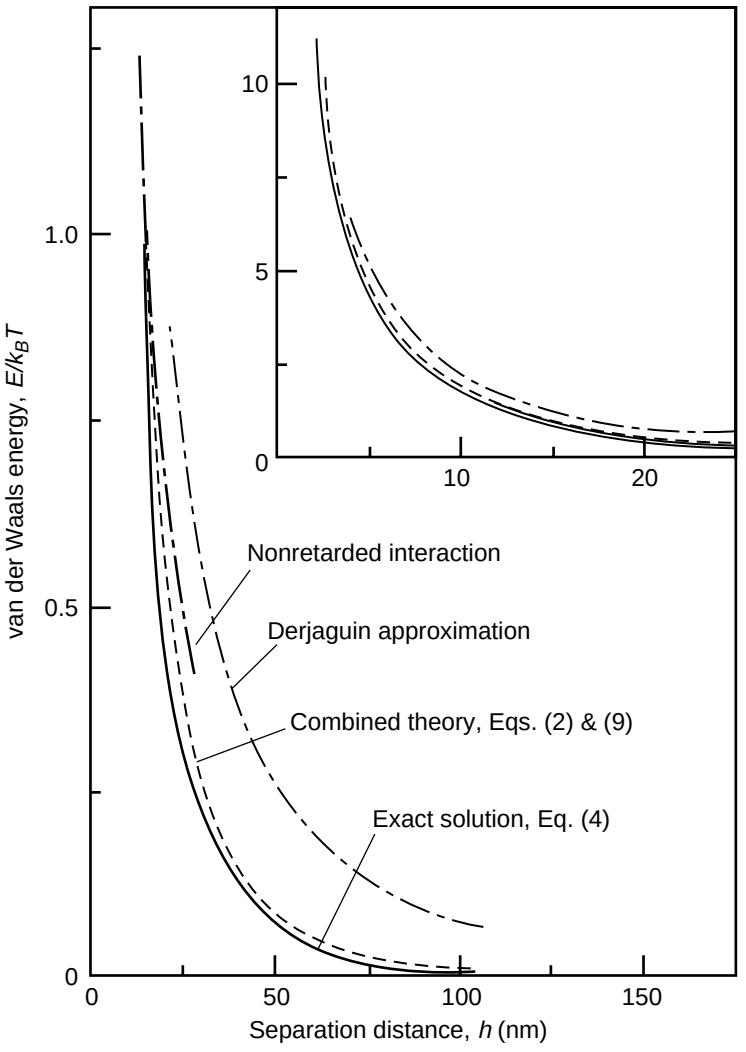


Fig. 1 Comparison between approximate models and the exact solution for the van der Waals energy for two polystyrene spheres of radius $R = 250$ nm in water. (From Ref.[10].)

most well-known approximation was provided by Derjaguin[15] who made the approximate estimation of the interaction energy or the force between two spherical particles or curved surfaces from the corresponding interaction energy per unit area between two infinite parallel flat plates. In terms of force, $F(h)$, between two spherical particles with radii R_1 and R_2, and energy per unit area between parallel flat plates,

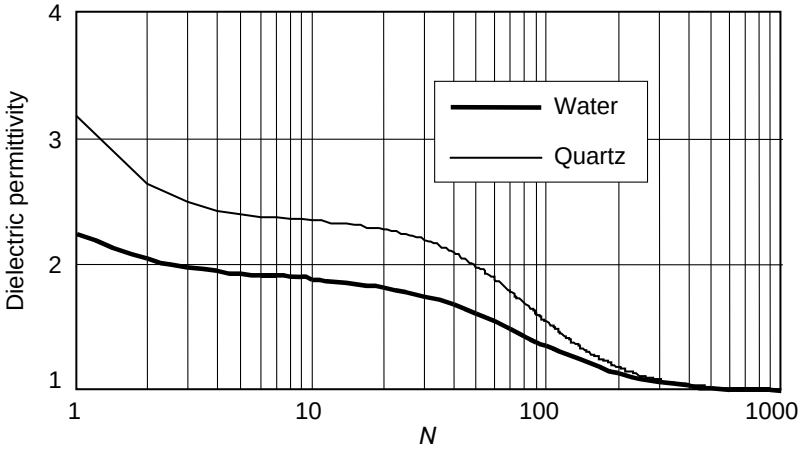


Fig. 2 Dielectric permittivity $\varepsilon(i\xi_N)$ vs. the index N of the equally spaced real frequency ξ_N for water and crystalline quartz. (From Ref.[10].)

Table 1 Parameters for the dielectric function $\varepsilon(i\xi)$ [Eq. (5)] for water at 20°C

Region	d_m	τ_m (s/rad)	f_j (rad^2/s^2)	ω_j (rad/s)	g_j (rad/s)
Microwave	75.5	0.94×10^{11}			
Infrared			9.57×10^{11}	3.19×10^{13}	2.28×10^{13}
			5.31×10^{12}	1.05×10^{14}	5.77×10^{13}
			1.97×10^{12}	1.40×10^{14}	4.25×10^{13}
			8.66×10^{12}	3.04×10^{14}	3.80×10^{13}
			2.13×10^{13}	6.38×10^{14}	8.50×10^{13}
Ultraviolet			4.95×10^{15}	1.25×10^{16}	9.57×10^{14}
			5.88×10^{15}	1.52×10^{16}	1.28×10^{15}
			1.82×10^{16}	1.73×10^{16}	3.11×10^{15}
			9.66×10^{16}	2.07×10^{16}	5.92×10^{15}
			1.73×10^{17}	2.70×10^{16}	1.11×10^{16}
			3.69×10^{16}	3.83×10^{16}	8.11×10^{15}

(From Refs.[10,13].)

$E^*(h)$, the Derjaguin approximation gives

$$F(h) = 2\pi \frac{R_1 R_2}{R_1 + R_2} E^*(h) \tag{7}$$

The interaction energy between the particles is then obtained by integrating Eq. (7) from infinity to the separation distance h. The Derjaguin approximation can be applied to van der Waals, double-layer, and many other interactions. For van der Waals interaction, either Eq. (1) or (3) for $E^*(h)$ can be substituted into Eq. (7) to determine the force between two spheres.

As long as the range of the interaction and the separation h is much less than the radius of curvature of the system, it is a valid approximation for interaction between surfaces quadratically curved in the vicinity of the point of closest approach. The condition makes curvature effects, higher than second order (via truncation of a Taylor series expansion), on the approximated energy significantly small.[16] However, the Derjaguin approximation should not be adequate for submicron-sized particles. Shown in Fig. 1 is the deviation of the Derjaguin approximation from Langbein's full macroscopic prediction described by Eq. (4) for van der Waals interaction between 250 nm radii particles. There are two important improvements of the Derjaguin approximation for submicron-sized particles.

Table 2 Parameters $\varepsilon(0)$, B, and ω in Eq. (6)

Materials	B	$\varepsilon(0)$	$\omega(10^{16}$ rad/s)
Water	1.887	80	2.068
Air	1	1	—
Crystalline quartz	2.359	4.29	2.032
Fused quartz	2.098	3.80	2.024
Fused silica	2.098	3.81	2.033
Calcite	2.516	8.20	1.897
Calcium fluoride	2.036	7.36	2.368
Sapphire	3.071	11.6	2.017
Polystyrene	2.447	2.56	1.393
Tetradecane	2.041	2.04	1.661

(From Refs.[10,12,14].)

The first improvement is based on the discovery that Eq. (2) of the Hamaker microscopic theory for spheres agrees with Eq. (4) of the continuum macroscopic theory when the Hamaker constant, A, in Eq. (2) is determined from Eq. (3) of the Lifshitz theory for parallel flat plates (Fig. 1). The combined Hamaker–Lifshitz function, $A(h)$, can be obtained by comparing the right-hand sides of Eqs. (1) and (3), giving

$$A_{132}(h) = -\frac{3k_BT}{2} \sum_{N=0}^{\infty}{}' \int_{x_N}^{\infty} x \ln[(1 - y_{13}y_{23}e^{-x}) \times (1 - z_{13}z_{23}e^{-x})]\mathrm{d}x \tag{8}$$

Eq. (8) can be simplified to provide an approximate prediction suitable for engineering calculations. Using Eq. (6), the simplification of Eq. (8) gives

$$A_{132}(h) = A^0_{132} + A^{\xi}_{132}(h) \tag{9}$$

The first and second terms on the right-hand side of Eq. (9) represent the zero-frequency (separation-independent) and the nonzero-frequency (dispersion) parts described by:[17]

$$A^0_{132} = \frac{3k_BT}{4} \sum_{m=1}^{\infty} \left\{ \frac{\varepsilon_1(0) - \varepsilon_3(0)}{\varepsilon_1(0) + \varepsilon_3(0)} \frac{\varepsilon_2(0) - \varepsilon_3(0)}{\varepsilon_2(0) + \varepsilon_3(0)} \right\}^m \Big/ m^3 \tag{10}$$

$$A^{\xi}_{132}(h) = \frac{3\hbar\omega}{8\sqrt{2}} \frac{(B_1 - B_3)(B_2 - B_3)}{(B_1 - B_2)} \times \left\{ \frac{I_2(h)}{\sqrt{B_2 + B_3}} - \frac{I_1(h)}{\sqrt{B_1 + B_3}} \right\} \quad (11)$$

Functions $I_1(h)$ and $I_2(h)$ in Eq. (11) describe the retardation effects and are defined by $I_j(h) = [1 + (h/\lambda_j)^q]^{-1/q}$, where $q = 1.185$. The characteristic wavelengths, λ_j, are measured in units of length by:$\lambda_j = \frac{c}{\pi^2\omega}\sqrt{\frac{2}{B_3(B_j + B_3)}}$, where c is the speed of light in vacuum. Unlike the London wavelength in the classical theories on retardation,[2] which is ~100 nm, λ_j is on the order of 5 nm, depending on the refractive index of the interacting materials. For two identical particles, Eq. (11) reduces to $A^{\xi}_{131}(h) = \frac{3\hbar\omega}{16\sqrt{2}} \frac{(B_1 - B_3)^2}{(B_1 + B_3)^{3/2}} \{1 + (h/\lambda)^q\}^{-1/q}$.

The van der Waals interaction energy in salt solutions is reduced.[4] In this case, Eq. (9) for the Hamaker–Lifshitz function is recast to give

$$A_{132}(\kappa, h) = A^0_{132}(1 + 2\kappa h)e^{-2\kappa h} + A^{\xi}_{132}(h) \quad (12)$$

where κ is the Debye constant. The zero-frequency term of the Hamaker–Lifshitz function is strongly influenced by salts whereas the dispersion (nonzero frequency) term remains unaffected, as the electrolyte ions do not respond to frequencies higher than 10^{14} rad/s. Therefore, the van der Waals forces across an electrolyte solution will effectively be determined solely by the London dispersion interaction.

ELECTROSTATIC DOUBLE-LAYER INTERACTION

Since particles in a polar solvent, like water, become electrically charged, electrostatic double-layer interaction (EDL) is important in determining particle–particle interactions. The Poisson–Boltzmann (PB) equation is used to describe the double-layer interaction. For $z : z$ valence salt solutions, the PB equation yields

$$\nabla^2\left(\frac{ez\psi}{k_BT}\right) = \kappa^2 \sinh\left(\frac{ez\psi}{k_BT}\right) \quad (13)$$

where ψ is the electrostatic potential in the solution and e is the charge of an electron. The Debye constant, κ, is defined by: $\kappa = \left\{\frac{2000cN_Ae^2z^2}{\varepsilon\varepsilon_0k_BT}\right\}^{1/2}$, where c is the salt concentration in mole/L, N_A the Avogadro number, ε_0 the permittivity of the vacuum, and ε is the dielectric constant (the relative permittivity) of the salt solution. The EDL force acting on a particle is determined by integrating the stress tensor, **T**, over the particle surface:[18]

$$\mathbf{F}_{edl} = \oint \mathbf{T} \cdot \mathrm{d}\boldsymbol{s} \quad (14)$$

where d$\boldsymbol{s}$ is the normal vector of the surface element. **T** is determined by the tensor of the hydrostatic (osmotic) pressure and the electric stress tensor by:

$$\mathbf{T} = \left[\varepsilon\varepsilon_0\kappa^2\left(\frac{k_BT}{ez}\right)^2 \sinh^2\left(\frac{ez_i\psi}{2k_BT}\right) + \frac{\varepsilon\varepsilon_0}{2}E^2\right]\mathbf{I} - \varepsilon\varepsilon_0\mathbf{EE} \quad (15)$$

where $\mathbf{E} = -\nabla\psi$ is the electric field, E the magnitude of **E**, **I** the unit tensor, and **EE** describes the dyadic product of the two vectors **E** and **E**. The integration of Eq. (14) gives the net force acting along the particle center-to-center line as by symmetry all components perpendicular to the centerline integrate to zero.

In Eq. (15), the electrostatic potential, ψ, is for the overlapping electric double layer of the interacting particles. Numerous models have been created to predict the overlapping field electrostatic potential between parallel plates. However, calculation of the EDL interaction for the common geometry of two spheres has not been satisfactorily resolved, due mainly to the nonlinear partial differential terms in Eq. (13) arising because of the three-dimensional geometry of the system. As a consequence, a number of approximate and numerical models have been developed for the calculation of the EDL interaction between two spheres. These models are briefly described below.

Approximate Models for EDL Interaction Between Two Spheres

The first approximate models are obtained by solving the linearized Eq. (13) for planar parallel surfaces and determining the EDL force and energy between spherical particles using the Derjaguin approximation. For low potential, the right-hand side of Eq. (13) can be linearized by $\sinh z = z + O(z^3)$. Eq. (13) reduces to a one-dimensional differential equation, which can be solved analytically using appropriate boundary conditions at the surfaces. The solution for ψ is then substituted into Eq. (15) to determine the pressure, which can be integrated to obtain the force and energy between spheres, using the Derjaguin approximation described by Eq. (7). The obtained models are summarized in Table 3.

Table 3 Results of Debye–Hückel linearization of Eq. (13) and Derjaguin approximation

Boundary conditions	Force and energy between spheres with radii R_1 and R_2
Constant surface potentials: $\psi(0) = \psi_1$ and $\psi(h) = \psi_2$ Hogg–Healy–Fuerstenau (HHF) model[19]	$F^{\psi} = \varepsilon\varepsilon_0\kappa\frac{2\pi R_1R_2}{R_1+R_2}\frac{2\psi_1\psi_2\exp(\kappa h) - \psi_1^2 - \psi_2^2}{\exp(2\kappa h) - 1}$ $E^{\psi} = \varepsilon\varepsilon_0\frac{\pi R_1R_2}{R_1 + R_2}[4\psi_1\psi_2\,\mathrm{a\,tanh}(e^{-\kappa h}) + (\psi_1^2 + \psi_2^2)\ln(1 - e^{-2\kappa h})]$ ψ_1 and ψ_2 are the particle surface potentials
Constant surface charges: $\left(\frac{d\psi}{dx}\right)_{x=0} = -\kappa\psi_{1\infty}$ $\left(\frac{d\psi}{dx}\right)_{x=h} = \kappa\psi_{2\infty}$	$F^{\sigma} = \varepsilon\varepsilon_0\kappa\frac{2\pi R_1R_2}{R_1 + R_2}\frac{2\psi_{1\infty}\psi_{2\infty}\exp(\kappa h) + \psi_{1\infty}^2 + \psi_{2\infty}^2}{\exp(2\kappa h) - 1}$ $E^{\sigma} = \varepsilon\varepsilon_0\frac{\pi R_1R_2}{R_1 + R_2}[4\psi_{1\infty}\psi_{2\infty}\mathrm{a\,tanh}(e^{-\kappa h}) - (\psi_{1\infty}^2 + \psi_{2\infty}^2)\ln(1 - e^{-2\kappa h})]$ $\psi_{1\infty}$ and $\psi_{2\infty}$ are the surface potential at infinite separation distance (of isolated particles)
Constant surface potential and charge: $\psi(0) = \psi_1$ $\left(\frac{d\psi}{dx}\right)_{x=h} = \kappa\psi_{2\infty}$	$F^{\sigma\psi} = \varepsilon\varepsilon_0\kappa\frac{2\pi R_1R_2}{R_1 + R_2}\frac{2\psi_1\psi_{2\infty}\exp(\kappa h) + \psi_1^2 - \psi_{2\infty}^2}{\exp(2\kappa h) + 1}$ $E^{\sigma\psi} = \varepsilon\varepsilon_0\frac{\pi R_1R_2}{R_1 + R_2}[4\psi_1\psi_{2\infty}\,\mathrm{a\,tan}\,(e^{-\kappa h}) + (\psi_1^2 - \psi_{2\infty}^2)\ln(1 + e^{-2\kappa h})]$

The second group of EDL models are obtained by solving the nonlinear PB Eq. (13) for planar parallel surfaces using elliptic functions and integrals,[20–22] and determining the EDL force and energy between spherical particles using the Derjaguin approximation.[15,20,22] While the models given in Table 3 are only applied to low surface potentials, the models with elliptic functions and integrals in this second group are not limited by the condition of low surface potentials. The interaction force and energy obtained for spherical particles are only restricted by the applicability of the Derjaguin approximation, i.e., the particle size must be significantly larger than the separation distance between the surfaces. This condition is not satisfied by submicron- and nano-sized particles because the range of double-layer interaction in water often exceeds a few 100 nm.

The third group of the approximate models includes various improvements of the Derjaguin approximation, linearization, and approximate solutions of PB Eq. (13) for spherical particles. The first improvement on the Derjaguin approximation for the interaction energy between identical spheres was probably obtained by the Debye–Hückel linearization and the superposition approximation,[23] given by:

$$E(h) = 4\pi\varepsilon\varepsilon_0 R\psi_s^2\frac{R + h}{2R + h} \times \ln\left[1 + \frac{R}{R + h}\exp(-\kappa h)\right] \quad (16)$$

For two spheres of different radii, R_1 and R_2, and at a constant surface charge interaction, the improved Derjaguin approximation in the framework of the Debye–Hückel linearization can be expressed as:[24]

$$E^{\psi}(h) = \varepsilon\varepsilon_0\frac{\pi R_1R_2}{R_1 + R_2 + h}[4\psi_1\psi_2\mathrm{a\,tanh}(e^{-\kappa h}) + (\psi_1^2 + \psi_2^2)\ln(1 - e^{-2\kappa h})] \quad (17)$$

Eq. (17) includes the separation distance in the sum of the particle radii. In the limit of large particles, namely, $R_1 + R_2 \gg h$, Eq. (17) reduces to the Hogg–Healy–Fuerstenau expression for $E^{\psi}(h)$, given in Table 3.

For the interaction between identical spheres at moderate surface potentials ($y_\infty \leq 4$), the improved Derjaguin approximation can be described by:

$$E^{\psi}(h) = \varepsilon\varepsilon_0\frac{64\pi R^2}{2R + h}\left(\frac{k_BT}{ez}\right)^2 \times \mathrm{a\,tanh}^2\left[e^{-\kappa h/2}\tanh\frac{y_\infty}{4}\right]e^{\kappa h}\ln(1 + e^{-\kappa h}) \quad (18)$$

Eq. (18) is claimed to be valid for all κh.[24] Note that in the limit of small surface potentials, Eq. (18) does not reduce exactly to Eq. (16), which was derived based on the superposition approximation typically used in the limit of large κh. However, if the particle radius is significantly larger than the range of the EDL interaction, Eqs. (16) and (18) in the limit of small potentials reduce to the expression:

$$E(h) = 2\pi\varepsilon\varepsilon_0 R\psi_s^2\ln[1 + \exp(-\kappa h)], \quad (19)$$

which was first derived for two identical spheres at the constant surface potential interaction using the Derjaguin approximation and the Debye–Hückel linearization by Derjaguin in 1939.[20]

There are two other developments in predicting electrostatic double-layer interactions between spheres. Firstly, an approximate expression for the correction to the Derjaguin approximation has been using a series expansion and the Debye–Hückel linearization.[25] Comparison with the exact numerical results has found that the derived expressions for the EDL interaction at constant surface potential are in good agreement with the exact numerical results. However, for constant surface charge a poorer approximation is obtained at small separation distances. This is most probably due to the inaccuracy of the Debye–Hückel linearization under the condition of the constant surface charge interaction. The study highlights the ongoing difficulty in improving the Derjaguin approximation for the EDL interaction at constant surface charge for particles in the colloidal size range. Secondly, the Derjaguin approximation is actually the zero-order approximation in terms of curvatures of the particle surfaces. Improvements to this model include the effect of curvature. For instance, in the case of interaction at constant surface potential, the additional corrections to the known HHF expression are the second and higher order terms of $1/(\kappa R)$.[26] However, the corrections often contain many terms and series expansions, approaching the complication of direct numerical computations.

Exact Numerical Solutions for EDL Interaction Between Two Spheres

The direct numerical solutions of PB equation for spheres have been reported by a number of researchers, including.[27] In the numerical computation, PB Eq. (13) is conveniently expressed and solved in the bispherical coordinates. Due to the rotational symmetry of the interaction along the centerline, Eq. (13) simplifies into

$$\frac{\partial}{\partial\xi}\left(\frac{\sin\eta}{\cosh\xi - \cos\eta}\frac{\partial\phi}{\partial\xi}\right) + \frac{\partial}{\partial\eta}\left(\frac{\sin\eta}{\cosh\xi - \cos\eta}\frac{\partial\phi}{\partial\eta}\right) = \frac{c^2\sin\eta}{(\cosh\xi - \cos\eta)^3}\sinh\phi \tag{20}$$

where the reduced potential, ϕ, is defined by $\phi = ez\psi/k_BT$. The bipolar coordinates (η, ξ) are directly related with the reduced cylindrical coordinates (r, z), where the z-axis is the particle inter-center line. One obtains: $r = \frac{c\sin\eta}{\cosh\xi - \cos\eta}$ and $z = \frac{c\sinh\xi}{\cosh\xi - \cos\eta}$. The origin of the cylindrical coordinates on the centerline is determined by the inter-center distance, $d = h + R_1 + R_2$, between the spheres and the distances, d_1 and d_2, of the center of each sphere to the origin. We have $d_1 = \frac{d^2 + R_1^2 - R_2^2}{2d}$ and $d_2 = -\frac{d^2 + R_2^2 - R_1^2}{2d}$, where R_1 and R_2 are the sphere radii, and $d_1 > 0$ and $d_2 < 0$. In the bipolar coordinates, the sphere surfaces are described by the constancy of the ξ-coordinate, i.e., by $\xi = \alpha_1$ for sphere 1 and $\xi = \alpha_2$ for sphere 2, where $\alpha_1 = \operatorname{acosh}\frac{d^2 + R_1^2 - R_2^2}{2dR_1}$ and $\alpha_2 = -\operatorname{acosh}\frac{d^2 + R_2^2 - R_1^2}{2dR_2}$. The length constant, c, in Eq. (20) is defined by $c = R_1\sinh\alpha_1 = -R_2\sinh\alpha_2$.

Eq. (20) can be discretized in the computational domain bounded by $\xi = \langle\alpha_2,\alpha_1\rangle$ and $\eta = \langle 0,\pi\rangle$. The discretization can be carried out using either the finite volume or finite difference scheme, producing a system of nonlinear equations for the reduced potential in the computational domain, which can be solved employing the Newton–Ralphson method or other relaxation techniques. The boundary conditions of constant surface potentials are described by $\phi = \phi_1$ at the surface of sphere 1 with $\xi = \alpha_1$ and $\eta = \langle 0, \pi\rangle$, and $\phi = \phi_2$ at the surface of sphere 2 with $\xi = \alpha_2$ and $\eta = \langle 0, \pi\rangle$. The boundary conditions of constant surface charge are described by $\left(\frac{\cosh\xi - \cos\eta}{c}\frac{\partial\phi}{\partial\xi}\right)_{\xi=\alpha_1,\alpha_2} = \pm\frac{\sigma_{1,2}ze}{\varepsilon\varepsilon_0\kappa k_BT}$ at the surface of spheres 1 and 2 with the surface charges, σ_1 and σ_2, respectively.

The general behavior of the numerical results of Eq. (20) for the potential distribution between the particle surfaces is shown in Fig. 3 for the constant surface potential interaction between two spherical particles with radii of 20 nm and surface potentials of 50 mV. The effect of the interaction is to distort the potential profiles around the particles. At a separation distance of about 10-particle radii (~200 nm), the potential profiles are almost identical to those around the isolated particles. However, the distortion of the potential profiles becomes stronger with decreasing the separation distance between the surfaces. The strong distortion of the potential profiles increases the repulsive interaction between the particles as shown below.

Once the potential field in the computational domain is available, the EDL force can be determined by integrating the stress over the sphere surface using Eqs. (14) and (15). In the bipolar coordinates, the normal vector of the surface element in Eq. (14) yields $\mathrm{d}s = -\mathbf{i}_\xi 2\pi\left(\frac{c}{\cosh\xi - \cos\eta}\right)^2\sin\eta\,\mathrm{d}\eta$, where the unit vector, $\mathbf{i}_\xi$, is pointed towards the surface of one of the spheres. The potential gradient is described by

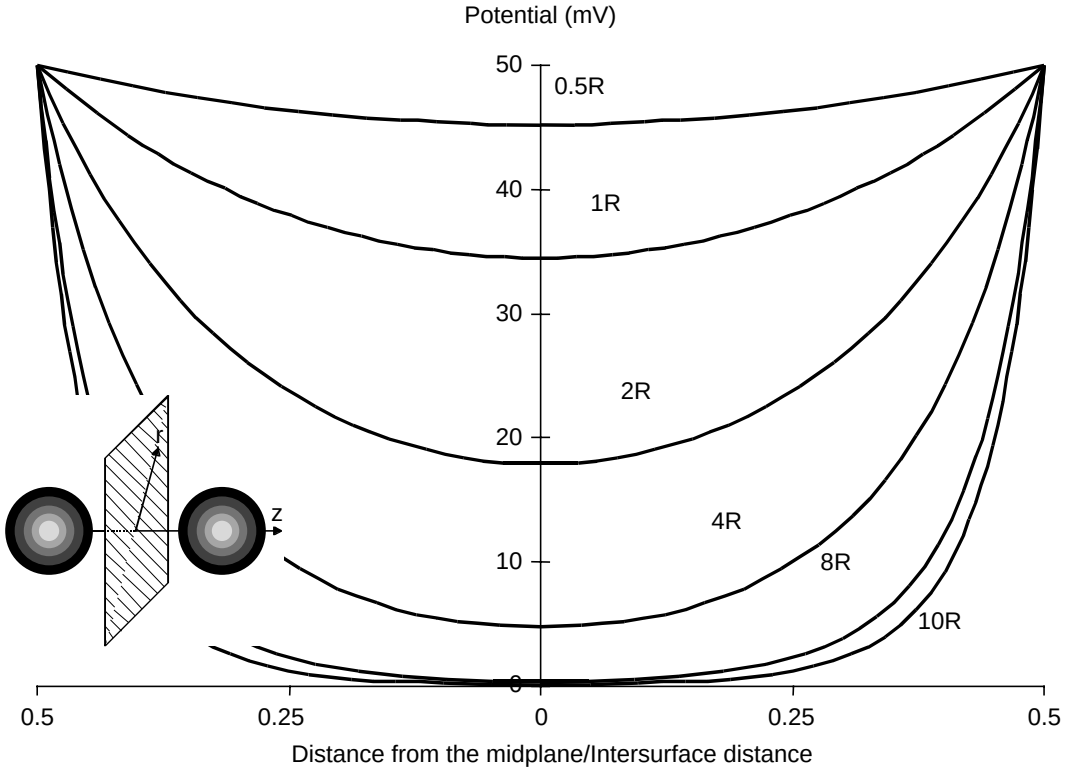


Fig. 3 Numerical results of Eq. (13) for the potential distribution on the inter-center line of two identical spherical particles with radii $R = 20$ nm and surface potentials of 50 mV, showing the distribution at the different inter-surface distances: 0.5, 1, 2, 4, 8, and 10R. The distance from the ($z = 0$) midplane is normalized by dividing by the inter-surface distance.

$\mathbf{E} = -\frac{\cosh\xi - \cos\eta}{c}\left(\frac{\partial\psi}{\partial\eta}\mathbf{i}_\eta + \frac{\partial\psi}{\partial\xi}\mathbf{i}_\xi\right)$, where $\mathbf{i}_\eta$ is another unit vector of the η-coordinate. Combining Eqs. (14) and (15) gives

$$\mathbf{F} = 2\pi\int_0^\pi \left[\left(\left\{\frac{c}{\cosh\xi - \cos\eta}\right\}^2 \Pi + \frac{1}{2}\varepsilon\varepsilon_0\left\{\left(\frac{\partial\psi}{\partial\xi}\right)^2 - \left(\frac{\partial\psi}{\partial\eta}\right)^2\right\}\right)\mathbf{i}_\xi + \varepsilon\varepsilon_0\frac{\partial\psi}{\partial\eta}\frac{\partial\psi}{\partial\xi}\mathbf{i}_\eta\right]\sin\eta\,\mathrm{d}\eta \tag{21}$$

where Π describes the first term (the osmotic pressure) in the brackets in Eq. (15). The net force acting on the inter-center line is obtained by multiplying Eq. (21) with the unit vector, $\mathbf{i}_z$, of the z-axis. Since $\mathbf{i}_\eta \cdot \mathbf{i}_z = \frac{\sin\eta\sinh\xi}{\cosh\xi - \cos\eta}$ and $\mathbf{i}_\xi \cdot \mathbf{i}_z = \frac{\cos\eta\cosh\xi - 1}{\cosh\xi - \cos\eta}$, Eq. (21) gives

$$F = 2\pi\int_0^\pi \left[\left(\left\{\frac{c}{\cosh\xi - \cos\eta}\right\}^2 \Pi + \frac{\varepsilon\varepsilon_0}{2}\left\{\left(\frac{\partial\psi}{\partial\eta}\right)^2 - \left(\frac{\partial\psi}{\partial\xi}\right)^2\right\}\right)\times\frac{1 - \cos\eta\cosh\xi}{\cosh\xi - \cos\eta} + \varepsilon\varepsilon_0\frac{\partial\psi}{\partial\eta}\frac{\partial\psi}{\partial\xi}\frac{\sin\eta\sinh\xi}{\cosh\xi - \cos\eta}\right]\sin\eta\,\mathrm{d}\eta \tag{22}$$

The integral in Eq. (22) is independent of the specific choice of the constant ξ-plane and, therefore, it can be numerically integrated along any surface that encloses one of the spheres and has a constant value for the ξ-coordinate. The surface $\xi = 0$ is particularly useful for the numerical integration because $\sinh\xi = 0$ at $\xi = 0$. For symmetrical interactions between two identical spheres, the surface of $\xi = 0$ is the plane of the symmetry, where the ξ-derivative of the potential is zero, and the numerical integration is significantly simplified.

Finally, knowing the force as a function of separation distance, the interaction energy, $E(h)$, can be determined by integrating the force with respect to the separation distance from infinity. The numerical integration can start at some cut-off distance as the available analytical solutions described by Eqs. (16)–(19) can be used when the particles are far apart and the interaction is weak.[20] In practice, the numerical

data for the interaction force versus separation distance are often limited and the integration for the interaction energy can be effectively carried out using the Gauss-Laguerre quadrature scheme. The same principle can be used to integrate Eq. (22) with moderate computing costs.

Exact numerical results are used to validate the available approximate models described by Eqs. (16)–(19). The comparison is shown in Fig. 4 for particles with scaled radii $R\kappa = 0.1$ and $R\kappa = 15$. The interaction energy was determined for two identical spheres in a $z:z$ electrolyte solution. The approximate solutions are given by Eqs. (16)–(19) and the equation for the HHF model given in Table 3. For the exact numerical solution, the full Poisson–Boltzmann equation was discretized and solved by the finite volume method. The results have been plotted for two particle sizes $\kappa R = 0.1$ (Fig. 4A) and $\kappa R = 15$ (Fig. 4B).

It can be seen that for both scaled particle sizes the predictions for $G(h)$ from Eqs. (16)–(18) match closely the exact numerical solution. For the result of the Derjaguin Eq. (19) and the HFF model given in Table 3, however, good agreement is obtained only for the $\kappa R = 15$ case. Further analysis (not shown here) indicates that the Derjaguin approximation should only be used for $\kappa R \geq 10$, which corresponds to spherical particles larger than a few microns in diameter. Although it can be seen from Fig. 4 that both Eqs. (16) and (17) provide very good approximations to the exact numerical solution, Eq. (17) is more widely applicable to double-layer interaction at constant and low surface potentials. It is also simpler in terms of the expressions involving separation. For these reasons, it is usually recommended to use the generalized HHF expression given by Eq. (17) for calculating the double-layer interaction energy between spheres.

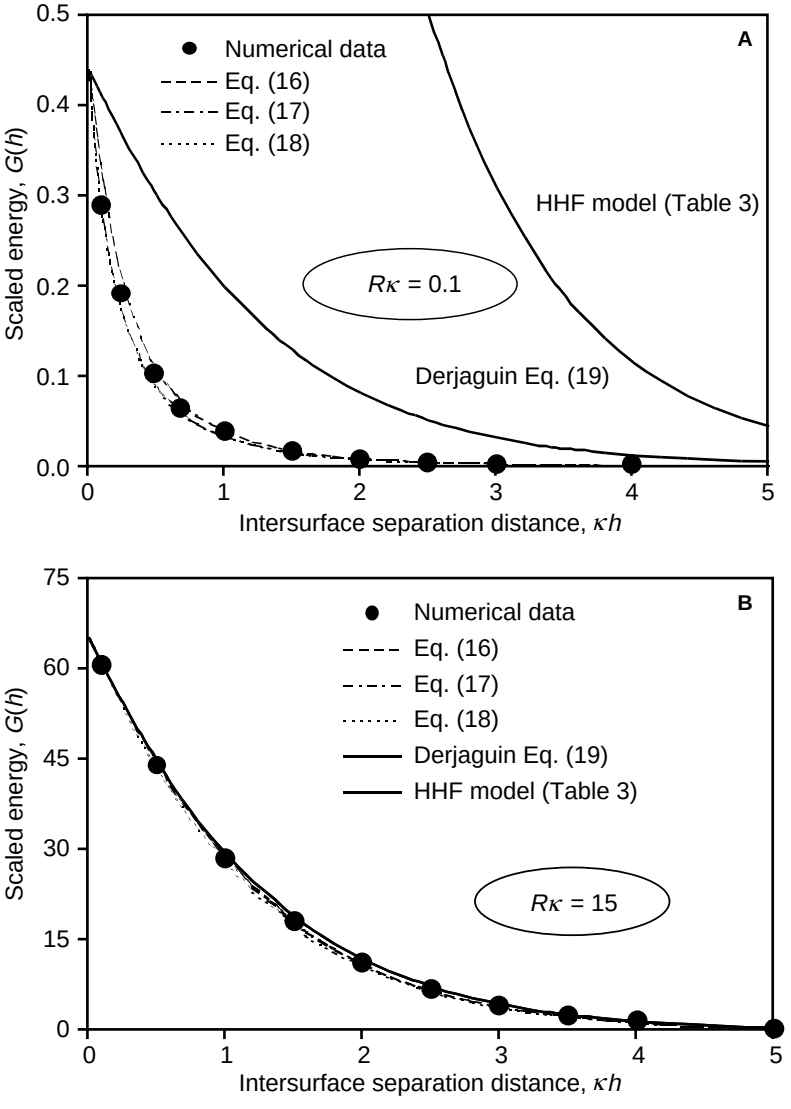


Fig. 4 Comparison of the model improvements on the Derjaguin approximation to the exact numerical computational results of the full Poisson–Boltzmann equation for two spheres with the scaled radius $R\kappa = 0.1$ and $R\kappa = 15$ and constant surface potential $\psi_s ez/(k_B T) = 1$. The scaled energy, $G(h)$, on the vertical axis is defined by $G(h) = E(h)/\left\{\frac{\varepsilon\varepsilon_0}{\kappa}\left(\frac{k_B T}{ez}\right)^2\right\}$.

LIMITS OF THE DERJAGUIN APPROXIMATION PROBED WITH AFM TIPS

The limits of the celebrated Derjaguin approximation for predicting forces between submicron-sized particles have been argued for some time. Now the approximation can be validated using the force data obtained for the interaction between the AFM tips on microfabricated cantilevers and the flat surfaces. The radius of curvature of the AFM tips is about 10 nm and provides the ideal geometry with small interaction forces. Fig. 5 shows an example for the forces measured with the graphite (HOPG) flat surfaces and the silicon nitride tips with the radius of curvature of about 7 nm in solution with different pH.

In Fig. 5, the interaction force between the tip and the HOPG surface, as predicted by the Derjaguin approximation, F_{DA}, is described by:[28]

$$F_{DA}(h) = 2\pi R E^*(h) \tag{23}$$

where R is the characteristic radius of the tip ($\sim$7 nm) and E^* is the interaction energy between flat surfaces per unit area. An EDL model and a hard-core repulsion model are used to describe the energy E^* by:[28]

$$E^*(h) = \frac{2\pi}{\varepsilon\varepsilon_0\kappa}(\sigma_1\sigma_2)\exp[-\kappa(h - h_0)] + \frac{const}{72\pi(h - h_0)^8} \tag{24}$$

where $\sigma_1\sigma_2$ is the product of the surface and tip charge densities, *const* a constant, and h_0 is an off-set separation distance due to the uncertainty of absolute zero separation distance in the force measurements with AFM.

The surface element integration (SEI) method provides an improvement on the interaction force between a spherical particle and a flat surface.[29] The SEI improves the Derjaguin approximation by replacing infinity in the integration of the Derjaguin approximation by a finite upper limit, leading to the following prediction for the interaction force, F_{SEI}, between the tip and the surface:

$$F_{SEI}(h) = 2\pi\int_0^R \frac{\partial}{\partial h}\{E^*(D_+) - E^*(D_-)\} r\,dr \tag{25}$$

where the energy E^* is given by Eq. (24) and $D_\pm = h + R \pm \sqrt{R^2 - r^2}$. The integration can be calculated numerically.

The solid lines in Fig. 5 show Eqs. (23) and (25) with the model parameters obtained by regression analysis. Despite the best-fit procedure employed, the Derjaguin approximation shows significant deviation from the experimental data at short separation distances. The Derjaguin approximation cannot model the interaction under the given conditions when the radius of the surface curvature is similar to the range of the double-layer forces ($1/\kappa = 9.6$ nm). In this case, the upper limit of the integration of the Derjaguin approximation cannot approach infinity and the SEI approximation, which considers the finite limit of the integration, agrees with the experimental data.

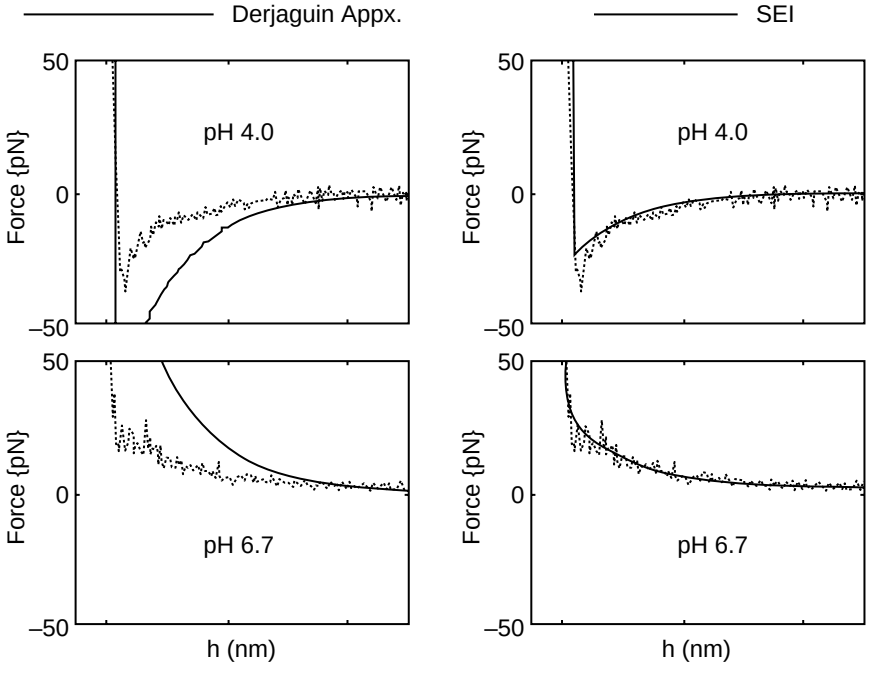


Fig. 5 Experimental data (dotted lines), and the Derjaguin approximation (solid lines on the left diagrams) and its SEI improved prediction (solid lines on the right diagrams) for the interaction force vs. separation distance between a silicon nitride AFM tip and an HOPG surface in 1 mM NaCl solutions ($1/\kappa = 9.6$ nm. (From Ref.[28].)

CONCLUSIONS

Improvements in the DLVO force predictions for the van der Waals and electrostatic double-layer interactions between small spherical particles have been critically reviewed. The van der Waals interaction energy between two spheres can be determined from the exact Langbein theory. The approximate equations for van der Waals interaction energy between two spheres can be obtained by combining the Hamaker microscopic and Lifshitz macroscopic theories. The Hamaker–Lifshitz function, $A_{132}(h)$, described by Eq. (12) depends on the separation distance (accounting for the retardation effects due to the finite speed of light) and salt concentration (accounting for the screening effects of electrolytes). The Hamaker–Lifshitz function is used in conjunction with Eq. (2) to determine the van der Waals interaction energy and force between small spherical particles.

It was shown in the analysis that the Poisson–Boltzmann Eq. (13) developed within the Gouy–Chapman theory provides a useful tool for predicting the EDL interaction between particles. Comparison with the numerical solution of the Poisson–Boltzmann equations and experimental data obtained with AFM tips shows that the Derjaguin approximation will fail to predict the double-layer interaction between nano-sized particles. The predictions using the approximate expressions described by Eqs. (16)–(18) for small particles were found to be an improvement over the Derjaguin approximation, when compared with the exact numerical data obtained by numerically solving the full Poisson–Boltzmann equation. It was also shown that the SEI technique provides significant improvement for predicting the force between a surface with nano-sized radius of curvature such as an AFM tip and a flat surface.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Australian Research Council, the National Science Foundation, and the Research Project Committee of the University of Newcastle for the financial support.

REFERENCES

1. Hamaker, H.C. The London-van der Waals attraction between spherical particles. Physica. **1937**, *4*, 1058–1072.
2. Casimir, H.B.G.; Polder, D. The influence of retardation on the London-van der Waals forces. Physical Review **1948**, *73*, 360–372; Overbeek, J.T.G. The interaction between colloidal particles. In *Colloid Science*; Kruyt, H.R., Ed.; Elsevier: Amsterdam, 1952; 266–270; Schenkel, J.H.; Kitchener, J.A. A test of the Derjaguin-Verwey-Overbeek theory with a colloidal suspension. Trans. Faraday Soc. **1960**, *56*, 161–173.
3. Lifshitz, E.M. The theory of molecular attractive forces between solid bodies. J. Exp. Theor. Phys. USSR **1955**, *29*, 83–94.
4. Mahanty, J.H.; Ninham, B.W. *Colloid Science: Dispersion Forces*; Academic Press: London, 1977.
5. Dzaloshinskii, I.E.; Lifshitz, E.M.; Pitaerskii, L.P. The general theory of van der Waals forces. Adv. Phys. **1961**, *10*, 165–209.
6. Langbein, D. *Van der Waals Attraction*; Springer Verlag: Berlin, 1974.
7. Love, J.D. On the van der Waals force between two spheres or a sphere and a wall. J. Chem. Soc. Faraday Trans. 2 **1977**, *73* (5), 669–688.
8. Kiefer, J.E.; Parsegian, V.A.; Weiss, G.H. Some convenient bounds and approximations for the many body Van der Waals attraction between two spheres. J. Colloid Interface Sci. **1978**, *67* (1), 140–153.
9. Pailthorpe, B.A.; Russel, W.B. The retarded van der Waals interaction between spheres. J. Colloid Interface Sci. **1982**, *89* (2), 563–566.
10. Nguyen, A.V. Improved approximation of water dielectric permittivity for calculation of Hamaker constants. J. Colloid Interface Sci. **2000**, *229* (2), 648–651.
11. Nguyen, A.V.; Schulze, H.J. *Colloidal Science of Flotation*; Marcel Dekker: New York, 2004.
12. Hough, D.B.; White, L.R. The calculation of Hamaker constants from Lifshitz theory with applications to wetting phenomena. Adv. Colloid Interface Sci. **1980**, *14* (1), 3–41.
13. Parsegian, V.A.; Weiss, G.H. Spectroscopic parameters for computation of van der Waals forces. J. Colloid Interface Sci. **1981**, *81*, 285–289; Roth, C.M.; Lenhoff, A.M. Improved parametric representation of water dielectric data for Lifshitz theory calculations. J. Colloid Interface Sci. **1996**, *179* (27), 637–639.
14. Israelachvili, J.N. *Intermolecular and Surface Forces*; Academic Press: London, 1992.
15. Derjaguin, B. Untersuchungen ueber die Reibung und Adhaesion IV. Theorie des Anhaften kleiner Teilchen. Kolloid-Z. **1934**, *69*, 155–166.
16. White, L.R. On the Deryaguin approximation for the interaction of macrobodies. J. Colloid Interface Sci. **1983**, *95* (1), 286–288.
17. Nguyen, A.V.; Evans, G.M.; Schulze, H.J. Prediction of van der Waals interaction in bubble-particle attachment in flotation. Int. J. Miner. Process. **2001**, *61* (3), 155–169.

18. Derjaguin, B.V. On the repulsive forces between charged colloid particles and on the theory of slow coagulation and stability of lyophobic sols. Trans. Faraday Soc. **1940**, *36*, 203–215; Bell, G.M.; Levine, S.; McCartney, L.N. Approximate methods of determining the double-layer free energy of interaction between two charged colloidal spheres. J. Colloid Interface Sci. **1970**, *33* (3), 333–360; Russel, W.B.; Saville, D.A.; Schowalter, W.R. *Colloidal Dispersions*; Cambridge University Press: Cambridge, 1989.
19. Hogg, R.; Healy, T.W.; Fuerstenau, D.W. Mutual coagulation of colloidal dispersions. Trans. Faraday Soc. **1966**, *62*, 1638–1651.
20. Derjaguin, B.V. *Theory of Stability of Colloids and Thin Films*; Plenum: New York, 1989.
21. Chan, D.Y.C. A simple algorithm for calculating electrical double layer interactions in asymmetric electrolytes. Poisson-Boltzmann theory. J. Colloid Interface Sci. **2002**, *245* (2), 307–310; Devereeux, O.F.; deBruyn, P.L. *Interaction of Plane Parallel Double Layers*; MIT Press: Cambridge, 1963.
22. Verwey, E.J.W.; Overbeek, J.T.G. *Theory of the Stability of Lyophobic Colloids*; Elsevier: Amsterdam, 1948.
23. McCartney, L.N.; Levine, S. An improvement on Derjaguin's expression at small potentials for the double layer interaction energy of two spherical colloidal particles. J. Colloid Interface Sci. **1969**, *30*, 345–362.
24. Sader, J.E.; Carnie, S.L.; Chan, D.Y.C. Accurate analytic formulas for the double-layer interaction between spheres. J. Colloid Interface Sci. **1995**, *171* (1), 46–54.
25. Glendinning, A.B.; Russel, W.B. The electrostatic repulsion between charged spheres from exact solutions to the linearized Poisson-Boltzmann equation. J. Colloid Interface Sci. **1983**, *93*, 95–111; Carnie, S.L.; Chan, D.Y.C. Interaction free energy between identical spherical colloidal particles: the linearized Poisson-Boltzmann theory. J. Colloid Interface Sci. **1993**, *155* (2), 297–312.
26. Chan, D.Y.C.; White, L.R. The electrostatic interaction between spherical colloidal particles–a comment on the paper by Barouch et al. J. Colloid Interface Sci. **1980**, *74* (1), 303–305; Ohshima, H.; Chan, D.Y.C.; Healy, T.W.; White, L.R. Improvement on the Hogg-Healy-Fuerstenau formulas for the interaction of dissimilar double layers. II. Curvature correction to the formula for the interaction of spheres. J. Colloid Interface Sci. **1983**, *92* (1), 232–242.
27. Chan, B.K.C.; Chan, D.Y.C. Electrical double-layer interaction between spherical colloidal particles: an exact solution. J. Colloid Interface Sci. **1983**, *92* (1), 281–283; Palkar, S.A.; Lenhoff, A.M. Energetic and entropic contributions to the interaction of unequal spherical double layers. J. Colloid Interface Sci. **1994**, *165* (1), 177–194; Qian, Y.; Bowen, W.R. Accuracy assessment of numerical solutions of the nonlinear Poisson-Boltzmann equation for charged colloidal particles. J. Colloid Interface Sci. **1998**, *201* (1), 7–12; Carnie, S.L.; Chan, D.Y.C.; Stankovich, J. Computation of forces between spherical colloidal particles: Nonlinear Poisson–Boltzmann theory. J. Colloid Interface Sci. **1994**, *165* (1), 116–128; Stankovich, J.; Carnie, S.L. Electrical double layer interaction between dissimilar spherical colloidal particles and between a sphere and a plate: nonlinear Poisson–Boltzmann theory. Langmuir **1996**, *12* (6), 1453–61.
28. Todd, B.A.; Eppell, S.J. Probing the limits of the Derjaguin approximation with scanning force microscopy. Langmuir **2004**, *20*, 4892–4897.
29. Bhattacharjee, S.; Elimelech, M. Surface element integration: a novel technique for evaluation of DLVO interaction between a particle and a flat plate. J. Colloid Interface Sci. **1997**, *193* (2), 273–285.

Pervaporation: Vapor Permeation

Nicholas Patrick Wynn
Sulzer Chemtech GmbH, Neunkirchen, Germany

INTRODUCTION

Pervaporation occupies a special niche in the chemical industry—it is the only membrane process primarily used to purify chemicals. At the start of 2000, about 100 pervaporation units were operating worldwide, most of them dehydrating solvents, such as ethanol and isopropanol. Now that pervaporation has been proven in these end-of-pipe applications, attention is being focused on large-scale separations that are closer to the chemical reaction step—to be more critical for production and to bring much greater benefits.

These developments result from the introduction of composite membranes, originally developed in the 1970s primarily, for desalination by reverse osmosis. Application of the same membrane fabrication techniques to pervaporation membranes radically improved their performance and spurred commercial utilization. Today, pervaporation and vapor permeation plants are widely used to dehydrate volatile organics and separate other mixtures, primarily in the pharmaceutical and fine chemical industries.

This shift in focus is accelerating the development of new, more robust membranes, with tailored performance, which can be used at higher temperatures. Over the next few years, pervaporation will be used increasingly in the CPI, HPI, and O&G industries, not just for dehydration, but also for organic–organic separations.

PERVAPORATION AND VAPOR PERMEATION PROCESSES

Pervaporation and vapor permeation (and gas permeation) are closely related processes and are characterized by generating a permeate in the vapor state. In this situation, the driving force for permeation of a particular component approximates very closely to the difference in partial vapor pressure of that component across the membrane. Because the pressure on the back-side of the membrane is low, almost all of the faster permeating component can be removed from the feed. The process purifies the feed by removing the faster permeating component. The product from the process is the retentate and the concentrated impurity is the permeate.

Contrast this with the more common pressure driven membrane processes, such as reverse osmosis, where liquid is present on both sides of the membrane. In this situation, the faster permeating component cannot be completely removed by permeation. The membrane acts like a strainer or filter, holding back the slower or non-permeating component. In this case, the product from the process is the pure permeate and the concentrated impurity is the retentate.

The best performing industrial membranes permeate water in preference to other components hence the pressure driven filtration and RO processes are used typically to purify water. In contrast, pervaporation and vapor permeation are commonly, but not exclusively, used to remove water from organics (Fig. 1).

Transport of a component through the membrane is best described by the so-called ''solution–diffusion model.'' This model is based on a component of the feed having a high affinity to the membrane being easily and preferentially adsorbed and dissolved in the dense membrane. Because of a concentration gradient, it migrates through the membrane by a diffusion process and is desorbed at the back-side of the membrane. Thus, the separation characteristic of the membrane is determined primarily by the different solubilities of components in the membrane material, and to a lesser extent by the different diffusivities (which may actually counteract the solubility differences).

Driving Force for Pervaporation

In pervaporation and vapor permeation processes the permeate leaves the membrane as a vapor, hence the driving force for permeation for a particular component is the difference in partial vapor pressure of that component across the membrane. The partial vapor pressures of the components at the feed side are fixed by the feed composition and temperature. Temperature can only be increased up to the operating limit of the membrane material, typically quite low for polymer membranes. Therefore, the driving force for the transport of matter through the membrane is applied principally by reducing the partial vapor pressure at the permeate side.

Different means have been proposed to effect this reduction of the permeate side partial vapor pressure:

1. The permeating vapor is condensed under vacuum, i.e., using a sufficiently low temperature.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008111

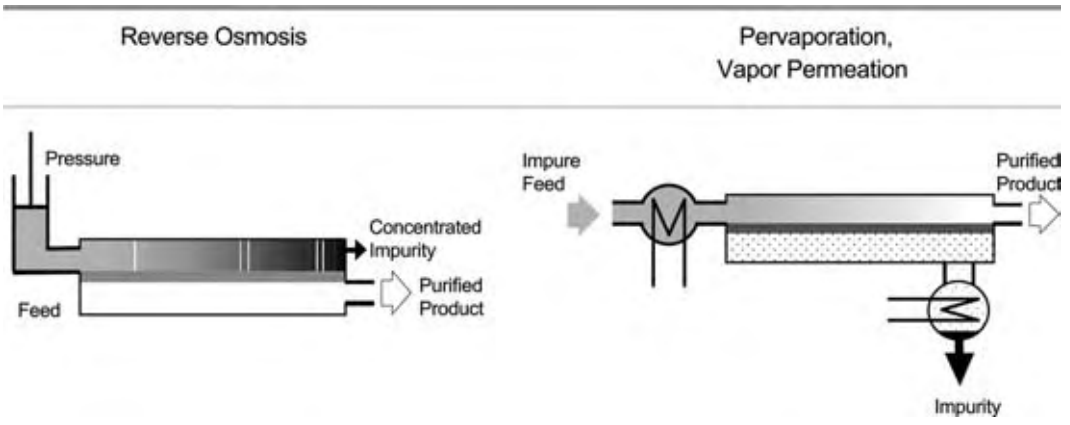


Fig. 1 Contrast pervaporation with reverse osmosis. *(View this art in color at www.dekker.com.)*

Only a small vacuum pump is then required to remove trace quantities of noncondensables.

2. All permeating vapor is removed by means of a vacuum pump. The vapor may be condensed after recompression at the pressure side of the pump.
3. The permeate side of the membrane is swept with an inert gas, to keep the partial vapor pressure of the faster permeating component sufficiently low. If the gas stream cannot be discarded, it has to be reconditioned and recycled.

In most industrial installations, the first option has been proven to be the most effective and economical process. However, it is preferable to pass all the permeate through a vacuum pump and condense them at atmospheric pressure, if the permeate quantity is low or if it contains a very volatile organic. Regarding the second option, sweep gas operation is rarely used, because sweeping right up into the pores of composite membranes is difficult.

Continuous Pervaporation Systems

Pervaporation is unusual amongst membrane processes because permeation through the membrane is coupled with a phase change from liquid to vapor. Thermodynamically, this is equivalent to a partial flash. If the membrane is operated adiabatically, this results in a cooling of the liquid on the pressure side of the membrane, which quickly reduces the driving force and permeation rate. This problem is overcome in older plants by splitting the total membrane area into a number of stages in series and installing interstage heaters. In more modern plants, steam heated tubular pervaporation modules that operate isothermally are more commonly used. Continuously operating pervaporation units run at steady state conditions (Fig. 2). Feed is passed continuously through the unit and the product leaves the unit at the desired water content. Wet feed is first heated in a recuperator, and fed to the membrane modules where it is further heated and dehydrated by pervaporation, and then it is cooled in the recuperator, depressurized and passed off the skid as product. Permeating water vapor is condensed in the permeate condenser, which is cooled by chilled water/glycol. A vacuum pump removes noncondensables from the system. Optionally, if the permeate flow is small, it can be passed through the vacuum pump and condensed at atmospheric pressure without using refrigeration.

Batch Pervaporation Systems

Pervaporation can also operate in batch mode, and this is done typically when testing membranes for small plants and for some larger multipurpose plants. Batch pervaporation systems are robust, well proven, and flexible in operation. The pumparound rate on batch systems is normally set high to give a low permeate quantity per pass. Pervaporative cooling effects are small, and such systems can be built with a single preheater and unheated modules (Fig. 3).

Hydrated solvent from a buffer tank is continuously circulated through the unit until the desired degree of dryness has been reached (Fig. 4). Such units are very flexible—different start and end water contents can be accommodated, and with multi-purpose membranes a range of solvents can be dehydrated. Note that batch pervaporation units are not suitable for feeds that may contain suspended solids or dissolved solids, which could precipitate as water is removed.

Vapor Permeation Systems

In vapor permeation, the feed is typically a saturated vapor mixture, or at least the permeating component

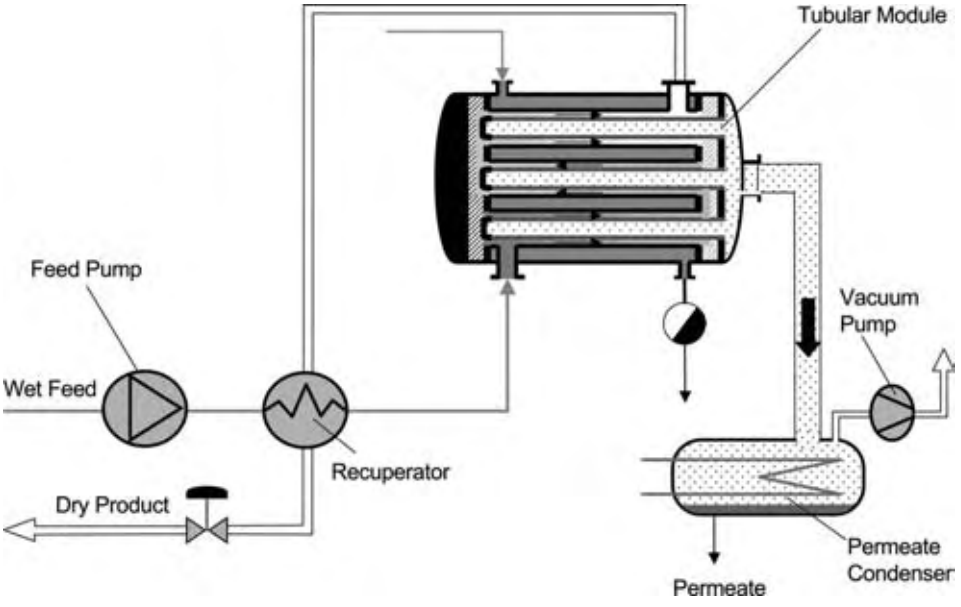


Fig. 2 Continuous pervaporation unit. *(View this art in color at www.dekker.com.)*

is close to saturation conditions. The temperature and pressure of the feed are linked by the vapor–liquid equilibrium. This process does not include a phase change, hence heating is unnecessary, and the membrane modules are simpler.

Vapor permeation is often a preferred technique to pervaporation, and liquid feeds are often evaporated, especially, to run the vapor permeation process. The evaporator is typically operated under pressure, giving a vapor feed at the maximum temperature and flux rate consistent with membrane stability. The advantages are a simpler plant and more reliable operation, because both dissolved and suspended solids are removed in the evaporator and cannot damage the membranes. The disadvantage is the extra energy required to evaporate the complete feed (Fig. 5).

Vapor Permeation Combined with Distillation

Vapor permeation is also used in combination with distillation feed of net overhead vapor from a distillation column directly to a vapor permeation plant is a very economical way of splitting azeotropes (Fig. 6). Typically, the column must be operated at pressure to provide a vapor overhead at the optimum temperature for permeation. Fig. 7 shows a vapor permeation plant for drying isopropanol directly coupled to a distillation column.

Process Selection

Selection of the optimum process is application specific, however, Table 1 gives some general guidelines (Fig. 8A–C).

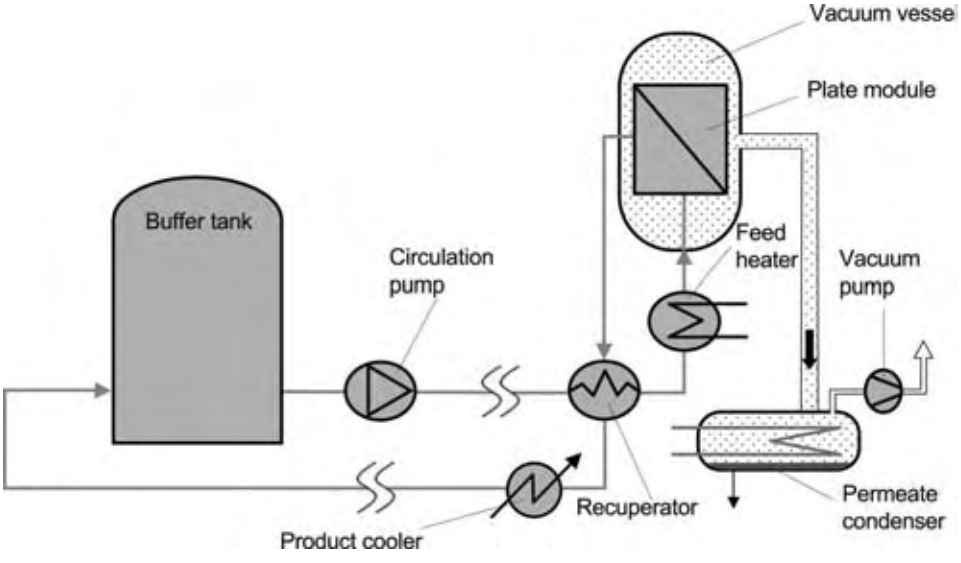


Fig. 3 Batch pervaporation unit. *(View this art in color at www.dekker.com.)*

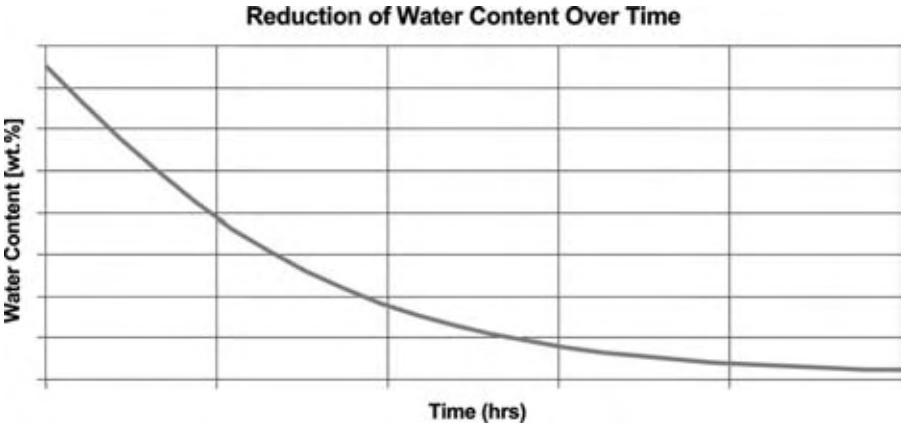


Fig. 4 Reduction of water content over time. *(View this art in color at www.dekker.com.)*

PERVAPORATION AND VAPOR PERMEATION MEMBRANES

Composite Polymer Membranes

Developments in pervaporation were boosted by the improvements in membrane structure and membrane manufacture resulting from R&D expenditures on desalination membranes in the 1970s. Two different types of pervaporation membrane were developed in the beginning of the 1980s:

- Hydrophilic membranes, used mainly for the removal of water from organic solvents and solvent mixtures, especially azeotropic mixtures.
- Organophilic membranes, preferentially dissolving and permeating organics, used for the removal of volatile organic components from water and gas streams.

Both membrane types are of composite construction, with adequate mechanical, chemical, and thermal stability despite very thin high flux separation layers. Because of the composite structure, flat sheet configurations are preferred. The substructure of both types of pervaporation membranes consists of a porous support membrane with an asymmetric pore structure, cast onto a support layer consisting of a woven or nonwoven fabric. The result is similar to an ultrafiltration membrane. On the exposed side of this porous substructure the pores have diameters in the order of 20–50 nm. Pore width on the fabric side is closer to the micrometer range. Structural polymers like polyacrylonitrile, polyetherimide, polysulfone, polyethersulfone, and polyvinylidenfluoride are used to form the cast porous support. Polyester, polypropylene, and similar fibers are used for the fabric support.

A thin dense separating layer (in the range of 0.5–5 μ thick) is then coated on this substructure. Different coating techniques are in use, but most commonly a solution of polymer in an appropriate solvent is spread onto the porous substructure. The solvent is evaporated, followed by further treatment to crosslink the polymer.

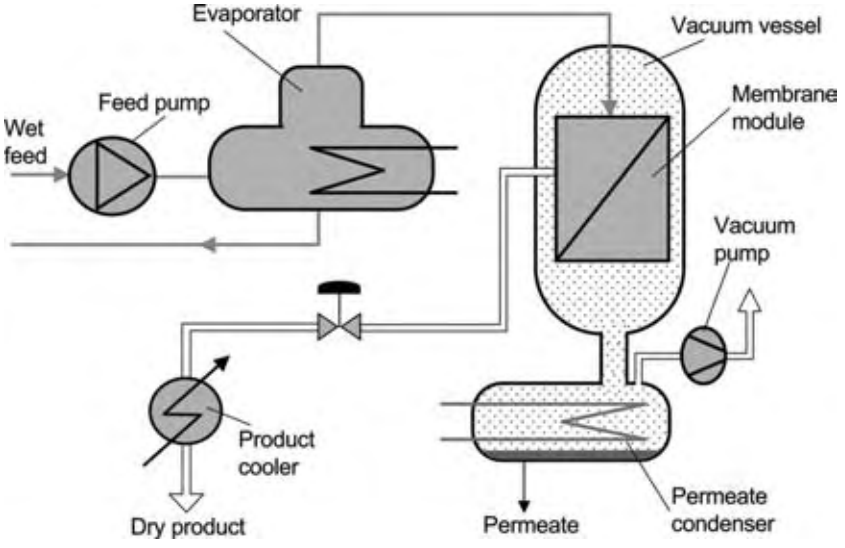


Fig. 5 Vapor permeation unit. *(View this art in color at www.dekker.com.)*

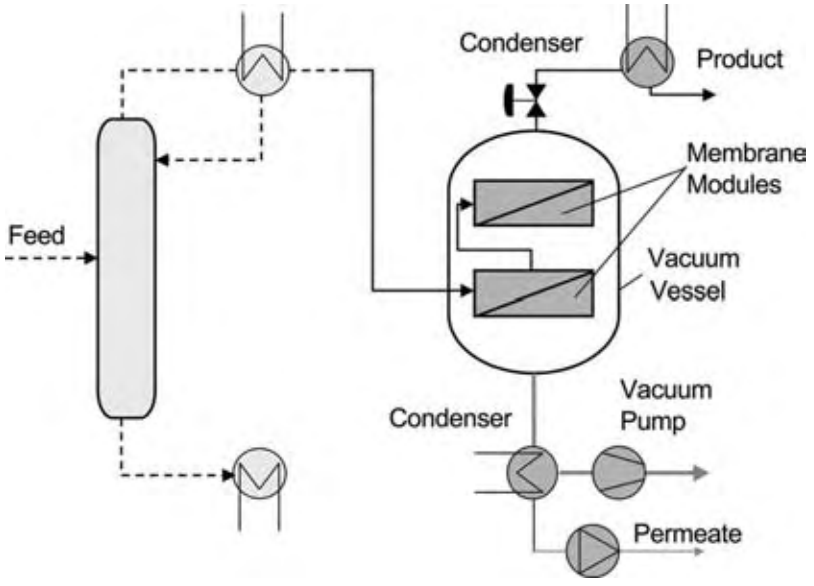


Fig. 6 Vapor permeation coupled with distillation.

In hydrophilic membranes the separating layer is most commonly made from crosslinked polyvinylalcohol (PVA), however, polyimides or natural polymers like chitosan or cellulose acetate (CA) are also used (Fig. 9).

In organophilic membranes the separation layer is typically formed from siloxanes like polydimethyl siloxane (PDMS) or polyoctylmethyl siloxane (POMS).

In recent years tailored polymeric membranes for organic–organic separations have been intensively developed. The large-scale separation of olefins from paraffins is of particular interest, e.g., propene from propane, also aromatics like benzene or toluene from aliphatic hydrocarbons or the separation of the xylene isomers. A number of smaller organic–organic separation plants are already operating, separating methanol from mixtures with hydrocarbons, ethers, and esters. The membranes used are actually of the hydrophilic type, through which the more polar methanol permeates quickly.

Ceramic Membranes

Ceramic membranes are now being used in pervaporation and promise higher operating temperatures and better chemical resistance than polymer membranes. The thermal, mechanical, and chemical stability of the cast porous substructure in polymeric membranes limits the operating range more than the stability of the actual separating layer. Ceramic substructures can be coated by crosslinked polymeric separating layers similar to those on polymeric substructures. In more recent developments inorganic separation layers are applied, either by coating the porous substructure with a layer of zeolite, or by reducing the size of the pore to molecular dimensions. The separation mechanism of these membranes is even more complex than that of polymeric separating layers, as molecular sieving effects, caused by shape and size of molecules, and molecule–surface interaction decide whether a component can pass through the membrane or will be retained. Flux rates can be considerably higher than with polymeric membranes. A number of ceramic membrane plants are already operating in dehydration applications (Fig. 10).

Modules for Pervaporation and Vapor Permeation

The basic module types originally used for ultrafiltration and reverse osmosis in water treatment have been adapted for pervaporation and vapor permeation.

In both processes the partial vapor pressure at the permeate side has to be reduced to fairly low values, especially when low final concentrations of permeating component have to be reached in the retentate. Large volumetric flows have to be accommodated on the permeate side with minimal pressure drop.

As any feed mixture will contain organic components at high concentration, mostly at elevated temperatures, chemical stability of all module components, such as spacer and potting material and glues is critical. So far two types of modules are most widely used.

Plate Modules

Plate modules are mainly used for dehydration applications, with stainless steel used as construction material for support plates for the membranes and for spacers. Gaskets are normally graphite based and

Fig. 7 Combination vapor permeation and distillation. *(View this art in color at www.dekker.com.)*

compatible with a wide range of organics. Permeate channels are open over the whole periphery of the modules, which are housed in a vacuum vessel directly piped to the permeate condenser. Alternative designs mimic plate heat exchangers, in which the supported membrane replaces the heat exchanger plates. Internal ducts channel feed, retentate, and permeate. The drawback with such modules is the risk of gasket failure releasing process fluids to the outside (Fig. 11).

Spiral Wound Modules

Spiral wound modules with stainless steel central tubes, but otherwise similar to those known for other membrane processes, are mainly used for removing small quantities of organics from water with organophilic membranes. Glues can more easily resist this environment because of the low organic content and low temperatures. Multiple spiral wound modules are housed inside pressure tubes and assembled in banks in conventional skids. In a special design, the sandwich structures of membranes, permeate and feed spacer is welded together and not spirally wrapped around the central tube, but are arranged as flat sheets on the central tube for the removal of the permeate (Fig. 12).

Tubular Modules

Tubular modules are now being introduced for pervaporation and vapor permeation with number of advantages. Ceramic membranes are preferably coated on tubular substrates; coating the outside of the tube allows high operating pressures on the feed side. The tube-in-tube design, with feed flowing in multiple passes through the annulus between external steel tubes and central ceramic tubes, combines high turbulence in the feed channels with low feed flows per unit membrane area (low loading). Steam heating the shell of such a module ensures isothermal operation and the low loading allows almost all the permeating component to be removed in a single module (Fig. 13).

Tubular pervaporation and vapor permeation modules are also under development to house polymer membranes and promise more predictable performance, easier membrane replacement, isothermal operation, and lower costs.

Table 1 General guidelines for process selection

	Continuous pervaporation	Batch pervaporation	Vapor permeation
Tolerance to dissolved or suspended solids	Poor—membranes damaged by solids	Poor—membranes damaged by solids	Blow down from evaporator or column avoid entrainment
Flexibility to change feed, operating conditions	Fair—feed rate cannot normally be increased	High—change batch start and end conditions, batch time	Low—operation also tied to pressure of vapor feed
Drying non-polar organics (high water activity)	Suitable because of high water partial pressure relative to concentration	Suitable because of high water partial pressure relative to concentration	Not suitable, large membrane areas required
Steam requirement	Low (Fig. 8A)	Low (Fig. 8B)	Moderate (Fig. 8C)

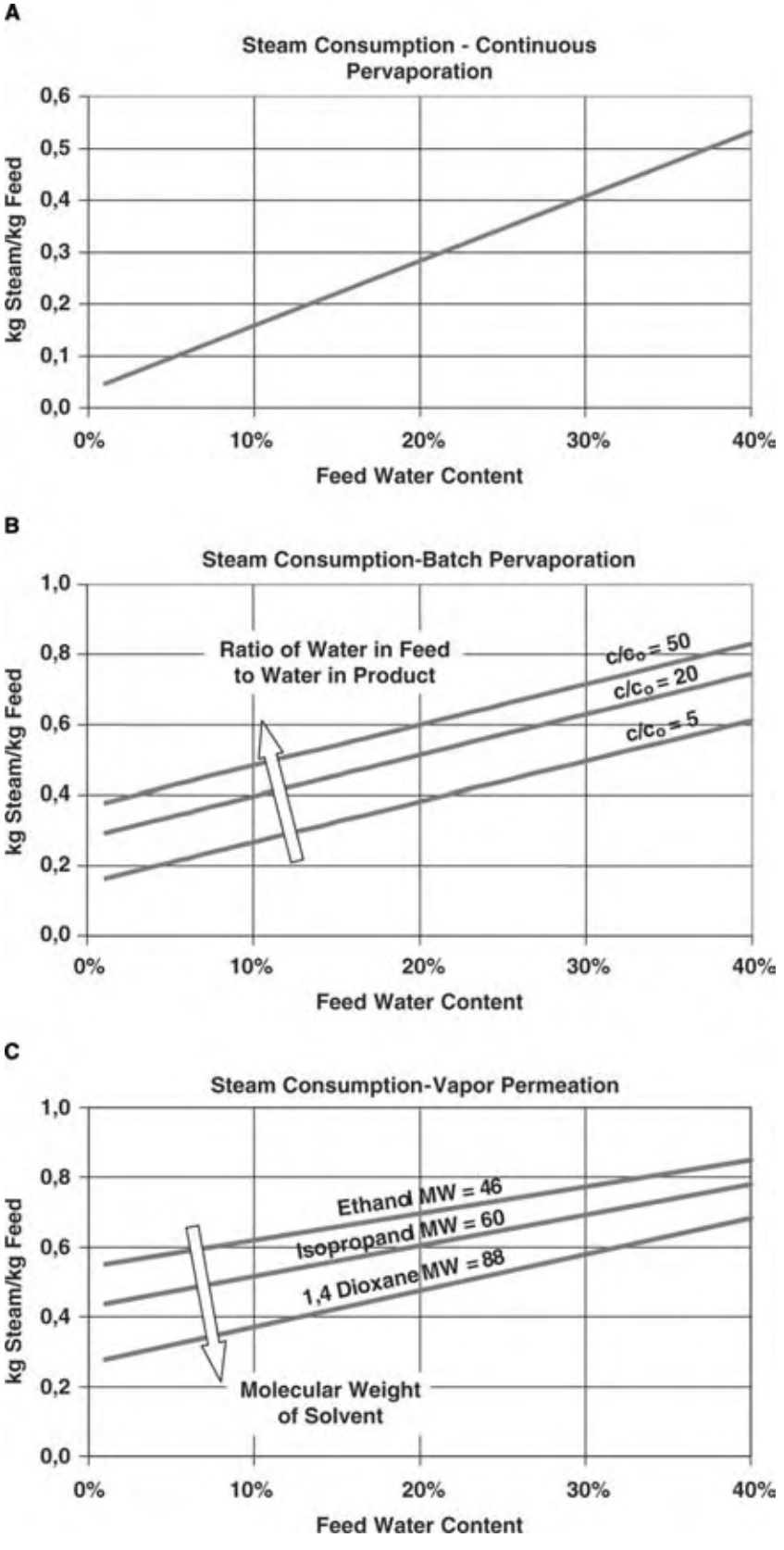


Fig. 8 Steam consumption. *(View this art in color at www.dekker.com.)*

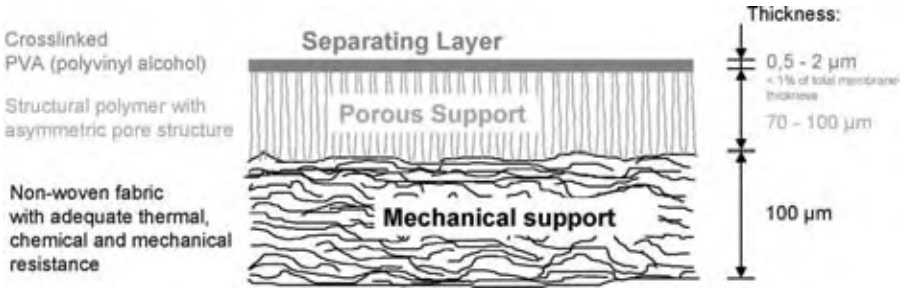


Fig. 9 Membrane structure. *(View this art in color at www.dekker.com.)*

Hollow fiber designs are only rarely used, generally with the feed flow inside the bore of the fiber. The more conventional arrangement with feed outside the fiber has the disadvantage of high permeate pressure loss.

MODELING PERVAPORATION

Membrane Testing and Data Correlation

Performance of a specific membrane with a particular feed is typically determined in a laboratory experiment. A heated quantity of feed is run over the membrane in a batch pervaporation test, and samples of feed and permeate are taken periodically and analyzed. Permeate rate is also measured. The temperature of the circulating liquid is thermostatically controlled and the permeate pressure is kept low, perhaps 5–10 mbar. Overall flux rates for permeating and nonpermeating components are determined by mass balance and plotted together with permeate composition against the feed composition.

From such results an Arrhenius temperature dependency can be correlated as follows:

$$\dot{n}_i'' = f(x_{iF}) \exp(-k/T_F)$$

Theory

In general, the flux of component i through a membrane is proportional to the membrane area and depends on the state of the feed (pressure and temperature), the composition of the feed, and the back pressure on the back-side of the membrane:

$$\dot{n}_i''/A = \dot{n}_i'' = f(P_F, T_F, \text{composition}_F, P_P)$$

In the case of pervaporation, because the feed is a liquid and the permeate is a vapor, the pressure of the feed is unimportant. The above equation can then be simplified to:

$$\dot{n}_i'' = f(T_F, \text{composition}_F, P_P)$$

For binary systems with low back pressures the equation can be further simplified to

$$\dot{n}_i'' = f(C_{iF}, T_F)$$

Corresponding to the data sets referred to above.

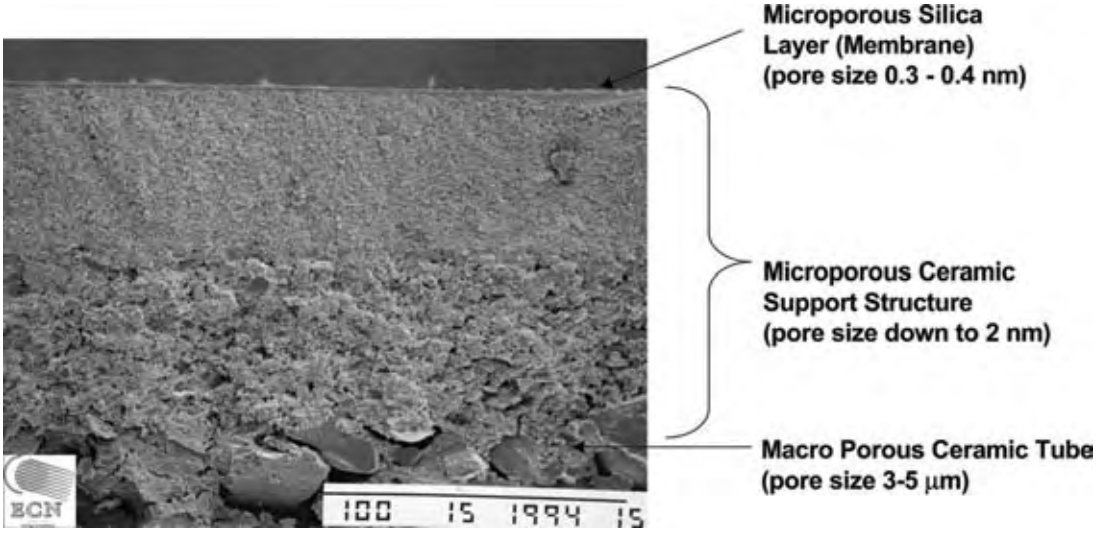


Fig. 10 Cross section of silica pervaporation membrane. *(View this art in color at www.dekker.com.)*

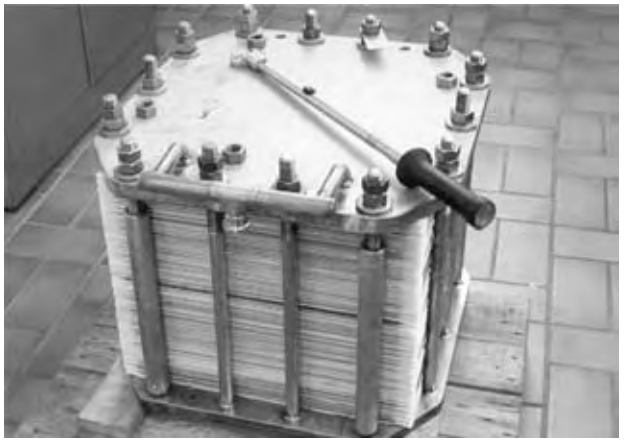


Fig. 11 Plate module. *(View this art in color at www.dekker.com.)*

Modeling for Process Simulators

With the advent of process simulation packages, modeling of pervaporation and vapor permeation processes in a user added subroutine allows these unit processes to be included in overall separation schemes right from the conceptual stage. This enables many different combinations of pervaporation, for example, distillation to be studied and the optimum operating parameters for the preferred configuration to be selected very quickly. Such parameters include membrane feed temperature, which strongly influences the flux rate and, therefore, the membrane area required and permeate pressure, which influences apparent selectivity, as well as permeate condensing temperature.

Most binary system/pervaporation membrane combinations are tested as above with low permeate pressures. The following relationship can be extracted from such results for both fast and slow permeating components:

$$\dot{n}_i'' = J_{oi}(x_{i,\mathrm{F}} p_i^o \gamma_{i,\mathrm{F}} - y_{i,\mathrm{P}} P_{\mathrm{P}})$$

Pervaporation membranes are commonly subject to swelling. This has the effect of increasing the permeance as the concentration of permeating component increases. In addition, permeance generally increases with temperature. Hence, J_{oi} is not a constant, but can be expressed as a function of temperature and of the partial vapor pressure of the permeating component, i.e.,

$$J_{oi} = f(T, p_i)$$

On inspection, it can be seen that this relationship is similar to the one developed above for membrane testing and data correlation. It is, however, easier and more useful when incorporated into generic simulation programs. Even when flux data is only available for the membrane in question with other feed mixtures, the thermodynamic data called up by the simulation

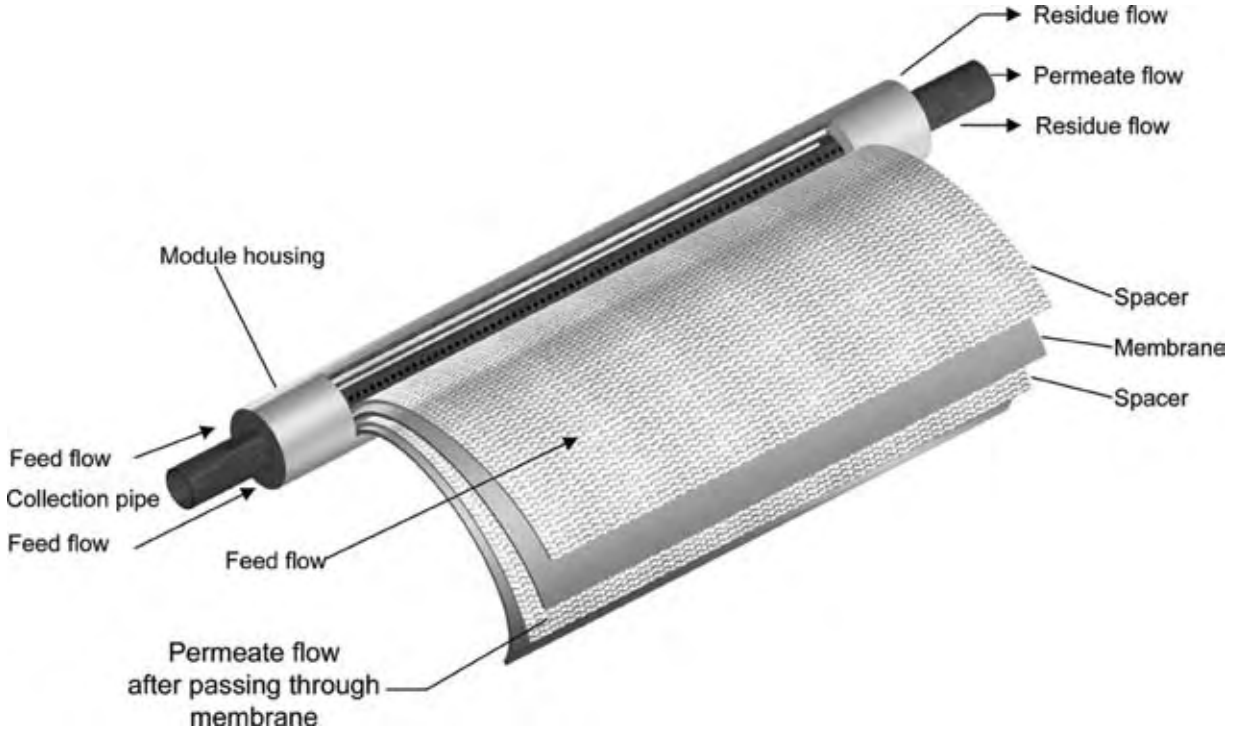


Fig. 12 Spiral wound module. *(View this art in color at www.dekker.com.)*

Fig. 13 High flux pervaporation module for silica membrane. *(View this art in color at www.dekker.com.)*

package (i.e., activity coefficients) allow a reasonable prediction to be made of expected flux rates.

APPLICATIONS OF PERVAPORATION AND VAPOR PERMEATION

Pervaporation is most competitive in situations where simple evaporation or distillation lacks separation power because of low volatility differences. It was first adopted for breaking azeotropes, primarily in solvent dehydration. These small-scale, end-of-pipe applications have served as a low risk, fast payback demonstration of a developing technology. The earliest plants were mainly continuous pervaporation for a dedicated service. Lately, multipurpose batch plants and continuous vapor permeation plants for impure feeds have also been built (Table 2).

Further applications with high growth potential listed above also serve to break azeotropes, or remove trace water. Removing water from condensation reactions can give multiple benefits in certain situations. The largest growth is now in separations that are integral to production in much larger plants. Accumulated experience, higher standards of reliability, and equipment development, which leverages economies of scale are now driving adoption of the technology in these situations.

Solvent Dehydration

Solvents are used for a variety of purposes in the pharmaceutical industry, e.g., for synthesis of pharmaceuticals, to precipitate materials from aqueous solutions, for cleaning purposes and for drying final products.

Spent solvents nearly always contain some water. Dehydration is an essential step in their recovery, but is difficult because most polar solvents form azeotropes with water. Final water removal by distillation is

Table 2 Applications of pervaporation and permeation

Application	Task	Processes	Membrane	Status
Solvent dehydration	Breaking azeotropes	Batch and continuous pervaporation, vapor permeation—often coupled with distillation	Hydrophilic, e.g., PVA polymer composite, ceramic	Well-established
Debottlenecking distillations	Breaking azeotropes	Continuous pervaporation and vapor permeation	Various	Growing
Water removal from condensation reactions	Breaking azeotropes, trace water removal	Batch and continuous pervaporation, vapor permeation—often coupled with distillation	Hydrophilic, e.g., PVA polymer composite	Growing
Methanol and ethanol removal	Breaking azeotropes	Batch and continuous pervaporation, vapor permeation—often coupled with distillation	Hydrophilic, e.g., PVA polymer composite	Growing
Organics removal from wastewater	Trace organic removal	Continuous pervaporation	Organophilic, e.g., PDMS, POMS polymer composite	Sporadic only
Separating hydrocarbons	Various	Continuous pervaporation	Various	Embryonic
Drying of reaction feed streams	Trace water removal	Continuous pervaporation	Hydrophilic, e.g., PVA polymer composite, ceramic	Embryonic

impossible or complicated. Entrainer use is not an option for pharmaceutical or fine chemical production, where stringent process certification rules out the introduction of potential sources of contamination.

Pervaporation dehydrates solvents without using any third substance or entrainer, simply, cheaply, and without problems and irrespective of vapor/liquid equilibria. On site solvent recovery by using pervaporation and vapor permeation is becoming standard practice in the pharmaceutical and chemical industries.

Solvents Commonly Dehydrated Using Pervaporation/Vapor Permeation

Isopropanol, ethanol: Standard applications for pervaporation, typically dehydrated from their azeotropes to fractions of a percent of water. Many continuous, batch and vapor permeation units are operating around the world.

Ethyl acetate, butyl acetate: Form azeotropes in the miscibility gap and were traditionally dehydrated by two distillation columns and a phase separator, however, with a massive recycle. Esters decompose in contact with zeolites. Pervaporation/vapor permeation is easily the best technique for dehydration.

THF: Easily dehydrated by pervaporation down to a few 100 ppm water. Traditional caustic washing is operationally messy, requiring a redistillation of the product. Pressure swing distillation requires high pressures and large recycles.

MEK: Pervaporation is again the preferred technique. Distillation is only possible with an entrainer because the azeotropic composition is nearly identical to the miscibility limit.

n-Butanol, n-propanol: Form azeotropes with high water content, hence the distillation/phase separation process involves massive recycle streams. Pervaporation plants are less costly to build and easier to operate.

Acetone: Does not azeotrope with water, but when distilled, a large reflux is required to get a half dry product. Pervaporation is ideal for final dehydration or for debottlenecking existing distillation systems. Very often pervaporation and vapor permeation plants are designed for operation with various solvents or with solvent mixtures.

Pervaporation and vapor permeation offer many benefits when dehydrating solvents:

- No introduction of additional chemicals—complete solvent dehydration by pervaporation membranes irrespective of azeotrope formation—no possibility of contamination.
- A choice of batch or continuous pervaporation systems, or continuous vapor permeation depending on the duty.
- Able to dehydrate esters without any decomposition.
- Low energy consumption.

Solvent Recovery from Mother Liquors

Spent solvents typically contain some water and are often saturated with dissolved material. Such mother liquors cannot be re-used without purification. Evaporation combined with vapor permeation is a powerful technique for purifying and dehydrating mother liquors.

The feed of spent solvent is evaporated and the resulting vapor is fed directly to a vapor permeation unit. Water vapor selectively permeates the membrane and is condensed under vacuum. The water free solvent vapor leaving the vapor permeation unit is condensed and is stored for reuse (product).

A blowdown is taken from the evaporator to prevent build-up of dissolved solids. This purge can be treated to recover valuable components.

Combining evaporation with vapor permeation gives the following benefits:

- Only vapor is fed to the membranes—no possibility of fouling and no possibility of solids carryover into the recovered solvent.
- Both evaporation and vapor permeation process steps are carried out in a single unit.

Debottlenecking Distillations

Debottlenecking pinched distillations

Distillation processes are driven by volatility differences. If these volatility differences are small, or become small under certain conditions, then columns need to operate with high reflux to achieve the desired separation. Because pervaporation/vapor permeation processes separate irrespective of volatility differences, they can be used very effectively to debottleneck pinched distillations.

Consider for example, the system acetone/water. Acetone is concentrated in the vapor phase at low concentrations, hence stripping of acetone from water is easy. At high concentrations this is not the case. Complete dehydration of acetone is difficult (Fig. 14).

Debottlenecking entrainer distillation systems

Existing entrainer distillation systems can also be effectively debottlenecked using pervaporation/vapor permeation (Fig. 15). Normally, the rectification column will be operating to give a product as close to the azeotrope as possible, running with a high reflux.

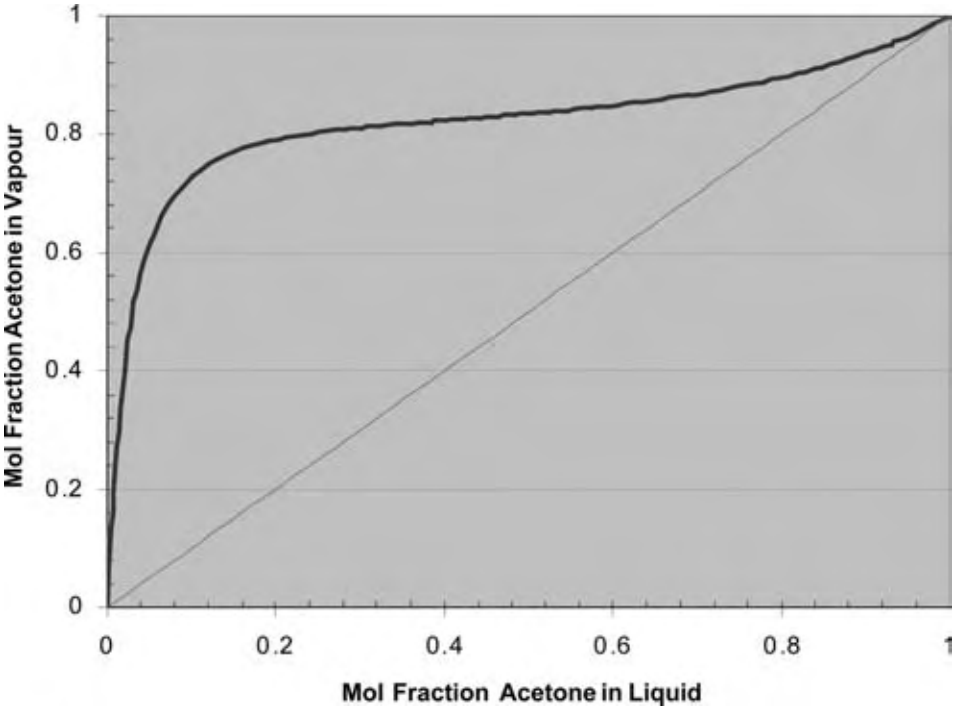


Fig. 14 Acetone/water distillation. *(View this art in color at www.dekker.com.)*

To debottleneck the system, reflux in the rectification column is reduced, giving more overhead product, but with a higher water content. The pervaporation unit is sized to remove enough water that the subsequent entrainer column is also unloaded. Both columns can then realize a significant capacity increase.

The pervaporation unit required for debottlenecking is relatively small, because the driving force for water permeation is high.

Adding a pervaporation/vapor permeation system to a pinched distillation can give the following benefits:

- Higher product capacity and reduced reflux.
- Significantly higher product purity.
- Reduced energy costs.

Recovery of tower side-draws

Trace components in distillation feeds can accumulate in the column if they form azeotropes with the other feed constituents, and if these azeotropes are intermediate boilers. This situation is alleviated by a side-draw at the tray with maximum concentration of

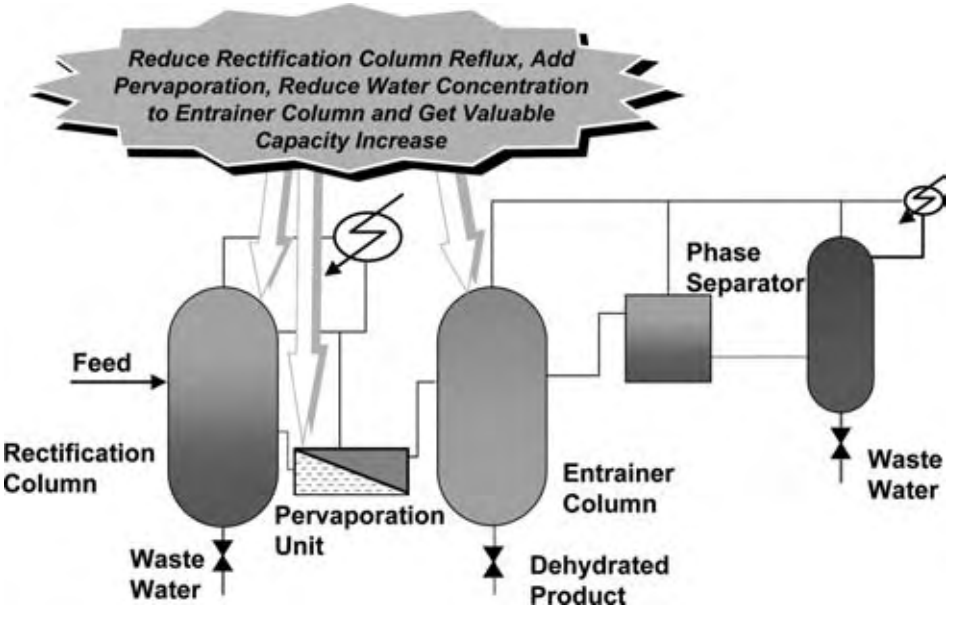


Fig. 15 Debottlenecking distillation systems. *(View this art in color at www.dekker.com.)*

azeotrope, leading to a loss of product, however. If these azeotropes can often be broken by pervaporation/vapor permeation, the product can be returned to the column, without the trace component. Yield is improved and column capacity can be increased as parasitic internal flows are reduced.

This technique has already been commercialized to debottleneck methanol distillation columns and has also been proposed for MTBE rectification (Fig. 16).

Water Removal from Condensation Reactions

Typical condensation reactions include:

Esterification:

$$R'\text{–}COOH + HO\text{–}R \rightleftharpoons R'COOR + H_2O$$

Acetalisation:

$$R'\text{–}CHO + 2HO\text{–}R \rightleftharpoons R'CH\text{–}(OR)_2 + H_2O$$

Ketalisation:

$$R'R''CO + 2HO\text{–}R \rightleftharpoons R'R''C(OR)_2 + H_2O$$

Condensation reactions are normally equilibrium limited, hence removal of co-product water reduces production costs in three ways:

- Purer products—less effort for product purification.
- Higher yield—lower reagent consumption.
- Faster reaction—greater reactor throughput.

Fig. 17 shows how on-line water removal shifts the equilibrium in a batch esterification.

The optimum scheme for water removal from these reactions depends on the relative volatilities of the reactants and products, and whether the units are operated batchwise or continuously. The examples below illustrate how pervaporation or vapor permeation can be used in particular situations.

Batch condensation reactions of ethyl and propyl alcohols

Unless producing volatile products, the typical procedure for these reactions is to dissolve the acid in an excess of alcohol, add catalyst, and then heat to drive off a water/alcohol mixture. This vapor is fed to a distillation column; the reaction water leaves as bottom product, and the alcohol/water azeotrope leaves as top product. At the beginning of the reaction, the azeotropic mixture is fed back to the reaction and the reaction still proceeds at a reasonable rate. As the reaction progresses, the concentration of product in the excess alcohol approaches an equlibrium because of the water in the reactor and the reaction slows. At this point, recycling of wet alcohol is discontinued and fresh dry alcohol is added instead. Gradually, co-product water is distilled away, correspondingly, more acid is reacted and a reasonably high yield is obtained, if sufficient time is available. The reaction product, excess alcohol, and unused acid are then separated and a further batch is started. Normally a batch is started with wet alcohol, generated at the end of the previous batch.

There are three ways in which pervaporation or vapor permeation can be used to enhance such reactions:

- *Offline pervaporation* (Fig. 18) is the simplest procedure and can be very effective. In this case, alcohol/water azeotrope from the top of the column is collected in a tank over the latter part of the

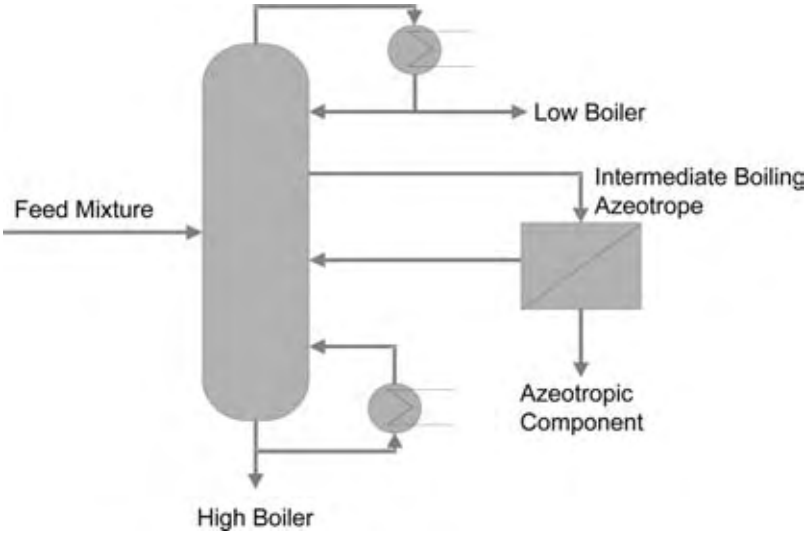


Fig. 16 Debottlenecking by splitting side-draw.

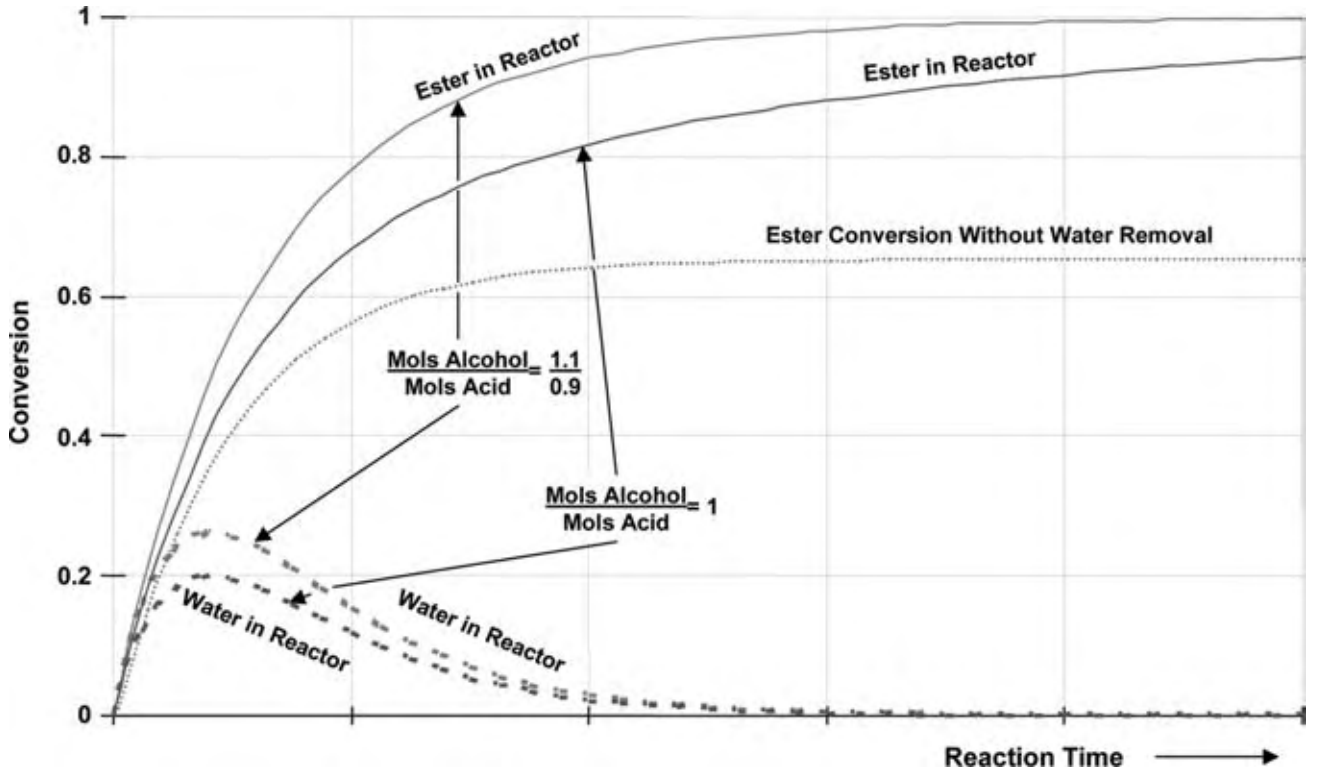


Fig. 17 Progression of a batch esterification (with equilibrium constant $K = 4$) with continuous water removal by pervaporation. *(View this art in color at www.dekker.com.)*

batch. The tank contents are continuously dehydrated by pervaporation and the resulting dry alcohol stored in a second tank. This material is fed to the process over the latter part of the batch.

- *Online vapor permeation* (Fig. 19) is used to remove water directly from the product of the distillation column. Dry alcohol is continuously fed back to the reaction.

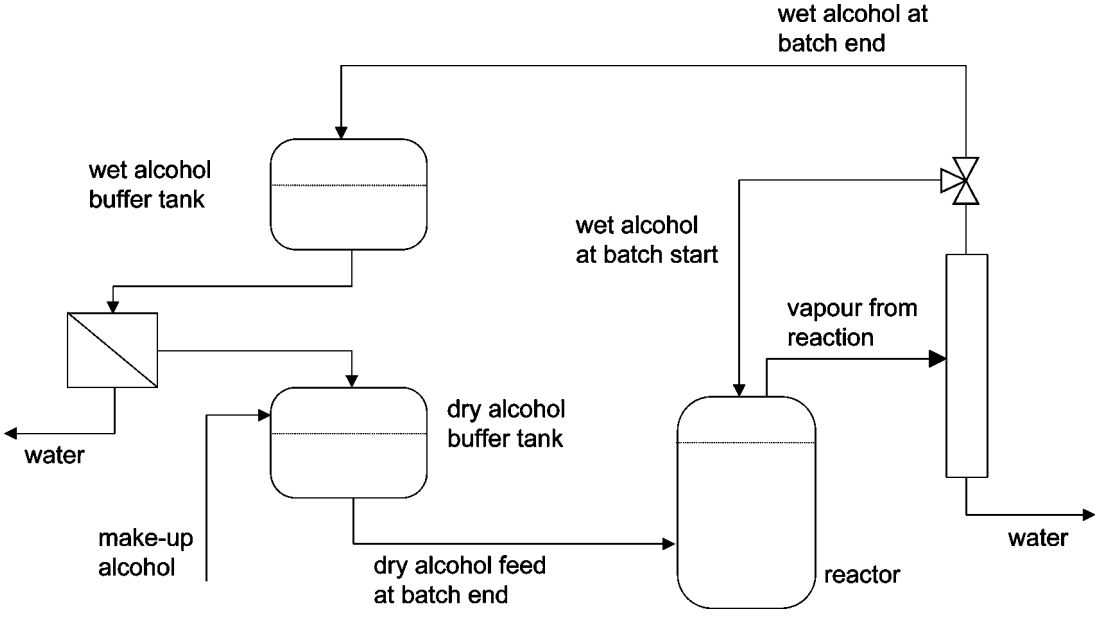


Fig. 18 Reaction water removal—offline pervaporation.

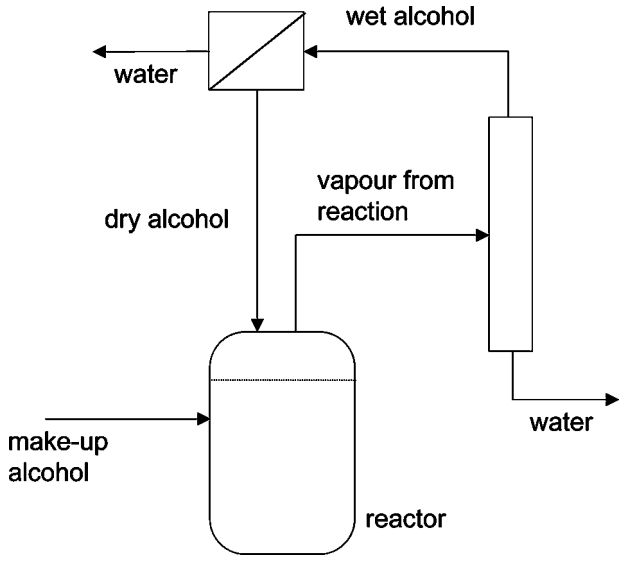


Fig. 19 Reaction water removal—online vapor permeation.

- *Online pervaporation* (Fig. 20) replaces the distillation column altogether. Reaction mixture is continuously pumped through the pervaporation unit at a high rate, and a drier stream is returned to the reactor.

In general, online pervaporation will give the most benefits, provided the membrane is able to handle the reaction mix. Such units are simple, and a high pump-around rate enables water to be removed very fast at the beginning of the batch, when it is being generated fastest. Although highly crosslinked, PVA membranes can cope with a few percent of sulfuric acid catalyst, and problems can arise with aggressive catalysts and with impurities in the feed. In addition, the pores in a composite membrane can become blocked, if the acid or product permeate the membrane faster than they can evaporate into the permeate vapor.

Online vapor permeation avoids these problems, but the unit is constrained to operate at the reaction pressure (normally atmospheric) and with the flowrate passing through the column. It is difficult to fully utilize such units throughout the cycle.

Offline pervaporation allows the membrane unit to be utilized at full capacity throughout the batch—in fact it can also be used when the reactor is not operating. In addition, such a unit is easily coupled to other reactors. Such units enhance the economy of a batch processing operation, while providing a high degree of flexibility.

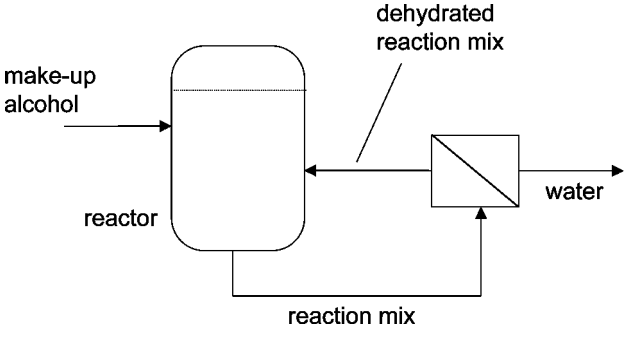


Fig. 20 Reaction water removal—online pervaporation.

Continuous production of ethyl and propyl esters of low volatility acids

The classical scheme for such a process is shown in Fig. 21. Acid is continuously fed into a reactor containing an excess of alcohol. A product mix is continuously drawn off from the reactor containing alcohol, water, ester, and some unreacted acid. Three columns are then used to remove product ester and co-product water. Unreacted alcohol is recycled back to the reaction.

In the first column, the least volatile component, unreacted acid, is taken out at the bottom and recycled back to the reaction. Alcohol, water, and ester pass overhead to the second column. Product ester is taken from the bottom of this column and water and alcohol taken overhead. The third column is used to remove water, again taken out at the bottom. The overhead product from this column is the alcohol/water azeotrope, which is recycled to the reactor.

This scheme has two major drawbacks:

- There is a high concentration of water in the reactor, requiring a large excess of alcohol to drive the reaction.
- The recovered ester is contaminated with trace quantities of water, which hydrolyse the product, decreasing its purity, quality, and usability in many situations.

Fig. 22 shows a reaction scheme enhanced by continuously removing water directly from the reactor. In this case, the water is removed from the vapor phase. A vapor stream is sparged from the reactor and circulated through a vapor permeation membrane module, where water is selectively permeated through the membranes. The membrane unit is sized, such that all the reaction water can be removed with the water/alcohol ratio just below the azeotropic composition.

Reaction mixture is similarly passed through the first column to remove excess acid. In the second column, ester is again taken out at the bottom. However, because relatively little water is in the feed to this column, the water is entrained out with the alcohol. The ester bottom product is contaminated with traces of alcohol instead of water, and hence will not hydrolyse.

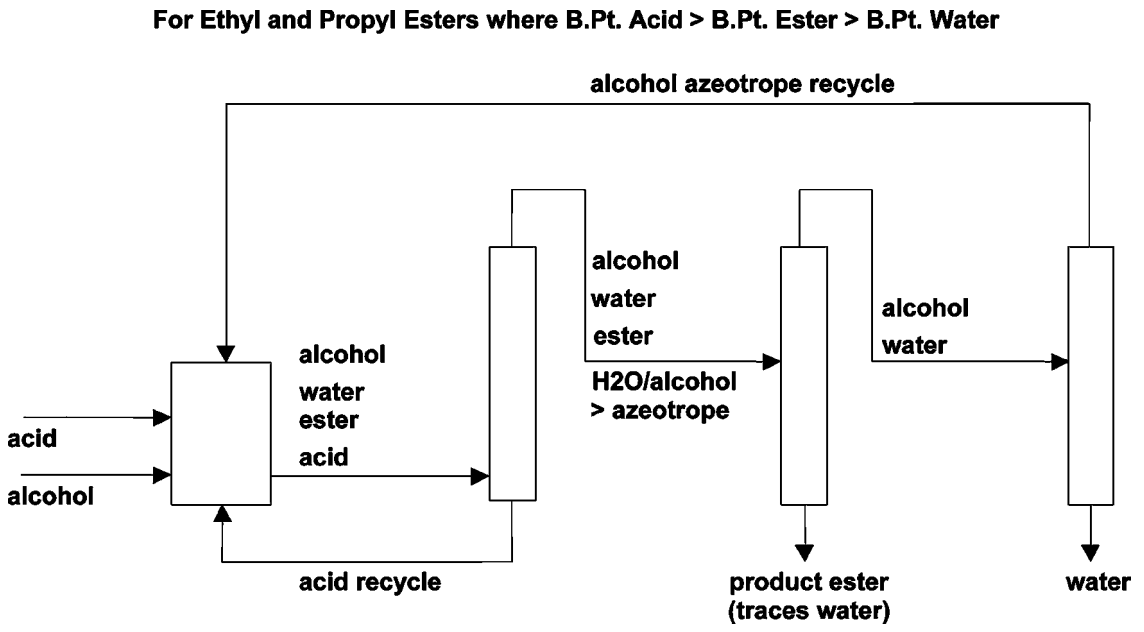


Fig. 21 Classical reaction/separation scheme.

Removing reaction water directly from the reactor improves reaction conditions—the reaction runs faster, lower residence times suffice, equipment costs are minimized, and side reactions are reduced. No third column is required and product quality is better.

Continuous esterification with heterogeneous catalysts

Some condensation reactions are carried out using heterogeneous catalysts. Continuously removing water from such systems brings concrete benefits, particularly in yield.

Fig. 23 shows a scheme for removing water from a continuous esterification process, which uses a heterogeneous catalyst. The process runs in four stages. Each stage includes a reactor, where the components are brought close to equilibrium over the catalyst. The mixture then flows through a pervaporation stage, where water generated in the reaction step is removed, shifting the reaction equilibrium. In the next reaction step, equilibrium is re-established and then again the reaction water is removed. In four such stages, the reaction will be pushed far over to the right and consume nearly all of the feed supplied. Such a procedure not only maximizes reagent usage, but also minimizes

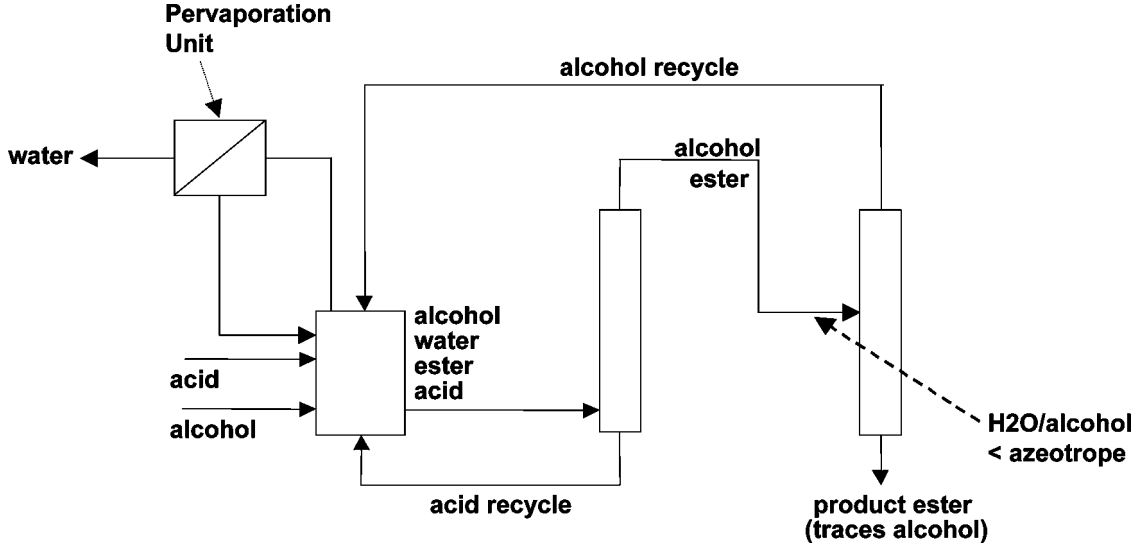


Fig. 22 Enhanced reaction/separation scheme.

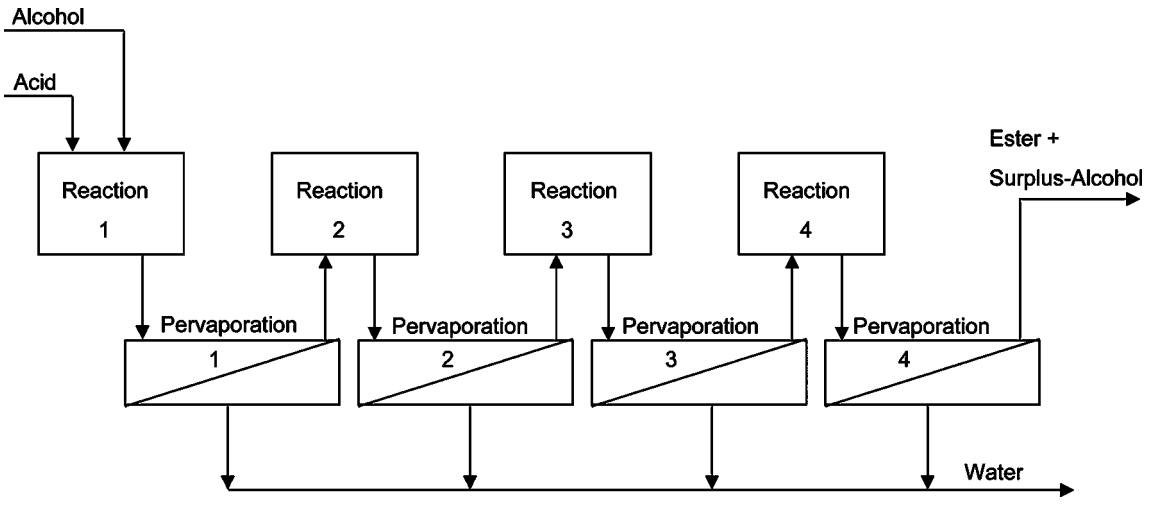


Fig. 23 Four stage reaction/pervaporation cascade.

the separation work required to purify the product. A number of plants of this type are operating.

Transesterifications using methyl esters

Many esters are made using transesterifications, because the milder conditions prevailing permit the reaction of components containing additional functional groups. In alcoholysis, a complex alcohol is reacted with a methyl ester, forming a complex ester and methanol. These reactions are all equilibrium limited, hence the reaction only proceeds if product or co-product is removed. Commonly the reaction mix is distilled to remove the methanol. However, methanol azeotropes react with many methyl esters, hence driving out the methanol also removes a reactant.

Separation of these azeotropic mixtures generated in such situations by traditional means is difficult. However, polymer membranes with a low degree of cross linking preferentially permeate methanol over less polar organics.

Fig. 24 shows a vapor permeation unit installed to remove methanol from the top of a column, treating the boil-off from a transesterification reactor. The column is operated to condense overhead product close to the azeotropic point. Condensed liquid is refluxed through the column. Net overhead vapor is passed through the vapor permeation unit, which is sized to

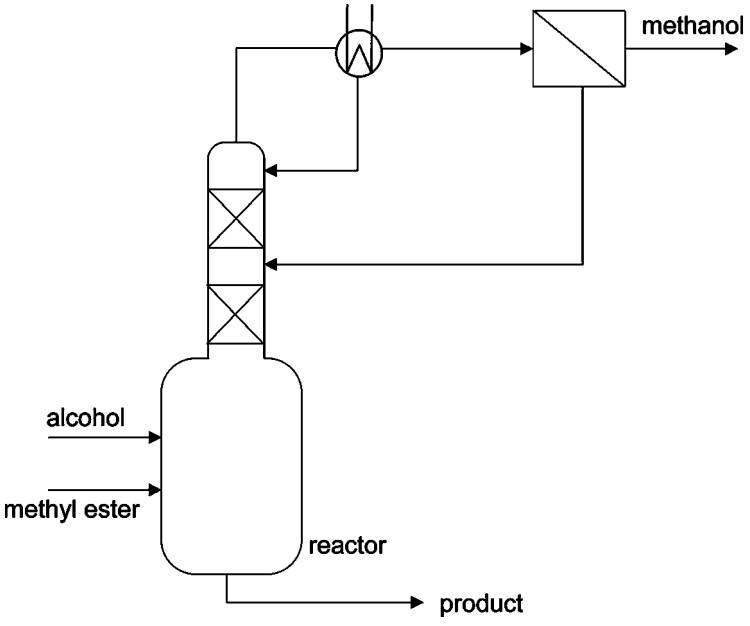


Fig. 24 Methanol removal from a transesterification.

Fig. 25 A pervaporation plant for methanol removal. *(View this art in color at www.dekker.com.)*

permeate methanol at the rate it is generated in the reaction. The recovered ester is recycled to the reaction. In the example shown, membrane area is saved by feeding ester, only partially depleted, in methanol back to the column. The remaining methanol is stripped from the stream as it passes down the column to the reactor.

Such enhancements directly impact the quantity of reagent required, and hence have a direct impact on the bottom line. Payback times for the separation equipment are typically extremely short.

Methanol and Ethanol Removal

Hydrophilic polymer membranes have also been developed, which will permeate methanol, ethanol, and to some extent isopropanol. These compounds can be removed from less polar organics, although membrane selectivity is not as high as when permeating water. Fig. 25 shows an industrial vapor permeation plant, which continuously removes methanol from a methyl ester/methanol azeotrope.

Methanol forms azeotropes with many substances, particularly esters, and often cannot be recovered from spent solvents or from reaction mixtures by simple distillation. Pervaporation provides a simple way to break these azeotropes. Used alone or in combination with distillation, such units provide an economical and reliable route to recover or remove methanol (Fig. 26).

A separation scheme for methanol removal from a methanol rich methanol/ethyl acetate mixture is shown below. The mixture is distilled to the azeotrope, taking out pure methanol as bottom product. The overhead stream is passed directly to a vapor permeation unit, which permeates a methanol rich stream. This stream is condensed and passed back to the methanol column, via the feed buffer (Fig. 27).

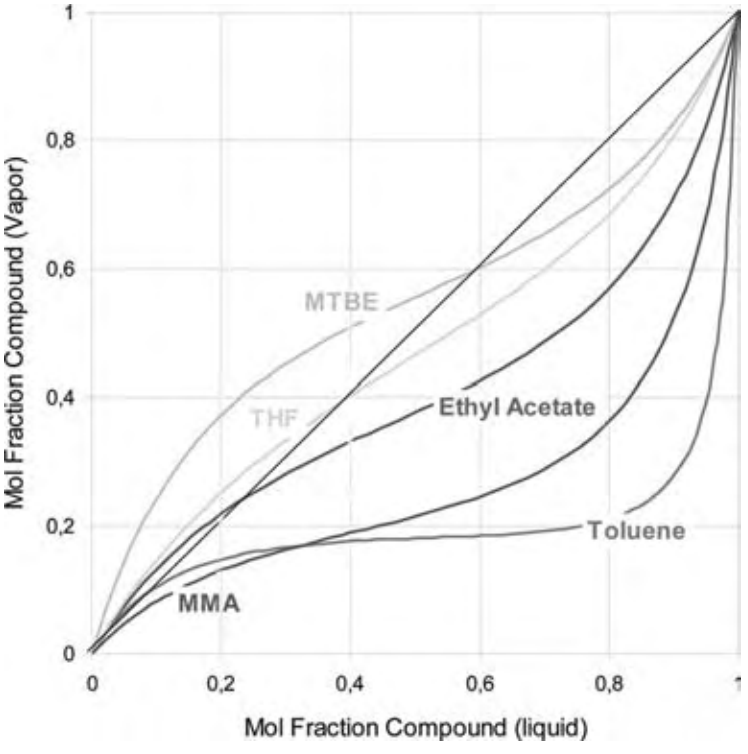


Fig. 26 Azeotrope splitting. *(View this art in color at www.dekker.com.)*

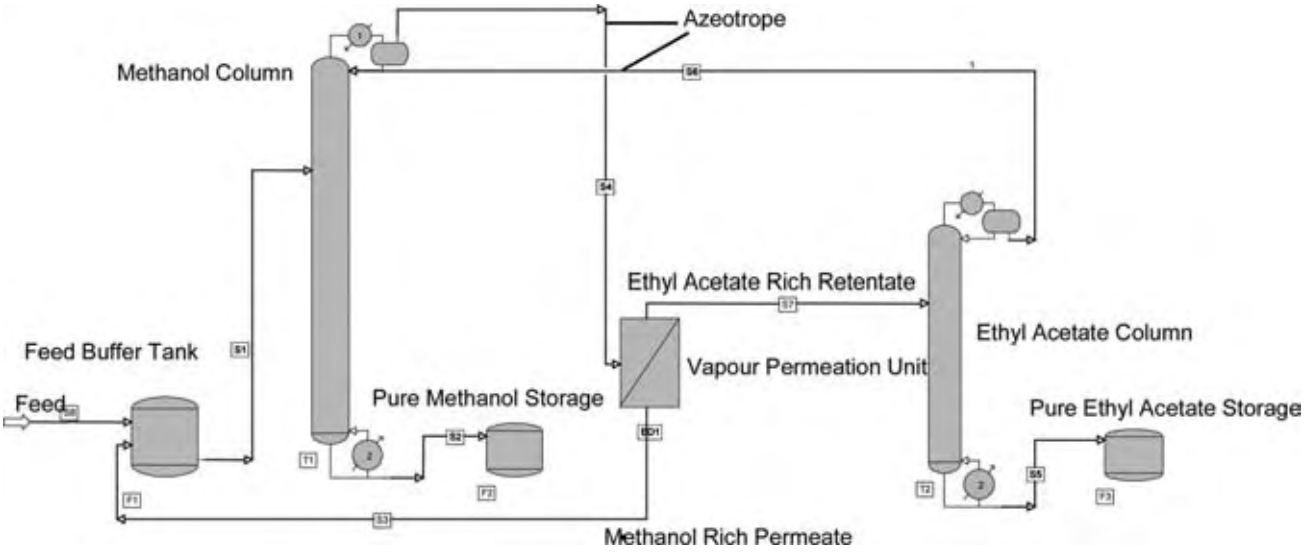


Fig. 27 Methanol recovery by azeotrope breaking—example methanol/ethyl acetate. *(View this art in color at www.dekker.com.)*

Retentate from the vapor permeation unit, strongly depleted in methanol, is fed directly to the ethyl acetate column. Pure ethyl acetate leaves this column as bottom product, while overhead azeotrope is sent to the vapor permeation unit.

Many solvent or ester/methanol mixtures can be separated using a similar scheme. If the feed is close to the azeotrope, then the methanol column can be dispensed with. If the capacity is small the purification column for the second component may not be required, depending on the desired purity. One of the most promising uses of this technology is removing methanol from the products of transesterifications.

Methanol removal from column side-draws

Institut Français du Pétrole (IFP) has developed a process, where pervaporation of methanol is used to debottleneck MTBE production. In the debutanizer columns used in MTBE processing, the MTBE/methanol azeotrope results in a concentration of methanol at a point midway between the feed tray and the

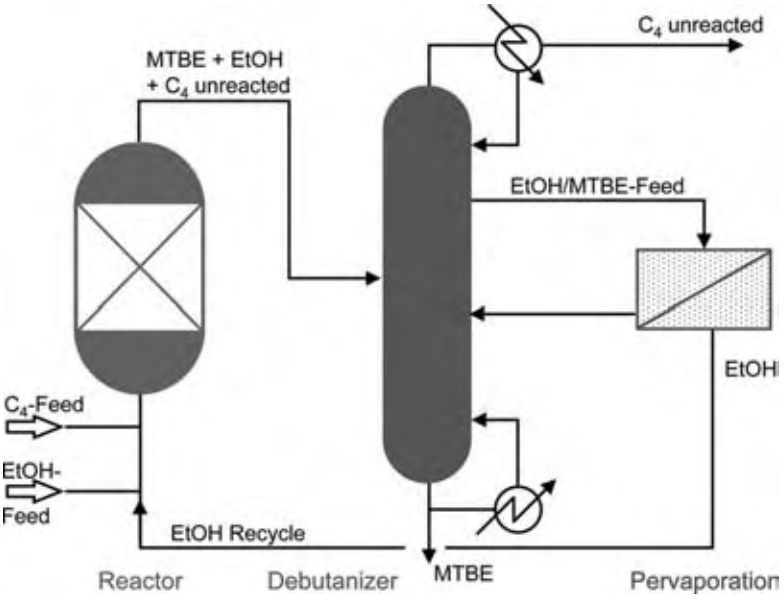


Fig. 28 MTBE-synthesis—optimized process with integrated pervaporation. *(View this art in color at www.dekker.com.)*

reboiler. Pervaporating methanol out of the process from a side-draw taken at this point results in methanol free MTBE as debutanizer bottom product (Fig. 28).

Separation systems, based on pervaporation/vapor permeation of methanol, offer the following benefits:

- Problem-free separation of methanol/organic mixtures irrespective of azeotrope formation.
- Avoids water wash for methanol removal.
- Low energy costs.

Organics Removal from Wastewater

Treatment of contaminated wastewater by pervaporation is superficially attractive, however, only a few commercial installations have been built. Phenol, which azeotropes with water at around 10 wt%, is typically recovered by extraction in large plants, where measures to eliminate extractant loss can be economically applied. Pervaporation is an attractive and cost competitive alternative for small plants. Organophilic membranes in spiral wound modules are used in these applications.

Separation of Hydrocarbons

A number of hydrocarbon separations have been intensely studied and piloted in recent years and commercialization is expected soon. Pervaporation is expected to be one of a number of proven options for sulfur and benzene removal from fuels and olefin/paraffin separations. These plants will use robust, specially engineered polymer membranes, installed in large-scale tubular modules.

Continuous Drying of Reaction Feed Streams

Mol sieves are routinely used for drying a number of reaction feeds to very low water levels. This is particularly important where expensive catalysts are susceptible to wet feed. Pervaporation has the advantage of continuous, steady operation, minimizing operational upsets, which can result from sieve regeneration. Simpler systems are especially attractive where toxic materials are involved.

CONCLUSIONS

The emerging applications outlined above use existing membranes with proven industrial performance. Development of new and better membranes, i.e., with higher fluxes, better selectivities, and broader chemical resistance is being pursued in a number of companies and institutions. These efforts will expand the areas where pervaporation is a viable option. Parallel developments in module design are also opening up new opportunities. In particular, tubular membranes allow the use of module geometries with superior heat and mass transfer characteristics.

These developments will have a wide impact. Reaction enhancement will be a major beneficiary, but a look at the simpler field of solvent dehydration shows that the innovation process is very application dependant. Pervaporation (with vapor permeation) is progressively displacing other techniques in solvent dehydration. Replacing entrainer distillation for drying ethanol and isopropanol, pervaporation at initial stages is always now preferred to techniques, where a third component must be added to shift equlibria. The handling of entrainers and/or calcium chloride or caustic with the attendant environmental risks and costs is no longer a viable option.

Pervaporation has also made inroads in drying solvents, where phase separation facilitates drying by distillation. Drying esters this way requires high reflux ratios, hence pervaporation is easily cost competitive. If the solvent mixes include alcohols, then the phases will not separate and distillation will not work at all.

The field of reaction enhancement is more complex and developments more difficult to predict. The ability of ceramic membranes to run at higher temperatures greatly increases the number of reactions, which can be the considered as enhancement candidates. These reactors can be further enhanced, for example, by using the ceramic tube to support a catalyst as well as a membrane. Equipment of this type is already under development.

NOMENCLATURE

i_i''	Membrane flux of component i per unit area (kmol/m^2/h)
$x_{i\mathrm{F}}$	Mol fraction of component i at a point on the feed side of the embrane
K	Constant for flux dependency on temperature (1/K)
T_{F}	Temperature at a point on the feed side of the membrane (K)
A	Membrane area (m^2)
P_{F}	Pressure at a point on the feed side of the membrane (bar)
P_{P}	Permeate pressure (bar)
J_{oi}	Flux coefficient for component i (kmol/m^2/h/bar)
p_i^o	Vapor pressure of the pure component i at temperature T_{F} (bar)
$\gamma_{i\mathrm{F}}$	Activity coefficient of component i

$y_{i,p}$ Mol fraction of component i in the permeate leaving the membrane at a specific point

p_j Vapor pressure of swelling component j at temperature T_F and mol fraction CiF (bar)

BIBLIOGRAPHY

Baker, R.W. *Membrane Technology and Applications;* McGraw Hill Professional Engineering, 2000.

Balko, J.; Wynn, N. Novel S-Brane™ technology for improved. Ultra-low sulphur gasoline economics. ERTC 7th Annual Meeting, Paris, November 2002.

Ballweg, A.H.; Brüschke, H.E.A.; Schneider, W.H.; Tusel, G.F.; Böddeker, K.W.; Wenzlaff, A. Pervaporation membranes. An economical method to replace conventional dehydration and rectification columns in ethanol distilleries. Fifth International Symposium on Alcohol Fuel Technology, Auckland, New Zealand, May 13–18, 1982.

Binning, R.C.; James, F.E. Permeation. A new commercial separation tool. Refin. Eng. **1958**, *30* (6), C14.

Brüschke, H.E.A.; Schneider, W.H.; Tusel, G.F. Pervaporation membrane for the separation of water and oxygen-containing simple organic solvents. European Workshop on Pervaporation, Nancy, France, Sept. 21–22, 1982.

Hauser, J.; Heintz, A.; Reinhard, G.A.; Schmittecker, B.; Wesslein, M.; Lichtenthaler, R.N. Sorption, diffusion, and pervaporation of water—alcohol mixtures in PVA-membranes. Proceedings of the 2nd International Conference on Pervaporation Processes, Bakish, R., Ed.; San Antonio, Englewood, 1987; 15 pp.

Kober, P.A. Pervaporation, perstillation, and percrystallization. J. Am. Chem. Soc. **1917**, *39*, 9444.

Martin, N. *Separating Azeotropic Mixtures,* Sulzer Technical Review, 3; 1998.

Martin, N. Rückgewinnung von Lösungsmitteln aus Mutterlaugen durch Dampfpermeation. Chemie Technik **1999**, *28*, 2.

Martin, N. Flexible and efficient. Batch pervaporation units for flexible dehydration. Process Eur. **2000**, *1*, 20–21.

Martin, N. Simple and effective. Remove methanol from organics using pervaporation. Process Worldwide **2002**, *4*, 52–53.

Naylor, T.De.V.; Zelaya, F.; Bratton, G.J. The BP-Kalsep pervaporation system. Proceedings of the 4th International Conferences on Pervaporation Processes, Ft. Lauderdale, Florida, Bakish, R., Ed.; Englewood, NJ, 1989; 428 pp.

Smallwood, I. *Solvent Recovery Handbook;* British Library Cataloguing in Publication Data, 1993.

Strathmann, H. *Industrial Utilization of Membranes and Membranes Processes*; University of Twente: The Netherlands.

van Veen, H.M.; van Delft, Y.C.; Engelen, C.W.R.; Pex, P.P.A.C. Dewatering of organics by pervaporation with silica membranes. Paper Presented at the 7th Aachener Membran Kolloquium. Aachen, Germany, March 9–11, 1999.

Wynn, N. Pervaporation and Vapor Permeation Processes in the Chemical Industry. CAV **1998**, *11*.

Wynn, N. Dehydration with silica pervaporation membranes. Sulzer Tech. Rev. **2000**, *3*.

Petroleum Refinery Distillation

Stanley Marple
Chemical Engineering Department, University of Houston, Houston, Texas, U.S.A.

INTRODUCTION

Distillation in the petroleum refinery is different from most distillation operations in two principal ways:

1. Thousands of molecular species are present. The separations are roughly by molecular weight, thus yielding a differentiation of physical properties.
2. In many cases, the distillation feed contains species of high molecular weight, which cannot be readily boiled in shell-and-tube heat exchangers such as those used in most distillation processes.

The first problem makes it difficult to predict distillation separations because the large number of species would overload computational systems. The problem is solved by first analyzing feed stocks by precision batch analytical distillation or equivalent gas chromatography. The stocks are then divided (on paper) into groups of narrow boiling-point range. As volatility is represented fairly accurately by atmospheric boiling point, each boiling-point range, e.g. 200–225°F, can be represented as a single pseudocomponent at the fraction's mean boiling point. Of course, in the lower boiling-point ranges, individual chemical species can be analyzed and entered as such into the distillation model. Then the refinery distillation operation can be studied as a familiar multicomponent distilling calculational model.

The second problem, providing reboil heat for heavy (high molecular weight) stocks, is handled by using fired heaters with short thermal exposure time to preheat the feedstocks. The distilling column is then set up primarily as a rectifying column with side-stripped draw streams to separate the products. Usually superheated steam is used to strip volatiles from each side-draw stream and the bottom product. Less often the strippers may be reboiled. In some operations, the feed preheat may be supplied by a catalytic or thermal reactor. Product separation of the heaviest fractions can be carried further by reheating the bottoms from the first distillation and rectifying the distillate under vacuum. In the interest of concise presentation, this article will be primarily devoted to refinery distillation separations handled by the methods described above, with a brief discussion only of other separations that use mostly well-known general distillation processing methods.

CRUDE OIL AND CRACKED PRODUCT DISTILLATION

In most refineries, only two basic schemes are used for crude oil distillation: single-column and two-column arrangements. In the single-column arrangement, liquid crude oil is pumped at high velocity through a series of preheat exchangers and through the coils of a fired preheat furnace, where partial vaporization takes place. The exit temperature from the fired heater is often in the vicinity of 720°F or 382°C, limited by thermal degradation. The two-phase stream is introduced into the feed zone of the crude tower (Fig. 1). The unvaporized liquid portion of the feed flows by gravity over a section of stripping trays on which superheated steam vaporizes any volatile components.

The pressure in the preheat train is maintained high, perhaps 500 psi, to prevent vaporization, because the flow must distribute accurately among parallel piping as it passes through the preheat furnace.

If the crude oil contains a substantial amount of water and volatile hydrocarbons, it may be best to reduce preheat system pressure by removing water and volatiles upstream of the warmest preheat exchangers. This separation may be done in a primary distilling column or even in a preflash vessel. Fig. 2 represents the resulting two-column scheme.

In the usual two-column scheme, the primary column depends on recovered heat for vapor formation; a few refiners use a fired heater reboiler. The main crude column runs at low pressure compared to the single-column situation. The main column in two-column systems is often called the ''secondary column'' or the ''atmospheric column.''

Comparison of Single-Column vs. Two-Column Crude Distillation

The choice of single-column vs. two-column crude distillation may depend to some extent on the feedstocks to be handled and on the choice and costs of downstream processes.

The single-column scheme will produce less residue, because the inclusion of water and volatiles in the main-column feed reduces the partial pressure of heavy oil vapor fractions; thus more of the heavy oil can

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120017366

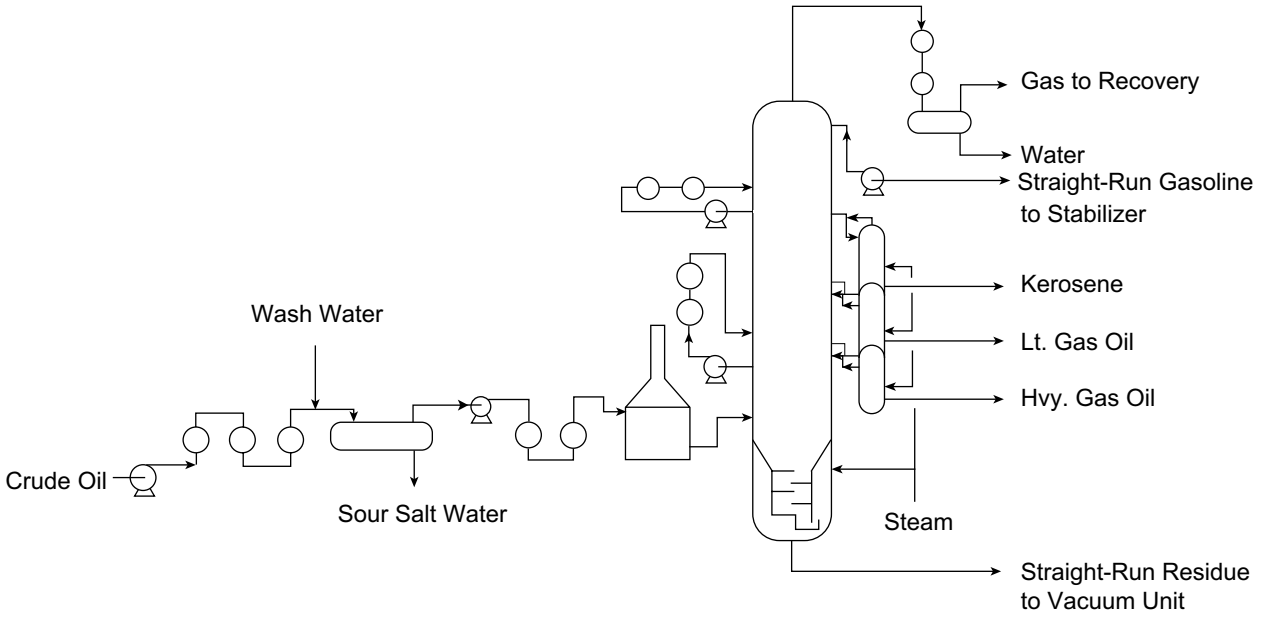


Fig. 1 Basic scheme of single-column crude distilling unit.

vaporize. This effect is similar to steam distillation in principle. The column temperatures for each scheme are the same, limited by thermal degradation.

Consider however the handling of gasoline and naphtha fractions. Most refiners split the naphtha into at least two streams: The lighter straight-run gasoline is often treated for sulfur removal (''sweetened'') by wet chemical processes, while the heavier stream may be hydro-desulfurized. In the two-column scheme, the straight-run gasoline may be an overhead product of the primary column. The heavy gasoline can go directly to hydrotreating as a top product of the main column.

However for the single-column scheme, both light and heavy naphtha fractions are fed to the main column. If they are taken together as overhead liquid product, there will need to be a downstream stabilizer and a naphtha splitter to produce the gasoline streams for treatment.

In the single-column scheme, some refiners take the heavy gasoline as a side draw off the upper main column. With this arrangement, there may be internal

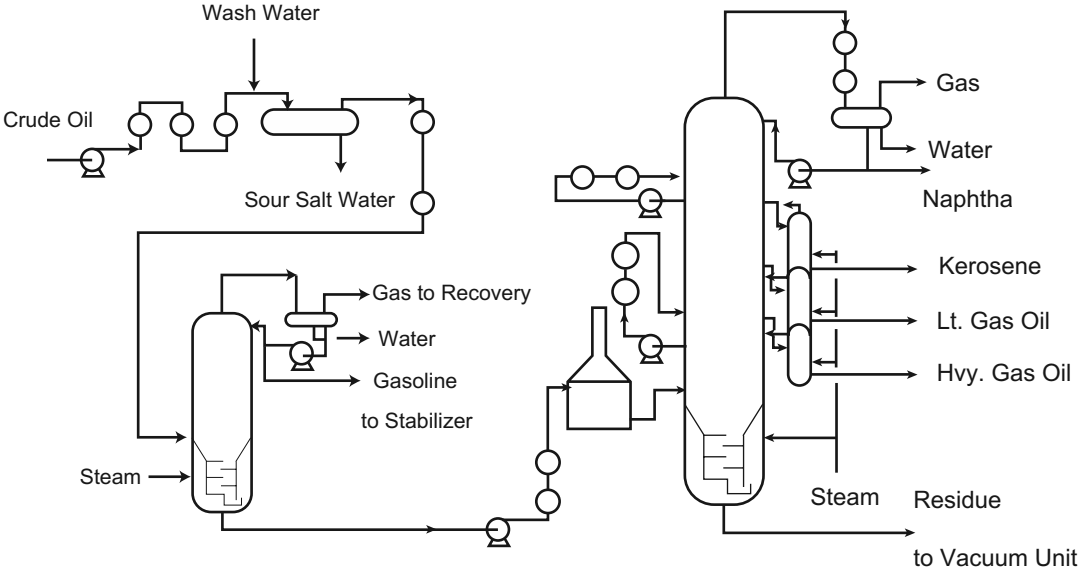


Fig. 2 Basic scheme of two-column crude distillation unit.

water condensation in the main column because of low tray temperatures in the upper column. Even though internal water condensation can be dealt with satisfactorily, it is a nuisance, which requires additional equipment cost.

The main column will be larger in the single-column units, hence somewhat more expensive.

Fuel cost will usually be higher for two-column operations, because more of the available heat from hot products must be used to provide vapor for the primary column. Less of this heat of vaporization is recoverable than in the single-column scheme because the overhead products are cooler in the two-column scheme.

It may be necessary to compare overall capital and operating costs to determine which type of crude scheme is best for a given situation. Occasionally a refiner will convert a single-column system to two columns as a means of expanding throughput.

When volatiles and water are removed in a flash vessel rather than by a primary column, it is sometimes expedient to direct the flash vessel overhead stream into the crude column. Unfortunately, the flashed stream is relatively cold and may contain color bodies from entrained crude residue. These characteristics favor the use of a primary distilling column rather than a flash vessel.

The preheat train

The warm product streams and pumparound heat removal streams from the main column are used to preheat the cold crude oil coming from storage. Obviously the heat recovery system is subject to optimization. If the crude stream is kept intact, there are two ways to optimize: 1) temperature approach on the cold end of each exchange service; and 2) order of the exchangers and number of breaks for the warm streams. This optimization may be done graphically as illustrated in Fig. 3 or mathematically.[1] Another degree of optimization is possible by splitting the crude stream into parallel trains of exchangers. Such splitting adds greatly to capital cost and increases the complexity of the operation.

Fig. 3 shows how the exchange train might be arranged in one typical case. Obviously any change made at the cold end of the train has an effect all the way along the system.

For example, if additional Btus are recovered at the cold end of the train, the crude oil temperature rise will hurt the heat recovery at the hot end of the train, so that the net heat saving might be only half as much as expected. Another complication is that the hot end of the train may require expensive metals to inhibit corrosion.

In operation, the crude oil preheat exchangers will gradually become fouled and lose effectiveness over a period of months. The piping of the exchange train needs to be arranged to permit cleaning on stream.

The crude oil desalter

The crude oil contains suspended droplets of salt water and silt, which must be removed. Demulsifying

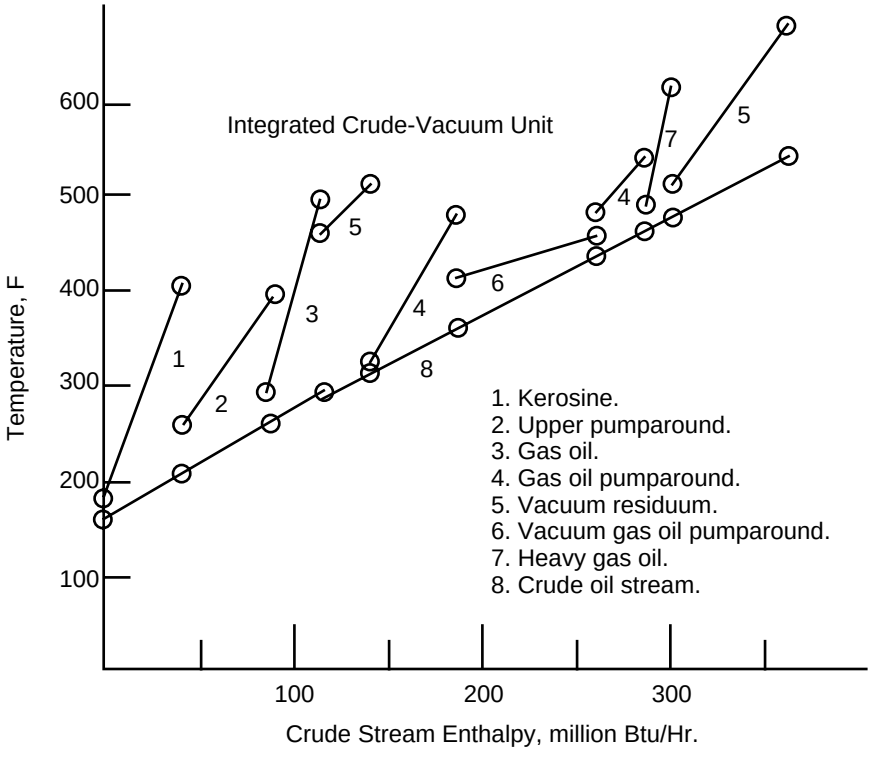


Fig. 3 Typical preheat train.

chemicals and wash water are mixed with the crude stream and it passes to a large horizontal cylinder, which may be equipped with an electrostatic separator, to remove the salt water. Sometimes a second stage of desalting is useful.

The fired crude heater

Design of the fired preheater is a complex subject and cannot be adequately described here. A few points are worth mentioning. Firstly, it is necessary to prevent vaporization before the flow control valves that divide the flow among parallel tube passes. Two-phase flow results in unreliable control; low flow is dangerous and high flow gives inadequate vaporization. Secondly, corrosion-resistant metallurgy is necessary and thirdly, the furnace must avoid hot spots on the tube walls.

And fourthly, it will be necessary to provide for periodic removal of solidified carbonaceous material on the inside walls of the tubes. This is done by controlled burning in the presence of steam; the coke tends to break up (spalling) and react with the oxygen and water vapor to produce CO, CO_2, and H_2O.

Sometimes the coke is too hard for chemical removal and must be drilled out. This decoking is required at intervals from several months to a year or more, depending on the composition of the crude oil and the temperature severity. The exit flow from the heater will normally contain liquid; care should be exercised to make sure that flow is in the mist-annular regime.

The crude column (Fig. 1)

The two-phase stream from the preheater flows into an enlarged feed zone, a vertical cylinder, which has the purpose of separating the liquid and vapor portions of the feed stream and providing well-distributed vapor flow to the trays above. The crude column will be constructed of carbon steel, but may require lining with high-chromium alloy, especially in the hot zones. The type of lining depends on the crude oil composition; naphthenic acidic or high-sulfur crudes are especially troublesome. The phase separation in the feed zone, usually called the flash zone, may be accomplished by gravity alone at low velocity or by centrifugal separation with tangential entry to the feed zone, or with curved internal vanes. If centrifugal flow is present, the swirl needs to be stopped by baffling in the flash zone so that an even vapor flow is presented to the wash trays just above.

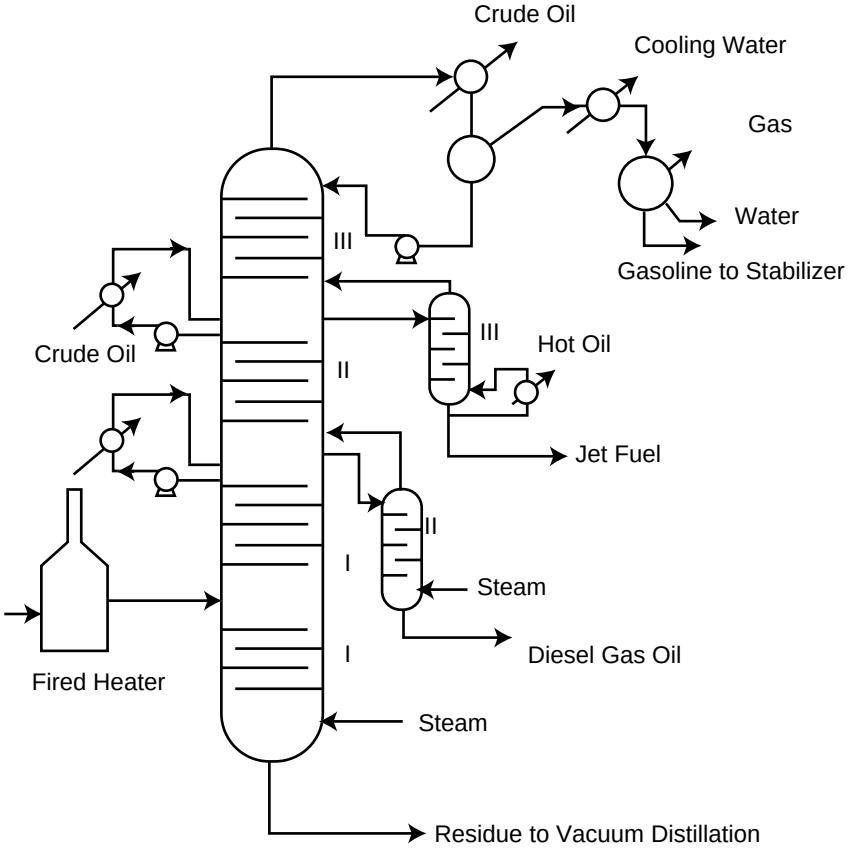


Fig. 4 A crude column as three two-product columns.

The side-product liquid streams shown in Figs. 1, 2 and 4 may be generated at each zone by liquid pumparounds, which also provide internal reflux for rectification. Of course, it would be possible to condense only at the top of the column, but the heat recovery would be greatly damaged because of the poor thermodynamic reversibility. Also, the hydraulic load in the upper column would be a problem.

The trays that carry the pumparound liquid are the heaviest loaded in the tower. Sometimes extra vertical spacing is provided at these trays, or high-capacity trays or packing may be used.

When heavy gasoline (e.g., 200°F to 400°F boiling range) is taken as a side-stream product, the upper trays may be in a water-condensing situation.

Sometimes internal water settling is provided; however, external settlers may be needed because the required settling time is often several minutes. The internal liquid water will be corrosive; tower and tray metallurgy must be corrosion-resistant when liquid water is present.

Side-product strippers (Fig. 1)

Each side-product liquid draw goes through a steam stripper, usually equipped with 5 or 10 trays, to prevent inclusion of more volatile hydrocarbons in the product. The side-product stripper may be reboiled for sharper separation. Optimum amounts of stripping steam will often be in the order of 2% or 3% w or more of each product rate. After stripping, the side products are cooled in the preheat train and cooled further in water or air coolers to safe storage temperatures. Heavy gas oil and residuum products often go to downstream cracking processes without cooling.

Upper crude column trays

As will be shown later, the trays in the main column are mostly rectifying sections for the side products; some of them are pumparound contact condensers. There is a wide variation in the industry of tray count between side draws, from 4 to 15, depending on the degree of rectification desired.

Two or three trays just above the flash zone are wash trays to prevent entrainment of dark color bodies into the heavy gas oil. The internal reflux flow on these trays should be limited to a few percent of the molal vapor rate, for good energy conservation and minimum residue handling. These trays are often equipped with anti-blowing baffles. The wash rate is sometimes controlled by using a total draw for the heaviest gas oil, then metering the wash oil rate.

Residue stripping trays

Five or ten trays at the bottom of the column are used for steam stripping the straight-run residue. Because this section is very hot, it is important to ensure that no liquid steam condensate ever enters this section during operation.

Top condensing sections

Overhead vapors may be cooled with cold crude oil, with cooling water or air coolers. When crude oil exchange is employed, extreme care is necessary to prevent leakage of crude oil into the hydrocarbon condensate. Filming or neutralizing amines or other chemicals may be added to the overhead vapor stream to prevent fouling and corrosion. If heat is recovered from the top condensing system, it is best to use two overhead liquid accumulators. The first would provide only warm top reflux for good heat economy.[2]

Another important factor is the size of the overhead product liquid accumulator. Ten minutes liquid residence time based on net product rate is common; however, aromatic gasoline fractions may require 5 to 8 min settling time for clarification because water vapor tends to condense in small droplets. The required accumulator size is also affected by the type of process equipment downstream.

Computer modeling

Computer modeling of the crude process is worthwhile for many reasons including: 1) initial design; 2) economic optimization of operation; and 3) control and adjustment of product compositions and operating costs. The first models were based on reducing the process to a combination of two-product distillation calculations because of the available computer power. More recently, the computer calculation is able to handle the crude tower as is; however, the two-product combinations are useful to guide technical supervision of the operation (Figs. 4 and 5).

The results of test runs or of computer modeling the separations in the crude tower system are represented as stream compositions in Fig. 6.

Each curve in Fig. 6 is analogous to a population distribution curve for the molecules in the crude feed and products. Each could also be represented as weight or volume distribution. The curves are constructed from the amounts of material with atmospheric boiling points in, for example, three-degree boiling range intervals in the precision distillation or gas chromatographic analyses.

It is apparent from the illustrations that the rectifying sections of each conceptual tower (I, II, or III) are

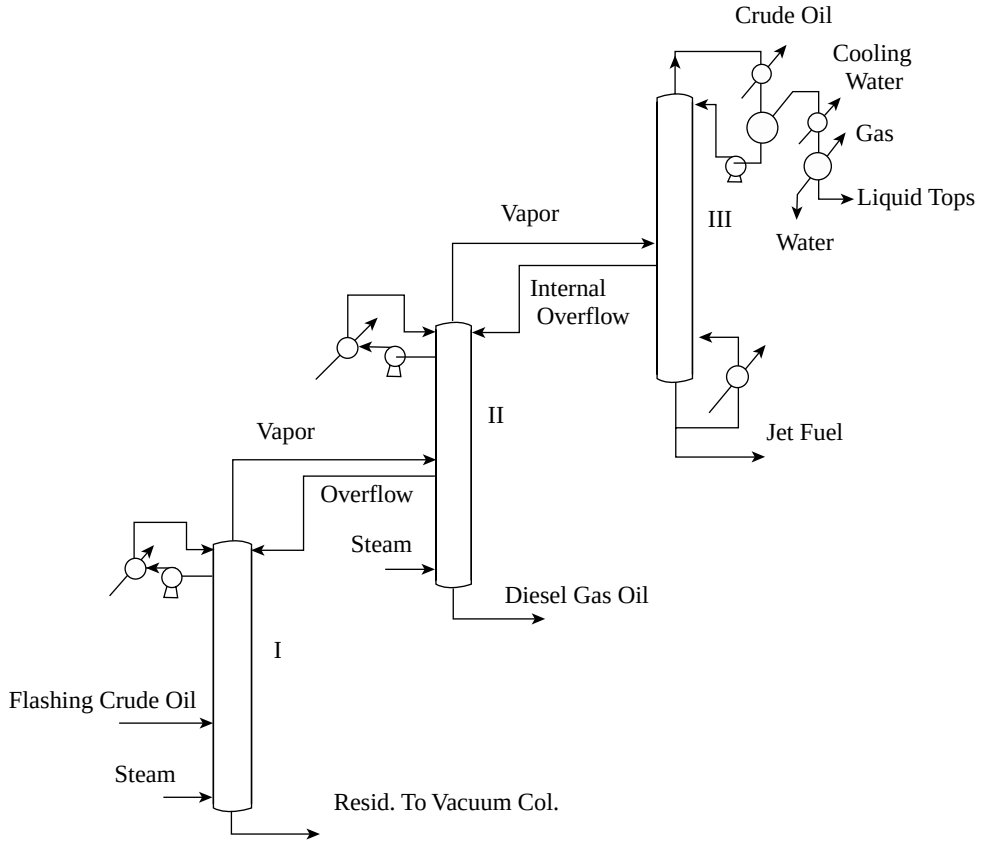


Fig. 5 A crude column as three two-product columns.

parts of the main column above the feed point. These sections are in rectification; i.e., each of them is the primary agent to tailor the back or high-boiling end of each side product leaving the system just above the rectifying trays in question. Similarly, it is apparent that the side-stripper trays have primary responsibility for the volatiles content of each product.

The modeling calculations are usually made with a commercial modeling calculation system, which includes calculation (from correlations) of physical properties and equilibrium stage separations. There are extensive empirical predictions of separations available[3–5] but many successful models are built on a direct use of semi-fundamental distillation calculations.

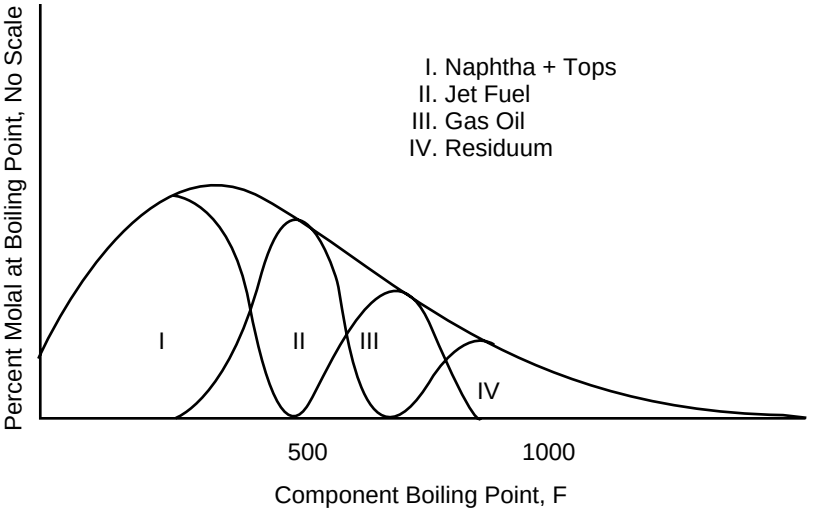


Fig. 6 Typical crude column cuts, molal distribution.

Regulation of crude tower operation

Three types of regulation may be identified as necessary for good operation:

Yield control is the primary control of product properties, accomplished by adjusting the product draw rates. These adjustments must be made in a fashion to maintain overall material balance, i.e., if operation is steady, at least two rates must be adjusted together to maintain steady state. The separation between side draws is sometimes defined as "cut point." This term has had several different definitions; a preferred definition is the atmospheric boiling point of the component equally divided between adjacent product draw streams.

Sharpness of separation can be adjusted by varying internal reflux to adjust removal of high-boiling components and stripping vapor rate to adjust removal of more volatile materials. Changing the side-product draw point has only a small effect on the boiling-point range of the side products. The major effect of tray count is on the degree of rectification of the side products.

As in material balance, heat balance must be maintained when internal reflux rates are being changed.

An example of the need for sharpness control might be the benefit of removing dicyclic hydrocarbons boiling at 400°F or above from the catalytic reforming feed, while maintaining naphtha yield. These dicyclics are known to foul platinum reforming catalysts with carbon.

Rectification of reformer feed can be improved by providing more internal reflux for the upper trays of the crude column, perhaps by using less pumparound heat removal above the jet fuel draw point.

Steadiness of operation is a complex subject. A steady operating pressure is important: when the column is making gas, back pressure control is the usual mode. Otherwise control of condensing rate is involved. Sometimes this is done by injecting non-condensible gas and controlling its removal rate. Smoothness of control requires defense against variation in vaporization rate; fuel composition may vary when refinery waste gas is being burned. Also, variation in condensing rate may require top tray temperature control.

A steady feed composition is important; sudden changes in crude type can be a challenge to operators.

Cracked Product Distillation

The distillation separation of products from cracking operations, whether catalytic or thermal, is similar in principle to the technology of main crude oil distillation columns. As the cracking reactors discharge a product at temperatures in the range of 700°F to 1200°F, no fired preheater is necessary. On the other hand, it is often necessary to add a desuperheating section to reduce the mostly vaporized feed mix to a temperature that will allow distillation without rapid coke formation.

Fig. 7 shows the feed zone of a typical catalytic cracking plant main fractionator. The reactor product is hot vapor; it is fed to the column through a large conduit, which may be several feet in diameter. The desuperheating is accomplished by direct contact in a zone equipped with simple baffle trays or large-size packing, against a recirculated liquid condensate stream, which is cooled by generating steam. There is a net make of heavy black slurry oil, which is drawn off, clarified by gravity or other settling, and sent to No. 6 fuel oil blending. It is necessary to maintain a substantial liquid rate on the desuperheating hardware to prevent carbon deposition.

Attempts to improve vapor distribution to the desuperheating packing or trays have usually been unsuccessful because of carbon laydown on the baffles.

The upper section of the cracking plant fractionator is very similar to the crude column arrangements previously discussed.

Sometimes when cracking throughput is expanded, it is expedient to add top condensing capacity to handle the additional condensing load. Unfortunately, heat recovery is thus wasted and column capacity may be challenged. Attention should be paid to expanding heat removal in the desuperheating and pumparound zones when cracking capacity is increased.

Since cracking processes make relatively more normally gaseous products than found in the usual crude distiller, there is usually a substantial gas compression section also. This gas compression section presents an opportunity to keep the column pressure low, perhaps only 5 or 10 psi gage, in turn allowing lower reactor and catalyst regenerator pressures, which in turn improve cracking yields and reduce regeneration air supply horsepower.

The compression often comprises two or more stages of centrifugal machines with interstage cooling and liquid separation. Discharge to the gas absorption plant will often be at pressures near 200 psi. The liquid interstage condensate may be added to the de-ethanizing absorber to reduce the need for recirculated absorption medium. Cracking plant overhead liquid accumulators are larger than in crude oil distillation, because the more aromatic cracked naphthas require more clarifying time to get rid of the entrained water.

The main column zone just above the slurry return has the primary duty of preventing catalyst particles from being entrained into the gas oil product via entrainment of slurry. Usually only two or three fractionating trays with a small liquid flow will be required below the heaviest gas oil draw tray.

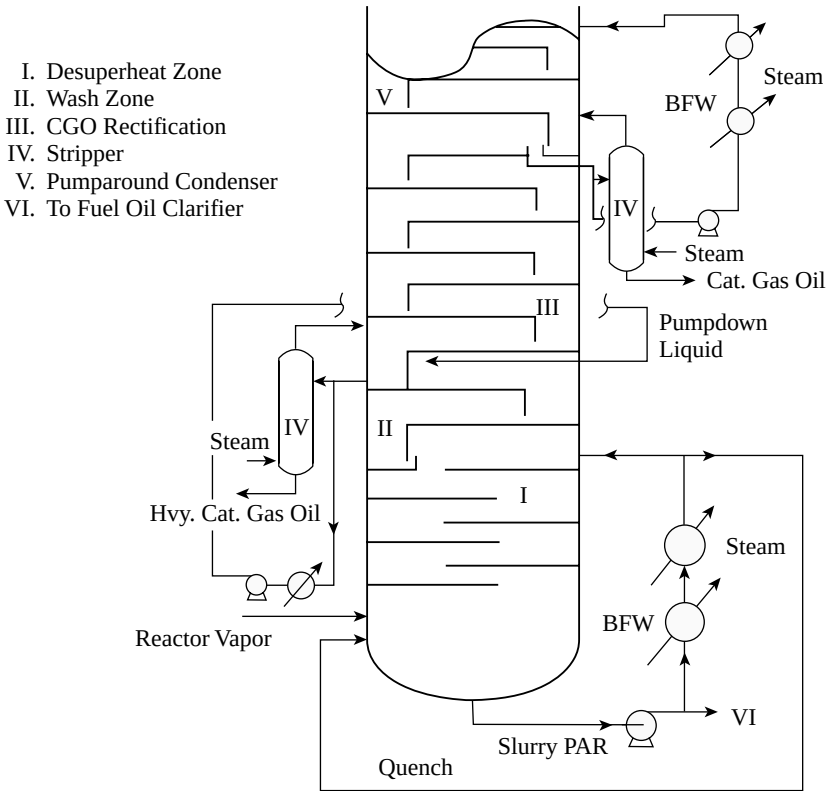


Fig. 7 Lower section of catalytic cracking main fractionator.

Operation of the desuperheating zone is particularly important. If the vapor leaving the zone is too hot, there will be carbon fouling of the fractionation trays. This temperature probably should not be above 700°F. The recirculated cooled slurry oil should not be too cool, however, because if it is as cool as 400°F, for example, the black oil will absorb some volatiles from the reactor product, leading to low flash points for the No. 6 fuel oil.

It is important to get adequate heat transfer in the zone; however, if the contact has too many theoretical stages, the liquid will leave the column too hot and may deposit solids in the downstream heat exchangers. One theoretical stage of contact for the entire zone is a feasible condition in most cases.

Many refiners use a direct injection of cooled slurry oil into the bottom of the column to prevent fouling the exit piping. However, if too much of the temperature moderation comes from this ''quench,'' the desuperheating trays or packing may be deprived of cooling and become plugged with coke. If possible, the temperature of the liquid leaving the bottom of the desuperheating zone should be measured directly. The exact limiting temperatures around the desuperheating zone will vary with different cracking conditions and feedstocks; however, it is well known that condensation reactions for heavy hydrocarbons accelerate above 700°F and usually cannot be tolerated at 800°F.

Coking process fractionators

In principle, contact coking process fractionators are similar to catalytic cracking process fractionators. However, sometimes the fractionator desuperheating zone is also used for preheating coke-drum feedstock. Contact coking fractionators are difficult to control because of the cyclic nature of the process. Also, thermally cracked heavy oil liquids are susceptible to further molecular condensation.

VACUUM DISTILLATION OF HIGH-BOILING OILS

Fuels-type Distillation

Fig. 8 is a schematic representation of a vacuum column, sometimes called a ''flasher,'' intended to produce asphaltene-free cracking feed and heavy black fuel oil. The fired preheat furnace is similar to the crude unit heater, except that tolerable tube skin temperatures are somewhat lower because of the absence of light hydrocarbons.

For lowest capital cost, the preheat furnace is usually a ''dry'' operation. Sometimes, however, two or three percent weight of superheated steam is added to the feed to give a higher internal velocity, thus reducing the rate of carbon laydown inside the tubes. This steam is termed ''velocity steam'' to differentiate

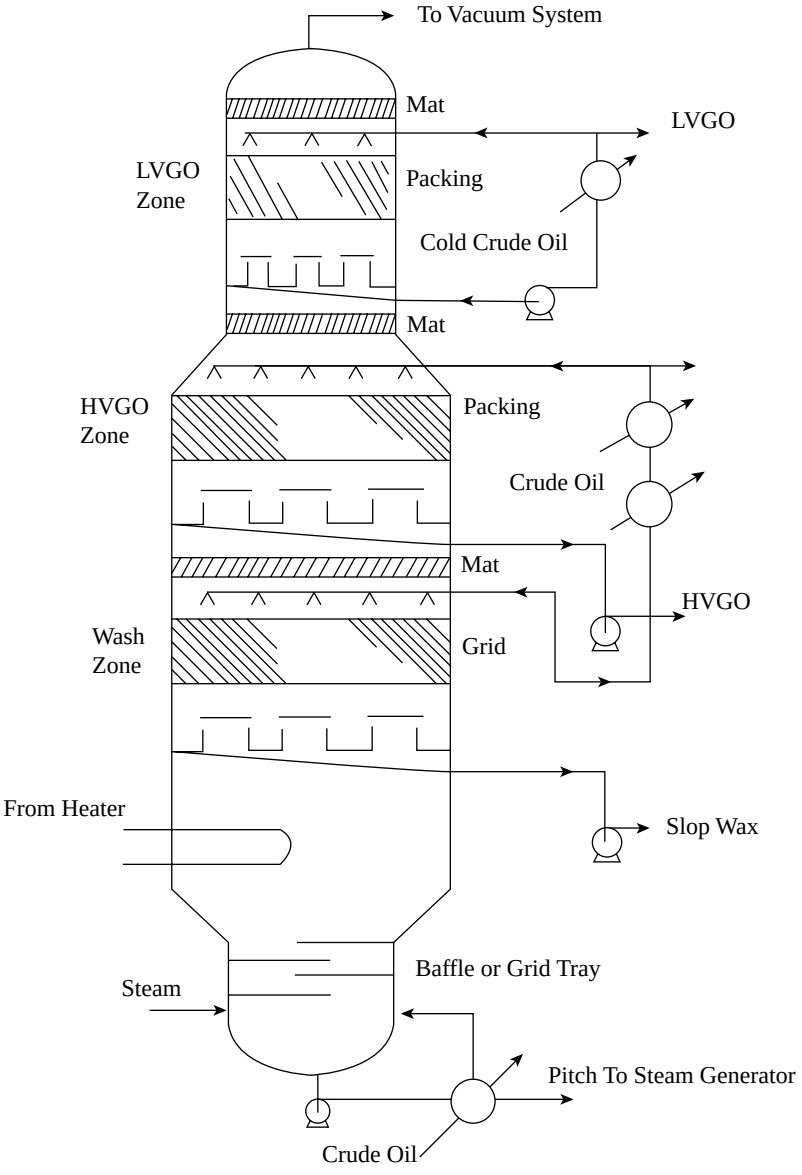


Fig. 8 Fuels-type vacuum column.

it from "dilution steam," which is sometimes introduced to reduce oil partial pressure.

The feed zone is intended to provide a primary separation of liquid residue droplets and vaporized gas oils. Some units utilize centrifugal action to throw the liquid against internal vanes or the vessel wall itself; other units rely on gravitational separation alone, usually at lower velocities than for the centrifugal units. It has been shown that at high velocities, approaching the velocity of sound in the two-phase mix, the liquid film on the surface of the feed piping assumes a wavy shape and that the vapor stream will strip off small droplets that are difficult to de-entrain. If velocities are kept far below the speed of sound, sometimes in the order of 400 ft/sec in this case, equipment sizes become very large.

If the feed zone is intended to be a centrifuge, the swirl should not impinge on the catch tray, because the Bernoulli effect at high vapor velocity can reverse flow through a chimney, thus causing maldistribution in the de-entrainment zone.

The de-entrainment zone is often composed of relatively open grid packing. It is intended more to catch tarry residue than to perform normal rectification. However, the temperatures are so high in this zone, often well over 700°F, that it is necessary to irrigate to prevent accumulation of coke. It is shown by the calculational models that much of the heavy vacuum gas oil usually used for irrigation will re-evaporate in the grid packing, so that the irrigation rate at the bottom part of the packing will be much less than the rate at the top. A minimum irrigation of around 0.3 gpm

per square foot of cross section at the bottom of the packing is sometimes recommended. However, the slop wax draw rate will often be much larger, because the liquid drawn includes the de-entrained vacuum residue, sometimes called ''pitch.'' The slop wax can be analyzed by high-temperature gas chromatography to establish what the real grid wash rate is. If insufficient wash is provided, the grid will eventually plug with coke.

Pumparound condensing zones

Frequently vacuum gas oils are condensed in packed sections. If several zones are used, there will be opportunity for better heat efficiency. Structured packing, sheet-metal type, has largely displaced trays for contact in vacuum pumparound zones because packing requires less pressure drop. Thus, flash zone pressures can be lower and less vacuum pitch can be produced. Flash zone pressures as low as 25 Torr (millimeters of mercury absolute) have become routine. For heat efficiency, a corrosion-resistant de-entrainment mesh pad is often installed above each pumparound packed zone.

The liquid reflux distributors above each packed zone are often arrays of solid-cone sprays; preferably the nozzles should be $\frac{1}{4}$ inch or larger in orifice, free passage, and pipe sizes. The spray liquid needs to be screened at all times. The sprays are not as good distributors as the drip-pan type, but they are economical and contribute to heat transfer without adding substantial pressure drop. The nozzles should be designed to operate at a flow rate that gives a good spray pattern without excessively small droplet sizes.

The vacuum at the top of the flash column is often produced by a sequence of three elevated steam-jet eductors with intermediate and final surface condensers to remove the steam. A liquid-ring compressor can be substituted for one or more eductors to conserve steam. Mechanical vacuum pumps are seldom used because of the corrosive nature of the off-gas. The noncondensible sour compressed gas and condensate are led away through a water seal for safety in case of steam system failure. The seal pot is equipped to skim off condensed oil continuously.

Lubricating Oil Vacuum Columns

These columns in principle are like crude columns, taking various lubricating oil stocks as side draws and using pumparound and water or air top condensing systems. The more recent installations use packing for internal condensing and rectification, for low pressure drop, and use in the order of 50 to 100 Torr as top column pressures. Vacuum systems are similar to those for fuels-type vacuum columns. Another feature is the reduced heat flux in the preheat furnace tubes, intended to reduce thermal degradation of products. Of course, cracking and color degradation can be corrected by hydrotreating. The distillates may be further treated by clay or acid treating, or by solvent extraction or hydrotreating to remove aromatic or asphaltic hydrocarbons. Normally the asphaltenes are found only in the vacuum bottom product. Wax is also removed by solvent crystallization and filtration, or by catalytic processes.

Usually, steam is added to the preheat system, either as velocity steam or after the preheater as dilution steam to get deeper volatilization. Each product is steam stripped as in crude oil distillation. In the case of naphthenic crude oils, the bottom product may be directed to asphalt specialties rather than being treated to make lubricating oil base stock.

RECOVERY AND PURIFICATION OF NORMALLY GASEOUS HYDROCARBONS

In the larger refineries, olefinic gas streams are recovered separately from paraffinic gases. The saturated gas plant collects its feed from crude oil distillation and from hydrogenation processes. The products might be hydrogen and methane, often fed to a hydrogen production reactor, ethane for ethylene plant feed, propane and butane for sale as fuel, and isobutane for alkylation plant feed to make high-octane gasoline.

The cracked gas plant, on the other hand, would likely produce a C_4 olefinic mixture for alkylation or chemical manufacture, a propylene product for the same purposes, a propane cut for fuel, and a C_2 fraction which is fed to chemical manufacture, or combined with hydrogen and methane used for hydrogen manufacture.

In a few modern plants, these separations are made by direct distillation, using refrigerated columns and condensers as used in ethylene plants. Most refiners, however, avoid deep refrigeration by absorbing the C_3^+ gases in naphtha or kerosene, using an arrangement such as that shown in Fig. 9. This arrangement is often seen in refineries designed by UOP. Operating pressures are fairly high, such as 200 psi, to assist absorption. The lean oil is sometimes refrigerated, but below about 50°F, hydrate plugging may occur.

The absorber and stripper shown in Fig. 9 can be combined into a single reboiled absorption column; however water condensed internally will cause bottom column corrosion and plugging, whereas in the scheme of Fig. 9, the water is removed at the feed drum. Tray efficiencies in absorption columns are much lower than in distillation columns because of the presence of noncondensible gases. Furthermore, there is a tendency to

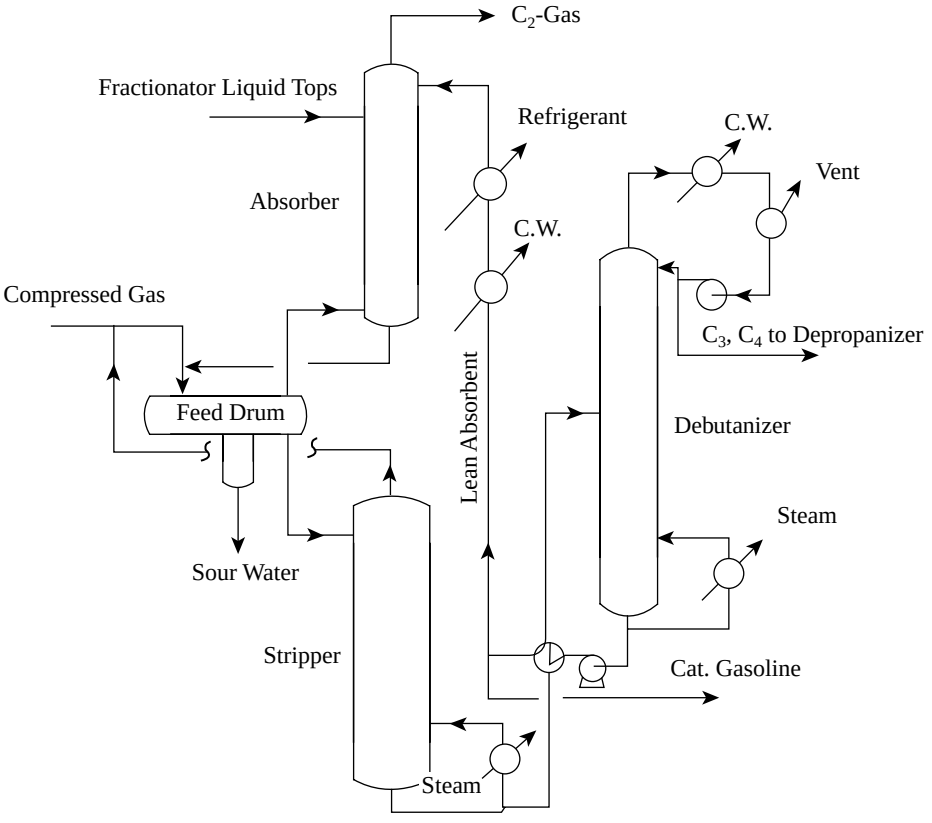


Fig. 9 Absorption-type de-ethanizing gas plant.

foaming in the upper part of the stripping section and in the absorber. System capacity factors as low as 0.6 or less should be expected.

In older refineries, high-pressure de-ethanizing distillation towers were often used, especially for straight-run crude oil distillation or thermal cracking plants.

Even at 500 psi plus, the critical properties of the ethane–ethylene and lighter top product made it difficult to maintain reflux without refrigeration. The advent of catalytic cracking and hydrocracking processes added large amounts of light gases, which made the absorption system of Fig. 9 more attractive.

HYDROPROCESS PRODUCT DISTILLATION

Most hydrotreating processes are equipped with reboiled product strippers to remove hydrogen and light ends produced in the reactions. These strippers are of normal design; however a few details may be noted. Because of the high bottom column temperatures required, fired-heater reboilers may be useful because they may allow vaporization with minimum exposure of bottom product to thermal degradation.

In the presence of hydrogen under pressure, the basic data systems such as SRK[6] and Peng-Robinson[7] fail to account properly for the presence of dissolved hydrogen in the liquid. Empirical systems based on Grayson–Streed[8] are often used for simulation.

Stripping columns may be equipped with a few rectifying trays and a small reflux flow to prevent loss of liquid product to the bulk hydrogen top gas. If the hydrotreater is upstream from a platinum-catalyst reforming unit, it may be advisable to reboil the stripper carefully to get a bone-dry feed to the catalytic reformer. In order to get good stripping and drying performance from the stripper, it may be advisable to preheat the stripper feed. In fact, preflashing the stripper feed is a method of increasing stripper capacity. Such flashing makes the rectifying trays particularly important.

SPECIAL SEPARATIONS

Solvents. Special boiling-range solvents are sometimes distilled from straight-run or other naphthas in the petroleum refinery. It may be tempting to prepare such solvents as part of the crude oil distillation process, but the small amounts usually required would be a costly nuisance to the very large crude oil distilling equipment. Hence smaller multi-plate columns are usually used. The design and modeling of these

columns is done using precision analytical distillation or gas chromatographic curves, splitting the curve into narrow boiling-point fractions similarly to the modeling of crude oil or cracked-product distillation.

Such solvents might be paint thinners, called Varnish Makers and Painters (V.M.&P.) naphthas or mineral spirits. There are many other specialty solvent fractions manufactured in petroleum refineries. The product specifications for these solvents as well as for many refinery products are often given in terms of the simple ASTM distillation boiling ranges. However, for best results in modeling, precision distillation analysis is used, with end-product correlation to compare with specifications. Such an approach is often used also for fuel fractions, even to the extent of controlling the operation with gas chromatographic precision distillation analysis.

Reforming and Isomerization Feedstocks. Some saturated fractions contain substantial amounts of low octane-number normal pentane or normal hexane. These chemical species can be isomerized for anti-knock performance enhancement. It is, however, useful to remove branched compounds from the feedstock by sharp distillation. A de-isohexanizer, for example, may need 60 or more actual valve plates and a high reflux ratio. The modeling of equipment for this purpose follows standard distillation design procedures, but because of the narrow relative volatilities, special care is needed with the basic data.

Alkylation-plant Feeds and Products. Separation between isobutane and normal butane for alkylation feed requires large plate counts, perhaps 60 or more, and high reflux ratios. Sometimes the butane splitter column is combined with the product stabilizer in a single large multi-product column.

Purification of Individual Molecular Species. The chemical process industries use large amounts of hydrocarbons, which can be individually recovered and purified from petroleum refinery fractions. Examples would include ethane, propene, butadiene, isoprene, benzene, toluene, and xylenes. Sometimes distillation must be supplemented with liquid–liquid extraction, as well as extractive and azeotropic distillation, to purify these materials. The technology of recovering individual chemical species will be dealt with in other parts of this encyclopedia.

BASIC DATA

The basic data used in modeling petroleum refinery distillations are usually obtained from proprietary systems of which many are marketed. For straight-run and paraffinic systems, Peng-Robinson[8] and Soave–Redlich–Kwong[7] systems are used to modify vapor–pressure and ideal system behavior. For the more aromatic cracked products, Chao–Seader[9] is often used. For species separations, a wealth of binary data and methods is available.

SAFETY

Injuries and damage are particularly to be guarded against. These occur in many ways; some of them are as follows:

1. When liquid water is mixed with hot oil, the resulting rapid evaporation may lead to violent vapor pressure surges. Such incidents with resulting equipment damage occur most frequently during shut-down or start-up procedures. The steam used for product stripping or for purging equipment may be the source of aqueous condensate, which may settle in unexpected places to cause trouble sooner or later. Extreme care is taken by refiners to avoid these incidents.
2. Combustible vapors allowed to mix with air may ignite violently. Prevention of these incidents is mostly obvious; however, undetected corrosion has sometimes led to vapor release with violent gas explosions. Vacuum distillation equipment must be protected against air intrusion, for example in case of steam supply failure.
3. Nitrogen or other inert gas, used for blanketing or purging, must be positively removed and equipment vented with air before personnel are allowed to enter equipment.
4. Warm vapors sometimes condense inside closed vessels; the resulting may collapse the vessel. Such incidents have occurred when a vessel was purged with steam, then closed in without venting. Similarly, a distillation column may go to vacuum if reboil fails but condensing continues.

CONCLUSIONS

Petroleum refinery distillation is like most chemical-plant distillation practice except for two differences: Firstly, the number of chemical species present is huge. Computer modeling of refinery distillation is often done by grouping components by boiling point to reduce the complexity of the calculation.

Secondly, the high-boiling hydrocarbons present in crude oil would make conventional reboilers subject to plugging with products from thermal degradation. To avoid this fouling, the feeds to distillation are often

partially prevaporized in fired heaters or other reactors at short residence time, followed by multi-product rectification towers.

Of course, the equipment sizes in refineries are very large.

The separation of C_1up to C_5 species is often done in complex absorption systems rather than by deep refrigeration.

ARTICLE OF FURTHER INTEREST

Bubble Cap Tray, p. 269.

REFERENCES

1. Townsend, D.W.; Linnhoff, B. Heat and power networks in process design: parts I and II. AIChE. J. **1983**, *29*, 742–748.
2. Bannon, R.P.; Marple, S. Heat recovery in hydrocarbon distillation. Chem. Eng. Progress, 1978, July, 41–45.
3. Watkins, R.N. *Petroleum Refinery Distillation*; Gulf Publishing Company: Houston, out of print.
4. Watkins, R.N. How to design crude distillation. Hydrocarbon Process. **1969**, *48* (12), 93–106.
5. Packie, J.W. Distillation equipment in the oil refining industry. AIChE. Trans. **1941**, *37*, 51–78.
6. Soave, G. Chem. Eng. Sci. **1972**, *27*, 1197.
7. Peng, D.Y.; Robinson, D.B. Ind. Eng. Chem. Fundam. **1976**, *15*, 59.
8. Grayson, H.G.; Streed, C.W. Vapor-liquid equilibria for high temperature, high pressure hydrogen-hydrocarbon systems, Proc. 6th World Petr. Congr., Frankfurt/Main III, Paper 20-PD7, 1963; 233–245.
9. Chao, K.C.; Seader, J.D. A general correlation of vapor-liquid equilibria in hydrocarbon mixtures. AIChE. J. **1961**, *7*, 598.

Phase Behavior of Hydrocarbon Mixtures

X.-Y. Zou
Oilphase-DBR, Schlumberger, Edmonton, Alberta, Canada

J. M. Shaw
Department of Chemical and Materials Engineering, University of Alberta, Edmonton, Alberta, Canada

INTRODUCTION

This contribution provides an introduction to the key building blocks, terminologies, and presentation methods used for describing, modeling, and discussing the phase behavior of hydrocarbon mixtures. The mechanics related to interpreting and constructing phase diagrams and phase behavior projections are reviewed. Solidification of components and the impact of systematic variations in the constituents present on observed phase behaviors are addressed explicitly. Solidification adds to the complexity of phase diagrams as it can also interfere with critical phenomena and suppress multiphase behavior that would otherwise be present. A more detailed discussion of phase diagrams for single component and binary systems can be found elsewhere.[1] While simple hydrocarbon mixtures are the focus of this contribution, the concepts illustrated provide insights for understanding the phase behavior of industrial fluids such as waxy and asphaltene rich hydrocarbon mixtures that pose a host of challenges for the petroleum industry.

OBSERVED PHASE BEHAVIORS

In addition to vapor (V), high-density liquid (L_2), or low-density liquid (L_1) phase behavior, reservoir fluids, oils, and other organic fluids exhibit a variety of multiphase behaviors and critical phenomena as noted in Fig. 1. These include liquid–vapor (L_1V or L_2V), liquid–liquid (L_1L_2), and liquid–liquid–vapor (L_1L_2V) phase behavior and associated critical phenomena. The designations $L = V$ or $L_1 = L_2$ mean that the two phases in question are critically identical. They possess the same values for density, composition, molar volume, viscosity, and all other physical properties, and the boundary between the phases disappears (Fig. 1). Two phases can also become critically identical in the presence of a third phase, giving rise to the so-called K and L points. A K point arises when a low-density liquid (L_1) becomes critically identical to a vapor in the presence of a high-density liquid (L_2). This is designated as $L_1 = V + L_2$. An L point arises when a low-density liquid (L_1) and a high-density liquid (L_2) become critically identical in the presence of a vapor phase. An L point is designated as $L_1 = L_2 + V$. Tricritical points, where three phases in equilibrium are also critically identical, are designated as $L_1 = L_2 = V$. Such critical points, while present in phase diagrams and phase projections, are rarely observed in practice. At low temperatures, solid phases such as asphaltenes and wax can, and frequently do, coexist with the fluid phases noted here and are discussed in later sections.

THE PHASE RULE

The maximum possible number of phases present at equilibrium for a given number of components is governed by the phase rule. This rule, first derived in its simplest form by Gibbs over a century ago, has been augmented over time to allow for magnetic and other field effects that increase the number of degrees of freedom (where relevant) and critical phenomena that reduce the number of degrees of freedom. In petroleum science and engineering, constraints imposed by critical phenomena are important and the relevant phase rule has become:

$$F = N - \pi + 2 \quad \text{for } k = 0 \tag{1a}$$

$$F = N - \pi + 2 - (k - 1) \quad \text{for } k = 2, 3, \ldots \tag{1b}$$

where F is the number of degrees of freedom, N is the number of components, π is the number of phases, and k is the number of phases that are critically identical. Thus, for a binary system, two noncritical phases may coexist over a range of compositions and pressures at fixed temperature, or over a range of temperatures and compositions at fixed pressure. In contrast, a binary system may exhibit two critical phases at specific compositions along a curve in pressure–temperature space as there is only one degree of freedom, i.e., once

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120040185

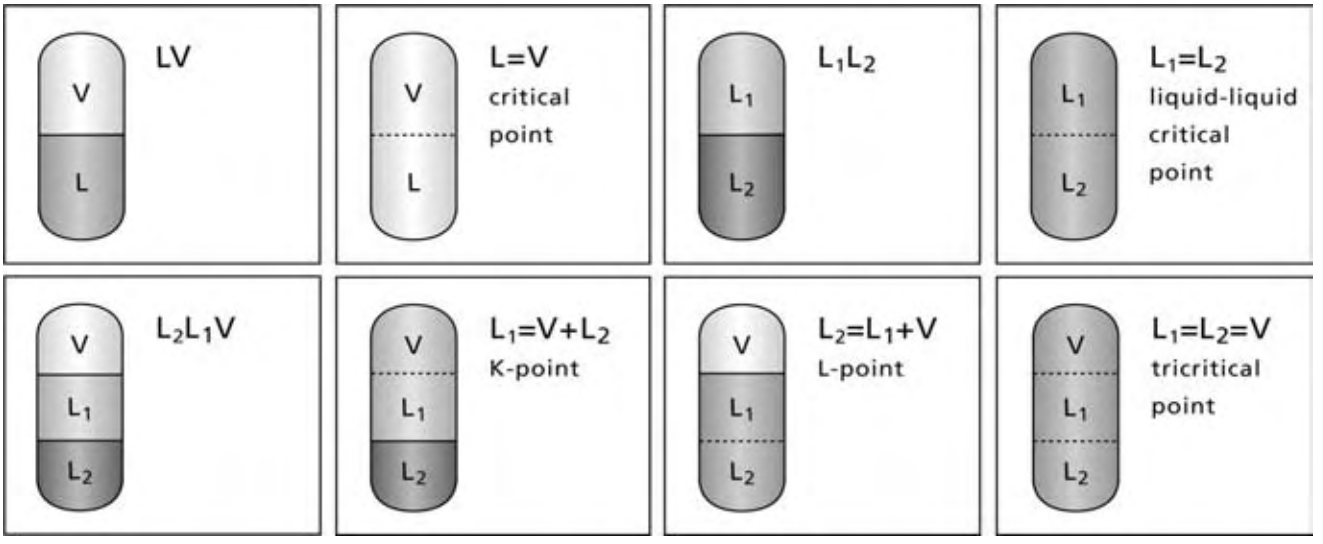


Fig. 1 Multiphase, fluid-only behaviors exhibited by hydrocarbon mixtures. Dashed phase boundaries indicate critically identical phases.

the composition is fixed the critical pressure and temperature are fixed as well. This curve is called a critical locus. Finally, binary mixtures may exhibit up to four noncritical phases in equilibrium, or up to three phases in equilibrium, when two of the phases are critically identical. Given the number and the diversity of components present in hydrocarbon mixtures, multiphase behavior is to be expected. For example, even the mixture methane + hexane exhibits L_1L_2V phase behavior.[2]

PROJECTIONS AND PHASE DIAGRAMS

Sets of observed phase behaviors are often presented as phase diagrams where pressure and composition are varied at fixed temperature (P–x diagrams), or temperature and composition are varied at fixed pressure (T–x diagrams), or pressure and temperature are varied at fixed composition (P–T diagrams). In order to present the phase behavior of binary mixtures, large numbers of these diagrams are required. Projections that capture key features of the phase behavior of a binary mixture, such as the vapor pressure curves of the pure components, critical phenomena, and multiphase behavior complement the more detailed diagrams regardless of composition, as one can sketch P–x and T–x phase diagrams from P–T projections. Two illustrative pressure–temperature projections, where the composition axis is collapsed into the pressure–temperature plane, are shown in Figs. 2 and 3 along with example P–x and T–x phase diagram constructions. The projection shown in Fig. 2, known as a Type I projection according to van Konynenburg and Scott[3] who proposed the first naming scheme for projections, is the simplest of the six possible binary projections. Mixtures such as methane + ethane or other pairs of similar compounds exhibit Type I phase behavior. The Type IV projection, shown in Fig. 3, captures key aspects of the complex phase behaviors exhibited by mixtures where substantial differences in molecular size, polarity, etc. result in only partial miscibility of the components—Type II through Type VI. For example, the carbon dioxide + tridecane binary exhibits Type IV phase behavior.[4]

HOW TO READ A P–T PHASE BEHAVIOR PROJECTION

In P–T projections, the composition axis is collapsed into the pressure–temperature plane. The vapor pressure curve for component A is labeled LV(A) and that for component B is labeled LV(B). These curves terminate at the component critical points (L = V) designated as hollow circles. In Fig. 2, dew pressure and bubble pressure curves for an intermediate composition "x" intersect at a point on the (L = V) critical locus where the liquid and vapor phases become critically identical. Normally, dew and bubble pressure curves are not shown in projections. They are shown here so that the construction of the related P–x at fixed T, and T–x at fixed P, phase diagrams is clearly illustrated. Each critical point on the critical locus corresponds to a fixed composition. Points close to the critical point of component A are critical points for mixtures with high concentrations of A, whereas points closer to the critical point of

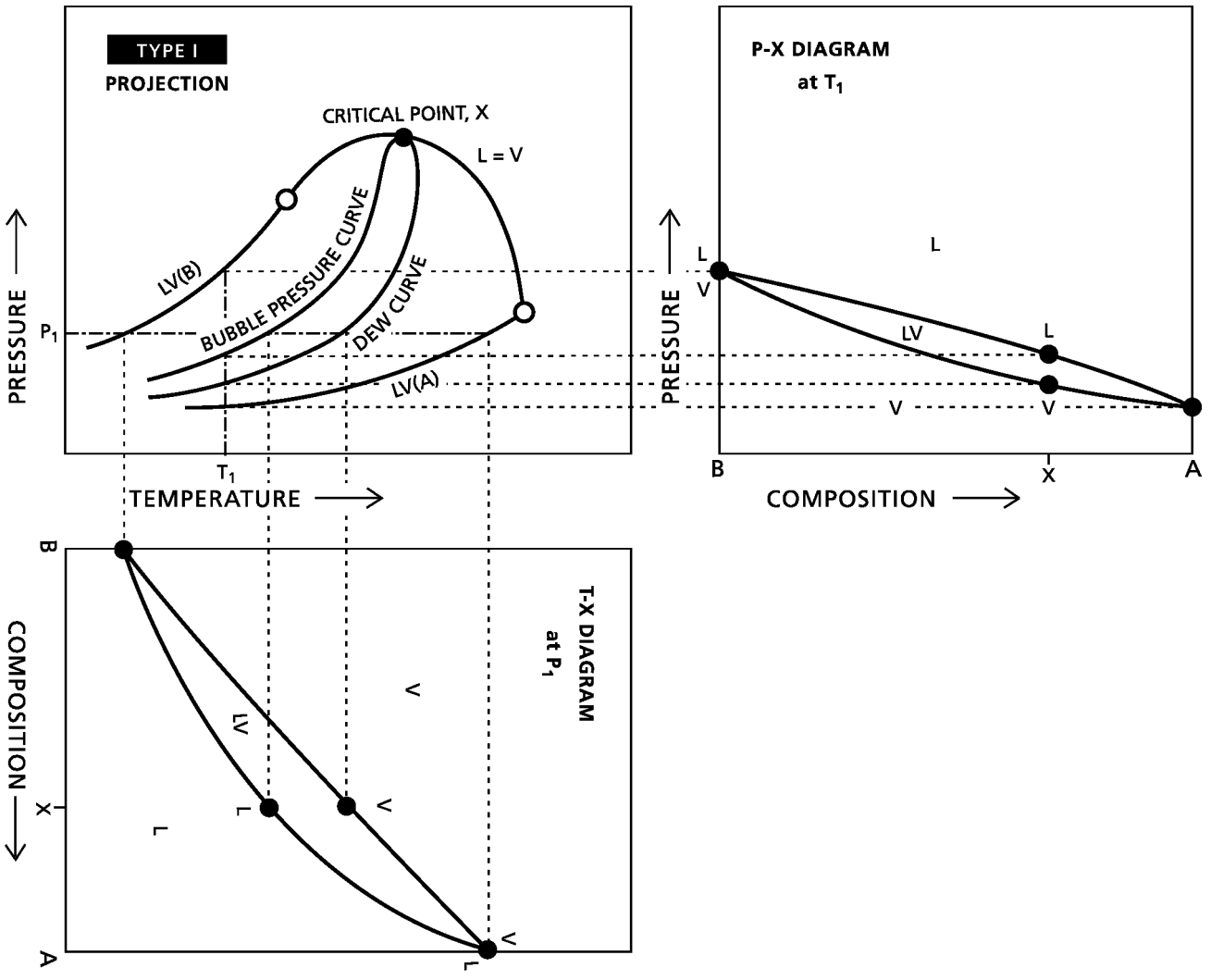


Fig. 2 Pressure–temperature phase projection and examples of pressure–composition and temperature–composition phase diagrams for a Type I binary mixture.

component B correspond to critical points for mixtures with high concentrations of B. Normally, the critical pressures for intermediate compositions are high relative to the component critical pressures.

CONSTRUCTION OF PHASE DIAGRAMS FROM PROJECTIONS AND EXPERIMENTAL OBSERVATIONS

With reference to the pressure–temperature projection in Fig. 2, we see that in order to construct a P–x diagram at T_1, a line parallel to the pressure axis passing from low pressure to high pressure is drawn and the points of intersection with features in the projection are noted. For the P–x at T_1 example, the construction line intersects the vapor pressure curve for pure A, the dew and bubble pressure curves for the mixture with composition x, and the vapor pressure curve for pure B. Each of these points is transcribed onto the P–x diagram and labeled accordingly. One then connects the dots! The points labeled V are connected one to the other sequentially in composition, as are the ones labeled L. As expected, all mixtures of A and B are a single-phase liquid, at high pressure, and a single-phase vapor at low pressure. A lens of liquid–vapor phase behavior arises at intermediate pressure. The boundary between the L and the LV regions is typically referred to as the bubble pressure curve, and the boundary between the LV and the V region is referred to as the dew pressure curve. A T–x at P_1 diagram is constructed in an analogous manner except that the line of construction is drawn parallel to the temperature axis.

In the Type IV pressure temperature projection illustrated in Fig. 3, the binary mixture exhibits partial miscibility at low temperature and near the critical

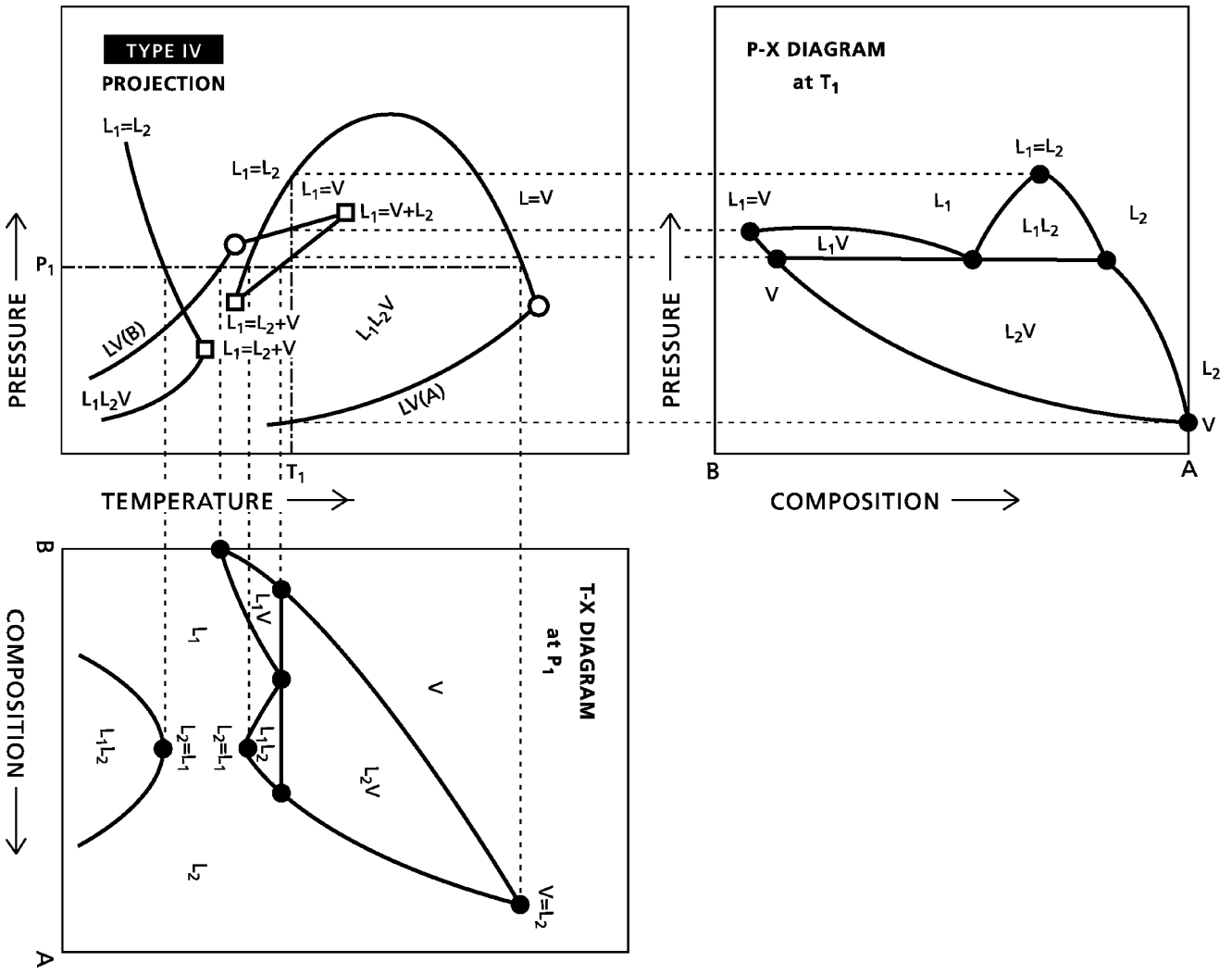


Fig. 3 Pressure–temperature phase projection and examples of pressure–composition and temperature–composition phase diagrams for a Type IV binary mixture.

region of component B. L_1L_2 and L_1L_2V phase behavior are observed in both regions in addition to the simpler V, L_2V, L_2, L_1, L_1V phase behaviors. The approach to P–x and T–x phase diagram construction is the same as for the Type I phase behavior case above. Let us construct the T–x diagram at P_1. The first curve encountered by the construction lines is an $L_1 = L_2$ critical locus. At temperatures below the point of intersection, the components A and B are not miscible. At higher temperatures they are mutually soluble. The vapor pressure curve for component B is encountered next followed by a second L_1L_2 critical locus that marks the beginning of the high temperature partial miscibility zone at P_1. The three-phase L_1L_2V line is then encountered. This is shown on the T–x diagram by the three dots for the compositions of V, L_1, and L_2 where the vapor comprises mostly B, the more volatile component, L_2 comprises mostly A, the least volatile component, and L_1 possesses an intermediate composition. The three dots are connected by a line to show that the three phases are in equilibrium, and to facilitate the construction of the phase diagram, as two and one phase behaviors surround the three-phase behavior. Finally, the $L = V$ locus is encountered at a temperature and a pressure close to the critical point of pure A. Again, the dots are connected sequentially in composition for the various phases and the T–x phase diagram emerges. Had other temperatures or pressures been chosen for the construction of the P–x and T–x phase diagrams, the phase behaviors encountered by varying temperature at fixed pressure or varying pressure at fixed temperature would have differed, as would the resulting phase diagrams, even though the underlying P–T projections are invariant.

Projections place phase diagrams or phase behavior observations in context and provide a clear basis for

the construction and verification of computational phase behavior models. For example, the set of observations obtained by heating a mixture comprising half A and half B at fixed pressure (P_1) in Fig. 3:

$$L_1L_2 \rightarrow L_2 \rightarrow L_1L_2 \rightarrow L_1L_2V \rightarrow L_2V \rightarrow V$$

is consistent with Type IV phase behavior and one would seek out critical phenomena [$L_1 = L_2$, $L_1 = V$, $L_2 = V$, $L_1 = L_2 + V$ (two points), $L_1 = V + L_2$], either computationally or experimentally, in order to define the phase behavior of the mixture. Other observation sets such as $L \rightarrow LV \rightarrow V$, obtained by heating an alternative composition (close to either pure "A" or pure "B") at the same fixed pressure (P_1), provide no cues for identifying the phase behavior of the mixture as a whole because they arise in all phase behavior types over ranges of compositions.

THE IMPACT OF SOLIDIFICATION ON OBSERVED PHASE BEHAVIORS (BINARY MIXTURES)

Generally, solid phases appear at temperatures below the melting temperature of one or both of the components. Solidification has a dramatic impact on the appearance of phase projections and diagrams.[4,10–12] For example, Figs. 4A and 4B illustrate the impact of solidification on the appearance of a Type I binary projection, i.e., the simplest binary case. In Fig. 4A, solid–liquid–vapor curves (S_ALV, S_BLV) emerge from the triple points of components A and B, respectively, and converge at a four-phase eutectic point where S_AS_BLV phase behavior arises. If the triple point of component A occurs at a temperature well above the critical temperature of component B, the S_ALV curve intersects the $L = V$ critical locus and the projection shown in Fig. 4B results. While the projections share a eutectic at low temperature, they are quite different at high temperature. For a Type IV binary, four projections, illustrated in Fig. 5A–D, arise depending on the respective values of the triple point temperatures and pressures of the components. In addition to four-phase eutectics (S_AS_BLV), four-phase ($S_AL_1L_2V$) points, referred to as Q points, also arise. L_1L_2V phase behavior, a characteristic of Type IV phase behavior, can be suppressed in whole or in part by solidification if the triple point temperature of component A is sufficiently high, a case that is treated elsewhere in detail.[11] The resulting projection, Fig. 5D, is identical to one of the possible projections for a Type I binary–Fig. 4B. Projections shown in Figs. 5C and 5D also correspond to those expected for a Type III binary, where in the absence of solidification L_1L_2V, phase behavior extends from the critical region of component B to low temperatures. It is frequently necessary to compare phase behavior observations for a homologous series of solvents (B_i) with fixed solute (A), or a homologous series of solutes (A_i) with fixed solute (B), in order to place observations in context.

THE IMPACT OF ASYMMETRY ON OBSERVED PHASE BEHAVIOR (BINARY MIXTURES)

The phase behavior exhibited by a mixture is related to the degree of asymmetry of the components, which is a measure of differences in size, structure, and polarity among components–the greater the differences, the greater the degree of asymmetry.[4,10,13] Raeissi et al.[13] highlight the role of asymmetry on the nature of the phase behavior exhibited by binary and pseudobinary mixtures of light gases (e.g., carbon dioxide, methane, propane) + members of various homologous series (alkyl-benzenes, *n*-alkanes, *n*-alcohols). Light gases tend to be miscible with the smallest members of homologous series (Type I). One observed sequence is that as the degree of asymmetry increases, the mixtures become only partially miscible at low temperatures where L_1L_2V curves appear in the pressure–temperature projections, but the mixtures remain miscible at elevated temperatures (Type II). As the degree of asymmetry is increased further, a second miscibility gap appears in the near critical region of the lighter component and one finds L_1L_2V phase behavior at both low and high temperatures. At intermediate temperatures, the mixtures remain miscible. This is referred to as Type IV phase behavior. If the degree of asymmetry is increased still further, the miscibility gaps at high and low temperature merge resulting in a single L_1L_2V curve stretching from low to high temperatures (Type III). For the carbon dioxide + *n*-alkanes series, the P–T projection for octane (C_8H_{16}) through dodecane ($C_{12}H_{26}$) is Type II, with tridecane ($C_{13}H_{28}$) Type IV (Fig. 5C), with tetradecane ($C_{14}H_{30}$) Type III (as in Fig. 5D). Other sequences of phase behavior transitions are observed for the other cases cited.

PHASE BEHAVIOR TRANSITIONS

Measurement and prediction of transitions from one type of phase behavior to another are key to our understanding of the physical properties of hydrocarbon mixtures. One can examine phase behavior type transitions for binary mixtures from the perspective of the solute or the solvent while varying the other. Both approaches are found in the literature. The solvent fixed approach is shown in Fig. 5 for carbon dioxide + *n*-alkane binary mixtures.[4] The anthracene

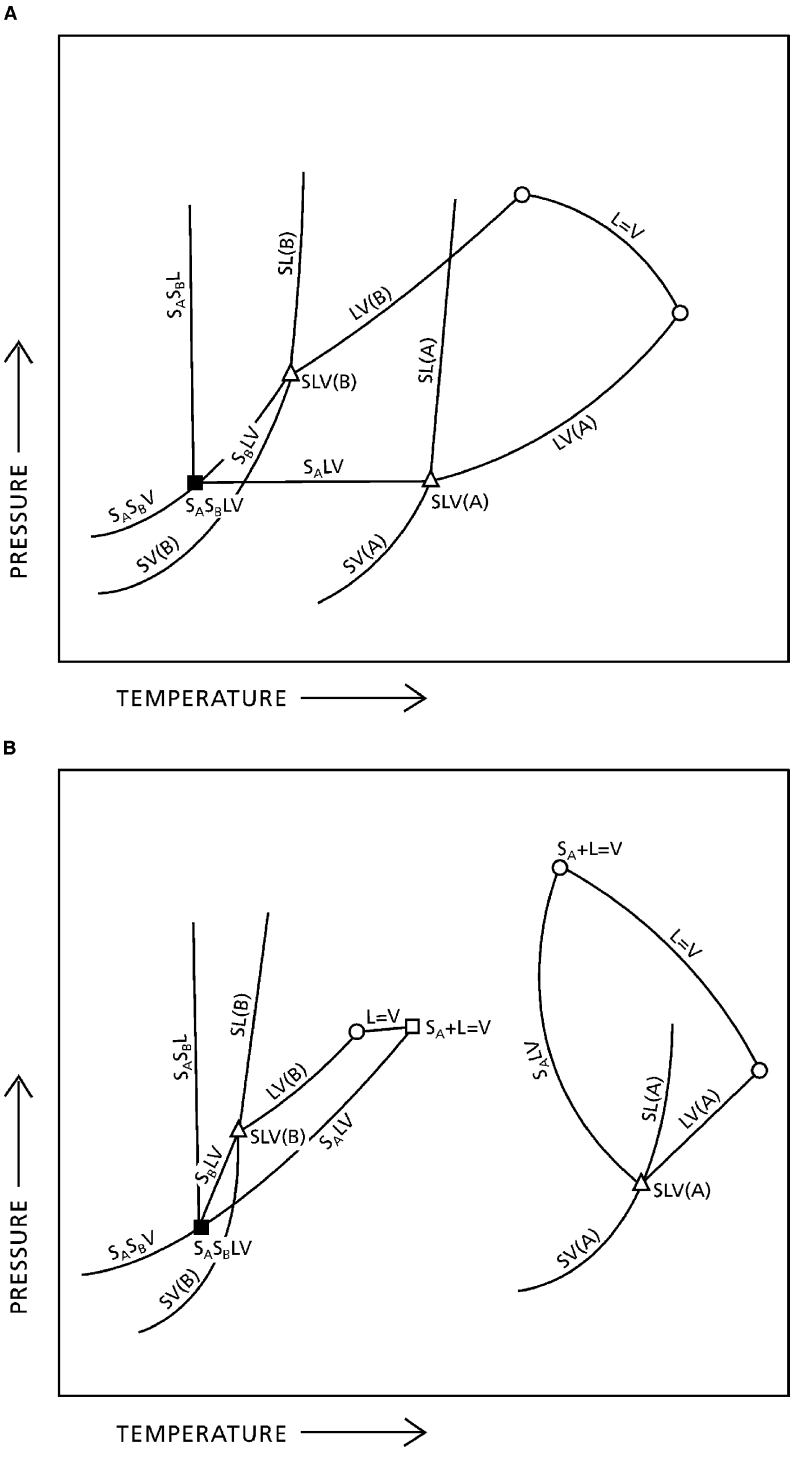


Fig. 4 Possible pressure–temperature projections for Type I binaries including solids.

(solute) + *n*-alkane (solvent) binary and pseudobinary case (Fig. 6) illustrates the latter approach.[9] Transitions from Type III (projection as in Figs. 4B and 5D) to Type IV (projection as in 5B) to Type V (Fig. 5C) to Type I (Fig. 4A) arise over the interval from anthracene + butane to anthracene + hexane.

Consequently, dramatic changes in phase behavior, with slight changes in either solvent or solute composition, arise commonly for hydrocarbon mixtures. Multiphase behavior can be introduced or suppressed as the examples illustrate. Such sharp transitions also pose a challenge for modeling as small errors in fluid phase behavior prediction, or the location of Q points, can yield phase behaviors and phase behavior transition sequences incompatible with experimental data.[9,11]

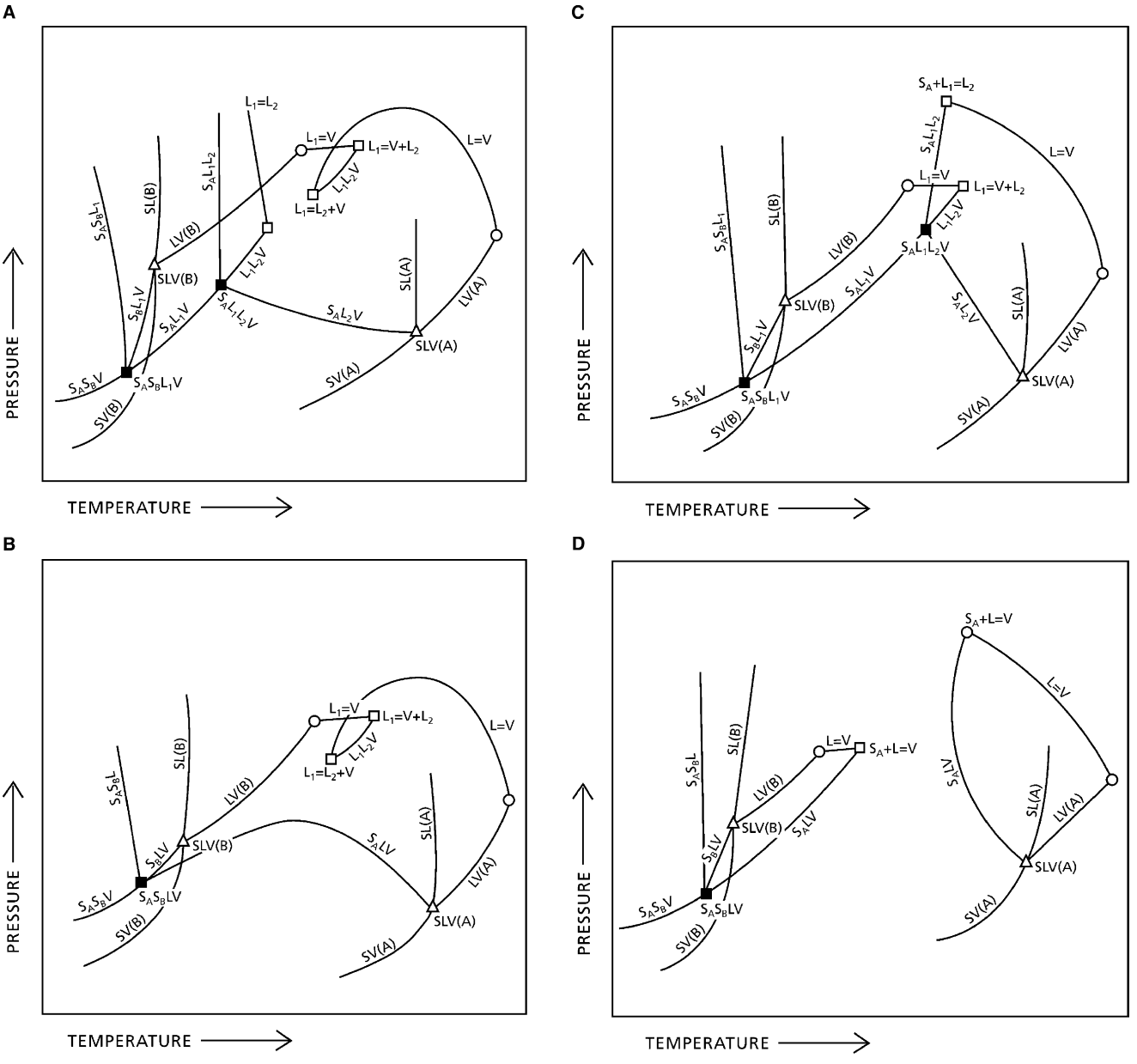


Fig. 5 Possible pressure–temperature projections for Type IV binaries including solids where the melting point of component A increases from A to D.

IMPLICATIONS FOR MULTICOMPONENT MIXTURES

Phase diagrams for multicomponent mixtures possess additional degrees of freedom and are inherently multidimensional. In practice, construction and interpretation of phase diagrams of multicomponent mixtures are similar to, and based on, those of binary mixtures.[14–17] The phase behavior of multicomponent mixtures can also be depicted as sections in PTx_1x_2-space, keeping one or more of the variables constant. A widely used section for ternary mixtures is an equilateral triangle composition diagram at fixed pressure and temperature (Fig. 7). Three vertexes (A, B, and C) represent the three pure components, respectively; a point on a side (AB, BC, or AC) represents a binary mixture (A + B, B + C, or A + C); a point within the triangle represents a ternary mixture (A + B + C). The region enclosed by triangle DEF (dash dot line) represents a three-phase zone. Points within the zone represent compositions of heterogeneous mixtures of three phases ($\alpha + \beta + \gamma$); the compositions of these three phases are fixed and represented by D, E, and F, respectively. The relative amounts of coexisting phases are determined from the component mass balances. As Fig. 7 illustrates, phase behaviors not present in binaries can arise within the body of the diagram

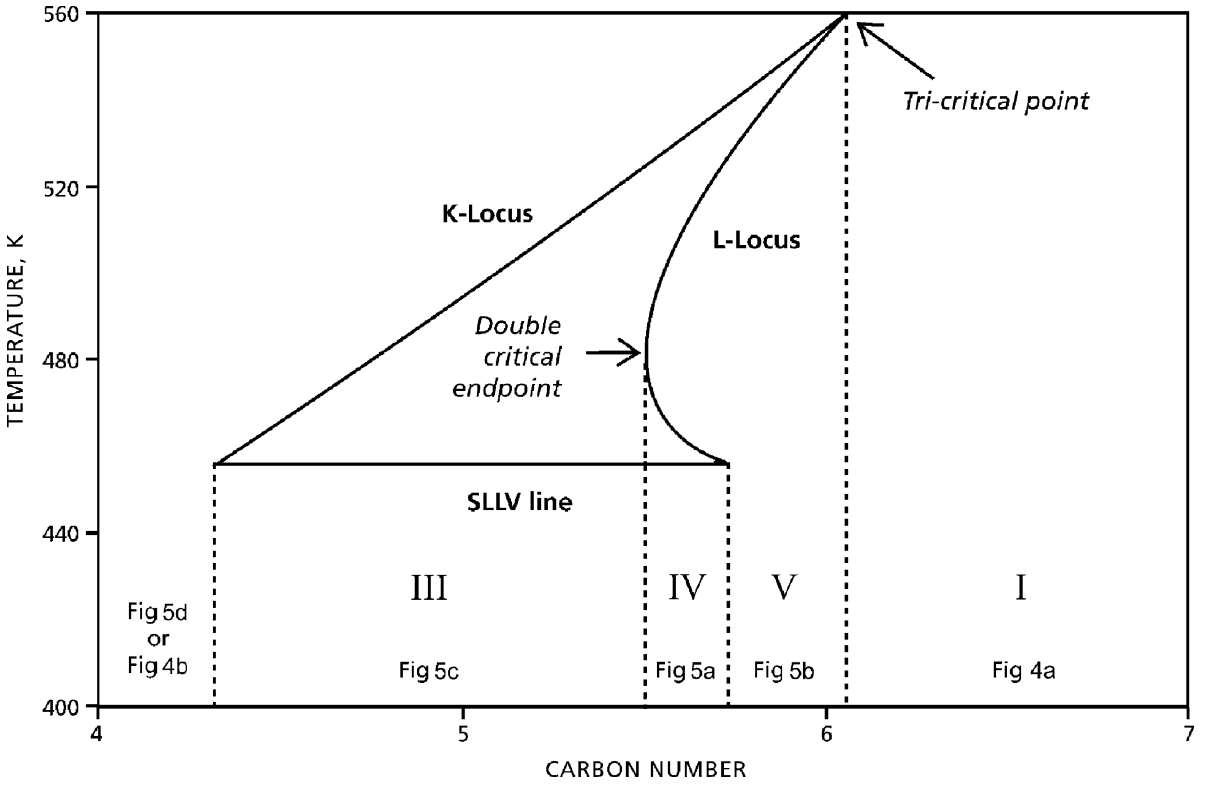


Fig. 6 Phase behavior transitions for the anthracene (solute) + *n*-alkane (solvent) system.

due to the additional degrees of freedom. For example, Heidemann and Abdel-Ghani[18] observed five co-existing phases (one vapor phase + four liquid phases) in phenol–water–hexane ternary mixtures whereas the constituent binaries exhibit a maximum of three phases in equilibrium! Stacks of triangular diagrams (prisms) are used to show the effect of pressure or temperature on the phase behavior of ternary mixtures.[19] Pyramids are used to describe the phase behavior of quaternary mixtures at fixed temperature and pressure where the vertices represent the four components; points on an edge represent binary mixtures; points on the triangular surfaces represent ternary mixtures; points within the pyramid represent quaternary mixtures.[20] Recently, Shaw[21] discussed the use of pyramid diagrams to describe the phase behavior of complex hydrocarbon mixtures.

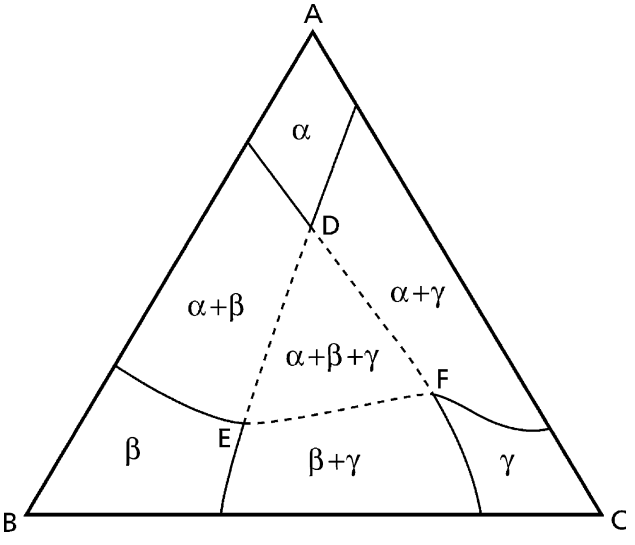


Fig. 7 An example of an equilateral triangle composition diagram for a ternary mixture at fixed pressure and temperature.

PHASE BEHAVIOR PREDICTION

The theory and conditions for phase equilibrium are well established. If more than one phase is present, then the chemical potential of a component is the same in all phases present. As chemical potential is linked functionally to the concepts of fugacity and activity, models for phase behavior prediction and correlation based on chemical potentials, fugacities, and activities have been developed. Historically, phase equilibrium calculations for hydrocarbon mixtures have been fragmented with liquid–vapor, liquid–liquid, and other phase equilibrium calculations, subject to separate and diverse treatments depending on the temperature, pressure, and component properties. Many of these methods and approaches arose to meet specific needs in the chemical process industries. Poling, Prausnitz,

O'Connell[24] provide an excellent survey of computation methods and restrictions.

From the 1980s onward, equation of state based models and cubic equation of state based computational methods in particular have come to dominate phase equilibrium calculations for hydrocarbon mixtures and are particularly useful for phase diagram and phase projection construction. van der Waals equation of state, for example, can be used to predict five of the six phase behavior types expected for binary mixtures including hydrocarbons.[3] The computations are intensive and involve minimizing Gibbs free energy for a mixture as a whole. Most algorithms employ Michelsen's tangent plane criterion[5,6] for this purpose. A large number of computational routines and options are now built into commercial process simulators. While these computations are frequently reliable for well-defined mixtures, one must be wary of the results obtained because, in general, errors can be substantial. This is particularly so at higher pressures where the number, nature, and composition of phases present at fixed composition can be predicted incorrectly,[7,8] or where the wrong phase behavior type can be predicted for a mixture as a whole.[9] Some equations of state also predict phase behaviors that do not arise in nature. Custom regression of interaction parameters based on (some) experimental data, and careful attention to the numerical methods employed, mitigate, but do not eliminate, these difficulties even for simple mixtures.

Pure solid + fluid phase equilibrium calculations are challenging but can, in principle, be modeled if the triple point of the pure solid and the enthalpy of fusion are known, the physical state of the solid does not change with temperature and pressure, and a chemical potential model (or equivalent), with known coefficients, for solid constituents is available. These conditions are rarely met even for simple mixtures and it is difficult to generalize multiphase behavior prediction results involving even well-defined solids.[23] The presence of polymorphs, solid–solid transitions, and solid compounds provide additional modeling challenges, for example, ice, gas hydrates, and solid hydrocarbons all have multiple forms.

IMPLICATIONS FOR HYDROCARBONS OF INDUSTRIAL INTEREST

Reservoir fluids and heavy oil + light hydrocarbon mixtures comprise numerous components from methane to asphaltenes and are often asymmetric in nature. These diverse constituents are difficult to identify at a molecular level[22] and are typically lumped into pseudocomponents. The reservoir fluid and heavy oil literatures diverge on how the lumping should be performed.[21] For reservoir fluids, the typical scheme is to treat low molar mass species individually, and intermediate hydrocarbons as pseudocomponents based on carbon number. For example, an ill-defined maltene fraction comprises one or more pseudocomponents and asphaltenes one or two others. For bitumen and heavy oil mixtures, the light components are typically ignored and one encounters pseudobinary, and more typically ternary, diagrams where the composition variables are asphaltenes, and one or two hydrocarbon fractions. Thus, the impact of the variability of the phase behavior of the high molar mass constituents on the phase behavior of reservoir fluids is largely ignored in one literature and the impact of low molar mass constituents on the phase behavior of heavy feedstocks is largely ignored in the other literature. The phase behavior of hydrocarbon mixtures is too sensitive for such an approach to be successful from a modeling perspective. As a consequence, the phase behavior of individual fluids must be identified primarily from experiments in order to obtain reliable information.

CONCLUSIONS

This contribution introduces phase behavior concepts, such as multiphase equilibria, critical phenomenon, and the phase rule that lay a foundation for understanding the phase behavior of hydrocarbon mixtures. The mechanics related to interpreting and constructing phase diagrams and projections are reviewed. The phase behavior of hydrocarbon mixtures is complex but falls within established qualitative frameworks that build on the work of van Konynenburg and Scott. Dramatic changes in phase behavior, with slight changes in either solvent or solute composition, characterize the phase behavior of asymmetric hydrocarbon mixtures. Solidification adds to the complexity of the phase behaviors exhibited through the introduction or the suppression of multiphase behavior. Regardless of their source, sharp phase behavior transitions pose challenges for process design and operation and hamper modeling efforts because small errors in fluid phase behavior prediction, or in the location of Q points, can yield phase behaviors and phase behavior transition sequences incompatible with experimental data, even for well-defined mixtures. Phase behaviors of multicomponent and ill-defined mixtures present additional challenges as phase behaviors not present in the constituent binary or ternary diagrams can arise. Composition variations within coarsely lumped fractions also influence phase behavior. The impact of such variations is not captured in existing databases and, as a consequence, the phase behavior of fluids of industrial interest is still approached primarily from experiment.

ACKNOWLEDGMENTS

We gratefully acknowledge the sponsors of the NSERC Industrial Research Chair in Petroleum Thermodynamics (Alberta Energy Research Institute, Albian Sands Energy Inc., Computer Modelling Group Ltd., ConocoPhillips Inc., Imperial Oil Resources, NEXEN Inc., Natural Resources Canada, Petroleum Society of the CIMM, Oilphase-DBR, Oilphase—a Schlumberger Company, Schlumberger, Syncrude Canada Ltd., NSERC) for their financial support.

REFERENCES

1. de Swaan Arons, J.; de Loos, Th. Phase behavior: phenomena, significance and models. In *Models for Thermodynamic and Phase Equilibria Calculations*; Sandler, S.I., Ed.; Marcel Dekker Inc.: New York, 1994; (chapter 5) 363–506.
2. Chen, R.J.J.; Chappelear, P.S.; Kobayashi, R. Dew-point loci for methane–hexane and methane–heptane binary systems. J. Chem. Eng. Data **1976**, *21* (2), 213–219.
3. van Konynenburgh, P.H.; Scott, R.L. Critical lines and phase equilibria in binary van der Waals mixtures. Philos. Trans. R. Soc. London **1980**, *298*, 495–540.
4. Miller, M.M.; Luks, K.D. Observations on the multiphase equilibria behavior of CO_2-rich and ethane-rich mixtures. Fluid Phase Equilibria **1989**, *44*, 295–304.
5. Michelsen, M.L. The isothermal flash problem. Part I: stability. Fluid Phase Equilibria **1982a**, *9*, 1–19.
6. Michelsen, M.L. The isothermal flash problem. Part II: phase split calculation. Fluid Phase Equilibria **1982a**, *9*, 21–40.
7. Baker, L.E.; Pierce, A.C.; Luks, C.D. Gibbs energy analysis of phase equilibria. SPE J. **1982**, *22*, 731–742.
8. Cartlidge, C.R.; Dukhedin-Lalla, L.; Rahimi, P.; Shaw, J.M. Preliminary phase diagrams for ABVB + dodecane + hydrogen. Fluid Phase Equilibria **1996**, *117*, 257–264.
9. Minicucci, D.; Zou, X.-Y.; Shaw, J.M. The impact of liquid–liquid–vapour phase behaviour on coke formation from model coke precursors. Fluid Phase Equilibria **2002**, 194–197, 353–360.
10. Peters, C.J. Multiphase equilibria in near-critical solvent. In *Supercritical Fluids: Fundamentals for Application*; Kiran, E., Levelt Sengers, J.M.H., Eds.; NATO ASI Series, Series E: Applied Sciences Vol. 273; Kluwer Academic Publishers, 1994; 117–145.
11. Wisniak, J.; Apelblat, A.; Segura, H. Prediction of gas–solid equilibrium using equations of state. Fluid Phase Equilibria **1998**, *147*, 45–64.
12. Poot, W.; Krüger, K.-M.; de Loos, T.W. High-pressure phase behavior of the binary systems (butane + adamantane) and (butane + diamantane). J. Chem. Thermodyn. **2003**, *35*, 583–596.
13. Raeissi, S.; Gauter, K.; Peters, C.J. Fluid multiphase behavior in quasi-binary mixtures of carbon dioxide and certain 1-alkanols. Fluid Phase Equilibria **1998**, *147*, 239–249.
14. Gregorowicz, J.; de Loos, Th.W.; de Swaan Arons, J. Unusual retrograde condensation in ternary hydrocarbon systems. Fluid Phase Equilibria **1992**, *73*, 109–115.
15. Gregorowicz, J.; de Loos, Th.W.; de Swaan Arons, J. Liquid–liquid–vapour phase equilibria in the system ethane + propane + eicosane: retrograde behavior of the heavy liquid phase. Fluid Phase Equilibria **1993**, *84*, 225–250.
16. Shaw, J.M.; de Loos, Th.W.; de Swaan Arons, J. Prediction of unusual retrograde condensation in model reservoir fluids. Fluid Phase Equilibria **1993**, *84*, 251–266.
17. Shaw, J.M.; de Loos, Th.W.; de Swaan Arons, J. An explanation for solid–liquid–liquid–vapor phase behavior in reservoir fluids. Petrol. Sci. Technol. **1997**, *15* (5&6), 503–521.
18. Heidemann, R.A.; Abdel-Ghani, R.M. A ternary system with five equilibrium phases. Chem. Eng. Sci. **2001**, *56* (24), 6873–6881.
19. de Swaan Arons, J.; de Loos, Th.W. Phase behaviour: phenomena, significance, and models. In *Models for Thermodynamic and Phase Equilibria Calculations*; Sandler, S.I., Ed.; Marcel Dekker, Inc., 1993; 363–506.
20. Ruiz, R.; Marcilla, A. Letter to the editor: phase behavior in liquid–liquid–solid quaternary systems. J. Chem. Eng. Data **1997**, *42*, 409–411.
21. Shaw, J.M. Towards common generalized phase diagrams for asphaltene containing hydrocarbon fluids. ACS, Petrol. Div. Prepr. **2002**, *47* (4), 338–342.
22. Hughey, C.A.; Rodgers, R.P.; Marshall, A.G. Resolution of 11000 compositionally distinct components in a single electrospray ionization fourrier transform ion cyclotron resonance mass spectrum of crude oil. Anal. Chem. **2002**, *74*, 4145–4149.
23. Shaw, J.M.; Béhar, E. SLLV phase behavior and phase diagram transitions in asymmetric hydrocarbon fluids. Fluid Phase Equilibria **2003**, *209*, 185–206.
24. Poling, B.E.; Prausnitz, J.M.; O'Connell, J.P. *The Properties of Gases and Liquids*, 5th Ed.; chapter 8, 2000.

Phase Equilibria

Karl B. Schnelle, Jr.
Department of Chemical Engineering, Vanderbilt University, Nashville, Tennessee, U.S.A.

INTRODUCTION

Our lives are intrinsically linked with mixtures. As scientists and engineers involved with the chemical processing industry, we are well aware of the nature of air and gasoline as mixtures. We must deal with these mixtures in our daily activities and professionally in our work as scientists and engineers. We are also concerned with the transfer of components to and from these mixtures to other mixtures. Oxygen from the air will transfer to the waters of our rivers and lakes, dissolving in them and providing a healthy mixture with the water to support aquatic life. If this does not happen, the waters may be of such a composition so as not to support the aquatic life and become polluted. If we are not aware of the nature of the combustion of gasoline, it is also possible for carbon monoxide formed in the combustion and entering the air to be transferred and dissolved in our blood stream thereby causing a toxic reaction and even death. The important thing here is the composition of the two phases, i.e., the air–water and air–blood as gasses and liquids where material is transferred from the one phase to the other. The transfer process continues from the one phase to the other until each phase reaches a constant value, which is said to be the equilibrium condition. In this condition, the composition of the components in each phase may be very different in from each other. When we separate the components in these mixtures by the processing methods of absorption, adsorption, distillation, extraction, etc., we need to have this difference established.

Phase equilibrium is the science that allows us to determine the physical properties of the phases that exist at this equilibrium condition. In general, phase equilibria deals with all combinations of vapor, liquid, and solid phases. This chapter will present the basic concepts and models to allow for the determination of the equilibrium compositions of these phases in combinations with each other. In this chapter, we will be not consider chemical and nuclear reactions, surface and tensile effects, or the effect of a gravitational or electromagnetic fields.

There are two textbook sources for the details of the concepts and models of phase equilibria. Smith, Van Ness, and Abbott[1] present the basic material in a book that is the most widely distributed textbook in chemical engineering. Advanced and very thorough information may be found in the text by Prausnitz, Lichtenthaler, and Gomes de Azevedo.[2]

FOUNDATIONS OF PHASE EQUILIBRIA

Measures of Composition

In phase equilibrium, mole fraction in each of the phases is the most common way to express composition that goes along with mole balances around the system being studied. There is also the possibility to use mass fraction for defining composition. Mass balances around the system can then be made. For some situations, absorption column design, for example, the composition may be expressed as mass of solute per mass of solvent.

Solution Equilibrium

It was the work of Josiah Willard Gibbs[3] that introduced the concept of the thermodynamics of multicomponent systems and applied the ideas to the behavior of chemical systems. A homogenous system is one in which the system properties are uniform throughout. An open system is one in which mass may be transferred between phases. We can then write the fundamental equation defining the Gibbs free energy function, G, for this system.

$$\mathrm{d}G \equiv -S\,\mathrm{d}T + V\,\mathrm{d}P + \sum_i \mu_i\,\mathrm{d}n_i \tag{1}$$

In Eq. (1), T is the temperature, P is the pressure, S is the entropy, V is the volume, n_i is the mole number of component i, and μ_i is the chemical potential of component i. In this case, the chemical potential is defined by Eq. (2).

$$\mu_i \equiv \left(\frac{\partial G}{\partial n_i}\right)_{T,P,n_j} \tag{2}$$

Consider the closed heterogeneous system made of two or more phases that are open systems. Equilibrium

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007962

is achieved when in each phase the temperature and pressure are the same and the chemical potential of each component is the same in all phases. The equality of chemical potential in every phase will define equilibrium for each component. Thus for m components and π phases, there are m times π phase equilibrium equations that can be written. The final one is Eq. (3); however, only $(m - 1)$ π are independent equations.

$$\mu_m^{(1)} = \mu_m^{(2)} = \ldots \mu_m^{(\pi)} \tag{3}$$

The Phase Rule

Intensive variables are those that do not depend on the mass of the system. In the case of a homogenous system specified by Eq. (2), the temperature, pressure, and the $m - 1$ mole fractions are intensive variables. Note that if x_i is the mole fraction of component i in a liquid solution, then $\sum_i x_i = 1.00$. Therefore, for m components only $m - 1$ mole fractions are required to define the system. Degrees of freedom, F, are the number of variables that can be independently fixed. In this case with π equal to the number of phases, there are $2 + (m - 1)(\pi)$ phase rule variables. The degrees of freedom are the difference between these phase rule variables and the number of independent equations noted above.

$$F = 2 - \pi + m \tag{4}$$

(If we allow P to be the number of phases and C to be the number of components, then $P + F = C + 2$, or in a small town like Grinders Switch, TN, the Police Force is the Chief plus Two officers.) Knowing the number of components in a system and the number of phases, we can now use this rule as guidance for the number of intensive variables we need to determine equilibrium for the system.

Chemical Potential and Fugacity

The chemical potential provides the fundamental criteria for determining phase equilibria. Like many thermodynamic functions, there is no absolute value for chemical potential. The Gibbs free energy function is related to both the enthalpy and entropy for which there is no absolute value. Moreover, there are some other undesirable properties of the chemical potential that make it less than suitable for practical calculations of phase equilibria. Thus, G.N. Lewis[4] introduced the concept of fugacity, which can be related to the chemical potential and has a relationship closer to real world intensive properties. With Lewis's definition, there still remains the problem of absolute value for the function. Thus, arbitrary reference states or standard states must be established to allow calculation to proceed. Eq. (5) presents the relationship between chemical potential, fugacity, and standard states for an isothermal change in solid, liquid, or gas, pure or mixed, ideal or real systems.

$$\mu_i - \mu_i^* = RT \ln \frac{f_i}{f_i^*} \tag{5}$$

The fugacity of the component in the solution is f_i. Both μ_i^* and f_i^* are values chosen at the standard state. Either can be arbitrarily chosen setting the value of the other, but they both cannot be chosen independently.

Activity and Activity Coefficient

The ratio f_i/f_i^* was called activity by Lewis. Therefore, the activity a_i is defined by Eq. (6)

$$a_i \equiv \frac{f_i}{f_i^*} \tag{6}$$

Activity compares the chemical potential of the component to the chemical potential of the component in the standard state. The fugacity of the component is determined by the temperature, pressure, and composition. In defining the standard state and deriving Eq. (2), the system was assumed to be isothermal. Therefore, the temperature of the standard state must be the same as the temperature of the component under study. However, the pressure and composition of the standard state may be set to best serve the situation under study.

Excess properties and activity coefficient

Liquid solutions are most readily dealt with by excess properties, which are defined for extensive properties and in this case for the Gibbs free energy by the difference between the real solution property G and the ideal solution property G. Eq. (7) is the excess Gibbs free energy.

$$G^{\mathrm{E}} = G - G^{\mathrm{id}} \tag{7}$$

Refer to Eq. (2) where the chemical potential is given as a function G at constant T, P, and composition n_j. Here, the subscript j indicates that all compositions but n_i are to be held constant along with T and P. Thus, Eq. (2) defines the partial molar Gibbs free energy $\overline{G}$.

Then, Eq. (8) defines the partial molar excess Gibbs free energy.

$$\overline{G^{\mathrm{E}}} = G - \overline{G^{\mathrm{id}}} \tag{8}$$

Note also from Eq. (2) for component i, that the chemical potential is related to the excess Gibbs free energy by Eq. (9).

$$\mu_i = \overline{G_i} \tag{9}$$

The activity coefficient can be defined by Eq. (10).

$$\gamma_i \equiv \frac{a_i}{x_i} \tag{10}$$

It can be shown that $\ln \gamma$ is a partial molar property in relation to G^E/RT. Then according to the characteristics of partial molar properties, Eq. (11) can be written.

$$\frac{G^E}{RT} = \sum x_i \ln \gamma_i \tag{11}$$

Another useful relation Eq. (12) can be found from Eq. (11) by differentiation.

$$\sum_i x_i d \ln \gamma_i = 0 \tag{12}$$

VAPOR–LIQUID EQUILIBRIUM (VLE)

Qualitative Behavior—Binary System

Equilibrium of a two component, i.e., a binary system can be readily depicted on a two dimensional figure. With two components and two phases, the degrees of freedom according to Eq. (4) would be two. If we choose to study an isothermal system, the concentration of both phases would be fixed at any pressure in the range of condensation. Conversely if we choose an isobaric system, the concentration of both phases would be fixed for any given temperature. Because many chemical separations processes run at constant pressure, we have chosen isobaric systems to illustrate. The following three figures are isobaric, temperature, x–y diagrams where x is the mole fraction in the liquid and y is the mole fraction in the vapor. For isobaric plots of binary data, the upper curve is the dewpoint and the lower curve is the bubblepoint. The dewpoint may be described for isobaric data as the temperature at which the first drop of liquid forms when cooling a constant composition solution from above the line. Conversely, when heating a constant composition liquid from below the curve, the bubblepoint is the temperature at which the first bubble appears. Fig. 1 is isobaric data for a simple binary solution, benzene and toluene, which closely obeys Raoult's law (discussed below). Fig. 2 is a more complex isobaric system, acetone and chlorobenzene, that does not obey Raoult's law. One of the more nonideal solutions and perhaps one of the most well-known azeotrope is acetone and chloroform. Due to hydrogen bonding (see Pauling[5]) to the OH^- ion, this solution, as do most of the water–alcohol solutions, forms an azeoptrope. An azeoptrope is where both the liquid and vapor phases have the same composition at the azeotropic temperature. There may be an upper or a lower boiling azeoptrope. Acetone and chloroform, shown in Fig. 3, are an upper boiling azeoptrope.

Conditions for Equilibrium

Fugacity for a vapor can be more precisely defined by Eq. (13).

$$\frac{f_i^v}{y_i P} \to 1 \quad \text{as } P \to 0 \tag{13}$$

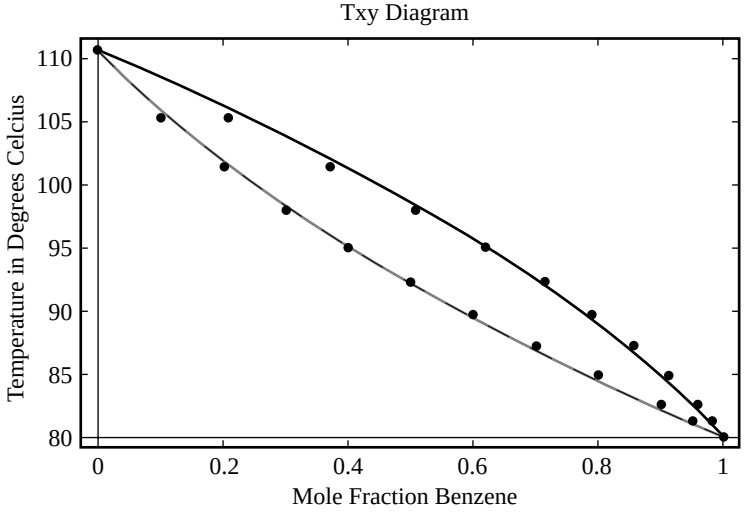


Fig. 1 Isobaric vapor–liquid equilibrium of benzene and toluene at 1.01325 Bars.

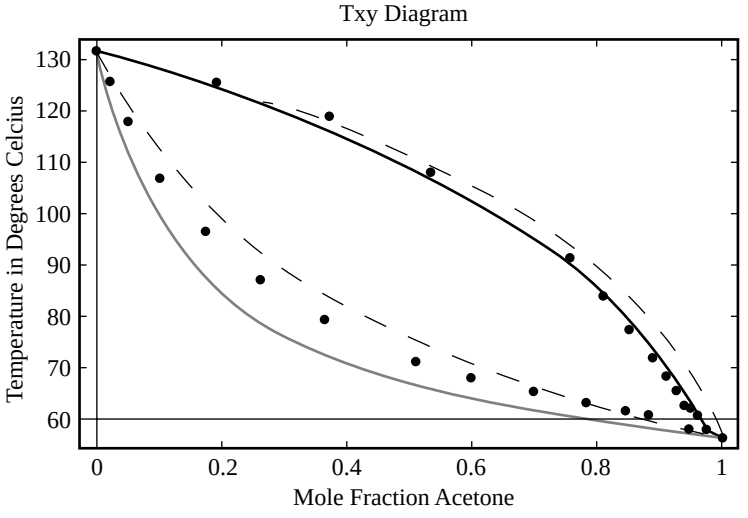


Fig. 2 Isobaric vapor–liquid equilibrium of acetone and chlorobenzene at 1.01325 Bars. Solid lines are calculated from Wilson's equation. Dashed lines are from Raoult's law.

Here, y_i is the mole fraction of component i in the vapor. As $P \rightarrow 0$ the vapor approaches an ideal gas. Therefore, for an ideal gas the fugacity is the partial pressure $y_i P$ as seen in Eq. (14).

$$f_i^v = y_i P \tag{14}$$

For phases, $\alpha,\beta,\ldots,\pi$, and component I, Eq. (5) leads to the fundamental equation of phase equilibria in terms of fugacity illustrated in Eq. (15).

$$f_i^\alpha = f_i^\beta = \cdots f_i^\pi \tag{15}$$

At equilibrium for a vapor and a liquid Eq. (16) expresses the equilibrium.

$$f_i^v = f_i^l \tag{16}$$

Raoult's Law

Raoult's law is the simplest quantitative expression for VLE. This law is based on the vapor phase being an ideal gas and the liquid phase being an ideal solution. Therefore, the vapor phase fugacity is given by Eq. (14). The effect of pressure on the liquid phase is very small, and since the vapor is ideal the liquid fugacity may be written as in Eq. (17). Here, P_i^{sat} is the

$$f_i^l = x_i P^{\text{sat}} \tag{17}$$

vapor pressure of component i at the temperature of the solution. Substituting Eqs. (14) and (17) into Eq. (16) results in Eq. (18) known as Raoult's law. Note that this law was

$$y_i P = x_i P_i^{\text{sat}} \tag{18}$$

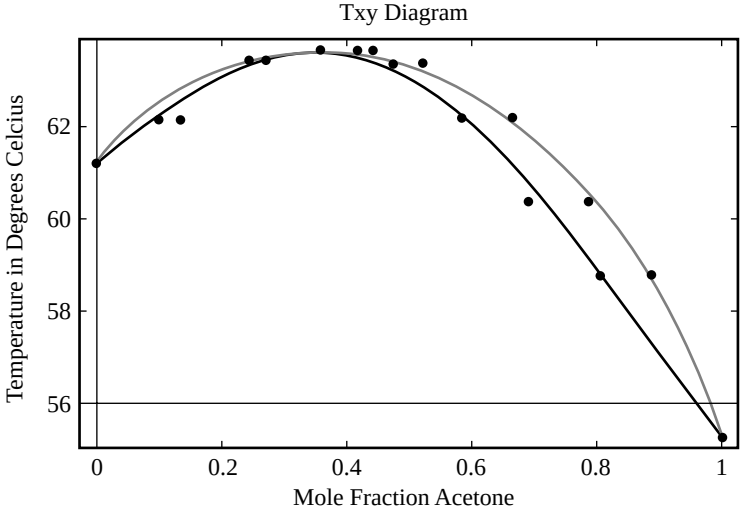


Fig. 3 Isobaric vapor–liquid equilibrium of acetone and chloroform at 0.9839 Bars. The solution is strongly hydrogen bonded, is non-ideal and does not obey Raoult's law. Solid lines are calculated from Wilson's equation.

developed essentially for an ideal solution where the vapor is an ideal gas. Therefore, Raoult's law is very limited in its usefulness. Essentially it is valid for solutions of chemically similar compounds at a pressure of up to 2 or 3 bars. As an example, see Fig. 1 above for benzene and toluene, which is a system that closely obeys Raoult's law. Raoult's law is insignificantly different than the actual data. In Fig. 2, the system is acetone and chlorobenzene, which does not obey Raoult's law. The system shown in Fig. 3 is acetone and chloroform, an azeoptrope, extremely nonideal, and cannot obey Raoult's law for the simple reason that there are three temperature points that must be satisfied where the vapor and liquid have the same composition, the two pure components and the azeoptrope. In the case of the two pure components, these temperature points are the boiling points at the pressure of this solution, 1.01325 bars.

Raoult's law can be applied only to systems where the components' vapor pressure is known. If the temperature of the system is above the critical temperature for any component, then there is no vapor pressure for that component and other means must be employed to determine compositions. This situation is described in the following section on Henry's law.

Henry's Law

In the case where the solute is at a temperature above its critical temperature, the liquid mixture cannot exist over the entire composition range. Assume that we are dealing with an ideal dilute solution where the solvent does not dissolve in the solute. We can write the fugacity for an ideal solution as in Eq. (19) where f_i is the fugacity of pure component i. Since we have a dilute

$$f_i^{id} = x_i f_i \tag{19}$$

solution, we can assume that no solvent dissolves in the solute. Let subscript 1 refer to the solvent and subscript 2 refer to the solute. Then as $x_2 \to 0$, Eq. (20) results in

$$\lim_{x_2 \to 0} \frac{f_1^{id}}{x_1} = f_1 \tag{20}$$

For component 2, the solute, we can define Henry's law constant by Eq. (21).

$$\lim_{x_2 \to 0} \frac{f_2^{id}}{x_2} = H_{2,1} \tag{21}$$

Henry's law for a dilute solution and for the situation of where the pressure is low enough that the ideal gas can apply is given by Eq. (22). Eq. (22) is very similar to Eq. (18),

$$y_2 P = x_2 H_{2,1} \tag{22}$$

Raoult's law, but now applied to dilute solutions. It is also apparent that Eq. (22) can be applied to the case where gases are only slightly soluble in a solvent. A classic case is to determine the saturation concentration of oxygen in natural waters and in waste treatment plant effluents.

The Case for Nonideal Liquid Solutions and Ideal Vapor Solutions

In the case of the ideal gas solution at moderate pressure, Eq. (16) defines the fugacity for the vapor. The activity was defined by Eq. (6), which can be combined with Eq. (10) and written as Eq. (23) for the liquid phase. If we now assume the pure component at the

$$\gamma_i^l = \frac{f_i^l}{x_i f_i^*} \tag{23}$$

pressure and temperature of the solution for the standard state of the liquid, f_i^* becomes the vapor pressure of component i. Vapor–liquid equilibrium can now be written as Eq. (26).

$$y_i P = x_i \gamma_i^l P_i^{\text{sat}} \tag{24}$$

This formulation is Raoult's law extended by the inclusion of the activity coefficient and is not much more difficult to use and it has much greater utility and accuracy at moderate pressures than Raoult's law. It is especially useful when correlations are available for liquid phase activity coefficients γ_i^l. Correlations for this purpose will be discussed in a later section. In the case of systems acetone–chlorobenzene and acetone and chloroform shown respectively in Figs. 2 and 3, the Wilson correlation equation was used to successfully fit the data. The Wilson correlation would also fit the data of benzene–toluene shown in Fig. 1; however, the fit would be no better than Raoult's law and is not illustrated in the figure.

Dealing with the Nonideal Case

The fugacity of the vapor phase can be accurately represented in terms of a fugacity coefficient defined by Eq. (25). For an ideal gas $\phi_{\text{i}} = 1.00$ and the vapor phase, fugacity can be

$$\phi_i^v = \frac{f_i^v}{y_i P} \tag{25}$$

represented by Eq. (16). A most exact representation for the vapor–liquid equilibrium can then be written as Eq. (26). A condensed phase such as a liquid phase

fugacity can be calculated

$$y_i^v \phi_i^v P = x_i \gamma_i^l f_i^l \tag{26}$$

from Eq. (27). Here, it is assumed that the specific volume, v_i^c, of the phase is very nearly

$$f_i^c = \phi_i^{\text{sat}} P_i^{\text{sat}} \exp\left[\frac{v_i^c (P - P_i^{\text{sat}})}{RT}\right] \tag{27}$$

constant over a large pressure increase. When the pressure difference in the exponential is small enough, the exponential will be nearly one. Thus at moderate pressures, the fugacity of a condensed phase is nearly equal to the vapor pressure. This approximation was used in Eq. (24) to write Raoult's law. Then when the pressure is low enough that the fugacity coefficient is nearly 1.00, Eq. (27) reduces to Raoult's law extended with the activity coefficient included.

The fugacity coefficient can be calculated from equations of state and the activity coefficient can be found from various correlations discussed in a later section.

The Equilibrium *K* Value

A useful formulation for VLE in separation calculations is the equilibrium K value defined in

$$K_i \equiv \frac{y_i}{x_i} = \frac{\gamma_i^l f_i^l}{\phi_i^v P} \tag{28}$$

Eq. (28). Under conditions of low pressure, K becomes Eq. (29), which for the case of

$$K_i = \frac{\gamma_i^l P_i^{\text{sat}}}{P} \tag{29}$$

an ideal solution can then be simplified to Eq. (30).

$$K_i = \frac{P_i^{\text{sat}}}{P} \tag{30}$$

Originally there were charts prepared to make use of these definitions. The DePriester Charts[6] for the homologous series of light hydrocarbons can serve as a basis with approximate validity for simple calculations of VLE situations.

Analyzing VLE Data to Determine Liquid Phase Properties

Extensive properties of solutions are those that depend upon the quantity of material in the system with which we are dealing. These properties can be readily handled by considering the contribution of each component separately using the concept of partial molar properties. A partial molar property of a particular component is that portion of the total extensive property of the solution. The partial molar Gibbs free energy can be written formally as in Eq. (31).

$$\overline{G} = \left[\frac{\partial(nG)}{\partial n_i}\right]_{T,P,n_{j\neq i}} \tag{31}$$

Based on Eq. (11), the partial molar excess Gibbs free energy will result in Eq. (32).

$$\frac{\overline{G^{\text{E}}}}{RT} = \left[\frac{\partial\left(\frac{nG^{\text{E}}}{RT}\right)}{\partial n_i}\right]_{T,P,n_{j\neq i}} = \ln \gamma_i \tag{32}$$

Therefore, if we can devise an equation for $\frac{G^{\text{E}}}{RT}$ as a function of the liquid mole fraction x_i, we can create an expression for the liquid activity coefficient, γ_i, as a function of the liquid mole fraction.

Representing VLE Data—Margules, van Laar, Wilson, and UNIQUAC

Wohl[7] proposed a general method for expressing excess Gibbs energies as a function of effective volume fractions. Liquid mole fractions can be related to the excess volume fractions. The advantage of Wohl's methods is that a rough physical significance could be attached to the parameters of the equations. Prausnitz[2] gives a detailed description of Wohl's method and the relation to equations representing the excess Gibbs energy. Common relationships resulting from Wohl are as follows:

Margules' equations are the best for molecules of similar size:

$$\frac{G^{\text{E}}}{x_1 x_2 RT} = A + (B - A)x_1 \tag{33}$$

$$\ln \lambda_1 = x_2^2[A + 2(B - A)x_1] \tag{34}$$

$$\ln \lambda_2 = x_1^2[A + 2(B - A)x_2] \tag{35}$$

Van Laar's equations are best for molecules of different sizes:

$$\frac{x_1 x_2 RT}{G^{\text{E}}} = \left(\frac{1}{A}\right) + \left(\frac{1}{B} - \frac{1}{A}\right)x_1 \tag{36}$$

$$\ln \gamma_1 = \frac{A x_2^2}{[x_2 + (A/B)x_1]^2} \tag{37}$$

$$\ln \gamma_2 = \frac{A x_1^2}{[x_1 + (A/B)x_2]^2} \tag{38}$$

Wilson and UNIQUAC Equations

There are many other equations, which have been proposed, that may or may not result from Wohl's method. Two of the most popular equations are the Wilson[8] and UNIQUAC by Abrams and Prausnitz.[9] These equations are based on the concept of local composition models, which was proposed by Wilson[8] in his paper. It is presumed in a solution that there are local compositions that differ from the overall mixture compositions. These local compositions result from differences in molecular size and intermolecular forces. The local compositions account for the short-range order and nonrandom molecular orientations.

Multicomponent Solutions

It is difficult to fit data with the local composition models due to their complex logarithmic forms. However, they are readily generalizable to multicomponent systems. Smith, Van Ness, and Abbott[1] and Prausnitz, Lichtenthaler, and Gomes de Azevedo[2] present the Wilson and UNIQUAC models extended for multicomponent solutions. They employ the constants from binary data. However, the constants are not unique in the sense that valid but different constants may be obtained from different sets of data. Again Wilson's equations cannot be employed for immiscible solutions, but the UNIQUAC model may be used to describe such solutions.

Wilson's equations: Good for polar or associating compounds

Wilson then derived Eq. (39) based on the local composition theory. Eqs. (40) and (41) for the activity coefficient result from Eq. (39).

$$\frac{G^{\mathrm{E}}}{RT} = -x_1 \ln(x_1 + x_2 G_{12}) - x_2 \ln(x_2 + x_1 G_{21}) \tag{39}$$

$$\ln \lambda_1 = -\ln(x_1 + x_2 G_{12}) + x_2 \left[\frac{G_{12}}{x_1 + x_2 G_{12}} - \frac{G_{21}}{x_2 + x_1 G_{21}} \right] \tag{40}$$

$$\ln \lambda_2 = -\ln(x_2 + x_1 G_{21}) + x_2 \left[\frac{G_{12}}{x_1 + x_2 G_{12}} - \frac{G_{21}}{x_2 + x_1 G_{21}} \right] \tag{41}$$

The parameters G_{12} and G_{21} are temperature dependant. A compilation of Wilson parameters can be found in Hirata, Ohe, and Nagahama.[10]

UNIQUAC Equations

The Universal Quasi-chemical Theory or UNIQUAC method of Abrams and Prausnitz[9] divides the excess Gibbs free energy into two parts. The dominant entropic contribution is described by a combinatorial part $\left(\frac{G^{\mathrm{E}}}{RT}\right)^{C}$. Intermolecular forces responsible for the enthalpy of mixing are described by a residual part $\left(\frac{G^{\mathrm{E}}}{RT}\right)^{R}$. The sizes and shapes of the molecule determine the combinatorial part, which is thus dependent on the compositions and requires only pure component data. Since the residual part depends on the intermolecular forces, two adjustable binary parameters are used to better describe the intermolecular forces. As the UNIQUAC equations are about as simple for multicomponent solutions as for binary solutions, the UNIQUAC equations for multicomponent solutions are given below. Species are identified by subscript i, subscript j is a dummy index. Here, q_i is a relative molecular surface area and r_i is a relative molecular volume. Both of these quantities are pure-species parameters.

$$\frac{G^{\mathrm{E}}}{RT} = \left(\frac{G^{\mathrm{E}}}{RT}\right)^{C} + \left(\frac{G^{\mathrm{E}}}{RT}\right)^{R} \tag{42}$$

$$\left(\frac{G^{\mathrm{E}}}{RT}\right)^{C} = \sum_i x_i \ln \frac{\phi_i}{x_i} + 5 \sum_i q_i x_i \ln \frac{\theta_i}{\phi_i} \tag{43}$$

$$\left(\frac{G^{\mathrm{E}}}{RT}\right)^{R} = -\sum_i q_i x_i \ln \left(\sum_j \theta_j \tau_{ji} \right) \tag{44}$$

$$\phi_i \equiv \frac{x_i r_i}{\sum_j x_j r_j} \tag{45}$$

$$\theta_i \equiv \frac{x_i q_i}{\sum_j x_j q_j} \tag{46}$$

In this case, $\tau_{ji} \neq \tau_{ij}$; however, when $j = i$, then $\tau_{ii} = \tau_{jj} = 1.0$. Temperature dependence enters into the equation through the parameter τ_{ji}. Gmehling, Onken, and Arlt[11] have found values

$$\tau_{ji} = \exp \frac{-(u_{ji} - u_{ii})}{RT} \tag{47}$$

for parameters $(u_{ij} - u_{ii})$ by regression of binary VLE data. The activity coefficients that follow can be found

from Eqs. (42), (43), and (44).

$$\ln \gamma_i = \ln \gamma_i^C + \ln \gamma_i^R \tag{48}$$

$$\ln \gamma_i^C = 1 \ln - J_i + J_i - 5q_i\left(1 - \frac{J_i}{L_i} + \ln \frac{J_i}{L_i}\right) \tag{49}$$

$$\ln \lambda_i^R = q_i\left(1 - \ln s_i - \sum_j \theta_j \frac{\tau_{ij}}{s_j}\right) \tag{50}$$

$$J_i = \frac{r_i}{\sum_j r_j x_j} \tag{51}$$

$$L_i = \frac{q_i}{\sum_j q_j x_j} \tag{52}$$

$$s_i = \sum_l \theta_l \tau_{li} \tag{53}$$

Parameters for the UNIQUAC equations can be found in Gmehling, Onken, and Arlt.[11]

UNIFAC—A Group Contribution Method

When it is necessary to estimate activity coefficients where no data or very limited data are, available, estimates may be made by using a group contribution method. In this case, a molecule is divided into functional groups, or subgroups of the molecule. These subgroups are assumed to act independently of the molecule in which they appear. Molecular interactions are accounted for by properly weighted sums of group interactions. Fredenslund, Jones, and Prausnitz[12] developed the method for UNIQUAC and named it Universal Functional Activity Coefficient or UNIFAC. Smith, Van Ness, and Abbott[1] reported the equations for the activity coefficients of multicomponent solutions and their parameters. These equations are very similar to the equations above for UNIQUAC. Recent sources for the group interaction parameters may be found in Hansen, Rasmussen, Fredenslund, Schiller, and Gmehling,[13] Gmehling, Li, and Schiller,[14] Fredenslund and Sørensen.[15]

Simplified Technique for Handling Binary Data

It is assumed that x, y, temperature, and pressure data are measured for a binary solution. The data can be either isothermal or isobaric. The measurements should be at a low enough pressure so that Eq. (54) will apply. The vapor–pressure relationship may be determined at the same time or found in the literature. The method is then to:

1. Calculate the activity coefficients from Eq. (24).

$$\gamma_1 = \frac{y_1 P}{x_1 P_1^{\text{sat}}} \qquad \gamma_2 = \frac{y_2 P}{x_2 P_2^{\text{sat}}} \tag{54}$$

2. Calculate the excess Gibbs energy function from Eq. (11).

$$\frac{G^{\text{E}}}{RT} = x_1 \ln \gamma_1 + x_2 \ln \gamma_2 \tag{55}$$

3. Plot $\frac{G^{\text{E}}}{x_1 x_2 RT}$ vs. x_1 to test Margules' equation and $\frac{x_1 x_2 RT}{G^{\text{E}}}$ vs. x_1 to test van Laar's equation.
4. Choose whichever plot offers the best fit to a straight line. Then, determine the parameters A and B of the activity coefficients from either Eq. (33) or (36).
5. Use Wilson's equation if neither plot produces a good fit to the data.

It would be best to examine a plot of the activity coefficients versus x_1 first. If these plots exhibit a maxima or a minima, neither van Laar's nor Wilson's equations are useful. Furthermore, Wilson's equations are not suitable for use in systems of limited solubility.

LIQUID–LIQUID EQUILIBRIUM (LLE)

Stability in Single Phase Systems

At a fixed temperature and pressure, the stable state of a system has a minimum Gibbs free energy, indicated in Eq. (56), which is the basic equation of equilibrium. A liquid mixture

$$(dG^t)_{T,P} = 0.0 \tag{56}$$

will split into two phases if it can lower this Gibbs energy. The Gibbs energy G^t is the total Gibbs energy. It can be related to the molar Gibbs energy, G, through Eq. (57). In this

$$G^t = nG < \sum_i n_i G_i \tag{57}$$

equation, the Gibbs energy G_i is for a pure component where n is the total moles in the solution.

At constant temperature and pressure, the Gibbs free energy, ΔG^{mix}, of mixing is then

$$\Delta G^{\text{mix}} = G - \sum_i x_i G_i < 0 \tag{58}$$

For a binary solution at constant temperature and pressure, a plot of ΔG^{mix} versus the mole fraction x_1 will result in a smooth curve concave downward where ΔG^{mix} is negative as long as the solution is stable. The condition for this concave curve to exist is given by Eq. (59).

$$\left(\frac{\partial^2 \Delta G^{\text{mix}}}{\partial x_1^2}\right)_{T,P} < 0 \tag{59}$$

If by decreasing the temperature the ΔG^{mix} becomes concave upward, the solution will split into two liquid phases. The composition of the phases may be determined from the plot.

Activity Coefficient Method Applied to LLE

In liquid–liquid equilibrium, equality of fugacity applies as it does in vapor–liquid equilibrium. Therefore, Eq. (15) applies to liquid–liquid equilibrium where α and β apply to the two liquid phases. Therefore,

$$f_i^{\alpha} = f_i^{\beta} \tag{60}$$

If the same standard state is chosen for both phases, Eq. (61) can be written as

$$x_i^{\alpha}\gamma_i^{\alpha} = x_i^{\beta}\gamma_i^{\beta} \tag{61}$$

The activity coefficients can be calculated from Wilson's equations or from UNIFAC if the parameters of the models are known. There are some parameters for UNIFAC in Magnussen, Rasmussen, and Fredenslund,[16] Gupte and Danner,[17] and Hooper, Michel, and Prausnitz.[18] These parameters are not as accurate as those for vapor–liquid equilibrium.

VAPOR–LIQUID–LIQUID EQUILIBRIUM (VLLE)

Under certain temperature and pressure conditions, it is possible for some binary solutions to form two liquid phases along with one vapor phase. Under these conditions, only one degree of freedom exits for the solution. Therefore, for a given pressure, the temperature and composition of all the phases are fixed. For a given temperature, the three phases will all occur in equilibrium. At a temperature greater than this, the system may be a single liquid; two phases, a vapor and liquid phase; or three phases, a vapor and two liquid phases. At low pressure, the composition of the three phase system may be calculated as shown in the following section.

Activity Coefficient Method Applied to VLLE

Eq. (24), modified Raoult's, can be applied to low pressure vapor–liquid–liquid equilibrium. For a given pressure, three phase pressure P, Eq. (24) can be rewritten for two components and for the phases α and β. The following four equations result.

$$x_1^{\alpha}\gamma_1^{\alpha}P_1^{\text{sat}} = y_1 P \tag{62}$$

$$x_1^{\beta}\gamma_1^{\beta}P_1^{\text{sat}} = y_1 P \tag{63}$$

$$x_2^{\alpha}\gamma_2^{\alpha}P_2^{\text{sat}} = y_2 P \tag{64}$$

$$x_2^{\beta}\gamma_2^{\beta}P_2^{\text{sat}} = y_2 P \tag{65}$$

When complete immiscibility occurs, Eqs. (63) and (64) can be added to give Eq. (66).

$$P = y_1 P + y_2 P = x_1^{\beta}\gamma_1^{\beta}P_1^{\text{sat}} + x_2^{\alpha}\gamma_2^{\alpha}P_2^{\text{sat}} \tag{66}$$

The vapor composition can be calculated from Eq. (66), which results in the following equation.

$$y_1 = \frac{x_1\gamma_1 P_1^{\text{sat}}}{P} \tag{67}$$

Thus with knowledge of the activity coefficients, all compositions of the three phases could be calculated.

SOLID–LIQUID EQUILIBRIUM (SLE)

Fugacity and Activity Coefficient Method Applied to SLE

In dealing with the solution of solids in liquids, the solubility or mole fraction of the solute in the solvent is the primary variable of interest. The solubility depends on the intermolecular forces between the solute and the solvent, which are represented by the activity coefficient. However, equally important in determining the solubility is the standard state to which the activity is referred and the fugacity of the pure solid. In writing the equilibrium equation, it is assumed that there is no solubility of the solvent in the solid. Then, Eq. (68) can be written.

$$f_{2(\text{pure solid})}^{s} = \gamma_2 x_2 f_2^{o} \tag{68}$$

The solubility is given by x_2, γ_2 is the activity coefficient of the liquid phase, and f_2^o is the standard state fugacity to which the activity coefficient is related. The solubility can then be found from Eq. (69), which illustrates the dependency of the solubility on the

ratio of the

$$x_2 = \frac{f_{2(\text{pure solid})}}{\gamma_2 f_2^o} \tag{69}$$

two fugacities.

Define the standard state as the pure sub-cooled liquid at the operating temperature. Then, f_2^o becomes the fugacity of the pure sub-cooled liquid, $f_{2(\text{pure, subcooled liquid})}$. Redefine the two fugacities as follows,

$$f^s_{2(\text{pure solid})} = f_2^s \quad \text{and} \quad f_{2(\text{pure sub-cooled liquid})} = f_2^{\text{SCL}}$$

By considering a path at the operating temperature where the solid passes through the triple point to a sub-cooled liquid, the ratio of the two fugacities in Eq. (69) can be written as follows.

$$\ln\frac{f_2^{SCL}}{f_2^s} = \frac{\Delta H_{fus}}{RT_t}\left(\frac{T_t}{T} - 1\right) - \frac{\Delta c_p}{R}\left(\frac{T_t}{T} - 1\right) + \frac{\Delta c_p}{R}\ln\frac{T_t}{T} \tag{70}$$

In Eq. (69), T_t is the triple point temperature, ΔH_{fus} the heat of fusion, and $\Delta c_{\text{p}} \equiv c_{\text{p(liquid)}} - c_{\text{p(solid)}}$ between T and T_t. Eq. (70) can be combined with Eq. (69) to find the solubility.

For an ideal solution $\gamma_2 = 1.0$ and in many cases, Δc_{p} can be neglected and the triple point temperature can be approximated by the melting point temperature T_{m}. Then, the solubility becomes Eq. (71).

$$\ln x_2 = -\frac{\Delta H_{fus}}{RT_m}\left(\frac{T_m}{T} - 1\right) \tag{71}$$

Prausnitz, Lichtenthaler, and Gomes de Azevedo[2] show several methods for determining the activity coefficient for SLE.

PHASE EQUILIBRIUM AT HIGH PRESSURE

Vapor–Liquid Equilibrium from Equations of State

Prausnitz, Lichtenthaler, and Gomes de Azevedo[2] suggest that the best way to approach high pressure vapor–liquid is through the fugacity and fugacity coefficient. Eq. (25) defines the fugacity coefficient for a vapor. In the same way, the fugacity coefficient can be defined for a liquid as follows,

$$\phi_i^L = \frac{f_i^L}{x_i P} \tag{72}$$

At equilibrium, the frugalities will be equal and the following equation will result.

$$\varphi_i^V y_i = \varphi_i^L x_i \tag{73}$$

It is then possible to calculate the fugacity coefficient from an equation of state applied to Eq. (74). In this equation, n_i is the moles of component i and n_T is the total number of

$$\ln\phi_{\text{i}}^{\text{V}} = \frac{1}{RT}\int_{V^V}^{\infty}\left[\left(\frac{\partial P}{\partial n_i}\right)_{T,P,n_{j\neq i}} - \frac{RT}{V}\right]dV - \ln\frac{PV^V}{n_T RT} \tag{74}$$

moles in the solution. A similar equation can be written for the liquid phase. The equation of state selected is applied to both the vapor and liquid fugacity coefficients. Since the calculation will involve at least binary mixtures and quite likely multicomponent mixtures, mixing rules for the equation of state parameters will need to be used. Smith, Van Ness, and Abbott[1] report this procedure for a cubic equation of state.

Solid–Vapor Equilibrium (SVE)—Supercritical Extraction

Important to many current processes is the ability of gasses at pressure above the critical to have increased solubility capability. A common example is the decaffeination of coffee beans by supercritical carbon dioxide. In this case, it is assumed that species 1 is the solute dissolved in species 2, which is therefore the solvent. It is assumed that the solvent is insoluble in the solute. Thus, the equilibrium can be written as follows,

$$f_1^{s(\text{pure})} = f_1^v \tag{75}$$

Then the solubility of the solute in the gas, y_1, can be written as in Eq. (76).

$$y_1 = \frac{P_1^{\text{sat}}\phi_1^{\text{sat}}}{P\varphi_1^{(\text{pure})}}\exp\left[\frac{V_1^s\left(P - P_1^{\text{sat}}\right)}{RT}\right] \tag{76}$$

The molar volume of the solid is V_1^s. The vapor phase nonidealities are reflected in the ratio of the fugacity coefficients $\frac{\phi_1^{\text{sat}}}{\phi_1^{(\text{pure})}}$. In most cases, the vapor pressure of the solute, P_1^{sat}, will be small enough to be neglected. An example of the use of this equation

to determine the solubility of naphthalene in carbon dioxide is given in Smith, Van Ness, and Abbott.[1]

SOME SPECIAL TOPICS

There are several areas of interest in chemical processing that should be reviewed in this chapter. These areas include retrograde condensation, polymers, and electrolytes. In each case, the material is more extensive than can be covered in any detail in this chapter. Reference should be made to Prausnitz, Lichtenthaler, and Gomes de Acevedo.[2] Furthermore, the principles covered in this chapter form the basis for viewing and calculating the phase equilibrium that applies. The subjects of polymers and electrolytes are so complex that they require a chapter by themselves. However, a brief statement follows describing retrograde condensation.

Retrograde Condensation

The vapor pressure curve forms the basis for the description of vapor–liquid equilibrium for a pure fluid. As the temperature increases, the vapor pressure curve for the vapor–liquid situation ends at the critical pressure. In the case of a binary or multicomponent solution, the critical point is not necessarily a maximum with respect to either temperature or pressure. It is then possible for a vapor or liquid to exist at temperature or pressures higher than the critical pressure of the mixture. At constant temperature, it is then possible for condensation to take place as the pressure is decreased. At constant pressure, condensation may take place as the temperature is increased. Vaporization can take place at constant temperature as the pressure is increased and decreased. This unusual behavior can be useful in some process situations, for example, in the recovery of natural gas from deep wells. If the conditions are right, liquefaction of the product stream is possible. At the same time, the heavier components of the mixture may be separated from the lighter components.

CONCLUSIONS

Phase equilibria are necessary components for any process that requires separation of materials to take place. Phase equilibria have always been necessary in the design of chemical and petroleum processing plants. They are also important in environmental situations like dissolved oxygen in natural waters and the design of air pollution control equipment. The subject is quite complicated and can only be briefly reviewed in this chapter. References provide the necessary information for a fuller understanding of the theories, correlations, and use of the methods of data handling.

ACKNOWLEDGMENTS

In preparation of the chapter, the author wants to acknowledge that his interest in phase equilibria was stimulated by Lawrence Nicholas Canjar. Professor Canjar was a great teacher, scholar, and a warm friend. He excelled especially in the field of chemical engineering thermodynamics. Through lectures and discussions, he opened the way for a lifelong interest of the author in teaching and using thermodynamics, especially Phase Equilibria. The author also wants to recognize Professor John Allen Tallmadge who preceded the author as an earlier student of Professor Canjar. It is their data that are presented in Fig. 3 of this chapter.

REFERENCES

1. Smith, J.M.; Van Ness, H.C.; Abbott, M.M. *Introduction to Chemical Engineering Thermodynamics*, 6th Ed.; McGraw Hill: New York, 2001.
2. Prausnitz, J.M.; Lichtenthaler, R.N.; Gomes de Azevado, E. *Molecular Thermodynamics of Fluid-Phase Equilibria*, 3rd Ed.; Prentice Hall: Upper Saddle River, New Jersey, 1999.
3. Gibbs, J.W. *The Scientific Papers of J. Willard Gibbs*; Dover Publications: New York, 1961; Vol. 1.
4. Lewis, G.N.; Randall, M. *Thermodynamics*; McGraw-Hill book Co. 1923.
5. Pauling, L. *The Nature of the Hydrogen Bond*, 3rd Ed.; Cornell University Press: New York, 1960.
6. DePriester, C.L. The K values for systems of light hydrocarbons. Chem. Eng. Pro. Symp. Ser. **1953**, *49* (7), 41–42.
7. Wohl, K. Thermodynamic evaluation of binary and ternary liquid systems. Trans. AIChE **1946**, *42*, 215–240.
8. Wilson, G.M. Vapor–liquid equilibrium. XI. A new expression for the excess free energy of mixing. J. Am Chem. Soc. **1964**, *86*, 127–130.
9. Abrams, D.S.; Prausnitz, J.M. Statistical thermodynamics of liquid mixtures: A new expression for the excess Gibbs energy of partly or completely miscible systems. AIChE J. **1975**, *21*, 116–128.
10. Hirata, M.; Ohe, S.; Nagahama, K. *Computer Aided Data Book of Vapor–Liquid Equilibria*; Elsevier Scientific Publishing Company: New York, 1975.

11. Gmehling, J.; Onken, U.; Arlt, W. *Vapor–Liquid Equilibrium Data Collection*; Chemistry Data Series, Vol. 1, parts 1–8, DECHEMA; Frankfurt am Main: 1974–1990.
12. Fredenslund, A.; Jones, R.L.; Prausnitz, J.M. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. AIChE J. **1975**, *21*, 1086–1099.
13. Hansen, H.K.; Rasmussen, P.; Fredenslund, A.; Schiller, M.; Gmehling, J. Vapor–liquid equilibria by UNIFAC group contribution. 5. Revision and extension. Ind. Eng. Chem. Res. **1991**, *30*, 2352–2355.
14. Gmehling, J.; Li, J.; Schiller, M. A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties. Ind. Eng. Chem. Res. **1993**, *32*, 178–193.
15. Fredenslund, A.; Sørensen, J.M. Group contribution estimation methods. In *Models for Thermodynamic and Phase-Equilibria Calculations*; Sandler, S.I., Ed.; Marcel Decker, Inc.: New York, 1994.
16. Magnussen, T.; Rasmussen, P.; Fredenslund, A. UNIFAC parameter table for prediction of liquid–liquid equilibriums. Ind. Eng. Chem. Res. Des. Dev. **1981**, *20*, 331–339.
17. Gupte, C.J.; Danner, R.P. Prediction of liquid–liquid equilibria with UNIFAC: A critical evaluation. Ind. Eng. Chem. Res. **1987**, *26*, 2036–2042.
18. Hooper, H.H.; Michel, S.; Prausnitz, J.M. Correlation of liquid–liquid equilibria for some water-organic liquid systems in the region 20–250°C. Ind. Eng. Chem. Res. **1988**, *27*, 2182–2187.

Phenolic Resins

Adriane G. Ludwick
Mohamed O. Abdalla
Department of Chemistry, Tuskegee University, Tuskegee, Alabama, U.S.A.

INTRODUCTION

This entry will provide an overview of the ''classical'' phenol–formaldehyde system. The fundamentals of this system will be described. The current applications of this classical system will be discussed. Resins prepared from structurally modified phenols will be examined and labeled as ''modified-classical'' phenol–formaldehyde systems. The effect of these modifications on the mechanical, thermal, and other properties will be examined. Finally, the considerable work on polymers that can be classified as ''nonclassical'' phenolic resins will be presented and the area of nanocomposites utilizing phenolic resins will be examined.

The aim of this entry is to present a balanced view of the current state of phenolic resins. The patent literature, refereed journals, and other general sources have been utilized to survey the classical and nonclassical phenolic resin systems as well as the nanocomposite systems. The article is pedagogical rather than comprehensive.

PHENOLICS CHEMISTRY

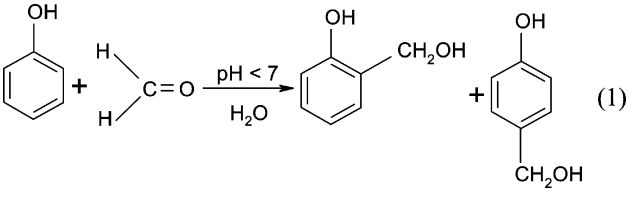
This section gives a brief overview of the chemistry of classical phenolics. Classical phenolics refer to the reaction of phenol and formaldehyde in water under basic or acidic conditions.

Phenol and Formaldehyde Reaction

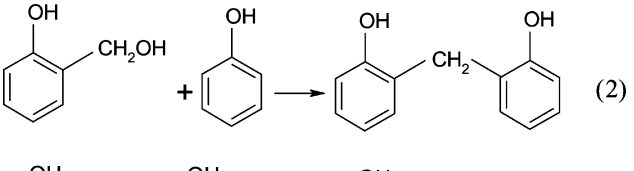
Phenol and formaldehyde react in water to give *ortho*- and *para*-substituted products (Scheme 1). The reaction mechanism is an electrophilic aromatic substitution.[1–3] The product depends on the molar ratio of phenol to formaldehyde, pH conditions, and temperature. The mechanism is essentially the same in acidic or basic aqueous solution, but the details of the reaction are different. The product in an acidic medium is known as a novolac and that in basic medium is known as a resol.

Novolac preparation and processing

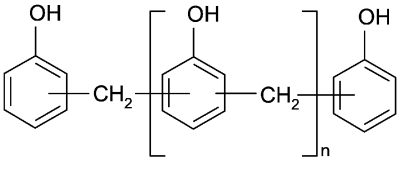
Typically, in the preparation of a novolac, the molar ratio of phenol and formaldehyde is greater than 1. Hence, the initial major product is a mono-substituted phenol [Eq. (1)]. Because the reaction is done under aqueous acidic conditions, the products shown in Eq. (1) are not isolated. Instead, a methylene bridge is formed between the phenyl rings [Eq. (2)]. In both Eqs. (1) and (2) the mechanism is an electrophilic aromatic substitution. Heating the system so as to promote removal of water and polymerization results in thermoplastic material known as novolac (**1**). This thermoplastic resin can be mixed with hexamethylenetetramine (formed from ammonia and formaldehyde) and stored until cure. Heating this system produces an excess of formaldehyde and ammonia. A cross-linked polymer results from the cure. The linkages are mostly methylene and amino groups.



(1)



(2)

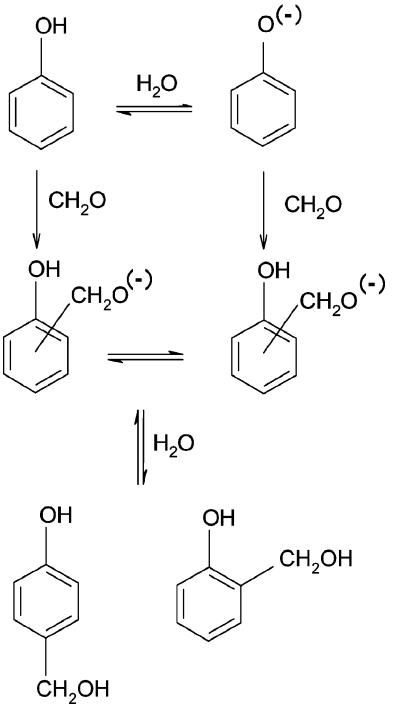


1

Resol preparation and processing

The chemistry of resol formation has many similarities to that of novolacs. However, because conditions utilize a phenol to formaldehyde ratio less than 1, mono-, di-, and tri-substitutions to the aromatic ring occur. Additionally, the aqueous medium is basic. Hence,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120027016



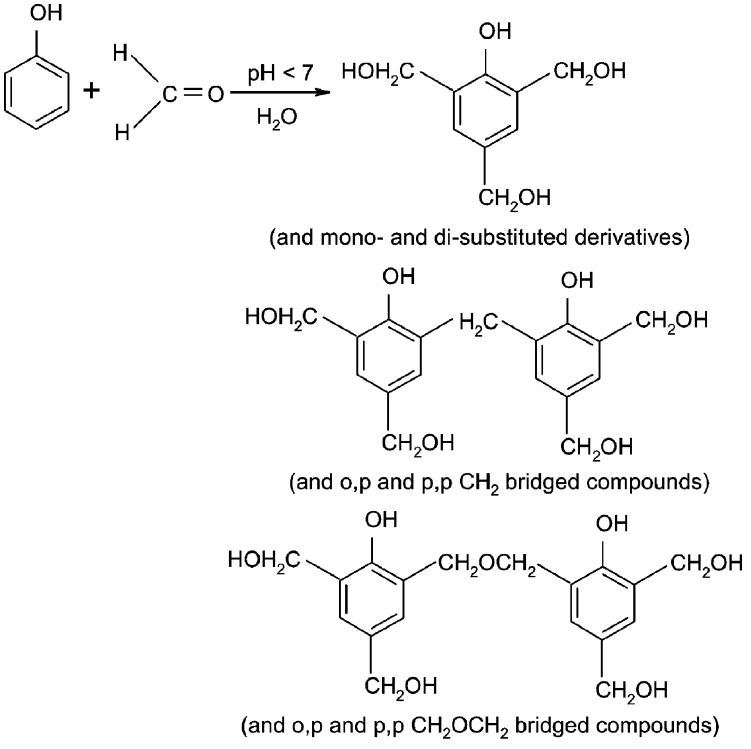
Scheme 1 General pattern for phenol–formaldehyde reaction.

these initial substitution products can be isolated (see Scheme 2). Cure of this system resulting in a cross-linked polymer occurs upon heating to temperatures above 100°C, typically 125–150°C.

Resol application

Using the classical resol chemistry, Asai and coworkers patented new molded materials.[4] Added to the resol phenolic resin are inorganic fiber, a silica powder, and a rubber component. The prepared phenolic resin composition was noted to have excellent resistance to abrasive wear and to heat shock, as well as excellent mechanical strength. As such, this material is very suitable for the fabrication of engine parts such as pulleys. These pulleys are currently fabricated from heavier and more expensive materials.

The resol phenolic resin is dimethylene-ether-based or methylol-based and either solid or liquid. The rubber component includes as possibilities: nitrile, acrylic, chloroprene, styrene–butadiene, or silicone. The inorganic fiber component includes glass, carbon, silicon carbide, or potassium titanate. The glass fiber may be preferably subjected to a surface treatment with a silane coupling agent to improve adhesion to the phenolic resin. Asai and coworkers optimized the amount of rubber in the resin composition to 1–15 parts by weight in 100 parts by weight of phenolic resin. Less than one part by weight of rubber detrimentally affected the elastic modulus of the product. More than 15 parts by weight of rubber resulted in a material with poor abrasive wear resistance. The inorganic fibers and silica powder composition were essential for the abrasive wear resistance properties of this phenolic resin mixture.



Scheme 2 General reaction for preparation of resol resins.

MODIFIED PHENOLIC RESINS

Phenol Derivatives

Modification of the phenol ring can affect both the kinetics and the properties of the phenol–formaldehyde system. Substitution at the *ortho* and/or *para* position of the phenol will limit the extent of cross-linking and hence the mechanical properties of the cured phenolic resin. A summary of current work on these modified-classical phenolic resins is given.

Bromine substitution

Including 4-bromophenol in the phenol–formaldehyde resol system impacts the cross-link density of the cured product. In a systematic study of this copolymer, a comparison was made among the polymers obtained using phenol only, a 9 : 1 mole ratio of phenol to 4-bromophenol and a 1 : 1 mole ratio of phenol to 4-bromophenol.[5] Comparisons included measurement of interlaminar shear strength and cone calorimetry tests of composites prepared using these phenolic resins and S2-glass fiber plain weave.

Table 1 demonstrates the effect on mechanical properties for the phenol–4-bromophenol systems. The substitution of bromine at the *para*-(4-) position on the phenol ring gives a significantly less brittle material than from the classical unsubstituted phenol. The cone calorimetry results (Table 2) indicate that the 1 : 1 phenol–4-bromophenol system has the best flammability properties. The 9 : 1 phenol–4-bromophenol and phenol systems give similar values. The hazard from the generation of hydrogen bromide upon burning a bromine-containing compound mitigates the usefulness of this material. However, this is a sound approach toward obtaining improved properties for phenolic systems.

The greater flexibility of the 1 : 1 phenol–4-bromophenol system indicated by these results is consistent with ^{13}C NMR observations. Analysis of the C–Br carbon absorptions confirmed that the phenol and 4-bromophenol rings are connected through primary bonds. Additionally, the 1 : 1 phenol–4-bromophenol resol contains the most ether linkages. The ether connections between the aromatic rings are more flexible than the methylene connections. Presumably, this situation persists in the cured system.

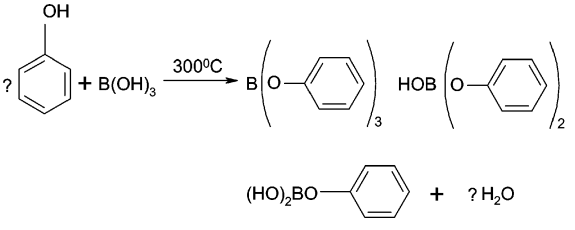
Table 1 Interlaminar shear strength properties of phenolic composites

Composite resin	Average maximum failure load (lb)	Average shear strength (psi)
P	73.7	951
9 : 1 P/4-BrP	90.8	1087
1 : 1 P/4-BrP	121.9	1443

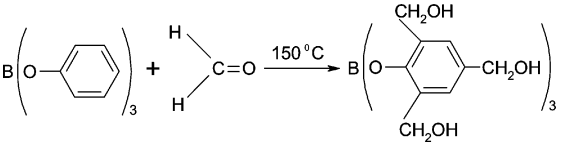
(With permission from Ref.[5].)

Boron-modified phenolic resin

Hirohata et al. reported a solid-state preparation of boron-modified phenolic resins (BPR).[6] A triphenyl borate ester (TPB) was prepared by reacting phenol with boric acid [Eq. (3)]. The TPB was then reacted with paraformaldehyde to produce the BPR [Eq. (4)]. The solid product was obtained by resinifying at 150°C. The resin produced by this method was difficult to process by conventional processing methods such as resin transfer molding. No structural characterization for the prepared TPB ester or resins was reported in Hirohata's study. However, thermal characterization of the hardened prepared resin showed the BPR to have superior resistance against thermo-oxidative degradation in comparison to the hardened regular and halogenated phenolic resins.



(3)



(4)

A more recent study by Abdalla, Ludwick, and Mitchell presented a different approach for the preparation for BPR.[7] The resin was prepared from the reaction of TPB and paraformaldehyde at three different resinifying temperatures, 130°C, 120°C, and 90°C. The BPR produced at 90°C melted upon reheating, which indicated promising processing applications for this resin. ^{1}H and ^{13}C NMR spectra of resins from the three resinifying temperatures had the same pattern of absorptions, similar to that of phenolic resins that did not contain boron. Comparison between the ^{1}H NMR spectra of the TPB and the BPR prepared at 120°C (see Fig. 1) showed that substitution of methylol groups occurred at the *ortho* and *para* positions of the ester phenyl rings (4.86–4.75 ppm). Aromatic, methylene and ether linkage protons were assigned at about 7.45–6.74 ppm, 4.93–3.36 ppm, and 5.30–4.91 ppm. The application of this system needs to be explored.

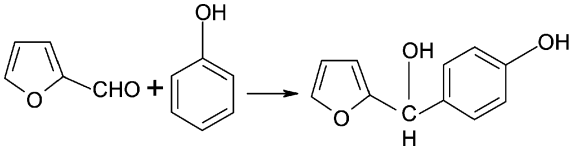
Table 2 Selected cone calorimetry results for phenolic composites

Composite resin	Time to ignition (sec)	Peak HRR (kW/m^2)	EHC (MJ/kg)	CO yield (g/g)
P	110	107	16.5	0.118
P/4-BrP 9 : 1	115	93	16.8	0.191
P/4-BrP 1 : 1	181.5	31	6.3	0.281

HRR, heat release rate; EHC, effective heat of combustion.
(From Ref.[5].)

Carbonyl Derivatives

Furfural is the aldehyde most commonly used as a replacement for formaldehyde in the preparation of phenolic resins.[8] Furfural is obtained from natural sources. The naturally abundant pentosans, obtained from agricultural and forestry wastes, are hydrolyzed by dilute sulfuric acid to give furfural. The furfural is then isolated by steam distillation. The reaction of furfural under alkaline conditions is similar to that of formaldehyde, yielding *ortho*- and *para*-hydroxyphenyl furfuryl alcohol [Eq. (5)]. The polymerization of this furfural resol proceeds by a mechanism that is not fully understood. It is possible that the scission of the furan ring occurs and that this ring takes part in the polymerization reaction, leading to a complex mixture of products. The cleavage of the furan ring occurs more readily under acidic conditions, giving, however, the same complex product as in the alkaline medium.



(5)

Furfural has replaced formaldehyde in phenol–resorcinol–formaldehyde resins.[9] These formaldehyde-containing resins have been used as cold-set, exterior grade structural wood adhesives for almost half a century. The use of furfural to prepare these resins has several advantages. In addition to longer resin shelf life, when furfural is used, the emission of formaldehyde is lowered. To enhance the reaction rate when furfural is used, a small amount of formaldehyde is added to the furfural system. Hence, the furfural system is not formaldehyde-free, but nearly so. Finally, furfural has a higher molecular weight than formaldehyde. This results in a smaller amount of the expensive resorcinol being required in the system.

Modified phenolic resin applications

Phenolic resins have been used as binders for mineral wool insulation products. Typically, the phenol–formaldehyde resol used in binders for wool insulation products has large amounts of free formaldehyde. This is due to the high molar ratio between formaldehyde and phenol typically used for insulation resins. Bohumila et al. patented that the addition of a small amount of sodium tetraborate during the early stage of the formation of a phenolic resin enhances its properties. Lower molecular weight products are obtained and the resin stability appears to be improved.[10] The advantage of a stable prepolymer is that it will reduce formaldehyde emissions during transfer and storage from the manufacturing site to the user's site.

Chiu, Shelast, and Lam patented a fast curing spray-dried phenol–formaldehyde composition that contains 0.02–0.09 mol of highly reactive phenolic compounds as curing accelerators per 100 parts of the solids of the phenolic resin.[11] The curing accelerators may include resorcinol, alkyl resorcinol, *m*-aminophenol, and phloroglucinol. The curing accelerator in the powder resin functions as a cross-linking agent. This cross-linking agent reacts with the methylol groups of the resole resin under the heat and pressure experienced during the manufacture of wood composite products. The fast curing powder resin composition may be made by first preparing a resol liquid phenol–formaldehyde resin that contains the desired amount of methylol groups and of the desired molecular weight. The second step includes eliminating the residual free formaldehyde with a scavenging agent. Then, the prepared liquid resol resin is mixed with the highly reactive phenolic compound. The mixture is spray-dried to prepare a powder resin composition without realizing a chemical reaction between the highly reactive phenolic compound and the phenol–formaldehyde resin. The spray-drying process stabilizes the highly reactive phenolic compound in the phenol–formaldehyde resin composition. Besides being a fast-curing product, Chiu and coworkers named several other advantages of their formulated powder resol including good shelf life, which causes reduction in the cost of manufacturing wood composites.[11] Moreover, the powder has a high tolerance of wood moisture content, which improves the quality of the manufactured wood composites.

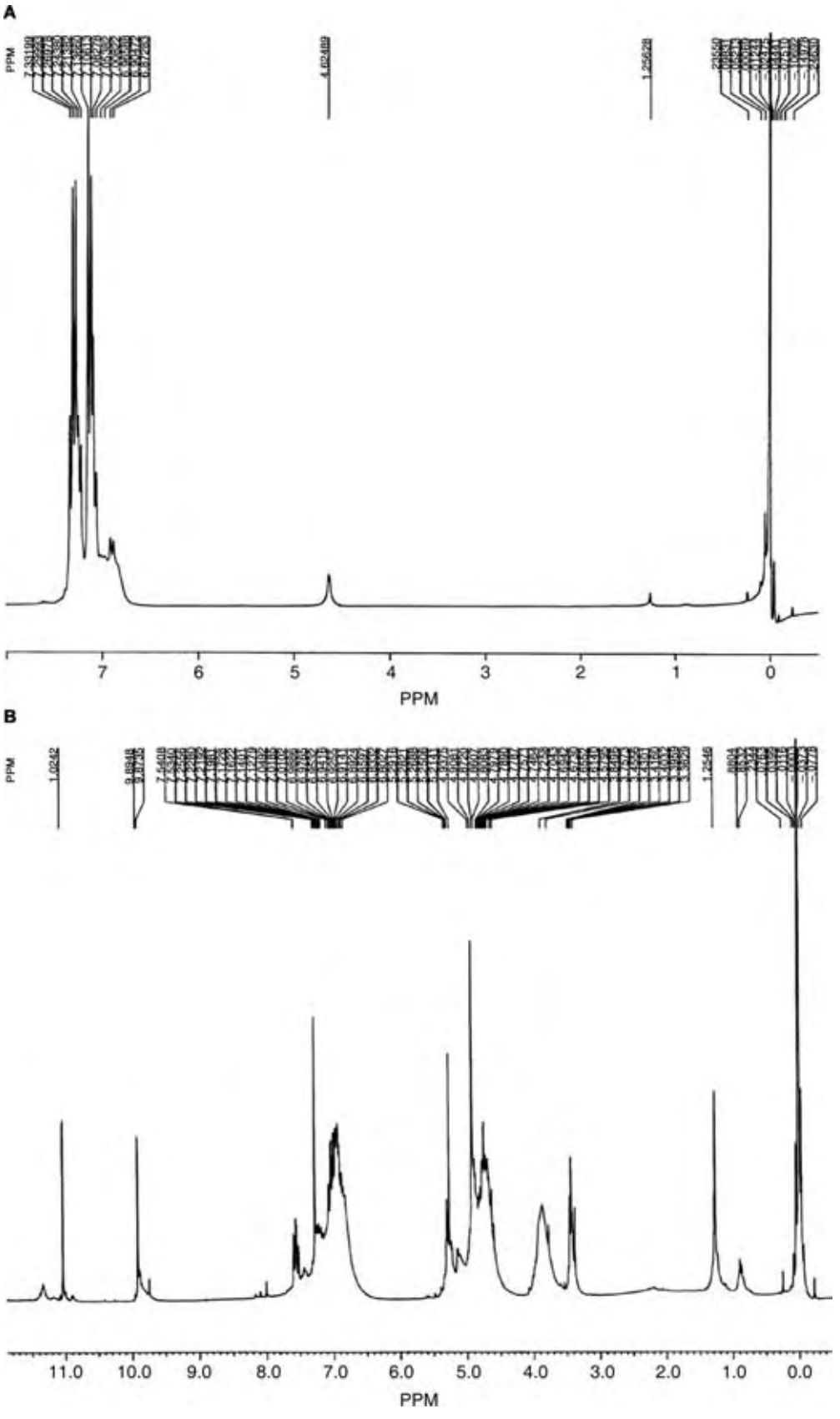


Fig. 1 [1]H NMR spectra of (A) pure TPB and (B) BPR prepared at 120°C. (From Abdalla, M. M.S. thesis, Tuskegee University, Tuskegee, AL, Aug 2000.)

NONCLASSICAL PHENOLIC SYSTEMS

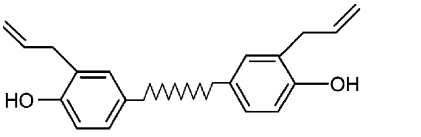
Addition-Cure Phenolic Resins

A recent review of addition-cure phenolic resins provides extensive information on the work in this area.[12] In this entry, parts of the review will be highlighted to provide a view of the importance of the work and its relationship to the classical and modified-classical phenolic resins.

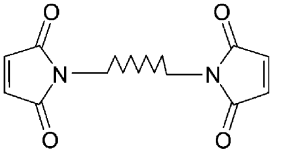
The cross-linked polymers obtained from the addition-cure approach are often a complex arrangement of atoms bonded in heterocyclic and carbocyclic rings. However, the objective in the preparation of these systems is not simplicity. It is to obtain systems with the desirable properties of phenolic resins retained and the undesirable properties improved or removed. Voids are an undesirable result in the synthesis of both resols and novolacs. Hence, addition-cure phenolic resins are designed to avoid this result. Ease and flexibility of processing are also sought in the addition-cure systems.

Diels–Alder systems

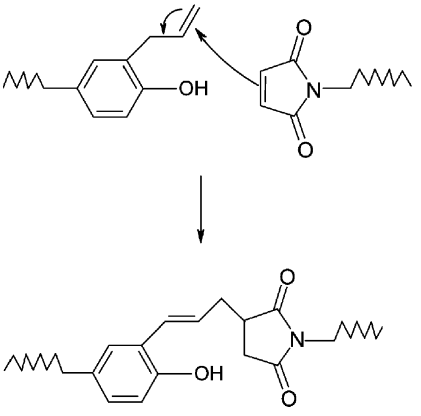
A well-studied addition-cure phenolic resin utilizes the Diels–Alder reaction for its preparation. The result is a complex polymer that originates from a diallyl phenol and a diimide (for example, see structures **2** and **3**, respectively). Initially, compounds **2** and **3** react across the double bond of the imide ring and the allylic group to generate a diene [see Eq. (6)]. Compound **2** then serves as the dienophile. The polymer forms from the reaction of the diene with **2**. The amount of **2** relative to the diene affects the properties of the resulting polymer, as does the cure temperature. Some of these properties will be summarized collectively, for all the addition-cure systems discussed in Table 3.



2



3



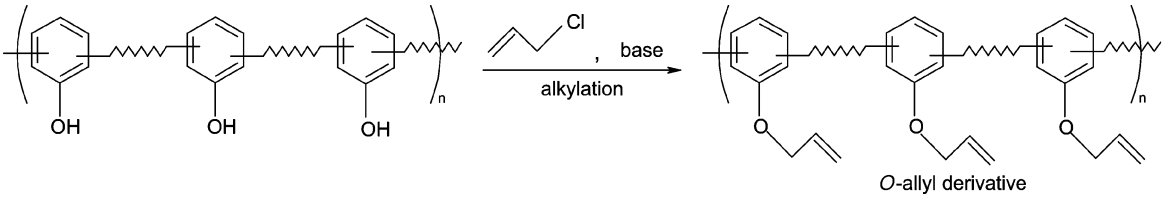
(6)

Rather than building the polymer through a succession of Diels–Alder reactions, the approach can be to utilize the hydroxyl groups on the backbone of a thermoplastic phenolic resin as the point of attachment for allylic groups. Heating rearranges the allylic attachment from the oxygen to the aromatic ring. Reaction of a diimide with this polymer produces the diene. The diene–dienophile reaction then proceeds to give the cross-linked addition-cure phenolic resin [see Eq. (7), formation of *O*-allyl derivative; Eq. (8), *C*-allyl

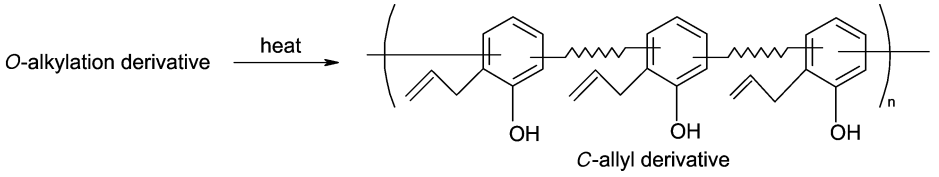
Table 3 Selected properties of addition cure polymers

	Bismaleimide/2,2′-diallyl bisphenol A[a]	Boron-containing alderenes[b]	Poly(bisoxazoline)/ Phenolic[c]	Phenolic/ epoxy[d]
Tensile strength (Mpa)	80–97	74–84	90	45.3[e]
Flexural strength (Mpa)	160–192	98–139	172–194	98.6[e]
Compressive strength (MPa)	220	—	237–256	—
HDT (°C)	273–295	—	—	—
T_g (°C)	195–217	278–330	—	96–151

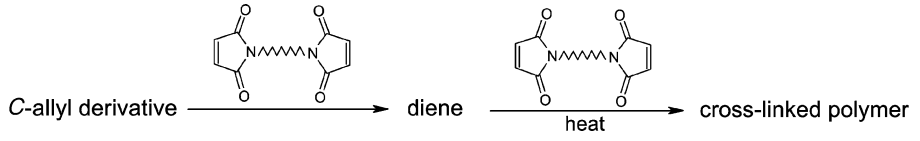
[a]For various molar ratios of reactants.
[b]For various boron ester structures.
[c]For various weight ratios of bisoxazoline to phenolic.
[d]For various weight ratios of phenolic to epoxy (50 : 50 to 80 : 20).
[e]Specific weight ratio of phenolic to epoxy (36 : 64).



(7)



(8)



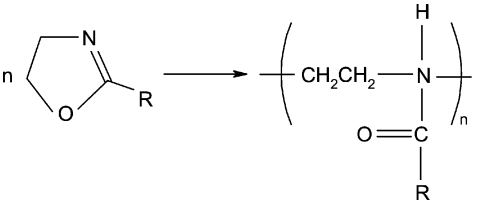
(9)

derivative formation; and Eq. (9), Diels–Alder reaction to give cross-linked polymer].

The thermoplastic phenolic resin in the above Eqs. (7)–(9) could be a novolac resin.[13,14] Similarly, the commercial polyparaalkyl phenol, Xylok, has been used in Diels–Alder reactions with diimides.[15,16] These systems have good processing properties, good mechanical properties, and superior thermal properties.

Oxazoline ring systems

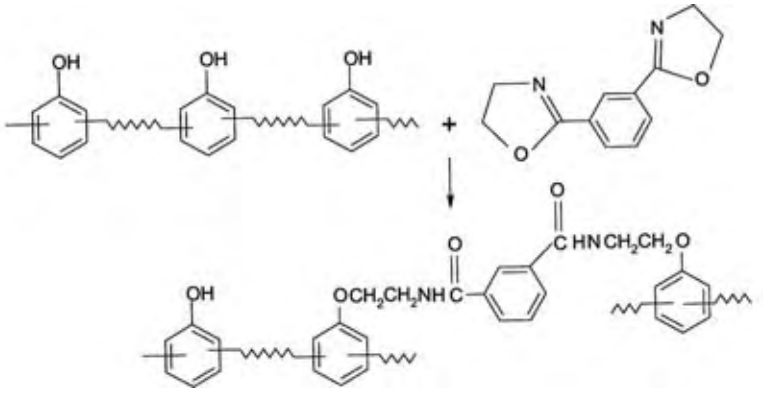
The versatile oxazoline ring has been utilized in reactions with novolacs. Oxazolines polymerize readily to give thermoplastics of various molecular weights [see Eq. (10)]. Reaction of a bisoxazoline with a novolac gives a cross-linked system that has properties dependent on the relative amounts of bisoxazoline to novolac [see below Eq. (11)]. This approach for an addition-cure phenolic resin has been commercialized.[17]



(10)

Epoxy systems

As an alternative to using hexamethylenetetramine as a cross-linking agent for novolacs, the hydroxyl group of the novolac can be reacted with epoxy reagents.[12,18,19] The variables to this approach are plentiful. The epoxy

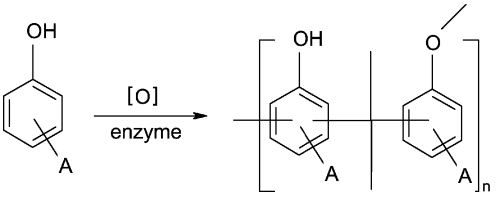


(11)

reagent can include as possibilities classical Bisphenol-A-based, halogen-containing epoxy systems, or epoxy systems with stereochemical features. Additionally, the relative amounts of novolac and epoxy reagent can be varied. A relatively larger amount of novolac will result in a decrease in T_g and an increase in toughness for the resulting cross-linked polymer. Flammability can be controlled either through the traditional novolac or through appropriate substitutions on the epoxy reagent. A schematic for the cross-linking is shown in Scheme 3. Some properties of these systems are included in Table 3.

Enzymatically Synthesized Phenolic Polymers

This class of phenolic polymers opens a wealth of potential systems.[20–22] The preparation [Eq. (12)] involves a phenol, an oxidizing agent, and an enzyme. The polymer obtained consists of phenylene and oxyphenylene units. The attachment A is limited only by the kinetics and thermodynamics of the reaction. In a recent study, A was CH_2OH.[23] The monomer, 4-hydroxybenzyl alcohol, was polymerized using hydrogen peroxide as the oxidizing agent and horseradish peroxidase as the catalyst. The resulting polymer therefore had a CH_2OH primary alcohol pendant. The alcohol functional group was reacted with the carboxylic acid functional group of 4-(1-pyrenyl) butanoic acid to give a pendant sensitive to explosive materials (e.g., dinitrotoluene). The approach of these authors would lend itself to other applications in, for example, agriculture (herbicide and pesticide attachments) and in the pharmaceutical field.

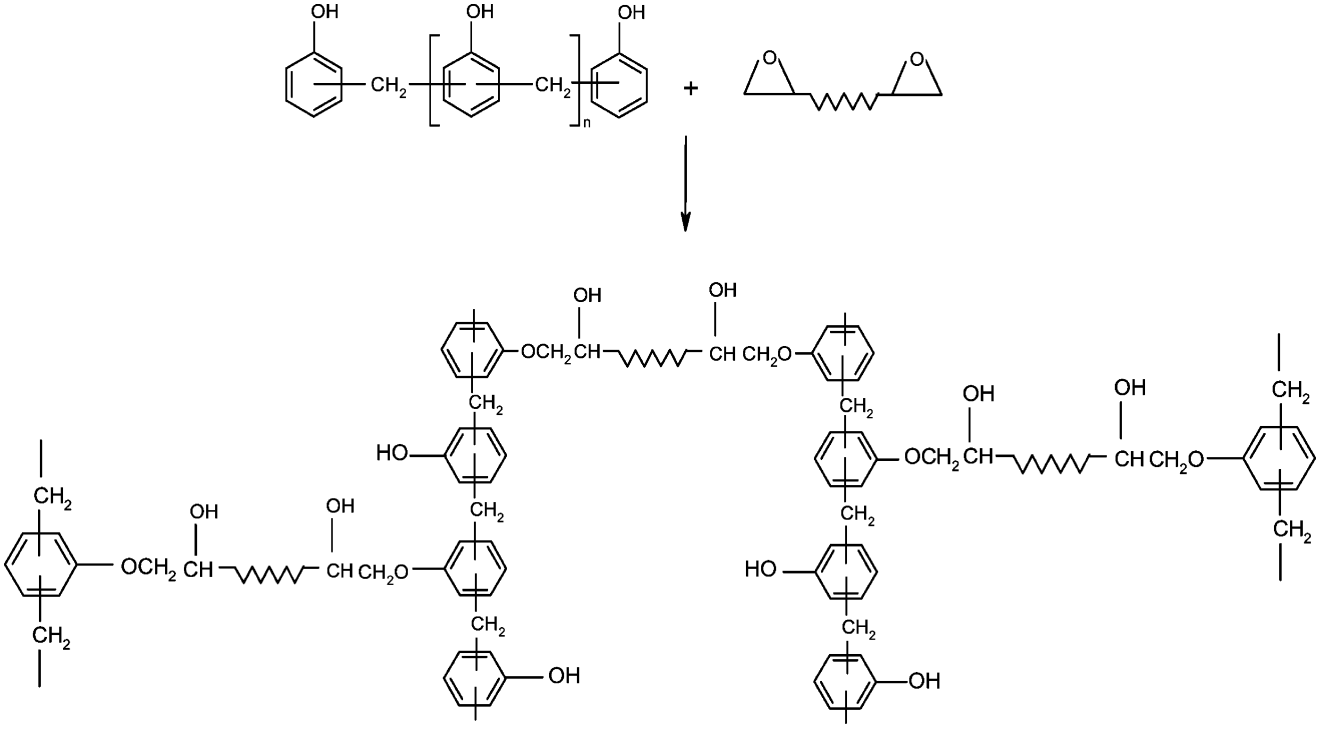


(12)

PHENOLIC RESIN NANOCOMPOSITES

General Background

The field of organic–inorganic hybrids or nanocomposites has attracted considerable attention as a method of enhancing polymer properties and extending the utility of polymers. The improvements occur through the use of molecular or nanoscale reinforcements rather than conventional particulate filled microcomposites. Among the organic–inorganic hybrids, researchers have focused mainly on the use of clay minerals as the polymer reinforcement. The clays are layered structures belonging to the general family of 2 : 1 layered silicate. The breakthrough for these materials occurred more than a decade ago when scientists from the



Scheme 3 Reaction of phenolic resins with epoxy.

Toyota Research Center reported unprecedented improvements in the mechanical properties of nylon–clay nanocomposites.[24] The Toyota Research group resolved the setback of the incompatibility of mixing between the hydrophilic-layered silicate and the organophilic polymers through chemical substitution on the surface of the clays. Typically, this is accomplished by exchanging the surface sodium cation with an organic cation, and thus producing organically modified layered silicate. As a result of the Toyota group's accomplishment, many studies on the preparation and characterization of various thermoplastic and thermoset polymer–layered silicate nanocomposites have been reported. All the nanocomposite systems reported showed new and improved properties when compared to their micro- and macro-composite counterparts and the neat polymers. The nanocomposites exhibit improved tensile strength and moduli, decreased thermal expansion coefficient, decreased gas permeability, enhanced ionic conductivity, and flammability.[24–29] Among phenolic resins, nanocomposite systems from novolacs have more reported successful results than resols.[30] One possible reason for this is that the novolac molecules have dimensions that fit easily between the silicate layers. Once the novolac molecules are in position between the silicate layers, this intercalated system can be reacted with a curing agent (e.g., hexamethylenetetramine) to give a cross-linked product. The control of exfoliation for this system is probably a function of the rate and extent of cure. For the resol system, the trisubstituted phenolic rings are larger than the phenolic rings of the novolac system. As a result, exfoliation of the silicate layers has probably begun before cure to a greater extent for the resol than for the novolac system. The distribution of the silicate layers in the resol system nanocomposite should be distinctly different from that for the novolac system nanocomposite. The control of exfoliation is also probably less for the resol system than for the novolac system. In other words, the resol system is probably more difficult to manage in the study of nanocomposites. Hence, the more easily managed novolac system has received more attention. Regardless, both systems should provide an interesting model for the molecular events in nanocomposite formation. The following two sections will present a summary of the studies that were done on phenolic resins reinforced with nanoparticles.

Novolac Nanocomposites

The earliest study cited for the preparation of novolac–layered silicate nanocomposites was reported by scientists from the Toyota Research Center.[31] Usuki et al. modified a sodium-type montmorillonite (MMT) clay with 4-aminophenol hydrochloride in water. The modified MMT was dispersed during the preparation of phenol and 37% formaldehyde solution in the presence of oxalic acid as catalyst. The prepared novolac–MMT nanocomposite was mixed with hexametylenetetramine as a curing agent and underwent press molding. The nanocomposite exhibited a heat distortion temperature of 210°C and a tensile modulus of 7.2 kg/m^2. These values were not specifically compared to those of other systems but the implication was that the nanosilicate improved the properties of the phenolic resin.

The thermoplastic nature of novolacs made it feasible to use melt intercalation to disperse nanoparticles into the polymer. The melt intercalation process involves heating the resin to a temperature just above its melting point, followed by dispersing the reinforcement in the molten resin by different mixing methods such as shear mixing. Lee and Giannelis were the first to report a preparation of a novolac–layered silicate nanocomposite via melt intercalation.[32] Choi, Chung, and Lee reported a more detailed study on preparation of novolac–layered silicate nanocomposites by melt intercalation.[30] They discussed the morphology and curing behaviors of these systems. They utilized two types of clays, natural MMT clay and synthetic fluorohectorite clay. The MMT clay was modified with dodecylammonium, octadecylammonium, and benzyldimethyloctadecylammonium salts and the fluorohectorite was modified with benzyldimethyloctadecylammonium salt. They also utilized the unmodified layered silicate for the purpose of comparison. From x-ray diffraction (XRD) data, they noted that the 12-carbon system widened the spacing between the silicate layers by 1.4 Å (12.2 Å for the pristine silicate layer spacing and 13.6 Å for the modified system spacing). The 18-carbon system widened this spacing by 6.1 Å (18.3 Å for this spacing). Attempts at melt intercalation by a novolac resin appeared very successful for the 18-carbon system. The spacing between the silicate layers increased by another 15.6 Å. Intercalation also appeared to have occurred for the pristine system (a further 3.3 Å increase in spacing). The 12-carbon system gave only a further 1.3 Å in spacing. The authors did not give any unifying rationale for these observations. However, it is possible that the micelle from the dodecylammonium salt distributes so as to occupy much of the space between the silicate layers. Hence, there is little room for the novolac resin. The octadecylammonium salt micelle may distribute so as to provide more open space for the resin. Upon cure with hexamethylenetetramine, the XRD data are interpreted to indicate that the spacing is maintained.

More recent work by Choi and coworkers involved study of octadecylammonium and the benzyldimethyloctadecylammonium salts of the previously discussed work and a bis(2-hydroxy ethyl)methyl tallow ammonium salt.[33] The tallow salt gave a slightly larger

increase in spacing between the silicate layers (19.4 Å) upon melt intercalation. The increase in spacing for the benzyl salt was approximately 18 Å in both studies. The authors examined the mechanical properties of these systems in comparison with the neat polymer. The octadecylammonium system showed no improvement up to 3 wt% of silicate, and then showed a decline in properties relative to the neat resin from 3–5 wt%. The other two modified systems showed significant improvement in mechanical properties relative to the neat polymer at all weight percents, but especially at 3 wt%. The tallow salt system gave the best mechanical properties.

Resol Nanocomposites

Byun, Choi, and Chung reported a method of preparation for resol–layered silicate nanocomposites by melt intercalation using unmodified Na-MMT and MMT modified with amino acids.[34] The amino acids utilized in Byun et al. study contained 3 carbons (β-alanine), 6 carbons (6-aminocaproic acid), and 12 carbons (12-aminododecanoic acid). Characterization by XRD and transmission electron microscopy (TEM) showed that all the prepared nanocomposites have partially exfoliated structures. Exfoliation of the layered silicate in the polymer is usually achieved through intercalation of the resin oligomers in the intergallery of the clay. Then, the intercalated oligomers undergo curing and the polymer molecules grow in size, which leads to complete disruption of the layered structures of the clay. However, in some polymer systems, including the one studied by Byun and coworkers, deintercalation of the polymer molecules from the intergallery of the layered silicate was observed to occur during the curing process. With the exception of the C6-MMT–resol system, Byun and coworkers observed deintercalation of the polymer to occur during the cure cycle for all the nanocomposite systems. Based on the findings about the degree of deintercalation, the C6-MMT was selected as the most compatible modified layered silicate with the resol polymer. The excellent compatibility between the C6-MMT and the resol prevented the deintercalation of the polymer from the intergallery of the layered silicate. At the initial stages of curing, the unmodified MMT–resol system showed a rapid expansion in the spacing between the silicate layers, which typically leads to exfoliation of these layered structures. This was attributed to the high compatibility between the hydrophilic unmodified MMT and the hydrophilic resol. However, at later stages of curing, this system showed the highest degree of deintercalation when compared to the other systems. This was attributed to the severe reduction in the compatibility between the resol and the unmodified MMT due to reduction of the methylol groups that are consumed during curing.

The mechanical properties of the C3, C6, and C12 nanocomposites were all significantly better than those of the neat phenolic resin, even if a very small amount of the silicate was used. Among the nanocomposites prepared, the organically modified MMT–resol systems showed better mechanical properties than those of the unmodified MMT–resol system. This improvement was attributed to the formation of an end-tethered structure due to the reaction of the carboxylic acid of the organic modifier with the methylol group of the phenolic resin. Thermogravimetric analysis reported by Byun and coworkers showed that the nanocomposite systems had similar thermal stability to that of the neat polymer.

Wang et al. reported preparation of resol nanocomposites by polymerizing phenol and formaldehyde in the presence of MMT modified with hydrochloric acid.[35] X-ray diffraction analysis for the acid-modified MMT–resol showed exfoliated structures for nanocomposites with low silicate contents (3 and 5 wt%) after 2.5 hr of curing time. However, in the XRD curve for the 10 wt% nanocomposite, a weak peak attributable to the (001) plane of MMT was observed. This suggested that complete exfoliation was not achieved in the 10 wt% nanocomposite. Both scanning electron microscopy and TEM analyses for the nanocomposite systems agreed with the XRD results. Wang et al. explained that exfoliation occurred in the acid-modified MMT–resol nanocomposites because of the presence of the protons in the intergallery of the MMT. The protons had a catalytic effect on the polymerization of the resol in the intergallery of the MMT, which resulted in the exfoliation of the layered structures. Dynamic mechanical analysis showed an approximately 60°C increase in the T_g of the nanocomposites in comparison to the neat resol. Similarly, a 53% improvement in the impact strength of the nanocomposites was reported.

CONCLUSIONS

Phenolic resins continue to be an important material at both the commercial and the research levels. These complex systems are fascinating not only because of their current usefulness, but also because of their potential to become even more useful materials of the future. The classical phenol–formaldehyde system is deceptively simple and lends itself to much variation depending on factors such as pH, molar ratio of reactants, preparation/cure temperature, and curing agents. The desire to enhance the properties of phenolic resins and expand the processing options has led to considerable work on modified-classical and nonclassical phenolic resins.

The modified-classical phenolic resins are particularly noteworthy for their effect on the mechanical and thermal properties of the cured resin. The processing of these systems is very similar to the classical systems. The nonclassical phenolic resins utilize phenol, but in many cases give a cured product with a chemical structure having little resemblance to the classical system. In future articles, it would be best to develop independent grouping for ''addition-cure phenolics resins.'' This redefinition of polymer types is not within the scope of this entry and instead, this interesting class of polymers has been viewed based on current designations.

The work on nanocomposites utilizing novolacs and resols can be considered preliminary and promising. The very subtle differences between novolacs and resols appear to have the potential to contribute to significantly different nanocomposite systems. The distinct separation of the silicate layers may provide the basis for the properties of nanocomposites prepared as novolacs or resols.

REFERENCES

1. Elias, H.G. *An Introduction to Plastics*; VCH: New York, 1993; 265–267.
2. Batzer, H.; Lohse, F. *Introduction to Macromolecular Chemistry*, 2nd Ed.; John Wiley & Sons: New York, 1979; 36–39.
3. Odian, G. *Principles of Polymerization*, 2nd Ed.; John Wiley & Sons: New York, 1981; 128–133, 139–140.
4. Asai, K.; Arai, H Phenolic Molding Composition. . U.S. Patent 6,716,907, Apr 6, 2004.
5. Bradford, I. Synthesis and Characterization of a Novel Brominated Phenolic Resin for Glass Fiber-Reinforced Composites. M.S. Thesis, Tuskegee University, Aug 2002.
6. Hirohata, T.; Misaki, T.; Yoshii, M. Bromine, chlorine and boron modified phenolic resins with excellent flame retardance and thermal stability. Zairyo (English translation) **1987**, *36* (401), 184–188.
7. Abdalla, M.O.; Ludwick, A.; Mitchell, T. Boron-modified phenolic resins for high performance applications. Polymer **2003**, *44* (24), 7353–7359.
8. Knop, A.; Pilato, L.A. Reaction mechanisms. In *Phenolic Resins: Chemistry, Applications, Performance and Future Direction*; Springer-Verlag: Berlin, 1985.
9. Pizzi, A.; Pasch, H.; Simon, C.; Rode, K. Structure of resorcinol, phenol, and furan resins by MALDI-TOF mass spectrometry and ^{13}C NMR. J. Appl. Polym. Sci. **2004**, *92*, 2665–2674.
10. Bohumila, Z.; Kwok, T.; Ruben, S Borate Modified Phenolic Resin for Insulation Materials. U.S. Patent 2004068083, Aug 4, 2004.
11. Chiu, S.T.; Shelast, C.M.; Lam, E.K U.S. Patent 6,608,162, Aug 19, 2003.
12. Nair, C.P.R. Advances in addition-cure phenolic resins. Prog. Polym. Sci. **2004**, *29*, 401–498.
13. Naito, S.; Saito, Y.; Shiomi, H. (Sumitomo Bakelite Co. Ltd). Thermosetting resin compositions. Japan Kokai Tokkyo Koho, JP. 01, 289,21 (89,289821), 1989; Chem. Abstr. **1990**, *112*, 159870.
14. Yan, Y.H.; Shi, X.M.; Lin, J.G.; Zhao, T.; Yu, Y.Z. Thermosetting resin system based on novolac and bismaleimide for resin transfer molding. J. Appl. Polym. Sci. **2002**, *83* (8), 1651–1657.
15. Stenzenberger, K.D.; Komig, P. New functionalized poly(aryleneetherketones) and their use as modifiers for bismaleimide resins. High Perform. Polym. **1993**, *5*, 123–137.
16. Gu, A.J.; Liang, G.Z.; Lan, L.W. Modification of polyaraalkyl phenolic resin and its copolymer with bismaleimide. J. Appl. Polym. Sci. **1996**, *59* (6), 975–990.
17. http://composite.about.com/library/PR/2001/blustm3.htm (accessed Sep 2004).
18. Tyberg, C.S.; Sankarapandian, M.; Bears, K.; Shih, P.; Loos, A.C.; Dillard, D.; McGrath, J.E.; Riffle, J.S.; Sorathia, V. Tough, void free, flame retardant phenolic matrix materials. Constr. Build. Mater. **1999**, *13*, 343–353.
19. Tyberg, C.S.; Sankarapandian, M.; Bears, K.; Shih, P.; Loos, A.C.; Dillard, D.; McGrath, J.E.; Riffle, J.S.; Sorathia, V. Structure-property relationship of void-free phenolic-epoxy materials. Polymer **2000**, *41* (13), 5053–5062.
20. Dordick, J.S.; Marletta, M.A.; Klibanov, A.M. Polymerization of phenols catalyzed by peroxidase in nonaqueous media. Biotechnol. Bioeng. **1987**, *30*, 31–36.
21. Kobayashi, S.; Shoda, S.; Uyama, H. Enzymatic polymerization and oligomerization. In *Synthesis/Polymer Engineering*; Adv. Polym. Sci. **1995**, *121*, 1–30.
22. Akkara, J.A.; Kaplan, D.L.; Vijay, T.J.; Tripathy, S.K. Enzyme-catalyzed polymerization. In *Polymeric Materials Encyclopedia*; Salamone, J.C., Ed.; CRC Press: Boca Raton, FL, 1996; Vol. 3, 2115–2125.
23. Kumar, V.; Dhawan, A.; Wang, X.; Parmar, V.S.; Samuelon, L.A.; Kumar, J.; Cholli, A.L. Post-coupling of enzymatically synthesized phenolic polymers for sensing hazardous and explosive type species. Polym. Prepr. **2003**, *44* (1), 1224–1225.
24. Kojima, Y.; Usuki, A.; Kawasumi, M.; Okada, A.; Fukushima, Y.; Kurauchi, T.; Kamigaito, O.

Mechanical properties of nylon 6-clay hybrid. J. Mater. Res. **1993**, *8* (5), 1185–1189.
25. Messersmith, P.B.; Giannelis, E.P. Synthesis and barrier properties of poly(ε-caprolactone)-layered silicate nanocomposites. J. Polym. Sci. Polym. Chem. **1995**, *33* (7), 1047.
26. Burnside, S.D., Giannelis, E.P. Synthesis and characterization of elastomeric nanocomposites. Proc. ACS Div. Polym. Mater. Sci. Eng. **1995**, *73*, 322.
27. Messersmith, P.B.; Giannelis, E.P. Synthesis, characterization, and properties of molecularly dispersed polymer-silicate nanocomposites. Chem. Mater. **1994**, *6*, 1719.
28. Vaia, R.A.; Vasudevan, S.; Krawiec, W.; Scanlon, L.G.; Giannelis, E.P. New polymer electrolyte nanocomposites: melt intercalation of poly(ethylene oxide) in mica-type silicates. Adv. Mater. **1995**, *7*, 154.
29. Gilman, J.W. Flammability and thermal stability studies of polymer layered-silicate (clay) nanocomposites. Appl. Clay Sci. **1999**, *15*, 31.
30. Choi, M.H.; Chung, I.J.; Lee, J.D. Morphology and curing behaviors of phenolic resin–layered silicate nanocomposites prepared by melt intercalation. Chem. Mater. **2000**, *12*, 2977–2983.
31. Usuki, A.; Mizutani, T.; Fukushima, Y.; Fujimito, M.; Fukumori, K.; Kojima, Y.; Sato, N.; Kurauchi, T.; Kamigaito, O Composite Material Containing a Layered Silicate. U.S. Patent 4,889,885, Dec 6, 1989.
32. Lee, J.D.; Giannelis, E.P. Synthesis and characterization of unsaturated polyester and phenolic resin nanocomposites. Polym. Prepr. Div. Polym. Chem. **1997**, *38* (2), 688–689.
33. Choi, M.H.; In, J. Mechanical and thermal properties of phenolic resin–layered silicate nanocomposites synthesized by melt intercalation. J. Appl. Polym. Sci. **2003**, *90* (9), 2316–2321.
34. Byun, H.Y.; Choi, M.H.; Chung, I.J. Synthesis and characterization of resol type phenolic resin/layered silicate nanocomposites. Chem. Mater. **2001**, *13*, 4221–4226.
35. Wang, H.; Zhao, T.; Yan, Y.; Yu, Y. Synthesis of resol–layered silicate nanocomposites by reaction exfoliation with acid-modified montmorillonite. J. Appl. Polym. Sci. **2004**, *92*, 791–797.

Photodegradation of Polymers

J. R. White
School of Chemical Engineering and Advanced Materials, University of Newcastle Upon Tyne, Newcastle Upon Tyne, U.K.

INTRODUCTION

Photodegradation of polymers is generally accepted as the main factor that limits their use in outdoor applications. It is caused by the ultraviolet (UV) component of solar radiation and is a serious problem in the sunniest climates. Artificial sources of UV also cause photodegradation and may disqualify polymers from use in certain applications. The sensitivity of polymers to artificial UV radiation has led to the development of accelerated aging laboratory equipment for assessing the weathering properties of polymers. Radiations such as γ-radiation that possesses greater energy than UV can also produce photodegradation and this may restrict the choice of materials used in certain nuclear installations. Photodegradation caused by γ-radiation is also of concern in medical implants such as acetabular cups used in hip prostheses because γ-radiation is an option for sterilization prior to inserting the implant. This is rather specialized, however, and the fundamental chemical and physical consequences of exposure to γ-radiation are not very different from those obtained with UV radiation; hence this entry concentrates exclusively on UV photodegradation. Most research into photodegradation has been conducted in atmospheric air because this relates to the majority of practical cases and is reflected on this entry. Although the weathering of polymers remains the focus of much research, the subject is well established, and for a balanced view of the scientific background and practical issues, the review provided by Davis and Sims in 1983 remains unsurpassed.[1]

The chemical mechanisms of degradation are dealt with first, followed by the physical consequences of the chemical changes. This leads to a consideration of the changes in engineering properties. Procedures for testing the photodegradability and weather resistance of polymers are then discussed. Much research has been directed into the development of photostabilizers that prolong the lifetime of polymers and this important and highly successful enterprise is dealt with next. Finally, some consideration is given to the recycling of photodegraded polymers.

CHEMICAL MECHANISMS OF PHOTODEGRADATION

Degradation Reactions

Polymer photodegradation proceeds by a radical chain process initiated either by dissociation caused by the collision of a UV photon with a polymer molecule or as the result of some impurity present, such as a polymerization catalyst.[2] Visible light photons and UV photons with wavelengths just less than the visible band are generally not energetic enough to cause the formation of radicals in polymers. The Earth's atmosphere scatters or absorbs all photons with wavelengths below about 290 nm, but photons with wavelengths in the range 290–340 nm are quite effective in forming radicals in many polymers. Lower wavelength (higher energy) photons may cause additional reactions but are not considered in this entry.

The chain reaction has four main stages:

1. *Initiation*, in which the long chain polymer molecule, PH, is converted into a radical:

$$PH + O_2 \rightarrow P^{\bullet} + HO_2^{\bullet} \quad (1)$$

There is a high probability that the radical will be located somewhere along the chain rather than at a terminal site.

2. *Propagation* also involves reaction with oxygen:

$$P^{\bullet} + O_2 \rightarrow POO^{\bullet} \quad (2)$$

$$POO^{\bullet} + PH \rightarrow POOH + P^{\bullet} \quad (3)$$

3. *Chain branching* involves hydroperoxides formed in reaction (2):

$$POOH \rightarrow PO^{\bullet} + HO^{\bullet} \quad (4)$$

$$2POOH \rightarrow POO^{\bullet} + PO^{\bullet} + H_2O \quad (5)$$

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007895

Radicals formed in (4) and (5) re-enter the propagation phase via the following reactions:

$$PO^{\bullet} + PH \rightarrow POH + P^{\bullet} \quad (6)$$

$$HO^{\bullet} + PH \rightarrow P^{\bullet} + H_2O \quad (7)$$

4. *Termination* involves the reaction of pairs of radicals:

$$P^{\bullet} + P^{\bullet} \rightarrow \text{inactive products} \quad (8)$$

$$P^{\bullet} + POO^{\bullet} \rightarrow \text{inactive products} \quad (9)$$

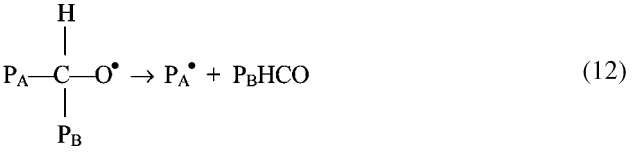
$$POO^{\bullet} + POO^{\bullet} \rightarrow \text{inactive products} \quad (10)$$

The reactions may be catalyzed by trace metal impurities such as polymerization catalyst residues or contaminants from processing machinery. The reaction scheme is similar to that in thermal oxidation, differing only in the method of initiation and in the relative reaction rates (which are likely to be quite different at the higher temperatures normally required for thermal initiation).

In solid polymers, the propagation and termination reactions are controlled by diffusion. Large macroradicals have very restricted motion in the solid phase, either because they are restrained by entanglements in the case of amorphous polymers, or are tethered by segments within the crystal phase in semicrystalline polymers. Spatial propagation of degradation can proceed by way of translational diffusion of low molecular weight radicals formed by the decomposition of peroxide or alkoxyl macroradicals. Decomposition of a peroxide radical, $POO^{\bullet}$, creates a small radical, $r^{\bullet}$, which then diffuses to a new site where it reacts with a hydroperoxide molecule to form a new macroradical in a radical exchange reaction:

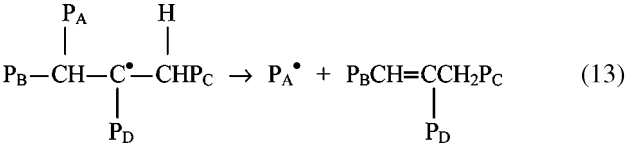
$$r^{\bullet} + POOH \rightarrow rH + POO^{\bullet} \quad (11)$$

Consequently, the peroxide radical effectively moves from the initiation site by a distance that depends on the diffusion rate of the small radical (and hence on the temperature, etc.) and on the reaction rate for reaction (11). In reactions (1)–(11), the long chain nature of the molecules is preserved almost unchanged. If the termination reaction (8) is by cross-linking rather than by disproportionation, then a significant change in the molecular size occurs and this will cause a significant change in the properties of the polymer. Analysis of photodegraded polymers usually shows that the predominant effect is chain scission, however, and this is not specifically covered by reactions (1)–(11). Scission releases chain entanglements in amorphous polymers and breaks tie molecules in semicrystalline polymers and, hence, has a very strong influence on the properties. Chain scission occurs through the decomposition of unstable species, especially hydroperoxides, and may be caused by UV radiation. Decomposition occurs by chain scission adjacent to the hydroperoxide group, at the position of the original site of the radical created in reaction (1). Hydroperoxides produced by reaction (3) or by other means can be decomposed by UV photons with wavelength below 360 nm giving a $PO^{\bullet}$ radical, as shown in reaction (4). β-Scission of the alkoxy radical produces two smaller molecules, one of which is a radical and may participate in further reactions of the kind shown earlier:



(12)

Alkyl radical decomposition produces a similar result:



(13)

Decomposition of hydroperoxides is a key feature in the degradation of polyolefins and may include reactions of the kind:

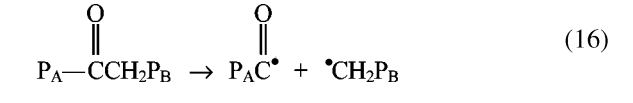
$$POOH \rightarrow P^{\bullet} + {}^{\bullet}OOH \quad (14)$$

$$POOH + PH \rightarrow PO^{\bullet} + P^{\bullet} + H_2O \quad (15)$$

These lead to chain scission through decomposition of radicals as in reactions (12) and (13).

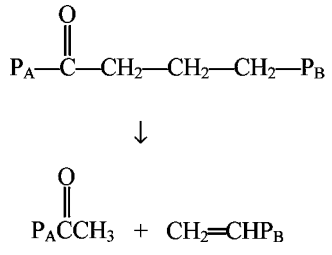
Carbonyl groups are frequently produced during polymer degradation making the material vulnerable to further deterioration because they are photolabile. Aldehyde and ketone carbonyl groups are common defects produced during polymer processing and may partake in photolytic decomposition reactions in the fabricated article, as in the following example for aliphatic ketones:

- *Norrish type I*:



(16)

- *Norrish type II*:



(17)

Kinetics of Polymer Photodegradation

The production of carbonyl groups can be used to determine the rate of oxidation caused by UV exposure. Fig. 1 shows the progressive increase in carbonyl group concentration measured using Fourier transform infrared (FTIR) in a series of polyethylene samples exposed to UV irradiation in the laboratory. It is frequently observed that there is an induction period during which there is no indication of change in many of the measures of degradation, including carbonyl group development. In Fig. 1, this is shown by the unpigmented polyethylene sample and also by the one containing the rutile form of TiO_2, a pigment that can provide some protection against UV irradiation. Both these samples also show evidence for autoacceleration whereby the rate of reaction increases progressively. This is hardly surprising in view of the role of reaction products in promoting further reactions, discussed in the earlier section. An alternative explanation for the acceleration in rate has been given by George et al.,[3] who proposed that infectious spreading of reaction is responsible. In their view, active species (e.g., free radicals) may migrate a finite distance before reacting and setting up a new zone of reaction. The third sample for which data are given in Fig. 1 is polyethylene pigmented with the anatase form of TiO_2. Anatase is a photocatalyst and is responsible for the different kinetic behavior. Finally, it is noted that carbonyl groups are both created and destroyed during photo-oxidation and that conclusions concerning the progress of degradation of the polymer should not be made based on carbonyl group accumulation alone.

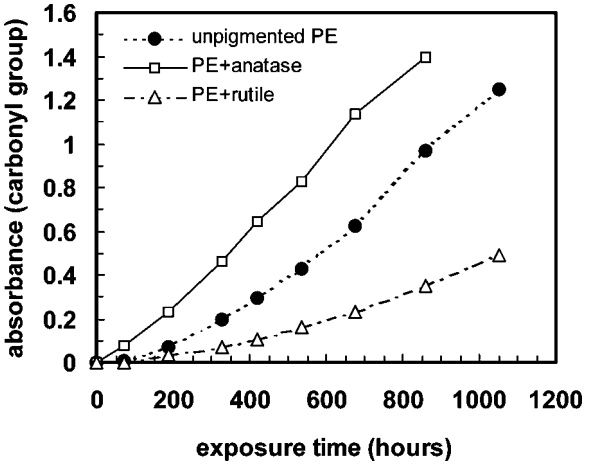


Fig. 1 Carbonyl group development in polymer samples exposed to UV illumination. The materials were an unpigmented polyethylene sample and similar samples pigmented with 5 phr TiO_2 pigment in the form of (i) anatase and (ii) rutile. (Data from Jin, C.Q. Ph.D. thesis, University of Newcastle upon Tyne, 2004. See also Jin, C.; Christensen, P.A.; Egerton, T.A.; White, J.R. Effect of anisotropy on photomechanical oxidation of polyethylene. Polymer **2003**, *44*, 5969–5981.)

A full kinetic analysis of photodegradation is very difficult because of the large number of reactions and reagents involved. Many of the reagents are both created and destroyed by reaction, and keeping track of their concentrations in a computed analysis is a formidable challenge. A similar problem is present in thermal degradation of polymers. Despite the difficulties, significant contributions have been made by Audouin et al.[4] and Gillen, Wise, and Cough.[5] This approach has been particularly valuable in the study of the depth profile of degradation that develops when reaction is oxygen diffusion limited, as discussed later.

Molecular Degradation Products

The feature most responsible for the special properties associated with polymers is their molecular weight. The molecular weight distribution (MWD) is an important characteristic to track the progress of photodegradation. Fig. 2 shows the MWD for a low-density polyethylene (LDPE) obtained using gel permeation chromatography (GPC) in the unexposed state and after laboratory UV exposure. The unexposed polymer had a fairly broad MWD. The MWD for the material at the exposed surface shifted very significantly to the left (lower molecular weight) showing that molecular chain scission was dominant. Smaller leftward shifts are shown for material deeper in from the exposed surface. In addition, a high molecular weight tail developed with a small but significant fraction of material having higher molecular mass than the largest molecular mass fraction in the unexposed sample. This is attributed to cross-linking.

Shyichuk has developed a Monte Carlo method of computer-aided simulation of the MWDs of degraded polymer derived from the MWD of the unexposed polymer and assuming scission and cross-linking are random events. The results of trial scission cross-linking concentrations are compared with measured MWDs for the exposed samples using the sum of the squares of the

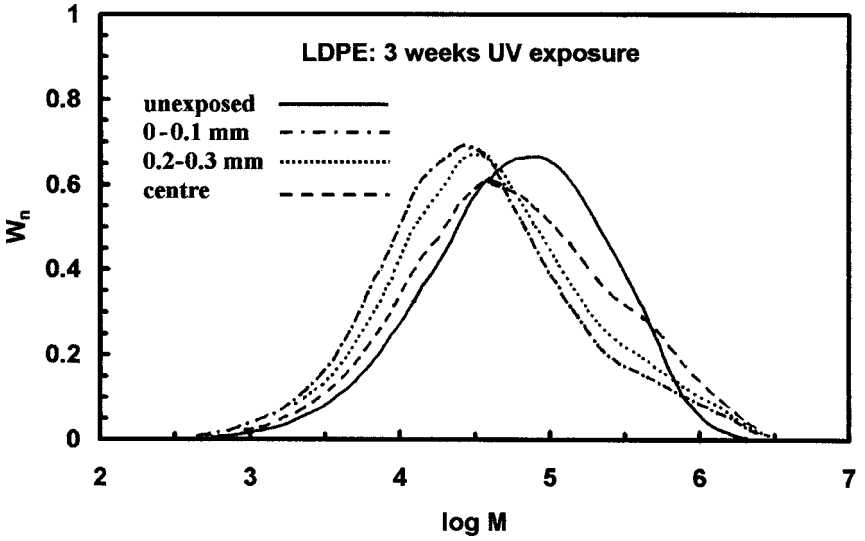


Fig. 2 Molecular weight distributions obtained from GPC analyses of layers of 0.1 mm thickness taken at different depths from the exposed surface of an LDPE bar after 3 weeks of UV exposure. The MWD for unexposed material is shown for comparison. (More details on the study from which these data are taken are given in Craig, I.H.; White, J.R.; Shyichuk, A.V.; Syrotynska, I. Photo-induced scission and cross-linking in LDPE, LLDPE, and HDPE. Polym. Eng. Sci. **2005**, *45*, 576–587.)

residuals at the data points to assess the closeness of fit and to determine the best values.[6] Fig. 3 shows results from this kind of analysis for polypropylene samples exposed in the laboratory for two different times. Samples were taken at different depths from the exposed surface of 3 mm thick bars. Scission dominated at all depths.

In addition to the macromolecular fragments, many small molecule species are formed during photodegradation. The smallest molecules migrate to the surface and are lost by a variety of mechanisms including volatilization. Although they no longer influence the properties of the material that remains behind, they can be used to monitor the progress of photodegradation in carefully controlled laboratory tests. Christensen, et al. have developed a method in which the sample under test is exposed to UV in a specially designed cell that allows simultaneous interrogation of the gaseous phase in an FTIR spectrometer.[7] Carbon dioxide evolved during photodegradation accumulates in the cell and the concentration is monitored continuously. This method was first applied to the study of paints[7] and its sensitivity can produce an assessment of the UV-oxidizability of a sample in a fraction of the time required using the carbonyl analysis method. This is shown by the results given in Fig. 4 for polyethylene samples. Figs. 1 and 4 are based on data obtained with the same

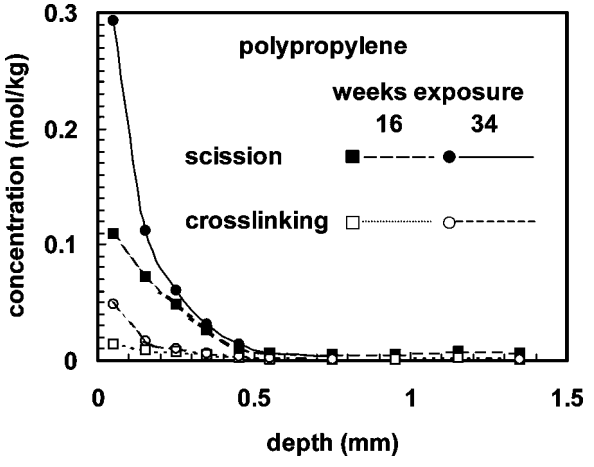


Fig. 3 Scission and crosslinking concentrations produced in polypropylene after 16 weeks and 34 weeks of UV exposure in the laboratory. Samples taken at different depths from the exposed surface. Each sample was 0.1 mm deep: the corresponding result is plotted at the midpoint of the layer. (More details in Shyichuk, A.V.; Turton, T.J.; White, J.R.; Syrotynska, I.D. Different degradability of two similar polypropylenes as revealed by macromolecule scission and crosslinking. Polym. Degrad. Stab. **2004**, *86,* 377–383.)

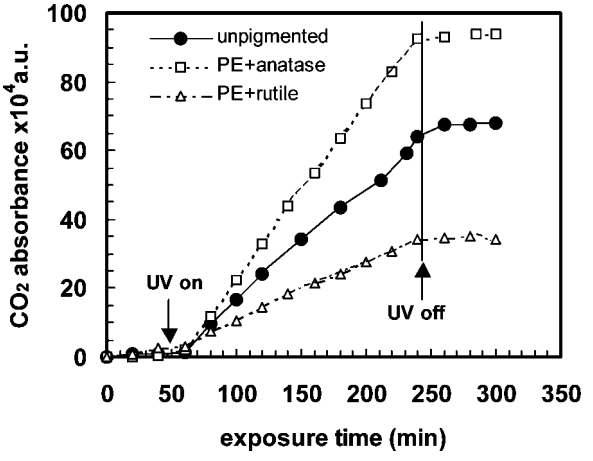


Fig. 4 Accumulated carbon dioxide emitted from polymer samples exposed to UV illumination from an unpigmented polyethylene sample and similar samples pigmented with 5 phr TiO_2 pigment in the form of (i) anatase and (ii) rutile. (Data from Jin, C.Q. Ph.D. thesis, University of Newcastle upon Tyne, 2004.)

materials and it is evident that the same ranking is obtained. Fig. 4 shows that a significant carbon dioxide signal was obtained almost as soon as the UV illumination was switched on: there was no apparent induction time. The rate of accumulation of CO_2 reduced almost to zero on switching off the illumination. The potential of this technique is not only for rapid assessment, it also permits convenient arrangement of experiments to test the effect of changing the atmosphere around the sample (especially by changing the oxygen concentration) or changing the spectral distribution in the illumination.

Environmental Effects

It is often observed that photodegradation of polymers occurs more rapidly in humid air than dry air. For this reason, tests should be conducted in controlled-humidity enclosures. Some popular commercial UV exposure cabinets include periodic water spray or temperature cycling that induces condensation of water onto the sample surface. The effects caused by this treatment are not confined to altering the reaction chemistry scheme and may include modifying the residual stress distribution because of chilling the surface. Pollutants, especially sulfur dioxide and oxides of nitrogen, may accelerate polymer degradation in the presence of UV. Ozone can attack polymers, especially if unsaturated. It is so aggressive toward rubber that perishing of rubber under tensile stress is often attributed to ozonation and the effect of photodegradation is disregarded as a possible cause. However, Maillo and White provided evidence for very significant photo-oxidation under these conditions.[8] They used a mask to partially cover the surface of rubber samples while being exposed to UV in the stretched state and showed that visible degradation was almost absent behind the mask, whereas in the illuminated zone immediately adjacent to the edge of the mask very substantial degradation occurred (Fig. 5).

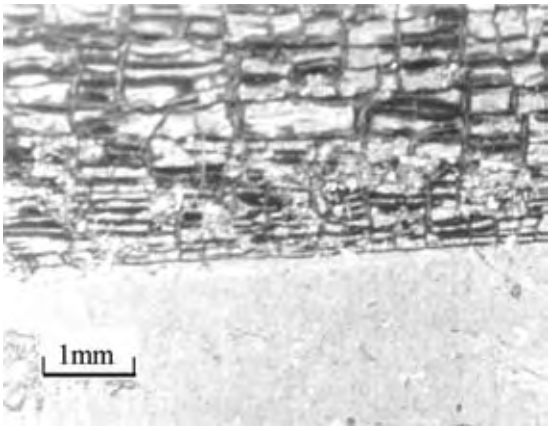


Fig. 5 Surface of blend of natural rubber and butadiene rubber (45 : 55) after UV exposure while under tensile stress (applied parallel to the vertical). The lower part of the sample was shielded from the UV illumination and suffered very much less visible degradation than the upper zone. (See also Ref.[8].)

PHYSICAL ASPECTS OF PHOTODEGRADATION

Oxygen Diffusion Limited Degradation: Depth Profiling

When many polymers are exposed to UV levels similar to those on a sunny day near to the equator, photodegradation reactions proceed very rapidly and the rate of oxygen consumption is very high. With some polymers, oxygen diffusing into the material from the exposed surface is consumed by reaction before it can penetrate very far. Therefore, under a period of sustained UV exposure, any oxygen that had accumulated in the interior of the sample prior to exposure is rapidly consumed and the reaction rate in the interior then becomes very slow because of oxygen starvation. Because of this, a strong degradation gradient develops, with the material at the surface becoming extensively degraded,[9] while that at depths of the order of 1 mm may be relatively unchanged. Examples of this can be seen in Figs. 2 and 3. If characterization measurements are conducted across the whole sample section,

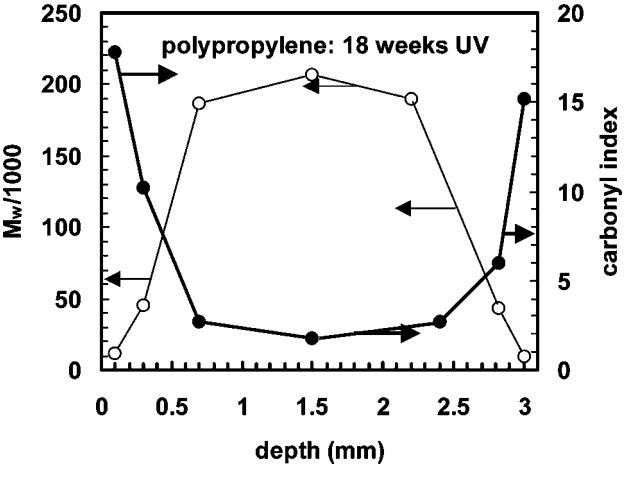


Fig. 6 Depth profiling in 3 mm thick polypropylene bar exposed to UV in the laboratory for 18 weeks. The exposed surface was at depth = 0. The steep gradients for both the carbonyl index and the weight average molecular weight at the back face indicate that oxygen starvation in the interior of the bar is the dominant factor rather than UV intensity. (Data from Rabello, M.S. Ph.D. thesis, University of Newcastle upon Tyne, 1996.)

through to the side that received no direct illumination (back face), it is sometimes observed that the degradation near to the back face is almost as advanced as that near to the exposed face (Fig. 6).[10–12] This can occur even when the UV levels in the interior and at the back face are very much less than those at the illuminated face because of absorption and scattering in the sample, a significant effect in polypropylene.[12] Attempts to take into account diffusion limited effects in models for the kinetics of photodegradation have reproduced the observed depth profiles fairly faithfully.[4,5,13]

Chemicrystallization

When a crystalline polymer is photodegraded, the oxidation occurs almost exclusively in the amorphous phase because oxygen can diffuse relatively freely through it but does not enter the crystal phase to a significant degree. When chain scission occurs, molecule segments that were previously relatively immobile because of entanglements or because of their attachment to a neighboring crystal become free to move. Some of them may rearrange and add to the growth faces of the crystals (or may even nucleate new crystals). Unless the photodegradation is very advanced, no damage accrues in the crystal phase and there is an increase in the crystallinity of the material. This is an example of chemicrystallization. An illustration of this phenomenon is given in Fig. 7, which shows data obtained for a linear low-density polyethylene in the form of 3 mm thick bars. Crystallinity measurements were made using differential scanning calorimetry on samples taken at different depths from the surface of an unirradiated bar and one exposed to UV in the laboratory for 6 weeks. Near the exposed surface the crystallinity in this material increased by up to 6%, a very substantial change, in view of the sensitivity of polymer properties to crystallinity. The change in the center of the 3 mm thick bar was very small, another example of oxygen diffusion-limited reaction. In Fig. 7, there appears to be a small drop in crystallinity after UV exposure but this is probably because of a small sample-to-sample variation and/or experimental error. The depth profile in crystallinity will contribute to the depth profile in other properties, such as density.

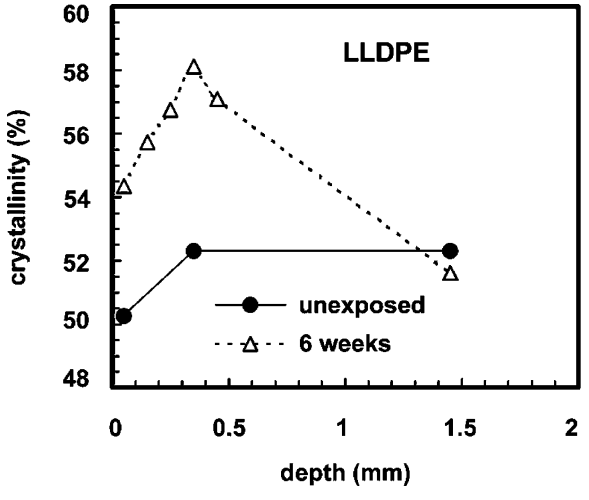


Fig. 7 Changes in crystallinity near to the surface of a linear lowdensity polyethylene bar exposed to UV in the laboratory for 6 weeks. Samples taken at different depths from the exposed surface. Each sample was 0.1 mm deep: the corresponding result is plotted at the midpoint of the layer. (More details in Craig, I.H.; White, J.R. Crystallization and chemicrystallization of recycled photo-degraded polyethylenes. Polym. Eng. Sci. **2005**, *45*, 588–595.)

ENGINEERING PROPERTIES—CONSEQUENCES OF PHOTODEGRADATION

Mechanical and Fracture Properties

The molecular changes that occur as a consequence of photodegradation almost always lead to a reduction in both strength and toughness of the polymer article. Entanglements in the amorphous phase of the polymer are essential for the mechanical integrity and obstruct the passage of a crack, and it is inevitable that their release by photoinduced chain scission will reduce the strength and the toughness of the material. It might be expected that photoinduced cross-linking would have the opposite effect, but it is generally observed that it also leads to embrittlement, presumably because it restricts molecular motion and reduces the range of deformation mechanisms available to produce toughening. The loss of toughness is sometimes very rapid and devastating. Polypropylene, normally a quite tough polymer, can effectively lose all toughness within a week of exposure to solar radiation in an equatorial region if it does not contain photostabilizers.

When the UV intensity is sufficient to produce depth profiling of the kind discussed earlier, it is the surface that becomes embrittled. Cracks nucleate and grow easily in the surface layer and rapidly reach the less degraded material in the interior (Fig. 8). If the embrittled zone is of sufficient depth, then the surface crack will constitute a critical stress-concentrating flaw and the crack will propagate into and through the rest of the section when sufficient stress is applied. This behavior follows closely the rules of fracture mechanics. After prolonged exposure, the surface may become so fragile that it forms multiple fractures (that may mutually unload) and/or it may have insufficient mechanical integrity to transmit stress into the underlying relatively undegraded material. At this point, it may

Fig. 8 Surface of polypropylene bar after 80 weeks of UV exposure followed by uniaxial tensile test. Bar axis is vertical. Cracks formed in degraded layer, perpendicular to stress axis. Degraded surface is very fragile and easily parted from underlying material that is less degraded. Degradation tends to be greatest nearer to the corner, possibly because oxygen can diffuse into the surface layers from two adjacent surfaces. (Courtesy of T.J. Turton. See also Refs.[16,24].)

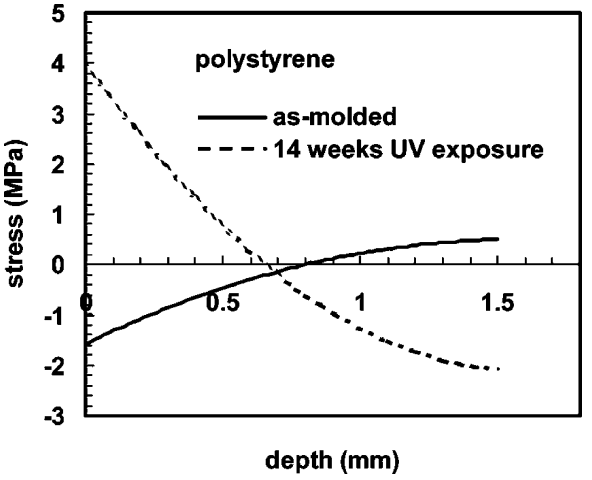


Fig. 9 Residual stress distributions in injection molded polystyrene bars in the as-molded state and after 14 weeks of ultraviolet (UV) exposure in the laboratory. The bars were approximately 3 mm thick and the distributions are shown as a function of depth from the (exposed) surface. (See Ref. [18])

be possible to observe some recovery of strength and toughness because the relatively undegraded material in the interior retains its strength and toughness, and supports the applied load.[14,15] Such recovery can be observed during a series of carefully controlled laboratory tests but is not of any practical value because, while in service, the component will fail as soon as it experiences a critical load and this is likely to happen when strength is near to the minimum and the component will not survive to reach the''recovered'' state.

Inspection of fracture surfaces of photodegraded polymers often reveals the presence of a smooth brittle zone near to the surface and a much more ductile fracture in the interior. The brittle zone depth often corresponds to that of the highly degraded depth obtained by FTIR or molecular mass measurements.[16]

Residual Stresses

Polymer articles made by injection molding, or any other molding operation in which a large temperature gradient exists while the material within the molding is solidifying, contain residual stresses.[17] When the molding is made from a hot melt that cools inside a cold mold, the stress in the interior is tensile because of thermal shrinkage of the core material as the molding cools; this is resisted by the previously solidified skin, which is then under compression. When photodegradation occurs, the material near the surface usually shrinks. In the case of a semicrystalline polymer, this is the result of chemicrystallization, the crystal density being significantly greater than the amorphous density in most polymers. A similar phenomenon has been observed to occur in polystyrene (Fig. 9 and Ref.[18]) and has been attributed to densification of the photodegraded material near to the surface caused by accelerated aging.[19] Similar changes in residual stresses have been observed in samples exposed to natural weathering.[20,21]

There are several reasons to be concerned by the changes in residual stress promoted by photodegradation. Flaws from which cracks grow and cause the ultimate failure of components (whatever material they are made from) are more likely to occur at the surface than in the interior. This is especially unfortunate because of the ease with which a critical strain can be applied at the surface while bending. The compressive stresses that are normally present at the surface in polymer moldings in the as-molded state discourage the formation and propagation of cracks and, therefore, toughen the material. Their loss is consequently detrimental and this is exacerbated if the stresses are reversed, becoming tensile at the surface, especially if the material at this position has become embrittled by photodegradation. If the changes in residual stress occur preferentially at one surface (e.g., the exposed surface), then the stresses through the section become imbalanced and the molding distorts; in some applications this will be unacceptable.

PHOTOSTABILIZATION

Most commercial polymers contain thermal stabilizers to protect against oxidation during processing at elevated

temperature, and many contain photostabilizers that can produce quite spectacular improvements in their lifetime in aggressive UV environments.[22] Protection against photo-oxidation can be provided by UV screening, UV absorption, excited state deactivation, free radical scavenging, and hydroperoxide decomposition.[22,23] UV screening may be provided by a reflective coating or by including pigments such as carbon black or TiO_2 (which are often present to provide other desirable properties).

Photodegradation is confined to the surface regions in the presence of such additives because UV penetration is restricted to very short distances. Some grades of TiO_2 photosensitize the material and must be used with caution. Ultraviolet absorption can be provided by additives such as hydroxybenzophenones, hydroxyphenylbenzotriazoles, hydroxyphenyltriazines, and derivatives of phenyl salicylates that are (almost 100%) transparent to visible light and do not alter the appearance of the product while having transmittance close to zero for all wavelengths below about 400 nm. The absorber must be able to dissipate the energy absorbed without damaging the neighboring polymer; as with screens, the presence of UV absorbers causes the UV intensity to drop rapidly as it penetrates into the material and degradation is concentrated even more sharply near the surface than is the case when no absorber is present. Excited state deactivation (or quenching) provides a method for the dissipation of energy absorbed by a polymer chromophore that would otherwise cause chain scission. As with UV absorbers, the quencher must be able to dissipate the acquired energy.

Radical scavengers limit photodegradation damage by breaking the oxidation chain. In the chain-breaking donor mechanism, a hydrogen atom is donated by (usually) a hindered phenol to an alkoxyperoxyl radical to form a hydroperoxide. Alkyl radicals are rendered ineffective by the chain-breaking acceptor mechanism for which suitable stabilizers include free radicals such as nitroxyls and phenoxyls that are relatively stable and do not themselves initiate reactions with the undamaged polymer. It is believed that some stabilizers, including the hindered amines, act via cyclic reaction paths so that they are regenerated and can be effective at lower concentrations or for longer periods of time than would otherwise be expected. This type of stabilizer can have a quite different effect on the depth profile to that produced by UV absorbers. The very much reduced reaction rate results in much slower oxygen consumption and there is ample opportunity for oxygen to diffuse into the interior of thick-walled articles. Therefore, oxidation is not diffusion-rate limited and proceeds at almost the same rate at all depths.[24] Thus, even though the reaction is slow, the rate of degradation at large depths from the surface is much greater than that in the absence of stabilizer.[24]

Hydroperoxides are potent photoinitiators and their decomposition into inactive products is an important method of stabilization in polyolefins for it prevents generation of radicals by reactions such as (4), (5), (14), and (15). Decomposition of hydroperoxides can be achieved by a reaction with phosphite esters or nickel chelates or by a catalytic action by a range of compounds including dithiocarbamates and mercaptobenzothiazoles. Decomposition of hydroperoxides formed during processing is important for the stabilization of PVC.

The distribution of stabilizers in a polymer article has a significant influence on their effectiveness. They have greatest influence when located at the surface. If they are lost by volatilization, surface migration, or consumed by reaction, it is an advantage if stabilizer can be replenished at the surface by diffusion from the interior, favoring low molecular mass stabilizers. Conversely, higher molecular mass molecules are less likely to be lost by volatilization or migration. The effect of stabilizer molecular mass has been studied by Gugumus.[25] Further immobilization of the stabilizer can be achieved by copolymerizing it with the polymer or by photografting to the polymer at a later stage. This has the additional advantage of ensuring that the stabilizer does not phase-separate from the polymer; solubility of stabilizers in polymers is often low, limiting the maximum useful concentration.

TESTING

Much of the testing of the photodegradation properties of polymers is conducted with the objective of determining their suitability for outdoor use, where solar radiation is usually the principal hazard. Natural exposure, even in an extreme climate, is too slow for many commercial requirements and suffers from the further objection that there are no simple ways to calibrate one climate against another. There are many strategies for accelerated artificial weathering in the laboratory, but it is difficult to calibrate one conditioning procedure—natural or artificial—against another. The purpose of testing is (i) to rank polymers according to their resistance to UV photodegradation; (ii) to determine whether additives, including stabilizers, have a beneficial or a detrimental effect; and (iii) to predict lifetime in service. The first two are relatively easy to achieve, though the ranking may differ (slightly) from one UV environment to another. The third one is extremely elusive and has been achieved only for very closely specified materials operating in very well defined environments.

The testing of photodegradation properties normally consists of a conditioning exposure of known characteristics followed by interrogation of the degraded sample by means of chemical analysis, structural/morphological characterization, and mechanical/engineering analysis.[26] Conditioning can be natural (usually in an extreme climate) or artificial. Acceleration of UV degradation can be achieved by using high intensities, though use of intensities higher than the maximum terrestrial UV levels may provoke degradation that is not only more rapid but different in kind to that obtained in service. Even less advisable is the use of UV sources with wavelengths lower than those in solar radiation as this may provoke chemical degradation pathways not found with higher wavelength radiation.[27] Popular chemical characterization techniques include FTIR, to determine the type and abundance of reaction species, and GPC to obtain the molecular mass distribution.

Morphological characterization can be conducted by light and electron microscope techniques and by X-ray diffraction and thermal analysis, often used to determine crystallinity. Standard mechanical tests can be used to determine strength, extension to break in tension, and toughness. Normally, a selection of characterization methods is used with samples exposed for selected periods. When mechanical tests are used, the exposure period increment must be fairly short, in case a recovery phenomenon is present (see the section ''Engineering Properties—Consequences of Photodegradation''). Other tests related to appearance, such as gloss or color, may be employed if relevant to the service application.

A departure from the conventional ''expose then test'' procedure in which these two activities are separated, has been described recently in which the carbon dioxide emitted as the result of photodegradation is collected and measured in a quasicontinuous manner while the exposure is in progress.[7]

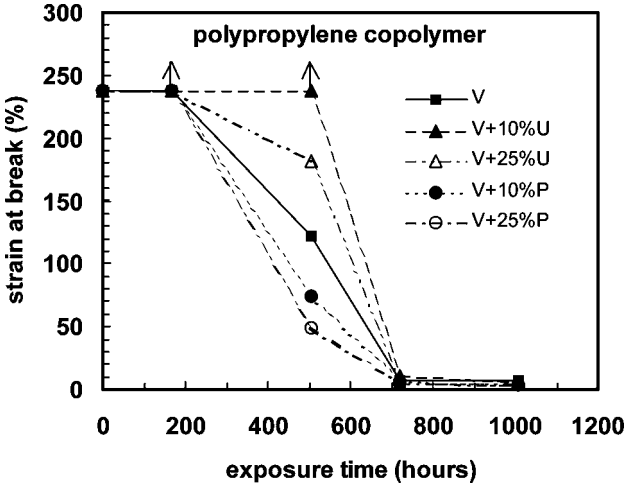


Fig. 10 Strain at break Vs. UV exposure time for polypropylene copolymer bars made from (a) virgin polymer (V); (b) virgin polymer + ''undegraded'' recyclate (U); (c) virgin polymer + photodegraded recyclate (P). Results for blends with recyclate contents of 10% and 25% are shown. (More details in Craig, I.H.; White, J.R. Mechanical properties of photo-degraded recycled photo-degraded polyolefins. J. Mater. Sci. **2006**, *in press*.)

RECYCLABILITY OF PHOTODEGRADED POLYMER

There is ever increasing pressure to recycle materials. Although thermoplastic polymers might be expected to be amenable to recycling, photodegraded polymers may present difficulties. Photodegradation creates active species that readily partake in further degradation reactions. Reprocessing may not destroy the activity. Even if the active species are lost by reaction during reprocessing (for example at elevated temperatures in mixing, extrusion, or molding), the reaction products will often be prodegradants and will be present to accelerate photodegradation of the recyclate during its new lifetime.[28,29] An example of prodegradation action is given in Fig. 10.

The addition of new stabilizer to recycled polymer (''restabilization'') is, therefore, required not only to replace the stabilizer consumed in the first life of the material, but also to attempt to combat the effect of the presence of prodegradants.[30]

CONCLUSIONS

Photodegradation of polymers is a complex process. Ultraviolet photo-oxidation creates free radicals that start a chain reaction, which can progress quite rapidly, causing rapid deterioration of mechanical properties. Degradation is often steeply graded from the exposed surface as the result of oxygen diffusion limited reaction. Photostabilizers are available and can produce spectacular improvements in polymer lifetime under UV exposure. The degradation products are often prodegradants and this places restrictions on the recyclability. Testing for photodegradability usually involves accelerated laboratory conditioning and this is difficult to calibrate against natural conditions to produce lifetime prediction.

REFERENCES

1. Davis, A.; Sims, D. *Weathering of Polymers*; Applied Science: London, 1983.

2. Rabek, J.F. *Photodegradation of Polymers (Physical Characteristics and Applications)*; Springer: Berlin, 1996.
3. Goss, B.G.S.; Blakey, I.; Barry, M.D.; George, G.A. Modelling of infectious spreading in heterogeneous polymer oxidation. II. Refinement of stochastic model and calibration using chemiluminescence of polypropylene. Polym. Degrad. Stab. **2001**, *74*, 523–532.
4. Audouin, L.; Gueguen, V.; Tcharkhtchi, A.; Verdu, J. ''Close loop'' mechanistic scheme for hydrocarbon polymer oxidation. J. Polym. Sci: A: Polym. Chem. **1995**, *33*, 921–927.
5. Gillen, K.T.; Wise, J.; Clough, R.L. General solution for the basic oxidation scheme. Polym. Degrad. Stab. **1995**, *47*, 149–161.
6. Shyichuk, A.V.; Lutsjak, V.S. A determination of rates ratio of simultaneous crosslinking and scission from MWD shape. Eur. Polym. J. **1995**, *31*, 631–634.
7. Christensen, P.A.; Dilks, A.; Egerton, T.A.; Temperley, J. Infrared spectroscopic evaluation of the photodegradation of paint. Part 1. The UV degradation of acrylic films pigmented with titanium dioxide. J. Mater. Sci. **1999**, *34*, 5689–5700.
8. Marcos Maillo, C.; White, J.R. A preliminary study of stress-aided photodegradation of rubber. Plast. Rubb. Compos. **1999**, *28*, 277–287.
9. Girois, S.; Delprat, P.; Audouin, L.; Verdu, J. Oxidation thickness profiles during the photooxidation of non-photostabilized polypropylene. Polym. Degrad. Stab. **1997**, *56*, 169–177.
10. Gardette, J.-L.; Gaumet, S.; Phillipart, J.L. Influence of the experimental conditions on the photooxidation of poly(vinylchloride). J. Appl. Polym. Sci. **1993**, *48*, 1885–1895.
11. Hoekstra, H.D.; Spoormaker, J.L.; Breen, J.; Audoouin, L.; Verdu, J. UV-exposure of stabilized and non-stabilized HDPE films: physicochemical characterization. Polym. Degrad. Stab. **1995**, *49*, 251–262.
12. O′Donnell, B.; White, J.R. Stress-accelerated photo-oxidation of polypropylene and glass fibre-reinforced polypropylene. Polym. Degrad. Stab. **1994**, *44*, 211–222.
13. Audouin, L.; Langlois, V.; Verdu; De Bruijn, J.C.M. Review: role of oxygen in polymer ageing: kinetic and mechanical aspects. J. Mater. Sci. **1994**, *29*, 569–583.
14. Rabello, M.S.; White, J.R. The role of physical structure and morphology on the photodegradation behaviour of polypropylene. Polym. Degrad. Stab. **1997**, *56*, 55–73.
15. Rabello, M.S.; White, J.R. Phototodegradation of polypropylene containing a nucleating agent. J. Appl. Polym. Sci. **1997**, *64*, 2505–2517.
16. Turton, T.J.; White, J.R. Observation of different photo-degradation behaviour in two similar polypropylenes. J. Mater. Sci. **2001**, *36*, 4617–4624.
17. White, J.R. Origins and measurement of internal stresses in plastics. Polym. Test. **1984**, *4*, 165–191 [Also appeared as Chapter 8 in *Measurement Techniques for Polymeric Solids*; Brown, R.P.; Read, B.E., Eds., Elsevier: Barking, 1984].
18. Li, T.; White, J.R. Residual stress distribution modification caused by weathering in polypropylene and polystyrene. Polym. Eng. Sci. **1997**, *37*, 321–328.
19. Arvidsson, A.; White, J.R. Densification of polystyrene under ultraviolet irradiation. J. Mater. Sci. Lett. **2001**, *20*, 2089–2090.
20. Qayyum, M.M.; White, J.R. Weathering of injection-moulded glassy polymers: changes in residual stress and fracture behaviour. J. Mater. Sci. **1985**, *20*, 2557–2574.
21. Qayyum, M.M.; White, J.R. The effect of weathering on residual stresses and mechanical properties in injection-moulded semi-crystalline polymers. J. Mater. Sci. **1986**, *21*, 2391–2402.
22. Zweifel, H. *Stabilization of Polymeric Materials*; Springer: Berlin, 1998.
23. Chirinos-Padrón, A.J.; Allen, N.S. Aspects of polymer stabilization. In *Handbook of Polymer Degradation*; Hamid, S.H., Amin, M.B., Maadhah, A.G., Eds.; Dekker: New York, 1992; Chapter 8, 261–303.
24. Turton, T.J.; White, J.R. Effect of stabilisers and pigment on photodegradation depth profiles in polypropylene. Polym. Degrad. Stab. **2001**, *74*, 559–568.
25. Gugumus, F. Aspects of the impact of stabilizer mass on performance in polymers. Part 1. Performance of low and high molecular mass HALS in polypropylene. Polym. Degrad. Stab. **1999**, *66*, 133–147.
26. Brown, R., Ed. *Handbook of Polymer Testing*; Rapra Technology: Shawbury, Shrewsbury, 1999.
27. Pospîšil, J.; Horák, Z.; Pilař, J.; Nešpürek, S.; Billingham, N.C.; Habicher, W.D. Effect of testing conditions on performance and durability of stabilisers in plastics. Polym. Polym. Compos. **2003**, *11*, 81–89.
28. Valenza, A.; La Mantia, F.P. Recycling of polymers. Part 1. Photo-oxidized polypropylene. Polym. Degrad. Stab. **1987**, *19*, 135–145.
29. Stevenson, W.; White, J.R. Photo-sensitivity of recycled photo-degraded polystyrene. J. Mater. Sci. **2002**, *37*, 1091–1100.
30. Pospîšil, J.; Sitek, F.A.; Pfaendner, R. Upgrading of recycled plastics by restabilization—an overview. Polym. Degrad. Stab. **1995**, *48*, 351–358.

Photoresists

Sergei V. Zelentsov
Nadezda V. Zelentsova
Chemical Department, Nizhny Novgorod State University, Nizhny Novgorod, Russia

INTRODUCTION

Since the appearance of the integrated circuit, photolithography (or photoresist technology) has been the main locomotive of economic and technological progress in the manufacturing of microelectronics devices. The ability of production of numerous copies of a circuit on a single silicon wafer in the batch-like technology has enabled the industry's exponential growth and the paramount improvement of the circuits. Gordon E. Moore[1] has made an observation that circuit densities of semiconductors had doubled every year for the past three decades and would continue to do so. Later, he revised his annual rate of circuit density doubling upward and has stated that every 18 months density-doubling is more suitable rate.

Photolithography can be defined as a method for transferring a desired pattern into a photosensitive material called photoresist, which having been placed on the top of the device layer, is able to change its chemical properties under the UV light exposure. Usually a photoresist contains a solvent, photosensitive chemicals (PAC), and a polymer binder. By changing the chemical properties of the photoresist in the exposed areas (it becomes either less resistant in the case of a positive photoresist or more resistant in the case of a negative photoresist against an etching procedure called development), it becomes possible to provide a prerequisite for a photoresist mask formation. Finally, regions of the substrate unprotected by photoresist are etched away; when the remaining photoresist is stripped away, the device layer becomes patterned as desired. Over the past 50 years, chemists have been able to provide a wide variety of photoresists to meet the constantly growing needs of microelectronics. The choice of materials to be utilized as components of photoresists and determination of the optimal parameters of their use in photolithography assumes thorough knowledge of the corresponding chemical mechanisms of the relief formation process. The trend in microelectronics toward printing features 0.20 μm and below has motivated researches to find new principles of the photoresist formulating, or to repair the old ones to be involved in the lithography at the 193-nm wavelength of argon fluoride excimer lasers, or even further at 157 nm and below. The chemical amplification and top-surface imaging concepts seem to be the most productive among the principles that have recently been developed. It is believed that they will be the grounds for the future levels of the technology.

Chemical nature of photoresists, the chemistries involved in the photolithography, the properties of photoresists are briefly described. The discussion includes diazonaphthoquinone/novolac positive photoresists, polymer-aromatic diazide negative photoresists, photopolymerizable compositions, chalcogenide glass using systems, chemically amplified photoresists, and photoresists with an image formation in a thin layer.

GENERAL SCHEME OF PHOTOLITHOGRAPHY

The photolithography usually consists of the following obligatory steps (see Fig. 1):

- *Wafer surface priming*. A surface treatment with a primer drives off moisture and improves adhesion of a photoresist. Typically, a wafer is heated in the presence of chemical vapor of the primer. Hexamethyldisilazane seems to be one of the best primers.
- *Coating of the wafer with a photoresist*. A small amount of the photoresist is dropped onto the center of the wafer and then spun at high speed to produce a uniform thin film.
- *Prebaking of the photoresist layer*. After the photoresist being coated on the wafer, a solvent should be driven off by a prebake.
- *Exposure of the photosensitive article*. The exposure step photographically transfers a pattern from a reticle or photomask to the photoresist coated on the wafer surface. Photomasks are glass plates with patterns made of opaque and transparent areas. A photomask will typically have the pattern for a few dice and will be stepped across the wafer exposing the pattern after each step. In order to ease a task of a photomask fabrication and make the process less defect sensitive, photomask patterns are either 5× or 4×, the size of the desired feature on the wafer, and the photomask pattern is optically shrunk before reaching the wafer.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120006819

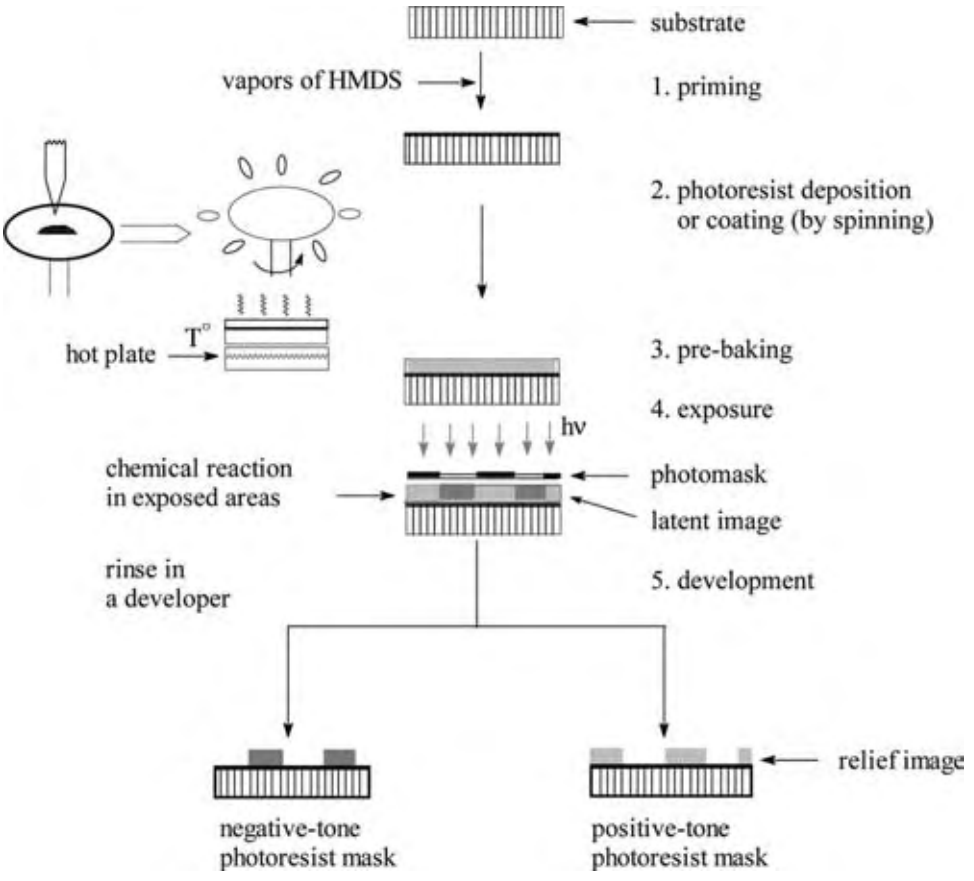


Fig. 1 Schematic presentation of a commonly used photolithography process. *(View this art in color at www.dekker.com.)*

- *Postexposure bake of the wafer*. A postexposure bake (PEB) improves contrast of the photoresist before its development. The PEB process causes three effects:[2] 1) diffusion of the PAC; 2) solvent evaporation; and 3) thermally induced chemical reactions. In general, the dissolution rate of a resist decreases as a function of a PEB temperature. PEB becomes more important for the photoresists with a chemical amplification (CA) feature. The photoresists need the PEB to complete chemical reactions initiated by exposure.
- *Development of a relief image in the photosensitive layer of a photoresist*. A development of a relief image in the photosensitive layer of the photoresist is the central step of a photolithography. The step suggests an implementation of physical and chemical procedures to use differences in properties of the exposed and unexposed parts of the photoresist for selective removal of only one type (i.e., either exposed or unexposed) of them and hence to produce a relief image in the photoresist. At the end of the development, the pattern from the reticle is replicated in the photoresist and a photoresist mask is formed.
- *Postbaking of a photoresist mask*. Postexposure high temperature bakes are used to stabilize the film prior to subsequent processing.

A photoresist mask formed by a photolithography is used to realize selective processing, for example, etching of a substrate through windows opened in the photoresist. After etching, the photoresist mask has to be stripped off and the wafer should be sent for further processing.

The most critical and expensive step of photolithography is an exposure. A typical exposure system consists of three main components: 1) a source of light at a particular wavelength; 2) a reticle photomask used as a master copy of the desired pattern through which the light is passed to transfer the pattern onto a photoresist-coated wafer sometimes by means of 3) projection optics. In early lithography systems, the photomask was placed directly on the wafer and illuminated by a lamp to expose the photoresist in a process known as contact printing. However, this process was not suitable for a batch production because the photomask gained distortions as a result of direct contact with the wafer. Later, the photomask was lifted just above

the wafer during exposure, in a method known as proximity printing. Proximity and contact printing transfer the photomask feature size to the wafer in a 1 : 1 fashion. In projection printing, the photomask pattern is demagnified as being transferred to the wafer. Projection optics is required to project a reduced image of the photomask onto the photoresist. Projection printing is the modern industry standard and promises to be the standard for at least in the nearest future.[3] Projection optics give rise to diffraction effects which in turn limit resolution, or minimum printable feature size of the lithography tool. Resolution obeys the Raleigh equation:

$$R = \frac{k_1 \lambda}{\mathrm{NA}} \tag{1}$$

where R is a resolution of projection optics used, λ the illumination wavelength, NA the numerical aperture of the final lens element, and k_1 a lithography level parameter, which effectively captures information about every other aspect of the lithography system. Usually k_1 belongs to the range 0.5–1.0. In the case of the standard process with a single-layer positive resist $k_1 = 0.8$; the use of modern techniques (such as multilayer resists, contrast-enhancing methods, top-imaging technologies, etc.) makes it possible to reach the limiting value of 0.5.

Next important characteristic of a projection lithography is the depth of focus (DOF) defined as a range over which the wafer can be moved along the optical axis in such a way that the image stays in focus:

$$\mathrm{DOF} = \frac{k_2 \lambda}{\mathrm{NA}^2} \tag{2}$$

where k_2 is a constant, which depends upon properties of projection optics, and has maximum value of 2 in the case of the idealized optics.

Usually, manufacturers of the projection optics for steppers destined for manufacturing of VLSI try to use the optics having minimal NA.

PHOTORESISTS, THEIR LITHOGRAPHIC PROPERTIES, CLASSIFICATION, AND MECHANISMS TO PRODUCE A RELIEF IMAGE

Lithography Properties of Photoresists

Photoresists can be characterized by several parameters such as sensitivity, resolution, contrast, etch resistance, etc.

Sensitivity

Sensitivity, S, to the UV radiation seems to be one of the most important properties of a photoresist. The value can be defined as dose (or exposure) of the UV light necessary to turn a photoresist into either insoluble (negative photoresist) or soluble (positive photoresist) compound.

The given definition is somewhat uncertain. Fig. 2 shows the dependence of thickness of a photoresist exposed by various UV light doses and developed by rinsing into a developer. There are three values that could characterize an exposure curve: $S_{\min}$, $S_{0.5}$, and $S_{1.0}$. These values of doses correspond, respectively, to the beginning of the relief image formation (frequently named as the critical or minimum dose), to obtaining of a photoresist mask with thickness being a half of that for a starting photoresist, and to reproducing thickness of the photoresist completely. They all are useful for photoresist characterization.

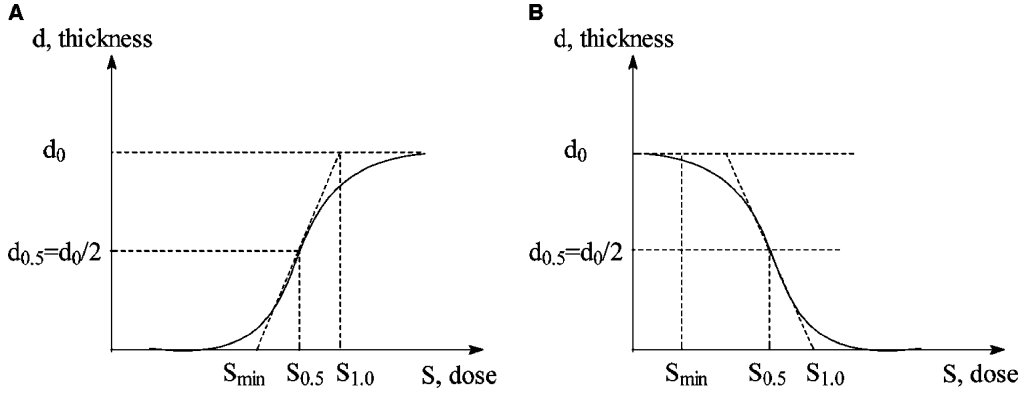


Fig. 2 Typical exposure or characteristical curves. d_0 is the maximum thickness of a resist layer: (A) a negative tone photoresist; (B) a positive tone photoresist.

Resolution

Resolution can be defined as the minimal or critical dimension (CD) of a photoresist mask element that is possible to develop. Sometimes resolution is considered to be a number of lines and intervals of equal width placed within an interval of 1 mm.

Contrast

Photoresist contrast, γ, is a measure of a resolving power of a photoresist. The photoresist should have high contrast to give high resolution, so that the UV light reflected from a substrate, or appeared because of diffraction effects during exposure should not influence upon a photoresist mask formation. The contrast correlates well with photoresist profile characteristics, but less so with CD reproduction.[4]

The measured contrast is a slope of a ''thickness obtained vs. dose exposed'' curve (characteristic curve). The theoretical value of contrast is the maximum slope of a plot of the logarithm of development rate vs. logarithm of exposure energy curve. Contrast can be determined as[5]

$$\gamma = \left| \log\left(\frac{S_{\min}}{S_{\max}}\right) \right|^{-1} \quad (3)$$

where $S_{\min}$ and $S_{\max}$ are doses corresponding to complete removal of the photoresist and to formation of areas with maximum thickness, respectively.

Etch resistance

A photoresist mask formed after development stage serves often as a mask for etching of a substrate. Resistance of the mask against the etchant of the substrate seems to be one of the main functional characteristics of the photoresist. The resistance can be measured as a time interval during which the photoresist mask withstands the etchant action, i.e., from the moment of the etching beginning up to the moment of appearance of such defects as partial destruction, undercutting, local point-like etching of the photoresist layer, etc. Etch-resistance can be written as

$$K = \frac{\nu_{\text{sub}}}{\nu_{\text{res}}} = \frac{h}{x} \quad (4)$$

where ν_{sub} and ν_{res} are rates of etching of a substrate and of a resist mask, respectively, and h is depth of the substrate etching, and x is width of side-etching of the resist mask.

Last two decades were period of a transition from a liquid toward dry etching. That is why plasma resistance is among the most important properties of a photoresist. In the case of fluorine-containing plasmas, etch resistance is defined as a ratio of the resist mask etching rate to the rate of SiO_2 etching. When a polymer substrate is to be etched, it is possible to use a ratio of the rate of the resist mask etching to the rate of polystyrene etching.

It is well known[6] that insertion of silicon-containing groups into a polymer enhances its plasma resistance greatly. The rate of a reactive ion etching in oxygen (RIE in O_2) in the stationary state, P, is proportional to the ratio of the silicon atom mass densities in the polymer, $\rho_{\text{pol.}}$, and in SiO_2, ρ_{oxide}:[7]

$$P = P_0 \frac{\rho_{\text{oxide}}}{\rho_{\text{polym.}}} \quad (5)$$

where P_0 is a rate of sputtering of SiO_2 being etched in RIE in O_2. Irrespective of the etch technology used, all polymers are etched with a constant rate after the starting transitional period. The difference between contents of silicon in the exposed and unexposed areas of the photoresist has to be not less than 10^{15}–10^{16} atoms per 1 cm^2 to ensure a successful RIE in O_2.[8]

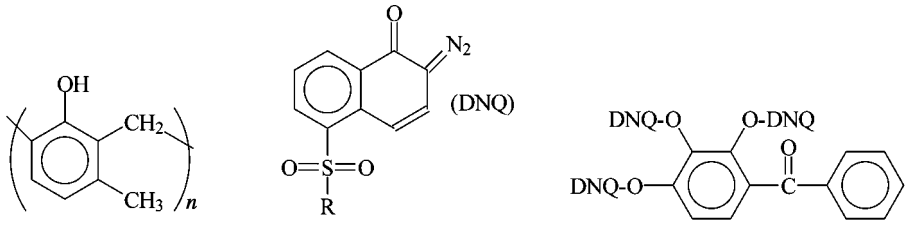
Many attempts have been made to find correlations between etch resistance of a polymer and its chemical nature. The thumb rule advices to use aromatic groups incorporated into a polymer to enhance etch resistance. However, about 20 years ago it was found[9] that aromatic rings were not strictly needed to improve polymer etch resistance. It was the carbon-to-hydrogen atoms ratio for a polymer that was related to its resistance in oxygen plasma. An increase in the ratio tends to produce a decreased rate of the RIE in O_2.

Application of a photoresist in a given technological process often imposes additional demands. It is extremely important to point out that choice of a photoresist should be dictated by a concrete technology.

Mechanisms of Relief Images Formation in Positive Photoresists

Diazidonaphthoquinone-containing photoresists

The diazido-*ortho*-naphthoquinone/novolac (DNQ/novolac) photoresists have been the workhorse for the semiconductor industry for over 30 years and still continue to be the most widely used photoresists today.[10] They may be utilized with different exposure wavelength technologies: I-line (365 nm), G-line (436 nm), and even 248 nm. Although productivity requirements for the technology with 248 nm exposure have demanded the transfer to photoresists with a chemical amplification, the DNQ–novolac resists still remain to be the most widely used and constantly finding new areas of implementation. The modern state-of-the-art in the microdevices technology gives an opportunity to use the DNQ–novolac resists to print features as small as 0.25 μm.[10] In addition, these photoresists in the form



1 **2** **3**

of thick film materials are used in the manufacture of thin film magnetic heads, micromachines, etc.

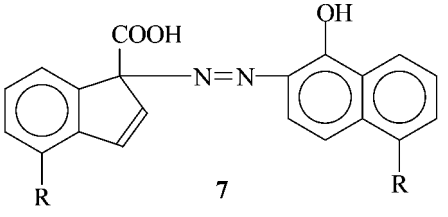
A DNQ–novolac photoresist usually consists of a novolac polymer and diazido-*ortho*-naphthoquinone [2-diazo-1(2H) -naphthalenone-5-sulfonate] derivative. Photoresist films are cast from solutions in organic solvents (such as propylene glycol monomethyl ether acetate). The polymer provides the desired physical characteristics of the material, such as good adhesion, film forming ability, and resistance to a plasma etching. Commercial photoresists contain a complex mixture of *meta*- and *para*-cresols having molecular weight of 960–2400 and 5–15 wt.% of DNQ.[11]

Pure novolac polymer **1** is dissolved in an aqueous base. An addition of DNQ **2** to the novolac leads to reduction of the dissolution rate of the polymer in the aqueous base, relative to the dissolution rate of the pure polymer. This difference can be of several orders of magnitude.[11] One of the reasons is that DNQ-fragment is hydrophobic by its nature. The second reason is that DNQ-fragments can and do form hydrogen bonds between nitrogen atoms of the diazo-group and hydrogen atoms of hydroxyl groups of the novolac molecules.[12,13] Although bond energies of the hydrogen bonds are not large and being of 10–12 kcal/mol, they are strong enough to inhibit a penetration of water molecules into the photoresist films. The hydrogen bonds formed between different novolac macromolecules serve as intermolecular physical cross-links.[12,13] Their importance to inhibit dissolution is supported by an observation that only use of PACs having not less than two DNQ groups (as it has place in the most frequently used DNQ-derivative **3**) makes it possible to inhibit the dissolution completely.[11]

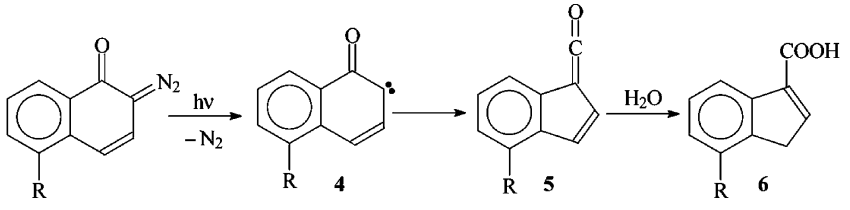
The UV light decomposes DNQ groups; the overall reaction scheme is shown here.[11,14]

Photolysis of the DNQ gives carbene **4** and molecular nitrogen; the carbene then rearranges into ketene **5**. The ketene is very reactive and interacts with trace amounts of water (constantly occurring in a novolac polymer) and produces indene carboxylic acid **6**. Besides, decomposition of DNQ leads to breaking of hydrogen bonds between DNQ and novolac macromolecules. In addition, gaseous nitrogen is evolved and produces microchannels in bulk photoresist. And at last, the acid formed is hydrophilic in contrast to the initial DNQ moieties being hydrophobic. All this promotes water molecules penetration into the exposed areas of the photoresist, and finally, the exposed areas become soluble in the aqueous base. It is this change in the dissolution rate of the photoresist, which allows for the formation of relief images in the positive photoresists, comprising of DNQ derivatives and novolac. Nowadays, aqueous solution of 0.26 N tetramethylammonium hydroxide is more frequently used as a developer.[11]

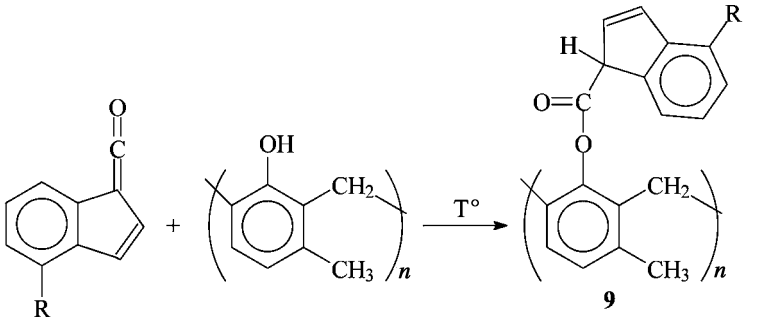
Ketene **5** can react not only with water but also take part in intermolecular dimerizations with unreacted DNQ derivatives, decarboxylation, and decarbonylation. In particular, coupling of DNQ and ketene or indene carboxylic acid gives rise to dye **7**.



7

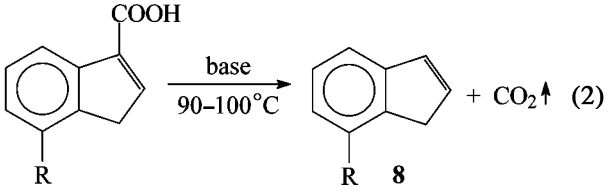


(1)



(3)

The presence of a base in the photoresist causes decarboxylation of the indene carboxylic acid.



(2)

Sometimes, the last reaction is used to obtain a negative-tone resist mask by means of the DNQ–novolac positive photoresist.[14] Imidazole, triethanol amine, and other bases can be used as tone-modifiers.

Ketene formed after exposure can react with novolac macromolecules, especially when water molecules are completely absent. The result of such interactions is crosslinking of novolac macromolecules.

Mechanisms of the Relief Images Formation in the Negative Photoresists

The most wide use in industry have been received by negative photoresists based upon 1) crosslinking of macromolecules by means of low molecular weight photosensitive bifunctional compounds; 2) photopolymerization; and 3) photochemically induced diffusion of metal atoms into chalcogenide glasses.

Negative photoresists possess very important advantages. They have

1. Large technological ranges of acceptable processing parameters; insensitivity to over development.
2. Good adhesion and wet etching resistance.
3. An ability to ''self-proximity corrections,'' i.e., unexposed areas can be ''recovered'' up to the desire dimensions, as a result of under etching in the course of subsequent isotropic wet etching.
4. A wider choice of components of the photoresist compositions and solvents to be used as developers.

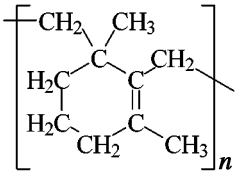
Unfortunately, the negative photoresists have some drawbacks:

1. Resolution is limited by the photoresist film thickness. A rule of thumb exists that typical CD of a photoresist mask obtained after development may be estimated as thrice of photoresist film thickness.
2. The relief formation in negative photoresists is very much inhibited by oxygen.
3. It is difficult to use them for the ''lift-off'' photolithography.

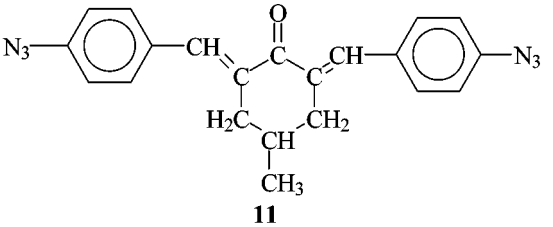
Crosslinking of polymers by diazides and related compounds

The most widespread photoresists of negative tone consist of organic polymers and aromatic diazides dissolved in an organic solvent. In 1970–1980, the photoresists were based on cyclized rubber **10** and 2,6-di (4′-azidobenzylidene) -4-methylcyclohexanone **11**.[14,15]

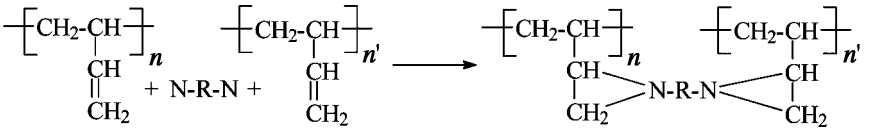
The relief formation in the photoresists can be presented by the following scheme. The diazide eliminates



10



11



(5)

nitrogen and turn into nitrene:

$$N_3\text{-R-}N_3 \xrightarrow[-2N_2]{h\nu} \bullet\dot{N}\text{-R-}\dot{N}\bullet \qquad (4)$$

The nitrene reacts with double bonds containing in polymer to produce ''diaziridine bridges'' or crosslinks.

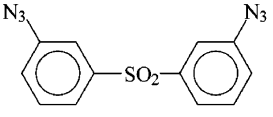
If there are no double bonds in polymer, then insertion of dinitrene into C–H bonds will have place.

Two crosslinks per macromolecule are needed to make a polymer insoluble. Quantum yield of a diazide photolysis is ca. 0.5, therefore, the critical dose of the photochemical polymer insolubilization should be about $50\,mJ/cm^2$.

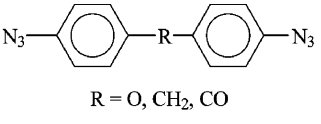
Oxygen usually inhibits the crosslink formation by trapping nitrene molecules to transform them into nitroso- or/and nitrocompounds.[14,16]

$$4N\text{-R-}N + 3O_2 \longrightarrow 2N\text{-R-NO} + 2N\text{-R-}NO_2 \qquad (7)$$

Photoresists, using cyclized rubber as a polymer and an aromatic diazide as a photoinitiator, swell strongly during development and have a very low contrast ($\gamma = 0.5$). Phenol- and cresolformaldehyde resins are frequently used as polymers and ''white azides'' as photosensitive components to avoid the last shortcoming. These azides **12** and **13** absorb at shorter wavelengths than the commonly used azide **11**.



12



R = O, CH_2, CO

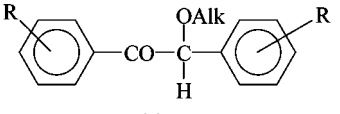
13

Absorbance of the photoresist is high (backscattering is absent), and a developer can underetch photoresist mask edges near a substrate. The latter gives a possibility to obtain the relief having high contrast and resolution. The underetching can produce images having a profile suitable for the ''lift-off'' lithography, but very rigorous control of the processing variable is desirable to produce such a profile.

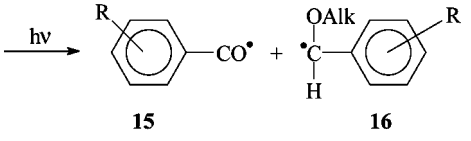
Photopolymerization

Photopolymerizable compositions have found their use in the production of thick (5–100 μm) relief images, especially in the field of printed plates and micromachinery. The compositions contain a monomer, matrix polymer (binder), and photoinitiator. The branched monomers are used most frequently. Many photosensitive compounds generating radical and ion centers upon UV irradiation are used as a part of the compositions. Polymethylmethacrylate and its derivatives are very often used as polymer binders.

The key component of a photopolymerizable composition is a photoinitiator. This compound can absorb UV radiation and then produce a pair of radicals, either by dissociation (a photoinitiator of the I type)[17]



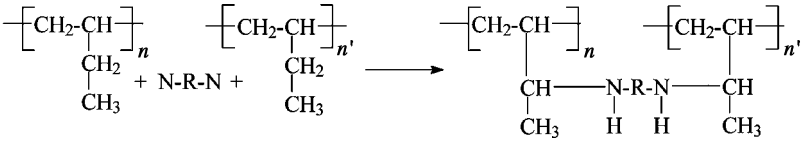
14



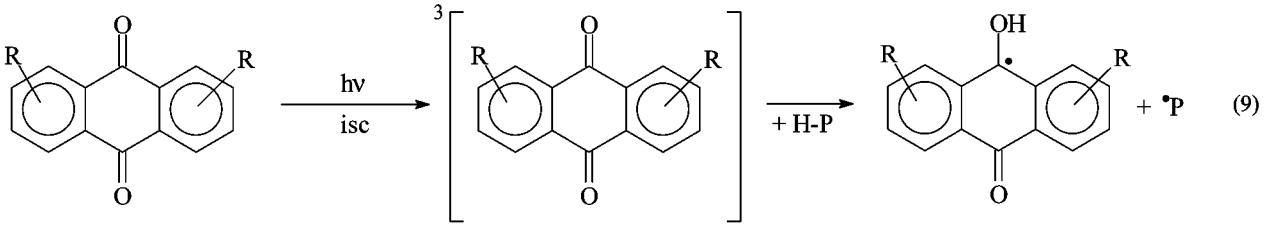
(8)

15 **16**

or by hydrogen atom abstraction (a photoinitiator of the II type).[17]



(6)



In the first case, the dissociation products **15** and **16** serve as centers initiating polymerization of monomers. In the second case, the II-type initiator absorbs the UV light and transfers into the singlet excited state and then, after intersystem crossing the molecule is converted into the triplet state **18**. The triplet molecule reacts with a hydrogen atom donor H–P to give ketyl radicals **19** and •P. The radicals initiate a polymerization process.

In general case

$$RM^{\bullet} + nM \longrightarrow R(M)_nM^{\bullet} \quad (11)$$

(the chain forming process)

$$R(M)_n^{\bullet} + R(M)_k^{\bullet} \longrightarrow R(M)_{n+k}R \quad (12)$$

(the terminating step)

The photopolymerizable composition can be used both as liquids to coat a substrate with light sensitive films and as dry photoresist films. In the latter case, the photoresist is a trilayer pie consisting of a top layer made of poly(ethylene terephthalate), a middle layer being the photopolymerizable composition, and an underlayer made of polyethylene. The underlayer is removed before exposure and the remained bilayer composite is laminated onto a clean substrate. The article is prebake to anneal a polymer layer and to enhance adhesion to the substrate. The selective exposure by the UV light through the top layer hardens the exposed areas of the photoresist. After exposure, the top layer is removed and the photoresist layer developed by rinsing into a developer solvent (trichloroethylene or the like).

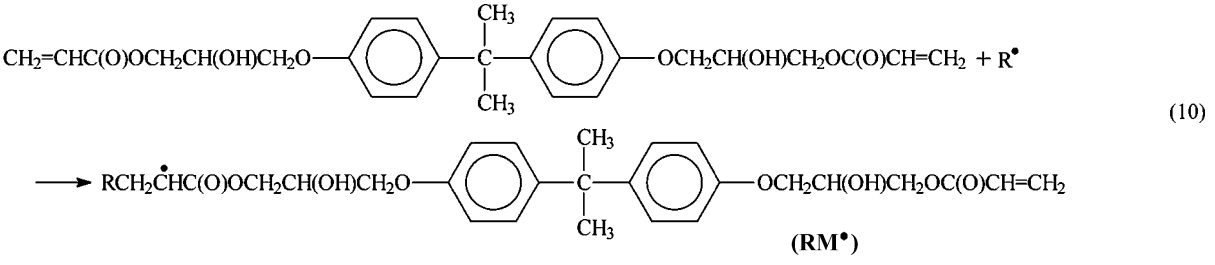
Among the most important advantages of the dry film photoresist are 1) easy coating; 2) convenience in processing; 3) processing stability; and 4) insensitivity to oxygen exposure. The dry film photoresist took their niche in the photoresist technologies of the electronics components not requiring high resolution.

Chalcogenide glasses

Inorganic photoresists constitute a new promising class of the photosensitive materials.[18–20]

The materials consist of an underlayer made of a chalcogenide glass such as As_2S_3, As_2Se_3, Ge–Se, etc. and a top thin layer of silver or its halogenide (AgCl). The UV light liberates silver atoms that diffuse into As_2S_3 layer (up to 10 nm depth) and produce alloys resistant to etching (this phenomenon is called as ''photodoping''). The UV light absorption is limited to a thin layer (because of very high absorbance of metal containing compositions), so high resolution may be obtained. Unfortunately, the UV light sensitivity of the resists is very low when liquid development is used. It can be enhanced tremendously (up to 50 mJ/cm^2) when etching in fluorine-containing plasmas is used.

For instance, Ge–Se composition deposited by CVD and doped by Ag has photosensitivity at wavelengths ranging from 200 to 450 nm. Upon exposure by the light Ag_2Se is formed as a result of photo initiated diffusion. Ag_2Se can be removed by treatment in an aqueous solution of KI. After the removal of the surplus



amount of silver the latent image is developed by plasma etching in SF_6 or freon gases.

The inorganic photoresists possesses some advantages:

1. They have very high contrast (up to 8) because of the contrast enhancement phenomena observed in the photoresists.[14,19,20]
2. The photoresists have very strong absorption even for thin films. The latter gives a possibility to use the photoresists as thin top layers in the bilayer photoresist systems.
3. The relief masks formed from the photoresists have very high resistance in oxygen plasma. This enables a use of the Ag/chalcogenide layer as a top layer in bilayer photoresists with high resolution and aspect ratio. The under layer of the photoresist is usually made of polymer such as polyimide.

The phenomenon of the contrast enhancement is caused by side-directed diffusion of silver atoms into the exposed areas because of the decrease of chemical potential in the areas. Silver atoms diffusion from the unexposed areas into the exposed ones compensates the influence of diffraction effects. As a result of diffusion, resolution may be considerably improved.

The use of inorganic photoresists should be advantageous for the technological processes because inorganic films can be deposited very homogeneously by a CVD procedure.

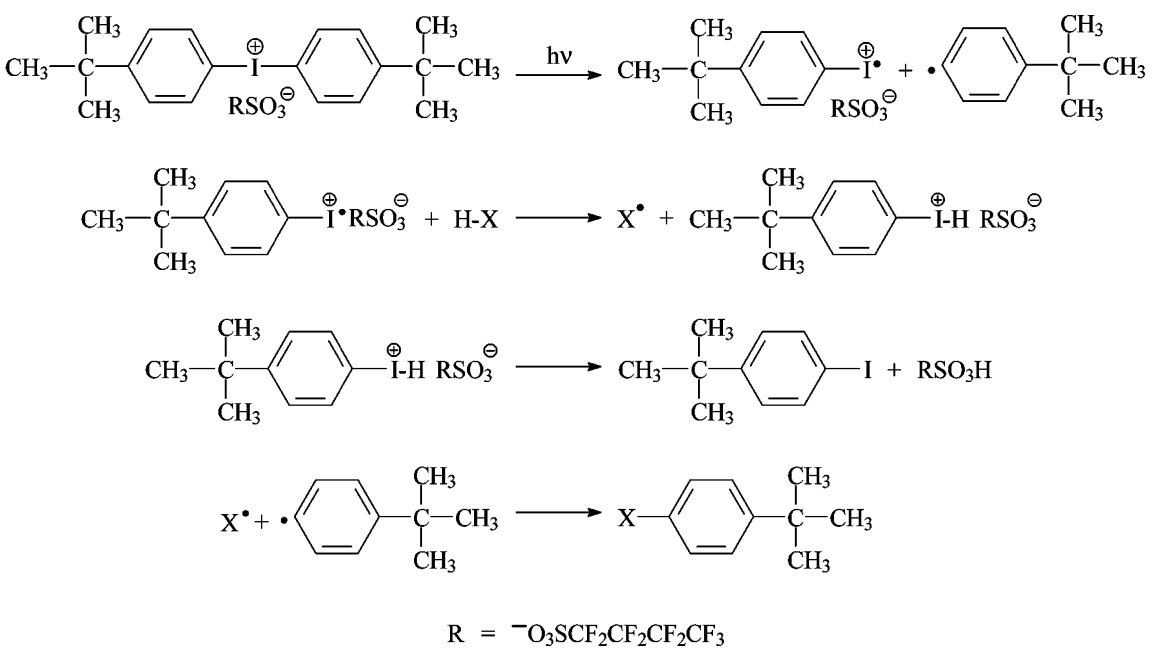
Chemically Amplified Photoresists

In 1979, Frechet and Willson put forward a very productive idea of a chemical amplification that was used in the development of a new generation of photoresists.[21] They decided to use a photoresist comprising of a photochemical acid generator (PAG) and a polymer that was able to switch from hydrophobic to hydrophilic in the course of acid catalyzed hydrolysis. The PAG reacts with light to produce an acid catalyst. During a subsequent postexposure bake, the catalyst diffuses and reacts with the polymer component, causing many reaction events in the polymer and recovers the acid catalyst. The acid molecules catalyze the deprotection reaction and provide a prerequisite for chemical amplification. The number of the reaction events initiated by single quantum absorption has been estimated to be of order of 100.[22]

The following reaction scheme can present a simplified mechanism occurring in the photoresists with chemical amplification chemistry.[21,23]

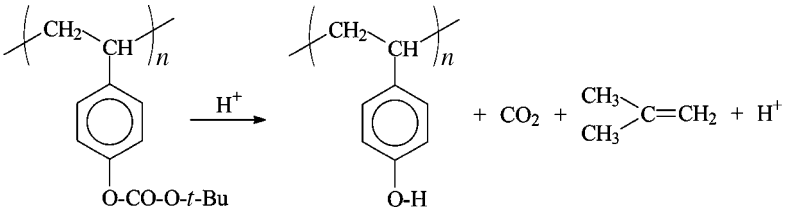
The photoresists are widely used in 248 and 193 nm photolithography and are generally considered as the most probable candidates for the next generation of photoresists with fine and ultra fine resolution.[21]

The problems met by chemically amplified photoresists are: a) poor stability to environmental contaminations such as airborne amines; b) sensitivity of lithographic parameters to PEB temperature variations; c) poor stability during storage both after coating and after exposure; d) side-directed diffusion



(13)

$R = {}^{-}O_3SCF_2CF_2CF_2CF_3$



(14)

of PAG causing broadening of formed lines; and e) formation of ''T'' shape profile.

The striking effect of airborne contaminations on the performance of the poly(4-*tert*-butoxycarbonyloxy styrene)/onium salt system has been demonstrated.[24] The effect can be observed when the coated wafers stand for 16 min in air containing only 15 ppb of an organic base. The cause of the effect is that the photogenerated acid can be neutralized by organic base.

Careful carbon filtration of the clean-room air has been used to remove airborne contamination effects.[24–26] To reduce the photolithographic performance parameters dependence upon the PEB temperature, it has been proposed[27] that the incorporating of the protecting hydrophility blocking groups having low activation energy for acid catalyzed hydrolysis (deblocking) should be used.

The most serious drawback of the chemically amplified resists is a considerable broadening of formed lines during PEB. Umbach et al. has shown that a thin (ca. 2 nm) electron beam exposure resulted in a ca. 40 nm feature with poly(4-*tert*-butoxycarbonyloxy styrene).[28] The high catalytic chain length also suggests that acid may migrate into areas that are unexposed. The side-directed acid diffusion is a reason of an image blur, line width altering, and so on. The problem can find its solving by using a base along with the polymer and PAG used. The base addition decreases side broadening, but it consumes some amount of the generated acid and the latter reduces both resolution and sensitivity of the photoresists.[29,30]

The chemical amplification idea appeared when it was necessary to develop photoresist material having high sensitivity, high resolution, and good plasma etch resistance. It was desirable primarily when the 248 nm exposure became the requirement of the industry. And the result of implementation of the idea was very good. Chemical amplification as a basic ideology of the photoresist creation partly worked at the 193 nm millstone.[21]

Photoresists with Thin Layer Imaging

The second productive idea useful in designing of modern photoresists is based upon ''thin film imaging'' (TFI).[31] The thin uppermost layer of a photoresist forms an image, which serves as a mask to transfer the pattern into the lower layers, to form an etch mask by means of the RIE in O_2. As thickness of the upper layer may be very low (of ca. 20–100 nm), the image in it can possess very high resolution. The RIE being anisotropic (i.e., the side etching is much slower than the in-depth one), permits to obtain high aspect ratio, and the pattern is transferred with high precision.

There are several realizations of the idea: bilayer lithographic processes, top-surface imaging through vapor phase silylation, digital top-surface imaging using vapor phase silylation, and imaging by means of selective graft-polymerization.

Bilayer lithographic processes

The TFI idea has found its application in developing of multilayer resists long ago. Later, the idea was used in the bilayer photoresists consisting of a thin silicon containing photosensitive imaging layer situated on the top of a crosslinked underlayer. After the UV exposure, and subsequent development the image formed in the upper layer is transferred (usually by means of RIE) into the underlayer. The remaining areas of the silicon-containing layer are oxidized by oxygen plasma to form plasma resistant oxide, which protects underlying material, while the unmasked areas of the underlayers are etched away. A simplified processing scheme of the processes using the bilayer photoresists is shown in Fig. 3 (the left-side scheme).

A modification of the bilayer photoresist processing is shown on the right side of Fig. 3. The scheme presents so-called ''Si-CARL'' (silicon chemical amplification of resist lines) process.[32] In the Si-CARL process, the upper layer is photosensitive, but does not contain silicon in its composition. In a typical process,[32] an anhydrade-containing photoresist thin imaging layer is selectively exposed by the UV light and then developed and exposed to a silylated agent. Silicon is incorporated into the remaining photoresist by use of amino-containing siloxanes, which react chemically with the anhidride functions and creates a silicon-containing skin on the surface of the image elements in the upper layer.

Disadvantage of both the processes shown in Fig. 3 lies in their complexity. Among their advantages are 1) good etch resistance of the photoresist masks

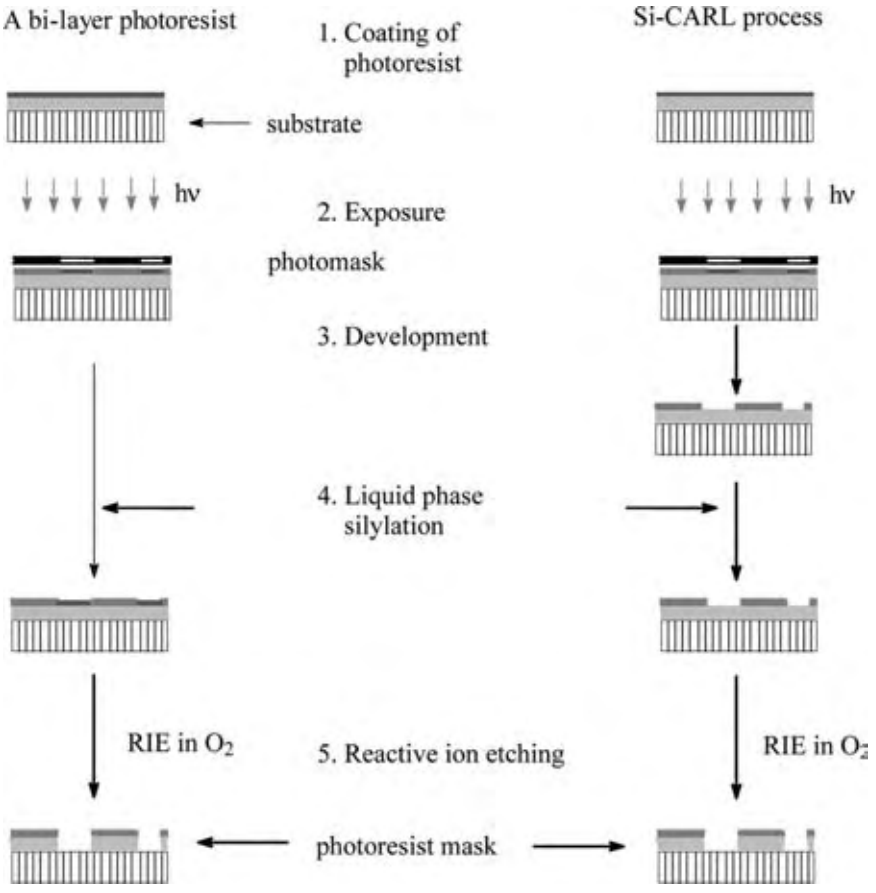


Fig. 3 Processing scheme of lithography with bilayer (left) and Si-CARL (right) photoresist. *(View this art in color at www.dekker.com.)*

coupled with high sensitivity of the upper photosensitive layer; 2) high aspect ratio of the masks formed; (3) the depth of focus for the photoresists is extremely large; and (4) the processes are simpler and cheaper than the trilayer photoresist technologies.

Top-surface imaging through vapor phase silylation

The complexity of the processes using the bilayer photoresists was reduced in single-layer photoresists working on the basis of the TFI idea. Diffusion enhanced silylated resist (DESIRE)-,[33,34] positive resist image by dry etching process (PRIME)-,[35,36] SUPER-,[37] and SAHR-technologies[38] use top-surface imaging of photoresists through their vapor phase silylation. Although the processes have been developed for the 365 or 248 nm lithography, there exist their modifications to work at shorter wavelengths.[39,40]

The DESIRE lithographic process is shown in Fig. 4. A DNQ/novolac photoresist is coated on a wafer (substrate). Exposure to the UV light causes the DNQ decomposition to give indene carboxylic acid. After the photoresist baked in vacuum at high temperature to form unstable ketene and the reaction of the ketene with phenolic group on the novolac resin, the unexposed regions of the photoresist become crosslinked. As a result, considerable difference in penetration of small molecules into the exposed and unexposed areas of the photoresist upper layer appears. The photoresist is then exposed to vapors of a reactive silylating agent (usually aminosilane). The compound permeates into the exposed regions of the upper layer (about 20–100 nm depth) of the photoresist, but not into the unexposed regions. The silylation agent reacts with free phenolic functions and transforms novolac macromolecules into silicon-containing ones. The silylated regions may then serve as etch masks in the course of subsequent RIE in O_2.

The most important advantages of the DESIRE photoresist technology are

- Greater simplicity of the process as compared to other TFI-technologies.
- Good compromise between high sensitivity of DNQ/novolac systems and high resolution obtained because of small thickness of the imaging layer and large anisotropy of RIE.
- Insensitivity to topology created on the substrate (absence of light scattering from a relief created on the substrate).
- Larger depth of focus.

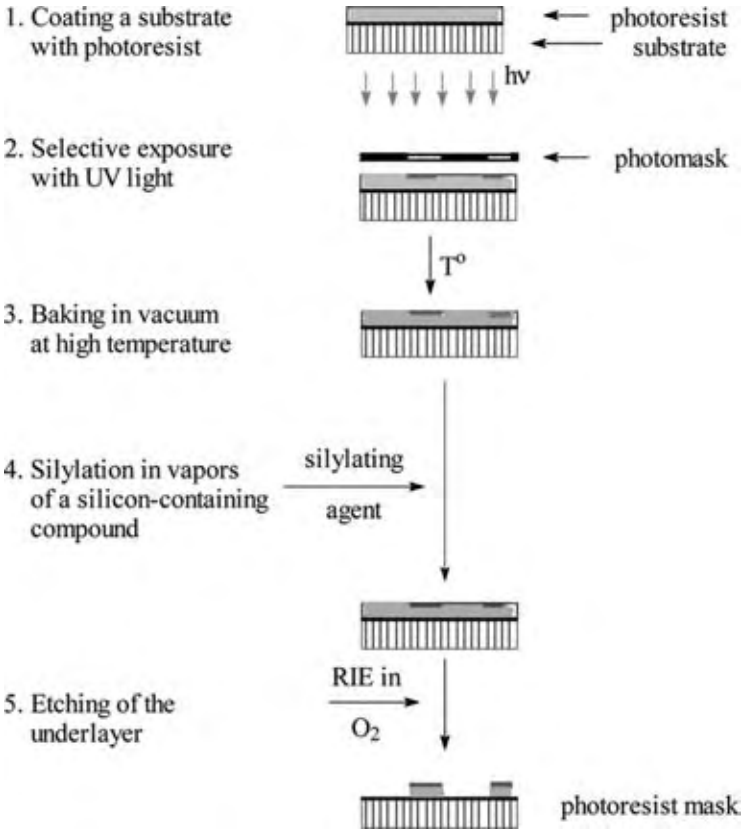


Fig. 4 Processing scheme for the DESIRE technology. *(View this art in color at www.dekker.com.)*

The main drawbacks of the DESIRE technology are the following points:

- "Skin" formation because of the silylating agent penetration not only into exposed, but unexposed regions.
- A "grass" formation (residues in open areas) caused by sputtering of silicon from silylated areas and transferring of it onto the open areas.
- Modified methods of cleanup and stripping of the photoresist mask is to be developed.
- The silylation should be controlled carefully because of cracking occurring when the content of silicon atoms incorporated into the photoresist is large enough and of not sufficient plasma resistance against RIE in O_2, when silylation process is stopped earlier than it should be.

It is possible to decrease the influence of "skin" or "grass" formation by etching away a top thin layer (ca. 20–30 nm), using halogen-containing plasma, just before the RIE in O_2.

A positive photoresist technology using polyvinylphenol layer as a photoresist and dimethylsilyldiamine as a silylating agent for 193 nm exposures has been developed.[39,41] The silicon atoms concentration in silylated regions with the said silylating agent in liquid phase is two times higher than that for the vapor phase silylation.[41] Resolution obtained by both methods is about 0.18 μm.[40,41] Diffusion rates correlate with the molar volume of the silylating agent, although the activation energy does not.[42]

The most frequently used silylation agents are hexamethyldisilazane, trimethylsilyldimethylamine, trimethylsilyldiethylamine, dimethylsilyldimethylamine, 1,1,4,4-tetramethyl-1,4-bis(*N*,*N*-dimethylamino)diethylene, and 1,1,3,3,5,5-hexamethylcyclotrisilazane.[29]

The PRIME technology being a variant of the DESIRE technology gives an opportunity to form openings down to 0.3 μm using the 248 nm exposure wavelength and hexamethyldisilazane as a silylating agent.[36]

Thus, the DESIRE technology gave a great impulse for penetration of the TFI idea into microdevice industry. It is not of large complexity, quite successful in implementation. The potential of the method does not seem to be used completely.

Digital top-surface imaging using a vapor phase silylation

A combination of the TFI and chemical amplification ideas leads to development of a new promising method

called ''digital top-surface imaging using a vapor phase silylation'' (DTSI).[43,44] The DTSI method was created to enhance the difference in reactivity toward a silylating agent between exposed and unexposed areas in the DESIRE technology. A photoresist in the DTSI technology contains poly(4-*tert*-butoxycarbonyloxystyrene) and PAG. After exposure by the UV light of the upper layer of the photoresist, a subsequent thermal treatment, exposure of the wafer to a silylating agent and RIE in O_2 good quality photoresist masks were formed.

The described photoresist technology is not greatly complicated when compared with the DESIRE technology, but it should and does possess a greater difference of the silicon atom components of exposed and unexposed areas of the photoresist, i.e., leads to better contrast and has improved production stability.

Both the DESIRE and DTSI processes might be good decisions for modern technology, but closely related with either DNQ/novolac or chemically amplified resists. The limited choice of materials to be used as the photoresist component and ''skin'' formation is the main drawbacks of the DTSI technology. The problems arisen could be party solved by the TFI approach realization in the form of the selective surface graft-polymerization.

Imaging by means of selective graft-polymerization

The idea being a basis for the method is simply to use a radical or ion latent image formed after selective exposure of a photosensitive layer to initiate graft-polymerization of monomers containing silicon and/or of improved plasma etch resistance.[45–52] The graft-polymerization has place on the interface between photoresist and silylating agent, or just nearby. Sometimes, the process is fast enough and produces a good conformal mask to etch the bulk photoresist by RIE in O_2. The processing scheme of the graft-polymerization lithography is shown in Fig. 5.

There exist several realizations of the graft-polymerization lithography (GPL).

Photolithography on the basis of the cationic graft-polymerization[46] uses an inert polymer film containing a PAQ coated onto a wafer. Exposing the photoresist to the UV radiation generates acid on its surface, to which a vapor-phase silicon-containing monomer is grafted. RIE in O_2 is used as a final step.[50,53,54]

One of the drawbacks of the method is inability to use metal-containing PAQ, such as $Ph_3S^+SbF_6^-$ or the like. There is no good substitution for the hexafluorometal-containing PAQ, but triphenylsulfonium-tris(trifluoromethylsulfonyl) methide. Most disadvantages caused by chemical amplification and postexposure heating have resulted in the widening of the large opening.

Another type of lithography uses a radical graft-polymerization.[55–58] In typical realization,[55] a negative resist structure can be created by applying a photoresist containing polyvinylacetate or poly(*iso*-butylene). A pattern-like exposure under the UV radiation creates radical centers. Exposure of the radical centers to oxygen converts them into peroxide or hydroperoxide groups. Subsequent thermally decomposing of the groups in presence of the silicon-containing

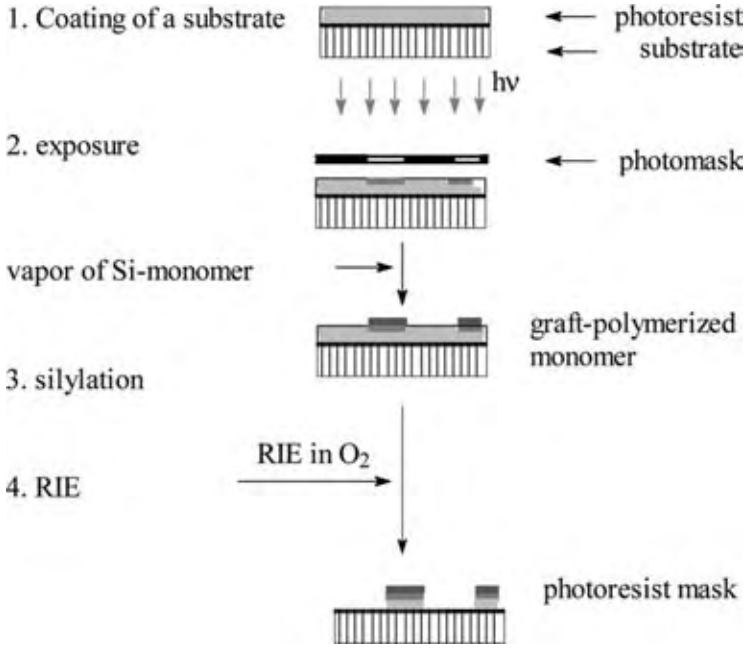


Fig. 5 Technology scheme of the photoresist technology using selective graft-polymerization. *(View this art in color at www.dekker.com.)*

monomer vapor and RIE in O_2, finishes the mask production process. The grafting monomer may be chosen from a wide class of silicon-containing monomers, such as vinyltrimethylsilane, γ-methacryloxyloxypropyltrimethoxysilane, etc.

The method described above serves as a prototype to create a universal imaging method, which will be in use to develop a new photoresist generation. Too high complexity of the method and not effective graft-polymerization of monomers from gas phase are among the main drawbacks of it.

To improve effectiveness of graft-polymerization and to decrease complexity of the lithography method, it is advantageous to use liquid or solid-phase polymerization strategies. One of the solving is to deposit a silicon-containing monomer onto a selectively exposed photoresist and then to heat it. The monomers will polymerize and an excess of them can be removed from the photoresist. The mask obtained can be used for the RIE in O_2.[52]

In conclusion, it should be stated that the TFI idea and the graft-polymerization imaging can be most desirable in the 193 and 157 nm technologies, and certainly will serve a prototype of a future photoresist.[59]

MODELING OR SIMULATING IN THE PHOTORESIST TECHNOLOGY

The photoresist technologies involve many steps; each can be characterized by many parameters. It is very difficult to optimize the processes. Previously used pilot experiments now proved to be very costly and difficult. It is not possible to optimize all parameters by performing large series of the lithography experiments, at all possible sets of the parameters. The problem may be simplified and sometimes completely solved by means of mathematical modeling and simulation.[60,61] Photoresist technology simulation provides the opportunity to perform experiments in a virtual environment that can be much faster and cheaper than the full-scale wafer experiments.

The process simulator PROLITH[62] is used to simulate a photolithography process and create a library of the input-technological parameters–profile relations. These profiles are then used to generate simulated diffraction responses, resulting in a library of diffraction-response-technological parameters–profile inter-relations.

There are some advantages provided by accurate simulation tools. They are not very expensive, not much time-consuming, and the ''experiments'' can be performed at more rich set of technological parameters, than the one usually being used for optimization, by means of the real experiments performing.

Today, software packages permitting to model an exposure step rigorously do exist, but the photochemical reaction steps and especially development step are still problems for theoretical description, and consequently a more primitive model must be used for their simulation. There seems to exist a good compromise suggesting strict descriptions of the exposure step (mainly by means of geometrical optics) and photochemical reactions (it is believed that the photochemical kinetics equations often permit to find their solutions or provide a ''constructive method'' of creating approximations in form of functional series[63]), and use semiempirical relations to describe development and plasma etching steps. The need for such an approach has been stated by the 2001 International Semiconductor Roadmap document that says ''there is a growing need to resist studies based on computational molecular modeling'' of the photoresist technologies.[3]

The further detail concerning the mathematical modeling of photoresist technologies may be found in literature.[64–68]

CONCLUSIONS

Photoresist technology is a novel technology of the 21st century. It finds more and more different applications in many fields of science and industry. The foremost of the technologies will be those using the thin film imaging idea.

The following textbooks and article can be useful for further reading on photoresists and their technology: Refs.[10,11,14,15,17,21,26,30,31,42,43,46,49,50,59].

REFERENCES

1. Moore, G.E. Lithography and the future of Moore's law. Proc. SPIE **1995**, *2440*, 2–17.
2. Toukhy, M.A.; Hansen, S.G. Influence of post exposure bake on resist contrast. Proc. SPIE **1994**, *2195*, 640–651.
3. http://public.itrs.net/Files/2002Update/Home.html (accessed January 2003).
4. Spragg, P.M.; Hurditch, R.J.; Toukhy, M.A.; Helbert, J.N.; Malhotra, S. Reliability of contrast and dissolution-rate-derived parameters as predictors of photoresist performance. Proc. SPIE **1991**, *1466*, 283–296.
5. Harada, K.; Tamamura, T.; Kogure, O. Delaited contrast (γ-value) measurements of positive electron resists. J. Electrochem. Soc. **1982**, *129* (11), 2576–2580.
6. Pederson, L.A. Structural compositions of polymers relative to their plasma etch characteristics. J. Electrochem. Soc. **1982**, *129* (1), 205–208.

7. Yurgensen, C.W.; Reichmanis, E.; Vasile, M. Experimental tests of the steady model for oxygen reactive etching. Proc. SPIE **1988**, *920*, 253–259.
8. Adesida, I.; Chinn, J.D.; Rathbun, L.; Wolf, E.D. Dry development of ion beam exposed PMMA resist. J. Vac. Sci. Technol. **1982**, *21* (2), 666–671.
9. Patterson, K.; Okoroanyanwu, U.; Shimokawa, T.; Cho, S.; Byers, J.; Willson, C.G. Improving the performance of 193 nm photoresists based on alicyclic polymers. Proc. SPIE **1998**, *3333*, 425–437.
10. Thompson, L.F.; Willson, C.G.; Bowden, M.J. *Introduction to Microlithography*, 2nd Ed.; American Chemical Society: Washington, DC, 1994.
11. Dammel, R.R. *Diazonaphthoquinone-Based Resists*; SPIE Optical Engineering Press: Bellingham, WA, vol. TT-11, 1993.
12. Roland, B.; Vandendriessche, J.; Lombaerts, R.; Denturck, B.; Jakus, C. Thermal crosslinking by unexposed naphthoquinone diazides as diffusion inhibition mechanism in the DESIRE process. Proc. SPIE **1988**, *920*, 120–127.
13. Vesser, R.J.; Schellenkens, J.P.W.; Reuhman-Huisken, M.S.; Van Ijzendoorn, L.J. Mechanism and kinetics of silylation of resist layers from the gas phase. Proc. Soc. Photo-Opt. Instrum. Eng. **1987**, *711*, 111–117.
14. Moreau, W.M. *Semiconductor Lithography: Principles, Practices, and Materials*; Plenum Publishing: New York, 1988.
15. DeForest, W.S. *Photoresists: Material and Processes*; McGraw Hill, 1975.
16. Zelentsov, S.V.; Zelentsova, N.V.; Zhezlov, A.B.; Oleinik, A.V. Photo-oxidation of aromatic azides. High Energy Chem. **2000**, *34* (3), 164–171.
17. Allen, N.S. Photoinitiators for UV visible curing of coatings: mechanisms and properties. J. Photochem. Photobiol. A: Chem. **1996**, *100*, 101–107.
18. Chapman, G.H.; Sarunic, M.V. Dry Multiplayer Inorganic Alloy Thermal Resist for Lithographic Processing and Image. PCT Patent WO 02/06897 A2, January 24, 2002.
19. Kaliteevskaya, N.A.; Seisyan, R.P. Contrast enhancement in image transfer via interaction of UV radiation with inorganic photoresist films. Semiconductor **2001**, *35* (2), 226–229.
20. Lavine, J.M.; Lis, S.A.; Masters, J.I. Submicron optical lithography using inorganic resists. Proc. Soc. Photo-Opt. Instrum. Eng. **1983**, *393*, 41–48.
21. MacDonald, S.A.; Willson, C.G.; Frechet, M.J.J. Chemical amplification in high resolution imaging systems. Acc. Chem. Res. **1994**, *27* (6), 151–158.
22. http://www.nist.gov/sigmax/poster03/tengjiao-abst.htm (accessed January 2003).
23. Dektar, J.L.; Hacker, N.P. Comparison of photochemistry of diarylchloronium, diarylbromonium, and diaryliodonium salt. J. Organic Chemistry **1991**, *56* (5), 1838–1844.
24. MacDonald, S.A.; Clecak, N.J.; Wendt, H.R.; Willson, C.G.; Snyder, C.D.; Knors, C.J.; Deyoe, N.B.; Maltabes, J.G.; Morrow, J.R.; McGuire, A.E.; Holmes, S.J. Airborne chemical contamination of a chemically amplified resist. Proc. SPIE **1991**, *1466*, 2–12.
25. Vigil, J.C.; Barrick, M.W.; Grafe, T.H. Contamination control for processing DUV chemically amplified photoresists. Proc. SPIE **1995**, *2438*, 626–643.
26. Kurihara, M.; Segawa, T.; Okuno, D.; Hayashi, N.; Sano, H. Performance of a chemically amplified positive resist for next-generation photomask fabrication. Proc. SPIE **1998**, *3412*, 279–291.
27. Cui, Z.; Prewett, P.D. Comparative study of AZPF514 and UVIII chemically amplified resists for electron beam nanolithography. Microelectron. Eng. **1999**, *46* (1–4), 255–258.
28. Umbach, C.P.; Broers, A.N.; Willson, C.G.; Koch, R.; Laibowitz, R.B. Nanolithography with an acid catalyzed resist. J. Vac. Sci. Technol. B **1988**, *6* (1), 319–322.
29. Dao, T.T.; Spence, C.A.; Hess, D.W. Study of silylation mechanisms and kinetics through variations in silylating agent and resin. Proc. SPIE **1991**, *1466*, 257–268.
30. Sheats, J.K., Smith, B.W., Eds. *Microlithography Science and Technology*; Marcel Dekker, Inc., 1998.
31. Seeger, D.E.; La Tulipe, D.C.; Kunz, R.R.; Garza, C.M.; Hanratty, M.A. Thin-film imaging: past, present, prognosis, IBM J. Res. Dev. **1997**, *41* (1/2), 81–88.
32. Sezi, R.; Sebald, M.; Leuschner, R.; Ahne, H.; Birkle, S.; Borndorfer, H. Benefits and prospects of aqueous silylation for novel dry developable high-resolution resists. Proc. SPIE **1990**, *1262*, 84–93.
33. Taira, K.; Takahashi, J.; Yanagihara, K. Effect of silulation condition on the silylated image in the DESIRE process, Proc. SPIE **1991**, *1466*, 570–582.
34. Coopmans, F.; Ronald, B. DESIRE: a novel dry developed resist system. Proc. SPIE **1986**, *631*, 34–39.
35. Pierrat, C.; Tedesco, S.; Vinet, F.; Mourier, T.; Lerme, M.; Dal'Zotto, B.; Guiber, C.J. PRIME process for deep UV and E-beam lithography. Microelectron. Eng. **1990**, *11* (1–4), 507–514.

36. Louis, D.; Laporte, P.; Molle, P.; Ullmann, H. Deep ultraviolet positive resist image by dry etching (DUV PRIME): a robust process for 0.3 μm contact holes. Proc. SPIE **1994**, *2195*, 497–505.
37. Mutsaers, C.M.J.; Vollenbroek, F.A.; Nijssen, W.P.M.; Visser, R.J. IMRE, BIM and SUPER using patternwise esterification. Microelectron. Eng. **1990**, *11* (1–4), 497–502.
38. Pavelchek, E.K.; Bohland, J.F.; Thackery, J.W.; Orsula, G.W.; Jones, S.K.; Dudley, B.W.; Bobbio, S.M.; Freeman, P.W. Silylated acid hardened resist. J. Vac. Sci. Technol. B **1990**, *8*, 1497.
39. Palmateer, S.C.; Kunz, R.R.; Horn, M.W.; Forte, A.R.; Rothschild, M. Optimization of a 193-nm silylation process for sub-0.25-um lithography. Proc. SPIE **1995**, *2438b*, 455–464.
40. Hutchinson, J.; Rao, V.; Zhang, G.; Pawloski, A.; Fonseca, C.; Chambers, J.; Holl, S.; Das, S.; Henderson, C.; Wheeler, D. Progress in 193-nm top-surface imaging process development. Proc. SPIE **1998**, *3333*, 165–175.
41. Maeda, K.; Ohfuji, T.; Aizaki, N.; Hasegawa, E. High-resolution surface imaging process using difunctional silylating reagent B(DMA)MS for ArF excimer laser lithography. Proc. SPIE **1995**, *2438b*, 465–473.
42. Hartney, M.A.; Johnson, D.W.; Spencer, A.C. Evaluation of phenolic resists for 193 nm surface imaging. Proc. SPIE **1991**, *1466*, 238–247.
43. MacDonald, S.A.; Schlosser, H.; Ito, H.; Clecak, N.; Willson, C.G. Plasma developable photoresist systems based on chemical amplification. Chem. Mater. **1991**, *3*, 435–442.
44. Ito, H.; MacDonald, S.A.; Miller, R.D.; Willson, C.G. Radiation Sensitive and Oxygen Plasma Developable Resist U.S. Patent 4,552,833, November 12, 1985.
45. Johnson, H.F.; Jamieson, A.T.; Ozair, S.N.; Sahban, N.; Farmer, N.; Hogan, Z.; MacDonald, S.; Willson, C.G. Material design and characterization for cationic graft polymerization lithography. Transport phenomenon in electronic materials processing, AIChE National Meeting, 2003, San Fransisco, CA, U.S.A.
46. Johnson, H.F.; Ozair, S.N.; Jamieson, A.T.; Trinque, B.C.; Brodsky, C.J.; Willson, C.G. Cationic graft polymerization lithography, Proc. SPIE **2003**, *5037*, 943–951.
47. Brodsky, C.J.; Trinque, B.C.; Johnson, H.F.; Willson, C.G. Advances in graft polymerization lithography, Proc. SPIE **2001**, *4343*, 416–420.
48. Jamieson, A.T.; Somervell, M.; Tran, H.V.; Hung, R.; MacDonald, S.; Willson, C.G. Top surface imaging at 157 nm, Proc. SPIE **2001**, *4343*, 406–416.
49. http://dot.che.gatech.edu/henderson/Research.htm (accessed September 2004).
50. http://willson.cm.utexas.edu/Research/Sub-Files/Cationic-Graft/index.htm (accessed October 2004).
51. Chen, X.C.; Tolbert, L.M.; Hess, D.W.; Henderson, C.L. New polymers for thin film lithography. Macromolecules **2001**, *34* (12), 4104–4108.
52. Zelentsov, S.V.; Zelentsova, N.V.; Oleinik, A.V. Novel aromatic azide applications in lithography, Book of abstracts, Part 1. In *Russian Conference Microelectronics'94*, Zvenigorod, 28 November–3 December, 1994, Moscow; 183–184.
53. Woods, J.G.; Rooney, J.M. Imaging Method for Vapor Deposited Photoresists of Anionically Polymerizable Monomer U.S. Patent 4,675,270, June 23, 1987.
54. Hult, A.; Ito, H.; MacDonald, S.A.; Willson, C.G. Process for Preparing Negative Relief Imaging with Cationic Photopolymerization U.S. Patent 4,551,418, November 5, 1985.
55. Brault, R.G. Graft Polymerized SiO_2 Lithographic Masks. PCT Patent WO 85/02030, May 5, 1985.
56. Morita, M.; Imamura, S.; Tamamura, T.; Kogure, O.; Murase, K. Direct pattern fabrication on silicon resin by vapor phase electron beam polymerization. J. Vac. Sci. Technol. B. **1983**, *1* (4), 1171–1173.
57. Morita, M.; Imamura, S.; Tamamura, T.; Kogure, O.; Murase, K. Dry developable multiplayer resist using direct pattern formation by electron beam-induce vapor-phase polymerization. J. Electrochem. Soc. **1984**, *131* (3), 653–654.
58. Steckl, A.J.; Moore, J.A.; Corelli, J.C.; Lin, W.-T. Image enhancement in high-resolution lithography through polymer grafting techniques. Symp. VLSI Technol., San Diego, 10–12 September 1984; Dig. Technol. Pap., Tokyo; 60–61.
59. Stewart, M.D.; Patterson, K.; Somervell, M.H.; Willson, C.G. Organic imaging materials: a view of the future. J. Phys. Org. Chem. **2000**, *13* (12), 767–774.
60. Gould, C. Advanced process control: basic functionality requirements for lithography. Proc. IEEE Advanced Semiconductor Manufacturing Conference XII **2001**, 49–53.
61. Zeidler, A.; Veenstra, K.J.; Zavecz, T.E. Advanced statistical process control: controlling sub-0.18 μm Lithography and other processes. Proc. SPIE **2001**, *4344*, 312–322.
62. Mack, C.A. PROLITH: a comprehensive optical lithography model. Proc. SPIE **1985**, *538*, 207–220.

63. Zelentsov, S.V.; Aranson, S.Kh.; Beliakov, L.A. Solving the problem of photochemical kinetics in a medium with low reagent mobilities. J. Math. Chem. **2003**, *33* (1), 39–54.
64. Dill, F.H.; Hornberger, W.P.; Hauge, P.S.; Shaw, J.M. Characterization of positive photoresists. IEEE Trans. Electron Devices **1975**, *22* (4), 445–452.
65. Henderson, C.L.; Pancholi, S.N.; Chowdury, S.A.; Willson, C.G.; Dammel, R.R. Photoresist characterization for lithography simulation. Part 2. Exposure parameter measurements. Proc. SPIE-Int. Soc. Eng. **1997**, *3049*, 816–828.
66. Henderson, C.L.; Tsiartas, P.C.; Pancholi, S.N.; Chowdury, S.A.; Dombrowski, K.D.; Willson, C.G.; Dammel, R.R. Photoresist characterization for lithography simulation. Part 3. Development parameters measurements. Proc. SPIE-Int. Soc. Eng. **1997**, *3049*, 805–815.
67. Henderson, C.L.; Scheer, S.A.; Tsiartas, P.C.; Rathsack, B.M.; Sagan, J.P.; Dammel, R.R.; Erdmann, A.; Willson, C.G. Modeling parameter extraction for DNQ-novolak thick film resists. Proc. SPIE-Int. Soc. Eng. **1998**, *3333*, 256–267.
68. Burns, S.; Schmid, G.; Tsiartas, P.; William, C.G. Advancements to the critical ionization dissolution model. J. Vac. Sci. Technol. B **2002**, *20* (2), 537–543.

Photovoltaic Materials

Richard Corkish
ARC Centre of Excellence for Advanced Silicon Photovoltaics and Photonics, University of New South Wales, Sydney, New South Wales, Australia

INTRODUCTION

Photovoltaic materials are photoconductive semiconductors able to absorb energy from photons (wave packets of light) to release electrons and holes (negative and positive charge carriers) from their host atoms and separate the different types of charge carriers. The energy quantum required to release an electron is termed the *bandgap* of the material and is in the range of 0.5–2 eV for commercially important solar cell materials. Freed electrons and the holes, which are essentially vacancies from which electrons have been excited, are mobile within the material and their motion is influenced by electric fields and diffusion. Photovoltaic cells contain some asymmetry that forces electrons and holes toward negative and positive contacts, respectively. That asymmetry is commonly a p–n junction, being the interface of p-type material that is doped with impurity atoms having a deficit of electrons relative to the host material and n-type material that is doped with atoms having an excess of electrons relative to the host. Hence, cells produce voltage and current (power) under illumination.

This entry gives a brief historical overview and discussion of relevant material properties, followed by a survey of the methods currently used to produce solar cell materials, with an emphasis on silicon.

HISTORY OF PHOTOVOLTAIC MATERIALS

A very large number of materials[1] have been discovered to be photovoltaically active since Alexandre–Edmond Becquerel noted the photoelectrochemical production of electricity from light.[2] Platinum, brass, or silver electrodes were inserted in a liquid electrolyte, one electrode was illuminated and current between the electrodes was observed with a galvanometer. The effect was (transiently) enhanced by coating the illuminated electrode with a semiconductor, such as copper oxide or silver halide. This discovery is generally acknowledged as the beginning of photovoltaics, although it would instead be termed photoelectrochemistry today. It was not until 1914 that a photoelectric, rather than entirely chemical, interpretation was made.

Selenium was discovered in 1876 to exhibit the photovoltaic effect in the solid state.[3] It was rediscovered more than once in Europe and in the United States until the 1930s and 1940s, by which time there were several companies manufacturing selenium photovoltaic cells as light sensors for photography and industrial applications. Selenium devices were then in competition with copper–cuprous oxide (Cu–Cu_2O) heterojunctions, in which the photovoltaic effect had been discovered in 1917.[1] The late 1920s and early 1930s was a time of intense activity and competition in the development of cuprous oxide rectifiers and photovoltaic cells. Commercial devices were produced from 1930 but were eventually surpassed by improved selenium devices in that decade.

Photovoltages have been identified in a range of other materials, including Ag_2S, MoS_2, Tl_2S, Ag_2S, Pb_2S, Ag_3AsS_3, Ag_5SbS_4, CdS, a range of BiI_3–PbI_2 alloys, and germanium.[1] However, these were unimportant, relative to selenium and cuprous oxide, and all were surpassed by the discovery of the effect in silicon in the 1940s.

Russel Ohl of Bell Laboratories discovered and filed patents in 1941 that were granted in 1946 and 1948[4] for the p–n junction photovoltaic effect in silicon. The early devices contained rectifying junctions that were accidentally formed by segregation of dopant impurities during cooling.

MATERIAL ISSUES FOR PHOTOVOLTAICS

Bandgap is a fundamental material property for photovoltaics. The bandgap restricts the range of wavelengths of light that can be absorbed by the material and limits the voltage it may produce. To a first approximation, only photons with energy in excess of the bandgap may be absorbed to produce electron–hole pairs, while the material is transparent to lower energy photons and each photon can produce only a single electron–hole pair, even those that carry much more than the bandgap energy. Hence, there is an optimization to maximize output power: A lower bandgap will allow the useful absorption of more photons and production of more current but the voltage will tend to be low. The optimum bandgap for reception of

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007896

the solar spectrum has been calculated to be in the range 1–1.5 eV.

Between the stages of initial creation of an electron–hole pair from absorption of a photon and eventual collection of the charge carriers at the contacts, the carriers are vulnerable to being lost through recombination. Recombination is encouraged by defects within the material, such as defects in crystal structure and impurities and by surfaces. Electrically active impurities and defects introduce allowed energy states into the otherwise forbidden bandgap energy range for charge carriers. Surfaces need to be passivated to reduce their impact on performance, for example, by thermally grown silicon dioxide on the surfaces of silicon cells.

Other important properties for photovoltaic materials are their refractive index, stability, brittleness, toxicity, crystal lattice constant, thermal expansion coefficient, temperatures required for processing into cells, energy investment for cell production, ability to be doped both types, level of technological knowledge and industrial maturity, cost, and abundance. Issues particular to passivation and the trapping of weakly absorbed light include the availability of compatible and affordable passivation and surface texturing methods.[5]

GROUP IV MATERIALS

Silicon is by far the most important semiconductor for electronics and photovoltaics since the 1960s and for the foreseeable future. It is used in many forms, with monocrystalline and multicrystalline (MC) self-supporting silicon wafers being dominant but also in the forms of deposited thin-films in amorphous or crystalline form. The structure of one commercially important crystalline silicon cell technology is shown in Fig. 1.

Metallurgical Grade Silicon

Silicon is produced in large quantities as metallurgical grade (MG) material for use in the metals and chemicals industries and some is further refined to produce ultrapure electronics grade (EG) silicon. MG silicon is made from lumps of quartz (10–100 mm) and carbon in the form of MG coal, woodchips, charcoal, or coke.[6] They are mixed and heated by an electric arc between carbon electrodes and the furnace body. The reduction of quartz to silicon proceeds according to the (principal) reaction,

$$SiO_2 + 2C \rightarrow Si\ (liquid) + 2CO\ (gas)$$

Additional, incidental reactions are discussed by Ceccaroli and Lohne.[6] The liquid crude silicon has 1–3% impurities, principally aluminum, calcium, iron, titanium, and carbon.

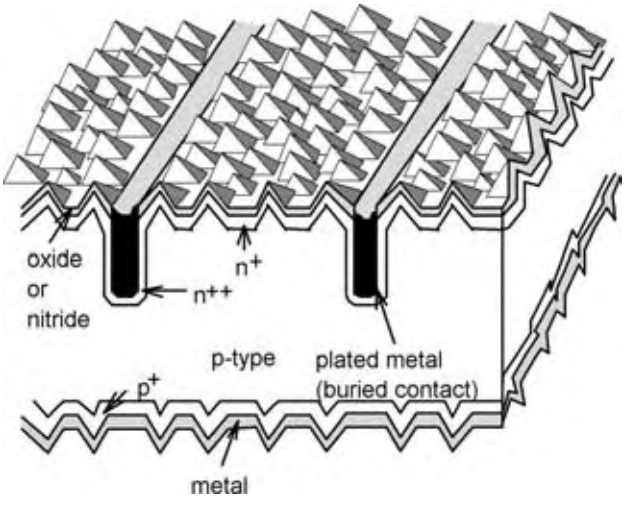


Fig. 1 Buried contact solar cell structure, an important wafer-based commercial silicon cell technology. Features include surface texturing for light trapping, diffusions front and rear and the front current grid buried in laser grooves. (Courtesy of UNSW Centre for Photovoltaic Engineering Image Library.)

Oxidizing gas and slag-forming materials, such as silica (SiO_2) sand, lime (CaO), limestone ($CaCO_3$), dolomite (CaO–MgO), or calcium fluoride (CaF_2) are added to the still-molten raw silicon in a refinement step. Less noble elements than silicon are oxidized and dissolved in the slag, which is removed. To avoid excessive heat losses, some of the silica can be provided directly by oxidation of some of the silicon melt instead of sand. Carbon is removed from the melt in the form of SiC precipitates.

The silicon is solidified in a mold, forming MC blocks with impurities congregating at grain boundaries. The blocks are then crushed into pieces up to 100 mm in size and the ''fines,'' which are both inconvenient for most subsequent processes and contaminated by impurities, are removed. An alternative process is to use granulation in water to avoid casting and crushing.

Semiconductor Grade Polycrystalline Silicon

Aside from experimental and developing processes,[7] there are three competing commercial processes for further refining metallurgical silicon into semiconductor grade polycrystalline silicon (polysilicon):[6,8] Siemens, Union Carbide Chemicals, and Ethyl Corporation methods. This industry exists primarily to support the electronics industry.

In the energy intensive Siemens process (Fig. 2), MG silicon and hydrochloric acid are combined in a

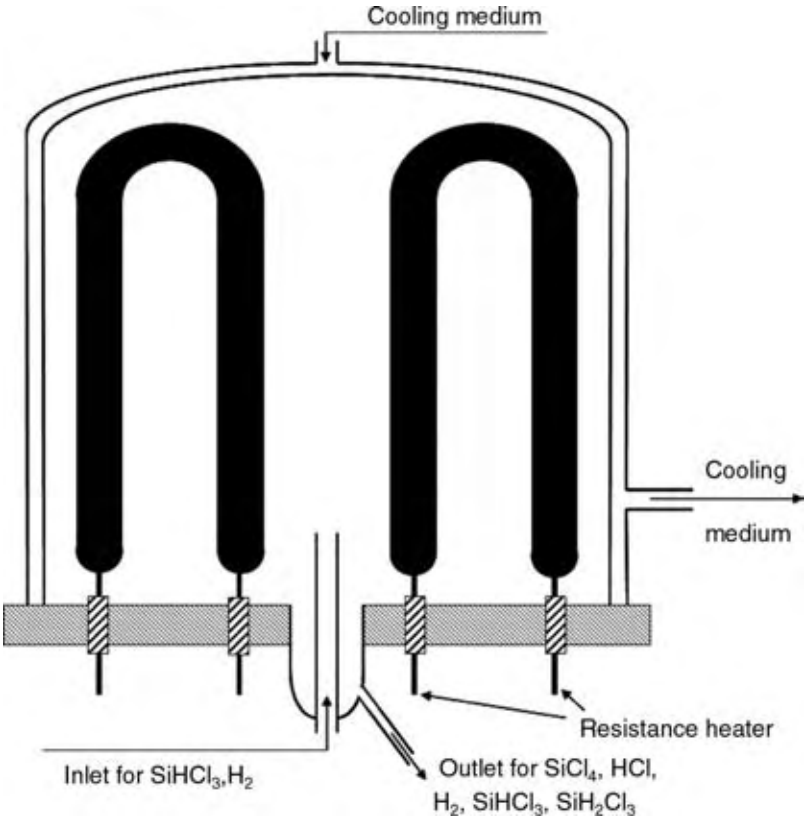


Fig. 2 Schematic of Siemens bell-jar reactor. (From Ref.[6].)

fluidized bed reactor to produce volatile trichlorosilane:

$$Si\ (solid) + 3HCl \rightarrow HSiCl_3 + H_2$$

Competing reactions result in other silanes, including $SiCl_4$, SiH_4, and SiH_2Cl_2, which are removed by fractional distillation in two stages.

The pure trichlorosilane is vaporized, mixed with hydrogen, decomposed into silicon seed rods, and heated to 1100°C. Deposition occurs on multiple inverted-U seed rods under steel bell jars (Fig. 2), cooled to avoid deposition on the inner surface. The main reactions are

$$2SiHCl_3 \rightarrow SiH_2Cl_2 + SiCl_4$$
$$SiH_2Cl_2 \rightarrow Si + 2HCl$$
$$H_2 + 2SiHCl_3 \rightarrow Si + 3HCl$$
$$HCl + SiHCl_3 \rightarrow SiCl_4 + H_2$$

Unfortunately, more of the source silicon results in $SiCl_4$ byproduct than in deposited polysilicon. The market for alternative uses for $SiCl_4$ has not kept up with polysilicon production rates, hence recycling into $SiHCl_3$ is now preferred.

One company is modifying the final polysilicon deposition step in the Siemens process.[7] The new method uses a graphite pipe heated to 1500°C, beyond the melting point of silicon. Liquid silicon precipitates from liquid trichlorosilane poured over the graphite and drips and hardens into granules. Another is reported to be trialing an alternative method to produce tiny polysilicon granules by blowing trichlorosilane into a fluidized bed reactor.[7]

The Union Carbide process for trichlorosilane production is an alternative. Tetrachlorosilane is hydrogenated through a bed of granular MG silicon, using one of the reactions commonly used for the recycling of the byproduct $SiCl_4$ in the Siemens process:

$$3SiCl_4 + 2H_2 + Si \rightarrow 4SiHCl_3$$

The trichlorosilane is separated by distillation from unreacted tetrachlorosilane, which is returned to the reactor. The trichlorosilane is then passed through quaternary ammonium ion exchange resins that catalyze the following two reactions,

$$2SiHCl_3 \rightarrow SiH_2Cl_2 + SiCl_4$$
$$3SiH_2Cl_2 \rightarrow SiH_4 + 2SiHCl_3$$

The monosilane, SiH_4, is distilled and pyrolyzed to deposit polysilicon on heated silicon seed rods in a metal bell jar,

$$SiH_4 \rightarrow 2H_2 + Si$$

and the $SiCl_4$ and $SiHCl_3$ are recycled to the earlier steps described above. The practices of recycling the byproducts and distilling each time results in a very high purity output and reduced input requirements, because the hydrogen and chlorine are both recycled.

The third commercial process for production of polysilicon was developed by the Ethyl Corporation. In this case, the raw material is not MG silicon but silicon fluoride, a waste product from fertilizer manufacture, which is reacted with lithium aluminum hydride or sodium aluminum hydride, with saleable byproduct of lithium aluminum fluoride or sodium aluminum fluoride,

$$SiF_4 + AlXH_4 \rightarrow SiH_4 + AlXF_4$$

where X is either Li or Na. The monosilane is pyrolized to polysilicon as above except that instead of a bell jar, the seed crystals are granules in a fluidized bed. A company is reported to be about to start production of granular polysilicon, especially for the photovoltaics market, from SiH_4 using fluidized bed reactors.[7] Another company instead, intends to precipitate silicon from silane onto hollow silicon cylinders at 800°C.

Other recent developments include one that avoids the conventional path of chemical purification of MG silicon, but uses a metallurgical refining step instead.[7] This process is claimed to be less reliant on capital and to reduce the energy requirement by 20% compared to the Siemens process. An experimental, two-stage carbothermic process is at the stage of pilot production and is under consideration.[7] Silicon carbide is made from quartz and carbon black powder in a plasma furnace. Then, the silicon carbide is converted to silicon in an electric arc furnace and the silicon is purified to remove residual carbon. The intention is to use high purity starting materials and thereby avoid one purification step altogether. Ceccaroli and Lohne[6] list a large number of alternative processes that have been or are being explored with the aim of increasing throughput, reducing energy demand, reducing costs, and facilitating continuous rather than batch operation, relative to the current commercial processes for production of polysilicon.

Crystalline Silicon

Until 1998, the silicon photovoltaics industry was able to rely on off-specification EG material, including rejects from polysilicon manufacture (material from aborted processes, faulty seeds, ends of seed rods close to the carbon contacts, fines and chips, dendritic growths, small granules, sawn sections of seed rods) and from crystal growth (heads and tails of monocrystalline boules, ingots from aborted processes, crucible leavings, pot scrap) that was effectively waste of the electronics industry. This amounted to 2000–3000 tons per year. However, the rapid growth of photovoltaics has resulted in demand exceeding that supply in recent years and shortages will occur until new capacity comes on line in 2006–2007.[7]

Purity requirements for silicon for solar cell manufacture are not as severe as for the electronics industry, with silicon with up to 4 ppm of carbon, 5 ppm of oxygen, 0.3 ppm of boron and 0.1 ppm of phosphorous, and metals having been used.[6] Carbon (Group IV) is an electrically neutral, substitutional impurity in silicon. It has smaller atoms and can be associated with species with larger lattice spacing, such as silicon dioxide. Oxygen, often originating from fused-quartz crucibles, is a common interstitial impurity. It is implicated in gradual degradation in silicon solar cells made from boron-doped (p-type) wafers.[9] The 3d transition metals, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, and Cu are additional common interstitial impurities. They often form complexes at structural crystal defects.

We adopt Basore's nomenclature (Table 1) for different forms of crystalline silicon.[10]

Single Crystal Silicon

Single crystal silicon wafers currently satisfy a significant but declining fraction of the solar cell market (36.4% in 2002, 32.2% in 2003) although there are signs of a possible resurgence.[11] Relative to MC, monocrystalline wafers have the advantages that they were already produced for other markets and they allow higher efficiency cells. On the other hand, they require higher quality feedstock, tend to be more expensive, and are circular, leading to lower packing densities in photovoltaic modules unless they are cut into squares or pseudo-squares.

Wafers grown by the Czochralski process[12,13] dominate the monocrystalline fraction of the photovoltaic market. The process has a purifying effect with respect to several important metallic impurities, to the extent that even poor quality starting material can produce good solar cells. The polysilicon source material is melted in a crucible under vacuum or inert

Table 1 Nomenclature for crystalline silicon

Crystallinity	Symbol	Grain size range
Single crystal	sc-Si	>10 cm
Multicrystalline	mc-Si	1 mm–10 cm
Polycrystalline	pc-Si	1 μm–1 mm
Microcrystalline	μc-Si	<1 μm

(From Ref.[10].)

gas and a seed crystal is dipped into the melt and slowly withdrawn vertically and rotated so that the liquid crystallizes to the seed (Fig. 3). The melt temperature and pulling speed are controlled to ensure a constant diameter of the growing monocrystalline boule and low dislocation density. The normal crystal orientation is ⟨100⟩ because this allows dislocations, which propagate on oblique (111) planes, to rapidly ''grow out'' as the top neck is forming and permits the use of simple NaOH anisotropic etching methods on solar cells for surface texturing to aid light-trapping and reduce reflection.

Very pure, low-defect crystals are produced by the alternative float zone process[13] in which a cylindrical polysilicon rod has a seed crystal melted into its lower end with an encircling inductive heating coil. The heater is then raised along the rod length, entraining a molten zone and a monocrystalline solidified region below. Impurities have higher solubility in the molten silicon and are carried with it to the top. The process may be repeated to enhance the purity and crystallinity.

Multicrystalline Silicon

Multicrystalline silicon wafers, with crystal sizes in the range 1–100 mm,[10] are currently the main workhorse for the photovoltaics industry. Hydrogen passivation steps used in cell manufacture in recent years, particularly those involved in the deposition of silicon nitride layers on front surfaces, have reduced the impact of electrically active impurities and defects in MC-Si cells and reduced the performance deficit relative to monocrystalline cells.

Three main processes are in use to produce MC-Si ingots—tricrystalline growth, the Bridgeman, and block casting methods, with Bridgeman more common than the others.[14] The tricrystalline process[11] results in ingots consisting of just three grains with all boundaries perpendicular to the (110) plane meeting at the ingot axis. Growth is carried out similar to the Czochralski method described above but a tri-Si seed is used and faster growth is possible. The pulling axis is parallel to the ⟨110⟩ orientation of each grain, hence dislocations do not naturally grow out during the top neck formation. The resulting wafers have higher mechanical strength than other MC wafers, allowing thinner wafers to be used for a given yield.

The Bridgeman and block casting methods both produce ingots with weights 250–300 kg. In the Bridgeman process, a nitride-coated crucible containing the silicon source material is slowly moved downward through an inductive heating zone so that the silicon melts and then solidifies from the bottom. The upward vertical movement of the liquid–solid interface, known as directional solidification, results in large columnar grains. The purpose of the nitride crucible coating is to allow slippage of the silicon as it expands and contracts.

In the block casting process, source silicon is melted in a first quartz crucible and poured into a second, nitride-coated crucible. Heaters are used to control cooling to produce columnar grains. Relative to Bridgeman, faster solidification is possible but at the expense of smaller grains.

The detrimental effects of grain boundaries on cell performance depend on their density and on their electrical activity, which depends on their decoration

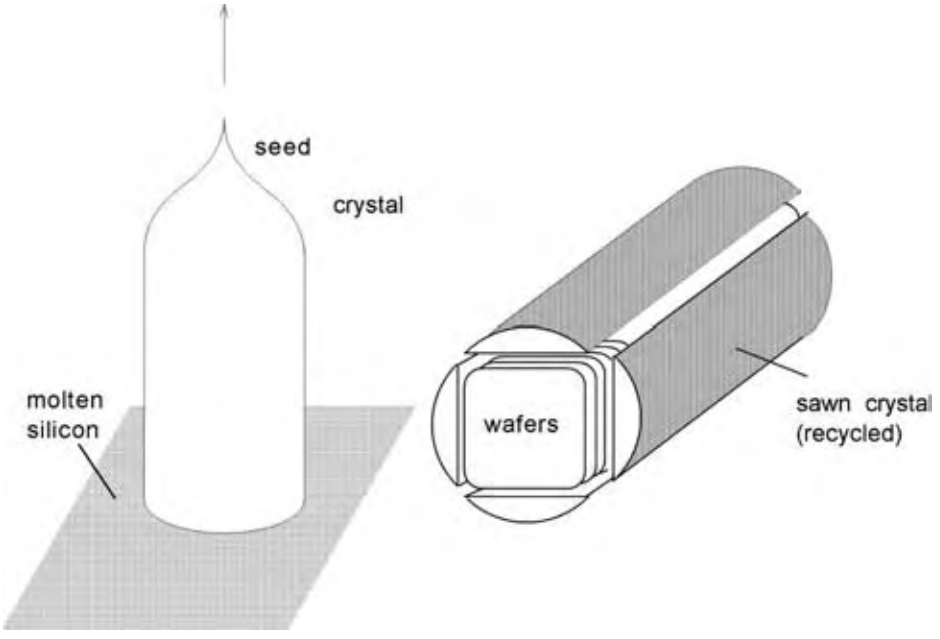


Fig. 3 Czochralski crystal pulling and wafering. (Courtesy of UNSW Centre for Photovoltaic Engineering Image Library.)

with impurities, particularly transition metals. Defects with more impurities facilitate additional recombination of charge carriers, which should be minimized for good cell performance. Keeping the liquid–solid interface planar has been shown to reduce grain boundary activity. Dislocations within crystal grains are also effective in reducing cell efficiency and these are reduced by limiting temperature inhomogeneities during cooling. Oxygen is incorporated into the ingot from the quartz crucible. Carbon from graphite heaters reacts with SiO to produce CO, which leads to carbon impurity incorporation in ingots, especially in the upper region. Excessive carbon can lead to SiC needles short circuiting the p–n junction in a cell. Crystallization is deliberately managed to segregate active metallic impurities into the top portion of an ingot, which is rejected.

Standard MC silicon is boron (p-type) doped to a concentration around $2 \times 10^{16}\,\mathrm{cm}^{-3}$, yielding electrical resistivity around 0.1–5 Ωcm. The doping is achieved by adding B_2O_3 to the raw silicon before melting. n-Type regions are produced during subsequent cell manufacture by diffusion of phosphorous, either by bubbling nitrogen carrier gas through $POCl_3$ and injecting into a quartz furnace or by phosphorous-containing pastes if a belt furnace is used.[14]

Numerous methods have been tried to produce silicon directly in thicknesses suitable for cell manufacture, thereby avoiding the wasteful process of sawing wafers from ingots.[12] These methods draw ribbons or foils of silicon from the melt either vertically or horizontally and result in MC material. Some of these methods are already commercialized and others are at various stages of development. *Edge-defined film growth* is a vertical film growth technique in which a seed crystal is lowered to the melt surface and a film is pulled upwards through a graphite die by capillary action. No die is used in the *dendritic web* method. As the seed is withdrawn, two dendrites propagate from its ends into the melt and the film grows as a meniscus. The *string ribbon* technique is similar except that two filaments, passing through the bottom of the crucible, are used to define the film edges. Horizontal growth is carried out on a moving, reusable graphite or ceramic substrate from a silicon reservoir in a die in the *ribbon growth on substrate* scheme. The ribbon, 100 μm or thicker, can be separated from the substrate by thermal expansion mismatch. In the *silicon film* process, the substrate of ceramic, graphite cloth, or stainless steel is retained as part of the solar module.

Development is underway on tiny (650–750 μm) spherical solar cells.[15,16] Each cell is formed in a silicon bead rather than from a wafer. To form the beads, droplets of molten silicon fall through vacuum in a 14 m tower, solidifying into MC spheres.

Thin-Film Crystalline Silicon

Thinner cells are sought to reduce material and energy usage, reduce bulk recombination, relax material quality requirements, and potentially allow lightweight and flexible cells.[5,17] One path is to use thinner wafers but production yield then tends to drop. Alternatively, the growth of crystalline thin-films on cheap substrates, particularly glass, is a major R&D goal.[5,18–23] The usual path is to deposit amorphous silicon or nanocrystalline silicon and crystallize it. The essential problem is that crystallization into large grains requires high temperatures, which can make some substrates unsuitable. If processing is restricted to low temperatures, it is possible to form nanocrystalline silicon simply by adjusting the hydrogen content in the normal deposition process for amorphous silicon.[18]

Various methods have been tried[5] to enlarge grains, including postdeposition recrystallization and aluminum-induced crystallization. The Kaneka group uses a glass substrate with reflective coating, chemical vapor deposition (CVD)-deposited n-type microcrystalline silicon, undoped polysilicon, and a p-type polysilicon layer. CSG Solar uses solid phase crystallization of amorphous silicon on glass, deposited by plasma enhanced chemical vapor deposition (PECVD), and intends to overcome the common problems with layer uniformity, durability, and yield.[24] Other methods in the research phase include cheaper deposition by evaporation and aluminum-induced crystallization, in which amorphous silicon deposited on an aluminum layer on a borosilicate glass substrate is heated and the aluminum migrates through the silicon layer, which crystallizes.[5,23]

Amorphous Silicon

Hydrogen is usually incorporated into amorphous silicon to reduce recombination at hanging bonds, forming a-Si : H.[17,18] It requires lower film thickness (~0.4 μm) than crystalline silicon for virtually complete absorption of sunlight.

However, a serious (although self-limiting and reversible by annealing for a few hours) light-induced degradation effect has severely restricted the promise of this technology. It is a result of metastable material defects created by the light. The hydrogen that is routinely incorporated into the material to neutralize unsatisfied atomic bonds also creates unsatisfied bonds under the influence of light. Degradation has been minimized by making the cells as thin as possible and this has encouraged the use of tandem structures to achieve adequate light absorption. Tandem cells are stacks of p–n junctions, each of which may have a different bandgap, each responding to a different section

of the solar spectrum, yielding higher overall efficiency. Triple-junction tandem modules, including an amorphous germanium junction, are commercially available.

a-Si : H is usually produced by various forms of PECVD of silane[17] at about 100–300°C. The film quality depends on many deposition parameters, including the substrate temperature, which controls the incorporation of hydrogen. The low temperature deposition technology allows plastics to be used as substrates in flexible and stick-on modules.

Sanyo's HIT (*heterojunction with intrinsic thin-layer*) technology combines crystalline and amorphous silicon. n-Type crystalline silicon wafers have layers of intrinsic and doped amorphous silicon applied to the front and rear. Contacts are made via transparent conducting oxide (TCO).

Germanium, Silicon Oxides, Silicon Nitrides, Silicon Carbides

Other Group IV materials play subsidiary or relatively minor roles in photovoltaics. Most important are silicon dioxide and silicon nitride that are used to passivate silicon surfaces and to reduce reflection of light. Native dioxide has been used for this task historically but silicon nitride applied by PECVD has recently gained increasing acceptance in the manufacture of MC solar cells because of the effect that incidental hydrogen incorporation has on passivation of grain boundaries. As mentioned above, amorphous germanium is used for low-bandgap cells beneath amorphous silicon in multijunction structures, to better match the solar spectrum, but amorphous silicon carbide has not been found to be of sufficient quality for high bandgap cells.[17] Silicon nanostructures (quantum wells, quantum dots, and superlattices) have been proposed to be formed in oxides, nitrides, and carbides in preliminary research into new solar cell structures such as all-silicon multijunction cells and hot carrier cells.[25–27] The cubic phase of silicon carbide has been studied as a way to use impurities, boron in this case, to advantage.[28]

GROUP III–V PHOTOVOLTAIC MATERIALS

High efficiency solar cells are produced from crystalline binary or multinary compounds of elements from Groups III and V, almost entirely for space applications. While a wide range of materials and methods have been employed, the monolithic tandem stack of a $Ga_xIn_{1-x}P$ cell, where $x \approx 0.516$, above GaAs and Ge cells is a commercially produced, indicative example.[29] Monolithic tandem structures present several challenges. The materials used in the various junctions should have similar thermal expansion coefficients and lattice constants and the sets of bandgaps should be a good match to the spectrum.

The lowest bandgap (Ge) cell is formed by diffusion of a dopant into a Ge substrate wafer by exposure to, for example, PH_3 or AsH_3. GaAs, which is almost lattice matched, is grown on the Ge substrate by metal oxide chemical vapor deposition (MOCVD) using trimethylgallium and arsine, and forms the middle cell. The top cell is made by MOCVD of GaInP, using the above gases and phosphine. An $Al_yIn_{1-y}P$, where $y \approx 0.532$, window layer reduces carrier recombination at the front surface. Se, from decomposition of H_2Se and Si from Si_2H_6 are n-type dopants and zinc from dimethyl or diethyl zinc, magnesium from cyclopentadienyl magnesium, and carbon from CCl_4 can dope III–V compounds p-type.

GROUP II–VI PHOTOVOLTAIC MATERIALS

The basic structure of polycrystalline cadmium telluride (CdTe) thin-film cells[18,30] has a glass superstrate and a layer of TCO as front contact, a near-transparent n-type cadmium sulfide (CdS) window layer, p-type CdTe, and a metallic rear contact. The CdTe is usually deposited by three families of techniques.[30] In the first group (vapor transport deposition, close space sublimation, physical vapor deposition, and sputtering) elemental vapors of Cd and Te condense and react on the substrate. In the second (electrodeposition), Cd^{2+} and $HTeO_2^+$ ions in acidic electrolyte are galvanically reduced at the surface:

$$HTeO_2^+ + 3H+ + 4e^- \rightarrow Te + 2H_2O;$$
$$Cd^{+2} + 2e^- \rightarrow Cd; \quad Cd + Te \rightarrow CdTe$$

The third family (MOCVD, spraying, and screen printing) involves the reaction of precursors at the surface. In MOCVD, the precursors may be dimethylcadmium and diisopropyltellurium in a hydrogen carrier. The spray method uses a slurry of CdTe, $CdCl_2$ in a propylene glycol carrier. A paste of Cd, Te, $CdCl_2$, and a binder is applied through a screen in the third method in this group.

Chlorine treatments, such as heating in the presence of $CdCl_2$ vapor, dipping in solutions of $CdCl_2$: CH_3OH or $CdCl_2 : H_2O$, or exposure to HCl or Cl_2, are commonly used following chlorine-free depositions.

CHALCOPYRITE PHOTOVOLTAIC MATERIALS

Nonstoichiometric copper indium gallium diselenide [Cu(In,Ga) Se_2—abbreviated as CIGS or CIS if the gallium is absent] is another promising thin-film PV material.[18,31] Large grains can be produced on foreign

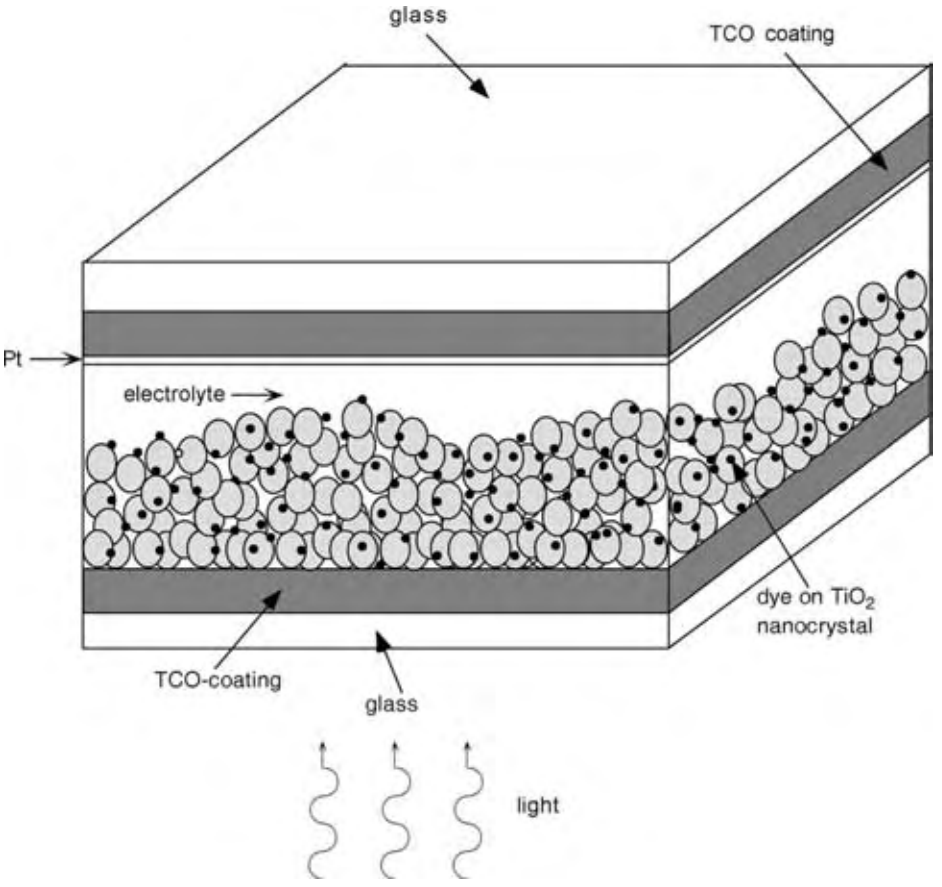


Fig. 4 Dye-sensitised nanocrystalline cell structure. (Courtesy of UNSW Centre for Photovoltaic Engineering Image Library.)

substrates, grain boundaries can be passivated, and the material is tolerant of deviations from exact ratios of the constituent elements. This material forms a heterojunction with an n-type CdS layer.

Several methods are used to deposit the films onto molybdenum-coated soda-lime glass substrates. Coevaporation of Cu, In, Se, and Ga from elemental sources onto a heated substrate, with careful control of the deposition rate of each, has achieved the highest efficiency. Another common method is selenization. A metal precursor layer of Cu, In, and Ga is deposited by, for example, sputtering and reacted in H_2Se or Se vapor. For layers without Ga, the final reaction in a series is

$$2InSe_2 + Cu_2Se + Se \rightarrow 2CuInSe_2$$

The CdS layer is grown from solution in a chemical bath of a cadmium salt, such as $CdSO_4$, $CdCl_2$, CdI_2, $Cd(CH_3COO)_2$, a ''complexing agent,'' such as ammonia and a sulfur source, commonly $SC(NH_2)_2$ (thiourea). The reaction,

$$Cd(NH)_3)_4{}^{2+} + SC(NH_2)_2 + 2OH^- \rightarrow CdS + H_2NCN + 4NH_3 + 2H_2O$$

occurs at room temperature.

A TCO layer of SnO_2, In_2O_3 : Sn (''ITO''), or ZnO interfaces to the metal electrical grid on the front of the cell. Deposition methods include dc sputtering, rf magnetron sputtering, and chemical vapor deposition.

DYE-SENSITIZED PHOTOVOLTAIC MATERIALS

Dye-sensitized nanocrystalline cells[32] are fundamentally different from the others discussed earlier in that they do not rely on semiconductor p–n junctions. Instead, they are electrochemical devices in which the optical absorption and carrier-collection processes are separated (Fig. 4).

Glass coated with TCO, normally SnO_2, has a porous film of TiO_2 (anatase phase), which is stable against photocorrosion, applied either as a colloidal solution or as a paste containing a polymer binder, such as polyethylene glycol, which is then sintered. The resulting nanoparticle film presents a large surface area of TiO_2, which is coated with a redox charge-transfer dye of ruthenium bipyridal or ruthenium terpyridine complexes.[32] The liquid electrolyte is usually an iodide/tri-iodide redox couple in a low viscosity nitrile solvent, such as acetonitrile, propionitrile, methoxyacetonitrile, or methoxypropionitrile.

A sputtered platinum counter-electrode on a TCO substrate may have its electrocatalytic activity, in the reduction of the tri-iodide ions, enhanced by creating Pt colloids from an alcoholic solution of H_2PtCl_6.

ORGANIC PHOTOVOLTAIC MATERIALS

Semiconducting polymers are in the R&D phase for several applications, including photovoltaics,[18,33,34] because they offer potential for very low cost and flexibility. Promising active materials include conjugated polymer–fullerene blends and blends of two conjugated (having conjugated carbon bonds in their backbone with alternating single and double or triple C–C bonds) polymers that form a carrier-separating heterojunction.

CONCLUSIONS

Silicon dominates the supply of material for solar cells now and for the foreseeable future for a range of practical reasons. Silicon is used in several forms, including sawn and grown wafers of single or MC material, crystalline spheres and thin-crystalline or amorphous films deposited on substrates or superstrates. The processes used to obtain MG and, subsequently, EG or solar grade silicon from raw quartz are briefly outlined, as are the fundamental processes used to make Group III–Group V, copper gallium indium diselenide, cadmium telluride, and dye-sensitized solar cells. With many competing technologies at different stages, the evolution of the industry is not clear but it is widely expected that thin-film technologies will become increasingly important relative to wafer-based methods. There is currently substantial developmental effort underway to incorporate the advantages of silicon into effective and affordable thin-film modules.

ACKNOWLEDGMENTS

The Centre for Advanced Silicon Photovoltaics and Photonics is supported by the Australian Research Council and the New South Wales Department of State and Regional Development. UNSW has a special interest in thin-film and wafer-based silicon technologies.

ARTICLES OF FURTHER INTEREST

Nanostructured Materials, p. 1825.
Renewable Energy, p. 2635.
Thin Film Science and Technology, p. 3061.

REFERENCES

1. Crossley, P.A.; Noel, G.T.; Wolf, M. *Review and Evaluation of Past Solar-Cell Development Efforts*; Contract Number NASW-1427 (AED R-3346); RCA Astro-Electronics Division for NASA: Washington, DC, 1968.
2. Becquerel, E. Recherches sur les efféts de la radiation chimique de la lumière solaire, au moyen des courants électriques. Bibliotheque Universelle de Geneve **1839**, *XXII*, 345–366.
3. Adams, W.G. The action of light on selenium (Abstract). Proc. R. Soc. Lond. **1876**, *25*, 113–117.
4. Ohl, R.S. Light-sensitive electric device. U.S. Patent 2,402,662, 25 June 1946; Light-Sensitive Electric Device Including Silicon. U.S. Patent 2,443,542, June 15, 1948.
5. Sopori, B. Thin film silicon solar cells. In *Handbook of Photovoltaic Science and Engineering*; Luque, A., Hegedus, S., Eds.; Wiley: Chichester, 2003; 307–357.
6. Ceccaroli, B.; Lohne, O. Solar grade silicon feedstock. In *Handbook of Photovoltaic Science and Engineering*; Luque, A., Hegedus, S., Eds.; Wiley: Chichester, 2003; 153–204.
7. Bernreuter, J. The delay. A lack of silicon supply for the photovoltaics industry over the next two to three years. Photon Int. **2004**, (5), 36–42.
8. Nijs, J.; Mertens, R.; Van Overstraeten, R.; Szlufcik, J.; Hukin, D.; Frisson, L. Energy payback time of crystalline silicon solar modules. Adv. Solar Energy **1998**, *11*, 291–327.
9. Schmidt, J. Light-induced degradation in crystalline silicon solar cells. Diff. Defect Data B: Solid State Phenom. **2004**, *95–96*, 187–196.
10. Basore, P.A. Defining terms for crystalline silicon solar cells. Prog. Photovolt.: Res. Appl. **1994**, *2* (2), 177–179.
11. Schmela, M. The return of mono? Market survey on silicon wafers and ingot. Photon Int. **2004**, (4), 34–38.
12. Koch, W.; Endrös, A.L.; Franke, D.; Häβler, C.; Kalejs, K.P.; Möller, H.J. Bulk crystal growth and wafering for PV. In *Handbook of Photovoltaic Science and Engineering*; Luque, A., Hegedus, S., Eds.; Wiley: Chichester, 2003; 205–254.
13. Goetzberger, A.; Knobloch, J.; Voss, B. *Crystalline Silicon Solar Cells*; Wiley: Chichester, 1998.
14. Tobías, I.; Del Cañizo, C.; Alonso, J. Crystalline silicon solar cells and modules. In *Handbook of Photovoltaic Science and Engineering*; Luque, A., Hegedus, S., Eds.; Wiley: Chichester, 2003; 255–306.
15. Nickel, J. Spheral versus Sphelar. Photon Int. **2004**, (2), 40–43.

16. Drewes, P. Spheral solar—a completely different photovoltaic technology. Proceedings of PV in Europe, Rome, Italy; WIP-Munich: Munich, 2002.
17. Deng, X.; Schiff, E.A. Amorphous silicon-based solar cells. In *Handbook of Photovoltaic Science and Engineering*; Luque, A., Hegedus, S., Eds.; Wiley: Chichester, 2003; 505–565.
18. von Roedern, B. Materials for solar energy. In *Encyclopedia of Energy*; Cleveland, C.J., Ed.; Elsevier: 2004; Vol. 5, 47–59.
19. Green, M.A. Thin-film photovoltaics. Adv. Solar Energy **2003**, *15*, 187–213.
20. Green, M.A. Crystalline and thin-film silicon solar cells: state of the art and future potential. Solar Energy **2003**, *74* (3), 181–192.
21. McCann, M.J.; Catchpole, K.R.; Weber, K.J.; Blakers, A.W. A review of thin-film crystalline silicon for solar cell applications. Part 1. Native substrates. Solar Energy Mater. Solar Cells **2001**, *68* (2), 135–171.
22. Catchpole, K.R.; McCann, M.J.; Weber, K.J.; Blakers, A.W. A review of thin-film crystalline silicon for solar cell applications. II. Foreign substrates. Solor Energy Mater. Solar Cells **2001**, *68* (2), 173–215.
23. Aberle, A.; Widenborg, P.I.; Song, D.; Straub, A.; Terry, M.L.; Walsh, T.; Sproul, A.; Campbell, P.; Inns, D.; Beilby, B.; Griffin, M.; Weber, J.; Huang, Y.; Kunz, O.; Gebs, R.; Martin-Brune, F.; Barroux, V.; Wenham, S.R. Recent advances in polycrystalline silicon thin film solar cells on glass at UNSW. Conference Record of the 31st IEEE Photovoltaics Specialists Conference, Orlando. FL, USA. Jan 3–7, 2005 (in press).
24. Basore, P.A. Simplified processing and improved efficiency of crystalline silicon on glass modules. Proceedings of the 19th European Photovoltaic Solar Energy Conference, Paris, France, June 7–11, 2004; WIP-Munich: Munich, 2004.
25. Green, M.A. *Third Generation Photovoltaics: Ultra-High Efficiency at Low Cost*; Springer Series in Photonics; Springer: New York, 2003.
26. Cho, E.C.; Cho, Y.H.; Trupke, T.; Corkish, R.; Conibeer, G.; Green, M.A. Silicon nanostructures for all-silicon tandem solar cells. Proceedings of the 19th European Photovoltaic Solar Energy Conference, Paris, 7–11 June, 2004; Hoffman, W., Bal, J.-L., Ossenbrink, H.A., Palz, W., Helm, P. Eds.; WIP-Munich and ETA-Florence: Paris, 2004; 235–238.
27. Jiang, C.W.; Cho, E.C.; Conibeer, G.; Green, M.A. Silicon quantum dots: application for energy selective contacts to hot carrier solar cells. Proceedings of the 19th European Photovoltaic Solar Energy Conference, Paris, 7–11 June, 2004; Hoffman, W., Bal, J.-L., Ossenbrink, H.A., Helm, P., Palz, W., Eds.; WIP-Munich and ETA-Florence: Paris, 2004; 80–83.
28. Richards, B.S.; Corkish, R.P.; Green, M.A. 3C–SiC as a future photovoltaic material. Proceedings of the 3rd World Conference on Photovoltaic Energy Conversion, Osaka, 11–18 May, 2003; Kurokawa, K., Kazmerski, L.L., McNelis, B., Yamaguchi, M., Wronski, C.D., Sinke, W.C., Eds.; Osaka, 2003; paper S1P-A7-17.
29. Olson, J.M.; Friedman, D.J.; Kurtz, S. High-efficiency III–V multijunction solar cells. In *Handbook of Photovoltaic Science and Engineering*; Luque, A., Hegedus, S., Eds.; Wiley: Chichester, 2003; 359–411.
30. McCandless, B.E.; Sites, J.R. Cadmium telluride solar cells. In *Handbook of Photovoltaic Science and Engineering*; Luque, A., Hegedus, S., Eds.; Wiley: Chichester, 2003; 617–662.
31. Shafarman, W.N.; Stolt, L. $Cu(InGa)Se_2$ solar cells. In *Handbook of Photovoltaic Science and Engineering*; Luque, A., Hegedus, S., Eds.; Wiley: Chichester, 2003; 617–662.
32. Hara, K.; Arakawa, H. Dye-sensitized solar cells. In *Handbook of Photovoltaic Science and Engineering*; Luque, A., Hegedus, S., Eds.; Wiley: Chichester, 2003; 663–700.
33. Halls, J.J.M.; Friend, R.H. Organic photovoltaic devices. In *Clean Electricity from Photovoltaics*; Archer, M.D., Hill, R., Eds.; Imperial College Press: London, 2001; Vol. 1, 377–443.
34. Brabec, C.J.; Dyakonov, V.; Parisi, J.; Sariciftci, N.S., Eds. *Organic Photovoltaics. Concepts and Realization*; Springer: Berlin, 2003.

Phytoremediation

Joel G. Burken
Department of Civil, Architectural and Environmental Engineering, University of Missouri–Rolla, Rolla, Missouri, U.S.A.

INTRODUCTION

Phytoremediation is the use of plants to treat contaminated soil or groundwater. Many different plant–contaminant or plant–microbe–contaminant interactions can lead to reducing the threat posed by contaminants. To best understand how plants can reduce contaminant threat, comprehending the role of plants and their interactions with the environment overall is important. If the plant processes are recognized, then the application of plants to remedy contaminated sites is clear and logical. In this work, the ecological role of plants will be outlined, and then the ways in which plant–environment interactions can be exploited in mitigating the threat of contaminants in our environment are discussed.

PLANT BACKGROUND

In considering the potential for phytoremediation in the simplest sense, thinking ''What does a plant do?'' is very important. Plants take in water, carbon dioxide, and light to carry out photosynthesis to generate oxygen and organic carbon. Plants also carry out ''dark reactions'' utilizing molecular oxygen (O_2), organic carbon from photosynthesis, and nutrients from their surroundings to survive and thrive. On a net basis considering both photosynthesis and dark reactions, plants generate molecular oxygen and organic carbon. These two products serve as the foundation of all higher life on earth.

Overall, these reactions do not involve specific interaction with organic or inorganic contaminants other than a select few inorganics that are used as nutrients. Plants are photoautotrophs, i.e., they use light for energy and inorganic carbon (CO_2) for carbon. This is contrary to other higher life forms such as animals and microorganisms that are chemoheterotrophs, i.e., using chemical reactions (primarily with organic carbon) for energy and organic carbon for carbon.

In metabolic and physical processes, however, plants interact greatly with the surrounding environment by extracting water from the subsurface, taking inorganic nutrients from the subsurface, depositing the organic carbon both aboveground (leaf litter, fallen biomass) and belowground (root die-off and exuded organics). Through these processes plants represent the dominant interaction between the aboveground and belowground environments. The interaction with the subsurface is of greatest interest as this environment is stable for many contaminants and is therefore the target for remediation efforts.

ENVIRONMENT BACKGROUND

The subsurface environment (near surface) is relatively stable in relation to the aboveground environment. Neither does the subsurface have the energy input in the forms of sunlight, precipitation, or wind as do most surface soils, nor does it have energy input from wave, wind, or currents of surface waters. Aboveground this energy input can greatly reduce compound concentrations through mixing and the resultant dispersion and dilution, resulting in the old adage ''The solution to pollution is dilution.'' The atmosphere is also highly oxidizing, with the presence of oxygen and photolytically generated radicals. Compounds in this environment are subject to numerous chemical and biological reactions. Plants provide a connection between the two environments and have many mechanisms that can be exploited in terms of impacting subsurface contaminants, both inorganic and organic. These mechanisms are exploited to result in different phytoremediation approaches. Approaches that target contamination by metals are termed as phytoextraction and phytostabilization. For organic contaminants, approaches are termed as rhizodegradation, phytodegradation, phytovolatilization, and phytostabilization. All these approaches are covered in detail below, with examples of their uses.

METALS

Phytoextraction

Plants actively uptake selected inorganics from the environment as nutrients. Nitrogen, phosphorous, and potassium are commonly the inorganics of the greatest need, and are the basis of most fertilizers.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016194

Many others are required as micronutrients at lower levels. Elevated levels of inorganics, such as metal species, that are not needed as nutrients can act as toxicants, i.e., contaminants. Plants have adapted the ability to colonize some of the harshest environments on earth, including sites contaminated by these metals. Two types of plants generally inhabit these sites, excluders and accumulators. Excluders have adapted to be very ''tight'' with respect to uptake of the toxic metals species. These plants have highly specialized transporters in their root membranes that allow them to be efficient in the uptake of nutrient species, yet not allowing the toxic species to enter. Most plants are not as efficient and the toxic species can ''leak'' into the plant and interrupt the vital metabolic activities. Other plants have developed resistance to the toxic effects of metals following uptake as they are then dealt with to avoid toxic effects. These plants are called hyperaccumulators and are the central approach to the process of phytoextraction. The term hyperaccumulator is somewhat ambiguous. Hyperaccumulator has been consistently used to describe a plant that accumulates metals, but the term has a variety of definitions. Hyperacculator has been used for plants that accumulate greater than 1% of dry mass as a metal, concentrate metals at levels 1000 times greater than the soil concentration in which they are grown, or accumulate metals at 100 times greater concentration than nonaccumulators.[1,2] In phytoextraction, plants are used to remove the contaminant metals from the subsurface and transport the metals to aboveground tissues that can be easily harvested and treated once they are enriched. Treatments include ashing and disposing of the high-metal-content ash or reentry into a smelting process.

The initial concept of using plants to accumulate metals is found in mining. Initial efforts to look at hyperaccumulators were focused on precious metals. Attempts to locate plants that accumulate gold or silver lead to the term phytomining. Efforts to locate such plants were largely unsuccessful. From these early search efforts, however, the field of geobotany sprang forward as did numerous interesting discoveries. Perhaps, one of most spectacular documentations of hyperaccumulation is a tree from New Caledonian (*Niemeyera acuminata*) that is locally named ''sève bleue.'' The sap of the tree is blue-green because of nickel being up to 25% of its dry weight—one of the plants with the highest metal content ever discovered. The capabilities of many hyperaccumulators, over 400 taxa of terrestrial plants have been identified as hyperaccumulators, are covered in a comprehensive review by Brooks.[2]

Hyperaccumulator levels vary in the amounts of metals accumulated. For example, plants must accumulate over 10,000 mg/kg for Zn, but levels of ''hyperaccumulator status'' for Pb and Cd are less than 100 mg/kg because of lower bioavailability and the lower levels found in nonaccumulating plants. Of the many hyperaccumulators identified, only a small number are well known and have application potential in phytoremediation. Members of the *Thlaspi* family are known hyperaccumulators. Zinc, nickel, and cadmium have been found at elevated levels in *Thlaspi caerulescens* and other members of the *Thlaspi* family. In addition, members of the *Alyssum* family have been identified as hyperaccumulators. In a full-scale application in Port Colborne, Ontario, Canada, a number of *Allysum* members were evaluated and *Alyssum murale* and *Alyssum corsicum* were noted as viable candidates.[3] The full-scale application was successful, and the plants were incinerated after the growing season leading to a Ni-rich ash, enriched enough to be more valuable than the conventional crops that could have been grown on the soil. Another hyperaccumulator that has been identified and applied in full-scale applications is *Brassica juncea*, known to accumulate Zn, Pb, and Cd among other metals.

Two limiting factors in hyperaccumulators are the limited availability of the metals and the small size of many hyperaccumulators. Many times the metals at these contaminated sites have limited availability for uptake. If a metal is available, the application of hyperaccumulators without amendment is applicable, resulting in continuous uptake over the growing season. To combat low availability, chelating agents have been used to increase availability and thereby plant uptake. This has been termed ''induced phytoextraction.''[4] Chelating agents that have been applied include ethylenediamine tetra-acetic acid (EDTA) and organic acids (citric, acetic, oxalic). These efforts have not only shown an increase in uptake, but also an elevated concern that not all of the mobilized metals will be captured by the plants. Full capture is questionable in all cases, making application of mobilizing agents unfavorable. A more in-depth evaluation of phytoextraction, including a listing of full-scale applications is presented by Salt, Smith, and Raskin.[4] The future of phytoextraction includes the advancements of molecular engineering. The transport and enzyme systems are under current study, and promotion of hyperaccumulation traits in heartier, larger-biomass plants is under development.[5]

Phytostabilization

The geobotany field has also led to advanced discoveries of metal tolerant plants. These plants are not accumulators, but rather excluders. These plants are able to tolerate elevated metal contents in soils and colonize the site to stabilize possible transport. The

stabilizing mechanisms include water, wind, and soil erosion abatement and hydraulic control. Erosion control is provided by reducing kinetic energy of the water and wind, thereby reducing the capability to carry contaminated soil. Such application of plants is not novel, as vegetation has been used for millennia to stabilize soil. In the case of phytostabilization the plant selection is the novel aspect, as the metal content of such sites greatly limits the number of plants that can survive and carry out the stabilization process. Plant selection and agronomics can establish stabilizing plants within growing seasons, whereas natural colonization may not occur in contaminated sites that are distant from naturally metal-rich soils where the tolerant plants evolved and thrive. Hydraulic control is provided as plants remove water from the subsurface, reducing the potential for leaching of dissolved metals to a lower soil, groundwater profile. In addition, the plants help to stabilize soil chemistry over longer periods. Plants provide soil organic carbon, which can act to provide binding sites and increased water holding capacity. Once plants are established, ecosystems can get a toehold and remediation and restoration of a site can proceed hand in hand.

ORGANICS

Uptake

Organic contaminants have considerably different transport and fate possibilities in plant–soil systems. Organic compounds are potentially uptaken into plants with the transpiration of water from the soil profile. Once uptaken, plants can degrade certain contaminants, or contaminants can bind within the plant or volatilize from the plant. Organics can also be degraded through metabolic processes of plant-associated bacteria. Among these processes phytoremediation of organics can be classified as phytodegradation, phytovolatilization, or rhizodegradation.

Root uptake has been proven to be an important pathway for contaminants with intermediate octanol–water partitioning coefficients (K_{ow}). Variable uptake of an organic compound by different plants has been observed. Plant species such as *Daucus carota* (carrot) and *Pastinaca sativa* (parsnip) with swollen storage roots did not translocate chemicals as well as expected from barley experiments.[6] While the lipid content was considered a factor, plant structure, root types, and other properties may all play a role. The effect of the chemical itself was best illustrated by the increasing root concentration factor (RCF) and the bell-shaped transpiration stream concentration factors (TSCF) relative to $\log K_{ow}$. The physiochemical properties of compounds, including the K_{ow}, solubility, and molecular weight, all influence uptake. The dominance of compounds' K_{ow} is understandable if the entrance mechanisms of compounds to plant cells are understood on the molecular level. Two main pathways exist for chemicals to enter into the root cells, the apoplastic pathway and the symplastic pathway. To enter the symplastic path chemicals are forced to cross a cell wall to enter a cell. Chemicals can then be transported from protoplast to protoplast via plasmodesmata, which are the minute cytoplasmic threads that extend through cell walls and connect adjacent living cells. The chemicals can transport through the Casparian strip and enter tracheid or vessel cells for transport in the xylem. In the apoplastic pathway, chemicals are transported along the cell walls without passing through them. Chemicals together with water can transport apoplastically as far as the endodermis where the water impermeable Casparian strip blocks the apoplastic pathway. Compounds must cross the plasma membranes and protoplasts of the endodermal cells to gain access to the xylem tissues. In all cases, chemicals have to pass through at least two membranes. As membranes are proteins and lipid bilayers and the octanol–water partitioning coefficient is a parameter to indicate the lipophilicity, the dominance of K_{ow} can be expected even though very complicated processes were considered.[7]

The description of the entry of compounds to plant cells above also accounts for the correlations of the octanol–water partitioning coefficient with the two most commonly used factors, namely, RCF and TSCF. These relationships are based on laboratory data and are essentially empirical. Root concentration factor is defined as mass compound per mass fresh weight of roots divided by mass compound per volume ambient solution (usually in μg/mL), and TSCF is defined as mass compound per volume of water transpired divided by mass compound per volume ambient solution. Root concentration factor correlates linearly with $\log K_{ow}$ when the compound's $\log K_{ow}$ is greater than 1 and is relatively steady when $\log K_{ow}$ is smaller than 1. The linear portion of the RCF curve gives a limiting value of 0.6–1.0 for polar compounds.[6] This lower limit of RCF may be due to the transport of chemicals to the intracellular and apoplastic water in the cortex of roots via advection, dispersion, and diffusion. The second process, the sorption to the lipid, hydrophobic constituents, is likely to be the main mechanism for nonpolar compounds, leading to the linear relation to $\log K_{ow}$. Root concentration factor and TSCF are thereby interrelated as a high RCF limits the uptake of hydrophobic compounds by sorption of the compounds to the root tissues, limiting the uptake into the vascular transport system.

Compounds with intermediate $\log K_{ow,}$ ranging from 1.5 to 3.0, are translocated most efficiently.

Compounds with lower hydrophobicity appear to have resistance to pass through the lipid membrane associated with epidermal layers of the roots. Hydrophobic compounds bind to root tissues before they enter the xylem tissues as noted above.

Phytovolatilization

Following uptake, compounds have many possible fates: volatilization to the atmosphere, irreversibly binding to biomass, or plant degradation. Volatilization can occur if the compound properties allow diffusion through tissues and volatilize at the plant–atmosphere interface. Volatilization decreases the concentration in the tissues near the surface, thereby creating a concentration gradient to promote diffusion from the deeper portions.

As discussed, water enters through the roots and enters the xylem tissues. Xylem tissues are long and have a hollow, narrow, tortuous geometry. In woody plants the proportions and arrangements of these tissues vary and the anatomies are species dependent. With age, older xylem tissues become less functional. In most tree species, the outer annular rings, i.e., sapwood, dominate transport of water and nutrients, and in phytoremediation, contaminant transport as well. A good example is the viability of hollow trees. As the center of a tree rots away, it does not interrupt the tree's transport activities. Overall, xylem tissues are metabolically inactive, i.e., ''dead.'' Therefore, no metabolism of organic compounds is expected in the xylem tissues during transport. Thereby, the transport of chemicals is dominated by the chemical and hydraulic properties.

In an investigation of mass distribution and volatilization of 11 different organic compounds ranging from nonvolatile to volatile following uptake, volatilization was clearly correlated with vapor pressure.[8] Compounds with a $V_p > 0.01$ atm were volatilized, with greater V_p relating to greater volatilization. Recent findings provide new insight into the uptake of volatile organic compounds (VOCs). Investigation of tree core samples from the Savannah River site in South Carolina showed clearly that plants do uptake trichloroethylene (TCE). The work done at the Savannah River site also indicated decreasing TCE concentrations along the transpiration path (i.e., with height up the trunk).[9]

Recent research has targeted contaminated water and plant interactions. In recent work, partitioning coefficients for contaminants and plant species used in phytoremediation have been determined in the hope of better understanding the fate of compounds, particularly VOCs. Volatile organic compounds were shown to partition between plant biomass, internal water, and the vapor phase present in small vials in a laboratory study.[10] In research related to the partitioning study, diffusion from tree stems and trunks was directly measured, revealing that VOCs are taken up and can then volatilize to the atmosphere before reaching the leaves.[11] However, these studies also showed that transport to the leaves was occurring. Modeling of these mechanisms has provided an understanding of how these mechanisms interact and expectations of phytovolatilization of VOCs in remediation applications.

Phytodegradation

Plants also carry out metabolic reactions that degrade contaminants. Xenobiotics undergo detoxification processes in plants as summarized in the ''green liver'' concept.[12] This detoxification process was initially studied to consider how herbicides are degraded. Studies revealed very similar processes to mammalian liver functions and the combined processes were termed ''green liver model.'' Compounds undergo three metabolism steps, namely: transformation, conjugation, and sequestration. Xenobiotics are first transformed by ring hydroxylation of aromatics, hydrolysis, oxidation, or reduction of chemical bonds. This step often involves placing a polar, reactive group into the xenobiotics and is called ''activation'' or ''Phase I'' reaction. A major class of enzymes in the activation process is the cytochrome P-450 monooxygenases, which play key roles in a wide range of oxidative metabolic reactions in various organisms from microbes to humans. Hundreds of P-450 gene sequences have been identified, showing the important biochemical roles of P-450. Lately, complementary DNAs from three *CPR* genes from hybrid poplar (*Populus trichocarpa* × *Populus deltoides*) were isolated.[13] Many other enzymes such as peroxidases, peroxygenases, and carboxylesterases are also active in transforming organics. While oxidations comprise the majority of activation reactions, reduction processes have been identified. Of these processes, the reduction of nitro groups is of special interest as nitroaromatic compounds and explosives have been investigated extensively. The contamination by nitroaromatics and explosives is vast at former ammunition plants and ranges, and spread out over acres. In the reactions to phytodegrade these compounds, the key process involved is the conversion of a nitro group to an amino analog. Many priority pollutants can experience both oxidation and reduction process in plants. Oxidative metabolites and reductive metabolites of trinitrotoluene have been identified in cell culture studies and intact plants.[14]

In the green liver metabolism, the transformation reactions are followed by conjugation or Phase II

reaction in which sugars, amino acids/amino acid derivatives or glutathione may be transferred to the activated xenobiotics. Some compounds, however, are conjugated directly without transformation. Those compounds usually contain hydroxyl, amino, or sulphydryl moieties that are already targets for conjugation. Usually, the appearance of OH^-, SH^-, $COOH^-$, NH_2^- functional groups in transformation triggers glycosyl transfer mediated by glycosyltransferases and the presence of halogen- and nitro- functional groups triggers glutathione conjugation catalyzed by glutathione-*S*-transferases. In either case, the conjugation changes the physiochemical properties of the transformed compounds and impedes their mobility and partitioning into or diffusion through biomembranes.

The conjugation is considered to be the key step in the detoxification. Specifically, sugar conjugation has been recognized as a major detoxification reaction in plants.[12] The main process in conjugation is the glycosylation following the formation of functions, such as OH^- or NH^-, in the transformation reactions. The most common groups of sugar conjugates derive from UDP-glucose as a cosubstrate. One other consideration is that xenobiotics or their metabolites in plants are potentially bioavailable when exposed to animals. Sugar conjugates can be remobilized in the digestion process. While the sugar conjugates are generally an intermediate, they can accumulate in the cytosol or vacuole.

Another extensively encountered conjugation involves glutathione, which is highly hydrophilic. Thus, the conjugation of glutathione of hydrophobic electrophilic compounds will result in the loss of the parent compounds' lipophilicity by converting them into products possessing bulk hydrophilic regions. However, the amphiphilic metabolites have been shown to be inhibitive to other cytoplasmic enzymes, and glutathione conjugation is the dominant process in the presence of halogen- and nitro- functional groups, which may account for the low storage of TCE and metabolites in plant tissues observed in ^{14}C-labeled experiments.[15]

While the primary conjugates are more or less stable products in the cytosol, they are usually further converted to more complex or simpler products. Primary glucose conjugates are frequently converted to glucosylglucosides or 6-*O*-malonyl-glycosides, which are more stable in the vacuole.[16] Glutathione conjugates are usually metabolized within a few days to cysteinyl conjugates, believed to localize in the vacuole. However, a large part of them will leave the vacuole and reach the cytosol and be broken down there, forming xenobiotic thiol as substrate for *S*-glucosyl transferase, and methylated thiol and sulfoxidation product, both of which are volatile and inclined to escape into the atmosphere.

Plant and mammalian metabolic action showed the most difference in the third step because of the absence of the excretion system in plants. While xenobiotic metabolites after the conjugation step are readily excretable in either urine or bile in animals, they are trapped in cell walls or vacuoles in plants. Many sugar conjugates seem to be converted to different products in apoplast and sequestered into cell wall compartments with hemicellulose, cellulose, pectin, and lignin while others are stored in the plant vacuole. For most glutathione conjugates, however, the metabolism seems not to end in the vacuole. They undergo further metabolism in the cell and end up finally in the apoplast, the rhizosphere, and even the atmosphere.[17] Therefore, the vacuole may be only an intermediate storage pool for glutathione conjugates. The accumulation in the plant cell wall might be the end points of metabolism in both cases.

While the conjugates are safely trapped to the cell wall as the plant is alive, the soluble conjugates in the vacuole may become liberated quickly after the plant is dead. Research showed that polymerization of xenobiotics into lignin did not prevent the bioavailability. Sandermann[12] demonstrated that dichloroanilin in wheat straw was liberated when feeding animals. Research about the availability of residues is mainly focused on pesticides, insecticides, and drugs. A similar research on ordinary contaminants will shed light on the study of phytoremediation.

An axenic poplar cell culture experiment exhibited conclusively that poplar cells are capable of transforming and mineralizing TCE without the involvement of microbial metabolism.[15] The metabolites of TCE in cell cultures include trichloroethanol, trichloroacetic acid, and dichloroacetic acid, which was the most predominant. Chloral hydrate was also found at levels lower than the detection limit and is a product of TCE oxidation by cytochrome P-450 oxygenase and the precursor of trichloroethanol and trichloroacetic acid in mammalian systems. The same metabolites were identified in field sites where vegetation was exposed to TCE contaminated groundwater and in laboratory studies.[18,19]

Overall, the green liver metabolism of contaminants in plant cells can be substantial. Limiting factors in the metabolism are transport related. One limitation is the inability of the compound to transport to metabolically active cells, being too hydrophobic or too hydrophilic. The other limiting transport phenomenon is the volatilization prior to reaching the leaves where degradation is prevalent. All these concepts and great detail regarding the specific degradation processes and enzymes appear in a recent review chapter on the topic of plant metabolism in phytoremediation.[20]

Rhizodegradation

The rhizosphere is the area under direct influence by plant roots. The definition of the rhizosphere is vague and has different meanings for different properties. For example, if the water table is lowered and the soil profile is dewatered, the entire soil profile could be defined as ''rhizosphere.'' In the most common usage, however, rhizosphere refers to the area around the individual roots that is greatly altered by the presence of the root, perhaps the most greatly influenced in terms of biology. This alteration of the biology can lead to enhanced biodegradation by soil microbes, termed rhizodegradation.

Rhizosphere differs greatly from bulk soil, void of roots. Primary processes that lead to the changes are the removal of water and addition of organic carbon. The fact that this changes the biology greatly should be no surprise, as carbon and water are indeed the themes central to biology. When looking for life in space, the presence of organic carbon and/or water is the main requirement. Within the rhizosphere, the microbial community is generally much more active. Estimates range from 10 to 10,000 times more microbial activity in the rhizosphere. While the idea of 1000 times more microbes may seem on the extreme side, plants have been known to deposit or ''exude'' huge amounts of organic carbon. Plants have been noted to deposit as much as 50% of the photosynthetic load to the rhizosphere. In considering that plants make up over 99% of the active biomass on earth, the greatly enhanced microbial community is understandable. Once the microbial enhancement was discovered in relation to bioremediation, the development of rhizodegradation and testing was a logical step. Since that time the field of rhizodegradation has increased in activity tremendously, and many specific plant–contaminant–inoculant combinations have been investigated.

Numerous studies reveal that plants promote the degradation of certain compounds in rhizosphere. For example, biodegradation of saturated fuel oil hydrocarbons increased by 20% in the presence of maize at the early growth stage, but no obvious differences were observed between planted and bare soil when plant growth stopped after 120 days, showing the increased degradation rate rather than the extent of degradation by plants in this case.[21] Work by Siciliano and Germida investigated numerous compounds, including chlorinated aromatics, and uncovered mechanisms of plant–contaminant interactions relating to the specific organics that plants exude.[22] Plants can exude compounds with similar structure as that of contaminants, thereby enriching the microbial population for degrader organisms. Such selective enrichment was observed for mulberry trees and polynuclear aromatic hydrocarbons (PAHs).[23] Mulberry trees were shown to exude large proportions of coumarins and flavonoids, which are common and naturally occurring PAHs. Research showed that a former industrial facility had been naturally colonized by mulberry trees. Other research has shown that not only exudation, but also root turnover is vital to rhizodegradation.[24]

The rhizosphere effect does not enhance the degradation rate or the extent of chemical degradation on all occasions. In the same way that a plant can selectively promote degrader organisms if the exudates are analogs for the contaminants, plants can select nondegrader organisms. Several experiments suggest that the degradation of petroleum hydrocarbons in soil may be unaffected by the rhizosphere. Boyle and Shann observed no difference regarding the mineralization of phenol and 2,4-dichlorophenol in rhizosphere and nonrhizosphere soils. In a study of diesel fuel degradation in rhizosphere soil, diesel range organic (DRO) compounds were decomposed most rapidly in soils with NPK fertilizer, regardless of plant cover.[25,26] In unvegetated pots amended with NPK fertilizer, 92.5% of DROs were decomposed after 50 days, and all DROs were removed at 150 days. No significant difference was noted between DRO decomposition in unvegetated soil treated with NPK compared to the soil cropped to legumes, which demonstrated faster removal of DROs than grasses. The addition of readily degradable carbon to the soil was found to stimulate microbial degradation of xenobiotics. However, considering the costs of operation and maintenance, phytoremediation still holds advantages. In addition, the addition of nutrients has to be maneuvered carefully. In the same study, NPK fertilizer treatment showed the highest removal rate of DRO, with urea being the second highest. The glucose treatment under mixed grasses resulted in lower DRO removal than in amended soil, indicating that the addition of nutrients may be a more effective way to enhance the degradation ability of indigenous microorganisms than the addition of degradable carbons.[26]

Inoculating rhizosphere soil with contaminant-degrading bacteria is another way to improve rhizodegradation. Combining a TNT-transforming bacterium and a tolerant plant was observed to decrease the concentrations of TNT in soil by 30% within a 5-week period, while the plant effect was not visible and its growth was reduced in the absence of the bacteria inoculant.[27] Meadow brome (*Bromus biebersainii*) did not stimulate degradation of 2-CBA in soil when 2-CBA was present at 200 mg/kg; however, extractable 2-CBA was reduced significantly when inoculated with a mixture of two pseudomonads.[28] The inoculation of bacteria improves the rhizodegradation ability by two

main mechanisms: increasing the numbers in the microorganism community as mentioned above or increasing the capacity of rhizosphere microorganisms to degrade contaminants. Inoculating Dahurian wild rye with a mixture of strains R75 and CB35 reduced 2-CBA contaminant by the increase of the degradation capacity of rhizosphere, but had no effect on total heterotrophic populations or on plant growth variables, indicating that inoculating did not increase the rhizosphere effect, but instead the degradation capacity of rhizosphere microorganisms.

Inoculation does not always affect the rhizosphere degradation capacity of plants favorably even though the rhizosphere community is enlarged. Researchers found that bacteria inoculation protects plants from the phytotoxic effects of pentachlorophenol and other compounds. However, protecting plants from contaminant phytotoxicity with bacteria inoculants does not always lead to contaminant degradation.[29] Sometimes, the inoculation process increased contaminant phytotoxicity and resulted in other adverse effects. Inoculating bacteria into certain plants led to the death of plants. Noninoculated Altai wild rye survived in 3-CBA contaminated soil but died when inoculated with degrader strains.[28] Plant exudates were demonstrated to impact the bacteria physiology and bacteria inoculants are known to alter plant physiology. Thus, plants might change bacteria in such a way that they produce phytotoxins or phytotoxic metabolites or, alternatively, bacteria may interfere with certain plant detoxification processes.[27]

Overall rhizosphere impacts help to improve soil properties and also stimulate the organic input to a subsurface environment and act as a water sink. These processes are generally helpful for soil properties and health. There are many other details and examples of rhizodegradation than are covered in this brief summary. A recently published book provides a detailed look at rhizodegradation.[30]

ANCILLARY BENEFITS

Phytoremediation offers many advantages over more accepted and older remediation technologies like pump and treat or excavation. One advantage is public acceptance. The general public accepts and embraces trees as a sign of a good environment. Adding plants to a contaminated site not only targets the remediation in a scientific, engineered approach, but also provides confidence that the site is safe to support vegetation. Phytoremediation also offers an ecological benefit. While many technologies void the site of all life and even all organic carbon, such as thermal heating, stream injection, advanced oxidation, phytoremediation fosters additional life, working to remediate and restore the contaminated site concurrently. This aspect is unique to phytoremediation. Energy input is also not needed once the system is established. The solar powered nature of phytoremediation makes it a prime candidate for remote and vast sites, where energy input is difficult to impossible or else is prohibitively costly.

CONCLUSIONS

Phytoremediation is promising and is proving to be a useful approach to remediating contaminated soils and groundwater. The multiple mechanisms involved in phytoremediation offer many different approaches, including stabilization, phytovolatilization, rhizodegradation, phytoaccumulation, and phytodegradation. These approaches allow designers the ability to target different contaminants in different media (soil or groundwater) by understanding the approach selected and the mechanisms that are utilized. Phytoremediation also offers numerous advantages that make it an attractive option in remediating contaminated sites.

REFERENCES

1. Baker, A.J.M.; Brooks, R.R. Terrestrial higher plants which hyperaccumulate metalic elements—a review of their distribution, ecology and phytochemistry. Biorecovery **1989**, *1*, 81–126.
2. Brooks, R.R. *Plants That Hyperaccumulate Heavy Metals: Their Role in Phytoremediation, Microbiology, Archaeology Mineral Exploration and Phytomining*; CAB International: Wallingford, U.K., 1998.
3. Kukier, U.; Peters, C.; Chaney, R.; Angle, J.S.; Roseberg, R. The effect of pH on metal accumulation in two alyssum species. J. Environ. Qual. **2004**, *33*, 2090–2102.
4. Salt, D.; Smith, R.; Raskin, I. Phytoremediation. Annu. Rev. Plant Physiol. Plant Mol. Biol. **1998**, *49*, 643–668.
5. Chaney, R.L.; Malik, M.; Li, Y.M.; Brown, S.L.; Brewer, E.P.; Angle, J.S.; Baker, A.J.M. Phytoremediation of soil metals. Curr. Opin. Biotechnol. **1997**, *8* (3), 279–284.
6. Briggs, G.G.; Bromilow, R.H.; Evans, A.A. Relationships between lipophilicity and root uptake and translocation of non-ionized chemicals by barley. Pestic. Sci. **1982**, *13*, 495–504.
7. Trapp, S. Modeling uptake into roots and subsequent translocation of neutral and ionisable organic compounds. Pest Manag. Sci. **2000**, *56*, 767–778.
8. Burken, J.G.; Schnoor, J.L. Distribution and volatilization of organic contaminants following

uptake by hybrid poplar trees. Int. J. Phytoremediation **1999**, *1* (2), 139–152.
9. Vroblesky, D.A.; Neitch, C.T.; Morris, J.T. Chlorinated ethenes from groundwater in tree trunks. Environ. Sci. Technol. **1999**, *33* (1), 510–515.
10. Ma, X.; Burken, J.G. VOCs fate and partitioning in vegetation: use of tree cores in groundwater analysis. Environ. Sci. Technol. **2002**, *36* (21), 4663–4668.
11. Ma, X.; Burken, J.G. Diffusion of TCE to the atmosphere in phytoremediation applications. Environ. Sci. Technol. **2003**, *37* (11), 2534–2539.
12. Sandermann, H. Plant metabolism of xenobiotics. Trends Biochem. Sci. **1992**, *17*, 82–84.
13. Ro, D.; Ehlting, J.; Douglas, C.J. Cloning, functional expression, and subcellular localization of multiple NADph-cytochrome P450 reductases from hybrid poplar. Plant Physiol. **2002**, *130*, 1837–1851.
14. Burken, J.G.; Shanks, J.V.; Thompson, P.L. Phytoremediation and plant metabolism of explosives and nitroaromatic compounds. In *Biodegradation of Nitroaromatic Compounds and Explosives*; Spain, J.C., Hughes, J.B., Knackmuss, H.J., Eds.; CRC Press: Boca Raton, FL, 2000; 239–275.
15. Newman, L.A.; Strand, S.E.; Choe, N.; Duffy, J.; Ekuan, G.; Ruszaj, M.; Shurtleff, B.; Wilmoth, J.; Heilman, P.E.; Gordon, M.P. Uptake and transformation of TCE by hybrid poplars. Environ. Sci. Technol. **1997**, *31* (4), 1062–1067.
16. Lamoureux, G.L.; Rusness, D.G.; Schroder, P. Metabolism of a diphenylether herbicide to a volatile thioanisole and polar sulfonic acid metabolites in spruce. Pestic. Biochem. Physiol **1993**, *47*, 8–20.
17. Schroder, P.; Harvey, P.J.; Schwitzguebel, J.P. Prospects for the phytoremediation of organic pollutants in Europe. Environ. Sci. Pollut. Res. **2002**, *9* (1), 1–3.
18. Doucette, W.; Bugbee, B.; Hayhurst, S.; Plaehn, W.; Downey, D.; Taffinder, S.; Edwards, R. Phytoremediation of dissolved-phase trichloroethylene using mature vegetation. The First International Conference on Remediation of Chlorinated and Recalcitrant Compounds, Battelle Press: Monterey, CA, 1998.
19. Schnabel, W.E.; Dietz, A.C.; Burken, J.G.; Schnoor, J.L.; Alvarez, P.J. Uptake and transformation of trhichloroethylene by edible garden plants. Water Res. **1997**, *31* (4), 816–824.
20. Burken, J.G. Uptake and metabolism of organic compounds: green-liver model. In *Phytoremediation: Degradation and Control of Contaminants*; McCutcheon, S.C., Schnoor, J.L., Eds.; John Wiley and Sons Inc.: Hoboken, NJ, 2003; 59–84.
21. Chaineau, C.H.; Morel, J.L.; Oudet, J. Biodegradation of fuel oil hydrocarbons in the rhizosphere of maize. J. Environ. Qual. **2000**, *29*, 569–578.
22. Siciliano, S.; Germida, J. Degradation of chlorinated benzoic acid mixtured by plant–bacteria associations. Environ. Toxicol. Chem. **1998**, *17* (4), 728–733.
23. Olson, P.; Fletcher, J. Ecological recovery of vegetation at a former industrial sludge basin and its implications to phytoremediation. Environ. Sci. Pollut. Res. **2000**, *7* (4), 195–204.
24. Leigh, M.; Fletcher, J.; Fu, X.; Schmitz, F. Root turnover: an important source of microbial substrates in rhizosphere remediation of recalcitrant contaminants. Environ. Sci. Technol. **2002**, *36* (7), 1579–1583.
25. Boyle, J.J.; Shann, J.R. Biodegradation of phenol, 2,4-DCP, 2,4-D, and 2,4,5-T in field-collected rhizosphere and nonrhizosphere soils. J. Environ. Qual. **1995**, *24* (4), 782–785.
26. Pichtel, J.; Liskanen, P. Degradation of diesel fuel in rhizosphere soil. Environ. Eng. Sci. **2001**, *18* (3), 145–157.
27. Siciliano, S.D.; Greer, C.W. Plant-bacterial combinations to phytoremediate soil contaminated with high concentrations of 2,4,6-trinitrotoluene. J. Environ. Qual. **2000**, *29*, 311–316.
28. Siciliano, S.; Germida, J. Mechanisms of phytoremediation: biochemical and ecological interactions between plants and bacteria. Environ. Rev. **1998**, *6*, 65–79.
29. Pfender, W.F. Bioremediation bacteria to protect plants in pentachlorophenol-contaminated soil. J. Environ. Qual. **1996**, *25* (6), 1256–1260.
30. Olson, P.; Reardon, K.; Pilon-Smits, E.A. Ecology of rhizosphere bioremediation. In *Phytoremediation: Degradation and Control of Contaminants*; McCutcheon, S.C., Schnoor, J.L., Eds.; John Wiley and Sons Inc.: Hoboken, NJ, 2003; 317–354.

Pilot Plant and Minipilot Units

Richard P. Palluzi
ExxonMobil Research and Engineering Company, Annandale, New Jersey, U.S.A.

INTRODUCTION

Pilot plants have a distinguished and long history in supporting research in chemical process industries. Called by a variety of names—miniunits, research units, pilot units, microunits, miniplants, etc.—they all have the same basic purpose: To allow research or data gathering about actual performance of a process under real world conditions faster and cheaper than is possible in a commercial facility. A pilot plant is a tool for investigating a process or a process related problem, on a manageable scale, in a realistic and timely manner. Its size (or scale) is intended to be just large enough for all the important factors to be evaluated, yet as small and simple and economical as possible. A pilot plant is a means to an end rather than an end itself.

TYPES OF PILOT PLANTS

Pilot plants can be classified in a number of ways with the most common being by size, purpose, or degree of automation. The different classifications are summarized in Tables 1–3. Classifications are merely indications of general properties of the units; most are so unique that each one is truly ''one of a kind.''

Laboratory units typically fit on a bench top or inside a laboratory hood and are typically piped in tubing (see Figs. 1 and 2). While traditionally they are manually operated and continuously attended, most are now computer automated and often operate unattended. Integrated pilot plants range in size from several pallets or frames to a small building, and are usually piped in tubing or small pipe (see Figs. 3 and 4). Most are computer controlled and designed for unattended operation. Semiwork units are the largest type of pilot plant, and often are at the lower end of a plant scale (see Fig. 5). They are usually piped in commercial-size pipe and tend to resemble commercial units in automation and operation. Classifying pilot plants by size is useful for their relative costs and scales, but a more useful measure of a pilot plant's value is its purpose.

Process simulation pilot plants usually replicate all the major process steps in the actual order likely to exist in the plant. The pilot-plant equipment will be similar, although smaller. Because virtually all of a process is simulated, there is less chance of a critical step being missed. Any part of a process can be investigated in this type of unit; even those not identified in advance. Data are usually easier to correlate and safer to scale up than for other types of pilot plants. It is generally easier to design (assuming an existing plant exists) and useful for the life of the process. However, a process simulation pilot plant is usually more complex, expensive to build, and labor intensive to operate and support.

Problem solving pilot plants tend to be narrower in scope and limited to only the area of interest. They are usually cheaper and faster to construct. Data can often be more readily developed through a customized and focused design. Many problems are also amenable to investigation through nonprocess simulation methods, which can further reduce the cost and/or complexity. However, problem solving pilot plants often cannot be easily adapted for other purposes limiting their long-term utility. Their focused approach requires careful problem identification and does not lend itself to investigating synergistic effects or unknown causes.

Basic research pilot plants are designed for acquiring fundamental knowledge about a process; their design is set by the data required. Their size varies from small bench top units to larger integrated facilities and their complexity is usually high. A basic research pilot plant has a very narrow but deep focus. It is generally very complex and expensive.

PILOT PLANT SELECTION

Selecting the right type of pilot plant requires a clear definition of the program's goals. An optimum selection process is shown in Fig. 6. (See also Refs.[1–3].)

PILOT PLANT LOCATION

Where a pilot plant is located can have major and lasting effects on both its initial construction cost and its annual operating cost. There are four basic types of locations:

- Separate buildings.
- Containment (blast) cells.
- Open bays.
- Laboratories.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007898

Table 1 Classification by size

Class	Characteristics	Cost range
Laboratory pilot plant Microunit Laboratory unit Miniunit Pilot unit	• Fits on a bench or in a laboratory hood • Usually occupies 2 m^2 or less • Small diameter tubing, hose, or pipe used (2–12 mm OD) • May be manual operated or highly automated • Often has significant associated off-line equipment	$10,000 to $250,000
Bench top unit Pilot plant Miniplant Integrated pilot plant	• Usually constructed in a laboratory, open bay, containment cell, or large walk in hood • Typically 4–25 m^2 • Moderate diameter tubing or piping used (6–50 mm) • Usually automated • May include remote feed and product systems	$100,000 to $1,000,000
Demonstration unit Prototype unit Semiworks Large scale pilot plant	• Usually in an exterior location or a dedicated building • Typically larger than 50 m^2, often much larger • Larger size tubing and pipe used (25–100 mm) • Usually highly automated • May include laboratory and office space as part of overall unit facility	$1,000,000 and above

Separate, dedicated buildings eliminate the possibility that a pilot plant's operations will interfere with another's. It allows each building to be optimized for the pilot plant's specific purpose. Safety concerns still exist as separate buildings, and often force operating personnel to work alone, out of sight of other personnel, which can raise concerns regarding their safety and the timeliness of a response to an accident. Constructing a building each time a new pilot plant is needed is also expensive and time consuming as is service and maintenance. The trend toward smaller units has made separate buildings uneconomical.

Isolating pilot plants by means of containment cells can be via steel barricades or reinforced-concrete bunkers. Typically, the walls or roofs are designed to blow out and vent the force of an explosion to a safe area. Total-containment cells, where the entire force of the release is confined in the cell, are much more expensive and complex. Containment cells are regarded as the safest possible location when all operations are

Table 2 Classification by purpose

Classification	Process characteristics	Other characteristics
Process simulation	• Mimics process operations step by step although not all steps may be included • Uses scaled down flow rates and operating conditions • Often uses smaller scale commercial equipment	• Usually larger than bench scale • Typically fairly complex
Problem solving	• Designed to address a specific process problem or issue • May mimic a portion of the process but often designed to investigate problem in a different manner • Flow rates and operating conditions usually mimic process but may allow operation at extremes for testing	• May be of any size or complexity • Usually more focused than a process simulation
Basic research	• Laboratory style operation with no relation to process except in chemistry • Specifically designed to allow data accumulation under conditions of interest • Process conditions mimicked but flow rates usually minimal	• Size varies greatly but usually larger than bench scale • Substantial analytical equipment usually involved • Usually constructed in a laboratory environment

Table 3 Classification by extent of automation

Classification	Process characteristics	Unit characteristics
Manual	All process conditions set and monitored manually by operators	• Poor repeatability • Loose process control • High operating labor costs • Manual data gathering
Local control	Most variables monitored and/or controlled by dedicated controllers	• Better repeatability • Better process control • Lower operating labor costs • Data gathering may be manual, automatic, or a mix
Automated	Computer controlled usually through a personal computer or programmable logic controller. Less frequently by a central control computer or as part of a distributed control system	• Better repeatability • Better process control • Lower operating labor costs • Data gathering automatic although some analytical instrumentation may remain off line • Some to all routine operations may be automated
Fully automated	Computer controlled usually through a personal computer (PC), programmable logic controller (PLC), central control computer or distributed control system (DCS)	• Best repeatability • Best process control • Lowest operating labor costs • Data gathering automatic including most or all analytical instrumentation • Most routine operations automated

Older systems, which used a mixture of manual operations, local controllers and computers, are now very rare. They represented a common step in the progress of automating pilot plants but, in general, have given way to either simpler units, which remain manual, or use local controllers and more complex systems, which are automated.

remotely controlled from outside the cell. Greater risks can be accepted as equipment, not personnel, are at risk. However, containment cells are very expensive ranging from \$100,000/m^2 to \$1,000,000/m^2 depending on the degree of containment and the layout. Cell design is complex, difficult to validate, and easily voided through improper unit placement or design. Estimating the blast potential of process failures is also difficult, as it requires one to predict the specific failure mode and conditions. Hence, the safety of the cell may be less than assumed. Pilot plant installation is costly, primarily because of the extensive automation required for remote control. Operating efficiency is limited by the restricted access; some routine entry is usually accepted even with attendant increased hazard.

An open bay is a large area having few, or no, vertical partitions that permit laying out a pilot plant in almost any configuration. Open bays are the most preferred pilot-plant layout because control and interlock systems are reliable enough to reduce the risks of a major accident to acceptable levels, alleviating the need for containment cells in most cases. Open bays are the least expensive areas to construct, as an open warehouse or a prefabricated metal building can serve as the basic structure. The openness enhances safety because operators are rarely alone or even out of sight. However, interactions between pilot plants need to be evaluated carefully. A small fire on a unit handling noncombustible materials can change from a minor hazard to a major accident if the unit is located next to one handling large quantities of flammable materials.

Laboratory areas, once primarily the home of only wet chemistry operations, are now increasingly housing permanent pilot plants. This is due largely because of reductions in the pilot plant size due to improved instrumentation and analytical techniques as well as by advances in scale up allowing larger leaps between piloting processes. Many pilot plants now easily fit inside a standard or an expanded laboratory hood. The high rate of ventilation in the hood substantially reduces the hazards of dealing with hazardous materials. The climate-controlled environment of most laboratories not only enhances operator comfort and productivity but allows the use of numerous specialty equipment that cannot withstand the rigors of an un-air conditioned open bay. Laboratory buildings are, however, very expensive to construct and even more costly to operate due to their need for once-through exhaust. While newer technologies (heat recovery systems, reduced flow hoods, outside make up air, etc.) hold out the hope of reducing these operating costs, their application to date has been mostly on a trial basis. Relative costs for construction and operating different types of pilot plant space arrangements are listed in Table 4.

Fig. 1 A laboratory scale pilot installed in a knee height hood. *(View this art in color at www.dekker.com.)*

Safe unit placement enhances the overall safety of a pilot plant. A containment cell or a separate building is not, however, a guarantee of safety. Other measures include: Increased process monitoring of safety specific variables, limiting inventories of hazardous feedstocks and products, detailed hazard analysis and risk assessments, and good design practices. One essential element is to ensure that the pilot plant has adequate space to support its operation. This requires a careful evaluation of the required space before the unit is constructed. While a detailed layout and a careful evaluation of the operating requirements remains the preferred method of determining the space needed, other estimating techniques are available.[4] The space required for storing and handling feedstocks and products frequently exceeds that for the pilot plant. Typical concerns requiring additional space include:

- Storage capacity is needed for process upsets.
- Recycled material storage may be needed.
- Storage due to the lead time for ordering feedstock and the hold time for disposing of product.

Fig. 2 A small reactor unit located in a bench top hood. *(View this art in color at www.dekker.com.)*

Fig. 3 An integrated pilot plant installed in an open bay. *(View this art in color at www.dekker.com.)*

- Space for supporting chemicals such as cleaning and flushing solvents.
- Room for storing drums, cans, and other containers.
- Maintenance space.
- Spare parts and off-line equipment space.
- Associated laboratory space.

DESIGN

The unique nature of designing a pilot plant arises from the smaller size and the lower level of design information available. A pilot plant is usually the first engineered version of a laboratory process. In many cases, the full range of operating parameters would not have been fully defined before design starts. This makes detailed design more difficult and costly as either a wider range of conditions must be accommodated, or costly and time-consuming modifications may be required as the actual ranges are established after testing. In addition, industry design specific rules are often either not available for a new process or their scale is too large for the proposed smaller pilot plant or microunit. Both are problems unlikely to be encountered in the design of a commercial unit.

Pilot plant design typically follows one of two approaches: Mimic the methodology used for the

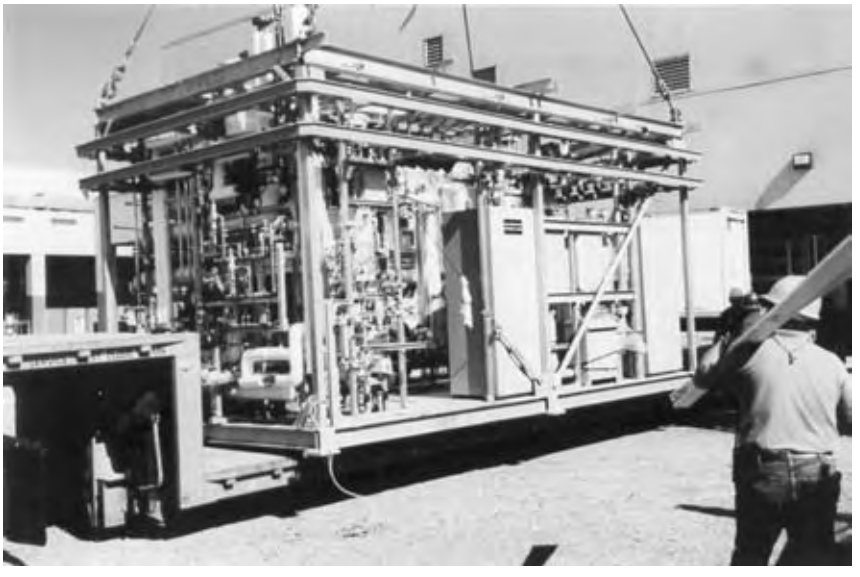


Fig. 4 A modular pilot plant in shipment from the construction shop to the field for installation. *(View this art in color at www.dekker.com.)*

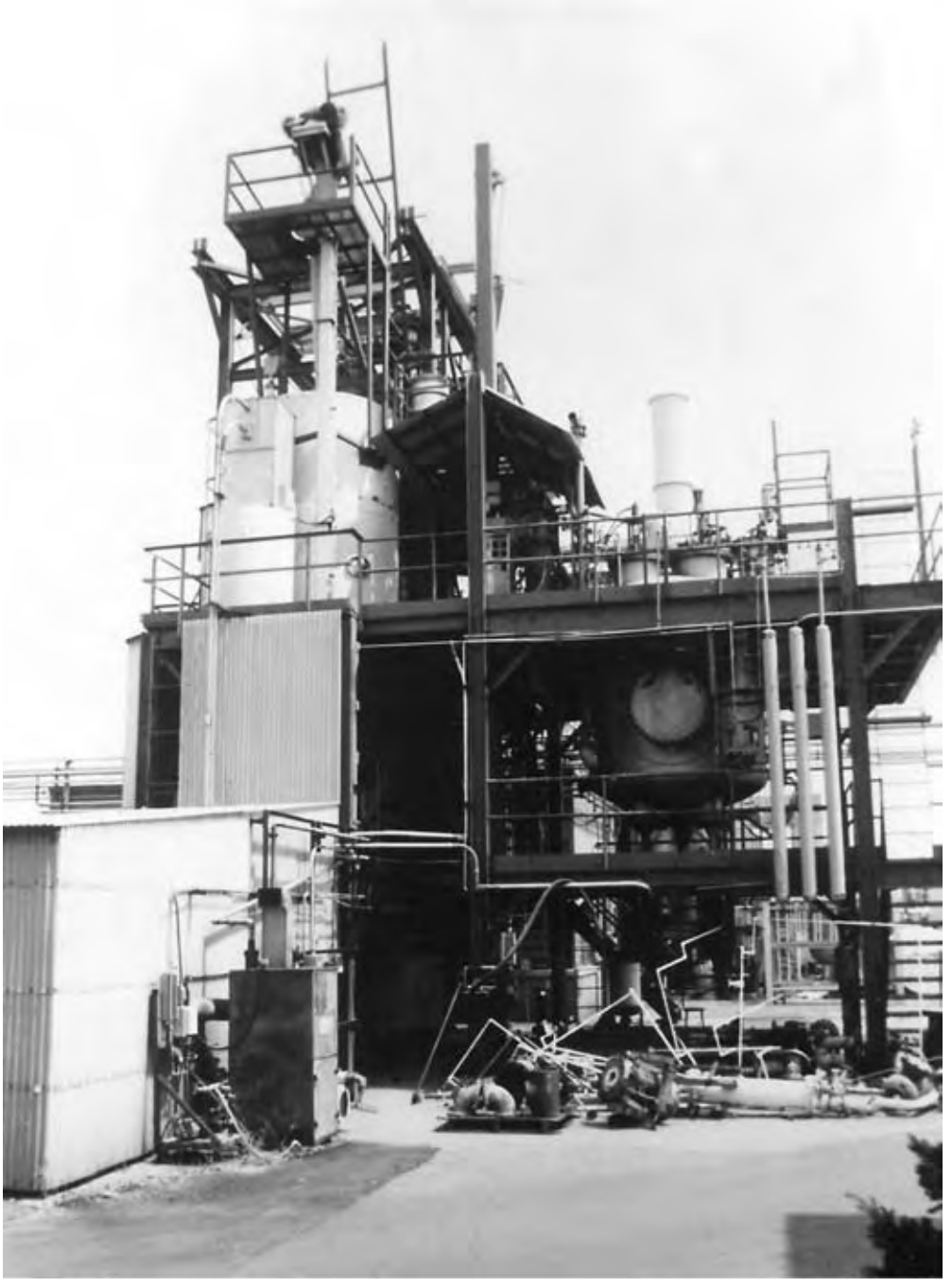


Fig. 5 A large semiworks pilot plant installed outside. *(View this art in color at www.dekker.com.)*

commercial plants or tailor the design specifically for the pilot plant. Either approach can work but the former often results in a larger and more complex design, while the latter can maximize the advantages of the pilot plant albeit at a higher level of risk in the design.

When mimicking the commercial design approach, the process steps, the basics of the process set, and the information for designing the equipment are known. Unfortunately, this approach is not always feasible particularly with new processes or novel technology. The proposed pilot scale may be outside the commercial design methodology because of the pilot plant's smaller size, different operating conditions, or novel process steps. Some full-scale operations are too expensive to reproduce on the pilot plant scale—e.g., solid handling and complex recycle systems. The commercial plant may utilize existing facilities, such as waste treating units and emission control equipment that cannot be duplicated economically in the pilot plant.

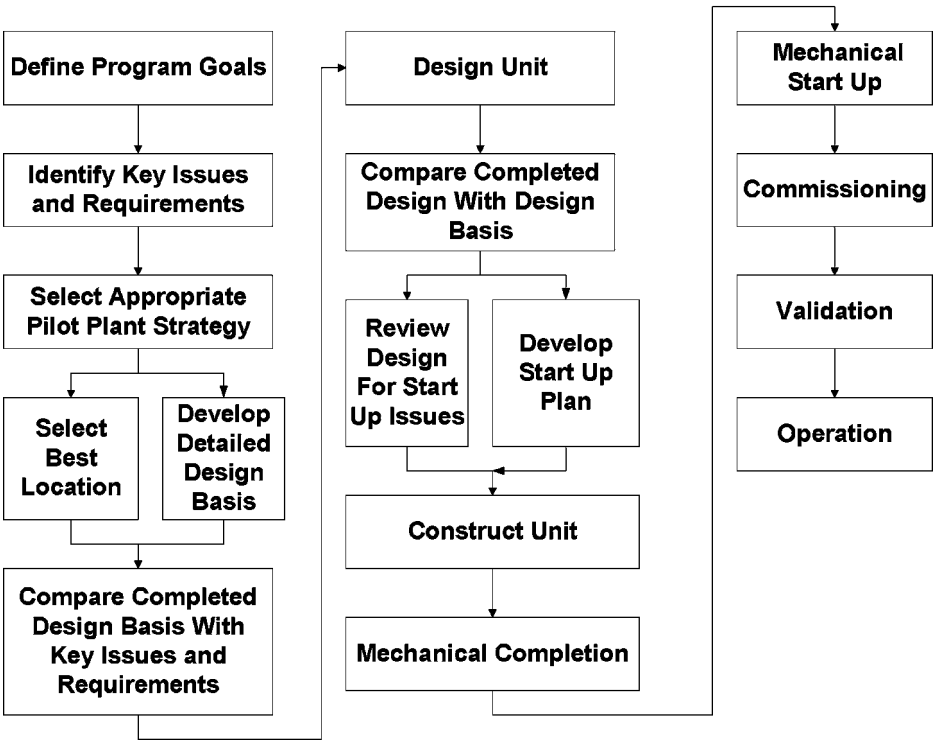


Fig. 6 A typical pilot plant selection process. *(View this art in color at www.dekker.com.)*

For a pilot plant specific design, conventional design techniques will be used whenever they can be matched to the scale and process as will commercial-scale design information. The equipment will be sized to try to maximize the advantages of the pilot plant, using the smaller scale to minimize or eliminate problem areas. This generally lowers the construction and operating costs. There are some problems with this approach as little information has been published on pilot plant design because most is considered proprietary. This makes the effectiveness of the design much more dependent on the skill level and the experience of the design engineer than in a commercial plant and increases the potential for design related problems arising later. It is still a much-pursued option as the potential construction and operating savings usually outweigh the risk. (See also Refs.[4–7].)

There are several options for those designing and building a pilot plant as summarized in Table 5. Each has advantages and disadvantages as shown. (See also Refs.[8–12].)

Two additional options exist in designing a pilot plant. The first is to rent space on an existing unit. Universities, specialty contractors, equipment manufacturers, and toll houses often have some level of existing pilot plans for rent for a test or a series of tests. This can minimize the lead time, as there is no design, construction, or start up that can give the organization access to a proven operation and reduce overall costs. Unfortunately, this option is not always viable. A suitable unit may not be

Table 4 Relative costs of constructing and operating different pilot plant space

Type of pilot plant space	Construction	Maintenance
Separate buildings	1.25–2.00	1.50–5.00
Steel cells	1.25–1.50	1.00–1.10
Barricaded concrete cells	3.00–10.00	1.10–1.25
Open bays[a]	1.00	1.00
Laboratory areas[b]	2.00–3.00	2.00–5.00

These figures are relative space costs for comparable structures in similar areas. They can vary widely across the country.
[a]All costs are relative to open bays, which for all practical purposes can be approximated by local costs to construct warehouse type space.
[b]Usually air conditioned unlike the other spaces.

Table 5 Options for design and construction

Design	Construction	Advantages	Disadvantages
In house	In house	• Most straightforward and simple • Information transfer minimized • Proprietary information protected • Reduced start up time • Improved in-house expertise developed • Reduced modification costs • Usually lowest risk option	• Requires adequate in-house resources not always available to all organizations • Requires skilled resources, design capability, construction staff, adequate resource time/availability • Often not a viable option for small organizations
In house	Contracted	• Retains most significant element affecting pilot plant success in house • Information transfer minimized • Proprietary information protected • Improved in-house expertise developed • Reduces demand on internal resources • Easier to locate a qualified construction contractor than a design contractor • Often faster to complete	• Final construction product is dependent on the skill of the contractor • Equipment and materials selection may be less than desirable unless specified • Locating qualified construction contractor is a tedious and time-consuming task • Construction supervision/follow up remains an issue • Higher cost (profit and overhead) • Some in-house follow up still required
Contracted	In house	• Reduces the need for in-house design resources • Compensates for lack of in-house design experience and/or expertise • Contractor design may be more efficient • Lump sum costs may be a viable option	• Rarely as effective as any of the other options • Locating a qualified design contractor • Lack of contractor expertise in specific area • Potential loss of proprietary information • Information transfer • Decreased operating efficiency • Higher start up costs • Higher maintenance costs • Higher cost (profit and overhead) Meeting industry and company • safety standards • Some in-house follow up still required
Contracted	Contracted	• Viable option for any size organization • Reduces the need for in-house resources to a minimum • Compensates for lack of in-house experience and/or expertise • May improve cost control • Contractor may be more efficient • Lump sum costs may be a viable option	• Locating a qualified contractor (both design and construction) • Lack of contractor expertise in specific area • Potential loss of proprietary information • Information transfer • Decreased operating efficiency • Higher start up costs • Higher maintenance costs • Higher cost (profit and overhead) • Meeting industry and company safety standards • Some in-house follow up still required

available, leading to inappropriate program compromises, expensive and/or time-consuming rework, and/or increased risk of failure. Even when a unit exists, it may not be available due to other commitments. Performing research outside the organization always raises some incremental risk of loss of proprietary information. There are also increased costs due to physical separation (travel and inefficiencies because of personnel absence), which is particularly a problem for extended programs. Safety can also be an issue if the facility owner and the organization do not share the same philosophy or concerns. Typically there is a potential disconnect as the unit owner knows the equipment safety issues but the client knows the materials concerns. Hence, an

Table 6 Types of safety reviews

Type	Purpose	Timing	Major issues
Conceptual or preliminary	• Identify potential problem areas early in design (''show stoppers'') • Ensure major safety concerns are addressed at the beginning of design • Identify problems without resolution	As early as possible in the project	• Location • Process operating conditions • Material hazards • Reactivity • Toxicity • Legal and code compliance
Preconstruction, design, or prebuild	• Review details of the proposed pilot plant • Approve final design for construction • Detailed safety review • Safety analysis or similar hazard assessment	Just before construction and materials procurement	• Detailed piping and instrumentation review • Risk assessment • Hazard analysis • Interfaces/interferences with other operations
Preoperational or prestart-up	• Review proposed operations • Examine the final construction product • Evaluate hitherto hidden concerns, interferences, or impacts • Review proposed operating procedures	Either immediately before or after start up[a]	• Operating procedures • Operating concerns • Construction deficiencies
Unit modification, control of change, or change of service	• Review the safety aspects of changes in equipment, procedures, or materials • Typically combine preconstruction and preoperational features • Prevent unauthorized, inappropriate, or poorly done modifications as they are the major cause of accidents	Whenever a change is envisioned	• Requires clear definition of what is a ''change'' • Information on existing systems and equipment must be up to date
Decommissioning, out of service, or end of service	• Ensure a unit is left in a safe condition when taken out of service • Avoid difficulties of disposing of equipmentcontaminated with unknown substances later	When a pilot plant is to be out of service for an extended period (typically 6 mo or more)	• De-energizing of equipment • Disposal of all hazardous materials • Cleaning all equipment • Identifying any long term concerns • Appropriate mothballing

[a]Before start up ensures no hazardous materials are introduced before a safety review and no operations occur before the construction product is inspected. After start up ensures start-up changes will be reviewed and accepts that actual operations are often too ill defined until this point for adequate review and many problems only appear during start-up.

independent client evaluation of all associated safety issues is critical. It is also important to reach resolution of some areas in advance such as who will handle and fund unit maintenance, who will set support priorities, who will operate the unit and ensure the accuracy, calibration, and certification of all key process and safety equipment.

Contracting the entire program, rather than just renting the equipment, reduces the in-house effort to a minimum and may result in reduced overall costs. However, it is difficult in this case to fully protect proprietary information as well as to capitalize on any fringe discoveries. The quality of the results is also highly dependent on the contractor's expertise. The distance from data can hide promising new areas and/or disguise existing problems. This approach is most viable where in-house expertise is lacking but is readily available outside. (See also Ref.[13].)

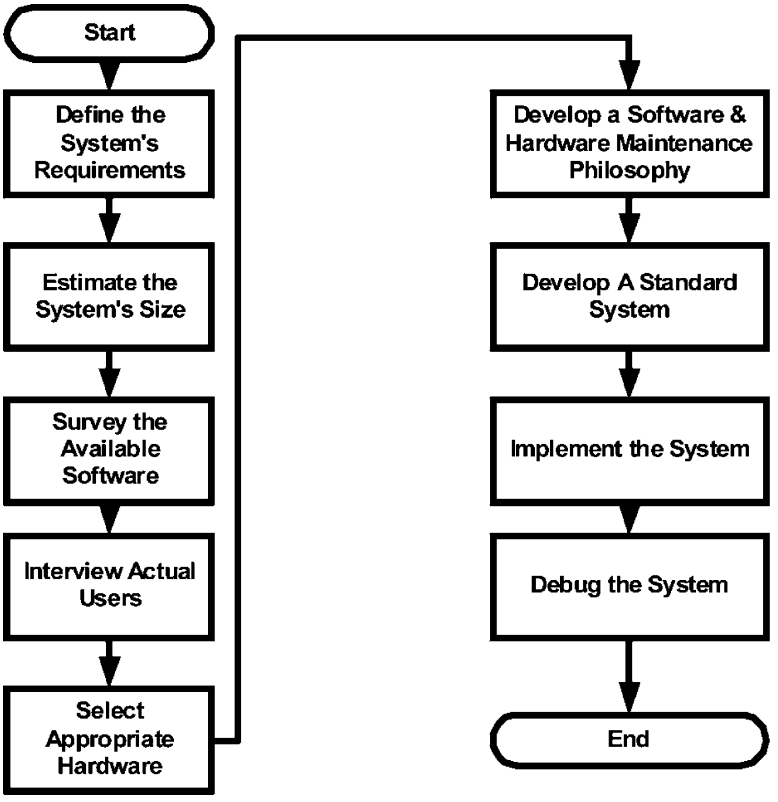


Fig. 7 Selecting a pilot plant control system. *(View this art in color at www.dekker.com.)*

SAFETY

Typically pilot plants undergo several safety reviews over the course of their conception. While names vary among organizations, the most common are shown in Table 6. (For a more detailed discussion of the numerous pilot plant and laboratory specific safety concerns, see Refs.[14–20].)

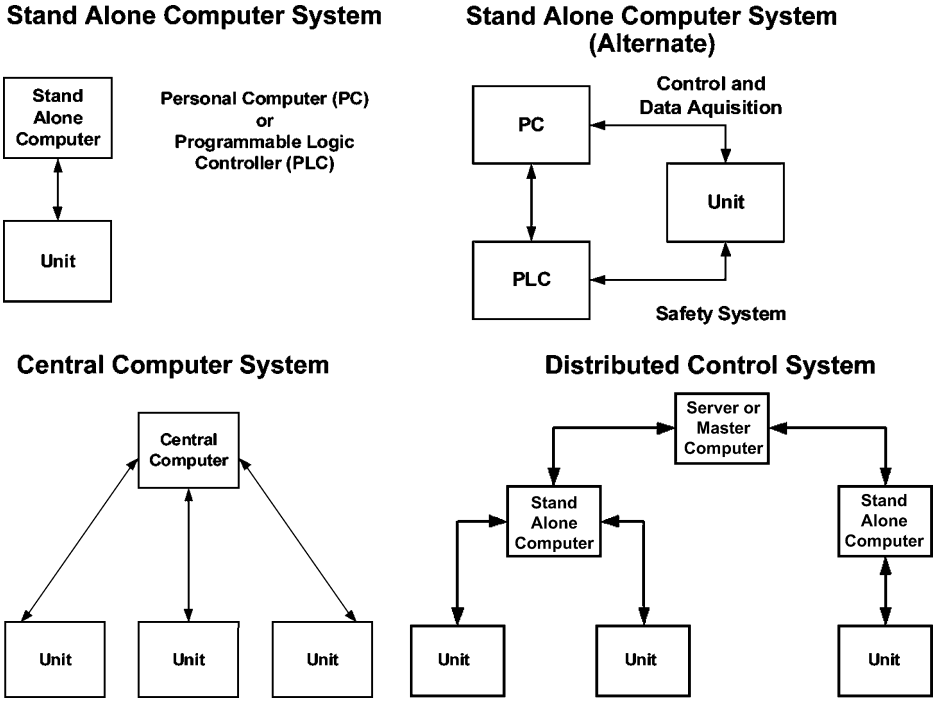


Fig. 8 Types of computer control systems. *(View this art in color at www.dekker.com.)*

Table 7 Computer control system analysis

Advantages	Disadvantages
Stand-alone computer system	
• Lowest cost per single unit • Easy and economical expansion • Separates pilot plants completely	• Standardization can become an issue • Each system's capacity can become a limiting factor • Many common components need to be reproduced • Marginal needs can be rarely satisfied • Control packages are frequently limited in size, scope, and capability • Operator access and screen displays may be limited, cumbersome, or less useful than desired • Data gathering and storage capabilities may be limited • Programming may require significant training and routine experience
Centralized computer system	
• Can be more economical for a large number of pilot plants • Usually provided with adequate programming support and maintenance services • Marginal needs are often justifiable across larger population • Usually practical to oversize all capabilities • Data gathering and control packages are more fully developed and useful • Data storage is rarely an issue	• Any failure affects everybody • Initial installation costs are much larger • Annual operating costs are very visible • Maintenance is difficult to schedule
Distributed control systems (DCS)	
• Has all the advantages of both a stand-alone and a central computer system	• Significantly more expensive to install, maintain, and support

CONTROL AND AUTOMATION

The typical process for selecting a pilot plant or a micro-unit control system is outlined in Fig. 7. This is often more difficult than immediately apparent because pilot plants are generally not operated long enough or enough are not constructed in a single location to permit side by side evaluations. The low construction rate in most organizations also leads to each new installation being treated as a unique entity. Hence most pilot plant control systems are selected from process plant experiences or literature searches rather than actual operating experience.

A key issue in selecting a pilot plant control system is to determine the degree of automation. Manual control is rarely suitable for any modern pilot plant due to the high cost of operating labor, the difficulty in ensuring consistent and quality results because of operator error and variability, and the difficulties in collecting data accurately and analyzing them later. In general, manual systems should be considered only if the control requirements are very small (one loop and less than five data points), the operation is a problem, which cannot be automated reliably or no support is available. Even then it will usually turn out to be a costly long-term choice. Automatic control systems are commonly selected because they reduce operating labor costs, increase accuracy through tighter process control, provide more rapid response and less operator bias, improve product quality, and reduce the potential for human error, which can improve both safety and operability. They also have the potential for automation with attendant savings. Automatic control systems do have some concerns. They lower operator attention, have a higher initial cost and a higher maintenance cost, and may increase support skill and staff levels. These are usually grossly outweighed by their advantages.

Automatic control systems can be based on individual stand-alone controllers or computer based systems. Individual controllers are easier to specify, readily available, and require less design and little or no programming. However, system integration is very difficult and data gathering remains a separate problem. System installation and maintenance are expensive and overall system

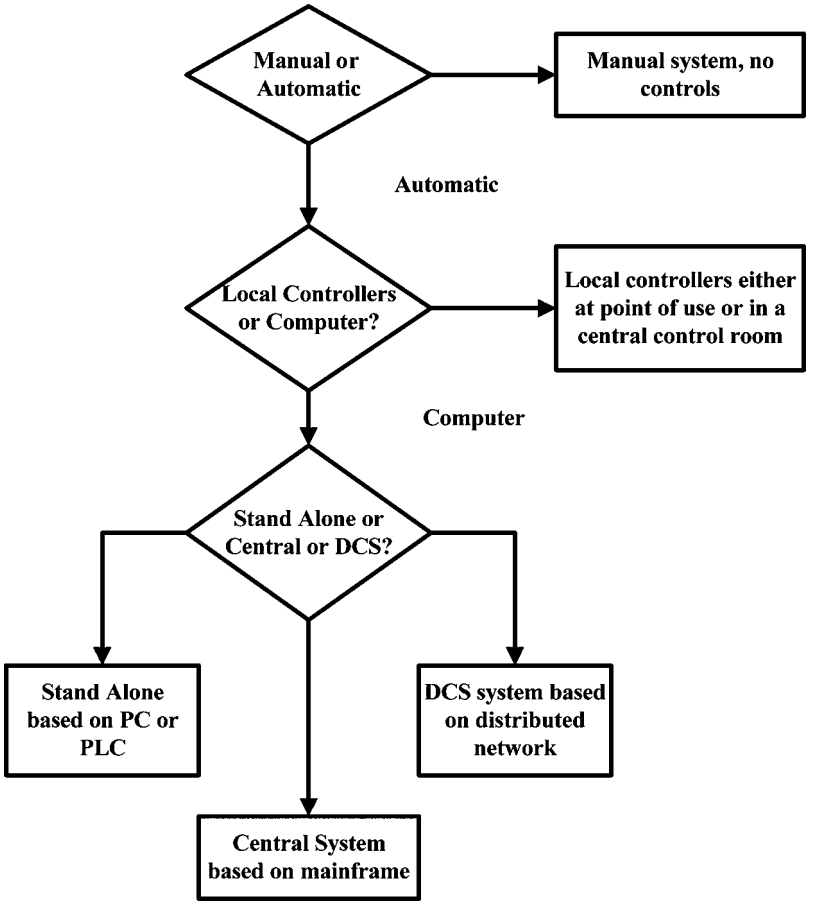


Fig. 9 Selecting the type of control system. *(View this art in color at www.dekker.com.)*

reliability is usually less than for a computer based system. Computer control systems are flexible, easier to specify and install, and simpler to modify. They inherently handle data gathering and provide the potential for easy automation. They are usually more expensive than local controllers but often not by much when the installation costs are included. Their hardware maintenance costs are less but they require programming, which can be expensive and require specialized support. Most new pilot plant control systems are computer based.

The different types of computer control systems are shown in Fig. 8, while their advantages and disadvantages listed in Table 7. Most new systems are stand-alone systems based on personal computers except for larger research organizations, which may be able to justify the more expensive distributed control systems (see Fig. 9 for a roadmap to selecting a pilot plant control system).

The first step in selecting a computer based control system is to determine the system requirements in terms of input and output signals. In most cases, this effort will have to start before the final process design is complete. Table 8 lists the common input and output requirements for typical system components.

The chosen system should have sufficient capability for the largest need envisioned. Room for expansion should be included. A typical computerized pilot plant grows by 20–50% per year during the first two years and by 10–30% thereafter. Select a control and data acquisition package first as it is the key element in obtaining a satisfactory system, then the computer hardware on which to run it. Most software control and data gathering packages can run on a wide variety of hardware platforms.

A pilot plant can be computer controlled in one of two ways: supervisory or direct digital control as shown in Fig. 10. In the first, a computer supervises another primary control element. In the second, the interfacing element is eliminated and the controller (e.g., a valve) is regulated directly. New computer-based automatic instrumentation is likely to be supervisory because it is faster and less expensive. (See also Refs.[21–23].)

COSTS

Estimating a pilot plant or a microunit's construction cost can be done in several different ways as summarized in Table 9. Pilot-plant cost estimating differs from that for full-size plants because little pilot-plant cost information has been published, the personnel

Table 8 Typical system requirements

	Inputs			Outputs	
Item	**TC**	**Analog**	**Digital**	**Analog**	**Digital**
Temperature control loops	1			1	
Flow control loops		1		1	
Pressure control loops		1		1	
Level control loops		1		1	
Other control loops		1		1	
Temperature switches			1		1
Flow switches			1		1
Pressure switches			1		1
Level switches			1		1
Other switches			1		1
Block valves					1
Motors			1		1
Analyzers		?	?		1
TC	1				
Instrument		1			
Temperature alarms	1				1
Flow alarms		1			1
Pressure alarms		1			1
Level alarms		1			1
Gas detection alarms			1		1
Fire/smoke alarms			1		1
Other alarms		?	?		1

["?" indicates either or both may be required.]

involved often less experienced in cost estimating, and every pilot plant is unique.

Similarity involves estimating the cost of a pilot plant from the cost to design and construct a similar unit. This method produces excellent results if the units are almost identical. Differences due to inflation can be accounted for reasonably through a published cost index, such as the *Chemical Engineering Plant Cost Index*. Unfortunately, few pilot plants are similar enough in all aspects to ensure an accurate estimate. Unless the cost base is large and the estimator experienced, these estimates are rarely more accurate than ± 50%. An estimate of such low accuracy is only useful for a preliminary screening.

In a cost-ratio estimate, the cost of an entire pilot plant (or a part of it) is related to a factor, such as

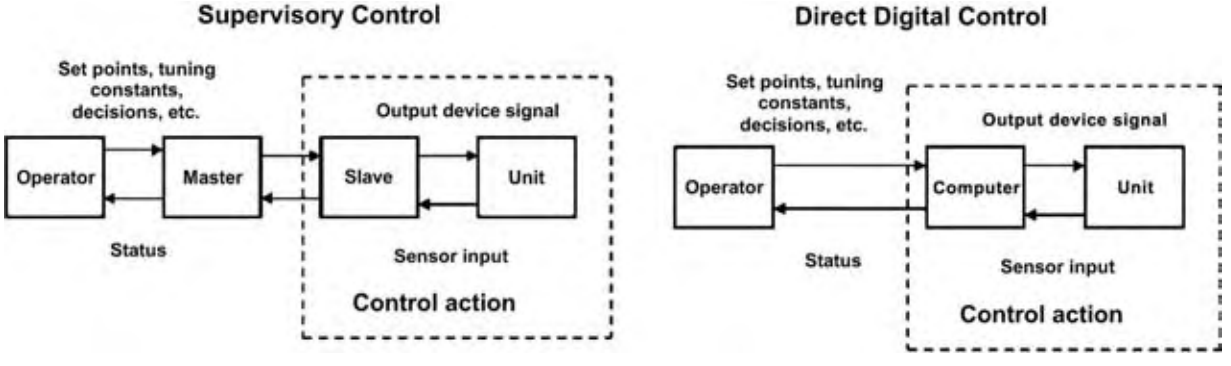


Fig. 10 Types of computer control. *(View this art in color at www.dekker.com.)*

Table 9 Pilot plant cost estimating methods

Factor	Similarity	Cost ratios	Details of time and materials required
Level of project definition	Low	Low–Medium	High
Site sensitivity	High	Low	Medium
Organization sensitivity	High	Low	Medium
Level of estimator experience	High	Medium	Medium–high
Cost accounting method sensitivity	Low	Medium	High
Ability to reconcile capabilities vs. costs	Low	Medium	High
Sensitivity to unknown factors	High	High	Medium
Sensitivity to single problem areas	High	Medium	Low
Information required	Low	Low–medium	Medium-high
Typical time required to develop (in weeks)	≤ 1	1–2	4–6
Typical cost of cost estimate as percentage of total project	≤ 1	1–3	3–5

Table 10 Typical pilot plant start up plan

Task description	Responsibility	Date
Arrange preliminary start-up planning meeting to develop start-up sequence	Project team/S/U team	
Arrange mechanical and technical support of start up	S/U team	
Field verification		
Update all piping and electrical prints to reflect all field changes	Project team	
Trace all lines and locate all valves, correcting the P&ID or unit as required	Operations team/Project team	
Tag all valves up to main utility shut-offs	Operations team	
Flushing		
Confirm all compression fittings are tight using a gage block as appropriate	PF	
Pressurize unit with nitrogen to 200 kpa and perform a low-pressure gross leak test. Correct any leaks found	PF/Operations team	
Prepare unit for flushing	PF/Operations team	
Disconnect all sensitive equipment	PF/Operations team	
Install jumpers around all disconnected equipment	PF/Operations team	
Install temporary lines to and from flushing cart or other solvent source	PF/Operations team	
Flush gas feeds	PF/Operations team	
Flush liquid feeds	PF/Operations team	
Flush gas product systems	PF/Operations team	
Flush liquid product systems	PF/Operations team	
Flush sampling systems	PF/Operations team	
Flush miscellaneous systems	PF/Operations team	
Dry all flushed lines with dry nitrogen for a minimum of ____ hours	PF/Operations team	
Reconnect all items disconnected for flushing	PF/Operations team	
Clean or replace all filters	PF/Operations team	
Calibrate all instrumentation	Operations team/IS	
Pressure transducers	Operations team/IS	
Mass flow meters	Operations team/IS	
Pressure switches	Operations team/IS	

(Continued)

Table 10 Typical pilot plant start up plan *(Continued)*

Task description	Responsibility	Date
Flow switches	Operations team/IS	
I/Ps	Operations team/IS	
P/I's	Operations team/IS	
Liquid level transducers	Operations team/IS	
Temperature switches	Operations team/IS	
Others	Operations team/IS	
Zero and span power controllers (SCRs)	IS/Operations team, Project team	
Check all thermocouples for polarity and position	IS/Operations team, Project team	
Verify all analog inputs to control system or computer	IS/Operations team, Project team	
Verify all analog outputs from control system or computer	IS/Operations team, Project team	
Verify all digital inputs from alarm system or computer	IS/Operations team, Project team	
Verify all digital outputs from alarm system or computer	IS/Operations team, Project team	
Leak testing		
Verify all fittings are at least hand tight	PF/Operations team	
Leak test gas feed systems	PF/Operations team	
Leak test liquid feed systems	PF/Operations team	
Leak test gas product systems	PF/Operations team	
Leak test liquid product systems	PF/Operations team	
Leak test sampling systems	PF/Operations team	
Leak test miscellaneous systems	PF/Operations team	
Leak test utility systems	PF/Operations team	
Leak test vent systems	PF/Operations team	
Leak test all heated components at operating temperatures	PF/Operations team	
Rotating equipment		
Verify pump rotation and operation	Operations team	
Produce calibration curve	Operations team	
Verify compressor rotation and operation	Operations team	
Verify compression capacity	Operations team	
Verify safety shut down system	Operations team	
Calibrate wet test meters	Operations team	
Verify purges to purge enclosures	Operations team	
Verify heater current draw	Operations team	
Verify ventilation to all enclosures and local vents	Operations team	
Seal all conduit seals	IS/Operations team	
Rough tune all control loops		
Commission unit in sections		
Feed system	Operations team	
Reactor system	Operations team	
Product separation system	Operations team	
Product collection system	Operations team	
Recycle system	Operations team	
Analytical system	Operations team	
Others	Operations team	
Conduct a dummy run under simulated conditions specified in advance	Operations team/S/U team/Project team	
Conduct a preoperational safety review	Operations team/Project team	
Update all prints to reflect as completed unit and file	Operations team/S/U team	

S/U Team, Start-up Team; Project Team, Project Construction Team; Operations Team, Operations Team; PF, Pipefitter Support; IS, Instrument Specialist Support.

Table 11 Typical maintenance times

Equipment maintenance	Typical annual work-hours (hr)
Instrumentation	
Controllers	0.5–2
Alarms	≤ 0.5
Manual stations	≤ 0.5
Annunciators	1.2
Indicators	0.5–1
Data loggers	2–4
Personal computers	4–12
Programmable logic controllers	8–24
Sensors and transmitters	
Differential pressure	≤ 0.5
Pressure	0.5–1
Thermal flow meters	8–12
Coriolis flow meters	12–16
Volumetric flow meters	≤ 0.5
Positive-displacement meters	0.5–1
Capacitance-resistance probes	8–24
Rotating equipment	
Plunger pumps	8–16
Centrifugal pumps	4–8
Gear pumps	4–8
Diaphragm pumps	8–12
Reciprocating compressors	8–16
Diaphragm compressors	2–12
Air actuators	0.5 or less
Analyzers	
Gas chromatographs	24–84
Liquid chromatographs	16–48
Mass specs	84–160
Density analyzers	8–16
Infrared analyzers	4–12
Moisture analyzers	24–84
Miscellaneous	
Control valves	2–24
Regulators	≤ 0.5
Relief valves	2–4

Note: These figures are averages over several years for reliable equipment in mild to moderate service. Older equipment, severe service, aggressive ambient conditions, and similar factors can raise these estimates significantly. These figures are based on trained and experienced personnel. New or untrained personnel typically require two to three times the maintenance-work time shown.

the cost of a major item of equipment or the number of control loops. Pilot-plant costs are rarely estimated this way because cost-ratio data for small equipment are not readily available, and most are developed by scaling down the known installed cost for the smallest item of equipment for a full-size plant. This frequently results in a highly inaccurate estimate because the cost of equipment below a certain size is often fixed. The accuracy of cost estimates based on cost ratios ranges from $\pm 25\%$ to $\pm 50\%$, depending on the nature of the ratio and the experience of the estimator.

Developing a detailed time and materials estimate requires dividing construction into a series of small tasks and estimating the labor and the materials required for each task. The narrower the scope, the higher the accuracy of the estimate. The accuracy ranges from $\pm 10\%$ to $\pm 25\%$. This is a tedious and time-consuming process but does produce the most accurate results and can serve as the framework for budgeting and controlling costs. (See also Refs.[24–27].) Operating, start up, and maintenance costs can also be estimated by a variety of methods.[1]

START UP

A pilot-plant startup typically takes longer than expected from scaling down the effort for a commercial plant because of the novelty of the process and the limited resources assigned.

The key to a successful startup is planning and organization beginning at the start of the project. The proposed design should be reviewed for start up issues such as how instrumentation will be calibrated and lines leak tested. This effort should then progress to making a detailed start up plan listing each task that must be performed in some detail and then estimating the resources and the effort required. This plan can then be used to develop a realistic start up schedule and resource list. It acts as the master plan for the entire start up and can be modified as the start up proceeds. Table 10 gives a typical generalized start up plan. An actual plan would be much more detailed. (For more details on pilot plant start up see Refs.[1,28,29].)

MAINTENANCE

Maintenance in a pilot plant often seems too insignificant or transitory to warrant organized attention. The reality is quite different. Few pilot plants operate for so short a time to require no maintenance. Annual maintenance costs are usually best expressed as a percentage of construction costs and vary widely due to differences in accounting practices and maintenance policies. In general, pilot plant maintenance costs can be estimated based on the technology and run length:

Conventional, 1–5 day runs	5–15%
Conventional, >5 day runs	10–20%
Severe service, 1–5 day runs	10–25%
Severe service, >5 day runs	20–40%

(See also Table 11.)

CONCLUSIONS

Pilot plant and microunit design and operation is a specialized area where information is often lacking. The trend is toward smaller, more heavily instrumented units that can be modified for a larger variety of programs.

REFERENCES

1. Palluzi, R.P. Choosing the right pilot plant. Chem. Eng. Prog. **1991**, 62.
2. Lowenstein, J.G. The pilot plant. Chem. Eng. **1985**, *62*.
3. Hockenhull, D.J.D. *Organization of a Pilot Plant for the Development of New Products;* Consultant to Glaxo Laboratories Ltd., 1991.
4. Palluzi, R.P. *Estimating Space Requirements in Pilot Plant Design, Construction and Operation*; McGraw-Hill: New York, 1992.
5. Davidson, R.S. Stop redesigning pilot plant systems. Chem. Eng. Prog. **1986**, *18*.
6. Randhava, R.; Lo, R.N. Modular pilot-plant technology. CEP **1982**, *76*.
7. Ginkel, S.L.; Olander, D.L. Stick to the basics of pilot plant design. Chem. Eng. Prog. **1991**, *27*.
8. Palluzi, R.P. Succeed at crash pilot plant construction. Chem. Eng. Prog. **1997**, 45.
9. Basu, P.K.; Quaadgras, J. Pharmaceutical pilot plants are different. Chem. Eng. Prog. **1997**, *66*.
10. Phillips, S.G. Skid-mounted pilot-plant facilities. Chem. Eng. Prog. **1990**, *15*.
11. Richardson, J.A. Using modular pilot plants to speed development of new processes. AIChE Meeting, May 1984.
12. Barkey, W.R.; Palluzi, R.P. Cost effective pilot plant design and construction. Chem. Eng. **2000**, 92.
13. Margolin, S.V. Why use outside R&D. Chemtech **1984**, *90*.
14. Palluzi, R.P. *Pilot Plant and Laboratory Safety*; McGraw-Hill: New York, 1994.
15. Palluzi, R.P. Develop and R&D safety standards program. Plant Safety. Chem. Eng. Prog. **1995**, 58.
16. Carr, J.W., Jr. Taking the guess work out of pilot plant safety. Chem. Eng. Prog. **1988**, 52.
17. Dore, J.C. Process safety—an integral part of pilot plants. Plant/Operat. Prog. **1988**, *223*.
18. Lesins, V.; Moritz, J.J. Develop realistic safety procedure for pilot plants. Chem. Eng. Prog. **1991**, *39*.
19. West, H.H.; Mannan, M.S. Make plants safer with a proper management of change program. Chem. Eng. Prog. **1998**, *25*.
20. Noren, A.R.; Naik, N. Keep pharmaceutical pilot plants safe. Chem. Eng. Prog. **1999**, *39*.
21. McClosky, R.J. Choosing the right process control system for a general purpose pilot plant facility. AIChE Meeting, Nov 1986.
22. Uitenham, L.; Mumjal, R. Choose the right control scheme for pilot plants. Chem. Eng. Prog. **1991**, *35*.
23. Spielmann, P.C.; Basu, P.K. Make the most of pilot plants DCSs. Chem. Eng. Prog. **1996**, *58*.
24. Uppal, K.B.; Van Gool, H. R&D phase—capital cost estimating. AACE Transact. **1992**, *7*, 1
25. DesJardin, M.; Collins, D.; Dietsche, L. *Early Stage Capital Estimates for Pilot Plants and Plants,* Dow Chemical Company.
26. Cressman, K.R. Pilot plant process costing. Cost Eng. **1983**, *25* (1), 31.
27. Yoslov, S. Cost estimating techniques for pilot plants. AACE Transact. **1992**, *7.1*.
28. Palluzi, R.P. Flush away equipment problems. Chem. Eng. **1989**, 81.
29. Palluzi, R.P. Leak testing. Chem. Eng. **1988**.

Pinch Design and Analysis

Robin Smith
Jin-Kuk Kim
Centre for Process Integration, University of Manchester, Manchester, U.K.

INTRODUCTION

Pinch analysis first emerged as an energy saving approach to the design of heat exchanger networks. The approach is soundly based in thermodynamics, but also allows economic criteria to be integrated with thermodynamics to provide practical and cost-effective solutions to energy saving problems. Since the early developments of the 1970s, pinch analysis has become established as the standard tool for the analysis and design of heat recovery systems. Many thousands of practical applications have been achieved.

Issues in the success of pinch analysis and design are:

- Pinch analysis exploits targeting before design. This allows objectives to be screened prior to detailed design. The ability to set scientifically based targets prior to design has been a key element in the success of the methodology.
- The methodology uses thermodynamic analysis and representational aids, but leaves the designer in full control of the decision making process.

The discussion here provides an introduction to the methods, but more detailed treatments are available.

DESIGN OF HEAT RECOVERY SYSTEMS

Fig. 1 shows a simple flowsheet. Some of the streams require heating (sinks for heat, or cold streams), whereas others require cooling (sources of heat, or hot streams). It is desirable to recover heat where possible from the streams requiring cooling to provide the heat for those requiring heating. Even for a simple system like the one shown in Fig. 1, the best arrangement for this heat recovery is not obvious. Streams requiring cooling need to be matched against those requiring heating, but taking account of the temperature differences available in the process. The second law of thermodynamics dictates that heat can only flow from a high-temperature source into a low-temperature sink. Moreover, there is usually a minimum acceptable temperature difference for heat transfer. Table 1 presents the details of the data from the simple example of Fig. 1. Each stream has an initial, or supply, temperature and a final, or target, temperature. In addition, there is an enthalpy change defined by the physical properties of the stream. Note that the enthalpy in Fig. 2A is a total enthalpy (i.e., the product of specific enthalpy and mass flow rate). Also, the heat capacity is a total heat capacity (i.e., the product of specific heat capacity and mass flow rate). In the simple examples in this article, it is assumed that the streams all have a constant heat capacity. With a small problem, such as the one in Fig. 1 and Table 1, trial and error might be used to try and match the hot streams and cold streams. Even for a small problem, such an approach is not guaranteed to give the best solution. A more systematic approach is required.

ENERGY TARGETS

The material and energy balance defines the heat sources (hot streams) and heat sinks (cold streams). To set energy targets, the streams are first plotted in temperature and enthalpy coordinates. Fig. 2A shows the two hot streams (sources of heat). The straight temperature–enthalpy profiles in Fig. 2A imply constant specific heat capacity. The positioning of the streams in the temperature–enthalpy plot is arbitrary on the enthalpy scale. This is because the reference enthalpy can be chosen to be any convenient value for each stream. The important specification for the stream is that the enthalpy change is fixed. Fig. 2B shows the two hot streams combined to give a composite hot curve. Where streams have a common temperature range, the enthalpies and heat capacity flow rates are combined. The hot composite curve, as shown in Fig. 2B, is the single-stream equivalent of the two separate streams in Fig. 2A in terms of temperature and enthalpy.

The corresponding cold streams (sinks for heat) are shown plotted in temperature–enthalpy coordinates in Fig. 3A. Fig. 3B shows the curves combined to give a composite cold curve. Again, where streams have a common temperature range, the enthalpies and heat capacity flow rates are combined. The cold composite curve, as shown in Fig. 3B, is the single-stream equivalent of the two streams in Fig. 3A in terms of temperature and enthalpy.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007899

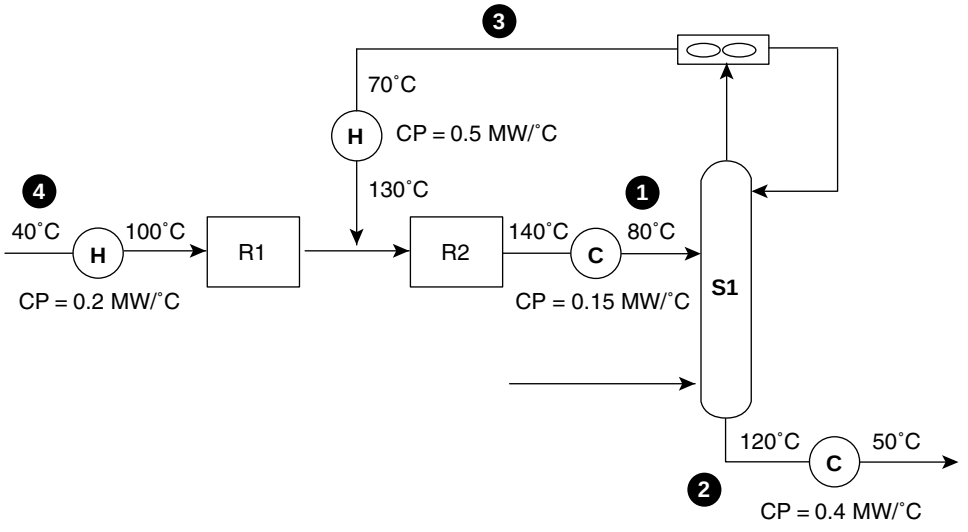


Fig. 1 A simple system with two hot streams and two cold streams.

The two curves can now be matched to determine how much heat can be recovered from the hot streams in the process to the cold streams in the process. The composite curves are shown in Fig. 4 matched such that there is a minimum temperature difference (ΔT_{min}) of 10°C. The hot composite curve must be above the cold composite curve by at least ΔT_{min} at all points for feasible heat transfer. The relative position of the two curves has been adjusted such that they are separated by a specified minimum temperature difference. In this case, $\Delta T_{min} = 10°C$. The overlap between the composite curves represents the heat recovery potential between the hot and cold streams in the process, shown in Fig. 4. The monotonic nature of the construction of the composite curves allows for maximum overlap between the curves. This in turn allows the construction to determine the maximum heat recovery potential (Q_{REC}). By maximizing the heat recovery, the residual demand for heating and cooling utilities is minimized. In Fig. 4, the part of the cold composite curve that projects beyond the start of the hot composite curve represents the external heating utilities required ($Q_{H_{min}}$). The part of the hot composite curve that projects beyond the start of the cold composite curve represents the minimum cold utility required ($Q_{C_{min}}$). The construction of the composite curves provides an energy target for the process and allows the best energy performance of the process to be determined prior to heat recovery network design.

In the case of the composite curves in Fig. 4, ΔT_{min} was chosen to be 10°C. What is the correct value of ΔT_{min}? In principle, the correct value of ΔT_{min} is determined by a capital–energy tradeoff, as illustrated in Fig. 5. The composite curves can be shifted horizontally relative to each other by adjusting the reference enthalpy of the curves. This allows ΔT_{min} to be varied. Small values of ΔT_{min} require small amounts of external heating and cooling utility. As ΔT_{min} increases, the consumption of both hot and cold utility increases to maintain the energy balance. The energy performance of the process therefore becomes poorer the larger the value of ΔT_{min}. On the other hand, the temperature differences for heat transfer between the hot and cold streams throughout the process become larger as ΔT_{min} increases. This means that the heat transfer area decreases, hence lowering the capital costs. Thus, as shown in Fig. 5, there is a capital–energy tradeoff. Combining the energy cost and capital cost allows an optimal value of ΔT_{min} to be identified. In practice, the shape of the optimization

Table 1 Stream data for the flow sheet in Fig. 1

Stream	Type	Supply temperature, T_S (°C)	Target temperature, T_T (°C)	ΔH (MW)	Heat capacity flow rate, CP (MW C^{-1})
1	Hot	140	80	–9.0	0.15
2	Hot	120	50	–28.0	0.4
3	Cold	70	130	30.0	0.5
4	Cold	40	100	12.0	0.2

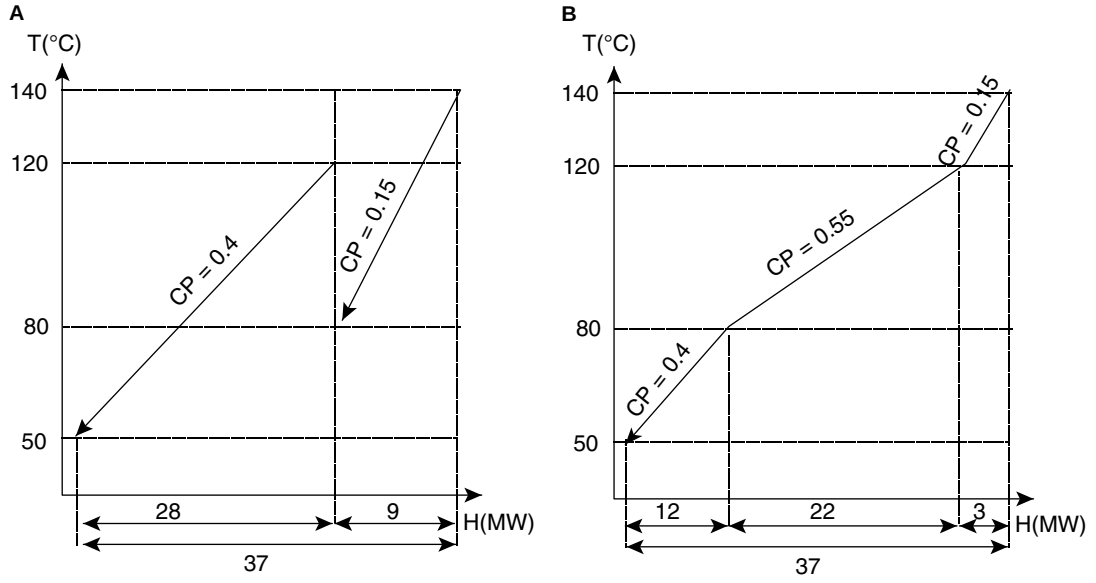


Fig. 2 Hot composite curve. (A) Individual temperature–enthalpy profiles for hot streams. (B) Combination of profiles for two hot streams.

curve tends to be quite flat, and such optimization is often not necessary. As long as the value of ΔT_{min} is somewhere in the region of the optimum point, there is little to be lost or gained in terms of the overall cost by small changes in ΔT_{min}. For chemical processes, a reasonable value of ΔT_{min} is usually around 10°C. A larger value is usually required for refinery processes (typically 20–30°C). For low-temperature processes, a reasonable value of ΔT_{min} is often lower (typically 5°C or less) because of the expense of providing cooling through refrigeration in low-temperature processes. Practical issues also restrict the choice of ΔT_{min}. For example, if shell-and-tube heat exchangers are to be used, then 10°C is, in any case, a practical minimum for equipment design. Although shell-and-tube heat exchangers can be designed with temperature differences below 10°C, special attention is required in their design, as the flow is not truly countercurrent in such devices. In contrast,

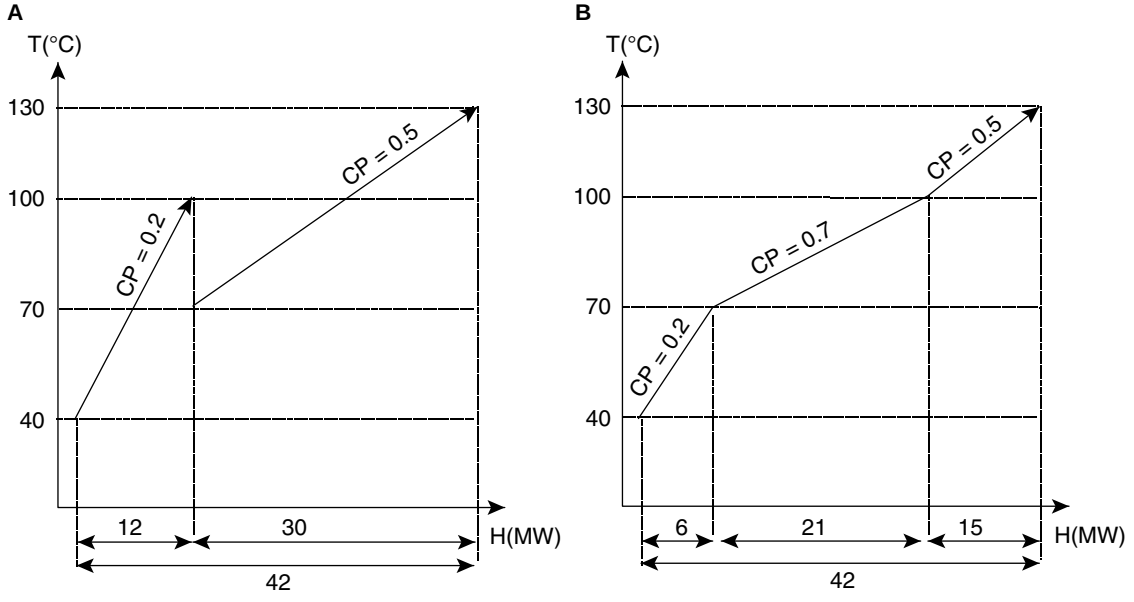


Fig. 3 Cold composite curve. (A) Individual temperature–enthalpy profiles for cold streams. (B) Combination of profiles for two cold streams.

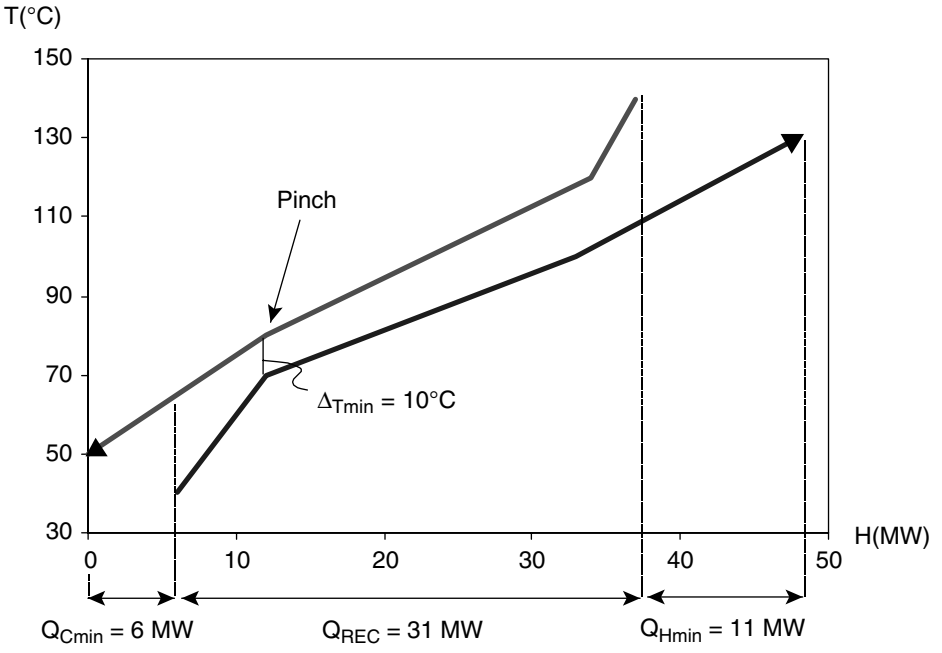


Fig. 4 Plotting the hot and cold composite curves together at $\Delta T_{\min} = 10°C$. *(View this art in color at www.dekker.com.)*

if, for example, plate heat exchangers are used, then the flow is more truly countercurrent, and a smaller temperature difference can be tolerated.

THE PINCH PRINCIPLE

The composite curves in Fig. 4 determine the target for maximum heat recovery and minimum hot and cold utility. The point of closest approach of the composite curves (where there is $\Delta T_{\min}$) limits their maximum heat recovery potential. This point of closest approach is known as the heat recovery pinch. However, not only does the pinch limit the heat recovery potential between the composite curves, it also is the key to achieving targets set by the composite curves in design.

To understand the true significance of the pinch, first divide the composite curves at the pinch, as illustrated in Fig. 6. Above the pinch (in temperature terms), the process is overall a heat sink. Heat recovery is possible from the composite hot stream to the composite cold stream above the pinch, but there is a residual amount of heat ($Q_{H_{min}}$) that is required to be imported from hot utility to satisfy the enthalpy imbalance. Below the pinch (in temperature terms), the process acts overall as a heat source. Heat recovery is possible from the composite hot curve to the cold composite curve below the pinch, but a residual amount of cooling ($Q_{C_{min}}$) is required to satisfy the enthalpy imbalance. Overall, the process above the pinch therefore acts as a heat sink, and that below the pinch acts overall as a heat source.

Now consider the possibility of transferring heat across the pinch from the overall process heat sink into the overall process heat source (Fig. 7). Temperature differences are such that it is feasible to transfer heat from hot streams above the pinch to cold streams below the pinch without violating the minimum temperature difference. On the other hand, it is not possible to transfer heat from the hot streams below the pinch to the cold streams above the pinch without either having negative temperature difference or violating the minimum temperature difference constraint. Thus, it is possible to transfer heat from above the

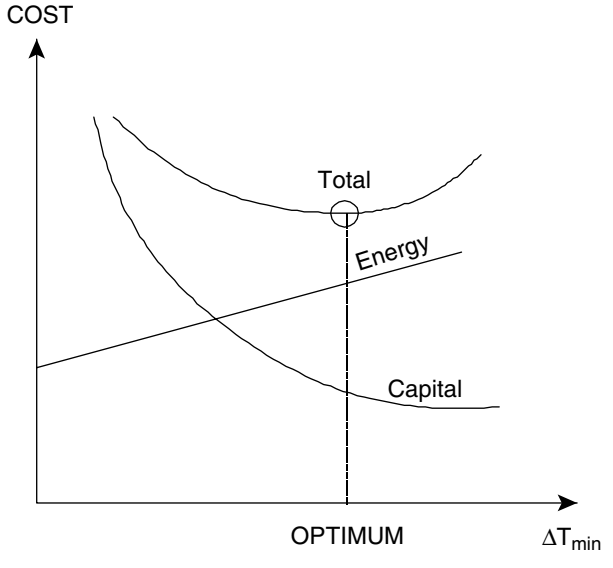


Fig. 5 Capital–energy tradeoff.

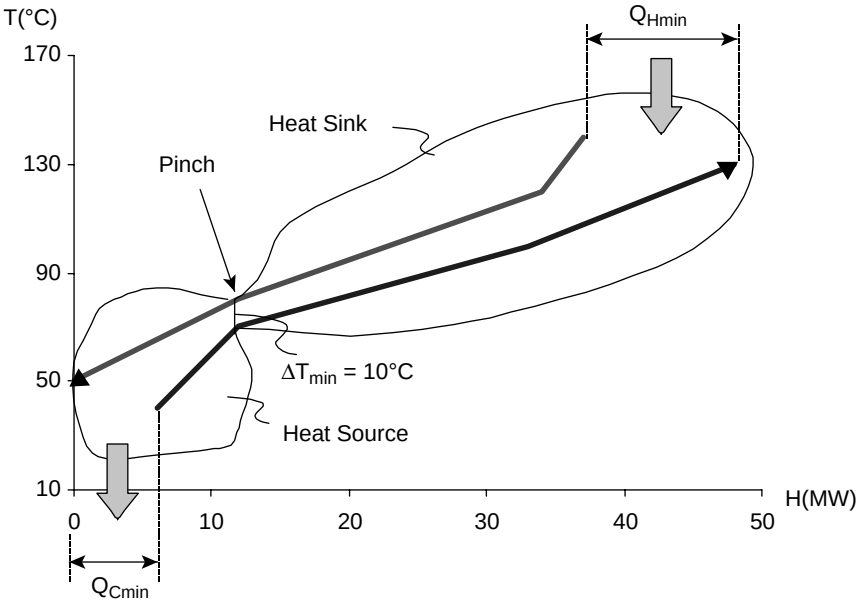


Fig. 6 The "pinch" divides the problem into a heat source and a heat sink. *(View this art in color at www.dekker.com.)*

pinch to below the pinch in process-to-process heat exchanges but not the other way around.

Having identified that it is feasible to transfer heat from above the pinch to below the pinch, suppose now that a quantity of heat XP is transferred in this way. Removal of the amount of heat XP from the system above the pinch now leads to a heat imbalance. It is not possible to correct this by transferring XP of heat from the below pinch to above the pinch, as this has been established to be infeasible. The only way to correct the enthalpy imbalance is to import an extra XP of heat from a hot utility. Having transferred XP of heat from above the pinch to below the pinch, the process below the pinch now has a surplus of even more heat than was previously the case. Thus, the result of transferring XP of heat from above the pinch to below the pinch is an increase of both the hot and the cold utility by that amount.

Process-to-process heat transfer is not the only way in which heat can be transferred across the pinch. There are basically three ways in which heat can be transferred across the pinch. The first, illustrated in Fig. 8A, is the one already discussed: transferring XP process-to-process across the pinch. Fig. 8B shows the use of hot utility below the pinch. Using hot utility to provide heat to cold streams below the pinch is feasible. However, $Q_{H_{min}}$ is still required to satisfy the enthalpy imbalance above the pinch, and the additional XP of heat transferred to the process below the pinch must ultimately be rejected to cold utility. This inappropriate use of hot utility below the pinch leads to an increase in the energy consumption. Finally, Fig. 8C shows an arrangement in which cold utility has been applied to process streams above the pinch. It is feasible to apply cold utility to the hot streams above the pinch. However, the consequence

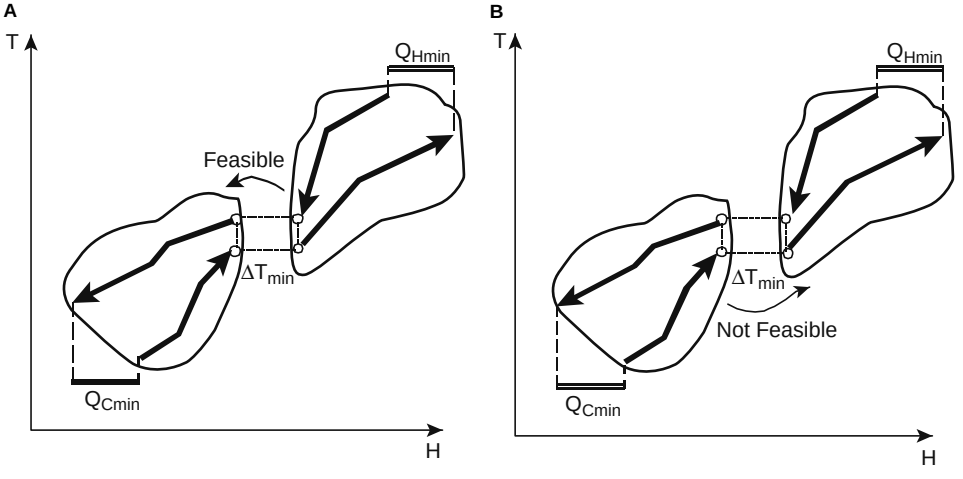


Fig. 7 Heat transfer across the pinch. (A) Heat transfer from above the pinch to below the pinch. (B) Heat transfer from below the pinch to above the pinch. *(View this art in color at www.dekker.com.)*

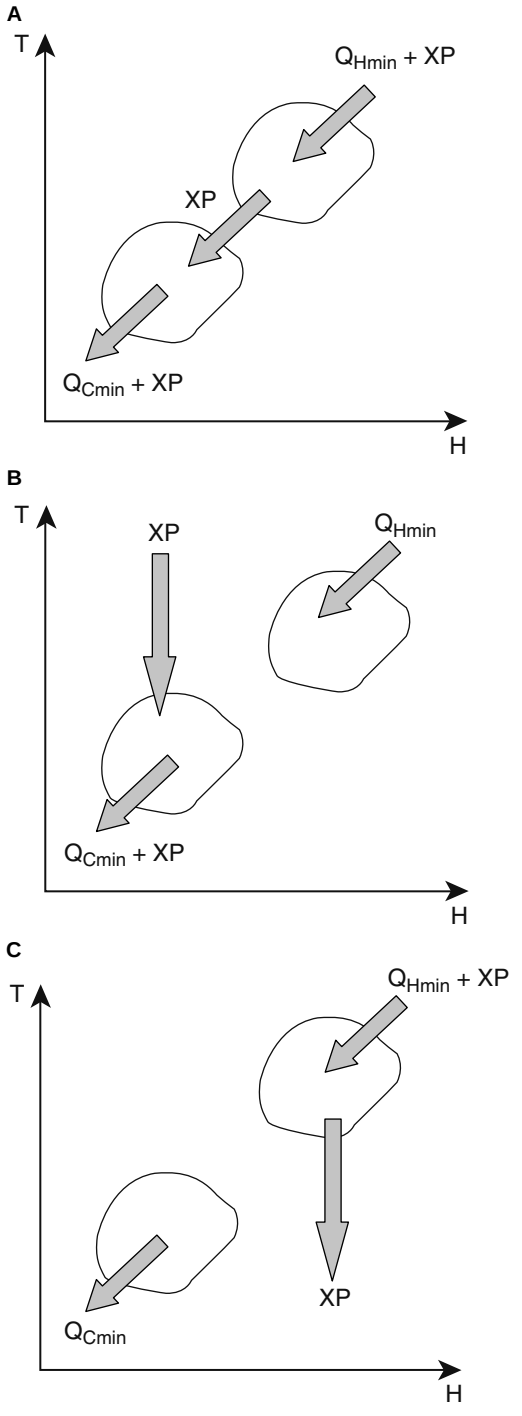


Fig. 8 Three ways of heat transfer across the pinch. (A) Process-to-process heat transfer across the pinch. (B) Use of hot utility below the pinch. (C) Use of cold utility above the pinch. *(View this art in color at www.dekker.com.)*

is now that the process above the pinch has a deficit of XP of heat that was previously being recovered from the hot streams to the cold streams above the pinch. This deficit cannot be corrected by transferring heat from below the pinch to above the pinch, for reasons already discussed. The only way this enthalpy imbalance can now be corrected is by importing an extra XP of heat from hot utility. The enthalpy balance below the pinch is maintained by $Q_{C_{min}}$ of cooling below the pinch. Thus, the inappropriate use of utilities leads to a corresponding increase in the consumption of both hot and cold utility.

If the condition of having no transfer of heat with a temperature difference smaller than ΔT_{min} is imposed, then each unit of heat transferred process-to-process from above the pinch to below the pinch or each unit of utility inappropriately used leads to a unit increase in the consumption of both hot and cold utility. Turning this argument around, to achieve the target for hot and cold utility set by the composite curves in design, the designer must not:

- Transfer heat across the pinch process-to-process
- Make use of inappropriate utility

Appropriate utilities are those required to satisfy the enthalpy imbalance in that part of the process.

THE PROBLEM TABLE

The composite curves provide a valuable tool both to set energy targets and to provide insights into the design problem. However, the composite curves are based on a graphical construction, and an alternative calculation, known as the *problem table*, can be developed to provide the energy targets. The problem table is developed here using the example from Table 1. The algorithm is in four steps.

Step 1. The first step in the problem table algorithm is to adjust the temperatures to include a finite temperature difference for heat transfer in the subsequent analysis. Hot streams are adjusted by reducing their temperature by $\Delta T_{min}/2$. Cold streams are adjusted by increasing their temperature by $\Delta T_{min}/2$. These adjusted temperatures are known as interval temperatures. They are not real temperatures, and the adjustment is made purely for the sake of the algorithm. If a hot and a cold stream have the same interval temperature, then they are in reality separated by just ΔT_{min}. Table 2 shows the supply and target temperatures of the streams from Table 1 adjusted for ΔT_{min}. In this case, the temperatures need to be adjusted by 5°C.

Step 2. In the second step, interval heat balances are created, as illustrated in Fig. 9. The interval temperatures created by adjusting the supply and target temperatures by $\Delta T_{min}/2$ are listed in descending order to create enthalpy intervals. The hot and cold streams are then superimposed on their appropriate temperature

Table 2 Shifted temperatures for the data in Table 1

Stream	Type	Supply temperature, T_S (°C)	Target temperature, T_T (°C)	Shifted supply temperature, T_S* (°C)	Shifted target temperature, T_T* (°C)
1	Hot	140	80	135	75
2	Hot	120	50	115	45
3	Cold	70	130	75	135
4	Cold	40	100	45	105

intervals. It should be noted in Fig. 9 that within each temperature interval, the hot and the cold streams are separated in temperature terms by just ΔT_{min}. This ensures that heat transfer is thermodynamically feasible within each temperature interval.

Step 3. The next stage in the algorithm is to perform a heat balance within each temperature interval. This heat balance is given by

$$\Delta H_{INTERVAL} = \Delta T_{INTERVAL}\left[\sum CP_C - \sum CP_H\right]$$

The enthalpy balance sums up the CPs of the cold streams within an enthalpy interval, sums up the CPs of the hot streams within the same interval, subtracts the latter from the former, and then multiplies the result by the temperature difference across the interval. The resulting heat balance might be a positive or a negative value. A positive value indicates that there is a net deficit of heat within a temperature interval. A negative value indicates that there is a net surplus of heat in an interval. The interval heat balances for all of the intervals are shown in Fig. 10. Such interval heat balances allow for heat recovery within each temperature interval. However, to complete the picture, the algorithm must consider heat recovery between intervals. Where an interval has a surplus of heat in Fig. 10, this should be used, where possible, to supply at least some of the deficit for those intervals that have a heat deficit. For heat transfer to be feasible, heat can only be transferred down the interval temperature scale. Fig. 11A shows heat cascaded down through the temperature intervals, assuming that there is no heat input into the highest temperature interval from utility. The heat flows between the temperature intervals show some positive and some negative values. A negative number indicates that heat needs to be transferred up the temperature scale, and this is not feasible. Thus, the cascade in Fig. 11A is not feasible, and some remedial action must be taken to correct the negative values. This can be done by increasing the heat flow through the entire system. Thus, heat should be added from hot utility to the interval with the highest temperature.

Step 4. To make the cascade in Fig. 11A feasible, enough heat must be added at the first temperature interval to make all of the flows either zero or positive. The largest negative heat flow in Fig. 11A is –11 MW. This must be added from hot utility to the first temperature interval. Fig. 11B shows the corresponding cascade. The amount of heat added in the first temperature interval corresponds to the target for minimum hot utility of 11 MW. The heat rejected from the last interval, 6 MW, corresponds to the target for cooling utility. These targets for hot and cold utility

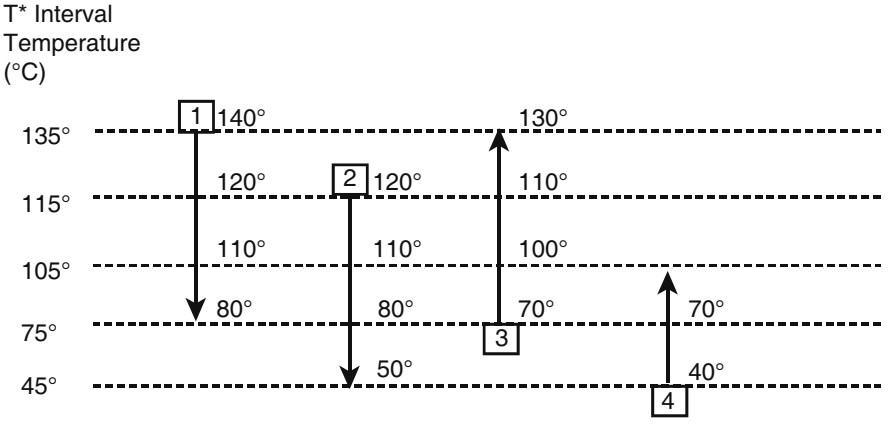


Fig. 9 The stream population for the data in Table 2. *(View this art in color at www.dekker.com.)*

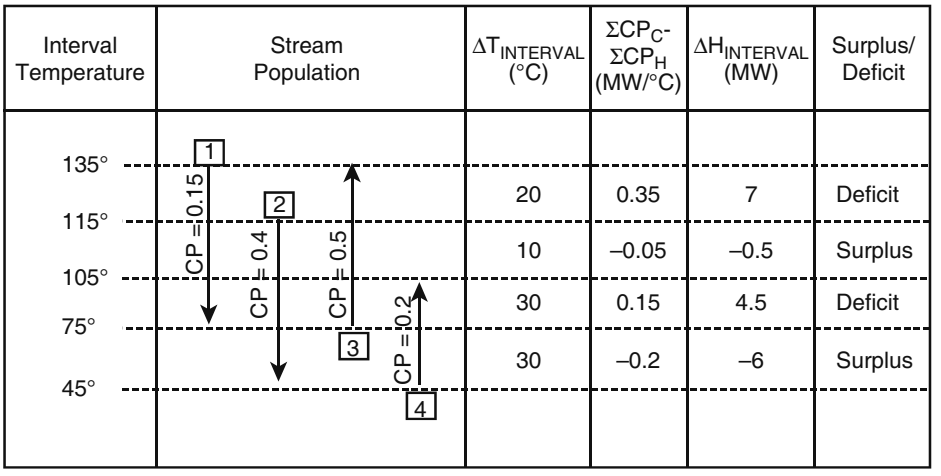


Fig. 10 Temperature interval heat balance. *(View this art in color at www.dekker.com.)*

agree with the targets generated from the composite curves.

The other feature to note regarding the cascade in Fig. 11B is the point where the heat flow becomes zero. This is at an interval temperature of 75°C. This corresponds to the location of the pinch. Given that the interval temperature for the pinch is 75°C, the hot stream pinch temperature is therefore 80°C and the cold stream pinch temperature 70°C. This agrees with the location of the pinch from the composite curves. Thus, the problem table algorithm provides the information required for minimum hot and cold utility targets and the location of the pinch that is critical in the next step, which is to design to achieve the targets.

THE PINCH DESIGN METHOD

Having established the hot and cold utility targets and the location of the pinch, the next step is to produce a design that will achieve the targets. The design procedure starts by setting up a design grid, as shown in Fig. 12. The hot streams are shown at the top of the diagram running from left to right and the cold streams at the bottom of the diagram running from right to left. A vertical line at the pinch temperature of 80°C for the hot streams and 70°C for the cold streams divides the grid into two parts. To the left is the part of the process above the pinch. To the right is the part of the process below the pinch.

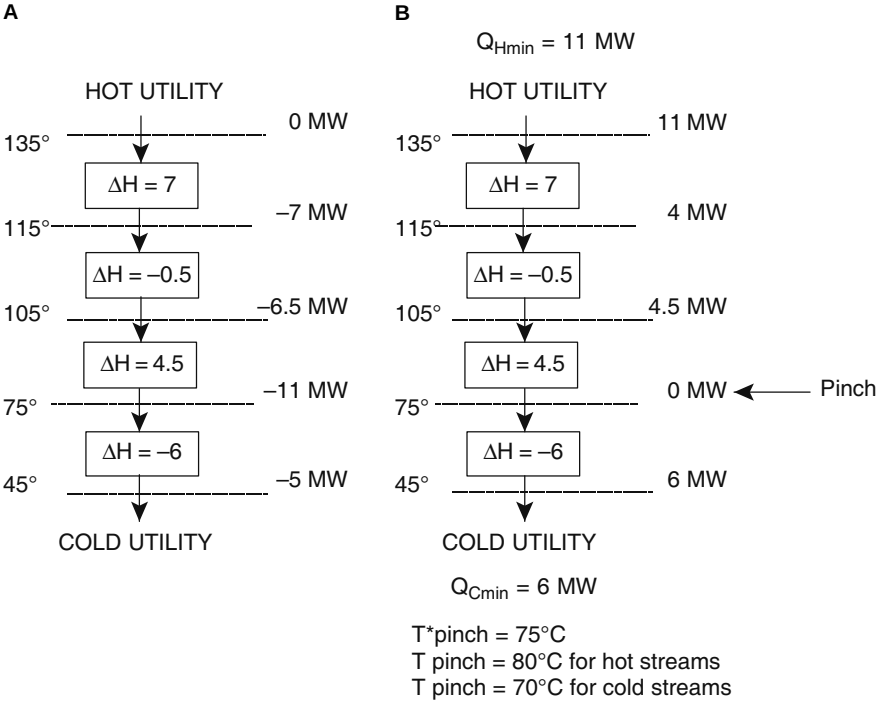


Fig. 11 The problem table cascade. (A) Cascade without heat input. (B) Cascade with heat added from hot utility. *(View this art in color at www.dekker.com.)*

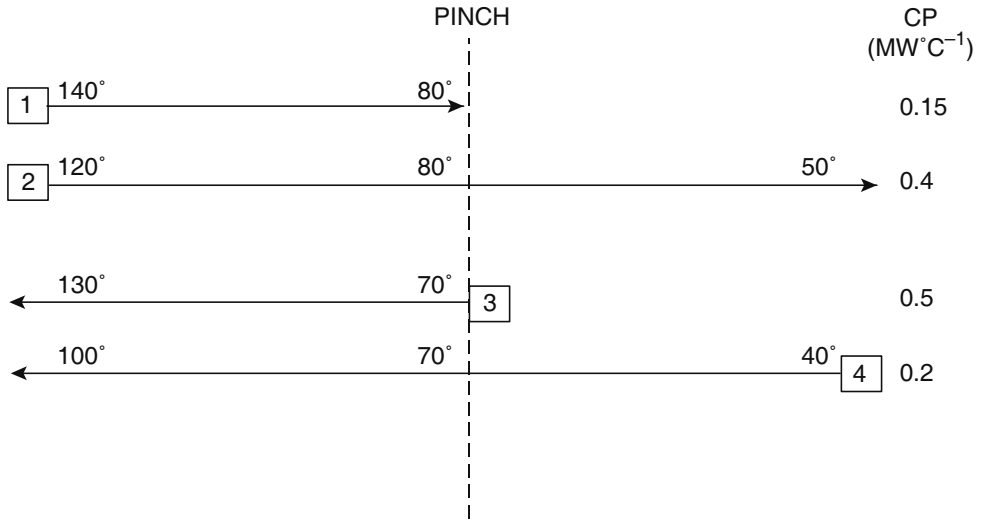


Fig. 12 Setting up a design grid. *(View this art in color at www.dekker.com.)*

To develop a design strategy, the temptation would be to start the design with the hottest hot stream and recover heat to a cold stream, gradually working from high temperatures to low temperatures. However, the composite curves in Figs. 4 and 6 show that the most constrained part of the design, where the temperature differences are the smallest, is around the heat recovery pinch. If the design was started at the highest temperatures, decisions would be made in a part of the design where there is freedom of choice, because of the large temperature differences. Working from this part of the design to the region of the pinch would mean working towards a more constrained part of the design. Decisions made away from the pinch might therefore create problems once the constrained part of the design around the pinch is approached. Thus, a sensible approach is to start the design where it is most constrained: around the pinch.

A heat exchanger in the grid diagram is represented by a vertical line joining two open circles on the streams that are being matched. Fig. 13A shows a match placed above the pinch, at the pinch. In Fig. 13A, the match is between Hot Stream 2 and Cold Stream 4. Because the match is at the pinch, the match starts with $\Delta T_{\min}$ at its cold end. The hot stream has a CP of 0.4 and the cold stream, a CP of 0.2. Because of the relative CPs of the two streams, as heat is transferred, moving away from the pinch the temperature differences become smaller and, therefore, infeasible. Fig. 13B shows an alternative arrangement whereby Hot Stream 2 is matched with Cold Stream 3. This time, the CP of the hot stream is 0.4 and that of the cold stream is 0.5. Starting with $\Delta T_{\min}$ at the pinch, because of the relative values of the CPs, now the temperature differences become wider as the heat duty increases moving away from the pinch: $CP_H \leq CP_C$. The argument only applies strictly for matches above the pinch, at the pinch.

Fig. 14 shows the part of the design below the pinch. In this particular case, because the heat exchanger network is simple, there is only one hot stream and one cold stream below the pinch. The match at the pinch shown in Fig. 14 is seen to be feasible with a hot stream CP of 0.4 matched against a cold stream CP of 0.2. By a reasoning analogous to the arguments for the CP inequality above the pinch, it can be concluded that below the pinch, at the pinch, $CP_H \geq CP_C$ for a feasible match.

To help identify suitable matches between hot and cold streams for the overall network, a CP table can be adopted, as illustrated in Fig. 15. The CP table lists the hot and cold streams in descending order of CP. This allows feasible matches between hot and cold streams at the pinch to be readily identified. Fig. 15(A) shows the hot and cold matches for the streams above and below the pinch. These satisfy the CP inequalities, and therefore the temperature differences throughout the matches are feasible.

The next question in the design is to determine how large these matches should be in terms of their duty. It would make sense for these matches to be as large as they can be to minimize the number of matches and reduce the complexity in capital costs of the design. The matches can only be made as big as the smaller of the duties on the two streams being matched. Thus, the tick-off heuristic can be used to minimize the number of matches by satisfying one of the streams in terms of its heat duty each time a match is placed. Fig. 15B shows the matches around the pinch with their duties maximized to satisfy some of the streams.

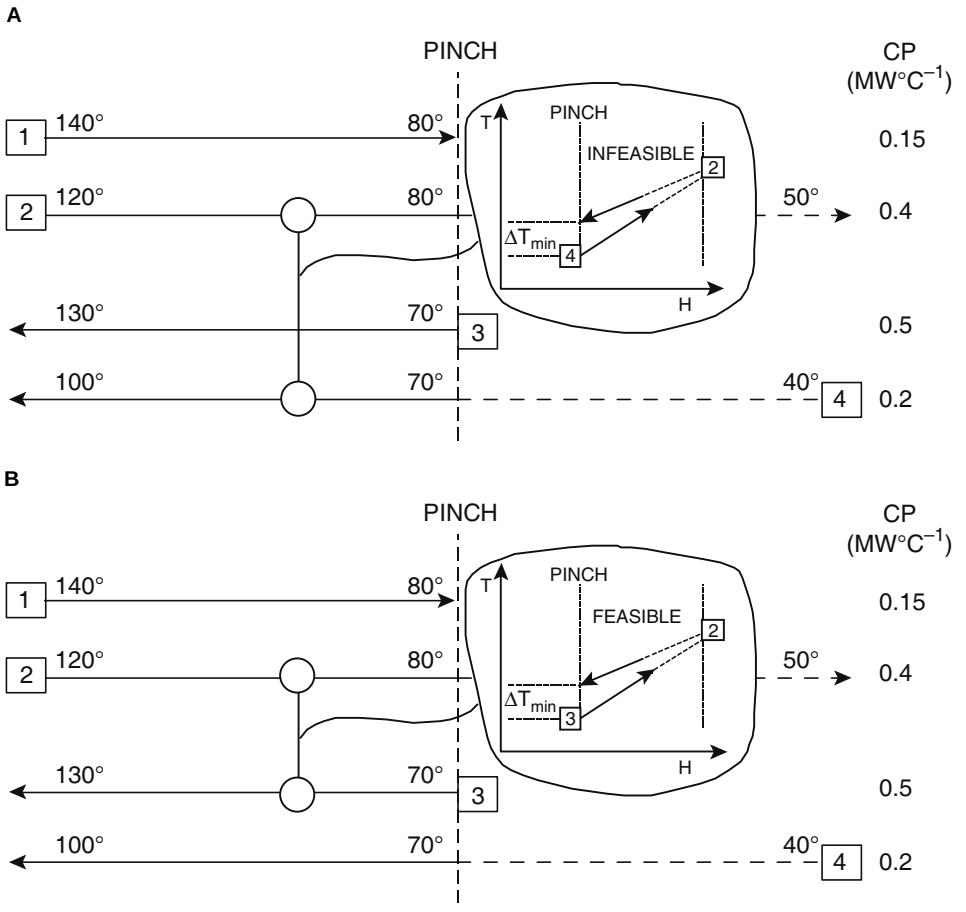


Fig. 13 Above the pinch, at the pinch, $CP_C \geq CP_H$ for a feasible match. (A) A match between hot stream 2 and cold stream 4. (B) A match between hot stream 2 and cold stream 3. *(View this art in color at www.dekker.com.)*

To complete the design, further matches are required, and also the placement of hot and cold utilities. In Fig. 15B, above the pinch, all of the hot streams must be reduced to pinch temperature by heat recovery. Thus, in Fig. 15B, a further match between a hot stream and a cold stream must be placed for the target to be achieved. This is shown in Fig. 15C, where an additional match has been placed above the pinch, maximizing its load. Once all of the hot streams above the pinch have been satisfied, it only remains to place hot utilities, as shown in Fig. 15C. Hot utilities in the grid diagram are represented by an open circle with 'H' on the stream to which the hot utility is applied.

Similar arguments apply below the pinch. In Fig. 15B, the match between Hot Stream 2 and Cold

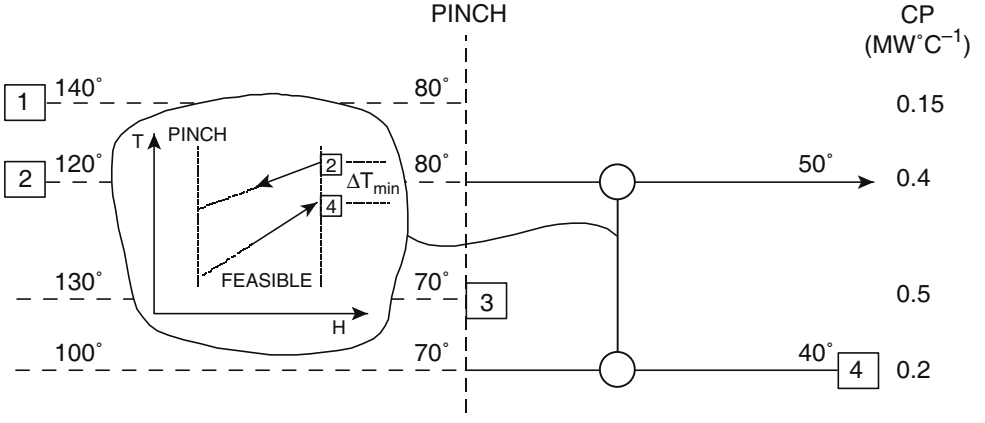


Fig. 14 Below the pinch, at the pinch, $CP_H \geq CP_C$ for a feasible match. *(View this art in color at www.dekker.com.)*

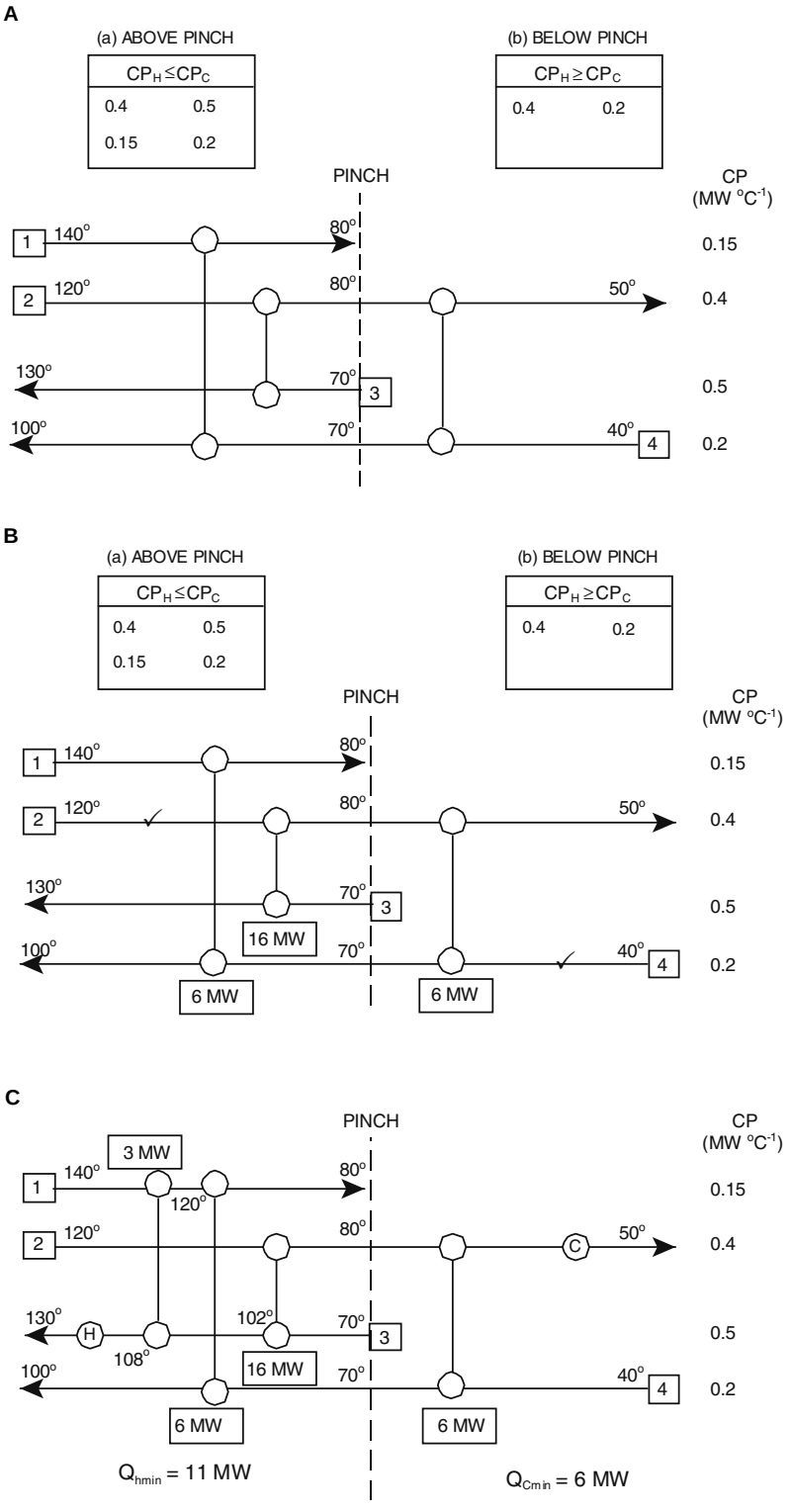


Fig. 15 The pinch design method. (A) CP tables follow feasible matches to be identified at the pinch. (B) Maximising the heat duties on the matches. (C) The completed design. *(View this art in color at www.dekker.com.)*

Stream 4 satisfies Stream 4 below the pinch once its duty has been maximized. Below the pinch, all of the cold streams must be heated to pinch temperature by heat recovery. Maximizing the duty on the match between Hot Stream 2 and Cold Stream 4 below the pinch allows Cold Stream 4 to be completely satisfied below the pinch. Thus, in this case, no further heat recovery matches are required below the pinch as all of the cold streams are satisfied. All that remains to complete the design is to apply coolers to the hot streams not satisfied by heat recovery. In the grid diagram, a cooler using cold utility is represented by

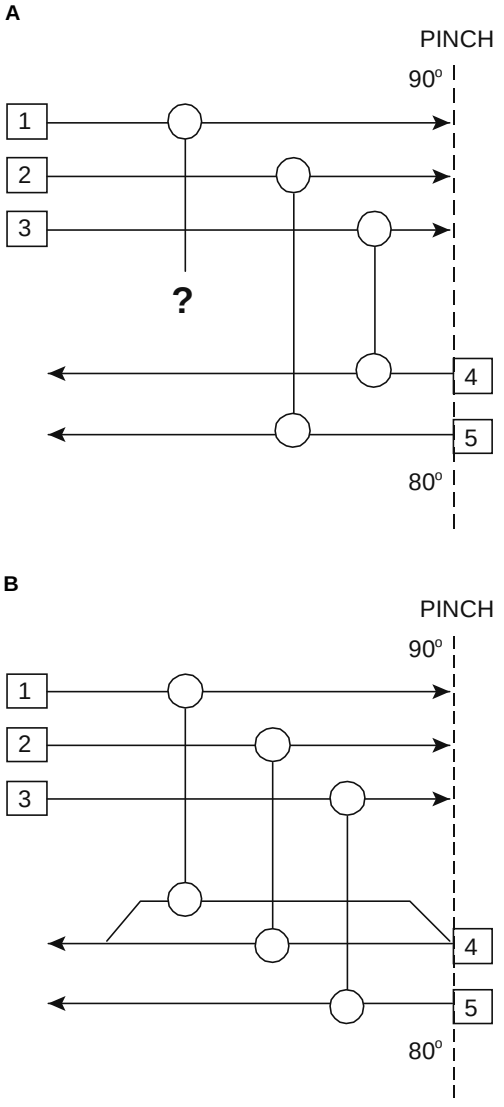


Fig. 16 Above the pinch, $N_C \geq N_H$. (A) An above-pinch part of a problem. (B) Splitting a cold stream allows all of the hot streams to be satisfied. *(View this art in color at www.dekker.com.)*

an open circle with a 'C' on the stream to which the cold utility is applied.

The completed design in Fig. 15C indicates a hot and cold utility requirement in agreement with the targets set by both the composite curves and the problem table analysis.

STREAM SPLITTING

Fig. 16A shows part of a problem above the pinch. The pinch temperature is 90°C for the hot streams and 80°C for the cold streams. The hot streams above the pinch must be cooled to a pinch temperature of 90°C by heat recovery. To reach 90°C, each of the three hot streams must be matched against a cold stream at 80°C at the pinch to have a feasible match. Unfortunately, there are three hot streams and only two cold streams. This means that one of the hot streams does not have a cold stream with the necessary 80°C to match against. This design problem can be solved by splitting a cold stream, as shown in Fig. 16B. Splitting a cold stream effectively creates another cold stream at 80°C. Now each of the hot streams can be matched against a cold stream at 80°C. The conclusion from this is that above the pinch, at the pinch, $N_H \leq N_C$.

Fig. 17A shows part of a problem below the pinch. In this case, three cold streams must be heated to the pinch temperature of 80°C without using utility. This

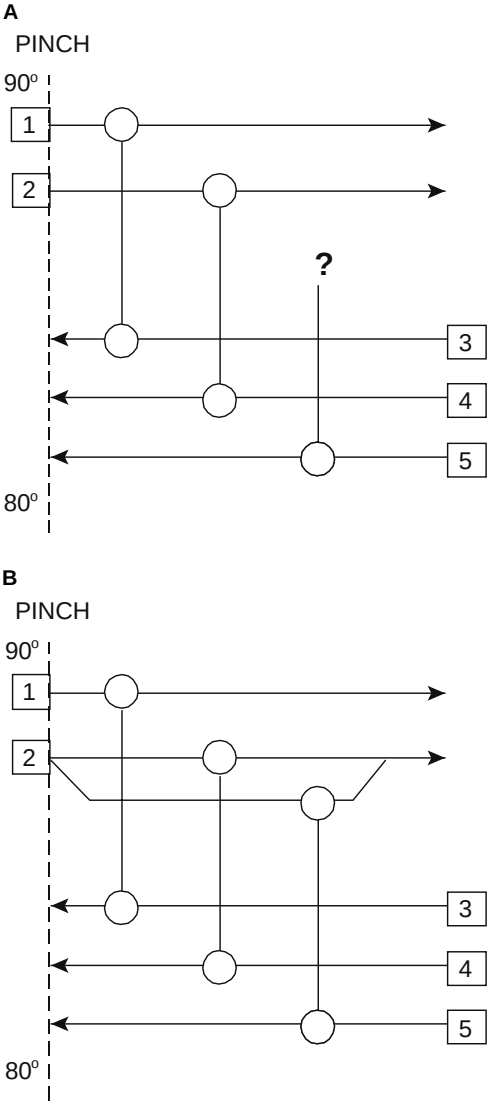


Fig. 17 Below the pinch, $N_H \geq N_C$. (A) A below-pinch part of a problem. (B) Splitting a hot stream allows all of the cold stream to be satisfied. *(View this art in color at www.dekker.com.)*

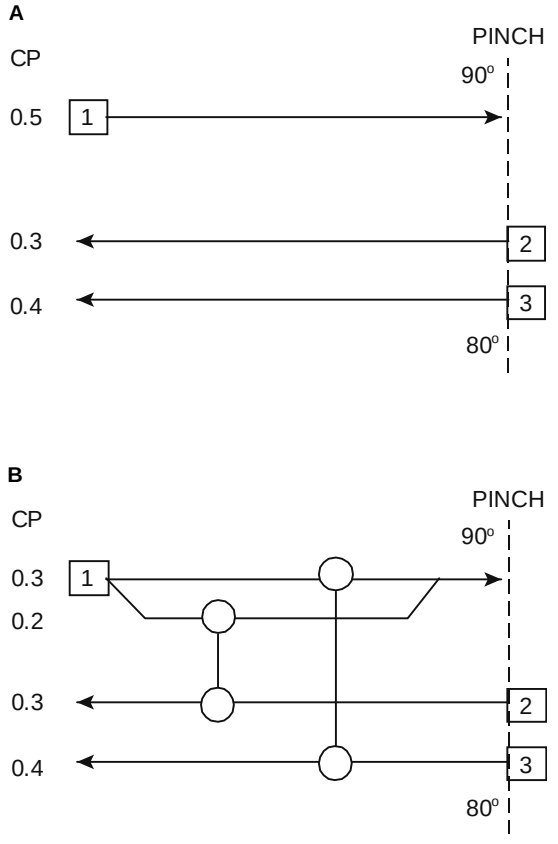


Fig. 18 Splitting a hot stream above the pinch for CP inequalities. (A) An above-pinch part of a problem in which the CP-inequalities cannot be satisfied. (B) Spilling a hot stream allow the CP-inequalities to be satisfied. *(View this art in color at www.dekker.com.)*

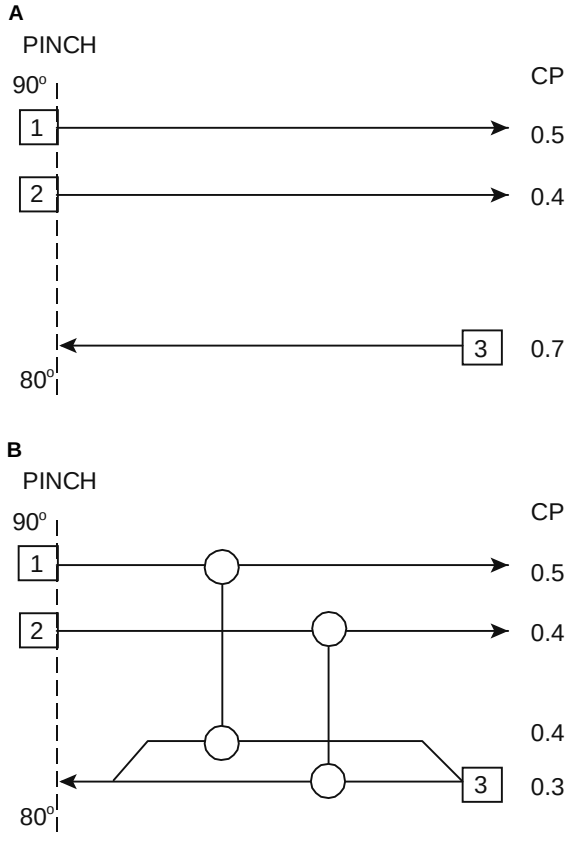


Fig. 19 Splitting a cold stream below the pinch for CP inequalities. (A) A below-pinch part of a problem in which the CP-inequalities cannot be satisfied. (B) Splitting a cold stream allow the CP-inequalities to be satisfied. *(View this art in color at www.dekker.com.)*

means that the three cold streams must be heated to 80°C by heat recovery. Two hot streams are available at 90°C. Thus, two of the cold streams can be satisfied but the third cannot. To solve this design problem requires another stream to be made available at 90°C. This can be achieved by splitting one of the hot streams, as shown in Fig. 17B. Now each of the three cold streams can be matched with a hot stream at a temperature of 90°C. The conclusion from this is that below the pinch, $N_H \geq N_C$.

Another above-pinch problem is shown in Fig. 18A. This time, there is one hot stream and there are two cold streams. Thus, as far as the number of streams is concerned, there should be no need to split a stream. However, the CP inequality also needs to be satisfied above the pinch: the hot stream, with a CP of 0.5, requires a cold stream with a greater CP. Neither of the two cold streams in Fig. 18A is large enough to satisfy the CP inequality. The design problem can be solved by splitting the hot stream into two branches, as shown in Fig. 18B. Splitting arbitrarily into two branches with a CP of 0.3 and 0.2 allows matches to be made that will satisfy the CP inequality. However, there is some freedom to change the branch flow rates.

Finally, Fig. 19A shows a below-pinch part of a problem. Cold Stream 3 has a CP of 0.7 and requires to be matched against a hot stream with a greater CP. Neither of the hot streams in Fig. 19A is large enough to satisfy the CP inequality. This time, the design problem is solved by splitting a cold stream into two branches, as shown in Fig. 19B. Now, the matches can be placed to satisfy the CP inequality required at the pinch, below the pinch.

Thus, there are two reasons why stream splitting is required at the pinch. The inequality regarding the number of streams must be satisfied; also, the CP inequality must be satisfied for each match.

THE GRAND COMPOSITE CURVE

So far, the discussion has related to problems in which there was one hot utility available and one cold utility

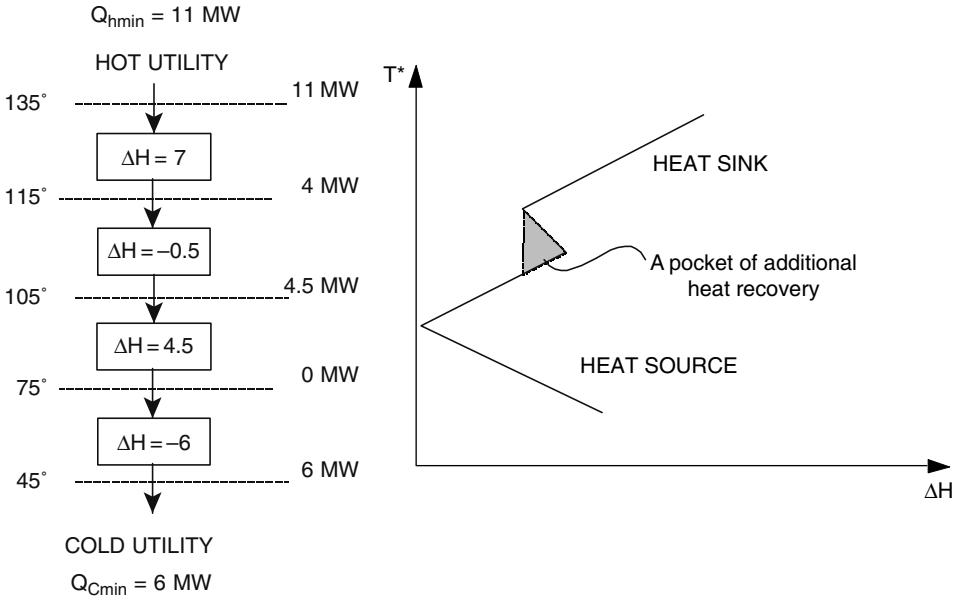


Fig. 20 The grand composite curve for the data in Table 1. *(View this art in color at www.dekker.com.)*

available. To deal with situations in which there are various utilities available, with different characteristics, another tool, known as the grand composite curve, is required. To construct the grand composite curve requires the heat flow from the problem table cascade to be plotted, as illustrated in Fig. 20. The grand composite curve represents the residual heating and cooling that are required after heat recovery has been carried

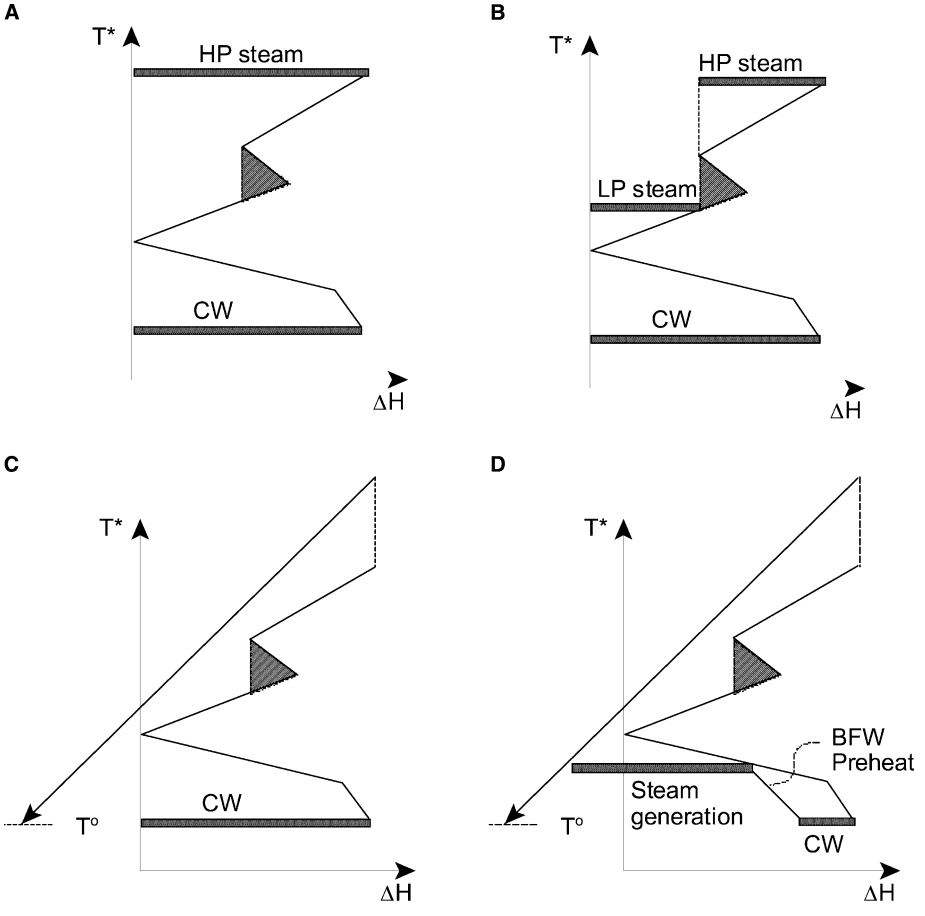


Fig. 21 Examples of the use of the grand composite curve. (A) High pressure stream and cooling water; (B) two steam levels and cooling water; (C) fuel gas and cooling water; (D) fuel gas, stream generation, and cooling water.

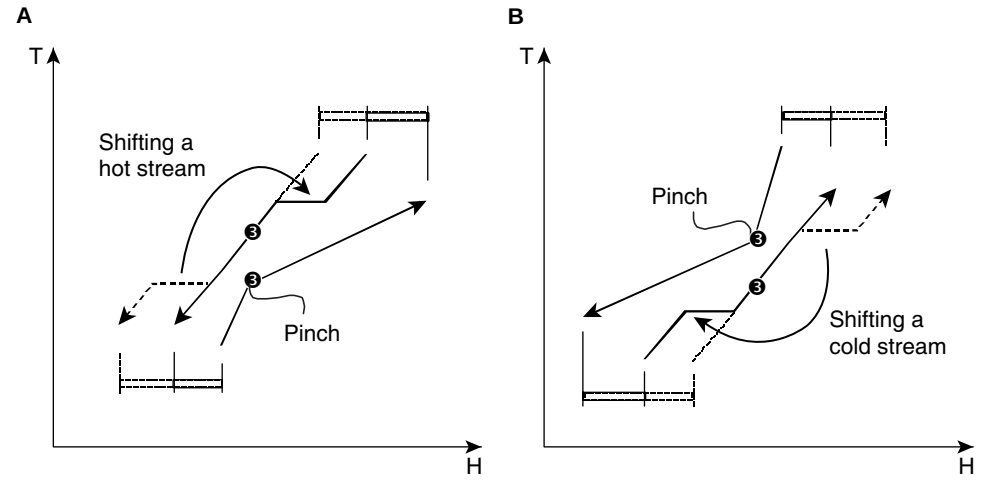


Fig. 22 Shifting streams across the pinch. (A) Shifting a hot stream from below to above the pinch reduces the utility requirements. (B) Shifting a cold stream from above to below the pinch reduces the utility requirements. *(View this art in color at www.dekker.com.)*

out. However, the heating and cooling are represented in terms of not only enthalpy, but also temperature. The grand composite curve provides a convenient tool to select the most appropriate mix of utilities for a problem.

In Fig. 20, it can be seen that there is a pocket where additional heat recovery takes place. This is in addition to the heat recovery that takes place throughout the cascade, as it is captured by the heat balance within each temperature interval. In the pocket, a local surplus of heat is used to satisfy a deficit of heat where possible.

Fig. 21 illustrates some examples of the use of the grand composite curve. Fig. 21(A) shows high-pressure (HP) steam and cooling water (CW) used to satisfy the heating and cooling requirements of a grand composite curve. In contrast, Fig. 21(B) shows the hot utility being satisfied by a mixture of HP and low-pressure (LP) steam. Fig. 21(C) shows yet another option in which a flue gas is used to provide the process hot utility. This flue gas could in principle be either that from a furnace or the exhaust from a gas turbine. The fourth option, in Fig. 21(D), shows the hot utility being satisfied by a furnace flue gas, but with cooling being taken up by LP steam generation, by a feedwater preheat and CW.

One point should be noted regarding the matching of utilities against the grand composite curve, as shown in Fig. 21. When a utility profile touches the grand composite curve, this does not imply that $\Delta T = 0$. The temperature shift included in the construction of the problem table means that the hot and cold streams are shifted by $\Delta T_{\min}/2$. This also applies to the hot and cold utilities. The result is that when a utility profile touches the grand composite curve, this implies that $\Delta T = \Delta T_{\min}$ rather than $\Delta T = 0$.

The grand composite curve is an extremely useful tool that can be used to screen and select different utilities and different utility mixes.

PROCESS CHANGES

The discussion so far has assumed that the supply and target temperatures and heat capacity flow rates of the streams in a heat exchanger network have been fixed. However, process flowsheets often feature degrees of freedom at the disposal of the designer that allow the performance of the heat exchanger network to be increased. Once the basic material and energy balance of the flowsheet have been fixed, all of the information required to carry out energy targeting is available (supply temperature, target temperature, and enthalpy as a function of temperature). There is often freedom to change the conditions within the material and energy balance within certain limits. Such changes might include, for example, changes in:

- Vaporizer and condenser pressures
- Distillation pressures
- Distillation reflux ratios
- Evaporator pressures
- Number of evaporator stages
- Reactor conversion
- Recycle flowrates

Fig. 22 illustrates how appropriate changes to the operating conditions within a flowsheet can create improvements to the targets for the process. In Fig. 22(A), a condensing stream, which is part of the hot composite curve below the pinch, has had its

pressure increased such that it is shifted from below the pinch to above the pinch. The effect is to reduce the cold utility and hot utility simultaneously. In general, shifting a hot stream from below the pinch to above the pinch causes an improvement in the targets for hot and cold utility.

The corresponding case for shifting a cold stream is shown in Fig. 22(B). This time, a vaporization stream, which is part of the cold composite curve above the pinch, has had its pressure decreased until it is below the pinch. This causes a decrease in the hot utility and the cold utility. Thus, any shift of a cold stream from above the pinch to below the pinch causes an improvement in the targets for the utilities. However, shifting streams relative to the pinch is only one way to bring about an improvement in the targets for utility consumption from process changes. In general,

- Increases in the total hot stream duty above the pinch
- Decreases in the total cold stream duty above the pinch
- Decreases in the total hot stream duty below the pinch
- Increases in the cold stream heat duty below the pinch

bring about a decrease in the utility requirements. These guidelines provide a reference for appropriate changes to improve the targets. Such process changes that improve heat recovery can be extremely important in creating cost-effective projects.

CONCLUSIONS

Pinch analysis exploits thermodynamics to allow energy systems to be targeted. These targets allow scoping and screening of different options. Composite curves can be used to provide insights into where process changes can improve targets. Process changes would mean exploiting the flexibility to change temperatures, flowrates, and pressures to alter the composite curves in a favorable way. At the same time, the grand composite curve can be used to screen different utility mixes. Once the most appropriate process conditions and mix of utilities have been chosen, the design of the heat exchanger network can proceed as described here.

BIBLIOGRAPHY

Cerda, J.; Westerberg, A.W.; Mason, D.; Linnhoff, B. Minimum utility usage in heat exchanger network synthesis—a transportation problem. Chem. Eng. Sci. **1983**, *38* (3), 373–387.

Hohman, E.C. Optimum Networks of Heat Exchange, Ph.D. Thesis; University of Southern California, 1971.

Linnhoff, B.; Hindmarsh, E. The pinch design method of heat exchanger networks. Chem. Eng. Sci. **1983**, *38* (5), 745–763.

Linnhoff, B.; Mason, D.R.; Wardle, I. Understanding heat exchanger networks. Comput. Chem. Eng. **1979**, *3*, 295–302.

Linnhoff, B.; Townsend, D.W.; Boland, D.; Hewitt, G.F.; Thomas, B.E.A.; Guy, A.R.; Marsland, R.H. *User Guide on Process Integration for the Efficient Use of Energy*; IChemE: Rugby, U.K., 1982.

Smith, R. *Chemical Process Design and Integration*; John Wiley & Sons Ltd: West Sussex, U.K., 2005.

Townsend, D.W.; Linnhoff, B. Heat and power networks in process design—Part I: Criteria for placement of heat engines and heat pumps in process networks. AIChE J. **1983**, *29* (5), 742–748.

Umeda, T.; Harada, T.; Shiroko, K. A thermodynamic approach to the synthesis of heat integration systems in chemical processes. Comput. Chem. Eng. **1979**, *3*, 273–282.

Pipeline Safety

Glenn B. DeWolf
URS Corporation, Austin, Texas, U.S.A.

INTRODUCTION

Pipelines are an essential part of the energy and process industries infrastructure of modern society. They are generally viewed as a safe and economical means of transportation, but the perception of relative safety is subjective. Pipelines transport hazardous liquids and gases from their sources to customers. Transmission pipelines carry products at high pressure for relatively long distances. Gas pipeline distribution systems distribute gas in local communities, typically, at lower pressures than transmission lines. Numerous industry standards and government regulations dealing with design, construction, operations, and maintenance are the basis of managing pipeline safety. Factors affecting safety are associated with the attributes of a pipeline system and the threats that the system faces to its containment integrity. This includes factors that affect the likelihood and the severity of the consequences of failure. Technology and practices are continuing to evolve for better safety management.

THE PIPELINE INFRASTRUCTURE

Pipelines are a vital part of the energy and chemical transportation infrastructure of modern society. They are generally viewed as a safe and economical means to deliver liquid and gas fuels and chemical raw materials to customers. However, the perception of their safety varies with segments of the population. The most common pipeline with which the average person is familiar is the gas main of the local gas distribution system with service connections to homes and businesses in communities. The overall pipeline infrastructure is much more and includes long-distance transmission pipelines in addition to local distribution system pipelines. Besides natural gas, transmission pipelines transport crude oil, gasoline, petroleum products, and chemical products.

In the United States, there are approximately 160,000 mi of liquid transmission pipelines, 200,000 mi of gas transmission pipelines, and over 1,000,000 mi of distribution lines, if the mileage of individual customer connections to municipal gas mains is included.[1,2]

New pipeline design and construction projects are an ongoing part of infrastructure expansion driven by population growth and economic development. Existing pipeline modifications and repairs are also ongoing to maintain the infrastructure. Safety considerations are fundamental to all parts of the pipeline lifecycle from planning through decommissioning and removal from service.

Distribution pipelines are mostly in populated areas and transmission lines may cross through or near multiple populated areas intermittently with sparsely populated areas as they carry the product long distances. Pipelines lie within a right-of-way land corridor, usually leased from the landowners. The right-of-way provides access for ongoing maintenance and future expansion of the systems.

SAFETY CONCERNS

Safety concerns are associated with the flammability and, sometimes, toxic properties of transported products, the high pipeline pressures (hundreds of pounds to over a thousand pounds per square inch), and the proximity to population or environmentally sensitive areas at some locations along the rights-of-way.

CAUSES OF FAILURE

Failure causes have been identified and classified to varying degrees of detail. Specific causes of failure are grouped by categories. Within each of these categories, there are specific causes. Classifications differ in the published technical literature, including regulatory reporting, but the various classifications account for the same major causes. One classification of causes identifies 22 threats in five categories that lead to pipeline failure:[3]

1. External forces
    a. Earth movement
    b. Heavy rains/floods
    c. Lightning
    d. Cold weather
    e. Third party damage
    f. Previously damaged pipe
    g. Vandalism

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007900

2. Corrosion
    a. External corrosion
    b. Internal corrosion
    c. Stress corrosion cracking
3. Materials
    a. Defective pipe seam
    b. Defective fabrication weld
    c. Defective girth weld
    d. Construction damage
    e. Defective pipe
4. Equipment and operations
    a. Malfunction of control/relief equipment
    b. Threads stripped, broken coupling
    c. Gasket/O-ring failure
    d. Seal/pump packing failure
    e. Incorrect operation
5. Other
    a. Miscellaneous
    b. Unknown

Similar compilations can be found in other technical literature sources.[4]

As causes have been identified, they have been investigated for better understanding and eventual resolution. Industry and government support specific studies to examine serious problems in-depth. For example, after a number of seam failures years ago, it was found that some pipe manufactured before 1970 that was seam-welded by electric resistance welding was subject to unacceptable rates of seam failure. A specialized study focused on this single issue.[5] Results of research and other new knowledge about pipeline integrity and safety are presented at pipeline conferences that involve the regulators, industry, and other stakeholders. By these forums, the operators, regulators, and others share their knowledge about lessons learned, needs, and ways to improve pipeline integrity, reliability, and safety.

Historical experience with pipeline failures reveals that the two single most common causes for failures are outside forces and corrosion. There are variations in individual systems. Accordingly, a significant part of the effort in achieving better pipeline safety is corrosion control and damage prevention directed at outside forces. The other causes of pipeline failure, individually, contribute to a much smaller proportion of pipeline incidents, but they are still addressed because their aggregate contribution matters.

Corrosion occurs either internally or externally. External corrosion can occur on both surface and buried portions of the pipeline. Aboveground, piping is protected by painting or specialized coatings. Belowground, piping is protected by specialized coatings and cathodic protection. Coatings can be of a number of materials and can be applied at the factory and in the field. At welded joints, the coating is field-applied. An important safety aspect of coating application is that the pipe is properly cleaned before coating, and the coating properly applied. Cathodic protection (CP) applies a voltage differential between the pipe and the soil in such a way that metal loss from exposed (uncoated) metal, from a coating defect, is prevented or reduced. Maintenance of cathodic protection involves periodic checks on the CP system components and on the pipe-to-soil voltage levels at various locations along the pipeline route. Deviations from specified conditions can be an indication of external corrosion. Where such corrosion is suspected, the pipe can be dug up, examined, and repaired or replaced according to need.

Internal corrosion occurs from chemically reactive constituents in the product, including contaminants that might accumulate in the system over time and moisture. Acidic constituents are the major contributors to internal corrosion. Internal corrosion is controlled by monitoring and controlling the product composition and by periodic cleaning of the line. This is done by passing a ''cleaning pig'' through the line and removing the accumulated dirt and debris. A cleaning pig is a plug-like device pushed through the line by fluid pressure. Special devices called pig launchers and pig receivers (pig traps) are installed at each end of a piggable pipe section or segment.

For outside forces, appropriate controls must address human-caused damage and natural forces damage. Human-caused damage can be classified as first, second, or third party according to whether it is self-inflicted by the pipeline operating company, the companies contractors, or third parties. Third-party damage is a major category for human-caused outside force. Other outside forces include natural forces, such as subsidence, seismic movement, landslides, frost heave, expansion and contraction of soils, erosion, flooding, and even lightning strikes for surface equipment. Proper design and construction for environmental and location factors related to outside forces followed by operating procedures addresses these threats. Procedures include patrols along the right of way, informing persons living or working near the pipeline, contractors who may be involved in work on or near the right of way, and other public. Damage prevention programs that require a notification call to the pipeline operator by a party doing work near, on or subsurface to the right of way are also a means to reduce the chances of pipeline damage. A major study discusses damage prevention issues for pipelines.[6] Some outside force damages cause an immediate failure. However, sometimes damage occurs but a failure does not occur until later, sometimes after years.

Other causes arise from manufacturing and assembly defects in pipe and other pipeline system components. Welds and other types of joints, fittings, valves, and other system components can fail and result in a product release.

INTEGRITY ASSESSMENT

Integrity assessment is an important preventive measure that can discover pipe wall conditions, which if left unremedied could lead to pipeline failure. Integrity assessment has been defined in recent years through both industry consensus standards and government regulation as comprising inspection and testing followed by evaluation of a pipeline using internal line inspection, hydrostatic or pressure testing, and direct assessment.[4,7,8]

Internal line inspection uses instrumented in-line devices (smart pigs or intelligent pigs) that traverse the interior of a line being inspected. They are used to find anomalies in pipe wall thickness caused by corrosion or externally caused impact damage. Electronically collected data are processed to provide results interpreted by trained experts that reveal the condition of the line. It can be used for both gas and liquid lines. Technology development continues to move in the direction of developing methods for smaller line diameters and for higher resolution. More information on pigging can be found in specialized publications.[9,10]

In hydrostatic pressure testing, the pipeline is filled with water and pressurized to a specified test pressure that exceeds the operating pressure of the pipeline. If there are latent defects that are weak, a failure occurs. If failure does not occur then the line has been shown to be safe at normal operating pressure.

Direct assessment follows a prescribed process found in industry standards.[11] Direct assessment visually examines the pipeline at predetermined locations where there is a cause to believe that some corrosion or outside force damage has occurred.

TECHNICAL FACTORS AFFECTING SAFETY

The causes of failure are associated with attributes of a pipeline system that either increase or decrease the likelihood that the causes will be realized. A full comprehension of technical factors related to these attributes requires consideration of all parts of the system. The system comprises not only the line pipe, but also the pump stations for liquid lines and compressor stations for gas lines, controls, measurement, and storage facilities. After the initial compressor or pump station, at the gas or liquid supply source, periodic booster compressor or pump stations may be found along a pipeline route to repressurize the line to overcome the line's pressure drop and maintain flow.

Operating pressure is one variable affecting the potential for failure and its consequences. The pressure profile of system depends on the product, product flow rate, pipeline diameter, distance, and terrain elevation profiles. Gathering and transmission pipeline systems operate at a pressure of many hundreds and even over a thousand pounds per square inch of pressure. Gas distribution systems operate at lower pressures, even reaching a few inches of water gage pressure, at the customer end. The safety consequences of a pipeline system failure depend on the product, pressure, size of system, part of system, and location of the failure. Pipeline safety practices recognize these differences as well as features common to all systems. At the highest level, pipeline safety comprises managing the risks of failure in terms of prevention and mitigation. Prevention measures apply to reducing the likelihood of a failure and mitigation measures to reduce the severity of consequences. Regulations, industry consensus standards, and sound industry practices all target various factors that have a bearing on the likelihood or severity of a pipeline failure and product release.

LIKELIHOOD FACTORS

Likelihood factors are associated with the system attributes associated with the individual causes of failure. Likelihood factors include such considerations as pipe coating, degree of cathodic protection, soil conditions, pipe age, and maintenance history for corrosion control and depth of burial surrounding land use and activities, one-call system effectiveness, effectiveness of system marker signs, and other factors for third-party damage potential, to cite some examples.[12]

There are variations of such lists found throughout the literature, but the differences are usually at the conceptual level at which causes are defined and differences in terminology rather than fundamental differences in understanding the causes.

The causes are influenced by the specific attributes of a pipeline system. Good discussions of these be found in the technical literature. As with the causes, the system attributes can also be defined at different levels of resolution and different terms are sometimes used by different experts. Fundamentally, however, the major relationships between attributes and causes are commonly understood. Ongoing research continues to refine current knowledge about the relationships and provides a basis for ongoing improvements in pipeline safety practices, whether technical or procedural.

The safety performance of pipeline systems, as determined from an analysis of the history of pipeline

incidents, reveals that corrosion and outside force damage are the significant causes of failures for the pipeline and contribute to failures also for other parts of the overall system. The exception is for plastic piping systems, where the corrosion mechanism is absent.

Pipeline safety management is fundamentally establishing relationships between the causes of failure and pipeline system attributes, then controlling the variables associated with those attributes to within prescribed limits. By controlling the variables associated with the causes, failures are prevented and safety goals based on prevention are achieved. A second component of pipeline safety is recognizing consequences and locating systems to minimize adverse consequences in the event of failure and providing appropriate emergency preparedness and response capabilities in the event of failure. Pipeline safety regulations establish the minimum standards governing pipeline safety. Industry standards and sound industry practice will at least meet these requirements and will in some cases exceed the minimum safety standards according to specific needs on individual pipeline systems and parts of those systems.

The body of engineering and management practices that result from regulations and standards cover the entire lifecycle of a pipeline system, including design, construction, operations, and maintenance. In the United States, federal statutes and associated regulations govern gas and hazardous liquid pipelines. The federal regulations apply to all pipelines, within certain definitions, and the jurisdiction for pipeline safety through regulation rests with either the federal government or the state governments depending on whether an interstate or intrastate pipeline is involved. Federal pipeline regulations are developed and administered by the U.S. Department of Transportation, Pipeline and Hazardous Materials Safety Administration (PHMSA), Office of Pipeline Safety (OPS). The regulations are found in Title 49, Code of Federal Regulations, Parts 190 through 199.[13] Part 192 applies specifically to gas pipelines and Part 195 to hazardous liquid pipelines. There are periodic updates and some regulatory initiatives are always under way. The regulations are developed under statutory authority provided by the Congress. The statutes for pipeline safety are periodically modified. Some major new requirements were added for pipeline integrity management in the Pipeline Safety Improvement Act of 2002.[14]

Industry consensus standards are sometimes incorporated into the regulations by reference and become a regulatory requirement. Examples of some these standards incorporated by reference in the regulations for gas pipelines are listed in Table 1. The regulations and standards establish an implied level of safety based on adherence to the specified requirements. For intrastate pipelines, under the enforcement jurisdiction of the individual states, state regulations can be more stringent, but not less stringent, than the federal regulations.

Besides regulatory development and enforcement, other OPS functions include pipeline safety data analysis based on data collected by OPS through annual and incident reports from the industry and from OPS inspections of pipeline systems, sponsoring of research, and training.

Two other federal authorities, the National Transportation Safety Board (NTSB) and the General Accounting Office (GAO), also have a role. The NTSB investigates reportable pipeline incidents to determine the causes and to make recommendations to the OPS for future prevention. The GAO audits the OPS for performance against its mission and makes recommendations for improvements, if needed. A report, issued by GAO in 2000, reviewed progress on pipeline safety issues by the OPS.[15]

The remainder of this entry discusses the role of industry standards and regulations in pipeline safety, technical factors, and other issues related to pipeline safety.

As noted above, safety issues are associated with the potential for significant environmental impacts for liquid lines, as well as fire or explosion. For gas lines, the primary issue is fire and explosion. In both cases, safety depends on sound design, construction, operations, and maintenance. Standards for such practices are found in the industrial technical literature and include various designs and operating standards of organizations, such as the American Petroleum Institute (API), the American Society of Mechanical Engineers (ASME), and the American Society of Testing and Materials (ASTM), and the National Association of Corrosion Engineers (NACE), among others.[16–19] Regulations incorporate some of these and other standards by reference, as shown in Table 1, and add additional requirements targeting specific risk factors that affect safety. Similar standards are incorporated into hazardous liquid regulations.

Important topics covered by design standards are materials of construction, sizing pipe for wall thickness and operating pressure, and various design factors that provide a specific safety margin. They also deal with corrosion prevention and protection. Construction standards deal with welding, joints, field fabrication, and assembly, structural support depth of burial, and other topics. For example, one of API's standards covers pipeline welding.[20] There are also operating and maintenance standards. For example, there is a standard for determining the remaining strength of pipe when there is some corrosion, to decide on continuing service, replacement, or repair.[21]

The source of first-hand experience with the pipeline systems is the industry. There is an ongoing effort

Table 1 Examples of standards incorporated by reference under the requirements of 49 CFR part 192

Organization	Standard
AGA	AGA Pipeline Research Committee, Project PR-3-805, ''A Modified Criterion for Evaluating the Remaining Strength of Corroded Pipe''
API	API Specification 5L ''Specification for Line Pipe'' API Recommended Practice 5L1 ''Recommended Practice for Railroad Transportation of Line Pipe'' API Specification 6D ''Specification for Pipeline Valves (Gate, Plug, Ball, and Check Valves)'' API Standard 1104 ''Welding of Pipelines and Related Facilities''
ASTM	ASTM Designation: A 53 ''Standard Specification for Pipe, Steel, Black and Hot-Dipped, Zinc-Coated, Welded and Seamless'' ASTM Designation A 106 ''Standard Specification for Seamless Carbon Steel Pipe for High-Temperature Service'' ASTM Designation: A 333/A 333M ''Standard Specification for Seamless and Welded Steel Pipe for Low-Temperature Service'' ASTM Designation: A 372/A 372M ''Standard Specification for Carbon and Alloy Steel Forgings for Thin-Walled Pressure Vessels'' ASTM Designation: A 381 ''Standard Specification for Metal-Arc-Welded Steel Pipe for Use with High-Pressure Transmission Systems'' ASTM Designation: A 671 ''Standard Specification for Electric-Fusion-Welded Steel Pipe for Atmospheric and Lower Temperatures'' ASTM Designation: A 672 ''Standard Specification for Electric-Fusion-Welded Steel Pipe for High-Pressure Service at Moderate Temperatures'' ASTM Designation A 691 ''Standard Specification for Carbon and Alloy Steel Pipe, Electric-Fusion-Welded for High-Pressure Service at High Temperatures'' ASTM Designation D638 ''Standard Test Method for Tensile Properties of Plastics'' ASTM Designation D2513 Standard Specification for Thermoplastic Gas Pressure Pipe, Tubing and Fittings'' ASTM Designation D 2517 ''Standard Specification for Reinforced Epoxy Resin Gas Pressure Pipe and Fittings'' ASTM Designation: F1055 ''Standard Specification for Electrofusion Type Polyethylene Fittings for Outside Diameter Controlled Polyethylene Pipe and Tubing''
ASME	ASME/ANSI B16.1 ''Cast Iron Pipe Flanges and Flanged Fittings'' ASME/ANSI B16.5 ''Pipe Flanges and Flanged Fittings''[3] ASME/ANSI B31G ''Manual for Determining the Remaining Strength of Corroded Pipelines'' ASME/ANSI B31.8 ''Gas Transmission and Distribution Piping Systems'' ASME Boiler and Pressure Vessel Code, Section I ''Power Boilers'' ASME Boiler and Pressure Vessel Code, Section VIII, Division 1 ''Pressure Vessels'' ASME Boiler and Pressure Vessel Code, Section VIII, Division 2 ''Pressure Vessels: Alternative Rules'' ASME Boiler and Pressure Vessel Code, Section IX ''Welding and Brazing Qualifications''
Manufacturers Standardization Society of the Valve and Fittings Industry, Inc. (MSS)	MSS SP44-96 ''Steel Pipe Line Flanges''
National Fire Protection Association (NFPA)	NFPA 30 ''Flammable and Combustible Liquids Code'' ANSI/NFPA 58 ''Standard for the Storage and Handling of Liquefied Petroleum Gases'' ANSI/NFPA 59 Standard for the Storage and Handling of Liquefied Petroleum Gases at Utility Gas Plants ANSI/NFPA 70 ''National Electrical Code''

between the OPS, industry, the state regulatory agencies, and outside stakeholders to better understand the causes of pipeline incident failure and modify risk control measures, that is, prevention and mitigation measures to reduce safety risks. These efforts are supported by industry organizations, which include, but are not limited to, the:

- American Gas Association (AGA)
- Association of Oil Pipelines (AOPL)
- American Petroleum Institute (API)
- American Society of Civil Engineers (ASCE)
- American Society of Mechanical Engineers (ASME)
- American National Standards Institute (ANSI)
- Interstate Natural Gas Pipeline Association of America (INGAA)
- National Association of Corrosion Engineers (NACE).

There are other specialized organizations and numerous contractors and vendors doing research to develop better materials, design methods, inspection and testing methods, and other practices to enhance the safety of pipelines. Most operators have multiple staff involved in various technical committees associated with these organizations, and various key technical people carry certifications of various types. For example, key corrosion personnel typically carry various certifications from NACE.

In the regulations, factors are addressed by specific requirements. As an example, some of the requirements for the operation and maintenance of natural gas pipelines under 49 CFR 192 Subparts L and M are shown in Table 2.

CONSEQUENCE FACTORS

The safety consequences of a pipeline failure depend on the product carried, pressure, diameter, time to shut off, distance between isolation points, and proximity to persons who could be affected. Historically, most of the harm from accidents has been to workers near the line. Some notable incidents have involved the public.

The product, conditions on the line, and conditions in the surroundings affect consequences. For example, high-pressure gas lines, which may fail by rupture, can result in a jet fire that is usually oriented vertically. The fire burns with intense heat and the primary threat is from the heat radiation within a certain distance from the fire. Flammable liquids can also burn by jet fires, but also offer the potential for a pool fire. The potential for a specific type of fire (or even a vapor explosion) varies with a number of conditions, including product properties, release conditions, and the presence of a suitable ignition source. There are releases of flammable materials for which ignition does not occur.

PIPELINE INTEGRITY MANAGEMENT FOR TRANSMISSION PIPELINES IN HIGH-CONSEQUENCE AREAS

New industry standards for pipeline integrity management began to appear in 1996.[22] New OPS regulations defining requirements for pipeline integrity management programs have been established since 2000 for both liquid and gas pipeline systems.[7,8] These standards and regulations have more explicitly defined requirements for integrity and, hence, safety management of pipelines. They define an overall process or management system, which explicitly calls for threat identification, risk-based evaluation, formal inspection, and testing referred to as pipeline integrity assessments, appropriate evaluation of the resulting data, identification and selection of prevention and mitigation measures, monitoring, and formal performance measurement based on defined performance measures. An important concept in these requirements is recognizing high-consequence areas (HCAs), based on the potential for safety and environmental impacts of a product release from a pipeline. High-consequence areas are defined in the regulations.

The purpose of formal pipeline integrity management is to better manage safety in HCAs. There are several types of HCAs and definitions differ between those designated for gas and liquid pipelines. An example for gas pipelines is:[7]

- All Class 3 and 4 locations (the definition of class locations for gas pipelines is already a part of the regulations and is based on differences in implied population density.) These are areas where there are at least 46 buildings intended for human occupancy or any buildings with four or more stories aboveground within 660 ft of the pipeline along any continuous mile of its length, respectively.
- A location not Class 3 or 4 that has an ''identified site'' within a potential impact circle of an impact radius calculated from a specified equation in the regulation.

Identified site is defined in the gas integrity management program (IMP) regulation. For liquid pipelines, HCAs include highly populated areas, other populated areas, navigable waterways, drinking water supplies, and ecologically sensitive areas.

Pipeline operators must implement an IMP, which includes periodic inspection, testing, and integration

Table 2 Example of some regulatory requirements for natural gas lines that address specific risk factors

Operation and maintenance activities	Regulatory citation
Procedural manual for operations, maintenance, and emergencies	192.605
Change in class location: required study	192.609
Change in class location: confirmation or revision of maximum allowable operating pressure	192.611
Underwater inspection and reburial of pipelines in the Gulf of Mexico and its inlets	192.612
Continuing surveillance	192.613
Damage prevention program	192.614
Emergency plans	192.615
Public education	192.616
Investigation of failures	192.617
Maximum allowable operating pressure	192.619, 192.621, 192.623
Odorization of gas	192.625
Tapping pipelines under pressure	192.627
Purging of pipelines	192.629
Replacement, repair, or removal from service of unsafe pipelines	192.703(b)
Prompt repair of hazardous leaks	192.703(c)
Patrolling	192.705, 192.721
Leak surveys	192.706, 192.723
Line markers	192.707
Record keeping	192.709
General requirements for repair procedures	192.711
Permanent field repair of imperfections and damages	192.713
Permanent field repair of welds	192.715
Permanent field repair of leaks	192.717
Testing of repairs	192.719
Test requirements for reinstating service lines	192.725
Abandonment or deactivation of facilities	192.727
Compressor stations: inspection and testing of relief devices	192.731
Compressor stations: storage of combustible materials	192.735
Compressor stations: gas detection	192.736
Pressure limiting and regulating stations: inspection and testing	192.739
Pressure limiting and regulating stations: telemetering or recording gauges (for gas pressure)	192.741
Pressure limiting and regulating stations: testing of relief devices	192.743
Valve maintenance	192.745, 192.747, and 192.749
Prevention of accidental ignition	192.751
Caulked bell and spigot joints	192.753
Protecting cast-iron pipelines	192.755
High-consequence areas	192.761

of information related to pipeline integrity into a risk-based process for future inspection and testing and for remedial actions of potentially unsafe conditions. The major elements of an IMP for gas pipelines, found in 49 CFR Part 192, are listed below.[7] There are similar or analogous elements for liquid lines.[8]

1. Identification of HCAs.
2. Baseline assessment plan.

3. Threats identification, which must include data integration, and risk assessment for setting priorities for integrity assessment and evaluating preventive and mitigative measures.
4. Direct assessment plan, if applicable.
5. Provision for remediating conditions found during integrity assessment.
6. Process for continual evaluation and assessment.
7. Plan for confirmatory direct assessment.
8. Provisions for adding preventive and mitigative measures to protect HCAs.
9. Performance plan as outlined in ASME/ANSI B31.8S, section 9, which includes performance measures.
10. Record keeping.
11. Management of change process according to ASME/ANSI B31.8S, section 11.
12. Quality assurance process according to ASME/ANSI B31.8S, section 12.
13. Communication plan, which includes elements of ASME/ANSI B31.8S, section 10 and procedures for addressing safety concerns raised by OPS or their state agency designee.
14. Procedures for providing (when requested), by electronic an other means, a copy of the operator's risk analysis or integrity management program to OPS or their state designee.
15. Procedures for ensuring that each integrity assessment is being conducted in a manner that minimizes environmental and safety risks.
16. Process for identification and assessment of newly defined HCAs.

RISK ASSESSMENT

Safety increases as risk decreases. Risk assessment is a tool that aids in understanding and making decisions about pipeline safety. Risk assessment is used in the pipeline industry as an aid in setting priorities for integrity assessment and for the application of other prevention and mitigation measures for pipeline integrity and safety. Risk assessment recognizes both the likelihood and the severity of consequences of failure through the classic equation:

$$\text{Risk} = \text{Likelihood} \times \text{Severity of consequences}$$

Common methods of risk assessment include relative risk numerical scoring methods and numerical probabilistic methods. With relative methods, scores and weighting factors are assigned to various risk factors and grouping of factors, based on system attributes, and an overall risk score is generated. The methods determine a score for likelihood of failure, the severity of the consequences of failure, and the overall risk. A pipeline is segmented and risks for each segment are determined to identify relative risk along a line. This aids in planning prevention and mitigation measures tailored to the differing levels of risk. Commercial software is available for some of the methods but individual operators often customize these packages for their specific systems. Operators also develop their own ranking systems. These systems vary in the way specific risk factors are accounted for and in the scoring methods, but contain many of the same factors.

REMEDIAL ACTIONS

Effective remedial actions are an important element in pipeline integrity and safety. They include repair, replacement, and relocation of pipe. Standard methods are used for repair of defects due to corrosion or outside force damage on pipelines. This includes line pipe, components, and pipe coatings. If the pipe is unsuited for repair, then replacement is made of appropriate lengths, which can range from a few feet to miles. In some cases, a solution to threats and potential adverse consequences is relocation. A new pipe is installed at a new location to deliver product between two specified points.

PREVENTION AND MITIGATION MEASURES

Integrity and safety involve application of prevention and mitigation measures to address the likelihood and consequence factors of risk, respectively. They can be classified broadly as engineering controls, people controls, and management controls.

Prevention and mitigation measures address the likelihood and severity of the consequences of failures, respectively. Under the pipeline integrity programs engendered by the new regulations, prevention and mitigation applied in HCAs might be greater than in non-HCA areas. For example, patrol frequency in a specific HCA could potentially be more frequent than in a non-HCA, or another HCA, if the threat of third-party damage to the system for that particular HCA suggested that an increased frequency would be an effective option for that particular locale in the system. A variety of enhanced prevention and mitigation measures could also be applied to other threats to pipeline integrity.

PERFORMANCE MONITORING AND MODIFICATION

Performance measurement is the way improvements in pipeline safety are determined. Changes in the

frequency and severity of incidents are a direct measure of performance.

This is a lagging measure rather than a leading measure. Leading measures are preferable in giving indications of progress or problems before a failure occurs. These include both technical measures and program and management measures. The latter deal with adherence to requirements defined for carrying out various prevention and mitigation activities. The former are associated with quantitative measures associated primarily with specific technical indicators of prevention. An example is the number or frequency of potential corrosion failures averted by the discovery and remediation of corrosion defects in the pipeline.

Applying lessons learned through the periodic performance evaluation leads to safety improvements.

SECURITY

As with other sectors of our society, the operating environment of the pipeline industry has been profoundly altered by September 11, 2001. Security has been elevated to a priority that previously had not been required when the major threat was an occasional and usually minor act of vandalism. In most cases, the threat with gas pipelines may be economic disruption more than public or environmental safety for most incident locations. The lines are buried through most of their runs and there may be more attractive surface targets. However, the threat is real and is being addressed. It is beyond the scope of this entry to elaborate in detail, but enhanced security measures are being evaluated and implemented throughout the industry.

CONCLUSIONS

Pipeline safety is associated with concerns about the effects of accidental release of fuels and chemicals transported under relatively high pressures in pipelines. There are a number of causes for pipeline failure. Increased understanding is leading to measures that are more specific to control specific causes of failure. The federal government has a significant role in pipeline safety through significant regulatory authority over pipelines. Industry and government have numerous institutions and mechanisms in place for meeting the mutual interest of preventing (and mitigating) pipeline incidents and maintaining pipeline safety. Among changes in the industry and regulatory establishments are more rigorous safety regulations, such as integrity management regulations, which comprise specific elements intended to improve safety when implemented by pipeline operation.

REFERENCES

1. http://ops.dot.gov/stats/po.htm (accessed Dec 2004).
2. http://ops.dot.gov/stats/GTANNUAL2.HTM (accessed Dec 2004).
3. Vieth, P.H.; Mesloh, R.E.; Kiefner, J.H. *Analysis of DOT Reportable Incidents for Gas Transmission and Gathering Pipelines—January 1, 1985 through December 31, 1994*; American Gas Association: Washington, DC, 1996; Catalog No. L51745.
4. American Society of Mechanical Engineers. *Managing System Integrity of Gas Pipelines*, *ASME Code for Pressure Piping*, *B31*; ASME B31.8S-2001 (Supplement to ASME B31.8); American Society of Mechanical Engineers: New York, 2002.
5. Fields, R.J.; Pugh, E.N.; Read, D.T.; Smith, J.H. *Assessment of the Performance and Reliability of Older ERW (Electric Resistance Welding) Pipelines*; U.S. Department of Commerce, National Technical Information Service: Springfield, VA, 1989.
6. U.S. Department of Transportation. *Common Ground: Study of One-Call Systems and Damage Prevention Best Practices*; Office of Pipeline Safety (OPS), Research and Special Programs Administration, U.S. Department of Transportation: Washington, DC; 20, 690.
7. Implementing Integrity Management—Final Rule (as amended), May 26, 2004, Pipeline Integrity Management in High Consequence Areas (Gas Transmission Pipelines), 49 CFR Part 192–Subpart O; http://primis.rsap.dot.gov/gasimp/docs/FinalRuleAmended_gas_full.pdf (accessed Oct 2004).
8. Pipeline safety pipeline integrity management in high consequence areas (hazardous liquid pipeline operators with 500 or more miles of pipeline). In *Federal Register*, 49 CFR Part 195; Department of Transportation, Research, and Special Projects Administration: Vol. 65 (232) 75, 378, Dec 1, 2000.
9. Pigging Products and Services Association. *An Introduction to Pipeline Pigging*; Gulf Publishing Company: Houston, TX, Feb 1995.
10. Tiratsoo, J.N.H., Ed. *Pipeline Pigging Technology*, 2nd Ed.; Gulf Publishing Company: Houston, TX, 1992.
11. NACE RP0502 Pipeline External Corrosion Direct Assessment Methodology; http://www.nace.org/product.asp?=SKU=51300%2DSG&ProdName=04187+External+Corrosion+Direct+Assessment (accessed Dec 2004).

12. Muhlbauer, W.K. *Pipeline Risk Management Manual*, 2nd Ed.; Gulf Publishing Company: Houston, TX, 1996.
13. Oct 1, 2003 Code of Federal Regulations, Parts 190–199; http://ops.dot.gov/regindex.htm (accessed Nov 2004).
14. Pipeline Safety Act of 2002; Public Law 107–355; Dec 17, 2002; http://ops.dot.gov/Pub_Law/107_cong_public_laws.pdf (accessed Nov 2004).
15. General Accounting Office (GAO). *The Office of Pipeline Safety is Changing, How it Oversees the Pipeline Industry*; United States General Accounting Office; Pipeline Safety, GAO/RCED-00-128, May 2000.
16. http://api.org (accessed Oct 2004).
17. http://asme.org (accessed Oct 2004).
18. http://astm.org (accessed Oct 2004).
19. http://www.NACE.org (accessed Dec 2004).
20. American Petroleum Institute. *Welding of Pipelines and Related Facilities*, 19th; API Standard 1104; American Petroleum Institute: Washington, DC, 1999.
21. American Society of Mechanical Engineers. *B31G-1991. Manual for Determining the Remaining Strength of Corroded Pipelines*, a supplement to B31, Code for Pressure Piping; American Society of Mechanical Engineers: Washington, DC, 1991; ASME B31G.
22. American Petroleum Institute. *Assurance of Hazardous Liquid Pipeline System Integrity*, 1st Ed.; American Petroleum Institute: Washington, DC, Aug 1996; API RP 1129.

Plant Metabolic Engineering

Eleanore T. Wurtzel
Department of Biological Sciences, Lehman College, The City University of New York, Bronx, New York, U.S.A.

Erich Grotewold
Department of Plant Cellular and Molecular Biology and Plant Biotechnology Center, The Ohio State University, Columbus, Ohio, U.S.A.

INTRODUCTION

The advent of recombinant DNA technologies over the past 20–30 yr has resulted in an impressive emergence of metabolic engineering as a powerful complement to classical chemical synthesis for the production of specialized pharmaceuticals and industrial compounds. The simplicity and low cost of growing plants makes them ideal systems for metabolic engineering. Here, we intend to provide a perspective of the advantages of plants for metabolic engineering, highlighting the significant hurdles that still need to be overcome. This entry begins by introducing plant metabolic engineering and by providing an overview of the main systems available, whole plants or plant cells in culture. We then describe the different approaches on hand for increasing flux through existing metabolic pathways and end by providing a perspective of the future of plants as chemical factories for the production of industrial materials and pharmaceuticals, or for removing pollutants from the environment.

PLANTS: THE NATURAL FACTORIES

Plant biotechnology is rapidly gaining momentum as a serious alternative to other biotechnological approaches for the production of fine chemicals, pharmaceuticals, and industrial products, for the cleaning of the environment through bioremediation while creating novel opportunities in agriculture. Plants are very attractive for biotechnological applications in large part because of the cheap availability of the main requirements for efficient plant growth, energy (sunlight), and carbon (CO_2). Man has also known how to grow, harvest, and process plant tissues since the beginning of modern agricultural practice. Plants contain an amazing diversity of genes encoding enzymes for thousands of distinct biochemical reactions and this natural genetic diversity dwarfs growing efforts in combinatorial chemistry to increase the available space of chemical entities. As a distinct advantage over synthetic chemical transformations and in contrast with microbes, plant cells are highly compartmentalized, facilitating, for example, the separation of substrates from the products. Thus, plants provide amazing opportunities for their use as natural factories.

PLANT METABOLIC ENGINEERING

In 1991, Bailey defined metabolic engineering as ''the improvement of cellular activities by manipulation of enzymatic, transport, and regulatory functions of the cell with the use of recombinant DNA technologies.''[1] Today, the availability of the complete genome sequence for several plants, together with the development of powerful techniques for the transformation and stable or transient expression of genes in plants brings plant metabolic engineering as a strong alternative to classical chemical synthesis for the production of pharmaceuticals and other important industrial compounds. Plant metabolic engineering involves the manipulation of existing metabolic pathways by either increasing or diverting flux to desired or from undesired products, respectively, or the generation of chemical entities not normally found in the plant production system (cells or whole plants, see next section) through the introduction of genes from other organisms. Essential elements in the toolbox of the metabolic engineer are mechanisms to eliminate or overexpress gene activity.

Strategies available to eliminate the activity of specific enzymes in a pathway involve one of several possible approaches:

1. Identification of a mutant gene for the corresponding enzyme. Traditionally, plant breeding has taken advantage of natural allelic diversity to introduce mutant alleles into the genetic background of interest. Alternatively, screens have been conducted to identify mutants with desired phenotypes. Transposons have provided powerful tools, facilitating the identification of mutants in any gene of interest through what is

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039300

currently known as ''reverse genetics.'' More recently, targeting-induced local lesions in genome (TILLING) have permitted the identification of single-nucleotide mutations in any gene in a randomly mutagenized plant population.[2]

4. Knocking out gene function by targeted RNA degradation. Double-stranded RNA interference (dsRNAi) today provides probably the preferred method to knock out gene function.[3] An emerging alternative to this is the use of RNase P-mediated RNA degradation.[4] Both of these approaches are dependent on the availability of methods for introducing and expressing RNA in the host plant.
5. Interfering with protein function using specific inhibitors or antibodies. There are many protein inhibitors of metabolic enzymes that, when overexpressed, could have the potential to inhibit specific enzymatic steps (e.g., Ref.[5]). Nonprotein inhibitors of metabolic enzymes are also extensively used, resulting in some potent plant herbicides for which resistance can be easily manipulated, or in the formation of new compounds, when nonessential pathways are inhibited.[6,7]

The expression of genes in plants or plant cells is dependent on several factors that include:

1. A method to introduce genes into the plant. Such methods are available for a number of plants. While *Agrobacterium*-mediated transformation provides a facile method and usually results in a low number of integration events, minimizing problems of cosuppression of endogenous genes, introduction of genes by particle bombardment has none of the host plant limitations imposed by the host preference of *Agrobacterium* and has been used in a broad range of plants.[8]
2. Promoters to direct gene expression in the appropriate spatial and temporal landscape. A large number of plant promoters capable of directing robust gene expression in almost every plant tissue have been identified.[9,10] An emerging alternative to using natural promoters is the generation of ''artificial'' or ''synthetic'' promoters that would provide the desired expression patterns.[11]
3. A source for the gene encoding the enzyme of interest. The completion (or near completion) of the sequencing of various plant genomes together with the availability of a large number of expressed sequence tags (ESTs) for plants where complete genome sequencing is impractical provide an enormous opportunity for the selection of genes encoding plant metabolic enzymes.

CHOICE OF PLANT SYSTEM FOR METABOLIC ENGINEERING

Whole Plants

Compared to animals, cells in culture (plant or animal), or microbes, whole plants provide a cheap and simple platform for large-scale mass production. Transgenic plants can be grown, maintained, and harvested using equipment already available from classical agricultural practices. The choice of a production plant for metabolic engineering depends on the specific metabolite to be produced and whether the necessary precursors are already present. Specialized organs (e.g., leaf hairs, glands, or trichomes) can be used both to sequester compounds and to provide for an accessible source for metabolite extraction.[12] There is a growing concern in using crop species for the production of phytochemicals, particularly pharmaceuticals, in areas where those crops are grown for food consumption purposes. Thus, plant species that do not hybridize with major crops or with wild relatives are gaining importance for the purposes of genetic manipulation.

Plant Cells in Culture

Cell culture offers the advantage of growth in minimal space without requirement for greenhouses and allows for development of uniform conditions to optimize phytochemical production. While cell culture may not always mimic the biochemical profile of cells as in the growing plant, it is possible to manipulate culture conditions, such that the appropriate array of genes are expressed and metabolites are biosynthesized and accumulated (e.g., treatment with elicitors and other stress inducers).[13] Cell lines may be transformed with genes encoding biosynthetic enzymes or regulatory factors and assessed for modification of metabolite profiles. Given that conditions are established for maximizing accumulation of the desired chemical constituents, plant cell culture is a useful alternative to growing differentiated plant tissues. In addition, cell culture may provide a platform to test genes that may later be utilized for the more time-consuming production of transgenic plants. For large-scale production of metabolites, whether they represent novel drugs, or unusual metabolites needed for performing nutritional studies, cell cultures are a good alternative to harvesting metabolites from rare plant species or from those refractory to transgenic manipulation.

Using Plant Genes in Microorganisms

Microorganisms (e.g., the gram-negative bacterium *Escherichia coli*) possess a biochemical profile similar

to plant plastids, such as chloroplasts. Because of the biochemical similarity between the microorganism and the plant plastid, the results obtained by genetic modification can be applied later to testing in plants. The advantage of using the microorganism initially is the savings in time as compared to production of a transgenic plant. Examples of this approach include application to manipulation of the carotenoid and isoprenoid biosynthetic pathways. This heterologous system has been used to test the function of plant transgenes encoding putative enzymes that utilize substrates produced by the bacterium.[14,15] Alternatively, *E. coli* cells have been used to screen for genes that may positively or negatively influence pathway flux.[16] These organisms have also proven to be expedient tools for testing potential strategies for manipulating pathway flux for pathways that operate in plant plastids.[17]

STRATEGIES FOR INCREASING FLUX OF EXISTING PATHWAYS

Manipulating the Activity of "Rate Limiting" Steps

Classically, flux constraints have largely been attributed to the presence of one or a few "rate limiting" or bottleneck enzymes in a pathway. The flux theory put forward by Kacser and Burns, and supported by elegant experimentation, questions the existence of single rate limiting enzymatic steps in biochemical pathways.[18,19] These studies provided the theory behind metabolic control analysis (MCA), which furnishes a tool to interpret why experiments aimed at manipulating metabolic pathways fail, largely a consequence of the distribution of the control of flux over multiple enzymes in a pathway, instead of one as the "rate limiting step" concept would suggest.[20] Nevertheless, several examples are available in the literature in which the expression of one or a few genes results in a significant increase in flux. For example, the expression of the chalcone isomerase enzyme in tomatoes resulted in a significant increase in the accumulation of flavonols, which are important nutraceutical components of the human diet.[21] Similarly, the overexpression of phytoene synthase in canola seeds dramatically increased flux into the carotenoid pathway.[22] Increasing the metabolic flux of upstream pathways that feed precursors for the synthesis of the desired compounds provides an additional strategy for engineering metabolism, with a good example provided by the increase of isoprenoid precursor for the manipulation of carotenoids.[23]

The experimental success in increasing metabolic flux by expressing one or a few key biosynthetic enzymes contrasts with the emerging notion that enzymes in metabolic pathways are often associated in large macromolecular complexes, or metabolons.[24] The most extreme example of a metabolon is the formation of a "metabolic channel" in which pathway intermediates are directly transferred between catalytic sites without diffusion. Clearly, if channeling is a general characteristic in metabolic pathways, the perturbation of the correct stoichiometry of the enzymes in a channel, for example, by overexpression, could have results opposite to what was expected. To what extent the formation of metabolons contributes to the distribution of the overall flux in a pathway over multiple enzymes, as suggested by MCA, remains to be established.

Metabolic Flux Analysis and Modeling

Metabolic flux is the rate at which the material is processed through a specific pathway. A general misconception is that an increase in flux through a pathway will result in the increased accumulation of the pathway products. This assumption is wrong because it fails to consider that pathway products can be the substrates for other chemical reactions. A case example is provided by the biosynthesis of amino acids. While the overall accumulation of amino acids can be increased by enhancing flux through their biosynthetic pathways, the accumulation of the free amino acids (i.e., those not incorporated into proteins) may not be affected. Thus, when investigating flux with the ultimate objective of modeling metabolic pathways, one needs to take into account the ultimate fate of the metabolites. The classical mass-balance approach requires measurements of the rates of substrate uptake, rates of metabolite secretion, and rates of biomass formation. Together with the estimation of metabolic pools and with knowledge of the metabolic networks involved, a fairly accurate model of flux through a pathway can be obtained. However, many plant metabolic pathways involve reversible reactions, multiple compartments, and parallel competing pathways that significantly complicate the mass-balance approach.[25] These complications can in part be solved by stable-isotope labeling under steady-state conditions, resulting in what is known as stable-isotope metabolic flux analysis, which has been applied to modeling flux though a number of plant metabolic pathways.[25]

Expression of Multiple Genes in Plants: Progress and Limitations

One way to overcome the limitations imposed by the flux theory is to overexpress multiple genes encoding enzymes of a pathway. There are significant limitations to introducing multiple genes in a transgenic plant and ensuring that they will all be expressed at the high

desired level, as exemplified by the attempts to increase the levels of monoterpenoid alkaloid biosynthesis by coexpressing tryptophan decarboxylase and strictosidine synthase.[26] In these studies, the analysis of a large number of transgenic tobacco plants showed a dramatic variation in the levels of expression of these two genes, largely a consequence of silencing and cosuppression.[26] While part of the problems encountered could be solved by minimizing the sequence homology between the transgenes and carefully controlling transgene copy number, significant hurdles prevent the ''stacking'' of genes in transgenic plants by conventional methods. Sequential sexual crossing of transgenic plants has permitted the stacking of up to three transgenes, and significant progress has been recently made in cotransformation techniques that permit the simultaneous insertion of multiple genes.[27,28] Emerging approaches to solve the problems associated with stacking transgenes include the expression of multiple genes from a single promoter creating translational fusions of multiple enzymes, which can then be cleaved into separate enzymes by the inclusion of convenient protease cleavage sites.[29] An alternative to the use of protease sites is the synthesis of a transcript, resembling a polycistronic mRNA, encoding multiple enzymes in a pathway. By having internal ribosome entry sites between the different subunits, translation of each enzyme can be achieved.[30] However, neither of these promising approaches has yet been utilized for the successful engineering of a plant metabolic pathway.

Diverting Flux Using Loss of Function Approaches

While the overexpression of metabolic enzymes provides a powerful tool in metabolic engineering, it is often as important to minimize flux through a pathway that results in an undesired product. As described before, several approaches are currently available for the downregulation or knockout of a specific enzymatic step. RNA interference (RNAi) is indeed emerging as a powerful complement to more classic loss of function methods.[31] But often, the most impressive pathway manipulation results are obtained when gain and loss-of-function approaches are combined to increase flux through one pathway while decreasing flux through a competing pathway. This combinatorial approach has proven to be very successful in modifying multiple lignin traits in aspen trees.[32]

Targeting Entire Pathways with Transcription Factors

According to MCA, increasing the speed of a rate limiting step in a biochemical pathway is often insufficient for a significant increase in the overall flux in the pathway. Metabolic pathways are often controlled by one or a few regulatory proteins that bind to the promoters of the genes encoding the enzymes in a pathway and activate their transcription. Thus, transcription factors provide an attractive alternative to the manipulation of single or multiple enzymes for increasing flux. Because of the conspicuous pigmentation provided by anthocyanin pigments, the flavonoid pathway provides one of the best-studied cases of control of a plant metabolic pathway.[33] Anthocyanin accumulation is regulated in many plant species by the concerted action of transcription factors corresponding to the MYB and bHLH families.[34] Not surprisingly, the regulators of anthocyanin biosynthesis have been among the first ones used to demonstrate the potential of transcription factors to activate entire pathways.[35] As examples of the successful application of this approach, the maize P1 (MYB family member) and C1 and R transcription factors have been used to activate separate branches of flavonoid biosynthesis in maize cultured cells, the ORCA3 (AP2 family member) to manipulate the accumulation of alkaloids in *Catharantus roseus* cultured cells, and the maize C1 and R to increase the accumulation of flavonols in tomato.[36–38] Transcription factors provide a powerful complement to other approaches to manipulate the accumulation of desired phytochemicals. As an example, a chimeric form of the maize C1 and R regulators of anthocyanin biosynthesis was combined with the silencing of a structural gene in the flavonol/anthocyanin pathway, flavanone 3-hydroxylase, to increase the content of phytoestrogenic isoflavones in soybean seeds.[39]

The identification of novel plant repressor domains, such as the EAR motif, expands the use of transcription factors to not only activate entire metabolic pathways, but also to inhibit them.[40] In addition to using natural transcription factors, an emerging approach is to artificially create transcription factors, primarily from the Zn-finger family, with known DNA-binding specificities.[41] Thus, the use of artificial DNA-binding domains fused to either transcriptional activation or repression motifs is likely to provide a powerful future tool for altering flux through entire metabolic pathways using a single transgene.

MAKING NEW COMPOUNDS IN PLANTS

Plants as a Source of Industrial Materials

Plants continue to be the unique source of a number of very important industrial materials that include wood, cotton, cork, and rubber. However, the advent of petroleum as a source of carbon for the chemical industry

has resulted in a significant reduction in the use of plants as the feedstock for organic industrial chemicals. In 1930, 30% of the organic chemicals were derived from plants, down to 1% in 1960.[42] Is there a future for metabolic engineering to revive the role of plants as a fundamental source of industrial materials? This question can only be answered by taking into account the economics involved in growing and harvesting the plants, extracting the desired compound, and separating it from other chemicals. Previous analyses highlighted the challenges associated with this.[43] One of the best-described examples of how plants could be used to manipulate industrial products is the formation of polyhydroxyalkanoates (PHA) in plants. These polyesters of 3-hydroxyacids have unique biodegradable and elastomeric properties, which make them highly desirable in a number of applications, such as the medical industry, and for making environmental friendly plastics. Polyhydroxyalkanoates accumulate in numerous bacteria and fungi for carbon storage, and the genes from bacteria were successfully introduced into the plastids of *Arabidopsis*, resulting in a significant accumulation of polymer.[27] As of today, the accumulation of PHAs of various compositions and chain lengths has been accomplished in a number of crop plants, yet significant hurdles need to be overcome to compete in price with cheaper plastics with similar properties obtained by chemical synthesis.[44]

Another plant polymer with multiple industrial applications, which is the subject of several efforts to modify its accumulation and properties, is starch. Starch is a polymer of two glucan polymers, the mainly linear amylose and the branched amylopectin. Depending on the proportion of amylose to amylopectin, and the length and branching of the chains, starches can have very different properties. Currently, a large number of enzymes are known in maize and other starch-producing plants that influence these fundamental properties of starches.[45]

Plants as Pharmaceutical Factories

Plant parts have a widespread use in traditional medicine (such as Indian ayurvedic medicine and Chinese herbal remedies), providing clues to bioactive principles with novel health and nutritional benefits that will contribute to resources for metabolic engineering. The small fraction of plants so far surveyed (<15% of terrestrial plants) has revealed an astonishing potential for the biosynthesis of small (<1000 Da) molecules, with over 100,000 of these phytochemicals already described.[46] These phytochemicals accumulate largely as a consequence of biotic and abiotic interactions of plants with their environment, and are mainly derived from what has been classically known as secondary metabolism. In addition to the important role that these compounds play as nutraceutical components of animal diet, many of these metabolites are bioactive, providing the basis for a large fraction of all the pharmaceuticals currently available in the market. Thus, a major thrust in plant metabolic engineering continues to be to exploit their huge biochemical diversity to make new compounds from existing ones or to increase the levels of compounds with biomedical or nutraceutical significance. Future genomic investigations of medicinal plants will yield novel enzymes and genes to further expand the repertoire available for metabolic engineering of either endogenous pathways or transfer to heterologous species for enhanced accumulation of novel drugs or lead compounds for further drug development, or production of novel enzymes for bioprocessing.

The exploitation of medicinal plants for metabolic engineering of undiscovered metabolites and metabolic pathways is still at its infancy and is likely to require the initial investment of molecular markers to provide plant genotype fingerprints followed by bioactivity guided fractionation to identify the target molecule(s) and further basic studies (precursor feeding, bioinformatics approaches to identify genes) to elucidate the relevant biosynthetic pathways.

Plants as Nutraceutical Factories

Modification of plants by metabolic engineering can have a profound impact on human health. For example, the grass family (Poaceae) members are the most important food crops worldwide, and include maize, wheat, barley, sorghum, pearl millet, and rice. The endosperm tissues of these crops serve as major food staples; unfortunately, they are deficient in nutritionally essential carotenoids. Endosperms of these food crops are generally low in provitamin A (1–10%) relative to nonprovitamin A carotenoids.[47] For example, the consumption of carotenoid-poor cereal crops is associated with vitamin A deficiency affecting 250 million children in developing countries.[48] Improved vitamin A nutrition could eliminate approximately 1.3–2.5 million annual deaths.[49] Carotenoids also have important health benefits in developed countries. In humans and animals, various carotenoids derived from plant sources act as antioxidants, protect against diseases such as cancer, heart, and eye diseases, and are important in vision, while other carotenoids are precursors to vitamin A and to retinoid compounds involved in development.[50–56] An alternative approach to alleviating vitamin A deficiency worldwide is to improve levels of provitamin A carotenoids in food staples such as corn, wheat, and rice by metabolic engineering. Preliminary success with metabolic engineering of the pathway in rice, tomato, tobacco, and canola points

to the potential of this approach.[57–60] Unexpected products in these transgenic plants, however, suggest that the technology is limited by current deficiencies in understanding of endogenous gene expression. Despite the preliminary successes using laboratory-optimized experimental transgenic materials, future integration of the pathway in local varieties will also entail pyramiding of multiple traits, a challenge that can be addressed by continuing research on pathway interactions and competition for common substrates.

A similar success was recently reported in the manipulation of folates (including tetrahydrofolate) in tomato fruits. Tetrahydrofolate is an essential cofactor for several one-carbon reactions including the DNA biosynthesis cycle and the methylation cycle, and hence, folate deficiency results in anemia. Diaz de la Garza and coworkers have recently shown that increasing up to 100-fold the biosynthesis of pteridine, a precursor for folate biosynthesis, resulted in a two-fold increase in the folate content in tomato fruit.[61] In this study, the investigators utilized a mammalian GTP cyclohydrolase I to bypass the negative feedback regulation featured in regulation of the corresponding plant enzyme.

Using Plant Compartments for Chemical Sequestration

Massive accumulation of metabolites may sometimes be problematic if transgenes are expressed constitutively throughout the plant. Several strategies can be used to overcome this problem. One approach is through the use of tissue-specific promoters, allowing for accumulation in either a specific organ or tissue. Alternatively, gene expression can be controlled in such a way that biosynthetic enzymes or metabolites are directed to specific cell compartments such as the vacuole or chloroplast. For the purpose of chloroplast accumulation, either the plastid genome itself may be engineered with the desired transgene(s) or nuclear-encoded transgenes may be outfitted with chloroplast-targeting signals. Powerful new plastid transformation methods have been developed over the past few years.[62] The expression of metabolic enzymes in plastids provides several advantages over the nuclear expression. Among these advantages is the higher level of expression possible, given the large number of plastids that a plant cell can harbor, and the impact on biosafety, given that in most crops, plastids are under maternal plastid inheritance, minimizing the risks associated with gene flow from pollen derived from genetically modified plants.

In addition to targeting metabolic pathways to specific organelles, it is conceivable that pathway intermediates or final products could be sequestered in specific subcellular compartments. This could have a significant impact on metabolite production by removing toxic compounds and increasing flux through the displacement of equilibrium. However, the challenge that remains is our ignorance of the possible pathways by which phytochemicals traffic within or between cells. It is well established that many phytochemicals are sequestered in the vacuole by modifications that include glutathione conjugations followed by the action of specific transporters, particularly from the ABC family.[63–65] However, it is becoming evident that, in addition to transporters, phytochemicals can traffic between compartments using specialized vesicles, minimizing the risks of undesired chemical reactions and decreasing their toxicity.[66,67] It is evident that once we learn more on the trafficking of phytochemicals, enormous opportunities will become available to manipulate these processes as an additional tool in the toolbox of the metabolic engineer.

EMERGING TECHNOLOGIES

Plant Diversity as a Source of New Genes

In the course of plant evolution, certain plant lineages have evolved unique biosynthetic pathways that offer opportunities to use genomic approaches to capture this biochemical diversity and provide tools for manipulating plants and other organisms that do not produce the compounds or where accumulation is blocked at a particular biosynthetic step. Even in a single species, such as maize, where many diverse lines are available, allelic variation offers opportunities to use molecular tools to direct breeding efforts using molecular probes for these particular alleles. For example, using association mapping, Buckler's group has identified specific candidate genes associated with starch accumulation and these results may be applied to metabolic engineering of starch accumulation by directing breeding efforts using molecular probes representing these novel alleles.[68]

Gene Shuffling and Directed Enzyme Evolution

Directed alteration of enzyme activity and substrate specificity can broaden the repertoire of metabolites produced in plants and other organisms. Error-prone DNA polymerases can be used for PCR amplification to generate mutations in plant genes, which are then expressed in bacteria or in other organisms that can be transformed with high efficiency, which are then screened for accumulation of novel compounds.[69] Gene shuffling technologies can produce a large number of enzyme variants, by shuffling fragments

from an existing library.[70,71] While the applications of gene shuffling are multiple and include selecting for enzymes with higher stability, the recent identification of new glyphosate *N*-acetyltransferase enzymes that resulted in increased resistance of various plants to the herbicide glyphosate provides a good example of the power of the approach.[72,73]

Need for New Technology

With the growing interest in metabolomics, a field that provides the necessary data for metabolic engineering, current methodologies are limited in giving resolution of metabolic constitution in complex organisms having multiple cell types and suborganellar compartments. Therefore, advancements in the fields of metabolomics and metabolic engineering will require new types of analytical technology that provide 3-D resolution overlaid with temporal variation. The future availability of such tools will have a far-reaching impact extending from plant metabolic engineering to medical research and diagnostics. Another area in which significant improvement is needed is in the integration of chemistry and biology, for example, to better establish how small molecules bind to proteins in enzyme–substrate or ligand–receptor relationships.

CONCLUSIONS

Despite the significant advances in the field of plant metabolic engineering, many of them described here, predicting the outcome of a metabolic engineering strategy (predictive metabolic engineering) remains a significant challenge.[74] This highlights the urgent need to continue to study fundamental aspects of plant metabolism including trafficking of small molecules and how those molecules interfere with fundamental cellular processes.

ACKNOWLEDGMENTS

We apologize to the many authors whom we have not referred to in this entry. Research in plant metabolic engineering in the Wurtzel Lab is supported by NIH (grant #S06-GM08225) and New York State and in the Grotewold Lab, by grants from the National Science Foundation (MCB-0130062) and the U.S. Department of Agriculture (NRICGP 2002-01267).

REFERENCES

1. Bailey, J.E. Toward a science of metabolic engineering. Science **1991**, *252* (5013), 1668–1675.
2. Till, B.J.; Colbert, T.; Tompa, R.; Enns, L.; Codomo, C.; Johnson, J.; Reynolds, S.H.; Henikoff, J.G.; Green, E.A.; Steine, M.N.; Comai, L.; Henikoff, S. High-throughput TILLING for functional genomics. In *Plant Functional Genomics: Methods & Protocols*; Grotewold, E., Ed.; Humana Press: Totowa, NJ, 2003; 205–220.
3. Baulcombe, D. RNA silencing in plants. Nature **2004**, *431* (7006), 356–363.
4. Rangarajan, S.; Raj, M.L.; Hernandez, J.M.; Grotewold, E.; Gopalan, V. RNase P as a tool for disruption of gene expression in maize cells. Biochem. J. **2004**, *380* (Pt 3), 611–616.
5. Rausch, T.; Greiner, S. Plant protein inhibitors of invertases. Biochim. Biophys. Acta **2004**, *1696* (2), 253–261.
6. Darmency, H. Genetics of herbicide resistance in weeds and crops. In *Herbicide Resistance in Plants*; Powles, S.B., Holtum, J.A.M., Eds.; CRC Press: Boca Raton, FL, 1994; 263–297.
7. Forkmann, G.; Martens, S. Metabolic engineering and applications of flavonoids. Curr. Opin. Biotechnol. **2001**, *12* (2), 155–160.
8. Twyman, R.M.; Christou, P. Plant transformation technology: particle bombardment. In *Handbook of Plant Biotechnology*; Christou, P., Klee, H., Eds.; John Wiley & Sons, Ltd.: Chichester, U.K., 2004; 263–289.
9. De Buck, S.; Depicker, A. Gene expression and level of expression. In *Handbook of Plant Biotechnology*; Christou, P., Klee, H., Eds.; John Wiley & Sons, Ltd.: Chichester, U.K., 2004; 331–345.
10. Mueller, A.E.; Wassenegger, M. Control and silencing of transgene expression. In *Handbook of Plant Biotechnology*; Christou, P., Klee, H., Eds.; Chichester, U.K., 2004; 291–330.
11. Bhullar, S.; Chakravarthy, S.; Advani, S.; Datta, S.; Pental, D.; Burma, P.K. Strategies for development of functionally equivalent promoters with minimum sequence homology for transgene expression in plants: cis-elements in a novel DNA context versus domain swapping. Plant Physiol. **2003**, *132*, 988–998.
12. Wagner, G.J.; Wang, E.; Shepherd, R.W. New approaches for studying and exploiting an old protuberance, the plant trichome. Ann. Bot. (Lond.) **2004**, *93* (1), 3–11.
13. Radman, R.; Saez, T.; Bucke, C.; Keshavarz, T. Elicitation of plants and microbial cell systems. Biotechnol. Appl. Biochem. **2003**, *37* (Pt 1), 91–102.
14. Matthews, P.D.; Luo, R.; Wurtzel, E.T. Maize phytoene desaturase and zetacarotene desaturase catalyze a poly-Z desaturation pathway: implications for genetic engineering of carotenoid

content among cereal crops. J. Exp. Bot. **2003**, *54*, 2215–2230.
15. Li, Z.H.; Matthews, P.D.; Burr, B.; Wurtzel, E.T. Cloning and characterization of a maize cDNA encoding phytoene desaturase, an enzyme of the carotenoid biosynthetic pathway. Plant Mol. Biol. **1996**, *30*, 269–279.
16. Gallagher, C.E.; Cervantes-Cervantes, M.; Wurtzel, E.T. Surrogate biochemistry: use of *Escherichia coli* to identify plant cDNAs that impact metabolic engineering of carotenoid accumulation. Appl. Microbiol. Biotechnol. **2003**, *60*, 713–719.
17. Matthews, P.D.; Wurtzel, E.T. Metabolic engineering of carotenoid accumulation in Escherichia coli by modulation of the isoprenoid precursor pool with expression of deoxyxylulose phosphate synthase. Appl. Microbiol. Biotechnol. **2000**, *53* (4), 396–400.
18. Kacser, H.; Burns, J.A. The control of flux. Symp. Soc. Exp. Biol. **1973**, *27*, 65–104.
19. Niederberger, P.; Prasad, R.; Miozzari, G.; Kacser, H. A strategy for increasing an in vivo flux by genetic manipulations. Biochem. J. **1992**, *287*, 473–479.
20. Morandini, P.; Salamini, F. Plant biotechnology and breeding: allied for years to come. Trends Plant. Sci. **2003**, *8* (2), 70–75.
21. Muir, S.R.; Collins, G.J.; Robinson, S.; Hughes, S.; Bovy, A.; De Vos, C.H.R.; van Tunen, A.; Verhoeyen, M.E. Over expression of petunia chalcone isomerase in tomato results in fruit containing increased levels of flavonols. Nat. Biotechnol. **2001**, *19*, 470–474.
22. Ravanello, M.P.; Ke, D.; Alvarez, J.; Huang, B.; Shewmaker, C.K. Coordinate expression of multiple bacterial carotenoid genes in canola leading to altered carotenoid production. Metab. Eng. **2003**, *5* (4), 255–263.
23. Botella-Pavia, P.; Besumbes, O.; Phillips, M.A.; Carretero-Paulet, L.; Boronat, A.; Rodriguez-Concepcion, M. Regulation of carotenoid biosynthesis in plants: evidence for a key role of hydroxymethylbutenyl diphosphate reductase in controlling the supply of plastidial isoprenoid precursors. Plant J. **2004**, *40* (2), 188–199.
24. Winkel, B.S.J. Metabolic channeling in plants. Annu. Rev. Plant Biol. **2004**, *55*, 85–107.
25. Schwender, J.; Ohlrogge, J.; Shachar-Hill, Y. Understanding flux in plant metabolic networks. Curr. Opin. Plant Biol. **2004**, *7* (3), 309–317.
26. Leech, M.J.; May, K.; Hallard, D.; Verpoorte, R.; De Luca, V.; Christou, P. Expression of two consecutive genes of a secondary metabolic pathway in transgenic tobacco: molecular diversity influences levels of expression and product accumulation. Plant Mol. Biol. **1998**, *38*, 765–774.
27. Nawrath, C.; Poirier, Y.; Somerville, C. Targeting of the polyhydroxybutyrate biosynthetic pathway to the plastids of Arabidopsis thaliana results in high levels of polymer accumulation. Proc. Natl. Acad. Sci. USA **1994**, *91* (26), 12760–12764.
28. Halpin, C.; Boerjan, W. Stacking transgenes in forest trees. Trends Plant Sci. **2003**, *8* (8), 363–365.
29. Beck von Bodman, S.; Domier, L.L.; Farrand, S.K. Expression of multiple eukaryotic genes from a single promoter in *Nicotiana*. Biotechnology **1995**, *13*, 587–591.
30. Martinez-Salas, E. Internal ribosome entry site biology and its use in expression vectors. Curr. Opin. Biotechnol. **1999**, *10* (5), 458–464.
31. Tang, G.; Galili, G. Using RNAi to improve plant nutritional value: from mechanism to application. Trends Biotechnol. **2004**, *22* (9), 463–469.
32. Li, L.; Zhou, Y.; Cheng, X.; Sun, J.; Marita, J.M.; Ralph, J.; Chiang, V.L. Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. Proc. Natl. Acad. Sci. USA **2003**, *100* (8), 4939–4944.
33. Mol, J.; Grotewold, E.; Koes, R. How genes paint flowers and seeds. Trends Plant Sci. **1998**, *3* (6), 212–217.
34. Irani, N.G.; Hernandez, J.M.; Grotewold, E. Regulation of anthocyanin pigmentation. Recent Adv. Phytochem. **2003**, *38*, 59–78.
35. Goff, S.A.; Klein, T.M.; Roth, B.A.; Fromm, M.E.; Cone, K.C.; Radicella, J.P.; Chandler, V.L. Transactivation of anthocyanin biosynthetic genes following transfer of *B* regulatory genes into maize tissues. EMBO J. **1990**, *9*, 2517–2522.
36. Grotewold, E.; Chamberlin, M.; Snook, M.; Siame, B.; Butler, L.; Swenson, J.; Maddock, S.; Clair, G.S.; Bowen, B. Engineering secondary metabolism in maize cells by ectopic expression of transcription factors. Plant Cell **1998**, *10* (5), 721–740.
37. van der Fits, L.; Memelink, J. ORCA3, a jasmonate-responsive transcriptional regulator of plant primary and secondary metabolism. Science **2000**, *0*, 1–3.
38. Bovy, A.; de Vos, R.; Kemper, M.; Schijlen, E.; Almenar Pertejo, M.; Muir, S.; Collins, G.; Robinson, S.; Verhoeyen, M.; Hughes, S.; Santos-Buelga, C.; van Tunen, A. High-flavonol tomatoes resulting from the heterologous expression of the maize transcription factor genes LC and C1. Plant Cell **2002**, *14* (10), 2509–2526.
39. Yu, O.; Shi, J.; Hession, A.O.; Maxwell, C.A.; McGonigle, B.; Odell, J.T. Metabolic engineering to increase isoflavone biosynthesis in soybean seed. Phytochemistry **2003**, *63* (7), 753–763.

40. Hiratsu, K.; Matsui, K.; Koyama, T.; Ohme-Takagi, M. Dominant repression of target genes by chimeric repressors that include the EAR motif, a repression domain, in *Arabidopsis*. Plant J. **2003**, *34*, 733–739.
41. Jantz, D.; Amann, B.T.; Gatto, G.J. Jr.; Berg, J.M. The design of functional DNA-binding proteins based on zinc finger domains. Chem. Rev. **2004**, *104* (2), 789–799.
42. Ohlrogge, J. Plant metabolic engineering: are we ready for phase two? Curr. Opin. Plant Biol. **1999**, *2* (2), 121–122.
43. Hitz, B. Commentary: economic aspects of transgenic crops which produce novel products. Curr. Opin. Plant. Biol. **1999**, *2* (2), 135–138.
44. Snell, K.D.; Peoples, O.P. Polyhydroxyalkanoate polymers and their production in transgenic plants. Metab. Eng. **2002**, *4* (1), 29–40.
45. Fernie, A.R.; Willmitzer, L. Carbohydrate metabolism. In *Handbook of Plant Biotechnology*; Christou, P., Klee, H., Eds.; John Wiley & Sons, Ltd.: Chichester, U.K., 2004; 525–557.
46. Verpoorte, R. Pharmacognosy in the new millennium: leadfinding and biotechnology. J. Pharm. Pharmacol. **2000**, *52* (3), 253–262.
47. Graham, R. Wheat: Research at Waite Agricultural Research Institute in Australia. CGIAR Micronutrients Project 1997, Update No. 2.
48. Underwood, B.A.; Arthur, P. The contribution of vitamin A to public health. FASEB J. **1996**, *10* (9), 1040–1049.
49. Humphrey, J.; West, K. Jr.; Sommer, A. Vitamin A deficiency and attributable mortality among under-5-year-olds. Bull. World Health Organ. **1992**, *72* (2), 225–232.
50. van den Berg, H.; Faulks, R.; Granado, H.F.; Hirschberg, J.; Olmedilla, B.; Sandmann, G.; Southon, S.; Stahl, W. The potential for the improvement of carotenoid levels in foods and the likely systemic effects. J. Sci. Food Agric. **2000**, *80* (7), 880–912.
51. Lee, C.; McCoon, P.; LeBowitz, J. Vitamin A value of sweet corn. J. Agric. Food Chem. **1981**, *29* (6), 1294–1295.
52. Bendich, A.; Olson, J. Biological actions of carotenoids. FASEB J. **1989**, *3*, 1927–1932.
53. Giovannucci, E.; Ascherio, A.; Rimm, E.B.; Stampfer, M.J.; Colditz, G.A.; Willett, W.C. Intake of carotenoids and retinol in relation to risk of prostate cancer. J. Natl. Cancer Inst. **1995**, *87* (23), 1767–1776.
54. Sommerburg, O.; Keunen, J.E.; Bird, A.C.; van Kuijk, F.J. Fruits and vegetables that are sources for lutein and zeaxanthin: the macular pigment in human eyes. Br. J. Ophthalmol. **1998**, *82* (8), 907–910.
55. Kohlmeier, L.; Kark, J.D.; Gomez-Gracia, E.; Martin, B.C.; Steck, S.E.; Kardinaal, A.F.; Ringstad, J.; Thamm, M.; Masaev, V.; Riemersma, R.; Martin-Moreno, J.M.; Huttunen, J.K.; Kok, F.J. Lycopene and myocardial infarction risk in the EURAMIC Study. Am. J. Epidemiol. **1997**, *146* (8), 618–626.
56. Kiefer, C.; Hessel, S.; Lampert, J.M.; Vogt, K.; Lederer, M.O.; Breithaupt, D.E.; von Lintig, J. Identification and characterization of a mammalian enzyme catalyzing the asymmetric oxidative cleavage of provitamin A. J. Biol. Chem. **2001**, *276* (17), 14110–14116.
57. Ye, X.; Al-Babili, S.; Kloti, A.; Zhang, J.; Lucca, P.; Beyer, P.; Potrykus, I. Engineering the provitamin A (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. Science **2000**, *287* (5451), 303–305.
58. Shewmaker, C.K.; Sheehy, J.A.; Daley, M.; Colburn, S.; Ke, D.Y. Seed-specific overexpression of phytoene synthase: increase in carotenoids and other metabolic effects. Plant J. **1999**, *20* (4), 401–412.
59. Mann, V.; Harker, M.; Pecker, I.; Hirschberg, J. Metabolic engineering of astaxanthin production in tobacco flowers. Nat. Biotechnol. **2000**, *18* (8), 888–892.
60. Rosati, C.; Aquilani, R.; Dharmapuri, S.; Pallara, P.; Marusic, C.; Tavazza, R.; Bouvier, F.; Camara, B.; Giuliano, G. Metabolic engineering of beta-carotene and lycopene content in tomato fruit. Plant J. **2000**, *24* (3), 413–419.
61. Diaz de la Garza, R.; Quinlivan, E.P.; Klaus, S.M.; Basset, G.J.; Gregory, J.F. III; Hanson, A.D. Folate biofortification in tomatoes by engineering the pteridine branch of folate synthesis. Proc. Natl. Acad. Sci. USA **2004**, *101* (38), 13720–13725.
62. van Bel, A.J.; Hibberd, J.; Prufer, D.; Knoblauch, M. Novel approach in plastid transformation. Curr. Opin. Biotechnol. **2001**, *12* (2), 144–149.
63. Edwards, R.; Dixon, D.P.; Walbot, V. Plant glutathione *S*-transferases: enzymes with multiple functions in sickness and in health. Trends Plant Sci. **2000**, *5* (5), 193–198.
64. Theodoulou, F.L. Plant ABC transporters. Biochem. Biophys. **2000**, *2000*, 79–103.
65. Rea, P.A.; Li, Z.-S.; Lu, Y.-P.; Drozdowicz, Y.M. From vacuolar GS-X pumps to multispecific ABC transporters. Ann. Rev. Plant Phys. **1998**, *49*, 727–760.
66. Lin, Y.; Irani, N.G.; Grotewold, E. Sub-cellular trafficking of phytochemicals using auto-fluorescent compounds in maize cells. BMC Plant Biol. **2003**, *3* (3), 10.
67. Grotewold, E. The challenges of moving chemicals within and out of cells: insights into the

transport of plant natural products. Planta **2004**, *219*, 906–909.
68. Wilson, L.M.; Whitt, S.R.; Ibanez, A.M.; Rocheford, T.R.; Goodman, M.M.; Buckler, E.S.T. Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell **2004**, *16* (10), 2719–2733.
69. Schmidt-Dannert, C.; Umeno, D.; Arnold, F. Molecular breeding of carotenoid biosynthetic pathways. Nat. Biotechnol. **2000**, *18* (7), 750–753.
70. Stemmer, W.P.C. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. Proc. Natl. Sci. USA **1994**, *91*, 10747–10751.
71. Jestin, J.L.; Kaminski, P.A. Directed enzyme evolution and selections for catalysis based on product formation. J. Biotechnol. **2004**, *113* (1–3), 85–103.
72. Eijsink, V.G.; Bjork, A.; Gaseidnes, S.; Sirevag, R.; Synstad, B.; van den Burg, B.; Vriend, G. Rational engineering of enzyme stability. J. Biotechnol. **2004**, *113* (1–3), 105–120.
73. Castle, L.A.; Siehl, D.L.; Gorton, R.; Patten, P.A.; Chen, Y.H.; Bertain, S.; Cho, H.J.; Duck, N.; Wong, J.; Liu, D.; Lassner, M.W. Discovery and directed evolution of a glyphosate tolerance gene. Science **2004**, *304* (5674), 1151–1154.
74. Sweetlove, L.J.; Last, R.L.; Fernie, A.R. Predictive metabolic engineering: a goal for systems biology. Plant Physiol. **2003**, *132*, 420–425.

Plasma Etching

David G. Retzloff
Department of Chemical Engineering, University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

A gas becomes a plasma when sufficient energy results in a significant number of gas molecules losing some or all of their outer shell electrons and become positively charged. The resulting state is an assembly of ions and electrons that are not bound together and can move more or less freely. The electrically charged gas is called an ionized gas. Sir William Crookes first studied this state of matter in 1879. Irvine Langumir was the first person to use the term plasma to describe this state of matter and it is now currently viewed as a fourth state of matter distinct from the well-known gas, liquid, and solid states of matter because of its special physical and electrical properties. The most commonly known forms of plasmas are fluorescent lamps, the sun and most star systems, lightning, the solar wind, the ionosphere around earth, and the Aurora Borealis to name a few.

The importance of plasma in plasma etching is due to the ''free'' electrons that are present. These electrons can be accelerated to very high energies in the presence of electromagnetic fields provided the gas is maintained at a sufficiently low pressure (vacuum) to minimize collisions with the gas and charged molecules. These energetic electrons can ''react'' with the gas molecules to produce chemical species (frequently halides such as F or Cl) capable of etching semiconductor materials such as silicon, gallium arsenide, and aluminum in an etching reaction. Liquid acids containing halides (such as HF and HCl) have already been known to chemically etch semiconductor materials. One of the advantages in using a plasma environment was the elimination of the liquid acid waste from the etching process. The second advantage of plasma etching is the directionality that is obtained in the etching process. Etching occurs principally in the direction normal to the surface of the semiconductor wafer. This directionality is due to the applied external electric field, which is normal to the wafer surface. The directionality occurs because the electrons and ions move along the electric field lines when colliding with the wafer surface, causing the surface to be reactive while the side walls that form remain relatively nonreactive. Etch directionality is practically impossible to achieve with liquid acid etching and is the principle reason that plasma etching is used almost exclusively in defining the various active regions in semiconductor processing. Photoresist techniques are used to define the regions on the wafer that are to be etched and the photoresist itself protects the areas on the wafer that are not to be etched.

HISTORY

Acid etching traces its origins back to the times of the Pharos in Egypt. It was used in medieval times to etch glass, swords, and other metallic objects. It was also used in the early days of manufacturing semiconductor solid-state devices. At that time, silicon was the principal material etched to define regions for active devices. It was common to etch with potassium hydroxide or hydrofluoric acid. Metal contacts were etched with various acids, for example aluminum contacts were etched with an acid mixture based on phosphoric and nitric acids. To make this procedure usable in semiconductor manufacturing, the surface that was to remain unetched had to be covered with a coating to protect it from the acid etch. The surface of the material being etched had to be exposed to the etching species. In medieval times, this was accomplished by covering the protected areas with a waxy coating. Such a technique, of course, would not be viable for the small structures needed in semiconductor manufacturing, so a new procedure called photolithography was developed to provide the spatial resolutions required. Photolithography uses a combination of photography and lithography. The surface of the part that is to be etched is covered with a material called photoresist. This is an organic polymer that becomes soluble in a developer after exposure to light. A mask that contains the image of the features to be etched is made. This mask is similar in function to the negative in photography. The mask is placed in the proximity of the surface to be etched. Light of the required wavelength is then focused on the mask and the image on the mask is transferred to the photoresist. The part with the photoresist coating is then placed in a developing solution to remove the resist that was exposed to the light, resulting in a pattern on the part that has exposed features that are being etched and covered (or protected)

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120026495

features that are not etched. This part is then placed in an etching bath to remove material by etching it away. Acid etching has no preferential direction for etching; as a result the feature being etched would be larger than the feature defined on the photoresist due to lateral etching. This is termed undercut. The time of etching and the experimentally determined etch rate were used to determine the required size of the image on the mask to obtain the desired feature size and depth on the part when etching was complete. In a production environment requiring precise control of feature size, this method has many drawbacks not the least of which is that the composition and strength of the etching solution change with time resulting in a varying etch rate. This was the state of affairs until 1979 when Winters and Coburn[1,2] studied the etching of silicon with XeF_2 as a source of atomic fluorine in the presence of an argon plasma. They observed that the silicon etch rate was very small for the Ar^+ ion beam alone and XeF_2 alone. Both of these measured etch rates were less than 6 Å/min. However, when both XeF_2 and the argon plasma were present, the etch rate rose to 55 Å/min. They quickly discovered that the plasma caused XeF_2 to dissociate to form fluorine that subsequently etched the silicon. Researchers studying plasma etching[1–7] observed that the etching was directional in the sense that almost all of the etching was in the direction of the DC component of the electromagnetic field at the surface of the wafer, which was perpendicular to the surface of the wafer. This observation led to precise pattern transfers from the mask to the wafer and ultimately opened the way for the development of submicron features on the surface of the wafer that are utilized in current manufacturing of VLSI semiconductor devices such as CPUs and dynamic memory. The most common and dominant use of plasma etching is in the production of semiconductor devices.

THE PLASMA REACTOR

In almost all commercial applications and certainly in the semiconductor industry, a plasma reactor at the basic level consists of two electrodes in a parallel plate or psuedoparallel plate configuration reminiscent of a parallel plate capacitor. Currently for advanced semiconductor designs, a single wafer parallel plate plasma reactor is used to achieve the maximum uniformity across the wafer during plasma etching. A simplified representation of this type of reactor is shown in Fig. 1. It consists of two electrodes in a parallel plate configuration located in a vacuum chamber. The electrodes are connected to an RF-power supply via an impedance matching network. The purpose of this tuning network is to maximize the transmitted power (and therefore minimize the reflected power) to the electromagnetic field that exists between the electrodes. The typical frequency for the RF-power supply is 13.56 MHz, which is dictated by the Federal Communications Commission so that there will not be any interference with the FM radio stations broadcasting in the area. A DC-power source can be used but is usually unsuitable when dielectric films may form at the wafer surface, which can extinguish a DC plasma discharge. Microwave power sources can also be used but these are found in special applications where there is a need for high-energy electrons in the plasma etch process. The RF-power generator is the most common power source in use. The other common reactor configuration is a hexagonal center electrode surrounded by a second electrode, which also serves as the outer wall of the vacuum chamber. This configuration in its most elementary form is depicted in Fig. 2. In this configuration, 24 wafers are processed simultaneously. The focus of the remainder of this presentation will be on the capacitively coupled parallel plate configuration with an RF-power supply for single wafer processing, as it is a common reactor configuration. However, advanced plasma etcher design for semiconductor processing now employs a second inductively coupled powered RF electrode to obtain a more uniform plasma environment across the wafer (see for example, http://www.oxfordplasma.de/technols/rie_icp.htm, http://cfd.plasmatrancetechno.com/, http://cfdplasma.trancetechno.com/ch2.pdf). Almost all of the material presented will apply to any plasma reactor configuration used for plasma etching. A vacuum environment for the plasma etcher is required so that the electromagnetic field can interact with the electrons present and accelerate them to very high velocities between collisions so that the electrons attain large translational energies sufficient to drive the required plasma reactions. There must be adequate time between collisions for the electrons to acquire the translational energy needed.

The feed gas to the reactor is composed of molecules that contain the etching species. Fluorine containing molecules such as CF_4, CF_3H, or SiF_6 are used to etch silicon. Molecules such as H_2 or O_2 may also be present in the feed gas to control the reaction pathways and will be subsequently discussed. This feed mixture is termed the recipe for the etching process. To etch aluminum or aluminum silicide, which are used to make the electrical contacts to the active device region, a chlorine containing gas such as BCl_3 is used.

THE PLASMA ETCHING PROCESS

The plasma etching process consists of two components. The first is the physics and chemistry that occur

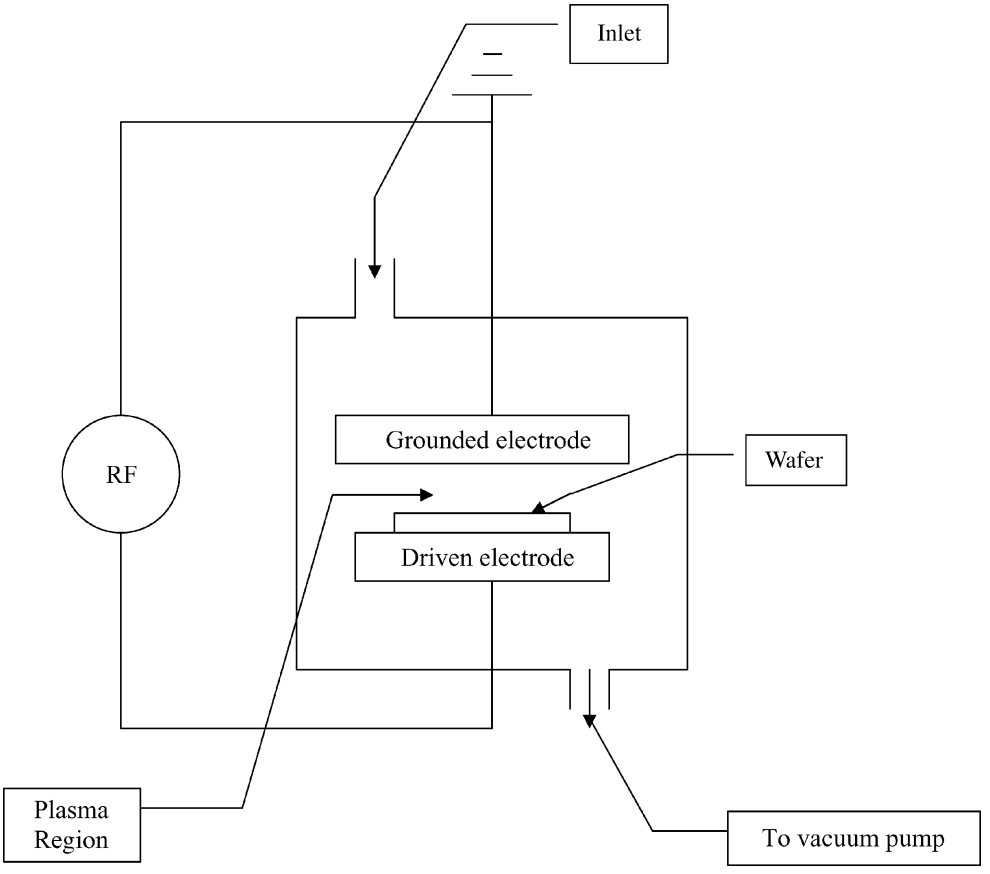


Fig. 1 Prototype single wafer parallel plate plasma etcher.

in the plasma, i.e., the physical volume between the two electrodes in the plasma reactor. The second is the physical and chemical processes that occur at the surface of the wafer (substrate). They will now be discussed in the order given above. For many commercial applications, the wafer is located on the driven electrode while the other electrode is connected to electrical ground. This configuration will be used for the purpose of discussion. The reverse configuration is also frequently found. The walls of the plasma reactor are normally grounded. To describe the etch process the space between the two electrodes is divided into three regions as depicted in Fig. 3. There is a positive space charge layer near the surface of the grounded and driven electrodes, which defines two of the three regions mentioned. There is also a positive space charge layer near the surface of any probe or wall that contacts the plasma. If one looks closely at these regions, they appear dark as the electrons in these regions have not been accelerated by the electromagnetic field to sufficiently high translational energy to achieve energy transfer to the neutral molecules by ''collision,'' which causes the outer shell electrons of the neutral molecules to make an excited state transition with an accompanying emission of radiation characteristic of that excited state as it relaxes to the ground state. The remainder of the space between the electrodes consists of a quasineutral region containing approximately equal numbers of positive and negative species and is termed the plasma. In this region there exist a large number of excited molecules, which relax back to the ground state with the emission of radiation causing the region to ''glow.'' Sometimes this region is called the glow discharge region. In this plasma region, the reaction chemistry necessary to produce the etching species occurs. A complete description of the chemistry involved is quite complex as can be seen in the work of Kushner[8] and Smolinsky and Flamm[9] for C_nF_m chemistry, which is used in the etching of silicon and silicon dioxide.

The plasma phase chemistry is most easily illustrated for a Freon 14 (CF_4) feed gas and provides the proto-type for understanding the processes in the plasma phase. The plasma chemistry must perform two tasks. First, it must generate the etching species and second, it must provide one or more additional electrons to sustain the plasma environment, as electrons are continually lost due to collisions with the walls of the reactor, recombination with positively charged ions, and flow out of the reactor towards

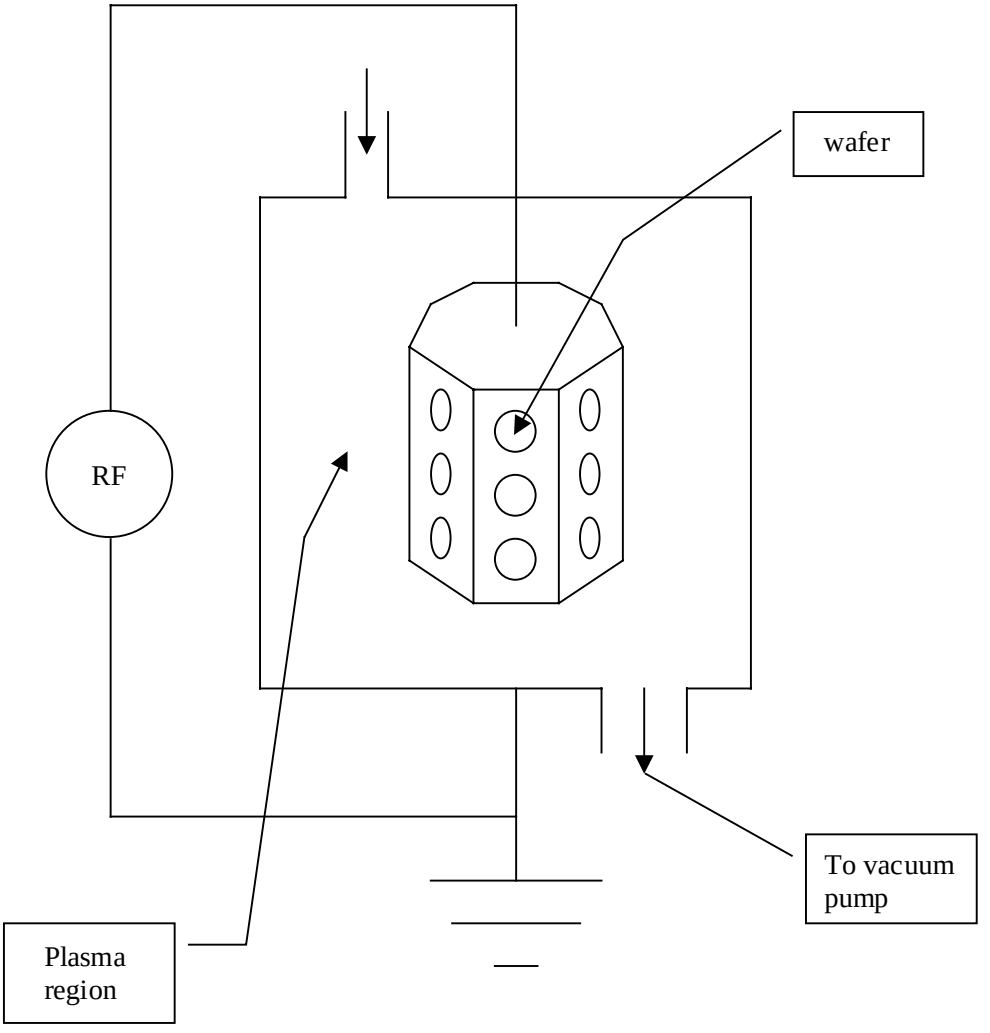


Fig. 2 Prototype hexode plasma etcher.

the vacuum pump. The following reaction occurs in the plasma:

$$CF_4 + e^- \rightarrow F + CF_3^+ + 2e^-$$

This is the simplest proto-type of the chemical processes that occur in the plasma phase. An overview of the basic plasma chemistry involved in a plasma etcher is given by Bell.[10] The characteristic glow of a plasma discharge is due to the emission or electromagnetic radiation due to the relaxation of an excited molecule to a ground or lower energy state. The glow process can be represented by a ''reaction'' of the following type:

$$CF_4 + e^- \rightarrow CF_4^* + e^-$$
$$CF_4^* \rightarrow CF_4 + hf$$

where f represents the frequency of emitted radiation and hf is the energy of the emitted radiation. This is a simplification of the processes that give rise to the glow because there are a number of molecular and atom species present in an industrial plasma that can be excited by collisions and relax back to a lower energy state by the emission of visible, higher frequency, and lower frequency radiations. The radiation field contains frequencies from all of these species as well. However, processes that result in higher frequency radiation are not significant in commercial etching.

For the plasma chemistry illustrated above, the chemistry at the surface of a silicon wafer is given by:

$$4F + Si \rightarrow SiF_4$$

This, of course, represents the overall chemical transformation that occurs and as such is a simplification of the detailed chemistry involved. A discussion

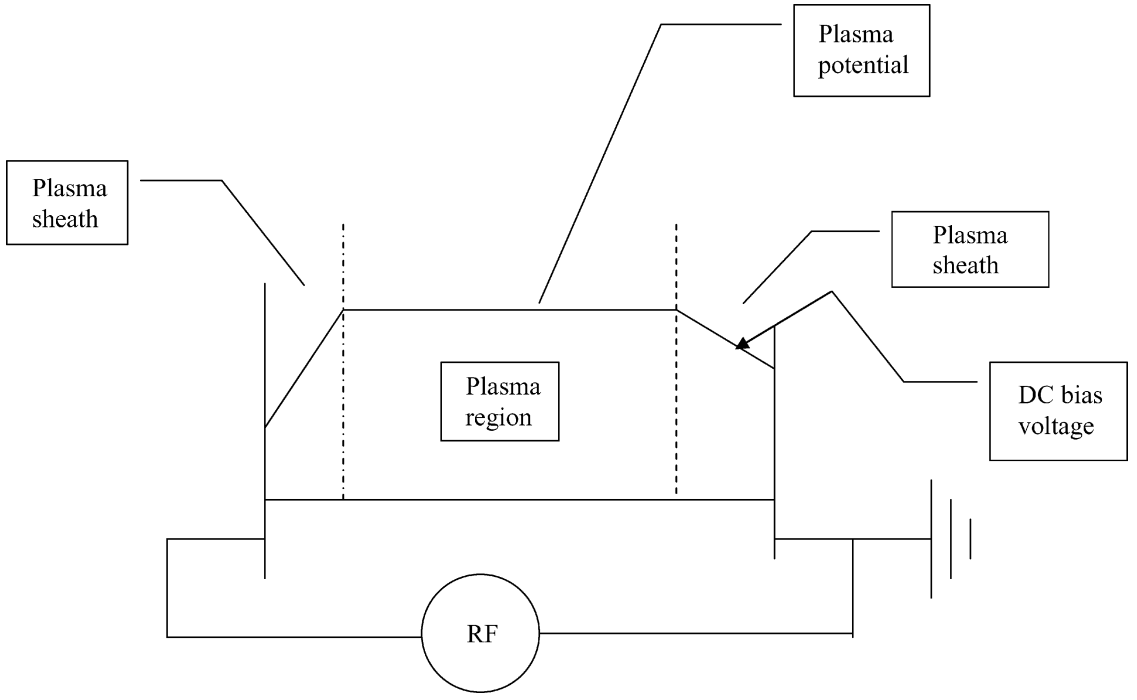


Fig. 3 DC potential, plasma, and sheath regions. The sheath region is a positive space charge region.

of the reaction mechanisms involved in the fluorine chemistry etching of silicon and silicon dioxide appears in.[2,11,12] The etch rate (reaction rate of silicon or silicon dioxide with fluorine at the surface of the wafer) for etching Si and SiO_2 has been measured and the possible chemical mechanisms that occur at the wafer surface discussed by Flamm and coworkers.[13] They report their results in terms of a standard Arrhenius form of a reaction rate with a first order dependence on the atomic fluorine concentration.

Two important observations have been made for etching SiO_2 and Si. First, the addition of O_2 to the feed increases the etch rates for both species dramatically and reaches a maximum at 12% O_2 for etching Si and 20% O_2 when etching SiO_2. However, for O_2 concentrations greater than 20% the etch rate for Si decreases much more rapidly than for SiO_2. Second, the addition of H_2 to the feed in amounts less than 40% does not effect the etch rate of SiO_2 but monotonically decreases the etch rate of Si to nearly zero at the 40% level.[14,15] Selectivities (this concept will be discussed later) for etching SiO_2 over Si of 40 : 1 are possible with CF_4/H_2. The increase in the etch rate of Si with the addition of O_2 is attributed to the following mechanism:

$$CF_4 + e^- \rightarrow F + CF_3^+ + 2e^-$$
$$CF_3^+ + O_2 + e^- \rightarrow 3F + CO_2$$
$$4F + Si \rightarrow SiF_4$$

The effects of H_2 on the Si an SiO_2 etch rates are due to the reactions:

$$CF_4 + e^- \rightarrow F + CF_3^+ + 2e^-$$
$$H_2 + 2F \rightarrow 2HF$$
$$4F + Si \rightarrow SiF_4$$

and

$$CF_4 + e^- \rightarrow F + CF_3^+ + 2e^-$$
$$4F + SiO_2 \rightarrow SiF_4 + 2O$$
$$H_2 + O \rightarrow H_2O$$

These mechanisms are grossly oversimplified and definitely not correct in detail but suggest the type of chemical competition that may occur.

A similar set of stoichiometric equations can be written for the etching of aluminum via chlorine chemistry, i.e.,

$$BCl_3 + e^- \rightarrow BCl_2^+ + Cl + 2e^-$$
$$3Cl + Al \rightarrow AlCl_3$$

See Tokunaga and Hess[16] for a discussion of aluminum etching using carbon tetrachloride.

Other materials that are normally etched in a microelectronics environment are poly-silicon, silicon nitride, doped poly-silicon, aluminum, gallium arsenide, and gold. The plasma chemistry involved is

analogous to that given previously, with the appropriate chemical species being present in the feed to provide the required etching species. A review of the chemicals and chemistry currently used in plasma etching for microelectronics applications appears in the work of Flamm and Donnelly.[17]

THEORETICAL CONSIDERATIONS IN PLASMA ETCHING

Most of the basic concepts in plasma etching are a result of the fact that the electron has an exceedingly small mass compared to atoms and molecules and the Lorentz force law for charged species in an electromagnetic field. The assumption of equilibrium or at least quasithermodynamic equilibrium plays a central role in our understanding of the plasma processes. This manifests itself in the use of the Maxwell-Boltzmann distribution to describe the energy distribution of the species in the plasma when describing particular physical phenomena at the elementary level in the plasma. This presentation begins with a review of several elementary physical principles. Charged species in an external electromagnetic field characterized by **E** and **B** fields experience a force called the Lorentz force given by $\mathbf{F} = q\mathbf{E} + q\mathbf{v} \times \mathbf{B}$ where q is the charge on the charged particle and **v** is the vector velocity of the charged particle. This causes the charged particle to accelerate. This accelerating charged particle then collides with other species present. The collisions tend to dissipate the translational energy of the charged particle and randomize the direction of motion. Thus, to accelerate the electrons to relative high translational energies to initiate plasma reactions, a vacuum environment is required to reduce the number of collisions per unit time and to allow an overall direction of motion to be established for the charged particles along the direction of the electric field. Typically, commercial plasma etching equipment operates in the high millitorr pressure range with an RF-power supply capable of up to 2 KW at a frequency of 13.56 MHz. It is an interesting aside that experimentally there clearly are sufficient electrons present to produce a plasma when the RP-power is first applied to the electrodes at low pressure as the plasma can be seen to glow. However, the two main mechanisms for producing such electrons, high field emission and thermionic emission, cannot account for the electron density necessary to produce the plasma. Thus, the true source of these initial electrons remains unknown.

Charge particles in an electromagnetic field move in a manner so as to produce an opposing induced electromagnetic field consistent with any constraints on the charged particles. This phenomenon is called shielding and also occurs in such areas as waveguides and ideal capacitors. Also, because electrons have such a small mass relative to the ions, they respond almost instantaneously to changes in the electromagnetic field while the positively charged ions to a first approximation remain stationary on the time scale of a 13.56 MHz RF-power supply. These two physical effects are responsible for the formation of an approximately equipotential region between the electrodes, which is approximately charge neutral on average, called the plasma, Debye shielding, the plasma sheath, the positive space charge, and the plasma potential. The equipotential region is a result of the shielding. Because the electrons have smaller mass, any surface at a fixed potential in contact with the plasma will initially experience a larger flux of electrons than positive ions, thus causing the surface to charge up negatively. This will continue until an opposing field builds up to oppose the motion of the electrons and results in equal fluxes of both electrons and positive ions. This is called Debye shielding and causes the potential of the surface to be lower than that of the plasma. The potential of the plasma is referred to as the plasma potential, V_p. This is true for all surfaces in contact with the plasma. Thus, the plasma potential represents the highest potential in the plasma etcher. If the surface is electrically floating with a potential V_f, the difference, $V_p - V_f$, is approximately 15 volts. The region between the plasma and any surface over which a potential difference occurs is called the plasma sheath. The width of the sheath is approximately the Debye length, which is on the order of 100 μm for typical plasmas. There is yet another feature of a plasma that occurs because of the very large mobility of the electrons relative to the ions. This is the self-bias induced in an RF-plasma. The RF generator produces a sinusoidal voltage waveform at the generator. However, the voltage waveform in the plasma etcher acquired a negative DC offset voltage. A high-voltage probe can measure this. This self-bias is on the order of half the applied RF peak-to-peak voltage. In addition, the sheath region acquires a net positive space charge. The relative sizes of the two electrodes in a plasma etcher directly influence the magnitude of the sheath voltage at these electrodes. It has been reported by Koenig[18] that the voltages at the two electrodes (labeled 1 and 2 in this formula) are given by

$$V_1/V_2 = A_2^4/A_1^4$$

This formula shows that the larger voltage sheath appears at the smaller electrode and is where the semiconductor wafer is normally located. Experimental measurements of these voltages suggest that the area ratio should be to a power smaller than 4 and greater than 1. One therefore expects sputtering to be greater

at the smaller electrode because sputtering at a surface is directly related to the magnitude of the voltage at that surface. Sputtering is the phenomenon of material ejection at a surface due to the collisions of atoms and molecules with this surface. All of these observations have been obtained by considering a simplified model that focuses on specific features of the processes that occur in a plasma.

A comprehensive model of the plasma etching environment is important for understanding how plasma etching works, improving the uniformity of the etch process, and in improving the design of a plasma etcher. Early efforts to model this process made an analogy with catalytic reactions for which a boundary layer formed at the surface due to the gas flow. The process was viewed in six steps: 1) A bulk flow of the feed gas (the etchant is produced in the bulk flow by chemical reaction) that forms a boundary layer at the wafer surface; 2) Diffusion of the etchant through the boundary layer to the wafer surface; 3) Adsorption of the etchant on the wafer surface; 4) Etchant reaction with the wafer surface; 5) Desorption of the reaction products from the wafer surface; and 6) Diffusion of the reaction products from the wafer surface to the bulk flow. This model evolved to a more fundamental description of a plasma etcher that started with kinetic theory and employed a Chapman-Enskog approach to obtain the continuum limit equations for mass, energy, and momentum. For an excellent treatment of this method, see the work of Jancel and Kahan.[19] These balance equations together with Maxwell's equations for the electromagnetic field provide a complete set of equations to describe the behavior of a plasma etcher. The Lorentz force law for charged particles contributes a term to the energy and momentum balances that is not usually present in most chemical engineering problems. There is a large literature on modeling a plasma etcher, which continues to increase every year. The general approach is illustrated in the work of several researchers.[20–29] Models of the etching process using the Boltzmann equations directly have also been proposed; however, for applications the continuum equations appear more suitable and thus a discussion of the Boltzmann equation approach to describing the etching process will not be pursued here. The interested reader may consult the work of Jensen and Graves,[20] Boeuf,[22] Tsai and Wu,[25] Kushner,[27,28] and Blaw et. al.[29] and the references therein for a discussion of these alternative models.

The models for a plasma etcher are usually one-dimensional (direction perpendicular to the electrodes in parallel plate geometry), use mobility limited description of the velocity field, use one energy equation that only involves the electron temperature, and write the electric potential in terms of Poisson's equation. This immediately eliminates spatial variations across the wafer, ignores temperature differences between the ions, neutrals, and the electrons, ignores the momentum exchange between the ions and the neutral molecules, and does not consider the fringing electromagnetic field that exists at the edge of the wafer when the wafer is approximately the same size as the electrode on which it resides. It is important to correctly predict the variation in the etch rate across the wafer as this has a direct influence on the final number of usable chips obtained from the wafer. Experimental measurements have established that in the plasma the electron temperature is on the order of 23,000 K, the ion temperature is 500 K, and the neutral temperature is 300 K. Inherent in these numbers is the assumption that the energy distribution for each species can be characterized by an equilibrium (Maxwell-Boltzmann) energy distribution, at least locally. This, of course, causes enormous fundamental problems with thermodynamic considerations. The concept of a local thermodynamic equilibrium has been advanced to address these problems but this concept has internal inconsistencies. The idea that there are species at distinctly different temperatures interacting and chemically reacting is fundamentally at odds with the concept of thermodynamic equilibrium. This also raises some interesting questions such as what does one write as an expression for the chemical rate of reaction? Specifically what is the energy dependence of the reaction rate expression? Is it Arrhenius in form? How does one treat species at different temperatures? Should one assign the concept of a phase to each species that has a characteristically different temperature, i.e., the electrons, the ions, and the neutral molecules? Of course, the phase in this context is not the same as the concept of gas, liquid, and solid. Rather, it is a conceptual way to describe the plasma by characterizing it as a collection of phases, each at a specific temperature, that interact both chemically and physically. The concept of temperature is retained, although it will require nonequilibrium thermodynamics to provide a suitable definition for temperature. The phases exchange energy and momentum and each phase moves with a characteristic vector velocity. The advantages of this approach are that it retains the usual concept of temperature and allows the energy balance for each phase to be written in terms of the temperature of that phase, the chemical reaction rates can be written down explicitly from the usual fundamental considerations, and the energy dependence of the chemical reaction rate is Arrhenius in form. The temperature that appears in this expression is a mass weight average of the temperature of the two species involved. This model reduces to models that have been previously used under appropriate assumptions. If we use the Freon-14 example and associate the electron with subscript 1, CF_3^+ with subscript 2, CF_4 with subscript 3,

and F with subscript 4, then we obtain the following model:

Mass Balance:

$$\partial n_1/\partial t + \nabla(n_1\mathbf{v}_1) = r_1$$
$$\partial n_2/\partial t + \nabla(n_2\mathbf{v}_2) = r_1$$
$$\partial n_3/\partial t + \nabla\cdot(n_3\mathbf{v}) = -r_1 + D_3\nabla^2 n_3$$
$$\partial n_4/\partial t + \nabla\cdot(n_4\mathbf{v}) = r_1 + D_4\nabla^2 n_4$$
$$\mathbf{v} = (\rho_3\mathbf{v}_3 + \rho_4\mathbf{v}_4)/(\rho_3 + \rho_4);$$
$$\rho_i = m_i n_i; \quad \rho = m_3 n_3 + m_4 n_4$$

Momentum Balance:

$$\rho_1(\partial\mathbf{v}_1/\partial t + \mathbf{v}_1\cdot\nabla\mathbf{v}_1) = \rho_1\mathbf{g} - n_1 e(\mathbf{E} + \mathbf{v}_1\times\mathbf{B}) + (\rho_1/\tau_1)(\mathbf{v}_2 - \mathbf{v}_1) + (\rho_1/\tau_1)(\mathbf{v} - \mathbf{v}_1)$$
$$\rho_2(\partial\mathbf{v}_2/\partial t + \mathbf{v}_2\cdot\nabla\mathbf{v}_2) = -\nabla p_2 + \rho_2\mathbf{g} - n_2 e(\mathbf{E} + \mathbf{v}_2\times\mathbf{B}) + (\rho_1/\tau_1)(\mathbf{v}_1 - \mathbf{v}_2) + (\rho_2/\tau_2)(\mathbf{v} - \mathbf{v}_2)$$
$$\rho(\partial\mathbf{v}/\partial t + \mathbf{v}\cdot\nabla\mathbf{v}) = -\nabla p + \rho\mathbf{g} + (\rho_1/\tau_3)(\mathbf{v}_1 - \mathbf{v}) + (\rho_2/\tau_2)(\mathbf{v}_2 - \mathbf{v}) + \nabla\cdot\tau$$
$$\nabla\cdot\tau = (\mu/3)\nabla(\nabla\cdot\mathbf{v}) + \mu\nabla^2\mathbf{v}$$

Energy Balance:

$$\rho_1(\partial U_1/\partial t + \mathbf{v}_1\cdot\nabla U_1) = -\nabla\cdot\mathbf{q}_1 + 2\rho_1(\delta_2/m_2)\nu_{12}(3k/3)(T_2 - T_1) + 2\rho_1(\delta_{34}/m_{34})\nu_{134}(3k/3)(T - T_1) - \mathbf{j}_1\cdot\mathbf{E}$$
$$\rho_2(\partial \mathrm{U}_2/\partial t + \mathbf{v}_2\cdot\nabla U_2) = -\nabla\cdot\mathbf{q}_2 - p_2\nabla\cdot\mathbf{v}_2 + 2\rho_2(\delta_2/m_2)\nu_{12}(3k/3)(T_1 - T_2) + 2\rho_2(\delta_{34}/m_{34})\nu_{134}(3k/3)(T - T_2) + \mathbf{j}_2\cdot\mathbf{E}$$
$$\rho(\partial U/\partial t + \mathbf{v}\cdot\nabla U) = -\nabla\cdot\mathbf{q} - p\nabla\cdot\mathbf{v} + 2\rho(\delta_2/m_{34})\nu_{134}(3k/3)(T_1 - T) + 2\rho(\delta_1/m_{34})\nu_{134}(3k/3)(T_2 - T)$$
$$U_i = (3/2)kT_i, \quad U = (3/2)kT_i,$$
$$\mathbf{q}_i = -k_i\nabla T_i, \quad \mathbf{q} = -\kappa\nabla T, \quad \mathbf{j}_i = en_i\mathbf{v}_i$$

Maxwell's Equations:

$$\nabla\cdot\mathbf{D} = 4\pi e(n_2 - n_1);$$
$$\nabla\times\mathbf{H} = (4\pi/c)\mathbf{j} + (1/c)\partial\mathbf{D}/\partial t;$$
$$\nabla\cdot\mathbf{B} = 0; \quad \nabla\times\mathbf{E} = (1/c)\partial\mathbf{B}/\partial t = 0;$$
$$\mathbf{j} = (\mathbf{j}_2 - \mathbf{j}_1)$$

The notation used is standard (see Ref.[19]) with the exceptions of n_i and r_1. The number density of species i is n_i and r_1 is the rate of appearance of number density for the electrons. The mass density is given by $n_i m_i$ with m_i being the mass of species i. Thus, the actual mass balance is obtained by multiplying each equation in the mass balance listed above by the mass of that species, m_i, which is constant.

These equations describe the plasma etcher. They must be supplemented by appropriate boundary conditions. Wafer etching occurs at the surface of the wafer and therefore appears as a boundary condition. Furthermore, because portions of the surface of the wafer are being removed during etching, this is a moving boundary value problem. Mathematical expressions for the reaction rate of etching can be found in the journal *Plasma Chemistry and Plasma Processing* for many materials as well as Refs.[13,30] for Si and SiO_2. It is clear that plasma etching is a difficult problem to formulate and solve. To adequately model the spatial variation in the etching at the surface of the wafer, the model must include the three spatial dimensions. One spatial dimension models will just not work.

We note that the radiation from the glow has not been considered in this model. To include this effect requires a mass balance for each excited species, a contribution in the momentum balances for each excited species, and additional terms in the energy balances that represent the energy lost via the radiation field. These effects are smaller than the terms that are retained in the model and thus represent "second order" effects. As such they are expected to change the solution to the problem slightly and have not been retained in the description.

To emphasize the complexity of the problem of modeling a plasma etcher, one more detail must be mentioned. Measurement of the temperatures of species in the plasma has been reported, which identify rotational-vibrational temperatures that are distinct from translational temperatures.[31,32] The second reference is significant in that it mentions the influence of rotational and vibrational energies (temperatures) on chemical reactivity. If one considers rotational and vibrational temperatures that are distinct from translational temperatures, then one is confronted with two questions. First, how is the reaction rate related to the rotational, vibrational, and translational temperatures? Second, how are the energy balances to be written for these distinct temperatures? It is obvious that there remain many fundamental challenges in developing a precise model of the plasma etcher. The value of such a model lies in its ability to guide the manufacturers of plasma etching equipment in the development of new designs to obtain better etching characteristics. It is also useful to the user in

identifying how to improve the current performance of a plasma etcher.

Fortunately, many improvements in plasma etcher designs and plasma etching applications can and are done through experimentation and development. This approach proceeds in concert with improvements in modeling and leads to the next generation designs for plasma etchers. Some fundamental considerations for plasma etcher performance are taken up in the next section.

PROCESS CONSIDERATIONS

The purpose of using a plasma etcher is to obtain well-defined patterns (features) on the surface of a substrate. Commercial applications of plasma etching occur almost exclusively in the microelectronics industry, in which case the substrate is a wafer. To achieve the desired resolution of the pattern being transferred to the wafer, the etching must occur in the vertical direction, i.e., perpendicular to the wafer in the areas not covered by the photoresist. There are two views of how this occurs. In the first case, both the ions and etching molecules are moving perpendicular to the wafer surface. In the case of the ions, this is due to the DC component of the electromagnetic field that exists in the sheath at the surface of the wafer and is perpendicular to the wafer surface. The etching species may be a neutral atom in which case it also is moving perpendicular to the wafer surface due to the collisions it experiences with the ions. These collisions impart the stated direction to the motion of the neutral atoms. Thus, these atoms and molecules collide only with the exposed wafer surface and not the sidewalls that develop as etching proceeds. This causes the exposed surface of the wafer to experience damage due to the collisions, which makes the area considerably more reactive than the sidewalls and results in vertical etching. The second point of view recognizes that the photoresist is also either being etched or sputtered off and reaction products form at and near the wafer surface. The products of the etch process then coat both the sidewalls and the wafer surface. However, the aforementioned sputtering action removes this coating at the surface of the wafer exposing it to further chemical etching. These are idealizations of the actual mechanism that gives rise to vertical etching. In actuality, both processes are likely to occur. The mechanisms for vertical etching have not been incorporated into a comprehensive model of the process. One obstacle is mathematically describing the relation of collisional damage to increased chemical reactivity. However, these models predict that the etch will be more vertical as the DC bias increases. To check this prediction, one has to quantify the idea of vertical etching in a way that can be experimentally measured. This is accomplished by defining the isotropy of etching and selectivity as follows. Let u denote the undercut, D the etch depth, d the thickness of the film being etched, and S denote the selectivity. The selectivity, S, is defined as the ratio of the etch rate (usually in μ/min) of the film

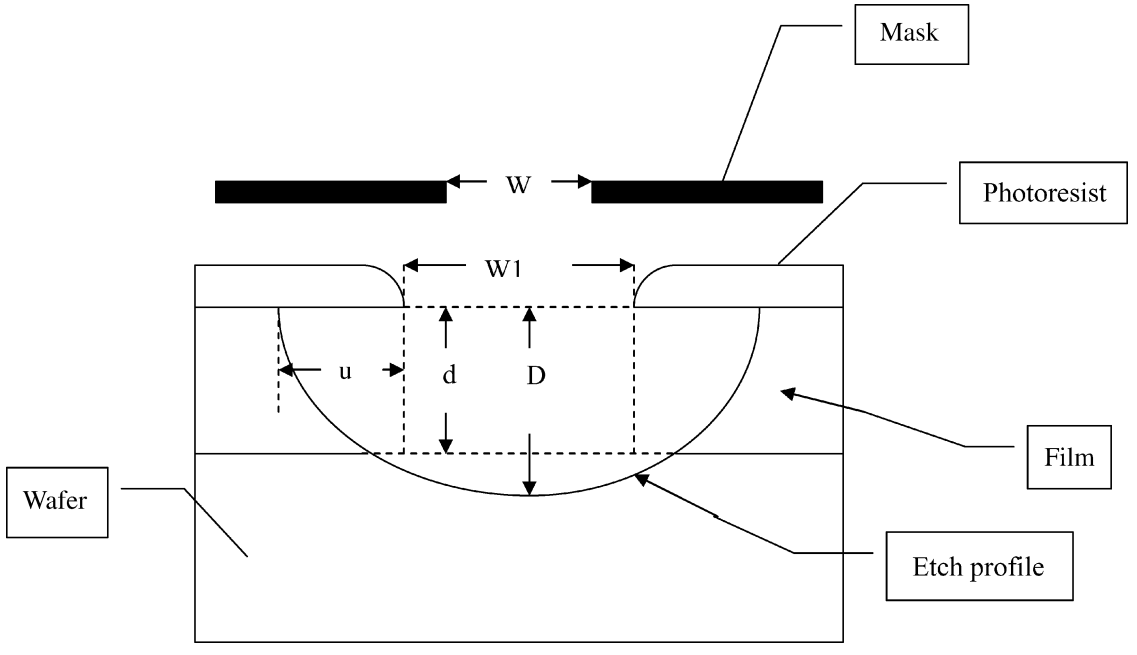


Fig. 4 Illustration of the physical quantities that determine the anisotropy and pattern fidelity. Here, u denotes the undercut, D the depth of the etch, and d denotes the thickness of the film being etched. The anisotropy, A, is defined as $A = u/D$.

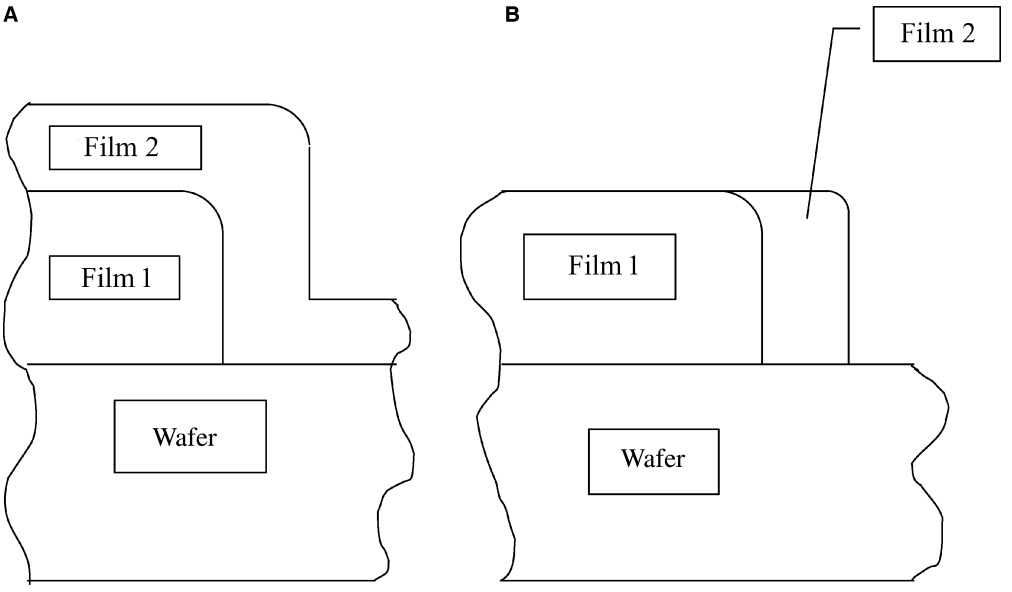


Fig. 5 (A) Depicts wafer prior to etching and (B) depicts the result of etching to endpoint. Removing the remainder of film 2 (residue) requires over-etching of both the wafer and film 1.

being etched to the etch rate of the material in the adjacent layer for the specific etching species being used. Selectivities of the order of 10 or greater are desirable because etch rates on the order of hundreds of angstroms per minute are common and selectivities of this order of magnitude are necessary to prevent excessive removal of second layer material leading to device failure. The anisotropy, A, is defined as $A = u/D$ (see Fig. 4). This is equivalent to defining the anisotropy as $A = 1 -$ (lateral etch rate/vertical etch rate). Thus, an anisotropy of zero corresponds to vertical etching only. One chooses a specific etch recipe (specific choice of feed composition to the etcher) and operating conditions to obtain anisotropy values less than 0.1 and $S \geq 10$. For $S \geq 100$, $d \approx D$ and the second layer from a practical point of view is not etched. By increasing the DC bias, smaller values of A are usually obtained; however, the selectivity frequently decreases.

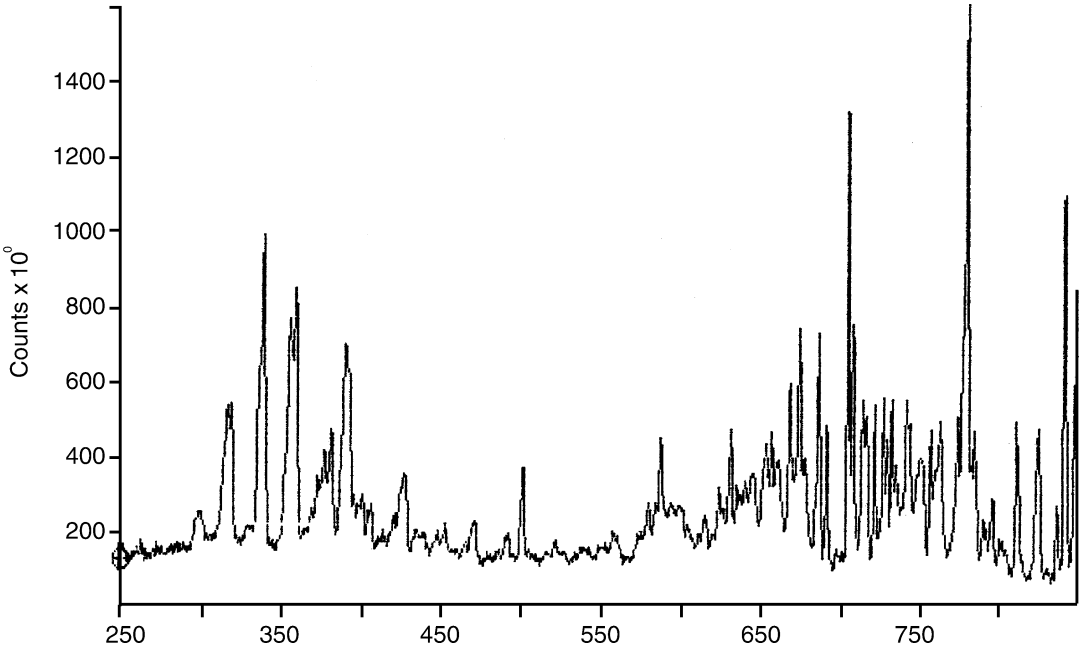


Fig. 6 A representative optical spectrum of a plasma etch process.

Thus, there is a trade-off between large S values and small A values when increasing the DC bias.

A second consideration is pattern fidelity, which is defined as the width of the feature in the photoresist before etching (which is also the width of the feature in the mask) minus the width of the feature in the photoresist after etching (which is the width of the feature etched into the surface of the wafer for a totally anisotropic etch). Pattern fidelity is controlled by both the chemistry of the process and the DC bias. Photoresists are known to undergo chemical reactions with the etching species as well as physical sputtering. Increasing the DC bias increases the sputtering effect while changing the recipe used in plasma etching can reduce photoresist etching. Pattern fidelity values must be very small for today's advanced microelectronics manufacturing with submicron geometries. The exact value is dictated by design considerations.

Over-etching is an important process consideration. The point at which a film has been etched to completion, i.e., when the film being etched is completely removed to the substrate, is called the endpoint. Over-etching occurs when etching continues beyond the endpoint. Over-etching occurs under three common circumstances. The first occurs when etching is not uniform across the wafer surface. In this case, the area of the wafer that etches the fastest will experience over-etch while the slowest etching portion of the wafer is being etched to completion. The second case occurs when there are two distinct film layers above the wafer surface as shown in Fig. 5. If one is to totally remove the top film (the ''residue'' of film 2), then both the substrate and film 1 will be over-etched. The third case occurs when the film layer is not uniform across the area being etched. The requirement that only an acceptably small amount of material may be etched away from the underlying surface during over-etching places lower limits on the values of the selectivity of material two to material one. The first two cases occur routinely in the complementary metal oxide semiconductor, [(CMOS), see Refs.[33,34]] process. The third case can also occur in CMOS fabrication.

Table 1 Emission wavelengths useful for endpoint detection

Film being etched	Wavelength (nm)
Resist	297.7, 483.5, 519.5
Si, poly-Si	704, 777
SiN, Si_3N_4	704, 674, 337
Aluminum	837, 396

The final process consideration to be discussed is endpoint detection. The objective is to determine when the material being etched has been completely removed exposing the surface of the layer beneath (endpoint) and to stop the etching at the endpoint. There are many methods to accomplish this but three in particular are in widespread use. The first method employs a spectrometer to measure the intensity of a specific spectral line. The spectral line chosen is one whose intensity will change significantly when the endpoint is reached. Thus, the etchant species or the etch products are viable candidates for the choice of spectral lines. The well-known 703.7 nm fluorine emission line is used in fluoride etching because of its strong intensity, the 337 nm N_2 line is used when etching silicon nitride, and the 837 nm line is used when etching aluminum. The endpoint is detected by observing when this change in intensity occurs. A representative spectrum for an etch process is shown in Fig. 6 and a short list of emission wavelengths useful for endpoint detection are given in Table 1.

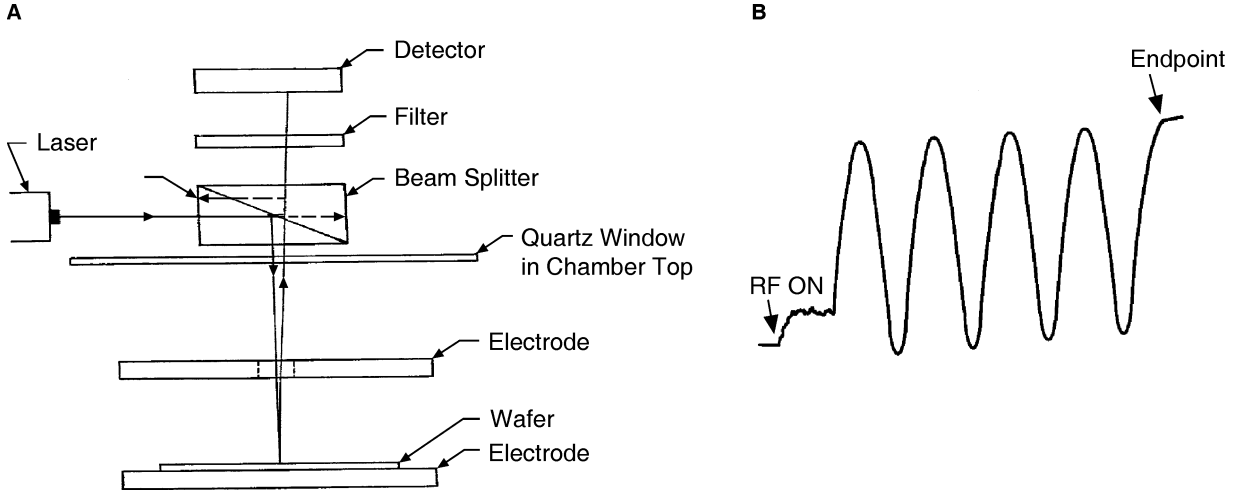


Fig. 7 (A) A common configuration for measuring the endpoint using laser interferometry and (B) a laser inteferogram depicting the point at which the RF-power is turned on as well as the endpoint for the etching process.

Table 2 Index of refraction for some common materials in semiconductor processing

Material	Index of refraction, m
SiO_2	1.46
Si_3N_4	2.05
Si	3.00
Positive resist	1.60

The second method uses laser interferometry to determine the endpoint. A laser emitting radiation at 632.8 nm is typically used. A common configuration found in plasma etching equipment is shown in Fig. 7; also shown is a typical inteferogram for the etching process. The formula for the film thickness etched is given by $t = n\lambda/2m$, where t is the thickness etched, n the number of periods (maxima or minima), λ the laser wavelength, and m is the index of refraction of the film being etched. The endpoint is usually identified as the point at which the inteferogram does not change with time. Values for the index of refraction for some common materials are given in Table 2.

The third method for endpoint detection uses a mass spectrometer that monitors a particular species whose concentration changes dramatically at the endpoint. In all three methods, the etch rate for a particular material can be determined. However, the interferometer technique has two advantages over the other methods. Changes in the etch rate with time can be detected and quantified, and significant etching of photoresist can be detected from the output of the strip chart recorder. In the next section, we take up the applications of plasma etching.

APPLICATIONS

As has already been stated previously, plasma etching is principally used in microelectronic fabrication. This is predominately due to cost considerations; the equipment for plasma etching comes with a high cost. The reason it is widely used in microelectronics manufacturing is due to the fineness of the features that can be obtained (submicron), the anisotropic nature of the etching process, and the elimination of post-treatment considerations for liquid wastes. The discussion of applications will focus on the enumeration of materials that have been etched using plasma etching and a brief presentation of etch rates that typify what occurs in plasma etching applications.

Listed in Table 3 are materials that have been plasma etched and the recipes (species in the feed gas) that have been used.

In Table 4, the etch rates and selectivities for some common materials are listed.

One should note that the actual values of the etch rates obtained are highly dependent on the composition of the feed gas, the applied RF-power, and the DC bias used. The values given in Table 4 are only indicative of what can be physically realized and have been measured in the author's laboratory. The same statement is true for the selectivities, which depend on the etch rates. There is a great deal of flexibility that can be advantageously used to obtain the desired etching conditions. Currently, the equipment manufacturers and research laboratories that focus on microelectronic manufacturing do this experimentally.

Although microelectronic applications dominate the use of plasma etching, there are many other

Table 3 Commonly etched materials and their etching recipe

Material	Plasma gas recipe
Si	CF_4, $CF_4 + O_2$, CCl_2F_2, SF_6, NF_3
SiO_2	CF_4, $CF_4 + O_2$, CCl_2F_2, $C_2F_6 + H_2$, $C_2F_6 + CF_3Cl$
Poly-Si	CF_4, $CF_4 + O_2$, $CF_3Cl + CF_4$, $C_2F_6 + CF_3Cl$
Si_3N_4	CF_4, $CF_4 + O_2$, SF_6
Al	CCl_4, BCl_3, $SiCl_4$
Al_2O_3	$CCl_4 + Ar$, BCl_3
GaAs	CCl_2F_2
Mo	$CF_4 + O_2$
Au	$C_2Cl_2F_2$
Pt	$CF_4 + O_2$, $C_2Cl_2F_4 + O_2$
Ti	CF_4
Ta	CF_4
W	CF_4, $CF_4 + O_2$
Cr	CCl_4
Cr_2O_3	$CCl_4 + Ar$

Table 4 Etch rates for common materials and some associated selectivities

Etch rate (A/min)	$CF_3Cl + CF_4$	CF_3Cl	$CF_4 + O_2$	$C_2F_6 + CHF_3$	$C_2F_6 + Cl_2$	$C_2F_6 + CF_3Cl$
Undoped poly Si		350			1055	995
SiO_2	50	60	70	40	120	200
p-doped Poly-Si	600	800	800	200	1175	1595
Photoresist	200	100	150	250	190	570
Si			1300			
Si_3N_4			1600	650		
P.S.G.			220	1500		
Selectivities						
p-doped poly Si/SiO_2		13	11	5	10	8
p-doped poly Si/resist		8	5	0.8	6	3
undoped poly Si/resist		4			6	2
undoped poly Si/SiO_2		6			9	5
Si_3N_4/P.S.G.			7			
P.S.G./SiO_2			3	38		

applications for plasma etching. Plasma etching is used for surface activation, cleaning, and modification of surface properties. It can make many surfaces hydrophilic, hydrophobic, bondable, lubricated, passivated, or abrasion resistant. These properties frequently find application in the medical equipment field. It can be used to etch plastic surfaces and act as a pretreatment in paint coatings or the gluing of plastics.

PLASMA DIAGNOSTICS

Plasma diagnostics consists of tools that can be used to characterize the plasma and thus provide information on the processes that are occurring with the plasma etcher. The main tools employed are spectroscopy, emission spectroscopy, mass spectrometry, and plasma probes. A detailed discussion on how each tool works is beyond the scope of this review. However, these tools are used to identify the species present in the plasma etcher, determine the temperature of the species, and determine the electron density.

CONCLUSIONS

Plasma etching is likely to remain a major component of microelectronic manufacturing for the future. Innovative ideas on adding additional inductively coupled electrodes and designing new feed nozzles to achieve a uniform plasma environment at the surface being etched will continue to lead to new plasma etcher designs. The driving force for these new designs will be the achievement of 100% yields for chips on a wafer as the design rules become smaller and smaller. Ultimately, the lower limit to the design rules will occur when the distance between active devices on the chip become the same order of magnitude as the molecular size of the wafer material. At the fundamental level, many questions remain to be answered before a comprehensive understanding of the plasma etching process can be achieved. There is active research on improving the description of the plasma etching process through modeling various aspects of this process. This work will continue to elucidate our understanding of this interesting and challenging process.

REFERENCES

1. Winters, H.F.; Coburn, J.W. The etching of silicon with XeF_2 vapor. Appl. Phys. Lett. **1979**, *50*, 3189–3196.
2. Coburn, J.W.; Winters, H.F. Plasma etching—a discussion of mechanisms. J. Vac. Sci. Technol. **1979**, *16*, 391–403.
3. Bondur, J.A. Dry process technology. J. Vac. Sci. Technol. **1976**, *13*, 1023–1029.
4. Somekh, S. Introduction to ion and plasma etching. J. Vac. Sci. Technol. **1976**, *13*, 1003–1007.
5. Lehmann, H.W.; Widmer, R. Profile control by reactive sputter etching. J. Vac. Sci. Technol. **1978**, *15*, 319–326.
6. Schwartz, G.C.; Schaible, P.M. Reactive ion etching of silicon. J. Vac. Sci. Technol. **1979**, *16*, 410–413.
7. Mogab, C.J.; Harshbarger, W.R. Abstract: Plasma-assisted etching for pattern transfer. J. Vac. Sci. Technol. **1979**, *16*, 408–409.

8. Kushner, M.J. A kinetic study of the plasma-etching process: I. A model for the etching of Si and SiO_2 in C_nF_m/H_2 and C_nF_m/O_2. J. Appl. Phys. **1982**, *53*, 2923–2938; IBID, A kinetic study of the plasma-etching process: II. Probe measurements of electron properties in an RF plasma-etching reactor. J. Appl. Phys. **1982**, 53, 2939–2946.
9. Smolinsky, G.; Flamm, D.L. The plasma oxidation of CF_4 in a tubular-alumina fast-flow reactor. J. Appl. Phys. **1979**, *50*, 4982–4987.
10. Bell, T. Alex. Abstract: Fundamentals of plasma chemistry. J. Vac. Sci. Technol. **1979**, *16*, 418–419.
11. Coburn, J.W.; Winters, H.F. Ion- and electron-assisted gas-surface chemistry—an important effect in plasma etching. J. Appl. Phys. **1979**, *50*, 3189–3196.
12. Coburn, J.W.; Winters, H.F. Abstract: Mechanisms in plasma etching. J. Vac. Sci. Technol. **1978**, *15*, 327–328.
13. Mogab, C.J.; Adams, A.C.; Flamm, D.L. Plasma etching of Si and SiO_2—the effect of oxygen addition to a CF_4 plasma. J. Appl. Phys. **1978**, *49*, 3796–3803.
14. Ephrath, L.M. Selective etching of silicon dioxide using reactive ion etching with CF_4/H_2. J. Electrochem. Soc. **1979**, *126*, 1419–1421.
15. Flamm, D.L.; Donnelly, V.M.; Mucha John, A. The reaction of flourine atoms with silicon. J. Appl. Phys. **1981**, *52*, 3633–3639.
16. Tokunaga, K.; Hess, D.W. Aluminum etching in carbon tetrachloride plasmas. J. Electrochem. Soc. **1980**, *127*, 928–932.
17. Flamm, D.L.; Donnelly, V.M. The design of plasma etchants. Plasma Chem. Plasma Process. **1981**, *1*, 317–363.
18. Koenig, H.R. US Patent 3,661,761, 1972.
19. Jancel, R.; Kahan, Th. *Electrodynamics of Plasmas*; John Wiley and Sons, 1966.
20. Graves, D.B.; Jensen, K.L. A continuum model of DC and RF discharges. IEEE Trans. Plasma Sci. **1986**, *PS-14*, 78–91.
21. Graves, D.B. Fluid model simulations of a 13.56 MHz r discharge: time and space dependence of rates of electron impact excitation. J. Appl. Phys. **1987**, *62*, 88–94.
22. Boeuf, J.P. Numerical model of rf glow discharges. Phys. Rev. A **1987**, *36*, 2782–2792.
23. Makabe, T.; Nakano, N.; Yamaguchi, Y. Modeling and diagnostics of the structure of RF glow discharges in Ar at 13.56 MHz. Phys. Rev. A. **1992**, *45*, 2520–2531.
24. Park, S.K.; Economou, D.J. Analysis of low pressure rf glow discharges using acontinuum model. J. Appl. Phys. **1990**, *68*, 3904–3915.
25. Tsai, J.H.; Wu, C. Two-dimensional simulations of RF glow discharges in N_2 and SF_6. Phys. Rev. A. **1990**, *41*, 5626–5644.
26. Kushner, M.J. Mechanisms for power deposition in Ar/SiH_4 capacitively coupled RF discharges. IEEE Trans. Plasma Sci. **1986**, *PS-14*, 188–196.
27. Kushner, M.J. Monte-carlo simulation of electron properties in RF parallel plate capacitively coupled discharges. J. Appl. Phys. **1983**, *54*, 4958–4965.
28. Richards, A.D.; Thompson, B.E.; Sawin, H.H. Continuum modeling of argon radio frequency glow discharges. Appl. Phys. Lett. **1987**, *50*, 492–494.
29. Blauw, M.A.; van der Drift, E.; Marcos, G.; Rhallabi, A. Modeling of fluorine-based high-density plasma etching of silicon trenches with oxygen sidewall passivation. J. Appl. Phys. **2003**, *94*, 6311–6318.
30. Ibbotson, D.E.; Flamm, D.L.; Mucha, J.A.; Donnelly, V.M. Comparison of XeF_2 and F-atom reactions with Si and SiO_2. Appl. Phys. Lett. **1984**, *44*, 1129–1131.
31. Stamou, S.; Mataras, D.; Rapakoulias, D. Spatial rotational temperature and emission intensity in silane plasma. J. Phy. D: Appl. Phy. **1998**, *31*, 2513–2520.
32. Viggiano, A.A. Ion molecule chemistry at high temperature: Derivation of rotational vibrational energy dependence; http://www.cstl.nist.gov/div838/kinetics2001/agenda/j_session/j5/j5.htm (accessed 2003).
33. Campbell, S.A. *The Science and Engineering of Microelectronic Fabrication*; Oxford University Press: New York, 1996.
34. Sze, S.M. *VLSI Technology*; McGraw-Hill: New York, 1983.

Plasma Polymerization Coating

Hirotsugu K. Yasuda
University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

When an organic vapor, such as methane, in low pressure (e.g., less than 1 torr) is subjected to an electromagnetic field, the electrical breakdown of the gas occurs, yielding a glow the color of which is characteristic to the gas. In the luminous gas phase, methane is activated and forms a polymeric deposition in the form of a coating on the surface of substrate placed in the glow. This process is termed ''plasma polymerization'' because the luminous gas phase or glow indicates the presence of plasma, and the process does not proceed without plasma. The strict definition of plasma is (at least partially) ionized gas, which maintains the electrical neutrality as a whole. The luminous gas phase in which plasma polymerization takes place, however, is not plasma in the strict sense.

Plasma polymerization is not a process to synthesize a polymer, but is a coating process, because the main product is in the form of a coating on the surfaces that are either in direct contact with the luminous gas phase or placed in the glow region of a reactor. Because of this, plasma polymerization is referred to as plasma polymerization coating. Plasma polymerization coating is unique in that a single process using a gas-phase monomer (starting material) achieves both the polymerization and the coating simultaneously. It is ideally suited for coating of a nanofilm (e.g., 10–100 nm), which is difficult to achieve by other conventional coating processes. Furthermore, the characteristics that could be attained by plasma polymerization coatings are unique in that they cannot be obtained by other conventional means of coating.

PLASMA, GLOW, AND PLASMA POLYMERIZATION

The glows in direct current (DC) discharge of Ar in a plasma polymerization reactor are schematically shown in Fig. 1. The size and the location of glows are dependent on the system pressure p and the distance d between the cathode and the anode. The parameter given by pd determines the glow characteristics including the breakdown voltage of discharge.

The distribution of electron temperature T_e (energy of electron) and electron density in a DC argon glow discharge in a plasma polymerization reactor is shown in Figs. 2 and 3, respectively.[1] The electron temperature rises as electrons are accelerated in the electric field. During this process, the number of electrons is relatively small. When the electron temperature reaches the maximum level as a function of the distance, T_e starts to drop significantly as the ionization of Ar atom occurs, which consumes the electron energy and creates an additional electron. The creation of excited species of Ar, which is responsible for the characteristic glow, occurs simultaneously with the ionization in the same location. The number of electrons starts to increase in this location and the density quickly increases as electrons are pulled toward the anode as depicted in Fig. 3. The true plasma state exists only in a narrow cross-sectional layer in the vicinity of the location where ionization of Ar occurs.

The main glow of Ar appears corresponding to the distance for the maximum electron temperature as observed in Fig. 2, implying that the glow recognized as the negative glow is associated with the ionization of Ar. The energy of photon-emitting species is in the vicinity of the ionization energy of Ar. Accordingly, the negative glow can be designated as ionization glow.

The ionization of organic molecules in glow discharge is not as simple as that of Ar. First of all, the ionization energy of greater than 10 eV is far above the bond energies of primary bonds involved in organic compounds. Typical bond energies are given in Table 1. Dissociation energy, metastable energy, and ionization energy for noble gases and diatomic gases are compared in Table 2. The low-energy electrons and/or off-centered collisions that cannot ionize molecules can break the bonds in organic molecules or create excited species that can trigger chemical reactions. These reactions are absent in the ionization of atoms. Furthermore, the removal of small molecules, such as H_2, HCl, HF, etc., out of an original organic molecule requires very little energy in comparison with the ionization energies.

It has been recently discovered that glow in DC glow discharge of organic compounds is significantly different from that of Ar.[2,3] In Ar discharge, the ionization glow is the primary glow and the cathode is virtually in the dark space. In strong contrast to Ar glow, the cathode in trimethylsilane (TMS) discharge is surrounded by the glow, which is weaker compared

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007902

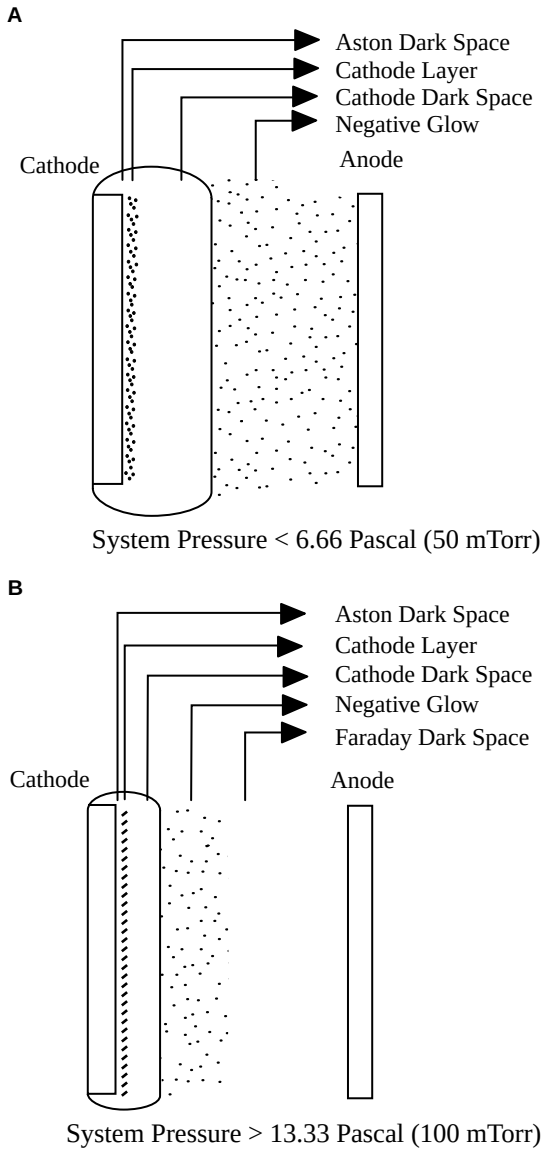


Fig. 1 Schematic view of nonmagnetron DC glow discharge of argon in the interelectrode space, interelectrode distance: 102 mm; the luminous regions are shown shaded.

to the negative glow observed in Ar discharge but is the primary glow of the system, especially at the inception of the glow discharge. The main glow adheres to the cathode surface, which means that the glow is caused by low-energy electrons and can be termed as dissociation glow in contrast to the ionization glow.

Simple molecular gases, such as O_2, N_2, CO_2, etc., do not create the dissociation glow, and photon-emitting species have energy at the level of ionization. The dissociation glow is observed only with organic gases that can be used in plasma polymerization coating, which implies that plasma polymerization proceeds by the dissociation of molecules caused by low-energy electrons but not by the ionization of molecules.

The glow discharge by alternating polarity power source up to approximately 100 kHz is essentially the same with DC discharge except that the two electrodes alternate their polarity. Accordingly, the deposition on an electrode drops to one-half of the deposition on the cathode of DC discharge.[4] As the frequency increases to radiofrequency (RF) range, an electron cannot complete the path from the cathode to the anode within the duration of one cycle, and the mode of motion changes from the linear motion to the oscillation according to the frequency of the power source. The oscillating electrons shift the dissociation glow from the electrode surface to the gas phase.

The glow in which plasma polymerization occurs is not a uniform luminous gas phase that could be viewed as multiple layers (as the onion layer structure). The location of the crucially important layer for plasma polymerization coating differs depending on the nature of gas, the frequency of electrical power source, the mode of coupling the electromagnetic energy to the gas phase, and the design factors of a reactor, that is, plasma, glow, and plasma polymerization are system dependent, and any of these terms cannot be described without specifying the details of the system. Where the substrates are placed with respect to the onion layered structure is an important factor in the processing of plasma polymerization coating.

GROWTH AND DEPOSITION MECHANISMS

As it can be recognized in the plasma polymerization coating of methane, no specific molecular structure is required to form plasma polymerization coating. The organic compounds that do not have any functional group, such as methane, can be polymerized by plasma polymerization and the normalized molar deposition rates of these nonmonomers are, by and large, the same as those for monomers that have double or triple bond, although some chemical structures influence the deposition rate in subtle manners. The growth mechanism in plasma polymerization is completely different from those of conventional polymerizations.[5]

Conventional polymerization does not occur in gas phase, particularly in vacuum, because of the limitation set by the ceiling temperature of polymerization, and there are only a few cases in which the deposition of polymeric material occurs in vacuum. Those exceptional cases are plasma polymerization and Parylene polymerization, which is also a vacuum polymerization coating process using a gaseous monomer. The common denominators for these two processes are: 1) the polymerizations yield solid-state polymer (in the form of film in most cases) from a gas phase monomer in vacuum and 2) the polymer formed by the processes

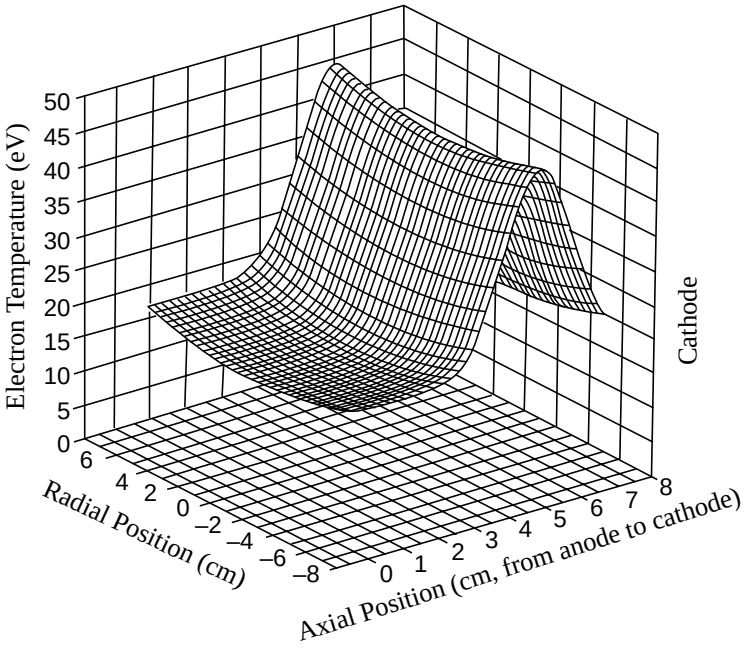


Fig. 2 Distribution profile of electron temperature in an argon DC glow discharge in a plasma polymerization reactor.

contains a large amount of dangling bonds (free radicals trapped in the solid polymer matrix).

Parylene polymerization proceeds with well-defined chemical species, whereas plasma polymerization proceeds via a variety of not well-defined chemical species, which are created in the luminous gas phase. The reactive species for Parylene polymerization is *para*-xylylene, which is bifunctional (i.e., biradicals), moderately reactive and relatively stable, and highly selective in chemical reactivity to its own structure.

Highly reactive (unstable) and nonselective species involved in plasma polymerization coating tend to react with any polymeric surface on which the species strike and form a polymer deposition with a high level of bonding or adhesion to the surface. Because of this aspect, plasma polymerization forms a thin film with a good adhesion to various kinds of substrate materials. Because of nonselective reactivity, the reactive species of plasma polymerization has poor penetration into small cavities, such as those of porous structure. Highly reactive species react with wall material at the entrance of a pore instead of penetrating into the pore.

Parylene polymer deposition has very poor adhesion to a smooth-surface substrate. A freestanding

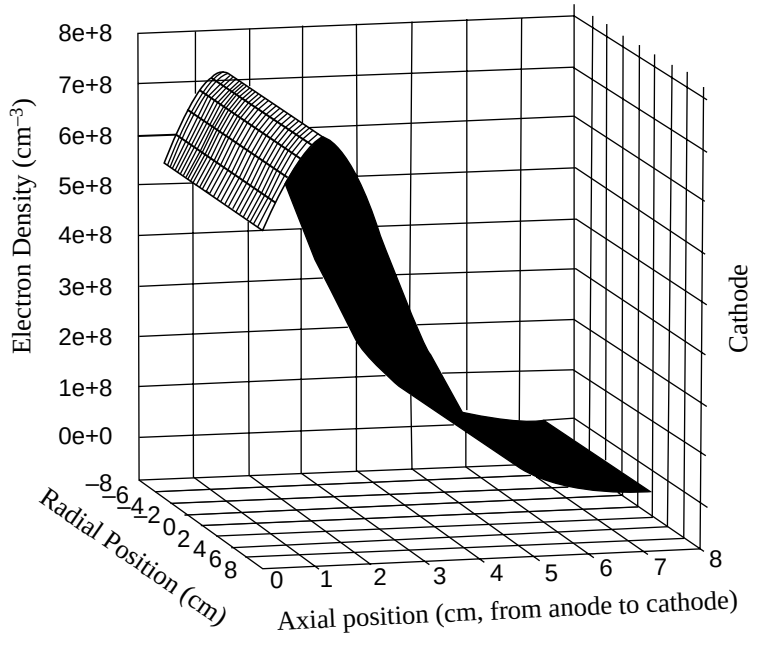


Fig. 3 Distribution profile of electron density in an argon DC glow discharge in a plasma polymerization reactor.

Table 1 Bond dissociation energy

Bond	Dissociation energy (eV)
C–C	3.61
C=C	6.35
	(2.74 for π bond)
C–H	4.30
C–N	3.17
C=N	9.26
C–O	3.74
C=O	7.78
C–F	5.35
C–Cl	3.52
N–H	4.04
C–H	4.83
O–O	1.52

ultrathin film can be obtained by peeling off the film from the substrate. On the other hand, if the substrate is porous the film adheres strongly to the substrate. Because of the high selectivity (toward the same kind of species), a *para*-xylylene molecule (in vacuum phase) can penetrate deep into the cavity without reacting with the wall material and waits for another *para*-xylylene to arrive. The second molecule reacts immediately without losing the overall reactivity because it is a bifunctional free radical. By repeating this step, Parylene deposition can penetrate deep into pores or microvoids existing on the substrate surface, yielding a very good adhesion to porous substrate by virtue of mechanical anchoring. Thus, Parylene film deposition can be characterized by 1) no adhesion to a smooth surface and 2) very good adhesion to porous substrate by forming an interpenetrating thin-layer film. These are completely opposite to the characteristics of plasma polymerization coatings.

Because of the well-defined chemical structure of the reactive intermediate, Parylene coating is semicrystalline; whereas plasma polymerization coating consists of an amorphous (noncrystalline) three-dimensional network because of the lack of well-defined chemical structure of reactive species. The amorphous nature of plasma polymerization coatings plays a crucially important role in protective coatings because there is no discernible defect or weak boundary, whereas the boundary between crystals and noncrystalline phase in a semicrystalline Parylene coating causes the intrusion of salts in a corrosive environment.[6]

In both the polymerizations, free radicals are the species that are responsible for the formation of bonds in the depositing materials. The growth mechanism, however, is not by the conventional chain-growth free-radical polymerization. In a conventional free-radical chain-growth polymerization, two free radicals and 10,000 monomer molecules yield a polymer with degree of polymerization 10,000, which does not contain free radicals. In contrast to this situation, in plasma polymerization and Parylene polymerization, 10,000 species with free radical(s) recombine to yield a polymer matrix that has an equivalent degree of polymerization, and contains numbers of unreacted free radicals (dangling bonds).

Table 2 Dissociation energy, metastable energy, and ionization energy for noble and diatomic gases

Gas	Dissociation energy (eV)	Metastable energy (eV)	Ionization energy (eV)
Ho	—	19.8	24.6
Ne	—	16.6	21.6
Ar	—	11.5	15.8
Kr	—	9.9	14.0
Xe	—	8.32	12.1
H_2	4.5	—	15.6
N_2	9.8	—	15.5
O_2	5.1	—	12.5

Rapid Step-Growth Polymerization Mechanism

The complex nature of plasma polymerization was explained by rapid step-growth polymerization (RSGP) in which recombination of free radicals constitutes the main mechanism to increase the size of molecules.[5] The presence of a large amount of dangling bonds, the change of the concentration of dangling bonds as a function of the duty cycle of pulsed discharge, and the dependence of the deposition rate on molecular structures of monomers, all support the RSGP mechanism. The increase of the dangling bonds because of pulsed discharge cannot be explained without RSGP mechanisms.

A schematic presentation of growth mechanisms is presented in Fig. 4. The actual growth mechanisms involve the participation of the surface or the third body in each reaction shown in the figure. To simplify the overall picture of reactions in luminous gas phase, however, the contribution of the surface is not shown in the figure. The role of ions in the growth reactions can be virtually ignored because the number of neutral species outnumber nearly 10^6 to 1.[7] How these reactions contribute to the deposition of materials is schematically illustrated in Fig. 5.

None of the reactions shown is a polymerization by itself. While repeating the steps via cycle-I or cycle-II, species with varying sizes deposit on the substrate surface. The once-deposited species increase in size by the same principle. The number of cycles repeated

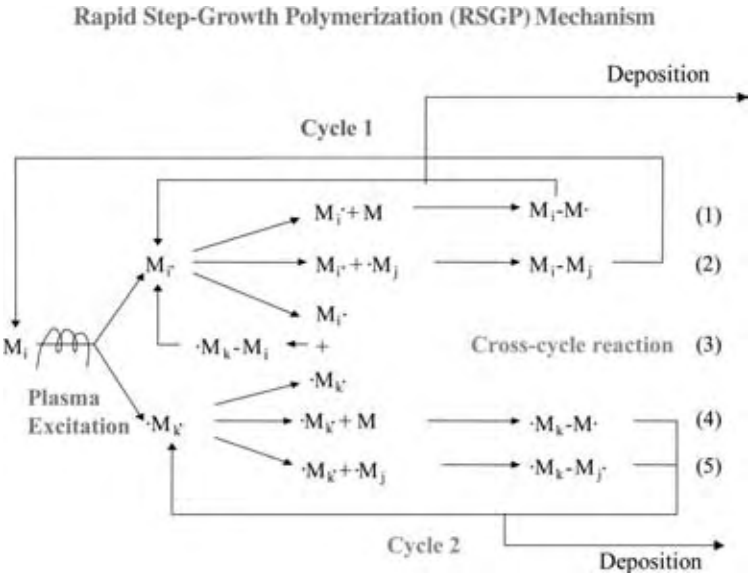


Fig. 4 Growth and deposition mechanisms in the luminous gas phase. *(View this art in color at www.dekker.com.)*

before the deposition occured can be expressed by the term kinetic path length. As the kinetic path length increases, the size of the gaseous species increases, and the saturation vapor pressure of the species decreases accordingly.

Influence of the Kinetic Path Length on Properties of Coating

An important feature of plasma polymerization is that the creation and the deactivation of reactive species (growth and deposition reactions) are not kinetically coupled. The reactive species is created irrespective of what happened to the created reactive species.

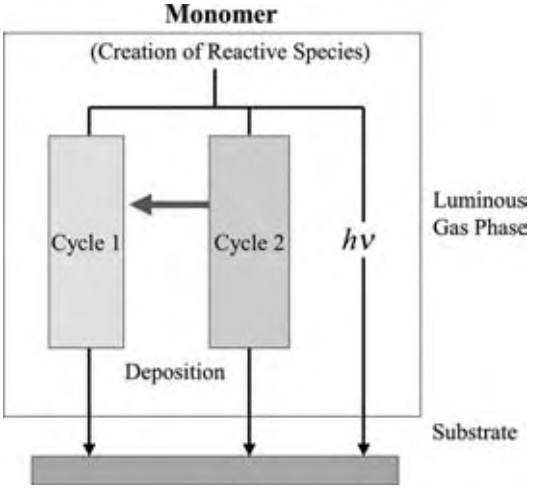


Fig. 5 Gas phase reactions and material deposition. *(View this art in color at www.dekker.com.)*

Consequently, if the temperature of a substrate is changed without changing plasma conditions, only the deposition (deactivation) process is influenced by the change. If one plots the deposition rates k at different substrate temperatures, T, as $\log k$ vs. $1/T$, it yields a negative slope indicating that the rate process cannot express the deposition rate. The change of deposition process alters the deposition rate as well as the chemical nature of plasma polymers under the identical plasma conditions, because the kinetic path length is altered.

This situation can be visualized by the analysis of plasma polymers of perfluoro-2-butyltetrahydrofuran (PFBTHF) formed at different substrate temperatures under two sets of discharge conditions.[8] Fig. 6 depicts the temperature dependence of deposition rate shown as plots of $\ln k$ vs. T. The electron spectroscopy for chemical analysis (ESCA C_{1s}) spectra of polymers deposited at different temperatures under different energy input levels are shown in Fig. 7. The details of ESCA C_{1s} spectra are shown in Table 3. The monomer is the same, and the discharge conditions are identical. Only the surface temperature of the thickness monitor is changed. Therefore, the luminous gas phase should be identical in all cases. Under such conditions, the following features of plasma polymerization coating are clearly seen: 1) the products obtained at different temperatures are not the same; 2) the temperature dependence of deposition rate is negative; and 3) a higher deposition rate at the lower temperature is obtained at the expense of the kinetic path length.

The most important factor, at a given substrate temperature, that influences the properties of plasma polymers from a monomer is the energy input level of the plasma polymerization process. The energy input level determines the extent of dissociation or

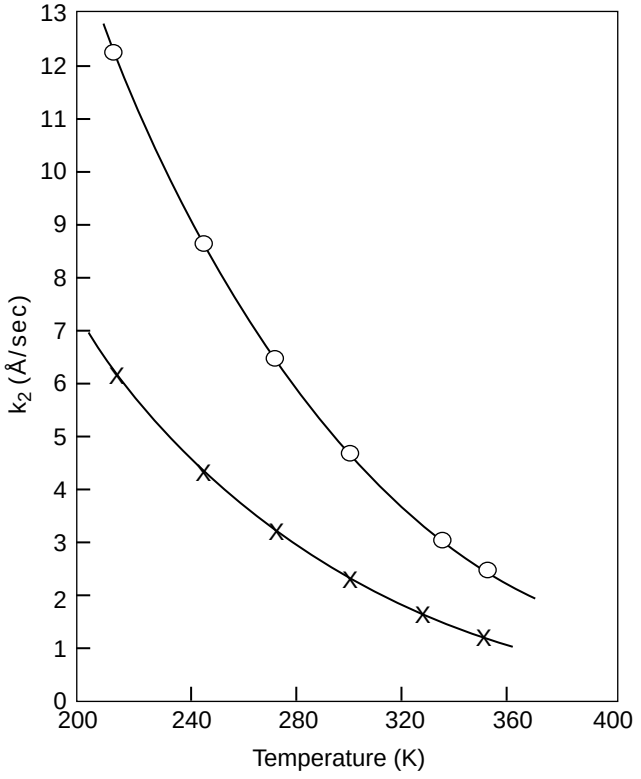


Fig. 6 Temperature dependence of the thickness growth rate k_2 for PFBTF. Flow rate [cm^3(STP)/min] and power (W): O, 1.04 (9.84 W); x, 0.54 (4–89 W). $W/FM = 12.8 \times 10^7$ J/kg.

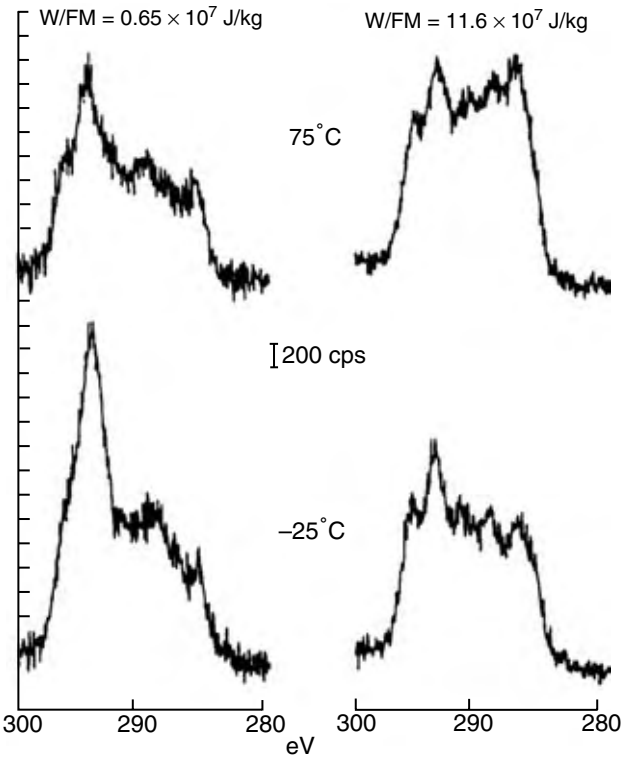


Fig. 7 The ESCA C_{1s} spectra of plasma polymers of PFBTF obtained at different W/FM and substrate temperatures.

scrambling of atoms in the monomer molecule. The retention of molecular structure of the original monomer decreases with the increasing level of energy input. The energy level in the luminous gas phase can be manifested by the parameter *W/FM*, where *W* is the discharge wattage, *F* is molar or volume flow rate, and *M* is molecular weight of the monomer.[5] The parameter has units of joules per kilogram, i.e., energy per mass (of monomer).

It is important to recognize that *W* is the energy input into the electrical discharge system, whereas *W/FM* is the energy input into the luminous gas phase in which plasma polymerization occurs. This subtle but very important difference could be visualized by the following analogy. A 10 W light bulb consumes 10 W of electrical energy. The bulb placed in a room hardly influences the temperature of the room. However, if the same bulb is placed in a small box used as an incubator, the bulb could be used as the sole heat source to control the temperature of the incubator. If the same bulb is kept in a hand it could cause burn. This means that 10 W itself does not mean anything with respect to the thermal effect of the system in which the bulb is placed. To see the effect of the light bulb on the temperature of air, it is necessary to divide the electrical energy input to the bulb by the total mass of air surrounding the bulb.

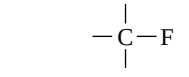
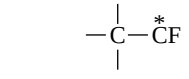
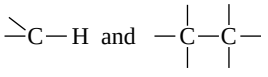
DEPENDENCE OF DEPOSITION RATE ON OPERATION PARAMETERS OF GLOW DISCHARGE

Domains of Plasma Polymerization

When the deposition rate is measured at a fixed flow rate and varying power input *W*, a line showing the dependence on *W* is obtained. However, when the same experiment is carried out at a different flow rate, another different line is obtained. Fig. 8 depicts the interrelated effect of *W* and *F* on the deposition rate plotted against *W*. At lower flow rates, the deposition rate reaches a plateau value at a very low value of *W* and becomes wattage independent. At a higher flow rate, the deposition rate reaches a plateau value at a higher value of *W*. The plateau region can be recognized as a monomer-deficient domain in which the flow rate of monomer is the rate-determining step. The domain before reaching a plateau value can be recognized as power-deficient domain in which the power input rate is the rate-determining step.

Table 3 Change in ESCA C_{1s} peaks of plasma polymers of perfluoro-2-butylterahydrofuran due to a change in substrate temperature and W/FM

W/FM [(J/kg) $\times$ 10^{-7}]	Substrate temperature (°C)	Area of component peaks (%)				
		1	**2**	**3**	**4**	**5**
0.65	75	16.0	21.1	19.4	26.6	17.0
	50	16.5	26.8	18.5	22.9	15.2
	25	17.3	27.2	19.5	22.7	13.3
	0	18.5	27.8	14.9	22.1	16.7
	−25	19.1	30.9	16.8	21.4	11.8
11.6	75	14.9	18.1	18.8	19.3	28.9
	50	15.0	18.8	18.3	20.0	28.1
	25	16.6	20.0	17.7	18.8	27.2
	0	16.7	20.2	17.5	19.4	26.2
	−25	17.8	21.0	17.8	18.8	24.6
	−49.5	18.0	21.2	17.1	19.8	24.0

	Peak	Approximate peak position (eV)	Approximate peak width (eV)
1	$-CF_3$	295	2.0
2	$-CF_2-$	293	1.9
3		291	2.3
		288	2.3
		286	2.8

When the same experiments were carried out as a function of flow rate at various fixed input power levels, *W*, the graph is as depicted in Fig. 9. The same data can be presented as a function of *W/FM*, as shown in Fig. 10. Based on *W/FM*, the material formation in plasma polymerization can be divided into two regimes: an energy-deficient regime and a monomer-deficient regime. In the energy-deficient domain, ample monomer is available but the power input rate is not sufficient. In this domain, the deposition rate increases with the power input. In the monomer-deficient (power saturated) domain, sufficient discharge power is available but the monomer feed-in rate is the determining factor for the deposition. These two domains cannot be identified based simply on the value of operational parameters. The domain can be identified only by the dependence of the deposition rate on operational parameters as depicted in the above figures.

The value of *W/FM* necessary to bring the plasma polymerization system into the monomer-deficient region is proportional to the value of the total bond energies divided by molecular weight, which can be considered as the specific bond energy of the monomer. This dependence clearly indicates that in the monomer-deficient region nearly all bonds in a monomer are broken, and because of this fact, the additional energy input does not influence the nature of plasma polymerization. The plasma polymerization in this region is a typical ''atomic'' polymerization, because the original monomer structure is nearly totally destroyed and what accounts for the polymer formation is the atoms that comprised the original monomer molecule but that are fragmented under the plasma conditions.

Thus, the value of $(W/FM)_c$ above which the deposition rate become independent of *W/FM*, which is characteristic of a monomer, can be given by

$$(W/FM)_c = \alpha\Phi$$

where $\Phi = \sum(\text{bond energy}/M)$ is the specific bond energy of the monomer and α the proportionality

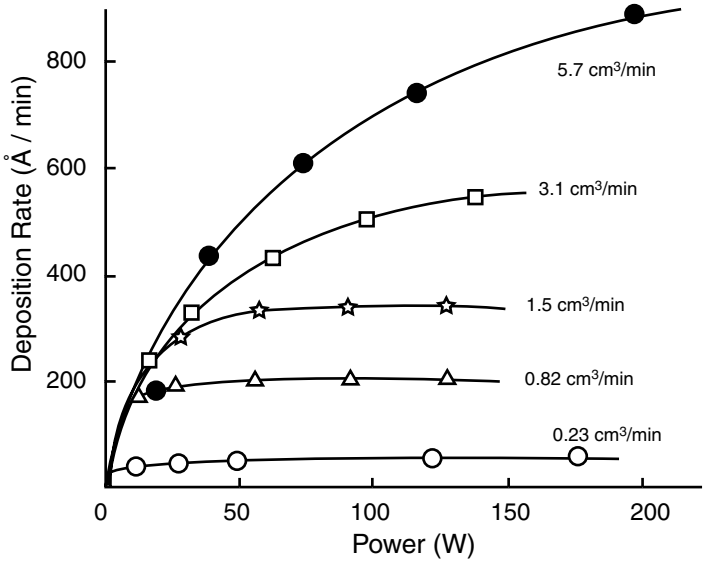


Fig. 8 Dependence of deposition rate of plasma polymer of tetramethyldisiloxane on discharge wattage at various monomer flow rates (cm^3/min).

constant for a given reactor. The value of α is roughly 20, although the value depends on the specific reactor system.[5] Thus, in the polymerization system employed, when the energy input in joules per kilogram exceeds 20 times the specific bond energy Φ (J/kg), the plasma polymerization becomes a typical atomic polymerization.

Specific Deposition Rate as a Function of *W/FM*

Most experiments start from the power-deficient domain, where the deposition of a plasma polymer can be expressed by the following expression.

$$k_1 = aW, \quad \text{where } a \text{ is a proportionality constant} \tag{1}$$

Dividing both sides of the equation by FM, one obtains

$$k_1/FM = k_0 = a(W/FM) \tag{2}$$

$$k_0 = a(W/FM) \tag{3}$$

where k_1 is the mass deposition rate and k_0 is the specific deposition rate, which is the normalized mass deposition rate.

The specific deposition rate k_0 is the only form of deposition rate that can be used to compare deposition characteristics of different monomers with different chemical structures and molecular weights under different discharge conditions (flow rate, system pressure, and discharge power). Similarly, W/FM can be considered as the normalized power input. When only one monomer is employed, k_1 can be used to establish the dependency of deposition rate on operational parameters. Even in such a simple case, k_1 cannot be expressed by a simple function of W or F, and its relationship to those parameters varies depending on the domain of plasma polymerization.

In the monomer-deficient domain, the deposition rate (asymptotic value) will increase as the flow rate is increased and shows a linear dependence on the monomer feed-in rate at a given discharge power and the system pressure (Fig. 13), i.e.,

$$k_1 = \beta(FM) \tag{4}$$

The relationship given by Eq. (4) is valid only in the monomer-deficient domain. The further increase of the

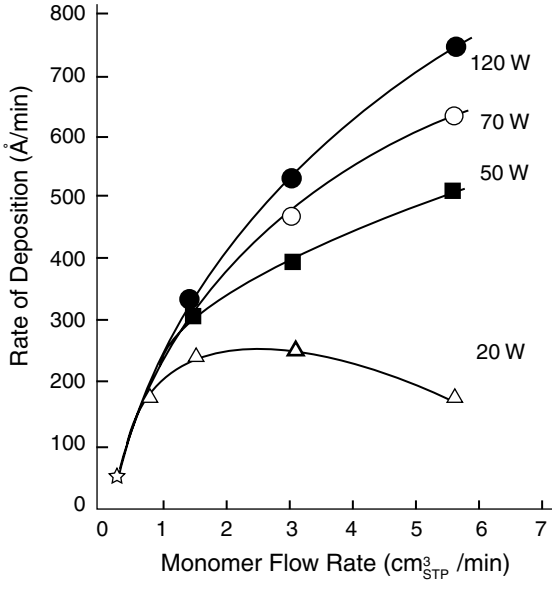


Fig. 9 Dependence of deposition rate of the plasma polymer of tetramethyldisiloxane on flow rate of the monomer.

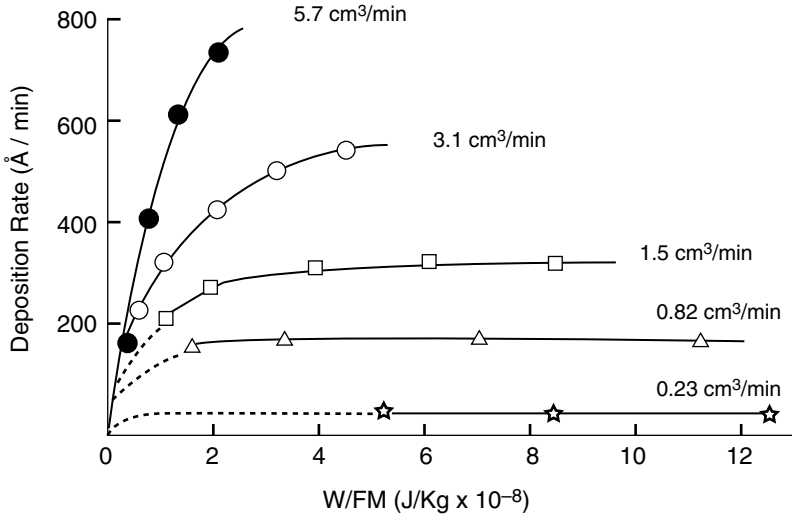


Fig. 10 Dependence of deposition rate of the plasma polymer of tetramethyldisiloxane on *W/FM* at various monomer flow rates.

flow rate (*FM*) will eventually decrease the deposition rate as the domain of plasma polymerization changes to the energy-deficient domain. Such a decrease can be expressed by Eq. (3).

The deposition rate depends on the composite parameter *W/FM*. Consequently, an increase in flow rate (at a given discharge power) has the same effect as decreasing the discharge power (at a given flow rate), and conversely, an increase of discharge power has the same effect as decreasing the flow rate. *W* and *F* cannot be regarded as independent parameters.

Fig. 11 illustrates how well the thickness growth rate, GR/*FM*, in 40 kHz and 13.5 MHz glow discharge of methane and *n*-butane, can be expressed as a function of the composite energy input parameter *W/FM*. Regardless of the mass of monomer, flow rate, and discharge wattage, a single line fits all data obtained in 40 kHz or 13.5 MHz plasma polymerization of hydrocarbons employed in which the deposition occurs on an electrically floating conductor or on a dielectric substrate placed in the glow.[9] Under the otherwise identical experimental setups, 13.5 MHz discharge yields a significantly higher deposition rate. This is due to the fact that the dissociation glow in RF discharge does not adhere to the electrode but exists in the gas phase.

The product of (deposition rate) and the deposition time *t* determine the film thickness. Hence, (*W/FM*)*t* is an important practical parameter to control the thickness of the deposition.[10] In many practical applications, in which the actual thickness of deposition is extremely difficult to measure, the overall functional character of the plasma polymerization process itself can be controlled by this parameter (*Wt/FM*).

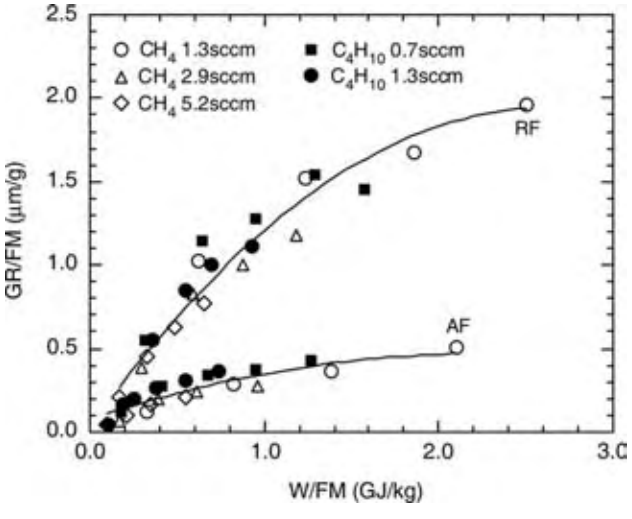


Fig. 11 Dependence of GR/*FM* on *W/FM* for 40 kHz (AF) and 13.56 MHz (RF) plasma polymerizations. Flow rate: methane, 1.3, 2.9, 5.2 sccm; *n*-butane, 0.7, 1.3 sccm.

Deposition in DC Discharge

In DC glow discharge of organic molecules, the dissociation glow adheres to the cathode surface. Consequently, the deposition onto the cathode surface is totally different from that which occurs in gas phase in most plasma polymerization coatings by HF and RF discharges. The best universal dependency for cathodic polymerization is found between k_1/M and the current density.[9] Fig. 13 depicts this relationship for all cathodic polymerization data, which were obtained in the same study, covering experimental parameters, such as flow rate, size of cathode, mass of hydrocarbon monomers, but at the same system pressure.

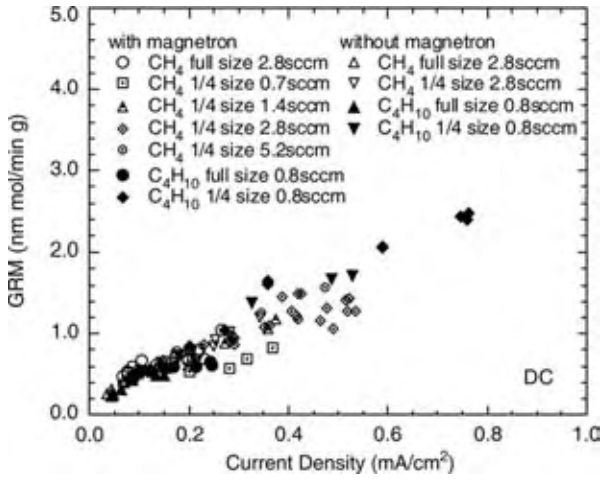


Fig. 12 A master curve for the relationship between GR/M and the current density for DC cathodic polymerization; data obtained under various conditions for methane and *n*-butane, at a fixed system pressure of 50 mtorr.

The cathodic deposition, in general cases, can be expressed by the following equation

$$k_1/[\mathrm{CM}] = \alpha_c[I] \quad (5)$$

$$k_1 = \alpha_c[I][\mathrm{CM}] \quad (6)$$

where $[I]$ is the current density, and [CM] is mass concentration of monomer in the cathode regions of a DC discharge.[4]

As the mass concentration of monomer in the cathode region depends on the system pressure, the deposition rate in the cathode region should depend on the system pressure p

$$k_1 = \alpha_c[I]p \quad (7)$$

The relationship given by Eq. (7) is a conspicuous deviation from that for the plasma polymerization (material formation in the luminous gas phase), in which the flow rate rather than the system pressure is the rate determining parameter and, hence, the deposition rate is independent of the system pressure (under a given flow rate).

MODES OF ELECTRICAL DISCHARGE PROCESS

DC Discharge

When the substrate is an electrically conducting material, DC discharge using the substrate as the cathode is a very useful and practical means to apply plasma polymerization coating on the substrate. The anode plate is a passive surface so far as plasma polymerization is concerned and is not necessary if the reactor is made of a metal.[11] The reactor wall could be used as the anode (grounded). The DC discharge has the great advantage that over 80% of plasma polymer deposits on the cathode because the dissociation glow is attached to the cathode surface, i.e., the deposition yield on substrate $Y_{p,s} > 0.8$. Depending on the size, number, and reactor design, $Y_{p,s}$ could be over 0.90. This means that the deposition on the undesirable place is minimal and could be easily controlled.

Alternating Current Discharge (Up to ~100 kHz)

An alternating current discharge, up to around 100 kHz, is essentially an alternating DC discharge and one electrode acts as the cathode in one-half of the discharge time and acts as the anode in the rest half.[4] In this mode of operation, substrates are placed in the interelectrode space. Consequently, the usable polymer deposition yield, Y_p in the interelectrode space, is less than 0.20. The deposition yield on substrates, $Y_{p,s}$, could be significantly lower than that. In DC and high-frequency discharge, the molecular dissociation glow, in which

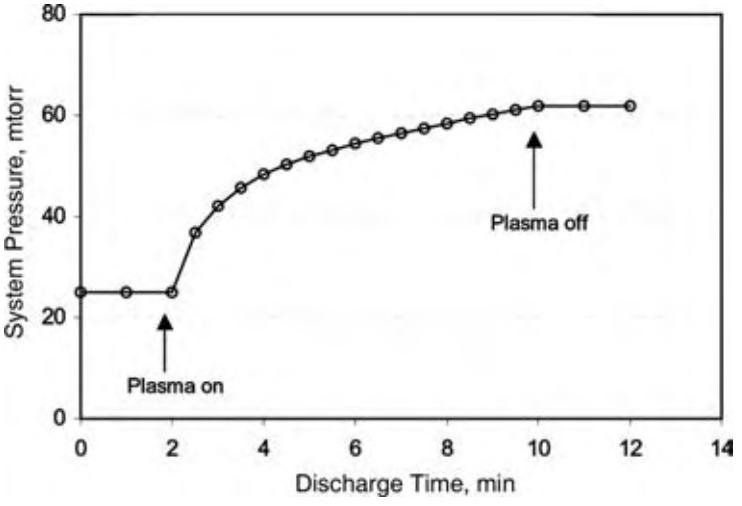


Fig. 13 Increase of system pressure during closed system plasma polymerization of TMS. Plasma conditions are: 25 mtorr, TMS, 2 panels of aluminum alloy, DC 1000 V.

the majority of chemically reactive species are created, adheres to the cathode, and consequently the interelectrode space is not a very active place unless magnetically enhanced electrode (magnetrons) is employed.

Radiofrequency Discharge

In RF discharge, the molecular dissociation glow no longer adheres to the electrode surface, and the gas phase near the electrode surface becomes the major site for the creation of chemically reactive species.[2,3] Consequently, the deposition yield on electrodes drops below 0.10, and the potential deposition yield in the interelectrode space increases to over 0.90. The deposition yield on substrate, $Y_{p,s}$, is significantly lower than the potential deposition yield because the volume of the glow expands beyond the interelectrode space as a function W/p and increases reactor wall contamination.

Because an electrode does not function as electrode in DC or alternating current high-frequency discharge, the electrode system could be kept outside a glass reactor (capacitive external electrodes) or a coil around a glass tube (inductively coupled external electrode) can be used to create plasma. These modes of coupling could be dealt as a factor in the system-dependent aspect of plasma polymerization, i.e., the basic plasma polymerization remains the same.

Pulsed RF Discharge

In pulsed RF discharge, for a certain length of time (e.g., 100 μsec) RF power is on and for another length of time (e.g., 900 μsec) power is off. During the on-period, the creation of chemically reactive species and their reactions (dissipation) occur simultaneously just as in the case of a continuous discharge. During the off-period, the creation of chemically reactive species ceases and only the dissipation reactions continues, if they could do so. The net effect is entirely dependent on the chemical structure of the monomer. If the monomer employed has the chemical structure, to which the existing free radicals can add on, the propagation via free-radical chain-growth polymerization occurs during the off-period.[5] In such a case, the pulsed discharge yields plasma coating that retains the original chemical structure of the monomer in higher degrees depending on the ratio of off-/on-period.

If the monomer contains functional groups of which one acts as an inhibitor of plasma polymerization (e.g., carboxylic acid), and another promotes the chain-growth polymerization (e.g., double bond), such as the case of acrylic acid, the net effect is seen as the greatly enhanced deposition rate because the poisoning effect that is caused by the fragmentation of the oxygen containing group in the plasma does not occur in the off-period. If the monomer does not contain these functional groups, the net effect is the reduced deposition rate more or less proportional to the fraction of on-time.

If the monomer contains a double bond or a triple bond, the concentration of dangling bonds in plasma polymerization coating increases significantly, because the cross-cycle reaction, shown as Eq. (3) in Fig. 4, ceases during the off-time and retains more bifunctional free radicals in the reaction system.[5] The increased dangling bonds could have a positive or negative effect on the coating depending on the objective of plasma polymerization coating.

While the pulsed discharge has the great advantage of retaining some monomer functional groups in the plasma polymer, in nearly all cases, it shortens the kinetic path length of plasma polymerization, which reduces the polymer characteristics of plasma polymerization coating and tends to reduce the adhesion of the coating. In other words, pulsed discharge tends to yield more or less conventional polymers at the expense of unique characteristics of plasma polymers. The advantage or disadvantage obtainable by the use of pulsed discharge entirely depends on the objective of plasma polymerization coating, particularly with respect to the types of plasma polymerization coating, which are described in a later section.

Microwave Discharge

In microwave discharge (frequency in the GHz range), the electromagnetic energy propagates on the surface, and the deposition yield on the substrate could be very high if the substrate surface could be used as the main energy propagation surface of microwave discharge. The waveguide is used externally similar to the external energy input in RF discharge. The coating of the inner surface of a plastic bottle by a microwave discharge is an example of the perfect utilization of microwave discharge in plasma polymerization coating. If numerous other substrates are placed in a bottle, however, the main energy propagation pattern is disturbed, and very uneven erratic coating occurs.

Magnetron Discharge

An electrode system equipped with a superimposed magnetic field is generally referred to as "magnetron."[12] The magnetron has been widely used in sputter deposition of metals and alloys; however, its use in plasma polymerization has been relatively uncommon.[13–15] The main use of magnetron is to

sputter the target material placed on the cathode for the subsequent physical deposition of the sputtered materials. In many applications of plasma polymerization, on the other hand, the inclusion of electrode material in the polymeric deposition has been viewed as contamination.

In certain applications of plasma polymerization, the incorporation of electrode material, particularly in a controlled and designed manner, by means of a magnetron discharge is extremely useful and becomes a great asset in plasma polymerization coating. For instance, a thin layer of plasma polymer of methane with a tailored gradient of copper has been shown to improve the adhesion of the thin layer to a copper substrate and also the adhesion of metal to a polymer film.[16,17] In general, applications of plasma polymerization coating, in which the metal contamination should be avoided, are important to select the electrode material that has low sputtering yield. Titanium has been used successfully in such cases.

The use of magnetron glow discharge for the formation of thin films by plasma polymerization has several obvious advantages over the use of glow discharges that do not have magnetic enhancement.[13] Perhaps, the most obvious advantage is the confinement of glow discharge in the lower-pressure regime of plasma polymerization, i.e., polymer deposition can be carried out at a system pressure <100 mtorr. Without a magnetic confinement of glow discharge, the glow expands as the system pressure decreases, often to the entire volume of a reactor. This expansion of glow discharge occurs even when small internal electrodes are used in a relatively large bell jar reactor. In this case, the deposition of material onto the target substrates becomes very inefficient, and the excessive deposition onto the reactor wall practically prohibits the use of such a reactor for the continuous coating of moving substrate in a large-scale operation.

Perhaps, the most important, but not so obvious, aspect of the magnetron glow discharge is that plasma polymerization coating can be performed in a low-pressure regime in which plasma polymerization by a nonmagnetron glow discharge is difficult or impossible. The magnetic enhancement lowers the overall impedance of glow discharge, and lowers the breakdown voltage.[13,18] Hence, the operation of glow discharge in the low-pressure regime becomes possible and practical. The low-pressure operation is favored in obtaining better barrier characteristics of plasma polymerization coating with larger kinetic path length.

Plasma polymerization in a low-pressure regime inevitably reduces the deposition rate of the process in the absence of the magnetic enhancement. However, the deposition rate obtainable by magnetron discharge is several folds higher than that obtained by nonmagnetron discharge, because the magnetron discharge effectively shifts the dissociation glow from the electrode surface to the toroidal glow. The reasonable deposition rates could be obtained in low-pressure regime by employing magnetron discharge. The relatively large-scale industrial applications (outside of microelectronic applications) of glow discharge polymerization employed the magnetron glow discharge.[19,20] An important advantageous factor of magnetron discharge, which is difficult to be appreciated or recognized by laboratory-scale experiments, is the minimum deposition on surfaces other than that of the substrate, i.e., the minimum reactor-wall contamination.

MODES OF OPERATION FOR PLASMA POLYMERIZATION COATING

Operational parameters that control plasma polymerization can be divided into two major categories: 1) characteristic parameters of a reactor, which can be altered but are not variable in each run, and 2) parameters that require adjustment for each run and often during a run. Size of the electrodes, distance between electrodes, and frequency of electric power are examples of parameters of the first category. Monomer flow rate, system pressure, and discharge power are examples of operational parameters of the second category. The parameters of the first category are important in designing a plasma polymerization reactor, but the parameters of the second category are critically important in executing plasma polymerization coating to produce the desired product under the restrictions set by the design of reactor.

Batch Operation

In a batch operation, substrates are placed in a reactor, and plasma polymerization coating is carried out as a unit operation. Repeating the same operation treats a large number of substrates. The batch processing is the primary mode for nearly all laboratory-scale operations. The batch processing can be done in a closed system or in a flow system. Because the number of molecules in a reactor under low pressure is small, it is often necessary to use a flow system to obtain a sufficient amount of coating.

The gas phase changes on the onset of glow discharge. The fragmentation of a molecule increases the total number of gas molecules in a system, and the polymerization or deposition of material decreases the total number of gas molecules. The balance between these two opposing reactions shows in the change of pressure of a flow system, if a throttle valve does not control the system pressure. In a batch operation,

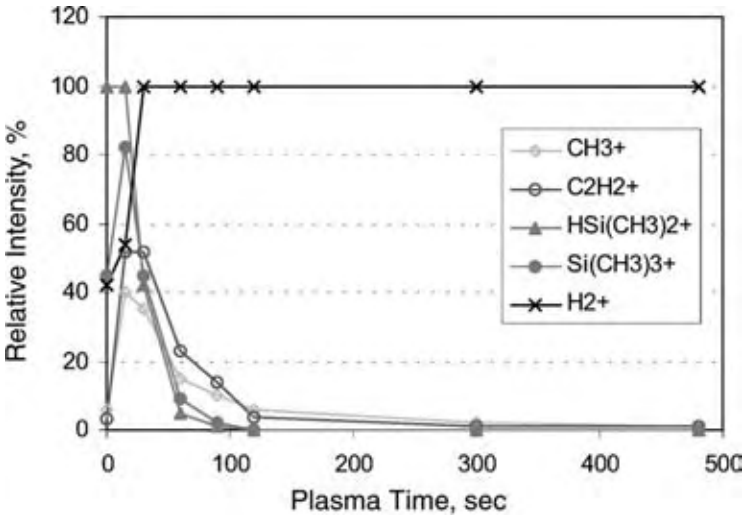


Fig. 14 Change of gas phase species in plasma of TMS with plasma time. Plasma conditions are: 25 mtorr, TMS, two panels of aluminum alloy, DC 1000 V. *(View this art in color at www.dekker.com.)*

therefore, steady-state plasma is not generally established in a relatively short operation time.

Closed System Operation

Very efficient plasma polymerization coating can be performed in a closed system reactor depending on the size of the reactor and the specificity of the deposition on the substrate. For instance, the deposition in DC discharge occurs mainly to the cathode surface and the closed system operation of DC plasma polymerization coating has a distinctive advantage that the total mass in the system is well defined and does not change by the electrical discharge, and high $Y_{p,s}$ can be achieved.

Fig. 13 depicts the change of system pressure during a closed system plasma polymerization of TMS. Fig. 14 depicts the change of gas composition during the same operation of closed system plasma polymerization coating. The easily polymerizable species disappears in the early stage of operation and after a certain time of operation the major gas phase in the reactor becomes hydrogen plasma.

The change of gas composition with the reaction time in a closed system leads to a composition-graded plasma coating, which is a great advantage in plasma polymerization coating of metals yielding excellent adhesion to the substrate as well as to the conventional coating that is applied on the plasma coating.[21] Fig. 15 depicts the comparison of X-ray photoelectron spectroscopy (XPS) cross-sectional profiles of TMS plasma polymer coatings. The coating prepared by a flow system operation shows a uniformly flat profile, whereas the coating prepared by a closed system operation shows a continuously changing profile, from Si-rich at the metal/coating interface to C-rich top surface of plasma polymerization coating.

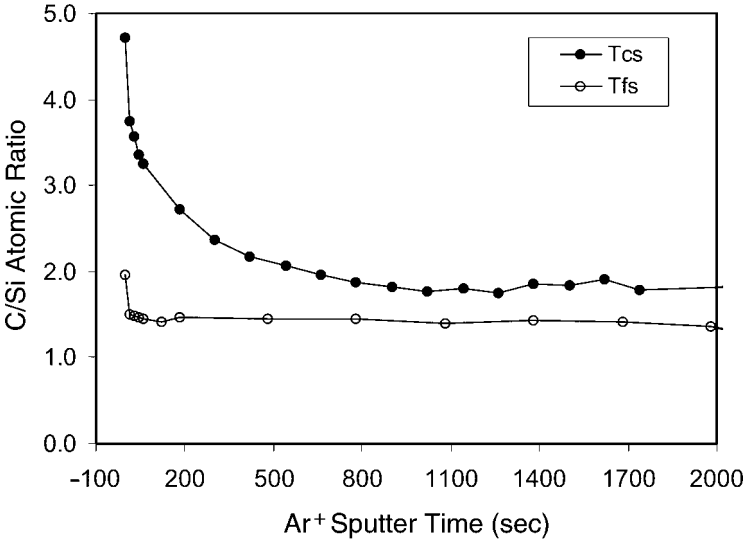


Fig. 15 C/Si ratios of plasma polymer films of TMS prepared in a flow system reactor (T_{fs}) and in a closed system reactor (T_{cs}), as generated by XPS depth profiling. *(View this art in color at www.dekker.com.)*

Continuous Operation

In the continuous processing, a steady-state flow of luminous gas is established and maintained for the duration of operation, e.g., 1 mo, without interruption. Because of the factors described above, it takes some time, e.g., 30 min, to establish a steady-state flow of luminous gas.[22] Once a steady-state is established, it is possible to maintain it for a sufficiently long period of time for a continuous processing. Substrates are fed into the steady-state flow of luminous gas in a cross-flow pattern. The rate of transport of substrate and the length of the path in the luminous gas phase determine the treatment time.

It is common that a substrate or a section of continuous substrate, such as fibers and films, passes multiple sets of luminous gas flow. Fig. 16 depicts a schematic diagram of continuous operation of plasma polymerization with two sets of luminous gas flow. A chamber that is pumped individually to avoid cross-contamination separates the two sets of luminous gas flow. Thus, multiple plasma polymerizations or treatments can be applied on a substrate according to this principle. A continuous substrate, such as fibers, tubes, and films, is fed vertically. The horizontal feeding of substrates, shown in Fig. 16, requires multiple substrates holding devices that travel through the reactor.

In a continuously operated flow system, those factors associated with a closed system or a batch-operated flow system mentioned above are virtually eliminated except at the very beginning of the operation. Therefore, the reproducibility of plasma polymerization obtained by continuous operation is superior to that obtained by the batch operation.

To achieve a continuous operation of plasma polymerization, it is necessary to consider the following factors: 1) deposition on electrodes; 2) deposition on reactor wall; and 3) flaking off of the deposition from these surfaces. Magnetron discharge confines the volume of luminous gas phase, which minimizes the deposition onto the reactor wall. It is a general practice to place removable liners on the walls of the reactor, which will be removed and replaced by new liners on each maintenance stoppage of the operation.

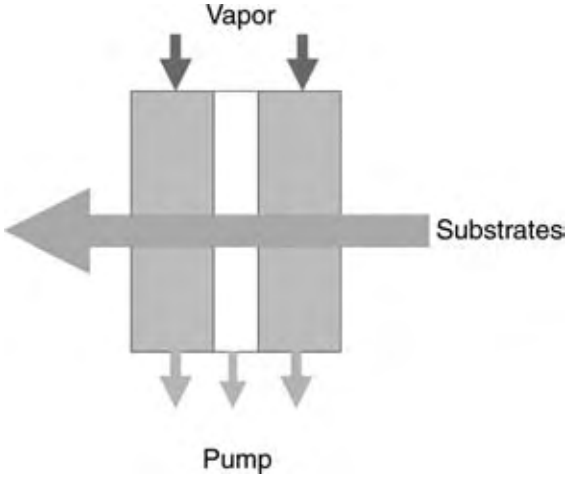


Fig. 16 Schematic representation of continuous operation of plasma polymerization coating. *(View this art in color at www.dekker.com.)*

GENERAL CHARACTERISTICS OF PLASMA POLYMERIZATION COATINGS

Plasma polymerization coating is system dependent and a monomer does not yield a well-defined polymer that can be identified by plasma polymer of the monomer. Plasma polymers formed at the high *W/FM* end of the power-deficient domain and also in the monomer-deficient domain are tight three-dimensional amorphous networks, which do not contain discernible functional groups (Type A plasma polymers). Plasma polymers formed at the low *W/FM* end of the power deficient domain could contain functional groups in the monomer, but the structure is much looser and often consists of oligomeric deposition (Type B plasma polymer). The coating that has a tight network structure with functional groups cannot be obtained by a single-step plasma polymerization of a monomer. The retention of functional groups of a monomer by some efforts, such as pulsed discharge, remote-plasma deposition, etc., can be achieved at the expense of the unique characteristics of the Type A plasma polymers, such as the good barrier characteristics and good adhesion to the substrate.

Type A plasma polymers have the characteristic internal stress built in the film, and the plasma polymerization coating acts as the tempered ultrathin layer on the substrate. The internal stress is caused by the wedging effect of the deposition process, and the total stress increases linearly with the thickness. As the thickness increases, the internal stress reaches the critical point beyond which the internal stress becomes greater than the cohesive force or the adhesive force of the plasma polymerized coating. Above the critical thickness, therefore, the coating cracks (not necessarily in macroscopic sense) or delaminates (buckles) from the substrate. Consequently, there exists a thickness limit of plasma polymerization coating. The tighter the structure, the smaller is the thickness limit.

The barrier characteristic of a thin layer is proportional to the value of (thickness)/(permeability) or (thickness) (transport resistance). Type B plasma polymers have a higher thickness limit but permeability is high. Although a thicker layer could be deposited

without creating cracks with Type B plasma polymers, the layer does not provide a good barrier because of the higher permeability due to the looser structure. Type A plasma polymers provide remarkably high barrier characteristics in spite of the small thickness limit. The barrier characteristics, however, decrease drastically as the coating thickness increases beyond the critical thickness, although the coating might visibly appear as an intact layer.[23]

A well-executed Type A plasma polymer coating is insoluble and infusible, and the surface is imperturbable to any change in the surrounding medium. Type B plasma polymer coatings, however, are often soluble in various solvents (including water) and, consequently, the surface is highly perturbable and the adhesion to the substrate is poorer than in Type A plasma polymer coatings. The frequently mentioned characteristics of plasma polymers, such as high cross-linking, insolubility and infusibility, stable network, etc., do not apply to Type B plasma polymer coatings.

CONCLUSIONS

Plasma polymerization coating is a unique nanofilm coating technology that yields a tight amorphous network structure of atoms contained in the monomer (Type A plasma polymers). Because of the unique growth mechanism of plasma polymerization, excellent adhesion of coating could be obtained in a practical manner.

Depending on the level of energy input, much loose structure with functional groups (Type B plasma polymers) could also be obtained; however, this type of coating does not share the unique characteristics of Type A coatings.

Type A plasma polymerization coatings could be the ultimate green processes that use the minimum amount of monomer and do not yield effluent. That environmental remediation processes are not required could more than offset the initial equipment cost in the overall cost of manufacturing.

REFERENCES

1. Tao, W.H.; Prelas, M.A.; Yasuda, H.K. Spatial distributions of electron density and electron temperature in direct current glow discharge. J. Vac. Sci. Technol. A. **1996**, *14*, 2113.
2. Yasuda, H.; Yu, Q. Molecular dissociation glow in plasma polymerization of trimethylsilane (TMS). J. Vac. Sci. Technol. **2004**, A *22* (3), 472–476.
3. Yasuda, H.; Yu, Q. Creation of polymerizable species in plasma polymerization. Plasma Chem. Plasma Process **2004**, *24*, 325–351.
4. Yasuda, H.K.; Yu, Q.S. Deposition of trimethylsilane in glow discharges. J. Vac. Sci. Technol. A **2001**, *19* (3), 773–781.
5. Yasuda, H. *Plasma Polymerization*; Academic Press: Orlando, FL, 1985.
6. Chen, M.; Yu, Q.S.; Reddy, C.M.; Yasuda, H.K. Model study investigating the role of interfacial factors in electrochemical impedance spectroscopy measurements. Corrosion **2000**, *56* (7), 709–721.
7. Kobayashi, H.; Bell, A.T.; Shen, M. J. Macromol. Sci. Chem. **1976**, *10*, 491.
8. Yasuda, H.; Wang, C.R. Plasma polymerization investigated by the substrate temperature dependence. J. Polym. Sci. Polym. Chem. Ed. **1985**, *23*, 87.
9. Miyama, M.; Yasuda, H. Direct current cathodic glow discharge polymerization of methane and butane. J. Appl. Polym. Sci. **1998**, *70*, 237–245.
10. Yu, I.; Yasuda, H. J. Appl. Polym. Sci. Appl. Polym. Symp. **1988**, *42*, 97.
11. Yu, Q.S.; Yasuda, H.K. Effects of cathode and anode on deposition of trimethylsilane in glow discharge. Plasma Polym. **2002**, *7*, 41–55.
12. Vossen, J.L.; Kern, W. *Thin Film Processes*; Academic Press: New York, 1978.
13. Sato, K.; Iriyama, Y.; Cho, D.L.; Yasuda, H. Plasma polymerization by magnetron glow discharge. I. Effect of magnetic field on breakdown of monomers in low pressure. J. Vac. Sci. Technol. A **1989**, *7* (2), 195.
14. Cho, D.L.; Yeh, Y.S.; Yasuda, H. Plasma polymerization by magnetron glow discharge. III. Effect of magnetic field on sputtering characteristics of electrode materials. J. Vac. Sci. Technol. A **1989**, *7*, 2960.
15. Sato, K.; Yeh, Y.S.; Yasuda, H. Plasma polymerization by magnetron glow discharge. II. Effect of magnetic field on chemical characteristics of plasma polymers of tetrafluoroethylene revealed by electron spectroscopy for chemical analysis. J. Vac. Sci. Technol. A **1989**, *7*, 3188.
16. Sun, B.K.; Cho, D.L.; O'Keefe, T.J.; Yasuda, H. Chemically graded metalization of nonconducting substrates by glow discharge plasma polymerization technique. In *Metallized Plastics 1*; Mittal, K.L., Susko, J.R., Eds.; Plenum Press: New York, 1989.
17. Yasuda, H.; O'Keefe, T.J.; Cho, D.L.; Sun, B.K. Mettallization of nonconducting substrates by

means of plasma polymerized transitional buffering films. J. Appl. Polym. Sci. Appl. Polym. Symp. **1990**, *46*, 243.

18. Yasuda, H.; Olcaytug, F. A study on electromagnetron for plasma polymerization. J. Vac. Sci. Technol. A **1991**, *9* (4), 2342.
19. Ross, D.L. Coatings for video discs. RCA Rev. **1978**, *39*, 136.
20. Leahy, M.F.; Kaganowicz, G. Solid State Technol. **1987**, *30* (4), 99.
21. Yu, Q.; Moffitt, C.E.; Wieliczka, D.M.; Yasuda, H. DC cathodic polymerization of trimethylsilane in a closed reactor system. J. Vac. Sci. Technol. A **2001**, *19* (5), 2163–2167.
22. Yasuda, H.; Inagaki, N. The initial and terminating stage of glow discharge polymerization investigated by thickness monitor. J. Appl. Polym. Sci. **1981**, *26*, 3557.
23. Kramer, P.W.; Yasuda, H. Effect of operational parameters on the air separation properties of composite hollow fiber membranes prepared by plasma polymerization of perfluorodimethyl cyclobutane. J. Appl. Polym. Sci. Appl. Polym. Symp. **1988**, *42*, 381.

Pollution Prevention

Ashok Kumar
Department of Civil Engineering, University of Toledo, Toledo, Ohio, U.S.A.

Harish G. Rao
LFR Inc., Elgin, Illinois, U.S.A.

Abhilash Vijayan
Charanya Varadarajan
Department of Civil Engineering, University of Toledo, Toledo, Ohio, U.S.A.

INTRODUCTION

During the 1960s and the early 1970s, cleaning up wastes created as a result of industrial and social activities was the foremost concern of environmental engineers and scientists. Several strategies were implemented to help clean or treat air emissions as well as contaminated soil and water resulting from activities dating as far back as the industrial revolution. In the early 1980s, the notion of protecting the environment from the effects of pollution-causing activities gained momentum. Environmental professionals and activists in the United States and Europe realized that the traditional approach of treating the waste after it was created (or end-of-pipe treatment) and cleaning up contaminated soil or groundwater (i.e., remedial treatment) to meet government standards were less desirable from societal, technical, and financial perspectives. A shift toward an approach that favored prevention of pollution, while paving the way for process efficiency and improved process design, was realized as the need of the day. The importance of preventing pollution at source, stressing the need for reduction in waste generation and associated pollution rather than treating the waste produced at the end of the process, then became recognized as a key component of sound environmental management. This along with the methodology for more efficient use of resources allows our ecosystems to meet and sustain the growing demands of human societies.

In 1990, with the passage of the National Pollution Prevention Act (PPA), the U.S. Environmental Protection Agency (USEPA) decided to make pollution prevention or P2 a top priority. In the United States, P2 includes source reduction and other practices that reduce or eliminate the creation of pollutants through increased efficiency and conservation of natural resources. Treating wastes was considered important, but efforts to prevent wastes from being generated was the priority. This entry provides an overview of the concept of P2. It is an attempt to offer an insight into the basic P2 principles and details the P2 procedure.

POLLUTION PREVENTION (P2)

P2 over the years has evolved from a concept to a proven technology to become a stand-alone and mainstream channel for eliminating the adverse effects of industrial and manufacturing activities on the environment or the ecosystem. The U.S. P2 initiatives originated during 1976–1979 with the adoption of the Pollution Prevention pays (3P) program by the 3M Corporation. North Carolina adopted waste minimization as a statewide priority for managing emissions from industry at around the same time. Most states followed suit, and were requiring P2 programs for industry, by the mid-1980s. By 1990, the Fortune 1000 U.S. corporations had adopted P2 as a primary policy. The shift in priorities in addressing the environmental issues kindled the transition from nearly 20–50 yr of conventional pollution control to the more proactive, preventive approach. Fig. 1 is an approximate timeline of the period from 1976 to 1996 when there were distinct explosions of success in P2 activities.[1–4] For detailed information on pollution prevention, see Bishop, Higgins, and Patel and Kumar.[5–7]

Definition

P2 is a term used primarily in the United States but a host of other terms with functional similarity are used globally and represent a globally accepted movement for improving business performance and providing a vision of a profitable, cleaner, and sustainable future focusing on a strategy of continuously reducing pollution and adverse environmental impact through source reduction—that is, eliminating waste at the source rather than treating it at the end of the process (end-of-pipe treatment).

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120006820

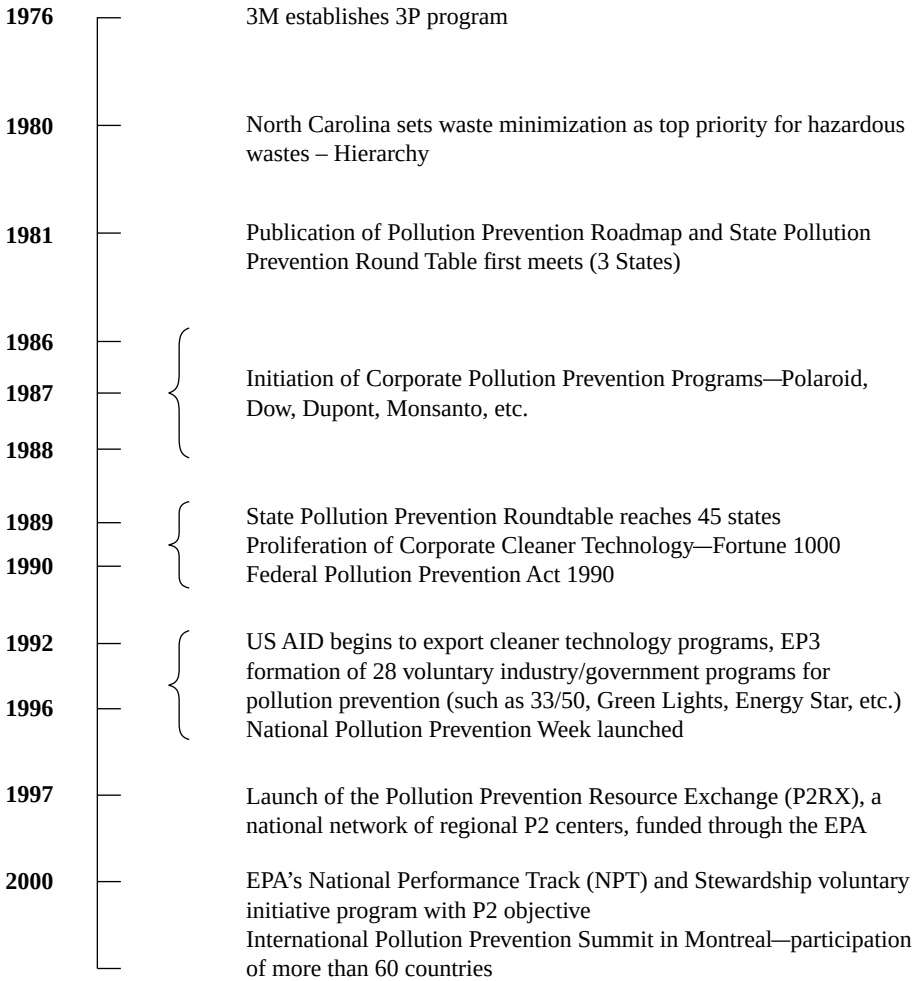


Fig. 1 Timeline of early P2 successes.

In the United States, the 1990 PPA was the first major national policy initiative aimed at promoting P2 as a fundamental component of environmental protection efforts. Under Section 6602 (b) of the 1990 PPA, Congress established a national policy that prioritized P2 over other means of reducing pollution. The policy and definitions discussed below are adapted from the USEPA's pollution prevention documents.[8–10] The policy highlighted the following aspects:

- Pollution should be prevented or reduced at the source whenever feasible.
- Pollution that cannot be prevented should be recycled in an environmentally safe manner whenever feasible.
- Pollution that cannot be prevented or recycled should be treated in an environmentally safe manner whenever feasible.
- Disposal or other release into the environment should be employed only as a last resort and should be conducted in an environmentally safe manner.

Fig. 2 is a representation of this hierarchy of priorities for managing waste.

The 1990 PPA defines ''pollution prevention'' mainly as ''source reduction,'' but it also includes other activities that reduce or eliminate the creation of pollutants through:

- Increased efficiency in the use of raw materials, energy, water, or other resources.
- Protection of natural resources by conservation.

The Act defines ''source reduction'' as any practice that

1. Reduces the amount of any hazardous substance, pollutant, or contaminant entering any waste stream or otherwise released into the environment (including fugitive emissions) prior to recycling, treatment, or disposal.
2. Reduces the hazards to public health and the environment associated with the release of such

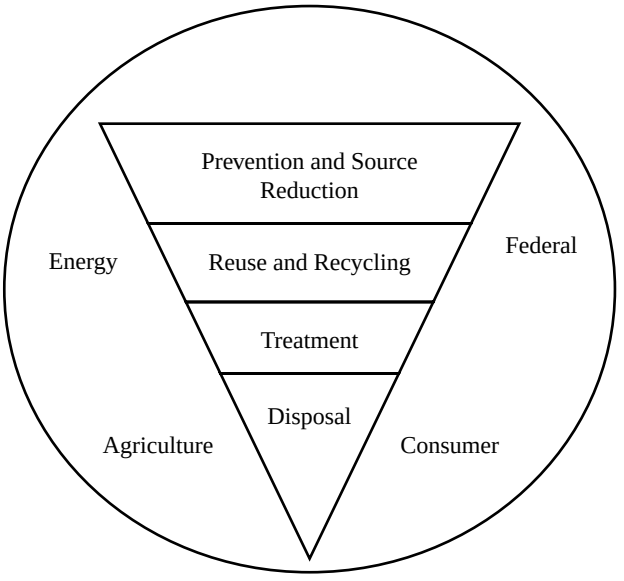


Fig. 2 Waste management structure for pollution-generating activities.

> substances, pollutants, or contaminants. Source reduction is achieved by equipment or technology modifications, process or procedure modifications, reformulation or redesign of products, substitution of raw materials, and improvements in housekeeping, maintenance, training, or inventory control. However, it does not include any practice which alters the physical, chemical, or biological characteristics or the volume of a hazardous substance, pollutant, or contaminant through a process or activity which itself is not integral to and necessary for the production of a product or the providing of a service.

Thus P2, as defined in the PPA, is a distinct approach that does not consider out-of-process recycling as a form of prevention, even though it can bring about environmental improvements and greater efficiency in resource use and conservation. This distinction is still debated between regulators and industry proponents, who assume recycling to be an equivalent of P2. In accordance with the hierarchy of preferred waste management options specified in the PPA. U.S. Environmental Protection Agency (EPA) has interpreted waste recycling is not eliminating the creation of pollution at the source, thereby posing a potentially higher hazard to workers, the environment and public health than the P2 management option and does not view recycling as synonymous with P2. However, in some cases, inprocess recycling is considered a form of P2, wherein the waste generated from a process is directly reincorporated into the same process.

As shown in Fig. 2, P2 approaches can be applied to all pollution-generating activities, including those found in the energy, agriculture, federal, consumer, as well as industrial sectors. P2 does not include practices that create new risks of concern.

Terms Synonymous with P2

P2 is a term widely known and accepted in North America, but is one of the many synonyms that are used in different parts of the world targeting environmental improvements and resource management. Some of these approaches that have functional similarity to P2 include:

Waste Minimization: Waste minimization is any effort to reduce or recycle the quantity of waste generated, and when feasible, to reduce or eliminate toxicity. It includes any source reduction or recycling activity undertaken to reduce the total volume or quantity of hazardous waste and/or the toxicity of hazardous waste, with the goal of minimizing threats to human health and the environment. Waste minimization focuses exclusively on solid wastes regulated under the Resource Conservation and Recovery Act (RCRA), and does not include treatment, unless the treatment is part of the recycling process. However, it is broader in scope than P2 to also include recycling and other means to reduce the amount of waste that must be treated or disposed of off- or onsite.

Sustainable Development: ''Sustainable development meets the needs of the present without compromising the ability of future generations to meet their own needs.'' P2 plays an important role in achieving the goal of sustainability as it emphasizes the efficient use of materials and resources with an environmental aspect added to it.

Best Management Practices: Best management practice is a process, technique, or innovative use of resources that has a proven record of success in providing significant improvement in cost, schedule, quality, performance, safety, environment, or other measurable factors that impact the health of an organization.

Cleaner Production: Cleaner production is the continuous application of an integrated preventive environmental strategy to processes, products, and services to increase overall efficiency and reduce risks to humans and the environment. Cleaner production can be applied to the processes used in any industry, to products themselves, and to various services provided in society.

Design for Environment: Design for environment means designing products and services with the environment in mind. It aims to minimize adverse environmental impacts throughout the entire life cycle of a product and integrates cleaner, cheaper, and smarter designs and solutions into everyday business practices.

Eco-efficiency: Delivery of competitively priced goods and services that satisfy human needs and bring quality of life, while progressively reducing ecological impacts and resource intensity throughout the life cycle, to a level at least in line with the earth's estimated carrying capacity. The World Business Council for Sustainable Development developed the concept of eco-efficiency to look at cost optimization in the context of positive environmental benefits.

Energy Conservation: Actions taken to reduce or more efficiently use energy, in an effort to preserve the environment and avoid depletion of energy resources.

Green Engineering: Green engineering is the design, commercialization, and use of processes and products that are both feasible and economical while minimizing 1) generation of pollution at the source and 2) risk to human health and the environment.

Green Chemistry: Green chemistry is the use of chemistry for pollution prevention, i.e., the design of chemical products and processes that reduce or eliminate the use and generation of hazardous substances. The EPA's Green Chemistry Program promotes the research, development, and implementation of innovative chemical technologies that prevent pollution in both a scientifically sound and a cost-effective manner.

All these concepts focus on improving the environmental condition. Use of a specific terminology is often linked to the particular forum in which the debate or discussion is occurring. Although these terms have subtle differences, they do share the major emphasis on prevention. For more detailed information on any of the concepts mentioned earlier, see Refs.[2,3,8,11–18].

Need for P2

The emphasis on pollution prevention is to deal with the present and future pollution. While pollution control is necessary to deal with contamination that has already occurred, the prevention of further contamination is more important. Environmental managers with the task of controlling pollution need an acceptable and measurable reason for exploring and eventually implementing a P2 project. The need for P2 implementation is chiefly to protect the environment while saving money and to save money while protecting the environment. The major advantages and benefits of developing and implementing a P2 project are as follows:

Reduced Compliance: Undertaking P2 projects reduces regulatory exposure and, in some cases, eliminates the need for expensive and time-consuming permits, manifesting, monitoring, and reporting. This has been a major thrust for waste reduction in industries.

Lowered Operating Cost: Pollution prevention activities often save money in the long run in terms of new material costs, waste treatment and disposal costs, and energy and operations costs. Many P2 projects have good returns on investment and short payback periods. Some P2 opportunities, such as reduction in toxic substances, can help reduce the costs of medical claims and disability leave. Costs associated with compliance and waste handling are also avoided if the waste or pollutant is prevented from being generated in the first place.

Enhanced Worker Health and Safety: Reduced use of toxic substances in the workplace is a major aspect of pollution prevention that improves air quality, water quality, and the safety of the work environment by reducing the likelihood of leaks, spills, and releases. These steps result in cost savings through preventing the loss of materials and decreased insurance rates by reducing medical claims and disability leave. Better labor relations also result from improved worker safety.

Increased Productivity: Pollution prevention improves plant productivity through more efficient use of raw materials, labor, equipment, and energy because of improved processes and operations. Small improvements to old and inefficient technologies and improved quality control of processes result in increased product yield and better quality.

Improved Environmental Quality and Reduced Future Liability: Reduction in waste discharges to the environment is the principal focus and benefit of P2 programs. However, many legal waste disposal and treatment methods often move environmental contaminants from one medium to another, thus resulting in environmental damage. This could result in future liability problems. Pollution prevention involving efficient methods and processes results in less waste, thus ensuring improved environmental protection and reduced long-term liability potential.

Improved Company Image and Community Relations: The USEPA publishes details of companies' waste and pollution prevention efforts through the Toxic Release Inventory. With a society aware of the environmental hazards associated with the all types of waste, it becomes increasingly difficult for big polluters to maintain an ongoing relationship with the community. The USEPA also publicly recognizes companies that make voluntary commitments to P2, so as to enhance their image in the community.

The section is adopted from the Ohio Pollution Prevention and Waste Minimization Planning Guidance Manual.[19]

METHODS FOR P2

Wastes could be reduced for any industry by a systematic analysis of processes to study the inputs

(materials, energy, and people) and activities creating wastes, which in turn requires the combined effort of the production, administration, and environmental specialist teams. One or more of the following methods can accomplish P2:[20–22]

- *Design changes*: Incorporation of environmental elements in the design of a product makes it more resource efficient and sustainable over the life cycle. Product redesign can involve extending the product's life span or enabling it to be completely recycled into a new product.
- *Process modification*: Process modifications involving reconfiguring the different steps in a process often help reduce production waste, cutting both pollution and waste management costs. This method of preventing pollution can involve reducing the total materials and energy used to manufacture a product. For example, a car wash company could reduce its rinse cycle by 5 sec, thereby conserving water and energy. Similarly, a car parts manufacturer could reduce one's paint use by utilizing more efficient painting techniques or machinery.
- *Materials substitution and product reformulation*: Use of alternative materials for production processes that can provide equivalent results while preventing costly hazardous waste generation, air emissions, and worker health risks is a sound approach to P2. Products containing toxic ingredients find their way into the ecosystem. Air, water, and soil pollution can be considerably reduced if the toxic substances are replaced with nontoxic ones. Product reformulation or redesigning a product may involve switching to new ingredients, but also includes a consideration of reducing packaging or reducing the amount of "ingredients" that go into a finished product. Reducing the amount of ingredients conserves not only resources but also energy.
- *Materials reuse*: Wastes generated from one process or industry can become the raw material for another. Materials reuse goes a long way in preventing pollution by: 1) reducing waste that needs to be disposed; 2) conserving resources; and 3) lessening the cost of materials.
- *Resource efficiency and conservation*: Using energy, water, raw materials, and other production inputs more efficiently conserves the resources apart from reducing pollution, improving productivity, and lowering the operational costs. Cutting down on the amount of water and energy use in day-to-day operations can result in reductions in the amount of greenhouse gases, particulate matter, and other pollutants that may be released into the atmosphere and into water discharges from a facility.
- *Good operating and work practices*: Rethinking day-to-day operations and maintenance activities can lead to P2 that can have a positive impact on the environment and uproot wasteful management practices that drive up costs and cause pollution. For example, volatile organic solvent emissions from a washer can be prevented during off hours if the washer is covered with a lid. Good operating practices such as: 1) inventory control 2) supply chain management; 3) preventive maintenance; 4) improved housekeeping; 5) employee training; 6) facility management; 7) material segregation, etc. play an important role in achieving P2. Educating staff on environmental issues is a very crucial work practice that help them practice P2 while carrying out their tasks. For example, employees can learn to prevent spills and leaks, to turn lights and equipment off when they are not being used, or to cut down on the amount of paper used in an office setting.
- *New clean technologies*: There is no real limit to the opportunities to practice P2, and scientists and engineers are constantly developing new technologies that enable humans to minimize adverse impacts on humans and the Earth's ecosystems that support all life. Switching to renewable energy sources to heat our homes or power our vehicles is an example of the use of a cleaner technology that can prevent pollution.

The four P2 case studies given below can provide ideas on how each of these methods can be used, depending on the nature of the operation and the goal of the P2 program.

These case studies were conducted during 2001–2003 for Ohio industries with the help of the National Institute of Standards and Technology Centers in Ohio, under a USEPA grant. Detailed information on these and other case studies is given in a report by Kumar et al.[23]

CASE STUDY 1: POLLUTION PREVENTION AT A METAL COATING FACILITY

P2 Goal

This study involved the pollution prevention assessment of a metal parts coating unit to bring about pollution prevention through waste reduction by making process changes, while maintaining the same costs.

Procedure

The study began with isolating sources of waste and pollution. This was followed up with suggestions to alleviate problems.

Sources of waste

Coating Various Customer Parts. Metal parts of various configurations are shipped to a metal coating facility (MCF) by customers for coating with various paints according to customer specifications. The customers of MCF have prepared the parts for painting, so no additional processing is necessary before painting is begun.

Each part is hand placed on an overhead chain conveyor that dips down into a trough with the appropriate paint. The parts are immersed for a few seconds to properly cover the part with the paint. As the time for immersion is completed, the part moves out of the paint trough, up an incline to a heated area for drying. When the part moves out of the trough, any excess paint will drip off onto a piece of plastic sheet. This paint currently dries on the plastic or accumulates into a small holding container, at the end of the incline, which is recovered. Plastic sheets are reused by color for other runs of the same color paint. Currently, paint dries on the sheets and can easily evaporate out of the small holding container.

Paint Sludge From the Dipping Tanks Paint used to coat the parts is adjusted by adding xylene to maintain proper viscosity. There is no mechanical agitation on the tank. Because of the configuration of the tank (low and long), it is difficult to provide manual agitation. The tank is stirred by hand whenever the xylene is added. This results in some settling of the paint, as mechanical agitation is more efficient than hand agitation. This settled paint ''sludge'' is placed in a drum for offsite disposal as a hazardous waste (Table 1).

Opportunities to Reduce Waste

Coating various customer parts

Careful consideration should be given to the placement of the parts on the hooks so that adequate paint coverage and rapid draining of excess paint are facilitated. This will require that each part be carefully analyzed to achieve these objectives.

The chain conveyor structure can be modified at a point where it leaves the dip tank to facilitate drainage. A sharp upward turn followed by a downward turn may increase the speed at which the excess paint drips off the part.

Paint sludge from the dipping tanks

As the paint drips off the parts in the line beyond the tank perimeter, it is directed toward an open top container. The open areas of the paint both in the painting tank and in the small holding container allow for airflow and drying. The surface area of the small holding container directly exposed to the air can be reduced by providing a lid that has a hole which is not larger than 2 in. in diameter and is tapered into the container. This will allow the paint to accumulate in the container rather than on the lid. Additionally, if the paint in the container has formed sludge in the bottom, thinner may be added along with mixing to decrease the viscosity of the sludge and, by careful adjustments of the thinner, may allow it to be reused as good raw material. Mechanical mixing is the preferred method for this activity. This should be a slow speed agitation that provides for mixing of the entire contents of the tank.

Expected Benefits and Results

It is estimated that emissions can decrease by up to 30% (from 42,774 to 29,942 lb/yr) with the implementation of these suggestions.

The hazardous waste issues will decrease with the reduction of the hazardous waste generation below 2200 lb/mo from 4527 lb/mo. By use of the previous recommendations and with the possibility that sludge may be recovered with thinner, the reduction can be significant. This will reduce the burden on the facility as well as the cost for paint and hazardous waste disposal. The cost reduction for hazardous waste disposal is estimated to be $25,550/yr. This cost reduction is affected by the reduction in evaporated paint and reduced sludge loss.

CASE STUDY 2: WASTE MINIMIZATION ASSESSMENT AT CLIENT COMPANY

P2 Goal

This project was intended to determine the best methods to help the client reduce the hazardous and

Table 1 Hazardous waste generated

Waste	Amount generated (£)	Amount reclaimed (£)	Amount disposed of (£)	% Reclaimed
Liquid pourable paint	22,200	0	22,200	0
Solid debris and liquid paint	20,300	0	20,300	0
3-yd boxes of paint sludge	11,820	0	11,820	0
Total	54,320	0	54,320	0

nonhazardous waste generated by the production operations.

Procedure

The facility waste reports, tours, and discussions helped gather data on the type and amount of generated waste associated with each manufacturing process.

Sources of Wastes

Nonhazardous waste

The client generated approximately 905.5 tons of solid nonhazardous waste in 2000. Only 49% of the generated waste was recycled, and the rest was sent to a landfill. Table 2 summarizes the solid waste generation/disposition.

Avenues for Nonhazardous Waste Minimization and Savings

1. Some of the raw material waste is from quality control problems when the mix is in the wet stage. Keeping the raw material in closed drums and by avoiding cross-contamination, it can be worked back into the manufacturing process within a 3–4 mo period, thereby avoiding generation of waste.
2. To respond to customer orders on a timely basis, the facility maintains a huge stock of raw material in bags (super sacks) for the two and one-half operational runs each year. Use of railcars to ship material during the runs would decrease the bag disposal cost and material-handling costs, and also improve quality. Quantitative analysis shows the elimination of 400 bags per year (a 46% reduction) by the use of two railcars per run.
3. An estimated 14,175 lb of raw material (unit price \$0.88/lb) is lost annually in the process of transportation and handling. This includes loss from damaged bags while loading and unloading from trucks, loss because of overflow while loading into the mixer, and loss owing to material left behind in emptied bags. Apart from the loss, transportation cost and labor cost are incurred in warehousing and delivery to the mixer. Railcar use would help avoid the loss of material and reduce cost. With a cost reduction of \$0.015/lb in shipment by railcars, \$12,000/yr can be saved in shipping 800,000 lb. Table 3 summarizes the savings from railcar use.
4. In one of the buildings, the drying equipment has a piece missing from the discharge side. This results in an estimated waste of 1050 lb/yr (\$924/yr), which can be easily saved by fixing the equipment.
5. Oil/grease used to lubricate the chain of bucket elevators has intermittently fallen on to the product rendering it a waste. By the use of a lubricant that does not drip, this waste could be eliminated.
6. Some of the feed hoppers for the mixers have an opening on one side to allow the control of flow out of the bag. A piece of removable Plexiglas in a slot on the side with the opening would cut down on what falls to the floor and still allow for control as needed.
7. The large dryer in Building 31 has part of the edge seal worn away allowing material to fall on the floor. Replacement or repair of this edge seal would stop this loss.

Table 2 Nonhazardous waste generation in 2000

Waste type	Amount generated (tons)	Amount reclaimed (tons)	Amount land filled (tons)	% Reclaimed
Cardboard	26	0	26	0
Pallets	114	114	0	100
Wood (pallet parts)	26	0	26	0
Fabric bags	77	0	77	0
Metal	222	222	0	100
Glass	7.5	0	7.5	0
Plastic	2	0	2	0
Office paper	30	0	30	0
Concrete	200	0	200	0
Raw materials	175	105	70	60
General trash	26	0	26	0
Total	905.5	441	464.5	49

Table 3 Railcar evaluation

Process	Amount saved	
	£	$
Offloading and warehousing	—	1,600
Delivering from warehouse to mixer	—	6,750
Bag damage	9,975	8,778
Bag overflow	1,000	880
Materials left in bags	3,200	2,816
Material cost savings	—	12,000
Total	14,175	32,824

8. Some of the conveyors to the bucket elevators on the furnaces do not have trays placed under them to catch any material that may fall off. An estimated loss of 15,750 lb/yr ($13,860) of material occurs as a result.
9. Approximately 44,000 lb of oats ($0.10/lb) are required annually for the cleanup of die from extruders. An estimated 40% of the oats could be reused one more time, saving $1760/yr, and resulting in 17,600 lb less of waste.

Hazardous waste

Table 4 shows the hazardous waste generated in 2001.

Avenues for Hazardous Waste Minimization and Savings

1. Material waste from furnaces 4 and 5 results in substantial hazardous waste generation because of the nature of the material. For furnace 4, there is a loss of 3500 lb/yr ($4.77/lb, $16,695/yr), and 7000 lb/yr for furnace 5 ($0.88/lb, $6,160/yr).
2. Improved housekeeping at waste generation sources could reduce the waste shipped offsite. Cross-contamination (mixing of nonhazardous waste with hazardous waste or vice versa) results in larger quantities of hazardous waste than are actually generated. Elimination of this would help achieve savings on disposal cost.
3. Fines and oversized custom products could be recycled, reducing waste. Metal loaded fines could be loaded to the wet mix, decreasing the amount of metal requiring impregnation in a future step.

Table 4 Hazardous waste generation in 2001

Waste	Amount generated (lb)
Miscellaneous dry waste	650,000
Water and sump sludge	52,070
Wastewater treatment sludge	200,000
Total	902,070

Recommendations

1. Switch to two railcars per run for raw material use to improve quality, decrease cost, and reduce labor.
2. Ensure that all offspecification drums of raw material wet mix are lined with a plastic bag, the bag is secured to prevent the wet mix from drying out, and training of all employees is continued on the importance of not mixing waste materials and proper labeling techniques.
3. Evaluate each wet mix for recycling. This material would be put back into the process at the wet mix portion. It can be used as a good intermediate or as a product substitute.
4. Use a different type of oil/grease or a different technique to lubricate bucket elevator chains.
5. For the custom product, continue with the investigations into using fines to make a product that can be used by the customer.
6. Reuse oats that are reusable when cleaning dyes.
7. Provide trays to capture any spillage under the conveyors and bucket elevators at all times.
8. Evaluate the gaskets and edge seal integrity on each machine before production begins, to minimize spills onto the floor.

Expected Benefits and Results. Table 5 summarizes the total estimated savings.

CASE STUDY 3: ENERGY SURVEY FOR A MANUFACTURING SERVICE CENTER

P2 Goal

An assessment to help reduce the spending on electricity bills and the amount of the pollutants produced was the highlight of the project.

P2 Procedure

The assessment procedure was broadly divided into the following two major components:

1. Data on the total electricity load were collected and compiled using the as-built drawings for lighting loads and the nameplate capacity of the motors for the major equipment and the air-conditioning loads. A spreadsheet was developed using the data mentioned earlier to conduct analysis of peak loads and energy loads under various scenarios.

Table 5 Potential savings

Improvements	Amount saved	
	£	$
Use of railcars	14,175	32,824
Use of trays under conveyors and bucket elevators	15,750	13,860
Reuse of oats for cleaning extruder dies	17,600	1,760
Lubrication type and method change for conveyors and bucket elevators	Included with others	Included with others
Hazardous waste reductions	75,500	27,730
Total	123,025	76,174

2. The utilization pattern of the lighting and equipment loads was studied by interviewing the various occupants of the manufacturing service center and observing the usage for a limited period. This part also involved measurement of light intensity at different locations in the manufacturing service center.

Observations

The study of the usage pattern revealed the following:

1. The employees of the manufacturing service center were not in the habit of shutting down their computers. The computer load was assessed to be the second highest contributor of electricity load.
2. The manufacturing service center had a shop floor that consisted of 54,400 W high-pressure sodium lamps. All these lights were kept on during office hours irrespective of occupancy of the shop. Ten of the lights were kept on for 24 hr/day, 7 days/week, 365 days/yr.
3. The light intensity as measured in some parts of the office was around 30–40% higher than the prescribed standards.

Recommendations

The following major recommendations were made:

1. The employees were asked to shut down the computers when leaving for the day. This alone would result in saving around 45,000 kWh on an annual basis.
2. The shop manager was asked to operate the shop lights on "as-needed" basis. This would result in an annual savings of around 20,000 kWh.

Apart from the major recommendations mentioned earlier, the employees were advised to use the lights in the conference room and other utility areas on an as-needed basis.

Results

Table 6 shows the amount of pollutants that will be reduced as a result of the recommendations.

CASE STUDY 4: ENERGY EFFICIENCY AND POLLUTION PREVENTION ASSESSMENT FOR CONCRETE BUILDING PRODUCTS MANUFACTURER

P2 Goal

An energy efficiency audit and a general pollution prevention assessment of the operations for a building industry, to determine if waste reduction and cost reduction are possible. Operations conducted at this facility include form fabrication, concrete mixing and pouring, and sand blasting and acid washing to impart the finish. These concrete slabs are formulated with a variety of sands and stone specific to the needs of the customer.

Table 6 Possible pollutant reduction

Name of the unit	CO_2 (lb/yr)	SO_2 (lb/yr)	NO_x (lb/yr)
Computers	81,000	1,033	348
Workshop	36,000	460	155
Total (for the facility)	117,000	1,493	503

P2 Procedure

An energy audit was conducted for the following systems:

- Hot water heater
- Boiler
- Compressed air
- Building lighting

A summary of the energy/water savings is provided in Table 7 for each of the systems examined during the assessment.

Domestic hot water heater

Current Status. The bathrooms and office area are supplied with hot water via a 32,000 Btu/hr residential hot water heater. The heater is at least 8 yr old and past its useful life, especially when heating hard water. During July and August of this year, gas usage averaged 9.5 ccf (100 cubic feet)/day. During this time, the only gas-using piece of equipment was the domestic hot water heater. If running at full capacity for the entire day, a properly operating heater of the same type should use only 7.7 ccf/day.

Recommendation. Replace the existing hot water heater with an instantaneous hot water heater to be installed in the men's bathroom fed from the existing piping.

Sioux boiler

Current Status. The Sioux boiler supplies hot water to the first mixing machine when needed. The burner has an output of 1,000,000 Btu/hr. The unit was built in 1986 and has an efficiency of 0.6.

Recommendation. Replace the Sioux boiler with a hot water heater similar to the one servicing the newer mixer but of a larger size because of the greater load.

Compressed air

Current Status. The compressed air systems are in the process of a major change out. General piping layout was discussed with plant personnel. Basic concepts in compressor room piping were discussed with the maintenance man doing the work. Leaks were tagged and the matching tags are presented with this report. On each, there is a location and an approximate leak rate.

Recommendation. Repair leaks that were tagged throughout the facility.

Lighting retrofit

Current Status. The production and shop floors are lit using 400 W high-intensity discharge (HID) lights. Because of overhead crane movement and the nature of production, lights must be kept at the ceiling level. There are also a few T12 task lights in the mixing and wood shop areas. In the offices and break rooms, inefficient T12 lighting technology is being used. The only lighting control is the standard on/off switches. Total lighting counts for the facility are: 400 W high pressure sodium (HPS)—76, 400 W metal halide (MH)—5, 4 ft four-lamp T-12 fixtures—65, and 4 ft two-lamp T-12 fixtures—11.

Recommendation. Replace all HID lights with more efficient 320 W fixtures throughout the plant. Change the offices and break room lighting from T12 to T8 lighting technology. To maximize efficiency, office lighting will not be a one-for-one change out. Lighting levels will be maintained at current standards after the retrofit. No controls will be added because at this time the company makes the effort to turn off lights when not in use.

BASIC STEPS TOWARD P2

The adoption of P2 as a core component of business management practice is a viable approach to both business success and environmental protection. Waste minimization or prevention can be made a routine part of daily business operations, at both small and large facilities. A little time and effort in following some P2 steps goes a long way toward success and preventing the creation of waste. The approach to P2 is compatible with the concept of environmental performance embodied in the Malcolm Baldridge model of rewarding businesses that raise awareness of the importance of quality and performance excellence as

Table 7 Potential energy savings identified by assessment

Proposed modification	Estimated annual energy savings	Estimated annual cost savings ($)	Cost to retrofit ($)	Payback (yr)
Change out hot water heater	2,738 ccf	2,382	2,850	1.2
Replace boiler with hot water heater	7,020 ccf	6,107	11,700	1.9
Fix compressed air leaks	6,200 kW/hr	486	450	0.9
Lighting retrofit	32,688 kW/hr	2,562	32,700	12.8

a competitive edge. The ideas of achieving environmental quality improvements using the traditional continuous quality improvement concept presented in the plan-do-check-act cycle is applicable to P2. Corporate policies and goals, plant policies and goals, environmental management tools, environmental auditing, information systems, employee awards, research and development, training and awareness, co-operation across plant boundaries are factors that lead to the greatest amounts of pollution prevention.[24] Fig. 3 shows a diagrammatic representation of the basic steps in the implementation of a P2 program in any industry.[6,8,19,25,26]

Each of the steps involved in the implementation of a P2 program and the potential benefits are described below:

1. Evaluate company policy and culture
   a. Establish an environmental management system with clear policies, goals, and objectives.
   b. Draft a brief written policy statement in support of a pollution prevention program.
   c. Justify changes in the process or design with cost savings potential and reasons that provide a win-win situation for all involved.
2. Identify waste and its source
   a. Characterize each waste stream—determine where the waste comes from, what processes generate it, and how much is being discarded.
   b. Examine all waste streams, including process wastes, hazardous wastes, nonhazardous wastes, solid wastes, and office waste.
   c. Study what is discarded in trashcans and dumpsters to determine what materials are being thrown away and consider what wastes are poured down the drain, such as rinse waters and process waters.
   d. Consider the costs of each stream, including the cost to purchase, dispose of, treat, or control it.
   e. Examine energy and water consumption on a facilitywise and process specific basis and look for high- and low-usage trends in your water and electric bills.
3. Set a target
   a. Set specific measurable and attainable goals to reduce specific types of wastes by some reasonable and feasible amount, for example, a reduction of process effluent by 10% or a reduction in energy consumption by 15%.
4. Identify and prioritize the opportunities
   a. Evaluate all the waste streams for possible reduction and come up with a list of measures that can be taken to reduce waste and improve process efficiency.
   b. Investigate opportunities for new products or ingredients that prevent waste generation.
   c. Prioritize the opportunities identified by considering
      i. Cost
      ii. Ease of implementation
      iii. Payback (and cost savings or cost avoidance)
      iv. Criteria deemed important by the organization, such as increased employee safety.
   d. Focus on opportunities that are easy to implement, have low capital investment, and reduce large volumes of waste.

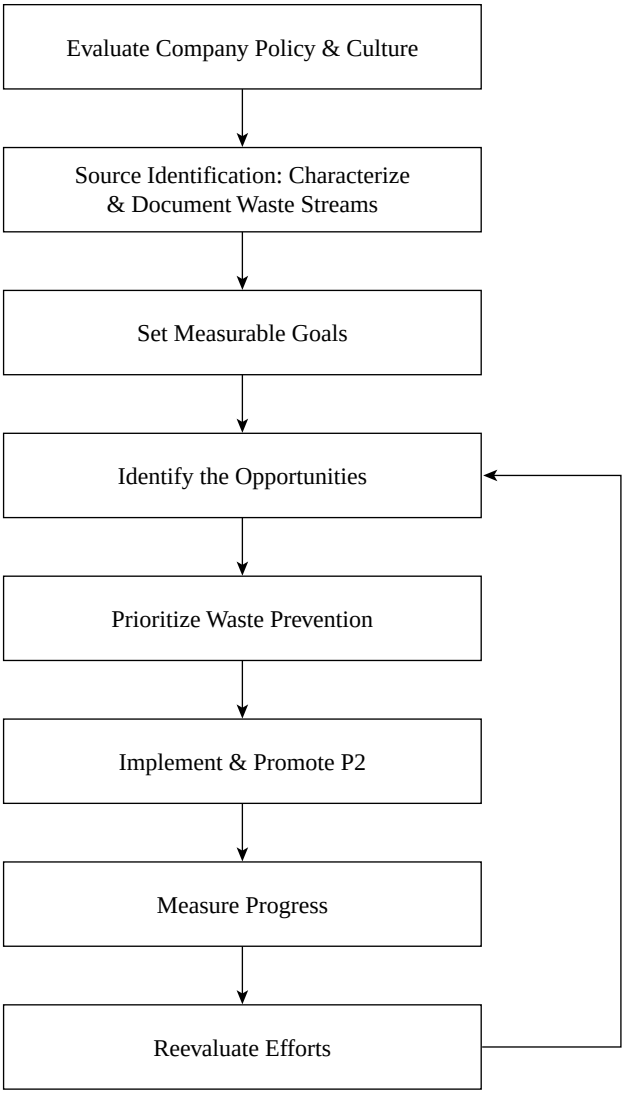


Fig. 3 Basic steps involved in a P2 program.

e. Identify the best methods or opportunities/avenues. For example, identify the practice or approach, such as process redesign, product redesign, or general operating procedures that would best enhance the efficiency and determine whether it is feasible.

5. Implement and promote P2
   a. Teach and train employees on P2 concepts on how to prevent waste from being created in the first place.
   b. Discuss the waste prevention policies and goals, and provide training to employees who must change the way they handle materials.
   c. Encourage employee involvement by asking for new suggestions and offering incentives.
6. Measure progress
   a. Quantify and track reductions and cost savings in volumes of waste produced and treated, energy and water use, raw materials consumed, toxic and greenhouse gas emissions, and other effluents.
   b. Qualitatively, claim the less tangible benefits such as improved public image, enhanced worker and workplace safety, continued community relationship, etc.
7. Re-evaluate
   a. Conduct regular assessments to identify additional waste prevention opportunities as part of a process of continual improvement. As long as wastes continue to be generated, there are opportunities to reduce it.

To derive the most benefits from the P2 program steps, various assessment tools have been developed by researchers and practitioners in the field.[27,28] Many of these tools for assessment and implementation of P2 have been developed and customized to suit a particular facility or organization. In particular, a systems approach to P2 assessment and the use of process mapping tools can be extremely useful in taking a global view of an organization's activities and how it impacts the people and the environment in which the activities are conducted.[29] The systematic approach of looking at an organization's current environmental and management systems and for deconstructing processes allows one to look at old processes in a new light. The systemwise review of processes also provides opportunities for a range of people with different functions—designers; production and operational staff, sales and marketing staff; supply chain vendors; customers and management—to come up with ideas and opportunities for actions to implement P2.

The success of P2 assessments can be evaluated in different ways. A common approach is to look at the type of assessment, the actions taken as a result of the assessment, and the P2 outcome of the actions taken. As the first step in this approach, summary information

Table 8 Company-specific P2 assessment summaries

Center	Type of assessment	Environmental impact considered	Types of pollution prevention addressed	Action by company (environmental improvement)	Reduction in the amount of pollutants released/energy consumed
Company 1	P2	Risk because of hazardous waste	Hazardous and nonhazardous wastes	Initial identification of substances of concern	Potential reduction of 75,000 lb of hazardous waste
Company 2	P2	Health risk because of chemicals	Toxic chemicals/PBT chemicals	Elimination of chemicals by their suppliers or selection of new supplies	38.2% reduction in pollutants (approx. 782 tons)
Company 3	P2	Wastewater	Reduction of blood and solids to wastewater	Began collecting blood and shipping to renderer for reuse	75,000 gal/yr blood
Company 4	Energy assessment/P2	Reduction in CO_2, SO_2, NO_x achieved by less power usage	Energy related air emissions	Installed laboratory A/C economizer, revised plant lighting layout	Reduction in energy use estimated at \$4,000 savings so far, projected \$30,000/yr total
Company 5	Lean manufacturing/P2	Reduction in time of production	Energy savings	Reduction in production time	Potential savings of \$2.5 million/yr

on various assessments may be collected from companies involved in P2 programs. A summary of assessments at five different companies is given in Table 8. The table also describes the environmental improvement as a result of the assessment and actions by the companies. The next step examines the P2 benefits derived from the assessments. The companies, where the P2 investigations were conducted, provided information as shown in Table 9 outlining which P2 benefits were derived as a result of the assessments. P2 assessments can also be evaluated by considering other metrics such as: 1) the number of recommendations implemented by the companies; 2) cost reductions after the implementation of the recommendations; and 3) risk reductions.

Potential Benefits

The potential benefits of P2 program implementation are varied and depend to a large degree on the level of effort made to systematically evaluate the policies, systems, designs, products and processes, and the commitment to take actions based on the assessment. The following major benefits may be seen as the outcome of successful P2 implementation:

- Cost savings
    - Source reduction, inprocess recycling, and improved energy efficiency reduce the amounts of raw materials and energy required.
    - Substituting hazardous chemicals with safer alternatives cuts purchasing cost and reduces pollution control costs.
    - Reduction of nonhazardous wastes reduces procurement and disposal expenses.
    - P2 activities cut down on the costs of complying with federal and state regulations and reporting requirements.
    - Undertaking P2 projects can reduce regulatory exposure and, in some cases, may eliminate the need for permits, manifesting, monitoring and reporting, and lower the cost of meeting expensive and time-consuming regulatory requirements.
- Reduced legal liability
    - Preventing the generation of wastes and emissions dangerous to the environment or human health protects a company from future liability under the Occupational Safety and Health Administration (OSHA) and federal and state environmental statutes.
    - Reducing wastes reduces long-term liability caused by improper disposal practices.
- Improved corporate image
    - P2 program can serve as an effective public relations tool. A company that demonstrates an active commitment to reducing its environmental impacts will have a more positive relationship with the local community and with its customers.
    - Improve marketing efforts by its environmental performance and its demonstrated concern for the health of people and the environment.
    - Establish itself as a responsible, reputable member of the community with a competitive advantage.
    - Be an industry leader with the ability to report on P2 activities and show measurable progress for inclusion in annual environmental reports.

Table 9 P2 benefits

Benefit	Company 1	Company 2	Company 3	Company 4	Company 5
Raw materials savings by recycling	Yes	Yes	Yes	N/A	No
Material substitution savings	Yes	Yes	N/A	Yes	N/A
Elimination of pollution control equipment	N/A	N/A	N/A	N/A	N/A
Improved process productivity	Yes	Yes	Yes	Yes	Yes
Better product quality	Yes	N/A	Yes	Yes	No
Reduced labor from pollution elimination	Yes	N/A	Yes	Yes	N/A
Reduction of permitting requirements	N/A	Yes	N/A	No	N/A
Reduction and/or elimination of offsite waste disposal	N/A	N/A	N/A	No	N/A
Reduction and/or elimination of offsite waste storage	Yes	N/A	N/A	Yes	N/A
Reduction of personal injury risks under OSHA	N/A	Yes	N/A	Yes	N/A
Permitting requirements minimized	N/A	N/A	N/A	Yes	N/A

  - Successful P2 program implementation can be used for receiving recognition from governmental and nongovernmental groups to showcase projects and be seen as a leader in environmental protection and sustainable practices.

- Improved worker safety
  - The P2 program can be used to improve worker health and safety.
  - Use of less harmful substances, reduced fugitive releases of solvents from manufacturing processes, and minimized waste to be handled and disposed of improve the occupational and environmental safety of workers and waste handlers.
  - Reducing or eliminating toxic waste emissions or discharges not only improves environmental quality, but can also reduce medical claims and disability leave requests from workers.

- Widespread use of P2
  - Success stories of the implementation of P2 in industries pave the way for more organizations to try P2 options and undertake changes and demonstrate environmental commitment and responsible stewardship.

Barriers and Challenges

Institutional, financial, and regulatory barriers to P2 adoption and full exploitation of its potential can exist in differing proportions depending on the type, size, and working culture of an industry or a facility where P2 is being considered. While a change in the attitude and behavior of the people involved with the company (i.e., the major stakeholders) helps overcome these barriers; behavioral change requires a thorough understanding of the target audience's attitudes, obstacles, and information sources.[19,30] A brief discussion of each of the major barriers follows.

Institutional barriers

1. Lack of employee participation in decision-making.
2. Employee inputs are not taken into account.
3. Organizational structure wherein the environmental management group is unaware or knows very little about the production sector's activities, i.e., no proper communication channel between different sectors in an industry or between different departments at a facility.
4. Upper management and environmental managers unaware of the cost-saving ideas and succumbing to conventional end-of-pipe solutions offered by entrenched consultants.
5. Customer demand for products resulting in toxic waste.
6. Lack of time and expertise to analyze causes of waste for small- and medium-sized industries.
7. Lack of company leadership, vision, or understanding of the company's environmental impacts and the absence of appropriate environmental policies and procedures that can benefit the business while being protective of the environment.

Financial barriers

1. Failure to understand the true costs of decisions, inclusive of the hidden costs of waste [hidden costs $= \sum$(treatment and disposal charges, permitting fees, record keeping, insurance, utility bills, and potential future environmental liability)].
2. Waste reduction proposals not able to clear the corporate internal investment hurdle threshold that a project must meet.
3. Reluctance of bankers to finance unconventional projects.

Regulatory barriers

1. Technological innovations in the current vigorously growing and competitive environment may not have been included or approved in the existing rules.
2. Companies focusing on conventional technologies geared toward compliance with single medium specific pollution control standards without being able to consider impacts on air, water, or soil media.
3. Failure of regulatory agency staff to allow innovative projects to be approved because of regulatory constraints or lack of understanding of a project's true benefits.

Integrating P2 into core business operations and understanding the full life cycle costs and benefits of decisions and increased communication among the key stakeholders are among the essential steps for overcoming these barriers.

CONCLUSIONS

The growth of P2 from its initial stages to the present stage has been a slow and steady path as with any

technical change. The progress of P2 in the United States in the last two decades has been very impressive, especially the manner in which environmental issues are being prioritized to take advantage of technological advances in the face of increased pressure to keep costs down. In addition, P2 practices that use a systems-based, quality-focused, systematic assessment approach are aiding in the creation of sustainable development practices to meet the triple bottom line—cost, environment, and social justice. The principle of a waste management hierarchy and the procedures for achieving P2 or green/cleaner production are generic across plant boundaries. Many organizational factors affect the likelihood of realizing the full potential of pollution prevention.[24] Availability of numerous case studies on P2 and a change in corporate thinking toward proactive environmental approaches show that the P2 concept has matured and is playing a major role in environmental protection without specific government regulation. The rapid development of new products using efficient processes that reduce pollution provides economic incentives for P2 propagation across a broader set of industries, which can lead to ever-faster cycles of review and implementation. P2 is a win-win opportunity as the forces of manufacturing interests and environmental responsibility come together.[4]

REFERENCES

1. de Bruijn, T.J.N.M; Hofman, P.S. Pollution prevention and industrial transformations evoking structural changes within companies. J. Cleaner Product. **2000**, *8*, 215–223.
2. UNEP. Cleaner Production—Related Concepts; http://www.uneptie.org/pc/cp/understanding_cp/related_concepts.htm (accessed Oct 2004).
3. The CDROM Greatest Hits Collection for Cleaner Production and Pollution Prevention; http://www.cleanerproduction.com/hits/greatest.htm (accessed Oct 2004).
4. Overcash, M. The evolution of US pollution prevention, 1976–2001: a unique chemical engineering contribution to the environment—a review. J. Chem. Technol. Biotechnol. **2002**, *77*, 1197–1205.
5. Bishop, P.L. *Pollution Prevention: Fundamentals and Practice*; McGraw-Hill Science/Engineering/Math, 1999; 768 pp.
6. Higgins, T.E. *Pollution Prevention Handbook*; Lewis Publishers: Boca Raton, 1995; 592 pp.
7. Patel, I; Kumar, A. A guide to information on pollution prevention on the World Wide Web. Environ. Prog. **2000**, *19* (4), W8–W13.
8. USEPA. P2 Home—Pollution Prevention Pays; http://www.epa.gov/p2/ (Oct 2004).
9. Pollution Prevention: EPA Statement of Definition; EPA/742/F-92/001; National Environmental Publications Information System, US EPA, 1992; 4 pp.
10. *101st Congress 2nd Session H.R. 5931: Pollution Prevention Act of 1990*, EPA/742/F-90/100; National Environmental Publications Information System, US EPA, 1990; 8 pp.
11. United Nations World Commission on Environment and Development; http://www.sustainable.doe.gov/overview/definitions.shtml (accessed 2004).
12. Hamner, B. What Is the Relationship Between Cleaner Production, Pollution Prevention, Waste Minimization and ISO 14000? The 1st Asian Conference on Cleaner Production in the Chemical Industry, Taipei, Taiwan, Dec 9–10, 1996; http://www.cleanerproduction.com/misc/Pubs/CP%20Concepts.html (accessed Oct 2004).
13. Best Manufacturing Practices; http://www.bmpcoe.org/ (accessed Oct 2004).
14. USEPA. Design for Environment; http://www.epa.gov/opptintr/dfe/ (accessed Oct 2004).
15. Allen, D.T.; Shonnard, D.R. *Green Engineering: Environmentally Conscious Design of Chemical Processes*; Prentice Hall PTR: New Jersey, 2002; 552 pp.
16. USEPA. Green Chemistry; http://www.epa.gov/gcc/whats_gc.html (accessed Oct 2004).
17. USEPA. Green Engineering; http://www.epa.gov/oppt/greenengineering/index.html (accessed Oct 2004).
18. Phipps, E. Pollution Prevention Concepts and Principles; http://www.umich.edu/~nppcpub/resources/p2.html (accessed Oct 2004).
19. Ohio Pollution Prevention and Waste Minimization Planning Guidance Manual; http://www.epa.state.oh.us/opp/guide/p2pch2.html (accessed Oct 2004).
20. Pollution Solutions; http://outreach.missouri.edu/polsol/p2meth.htm (accessed Oct 2004).
21. Practical Solutions for Economic and Environment Vitality; http://www.pprc.org/solutions.cfm (accessed Oct 2004).
22. Canadian Pollution Prevention Information Clearing House, Environment Canada; http://www.ec.gc.ca/cppic/youth/lplan/en/pos.cfm?prac=1 (accessed Oct 2004).

23. Kumar, A.; Vijayan, A.; Varadarajan, C.; Pendse, R.; Jampana, S. Pollution Prevention Assessments—Case Histories (2000–2003); http://p2tools.utoledo.edu/ (accessed Oct 2004).
24. Lober, D.J. Pollution prevention as corporate entrepreneurship. J. Organ. Change Manage. **1998**, *11* (1), 26–37.
25. *Pollution Prevention Fact Sheet: Setting up a Pollution Prevention Program*, EPA/742/F-93/004; National Environmental Publications Information System, US EPA, 1993; 2 pp.
26. Pollution Prevention Regional Information Center; http://p2ric.org/ (accessed Oct 2004).
27. Pollution Prevention Tools; http://p2tools.utoledo.edu/ (accessed Oct 2004).
28. Kumar, A.; Thomas, S. A Software Tool for Screening Analysis of Lean Practices. Environ. Prog. **2002**, *21* (3).
29. Pojasek, R.B. Pollution prevention redux, P2 finds new life through the quality connection. Pollut. Prev. Rev. **2000**, *10* (2), 49–52.
30. DiPeso, J. P2: putting environmental issues in a new light. Environ. Q. Manage. **2000**, *10* (1), 13–24.

Polyanhydrides

Maria P. Torres
Amy S. Determan
Surya K. Mallapragada
Balaji Narasimhan
Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa, U.S.A.

INTRODUCTION

Polyanhydrides are a class of bioerodible polymers that have shown excellent characteristics as drug delivery carriers. The properties of these biomaterials can be tailored to obtain desirable controlled release characteristics. Extensive research in this promising area of biomaterials is the focus of this entry. In the first part of the entry, the chemical structures and synthesis methods of various polyanhydrides are discussed. This is followed by a discussion of the physical, chemical, and thermal properties of polyanhydrides and their effect on the degradation mechanism of these materials. Finally, a description of drug release applications from polyanhydride systems is presented, highlighting their potential in biomedical applications.

BACKGROUND

The need for suitable materials for the delivery of drugs in a safe and controlled manner has led to the development of numerous biodegradable polymers. Controlled release of a variety of therapeutic agents has been achieved with the use of biodegradable polymeric devices. Research has focused on poly(α-hydroxy acids), poly(orthoesters), and poly(anhydrides). Poly(α-hydroxy acids), e.g., poly(lactic acid) (PLA) and poly(glycolic acid) (PGA), and copolymers undergo bulk erosion and the drug release kinetics from these carriers is not well defined. On the other hand, poly(orthoesters) and polyanhydrides undergo surface erosion with predictable kinetics. In the case of polyanhydrides, the degradation rates can be tailored to suit specific applications by changing the chemistry.

Polyanhydrides comprise monomer units connected by water-labile anhydride bonds. In the presence of water, the polymer is cleaved across the anhydride bond into two carboxylic acid groups (Fig. 1). It is precisely this hydrolytic instability that precluded their use in the textile industry in the 1950s and led researchers to suggest their potential as drug delivery carriers in the 1980s. Since then, polyanhydrides have been synthesized with a wide range of chemistries for a variety of biomedical applications.

The promising characteristics of polyanhydrides for biomedical applications rely on the surface erosion mechanism that translates into well-controlled release kinetics, where the drug release rate coincides with the degradation rate of the polymer. In an aqueous environment, the macromolecules at the surface break into smaller chains before water penetrates into the device. Thus, the drug is released as the polymer degrades. In contrast, bulk eroding polymers degrade slowly and water penetrates into the system much faster, having, in consequence, less predictable kinetics as the drug is released from the entire matrix. A comparison of surface and bulk erosion mechanisms is shown in Fig. 2.

Polyanhydride-based drug delivery devices (Gliadel®) have been approved by the Food and Drug Administration (FDA) for the treatment of brain tumors. This device is a polyanhydride wafer composed of sebacic acid and 1,3-bis(*p*-carboxyphenoxy) propane [1,3-bis (*p*-carboxyphenoxy) propane : poly(sebacic acid) (CPP : SA) copolymer in 20 : 80 molar ratio] loaded with the chemotherapeutic agent, carmustine, 1,3-bis[2-chloroethyl]-1-nitro-sourea (BCNU). Other potential applications of CPP : SA copolymers include the release of bethanechol for the treatment of Alzheimer's disease and the controlled release of insulin.[1] The treatment of osteomyelitis, which is a bone infection difficult to treat by conventional methods, has been carried out with 20 : 80 CPP : SA copolymer loaded with gentamicin sulfate.[1] Several chemotherapeutic drugs, local anesthetics, anticoagulants, neuroactive drugs, and vaccines have been delivered using polyanhydrides.

CLASSIFICATION

There are three major classes of polyanhydrides: aliphatic, unsaturated, and aromatic. The chemical structures are shown in Table 1.

Aliphatic Polyanhydrides

The first aliphatic polyanhydride was synthesized from the monomer adipic acid (AA), which is thermally unstable and forms cyclic dimers and polymeric rings

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039789

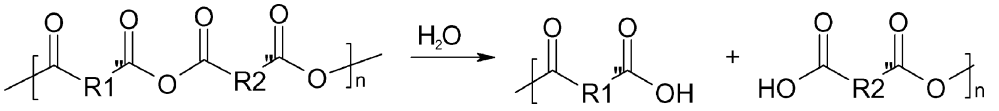


Fig. 1 Hydrolysis of polyanhydrides.

when heated at high temperatures. Poly(sebacic acid) (SA), the aliphatic polyanhydride most widely used in drug delivery applications at present, was synthesized for the first time in the 1930s.[2] Typical properties of aliphatic polyanhydrides include crystallinity, a melting temperature range of 50–90°C (increasing with monomer chain length), and solubility in chlorinated hydrocarbons. These degrade and are eliminated from the body within weeks. When copolymerized with aromatic polyanhydrides, the degradation time can be extended to several months as the aromatic composition increases. The most widely studied aliphatic–aromatic copolymer system is based on SA and 1,3-bis(*p*-carboxyphenoxy) propane (CPP).[3]

Unsaturated Polyanhydrides

The development of unsaturated polyanhydrides responded to the necessity of improving the mechanical properties of the polymers in applications such as the temporary replacement of bone.[4] Unsaturated polyanhydrides, prepared by melt or solution polymerization, include homopolymers of fumaric acid (FA), acetylenedicarboxylic acid (ACDA), and 4,4′-stilbenzenedicarboxylic acid (STDA). The chemical structures of poly(FA) and poly(ACDA) are shown in Table 1. These polymers are highly crystalline and insoluble in common organic solvents. The double bonds of these monomers make them suitable for further crosslinking to improve mechanical properties of polyanhydrides. When copolymerized with aliphatic diacids, less crystalline polymers with enhanced solubility in chlorinated solvents result.

Aromatic Polyanhydrides

The first aromatic polyanhydrides synthesized were poly(isophthalic acid) (IPA) and poly(terephthalic acid) (TA).[5] A few common aromatic polyanhydrides are shown in Table 1. Homopolymers of aromatic diacids are crystalline, insoluble in common organic solvents, and have melting points greater than 100°C. Their hydrophobicity results in a slow degradation rate that can last over a year in some cases. Thus aromatic polyanhydrides are not suitable for drug delivery when used as homopolymers. To overcome their slow degradation rates, they have been copolymerized with aliphatic diacids, i.e., CPP : SA copolymers, and with other aromatic monomers. The copolymers of the aromatic monomers TA and IPA are amorphous and soluble, with a faster degradation and a melting point below 120°C.[6]

Other Polyanhydride Chemistries

Although it is impossible to discuss in detail all the polyanhydrides that have been synthesized, some distinguishable classes are discussed here. Typical examples of novel classes of polyanhydrides include those derived from amino acids and fatty acids, and those modified by copolymerization with esters and ethers. The polyanhydrides derived from amino acids, including trimellitylimido glycine (shown in Table 1), pyromellitylimido alanine, and trimellitylimido l-tyrosine, have been copolymerized with aliphatic (SA) and aromatic (CPP and CPH) monomers to obtain enhanced degradation and improved mechanical strength because of the presence of the imide bond.[7] These polymers have been studied as vaccine carriers.[8] Some polyanhydrides have been synthesized from dimer and trimer unsaturated fatty acids, and form nonlinear hydrophobic fatty acid esters such as ricinoleic and maleic acid. Other classes of polyanhydrides include ones copolymerized with esters and ethers, which have been suggested as potential drug carriers in the last decade.[9,10] Uhrich and coworkers recently synthesized novel poly(anhydride-co-ester)s containing salicylic acid in the backbone.[11–14] The in vitro/in vivo release of salicylic acid (the active form of aspirin) was studied for the treatment of Crohn's disease and tuberculosis. Copolymers of aliphatic polyanhydrides with ε-caprolactone, trimethylene carbonate,

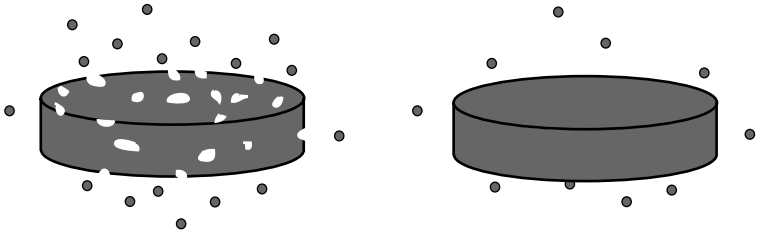
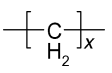
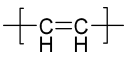
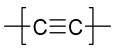
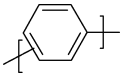
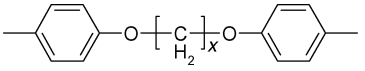
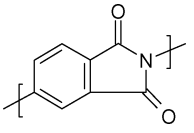
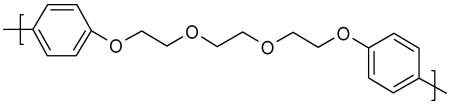


Fig. 2 Mechanisms of polymer erosion: bulk (left) and surface (right).

Table 1 Typical polyanhydrides used for drug delivery applications

Classification	R group	Examples
Aliphatic polyanhydrides		$x = 4$, Adipic anhydride (AA) $x = 8$, Sebacic anhydride (SA)
Unsaturated polyanhydrides		Fumaric anhydride (FA)
		Acetylenediacarboxylic anhydride (ACDA)
Aromatic polyanhydrides		meta: isophthalic anhydride (IPA) para: terephthalic anhydride (TA)
		$x = 1$, bis(*p*-Carboxyphenoxy)methane (CPM) $x = 3$, 1,3-bis(*p*-Carboxyphenoxy) propane (CPP) $x = 6$, 1,6-bis(*p*-Carboxyphenoxy)hexane (CPH)
Novel polyanhydrides		Trimellitylimido glycine (TMAgly)
		1,8-bis(*p*-Carboxyphenoxy)-3,6-dioxaoctane (CPTEG)

ethylene glycol,[15] and lactic acid have been synthesized. Several modifications of anhydride monomers have been carried out to obtain desired characteristics for particular applications.[16–18] An example is the incorporation of triethylene glycol (TEG) into an aromatic monomer (CPH) to enhance the hydrophilicity of the monomer, resulting in a faster degradation rate.[19] The resulting polymer (Table 1) is poly(1,8-bis(*p*-carboxyphenoxy)-3,6-dioxaoctane) (CPTEG).

SYNTHESIS

The most widely used method to synthesize polyanhydrides is melt condensation polymerization, which results in high molecular weight polymers.[20] Other methods include Schotten–Baumann condensation, dehydrative coupling, and ring opening polymerization.

Melt Polycondensation

The general process for melt polycondensation of polyanhydrides is shown in Fig. 3. It consists of reacting dicarboxylic acid monomers with an excess of acetic anhydride to form oligomers that are polymerized at high temperature under vacuum. The degree of polymerization is influenced by the monomer purity, the strength of the vacuum, the reaction temperature, and the reaction time. It has been found that for most polyanhydrides, the optimal polymerization temperature is in the range of 170–190°C.[21] In general, the condensation reaction is conducted for 2–3 hr, as significant depolymerization can occur after heating for longer periods.[22] With optimal conditions, molecular weights greater than 100,000 can be produced.

The polyanhydrides synthesized by melt condensation have fiber-forming properties in the molten state. They hydrolyze when exposed to air and this degradation is mainly controlled by the composition of the polymer. Homopolymers of aromatic monomers, such as CPH, degrade at a rate that is several orders of magnitude lower than that of homopolymers of aliphatic monomers.[22]

Several variations have been made to the melt condensation process. In the case of polymerization with propionic anhydride and butyric anhydride, harsh

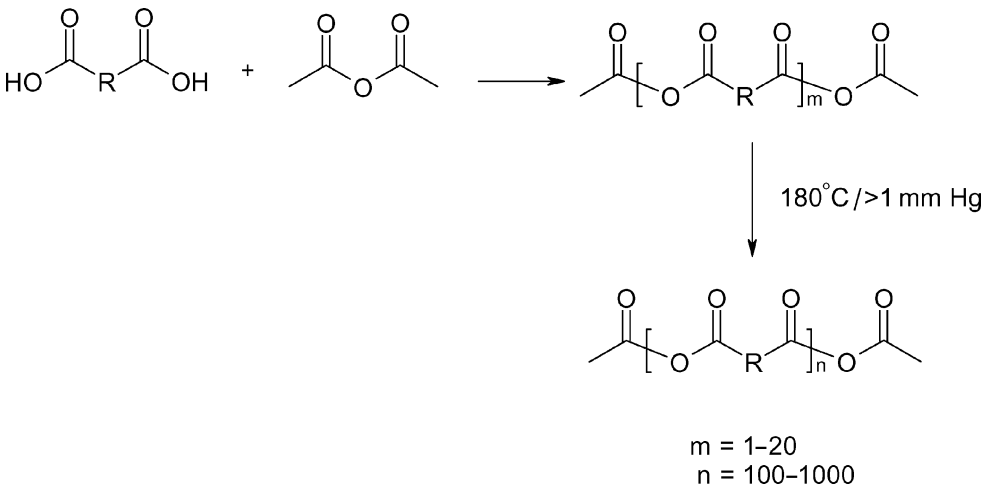


Fig. 3 Melt condensation polymerization of polyanhydrides.

conditions can be used for the removal of unreacted anhydride owing to the high boiling point of both chemicals.[6] A variety of catalysts have been used to polymerize polyanhydrides within 20–60 min; but the main disadvantage for biomedical applications is the potential toxicity from catalysts such as cadmium acetate, earth metal oxides, and $ZnEt_2$-H_2O.[6]

Schotten–Bauman Condensation

The Schotten–Bauman condensation produces polyanhydrides with moderate molecular weights by a dehydrochlorination reaction between a diacid chloride and a dicarboxylic acid.[21] The polymerization takes place by reacting the monomers for 1 hr at room temperature, and it can be conducted via solution or interfacial methods. Solvents that are used in solution polymerization include dichloromethane, chloroform, benzene, and ethyl ether. The degree of polymerization obtained with this method is approximately 20–30. Lower molecular weight products are obtained for less reactive monomers such as isophthaloyl chloride.

Polymerization conducted in aqueous interfacial systems suffers from hydrolytic decomposition. The decomposition reaction can be minimized when contact with water is avoided. In the case of polymerization in nonaqueous interfacial environments, products with number average molecular weights up to 5000 can be obtained.[22] Various aromatic polymers were prepared from the reaction of equimolar amounts of the acid dissolved in an aqueous base and the corresponding diacid chloride dissolved in an organic solvent. Reaction occurred between dibasic acid in one phase and an acid chloride in the other. Polar solvents for this reaction include dimethylformamide and 1,4-dicyanobutane.

Dehydrative Coupling

Another method to synthesize polyanhydrides is by dehydrative coupling of two carboxyl groups. Even though this method produces lower molecular weight products (mostly oligomers) compared with the methods described above, it is a single step polymerization where a dicarboxylic acid monomer can be directly converted into the polymer. Moreover, it can be conducted at low temperatures suitable for monomers that cannot resist harsh reaction conditions.

A number of dehydrative agents have been effective in coupling the carboxyl groups. The most effective agents are bis[2-oxo-3-oxazolidinyl]phosphinic chloride, *N*-phenylphosphoroamidochloridate, diphenyl chlorophosphate, diethyl phosphorobromidate, dicyclohexylcarbodiimide, chlorosulfonylisocyanate, and 1,4-phenylene diisocyanate.[22] In general, higher molecular weight polymers were obtained with polar solvents such as dichloromethane and chloroform. The major disadvantage of this method is the problematic isolation and purification of the final products while preventing hydrolytic decomposition.

Ring Opening Polymerization

Low molecular weight linear polymers undergo transformations between linear and cyclic forms. When a mixture of low and high molecular weight polymers is subjected to molecular distillation, cyclic monomers and dimers are distilled off and a high molecular weight polymer remains behind. The cyclic molecules are transformed to a polymer that contains large ring structures.[6]

Another variation of this process is the preparation of adipic acid from cyclic adipic anhydride

(oxepane-2,7-dione). The monomer is prepared by the reaction of adipic acid with acetic anhydride followed by catalytic depolymerization under vacuum. Factors that affect this reaction include temperature, reaction time, and the concentration of catalyst, if used. When catalyzed, reaction at 180°C for 30 min produced polymers with molecular weights up to 300,000. Uncatalyzed reactions that were carried out for more than 2 hr at 180°C yielded low molecular weight polymers.[6]

CHARACTERIZATION

To understand the properties that make polyanhydrides suitable drug carriers, their chemical, physical, and thermal behavior need to be characterized. This section discusses the methods to determine the chemical structure and composition, the molecular weight, the thermal properties, the phase behavior, the stability, and the erosion mechanism of polyanhydrides.

Chemical Structure and Composition

The technique most widely used for determining the chemical structure and composition of polyanhydrides is ^{1}H NMR spectroscopy. The chemical structure is assigned in accord with the chemical shifts characteristic of aliphatic and aromatic protons. The protons close to electronegative groups, i.e., aromatic groups, absorb at lower frequencies (6.5–8.5 ppm), while aliphatic protons absorb at higher frequencies (1–2 ppm).[23] ^{1}H NMR has also been used to determine the degree of randomness in polyanhydride copolymers.[24] By integration of NMR peaks it is possible to determine if a copolymer has a random or block-like structure. Other useful information obtainable from ^{1}H NMR spectra include details of the conversion of polymerization reactions, the actual composition of the polymer, the polymer molecular weight, and degradation rate.

Fourier transform infrared (FTIR) spectroscopy and Raman spectroscopy have also been used to authenticate polyanhydride structures. Aliphatic polymers absorb at 1740 and 1810 cm^{-1}, while aromatic polymers absorb at 1720 and 1780 cm^{-1}.[24] All the polyanhydrides show methylene bands because of deformation, stretching, rocking, and twisting. Aside from being used to ascertain polyanhydride structures, these techniques can be used to determine degradation progress, by monitoring the area of carboxylic acid peak (1770–1675 cm^{-1}) with respect to the characteristic anhydride peaks over time.

Molecular Weight

The molecular weight of polyanhydrides can be determined by gel permeation chromatography (GPC), viscosity measurements, and ^{1}H NMR spectra. Vapor pressure osmometry (VPO) cannot be used for molecular weight determination, as depolymerization occurs during the experiment. The weight average molecular weight (M_w) of polyanhydrides ranges from 5000 to 300,000. Typical polydispersity indexes are in the range of 2–15, which increases with molecular weight. Gel permeation chromatography determines the molecular weight relative to polystyrene standards. The intrinsic viscosity (η) is proportional to M_w, as shown by the Mark Houwink relationship for CPP : SA copolymer [Eq. (1)]. This relationship was calculated from viscosity experiments and M_w values from GPC.

$$[\eta]^{23^\circ C}_{CHCL_3} = 3.88 \times 10^{-7} \times M_w^{0.658} \quad (1)$$

An alternative way to estimate the molecular weight of polyanhydrides is by end group analysis from ^{1}H NMR spectra. The degree of polymerization can be calculated from the ratio of the area of the inner chain protons to the area of terminating groups. The number average degree of polymerization (DP) of CPP : SA copolymers is represented in Eq. (2), where (CPP) and (SA) depict the area of scaled inner chain protons, (Ac) represent the acetylated end group, and (SA^*) and (CPP^*) designate the carboxylic terminated polymer chain.[25]

$$DP = \frac{2[(CPP) + (SA)]}{[(Ac) + (SA^*) + (CPP^*)]} \quad (2)$$

Thermal Properties

The thermal transitions of polyanhydrides have been determined from differential scanning calorimetry (DSC). Differential scanning calorimetry thermal scans provide properties such as glass transition temperature (T_g), melting temperature (T_m), and heat of fusion (ΔH). It is important to know the values of T_g and T_m in the fabrication of drug delivery devices such as tablets and microspheres. While T_g determines the minimum temperature required for compression molding, T_m determines the minimum temperature necessary for injection molding or melt compression. A general decreasing trend in T_g's has been observed as methylene groups are added into the main chain of an anhydride monomer. As mentioned earlier, aliphatic polyanhydrides melt at temperatures below 100°C and aromatic polyanhydrides have melting points greater than 100°C.

It has been shown that the crystallinity of polymers affects erosion and drug release rates, because crystalline regions erode slower than amorphous ones.[26] Moreover, highly crystalline polyanhydrides affect the device morphology as it creates irregular external

surfaces. The crystallinity of polyanhydrides has been determined using x-ray diffraction, DSC, ^{1}H NMR spectroscopy, and small-angle x-ray scattering (SAXS). It has been demonstrated that homopolymers of aromatic and aliphatic diacids are crystalline. When copolymerized, polyanhydrides exhibited a decrease in crystallinity in copolymers of equimolar compositions, i.e., CPP : SA, CPH : SA, and FA : SA copolymers.[26] The ΔH from DSC thermographs exhibited a decrease as the copolymers approached equimolar compositions. This decrease in crystallinity is representative of the random behavior of the polymer chain, as determined by ^{1}H NMR spectra. In general, the copolymers rich in one monomer had higher crystallinity.

Phase Behavior

Polymer blends, which display distinct physical and chemical properties, are used for the design of materials for diverse applications. This variation in properties may lead to microphase separation, which in turn affects the drug release because of a thermodynamical partition of drugs between the phases, depending on their compatibility with the phase.[27] Research has shown that aliphatic, aromatic, and copolymers of anhydride monomers are miscible and the blends had a single melting temperature that was lower than that of the starting polymers.[6] On the other hand, polyanhydrides that are partially miscible with poly(orthoesters), poly(hydroxybutyric acids), and low molecular weight poly(esters) and have two melting temperatures are clearly indicative of the phase separation. Blends of polyanhydrides with poly(caprolactone) [poly(CL)] are completely immiscible. Degradation studies in blends of poly(CL) with poly(dodecanedioic anhydride) [poly(DD)] indicated that the anhydride component degraded rapidly and was released from the blend, without affecting the poly(CL) degradation.[24] Other studies include the characterization of microphase-separated copolymers of poly(SA) with poly(CPH) or poly(ethylene glycol).[28,29] The phase diagram for the poly(CPH)/poly(SA) blend system has been determined using SAXS, optical microscopy, and molecular simulations, while the blends of poly(SA) and PEG were characterized by DSC and infrared (IR) spectra. The poly(CPH)/poly(SA) system exhibits an upper critical solution temperature behavior.

Stability

The stability of polyanhydrides has been studied in solid state and in dry chloroform. Aromatic polyanhydrides such as poly(CPP), poly(CPH), and poly(CPM) maintained their original molecular weight for at least one year in solid state upon storage under dry argon or vacuum at 21°C. In contrast, aliphatic polyanhydrides such as poly(SA) have a high-degradation rate at the same storage conditions.[1] Studies performed with GPC revealed that M_w of polyanhydrides tends to undergo a rapid decrease initially, and later a constant stabilized decrease in molecular weight is observed. The decrease in molecular weight was explained by an internal anhydride interchange mechanism resulting in ring formation, as revealed by ^{1}H NMR. This mechanism was supported by the fact that the decrease in molecular weight was reversible and heating of the depolymerized polymer at 180°C for 20 min yielded the original high molecular weight polymer.[24] It is important to mention that polyanhydrides experienced significant weight loss when stored at ambient conditions in which water attacks the anhydride bonds.

The stability of polyanhydrides in a solution was studied using chloroform under dry nitrogen atmosphere at 37°C.[1] The aromatic polyanhydrides remained stable under these conditions during a three-day period, while copolymers with aliphatic SA had a significant molecular weight loss during the same time period. Therefore, polyanhydrides can be processed in a solution environment as long as the time is not extended more than this period.

γ-Irradiation methods have been utilized for the sterilization of polyanhydrides. In this technique, aliphatic and aromatic homo- and copolymers were irradiated at 2.5 Mrad, and the chemical structure as well as the physical properties were found to be the same before and after irradiation.[24] The studies showed that saturated polyanhydrides are stable during γ-irradiation, as a slight increase in molecular weight was observed. Electron paramagnetic resonance (EPR) spectroscopy was used to characterize free radicals in γ-sterilized polyanhydrides.[30] Polymers with high melting temperatures produced the highest yields of room temperature radicals, which in turn transform into less conjugated polyanhydrides that leads to lower molecular weight polymers.

Degradation and Erosion

Polymer erosion (i.e., mass loss) is a complex process that is determined by numerous factors, including the molecular weight loss (degradation), swelling, dissolution and diffusion of oligomers and monomers, and morphological changes.[31] Polyanhydrides undergo degradation prior to erosion, as a consequence of the chemical instability of the anhydride bond. Thus, degradation and erosion are limited to the surface, as water does not penetrate into the device.[32]

Erosion kinetics is complicated when the anhydride monomers of a copolymer system exhibit microphase separation that leads to the erosion of different phases

at different rates. The erosion of a fast eroding phase may leave the slow eroding phase intact.[33] At this point, the monomer solubility plays a major role in polyanhydride erosion kinetics, as monomers are accumulated in eroding zones of the matrix and its dissolution will depend on the pH of the microenvironment.[34] It is known that the saturation concentration of the monomers CPH, SA, and CPTEG is a function of pH and that, at a particular pH, the order of solubility of the monomers is CPTEG > SA > CPH. This provides valuable information when describing the drug release from polymers containing any of these monomers.[19,33]

POLYANHYDRIDE-BASED DRUG DELIVERY SYSTEMS

Biocompatibility

The biocompatibility of implantable polyanhydride disks was studied in the brain of rats, rabbits, monkeys, and eventually in human clinical trials.[35] Wafers of poly(CPP : SA) and poly(FAD : SA) were implanted in the frontal lobes of rats, rabbits, and monkeys. In all these studies, the animals receiving the implants showed no behavioral changes or neurological deficits, indicating that the polymers were not invoking a systemic or local toxicity. To determine how the body metabolized the poly(CPP : SA), radio-labeled copolymers were implanted in the brains of rats.[21] Seven days after the implantation, 40% of the ^{14}C SA-labeled polymer had been excreted as CO_2, 10% was excreted along with the urine, 2% with the feces, and 10% still in the implanted device. In the same period only 4% of the ^{14}C CPP-labeled polymer was excreted along with the urine and feces.

The biocompatibility of poly(CPP), poly(TA), and copolymers of CPP : SA and CPP : TA implanted in the corneas of rabbits was studied.[36,37] Six weeks after implantation, the cornea remained clear and showed no evidence of corneal edema or neovascularization, indicating biocompatibility of the polymer matrix implant.

Subcutaneous implants of 20 : 80 CPP : SA copolymer were administered in rats at doses of 40 and 120 times the size that is to be used in humans. The purpose of these experiments was to test the systemic toxicity of the polymers. Eight weeks after implanting the disks, the rats were sacrificed and their organs underwent histopathological evaluations. In general, there was little to no difference between the organs of the experimental group (receiving the implant) and those of the control group. Again, in all the cases the polymers underwent degradation and were found to cause minimal inflammation at the site of implantation. Thus, polyanhydrides are inert and suitable for in vivo drug delivery.[38]

Drug/Polymer Interactions

When selecting polymers as drug delivery carriers, it is necessary to establish whether the polymer will react with the incorporated or the released drug. Three factors need to be considered: the reactivity of the drug, the hydrophobicity of the drug, and the fabrication method. The reactivity of CPP : SA copolymer with the para substituted anilines: *p*-nitroaniline (PNA), *p*-bromoaniline, *p*-anisidine, and *p*-phenylenediamine was examined.[37] The model drugs were incorporated into the polymer matrix using injection and compression molding. When injection molding was used to encapsulate the drugs at 120°C, the more reactive drugs (*p*-bromoaniline, *p*-anisidine, and *p*-phenylenediamine) reacted with the polymer forming amides. However, when the drugs were incorporated into the polymer matrix using compression molding at room temperature, the drugs did not react with the polymer during the fabrication process.

The hydrophobicity of the drug can also influence the interactions between the drug and the polymer. When hydrophilic dyes (acid orange and brilliant blue) were encapsulated in polyanhydrides, the T_m's of the polymers were unchanged. When hydrophobic dyes (PNA and methyl red) were encapsulated, the T_m's of the polymers changed thereby indicating an interaction between the polymer and the drug.[39]

Device Fabrication

Polyanhydride drug delivery devices have been fabricated as implantable[40] and injectable devices. Implantable devices are fabricated by compression molding, melt compression, or solvent casting. The first step of compression molding is to obtain a fine powder of the drug and the polymer. The powders are physically mixed and placed in a piston mold. The wafer is formed by applying a pressure (typically 30 kpsi) and by heating the sample to a temperature 5–10°C above the T_g of the polymer.[1,41] One drawback of this method is the uneven distribution of the drug in the polymer, leading to poor reproducibility. The Gliadel system is a compression-molded wafer. To overcome the problem of uneven drug distribution, the drug and polymer are spray dried together to form microspheres. The microspheres are then compression molded to form the wafer.

An alternative to compression molding is melt compression. This procedure requires the polymer and drug to be heated 10°C above the T_m of the polymer, forming a viscous solution.[1,41] The solution can then be either placed in a conventional mold under low pressure or injection molded. This fabrication method results in an even drug distribution. However, the

elevated temperatures needed to melt the polymer could cause adverse reactions in temperature sensitive drugs such as proteins.

Solvent casting is done by codissolving or suspending the drug in the polymer solution. The solution is then poured into a flat open mold and cooled on dry ice. The resulting film is often fragile. If the drug is not soluble in the polymer, it will settle on the bottom of the film, leading to an uneven drug distribution.[27,30]

To form an injectable drug delivery device, the drug is loaded into polymer microspheres. Drug-loaded polyanhydride microspheres have been fabricated using different methods. The most common method is the solvent extraction method, which includes water/oil/water (w/o/w), water/oil/oil (w/o/o), or solid/oil/oil (s/o/o) (dependent on whether the drug is soluble in the polymer solvent). In the w/o/w method, the drug (typically proteins) is dissolved in an aqueous phase and then emulsified with a larger volume of polymer that is dissolved in an organic solvent, typically methylene chloride. The inner emulsion is then added to a larger volume of water that contains a surfactant, usually PVA, and allowed to stir for several hours to extract the solvents. In the case of w/o/o or s/o/o the outer aqueous PVA phase is replaced with an immiscible organic solvent, i.e., silicon oil. The spheres are typically collected by either centrifugation or filtration.

Microspheres can also be fabricated by the hot-melt procedure; however, this method is not ideal for encapsulating temperature sensitive drugs such as proteins.[1] Spray drying or atomizing the polymer and drug together can also be used to fabricate microspheres. This method requires the use of either a spray dryer or atomization.[1,42] In the case of a spray dryer, the polymer/drug suspension is pumped into the spray drier. As the suspension is sprayed, a stream of air causes the polymer spheres to harden. Microspheres are fabricated by atomization by passing the drug/polymer suspension through an atomizing nozzle. As the polymer/drug spheres leave the nozzle, they are collected in a bath of liquid nitrogen sitting on top of a frozen layer of ethanol.[42] The liquid nitrogen/ethanol bath is then stored at $-80^{\circ}C$ for three days. During this time the ethanol slowly thaws and the frozen microspheres fall into it. As the microspheres sit in the ethanol, the organic solvent (methylene chloride) slowly diffuses out, leaving solid spheres that could be collected by filtration.

In Vitro Release

The rate at which an encapsulated drug will be released from a polyanhydride device, either a wafer or a microsphere, is strongly dependent on polymer composition and drug distribution. Other factors that contribute to the release rate of drugs includes: fabrication technique, size/shape of the device, and pH of the surrounding media.

The hydrophobicity of a drug influences its distribution within the polymeric device.[27] *p*-Nitroaniline and disperse yellow have higher affinities for poly(CPH) and poly(SA), respectively. The two drugs were encapsulated in tablets of poly(CPH), poly(SA), and copolymers of the two to determine if the drugs would partition into the more favorable polymer microdomain. When the dominant polymer had a low affinity for the drug, a burst effect was seen. In the case of 50 : 50 CPH : SA copolymer, each drug followed the release of the monomers, indicating that the drug was partitioning into the more favorable domain.

The size of the device may also influence drug distribution and release rate.[43] Monodisperse microspheres of differing average diameters were studied to determine the influence of size of the device on the delivery. Smaller diameter microspheres showed a more prolonged release rate of drug than did microspheres that had a large diameter. As the diameter increased, the time it took for the microsphere to form by precipitation increased; thus, increasing the time for the drug to segregate toward the surface of the microspheres.

As the anhydride bonds in the polymer backbone are hydrolyzed, carboxylic acid is formed. The formation of the acidic degradation products reduces the local pH of the eroding device. The diffusion of the acidic degradation products away from the device is expedited when the device is in a basic solution. However, when the device is in an acidic solution, the erosion process is slowed significantly.[1,44]

Polyanhydrides have also been investigated as protein carriers.[45] Poly(SA) and 20 : 80 CPH : SA copolymer microspheres were found to conserve both the primary structure of the released protein [bovine serum albumin (BSA)] and the secondary structure of the encapsulated and released protein, and showed a sustained delivery for approximately 15 and 30 days, respectively. As the CPH content in the copolymer increased, the secondary structure of BSA was not conserved, as indicated by the steep decrease in the α-helix content.

In Vivo Delivery

The 20 : 80 CPP : SA copolymer was the first polyanhydride to be clinically tested in humans. The copolymer was used to encapsulate BCNU, a chemotherapeutic drug used to treat a fatal form of brain cancer known as glioblastoma multiforme. Along with the polymer, BCNU was codissolved, and the disks were fabricated by compression molding. The preclinical trials in rat, rabbit, dog, and monkey brains demonstrated the

effectiveness of the polymer in delivering an active drug that remained localized, minimizing systemic reaction to the drug.[46] The wafer, once implanted into the brain of the glioblastoma patients, releases the BCNU for approximately three weeks.[47] Scanning electron microscopy was used to monitor the erosion of the wafer both in vitro and in vivo.[48] It was found that the erosion of the wafer was controlled by diffusion of BCNU and erosion of the polymer. The delivery device was approved in 1996 by the USFDA for use in conjunction with surgery for patients suffering from recurrent glioblastoma. In 2003, the USFDA approved the use of the device in newly diagnosed advanced cases of malignant gliomas, in conjunction with surgery and radiation.

The use of 20 : 80 CPP : SA and 18 : 82 FAD : SA copolymers disks as drug delivery devices for carboplatin, a treatment for glioma, was also investigated in rodents.[49] The majority of the drug was released in seven days from the CPP : SA copolymer disk, and 65% of the drug was released from the FAD : SA copolymer disk in seven days. This method of delivery was more effective than systemic therapy and did not cause systemic toxicity.

A separate polyanhydride system has also been investigated for the treatment of osteomyelitis, a bone infection typically caused by bacteria.[50–52] Copolymer implants (50 : 50 FAD : SA) containing gentamicin were tested in the backs of rats, in the infected tarsocrural joints of horses, and in humans with infected prosthetic hips or knees.[51] In all the cases, the local delivery of gentamicin was successful and the systemic exposure to the drug was avoided.

CONCLUSIONS

Polyanhydrides are promising as biomaterials because they possess a unique combination of properties that include hydrolytically labile backbone, hydrophobic bulk, and chemistry that can be easily combined with other functional groups to design novel materials. These materials are primarily surface-erodible and offer the potential to stabilize protein drugs and sustain release from days to months. The microstructure characteristics of copolymer systems can be exploited to tailor drug release profiles. The versatility of polyanhydride chemistry promises a new class of drug release systems for specific applications.

REFERENCES

1. Tamada, J.; Langer, R. The development of polyanhydrides for drug delivery applications. J. Biomater. Sci. Polym. Ed. **1992**, *3* (4), 315–353.
2. Hill, J.; Carothers, W. Studies of polymerization and ring formation. XIV. A linear superpolyanhydride and a cyclic dimeric anhydride from sebacic acid. J. Am. Chem. Soc. **1932**, *54*, 1569–1579.
3. Domb, A.J.; Ron, E.; Langer, R. Poly(anhydrides) based on aliphatic–aromatic diacids. Macromolecules **1989**, *22*, 3200–3204.
4. Domb, A.J.; Mathiowitz, E.; Ron, E.; Giannos, S.; Langer, R. Polyanhydrides. IV. Unsaturated and crosslinked polyanhydrides. J. Polym. Sci. A **1991**, *29*, 571–579.
5. Bucher, J.; Slade, W.C. Anhydrides isophthalic and terephthalic acids. J. Am. Chem. Soc. **1909**, *31*, 1319–1321.
6. Domb, A.J.; Amselem, S.; Shah, J.; Maniar, M. Polyanhydrides: synthesis and characterization. Adv. Polym. Sci. **1993**, *107*, 94–141.
7. Staubli, A.; Ron, E.; Langer, R. Hydrolytically degradable amino acid containing polymers. J. Am. Chem. Soc. **1990**, *112*, 4419–4424.
8. Hanes, J.; Chiba, M.; Langer, R. Degradation of porous poly(anhydride-co-imide) microspheres and implications for controlled macromolecule delivery. Biomaterials **1998**, *19* (1–3), 163–172.
9. Jiang, H.L.; Zhu, K.J. Preparation, characterization and degradation characteristics of polyanhydrides containing poly(ethylene glycol). Polym. Int. **1999**, *48*, 47–52.
10. Weinberg, J.M.; Gitto, S.P.; Wooley, K.L. Synthesis and characterization of degradable poly(silyl ester). Macromolecules **1998**, *31*, 15–21.
11. Erdmann, L.; Macedo, B.; Uhrich, K.E. Degradable poly(anhydride ester) implants: effects of localized salicylic acid release on bone. Biomaterials **2000**, *21* (24), 2507–2512.
12. Erdmann, L.; Uhrich, K.E. Synthesis and degradation characteristics of salicylic acid-derived poly(anhydride-esters). Biomaterials **2000**, *21* (19), 1941–1946.
13. Schmeltzer, R.; Anastasiou, T.; Uhrich, K. Optimized synthesis of salicylate-based poly (anhydride-esters). Polym. Bull. **2003**, *49* (6), 441–448.
14. Anastasiou, T.; Uhrich, K. Aminosalicylate-based biodegradable polymers: synthesis and in vitro characterization of poly(anhydride-esters) and poly(anhydride-amides). J. Polym. Sci. A: Polym. Chem. **2003**, *41*, 3667–3679.
15. Jiang, H.; Zhu, K. Pulsatile protein release from a laminated device comprising polyanhydrides and pH-sensitive complexes. Int. J. Pharm. **2000**, *194* (1), 51–60.
16. Campo, C.J.; Anastasiou, T.; Uhrich, K.E. Polyanhydrides. The effects of ring substitution changes on polymer properties. Polym. Bull. (Berlin) **1999**, *42* (1), 61–68.

17. Anastasiou, T.J.; Uhrich, K.E. Novel polyanhydrides with enhanced thermal and solubility properties. Macromolecules **2000**, *33* (17), 6217–6221.
18. Anseth, K.S.; Shastri, V.R.; Langer, R. Photopolymerizable degradable polyanhydrides with osteocompatibility. Nature Biotech. **1999**, *17* (2), 156–159.
19. Torres, M.P.; Vogel, B.M.; Narasimhan, B.; Mallapragada, S.K. Synthesis and characterization of novel polyanhydrides with tailored erosion mechanisms. J. Biomed. Mater. Res. Submitted.
20. Domb, A.J.; Ehrenfreund, T.; Golenser, J.; Langer, R.; Israel, Z. Biodegradable polyanhydrides: synthesis and drug delivery applications. Biodegrad. Polym. **2003**, *2*, 121–151.
21. Domb, A.J.; Amselem, S.; Langer, R.; Maniar, M. Polyanhydrides as carriers of drugs. In *Biomedical Polymers Designed-to-Degrade Systems*; Shalaby, S.W., Ed.; Hanser Publishers: New York, 1994; 69–96.
22. Leong, K.W.; Simonte, V.; Langer, R. Synthesis of polyanhydrides: melt-polycondensation, dehydrochlorination, and dehydrative coupling. Macromolecules **1987**, *20* (4), 705–712.
23. Narasimhan, B.; Kipper, M.J. Surface-erodible biomaterials for drug delivery. Adv. Chem. Eng. **2004**, *29*, 169–218.
24. Kumar, N.; Langer, R.S.; Domb, A.J. Polyanhydrides: an overview. Adv. Drug Delivery Rev. **2002**, *54* (7), 889–910.
25. McCann, D.L.; Heatley, F.; D'Emanuele, A. Characterization of chemical structure and morphology of eroding polyanhydride copolymers by liquid-state and solid-state 1H n.m.r. Polymer **1999**, *40* (8), 2151–2162.
26. Mathiowitz, E.; Ron, E.; Mathiowitz, G.; Amato, C.; Langer, R. Morphological characterization of bioerodible polymers. 1. Crystallinity of polyanhydride copolymers. Macromolecules **1990**, *23*, 3212–3218.
27. Shen, E.; Kipper, M.J.; Dziadul, B.; Lim, M.-K.; Narasimhan, B. Mechanistic relationships between polymer microstructure and drug release kinetics in bioerodible polyanhydrides. J. Control. Release **2002**, *82* (1), 115–125.
28. Chan, C.-K.; Chu, I.-M. Phase behavior and miscibility in blends of poly(sebacic anhydride)/poly(ethylene glycol). Biomaterials **2002**, *23* (11), 2353–2358.
29. Kipper, M.J.; Seifert, S.; Thiyagarajan, P.; Narasimhan, B. Understanding polyanhydride blend phase behavior using scattering, microscopy, and molecular simulations. Polymer **2004**, *45* (10), 3329–3340.
30. Domb, A.J.; Elmalak, O.; Shastri, V.R.; Ta-Shma, Z.; Masters, D.M.; Ringel, I.; Teomim, D.; Langer, R. 8. Polyanhydrides. Drug Target. Delivery **1997**, *7*, 135–159.
31. Gopferich, A.; Tessmar, J. Polyanhydride degradation and erosion. Adv. Drug Delivery Rev. **2002**, *54* (7), 911–931.
32. Kumar, N.; Ravikumar, M.N.V.; Slivniak, R.; Krasko, M.Y.; Domb, A.J. Biodegradation of polyanhydrides. In *Biopolymers*; Matsumura, S., Steinbuechel, A., Eds.; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2003; Vol. 9, 423–456.
33. Kipper, M.; Narasimhan, B. Molecular description of erosion phenomena in biodegradable polymers. Macromolecules **2005**, *38* (5), 1989–1999.
34. Goepferich, A.; Langer, R. The influence of microstructure and monomer properties on the erosion mechanism of a class of polyanhydrides. J. Polym. Sci. A: Polym. Chem. **1993**, *31*, 2445–2458.
35. Katti, D.S.; Lakshmi, S.; Langer, R.; Laurencin, C.T. Toxicity, biodegradation and elimination of polyanhydrides. Adv. Drug Delivery Rev. **2002**, *54*, 933–961.
36. Brem, H.; Kader, A.; Epstein, J.I.; Tamargo, R.J.; Domb, A.; Langer, R.; Leong, K.W. Biocompatibility of a biodegradable, controlled-release polymer in the rabbit brain. Selec. Cancer Ther. **1989**, *5* (2), 55–65.
37. Leong, K.W.; D'Amore, P.; Marletta, M.; Langer, R. Bioerodible polyanhydrides as drug-carrier matrices. II. Biocompatibility and chemical reactivity. J. Biomed. Mater. Res. **1986**, *20*, 51–64.
38. Kumar, M.N.; Ravi, V.; Domb, A.J. Drug delivery, controlled. In *Encyclopedia of Biomaterials and Biomedical Engineering*; Wnek, G.E., Bowlin, G.L., Eds.; Marcel Dekker, Inc.: New York, 2004; 1, 467–477.
39. Shen, E.; Pizsczek, R.; Dziadul, B.; Narasimhan, B. Microphase separation in bioerodible copolymers for drug delivery. Biomaterials **2001**, *22* (3), 201–210.
40. Jain, J.P.; Mode, S.; Domb, A.J.; Kumar, N. Role of polyanhydrides as localized drug carriers. J. Control. Rel. **2005**, *103* (3), 541–563.
41. Chasing, M.; Lewis, D.; Langer, R. Polyanhydrides for controlled drug delivery. Biopharm. Manuf. **1988**, *2*, 33–39.
42. Gombotz, W.R.; Healy, M.S.; Brown, L.R. *A Very Low Temperature Casting of Controlled Release Microspheres*; Enzytech, Inc.: U.S.A., 1991.
43. Berkland, C.; Kipper, M.J.; Narasimhan, B.; Kim, K.K.; Pack, D.W. Microsphere size, precipitation

kinetics and drug distribution control drug release from biodegradable polyanhydride microspheres. J. Control. Release **2004**, *94* (1), 129–141.
44. Chasin, M.; Domb, A.; Ron, E.; Mathiowitz, E.; Langer, R.; Leong, K.; Laurencin, C.; Brem, H.; Grossman, S. Polyanhydrides as drug delivery systems. Drugs Pharm. Sci. **1990**, *45*, 43–70.
45. Determan, A.S.; Trewyn, B.G.; Lin, V.S.-Y.; Nilsen-Hamilton, M.; Narasimhan, B. Encapsulation, stabilization, and release of BSA-FITC from polyanhydride microspheres. J. Control. Release **2004**, *100* (1), 97–109.
46. Brem, H. Polymers to treat brain tumours. Biomaterials **1990**, *11*, 699–701.
47. Brem, H.; Mahaley, S.; Vick, N.A.; Black, K.L.; Schold, S.C.; Burger, P.C.; Friedman, A.H.; Ciric, I.S.; Eller, T.W.; Cozzens, J.W.; Kenealy, J.N. Interstitial chemotherapy with drug polymer implants for the treatment of recurrent gliomas. J. Neurosurg. **1991**, *74*, 441–446.
48. Dang, W.; Daviau, T.; Brem, H. Morphological characterization of polyanhydride biodegradable implant Gliadel during in vitro and in vivo erosion using scanning electron microscopy. Pharm. Res. **1996**, *13* (5), 683–691.
49. Olivi, A.; Ewend, M.G.; Utsuki, T.; Tyler, B.; Domb, A.J.; Brat, D.J.; Brem, H. Interstitial delivery of carboplatin via biodegradable polymers is effective against experimental glioma in the rat. Cancer Chemother. Pharmacol. **1996**, *39*, 90–96.
50. Perez, C.; Castellanos, I.J.; Costantino, H.R.; Al-Azzam, W.; Griebenow, K. Recent trends in stabilizing protein structure upon encasulation and release from bioerodible polymers. J. Pharm. Pharmacol. **2002**, *54*, 301–313.
51. Stephens, D.; Li, L.; Robinson, D.; Chen, S.; Chang, H.-C.; Liu, R.M.; Tian, Y.; Ginsburg, E.J.; Gao, X.; Stultz, T. Investigation of the in vitro release of gentamicin from a polyanhydride matrix. J. Control. Release **2000**, *63*, 305–317.
52. Li, L.C.; Deng, J.; Stephens, D. Polyanhydride implant for antibiotic delivery-from bench to the clinic. Adv. Drug Delivery Rev. **2002**, *54*, 963–986.

Polybutadiene

William L. Hergenrother
Bridgestone Americas Center for Research and Technology, Bridgestone/Firestone Research LLC, Akron, Ohio, U.S.A.

Mark DeDecker
Dan F. Graves
Firestone Polymers, Bridgestone/Firestone Inc., Akron, Ohio, U.S.A.

INTRODUCTION

As background for the preparation of this article, a Google™ search of ''polybutadiene'' (PB) was made and 24,700 hits were generated. The addition of other terms to the quoted search reduced the number of hits to 1668 for structural features, 660 for modifications, 525 for uses of blends, and 52 for polymerization variations. This still is an impressive and formidable number of potential references that show the significant activity and utilization of PB. This does not include styrene/butadiene (BD)/styrene triblock and random copolymers (SBR), that are separate subjects and will not be considered. The reason for the plethora of literature available on PB is derived from a combination of factors. Primarily the base monomer BD is abundant, inexpensive, and can be converted readily to a variety of different reactive polymeric structures.

OVERVIEW

The polymerization can be accomplished by the use of free-radicals and anionic or transition metal-mediated cationic systems. Gaseous, bulk, aqueous suspension, or a solution in organic solvents has been used for the processes mentioned. These processes are simplified by the ease of purification of the low-boiling BD monomer. Another factor that contributes to interest in the PBs is the tremendous range of microstructure composition and molecular weight that can be obtained in the homopolymerization to PB. With the multitude of conditions that have already been mentioned, further refinements in the catalyst or initiators used allow the obtainment of *cis*-1,4-, *trans*-1,4-, and 1,2-(vinyl) placement of BD in PB to vary in all proportions from 0% to nearly 100% of the microstructure placement desired. These variations are not insignificant in that the polymers can vary from low-temperature elastomers to high-melting thermoplastics.

The composition and distribution of the repeat structures within the polymer chain define the flexibility of the polymer chain.[1] Because many properties depend on this chain mobility, polymer composition is carefully controlled. In addition to chain mobility, polymer composition also defines the solubility parameter of the polymer, which is a critical property relative to the type of solvents in which the polymer is soluble, the ability of the polymer to accept and hold oil, and the relative compatibility of the polymer with other polymers.[2] Basically, these properties all involve polymer–solvent interaction with differences associated with both the structure and the molecular weight of the polymer and the solvent.

End-group functionalization of the above polymers with the presence of reactive double bonds along the backbone allows a tremendous variety of chemistry to be utilized, such that a myriad of PBs can be produced. A more manageable breakdown of the current significant activity for the use of PB is presented in the entry. In deference to a concise summary, emphasis is placed on the scientific literature over the past 5 yr and the published worldwide patent activity from 1981.

Molecular Weight

Typically, a polymer is composed of a mixture of many molecules of different lengths, each of which can contain a different distribution of incorporation of the possible repeat units. Most conveniently, an average is used to characterize these features. Reaction conditions and catalyst levels are varied to form everything from low-molecular weight oil to extremely viscous materials. Thus, polymers will range in the number of average molecular weight (M_n) from about 1 to >540 kg/mol, or an average degree of polymerization (DP), from about 18 to >10,000 monomer units per molecule. The bulk polymer viscosity increases to extremely high values as the chain length is increased. At very low chain length, this increase is linear with molecular weight until the chains are long enough to become entangled. Above the entanglement molecular weight, the viscosity increases to the 3.4–3.5th power of molecular weight.[3] In addition to viscosity, many

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007906

other physical properties increase in value proportionate to the increasing molecular weight of the polymer. The 3.5th power relationship of molecular weight to bulk viscosity rapidly causes a decreased ability to compound, extrude, and mold the polymer into potentially useful material.[4] As a result, some compromise must be reached or some creative new avenues must be exploited to obtain the best possible material. Many of these approaches are apparent in this entry.

Molecular Weight Distribution

Although there are many different statistical ways to describe any population, the M_n, weight average molecular weight (M_w), and polydispersity index (M_w/M_n) have been widely used because they are readily understood in physical terms, and they can be measured directly in the laboratory by size exclusion chromatography (SEC).[5,6] The polydispersity has a lower limit of 1 (i.e., all chains of exactly the same length) and can vary to 5 or greater. Typical values for commercial polymers range from 1.2 to 5, with those under 2 considered relatively narrow and those over 2.8 considered broad in distribution. The SEC measurement sometimes has been referred to as gel permeation chromatography. These measurements of molecular weights and distributions have become commonplace, such that automatic sampling of polymerization, processing, and final product testing are routinely done and calculated to provide data for process control.[6]

Branching

The concept of a polymer chain implies two ends per chain. However, because of the nature of the process that is used to form the polymer, the chain may contain one or more branch points resulting in multiple ends per chain. These ends can have a marked influence on polymer performance. Branching, molecular weight, and molecular weight distribution also have been shown to affect processability.[7] The optimum macrostructure often represents a compromise between processing and ultimate performance.

Network Structure

A critical requirement for obtaining the desirable properties from a rubbery material is the generation of a network structure. Charles Goodyear's discovery of ''vulcanization'' changed natural rubber from a material that became sticky when hot and brittle when cold to a material that could be used over a wide range of conditions, when he found a way to chemically connect the individual polymer chains into a three-dimensional network. This network now allows the support of considerable stress and retracts rapidly on release of the stress. Efforts to elucidate this process are still of interest.[8,9] The terms vulcanization, ''rubber cure,'' and ''cross-linking'' all refer to the same general phenomenon. For most rubber systems, the network is formed after the polymer is mixed with additive and curatives (compounded) and molded into the desired final shape. Once cross-linked, the material can no longer be processed. If cross-linking occurs prior to compounding or molding, the material is referred to as gelled, and it cannot be used. Most rubber is used in a compounded and cured form. There is an optimum cross-link density for many failure properties, such as tensile strength and tear.

Glass Transition Temperature

Glass transition temperature (T_g) is usually measured using thermal methods, such as differential scanning calorimetry, which look at the change in heat capacity when a material goes through the glass transition. Another very useful method is to use dynamic mechanical properties where the polymer is subjected to a temperature sweep in a dynamic mechanical spectrometer from a very low temperature ($-120^\circ C$) to well above room temperature ($+100^\circ C$). If the modulus is plotted vs. temperature, there is a sharp decrease as the polymer warms to above the T_g. There is also a peak in the energy loss property known as tan δ.

For a polymer to behave as a rubbery material, it is necessary for the chain segments to have great mobility. As the temperature is lowered, this ability of chain segments to move decreases until T_g is reached where all large-scale chain segment motion is prevented. Below this temperature, the rubber becomes a hard and brittle glassy material. Above this temperature amorphous plastics polymers, such as polystyrene, can exhibit rubbery character if the molecular weight is sufficiently high. It has been shown that the wet traction and wear of a passenger tire are directly dependent on the T_g.[2,10] Control of T_g is possible in PB by controlling the microstructure of the polymer chain. Vinyl placements restrict chain mobility and thus raise the T_g. The average microstructure of a PB polymer can be readily tailored to obtain any T_g of the final product within the range of 100% *cis*-1,4- to 100% of 1,2-microstructure.

Crystallinity

Crystallinity can be measured using the same thermal and dynamic mechanical methods used for measuring T_g; however, the melting transition (T_m) is much sharper and is at a higher temperature than the T_g. The majority of the PB is used as an elastomer for which

a low T_g is required. Syndiotatic 1,2-PB and crystalline high *trans*-1, 4-PB are exceptions and have been used as a reactive ''plastic'' that has a T_m in excess of 150°C. Very little emphasis is placed on these materials as it represents an extremely small portion of the utility of PB polymers and has been described adequately.[11]

PRODUCTION

The worldwide demand for elastomers was estimated to be 19.44 million metric tons for 2008, excluding NR latex materials.[a] Of this demand, 55% are synthetic elastomers of various kinds. Depending on their use elastomers have been classified into general purpose and speciality. The major general purpose elastomers are natural rubber, SBR rubber, PB rubber, isoprene rubber, and ethylene–propylene rubber. These rubbers are used in tires, mechanical goods, and similar applications. Specialty elastomers provide unique properties such as oil resistance or extreme heat stability. Although this differentiation is rather arbitrary, it also tends to classify the polymers according to the volumes used. About 74% of the rubbers used consist of SBR, butadiene rubber (BR), and ethylene–propylene rubber.

The worldwide consumption of PB in 2003 was about 2.1 million metric tons and is forecast to increase to 2.5 million metric tons by 2008.[a] The total U.S. consumption of BR is about 0.41 million metric tons, of which over 70%, as *cis*-1,4-PB (*cis*-PB), is used in the manufacture of tires. The next largest use of ~20% is mainly anionic PB for the plastic industry. The remainder is used in a variety of applications that is seen in the following sections.

A search of the worldwide patent base since 1981 revealed 9980 distinct patents that have been issued or are published applications of the utilization of PB. Computer technology is now available that will list and generate a map indicating the similarities among these patents.[b] Such a map with this broad database is seen in Fig. 1. Here, the peaks are labeled to indicate clusters of similar patents. For instance, the four most intense collections of points are shown as white peaks. Each peak represents the location of the maximum of over 400 patents related to the labeled category, while the next contour line shows the distribution of less closely related patents that also have a lower level of similarity to the peak.

[a]Worldwide Rubber Statistics 2004; International Institute of Synthetic Rubber Producers, Inc.: Houston; 9.

[b]These maps were created using MicroPatent, LLC's Aureka® IP analysis platform and Themescape® patent mapping tool. MicroPatent, LLC is a Thomson business. Aureka and Themescape are registered trademarks of Micropatent LLC.

RUBBER COMPOUNDING

The vulcanization discovered by Charles Goodyear was essential for the use of rubber products, but the impetus for rubber use was supplied by compounding the rubber with a vast number of materials. Without compounding and curing, the rubber utilization would only be a small fraction of what it is today. Any particular compounded rubber application has been optimized to produce very specific properties. These include appearance, processing, mechanical, electrical, chemical, and thermal properties, as well as cost. Developing such compounds requires a broad knowledge of material science and chemistry combined with experience. The use of designed experiments can greatly facilitate selecting the optimum compound formulation. The major components used in such compounds are curatives, reinforcing agents, fillers, plasticizers, and antidegradants.

Processing

Wide ranges of processes are used to convert a bale of rubber into an elastomeric product, such as a tire. The process generally starts with the compounding of the rubber with many of the previously stated ingredients. Many elastomeric stocks require more than one rubber to be included to obtain the desired performance goal. This has resulted in PB being added to many applications to impart specific key properties.

Rubber can be considered as a highly viscous state of matter that processes both as a liquid and as an elastomer. Mixing materials into rubber requires high shear, for which the simplest method is a double-roll mill, where the rubber is shear-mixed along with the other compounding ingredients in the bite of the mill. Large-scale mixing is most commonly done with a high-shear internal mixer called a Banbury. This mixing is a batch process although continuous internal mixers are also used.

Typically, the compounded rubber stocks need to be further processed for use. The process could be an injection or transfer molding into a hot mold, where it is cured. Extrusion of the rubber stock is used to make hose or tire treads and sidewalls. Another common process is calendaring, in which a fabric is passed through rolls where rubber is squeezed into the fabric to make fabric-reinforced rubber sheets for roofing membranes, belting or body plies for tires. The actual construction of the final product can be quite complex. For example, a tire contains in excess of six different compounded elastomeric stocks. All the components must be prepared and assembled with high precision so that the final cured product can operate smoothly at high speeds and last over 50,000 mi.

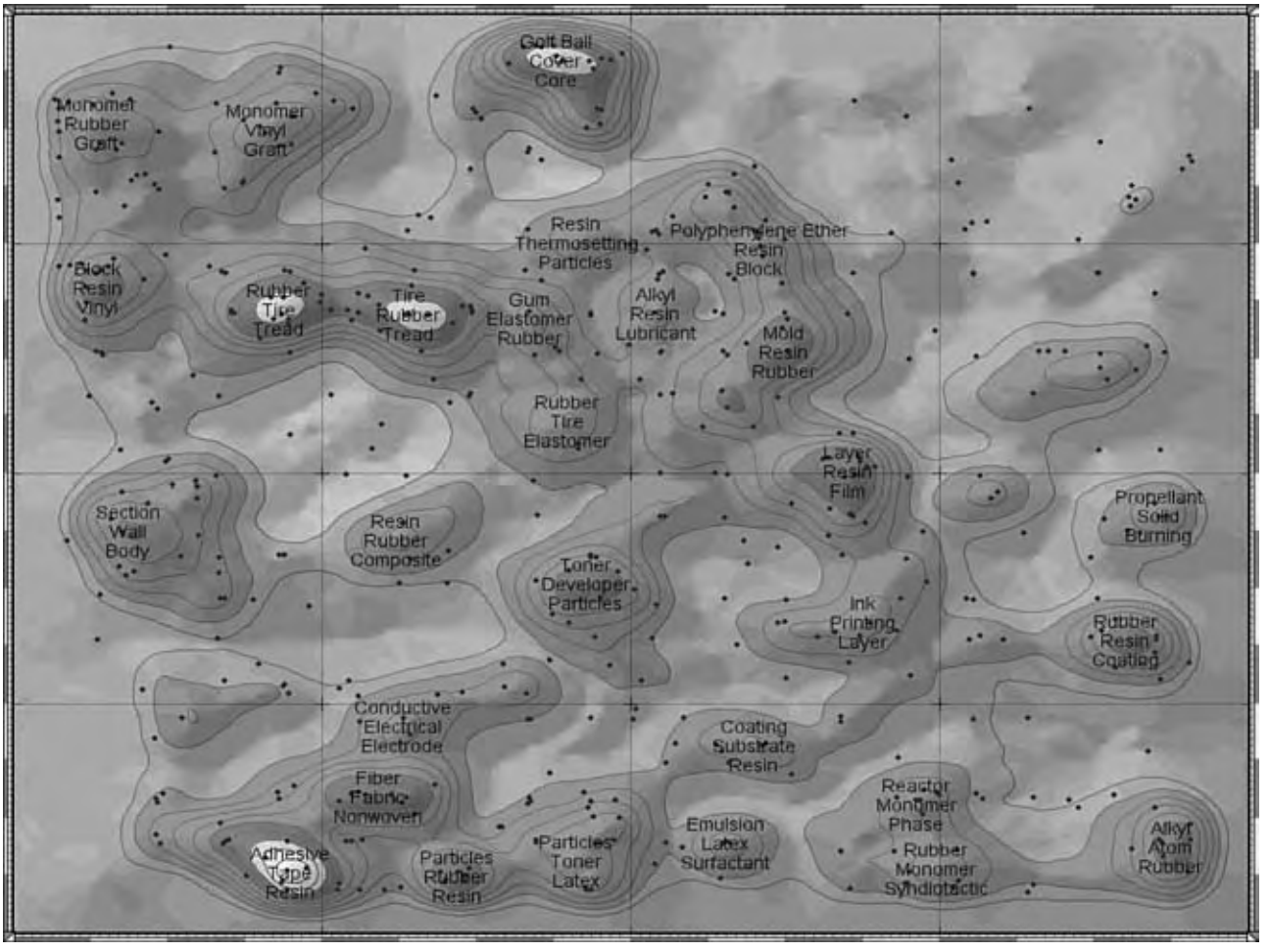


Fig. 1 Mapping of 9980 nonduplicated patents that have been published since 1981 involving the use of BR. *(View this art in color at www.dekker.com.)*

Monomer Production

Butadiene monomer can be produced by a number of different processes. The dominant method of production is as a by-product from the steam cracking of naphtha to produce ethylene. The BD is recovered from the C_4 fractions by extractive distillation.[12] Butadiene is generally produced by dehydrogenation or oxidative dehydrogenation of C_4 hydrocarbons.[13,14]

Polymer Production

Polybutadiene is usually produced by alkali metal or transition metal coordination solution processes, which allow the production of a variety of the desired DPs and microstructures. Alkali metal-based polymerization (usually alkyllithium) produces a product with about 36% *cis*-1,4-, 54% *trans*-1,4-, and 10% 1,2-PB. The polymerization process is conducted in an aliphatic hydrocarbon under an inert atmosphere in either a batch or a continuous mode. Because of the characteristics of this polymerization system, polymers of extremely narrow molecular weight distribution and low gel can be produced.[15] The narrowest distribution is produced via batch polymerization. The largest commercial use of anionic PB is in plastics modification. The PB is commonly characterized with a Mooney viscosity (ML_4) of about 52 with a 5.43% toluene solution viscosity at 25°C of about 160 cPs. Generally, other commercial PBs has an ML_4 range of 35–70 with a solution viscosity of 40–250 cPs.

Coupled (star branched) and end-functional PBs are possible using alkyllithium technology because of the presence of a living anion on the chain end. This anion is available for further reactions, which is discussed later. Upon the addition of polar agents (modifiers), such as ethers or amines, the alkyllithium initiators can produce PBs with vinyl contents up to 100%.[16] The vinyl content can be controlled by the ratio of

modifier to initiator and the polymerization temperature, with lower temperatures favoring increased vinyl formation. Even with high vinyl contents such polymers do not crystallize, because of the atactic nature of the vinyl units.

Both free-radical and anionic polymerization produce PBs that have mixed microstructures. Neither of these routes produces a rubber that can replace the high tack, green strength, or gum tensile of natural rubber. However, these PBs have been useful in many applications.

Emulsion polymerization

The formula developed to provide emulsion SBR (GR-S) during World War II was standardized with all rubber plants owned by the U.S. government.[17] Initiation occurs by the formation of a mercaptyl radical through reaction of the persulfate with the mercaptan, followed by chain propagation and growth of the polymer by the free-radical attack of the emulsified BD and styrene monomers. The molecular weight of the chain can be controlled by the concentration of mercaptan via a chain transfer mechanism. The growing polymeric radical chain can be terminated by reaction with mercaptan to form a new mercaptyl radical, which then can initiate an additional polymeric chain.[18] The terminated polymer produced above also can react with new growing chains, to lead to branching, which if the conversion of the monomer is too high can result in gel. Thus, above about 70% conversion the polymer becomes highly branched for use. Termination by the addition of free-radical scavengers, such as hydroquinone, is commonly used to stop polymerization before this point. The excess monomer is flashed off by steam stripping. Most of the emulsion-polymerized latex made today is SBR latex. However, PB latex is also made for specialized applications.

Solution Process

In 1955 investigators from the Firestone Tire and Rubber Company and the B. F. Goodrich Company announced the synthesis of polyisoprene with over 90% *cis*-1,4- structure. The work at Firestone was based on lithium metal catalysts, whereas the work at Goodrich was the result of using Ziegler–Natta-type coordination catalysts.[19,20]

The discovery of the ability of lithium-based catalysts to polymerize isoprene to give a high *cis*-1,4-polyisoprene was rapidly followed by the development of alkyllithium-based PB. The Firestone Tire and Rubber Company built the first commercial plant in 1960.

In addition to these engineering approaches to produce uniformly random SBR copolymer, the kinetics may be changed by the addition of polar modifiers. This action results in bringing the reactivity rates of the comonomers much closer together, but changes the microstructure of the PB portion. Whereas in the nonpolar system the vinyl content of the PB is around 10%, with polar additives vinyl contents of 20% and even higher are readily obtained. Care must be taken when polar modifiers are used as they can react with the growing chain, resulting in an undesired termination.[21,22]

High *cis*-PB is produced via solution processes using Ziegler–Natta-type transition metal catalysts. The major commercial catalysts of this type are based on titanium, cobalt, nickel, and neodymium.[23] Typically, transition metal is used in the form of a soluble metal salt, which can react with alkylaluminum or an alkylaluminum halide to reduce the metal to give the active species.[24–26] Because of the active nature of the catalyst mixture, the polymer solutions must be treated to deactivate or remove these materials from the final product. Many of these catalysts produce products with 90% or higher *cis*-1,4- content. The neodymium system is reported to produce the highest *cis*-(98–99%) with the most linear chain structure.[27] The highest branched *cis*-PB is produced with the Co system with Ni giving intermediate branching.[28–30]

All of the solution processes require high efficiency in recovering the solvent. The schematic in Fig. 2 shows how a typical solution polymerization plant is arranged. The most widely used process consists of termination of the polymerization and the addition of antioxidant to the polymer solution. The solution may be treated to remove initiator or catalyst residue and then transferred into an agitated steam-stripping vessel in which unreacted monomer and solvent are flashed off, leaving the rubber as particles (crumb) in water. The water/crumb slurry then is dewatered and dried. The recovered monomer/solvent is recirculated to a series of distillation columns to recover monomer and purify the solvent. As both the anionic and the coordination catalyst systems are highly sensitive to impurities such as water, the purification system is very critical for satisfactory process control.

The mapping shown in Fig. 1 includes references to SBR and styrene block copolymers in the PB search. Removing these citations from the database reduced the number to 4297, which can be seen mapped in Fig. 3. the area of high activity is centered on hydroxy terminated PB (HTPB). Low-M_n HTPB can be prepared by a variety of polymerization processes such as radical, anionic, or even using acyclic diene metathesis (ADMET).[31–33] The HTPB has a variety of uses as a propellant.[34–39] Other uses include reaction with epoxy resins, nylon, urethane, or even in the formulation of adhesives.[40–45] The use of HTPB as an oxygen scavenger in polyamide, polyvinyl alcohol, and multilayer

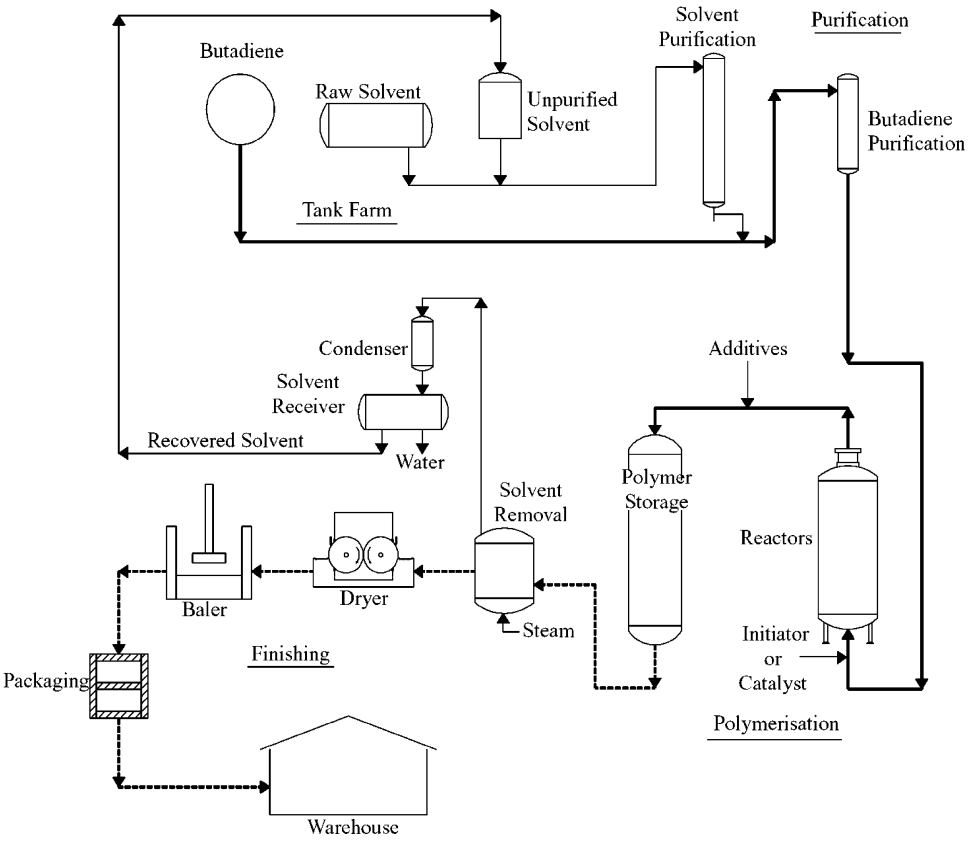


Fig. 2 Flow diagram for a typical solution process for the manufacture of PB. (Courtesy of Firestone Polymers, Akron, Ohio, U.S.A.)

composites or as a reactive component in plastic sheeting appears to be a growing market.[46–49] Another area of significant activity was in the production of syndiotatic PB with different metallic catalysts and additives.[50–53]

Now a closer look at the mapping of vinyl PB can be considered. A further refinement of the total PB citations revealed 926 references for vinyl PB, which are mapped in Fig. 4. Here, the aqueous preparation of syndiotatic 1,2-PB is seen along with the use of both syndiotatic and atactic PB in tire beads, silica containing treads, and even medical applications after cross-linking with peroxide or radiation.[54–62] Mixtures of catalysts have been used to allow amorphous 1,2-PB to be coproduced with syndiotatic PB in the same reactor.[63,64] Epoxidized vinyl PB can be used to reduce tire rolling resistance, toughen resin, and stabilize PVC.[65–67]

Further focusing of the mapping can be seen in Fig. 5, where 148 citations for high *cis*-PB are shown. The preparation of high *cis*-PB with Co, Nd, and Co with nitrogen ligands was revealed.[68–70] A significant use includes *cis*-PB as golf ball cores and covers with polyurethanes and unsaturated metal salts.[71–73] The market for high *cis*-PB in solid-core golf balls continues to grow and is very active. For this application, the polymer is compounded with zinc acrylate and the mixture is cured with peroxide.[74] The ionically cross-linked compound produced has outstanding resilience. In the last few years, the golf ball market has been shifting away from the traditional wound ball to these new solid-core balls, which use PB.

The more traditional application of the *cis*-PB continues in silica containing tire–tread applications with the majority of the polymer produced being used by the tire industry.[75–77] Cured PB has excellent low-temperature properties, high resiliency, and good abrasion resistance because of its low T_g. However, this same fundamental property also leads to very poor wet skid resistance. For this reason, PB is blended with other polymers such as natural rubber and SBR for use in tread compounds. In general, PB is a poorer-processing polymer than SBR, but this is not a problem when blended with other polymers. The very high *cis* polymers have the potential for strain-induced crystallization that can lead to improved green strength and increased cut growth resistance in the cured product. High *cis* PB is reported to have a melting point of 6°C.[78]

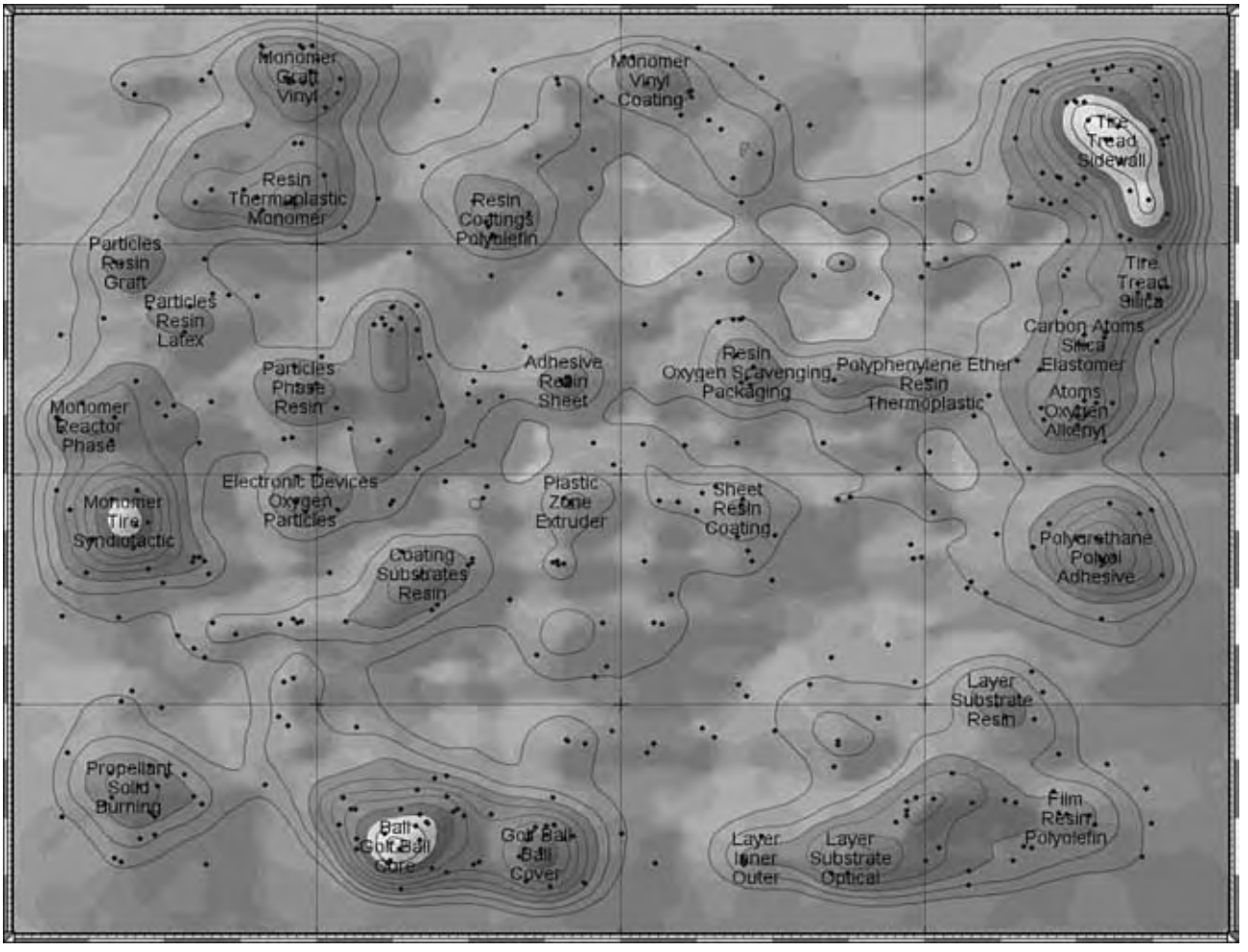


Fig. 3 Mapping of 4297 nonduplicated patent references since 1981 for the areas of use of BR without concern of SBR and block copolymer activity. *(View this art in color at www.dekker.com.)*

Fig. 6 shows the map from 109 references that cover grafting onto 1,4-PBs. The major use is as an impact modifier in plastics, in particular, high-impact polystyrene (HIPS) and acrylonitrile–BD–styrene resin (ABS).[79–83] In the HIPS application the rubber is dissolved in the styrene monomer, which is then polymerized via a free-radical mechanism. A complex series of phase changes occurs, resulting in small rubber particles containing even smaller polystyrene particles being incorporated into a polystyrene matrix. Because of the unique morphology that is formed, low levels of rubber (typically around 7%) provide rubbery particles having a volume fraction of 30–40%. This morphology leads to a high impact at very low rubber levels, providing good stiffness and hardness.[84] Major use of PB can be seen in the radical grafting to make HIPS and in the formation of ABS resins and adhesives.[83–86] Grafts of unsaturated carboxylic acids are used for high-modulus elastomers, reduction in air permeability, golf balls, and run flat tires.[87–92]

CHEMICAL ''PROMOTERS''

The promoter effect is the ability of a chemical additive to attach itself both to the polymer backbone and to the surface of the carbon black (CB) filler without causing polymer–polymer cross-linking. The best way to determine if a promoter has done its job is to look at the low-strain region of dynamic mechanical strain sweeps for decreased ''Payne Effect'' (Fig. 7). The degree of network structure can be measured from a plot of the elastic storage modulus (G') vs. strain where strain is varied from about 0.01% to 15% at 65°C. The dramatic drop in G' with strain as a result of the destruction of the CB network is known as the Payne Effect.[93] Reinforcing CBs at high loadings (>20 phr) set up a three-dimensional, chain-like network within the rubber matrix that increases the stiffness of the black-loaded compound and vulcanizes at strains <5%. This CB ''network structure'' is a major source of hysteresis in tire treads. The network is much more pronounced in the reactive super abrasion furnace

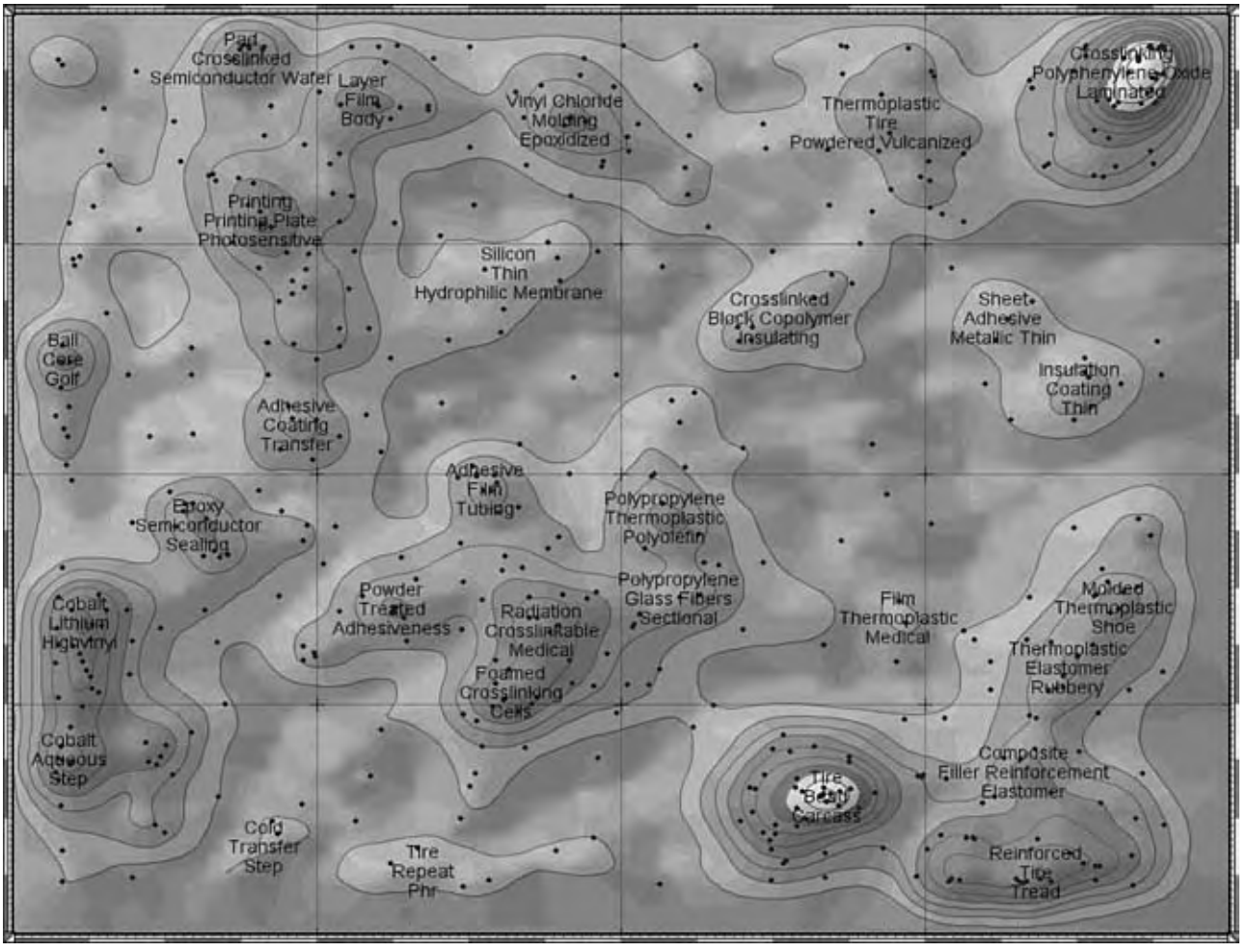


Fig. 4 A total of 926 citations for 1,2-(vinyl) BR are mapped. *(View this art in color at www.dekker.com.)*

(SAF) and high abrasion furnace (HAF) blacks and almost nonexistent with medium thermal (MT) black.

The hysteresis as measured by tan δ at about 65°C is used as a predictor of rolling resistance. As tan δ is the loss modulus (G'') divided by the elastic G', any additive that increases cross-link density would also reduce tan δ. This is why it is important to note the change in cross-link density, as predicted by G' or rheometer maximum torque when trying to determine whether a chemical additive really has a promoter effect or is just another cross-linking agent. Simply increasing cross-link density, without decreased Payne Effect, will always result in a significant drop in tear strength and may be an undesired increase in hardness. Compounds that have been successfully used include Nitrol, *N*-methyl-*p*-nitrosoaniline, benzofuroxan, and the heterocyclic di-*N*-oxides.[94–97] Despite the sharp decrease in tan δ and rolling resistance, these additives were only used briefly because their reactivity required special handling in production.

From a review of the promoter literature over the last 20 yr, it is clear that nitrogen compounds play a key role in achieving interaction with CB. The most active compounds seemed to involve some kind of nitrogen oxide chemistry that provide for the addition of the unsaturated rubber and a strong affinity for CB through a nitroxyl radical.[96,98]

Functional Solution BR

The driving force toward functional diene rubbers is to improve hysteresis properties for passenger tire treads. Historically, a classic trade-off of having to sacrifice one property to reduce another property was encountered. For instance, the reduction in hysteresis by lowering the T_g of the polymer resulted in a decrease in wet traction. As a result, another hysteresis-lowering mechanism was sought. The first promising approach was the attempt to chemically bond the elastomer to the filler, thereby reducing the agglomeration of the filler, which is a major source of hysteresis in highly filled rubber stocks. Initial efforts were directed toward the addition of reactive compounds during mixing that would show a promoter effect.

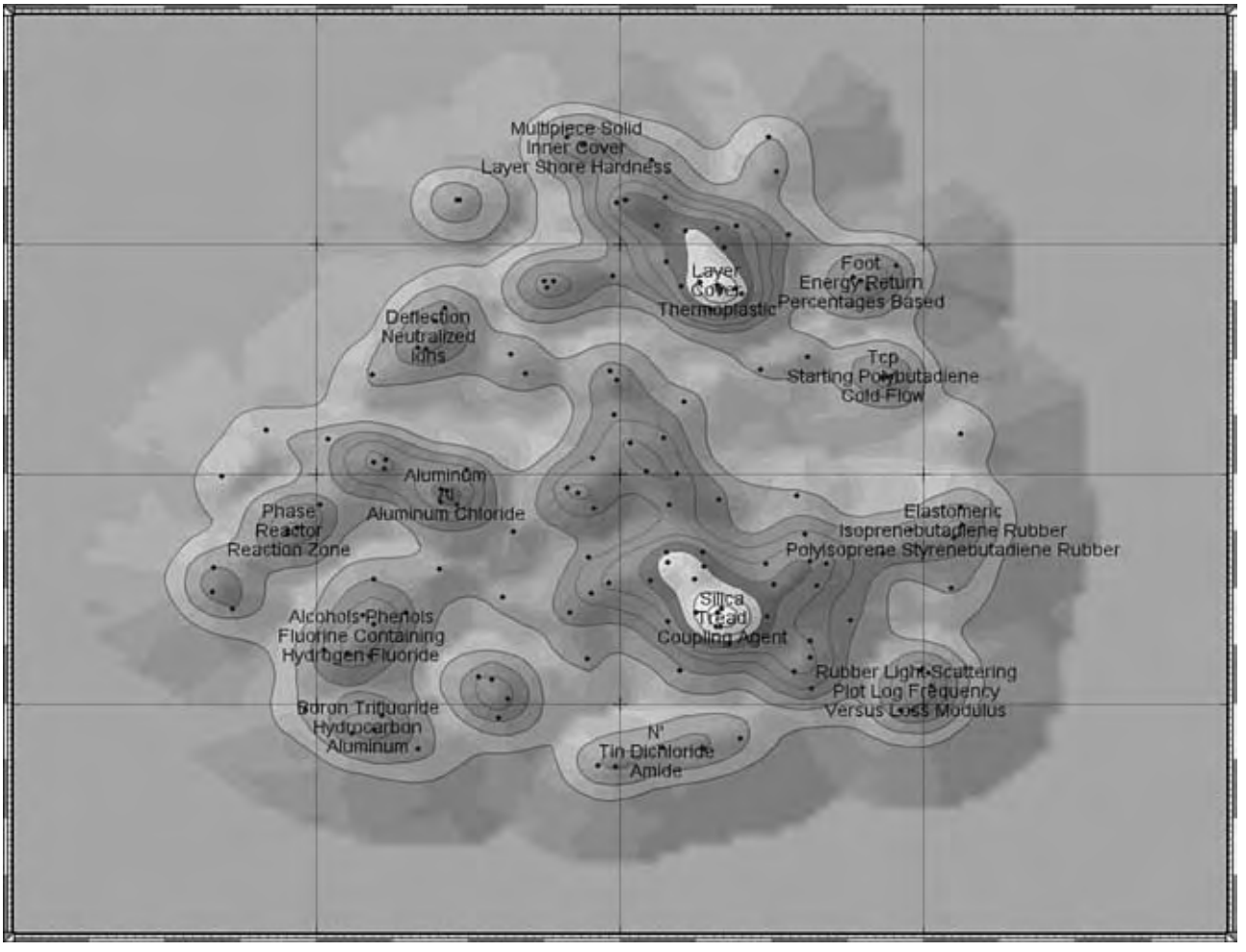


Fig. 5 Mapping of 148 citations for high *cis*-1,4-BR. *(View this art in color at www.dekker.com.)*

Anionic PB has a much narrower molecular weight distribution than an emulsion PB and gives lower hysteresis. However, the big advantage of anionic polymerization is the relatively stable growing chain ends, which can be chemically modified to improve interaction with CB and silica in tire compounds.[99] This modification can lead to a dramatic reduction in rolling resistance, which is critical for automotive manufacturers who must meet government-mandated fuel economy targets. The most active functional end groups contain either organotin or certain amines. Termination with tin tetrachloride is the easiest and the most popular method, which generates a four-armed star polymer. The polymer–tin bonds break down during mixing of the compound and both lower the compound viscosity and create active sites for reaction with CB surfaces. The creation of this ''carbon-bound rubber'' (Fig. 8) effectively prevents the formation of CB agglomerates, on a microscale, which is the source of a hysteretic three-dimensional network. The breaking of this network (Payne Effect) during the deformation of a tire tread is a major source of rolling resistance. Polymers with amine end-groups also show good activity with CB. Termination with silane esters is usually used to obtain interaction with silica fillers and provide a decrease in hysteresis. Recently, there have been efforts to make functional anionic initiators, some of which could be used to make low-hysteresis rubber.[100] Comparing an emulsion polymer to its solution counterpart and a chemically modified version of the solution polymer showed a 23% hysteresis reduction in going from emulsion to solution and an additional 15% reduction for the chemically modified polymer (representing an overall reduction of 38% of hysteresis).[101] To date, this type of chemical modification is best accomplished using anionic techniques.

UTILIZATION

The 101 citations covering functionalized PB are mapped in Fig. 9. These include epoxy, imide, and

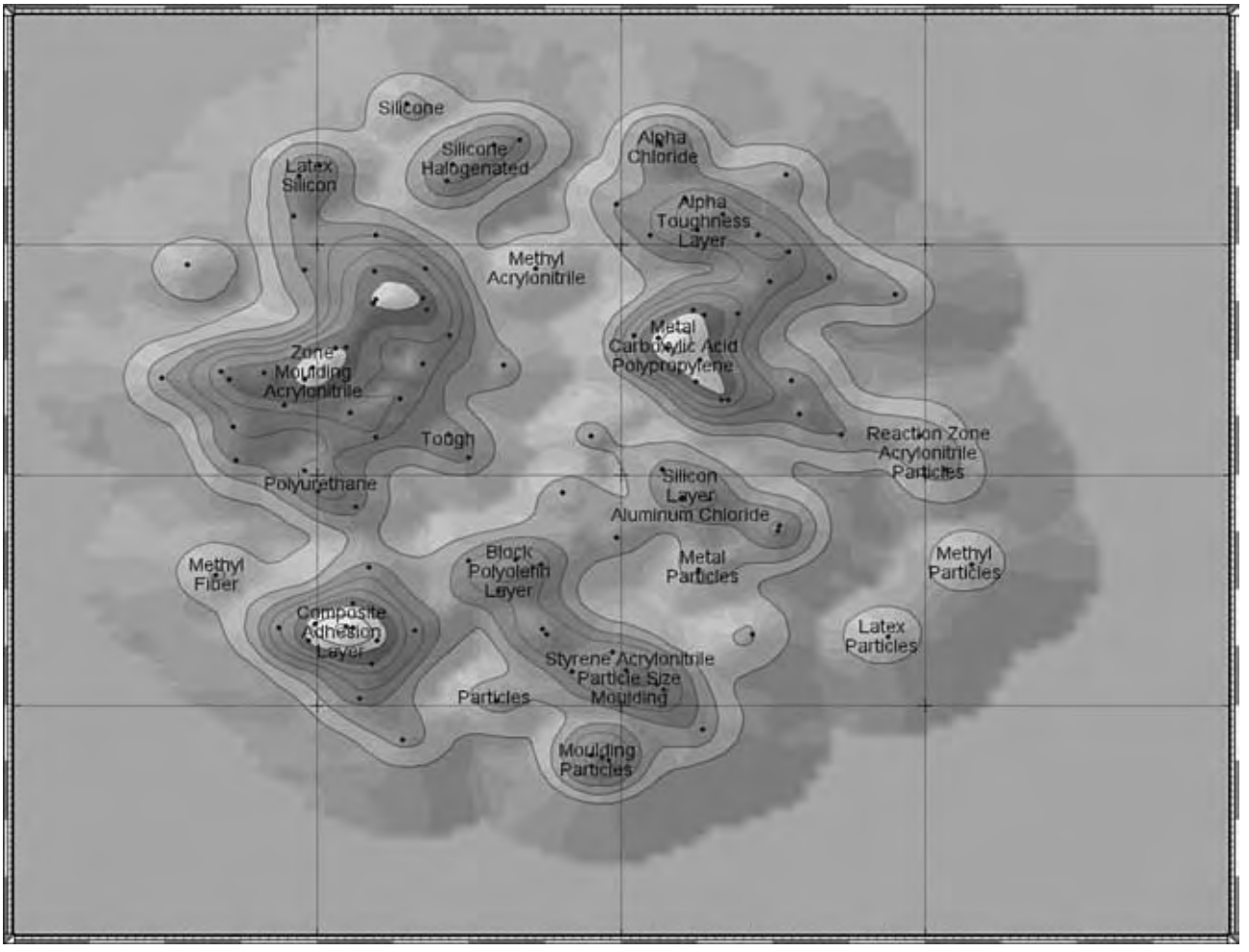


Fig. 6 Mapping of 109 references for grafted 1,4-BR. *(View this art in color at www.dekker.com.)*

modified HTPB to give toughened resins and the previously mentioned oxygen acceptors for bottles and barrier resins.[102,103] A number of citations were excluded by the searching protocol that eliminated any mention of SBR. As the functionalization chemistry is the same for SBR as for PB these citations must be included. Of particular interest is the use of functional lithium initiators containing tin, cyclic amines, *tert*-amino phenyl, or dithiane.[104–107] Termination of the lithium polymer with oxazoline, cyclic amines, and sulfur compounds, or by using high-T_g vinyl polymers also reduced hysteresis.[108–115]

Backbone functionalization with epoxide groups gave the modified PB that was discussed above. Additionally, the epoxidized PB has been used in sulfur- and amine-cured rubbers to give oil resistance.[116,117] The addition of hydrogen and chlorine or combination of alcohol and chlorine has been used as a roofing membrane.[118–120] Hydrogenated medium vinyl PB has also been reacted with maleic anhydride to give a sizing coating for glass fibers, and elastomeric impact modifier for molding resins.[121–124]

Tin-Coupled Solution PB

After the discovery that a simple tin-coupled PB showed a pronounced promoter effect, the work shifted from developing chemical additives to end-functionalizing anionic polymers. A group at Philips Petroleum Co., back in 1970, was producing tin-coupled PB using anionic techniques because the polymer–tin bonds would break down easily during mixing, yielding a better processing PB.[125] The living anionic PB was simply terminated with tin tetrachloride producing lithium chloride and a four-arm star polymer with tin at the center (Fig. 8). They noted that the improvement in heat buildup long before rolling resistance was an issue, but never related it to Payne Effect or CB–polymer interaction. They were simply trying to make an easy-processing PB. It was not until 1985 that a group at Bridgestone reported that tin-coupled polymers dramatically reduced the Payne Effect and increased carbon-bound rubber.[126] They concluded that reactive polymer chain ends, from the tin-terminated polymer, attached to the surface of CB. No mechanism was proposed for the attachment.

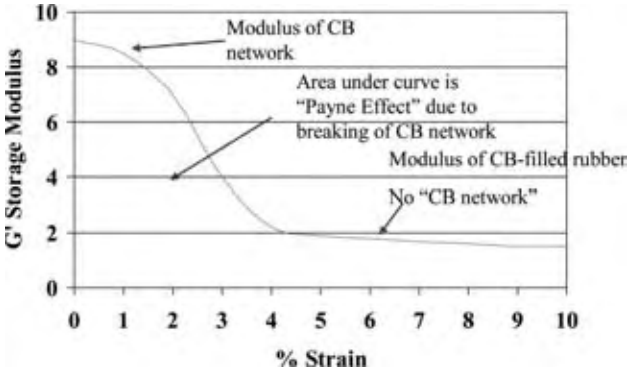


Fig. 7 The Payne Effect measured on a rheometrics dynamic analyzer (RDA II), using a 10 Hz strain sweep from 0.05% to 10% strain, 65°C, on a typical tire tread containing 50 phr HAF black. *(View this art in color at www.dekker.com.)*

Later, some Japanese synthetic rubber scientists believed the attachment involved quinone groups that are on the surface of highly reinforcing blacks blocks. This was based on the work done with allyltin compounds, and was later shown to involve a concerted reaction of the allyltin end with an *ortho*-quinone structure that is present on the reactive CB.[127,128] Tin-coupled diene elastomers are produced commercially by a number of polymer companies, including Firestone Polymers, and are still widely used today for low-rolling resistance treads, usually in a blend with natural rubber for improved tear strength. The structure for tin-coupled PB is shown in Fig. 8, along with a proposed reaction scheme for the polymer with the surface of a highly reinforcing CB.

Other Functional End-Groups for Improved Hysteresis

Since the discovery of the use of tin-coupled diene polymers for low-rolling resistance in about 1985, there is a significant effort to discover other end-grouping agents for anionic polymerization of PB and SBR, which would attach to CB. Although a myriad of terminating agents for anionic polymerization are possible only a few produce a functional group on the chain end that would reduce hysteresis in cured filled rubbers. The reaction of lithium PB with *N*-methyl pyrrolidinone, Michler's ketone (4,4′-*bis*-(diethylamino) benzophenone), or the Schiff Base from 4-dimethylamino benzaldehyde and 4-butyl aniline has given outstanding promoter effects in CB tread stocks.[129] The reduction of the 65°C tan δ by at least 20% was obtained with these compounds, indicating that significant reductions in rolling resistance would be observed in actual treads. Note that just as with the additives or chemical promoters, the most effective end groups contained either nitrogen or tin. It is shown repeatedly in studies on the promoter effect that CB seems to have a particular affinity toward compounds of tin or nitrogen (and their oxides).

Another important advantage of the above terminators is that the reaction with the polymeric live end proceeds cleanly with no coupling or side reactions producing a highly functional polymer. Many other potential terminators, like carbon dioxide, ketones, and aldehydes produce undesirable by-products and coupling of the polymer chains. To be cost-effective, the potential terminator must also survive the high temperature of the polymer cement at the peak of the

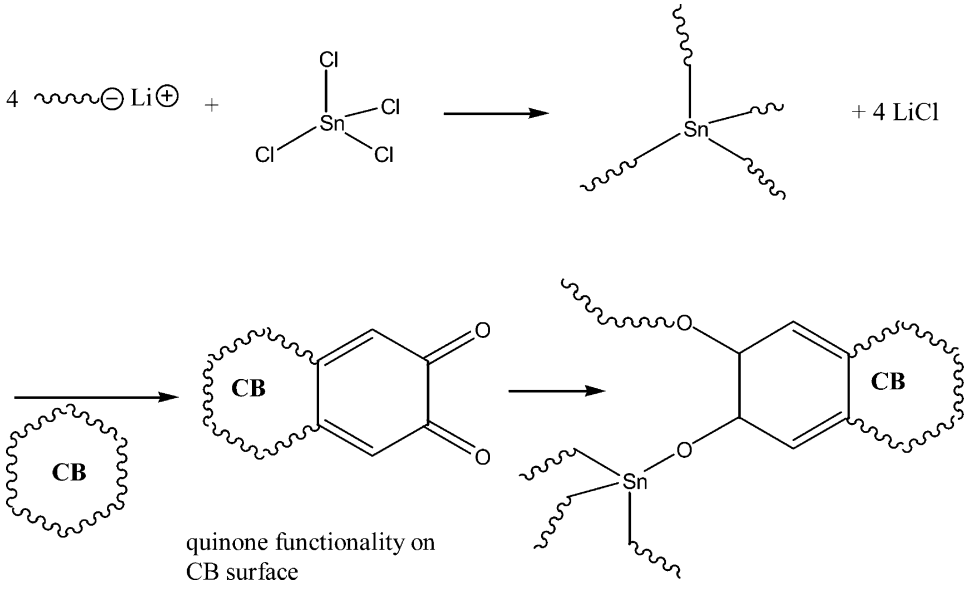


Fig. 8 Coupling of a living anionic polymer with tin tetrachloride followed by reaction with quinone functionality on CB.

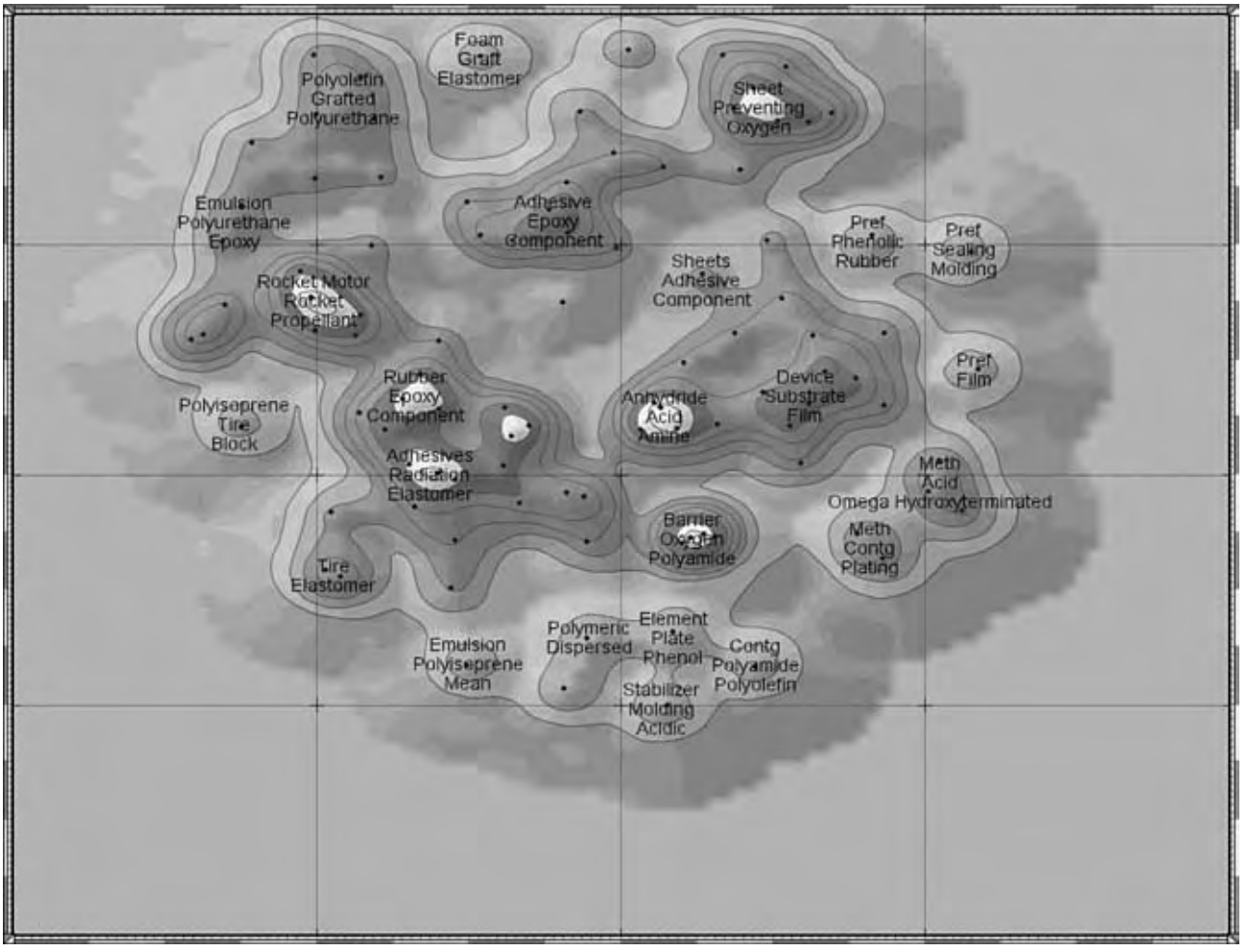


Fig. 9 Mapping of 101 patent citations for functionalized PB. *(View this art in color at www.dekker.com.)*

reaction (>100°C) and the steam desolventization process in the plant. If the polymer cement has to be cooled to low temperature prior to termination, or if the terminator is not hydrolytically stable, then the process will not be economically feasible. Tributyltin chloride has been shown to be another extremely effective promoter when used as an anionic terminator.[127] Usually, partial termination with tin tetrachloride, or other coupling agents is used to prevent cold flow problems when the above terminators are used.

Functional Anionic Initiators for Improved CB Interaction

Besides termination reactions to generate a promoter functionality on the tail of the polymer there have been a few initiators, which that have been used to attach a promoter group to the head of the polymer chain. Once again, the two most active groups were found to be certain amines and tin compounds. An even greater reduction of $\tan\delta$ and Payne Effect can be obtained with functional initiation probably owing to the fact that each polymer chain must have all of the head groups attached to a promoter. These initiators present the opportunity to also functionalize both chain ends by either coupling with tin tetrachloride or using some other active promoter-type terminator. The resulting difunctional polymer usually presents some processing challenges because of very high compound viscosities that can be controlled by slowing down the reaction of a portion of these functionalities during mixing. The amine head group could act like a macroaccelerator during sulfur curing, which results in reduced scorch time and higher cross-link density. This effect is not seen with the organotin initiator even though a pronounced decrease in Payne Effect is noted.

Tin–lithium initiator, also discovered by Bridgestone/Firestone, was found to be one of the most effective promoters for CB interaction.[104] The initiator was made from the reaction of TBT chloride with lithium metal producing tributyltinlithium. The polymerizations of SBR or PB proceeded identically to that with

butyllithium and the resultant polymer contained TBT groups on the head of every chain. When terminated with TBT chloride, the polymer had highly active TBT groups on both chain ends that produced very low tan δ and Payne Effect in black-filled stocks. The resulting low-rolling resistance and increased wear resistance were confirmed in actual tire tests.[130]

One interesting type of end-group attachment involves the addition of a free-radical generating chemical structure, which can be attached to the living anionic polymer through an epoxy group on one end of the functional chemical. This end-group then allows the functional PB to be added to a free-radical polymerization process to give HIPS or ABS with a modified structure. Under normal conditions of HIPS or ABS production, nonfunctional PB is dissolved in styrene monomer and the solution is then polymerized through a free-radical mechanism (with the additional of acrylonitrile, in the case of ABS). With a free-radical functionalized PB, block polymers are formed during the HIPS or ABS polymerization process. The formation of block polymers can lead to enhanced properties in the modified plastic compound. In some cases, the plastic compound can become clear. In other cases, the impact/gloss relationship can be improved.[131]

Promoters for Silica Stocks

In the 1990s a continuing push for lower rolling resistance led to using more silica to replace CB in passenger treads. Tread stocks for passenger tires can now contain from about 20% to almost 100% of the total filler as silica in a blend with CB to achieve low-rolling resistance. Obviously, the functional polymers developed for CB interaction would not necessarily be good with silica. For that reason, polymers were developed with organosilane end groups.[132]

REFERENCES

1. Rudin, A. *The Elements of Polymer Science and Engineering*; Academic Press: New York, 1982; 402.
2. Krause, S. Polymer compatibility. J. Macromol. Sci. C7 **1972**, *2*, 251.
3. Fox, T.G.; Grateh, S.; Loshaek, S. *Rheology—Theory and Applications*; Academic Press: New York, 1956; Vol. 1.
4. Billmeyer, F.W., Jr. *Textbook of Polymer Science*; John Wiley & Sons: New York, 1984.
5. Flory, P.J. *Principles of Polymer Chemistry*; Cornell University Press: Ithaca, NY, 1953; Chapter 7.
6. Rodriguez, F. *Principles of Polymer Systems*; McGraw-Hill: New York, 1982; 138.
7. Rudin, A. *The Elements of Polymer Science and Engineering*; Academic Press: New York, 1982; 126.
8. Damen, R.; Nieuwenhuizen, P.J.; Haasnoot, J.G.; Reedijk, J. Homogenous zinc (II) catalyst in accelerated vulcanization: V. The prevailing mechanism of crosslink formation in mercaptobenzothiazole systems. Rubber Chem. Technol. **2003**, *76*, 82.
9. Chapman, A.V.; Tinker, A.J. The effect of low molecular weight polybutadiene as processing aid on properties of silica-filled rubber compounds. Kautschuk Gummi K **2003**, *56*, 533.
10. Veith, A.G. A review of important factors affecting treadwear. Rubber Chem. Technol. **1992**, *65*, 601.
11. Halasa, A.F.; Massie Kirk-Othmer, J.M. Polybutadiene. In *Encyclopedia of Chemical Processing*; 4th Ed.; Wiley & Sons, Inc.: New York, NY, 1993; Vol. 8, 1031 pp.
12. Ponder, T. US butadiene co-production. Hydrocarbon Process. **1976**, *55* (10), 119.
13. Freidman, L.; Womeldorph, D.E.; Stevenson, D.H. Houdry dehydrogenation for olefin production. Proc. Am. Pet. Inst. **1958**, *38* (III), 203.
14. Welch, M. Butadiene via oxidative dehydrogenation. Hydrocarbon Process. **1978**, *57* (11), 131.
15. Adams, H.; Farhat, K.; Johnson, B. Gel permeation chromatography of polybutadiene. Ind. Eng. Chem. Prod. Dev. **1996**, *5* (2), 127.
16. Halasa, A.; Schulz, D.; Tate, D.; Mochel, V. *Advances in Organometallic Chemistry*; Stone, F., West, R., Eds.; Academic Press: New York, 1980; Vol. 18.
17. Dunbrook, R.F. Contribution of organic chemistry to the war effort, synthetic rubber. India Rubber World **1947**, *117*, 203.
18. Gardon, J.L. Mechanism of emulsion polymerization. Rubber Chem. Technol. **1970**, *43*, 74.
19. Stavely, F.W. Coral rubber a *cis*-polybutadiene. Ind. Eng. Chem. **1956**, *48*, 778.
20. Carlson, C.; Horne, S. Conjugated Polyolefin-Hydrocarbon Polymers and Co-polymers. U.S. Patent 3,728,325, Apr 17, 1973; GB 827,365, Feb 3, 1960; Goodrich/Gulf Chem.
21. Glasse, W.H. Spontaneous termination in living polymers. Prog. Polym. Sci. **1983**, *9*, 133.
22. Ogle, C.A.; Strickler, F.H.; Gordon, B., III. Reaction of poly(styryllithium) with tetrahydrofuran. Macromolecules **1993**, *26*, 5803.
23. Tate, D.; Bethea, T. Butadiene polymers. In *Encyclopedia of Polymer Science and Engineering*; Kroschwitz, J.I., Ed.; John Wiley & Sons: New York, 1985; Vol. 2, 537.

24. Thiele, S.K.H.; Wilson, D.R. Alternate transition metal complex based diene polymerization. J. Mater. Sci. Polym. Rev. **2004**, *C43*, 581.
25. Kaita, S.; Takeguchi, Y.; Hou, Z.M.; Nishiura, M.; Doi, Y.; Wakatsuki, Y. Pronounced enhancement brought about by substituents on the cyclopentadienyl ligand. Macromolecules **2003**, *36*, 7923.
26. Kaita, S.; Hou, Z.M.; Nishiura, M.; Doi, Y.; Kurazumi, J.; Horiuchi, A.C.; Wakatsuki, Y. Ultimately specific 1,4-*cis* polymerization of 1,3-butadiene with a novel gadolinium catalyst. Macromol. Rapid Commun. **2003**, *24*, 180.
27. Lauretti, F.; Gargani, L. Fatigue resistance of polybutadiene. Paper at the 27th Annual Meeting of the International Institute of Synthetic Rubber Producers, Jun 1987.
28. Mello, I.L.; Coutinho, F.M.B.; Soares, B.G.; Nunes, D.S.S.; Costa, M.A.S.; Maria, L.C.D. *cis*-1,4-Polymerization of butadiene with Ziegler–Natta catalyst systems based on neodymium. Quimia Nova **2004**, *27*, 277.
29. Nasirov, F.A. Bi-functional nickel- or cobalt-containing catalyst-stabilizers for polybutadiene production and stabilization. Iranian Polym. J. **2003**, *12*, 217.
30. Wu, X.G.; Wu, Z.H.; Liu, H.M.; Mo, K.; Liu, H.Q. Raising reactivity of Nd coordinate polymerization catalyst of non-homogenous phase by turbulent flow. J. Polym. Mater. **2002**, *19*, 321.
31. Sadeghi, G.M.M.; Morshedian, J.; Barikani, M. The effect of initiator-to-monomer ratio on the properties of polybutadiene-ol synthesized by free radical solution polymerization of 1,3-butadiene. Polym. Int. **2003**, *52*, 1083.
32. Morton, M.; Fetters, L.J.; Inomata, J.; Rubio, D.C.; Young, R.N. Synthesis and properties of uniform polyisoprene networks. I. Synthesis and characterization of α,ω-dihydroxy polyisoprene. Rubber Chem. Technol. **1976**, *49*, 303.
33. Tamura, H.; Nakayama, A. Functional telechelic polymer synthesis via ADMET polymerization. J. Macromol. Sci. Pure Appl. Chem. **2002**, *39*, 745.
34. Oberth, A.E. Energetic Plasticizers for Polybutadiene-Type Solid Propellant Binders. U.S. Patent 5,578,789, Nov 26, 1996; Aerojet General.
35. Sayles, D.C. A Processing Method for Increasing Propellant Burning Rate. U.S. Patent 4,655,860Apr 7, 1987; U.S. Army.
36. Stephens, W.D.; Nieder, E.G. Propellant Plasticizer. U.S. Patent 4,482,406, Nov 13, 1984; U.S. Air Force.
37. Teasdale, D.; Milanovic, V.; Chang, P.; Pister, K.S.J. Microrockets for smart dust. Smart Mater. Struct. **2001**, *10*, 1145.
38. Millett, J.C.F.; Bourne, N.K.; Akhaven, J. The response of hydroxy-terminated polybutadiene to one-dimensional shock loading. J. Appl. Polym. Phys. **2004**, *95*, 4722.
39. Siviour, C.R.; Gifford, M.J.; Walley, S.M.; Proud, W.G.; Field, J.E. Particle size effects on the mechanical properties of a polymer bounded explosive. J. Mater. Sci. **2004**, *39*, 1255.
40. Arikan, A.; Kaynak, C.; Tincer, T. Influence of liquid elastomeric additive on the behavior of short glass fiber reinforced epoxy. Polym. Compos. **2002**, *23*, 790.
41. Barcia, F.L.; Amaral, T.P.; Soares, B.G. Synthesis and properties of epoxy resin modified by epoxy-terminated liquid polybutadiene. Polymer **2003**, *44*, 5811.
42. Talon, O.; Couvercelle, J.P.; Bunel, C. Impact modification of polyamide-12 through blending with liquid polybutadiene. E-Polymers **2003**, *1*, 1–19.
43. Petrov, P.; Jankova, K.; Mateva, R. Polyamide-6-block-polybutadiene copolymers: synthesis and properties. J. Appl. Polym. Sci. **2003**, *89*, 711.
44. Wang, Z.F.; Wang, B.; Yang, Y.R.; Hu, C.P. Investigation of gas permeation and free volume hole properties of polyurethane membranes by positrons. In Positron Annihilation, ICPA-13, Proceedings, 2004; Vol. 445, 352.
45. Barcia, F.L.; Soares, B.G.; Sampaio, E. Adhesive properties of epoxy resin modified by end-functionalized liquid polybutadiene. J. Appl. Polym. Sci. **2004**, *93*, 2370.
46. Akkapeddi, M.K.; Kraft, T.J.; Socci, E.P. Oxygen Scavenging High Barrier Polyamide Compositions for Packaging Applications. U.S. Patent 6,423,776, Jul 23, 2002; Honeywell International.
47. Tsai, M.L.; Akkapeddi, M.K. Oxygen Scavenging Compositions Containing Ethylene Vinyl Alcohol Copolymers. U.S. Patent 6,793,994, Sep 21, 2004; Honeywell International.
48. Akkapeddi, M.K.; Socci, E.P.; Kraft, T.J.; Worley, D.C. Oxygen Scavenging Polyamide Compositions Suitable for PET Bottle Applications. U.S. Patent 6,610,234, Aug 26, 2003; Honeywell International.
49. Akifumi, K.; Yoshinori, K. Fouling Preventing Sheet. JP 2004-202,918, Jul 22, 2004; Toyo Ink.
50. Luo, S. Molybdenum-Based Catalyst Composition and Process for Controlling the Characteristics of Conjugated Diene Polymers. U.S. Patent 6,348,550, Feb 19, 2002; Bridgestone.
51. Luo, S.; Zak, D.E. Iron-Based Catalyst Composition for the Manufacture of Syndiotatic 1,2-Polybutadiene. U.S. 6,610,804, Aug 26, 2003; Bridgestone.

52. Luo, S. Catalyst Composition and Process for Controlling the Characteristics of Conjugated Diene Polymers. U.S. 6,288,183, Sep 11, 2001; Bridgestone.
53. Chen, Y.; Yang, D.C.; Hu, Y.M.; Zhang, X.Q. Effects of crystal growth conditions on morphology of crystalline syndiotatic 1,2-polybutadiene. Cryst. Growth Des. **2004**, *4*, 117.
54. Tsujimoto, N.; Hongyo, K.; Baba, Y.; Suzuki, M. Polybutadiene Composition and Process for Producing Same. U.S. Patent 5,468,822, Nov 21, 1995; UBE Ind.
55. Wong, T.H.; Cline, J.H. Syndiotatic 1,2-Polybutadiene Synthesis. U.S. Patent 5,986,026, Nov 16, 1999; Goodyear.
56. Osamu, K.; Nobuhiro, T. Production of Polybutadiene. JP 10-168,117, Jun 23, 1998; UBE Ind.
57. Hideaki, Y. Pneumatic Tire. JP 7-25,212, Jan 27, 1995; Bridgestone.
58. Hideaki, Y.; Shunji, A. Pneumatic Tire. JP 6-87,977, Mar 29, 1994; Bridgestone.
59. Zanzig, D.J.; Sandstrom, P.H.; Crawford, M.J.; Verthe, J.J.A. Tire with Reinforced Silica Tread. U.S. Patent 5,614,580, Mar 25, 1997; Goodyear.
60. Hubbell, J.K.; Crawford, M.J. Tire with Reinforced Silica Tread. European Patent 0,747,426, Jan 5, 2000; Goodyear.
61. Junji, K.; Koji, O.; Minoru, F.; Katsuaki, M. Radiation Crosslinkable Polymer Composition. JP 2002-348,479, Dec 4, 2002; JSR Corp.
62. Osami, S. Container for Cryopreservation. JP 2001-29,432, Feb 6, 2001; Terumo Corp.
63. Noboru, O.; Isamu, S.; Yoshito, Y. Novel Butadiene Polymer Composition. JP 59-124,943, Jul 19, 1984; Japan Synthetic Rubber.
64. Haruo, U.; Koichi, N.; Hidetomo, A.; Kazuya, J. Production of Reinforced Diene Rubber. JP 58-187,408, Nov 1, 1983; UBE Ind.
65. Takeuchi, Y.; Yoshimura, Y.; Ohshima, N.; Sakakibara, M. High Vinyl Polybutadiene or Styrene–Butadiene Copolymer. U.S. Patent 4,397,994, Aug 9, 1983; Japan Synthetic Rubber.
66. Yozo, K.; Yasuo, T.; Akira, S.; Munehiko, I. Molding Material for Synthetic Resin Molding. JP 57-177,046, Oct 30, 1982; Japan Synthetic Rubber.
67. Kazumichi, S.; Hiroshi, N.; Kenji, K. Flexible Vinyl Chloride Resin Film. JP 4-258,641, Sep 14, 1992; Chisso Corp.
68. Knauf, T.F.; Osman, A. Process for *cis*-Polybutadiene Production with Reduced Gel Formation. U.S. Patent 5,397,851, , Mar 14, 1995; Polysar Rubber.
69. Knauf, T.F.; Osman, A. Process for Polybutadiene Production Using Catalyst with High Activity. U.S. Patent 5,428,119Jun 27, 1995; Polysar Rubber.
70. Kerns, M.L.; Bowen, D.E.; Rodewald, S. Transition Metal Catalysts for Diene Polymerization. JP 2003-119,662, Jun 26, 2003; Goodyear.
71. Kato, A. Thread Wound Golf Ball. U.S. Patent 6,475,103, Nov 5, 2002; Sumitomo Rubber.
72. Bradley, W.R.; Simonutti, F.M. Golf Ball with a Cover which includes Polyurethane Rubber. U.S. Patent 6,152,836, Nov 28, 2000; Wilson Sporting Goods.
73. Moriyama, K.; Sano, Y.; Yoshida, K. Three Piece Solid Golf Ball. U.S. Patent 6,120,391, Sep 19, 2000; Sumitomo Rubber.
74. Maruko, T. Solid Golf Ball. U.S. Patent 6,071,201, Jun 6, 2000; Bridgestone Sports.
75. Ferrandino, M.P.; Hong, S.W.; McKenzie, G.T. Tire Tread Composition. U.S. Patent 5,569,697, Oct 29, 1996; Uniroyal Chemical.
76. Lucas, D.; Agostini, G.; Corvasce, F.G.; Hunt, J.O. Tire Tread for Ice Traction. U.S. Patent 5,967,211, Oct 19, 1999; Goodyear.
77. Stuhldreher, T.M. Kaolin Clay in Silica Tread Compounds Technical Field. U.S. Patent 6,080,809, Jun 27, 2000; Bridgestone/Firestone.
78. DeChirico, A.; Lamzani, P.; Eaggi, E.; Bruzzone, M. High *cis*-polybutadiene by uranium catalyst. Makromol. Chem. **1974**, *175*, 2029.
79. Pham, T.T.B.; Fellows, C.M.; Gilbert, R.G. Grafting of dodecyl methacrylate onto hydroxylated polybutadiene by miniemulsion polymerization. J. Polym. Sci. A. **2004**, *42*, 3404.
80. Cangialosi, D.; Lindsay, C.; McGrail, P.T.; Spadaro, G. Study of methyl methacrylate polymerization in the presence of rubbers. Eur. Polym. J. **2001**, *37*, 535.
81. Joseph, S.; Oommen, Z.; Thomas, S. Melt elasticity and extrudate characteristics of polystyrene/polybutadiene blends. Mater. Lett. **2002**, *53*, 268.
82. de Leon, R.D.; Morales, G.; Acuna, P.; Flores, R.F.; Robles, A.M. Effect of morphological parameters on toughness of rubber-modified polystyrene. Rev. Mex. Fis. **2004**, *50*, 85.
83. Baumgartner, E.; Hofmann, J.; Jung, R.H.; Moors, R. ABS Molding Materials having a Bimodal Rubber Particle Size Distribution. U.S. Patent 5,434,218, Jul 18, 1995; BASF.
84. Buckenell, C. *Toughened Plastics*; Applied Science: London, 1977.
85. Choe, E.W.; Forbes, C.E.; Filbey, J.A.; Sherriff, S.F. Rubber–Polyester Composites Including Polystyrene–Polyester Copolymers. U.S. Patent 5,597,651, Jan 28, 1997; Hoechst Celanese.

86. Guo, L.; Mounier, J.Y. Composite Thermoplastic–Elastomer Product. U.S. 6,652,937, Nov 25, 2003; Hutchinson.
87. Hergenrother, W.L.; Ravagnani, F.J.; Doshak, J.M. Poly (Metal Carboxylate) Grafted Rubbers with High Modulus. U.S. Patent 5,962,593, Oct 5, 1999; Bridgestone.
88. Freeman, R.M.; Hergenrother, W.L.; Ravagnani, F.J. High Modulus Low Hysteresis Rubber Compound for Pneumatic Tires. U.S. Patent 5,464,899, Nov 7, 1995; Bridgestone.
89. Freeman, R.M.; Hergenrother, W.L.; Ravagnani, F.J. High Modulus Low Hysteresis Rubber Compound for Pneumatic Tires. U.S. Patent 5,494,958, Mar 26, 1996; Bridgestone.
90. Morehart, C.L.; Ravagnani, F.J. Pneumatic Tire with Air Retention Carcass. U.S. Patent 5,484,005, Jan 16, 1996; Bridgestone.
91. Akihiko, H.; Kohei, T.; Toshiaki, T.; Yoshinori, S. Golf Ball. JP 8-103,515, Apr 23, 1996; Sumitomo Rubber.
92. Freeman, R.M.; Hergenrother, W.L.; Ravagnani, F.J. High Modulus Low Hysteresis Rubber Compound for Pneumatic Tires. U.S. Patent 5,494,091, Feb 27, 1996; Bridgestone.
93. Payne, A.R.; Watson, W.F. CB structure in rubber. Rubber Chem. Technol. **1963**, *36*, 147.
94. Hamad, P.; Walker, L. Nitrol, a chemical promoter. Rubber World **1971**, *164* (4), 51.
95. Morita, E. Reaction of *N*,4-dinitroso-*N*-methylaniline and polymers. Rubber Chem. Technol. **1976**, *49*, 1019.
96. Graves, D. Benzofuroxans as rubber additives. Rubber Chem. Technol. **1993**, *66*, 61.
97. Graves, D.F.; Engels, H.W. Rubber composition modified with heterocyclic di-N-oxides. U.S. Patent 4,822,845, Apr 18, 1989; Firestone.
98. Sullivan, A.B. Electron spin resonance studies of a stable aryl nitroso-olefin adduct, free radical. J. Org. Chem. **1966**, *31*, 2811.
99. Day, G.; Moore, D. Comparison of emulsion and solution SBR in tire performance. Paper at 26th Annual Meeting of the International Institute of Synthetic Rubber Producers, May 1985.
100. Quirk, R.P.; Jang, S.H. Recent advances in anionic synthesis of functionalized elastomers using functionalized alkyllithium initiators. Rubber Chem. Technol. **1996**, *69* (3), 444.
101. Oshima, N.; Salcacobore, M.; Tsutsumi, F. Newly developed solution SBR for low-rolling resistance tires. Paper at 27th Annual Meeting of the International Institute of Synthetic Rubber Producers, May 1986.
102. Ichirou, I. Epoxy Resin Composition Having Improved Adhesiveness and Cracking Resistance. JP 56-120,726, Sep 22, 2004; Toshiba Chem.
103. Schonfeld, R.; Schenkel, H.; Kuster, H. Impact-Resistant Epoxy Resin Composition. European Patent 1,272,587, May 19, 2004; Henkel.
104. Hergenrother, W.L.; Bethea, T.W.; Doshak, J.M. Tin Containing Elastomers and Products Having Reduced Hysteresis Properties. U.S. Patent 5,268,439, Dec 7, 1993; Bridgestone/Firestone.
105. Lawson, D.F.; Hergenrother, W.L.; Kerns, M.J. Tertiary Amines Containing Side-chain Organolithium Structures and Method for the Preparation Thereof. U.S. Patent 5,912,343, Jun 15, 1999; Bridgestone.
106. Hergenrother, W.L.; Doshak, J.M. Method of Preparing an Anionic Polymerization Initiator. U.S. Patent 5,238,893, Aug 24, 1993; Bridgestone.
107. Hogan, T.; Hergenrother, W.; Yan, Y.Y.; Lawson, D. The Use of Sulfur Containing Initiators for Anionic Polymerization of Monomers. WO 2004041870, May 21, 2004; Bridgestone.
108. Rademacher, C.M.; Hergenrother, W.L.; Graves, D.F.; Ulmer, J.D. Preparation of Low Hysteresis Rubber by Reacting a Lithium Polymer with Oxazoline Compounds. U.S. Patent 6,596,798, Jul 22, 2003; Bridgestone.
109. Lawson, D.F.; Hergenrother, W.L.; Kerns, M.L. Polymers, Elastomeric Compounds and Products Thereof, Derived from Novel Amine Compounds Containing Side-Chain Organolithium. U.S. Patent 5,786,441, Jul 28, 1998; Bridgestone.
110. Hogan, T.E.; Hergenrother, W.L. Synthesis and Use of Chain-Coupled Polymeric Sulfide Compounds in Rubber Formulations. U.S. Patent 6,806,307, Oct 19, 2004; Bridgestone.
111. Hergenrother, W.L.; Doshak, J.M. Method of Preparing Elastomers Having Reduced Hysteresis Properties with Sulfoxides. U.S. Patent 5,151,469, Sep 29, 1992; Bridgestone.
112. Hergenrother, W.L.; Rademacher, C.M.; Ulmer, J.D. Preparation of Low Hysteresis Rubber by Reacting a Lithium Polymer with a Sulfur Containing Reagent. U.S. Patent 6,579,949, Jun 17, 2003; Bridgestone.
113. Hergenrother, W.L.; Doshak, J.M. (Vinyl Sulfoxide)-Capped Elastomers and Compositions Having Reduced Hysteresis Properties. U.S. Patent 5,276,099, Jan 4, 1994; Bridgestone.
114. Hergenrother, W.L.; Morrow, M.E. Elastomers Having Reduced Hysteresis Prepared with Vinyl Imidazole. U.S. Patent 5,346,962, Sep 13, 1994; Bridgestone/Firestone.
115. Hergenrother, W.L.; Morrow, M.E. Elastomers Having Reduced Hysteresis Prepared with Vinyl Polycyclic Aromatic Hydrocarbons. U.S. Patent 5,349,024, Sep 20, 1994; Bridgestone/Firestone.

116. Rogers, M.B.; Botts, B.P.; Krishnan, R.M. Endless Two Part Rubber Track Comprised of Polyurethane Based Tread Component and Rubber Carcass Component and Vehicle Containing Such Track. U.S. Patent 6,769,746, Aug 3, 2004; Goodyear.
117. Yabe, T.; Uchiyama, K.; Aihara, N.; Yokoyama, K. Rubber Composition and Rubber Sealing Device. U.S. Patent 6,639,019, Oct 28, 2003; NSK Ltd.
118. Hergenrother, W.L.; Doshak, J.M. Process and Isomerizing Agent for Isomerizing Partially Hydrogenated Diene Polymers. U.S. Patent 5,314,967, May 24, 1994; Bridgestone/Firestone.
119. Davis, J.A.; Hergenrother, W.L.; McGillivary, D.R.; Valaitis, J.K. Substituted Polybutadiene Polymer and Roofing Membrane Formed Therefrom. U.S. Patent 5,126,384, Jun 30, 1992; Bridgestone/Firestone.
120. Hergenrother, W.L.; Davis, J.A.; McGillivary, D.R.; Valaitis, J.K. Haloalkoxylated Polybutadiene Polymer and Roofing Membrane Formed Therefrom. U.S. Patent 5,130,355, Jul 14, 1992; Bridgestone.
121. Hergenrother, W.L. Composites and Size Coated Glass Fibers Used Therein. U.S. Patent 4,537,917, Aug 27, 1985; Firestone.
122. Hergenrother, W.L. Process for Making Coated Glass Fiber Reinforced Composites. U.S. 4,524,040, Apr 2, 1985; Firestone.
123. Hergenrother, W.L.; Matlock, M.G.; Ambrose, R.J. Impact Resistant Polymeric Compositions Containing Polyamides, Maleic Anhydride Adducts of Hydrogenated Polymers, and Graft Copolymers Thereof. U.S. Patent 4,427,828, Jan 24, 1984; Firestone.
124. Hergenrother, W.L.; Matlock, M.G.; Ambrose, R.J. Impact Resistant Polymeric Compositions Containing Polyamides, Maleic Anhydride Adducts of Hydrogenated Polymers, and Graft Copolymers Thereof. U.S. Patent 4,508,874, Apr 2, 1985; Firestone.
125. Uraneck, C. Solution polymerized rubber with superior breakdown properties. J. Appl. Polym. Sci. **1970**, *14*, 1421.
126. Tsutsumi, F. Solution polymerized SBR terminated with tin compounds—tire application. Presented at International Rubber Conference, Kyoto, Japan, 1985.
127. Tsutsumi, F.; Sakibara, M.; Oshima, N. Structure and dynamic properties of solution SBR coupled with tin compounds. Rubber Chem. Technol. **1990**, *63*, 8.
128. Hergenrother, W.L.; Doshak, J.D.; Brumbaugh, D.R.; Bethea, T.W.; Oziomek, J. ^{119}Sn-NMR evidence for 2,1-initiation in the anionic polymerization of butadiene with trialklytin lithium. J. Polym. Sci., Polym. Chem. **1995**, *33*, 143.
129. Nagata, N.; Kobatake, T.; Wantanabe, H.; Uedia, A. Effect of chemical modification of solution-polymerized rubber on dynamic mechanical properties in carbon-black filled vulcanizates. Rubber Chem. Technol. **1987**, *60*, 837.
130. Bethea, T.; Hergenrother, W.L.; Clark, F.J.; Sarkar, S.B. Techniques to reduce tread hysteresis. Rubber Plast. News **1994**, Aug 29, 17–19.
131. Priddy, D.B.; Li, I.Q. In situ Block Copolymer Formation During Polymerization of a Vinyl Aromatic Monomer. U.S. Patent 5,721,320, Feb 24, 1998; Dow Chemical Co.
132. Shimizu, T. Diene Polymers and Copolymers Having an Alkoxysilane Group. U.S. Patent 5,508,333, Apr 16, 1996; Bridgestone.

Polycarbonate (PC)

Sarah E. Morgan
Jun Li
Department of Polymer Science, University of Southern Mississippi, Hattiesburg, Mississippi, U.S.A.

INTRODUCTION

Polycarbonate (PC), with its unique combination of toughness, transparency, processability, and temperature resistance, is one of the most important engineering thermoplastics in the world today. It is found in a wide range of products, from CDs and DVDs to greenhouses and optical lenses to automotive lighting and instrument panels. Polycarbonate forms a number of useful blends and composites and can be produced in a wide range of transparent and opaque colors, making it the material of choice for many products with special visual effects, like specially colored computer housings and personal communication devices. Bisphenol A polycarbonate (BPA-PC) is the most important commercial PC and is the primary subject of this review. Until recently, BPA-PC was produced via an interfacial polymerization process. Recently commercialized melt processes provide solvent- and phosgene-free production routes. This article provides a brief summary of the historical development of BPA-PC, an overview of commercial products and producers, a review of PC physical and mechanical properties, a description of current commercial production processes for PC as well as new process developments, an overview of processing and fabrication processes employed with PC, a summary of PC blends, copolymers, and applications, and a summary of PC literature.

BACKGROUND AND HISTORY

Interest in PC continues to grow, with over 44,000 current publications and patents, of which some 55% are patents.[1] Before 1960, there were fewer than 100 total publications and patents on PC. Annual new publications increased from around 200 per year in the 1960s to 2000 per year in the 1990s, with continued growth of approximately 1400 new patents published per year since 2000. Similarly, global production and consumption of PC continue to grow, with an estimated 1.5 million metric tons of global consumption in 2000[2] and a projected annual growth rate of 7% per year.[3] GE has the largest global market share of PC resin (sold under the trademark LEXAN® resin),[4,5] with Bayer (Makrolon®)[6] holding second place. Other major global producers of PC resin include Dow Chemical Company (CALIBRE®),[7] Teijin (Panlite®),[8] and Mitsubishi Engineering Plastics Corporation (Iupilon®).[9] There are frequent announcements of expansions, licenses, and plans for new production sites to serve emerging markets around the world.[10–12]

Polycarbonate based on 2,2-bis(4-hydroxyphenyl) propane (bisphenol A, BPA) is the PC of primary commercial interest (Fig. 1), but commercial production of BPA-PC did not begin until the 1960s.[13] The earliest reports of aromatic PC preparation, by Einhorn, appeared in the late 1800s and involved reaction of hydroquinone or resorcinol with phosgene in pyridine solution.[14] In 1902, Bischoff and Hedenstroem[15] reported a melt synthesis of these polymers via transesterification with diphenyl carbonate. Carothers and Van Natta[16] reported aliphatic PC synthesis in the 1930s. Aliphatic PCs display low melting points, strength, and durability and are generally of little commercial interest. The first commercial PC material, a cross-linkable allyl diglycol carbonate (CR-39®), was introduced by Pittsburgh Plate Glass Company (PPG) in the 1940s.[17]

Researchers at GE and Bayer independently developed commercially feasible synthetic processes for BPA-PC in the 1950s[18,19] and began commercial production in the early 1960s. Bayer was awarded the U.S. patent for PC produced via the interfacial polymerization process[20] and GE the U.S. patent for the melt transesterification process.[21] However, until recently, the major part of BPA-PC was produced via the interfacial process. Further information on the history of PC development can be found in previously published reviews.[22–28]

POLYCARBONATE STRUCTURE AND PROPERTIES

Aromatic PC based on BPA (Fig. 1) is the most important commercial PC. BPA-PC is amorphous and transparent, with a glass transition temperature (T_g) of approximately 150°C. The high T_g of BPA-PC is attributable to the bulky structure of the polymer,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007908

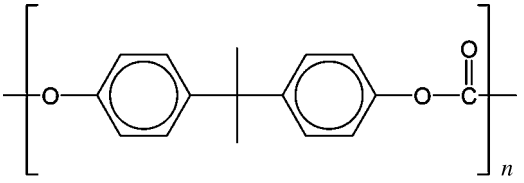


Fig. 1 Structure of BPA-PC.

which hinders segmental motion of the polymer chain. Unlike most amorphous polymers at temperatures below T_g, BPA-PC exhibits extraordinary impact strength and toughness. Its unusual low-temperature impact strength has been attributed by some authors to the presence of a low-temperature gamma transition that represents local scale motions of the polymer chain in the glassy state, which are thought to provide a mechanism for energy dissipation.[29–32] Alternative theories incorporate the importance of free volume and residual stresses on PC impact behavior (recently reviewed in Ref.[33]).[33–35] As shown in Fig. 2, a low-temperature gamma transition temperature is observed at around –100°C. Additional transitions observed are the β transition at around 60°C and the α transition (T_g) at around 150°C. The β transition temperature is dependent on the thermal and mechanical history of the test sample and is attributed to residual stresses in the polymer sample as well as to the state of aging of the sample.[36]

The mechanical, thermal, and processing parameters of PC depend on the molecular weight of the polymer. Typical commercial BPA-PCs for injection molding and extrusion applications have weight-average molecular weight (M_w) in the range of 17,000–36,000 as determined by light scattering. Molecular weight is often approximated from the intrinsic viscosity using the Mark–Houwink–Sakurada relationship:

$$[\eta] = KM_v^a$$

where $[\eta]$ is the intrinsic viscosity, M_v is the viscosity-average molecular weight, and K and a are constants for a given polymer/solvent combination at a given temperature. Molecular weight, polydispersity index, intrinsic viscosity, and melt viscosity data have been correlated for a range of PCs of differing molecular weights.[37] K and a values are reported for BPA-PC in a number of solvents.[22,38–40] In general, mechanical and thermal properties such as impact resistance and softening point increase with increasing molecular weight, as does melt viscosity (and thus resistance to flow). Mechanical properties increase rapidly with increasing molecular weight at low molecular weights, but at molecular weights higher than around 22,000 (intrinsic viscosity approximately 0.45 dl/g), only slight changes are observed. Melt viscosities, however, continue to increase, making melt processing increasingly difficult. To optimize the balance of processability with mechanical properties, the PC of lowest molecular weight that meets minimum mechanical property requirements should be chosen for a given application. Higher-viscosity resins are typically used for extrusion applications, where high melt viscosities are tolerated and high strengths are required. Very low viscosity resins are used in applications such as injection molding of CDs, where high flow is necessary but impact strength is of secondary importance. Melt flow rates, intrinsic viscosities, and molecular weights for typical commercial BPA-PCs are given in Table 1.

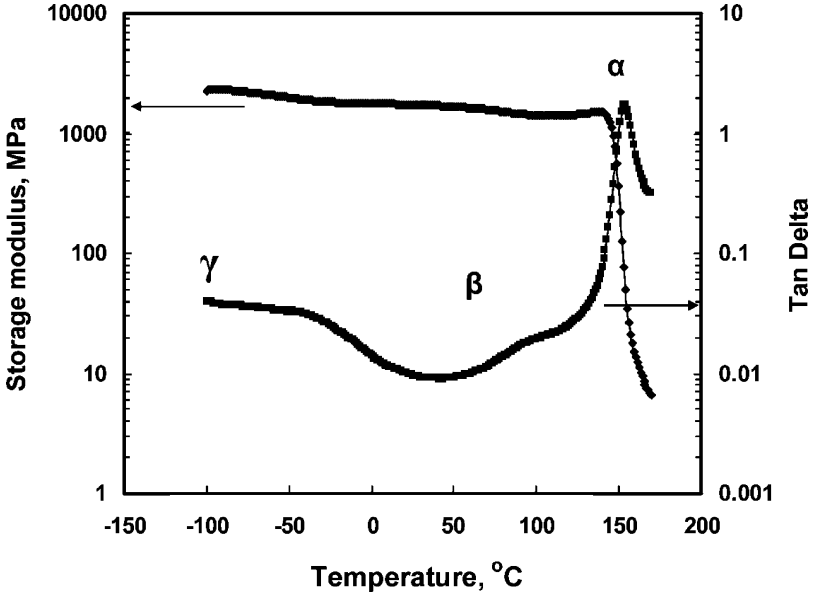


Fig. 2 Dynamic mechanical analysis data for LEXAN 101 PC resin. (Courtesy of GE.)

Table 1 Molecular weight and viscosity for typical commercial PC resins

GE LEXAN resin grade	Description	Melt flow index (g/min)[a]	Intrinsic viscosity (dl/g)[b]	M_w[c]
131	Ultrahigh viscosity	3.1	0.629	35,500
101	High viscosity	6.5	0.551	29,000
141	Medium viscosity	9.2	0.51	26,300
121	Low viscosity	16.2	0.454	21,200
HF1110	High flow	20.9	0.434	22,700
OQ1020	Optical quality	78	0.35	16,600

[a]300°C, 1.2 kgf.
[b]Intrinsic viscosity in methylene chloride at 25°C.
[c]Weight-average molecular weight by light scattering.
(From Ref.[25]. Courtesy of GE.)

Typical thermal and mechanical properties for commercial BPA-PCs of differing molecular weights are given in Table 2. Softening temperature increases with molecular weight for the high-flow (low molecular weight) polymers but is virtually constant for higher molecular weight materials. Similar trends are observed for mechanical properties such as notched izod impact and tensile strength.

Polycarbonate demonstrates excellent ductility below T_g, in contrast to most glassy polymers. The ductile/brittle transition in BPA-PC is a function of molecular weight, temperature, notch radius, sample thickness, rate of deformation, annealing, physical aging, and residual stresses.[33,41–44] Whereas standard-molecular-weight BPA-PCs are ductile to approximately –10°C, low molecular weight PCs are brittle at room temperature. Incorporation of bulky end caps, such as *p*-cumyl phenol, enhances ductility for low molecular weight PCs and improves physical aging performance.[45,46] Blends and copolymers of BPA-PC can be prepared with enhanced low-temperature ductility (Fig. 3).

Stress–strain curves measured over a range of temperatures for a typical BPA-PC are given in Fig. 4, illustrating the tough and ductile behavior of the polymer. The area under the stress–strain curve represents the energy per unit volume or toughness of the material. Typical values for standard BPA-PC samples measured at room temperature are approximately 65 J/m^3 for energy per unit volume, 62 MPa for yield stress, and elongation at break around 120% (Table 2). These relatively high values make PC a useful material for many engineering applications and have allowed PC and PC blends to replace metal in applications requiring high strength-to-weight ratios, for example in many automotive parts. Fillers such as chopped glass fiber are used in BPA-PC to provide enhanced strength and stiffness and dimensional stability for engineering applications.

BPA-PC is soluble in a number of chlorinated solvents, including methylene chloride (high solubility, 350 g/L at 25°C), chloroform, *cis*-1,2-dichloroethylene, and *sym*-tetrachloroethane, and is soluble in warm chlorobenzene and *o*-dichlorobenzene. Non-halogenated solvents for BPA-PC include tetrahydrofuran, dioxane, pyridine, and cresols. Polycarbonate is not soluble in aliphatic alcohols, esters, or ketones and exhibits very low water solubility or swellability (see Table 2) and excellent resistance to hydrolysis.[47,48] BPA-PCs are generally resistant to aqueous acids and bases, though exposure to base may lead to chemical etching, and prolonged exposure to strong acids or bases results in molecular weight degradation. Crazing and cracking occur in stressed PC samples on exposure to amines, strong base, aliphatic and aromatic hydrocarbons, and certain plasticizers.[49–51] Acetone induces rapid crystallization and cracking of molded PC parts. Copolymers and blends of BPA-PC may be prepared with enhanced chemical resistance. A silicone or acrylic hard coat is often used to coat PC sheet and film products (often with added UV stabilizer) to improve chemical, UV, and scratch resistance.

BPA-PCs exhibit plastic deformation at temperatures above 200°C. Polycarbonates are processed at temperatures ranging from 270°C to 345°C, depending on the molecular weight of the polymer.[52] At these processing temperatures, linear BPA-PCs exhibit largely Newtonian flow behavior over a wide range of shear rates. Fig. 5 shows the rheological performance of a medium-flow PC at three different temperatures. Branched PCs, prepared by incorporation of small amounts of polyfunctional phenolic groups,[53–60] demonstrate pseudoplastic behavior, as shown in Fig. 6. Thus, high molecular weight branched PCs are more easily melt-processed, making them ideal for extrusion and blow-molding applications. Copolymers and blends of BPA-PC may be prepared that provide enhanced flow behavior. Polycarbonate melt rheology has recently been reviewed.[61]

Polycarbonate displays excellent optical properties, with high transparency (transmission ~90%), low haze

Table 2 Physical and mechanical properties of typical commercial PC resins

	GE LEXAN resin grade					
Property	**OQ1022**	**HF1110**	**121**	**141**	**101**	**131**
Melt flow rate (g/10 min) (ASTM D1238, 1.2 kgf)	11[a]	25[b]	17.5[b]	10.5[b]	7[b]	3.5[b]
Vicat softening temperature (°C) (rate B/50, ASTM D1525)	140	N.R.[c]	154	154	154	154
Tensile properties (Type 1, 50 mm/min; ASTM D638)						
Tensile stress, yield (MPa)	61	62	62	62	62	62
Tensile stress, break (MPa)	51	66	69	69	69	66
Tensile strain, break (%)	100	120	125	130	135	110
Flexural modulus (MPa) (1.33 mm/min, 50 mm span; ASTM D790)	2460	2310	2340	2340	2340	2340
Impact properties						
Izod impact, notched (J/m) (23°C; ASTM D256)	550	641	694	801	908	908
Instrumented impact (J) (total energy, 23°C; ASTM D3763)	46	54	62	64	65	N.R.
Physical properties						
Specific gravity (ASTM D792)	1.19	1.2	1.2	1.2	1.2	1.2
Water absorption (%) (equilibrium, 23°C; ASTM D570)	0.2[d]	0.35	0.35	0.35	0.35	0.35
Light transmission (%) (ASTM D1003)	N.R.	88	88	88	88	88
Haze (%) (ASTM D1003)	N.R.	1	1	1	1	1
Refractive index (ASTM D542)	N.R.	N.R.	1.586	1.586	1.586	1.586

[a]250°C.
[b]300°C.
[c]Not reported.
[d]ISO 62.
(From Ref.[5]. Courtesy of GE.)

(<2%), high refractive index (1.588), low color, and the ability to be molded into articles with low birefringence. These properties, combined with its low density in comparison to glass, make PC ideal as a replacement for glass in applications such as optical lenses and windows. High-flow, high-purity PC grades are used in optical media applications, like CDs and DVDs. Additionally, PC is easily colored with dyes, pigments, and/or specialty additives (e.g., phosphorescent dyes, metallic particles) to provide a wide range of transparent, translucent, or opaque materials with desired visual effects.

BPA-PC absorbs in the UV region with a maximum at approximately 255 nm. On exposure to UV or sunlight, PC becomes yellow and exhibits surface degradation,[62] so for applications involving sun exposure, UV stabilizers are added.[63] Outdoor sheet applications generally employ a silicone or acrylate hard coat to provide improved weathering and scratch resistance.[64]

Pure BPA-PC shows excellent thermal stability at temperatures of up to 350°C. It is important to thoroughly dry the polymer before processing, and to ensure that impurities that might promote thermal oxidative degradation of the polymer, such as residual catalyst or metals, are removed. Generally, thermal oxidative stabilizers such as phosphites or phosphonites are added to prevent degradation and yellowing of the polymer during melt processing.[65]

Polycarbonate exhibits a relatively high oxygen index of 26. It may be further flame retarded by the addition of flame-retardant additives, including tetrabromo-BPA polymer, oligomer, or other brominated additives,[66,67] alkali metal salts,[68] polytetrafluoroethylene, phosphorus-containing additives, or silicones.[69] Glass fillers or other inert fillers may also provide improved performance in Underwriters Laboratories (UL) flame testing.

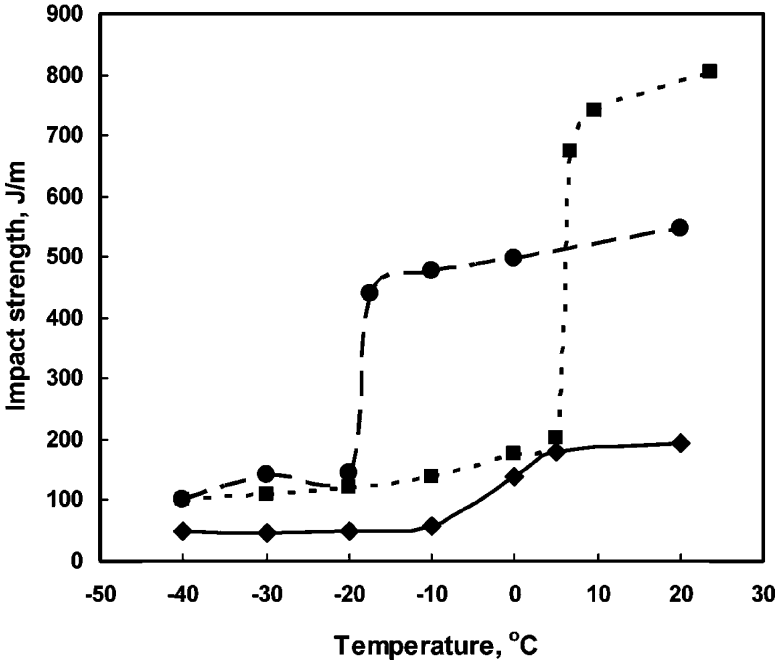


Fig. 3 Izod impact strength of PC (■), PC/ABS blend (●), and ABS (◆). (From Ref.[25]. Courtesy of GE.)

PRODUCTION PROCESSES

Interfacial Process

Until recently, virtually all commercial BPA-PC was prepared via the interfacial process (Fig. 7),[70–72] initially patented by Bayer.[18] This process involves polymerization of BPA monomer with phosgene in the presence of a catalyst at the interface of a water/organic solvent (typically methylene chloride) mixture. A monophenolic species, such as phenol or *p*-cumyl phenol, is added to control molecular weight. NaOH is added to neutralize the HCl formed and to maintain pH at reaction optimum (pH of 10–12). In a typical

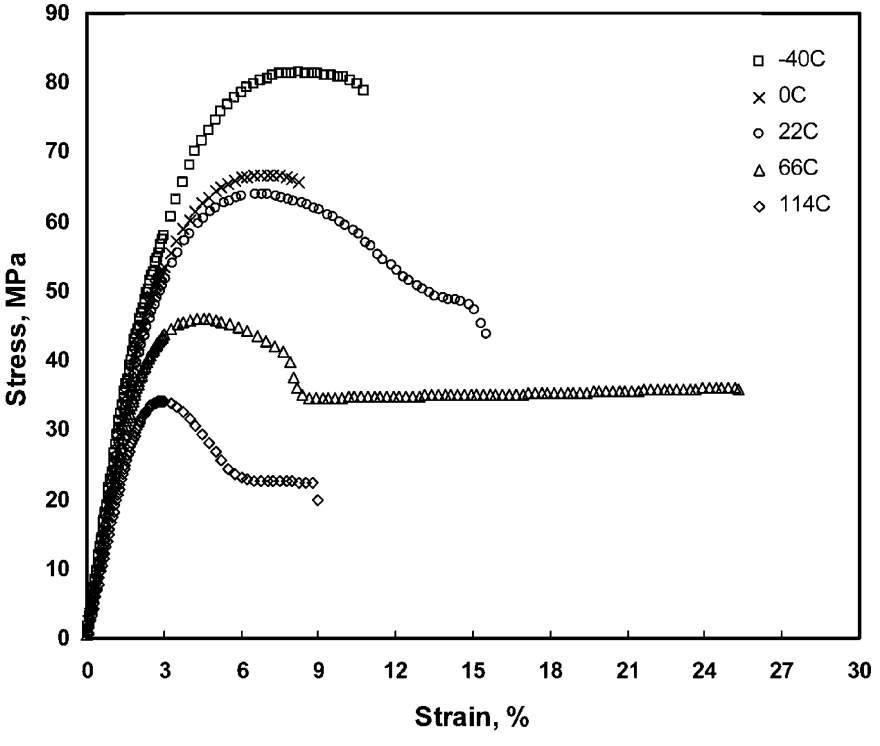


Fig. 4 Stress–strain curve for a typical commercial PC, LEXAN 141, measured at a strain rate of 0.0833%/sec. (Courtesy of GE.)

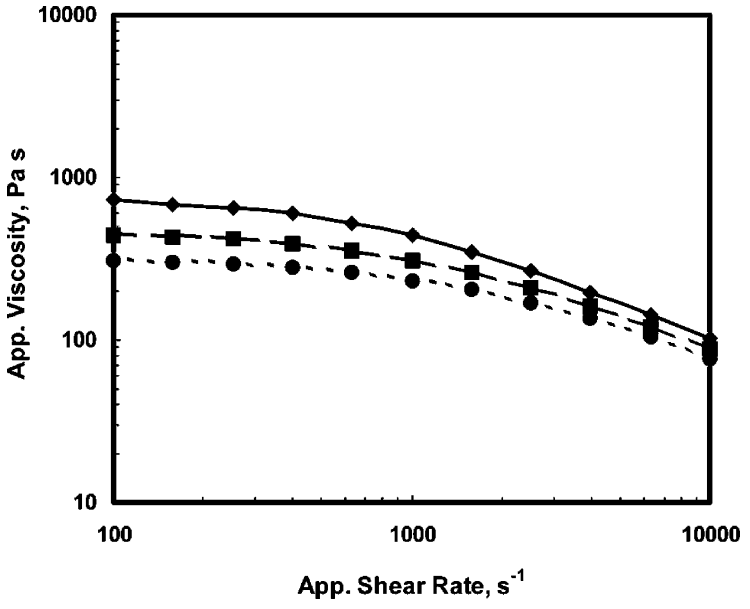


Fig. 5 Rheological performance of a medium-flow PC (LEXAN 141 resin) at three different temperatures (◆, 290°C; ■, 305°C; ●, 320°C). (Courtesy of GE.)

reaction, BPA is slurried in water with a monophenolic chain stopper (1–5%) and sodium hydroxide. Methylene chloride and a small amount of a tertiary amine catalyst are added and the mixture is stirred rapidly. Gaseous phosgene is added continuously to the stirred mixture until all phenolic groups are reacted, and NaOH is added as necessary to maintain pH. The reaction may be over-phosgenated to ensure complete reaction of all phenolic species. When the reaction is completed, the organic phase is separated from the brine phase, washed with acid to remove residual base and amine, and then washed with water to bring pH to neutral. The polymer may then be isolated via various methods, including melt devolatilization, anti-solvent precipitation, and spray-drying.[73–75]

The interfacial reaction proceeds in a complex, four-phase mixture (solid BPA, gaseous phosgene, organic phase in which polymer is soluble, and aqueous phase in which NaCl dissolves), making study and understanding of precise mechanisms difficult, although a number of models have been proposed.[71,76–82] Efficient mixing and control of pH are extremely important in controlling the reaction. A number of modifications to the basic process have appeared in the patent literature, including controlled addition of chain stoppers to control formation of oligomers[83,84] and processes to improve productivity and/or the quality[85–87] of the final product. Polycarbonate is produced industrially via both batch and continuous reactions.[24]

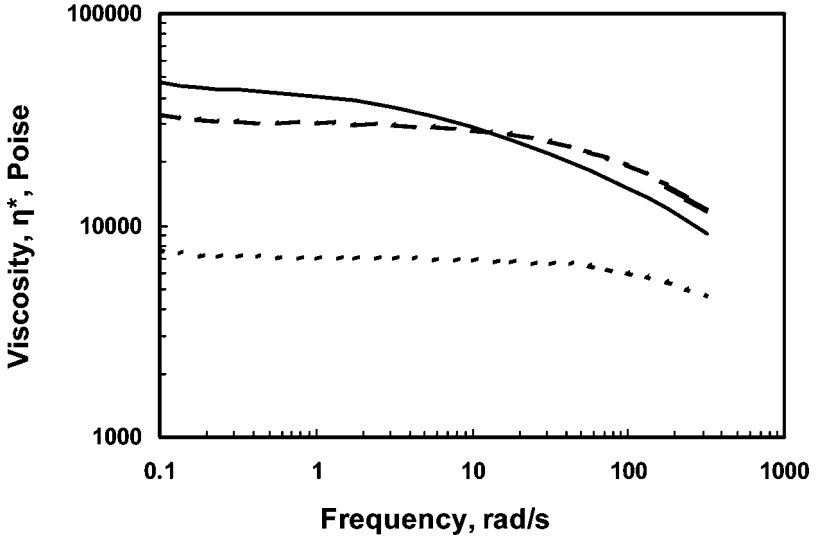


Fig. 6 Dynamic viscosity measurements of branched and linear PC resins, 530°F (·····, LEXAN 120 resin; – – –, LEXAN 130 resin; ——, LEXAN branched resin).

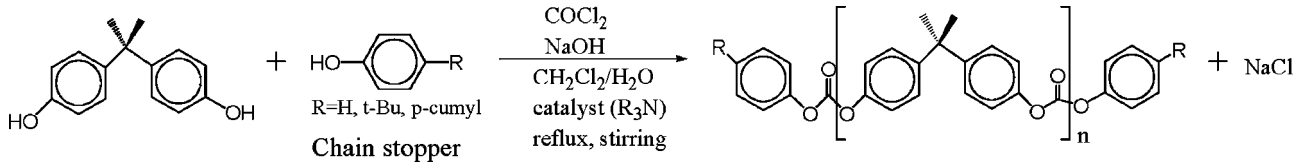


Fig. 7 Interfacial polymerization scheme for BPA-PC.

Melt or Transesterification Process

A melt process was originally developed by GE for the production of BPA-PC.[21] It was abandoned for commercial production in favor of the interfacial process because of production costs and because of the difficulty in obtaining polymer with high purity, high clarity, and low color. Work on improving the melt process continued, however, and the push to develop more ''environmentally friendly'' procedures (e.g., reduction of use of volatile organic solvents and phosgene) supplied further incentive for perfecting the process. Improvements have been achieved that allow production of BPA-PC with color and transparency of the same level as those of PC produced by the interfacial method.[25] GE currently operates a 150,000 ton per year melt production plant in Cartagena, Spain, and has announced plans to expand capacity to 300,000 metric tons per year at this site.[4,12] Additionally, GE operates a 35,000 ton per year plant in Chiba, Japan, originally built to test and validate the melt process technology.[25]

The melt polymerization process involves the base-catalyzed transesterification reaction of BPA with diphenyl carbonate (Fig. 8). A small amount (less than 0.01% molar) of basic catalyst such as Na, Li, K, or tetralkylammonium hydroxide or carbonate is used during the initial stages of the reaction. The reaction is performed under vacuum at 180–300°C. At later stages of the reaction, the temperature and the vacuum are increased (less than 1 mmHg) to remove phenol and drive the product to high molecular weight. Subsequently, the polymer becomes very viscous, and special devices, such as devolatilizing extruders, are required to ensure complete removal of phenol.

Because of the high temperatures experienced by the polymer in the melt process, initial attempts at producing melt BPA-PC resulted in highly colored product due to thermal oxidative reactions occurring during the process. Process improvements, including the use of highly pure starting materials, modified catalysts, and more reactive carbonates, have been developed that allow production of BPA-PC of high purity and transparency via the melt process.[88–96] Unlike the interfacial product, the melt product is an equilibrium product and does not undergo further transesterification reactions during subsequent melt processing. Thus, polydispersity and the amount of low molecular weight species (such as oligomers, cyclics, and diaryl carbonates) are lower for melt-produced BPA-PC than for interfacial PC. Another difference between interfacial and melt PC is that the melt product is not fully end-capped, so OH terminal groups are present in some amount in the melt product.

Initially, the melt process utilized diphenylcarbonate produced from phosgene.[97] Multiple processes are now available for entirely phosgene-free production.[98–102]

Asahi Kasei Corporation has developed a solventless, phosgene-free melt production process that uses CO_2 as a starting material and employs a reactor that uses gravity, rather than melt mixing equipment, to transfer molten polymer.[103–105] Chimei-Asahi has implemented the technology in a 50,000 ton per year commercial demonstration plant and has announced plans to expand production at their own plant and with other partners.[106]

Other Processes

There are a number of other processes for PC production that are under development or have limited current commercial application. Redistribution involves reaction of high molecular weight PC in an extruder with a catalyst that promotes transesterification.[107] The polymer undergoes ''redistribution'' to yield an equilibrium product similar to that obtained by the melt process, with reduced oligomer levels and low

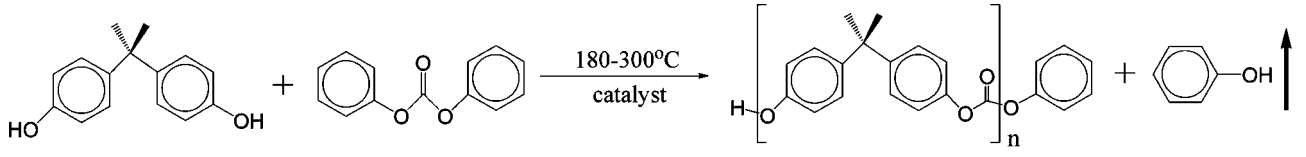


Fig. 8 Melt polymerization scheme for BPA-PC.

polydispersity index. Other monomers may also be added to the melt mixture to form copolymers or branched materials.

Ring opening polymerization of cyclic oligomers of BPA can be performed in the melt (200–300°C) or in solution at ambient temperatures.[108,109] This process can be used to obtain unique copolymers and ultra-high molecular weight polymers and enables processing techniques such as resin-transfer molding for production of composites.

Solid-state polymerization involves first-step production of low molecular weight polymer or oligomer via melt or interfacial process. The low molecular weight material is then crystallized in acetone. Basic catalyst is added and the material is heated above T_g but below crystalline melting temperature (T_m) to polymerize, and phenol is removed. The resulting polymer is melt-processed to remove crystallinity and form amorphous PC.[110,111]

Aliphatic PCs have traditionally been of low interest commercially because of their low strength and durability. There is increasing interest in the production of aliphatic PCs and their copolymers with cyclic esters for potential biomedical applications, due to their biocompatibility, low toxicity, and biodegradability, and the less acidic degradation product compared with conventional biodegradable polyesters such as poly(lactic acid). One example is poly(trimethylenecarbonate).[112–119] Recently, increasing attention has been paid to synthetic aliphatic PCs bearing functional groups to regulate hydrophilicity/hydrophobicity, permeability, bioresorption, and mechanical properties.[120–126] Preparation of functional aliphatic PCs via biocatalysis has been recently reviewed.[127,128]

POLYMER PROCESSING AND FABRICATION

Polycarbonate is readily processed by all major thermoplastic processing techniques, including injection molding, extrusion, blow molding, thermoforming, and solution casting. Polycarbonate should be thoroughly dried before melt processing, and generally a small amount of thermal oxidative stabilizer such as phosphite or phosphonite is added to prevent degradation and yellowing of the polymer.[65] Melt processing can be performed using single- or twin-screw extruders, with twin-screw generally preferred for highly filled products or blends. Polycarbonates are processed at temperatures ranging from 270°C to 345°C, depending on the molecular weight of the polymer and the specific application/processing technique. Recommended processing parameters for specific commercial grades are provided by PC producers.[5–7,52] Special processing techniques are required for high optical purity applications such as lenses and CDs, and it is important to prevent any contamination during the processing steps. Polycarbonate can be extruded into sheet, film, or tubing, and PC sheet can be thermoformed into complex shapes. Additionally, PC can be cold-formed, and rods or slabs can be machined into desired shapes. Secondary operations and part assembly for PC products have recently been reviewed.[129,130]

BLENDS AND COPOLYMERS

The most important commercial blends of BPA-PC are poly(acrylonitrile–butadiene–styrene) (PC/ABS) and polybutylene terephthalate (PC/PBT) or polyethylene terephthalate (PET). Commercial grades of PC/ABS include CYCOLOY® (GE), Bayblend® (Bayer), and PULSE® (Dow). PC/ABS blends exhibit improved flow and processability and enhanced low-temperature impact strength in comparison to PC (Fig. 3). These blends are widely used in applications requiring enhanced impact resistance, such as interior automotive parts and computer and electronics applications such as computer housings and cell phones. Non-halogenated flame-retardant PC/ABS blends are widely available. Poly(acrylic–styrene–acrylonitrile) (PC/ASA) blends (GELOY®, GE; Luran®, BASF) provide improved weatherability for outdoor applications such as exterior automotive parts, but exhibit reduced impact performance at low temperatures in comparison to PC/ABS. PC/PBT or PET blends (XENOY®, GE; Makroblend®, Bayer) provide enhanced chemical resistance and weatherability for applications such as lawn and garden equipment and automotive bumpers and fascias.

A number of commercially significant PC copolymers are produced. In addition to the previously discussed branched PCs (for extrusion and blow-molding applications) and copolymers of BPA with tetrabromo-BPA for enhanced flame retardancy, high-T_g polyester carbonate copolymers have been produced for a number of years (Bayer Apec®, GE LEXAN PPC; T_g approximately 190°C). Polyester carbonate copolymers can be produced via copolymerization of BPA with diacyl chlorides. Aromatic diacids produce high-T_g copolymers, while aliphatic diacids yield lower-T_g copolymers. A lower-T_g PC aliphatic polyester copolymer (GE LEXAN SP resin) exhibits enhanced flow and ductility in comparison to standard PC[131] and is useful for thin-wall injection molding applications requiring ductility and ease of melt processability, such as personal communication devices. GE has recently introduced two new PC copolymers, a PC–siloxane copolymer (LEXAN EXL) [132,133] and a copolymer of isophthalate terephthalate resorcinol with BPA (LEXAN SLX).[134] The PC–siloxane copolymers (LEXAN EXL resins) exhibit ultra-low temperature impact resistance (–60°C ductility), flame retardancy, and weatherability

for applications such as telecommunications and outdoor electrical equipment. Highly weatherable isophthalate terephthalate resorcinol copolymers (LEXAN SLX resins) undergo chemical rearrangement on exposure to UV light, yielding a protective weatherable surface layer for molded or extruded parts, and can replace paint in applications like exterior automotive body parts.

APPLICATIONS

Polycarbonate and its blends are used in a wide range of applications. Polycarbonate's clarity and optical properties combined with its toughness and impact resistance make it ideal for a large number of transparent products. Injection-molded transparent products include CDs and DVDs, optical lenses, sunglasses, protective eyeware, camera components, automotive lenses, and lighting. Extruded and thermoformed transparent PC products, including multilayer products with protective UV- and scratch-resistant surface layers, can be found in unbreakable windows for airplanes, trains, and schools, bullet-proof glass laminates, stadium roofs, skylights, and greenhouse roofs. Bayer and GE have established a joint venture to develop PC automotive glazing products.[135] Blow-molded applications for PC include refillable water bottles. Polycarbonate grades are increasingly used in medical applications, including blood reservoirs, membrane cartridges, filters, tubing connectors, syringes, and dental and surgical devices.[136] Medical-grade PC products can be sterilized via autoclaving, gamma sterilization, and radiation sterilization techniques. Transparent and translucent PCs with special colors and visual effects (e.g., metallic look or speckles, phosphorescence, light diffusion) are popular in applications such as computer housings and personal communication devices. Polycarbonate blends and filled products are used in opaque engineering thermoplastic applications. Key markets for PC/ABS blends include automotive applications such as instrument panels and interior components, and computer and business equipment including computer housing, printers, fax machines, personal pagers, and communication devices. PC/PBT blends are used in applications requiring chemical resistance, such as outdoor recreation equipment, lawn equipment, large structural parts, and automotive bumpers and fascias. Special grades of PC and PC blends are produced that meet global and U.S. agency standards for specific applications, including ECO-label products, UL requirements (for flame-retardant and electrical applications), FDA requirements (for food contact applications), automotive engineering requirements, and medical sterilization requirements. Continued developments in new copolymers, blends, and processes drive continued expansion of PC materials into new applications.

CONCLUSIONS

Polycarbonate is a unique and important engineering thermoplastic, with sustained growth in production, product and process development, publications, and patents. Its global consumption is projected to grow at a rate of 7% per year, with announced production expansions throughout the world. Polycarbonate's transparency, toughness, and temperature resistance, combined with its ability to form a large number of useful copolymers, blends, and filled and colored products, make it the polymer of choice for a wide range of applications. Continued research and development in PC production processes will ensure the continued growth of this important material.

ACKNOWLEDGMENTS

The authors gratefully acknowledge GE Advanced Materials and GE Global Research, especially Dr. Jim Scobbo and Dr. Andy Poslinski, for providing polycarbonate data and assistance in preparing this article.

REFERENCES

1. Results from search on SciFinder Scholar, American Chemical Society, September 2004.
2. Pardos, F. Engineering and specialty plastics in the world. 2002; http://pardos.marketing.free.fr/37.htm (accessed October 2004).
3. CMAI 2003 World cumene/phenol and polycarbonate analysis. 2003; http://www.cmaiglobal.com/news/WCPPA.pdf (accessed October 2004).
4. Westervelt, R. GE advanced materials moves forward. Chemical Week, March 24, 2004; http://www.chemweek.com (accessed October 2004).
5. http://www.gelexan.com/gelexan (accessed October 2004).
6. http://www.makrolon.com (accessed October 2004).
7. http://www.dow.com/engineeringplastics/prod/na/cal.htm (accessed October 2004).
8. http://www.teijin.co.jp/english/about/enterprise/index04.html (accessed October 2004).
9. http://www.m-ep.co.jp/mep-en/index.htm (accessed October 2004).
10. Alperowicz, N. Asahi Kasei licenses PC process to Russian firm. Chemical Week, September

15, 2004; http://www.chemweek.com (accessed October 2004).
11. Bayer News. Address by Werner Wenning. November 26, 2003; http://www.press.bayer.com/news/news.nsf/id/9F466177A10D76FEC1256DE9004C9D89 (accessed October 2004).
12. Esposito, F. GE adding capacity in Spain and China. Plastics News, December 8, 2003; http://www.plasticsnews.com (accessed October 2004).
13. King, J.A., Jr. Synthesis of polycarbonates. In *Handbook of Polycarbonate Science and Technology*; Plastics Engineering; 2000; Vol. 56, 7–26.
14. Einhorn, A. Liebigs Ann. Chem. **1898**, *300*, 135.
15. Bischoff, C.A.; Hedenstroem, A.V. Berichte **1902**, *35*, 3531.
16. Carothers, W.H.; van Natta, F.J. Polymerization and ring formation. III. Glycol esters of carbonic acid. J. Am. Chem. Soc. **1930**, *52*, 314–326.
17. http://corporate.ppg.com/PPG/opticalprod/en/monomers/products/default.htm (accessed October 2004).
18. Schnell, H.; Bottenbruch, L.; Krimm, H. Belgian Patent 532,543, 1954
19. Fox, D.W. Australian Patent 221,192, 1959
20. Schnell, H.; Bottenbruch, L.; Krimm, H. Thermoplastic Aromatic Polycarbonates and Their Manufacture . US Patent 3,028,365, April 3, 1962.
21. Fox, D.W. Aromatic Carbonate Resins and Preparation Thereof US Patent 3,153,008, October 13, 1964.
22. Christopher, W.F.; Fox, D.W. *Polycarbonates*; Reinhold Publishing: New York, 1962; 2 pp.
23. Schnell, H. *The Chemistry and Physics of Polycarbonates (Polymer Reviews)*; Interscience: New York, 1964; Vol. 9, 3–227.
24. Pham, H.T.; Munjal, S.; Bosnyak, C.P. Polycarbonates. In *Handbook of Thermoplastics*; Plastics Engineering; 1997; Vol. 41, 609–640.
25. Brunelle, D.J.; Kailasam, G. *Polycarbonates*, 2001, GE Global Research Technical Reports, http://www.crd.ge.com/cooltechnologies/biblio/2001crd136_bib.jsp (accessed October 2004).
26. Bostick, E.E. Introduction and historical background. In *Handbook of Polycarbonate Science and Technology*; Plastics Engineering; 2000; Vol. 56, 1–6.
27. Fox, D.W. Polycarbonates. In *Kirk-Othmer Encyclopedia of Chemical Technology*; 3rd Ed.; John Wiley and Sons, Inc: New York, 1982; 479–494.
28. Freitag, D.; Grigo, U.; Müller, P.; Nouvertné, W. Polycarbonates. In *Encyclopedia of Polymer Science and Engineering*, 2nd Ed.; Mark, H.F., Ed.; John Wiley and Sons: New York, 1988; Vol. 11, 648–718.
29. Xiao, C.; Yee, A. Scale of cooperative γ-relaxation of bisphenol A polycarbonate. Macromolecules **1992**, *25* (25), 6800–6809.
30. Floudas, G.; Higgins, J.S.; Meier, G.; Kremer, F.; Fischer, E.W. Dynamics of bisphenol A polycarbonate in the glassy and rubbery states as studied by neutron scattering and complementary techniques. Macromolecules **1993**, *26* (7), 1676–1682.
31. Jones, A.A. Molecular level model for motion and relaxation in glassy polycarbonate. Macromolecules **1985**, *18* (5), 902–906.
32. Wilkes, G.L. Mechanical properties. In *Polymer Characterization and Analysis*; Brady, R.F., Ed.; American Chemical Society/Oxford University Press: New York, 2003; 624–668.
33. LeGrand, D.G. Mechanical properties of polycarbonates. In *Handbook of Polycarbonate Science and Technology*; Plastics Engineering; 2000; Vol. 56, 107–130.
34. Bendler, J.T.; Shlesinger, M.F. Quantum molecular orbital calculations, Levy-stable distributions, and molecular relaxation in polycarbonate. In *Handbook of Polycarbonate Science and Technology*; Plastics Engineering; 2000; Vol. 56, 27–42.
35. Bendler, J.T.; Shlesinger, M.F. Defect-diffusion models of relaxation. J. Mol. Liq. **1987**, *36*, 37–46.
36. LeGrand, D.G.; Erhardt, P.F. Dynamic mechanical properties of polymers. Appl. Polym. Sci. **1969**, *13*, 1707–1719.
37. Bailly, C.; Daoust, D.; Legras, R.; Mercier, J.P.; Strazielle, C.; Lapp, A. On the molecular weight determination of bisphenol-A polycarbonate. Polymer **1986**, *27* (9), 410–415.
38. Berry, G.C.; Nomura, H.; Mayhan, K.G. Dilute solution studies on a polycarbonate in good and poor solvents. J. Polym. Sci. Polym. Phys. Ed. **1967**, *5* (1), 1–21.
39. De Chirico, A. Polycarbonate solutions in chloroform. Chim. Ind. **1960**, *42*, 248–251.
40. Tsuji, T.; Norisuye, T.; Fujita, H. Dilute solution of bisphenol A polycarbonate. Polym. J. **1975**, *7* (5), 558–569.
41. LeGrand, D.G. Crazing, yielding, and fracture of polymers 1. Ductile–brittle transition in polycarbonate. J. Appl. Polym. Sci. **1969**, *13*, 2129–2147.
42. Parvin, M.; Williams, J.G. Ductile–brittle fracture transitions in polycarbonate. Int. J. Fract. **1975**, *11* (6), 963–972.
43. Hyakutake, H.; Nishitani, H. Condition for determining the static yield and fracture of a polycarbonate plate specimen with notches. Eng. Fract. Mech. **1985**, *22* (3), 359–368.

44. Martin, G.C.; Gerberich, W.W. Temperature effects on fatigue crack growth in polycarbonate. J. Mater. Sci. **1976**, *11* (2), 231–238.
45. LeGrand, D.G.; Morgan, S.; McKenley, B.; McCloskey, P.; Snow, K.; Sybert, P.; Dardaris, D.; Caraher, J. The effect of endgroups and molecular weight on the thermal aging of polycarbonate resins. ANTEC Soc. Plastics Eng. **1994**, *52* (2), 2067–2070.
46. Okamoto, M. Polycarbonate-Made Optical Article US Patent 4,997,903, March 5, 1991.
47. Gardner, R.J.; Martin, J.R. Humid aging of plastics: effect of molecular weight on mechanical properties and fracture morphology of polycarbonate. J. Appl. Polym. Sci. **1979**, *24* (5), 1269–1280.
48. Pryde, C.A.; Kelleher, P.G.; Hellman, M.Y.; Wentz, R.P. The hydrolytic stability of some commercially available polycarbonates. Polym. Eng. Sci. **1982**, *22* (6), 370–375.
49. Kambour, R.P. Structure and properties of crazes in polycarbonate and other glassy polymers. J. Polym. Sci. **1964**, *5* (3), 143–155.
50. Kambour, R.P.; Romagosa, E.E.; Gruner, C.C. Bisphenol-A polycarbonate immersed in organic media. Swelling and response to stress. Macromolecules **1974**, *7* (2), 248–253.
51. White, S.A.; Weissman, S.R.; Kambour, R.P. Resistance of a polyetherimide to environmental stress crazing and cracking. J. Appl. Polym. Sci. **1982**, *27* (7), 2675–2682.
52. GE Plastics LEXAN® resin data sheets; http://www.geplastics.com/dsw/dsw (accessed October 2004).
53. Rawlings, H.L. Branching and Crosslinking of Polycarbonates US Patent 3,597,394, August 3, 1971.
54. Hedges, V. Randomly Branched Aromatic Polycarbonate from Triphenol US Patent 4,415,723, November 15, 1983.
55. Freitag, D.; Haberland, U.; Krimm, H. Tetraphenol Containing Polycarbonate US Patent 4,001,183, January 4, 1977.
56. Mark, V. Polyphenolic Compounds US Patent 4,514,334, April 30, 1985.
57. Mark, V.; Hedges, C.V. Branched Aromatic Polycarbonate from Tetraphenol US Patent 4,415,724, November 15, 1983.
58. Boden, E.P.; Krabbenhoft, H.O. Preparation of Branched Thermoplastic Polycarbonate from Polycarbonate and Polyhydric Phenol US Patent 4,888,400, December 19, 1989.
59. Mark, V.; Hedges, C.V. Branched Polycarbonate from Carboxy Containing Diphenol US Patent 4,562,242, December 31, 1985.
60. Scott, S.W. Process for Preparing a Branched Polycarbonate US Patent 4,001,184, January 4, 1977.
61. Jordan, T.C.; Richards, W.D. Polycarbonate melt rheology. In *Handbook of Polycarbonate Science and Technology*; Plastics Engineering; 2000; Vol. 56, 179–224.
62. Factor, A. Search for the sources of color in thermally aged, weathered, and γ-ray irradiated bisphenol A polycarbonate. Angew. Makromol. Chem. **1995**, *232*, 27–43.
63. Gugumus, F. Light stabilizers. In *Plastics Additives Handbook*, 3rd Eds.; Gachter, R., Muller, H., Klemchuk, P.P., Eds.; Hanser Publishers: New York, 1990; 129–262.
64. Medford, G.; Pickett, J.; Reynolds, C. The next generation of weatherable hardcoats for polycarbonate. PCI, 2001. (accessed October 2004).; http://www.pcimag.com/pci/cda/articleinformation/coverstory/bnpcoverstoryitem/0,29001,00 + en-uss_01dbc.html (accessed October 2004).
65. Gugumus, F. Autooxidation of synthetic polymers. In *Plastics Additives Handbook*, 3rd Ed.; Gachter, R., Muller, H., Klemchuk, P.P., Eds.; Hanser Publishers: New York, 1990; 1–104.
66. Mark, V.; Webb, J.L.; Williams, J.B. Flame Retardant Polycarbonate Composition US Patent 4,263,201, April 21, 1981.
67. Troitzsch, H.J. Flame retardants. In *Plastics Additives Handbook*, 3rd Ed.; Gachter, R., Muller, H., Klemchuk, P.P., Eds.; Hanser Publishers: New York, 1990; 709–748.
68. Morgan, S.E.; Davis, G.; Gijzen, E.; Lemmen, T. The performance of sulfonate salt-containing polycarbonate products in flammability testing. ANTEC Soc. Plastics Eng. **1994**, *3*, 2840–2845.
69. MacLaury, M.R.; Holub, F.H. Flame Retardant Compositions and Coated Article US Patent 4,273,691, June 16, 1981.
70. Vernaleken, H. Interfacial synthesis. In *Polymer Applications and Technology*; Millich, F., Carraher, C.E., Jr., Eds.; Marcel Dekker: New York, 1982; Vol. 2, 65 pp.
71. Gu, J.T.; Wang, C.S. The interfacial polycarbonate reactions. I. Defining the critical process parameters. J. Appl. Polym. Sci. **1992**, *44* (5), 849–857.
72. Dobkowski, Z.; Weilgosz, Z.; Krajewski, B. Molecular weight control of polycarbonates obtained by interfacial polycondensation. Angew. Makromol. Chem. **1974**, *39* (1), 7–20.
73. Walko, L.E.; Wallace, S.B.; Swanson, N.; Cook, R.M. Process for Converting Polymer Solutions into Granules US Patent 4,568,418, February 4, 1986.

74. Asoh, T.; Hosomi, T. Process for Producing Powder Aggregate Particle Formation from High Molecular Weight Polycarbonates European Patent 488,190, June 3, 1992.
75. Yamamoto, J. Purification of Polycarbonate Particles for Use in Memory Disks and Lenses Japanese Patent 63,218,730 A2, September 12, 1988.
76. King, J., Jr.; Bryant, G.L., Jr. Structures of two *N*-acyl triethyl ammonium salts and one simple triethyl ammonium salt. Acta Crystallogr. Sect. C Cryst. Struct. Commun. **1991**, *C47* (10), 2249–2252.
77. King, J.A., Jr.; Bryant, G.L., Jr. Preparation and characterization of crystalline *N*-acylammonium salts. J. Org. Chem. **1992**, *57* (19), 5136–5139.
78. Kosky, P.G.; Boden, E.P. The interfacial polycarbonate reaction: modeling the kinetics of carbamate side reactions. J. Polym. Sci. A. Polym. Chem. **1991**, *28* (6), 1507–1518.
79. Aquino, E.; Brittain, W.J.; Brunelle, D.J. Rates of carbonate bond formation: implications for macrocyclization. Polym. Int. **1994**, *33* (2), 161–164.
80. Aquino, E.; Brittain W.J., Brunelle, D.J. Mechanistic studies of carbonate macrocyclization. J. Polym. Sci. A Polym. Chem. **1994**, *32* (4), 741–746.
81. Mills, P.L. Design of multiphase gas–liquid polymerization reactors with application to polycarbonate polymerization. Ind. Eng. Chem. Process Des. Dev. **1986**, *25* (2), 575–586.
82. Mills, P.L. Analysis of multiphase polycarbonate polymerization in a semibatch reactor. Chem. Eng. Sci. **1986**, *41* (11), 2939–2952.
83. Munjal, S.; Wardlow, T.M.; Hall, A.F. Process for Preparing Polycarbonate Having Reduced Carbonate Byproduct Content US Patent 5,200,496, April 6, 1993.
84. Hirao, M.; Tonati, Y.; Ito, T.; Nakatsuka, M.; Yamahuchi, A. Aromatic Polycarbonate of Low Oligomer Content and Narrow Molecular Weight Distribution European Patent 600,447 A2, June 8, 1994.
85. Silva, J.M.; Kosky, P.G. The aqueous phase in the interfacial synthesis of polycarbonates. 2. Application of ionic equilibrium to the semibatch polycarbonate reaction. Ind. Eng. Chem. Res. **1991**, *30* (3), 468–474.
86. Megumi, T.; Yoshizaki, H.; Kondoh, S. Preparing Polycarbonate Oligomers by Reacting Phosgene with an Aqueous Solution Containing a Salt of a Dihydric Phenol in an Organic Solvent Which Comprises Previous Cooling the Aqueous Solution and Carrying Out the Reaction at a Low Temperature US Patent 4,255,557, March 10, 1981.
87. Becraft, M.L.; Ramsey, D.L. Reduced Phosgene Consumption During Preparation of Polycarbonate European Patent 644,220 A1, March 22, 1995.
88. Ko, A.W.; Starr, J.B., Jr Inhibition of Discoloration of Transesterification Polymers with Chromium, Nickel, Tantalum or Glass-Lined Reactor. US Patent 4,383,092, May 10, 1983.
89. Sakashita, T.; Shimoda, T. Process and Catalysts for the Preparation of Aromatic Polycarbonates with Low Color European Patent 351,168 A2, January 17, 1990.
90. Sakashita, T.; Shimoda, T. Process for the Production of Polycarbonates European Patent 360,578, March 28, 1990.
91. Sakashita, T.; Shimoda, T. Catalytic Process for Preparing Polycarbonates from Carbonic Acid Diester US Patent 5,026,817, June 25, 1991.
92. Sakashita, T.; Shimoda, T. Polycarbonate Containing Terminal Hydroxyl Groups, Low Sodium and Chlorine Content US Patent 5,276,129, January 4, 1994.
93. Sakashita, T.; Shimoda, T. Catalytic Melt Preparation of Polycarbonate with Subsequent Addition of Sulfonic Acid Compound US Patent 5,306,801, April 26, 1994.
94. Sakashita, T.; Shimoda, T.; Nagai, T. Optical Polycarbonate Compounds US Patent 5,276,109, January 4, 1994.
95. Kuze, S.; Okumura, R.; Kunishi, N. Polycarbonates and Their Manufacture European Patent 575,810 A2, December 29, 1993.
96. Kanno, T.; Fukuda, Y. Process for the Preparation of Polycarbonates by Melt-Polycondensation European Patent 592,900 A2, April 20, 1994.
97. Janatpour, M.; Shafer, S.J. Process for Making Diaryl Carbonates European Patent 228,672 A2, July 15, 1987.
98. Kim, W.B.; Joshi, U.A.; Lee, J.S. Making Polycarbonates without employing phosgene: an overview on catalytic chemistry of intermediate and precursor syntheses for polycarbonate. Ind. Eng. Chem. Res. **2004**, *43* (9), 1897–1914.
99. Rivetti, F.; Romano, U. Alcohol carbonylation with palladium(II) complexes, effects of ligands, carbon monoxide, pressure and added bases. J. Organomet. Chem. **1979**, *174* (2), 221–226.
100. Rivetti, F.; Romano, U. Alkoxy carbonyl complexes of palladium and their role in alcohol carbonylation. J. Organomet. Chem. **1979**, *154* (3), 323–326.
101. Delledonne, D.; Rivetti, F.; Romano, U. Process and Catalyst for Preparing Organic Carbonates US Patent 5,395,949, March 7, 1995.
102. Rivetti, F.; Romano, U.; Delledonne, D. Dimethyl carbonate and its production technology.

In *Green Chemistry*; ACS Symposium Series; 1996; Vol. 626, 70–80.
103. Komiya, K.; Kawakami, Y.; Okamoto, H. Wire-Wetting Fall Polymerization Process for the Production of Polycarbonates US Patent 5,589,564, December 31, 1996.
104. Fukuoka, S.; Kawamura, M.; Komiya, K.; Tojo, M.; Hachiya, H.; Hasegawa, K.; Aminaka, M.; Okamoto, H.; Fukawa, I.; Konno, S. A novel non-phosgene polycarbonate production process using by-product CO_2 as starting material. Green Chem. **2003**, *5* (5), 497–507.
105. Asahi-Kasei Press Release. Agreement to license polycarbonate technology. August 3; 2000. http://www.asahi-kasei.co.jp/asahi/en/news/2004/e040823.html (accessed October 2004).
106. Chimei-Asahi to expand polycarbonate capacity. Chemical Week, August 6, 2003. http://www.chemweek.com (accessed October 2004).
107. Campbell, A.J.; Dardaris, D., III; Faler, G.R.; McCloskey, P.J.; Evans, T.L. Redistribution of Organic Polycarbonate Compositions US Patent 5,414,057, May 9, 1995.
108. Brunelle, D.J.; Shannon, T.G. Preparation and polymerization of bisphenol A cyclic oligomeric carbonates. Macromolecules **1991**, *24* (11), 3035–3044.
109. Brunelle, D.J.; Boden, E.P.; Shannon, T.G. Remarkably selective formation of macrocyclic aromatic carbonates: versatile new intermediates for the synthesis of aromatic polycarbonates. J. Am. Chem. Soc. **1990**, *112* (6), 2399–2402.
110. Iyer, V.S.; Sehra, J.C.; Ravindranath, K.; Sivaram, S. Solid-state polymerization of poly (aryl carbonates): a facile route to high-molecular-weight polycarbonates. Macromolecules **1993**, *26* (5), 1186–1197.
111. Westeppe, U.; Freitag, D.; Fengler, G.; Grigo, U. Two-Step Manufacture of Cycloalkylidenediphenol-Based Polycarbonates German Patent 3,941,014 A1, June 13, 1991.
112. Bisht, K.S.; Svirkin, Y.Y.; Henderson, L.A.; Gross, R.A.; Kaplan, D.L.; Swift, G. Lipase-catalyzed ring-opening polymerization of trimethylene carbonate. Macromolecules **1997**, *30* (25), 7735–7742.
113. Huang, Q.; Shen, Z.; Zhang, Y.; Shen, Y.; Shen, L.; Yuan, H. Ring-opening copolymerization of trimethylene carbonate and D,L-lactide by rare earth chloride. Polym. J. **1998**, *30* (3), 168–170.
114. Chen, X.; McCarthy, S.P.; Gross, R.A. Preparation and characterization of polycarbonates from 2,4,8,10-tetraoxaspiro[5,5]undecane-3-one (DOXTC)-trimethylenecarbonate (TMC) ring-opening polymerizations. J. Appl. Polym. Sci. **1998**, *67* (3), 547–557.
115. Matsumura, S.; Tsukada, K.; Toshima, K. Novel lipase-catalyzed ring-opening copolymerization of lactide and trimethylene carbonate forming poly(ester carbonate)s. Int. J. Biol. Macromol. **1999**, *25* (1–3), 161–167.
116. Schappacher, M.; Fabre, T.; Mingotaud, A.F.; Soum, A. Study of a (trimethylenecarbonate-co-ε-caprolactone) polymer—Part 1: Preparation of a new nerve guide through controlled random copolymerization using rare earth catalysts. Biomaterials **2001**, *22* (21), 2849–2855.
117. Fabre, T.; Schappacher, M.; Bareille, R.; Dupuy, B.; Soum, A.; Bertrand-Barat, J.; Baquey, C. Study of a (trimethylenecarbonate-co-ε-caprolactone) polymer—Part 2. In vitro cytocompatibility analysis and in vivo ED1 cell response of a new nerve guide. Biomaterials **2001**, *22* (22), 2951–2958.
118. Yu, F.; Zhuo, R. Synthesis, characterization, and degradation behaviors of end-group-functionalized poly(trimethylene carbonate)s. Polym. J. **2003**, *35* (8), 671–676.
119. Ling, J.; Shen, Z.; Zhu, W. Synthesis, characterization, and mechanism studies on novel rare-earth calixarene complexes initiating ring-trimethylene carbonate. J. Polym. Sci. A Polym. Chem. **2003**, *41* (9), 1390–1399.
120. Chen, X.; McCarthy, S.P.; Gross, R.A. Synthesis, characterization, and epoxidation of an aliphatic polycarbonate from 2,2-(2-pentene-1,5-diyl)trimethylene carbonate (cHTC) ring-opening polymerization. Macromolecules **1997**, *30* (12), 3470–3476.
121. Vandenberg, E.J.; Tian, D. A new, crystalline high melting bis(hydroxymethyl)- polycarbonate and its acetone ketal for biomaterial applications. Macromolecules **1999**, *32* (11), 3613–3619.
122. Shen, Y.; Chen, X.; Gross, R.A. Aliphatic polycarbonates with controlled quantities of d-xylofuranose in the main chain. Macromolecules **1999**, *32* (12), 3891–3897.
123. Mizutani, M.; Arnold, S.C.; Matsuda, T. Liquid, phenylazide-end-capped copolymers of ε-caprolactone and trimethylene carbonate: preparation, photocuring characteristics, and surface layering. Biomacromolecules **2002**, *3* (4), 668–675.
124. Wang, X.L.; Zhuo, R.X.; Huang, S.W.; Liu, L.J.; He, F. Synthesis, characterization and in vitro cytotoxicity of poly[(5-benzyloxy-trimethylene carbonate)-co-(trimethylene carbonate)]. Macromol. Chem. Phys. **2002**, *203* (7), 985–990.
125. Ray, W.C., III; Grinstaff, M.W. Polycarbonate and poly(carbonate-ester)s synthesized from biocompatible building blocks of glycerol and

lactic acid. Macromolecules **2003**, *36* (10), 3557–3562.
126. Yang, J.; Hao, Q.; Liu, X.; Ba, C.; Cao, A. Novel biodegradable aliphatic poly(butylene succinate-co-cyclic carbonate)s with functionalizable carbonate building blocks. 1. Chemical synthesis and their structural and physical characterization. Biomacromolecules **2004**, *5* (1), 209–218.
127. Gross, R.A.; Kalra, B.; Kumar, A. Polyester and polycarbonate synthesis by in vitro enzyme catalysis. Appl. Microbiol. Biotechnol. **2001**, *55* (6), 655–660.
128. Bisht, K.S.; Al-Azemi, T.F. Biocatalytic synthesis of novel functional polycarbonates. In *Biocatalysis in Polymer Science*; ACS Symposium Series; Vol. 840, 156–171.
129. Poslinksi, A.J. Part assembly. In *Handbook of Polycarbonate Science and Technology*; Plastics Engineering; 2003; Vol. 56, 317–330.
130. Chao, H.S.; Burrell, M.C. Secondary finishing. In *Handbook of Polycarbonate Science and Technology*; Plastics Engineering; 2000; Vol. 56, 331–341.
131. Fontana, L.P.; Buckley, P.W.; Boden, E.P. Preparation of by the Bischloroformate Process of Soft Segment Polycarbonate US Patent 5,494,997, February 27, 1996.
132. Hoover, J.; Sybert, P.D. Compositions of Siloxane Polycarbonate Block Copolymers and High Heat Polycarbonates US Patent 5,455,310, October 3, 1995.
133. GE Plastics announces the next generation of polycarbonate: Clear Lexan® EXL (accessed October 2004).; resin.http://news.thomasnet.com/fullstory/21151 (accessed October 2004).
134. Naitove, M.H. New polymer uses sunlight to arm itself against UV damage. Plastics Technology 2004. http://www.plasticstechnology.com/articles/200402cu3.html (accessed October 2004).
135. http://www.exatec.de/default.aspx (accessed October 2004).
136. Powell, D.G. Medical applications of polycarbonate. Medical Plastics and Biomaterials 1998. http://www.devicelink.com/mpb/archive/98/09/003.html (accessed November 2004).

Polycyclic Aromatic Hydrocarbons (PAHs)

Teresa J. Cutright
Department of Civil Engineering, University of Akron, Akron, Ohio, U.S.A.

Sangchul Hwang
Department of Civil Engineering and Surveying, University of Puerto Rico, Mayagüez, Puerto Rico

INTRODUCTION

Polycyclic aromatic hydrocarbons (PAHs) are a class of organic compounds consisting of two or more benzene rings fused in a linear, angular, or clustered arrangement. They are formed predominantly as products of incomplete combustion. At high temperatures (e.g., in coking operations), low molecular weight PAHs are formed. Complex, multiringed PAHs occur at moderate temperatures. PAHs can also be formed over time even at low temperatures. Thus, it is not surprising that over 10,000 different PAH compounds have been identified. Due to the number of PAHs that can be formed, knowing how the ring structures are numbered and named can be critical. Fig. 1 illustrates three representative PAHs: pyrene, anthracene, and dibenz(*a*,*h*)anthracene. Typically, the PAH is shown oriented so that the greatest number of rings are in a horizontal row, with the majority being located in the top right-hand corner. As shown in Fig. 1(A), numbering starts with the first carbon position in the top-rightmost benzene ring and continues in a clockwise manner, skipping the carbons ''fused'' between rings.[1,2] The lowercase letters on the base components are used to distinguish different isomers as well as to designate where ring fusion occurs. For instance, the 3a notation in Fig. 1(A) designates a fused face after the third carbon. Using this approach, it is easy to see how the parent structure of anthracene [Fig. 1(C)] is the origin of dibenz(*a*,*h*)anthracene [Fig. 1(D)]. If there is only one ring in the top portion of the compound, numbering starts at what is typically considered the sixth carbon in a normal benzene ring [Fig. 1(E)]. Although there are a multitude of PAH structures, currently only 16 are on the U.S. Environmental Protection Agency's (USEPA) Priority Pollutants List. The specific PAHs on this list are shown in Fig. 2.

The increasing concern over PAHs stems from several of their characteristics that impact human health and the environment. The primary concerns are that PAHs: 1) are highly carcinogenic and toxic; 2) have a high potential for bioaccumulation; 3) have high occurrence of contamination; and 4) are recalcitrant to biodegradation.[3] This chapter provides an overview of the physical–chemical characteristics, sources, commercial uses, environmental regulations, fate and transport, and current remediation techniques of PAHs.

PHYSICOCHEMICAL CHARACTERISTICS

Table 1 lists the physicochemical characteristics of the 16 PAHs on the USEPA Priority Pollutants List. The extent to which a particular PAH exists in the gas or solid phase is dependent upon the compound's vapor pressure (i.e., Henry's constant) and boiling point as well as the ambient temperature. In general, an increase in molecular weight and number of ring structures leads to decreased PAH solubility and volatility, as well as increased hydrophobicity. The octanol–water partition coefficient (K_{ow}), a parameter used to describe the tendency of a compound to partition in octanol vs. water, can be used to estimate hydrophobicity. As the compound becomes more hydrophobic, K_{ow} increases, indicating that the compound preferentially sorbs to solid material vs. remaining in water. Thus, PAHs with higher molecular weights and multiple substitutions are more difficult to remediate. It is important to note that some of the physicochemical characteristics, particularly Henry's constant and solubility values, can vary depending on the estimation method used. Regardless of the source, the magnitude of the value should be similar.

SOURCES

Natural Sources

The first report on natural sources of PAHs was published in 1961 by Blumer,[4] who suggested that soil PAHs found in rural, remote areas were endogenous to the soil and not attributable to anthropogenic sources. Fifteen years later, it was found that PAHs were formed during thermal reactions associated with fossil-fuel and mineral production, natural burning of vegetation

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007910

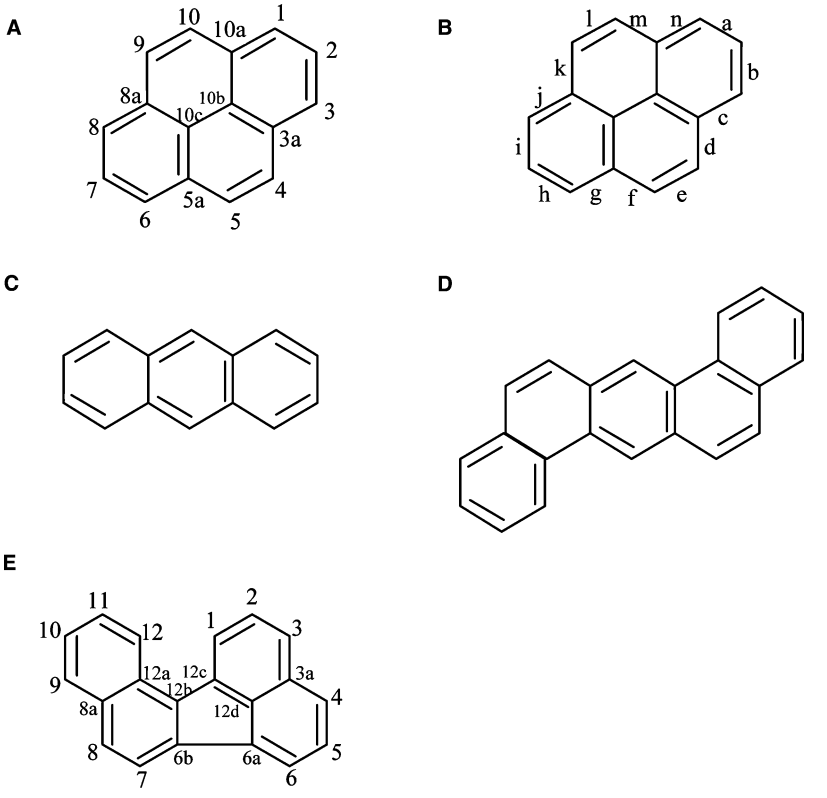


Fig. 1 Numbering and lettering approach for PAHs. (A) Numbering sequence of pyrene; (B) lettering of ''faces'' of pyrene; (C) anthracene; (D) dibenz(*a*,*h*)anthracene; (E) benzo(*j*)fluoranthene.

(i.e., forest fires), and volcanic eruptions.[5] PAHs can also by synthesized by algae[6] and terrestrial plants.[7]

Anthropogenic Sources

Although PAHs can occur naturally, their presence, particularly at high concentrations, is commonly associated with anthropogenic practices. As shown in Table 2, these processes include industrial petroleum processing (cracking, coking, gasification, etc.), accidental spills and leaks, and incomplete fuel combustion. One of the primary sources of PAH contamination is abandoned manufactured gas plant (MGP) sites. From the mid-1850s to 1950s, the primary source of heating and electricity was gas from fossil fuels.[8] To increase the BTU value of the ''manufactured'' town gas, light petroleum oils were cracked with coal (Fig. 3). During the cracking process, many of the heavier organic contaminants were condensed from the raw gas to form coal tars rich in PAHs. Like most of the unwanted byproducts of that era, the coal tar was typically disposed of onsite in unlined pits.[9] As the efficiency of electricity surpassed that of coal gasification, the MGP sites were abandoned, and their exact number and location were lost. Since early disposal pits were unlined, the PAHs sorbed to the soil, providing a long-term contamination source. Although PAHs are hydrophobic, the combination of quantity released and time since disposal has resulted in extensive contaminant migration in some areas.[10,11]

Present-day PAH contamination is also associated with wood-treatment industries, since PAHs are major components of creosote and anthracene oil—commonly used in wood-preserving pesticides. Other current industrial units that generate PAH contaminants are steel and carbon manufacturers, plasticizer and dye facilities, and some explosives generators. Generally, soil PAH concentration increases with increasing impact of industry, traffic, and domestic heating. For example, PAH concentration ranges from 5 mg/kg soil in an undeveloped area to 1.79×10^6 mg/kg at an oil refinery. In urban estuaries, PAH concentration can exceed 10^2 mg/kg.[12]

COMMERCIAL APPLICATIONS

Tar (i.e., pitch) is formed during combustion and contains a vast assortment of PAHs. The exact distribution depends on the variability of the feedstock and processing temperature. In the past, tar was used an asphalt additive and even as a roof sealer due to its viscosity. However, this and other PAH uses have diminished with time. The purification and distribution of most of the 16 PAHs on the EPA Priority

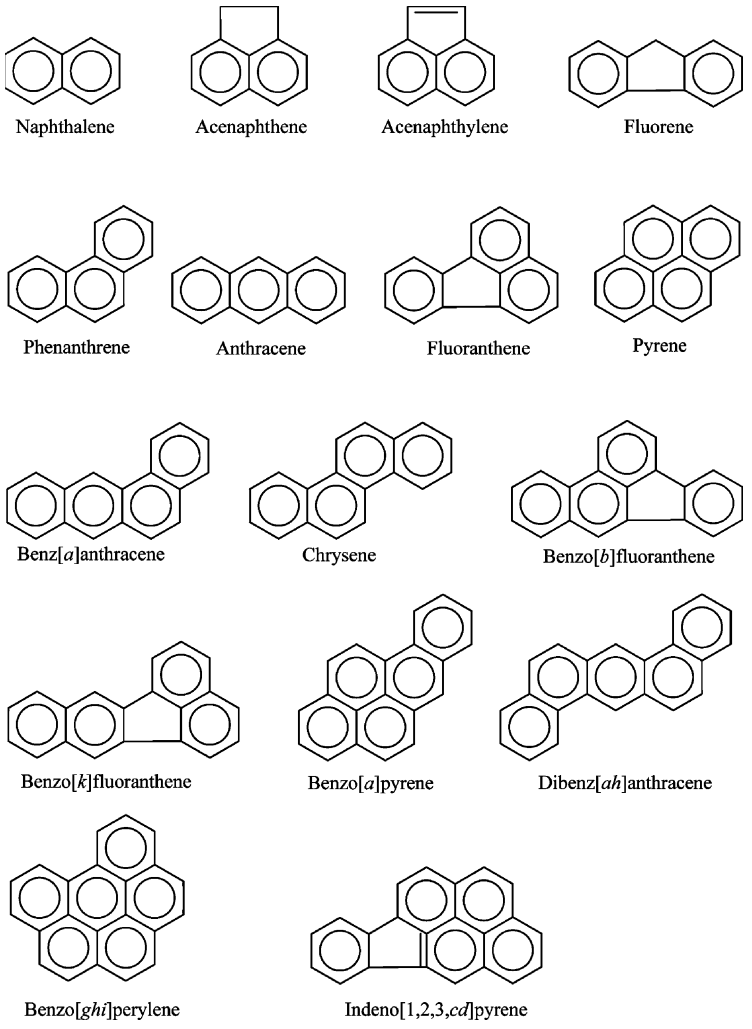


Fig. 2 Structures of PAHs on the EPA Priority Pollutants List.

Pollutants List is solely for research purposes. Currently, only a few of the PAHs generated during petroleum processing are used as raw material for other commercial applications. These include naphthalene, anthracene, phenanthrene, acenaphthene, fluorene, fluoranthene, and pyrene, with naphthalene and anthracene being the only two with large-scale commercial applications.

Naphthalene

Commercial uses

Naphthalene comprises two benzene rings and is the smallest PAH. The white, crystalline solid is also called mothballs, white tar, tar camphor, albocarbon, naphthaline, and naphthene. The most well-known commercial use of naphthalene in its pure form is as a household fumigant against moths. In 1989, naphthalene consumption as a moth repellent peaked at 12 million pounds. Since then, naphthalene has also been used as a solid block deodorizer for diaper pails and toilets, and as an antiseptic, expectorant, and anthelminthic. Approximately 65% of naphthalene is converted into phthalic anhydride.[13] Phthalic anhydride is used as an intermediate in the production of phthalate plasticizers, resins, phthaleins, dyes, pharmaceuticals, and insect repellents. It is also used in the production of the insecticide carbaryl, synthetic leather tanning agents, and surfactants. The consumption of naphthalene for the production of phthalic anhydride is expected to increase due to continued increase in demand. Another 10–15% of naphthalene is utilized for the manufacture of celluloid and hydronaphthalenes (for lubricants).[14–16] Naphthalene has had documented use as an additive in motor fuels, aluminum smelting, and soil fumigants as well. Recently, naphthalene has also been a commercial ingredient for the production of plasticizers, sulfonic acid, surfactants, and aromatic polyesters, and in the synthesis of salicylic acid and anthraquinone.[17,18]

Table 1 Physicochemical properties of PAHs on the EPA Priority Pollutants List

PAH	Structural formula	MW	Sol. (μg/L)	MP (°C)	BP (°C)	VP (Torr)	H (atm m^3/mol)	Density (g/cm^3)	log K_{ow}
Naphthalene	$C_{10}H_8$	128	31,700	80	218	4.92×10^{-2}	4.27×10^{-4}	1.03	3.37
Acenaphthylene	$C_{12}H_8$	152	3,470	92	265	2.0×10^{-2}	1.9×10^{-4}	0.899	4.07
Acenaphthene	$C_{12}H_{10}$	154	3,930	96	279	2.9×10^{-2}	2.4×10^{-4}	1.189	4.33
Fluorene	$C_{13}H_{10}$	166	1,980	116	293	1.3×10^{-2}	2.35×10^{-7}	1.203	4.18
Anthracene	$C_{14}H_{10}$	178	73	216	340	1.96×10^{-4}	1.8×10^{-6}	1.24	4.45
Phenanthrene	$C_{14}H_{10}$	178	1,290	101	340	6.8×10^{-4}	2.8×10^{-4}	0.98	4.46
Fluoranthene	$C_{16}H_{10}$	201	260	111	—	6.0×10^{-6}	10.5×10^{-6}	1.252	5.33
Pyrene	$C_{16}H_{10}$	202	135	149	360	6.85×10^{-7}	8.9×10^{-7}	1.271	5.32
Benz[*a*]anthracene	$C_{18}H_{12}$	228	14	158	400	5.0×10^{-9}	5.8×10^{-6}	1.274	5.61
Chrysene	$C_{18}H_{12}$	228	2	255	448	6.3×10^{-7}	1.05×10^{-6}	1.274	5.61
Benzo[*b*]fluoranthene	$C_{20}H_{12}$	252	1.2	167	—	5.0×10^{-7}	1.22×10^{-5}	—	6.57
Benzo[*k*]fluoranthene	$C_{20}H_{12}$	252	0.55	217	480	5.0×10^{-7}	3.87×10^{-5}	—	6.84
Benzo[*a*]pyrene	$C_{20}H_{12}$	252	3.8	179	496	5.0×10^{-7}	5.53×10^{-7}	1.35	6.04
Dibenz[*a*,*h*]anthracene	$C_{22}H_{14}$	278	2.49	262	—	1.0×10^{-10}	7.3×10^{-8}	1.282	5.97
Benzo[*g*,*h*,*i*]perylene	$C_{22}H_{12}$	276	0.26	222	—	1.0×10^{-10}	1.44×10^{-7}	—	7.23
Indeno[1,2,3,*c*,*d*]pyrene	$C_{22}H_{12}$	276	6.2	162	—	1.0×10^{-10}	2.96×10^{-20}	2.09	7.66

MW, molecular weight; Sol., solubility at 25°C; MP, melting point; BP boiling point; VP, vapor pressure at 20°C; H, Henry's constant; log K_{ow}, logarithm (to base 10) of partitioning coefficient.

Table 2 Anthropogenic activities associated with PAH production

Gasification/liquefaction of fossil fuels
Combustion of fossil fuels
Coke production
Catalytic cracking
Carbon black production and use
Asphalt production and use
Coal tar/coal tar pitch production and disposal
Refining/distillation of crude oil and crude-oil-derived products
Wood treatment processes
Wood preservative (e.g., creosote/anthracene oil) production
Fuel/oil storage, transportation, processing, use, and disposal
Landfills/waste dumps
Open burning of tires, refuse, coal, etc.
Incineration processes

Current production

Naphthalene is typically formed as a byproduct during coal gasification and liquefaction operations. By definition, coal gasification is the moderate- to high-temperature reaction of coal, air, steam, carbon dioxide, hydrogen, or a mixture of the components to generate a gaseous product.[19] The gaseous product is then used as either an energy source or a raw material for the subsequent production of industrial chemical or other liquid fuels. Depending on the coal source and temperature used, the resulting gaseous product can have a low, medium, or high BTU value. As such, the general schematic of an MGP is an example of an early coal gasification system. Additional information regarding gasification and liquefaction can be found in Speight's[10] and Perry and Green's[19] handbooks, as well as in the corresponding encyclopedia entries.

Petroleum gasification, liquefaction, and cracking processes generate byproducts that contain a large quantity of alkylnaphthalenes. These gaseous byproducts can be thermally or catalytically dealkylated to form naphthalene.[13,14] One approach is dealkylation of the gaseous stream in the presence of hydrogen. The bottoms of the dealkylator are sent to a series of four fractionators. Concentrated naphthalene is recovered from the bottom (400–450°F) of the second fractionator. The remaining two fractionators are used in conjunction with crystallizers to obtain purified fluorene and acenaphthene.

Acenaphthene and Acenaphthylene

Acenaphthene is produced by passing ethylene and benzene or naphthalene through a red-hot tube or by heating tetrahydroacenaphthene with sulfur to 180°C. Acenaphthene can also be obtained by high-temperature recovery of the distillate from concentrated tar. Once purified, acenaphthene is used as a dye intermediate, in the production of pharmaceuticals and plastics, and as a fungicide and insecticide. Acenaphthylene is typically manufactured by catalytic degradation of acenaphthene.

Fluorene and Fluoranthene

Fluorene is used in the formation of polyradicals for resins. It is also used as a raw material in the manufacture of dyes. In addition, it is used as a lining material to protect the interior of steel and ductile-iron drinking water pipes and storage tanks.

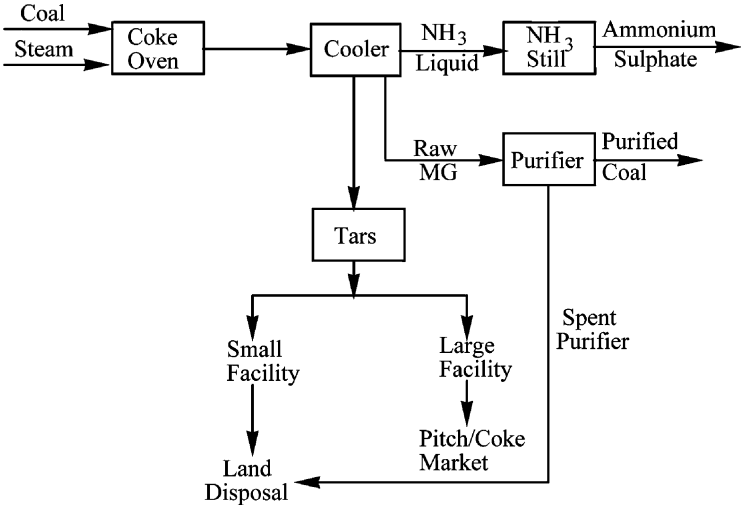


Fig. 3 Schematic of a generic MGP.

Anthracene and Phenanthrene

Anthracene and phenanthrene are stereoisomers that are crystals in pure form. Anthracene is a pale yellow crystal, while phenanthrene exhibits a yellow to brown hue. Besides its common name, anthracene is referred to as anthracin, green oil, or paranaphthalene. The compound is commercially produced by recovery from the coal tar distillation fraction known as anthracene oil or green oil. Anthracene is the key ingredient in the production of anthraquinone. However, it and phenanthrene are also used for the manufacture of dyes, fibers, plastics, and wood preservatives.[16,20] Phenanthrene, also known as phenanthrin, can be produced by high-temperature fractional distillation of coal tar oil. It is additionally used for the oxidation of diphenic acid for use in polymers,[13] as well as the production of chemical softeners, explosives, and some pharmaceuticals.[14,21,22] Recent research has extended the application of both isomers to scintillation counters, semiconductors, and photoconductors.[13,23]

REGULATIONS

As with most xenobiotics, PAHs are subject to a wide variety of environmental and occupational safety regulations. Due to space constraints, only a brief listing of the key regulations/codes is provided. A complete listing of these and other regulations can be found in the Code of Federal Regulations (CFR), standard regulation textbooks, and state and federal EPA homepages.

Current advances in epidemiological and toxicology studies have documented the potential harm that PAHs pose to human health. Therefore, the National Institute for Occupational Safety and Health (NIOSH) and the Occupational Safety and Health Administration (OSHA) have published exposure standards for individuals working in facilities that use or handle PAHs. For example:

- NIOSH: Recommends a workplace standard for coal tar products of $0.1\,mg/m^3$ as a 10-hr time-weighted average (TWA). Naphthalene has an exposure standard of 10 ppm or $50\,mg/m^3$ TWA.
- OSHA: Limits occupational exposure to coal tar pitch volatiles to $\leq 0.2\,mg/m^3$ as an 8-hr TWA. Coke oven emissions have a permissible exposure limit of less than $0.15\,mg/m^3$ as an 8-hr TWA.

Both NIOSH and OSHA have stricter exposure levels for individuals who conduct risk assessment studies at hazardous waste sites or work at incinerators. The exact exposure levels are PAH and site specific, and can be found in NIOSH and OSHA guides.

The Resource Conservation and Recovery Act (RCRA), Clean Water Act (CWA), and Clean Air Acts (CAAs) promulgate regulations to prevent future contamination by hazardous chemicals. Conversely, the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) and the Superfund Amendments and Reauthorization Act (SARA) are federal statutes that outline how existing/old contamination will be dealt with.

- RCRA: Hazardous waste disposal rules enable the EPA to regulate the manufacture, distribution, and final disposal of PAHs. It tracks each compound from cradle to grave.
- CWA: Contains guidelines of acceptable PAH levels in ground and drinking water.
- CAA amendment of 1990: Regulates air emissions containing PAHs. Specific industries impacted by this regulation include power plants, petroleum processing facilities, and steel manufacturing plants.
- CERCLA: Provides authority to enforce liabilities for the release of PAHs; established a fund for the clean up of existing sites; authorized EPA to initiate remedial actions; sets threshold constraints; and provides guidelines for acceptable remediation technologies.
- SARA: EPA includes 16 PAHs on a list of priority hazardous chemicals.

Drinking Water Standards and Health Advisories

The USEPA has established maximum contaminant levels (MCLs) for public water supplies to reduce the chance of adverse health effects from contaminated drinking water. Maximum contaminant levels are enforceable limits that public water supplies must meet and are lower than concentrations at which health effects have been observed. The only PAH with an established MCL is benz(a) pyrene, which is regulated at 0.2 parts per billion.

The USEPA[24] has, however, established health advisories (HAs) for several PAHs. An HA is defined as an estimate of acceptable drinking water levels for a chemical substance based on health effects information. An HA is not a legally enforceable federal standard but serves as technical guidance to assist federal, state, and local officials. Table 3 contains a partial listing of PAH HAs.

FATE AND TRANSPORT

This section includes an overview of the fate and transport of PAHs in air, water, and subsurface environments. Additional, more detailed information regarding PAH fate and transport can be found in publications by Hemond and Fechner-Levy,[25] Wilcke,[26] and Suess.[27]

Table 3 U.S. EPA health advisories for 16 PAHs (amounts for 10-kg child)

PAH	One-day (mg/L)[a]	10-day (mg/L)[b]	RfD (mg/kg day)[c]	DWEL (mg/L)[d]	Lifetime (mg/L)[e]	10^{-4} Cancer risk (mg/L)[f]
Naphthalene	0.5	0.5	0.02	0.7	0.1	—
Acenaphthene	—	—	0.06	2	—	—
Acenaphthylene	—	—	—	—	—	—
Fluorene	—	—	0.04	1	—	—
Phenanthrene	—	—	—	—	—	—
Anthracene	—	—	0.3	10	—	—
Fluoranthene	—	—	—	—	—	—
Pyrene	—	—	0.03	—	—	—
Benz[*a*]anthracene	—	—	—	—	—	—
Chrysene	—	—	—	—	—	—
Benzo[*b*]fluoranthene	—	—	—	—	—	—
Benzo[*k*]fluoranthene	—	—	—	—	—	—
Benzo[*a*]pyrene	—	—	—	—	—	0.002
Dibenz[*a*,*h*]anthracene	—	—	—	—	—	—
Benzo[*g*,*h*,*i*]perylene	—	—	—	—	—	—
Indeno[1,2,3,*c*,*d*]pyrene	—	—	—	—	—	—

[a]The concentration of a chemical in drinking water that is not expected to cause any adverse noncarcinogenic effects for up to one day of exposure.
[b]The concentration of a chemical in drinking water that is not expected to cause any adverse noncarcinogenic effects for up to 10 days of exposure.
[c]Reference dose. An estimate of a daily oral exposure to the human population that is likely to be without an appreciable risk of deleterious effects during a lifetime.
[d]Drinking water equivalent level. A lifetime exposure concentration protective of adverse, noncarcinogenic health effects that assumes that all of the exposure to a contaminant is from drinking water.
[e]The concentration of a chemical in drinking water that is not expected to cause any adverse noncarcinogenic effects for a lifetime of exposure.
[f]The concentration of a chemical in drinking water corresponding to an estimated lifetime cancer risk of 1 in 10,000.

Air

PAHs are present in the air as vapors and adsorbed to airborne particulate matters. In general, high molecular weight PAHs with at least five rings are present in airborne particulate matter, whereas low molecular weight PAHs with two to three rings are more abundant in the gas phase. This trend is dependent on each compound's specific gas-to-particle partition coefficients. The physical mechanism of PAH loss from atmosphere is dry and wet deposition, i.e., gravitational settling and scavenging by precipitation/water vapor. Conversely, photochemical transformations are considered the most dominant fate mechanism. For instance, PAHs have been found to react in the presence of light with atmospheric ozone, NO_x, SO_x, and OH radicals.

Seawater and Freshwater

PAHs are transported to aquatic environments via direct atmospheric deposition, runoff from PAH-contaminated land, and direct release into water. Approximately 70% of total benz(*a*) pyrene input contributes to atmospheric deposition to the marine environment, whereas over 75% of total river PAH is from runoffs from highways. The fate of PAHs in the aquatic environment is highly dependent on sorptive/desorptive phenomena as well as the suspended matter load. Low molecular weight PAHs can be removed from open waters by volatilization. Physicochemical interactions (i.e., sorption, desorption, biodegradation, etc.) with bottom sediments are a major pathway by which PAHs are removed from waters.

Sediments

Sediments act as a sink for PAHs entering aquatic ecosystems. However, they can also serve as a source of PAHs to aquatic organisms. In general, transport of PAHs to and within sediments is controlled by their hydrophobicity and particle sedimentation. In shallow waters and estuaries, PAHs present in surface sediments are subjected to biodegradation and photo-oxidation, whereas PAH degradation in buried sediments happens at a very slow rate, if at all it occurs.

Terrestrial Environments

The key factors controlling the fate and transport of PAHs in subsurface environments are sorption and desorption phenomena. The extent of these phenomena is known to depend on the type of the PAH, the sorbent characteristics, PAH aging period, and the presence of cosolute.[28] In general, as the PAH molecular weight and the amount of soil organic matter and clay minerals increase, sorption becomes stronger. Longer contact times (i.e., aging) between PAHs and sorbents result in further diffusion into soil components, physical entrapment in soil micropores, and the formation of strong bonds between PAHs and soil constituents.

REMEDIATION OF PAH-CONTAMINATED MEDIA

This section provides a brief overview of the PAH remediation technologies available. For additional information on each process mentioned in this section, readers are encouraged to refer to other encyclopedia entries as well as established texts.[10,29]

Thermal Decomposition

Incineration is technically the most effective technique in PAH remediation. However, it is cost prohibitive due to associated high excavation, transportation, and heating costs. Dealing with NIMBYs (Not In My Back Yard) and BANANAs (Build Absolutely Nothing Anywhere Near Anything) are major constraints to the use of this technology. Furthermore, future amendments to the Clean Air Act are expected to limit incineration applications to ''last resort'' technology.

Physicochemical Treatment

Volatile and soluble PAHs can be easily extracted with air or water from the soil phase, whereas resistant, high molecular weight PAHs are extracted with the help of media that enhance their extractability. Once PAHs are extracted, they are then treated with chemical processes such as oxidation and activated carbon adsorption.

Bioremediation

Both indigenous and exogenous micro-organisms have been involved in the success of microbial remediation of PAHs. Landfarming and composting are common ex situ methods of PAH bioremediation. The former uses a top layer of agricultural soil with some nutrient amendment to stimulate microbial activity. The latter is achieved by the addition of bulking agents (e.g., sawdust or woodcuttings) and regular mixing of compost piles.

Bioreactors are another ex situ method by which indigenous or exogenous, sometimes genetically modified, micro-organisms are inoculated. In general, high molecular weight PAHs persist because of: 1) low bioavailability (i.e., strong sorption affinity); 2) lack of the proper enzyme for indigenous micro-organisms to degrade them; 3) steric hindrance to enzyme attack; and 4) toxicity to micro-organisms.[30]

Phytoremediation can be placed in the bioremediation category as either an in situ or ex situ remediation technology, depending on the approach used. It uses plants and associated micro-organisms to extract, degrade, and stabilize PAHs. If the plants are used in place, the method is in situ. If the soil is excavated and taken to a greenhouse, then the method is classified as ex situ.

CONCLUSIONS

As outlined above, the formation of PAHs as a product of incomplete combustion has occurred since humans first used combustible material for cooking and heating. Production peaked during the mid- to late 1800s. At that time, PAHs were improperly disposed of, leaving society with a vast number of contaminated sites. Current advances in remediation technologies are making great strides in cleaning up these sites. The advent of environmental regulations and implementation of waste minimization practices has led to further decrease in PAH releases during manufacturing operations. Currently, only a few PAHs, predominantly naphthalene and anthracene, are used as raw materials in manufacturing processes.

Due to the adverse effects of PAHs on humans and other natural recipients, the fate and transport of PAHs have been increasingly investigated. Photochemical degradation, wet/dry deposition, sorption and desorption, and biodegradation are known to be the governing mechanisms for the fate and transport of PAHs. The dominant mechanism depends on the specific PAH–medium combination present. Strategies for PAH remediation depend on a large numbers of factors, including, but not limited to, the extent of contamination, long- and short-term effectiveness and performance, compliance with applicable requirements, implementability, cost, and acceptance by federal, state, and local communities.

ARTICLES OF FURTHER INTEREST

Advanced Oxidation, p. 41.
Bioremediation, p. 207.

REFERENCES

1. Harvey, R.G., Ed. *Polycyclic Aromatic Hydrocarbons: Chemistry and Carcinogenicity*; Cambridge University Press: New York, NY, 1991.
2. Loening, K.; Merrit, J.; Later, D.; Wright, W. *Polynuclear Aromatic Hydrocarbon Nomenclature Guide*; Battelle Press: Columbus, OH, 1990.
3. Bhattacharhi, A.; Freidman, G.M., Eds.; *Lecture Notes in Earth Sciences: Contaminated Sediment*; Springer-Verlag: New York, NY, 1989.
4. Blumer, M. Benzpyrenes in soil. Science **1961**, *134*, 474–475.
5. Blumer, M. Polycyclic aromatic compounds in nature. Sci. Am. **1976**, *234* (3), 35–45.
6. Borneff, F.; Selenka, H.; Kunte, H.; Maximos, A. Experimental studies on the formation of polycyclic aromatic hydrocarbons in plants. Env. Res. **1968**, *2*, 22–29.
7. Graf, W.; Diehl, H. Concerning the naturally caused normal level of carcinogenic aromatics and its cause. Arch. Hyg. **1966**, *150* (4), 49–59.
8. Yong, R.N.; Tousignant, L.P.; Leduc, R.; Chan, E.C.S. Disappearance of PAHs in a contaminated soil from Mascouche, Quebec. In *In-Situ Bioreclamation: Applications and Investigations for Hydrocarbon and Contaminated Site Remediation*; Hinchee, R.E., Olfenbuttel, R.E., Eds.; Butterworth-Heinemann Publishing: Boston, MA, 1991; 377–395.
9. Srivastava, V.J.; Kilbane, J.J.; Kelley, R.L.; Atkin, C.; Hayes, T.D. Biodegradation of old town gas sites. In *Gas, Oil and Coal Biotechnology*; Atkins, C., Smith, J., Eds.; Institute of Gas Technology: Chicago, IL, 1990.
10. Speight, J.G. *The Chemistry and Technology of Petroleum*, 2nd Ed.; Marcel Dekker, Inc.: New York, NY, 1991.
11. Bouchez, M.; Blanchet, D.; Vandecasteele, J.P. Degradation of PAHs by pure strains and by defined strain associations: inhibition phenomena and cometabolism. Appl. Microbiol. Biotechnol. **1995**, *43*, 156–164.
12. Cerniglia, C.E. Biodegradation of polycyclic aromatic hydrocarbons. Biodegradation **1992**, *3*, 351–368.
13. Gutman, I.; Cyvin, S.J. *Introduction to the Theory of Benzoid Hydrocarbons*; Springer-Verlag: New York, NY, 1989.
14. Kent, J.A., Ed.; *Riegel's Handbook of Industrial Chemistry*, 8th Ed.; Van Nostrand Publishing: New York, NY, 1983.
15. Okuno, M.; Takahashi, T. Oxidation Process and Layered Catalyst Containing Reactors for the Manufacture of Phthalic Anhydride from *ortho*-Xylene and/or Naphthalene European Patent 1120222000.
16. Sittig, M. *Aromatic Hydrocarbons: Manufacture and Technology*; Noyes Data Corp.: Ridge Park, NJ, 1976.
17. Baraeva, S.F.; Bryuske, S.I.; Matushkina, A.L.; Koloskova, V.N. Preparation and Isolation of 2-Amino-5-hydroxy-7-napthalenesulfonic Acid Russian Patent 19900315, 1990.
18. Mizunuma, T.; Iizuka, M.; Izumi, K. Industrial applications of surfactants derived from naphthalene. Spec. Pub. R. Soc. Chem. **1990**, *22* (2), 101–113.
19. Perry, R.H.; Green, D. *Perry's Chemical Engineering Handbook*, 6th Ed.; McGraw-Hill: New York, NY, 1983.
20. Nakamura, Y.; Kosho, N Manufacture of Anthracene-Type Polyesters with High Viscosity. Japanese Patent 19910325, 1991.
21. Lee, C.M.; Satterfield, C.N. Effect of ammonia on the hydrogenation of phenanthrene during the hydrodenitrogenation of quinoline. Energy Fuels **1993**, *7* (6), 978–980.
22. Wisziniowski, K. Utilization of phenanthrene in the industry. Chemik **1966**, *18* (8–9), 306–316.
23. Yada, S.; Kinoshita, H.; Takehara, N. Manufacture of Anthracene Polymer Semiconductor Materials Japanese Patent 1980703, 1989.
24. U.S. Environmental Protection Agency. In *Drinking Water Standards and Health Advisories*; EPA 822-B-00-001; Office of Water, 2000.
25. Hemond, H.F.; Fechner-Levy, E.J. *Chemical Fate and Transport in the Environment*; Academic Press: San Diego, CA, 2000.
26. Wilcke, W. Polycyclic aromatic hydrocarbons (PAHs) in soil—a review. J. Plant Nutr. Soil Sci. **2000**, *163*, 229–248.
27. Suess, M.J. The environmental load and cycle of polycyclic aromatic hydrocarbons. Science Tot. Env. **1976**, *6*, 239–250.
28. Hwang, S.; Cutright, T.J. Biodegradability of aged pyrene and phenanthrene in a natural soil. Chemosphere **2002**, *47*, 891–899.
29. LaGrega, M.D.; Buckingham, P.L.; Evans, J.C. *Hazardous Waste Management*, 2nd Ed.; McGraw-Hill: New York, NY, 2001.
30. Salanitro, J.P.; Dorn, P.B.; Huesemann, M.H.; Moore, K.O.; Rhodes, I.A.; Jackson, L.M.R.; Vipond, T.E.; Western, M.M.; Wianiewski, H.L. Crude oil hydrocarbon bioremediation and soil ecotoxicity assessment. Env. Sci. Technol. **1997**, *31*, 1769–1776.

Polymer Clay Nanocomposites

Hyoung J. Choi
Department of Polymer Science and Engineering, Inha University, Incheon, Korea

Ji W. Kim
Myung S. Jhon
Department of Chemical Engineering and Data Storage Systems Center, Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A.

INTRODUCTION

Material properties can be tuned systematically, by various methods such as blending and mixing, to control the synergy among pure materials, and thereby achieve desirable properties for engineering applications. Hence, particles, powders, and fibers are dispersed in pristine matrix materials to form ''composite'' or ''hybrid'' materials. Over the years, research on composite materials has spanned a broad range of scientific and technological areas, including polymer matrix composites, glass–ceramics, metal matrix composites, carbon–carbon composites, ceramic matrix composites, and intermetallic matrix composites. In composites, the major component controls the primary properties, although a minor component could significantly modify or enhance the composite's properties as well. Recently, a number of engineering fields have begun to adopt the new paradigm of nanoscience and technology for a variety of applications. This paradigm shift to ''nanocomposites'' has promoted the development of extraordinary properties by overcoming the inherent limitations of pure materials. Polymer composites, especially using the layered silicate clays, have recently attracted the attention of researchers, and are the focus of this review.

HISTORY

The term ''polymer clay nanocomposite (PCN)'' refers to a material composed of two-phase materials, where one phase (clay) is dispersed in the second phase (polymer matrix) at a nanometer level.[1,2] Composites exhibiting structural and compositional changes at the molecular scale have demonstrated several physical property enhancements that are otherwise unavailable in conventional composites. Of these, layered silicates have proven themselves vital as a reinforcing agent when dispersed into engineering plastics.[3] Nanolayers, however, are extremely difficult to disperse in polymer matrices because of their tendency for face-to-face stacking and inherent hydrophilic properties. Understanding the nature of the clay itself is important for being able to control the polymer–clay surface interactions. Organoclays dispersed in polymer matrices have shown profound effects, such as improved dimensionality, barrier properties, as well as flame retardation. Such benefits have led to the use of these composites in a broad range of applications, including the automotive and package industries,[4] as well as for commodity polymers: epoxies, polyurethane, polyimide, nitrile rubber, polyester, polypropylene, polystyrene, and polysiloxanes.[5] Polymer clay nanocomposites are categorized into three groups, according to the dispersion state of the clay in the polymer matrices, as shown in Fig. 1. Notice that Ray and Okamoto[1] recently classified the different state of PCN; e.g., intercalated, intercalated-and-flocculated, and flocculated.

PCNs' advantage over pristine systems is that they are lighter than conventional polymer/inorganic blends (because of a lower amount of heavy materials, i.e., clay), yet exhibits similar mechanical strengths. Further, PCNs demonstrate outstanding gas permeability resistances, without the special molecular design of polymeric multilayers. The advantages of PCNs extend to a broad range of applications (beyond conventional usages) including optics for electronic materials, high flammability resistance, and electroactive materials, as well as environmental recycling.[6] Polymer clay nanocomposites are also used as a ''model system'' for investigating the nanoscale confined polymer structures and dynamics.[7] The high molecular weight polymer dynamics in nanoscale restricted spaces have been extensively studied via conventional wisdoms, i.e., polymer dynamics confined to a highly entangled network, as well as static and dynamic conformation via computer simulations. Intercalation and exfoliation structures result in drastically different physicochemical responses than the conventional blends.

First, the history and definition for nanomaterials, the clay materials, and the preparation methods are introduced. Second, the characterization tools including the

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120006814

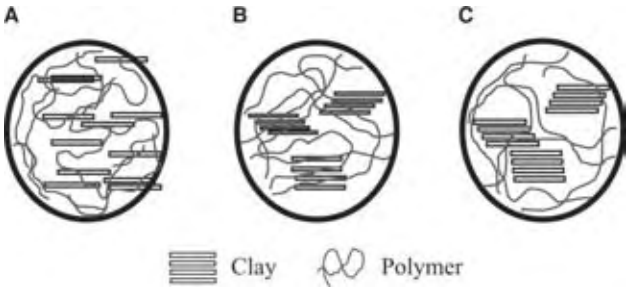


Fig. 1 The states of clay dispersion in polymer matrices: (A) exfoliated or delaminated nanocomposites; (B) intercalated nanocomposites; and (C) conventional blends.

X-ray diffraction (XRD), transmission electron microscope (TEM), and rheological measurements are discussed. Third, the static and dynamic structures of confined polymer in the gallery are examined. Finally, the commercial applications for the PCNs are given.

Preparation Methods

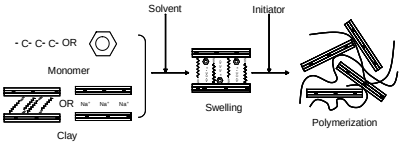
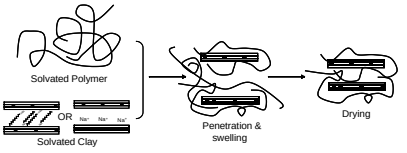
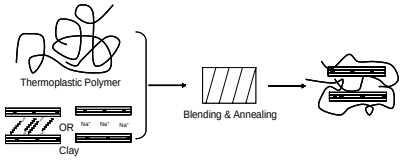
Polymer clay nanocomposite preparation methods are typically categorized as either solution or solution free (''melt''). The rule of thumb in preparation is: regardless of what is inserted into the clay layer, the layer space must be expanded first. While there are no significant differences in the preparation of different PCNs, concerns arise from the use of a solvent or interaction modification between the clay and polymer (or monomer).

The very first method used to synthesize PCNs was ''in situ polymerization'' [Table 1(I)], and it still remains as the most popular process for synthesizing thermoset PCNs. As usual, the organoclay is swollen first. The surface treatment of the organoclay and the swelling temperature are determined, depending on the polarity of the monomers. The polymerization location is a key control factor after the reaction is initiated. If the curing kinetics between the layers is dominant than outside the layers (i.e., intragallery polymerization is favored over extragallery one), the delamination of the clay is occurred. The driving force behind in situ polymerization is the polarity of the monomers. During the swelling stage, the high surface energy of the clay attracts polar monomers, so that they diffuse between the clay layers [see Table 1(I)]. At thermodynamic equilibrium, the diffusion terminates and the clay is swollen in the orientation, perpendicular to the alkylammonium ions.[8] When the polymerization is initiated, the monomers start to react with the curing agent. This reaction lowers the overall polarity of the intercalated molecules and additional polar molecules are forced into the clay layers. As a result, the organic molecules will eventually delaminate the clay.

The solution technique [Table 1(II)] is similar to that of the in situ process. Here, the organoclay is dispersed first in a toluene or *N*,*N*-dimethylformamide. The alkylammonium ions treated clays will then considerably swell to form gel structures. Later, polymer is added into the solution, which intercalates between the clay layers. Finally, the solvent is removed via evaporation (usually under a vacuum).

On the other hand, the solution free (or melt) intercalation process [Table 1(III)] was first introduced

Table 1 Advantages and disadvantages of synthesis methods for PCNs

Processes	Schematics	Advantages	Disadvantages	References
Solution				
(I) In situ		Polarity of the monomers		Epoxy and polyurethane[26,27]
(II) Solution		Low or no polarity	Require large quantity of solvent	Rubber and polypeptide[29,30]
(III) Solution free (or melt)		No solvent, mass production	Limited systems	Nylon-6, lactide, and polystyrene[3,25,28]

by Vaia, Ishii, and Giannelis.[3] This process consists of blending a molten thermoplastic with an organoclay to control the polymer–clay interactions. The mixture is then annealed above the glass transition temperature of the polymer to form a nanocomposite. One may notice that the entire polymer chains transporting between the clay layers is somewhat unexpected, as the radius of gyration of the polymer is an order of a magnitude larger than the interlamellar spacing.[9] As shown in Table 1(III), the polymer chains experience a drastic loss in configurational entropy during the intercalation process. The impetus behind this mechanism is the significant enthalpic contribution of the polymer/organoclay during the blending and annealing steps. This process has become increasingly popular because of its mass production capability. Creating plasticity without a solvent is difficult, and therefore, a limitation exists for the melt intercalation process.

Emulsion polymerization intercalation[10] is an alternative method of PCN synthesis. This method is similar to in situ polymerization, but has several differences and limitations. In the emulsion polymerization method, micelle helps the polymer to be monodisperse, yet the micelles are limited in their ability to penetrate into the layer. Nonetheless, this method is attractive because of its use of a micelle as a targeted material, which helps the monomer or prepolymer to penetrate into the silicate layers.

Clay Materials

While both polymer and clay are primary PCN components, we focus mainly on clay as it plays a subtle role in the formation of PCN.

The term ''layered silicates'' refers to natural clays or altered silicates, such as mica, laponite, and fluorohectorite. Because of its desirable layer charge density, montmorillonite (MMT), which was discovered by Damour and Salvetat (in 1847 in Montmorillon),[11] has become a widely used nanofiller clay nowadays. The ideal crystallographic MMT structure derived from the pyrophyllite consists of two fused silica (Si) tetrahedral sheets sandwiching an edge-shared octahedral sheet of either aluminum (Al) or magnesium (Mg) hydroxide. The sum of the single layer thickness and the interlayer spacing (the ''*d*-spacing'' or ''basal spacing'') of 9.6 Å represent the repeated unit of the multilayer material. Typical crystalline structures are characterized by XRD (refer to the following section). The layers can be considered as large aspect ratio lamellae (100–200 nm in length and 1 μm in diameter). Large and irregular aggregates (0.1–10 μm in diameter) are formed with five to ten lamellae associated by the interlayer ions in the primary particles (8–10 nm in transverse direction). These aggregates provide the structural basis for the clay. Depending on the silicates, the lateral dimensions of the layers vary from 30 nm to several microns. The van der Waals gaps, called ''interlayers'' or ''galleries,'' result from layer stacking.

The replacement of inorganic exchange cations mirror surfaces and polymer polarities expand the clay galleries. Depending on the clay charge density and the alkyl ion chains, different arrangements of the alkyl ion chains become possible. The alkyl ion chain is a surfactant, which could modify the surface property of clay layer. In general, the clay layers will be further apart, as the surfactant chain length becomes longer and the clay charge density becomes higher. This is expected as both these parameters modify the volume occupied by the intragallery surfactant. Depending on the clay charge density, the alkyl ion chains may lie parallel to the clay surface as a monolayer, a lateral bilayer, a pseudo-trilayer, or as an inclined paraffin structure (Fig. 2).

Large surfactant ions may adopt lipid bilayer orientations in the galleries at very high charge densities. The orientations of alkyl ion chains are observed from the XRD measurements. Molecular dynamics (MD) simulation also provide further insight into these orientations by observing a nano/macroscopic properties, including density, normal forces, chain configurations, and trans-gauche ratios. It has been found that a disordered liquid-like arrangement of chains is preferable for a monolayer, a lateral bilayer, and a pseudo-trilayer, with respective *d*-spacings of 1.32, 1.80, and 2.27 nm.[5] As expected, the alkyl ion chain head group, relative to the aliphatic portion, resides near the silicate surface. For the fully extended surfactant chain, the expected conformation is the trans (over gauche) just prior to the system progressing to the next highest layering pattern. This is expected, as the alkyl ion chains must be optimally packed under such dense surfactant concentrations.[5]

CHARACTERIZATION OF POLYMER/CLAY NANOCOMPOSITES

X-Ray Diffraction

Many standard methods exist for characterizing PCN structure. One method is XRD because of its ability to accurately evaluate the spacing between layers. Further, sample preparation is relatively easy and the X-ray analysis can be performed within a few hours. However, interpretation must be done carefully, as the lack of sensitivity in the analysis itself and equipment limitations can quickly lead to incorrect findings.

Due to experimental set-up errors (e.g., misalignments, counting limitations), the XRD analysis is complicated by the factors that the XRD analysis must

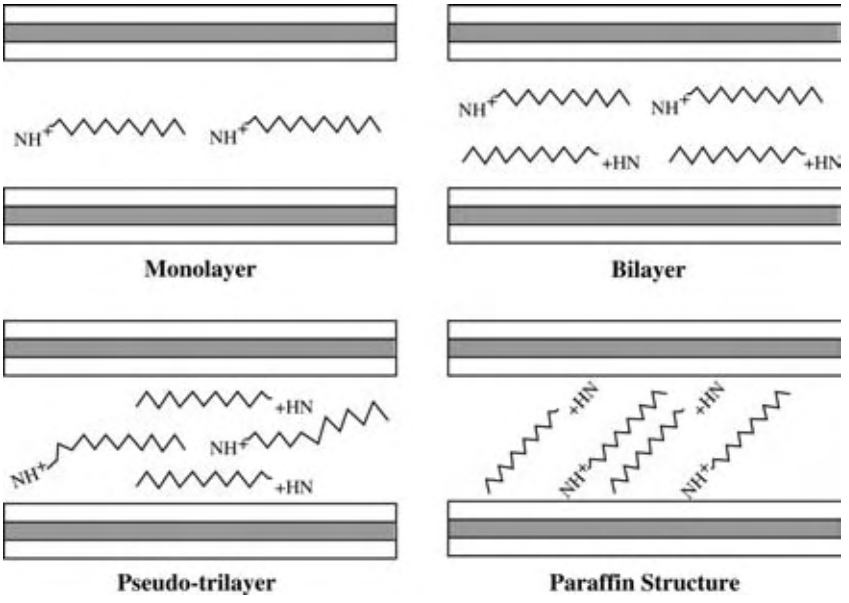


Fig. 2 Orientations of alkyl ion chains in the galleries of layered silicates with different clay charge densities.

be sensitive enough to detect the crystalline structure of the sample containing a small amount of clay, typically less than 10 wt%, otherwise, peaks in the diffraction pattern may not be observed, leading to the false diagnosis that a delaminated structure exists. Therefore, it is important to determine whether the crystallographic planes belonging to the clay layers are detectable by the XRD analysis, which is sensitive enough to detect (0 0 1) reflections. The XRD analysis should be performed at a low angle (diffraction angle $2\theta < 9^\circ$, i.e., $d > 9.8$ Å) to detect the (0 0 1) reflection to evaluate the d-spacing. Thus, the radiated surface may include the sample, as well as the sample holder, creating a large amount of noise and, in turn, complicate the interpretation of the XRD patterns. The X-ray penetration depth is inversely proportional to 2θ, implying that an analysis at a small θ will only reveal the structure presented in a thin layer near the surface (typically 0.1 μm for polymers). As a result, thin samples with large surface areas are recommended. The typical change in the (0 0 1) peak position for increased clay ratio in biodegradable polymer (BDP) matrices is shown in Fig. 3A.[12] Although a difference exists between pure polymer and PCNs, the change in the (0 0 1) peak position is not significant. Traditionally, the molecular structure of the PCNs has been elucidated using XRD and TEM in conjunction. Because of the periodic arrangement of the silicate layers, both in the pristine and intercalated structures, with periodicity of 1–4 nm and the presence of high atomic number species in the layers, the XRD method for determining the interlayer spacing becomes sufficient. However, in the absence of layer bundles, as for exfoliated or delaminated nanocomposites (as well as for disordered nanocomposites), XRD alone is not sufficient to provide definitive structural information.

TEM Analysis

The TEM is a complementary characterization tool to XRD. TEM produces a direct measurement of the spatial distribution of the layers, yet requires substantial skill in specimen preparation and analysis. TEM is a powerful technique for studying structures at the nanometer length scale, and is useful for verifying data obtained by XRD. TEM enables the precise observation of nanostructures at an extremely high spatial resolution (~0.2 nm). Specimen preparation, however, constitutes a major difficulty in TEM analysis. The specimens must be sufficiently thin to optimize the so-called mass–thickness contrast, which requires thicknesses to be on the order of 100 nm. One way to prepare specimens from polymer samples is through ultramicrotoming with either a glass or diamond knife. The ultra thin specimens, floating on the water surface, are collected on grids and dried. The grids are then transferred directly to the TEM column for observation. If the objective aperture is centered near the optical axis, then a bright background is visible (in the absence of a specimen), which is known as bright field imaging. Regions of the specimen that are thicker, or of higher density, will scatter the electron beam more and, therefore, appear darker in the image. As the clay has a higher density than the polymer, the clay layers become darker than the polymer. If the specimen is too thick, the polymer appears to be darker, which reduces the contrast with the clay. In addition to describing the spatial correlations of the layered silicates, TEM has been an extremely powerful technique for examining the homogeneity of the mixing process. A bright field TEM image of an organically modified layered silicate, intercalated with BDP is

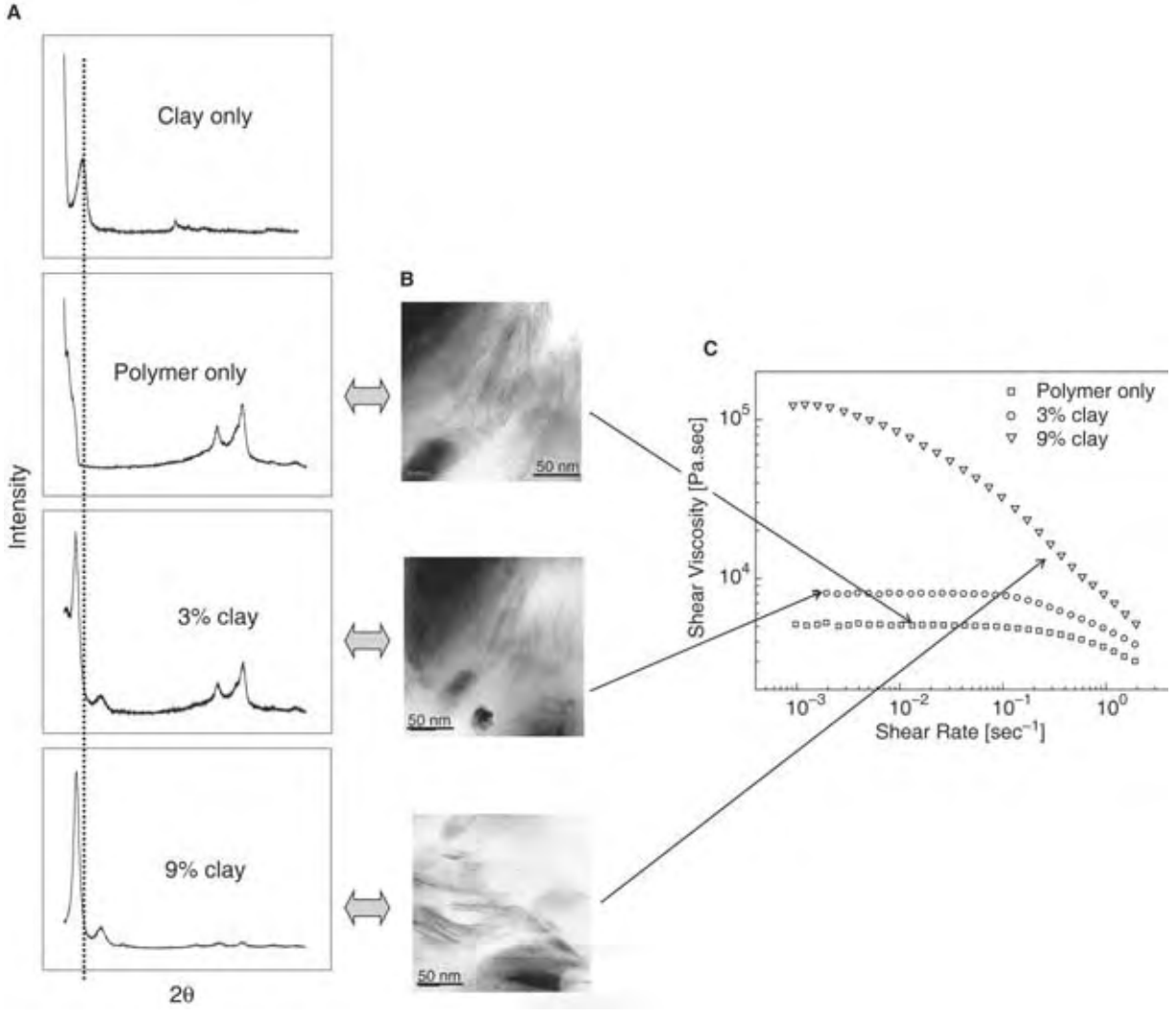


Fig. 3 Typical XRD patterns: (A) TEM images; (B) shear viscosity–shear rate relationship; and (C) BDP/clay nanocomposites. *(View this art in color at www.dekker.com.)*

shown in Fig. 3B. The periodically alternating dark and light bands, with a spacing of approximately 3 nm, represent the layers of silicate and the interlayers, respectively. The TEM also reveals the presence of individual crystallites consisting of several silicate layers, with bulk polymer filling the space between crystallites. Similar micrographs have also been obtained by the Toyota group[13] using monomer intercalation, followed by polymerization.

Rheological Measurement

Rheological measurement is another potential characterization method although an interpretation tool for the nanostructural information must be developed, before this becomes a viable characterization tool. Polymer clay nanocomposites are composed of two or more components; researchers studied their miscibility and dispersion. They investigated the rheological property change as a function of blend ratio between clay and polymer matrices. However, no significant impact from the PCN structure on the physical properties has been found. Despite this inconclusive observation, many attractions for the rheological characterization of PCNs, including processibility criteria, still exist.

The steady-shear rheological behaviors of a series of BDP/clay nanocomposites are shown in Fig. 3C.[12,14] The viscosity of the nanocomposites is enhanced considerably at low shear rates and increases monotonically with silicate loading (at a given shear rate). Furthermore, the intercalated nanocomposites display a shear-thinning behavior at low shear rates, whereas

the pure polymer exhibits Newtonian plateau in viscosity for a broad range of shear rates. At high shear rates (where the polymer displays a shear-thinning behavior), the nanocomposites also show shear-thinning characteristics. Similar trends are observed in Fig. 4 for the dynamic oscillatory measurements of BDP/clay nanocomposites,[12] where the slopes' transition to a flattened behavior is observed with an increase in the clay loading. The low-frequency response is indicative of solid-like behavior at high clay content. The slope and the absolute values of the modulus imply a supermolecular structural formation in the nanocomposites. In explanation of viscoelastic property of polymer, storage modulus G' represents the elasticity, and loss modulus G'' represent the fluidity, i.e., G' and G'' are able to represent the phase behavior of the PCN, such as solidlike and liquidlike. The melt behavior of BDP/clay nanocomposites is liquid-like at low frequencies ($G' < G''$), whereas solid-like behavior is found at high frequencies ($G' > G''$). Furthermore, the crossover frequency (ω_c) [defined as $G'(\omega_c) = G''(\omega_c)$] get smaller with the clay content. However, there exist a critical volume fraction of clay in the BDP/clay nanocomposites, where G' is always higher than G'', that is, the nanocomposites always behave like solid.

There are many other characterization methods (e.g., small-angle X-ray scattering, solid-state nuclear magnetic resonance, and Fourier-transformed infrared analysis) for investigating nanocomposite structure. These techniques are extensively reviewed in Ray and Okamoto.[1]

STATIC AND DYNAMIC POLYMER STRUCTURES

Organoclay–Polymer Interactions

If the polarity of the organoclay matches with the monomer or prepolymer, it intercalates into the galleries and pushes the clay layers out. Such behaviors have been observed for ε-caprolactam, epoxides, and polyethyleneoxide (PEO) that intercalate organoclay galleries as unreacted precursors. For long alkyl chain-exchanged organoclays, the swollen galleries show a d-spacing similar to a paraffin structure (Fig. 2). Even these ideally matched cases, however, do not necessarily form nanocomposites. The clay layers are only forced apart and no longer interacting through the alkyl ion chains. An ideal nanocomposite is formed upon polymerization (see Fig. 1A). The complete dispersion (or exfoliation) of the clay nanolayers yields composites with the highest degree of property enhancement. If the layers persist with a repeated stacking pattern with a gallery height less than two times the alkyl ion chains, then the final product will be an intercalated composite with regions of very high and low reinforced concentration (Fig. 1B). This nonuniform dispersion of nanolayers limits stress transfer throughout the composite, producing relatively less than optimal reinforcement. Although still poorer, conventionally scaled composites are formed if the clay and polymer are partially miscible (Fig. 1C). Here, the clay persists as tactoids of face-to-face stacking agglomerates throughout the polymer matrix. The poor dispersion of the reinforcing phase inhibits surface

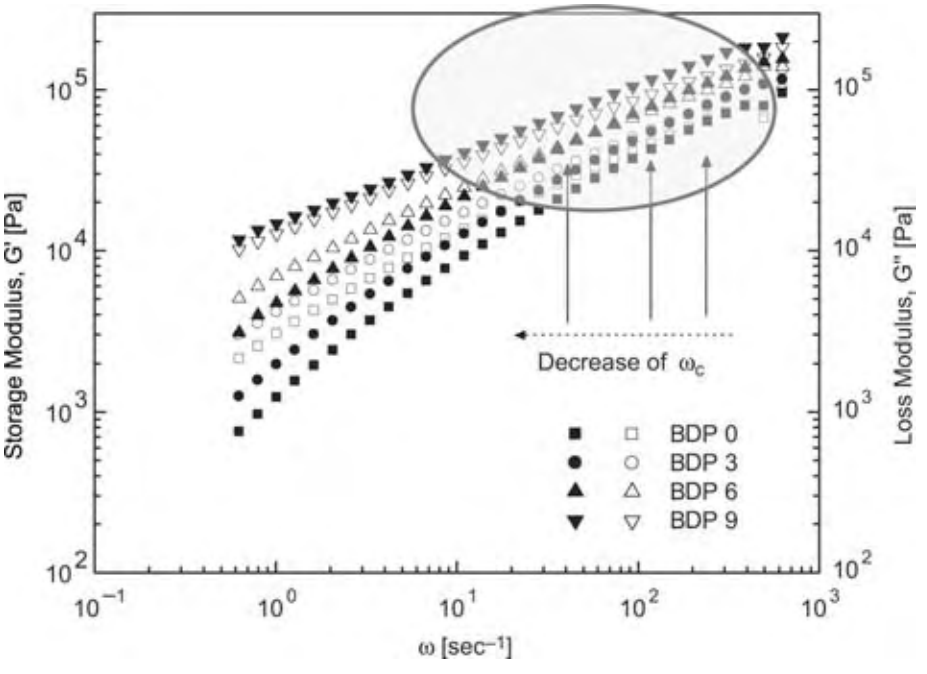


Fig. 4 Dynamic oscillatory test for BDP/clay nanocomposites (BDP series) [G' (filled), G' (open)]. The BDP 0, BDP 3, BDP 6, and BDP 9 indicate clay contents of 0, 3, 6, and 9%, respectively, in PCN. *(View this art in color at www.dekker.com.)*

contact between the polymers and clays, creating large regions of pure polymer. Therefore, it is imperative that the surface polarities of polymer and clay be matched, for the polymer to wet and intercalate the clay tactoids.

Nanoscopically Confined Polymer Dynamics

Polymer melts can intercalate as layered inorganic compounds without being assisted by shear or solvents. This is somewhat surprising, as the distances between the confined surfaces are comparable to the monomer size and are significantly smaller than the radius of gyration of the polymer. The reduction in free energy by the concentration gradient, during the intercalation, provides an ''enthalpic driving force'' for the polymer coils into the interlayer galleries. On the other hand, the configurational entropic force with chain stretching, in addition to other topological constrains and the adsorption on the surfaces, severely restricts the chains diffusion into pseudo-two dimensional slit. Using in situ XRD, Vaia et al.[15] studied the intercalation kinetics for a model polymer (monodisperse polystyrene, molecular weight = 30,000) in octadecylammonium modified fluorohectorite mixture annealed in situ at 160°C in a vacuum oven and observed that the intercalation kinetics is quite rapid, even under quiescent flow. Initially, the basal reflection was observed at $2\theta = 4.15^\circ$ (i.e., $d = 2.13$ nm). During annealing, the linearity of reflection is reduced progressively, whereas a new basal reflection develops at $2\theta = 2.82^\circ$ (i.e., $d = 3.13$ nm), which corresponds to the intercalated hybrid. The width of the original and final intercalated peaks appear to be similar, suggesting that the polystyrene melt intercalation does not drastically alter the coherence length, or disrupt the layering structure of the silicate crystallites.[15] The structure of PCN, in particular with organically modified layered silicates, is strongly dependent on the polymeric property, the charge carrying capacity of the layered silicate, as well as the chain length and structure of the cationic surfactant. However, both compatibility and hybrid structure for these nanocomposites are independent of the polymer molecular weight. The experimental results and a lattice based mean field theory are given in Vaia and Giannelis.[16]

The intercalation of polymer or prepolymer from the solution is described via minimum free energy principle. The driving force of polymer intercalation is the entropy from the solvent desorption. Several researchers investigated the thermodynamics properties of PCN with homo polymeric systems in a confined geometry. However, Lim et al.[17] investigated ternary systems, and explained that the intercalation distance of poly(methyl methacrylate) (PMMA)/organic-modified clay (OMMT) nanocomposite is larger than that for the PEO/OMMT mixture. Based on Flory–Huggins interaction parameters between polymers and silicate layers in a PEO/PMMA/OMMT mixture, the PMMA molecules have a larger affinity for the hydrophobic surfaces of OMMT than hydrophilic PEO.[17] The larger layer thickness and broadness of the peak in the PMMA/OMMT mixture are indicative of an improved intercalation for PMMA matrix over the OMMT surfaces. The relatively weak interaction between PEO and OMMT is because of the hydrophobic modification of the clay surface.

Furthermore, it is also noted that the interlayer distance of the PCN is related to the intrinsic polymer structure, such that the PEO molecule exhibits a straight chain structure and has high backbone flexibility because of the backbone oxygen atom. Also, the minimum thickness required to accommodate this polymer is much smaller than the minimum layer-thickness required for accommodating bulky PMMA molecules.

Computer Simulation in Confined Polymeric System

Giannelis, Krishnamoorti, and Manias have authored an excellent article on this subject.[14] A major challenge in developing nanocomposites for systems ranging from high-performance to commodity polymers is the lack of understanding in nanostructure-physical property relationship. Without this, progress in the nanocomposites area will largely depend on empirical rules, i.e., a cookbook approach. The large internal interfacial area between the polymer and the silicates, as well as the nanoscopic dimensions between the layered silicates differentiates PCNs from conventional blends and filled plastics. Researchers typically adopt thermodynamic and rheological analysis in bulk and extend their findings to the confined systems; however, it remains difficult to accurately implement confinement effects. Computer simulations will provide the structure and dynamics of the intercalated surfactant molecules, which will eventually assist the design of polymer-silicate nanocomposite systems. The dynamics of confined polymer is significantly different from the molecules in the local relaxation, the mobility, and the rheological properties. As a result, molecular simulation is desirable for modeling PCN, yet it requires tremendous computational effort.

The Monte Carlo (MC) and MD simulation, although still in their infantile stages, will provide in depth insight into the structure of nanocomposites at a nanoscopic level. For bulk polymeric systems, however, numerous mathematical modeling and simulation tools already exist. Manias et al.[2] examined polystyrene based nanocomposite via MD simulations,

and found the nonhomogeneity in density profile resulting from the confinement effect. For confined nanoscale systems, such as applications of PCN, only a limited body of literature exists. On the other hand, the static and dynamic behavior of a polymeric liquid film confined in a single wall case was studied extensively by our group,[18] using MC and MD (by systematically tuning polymer-wall interaction energy). In equilibrium case, we calculated the density profile and surface roughness (uniformity of the film) by MC simulation. MD simulation was used to calculate the transport coefficient and the diffusion coefficient, in particular. Relaxation process was also investigated via a correlation function. Here, the Kohlausch–Williams–Watts stretched exponential model was used, instead of single relaxation time model, to fit the time dependent correlation function. These findings can be generalized easily to the two-wall situation (as sketched in Fig. 5) and provide a fundamental understanding of nanoscopic structures and the molecular design for PCN.

APPLICATIONS

Improved Properties

In one example, the tensile strength of polyamide 6 was increased by 55% and the moduli by 90%, with the addition of only 4 wt% of delaminated clay. The enhanced tensile property of PCN suggests that nanocomposite performance is related to the degree of clay delamination, which increases the interaction between the clay layers and the polymers. Several explanations, based on the interfacial properties and the mobility of the polymer chains, have been given for this reinforcement. Kojima et al.[19] reported that the tensile modulus improvement for polyamide 6-clay hybrid originated from a constrained region, where the polymer chains have reduced mobility. The dispersion and delamination of the clay were the key factors for the reinforcement. The delaminated nanocomposite structure produces a substantial increase in modulus, yet it lowers the fracture toughness. A study on nylon 6-clay nanocomposite revealed that the fracture energy is ten times lower than the 4 wt% delaminated clay. It is thought that the reduction of plastic deformation for the constrained polymer matrix increases the brittleness of the nanocomposite. In the presence of microscale aggregates, significant plastic deformation and improvement in toughness was observed.[20] The results of tensile strength and elongation at break for PEO/clay nanocomposites vs. clay content are provided in Fig. 6A and 6B. A considerable improvement in tensile strength and a drastic decrease in elongation at break are observed as the clay content increase. The enhancement in tensile strength is directly affected by the reinforcement of PEO/clay naocomposite. The addition of clay in cross-linked matrices also triggers an increase in the elongation at break.

As described in the section ''Nanoscopically Confined Polymer Dynamics,'' an enthalpic force enables the polymer to penetrate into the clay layer (as shown in Fig. 7A), while the clay acts as a blockade to this penetration.

Another advantage of PCNs is the substantial decrease in permeability. The drastic decrease in permeability is attributed to the large aspect ratio of the clay layers, which increase the tortuous path of the gas as it diffuses into the nanocomposites.

As described in Fig. 7B(i) and 7B(ii), the relative tortuous path $\tau(\equiv d/d')$ increases with the aspect ratio of the clay layers. Measurements of the permeability of polyimide–clay hybrids, using clays with various aspect ratios, tend to verify this conjecture. However, in the case of water permeability, recent studies with the PCNs show a reduction by a factor of 40, when compared with the pristine polymer, contradicting this conjecture. Indeed, the aspect ratio of the clay layers in this material is smaller than that of the clay used in nylon 6-clay hybrids, causing a decrease in water permeability. From this, we conclude that the constraint resulting from the polymer chains structure in PCN may be the critical factor in determining the permeability. Low gas permeability for CO_2 and O_2 has advantage for food packaging (beverage or fatty food containers). These properties were examined in

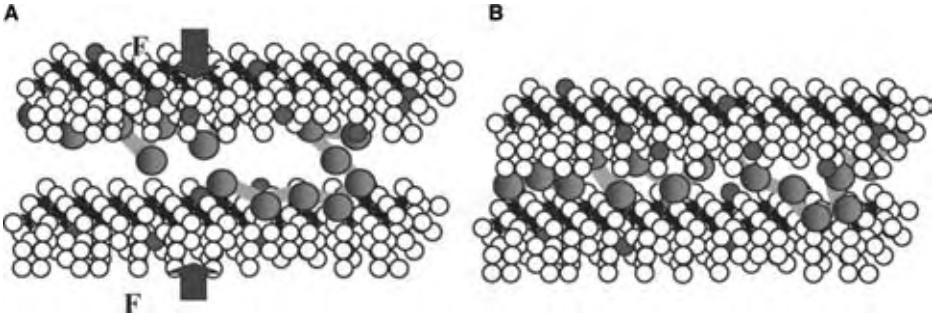


Fig. 5 Schematic diagram of reactive bead-spring model confined by two-site wall slit. Red spheres indicate polymer beads. Blue spheres in the slit wall represent the polar sites and white spheres represent the organically modified surface region: (A) before and (B) after compression.

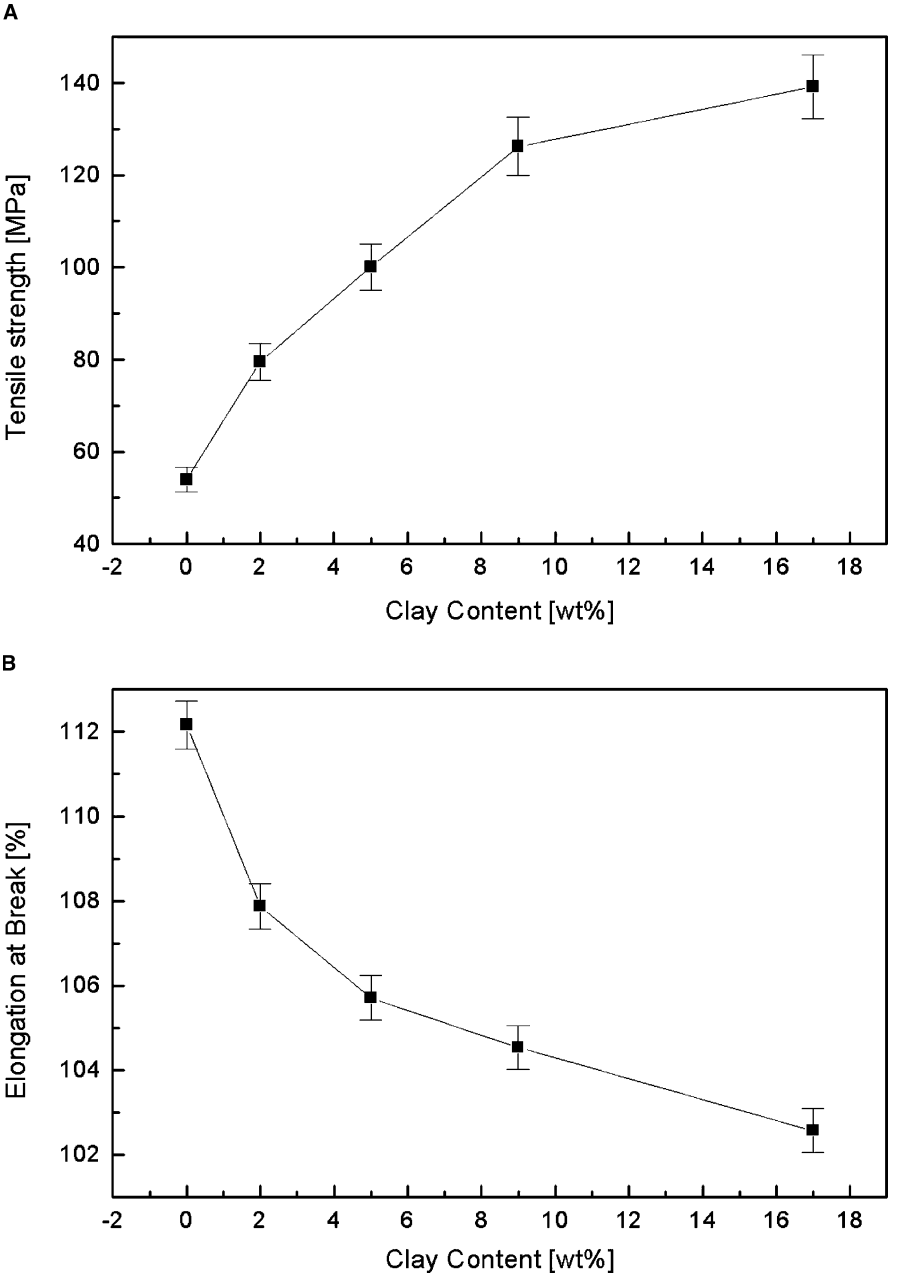


Fig. 6 (A) Tensile strength and elongation at break and (B) for PEO/clay nanocomposites.

comparison with the conventional filler type composites, and several companies have projected a market for PCNs as a packing material. The relationship between the rheological properties and microstructure of PCNs has been also examined, the results of which could complement the standard characterization methods for PCNs. Two delaminated hybrid systems prepared by in situ polymerization, poly(ε-caprolactone)/MMT and nylon 6/MMT, wherein the polymer chains are end-tethered to the silicate surface via cationic surfactants, are examined.[21] The large amplitude oscillatory shear measurements for the viscoelastic properties of the oriented nanocomposites were examined. Various molecular weights of the polymer chains, originating from the synthetic scheme used in the preparation of the nanocomposites, poses some concern for interpreting moduli variation with increasing silicate content. The average molecular weight decreased sharply at low silicate loading, and remained almost the same for nanocomposites, with 2 wt% or greater silicate loading. Despite the efforts made so far, state-of-the-art rheological measurements are insufficient to analyze the nanostructural information.

In the area of dispersion, it is believed that the shear force revokes the rearrangement of the silicate layers. This behavior increased the mechanical properties, but the rheological response is different from the

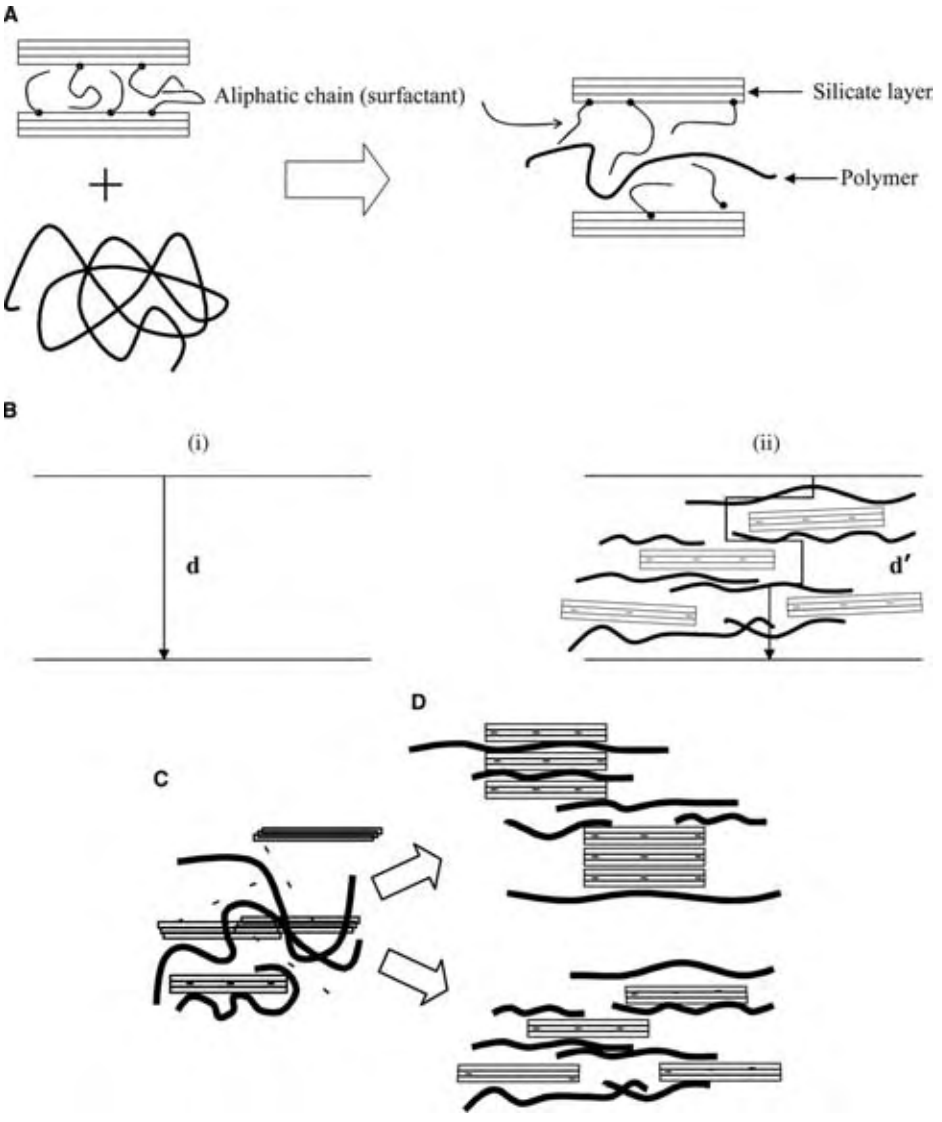


Fig. 7 (A) Polymer penetrates into the clay layer with enthalpic force; (B) (i) path without and (ii) with layered silicate clays. (C) A schematic of the PCN microstructure at equilibrium (D) and after shear.

dispersion characteristics of silicate layer in the polymer matrix. Further, the bundle of silicate layers was sheared differently (Fig. 7C and 7D). At present, it is difficult to measure the nanoscopic properties of PCN from the conventional rheological measurements alone.

An increase of thermal stability has also been observed for PEO/clay nanocomposites (Fig. 8). The nanocomposites with increasing clay content show substantially higher degradation temperatures than pure PEO. The start of the shift in the degradation temperature indicates a decrease in the oxygen permeability or a volatile degradation product, resulting from the homogeneously distributed clay sheets.

Polymer clay nanocomposites exhibit very low flammability.[22] For instance, the heat release rate during the combustion of polyamide 6-clay nanocomposite is reduced by 63% with a clay content of 5 wt%. The nanocomposite structure also enhances the property of the char through reinforcement of the char layer, indicating that nanocomposite parts preserve their shape during combustion. After 2 hr of ignition at 1000°C, it still is possible to distinguish the shape of the epoxy–clay nanocomposites samples prior to ignition. A study on polyetherimide nanocomposites demonstrated that intercalated nanocomposites showed the highest char yield and self-extinguishing behaviors. Thus, PCN has been recommended as a novel and environmentally friendly material for improving fire resistance.

Future Applications

Since the advent of PCNs, many researchers have attempted to commercialize them for a variety of engineering applications. The first commercialized nanocomposite product, developed by Toyota Central R&D,[4] was based on in-reactor processing of

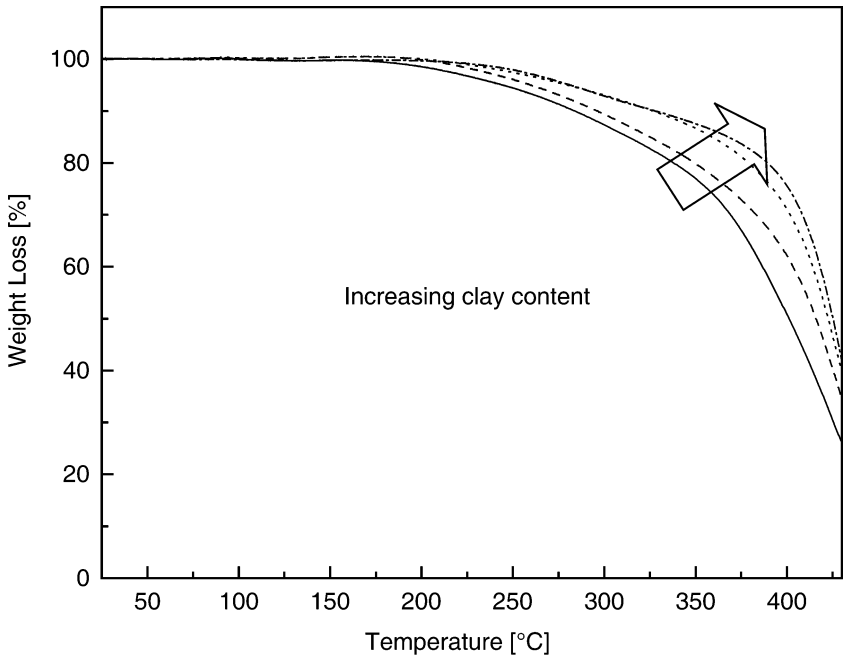


Fig. 8 Thermogravimetric analysis curves for PEO/clay nanocomposites.

caprolactam and MMT after an ion exchange with the hydrochloride salt in the aminolauric acid. While containing a minimal MMT, its properties (especially mechanical and thermal properties) were significantly improved as compared to conventional filler reinforced composites.

Technology for preparing nanocomposites directly via compounding has been investigated by Vaia, Ishii, and Giannelis.[3] Industrial R&D efforts have focused on process technology (e.g., melt or monomer exfoliation processes), as there are a number of polymers (e.g., polyolefins) that do not lend themselves to a monomer process. Nanocomposites with a variety of polymers, including polyacrylates or methacrylates, polystyrene, styrene-butadiene rubber, epoxy, polyester, and polyurethane, are amenable to the monomer process. The enhancement of mechanical properties, gas permeability resistance, and heat endurance are the primary objectives for the application of PCN, and their success will establish PCNs as a major commercial product.

Commercialization of PCN requires close consideration of potential environmental impacts, an improved understanding of synthesis mechanisms, and the PCN designs demanded by end-users.[23,24]

CONCLUSIONS

In this review, we examined the topics on PCNs in three categories: preparation, characterization, and application. Distinct from the previous review papers,[1,2] we introduced a spectrum of properties and characteristic methods, as well as our own research topics including rheology, MD in confined geometries, and applications. The preparation methods from the mass production view point were examined and the emerging role of rheology was also recognized. Although conventional rheological methods are important in processing, they may not give sufficient information for analyzing the nanocomposite structure of PCNs. The desirable properties, including improvement of mechanical properties, gas barriers, fire flammability, make commercialization and processing of PCN products viable. Future research should focus on better understanding and predicting environmental issues and must address these issues and specialized applications of PCNs, as well as their commercialization and marketability.

REFERENCES

1. Ray, S.S.; Okamoto, M. Polymer/layered silicate nanocomposites: a review from preparation to processing. Prog. Polym. Sci. **2003**, *28*, 1539–1641.
2. Manias, E.; Kuppa, V.; Yang, D.K.; Zax, D.B. Relaxation of polymers in 2 nm slit-pores: confinement induced segmental dynamics and suppression of the glass transition. Colloids Surf. A **2001**, *187–188*, 509–521.
3. Vaia, R.A.; Ishii, H.; Giannelis, E.P. Synthesis and properties of two-dimensional nanostructures

by direct intercalation of polymer melts in layered silicates. Chem. Mater. **1993**, *5*, 1694–1696.
4. Okada, A.; Kawasumi, M.; Usuki, A.; Kojima, Y.; Kurauchi, T.; Kamigaito, O. Nylon 6-clay hybrid. Mater. Res. Soc. Proc. **1990**, *171*, 45–50.
5. LeBaron, P.C.; Wang, Z.; Pinnavaia, T.J. Polymer-layered silicate nanocomposites: an overview. Appl. Clay Sci. **1999**, *15*, 11–29.
6. Ray, S.S.; Okamoto, M. Biodegradable polylactide and its nanocomposites: opening a new dimension for plastics and composites. Macromol. Rapid Commun. **2003**, *24* (14), 815–840.
7. Alexandre, M.; Dubois, P. Polymer-layered silicate nanocomposites: preparation, properties and uses of a new class of materials. Mater. Sci. Eng. **2000**, *28*, 1–63.
8. Messersmith, P.B.; Giannelis, E.P. Synthesis and characterization of layered silicate-epoxy nanocomposites. Chem. Mater. **1994**, *6*, 1719–1725.
9. Vaia, R.A.; Jandt, K.D.; Kramer, E.J.; Giannelis, E.P. Kinetics of polymer melt intercalation. Macromolecules **1995**, *28*, 8080–8085.
10. Kim, J.W.; Jang, L.W.; Choi, H.J.; Jhon, M.S. Physical and electroresponsive characteristics of the intercalated styrene-acrylonitrile copolymer/clay nanocomposite under applied electric fields. J. Appl. Polym. Sci. **2003**, *89*, 821–827.
11. Damour, A.A.; Salvetat, D. Et analyses sur un hydrosilicate d'alumine trouvé à montmorillon. Ann. Chim. Phys. Ser. **1847**, *21*, 376–383.
12. Lim, S.T.; Lee, C.H.; Choi, H.J.; Jhon, M.S. Solidlike transition of melt-intercalated biodegradable polymer/clay nanocomposite. J. Polym. Sci. Part B: Polym. Phys. **2003**, *41*, 2052–2061.
13. Yano, K.; Usuki, A.; Karauchi, T.; Kamigaito, O. Synthesis and properties of polyimide-clay hybrid. J. Polym. Sci. Polym. Chem. **1993**, *31*, 2493–2498.
14. Giannelis, E.P.; Krishnamoorti, R.; Manias, E. Polymer-silicate nanocomposites: model systems for confined polymers and polymer brushes. Adv. Polym. Sci. **1999**, *138*, 107–147.
15. Vaia, R.A.; Jandt, K.D.; Kramer, E.J.; Giannelis, E.P. Microstructural evolution of melt intercalated polymer-organically modified layered silicates nanocomposites. Chem. Mater. **1996**, *8*, 2628–2635.
16. Vaia, R.A.; Giannelis, E.P. Lattice model of polymer melt intercalation in organically-modified layered silicates. Macromolecules **1997**, *30*, 7990–7999.
17. Lim, S.K.; Kim, J.W.; Chin, I.; Kwon, Y.K.; Choi, H.J. Preparation and interaction characteristics of organically modified montmorillonite nanocomposite with miscible polymer blend of poly(ethylene oxide) and poly(methyl methacrylate). Chem. Mater. **2002**, *14*, 1989–1994.
18. Jhon, M.S. Physicochemical properties of nanostructured perfluoropolyether films. In *Advances in Chemical Physics*; Rice, S.A., Ed.; John Wiley: New York, NY, 2004; Vol. 129, 1–70.
19. Kojima, Y.; Usuki, A.; Kawasumi, M.; Okada, A.; Fukushima, Y.; Kurauchi, T.; Kamigaito, O. Mechanical properties of nylon 6-clay hybrid. J. Mater. Res. **1993**, *8*, 1185–1189.
20. Manias, E.; Touny, A.; Wu, L.; Strawhecker, K.; Lu, B.; Chung, T.C. Polypropylene/montmorillonite nanocomposites—review of the synthetic routes and materials properties. Chem. Mater. **2001**, *13*, 3516–3523.
21. Krishnamoorti, R.; Giannelis, E.P. Rheology of end-tethered polymer layered silicate nanocomposites. Macromolecules **1997**, *30*, 4097–4102.
22. Morgan, A.B.; Harris, J.D. Effects of organoclay soxhlet extraction on mechanical properties, flammability properties and organoclay dispersion of polypropylene nanocomposites. Polymer **2003**, *44*, 2313–2320.
23. Kim, B.H.; Jung, J.H.; Joo, J.; Epstein, A.J.; Mizoguchi, K.; Kim, J.W.; Choi, H.J. Nanocomposite of polyaniline and Na^+-montmorillonite clay. Macromolecules **2002**, *35*, 1419–1423.
24. Wu, J.; Lerner, M.M. Structural, thermal, and electrical characterization of layered nanocomposites derived from Na^+-montmorillonite and polyethers. Chem. Mater. **1993**, *5*, 835–838.
25. Kojima, Y.; Usuki, A.; Kawasumi, M.; Okada, A.; Kurauchi, T.; Kamigaito, O.; Kaji, K. Novel preferred orientation in injection-moled nylon 6-clay hybrid. J. Polym. Sci. Part B: Polym. Phys. **1995**, *33*, 1039–1045.
26. Kornmann, X.; Thomann, R.; Mülhaupt, R.; Finter, J.; Berglund, L. Synthesis of amine-cured, epoxy-layered silicate nanocomposites: the influence of the silicate surface modification on the properties. J. Appl. Polym. Sci. **2002**, *86*, 2643–2652.
27. Tien, Y.I.; Wei, K.H. High-tensile-property layered silicates/polyurethane nanocomposites by using reactive silicates as pseudo chain extenders. Macromolecules **2001**, *34*, 9045–9052.
28. Ray, S.S.; Yamada, K.; Okamoto, M.; Ueda, K. New polylactide-layered silicate nanocomposites. II. Concurrent improvements of material properties, biodegradability and melt rheology. Polymer **2003**, *44*, 857–866.
29. Pramanik, M.; Srivastava, S.K.; Samantaray, B.K.; Bhowmick, A.K. Rubber–clay nanocomposite by solution blending. J. Appl. Polym. Sci. **2003**, *87*, 2216–2220.
30. Krikorian, V.; Kurian, M.; Galvin, M.E.; Nowak, A.P.; Deming, T.J.; Pochan, D.J. Polypeptide-based nanocomposite: structure and properties of poly(L-lysine)/Na^+-montmorillonite. J. Polym. Sci. Part B: Polym. Phys. **2002**, *40*, 2579–2586.

Polymer Composites

Youssef K. Hamidi
M. Cengiz Altan
School of Aerospace and Mechanical Engineering, University of Oklahoma, Norman, Oklahoma, U.S.A.

Brian P. Grady
School of Chemical, Biological and Materials Engineering, University of Oklahoma, Norman, Oklahoma, U.S.A.

INTRODUCTION

Polymer composites are mixtures of a polymer, which is the major or continuous phase, and a filler, which can be either metal, ceramic, or even another polymer. Both thermosetting and thermoplastic resins can be used as the polymer phase; the former has the advantage of low viscosity while the latter has the advantage of the possibility of recycling and reuse. Essentially, all commercially important polymers have applications where the polymer is filled, although certainly some materials are more commonly filled than others. Typically, the reason that a particular polymer is a good or a bad candidate for use as the continuous phase of a composite is its ability to form strong interactions with a particular filler. Composites are a fast growing segment of the polymer industry; composites filled with materials having at least one dimension in the nanometer-size range such as nanoclays or nanotubes represent a step change in technology in the composites area. The purpose of this entry is to introduce the reader to this rather diverse field, with an emphasis on processing. For more information the reader is referred to one of the excellent monographs on the subject.[1–3]

MATERIALS

Thermoplastics

Global worldwide production of thermoplastics is approximately 100 billion kg/yr, or approximately 12 kg for every person on the planet. Only a small fraction of this amount is filled and used as a composite, but a small fraction of this large number is still a significant amount of material. By far the most important thermoplastic composites are made from flexible thermoplastics, i.e., semicrystalline materials with a glass transition temperature below room temperature. One significant exception to this generalization is polycarbonate; however, these glassy materials are, for a glass, flexible at room temperature. The reason for the use of fillers in flexible materials is to add stiffness, while the cost is typically a reduction in flexibility. Glassy polymers typically do not need more stiffness, while many applications require more stiffness from a flexible material. High-density polyethylene and polypropylene are used very commonly in polymer composites (see Fig. 1). These materials are commonly filled with a low-cost filler such as wood, clay, or glass. The decrease in flexibility and toughness caused by the introduction of filler is large in these materials because the adhesion between these materials and the filler is typically poor. The filler can be coated with a thin layer to improve the interaction between the filler and the polymer; however, this coating involves a significant cost. The cost of the composite, in the absence of a coating, can be significantly lower than the cost of the neat resin, which explains the popularity of these polymer composites. Good adhesion between relatively polar polymers and polar fillers yields higher performance composites. The most common polar filler is glass; while both polyamides (various types) and polycarbonate are commonly used as polar thermoplastic resins (see Fig. 1).

Thermosets

Unlike thermoplastics, which are simply melted, thermoset resins chemically react from low-viscosity liquids to solid materials during processing, a process termed curing. Structurally, thermosets differ from thermoplastics because of the presence of cross-links in the former, which means that thermosets cannot be reshaped or recycled once the chemical reaction occurs. One advantage of thermosets vs. thermoplastics is that wetting the filler becomes much easier with a low-viscosity material. By far the most common thermoset composite is automobile tires, which consist of a polymer made from styrene and butadiene monomers and carbon-black filler. The actual recipe used is much more complicated, and can include other monomers or polymers, as well as other fillers. In the absence of filler, the cured resin is rubbery at room temperature, which makes tires a

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039238

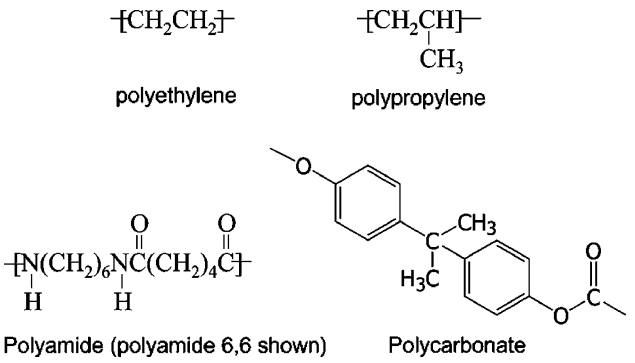


Fig. 1 Common thermoplastic resins used as the matrix phase in polymer composites.

unique thermosetting composite because most thermosetting composites are made with resins that are stiff at room temperature. Other types of resins commonly used as thermosets include epoxies, polyesters, and vinyl esters. Unlike thermoplastics, these categories are a bit misleading in the sense that various structures are classified under one heading. For example, approximately 10 different molecules with the epoxy functionality are important commercial products, and each has slightly different mechanical properties and/or environmental stability. The two monomers that typically combine to form an epoxy resin are an epoxide and an amine. Polyesters are made with polyester resins possessing unsaturation, and styrene. The structure of the polyester and the styrene (or other comonomer content) is varied in these systems. Finally, vinyl ester is a term used to describe bifunctional monomers with long organic structures between acrylic end groups.

Manufacturing Methods

The performance of a composite part depends not only on the materials selected but also on the process parameters used during manufacturing. Properties of the polymer matrix, quality of fiber–matrix adhesion, as well as composite microstructure and defects are functions of manufacturing. Various processes for manufacturing of commercial composite parts have been developed during the last 50 yr. While the aerospace, marine, and defense industries usually require manufacturing methods yielding high-performance composites, automotive and consumer goods industries focus on cost-effective, high-volume production methods. The type of manufacturing method also depends largely on whether the starting material is a low-viscosity uncured thermoset or a high-viscosity thermoplastic. This section briefly outlines the different composite manufacturing methods currently used in the composites industry.

Wet Layup

Wet layup consists of placing a layer of dry reinforcement inside a mold and then applying an uncured, low-viscosity thermoset resin as shown in Fig. 2. Woven glass fibers are the prevalent reinforcing preform utilized in layup processes, although carbon and aramid fibers are also used to a lesser extent.[3–5] Typical fiber volume fraction of composites manufactured via wet layup range between 30% and 50%. The resin can be poured, sprayed, or brushed on top of the preform layer either by hand or by machine. The fiber preform layer is rolled on or pressed after the application of resin to evenly distribute the resin and remove air pockets. Resin is applied on top after each layer of fiber mats is properly placed. This process is repeated until the desired thickness is reached. To provide a smooth surface finish on the mold side, a thin layer of mold release is often applied prior to starting the layup. Thereafter, pressure and heat are applied to allow the composite to cure. Pressure can be applied

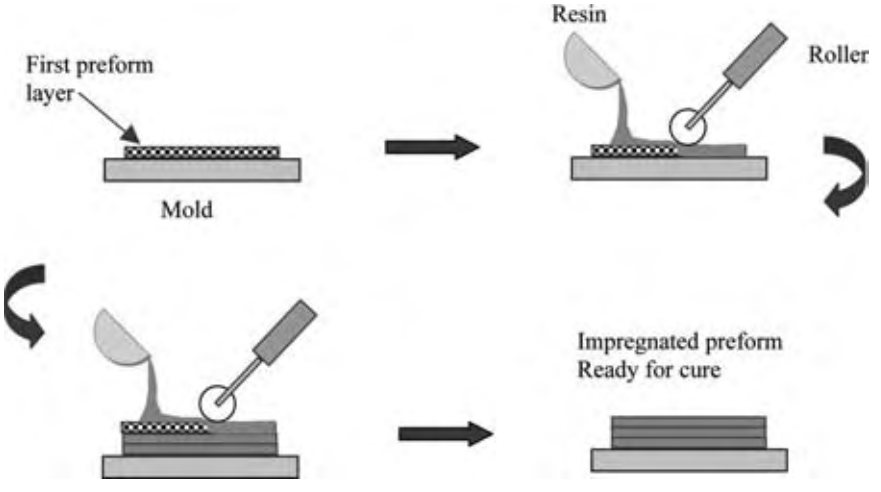


Fig. 2 Typical wet layup operation. *(View this art in color at www.dekker.com.)*

either mechanically or pneumatically; for more details see the section Molding Process. The material may not be cured to a final product; in some cases the material is partially cured to yield a product that is still soft but the resin has enough viscosity not to drip out of the material. This material is termed prepreg, and the prepreg can be shipped from one location to another where it is then formed and made into a final product.

Filament Winding

In the filament winding process, a continuous tape of polymer-impregnated fibers is wrapped over a rotating mandrel to form a composite part. Preform tapes can be either in preimpregnated or impregnated in a thermoset resin bath right before or right after winding. Successive layers are added at the same or different winding angles until the desired thickness is reached. The schematic of a typical filament winding process is shown in Fig. 3. Most winding machines operate similar to a lathe; the mandrel is mounted horizontally and rotates at a constant speed while the carriage delivering the fibers moves along the length of the composite part. Typical winding speeds range between 50 and 100 m/min. After the winding is complete, the composite part is allowed to cure either at room temperature, in an oven, or inside an autoclave. Steel and aluminum mandrels are usually used to facilitate the removal of the fabricated composite part after cure. However, inflatable and collapsible metal mandrels are also used in closed-end products.

Parts with diameters as small as 25 mm and as large as 6 m are being made by filament winding. In addition, a variety of fiber orientations can be achieved, leading to more control of design properties of the fabricated composites. Continuous fibers of glass, carbon, and aramid are used in filament winding while glass/polyester systems are more widely used because of their lower cost. The filament winding process can be automated yielding cost-effective, high-volume production. Most shapes fabricated by this technique are axisymmetric and include glass-fiber pipes, pressure vessels, rocket motor cases, sailboard masts, and other similar products.

Pultrusion

Pultrusion is a low-cost, high-volume process used to manufacture long, constant cross-section shapes from thermoset resins, and consists of pulling continuous dry fibers through a resin bath as shown in Fig. 4. Thereafter, impregnated fibers are pulled through a heated die for cure. The cross section of the die dictates the final shape of the product. The length and temperature of the die are determined by the pulling speed, the dimensions of the part, and the curing characteristics of the resin. At the exit of the heated die, the composite cools rapidly and 2–3% shrinkage is observed. The composite is continuously produced and a moving saw can be used to cut the part at the desired length without stopping the production.

Glass, carbon, and aramid fibers are used as unidirectional or fabric mat reinforcements, with E-glass/polyester being the most commonly used system.[3,6] The limitation of pultrusion is that only constant cross-section parts can be fabricated. However, a variety of hollow and solid profiles of any length can be manufactured.

Molding Processes

Compression molding

Compression molding consists of placing a predetermined amount of composite inside matching male

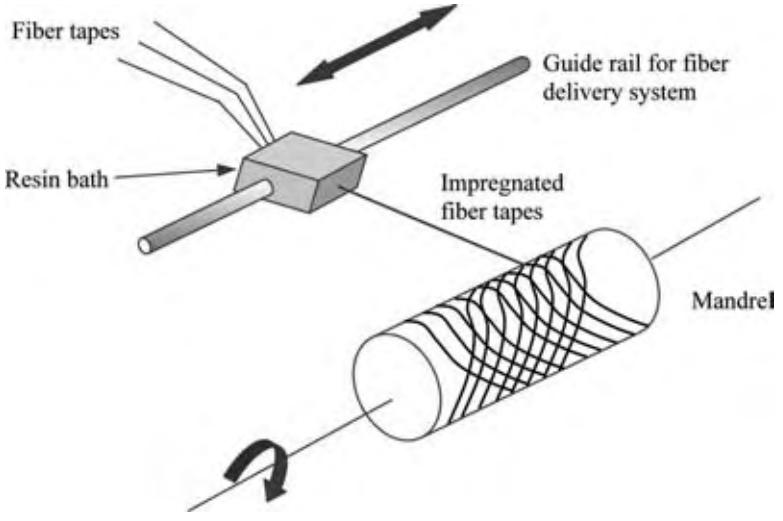


Fig. 3 Typical filament winding process schematic.

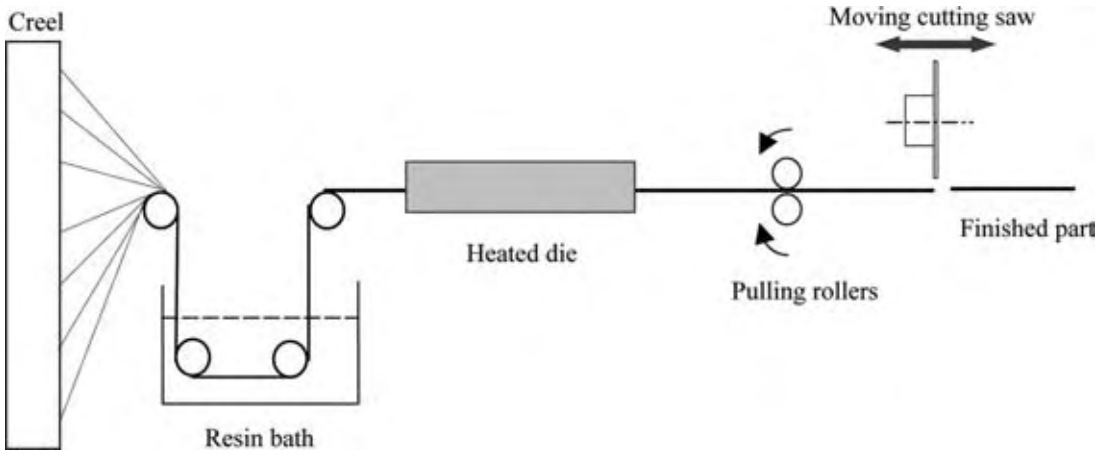


Fig. 4 Schematic of the pultrusion process.

and female metal molds. The mold walls are then heated, closed, and pressurized mechanically until the composite deforms into the desired shape. After the charge fills the mold cavity, the pressure is released and the molded part is ejected from the mold. A schematic of this process is presented in Fig. 5.

Discontinuous fibers are commonly used in compression molding with fiber contents often limited at 30% to maximize the surface quality of the final product while achieving complete filling of the mold cavity.[6,7] Sheet molding compound (SMC) and bulk molding compound (BMC) are the most commonly used materials in compression molding. Sheet molding compound is obtained by mixing chopped fibers, liquid resin, and fillers into a 2–5 mm-thick sheet product. A typical SMC compound contains glass fibers, polyester resin, and calcium carbonate.[3,6] Compression molding is also used for thermoplastics; however, injection molding is preferred unless mats are used. Compression molding usually requires a large initial investment and produces semistructural parts. The simplicity of the process minimizes part setup time, reduces secondary finishing, and allows high-volume production with a low scrap rate.

Autoclave molding

Instead of using mechanical force to pressurize a composite and force consolidation/shape, air pressure can be used if a low-viscosity thermoset is the resin. An autoclave is a large pressure vessel that allows the simultaneous application of heat and pressure during the manufacture of thermosetting composite laminates. Internal work space of commercial autoclaves can be as large as 10 m in diameter and 30 m in length. Fig. 6 depicts a small autoclave with a 90-cm diameter internal work space.

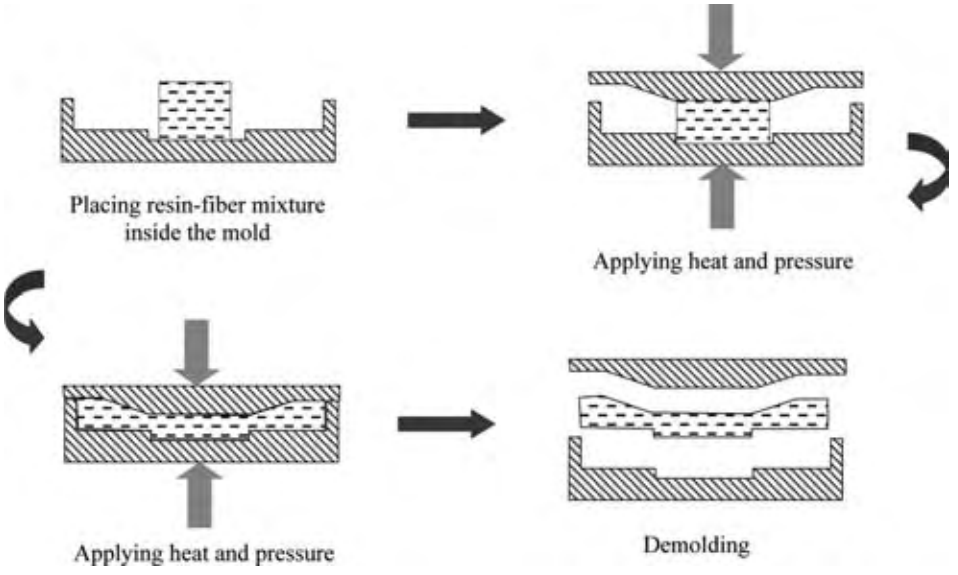


Fig. 5 Schematic of compression molding.

Fig. 6 Example of a small autoclave. *(View this art in color at www.dekker.com.)*

Typically, prepreg layers are cut to the desired shape and stacked in predefined orientations. Thereafter, the stacked prepregs are placed on a mold die and covered with a vacuum bag sealed at the edges. A vacuum is usually drawn before an external pressure of the order of 0.1–0.7 MPa is applied to the vacuum bag inside the autoclave. Simultaneously, the desired temperature profile is applied to ensure the complete cure of the composite part. An example of a temperature profile applied for 24 plies of AS4/3501-6 graphite/epoxy prepregs is depicted in Fig. 7. During most curing processes, an internal autoclave fan circulates the air inside ensuring uniform temperatures aided by convective heating. The external pressure and vacuum inside the bag create sufficient pressure differential to facilitate consolidation of the laminate and ensure good bonding between layers. During cure, the excess resin is absorbed by a bleeder placed above the prepregs within the vacuum bag. The application of vacuum also helps remove volatiles and excess air. Although autoclave processing requires high initial and operating costs, high-performance composites suitable for aerospace and similar structural applications are typically produced using this technique.

Injection molding

Injection molding is a high-volume, low-cost manufacturing process for thermoplastics, including thermoplastic composites, and is shown schematically in Fig. 8. Mixing of discontinuous fillers with a thermoplastic is the first step; this mixing typically occurs in a twin-screw extruder and the material is extruded and chopped into pellets. These pellets of thermoplastic resin mixed with chopped fibers or particles are then melted in a single-screw extruder, and the molten mixture is injected into a closed mold cavity. The injection pressure must be quite high to ensure complete filling and compensate for shrinkage due to cooling. After filling, the mold is then cooled down for solidification before the composite part is removed. Common volume fractions of particles range between 18% and 35%. Typical cycle times are on the order of 30 sec with the largest fraction of the time being spent during the cooling down step. Composites of complex geometries are conveniently fabricated to, essentially, their final shape. However, the required operating pressures (i.e., 50–100 MPa) limit injection molding to small- to medium-sized composite parts.[8]

Liquid composite molding

Liquid composite molding (LCM) processes such as resin transfer molding (RTM) have been long established in the automotive and aerospace industries as versatile technologies for manufacturing medium to large composite parts with complex geometries at low cost.[9–11] All LCM processes involve the injection of a liquid resin into a dry fiber perform, and are

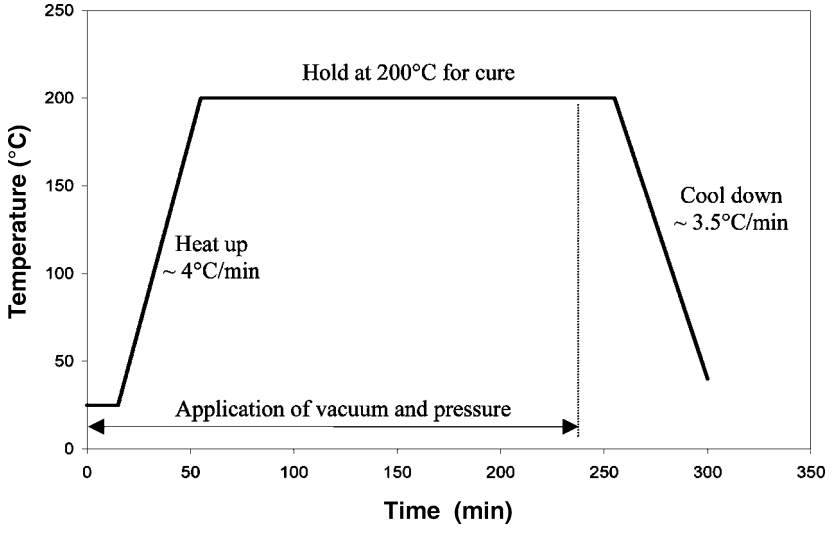


Fig. 7 Temperature profile utilized during an autoclave process for 24 plies of AS4/3501-6 graphite/epoxy prepegs.

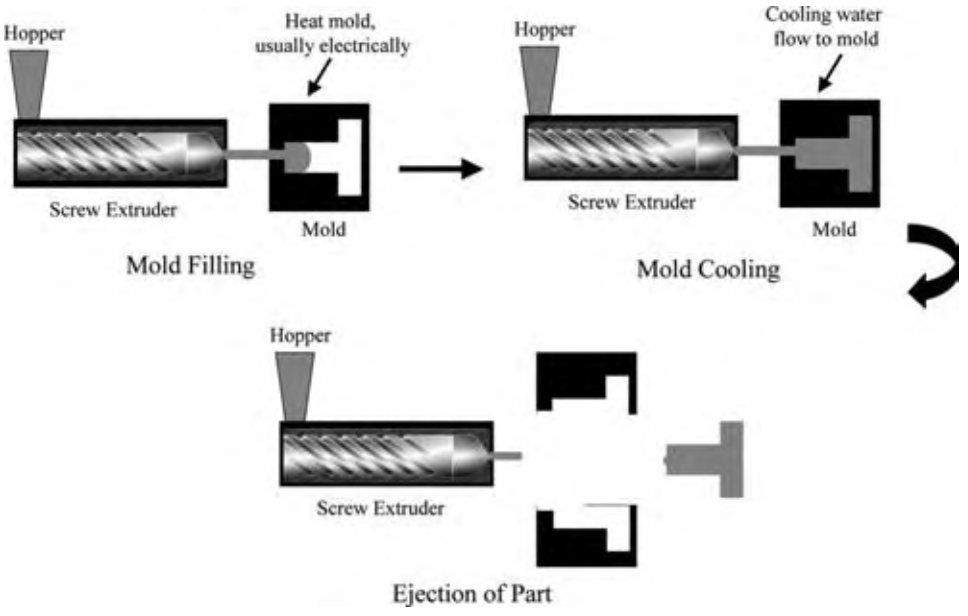


Fig. 8 Schematic of injection molding. *(View this art in color at www.dekker.com.)*

essentially the thermosetting equivalent to injection molding; except that a perform is typically placed in the mold prior to injection of the resin. Because of the lower viscosities of thermosets, the high pressures required for injection molding are not required for LCM. However, the cycle times are typically much longer because of the much longer time required for the part to harden, which in turn is due to the much slower kinetics of a chemical reaction in LCM vs. cooling in injection molding. Resin transfer molding, for instance, consists of injecting a reactive thermosetting resin into a closed mold cavity preloaded with a fibrous preform as shown in Fig. 9. Preforms used in LCM may consist of a 3D braided structure or multiple layers of 2D fabrics and mats. Glass, carbon, and aramid fibers are used as reinforcement, E-glass fibers being the most common.[3,11] High-performance carbon/epoxy systems are particularly used more often in the aerospace industry.[3,8,11] Resin transfer molding, can produce composite parts with volume fractions reaching 60%; however, typical applications may contain 25–35% fibers by volume. After the cure reaction is complete, the solidified composite part is removed from the mold. In other variants such as vacuum assisted resin transfer molding (VARTM), the impregnation is vacuum-driven, and thus half of the mold is often replaced with a vacuum bag. Because the pressure differential is much lower in VARTM, the cost of the mold can be reduced substantially, especially when fabricating larger parts.

PROPERTIES AND APPLICATIONS OF POLYMER COMPOSITES

Mechanical Properties

An increase in stiffness is probably the single most important reason that fillers are added to polymers. Stiffness is quantified using the modulus, which is the slope of a stress vs. strain curve at strain equals zero. The modulus, and other mechanical properties such as tensile strength and toughness, can be measured using different geometries; the two most common for composites are tensile and bending as shown in Fig. 10. Bending studies are typically more informative for fiber-reinforced composites because this geometry provides a better test of adhesion between the polymer and the filler, while tensile studies are more commonly used for particulate composites, if the matrix is flexible enough to allow tensile tests. For most systems, an increase in stiffness is typically accompanied by an unwanted decrease in flexibility and/or tensile

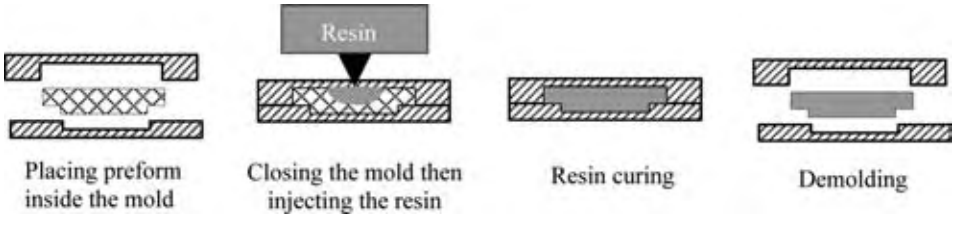


Fig. 9 Typical resin transfer molding process.

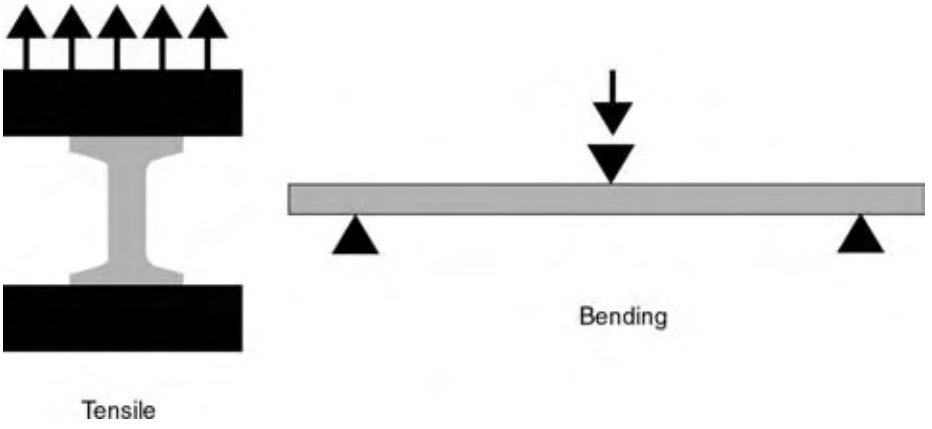


Fig. 10 Geometries for testing stress–strain of composites.

strength; and most important commercial fillers for a particular polymer have a lower reduction in these properties at a given volume fraction than other fillers. The applications of composites that depend primarily on mechanical property specifications are too numerous to list; some examples are airplane and automotive components. Other important mechanical properties that often justify the use of a filled system vs. one without a filler are abrasion resistance, for instance, automobile tires and resistance to creep, e.g., weight-bearing structural components.

Electrical and Thermal Conductivity

Another common reason to add a filler to a polymer is to increase either electrical conductivity or thermal conductivity. Polymers typically have electrical conductivity from 10^{-17} to 10^{-18} S/cm; though the addition of a moderately conductive filler such as carbon black conductivities of 10^{-2}–10^{0} S/cm are possible; highly conductive fillers such as silver can raise this value to 10^{1}–10^{2} S/cm. Applications include static dissipative devices and surge protectors. The impact of adding a highly thermally conductive filler to a polymer is much smaller at low-volume fractions vs. the impact of an electrically conductive filler on electrical conductivity. However, if a highly loaded stiff product is acceptable, polymer composites are capable of dissipating substantial amounts of heat.

Gas Diffusion and Fire Retardance

Nanotechnology has been identified as one of the key technologies that will lead to important advances in the quality of life, and important advances have been made using nanotechnology in the polymer composite area. The use of nanoclays; which have nanometer size thickness if exfoliated (i.e., separated apart), has enabled the manufacture of polymer-filled materials with vastly improved resistance to burning as well as vastly improved resistance to the diffusion of various gases. Applications that make use of the unique properties of composites filled with exfoliated clays are just beginning to appear. Manufacture of nanocomposites has proven to be quite difficult because agglomerates of clay platelets must be separated for the improved resistance to diffusion and flame to occur. Thermoplastic resins with specific clay–polymer interactions, e.g., nylon, have proven to be the most effective matrix for nanocomposites.

DEFECTS IN POLYMER COMPOSITES

The presence of defects in polymer composites adversely affects their mechanical properties and environmental resistance.[12–14] Conversely, environmental factors can lead to the worsening of already present defects. Defects in composites originate either during the manufacturing process as a result of incomplete wetting, improper process pressure, inadequate heating, or excessive temperature overshoot or during the service life of the composite because of external factors including cyclic loading, low-energy impact, moisture, and elevated temperatures. Poor choice of process parameters leads to voids, inclusions, resin burnout, and dry spots whereas external factors lead to defects such as subsurface damages, ply cracking, and delamination.

Defects Induced During Fabrication of Composites

Microvoids and dry spots

Depending on the manufacturing method, fiber type and content, and properties of the impregnating polymer, various mechanisms lead to the formation of large dry areas, called dry spots, and air inclusions or microvoids in the final composite part made with thermosetting resins. Common causes for void formation in polymer composites include volatilization of

dissolved gases or moisture in the resin during impregnation or curing, partial evaporation of mold releasing agent into the preform, and initial air bubble content in the resin mixture. In wet layup and pultrusion, uneven application of the resin is the primary cause of void and dry spot formation, whereas in autoclave curing of composite laminates, voids are mainly caused by entrapped air between layers in prepregs. In LCM processes, on the other hand, voids form during mold filling primarily because of mechanical entrapment of air at the advancing flow front.[15–19]

Void presence in composites, even in small amounts, is detrimental to their mechanical performance (Fig. 11). Judd and Wright reported that regardless of the manufacturing process, void presence induces reduction in interlaminar shear strength, tensile and flexural strength and modulus, torsional shear, fatigue resistance, and impact properties of a composite part.[12] Ghiorse indicated, for carbon/epoxy laminates, that each 1% increase in void content induced a 10% reduction in flexural and interlaminar shear strength, and a 5% reduction in flexural modulus.[13] Voidage is also known to affect both the rate and the equilibrium level of moisture absorption in composites.[14] Harper et al. reported that an increase from 1% to 5% in void volume fraction induces an increase of around 280% in the initial absorption rate, and 50% in the equilibrium mass gain for an AS4/3502 graphite/epoxy composite.[14]

Although it is not practically possible to manufacture void-free composites, different techniques have been successfully utilized for most manufacturing processes to fabricate composite parts with acceptable void content. Some of the void reduction techniques include: 1) degassing the resin mixture prior to impregnation; 2) utilization of vacuum to remove the air entrapped between prepreg layers; 3) proper selection of fiber/resin systems to lower gas emission during impregnation and cure; and 4) utilization of lower-viscosity resins at higher temperatures to facilitate the impregnation. Most industrial applications consider 1% void content as a threshold for acceptance of composite parts.[16,17] However, in highly structural applications, an acceptance threshold of 0.5% in void content is often applied.[20]

Wrinkling

A number of mechanisms are known to induce wrinkling in continuous-fiber polymer composites. These mechanisms include mismatch in thermal expansion between the fiber, matrix, and tool plate materials; the temperature changes experienced by the part during processing; and volumetric shrinkage of the matrix.[21–24] The difference between thermal expansion coefficients of fiber, matrix, and tool plate can be several orders of magnitude. Thus, residual stresses often develop during the cooling down of the part. The elevated cooling rates can lead to compressive stresses in the laminate surface, while slower cooling rates allow time for stresses to relax and avoid significant temperature gradients through the thickness of the part.[21]

Defects in injection molding

Incomplete filling and weld lines are the most common defects in injection-molded composites. High fiber contents and low heating temperatures often lead to highly viscous fiber–resin mixtures that require elevated injection pressures. If adequate pressure is not supplied or the solidified composite blocks parts of the injection gate or runners, only a fraction of the needed volume is injected into the mold cavity. This problem can usually be solved by using higher melting temperatures to reduce the mixture viscosity, and also by enlarging the injection gates. Another defect encountered in most injection-molded parts is the formation of weld lines.[25] When a weld line is formed by the merging of separate fluid fronts, fibers at both sides do not penetrate across the neighboring front, thus forming an unreinforced, structurally weak plane. This problem is often accentuated for parts containing longer fibers and higher fiber contents.[4] Fiber breakage is another defect affecting the performance of injection-molded parts. The mixing and shearing of the extruder screw often reduces the fiber length, yielding a composite part molded with reduced mechanical properties. Low screw speeds and injection pressure can be utilized to avoid excessive fiber breakage.[4]

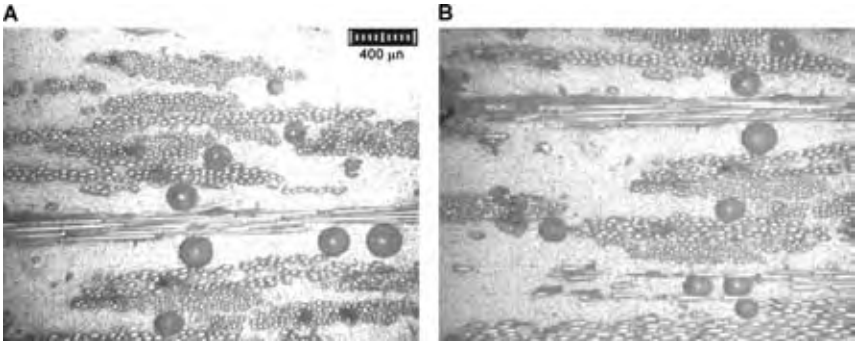


Fig. 11 Representative microscopic images of voids obtained from a glass/epoxy RTM composite.

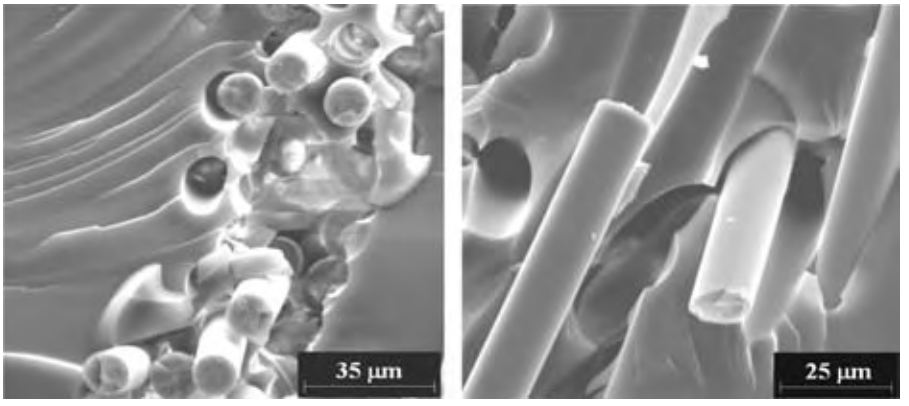


Fig. 12 Scanning electron micrographs of failed surfaces of an E-glass/epoxy composite.

Defects Induced During Service Life of Composites

Matrix cracks and fiber fractures

Although matrix cracks can be induced by residual stresses developed during curing, most matrix cracks are initiated during the service life of the composites.[23] Given that the matrix usually sustains a lower stress before failure compared to the fiber reinforcements, microcracks are often generated in the matrix after the application of high loads. Depending on the fiber orientation, content, and the direction of the applied load, multiple matrix cracks with different orientations can form inside the composite. If the matrix cracks propagate above a certain limit, fibers start supporting most of the applied load before they fail under different mechanisms. Short fibers frequently fail under pullout, i.e., extraction of the fiber from the matrix, while continuous fibers tend to fracture under excessive loads. Depending on the strength of the matrix–fiber adhesion, a combination of both mechanisms is usually present in fractured composites.[26] Environmental effects, such as the infusion of water, can lead to a reduction in matrix–fiber adhesion, which in turn leads to premature failure. Fig. 12 shows scanning electron micrographs of failed surfaces of an E-glass/epoxy composite.[27] Fibers failing under both pullout and fracture can be seen in Fig. 12.

Delamination

Delamination is the debonding of adjacent composite layers and is one of the most prevalent life-limiting defects in thermosetting composite laminates. Delamination is usually induced during service life of the composite from a low-energy impact or fatigue loading. However, delamination can also originate from other preexisting defects such as matrix cracks and voids. In addition to reducing the structural integrity of the material, delamination also results in deterioration of long-term performances.[28,29] Nondestructive evaluation methods have been commonly used to detect delaminations to prevent their excessive propagation leading to the final failure of the laminate.[30] Techniques such as x-ray and ultrasonic inspections are widely used, particularly in the aircraft industry where structural composite laminates are often used. Nondestructive evaluation tests usually help define regular inspection intervals and determine the need for repair.

REFERENCES

1. Astrom, B.T.; Astrom, T. *Manufacturing of Polymer Composites*; Chapman and Hall: New York, 1997.
2. Gay, D.; Hoa, S.V.; Tsai, S.W. *Composite Materials: Design and Applications*; CRC Press: Boca Raton, FL, 2002.
3. Mazumdar, S.K. *Composites Manufacturing: Materials, Product, and Process Engineering*; CRC Press: Boca Raton, FL, 2001.
4. Campbell, F.C. *Manufacturing Processes for Advanced Composites*; Elsevier: Amsterdam, 2004.
5. Hancox, N.L. *Fiber Composite Hybrid Materials*; MacMillan Publishing: New York, 1981.
6. Shibley, A.M. *Handbook of Composite Materials*; Van Nostrand Reinold, 1982.
7. Swanson, S.R. *Introduction to Design and Analysis with Advanced Composite Materials*; Prentice-Hall: Englewood Cliffs, NJ, 1997.
8. Bader, M.G; Smith, W.; Isham, A.B.; Rolston, J.A.; Metzner, A.B. Processing and fabrication technology. In *Delaware Composites Design Encyclopedia*; Technomic Publisher, 1990; Vol. 3.
9. Abraham, D.; McIlhagger, R. Investigations into various methods of liquid injection to achieve mouldings with minimum void contents and full wet out. Compos. Pt. A **1998**, *29* (5), 533–539.

10. Abraham, D.; Matthews, S.; McIlhagger, R. A comparison of physical properties of glass fiber epoxy composites produced by wet lay-up with autoclave consolidation and resin transfer moulding. Compos. Pt. A **1998**, *29* (7), 795–801.
11. Johnson, C.F. Resin transfer molding. In *Composite Materials Technology—Process and Properties*; Hanser Publisher, 1990.
12. Judd, N.C.W.; Wright, W.W. Voids and their effects on the mechanical properties of composites—an appraisal. SAMPE J. **1978**, *14* (1), 10–14.
13. Ghiorse, S.R. Effect of void content on the mechanical properties of carbon/epoxy laminates. SAMPE Q. **1993**, *23* (1), 54–59.
14. Harper, B.D.; Staab, G.H.; Chen, R.S. A note on the effect of voids upon the hygral and mechanical properties of as4/3502 graphite/epoxy. J. Compos. Mater. **1987**, *21*, 280–289.
15. Hamidi, Y.K.; Altan, M.C. Spatial variation of void morphology in resin transfer molded e-glass/epoxy composites. J. Mater. Sci. Lett. **2003**, *22* (24), 1813–1816.
16. Barraza, H.J.; Hamidi, Y.K.; Aktas, L.; O'Rear, E.A.; Altan, M.C. Porosity reduction in the high-speed processing of glass fiber composites by resin transfer molding (RTM). J. Compos. Mater. **2003**, *38* (3), 195–226.
17. Olivero, K.A.; Barraza, H.J.; O'Rear, E.A.; Altan, M.C. Effect of injection rate and post-fill cure pressure on resin transfer molded disks. J. Compos. Mater. **2002**, *36* (16), 2011–2028.
18. Mahale, A.D.; Prud'Homme, R.K.; Rebenfeld, L. Quantitative measurement of voids formed during liquid impregnation of nonwoven multifilament glass networks using an optical visualization technique. Polym. Eng. Sci. **1992**, *32* (5), 319–326.
19. Patel, N.; Rohatgi, V.; Lee, J.L. Micro scale flow behavior and void formation mechanism during impregnation through a unidirectional stitched fiberglass mats. Polym. Eng. Sci. **1995**, *35* (10), 837–851.
20. Goodwin, A.A.; Howe, C.A.; Paton, R.J. The role of voids in reducing the interlaminar shear strength in RTM laminates. Proceedings of ICCM-11, Australian Composite Structures Society, 1997; Vol. IV, 11–19.
21. Kugler, D.; Moon, T.J. Identification of the most significant processing parameters on the development of fiber waviness in thin laminates. J. Compos. Mater. **2002**, *36* (12), 1451–1479.
22. Meink, T.E.; Huybrechts, S.; Herman Shen, M.-H. Processing induced wrapage of filament wound composite cylindrical shells. J. Compos. Mater. **2002**, *36* (9), 1025–1047.
23. Dharia, A.K.; Hays, B.S.; Seferis, J.C. Evaluation of microcracking. In *Aerospace Composites Exposed to Thermal Cycling: Effect of Composite Lay-Up, Laminate Thickness, and Thermal Ramp*, 33rd International SAMPE Technical Conference, Nov 2001.
24. Coffin, D.W.; Pipes, R.B. Flange wrinkling. In *The Forming of Thermoplastic Composite Sheets*, AIP Conference Proceedings, Materials Processing and Design, 2004.
25. Papathanasiou, T.D.; Guell, D.C. *Flow-Induced Alignment in Composite Materials*; Woodhead Publishing, 1997.
26. Morley, J.G. *High-Performance Fiber Composites*; Academic Press, 1987.
27. Barraza, H.J. Admicellar-Polymerized Thin Elastomeric Coatings for Engineered Interfaces for Enhanced Composite Performance. Ph.D. dissertation, University of Oklahoma, 2003.
28. Reddy, A.D.; Rehfield, L.W.; Haag, R.S. Influence of prescribed delamination on stiffness-controlled behavior of composite laminates. Effects of defects in composite materials. ASTM STP **1984**, *863*, 71–83.
29. de Charentenay, F.X.; Harry, J.M.; Prel, Y.J.; Benzeggagh, M.L. Characterizing effect of delamination defect by mode i delamination test. Effects of defects in composite materials. ASTM STP **1984**, *863*, 84–103.
30. DiScalia, F.L.; Green, R.E. Jr. A hybrid non-contact ultrasonic system for sensing bond quality in tow-placed thermoplastic composites. J. Compos. Mater. **2000**, *34* (21), 1860–1880.

Polymeric Membranes

Takeshi Matsuura
Mehrdad Rafat
Department of Chemical Engineering, University of Ottawa, Ottawa, Ontario, Canada

INTRODUCTION

Membrane science can be divided into the following categories: material selection, material characterization and modification, membrane preparation, membrane characterization, membrane testing for applications, membrane transport, membrane module, and system design. This entry deals with the first three of the above-mentioned subjects, exclusively for polymeric materials. Polymeric properties that have to be considered in selecting polymeric materials for membrane preparation are identified. Together with the synthesis cost of polymeric materials, which is often prohibitively high, polymers are narrowed down in the selection procedure to a handful of technical polymers for practical applications. In the section for membrane preparation, a major portion is devoted to describe the methods to prepare asymmetric membranes, because most of the membranes that are currently used in industrial separation processes possess asymmetric structures. Methods for membrane surface modification are also described in recognition of the fact that the membrane performance is governed primarily by the surface of the membrane. Membrane characterization, membrane transport and membrane module, and system design are out of the scope of this entry. Examples of polymers that are currently used to manufacture industrial membranes are reviewed. They are mostly technical polymers that are mass-produced for purposes other than membrane manufacturing. The objective of this entry is, therefore, to review knowledge on polymeric properties that are relevant to membrane manufacturing, current technologies to prepare asymmetric membranes, and applications of polymeric membranes.

POLYMERS FOR MEMBRANES

Although membranes can be prepared from both ceramic and polymeric materials, ceramic materials have several advantages over polymeric materials, such as higher chemical and thermal stability. The market share of polymeric membranes is far greater than that of ceramic membranes because polymeric materials are easier to process and less expensive. A handful of technical polymers are currently used as membrane materials for 95% of all practical applications.[1] Polymeric materials that are used to prepare separation membranes are mostly organic compounds except for silicone rubbers that contain no carbon atoms.

Among others, chain flexibility/rigidity, crystallinity, hydrophilicity/hydrophobicity, molecular weight, chemical property, biodegradability, thermal property, mechanical property, and electrochemical property of polymers are considered to affect the performance of membranes most strongly. Moreover, these properties are often mutually interrelated.

Chain Flexibility/Rigidity

The presence of bulky pendant groups hinders the rotational motion around the main chain, while increasing the distance between them. Because of the increase in the interchain void space, small molecules tend to diffuse through the space more readily. On the other hand, the diffusion of large molecules through the interchain void space becomes more restricted because of the increase in chain rigidity. As a result, the selectivity of the polymeric membrane will increase, while the permeability of small molecules will also increase with an increase in the bulkiness of the pendant groups.[2,3] This is an important design criterion for gas separation membranes.

Crystallinity

Some polymers have a regular and symmetric nature in their molecular structures and tend to crystallize partially. Because the diffusion of permeant molecules through the crystalline region is far slower than in the surrounding noncrystalline region, increase in the degree of crystallinity leads to low permeability.

Hydrophilicity/Hydrophobicity

A measure for hydrophilicity/hydrophobicity of a polymer is given by the solubility parameter.[4]

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030631

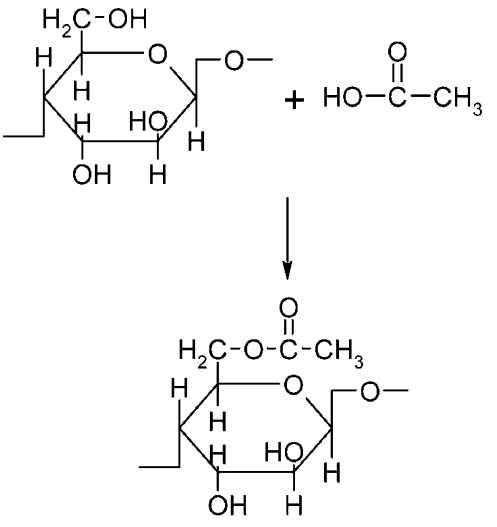
The solubility parameter is a parameter by which the nature and the magnitude of interaction forces working between molecules can be expressed. It is defined as the square root of cohesive energy density, $\Delta E/V$, where ΔE is the heat of vaporization (J/mol) and V is the molar volume (m^3/mol). Solubility parameters for some polymers are given in Table 1. Usually, a polymer of higher-solubility parameter is more hydrophilic. The interaction force working between two organic compounds, including polymers, becomes stronger when their solubility parameters are close to each other.

Molecular Weight

While the above three properties of polymers affect the permeation properties of the membrane directly, the effect of the molecular weight on permeation properties is still controversial, although there are some evidences that an increase in molecular weight enhances the permeation rate owing to an increase in chain stiffness.[5]

Chemical Property

Cellulose acetate is an ester of cellulose and acetic acid. Hence, hydrolysis takes place when the pH of the solution with which a cellulose acetate membrane is in contact is too high or too low, lowering the degree of acetylation, defined as the number of hydroxyl groups (total of three in one D-glucopyranose unit) that can be acetylated. Because a degree of acetylation above 2.5 is required for satisfactory salt rejection in seawater desalination, excessive hydrolysis results in poor membrane performance. The pH values between 5 and 7 should be maintained when cellulose acetate membranes are used.[6]



Formation of monosubstituted cellulose acetate

Aromatic polyamide, another polymeric material used for seawater desalination, can tolerate a wider pH range from 5 to 9. However, aromatic polyamide membranes are known to be susceptible to chlorination in the presence of chlorine in water.[3]

Another important chemical property of polymers is their ability to interact with organic solvents. Polymers should be dissolved in a solvent or a solvent blend when a membrane is prepared by the phase inversion technique or a thin polymer layer is coated on top of a substrate porous membrane. On the other hand, polymeric membranes are swollen by organic vapors or by organic solutes, if interaction between the polymer and the organic compounds is too strong, resulting in poor membrane performance. The strength of interaction between polymer and organic solvent is given by the Flory–Huggins (F–H) interaction

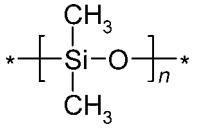
Table 1 Solubility parameters of polymers relevant for membrane preparation ($1\,cal^{1/2}/cm^{3/2} = 2.05\,J^{1/2}/cm^{3/2}$)

Polymer	Solubility parameter ($cal^{1/2}/cm^{3/2}$)	Polymer	Solubility parameter ($cal^{1/2}/cm^{3/2}$)
TFE	6.7	PEEK	12.8
PVDF	7.5	PES	13.9
PP	8.0	SPS	14.1
PE	8.6	PAN	14.4
PC	11.6	PEI	14.5
CTA	12.0	PA-300	15.0
NS-100	12.2	PA	15.9
PS (Udel)	12.6	PI	19.0
SPPO	12.6	PVA	19.1
CA-398	12.7	CE	24.1

coefficient.[7] The lower the coefficient, the stronger becomes the interaction. A relationship between solubility parameter and F–H interaction coefficient has been proposed.[8]

Thermal Property

Two states are distinguished for a polymeric material, the rubbery state and the glassy state. When a polymer is in a rubbery state, the polymer chain is very mobile because the segment can rotate freely along the segment bond. On the other hand, when a polymer is in a glassy state, the rotation is restricted by the presence of bulky pendant groups or by the strong interaction between chains. The transition from a glassy state to a rubbery state occurs at the glass transition temperature T_g, at which a sufficient amount of thermal energy is supplied to make the macromolecules more mobile. Many physical properties, such as density, tensile modulus, and permeability, show an abrupt change at T_g when the temperature is gradually changed. When T_g of a polymer is lower than room temperature, the polymer is in a rubbery state at room temperature and it is called rubbery polymer. On the other hand, when T_g of a polymer is higher than room temperature, the polymer is called glassy polymer. A typical rubbery polymer that is used for gas separation membranes is polydimethyl siloxane (PDMS). Gas permeability of PDMS is very high because of its high chain mobility. However, the selectivity is low.



Polydimethyl siloxane

Many polymers are used in their glassy states because of their enhanced selectivities in the glassy state. T_g should be sufficiently high for these polymers to be durable at a high operating temperature. Otherwise, the temperature may go beyond T_g and the selectivity of the membrane will dramatically decrease. Hence, T_g is one of the important criteria for glassy polymers. Some pertinent T_g values are listed in Table 2.[7,9]

Mechanical Property

The ratio of strain (deformation per unit length) caused by a stress (force per unit cross-sectional area) applied to a polymeric material is called Young's modulus. A sufficiently high Young's modulus is desired for a polymer when it is spun to hollow fibers.

Table 2 Glass transition temperatures of some polymers

Polymer	T_g(°C)	Polymer	T_g(°C)
PDMS	−125	PEEK	145
PE	−120	PC	150
Natural rubber	−72	PS	195
PVDF	−40	PEI	210
PP	−10	PPO	218
CA	80	PES	230
PVA	85	PAI[a]	295
PAN	120	PI	300
PTFE	126		

[a]PAI, polyamideimide.

Electrochemical Property

Membranes for electrodialysis and polymer electrolyte membrane fuel cell (PEMFC) have electric charges. Most of the nanofiltration membranes also carry negative charges. The content of electric charge in a polymer is given by ion-exchange capacity (meq (milliequivalent)/g of dry polymer).

PREPARATION OF POLYMERIC MEMBRANES

Membranes Without Asymmetric Structures

Track etching

A sheet of polycarbonate film moves underneath a radiation source and is irradiated by high-energy particles. The spots that are subjected to bombardment of the particles are degraded or chemically altered during this process. Then, the film undergoes an etching process in an alkaline or H_2O_2 bath, where the polymer is etched along the path of high-energy particles. Cylindrical pores of very uniform sizes are obtained by this method. The pore sizes are in the range of microfiltration (MF) membranes and the porosity is relatively small.[10]

Precipitation from the vapor phase

This method was developed by Zsigmondy in the early part of the last century. A cast polymer solution that consists of polymer and solvent is brought into nonsolvent vapor environment saturated with solvent vapor. While saturated solvent vapor suppresses the evaporation of solvent from the film, nonsolvent vapor diffuses into the film causing polymer coagulation. The formation of the dense skin layer does not occur. Pore sizes are uniform, perpendicular to the membrane surface.[11]

Asymmetric Structure of the Membrane

Most of the membranes that are used in industrial separation processes have an asymmetric structure. Membranes with asymmetric structures are called asymmetric membranes. Fig. 1 shows schematically a typical cross-sectional view of an asymmetric membrane.[12] As shown in the figure, an asymmetric membrane consists of two layers, i.e., one is a very thin dense layer at the top of the membrane and the other is a porous sublayer underneath the top dense layer (also called top skin layer). While the top dense layer governs the permeation properties of the membrane, the porous sublayer only provides the membrane with mechanical strength. When the material of the top skin layer and the porous sublayer are the same, the membrane is called integrally skinned asymmetric membrane. This type of membrane is made by the dry–wet phase inversion technique. When the polymer for the top skin layer is different from the polymer for the porous sublayer, the membrane is called composite membrane. The advantage of the composite membrane over the integrally skinned asymmetric membrane is that the material for the top skin layer and for the porous sublayer can be chosen separately to optimize the overall performance. This type of membrane is made by coating a thin layer on top of the surface of a porous substrate. Various coating techniques are available but the interfacial in situ polymerization method has been proven to be commercially the most successful.

Phase inversion technique—Preparation of integrally skinned asymmetric membranes

Phase inversion is a process in which a polymer is transformed from a liquid to a solid state. There are a number of methods to achieve phase inversion. Among others, the dry–wet phase inversion technique and the temperature induced phase separation (TIPS) are most commonly used in the industrial membrane manufacturing. The dry–wet phase inversion technique was applied by Loeb and Sourirajan in their development of the first cellulose acetate membrane for seawater desalination.[13] Therefore, this method is often called the Loeb–Sourirajan method. According to the Loeb–Sourirajan method, a polymer solution is prepared by mixing polymer, solvent, and sometimes even nonsolvent. The solution is then cast on a suitable surface by a doctor blade to a thickness of about 250 μm. After partial evaporation of the solvent, the cast film is immersed in a bath of nonsolvent medium, often called gelation medium. Because of a sequence of two desolvation steps, i.e., evaporation of solvent and solvent–nonsolvent exchange in the gelation bath, solidification of polymer film takes place. It is desirable to choose a solvent of strong dissolving power with high volatility. During the first step of desolvation by solvent evaporation, a thin skin layer of solid polymer is formed instantly at the top of the cast film owing to the loss of solvent. In the solvent–nonsolvent exchange process that follows, nonsolvent diffuses into, while solvent diffuses out of, the polymer solution film through the thin solid layer. The change in the composition of the polymer solution film during the solvent–nonsolvent exchange process, often called a composition path, is illustrated schematically in a triangular diagram that involves polymer–solvent–nonsolvent (Fig. 2).

At some moment, the content of solvent in the solution film becomes so low that the solvent no longer

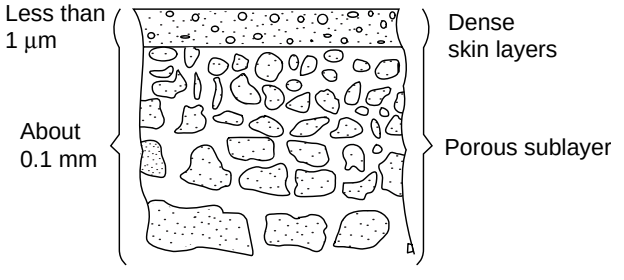


Fig. 1 Schematic representation of the cross section of an asymmetric membrane. (From Ref.[12].)

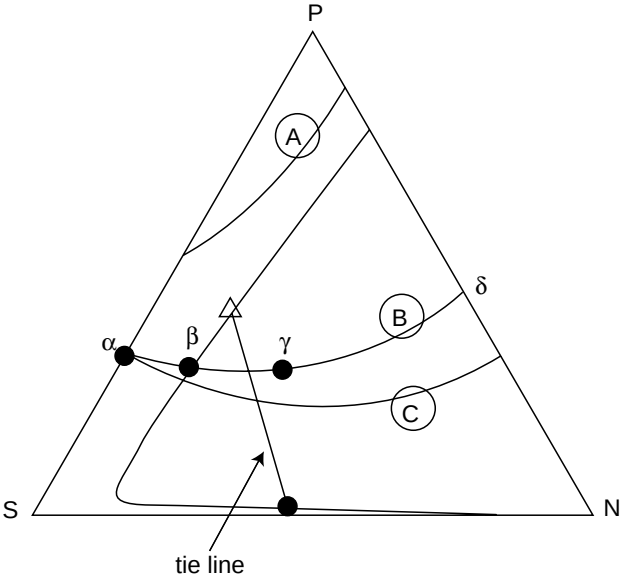


Fig. 2 Examples of the composition path on a triangular diagram. P = polymer, S = solvent, N = nonsolvent; composition paths A, B, and C depend on the ratio of solvent/nonsolvent flux rate during the solvent exchange. α, composition at the beginning of solvent exchange; β, composition at the point of crossing the phase boundary; γ, composition at the point of gelation; δ, composition at the end of solvent exchange. (From Ref.[12].)

is able to hold the polymer in one phase. Phase separation takes place at this moment, forming droplets of one liquid phase dispersed in the other continuous liquid phase. The moment of phase separation and the size and the number of the dispersed droplets depend on the nature of solvent and nonsolvent and the polymer solution composition. The control of the number and the size of the droplets will eventually control the structure of the porous substrate.[14]

The thin layer of solid polymer that forms during the first evaporation step becomes the top skin layer that will govern the selectivity and the flux of the membrane, while the porous structure that forms during the solvent–nonsolvent extraction step becomes the porous sublayer, providing the mechanical strength. Hence, the membrane obtained by the dry–wet phase inversion process is an integrally skinned asymmetric membrane. The top skin layer can also be made porous by lowering the polymer concentration in the casting solution and the solvent evaporation period. This is called, hereafter, porous skin layer. Ultrafiltration (UF) membranes have a porous skin layer. The asymmetric membranes can also be made in tubular form using a casting bob assembly and hollow fibers can be spun using a hollow fiber spinneret.[15]

Thermally induced phase separation

In this process phase inversion is introduced by lowering the temperature of the polymer solution. A polymer is mixed with a substance that acts as a solvent at a high temperature and the polymer solution is cast into a film. When the solution is cooled, it enters into an immiscible region due to the loss of solvent power. Liquid–liquid demixing occurs and the solution is separated into two phases, i.e., the polymer-lean phase is dispersed as droplets in the polymer-rich phase. Further, cooling causes gelation of polymer. Because the solvent is usually nonvolatile, it must be removed with a liquid that is miscible with the solvent but not miscible with the polymer. The membranes made by the TIPS method have pore sizes in the range of 0.1 and 1 μm and the pore structure is uniform in the depth direction.[16]

Membrane Surface Coating for Composite Membranes

Dip coating

An integrally skinned asymmetric membrane with a porous skin layer (hereafter called substrate membrane) is prepared from a polymer solution by applying the dry–wet phase inversion method and dried according to the method described later, before being dipped into a bath containing a dilute solution of another polymer. When the membrane is taken out of the bath, a thin layer of coating solution is deposited on top of the substrate membrane. The solvent is then removed by evaporation, leaving a thin layer of the latter polymer on top of the substrate membrane.

Interfacial in situ polymerization

This method, developed by Cadotte and coworkers of Film Tech in the 1970s, is currently most widely used to prepare high-performance reverse osmosis and nanofiltration membranes.[17] A thin selective layer is deposited on top of a porous substrate membrane by interfacial in situ polycondensation. There are a number of modifications of this method primarily based on the choice of the monomers.[18] However, for the sake of simplicity, the polycondensation procedure is described by a pair of diamine and diacid chloride monomers.

A diamine solution in water and a diacid chloride solution in hexane are prepared. A porous substrate membrane is then dipped into the aqueous solution of diamine. The pores at the top of the porous substrate membrane are filled with the aqueous solution in this process. The membrane is then immersed in the diacid chloride solution in hexane. Because water and hexane are not miscible, an interface is formed at the boundary of the two phases. Polycondensation of diamine and diacid chloride will take place at the interface, resulting in a very thin layer of polyamide. The preparation of composite membranes by the interfacial in situ polycondensation is schematically presented in Fig. 3.

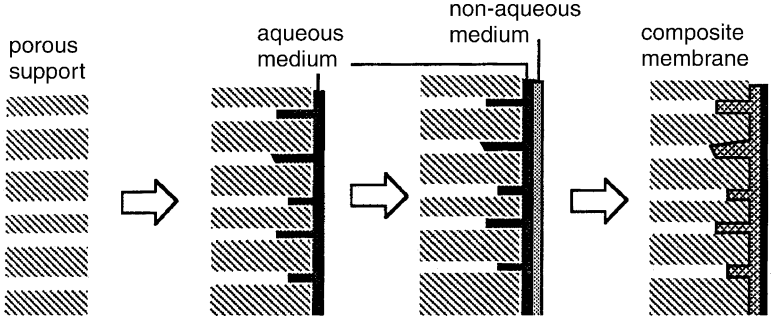


Fig. 3 Schematic representation of interfacial polycondensation to form a composite membrane. (From Ref.[3].)

There are a number of combinations for the choice of diamine and acid chloride monomers. For example, if trimesoyl chloride, which has three –COCl groups in an aromatic ring, is mixed with phthaloyl chloride, which has two –COCl groups, crosslinking will form between two main chains. Unreacted –COCl will become –COOH upon contact with water and the membrane will become negatively charged. Monomers with reactive groups other than amine and acid chloride can also be used.

Other Methods of Membrane Surface Modification

As mentioned above, the top skin layer governs the performance of a separation membrane. The surface deposition of contaminants from solutions or from gas mixtures is also affected by the surface properties of the membrane. This is particularly important when decline in the membrane flux with a prolonged operating period is observed, because it is often caused by the contaminant deposition. Hence, many attempts have been made to modify the membrane surface, aiming at prevention of contaminant deposition and maintenance of high flux. Several methods of surface modification are described below.

Chemical modification

The surface of a membrane can be modified by chemical reactions. For example, when the surface of a polyamide composite membrane is brought into contact with a strong hydrofluoric acid solution, the top polyamide layer becomes slightly thinner by a chemical reaction with hydrofluoric acid. As a result the flux increases considerably while the rejection of sodium chloride is unchanged or slightly increased.[19]

Plasma polymerization

When a vacuum is maintained inside a tubular reactor and a high-frequency field is applied outside, a glow discharge is generated inside the reactor (see Fig. 4). A plasma that consists of various ions, radicals, electrons, and molecules is formed in the glow discharge. When a porous substrate membrane is placed into the plasma, the surface of the membrane is subject to various changes corresponding to the property of plasma. The substrate surface can be etched and/or chemically active sites are introduced to the surface, and, upon contact with organic compounds, an irregular polymerization may occur at the substrate surface. This is called plasma polymerization.[20]

Graft polymerization

The surface of a porous substrate membrane is irradiated with γ-ray, which causes generation of radicals on the membrane surface. Then, the membrane is immersed into a monomer solution. The graft polymerization of the monomers is initiated at the membrane surface. By choosing a very hydrophilic monomer, the hydrophilicity of the surface is increased considerably.

Surface modification by surface modifying macromolecules

It is known that, in a polymer blend, thermodynamic incompatibility between polymers usually causes demixing of polymers. If the polymer is equilibrated in air, the polymer with the lowest surface energy (hydrophobic polymer) will concentrate at the air interface and reduce the system's interfacial tension as a consequence. The preferential adsorption of a polymer of lower surface tension at the surface was confirmed by a number of researchers for a miscible blend of two different polymers. Based on this concept, surface modifying macromolecules (SMMs) as surface-active additives were synthesized and blended into polymer solutions of polyethersulfone (PES). Depending on the hydrophobic or hydrophilic nature of the SMM, the membrane surface becomes either more hydrophobic or more hydrophilic than the base polymeric material.[21–23]

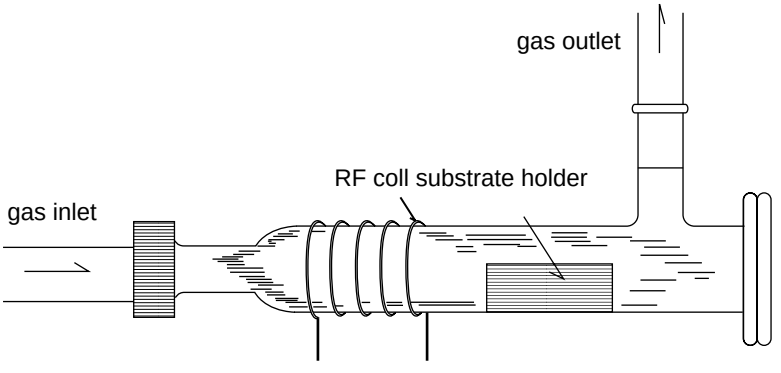


Fig. 4 Tubular reactor for plasma polymerization. (From Ref.[12].)

Membrane Drying

The wet cellulose acetate membranes prepared for reverse osmosis purposes can be used for gas separation when they are dried. The water in the cellulose acetate membrane cannot be evaporated in air, however, because the asymmetric structure of the membrane will collapse. Instead, the multistage solvent exchange and the evaporation method is applied. In this method, the water in the membrane is first replaced by a water-miscible solvent such as ethanol. Then, the first solvent is replaced by a second volatile solvent such as hexane. The second solvent is subsequently air-evaporated to obtain a dry membrane.[24,25] The reason for replacing water to hexane is to reduce the capillary force inside the pore so that the pore will not collapse during the drying process.

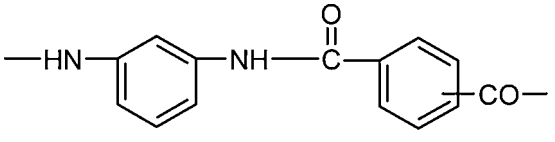
POLYMERIC MEMBRANES FOR MEMBRANE SEPARATION PROCESSES

Reverse Osmosis/Nanofiltration/ Ultrafiltration/Microfiltration

Reverse osmosis membranes

Cellulose acetate is the material for the first-generation reverse osmosis (RO) membranes. The announcement of cellulose acetate membranes for seawater desalination by Loeb and Sourirajan in 1960 triggered the applications of membrane separation processes in many industrial sectors. Cellulose acetate membranes are prepared by the dry–wet phase inversion technique.

Another polymeric material for RO is aromatic polyamide.[26]



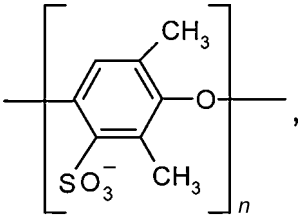
Aromatic polyamide polymer

As shown above, aromatic rings are connected by an amide linkage, –CONH–. While the aromatic ring attached to –NH– is *meta*-substituted, the ring attached to –CO– is the mixture of *meta*- and *para*-substitutions, which gives more flexibility to the polymeric material. Aromatic polyamide remains one of the most important materials for RO membranes because the thin selective layer of composite membranes is aromatic polyamide synthesized by interfacial in situ polymerization.

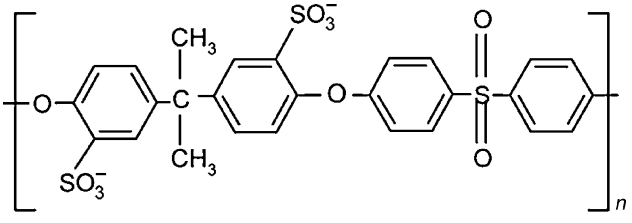
Nanofiltration membrane

Most of the nanofiltation (NF) membranes are negatively charged. As shown in interfacial polycondensation, trimesoyl (triacid) chloride is mixed with phthaloyl (diacid) chloride in the acidic component of polycondensation reaction. Although most of the carboxylic groups are consumed to form cross-linking, a small portion of carboxylic groups does not participate in the cross-linking reaction, becoming the source of the electric charge. Because –COOH becomes $–COO^-$ upon dissociation, the membranes are negatively charged. Because of the negative charge, anions are preferentially rejected by NF membranes.

Another method of preparing NF membranes is to dip-coat a thin layer of sulfonated polyphenylene oxide (SPPO), sulfonated polysulfone (SPS), or carboxylated polysulfone on a porous substrate membrane.[27–29]



SPPO



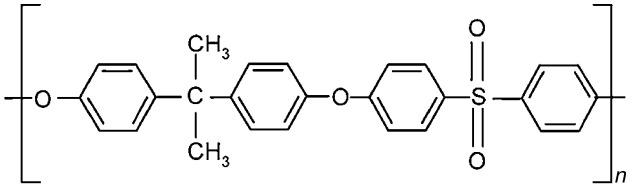
SPS

Sulfonic acid groups in SPPO and SPS also become negatively charged with $–SO_3^-$ groups upon dissociation. Sulfonic acid is a stronger acid than carboxylic acid.

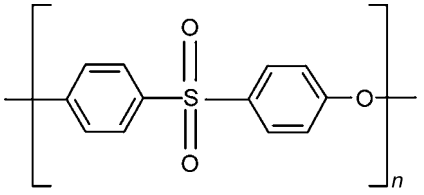
Ultrafiltration membrane

In contrast to the polymeric materials for RO and NF membranes, for which the macromolecular structures have much to do with their permeation properties such as salt rejection characteristics, the choice of membrane material for UF does not depend on the material's influence on the permeation properties. Membrane permeation properties are largely governed by the pore sizes and the pore size distributions of UF membranes. Rather, the thermal, chemical, mechanical, and biological stability is considered of greater importance.

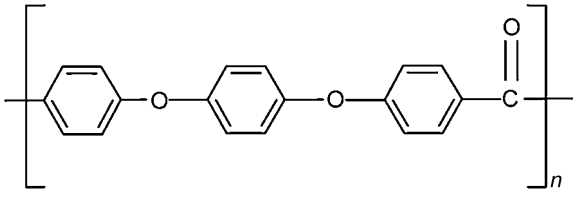
Typical UF membrane materials are polysulfone (PS), polyethersulfone (PES), polyetheretherketone (PEEK), cellulose acetate (CA), polyacrylonitrile (PAN), polyvinylidene fluoride (PVDF), polyimide (PI), and polyetherimide (PEI):



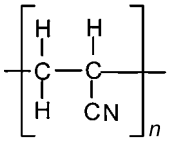
PS



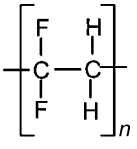
PES



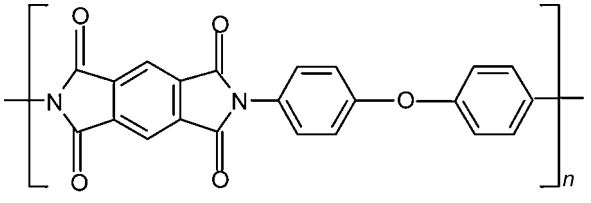
PEEK



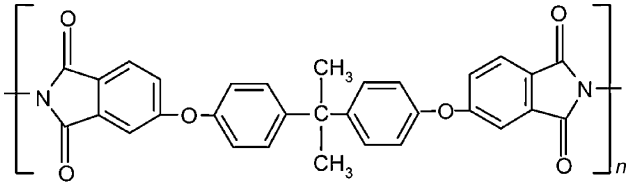
PAN



PVDF



PI, Kapton

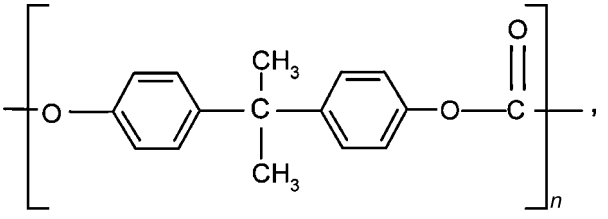


PEI, Ultem

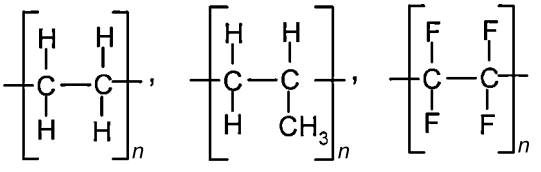
All of the above polymers have T_gs higher than 145°C except for cellulose acetate. They are also stable chemically and mechanically. Also, their biodegradability is low. The membranes are made by the dry–wet phase inversion technique.

Microfiltration membranes

Polymeric materials for MF membranes cover a very wide range from relatively hydrophilic to very hydrophobic materials. Typical hydrophilic materials are polysulfone (PS), polyethersulfone (PES), cellulose (CE) and cellulose acetate (CA), polyamide (PA), polyimide (PI), polyetherimide (PEI), and polycarbonate (PC). Typical hydrophobic materials are polyethylene (PE), polypropylene (PP), polytetrafluoroethylene (PTFE, Teflon), and polyvinylidene fluoride (PVDF).



PC

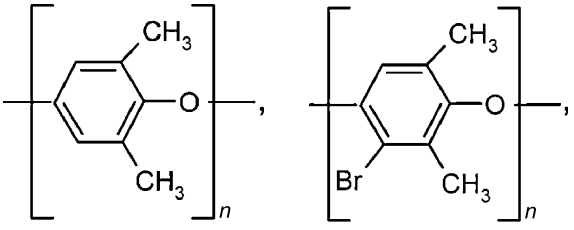


PE PP PTFE

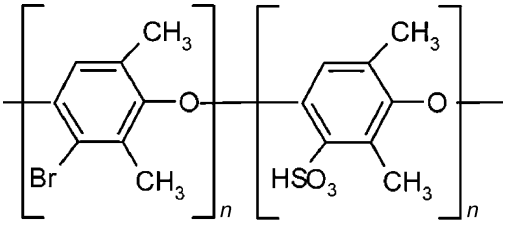
Hydrophilic MF membranes can be made by the dry–wet phase inversion technique. The latter technique is also applicable in making PVDF membranes. On the other hand, other hydrophobic MF membranes are made by the TIPS technique. In particular, semicrystalline PE, PP, and PTFE are stretched parallel to the direction of film extrusion so that the crystalline regions are aligned to the direction of stretch, while the noncrystalline region is ruptured, forming long and narrow pores. Hydrophobic membranes do not allow penetration of water into the pore until the transmembrane pressure drop reaches a threshold pressure called liquid entry pressure of water. These membranes can therefore be used for membrane distillation. Track-etching method is applied to make MF membranes from PC.

Gas and Vapor Separation

Gas and vapor separation membranes are classified into two categories. In the first one, rubbery polymers such as silicone rubber are used to take advantage of their high permeabilities, even though selectivities are rather moderate. Production of enriched oxygen for medical purposes is performed by this type of membrane with an oxygen/nitrogen selectivity of about 2. Asymmetric membranes made from glassy polymers such as cellulose acetate (CA), aromatic polyamide (PA), aromatic polyimides (PI), and polyphenylene oxide (PPO) and its derivatives belong to the second category.



PPO PPOBr



SPPOBr

These asymmetric membranes are made by the dry–wet phase inversion technique. Membranes must be dried before being used. Solvent exchange is necessary to dry cellulose acetate membranes. These membranes take advantage of the high selectivity of glassy polymers. The selective dense layer at the top of the membrane must be very thin so that a high flux can be achieved. They are used in a wide range of industrial gas separation processes such as hydrogen recovery from various chemical syntheses, sour gas removal from natural gas, and production of nitrogen-enriched air. For the asymmetric membranes to be effective in gas separation, the thin selective layer at the top of the membrane should be perfect and defect-free. This requirement is more stringent in gas separation membranes than in liquid separation membranes because defective pores cannot be automatically closed when the surface is in contact with dry gas. In contrast, the defective pores of RO and PV membranes can be closed by the swelling of the top skin layer when it is brought into contact with the feed liquid.

Because it is difficult to make a selective skin layer perfectly defect-free, a method was proposed by Henis and Tripodi to seal defective pores. Their method was applied to asymmetric PS membranes, which led to the production of commercial prism membranes.[30]

According to the method, a relatively thick silicone rubber layer is coated on a thin selective layer of an asymmetric PS membrane. The thickness of silicone rubber is about 1 μm while the effective thickness of the selective PS layer is 1/10 of 1 μm. While being coated, silicone rubber penetrates into the pores to plug them (Fig. 5). Thus, the feed gas is not allowed to leak through the defective pores. The selectivity of the membrane approaches that of the defect-free PS layer. Moreover, because the permeabilities of silicone rubber for gases are orders of magnitude higher than those of PS, the permeation rate is not affected very much even when a relatively thick silicone rubber layer is coated.

Membranes for vapor removal from air have a structure similar to the prism membrane, but they are prepared on a different principle.[31] Aromatic PEI is used to produce a porous substrate membrane by the dry–wet phase inversion method. This polymer was chosen over PS/PES because of the higher durability of PEI to organic vapors. Unlike an asymmetric PS substrate for the prism membrane, the top layer of asymmetric PEI membrane has a large number of pores, the size of which is equivalent to those of UF membranes. When a layer of silicone rubber is coated on the top layer of the porous substrate membrane, the silicone rubber layer will govern the selectivity and the porous support will provide only mechanical strength to the composite membrane. Because the permeabilities of water and organic vapors through the silicone

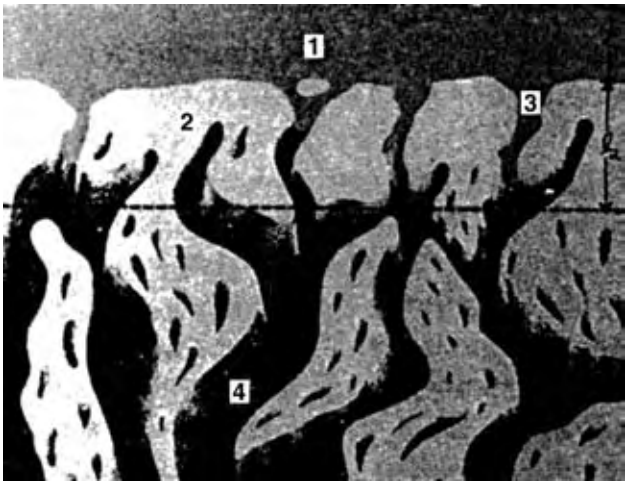
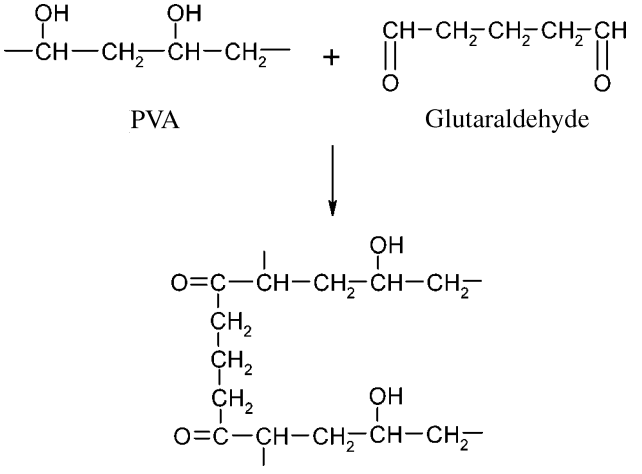


Fig. 5 Cross-sectional view of prism membrane. 1: silicone rubber layer; 2: skin layer of polysulfone; 3: pores filled with silicone rubber; 4: pores in the porous substrate membrane. (From Ref.[12].)

rubber layer are much greater than those of oxygen and nitrogen, these membranes are effective in dehumidification of air and removal of organic vapors from air.

Membranes for Pervaporation

Pervaporation (PV) membranes were developed for the dehydration of ethanol and other organic solvents. Therefore, the dense selective layer is made of polyvinyl alcohol, which is one of the most hydrophilic materials (see Table 2). Water is preferentially sorbed to polyvinyl alcohol and also preferentially transported. To suppress the excessive swelling of polymer in water, polyvinyl alcohol is partially cross-linked by dialdehydes such as glutaraldehyde.[32]



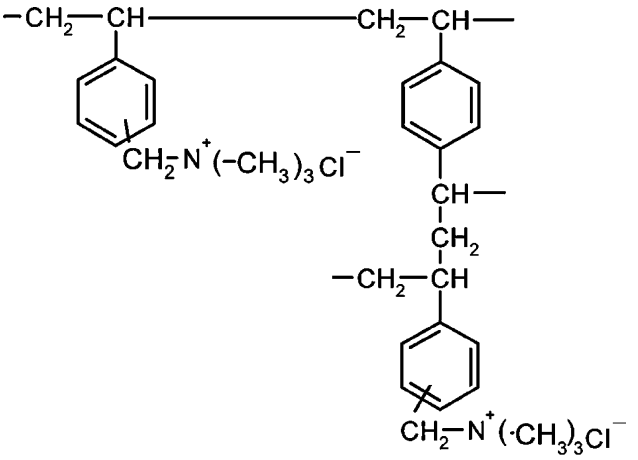
PVA Glutaraldehyde

Cross-linked PVA

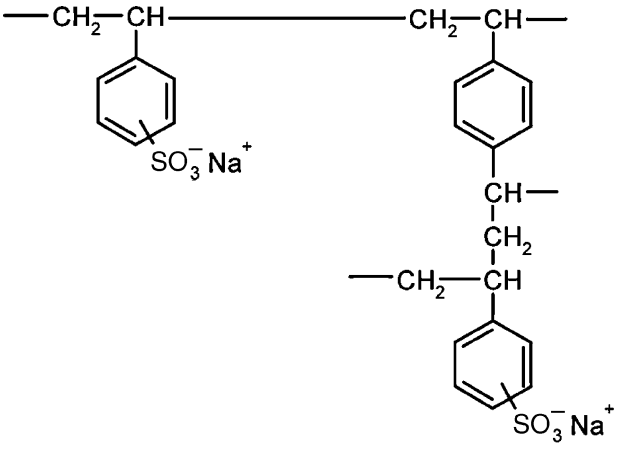
The dense polyvinyl alcohol layer is supported by a porous PAN substrate membrane. Polyelectrolyte material and chitosan, a natural product, are also potentially useful for dehydration by PV.[33,34] Silicone rubber membrane developed for the removal of organic vapors from air can also be used for the removal of volatile organic compounds (VOCs) from water by PV.[32] Because of the high hydrophobic nature of silicone rubber (see Table 2), VOCs are preferentially sorbed and transported through the membrane.

Electrodialysis

Membranes for electrodialysis (ED) are either positively or negatively charged. When a membrane is positively charged, it is called anion exchange membrane because only anions are allowed to permeate through the membrane, while a negatively charged membrane is called cationic membrane because only cations are allowed to permeate through the membrane. The base polymeric material is polystyrene cross-linked by divinylbenzene. Quarternary ammonium cations are attached to some aromatic rings of anionic membranes, while sulfonic groups or carboxylic groups are attached to some aromatic rings of cationic membranes.[35]



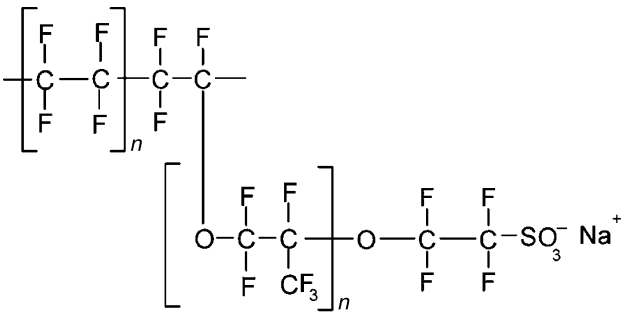
Anion selective membrane



Cation selective membrane

Fuel Cells

The development of new polymeric materials for polymer electrolyte fuel cell is one of the most active research areas, aiming at the new energy sources for electric cars and other devices. The mainstream of the material research for fuel cell is perfluoroalkyl sulfonic acid membranes such as Nafion, Acipex, and Flemion. The most well-known one is Nafion of Du Pont, which is derived from copolymers of tetrafluoroethylene and perfluorovinyl ether terminated by a sulfonic acid group.[35] Protons, when dissociated from the sulfonic acid groups in aqueous environment, become mobile and the membrane becomes a proton conducting electrolyte membrane.



Nafion

CONCLUSIONS

Many new polymers have been synthesized and tested for their permeation properties aiming at improvement in membrane performance for various applications. These efforts seem to continue in the future. However, only a handful of polymers, as shown in this entry, are currently being used as the materials for commercial membranes and they are not necessarily the polymers of the best permeation properties. This is mainly due to the cost factor that governs the membrane market. Transport properties of membranes are, on the other hand, known to be primarily governed by the membrane surface. Surface contamination, which may lead to deterioration in membrane performance, is also known to be governed by the membrane surface properties. Considering that only a small amount of polymer is required for surface coating, the future direction of R and D efforts will be focused on the development of new methods of surface coating and surface modification. This will allow us to utilize fully the potential of polymers for membrane materials.

REFERENCES

1. Peinemann, K. *Next Generation Membrane Materials*, Abstracts of the 15th Annual Meeting, Honolulu, HI, June 26–30, 2004.
2. Kim, T.H.; Koros, W.J.; Husk, G.R.; O'Brien, K.C. Relationship between gas separation properties of aromatic polyimides. J. Membr. Sci. **1988**, *37*, 45–62.
3. Mulder, M. *Basic Principles of Membrane Technology*; Kluwer Academic: Dordrecht, The Netherlands, 1991.
4. Van Krevelan, D.W. *Properties of Polymers*; Elsevier: Amsterdam, 1976.
5. Polotskaya, G.A.; Agranova, S.A.; Antonova, T.A.; Elyashevich, G.K. Gas transport and structural features of sulfonated poly(phenylene oxide). J. Appl. Polym. Sci. **1997**, *66*, 1439–1443.
6. Sourirajan, S. *Reverse Osmosis*; Academic Press: New York, 1970; 83.
7. Schuld, N.; Wolf, B.A. Polymer-solvent interaction parameters. In *Polymer Handbook*; Brandrup, J., Immergut, E.H., Grulke, E.A., Abe, A., Bloch, D.R., Eds.; Wiley-Interscience: New York, 1999; VII-247–VII-264.
8. Barton, A.F.M. *Handbook of Solubility Parameters and Other Cohesion Parameters*; CRC Press: Boca Raton, 1983; 265–266.
9. http://www.azom.com/details.asp?ArticleID=83&head=Thermoplastics%2B-%2BAn%2BI, page 4 of 5.
10. Zeman, L.J.; Zydney, A.L. *Microfiltration and Ultrafiltration Principles and Applications*; Marcel Dekker: New York, 1996; 164–165.
11. Zsigmondy, R.; Backman, W.Z. Seawater demeralization by means of an osmotic membrane. Anorg. Allgem. Chem. **1918**, *103*, 119.
12. Matsuura, T. *Synthetic Membranes and Membrane Separation Processes*; CRC Press: Boca Raton, 1994; 15–30.
13. Loeb, S.; Sourirajan, S. Adv. Chem. Ser. **1963**, *38*, 117.
14. Kesting, R.E. *Synthetic Polymeric Membranes*; McGraw-Hill: New York, 1971; 122–123.
15. Sourirajan, S.; Matsuura, T. *Reverse Osmosis/Ultrafiltration Process Principles*; National Research Council of Canada: Ottawa, 1985; 802–805.
16. Kesting, R.E. Phase inversion membranes. In *Materials Science of Synthetic Membranes*; ACS Symposium Series 269; Lloyd, D.R., Ed.; American Chemical Society: Washington, DC, 1985; 131–164.
17. Rozelle, L.T.; Cadotte, J.E.; Cobian, K.E.; Kopp, C.V., Jr. Nonpolysaccharide membranes for reverse osmosis: NS-100 membranes. In *Reverse Osmosis and Synthetic Membranes, Theory—Technology—Engineering*; Sourirajan, S., Ed.; National Research Council of Canada: Ottawa, 1977; 249–261.
18. Petersen, R.J. Composite reverse osmosis and nanofiltration membranes. J. Membr. Sci. **1993**, *83*, 81–150.
19. Kulkarni, A.; Mukherjee, D.; Mukherjee, D.; Gill, W.N. Reprocessing hydrofluoric acid etching solutions by reverse osmosis. Chem. Eng. Commun. **1994**, *129*, 53–68.
20. Hirotsu, T. Graft polymerized membranes of methacrylic acid by plasma for water-ethanol pervaporation. Ind. Eng. Chem. Res. **1987**, *26*, 1287–1290.
21. Suk, D.E.; Chowdhury, G.; Narbaitz, R.M.; Santerre, J.P.; Matsuura, T.; Glazier, G.;

Deslandes, Y. Study on the kinetics of surface migration of surface modifying macromolecules in membrane preparation. Macromolecules **2002**, *35*, 3017–3021.
22. Khayet, M.; Suk, D.E.; Narbaitz, R.M.; Santerre, J.P.; Matsuura, T. Study on surface modification by surface-modifying macromolecules and its applications in membrane-separation processes. J. Appl. Polym. Sci. **2003**, *89*, 2902–2916.
23. Hester, J.F.; Banerjee, P.; Won, Y.-Y.; Akthakul, A.; Acar, M.H.; Mayes, A.M. ATRP of amphiphilic graft copolymers based on PVDF and their use as membrane additives. Macromolecules **2002**, *35*, 7652–7661.
24. Lui, A.; Talbot, F.D.F; Sourirajan, S.; Fouda, A.E.; Matsuura, T. Studies on gas transport through dry cellulose acetate membranes prepared by solvent exchange technique. Sep. Sci. Technol. **1988**, *23*, 1839.
25. Gantzel, P.K.; Merten, U. Gas separation with high flux cellulose acetate membranes. Ind. Eng. Chem. Process Des. Dev. **1970**, *9*, 331–332.
26. Hoehn, H.H. Aromatic polyamide membrane. In *Materials Science of Synthetic Membranes*; ACS Symposium Series 26; Lloyd, D.R., Ed.; American Chemical Society: Washington, DC, 1985; 81–98.
27. Matsuura, T. Reverse osmosis and nanofiltration by composite polyphenylene oxide membranes. In *Polyphenylene Oxide and Modified Polyphenylene Oxide Membranes, Gas, Vapor and Liquid Separation*; Chowdhury, G., Kruczek, B., Matsuura, T., Eds.; Kluwer Academic: Boston, 2001; 181–212.
28. Allegrezza, A.E., Jr.; Parekh, B.S.; Parise, P.L.; Swiniarski, E.J.; White, J.L. Chlorine resistant polysufone reverse osmosis modules. Desalination **1987**, *64*, 285–304.
29. Guiver, M.D.; Tremblay, A.Y.; Tam, C.M. Reverse osmosis membranes from novel hydrophilic polysulfones. In *Advances in Reverse Osmosis and Ultrafiltration*; Matsuura, T., Sourirajan, S., Eds.; National Research Council of Canada: Ottawa, 1989; 53–70.
30. Henis, J.M.S.; Tripodi, M.K. Composite hollow fiber membranes for gas separation: resistance model approach. J. Membr. Sci. **1981**, *8*, 233–246.
31. Behling, R.-D.; Ohlrogge, K.; Peinemann, K.-V. The separation of hydrocarbons from waste vapor streams. In *Membrane Separations in Chemical Engineering*; A.I.Ch.E. Symposium Series 272; Fouda, A.E., Hazlett, J.D., Matsuura, T., Johnson, J., Eds.; American Institute of Chemical Engineering: New York, 1989; 68–73.
32. Koops, G.H.; Smolders, C.A. Estimation and evaluation of polymeric materials for pervaporation membranes. In *Pervaporation Membrane Separation Processes*; Huang, R.Y.M., Ed.; Elsevier: New York, 1991; 253–278.
33. Tsuyumoto, M.; Karakane, H.; Maeda, Y.; Tsugaya, H. Development of polyion complex hollow fiber membrane for separation of water ethanol mixtures. Desalination **1991**, *80*, 139–158.
34. Feng, X.S.; Huang, R.Y.M. Pervaporation with chitosan membranes I. Separation of water from ethylene glycol by a chitosan/polysulfone composite membrane. J. Membr. Sci. **1996**, *116*, 67–76.
35. Strathmann, H. Ion exchange membranes. In *Membrane Handbook*; Ho, W., Sirkar, K.K., Eds.; Van Nostrand Reinhold: New York, 1992; 230–245.

Polymerization Reactions: Modeling, Design, and Control

Kyu Yong Choi
Department of Chemical Engineering, University of Maryland, College Park, Maryland, U.S.A.

INTRODUCTION

Synthetic polymers are of enormous industrial importance, and we can hardly think of modern life without polymers. Although many of the commercially important polymers have been around for more than 50 years [e.g., poly(vinyl chloride), nylons, polyesters, polystyrene, polyethylene, etc.], their manufacturing processes are not completely free of technical problems or new challenges because of the rapidly changing and diversifying market environment, and the pressure for cost reductions and new product developments. Polymer manufacturers always face the problem of supplying polymer products that meet the customer specifications at the lowest possible production cost. There is also a continuing need for new polymeric materials and new applications in electronics, biotechnology, energy, and environmental industries. To meet such challenges, the role of polymer science and engineering will continue to be important.

Polymerization reactors are known to be difficult to design and control because of inherent complexities in polymerization kinetic and mechanisms, physical transport effects, and strong nonlinear reactor behaviors in industrial polymerization processes. Moreover, many of the process variables that affect important product quality indices cannot be measured online or can be measured only at low sampling frequencies with time delays, making the reactor control difficult. Many polymer manufacturers find that a better understanding of their existing polymerization reactions and processes would enable them to design more efficient polymerization technology and to develop improved or new products. In this regard, one of the important tools that would allow the polymerization manufacturers to advance polymerization technology is a quantitative process model. Indeed, there is a growing consensus in the polymer industry that a mathematical modeling of a polymerization process offers an invaluable opportunity to improve product quality and productivity, thereby enhancing the overall competitiveness. Several commercial process simulation packages utilizing computer-aided design tools are available for a variety of industrial polymerization processes.

BACKGROUND

Unlike other chemical reaction processes, polymerization reaction processes offer some unique problems or issues that must be considered in process design and control. They include:

1. Complex reaction kinetics and phase behavior: In many polymerization processes, the reaction mixture undergoes a significant phase change that affects the polymerization rate, polymer properties, and reactor operations (e.g., viscosity increase, particle formation, precipitation, etc.).
2. Material mixing and conveying: Serious nonidealities in micromixing and macromixing may occur at high conversions and high viscosities. Often, a large amount of mechanical energy is required for uniform mixing and product conveying. Fluid or particle sticking, or fouling of the reactor surfaces can severely affect the reactor operations.
3. Heat removal: Removal of reaction heat from a highly viscous polymeric fluid or a heterogeneous reaction mixture is often a critical reactor design and operational problem. In many industrial exothermic polymerization processes, reactor thermal runaway is the most serious potential hazard.
4. Polymer product quality control: The molecular architecture of a polymer is very sensitive to reaction environment. The actual customer specifications are often represented by nonmolecular parameters (e.g., tensile strength, impact strength, color, crack resistance, thermal stability, etc.) that must be somehow related to fundamental polymer properties such as molecular weight distribution, composition, composition distribution, branching, crosslinking, etc. Many of these properties are influenced by more than one reaction or process variable and hence, one needs to understand complex and nonlinear relations between reaction variables and fundamental polymer properties. The lack of online

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120018072

sensors to measure polymer properties makes the monitoring and control of polymerization reactors a special challenge.

A comprehensive review of the broad aspects of polymerization process modeling and its applications is not the objective here. In this entry, some critical issues related to the modeling, design, and control of polymerization reactors are discussed with some examples to illustrate modeling techniques and their applications to polymerization process optimization and control.

MODELING OF POLYMERIZATION REACTIONS

A modeling of a polymerization reactor process is needed for many reasons. Most of all, there is a need to develop a better understanding of the polymerization kinetics, chemical and physical transitions, and reactor behaviors under a given set of polymerization conditions and to use such an understanding to design improved reactor or process operations. For example, one can examine through process modeling how operating conditions of the reactor affect the polymerization rate and polymer properties. A reaction model is also useful in optimizing the process or reactor operations, and in designing advanced reactor control systems. For example, a model can be used to develop a model-based reactor control system or to evaluate various process control options even before a commercial plant is built. In modern polymerization plants, detailed polymerization process models are often used in conjunction with stochastic state estimation techniques (e.g., Kalman filter, extended Kalman filter, observer, etc.) to provide online estimates of process variables or polymer properties. Polymerization process models can also be used for training and education of production personnel. The research on modern mathematical modeling of polymerization reactions and polymerization processes was pioneered by two research groups: Ray and his coworkers at the University of Wisconsin-Madison, and Hamielec and his coworkers at the McMaster University in Canada. For a variety of industrially important polymerization systems, they developed solid frameworks for the modeling and analysis of polymerization kinetics and reactor dynamics.[1]

The mathematical modeling of polymerization reactions can be classified into three levels: microscale, mesoscale, and macroscale. In microscale modeling, polymerization kinetics and mechanisms are modeled on a molecular scale. The microscale model is represented by component population balances or rate equations and molecular weight moment equations. In mesoscale modeling, interfacial mass and heat transfer effects, micromixing, polymer particle size distribution, etc. are modeled on top of the microscale chemical kinetic model. Finally, in macroscale modeling, a reactor-scale system is modeled and the steady state and dynamic reactor behaviors are modeled and analyzed. For the simulation of a polymerization process, both microscale and mesoscale models must be incorporated into a macroscopic model.

Classification of Polymerization Systems

Polymerization reactions are classified broadly into chain-growth polymerization and step-growth polymerization by the way monomer linkages are formed. In chain-growth or addition polymerization, monomer molecules are incorporated successively into growing polymer chains bearing active propagating centers. The chain growth is terminated by termination or chain transfer reactions. If the chain termination rate is very low or controlled, a living or pseudoliving polymerization occurs (e.g., living anionic polymerization). Chain-growth polymerizations include: free radical polymerization, anionic and cationic polymerizations, group transfer polymerization, and coordination polymerization. In polycondensation or step-growth polymerization, polymer linkage occurs between the two polymers containing reactive end groups such as carboxylic acid group and hydroxyl group. It is interesting to note that a molecular weight–monomer conversion relation depends upon the type of polymerization mechanism. In free radical and coordination polymerizations, high polymer molecular weight is established at very low monomer conversions, whereas in step-growth polymerizations polymer molecular weight increases with an increase in monomer conversion (e.g., $\overline{X}_n = 1/(1 - p)$: $\overline{X}_n$ = degree of polymerization, p = conversion). Polymer molecular weight increases linearly with monomer conversion in a perfectly living polymerization system. Polymerization processes can also be classified by the types of reaction medium and kinetic mechanism as illustrated in Table 1.

Also, polymerization reactions are carried out in a variety of reactors including agitated batch reactors, continuous stirred tank reactors (CSTR), multizone autoclaves, loop reactors, tubular reactors, fluidized bed reactors, and a combination of these reactors.

Modeling Objectives

The main objectives in modeling polymerization reactions are to compute polymerization rate and polymer properties for various reaction conditions. These two types of model outputs are not separate but they are usually very closely related. For example, an increase in reaction temperature raises polymerization rate

Table 1 Classification of polymerization processes

	Kinetic mechanism			
Reaction medium	**Free radical**	**Ionic**	**Coordination**	**Condensation**
Homogeneous				
Solution bulk	Vinyl polymers (styrene, MMA, LDPE)	Polyethers (ethylene oxide), SBR	Polyolefins	Polyesters, polyamides
Heterogeneous				
Emulsion dispersion	Vinyl polymers (styrene, MMA, PVC)			
Suspension	Vinyl polymers (styrene, MMA, PVC)			
Precipitation	Vinyl polymers (PVC, PAN, PVDC)	Polyacetals, vinyls		Polyamides, solid state polymerization
Solid catalyzed			Polyolefins, syndiotactic polystyrene	

but decreases polymer molecular weight; an increase in catalyst concentration raises polymerization rate but decreases polymer molecular weight. From an industrial viewpoint, increasing polymer yield with a sacrifice in polymer quality is as unacceptable as improving the product quality with a sacrifice in polymer yield. Therefore, there is a need for a detailed understanding of the polymerization kinetics to devise a scheme to simultaneously achieve high productivity and desired polymer properties. Here, for process modeling purposes, we define the polymer properties as those that represent the polymer architecture: e.g., molecular weight distribution (MWD), molecular weight averages, copolymer composition, copolymer composition distribution (CCD), monomer sequence length distribution, short-chain and long-chain branching, crosslinking, stereoregularity, polymer particle size, particle size distribution (PSD), polymer morphology, etc. These ''fundamental'' polymer properties influence a polymer's physical, chemical, thermal, mechanical, and rheological properties.

Correlating the polymer molecular properties to end-use properties in a quantitative manner is still difficult. It is partly because most of the synthetic polymers are polydisperse and it is difficult to quantify the behaviors of a mixture of polymer molecules of different chain lengths in a processing environment.

Parameter Estimation

One important aspect of polymerization process modeling is that a modeling is never complete until all the relevant model parameters are determined or estimated. In fact, determining the model parameters using laboratory, pilot plant, or plant data is perhaps the most critical step for the successful development of a process model and at the same time it is the most time consuming, costly, and difficult process. The difficulty of parameter estimation comes from the multitude of reactions that affect each other and relevant kinetic parameters that are often masked by physical transport phenomena (e.g., diffusion, mass, and heat transfer effects). Quite often, a polymerization process model validated solely on the laboratory data fails to provide accurate predictions of the behavior of a large-scale plant reactor because reaction environments can be quite different (e.g., impurities, efficiency of mixing, etc.). It is also important to realize that extracting process data from plant operations for process modeling purpose is not as easy as one may expect. It is because few plant managers or plant engineers are willing to disturb normal commercial operations to generate some data for ''academic'' research and development. It would be important to develop a consensus between modelers and plant engineers: any efforts put into the development of a plant-scale polymerization reactor model will eventually benefit everyone involved because the model can help in improving the plant operations and product quality. In this regard, there is a need to develop some techniques of parameter estimation using plant data without disturbing normal commercial production, particularly in a large-scale continuous reactor process. Another point to be made in modeling an industrial polymerization process is that one must decide the level of sophistication of the model. Unlike academic polymerization kinetic models, industrial process models should be developed with clear objectives or purposes for a given set of constraints (e.g., personnel, cost, time). There is absolutely no reason for an industry to develop the most sophisticated and comprehensive process model in the world. For example, the goals of the

reactor/process modeling project must be clearly defined with specific applications laid down before the commencement of the modeling project. Sensible assumptions and/or simplifications are the keys to a successful modeling of an industrial polymerization process.

Modeling Procedure

In general, a polymerization process model consists of material balances (component rate equations), energy balances, and additional set of equations to calculate polymer properties (e.g., molecular weight moment equations). The kinetic equations for a typical linear addition polymerization process include initiation or catalytic site activation, chain propagation, chain termination, and chain transfer reactions. The typical reactions that occur in a homogeneous free radical polymerization of vinyl monomers and coordination polymerization of olefins are illustrated in Table 2.

The molecular weight distribution can be calculated by solving the mass balance equations for monomer(s), initiator (catalytic sites), and polymeric species with different chain lengths. When quasisteady state assumption is applied to live polymers or propagating active centers, the molecular weight distribution of live polymers is often represented by the Schultz–Flory most probable distribution. However, the calculation of the chain length distribution of dead polymers is in general quite complicated. For some special cases such as olefin polymerization with single site catalysts, the polymer molecular weight distribution follows the most probable distribution:

$$w(r) = \tau^2 r \exp(-\tau r) \tag{1}$$

where $w(r)$ is the weight fraction of polymers with chain length and τ is the ratio of chain transfer rates to propagation rate. The chain length distribution function is often combined with a reactor residence time distribution function to calculate the overall molecular weight distribution.

Instead of using a complete polymer chain length distribution, molecular weight averages are frequently used as a measure of molecular weight properties. The molecular weight averages can be relatively easily calculated by solving molecular weight moment equations derived from the polymer population balance equations. The polymer molecular weight moments are defined for live and dead polymers as follows:[2]

For homopolymerization:

$$\lambda_k \equiv \sum_{n=1}^{\infty} n^k M_n \tag{2}$$

where λ_k is the kth moment, n the number of monomer units in the polymer chain, and M_n the concentration of polymers with n repeat units (monomer units). Notice that the zeroth moment represents the total number of polymer molecules and the first moment represents the total polymer concentration. To calculate the polymer molecular weight averages, the differential equations for the first three leading moments should be derived and solved together with the rate equations for monomer, initiator (catalyst), and polymers. A detailed procedure of deriving the moment equations can be found elsewhere.[2] The variances of number average and weight average molecular weights can be calculated as follows:

$$\begin{aligned}\sigma_n^2 &= \frac{\sum_{i=1}^{\infty}\left(M_i - \overline{M}_n\right)^2 n_i}{\sum_{i=1}^{\infty} n_i} \\ &= \left[\frac{\lambda_2}{\lambda_0} - \left(\frac{\lambda_1}{\lambda_0}\right)^2\right] M_0^2\end{aligned} \tag{3}$$

$$\begin{aligned}\sigma_w^2 &= \frac{\sum_{i=1}^{\infty}\left(M_i - \overline{M}_w\right)^2 n_i M_i}{\sum_{i=1}^{\infty} n_i M_i} \\ &= \left[\frac{\lambda_3}{\lambda_1} - \left(\frac{\lambda_2}{\lambda_1}\right)^2\right] M_0^2\end{aligned} \tag{4}$$

where M_0 is the molecular weight of a repeating unit. $\overline{M}_n$ and $\overline{M}_w$ are the number average and weight average molecular weights, respectively. For a copolymer, unless it is an alternating copolymer, the degree of

Table 2 Reaction schemes for addition polymerizations

Free radical polymerization	Coordination polymerization
Initiation	Site activation
$I \xrightarrow{k_d} 2R$	$C_0^* + A \xrightarrow{k_a} C^*$
$R + M \xrightarrow{k_i} P_1$	
Propagation	Initiation
$P_1 + M \xrightarrow{k_p} P_2$	$C^* + M \xrightarrow{k_i} P_1$
$P_n + M \xrightarrow{k_p} P_{n+1} (n \geq 2)$	
Chain transfer	Propagation
$P_n + M \xrightarrow{k_{fm}} M_n + P_1$	$P_1 + M \xrightarrow{k_p} P_2$
$P_n + X \xrightarrow{k_{fx}} M_n + X$	$P_n + M \xrightarrow{k_p} P_{n+1} (n \geq 2)$
Chain termination	Chain transfer
$P_n + P_m \xrightarrow{k_{tc}} M_{n+m}$	$P_n + M \xrightarrow{k_{fm}} M_n + P_1$
$P_n + P_m \xrightarrow{k_{td}} M_n + M_m$	$P_n + X \xrightarrow{k_{fx}} M_n + C^*$
	$P_n \xrightarrow{k_{fs}} M_n + C^*$

I = initiator, R = primary radical, M = monomer, P_n = live polymer with chain length n, X = chain transfer agent, M_n = dead dead polymer with chain length n, C_0^* = active catalyst site, A = catalyst activator.

polymerization is not defined and the kth molecular weight moment of the copolymer is defined as

$$\lambda_k = \sum_{n=1}^{\infty}\sum_{m=1}^{\infty}(nw_1 + mw_2)^k \quad (5)$$

where w_1 is the molecular weight of monomer 1 (M_1) and w_2 is the molecular weight of monomer 2 (M_2). The number average and weight average molecular weights are calculated as: $\overline{M}_n = \frac{\lambda_1}{\lambda_0}$, $\overline{M}_w = \frac{\lambda_2}{\lambda_1}$. The ratio $\overline{M}_w/\overline{M}_n$ is a measure of MWD broadening and it is called the polydispersity. The mathematical techniques to derive the molecular weight moment equations can be found elsewhere.[2–4] For a linear and binary copolymerization system, the instantaneous chain length and composition distribution can be calculated using the modified Stockmayer bivariate distribution function given by:[5]

$$w(r,y) = (1 + y\delta)\tau^2 r\exp(-\tau r)\mathrm{d}r\frac{1}{\sqrt{2\pi\beta/r}}\exp\left(-\frac{y^2 r}{2\beta}\right)\mathrm{d}y \quad (6)$$

where $\beta = \overline{F}_1(1-\overline{F}_1)[1+4\overline{F}_1(1-\overline{F}_1)(r_1r_2-1)]^{1/2}$; $\delta = \frac{1-M_2/M_1}{M_2/M_1+\overline{F}_1(1-M_2/M_1)}$; $\overline{F}_1 =$ average mole fraction of monomer 1; M_1, M_2 = molecular weights of monomers 1 and 2; r = chain length; y = deviation from average copolymer composition ($=\overline{F}_1 - F_1$); and r_1, r_2 = reactivity ratios.

It should be pointed out that molecular weight averages do not necessarily represent the exact nature of polymer MWD. For example, it is possible that two polymer samples of different chain length distribution can have identical number and weight average molecular weights. In polymer industry, melt index (MI) is frequently used as a measure of polymer molecular weight or rheological properties. The melt index is the measurement of the flow rate in g/10 min of polymer flowing through a die at a given temperature under the action of a weight loaded onto a piston. Typical MI test conditions are 190/2.16, i.e., 190°C in temperature and 2.16 kg in weight. However, it is important to understand that MI is a viscosity at the particular shear rate and temperature employed in the test. Also, MI does not give any information about the elasticity of the polymer. Quite often in the polymer industry, MI is correlated with molecular weight averages (e.g., MI $= a\overline{M}_w^b$, where a and b are empirical parameters).

To illustrate the calculation of molecular weight averages using moment equations, let us consider a free radical polymerization of vinyl monomers. Table 3 shows the kinetic equations based on the kinetic scheme in Table 1 for initiator, monomer, and live and dead polymers; and the molecular weight moment equations for live and dead polymers. The molecular weight moment equations for other addition polymerization processes such as transition metal-catalyzed olefin polymerization can also be derived using the same method illustrated in Table 3.

By using the molecular weight moment equations, we can avoid the necessity of solving the infinite number of polymer population equations separately. A quasisteady state approximation is usually applied to live polymers and live molecular weight moments. Polymer molecular weight can also be calculated using the moment generating functions. If P_i denotes the discrete fraction of polymer of chain length i, the moment generating function is defined as $F_z \equiv \sum_{i=1}^{\infty} P_i z^i$. Then the kth moment is expressed as:

$$\lambda_k = \left[z\frac{\mathrm{d}^{(k)}}{\mathrm{d}z}F_z\right]_{z=1}$$

Heterogeneous Polymerization Systems

The modeling of heterogeneous polymerization systems is generally more complicated than that of the homogenous systems because mass and heat transfer effects between two or more immiscible phases must be considered. Industrially important heterogeneous polymerization reactions include emulsion polymerization, suspension polymerization, precipitation polymerization, and solid-catalyzed olefin polymerization. The general polymerization rate equation is represented simply as

$$R_\mathrm{p} = k_\mathrm{p} f([M],[A^*]) \quad (7)$$

where k_p is the polymerization rate constant, $[M]$ the monomer concentration, and $[A^*]$ the concentration of propagating centers. Both propagation rate constant and monomer concentration at active sites can be influenced by the presence of interfacial mass transfer effects. For example, in emulsion polymerization, the above general equation takes the following form:

$$R_\mathrm{p} = k_\mathrm{p} N_\mathrm{p} \overline{n} [M]_\mathrm{s} \quad (8)$$

where N_p is the number of polymer (latex) particles, $\overline{n}$ the average number of radicals per latex particle, and $[M]_\mathrm{s}$ the monomer concentration in polymer particles. Then, the question is how we can calculate these variables using the understanding of thermodynamics and the kinetics and mechanism of emulsion polymerization.[3,4,6] For example, to calculate the average number of radicals per polymer latex particle, we need

Table 3 Kinetic model equations for free radical polymerization

Kinetic equations	MW moment equations
$\frac{dI}{dt} = -k_d I$	$\frac{d\lambda_1^\ell}{dt} = k_i RM + (k_{fm}M + k_{fs}S)(P - \lambda_1^\ell) + k_p MP - (k_{tc} + k_{td})P\lambda_1^\ell$
$\frac{dR}{dt} = 2f_i k_d I - k_i RM$	$\frac{d\lambda_2^\ell}{dt} = k_i RM + (k_{fm}M + k_{fs}S)(P - \lambda_2^\ell) + k_p M(2\lambda_1^\ell + P) - (k_{tc} + k_{td})P\lambda_2^\ell$
$\frac{dM}{dt} = -k_i RM - k_p M\sum_{n=1}^{\infty} P_n - k_{fm}M\sum_{n=1}^{\infty} P_n$	$\frac{d\lambda_0^d}{dt} = \frac{1}{2}k_{tc}P^2 + k_{td}P(P - P_1) + (k_{fm}M + k_{fs}S)(P - P_1) = \frac{1}{2}k_{tc}P^2 + (k_{fm}M + k_{fs}S + k_{td}P)\alpha P$
$\frac{dP_1}{dt} = k_i RM - k_p MP_1 + (k_{fm}M + k_{fs}S)\sum_{n=2}^{\infty} P_n - (k_{tc} + k_{td})P_1\sum_{n=2}^{\infty} P_n$	$\frac{d\lambda_1^d}{dt} = \frac{1}{2}k_{tc}P^2(1-\alpha)^2\sum_{n=2}^{\infty} n(n-1)\alpha^{n-2} + k_{td}P^2(1-\alpha) \times \sum_{n=2}^{\infty} n\alpha^{n-1} + (k_{fm}M + k_{fs}S)P(1-\alpha)\sum_{n=2}^{\infty} n\alpha^{n-1} = \frac{1}{1-\alpha}\left[k_{tc}P^2 + (k_{fm}M + k_{fs}S + k_{td}P)P(2\alpha - \alpha^2)\right]$
$\frac{dP_n}{dt} = k_p M(P_{n-1} - P_n) - (k_{fm}M + k_{fs}S)P_n - (k_{tc} + k_{td})P_n\sum_{n=1}^{\infty} P_n (n \geq 2)$	$\frac{d\lambda_2^d}{dt} = \frac{P}{(1-\alpha)^2}[(k_{fm}M + k_{fs}S + k_{td}P)(\alpha^3 - 3\alpha^2 + 4\alpha) + k_{tc}P(\alpha + 2)]$
$\frac{dM_n}{dt} = (k_{fm}M + k_{fs}S)P_n + k_{td}P_n\sum_{n=1}^{\infty} P_n + \frac{1}{2}k_{tc}\sum_{m=1}^{n-1} P_m P_{n-m} (n \geq 2)$	where $\alpha \equiv \frac{k_p M}{k_p M + k_{fm}M + k_{fs}S + (k_{tc} + k_{td})P}$

(From Ref.[2])

to consider various physical transport processes such as radical transfer to and from a particle and radical termination inside and outside (aqueous phase) of a particle. To calculate the total number of latex particles, micellar nucleation as well as homogeneous nucleation kinetics needs to be understood. In emulsion polymerization, controlling the latex particle size distribution is one of the most important objectives because the particle size distribution can have a significant impact on the quality of coatings in final applications. Therefore, understanding the polymerization kinetics and complex thermodynamic and physical transport phenomena is more than a necessity in designing and operating emulsion polymerization processes. The kinetics of solid-catalyzed olefin polymerization is also quite complicated because of some physical effects such as catalyst fragmentation, monomer diffusion and sorption, and interfacial transfer resistances. In solid-catalyzed olefin polymerization, the rate equation is expressed as

$$R_p = k_p[M]_s[C^*] \tag{9}$$

where $[C^*]$ is the active catalytic site concentration and $[M]_s$ is the effective concentration of monomer at the catalytic site. Again, the question is how the reaction environment influences the rate constant, the monomer concentration at catalytic sites, and the active site concentration in the presence of intraparticle and interfacial mass transfer resistances, monomer sorption, disintegration of crystalline catalytic components, and heat transfer resistance.[7–10]

PARAMETER ESTIMATION IN POLYMERIZATION REACTION MODELING

One of the key tasks in polymerization reaction modeling is to determine or estimate the model parameters

including rate constants and relevant physical and thermodynamic parameters such as mass and heat transfer coefficients, diffusivity, density, heat capacity, active site concentrations, etc. Some kinetic parameters may change with a changing reaction environment. In transition metal-catalyzed olefin polymerizations, the kinetic parameters are catalyst dependent. Therefore, whenever a new catalyst is employed, a new set of kinetic parameters must be determined. Considering the fact that the properties of polyolefins are mostly dictated by the nature of the catalyst being used and that a large number of different types of catalysts are used for different polymer grades, we can easily understand the importance of having a well-established parameter estimation procedure that can be applied to any catalyst system. Quite often, a kinetic model validated on a small-scale laboratory reaction data provides poor predictions of a large-scale polymerization reactor. There are many reasons for such discrepancies but some notable reasons are differences in mixing performance (compositional and thermal heterogeneity), impurity levels, and mass and heat transfer resistances. In practice, the model parameters obtained from the laboratory data are used as a reference, and actual plant data are used to adjust the model parameters. To estimate the kinetic parameters, statistical experimental design techniques are highly recommended.

In some polymerization processes, developing a first principles model can be practically infeasible. For example, when several vinyl monomers are copolymerized using several free radical initiators, it is extremely difficult, if not impossible, to develop model equations to calculate the rate of polymerization and polymer properties such as molecular weight averages. Estimating the relevant kinetic parameters will be even more difficult than deriving the rate equations. In such cases, building a statistical model would be a more pragmatic approach. Of course, the statistical model should be used with some caution because it does not contain any physical or chemical information about the process in itself and the applicable process range can be quite narrow. Nevertheless, statistical models are frequently used in the polymer industry for quality control purposes.

CONTROL OF POLYMERIZATION REACTORS

Control of industrial polymerization reactors is a challenging task because, in general, control engineers lack rigorous polymerization process knowledge, process model, and rapid online or inline sensors to measure polymer properties. Exothermic polymerization processes often exhibit strongly nonlinear dynamic behaviors (e.g., multiple steady states, autonomous oscillations, limit cycles, parametric sensitivity, and thermal runaway), particularly when continuous stirred tank reactors are used.[11–15] Some polymerization processes are open loop unstable and susceptible to unmeasured disturbances or upsets even with a feedback controller in place. For example, in a transition metal-catalyzed olefin polymerization process, an unmeasured small amount of catalyst poisons can change the polymerization kinetics and hence the polymer yield and polymer properties. In worst cases, process disturbances may lead the reactor to instability.[16,17]

In polymerization processes, the primary goals of reactor control are to maintain stable reactor operations and main product quality indices at their target values. For an existing plant, improved reactor controls are needed to increase the polymer yield and to reduce production cost. As many of the polymer properties are hard to monitor online, first-level process variables are controlled. Typical reactor variables subject to control include polymerization temperature, pressure, feed rates of monomer(s), catalysts or initiators, chain transfer agents, solvents, etc. Both polymerization production rate and polymer properties are nonlinearly correlated and hence a polymerization process control system is inherently a multivariable control system. The primary control objective is to control temperature, pressure, and flow rates that can be readily measured online. The set points of these variables are often called the process recipe. In principle, as long as these variables are tightly controlled, consistent product quality can be warranted. However, in the presence of unexpected process disturbances or upsets, little can be done to correct the damages made on the product properties. The second-level control objectives include the direct control of polymer properties using online measurements or estimates of polymer quality indices. Any variations in the product quality can be corrected, in principle, if such quality indices are readily available during the polymerization. The reactor control objectives and control strategies may vary depending on the type of reactors. Table 4 illustrates typical control design issues for three different types of reactors.

Batch polymerization reactors are ideal to manufacture small volume polymers, specialty polymers, and polymers that are difficult to make in continuous reactors. Emulsion polymers, suspension polymers, and precipitation polymers are mostly made by batch polymerization processes. One of the disadvantages of a batch reactor is that the ratio of heat transfer surface area to reactor volume decreases as the reactor size is increased. For many polymer products made in batch reactors, the process economy improves with an increase in reactor size. Therefore, effective heat removal becomes a critical factor in designing and controlling a large-scale batch polymerization reactor.

In batch polymerization processes, examples of typical control objectives are 1) to maintain reactor

Table 4 Issues in reactor control designs

	Batch process	Semibatch process	Continuous process
Product type	Small volume specialty polymers; heterogeneous polymerization system (e.g., emulsion, suspension, precipitation)	Small volume specialty polymers; copolymers	Large volume commodity polymers; engineering polymers
Reaction phase	Liquid; liquid–solid	Liquid; liquid–solid	Liquid; gas phase; slurry phase
Typical reactor type	Mechanically agitated reactor with a heating jacket or condenser	Mechanically agitated reactor with a heating jacket or condenser	CSTR, tubular reactor, multiple CSTRs, fluidized bed reactor, loop reactor
Typical operations	Nonstationary dynamic process	Nonstationary dynamic process	Steady state
Control objectives	Design control trajectories; feedback tracking of control trajectories; time optimal control; temperature control	Design control trajectories; feedback tracking of control trajectories; time optimal control; optimal initial conditions	Design steady state operating conditions; maintaining steady state stability; grade transition controls; online property controls
Constraints	Batch time can be short; online property monitoring or product sampling can be difficult; reactor heat removal can be difficult for large reactors (small heat transfer area/reactor volume ratio); insufficient time for product quality analysis and corrective control	Batch time can be short; online property monitoring or product sampling can be difficult; reactor heat removal can be difficult for large reactors	Control failure leads to a large loss of product; online monitoring or direct properties control can be achieved; strong process nonlinearity; parametric sensitivity; oscillations; runaway reactions

stability (e.g., to avoid thermal runaway); 2) to obtain maximum polymer yield in minimal reaction time; and 3) to bring the product properties to their target values as closely as possible. In general, direct online control of polymer properties is not feasible in many batch polymerization processes. Even taking a sample can be quite a challenge in some high pressure batch reactor systems. A batch polymerization reactor should also be operated to maintain consistent batch-to-batch product quality and to maximize the product yield by increasing monomer conversion and/or reducing batch reaction time. The design of a batch polymerization reactor control consists of two stages: 1) offline design of a control trajectory (recipe); and 2) implementation and execution of the control trajectory. The control trajectory can be developed through experimentation, plant experience, or by using a process model.[18,19] A batch polymerization process is a multivariate and nonstationary or dynamic process. Quite often there may exist some conflicting control objectives (e.g., polymer yield, molecular weight, composition, batch reaction time) that require special treatments. For example, multiobjective dynamic optimization techniques can be used to develop optimal reactor operating policies or target control trajectories in the presence of conflicting control objectives.[20,21] Quite obviously, an accurate dynamic polymerization reactor model is a prerequisite for such advanced control designs. Many excellent dynamic optimization techniques have been developed and they are readily available to control engineers. A typical objective function (F) for reactor optimization takes the following form:

$$F = w_1 t_f + \sum_{i=1}^{N} w_i \left(\frac{Y_i}{Y_i^{\mathrm{d}}} \right)^2 \tag{10}$$

where w_i is the weighting factor, t_f the batch time, Y_i the product quality parameter (e.g., M_w, average copolymer composition, etc.), and Y_i^{d} the desired quality parameter value. The objective function is minimized

subject to process model and constraints:

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}(t), \mathbf{u}(t), t); \quad \mathbf{c}(\mathbf{x}(t), \mathbf{u}(t), t) \leq 0 \tag{11}$$

where $\mathbf{x}$ is the state variable (e.g., concentration, temperature, molecular weight moments, etc.) and $\mathbf{u}$ the control variable. Fig. 1 illustrates the design of optimal reactor temperature trajectories to obtain a desired molecular weight distribution in a batch methyl methacrylate polymerization process.[18] Here, the feasible sequential quadratic programming (FSQP) technique is used to find the sequence of optimal reactor temperature set points that will yield the best match between target and actual polymer chain length distributions at the end of the batch. The graphs on the left in Fig. 1 represent the computed sequence of reactor temperature set points at selected iterations, and the graphs on the right show the resulting chain length distribution compared with the target distribution. The final temperature set point program can be implemented and executed, or the trajectory can be updated online if timely measurements or estimates of polymer molecular weight distribution are available during the batch operation. Once the reactor control trajectory is designed, the next goal is to execute the trajectory as

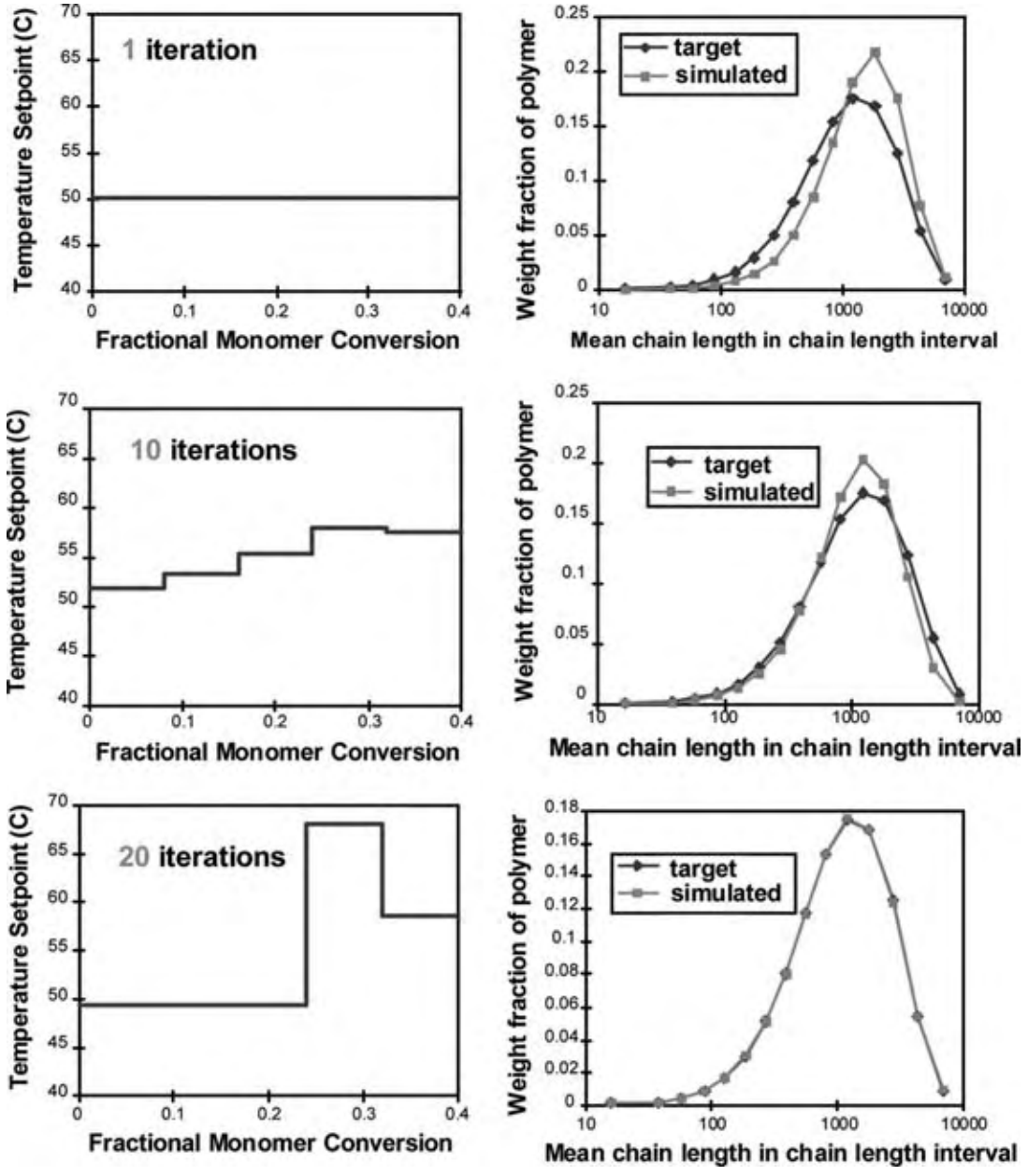


Fig. 1 Design of optimal batch polymerization reactor controls. *(View this art in color at www.dekker.com.)*

close to the design as possible. Traditional PID controllers are still widely used, but model predictive controllers (MPC) are also used in some polymerization processes where the use of advanced reactor control can be economically justified. In MPC, a process model is utilized to predict the output into the future and minimize the difference between the predicted model output and the desired output using some open loop objective function. The measurement is used to update the optimization problem for the next time step. The MPC algorithms are reasonably well developed and utilized in many chemical processes including polymerization processes.[22–27] As industrial polymerization processes exhibit strong nonlinearity, the application of linear model predictive control (LMPC) is often limited, particularly for grade transition control and for regulatory control. In nonlinear model predictive control (NMPC) algorithms, a nonlinear programming problem has to be solved online and hence computational load is generally quite heavy. To reduce the computational burden, a successive linearization of the original nonlinear process model can be used to approximate the nonlinear process behaviors.[23,25] As batch reactors are widely used to make specialty polymers of various grades, designing optimal batch scheduling is also an important issue.

Continuous reactors are operated at steady state, and hence the key objective in controlling a continuous polymerization reactor system is to maintain reactor stability in the presence of any process upsets and during normal steady state operations and grade transition operations. In some continuous polymerization processes such as liquid slurry olefin polymerizations, reactor fouling may develop over a period of time, gradually decreasing the control performance. In such a case, one may adjust the control variables to compensate for the changes in process characteristics. Then, some polymer property indices that are not directly measured or controlled online may drift from their specifications to result in poor product quality. Therefore, the development of online estimation techniques for polymer property indices that cannot be measured online or can be only measured offline with significant time delays becomes necessary.

In continuous polymerization processes, polymers of different properties are manufactured in a single product line. Therefore, the design of efficient grade transition controls and optimal production scheduling becomes another important control design objective. In some sense, the grade transition control problem is similar to the optimal design of a batch polymerization reactor control. Input–output models or transfer functional models are widely used to develop a multivariable control system. As such linear models are the approximations of highly nonlinear polymerization processes, the linear controllers have some limitations. Polymerization processes are multivariable systems in nature, and thus significant control loop interactions are expected in conventional feedback control systems. One of the most instructive examples that illustrate the value of a process model is the work by Congalidis et al. on the design of feedforward and feedback control of a continuous solution copolymerization reactor using a multivariable transfer function model.[28] They analyzed control structure/loop pairings using singular value decomposition and relative gain array. Then, loop pairings are determined and a combined feedforward/feedback strategy is developed for servo- and regulatory control problems. For a continuous terpolymerization process, Ogunnaike[27] illustrates an interesting control scheme that involves a two-tier system. In the first tier level, the flow rates of monomer, catalyst, solvent, and chain transfer agent are used to regulate reactant composition in the reactor. Then, at the second tier level, set points for the composition of the reactor contents are used at a less frequent update rate to regulate final product properties. In the second tier, an online, dynamic kinetic model running in parallel with the process supplies estimates of product properties. The model predictions are updated using delayed laboratory measurements and online stochastic filter.

Polymerization process control can benefit significantly from using online state estimation techniques. In general, online control of polymer properties such as molecular weight, MWD, copolymer composition, MI, density, etc. is difficult, mainly because of the lack of adequate online or in-process sensors.[29] Therefore, many of these polymer property parameters are controlled indirectly by controlling first-level process variables such as temperature, pressure, and the flow rates of various reactants, solvents, and catalysts. When some deviations in polymer properties are detected through laboratory sample analysis, certain reactor variables need to be adjusted. Extensive plant experience might be required to make such process adjustments, or model-based online state estimator can be used.

One alternative to the direct online measurement of polymer properties is to use a process model in conjunction with optimal state estimation techniques to predict the polymer properties. Indeed, several online state estimation techniques such as Kalman filters, nonlinear extended Kalman filters (EKF), and observers have been developed and applied to polymerization process systems.[30,31] In implementing the online state estimator, several issues arise. For example, the standard filtering algorithm needs to be modified to accommodate time-delayed offline measurements (e.g., MWD, composition, conversion). The estimation update frequency needs to be optimally selected to compensate for the model inaccuracy. Table 5 shows the extended Kalman filter algorithm with delayed offline measurements. Fig. 2 illustrates the use of online state estimator

Table 5 Extended Kalman filter algorithm with online and offline measurements

Process model	$\frac{dx}{dt} = f(x,u) + w(t), \quad w(t) \sim N[0, Q(t)]$ $x(0) \sim N[\tilde{x}_0, P_0]$
Online measurements	$y_{o,k} = h_0(x_k) + v_{0,k}; \quad v_{0,k} \sim N[0, R_{0,k}]$
Delayed measurements	$y_{d,k-\tau} = h_0(x_{k-\tau}) + v_{d,k-\tau}; \quad v_{d,k-\tau} \sim N[0, R_{d,k-\tau}]$
State estimation propagation	$\frac{d\hat{x}}{dt} = f[\hat{x}(t), u(t)]$
Error covariance propagation	$\frac{dP}{dt} = F(t)P(t) + P(t)F^T(t) + Q(t)$
State estimate update with online measurements	$\tilde{x}_k(+) = \tilde{x}_k(-) + K_k[y_{0,k} - h_0[\tilde{x}_k(-)]]$
Error covariance update with online measurements	$P_k(+) = [I - K_k H_{0,k}]P_k(-)$
Filter gain matrix with online measurements	$K_k = P_k H_{0,k}^T \left[H_{0,k} P_k(-) H_{0,k}^T + R_{0,k}\right]^{-1}$ where $F(t) = \left.\frac{\partial f[x(t), u(t)]}{\partial x(t)}\right\|_{x(t)=\tilde{x}(t)}$ $H_{0,k} = \left.\frac{\partial h_0(x_k)}{\partial x_k}\right\|_{x_k=\tilde{x}(-)}$

(i.e., EKF) with delayed molecular weight measurements in a continuous stirred tank styrene polymerization reactor.[30] In this particular simulation, the minimum offline measurement time is set as 30 min and irregular sampling intervals are assumed. Two models with different parameters are used for plant simulation (dotted lines) state estimation (solid lines). Fig. 2 clearly shows that polymer molecular weight can be estimated online with EKF and delayed offline measurements as long as a reasonably good process model is available. In general, frequent sample analysis is necessary when a relatively inaccurate process model is used for state estimation and less frequent sample analysis is necessary when a relatively accurate process model is used.

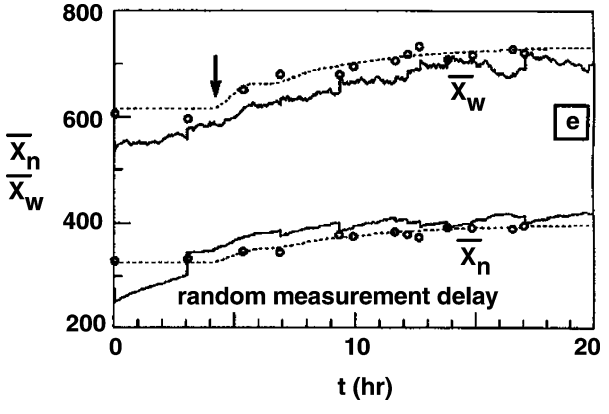


Fig. 2 Online state estimation simulations for a continuous stirred tank styrene polymerization reactor with extended Kalman filter.

The state estimation technique can also be incorporated into the design of optimal batch polymerization control system.[18,19] For example, a batch reaction time is divided into several control intervals, and the optimal control trajectory is updated online using the molecular weight estimates generated by a model/state state estimator. Of course, if batch reaction time is short, such feedback control of polymer properties would be practically difficult to implement. Nevertheless, the online stochastic estimation techniques and the model predictive control techniques offer promising new directions for the improved control of batch polymerization reactors.

CONCLUSIONS

Mathematical modeling is a powerful tool not only for the development of process understanding but also for that of the advanced reactor controls in polymerization processes. The modeling techniques for polymerization processes are reasonably well developed and several commercial simulation packages are available. The modeling of heterogeneous polymerizations such as precipitation polymerization and emulsion polymerization remains a challenge. In the past decade, excellent

design tools have also been developed for multiobjective optimization, online state estimation, and model predictive control. These techniques are gaining favorable acceptance in the polymer industry. One of the outstanding issues is to develop more efficient parameter estimation techniques, particularly in a plant environment without disturbing normal process operations.[31] Almost all the computational and modeling tools are available: the question is how one can use them wisely and creatively to produce the polymers of the highest quality at the lowest possible cost.

It is also instructive to understand that advancing the industrial polymerization process technology is often hampered because of the lack of fundamental research activities in academic and industrial laboratories. For example, in the manufacturing of poly(vinyl chloride), which is still an important and huge industry, controlling the polymer particle morphology and designing new properties are strongly needed. However, our understanding of polymerization kinetics and mechanisms remains little changed since 1980s because little research has been carried out on the polymerization of vinyl chloride in recent years. Many people believe that there should be nothing that is not understood about the polymers that are more than 50 years old, which is not quite true. With a tight economy, many companies can no longer afford advanced research and development on such old but profitable commercial polymers.

REFERENCES

1. Choi, K.Y. Continuous processes for radical vinyl polymerization. In *Handbook of Radical Vinyl Polymerization*; Mishra, M.K., Yagci, Y., Eds.; Marcel Dekker: New York, 1998; 275–297.
2. Ray, W.H. On the mathematical modeling of polymerization reactors. J. Macromol. Sci. Rev. Macromol. Chem. **1972**, *C8* (1), 1–56.
3. Schork, F.J.; Deshpande, P.B.; Leffew, K.W. *Control of Polymerization Reactors*; Marcel Dekker: New York, 1993.
4. Dotson, N.A.; Galvan, R.; Laurence, R.L.; Tirrell, M. *Polymerization Process Modeling*; VCH Publishers: New York, 1996.
5. Dube, M.A.; Soares, J.B.P.; Penlidis, A.; Hamielec, A.E. Mathematical modeling of multicomponent chain-growth polymerizations in batch, semibatch, and continuous reactors: a review. Ind. Eng. Chem. Res. **1997**, *36*, 966–1015.
6. Lovell, P.A.; El-Aasser, M.S. *Emulsion Polymerization and Emulsion Polymers*; Wiley: New York, 1997.
7. Floyd, S.; Choi, K.Y.; Taylor, T.W.; Ray, W.H. Polymerization of Olefins through heterogeneous catalysis III Polymer particle modeling with an analysis of intraparticle heat and mass transfer effect. J. Appl. Polym. Sci. **1986**, *32*, 2935–2960.
8. Floyd, S.; Choi, K.Y.; Taylor, T.W.; Ray, W.H. Polymerization of olefins through heterogeneous catalysis. IV. Modeling of heat and mass transfer resistance in the polymer particle boundary layer. J. Appl. Polym. Sci. **1986**, *31*, 2231–2265.
9. Floyd, S.; Heiskanen, T.; Taylor, T.W.; Ray, W.H.; Choi, K.Y. Polymerization of olefins through heterogeneous catalysis. VI. Effect of particle heat and mass transfer on polymerization behavior and polymer properties. J. Appl. Polym. Sci. **1987**, *33*, 1021–1065.
10. Hutchinson, R.A.; Ray, W.H. Polymerization of olefins through hetereogenous catalysis. VII. Particle ignition and extinction phenomena. J. Appl. Polym. Sci. **1987**, *34*, 657–676.
11. Choi, K.Y. Modeling of polymerization processes. In *Computer-Aided Design of Catalysts*; Becker, E.R., Pereira, C.J., Eds.; Marcel Dekker: New York, 1993; 335–389.
12. Kiparissides, C. Polymerization reactor modeling: a review of recent developments and future directions. Chem. Eng. Sci. **1996**, *51* (10), 1637–1659.
13. Kim, K.J.; Choi, K.Y.; Alexander, J.C. Dynamics of a CSTR for styrene polymerization initiated by a binary initiator system. Polym. Eng. Sci. **1990**, *30* (5), 279–290.
14. Kim, K.J.; Choi, K.Y.; Alexander, J.C. Dynamics of a cascade of two continuous stirred tank polymerization reactors with a binary initiator mixture. Polym. Eng. Sci. **1991**, *31* (5), 333–352.
15. Kim, K.J.; Choi, K.Y.; Alexander, J.C. Dynamics of a CSTR for styrene polymerization initiated by a binary initiator mixture. II. Effect of viscosity dependent heat transfer coefficient. Polym. Eng. Sci. **1992**, *32* (7), 494–505.
16. Choi, K.Y.; Ray, W.H. The dynamic behavior of fluidized bed reactors for solid catalyzed gas phase olefin polymerization. Chem. Eng. Sci. **1985**, *40* (12), 2261–2279.
17. Choi, K.Y.; Ray, W.H. The dynamic behavior of continuous stirred bed reactors for the solid catalyzed gas phase polymerization of propylene. Chem. Eng. Sci. **1986**, *43* (10), 2587–2604.
18. Crowley, T.J.; Choi, K.Y. Discrete optimal control of molecular weight distribution in a batch free radical polymerization process. Ind. Eng. Chem. Res. **1997**, *36*, 3676–3684.
19. Crowley, T.J.; Choi, K.Y. Control of copolymer hydrodynamic volume distribution in free radical copolymerization process. Comp. Chem. Eng. **1999**, *23*, 1153–1165.

20. Butala, D.N.; Fan, M.K.H.; Choi, K.Y. Multiobjective dynamic optimization of semibatch free radical copolymerization process with interactive CAD tools. Comp. Chem. Eng. **1988**, *12* (11), 1115–1127.
21. Butala, D.N.; Liang, W.R.; Choi, K.Y. Multiobjective dynamic optimization of batch free radical polymerization process by mixed initiator systems. J. Appl. Polym. Sci. **1992**, 1759–1778.
22. Peterson, T.; Hernandez, E.; Arkun, Y.; Schork, F.J. A nonlinear DMC algorithm and its application to a semibatch polymerization reactor. Chem. Eng. Sci. **1992**, *47* (4), 737–753.
23. Seki, H.; Ogawa, M.; Ooyama, S.; Akamatsu, K.; Ohshima, M.; Yang, W. Industrial application of a nonlinear model predictive control to polymerization reactors. Control Eng. Practice **2001**, *9* (8), 819–828.
24. Jeong, B.G.; Yoo, K.Y.; Rhee, H.K. Nonlinear model predictive control using a Wiener model of a continuous methyl methacrylate polymerization reactor. Ind. Eng. Chem. Res. **2001**, *40* (25), 5968–5977.
25. Young, R.E.; Bartusiak, R.D.; Fontaine, R.W. Evolution of an industrial nonlinear model predictive controller. AIChE Symp. Ser. **2002**, *98*, 342–351.
26. Doyle, F.J., III; Soroush, M.; Cordeiro, C. Control of product quality in polymerization processes. AIChE Symp. Ser. **2002**, *98*, 290–306.
27. Ogunnaike, B.A. The role of CACSD in contemporary industrial process control. IEEE Control Syst. **1995**, *15* (2), 41–53.
28. Congalidis, J.P.; Richards, J.R.; Ray, W.H. Feedforward and feedback control of a solution copolymerization reactor. AIChE J. **1989**, *35* (6), 891–903.
29. Chien, D.C.H.; Penlidis, A. On-line sensors for polymerization reactors. JMS Rev. Macromol. Chem. Phys. **1990**, *C30* (1), 1–42.
30. Kim, K.J.; Choi, K.Y. On-line estimation and control of a continuous stirred tank polymerization reactor. J. Proc. Control. **1991**, *1*, 96–110.
31. Sirohi, A.; Choi, K.Y. On-line parameter estimation in a continuous polymerization process. Ind. Eng. Chem. Res. **1996**, *35*, 1332–1343.

Polysaccharides

Anton Huber
PolySaccharide Initiative, Institut für Chemie, Karl-Franzens, Universität Graz, Graz, Austria

INTRODUCTION

The major part of annually assimilated biomass/renewable resources comprises polysaccharides/carbohydrates. These materials are produced, transformed, and utilized by all organisms in biosphere Earth. Although the most important producers with respect to quantities produced are green plants, polysaccharides are made by living organisms at all levels of organization by enzymatically catalyzed biochemical mechanisms in a complex aqueous environment. Estimations of global annually formed polysaccharides are in the range 100–300 Gt dry matter, which is approximately two magnitudes more than the estimated annual crude oil production of approximately 2 Gt.

Polysaccharides have no outstanding or obviously spectacular property, such as is the case with nucleic acids (genetic code) and many proteins (biocatalysts), but they have a wide range of functionalities in practically all biological transformation processes. They are involved in functionally structuring cell compartments by forming the following: matrices with varying resistance towards chemical, mechanical, or thermal stress; dynamic supermolecular soft gel structures and mucilages; rigid protective construction elements such as fibers and crystalline structures; decorative structures on surfaces; and interfaces with selective permeability.

In terms of material properties, polysaccharides are ''high-performance'' materials with superior features:

- High flexibility at the molecular level.
- Sufficient sensor and amplification quality.
- Appropriate ''smart'' response capability.
- Assembly and transformation at the molecular level.

The diversity of polysaccharides at the molecular level is due to the simple basic elements from which they are built. Abundant and ubiquitously available carbon (dioxide), oxygen, hydrogen, and H_2O, with contributions from phosphorus (P), sulfur (S), and nitrogen (N) at different states of oxidation (global distribution of C: 9.5%, H: 63%, O: 25.5%, N: 1.4%), are the building blocks for ''on-demand'' formation and transformation of natural carbohydrates, primarily by green plants and algae. The polysaccharides constituting carbohydrates contain simple functional groups: hemiacetal/hemiketal and alcoholic hydroxyl groups that can be oxidized or reduced. A chiral center at the hemiacetal/hemiketal position provides two anomeric forms (α, β) of each glycosyl residue. These have no noteworthy differences as regards usual laboratory chemistry; however, the chirality is of great significance during polymerization of the residues. The variety of polysaccharides additionally becomes practically unlimited when one considers, in combination, the heterogeneities in the degree of polymerization, branching characteristics, substitution patterns, and variations in the distribution of redox states along polysaccharide-constituting glycosyl residues.

The efficient sensor and amplification qualities of polysaccharides in different conditions in the aqueous environment are established by their pronounced and varied interactions with H_2O. This results in dynamic formation and disintegration of soft gel structures and extremely long-range interactions. In highly organized biological systems, specific glycosyl patterns on interfaces and surfaces of protein and lipid domains, such as on membranes and cell walls, can act as triggers for immune reactions. The formation and disintegration of polar and apolar glycosyl domains enable preferred docking to polar and apolar sites, respectively.

The ''smart'' response capability of polysaccharides upon application of stress allows polar and apolar domains to easily form or disintegrate due to variation of the order/disorder ratio at the molecular level. This finally results in variation of specific crystallinity index or crystalline/amorphous ratio at the macroscopic level. Significant variation of interactive properties may even be achieved by minor variation of branching characteristics, which changes surface/volume ratio and, hence, preferences for inter- or intramolecular stabilization. Additionally, a rather effective response option is variation of relative percentages of molar and mass fractions by limited degradation/reorganization or precipitation/dissolution transition.

Assembly and transformation of polysaccharides by molecular-scale tools is triggered by principles of self-organizing complex systems and include enzymatically catalyzed nonlinear formation/transformation and degradation based on molecular-level library command sequences (genetic code).

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039773

CONSTRUCTION AND NOMENCLATURE

Polysaccharides are formed by glycosidically linked carbohydrate (glycosyl) residues. As expected, the nomenclature of polysaccharides is based on the nomenclature of carbohydrates. The recommendations of the International Union of Pure and Applied Chemistry–International Union of Biochemistry and Molecular Biology Joint Commission on Biochemical Nomenclature (IUPAC–IUB JCBN) have been published[1–4] and are open to the public at http://www.chem.qmul.ac.uk/iupac/2carb/39.html.

The major basic building blocks for oligo- and polysaccharides (glycans) are D-glucopyranose (α-D-Glc*p*, β-D-Glc*p*) and D-fructofuranose (β-D-Fru*f*), either α- or β-glycosidically linked (Fig. 1). Any of the hydroxyl groups at each glycosyl residue (for the case of two (1→4)-linked glucosyl units: C2, C3, and C6 (Fig. 2A) in an oligo-/polysaccharide may be subject to oxidation/reduction, activation, substitution, or the formation of additional glycosidic linkages. In the formation of "second-order" glycosyl glycosides, branching is an important option; however, this is just one out of many possibilities, such as:

- Oxidation at the terminal secondary alcoholic hydroxyl group (C6-OH) to uronic acid (Fig. 2B).
- Substitution to increase charge density (polarity) with acetate groups [–O–C(=O)–CH_3] (Fig. 2C).
- Introduction of an amino (–NH_2) (Fig. 2D) or *N*-acetyl (Nac) (Fig. 2E) group at the C2-OH, compatible to peptide linkages.
- Substitution to reduce polarity/increase hydrophobicity by reducing –OH groups to the deoxy form (–H) or substitution with methyl groups (–CH_3).
- Activation by formation of sulfates [–O–(O=)S(=O)–O′] (Fig. 2F), phosphates [–O–(OH–)–P(=O)–O′] (Fig. 2G), or nucleosides.

There is no strict transition from oligo- to polysaccharides. However, oligosaccharides are typically

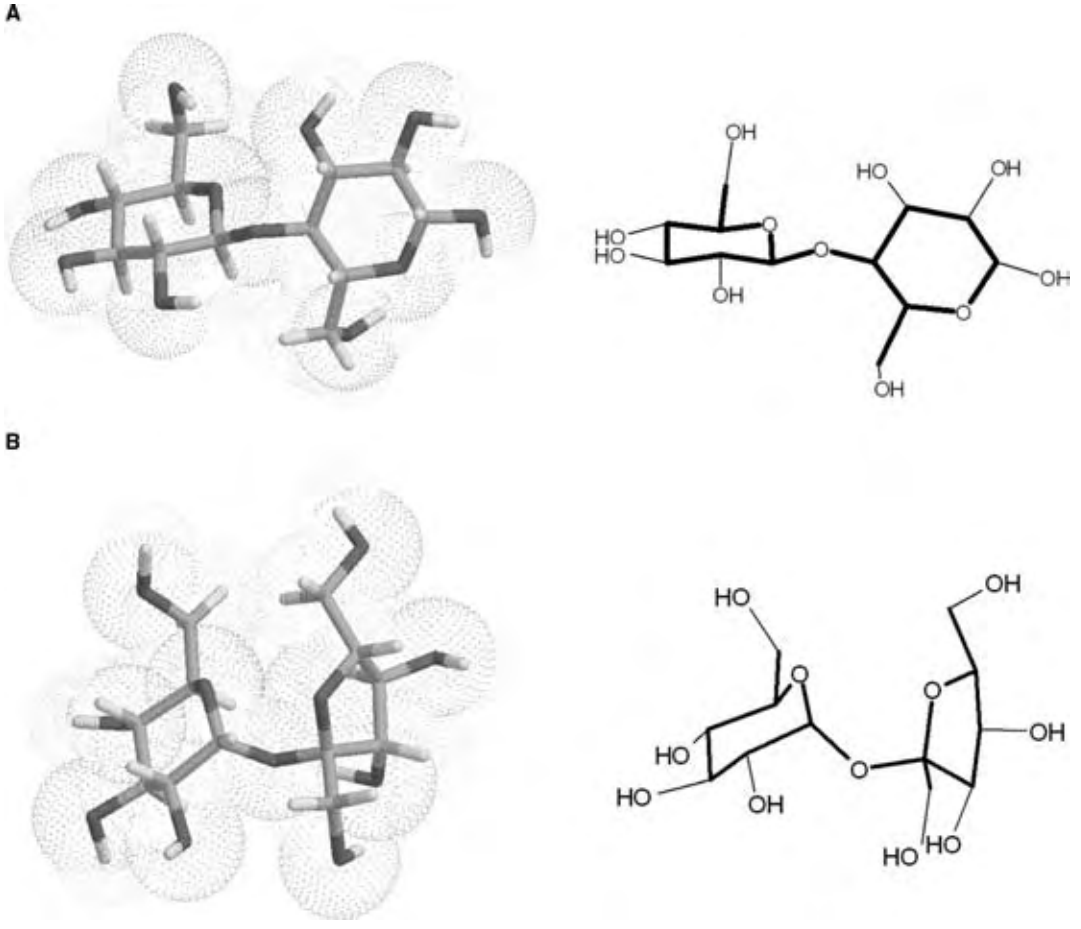


Fig. 1 Basic glycosyl residues and common glycosidic linkages: (A) β-D-Glc*p*-(1→4)-β-D-Glc*p*→β-D-glucopyranosyl-(1→4)-β-D-glucopyranose (cellobiose); (B) α-D-Glc*p*-(1→2)-β-D-Fru*f*→α-D-glucopyranosyl-(1→2)-β-D-fructofuranosid (sucrose). (Molecular modeling: SWEET, http://www.dkfz-heidelberg.de/spec/sweet2/doc/index.php. Chemistry: MDL ISIS/draw. Graphics: Macromedia Fireworks.) *(View this art in color at www.dekker.com.)*

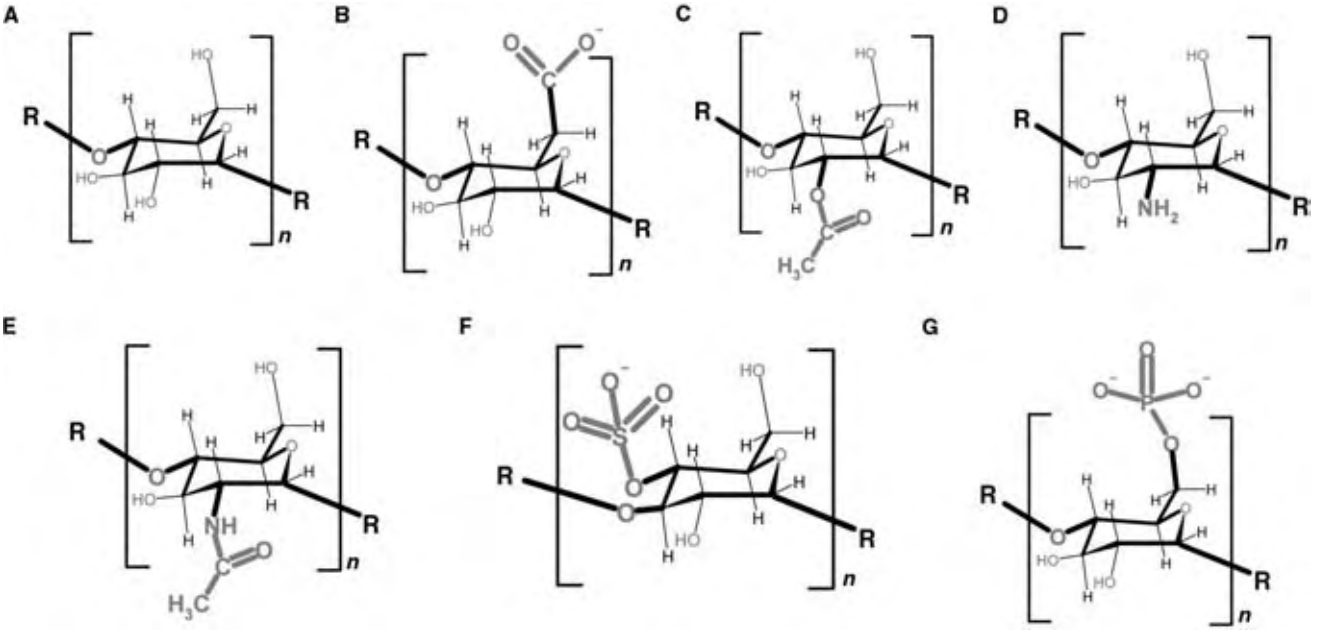


Fig. 2 Reactions at the hydroxyl groups of glycosyl residues of oligo-/polysaccharides. (A) Basic α(1→4) linked glycosyl residue. (B) Oxidation at C6 position to form uronic acid. (C) Oxidation/substitution at C2 position to form acetate. (D) Oxidation/substitution at C2 position to form glucosyl-2-amine. (E) Oxidation/substitution/compatibilization at C2 position to form glucosyl-2-*N*-acetyl. (F) Oxidation/substitution/compatibilization at C4 position compatibilization glycosyl-4-sulfate. (G) Oxidation/activation at C6 position compatibilization glucosyl-6-phosphate. (Molecular modeling: SWEET, http://www.dkfz-heidelberg.de/spec/sweet2/doc/index.php. Chemistry: MDL ISIS/draw.) *(View this art in color at www.dekker.com.)*

well-defined chemical compounds, whereas "polysaccharide" indicates multiple and superimposed heterogeneity, in particular, distributions in degree of polymerization, branching characteristics, and/or distribution of substituents and more or less oxidized glycosyl residues. Oligosaccharides and polysaccharides formed by sequences of identical or varying glycosyl residues are termed homo-oligomers/polymers and hetero-oligomers/polymers, respectively. The number of glycosyl residues is considered the degree of polymerization (dp) of a polysaccharide. However, if there is a repeating sequence consisting of several (different) glycosyl residues, this block is the reference for the degree of polymerization. The label for polysaccharides is obtained by replacing the terminal "-ose" of the glycosyl-residue parent compound with "-an." So, for instance, xylans are the homopolysaccharides formed from xylosyl residues. The terminal "-an" also indicates nonsubstituted polysaccharides. If, as for xylans, non-acetylated as well as partially acetylated forms occur, xylan refers to the non-acetylated material, and xylan acetate the acetylated derivatives. The major linkage in a polysaccharide may be indicated in the label, such as β-(1→4)-D-glycan for the glucosyl residues in cellulose. Polysaccharides entirely composed of uronic acid residues are labeled "-uronic acids." Hence, the generic name for a polysaccharide formed by glucuronyl residues is "glucuronan." However, well-established names such as cellulose, starch, inulin, chitin, amylose, and amylopectin are still used.

Although the basic qualities of polysaccharides are determined by the chemical functionality of the glycosyl residues constituting the polymers, there is another specifically polymer-related dominating property: a non-negligible excluded volume effect. Polysaccharides occupy volume; in particular, in aqueous environment, they transform rather large compartments into functional phase spaces. At the molecular level, this performance is controlled by many parameters, in particular, the following:

- Molecule geometry (they are present as regular helices and β-sheets or as irregular packed coils of varying density).
- The capacity to form polar/apolar domains.
- The optional presence of compatibility groups such as Nac.
- Pronounced and well-differentiated interactions with H_2O at different layers.
- Simple variation of local molar and mass concentrations and solubility status.

The kind of glycosyl residues and the anomeric type of glycosidic linkage in particular lead to the formation

of three basic polysaccharide conformations:

- Helical domains.
- Elongated β-sheet domains.
- Irregular, rather compact, structures of varying packing density.

The helical domains are typically stabilized by the association of several helices, internal H bonds, and strongly bound (immobilized) water. In aqueous medium, these polysaccharides tend to form dynamic supermolecular structures, in particular (soft) gels (Fig. 3A).

The elongated β-sheet domains are typically stabilized in supermolecular structures by a large number of intermolecular H bonds. Such polysaccharides are typically insoluble and form highly ordered compact structures with a rather high crystallinity index; solubility is increased if branches break the basic β-sheet symmetry (Fig. 3B).

The irregular, rather compact, structures of varying packing density are formed by more or less pronounced branching: Branches are symmetry breakers for both helical and β-sheet domains (Fig. 3C). Branching perturbs hydrophilic or hydrophobic domains and may shift material quality from ''crystalline'' to increasingly ''amorphous.'' This corresponds to a change in the preferences to stabilize their conformation: intermolecular for highly ordered helices and β-sheets; intramolecular for irregular branched structures. Whereas branching, in particular short-chain branching (scb), boosts the solubility of polysaccharides in aqueous medium, minor or no branching results in bad solubility and pronounced tendency to form supermolecular structures.

Polysaccharides may be classified in different ways. A reasonable classification refers to occurrence and dominant functionality:

- Surface/interface and bodywork structuring plant polysaccharides.
- Protective skeleton-forming chitin/chitosan of invertebrates such as arthropods.
- Extended energy repositories in eucaryotic cells, in particular those of green plants.
- Plant gums: ''response compounds'' for immediate protection upon damage or attacks and for regeneration.
- Microbial exopolysaccharides (EPSs) with cation-sequestering capabilities and protection and recognition activities.
- Functional glycosaminoglycans (GAGs), in particular, glycolipids, peptidoglycans, and proteoglycans participating in the modulation of biological process in eucaryotic cells.

POLYSACCHARIDES IN AQUEOUS ENVIRONMENT

Due to their high content of oxygen, polysaccharides are basically expected to be well soluble in polar solvents, in particular in aqueous medium. Indeed, polar groups, such as hydroxyl, uronyl, and phosphate groups, are potential partners for solvent H_2O molecules; they are also involved in the formation of inter- and intramolecular H bonds by –OH groups. An H bond is established by an H atom in a polar compound that is rather strongly attracted by two other atoms. The hydrogens in H_2O, for instance, are basically bound covalently to an oxygen atom (approximately $500\,kJ\,mol^{-1}$) but additionally may be attracted by oxygen atoms of neighboring water molecules or by similar polar attractor atoms in their environment. Such an attraction establishes a ''link,'' for instance, between two water molecules, via the H atom and is dominated by electrostatic forces (approximately 90%) with minor covalent contributions (approximately 10%). The approximately $23\,kJ\,mol^{-1}$ significantly exceeds van der Waals forces (approximately $1.3\,kJ\,mol^{-1}$). However, the electrostatic/covalent ratio is permanently fluctuating as, for instance, any influence that decreases the primary O–H bond length increases the covalent character. The de facto strength of each H bond is primarily controlled by the distance between H atom and attractor atom and decays exponentially, with minor angular dependence. Bonds exceeding 0.31 nm may be considered ''broken'' or ''not established.''[5–13]

In the natural environment, water is an integral component of polysaccharides and occurs in many forms:

- Strongly bound dense water, associated with ''amorphous,'' strongly hydrogen-bonding groups, in particular, charged groups such as uronyl, sulfate, phosphate, acetyl, or pyruvic.
- Weakly bound low-density water, associated with the dense water layers.
- Strongly bound low-density water, associated with oriented, strongly hydrogen-bonding groups, in particular, charged groups.
- Weakly bound low-density water, associated with hydrophobic groups or domains.
- Nonstabilized mobile water in hydration boundary conditions, limited to compartments within and in the vicinity of polysaccharides.

Due to these interactions, even for low polysaccharide concentrations, the aqueous medium becomes a functional phase space with stabilizing consequences for the affected compartment. This results in significantly reduced diffusion or even immobilization of

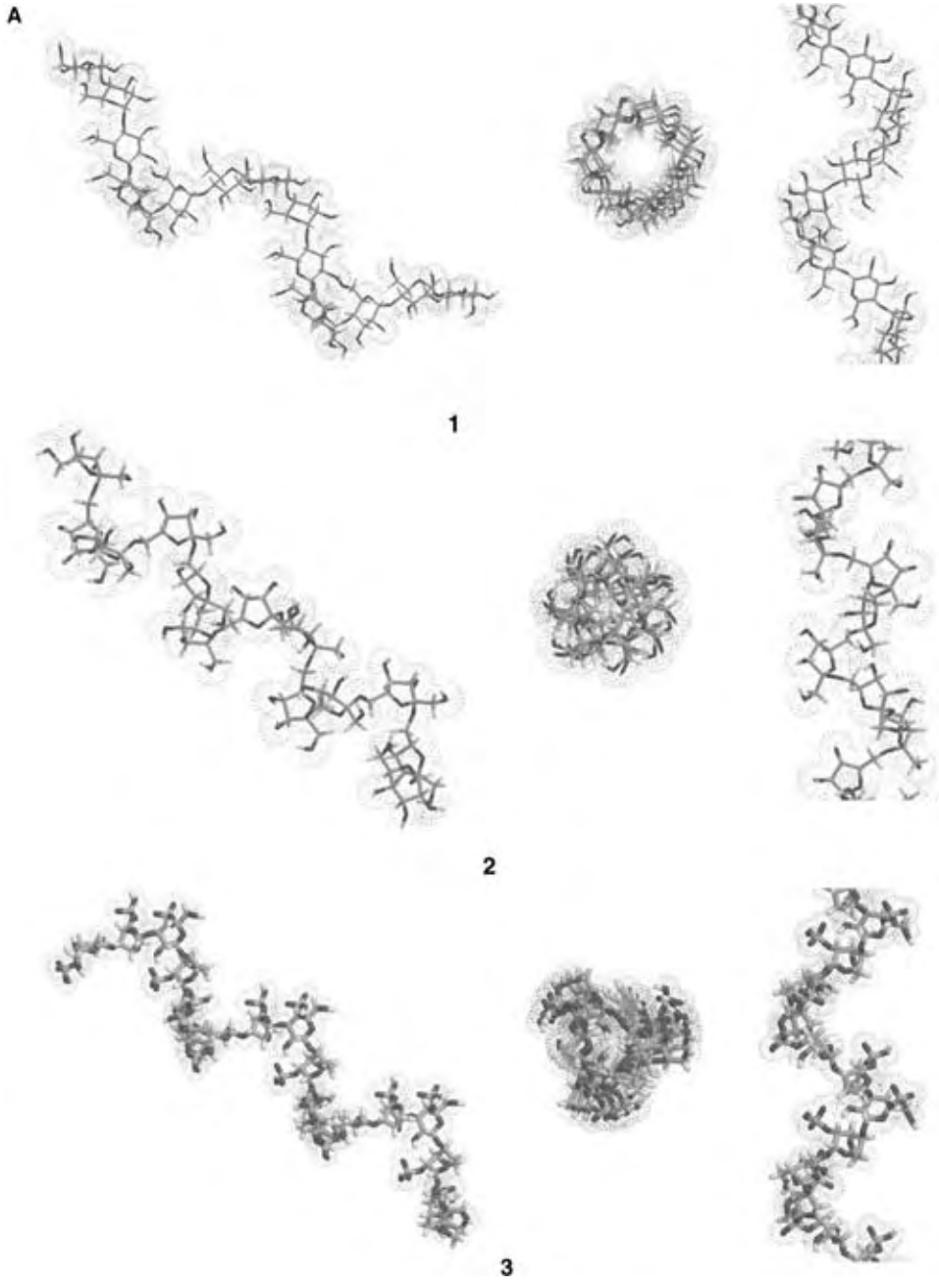


Fig. 3A The three basic polysaccharide conformations. (A) Polysaccharides forming helical structures. (**1**) Nonbranched starch component "amylose'': (1→4)-α-D-glucan; (**2**) levan: (2→6)-β-D-fructan; (**3**) carrageenan: κ form:→3)-β-D-Gal*p*-4-SO_4-(1→4)-α-D-3,6-anhydro-Gal*p*-(1→; ι form:→3)-β-D-Gal*p*-4-SO_4-(1→4)-α-D-3,6-anhydro-Gal*p*-2-SO_4-(1→; λ form:→3)-β-D-Gal*p*-2-SO_4-(1→4)-α-D-Gal*p*-2,6-SO_4-(1→. (B) Polysaccharides forming β-sheet structures. (**1**) Cellulose: (1→4)-β-D-glucan; (**2**) chitin/chitochitosan: (1→4)-β-D-Glc*p* 2-Nac/(1→4)-β-D-Glc*p* 2-NH_2; (**3**) guar: 1→4)-β-D-Man*p* backbone + each 1.5–2 Man*p* residue (6→1)-α-D-Gal*p*. (C) Polysaccharides forming irregular structures of varying packing density. (**1**) Glycogen/branched starch component "amylopectin'': (1→4)-αGlc*p* + (1→6)-αGlc*p*; (**2**) dextran: (1→6)-α-D-glucan + predominantly (1→3)-α-D-glucan branches + minor (1→4)-α-D-glucan and (1→2)-α-D-glucan branches; (**3**) sinistrin: mixeD-type and/or branched (2→1)-β-D-fructans and (2→6)-β-D-fructans. (Molecular modeling: SWEET, http://www.dkfz-heidelberg.de/spec/sweet2/doc/index.php. Chemistry: MDL ISIS/draw. Graphics: Macromedia Fireworks.) *(View this art in color at www.dekker.com.)* *(Continued next page.)*

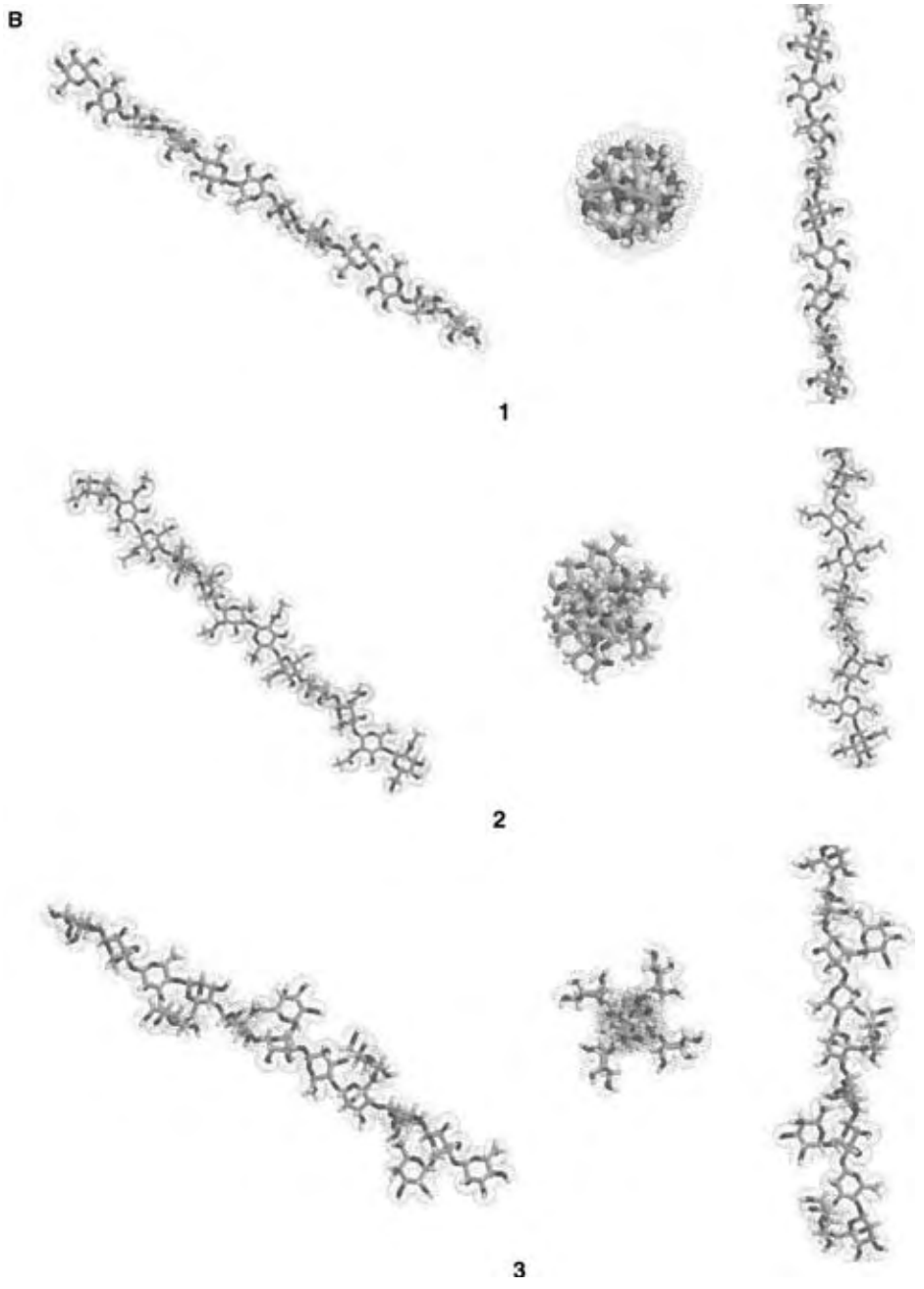


Fig. 3B *(Continued next page.)*

compounds. Phenomenologically, the formation of (soft) gels is observed. In particular, strongly bound water interacting with hydrophobic surfaces/interfaces forms junction zones and compartments of ''captured'' water clusters and increases the capability of polysaccharides to form supermolecular structures. Interaction of polar or charged groups located near hydrophobic domains with water results in isolation of these hydrophobic partitions by a zone of low-density water next to them, incapable of forming junction zones. Such structuring hydration makes polysaccharides more or less impenetrable to neighboring polymers. However, similarly hydrated polysaccharides tend to ''combine'' their hydration layers and may form rather huge supermolecular structures. If such aggregation is prevented, the alternative is phase separation and, hence, more (crystalline) or less (amorphous) ordered precipitation of polysaccharides. Hydration is controlled by equilibrium distances between polysaccharide molecules. Each equilibrium distance is the result of two antagonistic classes of forces: attractive short-range ($1/r^6$ decay) van der Waals forces and repulsive long-range ($1/r$ decay) electrostatic Coulomb forces (Fig. 4). Hence, the solubility

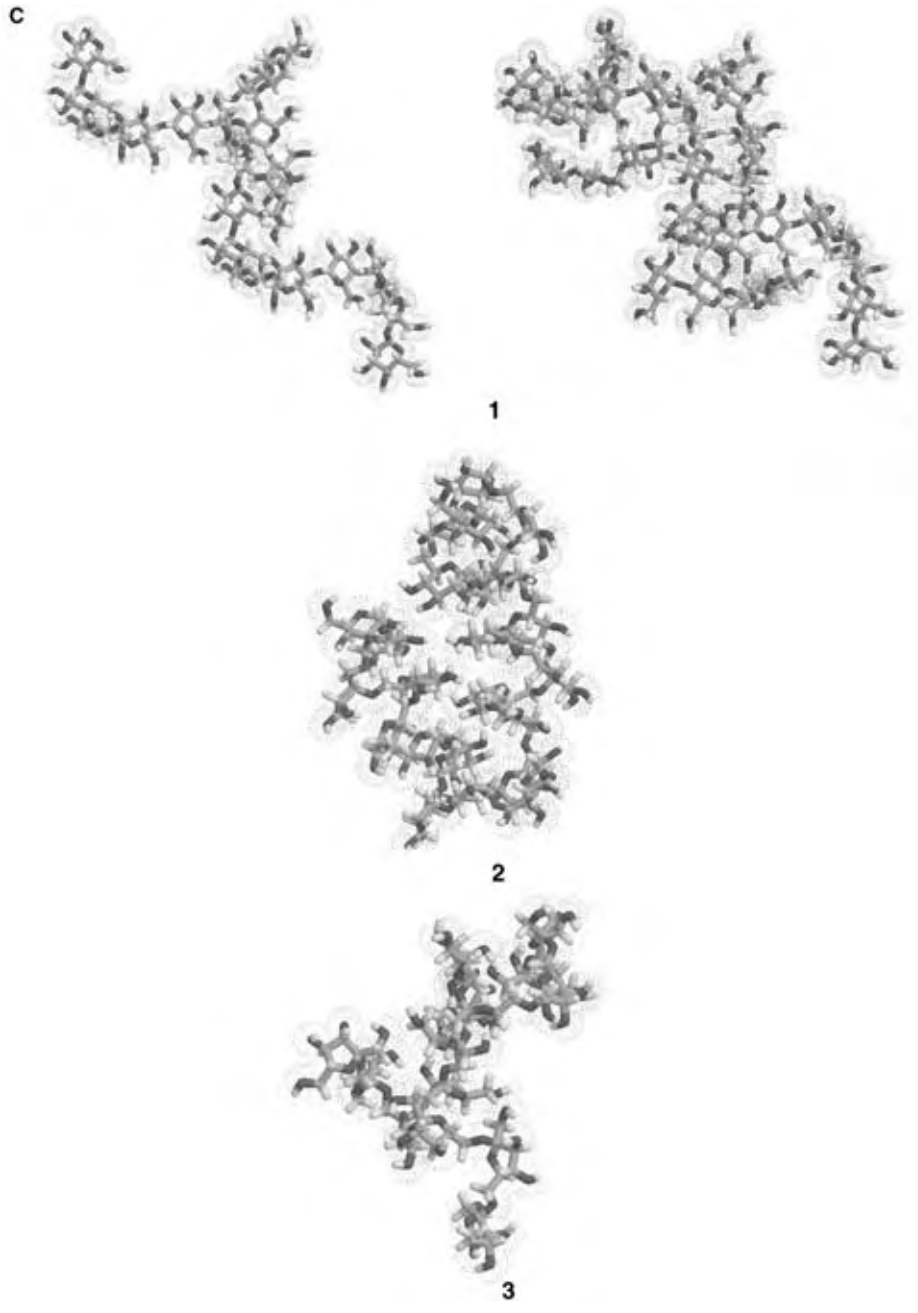


Fig. 3C *(Continued.)*

of polysaccharides in aqueous medium may be controlled by manipulation of electrostatic and van der Waals forces.

MATERIAL CHARACTERISTICS AND POLYSACCHARIDE CONCENTRATION

In particular with respect to the fact that H_2O is an integral component of the morphology and functionality of polysaccharide-containing compartments, the dependence of material characteristics on the de facto concentration or volume fraction of polysaccharides needs to be distinguished. Eq. (1) provides a scheme for a power law of controlling influences of de facto occupied/functionalized compartments (V_e):

$$V_e \propto \frac{ip}{c} md^{mc} \quad (1)$$

where V_e is the sphere-equivalent excluded volume (occupied/functionalized compartment), ip the interactive potential, c the concentration/volume fraction, md the molecular dimension (e.g., degree of polymerization), and mc the molecular conformation (e.g., helix, β-sheet, irregular).

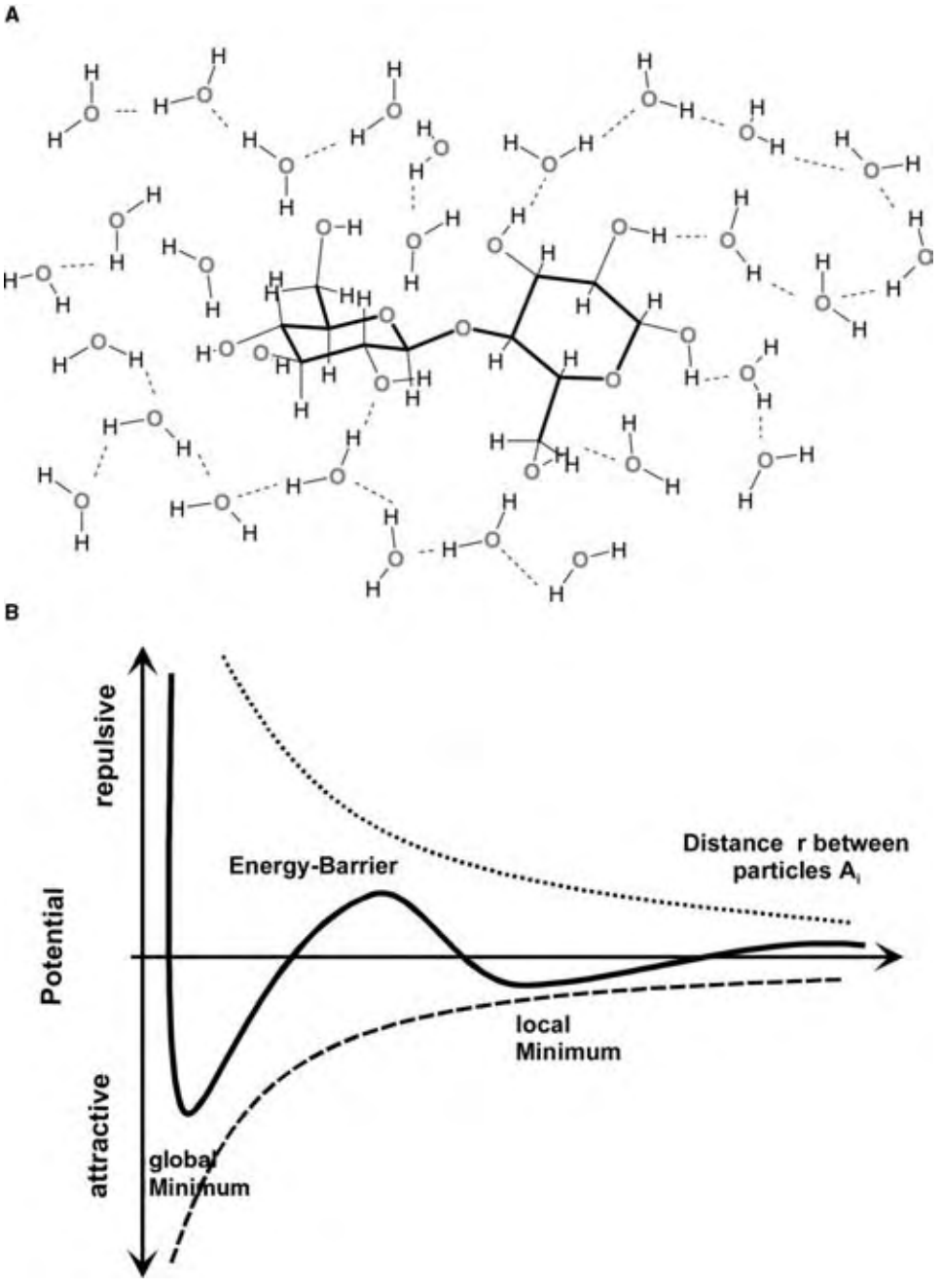


Fig. 4 Solubility of polysaccharides in aqueous medium. (A) Hydration/possible H bonds for a sequence of D-Glc*p* β(1→4)-D-Glc*p* (cellobiose). (B) Scheme for calibration of equilibrium distances due to repulsive electrostatic and attractive van der Waals forces. (Chemistry: MDL ISIS/draw. Graphics: Macromedia Fireworks.) *(View this art in color at www.dekker.com.)*

The correlation of material characteristics with de facto concentration of polysaccharides is best discussed for three regimes and limiting concentrations (Fig. 5).

In the limiting case of very dilute concentration (c_0), close to ideal conditions, interaction between polysaccharide molecules may be considered negligible, and material qualities are primarily correlated with isolated and independent conformations and dimensions of individual polysaccharide molecules.

In the range up to overlapping concentration ($c_0 - c^*$), macroscopic material qualities become increasingly influenced by polysaccharide–polysaccharide interaction phenomena. However, individual polysaccharide molecules typically do not overlap. Experimentally observed molecular and supermolecular characteristics shift away from values obtained at very dilute concentrations and become apparent values. Extrapolation with serial expansion coefficients with respect to polysaccharide concentration is an appropriate approach to obtain "ideal" ($c \rightarrow 0$) information from apparent values, and linear extrapolation works rather well, at least as a first-order approximation.

In the regime exceeding c^* ($c^* - \phi = 1$), concentration (c) is typically replaced by volume fraction (ϕ), and the polysaccharide fraction becomes equivalent

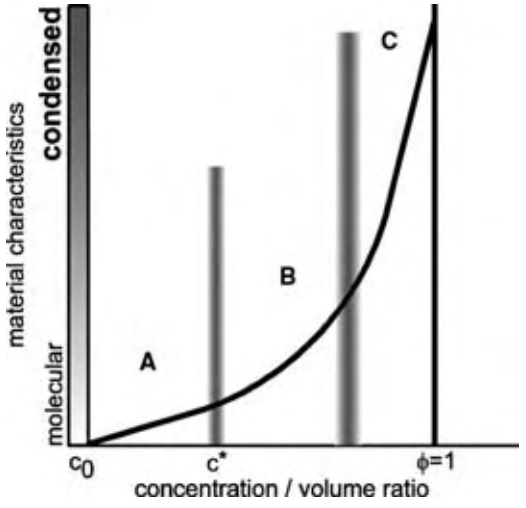


Fig. 5 Correlation of material characteristics with polysaccharide concentration. Concentration regimes [c_0: very dilute, close-to-ideal solutions; A: dilute; B: beyond overlapping concentration; C: dense systems (condensed/solid phase); volume ratio $\phi = 1$] and the consequences on observed material qualities: dependence on isolated molecule properties at close-to-ideal conditions (c_0) and increasing dominance of effects of supermolecular structures with increasing concentration. (Graphics: Macromedia Fireworks.) *(View this art in color at www.dekker.com.)*

in magnitude or even dominant with respect to total mass. Performance of such dense systems is dominated by long-range correlations, lifetime of correlations, formation and disintegration of domains, interfaces and surfaces, order of domains in terms of crystalline/amorphous ratio, and an increasing number of optional critical states with highly nonlinear response. In contrast to many other polymers, polysaccharides do not form melts at high volume ratio, but turn into interpenetrating H_2O-containing phases with complex dynamics.

Finally, for volume fractions of ϕ close to 1.0, polysaccharides typically form coexisting amorphous and crystalline phases, always containing several percent H_2O. Hence, $\phi = 1$ is practically never observed under natural conditions, and if H_2O content is reduced too much by technological processes, the system collapses and becomes ''insoluble'' forever.

HETEROGENEITY OF POLYSACCHARIDES: DISTRIBUTIONS AND MEAN VALUES

A polysaccharide is never a distinct compound, but typically consists of many similar fractions with respect to more or less any molecular characteristic. Hence, polysaccharide characteristics are characteristics of (superimposed) distributions and cannot be reduced to clearly defined, distinct values. So, for instance, polysaccharides do not have a single ''true'' molecular weight but consist of a molar mass distribution. As communication of distributions is rather tedious, mean values are often used instead, but these values must not be misunderstood as distinct ''true'' values. There are several ways to compute or determine mean values, but usually they are correlated with different moments of a distribution and indicated by either ''n'' (referring to number fractions of distributed components) or ''w'' (referring to mass fractions of distributed components). Additionally, each distribution of mass fractions may be transformed into distribution of molar fractions and vice versa if the molar mass of the fractions is known.

Hence, molecular weight, which is handled as molar mass distribution (MWD) or degree of polymerization distribution (dpD), is an important characteristic of polysaccharides (Fig. 6A). For practical purposes, molar mass distributions are expressed in either differential form (_d), with mass fractions [Eq. (2a)] or molar fractions [Eq. (2b)] represented by the normalized area (1.0), or integral form (_i), with accumulated percentages represented by the normalized height (maximum height = 1.0).

$$\int_0^\infty m(M)\mathrm{d}M = m\text{–MWD–d} = 1.0 \tag{2a}$$

$$\int_0^\infty n(M)\mathrm{d}M = n\text{–MWD–d} = 1.0 \tag{2b}$$

where MWD represents molecular weight distribution, m_ the mass fraction, n_ the molar (number) fraction, and _d the differential fraction.

Computation of the moments provides mean molar mass values, in particular, number-average molecular weight [M_n; Eq. (2c)] and weight-average molecular weight [M_w; Eq. (2d)]:

$$\text{number} - \text{average molecular weight} : \overline{M}_n = \frac{\sum n_i M_i}{\sum n_i} \tag{2c}$$

$$\text{weight} - \text{average molecular weight} : \overline{M}_w = \frac{\sum n_i M_i^2}{\sum n_i M_i} \tag{2d}$$

Additionally, molecular weight at the maximum of the differential form of distribution of mass fractions is often designated the peak molecular weight (M_p), and fraction percentages may indicate critical fractions of low- or high-molecular contributions.

Appropriate information about conformational aspects of dissolved polysaccharides may be obtained by means of a double logarithmic graph of molar mass

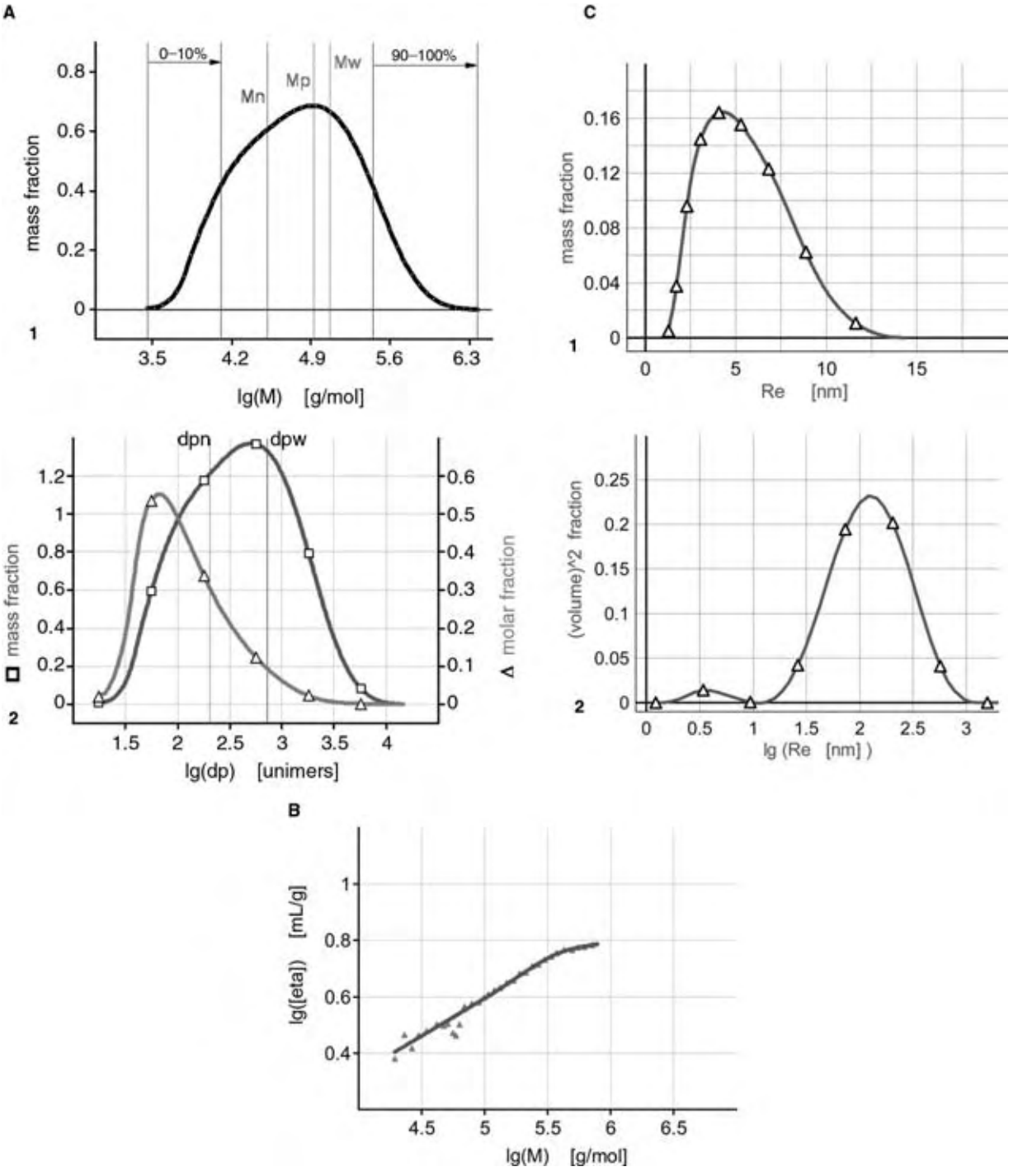


Fig. 6 (A) (**1**) Distributions, mean values, and fractions: normalized differential molar mass distribution of mass fractions (m_MWD_d); number-average molecular weight (M_n), weight-average molecular weight (M_w), molecular weight at maximum of differential molar mass distribution (M_p); 10% fraction of low molar mass components; 10% fraction of high molar mass components. (**2**) Degree of polymerization distribution derived from molar mass distribution: distribution of mass fractions (m_dpD_d), distribution of molar fraction (n_dpD_d); mean values: weight-average degree of polymerization; number-average degree of polymerization. (B) Conformation plot for aqueous, dissolved, highly short-chain, branched starch glucan: experimental data of corresponding lg(M) and lg([η]) values and least-squares fit to the data provide slope ($\equiv$ Staudinger–Mark–Houwink exponent $a = 0.29$) and intercept [$\equiv$ Staudinger–Mark–Houwink factor $K = 0.16$ (ml mol g^{-2})], indicating a rather compact, intramolecular-stabilized packing of the glucan coils. (C) (**1**) Dimensions of aqueous, dissolved, highly short-chain, branched starch glucan: sphere-equivalent radius distribution of occupied volume by constituting glucan molecules (geometric dimension) in the range of several nanometers and dimensions of coherent dynamics with contributions up to two magnitudes larger. (**2**) Transition from mass fraction to square of occupied volume and from normal to logarithmic scaling of dimensions zooms present supermolecular structures. (Data processing/graphics: a.h group CPCwin32.) *(View this art in color at www.dekker.com.)*

[lg(M)] and the corresponding intrinsic viscosity [lg([η])] (Fig. 6B). A strictly linear correlation is expected for a series of homologous polysaccharides, with the slope indicating the packing characteristics of the polymer coils: values close to 0.0 for hard/compact, spherical forms and values close to 2.0 for stiff, rodlike shapes. For irregular-shaped, uncharged, and branched polysaccharides, high packing density and, hence, small values for the slope of the conformation plot is expected; for polysaccharides with highly ordered structures established by either helices or β-sheets, or for polysaccharides with high charge density, high values for the slope are expected.

As polysaccharides tend to ''functionalize'' environmental phase space, for specification of polysaccharide dimensions, geometry and dynamics must be distinguished, although the transition is diffuse: Whereas the dimensions of polysaccharide molecules in terms of sphere equivalent radii of mean excluded volume for macroscopic periods are up to maybe 100 nm only, dynamics of coherent supermolecular structures provide sphere equivalent radii that are more than one magnitude larger and enter the micrometer range. However, these structures are ''hidden'' if mass fractions are taken for illustration. Their identification typically needs sophisticated detection and specific scaling, for instance, photon correlation spectroscopy and representation of detected populations with respect to the square of coherent (occupied) volumes (Fig. 6C).

For the case of branched polysaccharides, appropriate branching analysis is required to acquire information about general characteristics of the branching pattern (chemistry of branches, short-chain/long-chain branching, homogeneity of branching, secondary branching), percentage of branching, and mean distance between branching positions.

Controlled application of different kinds of stress, in particular, thermal, mechanical, and (bio)chemical impacts, supplies information about the response capabilities of solvent/polysaccharide systems and provides data on phase transitions, disintegration or reorganization phenomena, gel qualities, and degree of resistance or stability. But even there, any of the obtained characteristics may occur as distributions, illustrating that heterogeneity is a generic quality of polysaccharides.

SELECTED POLYSACCHARIDE SYSTEMS

Structuring Polysaccharides

Most structuring polysaccharides of higher plants are β-glycans,[14] in particular, the glycans in wood and cereals. The most abundant is a system consisting of cellulose, callose, arabinogalactans, hemicelluloses, and lignin, occurring, for instance, in wood.

Cellulose

Cellulose is a strictly nonbranched β(1→4)-linked glucan formed by the repeating dimer cellobiose (Fig. 7A). It forms microfibrils with a length of 100–x0 000 nm and a diameter between 2 and 20 nm. Cellulose fibers contain highly ordered (crystalline) and minor regular (amorphous) domains and, hence, form stiff construction material with limited flexibility.

Callose

The nonbranched β(1→3)-linked glucan callose (Fig. 7B) forms helices and is a transient cell wall component with high flexibility. It is found in growing zones; synthesis of callose is also an important response mechanism to applied physical or chemical stresses and pathogens.

Hemicelluloses

Hemicelluloses are a mix of β-linked heteroglycans with compositions that are strongly dependent on plant variety and species. They primarily consist of fucogalactoxyloglucans (XyG), glucuronoarabinoxylan (GAX), xylan (Xyl), mannan (Man), glucomannan, and galactomannan. Although not typical hemicelluloses, arabinogalactans, consisting of β(1→3)-D-galactopyranosyl residues with a high number of galactosyl and arabinosyl branches, are rather similar in functionality. Hemicellulose and arabinogalactans are closely associated with cellulose, callose, and lignin in primary and secondary layers of cell walls and are a kind of connective filling material (Fig. 7C).

Lignin/lignocellulose

Lignin primarily is formed by irreversible elimination of H_2O from carbohydrates and glycans. The actual chemical composition of lignin depends on the carbohydrate source. However, the constituting units are phenolic compounds containing alkyl residues, phenolic hydroxyl groups, carbonyl groups (aldehyde/ketone), and/or alcoholic hydroxyl groups. A three-dimensional skeleton is formed by irregular crosslinking of phenolic fragments and compounds via ether between alcoholic hydroxyl groups and by hemiacetal/hemiketal linkages of hydroxyl groups with carbonyl groups (Fig. 7D).

Pectin

Another structuring, ''compartment-connecting'' polysaccharide, in particular of fruit and vegetables, is

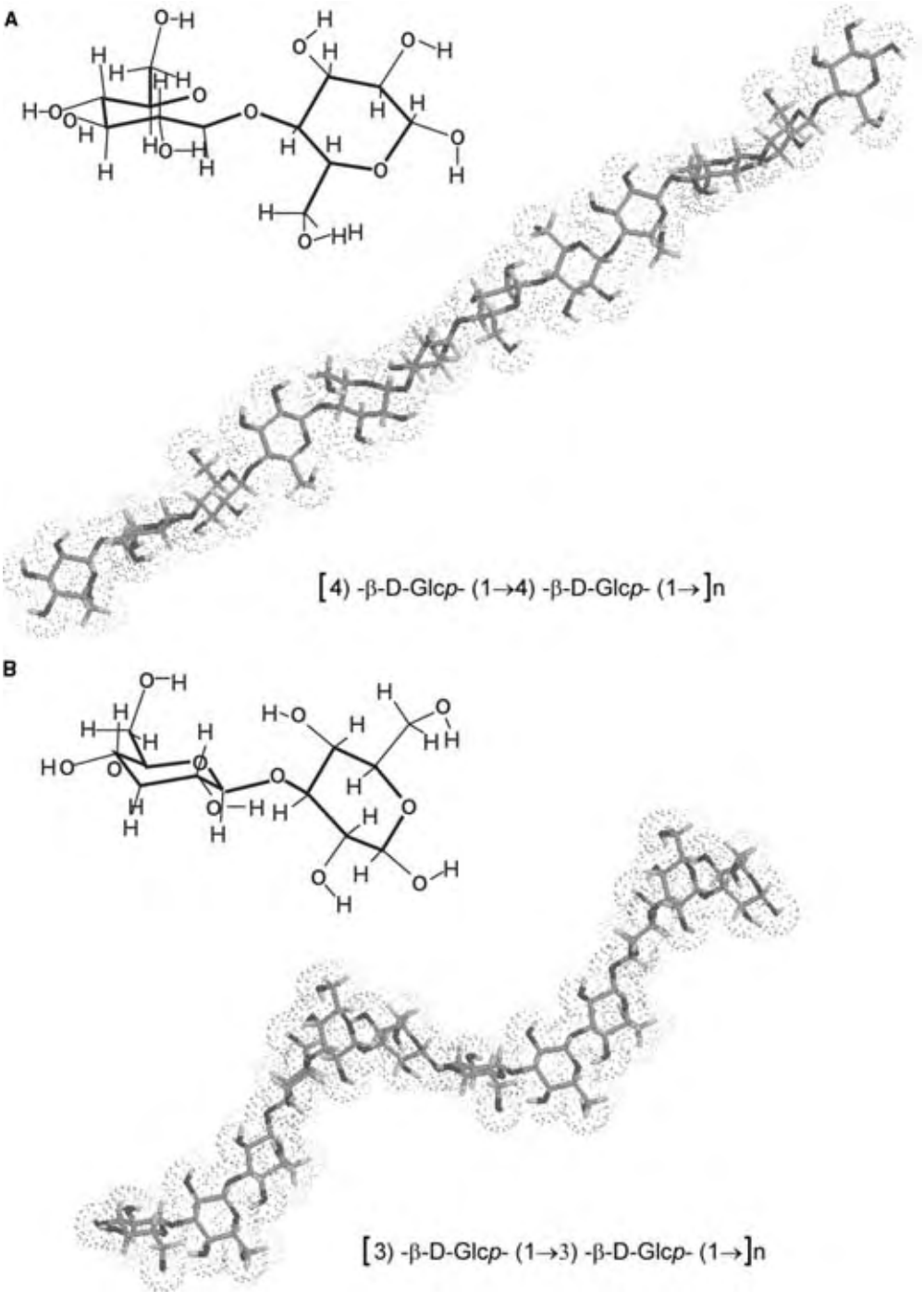


Fig. 7A, B Structuring polysaccharide system, particularly in wood (cellulose, callose, galactomannan, xylan, arabinogalactan, lignin). (A) Cellulose β-sheet sequence produced by β(1→4) glycosidically linked glycosyl residues forming highly ordered ''crystalline'' and minor regular/irregular ''amorphous'' fiber domains. (Molecular modeling: SWEET, http://www.dkfz-heidelberg.de/spec/sweet2/doc/index.php. Chemistry: MDL ISIS/draw. Graphics: Macromedia Fireworks.) (B) Callose helix sequence formed by β(1→3) glycosidically linked glycosyl residues. (Molecular modeling: SWEET, http://www.dkfz-heidelberg.de/spec/sweet2/doc/index.php. Chemistry: MDL ISIS/draw. Graphics: Macromedia Fireworks. Data processing: a.h group CPCwin32.) (C) Variably associated branched xylans and arabinoglycans. (Molecular modeling: SWEET, http://www.dkfz-heidelberg.de/spec/sweet2/doc/index.php. Chemistry: MDL ISIS/draw. Graphics: Macromedia Fireworks.) (D) (**1**) Basic phenolic compounds with alkyl, keto-, and aldehyde residues formed by irreversible elimination of H_2O from carbohydrates and glycans; (**2**) Crosslinked to alcoholic OH-groups by ether, hemiacetal, and hemiketal linkages forming an irregular 3dim skeleton. (Molecular modeling: The DundeePRODRG Server, http://davapc1.bioch.dundee.ac.uk/programs/prodrg/prodrg.html. Chemistry: MDL ISIS/draw.) *(View this art in color at www.dekker.com.)* *(Continued next page.)*

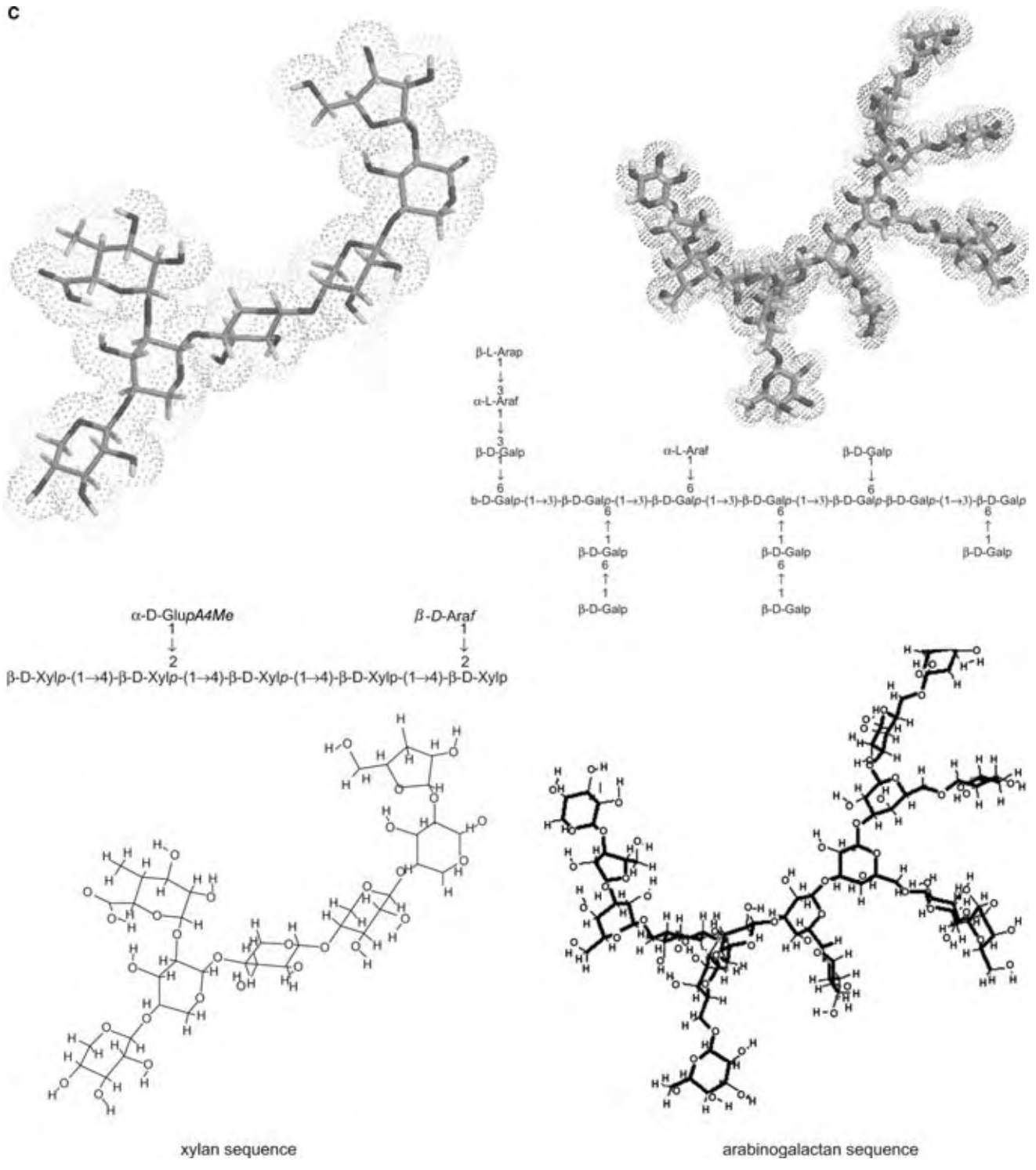
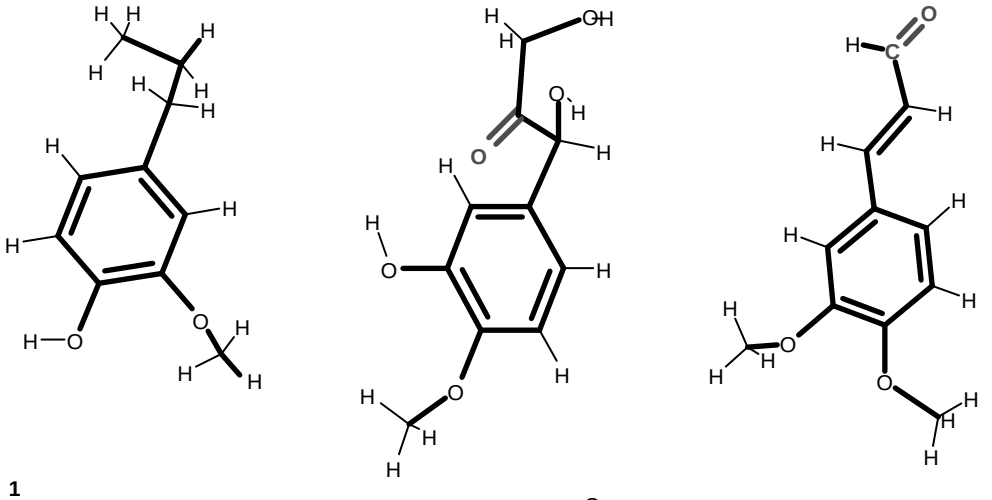
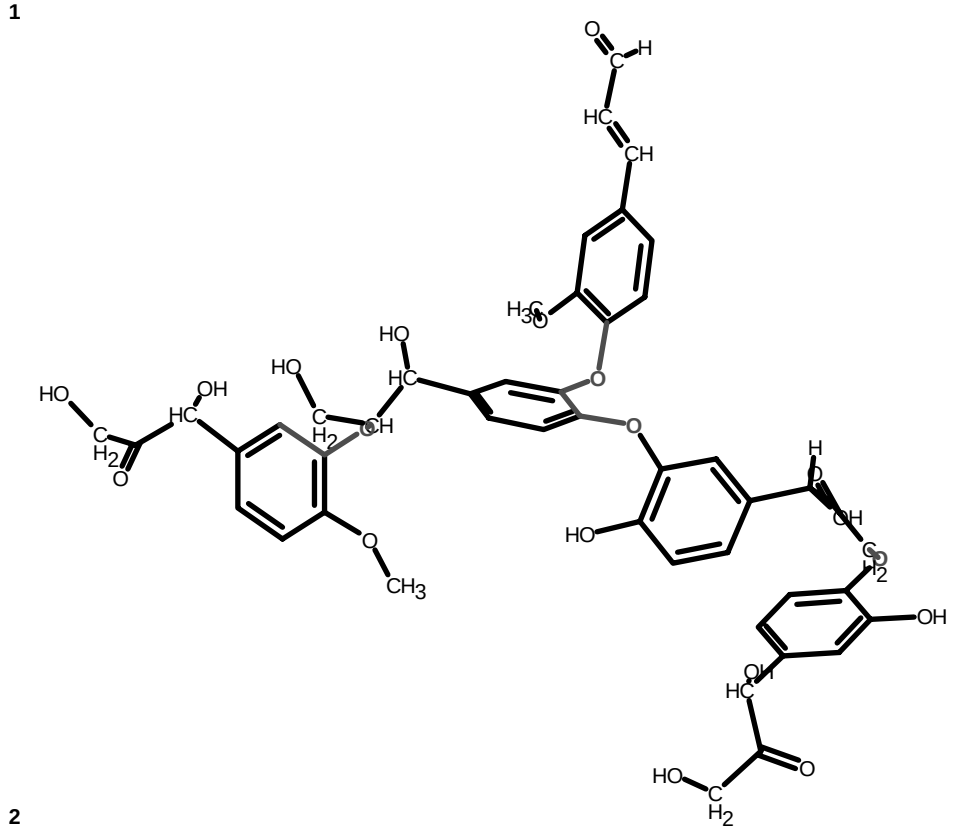


Fig. 7C *(Continued next page.)*

pectin, which consists of partially methylated α-(1→4)-D-galacturonyl (''smooth'') residues and alternating α-(1→2)-l-rhamnosyl-α-(1→4)-D-galacturonosyl (''hairy'') sections with dp 1–20 branches of neutral (primarily l-arabinose, D-galactose, D-xylose, and l-fucose) and charged [D-glucuronic acid, D-apiose, 3-deoxy-D-manno-2-octulosonic acid (Kdo), and 3-deoxy-D-lyxo-2-heptulosonic acid] glycosyl residues. Pectins are soluble in aqueous medium, but form sticky gels in the presence of divalent cations.



1



2

Fig. 7D *(Continued.)*

Chitin/chitosan

The major structuring polysaccharides of fungal cell walls and the exoskeletons of invertebrates, such as insects and crustaceans, are chitin/chitosan (Fig. 3B): (1→4)-β-D-Glc*p* 2-Nac/(1→4)-β-D-Glc*p* 2-NH_2. Chitin and chitosan differ in their degree of acetylation, and samples are typically specified by the ratio of glucosamine residues and the sum of glucosamine and *N*-acetyl-glucosamine residues.

Algae form and contain a high percentage of structuring polysaccharides that constitute cell wall layers and mucilage.

Agar

Most species of red algae (Rhodophyta) have a cell wall with an inner cellulose layer embedded in a matrix of agarose and agaropectin mucilage. This embedded

matrix causes the slippery texture of red seaweed and is commonly known as agar-agar or, simply, agar. Agarose is a virtually neutral polysaccharide, whereas agaropectin contains acidic components. Agarose is a galactan formed by the agarobiose repeating disaccharide [→3)-βD-Gal*p*-(1→4)-3,6-anhydro-αL-Gal*p*-(1→]. Agaropectin consists of an identical backbone and considerable amounts of acid groups such as sulfate, pyruvate, and uronyl residues. The composition of agar additionally varies with the source of the red algae.[15]

Alginate

In spite of the general low level of organization of algae, cellulose is the dominant structural element of their cell walls. However, in brown algae (Phaeophyta), alginic acid/alginates are important cell-wall components. Alginates are constructed from blocks of mannuronyl residues [→4)-β-D-Man*p*A-(1→] and guluronyl residues [→4)-α-L-Gul*p*A-(1→]. The β(1→4)-linked D-mannuronyl blocks form left-handed triple helices, whereas the α(1→4)-linked L-guluronyl blocks form double helices. The guluronate block has specific binding sites for divalent cations, which results in egg-box-like complexation and, hence, gel formation of alginates. Gel-forming capacity, however, strongly depends on block lengths and the kind of divalent cations present ($Mg^{2+} \ll Ca^{2+} < Sr^{2+} < Ba^{2+}$).[16]

Carrageenan

Carrageenan (Fig. 3A) is found in red seaweed (Rhodophyta), primarily in species *Chondrus*, *Gigartina*, *Eucheuma*, and *Phyllophora*. Carrageenan can be split into three major fractions: κ-carrageenan [→3)-β-D-Gal*p*-4-SO_4-(1→4)-α-D-3,6-anhydro-Gal*p*-(1→], which is insoluble and forms gel in the presence of aqueous KCl, ι-carrageenan [→3)-β-D-Gal*p*-4-SO_4-(1→4)-α-D-3,6-anhydro-Gal*p*-2-SO_4-(1→], which is soluble in aqueous K^+ solutions but forms gel in the presence of Ca^{2+}, and λ-carrageenan [→3)-β-D-Gal*p*-2-SO_4-(1→4)-α-D-Gal*p*-2,6-SO_4-(1→], which is soluble in the presence of both mono- and divalent cations. Carrageenans are flexible molecules that tend to form double-helical structures.[17]

Polysaccharides as Energy Repository

Energy-storage polysaccharides are formed by eucaryotic organisms during their growing and maturation periods as long external input is provided. If energy input is terminated (e.g., by harvesting or during starvation), the storage polysaccharides are degraded and provide energy to maintain metabolism. The most important storage polysaccharides of higher plants are the following.

Starch glucans

Starch glucans consist of nonbranched (nb) (Fig. 3A), long-chain branched (lcb), and short-chain branched (scb) (Fig. 3C) components in varying compositions. They are organized in granules with minor amounts of lipids, varying amounts of protein, and immobilized water and, depending on the source, might even be phosphorylated. In aqueous medium, starches, in particular, nb and lcb components, have a pronounced tendency to form supermolecular structures. Depending on the kind of branching pattern and degree of branching, starch glucans form highly ordered helices and multiple hexagonal (crystalline) supermolecular structures or transition structures from helices to irregular-shaped, increasingly densely packed amorphous coils.

Fructans

Fructans are oligomers and polymers formed by polymerizing fructose from transport metabolite sucrose on one out of three possible starter trioses: 1-kestose yields inulin-type (2→1)-β-D-fructans, 6-kestose yields levan-type (2→6)-β-D-fructans (Fig. 3A), and neo-kestose yields mixed-type and/or branched (2→1), (2→6)-β-D-fructans (Fig. 3C). Pronounced fructan metabolism is found in composites (chicory, Jerusalem artichoke), Liliaceae (onion, chives, garlic), cereals (wheat, barley, rye, oat), Asparagaceae (asparagus), Amaryllidaceae (banana), and Agavaceae (agave).

Plant Gum Polysaccharides

Plant gums are acidic polysaccharides in the aqueous environment that can form highly viscous gels in the presence of cations, in particular divalent Ca^{2+} and Mg^{2+}. The formation of plant gums is typically a response to damage or pathogen attack. The best-known gums are the following.

Gum arabic (gum acacia)

Gum arabic (gum acacia) is primarily formed by the gumnosis process by the species *Acacia senegal* (Mimosaceae) and consists of neutral glycosyl residues (L-rhamnose, L-arabinose, D-galactose) and glucuronyl residues (D-glucuronic acid, 4-methoxy-D-glucuronic acid). Gum arabic occurs as a hydroxyproline-rich glycoprotein (~2% protein) containing a motif of 19

amino acid residues (-ser-hyp[a]-hyp[a]-hyp[a]-thr-leu-ser-hyp[b]-ser-hyp[b]-thr-hyp-thr-hyp[a]-hyp[a]-hyp[a]-gly-pro-his-) with contiguous hydroxyprolines ([a]) attached to oligo-α-(1→3)-L-Ara*f* and noncontiguous hydroxyprolines ([b]) attached to a backbone of (1→3)-β-D-Gal*p* residues that is heavily branched with l-Rha*p*-D-Glc*p*A-L-Ara*f*-D-Gal*p* pentasaccharide.[18]

Traganth

Traganth is formed by leguminous species *Astragalus gummifer* primarily in the Near East consists of two fractions: water-soluble tragacanthin, constituted by D-Gal*p*A, D-Gal*p*, L-Fuc*p*, D-Xyl*p*, and L-Ara*f*, and water-insoluble bassorin, with predominantly D-galacturonyl methylesters.[19]

Guar/locust bean gum

Guar and locust bean gum (Fig. 3B) are galactomannans, formed by legume species *Cyamopsis tetragonolobus* (guar) primarily in Pakistan and India and *Ceratonia siliqua* (locust bean) primarily in the Mediterranean region. Guar and locust bean gum consist of a (1→4)-β-D-Man*p* backbone and statistically distributed short branches of (6→1)-α-D-Gal*p* (every 1.5–2 Man*p* residues) in case of guaran and branching at each fourth to fifth Man*p* residue in case of locust bean polysaccharide. Whereas the branches of guar are alternately disposed along the D-mannan backbone, those of locust bean gum are arranged in uniform blocks.[20]

Microbial Polysaccharides

More or less all micro-organisms form polysaccharides, either as cell wall (capsular polysaccharides, CPSs) and/or as metabolic products that are secreted into the extracellular medium (exopolysaccharides, EPSs). Capsular polysaccharides contribute to the morphology and stability of cell walls and also establish communication with adjacent cells and controlled interactions with (bio)chemical agents in the medium. Exopolysaccharides are involved in cell protection against toxic or limiting environments, sequestering of essential cations, colonization, and cellular recognition. The typically excellent water-binding capability of EPSs protects bacteria in low-moisture environments and inhibits amebic attack or phagocytosis.[21,22] In a laboratory environment, many microbial polysaccharides may be obtained rather pure and even in large, industrial-scale amounts by controlled biotechnological processing and optional prior genetic modification of microbes.

Xanthan

Xanthan is a (1→4)-β-D-Glc*p* backbone with oxidized trisaccharide branches (Man-Glc*p*A-Man) at each cellobiose residue industrially produced in large scales by *Xanthomonas campestris* from glucose, sucrose, starch, lactose, or whey feedstock. Particular due to its pronounced viscoelastic properties, xanthan is widely used as a thickener and stabilizer in food [it is listed as ''generally recognized as safe'' (GRAS) by the U.S. Food and Drug Administration (FDA) and non-food applications.[23] Under natural conditions, however, *Xanthomonas* species are plant pathogens, in particular for cultivated rutaceous species (including citrus), rice, beans, grape, and cotton, causing citrus canker.

β-Glucans

β-Glucans[24–27] are a group of basically β(1→3)-linked glucans from microbial sources, particularly of interest as they attach to specific iC3b receptors (CR3, CD11b/CD18) of phagocytic cells, stimulating phagocytosis and/or cytotoxic degranulation. Curdlan, produced by *Agrobacterium* spp. and *Alcaligenes* spp., is a strictly nonbranched β(1→3) glucan. Grifolan (GRN), made by *Grifola frondosa*, is a β(1→3)-glucan with (1→6)-β-D-Glc*p*-branches at approximately each third residue. Lentinan, made by *Lentinus edodes*, is a β(1→3)-glucan with two out of five glucosyl residues containing (1→6)-β-D-Glc*p*-branches. Schizophyllan, also known as SPG, sonifilan, or sizofilan, made by *Schizophyllum commune*, *Sclerotinia sclerotiorum glycan* (SSG), are β(1→3)-glucans with (1→6)-β-D-Glc*p*-branches at approximately each third residue. Scleroglucan, made by *Sclerotium glucanicum* from a mixed glucose/nitrate/mineral feedstock, is a β(1→3) glucan with β(1→6)-linked glucopyranosyl branches at approximately each third glycosyl residue. Beside its attractiveness as a pharmaceutical compound, scleroglucan is industrially produced on a large scale and used as drilling mud, asphalt emulsifier, adhesive, printing ink, and a component of cosmetics.

Dextran

Dextran (Fig. 3C) is a (1→6)-α-D-glucan with dominantly (1→3)-α-D-glucan branches and minor (1→4)- and (1→2)-α-D-glucan branches, industrially produced from sucrose feedstocks, primarily by *Leuconostoc mesenteroides* but also by *Streptococcus* and *Lactobacillus* spp. Crude dextran and dextran fractions are soluble in aqueous medium and have been applied in a wide range of pharmaceutical/clinical and food applications for decades.

Pullulan

Pullulan is a glucan of (1→6)-α-D-linked maltotriose residues produced by *Aureobasidium pullulans* from starch or sucrose feedstock.[28] It is well soluble in aqueous medium and forms films that possess thermal stability and are antistatic and elastic. Pullulan has adhesive properties and is directly compressible under heat with moisture. The polysaccharide is produced on an industrial scale and is used for multiple applications, including food, pharmaceuticals, and cosmetics. It has recently been classified as GRAS by the U.S. FDA.

Functional Glycosaminoglycans of Eucaryotic Cells

Together with proteins, nucleic acids, lipids, and water, glycans are basic constituents of eucaryotic cells. The short-term reserve compound glycogen, primarily in liver and muscle cells, is a pure form. Most functional glycosaminoglycans (GAGs), however, are linked to lipids (glycolipids) and proteins (peptidoglycans, proteoglycans) in the secretory pathway, where they participate in the modulation of biological processes and in the development and functioning of many physiologic systems. Approximately 1% of the genes in the mammalian genome are involved in the formation, transformation, and degradation of glycans. Many of these genes encode glycosyltransferases and glycosidases for the Golgi apparatus, where most of the surface/interface glycans and glycans of extracellular compartments are formed.[29]

Glycogen

Glycogen is an extremely α-(1→6)-short-chain branched α-(1→4)-D-glucan made by branching enzyme from prior-formed (1→4)-α-D-glucans. The major difference between glycogen of vertebrates and short-chain branched amylopectin of plants is the higher percentage of branching and statistical distribution of branching positions in vertebrates compared to clustering of fewer branches in plants. Glycogen is stored preferentially in liver and muscle cells as an immediately available short-time energy resource.

Glycosaminoglycans/proteoglycans

Glycosaminoglycans and proteoglycans[30–36] are part of mammalian core proteins with sulfated glycosaminoglycans. The core proteins direct biosynthesis, and, hence, the specific morphology and functionality of GAGs. The glycan moiety of GAGs is formed by disaccharides consisting of uronic acid or galactose and hexosamines. Chondroitin sulfate/dermatan sulfate consists of glucuronic acid/iduronic acid-β(1→3)-*N*-acetylgalactosamine sulfate and may be classified as chondroitin A [→4)-β-D-Glc*p*A-(1→3)-β-D-Gal*p*Nac-6-O-SO_3-(1→], chondroitin B, also known as dermatan sulfate [→4)-β-D-Ido*p*A-(1→3)-β-D-Gal*p*Nac-4-O-SO_3-(1→], and chondroitin C [→4)-β-D-Glc*p*A-(1→3)-β-D-Gal*p*Nac-4-O-SO_3-(1→]. Chondroitin sulfate is attached to proteoglycan's core protein via a Xyl-Gal-Gal trisaccharide; sulfate of type C additionally shows pronounced interaction with collagen and other extracellular molecules. The repeating unit of heparan sulfate/heparin heparin is [→4)-β-D-Glc*p*A-2-SO_3-(1–3)-β-D-Gal*p*-2NH-SO_3,6-SO_3-(1–4)-β-D-Ido*p*A-2-SO_3-(1–3)-β-D-Gal*p*-2NH-SO_3,6-SO_3-(1→]. As there is no structural difference between heparin and heparan sulfate, typically the amount of *N*-sulfatation is taken as the criterion for classification: heparan sulfate $<$ 50% N-O-SO_3 $>$ heparin. Keratan sulfate consists of the repeating disaccharide galactose-*N*-acetylglucosamine [→4)-β-D-Gal*p*-(1→4)-β-D-Gal*p*Nac-6-O-SO_3-(1→] and occurs in mammalian tissue in two forms, distinguished by molecular dimension and protein linkage (type I: small; *O*-glycosidically linked to Ser and Thr; type II: huge; *N*-glycosidically linked to Asp). Hyaluronan/hyaluronic acid is a GAG formed by the repeating disaccharide [→4)-β-D-Glc*p*A-(1→3)-β-D-Glc*p*Nac-(1→]. Although the biological function of hyaluronan is not completely clear, there are many tasks correlated with this most simple GAG. These comprise, for instance, the ability to immobilize different molecules/compounds at certain biological locations, support activities in cell morphogenesis and cell differentiation, inhibit cell activities on inflammation combined with simultaneous activation of phagocytosis, and protect cells against lymphocytes and viruses.

CONCLUSIONS

Polysaccharides are major renewable resources with a wide field of functional applications. They are made of glycosidically linked basic or redox-modified glycosyl residues by organisms at any level of organization on Earth. From an economic point of view, polysaccharides are abundantly available globally and, hence, are cheap raw materials. However, industrial processing is still limited to a few compounds such as cellulose and starch. From a biological point of view, polysaccharides are high-performance materials with enormous variability at molecular level, efficient sensor and amplification quality under varying conditions in an aqueous environment, smart response capability to applied stress, and assembly and transformation at the molecular level by molecular-scale tools triggered by the principles of self-organizing complex systems. This statement contains the potential explanation for

the problems faced during industrial handling of polysaccharides, as the huge variety of these materials, the unpredictable raw material qualities, and the high amount of ''impurities'' makes them an expensive raw material, because appropriate handling requires high-technology and variable processes.

The basic chemistry of polysaccharides is rather simple; however, drastically enhanced understanding of correlations between functional performance and molecular and supermolecular properties in natural systems is required, in particular information about:

- The capacity of plant polysaccharides to structure flexible and functional surfaces, interfaces, and biological compartments.
- Construction details of protective skeletons of invertebrates such as insects and crustaceans.
- The management of energy repositories in eucaryotic cells.
- Control mechanisms to establish protective gels and support regeneration upon damage or attack.
- Specific sequestration mechanisms from media and the capability to establish protection and recognition structures for microbial organisms.
- Mechanisms to modulate biological processes of eucaryotic cells.

In essence, the key to a better understanding of polysaccharide performance is understanding their interaction with water. Polysaccharides are capable of integrating water and forming functional phase spaces, and characterization of such systems requires consideration of geometric molecular-level dimensions as well as supermolecular distances of coherent dynamics.

Heterogeneity is a generic quality of polysaccharides: All characteristics occur as distributions and may be handled as distributions of molar fractions, referring to the number of distributed components, or distributions of mass fractions, referring to mass contributions of distributed components. In particular, for broad distributions, the difference between mass and molar distribution becomes significant and sometimes crucial. Molar mass distribution is a central piece in this puzzle of correlating molecular characteristics with polysaccharide performance. Additionally, optional branching characteristics, substitution patterns, and responses of aqueous polysaccharide systems to different kinds of applied stress need to be determined.

REFERENCES

1. IUPAC–IUB Joint Commission on Biochemical Nomenclature (JCBN). Polysaccharide nomenclature (Recommendations 1980). Pure Appl. Chem. **1982**, *54*, 1523–1526.
2. IUPAC–IUB Joint Commission on Biochemical Nomenclature (JCBN). Symbols for specifying the conformation of polysaccharide chains (Recommendations 1981). Pure Appl. Chem. **1983**, *55*, 1269–1272.
3. IUPAC–IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature of glycoproteins, glycopeptides and peptidoglycans. Pure Appl. Chem. **1988**, *60*, 1389–1394.
4. IUPAC–IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature of glycolipids. Pure Appl. Chem. **1997**, *69*, 2475–2487.
5. Chaplin, M.F. A proposal for the structuring of water. Biophys. Chem. **2000**, *83*, 211–221.
6. Kirschner, K.N.; Woods, R.J. Solvent interactions determine carbohydrate conformation. Publ. Natl. Acad. Sci. USA **2001**, *98*, 10541–10545.
7. Yaminsky, V.V.; Vogler, E.A. Hydrophobic hydration. Curr. Opin. Colloid Interf. Sci. **2001**, *6*, 342–349.
8. Widom, B.; Bhimalapuram, P.; Koga, K. The hydrophobic effect. Phys. Chem. Chem. Phys. **2003**, *5*, 3085–3093.
9. Du, Q.; Freysz, E.; Shen, Y.R. Surface vibrational spectroscopic studies of hydrogen bonding and hydrophobicity. Science **1994**, *264*, 826–828.
10. Scatena, L.F.; Brown, M.G.; Richmond, G.L. Water at hydrophobic surfaces: weak hydrogen bonding and strong orientation effects. Science **2001**, *292*, 908–912.
11. Muller, N. Is there a region of highly structured water around a nonpolar solute molecule? J. Solut. Chem. **1998**, *17*, 661–672.
12. Suresh, S.J.; Naik, V.M. Hydrogen bond thermodynamic properties of water from dielectric constant data. J. Chem. Phys. **2000**, *113*, 9727–9732.
13. Khan, A. A liquid water model: density variation from supercooled to superheated states, prediction of H-bonds, and temperature limits. J. Phys. Chem. **2000**, *104*, 11268–11274.
14. Karnezis, T.; McIntosh, M.; Wardak, A.Z.; Stanisich, V.A.; Stone, B.A. The biosynthesis of β-glycans. Trends Glycosci. Glycotechnol. **2000**, *12* (66), 211–227.
15. Lahaye, M. Chemistry and physico-chemistry of phycocolloids. Cah. Biol. Mar. **2001**, *42*, 137–157.
16. Draget, K.I.; Skjak-Braek, G.; Smidsrod, O. Alginic acid gels: the effect of alginate chemical composition and molecular weight. Carbohydr. Polym. **1994**, *25*, 31.
17. Lawson, C.J.; Rees, D.A. An enzyme for the metabolic control of polysaccharide conformation and function. Nature **1979**, *227*, 392.
18. Goodrum, L.J.; Patel, A.; Leykam, J.F.; Kieliszewski, M.J. Gum arabic glycoprotein contains

glycomodules of both extension and arabinogalactan-glycoproteins. Phytochemistry **2000**, *54*, 99.
19. Specifications for Identity and Purity of Certain Food Additives, FAO Food and Nutrition Paper No. 49; 1990; 735.
20. Hoffman, J.; Lindberg, B.; Painter, T. The distribution of D-galactose residues in guaran and locust-bean gum. Acta Chem. Scand. Ser. B **1976**, *30*, 365.
21. Looijesteijn, P.J.; Trapet, L.; de Vries, E.; Abee, T.; Hugenholtz, J. Physiological function of exopolysaccharides produced by Lactococcus lactis. Int. J. Food Sci. **2001**, *64*, 71–80.
22. Sutherland, I.W. Bacterial exopolysaccharides. Adv. Microb. Physiol. **1972**, *8*, 143–212.
23. Becker, A.; Katzen, F.; Pühler, A.; Ielpi, L. Xanthan gum biosynthesis and application: a biochemical–genetic perspective. Appl. Microbiol. Biotechnol. **1998**, *50*, 145–152.
24. Nakata, M.; Kawaguchi, T.; Kodama, Y.; Konno, A. Characterization of curdlan in aqueous sodium hydroxide. Polymer **1998**, *39*, 1475.
25. Adachi, Y.; Okazaki, M.; Ohno, N.; Yadomae, T. Enhancement of cytokine production by macrophages stimulated with (1→3)-beta-D-glucan, grifolan (GRN), isolated from Grifola frondosa. Biol. Pharm. Bull. **1994**, *17* (12), 1554–1560.
26. Sasaki, T.; Takasuka, N. Further study of the structure of lentinan, an antitumor polysaccharide from Lentinus edodes. Carbohydr. Res. **1976**, *47*, 99–104.
27. Rezzoug, S.A.; Maache-Rezzoug, Z.; Mazoyer, J.; Jeannin, M.; Allaf, K. Effect of instantaneous controlled pressure drop process on the hydration capacity of scleroglucan: optimisation of operating conditions by response surface methodology. Carbohydr. Polym. **2000**, *42*, 73–84.
28. McNiel, B.; Kristiansan, B. Temperature effects on polysaccharide formation by *Aureobasidium pullulans in stirred tank*. Enzyme Microb. Technol. **1990**, *12*, 521–526.
29. Lowe, J.B.; Marth, J.D. A genetic approach to mammalian glycan function. Annu. Rev. Biochem. **2003**, *72*, 643–691.
30. Lewis, S.; Crossman, M.; Flannelly, J.; Belcher, C.; Doherty, M.; Bayliss, M.T.; Mason, R.M. Chondroitin sulphation patterns in synovial fluid in osteoarthritis subsets. Ann. Rheum. Dis. **1999**, *58* (7), 441–445.
31. Murata, K.; Yokoyama, Y. Dermatan sulfate isomers in human articular cartilage characterized by high-performance liquid chromatography. Biochem. Int. **1987**, *15* (1), 87–94.
32. Fransson, L.A.; Carlstedt, I.; Coster, L.; Malmstrom, A. The functions of the heparan sulphate proteoglycans. Ciba Found. Symp. **1986**, *124*, 125–142.
33. Jalkanen, S.; Bargatze, R.F.; de los Toyos, J.; Butcher, E.C. Lymphocyte recognition of high endothelium: antibodies to distinct epitopes of an 85–95-kD glycoprotein antigen differentially inhibit lymphocyte binding to lymph node, mucosal, or synovial endothelial cells. J. Cell Biol. **1987**, *105* (2), 983–990.
34. Miyake, K.; Underhill, C.B.; Lesley, J.; Kincade, P.W. Hyaluronate can function as a cell adhesion molecule and CD44 participates in hyaluronate recognition. J. Exp. Med. **1990**, *172* (1), 69–75.
35. Fraser, J.R.; Laurent, T.C.; Laurent, U.B. Hyaluronan: its nature, distribution, functions and turnover. J. Intern. Med. **1997**, *242* (1), 27–33.
36. Toole, B.P. Hyaluronan in morphogenesis. J. Intern. Med. **1997**, *242* (1), 35–40.

Polyurethanes

Joanna D. Fromstein
Kimberly A. Woodhouse
Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, Ontario, Canada

INTRODUCTION

A polyurethane (PU) is a polymer containing recurring urethane linkages, formed through the reaction of a di- or polyisocyanate with a compound containing active hydrogen groups, usually a polyol.[1] The PU family of polymeric materials possesses a broad range of physical and chemical properties that are useful in varied applications such as insulating foam, stretchy synthetic leathers and textiles, paints, coatings, and adhesives.[2,3] This wide variety of applications is attributable to the vast array of chemicals that can be used as reactants in the PU synthesis. Polyurethanes play an important role in the plastics industry, representing roughly 6% of plastic products synthesized worldwide.[2] Polyurethanes have relatively low cost of manufacture and processing and are traditionally stable, resilient materials.[2] This, combined with the wide range of chemical and physical properties available, makes polyurethanes a good choice for many synthetic applications.

BACKGROUND

Polyurethanes represent approximately 6% of global plastic consumption, and 50% of the entire PU consumption occurs in North America and Europe.[2] Roughly 30% of PUs are used for furniture and mattresses, 15% for vehicles, and 13% for construction.[2] In all of these applications, PU foams are used. In fact, over 80% of polyurethane applications use foam materials.[2] Other forms of PU available include solid blocks, fibers, paints, adhesives, and coatings.

Polyurethane is a term used to represent the group of polymers that contain the urethane linkage, sometimes called an ''ester of carbamic acid'' (Fig. 1).[2] Polyurethanes are block copolymers of the form $(AB)_n$ that is formed by a polyisocyanate addition reaction. Some PUs are thermoplastics, while others are thermosets. The properties of the PU can be changed substantially by changing the chemical composition of the polymer's constituents, and thus the PU backbone itself. Polyurethanes can be made in a wide variety of grades, densities, and stiffnesses.[3]

Professor Otto Bayer discovered PUs in Germany in 1937,[2,4] during the development of a rival product for nylon (polyamide) fibers.[4] In this process, he discovered the diisocyanate addition polymerization reaction that is the basis of synthesizing PUs.[4] Bayer's first polymers were polyureas, prepared by reacting hexamethylene diisocyanate with hexamethylene diamine. However, the resulting product was infusible, and it did not form fibers or films very well. Shortly after, in 1938, a German patent was issued, covering linear PU preparation from glycols and diisocyanates.[4] Since then, significant advances have been made in the PU field, enabling the use of PUs for paints, adhesives, structural foams, rigid foams, flexible foams, and medical and thermoplastic elastomers.[2,3,5]

CHEMISTRY

Polyurethanes are formed when a diisocyanate (or polyisocyanate) is reacted with hydroxyl groups at a molar ratio of 2 or higher (isocyanate : hydroxyl).[3] When the polyol and polyisocyanate are combined in the presence of a suitable catalyst, the exothermic polymerization reaction begins spontaneously. This type of synthesis is an addition polymerization. Most polyols and polyisocyanates used for manufacturing PUs are liquid at standard room temperature.[3] Industrially, the PU synthesis reaction is rapid, and the product is a solid polymer. The reaction rate can be varied significantly by changing the catalyst type and concentration, facilitating the use of PUs in a variety of applications.[3]

Reactions of Isocyanates

Isocyanate (NCO) groups are energy-rich and highly reactive; they react exothermically with hydrogen-active compounds, epoxides, and other isocyanates.[2] Isocyanate groups will react most strongly with alcohol

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007921

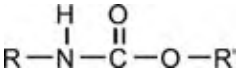


Fig. 1 The urethane linkage.

(hydroxyl) and amine groups. The reaction of an isocyanate with alcohol results in the formation of a urethane linkage with the rate depending on the chemical structure of both the isocyanate and the polyol (Fig. 2).[1] Aliphatic polyols are the most reactive, with primary hydroxyl groups having higher reactivity than secondary groups.[3] Phenol groups react significantly more slowly than aliphatic compounds. The resulting urethane groups are more susceptible to degradation upon heating.[3] This reversible reaction is used to advantage in the manufacture of ''blocked'' isocyanates.[3]

When isocyanates come into contact with water, gaseous carbon dioxide and a substituted urea are formed (Fig. 3). During foam production, this CO_2 is often used as the primary blowing agent.[2] When synthesizing PU products that are not intended to be foams, water contamination should be avoided at all costs.[2,3]

Isocyanates also react with primary and secondary amine compounds. Tertiary amines cannot react with isocyanates because they do not contain active hydrogen atoms, but they are powerful catalysts for many other isocyanate reactions.[3] Diamines are frequently used as chain extenders and curing agents in PU manufacture.[6] The addition of a diamine to the reaction mixture increases the overall reactivity during polymerization. The reaction between an isocyanate group and an amine results in the formation of a urea bond. The polyurea segments present in the finished PU serve to increase the potential for both covalent and hydrogen bond crosslinks within the polymer.

The reactivity of unhindered isocyanates with primary amines is considerably higher than that with primary alcohols. The reactivity of amines increases with the basicity of the amine, and aliphatic amines are much faster than aromatic amines.[3] Polar solvents can increase the reactivity of isocyanates through the stabilization and polarization of the isocyanate group.[3]

Isocyanate groups can react, under select conditions, with the active hydrogen atoms of urethane and urea linkages.[3] When an isocyanate reacts with a urea, it forms a biuret, whereas reaction with a urethane linkage results in an allophanate. These reactions are reversible, crosslinking reactions, and the kinetics are much faster for biuret formation.[3] Isocyanates can also react with compounds that do not contain active hydrogen, such as epoxides.[2] The products formed following the reaction of epoxides with isocyanates are used as embedding compounds and casting resins.[2]

Isocyanates can also react with other isocyanate molecules to form oligomers (Fig. 4).[3] This polymerization is more likely to occur in the presence of basic catalysts.[3] Isocyanate dimers, also called ''uretidinediones,'' can only be formed by aromatic isocyanates, and uretidinedione formation is inhibited by ortho substituents. Hence, only 4,4-diphenylmethanediisocyanate (MDI) dimerizes at room temperature, and its rate of formation is quite low.[3] At higher temperatures, MDI would form an insoluble polymeric material.[3] Trimers of isocyanates are also possible; these structures are called ''isocyanurates.''[3] Isocyanurates are formed by both aliphatic and aromatic isocyanates, and the resulting structure is highly stable to a temperature of approximately 270°C.[3] Isocyanurates give very stable branch points for crosslinking, unlike the reversible uretidinedione, biuret, and allophanate linkages.[3] Isocyanurate groups have been shown to decrease the flammability of rigid foams used for construction applications.[2,3] These isocyanurate groups are formed by polymerizing most of the isocyanate groups together to form isocyanurate ring structures.[3]

Polyurethane Synthesis

Polyurethanes can be synthesized using either a one-shot process or a two-stage, prepolymerization technique.[2] For the one-shot process, all chemicals are combined together at once and allowed to react. Although this method is faster and less labor-intensive, the final product tends to have a less organized structure than that achieved using the two-stage method.[2] In the

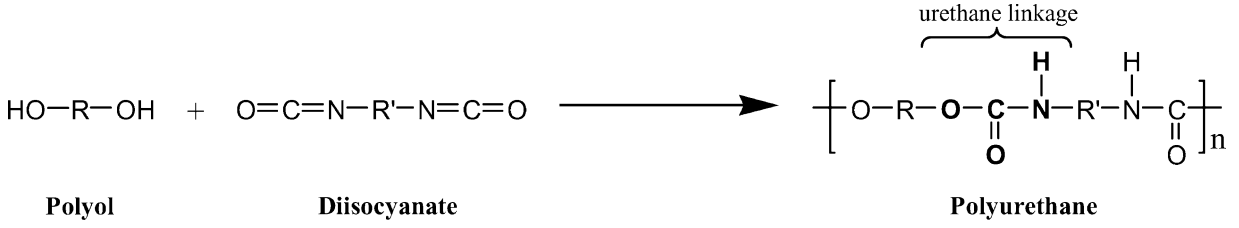


Fig. 2 The polyurethane addition reaction between a polyol and a diisocyanate.

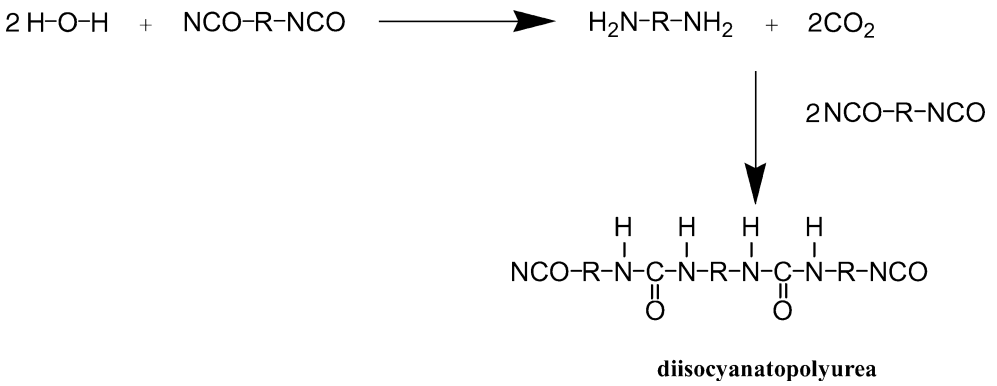


Fig. 3 The reaction between isocyanates and water, resulting in the formation of carbon dioxide gas and a substituted polyurea.

prepolymer method (Fig. 5), a diisocyanate is first allowed to react with a polyester or polyether diol at a 2 : 1 molar ratio. This results in a urethane-linked prepolymer that comprises the diol, flanked on either side by the isocyanate. In the second step of polymerization, the polymer is lengthened through the addition of a chain extender, which links up with remaining isocyanate groups. Diol chain extenders form PUs while diamine chain extenders result in the formation of PU ureas. The PUs synthesized using the two-step process are usually segmented PUs that possess a phase-segregated morphology consisting of long flexible polyester or polyether chains linked by rigid blocks.[2,3]

The isocyanate and chain extender form rigid blocks (hard segments), whereas the polyols give rise to flexible regions (soft segments).[2] By modifying the chemistry of the three reactants, segmented PUs with properties ranging from soft, jelly-like substances to hard rigid plastics can be made. Hard-segment blocks are in the order of 20 Å × 55 Å in size.[2,7]

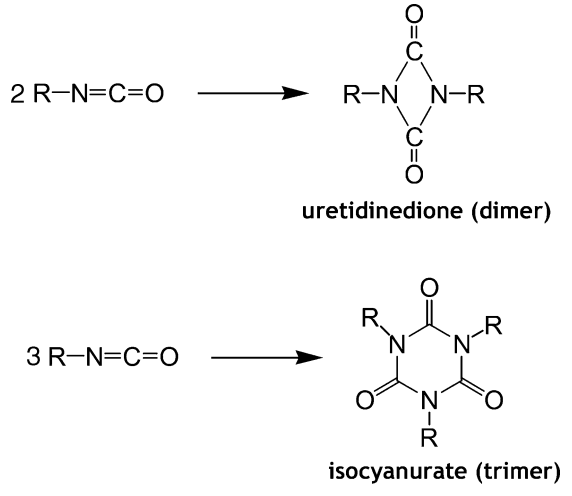


Fig. 4 Dimeric and trimeric structures formed by isocyanates when they react with other isocyanate groups.

The degree of hard-segment aggregation or separation is dependant upon the choice of polyether or polyester polyol, the weight ratio of the hard segment to that of the chosen polyol, the isocyanate, the processing parameters, and the synthesis reaction conditions.[3] Highly crosslinked rigid PUs are too tightly packed to allow for phase segregation to occur.

The Reactants

Polyisocyanates

Roughly 95% of all PUs are synthesized using either toluene diisocyanate (TDI) or MDI.[2,3] Both of these aromatic isocyanates result in the formation of strong materials with stiff polymeric chains. Of the two, MDI is more reactive and less volatile. Most TDI that are used commercially are composed of an 80 : 20 or 65 : 35 mixture of 2,4-TDI to 2,6-TDI. Both a monomer form and a polymerized form of MDI can be purchased.[3] Polymeric MDI can be reverted to monomeric diisocyanates by elevating the temperature.[3] Some specialty applications, particularly paints and coatings, use aliphatic polyisocyanates during synthesis.[2] These isocyanates are used to impart resistance to light, hydrolysis, and thermal degradation.

Polyols

There are two classes of polyols used in PU synthesis: polyesters and polyethers.[3] Approximately 90% of all PUs are manufactured using polyether polyols. Many polyols can be modified to have a higher degree of primary hydroxyl groups, or to contain one of polyurea or PU dispersions.[3] The primary polyols are more reactive and the dispersed polyurea and PU segments serve to increase the crosslinking of the finished PU.[3] Most of the polyether polyols

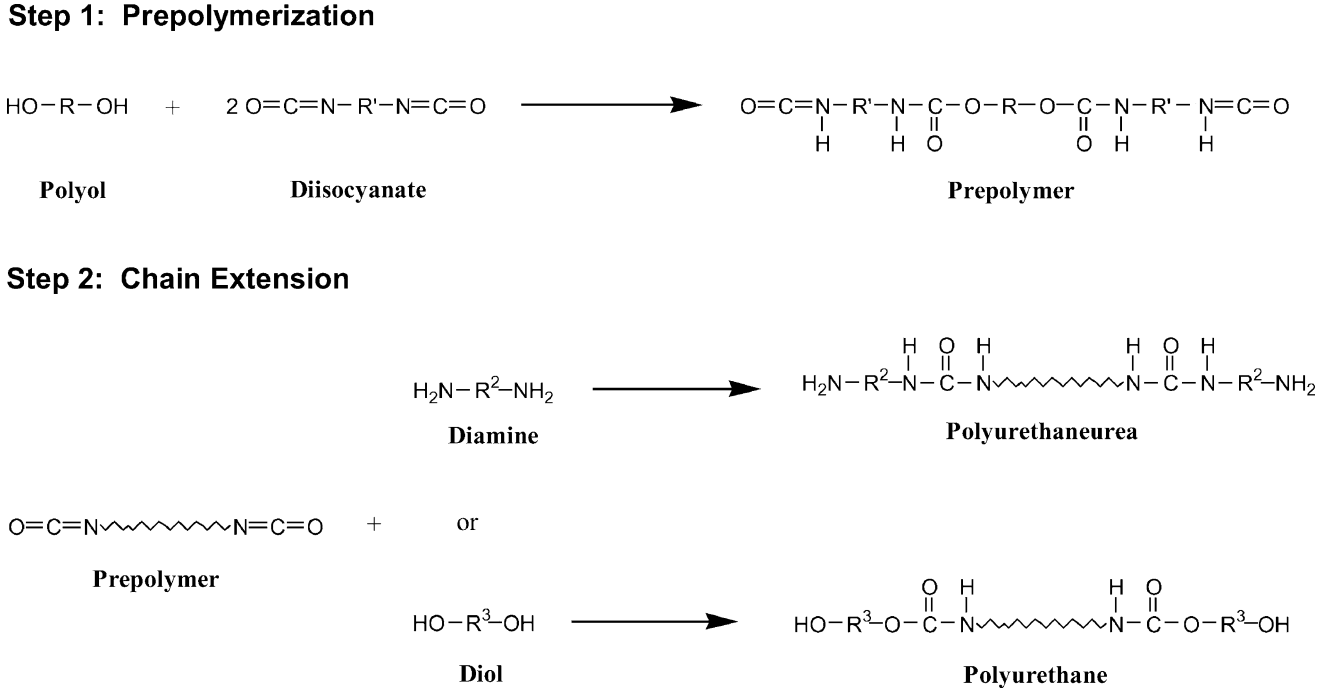


Fig. 5 The two-step PU synthesis method.

used in PU manufacture are made by the addition polymerization of alkylene oxides onto alcohol or amine initiators.[3]

Polyester polyols are more expensive to produce, more viscous, and form polymers that are more susceptible to hydrolysis. They do, however, impart high tensile strength and good resistance to flexing, abrasion, and photooxidation.[3] Hence, these polyester polyols, especially adipates, polycaprolactones, and polycarbonates, are commonly used for elastomers, coatings, and adhesives, where these properties are desirable.[3,4] Polyester polyols also form polymers that are resistant to dry cleaning solvents and high temperatures, making them attractive for use in textiles for the fashion industry.[4]

Chain Extenders and Crosslinkers

Chain extenders are low molecular weight polyols or polyamines that are used in flexible PUs to elongate the PU chain. They are difunctional glycols, diamines, or hydroxy amines.[3] Flexible polyurethanes such as flexible foams, elastomers, and reaction injection molding (RIM) processed PUs are made using chain extenders.[3] The chain extender's presence permits hard-segment segregation, resulting in an increase in both the modulus and the thermal stability.[3] Crosslinking agents are simply chain extenders with a functionality of 3 or more, providing extra crosslink sites. Some chain extenders, particularly diamines, will also behave as curing agents.[3]

Additives and Modifiers

A wide variety of catalysts have been used in PU production. Most catalysts are either tertiary amines or organo-metallic compounds.[3,8] The most commonly used catalysts in PU foam production are the tertiary amines, which catalyze the reaction between water and isocyanates.[3] The catalytic effect can be altered, depending on the chemical structure of the amine and on its basicity.[3] In addition, some tertiary amines catalyze chain branching through biuret linkages. This chain branching within the polyurea segments inhibits the formation of well-ordered secondary bonded domains, producing softer forms with higher resiliences. A similar effect can be achieved by using a catalyst that specifically encourages polymerization of the isocyanates, such as symmetrical triazine derivatives and alkali metal salts of carboxylic acids and phenols.[3] Organo-tin and other similar catalysts are used to encourage the reaction between hydroxyl groups and isocyanate groups. Usually, organo-metallic catalysts are used with aliphatic isocyanates, whereas tertiary amines are used in systems using aromatic isocyanates.[8] In the manufacture of some PU foams, a mixture of organo-metallic and tertiary amine catalysts is sometimes required to maintain a balance between the PU polymerization reaction and the reaction of water and isocyanate to form CO_2.[3,8]

Surfactants are often added to PU reactants to serve as emulsifiers; these polyether siloxanes improve the miscibilities of the reactants (polyisocyanate, polyol,

water, etc.) prior to synthesis.[2] Organo-silicon compounds are also used during foam production as ''foam stabilizers'' or ''cell regulators.''[2] These additives help to control the size and nature of the pores and stabilize the rising foam until it has cured.[2]

Other additives that may be used by PU manufacturers include curing agents, flame retardants, fillers, antiaging agents, colorants, antistatics, biocides, and release agents.[2]

MANUFACTURING AND PROCESSING

Commonly, polymerization coincides with the shaping of the PU product. In general, the polymerization process begins with the polyisocyanate and polyol being conveyed from their storage tanks to a working container. The ingredients are heated to the desired temperature and fed by metering units to the mixing head. Once mixed, the reactants are discharged into a mold or onto a substrate and allowed to react fully.[2]

There are two primary methods of PU production: continuous and discontinuous (batch). In the continuous process, the polyisocyanate and polyol are transferred from storage tanks into working containers, adjusted to the desired temperature, and then metered into the mixing head. From the mixing unit, the reaction mixture is ejected onto a substrate or into a mold, where it is allowed to react fully.

Cavity foaming and molding production are batch operations.[2] The minimum demold time depends on the reactivity of the foam system, the density of the foam, the temperature reached inside the block of foam, the size of the block, and the degree of overpack in the mold.[3]

Continuous manufacture is the most economical method of making large quantities of slabstock foam. The results also tend to be of better quality as it is easier to ensure uniformity of cell size and structure in this sort of operation. In this method, the foam reaction mixture is metered into a trough on a moving conveyer belt.[3] The sidewalls in this process must be rigid enough to withstand the pressure that occurs just after the foam has risen completely. Conveyer belts that move synchronously with the base belt are commonly used. Another conveyer that is height adjustable is used as a lid to ensure that a flat-topped foam block is formed.[2]

RIM

Reaction injection molding produces flexible foams of highly reactive PU systems using very short molding times (less than 1 sec in some cases).[2] This process uses high-pressure machines to distribute the individual reactants into the molds using closed circuit pipelines.[9] Each mold has a rigidly mounted self-cleaning mixing head attached to the clamping unit. Ground short glass fibers (140–240 μm) can be added to one of the components (commonly the polyol) to produce reinforced reaction injection molded PU.[2,9] Alternately, a glass mat can be placed in the mold and then surrounded by the reaction mix in a second step. This technique is called structural-RIM.[2] A more recent technique, long fiber injection, has been developed in which fibers that are 5–10 cm long are introduced directly into the mold with the reaction mix, all in one stage.[2]

Reaction injection molding differs from thermoplastic injection molding in numerous ways. First, the internal pressures in the molds and locking forces are one order of magnitude lower than those required for thermoplastic injection molding. In addition, the flow paths achieved with RIM are at least one order of magnitude higher.[2] Reaction injection molding is most commonly used for the production of parts with large area such as spoilers, fenders, and bumpers for motor vehicles.[9]

FOAMS

Approximately 80% of PU are manufactured as foams.[2] Polyurethane foams are lightweight materials, consisting predominantly (>98%) of air (and gases in the case of closed-cell foams).[2] Both the chemistry and the processing will have major effects on the final properties of the foam.[3] Polyurethane foams can be used to form composites with almost all flexible or rigid facings.[2] As the density of the foam increases, the suitability for demanding dynamic applications is improved.[3]

A number of higher functionality and viscosity MDIs are available for use in rigid foam manufacture.[3] Their higher viscosity improves the uniformity of the reaction mixture, resulting in a better pore structure.[2] The cells in these foams are less elongated, resulting in a final product that has a greater strength and higher dimensional stability.[2] These foams are also more fire resistant than those made using pure polymeric MDI.[3]

Glass-fiber incorporation increases the fire resistance of polyisocyanurate foam laminates. The glass fibers embedded in the foam prevent the development of deep fissures in the protective carbonaceous char that is formed when polyisocyanurate foam is exposed to high temperature flames.[3]

Flexible Foams

Flexible foams are used in mattresses, car seats, and ''comfy chairs.''[3] They are open-celled materials that

have a relatively low resistance to deformation under pressure.[2] Flexible foams can be classified as either polyether or polyester foams, depending on the chemical composition of the PU.[2] Approximately 95% of all foams manufactured are polyether foams.[2] Polyester-based PU foams are used for high load bearing applications because they have higher compressive and tensile strengths than their polyether counterparts.[2] Special polyester foams include highly elastic foams, also called ''high resilience'' (HR) or ''cold slabstock'' foams.[2] High resilience foams can be produced with flame retardants as well. These materials are called ''combustion-modified HR'' foams.[2] Other modified foams include supersoft, flame-laminated, and high-frequency welded foams.[2]

Flexible PU foams are synthesized by the reaction of a diisocyanate with a hydroxyl-terminated polyether or polyester and water, simultaneously.[1,3] When the water and the isocyanate come into contact, they react and produce heat and carbon dioxide. This CO_2 is trapped within the polymerizing material. As the gas bubbles expand, cells or pores are formed within the polymer, making a foam. This foaming process is also called chemical blowing, where the carbon dioxide created by the reaction behaves as the ''blowing agent.''[1] In addition to forming CO_2, reacting an isocyanate with water results in the formation of a substituted polyurea, which becomes the hard segment in the final PU foam.[3] The heat evolved by the reaction affects the rate of polymerization, helps the gases to expand, and aids in the curing of the finished product.[1,3] Chemical blowing also stiffens the PU by increasing the polyurea content of the polymer structure, and thus the secondary bonding between polymer chains.[3] The rates of polymerization and carbon dioxide formation are maintained in balance by selecting polyols and isocyanates of appropriate reactivities.[3] In addition, catalysts, foam stabilizers, and temperature control can also be used to help achieve the desired physical properties.[3] Some flexible foams even use additional ''physical'' blowing agents to help achieve the desired cellularity. Physical blowing agents are chemically inert liquids that have a low boiling point and will vaporize under the heat evolved by the reaction.[3] Unlike the chemical blowing agents, physical blowing agents tend to reduce the stiffness of the polymer by reducing the polymerization reaction temperature and the final curing temperature of the PU.[3]

More than half of the PUs produced worldwide are formed as low-density flexible foam.[3] In Europe, this material is made at a rate of 1 kg per person each year.[3] The bulk of this material is used in the furniture, bedding, and automotive industries.[3] These foams have high tensile strength and elongation to break values, and a HR.[3] The majority of flexible foam is fabricated as slabstock, with the remainder being molded or used as padding.[2]

There are two main types of flexible foam slabstock: polyester-based foams, used for technical and high elongation grade laminates or textiles; and polyether-based foams used for upholstery, HR, and flame retardance.[3] By varying the type of polyester or polyether, the length of the polyol chain, the structure and size of the hard segment, and the amount of blowing, the foam can be tailored to meet the required specifications.[2]

Over 90% of all flexible PU foam production is now based on polyether polyols and 80 : 20 TDI.[3] Some specialty foams have been developed recently using MDI, either alone or in combination with TDI. These new materials have been produced primarily for furniture and vehicle seating.[3]

Flexible foam slabstock is normally produced continuously.[1,2] To obtain reproducible materials, all ingredients must be conditioned to within 2°C of the manufacturer's requirements.[2] First, the reactants are metered from their working containers and conditioned to the appropriate temperature. Next, they are quickly mixed together in a mixing head. The mixer nucleates the reaction mixture, providing growth points for the bubbles.[2] This control of bubble formation is necessary to control the size, shape, and distribution of the cells within the foam. Nucleation is achieved by placing a fine dispersion of gas (usually air or nitrogen) in the reaction mixture.[2] The nucleation can be controlled by the agitator speed and by the pressure in the mixer. The mixed, nucleated reaction mixture is usually deposited in the mold or continuous trough within 1 sec after mixing begins. Next comes an induction, or ''cream'' time, that enables the bubbles to grow in size.[2] The reaction mixture is then extruded from the mixing head onto a rolling conveyor belt (up to 10 m/min), where the PU will react and form the final slabstock product. Outputs of 50–600 L/min of the reaction mix can be obtained; these materials have densities ranging from 15 to 60 kg/m^3.[2]

Once the slabstock has been made, the conveyor belts carry the foam to storage for 12–24 hr, to allow the materials to cool and cure.[2] Finally, the foams are transported via conveyor belt past cutters and trimmers that cut the foam to size. High speed water jet are used to cut a variety of shapes from the flexible foam.[2]

Flexible foams can be modified after production using post-treatments such as reticulation and impregnation.[3] Impregnation of low-density flexible foam can include a coloration of the foam by coating the material with a binder or surface coating that contains pigment.[3] The most common impregnation treatments are flame retardants such as aluminum hydroxide, which are applied by impregnating the foam with up

to 300 wt.% aluminum hydroxide in a synthetic latex binder. Other frequently impregnated coatings include antifungal, water-resistant, and ion-exchange resin coatings.[3]

Reticulation is an after-treatment that removes residual cell membranes to produce a foam with a skeletal rib structure. Reticulated foams are very effective filters for the removal of dust and fibers from air and other gases. They allow high flow rates and low pressure gradients (i.e., minimal energy consumption). Methods of reticulation include chemical hydrolysis and the use of an explosion flame front to melt the membranes.[2]

All low-density foams tend to be anisotropic. Although the nucleation bubbles start spherical, they have a tendency to elongate in the direction in which they rise. The fewer and larger the bubbles, the greater the elongation of the pores. Surface active agents can be used to prevent cells from rupturing prematurely under pressure.

Semirigid Foams

Semirigid foams are often used as padding.[2] They have a significantly higher density and compression strength and possess excellent mechanical damping capabilities. As a result, semirigid foams are frequently used as protective cushions in cars. A special, high-performance semirigid foam is known as energy absorbing foam, and it is used not only in the seats but also in car bumpers.[2] The manufacture of semirigid foams is very similar to that of flexible foams, but different reactant chemistries will be selected to create materials with more crosslinks.

Rigid Foams

Rigid PU foam is a highly effective thermal insulator, used frequently in refrigerators and as building insulation.[3] While flexible foams are slightly crosslinked materials, rigid foams are highly crosslinked.[2] Rigid foams have a relatively high resistance to deformation under pressure and particularly low heat conduction owing to the insulating gases trapped in the closed cells.[2] These materials can be used at temperatures ranging from -200 to $+150^{\circ}C$ and can even withstand short-term exposures of up to $250^{\circ}C$.[2] These properties enable special PU rigid foams to be thermoformable, despite their crosslinked nature.[2] Polyurethane rigid foams are rot proof and odor proof, and resistant to plasticizers, fuels, oils, and dilute acids and alkalis.[2] Moldings made of rigid foams are light and robust while remaining convenient and attractive. Rigid foams can also be used to fill cavities and form pipes and panels.[2]

Rigid foam slabstock is manufactured using both continuous and discontinuous processes.[3] The discontinuous method is restricted to systems in which a long cream time is needed at ordinary temperatures. In this process, a slow reacting foam mixture is stirred. The mixture is then poured into a mold, and a floating lid is placed over top of the material. The reaction mixture expands as the slabstock forms until the lid's preset height is reached. As a safety measure, if too much reaction mixture is added to the mold, the floating lid assembly has enough give to allow for overflow, thereby preventing damaging the mold with excessive pressure.[3]

Disadvantages of using batch (or discontinuous) manufacture include limited choice of chemical systems, high cost of labor, increased hazards because of frequent chemical handling, and loss of reaction mixture to the mixer and mixing bowl.[3] Many of these weaknesses can be reduced significantly by the use of machines for dispensing and mixing.[3] The cream time will be reduced compared to batch production, because of the energy input by the mixer into the system.[3] To manufacture a large, uniform block of foam using a highly reactive system, a high capacity/throughput dispensing machine is needed.[3]

Structural Foams

A special group of PU foams is structural foams that are made directly in their molds as either flexible or rigid foams.[2] Structural foams are PU moldings that have a porous core and a nonporous outer layer of the same material and have been made in a single pass.[2] These sandwich-style panels are used in construction applications.[2]

PU Elastomers

Polyurethane elastomers are used in shoe soles, sports equipment, bumpers, underwear, and other stretchy clothing.[3] A urethane elastomer is a linear block copolymer synthesized from a di- or polyisocyanate, a polyol, and (frequently) a chain extender. Traditionally, PU elastomers have been synthesized using the prepolymerization process. Solid PU elastomers have high elasticity and high resistance to wear under different stresses.[2] The high modulus of these materials is attributable to the phase segregation within the polymer. The hard segments serve to reinforce the strength of the material by forming hydrogen bond crosslinks with the surrounding soft segment matrix.[3,6] Polyurethane elastomers are resistant to light, oxygen, and UV radiation. Sometimes, additives are incorporated to protect the PU elastomer from hydrolysis.[2] Polyurethane elastomers can be made into porous foams as

well. These cellular elastomers are high-performance, elastically deformable flexible molded foams.[2]

Polyurethane elastomers are available as thermoplastic elastomers suitable for conventional thermoplastic processing, cast elastomers, elastomeric fibers, and coatings on textiles and other flexible substrates. Almost all of PU elastomers are based on the standard $(AB)_n$ block copolymer structure with alternating hard and soft segments.[6,7] The soft segments are composed of the polyester or polyether groups, while the hard segments are the reaction products of a polyfunctional (usually difunctional) isocyanate and a low molecular weight chain extender, either a diol or a diamine. Polyurethane elastomers can possess hardnesses ranging from 10° Shore A to over 60° Shore D.[3] As a group, fully cured PU elastomers are tough, resilient, abrasion-resistant, of high mechanical strength, and resistant to many solvents and chemicals.[3] However, PU elastomers are not particularly stable in the presence of strong acids or bases, oxidizing agents, and some highly polar solvents.[3]

Originally PU elastomers were synthesized using one of the aromatic isocyanates (MDI, TDI, etc.).[3] Recently, however, more focus has been placed on the use of distilled MDI and modified MDI products. HMDI and IPDI are used to make coatings for applications when high resistance to weathering and discoloration by light is important.[3]

The raw materials used for PU elastomer synthesis are often solids that first have to be melted, desiccated, and then degassed in the working container prior to use. Polyurethane elastomer synthesis usually involves the use of a chain extender as well. Diamine chain extenders increase the mechanical integrity of the PU by increasing the degree of phase segregation and physical physical crosslinking within the material.[3]

Polyurethane elastomers are often sold by the chemical companies as granules or pellets that can be used by manufacturers to create products.[3] These pellets and granules are processed using traditional methods like extrusion, injection molding, etc.

Polyurethane fibers (elastanes) can be made via either of the following two ways. In the first technique, the PU is first synthesized using the prepolymer method, and then the PU is spun from solution (approximately 30% PU) to form a filament or yarn.[2] The yarn can either be collected as one long strand or be cut as it is collected. The second method, called ''reactive spinning,'' involves synthesizing the polyurethane and forming the filament all at the same time using a reactive bath to spin the fibers.[2]

Thermoplastic PU elastomers (TPU) are a special group of PU elastomers that can be processed using the traditional economical techniques available for thermoplastic polymers.[2] These materials are designed to meet a wide variety of industrial needs, including injection molding, blow-molded parts, films, coverings, tubing, etc.[2] Polyether-based TPUs are usually composed of polytetramethyleneoxide diols and polytetrahydrofurans.[3] Thermoplastic PU elastomers synthesized using MDI appear to have a higher degree of phase separation than those made with TDI. Thermoplastic PU elastomers are frequently chain extended with a low molecular weight diol such as 1,4-butanediol. In cases where resistance to hydrolysis is required, 6-hydroxycaproic acid polyesters are used as the polyol.

Thermoplastic PU elastomers are produced both continuously and discontinuously using the one-shot process.[2] Once the components have been mixed in a reactor, they are dispensed onto a conveyor belt.[2] The TPU reacts fully and solidifies in the form of sheets which are chopped in a granulator and then processed into uniform pellets in an extruder. If the sheets are processed immediately, no drying time is needed before granulation. Otherwise, a drying step at approximately 100°C is incorporated for anywhere from 30 min to 2 hr. The production of TPU granules can also be carried out entirely in a twin-screw extruder (compression ratio 1 : 2) or in a single-screw extruder (compression ratio of 1 : 3).[2] Temperatures in the order of 180–250°C are commonly used for this method, with mold temperatures kept steady at 20–40°C using a water jacket. Finished products attain their optimal properties following a 15–20 hr curing step at 80–120°C, or 4–6 wk after storage if not treated following fabrication.[2]

Medical PUs are another subset of PU elastomers. Segmented PUs were first suggested for use in a biomedical application in 1967.[10] Early work with PU elastomers showed that these materials could be used for implants without causing a large, unwanted inflammatory response. The first medical devices made of PUs, however, were found to be susceptible to hydrolysis and degraded faster than desired.[6,10] From that time, new biostable materials have been developed for use as pacemaker leads, catheters, vascular grafts, stents, etc.[6,10–12] In addition to these stable materials, many researchers have developed intentionally biodegradable PU elastomers for use in tissue engineering.[6,13,14] These degradable materials are formed into highly interconnected three-dimensional scaffolds using new processing techniques, such as electrospinning or laser boring, and then used as the base to make replacement soft tissues.[12,14]

Specialty PUs

Polyurethane electrical embedding compounds exist in both flexible and rigid variations. These specialty PUs

have a long-term thermal stability. Another unique PU group is the PU sealants, which compete with commercially available silicone and polysulfide sealants.

Polyurethane rubbers are a special type of PU elastomer. The PU rubbers are highly resistant to oil and fuel and have high abrasion resistance and tensile strength similar to PU elastomers. Polyurethane rubbers maintain their flexibility even at a temperature as low as -40°C and are thermally stable up to 125°C. Polyurethane rubbers have better mechanical strength and resist abrasion, ozone, and oil better than other specialty rubbers such as acrylates and nitrile butadiene.[2]

Polyurethanes are also used to make paints, available as water-thinnable paints and powder coatings. Items painted with PU paints benefit not only from a lacquer film, but also from an improved chemical resistance.[2] Polyurethane paints are scratch, abrasion, and impact resistant, and they adhere well to a variety of surfaces.[2] Polyurethane adhesives are similarly good choices as they resist heat, water, solvents, grease, and oils.[2]

Polyurethanes have also been invaluable as coatings for textiles, paper, and leather. The ability of these lacquers to adhere to a wide variety of substrates, combined with the flexibility across a wide range of temperatures and high tensile strength and tear propagation resistance, makes PU coated materials attractive substitutes for PVC, nitrocellulose, and acrylates.[2]

Polyurethane fibers are another niche application. These ''elastanes,'' the basis of Lycra™, have nearly taken over the textile industry, displacing rubber threads (elastodienes) in the process.[2,15] The high popularity of PU fibers is attributable to the good tensile strength and elasticity of highly segmented polyurethanes. In addition, elastanes can be processed in a variety of sizes, either as continuous filaments (yarns) or as shorter fibers. Rubber threads, on the other hand, are available solely as monofilaments.[2]

Polyurethanes can be processed into microcapsules and gels. Polyurethane gels swell in water, absorbing up to 90% water, which then becomes trapped within the gel and cannot be squeezed back out.[2] Polyurethane microcapsules are fabricated by producing a very fine oil in water emulsion (droplet size approximately 3–10 μm) and by applying high shearing forces using ultrasonic dispersers.[2]

CONCLUSIONS

The development of PUs has had a monumental impact on our society. It is the ability to tailor-make a PU for each given application that accounts for their widespread use.[5] Since the discovery of PUs in the late 1930s, their flexible chemistry and diverse physical properties have led to the use of PUs in an astounding variety of applications. Future directions include degradable PUs for use in tissue engineering, designer polymers based on PU chemistry that combine proteins with synthetic polymers, and light weight composites for various applications.

REFERENCES

1. Woods, G. *Flexible Polyurethane Foams*; Applied Science Publishers Ltd: Essex, 1982.
2. Uhlig, K. *Discovering Polyurethanes*; Carl Hanser Verlag: Munich, 1999.
3. Woods, G. *The ICI Polyurethanes Book*, 2nd Ed.; ICI Polyurethanes and John Wiley & Sons: Chichester, 1990.
4. Frisch, K.C. Historical developments of polyurethanes. In *60 Years of Polyurethanes*; Kresta, J.E., Eldred, E.W., Eds.; Technomic Publishing Company, Inc.: Lancaster, PA, 1999; 1–21.
5. Kresta, J.E.; Eldred, E.W., Eds. *60 Years of Polyurethanes*; Technomic: Lancaster, PA, 1998; 400.
6. Lamba, N.M.K.; Woodhouse, K.A.; Cooper, S.L. *Polyurethanes in Biomedical Applications*; CRC Press LLC: Boca Raton, 1998.
7. Hepburn, C. *Polyurethane Elastomers*, 2nd Ed.; Elsevier Science Publishers Ltd.: Essex, U.K., 1992.
8. Silva, A.L.; Bordado, J.C. Recent developments in polyurethane catalysis: catalytic mechanisms review. Catal. Rev. **2004**, *46* (1), 31–51.
9. McCrum, N.G.; Buckley, C.P.; Bucknall, C.B. *Principles of Polymer Engineering*, 2nd Ed.; Oxford University Press, Inc.: New York, 1997.
10. Gunatillake, P.A., et al. Designing biostable polyurethane elastomers for biomedical implants. Aust. J. Chem. **2003**, *56*, 545–557.
11. Jung, B.O., et al. Preparation of polyurethane membrane dressing with various water vapor transmissions. J. Ind. Eng. Chem. **2002**, *8* (1), 57–63.
12. Zdrahala, R.J.; Zdrahala, I.J. Biomedical applications of polyurethanes: a review of past promises, present realities, and a vibrant future. J. Biomater. Appl. **1999**, *14* (1), 67–90.
13. Skarja, G.A.; Woodhouse, K.A. Synthesis and characterization of degradable polyurethane elastomers containing and amino acid-based chain extender. J. Biomater. Sci. Polym. Ed. **1998**, *9* (3), 271–295.
14. Fromstein, J.D.; Woodhouse, K.A. Elastomeric biodegradable polyurethane blends for soft tissue applications. J. Biomater. Sci. Polym. Ed., **2002**, *13* (4), 391–406.
15. Reisch, M. What's that stuff? Chem. Eng. News **1999**, *77* (7), 70.

Polyvinylidene Fluoride

Jeffrey H. Yen
Ramin Amin-Sanayei
Arkema Inc., King of Prussia, Pennsylvania, U.S.A.

INTRODUCTION

Polyvinylidene fluoride (PVDF) is produced by free radical polymerization of vinylidene fluoride ($CH_2{=}CF_2$). The demand for PVDF has been growing remarkably because of its outstanding properties such as:

- Mechanical strength and toughness.
- High abrasion resistance, enabling it to be used with slurries.
- High thermal stability under operating and processing application; PVDF does not degrade (discolor) when subjected to heat.
- High dielectric strength.
- High purity.
- Readily melt processable.
- Resistant to most chemicals and solvents.
- Resistant to ultraviolet and nuclear radiation.
- Resistant to weathering, because of its transparency and inertness to ultraviolet (UV) rays.
- Resistant to fungi.
- Low permeability to most gases and liquids.
- Low flame and smoke characteristics.

PVDF is a semicrystalline polymer: the crystalline fraction results in very low permeability to gases and fluids as well as low swelling in certain solvents, whilst retaining a high impact resistance, a good stress cracking resistance, and a very high dimensional stability. It offers a relatively wide application temperature, ranging from $-40^{\circ}C$ up to $+150^{\circ}C$. It is nonflammable and nontoxic; so it can be used in food industry.

PVDF is the third most widely used fluoropolymer, after polytetrafluoroethylene (PTFE) and fluorinated ethylene-propylene (FEP). The worldwide consumption of PVDF was approximately 15,000 metric tons in 2001 and is growing at an annual rate of ~6–8%. PVDF applications have been expanded over the past 40 years because of its unique physical properties, and have over 30 years of proven and field performance data on thermal, chemical, radiation, and weathering applications. PVDF applications include, but are not limited to, chemical processing of pipes and components, semiconductor, architectural finishes and coatings, electrical plenum, cable jacketing, offshore/petrochemical flexible piping, binders for lithium ion batteries, fuel cell membrane, polymer processing additives, decorative films and sheets, etc.

VINYLIDENE FLUORIDE: MONOMER

Chemical and Physical Properties

Vinylidene fluoride (VDF), also known as refrigerant 1132a, 1,1-difluoroethylene, 1,1-difluoroethene, vinylidene difluoride, etc., is a colorless, nearly odorless gas. Unlike the other fluoromonomers, VDF has long shelf life stability and does not self-initiate unwanted polymerization. Therefore, VDF can be stored, and shipped pure or with polymerization inhibitors. Its properties are summarized in Table 1.[1]

Manufacturing of Vinylidene Fluoride

Arkema (formerly Pennwalt or Atofina), Dow, DuPont, ICI, Kureha, etc. have developed a number of processes based on C_1 and C_2 feedstocks.

VDF from C_1 routes

Yields and product selectivity for the C_1 routes are inferior to C_2 routes, thus, are less attractive.

$$CH_3OH + CHClF_2 \rightarrow CH_2CF_2 + HCl + H_2O$$

$$CH_4 + CCl_2F_2 \rightarrow CH_2CF_2 + 2HCl$$

$$CH_4 + CHF_3 \rightarrow CH_2CF_2 + HF + H_2$$

$$CH_4 + CCLF_3 \rightarrow CH_2CF_2 + HCl + HF$$

$$CH_3Cl + CHClF_2 \rightarrow CH_2CF_2 + 2HCl$$

VDF from C_2 routes

Most commercial VDF processes are based on C_2 routes because of their superior yield to C_1 routes. Among the C_2 routes, VDF will be relatively purer

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007923

Table 1 Physical properties of vinylidene fluoride

Molecular weight	64.04
Melting point,°C	−144
Boiling point,°C	−84
Critical temperature,°C	30.1
Critical pressure, kPa	4434
Vapor pressure, Pa	133.32 at 21°C
Water solubility	0.018% at 25°C
Critical density, kg/m^3	417
Heat of formation, kJ/mol	−335
Specific heat, C_p(g) at 25°C, J/mol/°K	60
Viscosity, cP	0.01245 at 25°C

when the feedstock is methylchloroform:

$$CH_3CCl_3 + 2HF \rightarrow CH_3CClF_2 + 2HCl$$

$$CH_3CClF_2 \rightarrow CH_2CF_2 + HCl$$

Other routes commercially practiced to produce VDF are :

$$CH{=}CH + 2HF \rightarrow CH_3CHF_2$$

$$CH_3CHF_2 + Cl_2 \rightarrow CH_3CClF_2 + HCl$$

$$CH_3CClF_2 \rightarrow CH_2CF_2 + HCl$$

or

$$CH_2CCl_2 + 2HF \rightarrow CH_3CClF_2 + HCl$$

$$CH_3CClF_2 \rightarrow CH_2CF_2 + HCl$$

or

$$CH_3CF_3 \rightarrow CH_2CF_2 + HF$$

or

$$CClF_2CCl_3 + 3H_2 \rightarrow CH_2CF_2 + 4HCl$$

In the reaction path from chloro-difluoro-ethane (CDFE) to VDF, a small amount of chlorine works as a catalyst. The mixture is preheated and then fed into a metal nickel tubular reactor, which is housed in a high temperature fired furnace at 550°C–600°C. The reactor effluent is then cooled and enters HCl adsorption tower for the removal of HCl. The gaseous effluent from the HCl adsorption tower is scrubbed with dilute caustic soda and is dried in a packed molecular sieve column. It then enters a distillation column for further purification and recycling of CDFE back to the pyrolysis reactor.

Health, Environment, and Safety

VDF is a flammable gas. When inhaled, potential health effects include nausea, difficulty in breathing, vomiting, dizziness, tingling sensation, suffocation, convulsions, coma, etc. Short-term skin contact may cause blisters and frostbite. The effects on short-term eye contact include frostbite and blurred vision. Other health, environmental, and safety related ratings are given in Table 2.[1]

POLYVINYLIDENE FLUORIDE

Chemical and Physical Properties

PVDF is a linear partially fluorinate polymer containing 59.4 wt% fluorine and 3 wt% hydrogen. The high level of intrinsic crystallinity, typically near 60%, provides stiffness, toughness, and creep resistant properties. Incorporation of various fluorinated comonomers at low levels, typically about 5–20 wt%, enhances flexibility and clarity of PVDF by reducing the crystallinity, which in turn reduces the end use temperature rating. PVDF is commercially produced via free radical polymerization either with emulsion or suspension processes.

The spatial arrangement of the alternating CH_2 and CF_2 groups along the polymer backbone creates a high dipole moment that accounts for the unique polarity, unusually high dielectric constant, complex polymorphism, and high piezoelectric and pyroelectric activity of the polymer.[2] Because of its polar nature, PVDF is permeation resistant to chlorine and bromine. Both chlorine and bromine are known to permeate most commonly available commercial polymers.

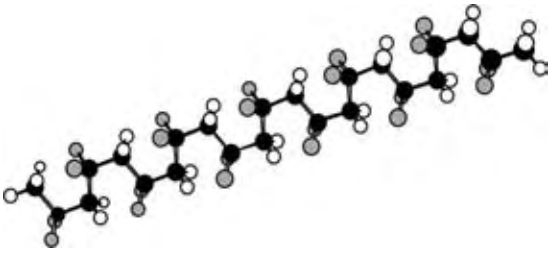
Typically, PVDF has four chain conformations:[3]

1. Alpha: The alpha form is the most common form of PVDF and is the most thermodynamically

Table 2 Environmental and safety related ratings

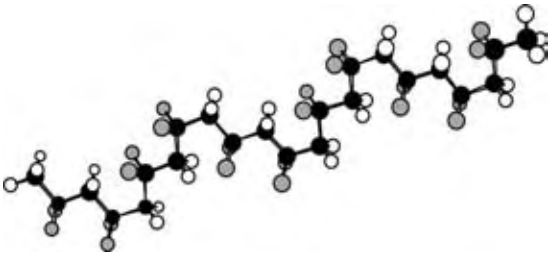
Lower flammable limit	5.5%
Upper flammable limit	21.3%
Autoignition temperature	640°C
ACGIH TWA	500 ppm
NIOSH recommended TWA 10 hr	1 ppm
NIOSH recommended ceiling	5 ppm
Octanol/water partition coefficient as log Pow	1.24

stable. It prevails on coating and normal melt processing of structural parts.



2. Beta: The beta form develops under mechanical deformation of melt-processed materials, typically at temperatures close to its melting transition. This structure provides some unique properties for PVDF: piezo- and pyro-electric activity.

3. Gamma: The gamma form arises infrequently.



4. Delta: The delta form develops through a distortion of one of the other forms under high electric fields.

(*Note*: In the above structures: grey = fluorine; white = hydrogen; and black = carbon.)

Resistance of PVDF to harsh acids, heat, ultraviolet rays, weathering, and oxidizing or high energy ionizing radiation environments, is the same as perfluoropolymers while its chemical resistance is not as broad. This is because the somewhat acidic hydrogen atoms along the chain are vulnerable to strong base attack.[4] This chemical susceptibility has, in fact, been turned into an advantage in preparing PVDF samples for adherence to a variety of substrates as well as introducing new chemical grafts to backbone. PVDF exhibits an unusual compatibility with other polymers having strong polar groups or carbonyl groups such as acrylics, and is particularly, miscible with short esters of methacrylic acid. This miscibility aspect led to the development of many alloys and mixed systems for coatings, membranes, and extruded parts.

Unlike conventional PTFE, PVDF can be molded through compression and injection processing and extruded into various shapes and configurations with conventional extruders. As a matter of fact, PVDF has been routinely processed on equipment that are also used for PVC and polyolefins.

PVDF Production

Polymerization chemistry

Free radical polymerization is the most widely used process in PVDF synthesis. It involves the reaction of VDF and other comonomers with active center followed by successive addition of monomer(s) under the condition in which monomers cannot react with each other without intervention of the active center. Active centers are generated by thermal decomposition of initiator and in some cases by photoinitiation of the catalyst. The average lifetime of each active center (free radical) is approximately few seconds depending on the degree of polymerization and the initiator concentration. For successful polymerization, the sequence of reaction must take place.

a. *Initiation Step*. VDF polymerization can be initiated with many different free radical initiators such as peroxides, peroxycarbonates, diacyl-peroxides, peroxyesters, and so on.

$$\mathrm{I} \rightarrow 2R^{\bullet}$$

b. *Propagation Step.* Successive monomer additions after the initiation step account for VDF consumption during polymerization and can be represented by

$$R_n^{\bullet} + \mathrm{VDF} \rightarrow R_{n+1}^{\bullet}$$

where n is the number of VDF units in the growing chain. The growing chain can undergo transfer reaction by abstracting hydrogen or halogen atom from a chain transfer agent(s) and that can be demonstrated by

$$R_n^{\bullet} + R' - X \rightarrow R_n - X + R'^{\bullet}$$

where $R_n - X$ is a dead chain and $R'^{\bullet}$ is a new radical that may or may not initiate new growing chain.

c. *Termination Step.* Termination can take place by the mutual annihilation of radical center to produce dead polymers.

$$R_n^{\bullet} + R_m^{\bullet} \rightarrow \text{dead polymer}$$

There are two possible points of addition on VDF for a propagating chain, either on carbon 1 (CF_2) or carbon 2 (CH_2). If each successive addition occurs in the same manner, the final polymer will have alternate carbon atoms. This type of addition will lead to head-to-tail arrangement. An inversion of in monomer addition will cause the formation of head-to-head and tail-to-tail addition. PVDF contains some head-to-head linkages because the F atoms are relatively small and their contribution to resonance stabilization of the growing polymer chain is not great. The head-to-tail placement, however, is the predominant successive propagation that favors both on steric and resonance grounds. Commercial grades exhibit a wide range of 3–7 mol% defect structures as determined usually by ^{19}F-NMR. The level increases with increasing polymerization temperature. The defect linkages affect significantly the crystallization processes and ultimate morphology.

Manufacturing process

The interesting and useful mechanical properties that are associated with PVDF depend on its molecular weight. The degree of polymerization of most commercial PVDF ranges from 1000 to 20,000 VDF repeating units. The heat of polymerization associated with propagation step is about 33 kcal/mol, which is much higher than that of a common vinyl monomer. This is because VDF is in a gaseous state and the resultant PVDF is solid. The common commercial polymerization reaction systems are emulsion and suspension polymerizations where the thermal and the viscosity problems can be managed. Use of nonaqueous media, such as CO_2, is also an active research area. Polymerization procedures, including temperature, pressure, recipe ingredients, monomer feeding strategy, and postpolymerization processing, are variables that influence product characteristics and quality.

Emulsion Polymerization. The temperature of the polymerization can vary depending on the characteristics of the initiator used, but it is typically between 30°C and 125°C. The pressure of the polymerization is typically between 1380 and 8275 kPa, but it can be higher if the equipment permits operation at higher pressure.

Surfactants used in the polymerization are water-soluble, halogenated surfactants. They are in particular fluorinated surfactant such as ammonium, substituted ammonium, quarternary ammonium, or alkali metal salts of perfluorinated or partially fluorinated alkyl carboxylates, monoalkyl phosphate esters, alkyl ether or polyether carboxylates, alkyl sulfonates, and alkyl sulfates.

A paraffin antifoulant may be employed, if desired, although it is not preferred, and any long-chain, saturated, hydrocarbon wax or oil may be used.

After the reactor has been charged with water, surfactant, and any optional paraffin antifoulant, it is brought to temperature and then pressurized with VDF and optional comonomer(s). The reaction pressure is maintained during polymerization course by adding monomer(s) to the reactor.

Chain-transfer agents range from alcohols, carbonate esters, ketones, carboxylate esters, to ethers compounds. Other classes of compounds, which serve as chain-transfer agents in the polymerization of VDF, are halocarbons and hydrohalocarbons such as chlorocarbons, hydrochlorocarbons, chlorofluorocarbons, and hydrochlorofluorocarbons.

The reaction can be started and maintained by the addition of any suitable initiator known for the polymerization of fluorinated monomers, including inorganic peroxides, ''redox'' combinations of oxidizing and reducing agents, and organic peroxides. Examples of typical inorganic peroxides are the ammonium or alkali metal salts of persulfates, which have useful activity in the 65°C–105°C temperature range. ''Redox'' systems can operate at even lower temperatures. Among the organic peroxides that can be used for the polymerization are the classes of dialkyl peroxides (di-*t*-butyl peroxide), peroxyesters (*t*-butyl peroxypivalate, *t*-butyl peroxyacetate), and peroxydicarbonates (di-*n*-propyl-peroxydicarbonate, diisopropyl-peroxydicarbonate).

Suspension Polymerization. Suspension polymerization of VDF can be carried out using an oil-soluble radical initiator such as peroxides, peroxycarbonates, diacyl-peroxides, or peroxyesters. Various dispersant agents such as cellulose derivatives or vinyl alcohol are used to prevent agglomeration during polymerization course. Coagulation of the polymer particle may occur at high comonomer loading when the polymer Tg is lower than the polymerization temperature. Suspension polymerization can be carried out using an oil-soluble initiator such as preoxy-dicarbonates or proxy. A suspension polymerization can also be conducted using a radical photoinitiator with UV–visible light. The polymerization temperature can be lowered because of photogeneration of radicals. Lowering the polymerization temperature reduces the number of defect structures in PVDF.

Other polymerization systems

VDF polymerization technology based on supercritical or liquid carbon dioxide as polymerization media has been reported.[5–8] This technology offers an advantage in the polymer isolation step where a clean dry polymer is produced simply by depressurization. The residual monomer(s) and CO_2 can be recycled back to the reactor. The PVDF having a multimodal molecular weight distribution is produced in CO_2. Adequate CO_2 density

Table 3 Typical design properties of PVDF homopolymers

Chemical and physical properties	
Molecular weight of repeat unit, g/mole	64.03
Melting temperature, °C	Typically 135°C–190°C
Amorphous density at 25°C, g/cc	1.74
Crystalline density at 25°C, g/cc	2.00
Water adsorption at equilibrium	0.015–0.05%
Mold shrinkage, %	2–3
Melt flow, g/10 min at 230°C and 5Kg	0.5–100
Mechanical properties	
Hardness, Rockwell M	75–120
Hardness, Shore D	75–78
Tensile strength, ultimate, MPa	36–55
Tensile strength, yield, MPa	20–70
Elongation at break, %	6–250
Elongation at yield, %	2–16
Tensile modulus, GPa	0.48–7
Flexural modulus, GPa	1.4–2
Flexural yield strength, MPa	14–65
Compressive yield strength, MPa	17–80
Izod impact, notched, J/cm	1–2
Izod impact, unnotched, J/cm	>15
Charpy impact, unnotched, J/cm^2	17.7–28
Charpy impact, notched low temp, J/cm^2	0.5–1.1
Charpy impact, unnotched low temp, J/cm^2	18.6–26
Charpy impact, notched, J/cm^2	0.8–11.8
Tensile impact strength, J/cm^2	140–210
Coefficient of friction	0.14–0.18
Tensile creep modulus, MPa at 1 hr	508–1600
Tensile creep modulus, MPa at 1000 hr	210–950
Electrical properties	
Electrical resistivity, ohm cm	10000–1.4 × 10^{16}
Surface resistance, ohm	10000–1.0 × 10^{16}
Dielectric constant,	>5.6
Dielectric constant, low frequency	>6.8
Dielectric strength, kV/mm	10–27
Dissipation factor	0.05–0.37
Dissipation factor, low frequency	0.024–0.071
Arc resistance, sec	50
Comparative tracking index, V	250–600
Thermal properties	
CTE, linear 20°C, μm/m-°C	100–177
CTE linear 20°C, transverse to flow, μm/m-°C	100
CTE linear 100°C, μm/m-°C	145
Heat capacity, J/g-°C	1.2–1.9
Thermal conductivity, W/m-K	0.17–0.2
Deflection temp at 0.46 MPa, °C	68–140
Deflection temp at 1.8 MPa, °C	48–125
Brittleness temp, °C	−62 ∼ −30
Glass temp, °C	−42 ∼ −25
Optical properties	
Refractive index, n_D at 25°C	1.42

for polymerization requires a pressure significantly higher (typically >100 bar) than a conventional emulsion polymerization. Fluorocarbons and partially fluorinated hydrocarbons with no significant chain-transfer activities are reported to be a suitable medium for VDF polymerization.[9]

A multistage emulsion polymerization of VDF and acrylics monomer facilitates production of alloys of PVDF and PVDF copolymers with acrylic polymers. The fluoropolymer portion is produced in the first stage having small particle size somewhere between 50 and 150 nm. Acrylic monomers with specific functionalities are generally added in the later stage under atmospheric pressure in the presence of an initiator.[10,11] The final morphology of the alloy depends upon the composition of the fluoropolymer particles, the acrylic monomer type, and the feed schedule. The structure of alloys can range from a core-shell, an interpenetrated structure, to an intimate blend morphology.[12–14]

Hydroxyl terminated telomers of VDF are prepared in methanol using *tert*-butyl peroxide as the initiator. Less than 10 mm thick film has been produce using microwave-stimulated, low-pressure plasma polymerization of VDF. An isoregic PVDF polymer with minimized head-to-head placement has been synthesized and reported.[15] Properties and synthesis of perdeuterated PVDF have also been pointed out.[16]

Polymer Properties

As mentioned previously, PVDF provides a number of outstanding chemical and physical properties, such as

mechanical strength and toughness, abrasion resistance, thermal stability, dielectric strength, melt processable, UV and radiation resistance, low gas permeability, flame resistance, etc. The design properties for PVDF homopolymers are summarized in Table 3.[17,18]

COMMERCIAL PVDF PRODUCERS AND MARKET

The process for PVDF polymerization was first patented by DuPont in the 1940s and acquired by Pennwalt in the 1950s. In 1961, Pennwalt introduced Kynar® PVDF into the market for several applications, including coatings, resins, etc. In 1965, Pennwalt's Calvert City plant came on stream producing a series of homopolymers with varying molecular weight. In the 1970s, Pennwalt introduced Kynar PVDF SL and ADS as low bake and air-dry touch-up paints, and Kynar PVDF 700 series for melt processing. Pennwalt was acquired by Elf Atochem in 1989. In 2002, TotalFina acquired Elf Aquitaine, and Elf Atochem was renamed as Atofina Chemicals. In 2004, Atofina was reorganized into Arkema Inc.

The production of VDF and PVDF is capital intensive as it requires R-142b as a monomer precursor, corrosion-resistant pyrolysis reactors, and low temperature distillation columns for the vinylidene fluoride monomer, high-pressure polymerization reactors, and finishing process. So far, mass production of PVDF has been limited to the United States, Europe, and Japan. Major PVDF producers are shown in Table 4.

The PVDF production in Thorofare, NJ, was originally built by Pennwalt and was diversified to Ausimont in 1999. In late 2001, Solvay started up its Decater, Alabama, plant with a capacity of 5000 MT. In 2002, Ausimont sold its Thorofare plant to Solvay and Solvay diversified its Decater plant to Dyneon (a 3M subsidiary). In 2003, Arkema expanded its PVDF production in Pierre-Benite, France.

The production capacity and market demand for PVDF in major producing and consuming regions are summarized in Table 5. [19]

Table 4 Major PVDF producers

Region	Producer	Location
North America	Arkema	Calvert City, Kentucky
	Solvay	Thorofare, New Jersey
	Dyneon	Decater, Alabama
Europe	Arkema	Pierre-Benite, France
	Solvay	Belgium
Asia	Kureha	Iwaka, Japan

APPLICATIONS OF PVDF

Both homo- and copolymers of vinylidene fluoride are melt processable with a broad operating window. Typical processing temperatures are 180–240°C for extrusion and for injection. At temperatures well above 300°C and approaching 350°C, thermal decomposition will start, and at about 400°C hydrogen fluoride evolution becomes a significant by-product, leading to dangerously high pressure build-up in the equipment. The onset of a thermal degradation will be evident as a darkening, or a discoloration of resin takes place, which in turn is a sign of forming the conjugated double bonds at the expense of dehydrofluorination of PVDF.

All the ordinary extrusion and molding equipment and techniques can be used to process PVDF into different shapes ranging from pipes, fitting, parts, solid rods and sheets, to thin films. Longevity of equipment can be extended using special alloys on the surface that is in contact with molten PVDF, but usually it is not necessary. PVDF based polymers are available commercially in a wide range of melt flow index and with additives to enhance processing or end use properties. However, both homopolymers and copolymers may be shaped from the molten state without extrusion aids or thermal stabilizers.

Industrial Applications

PVDF components are used extensively in the high purity semiconductor market (low extractible values), pulp and paper industry (chemically resistant to halogens and acids), nuclear waste processing (radiation and hot acid applications), and the general chemical processing industry (chemical and temperature applications). Fluoropolymers have also met specifications for food and pharmaceutical processing industries.

There are VF_2/HFP copolymers that are similar to PVDF resins in purity and chemical resistance, but possess unique qualities of additional chemical compatibility in high pH solutions, increased impact strength at ambient and colder temperatures, and increased clarity. The chemical usage guidelines for PVDF and PVDF FLEX can be found from PVDF manufacturers.[20]

PVDF can be fabricated into a wide range of components including:

- Pipes, fittings, and valves
- Pump assemblies
- Sheets and stock shapes
- Films
- Tubing

Table 5 PVDF production capacity and demand in major markets

	Year 2000		Year 2001	
Region	**Capacity (MT)**	**Demand (MT)**	**Capacity (MT)**	**Demand (MT)**
North America	16,300	9700	16,300	8700
Western Europe	7200	3900	7200	4100
Japan	1600	1520	1600	1400

- Tower packing
- Powder-coated metal
- Membranes
- Foam insulation

Offshore and Petrochemical Handling

Because of its exceptionally wide temperature tolerance (up to ~160°C), excellent UV and aging resistance, resistance to a wide range of hydrocarbon solvents and aggressive chemicals, high dielectric strength, flexural response, buoyancy and thermal insulation properties, PVDF is used in a broad spectrum of applications, ranging from sealing and thermal insulation in aggressive environments to pipeline pigs.

Architectural

PVDF resin has been used as the base resin in liquid coatings formulated by leading paint manufacturers. When formulated into a coating composition, which according to the license agreement contains a minimum of 70% by weight of PVDF resin, the resultant coatings exhibit superior

- Color retention
- Chalk resistance
- Gloss retention
- Chemical resistance
- Corrosion resistance
- Flexibility
- Stain resistance
- Overall exterior durability

Substrates

Liquid coatings containing PVDF resin are factory applied to properly cleaned, pretreated, and primed metal substrates, and then oven baked. Metal substrates most commonly coated with PVDF resin based finishes are aluminum, galvanized steel, and aluminum/zinc alloy plated steel. PVDF resin based coatings are typically applied to building components such as metal siding and roofing, storefront extrusions, curtain walls, louvers, skylights, and other miscellaneous metal trim and extrusions. Components can be either postformed from precoated coil stock, or spray coated after fabrication.

Color

Each of the PVDF brand licensees has established its own standard colors. Special colors, including metallics and exotics, are subject to approval by the licensees; however, a broad range of brighter colors are now available.

Composition and materials

Finishes based on PVDF resin are formulated by licensed formulators and contain, in addition to PVDF resin, solvents and high quality pigments. These high quality coating systems have a proven history of performance when exposed to severe ultraviolet radiation for extended periods (over 30 yr) of time.

Fig. 1 compares the performance of PVDF resin-based coating systems with competitive coatings for their weather resistance.[21,22] Panels coated with Kynar 500® and other coating systems were exposed on a South Florida test fence for 10–17 yr. The top of the panels was unexposed while the lower part was exposed to sunlight for 10–17 yr. It was concluded that Kynar 500 PVDF offers superior color retention, greater gloss retention, and better resistance to chalking than other systems tested.

Waterborne PVDF-based coatings

Recently, waterborne and environmentally favorable coatings have been developed based on fluoropolymer resin technology. Arkema is developing novel latex dispersions, using vinyl fluoride fluoropolymers and acrylic polymers similar to those used in PVDF architectural coatings. This new acrylic-modified fluoropolymer (AMF) technology can be used to make air-dry latex coatings with an inertness to photochemical attacklike traditional baked PVDF coatings.[23] In addition, compared to conventional PVDF paint formulations, such waterborne latex dispersions offer

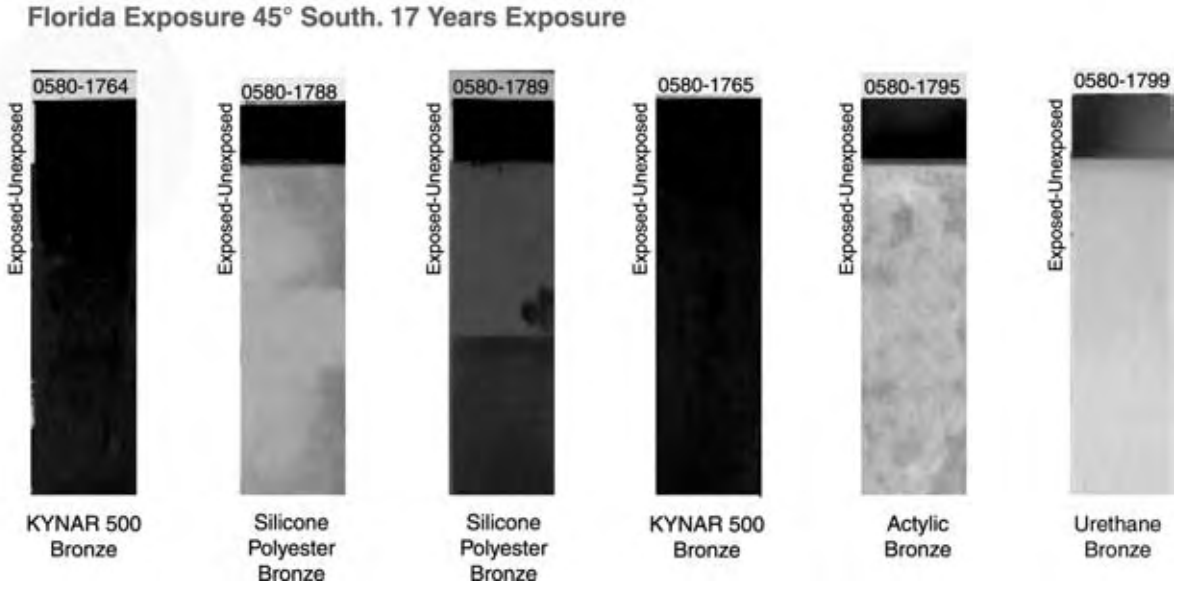


Fig. 1 Comparison of weather resistance of competitive coating systems. *(View this art in color at www.dekker.com.)*

a lower VOC emission while maintaining a comparable long term weathering performance.

Electrical Applications

PVDF/HFP fluoropolymer resin is the premium jacketing material for use on data, voice signal, alarm, and video low-voltage cable. Because the PVDF/HFP fluoropolymer exhibits minimal flame spread and smoke generation, cable jacketing with this fluoropolymer meets National Electrical Code (NFPA-70A) requirements for installation in building plenum areas without the need for conduit casing. In fact, to increase the safety factor, many insurance carriers recommend that their clients install plenum cable throughout in plenum and nonplenum building areas. VF_2/HFP fluoropolymer resin is extremely durable, and resistant to abrasion and mechanical damage during and after cable installation. The flexibility makes cable installation faster and easier. VF_2/HFP fluoropolymer is also a jacketing material for nonplenum cable applications where its flexibility; toughness; light weight; and resistance to chemical, thermal, corrosive, and ultraviolet attack are important. VF_2/HFP fluoropolymer resin is particularly well suited for use in hostile environments where corrosive atmospheres, thermal cycling, or mechanical stress to the cable are possible.

Lithium Ion Batteries and Fuel Cells

PVDF homopolymers and copolymers have gained success in the battery and fuel cell industry as binders for cathodes and anodes in lithium ion technology, and as battery separators in lithium-ion and fuel cell polymer technology. Their high electrochemical, thermal, and chemical stability, as well as ease of processing yields unmatched performance compared to other polymeric binders in lithium ion systems. The development of polymers for such challenging applications takes into account the adhesion properties of the polymer, crystallinity, solvent effects, solution behavior, and slurry preparation and processing.

Lithium ion batteries

Both the anode and the cathode are composed of a coating of the electrochemically active material onto a current collector (copper or aluminum). Another key component of the battery is the separator that physically separates the two electrodes and prevents contact between them. In the case of a liquid technology battery, a polyolefin separator is typically used and a liquid electrolyte is used to transport the Li ions from one side of the porous separator to the other. In the case of a polymer Li ion battery, a polymer, such as PVDF, is used to form a porous structure, which is then swollen with a Li^+ conducting liquid electrolyte.[24] This results in a gel-type electrolyte, which plays the dual role of electrolyte and separator, with no free liquid present.

Fuel cells

PVDF/graphite composites can lend themselves as the best materials for both molded and coated plates. A key component of a fuel cell stack is the bipolar plate. Also referred to as a water transport plate or a separator

plate, this plate functions to channel air (oxygen) to the cathode and hydrogen to the anode, while providing a method to carry away the heat of reaction and water formed in the process. The plates are the backbone of the fuel cell representing significant cost and weight. Operational fuel cells with bipolar plates produced from exotic materials such as titanium, niobium, tantalum, vanadium, etc. have proven useful for the testing of concepts but can never be used in a commercial operation. Machined graphite is also used. Viable processes for bipolar plate manufacture have explored a variety of low cost materials to replace the expensive metals or machined graphite.

Polymer Processing Additives

PVDF-based resins offer a range of fluorinated processing aids specifically designed to improve LLDPE and PP film extrusion, when used at 100–1000 ppm addition levels. Special grades of PVDF are effective in eliminating melt fracture (shark skin) and die build-up while allowing for higher outputs and reduced extruder pressure. This results in improved film quality and strength even when utilizing antiblock agents.[25]

PVDF-based processing aids can eliminate or minimize many of the melt instabilities that occur in high shear extrusion, mainly melt fracture. This allows a wider operating window for process parameters. This coating acts as a lubricant to reduce apparent viscosity, lower die pressure, and melt temperature, and increase film production rates.

Films and Foams

With the flexibility of the blown film extrusion process, one-step production of wide multilayer films is possible. Incorporating a built-in adhesive layer, PVDF films can be laminated directly onto various thermo and thermoset plastics or primed metal substrates. PVDF films can protect the esthetics of UV sensitive substrates while also providing a thermal, chemical, and abrasion-resistant barrier. Recently, white PVDF films have been commercialized that are of full opacity in the UV range of 290–400 μm.[26] Most recently, PVDF has been used in foams. The potential applications include flame-resistant insulation; filtration materials; exterior decorative, acoustical, or insulation materials, etc.[27]

USE, FABRICATION, PROCESSING, AND DISPERSION

PVDF is available commercially in a wide range of physical properties (e.g., melt flow rates, melting point, comonomer content, etc.) and a variety of forms, such as pellets, granulars, latex, fine powders, sheets, films, etc.

Extrusion

All the common extrusion techniques can be used to process PVDF into various products, such as rod, sheet, film, pipe, tubing, monofilament, wire insulation, etc.[28] No extrusion additives, lubricants, or heat stabilizers are needed. Equipment with materials of construction similar to those used for extruding PVC and polyolefins is applicable for PVDF. Extrusion temperatures vary from 210°C to 270°C depending on the type of PVDF resin and the shape being extruded. Extruders with L/D ratios of at least 20 are recommended. For very thin products, the temperature at the dietip can be as high as 320°C as required for heat polishing.

Water quenching can be used for wire insulation, tubing, and pipe. Sheet and film extruded from slit dies are frequently crystallized on polished steel rolls operating at 65°C–145°C. Extruded, blown, or flat film can be uniaxially or biaxially oriented to submil thickness. Monofilament is extrusion spun into a water bath, and then oriented and heat set at an elevated temperature.

Molding

PVDF resins are readily molded in conventional compression, transfer and injection molding equipment. Typical molding temperatures in the cylinder and nozzle are 180–240°C for injection types and 50–90°C for molds. As a crystalline polymer, PVDF typically shows relatively high mold shrinkage of about 2–3%, reflecting the high degree of crystallization and the difference between the solid and the molten densities.

Fusion Welding

PVDF components, such as pipe and tubing, may be joined by various fusion welding methods.[28] Compared to the solvent jointing method, fusion methods are preferred as the joints are stronger and processing itself is easier. Among the most common and efficient fusion welding methods are heat contact welding and hot gas/air using a welding rod. Contact welding is performed by holding two sections that have each been heated to create a molten layer of resin, in position with adequate pressure until the polymer has congealed and cooled. Hot as-welding using a welding road is a common way to construct PVDF lined components such as metals and fiber-reinforced plastics.

Temperatures required to adequately produce a molten layer of the surfaces vary from 210°C–290°C. Welds may be particularly susceptible to stress cracking in strongly alkali solutions where discoloration develops as a result of dehydrofluorination. VF_2/HFP fluoropolymers exhibit improved resistance to a high pH environment, where PVDF homopolymers have been known to become brittle over time.

Dispersion for Coating and Paint

Powder-grade PVDF is the key ingredient in the formulation of premium architectural coatings. The primary components of a PVDF-based coating system are: 1) PVDF resin (typically 15–30%); 2) acrylic resin (typically 6–15%); 3) pigment (typically 10–20%); 4) organic solvent (typically 40–65%); and 5) other additives (typically 2–5%), such as UV absorbers, antisettling additives, surfactants, etc.[3]

The coating formulation includes other ingredients, such as a pigment for color and opacity, modifiers and additives for performance enhancement, etc. Because the coating will be baked at elevated temperatures, the pigment must also be resistant to heat and chemicals. Nonorganic color pigments, such as ceramic compounds, are preferred.

In most cases, the substrates are metals, such as copper, aluminum, etc. The pretreatment of the metal substrate is critical to its performance. In some cases, a primer is applied to impart adhesion, corrosion, and delamination resistance.

Application is usually done through coil coating, spray coating, or powder coating. The coated material then undergoes a baking process, which evaporates the solvent, coalesces the PVDF particles, and fuses them together. The baking process is typically carried out at elevated temperatures above 165°C.

HEALTH AND SAFETY FACTORS

PVDF resins are relatively stable, nontoxic, odorless, and nonhazardous under typical handling conditions. Incidents, mechanical malfunctions, or human error may lead to thermal decomposition with evaluation of toxic hydrogen fluoride (HF) at high temperatures. Precautions must be taken to prevent excess inhalation and physical contact with hydrogen fluoride should decomposition take place. Unlike PVC, PVDF resins will stop decomposing when the heat source is removed and the temperature of the melt is allowed to fall to normal processing temperature. Generally speaking, PVDF has a limiting oxygen index (LOI) of ~44, i.e., an ~44% oxygen environment is needed for PVDF to continue to burn.

PVDF resins are stable when processed as recommended at temperatures between 190°C and 300°C.[19,28] Thermogravimetric analysis has shown that the initial thermal decomposition starts at approximately 375°C. At temperatures approaching 300°C, the onset of a thermal degradation will be evident as a darkening of the resin.

Additives, such as mica, glass fibers, titanium dioxide, and very finely divided metals, may catalyze thermal decomposition rates during processing and should be used with caution.

When handling PVDF powders, personal protective equipment should be used. The powder is not removed readily by dry cleaning; it should be brushed from clothing.

Similar to other fluoropolymers and fluoroelastomers, such as PTFE, FEP, PFA, etc., low-level perfluorinated surfactants or chemicals, such as ammonium perfluoro-octanoate (APFO), etc. may be used in some fluoropolymer production as an emulsifier. These perfluorinated compounds are mostly extremely stable, degrade slowly, and therefore persist in the environment. These surfactants have varying ecotoxicity profiles, and users should contact their supplier for a more detailed ecotox information for their particular product.[29] Industrial efforts are being made to reduce or even eliminate the use of such perfluorinated surfactants in their products and/or manufacturing processes.

CONCLUSIONS

PVDF is a high performance, melt-processible fluoropolymer with good mechanical strength and toughness, high thermal stability, high dielectric strength, excellent resistance to most chemicals, solvents, radiation and weathering, and low flame and smoke characters. Because of its excellent weatherability, PVDF-based architectural coatings have been used widely in high rise buildings as well as industrial and residential applications. Flexible VDF-based copolymers have been used in a wide spectrum of applications, such as plenum wire and cable, film, membrane for lithium batteries and fuel cells, etc. Owing to its outstanding mechanical, physical, chemical, electrical, weathering, and other properties, PVDF is among the cost-effective materials of choice when used in compatible applications.

ACKNOWLEDGMENTS

The authors wish to acknowledge Arkema Inc. for permission to publish this article. The technical and

market information provided by Dr. Jacques Komornicki, Dr. Digen Butala, Mr. David Seiler, Dr. Scott Gaboury, Dr. Mike Burchill, and other Arkema technical staff is greatly appreciated.

REFERENCES

1. (a) *Vinylidene Fluoride: International Chemical Safety Cards*; WHO/IPCS/ILO, 2001. (b) *MSDS for Vinylidene Fluoride*; Matheson Tri-Gas, Inc.: Parsippany, NJ, 2002.
2. Jungnickel, B.-J. *Polymeric Materials Encyclopedia*; CRC Press: New York, 1996; Vol. 11, 7115–7127.
3. Iezzi, R.A. Fluoropolymer coatings for architectural applications. In *Modern Fluoropolymers*; John Wiley & Sons, 1997.
4. Hinksman, P.; Isaac, D.H.; Morrissey, P. Environmental stren cracking of poly(vinylidene fluoride) and welds in alkaline solutions. Polym. Degrad. Stab. **2000**, *68* (2), 299–305.
5. Brothers, P.D European Patent 964009, December 15, 1999 24, 2004, (to E. I. du Pont de Nemours & Co. Inc.)
6. Mertdogan, C.A.; DiNoia, T.P.; McHugh, M.A. Impact of backbone architecture on the solubility of fluorocopolymers in supercritical CO_2 and halogenated supercritical solvents: comparison of poly(vinylidene fluoride-co-22 mol% hexafluoropropylene) and poly(tetrafluoroethylene-co-19 mol% hexafluoropropylene). Macromolecules **1997**, *30*, 7511–7515.
7. DeSimone, J.M.; Romack, T. World Patent 9628477, September 19, 1996 (to University of North Carolina at Chapel Hill)
8. Desimone, J.M.; Romack, T. US Patent 5674957, October 7, 1997 (to University of North Carolina at Chapel Hill)
9. Kruger, R.; Heilig, G.; Sommerfeld, C. World Patent 9948937, September 30, 1999 (to Bayer Aktiengesellschaft)
10. Amin-Sanayei, R.; Hedhil, L.; Wood, K. EC Patent SP-2000-3650, September 12, 2000 (to ARKEMA Inc).
11. Amin-Sanayei, R.; Granel, C.; Hedhil, L.; Wood, K. European Patent 03290237, January 30, 2003 (to Arkema Inc).
12. Gaboury, S.; Drujon, X.F. World Patent 9903900, January 28, 1999 (to Elf Atochem North America, Inc).
13. Kido, K.; Suzuki, F.; Kushida, K. Germany Patent 2952457, July 3, 1980 (to Kureha Chemical Industry Co, Ltd.).
14. Dohany, J. US Patent 4,141,873, February 27, 1982 (to Pennwalt Corp).
15. Cais, R.E. Preparation of Regioregular Polyfluoroethylenes and Products Thereof. US Patent 4,438,247, March 20, 1984 (to ATKT Technology).
16. Cais, R.E.; Kometani, J.M. Polymerization of vinylidene fluoride-d2. Minimal regiosequence and branch defects and assignment of preferred chain-growth direction from the deuterium isotope effect. Macromolecules **1984**, *17*, 1887.
17. *Premium Thermoplastics for Engineering Applications: Kynar® & Kynar® Flex PVDF*; Arkema Inc.: Philadelphia, PA, 2002.
18. Overview—Polyvinylidinefluoride (PVDF), Molded/Extruded, MatWeb Material Property Data. http://www.matweb.com. (accessed 2003).
19. Ring, K.-L. Fluoropolymers: CEH Marketing Research Report, SRI International, 2002.
20. *KYNAR® Chemical Resistance Chart for KYNAR® Homopolymer*; ARKEMA inc.: Philadelphia, PA, 2000. http://www.ARKEMAchemicals.com/kynarglobal/cpi-chem.cfm.
21. *Kynar® 500 PVDF Resin-Based Metal Coatings*; Arkema Inc.: Philadelphia, PA, 2003 & 2005.
22. *Case Studies in Performance: Kynar® 500 PVDF Resin-Based Coatings Versus Polyester Powder and Other Coatings*; Arkema Inc.: Philadelphia, PA, 2002.
23. Wood, K.; Guillot, C.; Cheucle, P.; Gaboury, S. The Effect of Formulation Additives on the Durability of New Fluoropolymer Latex Paints, 6th Nurnberg Congress; Vincentz Verlag: Hanover, Germany, April 2–4, 2001.
24. *KYNAR® PVDF for Lithium Batteries*; Arkema Inc.: Philadelphia, PA, 2002.
25. *KYNAR Flex® Polymer Processing Aids*; Arkema Inc.: Philadelphia, PA, 2001.
26. *Kynar® Film: White UV Opaque Kynar® Films & Colored Kynar® Films*; Arkema Inc.: Philadelphia, PA, 2003.
27. *KYNAR® PVDF Fluoropolymer-Based Foams*; Arkema Inc.: Philadelphia, PA, 2003.
28. *PVDF Fabrication*; Arkema Inc.: Philadelphia, PA, 2001.
29. *Guide to the Safe Handling of Fluoropolymer Dispersions*; Fluoropolymer Manufacturing Group, Society of the Plastics Industry, Inc., October 2001.

Porous Media

Karsten E. Thompson
Gordon A. and Mary Cain Department of Chemical Engineering, Louisiana State University, Baton Rouge, Louisiana, U.S.A.

INTRODUCTION

Porous media are a part of many chemical processing applications, including reaction engineering, chromatography, filtration, separations, catalysis, and more. They also impact related processes: the production of raw materials (hydrocarbons) from porous rock and the transport of chemical contaminants spilled into the environment.

The purpose of this entry is to provide an introduction to the morphology of porous media, the behavior of fluids in porous media, and modeling techniques. While selected equations are presented, the emphasis is on phenomenological behavior, the aim being to give the reader a basic foundation in the physics of transport before a more detailed study might be undertaken. Many important topics have been omitted due to space constraints; see the ''Conclusions'' section for recommended texts on these topics.

CONCEPTS AND DEFINITIONS

A porous medium is a material that is composed partly of solid phase, partly of void phase. In most porous media of interest, the void space is interconnected, which allows for fluid transport through the material. The term ''pore-scale'' implies behavior or analysis performed at a resolution where the void phase and solid phase can be distinguished. At this scale, the void phase is conceptually divided into pores (the larger voids, which provide its volume) and pore throats (constrictions that connect the pores), though the distinction is rarely black and white. In contrast, the continuum-scale approach is usually adopted in engineering practice. Continuum techniques treat the bulk porous medium as a single phase, which in turn requires spatially averaged parameters to be introduced that are intended to capture relevant characteristics of the pore-scale structure.

Unconsolidated porous media are loose packings. Examples include packed columns, sand, and loose fibrous materials. Conversely, consolidated media have a fixed, interconnected solid phase. Examples include sintered materials, reservoir rocks, and certain catalyst supports. The terms fibrous vs. granular media denote the aspect ratio of the particles (long and slender in fibrous materials). A medium's wettability indicates the preference of the solid phase for contact with one fluid as opposed to another. Clean quartz sand is water wet, while sand with adsorbed organic material can become oil wet. Most liquids behave as the wetting fluid relative to air, though mercury is an important exception. Wettability is quantified by the contact angle, which, in this entry, is defined as the angle between the fluid–solid boundary and the fluid–fluid interface at the three-phase contact point, measured through the wetting phase. This definition is not universal, however.

Porosity is one of the most important continuum-scale parameters. It is defined as the fraction of the total volume that comprises void space: $\varepsilon = V_{\text{void}}/V_{\text{total}}$. Equivalently, the solid volume fraction ($\phi = 1 - \varepsilon$) is generally used for fibrous materials or other open structures. The term microporosity implies that the particles in a porous medium are themselves porous, usually at a much smaller scale. A common example is porous catalyst in a packed-bed reactor.

Single-phase flow means that only one fluid phase is present in the void space (implying, most importantly, that no fluid–fluid interface exists). The superficial velocity v is a volume flux based on the *total* cross section of the material. (For flow in a packed column, $v = Q/A$, where Q is the volumetric flow rate and A is the column cross-sectional area.) The interstitial velocity, $v_{\text{int}} = v/\varepsilon$, better characterizes the average fluid velocity actually present in the voids.

MORPHOLOGY

The term ''morphology'' is used in a general sense to include any parameter that quantifies pore-scale structure: pore geometry, pore coordination number, grain shape, porosity, spatial correlations in pore size, etc. Pore morphology varies dramatically from one material to another and has a very strong influence on continuum-scale behavior.

Packed Beds

In industrial packed beds, particle shapes vary widely. Irregular shapes are often used to minimize frictional

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008116

losses, improve contacting or heat transfer, increase surface area, etc. However, spheres are also common (and are more amenable to mathematical analysis than other shapes); consequently, random sphere packings are the most carefully characterized of any specific porous medium. Random packings of uniform spheres exhibit a tight porosity range that falls between 0.36 and 0.42, independent of sphere diameter. Values of $\varepsilon = 0.37 - 0.38$ are most typical in practice, provided the particle-size distribution is tight. However, as polydispersity increases, porosity generally decreases.

The random close-packed (RCP) limit is the lowest porosity achievable without ordering, and has long been a subject of theoretical interest.[1] It is thought to be very close to 0.36, though an exact value has never been proved. Experimental determination of the limit is difficult due to edge effects (which increase the average porosity), local ordering, and the inability to obtain uniform-sized exact spheres. A classic study undertaken by Scott and Kilgour[2] involved packing spheres into a series of dimpled containers and then extrapolating the porosity results to an infinite radius to remove edge effects. Their results give $\varepsilon_{\min} = 0.3634 \pm 0.0005$. Much research has been performed using computer-generated packings, and limiting porosities have been achieved that are slightly lower than the experimental values.[3]

The packing structure near a wall differs from the bulk structure. This phenomenon has important implications with regard to fluid channeling and heat transfer near reactor walls. Fig. 1 is a graphic from a numerical simulation of flow in a chemical reactor.[4] The velocity quivers indicate preferential fluid flow near the walls. This particular simulation contains only a few particles across the reactor diameter. For larger ratios ($d_{\text{bed}}/d_{\text{particles}} > 30$), wall effects decay quite rapidly, penetrating approximately two sphere diameters in from the wall.[5]

In cases where the locations of individual spheres are known, detailed analyses of void structure can be carried out. This situation is restricted mainly to computer-generated packings, though one significant exception is the Finney packing,[1] which is an experimental mapping of ~8000 sphere locations in the interior of a random packing. Mellor[6] provides extensive statistical analysis of the Finney packing based on Delaunay and Voronoi discretizations.

The Delaunay tessellation is particularly important because it provides a tool to decompose the continuum void space into discrete pores, which is essential for pore-scale modeling.[7] An important drawback to using the Delaunay tessellation for extracting pore structure is that it leads to a fixed pore coordination number of 4, which is a geometric artifact; in a real packing, not all voids conform to the tessellation's tetrahedral structure. Experimentally, the average coordination number for a sphere packing has been measured to be $z = 4.3$.[8] A numerical analysis has been performed that constructs irregular polyhedral pores based on local geometry, as shown in Fig. 2. It suggests that this number should be higher, of the order of $z = 5.2$ for $\varepsilon = 0.38$.[9]

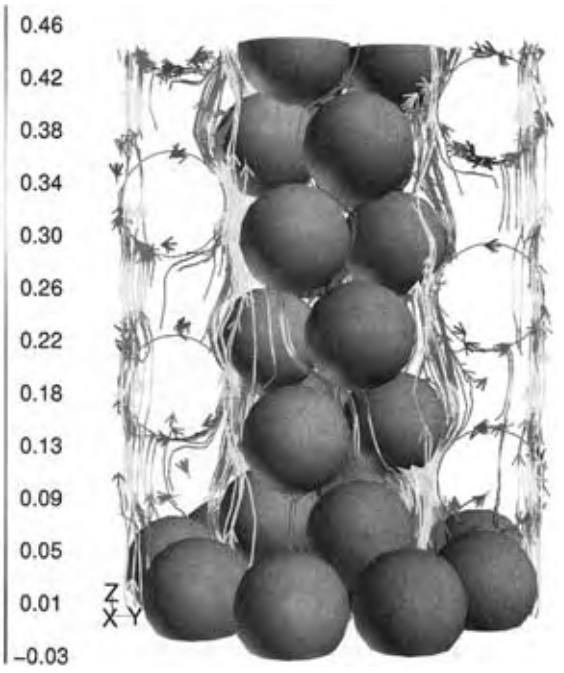


Fig. 1 Numerical simulation of flow in a packed-bed reactor. Velocity quivers illustrate near-wall channeling, which affects residence time distributions, heat transfer, etc. (From Ref.[4].)

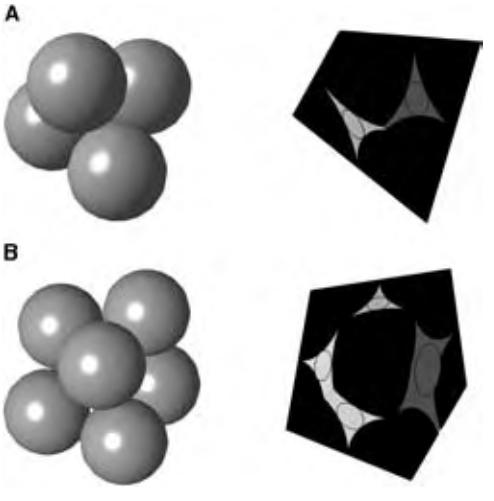


Fig. 2 Three-dimensional visualizations of pore structures in a random packing; sphere clusters on left; polyhedral discretizations of the same on right: (A) $N_s = 4$; $z = 4$ (tightest cluster possible); (B) $N_s = 6$; $z = 7$.

Consolidated Materials

Sandstones originate from loose packings and then become consolidated over time due to compaction, radial grain growth, and cementation at grain contacts. Despite these diagenetic processes, the original granular structure is evident in many sandstones. The final permeability depends primarily on the original grain size and the final porosity, and varies from the millidarcy range to the 1000-darcy range.

Carbonates are formed from irregular particles, and diagenesis has a more pronounced effect on the resulting structure. As such, they are more difficult to characterize using simple parameters such as grain- or pore-size distributions. Factors such as vugs, fractures, and other heterogeneities tend to have an adverse effect on their permeability.

Fibrous Materials

The defining feature of fibrous materials is their high-aspect-ratio particles, which in turn are responsible for significant differences from granular packings: The solid volume fraction can span nearly three orders of magnitude; anisotropy is the norm rather than the exception; and ordered fibrous materials are commonplace (e.g., woven fabrics and materials composed of aligned bundles, both of which are precursors for composite materials manufacturing).

Characterizing Morphology

Because mercury is nonwetting, mercury porosimetry is used to obtain capillary pressure curves during injection (see section on ''Immiscible Displacement''). This technique allows estimates of pore-size distribution to be made.

Thin-section analysis is an important (though destructive) tool. The void space is filled with epoxy or an equivalent compound, and the sample is sectioned and viewed by microscopy. The main advantage is the excellent resolution of detailed pore structure. However, sample preparation and image analysis are labor intensive. Additionally, three-dimensional structure can be obtained only by reconstruction from consecutive thin slices.

Though it is still an evolving technique, tomography is perhaps the most powerful tool for nondestructive analysis. Initial research in this area was performed using medical-grade equipment and gave only continuum-scale resolution. More recently, high-resolution microtomography has been developed, which can resolve detailed pore structure in many types of media.[10,11] Microtomography systems employ synchrotron beamlines or new-generation conventional X-ray sources and can achieve resolutions down to approximately 1 μm. Contrast in the images is derived from density difference (to resolve basic pore structure) or doping agents (to study fluid-phase distributions), both of which affect the absorptivity of the beam.

Raw data from microtomography consist of pixelized values of absorptivity. Depending on the contrast in the data, the pixelized mapping can often be used to generate images directly, as shown in Fig. 3. Otherwise, segmentation must be performed to map the continuous distribution of absorptivity values to discrete phases in the system. Even more challenging is extracting statistical information or reconstructing topology from the pixelized map. The most promising computational tool appears to be medial-axis analysis or a close variant.[10] It creates a skeleton of the void space, upon which the complete void structure or a fluid-phase distribution can be built.

SINGLE-PHASE FLOW

Equations for Fluid Motion

The Reynolds number for porous media is defined using an appropriate characteristic length, usually an average particle or pore diameter, so that $Re = \rho v D_p/\mu$. For single-phase flow at low Reynolds numbers ($Re < \sim 1$), the superficial velocity is linearly proportional to the applied force(s) driving the fluid flow. The most common equation for this linear relationship is Darcy's law:

$$\mathbf{v} = -\frac{1}{\mu}\mathbf{K} \cdot \nabla \mathscr{P} \tag{1}$$

where $\mathbf{v}$ is the superficial velocity, $\mathbf{K}$ is the permeability tensor, and $\mathscr{P}$ is the dynamic pressure. Eq. (1) is the

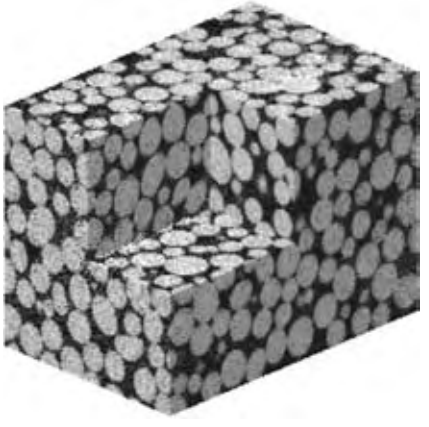


Fig. 3 Three-dimensional image of a packed bed of glass spheres obtained by synchrotron microtomography. (Courtesy of Clinton S. Willson.)

most general form of Darcy's law, valid for three-dimensional anisotropic porous media. In isotropic materials, the velocity and pressure gradient are collinear; hence, permeability can be assigned a scalar value. Additionally, in many practical situations (e.g., flow through a packed-bed reactor), the flow is macroscopically one-dimensional, and a convenient form of Darcy's law becomes

$$\frac{Q}{A} = \frac{K}{\mu}\frac{\Delta\mathscr{P}}{L} \tag{2}$$

The role of permeability in Darcy's law is to quantify the hydraulic conductivity of the medium. Permeability has dimensions $(\text{length})^2$, and is typically assigned units such as cm^2 or μm^2 in hydrology or materials science literature. The unit ''darcy,'' which has retained prominence in petroleum engineering, is based on an inconsistent set of traditional engineering units, and thus requires the awkward conversion $1\,\text{darcy} = 1\,\text{cm}^2\,\text{cp}/(\text{atm s}) = 9.8692 \times 10^{-9}\,\text{cm}^2$.

Beginning at approximately $Re = 1$, inertial forces become significant, and the linear relationship is no longer valid. The Forchheimer equation addresses this limitation by including a quadratic velocity term:

$$\frac{\Delta\mathscr{P}}{L} = a_1 v + a_2 v^2 \tag{3}$$

The Ergun equation (described subsequently) assumes this functionality and incorporates explicit expressions for the two coefficients. For $Re >> 1$, turbulent flow is observed. However, due to the distribution of length scales and velocities, and the large surface area for viscous dissipation, the onset of turbulence is gradual as Re is increased, rather than sudden.

In addition to the Reynolds-number limitations discussed above, a shortcoming of Darcy's law is that it does not allow boundary conditions to be imposed at porous media interfaces (e.g., a no-slip condition at a reactor wall). This problem is addressed by Brinkman's equation, which is a combination of the Stokes equation and Darcy's equation:

$$\nabla\mathscr{P} = \mu\nabla^2 v - \frac{\mu}{K}v \tag{4}$$

Brinkman's equation is restricted to higher-porosity materials: It is invalid for $\varepsilon < 1/3$,[12] and is best used for $\varepsilon > \sim 2/3$.[13]

Predicting Permeability, Pressure Drop, and Flow Rate

For porous media with unknown or complex pore structure, permeability must be determined experimentally. It can be measured by injecting fluid through a cylindrical packed bed or a cylindrical core sample and measuring the required pressure drop. Flow must be single phase (to avoid the very strong effects of multiphase flow on permeability). Additionally, edge effects must be avoided: Flow tends to channel near the edge of packed beds and around consolidated samples unless they are well contained. Methods for containment include cementing a sample into a flow cell with epoxy or placing it in a flexible sleeve and applying a very high overburden pressure to the sleeve.

Experimental measurements are time intensive and require special equipment, so empirical equations for permeability are of great value. Unfortunately, because the structures of porous media are so complex and varied, empirical techniques are quite specialized. A few important relationships are summarized below.

Flow in packed beds—The Ergun equation

For packed beds, a number of semi-empirical equations have been developed by approximating the packed bed as a bundle of tubes. Notably, this approach neglects heterogeneity, interconnectivity, and the converging–diverging nature of the pore space, but nonetheless it has proved effective for single-phase flow in packed beds because friction drag is the dominant effect.

The general approach is to define the traditional hydraulic radius, but for a more complex cross section for flow. For a porous medium, it is equivalent to the ratio of void volume to surface area, which is known for certain geometries, such as a packing of spheres:

$$\text{HR} = \frac{\varepsilon}{1-\varepsilon}\frac{D_\text{p}}{6} \tag{5}$$

The equivalent diameter ($D_\text{eq} = 4\text{HR}$) is used to define the Reynolds number and the frictional loss term in the mechanical energy balance:

$$Re = \frac{2}{3}\frac{\rho v D_\text{p}}{\mu}\frac{\varepsilon}{1-\varepsilon} \tag{6}$$

$$\mathscr{F} = 3f\frac{L}{D_\text{p}}\frac{(1-\varepsilon)}{\varepsilon}v^2 \tag{7}$$

At low Reynolds numbers, the friction factor for flow in tubes is $f = 16/\text{Re}$. For packed beds, both f and v are adjusted by a tortuosity factor to account for the fact that a fluid element travels a longer distance (winding through the pore space) than bed length L. Empiricism has led to adjustments in either the tortuosity or the final equation or both (depending on the derivation) to yield a well-accepted result for

low-Reynolds-number flow:

$$\mathscr{F} = 150\frac{v\mu L}{\rho D_p^2}\frac{(1-\varepsilon)^2}{\varepsilon^3} \quad (8)$$

At higher Reynolds numbers, the friction factor becomes roughly constant, and $\mathscr{F}$ is expected to take on a v^2 dependence. These arguments, and an empirical friction factor of $f = 7/12$, lead to

$$\mathscr{F} = 1.75\frac{v^2 L}{D_p}\frac{(1-\varepsilon)}{\varepsilon^3} \quad (9)$$

These two frictional loss terms can be used in the mechanical energy balance, leading to the Blake–Kozeny equation (for low Re) and the Burke–Plummer equation (for higher Re). Ergun combined the two to give the most well-known equation for flow through packed beds:

$$\frac{\Delta\mathscr{P}}{L} = 150\frac{v\mu}{D_p^2}\frac{(1-\varepsilon)^2}{\varepsilon^3} + 1.75\frac{\rho v^2}{D_p}\frac{(1-\varepsilon)}{\varepsilon^3} \quad (10)$$

(see Ref.[14] and the references therein). Because of its importance in chemical engineering, the Ergun equation has been studied extensively, and various adaptations can be found in the literature to improve predictions for specific media or flow conditions.

Single-phase flow—The Carman–Kozeny equation

The Carman–Kozeny equation is also a bundle-of-tubes model,[15] but is usually written in terms of permeability and, more importantly, retains specific surface area as an adjustable parameter. The latter difference makes it somewhat more general than the Ergun equation, and it has found widespread application in areas such as soil science and materials science.

Carman's approach begins with a general expression for laminar flow in a duct: $v = (\mathrm{HR})^2\Delta\mathscr{P}/(C\mu L)$. The constant C equals 2 for a cylindrical tube and varies between 1.78 and 3 for other regular cross sections.[15] Carman proposed $C = 2.5$ for the generally complex shapes in porous media and incorporated a tortuosity factor $L_{\mathrm{equiv}}/L = \sqrt{2}$, leading to the often-quoted Carman–Kozeny constant, $k_{\mathrm{CK}} \equiv C(L_{\mathrm{equiv}}/L)^2 = 5$. Retaining the specific surface S_0 (instead of assuming spherical particles) and substituting the duct velocity into Darcy's law gives the Carman–Kozeny equation:

$$K = \frac{1}{k_{\mathrm{CK}}S_0^2}\frac{\varepsilon^3}{(1-\varepsilon)^2} \quad (11)$$

Other values of k_{CK} have been measured for various porous media (e.g., see Ref.[16]). The fact that both k_{CK} and S_0 can be adjusted from one material to another makes this equation more flexible than the Ergun equation, though the tradeoff for this flexibility is the need to provide the two constants for a given porous medium.

Fibrous materials—The Jackson and James approach

Empirical permeability predictions are not as effective in fibrous materials as in packed beds due to wider variations in pore morphology. The Carman–Kozeny equation is used at times, in which case k_{CK} and S_0 must be inferred from what is known about the particular fiber structure or from available experimental data.

A more general approach was proposed by Jackson and James,[17] who derived a simple equation for the permeability of fibrous materials based on asymptotic expansions for flow perpendicular to and parallel to cylinders. Their equation for nondimensional permeability is

$$\frac{K}{R^2} = \frac{3}{20\phi}[-\ln\phi - 0.931 + \mathrm{O}(\ln\phi)^{-1}] \quad (12)$$

where R is the fiber radius. The same authors collected a large set of experimental permeabilities that span nearly three orders of magnitude in solid volume fraction, as shown in Fig. 4. These data indicate that order-of-magnitude nondimensional-permeability predictions can made by using ϕ as the sole parameter.

MULTIPHASE FLOW

Multiphase flow theory is very well developed because of applications such as interfacial separations, petroleum production, and vadose-zone hydrology; yet, it remains a mostly empirical science. This is a consequence of the complex physics of flow and because the parameters describing multiphase flow are highly sensitive to pore structure and interfacial configurations. Though three-phase flow is important in applications such as oil production and contaminant transport in the subsurface, we limit the following discussion to two-phase flow for simplicity.

Phase Distributions

At the macroscale, only the *average* fractional saturation of a phase, S_{w} or S_{nw}, is usually known (where, for instance, wetting-phase saturation $S_{\mathrm{w}} \equiv V_{\mathrm{wetting\ phase}}/V_{\mathrm{void}}$). However, the spatial distribution of phases at the microscale is important because of its strong effect on macroscopic behavior. A generic rule of thumb is

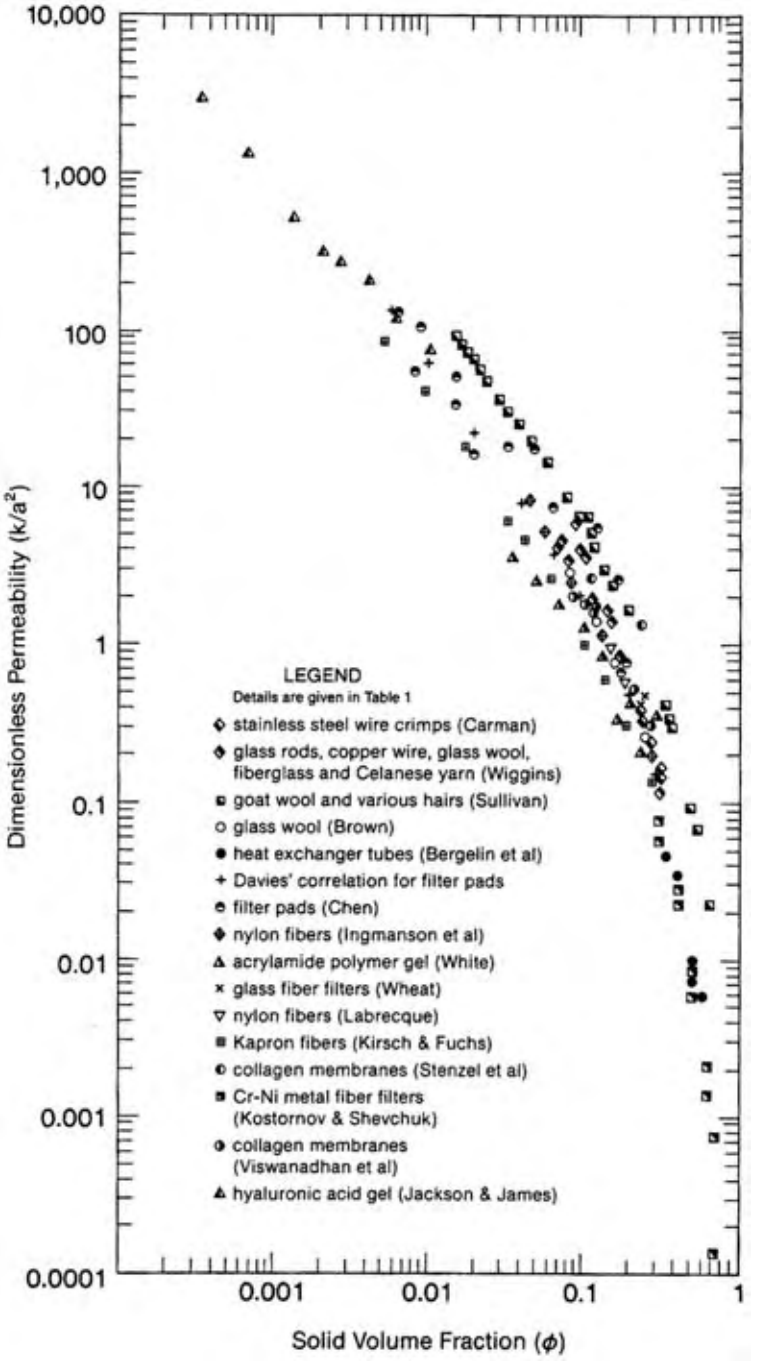


Fig. 4 Nondimensional permeability vs. solid volume fraction for a variety of fibrous porous media. (From Ref.[17].)

that nonwetting phase can be found in larger pores, while wetting phase fills the smaller pores and corners of the larger pores. Furthermore, in strongly wetting materials (i.e., contact angle $\approx 0^\circ$), a very thin film of wetting fluid coats all surfaces. While these generalizations oversimplify the real situation, they provide valuable insight into fundamental behavior.

Consider a nonspherical pore (any shape with angles, corners, cervices, etc.) that contains two immiscible fluids. The nonwetting fluid tends to reside in the center of the pore, reducing its contact area with the solid surface. The wetting fluid occupies the corners and crevices, where the specific surface area is larger. This configuration leads to a pressure difference between the two phases, which is called the capillary pressure. Under equilibrium conditions, it is related to the curvature of the interface according to

$$\Delta P_c = \sigma\left(\frac{1}{r_1} + \frac{1}{r_2}\right) \tag{13}$$

where σ is the interfacial tension. Note that the interface can have opposing radii of curvature r_1

and r_2—see the subsequent discussion of pendular rings. In general, a change in the fraction of wetting phase in the pore changes the interfacial curvature, and therefore must be accompanied by a capillary-pressure change. These simple arguments illustrate a fundamental relationship that exists between phase saturation and capillary pressure. In individual pores, the functionality is complex and of interest only in theory. However, when averaged over the entire medium, this functionality is called the capillary pressure curve and is a key component of multiphase flow modeling.

Because the natural wetting-phase configuration includes occupation of the tighter pore spaces, wetting phase can be drawn into pore throats and cause a disconnection in the nonwetting phase. This behavior leads to arguably the most significant difference between the two fluid phases: the tendency of fractions of the nonwetting phase to become disconnected. The resulting blobs or ganglia can occupy single pores or many pores. In percolation terminology, a ganglion is disconnected from its backbone fraction, meaning that there is no continuous path from the ganglion to the boundary of the porous medium through the nonwetting phase.

The consequences of this behavior are important: Without a connection to the porous-medium boundary, fluid from a ganglion can no longer be evacuated from the pore space by simple pressure-driven flow. Though disconnected ganglia can be mobilized, this is rare in applications where interfacial forces tend to overwhelm viscous or gravitational forces. Hence, formation of a disconnected fraction of nonwetting phase is often viewed as being synonymous with trapping of the nonwetting phase. This behavior is responsible for immobilizing organic-phase contaminants in groundwater and for the formation of trapped voids (defects) during injection molding of composite materials.

The implication of the foregoing discussion is that the wetting phase does *not* become isolated or trapped, which is a difficult statement to assess. Consider a packing of perfect glass spheres. At sphere–sphere contacts, wetting phase collects in pendular rings, which have opposing radii of curvature, as shown in Fig. 5.[18] By increasing the capillary pressure, the pendular ring can be forced to shrink in size, constricting toward its axis. Hence, *in a geometrical sense*, the wetting phase can become disconnected (i.e., it becomes so small that it touches no other pendular ring).

Two issues cloud this picture, however. First, for strongly wetting materials, thin wetting films cover all surfaces. These films provide connectivity in the strictest sense, but do not allow for measurable flow. Second, Dullien et al.[19] showed that if sphere surfaces are roughened, wetting fluid continues to drain even after the rings become geometrically disconnected.

Obviously, the issue of phase connectivity is complex, but as before, it can be summarized by certain useful rules of thumb. In real materials, the wetting phase generally remains interconnected (via a combination of crevice geometry and surface roughness). However, even when connected, it may be very slow to drain at low wetting-phase saturations, giving the appearance of a trapped phase over short time scales. Finally, true trapping of the wetting phase can occur under certain conditions. Factors conducive to this behavior include high-porosity media, smooth surfaces (e.g., synthetic fibers), and nonzero contact angles.

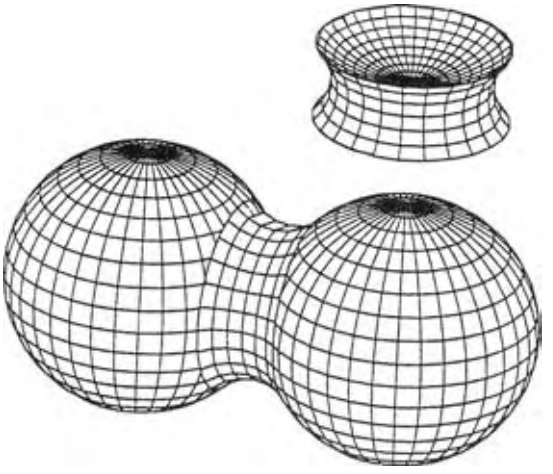


Fig. 5 Schematic of a pendular ring of wetting phase formed at a sphere–sphere contact (inset shows wetting phase only). (From Ref.[18].)

Immiscible Displacement

One phase displacing another from a porous medium is termed ''immiscible displacement.'' The process is usually inefficient from the standpoint of how much of the original phase can be displaced, which is the reason that large amounts of crude oil remain unrecovered in abandoned reservoirs.

Consider, first, displacement of a wetting phase by a nonwetting phase, which is termed drainage. A qualitative understanding of the process can be gained by considering invasion into a single pore, as shown schematically in Fig. 6. Frame B in this figure depicts a nonwetting phase poised at the constriction into an uninvaded pore. As the capillary pressure is increased, the interface deforms until its radius of curvature is small enough to pass through the constriction. At this point, the interface quickly moves across the downstream pore in a non-equilibrium process that has been termed a Haines jump. The dynamics of this jump are very rapid, even if the overall immiscible displacement is carried out very slowly so as to ensure equilibrium interface configurations otherwise.

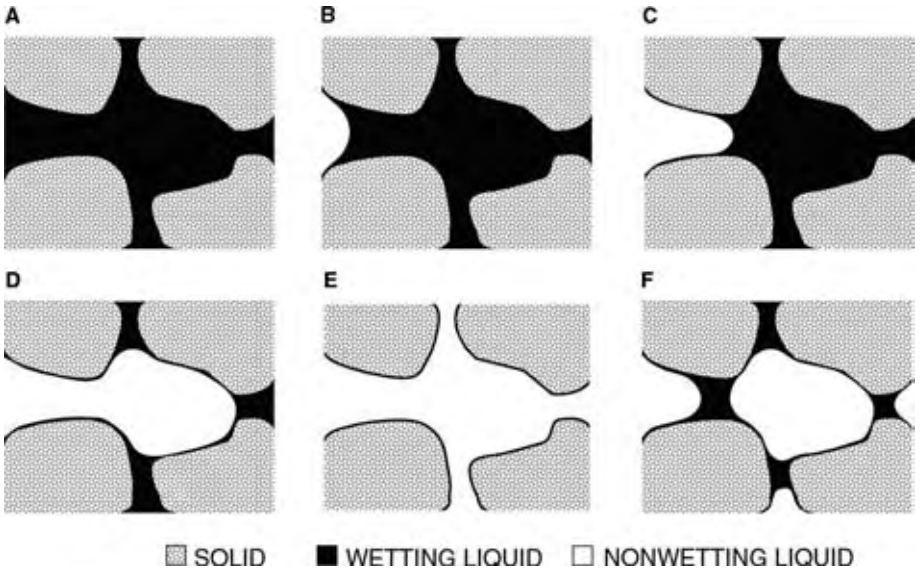


Fig. 6 Schematic of nonwetting phase invasion into a pore: (A) schematic representation of a pore formed by four grains; (B) low capillary pressure; (C) critical capillary pressure for invasion; (D) phase distributions after Haines jump; (E) high capillary pressure; (F) disconnection of the nonwetting phase after snap-off.

Despite the suddenness of these pore-scale events, the displacement process appears smooth at the macroscopic scale: As the capillary pressure is increased, the nonwetting phase flows into increasingly smaller pore spaces, and S_w is thereby reduced. The macroscopic functionality (capillary pressure vs. saturation) is the capillary pressure curve mentioned previously. Fig. 7 is an example capillary pressure curve obtained by injecting mercury into a dry packing of glass spheres.[20] (For the following discussion, view Fig. 7 rotated 90° counterclockwise, which is the more common method of presentation.) The upper curve is the primary drainage branch. It should be followed from right to left beginning at $S_w = 1$ (or $V_{free}/V_{sphere} \approx 0.58$). The initially steep ascent represents the invasion of a small fraction of larger pores that happen to be near the inlet face. The long plateau is characteristic of homogeneous porous media and represents invasion of the bulk of the pore space at an essentially constant capillary pressure. The steep rise at low wetting-phase saturations can be indicative of two different phenomena. In a material with distributed throat sizes, it can represent the fraction of void space composed of small pores (inaccessible except at high capillary pressures). In relatively uniform media, it corresponds to the wetting phase retreating into the tightest corners and crevices of the void space.

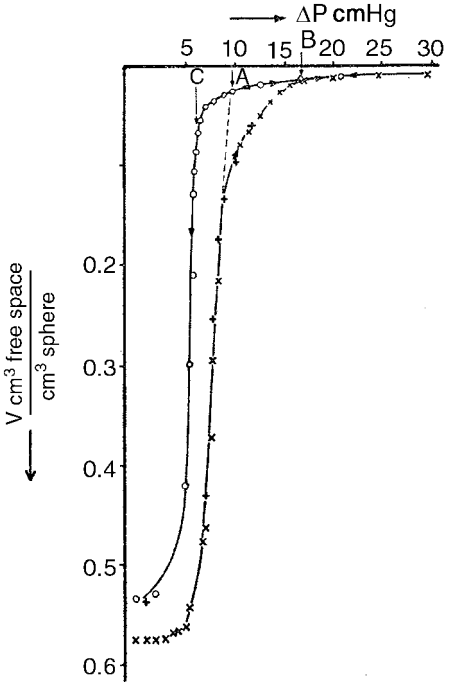


Fig. 7 Capillary pressure curve for mercury (nonwetting phase) intrusion into a packing of glass spheres. (From Ref.[20].)

The reverse process, displacement of the nonwetting phase by the wetting phase, is called imbibition. It is a dramatically different process than drainage. Consider the end of the drainage curve in Fig. 7 as the staring point for secondary imbibition. This point is represented by the Fig. 6E schematic, in which a large capillary pressure must be applied simply to keep the wetting films (black) squeezed into the corners of the pore space. If the capillary pressure is then decreased slowly, the radii of curvature must increase to maintain equilibrium [provided that an interconnected path exists to deliver wetting phase fluid into the pore! (See the discussion in the previous section.)] At some critical capillary pressure, the radii of curvature can no longer increase to maintain equilibrium (because of geometric constraints), the interface becomes unstable, and the nonwetting phase quickly retracts into a connected pore.

As with drainage, these sudden non-equilibrium retractions are masked at the continuum scale, and the process is simply exhibited by the steep descent (increasing S_w) on the imbibition capillary pressure curve. The plateau is lower than the drainage plateau because the instability that causes nonwetting-phase retraction from a pore usually occurs at a lower capillary pressure than that required to invade the same pore during drainage. Thus is derived the generalization that *pore* size controls imbibition, while *pore-throat* size controls drainage.

In reality, other factors complicate the imbibition process. First, retraction of the nonwetting phase can occur only if the nonwetting phase is connected to the backbone fraction. Second, the capillary pressure at which retraction occurs depends on the current configuration of the interface in the pore. This latter effect has been observed in simple geometries[21] but is nearly impossible to quantify in real media because of the complex, saddle-shaped configurations that interfaces adopt. Third, the entire process is rate dependent because of the need to deliver wetting fluid (sometimes through thin films) and evacuate nonwetting phase.

A final distinction worth noting is the occurrence of spontaneous imbibition, which is commonplace: for instance, water absorbing into a dry porous material such as cloth or wood. The physics of this process stem from the interfacial tension: When the wetting fluid contacts the dry material, the tension (in an average sense) is oriented into the porous medium because of the contact angle. When this tension is integrated over the perimeter of the three-phase contact line, a net force results that draws the wetting phase into the medium. Though spontaneous imbibition is rapid when a dry material first contacts a wetting phase, it slows as the net force from interfacial tension is counteracted by viscous forces behind the invading front. If the spontaneous imbibition is upward, it comes to equilibrium when the static head of wetting fluid equals the capillary pressure difference.

Stability of Displacement Phenomena

Immiscible displacements rarely proceed as uniform fronts. The most well-documented instability is viscous fingering, which occurs due to an unfavorable mobility ratio, meaning

$$M \equiv \frac{(k_r/\mu)_{\text{injected}}}{(k_r/\mu)_{\text{displaced}}} > 1 \tag{14}$$

where k_r are relative permeabilities (see the following section). Under these conditions, spatial disturbances in the injection front grow rather than being damped out. The result is that fingers of the injected fluid move well ahead of average front position $v\,\Delta t$. The onset of viscous fingering has been characterized using linear stability analysis (e.g., see Ref.[22]). The nonlinear regime, which is of the most practical importance, is characterized by tip splitting (which produces new fingers), shielding (which stunts the growth of smaller fingers), and stabilizing phenomena. Additionally, there is a strong coupling between viscous instabilities and permeability heterogeneity,[23] though this effect is difficult to generalize because it depends on the specific permeability field and the wavelength of the instability.

Other instabilities include gravity fingering (which is analogous to viscous fingering) and capillary fingering (which is unrelated phenomenologically but also leads to highly nonuniform displacement fronts).

Channeling is the tendency for fluid to respond to continuum-scale variations in permeability. It is sometimes confused with viscous fingering, in part because the two often occur simultaneously in real applications. However, channeling is observed over much larger scales than viscous fingering.

Equations for Multiphase Flow

During multiphase flow, the presence of a second phase in the pore space considerably reduces the apparent permeability of a medium. Mathematically, this behavior is quantified by the relative permeability k_r, which is used in a multiphase version of Darcy's law:

$$\mathbf{v}_w = -\frac{k_{r,w}K}{\mu_w}\nabla\mathscr{P}_w \tag{15}$$

$$\mathbf{v}_{nw} = -\frac{k_{r,nw}K}{\mu_{nw}}\nabla\mathscr{P}_{nw} \tag{16}$$

(Isotropic permeability is assumed for simplicity.) Relative permeability is dimensionless. It is simply a scaling factor that quantifies the reduction in permeability to a given phase due to the fact that it must share the void space.

Intuitively, a strong dependence on phase saturation should be expected: For example, if the nonwetting phase occupies a large fraction of the pore space, then the apparent conductivity to the wetting phase is low. What may be surprising is how strong this dependence is. For the wetting phase, in particular, k_r decreases over orders of magnitude as its saturation decreases. The reason is that wetting phase is relegated to smaller pores, corners, thin films, etc., which are poor conductors of fluid due to their large specific surface areas (and therefore high drag). Hence, as S_w decreases, there is a disproportionate loss of hydraulic conductivity.

For nonwetting-phase fluids, a converse argument could be made: that the relative permeability should

decrease slowly because the fluid occupies larger pores. While this effect is underlying, it is usually offset by other factors: more tortuous flow paths and larger dead-end and disconnected fractions of nonwetting phase. In general, both branches of a relative permeability curve are nonlinear and can exhibit strong hysteresis, as shown by the schematic in Fig. 8. (The data are often presented on a semi-log plot because of the large range of k_r values.) Examples of relative permeability curves for simple porous media can be found in Ref.[24].

Experimental measurement of relative permeability is problematic. Rigorously, experiments should be performed during steady-state multiphase flow: Two fluids are simultaneously injected into a porous medium, and the pressure drop over an interior section (to avoid large capillary-pressure gradients) is measured. However, there are drawbacks to this approach: First, obtaining steady-state multiphase flow is an exceptionally slow process because of subtle shifts in phase distributions en route to the final steady state. Second, determining the phase saturation is a significant challenge. Overall mass balances can be used, but exhibit poor precision. Tomography techniques (preferred) require specialized equipment. For these reasons, other techniques have been developed that allow relative permeability to be measured during a transient immiscible displacement.[25] These techniques require a priori assumptions about the multiphase behavior, but the tests are widely accepted because of the lack of good alternatives.

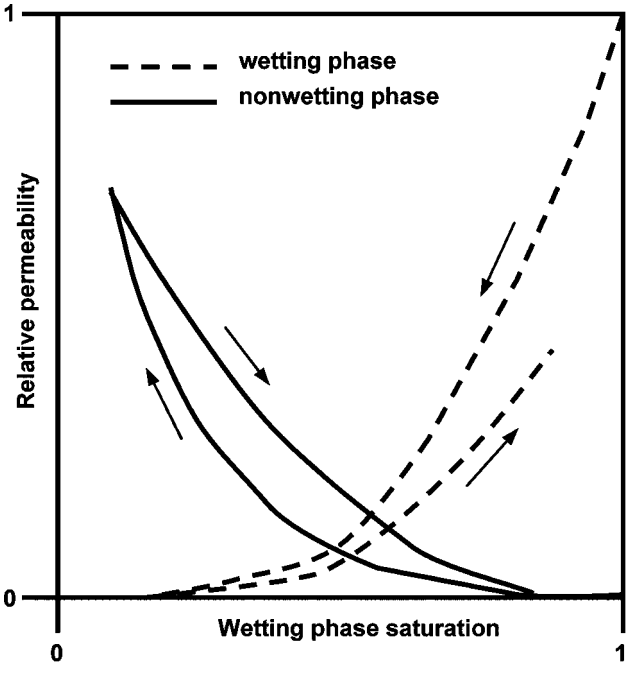


Fig. 8 Schematic of typical relative permeability curves for strongly wetting materials. Curves with positive slope are for the wetting phase. Arrows indicate the direction of change of phase saturation.

In the foregoing discussion, the assumption is that each phase is transported in its own percolating network by a pressure-driven flow mechanism. This is the generally accepted view of multiphase flow in subsurface applications, and is certainly true at low values of the capillary number ($Ca \equiv v\mu/\sigma$). However, blob mobilization is a dominant form of transport in many unit operations in chemical engineering, where the capillary number and Reynolds number are higher. In these cases, specialized correlations for multiphase flow should be used.

MISCIBLE DISPLACEMENT

Examples of miscible displacement are the intrusion of saltwater into fresh groundwater or a step change in feed composition in a chemical reactor. In one sense, miscible displacements are simpler processes than immiscible displacements because issues such as interfacial behavior and phase trapping are not relevant. However, they are complicated by hydrodynamic dispersion (which tends to smear the displacement front), and they are subject to similar viscous instabilities as those described earlier.

Phenomenologically, dispersion occurs because different solute molecules take different times to move from point A to point B in a convective flow. The difference in transit times is attributed to a combination of kinematic and dynamic effects.[26] Kinematic effects stem from tortuosity and the splitting and joining of stream tubes during flow, so that different particles traverse paths of different length. Dynamic effects stem from the distribution of velocities that a particle encounters on its path. Mechanical dispersion consists of the combined effects associated purely with convection (i.e., velocity variations due to hydrodynamics). Nonmechanical effects are associated with the diffusion of solutes from one streamline to another and into stagnant regions.[27]

Dispersion is most commonly modeled as a diffusive process. For flow in a packed column, dispersion is captured by the $D\nabla^2 C$ term in the one-dimensional convection–diffusion equation. The longitudinal dispersion coefficient, D, is a function of the Peclet number, $Pe = vR/D_m$ (where D_m is the molecular diffusion coefficient). At very low Peclet numbers, dispersion stems simply from molecular diffusion and is quantified by $D = D_m/(F\varepsilon)$, where F is the formation factor. At $Pe > \sim 10$, mechanical effects begin to dominate. In practice, D can be considered nearly proportional to Pe over a large range, though theoretical analysis has identified phenomena responsible for nonlinear scaling in this regime.[27] The actual value of the

dispersion coefficient depends on the structure of the material. Numerous tabular and empirical relationships have been developed that allow for its estimation. A summary can be found in Section 6.3.4 of Ref.[12].

A problem with the approach mentioned is that Fickian dispersion (i.e., dispersion that behaves in a diffusive manner) requires solutes to have sufficient time during their transit to be exposed to a representative sample of the velocity field. This requirement is not always satisfied, particularly in situations such as short packed beds, very high Peclet numbers, or media with severe heterogeneities. The latter case is common in fractured materials, where non-Fickian dispersion is nearly always observed. The issue has raised interesting research questions during recent years, and alternative mathematical techniques are only now emerging.[28]

MODELING FLOW IN POROUS MATERIALS

Continuum-Scale Modeling

For calculating one-dimensional flow behavior in a packed bed, the Ergun equation is generally sufficient. For single-phase, constant-density flow in two or three dimensions, Darcy's law can be substituted into the continuity equation ($\nabla \cdot \mathbf{v} = 0$) and simplified to obtain the Laplace equation:

$$\nabla^2 \mathscr{P} = 0 \tag{17}$$

After solving this boundary-value problem for pressure, the velocity field can easily be obtained in conjunction with Eq. (1).

Modeling solute transport requires the convection–dispersion–reaction equation to be used, written as

$$\frac{\partial C_A}{\partial t} = \nabla \cdot (\mathbf{D} \cdot \nabla C_A) + \mathbf{v} \cdot \nabla C_A + R_A \tag{18}$$

for the general case, or

$$\frac{\partial C_A}{\partial t} = D \frac{\mathrm{d}^2 C_A}{\mathrm{d}z^2} + v \frac{\mathrm{d}C_A}{\mathrm{d}z} + R_A \tag{19}$$

for axial flow in a reactor. C_A is the concentration of a reacting species and R_A is its reaction rate (generation). Except for the simplest cases (e.g., steady flow with a first order chemical reaction), a numerical solution is required.

Research on numerical modeling of large-scale flows has been driven by hydrologists and the oil industry. In the latter application, the main focus is on simulating multiphase displacements at the reservoir scale, and physical dispersion is generally unimportant. Hence, the mathematical approach is to write the unsteady continuity equation for each phase and substitute multiphase Darcy's law into the velocity terms. For two-phase flow, this approach gives two equations, with the following four unknowns: S_w, S_{nw}, P_w, and P_{nw}. Two auxiliary equations complete the system of equations: $S_w + S_{nw} = 1$, and $P_{nw} = P_w + P_c(S_w)$. The capillary pressure functionality in the second equation is obtained from the macroscopic capillary pressure curve. The equations also contain relative permeabilities that require explicit evaluation.

In practice, the resulting partial differential equations are usually solved using a finite-difference or finite-volume method. A general approach is to combine the equations so that one can solve implicitly for the pressure field P_w. The phase saturations are then updated implicitly or explicitly, depending on accuracy and stability issues. These ideas are described in a number of general texts (e.g., Ref.[29]). Recent advances in numerical reservoir simulation, including important issues related to upscaling, can be found in the Society of Petroleum Engineers conference proceedings and journals.

Pore-Scale Simulation

The advantage of continuum-scale simulation is the ability to capture large-scale behavior. Its disadvantage is that it is an empirical approach at heart: The fundamental details are wrapped into spatially averaged parameters such as permeabilities, mass-transfer coefficients, etc.

At the pore scale, the basic equations for fluid flow are the fundamental equations of motion. (And, once velocity is obtained it can be used in the convection terms of the energy- and mass-conservation equations.) Two significant hurdles exist when working with these equations. The first is the complexity of the boundary conditions. The second is the very small length scales over which numerical solutions can be obtained.

Volume averaging is a technique in which the fundamental equations are spatially averaged over a representative elementary volume (REV) of porous media.[30] This approach has provided insight into the relationship between fundamental physics and larger-scale behavior but is rarely used for studying transport in specific media in a deterministic sense.

Direct numerical solution using finite-difference, finite-element, and boundary-element methods have played important roles in porous-media research. During the last decade, the lattice-Boltzmann method[13] has emerged as a preferred method for many applications, particularly in the hydrology literature. Significant advantages include relatively simple

coding, the ability to simulate moderate-Reynolds-number flows, and the use of regular grids (which alleviates complex meshing problems). However, the method has certain drawbacks: Its ability to model interfacial phenomena quantitatively remains in question, and it is not amenable to local grid refinement.

Finally, pore-scale network modeling is a specialized technique that has evolved over the past half-century. It is an approximate technique, where the pore space is discretized into pores and pore throats, and local flow is approximated using Poiseuille flow equations. It retains many advantages of other pore-scale techniques: The physics of flow are incorporated using first principles, and it captures factors such as pore-size distribution, interconnectivity, and small-scale heterogeneity.[26] At the same time, significant gains in efficiency are realized by the approximations, so that much larger characteristic lengths can be modeled than with other numerical techniques. Modern network models are built upon physically representative networks of complex three-dimensional porous media and can be used for quantitative modeling in certain applications.[7,9]

CONCLUSIONS

Porous media pose interesting scientific challenges because of the interplay between a series of factors: heterogeneity, complex structure, and multiscale behavior. This entry provides only a brief introduction to important phenomenological behavior, with the expectation that this background will be of value during more in-depth reading.

Excellent texts are available that describe fundamental behavior, including Refs.[12,26,31]. Current research is often published in *Transport in Porous Media* (which tends to have a mathematical focus), *Water Resources Research*, *Journal of Fluid Mechanics*, and the traditional chemical engineering journals.

REFERENCES

1. Finney, J.L. Random packings and the structure of simple liquids. Proc. R. Soc. Lond. A. **1970**, *319*, 479.
2. Scott, G.D.; Kilgour, D.M. The density of random close packing of spheres. J. Phys. D. **1969**, *2*, 863.
3. Jodrey, W.S.; Tory, E.M. Computer simulation of close random packing of equal spheres. Phys. Rev. A. **1985**, *32* (4), 2347.
4. Dixon, A.G.; Nijemeisland, M. CFD as a design tool for fixed bed reactors. Ind. Eng. Chem. Res. **2001**, *40*, 5246.
5. Reyes, S.C.; Iglesia, E. Monte Carlo simulations of structural properties of packed beds. Chem. Eng. Sci. **1991**, *46* (4), 1089.
6. Mellor, D.W. Random Close Packing (RCP) of Equal Spheres Structure and Implications for Use as a Model Porous Medium. Ph.D. thesis, Open University, 1989.
7. Bryant, S.L.; Mellor, D.W.; Cade, C.A. Physically representative network models of transport in porous media. AIChE J. **1993**, *39* (3), 387.
8. Yanuka, M.; Dullien, F.A.L.; Elrick, D.E. Percolation processes in porous media Int. J. Colloid Interf. Sci. **1986**, *112* (1), 24.
9. Al-Raoush, R.; Thompson, K.E.; Willson, C.S. Comparison of network generation techniques for unconsolidated porous media. Soil Sci. Soc. Am. J. **2003**, *67*, 1687.
10. Lindquist, W.B.; Lee, S. Medial axis analysis of void structure in three-dimensional tomographic images of porous media. J. Geophys. Res. **1996**, *101* (B4), 8297.
11. Al-Raoush, R.I.; Willson, C.S. Extraction of physically-representative pore network from unconsolidated porous media systems using synchrotron microtomography. J. Hydrol. **2005**, *300* (1–4), 44.
12. Dullien, F.A.L. *Porous Media: Fluid Transport and Pore Structure*, 2nd Ed.; Academic Press: San Diego, 1992.
13. Koch, D.L.; Ladd, A.J.C. Moderate Reynolds number flows through periodic and random arrays of aligned cylinders. J. Fluid Mech. **1997**, *349*, 31.
14. Bird, R.B.; Stewart, W.E.; Lightfoot, E.N. *Transport Phenomena*, 2nd Ed.; Wiley: New York, 2002.
15. Carman, P.C. Permeability of saturated sands, soils and clays. J. Agric. Sci. **1939**, *29*, 262.
16. Wyllie, M.R.J.; Gregory, A.R. Fluid flow through unconsolidated porous aggregates. Ind. Eng. Chem. **1955**, *47* (7), 1379.
17. Jackson, G.W.; James, D.F. The permeability of fibrous porous media Can. J. Chem. Eng. **1986**, *64*, 364.
18. Gvirtzman, H.; Roberts, P.V. Pore scale spatial analysis of two immiscible fluids in porous media. Water Resour. Res. **1991**, *27* (6), 1165.
19. Dullien, F.A.L.; Zarcone, C.; Macdonald, I.F.; Collins, A.; Bochard, R.D.E. The effects of surface roughness on the capillary-pressure curves and the heights of capillary rise in glass bead packs. J. Colloid Interf. Sci. **1989**, *127*, 362.
20. Kruyer, S. The penetration of mercury and capillary condensation in packed spheres. Trans. Faraday Soc. **1958**, *54*, 1758.
21. Lenormand, R.; Zarcone, C.; Sarr, A. Mechanisms of the displacement of one fluid by another

in a network of capillary ducts. J. Fluid Mech. **1983**, *135*, 337.
22. Tan, C.T.; Homsy, G.M. Stability of miscible displacements in porous media: rectilinear flow. Phys. Fluids **1986**, *29* (11), 3549.
23. Chen, C.-Y.; Meiburg, E. Miscible porous media displacements in the quarter five-spot configuration. Part 2. Effect of heterogeneities. J. Fluid Mech. **1998**, *371*, 269.
24. Naar, J.; Wygal, R.J.; Henderson, J.H. Imbibition relative permeability in unconsolidated porous media. Soc. Pet. Eng. J. **1962**, *2* (1), 13.
25. Johnson, E.F.; Bossler, D.P.; Naumann, V.O. Calculation of relative permeability from displacement experiments. J. Pet. Tech. **1959**, *216*, 61.
26. Sahimi, M. *Flow and Transport in Porous Media and Fractured Rock*; VCH Press: Weinheim, Germany, 1995.
27. Koch, D.L.; Brady, J.F. Dispersion in fixed beds. J. Fluid Mech. **1985**, *154*, 399.
28. Berkowitz, B. *Dispersion in Heterogeneous Geological Formations*; Kluwer: Dordrecht, Netherlands, 2001.
29. Mattax, C.C.; Dalton, R.L. *Reservoir Simulation*; Society of Petroleum Engineers: Richardson, TX, 1990. Monograph Vol. 13.
30. Whitaker, S. Volume averaging of the transport equations. In *Fluid Transport in Porous Media*; du Plessis, P., Ed.; Computational Mechanics Publications: Southampton, U.K., 1997.
31. Bear, J. *Dynamics of Fluids in Porous Media*; Dover: New York, 1972.

Powder Coating Application Processes

Harry J. Lader
Harry Lader and Associates, Inc., Cleveland, Ohio, U.S.A.

INTRODUCTION

Powder coating originated in the 1950s as a method to coat parts by immersing preheated parts in a fluidized bed. This method was used primarily for corrosion resistance or electrical insulation since only thick films were possible. These powder coatings were dry-blended mixtures of pigmented epoxy resins, flow agents, and curing agents. Over time, electrostatic fluidized beds were developed that provided better control over film thickness and are now used mostly for coating small parts. Powder coatings were compounded using extruders in the 1960s. This development meant that powder coatings could now be applied as an electrostatic spray, instead of using fluidized beds. The resin of choice was epoxy due to its good storage stability, good flow properties, and reasonable appearance. Epoxy powders had the ability to retain an electrostatic charge that allowed the powder to be attracted to and adhere to a grounded part during the application process. The charged powder remained on the part until the powder was cured. The ability to formulate a powder coating as compounded particles spurred tremendous growth for the powder coating industry.

While compounded epoxy powders stimulated the development of the powder coating industry, epoxies were prone to chalking and gloss reduction because of poor stability to ultraviolet (UV) radiation for exterior applications. This property deficiency encouraged research on other resin systems and eventually led to a range of powder coatings that now include epoxy/polyester hybrids, TGIC (triglycidyl isocyanurate)–polyesters, polyurethanes, acrylics, in addition to epoxies. These powders are known as ''thermosetting powders'' as they crosslink on heating and comprise a majority of the powder coatings in use today. On the other hand, thermoplastic powders, which can also be used in electrostatic powder coating applications, do not crosslink on heating and therefore can be remelted. Examples of these types of powders include polyolefins, polyvinyl chloride, polyamides, and polyvinylidene fluoride.

In this entry, we review basic charging concepts, corona charging, internal charging, tribocharging, and electrostatic fluidized beds. Along with these methods, some of the properties of powders needed for their application are also discussed.

BACKGROUND

Advantages

The major advantage of powder coatings is that there are no solvents with which to contend; hence, compliance with environmental regulations is much easier. The problems and costs associated with them, such as toxicity, mixing, thinning, and labor, are eliminated while reducing time and material variability. Another major benefit is the high material utilization through recycling. While a primer is generally not required, the part must be properly cleaned to ensure a good quality coating.

After powder has been applied, a convection oven, infrared heating, or a combination of infrared followed by UV curing is used to cure the part. In some cases, induction heating of metals parts is used. Powder coated parts are fully cured when they leave the oven; hence the total processing time tends to be shorter as compared with wet coatings. This results in a reduction of rejects due to damage that can occur after painting. Powder coating allows the parts to be racked closer together on a conveyor allowing more parts to be processed in a shorter time. Powder coated parts are cured at elevated temperatures, usually between 160°C and 220°C (320–430°F), which generally results in a tougher, more chip-resistant coating.

Disadvantages

While there are many advantages to using powder coatings, there are also some disadvantages. Some of the challenges include the ability to quickly change colors, especially to take advantage of the benefits of recycling; color matching and uniformity; particle size variation during recycling that could lead to color variation, especially with regard to metallic powders; high cure temperatures; film thickness nonuniformities, especially at edges, corners, or in recessed areas; thin films; and good surface appearance with high transfer efficiency. Advances in equipment technology are addressing some of these challenges, while others are being improved through material technology.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007927

Powder Coating Markets

Powder coatings are used in a wide range of markets. Epoxy powders are typically used for functional applications such as steel pipelines and rebar that is used in concrete construction. These functional coatings typically have film thicknesses of over 250 μ (10 mil) and are usually applied onto preheated substrates. Many decorative coatings also serve a protective function, but have to be visually appealing as well. These coatings are typically applied between 25 and 100 μ (1–4 mil). The choice of powder depends on the environment in which the part is to be used. Today, polyester/epoxy hybrids comprise almost half of all the powders used worldwide. Their improved weatherability over epoxies and lower cost have allowed them to capture the largest percentage of the powder market. Polyesters have captured almost 40% of the world market, while epoxies have diminished to only about 10%. Polyurethane and acrylic powders are in low single digits.[1] Typical applications for powder coatings include metal furniture and fixtures, machinery, major and small appliances, automotive, electrical components and equipment, building and architectural, farm, and garden equipment.

The types of powder coatings available are gradually evolving with the advent of new powder chemistries that allow for numerous formulation variations. For example, powder coatings have been developed to replace porcelain in laundry applications, heat-resistant polymers for range door trim, and flexible powder coatings for refrigerator doors. This wide latitude in formulation capabilities means that there may be more than one resin chemistry used in a particular powder formulation.

OVERVIEW OF THE POWDER COATING APPLICATION PROCESS

Powder charging can use either a corona gun or a tribocharging gun. Other types of application devices are discussed later. Corona charging uses a gun applicator with an electrode at the tip, usually at a negative high voltage, while a tribogun does not need an external power supply as the powder is charged by friction. A fluidized bed or a box feeder is used to feed the powder to the gun. The powder in the fluidized hopper is mixed with dry compressed air through a porous fluidizing plate so that the powder acts like a liquid. A pump transports the powder through a delivery tube into a corona- or tribogun applicator. The charged powder is sprayed through the gun onto a part. Nozzles are used to shape the spray pattern to the part so that most of the powder attaches itself to the grounded part. The booth contains the powder. This powder can be recovered and transported back to the feed hopper, if necessary. Virgin powder is added to the system to maintain powder levels. (See Ref.[2] for an excellent practical guide on powder coating.)

Recently, the adoption of lean manufacturing practices, which focuses on meeting customers' needs to quickly deliver any quantity of any parts in specified colors, has resulted in many color changes daily. This need has encouraged equipment manufacturers to provide the ability to change powder color quickly. In addition, equipment improvements have increased transfer efficiency and have reduced the need to recover powder. By increasing first pass transfer efficiency, spray-to-waste is gradually becoming accepted.

POWDER CHARGING AND APPLICATION

Corona Charging

A basic knowledge of corona charging is important for understanding how powders charge, the limitations of charging efficiency, and how this impacts transfer efficiency and surface appearance. A corona discharge occurs when a high voltage is applied to an electrode in the presence of a ground. When the electric field at the electrode is very high and exceeds the breakdown of air, negative ions, present naturally in air, are split into both an electron and a neutral atom when they enter the high field region. The electron is driven away from the high electric field and gains enough energy to ionize an air molecule, creating a cascading effect. At the same time, naturally occurring positive ions are attracted to the negative point electrode, where they attain sufficient energy from the high electric field and release secondary electrons that add to the discharge. This process is self-sustaining above the threshold voltage. Light emitted during ionization creates a glow around the sharp electrode causing a "corona."[3]

Numerous negative ions are created when the electrode goes into corona. These ions will follow the field lines toward the grounded part. Fig. 1 is a schematic of a corona gun spraying powder and shows the electric field lines between the electrode and a flat part. The powder dispersed from the nozzle will enter the region between the electrode and the grounded part. Most of the air-diffused powder particles will attract ions and become charged. The electric field lines will accelerate the charged particles toward the grounded part and the free ions will also continue along the same path.

Fig. 1 also shows that the electric field lines concentrate on the sharp edges of the part. Some of the electric field lines actually wrap around the back of the target. In addition to the electric field lines, aerodynamics will also contribute to this "wrap" effect. Increased film thickness at the edge is usually referred

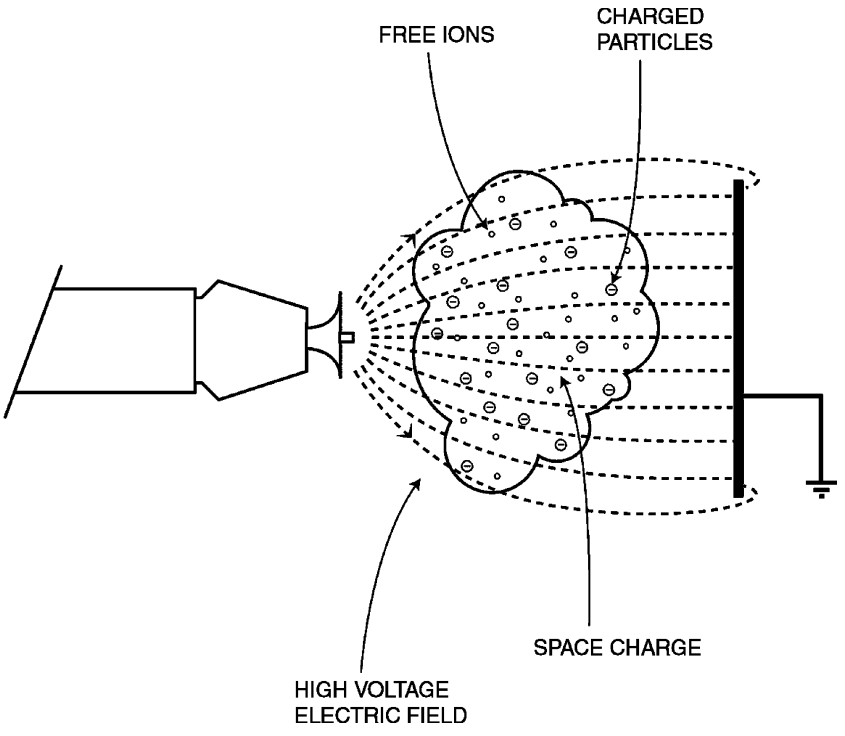


Fig. 1 Schematic of a corona gun with an electrode at high negative voltage with charged powder and ions in the presence of a grounded flat part.

to as the ''picture frame'' effect. The charged powder and free ions between the gun and ground comprise a charged cloud known as the ''space charge.'' This space charge creates its own electric field lines to the ground.

Fig. 2 is a schematic representation of the electric field lines from the gun to a part that has a protrusion at its center. In this case, the electric field lines will concentrate at the protrusion. The charged powder and free ions will build up faster here and contribute to back-ionization. Because the powder particles are strong dielectrics, their charge will be retained for many hours, depending on a number of factors such as powder resistivity, as well as the ambient relative humidity.

The maximum charge that can be acquired by a spherical, nonconducting particle in a uniform electric field was determined by Pauthenier[4] and is shown in Eq. (1):

$$q_{\max} = 4\pi\varepsilon_0 a^2 E p, \quad \text{where } p = \frac{3\varepsilon_r}{(\varepsilon_r + 2)} \tag{1}$$

In Eq. (1), ε_0 is the permittivity of free space (8.8×10^{-12} F/m), E is the maximum electric field before breakdown of air occurs (3×10^6 V/m), a is the particle radius in m, and ε_r is the relative dielectric constant of a powder particle. This equation shows that the maximum charge that can be acquired is proportional to the electric field and the square of the particle radius. It is also weakly dependent on the relative dielectric constant of the particle. The charge-to-mass ratio takes into account the mass of the particle and is determined by dividing Eq. (1) by the mass of a spherical particle, $(4/3)\ \pi a^3 \rho$, where ρ is the density of the

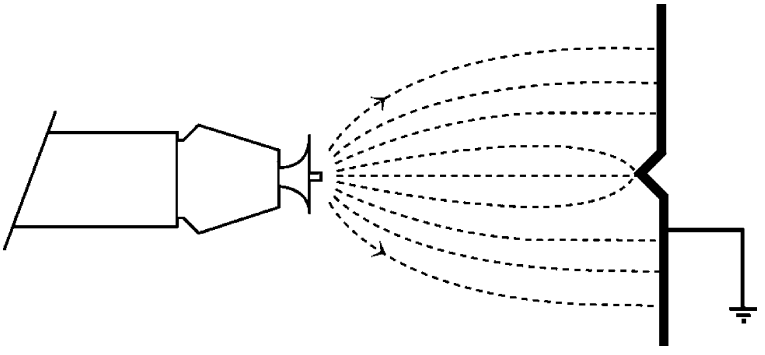


Fig. 2 Schematic of a corona gun with an electrode at high voltage in the presence of a part containing a protrusion at its center.

powder (kg/m^3), as shown in Eq. (2):

$$\frac{q_{max}}{m} = \frac{3\varepsilon_0 Ep}{a\rho} \quad (2)$$

Eq. (2) shows that the maximum charge-to-mass ratio is inversely proportional to the size of the particle. This equation shows that the maximum charge-to-mass ratio of a 100 μ particle will be 10 times lower than a 10 μ particle. It is also important to note that because the mass of a spherical particle is proportional to the third power of a particle's radius, the mass of one 100 μ spherical particle is the same as that of a thousand 10 μ particles.

The actual charge-to-mass ratio of a charged particle will depend on the mechanism of charging used, such as the electrical characteristics of the charging gun, as well as the surface chemistry and electrical resistivity of the particle, particle shape, ambient relative humidity, and the charging time. Powder particles are not perfectly spherical. Rather, they have a gravel-like shape and, as they have a high electrical resistivity, they will not be uniformly charged. Also, the particles spend only a finite amount of time in the charging zone. Therefore, the theoretical maximum charge-to-mass ratio cannot be achieved. Under ideal conditions, smaller particles can have a larger charge-to-mass ratio. However, when they are mixed with larger particles, the field lines will interact more strongly with the larger particles, making it more difficult for the smaller particles to become charged.

Aerodynamics and gravity also come into play. Smaller particles will tend to follow the airflow lines across the part while the larger particles, having a greater inertia, will generally move in the direction toward the part. Gravity will tend to act more strongly on larger particles, pulling them downward. While some of the coarser particles will bounce or fall off the target and be pulled downward by gravity, the finer particles will tend to overspray the target.

Eq. (3) shows how long it will take a particle to reach its maximum charge.

$$q(t) = \frac{q_{max}}{1 + \tau/t} \quad (3)$$

Here, τ is the time for a particle to reach half of its maximum charge and was calculated to be about 0.1 sec for a typical corona discharge.[5] After 1 sec, 90% of the maximum theoretical charge will be realized. Particles having a lower velocity will spend more time in the charging zone and will have a higher charge-to-mass ratio. Reducing the powder feed rate and its velocity will allow the powder cloud to be more diffuse and spend more time in the charging zone, improving powder charging that will result in more powder being deposited onto the target.

Corona charging is very widely used due to its ability to charge a wide range of powders at fairly high throughput. This charging method is relatively insensitive to ambient humidity, plus the equipment is robust for a wide variety of manufacturing processes.

Back-Ionization

Increasing the voltage on the electrode tends to increase transfer efficiency, up to a point. As voltage is increased, back-ionization also increases until powder deposition is reduced. This is due to the fact that less than 1% of the total number of ions generated contributes to charging the powder and more than 99% of the ions are deposited on the surface of the part and any other nearby grounds.[6] Ions are much smaller and have a much faster mobility than powder particles. As powder is sprayed onto the part, the particles adhere where little or no powder has previously been deposited. Back-ionization begins with as little as a single monolayer of powder.[7] As powder is deposited onto a surface, electrical discharges in the layer can be observed by using an image intensifier. Air that is trapped in the powder layer is highly stressed, which causes it to break down into both positive and negative ions at these discharge points. The positive ions move away from the part and "back" toward the negative electrode, reducing the efficiency of powder deposition onto the part by neutralizing the oncoming negatively charged powder. The effects of back-ionization can usually be seen in the uncured powder as "starring" or "orange peel" in the finished coating. In severe cases, back-ionization can cause craters and pinholes down to the substrate.

Equipment manufacturers have found ways to deal with this problem. One method simply uses a grounded metal rod or ring, known as an "ion collector," positioned behind the electrode but at a distance closer to the electrode than the distance between the electrode and the part. The idea is to create a stronger electric field between the electrode and the ion collector than between the electrode and grounded part, so that fewer ions will be attracted to the grounded part.[8] Internal charging is another way to reduce the effects of back-ionization. In this case, the ground electrode is located inside the powder applicator, instead of outside, which neutralizes a significant portion of free ions during the charging process.

A recent study was done to examine the effectiveness of ion collectors in reducing ion current and its effect on surface appearance.[9] A smooth ring and a ring with 11 sharp points for concentrating the electric field were investigated. The authors found that both

the charge-to-mass ratio and transfer efficiency correlate directly with an increase in ion current on the target. That is, a reduction in ion current responsible for back-ionization occurs at the expense of transfer efficiency. Interestingly, they also found an improvement in surface appearance with the smooth ring that collected fewer ions. Therefore, care needs to be taken when using ion collectors to reduce back-ionization while optimizing transfer efficiency and surface appearance.

Another technique to reduce back-ionization is to control the gun current. Controlling gun current helps reduce excessive free ions by automatically reducing the electrode voltage when the gun-to-part distance decreases. This method reduces back-ionization and the Faraday Cage effect (see below).

Free ions also affect the ability to apply a second layer of powder on top of a coating that has already been cured. This application is known as ''recoating.'' When charged powder and ions are applied to this surface, the charge builds rapidly as charge cannot bleed off. Therefore, minimizing the amount of free ions makes recoating easier to achieve.

The Faraday Cage Effect

As discussed earlier, electric field lines concentrate on sharp points or edges. Fig. 3 shows what happens to the electric field lines when the part has a recessed area or a channel. Here, the electric field lines concentrate on the edge of the corner. This means that the powder and ions will deposit more rapidly on these areas and will tend to go into back-ionization. Furthermore, as there are very few field lines inside the channel, the powder has difficulty penetrating there. Back-ionization at the edge of the recess will also neutralize incoming charged particles and prevent them from depositing inside the corners. The use of ion traps, controlling gun current, or an internal charging gun can be used to overcome the Faraday Cage effect.

In addition to the Faraday Cage effect, aerodynamic effects do not favor spraying inside a corner. Airflow considerations based on part geometry must also be taken into account to deliver the highly charged powder inside a corner or a channel, while not using so much air that the powder is prevented from depositing.

Internal Charging

Corona guns have the disadvantage of the Faraday Cage effect from strong electric fields and back-ionization due to excess ions. As seen later, tribocharging largely overcomes the problem of the Faraday Cage effect and back-ionization, but is sensitive to the chemistry of the powder and the extreme ambient relative humidity conditions. Internal charging guns can minimize these disadvantages by putting both the high-voltage electrode and the ground inside the gun.[8] Fig. 4 is a schematic of an internal charging gun. As both the charging electrode and the ground are inside the gun, few electric field lines will escape outside the gun minimizing back-ionization resulting in a smoother finish, similar to a tribogun. Furthermore, the powder does not have to be specially formulated as it does for the tribogun. External electrodes are used to improve transfer efficiency. Powder flow rates are similar to a conventional corona gun.[10] A patent on an internal charging gun[11] refers to a chamber used to slow down and swirl the powder to increase the time of the particles in the charging zone.

Tribocharging

This method charges powder as a result of frictional contact with the inside surfaces of a tribogun. No source of high voltage is required. Transfer of charge occurs when two materials that differ in their ability to accept electrons are brought into close contact with one another. Materials that have the ability to accept

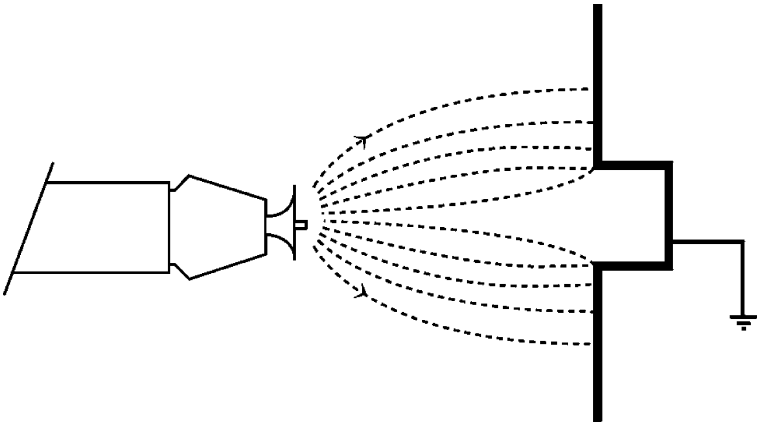


Fig. 3 Schematic of a corona gun with an electrode at high voltage in the presence of a part containing a recess.

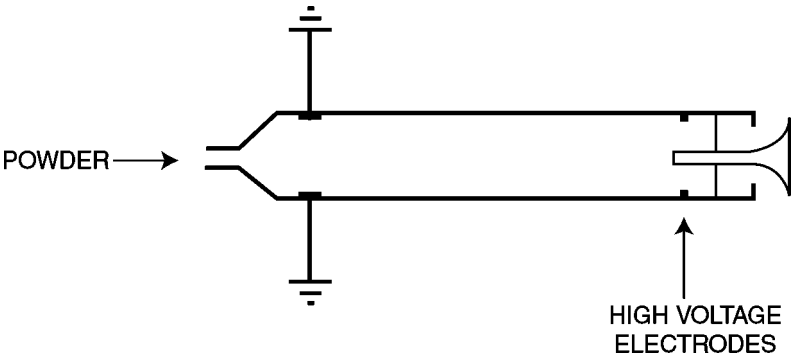


Fig. 4 Schematic of an internal charging gun.

electrons are called electronegative, while those that can donate electrons are called electropositive. The internal surfaces of a tribocharging gun are usually made of a material that is very electronegative, such as polytetrafluoroethylene or Teflon®.[12] This material accepts electrons from powder particles, causing the powder to charge positively. A tribogun typically uses additional air to help charge the powder. The path inside the gun is designed to ensure multiple contacts between the powder and the internal walls.

Fig. 5 shows a schematic of a tribocharging gun. As negative charge builds up inside the gun, a means is needed to bleed off this excess charge. This is accomplished by grounding the gun, sometimes through a microammeter to monitor or record the current. The current-to-ground is an indication of the amount of charging that occurs inside the gun. If the powder flow rate is constant, the charge-to-mass ratio can be determined by dividing the current (in μA) by the powder flow rate (in g/sec) to obtain μC/g.

Because triboguns have no charging electrode, the Faraday Cage effect is greatly reduced. This allows the powder to penetrate recesses, corners, and deep channels. In addition, as triboguns do not generate free ions, as corona guns do, they can coat parts that are not highly conductive more easily than a corona gun. This means that triboguns are better at recoating parts, as they do not build up charge as fast as corona guns. The resulting surface appearance is usually smoother and has considerably less back-ionization at a given film thickness. Triboguns, therefore, can build thicker films than corona guns before back-ionization becomes a problem.

The amount of charge on powder is determined by many factors including the composition of the powder, particle size, velocity of the powder on contact, and ambient relative humidity. Some powder formulations will tribocharge better than the others depending on their ability of being charged positive. Most polyamide and epoxy powders will tribocharge well. Some epoxy/polyesters, polyurethanes, and TGIC–polyester powders can also be tribocharged. Powder formulators can adjust the tribocharging properties of the powder through selection of the resin and the use of additives.

Another factor that controls charging efficiency is particle size distribution, which is discussed in more detail later. Larger particles will impact the tribosurface with greater energy, resulting in better charging. If the level of fine particles is too high, they will tend to build up in the reclaim system as they would not charge as effectively as larger particles. Typically, tribocharging powders have a median particle size between 30 and 45 μ, slightly greater than corona or internal charging guns. The level of fine particles (under 10 μ) should be minimized to enhance the charging process and also to minimize problems associated with fluidization.

Recently, a new tribocharging method has been developed that charges the powder negatively[13] instead of positively, as conventional tribocharging. This method relies on air jets to force the powder particles to contact the internal walls. In this case, the tribocharging surface gives up an electron to the powder. This method has the same advantages as conventional tribocharging. In addition, this charging gun can be

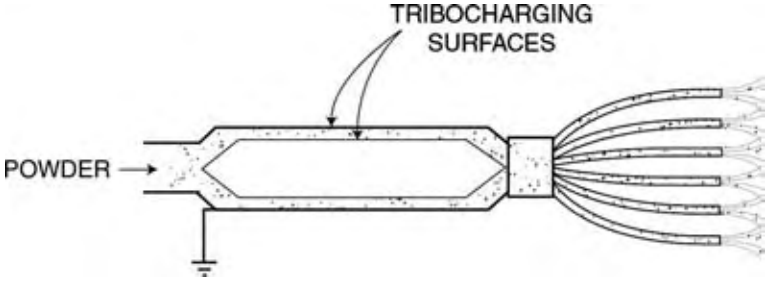


Fig. 5 Schematic of a tribocharging gun.

used together with a conventional corona gun for touch-up of difficult-to-coat areas, because the charged powder has the same negative polarity as a corona gun. However, it also suffers from the same disadvantages as conventional tribocharging in that the charging is sensitive to the type of powders that can be used. As with conventional tribocharging, most powder suppliers have learned how to optimize the powder formulation for negative tribocharging.

Nozzles

After the powder becomes charged, whether by corona, internal charging, or tribocharging, the mixture of air and powder particles is shaped into a spray pattern. The air velocity and the nozzle influence the spray pattern. Because corona charging uses a high voltage electrode, an electric field between the gun and the grounded part causes the charged powder and ions to follow the field lines to the part. If the part geometry is complex, electric field lines will not penetrate inside recesses and corners making it difficult for powder to penetrate these areas due to the Faraday Cage effect described earlier.

Flat spray (also called fan spray) nozzles can partially overcome this effect by forcing powder into recesses. However, if the velocity is too high, it can also blow powder off the part. Fig. 6 shows a perspective of a flat spray nozzle. The electrode is slightly recessed inside the nozzle and the powder emerging from the slot somewhat shields the electrode from the grounded workpiece. These factors combine to reduce the electric field between the gun and the part. This nozzle generates a directed fan spray pattern that has a pattern width of approximately 15 cm (6 in.) to more than 30 cm (12 in.). The flat spray nozzle is ideal for large flat areas. This nozzle is not recommended for wire goods because of its high powder velocity.

Conical nozzles use deflectors to generate a cone-shaped pattern as shown in the schematic in Fig. 7. Pattern widths can vary from as small as about 2.5 cm (1 in.) to approximately 60 cm (2 ft) in diameter, depending on the size and the radius of the deflector. The size of the pattern also depends on the powder flow rate. Conical nozzles generate a ''softer'' spray pattern because the deflector creates a partial vacuum that ''pulls'' the inner part of the cone pattern toward itself. This reduces the forward velocity of the powder. Conical nozzles are usually used in manual coating operations, where they are used for large flat parts and complex shapes.

Electrostatic forces dominate when the charged powder comes within about 10 mm (<0.5 in.) of the grounded part. As the conical nozzle has a lower forward velocity, the powder cloud tends to hover in front of the target, allowing the charged powder particles to be attracted to the grounded surface. In addition, the powder from a conical nozzle has a longer residence time in the charging zone allowing it to pick up more charge. In contrast, the flat spray nozzle is more directional, allowing aerodynamic forces to dominate making it easier for the powder to penetrate recessed areas. For both nozzles, the best transfer efficiency is achieved by good dispersion of the powder in air, reducing the velocity and the powder flow rate as low as possible for the specific application. Some powder guns swirl the air to control the velocity of powder, which improves particle charging.

Nozzles used for tribocharging can be very flexible, as tribocharging does not depend on a high voltage electrode at the front of the gun. The powder exiting the gun can be split into different streams and shaped in a multitude of ways allowing the powder to conform to the part geometry. The velocity of the powder from each ''finger'' of the nozzle is relatively low, allowing the gun to be closer to the part resulting in uniform

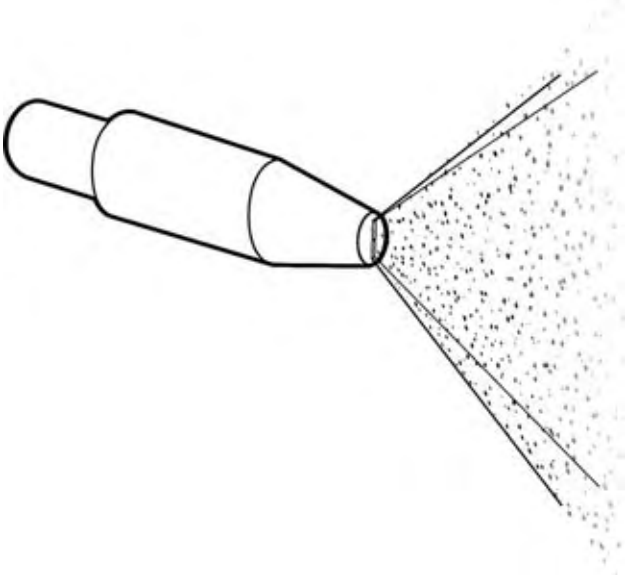


Fig. 6 Schematic of a flat spray (fan spray) nozzle.

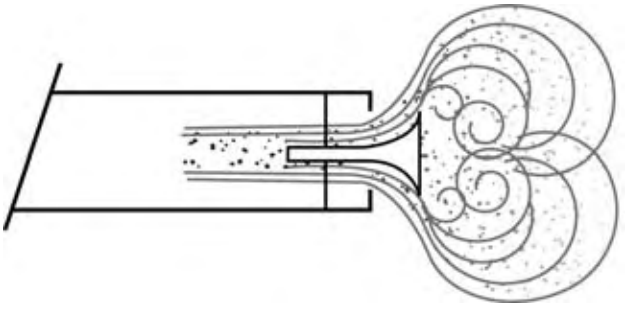


Fig. 7 Schematic of a conical nozzle.

powder deposition. Parts that can be coated using this technique are, for example, radiators, transformers, bicycles frames, or wire goods.

Bells and Discs

Powder can also be applied by bells or discs, similar to that used in the liquid paint industry. A powder bell uses a rotating turbine that rotates an enclosed bell head. The powder is delivered to the head where centrifugal forces disperse and atomize the powder. The particles are charged by corona either from the high voltage at the edge of the bell or from the external electrodes. Shaping air around the head controls the spray pattern and the forward velocity of the particles. The powder bell delivers high powder output uniformly over a wide flat area, making it ideal for the automotive industry.

Discs use large conical-type deflectors usually mounted in the vertical downward position so that powder is deflected in a horizontal position in a 360° radius. An omega (Ω) loop conveyor system normally surrounds the powder disc that is usually mounted on an oscillator or reciprocator for vertical up-and-down motion. The powder can be charged by a corona edge or by tribocharging guns. Typical applications for the powder disc are vertical extrusions or wire goods.

Electrostatic Fluidized Bed Technology

Powder application technology began with fluidized beds and evolved to electrostatic fluidized beds that are still in use today. Electrostatic fluidized beds are a method of powder application that does not use guns to apply the powder. Dry, compressed air passes through the charging media that has electrodes at a high voltage as shown in Fig. 8. This causes the air to become ionized. The ionized air is used to fluidize the powder and causes the powder to become charged. An electric field is created between the grounded part above and the charged powder cloud resulting in the powder being attracted to the part. As the grounded part becomes coated, additional charged powder will have difficulty in adhering to the coated surface due to back-ionization discussed earlier. The film thickness is controlled by voltage and exposure time.

POWDER RECYCLING

Particle Size

Any discussion of recycling powder must include information about particle size. As powders are similar to

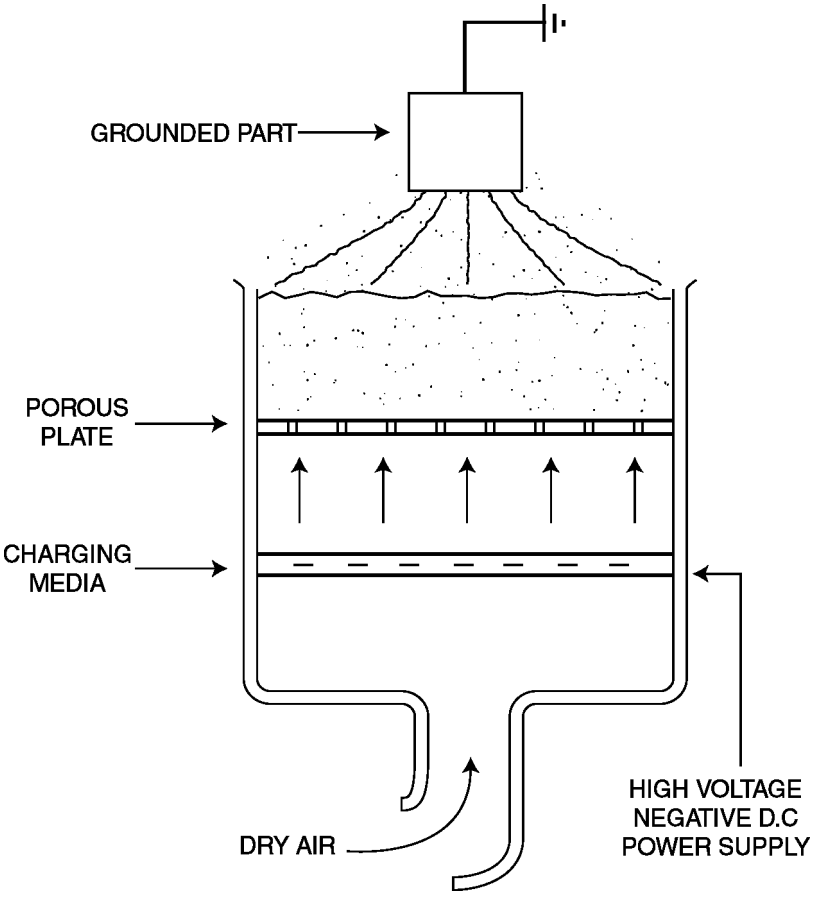


Fig. 8 Schematic of an electrostatic fluidized bed.

gravel-shaped particles, their size is not easily defined. There are many ways to define particle size, as well as different techniques that can give different results. An excellent review of this subject is presented by Rawle.[14] Low angle laser light scattering or commonly referred to as "light scattering" or "laser diffraction" is used extensively in the powder coating industry. Assumptions used in the older instruments that use the Fraunhofer approximation can make it difficult to compare the results between different instruments; the algorithms used to calculate particle size can vary depending on the type and model instrument that is used. Recently, particle size instruments have been developed that are based on Mie scattering theory, which gives more accurate results, as long as the refractive index of the powder and medium are known.

Laser diffraction methods generate a volume mean distribution. For laser light scattering techniques, the median particle size, commonly referred to as $D(v, 0.5)$, is defined as the value that divides half of the volume distribution, i.e., 50% of the distribution by volume is above this value and 50% is below this value. It is also common to define the volume percentage that is either above or below certain particle sizes. For example, particles below 10 μ agglomerate due to van der Waals forces causing fluidization and pumping problems. Therefore, it would be important to know the volume percentage of particles less than 10 μ. Coarse particles can also be a problem, especially when trying to obtain film uniformity; so it would be prudent to know the volume percentage of particles greater than, for example, 100 μ. Knowledge of these percentages helps to define the particle size distribution used.

Powder Recovery

Fig. 9 shows a simplified schematic of how powder is recycled in a system. After the powder is charged, the mixture of air and powder particles is sprayed onto a part. The powder that does not adhere to the part is contained in the spray booth and can be recycled. The oversprayed powder and the air mixture are drawn into the collector, where either a cartridge filter system or a cyclone (not shown) is used to separate the powder from the air. The collected powder is then transferred to the feed hopper where it is sieved (not shown) and remixed with virgin powder and pumped to the guns. Sieving the powder removes foreign particulates and conditions the powder prior to spraying. The particle size distribution in a cartridge filter system usually shifts toward finer particles. In a cyclone system the fines are scalped off, shifting the particle size distribution to the coarser side.

A model was developed that accurately predicts how particle size changes in a recovery system.[15] The basis of the model is that each particle size in a distribution has a certain probability of depositing onto the part. Probability factors for each particle size can be determined by comparing the particle size distribution of virgin powder to the powder that has been oversprayed one time. Factors calculated for one cycle are used to calculate the particle size distribution of subsequent cycles. In addition to these factors, a transfer efficiency value is needed to determine how much virgin powder needs to be added to maintain a constant level in the system. The particle size results calculated using this model were in good agreement with those measured from a production run of primer surfacer for exterior vehicle application. The model also showed

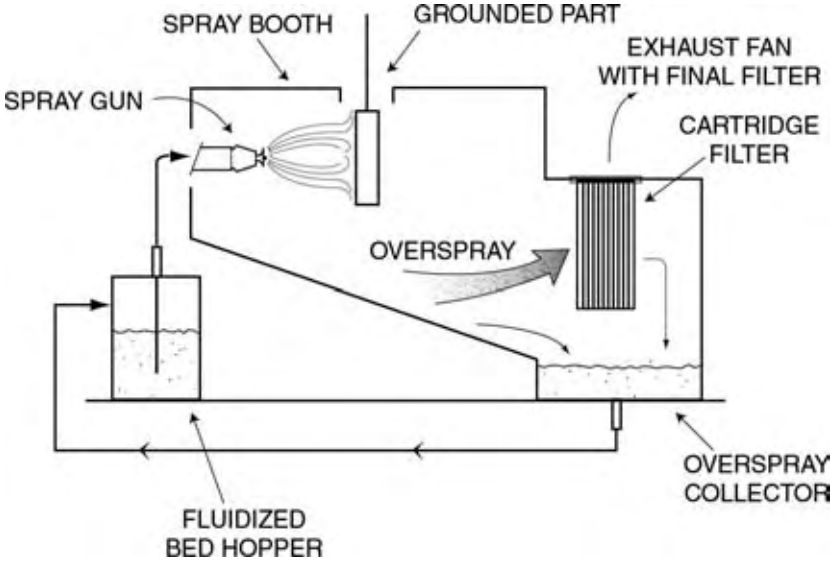


Fig. 9 Schematic of a powder recycling system.

that the median particle size in that system achieved steady state after only four complete cycles.

MEETING NEW CHALLENGES

New Equipment Technology

Recent developments in equipment technology are addressing some of the challenges of powder coatings such as fast color change and improved transfer efficiency. New powder feed technology has been developed[16,17] that allows greater powder loading at slower powder velocity to the spray guns using much less compressed air than a standard venturi pump. The advantages of this new technology are reduced powder consumption for better transfer efficiency using less energy, gentler transport of powder with less wear on the powder and the equipment, constant high delivery of powder over longer distances, and faster color change.

Fast color change developments have recently been made that utilize a two-canister system to apply color-specific powders for automotive applications.[18] While one canister is used for powder coating, the other one is getting ready for the next job. This allows a continuous operation with 8 sec claimed between jobs. Another recent development[19] consists of a color-changing manifold that uses automated valves to control color change. Powders from the different feed hoppers are attached to the manifold. The powder of choice is fed into the manifold that consists of a central channel. Air is used to purge the hose and block between colors, which allows for a color change of usually less than 8 sec.[10]

New Powder Technology

Much development work has produced powder coatings for decorative niche markets. They are now available in a wide range of gloss levels, colors, and specialized surface effects such as metallics, pearlescent, fluorescent, antique, hammertone, textured, wrinkled, and clearcoats. Metallic powder coatings have presented an application challenge, because some types of metallics are dry blended. Application equipment manufacturers have learnt how to apply these challenging materials.[20]

A major disadvantage of powder coatings has been their high cure temperature requirement. A great deal of progress has been accomplished over the past several years to reduce the cure temperature of powder. Conventional powders typically have cure temperatures between 160°C and 220°C (320–430°F) for about 20–30 min. These high cure temperatures and long cure times have prevented powder coatings from being used for plastics, wood, and medium-density fiberboard products, or metals with heat-sensitive parts such as low melting solder joints. Recent developments in curing technology have produced powders that can be melted at much lower temperatures of 100–140°C (212–285°F) for powder flow-out followed by curing with UV light.[21] Recent advances also include antimicrobial and high-temperature resistant coatings, improvements in weatherability, resistance to loss of gloss, improved mar- and scratch-resistance, and superdurable coatings. Nanotechnology will gradually play an increasing role as new powder coating compositions are developed having improved properties.[22]

CONCLUSIONS

An overview of powder coating application technology that includes the basics of powder charging has been described. From early developments, powder coating and powder application equipment technology have also been examined. Application methods have made great strides from its early beginnings using fluidized bed technology to provide thick coatings for corrosion protection and insulation. Today, the emphasis is on faster color change, higher transfer efficiency, and in some cases spray-to-waste. Powder coatings have evolved into the decorative markets that include specialty surface textures and appearances, antimicrobial powders, scratch-resistant, heat-resistant and low-cure temperature coatings. Application equipment has kept pace with new powder introductions. Improvements in powder coatings and application technology have steadily eliminated many of the drawbacks of using powder coatings. It is clear that equipment technology, as well as powder coating technology will continue to evolve in new directions creating new opportunities.

ACKNOWLEDGMENT

My sincere appreciation to Leah Lader for beautifully creating all the line drawings for the figures.

REFERENCES

1. Harris, S., Ed.; *Focus on Powder Coatings, Chemquest Powder Market Survey*; Elsevier Advanced Technology; 2004; Sept, 6–7.
2. Liberto, N.P., Ed.; *Powder Coating—The Complete Finisher's Handbook*, 2nd Ed.; The Powder Coating Institute: Alexandria, VA, 1999.

3. Cross, J.A. Electrification of solids and liquids. In *Electrostatics—Principles, Problems, and Applications*; Adam Hilger by IOP Publishing Ltd.: Bristol, England, 1987; 47.
4. Cross, J.A. Electrification of solids and liquids. In *Electrostatics—Principles, Problems, and Applications*; Adam Hilger by IOP Publishing Ltd.: Bristol, England, 1987; 51–52.
5. Cross, J.A. Electrification of solids and liquids. In *Electrostatics—Principles, Problems, and Applications*; Adam Hilger by IOP Publishing Ltd.: Bristol, England, 1987; 52.
6. Hughes, J.F. Electrostatic effects. In *Electrostatic Particle Charging—Industrial and Health Care Applications*; John Wiley & Sons, Inc.: New York, 1997; 30.
7. Hughes, J.F.; Ting, Y.C. Surface disruption phenomena in electrostatically deposited powder layers. In *Proceedings of the 12th Annual Meeting of the IEEE*; Industrial Applications Society: Los Angeles, CA, 1977; IEEE; 35-G, 906–909.
8. Campbell, D.H. Powder charging techniques. In *Symposium on Powder Coatings*; Birmingham, England, Apr 4–5, 1995; Paint Research Association, 1995.
9. Biris, A.S.; Mazumder, M.K.; Sims, R.A.; Yurteri, C.U.; Farmer, S.; Snodgrass, J. The effect of ring electrodes attachment to a corona gun on control of free ion concentration and back corona for improving powder paint appearance. IEEE Trans. Ind. Appl. **2003**, *39* (6), 1614–1621.
10. Milojevic, K. Multi-color powder application with internal charging guns: spray pattern and thickness uniformity control. In *21st International Conference on Automobile Body Finishing SURCAR*, Cannes, June 26–27, 2003; 165–174.
11. Börner, G.; Nienburg, H.-C.; Wittmann, J.; Böhme, H. Powder-Spraying Appliance. US Patent 6,254,684, July 3, 2001.
12. Teflon® is a registered trademark of DuPont.
13. Rehman, W.R.; Lader, H.J.; Messerly, J.W. Unipolarity Powder Coating Systems Including Improved Tribocharging and Corona Guns. US Patent 6,645,300, November 11, 2003.
14. Rawle, A. The importance of particle sizing to the coatings industry. Part 1: particle size measurement. Adv. Colour Sci. Technol. **2002**, *5* (1), 1–12.
15. Lader, H.J. Particle size modeling of powder paint in a recovery system. In *Powder Coating '94 Proceedings*; Powder Coating Institute, 1994; 254–267.
16. Moser, J. Soft spray: a new powder coating technology. JOT/Oberflaeche **2003**, *43* (10), 54.
17. Nordson signs licensing agreement with Borger. Finishing 2004, February.
18. GM honors top innovators of 2003. Powder Coating **2004**, *15* (6), 10.
19. Attinoto, R.A.; Ciarelli, G.J.; Koster, M.L.; Milojevic, D.K.; Rennie, C.M. Powder Paint Color Changer. US Patent 6,589,342, July 8, 2003.
20. Binder, J.; Lader, H.; Schroeder, J. Making the world a brighter place. Product Finishing 2003, (April), 64–68.
21. Udding-Louwrier, S.; Witte, F.M.; Sjoerd de Jong, E. Radiation curable powder coating system. Product Finishing (April), **1997**, 26–30; Buysens, K.; Tielemans, M.; Randoux, T. Radiation-curable coatings: a variety of technologies for a variety of applications. Surf. Coat. Int.—Part A **2003**, *2003* (05) 179–186; Christianson, R. Powder-coated wood OEM has bold ambitions. Wood and Wood Products 2004, July 1.
22. Anderson, L.G.; Barkac, K.A.; DeSaw, S.A.; Hartman, M.E.; Hayes, D.E.; Hockswender, T.R.; Kuster, K.L.; Nakajima, M.; Olson, K.G.; Sadvary, R.J.; Simpson, D.A.; Tyebjee, S.; Truman, F.W. Coating Compositions Having Improved Scratch Resistance, Coated Substrates and Methods Related Thereto. US Patent 6,803,408, October 12, 2004.

Power Factor

Peter R. Pujadó
UOP LLC, Des Plaines, Illinois, U.S.A.

INTRODUCTION

In an alternating current (AC) circuit, the power factor is the ratio of the real power delivered to the apparent power supplied, with the apparent power supplied being defined as the product of intensity times voltage. Therefore, the apparent power supplied is measured in ''volt-amps,'' while the real power delivered is measured in ''watts.''

The power factor arises from the dephasing of the electromotive force or voltage, in volts, and the current, in amperes. This dephasing results from the interaction of inductances and/or capacitances together with the resistance of the circuit. Because direct current is unaffected by inductances or capacitances, the power factor, cos φ, for direct current is always one, provided that the direct current is free of fluctuations. Similarly, if the circuit consists of resistance only, and is free of both inductances and capacitances, the power factor for AC is also one—this would be the case, for example, in the tungsten (wolfram, W) filament of a light bulb that has negligible inductance characteristics

DISCUSSION

The power factor, cos φ, is defined as

$$\cos\varphi = \frac{\text{watts}}{E^{*}I} = \frac{\text{watts}}{\text{volts}^{*}\text{amperes}} \quad (1)$$

This number can never be greater than one. Usually, it is given as a percentage, with 100% being the maximum theoretical power deliverable.

To some extent, the power factor is of more importance to the electric supply company than to the consumer. The consumer pays for the watts actually consumed, not the volt-amps supplied. A low power factor of, say, 50% is almost tantamount to saying that 50% of the energy supplied is returned to the supplier, thus increasing the supply and distribution costs. Because of this, consumers with low power factors may be required to raise them or to incur additional charges from the suppliers.

In the U.S.A., the Energy Policy Act (EPAct) of 1992 became effective on October 24, 1997, with some provisions delayed until October 24, 1999. The intent of the law was to reduce the rate of energy consumption in the U.S.A. by requiring the use of energy-efficient products. In order to accomplish this goal, the EPAct mandates that most industrial AC motors used in the U.S.A., whether imported or manufactured locally, must meet some minimum energy efficiency standards as defined in Table 12-10 of the National Electrical Manufacturers' Association (NEMA) Standard MG 1. The overall energy efficiency of a motor is the combination of its electrical and mechanical efficiencies.

In general, the EPAct regulations apply to all general-purpose motors that meet the following specifications (http://www.nema.org):

- single speed, polyphase T frame
- 1–200 horsepower (HP)
- 3600, 1800, and 1200 rpm
- foot mounted
- squirrel-cage induction motors, NEMA design A and B
- continuous rated
- operating at 230/460 V, constant 60 Hz line power (see NEMA Standard MG1)

	Nominal full-load efficiencies					
	Open motors			Enclosed motors		
HP	2 pole	4 pole	6 pole	2 pole	4 pole	6 pole
1.0	—	82.5	80.0	75.5	82.5	80.0
1.5	82.5	84.0	84.0	82.5	84.0	85.5
2.0	84.0	84.0	85.5	84.0	84.0	86.5
3.0	84.0	86.5	86.5	85.5	87.5	87.5
5.0	85.5	87.5	87.5	87.5	87.5	87.5
7.5	87.5	88.5	88.5	88.5	89.5	89.5
10.0	88.5	89.5	90.2	89.5	89.5	89.5
15.0	89.5	91.0	90.2	90.2	91.0	90.2
20.0	90.2	91.0	91.0	90.2	91.0	90.2
25.0	91.0	91.7	91.7	91.0	92.4	91.7
30.0	91.0	92.4	92.4	91.0	92.4	91.7
40.0	91.7	93.0	93.0	91.7	93.0	93.0
50.0	92.4	93.0	93.0	92.4	93.0	93.0
60.0	93.0	93.6	93.6	93.0	93.6	93.6
75.0	93.0	94.1	93.6	93.0	94.1	93.6
100.0	93.0	94.1	94.1	93.6	94.5	94.1
125.0	93.6	94.5	94.1	94.5	94.5	94.1
150.0	93.6	95.0	94.5	94.5	95.0	95.0
200.0	94.5	95.0	94.5	95.0	95.0	95.0

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007925

Other motors do not fall under the EPAct regulations. for example, footless designs, other horsepower ranges, 900 rpm and slower speeds, two-speed versions, 50 Hz, 200/400 and 500 V, NEMA design C and D types, U frame, single-phase motors, frame sizes 56 and smaller, and motors designed for inverter use only; covered motors may be used with inverters and still fall under the provisions of the EPAct. Underwriters' Laboratory listed motors may be handled on a case-by-case basis.

While motors are available that have significantly higher efficiencies than those specified by the EP Act, many existing industrial motors have much lower efficiencies than required.

In addition to the above, NEMA designates a special class of Premium Efficiency Electric Motors (2003) for three-phase induction motors applied to municipal or industrial applications for operation on voltages of 600 V or less, rated 500 HP or less, operating more than 2000 hr/yr at more than 75% of full load. For details, please refer to NEMA MG 1-1998 Table 12-12. For each HP rating, that table defines nominal and minimum efficiencies for 2-, 4-, and 6-pole open motors.

PURE RESISTANCE

The general term for the combined effect of resistance, inductance, and capacitance is impedance. Briefly, we shall describe each of the contributions and their net effect.

Resistance is the obstacle that a direct electric current has to overcome to circulate through a conductor under the driving force of the voltage or electromagnetic force. In its simple form, it can be defined as:

$$\text{Resistance} = R = \frac{E}{I} = \frac{\text{volts}}{\text{amperes}} = \text{ohms} \qquad (2)$$

RESISTANCE AND INDUCTANCE IN SERIES

Any changes or oscillations in an electric current generate a counterforce that opposes such changes. By definition, an AC changes continuously, with both the intensity and the voltage varying according to sinusoidal functions

$$\begin{aligned} I &= I_o \sin\theta \\ V &= V_o \sin\theta \end{aligned} \qquad (3)$$

If the AC flows through a short straight wire, any such counterforces are negligible. However, if the AC flows through a coil, the effects can be very significant. A coil, or any similar device, is said to generate "self-induction" or, in brief, it creates a circuit with inductance. Electric motors, for example, are wired through numerous coils with high inductance. Thus, electric motors, if not corrected, can have very low power factors.

In the presence of inductance we can still think of a reactance similar to the resistance, but now defined as:

$$\text{Reactance} = X_L = \frac{E}{I} = \frac{\text{volts}}{\text{amperes}} = \text{ohms} \qquad (4)$$

where

$$X_L = 2\pi f L \qquad (5)$$

with f being the frequency of the AC in hertz or cycles per second (typically 60 in the U.S.A. and 50 in Europe) and L being the inductance in henrys. L is a property of the circuit and depends on the coil (diameter and number of turns) and on the core material inside the coil. A circuit has an inductance of 1 henry when a variation of 1 ampere per second generates an electromotive force of 1 V. Correctly speaking, X_L is the "inductive reactance."

If the coil consists of a wire with negligible resistance, $R \sim 0$, the intensity I is delayed 90° with respect to the voltage V. This can be explained in terms of an induced counterforce proportional to the rate of variation in intensity.

$$E = L\frac{dI}{dt} + I^*R \qquad (6)$$

such that, if $R = 0$ and I varies according to $I = I_o \sin(2\pi ft)$, then

$$\begin{aligned} E &= L(dI/dt) = 2\pi f L I_o \cos(2\pi ft) \\ &= 2\pi f L\ I_o \sin(2\pi ft + \pi/2) \end{aligned} \qquad (7)$$

namely, E is advanced 90° with respect to the intensity, I.

$$E = 2\pi f L I = X_L I \quad \text{with} \quad X_L = 2\pi f L \qquad (8)$$

In such a circuit with pure inductance, $\cos\varphi = \cos 90° = 0$, such that the power supplied to the coil is stored in the electromagnetic field and is returned when the field disappears, with no net power delivery.

If the resistance, R, is not negligible, the wire can be regarded as consisting of a resistance and an inductance in series, and a more general impedance can be defined as follows:

$$\text{Impedance} = Z = \frac{E}{I} = \frac{\text{volts}}{\text{amperes}} = \text{ohms} \qquad (9)$$

where $Z = [R^2 + X_L{}^2]^{1/2}$, or the square root of the sum of the squares of resistance and inductive reactance, respectively.

In this case, the intensity I is delayed an angle φ with respect to the voltage, such that

$$\varphi = arc\tan(X_L/R) \tag{10}$$

The effective power delivered is $EI\cos\varphi$, with $\cos\varphi$ being the power factor.

If $R \sim 0$, $\varphi \sim 90^\circ$ as before, and the effective power delivered is zero.

If $X_L \sim 0$, $\varphi = 0^\circ$, and the effective power delivered is $E \times I$.

RESISTANCE AND CAPACITANCE IN SERIES

If a circuit in which AC flows was to consist of a pure capacitance, the intensity would be advanced with respect to the voltage by an angle of 90°, or the opposite of the effect of the inductance seen above.

The capacity of a circuit is defined as $C = Q/E$, with C in farads, Q being the electrical charge in coulombs, and E the voltage. One farad is the capacity of a capacitor or a circuit in which one coulomb of electrical charge creates a voltage difference of 1 V.

In the presence of capacitance we can define a capacitive reactance as follows:

$$\text{Reactance} = X_c = \frac{E}{I} = \frac{\text{volts}}{\text{amperes}} = \text{ohms} \tag{11}$$

where

$$X_c = \frac{1}{2\pi f C} \tag{12}$$

with f being the frequency of the AC and C the capacity of circuit measured in farads.

Following the same reasoning we used for inductance, if $E = E_o \sin(2\pi ft)$, we obtain:

$$\begin{aligned} I &= \frac{dQ}{dt} = C\frac{dE}{dt} = 2\pi f C E_o \cos(2\pi ft) \\ &= 2\pi f C E_o \sin(2\pi ft + \pi/2) \end{aligned} \tag{13}$$

With pure capacitance, the intensity is advanced by $\varphi = 90^\circ$ with respect to the voltage. Therefore, the net power delivered is zero. In this circuit,

$$E = X_c I \quad \text{with} \quad X_c = \frac{1}{2\pi f C} \tag{14}$$

If the circuit can be regarded as consisting of resistance and capacitance in series, then:

$$\text{Impedance} = Z = \frac{E}{I} = \frac{\text{volts}}{\text{amperes}} = \text{ohms} \tag{15}$$

where $Z = [R^2 + X_c{}^2]^{1/2}$, or the square root of the sum of the squares of resistance and capacitive reactance, respectively.

In this case,

$$\varphi = arc\tan(X_C/R) \tag{16}$$

RESISTANCE, INDUCTANCE, AND CAPACITANCE IN SERIES

We can equally define

$$\text{impedance} = Z = \frac{E}{I} = \frac{\text{volts}}{\text{amperes}} = \text{ohms} \tag{17}$$

where $Z = [R^2 + (X_L - X_C)^2]^{1/2}$, and $\varphi = arc\tan[(X_L - X_C)/R]$

In the resulting configuration, the intensity may be advanced or delayed with respect to the voltage depending on whether X_C is greater than or less than X_L.

If $X_C = X_L$, the intensity is at its maximum E/R, and the circuit is in resonance. Note that $X_C = X_L$ implies that

$$2\pi f L = \frac{1}{2\pi f C} \tag{18}$$

so that

$$f = \frac{1}{2\pi(L C)^{1/2}} \tag{19}$$

is the resonant frequency or the natural frequency of the circuit. If R is low, resonant circuits can lead to dangerously high internal voltages across the coils and capacitors, much higher than the externally applied voltage. A circuit with resistance, impedance, and capacitance in series can lead to hazardous situations if it is not properly monitored to avoid the onset of resonance.

OTHER CIRCUITS

Using the same analogy as resistance, any other impedance configurations can be structured with, for example, resistance, inductance, and capacitance all in parallel; or various combinations of impedances in series and in parallel. The analysis is analogous to that of resistances with direct current. For impedances in series, we start with the vector of intensity and locate the voltage vectors advanced or delayed depending on the type of impedance. For impedances in parallel,

we start with the voltage vector and advance or delay the intensity vectors as required.

Note that, although we indicated above that a short straight wire can be regarded as consisting solely of resistance, this is no longer true for long power lines. Apart from the power factor at the generating station, the magnetic field produced by each conductor will affect the other conductors and, in effect, create an inductance. The analysis is different for monophase or triphase transport systems but, in general, the inductive reactance will be a function of the distance between cables, the frequency of the AC, and the gauge and resistance of the cables. Significant power losses can arise in long AC power lines.

HOW TO IMPROVE THE POWER FACTOR

We have seen above that the dephasing angle φ between intensity and voltage can be modified by using various combinations of inductive and capacitive reactance.

The most usual situation that leads to low power factors is the use of inductive motors that, as the name implies, can introduce very large inductive reactance in the line. The electric load introduced by an inductive motor can be represented as a resistance and an inductance in series. This combination will have an intensity vector that will be delayed with respect to the voltage.

If we add a capacitor in parallel to the motor, we are imposing an intensity vector that will be advanced 90° with respect to the voltage. By using the right size capacitor we can decrease the angle by which the overall intensity is delayed with respect to the voltage and, therefore, increase $\cos\varphi$. A capacitor with only a few micro-farads (μF) usually suffices to increase the power factor from, say, 75% to approximately 85% or higher.

For example, let us assume that we wish to improve the power factor from $\cos\varphi_1$ to $\cos\varphi_2$, where $\cos\varphi_2 > \cos\varphi_1$ in a circuit that delivers W W, at E V, and at a frequency of f Hz (cycles/second)

In Fig. 1, I is the initial current in amperes, equivalent to the vector length $\underline{oa}$. Addition of a capacitor in parallel will create an intensity vector I_C advanced 90° with respect to the voltage, E. In order to calculate the new angle φ_2, all we need to do is set the distance $\underline{ab} = I_C$.

$$I = \frac{W}{E^* \cos\varphi_1} \tag{20}$$

$$\underline{ac} = I\sin\varphi_1$$
$$\underline{oc} = I\cos\varphi_1$$
$$\underline{bc} = \underline{oc}\tan\varphi_2$$

$$I_C = \underline{ab} = \underline{ac} - \underline{bc} = E/X_c = 2\pi f C E \tag{21}$$

$$\begin{aligned} C &= \frac{I_C}{2\pi f E} = I\frac{\sin\varphi_1 - \cos\varphi_1\tan\varphi_2}{2\pi f E} \\ &= W\frac{\tan\varphi_1 - \tan\varphi_2}{2\pi f E^2} \end{aligned} \tag{22}$$

The corrected line intensity will now be

$$I_L = \frac{W}{E^*\cos\varphi_2} = I\frac{\cos\varphi_1}{\cos\varphi_2} < I \tag{23}$$

As an example, if $W = 10$ kW, $E = 480$ V, $f = 60$ Hz, and $\cos\varphi_1 = 0.70$, and we wish to improve the power factor to $\cos\varphi_2 = 0.90$, we should add a capacitor in parallel to the load with a capacity $C = 61.7$ μF (micro-farads). The line current will decrease from 29.8 A to 23.1 A.

It should be remembered that both the efficiency and the power factor of a typical motor are a function of the load. Motors should be sized for optimum operation at 100% load on a continuous operation basis. At 100% load, the efficiency of a good motor in the 10–30 HP range will typically be somewhere between 89% and 92% (up to 93–94% for a NEMA premium motor). If the load is decreased to, say, 50%, the efficiency of the same motor may easily drop to 85% or lower.

If the load with low power factor were capacitive, we can achieve a similar correction of the power factor by adding an inductance in parallel to the load.

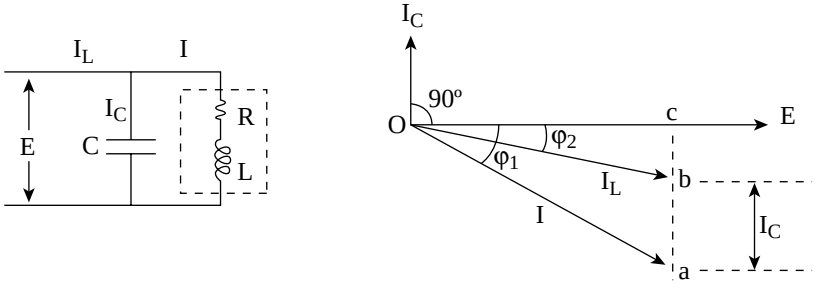


Fig. 1 Correction of the power factor.

CONCLUSIONS

The power factor, or potency factor, is a natural result of the intrinsic nature of AC and its interaction with self-induced magnetic fields. It can be corrected and minimized but it cannot be eliminated altogether, principally when subject to varying loads. Use of direct current may be a preferred solution when the dissipative losses from AC may become prohibitive.

BIBLIOGRAPHY

Bitter, F. *Currents, Fields, and Particles*; MIT Technology Press and John Wiley & Sons, 1956.

Dorf, R.C. *The Electrical Engineering Handbook*, 2nd Ed.; CRC Press, 1999.

Dorf, R.C. *CRC Handbook of Engineering Tables*. Electrical Engineering Handbook Series, CRC Press, 2003.

Fink, D.G.; Beaty, H.W.; Eds. *Standard Handbook for Electrical Engineers*, 14th Ed.; McGraw-Hill, 1999.

Gray, A.; Wallace, G.A. *Principles and Practice of Electrical Engineering*, 7th Ed.; McGraw-Hill Book Company, 1956.

http://www.nema.org.

http://www.motorsmatter.org.

http://www.energv.wsu.edu/cfdocs/mmdownload/mmdnld nmc2.cfm.

Seely, S. *RadioElectronics, Electrical and electronic engineering series*, McGraw-Hill Book Company, 1956.

Yarbrough, R.B. *Electrical Engineering Reference Manual*, 5th Ed.; Professional Publications, 1994.

Pressure-Relief Valve Design

Jonathan Francis
University of Central Lancashire, Preston, U.K.

INTRODUCTION

Protection against overpressure is essential for the safe operation of fluid power and process plant. Automatic pressure-relief valves are commonly used for this purpose. They work on the principle of a force balance. When the valve is shut the hydrostatic force tending to open it varies linearly with system gauge pressure, and the return force tending to keep the valve closed is (very nearly) a constant. If system pressure is too high then the hydrostatic force will surmount the return force opening the valve and venting the system.

This entry describes the design of relief valves. Engineering design of these protection devices is dominated by function, but is also influenced by manufacturing cost and assisted by consideration of scale and shape. The main aim of this entry is to describe the function, performance requirements, and control dynamics of spring-loaded automatic pressure-relief valves. A secondary aim is to discuss modern engineering design solutions and some relatively recent theoretical and practical developments affecting relief valve design.

TYPES OF RELIEF VALVE

Early Relief Valves

Fuller describes four types of automatic relief valve. Each differs by means of the return force applied, as illustrated in Fig. 1. The four types are:[1]

1. Mass lever
2. Tension spring lever
3. Deadweight direct load
4. Compression spring direct load.

The first valves were of the mass–lever type. Tension spring lever and deadweight direct load valves were later developed to overcome weaknesses inherent in the mass lever design: a calibrated tension spring provides evidence of set pressure; the deadweight direct load valve improved stability and reseating because of reduced tilt. However, an appropriate valve size was not easily determined until James Joule provided evidence of the energic nature of heat in the mid-19th century, to dispel the myth of ''phlogiston'' (a supposedly weightless fluid previously blamed for boiler explosions).

Emergence of the Direct-Acting Spring-Loaded Relief Valve

Also in the mid-19th century, the basic Cornish boiler was replaced as state-of-the-art by the Lancashire boiler. A design feature of the latter allowed subsequent passage of combustion products beneath and around the water drum, after leaving the fire tube (passing through the drum). Thermal efficiency of boilers was further improved by a number of later designs, such as the Cochrane vertical boiler, the economic boiler, and the package boiler. However, the greatest leap forward in terms of steam pressure and steaming rate came with the introduction of the water-tube boiler. With their small-diameter water tubes, these could operate at higher fluid pressures and provide a large heating surface to water volume ratio.

Deadweight and mass lever valves are impractical for use on high-pressure systems, being too heavy and cumbersome. Tension spring valves tilt more easily and for an equivalent lift of the valve disk, spring movement in a tension spring lever valve must be greater than that in a compression spring direct-load valve. The benefits of being lightweight and compact are reflected in the low cost of a compression spring direct-loaded valve, making it the most common for protection of pressurized systems.

Terminology

A primary aspect of relief valve design is to provide an increase of fluid force on the disk to raise it against the spring. When this is achieved without an increase of reservoir pressure, but as a result solely of the changing flow structure in a lifting valve, it is known as ''lift-assistance.'' Without lift-assistance, valves will open smoothly and linearly with overpressure. It is the lift-assistance that causes the classic popping action of valves used with compressible fluids (i.e., sudden and rapid opening).

Design for lift-assistance is associated with pressure recovery downstream of the valve seal. Generally, two approaches are adopted: 1) a valve arrangement with huddling chamber as shown in Fig. 2(A) and 2) a valve arrangement with a deflector plate as shown in Fig. 2(B). Both approaches require extra disk area outside the sealing area on which the recovered pressure may

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007929

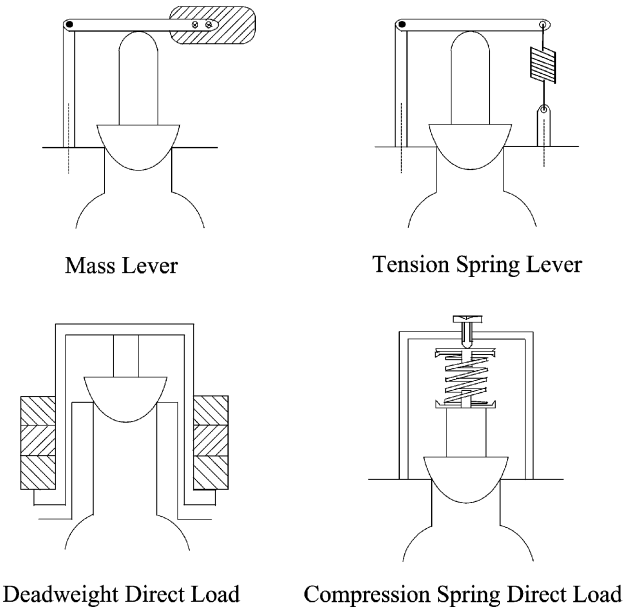


Fig. 1 Schematic illustrating four early types of automatic pressure-relief valve.

act to drive the disk upward. With the disk extended beyond the seal radius, a huddling chamber forms a gradually enlarging flow area immediately downstream of the seal to assist in pressure recovery. The addition of a deflector plate causes a dynamic pressure on the outer portion of the disk because of a change of momentum beneath it, but the volume between the deflector plate, body base, and seat support can also be considered a huddling chamber.

In the United Kingdom, relief valves were traditionally categorized by the degree of lift necessary to achieve the rated discharge. The division into ordinary-lift, high-lift, and full-lift categories is a throwback to BS 759 (obsolescent).[2] The subsequent standard BS 6759 and the recently introduced harmonized European standard BS EN ISO 4126-1 do not make any such distinction.[3,4] Ordinary-lift valves achieved their maximum disk lift at less than 1/12 the bore diameter and were generally used on liquid service. High-lift valves would attain lifts up to one-fourth the bore diameter and full-lift valves over one-fourth, making them suitable for gas or vapor service.

Modern terminology is derived from the United States, where valves lacking appreciable lift assistance, but which open smoothly to rated discharge at a low disk elevation (i.e., ordinary-lift valves), are referred to as "relief valves" or sometimes "pressure-relief valves." Valves with lift-assistance may be termed "safety valves" or "safety-relief valves," where the former might be used to describe a valve lifting to less than one-fourth the bore diameter (i.e., a high-lift valve) and the latter to describe valves achieving higher lifts (i.e., full-lift valves). A consistent use of terminology is no longer prevalent throughout either North America or Europe. In the International Standard, the phrase "safety valve" is used as a generic term to describe all types and classes of valves that open automatically to relieve excessive pressure.[4]

DOMINANT DESIGN CONSIDERATIONS

This entry focuses on the design of the product itself. The phenomenon of explosion (that it is meant to prevent) and the selection, sizing, or application of

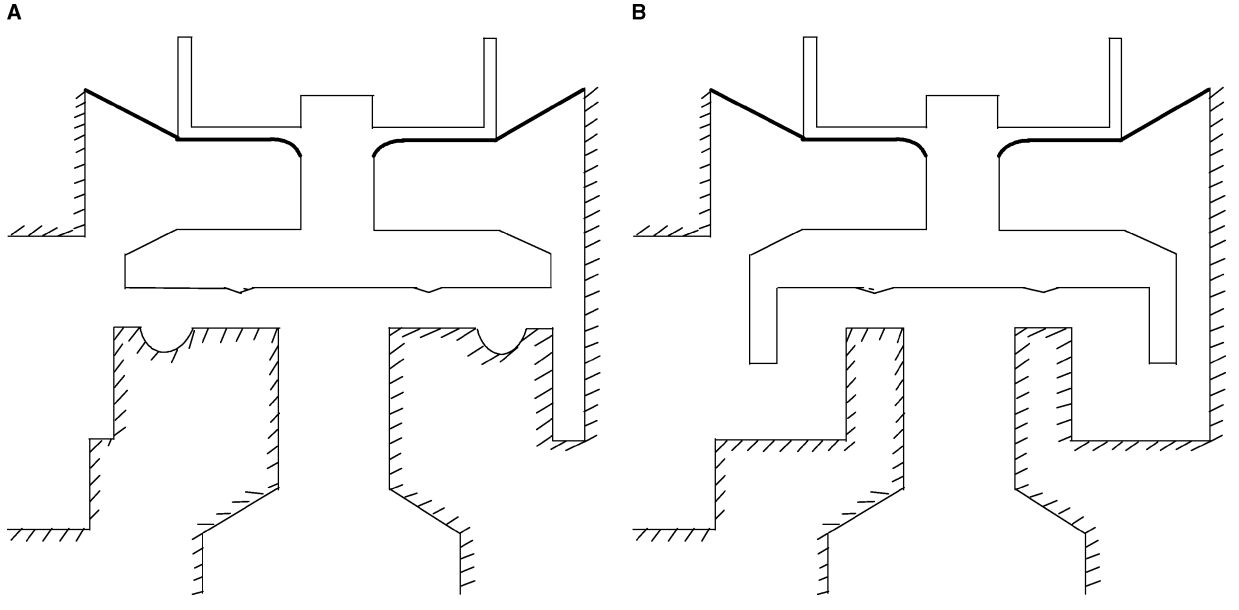


Fig. 2 (A) Illustrative valve arrangement with huddling chamber. (B) Illustrative valve arrangement with deflector plate/skirt.

the device are worthy of entire books in their own right. Readers are directed to alternative sources for information concerning the dynamics of reacting flows, design for fire safety, pressure vessel design, and bursting disk design.[5–12] They are also advised to study relevant standards and manufacturer's literature for information concerning sizing, type testing of relief valves, independent witness approval, statutory requirements, and application to potable water and flammable atmospheres.[4,13–15]

Valve Shape and Manufacture

Valve shape is constrained by its function. An entrance and an exit port are needed to pass fluid from an overpressurized system to a place of safe discharge. Between the two ports, a barrier is needed to prevent discharge under normal operating conditions (i.e., a disk and a seat) and a means to ensure that the barrier remains leakproof (i.e., a seal and hence a spring to provide a clamping force). There must be facility for the barrier/disk to move to allow passage of fluid in an emergency; and this movement must be initiated by system pressure.

The aim is to achieve functionality with minimum product cost. Generally, product cost can be maintained as low as possible by minimizing both material volume and machining of parts. Fig. 3 shows a sectional sketch of a prototype high-lift valve. The main body, moving parts, and flow passages are shown in full. The thick black line is a diaphragm that prevents fluid entering the spring chamber. This chamber is foreshortened at the top of the figure and its actual height significantly affects material and production cost. The shorter is the spring chamber, the less is the material needed and the cheaper is the bonnet casting. The height of the bonnet is controlled by spring wire thickness and pitch, as design lift must not be compromised by prior achievement of full coil compression. Spring design is beyond the scope of this entry, but note that the piston into which the spring would fit (spring not shown) is not wholly within the bonnet. The piston sits on the spindle collar and disk within the main body of the valve to reduce the necessary bonnet height; and body height cannot be reduced without compromising escape of fluid from the backpressure space above the disk.

Valve bodies are generally cast. The thickness of a valve wall for use on medium pressure systems (1–10 bar) is generally well in excess of that needed to meet pressure test standards. However, reduction in material volume for thinner walls would be more than offset by the additional difficulties of casting thin walls with thicker seat supports or screwed fixings. Ease of assembly also reduces cost.

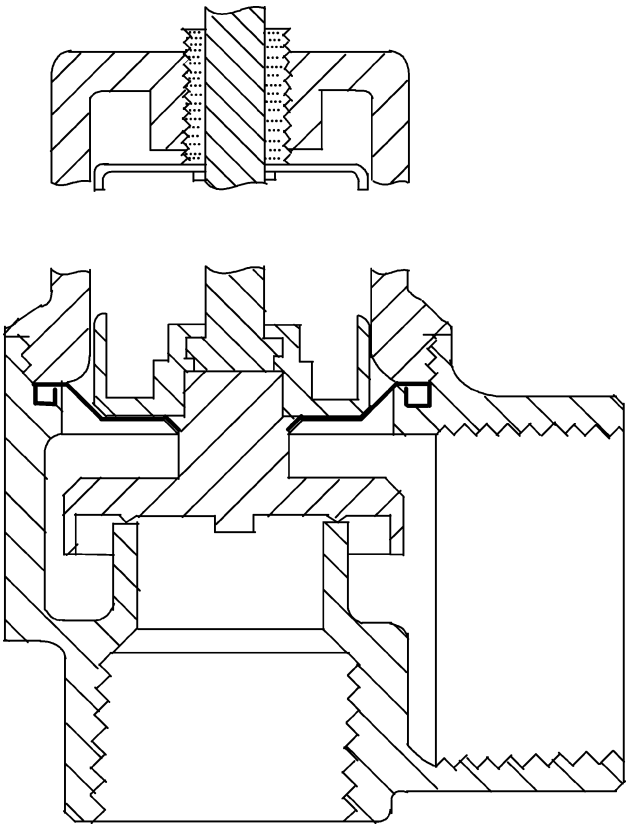


Fig. 3 Sectional sketch of a spring-loaded pressure-relief valve (spring not shown).

Careful examination of Fig. 3 reveals that the disk and its spindle connector would need to be made in two parts to enable the assembly of this prototype. A slightly larger access where the bonnet connects would enable a composite disk and spindle to be inserted from above. However, the spindle needs to be able to pivot within the disk mounting to enable the three-dimensional fluid forces to tilt the disk without causing excessive friction in the spindle-lift guides. Nevertheless, assembly would be simpler where the disk is inserted from above, and any reduction in diameter (and hence material) confined to the spring chamber.

Valves should be installed vertically as shown in Fig. 3. Care must be taken that the internal diameter of the pipe fixing is equal to or greater than the valve bore and the connector is short enough to limit pressure loss between the reservoir and bore to 3% of set pressure.[16] If the pipe-fixing diameter is too small then the valve may not discharge at full capacity. Moreover, instability may be induced because of excessive pressure drop in the pipe connector. Rapid and repetitive rise and fall of a valve disk is referred to as valve "chatter." It can also be induced by misapplication of a valve intended for gas service on liquid service.

Good design can be compromised by poor application or size selection. If a valve passes too little fluid

then the system will not be adequately protected. Conversely, if a valve is oversized then its disk may lift only a short distance and continually pass fluid. This is known as ''simmering.''

Sometimes the valve disk does not reseat properly, but continues to discharge a small quantity of fluid. This is known as ''feathering'' of the valve, and it can be caused by damage to the seal from particulate matter suspended in the relieving fluid. It can also be due to balance of those forces that control the functioning of the valve, which can be checked using the techniques described in the next section.

RELIEF VALVE FUNCTION

Valve Operation

The operating cycle

Fig. 4 illustrates an operating cycle. When system pressure is raised above working pressure, the seal relaxes under the influence of a gradually reduced clamping force. At a preset pressure, leakage past the seal is initiated; this is denoted by point ''a'' in Fig. 4. Further increase in pressure will cause the disk and seal to be lifted clear of the seat (e.g., point ''b''). In an ordinary-lift valve the initially smooth opening is sufficient to relieve excess pressure, so that once the system fault is rectified the relief valve will ideally shut by retracing the line to point a. In practice, friction in the spindle-lift guides will cause the valve to close at a lower pressure. In a high-lift or full-lift valve, the initial opening is not sufficient to relieve the overpressure, so that the valve continues to lift beyond point b. If the pressure is raised sufficiently then the disk–spring force system will become unstable (point ''c''), and the valve will open rapidly to point ''d.''

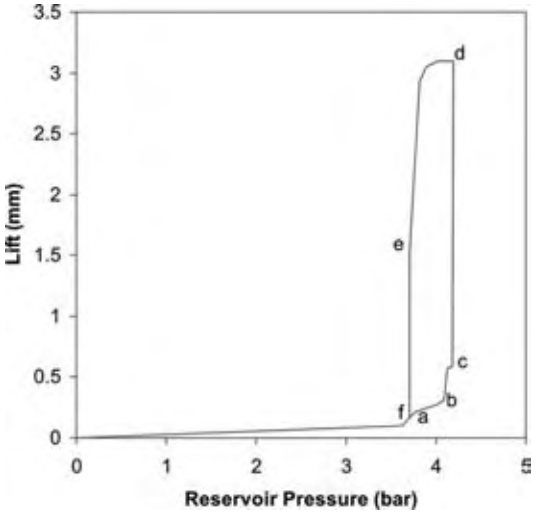


Fig. 4 Example of the operating cycle of a pressure-relief valve. *(View this art in color at www.dekker.com.)*

With pressure relief from point d, a gradual reduction in pressure and lift is seen until at point ''e'' the disk–spring force system once again becomes unstable, and the valve rapidly closes (point ''f''). Soft seals (such as neoprene) are desirable at low pressures because they distort more easily to form a seal. Hard seals (such as ptfe or viton) are needed to withstand the reaction forces at the seat on high-pressure systems, but do not so easily distort to prevent flow when slightly damaged. On some high-pressure applications, metal to metal seating is used because even hard sealing materials are prone to damage (either from the reaction forces or from the fluids in the system).

The pressure at which the valve begins to pass fluid is termed the ''set pressure'' and the pressure at which it opens rapidly to its highest lift is called the ''popping pressure.'' The difference between them is called the ''accumulation'' and the difference between set pressure and reseating pressure is called ''blowdown.'' Technical standards provide limitations for accumulation and blowdown and for most of the 20th century the designer's aim was to meet these standards, but recently focus has shifted so that more valves are being developed with the intention of minimizing fluid loss (or waste, see later).

Cycle synthesis

The foregoing operating cycle is controlled by the balance of forces acting on the valve disk. The force due to the fluid surrounding the disk and the force due to the spring compression are the main components in the force balance. As a first approximation, all other forces can be ignored. Fig. 5 is typical of the way in which the opposing forces of the fluid and spring vary with lift when the disk is artificially raised from its seat at constant pressure. Curves of fluid force can be identified for notable pressures in the operating cycle and compared with spring forces. Neglecting the mass of moving parts, the spring force on the point of lifting is equal to the product of sealing area and gauge reservoir pressure.

The cycle of Fig. 4 can be deduced from the graph of Fig. 5. Intersections a–f on Fig. 4 correspond to points a–f in Fig. 5.

The valve begins to leak at some pressure ($P1$), which is set by precompression of the spring. However, the spring force remains greater than the fluid force (at $P1$) for an initial lift, which explains why the valve does not pop open immediately on leaking. A gradual lift can be observed (say $L2$) for further small increases in

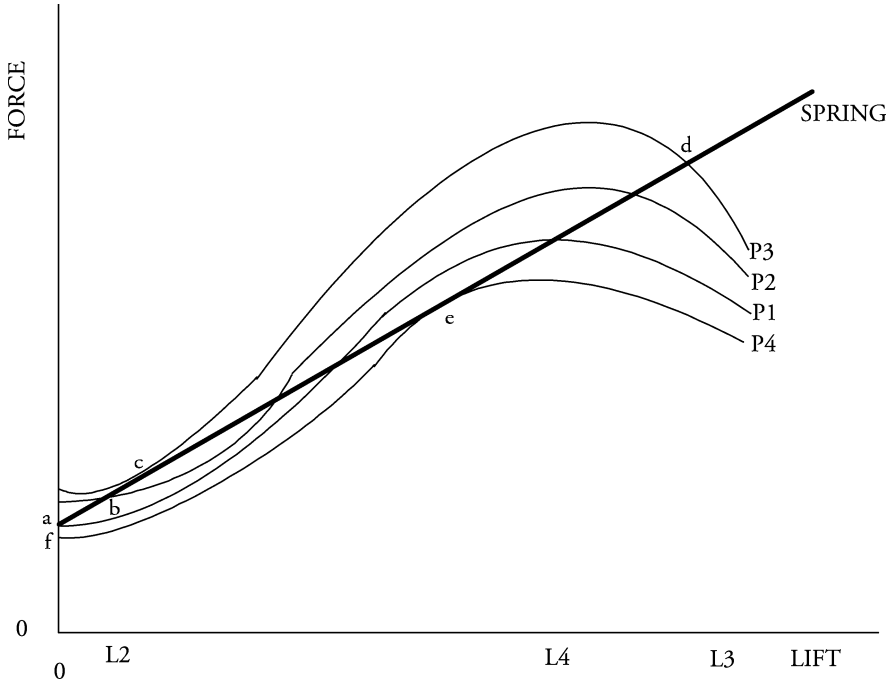


Fig. 5 Variation of spring and fluid forces with lift.

reservoir pressure (e.g., $P2$). However, when a particular pressure is reached ($P3$) the rate at which the fluid force increases with disk lift becomes greater than the spring rate, and the system becomes unstable. A further infinitesimal increase in pressure causes the valve to open quickly until the forces are once again in balance at a new position ($L3$).

With pressure relief, the disk elevation gradually reduces until another point of instability ($L4$) is reached, at which the valve suddenly closes.

A typical lifting transient has been found to be in the order of 50 msec. [17,18] In this time, the disk of a 20 mm nominal bore full-lift valve will move approximately 5 mm, while a particle of air travelling at sonic velocity would move in excess of 15 m. Even at a Mach number of 0.1 these rough figures indicate that a quasi-steady assumption is reasonable, as Bicen, Vafidis, and Whitelaw also discovered for internal combustion engine valves.[19]

It is known that the opening part of the operating cycle can be synthesized in the manner just described. The maximum lift position and popping pressure can be reliably predicted from force analysis or experimentally derived force plots.[20–22]

Force Characteristics

Force plots like those of Fig. 5 can be generated from steady forces measured in the absence of a spring. The spring is removed from the valve so that the (fluid induced) upward thrust can be measured with a cantilever load cell. The load cell is used as a lift stop, initially set at a low lift but then adjusted to gradually higher positions. Prestressed mild steel spindles should also be substituted for soft metal spindles, and designed so that the upward force on the piston reacts on a collar of the new spindle. This helps to reduce buckling and direct compression errors on supposed lift position during testing.

Individual plots need not be generated for all possible set pressures and valve sizes that a valve series is intended to cover. Instead, a force characteristic can be defined in which dimensionless force (or force factor) is plotted against dimensionless lift.

The moving parts of a valve are the disk, piston, and spindle. Fig. 6A shows a cutaway section of a general purpose relief valve when closed and Fig. 6B shows the forces experienced by the moving parts during the lifting transient.

$F1$ is the vertical component of the combined hydrostatic and hydrodynamic (fluid) force acting on the moving parts arising from the absolute pressure distribution and denoted by the shading on the wetted surfaces in Fig. 6B. $F2$ is the weight of the moving parts, $F3$ is the spring compression force, and $F4$ is the weight of the spring. $F5$ is the atmospheric pressure acting over the top of the piston. Neglecting friction in the guides, $F6$ is the viscous damping due to movement of fluid into and out of the space above the disk (and for a few valves, though not the one illustrated here, also due to an additional dashpot). $F7$ is due to the accelerating mass of moving parts and $F8$ is the clamping force or reaction at the seat in a closed valve.

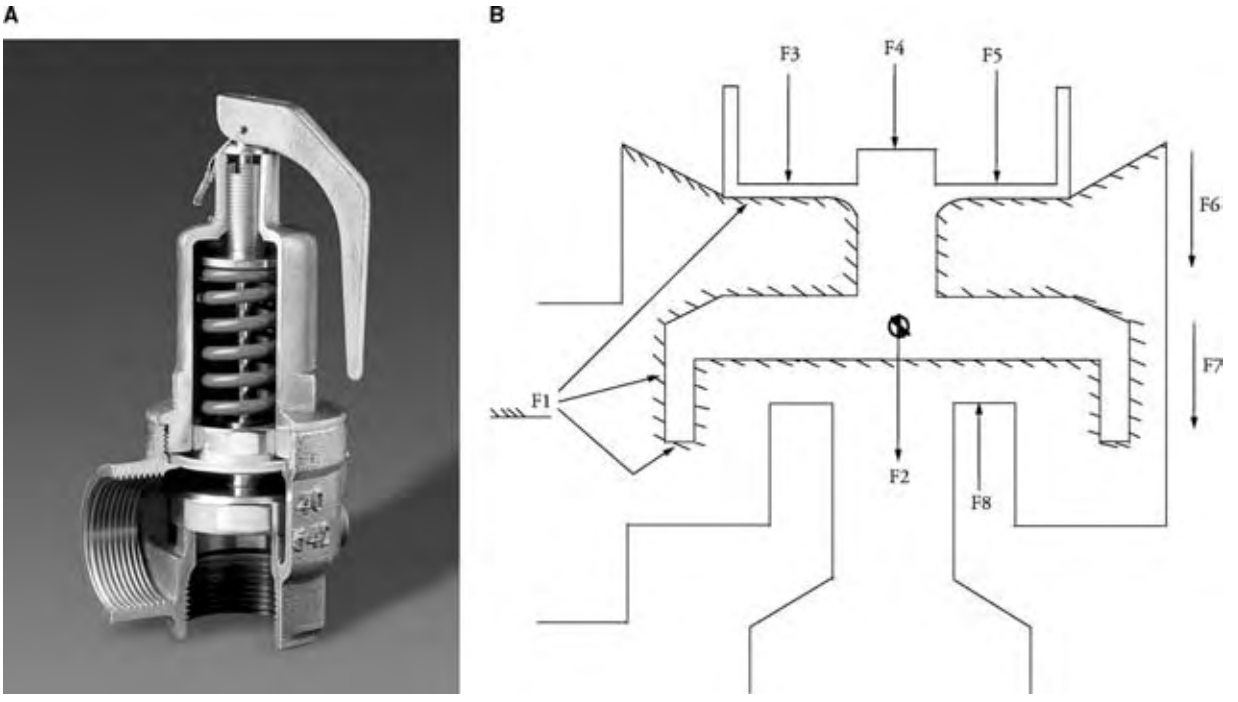


Fig. 6 (A) Cutaway view of a general-purpose relief valve. (Reproduced by kind permission of Delta Fluid Products Ltd.) (B) Forces acting on the disk, piston, and diaphragm assembly. *(View this art in color at www.dekker.com.)*

Under steady conditions in an open valve, $F6$, $F7$, and $F8$ all have a zero value, and

$$F1 = F2 + F3 + F4 + F5 \tag{1}$$

The valve spring is removed when testing with the load cell, which then bears the upward thrust ($F0$), and

$$F0 = F3 + F4 \tag{2}$$

and as F4 $\ll$ F3, it can be neglected. Thus, load cell thrust $F0$ is virtually the equal of spring compression force in a complete valve.

Spring compression cannot be easily measured, but if a touch-tight seal is assumed, then just prior to the disk rising clear of the seat, the spring compression force will be equal to the hydrostatic resultant on the disk, i.e.,

$$F3_{(X=0)} = (P_o - P_a)\frac{\pi}{4}D_s^2 \tag{3}$$

where $P_o - P_a$ is the gauge reservoir pressure and D_s is the diameter (not of the bore, but of the seal) and $X = 0$ represents a nil dimensionless lift.

At the point of lifting, the load cell force $F0$ will vary linearly with the reservoir gauge pressure. It will also vary linearly with the square of the sealing diameter (D_s), but bore diameter (D) is usually chosen as a nominal measure of valve size. The force factor can be defined as:

$$K_F = \frac{4xF0}{\pi(P_o - P_a)D^2} \tag{4}$$

A graph of the force factor (K_F) plotted against dimensionless lift (L/D) can be referred to as a "force characteristic." An example is shown in Fig. 7 for a particular valve tested at various pressure ratios on air service. For ideal flow, at pressure ratios below 0.528, the valve would be subject to choked flow and the pressure distribution on the moving parts, and hence the dimensionless force factor, would vary according to the position and strength of the shock wave downstream of the throat (see later), and hence according to the built-up backpressure value in any particular application. The design lift for the valve shown in Fig. 7 is $X = 0.167$, and within that range the results tend to collapse toward a single curve for lower pressure ratios. At excessive lifts, the diaphragm (protecting the spring chamber) can become trapped and impart inconsistent retarding forces on the disk.

The spring compression force can also be normalized by the hydrostatic resultant at $X = 0$, and the resulting line superimposed on the force characteristics, as shown in Fig. 8. Maximum lift occurs at the intersection of the force and spring characteristics. It is by this means that lift and popping pressure can be predicted across a range of valves and springs selected to achieve design lift. However, accuracy of these

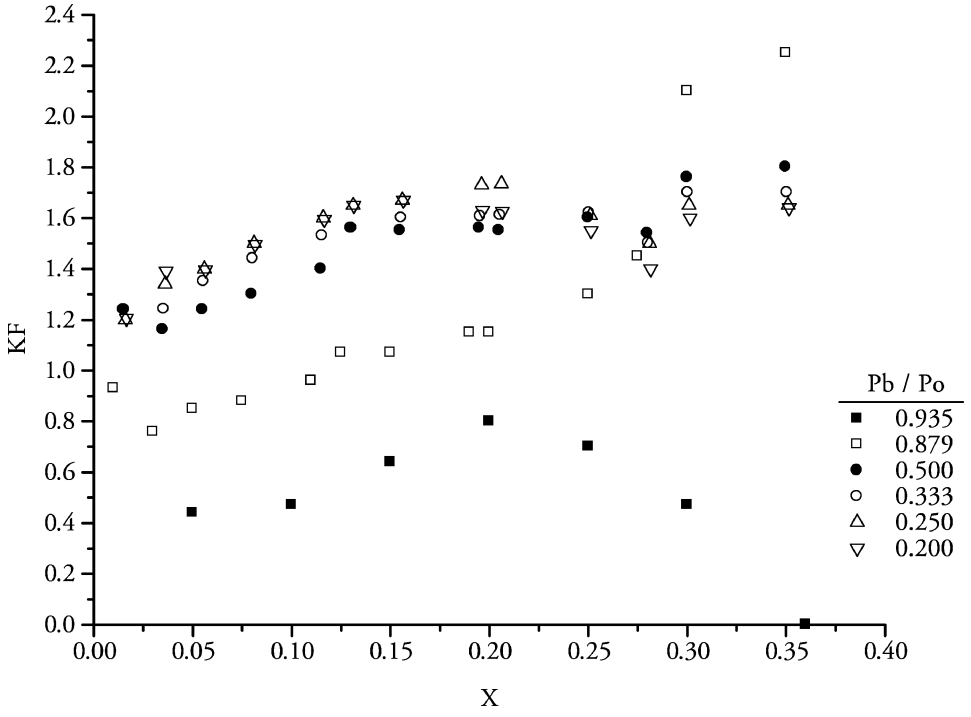


Fig. 7 Effect of pressure ratio on force characteristics of a high-lift relief valve.

predictions is limited by the degree of dynamic similarity that can be achieved, which is mainly dependent on geometrical similarity across the valve range.

Prediction of Blowdown

The compression spring direct loaded valve is reminiscent of a classic mass–spring–damper arrangement. As a result, there have been many publications purporting to simulate the dynamic response of spring-loaded relief valves using a linear second-order system, but few are compared with real valve activations.[23–25] Unfortunately, such simulations are prone to omitting important variables, especially when coupled with momentum models of fluid dynamics intended to predict the forcing function. In practice, the valve exhibits hysteresis, which enables it to open and close at different pressures. This is partly a result of Coulomb damping due to sliding friction and stiction in the valve guides, but may also be due to different flow structures beneath an opening and closing valve disk (i.e., a Coanda effect).

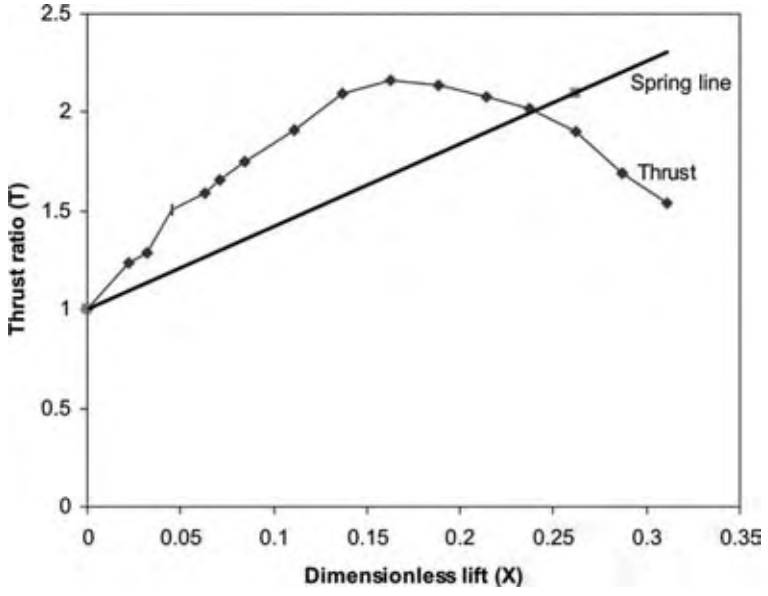


Fig. 8 Example of spring characteristic intersecting the force characteristic. *(View this art in color at www.dekker.com.)*

It is difficult to account for hysteresis in a second-order model because of its nonlinearity with viscous damping (disk movement in the relieving fluid), its lack of repeatability, and its dependence on built-up backpressure (and hence the particular application). However, where additional viscous damping is deliberately added, a linearized analysis will yield useful results.[26–28]

Hysteresis also prevents force plots generated for a gradually lifting valve from being used to synthesize the closing action of a valve; and so predict blowdown. Nevertheless, the shape of the force curve generated for a gradually closing valve is clearly related to blowdown (from Fig. 5); and the two force plots will be of similar shape where any Coanda effect is minor. It has been demonstrated that the shallower is the rise of the force curve above the spring line, the more difficult is lift prediction, but the smaller is the blowdown; a shape factor can then be used to empirically predict blowdown.[18]

VALVE FLOW

Dimensional Analysis

Flow-through relief valves have traditionally been analyzed by treating the valve as a convergent nozzle.[17,29,30] A discharge coefficient (k_d or k_m) is defined as the ratio of actual mass flow rate to the theoretical ideal mass flow rate through a one-dimensional convergent nozzle of exit area equal to the nominal bore of the tested valve. Measured discharge coefficients are derated by 10% (K_{dr}) to specify valve capacity.

According to Ref.[4] where the bore is treated as a nozzle of area (A) (mm^2), reservoir and backpressures (P_o and P_b, respectively) (bar) (abs.), and density of fluid (ρ) (kg/m^3), then valve capacity (Q_m) (kg/hr) can be found for liquids:

$$Q_m = 1.61 K_{dr} K_v A[\rho(P_o - P_a)] \tag{5}$$

where K_v is a correction factor for viscosity, usually assumed to be unity.

For compressible flows, two conditions exist: critical and subcritical flow. Neglecting friction, subcritical flow occurs when:

$$P_b/P_o > [2/(\gamma + 1)]^{\gamma/(\gamma - 1)} \tag{6}$$

where γ is the isentropic index at the reservoir conditions, or ratio of specific heats. Critical or choked compressible flow occurs when Eq. (6) is untrue. In the latter condition, the valve capacity can be found:[4]

$$Q_m = 0.2883 K_{dr} C A[\rho P_o]^{0.5} \tag{7}$$

where C is a complex function of the isentropic index. If flow is subcritical then the equation becomes:

$$Q_m = 0.2883 K_{dr} C A K_b[\rho P_o]^{0.5} \tag{8}$$

where K_b is a different complex function of the isentropic index. These complex functions can be found in Ref.[4] and their development can be traced from gas dynamics texts.[4,31,32]

However, the end of the inlet bore is not the minimum section of a valve flow passage. Even in the full-lift condition, separation of a shear layer at the leading edge of the seat will often make the mainstream flow contract further as it passes over the seat. Flow through the bore and that across the seat have also been modeled as a convergent nozzle followed by one-dimensional frictional adiabatic ''Fanno'' flow between parallel plates with successful prediction of the lifting transient (from forces) as the result.[33] However, from Fig. 3 it can be seen that the flow path generally converges toward the seal, and diverges at greater radii downstream of the seal. Where the seat has any degree of thickness, or where a deflector skirt or a huddling chamber exist then the transition from the bore to the main body appears as a convergent–divergent passage.

A convergent–divergent nozzle is a useful analogy by which is developed the scientific basis of relief valve flow for compressible fluids.[22] For liquid service, a convergent nozzle is sometimes considered a more appropriate analogy, but pressures measured beneath the disk of a valve subject to both incompressible air flow and choked compressible air flow are in keeping with those expected from a converging–diverging geometry.[34–36]

Compressible mass flow rate (m) through a valve can be affected by the following variables:

1. Component shape and finish (λ)
2. Disk lift (L)
3. Valve nominal bore (D)
4. Absolute reservoir pressure (P_o)
5. Absolute exhaust pressure (P_b)
6. Atmospheric pressure (P_a)
7. Reservoir viscosity (μ)
8. Isentropic index (γ)
9. Reservoir density (ρ_o).

Use of the Buckingham pi theory leads to a functional relationship, such as:

$$\frac{m}{D^2\sqrt{P_o \rho_o}} = \Phi(\Theta) \tag{9a}$$

The independent groups (Θ) include dimensionless lift (L/D), backpressure ratio (P_b/P_o), Reynolds number

(Re), and the isentropic index or ratio of specific heats (γ). The dependent group relates the mass of discharge to the valve size and to the reservoir conditions. It is suitable for assessing dynamical similarity under conditions of choked compressible flow. Substitution of P_o/RT_o for ρ_o in the dependent group produces a more conventional representation:

$$\frac{m\sqrt{R\,T_o}}{D^2 P_o} = \Phi(\Theta) \tag{9b}$$

where R is the characteristic gas constant and T_o is the absolute reservoir temperature.

An ideal gas is completely defined by the gas constant (R) and isentropic index (γ). For any such gas a critical flow factor (ψ) can be found assuming one-dimensional isentropic mass flow (m^*) through a minimum section or "throat" (A^*):

$$m^* \frac{\sqrt{T_o}}{A^* P_o} = \psi \tag{10}$$

For air entering a valve at room temperature $R = 284\,\text{kJ/kg/K}$, $\gamma = 1.4$, and $\psi = 0.04042$. However, real relief valves suffer from area contraction due to separated flow patterns, and the flow is neither one-dimensional nor isentropic. Consequently, the flow rate (m) will differ from that in an ideal duct having the same minimum section. A general form of discharge coefficient can then be defined by

$$\frac{m}{m^*} = m\frac{\sqrt{T_o}}{\psi P_o A^*} \tag{11}$$

If the bore is assumed to be the minimum area then the discharge coefficient may be designated K_M, where

$$K_M = \frac{m\sqrt{T_o}}{\psi P_o(\pi/4) D^2} \tag{12}$$

The derated coefficient (K_{dr}) is then 90% of K_M.

Flow Characteristics

A graph of discharge coefficient (K_M) plotted against dimensionless lift (L/D) is known as a "flow characteristic." Fig. 9 shows an example of flow characteristics plotted for various pressure ratios relieving from an air receiver, where P_b is the total backpressure at the valve exit port and P_o the reservoir total pressure.[22] According to ideal one-dimensional compressible flow theory, choked flow should occur at pressure ratios below 0.528. In practice, choked flow will occur at lower pressure ratios because of friction losses upstream of the curtain area. Nevertheless, the curves in Fig. 9 are close-together at pressure ratios below 0.333, and have collapsed to almost a single curve for ratios less than 0.25. Other work shows a similar effect.[29,37,38] One curve can therefore be used to represent many pressures.

In the medium pressure range (1–10 bar) the Re is almost always above 4000 and below 400,000, so its effect on flow pattern is almost nonexistent, and consequently, it has a minor effect on dimensionless flow characteristics. If the valve range is truly geometrically similar, then the discharge coefficient will be unaffected

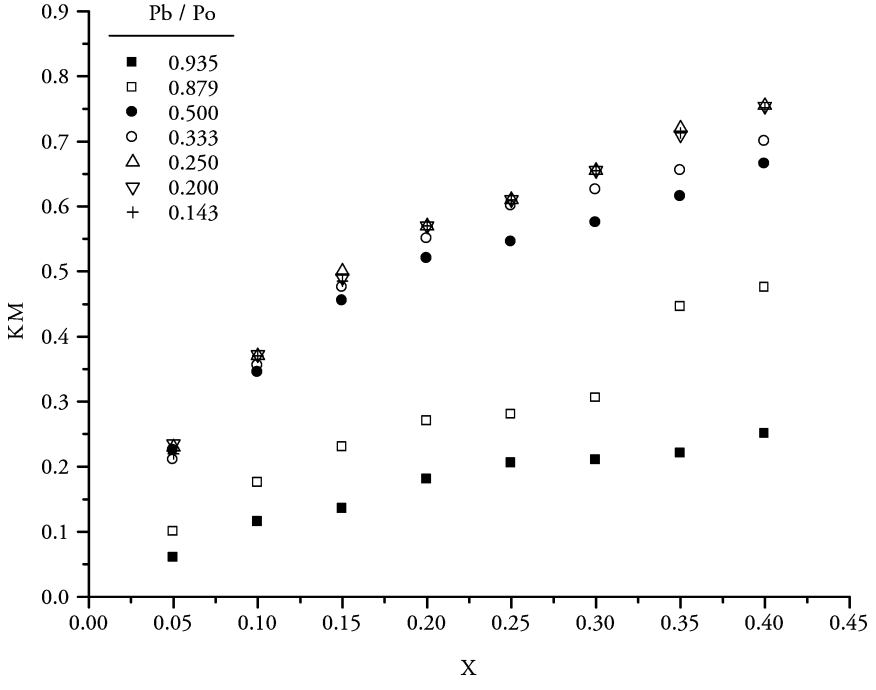


Fig. 9 Effect of pressure ratio on flow characteristics of a high-lift relief valve.

by changes in scale and in exhaust, atmospheric, or reservoir pressures. For a selected design (Θ) there remains only the one major variable: dimensionless lift (L/D).

Passage Shape

It has been known for many years that changes in slope and sudden discontinuities in flow characteristics of disk valves coincide with sudden alterations in separated flow patterns.[39] In relief valves, flow separates from the leading edge of the inlet bore and from the tip of the seal. As lift of a relief valve disk increases, the separated shear layer issuing from the leading edge of the seat reattaches further along the seat; and that from the seal tip reattaches to the disk closer to the skirt. By use of both computational fluid dynamics (CFD) software and oil-paint technique for incompressible flow at pressure ratio $P_b/P_o = 0.935$, Francis and Betts have shown that the apparent reduction of slope at midlift in relief valve flow characteristics (appearing in Fig. 9 between 0.10 and 0.15) coincides with the point at which reattachment to the rear edge of the seat ceases, so that at higher lifts the separated shear layer misses the rear edge by a considerable margin.[35] It was also found that reattachment to the disk ceased at a slightly lower lift.

A shock wave will exist downstream of the throat and is typically located where flow enters the huddling chamber. A sketched example is shown in Fig. 10. Evidence of shock wave location has been found from pressures measured on the disk surface.[36] Between the throat (usually close to the seal) and the shock wave is a supersonic region of low pressure that detracts from the general upward thrust. If the shock wave can be induced soon after the mainstream flow passes the seal then pressure recovery downstream of the shock wave can act on a greater area (and low pressure immediately upstream of the shock on a smaller area) to assist lifting. However, to avoid instability, the exit from the huddling chamber should not be so restricted as to form a second throat.

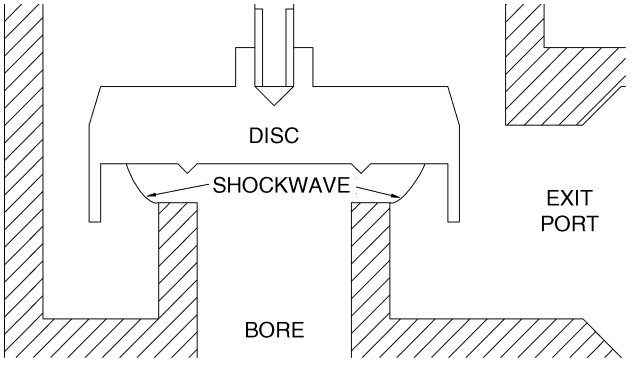


Fig. 10 Schematic showing likely position of shock wave downstream of the minimum section.

In fact, the gap at exit from the huddling chamber (or at the bottom of the skirt) is quite critical, as the pressure established in the huddling chamber significantly affects blowdown. This is why some valves have adjustable rings that can be moved up/down the outer surface of the seat support to adjust the size of gap at exit from the huddling chamber/skirt bottom edge (and so optimize performance). It is a balance between the desire to establish a high pressure to assist lifting and avoid too high a pressure that exacerbates blowdown.

MODERN DESIGN SOLUTIONS

Modern valve development is incremental. The techniques described in this entry are used to a greater or lesser extent by various designers, together with reliance on ''black art'' career expertise and prototyping by trial and error.

Theoretical Developments

Zonal models are not new in relief valve development. They are prone to oversimplification, but have been used successfully to approximate fluid forces on a disk and approximate the lifting transient.[30,40,41]

Fig. 11 reproduces the flow passage geometry of a relief valve. Note that the seal is modeled as a knife-edge. The fluid regions are split into three zones. Zone ''A'' is that region beneath the disk within the compass of the seal. Zone ''B'' is the remaining passage beneath the disk encompassed by the deflector skirt. Zone ''C'' is the backpressure region that affects both the piston and the disk top surface. In this model the skirt is presumed to be thin and thus the thrust from pressure acting on the lower surface of the skirt is lumped together with that from Zone B. Assuming that each zone can be represented by a pressure homogenous throughout the zone, then by analyzing the fluid mechanics and particularly the pressure losses through a valve, it is possible to provide a reasonable representation of upward thrust. A recent model has shown that despite lift and popping pressure being adequately predicted, the shape of the force characteristic (and hence blowdown) is not so well predicted.[41] Greater sophistication is required and this is now offered by field models.

Two areas of research that are ripe for making incremental improvements to knowledge of spring-loaded valves are: 1) use of CFD to identify beneficial changes to geometry and 2) CFD studies and theoretical or empirical modeling of two-phase flows to enable

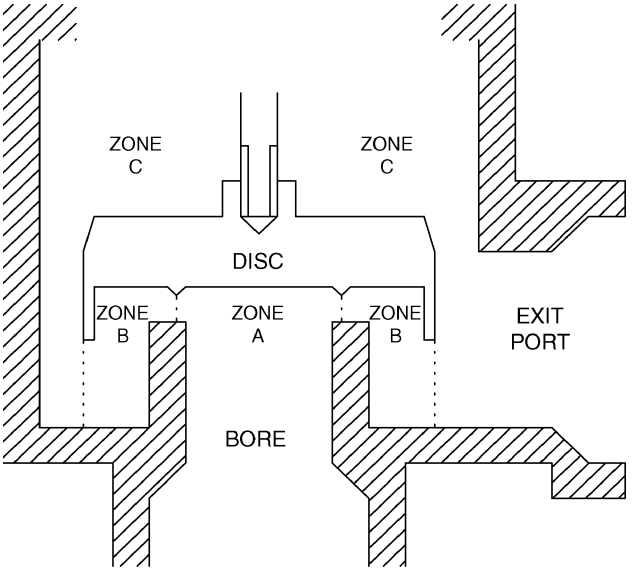


Fig. 11 Zones of homogenous pressure assumed in zone model of a high-lift valve.

appropriate sizing methods to be developed and applications skill to become more widespread. In the latter case, manufacturers often quote discharge capacities for single-phase flows but rarely for applications where two-phase flows can be expected.

Published use of CFD to analyze flow structure in relief valves is sparse and its potential for prototype simulation has not been realized.[35,42,43] However, the last 5–8 yr have seen improvements in meshing and processing capability of commercial solvers, as well as the turbulence models that can be selected, so that CFD is expected to be used more frequently in the development of engineering components. Some current (and early stage) work is aimed at improving the accumulation and blowdown characteristics (without blowdown rings or pilot operation), and investigating the potential for fluidic design within this context.[44]

Procedures for sizing relief valves for two-phase flows are still in the developmental stage but the potential safety advantages of promulgating sizing methods among suppliers make it a worthwhile pursuit.[45,46]

Practical Developments

Temperature relief valves for use on unvented hot-water systems are beyond the scope of this entry. Combined pressure and temperature relief valves became popular in the United Kingdom in the early 1990s, but these devices are merely pressure-relief valves with additional facility for actuation by a temperature-sensitive element. As the bore of a combined valve is not required to choke, the addition of a temperature-sensitive actuator passing through the bore does not meaningfully affect the valve operation or flow characteristic.

In the last 15 yr there has also been an increase in the use of full-bore valves with dual outlets. Full-bore valves are intended to ensure that the bore is subject to choked flow, maximizing the discharge capacity. Despite the outlet pipe connection usually being one nominal size up from the inlet connection, the flow passages within older designs of valve encouraged a buildup of backpressure above the disk by restricting flow toward the exit port. This tended to limit the valve lift during discharge. Bellows, used to protect the disk and spindle connector, also had the advantage of limiting the area of the disk top surface subjected to built-up backpressure. Moreover, a piston arrangement is still sometimes introduced to balance out the backpressure forces, as shown in Fig. 3, where any excess pressure acting on the upper surface of the disk also acts on the underside of the piston to further compress the spring that fits snugly into the piston. However, prevention of built-up backpressure is preferred to these methods of compensation. For example, by altering the shape of the disk top surface to increase the chamfer size, an improved flow and better k_{dr} were achieved by one manufacturer (S. Cooper, private communication). See Fig. 12 and compare with Fig. 6A.

Dual outlets are used to minimize built-up backpressure. They are usually used in high-pressure protection and where the maximum use of the bore is

Fig. 12 Cutaway view of a high-lift safety valve. (Reproduced by kind permission of Delta Fluid Products Ltd.) *(View this art in color at www.dekker.com.)*

desired without excessive restrictions downstream or where long discharge piping is needed to carry the fluid some distance to a safe collection point. Some dual outlet valves have opposing outlets to minimize the reaction forces that the valve mount needs to withstand. Others are actually twin valves mounted alongside each other on the same valve body (see Fig. 13).

An aspect of valve design that is now of growing importance is waste minimization. This is achieved through a combination of good sealing, closely controlled operating pressures, and correct sizing of valves. Good sealing is promoted by pilot operation.

Fig. 14 shows an example of a closed pilot-operated valve. Operation of the main valve is subject to operation of the pilot, which is itself a spring-loaded direct-acting relief valve, except that the valve is held open by the spring to allow system pressure to be bled to a chamber above the main valve disk. As the system pressure rises, so too does the pressure in the chamber above the main valve disk (and as this acts downward on a larger area than the main valve bore, it keeps the valve firmly clamped shut). The pilot valve is set to commence lifting at (say) 97% of the intended main valve set pressure. With a short travel, it reaches the seat and hence seals off the bleed line to the chamber above the main valve disk at (say) 98% of the set pressure. Further, increase in system pressure causes the pilot valve spindle to lift clear off its disk, opening a central drain to relieve pressure in the chamber above the main valve disk (see Fig. 15). Hence, the main valve lifts, and fully.

Fig. 13 Example of a twin valve mounted on the same bore supply. (Reproduced by kind permission of Delta Fluid Products Ltd.) *(View this art in color at www.dekker.com.)*

The advantages of this type of valve are considerable. Because the clamping force keeping the main valve shut is itself linearly dependent on system pressure, an increase in system pressure does not cause the seal to

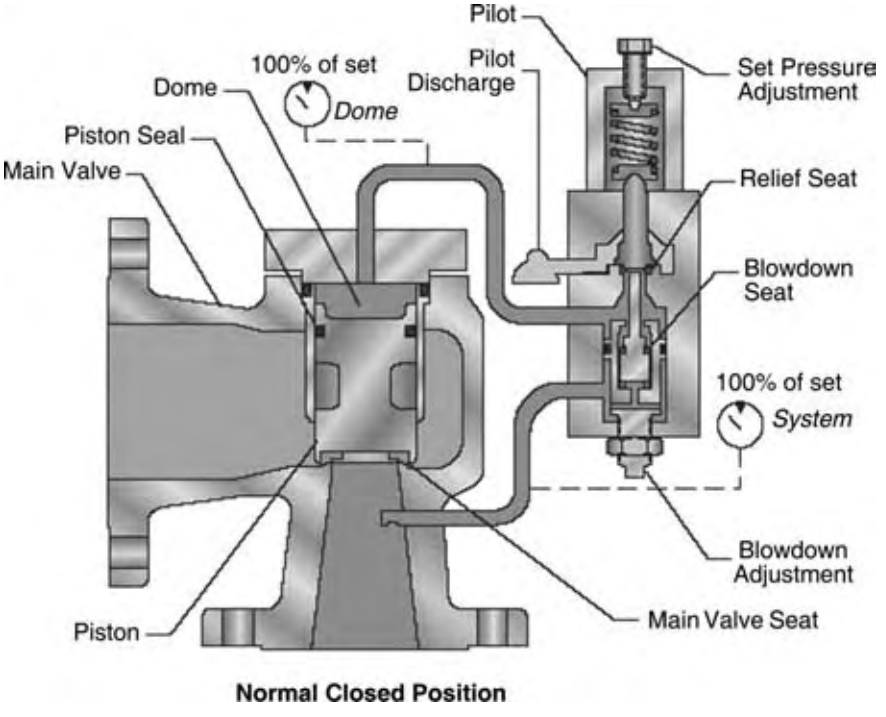


Fig. 14 Schematic of a pilot-operated valve in the closed position. (Reproduced by kind permission of Tyco Valves & Controls.) *(View this art in color at www.dekker.com.)*

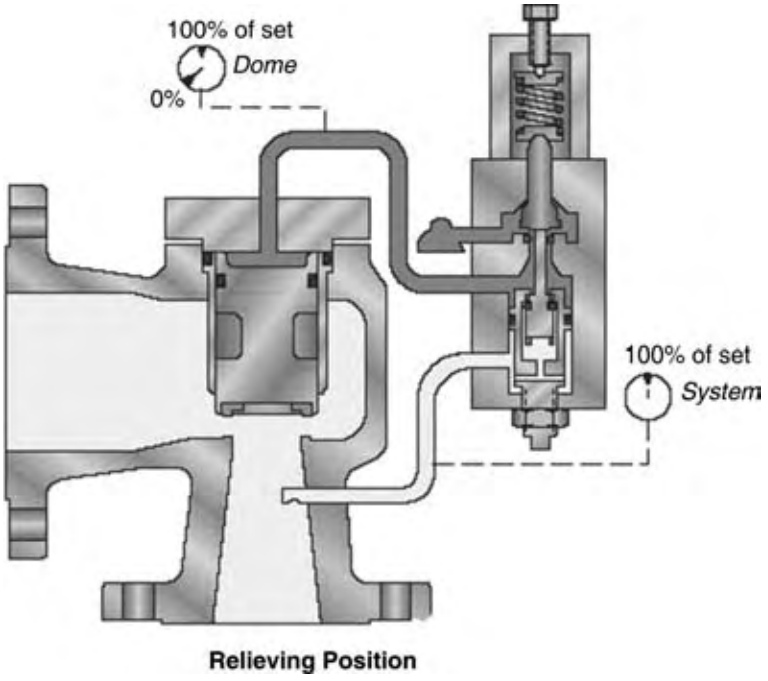


Fig. 15 Schematic of a pilot-operated valve when draining fluid from the upper chamber. (Reproduced by kind permission of Tyco Valves & Controls Ltd.) *(View this art in color at www.dekker.com.)*

relax. Hence, leakage is minimized until the set pressure is achieved. Moreover, when the main valve reaches set pressure it opens to full lift without an appreciable pressure accumulation, whereas a spring-loaded valve needs some accumulation above set pressure to raise the disk against its spring, to generate sufficient change of momentum beneath the disk to cause rapid opening. Pilot-operated valves are also thought to incur less hysteresis from friction and stiction in the moving part guides, partly owing to piston rings on the main valve disk sealing the upper chamber from the bore and partly because the pilot valve's travel and tilt are reduced compared to an equivalent-sized spring-loaded valve. As a result, the blowdown can be reduced.

Reduced accumulation, reduced blowdown, and improved sealing have given pilot-operated valves a reputation for minimizing the loss of system fluid when called on to operate. They require less overall mass than an equivalent-sized spring-loaded valve, and the pilot valve could potentially be sited remote from the main valve (but limited by lag in the bleed lines). However, they have more moving parts and more and narrower flow passages, making them more prone to failure through damage or blockages.

Finally, in the United Kingdom increasing regulation over the last 15 yr has driven the relief valve industry, but has led to few design improvements. Products sold 15 yr ago have proven suitable for meeting new statutory requirements for potable water and national regulations introduced to meet the European Union Pressure Equipment Directive (97/23/EC) and "ATEX" Directive (94/9/EC). Together with harmonization of Technical Standards across the European Union, these laws have led to more robust methods of independent witnessing of type tests and approval of products. However, the focus has been on providing documentary evidence of design and performance and not on improving already well-established and successful designs.

CONCLUSIONS

Design of spring-loaded pressure-relief valves is supported by a well-established mechanical engineering science base. With the use of established techniques, together with modern developments in zone and field modeling, there is potential for incremental improvements to be made to flow structure and stability or modulation of relief valves. However, for leading manufacturers supplying the mass market, the performance vs. cost balance must surely be close to its optimum.

Pilot-operated valves, and to a lesser extent spring-loaded valves offer opportunities for waste minimization. Advances in sealing technology and reductions in friction/stiction may continue to be made to reduce accumulation and blowdown. The potential advantages of fluidic design to this goal are as yet unclear.

ACKNOWLEDGMENTS

The author wishes to thank the following people who have contributed to the discussions for preparing

this entry: Dr. P. L. Betts, Senior Lecturer (retired), UMIST; Mr. S. Cooper, Technical Director, Delta Fluid Products; and a small group of people who, for personal or commercial reasons, wish to remain anonymous.

REFERENCES

1. Fuller, J.A. The history of safety valves leading to the development of British Standards for safety valves. IMechE Conference Publication C365/84: Safe Pressure Relief, 1984; 1–12.
2. BS 759 *Valves, Gauges and Other Safety Fittings for Application to Boilers and to Piping Installations for and in connection with Boilers, Part 2: Specification for Safety Valves (obsolescent)*. British Standards Institute: London, 1975.
3. BS 6759 *Safety Valves, Parts 1–3*. British Standards Institute: London, 1984.
4. BS EN ISO 4126-1. *Safety Devices for Protection Against Excessive Pressure, Part 1: Safety Valves*; British Standards Institute: London, 2004.
5. Williams, F.A. *Combustion Theory: The Fundamental Theory of Chemically Reacting Flow Systems*, 2nd Ed.; Addison-Wesley: Redwood City, CA, 1985.
6. Zeldovich, Y.B.; Barenblatt, G.I.; Librovich, V.B.; Makhviladze, G.M. *The Mathematical Theory of Combustion and Explosions;* translated from Russian by McNeill, D.H.; Consultants Bureau: New York, 1985.
7. API 520-1 *Sizing, Selection, and Installation of Pressure-Relieving Devices in Refineries, Part 1: Sizing and Selection*, 7th Ed.; American Petroleum Institute: Washington, DC, 2000.
8. Wong, W.Y. Fires, vessels, and the pressure relief valve. Chem. Eng. **2000**, 84–92.
9. BS EN 13445. *Unfired Pressure Vessels, Parts 1–7*; British Standards Institute: London, 2002.
10. ASME BPV *Boiler and Pressure Vessel Code,* Complete Code. American Society of Mechanical Engineers: New York, 2004.
11. BS EN ISO 4126-2. *Safety Devices for Protection Against Excessive Pressure, Part 2: Bursting Disc Safety Devices*; British Standards Institute: London, 2004.
12. Chung, P.W.H; Yang, S.H.; He, C.H. Conceptual design of pressure relief systems. J. Loss Prev. Process Ind. **2000**, *13* (6), 519–526.
13. API 526. *Flanged Steel Pressure Relief Valves*; American Petroleum Institute: Washington, DC, 2002.
14. Anon. *Water Regulations Guide*; Water Regulations Advisory Scheme: Gwent, 2002.
15. Anon. *Crosby Pressure Relief Valve Engineering Handbook*; Technical Document No. TP-V300, 1997. www.andersongreenwood.com/literature.asp.
16. Frommann, O.; Friedel, L. Analysis of safety valve chatter induced by pressure waves in gas flow. J Loss Prev. Process Ind. **1998**, *11* (4), 279–290.
17. Borzov, B.A.; Grishin, Y.M.; Kozlov, V.M. Study of the characteristics of angle-type safety valves. Soviet Energy Technol. **1984**, *12*, 45–49.
18. Francis, J. Pressure Relief Valves: An Improved Design Methodology. Ph.D. Thesis, University of Manchester Institute of Science and Technology, Manchester, U.K., 1994.
19. Bicen, A.F.; Vafidis, C.; Whitelaw, J.H. Steady and unsteady airflow through the intake valve of a reciprocating engine. Trans. ASME: J. Fluids Eng. **1985**, *107*, 413–420.
20. Follmer, B.; Zeller, H. The influence of pressure surges on the functioning of safety valves. In Proceedings of the 3rd International Conference on Pressure Surges, BHRA Fluid Engineering, 1980; 429–444.
21. Sterland, P.R. Understanding the behaviour of relief valves. IMechE Conference Publication C365/84: Safe Pressure Relief, 1984; 57–62.
22. Betts, P.L.; Francis, J. Design of high-lift pressure relief valves with non-adjustable blowdown for gas/steam. Proc. IMechE Pt. E: J. Process Mech. Eng. **1995**, *209* (1), 3–15.
23. Ray, A. Dynamic modelling and simulation of a relief valve. Simulation **1978**, *31* (5), 167–172.
24. Singh, A. An analytical study of the dynamics and stability of a spring loaded safety valve. Nucl. Eng. Des. **1982**, *72* (2), 197–204.
25. Kalra, S.P.; Liou, S.Y. Dynamic simulation of normal and off-normal flow discharge through safety relief valves. AIChE Symp. 236/80: Heat Transfer **1984**, 267–274.
26. Watton, J. The stability of a two-stage pressure relief valve. Proc. Am. Control Conf. **1989**, *2*, 1503–1507.
27. Ziegler, B.; Coppolani, P.; Garcia, J.L. Dynamic behaviour of check valves and safety valves with fluid interaction. Nucl. Eng. Des. **1990**, *124* (3), 391–401.
28. Dasgupta, K.; Karmakar, R. Modeling and dynamics of a single-stage pressure relief valve with directional damping. Simul. Model. Prac. Theory **2002**, *10* (1–2), 51–67.
29. Kolenko, N.N.; Dedkov, A.K.; Mulyukin, O.P.; Kondrashov, Y.I. Determination of the flow-rate factor of the medium in high-lift safety valves. Chem. Pet. Eng. **1984**, *20* (5), 244–246.
30. Kruisbrink, A.C.H. Modeling of safety and relief valves in waterhammer computer codes. Proceedings of the 3rd International Conference

Developments in Valves & Actuators for Fluid Control; Reproduced by Delft Hydraulics, 1990.
31. Zucrow, M.J.; Hoffman, J.D. *Gas Dynamics*; Wiley: New York, 1976; Vol. 1.
32. Yahya, S.M. *Fundamentals of Compressible Flow*; Wiley: New Delhi, India, 1982.
33. Parker, G.J. 'Pop' safety valves: a compressible flow analysis. Int. J. Heat Fluid Flow **1985**, *6* (4), 279–283.
34. Morris, S.D. Liquid flow through safety valves: diameter ratio effects on discharge coefficients, sizing, and stability. J. Loss Prev. Process Ind. **1996**, *9* (3), 217–224.
35. Francis, J.; Betts, P.L. Modeling incompressible flow in a pressure relief valve. Proc. IMechE Pt. E: J. Process Mech. Eng. **1997**, *211* (1), 83–93.
36. Betts, P.L.; Francis, J. Pressures beneath the disc of a compensated pressure relief valve for gas/vapour service. Proc. IMechE Pt. E: J. Process Mech. Eng. **1997**, *211* (3), 285–289.
37. Borzov, B.A.; Kozlov, V.M. Determining the critical pressure ratio in the flow section of a safety valve. Soviet Energy Technol. **1987**, *14*, 39–42.
38. Francis, J.; Betts, P.L. Backpressure in a high-lift compensated pressure relief valve subject to single-phase compressible flow. J. Loss Prev. Process Ind. **1998**, *11* (1), 55–66.
39. Kastner, L.J.; Williams, T.J.; White, J.B. Poppet inlet valve characteristics and their influence on the induction process. Proc. ImechE. **1963**, 178.
40. Lai, Y.S. Conventional spring loaded safety relief valves subjected to back pressure. Proceedings of ASME Pressure Vessels and Piping Conference: Pipeline Dynamics Valves, 1989; PVP 180, 111–118.
41. Francis, J. Zone modelling of high-lift pressure relief valves. Building Serv. Eng. Res. Technol. **2000**, *21* (4), 219–224.
42. Bilanin, A.J.; Teske, M.E. Modeling flow through spring-loaded safety valves. Proceedings of ASME Pressure Vessels and Piping Conference: Pipeline Dynamics Valves, 1990; PVP 190, 29–36.
43. Sethi, S.; Sabet, A. Predicting the fluid flow through a pressure relief valve using the CFD code phoenics. Proceedings of the 11th Annual Simulators' Conference: Simulation Multiconference, 1994; Vol. 26 (88), 229–235.
44. Francis, J. Internal Report, 2004.
45. Darby, R. On two-phase frozen and flashing flows in safety relief valves. J. Loss Prev. Process Ind. **2004**, *17* (4), 255–259.
46. Leung, J.C. A theory on the discharge coefficient for safety relief valves. J. Loss Prev. Process Ind. **2004**, *17* (4), 301–313.

Process Optimization

Ralph W. Pike
Department of Chemical Engineering, Louisiana State University, Baton Rouge, Louisiana, U.S.A.

INTRODUCTION

The objective of optimization is to select the best possible decision for a given set of circumstances without having to enumerate all the possibilities. For an operating plant, the objective is to determine operating conditions that give the maximum profit or the minimum cost. For plant design, the corresponding objective is to determine the process equipment size and minimum cost as measured by the net present value (NPV) and the rate of return. For operating plant and plant design optimization, a quantitative description is required in the form of a model or simulation that accurately predicts the performance of the actual plant. An economic model is required that gives the profit or the cost that can be maximized or minimized. The optimal design and operating conditions for the plant are determined using an optimization algorithm.[1]

There are mathematical programming languages that have the user enter the process and the economic model and then specify a solver (optimization algorithm). The resulting program is run, and the results are presented to the user in a readable form. Examples of these languages are General Algebraic Modeling System (GAMS), A Mathematical Programming Language (AMPL), LINDO/LINGO, and the Solver option in Excel.

The following sections describe the theoretical basis for mathematical programming and applications in linear programming (LP), nonlinear programming (NLP), mixed-integer programming (MIP), and multicriteria programming. Also, formulating economic and process models required for process optimization is summarized. Optimization methods take advantage of the mathematical structure of the economic and the process models to locate the optimum, possibly a global optimum. Models should be developed to use these capabilities to locate optima. For example, if linear equations can provide satisfactory representations of the economics and processes of the plant, then LP can be used to locate the global optimum. However, if the models are nonlinear, optimization methods can only guarantee a better point than the starting point, possibly a local optimum.

PROCESS MODELS

A process model or simulation of a plant is a set of equations that describe the operations of the plant and predict its performances. This set of equations includes material and energy balances, rate equations, and equilibrium relations. The material balances describe the conservation of mass in the individual units of the process, and the energy balance equations describe the conservation of energy in these units. In the material balances, there are terms that describe the rate of conversion of components by chemical reactions. These terms are given by the rate equations from chemical kinetics. In energy balances, there are terms that describe the exchange in energy with the surroundings and the work done by the unit. The energy exchange is described by rate equations for heat transfer, and the work performed is described by the method used for fluid movement, e.g., compressor.

A chemical plant includes tens to hundreds of process units, such as chemical reactors, heat exchangers, distillation columns, absorption towers, etc. For each unit, material and energy balances are used to relate input and output streams. Rate equations and equilibrium relations help describe the conversion of species, mass, and energy in the units. Collectively, these equations provide the equality constraints for the plant model.

The model or simulation of an operating plant requires a detailed model of the process units and their integration. This simulation is developed following the well-known onionskin approach. First, the chemical reactor is described; this is followed by developing the separation and recycle models that model the feed preparation and product purification unit operations. Then, pinch analysis is used to configure the heat exchanger network for minimum utility requirements. The simulation must predict the performance of the actual plant with sufficient accuracy that changes predicted by the model match the changes in the plant, called plant-model matching.

In the design of a new plant, the simulation requirements vary depending on the economic analysis. An order of magnitude economic analysis would use a block flow diagram with complete material and energy

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP–120018026

balances. A study estimate uses a process flow diagram (PFD), and the major equipment is sized. PFD is a major step up from a block flow diagram, and it represents all the equations (process model) necessary for the design of the process. All the major units are represented on the diagram with a unique number and a descriptive name. Also process streams are shown and identified with a number, along with a description of the process conditions and chemical composition included. In addition, all utility streams supplied to the major units are shown along with basic control loops illustrating the control strategy to operate the plant during normal operations. This diagram represents the material and energy balance equations, rate equations, and equilibrium relations.

PROCESS ECONOMICS

New plants, products, and technology require new capital, and most firms have limited resources. Consequently, investment decisions require capital budgeting. This plant profitability analysis is the evaluation and the selection of the best investments from a set of alternatives. Methods for evaluating investments include NPV and rate of return, among others, for private companies and benefit-cost ratio for public works projects. All these come under the purview of plant profitability analysis.

Risk is a part of the decision process also. The analysis of projects must incorporate the level of risk to be able to judge projects with high returns and high risks with those having lower returns and more certain outcomes.

A company typically has several projects competing for funds to be invested. The projects are ranked based on their NPV and risk. This is an economic decision problem. However, each project is an optimization problem by itself. For a valid comparison among projects, the optimum design is required for each project to have the maximum NPV.

Some terms used in economic decision analysis are given in Table 1,[2] and using an annual basis is

Table 1 Terms used in economic decision analysis on an annual basis

Sales (sales price S_p × product flow rate − mass per yr, m)	$S = S_p m$
Manufacturing costs	C_M
General expenses	C_G
Total product cost or total annual expenses	$C_T = C_M + C_G$
Purchased equipment cost	$C_{purchase}$
Installed equipment cost or fixed capital investment, FCI	$C_{installed} \sim (2.5–6.8)C_{purchase}$
Total plant cost or total capital investment, TCI	$C_{total\ plant} \sim 2.4C_{installed}$
Capital expenditure annually	C_{cap}
Depreciation and allowance for tax purposes	$D \sim C_{installed}$/economic life
Gross profit	$P_G = S - C_M - D$
Net annual income before taxes	$I_{net} = S - C_T$
Net annual profit before taxes	$P_{net} = P_G - C_G$
Net annual cash flow before taxes	$CF = I_{net} - C_{cap}$
Taxable income	$I_{net} - D$
Taxes (tax rate $t \sim 35\%$ of taxable income in United States)	$T = t(I_{net} - D)$
Net annual income after taxes	$I_{xt} = I_{net} - T$
Net annual profit after taxes	$P_{xt} = I_{xt} - D$
Net annual cash flow after taxes	$CF_{xt} = I_{xt} - C_{cap}$
Value added (sales − raw materials cost)	$P_{value\ added} = S - C_{raw\ matls}$
Earnings—net annual income after taxes from continuous operations excluding significant extraordinary and nonrecurring items	E
Profit margin − after tax earnings as a percentage of sales	P_{margin}

(From Ref.[2].)

standard. Sales (S), the income or revenues received from customers who will purchase the plant's products and by-products, is estimated using market research. Total product costs or total annual expenses of a plant (C_T), the sum of manufacturing costs (C_M), and general expenses (C_G) are estimated by cost engineering using a flowsheeting program and related information. Capital expenditure annually (C_{cap}) includes funds for working capital and plant additions and modifications.

The annual gross profit (P_G) is the sales less manufacturing costs and depreciation. Depreciation is a business expense. The annual net profit (P_{net}) is the gross profit less the general expenses. The annual net profit after taxes (P_{xt}) is the net income after taxes less depreciation.

Table 1 shows the net or cash annual income before taxes (I_{net}), which is the income from sales less total product costs. Plant equipment can be depreciated, and the taxable income is the net income less depreciation. Taxes (t) are paid on a sliding scale, and a corporate rate of 35% is used for estimation in the United States and approximately 50% in developed countries. Then the net income after taxes (I_{xt}) can be determined.

In profitability assessments, annual cash flows are more meaningful than net profit. The net annual cash flow after taxes (CF_{xt}) is the net profit after taxes less the annual expenditure of capital for additions and modifications. Net annual cash flows are used in discounted cash flow calculations to determine the NPV and the rate of return, which are two key measures used in economic decision analysis.

The following example illustrates the development of an economic model for the NPV, rate of return, and economic price using a preliminary design of a process to produce 100 million pounds per year of aniline from the reaction of phenol and ammonia. The following information has been developed.

Plant capacity	100 million pounds/yr
Plant installed cost	\$6.0 million
Total plant cost	\$11.9 million
Total product cost	\$46.3 million/yr
Annual capital expenditures for worn out equipment	\$0.5 million/yr
Estimated annual sales	\$53.2 million/yr
Economic life	10 yr
Tax rate	35%
Minimum attractive rate of return	15%
Depreciation	Straight-line method with no salvage value

Performing the following economic analysis determines:

Net annual income before taxes
= sales − total product cost
= \$53.2 − \$46.3 = \$6.9 million

Net annual cash flow before taxes
= net annual income before taxes
− annual capital expenditures for worn out equipment
= \$6.9 − \$0.5 = \$6.4 million/yr

Depreciation = plant installed cost/economic life
= \$6.0/10 yr = \$0.6/yr

Taxes = tax rate* taxable income
= tax rate*(net annual income before taxes − depreciation)
= 0.35(6.9 − 0.6) = \$2.2 million

Net annual income after taxes
= net annual income before taxes − taxes
= 6.9 − \$2.2 = \$4.7 million

Net annual cash flow after taxes
= net annual income after taxes
− annual capital expenditures for worn out equipment
= \$4.7 − \$0.5 = \$4.2 million

NPV based on the net annual cash flow after taxes and minimum attractive rate of return:

$$\begin{aligned} NPV &= -CF_0 + A\{[1 - (1 + i)^{-n}]/i\} \\ &= -\$11.9 + \$4.2\{[1 - (1.15)^{-10}]/0.15\} \\ &= \$9.2 \text{ million} \end{aligned}$$

Rate of return where the NPV is zero:

$$\begin{aligned} NPV &= -CF_0 + A\{[1 - (1 + i)^{-n}]/i\} = 0, \\ &\quad -\$11.9 + \$4.2\{[1 - (1 + i)^{-10}]/i\}, \\ i &= 33.3\% \end{aligned}$$

$$\begin{aligned} \text{Annual cost of capital} &= EUAC = P^*(A/P) \\ &= P^*\{i/[1 - (1 + i)^{-n}]\} \\ &= \$11.9\{0.15[1 - (1.15)^{-10}]\} \\ &= \$2.37 \text{ million/yr} \end{aligned}$$

Economic price
= (total product cost
+ annual cost of capital)/product rate
= ($46.3 million/yr + $2.37 million/yr)/
100 million pounds/yr = $0.49/lb

Total Cost Assessment

The business focus of chemical companies has moved from a regional to a global basis, and this has redefined how companies organize and view their activities. Tools such as total (full) cost assessment (accounting) (TCA), life cycle assessment, sustainable development, and eco-efficiency (economic and ecological) are being used. Total or full cost accounting identifies the real costs associated with a product or a process. It organizes different levels of costs and includes direct, indirect, associated, and societal. Direct and indirect costs include those associated with manufacturing. Associated costs include those associated with compliance, fines, penalties, and future liabilities. Societal costs are difficult to evaluate as there are no standard, agreed-upon methods to estimate them, and they can include consumer response and employee relations. The AIChE/CWRT TCA program[3] uses five types of costs. Type 1 costs are direct costs for the manufacturing site, type 2 are potentially hidden corporate and manufacturing site overhead costs, type 3 are future and contingent liability costs, type 4 are internal intangible costs, and type 5 are external or sustainable costs that the company does not pay directly including those borne by society and from deterioration of the environment by pollution within compliance regulations. Sustainable development is the concept that development should meet the needs of the present without sacrificing the ability of the future to meet its needs. External or sustainable costs are very difficult to quantify, and the TCA report gives some estimates for these costs from a study of environmental cost from pollutant discharge to air from electricity generation, e.g., $0.22–2.38/ton for CO, $0–3.25/ton for carbon dioxide.

OPTIMIZATION METHODS

In this section, the important aspects of the mathematical basis for optimization methods are described. This will provide the necessary background to understand the most widely used method, LP. Then descriptions of two more effective NLP methods are outlined: the generalized reduced gradient method and the successive LP method. Then methods for mixed-integer and multicriteria optimization problems are summarized.

Analytical Methods (Theory of Maxima and Minima)

The classical theory of maxima and minima (analytical methods) is concerned with finding the maxima or minima, i.e., extreme points of a function and it provides the theoretical basis for optimization methods and computer programs. This theory determines the values of the n independent variables $x_1, x_2, \ldots, x_n$ of a function, where it reaches maxima and minima points.

The necessary conditions have been developed by Kuhn and Tucker for a general nonlinear optimization problem with equality and inequality constraints.[1] This optimization problem written in terms of minimizing $y(\mathbf{x})$ is:

$$\text{minimize: } y(\mathbf{x}) \tag{1}$$

$$\text{subject to: } f_i(\mathbf{x}) \leq 0 \quad \text{for } i = 1, 2, \ldots, h \tag{2}$$

$$f_i(\mathbf{x}) = 0 \quad \text{for } i = h + 1, \ldots, m \tag{3}$$

where $y(\mathbf{x})$ and $f_i(\mathbf{x})$ are twice continuously differentiable real valued functions.

Any value of $\mathbf{x}$ that satisfies the constraint equations (2) and (3) is called a feasible solution to the problem. Then to locate points that can potentially be local minima of equation and satisfy the constraint equations (2) and (3), the Kuhn–Tucker necessary conditions are used. These conditions are written in terms of the Lagrangian function for the problem, which is:

$$L(\mathbf{x}, \lambda) = y(\mathbf{x}) + \sum_{i=1}^{h} \lambda_i \left[f_i(\mathbf{x}) + x_{n+i}^2 \right] + \sum_{i=h+1}^{m} \lambda_i f_i(\mathbf{x}) \tag{4}$$

where the x_{n+i}s are the surplus variables used to convert the inequality constraints to equalities.

The necessary conditions for a constrained minimum are given by the following theorem:

In order to minimize $y(\mathbf{x})$ subject to $f_i(\mathbf{x}) \leq 0$, $i = 1, 2, \ldots, h$ and $f_i(\mathbf{x}) = 0$, $i = h + 1, \ldots, m$, the necessary conditions for the existence of a relative minimum at $\mathbf{x}^*$ are:

1. $\dfrac{\delta y(\mathbf{x}^*)}{\delta x_j} + \sum_{i=1}^{h} \lambda_i \dfrac{\delta f_i(\mathbf{x}^*)}{\delta x_j} + \sum_{i=h+1}^{m} \lambda_i \dfrac{\delta f_i(\mathbf{x}^*)}{\delta x_j} = 0$
   for $j = 1, 2, \ldots, n$
2. $f_i(\mathbf{x}^*) \leq 0$ for $i = 1, 2, \ldots, h$
3. $f_i(\mathbf{x}^*) = 0$ for $i = h + 1, \ldots, m$
4. $\lambda_i f_i(\mathbf{x}^*) = 0$ for $i = 1, 2, \ldots, h$
5. $\lambda_i > 0$ for $i = 1, 2, \ldots, h$
6. λ_i is unrestricted in sign for $i = h + 1, \ldots, m$ (5)

Examining these conditions, the first one is setting the first partial derivatives of the Lagrangian function with respect to the independent variables $x_1, x_2, \ldots, x_n$ equal to zero to locate the Kuhn–Tucker point, $\mathbf{x}^*$. The second and third conditions repeat the inequality and the equality constraint equations that must be satisfied at the minimum of the constrained problem. The fourth condition is another way of expressing $\lambda_i x_{n+i} = 0$, $i = 1, 2, \ldots, h$ from setting the partial derivatives of the Lagrangian function with respect to the surplus variables equal to zero. Either $\lambda_i = 0$ and $x_{n+i} = 0$ (constraint is active) or $\lambda_i = 0$ and $x_{n+i} \neq 0$ (constraint is inactive). Thus, the product of the Lagrange multiplier and the constraint equation set equal to zero is an equivalent statement, and this is called the complementary slackness condition. The fifth condition comes from examining the equation $\partial y(\mathbf{x}^*)/\partial b_i = -\lambda_i$. The argument is that as b_i is increased the constraint region is enlarged; this cannot result in a higher value for $y(\mathbf{x}^*)$, the minimum in the region. However, it could result in a lower value of $y(\mathbf{x}^*)$, and correspondingly $\partial y(\mathbf{x}^*)/\partial b_i$ would be negative; as b_i increases, $y(\mathbf{x}^*)$ could decrease. Therefore, if $\partial y(\mathbf{x}^*)/\partial b_i$ is negative, then the Lagrange multiplier λ_i must be positive, for the equation to be satisfied. This condition is called a constraint qualification, as is discussed subsequently. For the sixth condition, it has been shown that the Lagrange multipliers associated with the equality constraints are unrestricted in sign; there is no argument comparable to the one given above for the Lagrange multipliers associated with the inequality constraints.

In summary, condition 1 gives a set of n algebraic equations, and conditions 2 and 3 give a set of m constraint equations. The inequality constraints are converted to equalities using h slack variables. A total of $n + m$ constraint equations are solved for n variables and m Lagrange multipliers that must satisfy the constraint qualification. Condition 4 determines the value of the h slack variables. This theorem gives an indirect problem in which a set of algebraic equations is solved for the optimum of a constrained optimization problem.

The Kuhn–Tucker point located using the necessary conditions described above requires sufficiency conditions to determine if it is a maximum, a minimum, or a saddle point. Tests for sufficiency conditions are determined by expanding the Lagrangian function in a Taylor series and neglecting third and higher order terms to give a function that contains only terms involving second partial derivatives evaluated at the Kuhn–Tucker point. This gives a differential quadratic form, and a test similar to the one for the unconstrained problem is obtained to determine if the Kuhn–Tucker point is a maximum, a minimum, or a saddle point. The sufficient conditions for the case of both inequality and equality constraints are more elaborate than if only equality constraints are involved. Further details are given in standard texts.[1,2,4]

Linear Programming

Of all the optimization methods, LP is the one that is most widely applied. The technique has been used for optimizing many diverse applications, including refineries and chemical plants, livestock feed blending, routing of aircraft, and scheduling their crews. Many industrial allocation and transportation problems can be optimized with this method. The application of LP has been successful, particularly in the cases of selecting the best set of values of the variables when a large number of inter-related choices exist. Often such problems involve a small improvement per unit of material flow times large production rates to have as the net result be a significant increase in the profit of the plant. A typical example is a large oil refinery where the stream flow rates are very large, and a small improvement per unit of product is multiplied by a very large flow rate to obtain a significant increase in profit for the refinery.

The term ''programming'' of LP does not refer to computer programming but to scheduling. Linear programming was developed about 1947, before the advent of the computer, when Dantzig recognized a generalization in the mathematics of scheduling and planning problems. Developments in LP have followed advances in digital computing, and now problems can be solved involving several hundred thousand independent variables and constraint equations.

As the name indicates, all the equations that are used in LP must be linear. Although this appears to be a severe restriction, there are many problems that can be cast in this context. In an LP formulation, the equation that determines the profit or the cost of operation is referred to as the objective function. It must have the form of the sum of linear terms. The equations that describe the limitations under which the system must operate are called the constraints. The variables must be nonnegative, i.e., positive or zero only.

Fig. 1 gives some geometric intuition about the mathematical structure of the problem and the way this structure can be used to find an optimal solution. The profit function is a plane and the highest point is the vertex $A = 10$, $B = 20$ with a profit of $P = 110$. The intersection of the profit function and planes of $P =$ constant gives a line on the profit function plane as shown for $P = 96$. This diagram emphasizes the fact that the profit function is a plane, and the maximum profit will be at the highest point on the plane and located on the boundary at the intersection of constraint equations, a vertex.

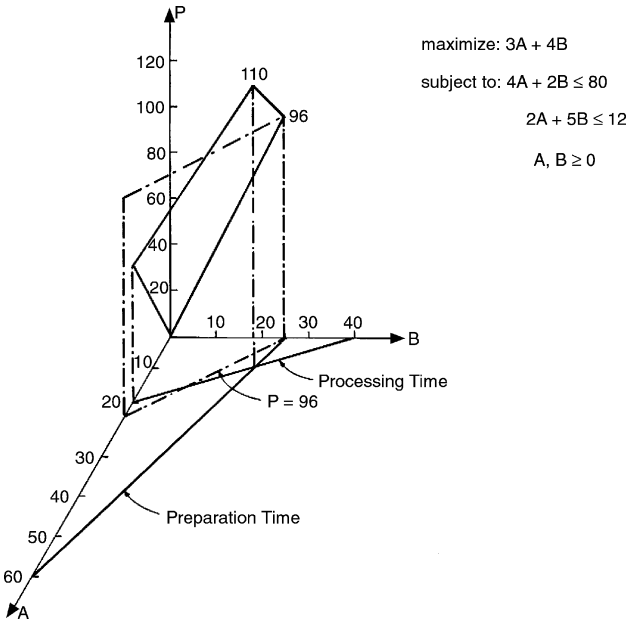


Fig. 1 Geometric representation of objective function and constraints of a LP problem (From Ref.[1].)

The general formulation of the LP problem can be written as:

$$\text{optimize:} \quad \sum_{j=1}^{n} c_j x_j$$

$$\text{subject to:} \quad \sum_{j=1}^{n} a_{ij} x_j \leq b_i \quad \text{for } i = 1,2,\ldots,h$$

$$\sum_{j=1}^{n} a_{ij} x_j = b_i \quad \text{for } i = h+1,\ldots,m \qquad (6)$$

The computational method to determine the optimum, the simplex method, requires equality constraints. The inequalities are converted to equalities by introducing slack and surplus variables. This is illustrated by converting the inequality equation given here to an equality by adding a slack variable x_3.

$$x_1 + x_2 \leq b \qquad (7)$$

$$x_1 + x_2 + x_3 = b \qquad (8)$$

If the inequality had been of greater than or equal to type, then a surplus variable would have been subtracted from the left-hand side of the equation to convert it to an equality.

A basic solution of the constraint equations is a solution obtained by adding slack and surplus variables to the inequality constraints, then by setting $(n - m)$ the variables equal to zero, and solving the constraint set for the remaining m variables. From this set of basic solutions, the groups of solutions where the values of the variables are all nonnegative are called basic feasible solutions; all the m variables are positive. There are several theorems that establish global optimality and relate the maximum or the minimum of the objective function to the basic feasible solutions of the constraint equations. An important result is:[1]

> If the objective function possesses a finite minimum, then at least one optimal solution is a basic feasible solution.

The standard procedure to solve an LP problem is the simplex algorithm.

The steps in this algorithm for maximizing the profit (objective) function are as follows:[1]

1. Place the problem in an LP format with linear constraint equations and linear objective function.
2. Introduce slack and surplus variables to convert inequalities to equalities and adjust the constraint equations to have positive right-hand sides.
3. Select an initial basic feasible solution. If all the constraint equations were inequalities of the "less than or equal to" form, the slack variables can be used as the initially feasible basis.
4. Perform algebraic manipulations to express the objective function in terms of variables that are not in the basis, i.e., are equal to zero. This determines the value of the objective function for the variables in the basis.
5. Inspect the objective function and select the variable with the largest positive coefficient to bring into the basis, i.e., make nonzero. If there are no positive coefficients, the maximum has been reached (automatic stopping feature of the algorithm).
6. Inspect the constraint equations to select the one to be used for algebraic manipulations to change the variable in the basis. The selection is made to have positive right-hand sides from the Gaussian elimination. This is necessary to guarantee that all the variables in the new basis will be positive. Use this equation to eliminate the variable selected in step 5 from all the other constraint equations.
7. Use the constraint equation selected in step 6 to eliminate the variable selected in step 5 from the objective function. This moves one of the variables previously in the basis to the objective function, and it is dropped from the basis, i.e., set equal to zero. Also this determines the new value of the objective function.
8. Repeat the procedure of steps 5–7 until all coefficients in the objective function are negative

and stop. If the procedure is continued past this point, then the value of the objective function would decrease. This is the automatic stopping feature of the algorithm.

The simplex method is called a direct method because it moves from one basic feasible solution to another until the optimum is reached. Examples are given to illustrate this method in standard texts,[1,5,6] and the algorithm can be modified to minimize rather than maximize. Also, artificial variables are added in computer programs to have an initially feasible basis, and these are removed using the big M method or the two-phase method. The revised simplex method is more efficient computationally, and it uses the same logic but works with the coefficients in the original problem. In addition, there is a dual problem that can be constructed from the primal (original) problem that may be easier to solve if it has fewer constraint equations than the original problem. Also, standard texts[1,5,6] discuss degeneracy and infeasible and unbounded solutions.

The procedure for formulating the LP problem for a plant or a process includes developing the objective function from the cost or the profit of the process or the plant and the constraint equations. The equality constraints are the material and energy balances, rate equations, and thermodynamic relations. The inequality constraints are the availability of raw materials, the demand for products, and equipment capacity limitations and conversion capabilities. A simple petroleum refinery has been used to demonstrate the development of the LP problem, and the optimal solution has been obtained using GAMS and Excel for comparison.[1] Linear Programming models of chemical plants and refineries have the form of very large and sparse matrices. These problems can have several hundred thousand constraint equations, and programs such as Aspen Technology's PIMS program are designed for obtaining solutions to these problems.

Sensitivity analysis or postoptimal analysis determines the range of values for a_{ij}, b_i, and c_j for the optimal solution to remain optimal. Changes in the right-hand side of the constraint equations, b_i, correspond to changes in the maximum capacity of a process unit or the availability of a raw material, for example. Changes in the coefficients of the objective function, c_j, correspond to changes of the cost or the sale price of the products. Changes in the coefficients of the constraint equations, a_{ij}, correspond to changes in mass or volumetric yields of a process. Knowing that these various coefficients and parameters can vary with-out changing the optimal solution reduces the number of times the LP problem must be solved.

Sensitivity analysis requires evaluating Lagrange multipliers using the following equation obtained from the classical theory of maxima and minima.

$$\lambda_i = -\sum_{k=1}^{m} \beta_{ki} c_c \quad \text{for } i = 1, 2, \ldots, m \tag{9}$$

where β_{ik} are the elements of the inverse of the matrix of the coefficients of the variables that are the basis from the original problem.[1]

The following Eqs. (10) and (11) give the changes of the variables in the basis and profit for changes Δb_i. Referring to Eq. (11), the units on the Lagrange multiplier are dollars per unit mass or volume, hence the name "shadow prices." For the optimal solution to remain optimal, $\mathbf{x}^*_{j,\text{new}} \geq 0$.

$$\mathbf{x}^*_{j,\text{new}} = \mathbf{x}^*_j + \sum_{i=1}^{m} \beta_{ji} \Delta b_i \quad \text{for } j = 1, 2, \ldots, m \tag{10}$$

$$p^*_{\text{new}} = p^* - \sum_{i=1}^{m} \lambda_i \Delta b_i \tag{11}$$

The following equation gives the new value of the reduced cost, $c'_{j,\text{new}}$, the coefficient of the variables not in the basis at the last step in the simplex algorithm.

$$c'_{j,\text{new}} = c'_j + \Delta c_j - \sum_{k=1}^{m} \Delta c_k \sum_{k=1}^{m} \Delta a_{ij} \beta_{ki} \quad \text{for } j = m + 1, \ldots, n \tag{12}$$

where Δc_j, is the change in nonbasic variable cost coefficients c_j, and Δc_k, is the change in the basic variables cost coefficients c_k. When maximizing, the new coefficients must remain negative for the variables not in the basis to have the optimal solution remain optimal, i.e., $c'_{j,\text{new}} < 0$.

Examples illustrating these procedures are given in standard texts.[1,5,6] Also, equations for changes in a_{ij} are given in these texts.

Nonlinear Programming

There are essentially six types of procedures to solve constrained nonlinear optimization problems. The three methods considered more successful are the successive LP, the successive quadratic programming, and the generalized reduced gradient method. These methods use different strategies but the same information to move from a starting point to the optimum, the first partial derivatives of the economic model, and constraints evaluated at the current point. Successive LP is used in a number of solvers including MINOS. Successive quadratic programming is the method of

choice for flowsheeting optimization, for example in Aspen Plus. The generalized reduced gradient method is the optimization program used in Excel's Solver. The other three have not proved as useful as these three, especially in problems with a large number of variables. These are penalty and barrier function methods, augmented Lagrangian functions, and the methods of feasible directions (or projections), which are sometimes called methods of restricted movement.

Successive LP solves a sequence of LP problems to move from a starting point to the optimum. Successive quadratic programming solves a sequence of quadratic programming problems to move from a starting point to the optimum. The generalized reduced gradient method performs a series of line searches in space in the direction of the reduced gradient to move from a starting point to the optimum. These methods require a feasible starting point and a stopping criterion.

Successive LP expands Eqs. (1–3) in a Taylor series about a feasible point $\mathbf{x}_k$ and retains the linear terms. This formulation gives an LP problem in Δx_j, which has to be modified using $\Delta x_j^+ = \Delta x_j - \Delta x_j^-$ where $\Delta x_j^+ = \Delta x_j$ if $\Delta x_j > 0$ and 0 if not, and $\Delta x_j^- = -\Delta x_j$ if $\Delta x_j < 0$ and 0 if not. This permits the algorithm, Eq. (13), to move in all directions from $\mathbf{x}_k$ to locate an optimum.

$$\begin{aligned}
\text{optimize:} \quad & \sum_{j=1}^{n} c_j \Delta x_j^+ - \sum_{j=1}^{n} c_j \Delta x_j^- = y - y(\mathbf{x}_k) \\
\text{subject to:} \quad & \sum_{j=1}^{n} a_{ij} \Delta x_j^+ - \sum_{j=1}^{n} a_{ij} \Delta x_j^- \le b_i - f_i(\mathbf{x}_k) \\
& \qquad \text{for } i = 1, 2, \ldots, m \\
& \Delta x_j^+ - \Delta x_j^- \le (u_j - x_{jk}) \\
& \qquad \text{for } i = 1, 2, \ldots, n \\
& \Delta x_j^+ - \Delta x_j^- \ge (l_j - x_{jk})
\end{aligned} \tag{13}$$

The bounds on the upper and lower limits on the variables are specified by $(u_j - x_{jk})$ and $(l_j - x_{jk})$. The value of the next point for linearizing is given by $x_{jk+1} = x_{jk} + \Delta x_j^+ - \Delta x_j^-$, and the procedure is started by specifying a feasible starting point $\mathbf{x}_0$ $(k = 0)$.

In the generalized reduced gradient method, the independent variables are separated into basic and nonbasic ones. There are m basic variables $\mathbf{x}_\mathrm{b}$, and $(n - m)$ nonbasic variables $\mathbf{x}_\mathrm{nb}$ from Eqs. (2) and (3) with the inequalities converted to equalities using slack and surplus variables. In theory, the m constraint equations could be solved for the m basic variables in terms of the $(n - m)$ nonbasic variables, i.e.,

$$f_i(\mathbf{x}) = f_i(\mathbf{x}_\mathrm{b}, \mathbf{x}_\mathrm{nb}) = 0 \quad \text{for } i = 1, 2, \ldots, m \tag{14}$$

Indicating the solution of $\mathbf{x}_\mathrm{b}$ in terms of $\mathbf{x}_\mathrm{nb}$ from Eq. (14) gives:

$$x_{i,\mathrm{b}} = f_i(\mathbf{x}_\mathrm{nb}) \quad \text{for } i = 1, 2, \ldots, m \tag{15}$$

Then, the economic model can be thought as a function of the nonbasic variables only, i.e., if the constraint Eq. (15) is used to eliminate the basic variables in Eq. (1), i.e.,

$$y(\mathbf{x}) = y(\mathbf{x}_\mathrm{b}, \mathbf{x}_\mathrm{nb}) = y(f(\mathbf{x}_\mathrm{nb}), \mathbf{x}_\mathrm{nb}) = Y(\mathbf{x}_\mathrm{nb}) \tag{16}$$

Expanding the above equation in a Taylor series about $\mathbf{x}_k$ and including only the first order terms gives:

$$\nabla^T Y(\mathbf{x}_k) \mathrm{d}\mathbf{x}_\mathrm{nb} = \nabla^T y_\mathrm{b}(\mathbf{x}_k) \mathrm{d}\mathbf{x}_\mathrm{b} + \nabla^T y_\mathrm{nb}(\mathbf{x}_k) \mathrm{d}\mathbf{x}_\mathrm{nb} \tag{17}$$

A Taylor series expansion of the constraint equations gives an equation that can be substituted into Eq. (17) to eliminate the basic variables.

$$\begin{aligned}
& \sum_{j=1}^{m} \frac{\partial f_i}{\partial x_j}(\mathbf{x}_k) \mathrm{d}x_{j,\mathrm{b}} + \sum_{j=m+1}^{n} \frac{\partial f_i}{\partial x_j}(\mathbf{x}_k) \mathrm{d}x_{j,\mathrm{nb}} = 0 \\
& \quad \text{for } i = 1, 2, \ldots, m
\end{aligned} \tag{18}$$

and Eq. (18) can be written in the matrix form as:

$$\mathbf{B}_\mathrm{b} \, \mathrm{d}\mathbf{x}_\mathrm{b} + \mathbf{B}_\mathrm{nb} \mathrm{d}\mathbf{x}_\mathrm{nb} = 0 \tag{19}$$

where $\mathbf{B}_\mathrm{b}$ as the matrix of the first partial derivatives of f_i associated with the basic variables and $\mathbf{B}_\mathrm{nb}$ as the matrix associated with the nonbasic variables.

Eq. (19) can be solved for $\mathrm{d}\mathbf{x}_\mathrm{b}$ and substituted into Eq. (17) to obtain the equation for the reduced gradient given below.

$$\nabla^T Y(\mathbf{x}_k) = \nabla^T y_\mathrm{nb}(\mathbf{x}_k) - \nabla^T y_b(\mathbf{x}_k) \mathbf{B}_\mathrm{b}^{-1} \mathbf{B}_\mathrm{nb} \tag{20}$$

Knowing the values of the first partial derivatives of the economic model and constraint equations at a feasible point, the reduced gradient can be computed by Eq. (20). Then the reduced gradient line is used to locate the optimum by a single variable search on α, the parameter of the reduced gradient line in Eq. (21).

$$\mathbf{x}_\mathrm{nb} = \mathbf{x}_{k,\mathrm{nb}} + \alpha \nabla^T Y(\mathbf{x}_k) \tag{21}$$

In taking trial steps as α is varied along the generalized reduced gradient line, the matrices $\mathbf{B}_\mathrm{b}$ and $\mathbf{B}_\mathrm{nb}$ must be evaluated along with the gradients $\nabla y_\mathrm{b}(\mathbf{x}_\mathrm{b})$ and $\nabla y_\mathrm{nb}(\mathbf{x}_\mathrm{k})$. This requires knowing both $\mathbf{x}_\mathrm{nb}$ and $\mathbf{x}_\mathrm{b}$ at each step. The values of $\mathbf{x}_\mathrm{nb}$ are obtained from Eq. (21). However, Eq. (15) must be solved for $\mathbf{x}_\mathrm{b}$; frequently, this must be done numerically using the

Newton–Raphson method, which gives Eq. (22).

$$\mathbf{x}_{i+1,\mathrm{b}} = \mathbf{x}_{i,\mathrm{b}} - \mathbf{B}_{\mathrm{b}}^{-1} f\left(\mathbf{x}_{i,\mathrm{b}}, \mathbf{x}_{\mathrm{nb}}\right) \tag{22}$$

where the values of the roots of the constraint equations (2) and (3) are sought for $\mathbf{x}_{\mathrm{b}}$, having computed $\mathbf{x}_{\mathrm{nb}}$ from Eq. (21). Thus, the derivatives computed for the generalized reduced gradient's $\mathbf{B}_{\mathrm{b}}$ matrix can be used in the Newton–Raphson root seeking procedure also.

Successive quadratic programming solves a sequence of quadratic programming problems. A quadratic programming problem has a quadratic economic model and linear constraints. To solve this problem, the Lagrangian function is formed from the quadratic economic model and linear constraints. Then, the Kuhn–Tucker conditions are applied to the Lagrangian function to obtain a set of linear equations. This set of linear equations can then be solved by the simplex method for the optimum of the quadratic programming problem.

Successive quadratic programming solves a NLP problem iteratively by constructing a quadratic approximation to the economic model and a linear approximation to the constraint equations. Then a sequence of quadratic programming problems are solved, and these intermediate solutions generate a series of points that must remain in the feasible region or sufficiently close to this region to converge to the optimum. The logic used with this method is to search along the line between the new and the previous point to maintain a feasible or a near feasible solution. An exact penalty function is used with the line search to adjust the step from one feasible point to the next feasible point.

The computational effort in evaluating the Hessian matrix is significant, and quasi-Newton approximations have been used to reduce this effort. The Wilson–Han–Powell method is an enhancement to successive quadratic programming where the Hessian matrix, $(\mathbf{q}_{jk})$, is replaced by a quasi-Newton update formula such as the BFGS algorithm. Consequently, only first partial derivative information is required, and this is obtained from finite difference approximations of the Lagrangian function.

The methods described in this section locate local optima, and different starting points are used frequently to determine if other local optima are found. Also, the methods can stall before reaching an optimum. Consequently, code testing is an ongoing effort to improve these algorithms. The Optimization Technology Center of the Argonne National Laboratory of the Department of Energy maintains a website that provides recent results.

Methods are being developed that can locate the global optimum of a NLP problem from industrial plants. One approach, αBB,[4] uses an upper bound and a lower bound that is a convex NLP with a global optimum. An iterative procedure is used to have the upper and the lower bounds converge to the global optimum. Another approach, interval analysis,[7] eliminates intervals on the independent variables to locate an increasingly smaller region that contains the global optimum. A third approach uses convexification.[8]

Mixed-Integer Programming

There are linear and NLP problems that require integer values for some or all the decision variables. For example, integer quantities are necessary for activities associated with machines, vehicles, or people. These problems with some of the variables having integer values are known as MIP problems.

Mixed-integer linear programming (MILP) problems require maximizing or minimizing a linear function subject to linear equality or inequality constraints with integer restrictions on some or all the variables. The mathematical statement of MILP can be expressed as:

$$\text{(MILP)} \quad \max\left\{cx + hy : \mathbf{A}x + \mathbf{G}y \leq b, \; x \in Z_+^n, y \in R_+^p\right\} \tag{23}$$

where Z^n is a set of n-dimensional vector of positive integers and R^p is a set of p-dimensional positive real vectors. The variables or unknowns are $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_p)$. $\mathbf{A}$ and $\mathbf{G}$ are $m \times n$ and $m \times p$ matrices, respectively. The objective function is $z = cx + hy$ with c and h being n and p ordered vectors, respectively.

The MILP has two special cases: LP that has all continuous variables and integer programming that has only integer variables. The mathematical statement of an integer LP problem is the same as the LP model, but with an additional restriction that the variables must take on integer values.

The two primary determinants of computational difficulty for an MILP problem are the number of integer variables and the structure of the problem. The number of integer variables is important, because the computational time increases tremendously as the number of integer variables increases.

The most widely used method for solving both integer and MIP problems is the branch-and-bound algorithm. Most commercial computer codes for solving integer programming problems use this approach. The method performs an efficient enumeration of a small fraction of the possible feasible integer solutions to locate the optimum.

The basic idea of the branch-and-bound technique is to divide and conquer. If the original problem is very

large, then it would be difficult to solve it directly; and hence it is divided into increasingly smaller subproblems until they can be solved easily or conquered. To divide (branch) the original problem into smaller subproblems, the entire set of feasible solutions is partitioned into increasingly smaller subsets; and for each one, an upper bound for the value of the objective function is obtained from the solutions within that subset (when maximizing). The conquering (fathoming) is done in two parts. Firstly, the bounds for the best solution in the subset are found; and then the subset is discarded if its bound indicates that it cannot possibly contain an optimal solution for the original problem. The subset with the highest upper bound is partitioned further into subsets. Their upper bounds are obtained in turn and used as before to exclude some of these subsets from further consideration. From all the remaining subsets, another one is selected for further partitioning and so forth. This process is repeated until a feasible solution is located such that the corresponding value of the objective function is greater than the upper bound for any of the other subsets. Such a feasible solution must be optimal because none of the subsets can contain a better solution. A detailed description of this algorithm is available in Refs.[2,4]

Design optimization problems frequently involve both continuous and binary variables, and have the form of mixed-integer nonlinear programming (MINLP) problems. The continuous variables represent the flow rates, temperature, pressures, etc., and binary variables represent the configuration of process units. These problems have been difficult to solve, and a significant amount of research has been spent developing algorithms that are effective in solving MINLP problems. The results have been improved algorithms implemented in relatively reliable computer programs. The mathematical form of a MINLP problem can be expressed as:

$$\begin{aligned} \text{minimize:} \quad & c^T y + f(x) \\ \text{subject to:} \quad & \mathbf{A}y + h(x) = 0 \\ & \mathbf{B}y + g(x) \leq 0 \\ & x \in X = \left\{x | x \in R^n, x^L \leq \mathbf{x} \leq x^U\right\} \\ & y \in Y = \{y | y \in \{0,1\}^m, \mathbf{A}y \leq \mathrm{a}\} \end{aligned} \quad (24)$$

where $\mathbf{x}$ is a vector of continuous variables and y is a set of binary variables that can be used to define the topology of the system representing the existence or nonexistence of different processing units. The nonlinearities in the economic and process models appear in the terms $f(x)$, $g(x)$, and $h(x)$.

MINLP problems can be solved using several algorithms including branch-and-bound, generalized Benders decomposition (GBD), and the Outer Approximation/Equality–Relaxation (OA/ER). These and other methods have advantages and disadvantages.[2] Floudas[4] described the GMIN-αBB algorithm that uses a branch-and-bound strategy to locate the global optimum of nonconvex MINLP problems, while Tawarmalani and Sahinidis[8] described the BARON algorithm that uses a branch-and-bound and convexification strategy. General Algebraic Modeling System includes the following solvers for MINLP problems, DICOPT, SBB, BARON (global optimizer), and solver manuals can be downloaded from the website, www.gams.com.

CONCLUSIONS

Multicriteria optimization adds the complication that there is more than one economic model to be optimized. One approach is to convert the multicriteria into a single criterion using weighting factors and then apply previously described methods. Another approach is to convert the economic criteria into constraints. The objective is to locate an optimal solution to one economic model that cannot be improved without deteriorating other economic models, called Pareto optimal solutions. The most difficult optimization problem to solve is a multicriteria, MINLP. Other important topics that are worth pursuing for process optimization include online optimization where data are obtained from the distributed control system of the plant, gross errors are removed, and the data are reconciled to satisfy material and energy balances. These reconciled data are used to update parameters to have plant-model matching, and then optimal set points are determined using economic optimization.[2] Dynamic programming converts a large, complicated optimization problem into a series of simpler, related problems that can be solved with the dynamic programming algorithm. Dynamic programming is very effective for optimization over time and optimal allocation problems.[1] A statistical approach divides the feasible region into a set of grid points. Then the economic model is evaluated at a number of these grid points selected randomly. Depending on the total number of grid points and the number of evaluations of the economic model, a statement can be made that the largest value found is in the top x percent with a probability of y.[1] These methods are called random search, genetic algorithms, and simulated annealing. Sensitivity analysis for NLP comparable to LP is not available, and Monte Carlo simulation is used to determine the sensitivity of the optimal solution to parameters in the process and economic model. Variational methods determine an optimal function rather than an optimal point and are based on the calculus of variations. Variational methods are the basis

for optimal control and are effective methods for approximated solutions to the second order partial differential equations arising from the conservation equations in momentum, energy, and mass transfer (transport phenomena).[1]

REFERENCES

1. Pike, R.W. *Optimization for Engineering Systems*; Van Nostrand Reinhold: New York, NY, 1984.
2. Pike, R.W. *Optimization for Engineering Systems*; 2nd Ed.; www.mpri.lsu.edu, Baton Rouge, LA, 2003.
3. Constable, D.; Arthur, D.L.; Jesse, H.; Duane, K.; Jill, M.; Jeff, M.; Lawrence, H.; Laura, T.; Earl, B.; Mark, P.; Ion, N.; Scott, B.; Kathy, K.; Paul, C. *Total Cost Assessment Methodology; Internal Managerial Decision Making Tool*; AIChE/CWRT; AIChE: New York, NY, Feb 10, 2000.
4. Floudas, C.A. *Deterministic Global Optimization*; Kluwer Academic Publishers: Boston, MA, 1999.
5. Hiller, F.S.; Lieberman, G.J. *Introduction to Operations Research*; McGraw-Hill: New York, NY, 1990.
6. Ecker, J.G.; Kupferschmid, M. *Introduction of Operations Research*; Wiley: New York, NY, 1988.
7. Hansen, E. *Global Optimization Using Interval Analysis*; Marcel Dekker, Inc.: New York, NY, 1992.
8. Tawarmalani, M.; Sahinidis, N.V. *Convexification and Global Optimization in Continuous and Mixed-Integer Programming*; Kluwer Academic Publishers, 2002.

Processing of Pharmaceuticals Using Dense Gas Technologies

P

Raffaella Mammucari
Fariba Dehghani
Neil R. Foster
School of Chemical Engineering and Industrial Chemistry, The University of New South Wales, Sydney, New South Wales, Australia

INTRODUCTION

Particle size and polymorphism are key characteristics of pharmaceutical compounds. They determine not just the ''processing behavior'' but also the therapeutic effects of active pharmaceutical ingredients and their formulation with excipients. Particle size and polymorphism determine the dissolution rate of poorly water soluble drugs, which, in turn, is related to bioavailability. Dimensional properties can determine the practicability of specific administration routes. Efficient lung delivery, for instance, is heavily dependent on the production of particulates with suitable fluid dynamic properties to ensure the efficient deposition of the active in the target site as well as consistency and uniformity of the active dose. It is widely recognized that the ideal aerodynamic particle size for lung delivery is between 1 and 5 μm. Nanosized particles are amenable to parenteral delivery in suspension, a mode of delivery particularly advantageous for actives with very low solubility in biological fluids since particles in this size range can cross the cellular membranes so that a preliminary dissolution stage is not required.

Particle design applied to pharmaceutical processing has the potential to improve the efficacy of current medications as well as to open the way to the use of alternative delivery routes. An example is the administration of drugs, such as insulin, that are subject to extensive gastrointestinal breakdown and thus cannot be administered orally. The alternative is parenteral administration, which has major side effects, especially in long-term or chronic conditions.

Conventional micronization techniques are not always adequately responsive to the demands of the pharmaceutical industry. Processing limitations can be related to high shear forces, high temperature, electrostatic charges, and difficulties of solvent removal that are the main disadvantages of jet milling, spray drying and ball milling, and liquid–liquid antisolvent processes, respectively. An emerging technology that has demonstrated the capability of generating micrometer- and nanoparticles of active pharmaceutical compounds utilizes the properties of dense gases (DGs).

Dense gases are fluids that are near or above their critical points, generally with reduced temperature and pressure between 0.9 and 1.2. Near-critical fluids exhibit distinctive properties such as low viscosity, high diffusivity, low surface tension, and solvation power. Because the properties of DGs are dependent on pressure and temperature, they offer a unique environment with tunable properties that is ideal for the formulation of pharmaceuticals.

PARTICLE FORMATION

Rapid Expansion of a Supercritical Solution

The rapid expansion of a supercritical solution (RESS) is a crystallization process that relies on the ability of DGs to dissolve the compound to be processed. Conceptually, the process comprises two stages: initially, the DG dissolves the solute and subsequently, the DG solution is depressurized through an expanding device. The abrupt pressure reduction dramatically reduces the dissolving power of the dense phase. The pressure wave propagates through the dense phase at the speed of sound inducing a fast nucleation of the solute and promoting the formation of micrometer size particles with narrow particle size distribution. The process generates uncontaminated particles. Pharmaceutical compounds exhibiting a significant solubility in the dense phase are amenable to RESS processing. The dissolution rate of the poorly water soluble anti-inflammatory drug ibuprofen was improved by micronizing the drug with the RESS process. The improvement was related to particle size reduction and the formation of amorphous fractions.[1] The enhancement of dissolution rate achieved by the RESS process can be higher than that obtained with other methods such as milling as indicated by the micronization of the antifungal griseofulvin.[2] Other examples of drugs micronized with the RESS process include

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120040350

β-sitosterol, theophylline, progesterone, medroxyprogesterone acetate, and stigmasterol.[2–5]

The RESS process has also been used to generate composite materials including pharmaceutical agents. Examples of drug–polymer formulations generated by the RESS process are the antiinflammatory naproxen and the cholesterol lowering drug, lovastatin, each processed with the biodegradable polymer poly(lactic acid) producing composite microparticles.[6–8]

To extend the applicability of the RESS process to compounds with low solubility in DG cosolvents have been used. The addition of small amounts of organic solvents can significantly increase the solvation power of carbon dioxide for many solutes. Commonly used modifiers are acetone and low molecular weight alcohols. The addition of 1 wt% acetone improves the solubility of poly-lactic acid in carbon dioxide by 500%.[9] The acetone-modified carbon dioxide was used to process poly-lactic acid by the RESS process. The product morphology varied greatly with the operating conditions, from irregular particles within 10 and 25 μm to microspheres of about 10 μm.[9] Commonly used modifiers are completely miscible with the DG under a wide range of conditions. However, a phase separation should be avoided during the depressurization stage, otherwise the liquid solvent may dissolve the particulate or induce coalescence with substantial loss of the product morphological characteristics.

Alternatively, modifiers that are not solvents for the products can be introduced in the system. Dense gas–modifier mixtures can exhibit higher solvent power than each of the components.[10,11] In such instances phase separation during the depressurization stage does not affect the particle morphology because the liquid phase present is a nonsolvent for the product. The process is known as RESS-N and has been applied to the processing of various polymers. Mishima et al. reported the formation of micrometer-sized composite particles including the proteins lysozyme and lipase and other compounds of pharmaceutical interest such as the active ingredients *p*-acetamidophenol, acetylsalicylic acid, 1,3-dimethylxanthine, flavone, and 3-hydroxyflavone. The coating agent—poly(ethylene glycol) (PEG) with different molecular weights—was dissolved in carbon dioxide modified with ethanol. Ethanol was preferred over other modifiers, such as methanol, propanol, acetone, and toluene, because of its significant cosolvency effect and low toxicity. The solubility of PEG with different molecular weights in carbon dioxide rose from below 1 wt% to about 13 wt% when 35 wt% ethanol was added to carbon dioxide at 35°C and 160 bar. The process was performed by using modified carbon dioxide to solubilize the polymer and the DG solution was used to suspend micrometer-sized protein. The suspension was depressurized through a capillary nozzle causing the precipitation of the polymer on the protein particles. Discrete microparticles were produced in all cases with a coating thickness that could be controlled through variation of the polymer concentration in the dense phase. The process was extended to the use of other coating agents: poly(methyl methacrylate), poly(lactic acid), poly(DL-lactide-*co*-glycolide), poly(DL-lactide-*co*-glycolide), poly(ethylene glycol)–poly(propylene glycol)–poly(ethylene glycol) triblock copolymer, and ethylcellulose;[10,11] however, such polymers exhibit higher solubility in liquid ethanol rather than in ethanol–carbon dioxide mixtures, therefore its definition as a nonsolvent in the process may not be proper.

The potential of the RESS process for the manufacture of pharmaceutical formulations with superior characteristics has been demonstrated. The implementation of the process on the production scale presents challenges related to the high gas/product ratios that are usually required, high operating pressures, complexity of particle recovery, and intense cooling effect during the expansion stage, which can result in nozzle blockage.[12] Batch mode RESS plants for the production of tens of kilograms per day are currently in operation.[13]

Particles from Gas Saturated Solutions

Particles from gas saturated solutions (PGSS) is a technique applicable to compounds that melt under exposure to DGs. Melting point depression by exposure to high-pressure fluids is a well-known phenomenon that has been observed in various systems. In the PGSS process the DG dissolves in the product and can induce a phase transition, from solid to fluid, and reduce melt viscosity. The melt is rapidly expanded through a nozzle. Associated with the rapid expansion is a sharp drop in temperature and the release of the gas that induces the crystallization of the product. The process can be extended to the processing of multiphase systems such as emulsions and suspensions. The PGSS process is particularly suitable for the processing of systems containing polymers that are often able to absorb large amounts of DGs.[14] The PGSS process can generate micrometer-sized particles and is suitable also for the processing of highly viscous and adhesive material.[15] By controlling the operating conditions, it is possible to produce highly differentiated products from microparticles to films.

Composite particles for controlled release of pharmaceutical compounds have been generated by the PGSS process. An example is the formation of 2–3 μm theophylline/hydrogenated palm oil (HPO) particles.[14] The process produced fine and dry powders operating at 60°C and at pressures between 120 and 180 bar. The morphology of the particulates varied with the

operating pressure and included spherical and needle-like particles.[14] Dissolution studies showed the burst release of 22–45% of the drug content followed by an almost complete release over 10 hr. The results indicated that a significant amount of theophylline was on the surface of the microparticles.[14]

It has to be noted that the drug concentration in the dissolution medium approximated its equilibrium solubility values,[14] thus a reduction of the dissolution rate vs. time cannot be conclusively related to the coformulation with HPO.

The PGSS process has also been successfully applied to the production of drug/polymer composites with the intent of enhancing the solubility and the dissolution rate of the active ingredient. Nifedipine and felodipine are antihypertensive drugs with very low solubility in water. The coprecipitation of each compound with the hydrophilic polymer PEG (MW 4000) was achieved by the PGSS process using carbon dioxide between 50°C and 70°C and at pressures below 200 bar. Both DG processed drug formulations exhibited dramatic dissolution rate improvements compared with untreated samples.[16]

The PGSS process is quite versatile and does not generally involve the use of liquid solvents. Compared with the RESS process it requires much lower DG consumption and operates at lower pressures. Particle recovery is easier than in the RESS process because bigger particles are formed from smaller volumes of DG. The reduced amount of DG required and the lower operating pressure substantially affect the installation and operation costs related to the size of the high-pressure vessels required and the recycling of the DG. The process has been used on the pilot scale for various compounds and has also been applied on a large scale for coating.[15,17] Knez and Weidner have estimated the processing costs to be between 0.15 and 0.60 Euro/kg. The operating cost of the PGSS process will obviously depend on the substrate as well as on the size of the plant. However, the cost estimate indicates that the process may also be suitable for the production of formulations that are not highly valuable.[15]

Gas AntiSolvent Precipitation

Most pharmaceutical compounds have low solubility in DGs and can be processed by DG antisolvent techniques. Dense gases can expand solutions, thereby lowering the solvent power of the solvent and triggering the precipitation of the solute thus acting as antisolvents. Various processes have been developed utilizing DGs as antisolvents. A major distinction is made between batch and semicontinuous processes. In the first case, a liquid solution is introduced into a high-pressure vessel. Subsequently, the system is pressurized with a DG that expands the solution to an extent that is dependent on the operating conditions. Stirring devices can eventually be used to improve mass transfer in the system. Once the final pressure is reached and the precipitate is formed the expanded solvent is purged from the system and the residual solvent is removed from the product by a flux of DG antisolvent. The system is then depressurized and the product collected. The process is commonly referred to as gas antisolvent precipitation (GAS).

Gas antisolvent processes can be performed in a semicontinuous mode. In this case the solution and the antisolvent are continuously introduced in the system until the desired amount of the product is formed. The introduction of the solution is then stopped and the DG flux extracts the residual solvent from the system. The system is then depressurized to enable collection of the product. The solution is generally introduced through an atomization nozzle that favors the prompt expansion of the solution and the formation of small particles. Different process configurations have been utilized, i.e., co- and countercurrent introduction of the solution and antisolvent fluxes and various nozzles have been designed. The process is referred to by different acronyms such as ASES (aerosol solvent extraction system), SAS (supercritical antisolvent), SEDS (solution enhanced dispersion by supercritical fluids), PCA (precipitation with a compressed fluid antisolvent), GASR (gas antisolvent recrystallization), GASP (gas antisolvent precipitation).

Dense gas antisolvent processes have been widely used for pharmaceutical applications because of the possibility of operating under mild conditions, thus preserving the properties of labile compounds, and because of their ability to generate micronsized particles with optimum morphology for various drug applications. Crystallization by DG antisolvent processing has a very broad applicability because many organic solvents, including low molecular weight alcohols, dimethyl formamide, acetonitrile, and methylene chloride, can be expanded by DGs.[18]

Processing Proteins from Aqueous Solutions

The range of applications of peptide drugs is very broad, and includes analgesia, cancer therapy, and infection treatment.[19]The processing and delivery of proteins are often challenging, as many protein molecules are physically and chemically unstable.[20,21] Once administered, proteins can be subject to enzymatic degradation, and present unfavorable permeation through cellular membranes.

Inhalation delivery of therapeutic peptides is a promising administration route to avoid the degradation pathways in the gastrointestinal tract and in the liver.

An optimum aerodynamic particle diameter for lung delivery is considered to be ~3 μm. Particles having aerodynamic size above 5 μm typically deposit in the upper part of the lung, which is not an efficient site for drug adsorption. Particles with aerodynamic diameter smaller than 1 μm can be expelled during exhalation. To deliver powders efficiently via inhalation, particles with defined size must be obtained and delivered.[22–24]

Conventional micronization techniques generally have drawbacks such as interbatch inconsistency, low yield, and broad particle size distribution.[25] Proteins micronized through these techniques can be thermally and chemically degraded.[25] For example, proteins can be unstable to lyophilization while spray drying and milling (capable of producing protein particles in the size range of 5–50 μm) can easily denature the product from the heat and shear forces involved.[26] Grinding with the use of high-velocity compressed air can generate particles in the range 1–10 μm, but there are difficulties in the product collection due to electrostatic charges.[26] Another limit of some conventional techniques is the considerable use of organic solvents, which results in high environmental impact and the presence of solvent residues in the final product.[25]

Dense gas antisolvent techniques are amenable to the precipitation of proteins because of the low solubility of these compounds in DGs such as carbon dioxide. Lysozyme, trypsin, myoglobin, and insulin are examples of peptides that have been precipitated from organic solutions using CO_2 as an antisolvent.[26–30] Both batch and semicontinuous DG antisolvent techniques have been used to precipitate proteins from organic solvents such as methanol, ethanol, and dimethyl sulfoxide.[31,32]

The use of organic solvents, however, can be detrimental to the integrity of the protein. Furthermore, the solubility of proteins is generally low in such solvents so that only low yields can be obtained.[25,33] Aqueous solutions provide a more favorable medium for protein stability and good solvating power so that the precipitation of proteins from aqueous solutions can be preferable to organic solvent systems.[25,33,34]

The precipitation of different proteins from water solutions using modified CO_2 has been described in various works. Lysozyme, trypsin, albumin, and recombinant human deoxyribonuclease (RHO) are examples of proteins that have been precipitated from water using carbon dioxide, modified with ethanol, as an antisolvent.[25,34] Because proteins are typically labile molecules, the ability of DG processing to maintain protein integrity is an important issue. The biochemical integrity of lysozyme, albumin, and insulin has been retained after DG processing, while trypsin and RHO were partially degraded.[25,34] The proteins were precipitated as microspheres in the inhalable size range. The percentage of respirable fraction (RF) suitable for respiratory delivery was protein dependent. Lysozyme could be precipitated as a powder with an RF of 63%, while the maximum RF for albumin was 45%.[25]

The effect of processing technique on the aerosol performance of insulin formulations prepared by DG antisolvent precipitation and spray drying has been investigated. Two types of insulin formulations were produced. They were 1 : 20 insulin–mannitol (IM) and 1 : 20 : 8 insulin–mannitol–citric acid (IMC). Mannitol is an excipient commonly used to improve aerodynamic performance of powders for inhalation delivery. Citric acid has been found effective in improving insulin absorption in the lungs.[35]

Dense gas IM samples were precipitated from dimethylsulfoxide solutions using carbon dioxide as an antisolvent, while spray dried products were generated from insulin suspensions in mannitol aqueous solutions. Dense gas and spray dried IMC samples were produced from aqueous solutions; in the DG antisolvent process carbon dioxide modified with ethanol was used as an antisolvent. The particle size of DG processed formulations was substantially larger than the corresponding spray dried material for both IM and IMC samples: 21.8 vs. 5.71 μm and 11.9 vs. 7.24 μm, respectively, as measured by laser diffraction of the dry powders. The aerosol performance of DG processed IMC formulations, however, was superior to the spray dried counterpart, as measured by impactor studies. In fact, the RF of the DG processed IMC was 47.6%, while spray dried IMC had an RF of 30.0%. The aerosol performance of IM samples followed an opposite trend with the DG product exhibiting an RF of 26.6% and the spray dried of 30.6%.[35]

Insulin bioavailability by intratracheal administration of the different formulations to rats was studied. The absorption of formulations containing citric acid was not influenced by the preparation technique. The DG processed IM samples had a marked improvement of insulin absorption compared with the spray dried material. The difference may have been ascribed to the faster dissolution rate presented by the DG processed material or by reversible changes in the secondary structure of the protein.[35] The reasons for the improved performance of DG processed IM samples were not ascertained. It has to be noted that the in vivo test did not rely on the aerodynamic properties of the powders for their delivery to the absorption site. An overall evaluation of the different formulations that combined aerodynamic properties and in vivo absorption results indicated that the DG processed formulations had the potential to be more effective than the corresponding spray dried products. Particularly interesting is the improvement of lung absorption registered for IM samples upon DG processing. The results indicate that improving the RF of DG

processed insulin is an achievable way to improve insulin absorption in the lungs without resorting to the use of absorption enhancers.[35]

Dense gas antisolvent precipitation processes have been put into operation on pilot and commercial scale for the production of various materials including pharmaceuticals.[13] Pilot scale micronization of the antibiotic amoxicillin from *N*-methylpyrrolidone (NMP) solution was conducted by DG antisolvent precipitation using carbon dioxide as an antisolvent.[36] Amoxicillin is used in the therapy of respiratory tract conditions. The pilot scale process was performed at 40°C and 150 bar processing 0.6 kg/hr of solution and delivering 20 kg/hr of antisolvent to the precipitation chamber. The process yield was about 90%. The large-scale process was able to reproduce the particle morphology obtained on the laboratory scale with high consistency. The effect of operating parameters observed on the bench scale, specifically solution concentration, was accurately replicated on the larger scale. Amorphous spherical particles were produced with average particle size within 0.3 and 1.2 μm, depending on the concentration of solute in the NMP.[36]

Highly consistent results were also obtained in the scale-up of the precipitation of nicotinic acid from methanol from a laboratory rig to a pilot plant. The process, performed at 90 bar and 90°C, produced rectangular prism particles between 4 and 5 μm on both manufacturing scales.[12,37]

The analgesic paracetamol has been processed in a pilot plant from dimethylformamide (DMF), ethanol, and acetone. The plant used had the capacity of 30 kg/hr of CO_2 and 0.6 m^3/hr solution and included an 8 L precipitation vessel. The operating temperature was fixed at 45°C while the pressure was varied between 100 and 200 bar. Particle size and morphology varied mostly with the nature of the solvent. Spherical particles were obtained from DMF and acetone, while needle-shaped particles could be produced from ethanol and acetone. Rhombic particles were produced only from ethanol. Particle size varied between 1 and 10 μm depending on the operating pressure.[38]

The cost of a semicontinuous antisolvent precipitation process on a large scale has been estimated between 47 and 97 Euros/Kg, based on a production of 1–8 tons/yr from a feed containing 5–10 wt% product. The authors concluded that a production in excess of 4 tons/yr would be necessary for the process to be economically viable.[39] The cost of the process has been estimated to be comparable to the cost of spray drying.[12] Overall, large-scale semicontinuous antisolvent precipitation processes appear to be an economically viable alternative for the micronization of high-value-added products.[39]

POLYMORPHISM IN DENSE GAS PROCESSING

Dense gas antisolvent processes are capable of producing pure polymorphic forms of compounds. The significance of this property is that crystal polymorphism is a critical factor in determining the biological activity of pharmaceuticals as well as their performance in delivery devices and the shelf life of drug formulations.[40] Additionally, different polymorphic forms of currently marketed drugs can be considered as new entities and may be patented, thus providing significant financial advantage to the assignee. Polymorphism is a common phenomenon in the pharmaceutical industry and affects the manufacturing of various classes of drugs including barbiturates and steroids.[41]

To overlook the importance of polymorphism for the properties of therapeutic compounds before commercialization can have considerable repercussions. In 1998, the Abbott company was brought into a market crisis because the anti-HIV drug ritonavir—commercialized since 1996 under the name Norvir—failed a dissolution specification.[42] Investigations led to the discovery that an unknown, thermodynamically more stable, and much less soluble, polymorphic form of ritonavir (RII) was crystallized upon storage. It was discovered that traces of RII during processing led to the production of unstable formulations in which the desired polymorph converted to RII. The composition of the drug formulation was adapted to the properties of RII to compensate for its lower solubility; however, the original polymorph was still preferable owing to production issues and, eventually, a selective crystallization technique that produced only the desired polymorph was developed.[42] In this case, the active ingredient in the final formulation was in solution, thus the polymorphic form from which the solution was generated did not impact the efficacy of the preparation. Thus modification of the formulation composition to ensure complete dissolution of both ritonavir polymorphic forms was an acceptable procedure.[42]

In cases where drug formulations containing more than one polymorph are marketed it is required that the composition is fixed in relation to each polymorphic form. The issue is complicated by the conversions between polymorphs in the solid state. It is clear that the conversion of a crystal structure in a more stable polymorph has to be inhibited, to preserve the composition of the drug formulation. In terms of morphological stability, the production of the more stable polymorph of an active pharmaceutical ingredient is the more convenient option; however, other issues may play critical roles.

The manufacturing of amorphous products may be considered because they usually present solubility and

dissolution rates significantly higher than the corresponding crystalline forms. Examples of amorphous drugs that have been accepted in the market are the antihypertensive quinapril HCl (Accupril®), the antiasthma zafirlukast (Accolate®), and the anti-AIDS nelfinavir masylate (Viracept®).[43] However, drug formulations containing amorphous fractions of products generally exhibit very low stability toward various degradation processes such as chemical degradation, particle aggregation, and crystallization so that the production of crystalline material is generally pursued when possible.[41]

Conventional processing usually requires two separate steps to produce a specific polymorph with controlled particle size. They are the crystallization of the polymorph and subsequent particle size reduction. The first step may include a purification stage (i.e., when liquid–liquid antisolvent crystallization is applied and the second step (i.e., milling) may generate amorphous fractions.[44] Crystallization techniques that may be used to produce particulates with high surface area, such as spray drying and freeze-drying, also can generate defective crystals and amorphous fractions.[41]

The RESS process has been used to produce a single polymorphic form of carbamazepine—an anticonvulsant drug—that is accepted by the Food and Drug Administration. Carbamazepine has four polymorphic forms of which only one is approved for commercial formulations. The RESS process generates micrometer particles with mean diameter smaller than 3 μm and a quite narrow particle size. The crystal structure of the product could be controlled by modification of operating pressure and temperature, which were varied between 170 and 200 bar and 35 and 75°C, respectively. All the RESS processed samples contained a higher fraction of amorphous material than raw carbamazepine.[45]

Dense gas antisolvent crystallization has been used to control the ratio of the polymorphic forms generated. The ability of DG processes to produce microparticles of pure polymorphs holds significant advantages compared with conventional processing. Beach et al. reported the precipitation of the antiasthma drug salmeterol xinafoate from acetone and ethanol solutions.[46] Sameterol xinafoate exists in two polymorphic forms (PI and PII). Conventional crystallization techniques generate PI with a small fraction of PII. By modifying the operating temperature and pressure, the authors could control the ratio between PI and PII in the product. Single polymorphic forms were produced by operating at pressures above 250 bar. The material produced was suitable for lung delivery, having a particle size between 2 and 5 μm, and exhibited notable stability.[46]

The antibacterial sulfathiazole presents five polymorphic structures. The modification of the operating conditions in a DG antisolvent process allowed control over product composition so that it was possible to produce three of the polymorphic forms separately. By DG antisolvent precipitation from saturated acetone solutions it was possible to produce amorphous material and two of the polymorphs. The effect of operating temperature was not significant. The formation of the different forms was determined by kinetic factors modified through changes of the solution flow rate. Sulfathiazole was also processed by the same DG antisolvent process from saturated methanol solutions. In methanol systems, the crystal structure of the precipitate was thermodynamically controlled and three of the polymorphic forms were individually produced by changing the operating temperature. The operating temperature (between 0°C and 120°C) was found to be the most important parameter in determining the formation of a certain polymorphic form from methanol solutions while changes in the solution flow rate had a minor effect and could be manipulated to vary the particle morphology. The flow rate of the antisolvent (carbon dioxide) and the operating pressure were 10 ml/min and 200 bar in all the experiments.[47]

Three polymorphic forms of the antiepileptic drug carbamazepine could be produced individually and as mixtures by the DG antisolvent precipitation by using carbon dioxide as an antisolvent. The parameters controlling the polymorphism of carbamazepine were the solvent (dichloromethane or methanol), temperature (within 45°C and 85°C), pressure (within 80 and 250 bar), and the level of supersaturation reached in the precipitation chamber (modified through changes of experimental pressure, temperature, solvent flow rate). In all experiments solution concentration and antisolvent flow rate were kept constant. Particle size and morphology were highly dependent on the polymorph form, which varied between needle-like particles over 60 μm long and plate-like particles about 10 μm wide.[48]

A modification of DG antisolvent precipitation, involving introducing water with the DG antisolvent, was developed to promote the formation of crystalline rather than amorphous products. The water and the antisolvent should be miscible and the mixture is used to expand the primary solvent to trigger particle precipitation. The same antisolvent/water stream can be used to condition the already formed particles after the flow of solution in the precipitation chamber is terminated. The process has been applied to the production of crystalline formoterol fumarate, an antiasthma drug. Dense gas antisolvent precipitation of formoterol fumarate from methanol solutions using carbon dioxide as an antisolvent generated totally amorphous particles. By introducing water to the carbon dioxide stream and further exposing the particles thus formed to the carbon dioxide–water flux, it was possible to generate the dihydrated formoterol fumarate polymorph. It has been claimed that the technique is suitable for processing a wide range of pharmacologically active ingredients,

excipients, and their combinations.[49] The process has been extended to the control of the polymorphic structure of preformed compounds, particularly to convert amorphous forms to crystalline. Application to the conditioning of amorphous lactose (120 bar and either 40°C or 70°C) resulted in substantial conversion to crystalline material as observed by thermal activity monitoring. The particle size before and after the process was monitored through Coulter counter analysis and was found to be preserved.[50] However, due to the hygroscopic nature of many amorphous compounds, resulting in substantial particle agglomeration upon exposure to humidity, the process is unlikely to generally preserve the original particle morphology. In particular, particle size measurements based on aerodynamic performance are likely to produce substantially different results before and after powder processing.

CONCLUSIONS

The unique properties of DGs, including tuneable solvation power and excellent mass transfer properties, are advantages that are beneficial in a wide range of processes applied to pharmaceutical compounds. Dense gas technology commonly utilizes carbon dioxide, which is nonflammable, nontoxic, and inert, and has moderate critical conditions. The properties of carbon dioxide are particularly suited to the processing of labile compounds such as many active pharmaceuticals.

The potential of DG technology for the engineering of micronized particles, with control over critical physical properties such as particle size and morphology and polymorphism, has been demonstrated. However, the widespread application of DG technology on the commercial scale is complicated by an incomplete understanding of the particle formation mechanisms and the consequent limited predictive ability. Additional issues are related to the relatively high implementation costs. Nevertheless, various DG processes have been successfully applied to the manufacture of compounds on the commercial scale. Further developments are expected from the improving knowledge of current DG techniques and fundamental studies in the area.

REFERENCES

1. Charoenchaitrakool, M.; Dehghani, F.; Foster, N.R.; Chan, H.K. Micronization by rapid expansion of supercritical solutions to enhance the dissolution rates of poorly water-soluble pharmaceuticals. Ind. Eng. Chem. Res. **2000**, *39* (12), 4794–4802.
2. Turk, M.; Hils, P.; Helfgen, B.; Schaber, K.; Martin, H.-J.; Wahl, M.A. Micronization of pharmaceutical substances by the rapid expansion of supercritical solutions (RESS): a promising method to improve bioavailability of poorly soluble pharmaceutical agents. J. Supercrit. Fluids **2002**, *22* (1), 75–84.
3. Subra, P.; Debenedetti, P. Application of RESS to several low molecular weight compounds. Process Technol Proc. **1996**, *12*, 49–54.
4. Alessi, P.; Cortesi, A.; Kikic, I.; Foster, N.R.; Macnaughton, S.J.; Colombo, I. Particle production of steroid drugs using supercritical fluid processing. Ind. Eng. Chem. Res. **1996**, *35* (12), 4718–4726.
5. Ohgaki, K.; Kobayashi, H.; Katayama, T.; Hirokawa, N. Whisker formation from jet of supercritical fluid solution. J. Supercrit. Fluids **1990**, *3* (3), 103–107.
6. Pathak, P.M.; Meziani, M.J.; Desai, T.; Sun, Y.-P. Nanosizing drug particles in supercritical fluid processing. J. Am. Chem. Soc. **2004**, *126* (35), 10842–10843.
7. Kim, J.-H.; Paxton, T.E.; Tomasko, D.L. Microencapsulation of naproxen using rapid expansion of supercritical solutions. Biotechnol. Prog. **1996**, *12* (5), 650–661.
8. Debenedetti, P.G.; Tom, J.W.; Yeo, S.; Lim, G.B. Application of supercritical fluids for the production of sustained delivery devices. J. Controlled Release **1993** *24* **(1–3)**, 27–44.
9. Tom, J.W.; Debenedetti, P.G. Formation of bioerodible polymeric microspheres and microparticles by rapid expansion of supercritical solutions. Biotechnol. Prog. **1991**, *7* (5), 403–411.
10. Matsuyama, K.; Mishima, K.; Hayashi, K.-I.; Ishikawa, H.; Matsuyama, H.; Harada, T. Formation of microcapsules of medicines by the rapid expansion of a supercritical solution with a nonsolvent. J. Appl. Polym. Sci. **2003**, *89* (3), 742–752.
11. Mishima, K.; Matsuyama, K.; Tanabe, D.; Yamauchi, S.; Young, T.J.; Johnston, K.P. Microencapsulation of proteins by rapid expansion of supercritical solution with a nonsolvent. Materials, interfaces, and electrochemical phenomena. AiChE J. **2000**, *46* (4), 857–865.
12. York, P. Strategies for particle design using supercritical fluid technologies. Pharm. Sci. Technol. Today **1999**, *2* (11), 430–440.
13. Clavier, J.Y.; Perrut, M. Scale-up issues for supercritical fluid processing in compliance with GMP. In *Supercritical Fluid Technology for Drug Product Development*; York, P., Kompella, U.B., Shekunov, B.Y., Eds.; Marcel Dekker, Inc.: New York, 2004.

14. Rodrigues, M.; Peiriço, N.; Matos, H.; de Azevedo, E.G.; Lobato, M.R.; Almeida, A.J. Microcomposites of theophylline/hydrogenated palm oil from a PGSS process for controlled drug delivery systems. J. Supercrit. Fluids **2004**, *29* (1–2), 175–184.
15. Knez, Z.; Weidner, E. Particles formation and particle design using supercritical fluids. Curr. Opin. Solid State Mater. Sci. **2003**, *7* (4–5), 353–361.
16. Kerc, J.; Srcic, S.; Knez, Z.; Sencar-Bozic, P. Micronization of drugs using supercritical carbon dioxide. Int. J. Pharm. **1999**, *182* (1), 33–39.
17. Jung, J.; Perrut, M. Particle design using supercritical fluids: literature and patent survey. J. Supercrit. Fluids **2001**, *20* (3), 179–219.
18. Foster, N.; Mammucari, R.; Dehghani, F.; Barrett, A.; Bezanehtak, K.; Coen, E.; Combes, G.; Meure, L.; Ng, A.; Regtop, H.L.; Tandya, A. Processing pharmaceutical compounds using dense-gas technology. Ind. Eng. Chem. Res. **2003**, *42* (3), 6476–6493.
19. Adjei, A.L.; Gupta, P.K. Eds. *Inhalation Delivery of Therapeutic Peptides and Proteins; Lung. Biol. Health Dis.* Marcel Dekker: New York, 1997.
20. Hageman, M.J. The role of moisture in protein stability. Drug Dev. Ind. Pharm. **1988.** *14* (14), 2047–2070.
21. Pikal, M.J.; Dellerman, K.M.; Roy, M.L.; Riggin, R.M. The effects of formulation variables on the stability of freeze-dried human growth hormone. Pharm. Res. **1991**, *8* (4), 427–436.
22. Gonda, I. Physicochemical principles in aerosol delivery. In *Topics in Pharmaceutical Sciences*; Midha, K.K., Crommelin, D.J.A., Eds.; Medpharm Scientific: Stuttgart, Germany, 1992.
23. Martonen, T.B.; Bell, K.A.; Phalen, R.F.; Wilson, A.F.; Ho, A. Growth rate measurements and deposition modeling of hygroscopic aerosols in human tracheobronchial models. Ann Occup. Hyg. **1982**, *26* (1–4), 93–108.
24. Newman, S.P.; Clarke, S.W. Therapeutic aerosols 1—physical and practical considerations. Thorax **1983**, *38* (12), 881–886.
25. Bustami, R.T.; Chan, H.-K.; Dehghani, F.; Foster, N.R. Generation of microparticles of proteins for aerosol delivery using high pressure modified carbon dioxide. Pharm. Res. **2000**, *17* (11), 1360–1366.
26. Chattopadhyay, P.; Gupta, R.B. Protein nanoparticles formation by supercritical antisolvent with enhanced mass transfer. AIChE J. **2002**, *48* (2), 235–244.
27. Muhrer, G.; Mazzotti, M. Precipitation of lysozyme nanoparticles from dimethyl sulfoxide using carbon dioxide as antisolvent. Biotechnol. Prog. **2003**, *19* (2), 549–556.
28. Moshashaee, S.; Bisrat, M.; Forbes, R.T.; Nyqvist, H.; York, P. Supercritical fluid processing of proteins. I: lysozyme precipitation from organic solution. Eur. J. Pharm. Sci. **2000**, *11* (3), 239–245.
29. Winters, M.A.; Debenedetti, P.G.; Carey, J.; Sparks, H.G.; Sane, S.U.; Przybycien, T.M. Long-term and high-temperature storage of supercritically-processed microparticulate protein powders. Pharm. Res. **1997**, *14* (10), 1370–1378.
30. Thiering, R.; Dehghani, F.; Dillow, A.; Foster, N.R. The influence of operating conditions on the dense gas precipitation of model proteins. J. Chem. Technol. Biotechnol. **2000**, *75* (1), 29–41.
31. Debenedetti, P.; Tom, J.W.; Yeo, S.D.; Lim, G.B. Application of supercritical fluids for the production of sustained delivery devices. J. Controlled Release **1993**, *24*, 27–44.
32. Debenedetti, P.G; Lim, G.B; Prud'homme, R.K Formation of Protein Microparticles by Antisolvent Precipitation. European Patent 5423141, Nov 13, 1992
33. Lalor, C.B.; Hickey, A.J. Generation and characterization of aerosols for drug delivery to the lungs. In *Inhalation Delivery of Therapeutical Peptides and Proteins*; Akwete, L.A., Gupta, P.K., Eds.; Marcel Dekker: New York, 1997; 261 pp.
34. Sloan, R.; Hollowood, M.E.; Humpreys, G.O.; Ashraf, W.; York, P. Supercritical fluid processing: preparation of stable protein particles. In Proc. of 5th meeting on Supercritical Fluids, Nice, France, 1998; I.N.P.L.
35. Todo, H.; Iida, K.; Okamoto, H.; Danjo, K. Improvement of insulin absorption from intratracheally administered dry powder prepared by supercritical carbon dioxide process. J. Pharm. Sci. **2003**, *92* (12), 2475–2486.
36. Reverchon, E.; De Marco, I.; Caputo, G.; Della Porta, G. Pilot scale micronization of amoxicillin by supercritical antisolvent precipitation. J. Supercrit. Fluids **2003**, *26* (1), 1–7.
37. Rehman, M.; Shekunov, B.Y.; York, P.; Colthorpe, P. Solubility and precipitation of nicotinic acid in supercritical carbon dioxide. J. Pharm. Sci. **2001**, *90* (10), 1570–1582.
38. Kroeber, H.; Teipel, U.; Krause, H. Micronization of organic materials by the PCA process. In Proc. of 4th Meeting on High-Pressure Process Technology and Chemical Engineering, Venice, Italy, Sep 22–25, 2002; The Italian Association of Chemical Engineering.

39. Rantakyla, M.; Aaltonen, O.; Hurme, M. Cost study of a supercritical antisolvent (SAS) production process. In Proc. of 4th International Symposium on High-Pressure Process Technology and Chemical Engineering, Venice, Italy, Sep 22–25, 2002; The Italian Association of Chemical Engineering.
40. Guillory, J.K. Generation of polymorphs, hydrates, solvates, and amorphous solids. In *Polymorhism in Pharmaceutical Solids*; Brittain, H.G., Ed.; Marcel Dekker: New York, 1999.
41. Shekunov, B.Y.; York, P. Crystallization processes in pharmaceutical technology and drug delivery design. J. Cryst Growth **2000**, *211* (1–4), 122–136.
42. Chemburkar, S.R.; Bauer, J.; Deming, K.; Spiwek, H.; Patel, K.; Morris, J.; Henry, R.; Spanton, S.; Dziki, W.; Porter, W.; Quick, J.; Bauer, P.; Donaubauer, J.; Narayanan, B.A.; Soldani, M.; Riley, D.; McFarland, K. Dealing with the impact of ritonavir polymorphs on the late stages of bulk drug process development. Organic Process Res. Dev. **2000**, 413–417.
43. Almarsson, O.; Gardner, C.R. Novel approaches to issues of developability. Curr. Drug Discov. **2003**, 21–24.
44. Tong, H.H.Y.; Shekunov, B.Y.; York, P.; Chow, A.H.L. Characterization of two polymorphs of salmeterol xinafoate crystallized from supercritical fluids. Pharm. Res. **2001**, *18* (6), 852–858.
45. Gosselin, P.M.; Thibert, R.; Preda, M.; McMullen, J.N. Polymorphic properties of micronized carbamazepine produced by RESS. Int. J. Pharm. **2003**, *252* (1–2), 225–233.
46. Beach, S.; Latham, D.; Sidgwick, C.; Hanna, M.; York, P. Control of the physical form of salmeterol xinafoate. Org. Process Res. Dev. **1999**, *3* (5), 370–376.
47. Kordikowski, A.; Shekunov, T.; York, P. Polymorph control of sulfathiazole in supercritical CO2. Pharm. Res. **2001**, *18* (5), 682–688.
48. Edwards, A.D.; Shekunov, B.Y.; Kordikowski, A.; Forbes, R.T.; York, P. Crystallization of pure anhydrous polymorphs of carbamazepine by solution enhanced dispersion with supercritical fluids (SEDS^{TM}). J. Pharm. Sci. **2001**, *90* (8), 1115–1124.
49. Bisrat, M.; Moshashee, S.; Nyqvist, H.; Demirbuker, M.A. Preparation of Crystalline Drug Particles in Solvated Forms. WO20000306132000, Nov 22, 1999.
50. Bisrat, M.; Moshashee, S.; Nyqvist, H.; Demirbuker, M.A. Process for Producing Drug Particles with a Converted Amorphous and/or Metastable Crystalline Region into Crystalline State. WO2000030614, Nov 22, 1999.

Propylene Production

Abdullah M. Aitani
King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

INTRODUCTION

Propylene is one of the main building blocks for petrochemicals and for clean fuel alkylate blends. It is used in the production of a wide variety of petrochemical products such as polypropylene, acrylonitrile, cumene, oxo-alcohols, propylene oxide, acrylic acid, isopropyl alcohol, and polygas chemicals. Polypropylene accounts for about half of the world propylene consumption, which consequently drives the demand. Other uses of propylene within a refinery include alkylation, catalytic polymerization, and dimerization for the production of high-octane gasoline blends. In general, propylene is supplied in three separate quality grades: refinery (~70%), chemical (~92 to 96%), and polymer (99.6%).

The two major commercial sources of propylene are ethylene steam cracker plants and refinery fluid catalytic cracking (FCC). Other routes include olefin metathesis and propane dehydrogenation. About 68% of world propylene is produced as coproduct to ethylene by the steam cracking of hydrocarbons and about 29% is produced as coproduct to gasoline in the FCC of gas oils. The remaining 3% is produced via propane dehydrogenation and olefin metathesis.[1,2]. Other emerging technologies include catalytic cracking of higher hydrocarbons, olefin interconversion, and methanol conversion. Table 1 presents a summary of world propylene sources and production data.

Current global production of propylene stands at about 54 million metric tons per year (tpy) and is valued roughly at $17 billion.[2] The bulk of propylene production and consumption is concentrated in North America, Western Europe, and Japan. These areas represent about 68% of world capacity and 70% of demand. Propylene demand is expected to grow fast and to nearly double in the next 10 yr, reaching more than 91 million tons by 2010 at a growth rate of 4.7%/yr.[2] Because world consumption is forecast to grow faster than production capacity, propylene has been termed as "olefin of the future."[1] This increase is driven by the demand for derivatives, especially polypropylene.

STEAM CRACKING OF HYDROCARBONS

The main route for producing light olefins, especially ethylene and propylene, is the steam cracking of hydrocarbons, which was commercialized in the 1950s. Currently, world capacity of ethylene steam crackers stands at about 112 million tpy and the amount of propylene coproduced in these steam crackers is about 36 million tpy. The capacity and the throughput of steam crackers are continuously growing due to the ever-increasing world demand for polymers (polyethylene and polypropylene in particular) and other olefin-based derivatives. Ethane, LPG, and naphtha are the dominant steam cracker feedstocks. Natural gas condensate is abundant in North America and the Middle East, while naphtha is commonly used in Asia and Europe.

Steam cracking reactions are principally bond breaking and a substantial amount of energy is needed to drive the reaction toward olefin production. The reaction is highly endothermic; so it is favored at higher temperatures and lower pressures. Superheated steam is used to reduce the partial pressure of the reacting hydrocarbons. It also reduces carbon deposits that are formed by the pyrolysis of hydrocarbons at high temperatures. Long chain hydrocarbons crack more easily than shorter chain compounds and require lower cracking temperatures.

A typical steam cracker consists of several identical pyrolysis furnaces in which the feed is cracked in the presence of steam as a diluent.[3] The cracked gases are quenched and then sent to the demethanizer to remove hydrogen and methane. The effluent is then treated to remove acetylene, and ethylene is separated in the ethylene fractionator. The bottom fraction is separated in the de-ethanizer into ethane and C_3^+, which is sent for further treatment to recover propylene and other olefins. Typical operating conditions of ethane steam cracker are 750–800°C, 1–1.2 atm, and steam/ethane ratio of 0.5. Liquid feeds are usually cracked at lower residence time and higher steam dilution ratios compared to gaseous feeds. Typical conditions for naphtha cracking are 800°C, 1 atm, steam/hydrocarbon ratio of 0.6–0.8, and a residence time of 0.35 sec. Liquid feedstocks produce a wide spectrum of coproducts including BTX aromatics that can be used in the production of variety of chemical derivatives.

In the cracking furnaces, the achievable propylene to ethylene ratio (P/E) is limited to a value of about 0.65, because at higher ratios the total olefin yield

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120037901

Table 1 Sources of propylene and world production data

Propylene source	World production in 2002 (million tpy)	Share (%)	Annual growth for 2002–2015 (%)
Steam crackers	35.9	68.0	4.3
Refinery FCC units	15.3	29.0	5.0
Metathesis/dehydrogenation	1.6	3.0	6.5
Total	52.8	100.0	4.7

(From Refs.[1,2].)

(i.e., the sum of ethylene and propylene) drops to an uneconomically low level. Due to the higher annual growth rates of propylene compared to ethylene, the highest possible P/E in steam crackers is very desirable. Table 2 shows typical propylene yields from steam cracking of different hydrocarbon feedstocks, and how the C_5^+ liquid byproducts increase dramatically with heavier feeds.[4]

FLUID CATALYTIC CRACKING

Fluid catalytic cracking units account for about 97% of propylene produced at the refineries as a coproduct to gasoline. Delayed cokers and visbreakers produce the remaining 3%. In some refineries, a fraction of the propylene and all light gases are normally diverted into sweet fuel gas. About 60% of FCC propylene is used in chemical production and the remaining is used in the production of high-octane gasoline blends. Fluid catalytic cracking is the largest refining process for the production of gasoline with a global capacity of more than 14.2 million bbl/d or 715 million tpy; about 50% of this capacity is in North America alone. Fluid catalytic cracking converts heavy oil feedstocks such as vacuum gas oil, residues, deasphalted oil into lighter products that are largely olefinic for light fractions and strongly aromatic for heavy fractions.[5]

A schematic flow diagram of a conventional FCC unit is shown in Fig. 1. The main FCC sections comprise feed injection system, reactor (riser), stripper, fractionator, and regenerator. A fluidized catalyst system is used to facilitate catalyst and heat transfer between the reactor and the regenerator. Cracking reactions are endothermic; the heat balance is obtained by the combustion of catalyst-deposited coke in the regenerator. In general, all cracking reactions are characterized by the production of appreciable amounts of corresponding olefins. The propylene yield from FCC unit is a function of the following parameters:

- Processing capacity of the FCC unit;
- Type of feedstock;
- Riser reactor outlet temperature (severity); and
- Fluid catalytic cracking catalyst type and additives.

To achieve higher outputs of light olefins, particularly propylene, the hydrogen content of the feedstock must be increased and the sulfur content reduced. This can be achieved by the utilization of low sulfur crude oils or using a higher performance feed hydrotreater upstream of the FCC unit. There are several commercial FCC processes currently employed with major differences in the method of catalyst handling.

Currently, it is not economical to build an FCC for on-purpose propylene production because of low

Table 2 Product yields from steam cracking of various hydrocarbons

	Gaseous feeds			Liquid feeds	
Product yield (wt% on unit)	Ethane	Propane	Butanes	Naphtha	Gas oil
H_2 and methane	13	28	24	26	23
Ethylene	80	45	37	30	25
Propylene	1.11	14.0	16.4	14.1	14.4
Butadiene	1.4	2	2	4.5	5
Mixed butenes	1.6	1	6.4	8	6
C_5^+	1.6	9	12.6	18.5	32
Propylene/ethylene (wt/wt)	0.03	0.3	0.5	0.4	0.6
Propylene (wt% of C_3)	86.7	58.3	99.0	98.3	96.7

(From Ref.[3].)

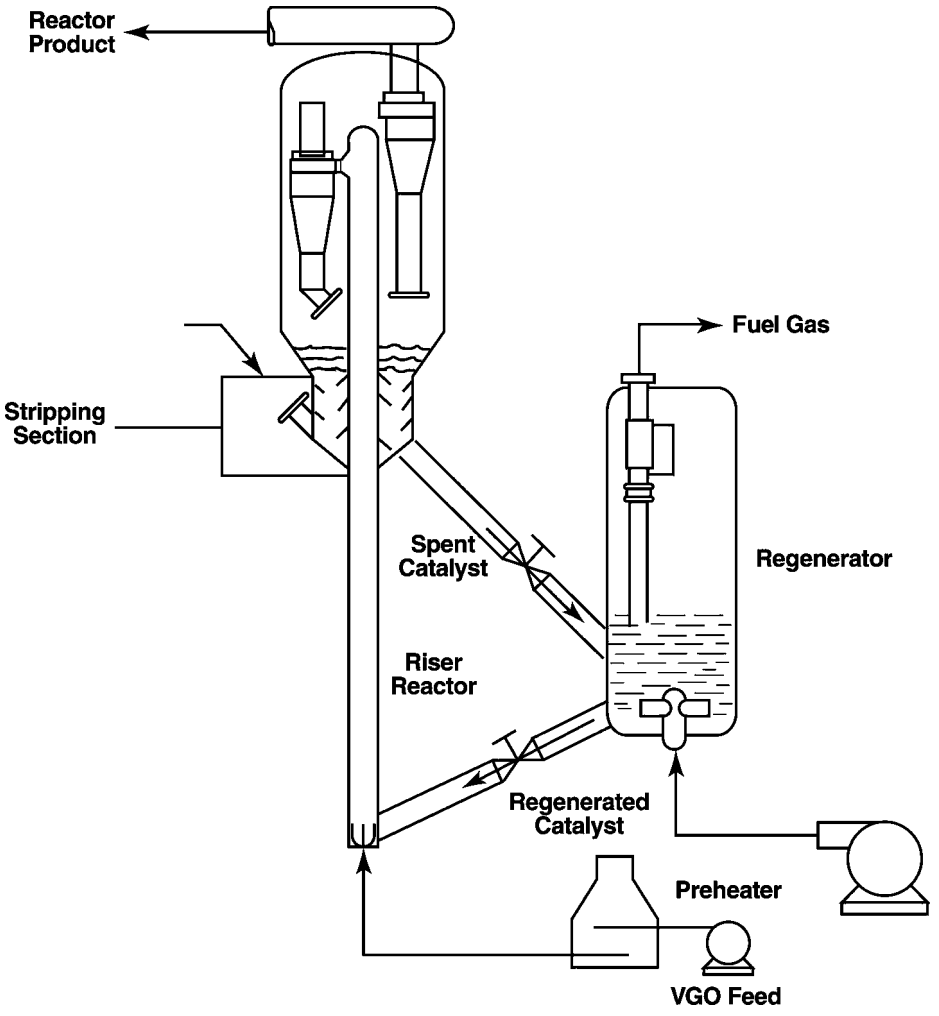


Fig. 1 Schematic flow diagram of a conventional FCC process.

propylene yield; however, it is economical to recover propylene from an FCC unit for chemical uses. Deep catalytic cracking, licensed by Stone & Webster/Shaw Shaw and RIPP/Sinopec, is the only high-olefin FCC process that reached the commercialization stage.[5] Recent catalyst developments have helped increase the propylene yield of conventional FCC from 4.5% to 10% or greater. In particular, the use of ZSM-5 catalyst additive is increasing to boost propylene production. Processes in various stages of development include PetroFCC (UOP), High Severity-FCC (KFUPM, JCCP, Saudi Aramco), Maxofin (KBR, ExxonMobil), Selective Component Cracking (Lummus), and IndMax (Indian Oil). Table 3 presents a comparison of the yield structure of conventional FCC and selected high-olefin technologies.[5–7]

Table 3 Product yields of conventional and emerging FCC processes

Parameter	FCC	DCC	PetroFCC	HS-FCC
Reaction temperature (°C)	500	530	590	600
Product yield (wt%)				
Ethylene	1.5	5.4	6.0	2.3
Propylene	4.8	14.3	22.0	15.9
Mixed butanes	6.9	14.7	14.0	17.4
Gasoline	51.5	39.0	28.0	37.8
Heavy and light oils	21.0	15.6	14.5	9.9
Coke	4.5	4.3	5.5	6.5

(From Refs.[5–7].)

Table 4 Typical properties of commercial propane dehydrogenation processes

Process	Licensor	Reactor type	Catalyst	Reaction conditions
Catofin	Lummus-Houdry	Fixed bed	Cr_2O_3/Al_2O_3	560–620°C, > 0.5 atm
FBD-3	Snamprogetti-Yarsintez	Fluidized bed	Cr_2O_3/Al_2O_3	540–590°C, 1 atm
Oleflex	UOP	Moving bed	Pt/Al_2O_3	550–650°C, > 1 atm
PDH	Linde-BASF-Statoil	Fixed bed	Cr_2O_3/Al_2O_3	590°C, > 1 atm
STAR	Krupp Uhde	Fixed bed	$Pt/Sn/Zn/Al_2O_3$	500°C, 3.5 atm

(From Ref.[8].)

PROPANE DEHYDROGENATION

Propane dehydrogenation is a highly endothermic process. High temperatures and relatively low pressures are used to get a reasonable conversion of propane. The reaction is equilibrium limited. The amount of olefin in the reactor effluent is dependent on the reactor outlet conditions. Thermal cracking reactions limit the maximum practical temperature, and pressure, therefore, becomes the dominant variable.

$$C_3H_8 \Rightarrow C_3H_6 + H_2 \quad \Delta H_{r,298\,K} = 129.4\,kJ/mol$$

Side reactions that occur simultaneously with the main reaction cause the formation of some light and heavy hydrocarbons that result in the deposition of a small amount of coke on the catalyst. Two different catalyst systems based on chromium and platinum are used within a temperature range of 500–650°C. Because of the fast deactivation by coke formation, different concepts have been used to enable regeneration of the catalyst.

Several commercial processes have been developed for the catalytic dehydrogenation of propane to propylene as presented in Table 4.[8] Of the seven commercial propane dehydrogenation plants in operation, six use UOP's Oleflex continuous moving-bed process. The other uses ABB Lummus' Catofin cyclic multiple-reactor system. Other processes include Krupp Uhde's STAR process, as well as technologies from Linde and Snamprogetti.[9]

Fig. 2 presents a typical schematic flow diagram for propane dehydrogenation. In a process with fixed-bed reactors, at least two reactors must be used so that the catalyst in one reactor can be regenerated without stopping the process. The advantage with a moving bed or a fluidized bed is that the catalyst can be continuously removed from the reactor and be regenerated. The disadvantage is that a separate regeneration unit is needed.

METATHESIS

Olefin metathesis is a useful reaction for the production of propylene from ethylene and butenes using certain transition-metal compound catalysts. The two main equilibrium reactions that take place simultaneously are metathesis and isomerization. Metathesis transforms the carbon–carbon double bond, a functional group that is unreactive toward many reagents that react with many other functional groups. New carbon–carbon double bonds are formed at or near room temperature even in aqueous media from starting materials. Because olefin metathesis is a reversible reaction, propylene can be produced from ethylene and butene-2. Metathesis can be added to steam crackers to enhance the production of propylene by the transformation of ethylene and the cracking of mixed butenes. Fig. 3 shows a schematic flow diagram of a typical metathesis process. Examples of metathesis

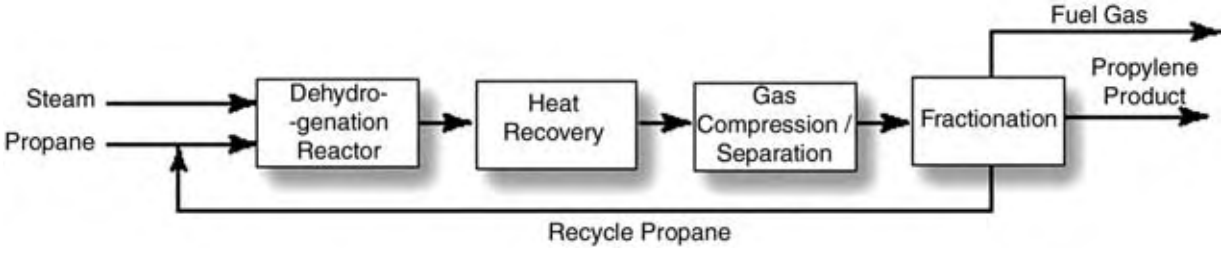


Fig. 2 Flow diagram of catalytic dehydrogenation of propane to propylene.

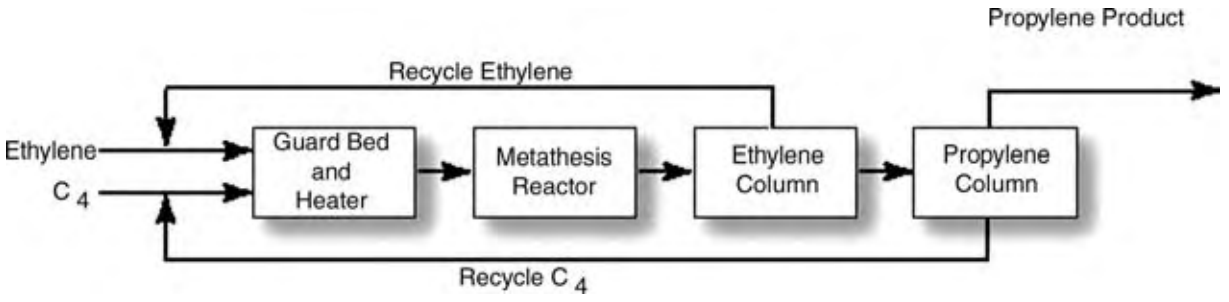


Fig. 3 Typical flow diagram of olefin metathesis process.

processes include the low temperature Meta-4 by Axens and the Olefins Conversion Technology (OCT) by ABB Lummus (acquired from Phillips).

EMERGING TECHNOLOGIES

Catalytic Cracking of Higher Hydrocarbons

A number of nonconventional technologies such as the catalytic cracking of naphtha are being explored as an alternative conventional steam cracking of hydrocarbons. The main objective is to achieve higher yields of light olefins and to lower capital and operating costs. Vniios has developed a conceptually new technology for the catalytic cracking (pyrolysis) of petroleum fractions using heterogeneous catalysts. The process key element is a potassium–vanadium catalyst doped with boron oxide and supported on a robust mullite–corundum carrier. Despite numerous patents and demonstration plant trials, none of these processes have been used in commercial olefin production. However, it is unlikely that these processes will replace steam cracking for ethylene production in the foreseeable future.

Olefin Interconversion

Several on-purpose propylene production technologies are based on the catalytic interconversion of C_4–C_8 olefins in a fixed- or fluidized-bed reactor configuration. The process can be added to conventional steam cracking plants or FCC units. All developed processes use shape-selective medium-pore zeolitic catalysts that crack olefin-rich streams to predominately ethylene and propylene. The fluidized-bed technologies are based on a reactor/regenerator design similar to conventional FCC. Reaction thermodynamics determine the product slate and the selectivity independent of feedstock sources. Lurgi developed a low-pressure catalytic process known as Propylur that uses adiabatic fixed-bed technology for converting higher olefins to propylene and ethylene in the presence of steam. Other technologies include Atofina/UOP's olefin cracking process (OCP), Lyondell/Kellogg's fluidized-bed Superflex process, Linde's fixed-bed catalytic cracking (FBCC), and ExxonMobil's olefin interconversion (MOI) and fluidized-bed propylene catalytic cracking (PCC).

Propylene from Methanol

Because the market for olefins currently greatly exceeds that for methanol production, olefin production could become an important new outlet for the potentially vast quantities of low-cost methanol. Methanol conversion produces a mixture of ethylene and propylene of various ratios or primarily propylene depending on the process. Currently, there are two processes for the production of propylene from methanol: the first process is methanol to olefin (MTO) process, developed by UOP and Hydro,

Table 5 Selection criteria for on-purpose propylene production processes

Selection criteria	Refinery DCC	Dehydrogenation	Metathesis	Olefin cracking	MTO/MTP
Feedstock	Gas oils	Propane	C_2–C_4 Olefins	C_4–C_8 Olefins	Methanol
Steam cracker integration	No	No	Yes	Yes	No
Investment	Moderate	High	Moderate	Moderate	Very high
Commercial units	Several	Several	One	None	None

(From Ref.[1].)

and the second process is methanol to propylene (MTP) process, developed by Lurgi. Both processes start from converting natural gas into methanol, which is then converted to light olefins. The processes are attractive as part of a complex for the utilization of stranded natural gas reserves.

CONCLUSIONS

Propylene is conventionally produced as a coproduct in steam cracking of naphtha or FCC of gas oil in refineries and to a lesser extent via propane dehydrogenation and olefin metathesis. To meet strong demand for propylene derivatives, research is ongoing to develop new on-purpose technologies such as catalytic cracking of hydrocarbons, olefin interconversion, and methanol conversion. The selection criteria for on-purpose propylene production processes are summarized in Table 5. These processes uncouple propylene production from steam crackers and refineries allowing higher production flexibility. However, they will remain more expensive than the coproduct routes and will continue to provide only marginal production volumes in niche circumstances. Much of the increase in propylene demand is expected to come from steam crackers, followed by refineries and dehydrogenation plants.

REFERENCES

1. Plotkin, J. Propylene: olefin of the future? Presented at the 21st Annual Petrochemical World Trade Conference, Chemical Week and Chemosystems: Houston, Sept, 2003.
2. Walther, M. Refinery sources will fill the future "propylene gap." Oil Gas J. **2003**, *101* (4), 65.
3. Matar, S.; Hatch, L. *Chemistry of Petrochemical Processes*, 2nd Ed.; Gulf Professional Publishing: Boston, 2001; 91 pp.
4. Kniel, L.; Winter, O.; Stork, K. *Ethylene: Keystone to the Petrochemical Industry*; Marcel Dekker: New York, 1980; 184 pp.
5. Letzsch, W. The role of the FCC in the refinery of the future. Presented at Akzo Nobel Technical Seminar on How to Meet Environmental Targets, Akzo Nobel Paris, Nov, 2002.
6. Houdek, J.; Hemler, C.; Pittman, R.; Upson, L. Developing a process for the new century. Petroleum Technology Quarterly **2001**, *6* (1), 141.
7. Ma'adhah, A.; Abul-Hamayel, M.; Aitani, A.; Ino, T.; Okuhara, T. Down-flowing FCC reactor increases propylene and gasoline make. Oil Gas J. **2000**, *98* (33), 66.
8. Farrauto, R.; Bartholomew, C. *Fundamentals of Industrial Catalytic Processes*; Blackie Academic and Professional: London, 1997; 411 pp.
9. Tullo, A. Propylene on demand. Chemical and Engineering News **2003**, *81* (50), 15.

Protein Design

Zhilei Chen
Center for Biophysics and Computational Biology, University of Illinois, Urbana, Illinois, U.S.A.

Huimin Zhao
Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.

INTRODUCTION

Protein design refers to the ability to alter protein structure to achieve the desired protein function. Of the two classes of proteins (binding proteins and catalytic proteins), catalytic proteins, or so-called enzymes, are of particular importance to chemical processing because they can be used as chemical catalysts. Enzymes are able to catalyze a broad range of chemical reactions with exquisite specificity and selectivity (stereo-, regio-, and chemo-). In addition, most enzymes are quite efficient catalysts and operate at mild conditions, resulting in less energy consumption. In the past several decades, enzymes have been increasingly employed to synthesize chemicals and materials in the pharmaceutical, chemical, food and agriculture industries. However, the number and diversity of the applications are modest compared to the total number of enzymes identified so far (~4000 enzymes). One main reason for this discrepancy is that naturally occurring enzymes are often not functionally optimal under process conditions in terms of their activity, stability, specificity, and selectivity. To overcome this limitation, tailor-made biocatalysts must be developed by protein design. This entry discusses the molecular tools of protein design and their applications in engineering naturally occurring enzymes into commercially viable biocatalysts for chemical processing. However, their applications in the development of therapeutic proteins and monoclonal antibodies are not discussed.

PROTEIN DESIGNER'S TOOLBOX

Prior to the advent of recombinant DNA technology, the ability to design proteins was limited to chemical modification methods in which specific residues in a protein are modified at the protein level by chemical agents. Different strategies such as atom replacement and segment reassembly have been used to alter enzyme substrate specificity, activity, cofactor requirement, and stability.[1] These methods can introduce a diverse range of functionality that does not occur in natural proteins. However, because only a few protein residues can be selectively modified chemically and the modifying process is rather tedious and time-consuming, these tools have not been widely used for the development of commercial enzymes.

The advent of recombinant DNA technology and polymerase chain reaction (PCR) technology has greatly changed the landscape of protein design. Numerous powerful protein design techniques have been developed in the past two decades, all of which target the modifications at the DNA level. Consequently, these protein design methods have been classified as genetic methods to distinguish them from the above chemical methods. Apart from this classification, all current protein design methods can be categorized into two general strategies: rational design and directed evolution. Rational design involves rational alterations of selected residues in a protein to cause predicted changes in function. It usually requires detailed knowledge of enzyme structure, function, and catalytic mechanism, which represents a bottom-up approach. In comparison, directed evolution, sometimes called irrational design, mimics the natural evolution process in the laboratory and involves repeated cycles of generating a library of different protein variants and selecting the variants with desired functions. Due to its combinatorial nature, it does not require any detailed structural and functional understanding of the target enzymes. Nonetheless, further characterization of the isolated variants may provide insights into protein structure and function. Thus, directed evolution represents a top-down approach. It should be noted that both rational design and directed evolution have been widely used in protein design.

Rational Design

With the aid of site-directed mutagenesis and the availability of enzyme structures solved by x-ray crystallography, many attempts involving rationally designing proteins were made in the 1980s. Early successful

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120018030

examples such as rational design of tyrosyl-transfer RNA synthetase and β-lactamase demonstrated the power of this approach in understanding enzyme catalysis and engineering enzyme activity and specificity.[2,3] The numerous studies that followed these early examples have led to the formation of a new field–protein engineering.[4] Notable examples include the design of a T4 lysozyme with increased stability by engineering disulfide bonds into the protein, the redesign of a lactate dehydrogenase into a malate dehydrogenase by introducing a single point mutation, and the design of subtilisin enzymes with increased activity and altered pH profiles by engineering surface charge.[5–7] However, owing to the sieving effect of publication (only the successful cases are published), rational design may look more "rational" than it actually is. The truth is that our current ability of rationally designing protein is still rather limited. It became apparent that, owing to the intricate and complex relationship between protein structure and function, rationally introduced mutations often have unexpected disastrous effects on enzyme stability and activity. Thus, to circumvent the unpredictable effects of individual residues, methods that are able to assess mutations at single sites or multiple sites in a large-scale format were developed in the early 1990s. This new trend ultimately led to the establishment of directed evolution as a powerful approach for protein design. The past 10 yr have witnessed many more publications on directed evolution than those on rational design. Nonetheless, with recent advances in genomics, bioinformatics and structural proteomics, our knowledge about proteins is rapidly expanding. Computational techniques are also playing an increasingly important role in protein rational design, especially with the availability of faster and cheaper computers. Thus, it is foreseeable that the ability to rationally design proteins will be significantly improved.

Site-directed mutagenesis

Site-directed mutagenesis is the most powerful and widely used rational design approach, which entails the precise modification of specific residue(s) in a given protein at the DNA level. These residue(s) are typically identified from the three-dimensional protein structure obtained by x-ray crystallography or NMR methods. In the absence of structure, a structural model of the target protein can be built from the three-dimensional structure of a related protein sharing high sequence homology with the target protein using homology-modeling programs such as Insight II (Accelrys Inc., San Diego, CA) and SYBYL®/Base (Tripos, Inc., St. Louis, MO). If these options are not available, sequence analysis programs such as BLAST and CLUSTALW (http://workbench.sdsc.edu/) can be used to identify the residues that are conserved among a family of homologous proteins, and are therefore presumed to be functionally important.

A large number of experimental methods have been devised for site-directed mutagenesis. The methods developed in the early 1980s all involve the use of single-stranded bacteriopbage DNA molecules carrying a target gene and chemically synthesized complementary oligonucleotides containing the desired nucleotide substitutions. After DNA annealing and synthesis, the newly synthesized DNA strand contains the target gene with the desired mutations. Although these methods are conceptually simple, the generation of single-stranded DNA is time-consuming and the mutagenesis efficiency (the frequency of the target genes with mutations) is relatively low (less than 50%). Thus, a few PCR-based mutagenesis methods have been developed to overcome these limitations. One of the most widely used methods is the overlap extension PCR mutagenesis method.[8] As illustrated in Fig. 1, four primers, A-F, A-R, M-F and M-R, are used. A-F and A-R correspond to the beginning and the end of the target gene sequence, respectively. M-F and M-R cover the region where a point mutation is desired. First, two independent rounds of PCRs are performed using the target gene as a template.

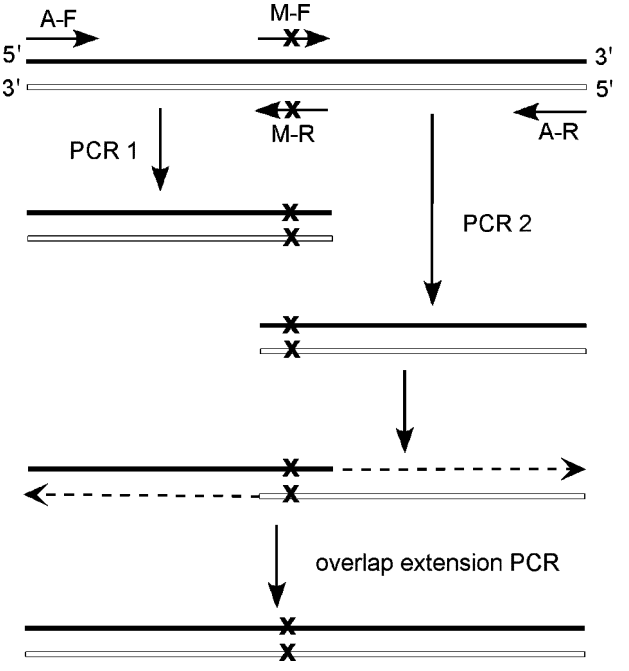


Fig. 1 Site-directed mutagenesis by splicing overlap extension (SOEing). Both the target gene and the PCR products are shown in double strands. Primers are shown as arrows and mutations in primers and products as ×. Dashed arrows indicate the directions of DNA extension by DNA polymerases.

Primers A-F and M-R are used as a pair to amplify the 5′ end of the target gene, and this amplified product will contain a point mutation near the 3′ end. Primers M-F and A-R are used together to amplify the 3′ end of the target gene, leaving a point mutation near the 5′ end of the product. Polymerane chain reaction products from both the reactions are purified and combined for another round of PCR without additional primers. This reaction is called overlap extension because the 3′ end of the first PCR product is complimentary to the 5′ end of the second PCR product, resulting in these two DNA ends priming each other. The resulting target gene with a point mutation will be further amplified by primers A-F and A-R and subcloned into a vector for protein expression.

Domain swapping

Novel protein functions are needed for many applications. Although site-directed mutagenesis is effective in altering protein functions, it often results in incremental improvement of protein functions rather than dramatically improved or novel protein functions because it can only modify protein sequences at single or multiple sites. It is expected that the creation of novel protein functions may require dramatic changes in protein structures. To address this limitation, one simple genetic method to introduce novel protein functions is to combine protein domains from different proteins, or so-called domain swapping. A protein domain is typically defined as a folding and/or functional unit of a protein. Because many large proteins contain several functional domains and domains tend to fold independently, the fusion of domains with different functions may result in a multifunctional protein with unique features.

One recent example of domain swapping is the design of a ligand-regulated DNA recombinase in which the ligand binding domain of human estrogen receptor was linked to the Cre recombinase.[9] The Cre recombinase can integrate a foreign gene into the *LoxP* site on the host genome while the human estrogen receptor is a nuclear hormone receptor that regulates the action of estrogen in different tissues and organs. The human estrogen receptor contains at least three distinct domains: a DNA-binding domain, a transactivation domain, and a ligand binding domain. The binding of estrogenic compounds to the ligand binding domain of the estrogen receptor will cause conformational changes in the estrogen receptor. The Cre recombinase and the estrogen receptor are two completely different proteins, yet the fusion protein created by domain swapping exhibits a protein function different from either parental protein: the recombinase activity is dependent on the ligand binding of the estrogen receptor. Such a fusion protein was used to reduce the toxicity to the proliferating *Drosophila* cells caused by the chronic expression of the Cre recombinase. It should be noted that for a domain swapping experiment to be successful, it is very important to define the exact domain boundaries within a protein. Two main methods have been developed: one is through sequence alignment of a pool of homologous proteins, and the other is through deletion experiments. These two methods usually give a similar but not identical definition of domain boundary. In many cases, because of the lack of complete understanding of protein functions, these slightly different definitions may lead to different results.

Directed Evolution

Although rational design has been demonstrated to be an effective protein design strategy, its success rate is not very high because of our limited understanding of protein folding, structure, function, and dynamics. Moreover, it is rather time-consuming owing to the need of a high-quality three-dimensional protein structure for experimental guidance. In contrast, directed evolution does not use any preconceived ideas about what is important and only relies on a very simple algorithm that nature has successfully been using for eons: diversification coupled with selection. In essence, directed evolution mimics natural Darwinian evolution in a laboratory environment. As shown in Fig. 2, genetic diversity is first introduced into a target gene through random mutagenesis and/or recombination. The library of mutant genes is then transformed into host cells in which the mutant genes are converted into their corresponding proteins. Functionally improved mutant proteins are identified through an appropriate selection or screening strategy. The same process will be repeated until the goal is achieved or no further improvement is possible. It should be noted that the host cells used for protein expression in a directed evolution experiment are laboratory microorganisms such as *Escherichia coli* and *Saccharomyces cerevisiae.* Expression of foreign proteins in these host organisms is enabled by recombinant DNA technology. These microorganisms are favored because of their rapid growth rates, the availability of many genetic engineering tools, and their well-known genetics.

It should also be noted that both screening methods and selection methods have their own advantages and disadvantages. Screening involves physically or chemically interrogating every mutant protein in a library individually, and is often implemented in a 96-well plate format using plate readers. As a result, its throughput is relatively low (the size of the library that can be screened is limited to $\sim 10^4$). However, because the screens are done in vitro with whole cells, cell

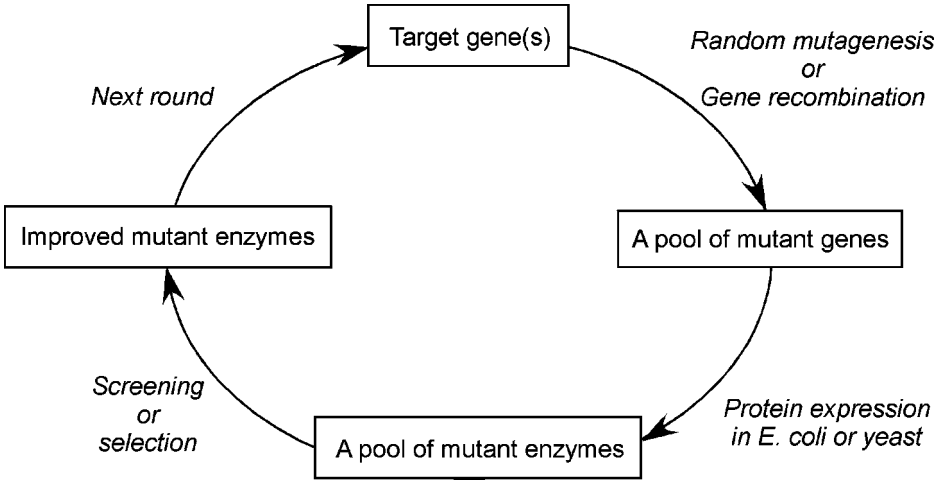


Fig. 2 The general scheme of directed evolution.

lysates, or partially purified enzymes, and often in much the same way the enzymes are traditionally assayed, the experimental conditions can be easily tailored to meet the reaction constraints such as non-natural environment or substrates, and the screens can be implemented very quickly. Moreover, multiple measurements can be made on each sample to check several key enzyme properties. Thus, screening is the most flexible sorting method in directed evolution. In comparison, selection involves linking the survival or growth of the microorganism to the target protein so that only the microorganisms possessing the desired enzyme function can grow. Consequently, a much larger library of enzyme mutants ($>10^6$) can be assessed by selection, and the size of the library is limited only by the cell transformation efficiency. Unfortunately, it is not an easy task to devise a selection method for a given protein in most cases because the desired enzyme function is often nonnatural and cannot be coupled to the growth and survival of the host organism. Even when a selection is available, because of the redundancy and complexity of genetic regulatory networks, host organisms can often create solutions that are not related to the targeted enzyme function. Thus, extra care must be taken to ensure that the positives are indeed the result of the mutations in the targeted enzyme. It is often advised to combine selection and screening if both methods are available.

A successful directed evolution experiment involves two key components: creating genetic diversity and developing a screening or selection method. In the past several years, many experimental methods have been developed to introduce genetic diversity into the target gene, all of which can be grouped into two categories: methods of random mutagenesis and methods of gene recombination. As shown in Fig. 3, random mutagenesis starts from a single parent gene and introduces new nucleotide substitutions randomly in the progeny genes, or inserts or deletes one or more nucleotides at random positions in the progeny genes. In contrast, gene recombination usually starts from a pool of mutants from a single gene or a pool of closely related parental genes of different origin and creates blockwise exchange of sequence information among the parent

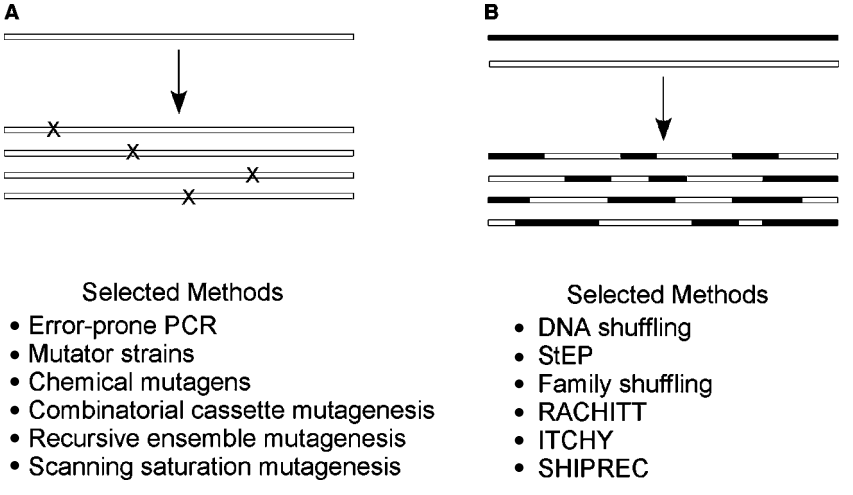


Fig. 3 The comparison between random mutagenesis methods and gene recombination methods. Random mutagenesis methods create a library of variants containing point mutations or insertions/deletions (represented by $\times$) from a single parental gene, whereas gene recombination methods create a library of chimeric variants via blockwise exchange of sequence information among the parental genes. A few representative methods that have been developed so far are listed.

genes. Recombination in a genetic sense means the breaking and rejoining of DNA fragments in new combinations. Both random mutagenesis and gene recombination are important natural evolutionary processes. A few of the most commonly used techniques for generation of diversity will be discussed below. For additional and more detailed discussions on various evolutionary methods, interested readers are referred to a recent review contributed by Zhao and Zha.[10]

Error-prone PCR

Because of its simplicity and efficiency, error-prone PCR is the most widely used random mutagenesis method. It is essentially a variation of the standard PCR with slightly modified reaction conditions. The reaction buffer used for standard PCR contains an equal molar amount of each dNTP and a certain concentration of Mg^{2+} (typically 1.5 mM). Mg^{2+} is very important for the activity of *Taq* DNA polymerase because it is directly involved in catalysis. During the DNA amplification process in a standard PCR, the chance of incorporating wrong nucleotides into the progeny genes by the *Taq* DNA polymerase is very low. However, this very small error rate can be dramatically increased by modifying the reaction buffer so as to force the *Taq* DNA polymerase to incorporate more incorrect nucleotides during amplification. These changes include: 1) use of unbalanced concentrations of dNTPs; 2) addition of $MnCl_2$ in addition to $MgCl_2$; and 3) use of a high concentration of Mg^{2+} (typically 7 mM). Owing to the limitations on library sorting imposed by screening or selection and the observation that most of the mutations are deleterious and beneficial mutations are rare, the mutation rate must be tuned to the power of the sorting method. Thus, only one or two amino acid substitutions are usually introduced to the target protein to increase the possibility of finding mutant proteins with the desired function. Fortunately, it was found that the error rate of error-prone PCR can be easily and precisely controlled by the Mn^{2+} concentration, which makes this method even more appealing.[11] The main disadvantage of error-prone PCR is that it can only access about six amino acid substitutions at a given residue position because of the degeneracy of the genetic code, thus reducing the potential diversity significantly. Moreover, the mutations are not truly random. For example, a common bias of error-prone PCR is the high occurrence of AG substitutions.

DNA shuffling and family shuffling

The method of DNA shuffling, also known as ''sexual PCR,'' is the first and most widely used gene recombination method, developed in 1994 by Stemmer.[12] As shown in Fig. 4, a pool of selected closely related genes containing point mutations are randomly digested with enzyme DNase I to obtain small double-stranded DNA fragments (20–50 bp). These DNA fragments are purified and reassembled into a full-length gene in a PCR-like reaction (without primers). Recombinogenic events occur when fragments derived from different parental genes prime one another. This reassembly mixture is then used as a template for a standard PCR reaction with primers flanking the gene of interest. The final amplified product will consist of a library of full-length genes containing recombined mutations from different parental genes.

It is noteworthy that these closely related genes are often mutants derived from a single parental gene. However, naturally occurring homologous genes sharing relatively high sequence identity (>70%) can also be recombined using DNA shuffling under modified reaction conditions, which is called ''family shuffling.''[13] It has been demonstrated that family shuffling can significantly accelerate the rate of functional enzyme improvement in comparison with

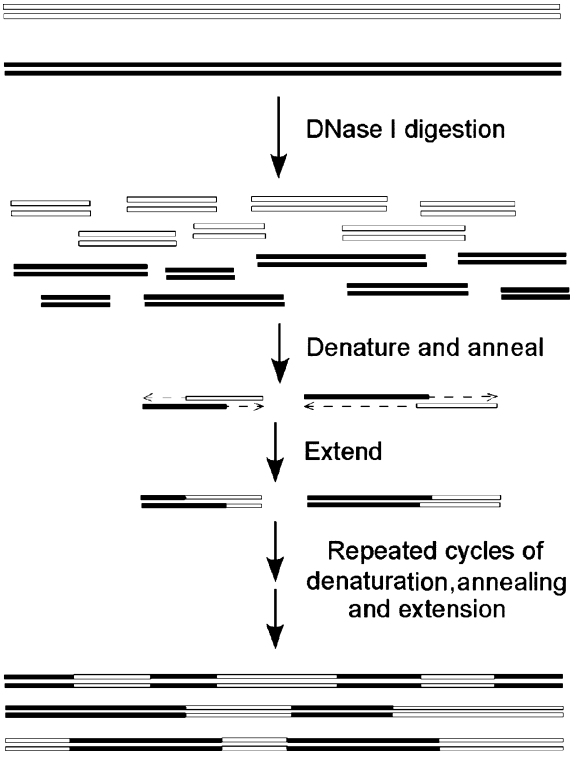


Fig. 4 DNA shuffling. For simplicity, only two homologous genes are shown. These two double-stranded genes are mixed in equal molar ratio followed by random fragmentation with DNase I. These short fragments are reassembled into full-length chimerical genes in a PCR-like process. The full-length genes may be amplified by a standard PCR and subcloned to an appropriate vector.

DNA shuffling or error-prone PCR. The power of family shuffling may arise from its ability to sparsely sample a larger portion of sequence space that is functionally rich because the parental genes have been preselected in nature to be functional and useful. Generally speaking, compared to error-prone PCR, the main advantages of DNA shuffling or family shuffling include its ability to rapidly accumulate beneficial mutations and remove deleterious mutations, and its ability to explore a larger portion of protein sequence space.

Nonhomologous DNA recombination

DNA shuffling or family shuffling relies on relatively high levels of sequence identity (more than 70%) to recombine genes in vitro. DNA shuffling also tends to generate crossovers only in regions of the highest sequence identity. In general, if the sequence identity is less than 70%, most of the shuffled progeny genes will be the same as parental genes. Given that many proteins having similar three-dimensional structures share low or no discernible sequence homology, homology-dependent methods for recombining genes may potentially limit the solutions to protein design problems. In fact, many lines of evidence from rational design and computational studies have indicated that functional proteins can be obtained by recombining genes with low sequence homology.[12]

Recently, several methods have been developed to recombine nonhomologous genes. One of them is the so-called sequence homology-independent protein recombination (SHIPREC) method.[14] As illustrated in Fig. 5, two parental nonhomologous genes of similar size are fused together through a DNA linker containing multiple restriction sites. This dimer gene is digested with DNase I to produce a pool of random-length fragments. Fragments of a length corresponding to either of the parental genes are isolated and treated with nuclease S1 to form blunt ends. These blunt-ended fragments will then be circularized by blunt-end ligation, followed by linearization with restriction digestion in the linker region to create a library of chimerical genes. The chimerical genes are subsequently cloned into an appropriate expression vector and transformed into a suitable host for further screening or selection. This method was used to recombine a membrane-associated human cytochrome P450 (1A2) and the heme domain of a soluble bacterial P450 (BM3), which resulted in variants of 1A2 enzymes that are properly folded and more soluble in the bacterial cytoplasm than the wild-type 1A2 enzymes.[14] One limitation of this method is that only two parental genes are shuffled and there is only one crossover in every chimerical gene.

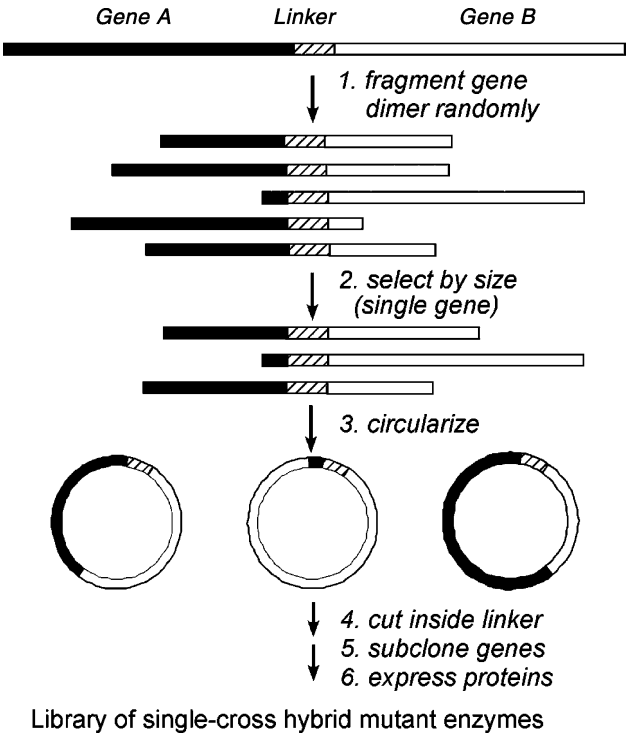


Fig. 5 Schematic representation of the SHIPREC method.

Hybrid Approaches Combining Rational Design and Directed Evolution

Although directed evolution is a powerful tool for protein design, it is limited by the number of protein variants that can be screened experimentally. As shown in Table 1, for a protein of typical size (300 amino acids), the library size of protein variants with each variant containing three mutations is over 3×10^{10}, which is too large to be examined experimentally. However, it is possible that even more simultaneous mutations are needed in a protein to achieve novel protein activity or drastic improvement of protein functions. On the other hand, although rational design has been used for more than two decades, our

Table 1 The potential mutant library size of a protein with 300 amino acids

Mutations	Potential library size[a]
1	5700
2	16, 190, 850
3	30, 557, 530, 900
4	43, 109, 036, 717, 175
5	48, 489, 044, 499, 478, 440

[a]The number of enzyme mutants can be calculated by the simple algorithm $N = 19^M \times 300!/[(300 - M)!M!]$ where M is the number of simultaneous mutations in a protein.

current understanding of protein structure and function still cannot guarantee the success of rational design approaches. Nonetheless, rational design may enable large jumps in the fitness landscape and create protein variants with novel but poor functions, which can then be optimized by directed evolution (Fig. 6). In particular, owing to their high speed, computational techniques may be used to perform in silico screening of protein variants with a library size of $\sim 10^{80}$ or to search vast regions of sequence space to identify the protein sites for more efficient directed evolution.[15,16] Thus, a hybrid approach combining the best of rational design and directed evolution may represent the most powerful approach for protein design.

One such hybrid approach has been recently used to create TEM-1 β-lactamase mutants with increased resistance to the antibiotic cefotaxime.[15] This approach involves a computational protein optimization algorithm called protein design automation (PDA) to reduce the sequence space by many orders of magnitude followed by experimental screening of these selected sequences. The algorithm starts with the structure of a target protein, selects all or a specified set of residues responsible for a specific protein function and predicts the optimal sequences consistent with the proper fold. In this study, 19 residues near the active site of TEM-1 β-lactamase were selected for optimization, which corresponds to a theoretical library of 20^{19} ($\sim 7 \times 10^{23}$) sequences. After computational prescreen by PDA, this library was reduced to 200,000 sequences, which were then constructed and experimentally screened. In a single round, several variants showing a 1280-fold increase in cefotaxime resistance were obtained. These variants contained multiple mutations that have not been discovered before.

EXAMPLES

The applications of enzymes as biocatalysts for industrial chemical processes are rapidly expanding. Like chemical catalysts, the ideal biocatalysts are those enzymes with high stability, activity, specificity, and/or enantioselectivity under process conditions. Unfortunately, naturally occurring enzymes are often deficient in one or more of these important aspects, and therefore cannot meet the stringent requirements of process conditions. In addition, because not all the chemical reactions can be catalyzed by enzymes, owing to either the use of unnatural substrates or the limitations in the chemistry of catalysis, the creation of novel enzyme functions is highly desired. Protein design has been used to address these limitations with great success. Some of these examples will be highlighted below.

Increasing Enzyme Activity and Stability

Both enzyme stability and activity are major concerns in all biocatalytic processes. Poor enzyme stability requires frequent replacement of catalysts while poor enzyme activity requires the use of large amounts of catalysts, both of them resulting in low productivity and unfavorable process economics. Moreover, a high reaction temperature is usually desired in a chemical process because it can result in high reaction rates, favorable equilibrium, and increased substrate solubility, but it also requires high enzyme stability. As naturally occurring enzymes have been evolved for specific biological functions under biological conditions (i.e., aqueous solutions, neutral pH and mild temperatures for most of the organisms), their intrinsic activity and stability are often not high enough for practical applications that may require enzymes to function at extreme pHs, high temperatures, or in organic cosolvents. Thus, further improvement of enzyme activity and stability may be required.

Many successful examples using rational design to engineer enzyme stability or activity have been reported. One of them is the above-mentioned redesign of a T4 lysozyme with increased stability by engineering disulfide bonds into the protein.[5] Another notable example is the design of a subtilisin BPN' mutant containing six site-directed mutations, which was 100-fold more stable than the wild-type enzyme in an aqueous environment and 50-fold more stable than the wild-type in anhydrous dimethylformamide.[17] Subtilisin enzymes have been used as additives in laundry detergents for stain hydrolysis and solubilization, and catalysts for the synthesis of peptides and enantioselective transformation of chiral alcohols, acids, and amines.

Directed evolution is also very effective in engineering enzyme stability and activity. Unlike rational design, which tends to improve one enzyme property at a time (in fact, attempts to rationally alter one enzyme property often disrupt other existing important characteristics), directed evolution may improve multiple enzyme properties simultaneously. For example, five rounds of directed evolution consisting of alternate cycles of error-prone PCR and in vitro gene recombination coupled with screening led to the isolation of a highly stable and active subtilisin E mutant.[11] This mutant contained eight thermo-stabilizing mutations, which were located all over the protein structure. It showed a >200-fold longer thermal inactivation half-life at 65°C, an 18°C higher temperature optima, and a >5-fold higher activity than the wild-type enzyme. Another impressive example is the simultaneous improvement of four distinct enzyme properties of subtilisin, including thermostability, activity in organic solvents, activity at pH 10, and activity at pH 5.5 by

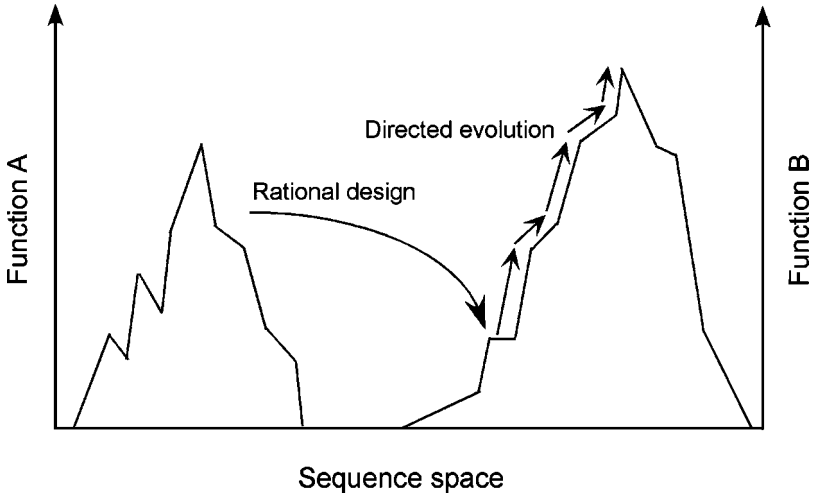


Fig. 6 Schematic representation of hybrid approaches combining rational design and directed evolution. Rational design may allow a large jump in sequence space from high fitness for function A to a low fitness for novel function B. The low fitness for function B might be readily optimized by directed evolution.

directed evolution.[18] Family shuffling was used to recombine 26 homologous subtilisin genes to create a library of chimerical subtilisin genes. Out of 654 active subtilisins, a few mutants showed significant improvement over any of the parental enzymes for each individual enzyme property.

Changing Substrate Specificity

Natural enzymes are optimized for their intended substrates, and often cannot accept nonnatural substrates that are required for a desired chemical process. In the past two decades, both rational design and directed evolution have been successfully used to alter substrate specificity of a large number of common classes of enzymes such as oxidoreductases, hydrolases, and transferases. One exemplary enzyme of particular interest is aspartate aminotransferase which catalyzes the interconversion of aspartate with its corresponding α-keto acid using pyridoxal phosphate as a cofactor. *E. coli* aspartate aminotransferase shares 43% sequence identity with *E. coli* tyrosine aminotransferase. Onuffer and Kirsch [19] used homology modeling to build a structural model of *E. coli* tyrosine aminotransferase based on the solved crystal structure of *E. coli* aspartate aminotransferase.[19] Structural comparison of the active sites of these two enzymes revealed six different positions between them. Mutagenesis of all six residues in aspartate aminotransferase by site-directed mutagenesis of those found in tyrosine aminotransferase successfully altered the substrate specificity of aspartate aminotransferase. The reactivity of the aspartate aminotransferase with phenylalanine was increased by over three orders of magnitude without sacrificing the high transamination activity with aspartate observed for both enzymes. In another case, directed evolution was used to convert *E. coli* aspartate aminotransferase into a valine aminotransferase.[20] A mutant enzyme with 17 amino acid substitutions was generated that shows a 2.1×10^6-fold increase in the catalytic efficiency for a nonnative substrate, valine. The crystal structure of the mutant enzyme indicated a remodeled active site and altered subunit interface caused by the cumulative effects of mutations. Most amazingly, only one of the mutations directly contacts the substrate, which underscores our limited understanding of enzyme substrate specificity. These mutations would be difficult, if not impossible, to be identified and introduced to the mutant enzyme by a rational design approach.

Improving Enantioselectivity

The production of enantiomerically pure compounds is of great importance to the chemical and pharmaceutical industries. Enzymes are chiral catalysts by nature and they have incredible potential for creating enantiomerically pure products. However, many existing natural enzymes show low degrees of enantioselectivity, which requires further improvement by protein design.

Because the molecular basis of enantioselectivity is poorly understood, directed evolution seems to be an excellent choice for engineering enantioselective biocatalysts. Several impressive examples have been documented. In a classical study, Reetz and coworkers[21] used error-prone PCR coupled with a 96-well plate based colorimetric screening method to increase the enantioselectivity of a *Pseudomonas aeruginosa* lipase toward 2-methyldecanoate.[21] After several rounds of directed evolution, the enantioselectivity of the lipase increased from $E = 1.04$ (2% enantiomeric excess) to $E = 25$ (90–93% enantiomeric excess, ee) (E is the enantioselectivity factor). Using a similar approach,

Arnold and coworkers even inverted the enantioselectivity of hydantoinase from D-selectivity (40% ee) to moderate L-preference (20% ee at 30% conversiton).[22] This evolved mutant is now being evaluated in an industrial chemical process at Degussa.

With the increasing knowledge of the molecular basis of enzyme enantioselectivity, rational design has also achieved some success. In one case, Rotticci et al. used molecular modeling to study the different binding modes for alcohol enantiomers in the active site of *Candida antarctica* lipase B and proposed a model for its enantioselectivity.[23] Site-directed mutagenesis was used to alter the active site residues causing unfavorable interactions between the substrate and the enzyme. A single mutation, Ser47Ala, resulted in improvement of the lipase-catalyzed resolution of l-chloro-2-octanol from $E = 14$ to $E = 28$. In another case, van Den Heuvel, Fraaije, and Van Berkel[24] studied the crystal structure of vanillyl-alcohol oxidase and identified a few important residues within the active site that might contribute to the (*R*)-selective formation of (*R*)-l-(4′-hydroxyphenyl) ethanol from 4-ethylphenol. A double mutant was constructed by site-directed mutagenesis, which shows inverted enantioselectivity [(*S*)-selective with 80% ee].

Creating de novo Catalytic Activity

Although natural enzymes can catalyze numerous chemical transformations such as oxidation, reduction, hydrolytic reactions, and carbon–carbon bond formation reactions, enzymes with novel catalytic activities are still needed for the application of enzymes in many industrial biocatalytic processes. The ultimate goal of protein design is to design de novo catalytic activity such that a biocatalyst can be readily obtained for any given chemical transformation. While most protein design so far has really been protein redesign in which the existing protein functions have been adapted under different regimes, a few successful attempts have been made toward this ultimate goal.

An impressive example is the creation of novel enzyme substrate specificity and activity by the DNA shuffling of two highly homologous triazine hydrolases, AtzA and TriA.[25] These two enzymes catalyze the dechlorination and deamination reaction of atrazine and aminoatrazine, respectively. Although they share limited overlap in substrate preference, they only differ by 9 out of 475 amino acids. After one round of DNA shuffling, several variants were found to hydrolyze substrates that were not substrates for either of the parental enzymes.

Another impressive example is the directed evolution of novel ampicillin-resistant activity from a functionally unrelated DNA fragment.[26] A DNA fragment from the genomic DNA library of *Pyrococcus furiosus* was shown to confer very low ampicillin resistance activity (β-lactamase activity) on *E. coli,* while *P. furiosus* itself does not have any β-lactamase activity. This ampicillin resistance activity was significantly enhanced after 50 rounds of DNA shuffling and screening at increasing ampicillin concentrations. The evolved DNA fragments also confer resistance to other drugs that inhibit bacterial cell-wall synthesis.

By taking advantage of the ever-increasing computing power, various computational techniques have been attempted to create de novo protein activity. A particularly impressive example is the creation of catalytic activity in a binding (catalytically inert) protein.[27] Mainly owing to its favorable protein expression properties and thermodynamic stability, rather than any structural similarity to a natural enzyme, *E. coli* thioredoxin was used as a protein scaffold to create enzyme activity concerning histidine-mediated nucleophilic hydrolysis of *p*-nitrophenyl acetate. The design strategy is based on the physical and chemical principles governing protein stability and catalytic mechanism. A protein design software ORBIT was used to perform an active site scan and identified two promising catalytic positions and the surrounding active-site mutations required for substrate binding. Two candidate mutants were constructed by site-directed mutagenesis and both of them showed catalytic activity significantly above the background. Although they are not particularly impressive catalysts, such mutants should be adequate starting points for directed evolution.

CONCLUSIONS

Protein design or engineering is a rapidly growing field of academic research and industrial practice. Its goals include not only addressing fundamental relationships among protein folding, structure, function, and dynamics, but also designing proteins with desired features for applications in pharmaceutical, chemical, agricultural, and food industries. Of particular interest to chemical processing is the application of protein engineering tools to the development of enzyme biocatalysts. Two distinct and yet complementary protein design approaches, rational design and directed evolution, have been successfully developed to engineer biocatalysts with altered or novel stability, activity, substrate specificity, selectivity, cofactor specificity, reaction chemistry, and pH optima. Recently, a third approach combining the best of rational design and directed evolution has also shown great promise in protein engineering, which will attract increasing attention in the near future. With recent advances in

structural genomics and proteomics, and the development of miniaturized and automated high throughput screening technologies, protein engineers will be equipped with more powerful tools to tackle the ever-challenging protein design problems, which will certainly accelerate the widespread adoption of biocatalysts in chemical processing.

ACKNOWLEDGMENT

We thank the National Science Foundation (grant BES-0348107 to HZ) for supporting our work on protein design of human estrogen receptors.

ARTICLE OF FURTHER INTEREST

Biocatalysis, p. 101.

REFERENCES

1. Qi, D.; Tann, C.M.; Haring, D.; Distefano, M.D. Generation of new enzymes via covalent modification of existing proteins. Chem. Rev. **2001**, *101* (10), 3081–3111.
2. Winter, G.; Fersht, A.R.; Wilkinson, A.J.; Zoller, M.; Smith, M. Redesigning enzyme structure by site-directed mutagenesis–tyrosyl transfer–RHA synthetase and ATP binding. Nature **1982**, *299* (5885), 756–758.
3. Sigal, I.S.; Harwood, B.G.; Arentzen, R. Thiol-beta-lactamase—replacement of the active-site serine of RTEM beta-lactamase by a cysteine residue. Proc. Natl. Acad. Sci. USA. **1982**, *79* (23), 7157–7160.
4. Brannigan, J.A.; Wilkinson, A.J. Protein engineering 20 years on. Nat. Rev. Mol. Cell. Biol. **2002**, *3* (12), 964–970.
5. Perry, L.J.; Wetzel, R. Disulfide bond engineered into T4 lysozyme: stabilization of the protein toward thermal inactivation. Science **1984**, *226* (4674), 555–557.
6. Wilks, H.M.; Hart, K.W.; Feeney, R.; Dunn, C.R.; Muirhead, H.; Chia, W.N.; Barstow, D.A.; Atkinson, T.; Clarke, A.R.; Holbrook, J.J. A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework. Science **1988**, *242* (4885), 1541–1544.
7. Russell, A.J.; Fersht, A.R. Rational modification of enzyme catalysis by engineering surface charge. Nature **1987**, *328* (6130), 496–500.
8. Horton, R.M.; Cai, Z.L.; Ho, S.N.; Pease, L.R. Gene splicing by overlap extension: tailor-made genes using the polymerase chain reaction. Biotechniques **1990**, *8* (5), 528–535.
9. Heidmann, D.; Lehner, C.F. Reduction of Cre recombinase toxicity in proliferating Drosophila cells by estrogen-dependent activity regulation. Dev. Genes. Evol. **2001**, *211* (8–9), 458–465.
10. Zhao, H.; Zha, W. *Enzyme Functionality*: *Design, Engineering and Screening*; Svendsen, A., Ed.; Marcel Dekker, Inc.: New York, 2003; 353–373.
11. Zhao, H.; Moore, J.C.; Volkov, A.A.; Arnold, F.H. In *Manual of Industrial Microbiology and Biotechnology*, 2nd Ed.; Demain, A.L., Davies, J.E., Eds.; ASM Press: Washington, DC, 1999; 597–604.
12. Stemmer, W.P. Rapid evolution of a protein in vitro by DNA shuffling. Nature **1994**, *370* (6488), 389–391.
13. Crameri, A.; Raillard, S.A.; Bermudez, E.; Stemmer, W.P. DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature **1998**, *391* (6664), 288-291
14. Sieber, V.; Martinez, C.A.; Arnold, F.H. Libraries of hybrid proteins from distantly related sequences. Nat. Biotechnol. **2001**, *19* (5), 456–460.
15. Hayes, R.J.; Bentzien, J.; Ary, M.L.; Hwang, M.Y.; Jacinto, J.M.; Vielmetter, J.; Kundu, A.; Dahiyat, B.I. Combining computational and experimental screening for rapid optimization of protein properties. Proc. Natl. Acad. Sci. USA. **2002**, *99* (25), 15926–15931.
16. Voigt, C.A.; Mayo, S.L.; Arnold, F.H.; Wang, Z.G. Computational method to reduce the search space for directed protein evolution. Proc. Natl. Acad. Sci. USA. **2001**, *98* (7), 3778–3783.
17. Wong, C.-H.; Chen, S.-T.; Hennen, W.J.; Bibbs, J.A.; Wang, Y.-F.; Liu, J.L.C.; Pantoliano, N.W.; Whitlow, M.; Bryan, P.N. Enzymes in organic synthesis—use of subtilisin and a highly stable mutant derived from multiple site-specific mutations. J. Am. Chem. Soc. **1990**, *112* (3), 945–953.
18. Ness, J.E.; Welch, M.; Giver, L.; Bueno, M.; Cherry, J.R.; Borchert, T.V.; Stemmer, W.P.; Minshull, J. DNA shuffling of subgenomic sequences of subtilisin. Nat. Biotechnol. **1999**, *17* (9), 893–896.
19. Onuffer, J.J.; Kirsch, J.F. Redesign of the substrate specificity of Escherichia coli aspartate aminotransferase to that of Escherichia coli tyrosine aminotransferase by homology modeling and site-directed mutagenesis. Protein Sci. **1995**, *4* (9), 1750–1757.
20. Oue, S.; Okamoto, A.; Yano, T.; Kagamiyama, H. Redesigning the substrate specificity of an enzyme by cumulative effects of the mutations

of non-active site residues. J. Biol. Chem. **1999**, *274* (4), 2344–2349.
21. Liebeton, K.; Zonta, A.; Schimossek, K.; Nardini, M.; Lang, D.; Dijkstra, B.W.; Reetz, M.T.; Jaeger, K.E. Directed evolution of an enantioselective lipase. Chem. Biol. **2000**, *7* (9), 709–718.
22. May, O.; Nguyen, P.T.; Arnold, F.H. Inverting enantioselectivity by directed evolution of hydantoinase for improved production of L-methionine. Nat. Biotechnol. **2000**, *18* (3), 317–320.
23. Rotticci, D.; Rotticci-Mulder, J.C.; Denman, S.; Norin, T.; Hult, K. Improved enantioselectivity of a lipase by rational protein engineering. Chembiochem. **2001**, *2* (10), 766–770.
24. Van Den Heuvel, R.H.; Fraaije, M.W.; Van Berkel, W.J. Direction of the reactivity of Vanillyl-alcohol oxidase with 4-alkylphenols. FEBS Lett **2000**, *481* (2), 109–112.
25. Raillard, S.; Krebber, A.; Chen, Y.; Ness, J.E.; Bermudez, E.; Trinidad, R.; Fullem, R.; Davis, C; Welch, M.; Seffernick, J.; Wackett, L.P.; Stemmer, W.P.; Minshull, J. Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes. Chem. Biol. **2001**, *8* (9), 891–898.
26. Yano, T.; Kagamiyama, H. Directed evolution of ampicillin-resistant activity from a functionally unrelated DNA fragment: a laboratory model of molecular evolution. Proc. Natl. Acad. Sci. USA. **2001**, *98* (3), 903–907.
27. Bolon, D.N.; Mayo, S.L. Enzyme-like proteins by computational design. Proc. Natl. Acad. Sci. USA. **2001**, *98* (25), 14274–14279.

Protein Folding: Biomedical Implications

Ajit Sadana
Department of Chemical Engineering, University of Mississippi,
Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.

Tuan Vo-Dinh
Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.

Nigil Satish Jeyashekar
Department of Chemical Engineering, University of Mississippi, University, Mississippi, U.S.A.

INTRODUCTION

The correct folding of proteins from the inactive folded form to the native (active) folded form has implications in the biotechnology industry, as well as in medical science. In both cases, improper folding results in aggregation that has deleterious results. In the biotechnology industry, the aggregates often lead to a substantial decrease in the yield of the desired active protein form. This can be critical in the economics of the process, especially when the final product is required in a very pure form and is expensive. Novel methods have been explored to unfold the aggregates and then fold them back to the desired and active form. The procedures involved are often expensive and substantially increase the bioprocessing cost and reduce the yield of the desired product. The cost of the downstream processing train can often be as high as 90% of the entire cost of the process, due to the increasing number of steps required.[1] A way to minimize the cost of the downstream process is to minimize the formation of aggregates. The formation of such protein aggregates in the human body leads to pathogenic diseases, some of which are fatal. Proteins misfold to form aggregates at various parts in the human body. Aggregate deposition in the brain lead to Alzheimer's disease, deposition in the knee or elbow joints lead to arthritis, etc. If the causes can be determined for the deposition of such aggregates, then perhaps their deposition could be eliminated, or at least minimized. For this, detailed structural and biochemical information about the protein is required.

Immunoglobulin (Ig) light chains are responsible for quite a few protein deposition diseases. Some immunoglobulin chains form amyloid deposits, whereas others do not. A protein conformational change in the partially folded protein is involved that subsequently leads to association and fibril formation. During the folding pathway of especially larger proteins, hydrophobic patches may be exposed.[2] These hydrophobic patches, especially in the restricted and constrained environments of the cell, may lead to misfolding and aggregate formation. Molecular chaperones minimize the off-pathway events that decrease the yield of the stable and native form of the protein,[3] and assist the unfolded conformation to fold correctly to the native form.

The entry reviews the kinetic approach used to analyze the formation of protein aggregates. The kinetic approach analyzes different folding kinetic mechanisms of proteins that misfold to form pathological aggregates. This would include simple series, series-parallel, and complex mechanisms. The thermodynamics approach describes the search for the structure with the low energy, low entropy, and most stable conformation. This approach facilitates an easy comparison of different mechanisms involved during the slow transition of the normal (healthy) state to pathological state. A few examples are discussed with respect to each mechanism. A unified model for folding is proposed and the kinetics is critically analyzed. A complete understanding of protein folding from the kinetic perspective can help prevent aggregate formation. This can then outline some general principles involved in designing therapeutic strategies to cure such intractable diseases.

PROTEIN FOLDING MECHANISMS

Protein folding mechanisms can be classified into three different categories: series, series-parallel, and complex mechanisms. In general, it is the series pathway that leads to the formation of the native and functional form of the protein. Series-parallel and complex mechanisms often contain ''off-pathway'' steps that generally yield aggregates, which lead to deposition. Such mechanisms that result in the deposition of pathological aggregates and in the elimination of such aggregates are our primary focus.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120022533

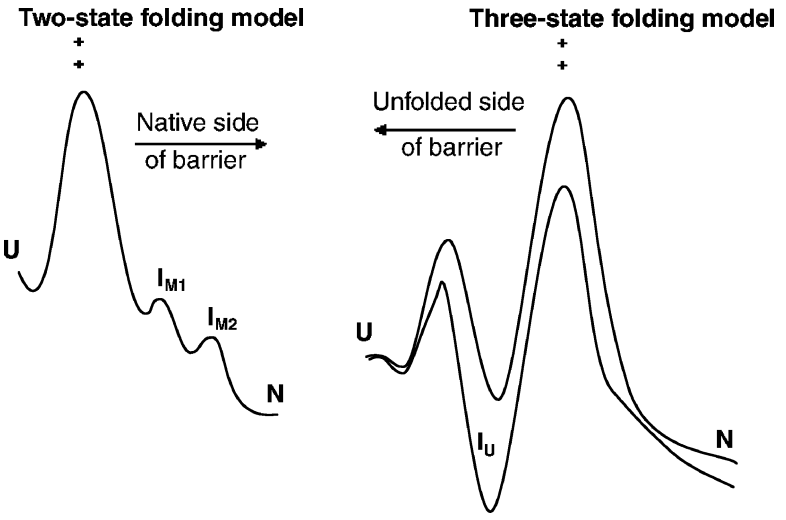


Fig. 1 Comparison of free-energy diagrams of two-state (left) and three-state (right) folding mechanisms. (From Ref.[4].)

Series Mechanisms

The series mechanism of protein folding may or may not involve the presence of stable and detectable intermediates. The following two-state mechanism is involved during the folding of small proteins:

$$U \rightleftharpoons N \tag{1}$$

Here U is the unfolded protein. N is the native form of the protein that exhibits a three-dimensional structure and a specific biological function. This mechanism exhibits the potential to provide the highest yield, and there is minimal possibility of aggregate formation because of the absence of off-pathway steps. An attempt was made to distinguish whether a two-state (no intermediate) or a three-state (defined kinetic intermediate) kinetic process is involved during ubiquitin folding.[4] The results indicate that a two-state folding model is adequate to describe the folding pathway. Free-energy diagrams were used to distinguish between a two-state and a three-state model.

Fig. 1 shows that in two-state folding, intermediates are formed on the native side of the barrier, indicating that the conversion from intermediates to the native states is a fast reaction and that conversion from the unfolded to the first intermediate is the rate-limiting step. Two-state folding behavior requires that any intermediate formed during the folding process, be more stable than the denatured state, as seen in the two-state folding model in Fig. 1 and that the intermediates be located on the native side of the rate-limiting barrier.[4]

In three-state folding, Fig. 1 indicates that the intermediates are formed on the unfolded side of the barrier indicating that the conversion from the intermediate to the native state is the rate-limiting step. The three-state mechanism was used when a detectable intermediate was present. Here "I" is the intermediate form.

$$U \rightleftharpoons I \rightleftharpoons N \tag{2}$$

This mechanism has been observed during the folding kinetics of phage 434 Cro protein.[5] The mechanism provides an insight into the links between the structures, stability, and folding mechanisms involved when small, completely helical proteins fold.

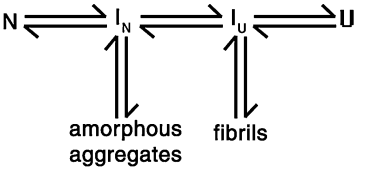
In general, the series pathway may be extended to involve "*n*" stable and detectable intermediates. Here $I_1, I_2, I_3, \ldots, I_n$ are the intermediate states:

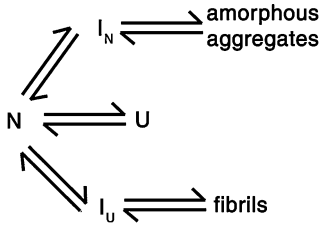
$$U \rightleftharpoons I_1 \rightleftharpoons I_2 \rightleftharpoons I_n \rightleftharpoons N \tag{3}$$

Series-Parallel Mechanisms

Series-parallel mechanisms often include off-pathway steps from the intermediates that sometimes deposit to form aggregates. In some proteins such aggregates are pathological. The biophysical properties and amyloidogenicity of the variable domain of a recombinant amyloidogenic light chain (spinal muscular atrophy, SMA) were analyzed.[6] The mechanistic schemes were proposed in Scheme 1 and Scheme 2.

In both the schemes, the intermediates lead to the formation of amorphous and fibrillar aggregates that causes light chain deposition disease and light chain





(or) AL amyloidosis (amyloid fibrils characterized by the presence of immunoglobulin light chains), respectively. Hence different forms of aggregates formed due to misfolding of the same protein may lead to different diseases.

Light chain or AL amyloidosis is a pathological condition arising from systemic extracellular deposition of monoclonal immunoglobulin light chain variable domains in the form of insoluble amyloid fibrils, especially in kidneys. Amyloid fibril formation from native proteins occurs via a conformational change leading to a partially folded intermediate conformation, whose subsequent association is a key step in fibrilization. Hence, partially folded intermediates are critical precursors for both light chain amyloid fibrils and amorphous aggregates.

Amyloid Aggregates

Major classes of amyloid diseases, e.g., Alzheimer's disease, Type II diabetes, and Parkinson's, are characterized by deposition of ordered protein aggregates termed amyloidosis.[7] Type II diabetes is an autoimmune disease. As the protein misfolds, it forms insoluble aggregates. The mechanism of amyloid fibril formation, an insoluble aggregate, is shown in Fig. 2.

Native state (N) is formed from the unfolded state through intermediate(s) (I). The intermediate(s) acts as a template and propagates to form ordered oligomeric intermediates or protofibrils (PF), which builds up to form stable, insoluble fibrils (F) as shown in Fig. 2. The reaction mechanism involved in this sequence is similar to initiation, propagation, and termination steps in a polymerization reaction. Amorphous aggregates, formed are in equilibrium with the intermediates. In other words, this process is reversible. However, amyloid fibrils are the byproducts of irreversible protein aggregation.

In vivo, the correct assembly of proteins is guided by a family of cellular proteins termed molecular chaperones, e.g., heat shock protein (HSP), nuleoplasmins, and chaperonins. Chaperones bind to the intermediate that tends to aggregate, and either assembles the intermediate to the native state or renders the intermediate void of further reaction to form an aggregate.[8] Normally, all proteins should fold without molecular chaperones. Proteins that tend to form aggregates, like those shown in the above mechanisms, bind to a chaperone to yield the native state.

Prion as Pathological Aggregates

Prion represents a distinct class of infectious agents that are proteinacious and devoid of nucleic acid.[9] The proposed mechanism for prion aggregation is shown in Fig. 3.[10]

Normal protease sensitive form (PrP^c) is formed from the unfolded protein via a prionogenic intermediate.

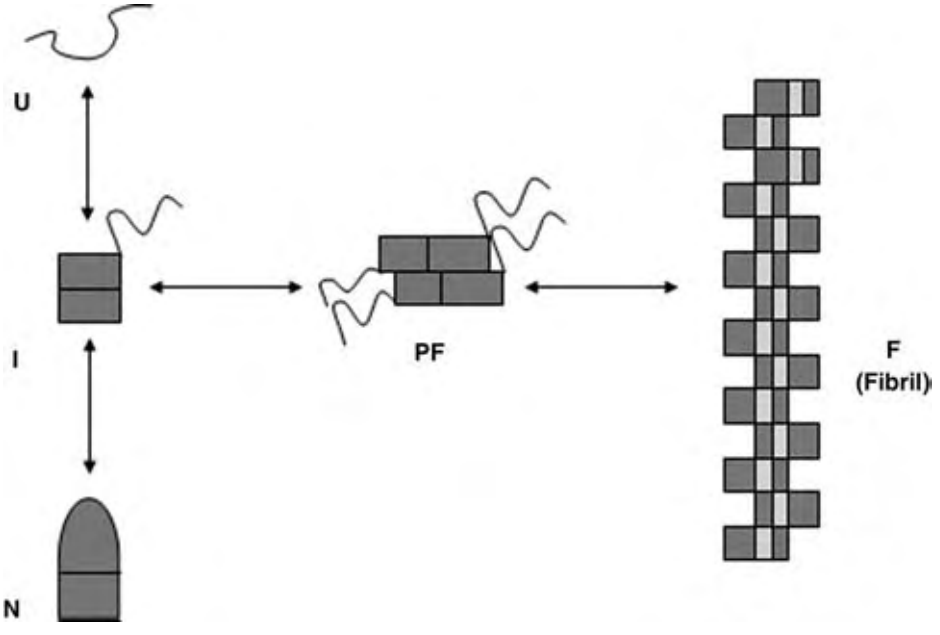


Fig. 2 Mechanism of amyloid fibril formation. (From Ref.[7].)

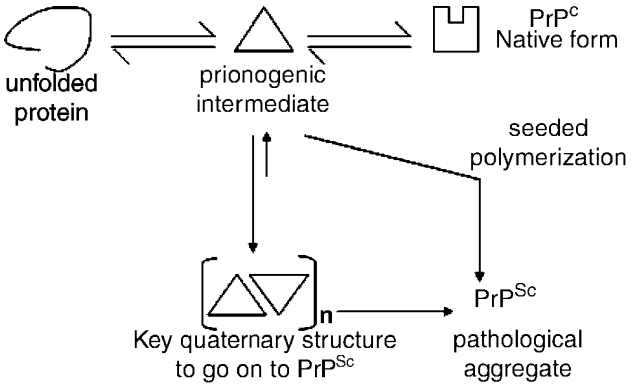


Fig. 3 Folding pathway of prion protein to form scrapie. (From Ref.[9].)

This misfolded form protease-resistant form (PrP^{Sc}) or scrapie may form by

i. Propagation of the prionogenic intermediate to form a template that terminates to form the PrP^{Sc} fibril.
ii. The PrP^{Sc} formed catalyzes the conversion of prionogenic intermediate, as well as the normal form of PrP^{C}. Since scrapie acts as a ''seed'' that catalyzes the conversion to the misfolded form, the conversion is termed as seeded polymerization.

PrP^{C}, which is the native form, acts as a self-chaperone to form PrP^{C}, termed autochaperone, which is similar to autocatalysis in chemical reactions.[11] Since PrP^{Sc} also converts the normal form of the PrP^{C} to the misfolded form, there exists a kinetic competition between these reactions. The model shown in Fig. 4 illustrates such kinetic competition representing the normal folding mechanism, misfolding via prionogenic intermediate, autochaperone folding, and seeded polymerization misfolding. Hence, the key step to prevent aggregate formation is to block the formation of PrP^{Sc} via the prionogenic intermediate step. The model is termed as autochaperone misfolding invasion.

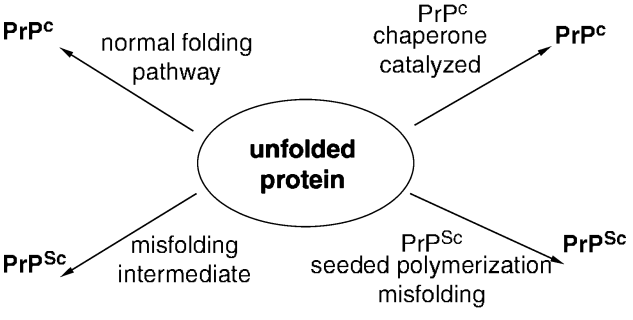


Fig. 4 Kinetic competition in prion folding–autochaperone invasion model. (From Ref.[11].)

The misfolded form, scrapie, results in a class of disease called spongiform encephalopathies, e.g., mad cow disease. Unlike other diseases caused due to protein misfolding, prion diseases are infectious. The infecting agent is the misfolded protein. Prionogenic diseases are transmitted across the species barriers, from cow to sheep to human beings, and are fatal.

Protein Triage Model: Repair Mechanism

The refolding and degradation of proteins and the role of chaperones and proteases in protein folding are analyzed.[12] The ''triage model'' describes the quality control mechanism within the human body. The mechanism describes the fate of an unfolded protein, which may fold either to the native state or form the aggregate. The protein that tends to misfold may bind with the chaperone. This undergoes protein remodeling and is mixed with the pool of unfolded and misfolded proteins as shown in Fig. 5.

A unique repair mechanism is involved. The misfolded protein gets bound to ubiquitin and ATP; a proteasome. A part of the protein-bound proteasome undergoes degradation releasing ATP, while the other part releases the remodeled protein and ubiquitin. This recirculates back to the pool of unfolded and misfolded proteins. Hence, the misfolded protein gets a chance to fold to its native functional form. The mechanism is termed post-translational quality control. The onset of pathological diseases occurs when protein escapes such quality control mechanisms and deposits into extracellular space. Protein aggregation associated with prion and amyloid diseases can be considered biologically relevant failures of post-translational quality control.

Complex Mechanisms

Mechanisms other than series and series-parallel mechanisms are classified under complex mechanisms. The following mechanism was proposed for β(25–35) amyloid-protein interaction:[13]

Amyloid fibrils of the β(25–35) peptide are a cytotoxic fragment of Alzheimer's β-peptide at position 25 and 35. The native, soluble, and folded state that is soluble interacts with the β(25–35) amyloid (step 1 in Fig. 6). The amyloid–protein complex forms aggregates and deposits (step 2). In some cases, the amyloid exerts a force to unfold the protein on the surface of the amyloid fibril (step 3). The unfolded or partially unfolded state has a stronger propensity to aggregate resulting in amorphous aggregates, which adhere strongly to the surface of the amyloid fibril and undergoes irreversible deposition.

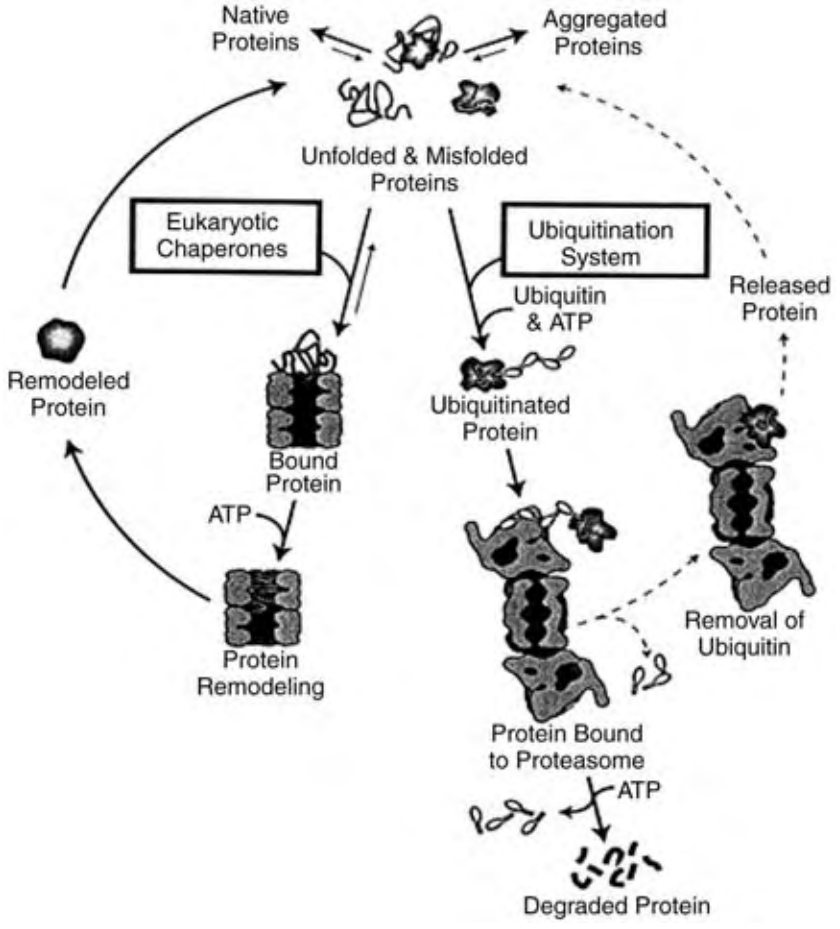


Fig. 5 Protein triage model–repair mechanism. (From Ref.[12].)

The β(25–35) interacts with the native protein, destroys its native conformation, and forms pathological aggregates. Its action is exactly opposite to that of a chaperone, which helps proteins fold correctly. Hence, β(25–35) is termed as an antichaperone. protein disulphide isomerase (PDI), a folding catalyst and chaperone, under certain conditions, can facilitate the misfolding and aggregation of its substrates during lysozyme folding. However, there is not enough evidence to prove that the aggregates are amyloidal in nature.[14] Although aggregation has been viewed as arising from nonspecific interactions, it is clear that specific interactions between the folding intermediates and/or misfolded proteins may be responsible for the

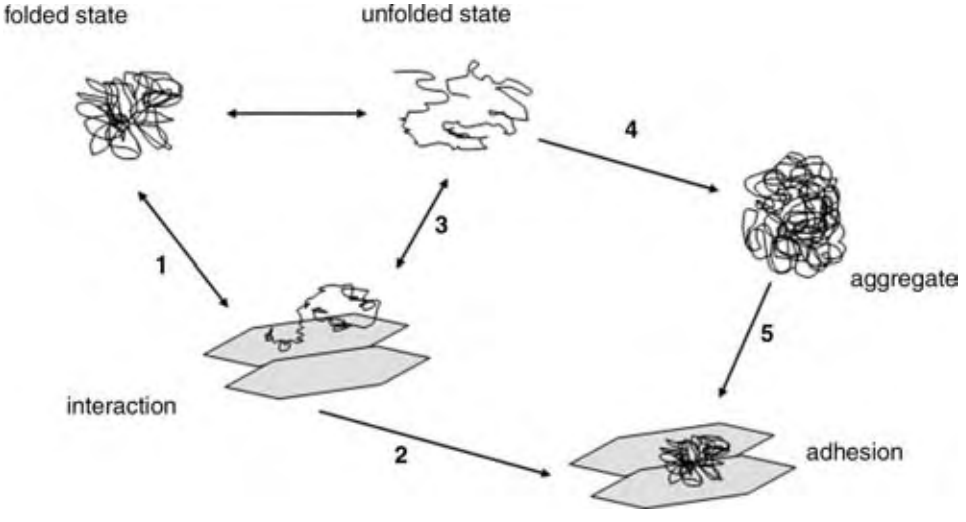


Fig. 6 Complex mechanism–β amyloid–protein interaction. (From Ref.[13].)

multiple interactions leading to the formation of large, insoluble aggregates.

UNIFIED MECHANISM FOR FOLDING

The unified mechanism for folding was presented in view of all the different categories of mechanisms. In most cases of aggregation, there is a kinetic competition between aggregation and other processes such as folding. The environmental conditions and the protein concentration(s) significantly affect the degree and rate of intermolecular association.[15] The mechanism is shown in Fig. 7.

In many cases, when aggregation occurs from a solution of the native protein, it is the partially folded intermediates in equilibrium with the native state that are the immediate precursors of the aggregates.

Longer-lived intermediates are more likely to lead to aggregation, because there is a greater chance of interaction with another such partially folded intermediates. In vivo, the molecular chaperones involved in preventing aggregation may become saturated, and thus there will not be enough free chaperones to bind to the newly synthesized protein, to help it fold to the native state.

Aggregation occurs when hydrophobic surfaces on proteins interact in an intermolecular manner. Three-dimensional propagation of such interaction leads to larger aggregates and deposits when the solubility limit is exceeded. In the unfolded state, the hydrophobic side chains are scattered around, whereas in intermediates there exists large patches of contiguous surface hydrophobicity. Hence, there is greater propensity for the partially folded intermediates to aggregate.

Two basic models have been proposed for amyloid fibril formation from the intermediates.[16] In the nucleation model, the intermediates cluster to form nuclei. Fibril is formed from the nucleus after the nucleus has reached a critical size. Such fibrils add on to its ends to form aggregates. In the polymerization model, intermediate peptide–PF complexes are formed and associate end to end or laterally to form fibrils. The amyloid fibrils also add to the existing PF and deposits, once the solubility limit is reached.

KINETICS OF AGGREGATION

Amyloid Fibrils

The origin of the essentially irreversible formation of fibril is that the peptide adds to the existing fibril, and then it is buried as other peptide(s) is added and the fibril continues to grow. Therefore, while the peptide at the reaction interface shows a reversible reorganization and deorganization with rate constants k_r and k_d, respectively, once the peptide is buried within the fibril, the rate of deorganization k_d is effectively zero.

Amorphous Aggregates

In amorphous deposits, the initial adsorption and desorption steps with associated rate constants k_a and k_d, respectively, are reversible. However, a peptide with the β-peptide deposit is essentially trapped making k_d effectively zero for those molecules that are not at the interface with the solution. Finally, within the amorphous deposit, the "β'' peptide that is not at the reaction interface in contact with the existing fibril has a rate of reorganization k_r, to the β-peptide that is effectively zero.

THERAPEUTIC STRATEGIES

In principle, masking the hydrophobic surfaces on intermediates and monomers can prevent aggregation. This would reduce the rate of aggregate deposition. A system of artificial chaperones (detergents and cyclodextrin), have been successfully used to minimize aggregation.[15] Increasing the stability of the native state prevents the backward reaction and does not favor the formation of the intermediates that lead to aggregate formation. Subsequently, this prevents

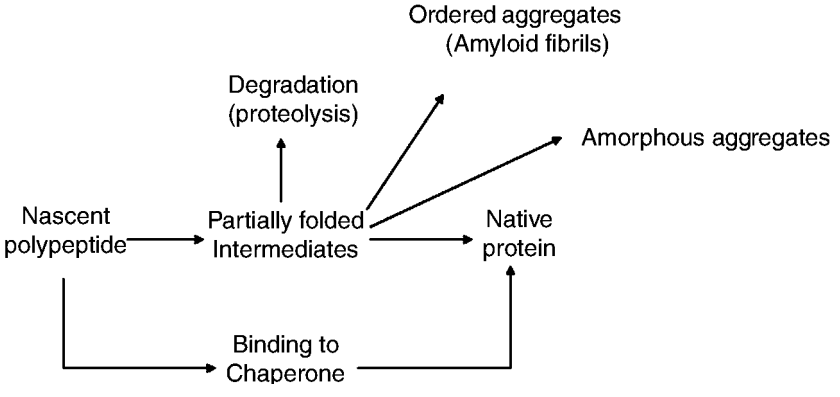


Fig. 7 Unified mechanism of protein folding. (From Ref.[15].)

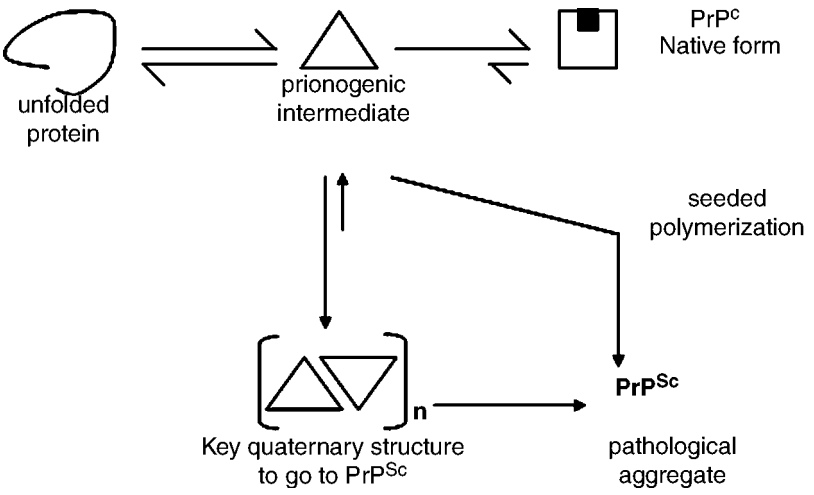


Fig. 8 Ligand binding to prion protein–a possible therapeutic strategy to prevent scrapie formation. (From Ref.[9].)

pathological deposits. Osmolytes and other stabilizing ligands (kosmotropes) may decrease the amount of aggregation. This presumably reflects differential stabilization of the native state. Osmolytes have been shown to inhibit prion formation in the scrapie system. This is consistent with the formation of prions.[15]

The stability of the native protein PrP^C is increased by ligand binding to the specific site on the protein molecule. This reduces the rate of conversion (backward reaction) of native protein to prionogenic intermediate that aggregates to form pathological PrP^{Sc}. The rate of amyloid formation is usually greatly enhanced by seeding. Hence, prevention of amyloid seeds will delay the onset of amyloid formation.

Therapeutic methods have been developed that are designed either to inhibit amyloid formation by blocking fibril growth. Diclofenac analogues that belong to the class of nonsteroidal anti-inflammatory drugs are transthyretin amyloid formation inhibitors[17] (refer to Fig. 8). Apart from these therapeutic strategies, surgical gene therapy and organ transplantation are the other methods of curing such diseases.[15]

THERMODYNAMIC PERSPECTIVE OF PROTEIN FOLDING

Kinetics and thermodynamics are always complementary for process study. While kinetics describes the rate of the process, thermodynamics analyzes its feasibility. As kinetic mechanisms describe the sequential search with the end state being a single protein having a particular function, a thermodynamic study involves

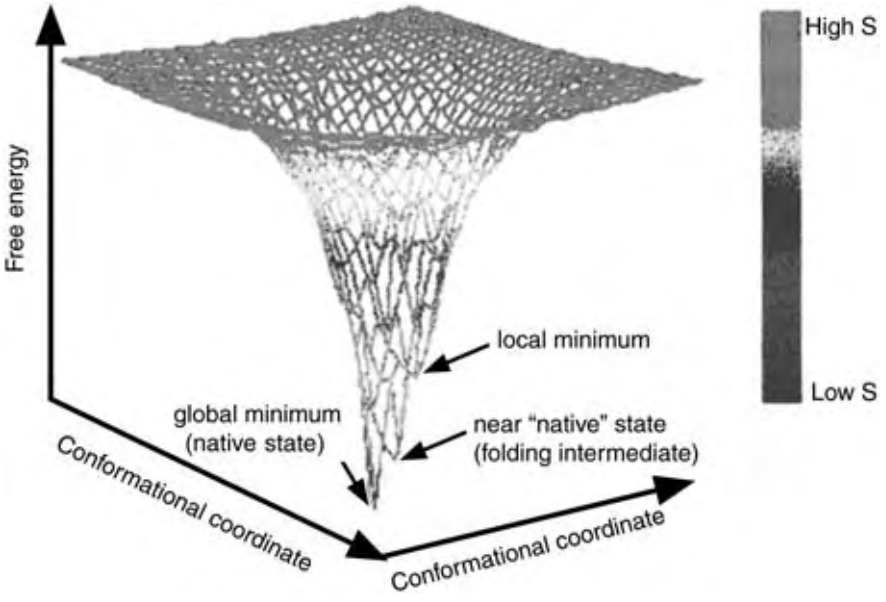


Fig. 9 Energy landscape diagram. (From Ref.[18].) *(View this art in color at www.dekker.com.)*

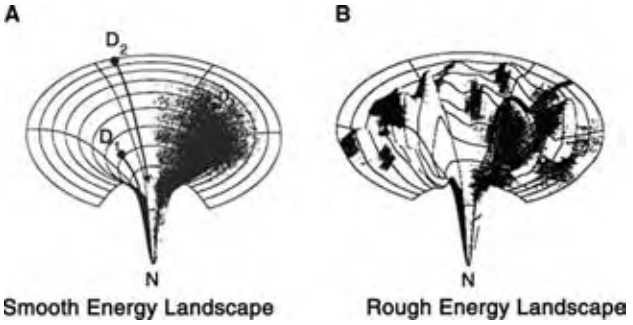


Fig. 10 (**A**,**B**) Comparison of smooth and rough energy landscape. (From Ref.[18].)

a conformational search with the end state being the single native structure of a given protein.

The thermodynamics of protein folding can be viewed as a ball rolling down a cone trying to reach a stable stationary position, which is the lowest point in the cone. In this case, the ball is the protein and the path traced by it, to reach the final position or the folded state, is the folding sequence. The cone represents the energy coordinates. Thus, protein folds to attain the least free-energy state. The plot is represented three dimensionally with the free-energy (z-axis) and conformational coordinates (x- and y-axis). Such plots or cones are referred to as energy landscapes.[18]

As we move down an energy landscape, many conformations have high energy and a few have a low energy. Conformations having high energy have high conformational entropy (high S), and states having low energy (native state and other deep minima) have low conformational entropy (low S) (refer to Fig. 9). Also, smoother funnel topography (surface) has implications about dynamics and time dependence, because the energy barriers are so small that the process happens quickly. For the smooth funnel, folding, kinetics should be fast and two-state, and for the rugged landscape folding, kinetics will be slower and more complex.

A protein can fold to only one native functional form. In the landscape, there should exist a point of least energy that corresponds to the native state. Such a point of least energy is called the global minima. However, there exist several local minima on the conformational coordinates on the energy landscape. Such points of local minima are ''kinetic traps,'' where folding might sometimes stop after attaining that minima. ''Kinetic traps'' refer to ''off-pathway'' proteins that may aggregate to form pathological aggregates. All such points of local minima can be joined together to form a parabolic curve called convex global underestimator (CGU).[18] Several CGU's can be constructed for a single landscape. The apex of the CGU corresponds to the lowest energy, and gives the global minimum. This is the ultimate native and functional state of the protein.

Figs. 10A and 10B show a section of a smooth energy landscape with three different proteins trying to attain a native state via different conformational pathways, and a section of another energy landscape with a rough topology, energy barriers, and local minima.

The transition of short protein from the unfolded to the folded state is accompanied by a drastic reduction of entropy. In this scenario, there is competition between the two minima of the free energy: the folded state with low energy and entropy, and the unfolded state corresponding to high energy and entropy. The folding mechanism can be viewed as multiple conformations racing down the funnel slopes in multiple paths, with some paths more heavily traveled than others.[19]

Fig. 11 represents the two-dimensional section of the energy landscape with the free energy plotted against the conformational coordinate. The CGU can

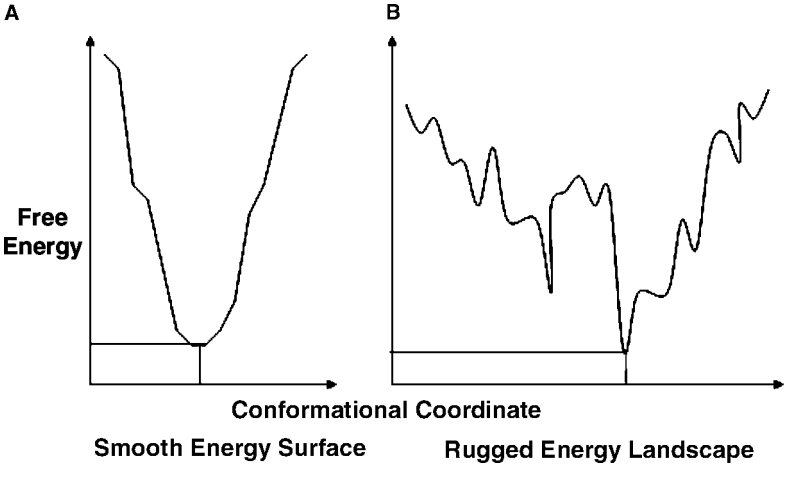


Fig. 11 Smooth and rough energy landscape fluctuations on a two dimensional coordinate. (From Ref.[18].)

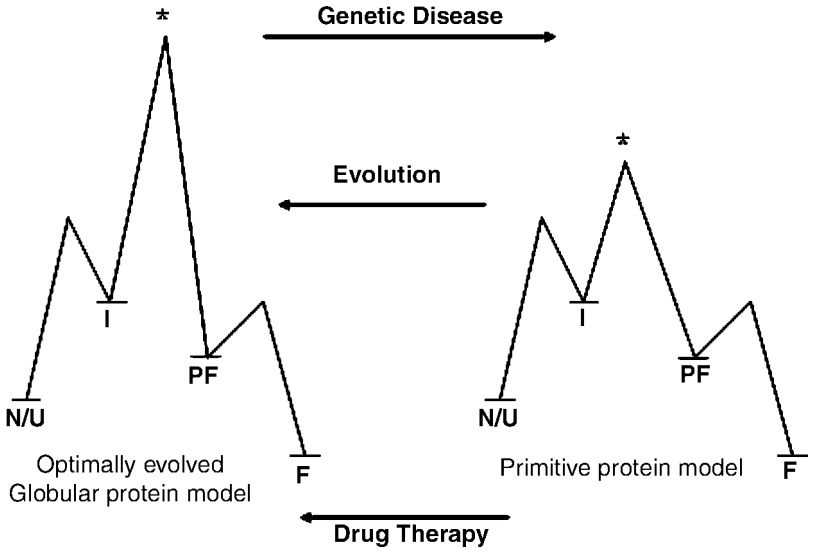


Fig. 12 Primitive vs. optimally evolved protein folding model. (From Ref.[7].)

be constructed for the rugged surface by joining all the points of local minima.[18] This would serve to find the global minima in the case of complex landscape and to convert the rugged surface to a smoother surface. Thus, the CGU removes all the local minima and the energy barrier. This also provides an easier and a correct route to reach the native state with the least free energy and entropy. This is a possible solution to complex landscapes.

KINETICS VS. THERMODYNAMICS

Pathological aggregates may be formed when a protein encounters local minima and fails to recover from it and folds further to a native state. A protein that folds through a complex mechanism must overcome all the energy barriers (up hills) and local minima (kinetic traps) to attain global minima (native state). The surface of the energy landscape would have more corrugations for a complex mechanism than either a series or a series-parallel mechanism. A series mechanism has a smoother surface where the local minima and the energy barriers are the minimum. Kinetically, the off-pathway steps have to be minimized to prevent aggregate formation. This implies a reduction in the number of local minima and the energy barrier towards the formation of the native state. In some cases, the native protein is unstable, and reverses to form aggregates via the intermediates. This could probably be because of the presence of a small energy barrier that the native protein can overcome to deposit into a ''deep'' local minima some where near the global minima. The role of chaperones would be similar to catalysts, where the activation energy to overcome the energy barrier and to rise out of the local minima is provided by the chaperone.

THERMODYNAMIC INERPRETATION OF THERAPEUTIC STRATEGY

Evolution has designed proteins to fold in a pathway that leads to their native form. However, proteins that tend to escape such pathways result in the deposition of aggregates. The free-energy diagram or thermodynamic diagrams for an optimally evolved protein-folding scheme vs. a primitive protein-folding scheme, which tends to form aggregates is shown in Fig. 12.[7]

Ribosomal production of the unfolded state leads to amyloid fibril formation in the primitive situation. In the optimally evolved globular protein, the model native state can be formed and fibril formation can be overcome by one of the three cases.[7]

Energy barrier to ordered aggregation that results from stabilization of the folded state.

Destabilization of the intermediate to prevent its build up.

An increase in the activation energy (*) for ordered oligomerization (I to PF), or the conversion of the intermediate to Protofilament (precursor to fibrils).

All therapeutic strategies are designed based on one of the three cases outlined above.

CONCLUSIONS

Protein-folding mechanisms are reviewed with the purpose of providing physical insights into the formation of fibrils and aggregates that deposit in different parts of the body and gradually lead to intractable diseases. The folding mechanisms were classified into three types. Examples in each category were given, and a unified mechanism for folding was proposed. The mechanism discussed in each category is related to a specific disease. Models and mechanisms of

aggregation and the kinetics of aggregate formation were presented. The focus was also on different therapeutic strategies, which prevent the β-peptide from precipitating and becoming toxic to nerve cells. The unified mechanism of protein folding was represented thermodynamically using energy landscapes. The regions relating to the formation of the native state and misfolded proteins were represented on the energy landscape. The entry describes the relation between the different kinetic mechanisms and its relation to the thermodynamic landscapes. The basic principle in the treatment of such intractable diseases caused due to misfolding lies in blocking or inhibiting the fibril formation. Thermodynamically, it refers to smoothening the surface of the energy landscape by reducing the kinetic traps or local minimas that lead to misfolding. Therapeutic strategies to cure diseases caused due to misfolded protein deposition can be designed and developed based on fundamental kinetic and thermodynamic principles. This entry is an effort to present mechanisms and methods that can be designed to block or prevent pathological aggregates deposition. This could be used to develop effective therapeutic strategies to cure diseases caused due to misfolding.

REFERENCES

1. Sadana, A. Applications and economics of bioseparation. In *Bioseparation of Proteins*: *Unfolding/Folding and Validations*; Satinder, A. Ed.; Separation Science and Technology series; Academic Press: San Diego, 1998; Vol. 1, 259–285.
2. Smoot, A.L.; Panda, M.; Brazil, B.T.; Buckle, A.M.; Fersht, A.R.; Horowitz, P.M. The binding of *bis*-ANS to the isolated GroEL apical domain fragment induces the formation of a folding intermediate with increased hydrophobic surface not observed in tetradecameric GroEL. Biochemistry **2001**, *40*, 4484–4492.
3. Chatellier, J.; Buckle, A.M.; Fresht, A.R. GroEL recognizes sequential and non-sequential linear structural motifs compatible with extended beta-strands and alpha-helices. J. Mol. Biol. **1999**, *292* (1), 163–172.
4. Krantz, B.A.; Sosnick, T.R. Distinguishing between two-state and three-state models for ubiquitin folding. Biochemistry **2000**, *39*, 11696–11701.
5. Laurents, D.V.; Corrales, S.; Elfas-Arnanz, M.; Sevilla, P.; Rico, M.; Padmanabhan, S. Folding kinetics of phage 434 Cro protein. Biochemistry **2000**, *39*, 13963–13973.
6. Khurana, R.; Gillespie, J.R.; Talapatia, A.; Minert, L.J.; Ionescuzanetti, C.; Millet, I.; Fink, A.L. Partially folded intermediates as critical precursors of light chain amyloid fibrils and amorphous aggregates. Biochemistry **2001**, *40*, 3525–3535.
7. Lansbury, P.T., Jr. Following nature's anti-amyloid strategy. Nat. Biotechnol. **2001**, *19*, 112–113.
8. Minton, A.P. Protein folding: thickening the broth. Curr. Biol. **2000**, *10*, R97–R99.
9. Caughey, B. Transmissible spongiform encephalopathies, amyloidoses and yeast prions: common threads? Nat. Med. **2000**, *6*, 751–754.
10. Kelly, J.W. The environmental dependency of protein folding best explains prion and amyloid diseases. Proc. Natl. Acad. Sci. USA **1998**, *95*, 930–932.
11. Liautard, J.P. Analytical background and discussion of the chaperone model of prion diseases. Biotheoretica **1999**, *47*, 219–238.
12. Wickner, S.; Maurizi, M.R.; Gottesman, S. Posttranslational quality control: folding, refolding, and degrading proteins. Science **1999**, *286*, 1888–1893.
13. Konno, T. Amyloid-induced aggregation and precipitation of soluble proteins: an electrostatic contribution of the Alzheimer's β(25–35) amyloid fibril. Biochemistry **2001**, *40*, 2148–2154.
14. Sidekari, V.; Gilbert, H.F. Mechanism of the antichaperone activity of protein disulphide isomerase: facilitated assembly of large, insoluble aggregates of denatured lysozyme and PDI. Biochemistry **2000**, *39*, 1180–1188.
15. Fink, A.L. Protein aggregation: folding aggregates, inclusion bodies and amyloid. Fold. Des. **1998**, *3*, R9–R23.
16. Massi, F.; Straub, J.E. Energy landscape theory for Alzheimer's amyloid β-peptide fibril elongation. Proteins: Struct. Funct. Genet. **2001**, *42*, 217–229.
17. Oza, V.B.; Smith, C.; Raman, P.; Koepf, E.K.; Lashuel, H.A.; Petrassi, H.M.; Chiang, K.P.; Powers, E.T.; Sachettinni, J.; Kelly, J.W. Synthesis, structure and activity of diclofenac analogues as transthyretin amyloid fibril inhibitors. J. Med. Chem. **2002**, *45*, 321–332.
18. Dill, K.A. Polymer principles and protein folding. Protein Sci. **1999**, *8*, 1166–1180.
19. Weissman, J.S.; Rye, H.S.; Fenton, W.A.; Beechem, J.M.; Horowich, A.L. Characterization of the active intermediate of a GroEL-GroES-mediated protein folding reaction. Cell **1996**, *84* (3), 481–490.

Protein Production in Transgenic Plants: Development and Commercialization

Wayne R. Curtis
Department of Chemical Engineering, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.

INTRODUCTION

The potential for producing commercially important proteins in plants has been obvious since the first reports of successful integration of foreign genes in plants over three decades ago. However, to date, only Prodigene has produced a commercial recombinant protein product in transgenic plants.[1] The motivations include reduced production cost, scalability, avoidance of human pathogens, and eukaryotic post-translational modification; however, numerous problems have slowed down the progress. The range of application of this technology is quite large, from industrial enzymes to injectable or diagnostic monoclonal antibodies. Although in this entry the emphasis is on therapeutic proteins, the basic concepts and associated citations are equally applicable to any protein of commercial interest.

The use of genetic engineering to improve food production is ongoing throughout the world. While many of these companies are predominantly in the field of agriculture biotechnology (Monsanto, Syngenta), virtually every, other chemical company has significant agricultural chemicals divisions and plant biotechnology efforts (Dow AgroSciences, Bayer CropScience, BASF Plant Science GmbH, and DuPont Agriculture & Nutrition). The efforts of these companies are (presumed to be) largely focused on issues such as crop improvement (e.g., nutrition, food processing, cultivar development, etc.) and productivity enhancement (e.g., insect resistance and herbicide tolerance). The focus of this entry is specifically on the use of plants for protein production with an emphasis on pharmaceuticals. Because of lack of information, any efforts within—and collaborations between—large pharmaceutical and agrichemical companies are omitted. It must be kept in mind that the efforts of these companies have not only created the enabling technologies associated with transgenic plants, but are pushing the boundaries of genetic modification in plants as well. If ''molecular farming'' becomes profitable, these companies can and will likely become the major players.

TRANSGENIC PLANTS FOR AGRONOMIC PRODUCTION

The introduction of DNA into plant cells to facilitate expression of heterologous proteins such as human therapeutics has become routine. Strategies for DNA delivery are improving and evolving, but the use of *Agrobacterium tumafaciens* remains the dominant technique. Agrobacterium is a natural pathogen of plants that has the ability to transfer DNA from a specialized tumor-inducing plasmid. Genetic engineering involves replacing a portion of the transfer DNA (T-DNA) with the heterologous gene of interest (which is carried out in *Escherichia coli* by standard molecular biology techniques). Exposure of the plant tissue to Agrobacterium containing the modified plasmid results in integration of the T-DNA into the plant chromosome.[2,3] Other techniques including the introduction of modified plant viruses or chloroplast transformation will also be discussed in this section because they similarly represent approaches for genetically engineering plants to produce proteins. These three approaches are shown schematically in Fig. 1.

Much of the work with transgenic plants has focused on species that are ''easy'' to transform such as the Solonaceous family that includes tobacco, tomato, and potato. Commercial targets include more difficult species including corn and soybeans. An important aspect of understanding the technological development of transgenic plants is that ease of transformation does not imply speed. Regeneration of fertile plants from a transformation event that is often a single cell takes months to complete. In addition, the size of transgenic plants and the space needed for subsequent maintenance and genetic crosses invariably result in the academic screening of relatively small populations of plants (10–20 plants). The variation among these transformants is often quite large (order of magnitude) and it is not unusual to have plants that do not express the transgene at all. Not surprisingly, more intensive efforts that screen larger populations of plants are able to identify higher levels of expression; nonetheless, the levels of heterologous protein

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030629

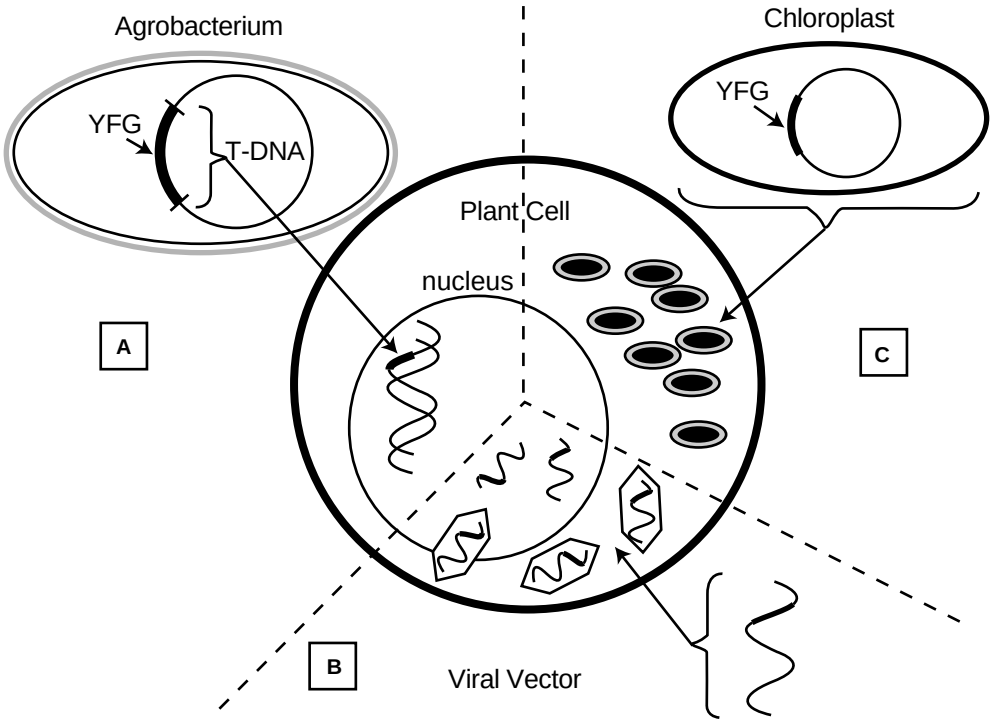


Fig. 1 Schematic of gene delivery strategies for producing protein in plants. (A) Agrobacterium where ''your favorite gene'' (YFG) is inserted into the bacterial T-DNA and infection results in transfer of the T-DNA to the plant chromosome; (B) DNA is inserted into the viral genome and the protein is produced as the virus amplifies inside the plant cell; (C) DNA is inserted into the chloroplast DNA resulting in protein expression inside the chloroplast

expression in plants are limited to relatively low levels (a few percent of soluble protein). One of the most important mechanisms limiting the expression is gene silencing. Technology is being developed using virus proteins, which inhibit this phenomenon as a strategy to overcome current limitations in expression levels.[4,5]

Stable Chromosomal Transformation

In the plant chromosome, each gene is driven by its own promoter. The most straightforward approach to transformation is the use of a constitutive promoter that expresses the heterologous gene ''all the time'' and in ''all'' tissues. The promoter of the cauliflower mosaic virus CaMV 35s protein is the most commonly used constitutive promoter.[6] Other promoters are gaining popularity for expression.[7] particularly for monocots (grasses). Use of specialized promoters opens up opportunities for controlling expression to specific organs or at specific times.[8] These different approaches to control gene expression are reflected in Table 2 among the technologies that have spawned companies. Wound-induced promoters were the strategy used by CropTech, Inc., where the protein was expressed only after the plant (tobacco) tissue was harvested and macerated. This provided a means of obtaining ''containment'' as the protein would be minimally expressed in the field. This strategy also exemplifies the advantage of an inducible promoter where toxic effects associated with expression can be minimized. Expression of proteins from genes integrated into the chromosome generates a protein extract that usually contains 99% plant proteins (Table 1).

The strategy of SemBioSys, Inc. is based on expression associated with oil bodies in seeds.[30] The premise of this approach is that the protein can be recovered from the majority of the contaminating proteins by partitioning with the oil fraction of the seed.[31] Seed specific expression is very attractive for commercial production because it is amenable to agronomic practices for harvest. Expression in seeds has also been shown to provide for long-term stable storage (years without refrigeration), and seeds are typically characterized by only a few ''contaminating'' seed storage proteins rather than the complex mixture found in other plant parts. Prodigene, Monsanto, and invariably many Agriscience groups have developed efficient strategies for corn and soy seed expression. The acquisition/affiliation of major seed companies by agrichemical companies (DeKalb and Asgrow by Monsanto, Northrup King and Ciba Seed by Novartis, and

Table 1 Constitutive expression of proteins in transgenic plant tissues (emphasis on therapeutics)

Protein	Expression level	Plant/tissue	Ref.	Comment
Transgenic plants				
Rabies antibody (LC/HC), monoclonal, human	3 μg/g FW(0.07% TSP)	Tobacco leaves	[9]	
Cytomegalovirus glycoprotein, 883aa	11.2 mg/g seed TSP, 1.07% seed protein	Tobacco seeds	[10]	Human gene, plants grown in mine shaft
Anti-Rhesus D IgGI antibody, 150 kDa	0.6% TSP	*Arabidopsis*	[11]	Glycosylated
IgG (MAK33)	0.75%/0.5% TSP	Potato leaves/tubers	[12]	Co-transformed roots
Virus structural S-protein, swine gastroenteritis	40 μg/g DW	Corn seed	[13]	Prodigene
Human IgG	1% TSP	Alfalfa *(Medicago saliva)*	[14]	
LT-B, *E. coli* enterotoxin	4–10 μg/g FW	Tuber tissue	[15]	
LT-B, *E. coli* enterotoxin	2.2% TSP (94.6 μg/g FW)[A]	Leaves, *N. tabacum*	[16]	Plant codon optimized
β-Glucuronidase	0.7% TSP	Corn seed	[17]	First commercial product
Avidin, 66 kDa	5.7%			Egg white protein
α-Interferon	600 mg/L	Lemna (duckweed)	BioLex web page	Basis of volume not clearly specified
Chloroplast expression				
Human serum albumin, 66.5 kDa, 585aa	11% TSP	Tobacco leaves	[18]	
B.t. insecticidal protein	46% TSP	Tobacco leaves	[19]	Highest protein in transgenic plant to date
Human growth hormone	7% TSP	Tobacco leaves	[20]	
Fluorescent protein fusion (FLARE-S)	18% TSP	Tobacco leaves	[21]	
Plant tissue culture				
Antibody, guys' 13, monoclonal	1.6 mg/g DW (6.5% TSP)	Tobacco	[22]	Root culture
Antibody, anti-cancer, (L6 sFv)	200 mg/L	Tobacco	[23]	Monsanto patent
Antibody, 1K2 receptor (sFv)	l00 mg/L			Secreted
TMV antibody, full-size murine, IgG,	45 μg/g FW(0.9% TSP),	Tobacco–Petite Havana	[24]	In cell walls, 80% recovered by digestion
Human αl-antitrypsin	100 mg/L	Rice suspension culture	[25]	
Human IL-4	0.28 mg/L	Tobacco	[26]	Secreted

(From Refs.[27–29].)

Table 2 Protein expression achieved using viral vectors

Protein/peptide	Size	Titer	Virus: host	Ref.	Comment
Rabies peptide-CP fusion, 29.3 kD	29.3 kDa	400 μg/g FW	AIMV tobacco leaves	[32]	ssRNA virus w/suppl. CP
Rabies peptide-CP fusion, 19.3 kD	19.3 kDa	50 μg/g FW	TMV: *N. benthamiana*		CP fusion
		60 μg/g FW	TMV: lettuce		CP fusion
Neutralizing epitope HIV-1	6aa	150 μg/10 g FW	Potato virus X	[33]	
Birch pollen allergen	17.5 kDa	2.53% TSP	TMV	[34]	
HIV-l p24			Tomato bush stunt virus	[35]	CP(−)
α-Trichosanthin		2% TSP	TMV	[36]	Anti-viral protein
Angiotensin-I converting enzyme	12aa	100 μg CP-fusion/g FW	TMV	[37]	CP/CP-fusion ''read-through''

Mycogen with Dow and Pioneer by Dupont) makes the perspective of plants as a chemical production platform a logical extension of these agrichemical businesses. However, seeds are also the dominant source of human nutrition.

Concerns for gene containment and public perception of genetically modified foods (and potential cross contamination) have placed a major roadblock in the path of these commercialization efforts. Examples of alternatives that achieve better-contained expression (e.g., potato tubers[12]) are given in Table 1. Transformed plant tissues in culture are also a contained system, where expression levels are comparable to those of other transgenic systems. As shown in Table 1, Monsanto developed and patented secreted antibody levels of 25–200 mg/L,[23] and reports for antibodies in tobacco cell suspension achieved 6.5% TSP.[22] More recent specialized production systems that can achieve containment include secretion from roots in hydroponics (Phytomedics, Inc.) and secretion from duckweed (BioLex). Phytomedics is based on a specialized promoter and secretion sequence that is induced by phosphate deprivation.[38] BioLex has reported prepurified protein titers of over 600 mg/L for α-interferon (www.biolex.com). According to news releases, these companies have each secured multimillion dollars of venture and government funding in 2003–2004, reflecting both progress and optimism for use of plants for protein production. It is improbable that any such containable plant production could compete with field-based production. Therefore, the continued development of these processes also reflects concern for the future of production of human therapeutics in the field. Growth of large quantities of transgenic plants in containment greenhouses and the use of underground mines for contained growth of plants expressing human cytomegalovirus glycoprotein[10] illustrate how containment concerns are impacting the development of this technology.

Viral Vectors

Genetic elements of viruses are ubiquitously used in genetic engineering to achieve supranatural expression in organisms. For viral vectors, the production system involves the use of a ''functional'' virus. The expression levels achieved by these systems are typically several percent of soluble protein—consistently higher than achieved by chromosomal integration (Table 2).

Viral vector technology launched a company called ''Large Scale Biology Corporation'' (LSBC^{TM}). Part of the commercialization plan is to grow tobacco plants and introduce the genetically engineered tobacco mosaic virus (TMV) into plants by high-pressure spraying. For field application, this system would require approval for ''release'' of a genetically modified organism (GMO) virus. There have also been concerns on the size of the protein that can be expressed efficiently without disrupting the virus as well as on the genetic integrity for an RNA virus. The extent to which these issues have been resolved is not clear; the role of virus-based expression of proteins has been significantly downplayed in current company literature.

The use of viral vectors has been impressively successful in expressing small proteins and epitopes as fusions with the viral coat protein. This approach also displays the protein on the surface of the viral capsid which has been utilized very effectively for achieving an immunization response in animal studies.[33,39] The initial use of this system appears to be as a development platform for protein production—much like baculavirus for mammalian protein production. The future of viral vectors beyond this stage depends upon issues of containment similar to field-grown transgenic plants. The cost-advantage of plant-based expression is severely reduced if plants must be produced in containment greenhouse conditions. Although studies have shown that genetically modified viruses have limited survival relative to their wild-type counterpart,

rational arguments of virus containment are subject to the same—sometimes nonscientific—scrutiny of the public (in this case it would include tobacco farmers whose crops could be a host). Viral vectors are discussed further under transient gene expression.

Chloroplast Transformation

Although chloroplast transformation is relatively new, it represents a significant breakthrough in plant transformation that has achieved expression levels that is several orders of magnitude higher than chromosomal integration.[40] In Table 1, reported expression levels of 7–46% of TSP are shown. Functional human growth hormone has been produced.[20] The highest expression level of 46% of leaf protein for the 65 kDa insecticidal ''Bt protein'' was associated with the precipitation of protein into ''inclusion bodies'' within the chloroplast. High levels of expression result from both the presence of many gene copies (200–10,000 plastids/cell) and the lack of gene regulation characteristic of the prokaryotic-like gene expression of the chloroplast. The prokaryotic nature of the chloroplast also confers limitations in desirable post-translational modification. Nonetheless, formation of disulfides and protein folding has been accomplished, and, because of its polycistronic nature, the chloroplast gene expression lends itself to introduction of multiple genes to facilitate modifications of gene expression that might be needed. At this time, technical information on gene expression from chloroplast transformation is very limited, as there is intense activity in development and patenting.

Patents and Politics/Start-Ups and Shut Downs

Successful genetic transformation in plants started with a flourish of patent applications in 1983 from Washington University, Max Planck Institute, and Monsanto. The application of Washington University took 17 years to be issued (U.S. 6,051,757, 4/18/2000). Max Planck Institute received its European patent September 5,1996 (EP 116,718). Ironically, Monsanto did not receive a U.S. patent based on these initial filings (though technically their initial approach to transformation could be considered superior). Attempting to understand the significance of these patents, or the dozens of related patents granted for plant transformation is outside the scope of this entry. The intellectual property landscape has become increasingly complex and includes a combination of techniques, genes, and genetic elements (promoters, terminators) so that understanding legal issues in genetic engineering of plants has become more difficult than understanding the technical aspects of generating transgenic plants.

In contrast to the general area of plant transformation, Monsanto has clearly had the dominant patent position in the niche area of pharmaceutical production in plants. In its acquisition of Calgene, Monsanto acquired the dominant patent in this field of human therapeutic protein expression in plants. Although the patent of Goodman et al. (U.S. 4,956,282, 9/11/1990) demonstrated and claimed expression of interferons in dicot plants, it became the pivotal enabling technology for mammalian gene expression in plants. Monsanto chose to exercise this patent (in conjunction with acquired Agracetus expertise) to pursue ''third party'' development of therapeutic molecules. However, on October 15, 2003, Monsanto announced dissolution of its Protein Technology business unit. In exiting this business, it stated that the decision was ''not a reflection on the performance of the business or the viability of the technology.'' It would appear that the obstacles that have slowed down the commercialization of this technology have diminished the useful patent life of their dominant patent (impending expiration 9/11/2007). Nonetheless, for the next few years, any company that intends to commercialize mammalian gene in transgenic plants will have to consider this patent as part of the ''freedom to operate'' (FTO). It is not clear how Monsanto plans to manage this lingering IP. It could either try to recover some value by licensing to a company that might want to get a ''headstart'' in this field or leverage this in public relations associated with GMO by sharing a more favorable ''limelight'' associated with the use of genetically engineered plants to improve human health. Either way, public opinion and liability will invariably play a large part in what happens in the next few years. The interesting consequence of an early dominant patent position for the area of mammalian genes in plants is that, postexpiration, there is equally broad FTO for commercialization. If issues associated with containment and public perception can be resolved, there are huge business opportunities. Commercialization of these technologies in non-U.S./EU countries such as India seems inevitable.

The fate of ''molecular farming'' is intrinsically linked to public perception associated with GMO and more specifically genetically modified food crops. ''Incidents'' associated with GM-plants have become familiar to the public: the contamination of tacos with Avantis' Starlink corn,[41] public outcry over the ''terminator gene'' to prevent seed replanting is often associated with Monsanto, who did not develop or ever own the technology,[42] and finally, the killing of monarch butterflies with B.t. *(Bacillus thuringiensis)* corn pollen in laboratories at Cornell,[43] all fuel concerns of the public and politicians. Reviews specifically discussing the details of safety are a better place to become acquainted with the technical details behind

these often misunderstood news stories.[44] These concerns became more tangible for Prodigene Inc., arguably the leader in technology to produce proteins in plants, because of their commercialization of several (non-mammalian) proteins from transgenic plants. While neither β-glucuronidase nor avidin appears to be current commercial products, these products provided much needed case studies on the processes associated with expression and purification of protein from transgenic corn.[17,45] These early successes were offset by a nearly disastrous USDA/FDA December 6, 2002, sanction for reported violation of the Plant Protection Act, where transgenic corn containing a protein being studied under an FDA-IND permit became mixed into a soybean crop as a result of ''volunteer'' growth the next year (FDA, release T02-46, 11/19/2002). In addition to a $250,000 fine, the company agreed to purchase half a million bushels of soybeans and pay for disposal as well as cleaning of the contaminated storage facility in Nebraska.[46] Besides the $3 MM price tag, this incident demonstrated the financial risk associated with ''molecular farming'' and validated concerns for containment of therapeutic molecules in food-crop species. Prodigene has survived this incident. In April 2004, Prodigene announced another commercial product (TrypZean™) a plant-produced recombinant bovine serine protease trypsin, and others are in development.[47] A final twist of irony is that Prodigene is one fourth owned by a publicly financed Iowa investment fund; therefore, Iowa taxpayers are indirectly gambling on the success of molecular farming. For field-based production (the basis of favorable economics), the strategies for gene containment must be incorporated in the initial stages of genetic engineering. There are many transgenic plant materials containing mammalian and human genes developed for commercialization that may remain indefinitely in laboratory seed collections. If issues of containment are dealt with to permit field-based production of therapeutic proteins in plants, Prodigene may benefit because its near demise has clearly refocused other efforts on more contained (and more expensive) production systems. Discoveries and associated patent rights have been a very important driving force for continued efforts in commercialization of pharmaceuticals in plants for ''startup'' ventures. A summary of some key patents and associated companies is provided in Table 3.

Many of the new companies have leveraged novelty of an expression platform (e.g., containment) as the competitiveness of the protein production platform. Epicyte, Inc. appears to have been an exception that developed proprietary expertise in antibody expression in plants without a discernable novel patent position (whether Epicyte obtained a license to operate from Monsanto is not clear; Epicyte was acquired by another ''startup,'' BioLex, Inc. in 2004). Because expression of mammalian genes in plants has been covered by the Calgene-Monsanto patent since 1990, much of the academic research in this area is not protected by intellectual property. This includes a wide variety of therapeutic proteins that have been expressed in plants. Several organizations at the academic–commercial boundary have made considerable advances toward commercialization: Fraunhofer IME, Aachen, or Schmallenberg Branches, Germany (and associated Research Center for Molecular Biotechnology in Newark, Delaware, U.S.A.) falls on the commercial boundary, while the Biotechnology Foundation Laboratories at Thomas Jefferson University, Philadelphia, Pennsylvania, U.S.A. and Boyce Thompson Institute for Plant Research (BTI), Cornell University Campus, Ithaca, New York, U.S.A. have more academic orientations. These groups have made tremendous public domain contributions to issues of glycosylation and pharmacokinetics of plant-expressed proteins (antibodies in particular) as reflected in the various tables and references of this entry.

There are numerous additional companies which have either faltered or altered their mission as the field of ''molecular farming'' continues to germinate. CropTech filed for bankruptcy in April 2003, while in the middle of a planned move from Blackburg Virginia to Charleston, South Carolina. Large Scale Biology Corporation has extensively diversified into providing a wide range of diagnostic and proteomic services to the applied plant biotechnology community. Both of these companies sought to utilize tobacco as the production platform; this alternative use of tobacco has had potential connections to Tobacco Settlement Funds to help spur their development. These funds can have continued impact on commercialization in the U.S.A. if they are managed as development funds by states looking to attract biotechnology industry.

The preceding brief summary of the current status of commercialization of protein production in plants is heavily focused on U.S. industry—owing to the knowledge of the author. It emphasizes the role of safety and public perception in addition to technical and economic feasibility as the factors dictating commercialization. Another very important influence on commercialization is the slow development time-scales for transgenic plants which is a motivation behind rapid expression systems as described in the next section.

RAPID PROTEIN EXPRESSION SYSTEMS FOR DEVELOPMENT

Developing a transgenic plant takes several months as a minimum—much longer for transformation that involves plant regeneration from embryogenic tissue.

Table 3 Key patents for protein expression in transgenic plant production systems (Note: The associated company relates to the company associated with commercialization of the patent)

Inventor(s)	Patent #	Filed: issued	Assignee	Company	Area
Goodman et al.	4,956,282	7/29/85 9/11/90	Calgene, Inc.	Monsanto	Mammalian proteins in plant cells
Moloney	5,650,554	12/30/94 7/22/97	SemBioSys, Inc.	SemBioSys	Protein on seed oil bodies to facilitate separation.
Cramer and Weissenborn	5,670,349	7/29/94 9/23/97	Virgina Tech Intellectual Properties, Inc.	CropTech	Post-harvest promoter (achieve containment)
Wesley et al.	5,804,694	11/6/95 9/8/98	Prodigene, Inc.	Prodigene	β-Glucuronidase production; first commercial product
Artnzen and Lam	5,914,123	6/7/95 6/22/99	Prodigene, Inc.	Prodigene	Edible vaccines in plants
Fitchen and Beachy	5,955,647	9/21/99 1/7/03	The Scripps Research Institute	Large Scale Biology Corp.	Proteins expressed in plants from tobacco mosaic virus (TMV)
Stomp	6,040,498	8/11/99 3/21/00	North Carolina State University	BioLex (+Epicyte)	Genetic engineered Lemna (duck weed) for protein production.
Raskin and Haran	6,544,789	3/13/00 4/8/03		Phytomedics, Inc.	Phosphate-starvation-induced promoter/hydroponic secretion.
Daniell and McFadden	6,680,426	9/19/02 1/20/04	Auburn University	—	Gene expression from plant chloroplasts

Techniques for expedited transformation of meristems and embryos risk chimeric tissues, but increase speed of obtaining tissue to examine stability and expression levels of the introduced gene. It can take several years to obtain sufficiently characterized seeds to permit production of milligram quantities of product for testing. This extended development time-frame tends to place plant-based production as a second generation production platform for established drugs rather than a primary development platform. This is unfortunate, as plant-based production should expand commercialization of new drugs because of lower production costs. Recently there have been several strategies to rapidly provide proteins expressed in plant tissues (Table 4) to speed up the development time-frame.

Laboratory techniques such as biolistics (gene gun) and DNA delivery to protoplasts are not discussed in this section, as they are inadequate for producing quantitative amounts of purified protein. The speed of the viral vectors does provide material in the time-frame of several weeks; therefore, this production system can provide both rapid and scaleable production. This system is described in the previous sections; it is still not clear how well the viral expression system can tolerate large proteins. Most of the reports concern epitopes or small proteins. In addition, the potential need for fusions with coat protein for efficient expression might also limit this system. The lack of reports to refute these concerns after many years suggests that there are serious limitations to using viral vectors for expressing some proteins.

Agrobacterium can facilitate transient expression in both intact and tissue cultured plants. Leaf infiltration involves introducing the Agrobacterium suspension into the intracellular spaces of the leaf, using pressure or vacuum.[52,53] This approach has been scaled to a level of a kilogram of leaves, sequentially vacuum infiltrated to permit recovery of 20 mg of purified full chimeric antibody used in pregnancy testing after only three days of coculture.[50] This system has also been successfully used to express single-chain antibodies and diabodies in tobacco leaves as well (Table 4). While this technique is scalable, it does require the production of tobacco leaf tissue as the substrate. Thus far, all reports have used either in vitro or greenhouse grown tobacco leaves; it would be interesting to see if field-grown material could also be used, as it would be less expensive.

Suspension cultured cells have also been used for transient expression.[54] This approach has been used to show that by combining two different Agrobacterium strains, the T-DNA from both (Fig. 1) can enter the cell and be expressed.[55] This approach is carried out in static culture where the cells are mixed with Agrobacterium and coculture takes place on an agar-solidified medium. This approach has been ''rediscovered'' recently in a slightly larger format,[45] where Agrobacterium are introduced into a lawn of plant cells created by growmg filtered plant cell suspension on solid media. Although this approach can provide small amounts of protein for functional testing, it is limited in terms of scaleup. This is particularly important because the levels of expression in transient culture are toward the lower end of what is observed in stable transformed lines (Tables 1 and 4). To alleviate this limitation, our laboratory has developed a liquid culture based transient expression system that can be scaled up in bioreactor systems.[56] This was accomplished by developing Agrobacterium auxotrophs[48] that were selected specifically for their inability to grow in liquid culture medium. Using this system, the Agrobacterium growth can be controlled and permit

Table 4 Transient expression of heterologous genes in plant tissues using Agrobacterium-delivered T-DNA (no viral vector)

Protein/peptide	Titer	Host: approach	Ref.
β-Glucuronidase	0.03% TSP	*Nicotiana glutinosa* cell suspensions; liquid co-culture w/Agro	Collens et al.[48]
β-Glucuronidase	Qualitative	Tobacco cell suspension, static co-culture 5 days with Agro; ~150 mg FW tissue/treatment	Fuentes et al.[49]
CryIAbl (insecticidal protein)	0.25 μg/ml extract		
Antibody, hepititus B surfice antigen	0.62 μg/ml extract		
HCG antibody, pregnancy test, scFv, 33 kDa	32 μ/g FW (purified)	Tobacco leaves, vacuum infiltrated Agro; cocultured for 3 days; ~1kg leaf tissue	Kathuria et al.[50]
Diabody, 66 kDa	40 μg/g FW		
full antibody, 170 kDa	20 μg/g FW		
Diabody, carcinoembryogenic antigen	12 μg/g FW	Tobacco leaves, vacuum infiltrated Agro; cocultured 3 days; ~1 kg leaf tissue	Vaquero et al.[51]

expression in liquid grown plant cell suspensions, germinating seeds or other plant tissues in a time-frame of several days. Scalability can be accomplished based on standard bioreactor technology, and biomass production can come from many different formats.

These rapid expression technologies have developed to try to reduce the time-frame for protein expression in plants. Protein is expressed in plant tissues in a small fraction of the time and effort it takes to generate transgenic plants and should prove valuable for rapidly producing quantitative amounts of protein for early stage testing of therapeutic proteins. The recent concerns for containment and potential advantages of non-animal-based production of proteins for safety reasons may also make these approaches a potential alternative to other contained production systems. This is particularly true if the expression levels can be enhanced. The potential is particularly high for quantities of protein that are quite small (e.g., diagnostics) where there is no real advantage to the scale of production that would be represented by field-based production. Because transient systems include both plant and bacterial proteins, they are inherently ''messier'' than transgenic plants. Purification becomes particularly important as described briefly in the next section.

PROTEIN PURIFICATION FROM PLANTS

The predominant focus of this entry is on the production of proteins, and specific details on protein purification can be found in references which deal more specifically with this topic.[57,58] Only general comments will be provided here as the separation strategy is invariably linked to the production system. Specific details of protein purification will be dependent on the molecule of interest and intended purpose.[59,60] In the case of edible vaccines, the hope is to eliminate purification. In this case, the protein being expressed contains the antigenic epitope with the intent of obtaining a mucosal antibody response. As noted above, the use of viral vectors with coat-protein fusions has particular advantages by presenting the antigen on the surface of the virus particle. Virus particles can also be the purified product, as it has been shown that injection of the chimeric plant virus can produce IgG response. Reviews on plant-based vaccines can provide the details of this very broad and complicated area of research. Industrial or commercial enzymes can be used at a minimum level of purification, while injectable therapeutics require absolutely meticulous purification and process validation with the FDA. If agronomic production of proteins becomes a reality, the great reduction in production costs will offset the costs of development and implementation of new process schemes for protein purification from plants. However, if requirements for containment impose cost restrictions on production (transgenic greenhouses, GMP process tanks, etc.), then the ability to tolerate more difficult separations will be reduced. Transgenic expression in leaf extracts provides a ''messy'' starting point for purification. This is particularly true for expression senescing tissue (e.g., CropTech) or plant tissues that have high content of phenolics or other compounds that can react with the proteins. Transgenic production in seeds permits separation from a subset of storage proteins. In corn, seed storage proteins (zeins) should be relatively easy to separate from many therapeutic protein candidates. The strategy of SemBioSys is a novel means of reducing ''hydraulic load'' by partitioning protein with oil bodies; however, this introduces a predominantly hydrophobic phase that is an atypical starting point for downstream purification. While the production of chloroplast ''protein crystals'' may facilitate high levels of production and initial separation, the ability to utilize this will depend on the characteristics of the protein and ability to recover functional protein. The secretion approach of Biomedics and BioLex produces dilute initial concentrations—but with the advantage of avoiding contamination with intracellular proteins.

CONCLUSIONS

Use of plants as a production platform for proteins is rapidly advancing. This is being driven by discoveries in control of gene expression in plants, perceived future needs of the pharmaceutical industry and maturing patent positions. Safety, gene containment and public acceptance are equally important inertial forces shaping the industry. In the next decade, commercialization of nontherapeutic (lower value) proteins in plants is inevitable as expression levels increase and the industry gains experience. Commercialization of biologically active (high value) human therapeutics is more problematic because of issues other than technical and economic feasibility.

REFERENCES

1. Horn, M.E.; Woodard, S.L.; Howard, J.A. Plant molecular farming: systems and products. Plant Cell Rep. **2004**, *22*, 711–720.
2. Bevan, M. Binary *Agrobacterium* vectors for plant transformation. Nucleic Acids Res. **1984**, *12* (22), 8711–8721.
3. Gelvin, S.B. Agrobacterium-mediated plant transformation: the biology behind the ''gene-jockeying'' tool. Microbiol. Mol. Biol. Rev. **2003**, *67* (1), 16–37.

4. Vance, V.; Vaucheret, H. RNA silencing in plants—defense and counterdefense. Science. **2001**, *292* (5525), 2277–2280.
5. Voinnet, O.; Rivas, S.; Mestre, P.; Baulcombe, D. An enhanced transient expression system in plants based on suppression of gene silencing by the pl9 protein of tomato bush stunt virus. Plant J. **2003**, *33*, 949–956.
6. Benfey, P.N.; Chua, N.H. The cauliflower mosiac virus 35S promoter: combinatorial regulation of transcription in plants. Science **1990**, *250*, 959–966.
7. Wang, J.; Oard, J.H. Rice ubiquitin promoters: deletion analysis and potential usefulness in plant transformation systems. Plant Cell Rep. **2003**, *22* (2), 129–134.
8. Potenza, C.; Aleman, L.; Sengupta-Gopalan, C. Targeting transgene expression in research, agricultural, and environmental applications: promoters used in plant transformation. In Vitro Cell Dev. Bio. Plant **2004**, *40*, 1–22.
9. Ko, K.; Tekoah, Y.; Rudd, P.M.; Harvey, D.J.; Dwek, R.A.; Spitsin, S.; Hanlon, C.A.; Rupprecht, C.; Dietzschold, B.; Golovkin, M.; Koprowski, H. Function and glycosylation of plant-derived antiviral monoclonal antibody. PNAS **2003**, *100*, 8013–8018.
10. Tackaberry, E.S.; Prior, P.; Bell, M.; Tocchi, M.; Porter, S.; Mehic, J.; Ganz, P.R.; Sardana, R.; Altosaar, I.; Dudani, A. Increased yield of heterologous viral glycoprotein in the seeds of homozygous transgenic tobacco plants cultivated underground. Genome. **2003**, *46*, 521–526.
11. Bouquin, T.; Thomsen, M.; Nielsen, L.K.; Green, T.H.; Mundy, J.; Dziegiel, M.H. Human anti-Rhesus D IgGI antibody produced in transgenic plants. Transgenic Res. **2002**, *11*, 115–222.
12. DeWilde, C.; Peeters, K.; Jacobs, A.; Peck, I.; Depicker, A. Expression of antibodies and Fab fragments in transgenic potato plants: a case study for bulk production in crop plants. Molec. Breeding **2002**, *9*, 271–282.
13. Streatfield, S.J.; Jilka, J.M.; Hood, E.E.; Turner, D.D.; Bailey, M.R.; Mayor, J.M.; Woodard, S.L.; Beifuss, K.K.; Horn, M.E.; Delaney, D.E.; Tizard, I.R.; Howard, J.A. Plant-based vaccines: unique advantages. Vaccine **2001**, *19*, 2742–2748.
14. Khoudi, H.; Laberge, S.; Ferullo, J.M.; Bazin, R.; Darveau, A.; Castonguay, Y.; Allard, G.; Leimieux, R.; Vezina, L.P. Production of a diagnostic monoclonal antibody in perennial alfalfa plants. Biotechnol. Bioeng. **1999**, *64*, 135–143.
15. Mason, H.S.; Haq, T.A.; Clements, J.D.; Arntzen, C.L. Edible vaccine protects mice against *Escherichia coli* heat-labile enterotoxin (LT): potatoes expressing a synthetic LT-B gene. Vaccine **1998**, *16*, 1336–1343.
16. Kang, T.-J.; Han, S.-C.; Jang, M.-O.; Kang, K.-H.; Jang, Y.-S.; Yang, M.-S. Enhanced expression of the B subunit of *Escherichia coli* heat-labile enterotoxin in tobacco by optimization of the coding. Appl. Biochem. Biotechnol. **2004**, *117* (3), 175–187.
17. Kusnadi, A.R.; Hood, E.E.; Witcher, D.R.; Howard, J.A.; Nikolov, Z.L. Production and purification of two recombinant proteins from transgenic com. Biotechnol. Prog. **1998**, *14*, 149–155.
18. Fernandez-San, M.A.; Mingo-Castel, A.; Miller, M.; Daniell, H.A. A chloroplast transgenic approach to hyper-express and purify human serum albumin, a protein highly susceptible to proteolytic degradation. Plant Biotechol. **2003**, *1*, 77–79.
19. DeCosa, B.; Moar, W.; Lee, S.-B.; Miller, M.; Daniell, H. Overexpression of the Bt cry2Aa2 operon in chloroplasts leads to formation of insecticidal crystals. Nature Biotechnol. **2001**, *19*, 71–74.
20. Staub, J.M.; Garcia, B.; Graves, J.; Hajdukiewicz, P.T.; Hunter, P.; Nehra, N.; Paradkar, V.; Schlittler, M.; Carroll, J.A.; Spatola, L.; Ward, D.; Ye, G.; Russell, D.A. High-yield production of a human therapeutic protein in tobacco chloroplasts. Nat. Biotechnol. **2000**, *18*, 333–338.
21. Khan, S.M.; Maliga, P. Fluorescent antibiotic resistance marker for tracking plastid transformation in higher plants. Nat. Biotechnol. **1999**, *17*, 910–915.
22. Sharp, J.M.; Doran, P.M. Strategies for enhancing monoclonal antibody accumulation in plant cell and organ cultures. Biotechnol. Prog. **2001**, *17*, 979–992.
23. Russell, D.; Fuller, J. Method for producing antibodies in plant cells. U.S. Patent 6,080,560, June 27, 2000.
24. Fischer, R.; Liao, Y.-C.; Drossard, J. Affinity-purification of a TMV-specific recombinant full-size antibody from a transgenic tobacco suspension culture. J. Immunol. Meth. **1999**, *226*, 1–10.
25. Terashima, M.; Murai, Y.; Kawamura, M.; Nakanishi, S.; Stoltz, T.; Chen, L.; Drohan, W.; Rodriguez, R.L.; Katoh, S. Production of functional human α1-antitrypsin by plant cell culture. Appl. Microbiol. Biotechnol. **1999**, *52*, 516–523.
26. Magnuson, N.S.; Linzmaier, P.M.; Reeves, R.; An, G.; HaGlass, K.; Lee, J. Secretion of biologically active human interleukin-2 and interleukin-4 from genetically modified tobacco cells in suspension culture. Protein Express Purif. **1998**, *13*, 45–52.

27. Koprowski, H.; Yusibov, V. The green revolution: plants as heterologous expression vectors. Vaccine. **2001**, *19*, 2735–2741.
28. Ma, J.K.-C.; Drake, P.M.; Christou, P. The production of recombinant pharmaceutical proteins in plants. Nature. **2003**, *4*, 794–805.
29. Fischer, R.; Stoger, E.; Schillberg, S.; Christou, P.; Twyman, R.M. Plant-based production of biopharmaceutical. Curr. Opin. Plant Biol. **2004**, *7*, 152–158.
30. Moloney, M.M.; Habibi, H.R. Expression of Somatotropin in Plant Seeds. U.S. Patent 6,288,304 Bl, September 11, 2001.
31. Parmenter, D.L.; Booth, J.G.; van Rooijen, G.J.; Moloney, M.M. Production of biologically active hiruden in plant seeds using oleosin partitioning. Plant Mol. Biol. **1995**, *29*, 1167–1180.
32. Yusibov, V.; Hooper, D.C.; Spitsin, S.V.; Fleysh, N.; Kean, R.B.; Mikheeva, T.; Deka, D.; Karasev, A.; Cox, S.; Randall, J.; Koprowski, H. Expression in plants and immunogenicity of plant virus-based experimental rabies vaccine. Vaccine. **2002**, *20* (25–26), 3155–3164.
33. Marusic, C.; Rizza, P.; Lattanzi, L.; Mancini, C.; Spada, M.; Belardelli, F.; Benvenuto, E.; Capone, I. Chimeric plant viruss particles as immunogens for inducing murine and human immune responses against human immunodeficiency virus type 1. J. Virol. **2001**, *75*, 8434–8439.
34. Krebitz, M.; Wiedermann, U.; Essl, D.; Steinkellner, H.; Wagner, B.; Turpen, T.H.; Ebner, C.; Scheiner, O.; Breiteneder, H. Rapid production of the major birch pollen allergen Bet v 1 *Nicotiana benthamiana* plants and its immunological *in vitro* and *in vivo* characterization. FASEB J. **2000**, *14*, 1279–1288.
35. Zhang, G.; Leung, C.; Murdin, L.; Rovinski, B.; White, K.A. In planta expression of HIV-1 p24 protein using an RNA plant virus-based expression vector. Molec. Biotechnol. **2000**, *14*, 99–107.
36. Kumagai, M.H.; Turpen, T.H.; Weinzettl, N.; Della-Cioppa, G.; Turpen, A.M.; Donson, J.; Hilf, M.E.; Grantham, G.L.; Dawson, W.O.; Chow, T.P.; Piatak, M.; Grill, L.K. Rapid, high-level expression of biologically active α-trichosanthin in transfected plants by an RNA viral vector. PNAS. **1993**, *90*, 427–430.
37. Hamamoto, H.; Sugiyama, Y.; Nakagawa, N.; Hashida, E.; Matsunaga, Y.; Takemoto, S.; Watanabe, Y.; Okada, Y. A new tobacco mosaic virus vector and its use for the systemic production of angiotensin-I-converting enzyme inhibitor in transgenic tobacco and tomato. Bio/Technology. **1993**, *11*, 930–932.
38. Borisjuk, N.V.; Borisjuk, L.G.; Logendral, S.; Petersen, F.; Gleba, Y.; Raskin, I. Production of recombinant proteins in plant root exudates. Nat. Biotechnol. **1999**, *17*, 466–469.
39. Koo, M.; Bendahmane, M.; Lettieri, G.A.; Paoletti, A.D.; Lane, T.E.; Fitchen, J.H.; Buchmeier, M.J.; Beachy, R.N. Protective immunity against murine hepatitis virus (MHV) induced by intranasal or subcutaneous administration of hybrids of tobacco mosaic virus that carries an MHV epitope. PNAS. **1999**, *96*, 7774–7779.
40. Daniell, H. New tools for chloroplast genetic engineering. Nat. Biotechnol. **1999**, *17*, 855–856.
41. Dorry, E. Taco dispute underscores need for standardized tests. Nat. Biotechnol. **2000**, *18*, 1136–1137.
42. Charles, D. *Lords of the Havest*. Perseus Publishing: Caimbridge, MA, 2001.
43. Gatehouse, A.M.; Ferry, N.; Raemaekers, R.J. The case of the monarch butterfly: a verdict is returned. Trends Genet. **2002**, *18* (5), 249–251.
44. Peterson, R.K.D.; Arntzen, C.J. On risk and plant-based biopharmaceuticals. Trends. Biotechnol. **2004**, *22*, 64–66.
45. Witcher, D.R.; Hood, E.E.; Peterson, D.; Bailey, M.; Bond, D.; Kusnadi, A.; Evangelista, R.; Nikolov, Z.; Wooge, C.; Mehigh, R.; Kappe, W.; Register, J.; Howard, J.A. Commercial production of beta-glucuronidase (GUS): a model system for the production of proteins in plants. Molec. Breeding. **1998**, *4* (4), 301–312.
46. Fox, J.L. Puzzling industry response to ProdiGene fiasco. Nat. Biotechnol. **2003**, *21*, 3–4.
47. Woodard, S.L.; Mayor, J.M.; Bailey, M.R.; Barker, D.K.; Love, R.T.; Lane, J.R.; Delaney, D.E.; McComas-Wagner, J.M.; Mallubhotla, H.D.; Hood, E.E.; Dangott, L.J.; Tichy, S.E.; Howard, J.A. Maize (Zea mays)-derived bovine trypsin: characterization of the first large-scale, commercial protein product from transgenic plants. Biotechnol. Appl. Biochem. **2003**, *38*, 123–130.
48. Collens, J.L.; Lee, D.R.; Seeman, A.M.; Curtis, W.R. Development of auxotrophic *Agrobacterium tumefaciens* for gene transfer in plant tissue culture. Biotechnol. Prog. **2004**, *20*, 890–896.
49. Fuentes, A.; Ramos, P.L.; Ayra, C.; Rodriguez, M.; Ramirez, N.; Pujol, M. Development of a highly efficient system for assessing recombinant gene expression in plant cell suspensions via *Agrobacterium tumefaciens* transformation. Biotechnol. Appl. Biochem. **2004**, *39*, 355–361.
50. Kathuria, S.; Sriraman, R.; Nath, R.; Sack, M.; Pal, R.; Artsaenko, O.; Talwar, G.P.; Fischer, R.; Finnern, R. Efficacy of plant-produced recombinant antibodies against HCG. Hum. Repro. **2002**, *17*, 2054–2061.

51. Vaquero, C.; Sack, M.; Schuster, F.; Firnnern, R.; Drossard, J.; Schumann, D.; Reimann, A.; Fischer, R. A carcinoembryonic antigen-specific diabody produced in tobacco. FASEB J. **2002**, *16* (3), 408–410.
52. Kapila, J.; De Rycke, R.; van Montagu, M.; Angenon, G. Agrobacterium mediated transient gene expression system for intact leaves. Plant Sci. **1996**, *122*, 101–108.
53. Yang, Y.; Li, R.; Qi, M. In vivo analysis of plant promoters and transcription factors by agroinfiltration of tobacco leaves. Plant J. **2000**, *222* (6), 543–551.
54. Narasimhulu, S.B.; Deng, Xb.; Sarria, R.; Gelvin, S.B. Early transcription of agrobacterium T-DNA genes in tobacco and maize. Plant Cell. **1996**, *8*, 873–886.
55. Gomord, V.; Fitchette-Lainé, A.-C.; Denmat, L.-A.; Michaud, D.; Faye, L. Production of Foreign proteins in tobacco cell suspension culture. In *Methods in Biotechnology, Recombinant Proteins from Plants*; Cunningham, C., Porter, A. J. R., Eds.; Humana Press Inc.: Totawa, NJ, 1998; Vol. 3, 155–165.
56. Curtis, W.R. Quantitative Transient Protein Expression in Plant Tissue Culture. U.S. Patent 6,740,526 Bl, May 25, 2004.
57. Desai, U.A.; Gur, G.; Daunert, S.; Babbitt, R.; Li, Q. Expression and affinity purification of recombinant proteins from plants. Protein Expr. Purif. **2002**, *25*, 195–202.
58. Menkhaus, T.J.; Pate, C.; Krech, A.; Glatz, C.E. Recombinant protein purification from pea. Biotechnol. Bioeng. **2004**, *86*, 108–114.
59. Evangelistra, R.L.; Kusnadi, A.R.; Howard, J.A.; Nikolov, Z.L. Process and economic evaluation of the extraction, and purification of recombinant β-Glucuronidase from transgenic com. Biotechnol. Prog. **1998**, *14*, 607–614.
60. Valdés, R.; Gómez, L.; Padilla, S.; Brito, J.; Reyes, B.; Albarez, T.; Mendoza, O.; Herrera, O.; Ferro, W.; Pujol, M.; Leal, V.; Lenares, M.; Hevia, Y.; Garcia, C.; Milá, L.; Garcia, O.; Sánchez, R.; Acosta, A.; Geada, D.; Paez, R.; Vega, J.L.; Borroto, C. Large-scale purification of an antibody directed against hepatitis B surface antigen from transgenic tobacco plants. Biochem. Biophys. Res. Comm. **2003**, *308*, 94–100.

Proton-Exchange Membrane Fuel Cells

P

Pyoungho Choi
Pradeep Haldar
Albany Nanotech, The College of Nanoscale Science and Engineering (CNSE), State University of New York, Albany, New York, U.S.A.

Ravindra Datta
Fuel Cell Center, Worcester Polytechnic Institute, Worcester, Massachusetts, U.S.A.

INTRODUCTION

This article provides a basic explanation of fuel cells, especially the fundamentals of proton-exchange membrane fuel cell (PEMFC). Starting with an introduction to fuel cells, it addresses principles, materials, fuel generation processes, and future technological challenges of PEMFC.

DEFINITION

A fuel cell is defined as an electrochemical device in which the chemical energy of a fuel is converted directly into electrical energy. The fuel is typically a conventional fuel, such as an alcohol or a hydrocarbon, or a substance derivable from it, e.g., hydrogen, which can be supplied continuously. The term directly implies that the device has an anode at which the fuel is electrocatalytically oxidized with the concomitant production of electrons, and a cathode at which a continuously supplied oxidant, typically oxygen, is reduced. The electrodes sandwich an electrolyte layer through which ions produced at the anode (e.g., H^+) or at the cathode (e.g., OH^-, O^{2-}, CO_3^{2-}) diffuse, but not electrons, which proceed through the external circuit where useful electrical work may be extracted. Fuel cells are distinct from batteries, wherein the fuel is typically a metal sealed into the device.

BRIEF HISTORY

The ''fuel cell effect'' was first discovered in 1838 (published in 1989) by C.F. Schoenbein who found the ''inverse electrolysis'' principle in his experiment using platinum electrodes immersed in dilute sulfuric acid solution.[1] However, the invention of the ''fuel cell'' is credited to W.R. Grove who demonstrated Schoenbein's discovery on a practical scale by inventing the ''gas battery'' during 1839–1845.[2–4] Fig. 1 shows an early experiment in which hydrogen and oxygen gases were produced by applied electric current and collected in the small tubes holding the electrode. When the electric current was stopped, a current in the reverse direction was observed, which was generated by the recombination of the gases on the platinum electrodes. The electrochemical reactions took place at gas–liquid–solid interface and the platinum electrodes acted as catalysts, as well as current collectors. W.R. Grove realized the importance of the ''surface of action'' and designed platinum electrodes with deposited platinum particles to increase the surface area of the reaction.[3] In 1889, the term ''fuel cell'' was coined by L. Mond and C. Langer, who utilized a three-dimensional porous electrode structure.[5] In 1894, the utilization of coal as a fuel for fuel cell was designed and coal-based fuel cell employing molten carbonate was developed in early 1900s, and solid oxide fuel cell, molten carbonate fuel cell, alkaline fuel cell were developed later, and phosphoric acid fuel cell was the last fuel cell invented in 1967.[6–8] In 1965, United Technologies Corporation (UTC) produced alkaline fuel cell for the Apollo Lunar Mission. In the late 1950s and early 1960s, General Electric (GE) started to work on fuel cells and developed the first fuel cell, based on a proton-exchange membrane (PEM) as the electrolyte.[9–11] The PEMs used were blends of inert polymer with a highly cross-linked polystyrene-based ionomer, sulfonated phenol-formaldehyde, and heterogeneous sulfonated divinylbenzene cross-linked polystyrene. However, these materials were found to be chemically degraded during the operation of the fuel cells. A solution came in mid-1960s through the collaborative efforts of GE and DuPont, resulting in the development of the Nafion® membrane. In August 1965, the Gemini 5 spacecraft used GE's PEMFC as a source of electrical power. In 1986, Los Alamos National Laboratory (LANL) demonstrated a high surface area electrocatalyst layer that included ionomer gel to increase the three-phase interface and, thus, reduced the amount of catalyst loading significantly without a loss in the cell performance.[12] In 1991, Ballard designed the ''serpentine'' flow field in bipolar plates

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007780

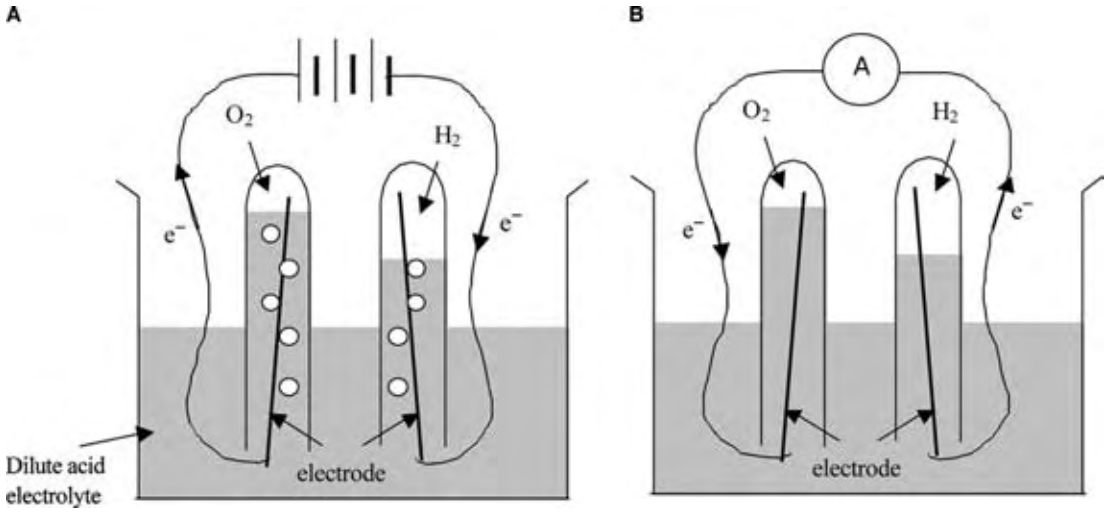


Fig. 1 Comparison of electrolysis (A) and fuel cell (B).

to facilitate water removal from the cathode, thereby improving performance through enhanced oxygen distribution to the cathode.[13] Since the middle of 1990s, a significant number of fuel cell programs have evolved. In recent years, almost all of the major auto-makers, e.g., General Motors (GM), Toyota, Daimler-Benz, Honda, Nissan, and Ford, have made substantial investment into the development of fuel cell technology for use in fuel cell vehicles (FCV), which would provide higher efficiency and require less maintenance than conventional engines. Fuel cells are also of interest for stationary energy generation because they can provide power directly at the site of use and avoid energy distribution losses. Recent progress in reducing the costs and improving the performance and durability of fuel cells promises that, in the near future, this revolutionary technology will be an important power generation system that is reliable, clean, and environment friendly.[14–19]

TYPES

Fuel cells are usually classified by the electrolyte employed in the system, which determines the operating temperature and, hence, the fuel that could be utilized. Table 1 summarizes the characteristics of different types of fuel cells.[20–22]

Alkaline Fuel Cell

Alkaline fuel cell (AFC) was used for Apollo and Space Shuttle program. Alkaline fuel cell employs liquid alkaline, e.g., KOH, as an electrolyte so that fuel, as well as air or oxygen, should be free of CO_2 because the strong alkaline electrolyte reacts with CO_2 to form carbonates, which reduces the ionic conductivity. Electrodes, e.g., Ni, Ag, and metal oxides, are relatively inexpensive compared to that of other fuel cells.

Table 1 Comparison of different types of fuel cells

Type	Fuel	Electrolyte	T (°C)	Charge carrier	Anode reaction	Cathode reaction
AFC	H_2	KOH	65–220	OH^-	$H_2 + 2OH^- \leftrightarrows 2H_2O + 2e^-$	$\frac{1}{2}O_2 + H_2O + 2e^- \leftrightarrows 2OH^-$
PEMFC	H_2	PEM	80	H^+	$H_2 \leftrightarrows 2H^+ + 2e^-$	$\frac{1}{2}O_2 + 2H^+ + 2e^- \leftrightarrows H_2O$
PAFC	H_2	H_3PO_4	205	H^+	$H_2 \leftrightarrows 2H^+ + 2e^-$	$\frac{1}{2}O_2 + 2H^+ + 2e^- \leftrightarrows H_2O$
MCFC	H_2, CO, CH_4	$LiCO_3$–K_2CO_3	650	CO_3^{2-}	$H_2 + CO_3^{2-} \leftrightarrows H_2O + CO_2 + 2e^-$	$\frac{1}{2}O_2 + CO_2 + 2e^- \leftrightarrows CO_3^{2-}$
SOFC	H_2, CO, CH_4	YSZ	600–1000	O^{2-}	$H_2 + O^{2-} \leftrightarrows H_2O + 2e^-$	$\frac{1}{2}O_2 + 2e^- \leftrightarrows O^{2-}$

Phosphoric Acid Fuel Cell

Phosphoric acid fuel cell (PAFC) was the first fuel cell to be commercialized. PAFC uses liquid phosphoric acid as an electrolyte, which is soaked in silicon carbide particle matrix using capillary action. PAFC is tolerant to CO_2 feed stream because the electrolyte is an acid. However, carbon monoxide poisons the Pt electrodes so that its concentration should be below 3–5 volume % in the feed stocks.

Proton-Exchange Membrane Fuel Cell

Proton-exchange membrane fuel cell (PEMFC) is receiving the main attention today. Proton-exchange membrane fuel cell employs a solid polymer with acid sites as an electrolyte so that the cell can easily be stacked. In the presence of water, the acid sites dissociate to provide the protons for conduction. Thus, the feed gases must be humidified to provide good proton conductivity. The management of water is one of the important issues in PEMFC. The operating temperature of PEMFC is below 100°C, typically around 80°C, which is set by the thermal stability and conductivity of polymeric PEMs. Some increase in operating temperature may be possible under pressurized conditions or with electrolytes that perform adequately at low relative humidity.

Direct Methanol Fuel Cell

Direct methanol fuel cell (DMFC) was developed in 1950s–1960s, based on the liquid alkaline or aqueous acid solution as the electrolyte. It converts the methanol directly into electricity, instead of using ''indirectly'' produced hydrogen from methanol through the reforming process. Today, DMFC commonly refers to as the one that employs PEM as the electrolyte. Fuel for DMFC is a dilute solution of methanol, usually 3–5 wt% in water. The size of DMFC can be considerably smaller than PEMFC because of the elimination of fuel processor, and complex humidification and heat management systems. The performance of DMFC is relatively low compared to that of PEMFC.

Molten Carbonate Fuel Cell

Molten carbonate fuel cell (MCFC) uses an electrolyte composed of a molten mixture of carbonate salts, e.g., lithium carbonate, potassium carbonate, and sodium carbonate, usually retained in a ceramic matrix, e.g., $LiAlO_2$. When heated to a temperature of around 650°C, these salts melt and become conductive to carbonate ions (CO_3^{2-}). Natural gas can be used directly without a fuel processor because of the high operating temperature (~650°C). Nickel (or Ni–Cr) and NiO have been used for anode and cathode materials for MCFC, respectively.

Solid Oxide Fuel Cell

Solid oxide fuel cell (SOFC) uses solid ceramic material, such as Y_2O_3 stabilized ZrO_2 (YSZ), as an electrolyte. As SOFC operates at high temperature (600–1000°C), a variety of fuels, e.g., hydrogen, methane, and carbon monoxide, can directly be utilized. The high temperature places severe constraints on material selection and results in difficult fabrication process. Co–ZrO (or Ni–ZrO) and SrO doped $LaMnO_3$ have often been used for anode and cathode materials, respectively.

PRINCIPLES OF PEMFC

Working Principle

A schematic diagram showing the nanostructure of the PEMFC is shown in Fig. 2.[23] PEMFC has received tremendous attention lately because of its potential in transportation and other applications, and is the focus of this review. The fuel, typically H_2, is fed continuously to the anode and the oxidant, usually O_2 from air, is fed continuously to the cathode. At the surface of the anode, the fuel is converted into protons (H^+) and electrons (e^-). The protons travel through a PEM, which prohibits the flow of electrons to the cathode. The electrons (e^-) are, thus, forced to travel through an external wire and deliver part of their energy to a ''load'' on their way to the cathode. At the cathode, the transferred protons and the energy depleted electrons combine with oxygen to produce water. Hydrogen and methanol are the common choices as fuels because of their relatively high activity at low temperatures (<100°C). Gaseous oxygen, or air, is the most common choice for the oxidant because it is readily and economically available. In principle, other fuels and oxidants can also be utilized. The electrochemical reactions take place at the surface of the electrode that is in contact with both the PEM as well as the backing of carbon paper or carbon cloth, which provides a conductive and porous medium that simultaneously allows the flow of reactant gases, liquid water produced, and electrons. The catalyst layer is usually 5–50 μm in thickness and consists of platinum nanoparticles, e.g., 2–4 nm in diameter, supported on the carbon black particle, e.g., 30 nm in diameter. The amount of the Pt on the carbon support is 10–40 wt%

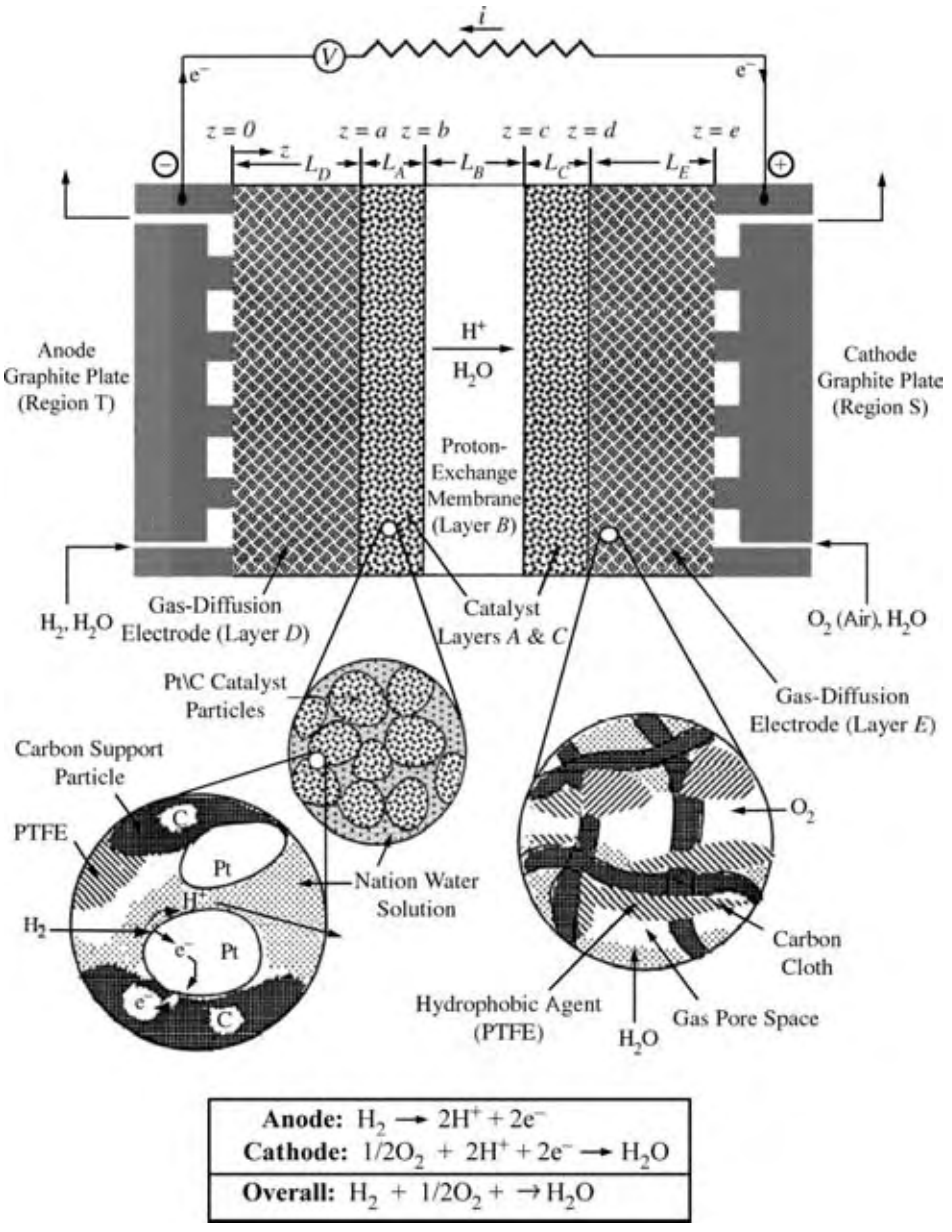


Fig. 2 A schematic representation of H_2/O_2 PEMFC.

and the loading of Pt is in the range of 0.1–0.4 mg/cm^2 MEA and 1–4 mg/cm^2 MEA for PEMFC and DMFC, respectively, where MEA (Membrane-Electrode-Assembly) refers to a proton exchange membrane, anode and cathode, and gas diffusion layers (GDLs). The electrode is mixed with an ionomer solution, e.g., Nafion, a common PEM, to provide a conduction medium for protons within the electrode layer. To bind the electrode particles together and facilitate the water management within MEA, the hydrophobic poly-tetrafluoroethylene (PTFE) solution may be added to the electrode. The protons produced at the anode travel through the membrane by virtue of the electric field created across the membrane between the anode and cathode.

Essential Concepts

Potential

An electrical potential exists across the interface between an electrode and an electrolyte, when partitioning of electrons and charged species occurs between the electrode and the electrolyte.[24–26] Consider the immersion of platinum electrode in acid electrolyte containing hydrogen ions as an example. Equilibrium would soon be established between the hydrogen ions in solution near the electrode, the electrons on the surface of the electrode, and the dissociated hydrogen atoms on the Pt surface

$$Pt + H^+ + e^- \leftrightarrows Pt - H \quad (1)$$

The H^+ represents a hydrated proton in the form of hydronium ion H_3O^+, which itself is hydrated. As a result, the electrons build up on the electrode surface in a layer, while the hydrated protons form an adjacent layer in the electrolyte, setting up a double layer of charges. Fig. 3 shows an example of the double layer formed near the negative electrode. The separation of charges leads to a potential difference between the electrode and electrolyte. The potential difference across the double layer at equilibrium would be the ideal potential of the electrode. However, the absolute potential between the electrode and electrolyte cannot readily be measured experimentally. Thus, the potential is typically given with reference to a reference electrode, where the potential corresponding to the reversible equilibrium between hydrogen gas at one atmosphere and hydrogen ion at unit activity is taken as zero at all temperatures. This defines the standard hydrogen potential (SHP) scale on which all other potentials are quoted.

Current

At equilibrium, Eq. (1) is in its dynamic condition, i.e., the rate of forward reaction equals the rate of the reverse reaction. If Eq. (1) departs from equilibrium, i.e., one of the directions of the reaction is faster than the other, then the net flow of protons and electrons, or current, develops. The anode is defined as the electrode at which the de-electronation reaction occurs and the cathode as the electrode at which the electronation reaction occurs. The rate of the electron transfer reaction can be written in terms of current, which is defined by the movement of electrical charges carried by electrons in an electronic conductor and by ions in an ionic conductor. The more the system is away from equilibrium, the higher the current. As the current is a measure of the rate of the reaction, the current is a function of the concentration of the species participating in the electrochemical reaction. The direction of conventional current is defined by the direction of flow for a positively-charged particle and, therefore, electron flow is opposite in direction to conventional current flow. Fig. 4 shows a schematic representation of the anode, the cathode, corresponding electron flows, net electron flow, current flows, and net current flow.

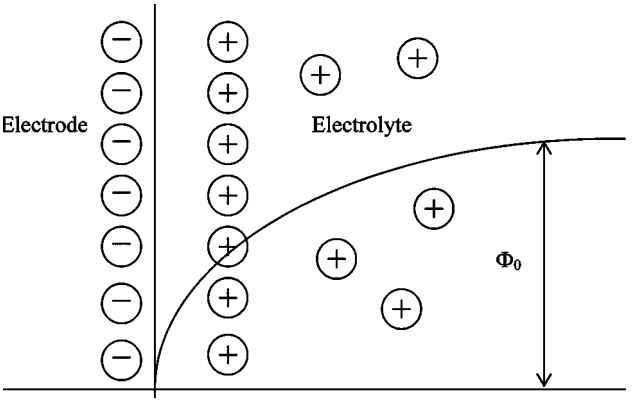


Fig. 3 A schematic of double layer at the surface of an electrode and potential gradient established between the layers of charges.

Overpotential

The overpotential η is defined as the difference in the potential of an electrode between its equilibrium and operating potentials when a current is flowing. Thus, higher overpotential leads to a higher current, the direction of which depends on the sign of the overpotential. The overpotential represents the extra potential energy needed to force the electrode reaction to proceed at a desired rate and in a desired direction. Charging the potential of metal electrode changes the energy of the surface electrons for the charge transfer by changing the position of Fermi level.[26,27] Fig. 5 illustrates the electron transfer by the applied potential, which alters the electron energy level of the metal electrode and, thus, facilitates the reduction or oxidation reaction. For a desired current density, the value of the overpotential also depends upon the inherent rate of the electrode reaction, i.e., a fast reaction will need a smaller overpotential than a slow reaction. The rate of an electrode reaction, of course, depends upon the temperature, potential, and activities of reacting species at the interface.

Electrochemical interface

Electrochemical reactions in a PEMFC occur at the ''three-phase interface,'' where the three necessary components for the reaction, i.e., reactant (e.g., gas), electron conductor (e.g., metal electrode), and ionic conductor (e.g., solid polymer electrolyte), meet. Fig. 6 schematically illustrates this in the PEMFC situation. The three components can actually meet, not only on the line defining the three-phase contact, but also in an area of two-phase contact, where gas molecules dissolve and diffuse through the electrolyte onto the electrode surface. As the electrochemical reaction depends on the concentrations of reacting species at the interface, the diffusion and solubility of the reactants and products in the electrolyte are important parameters in determining the overall rate of the electrochemical reactions. In practice, the polymer electrolyte, which is dissolved in solution, is mixed with electrode

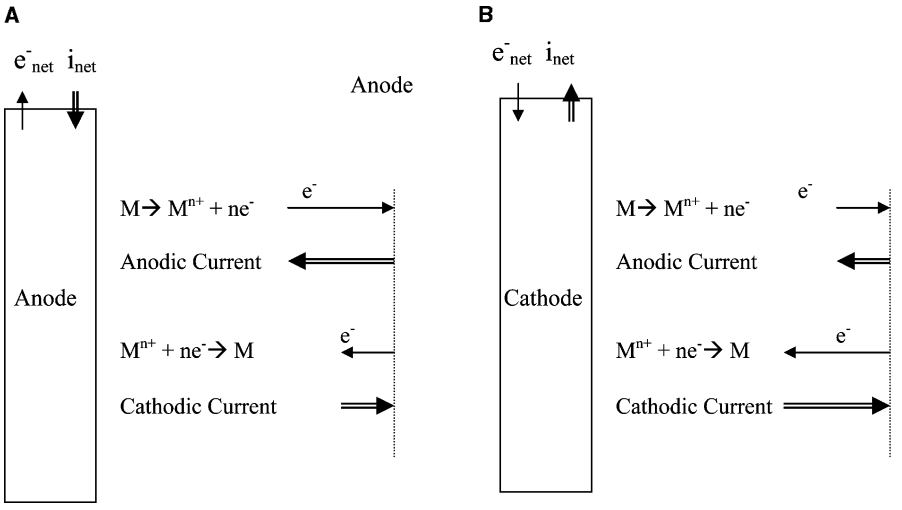


Fig. 4 A schematic representation of an anode (A) and a cathode (B) with their corresponding electron and current flows.

particles (Pt black or Pt/C) to increase the electrochemical interface to improve the overall rate of the reaction and thus, the performance of PEMFC. A thin layer (~1 nm thickness) of electrolyte is often envisioned to cover the electrode particle. However, the entire surface of electrode particle should not be covered by the electrolyte because this would preclude the electronic continuity. The electrode and the electrolyte phases need to maintain their continuity as conductors of charged reaction products, namely, electrons and protons, respectively.

Thermodynamics

A PEMFC converts the chemical energy of reactants into electrical energy by electrochemical reactions, which involve the transfer of electrons across the interface. Thermodynamic analysis describes fundamental aspects of the relationship between the thermodynamic and electric energies. An electrochemical reaction may be written in the generic form[23]

$$\sum_{\substack{i=1 \\ i \neq e^-}}^{n} \nu_{\rho i} A_i^{z_i} + \nu_{\rho e^-} e^- = 0 \tag{2}$$

where $\nu_{\rho i}$ is the stoichiometric coefficient of species i in reaction ρ, A_i is the chemical species i, z_i is the charge number of species i, $\nu_{\rho e^-}$ is stoichiometric coefficient of electrons in reaction ρ, and e^- represents the electrons participating in the reaction ρ. The electrochemical reactions should satisfy the charge balance, i.e.,

$$\sum_{\substack{i=1 \\ i \neq e^-}}^{n} \nu_{\rho i} Z_i = \nu_{\rho e^-} \tag{3}$$

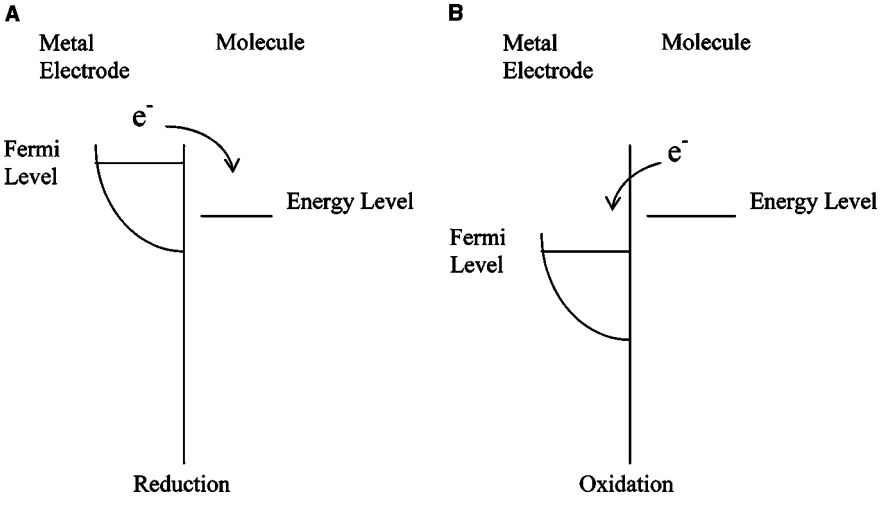


Fig. 5 Electron transfer at an inert metal electrode. The Fermi level of electrons in metal is altered by the applied potential so that the electron can move from the electrode to the molecule (A) or from the molecule to electrode (B).

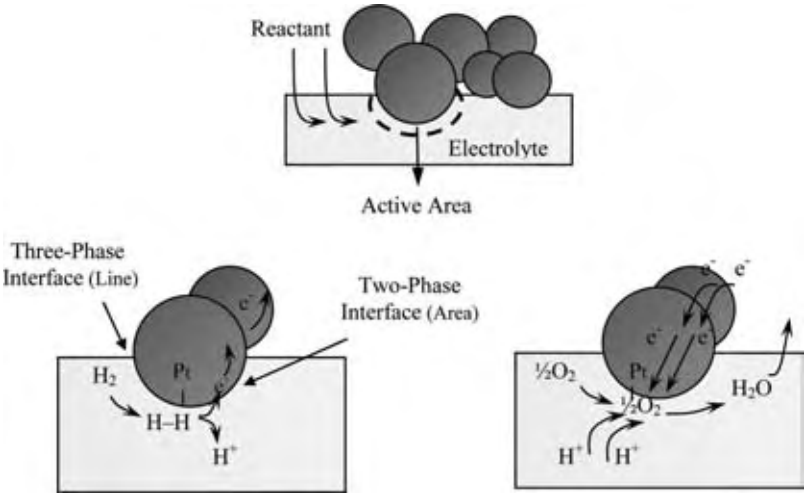


Fig. 6 A schematic illustration of the electrochemical interface and reactions at the anode and cathode in PEMFC. *(View this art in color at www.dekker.com.)*

The electrochemical reaction equilibrium condition in phase α can be written by separating the electrons from other reacting species as

$$\sum_{\substack{i=1 \\ i\neq e^-}}^{n} \nu_{\rho i}\mu_i^{e,\alpha} + \nu_{\rho e^-}\mu_{e^-}^{e,\alpha} = 0 \tag{4}$$

where $\mu_i^{e,\alpha}$ denotes the electrochemical potential of component i in phase α. When a metal (M) electrode and electrolyte solution (S) is in electrochemical reaction equilibrium, the electrochemical potential of a component i should be the same for both phases

$$\mu_i^{e,M} = \mu_i^{e,S} \tag{5}$$

Substitution of Eq. (5) into (4) gives

$$\sum_{\substack{i=1 \\ i\neq e^-}}^{n} \nu_{\rho i}\mu_i^{e,S} + \nu_{\rho e^-}\mu_{e^-}^{e,M} = 0 \tag{6}$$

The electrochemical potential of species i in phase α is the sum of its chemical potential and electrical potential

$$\mu_i^{e,\alpha} = \mu_i^{\alpha} + z_i F\phi^{\alpha} \tag{7}$$

where F is Faraday's constant (=96,485 C/equiv.) and ϕ^{α} is the electrical potential in phase α. Substitution of Eq. (7) into (6) gives

$$\sum_{\substack{i=1 \\ i\neq e^-}}^{n} \nu_{\rho i}\mu_i^{S} + \sum_{\substack{i=1 \\ i\neq e^-}}^{n} \nu_{\rho i}z_i(F\phi^{S}) + \nu_{\rho e^-}\mu_{e^-}^{M} + \nu_{\rho e^-}(-1)(F\phi^{M}) = 0 \tag{8}$$

Using the charge balance [Eq. (3)], Eq. (8) can be rearranged as

$$\Phi_0 = \frac{1}{\nu_{\rho e^-}F}\left(\sum_{\substack{i=1 \\ i\neq e^-}}^{n} \nu_{\rho i}\mu_i^{S} + \nu_{\rho e^-}\mu_{e^-}^{M}\right) \tag{9}$$

where the electrode potential Φ_0 is

$$\Phi_0 = \phi^{M} - \phi^{S} \tag{10}$$

Next, using the chemical potential of species i

$$\mu_i^{\alpha} = \mu_i^{0,\alpha} + RT\ln a_i^{\alpha} \tag{11}$$

where $\mu_i^{0,\alpha}$ is the chemical potential of species i when the activity a_i^{α} of i is 1, R is the gas constant, and T is the temperature. Substitution of this into Eq. (9) provides

$$\Phi_0 = \frac{1}{\nu_{\rho e^-}F}\left(\sum_{i=1}^{n} \nu_{\rho i}\mu_i^{0} + RT\sum_{\substack{i=1 \\ i\neq e^-}}^{n} \nu_{\rho i}\ln a_i^{S}\right) \tag{12}$$

where $a_{e^-}^{M} = 1$ is adopted. Defining the standard Gibbs free energy change for reaction ρ

$$\sum_{i=1}^{n} \nu_{\rho i}\mu_i^{0} = \Delta G_{\rho,\Phi=0}^{0} \tag{13}$$

Combining Eqs. (12) and (13) provides

$$\Phi_0 = \frac{1}{\nu_{\rho e^-} F}\left(\Delta G^0_{\rho,\Phi=0} + RT \ln \prod_{\substack{i=1 \\ i\neq e^-}}^{n} a_i^{\nu_{\rho i}}\right) \quad (14)$$

This reflects a relationship between the chemical energy ($\Delta G^0_{\rho,\Phi=0}$) and electrical potential (Φ_0) at equilibrium. For unit activities, Eq. (14) provides the standard equilibrium potential

$$\Phi^0_0 = \frac{\Delta G^0_{\rho,\Phi=0}}{\nu_{\rho e^-} F} \quad (15)$$

Thus, Eq. (14) may be expressed alternatively as

$$\Phi_0 = \Phi^0_0 + \frac{RT}{\nu_{\rho e^-} F} \ln \prod_{\substack{i=1 \\ i\neq e^-}}^{n} a_i^{\nu_{\rho i}} \quad (16)$$

Eq. (16) is commonly referred to as Nernst equation that provides a relationship between the standard potential at standard conditions (1 M, 1 atmosphere, and 298 K) and actual potential, which is the sum of the standard electrode potential and a correction term for the deviation from unit activities of the reactant and product species.

The electrochemical reactions and thermo dynamic relations are provided below for PEMFC and DMFC.

	$\Delta G^0_{\rho,\Phi=0}$ (kJ/mol)	Φ^0_0 (Volts)	
PEMFC			
Anode: $H_2(g) \leftrightarrows 2H^+ + 2e^-$	0.00	0.00	(17)
Cathode: $\frac{1}{2}O_2(g) + 2H^+ + 2e^- \leftrightarrows H_2O(l)$	−237.1	1.229	(18)
Overall: $H_2(g) + \frac{1}{2}O_2(g) \leftrightarrows H_2O(l)$	−237.1	1.229	(19)
DMFC:			
Anode: $CH_3OH(l) + H_2O(l) \leftrightarrows CO_2(g) + 6H^+ + 6e^-$	9.3	0.016	(20)
Cathode: $\frac{3}{2}O_2(g) + 6H^+ + 6e^- \leftrightarrows 3H_2O(l)$	−237.1	1.229	(21)
Overall: $CH_3OH(l) + \frac{3}{2}O_2(g) \leftrightarrows 2H_2O(l) + CO_2(g)$	−702	1.213	(22)

Nernst equations for the anode and cathode reaction of PEMFC, thus, are

$$\Phi_{0,A} = 0 + \frac{RT}{2F} \ln\left(\frac{a^2_{H^+}}{a_{H_2}}\right) \quad (23)$$

$$\Phi_{0,C} = 1.229 - \frac{RT}{2F} \ln\left(\frac{a_{H_2O}}{a^{1/2}_{O_2} a^2_{H^+}}\right) \quad (24)$$

For the overall reactions, the relation between the potential and activity is

$$V_0 = 1.229 + \frac{RT}{2F} \ln\left(\frac{a_{H_2} a^{1/2}_{O_2}}{a_{H_2O}}\right) \quad (25)$$

It is clear that the activity at the interface influences the potentials of both electrodes as well as terminal cell potential of PEMFC. Table 2 provides a list of standard potentials for electrode reactions and the overall reactions for the system of interest based on standard thermodynamic data.[28]

Kinetics

The rate of an electrochemical reaction ρ, r^*_ρ (mole/cm^2 catalyst surface area/s), as that of a chemical reaction, depends upon the temperature and activities of reacting species. In addition, in the case of the electrochemical reaction, the electric energy at the electrode–electrolyte interface also strongly influences the rate of the reaction. Thus, the rate of an electrochemical reaction is commonly written as the product of a reaction rate constant k^*_ρ and a function of activities of various species, a_A, a_B, ..., involved in the reaction

$$r^*_\rho = k^*_\rho(T,\Phi) \cdot f(a_i) \quad (26)$$

From the thermodynamic formulation of transition state theory (TTST),[29] the net rate for an elementary electrochemical reaction may be written as

$$r^*_\rho = \overrightarrow{r}^*_\rho - \overleftarrow{r}^*_\rho = \overrightarrow{k}^*_\rho \prod_{i=1}^{r} a_i^{-\nu_{\rho i}} - \overleftarrow{k}^*_\rho \prod_{i=r+1}^{n} a_i^{\nu_{\rho i}} \quad (27)$$

where $\overrightarrow{k}^*_\rho$ and $\overleftarrow{k}^*_\rho$ represent the rate constants for the forward and reverse reactions, respectively. It is assumed that the first r species are reactants and the remaining are products. The rate constant for the

Table 2 Thermodynamic values of H_2, O_2, and some hydrocarbon fuels of PEMFC interest

Reaction	ΔH (kJ/mol)	ΔG (kJ/mol)	Φ_0^0 (Volts)	Efficiency (%)
$H_2\ (g) \leftrightarrows H^+ + 2e^-$	0	0	0	
$\frac{1}{2}O_2\ (g) + 2H^+ + 2e^- \leftrightarrows H_2O\ (g)$		−228.6	1.185	
$H_2\ (g) + \frac{1}{2}O_2\ (g) \leftrightarrows H_2O\ (g)$	−241.8	−228.6	1.185	94.5
$H_2\ (g) + \frac{1}{2}O_2\ (g) \leftrightarrows H_2O\ (l)$	−285.8	−237.1	1.229	83.0
$CH_4\ (g) + 2H_2O\ (g) \leftrightarrows CO_2\ (g) + 8H^+ + 8e^-$		113.3	0.147	
$CH_4\ (g) + 2O_2\ (g) \leftrightarrows CO_2\ (g) + 2H_2O\ (g)$	−802.5	−801.1	1.038	99.8
$CH_3OH\ (g) + H_2O\ (g) \leftrightarrows CO_2\ (g) + 6H^+ + 6e^-$		−3.5	−0.006	
$CH_3OH\ (g) + \frac{3}{2}O_2\ (g) \leftrightarrows CO_2\ (g) + 2H_2O\ (g)$	−676.1	−689.3	1.191	102.0
$CH_3OH\ (l) + H_2O\ (l) \leftrightarrows CO_2\ (g) + 6H^+ + 6e^-$		9.3	0.016	
$CH_3OH\ (l) + \frac{3}{2}O_2\ (g) \leftrightarrows CO_2\ (g) + 2H_2O\ (l)$	−725.9	−702.0	1.213	96.7
$HCHO\ (g) + H_2O\ (g) \leftrightarrows CO_2\ (g) + 4H^+ + 4e^-$		−63.3	−0.164	
$HCHO\ (g) + O_2\ (g) \leftrightarrows CO_2\ (g) + H_2O\ (g)$	−526.7	−520.5	1.349	98.8
$HCOOH\ (l) \leftrightarrows CO_2\ (g) + 2H^+ + 2e^-$		−33.0	−0.171	
$HCOOH\ (l) + \frac{1}{2}O_2\ (g) \leftrightarrows CO_2\ (g) + H_2O\ (l)$	−254.3	−270.1	1.400	106.2
$C_2H_5OH\ (g) + 3H_2O\ (g) \leftrightarrows 2CO_2\ (g) + 12H^+ + 12e^-$		64.9	0.056	
$C_2H_5OH\ (g) + 3O_2\ (g) \leftrightarrows 2CO_2\ (g) + 3H_2O\ (g)$	−1277.6	−1306.7	1.129	102.3
$C_2H_5OH\ (l) + 3H_2O\ (l) \leftrightarrows 2CO_2\ (g) + 12H^+ + 12e^-$		97.3	0.084	
$C_2H_5OH\ (l) + 3O_2\ (g) \leftrightarrows 2CO_2\ (g) + 3H_2O\ (l)$	−1366.8	−1325.3	1.145	97.0

electrochemical reaction from TTST can be written as

$$k_\rho^* = \left(\kappa \frac{k_B T}{h\gamma^{\pm}}\right) \exp\left(\frac{-\Delta \overline{G}_\rho^{\neq,e}}{RT}\right) \tag{28}$$

where κ is the transmission coefficient, k_B is the Boltzmann constant, h is the Planck constant, $\gamma^{\pm}$ is the activity coefficient of transition state complex, and $\Delta \overline{G}_\rho^{\neq,e}$ is the Gibbs free energy of activation for electrochemical reaction. The Gibbs free energy can be divided into the chemical component and electrical component as

$$\Delta \overline{G}_\rho^{\neq,e} = \Delta \overline{G}_{\rho,\Phi=0}^{\neq} - \nu_{\rho e^-} \alpha_\rho \Phi \tag{29}$$

where $\Delta \overline{G}_{\rho,\Phi=0}^{\neq}$ represents the contribution of chemical part to the electrochemical activation energy, α_ρ is the transfer coefficient, i.e., the fraction of potential contributing to the electrical part of the activation energy of the forward reaction, and Φ represents the potential difference between the electrolyte and electrode. The Gibbs free energy of charged species is influenced by the electric potential formed between the electrode and the electrolyte through which the electrochemical charge transfer occurs. The transfer coefficient describes the position of maximum, or transition, in the diagram of the potential energy to distance along the reaction coordinates.[25] For oxidation reactions, the reaction rate constant can now be written as

$$\overrightarrow{k}_\rho^* = \left(\kappa \frac{k_B T}{h\gamma^{\pm}}\right) \exp\left(\frac{-\Delta \overrightarrow{G}_{\rho,\Phi=0}^{\neq} + \alpha_\rho \nu_{\rho e^-} F\Phi}{RT}\right) \tag{30}$$

$$\overleftarrow{k}_\rho^* = \left(\kappa \frac{k_B T}{h\gamma^{\pm}}\right) \times \exp\left(\frac{-\Delta \overleftarrow{G}_{\rho,\Phi=0}^{\neq} + (\alpha_\rho - 1)\nu_{\rho e^-} F\Phi}{RT}\right) \tag{31}$$

The rate of the electrochemical reactions is frequently expressed in terms of current density (A/cm^2) as

$$i^* = F\nu_{\rho e^-} r_\rho^* \tag{32}$$

This indicates that the anodic current is positive, while the cathodic current is negative. Using Eqs. (27) and (32) the net current density, or the measured current density, is

$$i^* = \overrightarrow{i^*} - \overleftarrow{i^*} \tag{33}$$

where $\overrightarrow{i^*}$ is the forward current density and $\overleftarrow{i^*}$ is the reverse current density for the electrochemical reaction ρ. When the potential of an electrode deviates from

equilibrium conditions, the overpotential η is given by

$$\eta = \Phi - \Phi_0 \tag{34}$$

Applying the rate constants [Eqs. (30) and (31)] to Eqs. (27), (32), and (33) gives the rate of forward and reverse reactions in terms of current densities $\overrightarrow{i^*}$ and $\overleftarrow{i^*}$

$$\overrightarrow{i^*} = F\nu_{\rho e^-}\left(\kappa\frac{k_B T}{h\gamma^{\pm}}\right)\exp\left(\frac{-\Delta\overrightarrow{G}^{\neq}_{\rho,\Phi=0}}{RT}\right) \times \exp\left(\frac{\alpha_\rho \nu_{\rho e^-} F\Phi_0}{RT}\right)\exp\left(\frac{\alpha_\rho \nu_{\rho e^-} F\eta}{RT}\right)\prod_{i=1}^{r} a_i^{-\nu_{\rho i}} \tag{35}$$

$$\overleftarrow{i^*} = F\nu_{\rho e^-}\left(\kappa\frac{k_B T}{h\gamma^{\pm}}\right)\exp\left(\frac{-\Delta\overleftarrow{G}^{\neq}_{\rho,\Phi=0}}{RT}\right) \times \exp\left(\frac{(\alpha_\rho - 1)\nu_{\rho e^-} F\Phi_0}{RT}\right) \times \exp\left(\frac{(\alpha_\rho - 1)\nu_{\rho e^-} F\eta}{RT}\right)\prod_{i=r+1}^{n} a_i^{\nu_{\rho i}} \tag{36}$$

Use of Eq. (33) and consolidating terms on the right hand side of Eqs. (35) and (36) except the overpotential part give

$$i^* = i_0^*\left[\exp\left(\frac{\alpha_\rho \nu_{\rho e^-} F\eta}{RT}\right) - \exp\left(\frac{(\alpha_\rho - 1)\nu_{\rho e^-} F\eta}{RT}\right)\right] \tag{37}$$

where i_0^* is the current density at equilibrium and commonly referred to as exchange current density, i.e.,

$$i_0^* = F\nu_{\rho e^-}\left(\kappa\frac{k_B T}{h\gamma^{\pm}}\right)\exp\left(\frac{-\Delta\overrightarrow{G}^{\neq}_{\rho,\Phi=0}}{RT}\right) \times \exp\left(\frac{\alpha_\rho \nu_{\rho e^-} F\Phi_0}{RT}\right)\prod_{i=1}^{r} a_i^{-\nu_{\rho i}} = F\nu_{\rho e^-}\left(\kappa\frac{k_B T}{h\gamma^{\pm}}\right)\exp\left(\frac{-\Delta\overleftarrow{G}^{\neq}_{\rho,\Phi=0}}{RT}\right) \times \exp\left(\frac{(\alpha_\rho - 1)\nu_{\rho e^-} F\Phi_0}{RT}\right)\prod_{i=r+1}^{n} a_i^{\nu_{\rho i}} \tag{38}$$

At electrochemical dynamic equilibrium, the net current is zero, i.e., the electron transfer process occurs at equal rates, both in the forward and in the reverse direction because of the formation of excess charges on both sides of electrode and electrolyte. The exchange current density is large when the chemical components of Gibbs free energies for forward and backward reactions are small [Eq. (38)]. Eq. (37), known as Butler-Volmer equation, is the most general expression describing the relation between the current and overpotential at a particular electrode.[24–27]

In PEMFC, the current density based on geometric area of MEA is frequently used for its convenience. The net current density i (A/cm^2 MEA) can be described as $i = \gamma_M i^*$ and $i_0 = \gamma_M i_0^*$, where γ_M is the roughness factor, i.e., the ratio of electrochemically active area per geometric MEA area. Thus, the current density based on geometric MEA area is given by

$$i = i_0\left\{\exp\left[\frac{\alpha_\rho \nu_{\rho e^-} F\eta}{RT}\right] - \exp\left[\frac{(\alpha_\rho - 1)\nu_{\rho e^-} F\eta}{RT}\right]\right\} \tag{39}$$

Therefore, the current density depends on the exchange current density (i_0), transfer coefficient (α_ρ), overpotential (η), and temperature (T). Fig. 7 represents typical current-overpotential curves based on Eq. (39). The net current is the result of the combined effects of the forward (anodic) and reverse (cathodic) currents. Although the Butler-Volmer equation for an electrochemical reaction in PEMFC is valid over the full potential range, simpler approximate equations may often be used for limited conditions. Thus, for the common value $\alpha_\rho = 1/2$, Eq. (39) becomes

$$i = 2i_0 \sinh\left(\frac{\nu_{\rho e^-} F\eta}{2RT}\right) \tag{40}$$

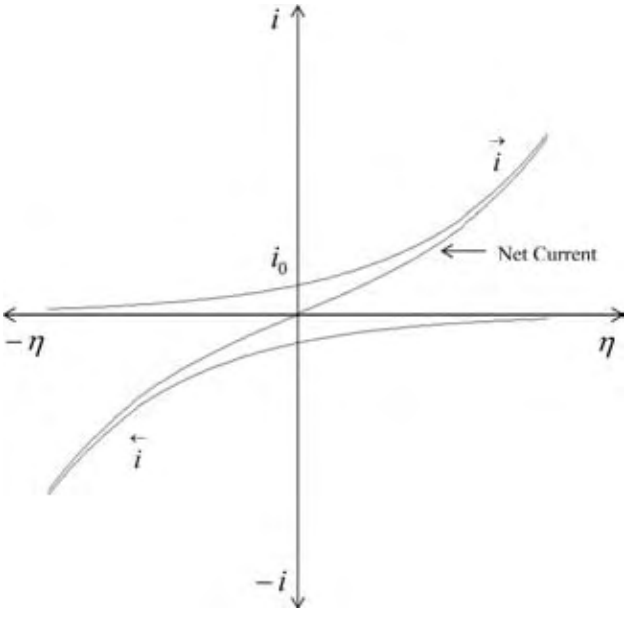


Fig. 7 A typical over-potential and current curve for electrochemical reactions. *(View this art in color at www.dekker.com.)*

For large overpotential η, i.e., $|\eta| \gg RT/F$, the second exponential term on the right hand side of Eq. (39) can be neglected compared with the first and gives

$$i \approx i_0 \exp\left(\frac{\alpha_\rho \nu_{\rho e^-} F\eta}{RT}\right) \tag{41}$$

Solving for η, Eq. (41) reduces to

$$\eta \approx \left(\frac{2.3RT}{\alpha_\rho \nu_{\rho e^-} F}\right) \log i - \left(\frac{2.3RT}{\alpha_\rho \nu_{\rho e^-} F}\right) \log i_0 = a + b \log i \tag{42}$$

Thus, the overpotential is logarithmically dependent on the current density. The parameter a and b are characteristics of electrochemical reactions and are readily obtained from a plot of η vs. $\log i$. The intercept of the straight line at $\eta = 0$ gives the exchange current density and the slope can provide α_ρ, the transfer coefficient. Eq. (42) is commonly referred to as the Tafel equation.

For very low overpotential, i.e., $|\eta| \ll RT/F$, the exponential term may be expanded in a series in Eq. (39) and approximated by a linear form

$$\eta \approx \left(\frac{RT}{\nu_{\rho e^-} F}\right) \frac{i}{i_0} \tag{43}$$

At very low overpotential region, the exchange current density can be readily obtained from the slope of the η versus i plot, which gives a line passing through the origin. It can be perceived from Eqs. (42) and (43) that electrochemical reaction with a large exchange current density needs low overpotential for a required current density, and vice versa.

Reaction Mechanisms

Hydrogen oxidation reaction

The hydrogen oxidation reaction (HOR) at the anode proceeds on Pt-based catalysts and is one of the simplest reaction systems.[30–36] Nonetheless, fundamental information of the mechanism and kinetics of HOR is still lacking. The most common mechanisms are the so-called Heyrovsky-Volmer and Tafel-Volmer mechanisms involving the following steps:

Tafel step:

$$S_T: \quad 2Pt_{(S)} + H_2 \leftrightarrows Pt\text{-}H_{ads} + Pt\text{-}H_{ads} \tag{44}$$

Heyrovsky step:

$$S_H: \quad Pt_{(S)} + H_2 \leftrightarrows Pt\text{-}H_{ads} + Pt + H^+ + e^- \tag{45}$$

Volmer step:

$$S_V: \quad Pt\text{-}H_{ads} \leftrightarrows Pt + H^+ + e^- \tag{46}$$

where $Pt_{(S)}$ is a free surface site on Pt and $Pt\text{-}H_{ads}$ is an adsorbed H-atom on the Pt active site. The Tafel-Volmer pathway involves S_T and S_V, i.e., the overall reaction (OR) is given by $OR = S_T + 2S_V$, while in the Heyrovsky-Volmer pathway, $OR = S_H + S_V$. The dissociative adsorption of H_2 is considered to be a rate-determining step. The overpotential for the HOR is relatively small at most practical current densities for a pure hydrogen feed. However, for reformed hydrocarbons, the anode feed may contain roughly 100 ppm CO, even after gas clean up in the fuel reforming section of a fuel cell power plant. Even at this low level, CO adsorbs very strongly on Pt surface, occupying the vast majority of the sites and, thus, increasing the anode overpotential substantially.[33,34] The CO coverage on the Pt surface depends on the temperature and CO concentration. In the presence of CO, a possible mechanism is[33]

$$Pt_{(S)} + CO \leftrightarrows Pt\text{-}CO_{ads} \text{ (high CO coverage)} \tag{47}$$

$$Pt_{(S)} + \frac{1}{2}H_2 \leftrightarrows (Pt\text{-}H_{ads}) \text{ (low H coverage)} \tag{48}$$

$$Pt_{(S)} + H_2O \leftrightarrows Pt\text{-}OH_{ads} + H^+ + e^- \tag{49}$$

$$Pt\text{-}(CO)_{ads} + Pt\text{-}OH_{ads} \leftrightarrows Pt_{(S)} + CO_2 + H^+ + e^- \tag{50}$$

Carbon monoxide preferably adsorbs on the Pt surface and inhibits the dissociative adsorption of H_2, which may be explained by the strong bond strength of Pt-CO compared to that of Pt-H. The last two reactions explain CO removal from the surface via electro-oxidation at high anode overpotentials. Alloying Pt with Ru improves CO tolerance through the facile removal of $Pt\text{-}(CO)_{ads}$ as[33,35,36]

$$Ru + H_2O \leftrightarrows Ru\text{-}OH_{ads} + H^+ + e^- \tag{51}$$

$$Ru\text{-}(OH)_{ads} + Pt\text{-}CO_{ads} \leftrightarrows Ru + Pt + CO_2 + H^+ + e^- \tag{52}$$

This enhancement is ascribed to both the activation of H_2O to OH_{ads} on Ru and a decrease in CO binding energy when Ru is added to the surface due to electronic effects of alloying (bifunctional mechanism). Raising the fuel cell temperature above 160°C can decrease the CO coverage remarkably. For a reformate gas containing CO, Pt–Ru electrodes are commonly used,[37,38] and a number of other Pt-alloys such as Pt–Mo,[39] Pt–Sn,[40] Pt–Rh,[41] and Pt–Re[42] have been studied for the HOR reaction.

Methanol oxidation reaction

The anode reaction mechanisms suggested for methanol oxidation reaction (MOR) in DMFC include:[43–46]

1. Absorption of methanol on the electrode;
2. Stepwise electro-dehydrogenation to eventually form adsorbed CO; and
3. Surface reaction of OH (resulting from water) with adsorbed CO to generate CO_2.

The electro-oxidation on Pt is, thus, considered to take place through the following steps:[43]

$$Pt_{(S)} + CH_3OH \leftrightarrows Pt\text{-}(CH_3OH)_{ads} \quad (53)$$

$$Pt\text{-}(CH_3OH)_{ads} \leftrightarrows Pt\text{-}(CH_2OH)_{ads} + H^+ + e^- \quad (54)$$

$$Pt\text{-}(CH_2OH)_{ads} \leftrightarrows Pt\text{-}(CHOH)_{ads} + H^+ + e^- \quad (55)$$

$$Pt\text{-}(CHOH)_{ads} \leftrightarrows Pt\text{-}(CHO)_{ads} + H^+ + e^- \quad (56)$$

$$Pt\text{-}(CHO)_{ads} \leftrightarrows Pt\text{-}(CO)_{ads} + H^+ + e^- \quad (57)$$

$$Pt_{(S)} + H_2O \leftrightarrows Pt\text{-}(OH)_{ads} + H^+ + e^- \quad (58)$$

$$Pt\text{-}(CO)_{ads} + Pt\text{-}(OH)_{ads} \leftrightarrows Pt\text{-}(COOH)_{ads} \quad (59)$$

$$Pt\text{-}(COOH)_{ads} \leftrightarrows Pt_{(S)} + CO_2 + H^+ + e^- \quad (60)$$

Thus, the reaction is poisoned by the formation of $Pt\text{-}(CO)_{ads}$ after complete methanol dehydrogenation. There have been intensive searches for other active materials, which can provide oxygen in its active form from water at low overpotentials, to increase the oxidation rate of chemisorbed CO. Pt–Ru alloys are reported as having excellent promotional effects,[47,48] and PtSn,[49] PtMo,[50] PtRuCo,[51] and PtRuIrOs[52] have also been studied for the MOR reaction of DMFC.

Oxygen reduction reaction

The oxygen reduction reaction (ORR) at the cathode

$$O_2 + 4H^+ + 4e^- \leftrightarrows 2H_2O \quad (61)$$

has also been studied extensively and a number of mechanisms have been proposed.[53–60] However, there is still considerable uncertainty in the ORR. A simple mechanism may be written for ORR in an acid electrolyte

$$Pt_{(S)} + O_2 \leftrightarrows Pt\text{-}(O_2)_{ads} \quad (62)$$

$$Pt\text{-}(O_2)_{ads} + H^+ + e^- \leftrightarrows Pt\text{-}(OOH)_{ads} \quad (63)$$

$$Pt\text{-}(OOH)_{ads} + H^+ + e^- \leftrightarrows Pt\text{-}(OHOH)_{ads} \quad (64)$$

$$Pt\text{-}(OHOH)_{ads} + H^+ + e^- \leftrightarrows Pt\text{-}(OH)_{ads} + H_2O \quad (65)$$

$$Pt\text{-}(OH)_{ads} + H^+ + e^- \leftrightarrows Pt_{(S)} + H_2O \quad (66)$$

The dissociative adsorption of the oxygen molecule is unlikely and the first electron transfer to the oxygen molecule [Eq. (63)] is considered to be the rate determining step.[59,60] The formation of H_2O_2 and various forms of platinum oxides on the surface reduces the theoretical potential obtainable by ORR. In fact, potential losses in PEMFC arise predominantly because of the sluggishness of the ORR. The structural sensitivity of ORR has been discussed[61–64] over the Pt-based electrodes. Pt/C is commonly used for the cathode material, and Pt-based alloys, such as PtCr[65] and PtNi,[66] have also been investigated.

Overall PEMFC Performance

The performance of PEMFC is often presented by the polarization curve that shows the voltage output as a function of current density. Fig. 8 shows a typical polarization curve of PEMFC. As the PEMFC processes charge-transfer reactions and the diffusion of the reactants to and products from the electrochemical interface, the transport and kinetics within the cell determine the polarization characteristics of PEMFC. In the practical PEMFC, the terminal cell potential V

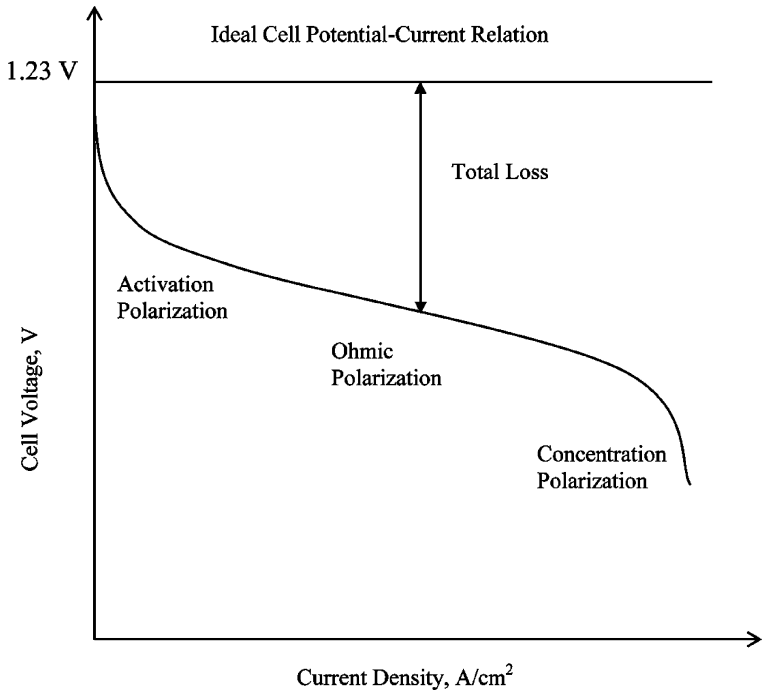


Fig. 8 A typical polarization curve of PEMFC.

(the output voltage) is a function of current density i, i.e., it decreases with increasing current density drawn from the cell because of the losses encountered in the various reactions and transport processes. Three major reasons of the potential loss are termed: concentration, activation, and ohmic overpotentials. First, the concentration overpotential is caused by the resistance of mass transport of reacting species. For example, the concentration of a reactant (e.g., O_2) in the vicinity of the interface may be lower than that in the bulk phase because of the hindrance to the transport of reactants, or the fact the products may not leave the surface instantaneously, e.g., H_2O in PEMFC. Second, the activation overpotential is caused by the finite rate of the electrochemical reactions, e.g., the reaction may proceed via several steps, one or more of which may be slow and irreversible. Third, ohmic overpotential is caused by the internal resistance of the phases to the transport of charged species. In most cases, the resistance of electrode to electronic conduction is relatively small compared to that of the electrolyte to ionic conduction because the electronic conductivity is much higher than the ionic conductivity in the electrolyte. A thin electrolyte layer having high proton conductivity, therefore, reduces the ohmic overpotential. In PEMFC, the transport of reactants across the electrolyte, i.e., hydrogen to cathode or oxygen to anode, may cause mixed potential at the electrodes, which may result in the departure from the ideal behavior expected at the electrodes. Under the conditions where all three types of overpotentials are considered, the terminal cell potential may be written as

$$V = V_0 - \sum_i^n |\eta_i| \tag{67}$$

where the second term on the right hand side includes all the three major types of overpotentials. The absolute symbol is used to prevent any confusion in the sign because the cathode side gives negative overpotentials due to negative currents by the electronation reactions. The overpotentials for PEMFC can, thus, be written as

$$\sum_i^n |\eta_i| = |\eta_{C,A}| + |\eta_{C,C}| + |\eta_{K,A}| + |\eta_{K,C}| + |\eta_M| + |\eta_I| \tag{68}$$

where the terms on the right hand side represent the concentration overpotential at the anode, concentration overpotential at the cathode, kinetic overpotential for an anode reaction, kinetic overpotential for a cathode reaction, membrane overpotential, and any interface contact overpotential, respectively. Fig. 9 schematically illustrates a PEMFC as an electrical network having an internal ideal voltage source coupled with the internal losses because of the resistances.

The overpotential increases with the current density and the potential–current relation of the cell can be obtained by solving transport and kinetic equations in the cell. Fig. 10 shows a schematic of

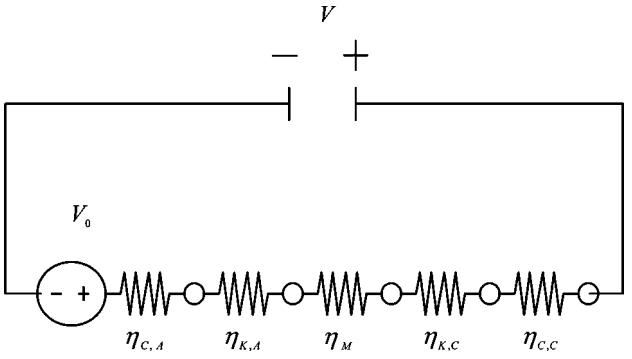


Fig. 9 A schematic representation of electrochemical reactions as an electrical network having an internal ideal potential source coupled with the losses because of the resistances. *(View this art in color at www.dekker.com.)*

the concentration distribution of reacting species across the MEA during PEMFC operation. A simple analytical model, based on Figs. 2 and 10, has been reported.[23] The basic conservation equations are presented as follows:

Anode chamber (Region *T*):

$$F_{\mathrm{T}}(c_{\mathrm{H}_2,T0} - c_{\mathrm{H}_2,T}) = N_{\mathrm{H}_2,z} \cdot A \tag{69}$$

where F_{T} (cm^3/sec) is the volumetric flow rate of the reactant (e.g., hydrogen) in the anode, $c_{\mathrm{H}_2,T0}$ and $c_{\mathrm{H}_2,T}$ are the concentration of H_2 entering and exiting the region *T*, respectively, $N_{\mathrm{H}_2,z}$ ($\mathrm{mol/cm}^2$/sec) is the flux of *i* in z-direction, and *A* (cm^2) is the geometric area of the MEA.

Gas diffusion layers (Layers *D* and *E*):

$$\frac{dN_{i,z}}{dz} = 0 \tag{70}$$

$$N_{i,z} = -D_i^{e,\alpha}\frac{dc_i}{dz} \tag{71}$$

where $D_i^{e,\alpha}$ denotes the effective diffusivity of species *i* in phase α. Fick's law is simply used for the diffusion of species *i* (e.g., hydrogen or oxygen) with an effective diffusion coefficient for the constant flux condition.

Electrodes (Layers *A* and *C*):

$$\frac{dN_{i,z}}{dz} = \nu_{\rho i} r'_{\rho} \tag{72}$$

$$N_{i,z} = -D_i^{e,\alpha}\left(\frac{dc_{i\alpha}}{dz} + \frac{z_i c_i F}{RT}\frac{d\phi^{\alpha}}{dz}\right) \tag{73}$$

where r'_{ρ} represent the rate per volume of electrode layer (rate/cm^3 catalyst layer). For electrochemical systems, the driving forces for diffusion are approximated as chemical potential gradients and electrostatic forces.

PEM electrolyte (Layer *B*):

$$\frac{di}{dz} = 0 \tag{74}$$

$$i = -\sigma_B \frac{d\phi^S}{dz} \tag{75}$$

where σ_B is the specific ionic (proton) conductivity of the electrolyte (layer *B*) and ϕ^S represents the potential of the solid electrolyte. If it is assumed that the specific proton conductivity remains constant, Eq. (75) becomes

$$\phi^{S,B}(b) - \phi^{S,B}(c) = \left(\frac{L_B}{\sigma_B}\right) \cdot i \tag{76}$$

where L_B is the thickness of the electrolyte. Hence, the overpotential increases with the current drawn and the thickness of electrolyte, and decreases with the specific proton conductivity.

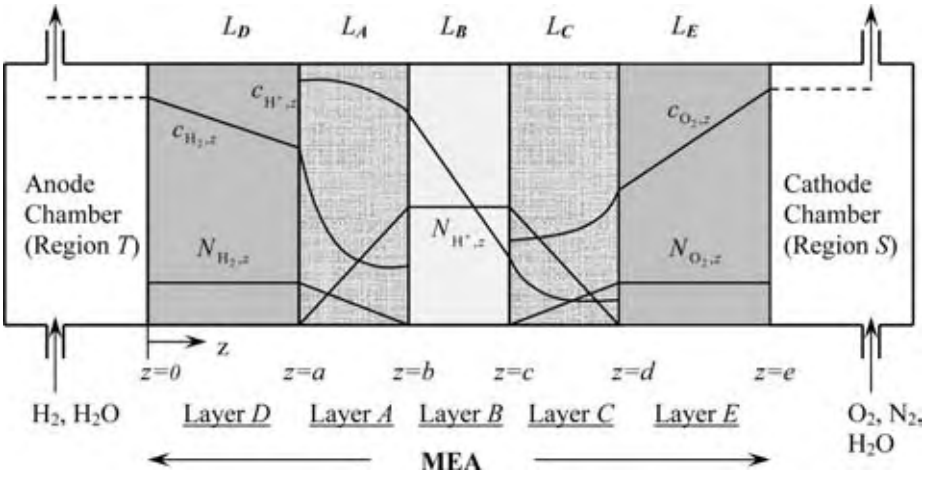


Fig. 10 A simplified one-dimensional picture of the concentration distribution of reacting species across MEA in the operating conditions of PEMFC. *(View this art in color at www.dekker.com.)*

Cathode chamber (Region S):

$$F_S(c_{O_2,S0} - c_{O_2,s}) = N_{O_2,z} \cdot A \tag{77}$$

where F_S is the volumetric flow rate of the oxidant (oxygen or air), and $c_{O_2,S0}$ and $c_{O_2,S}$ represent the concentration of oxygen at the inlet and outlet of the cathode chamber (region S), respectively. These equations can be solved for the case of no transport limitations in the catalyst layers, providing the overpotentials related to the current densities as

$$\eta_A = \eta_{C,A} + \eta_{K.A} = \frac{RT}{\alpha_A F}\sinh^{-1}\left\{\frac{1}{2}\left[\frac{i_A/i_{A,0}}{1 - i_A/i_{A,L}}\right]\right\} \tag{78}$$

$$-\eta_C = -(\eta_{C,C} + \eta_{K.C}) = \frac{RT}{\alpha_C F}\sinh^{-1}\left\{\frac{1}{2}\left[\frac{i_C/i_{C,0}}{1 - i_C/i_{C,L}}\right]\right\} \tag{79}$$

$$\eta_M = i\left(\frac{L_B}{\sigma_B}\right) \tag{80}$$

$$\eta_I = iR_I \tag{81}$$

where $i_{A,0}$ (or $i_{C,0}$) is the exchange current density at the anode (or cathode), $i_{A,L}$ (or $i_{C,L}$) is the limiting current density at the anode (or cathode), $\eta_{C,A}$ (or $\eta_{C,C}$) is the concentration overpotential because of the diffusion of hydrogen at the anode (or oxygen at the cathode), and $\eta_{K,A}$ (or $\eta_{K,C}$) is the overpotential because of the kinetics at the anode (or cathode) surface. Usually, $\eta_{K,C}$ is the dominant overpotential in PEMFC. The limiting current density is the maximum current density that can be achieved for an electrode reaction at a given concentration of the reactant. As the current density is increased, the surface concentration of the reactant is decreased by virtue of diffusional limitations. Finally, a limiting current situation is reached, when the further change of the electrode potential cannot increase the reactant flux because of diffusional limitations, i.e., no more reactant can be supplied to the electrode to sustain the desired current.

The terminal potential, or potential difference at terminals, can be obtained as

$$V = \frac{1}{F}(\mu^e_{e^-,A} - \mu^e_{e^-,C}) = \frac{1}{F}\left((\mu_{e^-,A} + (-1)F\phi^{M,A}) - (\mu_{e^-,C} + (-1)F\phi^{M,C})\right) \tag{82}$$

where $\mu^e_{e^-,A}$ and $\mu^e_{e^-,C}$ denote the electrochemical potential of electrons at the anode and cathode, respectively. If it is assumed that the chemical potential of electrons at the anode and cathode are equal, the terminal potential is given by

$$V = \phi^{M,C} - \phi^{M,A} \tag{83}$$

where $\phi^{M,C}$ and $\phi^{M,A}$ represent the potential of the metal (M) electrode at the cathode (C) and anode (A), respectively. It is shown that the electrode potential is given by

$$\Phi_M = \phi^M - \phi^S = \Phi_{0,M} + \eta_M \tag{84}$$

By the substitution of Eq. (84) into (83) and followed by the arrangement of terms, the terminal cell potential can be obtained as

$$V = (\Phi_{0,C} - \Phi_{0,A}) + \eta_C - \eta_A - (\phi^{S,A} - \phi^{S,C}) \tag{85}$$

where $(\Phi_{0,C} - \Phi_{0,A})$ is the open circuit potential V_0 and the last term on the right hand side can be written as

$$\phi^{S,A} - \phi^{S,C} = (\phi^{S,A} - \phi^{S,B}(b)) + (\phi^{S,B}(b) - \phi^{S,B}(c)) + (\phi^{S,B}(c) - \phi^{S,C}) \tag{86}$$

The first and third parentheses on the right hand side of Eq. (86) represent the electrodes and electrolyte contact resistances and the second term represent the bulk resistance of the electrolyte. Substitution of Eqs. (78–86) into (85) gives the following current–voltage relation, or polarization equation, for PEMFC

$$V = V_0 - \frac{RT}{\alpha_A F}\sinh^{-1}\left\{\frac{1}{2}\left[\frac{i/i_{A,0}}{1 - i/i_{A,L}}\right]\right\} - \frac{RT}{\alpha_C F}\sinh^{-1}\left\{\frac{1}{2}\left[\frac{i/i_{C,0}}{1 - i/i_{C,L}}\right]\right\} - i\left(\frac{L_B}{\sigma_B}\right) - iR_I \tag{87}$$

The anode overpotential for hydrogen oxidation reaction is low, i.e., $\eta_A \approx 0.05\,\text{V}$, so that Eq. (78) may be linearized, while the cathode overpotential is very high, i.e., $\eta_C \approx -0.4\,\text{V}$ in normal hydrogen fuel cells, and η_C vs. i is highly nonlinear, often approximated by the Tafel expression.

Fig. 11 shows a typical polarization curve for PEMFC along with potential losses for different regions of current densities. The corresponding power density can be obtained by multiplying i in Eq. (87), i.e., $P = V \cdot i$, where P is power density (W/cm^2). To produce a high power density, the polarization curve needs to be flattened by reducing the various types of the overpotentials. In the practical PEMFC, the operating terminal voltage is in the vicinity of 0.5–0.7 depending on the shape, e.g., flatness, of the

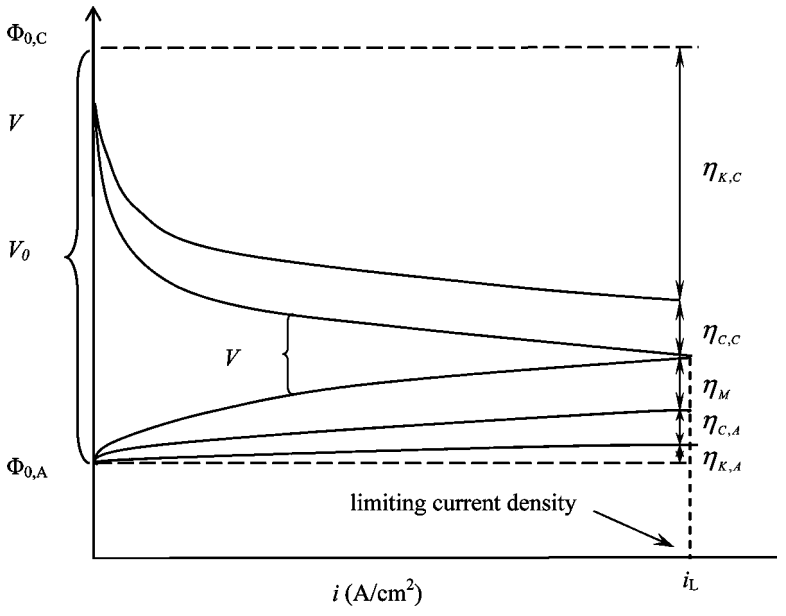


Fig. 11 A typical polarization curve for PEMFC along with potential losses because of the overpotentials with current densities.

polarization curve. This is discussed in more detail below.

In analogy to the linear Ohm's law, it may useful to define a *differential* specific resistance

$$R \equiv \frac{d(V_0 - V)}{di} \tag{88}$$

Use of Eqs. (67), (68), (78), and (79) gives

$$R = \frac{d}{di}(|\eta_C| + |\eta_A| + |\eta_M| + |\eta_I|) \tag{89}$$

Differentiation of overpotentials given in Eqs. (78)–(80) with respect to i provides the differential specific resistances at the anode and cathode as

$$R_A = \frac{RT}{(2\alpha_A F i_{A,0})(1 - i/i_{A,L})^2\left[1 + \frac{1}{4}\left(\frac{i/i_{A,0}}{1 - i/i_{A,L}}\right)^2\right]^{1/2}} \tag{90}$$

$$R_C = \frac{RT}{(2\alpha_C F i_{C,0})(1 - i/i_{C,L})^2\left[1 + \frac{1}{4}\left(\frac{i/i_{C,0}}{1 - i/i_{C,L}}\right)^2\right]^{1/2}} \tag{91}$$

and, of course, $R_M = L_M/\sigma_M$. It can be shown from a current versus differential resistance plot that the kinetic resistance at the cathode, the membrane resistance, and the diffusional resistance dominate at low, intermediate, and high current density regions, respectively.

Overall PEMFC Efficiency

The application of the first law of thermodynamics to a PEMFC provides the relation between enthalpy change and heat added to the system, as well as the electric work done by the system, in the absence of kinetic and potential energy changes

$$\Delta H = Q - W_E \tag{92}$$

where the electric work done by the PEMFC, $W_E = nFV$, where n is the number of electrons transferred. To determine the maximum useful electric work possible, Eq. (92) is applied to a reversible process

$$-W_{E,rev} = \Delta H - Q_{rev} \tag{93}$$

For a reversible change at constant temperature and pressure, the second law of thermodynamics provides the heat transferred to the system as

$$Q_{rev} = T\Delta S \tag{94}$$

Another relevant thermodynamic relation between the state functions is

$$\Delta G = \Delta H - T\Delta S \tag{95}$$

Combining Eqs. (93–95) gives

$$W_{E,rev} = -\Delta G = nFV_0 \tag{96}$$

Therefore, the changes in Gibbs free energy is equal to the maximum electrical work (or energy) attainable at the given temperature and pressure.

The efficiency of energy conversion devices is given by

$$\varepsilon = \frac{\text{useful energy output}}{\text{energy input}} = \frac{W_E}{-\Delta H} \quad (97)$$

If the PEMFC works ideally, the change of Gibbs free energy of the reaction can be completely converted to electrical energy. Thus, the ideal or thermodynamic efficiency of PEMFC can be given as

$$\varepsilon_0 = \frac{W_{E,rev}}{-\Delta H} = \frac{\Delta G}{\Delta H} = 1 - \frac{T\Delta S}{\Delta H} \quad (98)$$

If the change in the Gibbs free energy ΔG is greater than the changes in the enthalpy ΔH, the thermodynamic efficiency can exceed 1 (or 100%). Table 2 also lists the thermodynamic efficiencies for fuel cell reactions of interest under standard conditions.

From the data provided in the Table 2, the reversible heat generated in PEMFC

$$Q_{rev} = T\Delta S = \Delta H - \Delta G \quad (99)$$

is $-48.7\,\text{kJ/mol}$ for liquid water product at standard state PEMFC reactions, indicating that heat is released from the system. If the potentials of the anode and cathode are at their equilibrium values, i.e., $\Phi_{0,A}$ and $\Phi_{0,C}$, and also the terminal voltage V is independent of current, then the actual voltage would be the same as the ideal voltage, i.e., $V = V_0$. In practice, however, the operating potential V deviates significantly from the equilibrium value V_0 because of the reasons discussed above. The voltage efficiency is usually defined as

$$\varepsilon_V = \frac{V}{V_0} \quad (100)$$

The potential loss, i.e., the deviation of the overall potential V from V_0, is described in the previous section.

In addition to the potential losses, the current loss from the theoretical conversion of reactants could be significant. The current (or Faradic) efficiency is defined as

$$\varepsilon_C = \frac{I}{I_T} \quad (101)$$

where I_T and I represent the theoretical current calculated based on the consumed reactants and the actual current, respectively. This loss may be caused by side reactions, loss or accumulation of reaction intermediates, chemical decompositions of reactants, and chemical (non-Faradic) reactions between reactants because of dissolution and diffusion through electrolyte layer to the opposite electrode. The PEMFC efficiency is then the product of the efficiencies described above, i.e.,

$$\varepsilon = \varepsilon_0\varepsilon_V\varepsilon_C \quad (102)$$

Frequently, ε_C is close to 1.0 for PEMFC, while the thermodynamic efficiency ε_0 is fixed. Therefore, ε directly varies with the operating voltage. This indicates the great importance of operating at as high a potential as feasible. Thus, even though a smaller fuel cell would result if it is operated at around 0.5 V corresponding to the maximum in power density, it is common to operate it in the range of 0.6–0.7 V, where the efficiency is higher, even though the fuel cell is bigger for a given power density. Assuming $\varepsilon_C = 1.0$, the efficiency of PEMFC is given by

$$\varepsilon = \frac{nFV}{-\Delta H} = \frac{nF}{-\Delta H}(V_0 + \eta_C - \eta_A - \eta_{PEM} - \eta_I) \quad (103)$$

Using Eq. (87), the efficiency can be obtained, i.e., for $n = 2$, by

$$\varepsilon = \frac{2F}{-\Delta H}\left[V_0 - \frac{RT}{\alpha_A F}\sinh^{-1}\left\{\frac{1}{2}\left[\frac{i/i_{A,0}}{1 - i/i_{A,L}}\right]\right\} - \frac{RT}{\alpha_C F}\sinh^{-1}\left\{\frac{1}{2}\left[\frac{i/i_{C,0}}{1 - i/i_{C,L}}\right]\right\} - i\left(\frac{L_B}{\sigma_B}\right) - iR_I\right] \quad (104)$$

The heat generated by a fuel cell per kW electricity produced can be obtained by

$$\overline{Q} = \frac{1}{\varepsilon} - 1 \quad (105)$$

Also, the amount of hydrogen required to produce 1 kW electricity can be given by

$$\bar{n}_{H_2} = \frac{1}{-\Delta H}\left(\frac{1}{\varepsilon}\right) \quad (106)$$

Thus, the higher the operating voltage and efficiency, the lower the heat loss and hence, the smaller the heat exchanger, and the lower the amount of H_2 needed, and the smaller the balance of plant.

MATERIALS FOR PEMFC

Electrode

The materials for PEMFC electrodes should have good electrical conductivities and be stable in contact with electrolyte. Platinum-based electrodes have shown excellent electrochemical activities for PEMFC.

To reduce the cost of precious Pt metal as well as improve the electrochemical activity, nano-sized Pt particles are dispersed on supports, e.g., Vulcan XC-72 (30 nm), by various techniques.[67–70] The carbon-based support has sufficient electron conductivity and inhibits agglomeration of Pt particles. However, the growth of particles during the PEMFC operation may occur due to crystallite migration and local heat generated by the electrochemical reactions. Nanomaterials such as carbon nanotubes have recently been used as supports for platinum nanoparticles.[71] Fig. 12 shows TEM (transmission electron microscope) images of Pt nanoparticles supported on carbon black. There have been continuous efforts to reduce the amount of expensive platinum loadings for PEMFC electrodes.[72–75] As the electrochemical reactions occur only on the interface between the electrode and electrolyte, the electrodes should be designed to contact the electrolyte to provide high interface area with high rates of transport of the reacting species. It is estimated that even in state-of-the art designs, only roughly 20% of available platinum surface is utilized.[69] Thus, the development of new methodologies for producing a large interface area as well as the highly active electrode materials for electrochemical reactions one of the most challenging tasks in PEMFC. For a reformate containing CO, as mentioned previously, the platinum is usually alloyed with other metals, e.g., ruthenium, to tolerate to the CO-poisoning on the surface of the electrode. PtRu/C (Pt : Ru = 1 : 1 atomic ratio) electrode has shown an excellent performance for PEMFC containing 10–100 ppm CO, when compared with pure Pt.[32] For the DMFC, which has virtually the same problem of strongly adsorbed CO on the surface of the electrode, a number of the ratios of PtRu alloy have been studied to find an optimum composition for DMFC.[46]

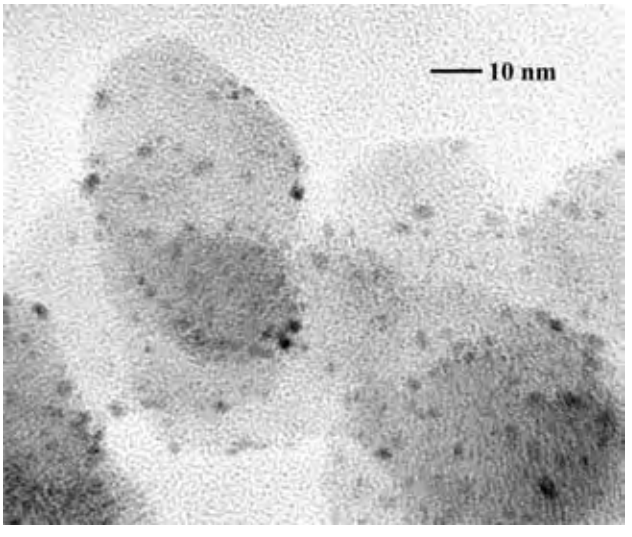


Fig. 12 A TEM image of 20% Pt/C nanoparticles taken at the author's laboratory.

Electrolyte

A solid polymer electrolyte, i.e., a PEM, conducts protons from anode to cathode and keeps the reactant gases from mixing directly. Suitable materials for electrolytes, thus, need to meet certain requirements such as stability (chemical, thermal, and mechanical) and low gas permeability over fuel cell operating conditions, in addition to excellent proton conductivity, low cost, and mechanical strength. The PEM electrolytes developed so far can be classified into three categories:[76–80] i) perfluorinated polymers, e.g., Nafion®; ii) partially fluorinated polymers, e.g., poly-α, β, β-trifluorostyrene; and iii) hydrocarbon polymers, e.g., poly(benzimidazole) or PBI. In addition, composite membranes, which can be obtained via modification of polymer (host membrane) by the incorporation of inorganic solid acids, may be identified as a new family of PEM.[80,81] Hydrocarbon membranes are relatively cost effective and the composite membranes are developed for high temperature PEMFC operation, which can provide improved kinetics at the electrodes, as well as higher tolerance to contaminant gases.[82,83]

Nafion and its other perfluorosulfonic acid (PFSA) relatives have so far demonstrated that they meet the key requirements, and hence have been widely used as PEM materials. However, they are expensive, effective at low temperature (100°C), highly permeable to liquid, and degraded during the operation. The properties and structure of Nafion have been extensively studied.[84,85] Fig. 13A shows a chemical structure of Nafion in which the backbone structure is similar to that of Teflon (PTFE). The backbone provides mechanical strength and is strongly hydrophobic, while the proton conducting sulfonic acid group is hydrophilic and, thus, nanoseparation is obtained when water is introduced into the Nafion. There have been a number of models for the nanostructures of Nafion including a cluster-network model, an elastic model, a three-phase model, and a variety of recent models based on a wide range of experimental techniques.[85] Nafion is classified by the equivalent weight (EW), defined by the weight of dry Nafion per mole of sulfonic acid groups, and an EW of 1100 is widely utilized. Nafion is commonly labeled with numbers such as Nafion 117 (or 115, or 112), referring that its EW is 1100 and thickness is 0.007 inch (or 0.005, or 0.002 inch). Nafion is usually pretreated prior to use in PEMFC and the pretreatment can also provide different forms, namely, N, S, and E forms. The N (normal) and S (shrunken) forms are the membranes

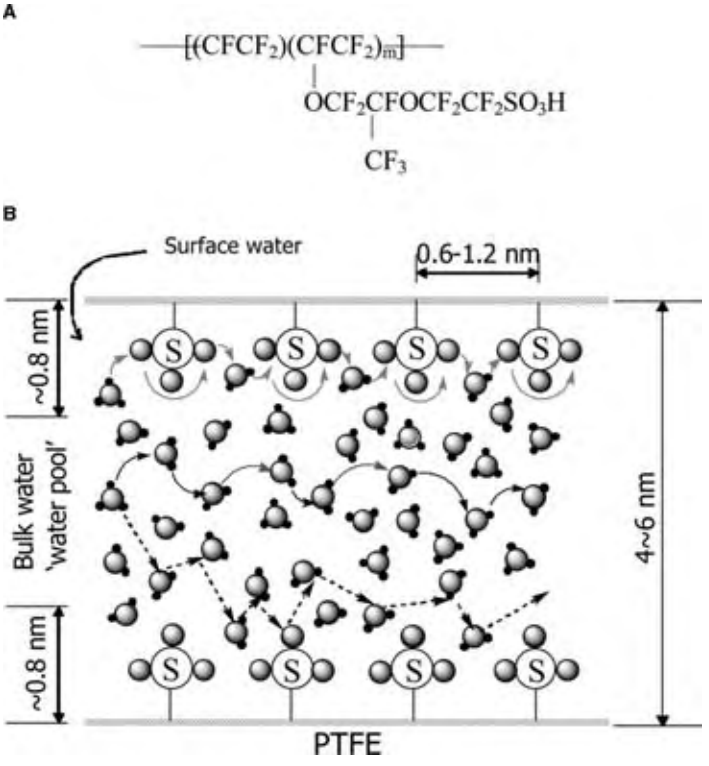


Fig. 13 A chemical structure (A) in which the relationship between EW and m is given by EW = 100 m + 446, and a simplified physical picture (B) showing proton transport, in a pore of Nafion. *(View this art in color at www.dekker.com.)*

dried at 80°C and 105°C, respectively. The E form is the one treated by increasing the temperature to around the glass transition temperature of Nafion (111°C) to allow the polymer chains to reorient themselves in the presence of water and then cleaned in a boiling 3% H_2O_2 solution, followed by boiling in 0.5 M H_2SO_4 to ensure full protonation, and finally in deionized water. The proton conductivity of Nafion is highly dependent upon the hydration level.[86,87] Thus, an understanding of water uptake in Nafion is very important to design electrolytes as well as to operate PEMFC at its optimum conditions. A thermodynamic model for water sorption has been developed in which the water uptake was determined by a balance between the internal pressure of water within the pores and the elasticity of the polymer matrix.[88] The state of water molecules in Nafion could be divided into two types, i.e., the strongly or chemically bound molecule to the sulfonic acid site and physically equilibrated water molecules. As the water generated at the cathode is not enough to humidify the membrane, water is usually supplied with the feed gases to provide proper proton conductivity. However, it is not desirable to keep too much water in the cell because of flooding that prohibits the transport of reactant gases to the active site for the electrochemical reaction. Thus, the management of water in the cell is one of the key issues to address to achieve enhanced fuel cell performance.[89–91] The distribution of water within the cell depends on the sorption characteristics of the polymer membrane, the drag coefficient of water molecule that accompany protons that transfer from anode to cathode, the diffusion of water because of concentration gradient within the cell, and the current density that affects the rate of generation of water at the cathode side by ORR, etc. The formation of water at the cathode and the drag of water molecules, i.e., 1–2.5 H_2O/H^+, may induce a back diffusion of water from the cathode to the anode.[92] There are some theoretical treatments of water transport in PEMFC.[93–95]

PEM separates the reactant gases and prohibits the transport of gases from the anode to the cathode, and vice versa. However, the gases dissolve and permeate through the PEM. The permeation of gases through the electrolyte is not favorable in a sense that it leads a loss of fuels and causes mixed potentials at both electrodes. However, it can provide appreciable concentration of reactants on the electrochemical interface by two phase contacts as explained previously. The solubility of hydrogen and oxygen gases in Nafion depends upon the humidification level and temperature.[96–98]

The mechanism of proton conduction in Nafion has been studied.[99–102] The proton transport can proceed via Grotthuss or structural diffusion mechanism that requires structural orientation of water molecules, and *en masse* or vehicle mechanism by which hydrated proton, e.g., H_3O^+, migrates by an ordinary diffusion process.[100,101] The interaction of protons with the membrane and water would determine the dynamics of proton environment, and thus, the proton conduction

through the membrane. Although there have been various proton transport mechanisms suggested,[103–106] a development of a complete mechanism of proton transport in PEM is still a challenging task because of the complex nanostructure and the inhomogeneous nature of PEM when hydrated. Fig. 13(B) shows a simplified physical picture of Nafion showing a possible proton transport mechanism in its fully hydrated state.[100]

The chemical stability of PEMs is an important matter for its long-term use in PEMFC. Unfortunately, PEMs are degraded during the course of PEMFC operation, which can be determined by the detection of the decomposition products, e.g., fluorine from Nafion, in the product water.[107–109] The crossover of oxygen gas from the cathode to anode may form H_2O_2 at the anode, which can decompose to give $^{\bullet}OH$ or $^{\bullet}OOH$ radicals.[107] The attack of these radicals may initiate the polymer degradation at the anode side, which is evidenced by the observation of the increase of the degradation rate, when operating at the open circuit, or very low current density conditions, which facilitate the diffusion of oxygen gas.[109] The radicals can attack any H-containing bonds in the polymer, e.g., for the end group of CF_2COOH[107,110]

$$R_f - CF_2COOH + {}^{\bullet}OH \rightarrow R_f - CF_2^{\bullet} + CO_2 + H_2O \quad (107)$$

$$R_f - CF_2^{\bullet} + {}^{\bullet}OH \rightarrow R_f - CF_2OH \rightarrow R_f - COF + HF \quad (108)$$

$$R_f - COF + H_2O \rightarrow R_f - COOH + HF \quad (109)$$

As the peroxide radical is considered as the main reason for the polymer degradation, a diluted solution of hydrogen peroxide (3% H_2O_2) was used to investigate the degradation process.[111] A more severe degradation at the cathode side, rather than anode side, of the membrane has been reported.[112] The radicals, such as $^{\bullet}OH$ or $^{\bullet}OOH$, which could be formed during ORR, may cause the degradation process at the cathode side of the membrane. The durability of PEM depends on the mechanical stability (swelling, thickness change), operating conditions, such as temperature, pressure, water management, hydration level, and impurities of reactant gases, and materials contacting the membrane, such as gasket and bipolar plates.

Gas Diffusion Layers

The gas diffusion layers, one next to the anode and the other next to the cathode, are usually made of a porous carbon paper or carbon cloth, typically 100 μm to 300 μm thick. Fig. 14 shows a porous GDL made of carbon paper, which is partially covered by catalyst layer. The porous nature of the backing layer ensures effective diffusion of feed and product components to and from the electrode on the MEA. The correct balance of hydrophobicity in the backing material, obtained by PTFE treatment, allows the appropriate amount of water vapor to reach the MEA, keeping the membrane humidified while allowing the liquid water produced at the cathode to leave the cell. The permeability of oxygen in the GDL affects the limiting current density of ORR, and thus the performance of PEMFC.[113]

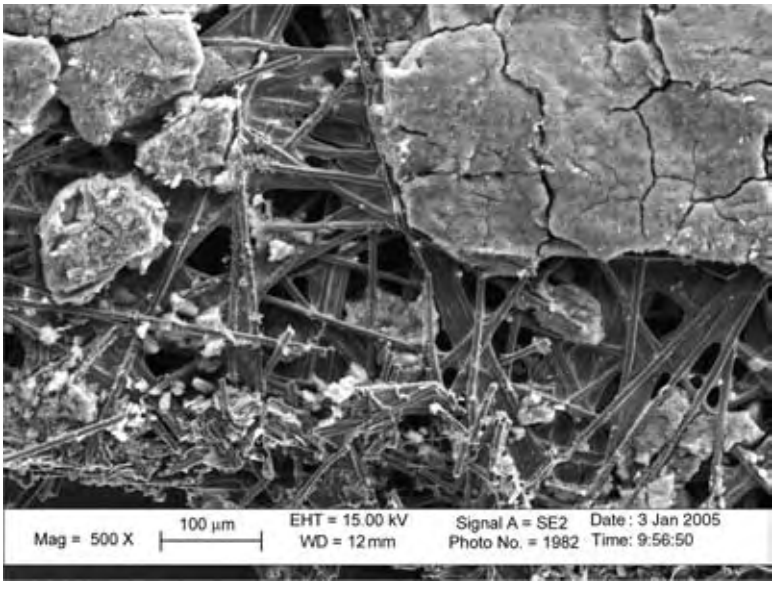


Fig. 14 A SEM (scanning electron microscopy) image of GDL made of carbon paper, which is partially covered by catalyst layer (taken at the author's laboratory).

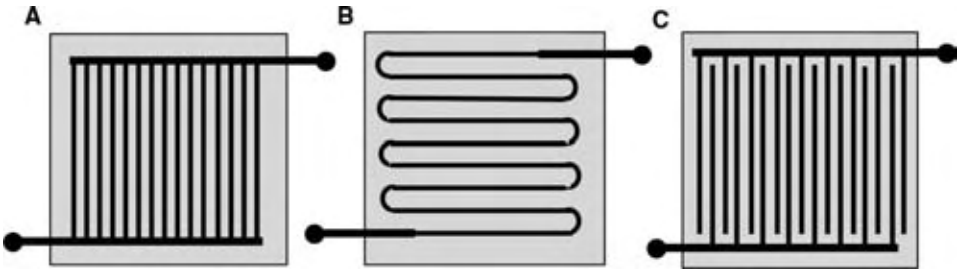


Fig. 15 Some topographies of bipolar plates: (A) parallel; (B) serpentine; and (C) interdigitated flow patterns.

Bipolar Plates

Two plates in a single cell provide a flow field for the feed stream and collect current in fuel cells. The plates should be made of lightweight, mechanically strong, chemically stable, gas impermeable, and electrically conducting materials; graphite (or metal) is commonly used.[114] The pattern of the flow field in a plate, as well as the width and depth of the channels, is very important for the efficiency of PEMFC.[115] A number of plate topologies have been developed including straight, serpentine, and interdigitated flow fields (Fig. 15). The design of the flow field also influences water supply to the membrane, water removal from the cathode, and distributions of current across the cell. In fact, several techniques have been developed to measure the distribution of current density in flow fields of an operating cell.[116,117]

Sealing Materials

Sealing in fuel cell is required to separate gases entering into the cell. The materials for sealing should withstand the acidic environment of electrolyte and be durable for long-term operation. Silicone rubber based materials are widely employed because of their elasticity and excellent heat resistance. Silicone based materials, however, are degraded during the operation due to the acidic environment and thermal stress in the cell. The decomposition product can affect the hydrophobic property of the electrode layer and transport of reacting species, and thus the performance of PEMFC.[118]

Fabrication of MEA

The efficiencies of electrochemical kinetics and transport of reaction species are determined by the effective utilization of the electrodes, membrane, and reactant gases in PEMFC systems.[119,120] The effective design and fabrication of the MEA is the heart of PEMFC technology. The porous gas diffusion layer allows the flow of reactant gases, and also acts as the electron collector while allowing liquid water transport. The electrode is often placed on one side of the GDL by a painting, spraying, or screen-printing technique. The MEA should be designed to maximize the electrode utilization through enhanced transport of protons, electrons, and feed gases and products. Nafion solution is usually added to the electrode to improve accessibility of protons and thus decreases the inactive sites for the electrochemical reaction. However, too much Nafion over electrodes decreases the porosity of the electrode and inhibits transport of reaction species and can also cause electronic insulation. Fig. 16 shows a focus ion beam (FIB) image of the electrode layer on GDL, which provides the surface and sub-surface

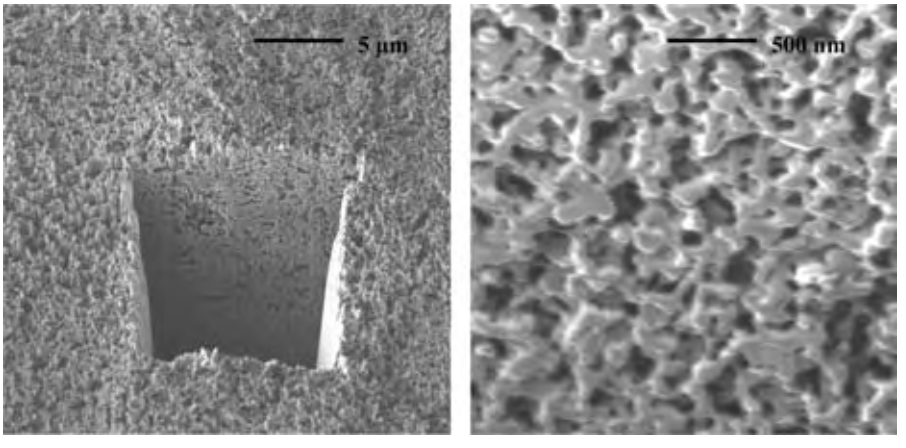


Fig. 16 A surface morphology of an electrode layer (~30 μm) and pore distribution of subsurface trenched (~10 μm) taken at the author's laboratory.

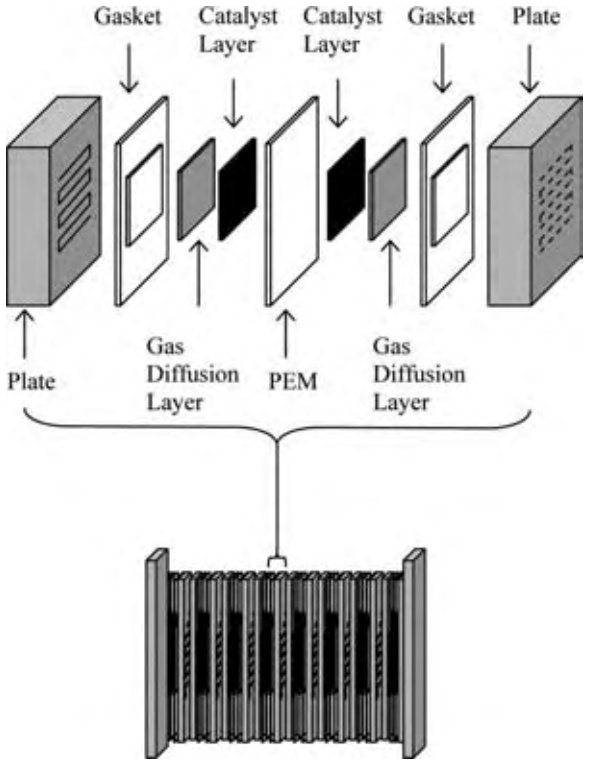


Fig. 17 A schematic diagram of a single fuel cell and a cell stack.

morphology, i.e., porosity distribution, agglomeration of any particular phase, and nanostructure. Design of nanostructure is necessary to optimize a number of properties, such as diffusivity of reactant and products, proton conductivity, electrical conductivity, and hydrophobicity. The preparation of MEA is usually completed by hot pressing the membrane at slightly above its glass transition temperature with the two electrodes in which two anode and cathode layers contact with the membrane. Fig. 17 shows the components of a single cell and a stack. The connection of single cells in a stack can be in series or parallel depending on the voltage and current requirements for specific applications. Large-scale production of MEA for PEMFC should consider a variety of material selection criteria and material processing aspects along with MEA fabrication technologies, which are optimized for the transport and electrochemical reactions.

FUELS FOR PEMFC

Fuels

Hydrogen and methanol are the common choice of fuels for PEMFC. Hydrogen is produced industrially by processing conventional hydrocarbon fuels. This process has been adopted for small-scale generation of hydrogen for use in fuel cells. Fig. 18 shows a general scheme of a hydrogen generation process for PEMFC applications. It consists of fuel reforming, water gas shift reaction, and deep cleaning of CO.[121–123] Steam reforming of methane is generally the most economic way to produce hydrogen industrially. Catalytic hydrodesulfurization and sulfur adsorption should precede the reforming reactions in the conventional plants for hydrogen generation from natural gas. Methanol, which is produced from syngas obtained by steam reforming of methane industrially,[124] can be used directly in DMFC or reformed to hydrogen for PEMFC. The choice of fuel in PEMFC is based on a number of factors, such as safety of fuel, fuel economy, cost for fuel storage, performance, infrastructure, lifecycle cost of transportation, emission, etc.[125]

Fuel Reforming

Fuel reforming can be performed by steam reforming,[126–128] partial oxidation,[128–130] and autothermal reforming.[131–133] The steam reforming reactions for hydrocarbons and alcohols are given below:

$$CH_4 + H_2O(g) = CO + 3H_2 \quad \Delta H = 205.9\,\text{kJ/mol} \tag{110}$$

$$C_mH_n + mH_2O = mCO + \left(m + \frac{n}{2}\right)H_2 \tag{111}$$

$$CH_3OH(g) + H_2O(g) = CO_2 + 3H_2 \quad \Delta H = 49.3\,\text{kJ/mol} \tag{112}$$

$$C_nH_{2n+1}OH + (2n - 1)H_2O = nCO_2 + 3nH_2 \tag{113}$$

Steam reforming reactions are highly endothermic ($\Delta H > 0$), and thus, heat needs to be supplied by the appropriate design of a reactor and a heat exchanger. Industrial methane steam reforming operates at high temperatures (~800°C) over Ni-based catalysts. However, the Ni-based catalysts have several drawbacks, including their high exothermicity for the reaction with air, high tendency for coke formation, and extremely low tolerance of sulfur. Steam reforming of methanol [Eq. (112)] operates at relatively low temperatures (200–400°C) with low steam-to-carbon ratios to produce a reformate with a high H_2 concentration. Cu-based catalysts are highly active for methanol steam reforming. However, the deactivation of Cu-based catalysts, when exposed to liquid water during shutdown, is a concern for PEMFC applications.

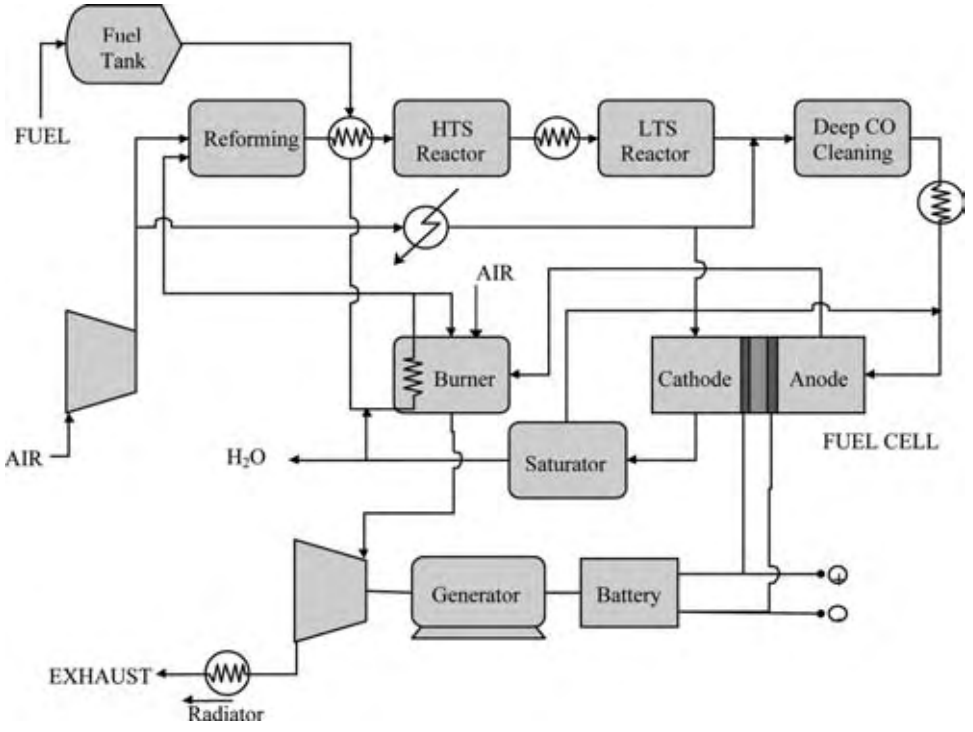


Fig. 18 A general scheme of hydrogen production for PEMFC application. *(View this art in color at www.dekker.com.)*

Partial oxidations of hydrocarbons and alcohols are given by the following:

$$CH_4 + \frac{1}{2}O_2 = CO + 2H_2 \quad \Delta H = -35.9\,\text{kJ/mol} \tag{114}$$

$$C_mH_n + \frac{m}{2}O_2 = mCO + \frac{n}{2}H_2 \tag{115}$$

$$CH_3OH(g) + \frac{1}{2}O_2 = CO_2 + 2H_2 \quad \Delta H = -192.5\,\text{kJ/mol} \tag{116}$$

$$C_nH_{2n+1}OH + \left(n - \frac{1}{2}\right)O_2 = nCO_2 + (n+1)H_2 \tag{117}$$

These partial oxidation reactions are exothermic and, thus, reformers are expected to be energy efficient and compact compared to the steam reforming. The partial oxidation of methane can be carried out by Ni, Co, Rh, and Pt group metals for the temperature range of 700–1000°C, while that of methanol has been studied over Cu-based catalysts in the temperature range of 200–300°C. Autothermal reforming of hydrocarbons and alcohols may be represented by the following:

$$CH_4 + aH_2O + bO_2 = cCO + dCO_2 + eH_2 \quad \Delta H = \sim 0\,\text{kJ/mol} \tag{118}$$

$$C_mH_n + \frac{m}{2}H_2O + \frac{m}{4}O_2 = mCO + \frac{1}{2}(m+n)H_2 \tag{119}$$

$$CH_3OH + a'H_2O + b'O_2 = c'CO + d'CO_2 + e'H_2 \quad \Delta H = \sim 0\,\text{kJ/mol} \tag{120}$$

$$C_nH_{2n+1}OH + \left(n - \frac{1}{2}\right)H_2O + \left(\frac{n}{2} - \frac{1}{4}\right)O_2 = nCO_2 + \left(2n + \frac{1}{2}\right)H_2 \tag{121}$$

Partial oxidation provides quick start-up and compactness, while steam reforming produces relatively high concentration of hydrogen in the product gas. The steam reforming is endothermic and the partial oxidation is exothermic so that the combination of these two reactions in appropriate proportion allows close to thermal neutrality, or adiabatic conditions at the desired temperature. HotSpot™ reactor systems have been developed for the autothermal reforming of methanol.[133]

Water Gas Shift Reaction (CO Cleanup)

The products from the reformer usually contain up to 12% CO depending upon the choice of fuels, catalysts, and operating temperatures. The concentration of CO is reduced further by the water gas shift reaction,[134–136] which is industrially performed in two stages in different temperature regions. First one is the high temperature (320–450°C) shift (HTS) reactor that reduces CO to the equilibrium value for a given set of reaction conditions over Fe_2O_3–Cr_2O_3 catalysts, and the second is low temperature (~200°C) shift (LTS) reactor that provides further reduction of CO over Cu/ZnO catalysts

$$\begin{aligned} CO + H_2O\,(g) &= H_2 + CO_2 \\ \Delta H &= -41.2\,\text{kJ/mol} \end{aligned} \tag{122}$$

The concentration of CO leaving the low temperature shift reactor can be reduced to the range of 0.1–1%, depending on operating conditions. However, it is still too high to be used directly for PEMFC applications.

Deep Cleaning of CO

For the long-term durability of PEMFC, the acceptable CO concentration appears to be 10–100 ppm. To meet the requirement, three possible reactions can be considered: preferential (or selective) oxidation,[137] methanation,[138] and Pd (or Pd alloy) membrane processes.[139–141] Preferential oxidation (PrOx) of CO can convert CO to CO_2, without excessive hydrogen oxidation (to water), to acceptable levels of CO using multi-stage reactors

$$\begin{aligned} CO + \frac{1}{2}O_2 &= CO_2 \\ \Delta H &= -283.0\,\text{kJ/mol} \end{aligned} \tag{123}$$

$$\begin{aligned} H_2 + \frac{1}{2}O_2 &= H_2O\,(g) \\ \Delta H &= -241.8\,\text{kJ/mol (undesirable)} \end{aligned} \tag{124}$$

Methanation can also be used for deep CO cleaning process

$$\begin{aligned} CO + 3H_2 &= CH_4 + H_2O\,(g) \\ \Delta H &= -205.9\,\text{kJ/mol} \end{aligned} \tag{125}$$

However, it requires three molecules of hydrogen for the removal of a CO molecule. Furthermore, the control of the reaction is difficult because of the presence of CO_2 created by the following undesirable reactions

$$\begin{aligned} CO + H_2O\,(g) &= CO_2 + H_2 \\ \Delta H &= -41.2\,\text{kJ/mol (undesirable)} \end{aligned} \tag{126}$$

$$\begin{aligned} CH_4 + 2H_2O\,(g) &= CO_2 + 4H_2 \\ \Delta H &= 164.7\,\text{kJ/mol (undesirable)} \end{aligned} \tag{127}$$

Thus, the methanation reaction appears to be impractical in conjunction with PEMFC, especially for on-board use in vehicles.

Palladium (or Pd alloy) membranes can be effectively used for CO removal or hydrogen separation because hydrogen diffuses through dense palladium membranes extremely selectively. Thin layers of Pd can be prepared over the steam reforming or water gas shift reactor so that only hydrogen can diffuse through the membrane. In principle, this combined system can represent the integration of reforming, water gas shift, and deep cleanup of CO in a compact way suitable for transportation applications. However, the need for a high pressure difference, relatively high temperature (300–400°C), and high cost are drawbacks of this process.

Hydrogen Storage

Hydrogen can be stored in the form of compressed gas[142] or as a liquid.[143] It is also stored in solids using adsorbents[144–146] and metal hydrides.[147,148] Hydrogen storage as a compressed gas in a high-pressure cylinder (200–300 bar) can be widely used because of its simplicity and low cost for storage. However, it requires a relatively large vessel volume compared to the other techniques because of the low density of hydrogen in its gas state. Hydrogen can also be stored as a liquid at the boiling point of hydrogen (−252.67°C) in highly insulated vacuum containers. This technique has higher volumetric hydrogen density. However, a large amount of energy is needed to liquefy hydrogen gas and a special vessel is also required to reduce evaporation of liquid hydrogen over a long period of time. Hydrogen can be stored in solid adsorbents, such as activated carbon[144] and nanostructured materials.[145,146] Some metals can absorb hydrogen under moderate pressures, forming reversible hydrogen compounds called hydrides, e.g., $NaBH_4$, $LiBH_4$, and $LaNi_5H_6$.[147,148] As hydrogen is stored in its atomic form in the metal alloy state, storing hydrogen in metal hydrides is volumetrically effective, but the weight of metal alloy along with vessel for the alloy is quite high compared to other techniques. The challenging issues in complex metal hydrides are the dense packing hydrogen atoms and scientific understanding

of the mechanisms and kinetics of the absorption and desorption.

FUTURE CHALLENGES OF PEMFC

Although there have been significant technological advancements in PEMFC, there are still some fundamental issues that need to be addressed before its anticipated commercialization. Some significant challenges in obtaining the performance enhancement of PEMFC include the development of: i) anodes tolerant to CO ($\sim$100 ppm) with noble metal loading lower than 0.1 mg/cm^2 MEA; ii) cathodes which reduce overpotentials with noble metal loading lower than 0.1mg /cm^2 MEA; iii) electrolytes which are relatively inexpensive, highly conductive, stable at high temperature (150–200°C), and/or impermeable to methanol; iv) electrolytes working without water so that the cumbersome hydration equipment is not required; v) efficient storage materials for hydrogen; vi) hydrogen generation processes optimized and reduced in size and complexity; and vii) low cost fuel cell stacks by increasing the effectiveness of all the components of fuel cells. Other concerns include the durability of PEMFC components and the ability to withstand temperature extremes common in North America, including well below freezing.

CONCLUSIONS

This article reviews fundamental knowledge and understanding of PEMFC. It is hoped that this review has provided useful information for PEMFC researchers and others who are interested in fuel cell systems. It is possible that power generation via fuel cells will become as common as that at present via heat engine that could provide substantial economic and environmental benefits.

REFERENCES

1. Schoenbein, C.F. On the voltaic polarization of certain solid and fluid substances. Phil. Mag. **1839**, *14*, 43–45.
2. Grove, W.R. On voltaic series and the combination of gases by platinum. Phil. Mag. **1839**, *14*, 127–130.
3. Grove, W.R. On a gas voltaic battery. Phil. Mag. **1842**, *21*, 417–420.
4. Grove, W.R. On the gas voltaic battery: Voltaic action of phosphorus, sulphur, and hydrocarbons. Proc. Roy. Soc. Lon. **1845**, *5*, 557–558.
5. Mond, L.; Langer, C. A new form of gas battery. Proc. Roy. Soc. Lond. **1989**, *46*, 296–304.
6. Appleby, A.J.; Foulkes, F.R. *Fuel Cell Handbook*; Van Nostrand Reinhold: New York, 1989.
7. Ketelaar, J.A.A. History. In *Fuel Cell Systems*; Blomen, L.J.M.J., Mugerwa, M.N., Eds.; Plenum Press: New York, 1993; 19–35.
8. Chen, E. History. In *Fuel Cell Technology Handbook*; Hoogers, G., Ed.; CRC Press: New York, 2003; 2-1–2-40.
9. Grubb, W.T. Batteries with solid ion-exchange membrane electrolytes. I. Secondary cells employing metal electrodes. J. Electrochem. Soc. **1959**, *106*, 275–278.
10. Grubb, W.T.; Niedrach, L.W. Batteries with solid ion-exchange membrane electrolytes. II. Low-temperature hydrogen-oxygen fuel cells. J. Electrochem. Soc. **1960**, *107*, 131–135.
11. Cairns, E.J.; Douglas, D.L.; Niedrach, L.S.W. Performance of fractional-watt ion-exchange membrane fuel cells. AIChE **1961**, *7*, 551–558.
12. Raistrick, I.D. Modified gas diffusion electrode for proton exchange membrane fuel cells. Proceedings of Symposium on Diapharms, Separators and Ion Exchange Membranes; White, R.E., Konishita, K., Van Zee, J.W., Burney, H.S., Eds.; Electrochemical Society: Pennington, NJ, 1986; PV 86-13, 172–178.
13. Walkins, D.S.; Dircks, K.W.; Epp, D.G. Novel Fuel Cell Fluid Flow Field Plate. U.S. Patent 4,988,583, January 29, 1991.
14. Carrette, L.; Friedrich, K.A.; Stimming, U. Fuel cells—fundamentals and applications. Fuel Cells **2001**, *1* (1), 5–39.
15. Costamagna, P.; Srinivasan, S. Quantum jumps in the PEMFC science and technology from the 1960s to the year 2000: Part I. Fundamental scientific aspects. J. Power Sources **2001**, *102* (1–2), 242–252.
16. Costamagna, P.; Srinivasan, S. Quantum jumps in the PEMFC science and technology from the 1960s to the year 2000: Part II. Engineering, technology development and application aspects. J. Power Sources **2001**, *102* (1–2), 253–269.
17. Wilkinson, D.P. Fuel cells. Interface **2001**, *10* (1), 22–25.
18. Stone, C.; Morrison, A.E. From curiosity to "power to change the world®". Solid State Ionics **2002**, *152–153*, 1–13.
19. Perry, M.L.; Fuller, T.F. A historical perspective of fuel cell technology in the 20th century. J. Electrochem. Soc. **2002**, *149* (7), S59–S67.
20. Acres, G.J.K. Recent advances in fuel cell technology and its applications. J. Power Sources **2001**, *10* (1–2), 60–66.

21. Haile, S.M. Fuel cell materials and components. Acta Materialia **2003**, *51* (19), 5981–6000.
22. Brandon, N.P.; Skinner, S.; Steele, B.C.H. Recent advances in materials for fuel cells. Annu. Rev. Mater. Res. **2003**, *33*, 183–213.
23. Thampan, T.; Malhotra, S.; Zhang, J.; Datta, R. PEM fuel cell as a membrane reactor. Catalysis Today **2001**, *67* (1–3), 15–32.
24. Atkins, P.W. *Physical Chemistry*, 4th Ed.; Oxford University Press: Oxford, UK, 1990.
25. Bockris, J.O'M.; Srinivasan, S. *Fuel Cells: Their Electrochemistry*; McGraw-Hill Book Company: New York, 1969.
26. Brett, C.M.A.; Brett, A.M.O. *Electrochemistry Principles, Methods, and Applications*; Oxford University Press: Oxford, 1993.
27. Bard, A.J.; Faulkner, L.R. *Electrochemical Methods*: *Fundamentals and Applications*, 2nd Ed.; John Wiley & Sons: New York, 2001.
28. Lide, D.R., Ed. *CRC Handbook of Chemistry and Physics*, 82nd Ed.; CRC Press LLC: New York, 2001, Sec. 5.
29. Conners, K.A. *Chemical Kinetics: The Study of Reaction Rates in Solution*; VCH: New York, 1990.
30. Markovic, N.M.; Ross, P.N. Surface science studies of model fuel cell electrocatalysts. Surf. Sci. Rep. **2002**, *45* (4–6), 117–229.
31. Conway, B.E.; Tilak, B.V. Interfacial processes involving electrocatalytic evolution and oxidation of H_2, and the role of chemisorbed H. Electrochim. Acta **2002**, *47* (22–23), 3571–3591.
32. Ralph, T.R.; Hogarth, M.P. Catalysis for low temperature fuel cells. Part II: The anode challenges. Platinum Metals Rev. **2002**, *46* (3), 117–135.
33. Vogel, W.; Lundquist, J.; Ross, P.; Stonehart, P. Reaction pathways and poisons—II: The rate controlling step for electrochemical oxidation of hydrogen on Pt in acid and poisoning of the reaction by CO. Electrochim. Acta **1975**, *20* (1), 79–93.
34. Gottesfeld, S.; Pafford, J. A new approach to the problem of carbon monoxide poisoning in fuel cells operating at low temperatures. J. Electrochem. Soc. **1988**, *135* (10), 2651–2652.
35. Wang, K.; Gasteiger, H.A.; Markovic, N.M.; Ross, P.N. On the reaction pathway for methanol and carbon monoxide electro-oxidation on Pt-Sn alloy versus Pt-Ru alloy surfaces. Electrochim. Acta **1996**, *41* (16), 2587–2593.
36. Gasteiger, H.A.; Markovic, N.M.; Ross, P.N. H_2 and CO electro-oxidation on well-characterized Pt, Ru, and Pt-Ru. 1. Rotating disk electrode studies of the pure gases including temperature effects. J. Phys. Chem. **1995**, *99*, 8290–8301.
37. Papageorgopoulos, D.C.; de Heer, M.P.; Keijzer, M.; Pieterse, J.A.Z.; de Bruijn, F.A. Nonalloyed carbon-supported Pt-Ru catalysts for PEMFC applications. J. Electrochem. Soc. **2004**, *151* (5), A763–A768.
38. Hajbolouri, F.; Andreaus, B.; Scherer, G.G.; Wokaun, A. CO tolerance of commercial Pt and Pt-Ru gas diffusion electrodes in polymer electrolyte fuel cells. Fuel Cells **2004**, *4* (3), 160–168.
39. Lee, S.J.; Mukerjee, S.; Ticianelli, E.A.; McBreen, J. Electrocatalysis of CO tolerance in hydrogen oxidation reaction in PEM fuel cells. Electrochim. Acta **1999**, *44* (19), 3283–3293.
40. Gasteiger, H.A.; Markovic, N.M.; Ross, P.N. Electro-oxidation of CO and H_2/CO mixtures on a well-characterized Pt_3Sn electrode surface. J. Phys. Chem. **1995**, *99* (22), 8945–8949.
41. Ross, P.N.; Kinoshita, K.; Scarpellino, A.J.; Stonehart, P. Electrocatalysis on binary alloys I. Oxidation of molecular hydrogen on supported Pt-Rh alloys. J. Electroanal. Chem. **1975**, *59* (2), 177–189.
42. Grgur, R.N.; Markovic, N.M.; Ross, P.N. Electro-oxidation of H_2, CO and H_2/CO mixtures on a well-characterized Pt-Re bulk alloy electrode and comparison with other Pt binary alloys. Electrochim. Acta **1999**, *43* (24), 3631–3635.
43. Hamnett, A. Mechanism and electrocatalysis in the direct methanol fuel cell. Catalysis Today **1997**, *38* (4), 445–457.
44. Gasteiger, H.A.; Markovic, N.; Ross, P.N.; Cairns, E.J. Temperature-dependent methanol electro-oxidation on well-characterized Pt-Ru alloys. J. Electrochem. Soc. **1994**, *141* (7), 1795–1803.
45. Sundmacher, K.; Schultz, T.; Zhou, S.; Scott, K.; Ginkel, M.; Gilles, E.D. Dynamics of the direct methanol fuel cell (DMFC): Experiments and model-based analysis. Chem. Eng. Sci. **2001**, *56* (2), 333–341.
46. Hogarth, M.P.; Ralph, T.R. Catalysis for low temperature fuel cells. Part III: Challenges for the direct methanol fuel cell. Platinum Metals Rev. **2002**, *46* (4), 146–164.
47. Dinh, H.N.; Ren, X.; Garzon, F.H.; Zelenay, P.; Gottesfeld, S. Electrocatalysis in direct methanol fuel cells: in-situ probing of Pt-Ru anode catalyst surfaces. J. Electroanal. Chem. **2000**, *491* (1–2), 222–233.
48. Lizcano-Valbuena, W.H.; Paganin, V.A.; Gonzalez, E.R. Methanol electro-oxidation on gas diffusion electrodes prepared with Pt-Ru/C catalysts. Electrochim. Acta **2002**, *47* (22–23), 3715–3722.

49. Haner, A.N.; Ross, R.N. Electrochemical oxidation of methanol on tin-modified platinum single-crystal surfaces. J. Phys. Chem. **1991**, *95* (9), 3740–3746.
50. Gotz, M.; Wendt, H. Binary and ternary anode catalyst formulations including the elements W, Sn and Mo for PEMFCs operated on methanol or reformate gas. Electrochim. Acta **1998**, *43* (24), 3637–3644.
51. Strasser, P.; Fan, Q.; Devenney, M.; Weinberg, W.H.; Liu, P.; Norskov, J.K. High throughput experimental and theoretical predictive screening of materials—a comparative study of search strategies for new fuel cell anode catalysts. J. Phys. Chem. B **2003**, *107* (40), 11,013–11,021.
52. Gurau, B.; Viswanathan, R.; Lafrenz, T.J.; Liu, R.; Ley, K.L.; Smotkin, E.S.; Reddington, E.; Sapienza, A.; Chan, B.C.; Mallouk, T.E.; Sarangapani, S. Structural and electrochemical characterization of binary, ternary, and quaternary platinum alloy catalysts for methanol electro-oxidation. J. Phys. Chem. B **1998**, *102* (49), 9997–10,003.
53. Wroblowa, H.S.; Pan, Y.C.; Razumney, G. Electroreduction of oxygen a new mechanistic criterion. J. Electroanal. Chem. **1976**, *69* (2), 195–201.
54. Sepa, D.B.; Vojnovic, V.; Damjanovic, A. Reaction intermediates as a controlling factor in the kinetics and mechanism of oxygen reduction at platinum electrodes. Electrochim. Acta **1981**, *26* (6), 781–793.
55. Konishita, K. *Electrochemical Oxygen Technology*; John Wiley & Sons: New York, 1992.
56. Lipkowski, J.; Ross, P.N. *Electrocatalysis*; Wiley/VCH: New York, 1998.
57. Ralph, T.R.; Hogarth, M.P. Catalysis for low temperature fuel cells. Part I: The cathode challenges. Platinum Metals Rev. **2002**, *46* (1), 3–14.
58. Wang, J.X.; Markovic, N.M.; Adzic, R.R. Kinetic analysis of oxygen reduction on Pt (111) in acid solutions: intrinsic kinetic parameters and anion adsorption effects. J. Phys. Chem. B **2004**, *108* (13), 4127–4133.
59. Markovic, N.M.; Schmidt, T.J.; Stamenkovic, V.; Ross, P.N. Oxygen reduction reaction on Pt and Pt bimetallic surfaces: a selective review. Fuel Cells **2001**, *1* (2), 105–116.
60. Anderson, A.B.; Albu, T.V. Catalytic effect of platinum on oxygen reduction. An ab initio model including electrode potential dependence. J. Electrochem. Soc. **2000**, *147* (11), 4229–4238.
61. Kinoshita, K. Particle size effects for oxygen reduction on highly dispersed platinum in acid electrolytes. J. Electrochem. Soc. **1990**, *137* (3), 845–848.
62. Watanabe, M.; Sei, H.; Stonehart, P. The influence of platinum crystallite size on the electroreduction of oxygen. J. Electroanal. Chem. **1989**, *261* (2), 375–387.
63. Mukerjee, S. Particle size and structural effects in platinum electrocatalysis. J. Appl. Electrochem. **1990**, *20* (4), 537–548.
64. Wilson, M.S.; Garzon, F.H.; Sickfacus, K.E.; Gottesfeld, S. Surface area loss of supported platinum in polymer electrolyte fuel cells. J. Electrochem. Soc. **1993**, *140* (10), 2872–2877.
65. Mukerjee, S.; Srinivasan, S. Enhanced electrocatalysis of oxygen reduction on platinum alloys in proton exchange membrane fuel cells. J. Electroanal. Chem. **1993**, *357* (1–2), 201–224.
66. Paulus, U.A.; Wokaun, A.; Scherer, G.G.; Schmidt, T.J.; Stamenkovic, V.; Radmilovic, V.; Markovic, N.M.; Ross, P.N. Oxygen reduction on carbon-supported Pt-Ni and Pt-Co alloy catalysts. J. Phys. Chem. B **2002**, *106* (16), 4181–4191.
67. Kinoshita, K.; Stonehart, P. Preparation and characterization of highly dispersed electrocatalytic materials. In *Modern Aspects of Electrochemistry*; Plenum Press: New York, 1977; Vol. 12, 183–266.
68. Hooger, G.; Thompsett, D. Catalysis in proton exchange membrane fuel cell technology. CATTECH **2000**, *3* (2), 106–124.
69. Lister, S.; Mclean, G. PEM fuel cell electrodes. J. Power Sources **2004**, *130* (1–2), 61–76.
70. Antolini, E. Review in applied electrochemistry number 54: recent development in polymer electrolyte fuel cell electrodes. J. Appl. Electrochem. **2004**, *34* (6), 563–576.
71. Matsumoto, T.; Komatsu, T.; Nakano, H.; Arai, K.; Nagashima, Y.; Yoo, E.; Yamazaki, T.; Kijima, M.; Shimizu, H.; Takasawa, Y.; Nakamura, J. Efficient usage of highly dispersed Pt on carbon nanotubes for electrode catalysts of polymer electrolyte fuel cells. Catalysis Today **2004**, *90* (3–4), 277–281.
72. Wilson, M.S.; Valerio, J.A.; Gottesfeld, S. Low platinum loading electrodes for polymer electrolyte fuel cells fabricated using thermoplastic ionomers. Electrochim. Acta **1995**, *40* (3), 355–363.
73. Kumar, G.S.; Raja, M.; Parthasarathy, S. High performance electrodes with very low platinum loading for polymer electrolyte fuel cells. Electrochim. Acta **1995**, *40* (3), 285–290.
74. Uchida, M.; Fukuoka, Y.; Sugawara, Y.; Ohara, H.; Ohta, A. Improved preparation process of very-low-platinum-loading electrodes for polymer electrolyte fuel cells. J. Electrochem. Soc. **1998**, *145* (11), 3708–3713.

75. Gasteiger, H.A.; Panels, J.E.; Yan, S.G. Dependence of PEM fuel cell performance on catalyst loading. J. Power Sources **2004**, *127* (1–2), 162–171.
76. Rikukawa, M.; Sanui, K. Proton conducting polymer electrolyte membranes based on hydrocarbon polymers. Prog. Polym. Sci. **2000**, *25* (10), 1463–1502.
77. Mehta, V.; Cooper, J.S. Review and analysis of fuel cell design and manufacturing. J. Power Sources **2003**, *114* (1), 32–53.
78. Roziere, J.; Jones, D.J. Non-fluorinated polymer materials for proton exchange membrane fuel cells. Annu. Rev. Mater. Res. **2003**, *33*, 503–555.
79. Alberti, G.; Casciola, M. Composite membranes for medium-temperature PEM fuel cells. Annu. Rev. Mater. Res. **2003**, *33*, 129–154.
80. Malhotra, S.; Datta, R. Membrane-supported nonvolatile acidic electrolytes allow higher temperature operation of proton-exchange membrane fuel cells. J. Electrochem. Soc. **1997**, *144* (2), L23–L26.
81. Savadogo, O. Emerging membranes for electrochemical systems: part II. High temperature composite membranes for polymer electrolyte fuel cell (PEFC) applications. J. Power Sources **2004**, *127* (1–2), 135–161.
82. Li, Q.; He, R.; Jensen, J.O.; Bjerrum, N.J. PBI-based polymer membranes for high temperature fuel cells – preparation, characterization and fuel cell demonstration. Fuel Cells **2004**, *4* (3), 147–159.
83. Yang, C.; Costamagna, P.; Srinivasan, S.; Benziger, J.; Bocarsly, A.B. Approaches and technical challenges to high temperature operation of proton exchange membrane fuel cells. J. Power Sources **2001**, *103* (1), 1–9.
84. Heitner-Wirguin, C. Recent advances in perfluorinated ionomer membranes: structure, properties and applications. J. Membr. Sci. **1996**, *120* (1), 1–33.
85. Mauritz, K.A.; Moore, R.B. State of understanding of Nafion. Chem. Rev. **2004**, *104* (10), 4535–4585.
86. Zawodzinski, T.A.; Springer, T.E.; Davey, J.; Jestel, R.; Lopez, C.; Valerio, J.; Gottesfeld, S. A comparative study of water uptake by and transport through ionomeric fuel cell membranes. J. Electrochem. Soc. **1993**, *140* (7), 1981–1985.
87. Hinatsu, J.T.; Mizuhata, M.; Takenaka, H. Water uptake of perfluorosulfonic acid membranes from liquid water and water vapor. J. Electrochem. Soc. **1994**, *141* (6), 1493–1498.
88. Choi, P.; Datta, R. Sorption in proton-exchange membranes: an explanation of Schroeder's paradox. J. Electrochem. Soc. **2003**, *150* (12), E601–E607.
89. Voss, H.H.; Wilkinson, D.P.; Pickup, P.G.; Johnson, M.C.; Basura, V. Anode water removal: a water management and diagonistic technique for solid polymer fuel cells. Electrochim. Acta **1995**, *40* (3), 321–328.
90. Janssen, G.J.M.; Overvelde, M.L.J. Water transport in the proton-exchange-membrane fuel cell: measurements of the effective drag coefficient. J. Power Sources **2001**, *101* (1), 117–125.
91. Shimpalee, S.; Greenway, S.; Spuckler, D.; Van Zee, J.W. Predicting water and current distributions in a commercial-size PEMFC. J. Power Sources **2004**, *135* (1–2), 79–87.
92. Zawodzinski, T.A.; Davey, J.; Valerio, J.; Gottesfeld, S. The water content dependence of electro-osmotic drag in proton-conducting polymer electrolytes. Electrochim. Acta **1995**, *40* (3), 297–302.
93. Okada, T. Theory for water management in membranes for polymer electrolyte fuel cells: part 1. The effect of impurity ions at the anode side on the membrane performances. J. Electroanal. Chem. **1999**, *465* (1), 1–17.
94. Okada, T. Theory for water management in membranes for polymer electrolyte fuel cells: part 2. The effect of impurity ions at the cathode side on the membrane performances. J. Electroanal. Chem. **1999**, *465* (1), 18–29.
95. Kulikovsky, A.A. The effect of cathodic water on performance of a polymer electrolyte fuel cell. Electrochim. Acta **2004**, *49* (28), 5187–5196.
96. Chiou, J.S.; Paul, D.R. Gas permeation in a dry Nafion membrane. Ind. Eng. Chem. Res. **1988**, *27* (11), 2161–2164.
97. Buchi, F.N.; Wakizoe, W.; Srinivasa, S. Microelectrode investigation of oxygen permeation in perfluorinated proton exchange membranes with different equivalent weights. J. Electrochem. Soc. **1996**, *143* (3), 927–932.
98. Gode, P.; Lindbergh, G.; Sundholm, G. In-situ measurements of gas permeability in fuel cell membranes using a cylindrical microelectrode. J. Electroanal. Chem. **2002**, *518* (2), 115–122.
99. Thampan, T.; Malhotra, S.; Tang, H.; Datta, R. Modeling of conductive transport in proton-exchange membranes for fuel cells. J. Electrochem. Soc. **2000**, *147* (9), 3242–3250.
100. Choi, P.; Jalani, N.H.; Datta, R. Thermodynamics and proton transport in Nafion II. Proton diffusion mechanisms and conductivity. J. Electrochem. Soc. **2005**, *152* (3), E123–E130.
101. Kreuer, K.D. Proton conductivity: materials and applications. Chem. Mater. **1996**, *8* (3), 610–641.

102. Kreuer, K.D.; Paddison, S.J.; Spohr, E.; Schuster, M. Transport in proton conductors for fuel-cell applications: simulation, elementary reactions, and phenomenology. Chem. Rev. **2004**, *104* (10), 4637–4678.
103. Paddison, S.J.; Paul, R.; Zawodzinski, T.A. A statistical mechanical model of proton and water transport in a proton exchange membrane. J. Electrochem. Soc. **2000**, *147* (2), 617–626.
104. Eikerling, M.; Kornyshev, A.A. Proton transfer in a single pore of a polymer electrolyte membrane. J. Electroanal. Chem. **2001**, *502* (1–2), 1–14.
105. Eikerling, M.; Kornyshev, A.A.; Kuznetsov, A.M.; Ulstrup, J.; Walbran, S. Mechanisms of proton conductance in polymer electrolyte membranes. J. Phys. Chem. B. **2001**, *105* (17), 3646–3662.
106. Jinnouchi, R.; Okazaki, K. Molecular dynamics study of transport phenomena in perfluorosulfonate ionomer membranes for polymer electrolyte fuel cells. J. Electrochem. Soc. **2003**, *150* (1), E66–E73.
107. Curtin, D.E.; Lousenberg, R.D.; Henry, T.J.; Tangeman, P.C.; Tisack, M.E. Advanced materials for improved PEMFC performance and life. J. Power Sources **2004**, *131* (1–2), 41–48.
108. Liu, W.; Ruth, K.; Rusch, G. Membrane durability in PEM fuel cells. J. New Mater. Electrochem. Syst. **2001**, *4*, 227–231.
109. Buchi, F.N.; Gupta, B.; Haas, O.; Scherer, G.G. Study of radiation-grafted FEP-G-polystyrene membranes as polymer electrolytes in fuel cells. Electrochim. Acta **1995**, *40* (3), 345–353.
110. Pianca, M.; Barchiesi, E.; Esposto, G.; Radice, S. End groups in fluoropolymers. J. Fluorine Chem. **1999**, *95* (1–2), 71–84.
111. Gode, P.; Ihonen, J.; Strandroth, A.; Ericson, H.; Lindbergh, G.; Paronen, M.; Sundholm, F.; Sundhohm, G.; Walsby, N. Membrane durability in a PEM fuel cell studied using PVDF based radiation grafted membranes. Fuel Cells **2003**, *3* (1–2), 21–27.
112. Ericson, H.; Kallio, T.; Lehtinen, T; Mattsson, B.; Sundholm, G.; Sundholm, F.; Jacobsson, P. Confocal raman spectroscopic investigations of fuel cell tested sulfonated styrene grafted poly (vinylidene fluoride) membranes. J. Electrochem. Soc. **2002**, *149* (2), A206–A211.
113. Williams, M.V.; Begg, E.; Bonville, L.; Kunz, H.R.; Fenton, J.M. Characterization of gas diffusion layers for PEMFC. J. Electrochem. Soc. **2004**, *151* (8), A1173–A1180.
114. Cooper, J.S. Design analysis of PEMFC bipolar plates considering stack manufacturing and environment impact. J. Power Sources **2004**, *129* (2), 152–169.
115. Arico, A.S.; Creti, P.; Baglio, V.; Modica, E.; Antonucci, V. Influence of flow field design on the performance of a direct methanol fuel cell. J. Power Sources **2000**, *91* (2), 202–209.
116. Mench, M.M.; Wang, C.Y. An in situ method for determination of current distribution in PEM fuel cells applied to a direct methanol fuel cell. J. Electrochem. Soc. **2003**, *150* (1), A79–A85.
117. Geiger, A.B.; Eckl, R.; Wokaun, A.; Scherer, G.G. An approach to measuring locally resolved currents in polymer electrolyte fuel cells. J. Electrochem. Soc. **2004**, *151* (3), A394–A398.
118. Schulze, M.; Knori, T.; Schneider, A.; Gulzow, E. Degradation of sealings for PEFC test cells during fuel cell operation. J. Power Sources **2004**, *127* (1–2), 222–229.
119. Weidner, J.W.; Sethuraman, V.A.; Van Zee, J.W. Engineering a membrane electrode assembly. Interface **2003**, *12* (4), 40–43.
120. Eikerling, M.; Ioselevich, A.S.; Kornshev, A.A. How good are the electrodes we use in PEFC? Fuel Cells **2004**, *4* (3), 131–140.
121. Trimm, D.L.; Onsan, Z.I. Onboard fuel conversion for hydrogen-fuel-cell-driven vehicles. Catal. Rev. Sci. Eng. **2001**, *43* (1–2), 31–84.
122. Song, C. Fuel processing for low-temperature and high-temperature fuel cells. Challenges, and opportunities for sustainable development in the 21st century. Catalysis Today **2002**, *77* (1–2), 17–49.
123. Farrauto, R.; Hwang, S.; Shore, L.; Ruettinger, W.; Lampert, J.; Giroux, T.; Liu, Y.; Ilinich, O. New material needs for hydrocarbon fuel processing: generating hydrogen for PEM fuel cell. Annu. Rev. Mater. Res. **2003**, *33*, 1–27.
124. Cheng, W.; Kung, H.H. *Methanol Production and Use*; Dekker: New York, 1994.
125. Adamson, K.A.; Pearson, P. Hydrogen and methanol: a comparison of safety, economics, efficiencies and emissions. J. Power Sources **2000**, *86* (1–2), 548–555.
126. Peters, R.; Dusterwald, H.G.; Hohlein, B. Investigation of a methanol reformer concept considering the particular impact of dynamics and long-term stability for use in a fuel-cell-powered passenger car. J. Power Sources **2000**, *86* (1–2), 507–514.
127. Lindstrom, B.; Pettersson, L.J. Hydrogen generation by steam reforming of methanol over copper-based catalysts for fuel cell applications. Int. J. Hydrogen Energy **2001**, *26* (9), 923–933.
128. Barz, D.P.; Tragner, U.K.; Schmidt, V.M.; Koschowitz, M. Thermodynamics of hydrogen generation from methane for domestic polymer electrolyte fuel cell systems. Fuel Cells **2004**, *3* (4), 199–207.

129. Freni, S.; Calogero, G.; Cavallaro, S. Hydrogen production from methane through catalytic partial oxidation reactions. J. Power Sources **2000**, *87* (1–2), 28–38.
130. de Smet, C.R.H.; de Croon, M.H.J.M.; Berger, R.J.; Marin, G.B.; Schouten, J.C. Design of adiabatic fixed-bed reactors for the partial oxidation of methane to synthesis gas. Application to production of methanol and hydrogen-for-fuel-cells. Chem. Eng. Sci. **2001**, *56* (16), 4849–4861.
131. Hagh, B.F. Optimization of autothermal reactor for maximum hydrogen production. Int. J. Hydrogen Energy **2003**, *28* (12), 1369–1377.
132. Lindstrom, B.; Pettersson, L.J. Development of a methanol-fueled reformer for fuel cell applications. J. Power Sources **2003**, *118* (1–2), 71–78.
133. Gary, P.G.; Petch, M.I. Advances with HotSpot^{TM} fuel processing. Platinum Metals Rev. **2000**, *44* (3), 108–111.
134. Newsome, D.S. The water-gas shift reaction. Catal. Rev. Sci. Eng. **1980**, *21* (2), 275–318.
135. Rhodes, C.; Hutchings, G.J.; Ward, A.M. Water-gas shift reaction: Finding the mechanistic boundary. Catalysis Today **1995**, *23* (1), 43–58.
136. Ruettinger, W.; Ilinich, O.; Farrauto, J. A new generation water gas shift catalysts for fuel cell applications. J. Power Sources **2003**, *118* (1–2), 61–65.
137. Ratnasamy, P.; Srinivas, D; Satyanarayana, C.V.V.; Manikandan, P.; Senthil Kumaran, R.S.; Sachin, M.; Shetti, V.N. Influence of the support on the preferential oxidation of CO in hydrogen-rich steam reformates over the CuO–CeO_2–ZrO_2 system. J. Catal. **2004**, *221* (2), 455–465.
138. Hooger, G. The fueling problem: fuel cell systems. In *Fuel Cell Technology Handbook*; Hoogers, G., Ed.; CRC Press: New York, 2003, 5-1–5-23.
139. Kikuchi, E. Membrane reactor application to hydrogen production. Catalysis Today **2000**, *56* (1–3), 97.
140. Wieland, I.S.; Melin, I.T.; Lamm, I.A. Membrane reactors for hydrogen production. Chem. Eng. Sci. **2002**, *57* (9), 1571–1576.
141. Lattner, J.R.; Harold, M.P. Comparison of conventional and membrane reactor fuel processors for hydrocarbon-based PEM fuel cell systems. Int. J. Hydrogen Energy **2004**, *29* (4), 393–417.
142. Tummala, M.; Krepec, T.; Ahmed, A.K.W. Optimization of thermocontrolled tank for hydrogen storage in vehicles. Int. J. Hydrogen Energy **1997**, *22* (5), 525–530.
143. Aceves, S.M.; Martinez-Frias, J.; Garcia-Villazana, O. Analytical and experimental evaluation of insulated pressure vessels for cryogenic hydrogen storage. Int. J. Hydrogen Energy **2000**, *25* (11), 1075–1085.
144. de la Casa-Lillo, M.A.; Lamari-Darkrim, F.; Cazorla-Amoros, D.; Linares-Solano, A. Hydrogen storage in activated carbons and activated carbon fibers. J. Phys. Chem. B **2002**, *106* (42), 10,930–10,934.
145. Lueking, A.; Yang, R.T. Hydrogen storage in carbon nanotubes; residual metal content and pretreatment temperature. AIChE **2004**, *49* (6), 1556–1568.
146. Seayad, A.M.; Antonelli, D.M. Recent advances in hydrogen storage in metal-containing inorganic nanostructures and related materials. Advanced Materials **2004**, *16* (9–10), 756–777.
147. Zaluska, A.; Zaluski, L.; Strom-Olsen, J.O. Structure, catalysis and atomic reactions on the nano-scale: A systematic approach to metal hydrides for hydrogen storage. Applied Physics A: Materials Science & Processing **2001**, *72* (2), 157–176.
148. Zuttel, A. Materials for hydrogen storage. Materials Today **2003**, *6* (9), 24–33.

Reactive Extrusion

Gerard T. Caneba
Department of Chemical Engineering, Michigan Technological University, Houghton, Michigan, U.S.A.

INTRODUCTION

In the last quarter of the 20th century, there has been a tremendous spurt in the manufacture and application of polymers. The rapid growth of the polymer industry since the 1970s has led to the development of new methods of polymer synthesis and processing. An operation where simultaneous polymer synthesis and processing occurs is called reactive processing. Reactive extrusion (REX) and reaction injection molding (RIM) are particular examples of reactive processing. In REX, polymerization and other reactions and formulations associated with polymers are carried out inside the extruder, while processing is underway.

In the 1960s, practical processes for several types of chemical reactions run in extruder reactors were developed in a number of industrial laboratories. Landmark examples of reactive processing include controlled rheology of polyolefins by controlled molecular weight degradation, halogenation of polyolefins, and grafting of acrylic acid and vinylsilanes. Various commercial products are being manufactured through reactive processing and more are being produced every year. Compounding is being done more effectively through reactive processing because of the added dimension of chemical reaction in using a processing type of equipment. Such an approach has resulted in better performing products and/or lower cost.

CHEMICAL REACTIONS WITHIN REX SYSTEMS

REX, also called reactive compounding, refers to performance of chemical reactions during the extrusion processing of polymers.[1] In this case, an extrusion device is used as a chemical reactor instead of being used only as a processing equipment. Fig. 1 shows a schematic of a typical REX operation.

An early published example of reactive processing was a description of bulk polymerization of caprolactam in an extruder to give nylon-6.[2] Intense activity in recent years, mostly in industrial laboratories and at extruder companies, has produced more than 600 patents and 60 published papers on the subject of reactive processing or REX.[3] Various types of polymer reactions that have been implemented through REX include:[4]

1. Bulk polymerizations, such as addition (free-radical- and ionic-based) and step-growth types.
2. Grafting polymerization by small molecules.
3. Interchain copolymer formation, based on chain cleavage, graft copolymerization, and end-group block copolymerization.
4. Coupling and branching.
5. Controlled degradation.
6. Polymer functionalization and functional group modification.

Bulk Polymerization

In this type of polymerization, an undiluted monomer or mixture of monomers is converted to a high molecular weight homopolymer or copolymer. The progress of the reaction is characterized by a large increase in the viscosity of the reacting mixture (from less than 50 Pa-s to greater than 1,000 Pa-s). Two reaction mechanisms of bulk polymerization have been implemented in REX: free-radical addition and condensation reaction chemistries.

Addition, also known as chain polymerization, occurs when the monomer units are added singly to a growing chain. The reaction takes place in three steps. The first is initiation, wherein a catalyst or initiator molecule (I) is broken down to give free radicals ($R^{\bullet}$ or $RO^{\bullet}$). A free radical activates a single monomer unit to produce the primary radical ($R'\text{-}M^{\bullet}$). The next step is propagation, wherein new monomer units are singly added to the radical ends. As the propagation reaction proceeds, these polymer radicals grow in length, and the molecular weight of the polymer increases. The third and final step is termination, in which two polymer radicals either combine or undergo a charge transfer to form dead polymer molecules. The former type of termination is called coupling or recombination, while the latter case is called disproportionation. An outline of the basic free-radical addition polymerization is shown below.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120027022

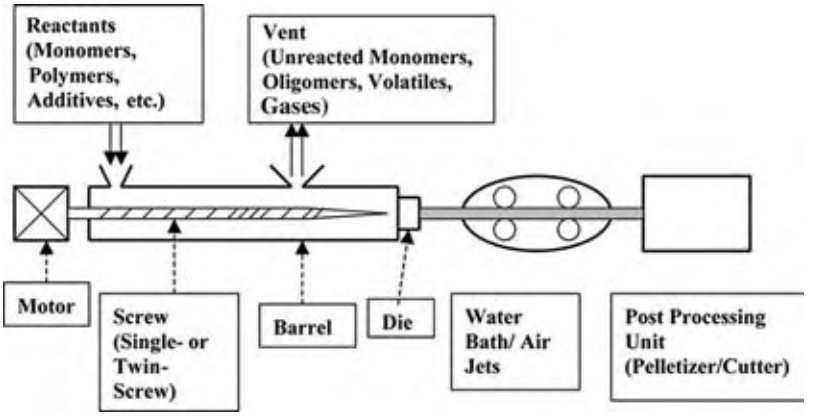


Fig. 1 Schematic of a reactive extrusion (REX) system.

Initiation:

$$I \Longrightarrow 2R^{\bullet} \text{ or } 2RO^{\bullet} + \text{Gas}$$

$$R^{\bullet} \text{ or } RO^{\bullet} + M \Longrightarrow R' - M^{\bullet}$$

Propagation:

$$R' - M^{\bullet} \Longrightarrow R' - M - M^{\bullet}$$

$$\vdots$$

$$R' - M_n^{\bullet} + M \Longrightarrow R' - M_{n+1}^{\bullet}$$

Termination:

$$R' - M_x^{\bullet} + R' - M_y^{\bullet} \Longrightarrow R' - M_{x+y} - R'$$

(Coupling or Recombination)

$$\Longrightarrow R' - M_{x+1} + R' - M_y - 1$$

(Disproportionation)

Chain Transfer:

$$R' - M_x^{\bullet} + C_T \Longrightarrow R' - M_x + C_T$$

The species C_T is the chain transfer agent, which can be monomer, solvent, initiator, and/or polymer molecules. The new radical site $C_T\bullet$ can in turn undergo further propagation, termination, and chain transfer reactions.

Bulk addition multipolymerization kinetics occurs when two monomers are employed. Bulk free-radical homopolymerizations and copolymerizations that are implemented in REX include:[4] a) styrene-acrylonitrile, styrene-methyl methacrylate, styrene-acrylamide; b) methyl methacrylate-acrylonitrile, ABS; c) acrylate ester mixtures; d) ethyl acrylate-methacrylic acid and mixtures with other monomers; e) methyl methacrylate; f) ε-caprolactone;[5,6] and, *n*-isopropylacrylamide-acrylic acid-alkyl acrylates.[7–11] To implement REX for bulk free-radical polymerizations, the highest possible operating temperature is used for maximum reaction rate. Both single (1.2 m diameter, 15.2 m long, at production rate in the order of 900 kg/h) and twin screws have been used, although better heat removal was obtained using twin intermeshing self-wiping screws. Sometimes, a low-viscosity prepolymer is produced using a CSTR leading to a REX reactor.

Step-growth polymerization occurs when two individual molecules react to form the monomer. Two monomers react to form a dimer. These dimers react to form tetramers, etc. If at each step a byproduct small molecule is excluded, such as water or an alcohol, then we have the so-called condensation polymerization. A scheme describing a polyesterification reaction involving the manufacture of poly ethylene terephthalate (PET), which is used to produce plastic beverage bottles, is shown below.

$$n\text{HOOC-}(C_6H_6)\text{-COOH} + n\text{HO-}(CH_2)_2\text{-OH} \Longrightarrow \text{^^^[-(CO)-}(C_6H_6)\text{-(CO)-O-}(CH_2)_2\text{-O-]}_n\text{^^^} + nH_2O$$

Terephphthalic Acid — Ethylene Glycol — PET — Water

Polyesters, polyetherimide, melamine-formaldehyde, polyurethanes, polyurethane-ureas and polyamides (nylons)[12,13] are examples of condensation polymers prepared by REX.[4] Because a small molecule is produced in condensation reactions, vent ports are employed. Between the ports are melt sealing screw sections, to prevent back mixing of volatilizing melt. Sealing screw sections are constructed by a right handed–left handed sequence of screw elements. Another chemistry feature of step-growth reactions is their sensitivity to errors in stoichiometric feed proportions. This poses problems with solid feed materials, which are normally converted to liquid form for more

accurate metering. In a representative process for the production of a polyurethane from 9 parts butane-1,4-diol chain extender, and 91 parts low molecular weight polyester (OH number = 51.7, prepared from adipic acid and butane-1,4-diol), and 35 parts 4,4′-diisocyanatodiphenylmethane, the reactants are fed as liquids into the 53-mm-diameter 1855-mm-long twin-screw extruder with corotating self-wiping screws. The extruder entry zone is at 90–120°C, the middle section at 180–260°C, and the last section at 100–180°C. The extruder has two to three kneading zones that are 240-mm long, to prevent the formation of gel inhomogeneities at 0.8–2.5 min residence times with throughputs of 10–100 kg/h and screw speeds of 70–130 rpm.

Polyureas have been produced in an extruder from where it comes out as a shaped product inside a mold.[14] This special type of REX is called RIM. Polyurea is formed (a) directly from a complex or (b) from the reaction of the complex with a second amine:

$$\text{Amine} + \text{Isocyanate} \Longleftrightarrow \text{Complex} \longrightarrow \text{Polyurea(a)}$$

or

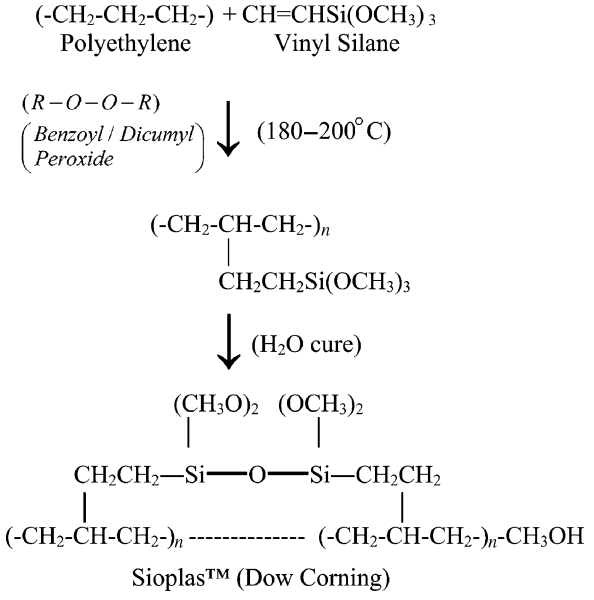
$$\text{Amine} + \text{Isocyanate} \Longleftrightarrow \text{Complex} + \text{Amine} \longrightarrow \text{Polyurea(b)}$$

Usually, methylene diisocyanate (MDI) is used with m-phenyl diamine (m-PDA) with hard (aromatic amine) or- soft (amine terminated polyether with MW of 500–5400 Daltons) chain extenders. With optimized formulations and extruder-mold conditions, cycle times of 3 to over 5 min can be obtained. With the use of internal mold release agents, the cycle time can be reduced to less than 2 min. Typical RIM operations for the production of automotive parts can run down to 1-min cycle time. These products range from the softer bumper covers to stiffer vertical body panels, such as fenders, doors, and quarter panels.

Bulk polymerizations are also implemented in REX systems in the presence of pre-existing polymers, to produce interpenetrating polymer networks (IPNs) (Refer discussion of IPNs in a later section).

Grafting Polymerization by Reactive Small Molecules

This involves the formation of graft copolymers from a reaction between polymers and monomers.[4] Monomer units can be propagated onto the polymer backbone to form a graft structure. Free radicals, air, or ionizing radiations are used to initiate the reaction. A schematic representation is shown below. A more specific case, the manufacture of vinyl silane (Sioplas™), is also been presented. Sioplas™ has been developed by Dow Corning and is used in wire coating and pipe insulation processes.



The REX process is normally carried out in two steps, with the first step occurring at 180–200°C. The mixture is passed through another REX reactor for mixing, and then through a final shaping REX reactor to complete the water curing reaction. The final cross-linked product (Sioplas™) is used as a wire or cable coating.

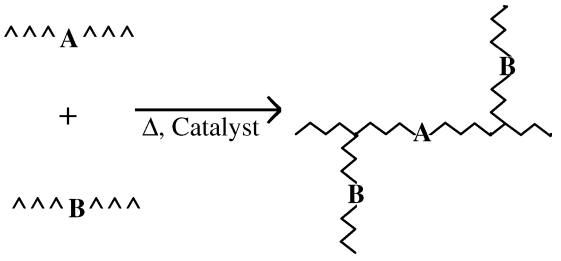
Grafting onto polyolefins (especially branched polyolefins) using free-radical sources is feasible because of the relative stability of these radical sites.[15] Moreover, propagation of certain monomers from radical sites is favored. Thus, it has been reported that acrylic acid, its derivatives and analogs (acrylic acid, methacrylic acid, glycidyl acrylate, butyl acrylate, methyl methacrylate, lauryl methacrylate, butyl methacrylate, polyethylene glycol methacrylates, 2-hydroxyethyl methacrylate, styrene, styrene-acrylonitrile, other styrene derivatives, and trialkoxysilane-containing monomers) have been grafted onto various polyolefins and copolymers (PE, PP, EP, EVA, ethylene-butyl acrylate, natural rubber, EPDM, PB, poly(ethylene oxide)) using REX operations.[4,16–19]

Interchain Copolymer Formation

This involves the reaction between two or more polymers to form a copolymer.[4] Interchain copolymer

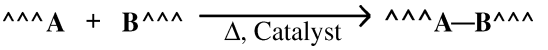
formation is further subdivided into the following types:

i. Graft copolymerization:



A popular example involves the grafting of maleic anhydride (MA) onto polypropylene (PP) using peroxide initiators. MA grafts become reactive sites along the PP chains resulting in the first polymer substrate. A second polymer substrate (Nylon 6) is intensely mixed with it to produce rubber-toughened nylon.[20] Other graft copolymers produced in this scheme involve ethylene-vinyl acetate-g-glycidyl methacrylate/styrene-maleic anhydride, and ethyl acrylate-vinylbenzyl chloride copolymer/Nylon 6,6 first substrate/second substrate combinations. Another system developed by Dow Chemical Co. involves a reactive polystyrene, RPS, as the first substrate, which is a polystyrene with approximately 1% reactive oxazoline groups distributed randomly along the chain. The oxazoline ring, which is reactive to both weak and strong nucleophiles, can react with the polymers that are terminated with amine, mercaptan, hydroxyl, epoxy, anhydride, and carboxylic acid groups.[4]

ii. Block copolymerization by end-group reaction:



For example, Nylon 6,6 can be reacted with PET in a reactive extruder to produce a block copolymer. However, as the concentration of A and B groups is relatively low, long residence times are required. Thus, it is more economical to form the block copolymer by other methods.[3]

Coupling/Branching

This type of reaction is one in which a homopolymer interacts with a polyfunctional agent to build molecular weight by chain extension or branching.[4] In Table 1, some examples of these reactions are tabulated.

For the Nylon 6,6 chain extension reaction,[21] a 5-cm single-screw extruder was used with a 5-min residence time. As the coupling agent is multifunctional, some branching can occur. Stoichiometric control is important to produce the required exact amount of branching. If no branching is required, a slight excess of the coupling agent has been used.

Controlled degradation

Controlled degradation is a reaction in which a reduction in polymer molecular weight is effected to meet the specific performance criteria.[4] One such application of controlled degradation is in the preparation

Table 1 Examples of coupling/branching reactions implemented with REX

Polymer(s)	Agent(s)	Temperature (°C)	Result
Nylon -6, -66[21]	Triphenyl phosphite (1% by weight)	250–300	Increase in molecular wt
PET[22]	Triphenyl phosphite (0.5–2.5% by weight)	265–285	A 40% increase in molecular wt
Linear PC[23]	1,1,1-tri(4-hydroxy-phenyl) ethane, 2,2′5,5′-tetra(4-hydroxyphenyl) hexane, trimellitic anhydride, trimellitic acid, trimellitoyl trichloride, 4-chloroformyl phthalic anhydride, pyromellitic acid, pyromellitic dianhydride, mellitic acid, mellitic anhydride, trimesic acid, benzophenonetetracarboxylic acid and benzophenonetetracarboxylic anhydride (0.1–10 % by weight)	100–400	Increased branching and crosslinking
Linear PC[24]	Triacrylate and peroxide	100–350	Increased branching

of lower-viscosity resins for injection molding.

$$\text{Polypropylene} \xrightarrow[Air,290^{\circ}C]{} \text{Decrease in molecular weight, viscosity, and elasticity (BASF, Exxon)}$$

$$\text{Polypropylene} \xrightarrow[(220-240^{\circ}C,Dialkyl-Peroxide)]{} \text{Increase in molecular weight, viscosity, and elasticity.}$$

$$\text{Poly(ethylene terphthalate)} \xrightarrow[(265-275^{\circ}C,0.19\%\ Ethylene-Glycol)]{} \text{20\% decrease in viscosity (Eastman Kodak).}$$

Controlled degradation of PP (BASF, Exxon) was done using a single-screw extruder with an air and polymer feed. The Eastman Kodak PET degradation process involves the coextrusion of PET and ethylene glycol.

In a more recent work, REX was used to produce levulinic acid by controlled degradation of Starch.[25]

Polymer Functionalization and Functional Group Modification

This application area involves the introduction of functional groups onto polymer backbone, end group, side chain, or modification of existing functional groups.[4] Thus, it is an intermediate result of some of the above-mentioned REX reactions, such as grafting and coupling. Chlorinated PVC (CPVC) is an end product from this type of reaction, wherein chlorine gas is reacted with PVC or LDPE. Other polyolefins, such as butyl rubber, high density or linear low density polyethylene, EP rubber, EPDM, EVA, or polyisobutylene, have been used as feedstocks for the formation of various chlorinated products. Challenges of this type of operation include widely varying viscosities, corrosion, and removal of hydrogen and hydrogen chloride gases. The reaction is conducted in a 5-cm extruder with nonintermeshing counter-rotating twin screw made from Hastelloy. Gas removal zones are isolated by melt seals from reverse-flight screw elements. Chlorine incorporations of 0.38–3.64% have been realized from these other polyolefin materials at 90–300 rpm, maximum temperatures of 140–220°C, and yields of 25–83 kg/h.

Other examples of functionalization/functional group modifications systems implemented in REX include:

1. Saponification of EVA with sodium methoxide in methanol using a twin-screw extruder.
2. Capping of carboxylic acid end groups, such as PET, with 1 wt% phenyl glycidyl ether at 255°C by coextrusion in a twin-screw extruder.
3. Hydrolysis of acyl fluoride end groups on hexafluoropropylene-tetrafluoroethylene copolymer with 1% water in a corrosion-resistant 28-mm intermeshing twin-screw extruder at 360°C.
4. Cyclicization of pendant carboxylic acid groups, such as the ammonia treatment of poly(methyl methacrylate) at 310°C to give ring closure of adjacent carboxylic acid unite to produce 2,6-piperidinedione (glutarimide) units in the chain.
5. Introduction of hydroperoxide groups into polyethylene by coextrusion with air under pressure at 190°C in a 20-mm extruder with 3-min residence time.
6. Grafting of muconic acid onto polyolefins.[26]
7. Conversion of polyketones to polyesters with organic acid.[27]
8. Conversion of poly(acrylic acid) to poly(*n*-isopropylacrylamide).[7–11]

Reactive compounding using REX[28,29] could involve the above-mentioned reactions. In a number of instances, the reactions involved are not well defined and/or is a complicated combination of the above-mentioned chemistries.[30–34] Most recently, REX has been used to make nanosilicates (nanoclay) compatible,[35] which signals a new emerging area of nanocomposite (nanometer-scale particles in polymers) manufacture using REX methods.

ADVANTAGES AND DISADVANTAGES OF REX

REX has certain advantages over other types of reactors, although it has its own set of limitations.[36,37] It involves higher surface-to-volume ratio compared to a stirred-tank reactor, resulting in better mixing, homogenization, and even temperature control. Consequently, there is better distribution of reactants, resulting in faster conversion rates. Also, REX involves easier venting of volatiles and unreacted monomer, and less chance of stagnation and thermal degradation. Because of the capability to impart relatively high mixing power per fluid volume, REX has the ability to handle highly viscous fluids, such as high polymers. Sections of screw elements can be constructed in series in REX operations, which results in operational flexibility and versatility, cleanliness, and ability to create specific reaction zones. If the operation is possible at all, the normal result is a lower process investment cost.

A disadvantage of REX operation is its difficulty with low-viscosity fluids (<10 Pa-s) and temperature control especially for highly exothermic reactions. Usually, a critical design criterion is that REX is not feasible beyond the residence times greater than

1000 s. Finally, there is no such thing as a multipurpose REX system, as their designs are specific to the chemistry and material properties of materials being handled.

Equipment

Reactive processing is limited to polymerization or chemical reactions of polymers in conventional single-screw or twin-screw extruders, excluding processes in oscillatory kneaders, Banbury-type continuous mixers, or Diskpack equipment. Emphasis is placed on continuous processes that have been implemented commercially or that can serve as models for commercial purposes.

A reactive extruder may be considered to be a horizontal reactor with one or two internal screws for conveying reactant polymer or monomer in the form of a solid or slurry, melt, or liquid. The most common reactants are polymer or prepolymer melts and gaseous, liquid, or molten low molecular weight compounds.

A particular advantage of the extruder as a chemical reactor in this context is the absence of a solvent as the reaction medium. No solvent stripping or recovery process is required, and the product contamination by solvent impurities is avoided. However it should be noted that minor amounts of small molecule species can be used in REX operations. In this case, REX also allows stripping of these components in isolated sections of the extruder.

Because of their versatility, most extruder reactors are twin-screw extruders that possess a segmented barrel, each segment of which can be individually heated or cooled externally. In addition to external heating, a molten material may be shear heated by the resistance of the viscous material to the conveying motion of the screw; these processes provide energy for the chemical reaction. Extruder screws often have specialized sections or configurations, e.g., high shear mixing sections. Twin-screw extruder screws may be equipped with interchangeable screw elements that provide different degrees of mixing and surface area exposure by varying the depth between screw flights, the individual flight thickness, and the direction and degree of flight pitch. Kneading blocks may be included as screw elements to provide intensive mixing. By varying the external heating, the screw element configuration, the clearance between screw and barrel wall in individual barrel segments, the total energy and the degree of mixing of material in each barrel segment may be varied. In this way, an extruder may be transformed into a chemical reactor with controlled reaction zones made up of individual barrel segments. In each of these segments sequential chemical processes can occur.

PROCESS PROCEDURE AND CONDITIONS

In a typical REX process, the reactants are fed into the extruder feed throat, where the material is usually heated to initiate reaction or increase the reaction rate. The reactant mixture is conveyed through sequential barrel segments, where the degree of mixing and specific energy input bring the reaction to the desired degree of completion, within the limits of residence time in the extruder. At this stage, the reaction may be quenched by cooling or addition of a catalyst quencher where applicable, and volatile by-products or excess reactants may be removed. Molten polymer is forced from the extruder through a die with one or more openings. The geometry of the die openings is one factor determining the pressure against which the extruder has to pump by the conveying motion of the internal screw. Polymer melt issuing from the die is usually rapidly cooled by contact with a fluid medium such as water. Cooling and solidification quench any chemical reaction that still occurs but may not have been particularly quenched in the extruder.

An advantage of an extrusion device as a reactor is the combination of several chemical process operations into one piece of equipment with accompanying high space-time yields of product. An extruder reactor is ideally suited for continuous production of material after equilibrium is established in the extruder barrel for the desired chemical processes.

For extruder reactors typical operating conditions are 70–500°C. Although this entire range of temperatures may extend over the length of an extruder, the temperature differential between adjacent barrel segments is often greater than 100°C because of slow heat transfer to and from reactant material. Typical extruder residence times are 10–600s. Residence time and, hence, the time available for chemical reaction is determined by extruder length, rate of introduction of reactants, and screw speed. Often greater versatility may be achieved by running a chemical reaction in two or more extruders connected in series.

By providing individual barrel segments with external openings, it is possible to introduce solid, liquid, or gaseous reactants at specified points in the chemical process. Their residence time in the reactor is controlled by the distance between the injection point and the die. For example, a heat-sensitive reactant may be introduced at a barrel segment near the die to minimize residence time. Similarly, an inert gas may be introduced at the extruder feed throat to protect a process from atmospheric oxygen. Volatile by-products from chemical reactions or excess reactants may be easily removed by applying a vacuum to the appropriate barrel segment and providing the appropriate screw segment with proper flight depth and geometry to ensure efficient surface renewal and exposure of the reaction

mixture to low pressure. The individual reaction zones at different pressures may be segregated from one and another by melt seals formed by restricting the flow of material both before and after the zone. Melt seals are commonly formed by using reverse flight and/or shallow screw flight segments that prevent back mixing of material in adjacent zones. Typical pressures encountered in REX processes are 0–50 MPa (7250 psi).

With the highly efficient mixing provided by an extruder, it is possible for two materials with great differences in viscosity to react. In the extreme case, molten polyolefins have been chlorinated using chlorine gas in a commercial extrusion process developed at Exxon.[20] A key to the success of this process is the design of the screw element geometry to give highly efficient surface renewal of polyolefin in contact with chlorine gas in a pressurized reaction zone segregated from the remainder of the extruder by formation of melt seals.

INTERPENETRATING POLYMER NETWORK SYSTEMS

A very convenient use of reactive processing equipment systems is the formation of IPNs. Traditionally, IPNs are synthesized by swelling a crosslinked polymer (Polymer I) with a second monomer together with crosslinking and activating agents, and then polymerizing the monomer in situ[38,39] to form Polymer II. An IPN can be distinguished from simple polymer blends, blocks, and grafts in two ways: 1) An IPN swells, but does not dissolve in solvents and 2) creep and flow are suppressed. Due to the wide range of possibilities covered by the working definition of IPN and its extensions, it is obvious that polymer scientists and engineers have barely scratched the surface of what is rapidly becoming one of the major areas of technological learning. In terms of mechanical properties, desirable characteristics of the two polymeric materials making up the IPN seem to have been preserved.

A number of variations of the above-mentioned full IPNs have also been stated in the literature.[40] One of them involves having either Polymer I or II as linear (not crosslinked) polymer, in which case it is called semi-IPN. The other variation involves the formation of Polymer I and II simultaneously through two noninterfering polymerization processes (such as stepwise and chain polymerizations); in which case it is called simultaneous IPN (SIN). If a linear polymer is formed simultaneously with a crosslinked polymer, then we have a semisimultaneous IPN (semi-SINS). Still another type is taking a mixture of two linear polymers, and crosslinking both components simultaneously, in which case it is called interpenetrating elastomeric network (IEN). The common feature of all these types of IPNs is that a monomer system is being polymerized in the presence of another polymer.

If physical crosslinks are used in an IPN system, then we would have a thermoplastic IPN.[41] Up to this time, few IPNs or SINs have been made where both components are elastomeric. One of these rare works involves the crosslinking of cis-1,4-polyisoprene (Elastollan) in thermoplastic polyurethanes (Morthane).[42]

Examples of IPN systems in which the pre-existing polymer was usually an elastomer include: a) ethyl acrylate, methyl methacrylate, or styrene and tetraethylene glycol dimethacrylate (TEGDM) as crosslinking agent was polymerized with dissolved benzoin or dicumyl peroxide by exposure to ultraviolet (UV) light or application of heat, respectively;[43] (b) butadiene crosslinked with peroxides using potassium persulfate as initiator in an emulsion;[44] (c) castor oil crosslinked by tolylene diisocyanate (TDI);[45] (d) styrene polymerized in SBR by thermal polymerization techniques;[44] (e) simultaneous IPN of castor oil crosslinked with TDI and styrene crosslinked with divinylbenzene (DVB) using benzoin and UV light.[45]

A work in IPN systems involving a Hytrel^{TM} copolyester includes crosslinking of methyl methacrylate (20%) with ethylene glycol dimethacrylate (1%) and an initiator in Hytrel 7246 fibers (3.1%) and alumina filler (61.9%),[46] wherein the product exhibited good impact and crack resistance. Another work involves the vulcanizing of acetal resins (Delrin 100F, 80%) with 0.01–5 phr isocyanates or isothiocyanates (Desmodur TT) in Hytrel 4056 (20%)[47] to produce high impact strength mixtures. The work in which chloroprene rubbers (Denka Chloroprene DCR-35, 50 parts) is vulcanized with diglycidyl phthalate (5 parts) in Hytrel HTC2551 (50 parts) resulted in tensile strength of 1.88 MPa, elongation of 390%, and Shore D hardness of 35 compared with 0.45 MPa, 40%, and 31, respectively, for a similar test piece without Hytrel.[48] An interesting type of IPN work with Hytrel involves the radiocuring (1.5 MeV of electron beam, Mrad) of poly(vinylidene fluoride) (30 parts) with triallyl isocyanurate (3 parts) in Hytrel (70 parts), which gave good flexibility, transparency, and flame resistance.[49]

PROCESS DESIGN CONSIDERATIONS

It is usually not possible to employ a simple commercial single- or twin-screw extruder for a given chemical reaction without considerable modification. For optimum operation, an extruder reactor must be custom designed with specific knowledge of the type of chemical reaction desired. Numerous parameters must be considered, such as the number of barrel segments and barrel diameter; length of the conveying screw; screw element design in

the individual barrel segments; venting locations; control of by-products and excess reactants; heat transfer; feed injection points; feed-rate control; feed quality, metallurgy, and corrosion; product quenching, cooling, and forming; and online analyzers.

The limited residence time available for reaction in an extruder is often a disadvantage. Although about 30 min is the maximum, less than 5-min residence time is usually observed in commercial extruders with realistic barrel lengths. Thus, the kinetics of the desired chemical reaction must be such that about 1–5 min would be the sufficient time for complete reaction, although when applicable, catalysis frequently shortens the time required for reaction.

Despite the advantages of operating without a solvent, there are disadvantages as well. The beneficial effects on the rate and outcome of reaction often observed in inorganic and polymer reactions by changing solvent polarity, hydrogen bonding ability, boiling point, and other parameters are not easily duplicated in a melt process. The absence of a liquid phase in most REX processes may make heat transfer inefficient and difficult to control in exothermic chemical reactions. Liquids may be highly effective heat-transfer agents, whereas most polymer melts have only low thermal conductivity. However, as a common characteristic of extruder reactors is the very high surface to volume ratio of reactant mixture, problems involving heat transfer are easily overcome. When liquid reactants are used, a certain minimum viscosity is required for efficient conveyance through an extruder. Variation in screw geometry and sequential extruders may be used to overcome problems with low-viscosity bulk reactants.

Small commercial continuous extruders require a minimum of 100–1000 g of material for a typical experiment. Such extruder reactions may not be convenient for running a large matrix of screening experiments to optimize processing conditions, especially when limited amounts of valuable starting materials are involved. Most commercial REX processes are simple one-step chemical reactions between inexpensive and readily available starting materials. REX is particularly suited for ''just-in-time'' inventory control of a commercial polymer product because the total time involved in reaction and product finishing is often much shorter than in conventional continuous processes.

CONCLUSIONS

The process REX involves chemical reactions within an extruder equipment and has several advantages over conventional manufacturing methods. Not only does it mean a reduction in equipment cost, but also unique advantages have been realized in cases wherein relatively fast chemical reactions occur in viscous fluids. However, limitations have been observed, such as limited residence times and inability to handle thin fluids. In general, if possible, REX is a method that should always be given serious consideration.

REFERENCES

1. Janssen, L.P.B.M. *Reactive Extrusion Systems*; Marcel Dekker, 2004.
2. Illing, G. Direct extrusion of nylon products from lactams. Modern Plastics **1969**, *46* (18), 73–476.
3. Kowalski, R.C. ACS Polym. Div. Topical Workshop on Polymerization and Polymer Modification by Reactive Processing. Bermuda, May 21–24, 1985.
4. Brown O. Reactive extrusion. In *Encyclopedia of Polymer Science and Technology*, 2nd Ed.; John Wiley and Sons, Inc., 1988; Vol. 14, 169–189.
5. Wauthier, H. Process for the Manufacture of Poly(ε-caprolactones) and Poly(ε-caprolactones) Which Have High Molecular Masses Obtainable by this Process U.S. Patent No. 5,468,837, Nov 21, 1995.
6. Wauthier, H. Process for the Continuous Manufacture of Poly(ε-caprolactones) U.S. Patent No. 5,656,718Aug 12, 1997.
7. Mumick, P.S.; Chiang, Y.; Wang, J.H. Temperature-sensitive Polymers and Water-dispersible Products Containing the Polymers U.S. Patent No. 5,969,052, Oct 19, 1999.
8. Mumick, P.S.; Chiang, Y.; Wang, J.H. Temperature-sensitive Polymers and Water-dispersible Products Containing the Polymers U.S. Patent No. 6,277,768, Aug 21, 2001.
9. Mumick, P.S.; Chiang, Y.; Wang, J.H. Temperature-sensitive Polymers and Water-dispersible Products Containing the Polymers U.S. Patent No. 6,410,155, June 25, 2002.
10. Mumick, P.S.; Chiang, Y.; Wang, J.H. Temperature-sensitive Polymers and Water-dispersible Products Containing the Polymers U.S. Patent No. 6,410,644, June 25, 2002.
11. Mumick, P.S.; Chiang, Y.; Wang, J.H. Temperature-sensitive Polymers and Water-dispersible Products Containing the Polymers U.S. Patent No. 6,451,429, Sept 17, 2002.
12. Yuo, W.-B.; Lee, S.-Y.; Lee, M.-S. Nylon-6 synthesis with Water, Alkali Metal Hypophosphite Catalyst and Amine Co-Catalyst U.S. Patent No. 5,264,541, Nov 23, 1993.

13. Yuo, W.-B.; Lee, S.-Y.; Lee, M.-S. Nylon-6 Catalyst SystemUS. Patent No. 5,290,747, Mar 1, 1993.
14. Priester Jr., R.D.; Vespoli, N.P.; Martinez III, F.C. Polyureas. In *Reaction Polymers – Chemistry, Technology, Applications, Markets*; Gum, W.F., Riese, W., Ulrich, H., Eds.; Hanser Publishers: New York, 1992; Section IIIC.
15. Morrison, R.T.; Boyd, R.N. *Organic Chemistry*; Allyn and Bacon, Inc.: Boston, 1983; 109–110.
16. Wang, J.H.; Schertz, D.M. Method of Modifying Poly(ethylene oxide) U.S. Patent No. 6,117,947, Sept 12, 2000.
17. Wang, J.H.; Schertz, D.M. Reactive Extrusion Method of Making Inverse Phase Blends of Poly (ethylene oxide) and Polyolefin U.S. Patent No. 6,225,406, May 1, 2001.
18. Johnson, E.D.; Wideman, G.J.; Wang, J.H. Moisture-induced Poly(Ethylene Oxide) Gel, Method of Making Same and Articles Using Same U.S. Patent No. 6,790,519, Sept 14, 2004.
19. Soerens, D.A.; Johnson, E.D.; Wideman, G.J.; Wang, J.H. Modified Poly(ethylene oxide), Method of Making Same and Articles Using Same U.S. Patent No. 6,610,793, Aug 26, 2003.
20. Xanthos, M.; Dagli, S.S. Compatibilization of polymer blends by reactive extrusion. Polym. Eng. Sci. **1991**, *31*, 929–935.
21. Aharoni, S.M.; Hammond, W.B.; Szobota, J.S.; Masilamani, D. Reactions in the presence of organic phosphates: I. High temperature amidation in the absence of solvents. J. Polym. Sci. Polym. Chem. **1984**, *22*, 2657.
22. Aharoni, S.M.; Forbes, C.E.; Hammond, W.B.; Hindenlang, D.M.; Mares, F.; O'Brien, K.; Sedgewick, R.D. High temperature reactions of hydroxyl and carboxyl PET chain end groups in the presence of aromatic phosphate. J. Polym. Sci. Part A: Polym. Chem. **1986**, *24*, 1281.
23. Hoeks, T.L.; Kusters, A.A.M.; Lin, Y.-G.; McCloskey, P.J.; Mestanza, R.; Wu, P.-P. Method for Branching and Crosslinking a Polycarbonate Resin U.S. Patent No. 6,087,468, July 11, 2000.
24. Mestanza, R.; Hoeks, T.L.; Kusters, A.A.M.; De Bont, J.J. Branched Polycarbonate Produced by Reactive Extrusion U.S. Patent No. 6,022,941, Feb 8, 2000.
25. Ghorpade, V.M.; Hanna, M. Method and Apparatus for Production of Levulinic Acid Via Reactive ExtrusionUS. Patent No. 5,859,263, Jan 12, 1999.
26. Chen Sr., P.N.; Glick, M.M.; Jaffe, M.M.; Forschirm, A. Muconic Acid Grafted Polyolefin Compatibilizers U.S. Patent No. 5,173,541, Dec 22, 1992.
27. Austin, R.G.; Patton, T.L Process for Converting Polyketones to Polyesters with Organic Peroxy Acid. U.S. Patent No. 5,180,797, Jan 19, 1993.
28. Reactive compounding done on-line with extrusion. In *Modern Plastics*; 1987; 12–15.
29. Kreisher, K. Custom compounders get flexible with reactive extrusion methods. In *Modern Plastics*; 1990; 36–38.
30. White, J.; Kim, B.-J. Block Copolymers of Lactone and Lactam, Compatibilizing Agents, and Compatibilized Blends U.S. Patent No. 6,486,257, Nov 26, 2002.
31. White, J.; Kim, B.-J. Block Copolymers of Lactone and Lactam, Compatibilizing Agents, and Compatibilized Blends U.S. Patent No. 6,835,774, Dec 28, 2004.
32. White, J ;Kim, B.-J. Block Copolymers of Lactone and Lactam, Compatibilizing Agents, and Compatibilized Blends. U.S. Patent No. 6,747,096, June 8, 2004.
33. Yoo, Y.-D.; Kim, Y.-W.; Cho, W.Y. Biodegradable polyethylene composition chemically bonded with starch and a process for preparing thereof U.S. Patent No. 5,461,094, Oct 24, 1995.
34. Shah, S.D.; Kakarala, S.; Schneider, J.E. Thermoplastic Polymer Alloy Compositions and In-line Compounding Process for Direct Sheet Extrusion of Sheets Prepared from the Thermoplastic Polymer Alloy Compositions U.S. Patent No. 6,153,680, Nov 28, 2000.
35. Kalambur, S.B.; Syed, SH.R. Starch-based nanocomposites by reactive extrusion processing. Polym. International **2004**, *53* (10), 1413–1416.
36. Todd, D.B. Consider reactive extrusion. Chem. Eng. Prog. **1992**, 72–74.
37. Sneller, J.A. Reactive processing: new era of innovation begins in resin production. In *Modern Plastics*; 1985; 56–60.
38. Newman, N.F.; Kowalski, R.C. Process for the Manufacture of Halogenated Elastomers U.S. Patent No. 4,384,072, May 17, 1983.
39. Sperling, L.H. *Encyclopedia of Polymer Science and Technology*, 1st Ed.; John Wiley and Sons, Inc.
40. Sperling, L.H.; Amts, R.R. Simultaneous interpenetration networks. J. Appl. Polym. Sci. **1971**, *15* (9), 2317–2319.
41. Sperling, L.H. *Interpenetrating Polymer Networks and Related Materials*; Plenum Press: New York.
42. Sperling, L.H. Interpenetrating polymer networks. In *Comprehensive Polymer Science–The Synthesis Characterization, Reactions and Applications of Polymers*; Allen, G., Bevington, J.C., Eds.; Pergamon Press: New York, 1989; Vol. 6 (Polymer Reactions), 423–436.
43. Mishra, V.; Murphy, C.J.; Sperling, L.H. Interpenetrating polymer networks based on thermoplastic polyurethanes (TPUs) and cis-1,4 polyisoprene. J. Appl. Polym. Sci. **1994**, *53*, 1425–1434.

44. Huelck, V.; Thomas, D.A.; Sperling, L.H. Interpenetrating polymer networks of poly(ethyl acrylate) and poly(styrene-co-methyl methacrylate). I. Morphology via electron microscopy and II. Physical and mechanical behavior. Macromolecules **1972**, *5* (4), 340–348.
45. Donatelli, A.A.; Sperling, L.H.; Thomas, D.A. Interpenetrating polymer networks based on SBR/PS. 1. Control of morphology by level of crosslinking and 2. Influence of synthetic detail and morphology on mechanical behavior. Macromolecules **1976**, *9* (4), 671, 676.
46. Yenwo, G.M.; Manson, J.A.; Pulido, J.; Sperling, L.H.; Conde, A.; Devia, N. Castor-oil-based interpenetrating polymer networks: synthesis and characterization. J. Appl. Polym. Sci. **1977**, *21* (6), 153.
47. Howard E.G., Jr. Impact-Resistant Heat-Formable Filler-Containing Polymer/Elastomeric Fiber Composites U.S. Patent Applied 87-96098, Sept 11, 1987.
48. Sasaki, K.; Yazaki, Y Block Copolyether Ester Compositions. Japanese Patent No. 61,091,245, May 9 Showa, 1986.
49. Shingyochi, K.; Kashiwazaki, S.; Ando, Y.; ana Yamazaki, M Radiocurable Polymer Blends. Japanese Patent No. 60,137,918 A2, July 22 Showa, 1985.

Reactive Separation

Vincent G. Gomes
University of Sydney, Sydney, New South Wales, Australia

INTRODUCTION

Many industrially important chemical reactions are limited by the equilibrium conversion of reactants within a feed and product mix. An important recent innovation in process design and operation is to combine the reaction and separation steps in a single unit operation known as reactive separation or integrated reactive separation and the process unit is called a multifunctional reactor. This combination has been recognized by the process industries as having favorable economics and as an important means for implementing process intensification for certain classes of reacting systems. This entry primarily provides an overview of reactive separation processes in use or under development and describes the design, operation, optimization, and control issues associated with selected classes of reactive separations. Particular emphasis is placed on reactions involving gas/liquid and gas/liquid/solid systems, both mediated by catalysts. The entry includes recent results and important on-going innovations in this fast-developing field.

PROCESS DESCRIPTION

A new class of unit operation, known as reactive separation or integrated reactive separation using a multifunctional reactor, is currently under various stages of development and implementation.[1–8] Process intensification presents one of the most important trends in today's process technology. It consists of the development of innovative processes that offer drastic improvements in chemical manufacturing and processing, substantially decreasing equipment volume, energy consumption, or waste formation, and ultimately leading to cheaper, safer, sustainable technologies.

Reactive separation processes are of significant industrial importance owing to the potential opportunities in generating novel products and their substantial technical and commercial advantages. Some of the advantages of conducting reaction and separation simultaneously include:

- *Improved conversion*: High conversion (close to 100% and beyond equilibrium) can be achieved because the removal of one or more products from the reaction phase could force the equilibrium to move to a higher level of conversion.
- *Improved selectivity*: Owing to the removal of products and establishment of favorable conditions for the main reaction, there is often reduction in the rates of side reactions.
- *Catalyst requirement*: Significantly reduced catalyst is required for the same degree of conversion.
- *Achieving difficult separations*: The process can help separate products that otherwise could not be separated, e.g., avoidance of azeotropes in multiple product streams in multicomponent systems or mixtures.
- *Heat transfer integration*: If the reaction is exothermic, the heat of the reaction can be used to reduce the overall heat duty required for separation. Some processes can be run at less severe conditions, resulting in longer catalyst life.
- *Multifunctional reactor*: Novel reactive separation design enables carrying out of several reactions in distinct dedicated zones.
- *Hot spots and runaways*: Avoidance of hot spots and runaways can be achieved by using milder conditions and thermal integration.
- *Cost*: Significant reduction in capital and operating costs can be achieved through simplification, reduction, or elimination of the usually expensive separation systems.

An important step is to decide on the separation methods that can be used to improve the performance of the reaction. The selection of a suitable separation method may be an iterative process and requires satisfaction of several factors, such as:

- Nature of the reaction
- Phases present
- Type of catalyst
- Operating conditions
- Residence time requirement
- Scaling up to large flows
- Removing of desired components and effect on other components
- Aspects such as fouling or foaming
- Process optimal condition mismatch for separation and reaction
- Process flexibility and control despite reduction in degrees of freedom.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120041308

Nominally, integration of reaction and separation may take place only at the equipment level, without introducing new functional interrelations between the operations. In that case, neither does the reaction have any influence on the separation, nor has the separation process any effect upon the reaction. The desired outcomes of such a combination are a smaller inventory, compact plant layout, and improved energy management. The Urea 2000+ technology represents a typical example of such a noninteracting integration.[7] In this process, the carbamate condenser, the urea reactor, and the inert scrubber have been successfully combined in a single vessel, the pool reactor. The integration resulted in a considerably smaller, less energy intensive, and relatively low-cost plant (height of equipment decreased 2.5 times). However, interrelations between the reaction and the other operations remained as in the conventional technology. In most cases, however, the reaction and separation are integrated to benefit from the interactions between the two, for instance, a shift of the product composition beyond the equilibrium via in situ separation, or enhancement of separation efficiency via chemical reaction.

Studies on combined reactive processes with separations have been carried out primarily for the following interacting configurations:

- Reaction with distillation
- Reaction with adsorption
- Reaction with permeation
- Reaction with absorption
- Reaction with leaching
- Reaction with extraction
- Reaction with crystallization.

A sample of the integrated processes is shown in Fig. 1.

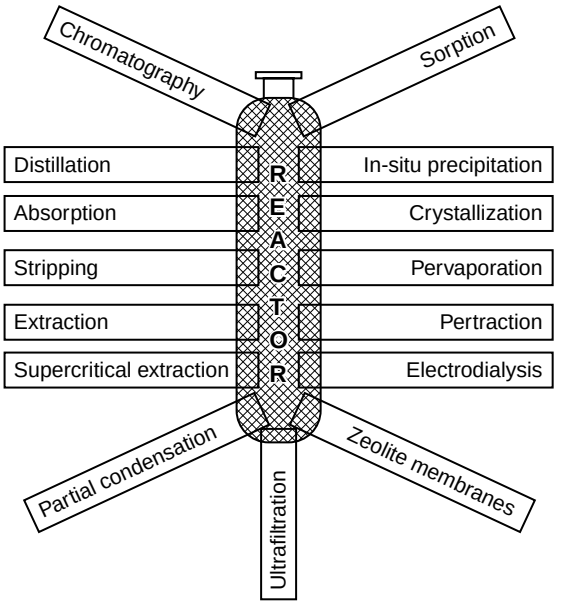


Fig. 1 In situ separation functions integrated into the reactor. (From Ref.[4].)

REACTIVE DISTILLATION

Overview

Distillation is a process of physically separating a mixture of liquids into two or more products that have different boiling points by preferentially boiling the more volatile components out of the mixture. Conversely, if a vapor is cooled, the less volatile (i.e., higher boiling point) material has a greater tendency to condense in a greater proportion than the more volatile material. Despite its antiquity, distillation is still a major unit operation in chemical plants due to certain fundamental kinetic and thermodynamic reasons:

- From a kinetic view, in distillation, mass transfer is limited only by the diffusional resistances on either side of the vapor–liquid interface because no inerts are present. Distillation, therefore, has the potential to deliver high mass transfer rates.
- From a thermodynamic view, typical efficiency is about 10% for distillation (can be improved with intercondensers). Most other separation processes are not more efficient.
- Distillation is cost-effective for separating liquid mixtures, except when:

  (a) The difference in volatility between components is small
  (b) A small quantity of feed is to be vaporized to recover small quantities of products
  (c) A component is thermally unstable even under vacuum
  (d) The mixture has highly corrosive or fouling tendencies.

The term reactive distillation (RD) refers to both catalyzed and uncatalyzed reaction systems. Catalytic distillation systems may use a homogenous or heterogenous catalyst to accelerate the reaction. Reactive distillation is a well-known example of reactive separation process, and is used commercially. The first patent and early journal articles deal mainly with homogenously catalyzed reactions such as esterifications, transesterifications, and hydrolysis.[9] Heterogenous catalysis with RD is a more recent development. The key advantages for a properly designed RD column are complete conversion of reactants and attainment of high selectivity. An example of the benefits of RD is the acid catalyzed production of methyl acetate by

Eastman Chemical, regarded as a classic illustration of the process intensification.[10,11] The reaction was traditionally carried out using the processing scheme shown in Fig. 2(A), which consists of one reactor and a train of nine distillation columns. In the RD implementation, only one column is required and nearly 100% conversion is achieved with significantly reduced capital and operating costs (Fig. 2B).

A substantial number of studies have been carried out with RD because of the significant benefits of the distillation process and its widespread use.[12–23] Other known applications of RD include:

- Manufacture of ethers (MTBE, ETBE, TAME).
- Hydration of ethylene oxide to monoethylene glycol.
- Selective hydrogenations of dienes and aromatics.

Processes in which RD may become potentially useful include:

- Decomposition of ethers to high-purity olefins
- Dimerization
- Alkylation of aromatics and aliphatics, e.g., ethylbenzene from ethylene and benzene, cumene from propylene and benzene, alkylation of isobutane with normal butenes
- Esterifications, e.g., ethyl acetate from ethanol and acetic acid
- Hydroisomerizations
- Hydrolyses
- Dehydrations of ethers to alcohols
- Oxidative dehydrogenations
- Carbonylations, e.g., *n*-butanol from propylene and syngas
- C1 chemistry reactions, e.g., methylal from formaldehyde and methanol

Design and Operational Considerations

A pilot scale RD column is shown in Fig. 3. Some of the constraints and difficulties in RD implementation are:[19]

- *Volatility constraints*: The reagents and products must have suitable volatility to maintain high concentrations of reactants and low concentrations of products in the reaction zone.
- *Residence time requirement*: If the residence time for the reaction is long, a large column size and large tray holdups will be needed.
- *Scale-up to large flows*: For large flow rates, liquid distribution problems arise.
- *Process condition mismatch*: Optimum conditions of temperature and pressure for distillation may be distant to optimal conditions for reaction and vice versa.
- *Construction*: Ease of installation, containment, and removal of the RD equipment and the catalyst are important.
- *Catalyst/liquid contact*: Good liquid distribution (avoidance of channeling and good radial dispersion

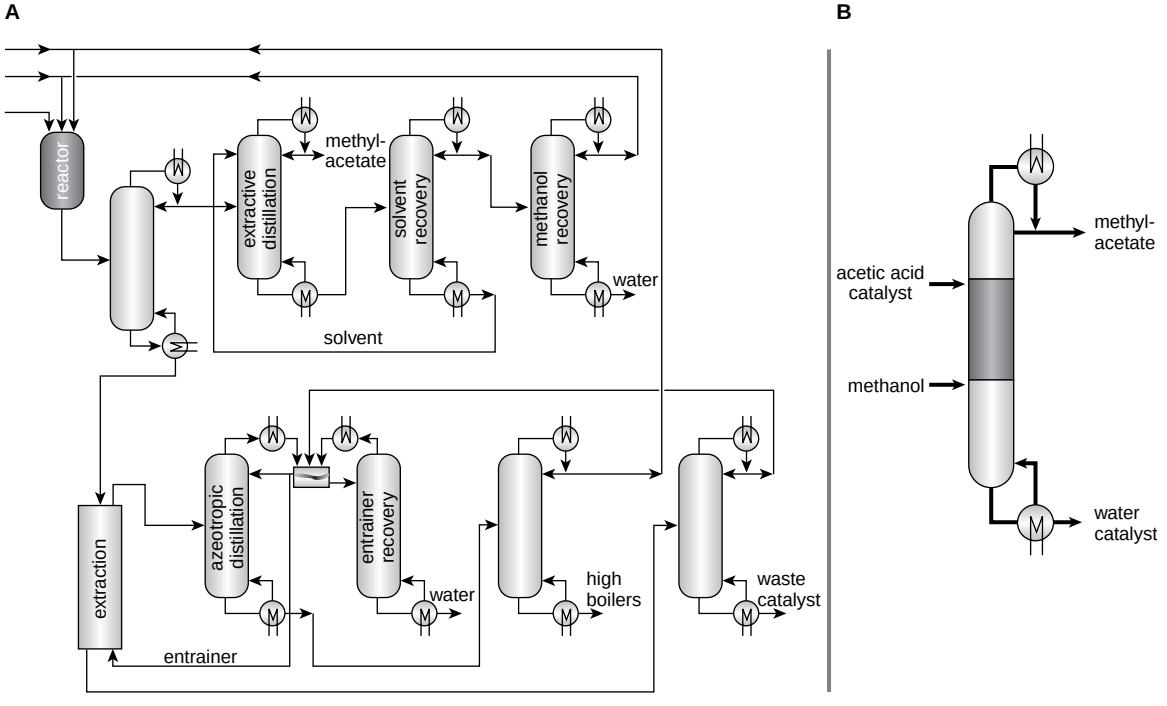


Fig. 2 Plant integration in methyl acetate separative reactor process by Eastman Chemical. (From Ref.[11].)

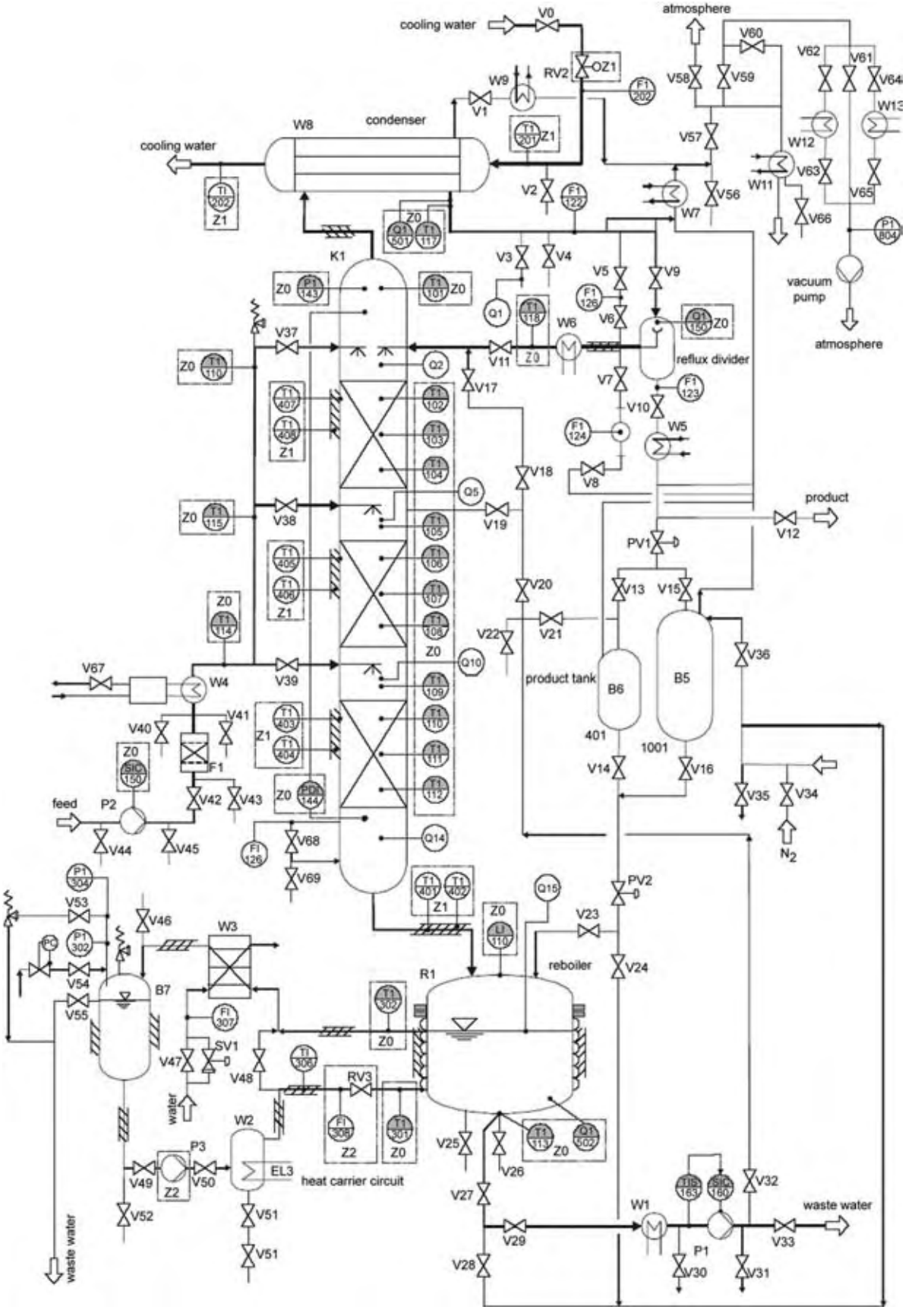


Fig. 3 Flowsheet of RD pilot plant. (From Ref.[22].)

of liquid) through the catalyst bed is required to ensure efficient mass transfer and to avoid reactor hot spots and runaways with even catalyst aging.
- *Vapor/liquid contact*: For fast reaction rates and equilibrium-limited reactions, the required size of the reactive zone is strongly influenced by vapor–liquid contact efficiency.
- *Column pressure drop*: This problem arises because of the need to use small catalyst particles (1–3 mm) through the catalytically packed reactive section to avoid intraparticle diffusional limitations.
- *Liquid holdup*: Liquid holdup, mean residence time, and liquid residence time distribution are important in determining conversion and selectivity.
- *Catalyst deactivation*: Catalyst deactivation is often accounted for during design by use of excess catalyst, and increase in reaction severity by increasing reflux (for increased residence time) or by increasing reaction temperature.
- Process flexibility and control aspects due to reduction in the degrees of freedom.

Hardware Considerations

The hardware issues regarding RD columns focus on the column internals such as the packings (random or structured packing, Fig. 4), distillation trays, and downcomers. Typical hardware design information is given in the standard sources as for conventional distillation.[24] Some of the important issues on hardware design aspects are discussed below.[19]

Catalytically packed columns

The catalyst particle sizes used in RD are usually of 1–3 mm range to avoid intraparticle diffusion limitations. To overcome the flooding limitations, the catalyst particles are contained within wire gauze envelopes. Most commonly, the catalyst envelopes are packed inside the column. Various shapes of catalyst envelopes have been patented. Some of these structures include:[18]

- Porous spheres filled with catalysts.
- Cylindrical envelopes with catalyst inside them.
- Wire gauze envelopes with various shapes: spheres, tablets, or doughnuts.
- Horizontally disposed wire mesh gutters filled with catalyst.
- Horizontally disposed wire mesh tubes containing catalyst.
- *Catalyst particles enclosed in bales*: Catalysts are loaded in pockets sewn into a fiberglass cloth. The resulting belt or catalyst quilt is rolled with layers of steel mesh to form a cylinder of catalyst bales, which are piled up to the required height. After the catalyst is spent, the column is shut down and the bales are replaced with the ones containing fresh catalyst.
- *Catalyst particles sandwiched between sheets of wire gauze*: These structures [licensed by Sulzer (KATAPAK-S) and Koch–Glitsch (KATAMAX)] consist of two pieces of crimped wire gauze sealed around the edge, forming a pocket 1–5 cm wide between the two screens. The catalyst sandwiches or wafers are bound in cubes, which are installed as monoliths inside the column to the required height. When the catalyst is spent, the packing is replaced with the ones containing fresh catalysts. An advantage of these structures over the catalyst bales is that the radial dispersion is about an order of magnitude higher.

Trays and downcomers

The catalyst envelopes placed in a trayed RD column are designed with various configurations:[18]

- Vertical envelopes, placed along the direction of the liquid flow path across a tray, are almost

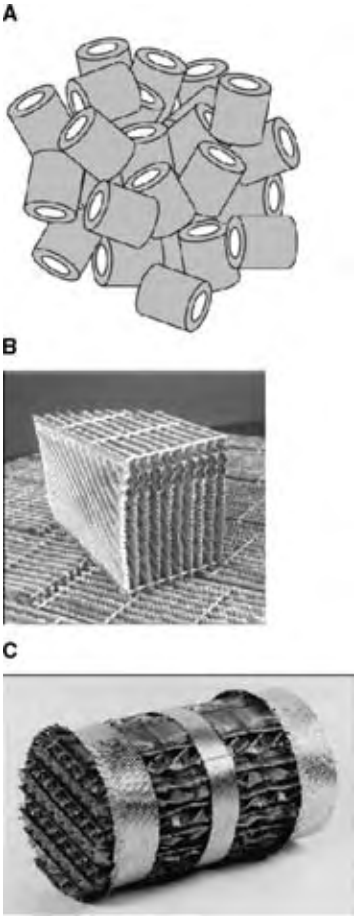


Fig. 4 (A) Random packings (Raschig rings). (B) Structured packings, KATAPAK (Sulzer). (C) Structured packings, MULTIPAK (Sulzer).

completely immersed in the froth on the tray, ensuring good contact between liquid and catalyst.
- Catalyst envelopes are placed within the downcomers. The primary drawback with installing the catalyst within downcomers is the limited volume available for catalyst inventory. Each stage is regarded as a reaction device (downcomer) followed by a separation section (froth on the tray).
- Catalyst envelopes are placed near the exit of the downcomer. Catalyst inventory is limited as the vapor does not pass through the catalyst envelopes.
- Trays and packed catalyst sections are used on alternate stages. The vapor flows through the packed section through a central chimney without contacting the catalyst. The liquid from the separation trays is distributed evenly into the packed reactive section below by a distribution device.
- Designs proposed for tray columns with catalyst containing pockets or regions that are fluidized by the liquid. Catalyst attrition is a concern in a fluidized bed and is taken care of by filtration of the liquid and by makeup of the catalyst.

Modeling and Control

The rate-based model often employed is based on the two-film theory and comprises the material and energy balances of a differential element of the vapor and of the liquid phase. The dynamic component balances for the liquid and the vapor are given by:[22]

$$\frac{\partial U_{\mathrm{L}i}}{\partial t} = \frac{D_{\mathrm{ax,L}}}{u_{\mathrm{L}}}\frac{\partial^2\left(Lx_i^{\mathrm{B}}\right)}{\partial z^2} - \frac{\partial\left(Lx_i^{\mathrm{B}}\right)}{\partial z} + \left(N_{\mathrm{L}i}a^i + N_{\mathrm{S}i}a^{\mathrm{cat}}\right)A_{\mathrm{c}} \quad i = 1,\ldots,\mathrm{nc}$$

$$\frac{\partial\left(Vy_i^{\mathrm{B}}\right)}{\partial z} - N_{\mathrm{L}i}a^iA_{\mathrm{c}} = 0$$

The dynamic energy balances are:

$$\frac{\partial E_{\mathrm{L}}}{\partial t} = \frac{\alpha_{\mathrm{ax,L}}}{u_{\mathrm{L}}}\frac{\partial^2(Lh_{\mathrm{L}})}{\partial z^2} + \frac{\partial(Lh_{\mathrm{L}})}{\partial z} + \left(Q_{\mathrm{L}}a^i + Q_{\mathrm{S}}a^{\mathrm{cat}}\right)A_{\mathrm{c}} - Q_{\mathrm{L}}^{\mathrm{HL}} \quad i = 1,\ldots,\mathrm{nc}$$

$$\frac{\partial(Vh_v)}{\partial z} - Q_{\mathrm{L}}a^iA_{\mathrm{c}} - Q_{\mathrm{V}}^{\mathrm{HL}} = 0$$

The symbols in the above equations denote the following: a is specific gas–liquid interfacial area (m^2/m^3); A is column cross-section (m^2); u_{L} is liquid velocity (m/sec); U is specific molar holdup (mol/m); V is vapor molar flow rate (mol/sec); L is liquid molar flow rate (mol/sec); x is liquid mole fraction; y is vapor mole fraction; Q is molar enthalpy (J/mol); N is molar flux [mol/(m^2 sec)]; Q is heat flux (W/m^2); $D_{\mathrm{ax,L}}$ is axial dispersion coefficient (m/sec^2); E is specific energy holdup (J/m); t is time (sec); z is axial coordinate (m); α is thermal dispersion coefficient (m^2/sec); nc is the number of components; subscripts i, j, k are component indices; superscript B is bulk phase.

A widely studied system is the synthesis of methyl acetate from methanol and acetic acid. This is a slightly exothermic, equilibrium limited liquid phase reaction:

$$CH_3OH + CH_3COOH \rightarrow CH_3COOCH_3 + H_2O$$

The low equilibrium constant and the strongly nonideal behavior causes the formation of the binary azeotropes: methyl acetate/methanol and methyl acetate/water, which pose challenges. As the column is operated at atmospheric pressure the vapor holdup is negligible. Moreover, reaction rates are taken into account by considering the solid catalyst phase and diffusion within the catalyst is neglected. The diffusional interactions such as osmotic or reverse diffusion that may occur are included in the Maxwell–Stefan equations. Modeling approaches for the simulation of the transient behavior of catalytic distillation columns require detailed information about the hydrodynamics of the column internals as structured packings and liquid distributors as well as the column periphery such as the top and bottom sections. A modified Langmuir–Hinshelwood–Hougen–Watson-based model was refitted and used with binary adsorption data.

Fig. 5 shows good agreement between the experimental and simulation results of dynamic liquid bulk concentrations. Because of its complexity the rate-based model is not suitable for controller design and optimization of the RD process. Therefore, an extended equilibrium stage model, which includes a reaction kinetic, is used for these tasks. Fig. 6 shows comparisons of simulation results of the rate-based model (RBA) and the equilibrium stage model for a typical trajectory of input variables. The dynamic behavior is covered well by the simplified model and the deviations between the absolute values are acceptable for control purposes. The advantage of substantially reduced computing time motivates the use of the simplified model for control and optimization purposes.

REACTIVE ADSORPTION

Overview

Adsorption separation combined with reaction has advantages similar to those described earlier for reactive separations. The primary advantages are, again,

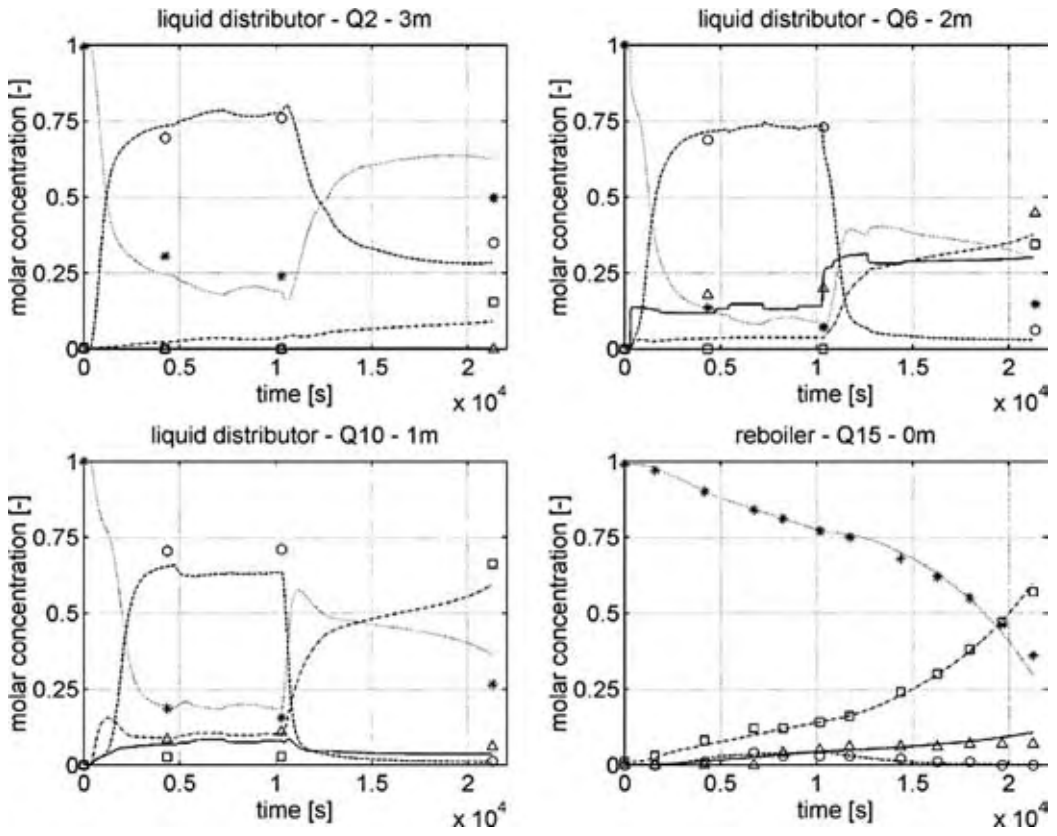


Fig. 5 Dynamic liquid bulk concentrations: experimental and simulation results. (Δ, acetic acid; ∗, methanol; ○, methyl acetate; □, water.) (From Ref.[22].)

significantly high conversion and selectivity and lower costs overall. Some suitable applications are as follows:

- Equilibrium- or selectivity-limited reactions.
- Reactions in which products can be separated by adsorption.
- Conditions for reaction and adsorption are compatible.
- Reaction with adsorption may be preferable when azeotropes or close-boiling products are obtained.
- Rapid mass transfer and low dispersion are desirable characteristics.
- *Heterogenous or homogenous*: The catalyst can be either a solid or a fluid. If a solid, the catalyst or adsorbent could be on the same particle or on discrete particles.

Specific applications include:

- Equilibrium-limited reactions

  Hydrogenation of 1,3,5-trimethylbenzene
  Methanol synthesis
  Ethyl acetate synthesis
  Methyl acetate synthesis
  Propene metathesis.

- Selectivity-limited reactions

  Isomerization of light paraffins
  Methane oxidative coupling.

Reactors

The types of reactors often used for integrated sorption with reaction are:

- Pressure swing reactors (PSR) or sorption enhanced reaction process (SERP)
- Chromatographic reactors
- Moving-bed reactors (MBR)
- Simulated moving-bed reactors (SMBR)
- Trickle bed reactors (TBR)

Currently, the application of adsorption-based processes to reaction systems are of considerable interest.[26–41] Hydrogen production from hydrocarbons and dehydrogenation are important industrial reactions, for example, the catalytic steam-methane reactor (SMR):

$$CH_4 + H_2O \rightarrow CO + 3H_2$$

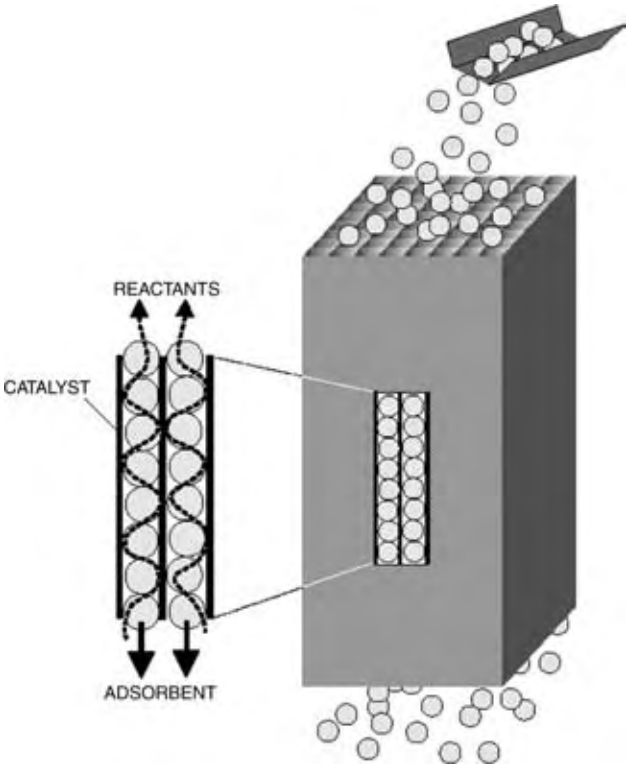


Fig. 6 Monolithic catalyst in moving-bed reactive adsorption system. (From Ref.[25].)

Steam–methane reaction is usually carried out at 50–600 psig and 750–900°C. The reactor effluent (70–72% H_2, 6–8% CH_4, 8–10% CO, 10–14% CO_2, and dry basis) is cooled and fed to another catalytic reactor for the water-gas shift reaction. The SERP concept for H_2 production by SMR has been experimentally demonstrated at a temperature of 450°C and a pressure of 55 psig (480 kPa). The reactor directly produced a 95 mol% pure H_2 stream containing 5 mol% CH_4 and less than 50 ppm of carbon oxide impurities using a feed gas containing steam and methane in the ratio of 6 : 1. The CH_4 to H_2 conversion was 82%. A conventional SMR reactor operated at 450°C under equivalent feed gas conditions would produce a product stream containing 53% H_2, 34% CH_4, and 13% carbon oxides with a CH_4 to H_2 conversion of only 28%. A temperature of 650°C would be necessary for a conventional SMR reactor to match the conversion of the SERP concept operated at 450°C using the same feed gas conditions. The SMR reactor effluent in that case would be 75.5% H_2, 4.4% CH_4, and 20.1% carbon oxides. Thus, the SERP operation provides substantial advantages.

The usual PSR or SERP reactors combine pressure swing adsorber with a periodic flow-forced fixed-bed reactor, and involve feeding reactants in a reactor having a mixture of catalysts and sorbents. The sorbents selectively remove some of the reaction products via physical and/or chemical sorption. Processes involving heterogenous catalysis reactions are by far the most dominant (>90%) in the chemical industry. Thus, considerable techno-economic incentives exist in combining chemical reaction and separation of products in a single unit operation. However, some of the challenges include a lack of validated tools for design and scaling of these reactors, a lack of knowledge about materials to mediate separation, and major gaps in techno-economic evaluation.

The simulated moving-bed reactor integrates continuous countercurrent chromatographic separation with chemical reaction. Such a combination allows higher conversions and better yield to be achieved by separating educts and products of an equilibrium reaction from each other. The movement of the bed with regard to the reactants inlets/outlets is usually realized in a rotating system. Another variant of the rotating cylindrical annulus chromatographic reactor is having the inlets of the mobile phase uniformly distributed along the annular bed entrance, while the feed stream is stationary. Because of the rotation of the reactor, the selectively adsorbed species traverse different helical paths through the bed and are collected at fixed locations. Alternatively, the reactor can be held stationary and the feed rotated. The hydrolysis of aqueous methyl formate and dehydrogenation of cyclohexane to benzene have been investigated using this technique.[26]

Another variant of adsorptive reactor is the gas–solid–solid trickle flow reactor, in which fine adsorbents trickle through the fixed bed of the catalyst, removing selectively one or more products from the reaction zone.[42] In case of methanol synthesis this leads to conversions significantly exceeding the equilibrium conversions under given conditions. The economics of methanol process based on this reactor was compared with a conventional low-pressure Lurgi process. For a 1000 tons/day production, the new technology offered considerable reductions in cooling water use (50%), recirculation energy (70%), raw materials (12%), and catalyst amount (70%). A further improvement of the concept has been in applying a moving bed of adsorbent through parallel channels of a monolithic catalyst, as shown in Fig. 6.[25]

Modeling of PSR

A fundamental understanding of sorption processes requires a detailed mechanistic knowledge of the equilibria, kinetics, and dynamics of the sorption process. The PSR is a cyclic batch process for which adsorption is carried out at a relatively higher pressure and desorption (regeneration) is accomplished at a lower pressure, generally using part of the product from the adsorption

step. During the high-pressure adsorption step, the preferentially adsorbed species are retained in the column, leaving the purified raffinate product in the gas phase. During desorption (regeneration), the adsorbed species are removed from the adsorbent, thus regenerating the column for use in the next cycle. The two-bed PSR (Fig. 7) involves four main steps:[39]

- *Pressurization* (*step 1*): One column is pressurized to the operating pressure of the adsorption step with feed gas entering one end, while keeping the other end closed. During this stage, the species with the lower sorption affinity is pushed toward, and hence is enriched, at the closed end.
- *Adsorption* (*step 2*): The concentration wave front developed at the pressurization step travels along the column during this stage such that the raffinate mainly contains the less strongly adsorbed species. A portion of it is used as a feed into the other column for desorption, while the remainder is delivered as a product.
- *Blowdown* (*step 3*): The pressure of the bed is lowered to atmospheric or subatmospheric pressure. The countercurrent direction of the flow reverses the direction of the traveling concentration wave front of the less sorbed species. As a result, the product end of the bed is not contaminated with the more strongly adsorbed species.
- *Countercurrent purge* (*step 4*): The raffinate gas from the adsorption step of the other bed is used to desorb the adsorbate and to remove it from the voids so that its amount within the bed is reduced at the start of the next cycle. Similar to blowdown, this step ensures that most of the strongly adsorbed species is removed from the product end.

The mathematical model for this process was derived based on the assumptions that the effect of thermal excursions is negligible, the total pressure remains constant during the high- or low-pressure steps, the fluid velocity variation along the bed is determined by the mass balance, the flow is governed by the axially dispersed plug flow model, the equilibrium relations are given by the extended Langmuir isotherm, the mass transfer rates are represented by the linear driving force (LDF) equations, and the rate coefficients are the same for both the high- and the low-pressure steps. The dynamic behavior of flow and species concentrations in the two columns is given below.

Component mass balance for reactants in the gas phase:

$$\frac{\partial c_{ij}}{\partial t} - D_Z \frac{\partial^2 c_{ij}}{\partial z^2} \pm v_j \frac{\partial c_{ij}}{\partial z} \pm c_{ij} \frac{\partial v_j}{\partial z} + \frac{1 - \varepsilon}{\varepsilon} \frac{\partial \overline{q}_{ij}}{\partial t} - v_i \rho_s R_i = 0 \tag{1}$$

Component mass balance for products in the gas phase:

$$\frac{\partial c_{ij}}{\partial t} - D_Z \frac{\partial^2 c_{ij}}{\partial z^2} \pm v_j \frac{\partial c_{ij}}{\partial z} \pm c_{ij} \frac{\partial v_j}{\partial z} + \frac{1 - \varepsilon}{\varepsilon} \frac{\partial \overline{q}_{ij}}{\partial t} + v_i \rho_s R_i = 0 \tag{2}$$

Overall mass balance:

$$\pm C_j \frac{\partial v_j}{\partial z} + \frac{\partial C_j}{\partial t} + \frac{(1 - \varepsilon)}{\varepsilon} \sum_i \frac{\partial \overline{q}_{ij}}{\partial t} = 0 \tag{3}$$

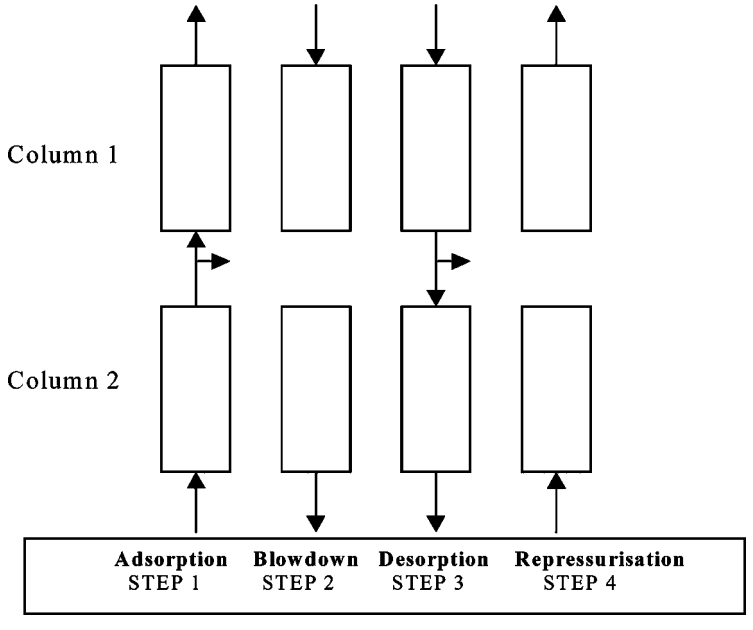


Fig. 7 Schematic diagram of a two-bed PSA cycle.

The mass transfer rate (from LDF or LDF model) across the external film is given by:

$$\frac{\partial \bar{q}_{ij}}{\partial t} = k_i(q^*_{ij} - q_{ij}) \quad (4)$$

$$q^*_{ij} = q_{is}\frac{b_i c_{ij}}{1 + \sum_i b_i c_{ij}} \quad (5)$$

Continuity condition:

$$C_j = \sum c_{ij} \quad (6)$$

Initial conditions:

$$c_{ij}(z = 0) = 0; \quad \bar{q}_{ij}(z = 0) = 0 \quad (7)$$

Boundary conditions for fluid concentration:

$$D_Z\left[\frac{\partial c_{ij}}{\partial z}\right]_{z=0} = -v_j|_{z=0}(c_{ij}|_{z=0^-} - c_{ij}|_{z=0^+})$$

$$\frac{\partial c_{ij}}{\partial z}_{(z=L)} = 0 \quad (j = 1 \text{ and } 2) \quad (8)$$

$$\frac{\partial c_{ij}}{\partial z}_{z=0} = 0; \quad \frac{\partial c_{ij}}{\partial z}_{(z=L)} = 0 \qquad (j = 3) \quad (9)$$

$$D_Z\left[\frac{\partial c_{ij}}{\partial z}\right]_{z=L} = -v_j|_{z=L}(c_{ij}|_{z=L^-} - c_{ij}|_{z-L^+})$$

$$\frac{\partial c_{ij}}{\partial z}_{(z=0)} = 0 \quad (j = 4) \quad (10)$$

Boundary conditions for fluid velocity:

$$v_j(z = L) = 0 \quad (j = 1 \text{ and } 3) \quad (11)$$

$$v_j(z = 0) = v_{0j} \quad (j = 2) \quad (12)$$

$$v_j(z = L) = v_{0j} \quad (j = 4) \quad (13)$$

$$\frac{\partial c_{ij}}{\partial z}_{(z=L)} = 0 \quad (j = 2) \quad (14)$$

In the equations above, subscript "i" denotes a gas species, subscript "j" represents a specific step; t denotes time; z is bed axial distance; c is gas phase mole fraction; q denotes adsorbed phase mole fraction; v is reaction stoichiometry for the species; R_i is the rate of reaction; v is interstitial fluid velocity; v_0 is inlet gas velocity; D_Z is dispersion coefficient; k is LDF coefficient; b is sorption coefficient; ε is bed voidage. For both the pressurization ($j = 1$) and the adsorption ($j = 2$) steps, the ($\pm$) signs are replaced by ($+$), while for the blowdown ($j = 3$) and purge ($j = 3$) steps, the ($\pm$) signs are replaced by the ($-$) sign. The ($+$) sign denotes flow from $z = 0$ to L, while the ($-$) sign denotes flow from $z = L$ to 0. The expressions containing $\bar{b}$ denote adsorbed phase concentration averaged over the adsorbent particle; $c_{|z=0^-}$ appearing in the boundary condition represents the concentration of the species just prior to feeding into the column inlet, while $c_{|z=0^+}$ denotes the concentration just inside the column.

Reactor Operation with and without Periodic Separation

The enhancement of conversion through the use of PSR was tested for a metathesis reaction, which is industrially relevant in producing ethene from propene and vice versa.[41]

$$2CH_2 = CH - CH_3 \leftrightarrow CH_2 = CH_2 + CH_3 - CH = CH - CH_3$$

The importance of this reaction is in maximizing the production of ethene (via forward reaction) or that of propene (via reverse reaction), depending on whether the demand for poly(ethylene) exceeds that for poly(propylene) or vice versa.

Initially, a fixed-bed reactor with no separation was tested. The reactor was packed only with the catalyst and inert filler material with no adsorbent. Further, the feed gas propene (9% by volume) was mixed with helium as a carrier. No separation via PSR operation was imposed on the reactor contents. Fig. 8 shows that the steady-state conversion of propene is significantly less than 1% (about 0.03%) and the product recovery is almost negligible. The slight time lead in simulation is due to nonaccounting of the dead time in experiments. The estimated equilibrium conversion from thermodynamics is about 24% based on pure propene feed. Thus, the conversion in a fixed-bed reactor is far below the equilibrium conversion.

The conversion could be enhanced for the forward reaction if the reverse reaction involving ethene and 2-butene is minimized. Further, since 2-butene desorption is a controlling factor due to its strong sorptive properties, 2-butene removal, in particular, will allow improved rate of reaction and product separation with the use of a sorbent such as γ-Al_2O_3. The effect of the PSR operation (cycle time = 40 sec) on the reactor performance was tested through simulations and by conducting experiments. Step inputs in inlet feed composition containing propene with helium carrier were conducted with clean sorbent beds and constant total gas flow rates. The results, compared with theoretical predictions, are shown in Fig. 9. Because ethene has

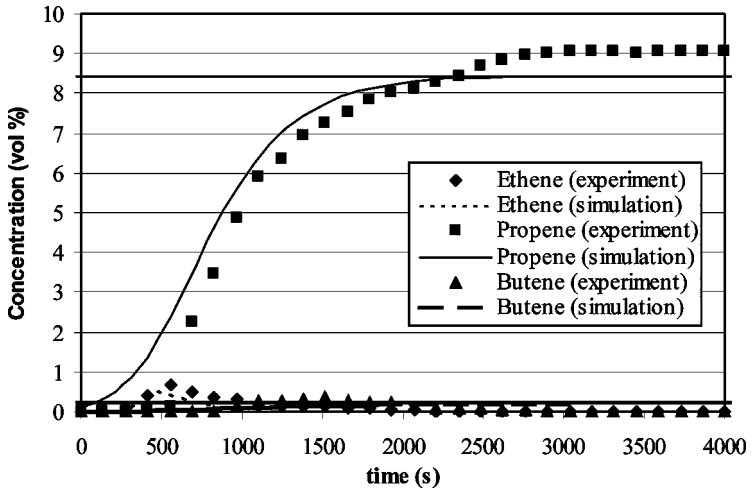


Fig. 8 Simulations and experimental results: gas phase exit composition for propene metathesis in fixed-bed catalytic reactor without pressure swing effect. (From Ref.[41].)

the lowest affinity toward γ-alumina, it is the first species to elute, followed by propene.

Both the simulation and the experimental results show that ethene is enriched to around 25% before stabilizing to a value of 6%. Simulation results are also close to those obtained from experiments. Simulation results, as a function of bed length and time, show that ethene, owing to its low sorption affinity is substantially affected by the sorption–desorption behavior of propene and 2-butene. Initially, the small amounts of ethene formed are obtained at the exit with no other component. As the reaction proceeds, additional 2-butene and propene are preferentially adsorbed and ethene is largely displaced from the adsorbent.

However, as 2-butene and propene concentration increases, re-equilibration of gases occurs in the sorbed phase with the result that ethene exit gas concentration decreases to attain a cyclic steady state (CSS) value of about 0.1 in about cycle 10. It may be noted that the steady-state exit concentrations of the desired product, ethene, is enhanced substantially over the case with no PSR, as described earlier. Further, the output of coproduct, 2-butene, though greater than the case without PSR, makes up a much smaller proportion of the total product mix at the column outlet. Thus, both the total yield of the product and the proportion of the desirable product have been increased by using PSR.

Performance Indicators

Ethene fraction in product stream at column exit (C_E):

$$C_{\mathrm{E}} = \frac{\text{Amount of ethene in product}}{\text{Amount of (ethene + propene + butene) in product}}$$

Recovery of ethene (RE_E):

$$\mathrm{RE_E} = \frac{\text{Amount of ethene in product}}{\text{Total ethene produced by the catalyst}}$$

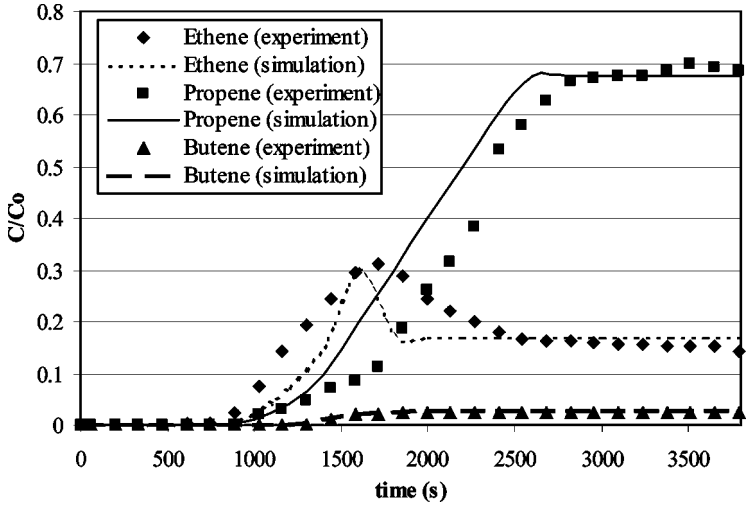


Fig. 9 Simulations and experimental results: gas phase exit composition for propene metathesis in fixed-bed catalytic reactor with PSR (cycle time = 240 sec, Co = 1.2×10^{-4} mol/cm^3, $C_{He} = 8 \times 10^{-5}$ mol/cm^3). (From Ref.[41] with permission.)

Butene–ethene ratio (B/E): It measures the extent of separation of the two products within the adsorption columns. If butene is considered as a by-product, this ratio also relates to product impurity.

Yield of ethene (Y_E): Owing to the adsorption of both the reactants and the products, the yield of ethene, rather than propene conversion, is used:

$$Y_{\mathrm{E}} = \frac{\text{Amount of ethene at column exit}}{\text{Ethene generated via metathesis with no physical adsorption effects}}$$

Catalyst/adsorbent productivity (Pd): It is the molar amount of ethene collected per unit amount of catalyst and adsorbent per unit time.

The performance indicators for the three cases of reaction are shown in Table 1. The analysis shows that the use of pressure swing reaction in propene metathesis has several advantages. First, a higher proportion of ethene is obtained in the product stream with pressure swing reaction. Based on a cycle time of 240 sec, ethene constitutes about 20% of the product stream, while it is about 7% without pressure swing and 4% in the absence of the 13× adsorbent. Further, the separation between ethene and 2-butene in the product stream for pressure swing reaction produces four times as much ethene as that of 2-butene. In contrast, the ethene and 2-butene gases are obtained in almost equal amounts from the other alternative methods investigated. If 2-butene is regarded as a by-product, the percentage of impurity of the product stream will be reduced with the use of pressure swing reaction. Further, the total productivity of the catalyst and the adsorbent system is increased significantly by PSR. Propene metathesis is favored by the use of pressure swing, with an increase in ethene yield by a factor of 3. The adsorbent is regenerated in situ by pressure swing to increase the total adsorption capacity. Hence, the product species, 2-butene in particular, is continuously adsorbed and removed from the gas stream, resulting in an increase in reaction rate.

The effects of changes in total cycle time and in propene feed rate were also investigated with the help of simulations. For enhanced product purity, PSR with a relatively short cycle time is favored, because the reaction is enhanced by the increased frequency of adsorbent regeneration, even though the adsorption is equilibrium controlled. This is demonstrated in Table 1, where ethene fraction, ethene yield, and adsorbent productivity are almost doubled as the total cycle time is decreased from 440 to 120 sec. However, the differences between these indicators are marginal for cycle times of 120 and 240 sec. The recovery increases on increasing the cycle time with decrease in product purity. Further, the separation of 2-butene from the product stream becomes less efficient, as B/E ratio is increased from 0.11 to 0.37.

OTHER REACTIVE SEPARATIONS

Membranes

The use of perm-selective membranes for separation in conjunction with reactors is gaining substantial attention.[43–46] An example of the various functions a membrane may have is shown in Fig. 10.[45] The area of membrane reactors is replete with interesting ideas, for example, heat- and mass-integrated combination of hydrogenation and dehydrogenation processes in a single membrane unit. However, no large-scale industrial applications of catalytic membrane reactors have been reported so far, the main reason being the relatively high price of membrane units, in addition to low permeability, sealing difficulty, as well as mechanical and thermal fragility of membranes. Potential areas for applications of catalytic membrane reactors are:

- Methane steam reforming
- Dehydrogenation, e.g., ethane to ethene, ethylbenzene to styrene
- Water-gas shift reaction

Table 1 Reaction/separation performance with variable process conditions

	Reactor without adsorbent and without PSR	Reactor with adsorbent and without PSR	Reactor operation with PSR (cycle time: 240 sec)
C_{E} (%)	4.0	7.4	19.7
$\mathrm{RE_E}$(%)			55.5
B/E	1	1	0.25
Y_{E} (%)	7.9	10.4	30.7
Pd $\times$ 10^4 (mol/kg/sec)	2.4	2.4	5.8

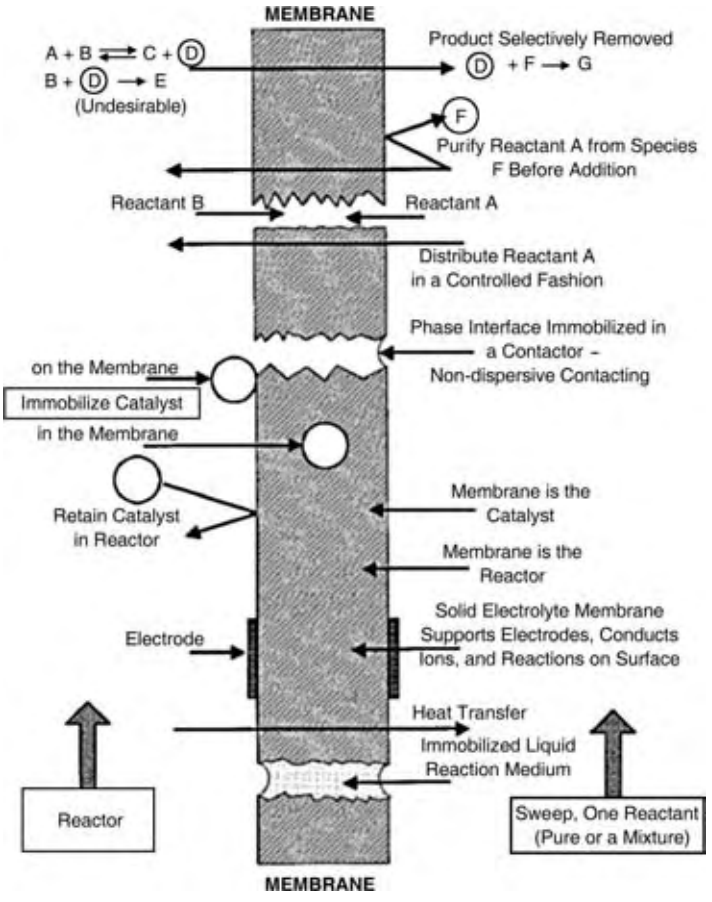


Fig. 10 Membrane functions in chemical reactor. (From Ref.[45].)

- Selective oxidations, e.g., butane to maleic anhydride, ethylene to ethylene oxide
- Oxidative dehydrogenation of hydrocarbons
- Oxidative coupling of methane.

Membranes are being frequently employed in the manufacturing of pharmaceuticals in combination with a bioreactor for enzymatic reactions. In DSM such a combination has been studied for the production of *S*-ibuprofen, via the hydrolysis of the (*R,S*)-ibuprofen methylester coupled to a racemization of the unwanted enantiomer.[46] The esterase used for the above conversion is strongly deactivated by the product. To solve this problem, an ultrafiltration membrane unit has been coupled to the reactor, to remove in situ the product formed. The application of the ultrafiltration has led to a twofold increase of the conversion/productivity, as shown in Fig. 11.

Extraction

Reactive extraction processes include simultaneous reaction and liquid phase separation. The immiscibility may occur naturally within the reactive system or is introduced deliberately by adding solvents. Reactive extraction can be used for significant improvements in selectivities or yields of desired products in multireaction systems, thereby reducing recycle flows and waste formation. The combination of reaction with liquid/liquid extraction can also be used for the separation of waste by-products that are difficult to separate using conventional techniques.[47,48]

Crystallization

Reactive crystallization or precipitation processes of industrial relevance include liquid-phase oxidation of *para*-xylene to terephthalic acid, acidic hydrolysis of sodium salicylate to salicylic acid, and the absorption of ammonia in aqueous sulfuric acid to form ammonium sulfate.[49] A special type of reactive crystallization is the diastereomeric crystallization, widely applied in the pharmaceutical industry for the resolution of the enantiomers. Here, the racemate is reacted with a specific optically active material (resolving agent), to produce two diastereomeric derivatives (usually salts), which are separated by crystallization. Diastereomeric crystallization is commonly used in

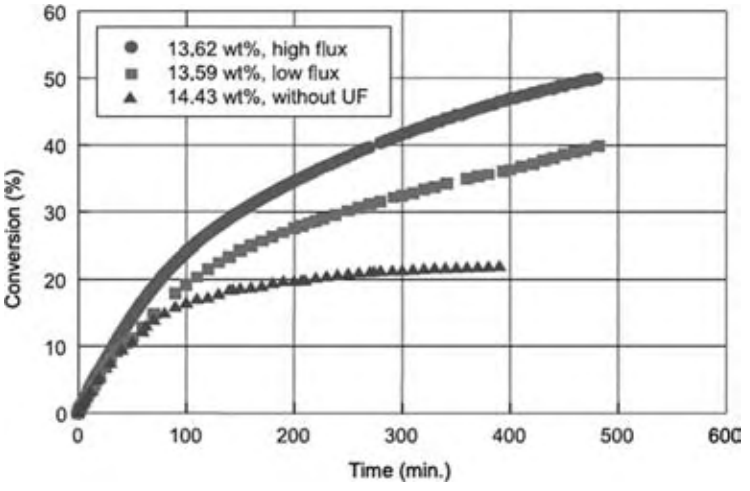


Fig. 11 Ibuprofen methylester conversion as a function of time, with and without integrated ultrafiltration unit. (From Ref.[46].)

the production of a number of pharmaceuticals, such as ampicillin, ethambutol, chloramphenicol, diltiazem, fosfomycin, and naproxen. Recently, reactive precipitation has successfully been used for the production of nanosized cubic particles of $CaCO_3$. Ultrafine particles with a mean size of 15–40 nm and a narrow size distribution were produced by the carbonation of lime suspension in a rotating packed-bed reactor.[50] The reaction times were 4–10-fold shorter than the corresponding reaction times in a conventional stirred-tank unit.

Absorption

Reactive absorption is the most widely applied type of a reactive separation process. It is used for production of a number of bulk chemicals, such as nitric and sulfuric acid. It is also often employed in gas purification, e.g., removal of carbon dioxide or hydrogen sulfide. Other areas of application include olefin/paraffin separations, where reactive absorption with reversible chemical complexation is a promising alternative to cryogenic distillation.

CONCLUSIONS

There are significant gains to be achieved through multiunctional reactor design concepts. There are major opportunities in partial or discretized multifunctionality. Rapid progress is expected in the application of RD, reactive adsorption, and membrane reactors, including bioreactors. Reactive absorption and adsorption processes may further be intensified by the use of rotating equipment, in which high-gravity fields enhance the mass transfer rates and which has almost exclusively been used for nonreactive separations. A type of rotating equipment has already been successfully applied on a large scale in deaeration of water in oil fields, by stripping it with natural gas in a rotating bed; another is offered commercially for countercurrent liquid/solid adsorption/ion exchange.[50] First commercialization of the reactive stripping in rotating equipment has also taken place.[51] Combination of Higee technology with reactive separations may lead to significant compacting of the process equipment resulting in smaller, cleaner, and more energy-efficient chemical plants.

The advantages of reactive separation are unfulfilled if design or operation is not proximate to the optimal requirements for the various integrated coprocesses. The problems often relate to incompatibility of operating conditions for separation. The studies cited demonstrate that high conversions are achievable at reduced temperatures. The problems lie in managing the requirements of the two processes, or of identifying separation compatible with the demands of the catalytic reaction. Forcing conditions for the combined process may not work economically. For example, a detailed process economic evaluation of RD for toluene disproportionation showed that the potential benefits of process simplification offered from the multifunctionality of RD are negated by the need to design the unit at conditions that are optimal for neither reaction nor distillation.[52]

ACKNOWLEDGMENTS

The author wishes to acknowledge permission granted by Elsevier (Figs. 1–3, 5, 6, 11), the Berkeley Electronic Press (Figs. 8, 9), and the American Chemical Society (Fig. 10) for reprinting figures referred in text.

REFERENCES

1. Stankiewicz, A.; Moulijn, J.A. Process intensification: transforming chemical engineering. Chem. Eng. Prog. **2000**, *96*, 22–34.
2. Agar, D.W. Multifunctional reactors: old preconceptions and new dimensions. Chem. Eng. Sci. **1999**, *54*, 1299–1305.
3. Kulprathipanja, S., Ed. *Reactive Separation Processes*; Taylor & Francis: Philadelphia, 2001.
4. Krishna, R. Reactive separations: more ways to skin a cat. Chem. Eng. Sci. **2002**, *57*, 1491–1504.
5. Schembecker, G.; Tlatlik, S. Process synthesis for reactive separations. Chem. Eng. Process. **2003**, *42*, 179–189.
6. Stankiewicz, A. Reactive separations for process intensification: an industrial perspective. Chem. Eng. Process. **2003**, *42*, 137–144.
7. Stamicarbon, B.V. A low cost design for urea. Nitrogen **1996**, *222*, 29–31.
8. DeGarmo, J.L.; Parulekar, V.N.; Pinjala, V. Consider reactive distillation. Chem. Eng. Prog. **1992**, *88*, 43–50.
9. Backhaus, A.A Continuous Processes for the Manufacture of Esters. U.S. Patent 1,400,849, 1921.
10. Siirola, J.J. An industrial perspective on process synthesis. AIChE. Symp. Ser. **1995**, *91* (304), 222–233.
11. Schoenmakers, H.G.; Bessling, B. Reactive and catalytic distillation from an industrial perspective. Chem. Eng. Process. **2003**, *42*, 145–155.
12. Stadig, W.P. Catalytic distillation: combining chemical reaction with product separation. Chem. Proc. **1987**, *50*, 27–32.
13. Fuchigami, Y. Hydrolysis of methyl acetate in distillation column packed with reactive packing of ion exchange resin. J. Chem. Eng. Jpn. **1990**, *23*, 354–358.
14. Doherty, M.F.; Buzad, G. Reactive distillation by design. Trans. Inst. Chem. Eng. **1992**, *70*, 448–458.
15. Sundmacher, K.; Hoffmann, U. Activity evaluation of a catalytic distillation packing for MTBE production. Chem. Eng. Technol. **1993**, *16*, 279–289.
16. Kenig, E.; Jakobsson, K.; Banik, P.; Aittamaa, J.; Gorak, A.; Koskinen, M.; Wettmann, P. An integrated tool for synthesis and design of reactive distillation. Chem. Eng. Sci. **1999**, *54*, 1347–1352.
17. Ben Armor, H.; Halloin, V.L. Methanol synthesis in a multi-functional reactor. Chem. Eng. Sci. **1999**, *54*, 1419–1423.
18. Taylor, R.; Krishna, R. Modelling reactive distillation. Chem. Eng. Sci. **2000**, *55*, 5183–5529.
19. Towler, G.P.; Frey, S.J. Reactive distillation. In *Reactive Separation Processes*; Kulprathi-panja, S., Ed.; Taylor & Francis: Philadelphia, 2001.
20. Luo, H.-P.; Xiao, W.-D. A reactive distillation process for a cascade and azeotropic reacting system: carbonylation of ethanol with dimethyl carbonate. Chem. Eng. Sci. **2001**, *56*, 403–410.
21. Noeres, C.; Kenig, E.Y.; Gorak, A. Modelling of reactive separation processes: reactive absorption and reactive distillation. Chem. Eng. Process. **2003**, *42*, 157–178.
22. Noeres, C.; Dadhe, K.; Gesthuisen, R.; Engell, S.; Gorak, A. Model-based design, control and optimisation of catalytic distillation processes. Chem. Eng. Process. **2004**, *43*, 421–434.
23. Engell, S.; Fernholz, G. Control of a reactive separation process. Chem. Eng. Process. **2003**, *42*, 201–210.
24. Stichlmair, J.G.; Fair, J.R. *Distillation Principles and Practice*; Wiley–VCH: New York, 1998.
25. Kapteijn, F.; Heiszwolf, J.J.; Nijhuis, T.A.; Moulijn, J.A. New non-traditional multiphase catalytic reactors based on monolithic structures. Catalysis Today **2001**, *66* (204), 133–144.
26. Fish, B.; Carr, R.; Aris, R. The continuous countercurrent moving bed chromatographic reactor. Chem. Eng. Sci. **1986**, *41*, 661.
27. Carr, R.W. Continuous reaction chromatography. In *Preparative and Production Scale Chromatography*; Ganetsos, G., Barker, P.E., Eds.; Marcel Dekker: New York, 1993; 421–447.
28. Mazotti, M.; Kruglov, A.; Neri, B.; Gelosa, D.; Morbidelli, M. A continuous chromatographic reactor: SMBR. Chem. Eng. Sci. **1996**, *51*, 1827–1836.
29. Vaporciyan, G.G.; Kadlec, R.H. Periodic separating reactors: experiments and theory. AIChE J. **1989**, *35* (5), 831.
30. Chatsiriwech, D.; Alpay, E.; Kershenbaum, L.S.; Hull, C.P.; Kirkby, N.F. Enhancement of catalytic reaction by pressure swing adsorption. Catal. Today **1994**, *20*, 351.
31. Alpay, E.; Chatsiriwech, D.; Kershenbaum, L.S.; Hull, C.P.; Kirkby, N.F. Combined reaction and separation in pressure swing processes. Chem. Eng. Sci. **1994**, *49*, 5845.
32. Sheikh, J.; Kershenbaum, L.S.; Alpay, E. 1-Butene dehydrogenation in rapid pressure swing reaction processes. Chem. Eng. Sci. **2001**, *56*, 1511.
33. Goto, S.T.; Tagawa, T.; Omiya, T. Dehydrogenation of cyclohexane in a PSA reactor using hydrogen occlusion. Chem. Eng. Essays (Jpn.) **1993**, *19* (6), 978.
34. Han, C.; Harrison, D.P. Simultaneous shift reaction and carbon dioxide separation for the

production of hydrogen. Chem. Eng. Sci. **1994**, *49*, 5875.
35. Lu, Z.P.; Rodrigues, A.E. Pressure swing adsorption reactors: simulation of three-step one-bed process. AIChE J. **1994**, *40* (7), 1118.
36. Xiu, G.; Soares, J.L.; Li, P.; Rodrigues, A.E. Simulation of five-step one-bed sorption-enhanced reaction process. AIChE J. **2002**, *48* (12), 2817–2832.
37. Sircar, S.; Carvill, B.T.; Hufton, J.R.; Anand, M. Sorption enhanced reaction process. AIChE J. **1996**, *42*, 2765.
38. Hufton, J.R.; Mayorga, S.; Sircar, S. Sorption-enhanced reaction process for hydrogen production. AIChE J. **1999**, *45* (2), 248.
39. Ruthven, D.M.; Farooq, S.; Knaebel, K. *Pressure Swing Adsorption*; VCH: New York, 1994.
40. Gomes, V.G.; Fuller, O.M. Dynamics of propene metathesis: physisorption and diffusion in heterogeneous catalysis. AIChE J. **1996**, *42* (1), 204–214.
41. Gomes, V.G.; Yee, K.W. A periodic separating reactor for propene metathesis. Chem. Eng. Sci. **2002**, *57* (18), 3839.
42. Westerterp, K.R.; Bodewes, T.N.; Vrijland, M.S.; Kuczynski, M. Two new methanol converters. Hydrocarbon Proc. **1988**, *67*, 69–73.
43. Ettouney, H.M.; Masiar, B.; Bouhamra, S.; Hughes, R. High temperature CO shift conversion using catalytic membrane reactors. Chem. Eng. Res. Des. **1996**, *74*, 649.
44. Balachandran, U.; Dusek, J.T.; Maiya, P.S.; Ma, B.; Mierill, R.L.; Kleefisil, M.S.; Udovich, C.A. Ceramic membrane reactor for converting methane to syngas. Catal. Today **1997**, *36*, 265.
45. Sirkar, K.K.; Shanbhag, P.V.; Kovvali, A.S. Membrane in a reactor: a functional perspective. Ind. Eng. Chem. Res. **1999**, *38*, 3715–3737.
46. Cauwenberg, V.; Vergossen, P.; Stankiewicz, A.; Kierkels, H. Integration of reaction and separation in manufacturing of pharmaceuticals: membrane-mediate production of S-ibuprofen. Chem. Eng. Sci. **1999**, *54*, 1473–1477.
47. Minotti, M.; Doherty, M.F.; Malone, M.F. Design for simultaneous reaction and liquid-liquid extraction. Ind. Eng. Chem. Res. **1998**, *37*, 4746–4755.
48. Samant, K.D.; Ng, K.M. Systematic development of extractive reaction process. Chem. Eng. Technol. **1999**, *22*, 877–880.
49. Kelkar, V.V.; Ng, K.M. Design of reactive crystallization systems incorporating kinetics and mass-transfer effects. AIChE J. **1999**, *45*, 69–81.
50. Chen, J.; Wang, Y.; Zheng, C. Synthesis of nanoparticles of $CaCO_3$ in a novel reactor. In *BHR Group Conf. Series,* Publ. No. 28; Semel, J., Ed.; Second International Conference on Process Intensification in Practice; Mechanical Engineering Publications Limited: London, 1997; 157–164.
51. Bisschops, M.A.T.; Van Hateren, S.H.; Luyben, K.Ch.A.M.; Van der Wielen, L.A.M. Mass transfer performance of centrifugal adsorption technology. Ind. Eng. Chem. Res. **2000**, *39*, 4376–4382.
52. Stitt, E.H. Multifunctional reactors? 'Up to a point lord copper'. Trans. IChemE Pt. A Chem. Eng. Res. Des. **2004**, *82* (A2), 129–139.

Reactor Engineering

Ken K. Robinson
Mega-Carbon Company, St. Charles, Illinois, U.S.A.

INTRODUCTION

Reactor and reaction engineering play a vital role in petroleum and chemical processing. The aim of this article is to acquaint the reader with the interaction between reactor design/selection and the characteristics of the chemical reaction of interest. Reactor selection and design are the basis of economical and safe operation. Chemical reactions in petroleum refining include a huge spectrum of unique properties. This includes how the reactants are contacted, whether a catalyst is used, how much heat is evolved or absorbed, and how fast the reaction takes place. This article guides the reader in selecting and designing reactors that will best carry out the reactions of interest. The reactor types discussed focus on those in a petroleum refinery, but many can be used in chemical processing as well.

This article first describes the ideal reactor types, namely batch, plug flow, and completely mixed reactors. Then, the petroleum reactors are discussed based on whether the reaction occurs in the vapor, liquid, or mixed vapor–liquid phase. More specifically, the naphtha-processing reactors are examined first, then gradually moving to heavier hydrocarbons, like kerosene and distillate, that react partially in the liquid and gas phases, and finally ending with a discussion on reactors processing heavy hydrocarbons like petroleum residuum, which reacts completely in the liquid phase.

THE IMPORTANCE OF REACTOR ENGINEERING

Commercial refineries represent huge investments in capital. Small errors in equipment sizing or yield translate to millions of dollars in unnecessary expense. Hence, it is extremely important to do the best job possible on sizing, modeling, and specifying reaction conditions for petroleum refining reactors. We need to have a clear understanding of reactors in the three stages of development. At the earliest stage, the laboratory reactors explore new reaction conditions, catalyst formulations, or feedstock types. Reaction kinetics may also be studied at this stage and reaction rate models developed, making sure that there are no confounding factors of the experimental test that will not scale appropriately to the commercial reactor.

The second stage of development is typically aimed at mimicking commercial operations by employing recycle streams to achieve realistic simulations of the integrated process. Isothermal conditions are usually maintained in the reactor, but if heat release is a concern, such as residuum hydrotreating, then it is wise to run adiabatically so that the adiabatic reaction temperature can be established and also how much heat must be removed in the final commercial design. Defining catalyst deactivation, yield patterns, and how various feed types influence the process are typical aspects to explore.

Finally, we move to commercial scale and adiabatic operation. We need to design the commercial reactor so that it is sized properly, can be started up and shut down safely, and can be operated confidently under steady-state conditions. It is a formidable problem for reaction engineers, but if they are careful and rigorous, the end product will be a success.

The flow diagram of a gasoline-orientated refinery is shown in Fig. 1. Most important, we look at the numerous reactors, which can convert feedstocks, sometimes catalytically and at other times with thermal processing to more valuable products. The reactors are summarized in Table 1 with some of the general characteristics.

The large spectrum of reactors is shown in Fig. 2 with the vertical axis showing the progression from the simplest types such as a delayed coker (a semibatch reactor) to the highly complex fluid catalytic cracking (FCC) unit, which has both the reaction phase and the catalyst being transported through the reactor.

REACTOR TYPES/MODELS

Ideal Reactors—A Brief Review

Many types of reactors have entered the field of petroleum refining, but they can be roughly divided into three types: 1) batch; 2) continuous stirred tank reactor (CSTR); and 3) continuous plug flow. In small-scale studies, the researcher may use a simple pipe reactor, which is operated batchwise. The CSTR reactor is used in small-scale studies for kinetic studies

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120040388

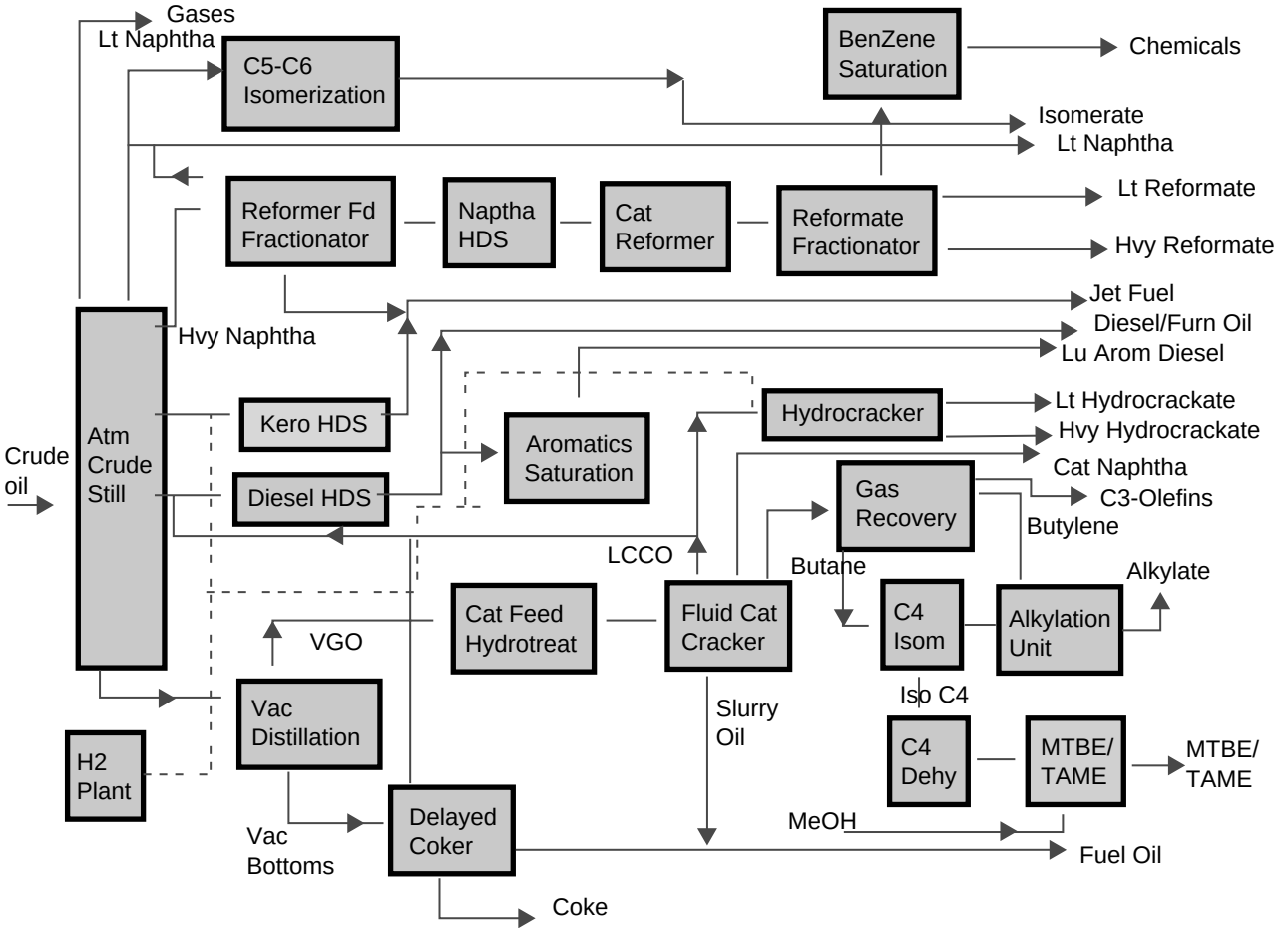


Fig. 1 Flow diagram of a gasoline-oriented refinery. *(View this art in color at www.dekker.com.)*

because the reaction rate is derived directly from the inlet and out concentrations, and it may simulate operation in a larger commercial reactor such as an ebullated bed where the high recycle rate approximates complete mixing. For continuous processing, almost any petroleum fraction may be fed over a fixed bed of catalyst in a plug flow reactor, with vapor phase operation for naphthas and trickle phase for distillates and residuum. The reactor performance equations for these three ideal reactors are given later.

A microbatch reactor (−5 ml) such as the tubing bomb reactor is a common, inexpensive device to develop data. The reactants and, optionally, the catalyst are changed in the small reactor, sealed, and then pressured. To start the reaction, the tubing bomb is typically immersed in a heated fluidized sand bath for a specified length of time with agitation. Shortly after immersion in the heated sand bath, the reactor pressure is increased to a final level, close to commercial conditions. To stop the reaction, the microreactor is pulled out of the heated bath and rapidly quenched in a cooling fluid. The defining equations are given in Eq. (1).

Batch reactor equations

$$t = C_{A_o}\int_o^x \frac{dx}{-r_A} \quad t = \frac{N_{A_o}}{W}\int_o^x \frac{dx}{-r'_A} \tag{1}$$

(thermal) (catalytic)

where

t = time

C_{A_o} = feed concentration of A

x = fractional conversion of A

r_A = reaction rate (mol/hr/volume of the reactor), $\frac{1}{V}\frac{dN_A}{dt}$

r'_A = reaction rate (mol/hr/weight of the catalyst), $\frac{1}{W}\frac{dN_A}{dt}$

W = catalyst weight

N_{A_o} = molar feed rate of A.

The CSTR operates continuously and is frequently a better tool to obtain kinetic data, if one can afford

Table 1 Types of reactors in a petroleum refinery

Reactor(s)	Type	Purpose
Naphtha hydrotreater	Vapor phase catalytic	Remove S and N from catalytic and virgin naphtha before catalytic reforming
Catalytic reformer	Vapor phase catalytic	Convert paraffins to higher-octane aromatics and isoparaffins
Alkylation unit	Catalytic liquid phase (H_2SO_4 or HF)	Combine isoparaffins with olefins to gasoline
Distillate hydrotreating cat-feed hydrotreater	Trickle phase catalytic	Remove S and N and saturate aromatics
Hydrocracker	Trickle phase catalytic	Convert gas oils, coker gas oil, and light catalytic cycle oil to lighter products
Fluid catalytic cracker unit	Vapor phase catalytic	Convert vacuum gas oil to catalytic naphtha
Coker	Semibatch thermal	Convert residuum to gas oil and coke
Residuum hydrotreater	Trickle phase catalytic/thermal	Convert heavy residuum to lighter distillates, removing metals (Ni, V), S, and N

the time and expense of setting one up. The reaction rate can be derived directly from experimental data as pointed out by Mahoney.[1] The reactor has no concentration or temperature gradients, and conversion is controlled by changes in feed rate or reactor temperature. A unique aspect of petroleum refining, relative to chemical processing, is that the streams are mixtures of thousands of compounds. Thus, it is not easy to work with moles as is common for a simple feed. We must, therefore, frequently use mass rather than moles for the reactor design equations. A common convention is to replace space time (J = reactor volume/volumetric feed rate) with reciprocal space velocity, using 1/WHSV for mass of catalyst/time/mass of feed or 1/LHSV for volume of catalyst/time/volume of feed. The reactor performance equation for a CSTR is given in Eq. (2).

CSTR reactor

$$\frac{V}{Q} = \tau = \frac{C_{A_o} X}{r_A} \quad \frac{W}{F_{A_o}} = \frac{X}{-r'_A} \quad \frac{1}{\text{WHSV}} = \frac{X}{-r'''_A} \tag{2}$$

(thermal) (catalytic)

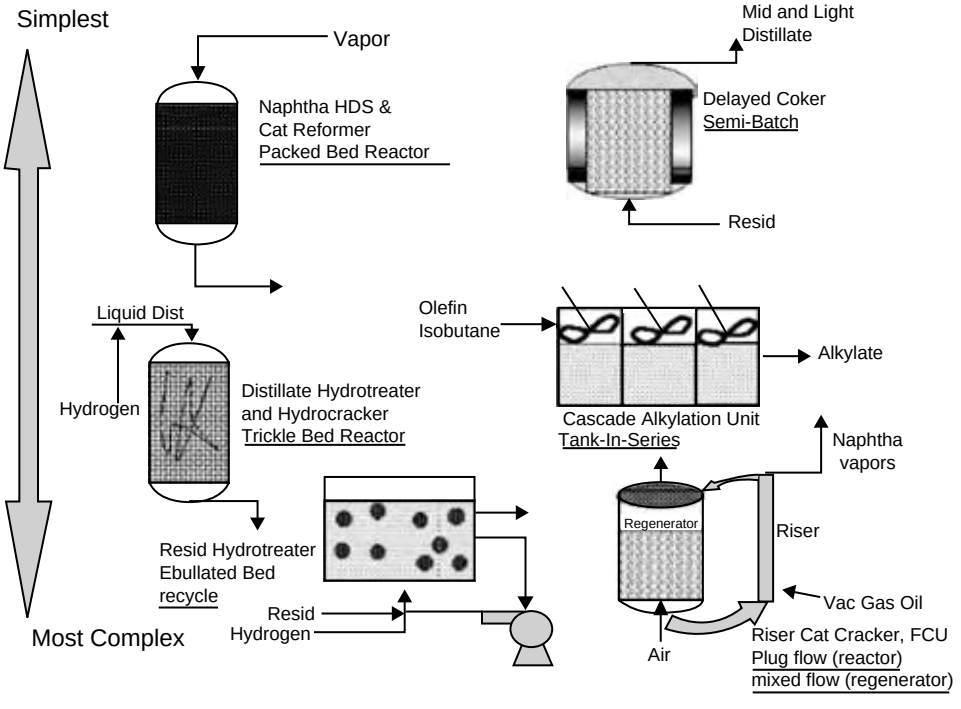


Fig. 2 Petroleum refining reactors. *(View this art in color at www.dekker.com.)*

where

τ = V/Q, volume of the reactor/volumetric flow rate
F_{A_o} = molar feed rate of A
r'''_A = $g/hr/g$ catalyst
WHSV = weight hourly space velocity (g total feed/hr/g catalyst).

The plug flow reactor is probably the most commonly used reactor in catalyst evaluation because it is simply a tube filled with catalyst that reactants are fed into. However, for catalyst evaluation, it is difficult to measure the reaction rate because concentration changes along the axis, and there are frequently temperature gradients, too. Furthermore, because the fluid velocity next to the catalyst is low, the chance for mass transfer limitations through the film around the catalyst is high. Eq. (3) is the reactor performance equation for a plug flow reactor.

Integral plug flow

$$\frac{V}{Q} = \tau = C_{A_o}\int \frac{dx}{-r_A} \quad \text{(thermal)}$$

or

$$\frac{W}{F_{A_o}} = \int \frac{dx}{-r'_A} \quad \frac{1}{\text{WHSV}} = \int \frac{dx}{-r'''_A} \quad \text{(catalyst)} \quad (3)$$

Commercial Reactor Design and Scale-up

The design of larger commercial reactors provides a significant challenge because heat effects are typically substantial and vary with the endothermic cat-cracking or reforming reactions to the highly exothermic hydrotreating and hydrocracking reactions; the flow regime deviates from the ideals of plug flow and perfect mixing. We examine commercial reactors in increasing order of complexity.

Heat Aspects

An energy balance is given below in Eq. (4) and is shown graphically in Fig. 3 for the various types of refining reactions, with the graph developed by Levenspiel.[2]

Energy balance for reactor

$$X = \frac{C_p \Delta T}{-\Delta H_r} \quad (4)$$

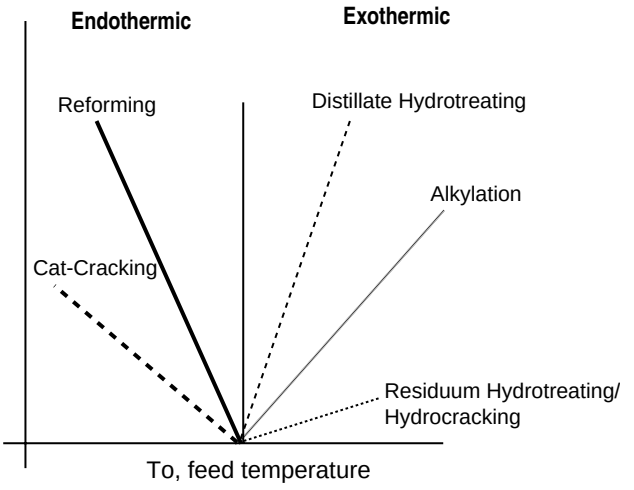


Fig. 3 Heat effects.

where

C_p = molar heat capacity of stream (BTU/lb/mol/°F)
ΔT = temperature difference, outlet minus inlet
$-\Delta H_r$ = heat of reaction (BTU/lb/mol).

Endothermic Reactions (Cat-Cracking, Reforming, Coking). In Fig. 3, we see that the fluid cracking unit (FCU) requires heat input to maintain the reaction temperature in the cracking zone and is shown on the far right in the endothermic region. Burning coke off the catalyst in the regenerator provides this heat and the recirculating catalyst transfers that energy to the cracking reaction in the riser of the FCU. Cat-cracking involves bond breakage and may be classified as endothermic. The second reaction that is highly endothermic is catalytic reforming. The dehydrogenation reaction is very endothermic and requires a reactor system of three to four reactors in series, with interstage heating between the reactors because the reaction temperature drop in each stage must be increased so that the reaction rate does not slow down too much.

Thermally Neutral (Isomerization). Isomerization, represented by the vertical line in the diagram, involves skeletal rearrangement of molecules but no change in the molecular weight. Thus, the reaction does not cause any cooling or heating of the feed stream.

Exothermic (Hydrotreating, Hydrocracking, Alkylation). Many of the petroleum refining reactions are exothermic. Hydrocracking is very exothermic

owing to aromatic saturation reactions. Although the molecular weight is reduced by the cracking reaction, this is preceded by hydrogenation (HDN) reactions, for example, aromatic ring saturation, which is necessary before the ring opening can occur. Alkylation is also quite exothermic because higher molecular weight compounds are formed from isobutane and olefins. Distillate and naphtha hydrotreating also release heat when organo-sulfur and nitrogen compounds (i.e., dibenzothiophene and pyridine) are converted to hydrogen sulfide and ammonia, respectively.

Naphtha/light hydrocarbon processing

Catalytic Reforming of Naphtha. Catalytic reformers may be differentiated as to mechanical design. Those that do not employ on-stream regeneration of the catalyst are fixed-bed processes and are called semiregenerative. The regenerative or cyclic processes include the original hydroforming process (Exxon), ultraforming (Amoco), and power forming (Exxon); they are characterized by a ''swing'' reactor in the multiple reactor train that is substituted for the reactor, which is off-line during regeneration (see Fig. 4).[3,4] The fluid hydroforming employs fluidized bed and the thermofor and hyperforming processes employ moving beds of catalyst. Universal Oil Products has developed a unique reforming process, called CCR for continuous catalyst regeneration, which integrates the reactor with the regeneration process in a stacked reactor concept. It makes use of a radial flow reactor with a moving bed of catalyst. The CCR platforming process (see Fig. 5) utilizes naphtha boiling in the range of 180–400 °F to produce high-octane gasoline or petrochemical precursors. This technology is one of the world's leading reforming processes, with many units operating currently. The platforming process provides refiners with proven, ultra-low-pressure (50 psig reactor pressure) operation and the highest reforming yields. In gasoline applications, the platforming process produces high-octane reformate for unleaded gasoline blending. In reformulated gasoline applications, the low-pressure, low-severity platforming process restores a refinery's hydrogen balance by maximizing the yield of hydrogen, even at the required low-octane severities. With appropriate prefractionation, the process produces the low-benzene, low-vapor-pressure material mandated by reformulated gasoline. The principal problem in the design of catalytic reformers is heat balance, with many of the reactions that produce aromatics being very endothermic. Cyclic mode is preferred for high aromatics production, while semiregenerative for moderate to high severity. Low-octane naphtha is converted to high-octane reformate by dehydrogenation and dehydrocyclization of paraffins to make an aromatic-rich product. Fig. 4 shows cyclic reforming where any reactor can be isolated, regenerated, and placed back on-stream. Typical yields and process conditions for these two processes are listed in Tables 2 and 3.

Catalysts

Reforming catalyst was developed essentially for the production of high-octane blending components from low-octane naphthas. Hydrogenation–dehydrogenation and acidic catalytic components are required for the various reactions that produce high-octane materials. These components are supported on a suitable base. Reforming catalyst has generally used platinum for the HDN–dehydrogenation function and chloride active sites for the acidic function, both supported on gamma alumina. The desirable reactions for octane production are listed below:

1. Dehydrogenation of cyclohexanes to aromatics.
2. Dehydroisomerization of cyclopentanes to aromatics.

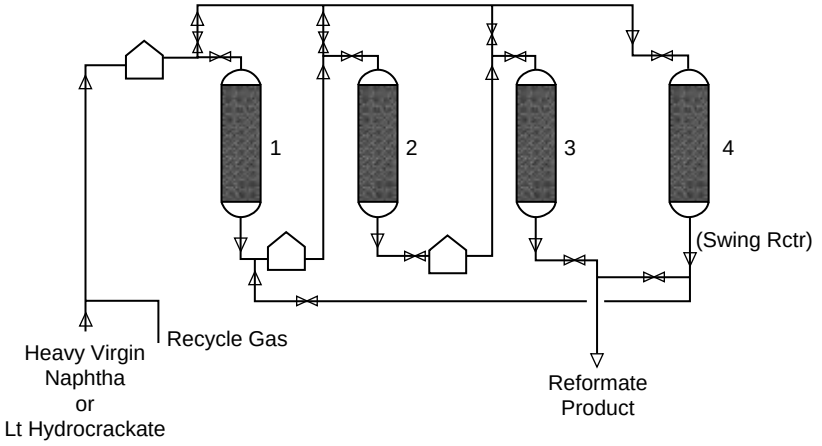


Fig. 4 Catalytic reforming unit (semiregenerative or cyclic).

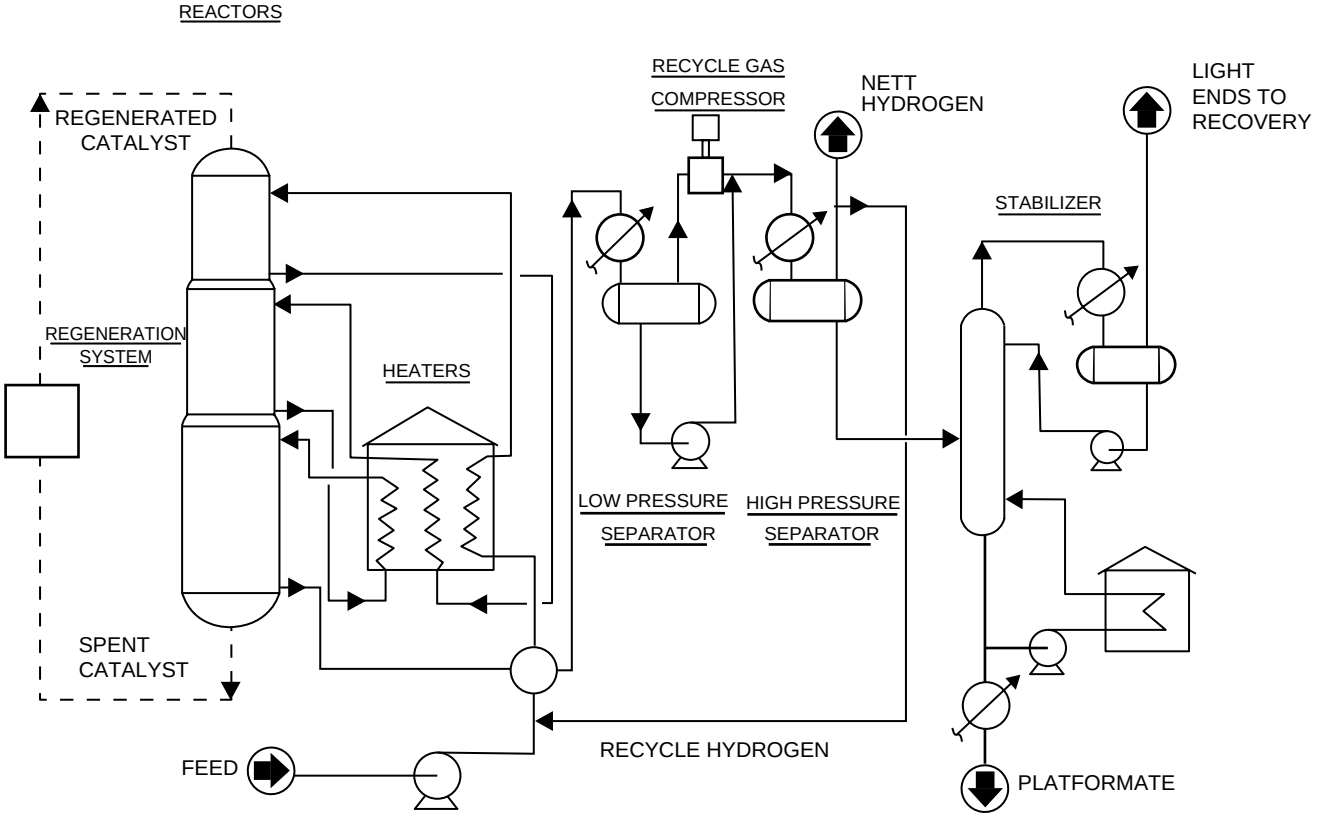


Fig. 5 CCR reforming process.

3. Dehydrocyclization of paraffins to aromatics.
4. Isomerization of *n*-paraffins to isoparaffins.
5. Hydrocracking of low-octane, long-chain paraffins.

Reformer reactor variables

Temperature. Normally, inlet temperatures range between 850 and 1000 °F. The temperature at which the catalyst beds are held is the primary variable available to the refiner to control product quality. Very high temperatures, above 1000 °F, can cause thermal reactions that will decrease reformate yields and increase catalyst deactivation from coke formation.

Space Velocity. Space velocity is defined as volume of naphtha processed per hour per volume of catalyst (or weight of naphtha per hour per weight of catalyst). It determines the limits of product quality (i.e., octane number). The greater the space velocity, the lower the limit, or maximum octane possible. For highly naphthenic feedstocks, high space velocity can be used. For more paraffinic feedstocks, lower space velocity is required to achieve the desired octane number in the product.

Pressure. Pressure affects dehydrogenation and hydrocracking reactions. Increasing the pressure will increase hydrocracking but adversely affect equilibrium aromatic

Table 2 Yields for semiregenerative and cyclic reforming

	Yield	Semiregenerative	Cyclic
57 °API naphtha 30 vol% cycloparaffins 57 vol% paraffins 53 Research octane No. (RON)	H_2 (%)	2.3	2.6
	C1–C4 (%)	13.1	11.2
	C5+ (%)	78.5	79.1
	RON	99	101

Table 3 Process conditions for semiregenerative and cyclic reforming

Process variable	Semiregenerative mode	Cyclic mode
Pressure (psig)	150 (low severity)	350
Temperature (°F)	960	960
Feed rate, WHSV	5.5	5.5
Catalyst	Pt/Re (0.4/0.4%)	Pt/Re (0.4/0.4%)
Octane, $(R + M)/2$	93	98

formation. Higher pressure reduces carbon deposition on the catalyst and slows down deactivation.

Reformer reactor design aspects

The principal problem in the design of catalytic reformers is heat balance. Reactions that produce aromatics are very endothermic, being partially offset by exothermic hydrocracking reactions. Large amounts of heat must be supplied to the reaction zone to keep the temperature high enough. In fixed-bed units the heat is supplied by reactors in series with intermediate reheating. The largest temperature drop is in the first reactor.

Reactor models

The plug flow model is frequently used to describe catalytic reformers. Referring to Eq. (5), we see that one way to calculate the amount of catalyst needed for a specified conversion level and naphtha feed rate is to graphically integrate the expression on the right-hand side of the equation. We do this by plotting 1/rate vs. the fractional conversion, and then compute the area. This area yields a value for 1/WHSV, the ratio of the catalyst charge to the oil flow rate. For a multiple-stage reactor, with an endothermic reaction, we have the plots shown in Fig. 6 for three reactors in series. As the temperature decreases in the reactor, the rate is correspondingly lowered and causes the 1/rate curves to increase as conversion, x, increases. The interstage heaters, between each reactor, bring the naphtha temperature up for the next reactor stage, so that the reaction rate is maintained high enough. The size of each stage can be computed from the area and the total area representing either W/Fao or 1/WHSV (reciprocal space velocity).

Reactor performance equations for a plug flow reactor

$$\frac{\mathrm{d}x}{\mathrm{d}(W/\mathrm{Fao})} = -r'_{\mathrm{A}} \quad \text{or} \quad \frac{\mathrm{d}x}{\mathrm{d}(1/\mathrm{WHSV})} = -r''_{\mathrm{A}}$$
$$W/\mathrm{Fao} = \int \frac{\mathrm{d}x}{-r'_{\mathrm{A}}} \quad \text{or} \quad (1/\mathrm{WHSV}) = \int \frac{\mathrm{d}x}{-r''_{\mathrm{A}}} \qquad (5)$$

where

x = fractional conversion
W = catalyst charge (lb)
Fao = molar feed rate (mol/hr)
WHSV = weight hourly space velocity (lb oil/hr/lb catalyst)
r'_{A} = reaction rate (mol/hr/lb catalyst)
r''_{A} = reaction rate (lb/hr/lb catalyst).

The energy balance equations for a plug flow reactor are given below in Eq. (6) and relate temperature change to either reactor length, z, or reciprocal space velocity, 1/WHSV.

Energy balance equations for plug flow reactor

$$\frac{\mathrm{d}T}{\mathrm{d}z} = \frac{\rho_{\mathrm{b}} \sum(-\Delta H_{\mathrm{r}})_{\mathrm{i}}\, r_i}{GC_{\mathrm{P}}}$$
$$\frac{\mathrm{d}T}{\mathrm{d}(1/\mathrm{WHSV})} = \frac{\sum \Delta H_{\mathrm{r}}\, r_{\mathrm{i}}}{C_{\mathrm{P}}} = \frac{\sum \alpha_{\mathrm{i}}\, \mathrm{d}x}{\mathrm{d}(1/\mathrm{WHSV})} \qquad (6)$$

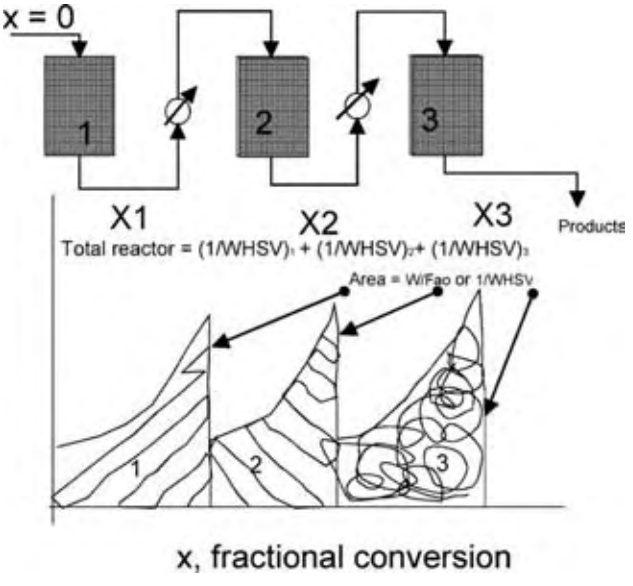


Fig. 6 Plug flow reactors in series.

where

T = temperature (F)
z = reactor length (ft)
ρ_b = bulk density of catalyst
G = mass velocity of feed (lb/hr/ft^2)
C_p = heat capacity of feed stream (BTU/lb)
ΔH_r = heat of reaction (BTU/lb/mol)
r_i = reaction rate of species i
WHSV = space velocity (lb naphtha/hr/lb catalyst)
α_i = adiabatic temperature rise for reaction i
x = fractional conversion of species i.

If the flow regime in the reactor has some mixing and/or dispersion and deviates slightly from plug flow conditions, then the mass balance can account for this with an ''axial dispersion'' model, a one-parameter model.[5] The material balance is shown below in Eq. (7).

Dispersion model for small amounts of mixing

$$\frac{D}{uL}\frac{d^2x}{dZ^2} - \frac{dx}{dZ} + \frac{\tau\rho_b r'_A}{C_{A_o}} = 0 \quad x = 0 \text{ at } Z = 0$$
$$\frac{D}{uL}\frac{d^2x}{dZ^2} - \frac{dx}{dZ} + \frac{\rho_f r'_A}{\text{WHSV } C_{A_o}} = 0 \qquad (7)$$
$$\frac{dx}{dZ} = 0 \text{ at } Z = 1$$

where

x_A = conversion
Z = axial position in reactor, z/L (dimensionless)
u = superficial velocity, Q/A
L = reactor length
(D/uL) = dispersion number, 1/Peclet (reactor)
Peclet = $Pe'(d_p/L)$
d_p = catalyst particle diameter
Pe' = particle Peclet = $20/(Re/\text{Schmidt}) + 1/2$
ρ_f = fluid density
ρ_b = bulk density of catalyst
r'_A = reaction rate of A (mol/hr/ft^{3}reactor)
τ = space time (hr)
C_{A_o} = feed concentration
z = $G/(\rho_b\text{WHSV})$(ft)
G = mass velocity, $\rho_f u$ (lb/hr/ft^2).

The second-order differential equation is solved with a numerical differential equation solver. The dispersion number is estimated by first predicting the particle Peclet number, Pe', from the equation above. Then, the value of the reactor Peclet number, Pe, is predicted from the particle Peclet number Pe' by multiplying by the ratio of the particle diameter, d_p, to the reactor length, L. Pe is the only parameter required to solve the dispersion model equation.

Alkylation

In sulfuric acid alkylation, olefins and isobutane react to form a gasoline blending component at around 45°F.[6] This reaction only occurs in the acid phase. Olefins are extremely soluble in sulfuric acid; isobutane is only slightly soluble. Olefins will polymerize in the absence of isobutane.

The polymerization reaction competes with the alkylation reaction:

Olefin + olefin = polymer
Olefin + isobutene = alkylate

A simplified sketch of an alkylation unit is given in Fig. 7. Olefin and isobutane are charged to a refrigerated stirred reactor. Acid in the reactor effluent is removed in a settler and recycled to the reactor. The most commonly used reactor in sulfuric acid alkylation is the Stratco contactor.[7] The principal advantage of the Stratco contactor is the high isobutane concentration in the reactor effluent. This is achieved by operating the contactor at a pressure sufficient to suppress vaporization of the isobutane refrigerant recycle. The heat of the alkylation reaction is removed indirectly by partially vaporizing the settler effluent through the tube bundle shown in Fig. 8.

The Kellogg cascade autorefrigeration unit, shown in Fig. 9, differs from the Stratco process in that the refrigeration is provided in situ by allowing a proportion of the reactants to vaporize within the reaction zone. In the cascade system, dilute olefin concentrations are obtained by splitting the olefin feed into a number of parallel streams, each of which is fed to a separate reaction compartment containing its own mixer. Isobutane passes through all of the compartments in series, but the olefin-containing feed is divided among the several components. The Kellogg cascade alkylation unit can be approximated by a tank-in-series reactor model.

Any olefin-containing hydrocarbon stream may be used to alkylate isobutane. Butenes are the usual alkylating agents, but propylene is also used, and ethylene and pentenes are employed to a limited extent.

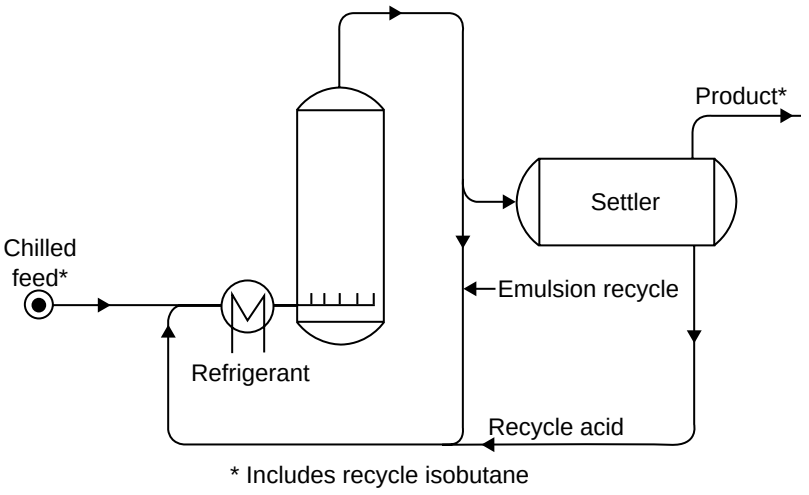


Fig. 7 Simplified sketch of an alkylation unit.

The chief sources of olefins are cracking operations, especially catalytic cracking. However, olefins can be produced by the dehydrogenation of paraffins; butanes are dehydrogenated commercially to provide feeds to alkylation. Isobutane is obtained from crude oils, cracking operations, catalytic reformers, and natural gas. To supplement these sources, *n*-butane is sometimes isomerized. Only small concentrations of diolefins are permissible in feeds to alkylation, particularly for sulfuric acid catalyst. Diolefins increase the consumption of acid.

Although the majority of early alkylation units used sulfuric acid as a catalyst, there has been a trend over recent years in favor of the hydrofluoric (HF) process. Today, the installed capacity of each process is roughly the same. The HF alkylation process is offered by two licensors, Universal Oil Products and Phillips Petroleum. The reaction is carried out at temperatures between 75 and 115 °F, which is considerably higher than the sulfuric acid process.

Rate Models for Alkylation. The rate model to describe alkylation of isobutane with olefins is complicated by the presence of two phases, sulfuric acid and hydrocarbon, and the transfer of reactants between these two phases. As mentioned earlier, the alkylation reaction takes place in the acid phase, with olefins highly soluble in the acid and isobutane only moderately soluble. The isobutane is fed in large excess to compensate for the lower solubility and the olefin concentration must be kept low or it will polymerize and form ''red oil,'' consuming the sulfuric acid catalyst. In Fig. 10, we have shown the film model, with olefins in the bulk gas phase, transferring first through the gas film, then dissolving in the acid, transferring through the liquid film, and finally entering the main body of the bulk liquid acid phase. The rate expression must account for the mass transport across both the gas and the liquid films as well as the alkylation reaction in the acid phase.

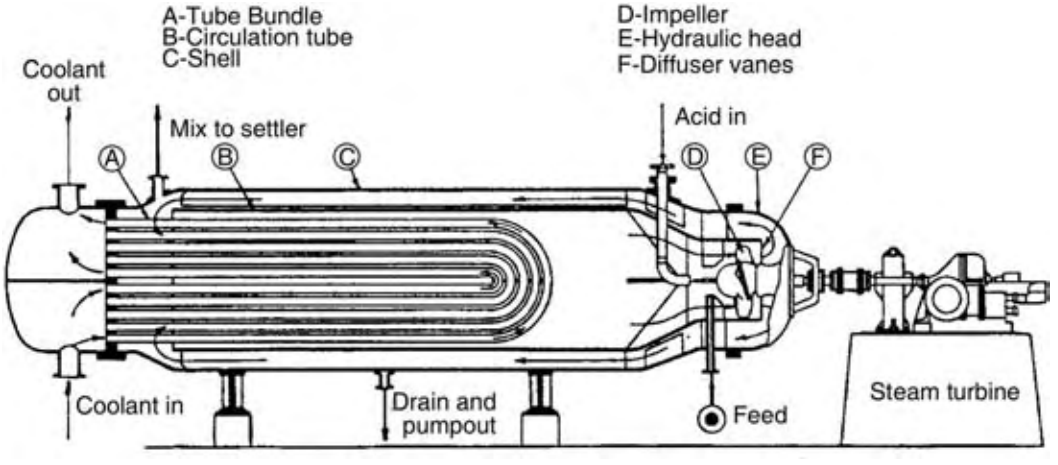


Fig. 8 Stratco contactor for alkylation.

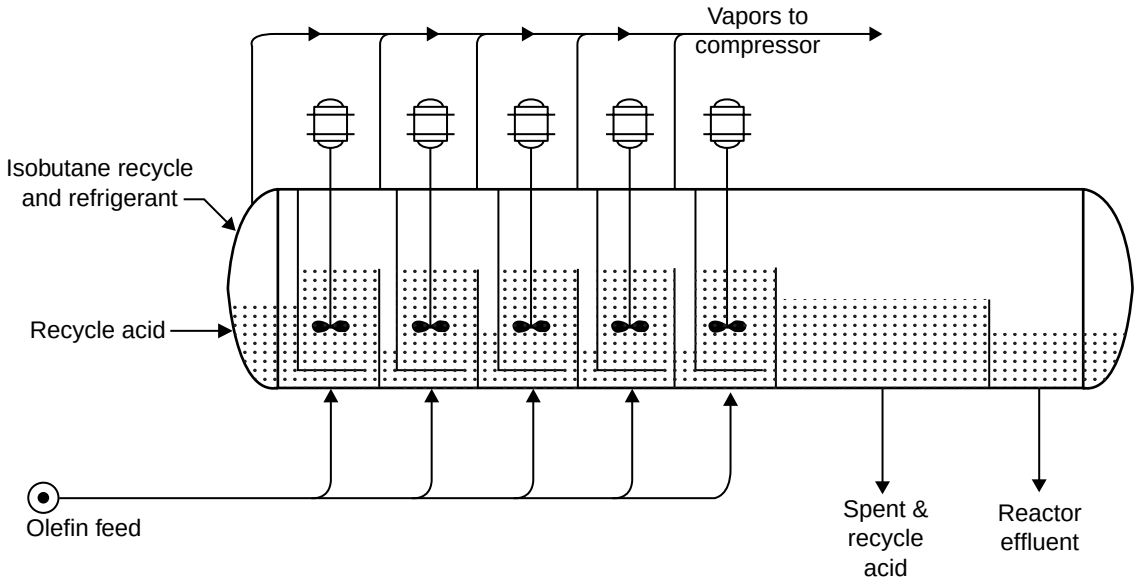


Fig. 9 The Kellogg cascade autorefrigeration unit for alkylation.

Levenspiel has presented a thorough analysis of the many situations involving multiphase kinetics and the equation that applies to alkylation is given below:[8]

$$-r_o''' = \frac{P_o}{1/(k_{og}\,a) + H_o/(k_{ol} \sim aE) + H_o/(k\,C_{iso} f_l)} \quad (8)$$

where

$-r_o'''$ = reaction rate (mol/hr/ft^3 reactor)
P_o = olefin partial pressure (atm)
k_{og} = mass transfer coefficient for gas film (mol/ft/hr/atm)
a = superficial area (ft^2/ft^3 reactor)
H_o = Henry's law constant (atm/ft^3/mol)
k_{ol} = liquid film mass transfer coefficient
E = liquid film enhancement factor,

$$\frac{\text{rate of transfer with chemical reaction}}{\text{rate for straight mass transfer}}$$

k = reaction rate constant (ft^3 liquid/mol/hr)
C_{iso} = liquid concentration (mol/ft^3)
f_l = fraction of liquid phase in the reactor.

Reactor design equations

The contacting pattern in two types of alkylation units is shown in Fig. 11. More specifically, we see that the Stratco unit with the single mixer on one end is approximated by a single mixed tank, as shown in the upper part of the figure. However, the Kellogg cascade unit has a series of compartments with mixers and olefin is sparged into each compartment to keep the concentration low so that it reacts with the isobutane rather than polymerizing. The tank-in-series model may be used to model this type of unit and this is shown in the lower part of the figure. A mass balance can be made for a stirred tank reactor readily because the composition is the same everywhere in the vessel.

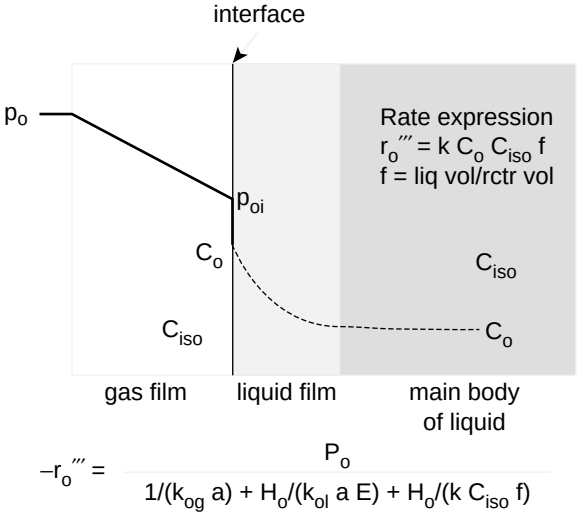


Fig. 10 Transfer of olefin into acid phase, and then reaction with I-C4. *(View this art in color at www.dekker.com.)*

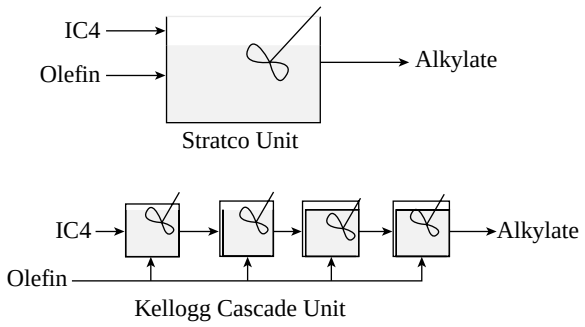


Fig. 11 Contacting patterns for alkylation. *(View this art in color at www.dekker.com.)*

For the case of mass transfer and reaction in an agitated tank contactor, the balance is simply olefin lost by gas = disappearance of olefin by reaction.

Thus, the equation below is the reactor performance equation for a single stirred tank.

Reactor performance for single stirred tank

$$\frac{F_g}{\prod}(P_{o\,in} - P_{o\,out}) = (-r_o''')|_{at\ exit}\ V_r \qquad (9)$$

where

F_g = molar flow rate of the gas
$\prod$ = total pressure
$P_{o\,in}$ = partial pressure of olefin into reactor
$P_{o\,out}$ = partial pressure of olefin leaving reactor
$-r_o'''$ = overall rate of conversion of olefinto alkylate
V_r = total volume of reactor.

To estimate the reactor volume, V_r, the solution is direct. Evaluate the overall rate, r_o''', from known stream compositions. Then, solve for the reactor volume. For several tanks in series, each tank is calculated separately and then the individual tank volumes are summed up to yield the total volume.

DISTILLATE PROCESSING

Hydrotreating

Hydrodesulfurization (HDS), HDN, and HDN of petroleum distillates (i.e., light cycle oil, kerosene, and light gas oil) can be accomplished in a down-flow trickle bed reactor, filled with cobalt–molybdenum (Co–Mo) or nickel–molybdenum (Ni–Mo) catalyst. Hydrocarbon feed and hydrogen-rich gas are mixed, heated, and contacted with the catalyst. The outlet stream is cooled and separated, with hydrogen-rich gas recycled and additional makeup hydrogen blended into the feed stream; the liquid is stripped to remove dissolved hydrogen sulfide and ammonia and then sent to storage. The process flow is shown in Fig. 12 for a distillate hydrotreating unit. In Table 4 are shown the reactor conditions for three streams varying in boiling point (i.e., naphtha, light catalytic cycle oil, and vacuum gas oil) and impurity content.

In petroleum hydrotreating, both vapor phase and trickle bed reactors are used for catalytic hydrotreating of petroleum fractions. Under normal processing conditions, naphthas are completely in the vapor phase while higher boiling fractions are hydrotreated in a mixed-phase trickle bed reactor. These two types of catalytic reactors are discussed extensively.

Fixed-bed catalytic reactors are widely applied to reaction systems in which the reactants are present in a single vapor phase. The scale-up and performance of commercial reactors can be predicted from experiments in small-scale reactors. On the other hand, the mixed-phase trickle bed reactor is considerably more complex to analyze and scale up. The performance of trickle bed reactors is influenced by many factors associated with mixed-phase (gas–liquid–solid) processing. Some of

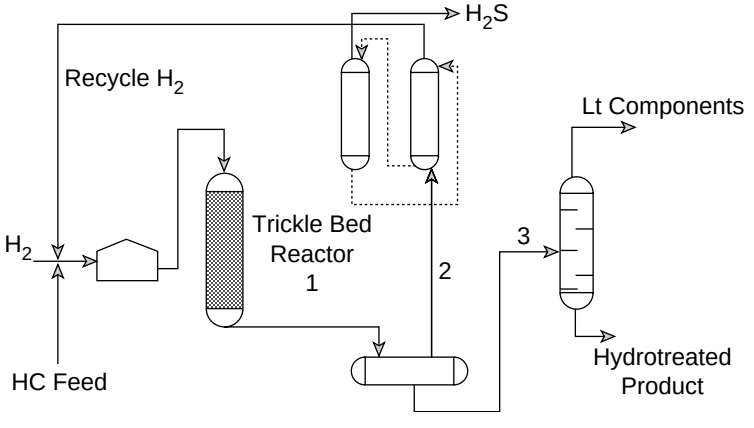


Fig. 12 Distillate hydrotreating.

Table 4 Types of hydrotreaters

	Type of reactor			
	Vapor phase	Trickle bed		
Purpose	**Reformer feed treating**	**Distillate HDS**		**Cat-feed hydrotreat**
Feed	Virgin/cracked naphtha	Light cycle oil, diesel	Kerosene	Vacuum gas oil
Gravity (°API)[a]	62	26	45	17
Boiling range (°F)	175–360	310–660	302–490	650–1000
Sulfur (%)	0.001/0.1	1.4	0.5	1.4
Nitrogen (ppm)	4	400	10	6000
Bromine (No.)	0/19	30	—	—
H_2 consumption (SCFB)[b]	80	350	150	700
Temperature (°F)	550	625	600	700
Pressure (psig)	500	400	350	1000
WHSV (lb/hr/lb catalyst)	6	4	4	1
Catalyst		Co–Mo/alumina		Ni–Mo/alumina
Recycle gas rate (SCFB)	500	700	400	1500

[a]Specific gravity = 141.5/(°API + 131.5).
[b]SCFB = Std. ft^3 per barrel

these include gas–liquid distribution, catalyst contacting, and mass transport resistances in the gas and liquid phases. To appreciate the complexity of this type of reactor, a step-by-step description of the reaction process follows:

1. *Gas–liquid distribution*: Oil and hydrogen are fed into the top of the reactor and distributed uniformly over the catalyst bed. Further, down the catalyst bed, flow maldistribution (i.e., channeling, wall flow) may take place owing to nonuniform bed properties.
2. *Catalyst contacting*: Catalyst particles must be actively contacted with fresh reactant to effectively utilize the catalyst. Stagnant regions on the catalyst particles impair overall reactor performance.
3. *Transport of reactants to catalyst*: To supply the catalyst with fresh reactant, several mass transfer steps are involved. The hydrogen must first transfer through the gas–liquid interface before it is in the bulk liquid phase. Then, dissolved hydrogen and oil reactant (sulfur, nitrogen, aromatics) in the bulk phase transfer through the liquid film around the catalyst to its outer surface. Finally, the reactants diffuse into the catalyst pores.
4. *Catalytic reaction*: The reactant molecules chemisorb on the active sites of the catalyst surface, react, and then desorb off the active site.
5. *Transport of reactants from catalyst*: Reaction products diffuse out of the catalyst pores, then through the film around the catalyst to the bulk phase, and finally the gaseous product (H_2S, NH_3, or H_2O) transfers through the gas–liquid interface into the bulk gas phase.

Heat transfer is also an important consideration in a commercial operation, because reactors are operated adiabatically. The heat release from desulfurization, polyaromatic saturation, etc., produces a temperature gradient along the catalyst bed, the magnitude directly proportional to hydrogen consumption. Temperature control in a multibed commercial reactor frequently employs ''cold-shot cooling'' between the beds with hydrogen to limit the maximum temperature rise. Other cooling means such as interreactor heat exchangers are also used. On the catalyst particle scale, heat transport considerations are usually not important. Interphase (film) and intraparticle temperature gradients are not large because the liquid phase effectively transfers heat from the particle.

General Kinetic Rate Model for Mixed-Phase Catalytic Systems

In general, the rate equation for heterogeneous reactions accounts for more than one rate process. This leads one to ask how processes involving both physical transport and reaction steps can be incorporated into one overall

rate expression. We can make an analogy with conductive heat transfer, in which a combination of different resistances is applied to chemical systems. In Fig. 13, we see that when catalyst particles are immersed in a liquid, there are two film resistances, one at the gas–liquid interface and the other on the catalyst surface, that must be accounted for plus the catalytic surface reaction, which can be influenced by pore diffusion. Thus, we can write the following equation to account for all of these physical transport and reaction steps.

$$H_2 + D \rightarrow \text{Products}$$

$$-r_{H_2} = \frac{P_{H_2}}{H/(k_l a) + H/(k' a_s) + H/(\epsilon k C_D)} \quad (10)$$

where

P_{H_2} = hydrogen pressure
H = Henry's law constant (atm/ft^3/mol)
k_l = liquid film mass transfer coefficient
a = interfacial area between gas and liquid
a_s = catalyst external area
k' = catalyst film mass transfer coefficient
ϵ = catalyst effectiveness factor for pore diffusion effects
k = rate constant for catalytic reaction
C_D = concentration of D in liquid phase.

Plug Flow Reactor Model for Mixed-Phase Reactor

The general reactor model for a trickle bed reactor was derived by Frey and Mosby and appropriately accounts for partial vaporization of the liquid hydrocarbon as it enters the reactor; a plug flow model is used.[9] Some of the hydrocarbon travels downward through the liquid phase, while the portion that is vaporized travels with the hydrogen-rich phase. The reactor model is given below in Eq. (11) and the energy balance equation follows it.

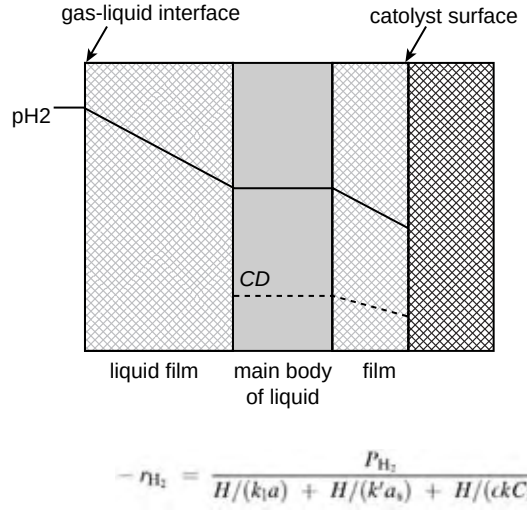


Fig. 13 Multiphase resistances in distillate hydrotreating. *(View this art in color at www.dekker.com.)*

Rate equation

$$\frac{dx_1}{d(1/WHSV)} = \frac{\text{reaction rate of species } i}{P_{io}[(a + H/O)\frac{P_v}{P} + (1 - a)]} \quad (11)$$

where

P_{io} = feed partial pressure of i (atm)
a = fractional vaporization of liquid feed
x_i = fractional conversion of i
H = mol gas/barrel of oil
O = mol oil/barrel of oil
P_v = vapor pressure of oil
P = total pressure
r_i = reaction rate of i (lb/mol/hr/lb catalyst)
WHSV = space velocity (lb oil/hr/lb catalyst).

Energy balance equation

$$\frac{dT}{d(1/WHSV)} = \frac{\sum \alpha_i \, dx_i}{d(1/WHSV)}$$

where α_i = adiabatic temperature rise per unit of conversion of i.

Hydrodesulfurization Kinetics

Reactions for removing sulfur, nitrogen, and oxygen, which are present as organic compounds, are characterized by replacement of the nonhydrocarbon component with hydrogen. The nonhydrocarbon component is hydrogenated (to H_2S, NH_3, and H_2O) and subsequently removed by stripping and fractionation. The desulfurization reaction for single compounds (i.e., benzothiophene and dibenzothiophene) can be modelled with the Frey–Mosby relationship, which has a Langmuir–Hinshelwood form and accounts for inhibition by hydrogen sulfide and aromatics.[10] This works well for single-sulfur compounds, but it should be pointed out that overall desulfurization kinetics, which includes compounds with a large range in reactivity, often behaves more like second order [see Eq. (12)]:

Rate expression

$$-r_s = \frac{k_s \, \text{Act} P_s \, P_{H_2}}{(1 + K_{HS} P_{HS} + K_{AR} P_{AR})^2} \quad (12)$$

Performance equation for plug flow reactor

$$\ln\frac{1}{1-x} = \frac{k_s\,\mathrm{Act}\,P_{H_2}}{\mathrm{WHSV}\left[1-a+(a+H/O)\left(\frac{P_s}{P}\right)\right]\times\left[1+K_{HS}P_{HS}+K_{AR}P_{AR}\right]^2}$$

where Act = catalyst activity.

Denitrogenation Rate Equation

Nitrogen removal is generally more difficult than desulfurization, for several reasons, for instance: 1) side reactions yield nitrogen compounds more difficult to remove than the original ones (indole ⇒ quinoline and carbazole derivatives) and 2) heterocyclic, nitrogen-containing rings must saturate during the hydrogenolysis, and the presence of large side groups on the ring appears to hinder the reaction sterically. Ho has a comprehensive review on hydrodenitrogenation catalysis that covers many topics.[11] The rate of denitrogenation is frequently described as being first order in nitrogen and hydrogen, as discussed by Satterfield and Yang, and the rate equation is given below as Eq. (13):[12]

$$-r_N = K_N\,\mathrm{Act}\,P_N P_{H_2} \tag{13}$$

where

r_N = rate of denitrogenation
Act = catalyst activity
P_N = partial pressure of nitrogen compounds
P_{H_2} = partial pressure of hydrogen.

Nitrogen becomes increasingly harder to remove as the number of aromatic rings increase. More specifically, as we go from single-ring heterocyclic compounds such as pyridine to multiring compounds like quinoline, nitrogen removal is significantly more difficult.

Aromatic Saturation Rate Equation (Equilibrium Limited)

The aromatic saturation rate equation is similar in form to denitrogenation with the exception that it is strongly influenced by chemical equilibrium, instead of the irreversible form. Levenspiel has developed an integrated expression for reversible reactions and this is provided in Eq. (14):[2]

$$-r_A = k_A\,\mathrm{Act}[P_A P_{H_2}^n - (1/K)P_{sat}]$$
$$K_o = \frac{k_A}{k_{sat}} = e^{-\Delta G_o/RT_o} \tag{14}$$
$$\ln(K/K_o) = \frac{-\Delta H_R}{R}\left[\frac{1}{T}-\frac{1}{T_o}\right]$$

where

ΔG_o = free energy of formation at T_o
ΔH_R = heat of reaction (− is exothermic, + is endothermic)
T_o = temperature under standard conditions
T = reaction temperature
P_{sat} = saturated aromatic pressure
P_A = aromatic pressure
k_A = rate constant of reaction
Act = catalyst activity for aromatic saturation
R = gas constant
K = equilibrium constant.

$$\ln\left(\frac{1}{1-X/X_e}\right) = \frac{k_A\mathrm{Act}(M+1)P_{H_2}^n}{\mathrm{WHSV}(M+X_e)\left[1+a+(a+H/O)\frac{P_s}{P}\right]}$$

where

M = ratio of saturates/aromatics in the feed stream and
X_e = equilibrium conversion value.

Hydrogen Consumption Estimates for Hydrotreating

The hydrogen used in hydrotreating can be predicted from the set of equations given in Table 5, which are based on stoichiometry of model reactions. The formulas correspond to 1, 3, and 5 mol of hydrogen per mole of olefin, sulfur, and nitrogen, respectively.

DISTILLATE HYDROCRACKING

Since the mid-1960s, hydrocracking has become a major refining process. It is one of the most versatile of modern petroleum processes. Flexibility of operation, with respect to both feedstock and product, has been reported. This flexibility in operation may be related to the development of specific families of the catalyst, the design of processing schemes that allow the catalysts to function efficiently, and the optimum refining relationships between hydrocracking and other refining processes. Whereas the commercial feedstocks range from naphtha to residua, there is a wide choice for the product of a hydrocracker.

Table 5 Calculation of hydrogen consumption

Reaction	Constant	Feed density (lb/gal)	Feed impurity	Conversion	Hydrogen consumed (SCFB = SCFB1 + SCFB2 + SCFB3)
Desulfurization	0.15	lb/gal	%S	% Desulfurization	SCFB1
Denitrogenation	0.57	lb/gal	%N	% Denitrogenation	SCFB2
Olefin saturation	0.10	lb/gal	Bromine number	% Saturation	SCFB3

SCFB = Std. ft^3 per barrel

The hydrocracking catalyst is dual functional: 1) cracking of high molecular weight hydrocarbons and 2) HDN of the unsaturates either formed during the cracking step or otherwise present in the feedstock. A typical cracking catalyst is silica–alumina and a base or noble metal serves as a hydrogenating catalyst. In a way, HDN helps cracking. The metal HDN sites keep the acid sites of the cracking catalyst clean and active by HDN of the coke precursors. During the past several years, interest in zeolite catalyst has been relatively high. Zeolites X and Y, and mordenite have been the center of attraction. [13] Zeolite, being a superior cracking catalyst compared to amorphous silica–alumina, permits a decrease in reaction temperature for the same level of conversion. The major HDN components are platinum, nickel, palladium, molybdenum, cobalt, etc. These may be altered by promotion with another metal. Inorganic salts, water, metals, and organic compounds of sulfur, nitrogen, or oxygen in petroleum act as poisons for the hydrocracking catalysts. Sulfur compounds inhibit the HDN component while nitrogen compounds inhibit the cracking component of the catalyst. Metal contaminants are deposited on the catalyst to deactivate it. These deposits, when in the active state, promote various dehydrogenation reactions and increase coke-producing tendency.

Types of Hydrocracking

Depending on the feedstock used, two types of hydrocracking are practiced industrially. If the feedstock is a heavy distillate obtained from a straight-run refining or cracking operation, it is called distillate hydrocracking. Residual hydrocracking is the name given to the process if the feedstock happens to be the residue of straight-run refining. The residues are usually lower in API gravity and higher in carbon residues and carbon/hydrogen ratio as compared to distillates. Galbreath and Van Driesen have shown that residual hydrocracking is clearly a different process from distillate hydrocracking; a different type of catalyst is used at a relatively high temperature.[14] We will discuss residual hydrocracking in a later section of this entry. The various distillate-hydrocracking processes include Unicracking, Isocracking, Houdry-Gulf, Isomax, Ultracracking, and BASF-IFP. They can be operated in either a single- or a two-stage process and the product slate can be adjusted by catalyst selection, reaction temperature, and staging. Most of these processes recycle the unconverted higher boiling product to extinction. Unicracking is characterized by the use of a molecular sieve catalyst but many of these other processes are converting to this sieve catalyst technology. Typical conditions for distillate hydrocracking include:

- Pressure: 1000–3000 psig
- Temperature: 575–825 °F
- Liquid space velocity: 0.3–2.0 × volume of feed/volume of catalyst/hr
- Recycle gas rate: 3000 std ft^3 per barrel.

A typical single-stage process is shown in Fig. 14 with the hydrocracking reactor in the middle and a pretreating section on the far left to convert sulfur, nitrogen, and oxygen compounds to hydrogen sulfide, ammonia, and water. The effluent then flows directly to the hydrocracking reactor. Interstage quenching of recycle gas between catalyst beds controls heat release and maintains temperature control.

A diagram of a five-bed hydrocracker is shown in Fig. 15. The dimensions of this particular commercial reactor were 10 ft diameter with 8 in.-thick walls. The design of hydrocracking reactors must be carefully considered on account of the severe operating conditions to which they are subjected, namely:

- High total pressure and hydrogen partial pressure.
- High hydrogen sulfide and ammonia partial pressure from the HDS and HDN reactions.
- High operating temperatures.
- High exothermic heat of reaction.

The heat of reaction is a function of the amount of hydrogen consumed and is normally in the range

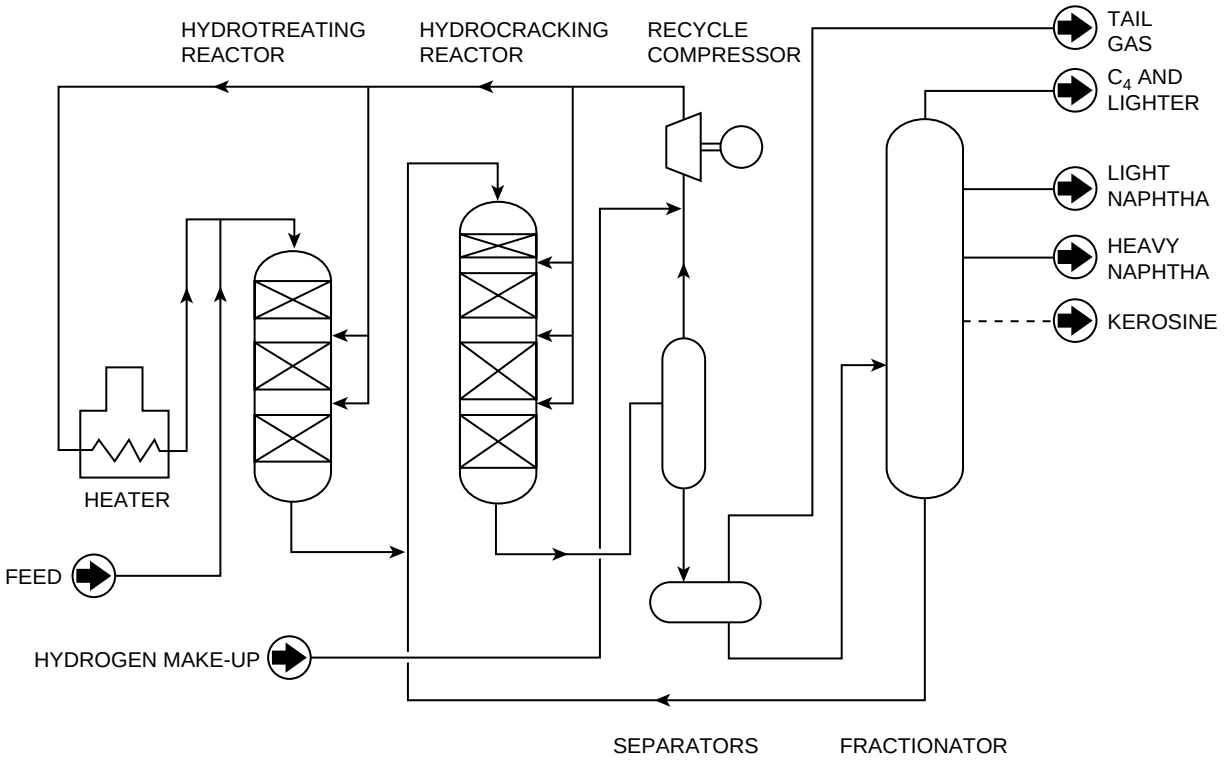


Fig. 14 Hydrocracking process with recycle.

of 490–630 kcal/m^3 of hydrogen. To control the temperature in the reactor, the catalyst charge is split into a number of separate beds with means of cooling the effluent from each bed. This is normally achieved by injection of cold recycle gas between beds or injection of cold recycle liquid between the catalyst beds. Gas injection, the first alternative, is the most favored. The reactor design equations, given for the trickle bed reactor in the previous section on hydrotreating, should be referred to for design of the hydrocracker. Unfortunately, we are not able to provide rate equations to describe the cracking reaction as this information is highly proprietary and is specific to the catalyst and feedstock used.

HEAVY GAS OIL CONVERSION

Fluid Cat Cracking

Fluid catalytic cracking is a process for the conversion of straight-run atmospheric gas oil, vacuum gas oil, and some atmospheric residues and heavy stocks into high-octane gasoline, light fuel oils, and olefin-rich light gases. Basically, the cracking reactions are carried out in a vertical reactor vessel in which vaporized oil rises and carries along with it, in intimate contact, small fluidized catalyst particles (see Fig. 16). The reactions are very rapid and only a few seconds of contact time is necessary for most applications. Simultaneously, with the desired reactions, a carbonaceous material with low H/C ratio, ''coke'' deposits on the catalyst and deactivates it. The spent catalyst and the converted oils are separated and the catalyst passed down-flow to a separate chamber, the regenerator, where the coke is combusted, reactivating the catalyst. The regenerated catalyst is then conveyed down-flow to the bottom of the reactor riser, where the cycle begins again. Control of the regenerator temperature is based on coke made in the reactor.

A major breakthrough in catalyst technology occurred in the mid-1960s with the development of zeolitic catalysts. These sieve catalysts demonstrated superior activity, gasoline selectivity, and stability characteristics compared to silica–alumina catalysts, then in use. The continuing advances in both activity and process design resulted in the advanced design concept of riser cracking in which all of the cracking reaction occurs in the dilute phase in the riser. The key to all-riser cracking is the design of a quenching system that stops the cracking reaction at the optimum yield of desired products. Operating conditions for a typical FCU are shown in Table 6 and how it might be altered is as follows.

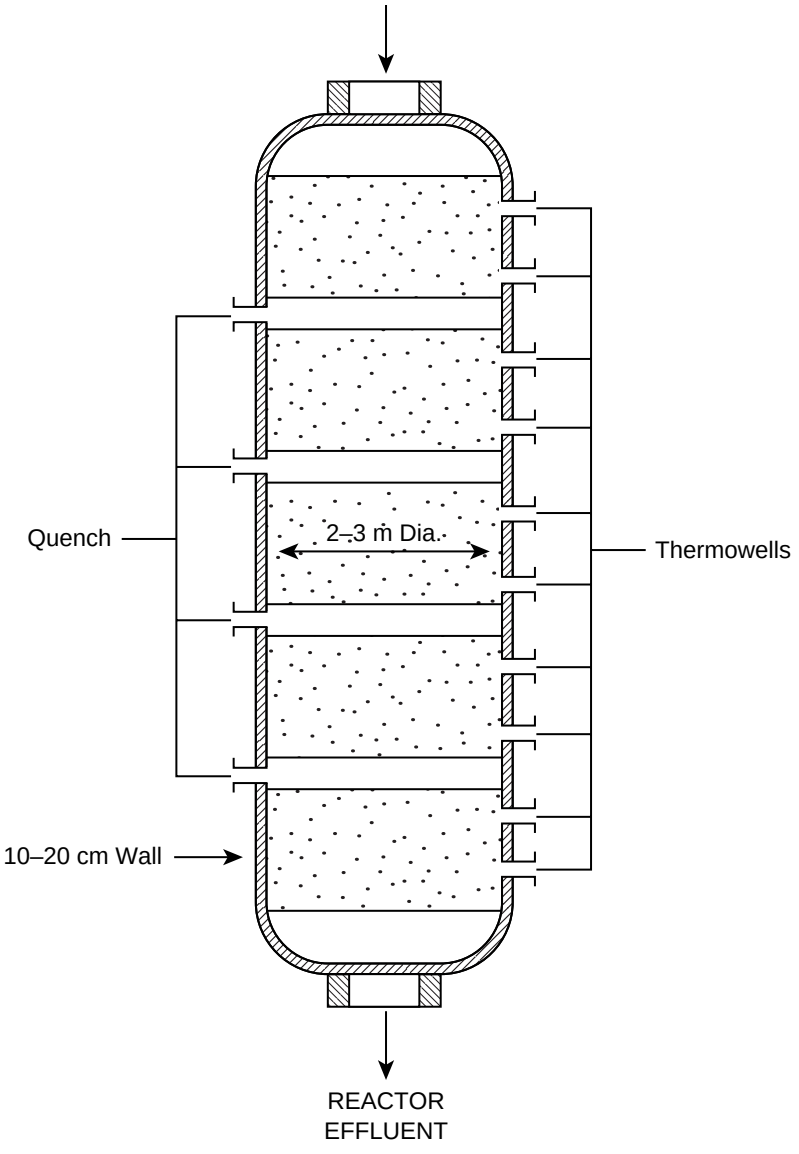


Fig. 15 Five-bed hydrocracker.

Every FCC complex contains the following sections:

1. *Reactor and regenerator*: Reactor cracks feed to light hydrocarbons, gasoline, middle distillates, coke, hydrogen, and hydrogen sulfide. In the regenerator, the circulating catalyst is reactivated by burning off the coke at high temperatures, releasing CO_2, CO, water, and SO_2.
2. *Main fractionation column*: Reactor effluent is separated into various products. The overhead includes gasoline and lighter material. The heavy liquid products, heavier naphtha, and cycle oils are separated as side cuts and slurry oil is separated as a bottom product.
3. *Gas concentration unit*: Unstable gasoline and lighter products from the main fractionation overhead are separated into fuel gas, C1–C4 for alkylation, and debutanized gasoline.

New concepts for FCC

Quick-Quench Cracking. Dilute phase cracking in the riser offers significant advantages over dense-bed cracking. As much as 90% of the feed conversion takes place in the short-contact-time condition in the riser, while the remaining 10% conversion in the dense-bed reactor is accompanied by overcracking of the desirable products. The introduction of the high-activity zeolitic catalysts demonstrated the need for short-contact-time cracking. UOP modified the design of the riser–reactor system. The rise was extended axially well into the reactor shell. The catalyst–hydrocarbon mixture exiting the reactor was directed through a simple device to disengage the hot catalyst from the products. In addition, the catalyst in the dense bed was also reduced. This quenched the cracking reactions at the riser outlet.

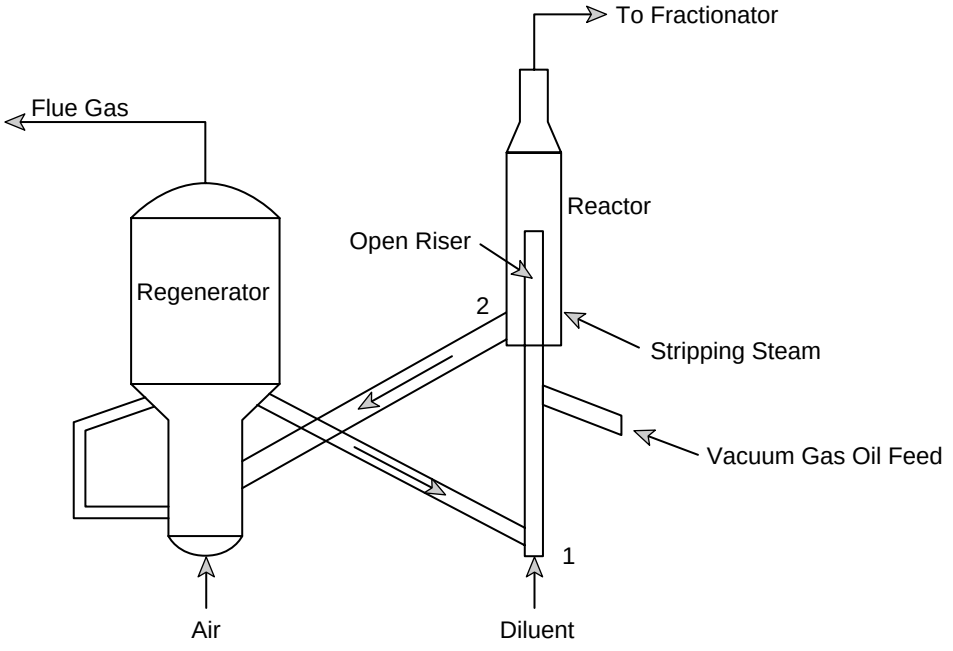


Fig. 16 Fluid cat-cracking unit.

CO Combustion. Complete or partial CO combustion will often benefit FCU operation. In conventional operation, the carbon from the spent catalyst is converted to equimolar quantities of carbon oxides. If the remaining CO is converted to carbon dioxide, the potential heat can be increased by 40–50%. This combustion can be promoted by adding very small amounts of platinum to the cracking catalyst. Complete CO combustion results in the regenerator flue gas containing less than 500 ppm CO, eliminating the need for a CO boiler. The increased heat release produces a higher regenerated catalyst temperature. Therefore, less catalyst recirculation is required to satisfy the reactor heat demand. Coke yield is also decreased and correspondingly is matched on a weight-for-weight basis with an increase in FCC products.

Types of FCU Designs

Model II "down-flow"

The Model II "down-flow" design (Fig. 17) was the first improvement to the Model I "up-flow" unit. The principal features of Model II include a reactor vessel near ground level with the catalyst regenerator offset and above it. A rather short transfer line carries both catalyst and hydrocarbon vapor to a dense-bed reactor. Double slide valves are used at various points in the unit, and the long regenerated-catalyst standpipe causes major operating problems. Commercial evidence indicated that although conversions were relatively low (40–55 vol%), a large portion of the cracking took place in the short transfer line.

Table 6 Process conditions for FCC

Operating variable	Current	Future
Pressure (psig)	20–30	20–30
Reactor temperature (°F)	925	10–25
Regenerator temperature (°F)	1200	1200
Catalyst/oil	12	14
Conversion (%)	80	85
Coke yield (%)	5	4
Catalyst makeup (lb/barrel)	0.15	0.15
Catalyst type	Ultrastable Y sieve	ZSM-5 additive for light olefins + desulfurization additive to capture S as H_2S rather than SO_2

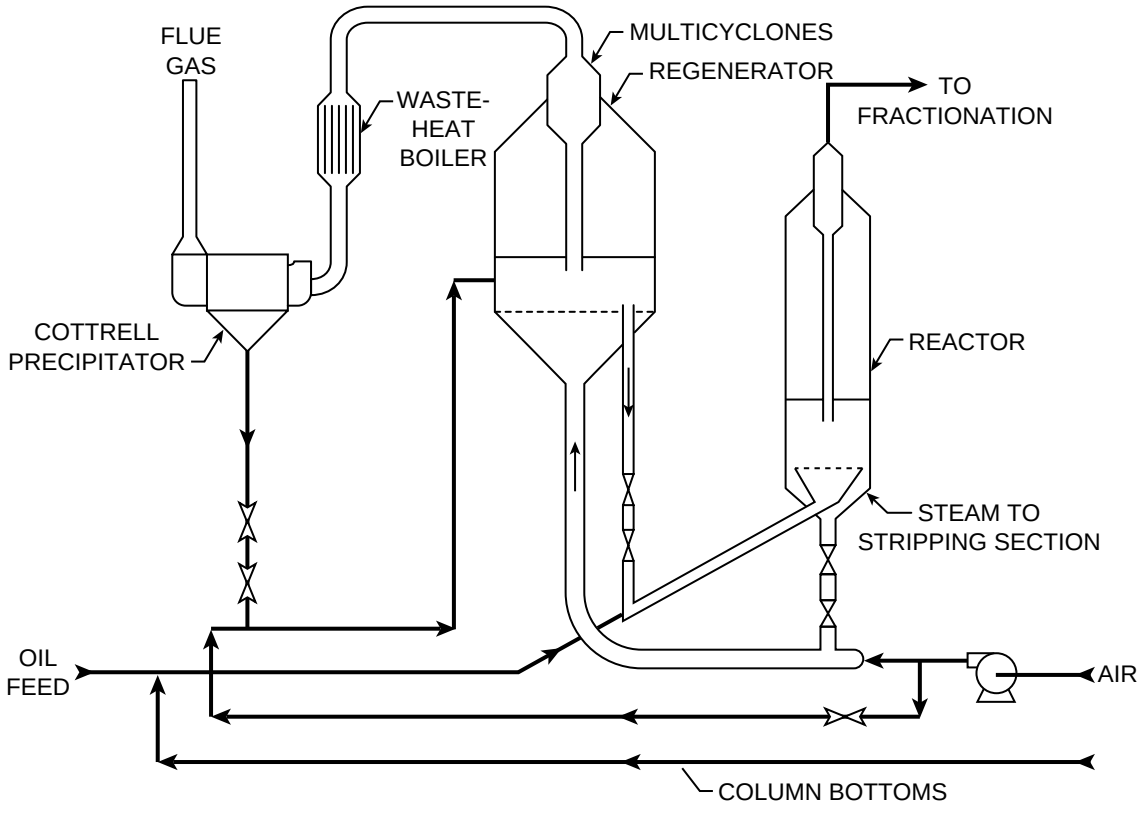


Fig. 17 Model II "down-flow" catalytic cracker.

UOP "stacked" FCU

Following the war, the "stacked" FCU design (Fig. 18), which featured a low-pressure reactor stacked directly above a high-pressure regenerator, was commercialized. The design was a major step toward shifting the cracking reaction from the dense phase of the catalyst to the dilute phase.

UOP "side-by-side" FCU

In the mid-1950s, the "straight riser" or "side-by-side" design (Fig. 19) was introduced. The regenerator was now near ground level with the reactor to the side in an elevated position. Regenerated catalyst, fresh feed, and recycle were directed to the reactor by means of a long, straight riser located directly below the reactor; product yields and selectivity were substantially improved.

Reactor Performance Equations for the "Reactor–Regenerator" System

If the catalyst deactivates rapidly, such as in the FCU, and it is circulated between a regenerator and a reactor, we need to properly account for the catalyst age in the reaction kinetic expression. Some simplifying assumptions are made regarding the catalyst deactivation rate model and the cracking conversion model. However, this allows us to clearly address the coupling of these two processing units and more elegant kinetics can be used for the cracking reactions to upgrade this model, as shown in Eq. 15:

Performance equations for the simplest kinetics reactor (plug flow)–regenerator (mixed flow)

A – R

Cracking reaction (first order):

$$-r'_{\mathrm{A}} = k'aC_{\mathrm{A}} \quad \text{and} \quad R = k'\tau' \tag{15}$$

Deactivation (first order):

$$-\frac{\mathrm{d}a}{\mathrm{d}t} = k_{\mathrm{d}}a \quad \text{and} \quad R_{\mathrm{d}} = k_{\mathrm{d}}\bar{t}_{\mathrm{s1}}$$

Catalyst regeneration:

$$\frac{\mathrm{d}a}{\mathrm{d}t} = k_{\mathrm{r}}(1 - a) \quad \text{and} \quad R_{\mathrm{r}} = k_r\bar{t}_{\mathrm{s2}}$$

Solids in reactor:

$$\overline{a}_1 = \left[R_{\mathrm{r}}\left(1 - \mathrm{e}^{-R_{\mathrm{d}}}\right)\right] / \left[R_{\mathrm{d}}\left(1 + R_{\mathrm{r}} - \mathrm{e}^{-R_{\mathrm{d}}}\right)\right]$$

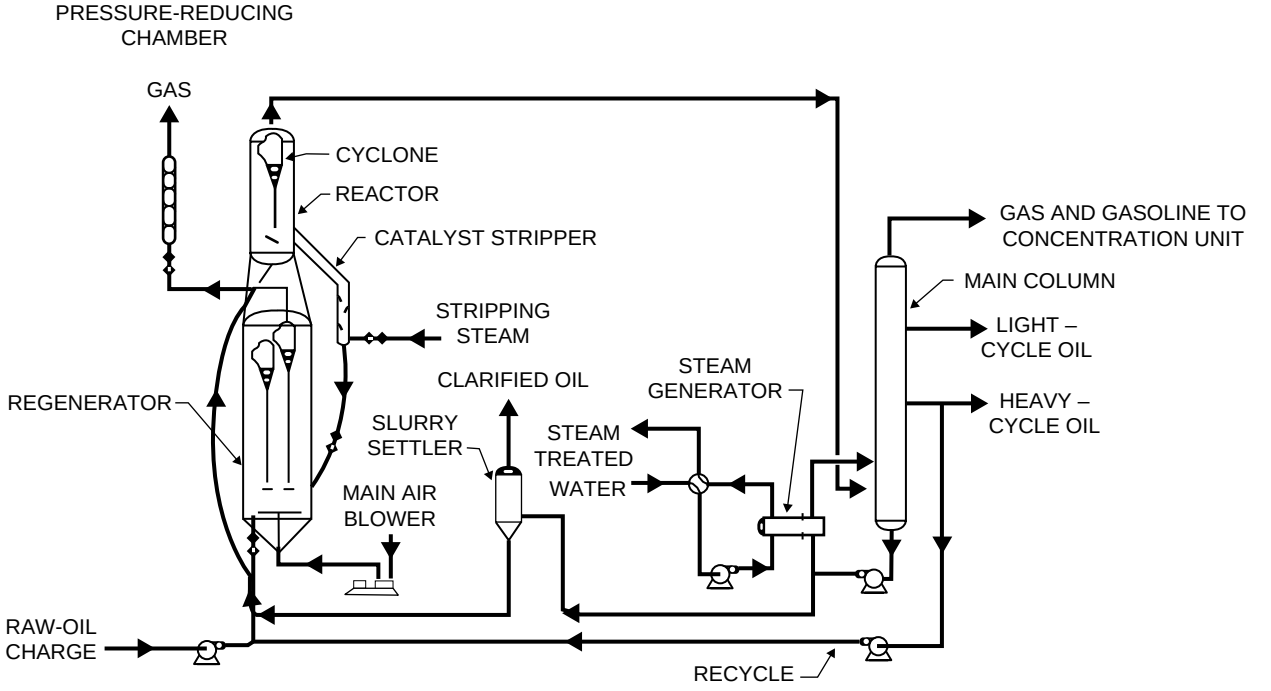


Fig. 18 The ''stacked'' FCU design.

$$\text{Solids in regenerator:} \quad a_2 = R_r/\left[1 + R_r - e^{-R_d}\right]$$

$$\ln\frac{1}{(1-X)} = \frac{RR_r(1 - e^{-R_d})}{R_d(1 + R_r - e^{-R_d})}$$

where

X = fractional conversion of A
a = catalyst activity
k' = rate constant
τ' = space time in reactor $(WC_{A_o})/F_{A_o}$
k_d = deactivation rate constant
$\bar{t}_{s1} = W_1/F_s$ and $\bar{t}_{s2} = W_2/F_s$
W_1 and W_2 = catalyst in reactor and regenerator
F_s = solids circulation rate.

The nomenclature for the above set of equations is given for the reactor–regenerator system in Fig. 20.

RESIDUUM PROCESSING

Residuum Hydroconversion

Residuum hydroconversion is a special case of hydrocracking. Compared to distillate hydrocracking, a significantly different catalyst is used, frequently low metal loading with Mo, with a special pore size distribution less subject to pore plugging by coke. More specifically, a bimodal pore size distribution is frequently used with large macrofeeder (GT. 1000 A°) pores along with smaller mesopores.[15] Because residuum frequently has high levels of vanadium and nickel in it, the catalyst activity may actually increase initially and then slowly decrease as the promotional effects of the metals adsorbed on the hydrocracking catalyst are counterbalanced by deactivation because of coking. The resid hydroconversion reactors may be either trickle bed or moving bed. Because of the high rates of catalyst coking, most moving bed residuum conversion reactors have a capability for fresh catalyst addition and spent catalyst withdrawal while the reactor is operating.

Typical conditions used for the hydroconversion of residuum include:

- Pressure: 2000–2500 psig
- Temperature: 725–800 °F
- Space velocity: 0.4–0.8 × volume of oil/hr/volume of catalyst.

Process Kinetic Models

S. J. Khang and J. F. Mosby (personal communication) have presented process models to describe the removal of sulfur, vanadium, and nickel in residuum hydroprocessing. The kinetic models are given below in Eq. (16):

Desulfurization rate, second order, 1000°F inhibited

$$-r_s = ak_s(1 - C_A/C_{A_o})^m \quad S^2$$

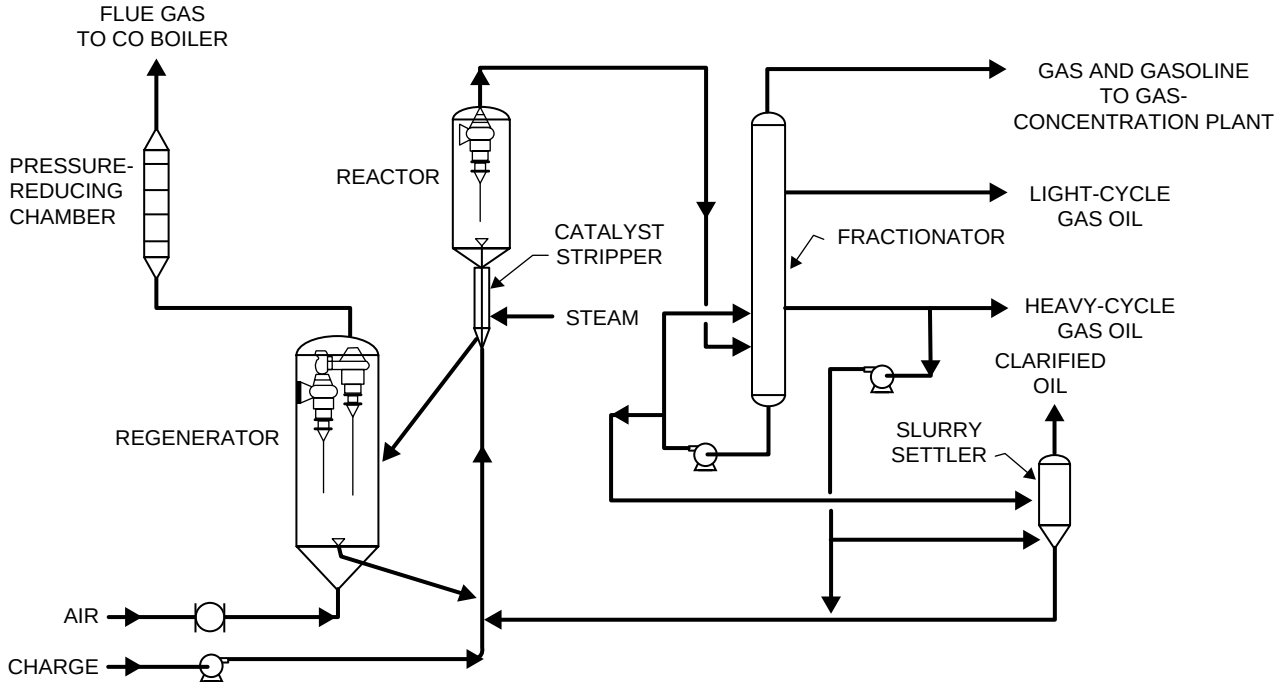


Fig. 19 The ''straight riser'' or ''side-by-side'' FCU design.

Devanadation rate, first order, 1000°F inhibited

$$-r_{\mathrm{v}} = ak_{\mathrm{v}}(1 - C_{\mathrm{A}}/C_{\mathrm{A_o}})^{n} \quad \mathrm{V}$$

Denickelation rate, first order, 1000°F inhibited

$$-r_{\mathrm{N}} = ak_{\mathrm{N}}(1 - C_{\mathrm{A}}/C_{\mathrm{A_o}})^{q} \quad \mathrm{N} \tag{16}$$

where m, n, and q are the orders indicating the degree of inhibition by 1000°F + materials.

Ebullated Bed Reactor for Resid Hydroconversion

Two moving bed processes are available for license, and include the H-Oil process developed by Hydrocarbon Research Inc. and the LC-Fining process developed by Cities Service and C-E Lummus.

The H-Oil reactor (Fig. 21) is rather unique and is called an ebullated bed catalytic reactor.[16,17] A recycle pump, located either internally or externally, circulates the reactor fluids down through a central downcomer and then upward through a distributor plate and into the ebullated catalyst bed. The reactor is usually well insulated and operated adiabatically. Frequently, the reactor-mixing pattern is defined as backmixed, but this is not strictly true. A better description of the flow pattern is ''dispersed plug flow with recycle.'' Thus, the reactor equations for the axial dispersion model are modified appropriately to account for recycle conditions.

The schematic in Fig. 22 shows the key elements of recycling a portion of the exit stream with the feed stream and how that affects the feed concentration(s). The recycle increases the superficial velocity, u, and changes the feed concentration owing to dilution with the product stream.

Although the H-Oil reactor is loaded with catalyst, not all of the reactions are catalyzed; some are thermal reactions, like thermal cracking, which depend on liquid holdup and not on how much catalyst is present. Thus, the material balance equations need to be divided into two categories, one set for the noncatalytic thermal reactions and another set for the catalytic reactions. A convenient parameter to use is the thermal volume/catalytic volume ratio, T/C, which is the ratio of liquid holdup to catalyst volume. In a commercial ebullated bed, this ratio is close to 1.0 under ebullation conditions. Consequently, the material balance equations for the catalytic reactions with no recycle are given in Eq. (17):

Dispersion model with no recycling catalytic reaction—only material balance

$$\left[\frac{D}{uL}\right]\frac{\mathrm{d}^2C_{\mathrm{A}}}{\mathrm{d}z^2} - \frac{\mathrm{d}C_{\mathrm{A}}}{\mathrm{d}z} + \frac{\tau r'_{\mathrm{A}}\rho_{\mathrm{B}}}{(T/C)} = 0 \begin{cases} C_{\mathrm{A}} = C_{\mathrm{A_o}}, & z = 0 \\ C_{\mathrm{A}} = C_{\mathrm{A_f}}, & z = 1 \end{cases} \tag{17}$$

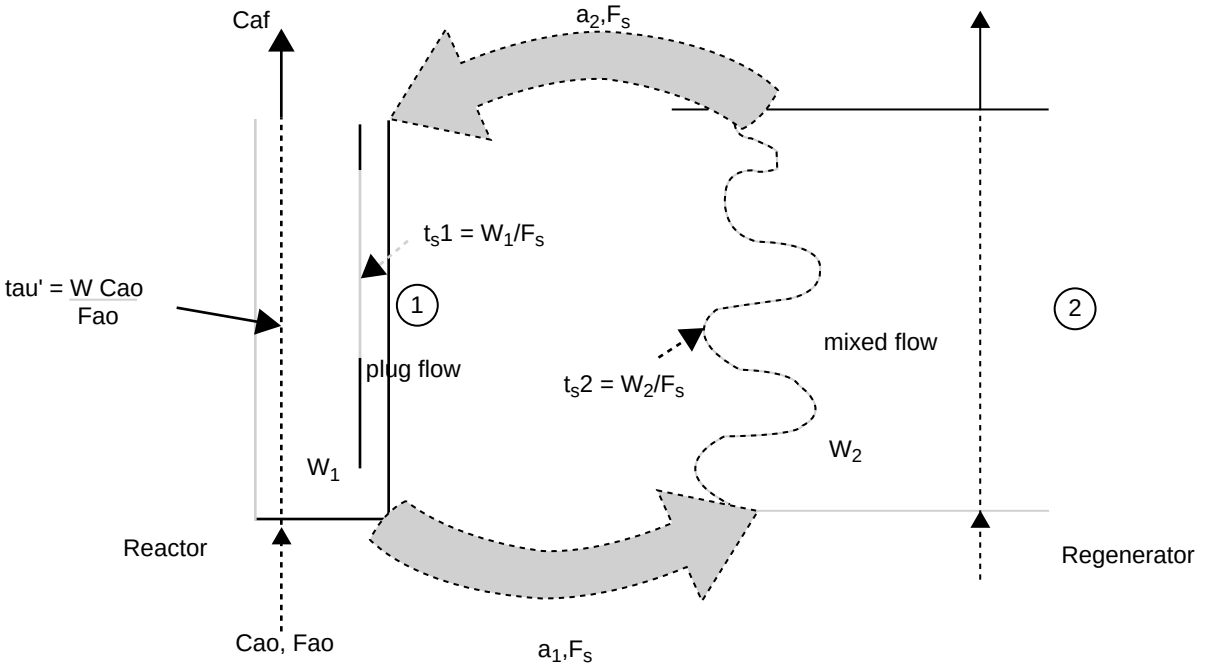


Fig. 20 Reactor–regenerator system. *(View this art in color at www.dekker.com.)*

where

T/C = thermal volume to catalytic volume
C_A = concentration
z = axial position in reactor, l/L (dimensionless)
D = diffusivity
u = superficial velocity, Q/A
L = reactor length
(D/uL) = dispersion number
r'_A = reaction rate of A (mol/hr/weight of catalyst)
τ = space time, L/u
ρ_B = bulk density of catalyst.

The final set of equations for the H-Oil reactor can now be written to account for the recycle situation, and thermal reactions in concert with catalytic reactions.

The model for the H-Oil reactor introduces two complications beyond the axial dispersion model. First, the boundary conditions are modified to account for the recycle and second, the catalyst in the reactor means that both thermal and catalytic reactions are occurring simultaneously. The set of equations given in Eq. (18) are solved numerically with a differential equation solver. This allows the reactor size to be determined by integrating along dimensionless distance z until concentration C_A equals the desired final concentration C_{A_f}.

Dispersion model with recycling catalytic and thermal reaction material balance

$$\left[\frac{D}{uL}\right]\frac{d^2C_A}{dz^2} - \frac{dC_A}{dz} + \tau\left[r_A + \frac{r'_A\rho_B}{(T/C)}\right] = 0$$

$$\begin{cases} C_{A_1} = \dfrac{(C_{A_o} + RC_{A_f})}{(1+R)}, & z = 0 \\ C_A = C_{A_f}, & z = 1 \end{cases} \tag{18}$$

Energy balance

$$\rho u C_P \frac{dT}{dz} = (r_A - \Delta H_R) + (r'_A - \Delta H'_R)$$
$$T = T_o \quad \text{at} \quad z = 0$$

Two- or More-Stage Residuum Hydroconversion

A two-stage residuum hydrocracker represents one additional level of complexity beyond the single-stage

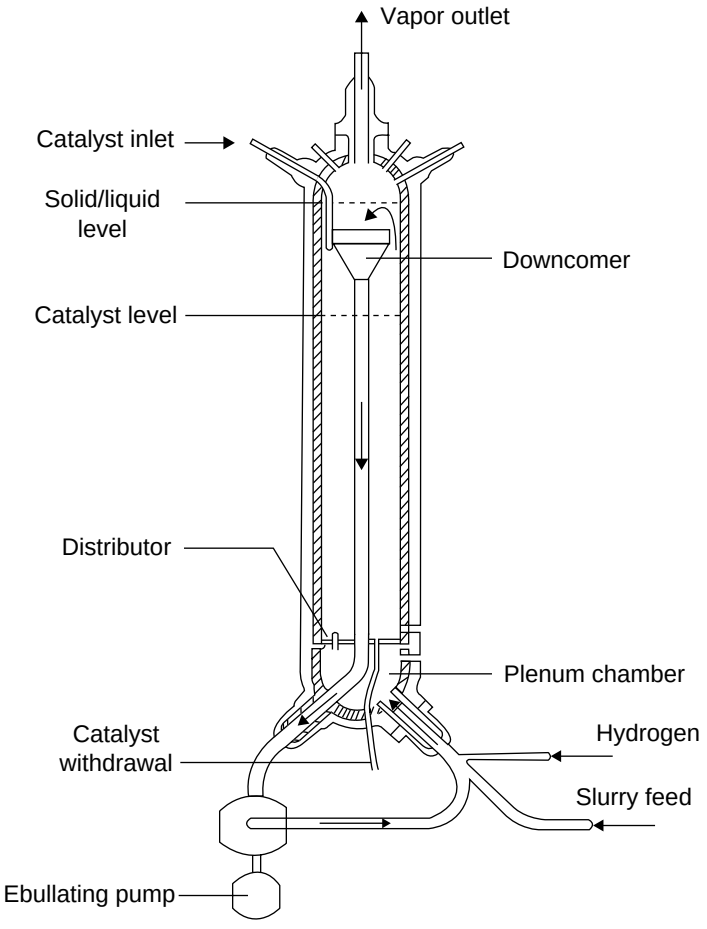


Fig. 21 Ebullated bed reactor.

H-Oil reactor. Now the exit stream from the first reactor becomes the feed stream for the second reactor. The reactor system may be visualized as shown in Fig. 23.

The approach to solving the above reactor system is to first develop the exit values for Reactor 1 and then to use these outlet variables as input for Reactor 2. The reactors will usually be operated at different temperature levels with stage 2 cooler than stage 1 so that the equilibrium for aromatic saturation reactions is favorably shifted. The interstage cooler is shown in the diagram for this reason. Furthermore, the second-stage catalyst may be different.

DELAYED COKING

Delayed coking is a semicontinuous process that is applied to the conversion of most types of residuum. Delayed coking is an endothermic process with the furnace supplying the necessary heat to complete the coking reaction in the coke drum. A diagram of the delayed coking process is given in Fig. 24, showing two coke drums, one active and the other off-line with the coke being cut out of it with high-pressure water nozzles. Hot fresh liquid feed is charged to the fractionator two to four trays from the bottom vapor zone. Any lighter material is stripped from the fresh feed

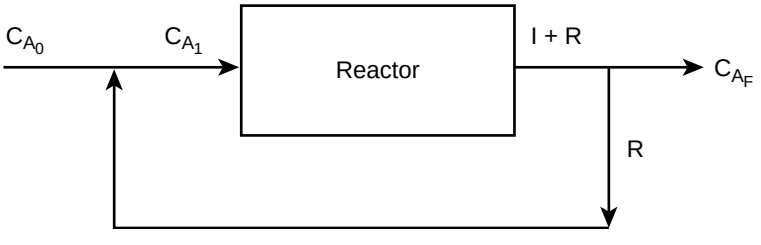


Fig. 22 Recycle reactor.

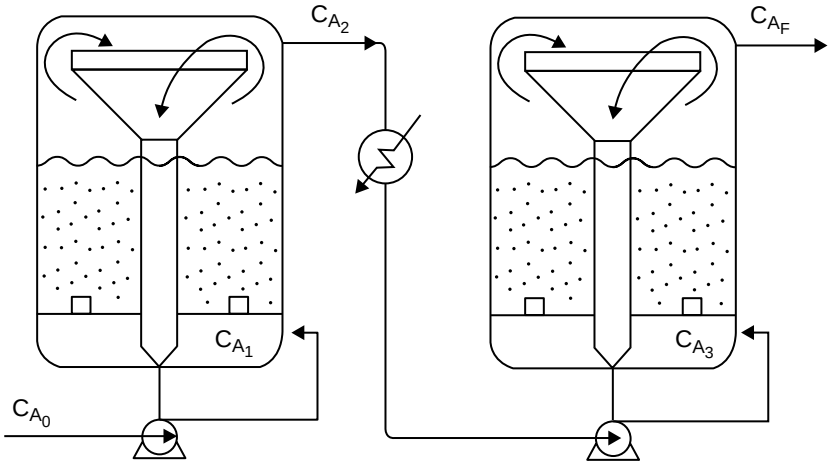


Fig. 23 Two stage residuum hydrocracker.

and then combined with the unconverted residuum to be fed through a heater and into the active coke drum. Steam is usually introduced in the heater tubes of the furnace to control velocities and prevent coke deposition in the heater tubes. The unvaporized portion of the heater outlet settles out in the coke drum where the combined effect of retention time and temperature causes the formation of coke. The exact mechanism of delayed coking is complex and it is not possible to determine the reaction steps involved; however, three distinct steps take place:

- Partial vaporization and mild cracking of the feed as it passes through the furnace.
- Cracking of the vapor as it passes through the coke drum.
- Successive cracking and polymerization of the heavy liquid trapped in the drum until it is converted to vapor and coke.

Three basic operating variables contribute to the quality and yields of delayed coking products. They are temperature, pressure, and recycle ratio:

- Heater outlet temperature: 875–975 °F
- Top coke drum pressure: 15–150 psig
- Recycle ratio, volume of recycle/volume of fresh feed: 0.05–2.

Temperature is used to control the volatile combustible material content of the coke product. The current trend is to produce coke with volatile matter between 6% and 8%. Pressure tends to retain more of the heavy hydrocarbons in the coke drum. This increases the coke yield and slightly increases the gas yield while decreasing pentane and heavier liquid product yield. Recycle ratio has the same general effect as pressure. As the recycle ratio is increased, the coke and gas yields increase and pentane, etc. decrease. The recycle ratio is primarily used to control the end point of the coker gas oil.

A typical coking cycle is as follows:

- Fill the drum with coke while it is in the ''active'' cycle—24 hr.
- Switch from active drum to inactive drum and steam out—8 hr.
- Cool with water and fill the entire coke drum—3 hr.
- Drain water from drum—2 hr.
- Unhead and decoke by drilling a 4 ft pilot hole and cutting—5 hr.
- Replace heads and test with steam—2 hr.
- Heat up by backing vapors from active drum to top of inactive drum—7 hr.
- Spare time for slippage of other steps—2 hr.
- Total—48 hr (24 hr cycle).

Yields for Delayed Coking

If the Conradson Carbon (Con Carbon) is known, the equations in Table 7 may be used.[18]

Table 7 Delayed coking yield equations

Coke (wt%)	1.6 × wt% Con Carbon
Gas (C4–) (wt%)	7.8 + 0.144 wt% Con Carbon
Gasoline (wt%)	11.29 + 0.343 wt% Con Carbon
Gas oil (wt%)	100 − wt% coke − wt% gas − wt% gasoline
Gasoline (vol%)	186.5/(131.5 + °API) × gasoline wt%
Gas oil (vol%)	155.5/(131.5 + °API) × gas oil wt%

Reaction Model for Residuum Coking

Levenspiel has developed an analysis for reactions where the phase changes from a liquid to a solid.[19] It is an adaptation of the Prout–Tompkins model and assumes that all coke nuclei are present at the start and no new ones are formed later.[20] Product grows from the nuclei along fracture places. The model is characterized by conversion first being slow because the boundaries between the reactant and product are very small. Growth then speeds up and eventually slows down as advancing product fronts meet. Please refer to Fig. 25 for the model description. This sounds strikingly similar to the situation in a delayed coker and thus it is proposed for use in the delayed coking process. Because little kinetic analysis has been done on this ''brute force'' process, it seems that some type of model is definitely needed for one to be able to predict performance and make logical judgments on process improvements. The reaction model takes the following form in Eq. (19):

$$\ln\left(\frac{X_R}{1-X_R}\right) = Mt - N \tag{19}$$

where

X_R = residuum conversion
M = characteristic slope
t = time
N = intersection point of plot where $t_i = N/M$.

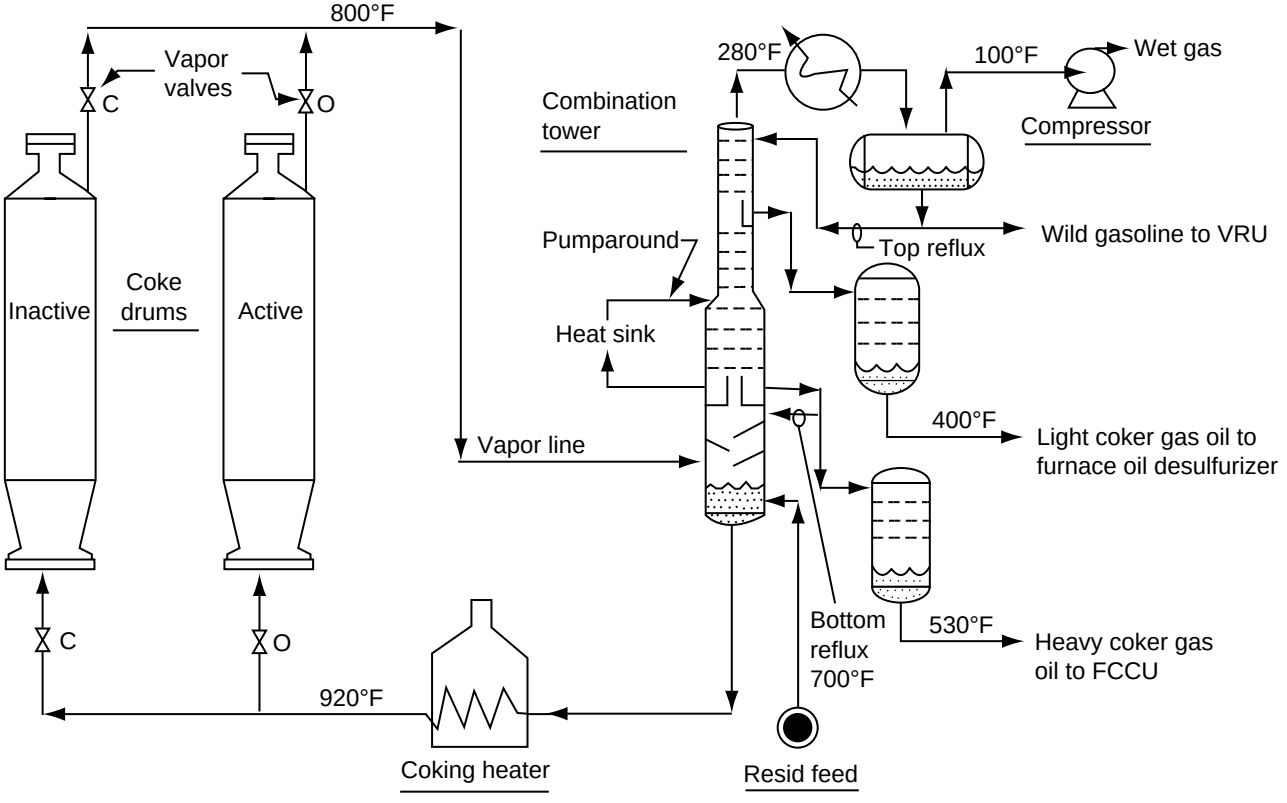


Fig. 24 Delayed coking process.

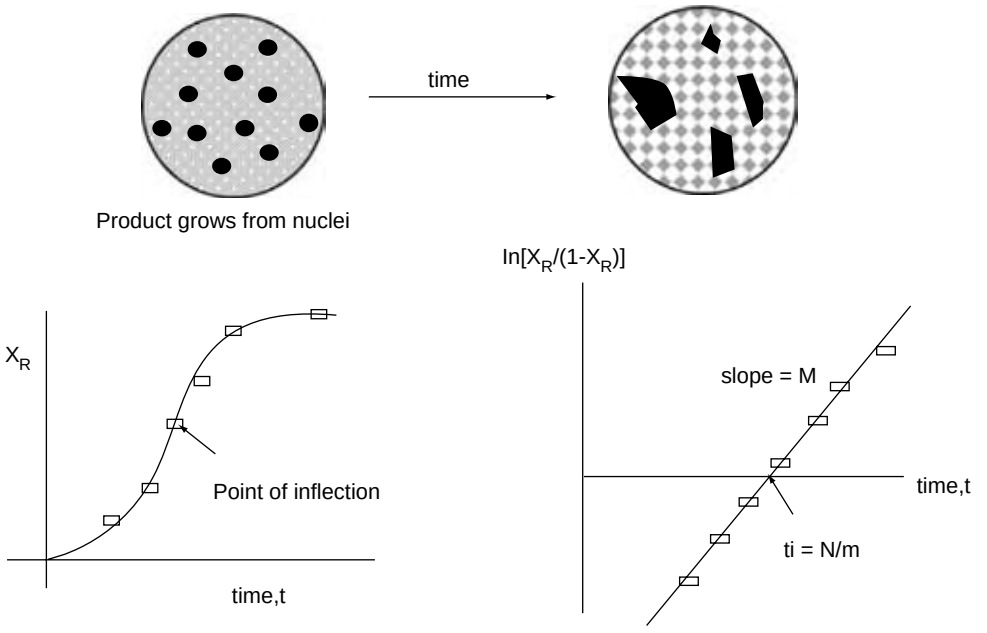


Fig. 25 Phase change model for coking. *(View this art in color at www.dekker.com.)*

In summary, a basic model can be used to interpret the coking data and establish how the coke is being formed as a function of cycle time in the coke drum.

CONCLUSIONS

The various reactors used in petroleum refining have been described along with the defining reactor performance equations and, occasionally, energy balances. The sizing of these reactors can be accomplished with these equations but, of course, for the detailed reactor design for materials of construction, process control, start-up/shut-down, we need to dig much deeper. Still, the methodology presented is the basic starting point to the final design and gives the engineer the tools to begin the design process. For catalytic reactions, this means how much catalyst is needed, and for noncatalytic reactions, it establishes the reactor size given the feed flow rate to the conversion unit. Although the focus has been on petroleum refining reactors, it should be noted that many of the reactors in chemical processing are very similar, where one has a well-defined feed composition rather than a mixture of many different hydrocarbons.

REFERENCES

1. Mahoney, J.A.; Robinson, K.; Meyers, E. Catalyst evaluation with the gradientless reactor. CHEMTECH **1978**, Dec, 758.
2. Levenspiel, O. *Chemical Reaction Engineering*, 3rd Ed.; John Wiley and Sons, 1999.
3. Gary, J.H.; Handwerk, G.E. *Petroleum Refining, Technology and Economics*, 2nd Ed.; Marcel Dekker, 1963.
4. Hengstebeck, R.J. *Petroleum Processing*; McGraw-Hill: New York, 1959.
5. Butt, J.B. *Reaction Kinetics and Reactor Design*, 2nd Ed.; Marcel Dekker, 2000.
6. Albright, L.F. Alkylation of isobutane with light olefins using sulfuric acid. Ind. Eng. Chem. **1970**, *9* (3).
7. Lieberman, N. *Troubleshooting Process Operations*; 3rd Ed; Pennwell Publishing Company, 1985.
8. Levenspiel, O. *The Chemical Reactor Omnibook;* Oregon State University Book Stores, Jan 1993.
9. Frey, C.G.; Mosby, J.F. Kinetics of hydrodesulfurization. Chem. Eng. Prog. **1967**, *63* (9).
10. Carberry, J.J. The contributions of heterogeneous catalysis to catalytic reaction engineering. Chem. Eng. Prog. **1988**, Feb.
11. Ho T.P. Hydenitrogenation catalysis. Catal. Rev. **1988**, *30* (1), 117–160.
12. Satterfield, C.N.; Yang, S.H. Ind. Eng. Chem. Process Des. Dev. **1981**, *20*, 53.
13. Pickert, P.E.; Bolton, A.P.; Lanewala, M.A. Molecular sieve zeolites: trendsetters in heterogeneous catalysis. Chem. Eng. **1968**, Jul 29.

14. Galbreath, R.B.; Van Driesen, R.P. Proc. 8th World Pet. Congr. **1971**, *4*, 129.
15. Fulton, J.W. Making the catalyst. Chem. Eng. **1986**, Jul 7.
16. Johnson, A.R.; Alpert, S.B.; Lehman, L.M.; Refinery applications of the H-oil process. 33rd Midyear Meeting of the American Petroleum Institutes Division of Refining, Philadelphia, PA, May 16, 1968.
17. Celestinos, J.A.; Zermeno, R.G.; Van Driesen, R.P. Oil Gas J. **1975**, *46* (48), 127–134.
18. Gary, J.H.; Handwerk, G.E. *Petroleum Refining, Technology and Economics*; Marcel Dekker, 1984.
19. Levenspiel, O. *The Chemical Reactor Omnibook*; OSU Book Stores Inc., 1993.
20. Prout, E.G.; Tompkins, F.C. Trans. Faraday Soc. **1944**, *40*, 488.

Real-Time Optimization: Status, Issues, and Opportunities

R

J. F. Forbes
Department of Chemical and Materials Engineering, University of Alberta, Edmonton, Alberta, Canada

T. E. Marlin
Department of Chemical Engineering, McMaster University, Hamilton, Ontario, Canada

W. S. Yip
Suncor Energy Inc., Fort McMurray, Alberta, Canada

INTRODUCTION

Process control, involving both closed-loop automatic control and advisory systems, contributes to the safe and profitable production of materials with consistently high product qualities at specified production rates. In many process plants considerable opportunity exists for further improvement of plant operations by increasing profit; this increased profit is achieved through Real-time operations optimization (RTO). Real-time operations optimization is an integral element of the typical hierarchical plant automation system shown in Fig. 1. The process control level achieves safety, product quality, and production rate goals, and all closed-loop technology including multivariable control (e.g., model predictive control) is implemented at this level. The RTO level is the first level at which economics is considered explicitly in operations decisions, i.e., in control calculations. This level addresses the short-term decisions on a time scale of hours to a few days. The third level provides integration of multiple RTO systems, which would be required for very large plants that cannot be modeled in a single system. The fourth level addresses longer-term issues like material inventories and production rate targets. All levels utilize process measurements as inputs to their feedback loops. Each higher level provides guidance to the subsequent lower level; for example, the RTO outputs could be set points to the process controllers.

The structure in Fig. 1 is essentially a cascade design. This structure is appropriate because of the differences in the plant dynamics and disturbance frequencies associated with the decisions at each level. Control responds to rapid disturbances and requires execution periods on the order of 1 sec. Real time optimization typically responds to disturbances occurring every few hours or slower, and it requires tens of minutes to compute. The higher levels respond to disturbances occurring every few days. While the cascade structure provides timely responses, it introduces interactions among decision makers that can lead to suboptimal results; therefore, coordinating the levels becomes a challenge.

This entry is organized in the following manner: the first section provides a thorough overview of the process operations challenges that RTO addresses, the potential opportunities that may be realized from effective RTO and the range of applicability of RTO; the second section focuses on the architecture of a typical RTO system, provides a summary of the current state of component technologies of an RTO system and areas where further development would be beneficial; the third section provides a discussion of possible future directions for RTO research and an introduction to some emerging RTO technologies; the entry closes with some concluding remarks.

The Fundamental Challenge in RTO

Process plants have benefited from the application of process control for many decades. Therefore, one might ask, ''What is the benefit for optimization beyond successful process control?'' The answer lies in the fundamental difference between control and optimization. In process control, the goal is assumed to be known, and the challenge is to achieve the goal. For example, the set point of each controller is specified, and the control performance is measured as the deviation on the measured variables from its set point.[1] In addition, the relationship (gain) between the adjusted variable and the controlled variable has an unchanging sign.

In optimization, the goal is defined in a very general manner through the objective function, constraints, and parameter values. Changes in the adjusted variables that increase profit can depend on the scenario. Stated another way, the gain between the adjusted variables and the profit can change magnitude significantly and

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030615

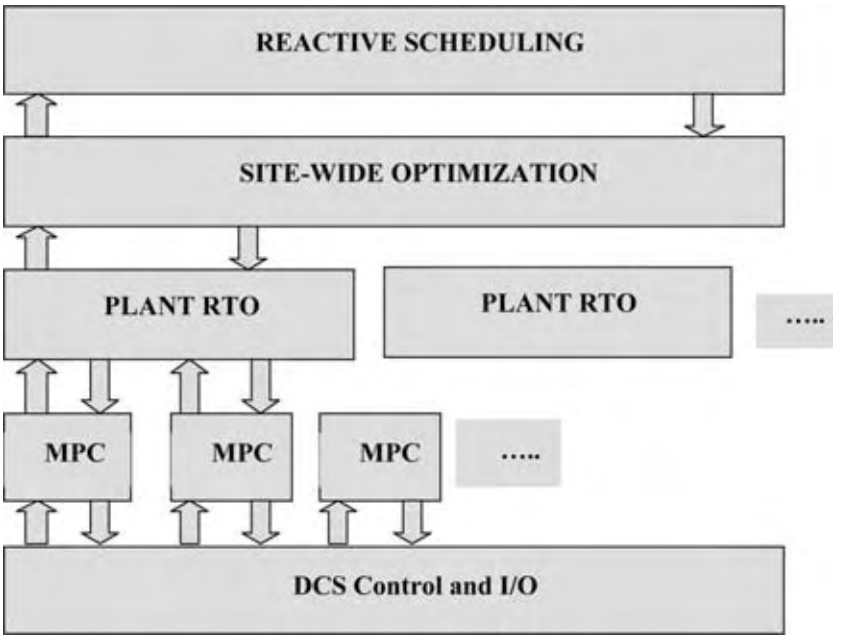


Fig. 1 Plant decision-making hierarchy.

can change sign. Thus, optimization is appropriate when the correct direction and distance for changes in adjustable variables cannot be determined through prior analysis, because they depend on the specific scenario. To achieve this demanding goal, optimization requires substantially better information about the process than does feedback control.

The RTO Opportunity

A key challenge for the engineer is to identify plants that will benefit from the application of RTO. Some key factors in selecting good candidates are given in the following with examples for an ethylene plant shown in Fig. 2. The process technology is covered in for example, Grantom, and Royer, and operation optimization for an ethylene plant is presented in Yang and Waldman.[2,3] Increased plant profit is possible via real-time optimization when the following criteria are satisfied.

1. Additional adjustable optimization variables exist after higher-priority safety, product quality, and production rate objectives have been achieved. Typically, many operating policies will achieve

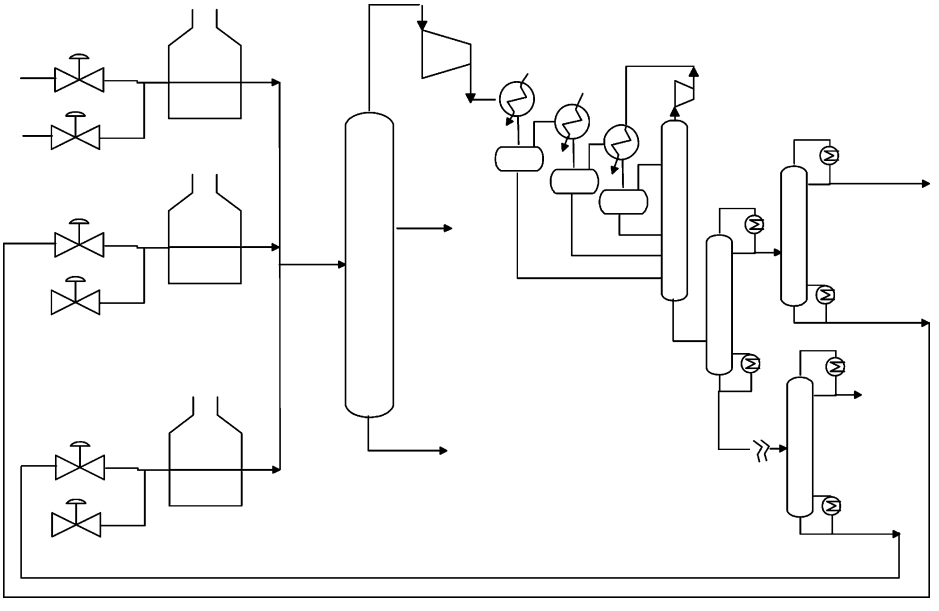


Fig. 2 Schematic of an ethylene production plant.

the same production and product quality because the plant contains more adjustable variables (final control elements) than control objectives. These final control elements are included to provide operating flexibility. For example, several boilers are installed, rather than one large boiler, which provides the ability to generate steam efficiently over a wide range of demands. In addition, one of the multiple boilers can be shut down (either for maintenance or for emergency shutdown) without interrupting the operation of the plant. As a result, optimization can reduce fuel consumption by adjusting steam production in the multiple boilers.

For an ethylene plant, each reactor has three key operating variables; outlet temperature, steam dilution, and feed flow rate. Plants have on the order of 24 parallel reactors (with several within each heater). In addition, the separation section has many variables that can be adjusted for optimization, including recycle compositions, distillation pressures, and refrigeration temperatures. Thus, a plant often has 100 variables or more for optimization.

2. The profit changes significantly as values of the optimization variables are changed. Naturally, no optimization is justified if the values of the optimization variables do not significantly influence profit. For the boiler example discussed above, similar boilers could have essentially the same efficiency relationship with the rate of steam production. If these efficiency curves were concave (as they usually are) optimum profit would be achieved at equal production by every boiler.

   For a typical ethylene plant, the cost of naphtha feed and fuel would be on the order of U.S.$ 260 million/yr. thus, a small percent decrease would yield a significant improvement in profit.
3. Disturbances occur frequently enough for real-time adjustments to be required. Real-time operations optimization improves profit by responding quickly to disturbances when compared to conventional off-line optimization, which involves a person inputting the data and analyzing the results. The importance of the disturbance frequency is represented in Fig. 3, in which the loss of profit for perfect optimization is plotted vs. disturbance frequency. (The loss of profit is used so that the figure resembles the common Bode plot used in automatic control, e.g., Ref.[4].) The dot–dash line represents the profit achieved if no optimization were performed, which is unlikely but provides a base line for the worst performance. Naturally, the profit losses would be large for most disturbance frequencies; however, the effect of all frequencies would diminish when the frequency is high, beyond the corner frequency of the process, where the process ''filters'' the disturbances.

   The dashed line represents the common situation of infrequent, off-line optimization. Performance is good and the profit losses are small for disturbances that occur at a much lower frequency (longer period) than the off-line analysis. The off-line analysis is not effective for disturbances occurring faster than the off-line analysis frequency, and as discussed, the disturbance effects decrease for a frequency beyond process corner frequency. Note that the off-line analysis performance requires the technology and software tools described here for RTO; if off-line analysis is performed using simplified technology, profit losses are likely to occur even at very low disturbance frequencies.

   Finally, the RTO system extends the range of good performance to much higher frequencies because of its much higher frequency of execution. The RTO performance is good until the frequency reaches the resonant frequency of the plant (under closed-loop control). Note that this resonant frequency is primarily determined by the plant dynamics, although it is affected by the process control design and tuning. Thus, RTO offers the potential for increased profit for disturbance frequencies between the off-line frequency and the resonance frequency.
4. Determining the proper values for the optimization variables is too complex to be achieved by selecting from standard operating procedures. Steps 2 and 3 above establish that a potential benefit exists for changing plant operations frequently. However, the optimum operating policies must be evaluated to determine if they could be achieved through process control or whether they require frequent optimization. Approaches for designing control that (approximately) maximizes profit is addressed through ''self-optimizing control.'' (e.g., Ref.[5].) Examples of self-optimizing control include maintaining the proper ratio of steam to hydrocarbon feed to the cracking furnace reactors and minimizing the pressure in a distillation tower.

The Use of Models in RTO

Optimization can be achieved using various combinations of models and plant data to represent the plant. The correct approach depends on the specific application. Some small-scale optimizations can be addressed using direct search methods. (e.g., Ref.[6–10].) These

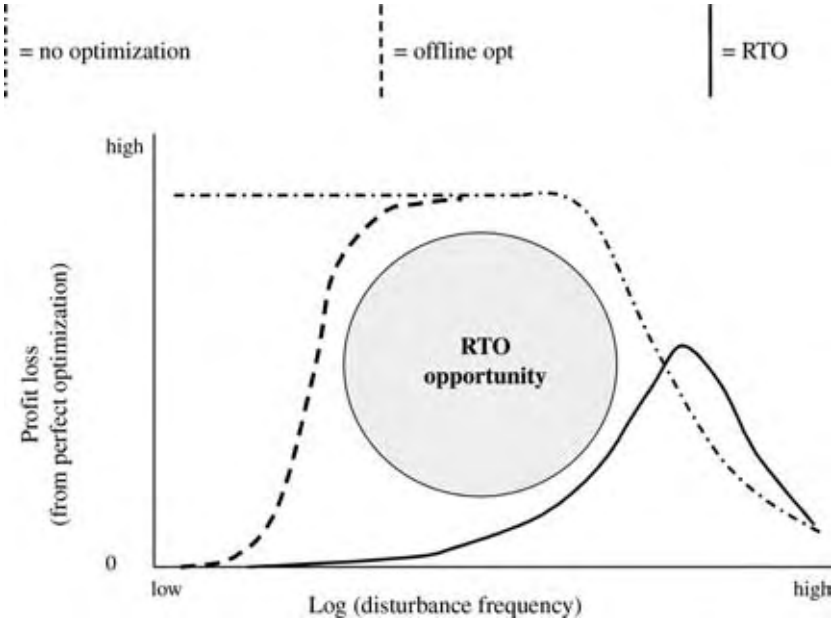


Fig. 3 The effect of disturbance frequency on optimization performance. *(View this art in color at www.dekker.com.)*

methods introduce designed experiments to determine the gradient of the profit. By the nature of this approach, the gradient estimate is only locally valid and corrupted by disturbances and measurement noise. In addition, the experiments can be time-consuming, especially when the method requires that a steady state be achieved between experiments. However, this approach can be appropriate when the following criteria are satisfied.

- Relatively few variables are adjusted by the optimizer.
- A good estimate of profit can be calculated using the values of measured variables.
- The plant achieves steady state quickly.
- Experiments of small magnitude do not seriously degrade product quality.

In contrast, plant-wide optimization involves a large number of variables throughout a slowly responding plant. For this situation, a model-based method is required to approach the best conditions within reasonable times. A fundamental process model provides information on the effects of optimization variables on the profit and potential inequality constraints over a wide range of conditions. A model-based approach is required for the following situations.

- A large number of variables are adjusted by the optimizer.
- The profit and potential inequality constraints are affected by variables that are not readily measured in real time, but can be predicted with reasonable accuracy.
- The plant responds slowly to changes in operation.
- Strong interactions occur between variables, so that sequential optimization of individual variables is not appropriate.
- Plant operations change frequently, requiring frequent re-optimization.

As an example of the importance of integrated models, the distillation column in Fig. 4 is considered. The optimizer must determine the effect on energy costs of a change in the reboiler heat transfer. In a typical process, the steam is generated in boilers that are owned by the plant and supply steam at several pressures for numerous consumers. The change in reboiler steam flow rate influences both steam levels, the steam flow through the turbine and the steam production in the boiler. Clearly, the optimizer must include models of the distillation process and the entire steam system; only with this complete model can a change in reboiler duty be related to the change in fuel purchase.

The models used in RTO are more complete than those typically used for plant design, because the equipment performance calculations are integrated with the flowsheet material and energy balances. Because the plant equipment exists, it must be modeled as it performs. One important reason for detailed performance models is the change in performance as the optimizer adjusts plant operation; for example, the shell and tube exchanger overall heat transfer coefficient depends on the flow rates. Another reason is the limitations imposed by equipment on the feasible operating window. For example, flows are limited by pumping capacities, separation by distillation tray hydraulic limits, and heat transfer by area and temperature driving force.

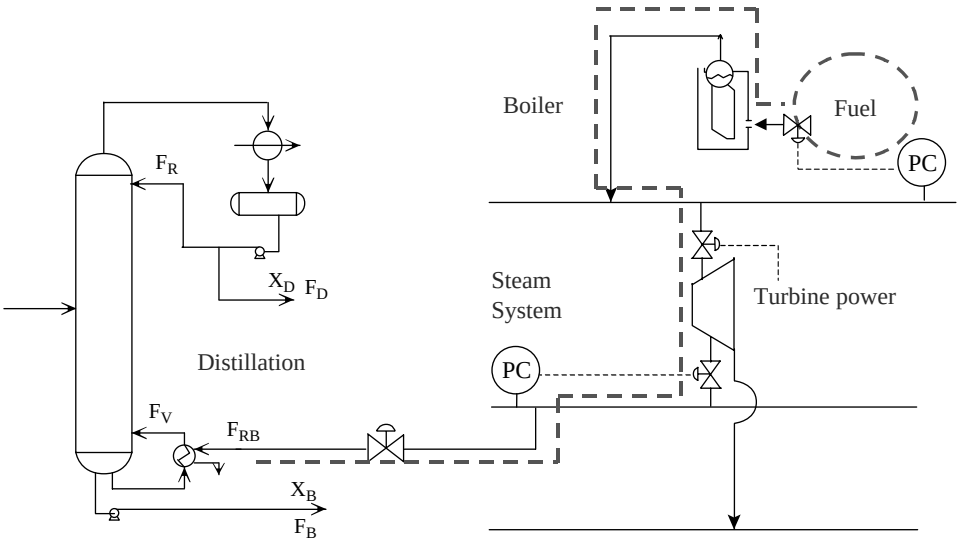


Fig. 4 The integrated effect of a process change on the plant and utility system. *(View this art in color at www.dekker.com.)*

Another important facet of RTO models is the inclusion of the effects of the lower-level process control. The controls are included because they influence the total effect of changes in optimization variables on the profit. A simple example demonstrating the importance of control is shown in Fig. 5. The fuel gas system pressure is controlled by adjusting one of the two sources of purchased fuel. Because the two sources have different prices, the selection of the manipulated fuel determines the incremental price of the fuel. The control design determines that the incremental prices is \$2/GJ. Note that this incremental price is different from the average price. Also, the price for reducing fuel consumption is lower than if the higher price fuel had been adjusted.

Naturally, even the best available fundamental models do not match the true plant behavior exactly. Therefore, RTO employs feedback to improve the approach to the true plant optimum by updating selected model parameters using current measurements. The performance of an RTO system depends strongly on the availability of key measurements, selection of updated parameters, and fidelity of the fundamental model structure and constant parameters.

RTO SYSTEMS ARCHITECTURE

A typical structure of a closed-loop RTO system is shown in Fig. 6, which consists of subsystems for data validation, model updating, model-based optimization, and command conditioning.[11] Raw measurements taken from the plant are filtered and checked for reliability in the data validation subsystem. Because, model-based RTO systems rely on model updating to correct for modeling errors and disturbances, an effective model updating system is required to ensure that the RTO system tracks the changing optimal operations closely. Model updating, most commonly effected via on-line estimation of some set of model parameters, uses the validated data. The updated process model is then used by the model-based optimization

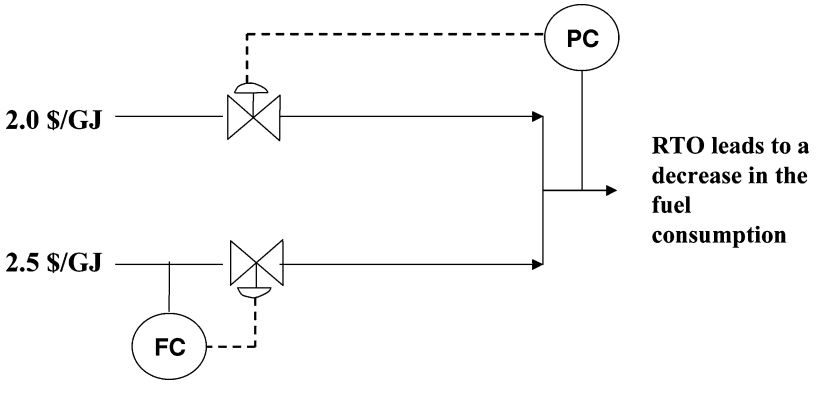


Fig. 5 Determining the incremental cost of fuel requires a model for the control system.

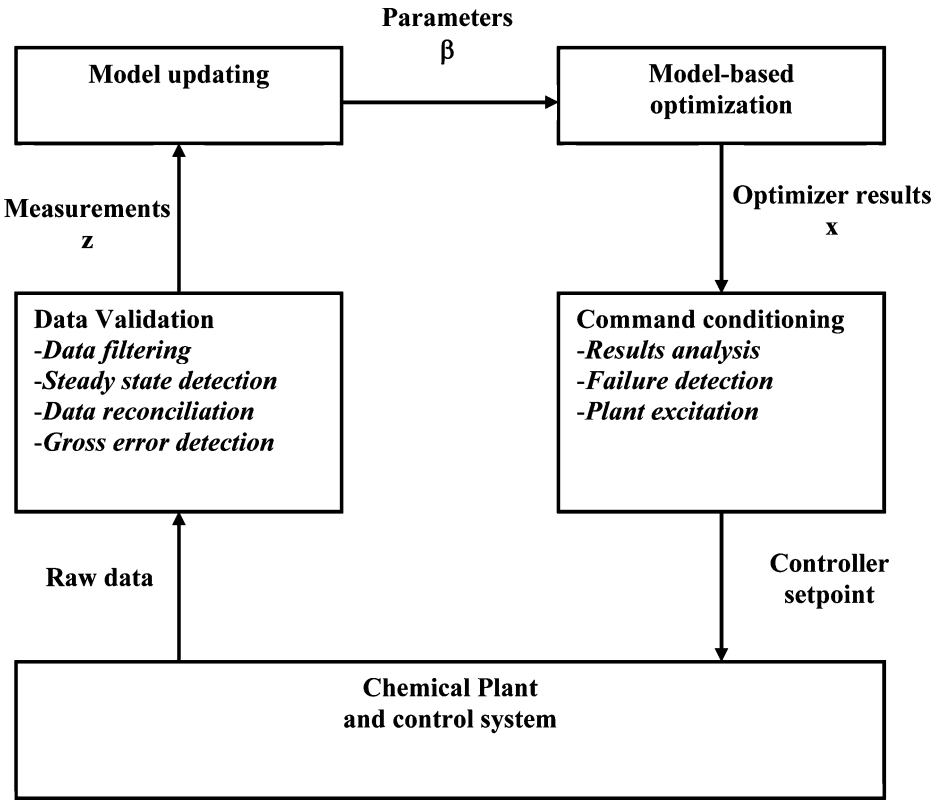


Fig. 6 Closed-loop RTO system.

subsystem to determine the optimum operating policy. The optimization results are analyzed and validated in the command conditioning subsystem before being forwarded to process controllers for implementation. In the following subsections, the details of individual subsystems, which represent the current available technologies for RTO implementation, are described.

Data Processing

The performance of an RTO system depends on the ability of the model to accurately reflect plant behavior, which is observed via process measurements that can be unreliable and subject to disturbances. Therefore, plant measurements are processed and validated in the data validation subsystem before they are used for model updating. The data validation process consists of three steps: 1) steady-state detection; 2) measurement filtering; and 3) data reconciliation and gross error detection.

Steady-state detection

Most industrial RTO applications consider steady-state optimization. Thus, measurements should represent steady-state plant behavior when used to update the RTO model. The steady-state detection module determines if a steady state has been reached based on plant measurements by one of the various statistical methods. Control charts in statistical process control, such as the Shewart chart,[12] can be used to monitor the process variables. If the variations of the process variables are within the control limits, the process is considered at steady-state. Steady state can also be detected by Hotelling T^2 test analysis of variance (ANOVA).[13–15] These techniques identify if the difference between two sample means taken in two consecutive periods is due to common cause variability or not. If the sample means are statistically equal, the process is considered at steady state.

Measurement filtering

Measurements are often corrupted by high frequency, zero-mean, random variations (i.e., measurement noise). These variations can be caused by high-frequency disturbances from the process (e.g., vortices shedding in flow measurement) or electrical interference during signal transmission. Plant measurements should be filtered to reduce the impact of high-frequency variations on the quality of the estimated parameters and as a result, the predicted optimal plant operation.

The filter should be able to attenuate spurious high-frequency variations without affecting the low-frequency variation to which the RTO system must respond. Such

low-pass filtering can be achieved by a variety of approaches, including averaging a set of sampled measurements. Commonly used averaging filters are the moving average and exponentially weighted moving average filters. In the moving average filter, a moving window with fixed size is used to keep a record of the most recent data sets. The filtered signal is calculated by averaging these equally weighted data sets. In the exponentially weighted moving average filter, weightings for the data sets are chosen so that older data sets are discounted in an exponential manner, and the filtered signal is calculated as the weighted mean of the data sets (cf. Ref.[16]).

Gross error detection and data reconciliation

After the process measurements have been initially processed, they should be analyzed for the presence of gross errors before being used in model updating. Gross errors are systematic errors, which can result in measurement bias. Gross errors can be caused by instrument miscalibration, changes in process conditions (fluid density measured by an orifice meter), inappropriate sensor technology (thermocouples for accurate temperature measurement), and so forth. Measurements that are contaminated with gross errors should be removed from the data set before further processing. Measurement data sets that have been deemed to be free of gross errors are then reconciled to ensure their consistency with respect to conservation laws, e.g., material balances.

The measurement adjustments required to satisfy the conservation law(s) are used to detect gross errors in the measurements and to ensure that the data set used for model updating is self-consistent.

Data reconciliation can be formulated as an optimization problem. The adjusted values of the measurements are determined by minimizing the weighted sum of squared measurement adjustments (i.e., the difference between measured values of the process variables and the process variable values that would satisfy the requisite conservation laws) subject to the conservation laws:[17]

$$\text{Minimize } \hat{\mathbf{z}},\ \mathbf{x} \quad (\mathbf{z} - \hat{\mathbf{z}})^{\mathrm{T}}\mathbf{Q}^{-1}(\mathbf{z} - \hat{\mathbf{z}}) \tag{1a}$$

$$\text{subject to} \quad \mathbf{B}(\hat{\mathbf{z}}, \mathbf{x}) = 0 \tag{1b}$$

where z is the vector of steady-state measurements, $\hat{\mathbf{z}}$ is the vector of reconciled process variables, Q is the measurement covariance matrix, x is the vector of unmeasured process variables, and the constraint in Eq. (1b) represents the steady-state balances. In this formulation, it is assumed that measurement errors are random errors. It is important to note that data reconciliation can be performed only when the set of process measurements contains sufficient redundancy.

Gross errors can be detected using statistical tests on the measurement adjustments. If there are no gross errors in the measurements, the measurement adjustments should be statistically indistinguishable from 0. Gross errors can be detected using a range of statistical tests on the adjustment(s) to the measurements.[17–22] For example, the sum of squared adjustments to all flow rates can be compared with a chi-squared statistic to test the hypothesis that the adjustments are all randomly distributed within their variances. Identification of gross errors via the typical statistical testing approaches requires iteration, as the underlying assumption of the available methods is that measurement data contain only random errors. Therefore, when gross errors are detected, suspect measurements are removed from the measurement data set and the procedure is repeated until no identifiable gross errors remain or there are insufficient measurements available to perform the tests.

Recently, simultaneous procedures for data reconciliation and gross error detection have been developed based on the maximum likelihood principle.[23–25] This formulation combines the univariate distributions for random errors and gross errors to account for the possible occurrence of gross errors in the measurements; however, an a priori knowledge of the gross error distributions is required for these methods. In these procedures each measurement error is tested using a combination of the assumed distributions to detect gross errors. These combined data reconciliation/gross error detection approaches have not been extended to handle highly correlated, jointly distributed gross errors.

Model Updating

Because plant–model mismatch is inevitable, process models must be updated on-line, and this is usually accomplished using validated process measurements. Most commonly, model updating is limited to a selected set of model parameters (e.g., heat transfer coefficients, reaction rate constants, and unmeasured feed compositions). The selection of parameters for updating is guided by the key insight that the optimizer uses derivatives in finding the economic optimum operation. Therefore, the updated parameters should change significantly and improve the accuracy of the important variables used in economic optimization. After the parameters have been selected, sensors can be located to ensure that the updated parameters can be calculated with sufficient accuracy. Selection of parameters and measurements for updating has been addressed by several researchers.[26–29]

Some commercial RTO systems either "back-calculate" model parameters from available process measurements or use some form of least-squares parameter estimation.[27] The back-calculation approach has been shown to be less robust to errors in the data and model than the least-squares approach and, as a result, is less desirable in RTO applications. The least-squares model updating problem can be solved by a number of nonlinear programming solvers, and the estimated parameters used for model-based optimization.

As an alternative to a sequential approach (gross error detection followed by parameter estimation), simultaneous data reconciliation and parameter estimation (DRPE) has been espoused by a number of researchers and has been tested in industrial application.[30–34] In the DRPE formulation, parameters are estimated by minimizing the difference between the measured and calculated values of process variables in a least-squares sense:

$$\text{Minimize } \beta, \hat{z}, x \quad (z - \hat{z})^{T} Q^{-1} (z - \hat{z}) \tag{2a}$$

$$\text{subject to} \quad f(\hat{z}, x, \alpha, \beta) = 0 \tag{2b}$$

where z is the vector of measured variables, $\hat{z}$ is the vector of adjusted measurements, α is the vector of fixed parameters, β is the vector of estimated parameters, x is the vector of unmeasured process variables, and the constraints in Eq. (2b) represent the process model. This model updating problem can be solved by any nonlinear programming solvers, and the estimated parameters are used for model-based optimization.

The main difference between the least-squares DRPE problem Eq. (2) and conventional least-squares parameter estimation is that in DRPE the measured variables are estimated, as well as the parameters and unmeasured process variables. Thus, the combined DRPE problem in Eq. (2) can become ill-conditioned because of missing measurements, changes in operating points, and so forth. In an ill-conditioned parameter estimation problem, measurement errors will be amplified in the model updating subsystem causing RTO performance degradation. Miletic and Marlin developed the updater diagnosis strategy to detect the ill-conditioned parameter estimation problem by checking the condition number of the covariance matrix of the estimated parameters.[35] If the ill-conditioned parameter estimation problem is identified, a recovery strategy of fixing one of the model parameters at its previous value and reestimating the other parameters is proposed by the authors. The choice of the parameter to be fixed can be made by evaluating each parameter's contribution to the ill-conditioning of the parameter estimation problem using singular value decomposition. The parameter that contributes the most to ill-conditioning should be fixed. The procedure is repeated until the parameter estimation problem becomes well conditioned.

Yip and Marlin showed that RTO performance can be improved through the use of multiple data sets for updating.[36] In this approach, a moving window of fixed size is used to keep a record of the current and most recent data sets for updating. Using multiple data sets increases the parameter observability and reduces the variability of the parameter estimates.

Economic Optimization

The updated model can be used to determine the optimum values for the optimization variables. The general formulation of the economic optimization problem is:

$$\begin{aligned} &\max_{x,y} P(\beta, x, y) \\ &s.t. \\ &f(\beta, x, y) = 0 \\ &g(\beta, x, y) \le 0 \\ &x_{\min} \le x \le x_{\max} \\ &y_{\min} \le y \le y_{\max} \end{aligned} \tag{3}$$

The symbols are P for profit, f for equality constraints, g for inequality constraints, x for optimization variables, y for dependent variables, and β (constant) for updated parameters. The objective function is a scalar measure of plant profit; it is usually the instantaneous profit (\$/hr), because the optimization variables do not involve the time value of money. Typical equality constraints include material and energy balances, heat and mass transfer relationships, and thermodynamic and kinetic models, and typical inequality constraints include equipment limitations limit compressor horsepower, and distillation tray hydraulics. The optimization variables are flow rates, pressures, temperatures, and other variables that can be manipulated directly. The dependent variables involve intermediate values required for the detailed models; for example, all distillation tray compositions, flow rates, and temperatures. Because of the fundamental models often used in RTO, the number of dependent variables can be quite large, on the order of hundreds of thousands.

Results Processing

Once the RTO system has calculated the optimal operating variables for the plant or system, these

optimal results must be analyzed to ensure that a number of critical issues are addressed, including:

1. The results are still relevant, given that RTO computations can take a substantial amount of time relative to the dominant plant dynamics.
2. The results may be corrupted by sensor noise (and other common cause sources of variation) propagating throughout the RTO loop.
3. The results may not represent a sufficiently large improvement in the operation to merit the recommended change.
4. The results may lead to a decrease in operating efficiency due to plant–model mismatch.
5. The results may not provide sufficient ''excitation'' to improve model accuracy in subsequent model updating cycles.

To address these issues, results processing must include functions for results analysis, failure detection, and information generation.

Results analysis

Of the five issues identified in the preamble to this section, results analysis addresses the first three, which focus on the relevance and validity of the calculated optimal operating targets.

Validity Checking. The main idea is to ensure that the plant has not undergone any substantive change (e.g., a change in feed conditions or process equipment being brought on- or off-line) during the time the RTO calculations took to execute. In general, it would not be appropriate to implement the results of an RTO calculation if model parameters, which were determined during RTO calculations, are no longer valid because of some set of system changes that have occurred during the course of these RTO calculations.

Determining whether the RTO results are still relevant to system operation most often involves testing to see if the steady-state operation has changed during the execution of the RTO calculations. This usually involves testing whether some set of preidentified set of process variables has significantly deviated from their measured values prior to the start of RTO calculations. No literature has developed on setting the thresholds, and as a result, practitioners rely solely on experience. If the plant is not at steady state, no variables are changed by RTO; it is especially important to recognize that outputting a partial solution could lead to suboptimal and even unfeasible plant behavior.

Significance Testing. The second issue that the results analysis subsystem must address is whether the difference between the RTO results and the current operation is due to the propagation of some ''common cause'' variability (e.g., sensor noise), which has resulted in a spurious change to the calculated optimal operation.

Miletic and Marlin proposed a statistical test to determine whether the difference between the current process operation and that predicted by the RTO system is statistically significant with respect to the common cause variability that propagates around the RTO loop.[35] In their work, Miletic and Marlin focus on the predicted values for the controller set points. When the RTO economic optimization problem is subject to limits on the amount that it will be allowed to change the process operation, this approach can fail. Zhang, Nadler and Forbes extended the approach to handle the situation where RTO changes are limited by constraints and proposed a two-step procedure for analyzing the statistical significance of RTO predictions.[29] The above statistical test is valid only when the common cause variability of the optimization results does not result in changes to the active set, so that these variations can be characterized by an ellipsoidal confidence region. Yip and Marlin developed a result analysis strategy to handle the situation where variations in the optimization results could result in active set changes.[37]

The final issue that the results analysis subsystem must address is whether the calculated change in optimal operation will produce a sufficient change in process performance that merits the change. Any RTO prediction, which is not restricted by move limits, that does not produce a significant change in the process performance should not be implemented, as it may contribute to unnecessary process variation and possibly disruptions to effective process operation. This can occur when the process performance is (locally) insensitive to changes in the RTO decision variables (i.e., process controller set points). Statistical tests similar to those used in checking the statistical significance of changes to the RTO decision variables could be employed; however, little work has been done in developing effective methods for testing the significance in predicted changes in process performance. Practitioners typically use ''common sense'' thresholds in the profit based on estimated model accuracy and meaningful profit changes.

RTO: Fault Detection and Diagnosis. Halim proposed a method that uses gradient information from the plant, which is obtained via plant experiments, and the model to both quantify plant–model mismatch and monitor RTO system performance.[38] In this work, the angle between the two profit gradients of the model-based RTO system and that obtained via plant experiments is used as an indicator of RTO

performance and the presence of faults. The source of the faults is determined by comparing input–output sensitivity matrices obtained from the plant and the model, using matrix decomposition techniques.

Current industrial technology involves monitoring the results from the DRPE updating calculation. Heuristic rules are used to determine whether measurement adjustments are ''too large'' or parameter values are ''beyond anticipated ranges.'' When results are questionable, RTO output is typically halted until the results return to acceptable values.

Information Generation. Pfaff studied the problem of ensuring that predictions in subsequent RTO intervals increase in accuracy.[39] In this work, he proposed an expansion of the command conditioning subsystem to evaluate whether the predicted RTO results generate sufficient information to increase the accuracy of parameter estimates in future RTO calculations. When the RTO predictions are not sufficiently information-rich (i.e., do not contain adequate excitation) an experimental design calculation, which uses an A-optimal criterion to reduce the uncertainty associated with decision variables, is employed to produce small adjustments to the RTO predictions. The resulting process controller set points sacrifice short-term profit for improved accuracy and greater profit in future RTO intervals. A profit-based experimental design criterion has been developed by Yip and Marlin which considers short-term profit loss during experimentation and future profit gain after experimentation to design the plant excitation.[40]

Apparently, no commercial application introduces plant changes to improve the parameter updating in model-based RTO. In some cases, the actions of the RTO system in responding to disturbances provide adequate variability.

Results analysis procedure

Any comprehensive results analysis procedure should involve five stages:

1. Relevance checking to determine whether the basis of RTO calculations has significantly changed (i.e., the process operation remains at the same steady state as before the initiation of RTO calculations).
2. Significance testing of predicted changes to the RTO decision variables (controller set points).
3. Significance testing to determine whether the RTO results will produce a meaningful change in system performance (e.g., process profitability).
4. An examination of the RTO results to determine whether the RTO system has failed and, if so, a determination of the source of the failure.
5. A determination of whether the predicted RTO results will produce sufficient excitation to ensure improving efficacy of the model updating system and, if not, an injection of an excitation signal into the RTO predictions.

If the RTO results fail any of the first four tests they should not be implemented and the RTO calculations should be reexecuted at the appropriate time.

OPTIMIZATION

At the heart of any RTO system is the optimization subsystem, which may be used to solve a number of problems including: data reconciliation, parameter estimation, and the determination of the optimal plant operating conditions or set points. Each of these optimization tasks has similar characteristics and it is these characteristics that guide the selection of appropriate solver technologies. Each of the optimization problems encountered within an RTO system is based on the process model, or at least a portion of it, and most often involves a very small number of degrees of freedom (e.g., less than 100 independent decision variables) in comparison to model size and total number of model variables. Furthermore, the optimization problems encountered within RTO systems are typically, but not always, nonlinear. For the purposes of this discussion, the prototypical optimization problem will be considered to be nonlinear with continuous variables, will be based on a model that may contain hundreds of thousands of variables and equations, will contain inequality constraints and bounds, and will have fewer than 1000 degrees of freedom.

Solution methods for optimization problems that involve only continuous variables can be divided into two broad classes: derivative-free methods (e.g., pattern search and stochastic search methods) and derivative-based methods (e.g., barrier function techniques and sequential quadratic programming). Because the optimization problems of concern in RTO are typically of reasonably large scale, must be solved on-line in relatively small amounts of time and derivative-free methods, and generally have much higher computational requirements than derivative-based methods, the solvers contained in most RTO systems use derivative-based techniques. Note that in these solvers the first derivatives are evaluated analytically and the second derivatives are approximated by various updating techniques (e.g., BFGS update).

There are a range of candidate techniques that could be used in the solution of the optimization problems that arise in RTO and comprehensive treatments of these can be found in a wide range of texts (e.g., Ref.[41–45]). Although early implementation of

RTO was based on solvers such as MINOS, which uses a reduced Lagrangian approach, most solvers used for RTO are based on sequential quadratic programming (SQP).[46]

Sequential quadratic programming is an iterative technique, wherein the optimization problem to be solved is locally approximated by a quadratic programming (QP) problem at each iteration. The solution to the local QP problem yields a search direction along which the next iterate is found. The process of local approximation, solution of the QP approximation for the search direction, and line search are repeated until a solution is found (i.e., convergence tolerances are satisfied). A thorough treatment of SQP can be found in Chapter 18 of Nocedal and Wright.[41] What makes the SQP approach particularly attractive for RTO applications is that the main computational steps (i.e., the solution of the QP approximation and the line search) occur only in the space of the degrees of freedom, which is usually orders of magnitude smaller in dimension that the full optimization problem space. This reduction in the dimension of the search space provides the opportunity to solve larger-scale RTO problems in a reasonable period of time (e.g., between 10 and 20 min) on currently available commercial computing platforms.

There are a number of variations on the basic SQP framework. These can differ in the heuristics that are used in initiating the solver, dealing with feasibility and convergence issues, and in the underlying QP solver. Some of these issues are discussed in Nocedal and Wright.[41] An extensive bibliography on the solution of the QP problems can be found in Gould and Toint.[47] Finally, there are emerging approaches that are based on interior point methods that show promise for RTO applications, (e.g., Ref.[48].) but have not yet been incorporated into commercially available RTO offerings.

FUTURE DIRECTIONS FOR RTO

A successful RTO application results from a series of judicious system design decisions (e.g., sensor system design, process model development, model updating strategy selection, and so forth). Although there has been research in most of these areas and some industrial practice has developed, a number of critical issues remain. In the current design approaches, design decisions, such as the selection of adjustable parameters or the sensor system design, are made either based on experience or using the expected RTO performance for different alternatives. Comprehensive and systematic approaches to making RTO system's design decisions are not yet available. Further, the RTO performance metrics on which such systematic approaches would be based currently use linear approximation to estimate performance and, as a result, are only valid in a small region near the test point. Incorporation of nonlinearity into RTO performance would help the designers make more informed design decisions and may decrease the design effort, because fewer nominal operating points would have to be tested.

Effective RTO design requires that a number of issues be addressed, including measurement selection and sensor system design, model fidelity and updating strategies, and so forth. Much of the available literature has tended to focus on the interdependence of parameter uncertainty and measurements availability and accuracy. A number of critical open questions remain, such as: what level of accuracy is required in the process model or parameter estimates; what is the benefit of improving the accuracy; how can the economic benefit of RTO be assessed in a systematic manner.

A well-designed RTO system will encounter performance degradation over time and therefore, on-line RTO performance monitoring and diagnosis are necessary to detect performance degradation. Although there has been some work in the detection of RTO performance degradation, it is only the beginning and further work is required. Detection of performance degradation or other RTO system faults is only the first step and effective methods to diagnose the root cause of the system failure are crucial to the success of any complex automation system. Plant operation's personnel cannot be expected to possess the skill set to effectively diagnose RTO system faults and it is unlikely operating companies will maintain a high level of RTO expertise. Thus, effective subsystems for RTO fault detection and diagnosis are essential in ensuring the widespread and effective use of the technology.

Requirements for higher performance at the plant-wide and, perhaps, the company-wide level are driving an emerging interest in plant-wide RTO. This is optimization on a very large scale. Such an ambitious undertaking cannot be handled as a simple extension of unit level RTO, as the resulting system would be extraordinarily difficult to build, implement and maintain. Rather than monolithic RTO, a decentralized approach is required. Because most continuous processing plants (e.g., oil refineries, chemical plants, etc.) are highly integrated with respect to the flow of material and energy, the operations of individual process units can influence the behavior of many other units. If a purely decentralized approach was taken where optimization problems were solved independently and the inter-dependence of the process unit operations ignored, the resulting plant-wide operation is most likely to be suboptimal and may not represent an

overall performance improvement. Plant-wide optimum operation could be achieved either by a monolithic RTO system whose scope is the entire plant or by coordinating the interacting unit-level RTO systems. The coordinator acts as a channel for the decentralized RTO systems to communicate with each other and exchange information, so that the interdependence of process unit operations is properly integrated into the decisions made by the unit-level RTO systems. Such a distributed approach to RTO will provide much higher levels of system integrity and reliability, drastically increase the ease of maintainability, while requiring little loss in system performance. A similar coordination scheme can be developed for coordinating the optimization problems vertically across the different layers of the hierarchy shown in Fig. 1.

The speed, accuracy, and reliability of an RTO system largely depend on the optimization algorithms at its heart. Most current SQP algorithms employ active set strategies to solve the optimization problems and these can be inefficient for problems with large numbers of inequality constraints. Application of interior point methods in commercial RTO offerings may hold promise for reducing the computation time in comparison to active set methods, but research is only just beginning to yield interior point algorithms that can challenge the more established approaches in terms of speed. Furthermore, because most RTO applications incorporate nonlinear process models, there is a possibility of multiple local optima. Little consideration has been given to this problem, despite the growing research in global optimization. As computer speed increases and more efficient algorithms become available, this issue should be addressed.

A recent resurgence of interest in self-optimizing or extremum-seeking control may result in the convergence of high-performance control and RTO technologies into a single application. This extremum-seeking research, has merged advanced process control techniques and RTO into a single technology.[49–51] The resulting approach has been termed economic optimizing control (EOC). This EOC approach is a fusion of nonlinear adaptive control and optimization, which achieves a high level of performance by simultaneously steering the process in a direction that maximizes the estimated profit function while continually updating the process parameters. As a result, EOC efficiently tracks optimal process operations as they change due to external driving forces. One of the powerful features of EOC is that it results in a relatively simple control law, which may drastically reduce the on-line computational requirements for RTO. Further, EOC promises to simplify the implementation and maintenance of the overall process automation system. To date, EOC has only been tested on comparatively small-scale simulations, under nearly ideal conditions. A considerable amount of further work is required to understand how this emerging technology will fare in industrial-scale application.

CONCLUSIONS

Real-time optimization or on-line optimization has proven itself to be an effective form of process optimization for delivering lasting economic benefit. The objective of this entry is to provide the reader with a concise, yet comprehensive introduction to RTO that discusses the challenges, benefits, and opportunities for RTO; the range of applicability of RTO; the current status of RTO technology and the research on which this technology is based.

Research and development of RTO technology continues and a number of the key directions were discussed in this entry; however as is pointed out, a vital area that has been largely overlooked, yet is critical for the long-term success of this model-based automation technology, is fault detection and diagnosis for RTO. Real-time optimization systems are composed of a complex, interacting set of subsystems, which are each crucial to the success of the system. The current RTO technology requires specialist skills to troubleshoot system performance problems and other forms of systems' failure. To ensure that the promise of RTO is realized and the widespread use of this automation technology, methods must be developed to aid plant personnel in quickly detecting and accurately diagnosing RTO system failures, without requiring specialty skill sets.

The performance of an RTO system is highly dependent on the effectiveness of the control system that implements RTO commands. The current approach to process optimization and control is to separate the two technologies (i.e., RTO and high performance process control) into two layers that are connected in a hierarchical structure. Some merging of the two technologies has occurred in commercial model predictive control offerings, but only to a limited extent. The emergence of extremum-seeking controllers that are specifically designed to maximize economic performance may represent the next step in the evolution of RTO by providing the framework for merger of process optimization and control.

REFERENCES

1. Desborough, L.; Harris, T. Performance Assessment measures for univariate feedforward/feedback control. Can. J. Chem. Eng. **1993**, *71*, 605–616.
2. Grantom, R.; Royer, D. Ethylene. In *Ullmann's Encyclopedia of Industrial Chemistry*, 5th Ed.;

Wolfgang, G., Ed.; VCH Verlagsgesellshaft mbH: Weinheim, Germany, 1987.
3. Yang, C.H.; Waldman, B. On-line optimization boosts ethylene Profits. Oil Gas J. **1982**, *80*, 104–109.
4. Marlin, T. Process control. In *Designing Processes and Control Systems for Dynamic Performance*; McGraw-Hill: New York, 2000.
5. Skogestad, S. Plantwide Control: the search for the self-optimizing control structure. J. Process Control **2000**, *10*, 487–507.
6. Box, G.E.P.; Chanmugam, J. Adaptive Optimization of continuous processes. I&EC Fund **1962**, *1* (1), 2–16.
7. Bamberger, W.; Isermann, R. Adaptive on-line steady-state optimization of slow dynamic processes. Automatica **1978**, *14*, 223–230.
8. Roberts, P.D. An algorithm for steady-state optimization and parameter estimation. Int. J. Syst. Sci. **1979**, *10* (7), 719–734.
9. Bozenhardt, H. Hyperplane: a case history, Proceedings of the Fifth Annual Control Engineering Conference, May 1986.
10. Golden, M.P.; Ydstie, B.E. Adaptive extremum control using approximate process models. AIChE J. **1989**, *35* (7), 1157–1169.
11. Marlin, T.E.; Hrymak, A.N. Real-time optimization of continuous processes. Fifth International Conference on Chemical Process Control, Lake Tahoe, NV, Jan, 1996, Kantor, J.C., Garcia, C.E. Carnahan B. Eds.; AIChE: New York, 1997, 156–164.
12. Contino, A.V. Improve plant performance via statistical process control. Chem. Eng. **1987**, Jul, 95–102.
13. Narasimhan, S.; Mah, R.S.H.; Tamhane, A.C.; Woodward, J.W.; Hale, J.C. A composite statistical test for detecting changes of steady states. AIChE J. **1986**, *32* (9), 1409–1418.
14. Narasimhan, S.; Kao, C.H.; Mah, R.S.H. Detecting changes of steady states using the mathematical theory of evidence. AIChE J. **1987**, *33* (11), 1930–1932.
15. Box, G.E.P.; Hunter, W.P.; Hunter, J.S. *Statistics for Experimenters*; Wiley: New York, 1978.
16. Durbin, J.; Koopman, S.J. *Time Series Analysis by State Space Methods*; Oxford Univeristy Press: Oxford, U.K., 2001.
17. Kim, I.W.; Kang, M.S.; Park, S.; Edgar, T.F. Robust data reconciliation and gross error detection: the modified MIMT using NLP. Comput. Chem. Eng. **1997**, *21* (7), 775–782.
18. Crowe, C.M.; Gracia Campos, Y.A.; Hrymak, A.N. Reconciliation of process flow rates by matrix projection part I: linear case. AIChE J. **1983**, *29* (6), 881–888.
19. Crowe, C.M. Reconciliation of process flow rates by matrix projection part II: the nonlinear case. AIChE J. **1986**, *32* (4), 616–623.
20. Serth, R.W.; Heenan, W.A. Gross error detection and data reconciliation in steam-metering systems. AIChE J. **1986**, *32* (5), 733–742.
21. Tong, H.; Crowe, C.M. Detection of gross errors in data reconciliation by principal component analysis. AIChE J. **1995**, *41* (7), 1712–1722.
22. Tong, H.W.; Crowe, C.M. Detecting persistent gross errors by sequential analysis of principal components. AIChE J. **1997**, *43* (5), 1242–1249.
23. Tjoa, I.B.; Biegler, L.T. Simultaneous strategies for data reconciliation and gross errors detection of nonlinear system. Comput. Chem. Eng. **1991**, *15* (15), 679–690.
24. Johnson, L.P.M.; Kramer, M.A.; Maximum likelihood data rectification: steady-state systems. AIChE J. **1995**, *41* (1), 2415–2426.
25. Ozyurt, D.B.; Pike, R.W. Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes. Comput. Chem. Eng. **2004**, *28*, 381–402.
26. Forbes, J.F.; Marlin, T.E. Design cost: a systematic approach to technology selection for model-based real-time optimization systems. Comput. Chem. Eng. **1996**, *20* (6/7), 717–734.
27. Fraleigh, L.M.; Guay, M.; Forbes, J.F. Sensor selection for model-based real-time optimization: relating design of experiments to design cost. J. Process Control **2003**, *13*, 667–678.
28. Loeblein, C.; Perkins, J.D. Economic analysis of different structures of on-Line process optimization systems. Comput. Chem. Eng. **1998**, *22* (9), 1257–1269.
29. Zhang, Y.; Nadler, D.; Forbes, J.F. Results analysis for trust constrained real-time optimization. J. Process Control **2001**, *11* (3), 329–341.
30. Bard, Y. *Nonlinear Parameter Estimation*; Academic Press: New York, 1974.
31. Kim, I.W.; Liebman, M.J.; Edgar, T.F. Robust error-in-variables estimation using nonlinear programming techniques. AIChE J. **1990**, *36* (7), 985–993.
32. Kim, I.W.; Liebman, M.J.; Bell, N.H. Parameter estimation for a laboratory water- gas-shift reactor using a nonlinear error-in-variables method. Comput. Chem. Eng. **1991**, *15* (5), 361–367.
33. MacDonald, R.; Howat, C.S.; Data reconciliation and parameter estimation in plant performance Analysis. AIChE J. **1998**, *34* (1), 1–8.
34. Hardin, M.B.; Joshi, A.; Jones, J.D. Rigorous crude unit optimization. NPRA Computer Conference, 1995.

35. Miletic, I.; Marlin, T.E. Results diagnosis for real-time process operations optimizations. Comput. Chem. Eng. **1998**, *22*, S475–S482.
36. Yip, W.S.; Martin, T.E. Multiple data sets for model updating in real-time operations optimization. Comput. Chem. Eng. **2002**, *26*, 1345–1362.
37. Yip, W.S.; Marlin, T.E. Results analysis for a Constrained real-time optimization. International Symposium on Advanced Control of Chemical Processes, ADCHEM 2003, Hong Kong; Jan 2004, 501–506.
38. Halim, A. Detection and Diagnosis of Plant-Model Mismatch for Real-Time Optimization. M.Sc. thesis, University of Alberta, Edmonton, 2003.
39. Pfaff, G. Generating Information for Real-Time Optimization. M.Sc. Thesis, University of Alberta, Edmonton, 2001.
40. Yip, W.S.; Marlin, T.E. Designing plant experiments for real-time optimization systems. Control Eng. Prac. **2003**, *11*, 837–845.
41. Nocedal, J.; Wright, S.J. *Numerical Optimization*; Springer-Verlag: New York, 1999.
42. Bonnans, J.F.; Gilbert, J.C.; Lemaréchal, C.; Sagastizábal, C.A. *Numerical Optimization*; Springer-Verlag: Berlin, 2003.
43. Nash, S.G.; Sofer, A. *Linear and Nonlinear Programming*; McGraw-Hill: New York, 1996.
44. Bertsekas, D.P. *Nonlinear Programming*; Athena Scientific: Belmont, 1995.
45. Fletcher, R. *Practical Methods of Optimization*, 2nd Ed.; Wiley: New York, 1987.
46. Murtagh, B.A.; Saunders, M.A. *MINOS 5.4 User's Guide*, Technical Report SOL 83-20R; Stanford University, 1995.
47. Gould, N.I.M.; Toint, P.L. *A Quadratic Programming Bibliography,* Report 01/02; Rutherford Appleton Laboratory, Computational Sciences and Engineering Department; Chilton, 2003.
48. Wächter, A.; Biegler, L.T. *On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming*, Research Report; IBM T, J. Watson Research Center; Yorktown, 2004.
49. Ariyur, K.B.; Krstic, M. *Real-Time Optimization by Extremum-Seeking Control*; Wiley: New Jersey, 2003.
50. Guay, M.; Zhang, T. Adaptive extremum seeking control of nonlinear dynamic systems with parametric uncertainties. Automatica **2003**, *39* (7), 1283–1294.
51. DeHaan, D.; Guay, M. Extremum seeking control of nonlinear systems with parametric uncertainties and state constraints. Proceedings of American Control Conference, Boston, MA, Jun 30 to Jul 2, 2004.

Recent Advances in Catalytic Distillation

R

Flora T. T. Ng
Department of Chemical Engineering, University of Waterloo, Waterloo, Ontario, Canada

INTRODUCTION

Catalytic distillation (CD) is a novel green reactor technology that combines a heterogeneous catalytic reaction and separation via distillation in a single distillation column. It is an excellent example of process intensification. There are many possible advantages in carrying out a chemical process using CD. They include enhanced product yield and selectivity, reduction of capital and operating costs, enhanced catalyst lifetime, reduction of waste treatment streams, and higher energy efficiency. Owing to the current concern of the impact of greenhouse gases such as carbon dioxide on the environment, a major benefit of CD is related to the utilization of the reaction heat for distillation, which reduces the energy consumption and hence reduces the production of greenhouse gases. This review will give an introduction to the fundamental aspects of CD and its application in different chemical processes such as etherification, hydration and dehydration, alkylation, hydrogenation, desulfurization, esterification, dimerization and oligomerization, and aldol condensation. Some examples of the recent advances and utilization of CD in the chemical and petrochemical industries are provided.

GREEN ENGINEERING AND CD

Green engineering is a new paradigm for the chemical processing industry. The principles of green engineering encompass energy efficiency, reduction of greenhouse gases and waste treatment streams.[1] Green engineering may be achieved in a number of ways including green chemistry, atom efficiency, and process intensification. Process intensification is one of the most important trends in chemical process engineering.[2,3] Integration of reaction and separation is one excellent example of process intensification. Catalytic distillation combines a heterogeneous catalytic reaction and separation via distillation in the same distillation column and is an excellent example of green engineering and process intensification.[4,5] Catalytic distillation is a novel green reactor technology, which has many potential industrial process applications.[4,6] The combination of a reaction (noncatalyzed and homogeneously catalyzed) and separation via distillation is defined as reactive distillation (RD) in this entry and will not be reviewed in detail here.

A schematic configuration for CD is shown in Fig. 1. It consists of a conventional distillation column, a condenser, reboiler, inlets, and outlets for the feed and product. A distinctive feature of the CD column is the placement of heterogeneous catalysts within the distillation column. The nature and placement of heterogeneous catalysts in the CD column is a key parameter to determine in a CD process. Besides catalyzing the reaction, the heterogeneous catalyst could also provide the interfacial area for liquid/vapor separation and distillation. Owing to the high-pressure drop in a distillation column when conventional catalyst particles were placed in it, the commercial implementation of a CD process was not feasible until a method of placing catalyst particles in a fiberglass bag was patented in 1980 by Smith in Texas, U.S.A.[7] This method of placing catalyst particles in a fiberglass bag is also known as ''Texas Tea Bags'' where the fiberglass bags are rolled in bundles and wrapped around with demister wire to provide void space for vapor flow so as to result in a low-pressure drop in the CD process. The fiberglass bags also reduce corrosion in the reactor because the acidic resins will not be in direct contact with the reactor wall. Since the invention of the Texas Tea Bags, a number of new CD packings have been patented or published.[4] There are a number of new patents on novel CD packings but this aspect will be not reviewed in this entry.

Smith also patented the CD process and its application for the production of methyl-*tert*-butyl ether (MTBE).[7–10] In 1981, Charter Oil refinery (now known as Valero Energy) in Houston, TX, U.S.A., demonstrated the first commercial application of CD for the production of MTBE. The commercial success of the application of CD for the production of MTBE led to the development of other CD processes for the refining, petrochemical, and chemical industries. In the past 6 yr, there is an exponential increase in the number of patents and papers appearing on various aspects of CD such as process development, process design and modeling, and the improvement of mass transfer characteristics of the catalyst packing. This entry will provide a review of the benefits and the most significant recent advances and industrial applications of CD.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007965

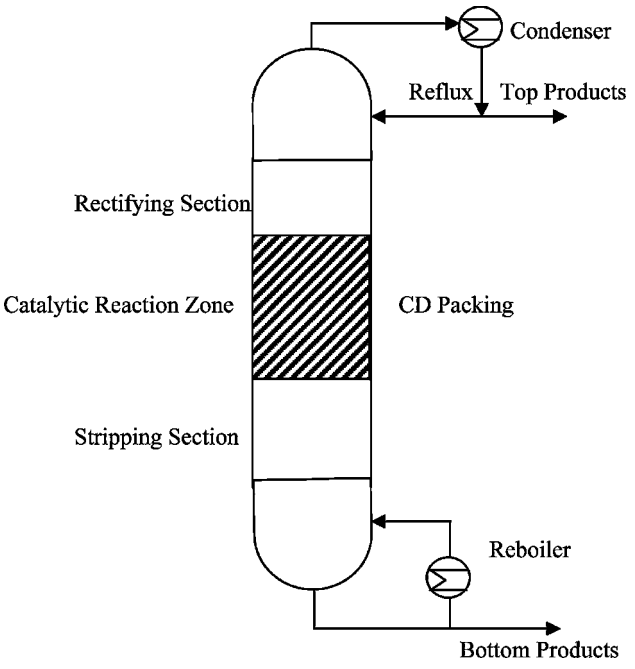


Fig. 1 A schematic for a CD process.

BENEFITS OF CD

Catalytic distillation provides the benefits of green engineering because of the inherent advantages of carrying out a heterogeneous reaction with separation via distillation in the same distillation column. It should be noted that in RD, there is the problem of separation of the homogeneous catalyst and also the corrosion problem associated particularly with the use of acidic or chloride containing homogeneous catalysts. The benefits of combining reaction with separation in CD include energy and capital savings, enhanced conversion and selectivity, longer catalyst lifetime, and reduction of waste streams.[4,6] The following describe the potential benefits in more detail:

1. In situ reaction and separation provides significant energy savings for exothermic processes. The heat generated in the reaction zone in the CD column is used to vaporize the liquid surrounding the catalyst and hence the reboiler duty in the distillation column can be reduced.
2. Process intensification reduces the number of process equipment and hence reduces the capital cost.
3. Increased conversion for equilibrium limited reactions due to the continuous removal of products via distillation as predicted by Le Chatelier's principle.
4. Improved product selectivity for consecutive and parallel reactions due to possible selective removal of the desired product via distillation.
5. Reduction of downstream waste treatment or separation stream due to the enhanced conversion and selectivity.
6. Improved catalyst lifetime due to the use of reaction heat for distillation and hence a reduction of hot spots and sintering of the catalyst. The reflux could also remove the foulant or coke precursors from the catalyst. It is also possible to eliminate the catalyst poisons or impurities by selecting a suitable feed position in the CD column.
7. Less reactor fouling for corrosive type of catalyst because the catalyst used is normally held within some structures and hence will not be in direct contact with the reactor wall.
8. Catalytic distillation could be used to break azeotropes.
9. Catalytic distillation provides a built-in safety feature for exothermic reactions because the reaction heat is used for distillation and hence reduces the risk of runaway reactions. Some CD reactions, e.g., hydrogenation, could be carried out at lower pressure than in the conventional reactor and hence it provides an added safety feature.

GUIDELINES FOR CD PROCESS SELECTION

Catalytic distillation is a multifunctional reactor, which combines a catalytic reaction and separation via distillation in a single distillation equipment. A CD reactor combines the characteristics of a countercurrent (liquid/vapor) flow reactor together with separation via distillation. Although CD has many advantages compared to conventional reactors, not all reactions could take advantage of CD. The following are some of the criteria to consider when evaluating whether a reaction could benefit from being carried out in a CD mode:

1. The reaction preferably occurs in the liquid phase at a reasonable rate because the catalyst particles are mostly wetted by the boiling liquid. The reaction temperature should be less than the critical temperature of at least one of the components because both liquid and vapor phases are required for distillation.
2. The reactants and products can be separated by distillation.
3. Very endothermic processes are not as suitable for CD because the inherent advantage of using the reaction heat for distillation cannot be achieved.
4. A catalyst with high stability and a long lifetime is required because it is normally packed in

some kind of distillation packing and it is very labor intensive and expensive to change it. Generally, the catalyst should have a lifetime of 1–2 yr to be commercially viable.

5. The temperature profile in a CD column is controlled by column pressure and is not isothermal. The kinetics of the reaction should be known prior to the CD experiments so that the CD packing could be located in the section of the column where the temperature is optimal for production of a specified product.

APPLICATIONS OF CD

The first comprehensive review of the application of CD up to the end of 1996 was published by our research group in CHEMTECH.[6] Ng and Rempel also wrote a chapter on CD in the *Encyclopedia of Catalysis*, which included literature information up to the end of December 2000.[4] A review on RD (including both homogeneous and heterogeneous catalyzed reactions) appeared recently.[11] This current review on CD will highlight the commercial CD processes and discuss the advances in CD, particularly since 2000, and includes the literature up to October 2004.

The first commercial and most well-known application of CD was in the production of MTBE. Besides etherification for the production of MTBE, CD could be applied in a number of processes such as alkylation, hydrogenation, isomerization, esterification, desulfurization, aldol condensation, oligomerization, hydration, hydrolysis, amination, and halogenation. Catalytic Distillation Technology (CDTECH), a partnership between ABB Lummus Global and Chemical Research and Licensing, is the leader in the development and commercialization of CD processes particularly related to the refining, petrochemical, and chemical industries. However, there are many more potential applications of CD that could be developed.

Etherification

Because of the reformulated gasoline program introduced by the U.S. government to amend the Clean Air Act in 1990 to reduce emissions from vehicles, oxygenates such as MTBE, ethyl-*tert*-butyl ether (ETBE), and tert-amyl methyl ether (TAME) were favored as some of the components to increase the octane number of gasoline. In 2002, there were over 60 MTBE units and 10 TAME units worldwide. Methyl-*tert*-butyl ether is now phased out in California and possibly banned in the United States nationwide because of its discovery in groundwater. Nevertheless, the experience gained from the successful application of CD for the production of MTBE is applicable to the development of other CD processes and hence will be discussed in some detail here.

Methyl-*tert*-butyl ether is produced from the acid catalyzed reaction between methanol and isobutylene:

$$CH_3OH + CH_2{=}C(CH_3)_2 \leftrightarrow CH_3OC(CH_3)_3 \qquad (1)$$

The first CD packing used in the CD process for MTBE process consisted of an acidic ion-exchange resin such as Amberlyst 15. These macroporous resin beads are packed into fiberglass bags wrapped with demister wire, which are subsequently arranged into bales of Texas Tea Bags that are inserted into a certain location in the CD column.

The advantages of this CD process include increased catalyst lifetime, higher energy efficiency, because the heat released in the adiabatic reactor also preheats the feed and vaporizes some of the feed for the CD column, and a simpler process layout.[4] Another important advantage is that isobutylene conversions up to 99.99% could be achieved in the CD process, whereas only 95–97% isobutylene conversion was achieved with a conventional process configuration (i.e., reactor followed by separation in a distillation column).[4] Although the CD process for MTBE production provides only a slight improvement in the conversion of isobutylene because the equilibrium constant for MTBE favors high isobutylene conversions, it should be emphasized that for a large volume production of fuel additives, even an increase of 2–3% conversion results in significant economic advantages.

An unusual feature of the CD process for MTBE production is that it is recovered from the bottom of the CD column even though its normal boiling point (55°C) is less than the boiling point of methanol (64.5°C). This observation is attributed to the formation of a minimum boiling azeotrope from methanol and MTBE. Apparently, if sufficient quantities of MTBE were accumulated in the CD column, it would ''lift'' the methanol into the reaction zone of the column resulting in a higher methanol conversion. This unusual behavior is believed to be responsible for the multiple steady states observed in the MTBE synthesis shown in process simulation and optimization studies and verified experimentally.[12,13]

Even though new MTBE plants are not being built in North America, CD is used currently for the production of other oxygenates such as TAME or ETBE, which being less soluble in water (i.e., less tendency to be transported in groundwater) are being considered or used as alternative octane enhancers. Ethyl-*tert*-butyl ether is produced from the etherification of ethanol and isobutylene while TAME production requires isoamylene and methanol. A simulation of the industrial production of a ''green'' octane enhancer from

etherification using bioethanol was reported recently.[14] The first commercial TAME plant came onstream at Star Enterprise at Convent, LA, U.S.A., in 1992. It was reported that a CD process for the production of TAME achieves over 90% isoamylene conversion compared to less than 70% conversion obtainable with fixed-bed reactors.[15] A recent patent disclosed an integrated process for the production of TAME from a light naphtha stream derived from a fluid catalytic cracking (FCC) unit.[16] This process involves hydrotreating and isomerization of the light naphtha fraction prior to the reaction of isoamylenes with methanol in the presence of an acidic ion-exchange resin. Recycling of the overhead from an etherification CD column containing unreacted alcohol and isoolefin to improve the conversion of an etherification process was also patented.[17] Catalytic Distillation Technology has a patented CD Taee® technology for the production of *tert*-amyl ethyl ether (TAEE) via the etherification of isoamylene with ethanol. The CD Taee process is based on a two-step reactor design, consisting of a boiling point fixed-bed reactor followed by final conversion in a CD column (Fig. 2). It is stated that the process design retains the heat of reaction as latent heat and reduces the heat input for the subsequent fractionation.

A large number of papers have been published on the process modeling and optimization of the etherification process. More details could be found in a handbook.[18] The most important aspect of process improvement is catalyst improvement because the Amberlyst ion-exchange resin used in the MTBE synthesis has an upper thermal stability limit of less than 100°C and there is a need to develop other acidic catalysts with higher thermal stability. Some of the recent papers have described the use of zeolites, particularly β-zeolites and β-zeolite membranes, as the acidic catalysts for the production of propylene glycol monoethyl ether and ETBE.[19,20]

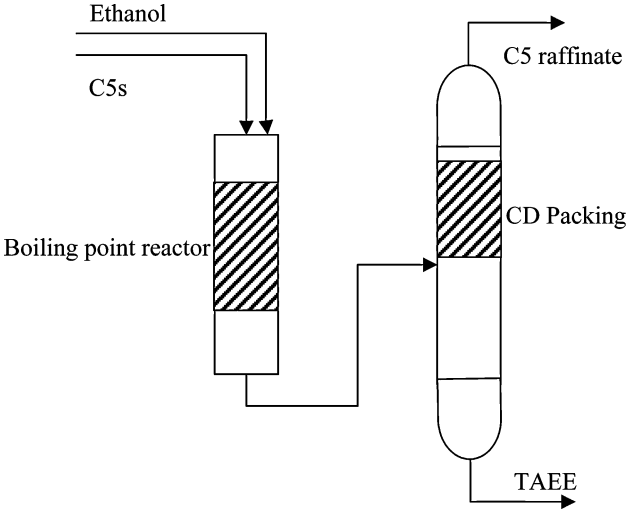


Fig. 2 A schematic of a CD Taee process.

Etherification of propylene and isopropyl alcohol to produce diisopropyl ether, an octane enhancer, has been patented as a two-stage process.[21] The first step involves the hydration of propylene to isopropyl alcohol using acidic ion-exchange resins or acidic zeolites and an optional cosolvent and the second step involves the etherification of propylene and isopropyl alcohol using an acidic catalyst such as Amberlyst 36 in a CD column.

A potential commercial development of oxygenates is related to the use of ethers to enhance the cetane number of diesel fuel. Because the solubility of ethers in water decreases with an increase in the molecular weight, there is a potential application of higher-molecular-weight oxygenates in the diesel fuel pool as their negative environmental impact on the groundwater diminishes should any leakage or spillage occur.

Hydration

Hydration of olefins to alcohols is equilibrium limited and hence CD is potentially suitable for such applications. The catalysts used for the process are acidic catalysts such as cation-exchange resins or zeolites. The hydration of isobutylene to produce *tert*-butyl alcohol via CD results in a higher conversion and there is no need to recycle the water.[22] The hydration process is catalyzed by acidic ion-exchanged resins at 85°C and about 1200 kPa. The CD process configuration involves feeding the isobutylene below the catalyst zone and the water is fed above the catalyst zone. Flooding of the reaction zone is introduced in the process to improve the contact between the catalyst and the liquid and to ensure that the water is in constant contact with the catalyst sites. Flooding of the catalyst zone apparently improves the catalyst lifetime and performance because catalyst deactivation is caused by mass transfer and liquid distribution problems. Some recent publications on the hydration of isobutylene include a patent and a study of the kinetics of the hydration process and discussions on the merits of the application of CD for hydration.[23,24]

Hydration of propylene to isopropyl alcohol or hydration of ethylene to ethanol requires much more severe reaction conditions, which are not as suitable for distillation unless a more active catalyst could be developed. A recent design of a CD process for the production of isopropyl alcohol via hydration of propylene indicates that high-purity isopropanol (99.9%) could be achieved operating at 2 MPa; however, the viability of the catalysts proposed for the process is yet to be determined.[25]

The hydration of ethylene oxide is a consecutive reaction that produces diethylene glycol from the consecutive reaction of ethylene glycol with ethylene oxide. Therefore, CD could be used to produce the desired ethylene glycol in high yield and selectivity. Recently, the hydration of ethylene oxide to produce ethylene glycol with conversion close to 100% and >98% selectivity was reported.[26] The synthesis of cyclohexanol by the selective hydration of cyclohexene in a mixture of benzene, cyclohexene, and cyclohexane with very close boiling points was reported.[27] This process configuration involves feeding a slurry containing hydrophilic H-ZSM5 zeolite particles into an RD column. The catalyst particles were recovered in the aqueous phase.

Dehydration

The production of diisopropyl ether by the dehydration of isopropyl alcohol and the simultaneous removal of the product diisopropyl ether and water at the reaction zone was reported to increase the conversion of this equilibrium limited reaction.[21] The dehydration of methanol to produce dimethyl ether was also reported.[28]

The dehydration of *tert*-butanol to produce isobutylene using β-zeolite, hydrofluorine (HF)-treated zeolite, and HF-treated montmorillonite has been reported to give a higher conversion of *tert*-butanol due to the in situ separation via CD.[29–31] It was reported that at a pressure of 80 kPa and 64–99°C using β-zeolite, conversion of 99% and selectivity of 94.5% were obtained.[31] The preparation of isobutene from a slurry CD process by feeding fine particles of acid cation-exchange resins with the *tert*-butanol was reported to give a 99.99% conversion.[32] Recently, the production of high-purity (>98%) tetrahydrofuran from the dehydration of 1,4-butanediol using cation-exchange resins was reported. Furthermore, this process was carried out using a new technique involving the use of a dual column operating at different pressures.[33,34]

Alkylation

An important industrial application of CD is the alkylation of benzene with ethylene or propylene to produce ethylbenzene or cumene, respectively, using acidic ion-exchange resins such as Amberlyst or zeolites operating at 130–5065 kPa and 80–500°C.[35] Cumene is a chemical intermediate for the production of phenol, acetone, and alpha-methyl styrene, which are used to produce resins and solvents. Ethylbenzene is an intermediate for styrene, an important monomer for polymers. Alkylation of benzene could also be used to reduce the carcinogenic benzene content of gasoline. A number of advantages of CD were obtained for the exothermic alkylation process and particularly noteworthy is the increased catalyst lifetime and enhanced selectivity to monoalkylated rather than dialkylated or trialkylated product. Catalytic Distillation Technology commercialized the production of ethylbenzene using the CD EB® technology in 1994 at the Mitsubishi Petrochemical in Yokkaichi, Japan. The CD Cumene® process was first brought onstream in 2000 at a capacity of 270,000 MTA by Formosa Chemicals and Fibre Corporation, Taiwan, and was expanded to double the capacity since 2004.

A schematic of a CD process for cumene production is shown in Fig. 3. The CDTECH process uses a zeolite catalyst in one of its patented CD structures and the product yield exceeds 99.5% purity with 99.9% selectivity to cumene. The high selectivity to cumene is achieved by controlling a low propylene concentration in the reaction section using a combination of process parameters such as system pressure, location of catalyst zone, and feedpoint. A low propylene concentration will result in a low propylene oligomerization rate and hence will reduce the amount of diisopropylbenzene and triisopropylbenzene produced by the consecutive reactions. One interesting aspect of this process is to recycle the diisopropylbenzene and triisopropyl benzene where transalkylation with benzene produces more cumene.

Recently, a novel CD process for the alkylation of benzene with propylene using suspended catalyst rather than encasing the catalyst in a rigid structure with 100% conversion for propylene with more than 90% selectivity to high-purity cumene was reported.[36] An improvement of the suspension CD by simultaneous alkylation and transalkylation for producing

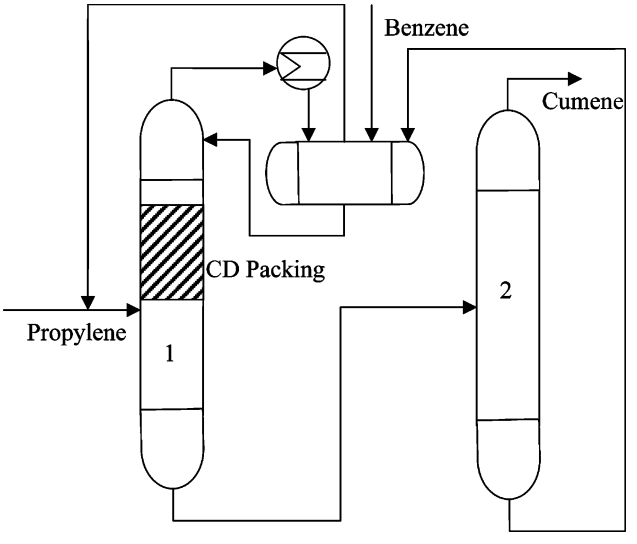


Fig. 3 A schematic of a CD process for cumene production: 1) CD column and 2) distillation column.

cumene was also reported using a modified zeolite catalyst.[37] Suspension CD involves the use of small catalyst particles inside the distillation column and the separated catalyst particles can be recycled and/or regenerated before recycling. Although suspension CD overcomes the difficulty of getting suitable CD packings, it requires the separation and recycling of the catalyst particles, which will increase the cost of the process. Currently, there is no commercial process using suspension CD.

The alkylation of benzene with C2–C4 olefins has also been patented using the acid form of the β-zeolite.[38] Selective alkylation of toluene with C2–C4 olefins using acidic zeolites such as Y-, β-, or Ω-zeolites followed by transalkylation with benzene has also been patented.[39] A recent paper reported an increased catalyst lifetime for the alkylation of benzene with ethylene at a relatively high ethylene to benzene ratio using Y-zeolite contained in high-efficiency packing.[40] This paper clearly shows that many of the advantages of CD are due to the fact that selectivity greater than 99.5% to ethylbenzene at a benzene to ethylene molar ratio of 1.5 : 2 was achieved. A recent patent on the use of enriched or pure ethylene at a benzene/ethylene ratio of 1 : 1 for the production of ethylbenzene was granted.[41]

The alkylation of FCC off-gas with benzene to produce ethylbenzene by CD was recently reported.[42] A rectified RD configuration could be used to increase the yield of monoalkylated product.[43]

Besides the production of cumene and ethylbenzene, there are a number of recent reports on the production of linear alkylbenzene, precursors to detergents, via the alkylation of benzene with C6–C18 olefins. One process uses suspension CD and essentially 100% conversion of olefin at low temperatures of 90–100°C was obtained.[44] An HF-treated mordenite used in the alkylation of benzene and C10–C14 olefins was found to give a 74–84% selectivity to linear alkylbenzene containing 80% 2-phenyl isomer.[45] A new patent on the alkylation of aromatic hydrocarbons such as benzene and cumene with straight-chain C6–C20 olefins on acidic catalyst such as zeolites or fluorine-treated zeolite catalyst packed in a Katamax-type packing was granted.[46] A patent application on the manufacture of xylenes from reformate by RD also appeared and higher than equilibrium amounts of *para*-xylene were claimed.[47]

Hydrogenation

Catalytic distillation hydrogenation is one of the more recent applications of CD that was commercialized by CDTECH for the selective hydrogenation of dienes in C4–C6 streams and the saturation of benzene in the aromatics under the trade name CD Hydro®. Commercial applications of CD Hydro include selective hydrogenation of butadiene in a mixed C4 refinery stream, selective hydrogenation of pentadiene and hexadiene, hydrogenation of benzene to cyclohexane and hydrogenation of acetylene in a C4 stream.[48] A process for the hydrogenation of benzene in a reformate stream was commercialized in Texaco's refinery at Bakersfield, TX, U.S.A., to produce gasoline with less than 10 ppm benzene in 1995. Conventional Ni or Pd catalyst supported on alumina placed in a distillation packing consisting of flexible, semirigid open-mesh tubular element was used as CD packing. The process operates at 1378–1723 kPa and 149–204°C compared to 2412–3101 kPa and 260°C for a fixed-bed hydrogenation process. The highly exothermic reaction is controlled by recycling the condensed overhead to the column and therefore safer than conventional fixed-bed hydrogenation processes, which require a high mass flow rate or efficient heat exchangers to remove the reaction heat. The heat of hydrogenation could also be used for vaporization of the feed in the catalyst zone resulting in a near-isothermal operation, which improves catalyst lifetime since the sintering of catalyst is reduced. It was reported that the CD Hydro process could be installed at about 25% less capital cost than a conventional reformate splitter followed by a fixed-bed hydrogenation unit.

It should be noted that there are a number of benefits associated with a hydrogenation process carried out in a CD column. Besides significant energy savings and hence reduction of greenhouse gases such as carbon dioxide, there is improved process safety due to the lower temperature and pressure required for a CD hydrogenation process and the near-isothermal temperature profile in the reaction zone, which reduces the possibility of runaway reactions. The lower pressure requirement also eliminates the need for a hydrogen compressor. In a recent patent on the hydrogenation of cyclopentadiene, it was shown that the pressure required for hydrogenation in a CD column was 8 psig and 175°F whereas greater than 200 psig and 200–450°F were used in a conventional hydrogenation reactor to achieve 99% conversion.[49] It was also reported that the catalyst was found to be more stable in the selective hydrogenation of methylacetylene and propadiene in a C3 fraction in a CD column than in a trickle-bed reactor.[50] An extended catalyst lifetime in the hydrogenation of C3 fraction was also reported.[51] The phenomena of lower hydrogen pressure and temperature required for hydrogenation in a CD column is particularly noteworthy. Experimental and modeling studies are being carried out by our research group to understand this distinctive feature of CD hydrogenation. This is likely related to the enhanced mass transfer of hydrogen to the catalyst in a CD process.

In 2001, a CD process for the production of 2,2,4-trimethylpentane with 100 octane rating from the dimerization of isobutene and hydrogenation of the octenes using a Pd/cation-exchange resin was disclosed, but no data were given in the patent.[52]

Hydrogenation in a CD column has been now applied to the hydrocracking and hydrotreating processes for vacuum gas oil to produce diesel and lighter distillates using different conventional hydrogenation catalysts such as Ni/Mo on alumina and Pd/alumina/zeolite.[53] Another related patent disclosed that hydrocracking together with posttreatment of the hydrocracked fraction via RD resulted in the reduction of hydrogen consumption and reduction of the overall reactor and catalyst volumes for a given level of performance.[54] It appears that the application of CD to the treatment of petroleum fractions will gain more attention in the near future.

An improvement on the catalytic hydrogenation of acetylenes and dienes in the C2–C5 fraction in a thermally cracked feed stream without significantly hydrogenating the C2 and C3 olefins was achieved using a combination of CD and fixed-bed catalytic steps.[55]

Desulfurization

With the new legislation on the reduction of S content in gasoline and diesel worldwide, production of low-S gasoline and diesel fuel is of great importance. Catalytic distillation was found to be particularly effective for the desulfurization of naphtha and gasoline produced from FCC.[4,56] In 2000, a new commercial CDTECH process for the desulfurization of FCC gasoline (CD HDS®) came onstream at Motiva in Texas, U.S.A. CD HDS takes advantage of the fact that the light olefins are fractionated to the top of the column while the benzothiophenes and substituted thiophenes are fractioned to the bottom of the column. Therefore, CD HDS minimizes the hydrogenation of light olefins and preserves the octane number of gasoline. The catalysts used are conventional hydrodesulfurization (HDS) catalysts such as Co/Mo and Ni/Mo on alumina supports, which are put inside catalyst structures composed of flexible, semirigid open-mesh tubular material such as stainless steel wire mesh.[57] A CD process, CD Hydro/CD HDS, utilizing two-distillation columns for the desulfurization of a naphtha stream with a minimum loss of olefins and octane number was disclosed.[58] The naphtha is fed to the first column containing hydrogenation catalysts such as Ni sulfide and Pd oxide to catalyze the reaction of mercaptans with some diolefins to form olefinic thioethers, which are then sent to the second distillation column together with the heavy sulfur compounds where desulfurization occurs (Fig. 4). This CD process produces gasoline with a minimal mercaptan content and eliminates the caustic treatment process used to remove mercaptans. It was reported that a combination of CD Hydro and CD HDS could reduce the FCC gasoline sulfur by 90% while the octane loss is very minimal.

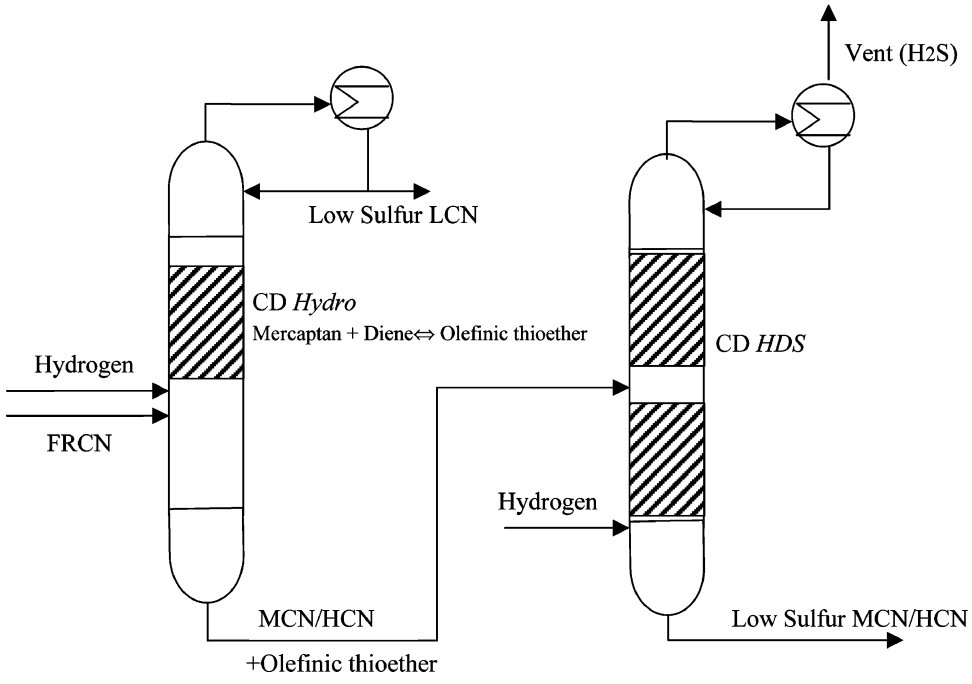


Fig. 4 A schematic of a CD Hydro/CD HDS process for production of low-sulfur gasoline. FRCN, full-range catalytic naphtha; LCN, light catalytic naphtha; MCN, medium catalytic naphtha; HCN, heavy catalytic naphtha.

This combined process could produce low-sulfur gasoline at half the current cost of desulfurization and won the Brian Davis Refining Technology award in 1999. The start-up of the first CD Hydro/CD HDS unit was at Irving Oil's St John, New Brunswick Refinery, Canada, in 1999 and produced less than 150 ppm S.[59] In 2000, Irving Oil received a U.S. Environmental Award. In 2004, the CD Hydro/CD HDS process was used to process about 1.0 million barrels per day of gasoline for mercaptans and sulfur removal in refineries all over the world.

Recently, a review of the production of ultra-low-sulfur gasoline, a patent application for the desulfurization of full-range naphtha by thioetherification and HDS, and the sulfidation of HDS catalysts utilizing CD also have appeared.[60–62] A recent paper on the comparison between the HDS in a conventional trickle-bed reactor and CD for light gas oil HDS concluded that the CD process for HDS is more efficient because the distillation process reduces the inhibitory effects of the H_2S and allows the HDS of the more refractory sulfur compounds to produce a low-sulfur diesel fuel and enhanced catalyst lifetime.[63] Application of CD for the HDS of diesel fuel is more challenging owing to the high boiling point of the diesel fraction although a design for the ultra-low-sulfur diesel production was published.[64]

A stripped RD column could be applied in the hydrocracking and hydrotreating of petroleum to avoid excessive hydrocracking or hydrotreating of lighter products.[43] If the desired products are of higher molecular weight, a rectified RD column can be used.

Esterification

There are many examples of the application of CD or RD for esterification.[4,11] Esterification of methanol or ethanol with acetic acid forms methyl acetate or ethyl acetate, respectively. Methyl acetate is important in the manufacture of polyesters and is an important solvent for cellulose while ethyl acetate is used in inks, fragrances, and pharmaceuticals. The manufacture of high-purity methyl acetate is difficult because of the equilibrium limitation and also the formation of azeotropes. The production of methyl acetate by Eastman Chemical Co. was the first commercial application of RD using a homogeneous liquid acid catalyst. Only one RD column and two smaller columns for processing sidestreams are required while in the conventional methyl acetate synthesis, two reactors and eight distillation columns are required.

Catalytic distillation was also used to produce methyl acetate. A macroporous acidic ion-exchange resin fixed in an open cross-flow structure packing was used in the CD column where the acetic acid was fed above the catalyst zone and the methanol was fed below the catalyst zone.[65] Because the acidic ion-exchange resin provides at least 20 times more protons per unit volume of reacting liquid than a homogeneous catalyst, the residence time to achieve a given level of conversion can be reduced. The process operates at atmospheric pressure with virtually complete conversion of methanol and acetic acid in equimolar proportions and only 10 theoretical plates are required. The CD process also eliminates the need for downstream treatment of liquid acid. A pilot scale production of butyl acetate in a CD process using a Katapak in combination with a CY packing from Sulzer Chemtech containing an acidic ion-exchange resin was reported.[66]

In the past few years, a large number of CD processes on esterification using Katapak packing such as Katapak-S or Multipak packing together with solid acid catalyst such as Amberlyst 15 have been reviewed in detail and hence these processes will not be reviewed here.[11] All these processes operate at mild temperature and pressure with conversions and selectivities in excess of 95% and in some cases close to 100%.

Dimerization and Oligomerization

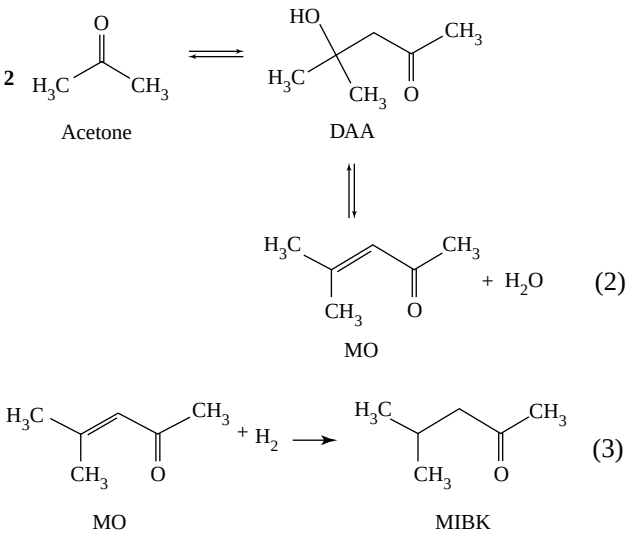
The oligomerization of olefins is an exothermic consecutive reaction, which benefits from the application of CD for enhanced selectivity to intermediate products. Catalytic distillation plays a particularly important role in enhancing the catalyst lifetime because in situ separation reduces the undesirable high-molecular-weight oligomers or polymers, which will form coke and deactivate the catalyst. The use of reaction heat for distillation also reduces the formation of hot spots and catalyst deactivation due to sintering.

Because of the phase out of MTBE, octenes will take on a more important role in the gasoline pool. The selective dimerization of C4 in a raffinate stream to octenes from the steam cracking of naphtha could be used to produce octane enhancers. In addition, there is an existing source of isobutylene due to the phase out of MTBE, which could be selectively dimerized to 2,4,4-trimethylpentene-1 and 2,4,4-trimethylpentene-2, which on hydrogenation will produce 2,2,4-trimethylpentane with an octane number of 100. Smith et al. have patented a CD process for dimerization using an acidic cation-exchange resin.[67] A patent application on the dimerization of light olefins using a combination of RD and conventional reactors was filed.[68] We have studied the oligomerization of butenes in a CD column. We found that when a Ni-zeolite was encased in a fiberglass bag as a CD packing, owing to the mass transfer resistance of the fiberglass bags, rapid catalyst deactivation due to the formation of

higher oligomers on the catalyst was observed. A novel CD packing containing Ni was found to be very effective for the selective dimerization of butenes to octenes.[69] Such CD packing containing Group VIII metals such as Pd was also very active for the hydrogenation of the C8 alkenes. We found that a combination of dimerization and hydrogenation CD packing provides a high conversion of isobutylene with high selectivity to 2,2,4 trimethylpentane and long catalyst lifetime time (Fig. 5). It was also reported that the dimerization of isobutene was more selective in the presence of *tert*-butyl alcohol.[70] A patent application on the production of oligomer from isobutane with CD that combines dehydrogenation and oligomerization was reported.[71] A process for the dimerization of isobutene and the hydrogenation in the presence of S compounds was patented.[72] Recently, the use of CD to produce a lubebase stock from lower-molecular-weight feedstock and acidic catalyst such as zeolites was disclosed.[73] A CD process for improving yields of higher-molecular-weight olefins from lower-molecular-weight olefins using isomerization and disproportionation was also patented.[74]

Aldol Condensation and the Production of Methyl Isobutyl Ketone

Aldol condensation of acetone produces diacetone alcohol (DAA), an environmentally friendlier solvent due to its low volatility and high boiling point, 165.6°C. Aldol condensation of acetone to produce DAA is strongly limited by equilibrium. At the boiling point of acetone, the equilibrium conversion of acetone is only 4.3%. Diacetone alcohol also undergoes dehydration to mesityl oxide (MO).



Experimental studies in our laboratory on the aldol condensation were carried out in a 1 in. CD column using Ambelite IRA-900 anion-exchange resin housed in fiberglass bags. The reboiler duty, which affected the flow rates, was found to play an important role in the selectivity to DAA.[75,76] A rate-based three-phase CD model was developed, which accurately predicts the yield and selectivity obtained under steady-state and transient conditions.[77–79] Model predictions and experimental data indicate that the production of DAA is external mass transfer controlled while the production of MO is kinetically controlled. The external mass transfer resistance was caused by the fiberglass bags. Recently,

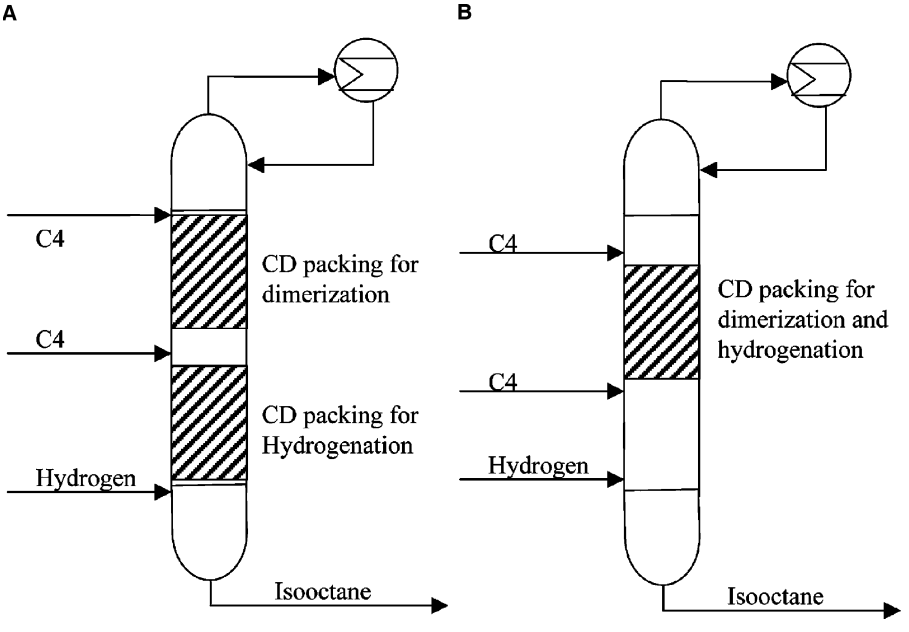


Fig. 5 A schematic of a CD process for the production of isooctane. (A) Dimerization and hydrogenation in two separate zone. (B) Dimerization and hydrogenation in the same zone.

we prepared beryl saddles coated with MgO and used it as CD packing without any fiberglass bags; a higher selectivity to DAA was obtained.

A one step synthesis of methyl isobutyl ketone (MIBK) via the aldol condensation of acetone and the in situ hydrogenation of MO was patented.[80,81] However, it is clear that a CD process for a one-step synthesis of MIBK from acetone and hydrogen is very complicated because MO can undergo a large number of reactions such as the production of higher-molecular-weight phorones besides being hydrogenated to MIBK. In addition, the presence of azeotropes and phase separations due to the production of water from the dehydration of DAA makes the one-step synthesis of MIBK from the aldol condensation of acetone very challenging.

OTHER CD APPLICATIONS

Besides the above CD processes, some of the recent novel applications of CD are outlined below. The production of amines from the hydrogenation of aniline and the selective production of diethanolamine from the reaction of monoethanolamine and ethylene oxide have been reviewed.[4] A patent on the production of phenol from cumene hydroperoxide disclosed that solid acid catalysts such as zeolites, ion-exchange resins achieved 100% conversion with about 60% selectivity to phenol at 50–90°C and 0–10 psig.[82] This process utilizes the heat of the decomposition of cumene hydroperoxide to effect the separation of the lower boiling components and hence reduces the energy cost and carbon dioxide emissions.

An interesting application of CD for the production of methanol from syngas (a mixture of CO and H_2) was patented in 1999.[83] Recently, another process on the methanol production from syngas using a CD column and multiple distillation stages, at least three, was disclosed.[84] This process utilizes CD to increase the conversion of synthesis gas to methanol beyond the equilibrium limitation using catalysts such as Cu–Zn, or Raney™ in the form of sponge.

A patent on the application of CD to produce alpha-olefins via the isomerization of internal olefins was granted.[85]

A new CD process for the production of vinyl acetate from acetic acid, ethylene, and oxygen using a Pd-type catalyst at 338–420 K, 2–5 bar was disclosed.[86] This illustrates the wide-ranging possibilities for the application of CD in a variety of processes for the chemical, petrochemical, and petroleum industry. The production of acetic acid from the carbonylation of dimethyl ether or methanol using RD and homogeneous catalyst was also patented.[87]

Applications of CD for Separations

Catalytic distillation can also be used for selective separations such as the separation of piperidine from *n*-amylamine, separation of isobutylene in a C4 stream, and removal of acetic acid from dilute aqueous streams.[4] The application of CD for separations will not be reviewed in this article. The potential use of RD for the separation of chiral compounds is very noteworthy although no corresponding CD process was reported.[88]

CONCLUSIONS

Catalytic distillation is a rapidly developing field with applications in many processes for the chemical, petrochemical, and petroleum industry. It is emerging as a tool for green process and technology innovations because it utilizes process intensification to achieve energy efficiency and the reduction of greenhouse gases.

With the advances in the fundamental understanding of the reaction engineering aspects of CD based on both experimental research and mathematical modeling, CD is becoming a valuable tool for the development of new green chemical and petrochemical processes. Some challenges for CD include the development of robust catalysts or CD packings that will have the mechanical strength to withstand high liquid/vapor flow rates in a distillation column and modeling the multiphase flow characteristics in a CD column using computational fluid dynamics. Another important aspect is to understand the behavior of noncondensable gases such as hydrogen in the CD hydrogenations. We are currently carrying out both experimental and mathematical modeling of a new CD process for the production of isooctane from isobutene.

REFERENCES

1. Anastas, P.T.; Zimmerman, J.B. Design through the 12 principles of green engineering. Environ. Sci. Technol. **2003**, *37* (3), 94A–101A.
2. Dautzenberg, F.M.; Mukherjee, M. Process intensification using multifunctional reactors. Chem. Eng. Sci. **2001**, *56*, 251–267.
3. Stankiewicz, A. Reactive separations for process intensification: an industrial perspective. Chem. Eng. Process. **2003**, *42*, 137–144.
4. Ng, F.T.T.; Rempel, G.L. Catalytic distillation. In *Encyclopedia of Catalysis*; John Wiley & Sons Inc.: New York, 2002.

5. Malone, M.F.; Huss, R.S.; Doherty, M.F. Green chemical engineering aspect of reactive distillation. Environ. Sci. Technol. **2003**, *37*, 5325–5329.
6. Podrebarac, G.G.; Ng, F.T.T.; Rempel, G.L. More uses for catalytic distillation. CHEMTECH **1997**, *27*, 37–45.
7. Smith, L.A., Jr. Catalytic Distillation Process U.S. Patent 4,232,177, Nov 4, 1980.
8. Smith, L.A., Jr. Catalytic Distillation Process U.S. Patent 4,307,254, Dec 22, 1981.
9. Smith, L.A., Jr. Catalytic Distillation Process and Catalyst, Canadian Patent 1,125,728, Jun 15, 1982
10. Smith, L.A., Jr. Catalytic Distillation Process U.S. Patent 4,336,407, Jun 22, 1982.
11. Hiwale, R.S.; Bhate, N.V.; Mahajan, Y.S.; Mahajani, S.M. Industrial applications of reactive distillation: recent trends. Int. J. Chem. Reactor Eng. **2004**, http:/bepress.com/ijcre/vol2/R1.
12. Taylor, R.; Krishna, R. Modelling reactive distillation. Chem. Eng. Sci. **2000**, 5183–5229.
13. Mohl, K.; Kienle, A.; Gilles, E.; Rapmund, P; Sundmacher, K.; Hoffmann, U. Steady-state multiplicities in reactive distillation columns for the production of fuel ethers MTBE and TAME: theoretical analysis and experimental verification. Chem. Eng. Sci. **1999**, *54*, 1029–1043.
14. Quitain, A.T.; Itoh, H; Goto, S. Reactive distillation for synthesizing ethyl tert-butyl ether from bioethanol. In *Reaction Engineering for Pollution Prevention*; Abraham, M.A., Hesketh, R.P., Eds.; Elsevier Science B.V.: Amsterdam, Netherlands, 2000; 237–246.
15. Quitain, A.; Itoh, H.; Goto, S. Industrial-scale simulation of proposed process for synthesizing ethyl tert-butyl ether from bioethanol. J. Chem. Eng. Jpn. **1999**, *32*, 539–543.
16. Adams, J.R.; Smith , L.A., Jr.; Hearn, D; Jones, E.M., Jr.; Arganbright, R.P. Integrated Process for the Production of Tame U.S. Patent 5,792,891, Aug 11, 1996.
17. Bakshi, A.; Hickey, T.P. Etherification Process U.S. Patent 5,919,989, Jul 6, 1999.
18. Hamid, S.H. *Handbook of MTBE and Other Gasoline Oxygenates*; Marcel Dekker, 2004.
19. Wang, L.; Li, J. Study on synthesis of propylene glycol monoethyl ether with β zeolite as catalysts by catalytic distillation. Jingxi Shiyou Huagong **2003**, *5*, 17–19.
20. Yu, S-B.; Li, Y-H.; Chen, H-F. ETBE reactive distillation on supported β zeolite membrane. Shiyou Zuebao Shiyou Jiagong **2003**, *19* (5), 58–62.
21. Marker, T.L.; Frank, G.A.; Barger, P.T.; Hammershaimb, H.V. Two-Stage Process for Producing Diisopropyl Ether Using Catalytic Distillation U.S. Patent 5,744,645, Apr 28, 1998.
22. Smith, L.A., Jr. Method for Operating a Catalytic Distillation Process U.S. Patent 5,221,441, Jul 22, 1993.
23. Reusch, D.; Beckmann, A.; Nierlich, F.; Tuchlenski, A. Catalytic Hydration Procedure for the Production of Tert-Butanol from Isobutylene and Water and using Reactive Distillation German Patent DE10260991(A1), Jul 8, 2004.
24. Zhang, C.M.; Adesina, A.A.; Wainwright, M.S. Isobutene hydration over Amberlyst-15 in a slurry reactor. Chem. Eng. Proc. **2003**, *42* (12), 985–991.
25. Xu, Y.; Chuang, K.T.; Sanger, A.R. Design of a process for production of isopropyl alcohol by hydration of propylene in a catalytic distillation column. Inst. Chem. Eng. Trans. IChem. E **2002**, *80*, Part A, 686–694.
26. Liu, H.; Qu, Y.; Wang, W. Simulation of hydration process of water and ethylene oxide by reactive distillation to produce glycol. Beijing Huagong Daxue Xuebao Ziran Kexueban **2002**, *29* (1), 18–20.
27. Frank, S.; Qi, Z.; Sundmacher, K. Synthesis of cyclohexanol by three-phase reactive distillation: influence of kinetics on phase equilibria. Chem. Eng. Sci. **2002**, *57* (9), 1511–1520.
28. An, W.; Chuang, K.T.; Sanger, A.R. Dehydration of methanol to dimethyl ether by catalytic distillation. Can. J. Chem. Eng. **2004**, *82* (5), 948–955.
29. Abella, L.C.; Gaspillo, P.D.; Itoh, H.; Goto, S. Dehydration of tert-butyl alcohol in reactive distillation. J. Chem. Eng. Jpn. **2000**, *33* (2), 351.
30. Gotze, L.; Bailer, O.; Moritz, P.; Von, S. Reactive distillation with KATAPACK. Catal. Today **2001**, *69* (1–4), 201–208.
31. Knifton, J.F.; Sanderson, J.R.; Stockton, M.E. Tert-butanol dehydration to isobutylene via reactive distillation. Catal. Lett. **2001**, *73* (1), 55–57.
32. Xue, Y.; Xu, J.; Xu, C.; Zhou, M.; Sheng, Z.; Shang, J. Preparation of isobutene by slurry catalytic distillation process. Huaxue Gongye Yu Gongcheng (Tianjin, China) **2003**, *20* (5), 251–255, 274.
33. Liu, Q.; Xiao, J. Synthesis of tetrahydrofuran from butanediol by pressure-sensitive reactive distillation. Jisuanji Yu Yingyong Huaxue **2001**, *18* (2), 123–126.
34. Liu, Q; Zhang, F.; Gao, H. Synthesis of tetrahydrofuran from butanediol by reactive distillation. Huaxue Gongcheng **2002**, *30* (2), 75–78.
35. Smith, L.A., Jr. Alkylation of Organic Aromatic Compounds U.S. Patent 4,849,569, Jul 19, 1989.
36. Wen, L.; Min, E.; Pang, G.; Yu, W. Synthesis of cumene by suspension catalytic distillation process. Huagong Xuebao (Chinese Ed.) **2000**, *51* (1), 115–119.

37. Lei, Z.; Li, C.; Li, J.; Chen, B. Suspension catalytic distillation of simultaneous alkylation and transalkylation for producing cumene. Sep. Purif. Technol. **2004**, *34* (1–3), 265–271.
38. Hendirksen, D.E.; Lattner, J.R.; Johannes, M.; Janssen, M.J.G. Alkylation Process Using Zeolite Beta U.S. Patent 6,002,057, Dec 14, 1999.
39. Chen, J. Process for the Production of Alkyl Benzene U.S. Patent 5,866,736, Feb 2, 1999.
40. Qi, Z.; Zhang, R. Alkylation of benzene with ethylene in a packed reactive distillation column. Ind. Eng. Chem. Res. **2004**, *43* (15), 4105–4111.
41. Zhang, J.; Li, D.; Fu, J.; Cao, G. Process and Apparatus for Preparation of Ethylbenzene by Alkylation of Benzene with Dilute Ethylene Contained in Dry Gas by Catalytic Distillation U.S. Patent 6,504,071, Jan 7, 2003.
42. Xu, L.; Wang, Q.; Liu, W.; Xie, S.; Sun, X.; Bai, J.; Zhang, S.; Wang, C. Thermodynamics of ethylbenzene synthesis through alkylation of FCC off-gas with benzene by catalytic distillation. Cuihua Xuebao **2003**, *24* (1), 73–78.
43. Tung, P. Dimensions in Reactive Distillation Technology U.S. Patent 6,500,309, Dec 31, 2002.
44. Wang, E.; Li, C. Simulation of suspension catalytic distillation for the synthesis of linear alkyl benzene. Chin. J. Chem. Eng. **2003**, *11* (5), 520–525.
45. Knifton, J.F.; Anantaneni, P.R.; Dai, P.E.; Stockton, M.E. Reactive distillation for sustainable, high 2-phenyl LAB production. Catal. Today **2003**, *79–80*, 77–82.
46. Winder, J.B.; Wharry, D.L.; Schell, J.R.; Brown, M.J.; Murray, J.L.; Howe, R.C.; Sorensen, W.L.; Szura, D.P. Reactive Distillation Process for the Alkylation of Aromatic Hydrocarbons U.S. Patent 6,642,425, Nov 4, 2003.
47. Feng, X.; Buchanan, J.S.; Crane, R.A.; Dakka, J.M.; Iaccino, L.L.; Mohr, G.D. Manufacture of xylenes by reactive distillation methylation of a reformate. PCT Int. **2003**, WO2004000768(A1), Dec 31.
48. Rock, K.; Gildert, G.R.; McGuirk, T. Catalytic distillation extends its reach. Chem. Eng. (New York) **1997**, *104* (7), 78–84.
49. Silvergerg, S.E.; Lattner, J.R.; Sanchez, L.E. Use of hydrogenative catalytic distillation to produce cyclopentane and/or cyclopentene from dicyclopentadiene. PCT Int. **2000**, WO2000029358(A1), May 25.
50. Yu, Z.; Gao, B.; Zhang, J. Selective hydrogenation of C3 fraction in catalytic distillation column. Xiandai Huagong **2001**, *21* (7), 23–26.
51. Li, B. Study on C3 catalytic distillation hydrogenation technology. Shiyou Huagong Sheji **2003**, *20* (4), 33–35.
52. Gildert, G.R.; Loescher, M.E. Catalytic Distillation Process for the Production of C8 Alkanes U.S. Patent 6,274,783, Aug 14, 2001.
53. Louie, W.S.; Mukherjee, U.K.; Hamilton, G.L. Hydrocracking of Vacuum Gas Oils and Other Gas Oils in Cocurrent-Countercurrent Reaction Flow System with Post-Treatment Reactive Distillation U.S. Patent 6,514,403, Feb 4, 2003.
54. Mukherjee, U.K.; Louie, W.S. Hydrocracking of Vacuum Gas and Other Oils Using a Post-Treatment Reactive Distillation System U.S. Patent 6,547,956, Apr 15, 2003.
55. Gartside, R.J.; Haines, R.I.; Skourlis, T.; Sumner, C. Olefin Plant Recovery System Employing a Combination of Catalytic Distillation and Fixed Bed Catalytic Steps U.S. Patent 6,759,562, Jul 6, 2004.
56. Rock, K.L.; Foley, R.; Putman, H.M. Improvements in FCC gasoline desulfurization via catalytic distillation. 1998 Annual Meeting, National Petroleum Refiners Association, Mar 15–17, 1998.
57. Hearn, D. Catalytic Distillation Machine U.S. Patent 5,266,546, Nov 30, 1993.
58. Hearn, D.; Putman, H.M. Gasoline Desulfurization Process U.S. Patent 5,597,476, Jan 28, 1997.
59. Gardner, R.; Schwarz, E.A.; Rock, K.L. Start up of first CDHydro/CDHDS unit at Irving Oil's Saint John, New Brunswick Refinery, NPRA, New Orleans, LA, Mar 18–20, 2001, AM-01-39.
60. Rock, K.L. Ultra low sulfur gasoline via catalytic distillation. In *Pre-Print Archive*, American Institute of Chemical Engineers (Spring National Meeting), New Orleans, LA, Mar 11–14, 2002; American Institute of Chemical Engineers: New York, 2002; 946–951.
61. Podrebarac, G.G.; Gildert, G.R. Catalytic distillation for desulfurization of full-range naphtha by thioetherification and hydrodesulfurization. PCT Int. Appl. **2002**, WO2002066580(A1), Aug 29.
62. Loescher, M.E.; Podrebarac, G.G.; Ho, P.K. Reactive distillation column for sulfidation of petroleum hydrodesulfurization catalysts. PCT Int. Appl. **2002**, WO2002062471A2, Aug 15.
63. Vargas-Villamil, F.D.; Marroquin, J.O.; de la Paz, C.; Rodriguez, E. A catalytic distillation process for light gas oil hydrodesulfurization. Chem. Eng. Process. **2004**, *43* (10), 1309–1316.
64. Perez-Cisneros, E.S.; Granados-Aguilar, S.A.; Huitzil-Melendez, P.; Viveros-Garcia, T. Design of a reactive distillation process for ultra-low sulfur diesel production. Comput. Aided Chem. Eng. **2002**, *10*, 301–306.
65. Krafczyk, J.; Gmehling, J. Application of catalyst packings for the manufacture of methyl acetate

by reactive rectification. Chem. Ing. Techn. **1994**, *66* (10), 1372–1375.
66. Hanika, J.; Kolena, J.; Smejkal, Q. Butylacetate via reactive distillation—modelling and experiment. Chem. Eng. Sci. **1999**, *54* (21), 5205–5209.
67. Smith, L.A., Jr.; Hearn, D; Jones, E.M., Jr. Oligomerization Process U.S. Patent 5,003,124, Mar 26, 1991.
68. Nurminen, M.; Pyhalahti, A.; Siira, P.; Tiitta, M. Process for Dimerizing Light Olefins U.S. Patent 2004/0181106 A1, Sep 16, 2004.
69. Ng, F.T.T.; Huang, C.; Podrebarac, G; Nkosi, Rempel; G.L.. Selective dimerization of butenes to octenes via catalytic distillation. Conference Preceedings of the 7th World Congress of Chemical Engineering, Glasgow, Scotland, Jul 10–14, 2005.
70. Loescher, M.E. Recovery of Tertiary Butyl Alcohol U.S. Patent 6,596,913, Jul 22, 2003.
71. Vora, B.V.; Hammershaimb, H.U. Oligomer Production from Isobutene With Catalytic Distillation U.S. Patent 6,025,533 (A), Feb 15, 2000.
72. O'Rear, R.; Dennis, J. Process for Making a Lube Base Stock from a Lower Molecular Weight Feedstock in a Catalytic Distillation Unit U.S. Patent 6,398,946, Jun 4, 2002.
73. Abazajian, A.N. Process for Improved Yields of Higher Molecular Weight-Olefins from Lower-Olefins Using Isomerization and Disproportionation in a Reactive Distillation Column U.S. Patent 6,518,469, Feb 11, 2003.
74. Di Gerolamo, M.; Catani, R.; Marchionna, M. Process for the Hydrogenation of Branched Olefins Deriving from the Dimerization of Isobutene U.S. Patent 0078462A1, Apr 24, 2003.
75. Podrebarac, G.G.; Ng, F.T.T.; Rempel, G.L. The production of diacetone alcohol with catalytic distillation. Part I. Catalytic distillation experiments. Chem. Eng. Sci. **1998**, *53* (5), 1067–1075.
76. Podrebarac, G.G.; Ng, F.T.T.; Rempel, G.L. The production of diacetone alcohol with catalytic distillation. Part II. A rate-based catalytic distillation model for the reaction zone. Chem. Eng. Sci. **1998**, *53* (5), 1077–1088.
77. Huang, C.; Yang, L.; Ng, F.T.T.; Rempel, G.L. Application of catalytic distillation for the aldol condensation of acetone: a rate-based model in simulating the catalytic distillation performance under steady-state operations. Chem. Eng. Sci. **1998**, *53* (19), 3489–3499.
78. Huang, C.; Ng, F.T.T.; Rempel, G.L. Application of catalytic distillation for the aldol condensation of acetone: the effect of the mass transfer and kinetic rates on the yield and selectivity. Chem. Eng. Sci. **2000**, *55* (23), 5919–5931.
79. Zheng, Y.; Ng, F.T.T.; Rempel, G.L. Catalytic distillation: a three-phase nonequilibrium model for the simulation of the aldol condensation of acetone. Ind. Eng. Chem. Res. **2001**, *40* (23), 5342–5349.
80. Lawson, K.H.; Nkosi, B. Production of MIBK Using Catalytic Distillation Technology U.S. Patent 6,008,416, Dec 28, 1999.
81. Saayman, N.; Lund, G.J.; Kindersmans, S. Process for Production of MIBK Using CD Technology U.S. Patent 6,518,462, Feb 11, 2003.
82. Levin, D.; Santiesteban, J.G. Production of phenol using reactive distillation. PCT Int. Appl. **2001**, WO02001046102A1, Jun 28.
83. Memphos, S.P.; Groten, W.A.; Adams, J.K. Process for Production of Methanol U.S. Patent 5,886,055, Mar 23, 1999.
84. Allison, J.D.; Wright, H.A.; Harkins, T.H.; Jack, D.S. Use of Catalytic Distillation Reactors for Alcohol Manufacture U.S. Patent 6,723,886, Apr 20, 2004.
85. Powers, D.H. Alpha Olefin Production U.S. Patent 6,768,038, Jun 24, 2004.
86. Adams, J; Groten, W.; Nemphos, S. Process for Vinyl Acetate U.S. Patent 6,620,965, Sep 16, 2003.
87. Voss, B. Acetic Acid Reactive Distillation Based on DME/Methanol Carbonylation U.S. Patent 6,175,039, Jan 16, 2001.
88. Okasinksi, M.J.; Doherty, M.F. Simultaneous kinetic resolution of chiral propylene oxide and propylene glycol in a continuous reactive distillation column. Chem. Eng. Sci. **2003**, *58*, 1289–1300.

Recycling of Spent Tires

Roger N. Beers
David A. Benko
The Goodyear Tire & Rubber Company, Akron, Ohio, U.S.A.

INTRODUCTION

Automobiles are an important part of our modern life providing mobility, status, enjoyment, and numerous other things. The tires from these cars and other vehicles are mostly discarded after their first useful life is consumed, even though a substantial amount of the original tire remains. Because tires are vulcanized under heat and pressure to form an extremely durable thermoset composite of high-molecular weight polymers, organic and inorganic fillers, steel, fabric, and rubber chemicals, it is very difficult to reverse the process and recover the original materials in the same form as when they started. Simply discarding the scrap tires is a waste of resources and causes environmental problems if landfilled or kept in tire stockpiles. Through the 1990s U.S. motor vehicle registrations have risen, greatly increasing the number of tires in use. Advancements in tire durability and treadwear have increased the tire mileage expectations, but large numbers of tires are still being removed from service. A great amount of effort is being exerted by the tire and rubber companies, government bodies, waste management companies, recyclers, and other interested parties to address this scrap tire challenge. New and innovative uses for scrap tires are being developed, and traditional uses are being expanded by improved quality products, improved processing, and changing government and consumer attitudes toward recycled products. This entry shows the major uses for scrap tires now and the amount of scrap tires disposed of by those processes. Reviews of rubber recycling, technical aspects of reprocessing vulcanized rubber, and recovery and reuse of rubber have been published previously and provide more information on this topic.[1–6]

COMPOSITION

The number of scrap tires generated annually is about 281 million.[7] Passenger tires account for 84% of the tires scrapped in a year, while light and heavy trucks contribute 15%, and the remaining 1% is generated from aircraft, heavy equipment, and off-road tires.

The approximate weight of scrap tires is 5.86 million tons.[7] The average scrap passenger tire weighs 20 lb, light truck tires weigh 30 lb, heavy truck tires weigh 100 lb, with the largest weighing up to 10,000 lb. For a typical spent passenger tire, 12–13 lb of rubber can be recovered. This rubber is composed of about 35% natural rubber and 65% synthetic polymers, while in truck tires the percentages are reversed with about 35% synthetic polymers and 65% natural rubber.[8] Rubber polymers found in tires are polyisoprene (mostly natural and some synthetic), polybutadiene, styrene–butadiene (solution and emulsion), and butyl. The other major components of scrap tires are carbon black, inorganic fillers, polyester or nylon fabric for ply, wire for belts and beads, and a variety of rubber chemicals comprising sulfur, accelerators, antidegradents, oils, and waxes. See Fig. 1 for the amounts of these materials.[9] A tire is a highly engineered and a complex composite that is difficult to physically degrade and nearly impossible to devulcanize to reclaim the original starting materials. Major emphasis has been put on using the cured rubber in some form (whole tires, cut tires, crumb) or to burn the tires to recover the energy value of all the components.

ENVIRONMENTAL CONCERNS

Scrap tires account for about 1.8% of the solid waste generated annually in the United States.[7] Besides, the tires taken out of service each year, approximately 300 million tires, are stockpiled around the country as of 2001.[7] These stockpiles are breeding grounds for mosquitoes, fire hazards, unsightly, and are liabilities. Fortunately, since the early 1990s this number has been reduced from the 1 billion tires reported stockpiled. This decline in stockpile estimates was based on three factors. Aggressive programs in some states to clean up stockpiles, improvements in estimating stockpile size, and the loss of tires in tire fires. This large number of tires provides a huge challenge for tire companies, environmentalists, government agencies, and businesses to utilize efficiently.

HISTORY

Rubber reclaiming or recycling is nearly as old as the rubber industry itself. In the early 1900s rubber was

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007939

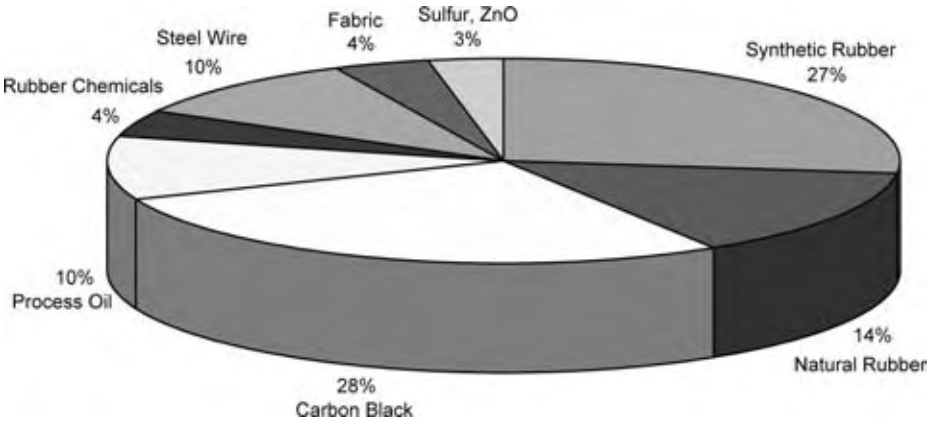


Fig. 1 Composition of average radial passenger tire by weight. *(View this art in color at www.dekker.com.)*

scarce and very expensive, so it was reused as much as was practical. The tire splitting industry dates back to about 1915.[3] Tires were split and made into shims, gaskets, mats, dock fenders, and other useful items. Before World War II all the rubber used was natural rubber, which was reclaimed with heat and chemicals and reused at levels up to 20% in new rubber compounds. Retreading of passenger tires was used extensively through the 1930s and 1940s. In the 1960s, recycling dropped to about 20%.[3] Cheap oil imports, larger uses of synthetic rubber, and the advent of steel-belted radials along with changing consumer attitudes led to a steady decline in tire recycling. In the 1970s interest in recycling spent tires increased as tire companies and others realized that discarding these high-energy containing materials was economically and environmentally unsound. Whole scrap tires were used as floating breakwaters, tire reefs, and highway crash barriers to take advantage of their energy absorbing properties and resistance to environmental degradation. Emphasis was placed on manufacturing increased quantities of shredded and ground rubber of different sizes and composition for use in engineering applications, such as road bases, erosion control, and asphalt modifiers. Finely ground crumb rubber was also used in tires and automotive applications. This was also the time when efforts began to use the high energy content of tires to generate steam and electricity. Pyrolysis of scrap tires was conducted under different processing conditions to degrade the tires to recover a mixture of gas, oil, and carbon black for sale and reuse. Although much work and investigation of the pyrolysis process was performed, it was not economically viable. Some of these recycling processes are still being used and have been enhanced by new technology, while others have declined significantly or are no longer used. The major uses for scrap tires today will be described in the remainder of this entry.

UTILIZATION OF RESOURCES

There are many new and innovative ways to utilize the resources offered by scrap tires. The reuse/recovery rate for scrap tires is about 78% with another 8% being retreaded.[7] There are five methods available to solve the problems associated with scrap tires:

Reduce: The most desirable option is to reduce the number of tires that enter the waste stream. Manufacturers have reduced the number of scrap tires by selling more durable extended mileage tires, with average tread life of 40,000–50,000 mi. Compared to the average mileage expectations of 25,000–27,000 in the 1970s, the industry has prevented millions of tires annually from becoming scrap tires.

Reuse: Reusing the tire is the next best alternative. Tire casings can be retreaded or made into various miscellaneous products. Aircraft tires can be retreaded up to 12 times and the average truck tire can be retreaded two or three times. About 16.2 million tires were retreaded in 2001 if passenger, light, medium, and heavy trucks are included. Casings can be used for products, such as muffler hangers, snow blower blades, plant holders, and other miscellaneous engineering applications. A large amount of casings are shredded, the wire and the fabric removed, and the rubber ground into various size crumbs. Larger-size tire shreds and tire chips are used in civil engineering uses. The smaller size crumb rubber is used in rubber and plastic articles and in asphalt.

Recycle: Today there is a great emphasis on devulcanization methods to recycle the rubber in tires and other rubber products. Some of these new devulcanization methods use supercritical fluid technology, ultrasonic techniques, microwave energy, and biological modification.[10–28] These methods are explained in detail in the entry, which deals primarily with devulcanization.

Recover: Much emphasis is being directed toward recovering the energy from the materials remaining

Table 1 2001 Annual utilization of scrap tires

Usage for scrap tires	Millions of tires
All reuse	218
Energy recovery, TDF	115
Civil engineering	40
Ground rubber	33
Exports	18
Fabricated products	8
Miscellaneous	7

(From Ref.[7].)

in the spent tires. The use of old tires to provide electricity and energy for cement kilns and papermaking are economically and environmentally sound solutions.

Landfill: Whole tires create problems in landfills, but shredded tires do not. They take up less space, do not readily burn, and do not provide habitat for mosquitoes. Landfilling of tires is a poor alternative to the options presented above and is simply disposal with no intent to get any value from the tires.

The major uses of scrap tires are divided into categories as shown in Table 1. About 218 million tires were reused in 2001, with the majority going to energy recovery. The next two largest usage categories were for civil engineering applications and ground rubber applications. All these uses will be described in detail.

GROUND OR CRUMB RUBBER

Tires are shredded and reduced in size for most reuse and recycle applications. The larger sizes are called tire shreds and tire chips. Tire shreds are basically flat irregularly shaped tire chunks ranging in size from 18 to 1 in. with the majority being between 4 and 8 in.[29]. These are used mainly for construction and civil engineering applications. Tire chips are more finely and uniformly sized, ranging from 3 to 0.5 in. in size. Tire chips are used extensively in civil engineering and tire-derived fuel (TDF). Ground rubber and crumb rubber are used interchangeably, but actually ground rubber particles are intermediate in size between tire chips and crumb rubber, ranging from 3/8 in. to 20 mesh. Crumb rubber particles range from 4 to about 200 mesh, with 20 to 80 mesh particles used the most.[30] Ground and crumb rubber are free of fabric, wire, and other contaminants. The smaller the particle size, the more expensive the rubber. In 2001, the average price of 10 mesh crumb rubber was about 13 cents/lb, and 80 mesh crumb rubber averaged about 31 cents/lb.[31]

There are three main methods for producing crumb rubber: dry ambient grinding, wet grinding, and cryogenic grinding.[4] Dry ambient grinding uses serrated grinders to grind various types of scrap and spent rubber articles. Normally, any wire or steel has been separated previously. The size of the rubber particles from this process usually ranges between 10 and 30 mesh. Even though this process is considered an ambient process, the rubber particles are exposed to considerable heat that is generated in the grinding process. Ambient grinding produces irregular-shaped particles with small appendages caused by the softening and tearing during grinding. Wet or solution grinding uses water and starts with 10–20 mesh rubber. These particles are ground between two wheels with the water enhancing the lubricity and cooling. Particles as fine as 200 mesh are obtained and size is related to the grinding time. This process gives fairly smooth particles. Scrap rubber can also be reduced in size by cryogenic grinding. Small pieces of rubber are placed in liquid nitrogen and ground into fine powder generally ranging from 30 to 100 mesh. Surface morphology is different from ambient grinding because the particles are fractured and surface oxidation is believed to be less.[32] Table 2 shows the amount of ground rubber used in North America in 2001.[33] This amount includes the market for the United States and Canada. The ground-rubber market consists of ground scrap

Table 2 2001 North American crumb rubber markets

Market	Millions of pounds
Asphalt modifications	292
Molded products	307
Sports surfacing	141
Tires/automotive	112
Surface modified/reclaim	36
Plastic blends	38
Animal bedding	37
Construction	28
Other	5
Total	996

(From Ref.[33].)

Table 3 Tire crumb use in the United States (millions of pounds)

Market	1995	2000
Asphalt modification	97	261
Rubber products	74	250
Surfaces	35	113
Automotive	52	100
Miscellaneous	11	143
Total	269	867

(From Ref.[1].)

tires and tire buffings. Table 3 shows the amounts of crumb rubber used in the United States only for 1995 and 2000 and shows the growth taking place in this market.[1] Although estimates vary slightly, at least 33 million scrap tires were used in this market in 2001 and the rest was obtained from tire buffings. No attempt is made to differentiate between tire buffings and scrap tires in identifying markets.

Rubber Modified Asphalt

Ground tire rubber is used as an additive in various types of asphalt pavement constructions. There are two main processes in use: the dry and the wet process, and modifications of both exist. In the dry process, ground rubber ranging in size from 0.25 in. to 20 mesh can be used as an aggregate substitute at a loading of 3–4 wt%.[34] The wet process uses 30–100 mesh crumb rubber, which is added to modify the physical and the chemical properties of the binder. Levels of crumb rubber in the wet process can vary, but as the level rises above 5–10%, the cost increases significantly compared to conventional pavement.[35] A recent article shows that 80 mesh crumb rubber is very good for use in both wet and dry processes.[36] The small particle size reacts faster with the asphalt to form a homogenous mix and is compatible in all types of asphalt equipment.

A study conducted in Canada used three types of crumb rubber at 10% by weight with two commonly used asphalt binders.[37] The rubber-modified asphalts (RMAs) prepared were evaluated for resistance to forming ruts and low-temperature cracking. Compared to conventional asphalt, the crumb rubber modifications significantly improved the high-temperature performance and moderately improved the low-temperature cracking. Both low- and high-temperature responses revealed improvement as the particle size of the rubber was reduced from 30 to 80 mesh and lower. The wet process modified binder is commonly referred to as asphalt rubber and is also used in seal coats, stress absorbing membranes, waterproofing membranes, and hot mix binders. The level of use of ground rubber varies according to the method, but in 2001 about 290 million pounds was used for asphalt construction.[33] The rubber improves asphalt ductility and increases the temperature at which asphalt softens, thus improving durability and fatigue resistance. The rubber also improves the skid resistance of the asphalt pavement.

A review paper by Van Kirk indicated that there is sufficient field experience to show the advantages of crumb rubber in RMA.[38] The main advantages are that pavement thickness can be reduced with no loss in service life, pavement cracking is decreased, and maintenance cost is reduced. The use of RMA has demonstrated that it is quieter, smoother, reducing overall cost, and a good method of using scrap tires.

California, Arizona, Texas, and Florida promote the use of ground rubber in asphalt. Several other states are evaluating the use of RMA and may utilize this method. Growth in this area for crumb rubber depends on a number of factors: DOT support, additional performance and cost data, and a readily available supply of crumb rubber.

Molded Products

Molded products are another large market for ground and crumb rubber from scrap tires. The size-reduced rubber can be molded, extruded, or processed in various ways with other thermosets or virgin rubber. Automotive products, tires, and plastic blends are not included in this category and will be discussed as separate categories. Crumb rubber extends and modifies the molded products, in particular lowering cost, improving impact resistance, lowering shrinkage, and improving processing. This market should experience relative growth because of improved product quality, more manufacturer and consumer interest, campaigns to buy recycled materials, and state tire program incentives. In 2001, about 300 million pounds was used in this area.[33] Some typical products are floor tile, mats, water hoses, sound proofing panels, bin liners, and belt covers.

Tires and Automotive Applications

Great effort is being expended to find ways to use increasing amounts of scrap tires in the manufacture of new tires. Automotive companies are strongly encouraging tire producers to increase the recycle content in the new tires they purchase. Ford is currently trying to have the tire recycle content raised to 10% by weight with the ultimate goal being 25% recycle. Many new technologies are being investigated to enable higher use of tire-derived rubber without compromising the stringent physical properties needed in a radial tire. Currently, about 10% is the maximum amount of reprocessed rubber used in certain compounds, and the overall tire content is around 3%.[7] Recycle is used in passenger, farm, and light and heavy truck tires. Tire components that commonly contain reprocessed rubber are the innerliner, treads, sidewall, ply, and barrier compounds. Currently, only about 25% of the total reprocessed rubber used in tires comes from scrap tires. The remainder comes from curing bladders, inner tubes, and tread peels. The curing bladders and inner tubes are used almost exclusively in the tire liners because they are butyl and provide excellent

Table 4 Tread compound properties containing crumb rubber

	Control (0 crumb)	20 phr[a] crumb	20 phr crumb	40 phr crumb
Tensile[b]	100	85	80	75
300% mod	100	91	82	77
100C rebound	100	94	93	91
Mooney scorch	100	93	81	79
Cure amount	100	92	79	78
Abrasion	100	90	83	68
Heat buildup	100	89	86	78
Viscosity	100	73	73	56

[a]phr is parts by weight per hundred parts of rubber in compound.
[b]Lower figures indicate worse performance.
(From Ref.[39].)

air retention properties. Also, butyl is incompatible with other tire elastomers.

Typical particle sizes used range from 20 to 80 mesh. The 80 mesh particles provide smoother compounds and possibly slightly improved properties. The effect of ground rubber recycle on most rubber compound properties is negative, which prevents higher usage. Ground rubber is used as an extender or filler to reduce costs and help processing. Properties that are impacted negatively are compound viscosity, tensile strength, modulus, elongation, scorch, abrasion resistance, cure, and hysteresis.[39] Typical effects of loading on key compound properties are shown in Table 4. The table shows how compound properties change as the amount of 40 mesh ground scrap tire rubber is added to a typical passenger tire tread.[39] The control contains no crumb rubber and 20, 30, and 40 parts of crumb rubber are added to the other tread formulations. The data are normalized with the control given a value of 100, and the other compounds compared to that value. Processing improvements generally noted are improved mold flow, better air bleed, and reduced shrinkage.

There are several reasons for the reduced properties. The hard ground rubber particles contain accelerator fragments from being cured previously and these fragments migrate into the virgin rubber and cause it to cure faster with less scorch protection.[40] Sulfur from the virgin rubber compound also migrates into the cured particles making them harder and causing the matrix to be cured to a lower amount, reducing properties. Another reason attributed to the reduced strength is poor interfacial bonding between the crumb rubber and the virgin rubber matrix.[41] Various surface treatments, both chemical and physical, additives, new processing methods, and compound modifications are being developed to improve the physical properties.[42–45] Table 5 shows the results of treated and untreated crumb when added at 20% to a natural rubber compound. In this case, the crumb is also natural rubber and the treated portion was surface activated with a two-step process developed by Vredestein rubber recycling.[44] The treated crumb compound was significantly better than the untreated crumb compound.

Automotive uses besides tires include belts, hoses, friction materials, mud flaps, air deflectors, brake pedal covers, caulks, and coatings. Like tires, the industry would like to see more recycled content in these applications. A total of about 112 million pounds of ground rubber is consumed in these applications.[33]

Sports Surfacing

This market segment has undergone a huge growth rate in the past several years. Examples are the use of

Table 5 Comparison of treated and untreated crumb rubber in NR compound

	Control (no crumb)	20% treated crumb	20% untreated crumb
Tear strength (N)	93.5	83.2	63.8
Abrasion (mm^3 lost)	85	80.8	99.5
Compression set (%)	39	43.7	52.3
Rebound (%)	51	49	24.4
Tensile (MPa)	24	20.4	17

(From Ref.[44].)

rubber in running track material, in grass surfaced playing areas, in stadium playing surfaces, and as turf topdressing. The main reasons for using ground or crumb rubber for these type surfaces is to reduce injuries, improve drainage, reduce damage to fields, and improve turf viability. Crumb rubber also retards weed growth, does not attract insects or rodents, and has a higher cushioning effect than other fill materials. The Detroit Lions football team's practice field contains a rubber soil material that has allowed practice in all weather conditions and has reduced player injuries. Another use is in horse arenas, where it gives the animals better footing and reduces injuries to the horse and to the rider. One company produces a granular product $3/8 \times 1/8$ in. in size for the playground market.[46] The material is put down in a layer 6 in. thick and provides greater shock absorbing characteristics than other playground surfaces, such as sand or mulch. Approximately 140 million pounds of ground rubber was used for some type of sport surfacing in 2001.[33] The increased use was mostly because of the improved quality of the surfacing, better application techniques, and the safety benefits. These type materials are finding use in golf courses and public parks.

Plastic Blends

Thermoplastic materials are manufactured using many processes, such as injection molding, extrusion, hot stamping, and others, that use heat and pressure to form a particular product or shape. In most cases, the thermoplastic is pelletized and crumb rubber can be readily blended in the composition during processing. The crumb rubber is used to improve impact resistance, improve processing, extend the material, or provide some performance improvement to the finished product. In many cases, the crumb rubber surfaces are treated with silanes, chemicals, such as ethylene–glycidyl methacrylate, compatibilizing agents, high-energy radiation like plasma etching, or other novel techniques to increase the adhesion with the thermoplastic matrix and improve compatibility.[47–50] Plastic blends with rubber are used for a great number of products, such as body panels for golf carts and lawn tractors, sports equipment, toys, cable and wire insulation, and truck bed liners. Crumb rubber is even used to make thermoplastic elastomers (TPE) by alloying 120 mesh rubber with an olefinic matrix.[51,52] Before alloying, the rubber powder undergoes a proprietary surface devulcanization process and compatibilization technique to enhance the blending of the rubber with a polyethylene/polypropylene mix to form the TPE. These applications accounted for about 38 million pounds in 2001 and there appears to be significant market potential for this application because of increased research and development on surface treated rubber.[33]

Construction

Ground rubber finds many uses in the construction industry because of its excellent weathering, chemical resistance, and nonbiodegradable properties. Some uses are paving or tiles for driveways, shingles and roofing materials, thermal and acoustic insulation panels, plastic lumber, flooring materials, and other building materials. There has been substantial growth in the manufacture of floor tiles and sidewalk pads, where crumb rubber is used with urethane binders for cost and injury reduction. About 28 million pounds of ground rubber was used for these applications in 2001.

Surface Modification/Reclaim

This category involves processing used rubber products into a form that can be incorporated into virgin rubber compounds. Ground rubber is the starting material for the chemical process that severs the cross-links and polymer chains by chemical or steam digestion to make reclaim. Other reprocessed rubber uses crumb rubber and modifies the surface by chemical or physical means to enhance the adhesion and compatibility to improve properties when mixed with virgin rubber compounds. About 36 million pounds of ground rubber is used for these applications.[33] The major benefit of using reclaim is lower cost and the improvements in processing it can add to the compounds. Compounds containing reclaim have lower nerve, which translates into better extrusion, increased calendar speed, and improved flow and mold filling. Properties like green strength and tensile are lowered when reclaim is used. Today some whole tire reclaim is used in low performance requirements like mats and extruded rubber products. The major rubber reclaimed is butyl that comes from truck inner tubes.

Crumb rubber can be ground into small particles from 30 to 200 mesh and used in virgin rubber compounds primarily as a filler. Smaller particle sizes give optimum surface smoothness and slightly higher physicals. Mesh sizes from 30 to 80 can lower compound cost, improve mold flow, help air bleed, and reduce shrinkage. However, crumb rubber generally increases compound viscosity and reduces strength and modulus. In low performance requirements, levels from 5% to 10% are commonly used. To increase properties, the crumb rubber is treated to enable the virgin rubber to bond better. New surface treatments are gaining, particularly in the tire market.

Animal Bedding

Ground rubber is used in agriculture for animal bedding. Ground rubber is used because it does not hold moisture, is relatively cheap, is not readily biodegradable, and does not harbor insects and pests. Mats are also made with crumb rubber for use by animals. Over 35 million pounds of ground rubber is used for this application.[33]

These are the main uses for ground and crumb rubber. In addition, another half-million tires or less are used for miscellaneous applications as diverse as can be imagined. The exact numbers of tires used for each application are as up-to-date as possible and as accurate as the data available. Overlaps in category use and some estimates in the number of actual tire units consumed may be slightly different from other published figures. The source of most of the information is the Scrap Tire Management Council that was formed to monitor and promote disposal and reuse of scrap tires.

CIVIL ENGINEERING

Civil engineering includes a wide range of uses for scrap tires. The five main applications include lightweight fill for embankments, subgrade insulation for roads, backfill for walls and bridge abutments, landfill construction and operation, and septic system drain fields. In nearly all the applications, the tire chips or tire shreds typically replace currently used construction materials, such as soil, clean fill, drainage aggregates and lightweight fill materials. The benefits of using the tire chips and shreds are lower density, improved drainage properties, better thermal insulation, and lower cost. Tire stockpiles are good for these applications, as the presence of dirt on the tires is not a problem. The civil engineering area is growing and about 40 million tires were used in 2001.[7]

Lightweight Fill for Embankments

Work has been done in a number of states using tire shreds as a subgrade fill in the construction of highway embankments and other fill projects. Tire shreds weighs considerably less than conventional soil fill and this allows construction on weak compression foundation. The use of tire shreds for most projects is considerably cheaper than alternatives, such as expanded shale aggregate and polystyrene insulation blocks.[29,53] Scrap tire materials have also been used to retain forest roads, protect coastal roads from erosion, enhance the stability of steep slopes along highways, and reinforce shoulder areas. The importance of this application is illustrated by a project in Maine, which used 1.2 million tires as lightweight fill for the construction of two highway embankments on weak clay and saved about $300,000.[7]

Subgrade Insulation for Roads

In northern climates, one of the problems affecting roads is the water released when the subgrade soils thaw in the spring. To prevent this, tire shreds are used as subgrade insulation. The insulation provided by a 6–12 in. layer of tire shreds keeps the subgrade soils thawed during the winter. Tire shreds also allow excess water to drain from beneath the road.[54]

Backfill for Walls and Bridge Abutments

Several projects have been constructed using tire shreds as backfill for walls and bridge abutments. The weight of the tire shreds produces low horizontal pressure on the wall allowing thinner walls at lower cost. Another advantage is that tire shreds provide better drainage and good thermal insulation, which eliminates problems with water or frost buildup behind the wall.[55]

Landfill Construction and Operation

Many landfill operators have found that scrap tires can be used beneficially in the construction and operation of the landfill. Tire shreds are used in building leachate collection basins in new landfills and they can also be used as the drainage layer in landfill caps. Another application is mixing tire shreds with soil as daily cover material.

Septic System Drain Fields

Tire shreds and chips are used as a replacement for stone in septic system leach fields. The lower density of the tire shreds greatly reduces the expense and the labor to construct leach fields, while providing equal performance to stone and gravel. The number of states allowing scrap tires in septic systems is increasing. As in all the civil engineering applications, there is concern about the environmental impact of using scrap tires. One of the major issues is groundwater contaminants and the effect of tire shreds on water quality. Studies have shown that tire shreds placed above or below the water table pose no significant health or environmental risk.[7,56] Neutral pH groundwater does not increase the concentration of metals and no detectable organic leachates are found from the tire shreds. Under some conditions, the steel belts contained in the tire

shreds may increase the levels of iron and manganese, but these are not harmful. These concerns affect the laws related to scrap tire use, and along with other issues like cost and availability, determine the use and growth. Each potential civil engineering use brings with it a particular set of technical, environmental, and economic constraints that must be fully evaluated before the application is readily acceptable. The potential use of scrap tires in civil engineering is very substantial and this will continue to grow.

ENERGY MARKET

Tires have a tremendous market potential as fuel. Energy recovery was the largest market in 2001, representing about 115 million scrap tires annually.[7] On average, 97% of a tire's content contributes directly to energy value. The average radial passenger tire derives its energy content from the following ingredients:

- *Synthetic rubber* is produced from crude oil, a high-energy fuel much cleaner than coal.
- *Natural rubber* is a renewable energy source that is harvested from trees.
- *Carbon black* also comes from crude oil and has a high energy value.
- *Rubber chemicals*, *oils, and organic fibers* are all derived from crude oil and contribute to the energy value.
- *Steel belts and beads* oxidize at high temperatures to produce 3500 Btu/lb.

Energy from scrap tires is being recovered in the United States and many other countries worldwide with beneficial results. Burning one passenger tire produces approximately 300,000 Btu, which is roughly equivalent to 2 gal of fuel oil or 23 lb of coal. Whole scrap tires average 15,000 Btu/lb and most coal averages 12,000–12,500 Btu/lb (Table 6).[57] Tires contain less nitrogen than coal, and there is less sulfur in tires than most types of coal. Tires burned under proper conditions reduce solid waste and air emissions, and produce little or no smoke or fumes.

Table 6 Comparative energy values

Fuel	Btu/lb
Petroleum coke	13,700
Bituminous coal	12,750
Subbituminous coal	10,500
Lignite coal	7,300
TDF	14,000–15,500

(From Ref.[57].)

Cement, electricity, and paper producers burn scrap tires as a supplemental fuel, because tires have more energy value per pound than most coal and burn more evenly. Scrap tire combustion decreases the emission of particulates, oxides of nitrogen, and carbon dioxide compared to coal. Whole tires can be burned in some cases or TDF is created by shredding scrap tires and processing them into tire chips. Typically, these tire chips range in size between 2 and 4 in., but many facilities are shifting to smaller fuel chips 2 in. by 2 in. or slightly smaller. These smaller chips contain less steel and can reduce handling problems and lower ash disposal. Boiler design and the production process determine the allowable wire content level. Tire-derived fuel is easy to handle, store, and feed into combustion devices. It is also easily blended with conventional fuels. Using scrap tires as fuel totally eliminates the scrap tire, while it decreases the dependence on foreign oil supplies and conserves precious natural resources.

The total use of scrap tires for energy recovery has declined over the past several years from 1996 when approximately 135 million scrap tires were consumed.[7] The use for energy recovery seems to have stabilized with the potential for slow growth over the next several years. It must be remembered that the energy market is very complex and many other factors affect the use of scrap tires for energy recovery. Factors that impact this area are air pollution laws, states subsidies for use of scrap tires, energy deregulation, economic conditions, product quality, and costs. All of these issues plus others can affect the economics and help or hinder the use of scrap tires in these energy applications.

Cement Industry

There are 36 cement kiln locations that processed tires for fuel in 2001 with the majority burning whole tires and the rest tire chips.[7] The original reason cement kilns began using scrap tires was to reduce fuel costs. A big advantage of using scrap tires in cement kilns is that there is no solid waste disposal. The tires are completely consumed and become a part of the final product. Other advantages are less carbon dioxide, nitrogen oxides, and sulfur dioxide. The wire of the tire provides iron oxide, which is one of the raw materials required. Also, limestone, which is a raw material for the cement, neutralizes sulfur from the tires. The amount of scrap tires used as a percentage of the total fuel consumed can range from 5% to 25%. The cement industry is the largest user of scrap tires for energy and will continue to consume significant amounts. Two important issues that will affect the amount of scrap tires used are USEPA focus on nitrogen oxides and the demand for cement.

Paper and Pulp Mills

There are 18 facilities that use scrap tires for pulp and paper production.[7] Tire-derived fuel is mixed with wood waste to produce steam. The TDF improves the combustion efficiency and displaces fossil fuels. A significant portion of Oregon's scrap tires goes to this market. The volume of scrap tires used for this application has fallen from its high in 1996 much like cement industry use. It should be noted that the decrease in TDF consumption was never attributed to increased air pollution or environmental degradation.

Utility Boilers

There are 11 plants using scrap tires to produce electricity for sale and they consume the second largest amount after the cement industry.[7] Two factors that have impacted the use of TDF and will continue to impact its use are the electric industry deregulation and the implementation of Clean Air Act amendments. Until all the issues are sorted out, further expansion in this area is unlikely.

Industrial Boilers

Industrial boilers use TDF to generate steam. There are 17 facilities that use scrap tires, although it is usually in a limited degree.[7] In this market segment, the use of TDF is primarily a function of the solid wastes available and represents about 2–5% of the fuel supply. Still, tires are used because the economics and the emissions are favorable. A small increase in the use of TDF should occur in the industrial boiler market.

EXPORT OF TIRES

Tires with adequate tread and/or retreadable tires are regularly exported from the United States. Used tires are regularly sold in these markets in many other parts of the world. Routinely, slightly more than 1 million tires/mo are exported or about 15 million tires/yr by best estimate.[7]

FABRICATED PRODUCTS

The fabricated products market encompasses those products made by cutting, punching, or stamping from scrap tire carcasses. This is one of the oldest methods of reuse and is not changing. This market has a large array of products, all of which take advantage of the toughness and durability of tire carcass materials. Examples of these products are dock bumpers, muffler hangers, mat components, and snowblower blades. These entire applications generally only use bias ply tires, or tires with no steel belts. This market will remain the same and uses about 8 million scrap tires.[7]

MISCELLANEOUS USES

There are a wide variety of uses for scrap tires that do not fit neatly into any of the preceding categories. The total in this catchall category is about 7 million scrap tires. These include agricultural uses where tires are used to construct stock feeders, protect fence posts, weigh down covers, erosion control, and other uses. Usage can also include swings, planters, or other more imaginative and innovative uses.

CONCLUSIONS

Among all the methods available to manage the increasing number of scrap tires, reuse, recycle, and recovery offer the most potential to arrive at a program that consumes all the scrap tires generated each year and steadily decreases the existing tire piles. Reuse includes the civil engineering and ground rubber markets that are steadily increasing the volumes of tires they consume in highway construction projects, asphalt modifications, sports surfaces, and molded products. New technology and innovations are driving the increased use of scrap tires, but there are other factors that impact use of scrap tires. State regulations on landfills, incentive programs to use tires, establishment of an infrastructure to collect, market and distribute scrap tires, and consumer attitudes also affect the number consumed.

Reclaim or recovery of the polymers and the carbon black in the tires has great potential for providing new sources of materials to be used again to manufacture tires or other rubber products. Scientists are working on new ways to devulcanize rubber and success has been reported by using ultrasonic, microwave, supercritical fluids, and mechanical means. True devulcanization would allow much more of the recaptured polymers to be used in tires and/or other demanding applications. Recovery of energy is still the largest market for scrap tires today, even though this important market that has shown some decline in volume since 1996. New boiler technology and processing enhancements to burn scrap tires are emerging and should solve any technical problems related to this application. Again, air pollution laws, energy deregulation, and economic issues are just as important as new technology in expanding the energy recovery market.

All these forces will continue to challenge the industry for ways to consume spent tires. Currently,

about 78% of the tires removed from service each year are being reused and that number should continue to grow as new technology and greater acceptance of recycled products occurs.

REFERENCES

1. Myhre, M.; MacKillop, D.A. Rubber recycling. Rubber Chem. Technol. **2002**, *75*, 429–474.
2. Crane, G.; Elefritz, R.A.; Kay, E.L.; Laman, J.R. Scrap tire disposal procedures. Rubber Chem. Technol. **1978**, *51*, 577.
3. Beckman, J.A.; Crane, J.; Kay, E.L.; Laman, J.R. Scrap tire disposal. Rubber Chem. Technol. **1974**, *47*, 597–623.
4. Klingensmith, W. Recycling, production, and use of reprocessed rubbers. Rubber World **1991**, *203* (6), 16.
5. Warner, W.C. Methods of devulcanization. Rubber Chem. Technol. **1994**, *67*, 559–566.
6. Adhikari, B.; De, D.; Maiti, S. Reclamation and recycling of waste rubber. Prog. Polym. Sci. **2000**, *25*, 909–948.
7. Rubber Manufacturers Association; http://www.rma.org/scraptires/facts_figures.html (accessed July 2004).
8. Schnormeier, R. *Recycled Tire Rubber in Asphalt*; 71st Annual Meeting Transportation Research Board: Washington, DC, 1992.
9. Goodyear Tire & Rubber Company, Internal Document, 1999.
10. Beers, R.N.; Benko, D.A. U.S. Patent 6,387,965, May 14, 2002.
11. Isayev, A.I.; Chen, J.; Tukachinsky, A. Novel ultrasonic technology for devulcanization of waste rubbers. Rubber Chem. Technol. **1995**, *68*, 267.
12. Isayev, A.I.; Kim, S.H.; Yushanov, P. Ultrasonic devulcanization of SBR rubber: experimentation and modeling based on cavitation and percolation theories. Rubber Chem. Technol. **1998**, *71*, 168.
13. Isayev, A.I.; Yashin, V.V. A model for rubber degradation under ultrasonic treatment: part I. Acoustic cavitation in viscoelastic solid. Rubber Chem. Technol. **1999**, *72*, 741.
14. Boron, T.; Klingensmith, W.; Roberson, P. Ultrasonic devulcanization of tire compounds. Tire Technol. Int. **1996**, 82–84.
15. Isayev, A.I.; Kim, S.H.; Levin, V.Y. Superior mechanical properties of reclaimed SBR with bimodal network. Rubber Chem. Technol. **1997**, *70*, 194.
16. Pelofsky, A.H. U.S. Patent 3,725,314, 1973.
17. Mangaraj, D.; Senapati, N. U.S. Patent 4,548,771, 1985.
18. Isayev, A.I. U.S. Patent 5,258,413, 1993.
19. Isayev, A.I.; Chen, J. U.S. Patent 5,284,625, 1994.
20. Isayev, A.I.; Schworm, D.; Tukachinsky, A. Devulcanization of waste tire rubber by powerful ultrasound. Rubber Chem. Technol. **1996**, *69*, 92.
21. Isayev, A.I.; Chen, J.; Yushanov, S.P. Ultrasonic devulcanization of rubber vulcanizates. II. Simulation and experiment. J. Appl. Polym. Sci. **1997**, *59*, 815.
22. Makrov, V.M.; Drozdovski, V.F. *Reprocessing of Tires and Rubber Wastes*; Ellis Horwood: New York, 1991.
23. Novotny, D.S.; Marsh, R.L.; Masters, F.C.; Tally, D.N. U.S. Patent 4,104,205, 1978.
24. Clifford, M.L. U.S. Patent 4,130,616, 1978.
25. Tyler, K.A.; Cerny, G.L. U.S. Patent 4,459,450, 1984.
26. Hunt, J.R.; Hall, D. U.S. Patent 5,362,759, 1994.
27. Wicks, G.G.; Schulz, R.L.; Clark, D.E.; Folz, D.C. U.S. Patent 6,420,457, 2002.
28. Romine, R.A.; Snowden-Swan, L.J. U.S. Patent 5,597,851, Jan 28, 1997.
29. Read, J.; Dodson, T.; Thomas, J. *Experimental Project—Use of Shredded Tires for Lightweight Fill*; Oregon Department of Transportation: Salem, OR, 1991.
30. Heitzman, M. *Design and Construction of Asphalt Paving Materials with Crumb Rubber*; Transportation Research Board: Washington, DC, 1992.
31. *Price and Market Surveys*; Recycling Research Institute.
32. Kohler, R.; O'Neill, J. Rubber Division Meeting, ACS, Cleveland, OH, Oct 17–20, 1995.
33. Recycle Research Institute, 2002 Scrap Tire & Rubber Users Directory, 2002.
34. Morrison, G.R.; Hesp, S.A.M. New look at rubber-modified asphalt binders. J. Mater. Sci. **1995**, *30*, 2584.
35. Tarricone, P. Recycled roads. Civil Eng. **1993**, *63* (4), 46.
36. Rouse, M.W. Application of CRM in asphaltic materials. Rubber World **1995**, *212* (2), 23.
37. Coomarasamy, A.; Hesp, S.A.M. Performance of scrap tire rubber modified asphalt paving mixes. Rubber World **1998**, *218* (2), 26–32.
38. Van, K.J. Rubber Division Meeting, ACS, Orlando, FL, Sep 21–24, 1999.
39. Goodyear Tire & Rubber Company, Internal Technical Evaluation, 2001.
40. Hamed, G.R.; Gibala, D.; Zhao, J. Tensile behavior of an sbr vulcanizate containing a single rubber particle. Rubber Chem. Technol. **1998**, *71*, 861.
41. Myhre, M.; MacKillop, D.A. Modification of crumb rubber to enhance physical properties of recycled rubber products. Rubber World **1996**, *214* (2), 42–46.

42. Peter, J.; Schmidt, P.; Mahlke, D. U.S. Patent 5,844,043, 1998.
43. Kim, J.K.; Burford, R.P. Study on powder utilization of waste tires as a filler in rubber compounding. Rubber Chem. Technol. **1998**, *71*, 1028.
44. Dierkes, W. Rubber World **1996**, *214* (2), 25.
45. Fuhrmann, I. Karger-Kocsis. Promising approach to functionalisation of ground tyre rubber-photochemically induced grafting: Short Communication. J. Plast. Rubber Compos. **1999**, *28*, 500.
46. http://www.gtrcrumbrubber.com/products.html (accessed Oct 2002).
47. Yasuda, H.; Marsh, B.; Brandt, S.; Reilly, C.N. ESCA study of polymer surfaces treated by plasma. J. Polym. Sci. Chem. Ed. **1977**, *15*, 991.
48. Evans, J.M. Nitrogen corona activation of polyethylene. J. Adhesion **1973**, *5*, 1.
49. Baker, W.E.; Rajalingham, P. The role of functional polymers in ground rubber tire-polyethylene composite. Rubber Chem. Technol. **1992**, *65*, 908.
50. Xu, Z.; Losur, N.S.; Gardner, S.D. Epoxy resin filled with tire rubber particles modified by plasma surface treatment. J. Adv. Mater. **1998**, *30*, 11.
51. Jury, J.R.; Chien, A. U.S. Patent 6,262,175, Jul 17, 2001.
52. Burgoyne, J.; Fisher, J.; Jury, J. U.S.Patent 5,510,419, Apr 23, 1996.
53. Bosscher, P.J.; Edil, T.B.; Eldin, N.N. *Construction and Performance of a Shredded Waste Tire Test Embankment*, 71st Annual Meeting; Transportation Research Board: Washington, DC, Jan, 1992.
54. Humprey, D.N.; Eaton, R.A. *Tire Chips as Subgrade Insulation—Field Trial,* Proceedings of the Symposium on Recovery and Effective Reuse of Discarded Materials and By-Products for Construction of Highway Facilities, Federal Highway Administration, Denver, CO, 1993.
55. Humprey, D.N.; Sandford, T.C.; Cribbs, M.M.; Manion, W.P. *Shear Strength and Compressibility of Tire Chips for Use as Retaining Wall Backfill*, 72nd Annual Meeting; Transportation Research Board: Washington, DC, Jan 1993.
56. Sengupta, S; Miller, J. *Investigation of Tire Shreds for Use in Residential Subsurface Leaching Field Systems: A Field Scale Study*; Department of Civil Engineering, UMass Dartmouth: Dartmouth, MA, 2000.
57. Rouse Rubber Industries, TDF Fact Sheet.

Reformulated Gasoline

A. K. Dalai
D. Ferdous
Catalysis and Chemical Reaction Engineering Laboratories, Department of Chemical Engineering, University of Saskatchewan, Saskatoon, Canada

INTRODUCTION

The purpose of introducing reformulated gasoline (RFG) is to improve the air quality to reduce motor vehicle emissions of toxic and tropospheric ozone-forming compounds. The RFG regulations involve the greatest reduction in emissions of ozone-forming volatile organic compounds (VOCs) and toxic air pollutants (TAP) through the reformulation of conventional gasoline. They also take into consideration the cost of achieving such reduction.

BACKGROUND

Prior to 1995, around 85 million people (one-third U.S. population) lived in urban areas. The air quality of these areas violated federal public health standards, largely because of automotive pollutants. Gasoline and diesel-fueled cars, trucks, and buses produced half of all air pollution in the United States. This air pollution includes 66% of airborne carbon monoxide, 31% of smog-forming hydrocarbons, and 43% of lung-damaging nitrogen oxide.

On January 1, 1995, nearly one-third of the motorists in the United States found something different at the gas pump, RFG. This gasoline is designed to reduce the environmental impact of burning the fuel so that air quality meets public health standards of the national Clean Air Act amendments of 1990. The CAA was designed to address the issues related to fuel quality improving gasoline. This required that the RFG to be used in cities with the worst smog pollution should contain oxygen to reduce harmful emissions of ozone. ARCO first introduced RFG in 1989 in Southern California. In the RFG, benzene and aromatic levels were reduced, vapor pressure (or volatility) decreased, and the oxygenates were added to improve combustion and other fuel properties.

Reformulated and oxygenated fuels have been already used at different regions of the United States where the ozone level limit has been exceeded. As indicated earlier, one way to reduce air pollution from cars and trucks is to use an RFG that is designed to burn cleaner, while the CAA requires cities with the worst smog pollution to use RFG, but other cities with smog problems may choose to use RFG. The RFG program is a significant step toward cleaning the air we breathe, and a significant component of the country's smog reduction strategy. The RFG is currently used in 17 states and the District of Columbia in the United States. Nowadays, about 30% of gasoline sold in the United States is reformulated. Each oil company prepares its own formula to meet federal emission reduction standards. The RFG's air quality benefits and other industrial and transportation controls aimed at smog reduction are responsible for the long-term downward trend in U.S. smog. In North America and other countries, increasingly stringent fuel specifications for RFG are mandated.

REFORMULATED GASOLINE AND ITS GOAL

Reformulated gasoline, referred to as ''clean gasoline'' is designed to reduce both exhaust and evaporative emissions from vehicles. This fuel offers a way to reduce a variety of gasoline emissions without developing alternative fuel technologies to replace oil derived fuels and internal combustion engine. The definition of RFG includes the following key factors:[1]

- A 15% reduction in toxic chemicals, including VOCs.
- A maximum of 1% of the carcinogen benzene (a 50% reduction) and a maximum of 25% for aromatic hydrocarbons (compared with 40% previously).
- A minimum content of 2% oxygen, by weight, to promote cleaner combustion, especially to reduce carbon monoxide emission.

The sunlight driven complex reactions involving VOCs and oxides of nitrogen (NO_x) are responsible for atmospheric ozone formation. The purpose of using reformulated fuels is to reduce both ozone-forming VOC emissions and air toxic emissions from vehicles. The key constraints that are introduced by Phase I in 1995 and II in 2000 are reduction in VOCs, NO_x, and toxics. While each of theses variables has complex

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007943

interactions with other gasoline properties, each is affected by a dominating component.[2] Volatile organic compounds, NO_x, and toxics are the functions of Reid vapor pressures (RVP), nitrogen content, and benzene content, respectively.

Beginning in January 1995, Phase I RFG was intended to provide a 15–17% reduction in both ozone-forming VOC emissions and air toxic emissions from vehicles. In the beginning of January 2000, Phase II RFG intended to provide a 25–29% reduction in VOC emissions, 20–22% reduction in air toxics emissions, and 5–7% reduction in nitrogen oxide emissions from vehicles.[3] This phase-wise introduction of RFG indicated that fuel parameters reduced a vehicle's pollutant emission. Phase I of California's RFG program began in 1992 and required reduced gasoline vapor pressure during the summer ozone season, use of detergent additives to control engine deposits, and elimination of lead-based antiknock additives.[4] The current EPA summer vapor pressure maximum for most areas of the United States is 9.0 psi. A lower summer maximum of 7.8 psi has been in effect in ozone nonattainment areas in VOC Control Region 1, such as California, since 1992.[5] However, Phase 2 of the California RFG program that took effect in the first half of 1996 required more extensive changes to gasoline properties.[6]

A number of programs have assessed the effect of the use of RFG on vehicle emissions; the most comprehensive of these is the Auto/Oil Air Quality Improvement Research Program.[4] Their task of assessing the air quality benefits of RFGs was relatively difficult, compared to the other evaluations of control measures that have typically been undertaken for regulatory purposes.[7]

The tailpipe emissions of the major pollutants regulated under the CAA such as carbon monoxide, hydrocarbons, and nitrogen oxide were reduced by 10–20% using RFG in conventional automobiles compared to that with conventional gasoline. The changes in gasoline properties included an increase in oxygen content, decrease in alkenes, aromatics, benzene, and sulfur content, reduction in volatility, and decrease in distillation temperatures.[4] Evaporative VOC emissions can be reduced by limiting the vapor pressure of RFG during the summer months. The limit is lower for the southern part of states because higher ambient air temperatures increase evaporative losses. The RFG vapor pressure limits are lower than the current Environmental Protection Agency (EPA) limits. Lowering RVP increased the refiner's cost of producing gasoline because low-cost normal butane must be removed from the gasoline pool. Phase 2 required approximately a 1.3 psi reduction in RVP (from 8 to 6.7 psi) in northern control areas (Region 2) and a 0.4 psi reduction (from 7.1 to 6.7 psi) in southern areas (Region 1) from Phase 1 RFG levels during the summer months. Reid vapor pressure plays an important role in reducing VOC emissions. However, reduction in RVP alone is not enough to achieve the required Phase 2 VOC reduction. A reduction in VP to 6.7 psi will reduce VOC emissions by about 24% in Region 1 and 22% in Region 2, well below the 29% and 27.4% required in Regions 1 and 2, respectively. Reduction in sulfur from 300 to 140 ppm will yield an additional reduction of 1.9%. Lowering aromatics from 32 to 26 vol% in Phase 1 provided an additional 1.5% VOC reduction that is still not enough to meet Phase 2 VOC reduction. The final necessary emission reductions must come from increasing E200, E300, and olefins, without violating the NO_x emission reduction requirement.[6]

TOTAL OXYGEN CONTENTS AND TYPE OF OXYGENATES

Oxygenates represent a key component of RFG. This fuel is required to contain an average of 2.1% oxygen by weight. The source of this oxygen is oxygenates organic combustible liquids containing an oxygen atom in their structure. While methyl tertiary-butyl ether (MTBE) and ethyl alcohol (ethanol) have been the oxygenates most commonly used in gasoline, ethyl tertiary-butyl ether (ETBE) and tertiary-amyl methyl ether (TAME) are likely to be used in RFG as well.[5] As the different oxygenated compounds contain different amounts of oxygen, the amount of oxygenate needed to meet the RFG oxygen requirement depends on which oxygenate the refiner/blender uses. Initially, it was believed that the oxygen requirement would advance the RFG goals but lately on subsequent testing has shown that oxygen content has very little effect on VOC or toxic emissions.

Oxygenates are either alcohol or ether. The purpose of oxygenates in gasoline is to reduce CO and hydrocarbons emission by assisting fuel combustion. Since 1979, oxygenates have been added to the fuel in the form of MTBE and ethanol as octane enhancers to replace lead in limited areas of the country. During the 1980s, oxygenates were used on a large scale because of the implementation of oxygenated gasoline programs to control emissions of CO in cold weather. In 1990, the CAA Amendments required the use of a high percentage of oxygenates in gasoline. For example, seasonally 15% MTBE by volume or year-round 11% MTBE is added in parts of the United States where CO levels in the winter or ozone concentrations in the summer exceed the National Ambient Air Quality Standards. Several researchers have studied the effect of oxygenates in gasoline on pollutant emissions. Kirchstetter et al. reported no change in NO_x emissions

and decreases in CO and VOC for the oxygenated gasoline.[8] On the other hand, Keller et al. reported no significant change in exhaust emissions from advanced technology vehicles when comparing oxygenated and nonoxygenated gasolines that meet all other California Phase 2 RFG standards.[9] The emission reduction benefit of oxygenated gasolines is highest for older vehicles with open loop control. Newer vehicles with computerized closed-loop control get a lesser benefit, and there is only a small benefit for advanced technology vehicles. These vehicles have a feedback emission control system that takes away the majority of the effect of the oxygenate.[10] However, the oxygenate requirement remains because of the following reasons: 1) it is an octane booster; 2) it is helpful to older vehicles; and 3) it is domestically produced, displacing imported petroleum. The most commonly used oxygenate in the United States was MTBE because of its compatible blending properties, high octane, low vapor pressure, availability, and low cost. However, this oxygenate is being phased out because of its detection at low levels in groundwater and possible toxicity.[10] Several studies have been conducted to find an alternative to MTBE.[11,12] At this moment, ethanol is the only approved replacement.

DIFFERENT PHASES OF REFORMULATED GASOLINE PROGRAM

Phase I of RFG

Phase I of RFG was introduced on January 1, 1995. It contains both formula standards and performance standard.[13] These standards for RFG are described below.

Formula standards

The formula standards for RFG are as follows: maximum of 1 wt% benzene, maximum of 25 wt% total aromatic content, 2 wt% minimum of oxygen, no metals and detergent additives.

Performance standards

Because of the various needs for suppliers, consumers, regulators and the complexity of predicting fuel emissions, the development of performance standard involved different groups such as the EPA along with the petroleum industry, auto manufacturers, oxygenate producers, the Department of Energy, various environmental groups, and state air pollution directors, etc.[13] Two models were developed by the EPA to predict emission performance from fuel composition. The models are described below:

Simple Model. The simple model was available for use from 1995 to 1997. In the simple model it was believed that RVP, oxygen, benzene, and aromatics will have a significant effect on emissions. However, the effects of aromatics on VOC and NO_x emissions were unclear. During the ''high-ozone season'' (summertime), the performance characteristics of the RFG were estimated to reduce total car VOCs and total toxic emissions by 15% of the 1990 baseline levels, or to meet the equivalence of a formula fuel performance, whichever is more stringent.[13]

The simple model was not directly used for monitoring VOC emissions. The EPA determined that a fuel with an RVP of 8.1 psi and 2.0 wt% oxygen would be sufficient to achieve the minimum 15% VOC emission reduction. Baseline emission levels are from 1990 model year (MY) vehicles operated on a baseline gasoline (Table 1). The baseline fuel, derived from average survey data of the compositions of gasolines in the United States in 1989, is shown in Table 1. This baseline fuel was defined to indicate a performance standard and not a formula composition standard. To achieve equivalency certification, refiner and importer's gasoline must comply with the CAA requirements for emissions as shown in Table 2.[14]

The simple model allows a wide variety of gasolines to comply with the regulations without specifying or limiting any other gasoline properties. For example, the formula states that oxygen content be at least 2 wt% but it does not stipulate the form of oxygen. This leaves the choice to the refiner as long as the fuel is blended to contain 2 wt% oxygen. Also, the RFG must produce no increase in NO_x emissions because NO_x and hydrocarbons react in the lower atmosphere or troposphere to form ozone in the presence of sunlight.[13]

Table 1 Baseline gasoline properties defined by the EPA in 1989

Property	Summer	Winter
API gravity	57.4	60.4
RVP (psi)	8.7	11.7
IBP (°F)	91	87
10% (°F)	128	111
50% (°F)	218	199
90% (°F)	330	332
EBP (°F)	415	404
Aromatics (vol%)	32	26.4
Olefins (vol%)	9.2	11.9
Saturates (vol%)	58.8	61.7
Sulfur (ppm)	339	338
Benzene (vol%)	1.53	1.62
Octane $(R + M)/2$	87.3	88.1

R, rated; *M*, measured.
(From Ref.[13].)

Table 2 Simple model baseline emissions

	Summer		
	Region 1	**Region 2**	**Winter**
Exhaust VOCs (g/mL)	0.444	0.444	0.656
Nonexhaust VOCs (g/mL)	0.858	0.766	0
Total VOCs (g/mL)	1.3	1.21	0.656
Exhaust benzene (mg/mL)	30.1	30.1	40.9
Evaporated benzene	4.3	3.8	0.0
Running loss benzene	4.9	4.5	0.0
Refueling benzene	0.4	0.4	0.0
1,3-Butadiene	2.5	2.5	3.6
Formaldehyde	5.6	5.6	5.6
Acetaldehyde	4.0	4.0	4.0
POMs	1.4	1.4	1.4
TAPs (mg/mL)	53.2	52.3	55.5

Region 1 is southern areas of the United States typically covered by ASTM class B during summer. Region 2 is northern areas of the United States typically covered by ASTM class C during summer. (From Ref.[13].)

Complex Model. A complex model is a set of statistically derived equations that relate fuel properties to vehicle emissions. This model became mandatory in 1998. The simple model calculates emission based on a fuel's RVP, oxygen, aromatic, and benzene content whereas the complex model adds four more variables (sulfur, olefin, and the 200°F and 300°F distillation volume fractions) to the equation. This model is based on the data collected from programs conducted around the United States. The database was made up of over 200 test fuels, 500 automobiles, and 5000 emission testings.[15] The complex model can be divided into two portions: exhaust and nonexhaust. The nonexhaust VOC was derived directly from the simple model approach where the nonexhaust benzene was modeled as a weight fraction of nonexhaust VOC from the headspace model of General Motors. The exhaust model was based on 19 different test programs.[13]

Phase II of RFG

Phase II of the RFG program began at refineries on December 1, 1999, and at retails outlets beginning January 1, 2000. This part of the RFG requirements used the complex model with stricter standards, as follows: a 25% reduction of VOCs and toxic emissions from baseline levels, with a latitude granted to the EPA encompassing consideration of technological feasibility and cost.[13] The Phase II complex emissions model uses the same variables as the Phase I complex emissions model. However, the estimated emissions using the Phase II model are different from those predicted by the Phase I model. An approximate comparison between Phase I and Phase II is given in Table 3. The VOC, NO_x, and TAP emissions reduction performance standards under Phase I using Phase I complex emissions model and under Phase II using Phase II complex emissions model are not directly comparable because of the differences between the Phase I and Phase II complex emissions models.[6] From Table 3, it was observed that Phase I winter RFG comes very close to meeting the Phase II RFG requirement.

AIR QUALITY BENEFITS OF USING RFG

The emissions from vehicles impact our environment in three different ways. They impact our air quality by contributing toxins to the air, by interacting with other emissions and sunlight to form ground-level ozone (smog) and by staying in the atmosphere as greenhouse gases.[16] The literature of the actual emissions from oxygenated fuels is quite limited. In 1987, the EPA found that oxyfuels, especially gashols, can reduce CO emissions by 10–30% in high-altitude areas, whereas in 1988 the EPA reported that CO reductions depend on the percentage of oxygen in the gasoline along with engine and exhaust-system technology. Analysis of fuel data submitted to the EPA by industry for compliance purposes shows that emission reductions from the RFG program have been more than the program requires each year since 1995. The Auto/Oil Air Quality Improvement Research Program (Auto/Oil) study highlights changes in fuel formulations that reduce automotive pollutants, especially photochemical ozone precursors and mobile air toxic emissions. This study concluded that California Phase 2 RFG reduces fleet average hydrocarbon emissions by 10–27% compared to industry average gasoline.[17] The estimated reductions in total air toxics emissions from the use of California Phase 2 RFG relative to conventional gasoline range from 9% to 32%, largely because of the reduced benzene and aromatic content of California Phase 2 RFG.[18] This gasoline is associated with significant decreases in benzene emissions and increases in formaldehyde emissions, as well as minor decreases in 1,3-butadiane and acetaldehyde emissions.[19]

The Auto/Oil study compares the emissions from vehicles using nonoxygenated RFG to vehicles using 11 vol% RFG.[18,20] This study showed only a 13% decrease in formaldehyde tailpipe emissions with the oxygenate-free fuel. The Auto/Oil study also evaluates engine-out and tailpipe (exhaust) emissions for 157

Table 3 Reformulated gasoline averaging standards

	RFG Phase 1 (January 1995–December 1999)			RFG Phase 2 (January 2000)		
	Summer region 1	Summer region 2	Winter	Summer region 1	Summer region 2	Winter
Product quality standard						
Oxygen (wt% min)	2.1	2.1	2.1	2.1	2.1	2.1
Benzene (vol% max)	0.95	0.95	0.95	0.95	0.95	0.95
Performance standards (using Phase II complex emissions model), percent reduction required						
Toxic air pollutants (TAP) (%)	18.5	17.8	17.3	21.5	21.5	21.5
Volatile organic compounds (%)	20.8	10.5	n.a[a]	29.0	27.4	n.a[a]
Nitrogen oxides (%)	1.4	1.6	1.7	6.8	6.8	1.5

[a]Not applicable.
(From Refs.[6,16].)

hydrocarbons for different fuel formations such as California Phase 2 RFG with MTBE, nonoxygenated California Phase 2 RFG, and a conventional nonoxygenated gasoline.[20] This study shows that California Phase 2 RFG with MTBE is associated with large increases in MTBE emissions and a lesser increase in isobutylene emissions and smaller increases in formaldehyde, acetone, and propadiene emissions compared to nonoxygenated RFG. Kirchstetter et al. compared the on-road emissions associated with California Phase 2 RFG and a California Phase 1 RFG.[4,21] The comparison was made by measuring the tunnel emissions for two sequential years when different statewide gasoline formulation standards were in effect. On the basis of representative samples of service stations, California Phase 2 RFG contained, on an average, 10.7 vol% MTBE in the summer of 1996 compared to 1.0 vol% MTBE for the Phase 1 formulation in the summer of 1995. Kirchstetter et al. also observed an 18% reduction in CO emissions, a 22% reduction in nonmethane organic carbon emissions, and a 6% decrease in NO_x emissions from the summer of 1995 (California Phase 1 RFG) to the summer of 1996 (California Phase 2 RFG).[4,21] Gertler et al. performed a tunnel study in the Sepulveda Tunnel in Los Angeles in October 1995 (Phase 1 California RFG) and July 1996 (California Phase 2 RFG).[22] A significant decrease in CO and NO_x vehicle emission was observed for 1996 California Phase 2 RFG compared to the California Phase 1 RFG measurements in 1995. In their study, the most important change in the hydrocarbons measured from the 2 yr was a nearly 50% increase in the faction of MTBE emissions.

Case Study 1

In Mexico, 76.7% of the population lives in urban areas. It has three mega cities with a population ranging from 1.5 to 8.0 million. In particular, Metropolitan Area of Mexico City (MAMC) alone has more than 18 million people. This urban region is expected to be one of the 20 largest urban regions of the world by the year 2015.[23,24] In the year of 2001 Mexico started a program to limit the exhaust emissions on new gasoline vehicles. Schifter et al. studied the fuel reformulation on the vehicle exhaust emissions in MAMC.[24] In their work, they prepared 15 gasoline blends using the current refinery streams available in the country. The RVP, sulfur, aromatic, olefin, and oxygenated molecule (MTBE or ethanol) were the main parameters considered in their studies. The United States Federal Test Procedure and FTP were used to evaluate total hydrocarbons, nitrogen oxides, and toxic exhaust emissions. The statistical design and analysis methods similar to those employed in the Auto/Oil Air Quality Research program were used in this work. In their work, different ranges of MYs were used, which acted as a surrogate for different emission control technologies. Thirty vehicles such as GT-1 (GM: Cutlass-89; Chrysler: Spirit-90; VW: Sedan-90; Nissan: Tsuru-90), GT-2 (Chrysler: Spirit-93; Nissan: Tsuru-93; GM: Cavalier-94; VW: Jetta-98; Nissan: Sentra-98; GM: Monza-98; Ford: Pick-up-98; Ford: Escort-98; Chrysler: Stratus-98), and GT-3 (Nissan: Tsuru-90; Chrysler: Neon-99; GM: Chevy-99; Chrysler: Stratus-99; Dodge: Pick-up-99; Nissan: Pick-up-2000; VW: Pointer-2000; Nissan: Sentra-2001; GM: Monza-2001; Ford: Fiesta-2002; GM: Astra-2002; Ford: Focus-2001; VW: Jetta-2001; VW: Sedan-2001) were randomly recruited in areas of the MAMC of different socioeconomic indicators. Addition of MTBE deceased CO emissions to some extent in all cases. However, it was statistically significant in the case of the GT1 fleet. Substitution of MTBE by ethanol decreased CO emissions in the entire fleet and it was statistically significant only for GT-1 and GT-2 vehicles. Methyl

tertiary butyl ether had a statistically significant impact on total hydrocarbon emissions for the entire fleet. A similar impact on GT1 and GT2 was observed when ethanol was added, whereas its effect on GT3 was significant. Increase in RVP resulted in lower emissions, which was statistically significant on GT-2 and GT-3 fleets. Increase in olefins and aromatics did not have any significant impact. However, concerning NO_x emissions, the only statistically significant impact was that increasing olefins raises emissions on GT-1. For selected fuels (see Table 4) a comparison of calculated straight average values of VOCs, TAP, and NO_x was made using the EPA Phase 1 and Phase 2 complex model. For complex model predictions, see Table 5. From Table 5 it was observed that F-10 had the highest reduction among the tested fuels with the 1990 MY's American fleet.

Case Study 2

The effects of the use of reformulated and oxygenated gasoline fuel blends on the air quality of European city were studied by Vinuesa et al.[25] The urban region of the Strasbourg–Kehl area (GSKA) in the middle of the upper Rhine valley was chosen as the area of investigation. The part of the valley used in this work was flat, 29 km wide, and 33 km long. The chosen period of investigation was May 9–15, 1998, which corresponded to a photochemical ozone pollution episode with low synoptic wind and high temperature (3–5°C above the seasonal average). Their study focused mainly on May 11 when the maximum value of 193 $\mu g/m^3$ of ozone was measured in the center of Strasbourg. Ethyl-*t*-butyl ether was used as an oxygenated compound in their work. The daily emissions of CO, VOC, and NO_x (in milligrams or tons) for May 11, 1998, are given in Table 6. The goal of their work was to examine the differences in pollutant concentrations predicted by the model when the gasoline vehicles used different fuel blends. The wind fields were exactly the same in the reference case and in different scenarios and consequently the concentration levels were explained in a similar way. The effects of alternative fuels on the urban air quality were studied by building emission scenarios based on available emission factors. From their work they observed that the reformulation directly decreased the aromatic fraction in the composition of the fuel and the oxygenated compounds decreased this part by dilution. They also observed that the use of alternative fuels considerably reduced the direct emissions of total VOC by up to 30% and 45%, respectively. For NO and NO_2 the situation was less clear. For all the scenarios, the simulated ozone levels were slightly lowered in the range of 1–5% and the maximum effect was obtained when all the vehicles used these alternative fuel blends. When the percentage fell to 80%, the reduction of ozone became less than 3%. However, from their work they concluded that the use of reformulated and oxygenated fuel blends led to some improvement of the air quality at a local scale but it did not lead to the drastic changes even when all the fleets use various alternative fuels.

Table 4 Characteristics of gasoline motor vehicle fuels

	Fuel			
Properties	**I/A**[a]	**MA**[a]	**F-10**[a]	**F-13**[a]
Aromatics (vol%)	28.0	24.1	19.8	40.3
Olefins (vol%)	13.5	9.0	5.0	4.8
Oxygen, as MTBE (wt%)	0.34	1.21	1.03	1.14
Benzene (vol%)	1.1	1.0	0.6	1.1
RVP (psi)	8.88	7.66	6.62	8.06
RON	91	91	91	93
MON	84	84	84	85
(RON + MON)/2	88	87	88	89
D-86 distillation				
IBP (°C)	35.9	37.9	40.5	39.8
10% (°C)	57.9	59.5	68.4	65.9
50% (°C)	97	101	104	111
90% (°C)	167	163	168	162
EP% (°C)	203	205	205	200
Sulfur (ppm)	724	403	89	34

[a]Fuel code names.
(From Ref.[24].)

Table 5 Comparison of Schifter et al.'s work with the EPA complex model

Fuel	Total VOC	NO_x	Total toxics
RFG-1[a]	−17.0	−1.5	−17.0
RFG-2[b]	−27.0	−7.0	−22.0
I/A	3.82	9.36	−1.66
MA	−17.44	−0.13	−19.86
F-10	−29.68	−12.8	−33.94
F-13	−5.74	−12.93	12.84

[a]Phase 1 complex model.
[b]Phase 2 complex model.
(From Ref.[24].)

Impacts of Oxygenated Compounds on the Environment

In RFG, the most commonly used oxygenates are MTBE (methyl tertiary butyl ether) and ethanol. As a component of RFG, MTBE plays a role in reducing total hydrocarbons, air toxics, and CO.[18] However, it is also responsible for certain pollutant emissions.

Table 6 Daily emissions of CO, VOC, and NO_x for 11 May 1998

	CO emisssions		VOC emissions		NO_x emission	
Sources	**mg**	**% CO**	**mg**	**% VOC**	**mg**	**% NO_x**
Biogenic (forest)			4.14	15.0		
Residential housing	12.9	18.7	1.52	5.5	0.77	2.4
Residential solvent use			5.3			
Ternary sector housing	0.17	0.3	0.02	0.1	0.36	1.1
Industries (point sources)	0.81	1.2	6.05	22.0	5.25	16.5
Other industries	0.17	0.2	0.04	0.1	1.10	3.5
Petrol station (distribution)			0.85	3.1		
Petrol station (storage)			0.69	2.5		
Road traffic (line sources)	44.6	64.7	7.07	25.7	19.8	62.2
Road traffic (surface sources)	6.0	8.7	1.00	3.6	2.1	6.4
			2.28	8.3		
Road traffic evaporation	3.7	5.4	2.18	7.9	0.75	2.4
Air traffic	0.07	0.1	0.03	0.1	0.14	0.4
Railway traffic	0.53	0.8	0.22	0.8	1.62	5.1
Fluvial traffic	68.9	100	27.5	100	31.9	100
Total						

(From Ref.[25].)

Methyl tertiary butyl ether enters the environment in different ways such as auto emissions, evaporative losses from gasoline stations and vehicles, storage tank releases, pipeline leaks, and accidental spills, and refinery stock releases. When gasoline is released into air, a significant portion exists in air and a small portion enters into soil and water.[26] When it is released into water a significant portion remains dissolved in surface water, with some partitioning into air and a much smaller amount into soil.[27] Because of its high solubility, it moves through the soil and into the groundwater more rapidly than other chemicals.[26]

Time spent at the service station, driving cars, at the parking garage and in homes with attached garage are the nonoccupational sources of gasoline emission to the environment and human exposure occurs through inhalation during fueling of automobiles.[28] Also, leakage from stationary sources (under ground storage tank) to soil or groundwater can contaminate the water supply. Methyl tertiary butyl ether has a strong taste and odor. A small amount of MTBE in water can make the water supply distasteful. Dermal contact (exposure of skin to toxic materials) of gasoline may occur though accidental spills of MTBE-blended gasoline or through the use of gasoline as a solvent.[29]

The estimated arithmetic mean occupational dose via air is in the range of 0.1–1.0 mg/kg/day, whereas doses from residential exposures, communicating, and refueling are in the range of 0.0004–0.006 mg/kg/day.[30] During refueling of a car, the concentrations of MTBE range from less than 1 ppm to 4 ppm (within the breathing zone) and from 0.01 ppm to 0.1 ppm inside cars (1 ppm = 3.57 mg/m^3 at 25°C, 1 atm).[31] However, the reference dose imposed by the EPA is 42 mg/kg/day.

The data on the presence of MTBE in drinking water are very limited. The most extensive monitoring data on MTBE in drinking water are available from California. In January 2000, 1444 systems had tested 6492 sources of drinking water. Methyl tertiary butyl ether was detected in 52 (0.8%) of these sources, including 31 of 6076 groundwater sources (0.5%) and 21 of 416 surface water sources (5%). Overall, 30 (2.1%) of the 1444 public water systems reported detection of MTBE in at least one of their drinking water sources. Although the state database did not include some contaminated wells that have been closed, very few sources had MTBE concentrations exceeding the EPA taste and odor drinking water recommendation of 20–40 μg/L.[26]

ALTERNATIVES OF MTBE

Because of the adverse effects of MTBE, ethanol is now considered as a promising alternative oxygenate. Ethanol is now marketed and distributed in every state in the United States. Ethanol is safe, biodegradable, and renewable. It also does not harm drinking water resources. Currently, around 80% of gasoline in California is blended with ethanol. At this time, the total ethanol consumption in California is 950 million

gallons per year, whereas in 2002 the consumption was 100 million gallons.[32]

But other concern also arises in the case of ethanol. It raises vapor pressure of the ethanol–gasoline mixture, resulting in increased evaporative emissions.[10] Also, there are cost and supply issues if a large amount of ethanol is required.[11,12]

CONCLUSIONS

The RFG program is a significant step toward cleaning the air we breathe. Evaporative VOC emissions can be reduced by limiting the vapor pressure of RFG during the summer months. The RFG program reduces emissions of TAP such as benzene, a known human carcinogen. It is possible to reduce the tailpipe emissions of the major pollutants regulated under the CAA-carbon monoxide, hydrocarbons, and nitrogen oxide by 10–20% using RFG. Methyl tertiary butyl ether oxygenate is being phased out because of its adverse effects on the environment. Ethanol is now considered as a promising alternative to MTBE.

NOMENCLATURE

API	American Petroleum Institute
ASTM	American Standards for Testing and Materials
CAA	Clean Air Act
EBP	End boiling point
ECS	Emission control system
ETBE	Ethyl tertiary butyl ether
E200	Percentage of fuel evaporated at 200°F
E300	Percentage of fuel evaporated at 300°F
MTBE	Methyl tertiary butyl ether
IBP	Initial boiling point
POM	Particulate organic matter
RFG	Reformulated gasoline
RVP	Reid vapor pressure
TAP	Toxic air pollutants
TAME	Tertiary amyl methyl ether
VOC	Volatile organic compound

REFERENCES

1. Cannon, J.S.; Azimi, L.S. Reformulated gasoline: cleaner air on the road to nowhere. Int. Assoc. Hydrogen Energy **1995**, *20* (12), 987–994.
2. Treiber, S.; McLeod, R.S.; Faitakis, Y.; Hutchings, R.L. The challenge to conventional blending technology. Hydrocarbon Process. **1998**, *77* (6), 101–113.
3. Bowman, F.M.; Seinfeld, J.H. Atmospheric chemistry of alternate fuels and reformulated gasoline components. Prog. Energy Combust. Sci. **1995**, *21*, 387–417.
4. Cohen, J.P.; Yarwood, G.; Noda, A.M.; Pollak, A.K.; Morris, R.E. *Auto/Oil Air Quality Improvement Research Program: Development of Emissions Reactivity Values for Phase II Results,* SYSAPP94-94; Systems Applications International: San Rafael, CA, July 1994.
5. http://www.chevron.com/prodserv/fuels/bulletin/fed-refm/fg-char.shtml.
6. Lidderdale, T.; Bohn, A. Demand and price outlook for Phase 2 reformulated gasoline, 2000. Energy Inf. Admin./Pet. Supply Monthly, **1999** Apr.
7. Yang, Y.-J.; Milford, J.B. Quantification of uncertainty in reactivity adjustment fcators from reformulated gasolines and methanol fuels. Environ. Sci. Technol. **1998**, *30*, 196–203.
8. Kirchstetter, T.W.; Singer, B.C.; Harley, R.A.; Kendall, G.R.; Chan, W. Impact of oxygenated gasoline use on California light duty vehicle emissions. ES & T **1996**, *30*, 661–670.
9. Keller, A.; Froines, J.; Koshland, C.; Reuter, J.; Suffet, I.; Last, J. *Health and Environmental Assessment of MTBE in Gasoline*, Report to the Governor and Legislature of the State of California as Sponsored by SB521, Vol. 1; Summary and Recommendations, Nov 1998.
10. Maclean, H.L.; Lave, L.B. Evaluating automobile fuel/propulsion system technologies. Prog. Energy Combust. Sci. **2003**, *29*, 1–69.
11. California Energy Commission. *Supply and Cost of Alternatives to MTBE in Gasoline*, Staff Report, Oct 1998; P300-98-013.
12. California Energy Commission. *Supply and Cost of Alternatives to MTBE in Gasoline*, Technical Appendices, Staff Report, October 1998; P300-98-013B.
13. Khan, R.M.; Popiel, E.; Reynolds, G.J. Designing fuel for the environment: reformulated gasoline. Energy Sour. **1996**, *18*, 513–523.
14. Federal Register. Regulation of fuels and fuels additives; standards for reformulated and conventional gasoline. Fed. Reg. **1994**, *59* (32), 7716
15. Korotney, D. The complex model made simple, Paper PETR-4. American Chemical Society National Meeting, Washington, DC, Aug 1994.
16. Rask, K. Clean air policy and oxygenated fuels: do we get what we pay for. Energy Econ. **2004**, *26*, 161–177.
17. Auto/Oil Air Quality Research Improvement Program (Auto/Oil). *Gasoline Reformulation and Vehicle Technology Effects on Exhaust Emissions*, Technical Bulletin No. 17, 1995.

18. Auto/Oil Air Quality Research Improvement Program (Auto/Oil). Program Final Report, 1997.
19. Franklin, P.M.; Koshland, C.P.; Lucas, D.; Sawyer, R.F. Evaluation of combustion by-products of MTBE as a component of reformulated gasoline. Chemosphere **2001**, *42*, 861–872.
20. Auto/Oil Air Quality Research Improvement Program (Auto/Oil). *Phase I and Phase II Test Data Public Release on Compact Disc*; Systems Application International, Inc.: San Rafael, CA, 1996.
21. Kirchstetter, T.W.; Singer, B.C.; Harley, R.A. Impact of California reformulated gasoline on motor vehicle emissions. 1. Mass emission rates. Environ. Sci. Technol. **1999**, *33*, 318–328.
22. Gertler, A.W.; Sagebiel, J.C.; Dippel, W.A.; O'Connor, C.M. The impact of California phase 2 reformulated gasoline on real world vehicle emissions. J. Air Waste Manage. Assoc. **1999**, *49*, 1339–1346.
23. Bravo, H.A.; Torres, R.J. The usefulness of air quality monitoring and air quality impact studies before the introduction of reformulated gasolines in developing countries. Mexico city, a real case study. Atmos. Environ. **2000**, *34*, 499–506.
24. Schifter, I.; Diaz, L.; Vera, M.; Guzman, E.; Lopez-S, E. Fuel formulation and vehicle exhaust emissions in Mexico. Fuel **2004**, *83*, 2065–2074.
25. Vinuesa, J.-F.; Mirabel, P.; Pnche, J.-L. Air quality effects of using reformulated and oxygenated gasoline fuel blends: application to the Strasbourg area (F). Atmos. Environ. **2003**, *37*, 1757–1774.
26. Ahmed, F.E. Toxicology and human health effects following exposure to oxygenated or reformulated gasoline. Toxicol. Lett. **2001**, *123*, 89–113.
27. WHO (World Health Organization of the United Nations). *Environmental Health Criteria 206: Methyl Tertiary-Butyl Ether*; WHO: Geneva, Switzerland, 1998.
28. Dourson, M.L.; Felter, S.P. Route-to route extrapolation of the toxic potency of MTBE. Risk Anal. **1997**, *17*, 717–725.
29. NRC (National Research Council). *Toxicological and Performance Aspects of Oxygenated Motor Vehicle Fuels*; National Academy Press: Washington, DC, 1996.
30. Brown, S.L. Atmospheric and potable water exposures to methyl tert-butyl ether (MTBE). Reg. Toxicol. Pharmacol. **1997**, *25*, 256–276.
31. Hartle, R. Exposure to methyl tert-butyl ether and benzene among station attendants and operators. Environ. Health Perspect. **1993**, *101* (6), 23–26.
32. http://www.ethanolrfa.org/leg_position_mtbe.shtml.

Renewable Energy

Gareth P. Harrison
Institute for Energy Systems, School of Engineering and Electronics, University of Edinburgh, Edinburgh, U.K.

INTRODUCTION

Climate change is widely regarded as one of the most significant global challenges in the 21st century. With conventional fossil fuel generation a major contributor, the expansion of renewable energy use is a critical element of the strategy to lower emissions of greenhouse gases. Many countries have set challenging targets for renewable use and deployment driven by the environmental, sustainability, and security benefits that may be attributed to renewables.

Here, these arguments are reviewed along with the renewable technologies available. The challenges for integrating these variable and often intermittent generating sources are highlighted and alternative applications in creating new energy vectors and in direct use are explored. Finally, the economics of renewables are examined with particular reference to the limitations of traditional comparisons with conventional sources and the impact of more robust techniques on the relative cost of fossil fuels and renewables.

THE DRIVE FOR RENEWABLE ENERGY

The case for promoting renewable energy revolves around the three key benefits associated with it:

1. Nonpolluting
2. Infinite reserves
3. Security of supply

Pollution

The primary argument in favor of renewable energy is that it does not entail the release of chemical pollutants to convert the energy. This is in contrast to fossil fuels, which emit a range of pollutants from carbon dioxide (CO_2) to the components of acid rain to ash. Table 1 shows the pollutants from typical 2000 MW plants.

The combustion of fossil fuels produces the majority of anthropogenic CO_2 with transport and power generation the largest sources. CO_2 is widely accepted as the major cause of climate change, which, if left unchecked, is predicted to lead to a global temperature rise of between 1.5°C and 5.8°C by the end of the century (Fig. 1).[2] This, and accompanying changes in other climate variables (precipitation, wind speed, etc), will have impacts on many sectors ranging from agriculture to human health that will be seen at local, regional, and global scales. In responding to this, many industrialized countries have signed the Kyoto Protocol requiring cuts in CO_2 and other gases by 2010, which became legally binding on participants in February 2005 following ratification by Russia. A difficulty for the Kyoto agreement is the refusal of the U.S.A. and Australia to ratify it, partly because large developing nations like China and India escape emissions limits. In July 2005, the U.S.A. and five Asia-Pacific states announced a voluntary pact to reduce emissions by developing new technology like 'clean coal;'[3] the effectiveness of this nonbinding agreement is disputed, however. These aggrements are the beginning of a longer process, with the United Kingdom Royal Commission on Environmental Pollution recommending CO_2 cuts of 60% by 2050 to limit the eventual rise of greenhouse gas concentrations to twice the preindustrial level.[4] Achieving the reductions required by Kyoto and the longer-term targets will not be straightforward. To date, modest CO_2 emissions have been achieved through switching to less carbon-intensive fuels like natural gas (albeit justified on a cost basis) but, ultimately, emission-free energy is required. While continued fossil fuel use will be possible if large-scale capture and sequestration of carbon can be achieved, the development and deployment of truly carbon-free energy sources like hydrogen, nuclear, and renewables will be critical.

Fossil-fueled power stations produce significant quantities of sulfur dioxide (SO_2) and nitrogen oxides (NO_x), which are precipitated as "acid rain" across wide areas and legislation such as the European Union Large Combustion Plant Directive has imposed emission limits in response. SO_2 emissions can be reduced by a range of techniques including the treatment of coal before ignition, combustion processes that collect sulfur in the ash, and SO_2 removal from the combustion product gases, known as flue gas desulfurization (FGD). Although effective, FGD is expensive, particularly so for retrofit, and reduces station efficiency.[5] While road transport is the major cause of NO_x emissions, power

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007945

Table 1 Typical emissions from 2000 MW fossil-fueled power stations (in ktons/yr)

Pollutant	Conventional coal (no FGD)	Conventional oil	Combined-cycle gas turbine
Carbon dioxide	11,000	9,000	6,000
Sulfur dioxide	150	170	~0
Nitrogen oxides	45	32	10
Airborne particulates	7	3	~0
Solid waste and ash	840	~0	~0
Ionizing radiation (Bq)	10^{11}	10^{9}	10^{12}

(From Ref.[1].)

stations still contribute a significant fraction. A wide variety of NO_x control technologies have been developed (e.g., flue gas scrubbing with alkalis) but again are expensive. Despite not deploying NO_x scrubbing, the United Kingdom has achieved reductions with greater use of gas-fired plant and the retrofit of low-NO_x burners to existing coal stations. Use of renewables can avoid these costly removal processes.

Radiation is emitted from nuclear and, because of the presence of trace elements, from fossil-fueled power stations. The amounts are generally small, with public exposure similar to that of background radiation levels. However, the major issues surrounding nuclear power are how to deal with radioactive waste products and how to avoid proliferation of radioactive material. Until these issues are resolved and a clear lead given by governments, new nuclear plant is unlikely to be built, particularly not by the private sector. Once again, renewables can begin to fill the gap.

Finite Resources

There is also the issue of sustainability of fuel supplies. Fossil fuels are the product of sedimentation processes over millions of years, but they will have been consumed in a matter of centuries. Table 2 shows recent estimates of global proven fossil fuel reserves and an estimate of their remaining lifetime assuming current usage rates. While there is uncertainty over the remaining reserves, particularly given that further oil and gas fields will be identified, there is little doubt that with rapidly increasing demand for oil and gas and increasing exploitation costs (in financial and energy terms) supplies will be limited to several decades.

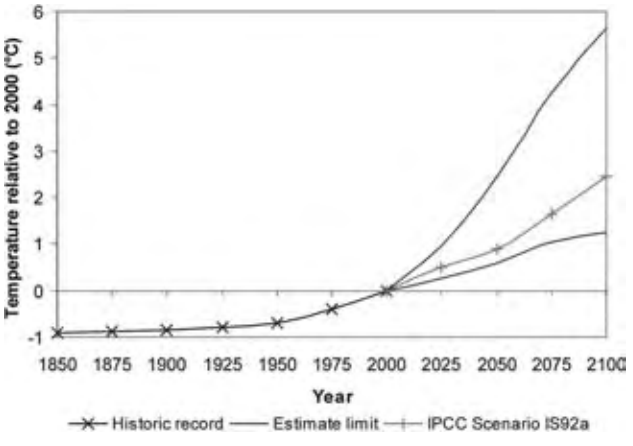


Fig. 1 History and range of future temperature rise. (From Ref.[2].) *(View this art in color at www.dekker.com.)*

Fortunately, there are sufficient renewable energy resources to meet our energy needs many times over, although in many cases the means of doing this in a technically and economically efficient manner is still under development. Table 3 shows an estimate of the power available from a range of renewables. To put these in perspective, global installed electrical capacity is currently just over 3000 GW (3 TW) and is anticipated to grow by some 120 GW/yr.

Although biomass is currently the most heavily used renewable in overall energy terms, very little is used to generate electricity. Accordingly, hydropower is the number one renewable in electricity terms and contributes around 19% of global supply. Installed capacity is currently around 650 GW with a further 100 GW under construction. In recent years wind has seen the largest growth and, with 8 GW added in 2003 alone, global installed capacity is now almost 40 GW; forecasts suggest an installed capacity of 1250 GW by 2020 meeting 12% of electrical needs.[8]

Security

An increasingly strong argument in favor of renewables is their positive effects on security by spreading risk. At the crudest level this can be interpreted as

Table 2 Proven fossil fuel reserves and reserve/production ratio

Fuel	Proven reserve	Reserve to production ratio
Oil	156 billion tonnes	41
Gas	176 trillion m^3	67
Coal	984 billion tonnes	192

(From Ref.[6].)

Table 3 Estimate of global renewable energy resource

Resource	Estimate of recoverable resource	Resource base
Solar radiation	1,000 TW	90,000 TW
Wind	10 TW	1,200 TW
Wave	0.5 TW	3 TW
Tides	0.1 TW	30 TW
Geothermal (heat flow)	—	30 TW
Biomass	1,150 TW	450 TW yr

(From Ref.[7].)

increasing energy self-sufficiency and lowering the risk of supply interruptions by accident, terrorist activity, or the actions of politically unstable fossil fuel exporting nations. In the U.K., fuel diversity is often indicated using the Shannon–Weiner index, which measures the logarithmic weighting of fuel technologies. Hence, by increasing the variety of fuel sources through new types and increasing volumes of renewables, diversity and, by implication, security increase.

Security can also be measured in terms of the economic implications of exposure to volatile fossil fuel prices. Studies suggests that fossil fuel price volatility has a range of negative effects on economic activity including impacts on employment levels and asset values; for example, a 10% oil price spike is estimated to reduce economic growth in the United States alone by as much as $200 billion over the following year.[9] This dwarfs current and future investment to commercialize renewables estimated at around $125 billion between 2001 and 2010.[10] The deployment of renewables reduces exposure to fossil price risk and this can be shown to lower overall cost.[11] This perhaps surprising result arises as the costs of renewables do not have any correlation with fossil fuel price changes. The addition of renewables creates a diversified generation portfolio that serves to lower overall generating costs for a given level of risk.

BRIEF REVIEW OF RENEWABLE ENERGY

There are a range of renewable energy sources currently in use in electricity systems or with the potential to contribute significantly in the future. The following energy sources are briefly reviewed and their merits explored:

- Hydropower
- Solar energy
- Wind energy
- Geothermal energy
- Biomass
- Tidal power
- Wave energy

Hydropower

Hydropower is the conversion of the gravitational potential energy of water into electricity. The power generated depends on the water flow rate and the height through which the water falls—the "head"—and is roughly 10 times the product of the two quantities. Global installed capacity is approximately 650 GW and produced 2700 TW hr in 2003 (around 19% of primary energy).[6] The estimated technical potential is around 14,000 TW hr/yr but the economically exploitable potential is between 40% and 65% of that, with much of the more economic plant already installed. The pattern of availability and exploitation varies significantly with countries like Norway producing 99% of their electricity from it while others little or none. The more developed regions have exploited far greater proportions of their resource while only around a small fraction of Asia's much larger potential has been tapped.

Hydropower is characterized by high initial capital costs offset by a long lifetime (civil engineering works often last more than 50 yr), high reliability, and low operational costs. Hydro is generally defined on the basis of its installed capacity as Table 4 indicates.

Large hydro is often defined as a grid-connected scheme in excess of 20 MW capacity and would normally possess a dam and a storage reservoir (Fig. 2). The largest scheme, the Three Gorges project on the Yangtze, will shortly have a capacity of 18 GW. Most large-scale schemes were developed prior to 1990 and the potential for new large schemes is now rather limited given that there are fewer commercially attractive sites still available, but perhaps more importantly, because of opposition on environmental grounds mainly as a result of the flooding of land to create the storage reservoir.

Fig. 2 Typical embankment dam in Texas, U.S.A. *(View this art in color at www.dekker.com.)*

Table 4 Hydro capacity definitions

Type	Capacity
Micro hydro	<100 kW
Mini hydro	<1 MW
Small hydro	<20 MW
Large hydro	>20 MW

Pumped storage schemes are large hydro schemes that are designed with a storage reservoir that is larger than justified by the natural flow of water into it. Alternatively, they may consist of a pair of reservoirs that are connected by appropriate penstocks. They operate such that when electricity demand is high, water is released from the upper reservoir driving the generators and exporting power to the grid; when electricity demand is low, the generators draw electricity from the grid (i.e., become motors), which are used to drive the turbines in reverse and pump water back into the reservoir. Pumped storage schemes play a major role in many countries in smoothing out imbalances between supply and demand and providing rapid response capabilities that help stabilize the grid. Their economics are based partly on these balancing and response capabilities but also on the fact that electricity at low demand is cheaper than at peak.

Schemes of less than 20 MW are relatively common and can be extremely competitive with minimal environmental impacts. While these smaller schemes currently contribute only about 3% to global hydropower capacity, they make a significant contribution in many regions of the world, especially in rural or remote regions where other conventional sources of power are less readily available. Small schemes can be associated with a dam and storage reservoir or can consist of short diversion schemes that are termed ''run-of-river.'' Small and micro hydro can also provide competitive power and while per-unit capital costs are higher than for larger schemes these can often be reduced by use of existing structures or by refurbishing existing plant and equipment. The cost of generating power from smaller hydro schemes depends on site characteristics and the hydraulic head (with economic viability decreasing as the head reduces).

Solar Energy

Most renewable energy sources are, ultimately, concentrated forms of solar energy. However, there are two distinct approaches to generating electricity directly:

- Solar thermal engines
- Photovoltaic (PV) cells

Solar thermal engines use a solar collector to create a temperature sufficient to raise steam and drive a turbine-generator. There are a wide range of prototype stations around the world that, unsurprisingly, tend to be in areas that are hot, dry, and sunny. The schemes use some form of solar collector to concentrate the sun's rays to a point or series of points. These arrangements include parabolic dishes, solar collector fields that use long lines of planar parabolic mirrors, and power towers that rely on the surrounding heliostats to reflect light into the power chamber. The devices range from a few hundred kilowatt to several hundred megawatt schemes that are cofired with gas. A different approach makes use of the smaller temperature differential at the ocean surface and at great depth. Ocean thermal energy (OTEC) pumps water from both 1000 m at low temperature and at the surface to drive a power-vapor cycle.[12]

Photovoltaic (PV) cells are semiconductor devices that convert sunlight directly into electricity. It is a well-established technology particularly for sites that are far from the distribution network. The cost of PV has dropped significantly and is a promising, low-risk technology—a fact indicated by the involvement of oil companies in their large-scale manufacture. The greatest potential for developing PV is to integrate it into buildings as PV modules can substitute for roof or façade elements, hence reducing net costs, can be operationally integrated and, being relatively unobtrusive, are unlikely to face public opposition.[13]

Wind Energy

Winds are the product of pressure differences in the atmosphere and the wind speed at a given location continuously varies. The power available from the wind is proportional to the cube of the wind speed. Accordingly, wind turbines are generally sited in areas of high mean wind speed and development of taller, larger-capacity turbines has occurred to harness the greater wind speeds at height (owing to the wind shear effect near the ground). Wind speed is quite unpredictable and at a given point fluctuates second by second and over longer periods of time; given the power law, individual wind turbine output varies significantly. Fortunately, wind turbines are commonly connected in substantial groups with aggregation having a marked smoothing effect.[13] An enormous variety of designs have been created to extract power from the wind but the three-bladed horizontal axis (HAWT) model is now the industry standard (Fig. 3).

A desire for economies of scale and the harnessing of higher wind speeds has seen turbine capacities grow significantly to the current standard of around 2 MW. The increase in blade diameter has pushed the overall

Fig. 3 Lamb Rigg wind farm, U.K. *(View this art in color at www.dekker.com.)*

height of these machines to around 130 m. Larger machines are being used for offshore wind developments of which there are several in the United Kingdom and Denmark. Offshore sites tend to benefit from higher wind speeds and lower turbulence levels, resulting in higher energy capture. There are several other advantages of offshore wind as well as a number of negative aspects as Table 5 indicates.

Despite the trend toward larger generators there are a wide range of smaller devices available, including those in the tens of kilowatt range as well as a number of microturbines for battery charging, water heating, as well as newer grid-connected versions (e.g., Ref.[14]) which are designed to be connected to domestic property.

Table 5 Advantages and disadvantages of offshore wind

Advantages	Disadvantages
Reduced visual impact	Higher capital costs
Higher mean wind speed	Access restrictions in poor weather
Reduced wind turbulence	Submarine cables required
Low towers due to low wind shear	

A number of different approaches have been taken by wind turbine manufacturers in converting the rotational motion into electricity. Some of the main differences include:

- Fixed- or variable-speed operation
- Direct drive generators or the use of a gearbox
- Stall or pitch regulation

Fixed-speed wind turbines usually use an induction generator via a gearbox and capacitors are required to improve the poor power factor.[15] Variable-speed operation increases energy capture by allowing maximum efficiency over a wide speed range. This necessitates a power electronic converter and while this increases losses the rotor acts as a flywheel smoothing the output. Power limitation in high wind speeds is achieved by shedding load either by stall control, which relies on the aerodynamics, or pitch control, which actively turns the blade away from the optimal position. Beyond the cutout speed, which is normally around 25 m/sec, the turbine shuts down to minimize damage.

Geothermal Energy

Geothermal energy is sourced from the heat flowing from the interior of the Earth to the surface. While the overall energy flow is small relative to that of solar heating, in some places the heat flow is concentrated sufficiently to allow the exploitation of the steam and hot water in the ground to drive thermal power stations and produce electricity. Strictly speaking, geothermal energy is nonrenewable on human lifetimes, but it is often grouped with renewables as it is a natural energy flow rather than the exploitation of stored chemical energy.[16]

The major advantage of geothermal energy is that it produces controllable and predictable power and the plant exhibits high availability and fairly competitive capital and operational cost. Other than site preparation, the longer-term concerns with geothermal plant include ground subsidence, induced earthquakes, and the release of pollutant gases. The pollutants include carbon dioxide, methane, as well as hydrogen sulphide, the source of the rotten eggs smell often associated with geothermal plant. In most cases these concerns tend to be related to older plant where the reinjection of cooler water is not practiced.[16]

Biomass

Biomass is the Earth's living matter and is an enormous store of energy. Historically, biomass was the sole fuel source with material burned for heat and

animal and vegetable fats used for lighting. The primary technology was the processing of wood to produce charcoal and allowed temperatures sufficient to extract metals from their ores. While such technologies have been largely superseded in the developed world, biofuels still represent around 14% of global primary energy consumption and over a third of consumption in developing nations.[17]

A whole range of materials can be used as biomass including wood, straw, and sewage. The majority decompose quickly so are not very good long-term energy stores. In addition, their low energy density (relative to fossil fuels) increases transport costs. This means that biomass power generating units are relatively small compared to conventional plant, relying on local supply chains for feedstock. A range of processes can be used to extract the energy in the biomass:[17]

- Direct combustion of raw biomass.
- Combustion following physical processing (e.g., chipping, drying).
- Thermal/chemical processing to concentrate the fuel.
- Biological processing such as anaerobic digestion or fermentation.

The direct product of each of these products is heat, which may be used locally for heating or chemical processing, to raise steam to generate electricity or a combination of the two as combined heat and power. As such, modern biomass generation will use electrical technologies common in other thermal generating plants. This is particularly true where biomass (e.g., solid processed sewage) is cofired with coal; this is becoming increasingly common in the United Kingdom.

While biomass generally involves the burning of material it differs from fossil fuels in that no more heat or carbon dioxide is produced than would be produced by natural processes. As such, it is referred to as ''carbon neutral.'' However, it still produces CO_2 and a range of other pollutants including nitrogen oxides, ash, but virtually no SO_2. A further environmental benefit is that the combustion of landfill gas avoids accidental explosions and prevents the release of methane, which is more potent as a greenhouse gas.

Tidal Power

Tidal power can use either conventional or new technology to extract energy from the tides. It is usually best deployed in areas where there is a high tidal range, which includes many areas of the United Kingdom as well as the United States, New Zealand, and parts of the west coast of India.

The conventional approach to extracting tidal energy is to construct a barrage across an estuary or a bay. As the tide rises (floods) and falls (ebbs), it creates a height differential between the inner and outer walls of the barrage. Water can then flow through turbines installed in the barrage and drive generators. Some tidal barrages operate on both the rising and the falling tide, but others, particularly estuarine barrages, are designed to operate purely on the falling tide. With basic ebb or flood generation the installed capacity is used for only short periods of 3–6 hr in each tidal cycle producing a block of power that may or may not coincide with high demand. Fortunately, a degree of smoothing of the power output can be achieved and tides can be predicted to a high degree of accuracy. Technology for tidal barrages is essentially the same for hydro schemes with the barrage and axial flow turbines of similar construction to those used in low-head dams.

Very few tidal barrages are in existence worldwide with the exception of the 240 MW La Rance scheme in France and smaller installations in Nova Scotia, Russia, and China. The United Kingdom has a number of attractive sites because of its high tidal ranges including the Severn (8 GW potential) and the Mersey (700 MW).[18] The scale of these installations and the associated high capital costs mean that the necessary investment is unlikely to be forthcoming within the current privatized electricity supply industry. Their environmental impact is also a major constraint on large-scale developments.

The construction of a large barrier across an estuary would have a major impact on the local estuary although there would be benefits as well as detrimental effects. The La Rance scheme caused the local ecosystem to collapse, although it has regenerated.[18] There would be a tendency for the water behind the barrage to become less saline as a result of lower seawater inflows, which would tend to allow freshwater flora and fauna to extend seaward. Lower tidal current velocities, particularly on the ebb tide, will tend to reduce sediment erosion or increase sedimentation. A reduction in suspended sediments allows increased sunlight penetration stimulating biological productivity and water level changes would also impact on the mud flats, home to wading sea birds.[18]

A more benign approach is to extract energy from the tidal flows that occur between headlands and islands or in and out of estuaries. The power available in these tidal streams varies with the cube of the current velocity and while sea currents are typically around 3 m/sec, much lower than the minimum velocities required for wind turbines ($\sim$7 m/sec), the density of seawater is such that the output of tidal stream devices is much higher than equivalently sized wind generators. The energy flows are significant with around 7.5 GW of accessible resource in Scotland alone.[19]

Fig. 4 Impression of a large-scale vertical-axis tidal rotor. *(View this art in color at www.dekker.com.)*

Tidal stream technology, which directly exploits these currents, is relatively new but is presently generating considerable interest. Turbine rotors can be used to extract energy from the flows, much as a wind turbine does. Alternative designs use oscillating aerofoils or rotating vertical hydrofoils (Fig. 4) to pump oil and generate power. Prototype devices currently on test in the United Kingdom include the 300 kW SeaFlow turbine possessing an 11 m-diameter rotor attached to a monopile.[20] The tidal turbines can be arranged singly or in arrays, allowing a range of power outputs that avoid the massive bursts of power associated with barrage technology. In common with barrages, the output will be variable but to a great extent predictable. In addition to monopiles, tidal devices can be moored to the seabed or perhaps more cleverly secured by an ''active'' mooring device designed to allow the easy recovery of the tidal device itself. The ''Snail'' (developed at the United Kingdom's Robert Gordon University) possesses hydrofoils that press the device onto the seabed.

Wave Energy

The worldwide wave power resource potential is huge with global power potential estimated to be up to 10 TW, which is the same order of magnitude as world electrical energy consumption. The power in the waves varies with the square of the wave height and is proportional to the period. Island countries are well placed to exploit wave energy with what is considered a huge accessible resource. However, when practical constraints such as conversion efficiency, shipping lanes, and environmental restrictions are taken into account, the resulting practicable resource is significantly reduced.

Waves are primarily driven by the wind with the best wave climates found in the temperate zones (30–60° latitude) where strong storms occur. The best locations include the United Kingdom, Ireland, Portugal, and the Canadian west coast.[21] Attractive wave climates are also found within ±30° latitude where trade winds blow as the lower power levels are compensated by the lower wave power variability.

Since the early 1970s there has been research into developing means of harnessing the power of the waves and various machines have been developed. These fall broadly into three categories:

1. Oscillating water column (OWC) devices, which channel waves into constricted chambers so that as the waves flow in and out of the chamber, they force air in and out of the chamber. These air flows are in turn channeled through a Wells turbine (which turns in the same direction irrespective of the air flow direction), which is used to drive a generator. This type of machine is generally designed to be fixed on or near the shore (or for incorporation into breakwaters) although several floating OWCs are at the prototype stage. This kind of machine is the most advanced and is particularly advantageous when incorporated into coastal protection. The U.K.-based company Wavegen built a shoreline OWC device known as the ''Limpet'' on the coast of Islay in Scotland, the first grid-connected wave device.
2. Fixed or semifixed machines that harness the hydraulic head in the water that occurs at a submerged point as the wave passes over that point. The pressure differential is used by a variety of means to cause a fluid to flow in a circuit, which is then used to drive a turbine and generator. The Archimedes Wave Swing is such a device.
3. Devices that use buoyancy to cause relative movement to indirectly drive a generator. Examples include the heaving buoy, rafts, and the original wave device, the ''Duck'' designed by Professor Stephen Salter at the University of Edinburgh (Fig. 5). An Edinburgh spin-out company, Ocean Power Delivery, has developed the ''Pelamis,'' which resembles (and indeed is named after) a sea snake in that it is a long-articulated series of tubes that flex as waves run along the length of the device.[22] This device has recently been connected to the grid at the European Marine Energy Test Centre in Orkney, Scotland.

Wave energy converters will be among the least environmentally harmful energy sources with minimal chemical and shipping hazard risk and zero visual impact for offshore devices (although there are some with shoreline-based devices).[21] There are, however, as yet unconfirmed suggestions of low-frequency effects on marine mammals.

Fig. 5 Early artist's impression of Salter's Duck. *(View this art in color at www.dekker.com.)*

EFFECTIVE USE OF RENEWABLES

Grid Connected

There are difficulties in absorbing the variable and often unpredictable energy of renewable energy sources into the electrical system.[23] At the root of the problem is the fact that electrical networks were designed to convey power from large, centrally dispatched thermal and hydropower plants via the high-voltage supergrids to customers at lower voltages. The capacities of many renewable sources and their geographical locations means they are being connected at lower-voltage distribution levels. The design of the networks together with geographic remoteness from load centers creates a range of technical difficulties such as changes to power flows, voltage variations, and rising fault levels among others. A significant academic and industrial research and development effort is concerned with identifying and mitigating the impact of embedded or distributed generation on the electricity network.

A further issue relates to the intermittent nature of the resources, particularly wind, and the ability of the electrical system as a whole to respond to rapidly changing power input. Studies in the United Kingdom suggest that at penetrations of 10% there is virtually no impact on network operation.[24] At 20% and above there are issues of raised reserve in the form of partly loaded conventional generators as well as the standing capacity required for periods of calm and this has economic implications. Denmark currently has well in excess of 20% although this to some extent is due to their connections with the German and Nordic networks that provide much of the response required. The introduction of other renewables is anticipated to partly mitigate the issue of wind intermittency as they are generally more predictable and less variable. Further research is under way in the United Kingdom and elsewhere, although it should be noted that other renewables will take many years to make a significant contribution.[25]

Overall, the needs of current electricity systems to maintain power balances on a second-by-second basis and the current approach of trying to get renewable generators to mimic thermal generation creates a range of technical and economic problems for renewables. These range from strict technical requirements for generator capability to network planning and protection, which conspire to make the connection of renewables appear fairly expensive. This is compounded by electricity markets that operate around "firm" power and penalize intermittent sources. An alternative approach that does not try to "shoehorn" renewables into conventional generation in embracing a more decentralized network will serve to mitigate many of these problems.[26]

Alternative Energy Vectors

One of the means of tackling the intermittency issue is to provide a degree of energy storage that accepts fluctuating inputs while providing a firm electrical output. A range of storage technologies are in development including flywheels, superconducting coils, and compressed air and oil. While these offer storage solutions on relatively short time frames, the energy density of chemical fuels provides a longer-term storage opportunity with the added benefit of creating new energy delivery vectors.

The most commonly mentioned is hydrogen in the context of a new hydrogen economy to supplant the existing carbon-based economy. Iceland has set itself the target of becoming the first all-hydrogen economy and has one of the first hydrogen refueling stations. Recently, Norsk Hydro has developed one of the Norwegian islands into an all-hydrogen system.

While it is possible to create hydrogen (H_2) by reforming natural gas this will require the release of CO_2, which will be difficult to sequester if it is performed on a decentralized basis. However, the electricity from carbon-free nuclear and renewables can be used to electrolyze water (H_2O) forming oxygen (O_2) as a by-product:

$$H_2O + \text{electricity} \rightarrow H_2 + \frac{1}{2}O_2$$

The process is appealing as H_2 is excellent for fuel cells in both electricity generation and vehicles and overall energy conversion efficiencies are over 60%.[27] The downside is that large-scale storage and transport of H_2 is difficult although research is looking into how

best to store hydrogen, e.g., in carbon nanotubes and prototype H_2-fueled vehicles are in existence.

An alternative or intermediate approach to the development of a full-scale hydrogen economy would be to convert the hydrogen into hydrogen-rich liquid fuels such as methanol, ethanol, or even longer-chain hydrocarbons.[27] These methods are proven to pilot stage or further.[28] Each approach requires a readily available source of CO_2—in the flues of fossil fuel power stations. This would not only assist in achieving high levels of renewables through chemical buffering but would allow continued use of fossil fuel plant and recycle CO_2 as a hydrogen carrier fuel. However, it would still require the capture of CO_2 from flue gases.

Direct Use

Renewables can also be used directly to provide useful work in chemical processes and elsewhere without the need to integrate with the electrical network. One approach is to use the mechanical energy from renewables such as wind to drive heat pumps to abstract thermal energy from the ambient airstream (or other low-grade heat source such as the ocean). The resulting heat may be delivered at a useful temperature for drying, steam raising, or other thermal processes.[29] Processes such as desalination that currently require significant fossil fuel usage could be achieved using renewables such as wind and wave.[30,31]

ECONOMICS OF RENEWABLE ENERGY

Probably the most controversial aspects of renewable energy are their apparent cost relative to conventional forms of generation. It is commonly stated that fossil-fueled generation has a lower cost than renewables with, for example, CCGT costing 2–3 p/kW hr while wind costs 4–5 p/kW hr and other renewables more. Fig. 6 shows such a comparison as was carried out for the United Kingdom's Royal Academy of Engineering; these suggest that it would be more economic to invest in gas and nuclear rather than wind and renewables and that the increasing use of renewables will raise electricity prices.[32] There are several critical flaws underlying such interpretations and analyses.

First, the analyses consider the direct costs associated with generation and ignore the social and environmental cost associated with them. The exclusion of these ''external'' costs means that fossil fuels appear cheaper than their true cost. Inclusion of damage costs for SO_2, NO_x, and, importantly, CO_2 makes a large difference even if they might be difficult to predict. Estimates of the ''cost'' of CO_2 vary significantly but £30/tonne—at the lower end—is often used in comparisons such as that in Fig. 6, which is seen to raise coal generating costs by almost 1 p/kW hr.[32]

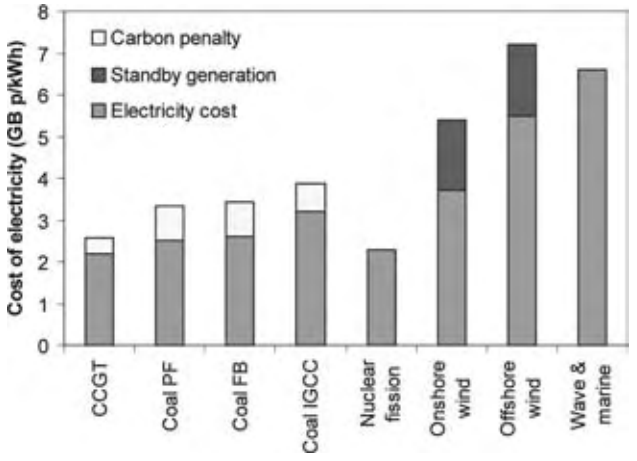


Fig. 6 Levelized electricity cost for generating technologies with/without carbon and standby generation for periods of no wind. (From Ref.[32].) *(View this art in color at www.dekker.com.)*

Second, the issue of subsidy for renewables is often used as an argument against their deployment and development and it should be removed to ''level the playing field.'' It should be remembered that fossil fuels and particularly nuclear power are mature technologies and have benefited from many decades of development and public subsidy. Although hydro is mature, wind is only just approaching maturity and newer renewables like wave and tidal stream are yet to receive significant development funding. As such, expecting these newer technologies to compete directly is not a reasonable or socially responsible approach; once developed subsidy will not be required.

Third, the standard approach of comparing technologies is inappropriate given the diversity of technologies and an increasingly market-orientated power sector. It is common practice to use the discounted cash flow methodology to calculate the levelized unit cost (p/kW hr) of a given technology. It is given by the ratio of the capital and recurring cost stream and the energy output stream, where both streams are discounted at the same rate. The International Energy Agency favors a 7% discount rate while it is common in the United Kingdom to use 8% and 15% to represent nominal state and private sector discount rates. Renewables are, in general, characterized by relatively high capital costs but compensated by low recurring operations and maintenance and zero fuel costs. The use of discounting in this manner means that the full renewable capital costs are seen while the often significant fossil fuel costs are reduced; this automatically biases the outcome in favor of the fossil technology. While nuclear also has large capital costs, discounting

Table 6 Standard and risk-adjusted leveled generation costs

Technology	Standard approach (U.S. cents/kW hr)	Risk-adjusted approach (U.S. cents/kW hr)	(%) Change
Geothermal energy	3.3	3.1	−6
Biomass	2.8	3.2	14
Wind	4	3.6	−10
Large hydro	5	4.7	−6
Nuclear	4	5.5	38
Pulverized coal	3.2	6.7	109
CCGT	3	7.2	140
Solar thermal	11.9	11.1	−6

(From Ref.[33].)

lowers the deferred decommissioning costs to a negligible level—again biasing the outcome.

Historically, the levelized cost methodology was used to compare financially similar alternatives, that of hydro with high capital cost and zero fuel costs and coal with similar capital costs and very low fuel costs. As fuel prices have risen and different technologies have become available the comparison has become less and less appropriate and is creating perverse outcomes. The problem stems from the fact that discount rate is, by definition, a reflection of risk. By applying the same discount rate to each generating technology it implies that the risks for each technology are identical. This is clearly not the case.

A more sophisticated approach is to use risk-adjusted discounted cash flow methods that properly account for the nondiversifiable risk associated with fossil fuel price volatility and the lower diversifiable risk associated with capital projects.[11] Application of this more robust financial analysis leads to a significant change in the costs of technologies. Table 6 suggests that the renewable technologies become slightly less costly while nuclear and fossil fuels, particularly gas, become significantly more expensive (in some cases the cost more than doubles). Furthermore, the cost rank order alters such that, with the exception of solar thermal, renewables become cheaper than the fossil fueled and nuclear options.

CONCLUSIONS

This entry outlines the arguments in favor of developing and deploying renewable energy technologies in terms of its environmental, sustainability, and security benefits. A range of technologies are briefly examined before the difficulties of integrating these variable and often intermittent generating sources are highlighted along with alternative uses for creating new energy vectors and in direct-use applications. Finally, the economics of renewables were explored with reference to the limitations in traditional comparisons with conventional sources.

REFERENCES

1. Thorogood, K.S.V. Electrical power engineering systems: an assessment of their impact on global climatic change. Eng. Sci. Educ. J. **1998**, *7* (3), 113–121.
2. Intergovernmental Panel on Climate Change. *Climate Change 2001: The Scientific Basis*; Cambridge University Press: Cambridge, 2001.
3. U.S. Department of State, 'New Asia-Pacific Partnership on Clean Development, Energy Security, and Climate Change', Press Release, July 27, 2005; http://www.state.gov/g/oes/rls/prsrl/50334.htm.
4. Royal Commission on Environmental Pollution. *Energy—The Changing Climate*; Stationary Office: London, 2000.
5. Bellhouse, G.M. Coal Purchase Analysis in the Electricity Supply Industry. PhD Thesis, University of Edinburgh, Edinburgh, 1997.
6. BP plc. *Statistical Review of World Energy 2004*; BP plc: London, 2004.
7. Sørensen, B. Renewable energy: a technical overview. Energy Policy **1991**, *19* (4), 386–391.
8. Greenpeace; European Wind Energy Association. *Wind Force 12: A Blueprint to Achieve 12% of the World's Electricity from Wind Power by 2020*; 2004.
9. Sauter, R.; Awerbuch, S. *Oil Price Volatility and Economic Activity: A Survey and Literature Review*; IEA Working Paper; IEA: Paris, 2002.
10. Final Report of G8 Renewable Energy Task Force. 2001; http://www.g7.utoronto.ca/energy/G8_report_2001.pdf (accessed Oct 2004).
11. Awerbuch, S.; Berger, M. *EU Energy Diversity and Security: Applying Portfolio Theory to Elec-*

tricity Planning and Policy-Making; International Energy Agency: Paris, 2003; http://library.iea.org/dbtw-pd/textbase/papers/2003/port.pdf.
12. Everett, B. Solar thermal energy. In *Renewable Energy: Power for a Sustainable Future*; Boyle, G., Ed.; Oxford University Press: Oxford, 1998; 41–88.
13. Infield, D.; Rowley, P. *Renewable Energy: Technology Considerations and Electricity Integration*; CREST, Loughborough University, U.K.; http://crestdl.lboro.ac.uk/courses/foundation/RSocChem%20Review%20article.doc (accessed Oct 2004).
14. Renewable Devices Ltd.; www.renewabledevices.com.
15. Jenkins, N.; Allan, R.N.; Crossley, P.; Kirschen, D.; Strbac, G. *Embedded Generation*; IEE Power and Energy Series 31; 2000.
16. Brown, G.; Garnish, J. Geothermal Energy. In *Renewable Energy: Power for a Sustainable Future*; Boyle, G., Ed.; Oxford University Press: Oxford, 1998; 341–382.
17. Ramage, J.; Scurlock, J. Biomass. In *Renewable Energy: Power for a Sustainable Future*; Boyle, G., Ed.; Oxford University Press: Oxford, 1998; 137–182.
18. Elliott, D. Tidal power. In *Renewable Energy: Power for a Sustainable Future*; Boyle, G., Ed.; Oxford University Press: Oxford, 1998; 226–314.
19. Garrad Hassan and Partners. In *Scotland's Renewable Resource*; Scottish Executive: Edinburgh, 2001.
20. Marine Current Turbines Ltd.; http://www.marineturbines.com/.
21. Duckers, L. Wave energy. In *Renewable Energy: Power for a Sustainable Future*; Boyle, G., Ed.; Oxford University Press: Oxford, 1998; 315–352.
22. Ocean Power Delivery Ltd. website at; http://www.oceanpd.com.
23. Harrison, G.P.; Kiprakis, A.E.; Wallace, A.R. Network integration of mini-hydro. Re-Gen. **2003**, *1* (1), 56–63.
24. ILEX Energy Consulting; Strbac, G. *Quantifying the System Costs of Additional Renewables*; Department of Trade & Industry: London, 2003.
25. Boehme, T.; Taylor, J.; Wallace, A.R.; Bialek, J.W.; Harrison, G.P. *Matching Renewables with Demand*; Institute for Energy Systems Working Paper; University of Edinburgh: Edinburgh, 2004.
26. Awerbuch, S. Restructuring electricity networks. Cogen. On-Site Power Prod. **2004**, *5* (4), 43–49.
27. Mignard, D.; Sahibzada, M.; Duthie, J.; Whittington, H.W. Methanol synthesis from flue-gas CO_2 and renewable electricity: a feasibility study. Int. J. Hydrogen Energy **2003**, *28*, 455–464.
28. Mignard, D.; Pritchard, C. Squaring the circle: sequestration of CO_2 as liquid fuel. In *Greenhouse Gas Control Technologies*; Vancouver, Sep 5–9, 2004.
29. Pritchard, C.; Low, R.E. A self-regulating heat pump to utilise wind and wave energy resources. Energy Sources **1990**, *12*, 12–24.
30. Infield, D.G.; Rahal, Z. Computer modelling of a large-scale stand-alone wind-powered desalination plant. In *Wind Energy Conversion from Theory to Practice*; Hunter, R., Ed.; Mechanical Engineering Publications Ltd.: Edinburgh, 1997, 129–134.
31. Crerar, A.C.; Low, R.E.; Pritchard, C. Wave powered desalination. Desalination **1987**, *67*, 127–137.
32. PB Power. *The Costs of Generating Electricity*; Royal Academy of Engineering: London, 2004.
33. Sharp, T. And the winner is ...? Int. Water Power Dam Constr. **2004**, *56* (3), 12–16.

Reprocessing of Domestic Spent Nuclear Fuel

R

Truman S. Storvick
Chemical Engineering Department, University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

The story of nuclear fuel reprocessing begins with the production of weapons grade plutonium. Natural uranium strategically placed in a large block of graphite gives sustained nuclear fission of uranium 235 (^{235}U) with some excess neutrons captured by uranium 238 (^{238}U) to produce plutonium 239 (^{239}Pu). The successful chemical process to isolate nearly pure weapons grade plutonium is the first production-scale nuclear fuel reprocessing. The military program to bridle the rate fission energy is released to power submarines led to domestic nuclear power plants serving electric utilities. It is the spent fuel from these power plants that is the object of domestic spent fuel reprocessing. There are two reasons to consider reprocessing spent nuclear fuel. The mass of radiotoxic material that accumulates using the current ''once through'' nuclear fuel cycle and the thousands of years this spent fuel remains hazardous can both be significantly reduced. Second, a major fraction of the energy content of the natural uranium remains when it is a declared ''spent fuel.'' These fuel values can be recovered and with proposed advanced nuclear reactor designs the mass and burial time for the remaining radioactive waste can be significantly reduced.

BEGINNINGS OF NUCLEAR FUEL PROCESSING

The remarkable story of the development of the nuclear power industry begins when Lise Meitner and Otto Hahn interpreted the capture of neutrons by a uranium atom nucleus (the ^{235}U isotope in natural uranium) followed by spontaneous splitting of the nucleus to form smaller atom nuclei. They called this process fission in their article published in *Nature* on February 6, 1939. The huge energy release by nuclear fission suggested that as each fission produces two or more neutrons, it should be possible to bring together a critical mass of fissionable material that would produce a chain reaction. The resulting explosion would be millions of times more powerful than the energy released by an equal mass of any known chemical explosive.

The curtain of secrecy dropped during World War II and the race to produce this nuclear explosive accelerated. Progress to separate nearly pure ^{235}U from natural uranium was under way, but proving to be very difficult. In January 1941, Seaborg and coworkers reported the discovery of the new element ^{239}Pu with a fission capture cross section greater than ^{235}U, suggesting that it would be an even better explosive.[1] A continuous source of neutrons would produce the ^{239}Pu by neutron capture of the ^{238}U in natural uranium. Chemical separation of ^{239}Pu from the uranium and fission products would be easier than separating the isotopes of natural uranium to produce weapons grade ^{235}U.

The emphasis on production of pure plutonium for weapons continued throughout the Cold War period. The successful control of the chain reaction to power submarines to cruise underwater for weeks rather than hours for the diesel–electric powered substitutes opened up a new option. The contractors that built the nuclear reactors and steam turbines for submarines also served the electric power utilities. Extension of this nuclear technology to domestic electric power production was obvious. President Eisenhower's Atoms for Peace initiative in 1954 opened the way for the nuclear power plants that followed that currently supply about 20% of the electric power to the U.S. power grid.

Plutonium Production by Precipitation

The initial laboratory experiments to isolate plutonium from irradiated uranium were based on precipitation. Plutonium was discovered and first produced by irradiation of uranyl nitrate hexahydrate with neutrons.[1] The Berkeley 60 in. cyclotron provided a 12 meV beam of deuterons directed onto a beryllium target producing neutrons by the (d,n) reaction. Cunningham and Werner showed that LaF_3 precipitate was an efficient plutonium carrier for Pu^{4+} but not Pu^{6+}.[2] This produced the first 2.8 μg pure plutonium metal. A 0.5 μg sample of ^{239}Pu was used to determine the fission cross section.[3] It was about 50% greater than ^{235}U making ^{239}Pu an attractive alternative to ^{235}U as nuclear bomb material.

Microscopic samples were used to determine the half-life of the alpha decay of ^{239}Pu to be $(2.43 \pm 0.12) \times 10^4$ yr compared to the current accepted value of 24,110 yr. The isotopic sequence to plutonium, ^{238}U + neutron (immediate release of gamma energy

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120022535

equal to the binding energy of the neutron), produced ^{239}U (half-life 23 min), which decays to ^{239}Np (half-life 2.4 days), yielding ^{239}Pu (with a half-life of 24,110 yr). The oxidation states of plutonium were determined using submicrogram quantities of plutonium and α-particle counting with samples too small for chemical analysis.

Fermi's Stagg Field demonstrated in 1942 that a properly configured graphite pile fueled with natural uranium produced a sustained fission reaction necessary to produce plutonium by neutron capture by ^{238}U.[4] This nuclear reactor demonstration promoted plutonium production to the fast track. The decision to build the Hanford plutonium production facility was made using an unprecedented engineering scale factor of about 10^9 going from micrograms to kilograms of Pu and reactor thermal energy release from watts to megawatts.

The design of the production-scale reactor began before the chemical isolation of plutonium was known. Eugene Wigner, an accomplished theoretical physicist who studied engineering in Europe, led the group that designed the nuclear pile, which became the Hanford reactor.[5] It was Wigner's computational procedures that were used to fix the dimensions of the graphite moderator, the distribution of the uranium metal fuel elements, and selected water to cool the reactor. Crawford Greenewalt, the leading engineer for the Du Pont Company in charge of plutonium production, played an important role in this project when everything was new and untried with enormous pressure to speed the project to completion.

Each of the Hanford reactors was a 28 (diameter) by 36 ft long pure graphite cylinder on its side, over 1200 tons of graphite. More than 2000 holes passed though the faces of the graphite and these were lined with aluminum tubes. About 250 tons of cylindrical uranium slugs encapsulated in aluminum were placed in the tubes in the graphite pile with control rods strategically located among the uranium fuel load. The aluminum tubes were also cooling water channels that removed approximately 250,000 kW of thermal energy. The nuclear pile was run for 100 days, a time that transmuted about one out of 4000 ^{238}U atoms to plutonium. The fuel slugs were pushed out of the graphite pile by fresh fuel and dropped into a pool of water that provided shielding from the intense radiation. After 60 days, the irradiated slugs were moved to chemical separation.

The laboratory methods for recovering and purifying plutonium were not suited for large-scale production. The ether extraction used on the cyclotron-produced plutonium is a safety hazard. The lanthanum fluoride precipitation step yielded a gelatinous product that is difficult to process and the aqueous fluorides produced serious equipment corrosion. Other processes were investigated, but precipitation was selected as the best alternative for quick development when early in 1943, S.G. Thompson showed that $BiPO_4$ precipitate would strongly carry Pu^{4+}.[6] The $BiPO_4$ precipitate is crystalline and easily collected by filtration or with a centrifuge. The crystals readily dissolve in strong nitric acid. This process was selected to recover plutonium from the irradiated uranium fuel from the Hanford graphite pile reactors.[7]

The aluminum cladding was removed and the fuel elements dissolved in nitric acid and treated with $NaNO_2$ to reduce the plutonium to Pu^{4+}. Sulfuric acid was added to complex the uranyl nitrate to prevent it from precipitating with the plutonium. The $BiPO_4$ precipitate carried essentially all the plutonium while most of the fission products and the U^{6+} remained in solution in this extraction step. The centrifuged precipitate was dissolved in nitric acid and $NaBiO_3$ added as an oxidant and a source of bismuth. The oxidized plutonium, Pu^{6+}, remains soluble and the fission products carried by the $BiPO_4$ precipitate were removed.

The sequence: plutonium reduction, $BiPO_4$ plutonium precipitation, oxidation of the plutonium to dissolve the precipitate, and precipitate by-products was repeated three times.[8] The final by-product precipitation with $BiPO_4$ was followed by a by-product precipitation with LaF_3 to remove traces of fission products and uranium. The Pu^{6+} in solution was reduced and precipitated with LaF_3 for the final Pu^{4+} precipitation. This precipitate was treated with KOH to give a mixture soluble in HNO_3. Two precipitation steps with H_2O_2 separated the plutonium and lanthanum with the product plutonium dissolved in nitric acid. The overall decontamination factor for this batch process was 10^7 with 97–98% plutonium recovery. This solution was processed to obtain PuO_2, which was reduced to metal and further refined to weapons grade plutonium.

CONTINUOUS PROCESS FOR PLUTONIUM PRODUCTION

Early consideration was given to developing a continuous process to produce weapons grade plutonium. Countercurrent extraction was an advanced chemical engineering technology but the equipment is complex and remote operation and maintenance required by radiation shielding made the first choice batch precipitation process. Following World War II and moving into the Cold War period, plutonium production remained a high priority and the Purex (plutonium uranium reduction extraction) was developed for processing irradiated uranium fuel.[9,10]

The Purex process feed was the uranium slugs irradiated from the graphite pile reactors. The aluminum

cladding on the irradiated uranium slugs was removed and the uranium dissolved in nitric acid. The solution contains uranium as uranoyl nitrate, $UO_2(NO_3)_2$, and plutonyl nitrates with the plutonium in both the tetravalent and the hexavalent states. Most of the fission products go into solution, but solid fines form when solubility limits are exceeded and were removed.

Design of nuclear fuel processing equipment must avoid any possibility of fissionable material producing a chain reaction.[11] The concentration and purity of fissile material contained in solution, ^{235}U and ^{239}Pu in particular; location of equipment materials that serve as neutron reflectors, moderators, and poisons; equipment shape; and volume are the key equipment design parameters. The plant must be designed to provide radiation shielding and prevent release of radioactive material to the surroundings. Equipment design and an operator culture that always follows approved process procedures are required to ensure safe reprocessing plant operation.

The dissolved metals and the extractor upstream wash waters are combined and analyzed. The uranium and plutonium concentrations are used to adjust the pH of the solution to optimize the separation from the fission products. Solutions about 1 M in uranium and plutonium and 2–3 M HNO_3 were the feed for the extraction.

Tributyl phosphate (TBP) is the most versatile agent for the extraction of uranium and plutonium from fission products and also for separation of plutonium from uranium. A 30% solution of TBP in normal hydrocarbons is the usual organic phase in the Purex process. The U and Pu are extracted by TBP as unionized complexes, free of hydration water, and solvated by a definite number of TBP molecules. The complex-forming reactions proceed according to the following:[12]

$$UO_2^{2+}(aq) + 2NO_3^-(aq) + 2TBP(org) \rightarrow UO_2(NO_3)_2 \cdot 2TBP(org)$$

$$PuO_2^{2+}(aq) + 2NO_3^-(aq) + 2TBP(org) \rightarrow PuO_2(NO_3)_2 \cdot 2TBP(org)$$

$$Pu^{4+}(aq) + 4NO_3^-(aq) + 2TBP(org) \rightarrow PuO_2(NO_3)_4 \cdot 2TBP(org)$$

The initial countercurrent extraction stages transfer the U and Pu to the organic or TBP phase with the fission products and the higher actinides remaining in the aqueous phase. Beyond the feed point, the organic phase continues to extraction stages where it is scrubbed with acid to remove traces of the fission products. Americium and curium in the +3 state stay in the aqueous phase. Neptunium is partially in the +6 (extractable) and the +5 state, so it splits between the organic and the aqueous phases. Careful control of the solution pH and metal ion concentration maintains the plutonium in the +4 state leading to essentially complete extraction of the plutonium and uranium values into the organic phase.

Solvent degradation by radiation damage is a problem during the initial separation, so the temperature is kept low and centrifuge extraction units minimize the contact time between the aqueous and organic phases.[9] The aqueous raffinate containing the fission products is processed to reduce the water content and recover nitric acid values. Volatile radioactive fission products released during extraction must be contained for disposal. This fission product stream accumulated during military production of weapons grade plutonium and the selection of a final form for permanent disposal is still pending. Weapons grade plutonium production is the source of most of the legacy wastes.

Separation of plutonium from the uranium in the TBP phase proceeds by countercurrent stripping of the plutonium into dilute nitric acid, which reduces the Pu to the Pu^{3+} soluble species. The uranium remains in the organic phase. This countercurrent stripping section is followed by a scrubbing section that completes the removal of the plutonium from the uranium held in the TBP phase.

Plutonium purification proceeds by reducing the aqueous phase pH that oxidizes the plutonium to Pu^{4+}, which then extracts into the TBP phase. Impurities stay in the aqueous phase. The TBP phase stripping/extraction cycle is repeated to complete the plutonium purification. The uranium is purified using the same TBP/nitric acid extraction/stripping cycle. Careful control of the each element's oxidation state in the extraction cascade produces the plant-scale separations of uranium from plutonium of 10^6. Fission product decontamination factor was 10^8. The plutonium and uranium recovery is about 99.9% with 95% of the nitric acid values and 99.7% of the organic solvent recycled.[12]

An advantage of the Purex process is the low salt content of the aqueous waste stream so the liquid volume can be reduced by evaporation. The HNO_3 values are recycled to the process. The Purex process produces nearly pure plutonium and recovers uranium and it is the process that has been adapted to treat domestic spent nuclear fuel.

CHARACTERIZATION OF DOMESTIC NUCLEAR FUEL

There are two major sources of radioactive materials in the United States that must be isolated to protect the public. The spent nuclear fuel accumulating at each of more than 100 power plant sites is the largest source.

Second is the Department of Defense legacy wastes, residues from weapons production, and the experimental nuclear reactor experiments to produce military nuclear power systems. These wastes are a heterogenous mix of solids and liquids, each requiring special handling and processing designed to reduce the total mass and isolate the radioisotopes for safe storage. Military legacy wastes require special treatment.

Nuclear power plants in the United States use light water moderated nuclear reactors (LWR) that produce the steam to generate electricity. The fuel elements for boiling water reactors and pressurized water reactors (PWR) are nearly the same. The fuel is uranium dioxide enriched with 3–4% ^{235}U and this produces a nearly uniform spent fuel, which would be the feed for domestic fuel reprocessing.

Ceramic UO_2 fuel pellets are machined to uniform size cylinders [1.26 cm (diameters) $\times$ 1.26 cm (length)], placed in zircaloy tubes about 4 m long, pressurized with helium, and sealed with free space at the top and bottom to collect fission product gases.[13] The tubes are set in fuel bundles with 15 $\times$ 15 or 17 $\times$ 17 arrays supported by end plates and intermediate spacers to secure each tube in a fixed position. The PWR fuel bundles provide flow channels for cooling water and the stable geometry ensures the neutron path length and moderates the neutrons to the thermal energy range necessary to sustain fission. Some of the grid positions are used for instrumentation and control rods. Each fuel assembly weighs about 658 kg and contains 523 kg of UO_2 (461 kg of U metal).[14] The total fuel loading varies with the reactor design, so for a PWR the core loading may be between 80 and 120 metric tons of the fuel. A fuel bundle stays in the reactor between 3 and 4 yr and the spent fuel is stored in water pools located on the power plant site. This is the current once through fuel cycle for all U.S. reactors.

There are more than 45,000 MTIHM [metric tons (tons) of initial heavy metal—actinides—elements with atomic number greater than 89] of spent nuclear fuel in storage and it is accumulating at the rate of about 2000 tons/yr. Without reactor operating license extensions for the reactors in service, the nuclear power production in the United States will end in about 2030 with the spent fuel inventory approaching 87,000 MTIHM. The current mandated limit for the Yucca Mountain repository is 63,000 MTIHM. Additional spent fuel storage will be required or there must be fuel reprocessing to meet the Yucca Mountain storage limit.

Irradiated uranium fuel elements contain fission product isotopes, activation products produced by atoms exposed to the intense neutron, α, β, and γ radiation in the reactor core, and actinides, produced by neutron capture of the nuclei of atoms with atomic number in the actinide series. The spent fuel assemblies are stored in water pools for γ-radiation shielding and the water is the heat sink for the thermal energy release from radioactive decay. Boron ions added to the water absorb decay neutrons.

Three characteristics of irradiated LWR low-enriched uranium spent fuel that are significant in fuel reprocessing are the total mass of the fission products and the transuranic elements, the thermal energy released by spontaneous radioactive decay that requires cooling, and the neutron and γ radiation released that requires biological shielding. The irradiation time and intensity when the fuel is in the reactor determine the isotopic composition of the spent fuel. There are more than 350 nuclides in the fuel immediately following reactor shutdown.[15] Many of these have very short half-lives that quickly follow a spontaneous decay sequence to more stable nuclides.[16] The total mass of all the isotopes remains essentially constant, but the rate of thermal energy release and the radiation intensity decrease as the spent fuel ages. Table 1 lists the major radioactive isotopes that require transmutation or long-term storage.

DOMESTIC FUEL REPROCESSING

The Purex process is the only one with commercial-scale operating history. This military technology has been modified to treat domestic spent nuclear fuel. The focus of international commercial efforts to reprocess domestic spent fuel is to recover uranium and plutonium in the spent fuel.

Table 1 Major components in LWR of spent nuclear fuel

Major components
955.4 kg U
8.5 kg Pu (5.1 kg ^{239}Pu)
Minor actinides
0.5 kg ^{237}Np
0.6 kg Am
0.02 kg Cm
Long-lived fission products
0.2 kg ^{129}I
0.8 kg ^{99}Tc
0.7 kg ^{93}Zr
0.3 kg ^{135}Cs
Short-lived isotopes
1.0 kg ^{137}Cs
0.7 kg ^{90}Sr
Stable isotopes
10.1 kg lanthanides
21.8 kg other stable isotopes

Basis, 1 metric ton, irradiation, 33,000 MWD/MT, 10 yr cooling. (From Ref.[17].)

The extended radiation time for the domestic fuel increases the quantity of fission products and the higher actinides. Pure plutonium product poses nuclear weapons proliferation risk and is the primary reason reprocessing is not practiced in the United States. The modified PUREX process has been practiced on an industrial scale in Europe and supports the production of mixed uranium–plutonium fuel. Blended UO_2 and PuO_2 powder is compacted and sinter to form the mixed oxide (MOX) fuel pellets much like the enriched UO_2 fuel. Natural and depleted uranium can be used to prepare MOX fuel and is the demonstrated option to recover fuel values from spent fuel.

Decladding the fuel pellets is the first step in domestic fuel reprocessing.[18,19] A practical method is to chop the fuel bundles into short lengths, 2.5–5.0 cm long, so that the metal oxide fuel pellets can be dissolved in strong nitric acid. The helium, which is in each fuel pin, and about 10% of the fission product gases are released during this step. The chopped material can be oxidized in a kiln to produce U_3O_8 from the UO_2. This increases the solid volume that pulverizes the ceramic fuel and releases occluded gases, and the tritium oxidizes to tritium water. Off-gas treatment must be in place to retain dust and collect the iodine, krypton, xenon, and tritium water that are released.

The second step is to dissolve the metal oxide fuel using strong nitric acid. The object is to bring all the fission products, uranium, and transuranics, into solution to feed the extraction process. Some of the fission products exceed solubility limits and the fine solids formed must be removed before extraction. Provisions to recover nitrogen oxides and collect gaseous fission products released during this step must be in place. The stainless steel and zircaloy fuel jackets from the fuel assemblies do not dissolve and are separated from the solution, washed, checked for radioactivity, and packaged for disposal as low-level radioactive waste.

The classic Purex process produces pure uranium and plutonium from the feed solution. The Purex waste stream contains the fission products and the other actinides totaling about 4% of the heavy metal content of the spent fuel. Reagents that form solids are avoided because they increase the mass of the waste stream to long-term storage. The actinides have very long half-lives that increase the solid waste storage time. Fission products that produce thermal energy complicate the waste storage problem. The high-energy γ radiation always requires shielding for handling, transporting, and storing the solid wastes. Technical, social, and political solutions to the spent fuel-storage problem remain even if there is no long-term future for nuclear energy. Plans to expand use of nuclear energy must also address the accumulated spent fuel inventory as new reactor fuel cycles evolve.

NUCLEAR ENERGY FUTURE

Sustainability is the keyword in discussions of the energy future to meet increasing global demands for thermal energy and electricity. There is a general international agreement that nuclear energy will play a role in supplying future energy demand. Nuclear power is the only technology now available that can produce economic, base load quantities of electricity without emitting pollutants associated with global climate change. About one-fifth of the U.S. electric power is now produced economically, safely, and reliably by nuclear power plants. The next 50 yr will see the end of the operating licenses of these reactors and that generating capacity must be replaced. Finding an acceptable and safe way to dispose of the spent nuclear fuel that remains radiotoxic for several hundred thousand years represents a serious open technical issue that must be addressed.

The National Energy Policy of 2000 includes initiatives that recommend deployment of additional nuclear power plants. This policy calls for expanded international cooperation to develop advanced technologies to improve reactor thermal efficiencies, recover fuel values remaining in spent nuclear fuel, and improve processing of fission products to reduced isolation times for repository disposal. The Advanced Fuel Cycle Initiative under the Department of Energy has been approved.[20] The task includes optimizing the cost and performance of the Yucca Mountain spent fuel repository, providing nuclear weapons proliferation resistant recovery of energy contained in spent fuel, and supporting the development of Generation IV nuclear reactor systems. Reduction in the volume of radioactive waste for burial and new reactor designs to fission or transmute the long-lived radioisotopes are the major objectives of this effort. Separation processes that isolate long-lived isotopes are the key to spent fuel management. Combining separation with new generation reactors that produce energy as they fission or transmute the long-lived isotopes is an important long-term goal. The World Nuclear Association web site provides current information on international nuclear research, technology, economics, and power plants under construction.

Proposed Generation IV Reactors

The international group has identified six Generation IV reactor systems for development.[21] All of these reactors should be ready for deployment by 2030. The fast neutron spectrum reactors can use the fuel values of all of the fissile and fertile transuranic isotopes in reprocessed fuel. This does not occur in the current thermal spectrum reactors. Producing energy

by fission of the transuranics rather than treating them as waste significantly reduces the long-term waste storage problem.

There are three fast-flux reactors proposed for development: the sodium cooled, the gas cooled, and the lead cooled. The fission cross sections for fast neutrons (high-energy spectrum neutrons) for all of the fissile actinides are nearly the same so the fast-flux reactors use all of the fissile actinides as fuel. The fast-flux isotopic fission cross sections are smaller than for thermal neutrons so the fraction of fissile isotopes (e.g., ^{235}U, ^{239}Pu, ^{241}Pu) in the fuel must be in the range of 10–20%.[22] The fertile isotopes (e.g., ^{238}U, ^{240}Pu, ^{242}Pu) have fast-flux neutron capture cross sections that are also high so neutron capture produces additional fissile material. Using these long-lived isotopes as fuel removes them from the waste stream. Converting fertile isotopes to fissile adds to the energy produced and is the practical way to use all of the natural uranium as fuel.

Liquid metal cooled reactor development began during the mid-1940s. Early experimental reactors led to the Experimental Breeder Reactor-II (EBR-II) designed to include complete on-site fuel reprocessing, fuel fabrication, and electricity generation. Experimental Breeder Reactor-II went critical in 1963 with metallic alloy fuel pins in stainless steel jackets cooled by liquid sodium circulated through a heat exchanger to produce steam.[23] During the 30 yr the reactor was in service, it was the test bed for new fuel assemblies for advanced reactor systems. The inherent safety of the EBR-II was demonstrated with the reactor coasting to shut down without operator intervention with safety systems off or with a total loss of coolant. The fuel cycle facility processed over 30,000 fuel elements and produced more than 300 subassemblies. New fuel irradiation tests were part of the experimental program. Initially designed as a fuel breeder reactor, it was successfully converted to a plutonium burner reactor. The EBR-II program produced much of the technical data that inform the fast-flux Generation IV reactor designs.

The thermal and nuclear properties of sodium (it scatters neutrons without absorbing them) made it the heat exchange fluid of choice for fast-flux reactors in spite of its nasty chemical properties when exposed to air or water. The French Superphenix, a commercial-scale sodium cooled reactor, was beset with technical problems, but demonstrated that fast-flux reactors can produce electric power at the 1000 MW level.

Liquid lead cooled reactors have been demonstrated by the Russian nuclear community. Lead has a higher melting point than sodium, an advantage for a reactor to operate at high temperature. At these higher temperatures, lead is corrosive to the steel structures and other metal components in the reactor core. Material selection and development are required for full deployment of this technology.

Helium cooling was seriously considered in 1943 for the Hanford plutonium production piles.[5] Helium cooled fast-flux reactors may be the first Generation IV reactors deployed.[20] The fuel forms are designed to operate at temperatures greater than 850 °C and they must contain the fission products. Candidates include composite ceramic fuel, advanced fuel particles, or ceramic clad elements of actinide compounds. The reactor core configuration might be based on prismatic blocks, pins, plates, or spheres. Electricity would be generated using a direct cycle gas turbine and compressor in a Brayton cycle. High-temperature thermal energy could be used for chemical processing such as production of hydrogen.

The fast-spectrum reactors with full recycle of actinides would be designed with on-site spent fuel reprocessing and fuel fabrication to minimize the on-site inventory of long-lived radioactive waste. Modern robotic equipment can be used for reactor refueling and for fuel reprocessing. The spent fuel reprocessing and fuel fabrication facilities must be developed to close the nuclear fuel cycle and use all the energy available in natural uranium.

Reprocessing Advanced Spent Fuel

The classical Purex process was designed to produce nearly pure uranium and plutonium. The Chemical Engineering Division of Argonne National Laboratory has demonstrated UREX+, an advanced aqueous process with five extraction trains that split commercial reactor spent fuel into five streams: 1) a nearly pure uranium stream (95.5% of the heavy metal in the spent fuel); 2) technetium sent to transmutation (0.08%); 3) Pu/Np converted to MOX fuel for LWR fuel and Am/Cm for transmutation or fast-flux reactor fuel (0.962%); 4) Cs/Sb decay heat producers sent to interim decay storage (0.017%); and 5) a mixed fission product stream (3.44%) composed of gases and solids incorporated into a waste form for geological repository disposal.[24] The percentages shown are computed from Table 1.

Developments during the EBR-II program included dry (nonaqueous) electrochemical separation of uranium (relatively high-enriched ^{235}U) from the spent fuel. These metallic fuel elements were chopped into an anode basket, submerged in molten salt (LiCl/KCl eutectic) and the uranium electrochemically deposited on an inert metal cathode. The stainless steel fuel casings, the salt containing the higher actinides, and fission products were immobilized (glassified) for disposal.

Recent advances on pyrometallurgical technology have shown that metallic reactor fuel can be processed

using reductive extraction between the salt phase (adding Li to reduce the metal chloride) and collecting the metals as an alloy of cadmium or bismuth.[25] The liquid metal alloy becomes the anode in an electrochemical cell collecting the mixed actinides metals on an inert cathode. Remote processing of highly radioactive, ''young'' spent fuel is possible because there is no radiation damage to the inorganic solvents.

Development of a one-step electrochemical process to reduce LWR spent fuel (UO_2) to metallic form is under development at Argonne National Laboratory.[26] The transuranic metals are separated by pyroprocessing technology. This reduces waste that requires repository disposal. The transuranic metals could be cast into fuel suitable for the Generation IV reactors or transmuted reducing them to fission products.

Economic factors will certainly impact the future of nuclear power. Members of the University of Chicago Department of Economics and the Harris School of Public Policy participated in a detailed study of the economic factors affecting the future of nuclear power.[27] In the long term nuclear power can be competitive with other sources given initial financial incentives.

CONCLUSIONS

In the 60 yr since discovery, plutonium is the most studied element and it will be the key to unlocking uranium as an energy source into the future.[28] It is only 50 yr since nuclear power was introduced. Commercial power plants use 1970s reactor technology. It is anticipated that the technical problems of nuclear fuel reprocessing and Generation IV nuclear reactor deployment can be solved. The Department of Energy Nuclear Power 2010 program represents a government–industry study designed to map strategy for the future of nuclear power.

The University of Chicago study has shown that advanced nuclear power production can be economically competitive with fossil fuel plants. The second 50 yr of nuclear power can be a period when research provides answers to safe nuclear power production and safe disposal of the nuclear wastes. The remarkably rapid development of nuclear science and technology in the 1940s can now proceed at a slow, deliberate pace. The technical research must be followed by government–industry cost-shared reactor deployment program. The initial economic barrier of capital funds for new facilities plus the extra engineering and financing costs that come with building and operating the expensive first-of-a-kind new nuclear plants must be shared. The national and international economics, culture, and politics will shape the path to future energy sources. Remember to train the next generation of nuclear scientists and technologists that can make it happen.

ACKNOWLEDGMENT

The hospitality and encouragement of the Chemical Engineering Department faculty made this work possible.

R

REFERENCES

1. Roland, L.K., Jerry, B.G., Gary, T., Eds.; *The Plutonium Story: Journals of Professor Glenn T Seaborg, 1939–1946*; Benefiel, Battelle Press: Columbus, OH, 1994.
2. Cunningham, B.B.; Werner, L.B. The first separation of the synthetic element $_{94}PU^{239}$. In *The Transuranium Elements, Research Papers*; Seaborg, G.T., Katz, J.J, Manningm, W.M., Eds.; McGraw-Hill Book Company: New York, 1949; 51–78.
3. Kennedy, J.; Seaborg, G.; Segre, E.; Wahl, A. Properties of 94(239). Phys. Rev. **1946**, *70*, 555.
4. Glasstone, S. The utilization of nuclear energy. In *Sourcebook on Atomic Energy*, 3rd Ed.; D. Van Nostrand Company, Inc.: Princeton, NJ, 1967; 520–522.
5. Rhodes, R. Different animals. In *The Making of the Atomic Bomb*; Simon and Schuster: New York, 1986; 497–499.
6. Thompson, S.G.; Seaborg, G.L. The first use of bismuth phosphate for separating plutonium from uranium and fission products. Prog. Nucl. Energy Ser. 3 **1956**, *1*, 163.
7. Wymer, R.G.; Vondra, B.L. Chemical aspects of LWR fuel reprocessing. In *Light Water Reactor Nuclear Fuel Cycle*; CRC Press, Inc.: Boca Raton, FL, 1981; 77–78.
8. Stevenson, R.L.; Smith, P.E. Aqueous separations. In *Reactor Handbook*; Stoller, S.M., Richards, R.B., Eds.; Interscience Publishers, Inc.: New York, 1961; Vol. 2, 227–228.
9. Wymer, R.G.; Vondra, B.L. PUREX solvent extraction chemistry. In *Light Water Reactor Nuclear Fuel Cycle*; CRC Press, Inc.: Boca Raton, FL, 1981; 103–162.
10. Benedict, M; Pigford, T.H.; Levi, H.W. The PUREX process. In *Nuclear Chemical Engineering*, 2nd Ed.; McGraw-Hill: New York, 1981; 466–514.
11. Benedict, M; Pigford, T.H.; Levi, H.W. Prevention of criticality in processing plants. In *Nuclear Chemical Engineering*, 2nd Ed.; McGraw-Hill: New York, 1981; 547–455.
12. Katz, J.J.; Seaborg, G.L.; Morss, L.R. PUREX process. In *The Chemistry of Actinide Elements*, 2nd Ed.; Chapman and Hall: New York, 1986; Vol. 1, 525–534.

13. Cochran, R.; Tsoulfanidis, N. Fuel design and fabrication. In *The Nuclear Fuel Cycle: Analysis and Management*; American Nuclear Society: LaGrange Park, IL, 1999; 77–104.
14. Wymer, R.G.; Vondra, B.L. Table 1 Physical characteristics of a reference PWR fuel assembly. In *Light Water Reactor Nuclear Fuel Cycle*; CRC Press, Inc.: Boca Raton, FL, 1981; 65 pp.
15. Cochran, R.; Tsoulfanidis, N. Fuel design and fabrication. In *The Nuclear Fuel Cycle: Analysis and Management*; American Nuclear Society: LaGrange Park, IL, 1999; 290–292.
16. Wymer, R.G.; Vondra, B.L. Historical development of the chemistry of aqueous reprocessing methods. In *Light Water Reactor Nuclear Fuel Cycle*; CRC Press, Inc.: Boca Raton, FL, 1981; 64–72.
17. Bennett, R.G. Alternative to direct disposal: advanced nuclear fuel cycles. In 2003 NAE National Meeting Symposium in Honor of Foreign Secretary Harold K. Forsen, Feb 6, 2003; slide 9.
18. Benedict, M.; Pigford, T.H.; Levi, H.W. The Purex process. In *Nuclear Chemical Engineering*, 2nd Ed.; McGraw-Hill: New York, 1981; 466–514.
19. Glasstone, S.G.; Sesonske, A. The reprocessing option. In *Nuclear Reactor Engineering*, 4th Ed.; Chapman & Hall: New York, 1994; 648–659.
20. *The Future Path of Advanced Spent Fuel Treatment and Transmutation Research*, Report to Congress on Advanced Fuel Cycle Initiative; U.S. Department of Energy Office of Nuclear Energy, Science, and Technology, Jan 2003.
21. *The Future Path of Advanced Spent Fuel Treatment and Transmutation Research*, Appendix B, Report to Congress on Advanced Fuel Cycle Initiative; Oct 2002.
22. Waltar, A.E.; Reynolds, A.B. Selection of core materials and parameters. In *Fast Breeder Reactors*; Pergamon Press: New York, 1981; 44–48.
23. EBR-II; http://www.anlw.anl.gov/anlw_history/reactors/ebr_ii.html.
24. Vandegrift, G.; Regalbuto, M.; Aase, S.; Bakel, A.; Battisti, T.; Bowers, D.; Byrnes, J.; Clark, M.; Emery, J.; Falkenberg, J.; Gelis, A.; Pereira, C.; Hafenrichter, L.; Tsai, Y.; Quigley, K.; Vander Pol, M. Designing and demonstration of the UREX+ process using spent nuclear fuel. In ATALANTE 2004, Advances in Future Nuclear Fuel Cycles, Jun 21–24, 2004, Nimes, France.
25. Uozumi, K.; Kinoshita, K.; Inoue, T.; Fusselman, S.; Grimmett, D.; Roy, J; Storvick, T.; Krueger, C.; Nabelek, C. Pyrometallurgical partitioning of uranium and transuranic elements from rare earth elements by electrorefining and reductive extraction. J. Nucl. Sci. Techol. **2001**, *38* (1), 36–44.
26. Argonne National Laboratory, Chemical Engineering Division; http://www.cmt.anl.gov/science-technology/nuclear/default.shtml.
27. The Economic Future of Nuclear Power; University of Chicago, Aug 2004.
28. Hoffman, D.H. *Advances in Plutonium Chemistry 1967–2000*; American Nuclear Society: La Grange, IL, 2002.

Resid Conversion

James G. Speight
CD & W Inc., Laramie, Wyoming, U.S.A.

INTRODUCTION

A residuum (pl. residua, also shortened to resid, pl. resids) is the residue obtained from petroleum after nondestructive distillation has removed all the volatile materials. The temperature of the distillation is usually maintained below 350°C (660°F) because the rate of thermal decomposition of petroleum constituents is minimal below this temperature but the rate of thermal decomposition of petroleum constituents is substantial above 350°C (660°F). If the temperature of the distillation unit rises above 350°C (660°F), as happens in certain units where temperatures up to 395°C (740°F) are known to occur, cracking can be controlled by adjustment of the residence time. This entry introduces some of the basic chemistry behind the synthesis and conversion processes.

DEFINITION

Residua are black, viscous materials and are obtained by the distillation of a crude oil under atmospheric pressure (atmospheric residuum) or under reduced pressure (vacuum residuum, 25–100 mm Hg). They may be liquid at room temperature (generally atmospheric residua) or almost solid (generally vacuum residua) depending on the nature of the crude oil. Resids have high molecular weights (usually in excess of 800 number average molecular weight for vacuum resids) and they contain polynuclear aromatic compounds as well as heteroatom compounds of nitrogen, oxygen, and sulfur. Metal compounds also occur in residua.

When a residuum is obtained from crude oil and thermal decomposition has commenced, it is more usual to refer to this product as pitch. The differences between conventional petroleum and the related residua are due to the relative amounts of various constituents present, which are removed or remain by virtue of their relative volatility.

The chemical composition of a residuum is complex. Physical methods of fractionation usually indicate high proportions of asphaltenes and resins, even in amounts up to 50% (or higher) of the residuum. In addition, the presence of ash-forming metallic constituents, including such organometallic compounds as those of vanadium and nickel, is also a distinguishing feature of residua and the heavier oils. Furthermore, the deeper the cut into the crude oil, the greater the concentration of sulfur and metals in the residuum and the greater the deterioration in physical properties.

It is the polynuclear aromatic system and the heteroatom species that pose the greatest limitation to resid conversion. The polynuclear aromatic and heteroatom species: 1) have high thermal stability; 2) poison catalysts by deposition of heteroatoms and metals; and 3) readily form coke that deposits on the catalyst and detracts from liquid production.

Conversion Chemistry

Understanding the chemical transformations of the macromolecules that occur in resids during conversion is greatly limited by the wide diversity (over a million chemical structures) of these complex macromolecules.

Therefore, in a mixture as complex as petroleum, the reaction processes can only be generalized because of the difficulties in analyzing not only the products but also the feedstock as well as the intricate and complex nature of the molecules that make up the feedstock. The formation of coke from the higher molecular weight and polar constituents of a given feedstock is detrimental to process efficiency and to catalyst performance. One method by which the process chemistry can be rationalized is to separate the resid and its conversion products into fractions using solubility/insolubility in volatile liquids as well as adsorption/desorption on solids. In this way a number of resids and resid conversion products were separated into coke (toluene insoluble), asphaltenes (toluene soluble/*n*-heptane insoluble), resins (*n*-heptane soluble, adsorbs on alumina), aromatics (*n*-heptane soluble, does not adsorb on alumina), and saturates (*n*-heptane soluble, does not adsorb on alumina).

The asphaltene constituents produce the highest amount of coke (relative to the other fractions of the resid). The formation of a coke-like substance during resid upgrading is dependent on several factors: 1) the degree of polynuclear condensation in the feedstock; 2) the average number of alkyl groups on the polynuclear aromatic systems; and 3) the hydrogen-to-carbon atomic ratio of the pentane-insoluble/heptane-soluble fraction.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120022536

Nitrogen species also appear to contribute to the pattern of the thermolysis. For example, the hydrogen or carbon–carbon bonds adjacent to ring nitrogen undergo thermolysis quite readily, as if promoted by the presence of the nitrogen atom. If it can be assumed that heterocyclic nitrogen plays a similar role in the thermolysis of asphaltenes, the initial reactions, therefore, involve thermolysis of aromatic-alkyl bonds that are enhanced by the presence of heterocyclic nitrogen. An ensuing series of secondary reactions, such as aromatization of naphthenic species and condensation of the aromatic ring systems, then lead to the production of coke. Thus, the initial step in the formation of coke from asphaltenes is the formation of volatile hydrocarbon fragments and nonvolatile heteroatom-containing systems.

The prevalent thinking is that the asphaltene nuclear fragments become progressively more polar as the paraffinic fragments are stripped from the ring systems by scission of the bonds (preferentially) between the carbon atoms alpha and beta to the aromatic rings. The polynuclear aromatic systems that have been denuded of the attendant hydrocarbon moieties are somewhat less soluble in the surrounding hydrocarbon medium than their parent systems. Two factors are operative in determining the solubility of the polynuclear aromatic systems in the liquid product. The alkyl moieties that have a solubilizing effect have been removed and there is also enrichment of the liquid medium in paraffinic constituents. There is an analogy with the deasphalting process, except that the paraffinic material is a product of the thermal decomposition of the asphaltene molecules and is formed in situ rather than being added separately.

In this model, there are four common features that apply to resid conversion and these features are: 1) an induction period prior to coke formation; 2) a maximum concentration of asphaltenes in the reacting liquid; 3) a decrease in the asphaltene concentration that parallels the decrease in heptane-soluble material; and 4) the high reactivity of the unconverted asphaltenes. Thus, the model can be represented as:

$$\mathrm{H}^{+} = a\mathrm{A}^{*} + (1 - a)\mathrm{V} \qquad \text{reaction constant } k_{\mathrm{H}}$$

$$\mathrm{A}^{+} = m\mathrm{A}^{*} + n\mathrm{H}^{*} + (1 - m - n)\mathrm{V}$$

reaction constant k_{A}

$$\text{Solubility limit: } \mathrm{A}^{*}_{\mathrm{max}} = S_{\mathrm{L}}(\mathrm{H}^{+} + \mathrm{H}^{*})$$

$$\mathrm{A}^{*}_{\mathrm{ex}} = \mathrm{A}^{*} - \mathrm{A}^{*}_{\mathrm{max}}$$

$$\mathrm{A}^{*}_{\mathrm{ex}} = (1 - y)\mathrm{TI} + y\mathrm{H}^{*}$$

where a is a stoichiometric coefficient; A^{+}, reactant asphaltenes; A^{*}, asphaltene cores; $\mathrm{A}^{*}_{\mathrm{max}}$, maximum asphaltene cores that can be held in solution; $\mathrm{A}^{*}_{\mathrm{ex}}$, excess asphaltene cores beyond what can be held in solution; H^{+}, reactant, nonvolatile heptane-soluble materials; H^{*}, product, nonvolatile heptane-soluble materials; k_{A}, first-order reaction rate constant for reactant asphaltene thermolysis (min^{-1}); k_{H}, first-order reaction rate constant for the thermolysis of reactant heptane-soluble materials (min^{-1}); m, stoichiometric coefficient; n, stoichiometric coefficient; S_{L}, solubility limit (wt%/wt%); TI, toluene-insoluble coke; and V, volatile materials. The first two parallel first-order reactions for the thermolysis of unreacted heptane-soluble materials and unreacted asphaltenes are the only reactions that occur during the coke induction period.

Conversion Processes

Resid processes can be made up of combinations of separations, thermal conversion, hydroconversion or hydrotreating, and fluid catalytic cracking. Distillation is the most common separation choice because of its low cost. However, fuel deasphalting is becoming more popular because it can separate the saturate constituents and small ring aromatics out of resids. Nevertheless, the selection of the conversion process depends on the quality of the resid or the selection of the crude for the refinery depends on the resid conversion process. Resids high in saturates and small ring aromatics and low in resins and asphaltenes are more appropriate for catalytic processing. On the other hand, resids high in resins and asphaltenes are more suited for coking.

Visbreaking

Visbreaking is a relatively mild thermal (noncatalytic) cracking process that is used to reduce the viscosity of residua. A visbreaker reactor may be similar to a delayed coker with a furnace tube followed by a soaker drum. However, the drum is much smaller in volume to limit the residence time with the entire liquid product flowing overhead. Alternatively, the entire visbreaker may be a long tube coiled within a furnace. Coke formation can occur and the coke accumulates on visbreaker walls; periodic decoking (cleaning) is necessary.

The visbreaking process uses the approach of mild thermal cracking to improve the viscosity characteristics of a residuum without attempting significant conversion to distillates. Low residence times are required to avoid gas and coke production and the hot liquid is quenched as it exits the reactor. Visbreaking conditions range from 450°C (840°F) to about 510°C (950°F) at the heating coil outlet with pressures varying from 50 to 300 psi. There are a number of different configurations for visbreaking units that depend on the product slate and refinery requirements.

In general, the resid is passed through a furnace where it is heated to a temperature of 480°C (895°F) under an outlet pressure of about 100 psi. The heating coils in the furnace are arranged to provide a soaking section of low heat density, where the charge remains until the visbreaking reactions are completed. Cracking is initiated in the liquid phase and continues in the reaction chamber and the cracked products are then passed into a flash-distillation chamber. The overhead material from this chamber is then fractionated to produce a low-quality gasoline as an overhead product and light gas oil as bottoms.

The reactions are not allowed to proceed to completion; the hot liquid reaction mix is quenched by the addition of light gas oil and then sent to a vacuum fractionator. This yields a heavy gas oil distillate and a residuum of reduced viscosity. Quench oil may also be used to terminate the reactions. An alternative process design uses lower furnace temperatures and longer reaction (contact) time, which is achieved by installing a soaking drum between the furnace and the fractionator. The disadvantage of this approach is the high potential for coke deposition in the soaking drum and the subsequent need to remove the coke.

The reduction in viscosity of distillation residua tends to reach a limiting value with conversion, although the total product viscosity can continue to decrease. The minimum viscosity of the unconverted residue can lie outside the range of allowable conversion if sediment begins to form. When shipment of the visbreaker product by pipeline is the process objective, addition of a diluent such as gas condensate can be used to achieve a further reduction in viscosity.

Conversion of residua in visbreaking follows first-order reaction kinetics. The minimum viscosity of the unconverted residue can lie outside the range of allowable conversion if sediment begins to form. When pipelining of the visbreaker product is the process objective, addition of a diluent such as gas condensate can be used to achieve a further reduction in viscosity. However, the main limitation of the visbreaking process, and for that matter all thermal processes, is that the products can be unstable. Thermal cracking at low pressure gives olefins, particularly in the naphtha fraction. These olefins give a very unstable product, which tends to undergo secondary reactions to form gum and intractable residua. Modification of visbreaking (thermal) chemistry is possible and more stable products are produced by application of the hydrogen donor solvents as in the hydrogen donor visbreaking process. There also remains the (potential) issue of sediment disposal, especially when the sediment is too contaminated for further use.

Important variables in the visbreaking process include temperature, pressure, and residence time. Any one variable can be changed (within predetermined limits) to alter the product slate. For example, raising the heater outlet temperature increases the yield of distillates and gaseous products.

Coking

Coking is a thermal process for the continuous conversion of resids into lighter products. The feedstock can be reduced crude, straight-run residua, or cracked residua, and the products are gases, naphtha, fuel oil, gas oil, and coke. The gas oil may be the major product of a coking operation and serves primarily as a feedstock for catalytic cracking units. The coke obtained is usually used as fuel, but processing marketing for specialty uses, such as electrode manufacture, production of chemicals, and metallurgical coke, is also possible and increases the value of the coke. For these uses, the coke may require treatment to remove sulfur and metal impurities.

In the coking processes the resid is heated to high temperatures (450–500°C) with the light boiling products distilling overhead and leaving the solid coke behind. Coking is the most popular conversion choice for resids as it reactively separates the polynuclear aromatic systems and the heteroatoms) into a low-value by-product (coke) that is counterbalanced by an increase of hydrogen content in the liquid product.

The yield of coke in a given coking process tends to be proportional to the carbon residue content of the feed (measured as the Conradson carbon residue), which is, in turn, related to the atomic hydrogen/carbon ratio.

Delayed Coking

The delayed coking process is the oldest and the most popular choice for resid upgrading. It is a semicontinuous process that can be applied to the conversion of most types of resids and residua.

The feedstock is introduced into a furnace whose outlet temperature varies from 480°C to 515°C (895–960°F). The heated feedstock enters one of a pair of coking drums where the cracking reactions continue. The cracked products leave as overheads and the coke deposits form on the inner surface of the drum. To give continuous operation, two drums are used; while one is on steam, the other is being cleaned. The temperature in the coke drum ranges from 415°C to 450°C (780–840°F) at pressures from 15 to 90 psi. Overhead products go to the fractionator, where naphtha and heating oil fractions are recovered. The heavy recycle material is combined with preheated fresh feed and returned to the reactor.

The coke drum is usually onstream for about 24 hr before becoming filled with porous coke. The following

procedure is used to remove the coke: 1) the coke deposit is cooled with water; 2) one of the heads of the coking drum is removed to permit the drilling of a hole through the center of the deposit; and 3) a hydraulic cutting device that uses multiple high-pressure water jets is inserted into the hole and the wet coke is removed from the drum. Normally, 24 hr is required to complete the cleaning operation and to prepare the coke drum for subsequent use onstream.

The Selective Yield Delayed Coking (SYDEC) process allows adjustment of the yields of coke and liquid products by adjusting the temperature. For example, increasing the temperature increases the yield of liquid products and the gas oil end point. Increasing the pressure and/or the recycle ratio increases the gas yield and the coke yield whereas the yield of liquid product and the gas oil end point are decreased.

Fluid Coking and Flexicoking

Fluid coking is a continuous process that employs the fluidized solids technique for the conversion of heavy low-grade feedstocks to lighter, more valuable products. The feedstocks charged to a fluid coker may be any type of resid or residuum where the carbon residue falls into the range 5–50 wt% or those materials having an API gravity less than 20.

The feedstock is sprayed onto a fluidized bed of hot coke particles that are agitated by the gaseous products rising through the bed in the reactor. The fluidized solids technique permits the use of higher temperatures (than delayed coking) but without the usual over abundance of coke formation because of the shorter contact times with the result that higher yields of liquid products are produced.

The residuum is decomposed by being sprayed into a fluidized bed of hot, fine coke particles, which permits the coking reactions to be conducted at higher temperatures and shorter contact times than can be employed in delayed coking. Moreover, these conditions result in decreased yields of coke; greater quantities of more valuable liquid product are recovered in the fluid coking process. The process employs two vessels, a reactor and a burner with coke particles circulated between the two to transfer heat (generated by burning excess coke in the burner) to the reactor. The preheated feedstock (260–370°C; 500–700°F) is injected directly onto the hot coke (480–565°C; 895–1050°F), at approximately atmospheric pressure, to crack to additional coke and liquid products that leave the reactor as overhead oil.

In the reactor, the coke particles flow down through the vessel into a stripping zone at the bottom. Steam displaces the product vapors between the particles, and the coke then flows into a riser that leads to the burner. Steam is added to the riser to reduce the solids loading and to induce upward flow. The average bed temperature in the burner is 590–650°C (1095–1200°F), and air is added as needed to maintain the temperature by burning part of the product coke. The pressure in the burner may range from 5 to 25 psi. Flue gases from the burner bed pass through cyclones and discharge to the stack. Hot coke from the bed is returned to the reactor through a second riser assembly.

The yields of products are determined by the feed properties, the temperature of the fluid bed, and the residence time in the bed. The use of a fluidized bed reduces the residence time of the vapor-phase products in comparison to delayed coking, which in turn reduces cracking reactions. The yield of coke is thereby reduced, and the yield of gas oil and olefins increased. An increase of 5°C (9°F) in the operating temperature of the fluid-bed reactor typically increases gas yield by 1% wt/wt and naphtha by about 1% wt/wt.

The bed temperature in the burner is of the order of 590–650°C (1095–1200°F), and any excess coke that is not removed as part of the burning is periodically removed from the burner. The coke yield from the process may be as little as 1.2 or as high as 1.7 times the carbon residue of the feedstock. As with delayed coking, the fluid coking process is capable of producing liquid products with substantially lower sulfur contents than the feedstock, but part of the sulfur in the feedstock is concentrated in the coke. There is elimination of sulfur into the gaseous products but uses for the coke depend very much on the amount of feedstock sulfur.

Flexicoking

Flexicoking is a direct descendent of fluid coking and uses the same configuration as the fluid coker but includes a gasification section in which excess coke can be gasified to produce refinery fuel gas. The flexicoking process was designed during the late 1960s and the 1970s as a means by which excess coke could be reduced in view of the gradual incursion of the heavier feedstocks in refinery operations. Such feedstocks are notorious for producing high yields of coke (>15% by weight) in thermal and catalytic operations.

In the process, excess coke is converted to a low-heating value gas in a fluid bed gasifier with steam and air. The air is supplied to the gasifier to maintain temperatures of 830–1000°C (1525–1830°F), but is insufficient to burn all of the coke. Under these reducing conditions, the sulfur in the coke is converted to hydrogen sulfide, which can be scrubbed from the gas prior to combustion.

Yields of liquid products from flexicoking are the same as from fluid coking because the coking reactor is unaltered. As with fluid coking, the extent of desulfurization depends on the sulfur content of the feedstock as well as on the chemical nature of the sulfur in the feedstock.

Catalytic Cracking

Catalytic cracking is a conversion process that can be applied to resids. It is one of the several practical applications used in a refinery that employ a catalyst to improve process efficiency. The original incentive to develop cracking processes arose from the need to increase gasoline supplies. Because cracking could virtually double the volume of gasoline from a barrel of crude oil, the purpose of cracking was wholly justified. The process employs a variety of reactors with bed types varying from fixed beds to moving beds to fluidized beds.

The fixed-bed process was the first to be used commercially and uses a static bed of catalyst in several reactors, which allows a continuous flow of feedstock to be maintained. Thus, the cycle of operations consists of: 1) flow of feedstock through the catalyst bed; 2) discontinuance of feedstock flow and removal of coke from the catalyst by burning; and 3) insertion of the reactor onstream. The moving-bed process uses a reaction vessel in which cracking takes place and a kiln in which the spent catalyst is regenerated; catalyst movement between the vessels is provided by various means. The fluid-bed process differs from the fixed-bed and moving-bed processes insofar as the powdered catalyst is circulated essentially as a fluid with the feedstock. The several fluid catalytic cracking processes in use differ primarily in mechanical design. Side-by-side, reactor–regenerator configuration or the reactor either above or below the regenerator is the main mechanical variation. From a flow standpoint, all fluid catalytic cracking processes contact the feedstock and any recycle streams with the finely divided catalyst in the reactor.

Catalytic cracking using a fluidized bed is the most popular form of cracking and is the emphasis of this section. To reaffirm this statement, there are more than 370 fluid catalytic cracking units in use worldwide with the capacity to produce more than 460,000,000 gal of gasoline from heavier feedstocks.

In the fluidized-bed (fluid catalytic cracking) units, the cracking reactions take place in the riser to form products, including coke. In the riser, the catalyst and the feedstock and products rise up the reactor pipe, and because the reactions are predominantly endothermic, the reaction temperature declines from bottom to top. At the top of the riser, the mixture enters a solid–gas separator, and the product vapors are led away. The coked catalyst passes to the stripper where steam is added and unreacted/reacted feedstocks adsorbed on the catalyst are released. The stripped catalyst is then directed into the regenerator where air is added and the combustion of coke on the catalyst (and any feedstock/product still adsorbed that was not stripped) occurs with the liberation of heat. Regenerator temperatures are typically 705–760°C (1300–1400°F). Heat exchangers and the circulating catalyst are used to retain and capture the regeneration heat for use in preheating the feedstock.

The entire catalyst inventory is continually circulated through the three parts of the unit. The catalyst residence time in the riser reactor section is typically 1–3 sec (with current trends to even shorter residence times), and the entire reactor–stripper–regenerator cycle is less than 10 min. To achieve cycle times of this order, catalyst circulation rates as high as 1 ton/sec are observed in large units.

The catalyst, which may be an activated natural or synthetic material, is employed in bead, pellet, or microspherical form and can be used as a fixed-bed, moving-bed, or fluid-bed configurations. Moving-bed units often employ catalysts in the form of beads or cylinders ($\frac{1}{8}$ – $\frac{1}{4}$ in. diameter). Fluid-bed units usually employ the catalyst in much smaller form where particle sizes may be of the order of 50 μm (50×10^{-4} cm).

Catalytic cracking in the usual commercial process involves contacting the feedstock with a catalyst under suitable conditions of temperature, pressure, and residence time. Catalytic cracking processes produce coke and overheads (volatile materials). The coke deposits on the catalyst and, therefore, markedly decreases catalyst activity. This problem is surmounted by transference of the catalyst from the reactor to a vessel (regenerator) where the coke is burned from the catalyst in air; movement of the catalyst between the vessels is continuous, thereby maintaining feedstock flow and production of products. The reactor temperature is usually on the order of 465–510°C (870–950°F) and pressures in the range 10–16 psi. The regenerator temperature may be substantially higher at 565–675°C (1050–1250°F) and 8–10 psi. The feedstock temperatures vary from 315°C to 455°C (600–850°F) at the point of entry to the reactor.

Although catalytic cracking was originally designed to convert gas oil to gasoline, process modifications have allowed the feedstock types to include resids and residua. The manner in which a feedstock will crack depends on a number of process variables such as feedstock composition, boiling range, structural types present, reactor conditions, and catalyst type. Thus, it is not surprising that the heavier feedstocks present special problems when used to feed a catalytic cracker. The susceptibility of the catalyst to the various

nitrogen compounds, as well as the metals (iron, nickel, vanadium, and copper) in the feedstock, represents a major drawback in using resids and residua as catalytic cracker feedstock. For example, if the concentration of metals (expressed in ppm) on the catalyst fulfills the following relation:

$$4V + 14Ni + Fe + Cu > 1000$$

the catalyst is severely contaminated and may actually have to be replaced.

Nevertheless, it is possible to process the heavier feedstocks in catalytic crackers and bring about some degree of desulfurization in the process. The degree of desulfurization depends on the amount of sulfur in the feedstock. It is generally believed that to optimize use of a catalytic cracking unit, feedstocks should be treated to remove excess high molecular weight material and metals by processes such as visbreaking, coking, or deasphalting to prolong catalyst activity.

Sulfur compounds are changed in such a way that the cumulative sulfur content of the liquid and nonvolatile products is lower than the sulfur content of the original feedstock. The decomposition of sulfur constituents into hydrocarbons and hydrogen sulfide (or other gaseous sulfur products) occurs.

Hydroconversion

Hydroconversion (variously referred to as hydrotreating and hydrocracking) is a group of refining processes in which the feedstock is heated with hydrogen under pressure. The outcome is the conversion of resids to a range of products.

Hydrotreating is generally used for the purpose of improving product quality without appreciable alteration of the boiling range. Mild processing conditions are employed so that only the more unstable materials are attacked. Thus, nitrogen, sulfur, and oxygen compounds undergo hydrogenolysis to split ammonia, hydrogen sulfide, and water, respectively. Olefins are saturated, and unstable compounds, such as di-olefins, which might lead to the formation of gums or insoluble materials, are converted to more stable compounds. Heavy metals present in the feedstock are also usually removed during hydrogen processing.

On the other hand, hydrocracking is a process in which thermal decomposition is extensive and the hydrogen assists in the removal of the heteroatoms as well as mitigates the coke formation that usually accompanies thermal cracking of high molecular weight polar constituents. Hydrocracking is similar to catalytic cracking, with hydrogenation superimposed and with the reactions taking place either simultaneously or sequentially. Hydrocracking was initially used to upgrade low-value distillate feedstocks, such as cycle oils (highly aromatic products from a catalytic cracker that usually are not recycled to extinction for economic reasons), thermal and coker gas oils, and heavy-cracked and straight-run naphtha. These feedstocks are difficult to process by either catalytic cracking or reforming, because they are usually characterized by a high polycyclic aromatic content and/or by high concentrations of the two principal catalyst poisons, sulfur, and nitrogen compounds.

Hydrocracking is perhaps the single most significant advance in petroleum refining technology over the last several decades. It is essentially an efficient thermal catalytic method of converting resids to lower boiling products. However, hydrocracking should not be regarded as a competitor for catalytic cracking. Catalytic cracking units normally use virgin gas oils as feedstocks; hydrocracking feedstock usually consists of refractive gas oils derived from cracking and coking operations. Hydrocracking is a supplement to, rather than a replacement for, catalytic cracking.

The problems encountered in hydrocracking resids can be directly equated to the amount of complex, higher boiling constituents that may require pretreatment. Processing these feedstocks is not merely a matter of applying know-how derived from refining conventional crude oils but also requires knowledge of composition.

The choice of processing schemes for a given hydrocracking application depends on the nature of the feedstock as well as the product requirements. The single-stage process can be used to produce gasoline but is more often used to produce middle distillate from heavy vacuum gas oils. The two-stage process was developed primarily to produce high yields of gasoline from straight-run gas oil, and the first stage may actually be a purification step to remove sulfur-containing (as well as nitrogen-containing) organic materials. Both processes use an extinction–recycling technique to maximize the yields of the desired product. Significant conversion of resids can be accomplished by hydrocracking at high severity. For some applications, the products boiling up to 340°C (650°F) can be blended to give the desired final product.

Product yields and hydrogen consumption vary with the feedstock but the process can be used in many different ways. One of the major advantages of hydrocracking is that it may be used to process the higher-boiling refractory feedstocks that may be produced by catalytic cracking or by any of the coking processes. Hydrocracking can also be applied to the conversion of the more difficult feedstocks, such as residua and resides, and there are a variety of hydrocracking processes that are designed specifically for this particular use.

Hydrocracking is an extremely versatile process that can be utilized in many different ways. One of the

advantages of hydrocracking is its ability to break down high-boiling aromatic stocks produced by catalytic cracking or coking. This is particularly desirable when maximum gasoline and minimum fuel oil must be made.

However, it must not be forgotten that product distribution and quality vary considerably depending on the nature of the feedstock constituents as well as on the process. In modern refineries, hydrocracking is one of several process options that can be applied to the production of liquid fuels from the heavier feedstocks. The most important aspect of the modern refinery operation is the desired product slate, which dictates the matching of a process with any particular feedstock to overcome differences in feedstock composition.

The reactions that are used to chemically define the processes (i.e., cracking and subsequent hydrogenation of the fragments, hydrogenation of unsaturated material, hydrodesulfurization, hydrodenitrogenation) may all occur. Hydrocracking a feedstock will, in all likelihood, be accompanied by hydrodesulfurization, thereby producing not only low-boiling products but also low-boiling products that are low in sulfur.

Hydroprocesses offer direct desulfurization of resids and high conversion. These processes were not originally designed for resids and yet the evolution of the various processes and catalysts has seen their application to resids and residua. Application of a hydroprocess as a pretreatment process has some merit. Changing the character of the feedstock constituents can make for easier conversion in a later stage of the refining sequence. The advantages are:

1. The products require less finishing.
2. Sulfur is removed from the catalytic cracking feedstock, and corrosion is reduced in the cracking unit.
3. Carbon formation during cracking is reduced and higher conversions result.
4. The catalytic cracking quality of the gas oil fraction is improved.

The downside to the direct application of hydroprocesses to the resids is always the cost of hydrogen and the short lifetime of the catalyst (which in effect is also a cost). There is the potentially wasteful use of hydrogen with hydrogen sinks within the feedstock whereupon hydrogen is used, but with little, if any, effect on the product character.

Fixed-Bed Processes

Because metal removal is one of the fastest reactions and the metals accumulate in the pores of supported catalysts, it is common to have a guard bed in front of the fixed bed. When insufficient demetallization activity occurs in the guard bed, the feed is switched to a second guard bed with fresh catalyst and the catalyst is replaced in the first guard bed. Thus, the fixed bed is protected from deposition of metals. To hydrogenate the largest macromolecules in the resid, the asphaltenes, some or all of the catalysts need to have pores 50–100 μm in diameter. Even with these precautions it is difficult to get longer than 1 yr run lengths on fixed-bed hydroconversion units with vacuum resid feeds and conversions to volatile liquids of 50% or more. This is due to catalyst deactivation with coke or to coke and sediment formation downstream of the reactor.

Ebullating Bed Processes

The LC Finer and H-Oil units mechanically ebullate the catalyst so that it can be mixed and replaced onstream. The conversion is greatly dependent on the feed but conversions of vacuum resids to volatile liquids of the order of 70% is possible. Often these units are limited by the deposition of coke and sediment downstream of the reactor in hot and cold separators.

Dispersed Catalysts Processes

If one cannot diffuse the asphaltenes to the catalyst, why not diffuse the catalyst to the asphaltenes? Dispersed catalysts also can be continuously added in sufficiently low enough amounts (i.e., 100 ppm) to consider them throwaway catalysts with the carbonaceous by-product. However, economics usually dictate some form of catalyst recycle to minimize catalyst cost. Nevertheless, by designing the reactor to maximize the solubility of the converted asphaltenes, the conversion of vacuum resids to gas and volatile liquids can be above 95% with greater than 85% volatile liquids. However, the last 5–10% conversion may not be worth the cost of hydrogen and reactor volume to produce hydrocarbon gases and very aromatic liquids from this incremental conversion. The answer depends on the value and use of the unconverted carbonaceous liquid by-product.

CONCLUSIONS

Resid conversion is now in a significant transition period as the demand for transportation fuels increases. To satisfy the changing pattern of product demand, significant investments in resid conversion processes

will be necessary and technologies are needed that will take resid conversion beyond current limits and, at the same time, reduce the amount of coke and other nonessential products.

More options are now being sought to increase process efficiency in terms of the yields of the desired products. New processes for resid conversion probably will be used in conjunction with visbreaking and coking options with some degree of hydroprocessing as a primary conversion step. In addition, other processes such as asphalt coking technology, the Cherry-P, or the eureka process may replace or, more likely, augment the deasphalting units in many refineries.

There remains room for improving coking and hydroconversion processes by reducing hydrocarbon gas formation, inhibiting the formation of polynuclear aromatic compounds not originally present in the resid, and separating an intermediate quality (low in cores) fraction before or during conversion. In addition, the challenge for hydroconversion is to take advantage of the nickel and vanadium in the resid to generate an in situ dispersed catalyst and to eliminate catalyst cost. Finally, resid catalytic cracking needs to move to poorer quality and lower-cost feeds by making more tolerant catalysts.

BIBLIOGRAPHY

Dolbear, G.E. *Petroleum Chemistry and Refining*; Speight, J.G., Ed.; Taylor & Francis: Washington, DC, 1998; 7 pp.

Gray, M.R. *Upgrading Petroleum Residues and Resids*; Marcel Dekker Inc.: New York, 1994.

Speight, J.G. *The Chemistry and Technology of Petroleum*, 3rd Ed.; Marcel Dekker Inc.: New York, 1999.

Speight, J.G. *The Desulfurization of Heavy Oils and Residua*, 2nd Ed.; Marcel Dekker Inc.: New York, 1999.

Speight, J.G.; Ozum, B. *Petroleum Refining Processes*; Marcel Dekker Inc.: New York, 2002; 538–587.

Wiehe, I.A. A solvent-resid phase diagram for tracking resid conversion. Ind. Eng. Chem. Res. **1992**, *31*, 530.

Wiehe, I.A.; Liang, K.S. Fuel Sci. Technol. Int. **1996**, *14*, 289.

Rheology of Cellulose Liquid Crystalline Polymers

R

Qizhou Dai
Faculty of Foresty, Biomaterials Chemistry, The University of British Columbia, Vancouver, British Columbia, Canada

Richard Gilbert
Department of Wood and Paper Science, North Carolina State University, Raleigh, North Carolina, U.S.A.

John F. Kadla
Faculty of Foresty, Biomaterials Chemistry, The University of British Columbia, Vancouver, British Columbia, Canada

INTRODUCTION

Cellulose and many of its derivatives form liquid crystalline phases in solutions and melts. Because of the chirality of the cellulose backbone, cellulosic liquid crystalline phases form chiral nematic structures. Chiral nematic mesophases possess a unique structure in which the alignment of molecular sheets is at a slight angle to one another resulting in a helicoidal supramolecular structure. It can be described by a pitch p (or its inverse, the twist p^{-1}); $p = \lambda_0/\tilde{n}$, where λ_0 is the reflection wavelength and $\tilde{n}$ is the mean refractive index of a sheet, and the corresponding handedness of the twist; right-handed helicoidal structures being assigned to a positive pitch ($p > 0$) and left-handed helicoidal structures to a negative pitch ($p < 0$). This type of supramolecular arrangement results in unique optical and physical properties. As a result, many lyotropic cellulosic systems have been developed with potential applications ranging from liquid crystalline displays to high modulus, high strength regenerated cellulose fibers. In recent years, systems that display a reversal of their helical twisting sense by solvent or chemical modification have been of particular interest.

The viscosity of these chiral nematic mesophases is concentration-dependent and strongly influenced by shear rate. When the shear rate is high enough, the polymer chains will orient along the shear direction. The chiral nematic structure changes to a flow-aligned nematic-like phase and the dependence of viscosity on concentration decreases. Undoubtedly during the shear process, shear forces cause an orientation in the shear direction of the twisted macromolecules of the chiral nematic structure. A quasi-nematic phase is formed with some dispersed chiral nematic domains. Upon removal of the shear stress, the thermodynamically more stable chiral nematic phase will reform leading to a disruption in the macromolecular orientation within the forming products, and as a result lower than expected modulus and strength. In this entry, we report on the transient rheological behavior of lyotropic cellulosic mesophases with varying pitch and handedness and discuss the relationship between the chiro-optical properties and relaxation behavior.

CELLULOSIC LIQUID CRYSTALS

Cellulose is the most abundant natural polymer. Cellulose and its derivatives (cellulosics) have played an important role in developing and establishing the current concepts and industrial applications of polymer science. Current interests are based on their versatile properties, biodegradability, and status as a renewable resource (Fig. 1).

Discovery of the formation of liquid crystalline solutions by cellulosics in the mid-1970s[1] has resulted in attempts to develop new cellulosics products with properties superior to those of conventional cellulosic. Following the first observation of mesophases formed in aqueous solutions of hydroxypropyl cellulose (HPC),[1] a variety of other cellulose derivatives have been reported to form liquid crystals.[2–7] Liquid crystalline solutions of cellulose and its derivatives[8–12] provide a potential route to high-modulus and high-tenacity cellulosic fibers, films, and other high-performance products.

Efforts to make high-performance fibers and films from cellulosic mesophases have been made.[12–17] For example, cellulose fibers produced from cellulosic mesophases show properties superior to those of commercially available fibers.[16,18,19] Although these fibers are superior to commercial products, their physical properties are lower than theoretically predicted. This is in part because of that the ordered structures

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120040351

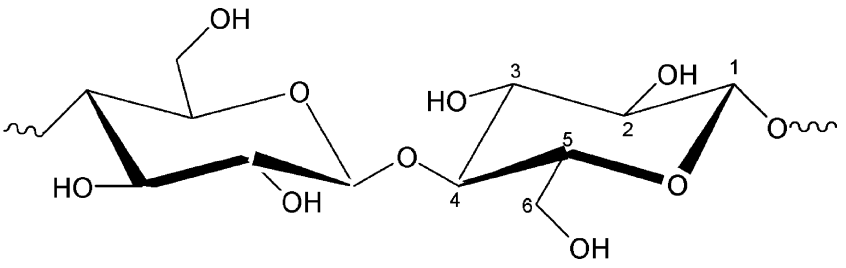


Fig. 1 Structural representation of cellulose.

are not well preserved because of the orientation relaxation during the processing steps.

Liquid crystalline (LC) solutions of cellulose derivatives form chiral nematic (cholesteric) phases.[6,20] Chiral nematic phases are formed when optically active molecules are incorporated into the nematic state. A fingerprint texture is generally observed under crossed polarizers for chiral nematic liquid crystals when the axis of the helicoidal structure is perpendicular to the incident light (Fig. 2).

In general, chiral nematic polymer liquid crystals (LCP) cannot form monodomains in which the rodlike polymers have a spatially uniform orientation within the sample. Typically, because of the high density of orientational defects, the LCPs are textured, with a distribution of polymer orientation. Microscopically, the polymer chains have a preferred orientation with a relatively narrow distribution around the average orientation. Macroscopically, the variation in space of the orientation results in a domain structure. Defects and orientational variations give rise to the polydomain texture and the overall LCP sample may be randomly ordered (Fig. 3).

Lyotropic liquid crystalline cellulose derivatives exhibit unique optical properties because of their helicoidal supramolecular structure.[21] The chiro-optical properties of the helicoidal structure can be described by a pitch p (or its inverse, the twist p^{-1}); $p = \lambda_0/\tilde{n}$, where λ_0 is the reflection wavelength and $\tilde{n}$ is the mean refractive index of a sheet, and the corresponding handedness of the twist; right-handed helicoidal structure being assigned to a positive pitch ($p > 0$) and left-handed helicoidal structures to a negative pitch ($p < 0$).[5,6,22] The nematic mesophase can be considered as a critical state of chiral nematic phase, in which the helicoidal structure has infinite pitch and no handedness. The supramolecular structure of chiral nematic mesophase indicates the effects of polymer–solvent and polymer–polymer interactions in the mesophase (Fig. 4).

The effect of polymer–solvent interactions on the mesophase can be derived from the rigidity of the polymer chain, the critical concentration to form liquid crystalline phase, and relaxation studies. After shearing a rigid polymeric liquid crystal, a banded texture is formed in which the direction of the bands

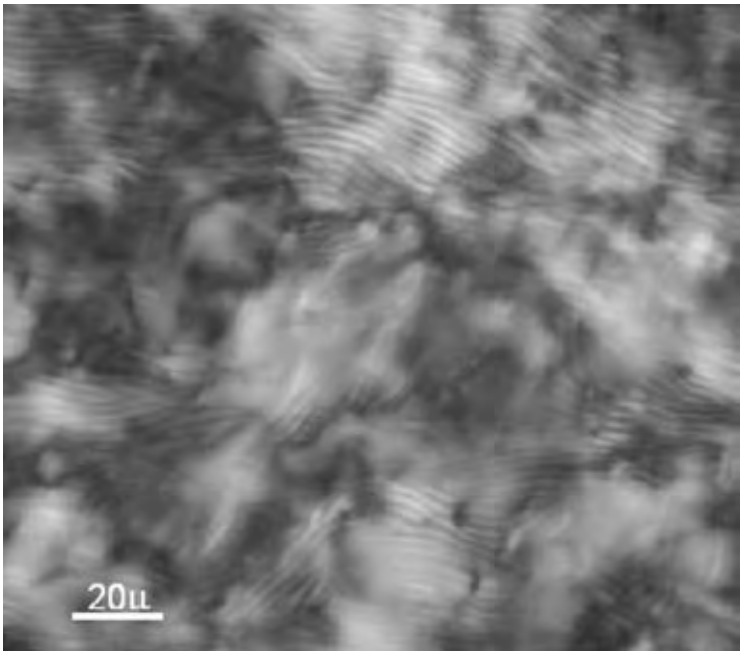


Fig. 2 Fingerprint pattern for chiral nematic mesophase of 40% EC in *m*-cresol. *(View this art in color at www.dekker.com.)*

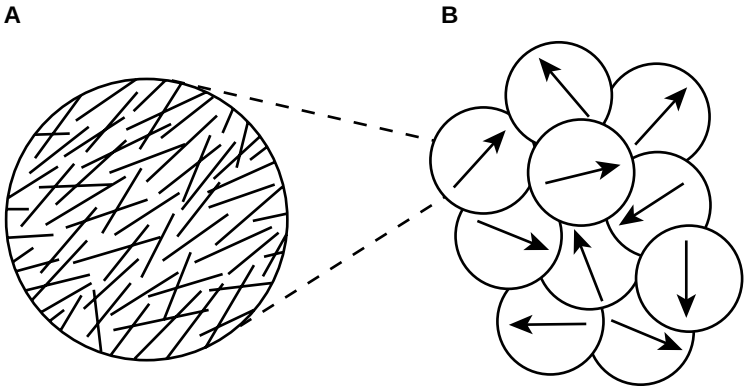


Fig. 3 Hierarchy of the distribution of the orientation of the molecules in a polydomain LCP. (A) Microscopic and (B) mesoscopic.

is perpendicular to the shearing direction after removing the shear force.[23,24] The banded structure forms because of a periodic packing of rigid/semirigid polymer chains in a zigzag fashion, which is believed to be owing to the relaxation of the polymer orientation.[25] For lyotropic LC solutions the banded texture is a transient phenomenon. After a period of time, the polymer chains fully relax and the banded texture disappears. The bandwidth and the relaxation time of the bands depend on the rigidity and pitch of the polymer chain resulting from specific polymer–solvent interactions. In fact, at the same concentration ethyl cellulose (EC) chains are more rigid in dichloroacetic acid (DCA) than in acrylic acid or glacial acetic acid and the pitch of EC/DCA mesophase is much higher than that of the EC/acetic acid[26] or acrylic acid,[27] resulting in wider bands and longer relaxation times for EC/DCA mesophase (Fig. 5).[28]

However, the rigidity of the polymer chain is not the only factor to determine the pitch of the mesophase. Guo and Gray[29] found that for acetylethylcellulose (AEC) in chloroform, with the increase of degree of acetylation, the rigidity of the cellulose chain increases monotonically. However, the pitch of the mesophase first increases to infinity and then decreases with an inversion of the handedness of the chiral nematic phase, from left-handed to right-handed.

As cellulosics are chiral, there must also be a chiral contribution to the interactions between the rods in the mesophase leading to the chiral nematic structure. Several theories have been proposed to account for the ''twisting'' force between chiral rodlike mesogens in the liquid crystalline phase. Goossens was the first to propose that the chiral nematic structure is the result of anisotropic dispersion energy between chiral mesogens.[30] Samulski and Samulski proposed that the macroscopic twist sense was dependent on the dielectric constant of the medium and the chirality of the constituent molecules.[31] They determined that the introduction of an asymmetric dispersion energy results in adjacent mesogens having a slight twist relative to one another, and that increasing temperature

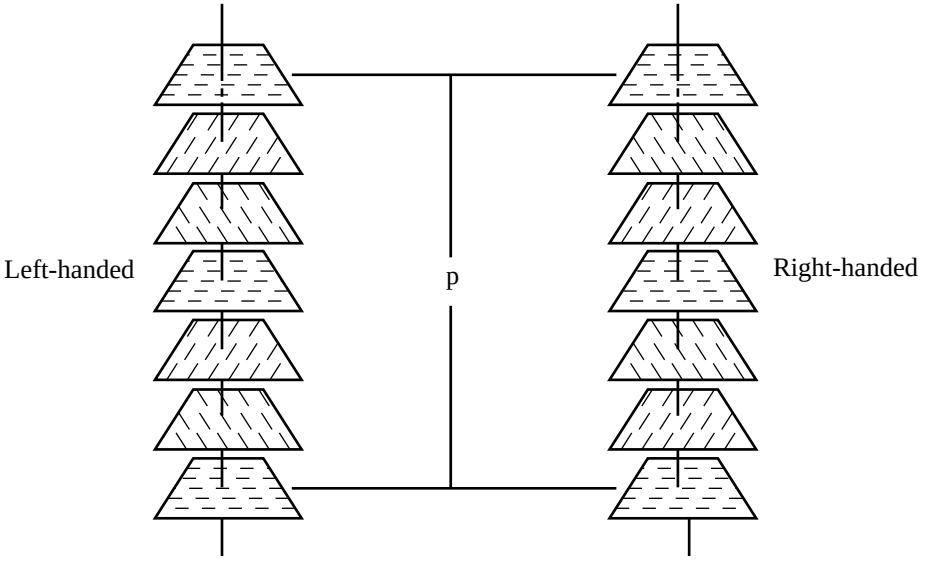


Fig. 4 Helicoidal structure of chiral nematic mesophases.

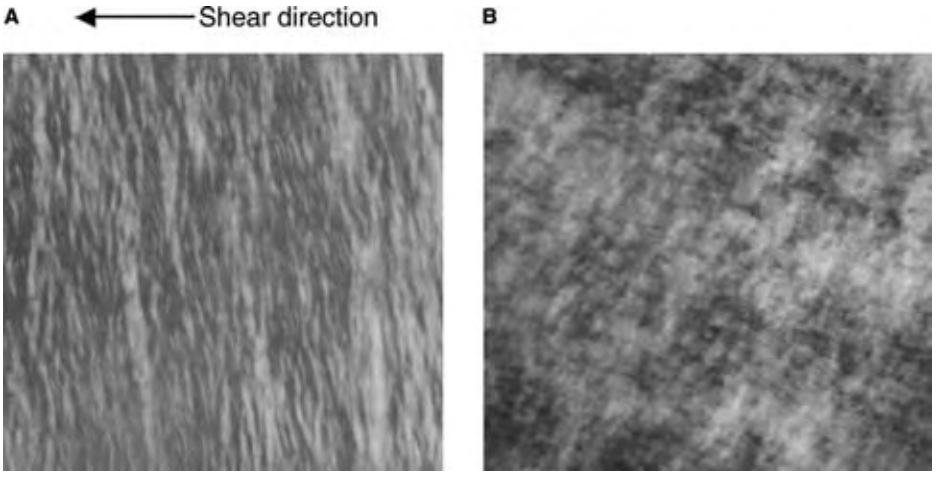


Fig. 5 Banded texture for 45% EC in *m*-cresol: (A) 1 min after shear and (B) 75 min after shear. *(View this art in color at www.dekker.com.)*

increases the frequency of molecular rotation, causing a decrease in the dispersion energy and an increase in the pitch.

The observed handedness and pitch of the helicoidal structure is not only sensitive to temperature, but also depends on the concentration and structure of the components.[26,32–37] According to Lin-Liu, Shih, and Woo[38] the temperature dependence of pitch for chiral nematic polymers does not seem to follow any particular pattern. It is believed that as temperature is increased, specific interactions, e.g., hydrogen bonding, whether inter- or intramolecular or polymer–solvent interactions are destroyed. The polymer chains become more flexible and the side groups more easily relaxed, thereby changing the physical properties of the chiral nematic structure. Similarly, an increase in concentration leads to a decrease in pitch for most lyotropic cellulosic liquid crystals with the exception of cellulose tricarbanilate (CTC) in ethyl methyl ketone, 2-pentanone,[35] or triethylene glycol monoether[34] and the chlorophenyl urethane derivative in diethylene glycol monoether.[39]

Numerous investigators have proposed theories for the temperature dependence of pitch.[40–44] However, none completely explain the experimental results obtained for the various cellulosic systems. Of those proposed, Osipov's approach, which is based on a molecular statistical theory,[42] takes into account steric and chiral interactions in solution to predict the influence of temperature and solvent on the pitch and twist sense of cellulose derivatives.[43] Assuming the cellulosic chain adopts a "twisted belt" as opposed to a helix, and the persistence length, l, of the cellulose chains is much smaller than that of a rigid chain, the twisting power could be expressed by

$$2\pi/P = -\rho^2(\chi - \lambda kT)/2K_{22} \quad (1)$$

where χ is related to the attractive interaction between chains, P the pitch, k the Boltzmann constant, K_{22} the twist elastic constant, ρ the number density of rigid segments, $\rho = cL/l_0$, where c is the number density of the macromolecular chains, l_0 the length of the segments, L the total length of the chain ($L > l > l_0$), and the pseudo-scalar parameter λ is determined by the steric repulsion between chains.

The theory predicts that the handedness of cellulosic liquid crystalline solutions, designated by the sign of the pitch, depends not only on temperature (T) and on steric repulsion of the chain (λ), but also on an attractive interaction parameter, χ, which depends on the nature of the solvent. The chiral forces are balanced when $(\chi - \lambda kT) = 0$. In this compensated condition, the pitch of the mesophase should become infinite, and the mesophase resembles a normal nematic phase.

Besides degree of substitution, nature of substituents, solvent, concentration, and temperature, other factors that change the polymer–solvent interactions can affect the pitch of lyotropic cellulosic mesophases. Doping inorganic salts[45,46] or small chiral molecules[47] into the lyotropic mesophase changes the polymer–solvent interactions. As the results, the pitch of the chiral nematic mesophase changes accordingly.

RHEOLOGY OF CELLULOSE LIQUID CRYSTALLINE SOLUTIONS

The microstructure and polymer–solvent interactions of lyotropic cellulosic mesophases can be derived from rheological studies. The lyotropic LCP solution is a complicated system and a wide range of unusual rheological phenomena have been observed.

Because of the wide range of technological applications and complexity of liquid crystal polymers, there is a strong need for further advance in the understanding of their rheological phenomena. Several theories have been introduced to predict the rheological

behavior of LCPs. The classical Leslie–Ericksen (LE) continuum theory uses a rigid rodlike director to describe the local molecular orientation.[48,49] The rheological behavior is the result of the interaction between the director field and applied flow field. Leslie–Ericksen theory is applicable when the flow field only affects the direction of molecular orientation without changing the degree of molecular alignment. This theory has been successful in discussing the flow of low molecular weight nematic liquid crystals. For LCP, it is applicable only at very low shear rates in defect-free samples.[50]

Doi molecular theory adds a probability density function of molecular orientation to model rigid rodlike polymer molecules.[51] This model is capable of describing the local molecular orientation distribution and nonlinear viscoelastic phenomena. Doi theory successfully predicts director tumbling in the linear regime and two sign changes in the first normal stress difference,[52] as will be discussed later. However, because this theory assumes a uniform spatial structure, it is unable to describe textured LCPs.

Larson and Doi introduced a mesoscopic polydomain model based on LE theory. This model includes a domain orientation distribution function and incorporates director tumbling, distortional elasticity, and texture size.[53] Larson–Doi model can qualitatively predict the steady flow behavior and transient behavior. However, discrepancies between the theoretical predictions and the experiments of model systems were observed, especially when the shear history includes rest periods.[50] This model is restricted to low shear rates without perturbing the molecular orientation distribution function in each domain.[52]

The theory for LCP rheology is still evolving. Rey and Tsuji[54] proposed a complete tensor theory to take into account three major effects of LCP flow: 1) short range order elasticity; 2) long range order elasticity; and 3) viscous flow. This theory is based on the unification of the classical nonequilibrium theories. The complete theory can predict more complex phenomena of LCP behavior, e.g., banded texture, defect generation, and coarsening. Extensive and detailed reviews of LCP rheology theory were provided by Larson,[55,56] Burghardt,[52] Rey and Denn,[57] and Marrucci and Greco.[58]

All the above theories are derived for rigid rod nematic liquid crystal systems. The rheological behavior of chiral nematic liquid crystals is more complex and less understood than that of nematic systems. Rey introduced a model based on rigid rod chiral nematic liquid crystals to describe permeation shear flow and small amplitude oscillatory shear flow.[59–61] The model can predict some common phenomena of chiral nematic liquid crystals, e.g., the three-region apparent viscosity curve,[62] oscillatory underdamped stress relaxation,[62] helix uncoiling,[63] and the higher apparent viscosity, as compared to nematic LCPs.[64] However, cellulosics are semiflexible polymers. Experimental results have shown that semiflexible HPC/H_2O system[65] has different rheological behavior from that of the more rigid poly(γ-benzyl-glutamate)/*m*-cresol system.[66] So far, no theory has been proposed to model the rheology of semiflexible chiral nematic LCPs.

Following the first observation of mesophase formation of HPC/H_2O, the rheology of this system has been widely investigated. Rheological studies of other cellulosic mesophases, e.g., cellulose,[67] ethylcellulose,[26,68,69] and cellulose tricarbanilate,[70] have also been reported.

Concentration Dependence of Viscosity

Isotropic solutions exhibit a monotonic increase in shear viscosity with increasing concentration. The viscosity increases to a maximum when the isotropic to anisotropic transition is approached. Upon formation of the anisotropic phase, the viscosity begins to decrease, after which the viscosity increases strongly as the concentration continues to increase (Fig. 6). In the isotropic state, the hydrodynamic volume is large because of the random polymer orientation. This restricts the polymer diffusivity and causes an increase in viscosity. In the anisotropic phase, the aligned polymer leads to a small hydrodynamic volume and a decrease in viscosity as rotational diffusion is much easier with a net orientation.

Zugenmaier found that the correlation of the viscosity maximum with the formation of the anisotropic phase was only valid when the shear rate was low

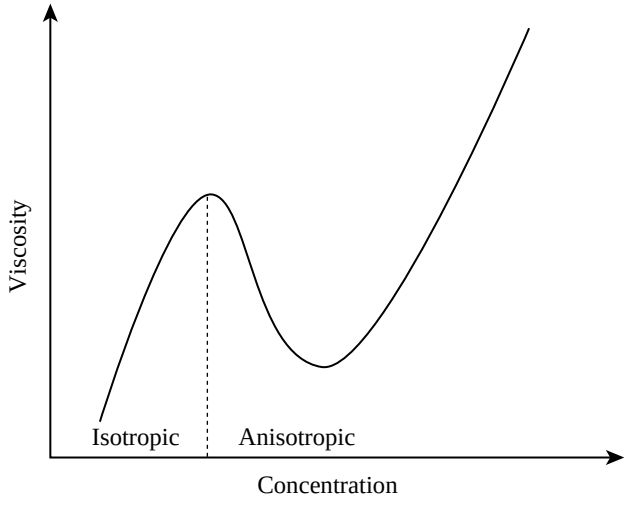


Fig. 6 Viscosity as a function of concentration for lyotropic LCPs.

(shear rate $\rightarrow$ 0).[70] At high shear rate, the concentration at which the maximum in viscosity occurs decreases. When the shear rate is high enough to cause shear-induced orientation (pseudo-nematic phase), entanglements of the random distributed polymers are released. The viscosity maximum disappears and a steady increase of viscosity vs. concentration is observed. This indicates that the viscosity of the nematic or pseudo-nematic mesophase is less sensitive to concentration than that of the chiral nematic phase.[70]

Steady Flow Behavior

Steady flow behavior is one of the most thoroughly studied rheological properties. Onogi and Asada[71] hypothesized the universal existence of three shear flow regimes to describe the viscosity of LCPs: a shear thinning regime at low shear rates (regime I), a Newtonian plateau at intermediate shear rates (regime II), and another shear thinning regime at high shear rates (regime III) (Fig. 7). The three-regime curve was observed in the anisotropic aqueous HPC solution[72] and HPC in acetic acid.[73] Asada proposed that some LCP systems do not exhibit all three regimes because not every regime lies in the accessible shear rate range.

In those systems that exhibit all three regimes, regime I is generally believed to reflect a defect structure, or stacked polydomain texture. The texture is a supramolecular disorder of the nematic structure and is composed of nematic domains with little or no macroscopic orientation.[72,74] This region is characterized by distortional elasticity associated with spatial variation in the director (average local molecular orientation) field. Regime II reflects a ''dispersed polydomain'' structure and regime III is believed to be characterized by nonlinear effects of flow on the molecular orientation. In regimes I and II, the flow is not strong enough to affect the molecular orientation. In regime III, the flow field is very strong and shear-induced molecular orientation becomes important. According to birefringence measurements for anisotropic HPC/H_2O solutions[65] and HPC/*m*-cresol solutions,[75] the molecular orientation is a monotonically increasing function of the steady state shear rate.

We found that the steady state flow of lyotropic AEC in acrylic acid (AA) showed three-regime curve, as shown in Fig. 8. Different AECs had different DA and the mesophases had different chiro-optical properties. The properties of different AECs are listed in Table 1. All the mesophases showed a typical three-regime curve. With the increase of shear rate, a shear-thinning is followed by a Newtonian plateau, and then followed by another shear-thinning region at high shear rate. The Newtonian plateau is in the shear rate range 1–5 S^{-1}. We also found that with the increase of pitch, the flow curve becomes flat and the three regimes are not as distinct as the one with smaller pitch. Even though several cellulosic mesophase systems show a three-regime flow curve, in fact, it appears that all three regimes are not typically accessible for all LCPs including cellulosics. Whether the three-regime flow curve is indeed universal or not remains a subject of debate.

The most striking phenomenon in the steady flow of LCPs is the occurrence of a negative first normal stress difference (N_1)—two sign changes in N_1 as a function of shear rate. In contrast, isotropic solutions only exhibit positive N_1 at all shear rates. Negative N_1 has been reported in HPC/H_2O[76–78] and HPC/*m*-cresol systems.[79]

The negative N_1 is the result of the coupling of molecular tumbling under flow and the local molecular-orientation distribution.[76] At low shear rates, the director tumbles with the flow and N_1 will be positive. At intermediate shear rates, nonlinear viscoelastic effects are important. The director tumbling competes with the steady director alignment along

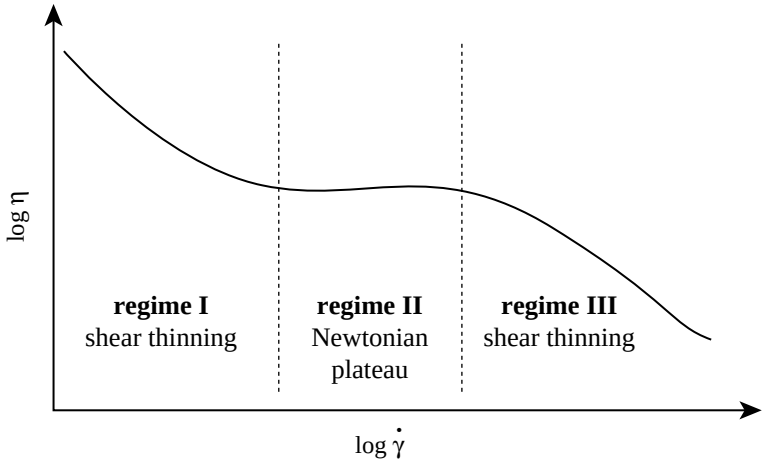


Fig. 7 Three-regime steady state shear viscosity for LCPs.

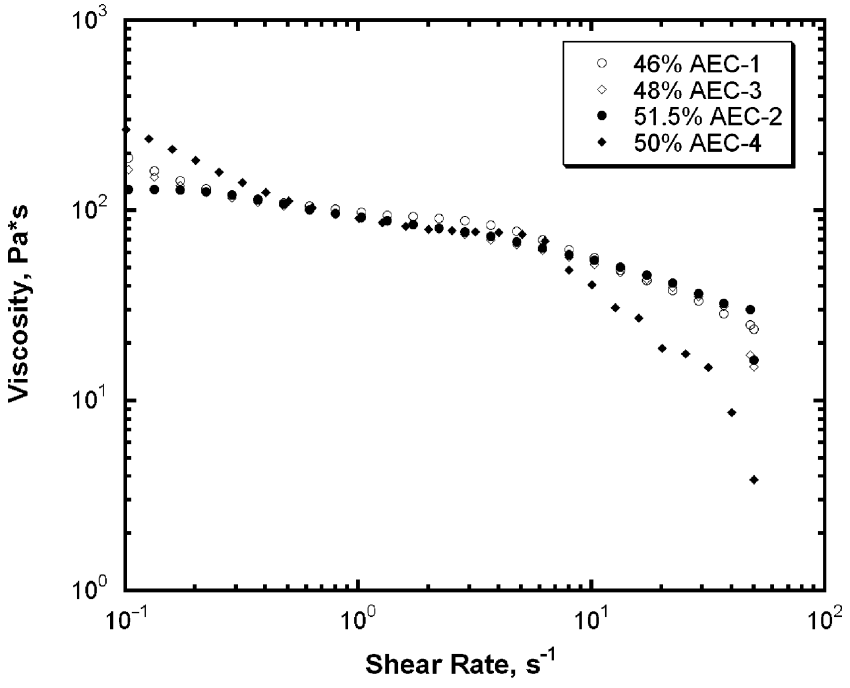


Fig. 8 Viscosity as a function of shear rate of AEC/AA solutions. (The properties of AEC samples are given in Table 1.)

the flow and the director oscillates about a steady value, N_1 becomes negative. At very high shear rates, the director aligns along the flow and N_1 is positive again. It is generally believed that the shear rate at which N_1 is minimum halts director tumbling and aligns the molecules.[79] However, negative N_1 behavior disappears at very high polymer concentration because of polymer–polymer friction interactions.[76]

Relaxation Behavior

The relaxation behavior, or the transient behavior of cellulosic liquid crystalline solutions upon the cessation of steady flow, is unique for LCPs. There are two kinds of relaxation. The bulk stresses relax quickly while the structures relax over a much longer time. Mewis and Moldenaers suggested that two levels of structures exist.[80] Stress relaxation reflects fast relaxation at the molecule level and is independent of the previous shear rate. Structure relaxation reflects the gradual change of the textures. This slow process is unique in that its time scale is inversely proportional to the previous shear rate. When the strain recovery (or recoil) after cessation of steady state flow is plotted against time multiplied by the previous shear rate, the curves superimpose one another.[50]

Table 1 Chiro-optical properties of lyotropic 50% AEC/AA solutions with different degrees of acetylation

AEC sample	Sample description	Degree of acetylation	Pitch
AEC-1	Low DA	0.13	−360 nm
AEC-2	Medium DA	0.34	−7.4 μm
AEC-3	AEC/EC mixture	0.34	+1.2 μm
AEC-4	High DA	0.50	+370 nm

It is generally believed that under high flow rate, the chiral nematic mesophase aligns along the flow direction and uncoils to form a nematic structure. Upon cessation of flow, the chiral nematic phase will reform and the molecular orientation will decrease.[81] By using evolution of the dynamic moduli[82] and birefringence[52] as a function of time, the structural change can be investigated. For lyotropic HPC/H_2O solutions, upon flow cessation, molecular orientation decreases to a globally isotropic condition at all rates[52] and the dynamic moduli increase to a maximum.[82] However, the final relaxed state depends on the shear history. After high shear rates, the solutions evolve toward an ''equilibrium'' state with a high modulus, while after low shear rates, the solutions relax to the ''equilibrium'' state with a low modulus.[65] The low-modulus state is ordered and evolves out of a state that has no macroscopic order upon cessation of flow. The high-modulus state is much less ordered, although it evolves from a rather well flow-aligned state upon the cessation of flow.[83]

At high shear rates the chiral nematic structure changes to a flow-induced nematic phase. However, the shear-oriented phase is easy to disrupt after removing the shear force. This is because of the driving force for the liquid crystalline solution to form the more thermodynamically stable chiral nematic structure.[82] Relaxation in a pseudo-nematic lyotropic

solution, after being oriented along the high shear rate direction, will persist for a long period of time, as the driving force to reform helicoidal structure may be limited. So far, there are a few reports that discuss the relationship between the relaxation behavior and chiro-optical properties of the mesophases. In the case of a lyotropic polypeptide system, poly(γ-benzyl-glutamate) (PBG) solutions containing a single optical isomer of either poly(γ-benzyl-l-glutamate) (PBLG) or poly(γ-benzyl-d-glutamate) (PBDG) or in the case of a racemic mixture of these two, there is no support for the idea that the chiro-optical properties of the chiral nematic PBG/*m*-cresol solution has any significant effects on the relaxation behavior.[52] This may be because of the fact that the mesophases of both optical isomers and the racemic mixture have a large pitch and are nematic-like.

In our recent study, relaxation behaviors upon shear force removal were investigated for lyotropic solutions of AEC in AA with different chiro-optical properties.[84] Ethylcellulose and AEC can form chiral nematic liquid crystalline solutions in many solvents and were found to change handedness as well as pitch with the change of degree of acetylation. In acrylic acid, EC forms a left-handed mesophase, while fully acetylated AEC forms a right-handed mesophase.[85] At a certain critical degree of acetylation (DA*) the pitch becomes infinite and an untwisted, nematic-like structure forms. When the degree of acetylation of AEC increases across DA*, the handedness of cholesteric liquid crystalline solutions changes from left-handedness to right-handedness. The pitch of the lyotropic solution of EC or fully acetylated AEC in AA is about one-tenth that of PBG solutions. Therefore, the effect of chiro-optical properties on relaxation behavior is expected to be more pronounced in AEC system than in PBG system.

The values of pitch for lyotropic AEC solutions in AA (50% w/w) are given in Table 1, where AEC-2 is a pure AEC with medium DA and AEC-3 is the mixture of EC and fully acetated AEC, which has the same average DA as AEC-2. However, in the liquid crystalline solutions, even at the same concentration, AEC-2 and AEC-3 have different pitch and handedness. This phenomenon was also observed in the lyotropic AEC/chloroform system.[86] The difference in chiro-optical properties may come from the complex interactions of multiple chiral centers present in each repeating unit of the cellulose chain, not from simple racemic mixtures as in the PBG system.

At the same concentration, the viscosities of AEC solutions are different such that the larger the pitch, the smaller the viscosity.[29] To eliminate the influence of viscosity of LCP solutions on the relaxation behaviors, the concentrations of AEC solutions were adjusted so that all solutions had similar viscosities under shear. As shown in Fig. 8, when the shear rate is between 0.5 and 6 s^{-1}, the four AEC solutions had the same viscosity.

Stress relaxation

Stress relaxation reflects fast relaxation at the molecular level and is independent of the previous shear rate. In the tumbling region (low pre-shear rate region), the final part of the stress relaxation curve is determined by the textural contribution because its time of evolution scales with the previous shear rate. This reflects the decrease in domain size with shear rate, which controls the scaling relations. In the flow align region, the textural contribution disappears and fast relaxation becomes nearly independent of shear rate.[87]

As shown in Fig. 9, the stress relaxation curves of all AEC/AA solutions collapse into one curve when the solutions were presheared with the same rate. Because the stress relaxation is at the molecular level and the chiro-optical properties reflect the suprastructural level, it is expected that the lyotropic solutions with different chiro-optical properties have the same stress relaxation behavior in both the tumbling and flow-align regions.

Dynamic modulus evolution upon cessation of flow

The development of storage and loss moduli after flow cessation is a useful tool to analyze structural relaxations on LCPs. Upon flow cessation, the flow-induced orientation is lost. The evolution of the moduli with time is because of the reformation of a chiral nematic phase that had become nematic under flow. In Fig. 10, the relaxation in complex modulus with time for various AEC samples is shown.

As seen in Fig. 10, the larger the pitch, the slower the rate the modulus evolution. The modulus of the right-handed mesophase (AEC-4) developed faster than that of the left-handed mesophase (AEC-1) with similar pitch. The difference in relaxation behavior of the AEC solutions may be because of the smaller driving force to reform helicoidal structures of chiral nematic phases from flow-induced nematic mesophases for the mesophases with larger pitch (nematic-like).

The reformation of chiral nematic phase from nematic phase sometimes goes through the transient state of band texture. The formation of band texture is also a relaxation phenomenon. For HPC/H_2O solutions, the band texture is only observed when the molecules have been well orientated in the shear direction.[76] A critical lower shear rate limit exists because of the stability of chiral nematic textures.

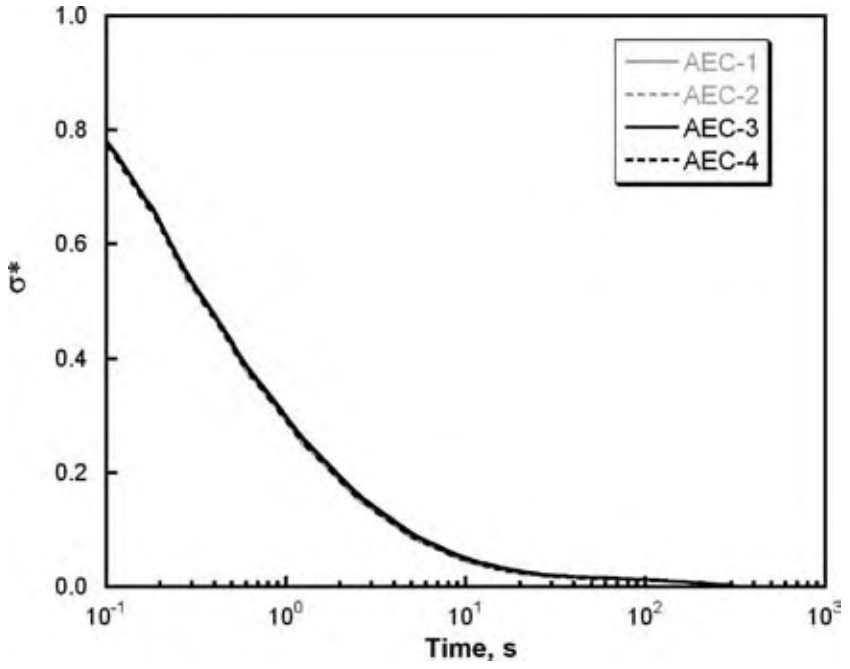


Fig. 9 Reduced shear stress $\sigma^*(\sigma^* = [\sigma(t) - \sigma_{fin}]/(\sigma_{in} - \sigma_{fin}))$ vs. time for all AEC/AA solutions.

An upper shear rate limit also exists above which no bands are formed.[88] The upper critical shear rate is associated with the flow-aligning region of molecular dynamics. Band texture formation is driven by the release of energy, which has been stored in the mesophase during shear.[76] The evolution of the band texture depends on the previous shear rate applied to mesophase. The rate of evolution of the band texture will increase, remain constant, or decrease according to whether the mesophase is sheared at low, intermediate, or high rates.[89] When the band texture appears during recoil, the appearance of the band texture stops the strain recovery process until the band texture disappears, and then recovery continues. Therefore, the presence of the band texture during recoil enhances the strain recovery.[90]

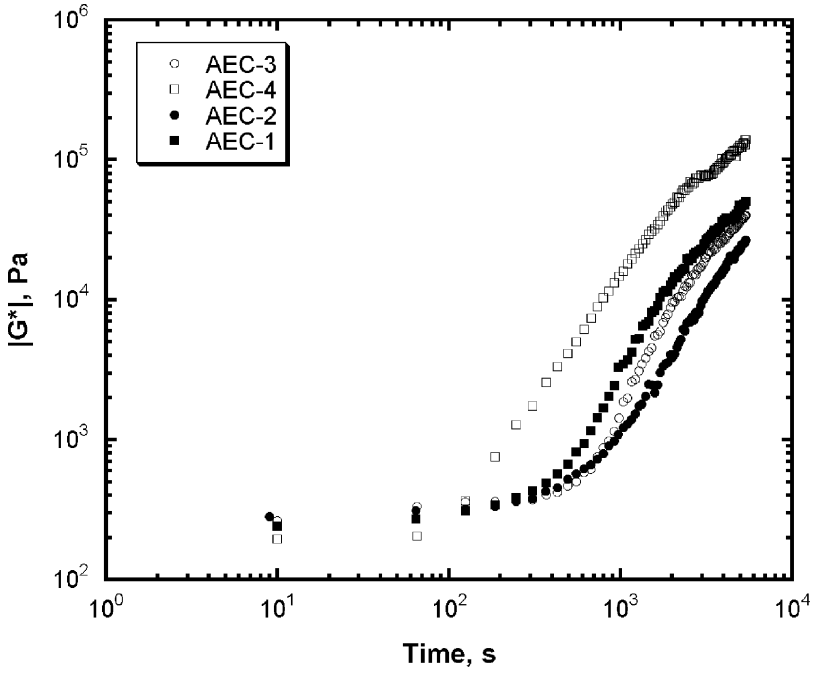


Fig. 10 Complex modulus evolution of AEC solutions vs. time.

CONCLUSIONS

Over the last three decades extensive information pertaining to the properties and behaviors of liquid crystalline cellulose derivatives has been obtained. However, the fact remains that fibers and films produced from cellulosic mesophases are inferior to those predicted by theory. A number of questions still remain surrounding these unique polymer systems. For example, what role, if any, do the pitch and handedness of the helicoidal structure play in determining the transient rheological behavior of lyotropic cellulosic mesophases. A correlation between the chiro-optical properties and relaxation behavior should be established. From our recent study, the relaxation behavior of lyotropic liquid crystalline solutions of (acetyl) (ethyl) cellulose was affected by the chiro-optical properties of the mesophase. In stress relaxation, handedness and pitch had no obvious effect. However, the development of complex modulus upon flow cessation was affected by pitch and handedness; the larger the pitch, the slower the rate the moduli evolved. Interestingly, however, the moduli of the right-handed mesophase developed faster than that of the left-handed one. Under flow, the cellulosic mesophases become nematic (shear aligned). Therefore, in samples with large pitch (nematic-like), the driving force to reform the helicoidal structures of chiral nematic phases upon removal of shear will be smaller than those with smaller pitch.

REFERENCES

1. Werbowyj, R.S.; Gray, D.G. Liquid crystalline structure in aqueous hydroxypropyl cellulose solutions. Molec. Crystals Liquid Crystals **1976**, *34* (4), 97–103.
2. Patton, P.; Gilbert, R.D. Anisotropic solutions of methylol cellulose. Polym. Preprints (American Chemical Society, Division of Polymer Chemistry) **1983**, *24* (2), 266–267.
3. Bhadani, S.N.; Gray, D.G. Cellulose-based liquid-crystalline polymers; esters of (hydroxypropyl) cellulose. Molec. Crystals Liquid Crystals **1983**, *99* (1–4), 29–38.
4. Laivins, G.V.; Gray, D.G. Liquid crystalline phase transition of a semiflexible polymer: acetoxypropyl cellulose. Macromolecules **1985**, *18* (9), 1753–1759.
5. Gilbert, R.D. Cellulose and cellulose derivatives as liquid crystals. ACS Symp. Ser. **1990**, *433*, 259–272.
6. Gray, D.G.; Harkness, B.R. Chiral nematic mesophase of lyotropic and thermotropic cellulose derivatives. In *Liquid Crystalline and Mesomorphic Polymers*; Shibev, V.P., Lam, L., Eds.; Springer-Verlag: New York, 1993; 298–323.
7. Sixou, P.; Bosch, A.T. Lyotropic liquid crystal solutions of cellulose derivatives. In *Cellulose: Structure, Modification and Hydrolysis*; Young, R.A., Rowell, R.M., Eds.; Wiley: New York, 1986; 205–219.
8. Chanzy, H.; Chaunis, S.; Monzie, P. Oriented cellulose films and fibers from a mesophase system. J. Polym. Sci. Polym. Phys. Ed. **1980**, *18* (5), 1137–1144.
9. Patel, D.L.; Gilbert, R.D. Mesomorphic solutions of cellulose triacetate in halogenated organic acids and mixtures of trifluoroacetic acid and dichloromethane. J. Polym. Sci. Polym. Phys. Ed. **1981**, *19* (9), 1449–1460.
10. Patel, D.L.; Gilbert, R.D. Lyotropic mesomorphic formation of cellulose in trifluoroacetic acid/chlorinated-alkane solvent mixtures at room temperature. J. Polym. Sci. Polym. Phys. Ed. **1981**, *19* (8), 1231–1236.
11. Chen, Y.S.; Cuculo, J.A. Lyotropic mesophase of cellulose in ammonia/ammonium thiocyanate solution. J. Polym. Sci. A: Polym. Chem. **1986**, *24* (9), 2075–2084.
12. Conio, G.; Corazza, P.; Bianchi, E.; Tealdi, A.; Ciferri, A. Phase equilibria of cellulose in *N,N*-dimethylacetamide/lithium chloride solutions. J. Polym. Sci. Polym. Lett. **1984**, *22* (5), 273–277.
13. Yang, K.S.; Theil, M.H.; Cuculo, J.A. Lyotropic mesophases of cellulose in the ammonia–ammonium thiocyanate solvent system. Effects of system composition on phase types. ACS Symp. Ser. **1989**, *384*, 156–183.
14. Yang, K.S.; Cuculo, J.A. Formation and characterization of the fibers and films from mesophase solutions of cellulose in ammonia/ammonium thiocyanate solvent. Polymer **1992**, *33* (1), 170–174.
15. Gilbert, R.D. Making strong cellulose fibers. Chemtech **1995**, *25* (11), 44–48.
16. Gilbert, R.D.; Hu, X.; Fornes, R.E. Preparation of high-strength/high-modulus regenerated cellulose fibers from lyotropic mesophases. J. Appl. Polym. Sci. **1995**, *58* (8), 1365–1370.
17. Bianchi, E.; Ciferri, A.; Conio, G.; Tealdi, A. Fiber formation from liquid-crystalline precursors. II. Cellulose in *N,N*-dimethylacetamide-lithium chloride. J. Polym. Sci. B: Polym. Phys. **1989**, *27* (7), 1477–1484.
18. O'Brien, J.P. Cellulosic Fibers from Anisotropic Solutions. US Patent 4,464,323, 1984.
19. O'Brien, J.P. High Strength Cellulosic Fibers. US Patent 4,501,886, 1986.
20. Gilbert, R.D. *Polymeric Materials Encyclopedia*; Salamone, J.C., Ed.; CRC Press: Boca Raton, 1996.

21. Guo, J.X.; Gray, D.G. Lyotropic cellulosics liquid crystals. In *Cellulose Polymers, Blends and Composites*; Gilbert, R.D., Ed.; Hanser: New York, 1994; 25–46.
22. Zugenmaier, P. Structural investigation on some cellulose derivatives in the crystalline and liquid crystalline state. In *Cellulose*: *Structure, Modification, and Hydrolysis*; Young, R.A., Rowell, R.M., Eds.; Wiley: New York, 1986; 221–245.
23. Suto, S.; Tateyama, S. Transient shear response of liquid crystal-forming hydroxypropyl cellulose solution in dimethylacetamide. I. Stress growth and relaxation behavior. J. Appl. Polym. Sci. **1994**, *53* (2), 161–168.
24. Suto, S.; Kohmoto, K.; Abe, A. Transient shear response of liquid crystal-forming hydroxypropyl cellulose solution in dimethylacetamide. II. Correlation between band formation and stress relaxation. J. Appl. Polym. Sci. **1994**, *53* (2), 169–178.
25. Nishio, Y.; Yamane, Y.; Takahashi, T. Morphological studies of liquid-crystalline cellulose derivatives. II. Hydroxypropyl cellulose films prepared from liquid-crystalline aqueous solutions. J. Polym. Sci. Polym. Phys. Ed. **1985**, *23* (5), 1053–1064.
26. Zugenmaier, P.; Haurand, P. Structural and rheological investigations on the lyotropic, liquid-crystalline system: *O*-ethylcellulose-glacial acetic acid-dichloroacetic acid. Carbohydr. Res. **1987**, *160*, 369–380.
27. Shimamoto, S.; Gray, D.G. A method to preserve the chiral nematic order of lyotropic ethylcellulose and (acetyl)(ethyl)cellulose mesophases in solid films. Chem. Mater. **1998**, *10* (6), 1720–1726.
28. Zhao, C.T.; Zhang, G.L.; Cai, B.L.; Xu, M. Solvent composition dependence of band morphology in sheared lyotropic ethyl cellulose liquid crystals. Macromol. Chem. Phys. **1998**, *199* (8), 1485–1488.
29. Guo, J.X.; Gray, D.G. Effect of degree of acetylation and solvent on the chiroptical properties of lyotropic (acetyl)(ethyl) cellulose solutions. J. Polym. Sci. B: Polym. Phys. **1994**, *32* (15), 2529–2537.
30. Goossens, W.J. Molecular theory of cholesteric phase and of twisting power of optically active molecules in a nematic liquid crystal. Molec. Crystals Liquid Crystals **1971**, *12* (3), 237.
31. Samulski, T.V.; Samulski, E.T. van der Waals–Lifshitz forces in lyotropic polypeptide liquid crystals. J. Chem. Phys. **1977**, *67* (2), 824–830.
32. Siekmeyer, M.; Steinmeier, H.; Zugenmaier, P. Supermolecular liquid-crystalline structures from highly concentrated cellulose derivative solutions. Angew. Makromol. Chem. **1989**, *166/167*, 131–138.
33. Siekmeyer, M.; Zugenmaier, P. Investigations of molar mass dependence of the lyotropic liquid-crystalline system: cellulose tricarbanilate/diethylene glycol monoethyl ether. Makromol. Chem. Rapid Commun. **1987**, *8* (10), 511–517.
34. Siekmeyer, M.; Zugenmaier, P. Solvent dependence of lyotropic liquid-crystalline phases of cellulose tricarbanilate. Makromol. Chem. **1990**, *191* (5), 1177–1196.
35. Vogt, U.; Zugenmaier, P. Structural models for some liquid crystalline cellulose derivatives. Berichte der Bunsen-Gesellschaft **1985**, *89* (11), 1217–1224.
36. Guo, J.X.; Gray, D.G. Preparation, characterization, and mesophase formation of esters of ethyl cellulose and methyl cellulose. J. Polym. Sci. A: Polym. Chem. **1994**, *32* (5), 889–896.
37. Harkness, B.R.; Gray, D.G. Left- and right-handed chiral nematic mesophase of (trityl)(alkyl)cellulose derivatives. Can. J. Chem. **1990**, *68* (7), 1135–1139.
38. Lin-Liu, Y.R.; Shih, Y.M.; Woo, C.W. Molecular theory of cholesteric liquid crystals and cholesteric mixtures. Phys. Rev. A: Atomic, Mol. Optical Phys. **1977**, *15* (6), 2550–2557.
39. Siekmeyer, M.; Steinmeier, H.; Zugenmaier, P. Structural investigations and phase behavior of a ternary lyotropic liquid-crystalline cellulosic system: cellulose tricarbanilate/3-chlorophenylurethane of cellulose/triethylene glycol monomethyl ether. Makromol. Chem. **1989**, *190* (5), 1037–1045.
40. Kimura, H.; Nakano, H. Orientational phase transition in the system of flexible molecules. J. Phys. Soc. Jpn. **1979**, *46* (6), 1695–1700.
41. Kimura, H.; Hosino, M.; Nakano, H. Statistical theory of cholesteric ordering in hard-rod fluids and liquid crystalline properties of polypeptide solutions. J. Phys. Soc. Jpn. **1982**, *51* (5), 1584–1590.
42. Osipov, M.A. Molecular theory of solvent effect on cholesteric ordering in lyotropic polypeptide liquid crystals. Chem. Phys. **1985**, *96* (2), 259–270.
43. Osipov, M.A. Theory for cholesteric ordering in lyotropic liquid crystals. Nuovo Cimento della Societa Italiana di Fisica, D. Condensed Matter, Atomic. Molec. Chem. Phys. Fluids, Plasmas, Biophys. **1988**, *10D* (11), 1249–1262.
44. Osipov, M.A. Molecular theory of cholesteric polymers. In *Liquid Crystalline and Mesophoric Polymers*; Shibev, V.P., Lam, L., Eds.; Springer-Verlag: New York, 1994; 1–25.

45. Chiba, R.; Nishio, Y.; Miyashita, Y. Electrooptical behavior of liquid-crystalline (hydroxypropyl)-cellulose/inorganic salt aqueous solutions. Macromolecules **2003**, *36* (5), 1706–1712.
46. Nishio, Y.; Chiba, R.; Miyashita, Y.; Oshima, K.; Miyajima, T.; Kimura, N.; Suzuki, H. Salt addition effects on mesophase structure and optical properties of aqueous hydroxypropyl cellulose solutions. Polym. J. **2002**, *34* (3), 149–157.
47. Shiau, C.C.; Labes, M.M. Control of the pitch of synthetic polypeptide lyotropic phases utilizing a chiral solvent. Molec. Crystals Liquid Crystals **1985**, *124* (1–4), 125–130.
48. Ericksen, J.L. Equilibrium theories of crystals. In *Advances in Liquid Crystals 2*; Brown, G.H., Ed.; Academic Press: London, 1976; 233–298.
49. Leslie, F.M. Theory of flow phenomena in liquid crystals. In *Advances in Liquid Crystals 4*; Brown, G.H., Ed.; Academic Press: London, 1979; 1–81.
50. Vermant, J.M.; Moldenaers, P.; Mewis, J. An evaluation of the Larson–Doi model for liquid crystalline polymers using recoil. Rheol. Acta **1999**, *38* (6), 537–547.
51. Doi, M.; Edwards, S.F. *The Theory of Polymer Dynamics*; Oxford University Press: London, 1986.
52. Burghardt, W.R. Molecular orientation and rheology in sheared lyotropic liquid crystalline polymers. Macromol. Chem. Phys. **1998**, *199* (4), 471–488.
53. Larson, R.G.; Doi, M. Mesoscopic domain theory for textured liquid-crystalline polymers. J. Rheol. **1991**, *35* (4), 539–563.
54. Rey, A.D.; Tsuji, T. Recent advances in theoretical liquid crystal rheology. Macromol. Theory Simul. **1998**, *7* (6), 623–639.
55. Larson, R.G. *Constitutive Equations for Polymer Melts and Solutions*; Boston: Butterworths, 1988.
56. Larson, R.G. *The Structure and Rheology of Complex Fluids*; Oxford University Press: New York, 1999.
57. Rey, A.D.; Denn, M.M. Dynamic phenomena in liquid-crystalline materials. Annu. Rev. Fluid Mech. **2002**, *34*, 233–266.
58. Marrucci, G.; Greco, F. Flow behavior of liquid crystalline polymers. Adv. Chem. Phys. **1993**, *86*, 331–303.
59. Rey, A.D. Generalized cholesteric permeation flows. Phys. Rev. E **2002**, *65* (2), No. 022701.
60. Rey, A.D. Theory of linear viscoelasticity of cholesteric liquid crystals. J. Rheol. **2000**, *44* (4), 855–869.
61. Rey, A.D. Theory of linear viscoelasticity of chiral liquid crystals. Rheol. Acta **1996**, *35* (5), 400–409.
62. Rey, A.D. Structural transformations and viscoelastic response of sheared fingerprint cholesteric textures. J. Non-Newtonian Fluid Mech. **1996**, *64* (2–3), 207–227.
63. Rey, A.D. Flow alignment in the helix uncoiling of sheared cholesteric liquid crystals. Phys. Rev. E **1996**, *53* (4), 4198–4201.
64. Rey, A.D. Simple shear and small amplitude oscillatory rectilinear shear permeation flows of cholesteric liquid crystals. J. Rheol. **2002**, *46* (1), 225–240.
65. Hongladarom, K.; Secakusuma, V.; Burghardt, W.R. Relation between molecular orientation and rheology in lyotropic hydroxypropyl cellulose solutions. J. Rheol. **1994**, *38* (5), 505–523.
66. Hongladarom, K.; Burghardt, W.R. Molecular alignment of polymer liquid crystals in shear flows. 2. Transient flow behavior in poly(benzyl glutamate) solutions. Macromolecules **1993**, *26* (4), 785–794.
67. Navard, P.; Haudin, J.M. Rheological behavior of isotropic and lyotropic solutions of cellulose. In *Cellulose: Structure, Modification, and Hydrolysis*; Young, R.A., Rowell, R.M., Eds.; Wiley: New York, 1985; 247–261.
68. Santamaria, A.; Lizaso, M.I.; Munoz, M.E. Rheology of ethyl cellulose solutions. Macromol. Symp. **1997**, *114* (Polymer–Solvent Complexes), 109–119.
69. Lizaso, I.; Munoz, M.E.; Santamaria, A. Transient rheological behavior of lyotropic solutions of ethyl cellulose in *m*-cresol. Rheol. Acta **1999**, *38* (2), 108–116.
70. Zugenmaier, P. Polymer solvent interaction in lyotropic liquid crystalline cellulose derivative systems. In *Cellulose Polymers, Blends and Composites*; Gilbert, R.D., Ed.; Hanser: New York, 1994; 71–94.
71. Onogi, S.; Asada, T. Rheology and rheooptics of polymer liquid crystals. In *Rheology, Vol. 1: Principles*; Astarita, G., Marrucci, G., Nicolais, L., Eds.; Plenum: New York, 1980; 127–147.
72. Walker, L.; Wagner, N. Rheology of region I flow in a lyotropic liquid-crystal polymer: the effects of defect texture. J. Rheol. **1994**, *38* (5), 1525–1547.
73. Metzner, A.B.; Prilutski, G.M. Rheological properties of polymeric liquid crystals. J. Rheol. **1986**, *30* (3), 661–691.
74. Ugaz, V.M.; Cinader, D.K.; Burghardt, W.R. Origins of region I shear thinning in model lyotropic liquid crystalline polymers. Macromolecules **1997**, *30* (5), 1527–1530.
75. Caputo, F.E.; Burghardt, W.R. Real-time 1-2 plane SAXS measurements of molecular orienta-

tion in sheared liquid crystalline polymers. Macromolecules **2001**, *34* (19), 6684–6694.
76. Ernst, B.; Navard, P. Band textures in mesomorphic (hydroxypropyl) cellulose solutions. Macromolecules **1989**, *22* (3), 1419–1422.
77. Grizzuti, N.; Cavella, S.; Cicarelli, P. Transient and steady-state rheology of liquid crystalline hydroxypropylcellulose solution. J. Rheol. **1990**, *34* (8), 1293–1310.
78. Baek, S.G.; Magda, J.J.; Larson, R.G.; Hudson, S.D. Rheological differences among liquid-crystalline polymers. II. Disappearance of negative N1 in densely packed lyotropes and thermotropes. J. Rheol. **1994**, *38* (5), 1473–1503.
79. Baek, S.G.; Magda, J.J.; Cementwala, S. Normal stress difference in liquid-crystalline hydroxypropylcellulose solutions. J. Rheol. **1993**, *37* (5), 935–945.
80. Mewis, J.; Moldenaers, P. Transient rheological behavior of a lyotropic polymeric liquid crystal. Molec. Crystals Liquid Crystals **1987**, *153* (A), 291–300.
81. Asada, T.; Toda, K.; Onogi, S. Deformation and structural re-formation of lyotropic cholesteric liquid crystal of hydroxylpropeycellulose + water system. Molec. Crystals Liquid Crystals **1981**, *68*, 231–246.
82. Grizzuti, N.; Moldenaers, P.; Mortier, M.; Mewis, J. On the time-dependency of the flow-induced dynamic moduli of a liquid-crystalline hydroxypropyl cellulose solution. Rheol. Acta **1993**, *32* (3), 218–226.
83. Godinho, M.H.; van der Klink, J.J.; Martins, A.F. Shear-history dependent ''equilibrium'' states of liquid-crystalline hydroxypropylcellulose solutions detected by rheo-nuclear magnetic resonance. J. Phys.: Condensed Matter **2003**, *15* (32), 5461–5468.
84. Dai, Q.G.; Gilbert, R.D.; Khan, S.A.; Kadla, J.F. Relaxation behaviors of lyotropic (acetyl)(ethyl) cellulose/acrylic acid solutions with different chiro-optical properties. Polym. Preprints (ACS, Div Polym. Chem.) **2004**, *45* (1), 824–825.
85. Shimamoto, S.; Uraki, Y.; Sano, Y. Optical properties and photopolymerization of liquid crystalline (acetyl) (ethyl) cellulose/acrylic acid system. Cellulose **2000**, *7* (4), 347–358.
86. Cowie, J.M.G.; Rodden, G.I. Blending as a method of tuning reflection wavelength and helical twisting sense in films and composites of liquid crystalline cellulose derivatives. Polymer **2002**, *43* (12), 3415–3419.
87. Mewis, J.M.; Vermant, J.; Moldenaers, P. Experimental evidence for the existence of a wagging regime in polymeric liquid crystals. Macromolecules **1997**, *30* (5), 1323–1328.
88. Vermant, J.M.; Mewis, J.; Picken, S.J. Band formation upon cessation of flow in liquid-crystalline polymers. J. Rheol. **1994**, *38* (5), 1571–1589.
89. Harrison, P.; Navard, P.; Cidade, M.T. Investigation of the band texture occurring in acetoxypropyl cellulose thermotropic liquid crystalline polymer using rheo-optical, rheological, and light scattering techniques. Rheol. Acta **1999**, *38* (6), 594–605.
90. Riti, J.B.; Navard, P. Textures during recoil of anisotropic hydroxypropyl cellulose solutions. J. Rheol. **1998**, *42* (2), 225–237.

Rotational Molding of Polymers

Céline T. Bellehumeur
Department of Chemical and Petroleum Engineering, University of Calgary, Calgary, Alberta, Canada

INTRODUCTION

Rotational molding, also known as rotomolding or rotocasting, is used to fabricate hollow plastic parts such as tanks used for chemical, agricultural, and automotive applications, toys, boat shells, and outdoor furniture. This process offers advantages over other processes in the fabrication of extremely large containers and of parts with intricate contours and free of weld lines. It is economically suited for short-run production and very large hollow articles, or for parts that have special constraints, which cannot be molded using other techniques. On the other hand, there are several disadvantages related to this process. The cycle time is long and typically ranges from 10 to 40 min, depending on the product size and molding conditions. The material cost is usually high due to the need for pulverization, and the choice of resins used in rotational molding is quite limited with polyolefins dominating the market. The process remains labor intensive, although some progress has been seen over the years in its automation. The rotational molding process has caught the attention of several major resin suppliers as well as mold and machine suppliers. The last few decades have also seen growing interest among researchers and designers, leading to significant technological advances through process modification, process control, part design, and the development of new materials.

In this entry, we focus on fundamental phenomena that control the quality of the end product from a processing and material perspective. The information will provide the reader with a better understanding of the process, which is necessary for the formulation of successful strategies in the development and processing of new resins for rotational molding applications. First, a detailed description of the rotational molding process is presented. This is followed by a discussion about particular features of the process, namely, melt densification and melt solidification phenomena. Recent technological advances in the development of resins new to rotational molding are also reviewed and presented in the last section of this entry.

ROTATIONAL MOLDING PROCESS

Process Description

The concept of rotational molding is illustrated in Fig. 1. The polymer is first loaded into a mold, which is then heated and rotated about its two primary axes (Figs. 1A and 1B). During the heating process, the tumbling powder gradually melts and sticks to the mold surface. Heating is continued after the powder has melted until complete densification is achieved. The mold is then cooled and the molded part is removed once it has reached a temperature safe for manual handling (Figs. 1C and 1D). Unlike most other polymer processes, rotational molding does not utilize pressure to force the melt into shape but relies primarily on the gradual deposition and adhesion of the polymer onto the mold. The mold rotation speed is relatively low (4–20 rpm) and the uniform coverage of the mold surface is ensured by the biaxial motion of the mold. The heating and cooling of the mold rely mostly on convective transport of energy. The most common method of heating is by means of gas combustion, while cooling relies on exposing the mold to forced air flow, water mist, water spray, or a combination of these methods.

In the rotational molding process, the resin is used mostly in powder form, the majority of which is produced through grinding between rotating metal plates. The quality of the powder depends on the pulverization process variables, such as the gap size between the plates, the temperature and blade conditions. Past studies have shown that the powder characteristics, namely, the shape, size, and size distribution, influence both the rotational molding cycle and the final product performance.[1,2] The quality of the powder affects heat transfer, the coverage of the mold cavity, as well as the initial size of bubbles during the melt deposition process. Simple screening procedures for the quality assessment of polymeric powder for rotational molding include particle size distribution, dry flow index, and bulk density. Typically, 35-mesh powder with a particle size distribution ranging from 100 to 500 μm is used in rotational molding as it provides a good balance between pulverization cost, powder flow characteristics, and powder packing

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007948

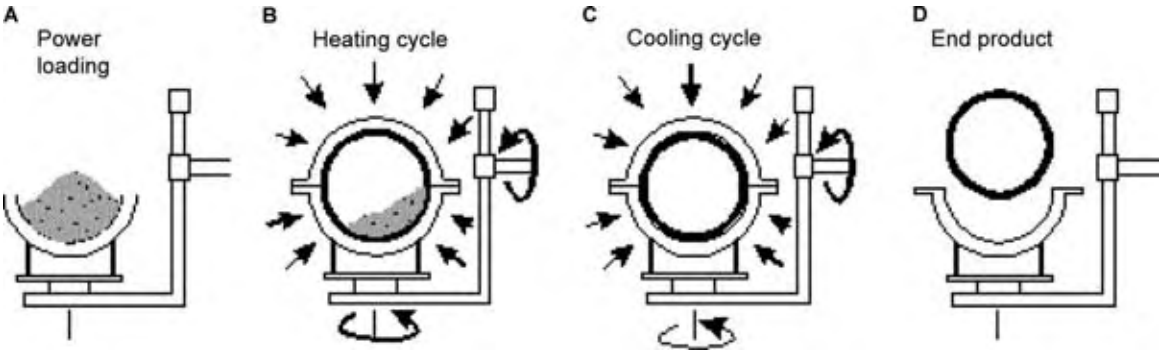


Fig. 1 The rotational molding process.

density. Both the flow index, which quantifies the ease of flow, and the bulk density are dependent on particle shape and size distribution.

Devices have been commercialized that allow for monitoring the temperature profile of the mold and the mold cavity during the molding process.[3] Such capability has led to important advances in the development of control protocols and process design. During the rotational molding cycle, all polymers exhibit a characteristic mold inside air temperature profile (see Fig. 2). Changes in the temperature profile are characteristic of key transitions in the processing of polymers. In Fig. 2, the region between points A and B represents the induction stage, where the solid polymer particles tumble freely in the mold and are heated primarily through conduction. Point B corresponds to the onset of powder adherence onto the mold surface and coalescence between individual particles, while point C marks the end of the melting phase. During the stage between B and C, semicrystalline polymers will undergo a melting transition that results in a reduction in the heat transfer to the mold inside air cavity and thus a reduction in the slope of the temperature profile. Such a change in the slope of the temperature profile is not as significant for amorphous polymers. The insulating effect of the deposited layer of polymer on the mold surface also causes a reduction in the heat transfer to the inner air. As the polymer particles adhere to the mold surface, pockets of air are trapped within the melt and form bubbles. The region from C to D corresponds to the fusion stage, where the bubbles undergo dissolution. Point D is often referred to as the peak internal air temperature (PIAT). The region between D and E characterizes the melt cooling stage. Point E indicates the residence time at which the melt undergoes recrystallization.

Molding Cycle Time

The rotational molding process is inherently transient, and the selection of optimal molding conditions will depend on several factors such as mold size, shape, and material, charge of the polymer, polymer thermal and rheological properties, convection conditions inside the oven, and cooling conditions. The measurement of the mold inside air temperature has been established as one of the means to control the molding cycle. Correlations between the PIAT and the mechanical properties of the molded part were first recognized by Crawford and Nugent.[4] Over the years, it has become a standard procedure to use the inside air temperature of the mold as an indicator in the determination of the time required for the completion of the heating cycle (melt densification) prior to the occurrence of polymer degradation upon exposure to high temperatures.

One important challenge in rotational molding is maximizing the mechanical properties of the end product while minimizing the molding cycle time. In many rotational molding applications, stiffness, dimensional stability, chemical resistance, and impact properties determine the performance of the final product. Processability encompasses the ability of a given resin to produce a final part with adequate properties when processed under a wide range of conditions (e.g., oven temperature, heating time, cooling conditions). The coalescence and melt solidification phenomena play a major role in both the molding cycle and the properties of the final parts. Resins with poor coalescence behavior require a longer heating time in the molding cycle, or the result will be the production of parts with large and numerous bubbles, which negatively impact the properties and appearance of the final product. The melt solidification behavior of the resins dictates the development of residual stresses and morphological features in the molded parts, which in turn contribute to the performance of the final products in commercial applications.

Over the years, guidelines that have to do with processing conditions, process control, powder quality, and material properties have been defined to optimize the process. A great deal of study has been carried out by Crawford's group, resulting in a number of

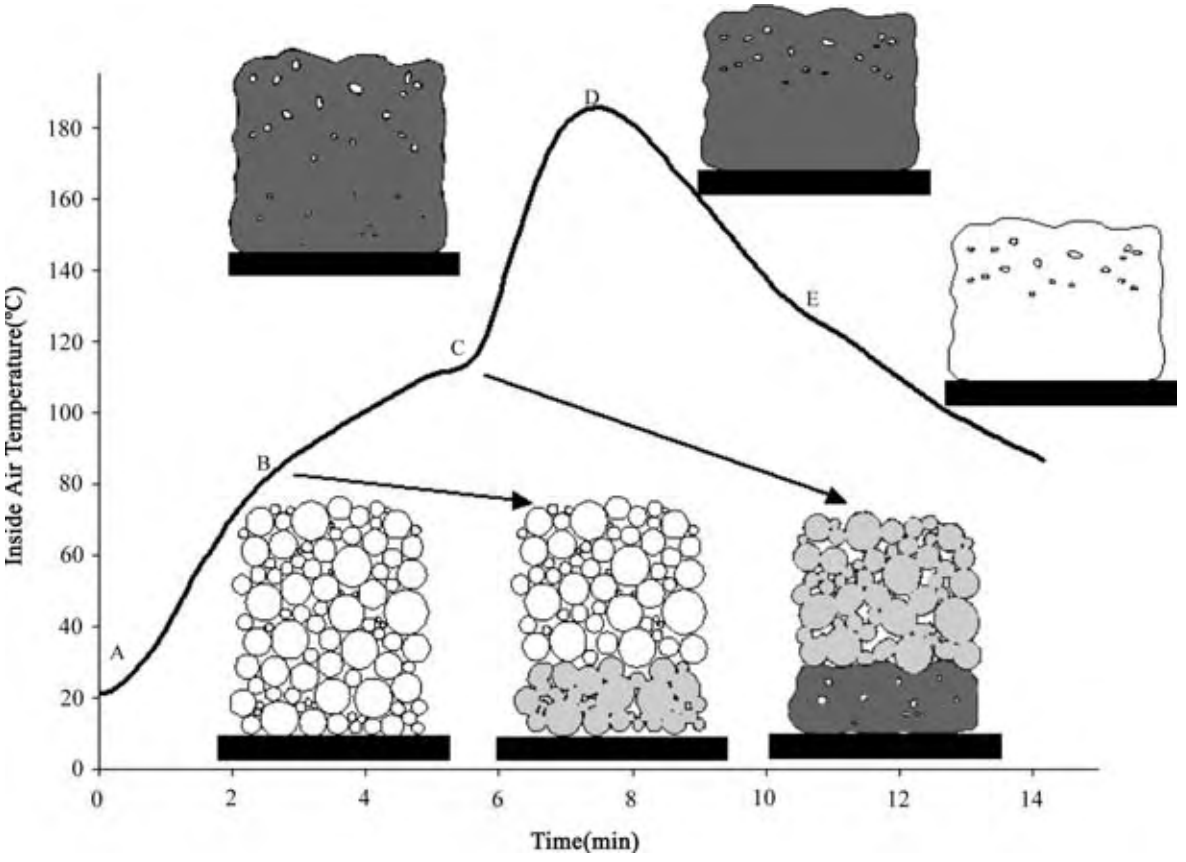


Fig. 2 Typical inside air temperature profile of mold.

patented technologies. The continuous monitoring of the inside air temperature as well as the inside pressure of the mold has been shown to provide useful information about the state of the material during the molding process and has been implemented into control strategies.[3,5] Strategies to reduce the molding time include preheating the powder and the mold, as well as using internal mold pressure in the late stage of the heating cycle for the removal of bubbles formed.[5] Along the same lines, mold pressurization during the cooling stage can significantly help reduce the cooling cycle time and also lead to improved dimensional stability of the molded product.[5,6] The identification of key material properties and their effect on the molding cycle and the performance of the final product are discussed in more detail in the following sections.

MELT DENSIFICATION

In rotational molding, two bulk movements dominate the melt deposition and densification: 1) the adhesion of particles on the mold surface and 2) the melting and collapse of the particulate structure.[7,8] The latter movement causes the entrapment of air into the melt and, thus, the formation of bubbles. Two distinct stages have been identified in the densification process: 1) particle coalescence and 2) bubble dissolution, as illustrated in Fig. 2.

Powder Coalescence

The coalescence of polymers is driven by the work of surface tension, which counteracts the viscous dissipation associated with the molecular diffusion within the coalescing domain. This phenomenon is often referred to in the literature as polymer sintering. In the rotational molding process, coalescence occurs at temperatures above that of the material melting point when dealing with semicrystalline polymers, or above the glass transition temperature for amorphous resins. The first analytical model describing the coalescence process was proposed by Frenkel:[9]

$$\frac{y}{a} = \left(\frac{\Gamma t}{\eta a}\right)^{1/2} \quad (1)$$

where y, a, Γ, t, and η are the neck radius formed between the particles, the radius of the particles,

material surface tension, time, and material viscosity, respectively. The model is based on the balance of the work of surface tension and the viscous dissipation. All other forces, including gravity, are neglected. The shape of the two spheres evolves, as shown in Fig. 3.

The validity of Frenkel's model is limited to Newtonian flow and can only be used to predict the early stage of the coalescence process, when the diameter of the two spherical particles remains nearly unchanged. The inadequacy of a Newtonian model in describing the coalescence of polymers was also demonstrated in other studies, as reviewed by Mazur, and has led to the development of models as well as alternative methods for the characterization of the coalescence behavior of polymers for rotational molding applications.[10] Based on theoretical and experimental analyses of the coalescence phenomenon, the material properties of primary interest in the evaluation of resin coalescence behavior in rotational molding have been identified as the resin viscosity, surface tension, and elasticity.

Most practitioners define the flow behavior of polymers based on the melt flow index; however, this property is not entirely relevant to the rotational molding process because it is essentially a shear-free and pressure-free process. The use of zero-shear viscosity has been proposed as a better way to assess the coalescence behavior of resins. Resins with lower zero-shear viscosity coalesce at a faster rate and can thus be processed using a shorter molding cycle.[11] The coalescence of individual powder particles is initiated as the particles stick and melt onto the mold surface or melt front. As the melt deposition process continues, pockets of air remain trapped between partially fused particles and lead to the formation of bubbles. In the rotational molding process, the coalescence of particles occurs at a temperature range close to the melting point of the material; thus, from a processing standpoint, low values of zero-shear viscosity at low temperatures (i.e., close to the temperature at which the particles adhere to the mold surface) are preferable.

In addition to zero-shear viscosity, material elasticity plays an important role in determining coalescence behavior in the rotational molding process.[10,12,13] For rotational molding grade material, this effect was first recognized from the experimental results obtained with impact modified polypropylenes and is illustrated in Fig. 4.[12] In this work, the relative elasticity of resins was evaluated based on stress relaxation measurements. The time required for the complete relaxation of the material was used as a measure of the melt elasticity. Resins with increased elasticity, but otherwise comparable zero-shear viscosities, produced parts with higher bubble content and, thus, lower density (see Fig. 4). In more recent studies, the relative elasticity of a resin has been evaluated from rheological oscillatory measurement, by comparing the ratio of the loss modulus over the elastic modulus (G''/G'). When comparing resins with similar molecular weight, low values of G''/G' at low frequencies are indicative of high elasticity, which translates into lower coalescence rates. The relative elasticity alone rarely explains poor coalescence behavior because resins with increased elasticity will often show variations in thermal and diffusion properties that are critical in the densification process. For instance, changes in the molecular structures leading to the broadening of the melting range can be detrimental to the coalescence process. Particles that soften at low temperatures and adhere to each other prior to a substantial movement of the bulk of the material cause the entrapment of large bubbles, as illustrated in the results presented in the literature.[8,14] Moreover, the presence of heterogeneities in the molecular formulation that cause the formation of segregated amorphous regions can seriously affect molecular interdiffusion and, thus, the coalescence process.

Molecular structure has an evident influence on a resin's rheological characteristics and expected coalescence behavior. Polymers with higher molecular weight have a slower coalescing rate because viscosity increases with molecular weight. Narrow molecular weight distribution results in lower elasticity and lower viscosity at low deformation rates, which is beneficial to the coalescence process. The comonomer content in polyethylene copolymers has mixed effects on the coalescence process. The increase of comonomer content generally results in a reduction in the melting temperature and the heat of fusion, which lead to earlier melting and onset of coalescence in transient molding processes. However, the incorporation of short-chain branches on linear polymer chains causes a reduction in the coalescing rate, as illustrated in Fig. 5. The effect of short-chain branches content on the coalescence rate may originate with the differences in the chain mobility due to variations of the chain linearity. The more linear the chains, the faster they are expected to diffuse.

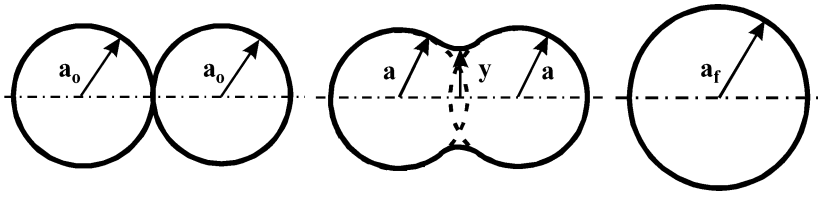


Fig. 3 Shape evolution of two coalescing spheres.

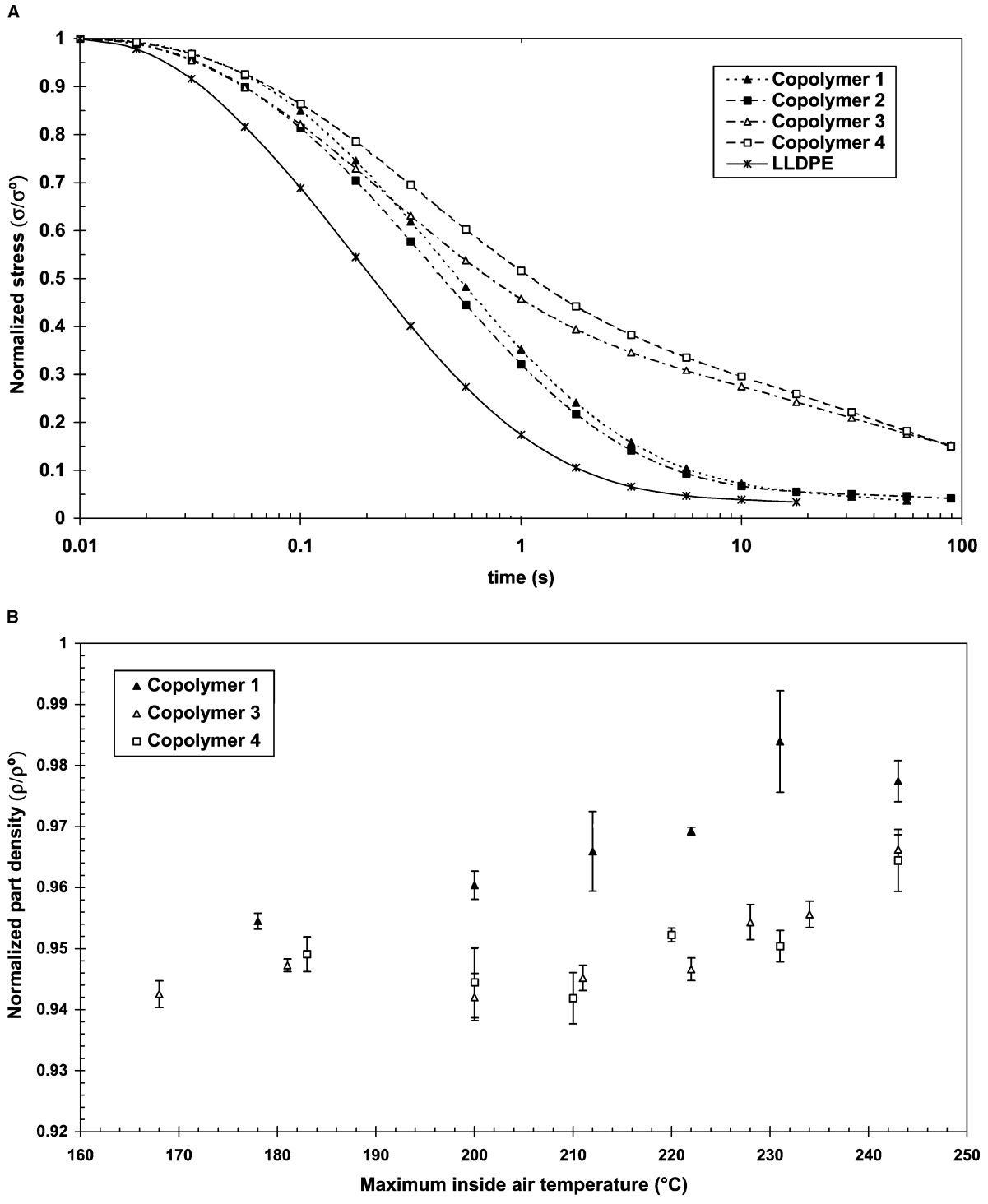


Fig. 4 (A) Stress relaxation curve of rotational molding grade ethylene copolymers. (B) Part density, which is indicative of the bubble content, measured after interrupting the rotational molding heating cycle at different heating times. (From Ref.[12].)

Variations in chemical composition and molecular structure can also have important repercussions on the surface energy of the material. Several techniques have been proposed for the experimental determination of surface tension, the sessile and pendant drop methods being the most promising and commonly used techniques. The common limitation of these techniques is that the melt surface energy is not directly measured; thus, the validity of the results depends on the underlining assumptions of the models used to fit the experimental observations. While surface tension has long been recognized as a controlling parameter in polymer

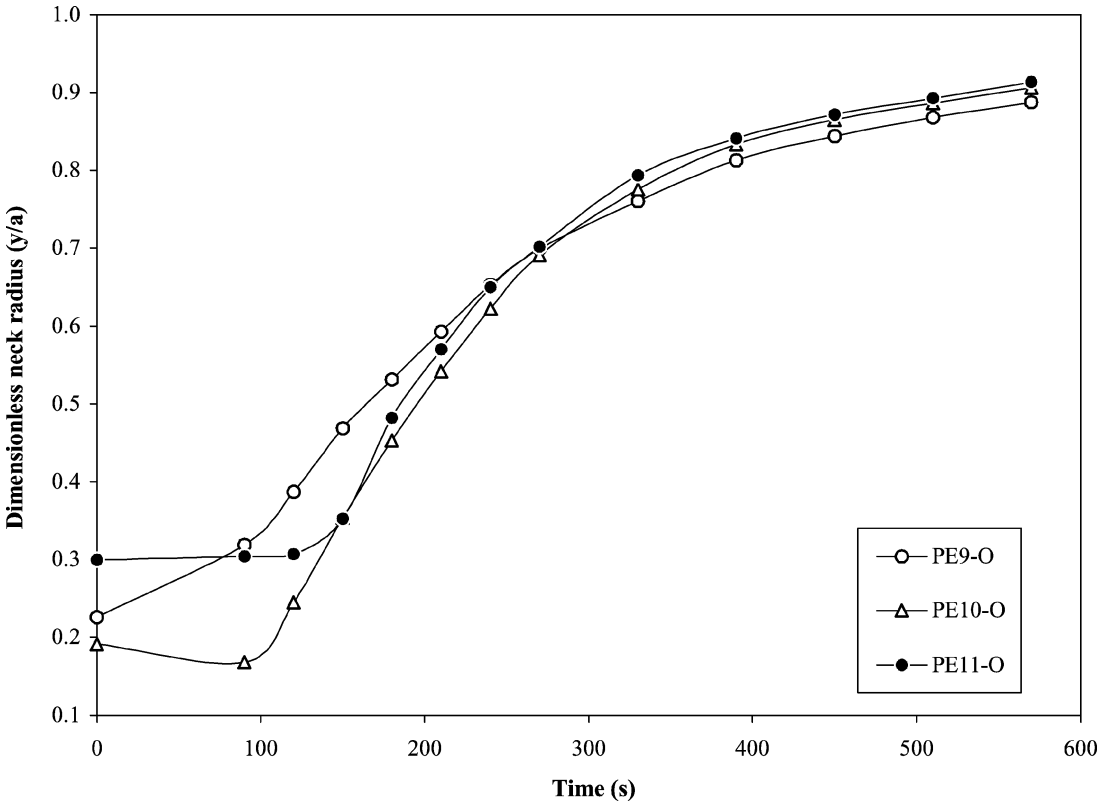


Fig. 5 Effect of comonomer content on the coalescence behavior of polyethylene copolymers at ramped temperature (111–226.5°C at 11°C/min). The relative methyl content for PE9-O, PE10-O, and PE11-O is 0.95, 0.56, and 0.15, respectively. (From Ref.[15].)

coalescence, limited work has been done in determining its effect on the processability of the resin in the rotational molding process.[16] The primary reason is that the measurement of the interfacial energy of polymer melts is challenging, owing to the viscous nature of the material and the potential for thermal degradation of the material upon exposure to high temperature and the long time scale of the experiments that have been defined for this type of measurement.

Dissolution of Bubbles

The level of coalescence between particles, the size of the particles, and the packing arrangement dictate the size of air cavities and, thus, the size of the bubble initially formed in the melt. Once formed, the bubbles remain stationary in the melt.[7] A relatively small bubble diameter, combined with the high viscosity of the melt, prevents the movement of the bubbles into the melt. The bubble removal is known to be a diffusion-controlled process. The identification of key parameters in the dissolution of bubbles formed in the melt has been done using a theoretical model that describes this process. The disappearance of the air bubble formed into the melt was modeled based on the analytical solution presented by Gogos:[17]

$$\left(R_0^2 - R^2\right) + \alpha_1(R_0 - R) + \alpha_1\alpha_2 \ln\left[\frac{R + \alpha_2}{R_0 + \alpha_2}\right] = 2D\frac{c_{s,P_\infty} - c_\infty}{\rho_\infty}t \tag{2}$$

where R_0 and R are the initial bubble radius and the bubble radius for the time interval t, respectively, and D is the diffusion coefficient of the gas into the melt (see Fig. 6).

The terms α_1 and α_2 in Eq. (2) are defined as follows:

$$\alpha_1 = \frac{4\Gamma}{R_g T}\left[\frac{2}{3} - \frac{c_{s,P\infty}}{c_{s,P\infty} - c_\infty}\right] \tag{3}$$

$$\alpha_2 = \frac{2\Gamma}{R_g T}\frac{c_{s,P\infty}}{\rho_\infty(c_{s,P\infty} - c_\infty)} \tag{4}$$

with R_g and T being the ideal gas law constant and the temperature of the melt–bubble system, respectively. The terms ρ_b and ρ_∞ symbolize the density of gas at the bubble/polymer melt interface for nonzero and zero surface tensions, respectively. The terms c_∞ and $c_{s,P\infty}$, which symbolize the dissolved gas concentration

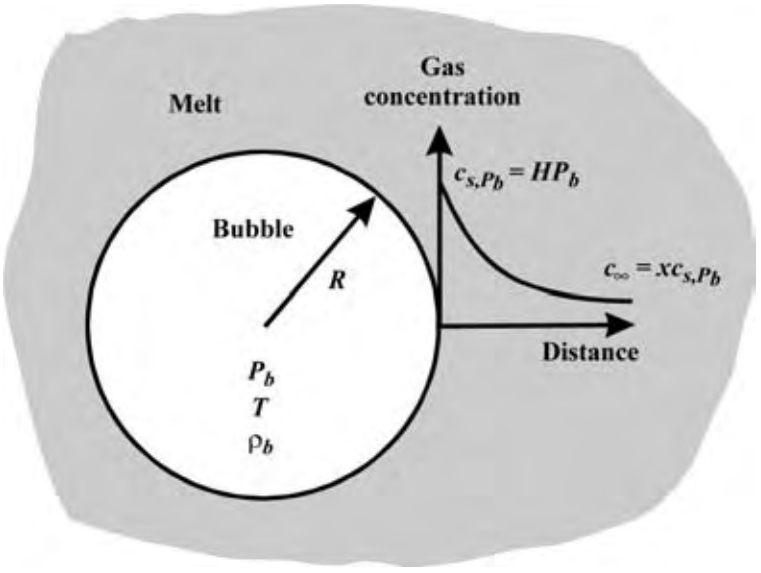


Fig. 6 Schematic of gas bubble dissolving into polymer melt with C_{s,P_b} being the dissolved gas concentration at the bubble/melt interface, c_∞ the dissolved gas concentration in the melt when partially saturated, x the degree of saturation in the melt, H Henry's law constant, P_b the bubble pressure, R the bubble radius, T the temperature, and ρ_b the density of the gas in the bubble.

in the melt when partially saturated and saturated for zero surface tension, respectively, are determined as follows:

$$c_{s,P_b} - c_{s,P\infty} = \frac{2\Gamma}{R_g T_P} \frac{c_{s,P\infty}}{R\rho_\infty} \quad (5)$$

$$c_\infty = xc_{s,P_b} \quad (6)$$

The dissolved gas concentration at the bubble/melt interface, c_{s,P_b}, can be related to the bubble pressure P_b through Henry's law. Gogos compared his model's predictions with the experimental data produced by Spence.[18] The model predictions fit very well with experimental data when selecting a degree of saturation close to 100% (91.9–99.6%). An alternative approach has been proposed by Kontopoulou and Vlachopoulos, who modeled the dynamics of bubble dissolution into the melt using conservation of mass and momentum.[19]

Factors that can affect the rate of bubble dissolution are the initial size of the bubbles, the temperature, the melt diffusion characteristics, the surface tension, and the solubility characteristics. The rate of consumption of oxygen due to degradation reactions probably contributes to the disappearance of bubbles, but this phenomenon has not yet been considered in the analysis of the process. It has been shown that within the range of viscosity typical to polyethylene rotational molding grade resins, the effect of viscosity on the dissolution of bubbles is marginal.[19] The level of air saturation in the polymer melt, however, was found to be crucial. While it cannot be determined accurately in the process, small variations in the level of saturation were found to cause large changes in the dynamics of the bubble dissolution.[17,19] Interfacial tension was shown to be important under conditions where the melt saturation level is close to unity.[17] The model proposed by Gogos was also useful in providing an explanation for the effect of mold pressurization on the bubble dissolution process. A sudden increase in the melt pressure causes a sudden shrinkage of the bubble. It also affects the level of melt saturation and causes an instep concentration gradient at the bubble/melt interface (see Fig. 6), leading to an increased rate of dissolution and faster removal of bubbles, as illustrated in Fig. 7. Experimental and theoretical results clearly showed that a small increase in the mold cavity pressure can lead to a significant reduction in the heating time required for complete melt densification in the rotational molding process.

Nonisothermal Densification

The densification process in rotational molding has been studied from both fundamental and practical perspectives. The models presented in the previous sections have furthered the understanding of the densification process in rotational molding; however, their use has been limited to the prediction of the densification process carried out under isothermal conditions. It is well known that heat transfer, powder coalescence, bubble formation, and bubble dissolution are collectively important in rotational molding; however, very few studies have addressed all aspects in modeling the densification process in rotational molding.

Heat transfer in the rotational molding process was first modeled by Rao and Throne.[20] Since this initial attempt at modeling the heating cycle, several studies

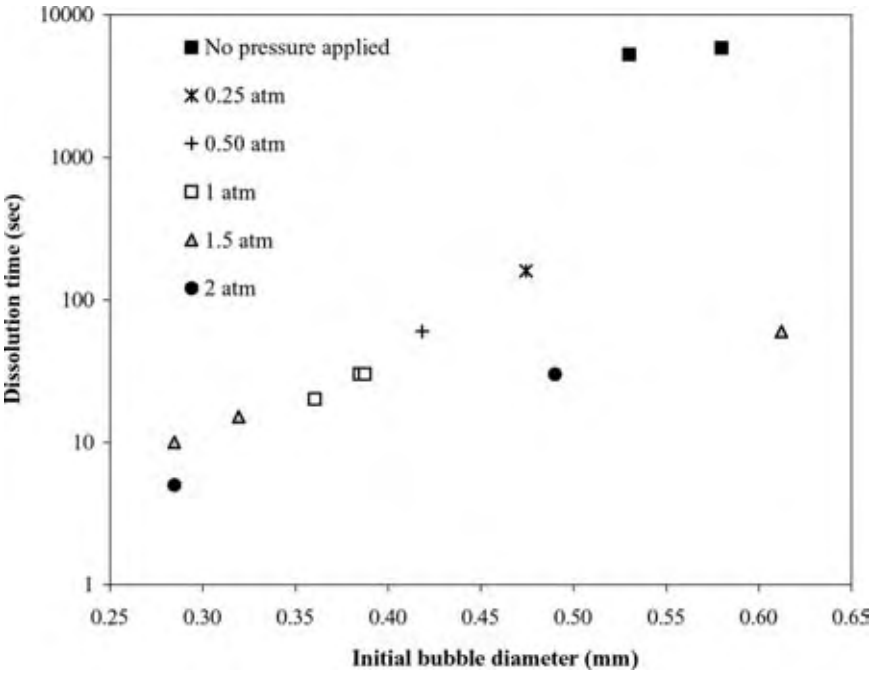


Fig. 7 Effect of mold cavity pressurization on the bubble dissolution time, based on experimental data presented by Spence. (From Ref.[18].)

have been undertaken to improve the model with consideration of different powder flow patterns, multidirectional heat transfer, complex mold geometry, internal heating, and the temperature dependence of the material properties. The thermal resistance within the mold wall was usually neglected. Moreover, it was generally assumed that the powder remains well mixed under conditions typically used in the rotational molding process. This assumption greatly simplified the modeling of the process, and model predictions have been found to be in good agreement with experimental results published in the literature.[21] However, the powder flow pattern may be better modeled using an avalanche or rolling type of flow, based on experimental evidence of the development of segregation patterns.[22,23] In most models presented in the literature, the mold curvature was neglected and only unidirectional heat transfer was considered. However, the trend toward the fabrication of more technically challenging parts with tighter specifications has provided some impetus to develop models that consider multidirectional heat transfer.[24] Moreover, the consideration of the kinematic of the mold rotation has been an important milestone in the development of commercial simulations that are used as design tools in the selection of rotation speed and heating conditions to ensure uniform melt deposition onto the mold surface for given mold geometry and polymer thermal properties.

The combined effect of heat transfer and sintering on the nonisothermal densification of polymer powder in rotational molding has been examined by Bellehumeur and Tiang.[25] For modeling purposes, the heating cycle was divided into three regimes: 1) heating of the mold and its contents until the mold temperature reaches the polymer melting point; 2) the polymer powder gradually melts and deposits on the mold surface; and 3) starts once all the powder has disappeared. In modeling the early stage of melt densification, the powder deposition process was treated as occurring in a layer-by-layer manner. The coalescence time allocated for each layer was determined by the rate of melt deposition. The initial size of the bubbles formed in the melt was determined by the neck growth achieved between particles, the size of the particles, and their packing arrangement. Bellehumeur and Tiang examined the relative importance of the powder characteristics and the material rheological properties on the melt densification process.[25] Their results showed that the initial size of the bubbles formed into the melt is primarily controlled by the powder particle size and packing arrangement (Fig. 8). This observation highlights the importance of powder quality on the molding cycle and the quality of the final product.

The time required by the melt deposition process is roughly proportional to the square of the molded part thickness and is significantly faster for polymers with lower melting points. A decrease in the heat of fusion also accelerates the deposition process, though not as much as a decrease in the melting temperature. A consequence of a fast melt deposition is that larger bubbles are formed in the melt, making them more difficult to remove during the heating cycle and negatively affecting the densification process. This problem can be surmounted with a reduction in the oven temperature; however, this reduction results in an increase in the total molding cycle time.

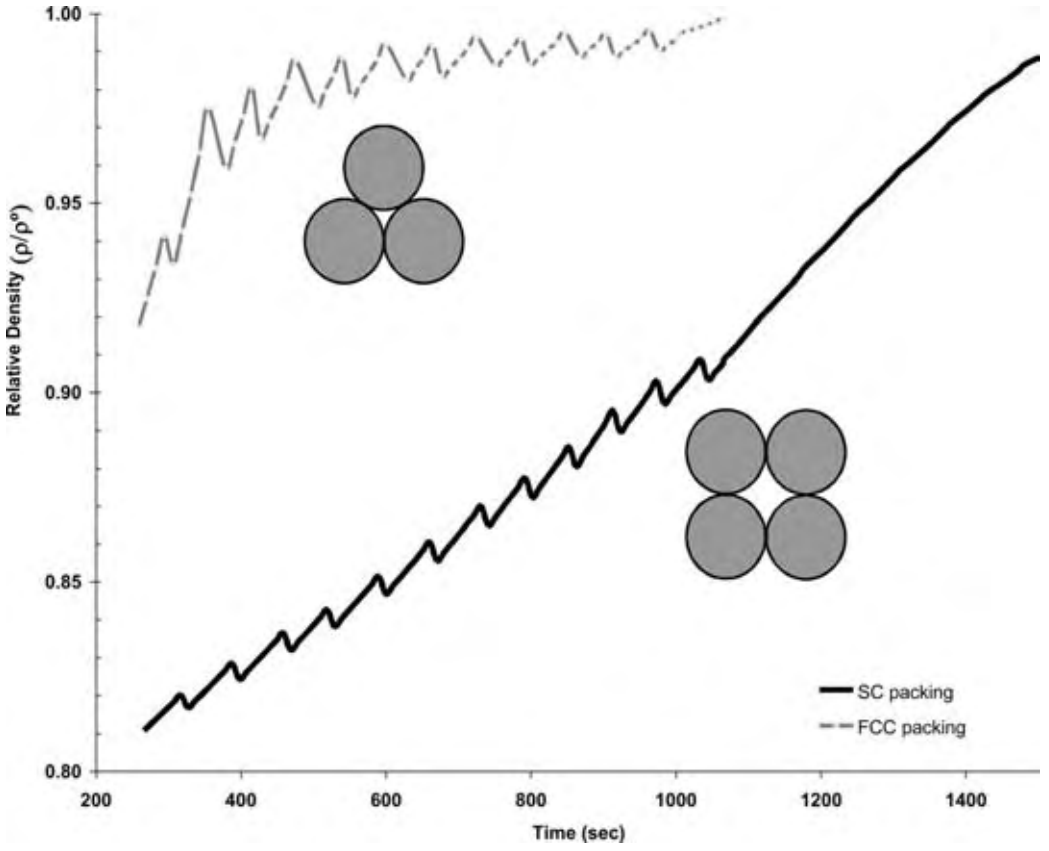


Fig. 8 Effect of the particle packing arrangement on the relative density of a rotational molding grade polyethylene with heating time. (From Ref.[25].)

MELT SOLIDIFICATION

The solidification of the polymer melt in rotational molding is relatively slow, in comparison to other processes, and is estimated to be in the range of 10–30°C/min. Moreover, the melt solidification is gradual and nonuniform across the molded part thickness, leading to important variations in the morphological features, as illustrated in Fig. 9, and dictating the properties and overall performance of the final product. The effects are more dramatic for resins with slower crystallization rates, such as polypropylene, compared to that observed with polyethylene.

Many thermoplastics used in rotational molding applications are semicrystalline in nature. The temperature and rate of crystallization vary with the material composition and molecular structure. Polymers that tend to crystallize have flexible backbones, regularly ordered atomic structure, and either small or no pendant groups. Crystallization occurring at higher temperatures is associated with a lower nucleation rate and the formation of coarse morphological structures. This usually results in the formation of structures that have a higher degree of stability and perfection but a lower failure strain, because the increase in the size of the spherulitic structure (common in polyethylene and polypropylene) is accompanied by a reduction in the interspherulitic boundary links. This, in turn, allows for the easier transmission of energy through the material and thus causes a loss in the ductility. This problem can be alleviated by using a faster cooling rate, which, on the other hand, leads to the generation of residual stresses in the parts. Because the material in contact with the mold is cooled at a faster rate than that near the inner surface, an asymmetrical residual stress profile is generated across the molded part thickness, inducing a bending moment of the part. If these induced residual stresses are high enough to overcome the structural integrity of the part, the molded part will deform and warp. Other factors that affect the development of residual stresses and the dimensional stability of the molded part include the material density, molded part thickness, mold material and thickness, rotation speed, and the application of mold release. The presence of residual stresses is associated not only with dimensional stability but also with the overall integrity of the molded part and is known to most severely affect the environmental stress cracking resistance of the product.

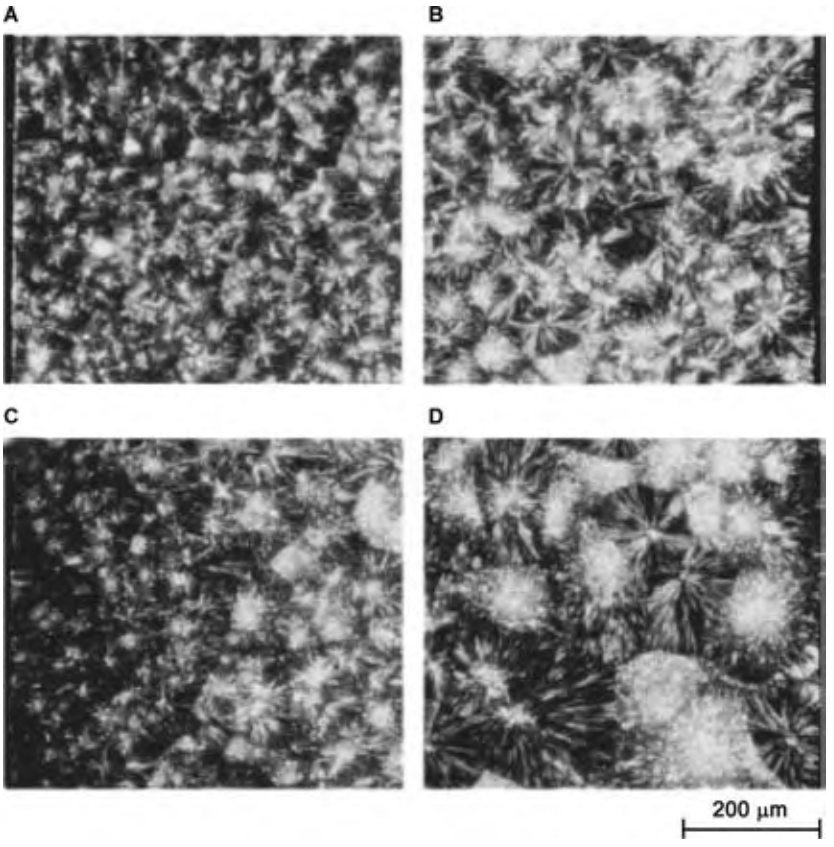


Fig. 9 Microphotograph of rotationally molded polypropylene parts subjected to water spray cooling (A, mold surface and B, inner surface) and parts initially subjected to water spray and subsequently air flow in the range of temperatures where crystallization occurs (C, mold surface and D, inner surface).[26]

In the rotational molding process, the polymer is subjected to high temperatures for relatively long periods of time to allow for melt deposition and full densification of the powder particles. As a consequence, thermo-oxidative degradation reactions can be severe if the material is not adequately stabilized. Degradation can often be determined by visual inspection of the molded part (change in color). Subtle changes can be detected through polarized or fluorescence microscopy (illustrated in Fig. 10) or infrared spectroscopy for the detection of carbonyl groups. The rheological characterization of specimens collected from the molded part can also provide useful information for degradation reactions can cause either chain scission or cross-linking reaction within the materials. Signs of degradation are usually first seen on the inner surface of the molded part for it is exposed to air during the entire molding process. Degradation reactions can be minimized with the use of an inert atmosphere in the mold. Excessive degradation of the material due to long exposure at high temperatures leads to the deterioration of the morphology (spherulitic structure) and the reduction in the mechanical properties of molded parts.[26,27]

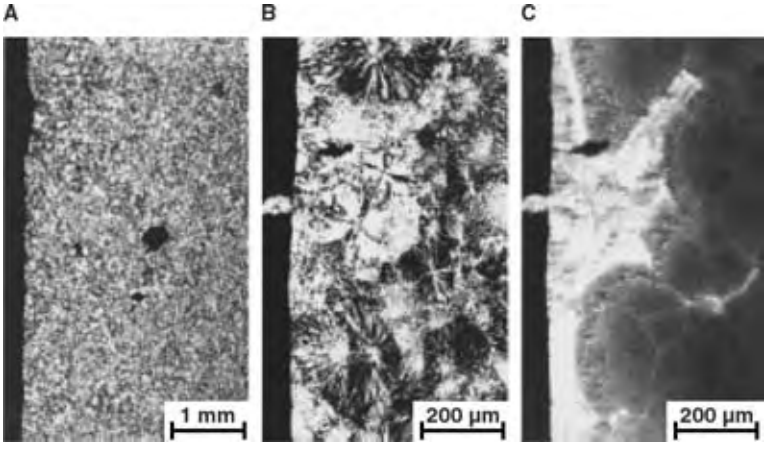


Fig. 10 Microstructure of rotationally molded polypropylene samples: (A) undercured specimen viewed under polarized light microscopy; (B) overcured specimen viewed under polarized light microscopy; and (C) overcured specimen viewed under fluorescence microscopy. In these pictures, the degraded layer shows higher birefringence. (From Ref.[26].)

Despite the importance of the melt solidification phenomena in rotational molding, only a few studies have been conducted with respect to modeling the cooling stage of the rotational molding process. One of the challenges resides in the nonlinearity of the model, which arises due to the unknown position of the moving solid/melt interface, and which is very similar to that encountered in modeling the heating stage of the rotational molding process. Throne pioneered this topic and investigated the various factors influencing the cooling rates.[28] Further work has been conducted to model this stage of the process with consideration of shrinkage, which can also greatly affect the molding cycle time.[24] Methodologies have also been proposed for the development of morphological features during processing based on half-time crystallization data, determined from differential scanning calorimetry, and spherulitic growth rate, which was obtained from optical microscopy experiments.[29] These are important steps toward the establishment of material–structure–properties–processing relationships for the rotational molding process.

TECHNOLOGICAL ADVANCES IN THE DEVELOPMENT OF NEW MATERIALS

Polyethylenes account for over 80% of all rotational molding production. They dominate the rotational molding market because of their thermal stability, availability in powder form, and relatively low cost. Over the last decade, polyolefins based on metallocene and single-site catalyst technologies have found their way into the rotational molding market. These technologies allow for better control of the molecular structure, which can have important repercussions on the processing behavior of the material as well as on the physical, chemical, and mechanical performance of the molded parts. Early results obtained with metallocene polyolefins suggested that these resins could be processed using a shorter heating time in rotational molding. In most studies, the reduction in the heating cycle was attributed to differences in the thermal properties (melting point and heat of fusion) between metallocene and Ziegler–Natta resins. However, warpage was commonly reported as a problem associated with using metallocene-based polyethylene. A more recent study showed that, with careful control of the molecular structure and the introduction of a certain level of heterogeneities, a significant reduction in the molding cycle time (up to 30%) could be achieved while maintaining adequate physical and mechanical properties.[30]

One important limitation in expanding the markets for rotational molding applications is the small range of resins suitable for the process. One cause for this limited choice of resins used commercially resides in poor processability behavior and the cost of pulverization, which is particularly important for resins that require cryogenic conditions. Alternatives to the use of powder particles in rotational molding have been proposed, thus eliminating the need for pulverization. The underwater pelletizing technology can be used to produce small cylindrical particles called micropellets, which are being used increasingly for rotational molding applications. It is well established that micropellets have better flow properties than powder particles. This characteristic has a significant impact on the thickness uniformity of molded parts and mold filling for complex geometries. However, one disadvantage of using micropellets is their narrow particle size distribution and their larger particle size. The packing of micropellets during the melt deposition process is such that larger bubbles are formed, compared to that seen when using a 35-mesh powder. These larger bubbles require a longer exposure to high temperatures for their complete dissolution.

Recent years have seen increased interest in the development of new rotational molding grade resins, which include polypropylene, acrylonitrile butadiene styrene, acetals, polyamides, thermoplastic foams, polyolefin blends, polyolefin plastomers, thermotropic liquid crystal polymers and nanocomposites. Two major concerns exist for the development of new resins: 1) obtaining adequate product performance and 2) achieving adequate processability. The most important roadblock in the development of many new resins, such as styrenic copolymers and polypropylene, resides in the fact that good impact properties can only be obtained with the incorporation of an impact modifier, often taking the form of a copolymer. Unfortunately, such a change in the material formulation often results in an increase in the material's relative elasticity, which is detrimental to the densification process and results in the production of parts with high bubble content and poor surface finish. Similarly, while the successful incorporation of nanoparticles in a polymer matrix can lead to the improvements of properties such as barrier resistance, it can also result in a dramatic increase in the linear viscoelastic properties. This proves to be an important drawback because increased zero-shear viscosity and elasticity are detrimental to the processability of the material. The traditional approach to solving this problem has consisted in varying the parameters known to affect properties that are key to the coalescence process (i.e. viscosity, elasticity, surface tension) with the incorporation of lubricants and other additives.[31] Promising results were also obtained for processing polyolefin plastomers with the uniform incorporation of the copolymer using metallocene catalyst technologies.[14] Alternative approaches have also included the careful selection of processing conditions that favor the completion of the coalescence

process between individual particles prior to the formation of bubbles into the melt.[32]

CONCLUSIONS

The rotational molding industry has seen a steady growth in the last two decades with the development of new process control strategies, machinery, and, to some extent, automatization. Yet, the development of new rotational molding applications is restricted for the processing cycle times are long and the choice of resins that can economically be used in this process is shockingly limited. The development of new resins is possible only if adequate product performance can be obtained while maintaining adequate processability. This can only be achieved with a good understanding of the phenomena that govern the quality of the end product. From a material formulation perspective, the key aspects to consider are the melt coalescence and melt solidification. While there has been significant progress in the development of methods for the rapid assessment of the processability and performance of resins, further work is required in defining strategies to overcome many of the process limitations.

REFERENCES

1. Throne, J.L.; Sohn, M.-S. Characterization of rotational molding grade polyethylene powders. Adv. Polym. Technol. **1989**, *9* (3), 181–192.
2. McDaid, J.; Crawford, R.J. The grinding of polyethylene powders for use in rotational molding. Proceedings of the Society of Plastics Engineers Annual Technical Conference, Atlanta, GA, Apr 26–30, 1998; The Society of Plastics Engineers: Brookfield, 1998.
3. Crawford, R.J.; Nugent, P.J. A new process control system for rotational moulding. Plast. Rubber Comp. Process. Appl. **1992**, *17* (1), 23–31.
4. Crawford, R.J.; Nugent, P.J. Impact strength of rotationally moulded polyethylene articles. Plast. Rubber Comp. Process. Appl. **1992**, *17* (1), 33–41.
5. Crawford, R.; Cramez, M.C.; Oliveira, J.J.; Spence, A. The importance of monitoring mold pressure during rotational molding. Proceedings of the Society of Plastics Engineers Annual Technical Conference, San Francisco, CA, May 5–9, 2002; The Society of Plastics Engineers: Brookfield, 2002.
6. Chen, C.H.; White, J.L.; Ohta, Y. Mold pressurization as a method to reduce warpage in rotational molding of polyethylene. Polym. Eng. Sci. **1990**, *30* (23), 1523–1528.
7. Crawford, R.J.; Scott, J.A. An experimental study of heat transfer during rotational moulding of plastics. Plast. Rubber Process. Appl. **1985**, *5* (3), 239–248.
8. Kontopoulou, M.; Vlachopoulos, J. Melting and densification of thermoplastic powders. Polym. Eng. Sci. **2001**, *41*, 155–169.
9. Frenkel, J. Viscous flow of crystalline bodies under the action of surface tension. J. Phys. **1945**, *9* (5), 385–391.
10. Mazur, S. Coalescence of polymer particles. In *Polymer Powder Technology*, 1st Ed.; German, R.M., Messing, G.L., Cornwall, R.G., Eds.; John Wiley & Sons: New York, 1995; 157–216.
11. Bellehumeur, C.T.; Bisaria, M.K.; Vlachopoulos, J. An experimental study and model assessment of polymer sintering. Polym. Eng. Sci. **1996**, *36* (17), 2198–2207.
12. Kontopoulou, M.; Bisaria, M.; Vlachopoulos, J. An experimental study of rotational molding of polypropylene/polyethylene copolymer. Int. Polym. Process. **1997**, *12*, 165–173.
13. Bellehumeur, C.T.; Kontopoulou, M.; Vlanchopoulos, J. The role of viscoelasticity in polymer sintering. Rheol. Acta **1998**, *37* (3), 270–278.
14. Wang, W.Q.; Kontopoulou, M. Effect of molecular structure on the rotational molding characteristics of ultra-low-density-ethylene-α-olefin copolymers. Polym. Eng. Sci. **2004**, *44* (3), 496–508.
15. Guillén-Castellanos, S.A.; Bellehumeur, C.T.; Weber, M. The effect of molecular structure on the coalescence of ethylene-α-olefin copolymers. Proceedings of the Society of Plastics Engineers Annual Technical Conference, San Francisco, CA, May 5–9, 2002; The Society of Plastics Engineers: Brookfield, 2002.
16. Tinson, A.; Takacs, E.; Vlachopoulos, J. The effect of surface tension on the sintering of polyethylene copolymers and blends in rotational molding. Proceedings of the Society of Plastics Engineers Annual Technical Conference, Chicago, IL, May 16–20, 2004; The Society of Plastics Engineers: Brookfield, 2004.
17. Gogos, G. Bubble removal in rotational molding. Polym. Eng. Sci. **2004**, *44* (2), 388–394.
18. Spence, A. Analysis of Bubble Formation and Removal in Rotationally Moulded Products. PhD thesis, Department of Mechanical and Manufacturing Engineering, Queen's University in Belfast, Belfast, Northern Ireland, 1994.
19. Kontopoulou, M.; Vlachopoulos, J. Bubble dissolution in molten polymers and its role in rotational molding. Polym. Eng. Sci. **1999**, *39* (7), 1189–1198.
20. Rao, M.A.; Throne, J.L. Principles of rotational molding. Polym. Eng. Sci. **1972**, *12* (4), 237–264.

21. Nugent, P. A Study of Heat Transfer and Process Control in the Rotational Moulding of Polymer Powders. PhD thesis, Department of Mechanical and Manufacturing Engineering, Queen's University in Belfast, Belfast, Northern Ireland, 1990.
22. Pop-Iliev, R.; Park, C.B. Single-step rotational foam molding of skin-surrounded polyethylene foams. J. Cell. Plast. **2003**, *39* (1), 49–58.
23. Olinek, J.; Anand, C.; Bellehumeur, C.T. Experimental study on the flow of polyethylene powder particles in rotational molding. Polym. Eng. Sci. **2005**, *45* (1), 62–73.
24. Gogos, G.; Liu, X.; Olson, L.G. Cycle time predictions for the rotational molding process with and without mold/part separation. Polym. Eng. Sci. **1999**, *39* (4), 617–629.
25. Bellehumeur, C.T.; Tiang, J.S. Simulation of non-isothermal melt densification of polyethylene in rotational molding. Polym. Eng. Sci. **2002**, *42* (1), 215–229.
26. Cramez, M.C.; Oliveira, M.J.; Crawford, R.J. Effect of nucleating agents and cooling rate on the microstructure and properties of a rotational moulding grade polypropylene. J. Mater. Sci. **2001**, *36* (9), 2151–2161.
27. Cramez, M.C.; Oliveira, M.J.; Crawford, R.J.; Apostolov, A.A.; Krumova, M. Rotationally molded polyethylene: structural characterization by x-ray and microhardness measurements. Adv. Polym. Technol. **2001**, *20* (2), 116–124.
28. Throne, J.L. Some factors influencing cooling rates of rotational molding parts. Polym. Eng. Sci. **1972**, *12* (5), 335–339.
29. Martin, J.A.; Cramez, M.C.; Oliveira, M.J.; Crawford, R.J. Prediction of spherulite size in rotationally molded polypropylene. J. Macromol. Sci. B **2003**, *B42* (2), 367–385.
30. Hay, H.; Weber, M.; Donaldson, R.; Gibbons, I.; Bellehumeur, C. Single site polyethylene resins with enhanced processability for rotational molding applications. Proceedings of the Society of Plastics Engineers Annual Technical Conference, Chicago, IL, May 16–20, 2004; The Society of Plastics Engineers: Brookfield, 2004.
31. Chaudhary, B.I.; Takács, E.; Vlachopoulos, J. Processing enhancers for rotational molding of polyethylene. Polym. Eng. Sci. **2001**, *41* (10), 1731–1742.
32. Scribben, E.; Baird, D. The rotational molding of a thermotropic liquid crystalline polymer. Polym. Eng. Sci. **2005**, *45* (3), 410–423.

Rubber Devulcanization

David A. Benko
Roger N. Beers
The Goodyear Tire & Rubber Company, Akron, Ohio, U.S.A.

INTRODUCTION

During the preparation of many rubber articles such as tires a variety of additives including sulfur, accelerators, and activators are used to induce sulfur cross-links between the rubber chains. These are formed during the heating or curing process and together with reinforcing fillers give the article its strength. Table 1 shows the composition of a rubber compound typical of that used in a body ply of a radial passenger tire.

The cross-links are composed of structures containing various sulfide lengths, usually described as monosulfidic, disulfidic, or polysulfidic, which are determined by the type, amount, and proportion of additives used in the rubber compound.

Devulcanization in a sulfur cured rubber is defined as the cleavage of the mono-, di-, and polysulfidic cross-links formed during vulcanization (Fig. 1). The vulcanization process is irreversible and additional heating induces changes in the network with a shift toward shorter cross-links but does not devulcanize the compound. Other methods are therefore needed to induce devulcanization.

BACKGROUND INFORMATION

Because of the large numbers of used tires existing in stockpiles throughout the United States it is desirable to recycle these tires back into new products. Unfortunately, simply grinding up the used tire and adding the resultant material to virgin rubber causes a significant drop in properties including reducing the strength and flexibility of the compound.

The need for a commercially viable devulcanization process has existed for years and continues to be a goal. Devulcanization without rubber chain degradation offers the potential for recycling used tires back into new products without sacrificing performance.

Considerable effort has been directed to solving this problem.[1,2] Although chemical probes have been developed that selectively cleave carbon–sulfur and sulfur–sulfur bonds but not carbon–carbon bonds, most of the effort on devulcanization processes has been focused on providing a usable form of rubber suitable for use as a reclaimed material in new articles.[3] In most cases direct evidence for actual devulcanization, i.e., breaking sulfur–sulfur and carbon–sulfur bonds without polymer chain scission, is lacking. However, in many instances the so-called devulcanization process increases the suitability for reuse (Fig. 2).

This article will review the classical chemical methods of devulcanization and their inherent limitations. The newer methods of devulcanization currently under development will also be reviewed. These include devulcanization using microbes, microwave, ultrasonically induced devulcanization, and devulcanization using supercritical fluids.

CHEMICAL DEVULCANIZATION

A large number of chemical devulcanization agents for natural and synthetic rubbers have been developed. These include phosphines and phosphates, numerous sulfides and mercaptans, metal salts such as methyl iodide, phenyl lithium, lithium aluminum hydride, and phase-transfer catalysts.[3–25]

Included in the list of sulfides and mercaptans are diphenyl disulfide, dibenzyl disulfide, diamyl disulfide, bis(alkoxy aryl) disulfides, butyl mercaptan and thiophenols, xylene thiols and other mercaptans, phenol sulfides and disulfides.[6–13,26,27]

Cook and coworkers reported the preparation, evaluation, and structural correlation of alkyl phenol sulfides as devulcanizing agents for styrene–butadiene rubber (SBR).[13] The effect of these alkyl phenol sulfides as reclaiming agent was compared with that of many aromatic thiols. Some *N*,*N*-dialkyl aryl amine sulfides were shown to be highly active reclaiming agents for vulcanized SBR in both neutral and alkaline reclaiming processes.[12]

A review of the science and technology of reclaimed rubber was published by Le Beau in 1967.[14] Knorr has shown the action of diaryl disulfide on the natural and synthetic rubber scraps of technical goods.[15]

A reaction mechanism for the breaking of polysulfide bonds in the presence of propane thiol/piperidine probe was proposed by Saville and Watson.[28] The thiol-amine combination gives an associate, possibly piperidinium propane-2-thiolate ion pair where sulfur atoms enhance the nucleophilic character that is responsible for cleaving

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008034

Table 1 Typical tire compound

Ingredient	Amount[a]
Diene rubber[b]	100
Reinforcing filler	50–60
Oils	5–15
Antioxidants	1–2
Zinc oxide	5
Accelerators	1–2
Sulfur	2–3

[a]Amounts are parts by weight per 100 parts by weight rubber.
[b]Diene rubber includes natural rubber, polyisoprene rubber, butadiene rubber, and styrene–butadiene rubber.

organic polysulfide linkages. The cleavage of polysulfide cross-links takes place possibly as a result of p_π–d_π delocalization of the displaced sigma electron pair of RSS^-. Campbell found hexane-1-thiol to be more reactive and thus capable of cleaving both polysulfide and disulfide linkages while leaving the monosulfides intact (Fig. 3).[29]

Moore and Trego described the use of triphenyl phosphine and di-*N*-butyl phosphate as chemical probes to establish a cross-link network structure in rubber vulcanizates.[4,5] Triphenyl phosphine and trialkyl phosphates cleave di- and polysulfide, as illustrated in Fig. 4.

Studebaker identified the use of lithium aluminium hydride as a chemical probe.[23,24] Under the right conditions it cleaves poly- and disulfide bonds, leaving the monosulfide intact. Lithium aluminum hydride reacts with polysulfides in an etheral solvent at moderate temperatures and then with a weak acid, the terminal groups are liberated as thiols and interior sulfur atoms are converted to hydrogen sulfide.[30–39] Lithium aluminum hydride under appropriate reaction conditions also cleaves disulfide bonds in organic disulfide, which is structurally related to cross-links, into two thiol groups.[32]

Similarly Gregg and Katrenick found that phenyl lithium will cleave polysulfides and disulfides but not monosulfide linkages.[21,22]

Methyl iodide can be employed to estimate monosulfide linkages in vulcanized natural rubber (NR).[33,34] After swelling the rubber in methyl iodide for several days the level of network bound iodine after reaction would reflect the concentration of monosulfide groups because the simple saturated monosulfide group reacts as shown in Fig. 5.

Simple disulfides reacted very slowly with methyl iodide but their reaction and those of monosulfides could be catalyzed by mercuric iodide.[35,36]

Anderson patented the reclaiming of sulfur vulcanized rubber in the presence of oil, water vapor, and a mixture of aryl disulfides (diphenyl disulfide, dicresyl disulfide, and dixylyl disulfide) at elevated temperature and pressure.[40]

Desulfurization of suspended rubber vulcanizate crumb (10–30 mesh) was carried out in a solvent such as toluene, naphtha, benzene, or cyclohexane in the presence of sodium.[41] The alkali metal cleaves mono, di- and polysulfidic cross-linkages of the swelled and suspended vulcanized rubber crumb at around 300°C in the absence of oxygen. As claimed by the authors, such treatment yielded a rubber polymer having a molecular weight substantially equal to that of rubber prior to vulcanization.

2-Mercaptobenzothiazole was also found to be effective as a reclaiming agent.[42] In this process

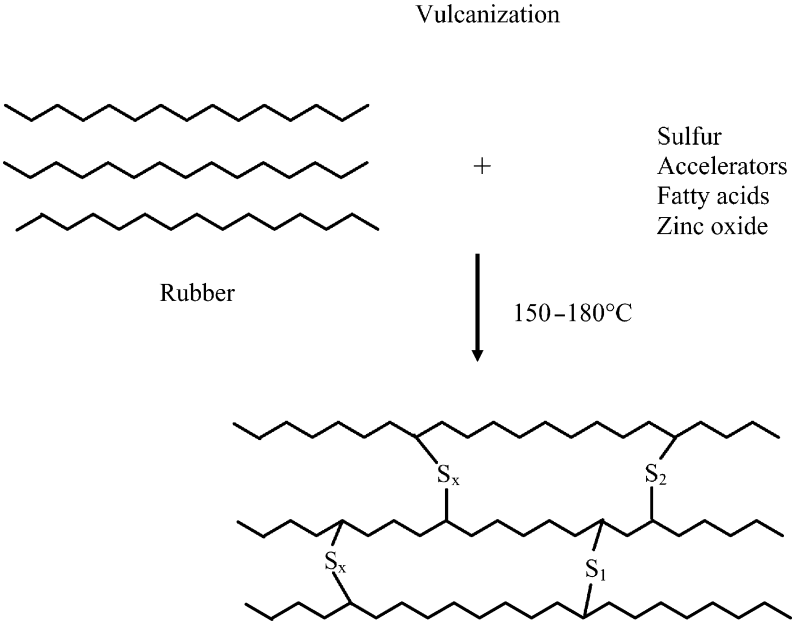


Fig. 1 Vulcanization.

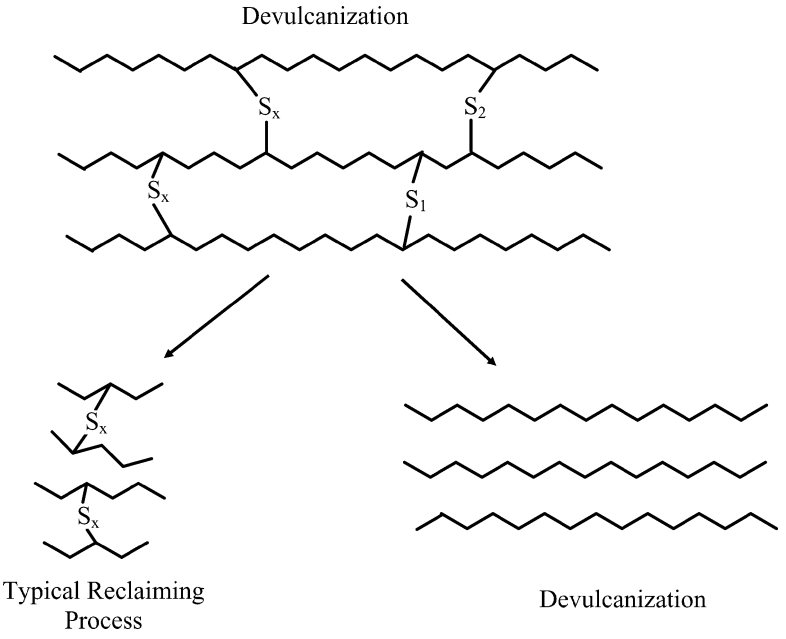


Fig. 2 Devulcanization.

powder rubber from waste tires was kneaded with process oil in the presence of 2-mercaptobenzothiazole or its cyclohexylamine salt to give reclaim rubber.

Vehicle tire scrap containing polyisoprene rubber, SBR, and butadiene rubber was devulcanized by low-temperature phase-transfer catalyst. Both the devulcanizing agent composition and the process were patented. The novelty of this process lies in the use of low-temperature phase-transfer catalyst and a process temperature lower than 150°C. The devulcanized rubber of this invention is distinguishable from conventional reclaimed rubber in that the devulcanized rubber is substantially free from polysulfide cross-links, which are selectively broken during the process with negligible main chain scission.[25,43]

Microwave Method

The microwave method is useful because it provides an economical, ecologically sound method of reusing elastomeric waste to return it to the same process and products in which it was originally generated. In this technique a controlled dose of microwave energy is used at a specified frequency and energy level to cleave carbon–carbon bonds.[44–46] Thus, in this process elastomer waste can be reclaimed without complete depolymerization to a material capable of being recompounded and revulcanized having physical properties essentially equivalent to the original vulcanizate.

In typical commercial processes the ''devulcanized'' rubber is not degraded. In this process it is claimed that

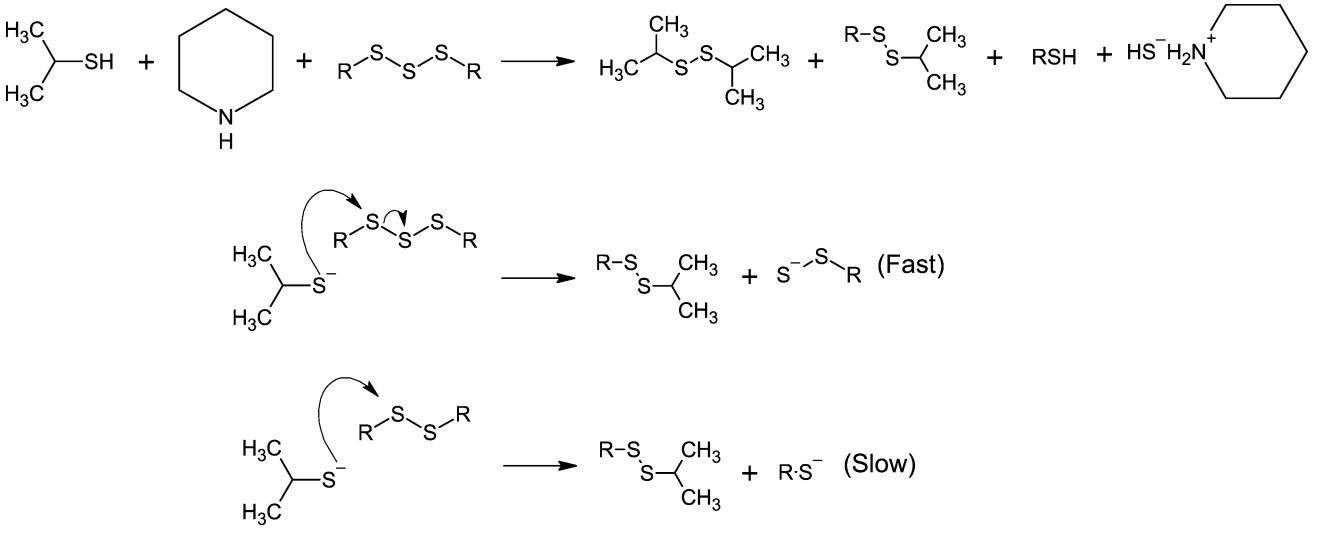


Fig. 3 Polysulfide bond breaking with thiols.

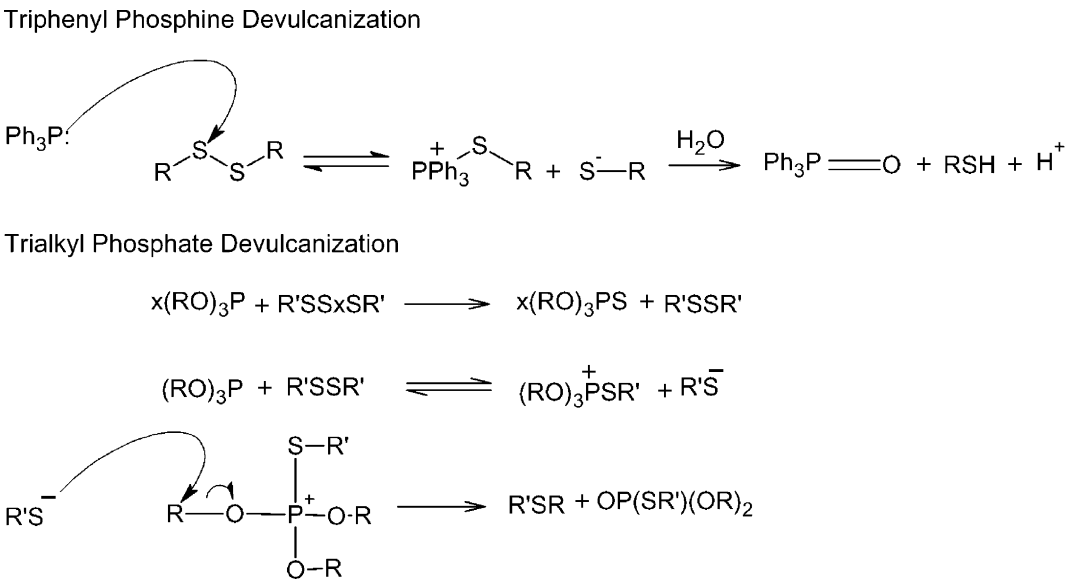


Fig. 4 Phosphine and phosphite devulcanization.

sulfur vulcanized elastomer containing polar groups is suitable for microwave devulcanization. Tyler and Cerny claimed their microwave devulcanization process as a method of pollution controlled reclaiming of sulfur vulcanized elastomer containing polar groups.[47] The microwave energy devulcanization device generates heat at a temperature in excess of 260°C to yield a mass, which is fed to an extruder that extrudes the rubber at a temperature of 90–125°C. The extrudate can be used per se as a compounding stock. Another process was developed for reclaiming waste elastomers by microwave radiation. The process involves the impregnation of the waste rubber with an essential oil and then heat treating the impregnated material under reduced pressure with microwave radiation.[48]

The waste material must be polar so that the microwave energy will generate the heat necessary to devulcanize. Microwave energy between 915 and 2450 MHz and between 41 and 177 W/hr/lb is sufficient to sever all cross-link bonds but insufficient to sever polymer chain degradation. The cost of devulcanized

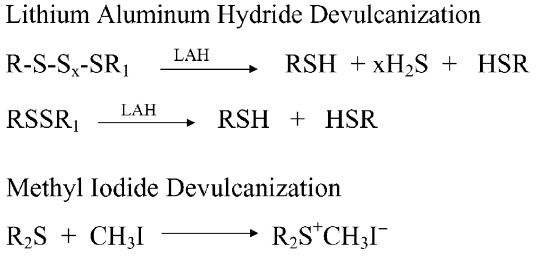


Fig. 5 Lithium aluminum hydride and methyl iodide devulcanization.

hose and inner tube material by the microwave method is only a fraction of the cost of the original compound. The transformation from waste to refined stock ready for remixing takes place in only 5 min with usually 90–95% recovery of the rubber. Therefore, it appears that this microwave technique is a unique method of reclaiming in terms of properties and fastness of the process.

Ultrasonic Method

After the microwave techniques, ultrasonic energy was used for the devulcanization of cross-linked rubber. The first work with ultrasonic energy was reported by Pelofsky in 1973, which was patented.[49] In this process solid rubber articles such as tires are immersed into a liquid, which is then kept under a source of ultrasonic energy whereby the bulk rubber effectively disintegrates upon contact and dissolves into liquid. In this process ultrasonic irradiation is in the range of about 20 kHz and at a power intensity of greater than 100 W.

Ultrasonic reclaiming of NR was reported by Okuda and Hatano in 1987, which was also patented.[50] They subjected the NR vulcanizate to 50 kHz ultrasonic energy for 20 min to achieve devulcanization followed by revulcanization and obtained reclaimed rubber with similar properties to those of original rubber. Isayev and coworkers reported in a number of publications the phenomenon of devulcanization by ultrasound energy and they also patented their developments.[51–59]

The devulcanization process requires a high energy level to break carbon–sulfur and sulfur–sulfur bonds.[1] An ultrasonic field creates high-frequency extension–contraction stresses in various media.[1] Isayev and his group also made a percolation simulation of the network degradation during ultrasound devulcanization in which they claimed an excellent agreement of experimental data for SBR and ground rubber from used tires (GRT) with the predicted dependence of the gel fraction of devulcanized rubber on cross-link density. Curing behavior, rheological properties, structural characteristics of devulcanized rubber from model SBR and GRT rubbers, as well as mechanical properties of vulcanized rubber samples were studied and a possible mechanism of devulcanization was also discussed. They characterized the degree of devulcanization by the measurement of cross-link density and gel fraction of the devulcanized rubber. Later, they published on the ultrasound devulcanization of sulfur vulcanized SBR and on vulcanization of ultrasonically devulcanized SBR elastomers.[60]

Isayev and coworkers studied the devulcanization of SBR at various temperatures, viz., 121°C, 149°C, and 176°C, different clearances at various flow rates, and the ultrasonic oscillation amplitudes.[51–53] The extent of devulcanization was studied by measuring percentage and cross-link density of the gel fraction. It was reported that both the cross-link density and the gel fraction decrease in the devulcanization process. For original ground rubber tire the measured gel fraction is 83% and cross-link density of gel is 0.21 kmol/m^3, but after ultrasound treatment at 121°C barrel temperature it reduces to 64–65% with cross-link density of 0.02 kmol/m^3. The cross-link density also decreases with higher residence time in the treatment zone and with higher specific ultrasonic energy.

The mechanical properties of the revulcanized sample were also studied. With decrease in the cross-link density of the devulcanized rubber, the tensile strength of revulcanized samples varies from 1.5 to 10.5 MPa and elongation at break varies from 130% to 250%. Based on the results of mechanical properties, Isayev et al. proposed that the devulcanized sample having a cross-link density lower than 0.06 kmol/m^3 can be regarded as overtreated, and samples with cross-link density higher than 0.10°kmol/m^3 can be regarded as undertreated.[52] Thus, overtreatment causes main chain breakage and undertreatment causes insufficient devulcanization. They also reported that ultrasound treatment of SBR results in low molecular weights of the sol fraction: $M_n = 2 - 4 \times 10^3$.[53] Ultrasonic devulcanization, therefore, causes significant degradation of polymer chains. A simple model based on a purely topological consideration was proposed and simulation of the process was carried out.[61–64] In the model they have assumed a breakup of the main chain bond and cross-link bonds as independent random events. Such random scission of cross-links and main chain results in the formation of soluble branched rubber chains regarded as fragmented gel structure or microgel. It is found that during ultrasound devulcanization the molecular weight of sol fraction decreases, from which it may be understood that during ultrasound treatment not only C–S or S–S bonds but also C–C bonds break. Isayev et al. suggested a revulcanization scheme.[55] They concluded that devulcanized rubber contained a larger amount of sulfidized molecules that were responsible for cross-linking during revulcanization.

Biotechnological Processes

Biodegradation of NR was recognized as early as 1914.[65] But an effective process was not discovered because rubber is a hydrophobic substance that is subject to attack only at the surface, rubber vulcanizates are highly cross-linked and highly branched, and rubber vulcanizates contain a large number of biologically active additives that retard biodegradation. Various biodegradation processes were tried over the years since, with some success reported, particularly after about 1985.

Fermentation methods were studied extensively at Rutgers University by Nickerson and coworkers and by Elmer in the 1970s on ground scrap tires.[65] Their work was aimed at producing commercially valuable products via the fermentation process from scrap tires. This work led to many important findings and showed that although technically feasible, this was not a solution to get rid of scrap tires. Even under ideal conditions only minor reduction of the rubber occurred.

An interesting recent approach was reported in a patent application to utilize a chemolithiotrope bacterium in aqueous suspension for attacking powder elastomers on the surface only, so that after mixing with virgin rubber, diffusion of soluble polymer chains is facilitated and bonding during vulcanization becomes again possible.[66,67]

A biotechnological process was developed by Straube et al. for the devulcanization of scrap rubber by holding the comminuted scrap rubber in a bacterial suspension of chemolithotropic microorganisms with a supply of air until elemental sulfur or sulfuric acid is separated.[68] This seems to be an interesting process, which obtains reclaim rubber and sulfur in a simplified manner.

The biodegradation of the *cis*-1,4-polyisoprene chain was achieved by Tsuchi, Suzuki, and Takeda.[69–71] They used bacterium that belonged to the genus *Nacardia* and led to considerable weight loss of different soft-type NR-vulcanizates. The microbial

desulfurization or devulcanization of particle surfaces was investigated to increase the possibility of producing high-quality rubber products containing a larger percentage of recycled rubber.[72]

In a typical process rubber powder, mainly SBR of old tires with 1.6% sulfur, was treated with different species of *Thiobacillus*, i.e., *Thiobacillus ferrooxidans*, *Thiobacillus thiooxidans*, and *Thiobacillus thioparus* in shake flasks and in a laboratory reactor. The sulfur oxidation depends to a large extent on the particle size. The best results were obtained with *T. thioparus* with a particle size of 100–200 μm. Of the total sulfur of the rubber powder, 4.7% was oxidized to sulfate within 40 days.[72]

In a recent paper Steinbuchel studied the biological attack of microorganisms on rubber materials to evaluate the possible contributions of biotechnology for the development and recycling of used rubber products.[73] Adaptation of microbial enrichment cultures with tire crumb material for several months resulted in enhanced growth of microorganisms, especially for NR and SBR.

Romine and Snowden-Swan developed a process for exposing fine powdered rubber to an enzyme that attacks the sulfur cross-links on the surface of the rubber particles.[74] This process is carried out at ambient temperature and takes approximately 3 days. The devulcanization reaction is halted at the sulfone ($R–SO_2–R$) or sulfoxide (R–SO–R) stages and not allowed to proceed to the unwanted sulfate ($R–OSO_2–R$) form. The sulfone and sulfoxide forms are reactive with virgin rubber and allow more of the devulcanized rubber to be used. The specific thiophyllic microbes used include *T. ferrooxidans*, *T. thiooxidans*, *Rhodococcus rhodochrous*, and *Sulfolobus acidocaldarius*. About 10–20% by weight of this devulcanized rubber can be added to a virgin rubber matrix. Rubber particles processed with *S. acidocaldarius* exhibited an increase of 15% in the modulus of elasticity when added at 15% by weight to a virgin rubber compound.

A patent was issued to Fliermans and Wicks in 2002 for combining microwave and biological techniques to treat vulcanized rubber particles.[75] Samples of 40 mesh crumb rubber were incubated with bacillus-type bacterium. The best results were obtained with a 10/40 volume ratio of bacterium to crumb rubber at a temperature of 60–65°C for a 20 hr treatment period. The devulcanized rubber was evaluated in a tire tread compound at 20% with definite improvement over untreated crumb in terms of Mooney viscosity, tensile strength, and elongation. It was found that by treating the biologically devulcanized rubber with microwave energy, the overall properties of the resulting cured rubber are improved over the comparable control mix using either untreated crumb rubber or rubber treated solely with the biotreatment or microwave protocols.

Supercritical Fluid Devulcanization

Hunt and Kovolak discovered that cured rubber could be devulcanized by heating with 2-butanol under supercritical conditions.[76] By heating rubber compounds to at least 150°C under a pressure of 3.4×10^6 Pa in the presence of 2-butanol the molecular weight of the rubber was maintained at a relatively high level and its microstructure was unchanged.

For example, sulfur-cured SBR samples that contained no filler, carbon black, silica, or a combination of carbon black and silica were heated with the 2-butanol under supercritical conditions. The SBR had an original weight average molecular weight of about 400,000. The weight average molecular weights of the devulcanized SBR samples recovered are reported in Table 2.

Several related alcohols were also investigated and were found to be less effective, although most did induce devulcanization. Table 3 shows the cumulative amounts of SBR polymer recovered from a cured sample after heating at various temperatures. It can be seen from Table 3 that 2-butanol was far better than any of the other alcohols evaluated.

Benko and Beers found that by treating ground rubber obtained from used tires in a similar process they could devulcanize the surface of the ground particle.[77–79] This enabled its reuse as new rubber compounds with only minimal loss of properties.

Although 2-butanol is a preferred solvent, a number of other alcohols and ketones are described. The alcohol or ketone employed as the solvent will have a critical temperature that is within the range of about 200°C to about 350°C. It is preferred for the alcohol or the ketone used as the solvent to have a critical temperature that is within the range of about 250°C to about 320°C. The term ''critical temperature'' is defined as the temperature above which the gas of a compound (the alcohol or the ketone) cannot be liquefied by the application of pressure. Some representative

Table 2 Polymer molecular weights after devulcanization

Example	Filler	Molecular weight[a]
11	No filler	181,000
12	No filler	186,000
13	Silica	244,000
14	Silica	293,000
15	Carbon black	197,000
16	Carbon black	216,000
17	Carbon black/silica	177,000
18	Carbon black/silica	177,000

[a]The molecular weights reported are weight average molecular weights.

Table 3 Polymer recovery with alcohols

		Cumulative polymer recovery (%)			
Example	**Alcohol**	**150°C**	**200°C**	**250°C**	**300°C**
1	2-Butanol	38	82	90	93
2	2-Butanol	40	70	85	92
3	Methanol	2	3	4	7
4	Ethanol	2	4	9	20
5	1-Propanol	3	16	43	69
6	2-Propanol	2	7	13	25
7	1-Butanol	4	19	57	86
8	Isobutyl alcohol	2	10	44	74
9	1-Pentanol	3	11	42	89
10	4-Methyl-2-pentanol	2	11	33	68

(From Ref.[76].)

examples of alcohols that can be used include methanol, ethanol, allyl alcohol, 1-propanol, isopropyl alcohol, *n*-butanol, *iso*-butanol, 2-butanol, *tert*-butanol, 1-pentanol, 2-methyl-1-butanol, 3-methyl-1-butanol, 3-methyl-2-butanol, 2,2-dimethyl-1-propanol, methyl isobutyl ketone, and 1-hexanol. Some representative examples of ketones that can be used include acetone, methyl ethyl ketone, methyl *n*-propyl ketone, methyl isopropyl ketone, and diethyl ketone. Mixtures of such alcohols and ketones can be utilized as the solvent.

In this series of experiments, whole tire reclaim rubber was ground to a particle size of 40 mesh (about 420 μm) and the surface of the ground crumb rubber was then devulcanized. The surface devulcanization was carried out in 2-butanol under the conditions of time, pressure, and temperature specified in Table 3. Then, the samples of surface devulcanized reclaimed rubber made were analyzed by a thermogravimetric technique to determine the volatile content and the polymer content. The results of this analysis are also reported in Table 4 along with the analysis of a control that was not subjected to the devulcanization procedure.

The samples of surface devulcanized reclaimed rubber made in this series of experiments were then compounded with a blend of virgin rubbers and cured. The blends were made by mixing 20 parts per 100 parts of rubber (phr) of the surface devulcanized reclaimed rubber samples with 70 phr of Plioflex® 1712 SBR, 30 phr of Budene® 1254 polybutadiene rubber, about 9 phr of aromatic oil, about 70 phr of carbon black, about 2 phr of stearic acid, about 4 phr of wax, about 1 phr of accelerator, about 2 phr of zinc oxide, about 1.5 phr of sulfur, and about 1 phr of antioxidant. The Plioflex® 1712 has a bound styrene content of about 28.5% and was oil extended with about 37.5% of an

Table 4 Surface devulcanization

Example	**Temperature (°C)**	**Pressure (psig)**	**Time (min)**	**Volatiles (%)**	**Polymer (%)**
19	270	900	20	41.61	11.8
20	270	900	40	38.75	9.16
21	270	1500	20	15.06	31.24
22	270	1500	40	21.36	23.07
23	300	900	20	38.09	18.54
24	300	900	40	46.08	10.96
25	300	1500	20	36.78	16.58
26	300	1500	40	37.44	7.28
27	285	1200	30	35.77	19.56
28	285	1200	30	36.68	20.23
Control				12.26	53.28

(From Ref.[77].)

Table 5 Cure properties of devulcanized rubber

Example	Torque (dN)[a]	Ts1 (min)[b]	T25 (min)[c]	T90 (min)[d]
19	17	7.3	9.1	18.5
20	18.2	7	8.7	16.7
21	16.3	7	9	19
22	16.3	7	8.7	19.5
23	16.4	6.7	8.3	17.7
24	16.9	6.8	8.4	19.5
25	18	6.5	8.3	18.5
26	17	6.5	8.1	17
27	17.5	6.5	8.2	16.5
28	16.6	6.8	8.6	19
Original	16.6	5.3	6.6	14.9
None	19	6.4	8.3	7.2

Properties are obtained from a cure rheometer at 150°C.
[a]Minimum torque to maximum torque.
[b]Time to a 1 dN increase in torque.
[c]Time to 25% of maximum torque.
[d]Time to 90% of maximim torque.
(From Ref.[77].)

aromatic oil. The blends were then cured at 150°C for 20 min. The cure properties of the blends are reported in Table 5 and the physical properties of the cured blends are reported in Table 6.

As can be seen from Table 5, the blends made with the surface devulcanized reclaimed rubber crumb did not have cure characteristics that differed substantially from the blend made without including any of the reclaimed rubber (the series labeled ''None''). In fact, the blends made with the surface devulcanized reclaimed rubber crumb had cure characteristics that were more similar to those made with no ground rubber than they were to those made with untreated whole tire reclaim rubber (the series labeled ''Original'').

As can be seen from Table 6, the physical properties of some of the cured blends made with the surface devulcanized reclaimed rubber crumb samples were equivalent to those made with only virgin rubber. For instance, the 100% modulus, 300% modulus, and percent elongation measured in Examples 19 and 20 were very similar to those found in the control using only virgin rubber (the series labeled ''None'').

Table 6 Physical properties of devulcanized rubber

Example	Tensile strength (MPa)	Elongation (%)	100% Modulus (MPa)	300% Modulus (MPa)
19	15.4	747	1.23	4.45
20	15.3	701	1.33	4.93
21	13	775	1.10	3.39
22	13.7	798	1.11	3.46
23	15.1	814	1.09	3.54
24	15.2	778	1.11	3.88
25	16.1	764	1.19	4.49
26	15.4	738	1.21	4.43
27	15.2	789	1.13	3.88
28	14.8	791	1.11	3.74
Original	14.5	661	1.16	4.10
None	18.6	757	1.28	4.93

(From Ref.[77].)

CONCLUSIONS

Because of the large numbers of used tires existing in stockpiles it is desirable to recycle these tires back into new products. Unfortunately, simply grinding up the used tire and adding the resultant material into virgin rubber causes a significant drop in properties including reducing the strength and flexibility of the compound.

The need for a commercially viable devulcanization process has existed for years and continues to be a goal. Devulcanization without rubber chain degradation offers the potential for recycling used tires back into new products without sacrificing performance.

Chemical processes have been known for many years but are either ineffective in selectively cleaving cross-links or commercially not viable. Recent developments may eventually provide a true devulcanization process that is commercially viable. These include ultrasonic devulcanization, biotechnological devulcanization, and supercritical fluid devulcanization.

REFERENCES

1. Warner, W.C. Methods of devulcanization. Rubber Chem. Technol. **1994**, *67*, 559.
2. Adhikari, B.; De, D.; Maiti, S. Reclamation and recycling of waste rubber. Prog. Polym. Sci. **2000**, *25*, 909.
3. Schnecko, H. Kautschuk Gummi Kunststoffe **1994**, *47*, 885.
4. Moore, C.G.; Trego, B.R. Structural characterization of vulcanizates. Part IV. Use of triphenylphosphine and sodium di-n-butyl phosphite to determine the structures of sulfur linkages in natural rubber, cis-1,4-polyisoprene, and ethylene-propylene rubber vulcanizate networks. J. Appl. Polym. Sci. **1964**, *8*, 1957.
5. Moore, C.G.; Trego, B.R. Structural characterization of vulcanizates. Part II. Use of triphenylphosphine to determine the structures of sulfur linkages in unaccelerated natural rubber-sulfur vulcanizate networks. J. Appl. Polym. Sci. **1961**, *5*, 299.
6. Tewksbury, L.B., Jr.; Howland, LH. Canadian Patent 2,469,529, 1949.
7. Rebmann, A. Swiss Patent 215,952, 1948.
8. Sverdrup, E.F.; Elgin, J.C. U.S. Patent 2,415,449, 1947.
9. Sverdrup, E.F.; Elgin, J.C. Canadian Patent 452,085, 1948.
10. Cotton, F.H.; Gibbons, P.A. U.S. Patent 2,408,296, 1946.
11. U.S. Patents 2,211,592, 1940; 2,414,145, 1947; 2,467,789, 1949; 2,471,866, 1949.
12. U.S. Patents 2,193,624, 1940; 2,359,122, 1944; 2,333,810, 1943; 2,363,873, 1944; 2,372,584, 1945; 2,469,529, 1949.
13. Webb, F.J.; Cook, W.S.; Albert, H.E.; Smith, G.E.P., Jr. Arylamine sulfide catalysts in reclaiming GR-S vulcanizates. Ind. Eng. Chem. **1954**, *46*, 1711.
14. Le Beau, D.S. Rubber Chem. Technol. **1967**, *40*, 217.
15. Knorr, K. Kautschuk Gummi Kunststoffe **1994**, *47*, 54.
16. Selker, M.L.; Kemp, A.R. Sulfur linkage in vulcanized rubbers. Ind. Eng. Chem. **1944**, *36*, 16.
17. Selker, M.L. Reaction of methyl iodide with sulfur compounds. Ind. Eng. Chem. **1948**, *40*, 1467.
18. Selker, M.L.; Kemp, A.R. Sulfur linkage in vulcanized rubber acetone extraction of vulcanizates. Rubber Chem. Technol. **1949**, *22*, 8.
19. Moore, C.G. J. Polym. Sci. **1958**, *32*, 503.
20. Manik, S.P.; Banerjee, S. Sulfenamide accelerated sulfur vulcanization of natural rubber in presence and absence of dicumyl peroxide. Rubber Chem. Technol. **1970**, *40*, 1311.
21. Gregg E.C., Jr.; Katrenick, S.E. Chemical structures in cis-1,4-polybutadiene vulcanizates. Model compound approach. Rubber Chem. Technol. **1970**, *43*, 549.
22. Gregg E.C., Jr. Sulfur crosslinks in polybutadiene vulcanizates. Rubber Chem. Technol. **1969**, *42*, 1136.
23. Studebaker, M.L. Lithium aluminum hydride analysis of sulfur—cured vulcanizates. Rubber Chem. Technol. **1970**, *43*, 624.
24. Studebaker, M.L.; Nabors, L.G. Sulfur group analyses in natural rubber vulcanizates. Rubber Chem. Technol. **1959**, *32*, 941.
25. Nichols, P.P. The scission of polysulfide crosslinks in scrap rubber particles through phase transfer catalysis. Rubber Chem. Technol. **1982**, *55*, 1499.
26. Elgin, J.C. Canadian Patent 456,789, 1949.
27. Sverdrup, E.F. U.S. Patent 2,494,593, 1949.
28. Saville, B.; Watson, A.A. Structural characterization of sulfur-vulcanized rubber networks. Rubber Chem. Technol. **1967**, *40*, 100.
29. Campbell, D.S. Structural characterization of vulcanizates part X. Thiol-disulfide interchange for cleaving disulfide crosslinks in natural rubber vulcanizates. Rubber Chem. Technol. **1970**, *43*, 210.
30. Arnold, R.C.; Lien, A.P.; Alm, R.M. The action of lithium aluminum hydride on organic disulfides. J. Am. Chem. Soc. **1950**, *72*, 731.
31. Farmer, R.H.; Ford, J.F.; Lyons, J.A. J. Appl. Chem. **1954**, *4*, 554.
32. Porter, M.; Saville, B.; Watson, A.A. J. Chem. Soc. **1963**, 346.

33. Meyer, K.H.; Hohenemser, W. Helv. Chim. Acta **1935**, *18*, 1061.
34. Meyer, K.H.; Hohenemser, W. Contribution to the study of the vulcanization reaction. Rubber Chem. Technol. **1936**, *9*, 201.
35. Hilditch, T.P.; Smiles, S. J. Chem. Soc. **1907**, *91*, 1394.
36. Steinkopf, W.; Muller, S. Ber **1923**, *56B*, 1926.
37. Klinger, H.; Maason, A. Ann. **1898**, *252*, 241.
38. Brjuchonenko, A. Ber. **1899**, *31*, 3176.
39. Bateman, L.; Glazebrook, R.W.; Moore, C.G.; Porter, M.; Ross, G.W.; Saville, R.W. J. Chem. Soc. **1958**, 2838.
40. Anderson, E., Jr. U.S. Patent 4,544,675, 1985.
41. Myers, R.D.; Nicholson, P.; MacLeod, J.B.; Moir, M.E. U.S. Patent 5,602,186, 1997.
42. Okamoto, H.; Inagaki, S.; Onauchi, Y.; Furukawa, J. Nippon Gomu Kyokaishi **1979**, *52* (12), 774.
43. Nicholas, P.P. U.S. Patent 4,161,464, 1979.
44. Novotny, D.S.; Marsh, R.L.; Masters, F.C.; Tally, D.N. U.S. Patent 4,104,205, 1978.
45. Fix, S.R. Elastomerics **1980**, *112* (6), 38.
46. Makrov, V.M.; Drozdovski, V.F. *Reprocessing of Tires and Rubber Wastes*; Ellis Horwood: New York, 1991.
47. Tyler, K.A.; Cerny, G.L. U.S. Patent 4,459,450, 1984.
48. Hunt, J.R.; Hall, D. U.S. Patent 5,362,759, 1994.
49. Pelofsky, A.H. U.S. Patent 3,725,314, 1973.
50. Okuda, M; Hatano, Y. Japanese Patent 62,121,741, 1987.
51. Isayev, A.I.; Chen, J.; Tukachinsky, A. Novel ultrasonic technology for devulcanization of waste rubbers. Rubber Chem. Technol. **1995**, *68*, 267.
52. Tukachinsky, A.; Schworm, D.; Isayev, A.I. Devulcanization of waste tire rubber by powerful ultrasound. Rubber Chem. Technol. **1996**, *69*, 92.
53. Levin, V.Yu.; Kim, S.H.; Isayev, A.I.; Massey, J.; Von Meerwall, E. Ultrasound devulcanization of sulfur vulcanized SBR: crosslink density and molecular mobility. Rubber Chem. Technol. **1996**, *69*, 104.
54. Isayev, A.I.; Yushanov, S.P.; Kim, S.H.; Levin, V.Yu. Rheo Acta **1996**, *35*, 616.
55. Isayev, A.I.; Kim, S.H.; Levin, V.Yu. Superior mechanical properties of reclaimed sbr with bimodal network. Rubber Chem. Technol. **1997**, *70*, 194.
56. Levin, V.Yu.; Kim, S.H.; Isayev, A.I. Vulcanization of ultrasonically devulcanized SBR elastomers. Rubber Chem. Technol. **1997**, *70*, 120.
57. Johnston, S.T.; Massey, J.; VonMeerwall, E.; Kim, S.H.; Levin, V.Yu.; Isayev, A.I. Ultrasound devulcanization of SBR: molecular mobility of gel and sol. Rubber Chem. Technol. **1997**, *70*, 183.
58. Isayev, A.I. U.S. Patent 5,258,413, 1993.
59. Isayev, A.I.; Chen, J. U.S. Patent 5,284,625, 1994.
60. Tapale, M.; Isayev, A.I. Continuous ultrasonic devulcanization of unfilled NR vulcanizates. J. Appl. Polym. Sci. **1998**, *70*, 2007.
61. Isayev, A.I.; Yushanov, S.P.; Chen, J. Ultrasonic devulcanization of rubber vulcanizates. I. Process model. J. Appl. Polym. Sci. **1996**, *59*, 803.
62. Isayev, A.I.; Yushanov, S.P.; Chen, J. Ultrasonic devulcanization of rubber vulcanizates. II. Simulation and experiment. J. Appl. Polym. Sci. **1997**, *59*, 815.
63. Isayev, A.I.; Yushanov, S.P.; Schworm, D.; Tukachinsky, A. Plast. Rubber Compos. Process Appl. **1996**, *25*, 1.
64. Yushanov, S.P.; Isayev, A.I.; Levin, V.Yu. Percolation simulation of the network degradation during ultrasonic devulcanization. J. Polym. Sci. Phys. Ed. **1996**, *34*, 2409.
65. Beckman, J.A.; Crane, G.; Kay, E.L.; Laman, J.R. Scrap tire disposal. Rubber Chem. Technol. **1974**, *47*, 597.
66. Merseburg, T.H.; Neumann, W., et al. DEO 4042009, German Patent, Jun 6, 1992.
67. Loffler, M. Deutsche Kautschuk Tagung dkt '94, Jun 27–30, 1994.
68. Straube, G.; Straube, E.; Neumann, W.; Ruckauf, H.; Forkmann, R.; Loffler, M. U.S. Patent 5,275,948, 1994.
69. Tsuchii, A.; Suzuki, T.; Takeda, K. Appl. Environ. Microbiol. **1985**, *50*, 965.
70. Tsuchii, A.; Takeda, K.; Tokiwa, Y. Degradation of the rubber in truck tires by a strain of Nocardia. Biodegradation **1997**, *7*, 405.
71. Tsuchii, A.; Takeda, K. Appl. Environ. Microbiol. **1990**, *56*, 269.
72. Loffler, M.; Straube, G.; Straube, E. Biohydrometall. Technol. Proc. Int. Biohydrometall. Symp. **1993**, *2*, 673.
73. Linos, A.; Steinbuchel, A. Kautschuk Gummi Kunststoffe **1998**, *51*, 496.
74. Romine, R.A.; Snowden-Swan, L.J. U.S. Patent 5597851, Jan 28, 1970.
75. Fliermans, C.B.; Wicks, G.G. U.S. Patent 6,407,144, Jun 18, 2002.
76. Hunt, L.K.; Kovalak, R.R. U.S. Patent 5891926, 1999.
77. Benko, D.A.; Beers, R.N. U.S. Patent 6380269, 2002.
78. Benko, D.A.; Beers, R.N. U.S. Patent 6387965, 2002.
79. Benko, D.A.; Beers, R.N. U.S. Patent 6462099, 2002.

Scrubbers

S. Komar Kawatra
Department of Chemical Engineering, Michigan Technological University, Houghton, Michigan, U.S.A.

INTRODUCTION

Scrubbers are pollution control devices, which remove pollutants from gas streams, particularly from the combustion gases produced by facilities such as coal-fired power plants. Scrubbers may use absorbents in slurry or solution form (wet scrubbers), or in powder form (dry scrubbers), with wet scrubbers being more often used. Modern scrubbers can consist of several cleaning steps such as dust separation, sulfur removal, capture of mercury, lead, or other heavy metals, and breakdown of nitrogen oxides. Scrubbers of the future will also need to capture and sequester carbon dioxide. Currently, the most common application of scrubbers is still the removal of sulfur, frequently combined with capture of fly-ash.

POLLUTANTS REMOVED BY SCRUBBERS

Sulfur

In sulfur scrubbers, the sulfur oxides that are produced during coal combustion are removed from the combustion gases using an absorptive chemical, such as calcium oxide. As nitrogen oxides are also produced during combustion, it is possible to develop combined desulfurization methods that can remove both sulfur dioxides and nitrogen oxides.[1] Postcombustion absorbers capture the sulfur in a solid form for disposal. There are two basic approaches to capturing the sulfur: 1) divert the combustion gases through a scrubber unit that is separate from the combustor and 2) design the combustor so that the sulfur absorbent can be injected directly along with the fuel and sulfur is captured in the combustion chamber. Sulfur scrubbers have been widely used by the industry for some time and have been highly developed. Separate sulfur scrubber units have the advantage that they can be used to retrofit existing plants.[2] There are several types of scrubber available, and many different chemical reactions have been used for the extraction of sulfur oxides, as shown in Table 1.

Nitrogen Oxides

Nitrogen oxides are one of the primary sources of acid rain that are emitted during combustion. There are several different oxides produced, which are collectively referred to as NO_x. These oxides are generated from two mechanisms:[16]

1. *Thermal NO_x*. This is formed directly from the nitrogen in the combustion air because of the high temperatures and the presence of oxygen. Formation by this mechanism is strongly affected by temperature and residence time, with significant amounts produced at temperatures above 1200°C. Unfortunately, the factors that generally lead to complete combustion (high temperatures, long residence times, and thorough mixing of fuel and air) all tend to promote thermal NO_x production. NO_x from this source is usually controlled by some combination of reductions in flame temperatures, staged combustion, and flue-gas recirculation, which tend to slightly degrade the combustion efficiency.
2. *Fuel NO_x*. This comes from combustion of nitrogen that is contained in the chemical structure of the fuel. This can account for up to 50% of the NO_x from oil, and as much as 80% of the NO_x from coal. Conversion of the nitrogen that was chemically combined with the fuel into NO_x is strongly dependent on the fuel/air stoichiometry, but is relatively unaffected by combustion temperature, unlike thermal NO_x, which is strongly affected by combustion temperature. Fuel NO_x emissions are reduced by burning the fuel with low oxygen availability, causing the nitrogen to form N_2 in preference to NO_x.

Once the NO_x has been formed in the combustion process, if it is not removed it largely converts to NO_2. This is the source of the brownish plume often seen from power plant stack discharges. Unlike sulfur, it is not easily reacted with absorbents to form a solid sludge. Instead, it is catalytically reduced with ammonia or urea to form N_2 and water. This catalytic reduction requires injection of the reductant into the flue gases within a particular temperature window, with adequate residence time and catalytic surfaces to complete the reduction process. This can be integrated with scrubbers for other materials, such as sulfur and

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007955

Table 1 Absorbents that have been studied for removal of sulfur oxides from coal combustion gases

Absorbent	Regenerable?	Product	Reference
$CaCO_3$	No	$CaSO_3$ or $CaSO_4$	[3]
$Ca(OH)_2$	No	$CaSO_3$ or $CaSO_4$	
CaO	No	$CaSO_3$ or $CaSO_4$	
MgO	Yes, heat to release SO_2	SO_2	[4]
Na_2SO_3	Yes, heat to release SO_2	SO_2	[3]
$CeO_2 + Al_2O_3$	Yes, heat to release SO_2	SO_2	[5]
$Ca(OH)_2$ + fly-ash	No	Calcium silicate sulfates	[6]
$CuO + MnO_2$	Yes, heat with H_2 at 200–560°C	H_2S	[7]
$Ca(OH)_2$ + methanol	No	$CaSO_3$ or $CaSO_4$	[8]
CaO + MgO	No	$CaSO_3$ and $MgSO_3$	[9]
$NaHCO_3$	Yes, react with $Ca(OH)_2$	$CaSO_4$	[10]
Gaseous oxidation	Nothing to regenerate	H_2SO_4	[11]
$CaCO_3$ + NaCl	No	$CaSO_3$ or $CaSO_4$	[12]
Chalk	No	$CaSO_3$ or $CaSO_4$	[13]
Cement flue dust	No	Calcium silicate sulfates	
Alkali + Al_2O_3	Yes, heat with CO at 700–800°C	S, SO_2	[14]
Ca–Mg Acetate	No	$CaSO_3$ or $CaSO_4$	[15]

Some are low-cost throw-away absorbents, while others are regenerable and convert the sulfur oxides into marketable products.

fly-ash, which provide the necessary residence time and control of process conditions.

Carbon Dioxide

A variety of methods for sequestration of CO_2 from combustion of fossil fuels have been suggested in the literature to deal with the estimated 1583 million metric tons carbon equivalent annually released into the atmosphere in the U.S.A.[17,18] These methods include use of photoautotrophic organisms to convert it to biomass, deep-ocean disposal as a liquid, reaction with minerals to form stable carbonates, and industrial utilization. All of these sequestration methods can be made more efficient if the carbon dioxide is first separated from the flue gases and concentrated. The ideal approach would be for carbon dioxide to be absorbed from the flue gas, with the CO_2-laden absorbent then regenerated by a rapid, low-energy process, releasing the CO_2 in a purified, concentrated form. This CO_2 could then be sequestered or utilized by any appropriate method.

To date, all commercial plants for separation and concentration of CO_2 use chemical absorption with a monoethanolamine solvent. This solvent was developed over 60 years ago to remove acid gases, such as CO_2 and H_2S, from natural gas streams. Monoethanolamine absorption is popular for the existing markets for high-purity CO_2, because it produces a very high-grade product. The process has also been used to remove CO_2 from flue gases, but it had to be modified to resist solvent degradation and equipment corrosion. Also, to minimize reagent costs, the solvent strength was kept relatively low. This resulted in large equipment sizes and high regeneration energy requirements. Other CO_2-absorption strategies have tended to make use of advanced technologies such as pressure-swing absorption, membrane technology, and hollow-fiber permeators. These are all highly efficient at producing a high-purity CO_2 stream, but are expensive to apply on a large scale. As a result, the technologies that are receiving the most attention are not sufficiently economical for wholesale capture of CO_2 from fossil fuel combustion, and a lower-cost absorbent is needed for flue-gas CO_2 fixation applications.

A survey of the technologies used for capturing sulfur oxides has indicated that certain low-cost commodity chemicals used for capturing sulfur could also be useful for capturing carbon dioxide.[19] One of these absorbents can be used to produce an aqueous solution which will efficiently capture carbon dioxide at temperatures less than approximately 25°C, as shown in Fig. 1. This solution can then be completely regenerated at only 100°C, releasing concentrated carbon dioxide that can be easily utilized or permanently sequestered.[20]

Mercury/Heavy Metals

Even though coal is not enriched in mercury relative to the components of the rest of the Earth's crust,

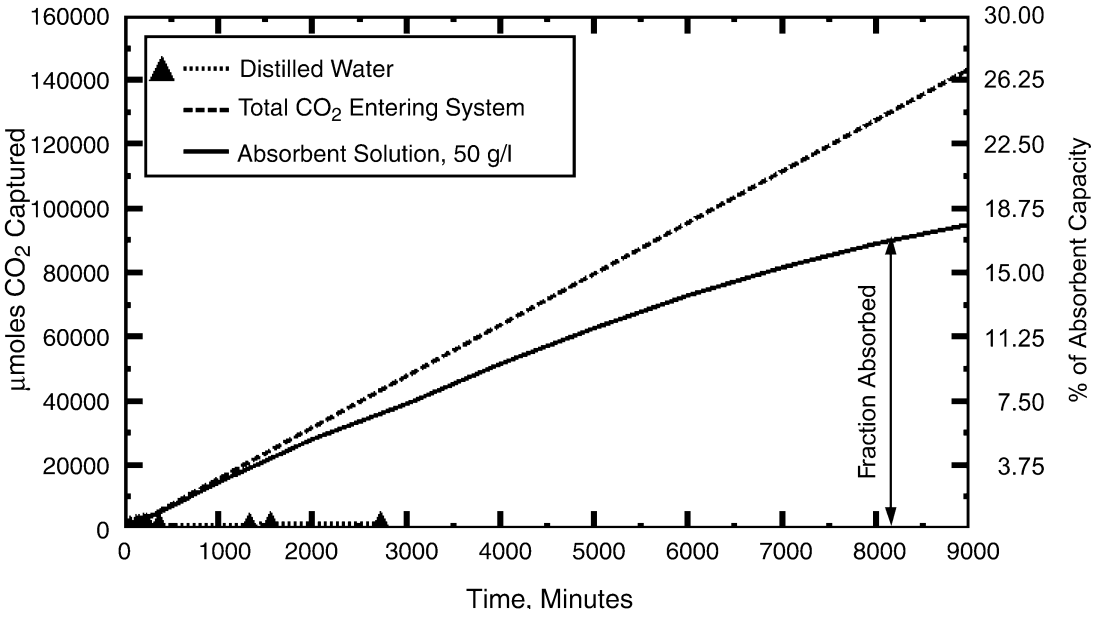


Fig. 1 Capture of carbon dioxide by a 50 g/L solution of absorbent in a 500 ml batch absorber. The gas being treated was ambient air, at a carbon dioxide concentration of 403 µmoles/mole and a flow rate of 0.96 slpm. The fact that this absorbent can reduce CO_2 levels to less than that of ambient air shows that a properly designed countercurrent scrubbing unit could use this solution to keep CO_2 levels in a combustion gas stream from exceeding the normal levels present in the atmosphere. It should be noted that, under these conditions, the ability of pure water to absorb carbon dioxide was negligible.

combustion of coal is nevertheless a significant source of mercury emissions. This is because of the high volatility of metallic mercury and its compounds, which are efficiently vaporized by the heat of coal combustion. As a result, mercury is emitted as a trace component at levels of a few parts per billion in coal combustion gases. As mercury is highly toxic and tends to accumulate in the ecosystem, it has become clear that coal-fired power plants must prevent even these highly diluted mercury emissions from escaping into the environment. However, it is highly uneconomical to have a dedicated scrubber that captures only mercury. This is because the need to contact the entire volume of gases with mercury absorbent results in the need for an absorbent mass that is thousands of times greater than the mass of the mercury captured. It is much more feasible to ''piggy-back'' the mercury capture capacity with scrubbers for other, higher-concentration pollutants produced by the plant. Possibilities include

1. Injection of activated carbon particles into the flue gas to absorb the mercury or operation of the coal combustion to convert a portion of the coal into activated charcoal. The mercury-laden charcoal is then captured by the fly-ash removal system (which may be an electrostatic precipitator, filter, or dust scrubber).
2. Capture of mercury in the sludge from a sulfur oxide scrubber.
3. Capture and sequestration along with carbon dioxide.

To capture mercury simultaneously with capture of sulfur oxides and carbon dioxide, it is necessary to ensure that the mercury is in a form that will be absorbed by the scrubber. The reducing environment of coal combustion results in a large fraction of the vaporized mercury being in the elemental state, which is unreactive, volatile, and can be transported worldwide over a period of years before it finally oxidizes and precipitates as a contaminant. It has been determined that, if mercury is in an oxidized state, it is readily captured by many existing scrubbers. This is because of oxidized mercury having a much higher solubility in water than elemental mercury, and a greatly increased reactivity with scrubbing agents. Therefore, the key to efficient capture of mercury is to ensure that it is in the oxidized state while it passes through the scrubber. This will require that the scrubber actively promote the oxidation of mercury. However, such oxidation cannot be achieved by the current generation of scrubbers.

WET SCRUBBERS

Wet scrubbers use a slurry or solution of a sulfur absorbent in water, which is generally contacted with the flue gases using a scrubber tower such as that

shown in Fig. 2. Flue gases rise through the tower through a falling spray of absorbent. The absorbent spray removes the sulfur oxides from the gases and collects in the base of the tower, where it is removed. The mist eliminator at the top of the column is to prevent fine droplets of absorbent from being carried up the stack, where they could cause corrosion or deposition problems in the stack or be released as particulate pollution.

In general, when designing a wet scrubber for capturing pollutants from gases, there are several steps that should be taken into consideration: 1) selection of absorbent material; 2) equilibrium data evaluation; 3) estimation of operating data, including mass and energy balances; 4) absorption column selection; 5) column diameter calculation; 6) estimation of column height and/or number of plates or other transfer units; 7) determination of pressure drop through the column.[21] In the case of wet flue-gas scrubbers, the units are fairly standardized and much of this work has already been done.

Absorbents

The most common absorbents are lime (calcium hydroxide) and limestone (calcium carbonate) slurries. Limestone is the preferred absorbent in many modern scrubbers, because of its low cost compared with lime and other absorbents.[22,23] However, lime is also used because of its higher reactivity, which allows it to absorb sulfur more rapidly. This makes it possible to use smaller scrubbers to treat a given quantity of gas when lime is the absorbent. When lime or limestone are used as the sulfur absorbents, the following reactions are believed to occur:[24]

- sulfur dioxide hydration:

$$SO_2(g) + H_2O(l) \Longrightarrow H_2SO_3(aq) \quad (1)$$

$$H_2SO_3(aq) \Longrightarrow H^+(aq) + HSO_3^-(aq) \quad (2)$$

- lime reactions:

$$Ca(OH)_2(s) \Longrightarrow Ca^{2+}(aq) + 2OH^-(aq) \quad (3)$$

$$Ca^{2+}(aq) + HSO_3^-(aq) + \frac{1}{2}H_2O(l) \Longrightarrow CaSO_3 \cdot \frac{1}{2}H_2O(s) + H^+(aq) \quad (4)$$

$$H^+(aq) + OH^-(aq) \Longrightarrow H_2O(l) \quad (5)$$

overall:

$$Ca(OH)_2(s) + SO_2(g) \Longrightarrow CaSO_3 \cdot \frac{1}{2}H_2O(s) + \frac{1}{2}H_2O(l) \quad (6)$$

- limestone reactions:

$$H^+(aq) + CaCO_3(s) \Longrightarrow Ca^{2+}(aq) + HCO_3^-(aq) \quad (7)$$

$$Ca^{2+}(aq) + HSO_3^-(aq) + \frac{1}{2}H_2O(l) \Longrightarrow CaSO_3 \cdot \frac{1}{2}H_2O(s) + H^+(aq) \quad (8)$$

$$H^+(aq) + HCO_3^-(aq) \Longrightarrow H_2CO_3(aq) \quad (9)$$

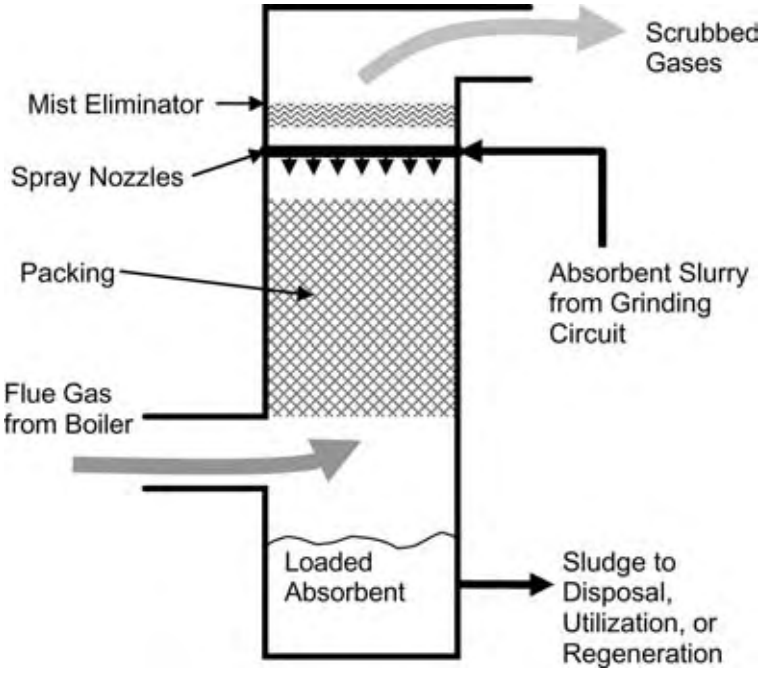


Fig. 2 Basic schematic of a wet scrubber column. Absorbent slurry percolates down through the packing, while the flue gases flow upward. The most common absorbents for sulfur oxides are limestone (calcium carbonate), lime (calcium hydroxide), and magnesium-enhanced lime made from dolomite. The sulfur-bearing sludge for some scrubbers is market-grade gypsum, but for other scrubbers it is a waste product that must be landfilled.

$$H_2CO_3(aq) \Longrightarrow CO_2(g) + H_2O(l) \quad (10)$$

overall:

$$CaCO_3(s) + SO_2(g) + \frac{1}{2}H_2O(l) \Longrightarrow CaSO_3 \cdot \frac{1}{2}H_2O(s) + CO_2(g) \quad (11)$$

The solid product from each of these sets of reactions is primarily calcium sulfite hemihydrate ($CaSO_3 \cdot \frac{1}{2}H_2O$), which has been confirmed by x-ray diffraction analysis of scrubber sludges.[25,26] A similar set of reactions collects sulfur trioxide (SO_3) from the flue gases, forming gypsum ($CaSO_4 \cdot 2H_2O$) as the solid product, but under normal boiler conditions sulfur trioxide makes up only about 0.5% of the total sulfur oxides, and so its removal is less important than the removal of sulfur dioxide.[24,27]

Equilibrium and Operating Data

The factor that determines the size of a wet scrubber needed to produce a given capacity is the mass transfer rate of sulfur dioxide from the gas phase to the liquid phase, as shown in Eq. (1).[16] This mass transfer rate can be expressed as:

$$N_g = \int \frac{dy}{y - y^*} = \int \frac{k_g a}{G} dV \quad (12)$$

where G is the molar gas flow rate (moles/s), y the mole fraction of sulfur dioxide in flue gas, k_g the gas film mass transfer coefficient (moles/m^2s), a the interfacial surface area (m^2/m^3), y^* the equilibrium sulfur dioxide concentration at the gas/liquid interface, V the volume of the gas/liquid intermixing region (m^3), and N_g the number of gas-phase transfer units (dimensionless).

As not all of the parameters needed can be calculated in advance (k_g can only be approximated, and a must be determined experimentally), the gas-phase mass transfer rate for a given scrubber design must be determined experimentally by operating the scrubber under conditions where y^* approaches zero. Under these conditions, Eq. (12) can be integrated to give:

$$N_g = -\ln(1 - E) = k_g a \frac{V}{G} \quad (13)$$

where E is the overall sulfur dioxide absorption fractional efficiency.

The number of gas-phase transfer units that exist in a given column design depends on a number of factors, including: 1) slurry spray rate; 2) droplet size and distribution; 3) gas-phase residence time, which is controlled by the height of the spray zone; 4) liquid residence time; 5) wall effects; and 6) gas flow distribution.[16]

In a limestone-based wet scrubber, the dissolution of calcium carbonate [Eq. (7)] is the primary rate-limiting reaction, because of the low solubility of calcium carbonate. The rate for the dissolution reaction can be expressed as:

$$\frac{d[CaCO_3]}{dt} = k_c([H^+] - [H^+]_{eq}) \cdot S_{spc}[CaCO_3] \quad (14)$$

where $[CaCO_3]$ is the calcium carbonate concentration in the slurry (moles/l), $[H^+]$ the hydrogen ion concentration (moles/l), $[H^+]_{eq}$ the equilibrium hydrogen ion concentration at the limestone surface (moles/l), S_{spc} the specific surface area of the limestone in the slurry, and k_c the reaction rate constant.

The limestone dissolution rates at various pH values and partial pressures of carbon dioxide are shown in Fig. 3.

Slurry–Gas Contact

The heart of a scrubber column is the slurry–gas contact zone, where gases are intimately combined with the absorbent slurry so that the pollutants can be captured by the reactions given previously. There are a number of possible methods for designing the contacting zone, including sprays, crossflow plates, baffle plates, counterflow plates, and packed columns. These all have the purpose of maximizing the interfacial area between the gas phase and the liquid phase, to allow rapid transport of gases across the surface.[29]

A serious problem occurs in many wet scrubbers if the sulfur dioxide is partially oxidized to sulfur trioxide. In this case, the main precipitate is calcium sulfite hemihydrate, with up to 15% calcium sulfate in solid solution in the sulfite particles. If more than 15% is as calcium sulfate, then it can no longer precipitate with the sulfite crystals, and instead precipitates as separate crystals of gypsum. However, there is a shortage of gypsum seed crystals in the slurry in this situation, and so much of the gypsum crystallizes in the scrubber, particularly in the slurry–gas contact zone. This can rapidly plug the scrubber and must be avoided. Many plants prevent this problem by adding thiosulfate ($S_2O_4^{-2}$) to the scrubber slurry as a reducing agent. This prevents the oxidation of sulfur dioxide and thus eliminates the formation of gypsum and the buildup of gypsum scale in the scrubber. A second solution to the plugging problem is to completely oxidize the calcium sulfite to gypsum, which provides more seed crystals for the gypsum and also prevents plugging.

When the solid sludge is removed from the scrubber as unoxidized calcium sulfite, as is done in many older

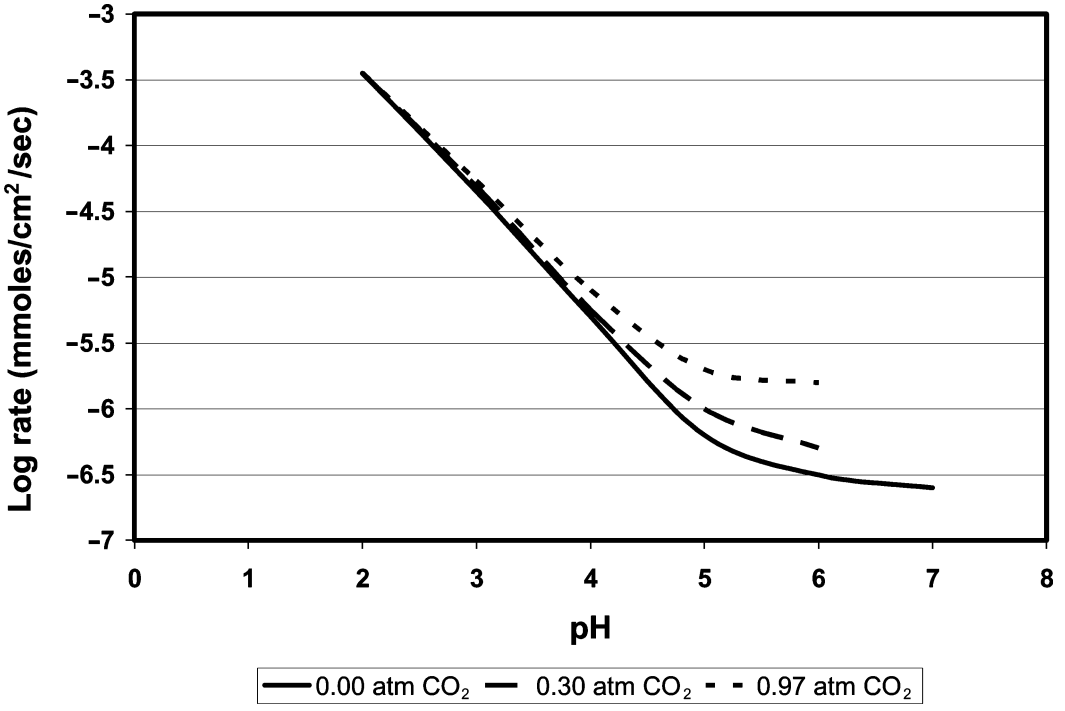


Fig. 3 Limestone dissolution rates as a function of pH and carbon dioxide partial pressure, at a temperature of 25°C. (From Ref.[28].)

scrubbers, it has no market value and must be disposed of by landfilling. The more advanced scrubbers include an oxidation step, which converts the sulfite to sulfate, and the solid product is then gypsum.[23] If it is sufficiently pure, this synthetic gypsum can be marketed to make plaster, wallboard, cement, and other construction products. The major barrier to widespread marketing of scrubber sludge is that it contains a number of impurities, and is of uneven quality, which makes it unattractive for most purposes in its raw form. Potential users are therefore not eager to purchase the material, even with significant price breaks.[30,31] Typical quality requirements for gypsum for use in wallboard manufacture are given in Table 2. For the gypsum to be salable, it should meet or exceed these requirements.

A complete circuit for an advanced scrubber is shown in Fig. 4, which includes oxidation of the sludge to form gypsum.[9] In this circuit, limestone is first reduced to a fine particle size by a grinding mill, producing a slurry. The slurry is then added to the absorber tank, and pumped into the scrubber tower. A portion of the descending absorbent is diverted back to the absorber tank, which provides more time for the sulfur dioxide and limestone to react. The remaining absorbent collects in the base of the tower, where it is oxidized by injected air while being recirculated in the lower portion of the scrubber. A portion of the absorbent is continuously drawn off to a hydrocyclone, which separates the gypsum particles from the absorbent slurry, and returns the liquid to the scrubber.

The most efficient wet-scrubber technology, from the standpoint of sulfur removal efficiency and equipment size, is the magnesium-enhanced lime process. This type of scrubber uses lime that contains up to 12% magnesium, which increases the absorption capacity of the lime to approximately 10–15 times that of the limestone scrubbers described previously. The principal advantage is that the magnesium-enhanced lime is soluble enough that the SO_2 removal is governed by the degree of gas–liquid contact in the

Table 2 Impurity limits for gypsum for use in plaster or wallboard manufacture

Impurity	Maximum wt.%
Fe_2O_3	1.5
SiO_2	1.0
MgO	0.1
K_2O	0.1
Na_2O	0.04
Cl	0.01
CO_3	1.5
SO_2	0.25
Moisture	8.0

(From Ref.[32].)

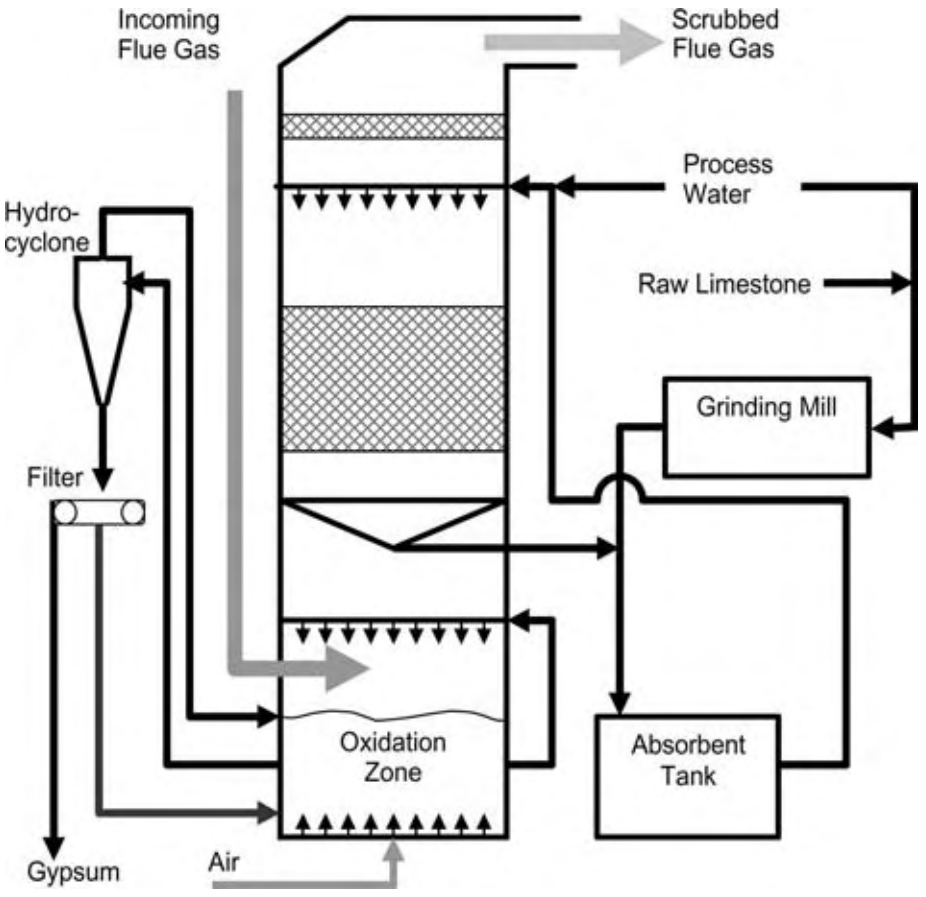


Fig. 4 Circuit for a wet limestone scrubber, with oxidation of the solids to gypsum. The absorbent tank simplifies control of the process, while the hydrocyclone and filter remove coarse gypsum particles.

scrubber, and not by the degree of absorbent dissolution as is the case with limestone. A disadvantage is that the magnesium-enhanced lime is comparatively expensive because it must be calcined by heating before use. Also, the magnesium specifically inhibits the formation of gypsum, which helps to prevent plugging and scaling, but also results in the sludge being an unmarketable sulfite sludge instead of gypsum. Finally, the magnesium content of the sludge is too high for use in most synthetic gypsum markets.

Power Consumption Example

Assume a 180 MW boiler burning coal with 2.5% sulfur by weight, and a heating value of 12,767 BTU/lb. An appropriate limestone scrubber with forced oxidation would operate with a liquid/gas ratio of 130 gal. liquid per 1000 ft.^{3} of flue gas, and a pressure drop of 5 in. water. Such a scrubber would consume 2.549 MW to operate, with the breakdown as shown in Table 3. This corresponds to 1.42% of the total power output of the plant. Such a scrubber would remove approximately 93% of the sulfur, while consuming approximately 13,000 lb/hr of limestone being added at 35% solids.[16]

Other Types of Wet Scrubber

In addition to lime and limestone, a number of other absorbents have been used to improve the efficiency of sulfur removal or to recover the sulfur in a marketable form while regenerating the absorbent. Next-generation scrubbers are therefore under development to improve the efficiency and reduce the quantity of unmarketable waste products.[33–35] Several of the scrubber technologies that use other absorbents are listed here:

- *Dual alkali process.* In this process, the absorption of the sulfur dioxide is first carried out using a solution of a sodium alkali, such as NaOH, Na_2CO_3, or Na_2SO_3. Because these are all very soluble in water, they can absorb the sulfur dioxide very rapidly and completely, and can be easily oxidized afterward. Also, the absorbent is a clear liquid rather than a slurry, and so the problems with scaling and plugging of the scrubber are much reduced.[10] The oxidized sulfur-bearing alkali is then circulated to a vessel where it reacts with lime or limestone, which precipitates the sulfur as calcium sulfate and regenerates the sodium alkali. A flow diagram of the

Table 3 Typical power requirements for limestone scrubber with forced oxidation of sludge to gypsum

	Average power (KW)
Absorber system	
Oxidation air blower	375
Absorber recirculation pump 1	312
Absorber recirculation pump 2	367
Absorber recirculation pump 3	380
Absorber recirculation tank agitators	60
Mist eliminator wash water pump	19
Misc. pumps and agitators	24
Subtotal	1537
Dewatering area	
Vacuum pump for filter	55
Filter wash water tank heater	16
Reclaim water pump	14
Hydrocyclone overflow pump	15
Filter feed tank agitator	7
Clarifier overflow sump pump	6
Misc. pumps and agitators	13
Subtotal	126
Reagent preparation	
Ball mill drive	220
Mill product tank pump	5
Limestone feed tank agitator	25
Misc. pumps and agitators	6
Subtotal	256
Other systems	
General instrument air	50
Differential induced draft fan power	580
Subtotal	630
Total power used by scrubber system	2549

(From Ref.[16].)

process is shown in Fig. 5. The dual-alkali process is reported to be stable and resistant to disturbances, and to be capable of removing more than 99% of the sulfur dioxide from flue gases.[10,36]

- *Wellman–Lord process.* This is a regenerable-sorbent process, producing SO_2 gas, which can be sold for industrial uses. It uses a solution of sodium sulfite (Na_2SO_3), which absorbs SO_2 and becomes a sodium bisulfite solution ($NaHSO_3$). The sodium bisulfite is then decomposed in a forced circulation evaporator, releasing the SO_2 at sufficiently high concentration to be compressed and sold as SO_2 gas, or used for producing elemental sulfur or sulfuric acid.[3]
- *Magnesium oxide process.* The magnesium oxide slurry is used to collect SO_2, and the resulting magnesium sulfite is thermally treated to release the SO_2 and regenerate the absorbent, as shown in Fig. 6. Like the Wellman–Lord process, this process is relatively complex and has a capital cost about 14% higher than that of limestone scrubbers.[4] It is therefore only economically viable when there is a reliable market for the by-products.[38]

SPRAY-DRY SCRUBBERS

Spray-dry scrubbers are an alternative to conventional wet scrubbers. In this type of scrubber, an alkaline slurry or solution is sprayed in fine droplets into a reaction vessel, along with the flue gas. The droplets rapidly react with the sulfur dioxide while drying to a fine powder of sulfite salts. This powder is entrained in the gas stream, and is carried to a dust precipitator where it is collected, as shown in Fig. 7. Most of the sulfur dioxide is collected in liquid-phase reactions while the droplets are drying, but 10–15% additional sulfur dioxide can be absorbed in gas/solid reactions, as the absorbent powder is swept through the ductwork and particulate collector. These are cocurrent devices, and so the limestone utilization and sulfur removal efficiency are inherently lower than those of countercurrent devices such as wet scrubbers. Partial recycle of the sorbent is often used to improve the sorbent utilization.

It is typical to install the spray-dryer before the plant fly-ash collector, so that the existing dust control equipment can be used to collect the used absorbent. Slaked lime [$Ca(OH)_2$] is the most common absorbent, although sodium carbonate (Na_2CO_3) is used in some plants. Spray-dryers have also been used with regenerable magnesium oxide absorbent.[4]

Spray-dryers are simpler and more compact than conventional wet scrubbers and have a lower capital and operating cost. Also, they do not produce large quantities of wastewater, and the spent absorbent is dry, thereby eliminating the need for thickening and filtration of the sludge. However, if the same dust precipitator is used for both the fly-ash and the spray-dryer product, the mixture of fly-ash and spent absorbent that they produce is unmarketable, and must be disposed of. Also, they require more expensive absorbents than conventional wet scrubbers. They are most suitable for retrofitting small plants that burn medium-sulfur coals, where capital costs and space restrictions are more of a consideration.[38]

VENTURI SCRUBBERS

Venturi scrubbers are mainly used for collecting fine particulates (such as fly-ash) from gas streams,[39] but they have also been adapted for absorption of sulfur dioxide.[40] These units are mechanically very simple, consisting of a reducing inlet with liquid sprays,

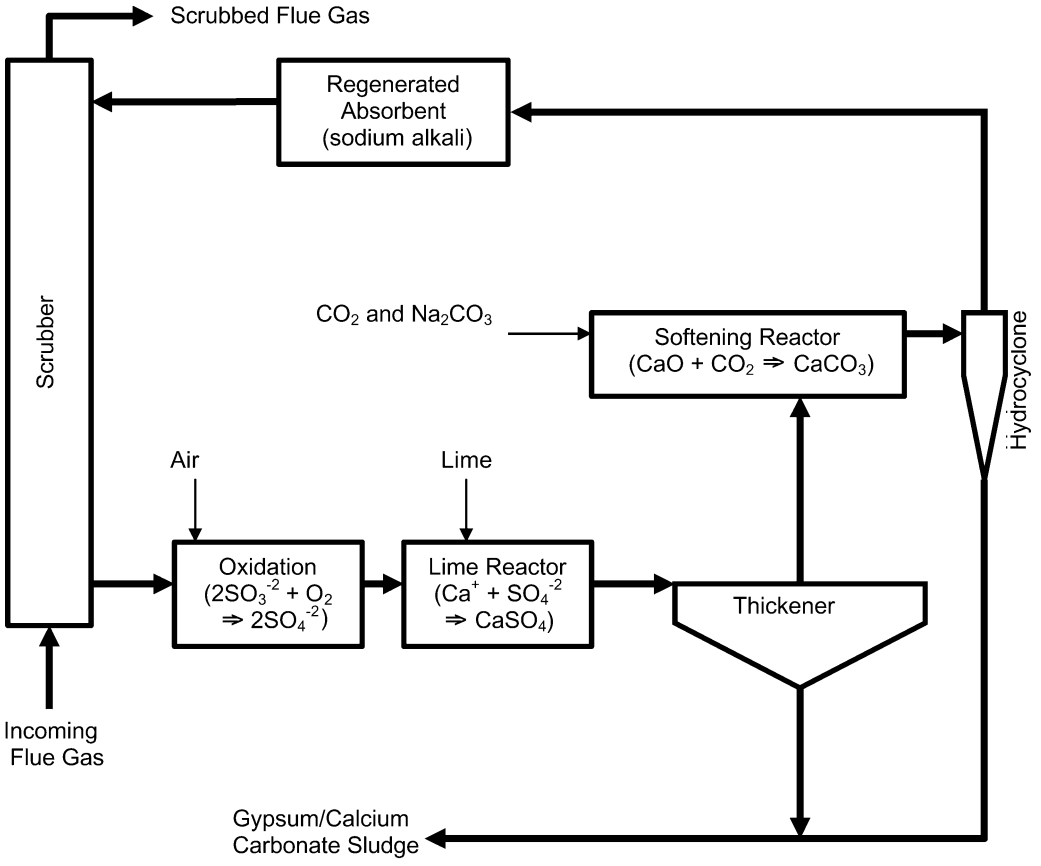


Fig. 5 Flow diagram of the dual-alkali scrubber process, using lime to regenerate the sodium alkali. The clarified liquid from the thickener contains dissolved calcium sulfate, which would produce calcium carbonate scale in the scrubber when it contacts the carbon dioxide in the flue gas. It is therefore precipitated in the softening reactor by a combination of carbon dioxide and sodium carbonate, and the resulting calcium carbonate precipitate is removed by the hydrocyclone.

a narrow throat where the gas/liquid contact occurs, and an expanding region, as shown in Fig. 8. The flue gases are injected into the venturi at the contracting inlet, along with an absorbent, such as lime slurry. The gas accelerates to high speed as it enters the throat, and atomizes the absorbent, providing good gas/liquid contact. The gas and the atomized liquid then expand and slow in the expander region and are then diverted to a mist eliminator to separate the liquid droplets from the scrubbed gases.

Venturi scrubbers have lower capital cost than other types of scrubber because they are mechanically simple; but they have a high energy consumption because of the need for pressurizing the gas to force it through the venturi. They also double as a fly-ash collection device, and so there is no need for separate scrubbers and fly-ash collectors when these units are used.[40] Because venturi scrubbers are cocurrent devices, with both the flue gas and absorbent traveling in the same direction, they cannot remove sulfur dioxide as completely as countercurrent devices, such as wet-scrubber towers, do.[29,41] They also produce a wet mixture of fly-ash and alkaline absorbent, which is unmarketable and can form a cement-like substance upon disposal.

A number of theoretical studies of venturi performance have been made to produce theoretical models that can predict performance from first principles. One of the key areas of uncertainty has been the droplet size formed by the venturi. Typically, this is estimated using the Nukiyama and Tanasawa equation to estimate the surface-mean droplet diameter:[39]

$$D_0 = \frac{1920\sqrt{\sigma_L}}{V_0\sqrt{\rho_L/62.3}} + 75.4\left(\frac{\mu_L}{\sqrt{\sigma_L\rho_L/62.3}}\right)^{0.45}\left(\frac{1000Q_L}{Q_G}\right)^{1.5} \quad (15)$$

where D_0 is the drop diameter (μm), V_0 the gas velocity (ft/s), σ_L the liquid surface tension (dynes/cm), ρ_L the liquid density (lb/ft^3), μ_L the liquid viscosity (centipoises), Q_L the liquid flow rate (ft^3/s), and Q_G the gas flow rate (ft^3/s).

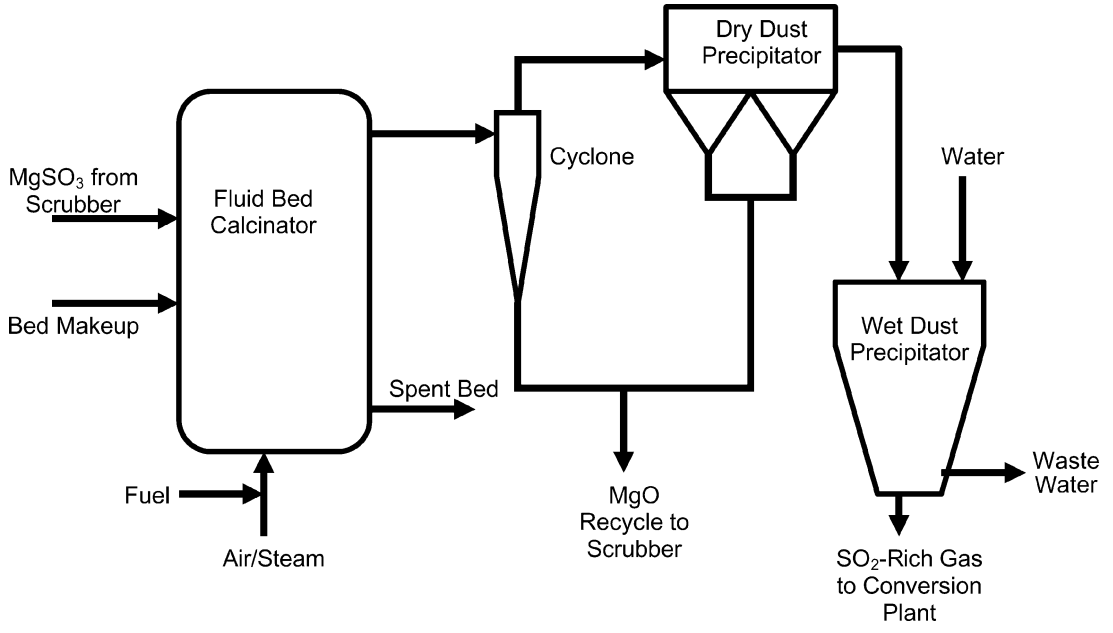


Fig. 6 Magnesium oxide regeneration and sulfur dioxide recovery section for a magnesium oxide scrubber. (From Ref.[37].)

This equation unfortunately has a number of severe shortcomings, including its lack of dimensional homogeneity and the fact that the effect of nozzle size on the drop diameter is not defined. It was originally derived for small diameter atomizing nozzles, and drop sizes reported by various investigators have varied from the equation prediction by as much as two to three times, and so this approach has not been highly successful.

In general, the ability of these scrubbers to collect particles improves, as the energy input to the system increases, and so collection of very fine particles requires increased flow rates and operating pressures. This has lead to a fairly useful scrubber design method based on the dissipation of power in the gas–liquid contactor. A number of studies have concluded that the collection efficiency of a scrubber on a given dust

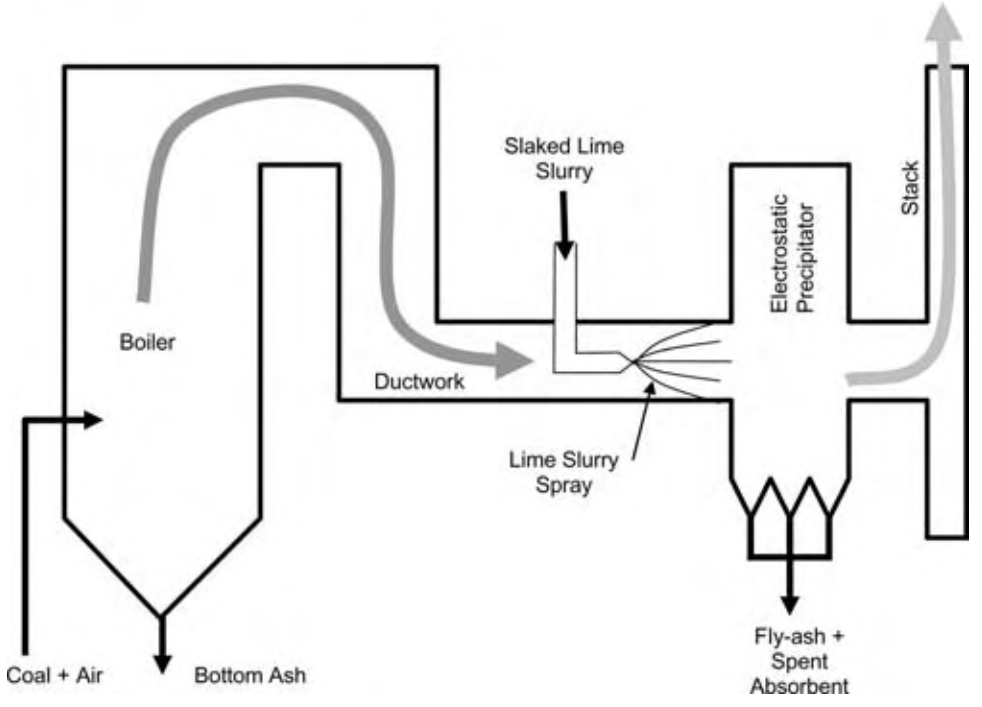


Fig. 7 Basic configuration for a single-stage spray-dry sulfur absorber, with no recycle of absorbent.

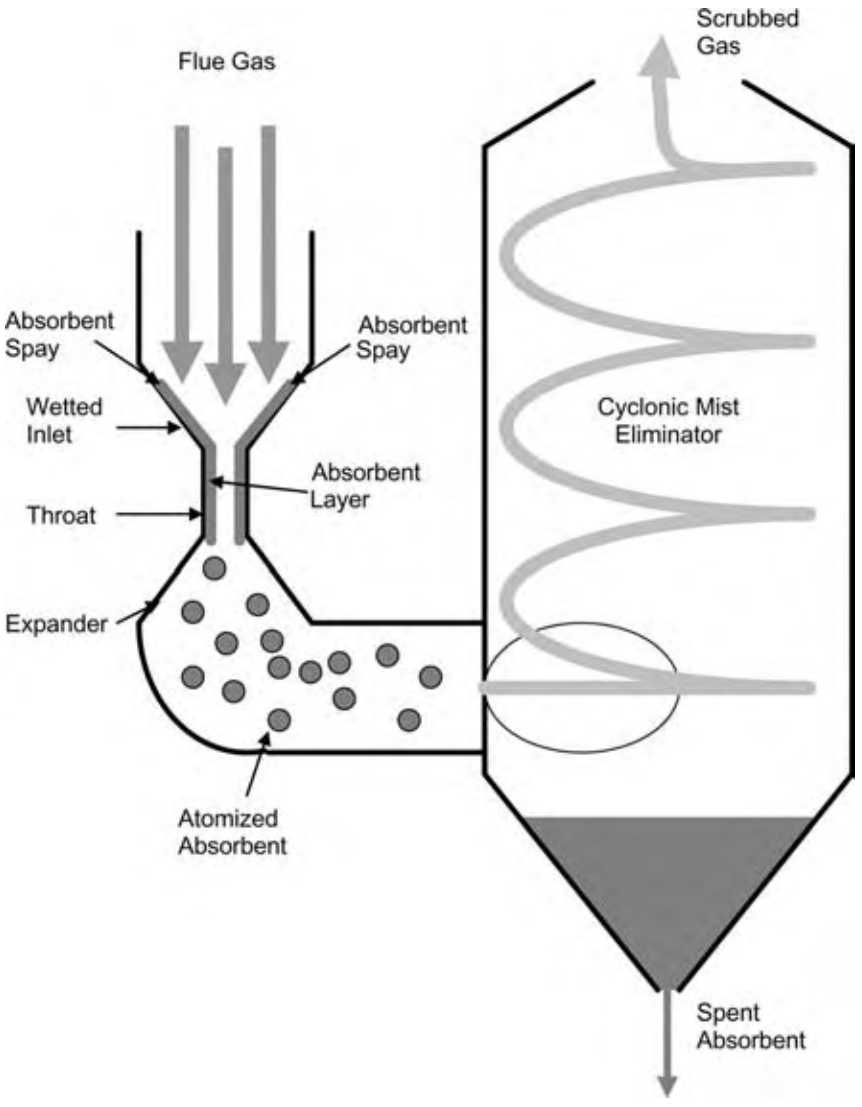


Fig. 8 Schematic diagram of a venturi scrubber with a cyclonic mist eliminator. The wetted inlet, throat, and expander make up the venturi section.

is dependent only on the contacting power, with only minor effects from the size or geometry of the scrubber.[39] The contacting power is commonly given in units of MJ/1000 m^3, and may conveniently be broken into three parts:

1. Gas-phase contacting power, P_G. In SI units, this is equal to the effective friction loss of the gas as it flows through the contactor, in kPa.
2. Liquid-phase contacting power. This is given by the equation $P_L = p_f(Q_L/Q_G)$, where p_f is the nozzle feed pressure in kPa, and Q_L and Q_G are the liquid and gas flow rates, respectively, in m^3/s.
3. Additional power supplied separately, P_M, such as by a power-driven rotor.

The total contacting power is then $P_T = P_G + P_L + P_M$. The collection efficiency, expressed as the number of equivalent contacting units (N_t), then becomes:

$$N_t = aP_T^{\gamma} \tag{16}$$

where α and γ are empirical constants that depend on the character of the particles being collected.[39]

DRY SORBENT INJECTION

Dry sorbent injection is very similar to the use of spray-dryers, except that the sorbent is injected as a dry powder rather than as an atomized slurry.[42,43] The most common sorbent is hydrated lime, but other sorbents can also be used. The sorbent is usually injected directly into existing ductwork, and so the amount of space required is negligible compared with that of other flue-gas desulfurization processes. This makes dry sorbent injection a very low-cost option.

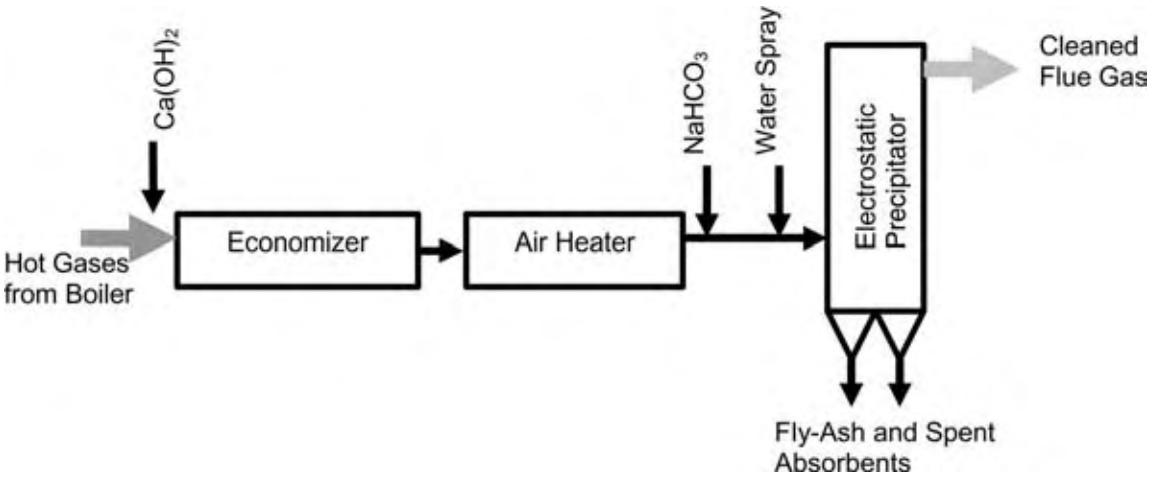


Fig. 9 Integrated dry injection process utilizing both calcium hydroxide and sodium bicarbonate to reduce sulfur dioxide and nitrogen oxide emissions. Calcium hydroxide is added to the hot flue gases before they are cooled in the economizer and combustion air heater, while sodium bicarbonate is added to the cooled gases before they enter the electrostatic precipitator. (From Ref.[44].)

Unfortunately, the reactivity of dry absorbents is much lower than that of absorbent slurries or solutions, and so dry sorbent injection is only suitable for applications where less than 70% of the sulfur dioxide needs to be removed from the flue gases.[38]

In cases where hydrated lime does not remove enough of the sulfur dioxide, but economics makes more efficient absorbents impractical, a two-stage absorbent injection scheme can be used, as shown in Fig. 9. Here, the relatively low-cost calcium hydroxide is used to remove the bulk of the sulfur dioxide, and is then followed by a spray of more effective (but higher-cost) sodium bicarbonate. In addition to further reduction of the sulfur dioxide content, the sodium bicarbonate spray also reduces the content of nitrogen oxides.

It is also possible to use limestone in dry sorbent injection, as is done in the limestone injection multistage burner (LIMB) system (Fig. 10). In this system, pulverized limestone is injected into the boiler directly, where the temperature is high enough to flash-calcine

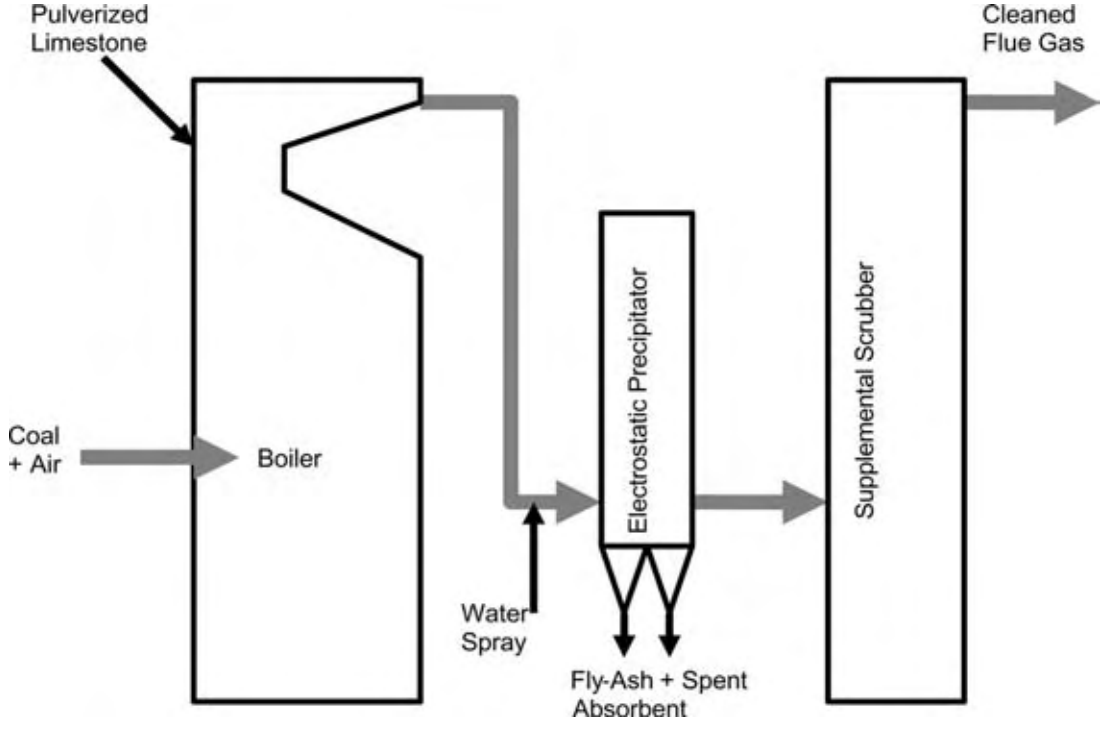


Fig. 10 The basic integrated LIMB dry sorbent injection system.

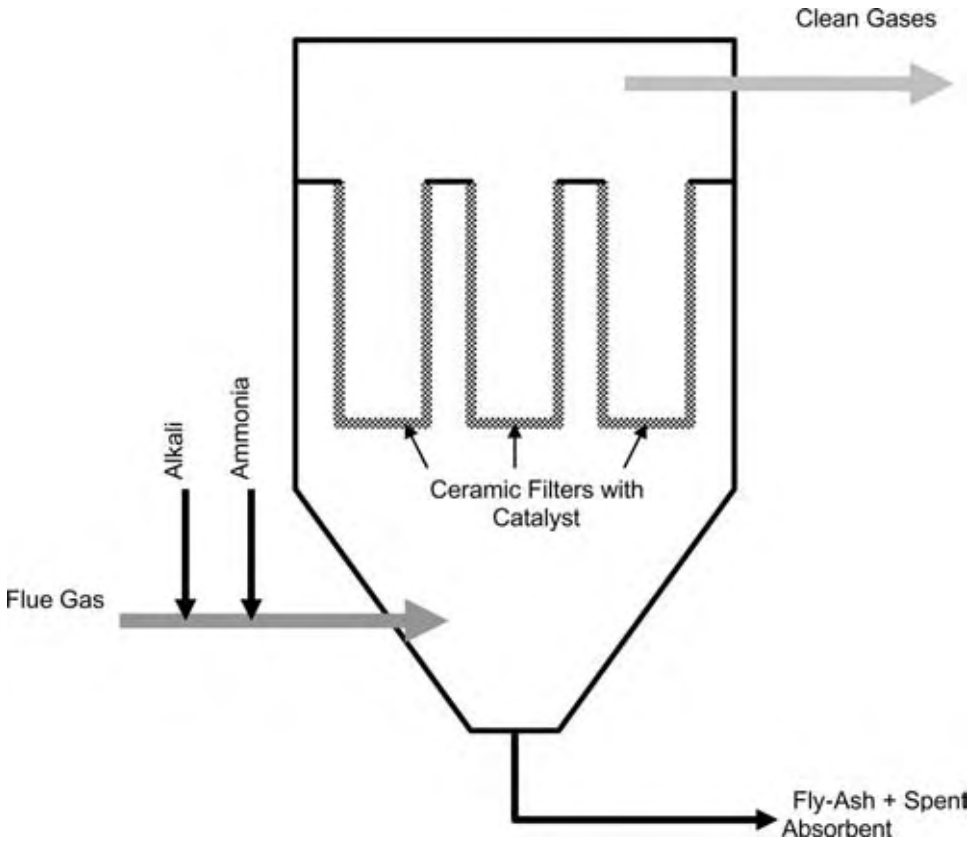


Fig. 11 Schematic of the SNRB catalytic baghouse. (From Ref.[46].)

the $CaCO_3$ to CaO. The CaO dust is carried off with the flue gases until the temperature drops enough for $CaSO_3$ to become stable. The CaO then captures the SO_2 to form $CaSO_3$, which is removed by the electrostatic precipitators, along with the fly-ash. If necessary, a second, wet SO_2 scrubber is used to finish the sulfur removal.[37]

A related technology, the SOx-NOx-Rox Box™, or SNRB, also destroys nitrogen oxides while removing sulfur dioxide.[45,46] This unit is a replacement for electrostatic precipitators and is installed in the flue-gas stream between the economizer and the combustion air heater, where the flue gases are still hot. An alkali is injected as sulfur absorbent, along with anhydrous ammonia. The gases then enter the high-temperature catalytic baghouse, which consists of catalyst-impregnated ceramic filter ''bags,'' as shown in Fig. 11. The ceramic filters capture and remove the sulfur-loaded absorbent, and the catalyst in the filters catalyzes the reduction of the nitrogen oxides by the ammonia, producing nitrogen gas and water.

SNOX SYSTEM

The SNOX system is designed for removing both sulfur oxides and nitrogen oxides from flue gases, and is unusual in that it does not use an alkali as an absorbent to collect the sulfur dioxide. Instead, it oxidizes the sulfur dioxide to sulfur trioxide and uses a special condenser to collect the sulfur trioxide as marketable sulfuric acid. This is combined with catalytic destruction of the nitrogen oxides with ammonia, producing the overall circuit shown in Fig. 12. The system is reported to be capable of better than 90% removal of both SO_2 and NO_x, while producing sulfuric acid at 93–95% concentration.[11]

RELATIVE COSTS OF SCRUBBERS

A comparison of the estimated sulfur removal ability and costs for various postcombustion processes is given in Table 4. It can be seen from this table that there is considerable variation in costs, depending on factors such as the size of the plant, fraction of the time that the plant operates at full capacity, ability to retrofit existing facilities, differences in fuel prices, sulfur content of the coal, and ability of existing equipment such as electrostatic precipitators to cope with changes in the process. Other techniques for reducing sulfur emissions, such as fuel switching or advanced combustion technologies, are included for comparison.

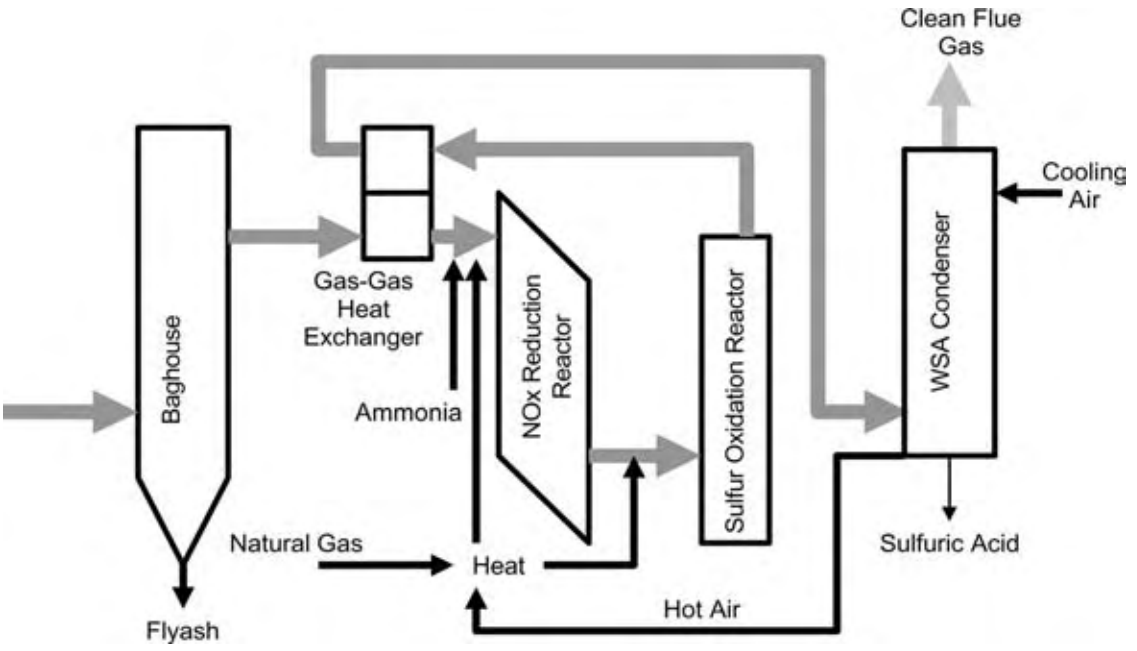


Fig. 12 Diagram of the SNOX system for production of sulfuric acid from flue gases while simultaneously destroying nitrogen oxides. (From Ref.[11].)

In general, fuel switching is currently the cheapest option at the current price difference of $10/ton between high-sulfur and low-sulfur coals. However, as low-sulfur coal prices increase, this will become less economical. Dry processes are the next most economical option, but they may not be able to meet emission control standards. The various types of scrubber have broadly similar costs, and the choice will depend on the specific plant under consideration. It appears from Table 4 that advanced combustion technologies such as integrated gasification combined cycle (IGCC) and fluidized bed combustion are drastically more expensive than scrubbers, but it should be remembered that the comparison is between retrofitting existing plants (for the scrubbers) and building new plants (for the IGCC and fluidized bed). The price differential between the two types of pollution control process is much less for completely new plants or for repowering existing plants.[38]

In general, the performance of a scrubber process will be improved if the sulfur content of the feed coal

Table 4 Cost comparison of various SO_2 control options

Technology	% SO_2 removal	Capital costs ($/kW)			Operating costs (mills/kWhr)			Most sensitive parameters
		Low	Base	High	Low	Base	High	
Fuel switching/blending	2–80	20	28	30	3	6	13	%S, CF, FPD, SCA
Lime/limestone FGD	90	120	240	520	5	16	150	MW, RF, CF, %S
Lime spray drying with existing ESP	76	70	170	540	3	10	130	MW, RF, CF, %S, SCA
Lime spray drying with new fabric filter	86	140	240	620	5	13	150	MW, RF, CF, %S
Integrated gasification combined cycle	95	1710	2100	2800	44	91	605	MW, CF, heat rate
Atmospheric fluid bed combustion	90	1360	1680	2250	40	80	480	MW, CF, heat rate
Dry sorbent injection	70	25	50	110	2	6	40	MW, CF, %S, SCA

CF, capacity factor; FPD, fuel price differential; MW, size of plant(MW); %S, fuel sulfur content; RF, retrofit factor; SCA, specific collection area of electrostatic precipitator (ESP).
(From Ref.[47].)

is reduced. A lower sulfur feed provides the following benefits:

1. Gas-scrubbing processes require less absorbent.
2. A lower percentage of the total SO_2 can be removed and still reach emission targets, allowing the less-efficient, lower-cost technologies to be used.
3. The quantity of desulfurization waste which must be disposed of is reduced.

It is therefore beneficial to pretreat the coal to remove as much sulfur as is practical before combustion, so that the expense of postcombustion desulfurization can be reduced. Combined with the other benefits of coal cleaning, it is evident that precombustion coal treatment is valuable even when it is not sufficient to completely desulfurize the coal by itself.

CONCLUSIONS

A wide range of postcombustion technologies are available for reducing the emissions of sulfur and other pollutants from coal-fired power plants. These include both wet and dry scrubbers, which produce either sulfur-bearing wastes or marketable by-products. In general, the less-expensive techniques are also the least effective for reducing emissions, and so the choice of which process to use depends on the quantity of pollutants in the fuel, the types of pollutants, and the emissions target, as well as on many other factors such as plant size, whether it is a retrofit of an existing plant or a new plant, availability of cleaner fuel, cost of waste disposal, and availability of markets for by-products.

Spray-dryers or dry sorbent injection are a good choice if there is little space in the plant for installing new equipment, there is no market for by-products, and the plant emissions can be brought into regulatory compliance with relatively modest reductions in emissions.

Wet scrubbers are well suited for plants with significant room for expansion, and which need to remove a large proportion of the pollutants from their emissions. Scrubbers with throw-away absorbents are the best choice when waste disposal is cheap, while those with regenerable absorbents and/or marketable by-products are the best choice when waste disposal is expensive and markets for by-products are nearby.

Precombustion and postcombustion pollutant removal technologies should not be considered to be in competition. Rather, they are complementary technologies that should be used together for maximum benefit. Many precombustion processes are low in cost, but cannot remove all of the pollutants, while postcombustion treatment can often capture nearly all of the emissions from the combustion gases, but their costs increase as the coal pollutant content increases. By using both types of process together, the maximum reduction of emissions can be achieved at the minimum cost.

REFERENCES

1. Anonymous. Will combined SO_2/NO_x processes find a niche in the market? Power **1990**, *134*, 26–28.
2. Ellison, W. Today's FGD systems satisfy retrofit needs for 1990s. Power **1991**, *135*, 101–106.
3. Couch, G.R. *Power from Coal—Where to Remove Impurities? Report No. IEACR/82*; IEA Coal Research: London, 1995.
4. Burnett, T.A.; Wells, W.L. Conceptual design and economics of and improved magnesium oxide flue gas desulfurization process. In *Flue Gas Desulfurization*; Hudson, Wells, Eds.; American Chemical Society: Washington, DC, 1982; 381–411.
5. Hedges, S.W.; Yeh, J.T. Kinetics of sulfur dioxide uptake on supported cerium oxide sorbents. Environ. Progr. **1992**, *11* (2), 98–103.
6. Kind, K.K.; Wasserman, P.D.; Rochelle, G.T. Effects of salts on preparation and use of calcium silicates for flue gas desulfurization. Environ. Sci. Technol. **1994**, *28* (2), 277–283.
7. Bjornbom, E.N.; Druesne, S.; Zwinkels, M.F.M.; Jaras, S.G. Study on the regeneration of copper–manganese sorbent for removal of sulfur dioxide from flue gases. Ind. Eng. Chem. Res. **1995**, *34* (5), 1853–1858.
8. Withum, J.A.; Yoon, H. Treatment of hydrated lime with methanol for in-duct desulfurization sorbent improvement. Environ. Sci. Technol. **1989**, *23* (7), 821–827.
9. Makansi, J. Controlling SO_2 emissions. Power **1993**, *137*, 23–56.
10. Valencia, J.A. The limestone dual alkali process for flue gas desulfurization. In *Flue Gas Desulfurization*; Hudson, Wells, Eds.; American Chemical Society: Washington, DC, 1982; 325–347.
11. Durrani, S.M. The SNOX process: a success story. Environ. Sci. Technol. **1994**, *28* (2), 88A–90A.
12. Bulewicz, E.M.; Janicka, E. Catalytic effect of NaCl on flue-gas desulphurization by limestone-based sorbents during the FB combustion of coal. J. Inst. Energy **1990**, *63*, 124–130.
13. Dennis, J.S.; Hayhurst, A.N. Alternative sorbents for flue-gas desulphurization, especially in fluidized-bed combustors. J. Inst. Energy **1989**, *62*, 202–207.

14. Gavalas, G.R.; Edelstein, S.; Flytzani-Stephanopoulous, M.; Weston, T.A. Alkali-alumina sorbents for high-temperature removal of SO_2. AIChE J. **1987**, *33* (2), 258–266.
15. Levendis, Y.A.; Zhu, W.; Wise, D.L.; Simons, G.A. Effectiveness of calcium magnesium acetate as an SO_x sorbent in coal combustion. AIChE J. **1993**, *39* (5), 761–773.
16. Babcock; Wilcox. Nitrogen oxides control, Chapter 34, Sulfur dioxide control, Chapter 35. In *Steam, Its Generation and Use*, 40th Ed.; The Babcock and Wilcox Company, 1992; 35-1–35-15.
17. EIA. Emissions of Greenhouse Gases in the United States 2000. Energy Information Administration, DOE/EIA-0573 (2000), 2001. ftp://ftp.eia.doe.gov/pub/oiaf/1605/cdrom/pdf/ggrpt/057300.pdf.
18. Herzog, H.; Drake, E.; Adams, E. CO_2 Capture, Reuse, and Storage Technologies for Mitigating Global Climate Changes. White Paper—Final Report, DOE Order No. DE-AF22-96PC01257, 1997.
19. Kawatra, S.K.; Eisele, T.C. *Coal Desulfurization*; Taylor and Francis: London, 2001.
20. Seidell, A. *Solubilities of Inorganic and Organic Compounds: A Compilation of Quantitative Solubility Data from the Periodical Literature*; D. Van Nostrand Company: New York; 1919.
21. Theodore, L.; Buonicore, A.J.; McKenna, J.D.; Kugelman, I.J.; Jeris, J.S.; Santoleri, J.J.; McGowan, T.F. Waste management, Section 25. In *Perry's Chemical Engineers' Handbook*, 7th Ed.; Perry, R.H., Green, D.W., Maloney, J.O., Eds.; McGraw-Hill: New York, 1999.
22. Bryan, R.R.; Smith, A.A.; Farmer, C. Texas plant demonstrates viability of coal option. Power **1993**, *137*, 57–64.
23. Dettmer, R. Sans sulphur: the Drax FGD project. IEE Rev. **1994**, *40* (2), 88–89.
24. Fellman, R.T.; Cheremisinoff, P.N. Lime/limestone scrubbing for SO_2 removal, Chapter 12. In *Air Pollution Control and Design for Industry*; Cheremisinoff, Ed.; Marcel Dekker: New York, 1993; 339–357.
25. Kawatra, S.K.; Eisele, T.C. *Separation of Flue-Gas Scrubber Sludge into Marketable Products, First Quarterly Technical Progress Report*; U.S. Department of Energy, Pittsburgh Energy Technology Center, Contract No. DE-FG22-93PC93214, 1993.
26. Kawatra, S.K.; Eisele, T.C.; Banerjee, D.D. Recovery of gypsum and limestone from scrubber sludge by water-only cyclone and conventional froth flotation. In *New Remediation Technology in the Changing Environmental Arena*; Scheiner, et al., Eds.; Society for Mining, Metallurgy, and Exploration, Inc.: Littleton, CO, 1995; 99–104.
27. Eisele, T.C.; Kawatra, S.K. Separation of the components of flue gas scrubber sludge by froth flotation. In *Proceedings of the 19th International Mineral Processing Congress*; Vol. 4, Chapter 33; Society of Mining, Metallurgy, and Exploration: Littleton, CO, 1995; 163–166.
28. Plummer, L.N.; Wigley, T.M.L.; Parkhurst, D.L. The kinetics of calcite dissolution in CO_2–water systems at 5°C to 60°C and 0.0 to 1.0 atm CO_2. Am. J. Sci. **1978**, *278*, 179–216.
29. Fair, J.D.; Steinmeyer, D.E.; Penney, W.R.; Crocker, B.B. Gas absorption and gas–liquid system design, Section 14. In *Perry's Chemical Engineers' Handbook*, 7th Ed.; Perry, R.H., Green, D.W., Maloney, J.O., Eds.; McGraw-Hill: New York, 1999.
30. Ellison, W.; Hammer, E.L. FGD gypsum use penetrates U.S. wallboard industry. Power **1988**, *132*, 29–33.
31. Van der Brugghen, F.W.; Koppins-Odink, J.M. Flue Gas cleaning in power stations in the Netherlands. In *Proceedings: First Combined Flue Gas Desulfurization and Dry SO_2 Control Symposium*; Electric Power Research Institute, Paper 1-5, 1989, 1/69–1/94.
32. Shoop, K.J.; Blystone, S.S.; Kawatra, S.K. Zeta potential measurements of the components of wet flue-gas scrubber sludge. In *Proceedings of the 1996 SME Annual Meeting*, Phoenix, AZ, Preprint No. 96–99, 1996.
33. Feeney, S. Upgrade scrubbers to improve performance. Power **1995**, *139*, 32–37.
34. Anonymous. DOE kicks off clean air program to boost performance of scrubbers. Fossil Energy Rev.: U.S. Department of Energy, **1992**; October–December, 18–19.
35. Anonymous. Southern company begins testing 2nd generation clean coal scrubber. Fossil Energy Rev.: U.S. Department of Energy, **1992**; October–December, 28–29, 39.
36. Hodges, G.J.; Roset, G.K.; Woodland, L.R.; Stevenson, J.A. Dual-alkali scrubbing at stillwater mining company. Mining Eng. **1992**, *44* (10), 1269–1271.
37. Chiang, S.-H.; Cobb, J.T., Jr. Coal conversion processes (desulfurization). In *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th Ed.; John Wiley & Sons, 1993; Vol. 6, 511–540.
38. Vernon, J.L.; Jones, T. *Sulphur and Coal, Report No. IEACR/57*; IEA Coal Research: London, 1993.
39. Pell, M.; Dunson, J.B. Gas–solid operations and equipment, Section 17. In *Perry's Chemical Engineers' Handbook*, 7th Ed.; Perry, R.H., Green, D.W., Maloney, J.O., Eds.; McGraw-Hill: New York, 1999.

40. Brady, J.D.; Legatski, L.K. Venturi scrubbers, Chapter 11. In *Air Pollution Control and Design for Industry*; Cheremisinoff, Ed.; Marcel Dekker: New York, 1993; 339–357.
41. Porter, H.F. Solids drying and gas–solid systems. In *Perry's Chemical Engineers' Handbook*, 6th Ed.; McGraw-Hill, 1984; 20/93.
42. Yeh, J.T.; Demski, R.J.; Joubert, J.I. Control of SO_2 emissions by dry sorbent injection. In *Flue Gas Desulfurization*; Hudson, Wells, Eds.; American Chemical Society: Washington, DC, 1982; 349–368.
43. Stouffer, M.R.; Rosenhoover, W.A.; Withum, J.A. Advanced coolside desulfurization process. Environ. Progr. **1993**, *12* (2), 133–139.
44. Helfritch, D.; Bortz, S.; Beittel, R.; Bergman, P.; Toole-O'Neil, B. Combined SO_2 and NO_x removal by means of dry sorbent injection. Environ. Progr. **1992**, *11* (1), 7–10.
45. Makansi, J. DOE's clean coal program: what industry has learned. Power **1992**, *136*, 56–61.
46. Kudlac, G.A.; Farthing, G.A.; Szymanski, T.; Corbett, R. SNRB catalytic baghouse laboratory pilot testing. Environ. Progr. **1992**, *11* (1), 33–38.
47. White, D.M.; Maibodi, M. Assessment of Control Technologies fro Reducing Emissions of SO_2 and NO_x from Existing Coal-fired Utility Boilers, Project Summary, Environmental Protection Agency, Report No. EPA/600/S7-90/018, 1991.

Six Sigma Design: An Overview of Design for Six Sigma (DFSS)

Sean A. Curran
Kwok-Wai Lem
Steve Sund
Mina Gabriel
Honeywell International, Inc., Morristown, New Jersey, U.S.A.

INTRODUCTION

Traditional six sigma methodology has become a standard process optimization tool for process industries. However, it has become clear that the ''Holy Grail'' of six sigma, 3.4 defects per million opportunities (DPMO), is simply unachievable after the fact. Consequently, there has been a growing movement to implement six sigma design usually called design for six sigma (DFSS). This methodology begins with defining customer needs and leads to the development of robust processes to deliver those needs.

In this entry, we introduce the DFSS approach to product/process development. This is based on a systematic application of powerful tools to define customer requirements and relate them to producer capability. This is followed by a detailed discussion of the various tools and their use. We show how the linkage of customer needs to product requirements and subsequent linkage of product requirements to process requirements drive to the development of robust processes.

DEFINITIONS AND THEORY

What Does DFSS Mean?

Ultimately any supplier's objective is to deliver on customer expectations 100% of the time.[1,2] DFSS is a systematic approach to develop processes that are capable of delivering on those expectations.[3–5] Ideally, all products would behave the same way all the time. In the real world, errors (i.e., defects) can and do occur because all processes exhibit variation. Consequently, all products manufactured via those processes will exhibit variation.[6] DFSS is a systematic methodology to predict that variability and the ability to meet customer needs. The methodology applies a suite of tools to define customer expectations and couples them directly with manufacturing capabilities so that the customer is always satisfied, and the manufacturer can earn adequate returns.

What Is Sigma?

The simple answer is that it is a measure of variability. Generally, in chemical processing we deal with normally distributed variables such as temperatures, flow rates, pressures, purity, mechanical strength, etc. These variables will have a nominal value or a specific setting and some level of variation. The average of all values is referred to as the mean (μ) and the spread is defined by the standard deviation, sigma (σ). We can describe these distributions statistically and predict the probability of a certain value occurring as illustrated in Fig. 1. In cases where the variables are not normally distributed, such as high purity materials where the distribution becomes highly skewed as one approaches the physical limits of 100% purity or 0% impurity, the data must be handled appropriately. The correct distribution can be modeled, or the data can be transformed into a normal distribution by using averages of multiple measurements. The central limit theorem predicts that such averages will be normally distributed. Strictly speaking, one cannot know the true mean and standard deviation unless the entire population is measured. In practice, we use a sample to predict the true mean and standard deviation. Consequently, there is always a finite probability that the sample chosen does not represent the true population.

What Is Six Sigma?

The discussion of sigma (σ) given earlier deals with the natural variation in the process.[1,7–10] Consequently, we denote the measurement of σ as a measure of the voice of the process. This is clearly an internally focused assessment of process capability, i.e., the ''voice of the process.'' The probability of a defect in the process, i.e., failure to meet customer expectations, is commonly defined in terms of DPMO. The focus of DFSS is to achieve a minimum of six standard deviations between the mean and the nearest specification ($\pm 6\sigma$), corresponding to 3.4 DPMO. Because the specifications are set to meet customer needs, six sigma in this

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016185

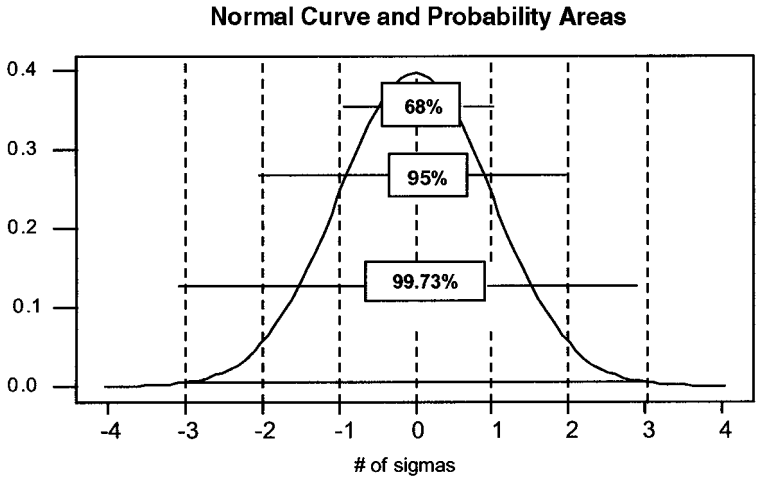


Fig. 1 Normal probability curve. *(View this art in color at www.dekker.com.)*

context is the ratio between the voice of the customer and the voice of the process. This ratio is the sigma score, commonly referred to as "Z" score and is calculated as shown in Eq. (1), where μ is the process mean, σ is the process standard deviation, USL is the upper specification limit, LSL is the lower specification limit, and the sigma score (Z) is the smaller of the calculated values. Brief inspection of the equation shows that Z is dimensionless.

$$Z = \frac{\min(\text{USL} - \mu, \mu - \text{LSL})}{\sigma} \quad (1)$$

In effect this calculation normalizes all processes to a number of sigmas rather than absolute values. The relationship between DPMO and Z score is illustrated in Fig. 2. Z refers to the white area under the curve and the shaded area shows the area where there is a probability of failure. DPMO is the integration of the shaded areas, i.e., the proportion of the results beyond the calculated Z value. Obviously, as Z increases the defective part of the distribution shrinks. The exact probability associated with a specific Z score can be easily obtained from Z score tables or calculated with common software packages such as Excel® or statistical packages such as Minitab®.[a] In the case of Excel, such calculations are not part of a standard package, but macros can easily be written to perform the needed calculations.

Why Six Sigma?

Naturally, all producers wish to minimize their production cost while maintaining quality. Likewise, all customers desire 100% defect free products at minimum cost. Infinite repetitions of product enhancements are only useful if they provide value to the customer. Further improvements add cost, without adding value. Brief inspection of a Z score table shows that a Z score of 6 actually corresponds to a probability of 10^{-9} for a failure. However, most studies of process capabilities are relatively short term in nature. After all, the supplier's interest is in making product, not continuously evaluating the process. Experience has shown that over time, processes will drift off the original mean. Typically this drift amounts to 1.5σ (for example, see p. 512 of Ref.[5]), so one compensates for the long-term variation by subtracting 1.5 from the short-term sigma score.[1,5] In effect, a short-term 6σ process becomes a long-term 4.5σ process, and this corresponds to the often quoted 3.4 DPMO.

What is a CTQ?

The acronym CTQ stands for critical to quality. At the highest level, CTQs are those benefits that accrue to the customer when they use a product/process. Top level CTQs are not specific to either product or process.

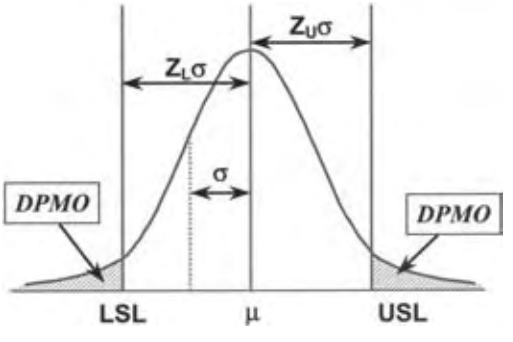


Fig. 2 Relationship of Z score with DPMO. *(View this art in color at www.dekker.com.)*

[a]Excel® is a registered trademark of Microsoft, Inc. and Minitab® is a registered trademark of Minitab®, Inc.

These are the performance attributes that the customer is actually paying for such as load bearing capacity, stability in a harsh environment, etc. In the following section, we discuss requirement flow down. As those requirements flow down, we establish lower level CTQs. These are those process and/or product requirements that are critical for delivering the customer CTQs, i.e., those parameters one must control in order to deliver the desired performance.

DFSS TOOLS

Quality Function Deployment

Once the customer needs (top level CTQs) have been defined a DFSS team needs to determine how process parameters that can be addressed, such as raw materials, process control parameters, etc. are related to the ability to deliver customer CTQs. This is done by generating a series of quality functional deployment (QFD) matrices as sketched in Fig. 3. There is no hard and fast rule on how many levels a QFD needs. Frequently, polymer and chemical QFDs have only three levels as there are no subassemblies, and consequently, there is no need for a ''parts'' QFD because process specifications relate directly to the process design CTQs.

We discuss QFD at some length because it is one of the pillars of a successful DFSS effort. A proper QFD relates customer CTQs to product features and assesses various competing alternatives for delivering the desired features.[4,5] Quality functional deployment then relates product features to product design, product design to process design, and finally process design to process specifications. Naturally, any product has an array of features and customer needs are rarely monofunctional. A successful DFSS effort focuses on the most critical features. Quality functional deployment provides a mechanism for defining product design CTQs based on customer CTQs, then defining process design CTQs based on product design requirements and process specification based on process CTQs, as illustrated in Fig. 4. Usually, the most difficult part of a QFD is the definition of customer needs. Rarely, if ever, is the first attempt at a QFD complete. Customers frequently do not know their total needs, and can only detect them as deficiencies, well after the initial QFD definition. The only way to compensate for this initial failing is to maintain close customer contact and treat the QFD as a living document. Once initial designs are developed, customer feedback must be used to update and refine the QFD.

The first matrix (QFD1) relates the customer needs, i.e., top level CTQs, to the required functions and features a product needs to deliver those CTQs.

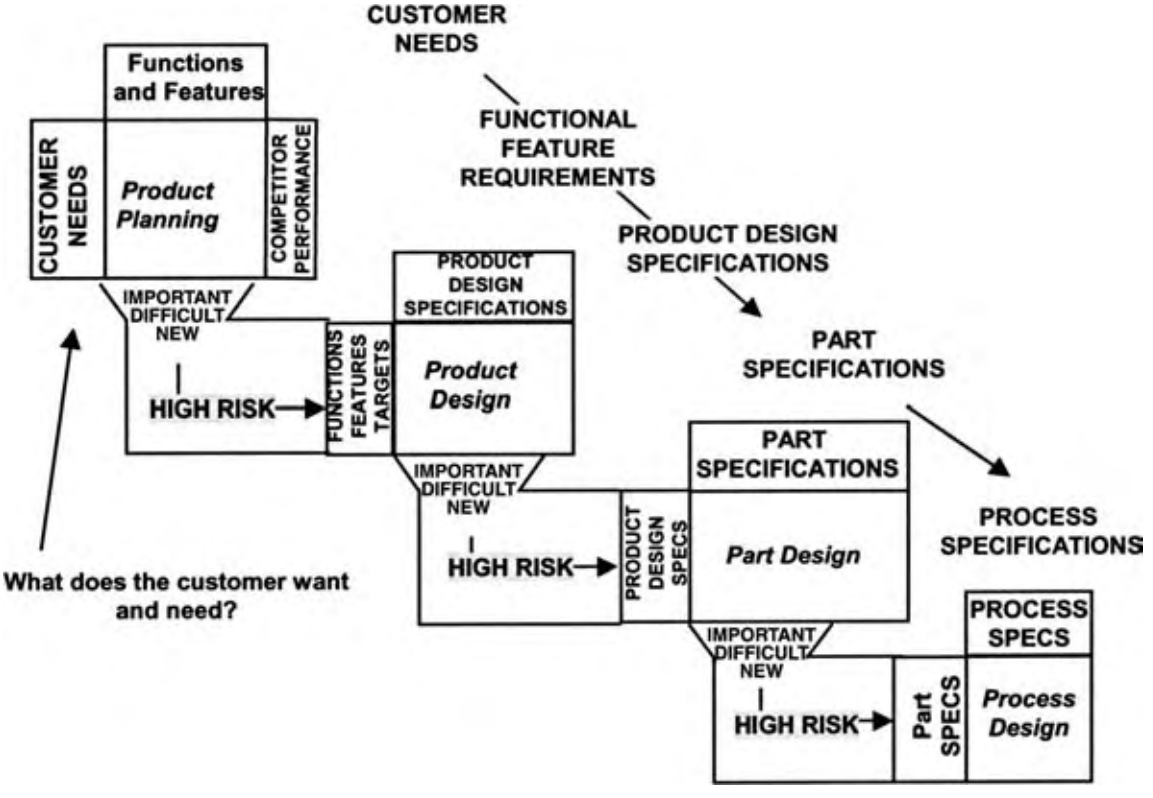


Fig. 3 Linkage of customer needs to process controls via QFD. *(View this art in color at www.dekker.com.)*

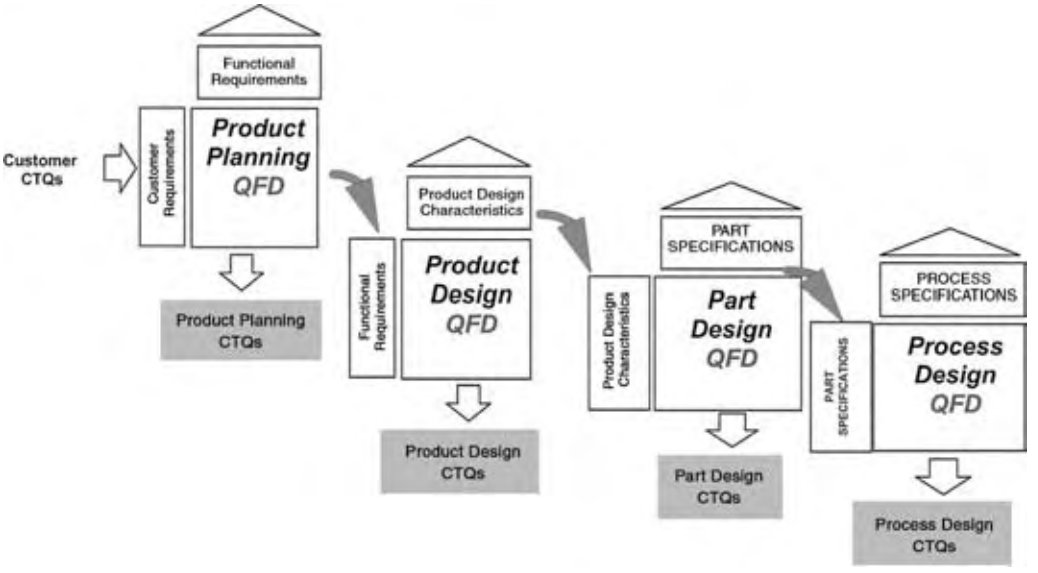


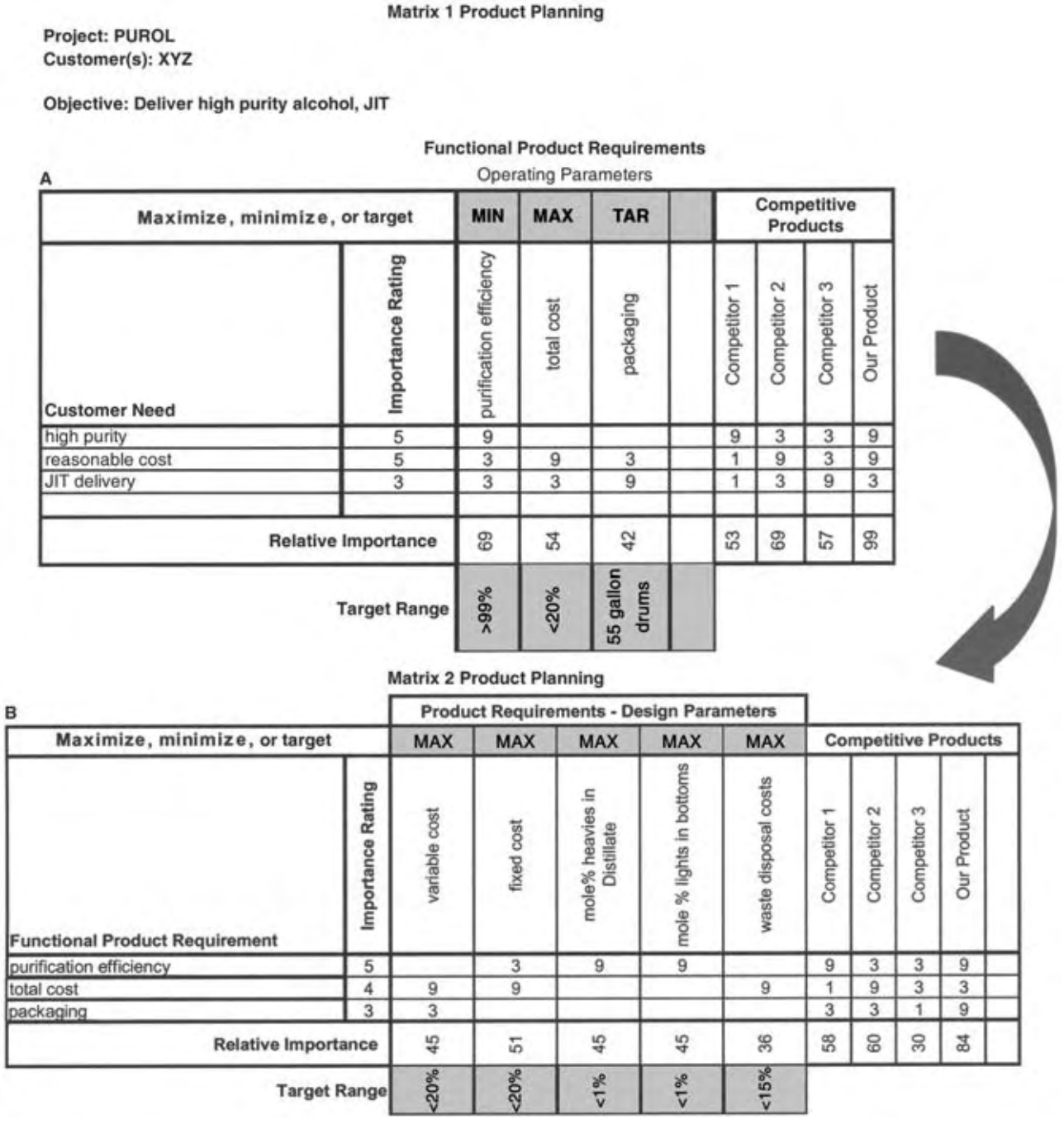
Fig. 4 Flow down of customer requirements to process specifications. *(View this art in color at www.dekker.com.)*

For example, if the customer CTQ is stability in harsh environments, the related features might be the ability to withstand 100°C exposure at pH 1 for 1000 hr. Note that the features still have not defined a specific product or process, but they have defined measurable criteria that relate to the desired CTQ. At this stage, the DFSS team assesses what competitive alternatives exist that might deliver the same features, and proceeds to flow the functional requirements downward to eventually define process specifications.

We do not have the space here to explore all the aspects of QFD, but there are excellent sources available in print and online that provide this kind of detail.[4,5,11–13] We now delve briefly into the mechanics of the QFD. This is a tool that is simple to understand, but can be very difficult in practice.

An illustrative example of a typical QFD is given in Figs. 5 and 6. A customer desires high purity alcohol, at reasonable cost with just in time delivery. The five steps in a QFD1 (Fig. 5A) are:

1. Assess customer needs, typically through customer interviews, surveys, trade groups, marketing contacts, etc. Once the needs have been identified they are prioritized via numerical rankings. A variety of approaches to these rankings can be used. In Fig. 5A, we have used a 1,3,5 ranking for denoting importance. The customer must absolutely have high purity and reasonable cost; hence these needs are rated 5. The delivery schedule is important, but there is some flexibility generating a rating of 3. The exact ranking scheme is not critical, provided the DFSS team understands how important these performance CTQs are to the customer.
2. Determine what types of existing solutions and competitive approaches already exist to deliver the desired product features and contrast them with the DFSS team's potential solution. Again a numerical ranking is used to assess how well each of the alternatives addresses the customer CTQs. In Fig. 5A, we have used a logarithmic 1,3,9 scale for these ratings. Competitor 1 can deliver high purity alcohol, but at a high cost and poor delivery schedule. Competitor 2 has low cost, but lower purity. The rating of each alternative is multiplied by their importance rating and their products are summed to provide an overall weighted performance rating. Thus Competitor 1's overall rating is 53 $\{(9 \times 5) + (1 \times 5) + (1 \times 3)\}$. This exercise allows the team to quickly assess how big the gap is between potential solutions, as well as defining potential strong points.
3. The DFSS team assesses what types of product features are needed to deliver the required customer benefit, and how they will measure those features—measurement system, targets, and ranges. Those items are listed as functional product requirements.
4. The DFSS team evaluates how strong the relationship is between the functional requirements and each customer CTQ. The DFSS team defines the rating scale for the relationship strength, although in our experience a logarithmic 0,1,3,9 scale, with 9 being the strongest relationship, works quite well. This type of ranking helps to rapidly make an assessment rather than arguing about which relationships are slightly stronger than the others.



Matrix 1 Product Planning

Project: PUROL
Customer(s): XYZ

Objective: Deliver high purity alcohol, JIT

A

Functional Product Requirements
Operating Parameters

Maximize, minimize, or target		MIN	MAX	TAR		Competitive Products			
Customer Need	Importance Rating	purification efficiency	total cost	packaging		Competitor 1	Competitor 2	Competitor 3	Our Product
high purity	5	9				9	3	3	9
reasonable cost	5	3	9	3		1	9	3	9
JIT delivery	3	3	3	9		1	3	9	3
Relative Importance		69	54	42		53	69	57	99
Target Range		>99%	<20%	55 gallon drums					

Matrix 2 Product Planning

B

		Product Requirements - Design Parameters									
Maximize, minimize, or target		MAX	MAX	MAX	MAX	MAX	Competitive Products				
Functional Product Requirement	Importance Rating	variable cost	fixed cost	mole% heavies in Distillate	mole % lights in bottoms	waste disposal costs	Competitor 1	Competitor 2	Competitor 3	Our Product	
purification efficiency	5		3	9	9		9	3	3	9	
total cost	4	9	9			9	1	9	3	3	
packaging	3	3					3	3	1	9	
Relative Importance		45	51	45	45	36	58	60	30	84	
Target Range		<20%	<20%	<1%	<1%	<15%					

Fig. 5 QFD1 and QFD2 for high purity alcohol. *(View this art in color at www.dekker.com.)*

5. Each relationship strength is multiplied by each CTQ importance rating and the products are summed to give a weighted relative importance for each feature.

Examination of Fig. 5 shows how all these different pieces of information are correlated to generate a numerical ranking of functional product requirement importance. Clearly, QFD is at best a ''semi'' quantitative tool, particularly at the top level. CTQ importance ratings are based on the customer's opinion. The strengths of the relationships are based on the DFSS team's experience and knowledge of what types of features will drive performance. Nevertheless, this is an extremely powerful tool because it focuses the team on what the product should be designed to do for the customer rather than what their existing product can do.

Once the team has completed the prioritization of functional requirements, they proceed to QFD2 (Fig. 5B) where the process is repeated, but this time

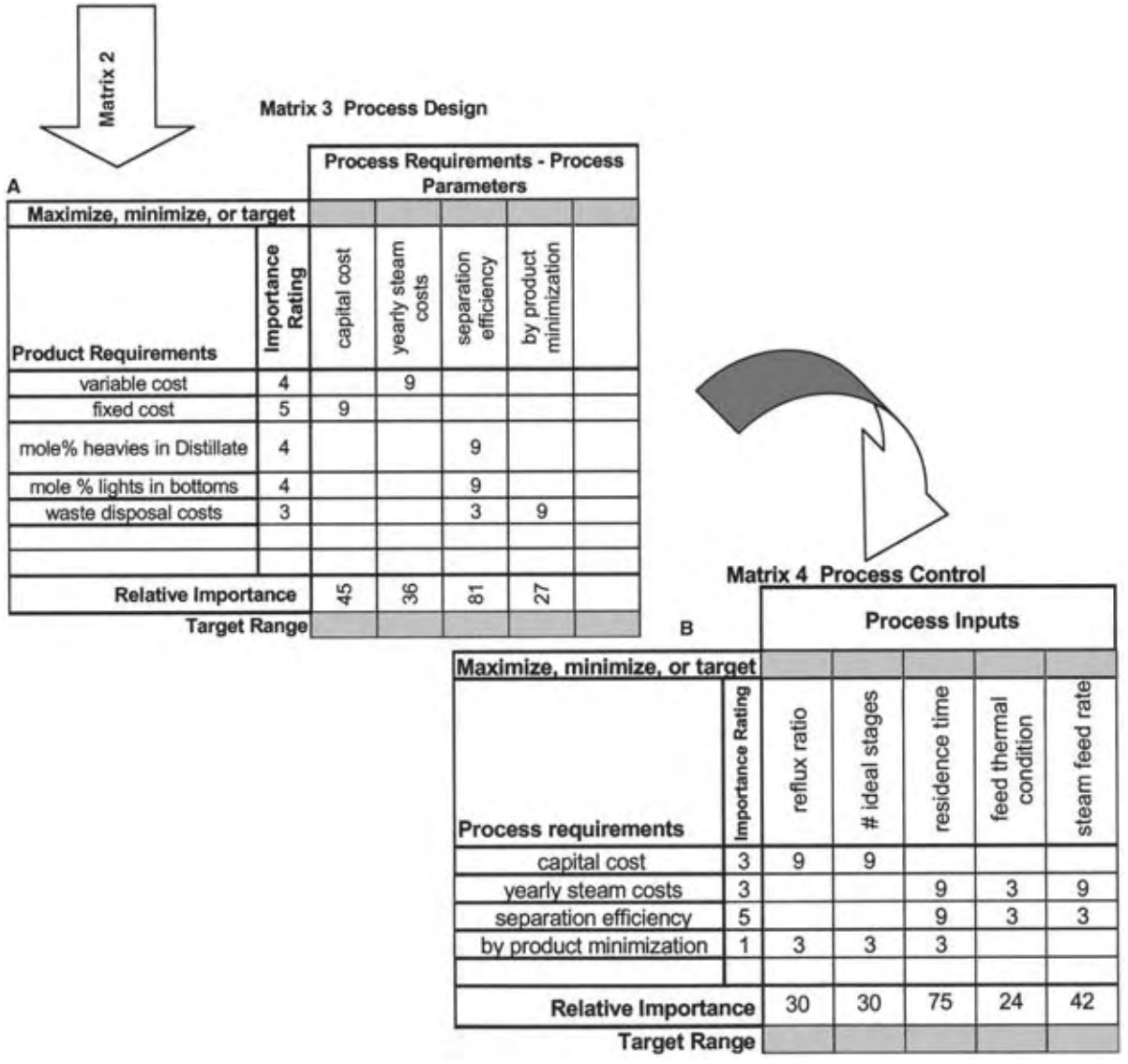


Fig. 6 QFD3 and QFD4 for high purity alcohol. *(View this art in color at www.dekker.com.)*

linking functional product requirements with product design requirements. The prioritized functional requirements are reranked, typically on a 1–5 scale; the product design requirements to deliver those functions are defined; potential competitive approaches are defined; and the numerical evaluations in terms of competitive performance and relationship strength are evaluated. Once the product design requirements are defined and prioritized, they are linked to the process requirements via QFD3 as shown in Fig. 6A. The process is repeated in QFD4 to link process requirements to process controls (Fig. 6B). Note that there are no operating ranges specified in Fig. 6. If one is designing a new process or product, the design requirements may well be known while the required process parameter settings to reach those design requirements need to be evaluated. The strength of the QFD is that it helps illustrate those process parameters that need to be defined. The approach for defining them is discussed in the sections on transfer functions, design of experiments (DOE), and setting specifications.

Transfer Functions

Simply put, the DFSS definition of a transfer function is:

$$\begin{aligned} Y &= f(x_1, x_2, x_3, \ldots, x_n), \\ S_Y &= g(S_{x_1}, x1, S_{x_2}, x2, \ldots) \end{aligned} \tag{2}$$

This shows that there is a definable relationship between the output of a process and its inputs, and between the variability of the inputs and variability

in the output. A simple illustration is given in Fig. 7. Two aspects of a given feature are important: first its direct effect on the output (Y), second the variability of the feature and the consequent variability of the output (S_Y). Variability is unavoidable as raw material lots change, processes change, and environmental influences affect a process. Therefore, simply defining the targets is only half the story.

As with QFD, transfer functions are an essential pillar of the DFSS methodology. In fact once QFD has defined what is required, then the transfer functions are a mathematical model of how the process can deliver customer CTQs. If we understand how process settings affect product features and how features affect performance, then it is feasible to design a product that will meet customer requirements, with yields matching model predictions. If we do not understand these relationships, then the probability of success diminishes rapidly.

There are two sets of transfer functions one should be concerned with: target settings and effects of variability.[4,14] These are illustrated in Fig. 8, which shows the relationship between controllable inputs and their variability with the desired output and its variability. Thus, QFD defines how one flows the customer CTQs downwards, while the transfer functions define how one predicts process capability and defines the critical control points.[4,5] The transfer functions allow the team to establish this linkage early in the development, rather than try to do it once products are in production. In addition when customer requirements change, the team does not have to begin a new project. With the transfer functions in hand, they can quickly evaluate capability and predicted reliability for the required process changes.

The most difficult part of using transfer functions is defining them in the first place. Once they exist, they are simple to use. The process map in Fig. 9 illustrates schematically how one can define transfer functions. These functions can take a variety of forms and arise from a variety of sources as illustrated in Fig. 10. The functions can be derived from known mechanistic models, e.g., based on the ideal gas laws, one can predict gas pressure at a given temperature, or an Arrhenius equation can be used to predict kinetics. The appropriate model parameters may be known a priori or can be determined by experimentation. One can also develop empirical models based on experimental data. Mechanistic models are preferable, because they explain why the output depends on the input and there is a clear understanding of the mechanism. Empirical models are also quite useful with the caveat that they are based on an observation of correlation. There is always a possibility that the observed correlation is serendipitous; so the DFSS team must carefully review the model to be sure that the observed function is reasonable. In practice, empirical models are more prevalent in complex chemical systems, but it is essential that the model only be used for the same experimental space that was used in the model. Extrapolation outside this area is at best highly questionable. If the team wishes to extrapolate, then additional experimentation is required to make sure the model fits in the new region.

We illustrate transfer function development with the high purity alcohol example as mentioned earlier. Assuming we have an empirical function, with known parameters, we can formulate a mathematical model to product performance and cost. Using the approach of Chang,[15] we can solve the short cut distillation equations of Hengstebeck, Geddes, Fenske, Underwood, and Gilliland.[16–21] This provides the transfer functions for technical requirements. We then need to incorporate the financial requirements. We have developed a Monte Carlo simulation program to couple the transfer functions for each level of the QFD leading to a final result, which optimizes for technical specifications while minimizing the cost.

The program obtains a standard normal value and transforms based on the input mean and standard deviation for each of the input variables. All these

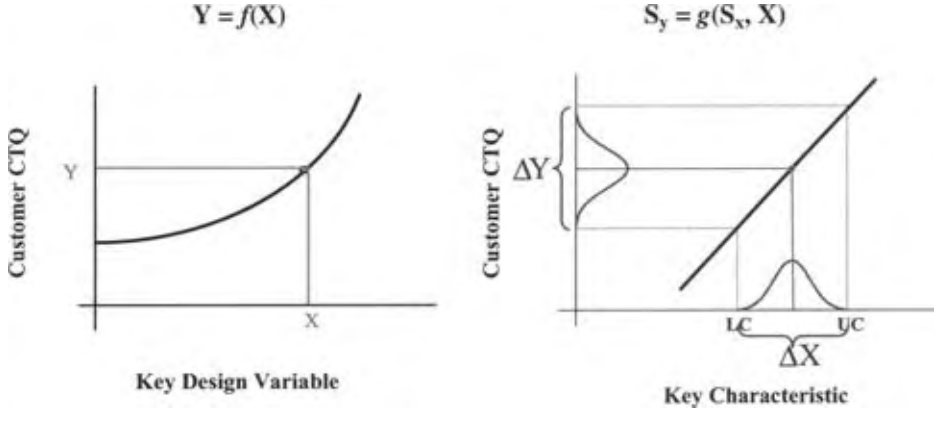


Fig. 7 Schematic illustration of transfer functions for targets and variability in a key variable. *(View this art in color at www.dekker.com.)*

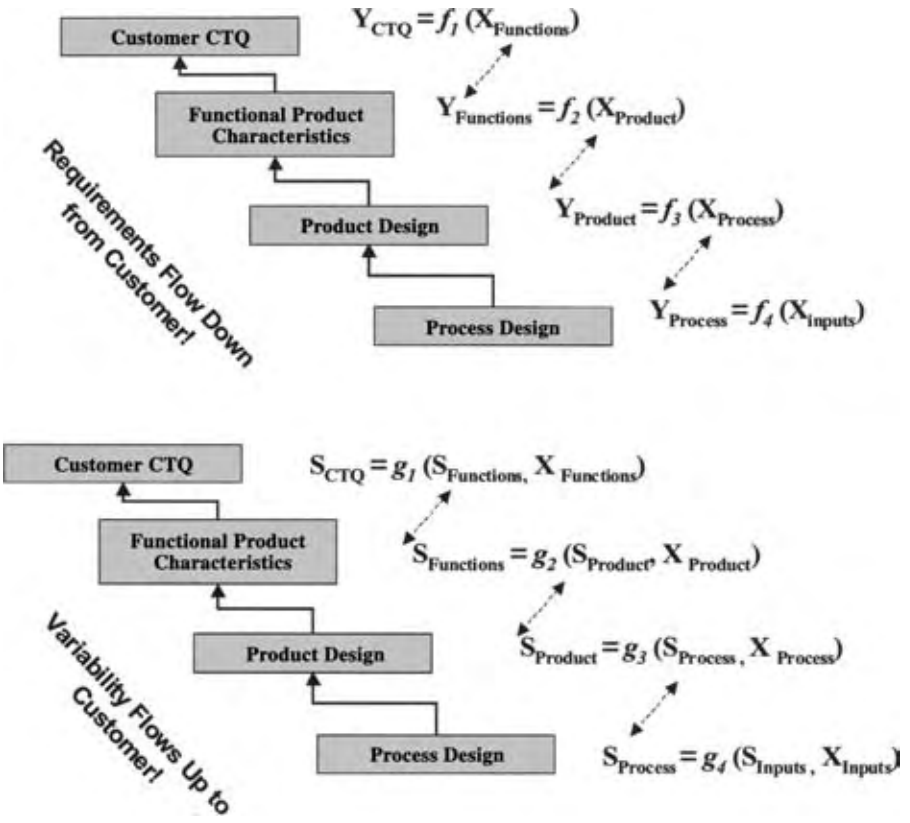


Fig. 8 Linkage of transfer functions from process parameters through customer CTQs. *(View this art in color at www.dekker.com.)*

are sent to the trays subroutine to calculate ideal stages (IS). These are then stored in an array of responses, and statistics such as mean, standard deviation, skewness, and kurtosis are calculated on the results. In addition, a factor response summary is calculated for each of the four input variables. The method used is that found in Refs.[22,23] The results of this simulation are shown in Fig. 11. We then extended the model to incor-

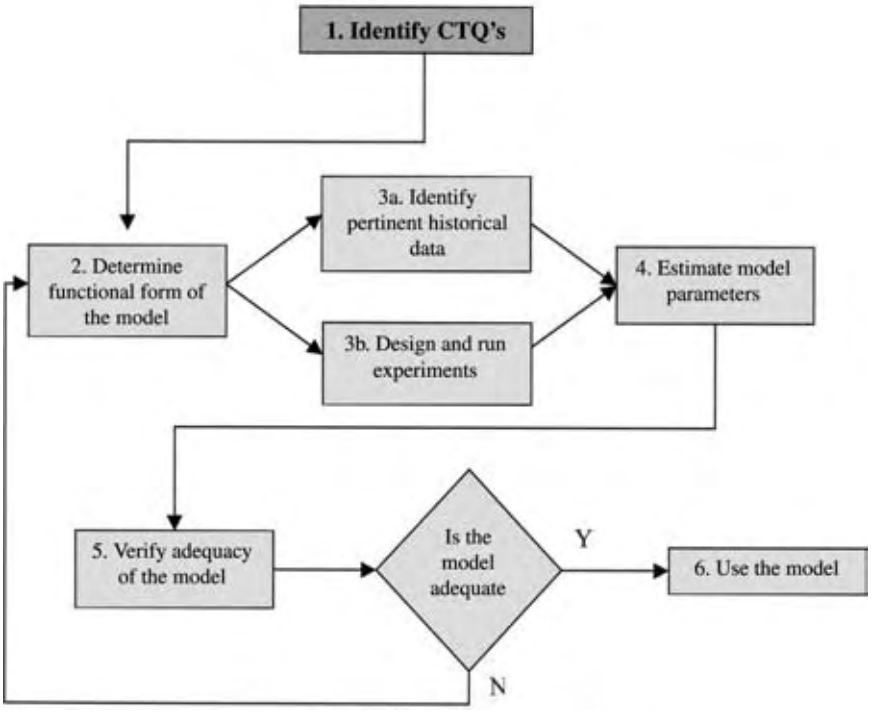


Fig. 9 Process map for creating transfer functions. *(View this art in color at www.dekker.com.)*

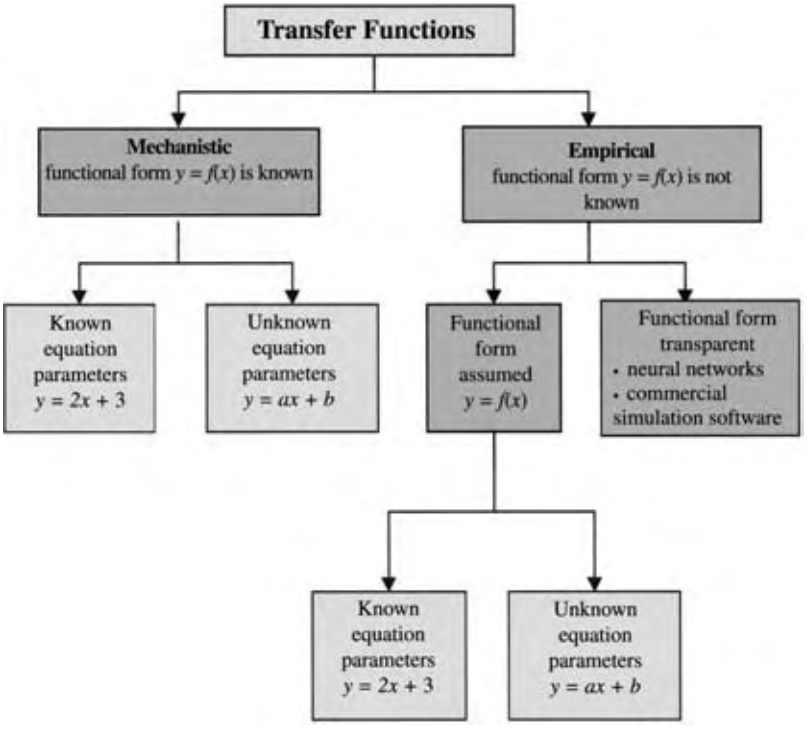


Fig. 10 Types of transfer functions. *(View this art in color at www.dekker.com.)*

porate those factors related to cost. This was of more interest, because it incorporates multiple levels of transfer function. In effect, we can manipulate the input variables and see how the ultimate outputs, which are four levels higher in the QFD, are affected.

This comprehensive model uses the same method described earlier; transforming a standard normal value using the first four means and standard deviations as depicted in Fig. 11.

Submission to the trays subroutine returns the number of IS and reflux ratio (RR). These are then used to calculate total cost (TC) of the column based on the algorithm (i.e., transfer functions) shown below.

Capital cost (CC) [$/yr] is calculated by Eq. (3), which is a function of the number of IS and RR.

$$CC = 1353278.3 + 44483.4 \times IS - 949840.14/RR \quad (3)$$

Fixed cost (FC) [$/yr] is calculated by Eq. (4), which is a function of CC.

$$FC = (0.0000759 + 0.0933 \times CC \times CC)^{1/2} \quad (4)$$

Yearly steam cost (YSC) [$/yr] is calculated based on 8000 hr in a year and Eq. (5) is a function of steam rate (SR) and steam cost (SC).

$$YSC = SR \times 8000 \times SC \quad (5)$$

Variable cost (VC) is calculated by Eq. (6) and is a function of YSC.

$$VC = 327322 + 1.31 \times YSC \quad (6)$$

Finally, TC [$/yr] is a function of FC and VC, and is calculated by Eq. (7).

$$TC = VC + FC \quad (7)$$

The results of this simulation are shown in Fig. 12.

Design FMEA

Simply knowing the required controls is insufficient to guarantee consistent performance. The DFSS team must design controls and safeguards into the system. The primary tool in this effort is the design failure modes and effects analysis (DFMEA). Failure mode and effects analysis (FMEA) is a well-known tool, which has been extensively used in a wide range of industries.[1–5,7–8] It is an approach that allows prioritization based on the highest failure risks in the development process. Failure mode and effects analysis couples the severity, frequency, and detectability of failure to meet customer CTQs to assess overall risk of a given failure type. There are different types of FMEA for manufacturing processes, administrative processes, and design processes. A DFSS team will

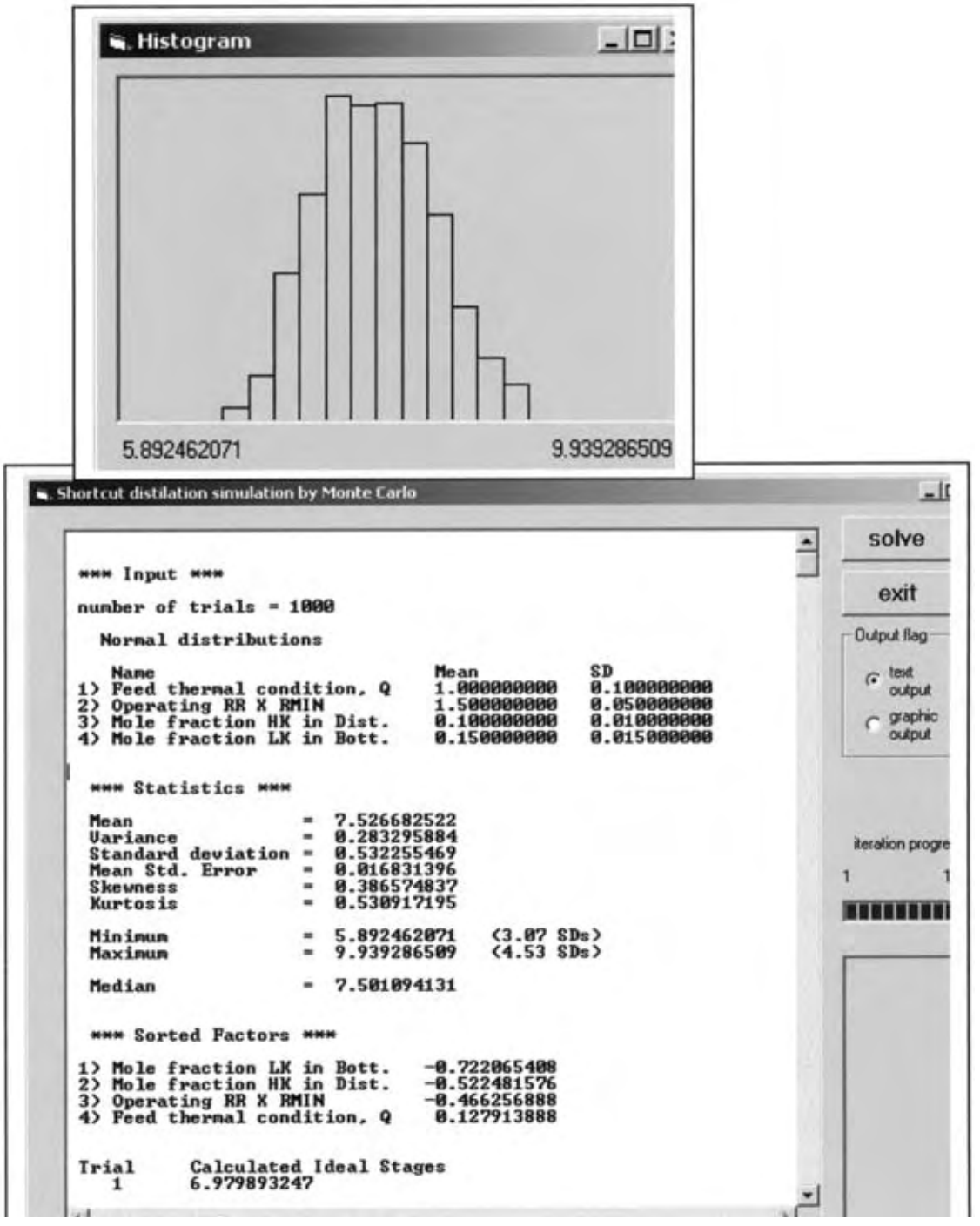


Fig. 11 Monte Carlo simulation of separation efficiency for alcohol purification. *(View this art in color at www.dekker.com.)*

be involved in the DFMEA and at the very least a preliminary manufacturing FMEA.

A typical design FMEA is given in Fig. 13 for the alcohol example discussed earlier. The headings across the top of the FMEA define the types of information it requires. The DFSS team must evaluate the risk associated with failure of the CTQs. We discuss the functional requirement FMEA, but the same approach applies to any FMEA. The first column specifies the requirement that is being evaluated. The second column specifies the failure mode. A failure mode is the way in which the requirement is not met. The third column specifies the effect on the customer if that requirement is not been met. The next column is a numerical rating of severity, i.e., how badly does it affect the customer. The fifth column addresses causes

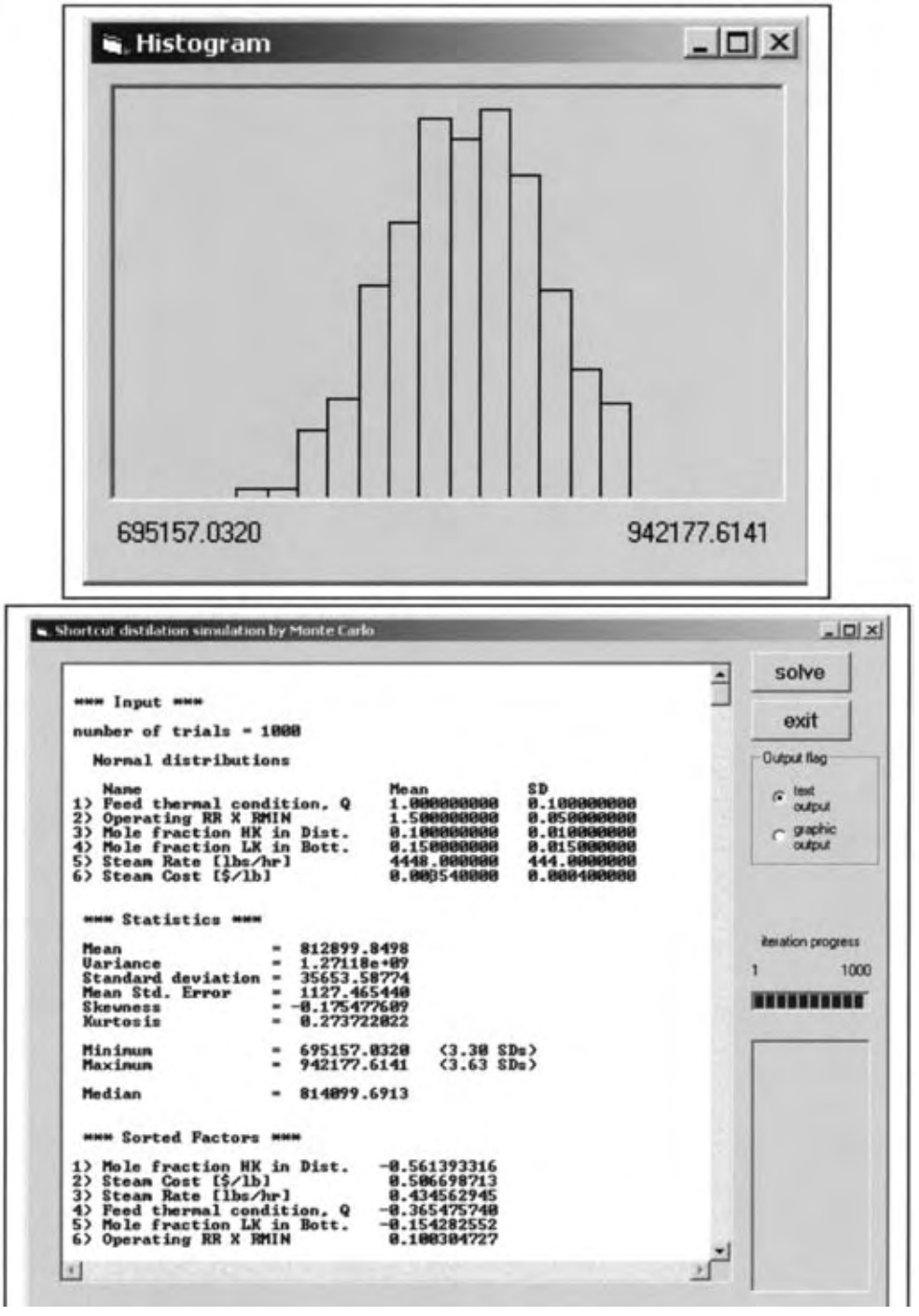


Fig. 12 Monte Carlo simulation for separation efficiency, incorporating cost transfer function. *(View this art in color at www.dekker.com.)*

for the failure mode. The next is a numerical rating of the likelihood for the occurrence of this cause. The seventh column is an assessment of the types of systems being employed to determine the occurrence of failure and to eliminate it before product release. The next is a numerical rating of the effectiveness of that particular detection system. The total risk is computed by multiplying the severity rating by the frequency rating by the detectability rating, generating the next column titled the risk priority number (RPN). The remaining columns are for tracking what corrective actions are needed, who is responsible, and the effect on the RPN after taking the required actions. Inspection of the RPN values immediately focuses the DFSS

Product Design Failure Modes and Effects Analysis (FMEA)

Date: 1/24/2001
Project Name: High Purity Alcohol
Responsible: T. Clarke
Rev. 2

Functional Product Requirement	Potential Failure Mode	Potential Failure Effects	SEV	Potential Causes/ Mechanisms	OCC	Current Design Evaluation or Control	DET	RPN	Actions Recommended	Responsibility	Actions Taken	SEV	OCC	DET	RPN
What is the functional product requirement under consideration?	In what ways could the functional product requirement fail to be fully met?	What would be the impact of failure mode on the customer (internal or external)?	How severe is the effect to the customer?	What could cause the failure mode to occur?	What is the likelihood that the cause will occur?	What methods, tools, or measures will discover the cause before design release?	How difficult is it to detect the cause of failure mode?	Risk Priority Number (SEV X OCC X DET)	What are the actions for reducing the occurrence of the cause or improving detection? (Should have actions only on high RPNs or easy fixes.)	Who is responsible for the recommended action?	What were the completed actions taken and the recalculated RPN? (Be sure to include completion month/year.)	How severe is the effect to the customer?	What is the likelihood that the cause will occur?	How difficult is it to detect the cause of failure mode?	Risk Priority Number (SEV X OCC X DET)
Total cost	Steam cost higher than expected	High variable cost	9	rise in energy cost	9		9	729	Negotiate long term supply	Supply Chain					
Total cost	Overdesign - too many stages	High fixed cost	9	Poor Model	3		9	243	Define model	T. Barry	Response Surface DoE, 7/31/01	9	1	1	9
purification efficiency	Heavies in Distillate	Product contamination	9	Short residence time	3	GC on product	3	81	Install sensor on line	M. Collins	Flow sensor designed and tested, 9/15/01	9	1	1	9
purification efficiency	alcohol in bottoms	raise cost	9	Short residence time	3	IR on bottoms	1	27	optimize flow	M. Collins	DoE to set tolerances, 1/15/02	9	1	1	9
Packaging	Wrong size	Slow delivery	3	No high purity drums	3	inventory control	1	9							

Fig. 13 Design failure mode and effects analysis for alcohol purification. *(View this art in color at www.dekker.com.)*

team on the highest risk failure modes, which in this illustration are rise in energy cost and poor model. This is not to imply that the other failure modes are unimportant. It is simply a mechanism to reduce or eliminate the highest risks. Failure mode and effects analysis is an iterative process. Once the high RPNs are reduced, the DFSS team needs to determine if the overall risk is sufficiently small to justify product manufacturing. If not, then the ongoing efforts taken must be aimed at reducing the other high risk items.

Developing an FMEA is a DFSS team effort. The numerical ratings for severity, detectability and frequency must be agreed on by the team, so that everyone agrees on the definitions. Otherwise one team member may consider the effect to be mild, another moderate, and yet another severe. As with QFD, the numbers here are ''semi'' quantitative, and round table team discussions help to achieve a more accurate picture of the true risks associated with the specific failure modes. If possible, the team should formalize the rating system as much as possible. A sample of such a formalized scale is given in Fig. 14.

Experimental Design and Process Simulation

Design of experiments is a fundamental tool in DFSS.[4,5] Systematic experimentation is frequently needed to determine the functional form of a transfer function.[14,24–29] This applies in particular to chemical reactions and polymer processing, where predicting yields of complex, multistep processes is difficult. DOE has been widely used to determine optimized settings for chemical processes. In DFSS one must also evaluate the effect of input variation on the output. There are various types of DOE that address this. Taguchi designs systematically evaluate the effect of noise variables on the outputs. Schematically, this approach is illustrated in Fig. 15A for a two-factor design, with two levels for each factor and two noise variables. The corners of the large square (the outer array) define the factor settings. Around each corner is a smaller square (the inner array) with the various settings for the noise factors. For example, the factors DV1 and DV2 might be RR and residence time. The noise variable might be steam feed rate and lot to lot

Rating	Effect (Severity of the Impact on Performance)	Cause (Likelihood of Occurance)	Control (Difficulty of Detection; Likelihood of "Escape")
10	Very high severity; hazardous without warning	Very High; failure is almost inevitable	No control in place; Cannot detect
9	Very high severity; hazardous with warning		Very remote chance of detection
8	Loss of primary function performance	High; repeated failures	Check for expected results; Remote chance of detection
7	Reduced primary function performance		Very low chance of detection
6	Loss of secondary function performance	Moderate; occasional failures	Low chance of detection
5	Reduced secondary function performance		Moderate chance of detection
4	Minor defect noticed by most customers		Moderately high chance of detection
3	Minor defect noticed by some customers	Low; relatively few failures	High chance of detection
2	Very minor defect detected by discriminating customers		Very high chance of detection
1	No effect	Remote; failure is unlikely	Prevent root cause; Almost certain detection

Fig. 14 Typical rating scale used for FMEA. *(View this art in color at www.dekker.com.)*

changes in raw materials. Minimizing the influence of noise variables then makes the design inherently more robust, i.e., stable to external influences. The major advantage for the Taguchi designs is the direct assessment of the effects of noise variables. Frequently, experimenters refrain from using these designs because they require a large number of experimental runs. Note that this two level two-factor design requires 16 experiments to define a main effects model.

An alternative approach is the use of response surface or mixture designs to determine the transfer function. A schematic of a two-factor response surface design is given in Fig. 15B. This is accomplished in nine experimental runs and evaluates main effects, as well as factor interactions. If the variability of the inputs is known, then one can model the predicted variability using propagation of error techniques or Monte Carlo simulations based on the transfer function. The major advantage relative to the Taguchi approach is the need for significantly fewer experimental runs.[24,25] However, this approach assumes that the experiment has adequately mapped the experimental space, and that the input variations are known.

TRANSITION TO MANUFACTURING

Setting Specifications

The achievement of six sigma is typically a balance of the cost and customer requirement. Consequently, there are two ways to achieve six sigma. One can develop a

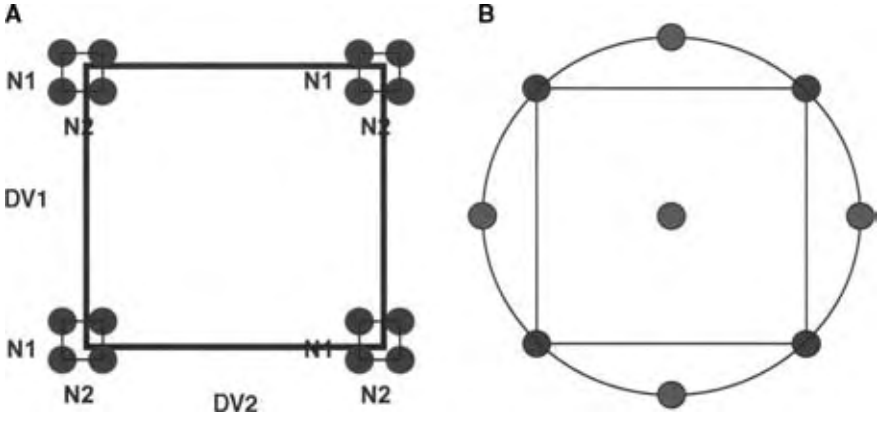


Fig. 15 Schematic illustration of different robust experimental designs: (A) Taguchi design and (B) response surface design. *(View this art in color at www.dekker.com.)*

Six Sigma Plus Control Plan

Product: High Purity Alcohol
Key Contact: T. Clarke
Phone: 555-1916

Core Team: T. Clarke, P. Pearse, J. Connolly

Date (Orig): 4/24/2001
Date (Rev): 3/17/2002
Rev. 1.1

Current Control Plan

Process	Process Step	Input	Output	Process Spec (LSL, USL, Target)	Cpk / Date (Sample Size)	Measurement System	%R&R or P/T	Current Control Method (from FMEA)	Who	Where	When	Reaction Plan
Distillation	Vaporization	feed thermal condition		Q=1 +/- .3	1.3	IR	5% R+R	pre-heat power	Operator	Control room	start of run	power to heat
Distillation	Vaporization	Steam feed		2000-7000lb/hr	1.7	Thermocouple	10%	Mass flow controller	Operator	Steam line	1/shift	shut down
Distillation	Vaporization	Residence time		10-15 min	1.8	Flow sensor	1%	on line sensor	Operator	Column inlet	Constant feedback to DCS	Adjust flow

Fig. 16 Sample control plan for high purity alcohol production. *(View this art in color at www.dekker.com.)*

very narrow process, which typically drives up the cost. Alternatively one can set wide specifications, but the customer CTQs must still be met. In DFSS one attempts to predict the variability as discussed earlier. This allows a prediction of achievable specifications.[30]

In practice, variation in customer CTQs is driven by multiple factors. For example, if the customer CTQ is a water white polymer (i.e., a transparent, colorless polymer), variations might depend on melt temperature, vacuum levels, quench speed, stabilizer content, etc. The strength of DFSS is that one will predict the set of conditions that minimize the variation around the target level of yellowing. Thus one might opt for rapid quenching and higher stabilizer because the oxygen level cannot be maintained at a sufficiently low level.

Finally, one can set specifications based on known process capability. If a sufficient number of lots have been produced, one can statistically evaluate the variation and set the specification limits outside the statistical control limits. Typical DFSS efforts do not lead to a sufficient number of sample lots for such a statistical evaluation. However, once the product is scaled up and in production, the specifications should be re-evaluated to make sure they will remain at six sigma performance.

Establishing Control Plans

The final critical piece of the DFSS puzzle is a control plan.[1–5,8–10] This is a document that specifies which are the critical inputs, where their limits need to be, and how they will be evaluated. Such a document should be part of any transfer to manufacturing. It needs to specify standard operating procedures that are required and what the capability of the measurement system is. DFSS teams should preferably include, but at the very least work with, manufacturing to make sure that any product or process designed can, in fact, be produced. Consequently, a preliminary control plan should be developed in parallel with the assessment of transfer functions and the FMEA. It should then be updated as more knowledge is developed about the control points and product performance. A good control plan has all the elements required to assess performance and take action to prevent failures. It includes the process points to be monitored, the specification limits, the measurement technique and measurement capabilities, as well as responsibility to take action. A typical control plan for our alcohol example is shown in Fig. 16.

CONCLUSIONS

Design for six sigma is not a radical new idea. It is in fact simply a codification of the scientific method. DFSS is a systematic application of good scientific and engineering practices. The tools employed are very powerful if used correctly. Fundamentally, DFSS is a road map and set of tools to provide the researcher with clear insight into customer needs and the customer with clear insight into producer capabilities.

ACKNOWLEDGMENTS

The authors thank Tony Signorelli, Janice Sund, Ross Rapoport, Alan Levy, Bill Hill, Dick Johnson, and

Páudraig Curran for their helpful comments and suggestions.

ARTICLE OF FURTHER INTEREST

Thermosets: Materials, Processes, and Waste Minimization, p. 3031.

REFERENCES

1. Harry, M.J.; Lawson, J.R. *Six Sigma Producibility Analysis and Process Characterization*; Addison-Wesley: New York, 1992.
2. Harry, M.J.; Schroeder, R. *Six Sigma: The Breakthrough Management Strategy Revolutionizing the World's Top Corporations*; Doubleday & Company: New York, 1999.
3. Crosby, P. *Quality Is Free: The Art of Making Quality Certain*; Reissue Ed.; Mentor Books, 1992.
4. Berryman, M.L. Transform your organization into one that's world class. Six Sigma Forum Mag. **2002**, *2* (1).
5. Creveling, C.M.; Slutsky, J.L.; Antis, D., Jr. *Design for Six Sigma—In Technology and Product Development*; Pearson Education, Inc.: New Jersey, 2003.
6. Senge, P.M.; Carstedt, G. Innovating our way to the next industrial revolution. MIT Sloan Manage. Rev., (Winter), 24–38.
7. Pande, P.S.; Holpp, L.H. *What is Six Sigma*? 1st Ed.; McGraw-Hill Trade, 2001.
8. Chowdhury, S. *Design for Six Sigma—The Revolutionary Process for Achieving Extraordinary Profits*; Dearborn Trade Publishing: Chicago, 2002.
9. Grant, E.L.; Leavenworth, R.S. *Statistical Quality Control*, 4th Ed.; McGraw-Hill: New York, 1988.
10. Pyzdek, T. *The Six Sigma Handbook*, 1st Ed.; McGraw-Hill Trade, 2000.
11. Cohen, L. *Quality Function Deployment: How to Make QFD Work for You*; Addison-Wesley, 1995.
12. Terninko, J. *Step-by-Step QFD: Customer-Driven Product Design*; CRC Press, 1997.
13. ReVelle, J.B.; Moran, J.W.; Cox, C.A. *The QFD Handbook*; Wiley & Sons: New York, 1997.
14. Haugen, E.B. *Probabilistic Mechanical Design*; Wiley & Sons: New York, 1980.
15. Chang, H.-Y. Computer aids short-cut distillation design. Hydrocarbon Process. **1980**, *59* (8), 79–82.
16. Hengstebeck, R.J. A simplified method for solving multicomponent distillation problems. Trans. AIChE **1946**, *42*, 309–329.
17. Geddes, R.L. Computation of petroleum fractionation—estimation of A.S.T.M. distillation curves from true boiling-point distillation analyses. Ind. Eng. Chem. **1941**, *33*, 795–801.
18. Thiele, E.W.; Geddes, R.L. Computation of distillation apparatus for hydrocarbon mixtures. Journal of Industrial and Engineering Chemistry **1993**, *25*, 289–295.
19. Fenske, M.R. Fractionation of straight-run Pennsylvania gasoline. Ind. Eng. Chem. **1932**, *24*, 482–485.
20. Underwood, A.J.V. Fractional distillation of multicomponent mixtures. Chem. Eng. Prog. **1948**, *44* (8), 603–614.
21. Gilliland, E.R. Multicomponent rectification: minimum reflux ratio. Ind. Eng. Chem. **1940**, *32*, 1101–1106.
22. Beyer, W.H., Ed. *CRC Handbook of Tables for Probability and Statistics*; The Chemical Rubber Co.: Cleveland, OH, 1966; 228 pp.
23. Suhir, E. *Applied Probability for Engineers and Scientists*; McGraw-Hill: New York, 1997.
24. Montgomery, D.C. *Design and Analysis of Experiments*, 3rd Ed.; Wiley & Sons: New York, 1991.
25. Del Vecchio, R.J. *Understanding Design of Experiments: A Primer for Technologists*; Hanser Publishers: Munich, 1997.
26. Box, G.E.P.; Hunter, W.G.; Hunter, J.S. *Statistics for Experimenters—An Introduction to Design, Data Analysis, and Model Building*; Wiley & Sons: New York, 1978.
27. Hicks, C.R. *Fundamental Concepts in Design of Experiments*, 4th Ed.; Saunders College Publishing: New York, 1993.
28. Ross, P.J. *Taguchi Techniques for Quality Engineering*, 2nd Ed.; McGraw-Hill Professional, 1995.
29. Roy, R.K. *Design of Experiments Using The Taguchi Approach: 16 Steps to Product and Process Improvement*; Interscience, 2001.
30. Brown, B. Chair Specifications for the Chemical and Process Industries: A Manual for Development and Use; ASQC Chemical and Process Industries Division Chemical Interest Committee ASQC, Quality Press: Wisconsin, 1996.

Size Reduction

Sunil Kesavan
Akebono Corporation, Farmington Hills, Michigan, U.S.A.

INTRODUCTION

The comminution or size reduction process is an important unit operation in the process industries in which solid materials are broken or cut into smaller sizes by the application of mechanical stress. Solids can be reduced in size by crushers and grinders which employ compression, impact, attrition, shear, or combinations thereof. Depending on the degree of particle size reduction desired, the end result is achieved in one or several steps. Devices that are used for size reduction operations can be classified into primary and secondary crushers, grinders, and pulverizers. Explosive blasting is used in many instances for primary size reduction of ore formations into sizes workable by primary crushing machines. The primary crushers are slow-speed machines that reduce the run-of-mine product into 15–25 cm lumps. A secondary crusher reduces these lumps to 5 mm product. Grinders reduce the products of crushing operations into powder. An intermediate grinder typically produces a product that passes a 40-mesh screen. Fine grinders reduce feed into product that passes a 200-mesh screen. Ultrafine grinders can convert the product of secondary crushers into 1–10 μ product. Cutting machines produce particles with a definite size and shape in the 2–10 mm size range.

Size reduction machines can be classified as follows:

a. Primary and secondary crushers
    1. Gyratory crushers
    2. Jaw crushers
    3. Roll crushers
b. Intermediate and fine grinders
    1. Impact mills
        a. Hammer mills
        b. Centrifugal pin mill
    2. Attrition mills
    3. Tumbling mills
        a. Ball and pebble mills
        b. Rod mills
        c. Tube mills; compartment mills
    4. Rolling-compression mills
        a. Ring roll mills
        b. Bowl mills
c. Ultrafine grinders
    1. Fluid-energy mills
    2. Agitated mills
    3. Impact mills with internal classification
d. Cutting machines
    1. Knife cutters, dicers, slitters.

GRINDING ENERGY REQUIREMENTS

Given the relatively large power consumption accompanying size reduction, it is important to quantify the energy requirements of these unit processes. Comminution theory dates back to the works of Kick[1] and Rittinger[2] that were done in Germany. A generalized differential equation that governs crushing can be written as:

$$dW = -K(dD/D^n) \quad (1)$$

where W is the energy input, K and n are constants, and D is the particle size.

Eq. (1) represents Kick's law when $n = 1$ and Rittinger's law when $n = 2$.

Kick's law[1] essentially states that the work required to obtain a given reduction ratio is the same irrespective of starting size. According to Rittinger's law[2], work is proportional to surface created. Rittinger's and Kick's laws are only useful over a limited particle size range and are not utilized today.

Bond[3] proposed that work is inversely proportional to the square root of particle diameter. Bond's theory of comminution is represented by Eq. (1) when $n = 1.5$ and can be written as:

$$W = 100W_i(1/D_P^{0.5} - 1/D_F^{0.5})$$

where W_i is the Bond work index or work required to reduce a unit weight of product from a theoretical infinite size to a product with 80% passing 100 μ. Extensive data are available on the work index and the Bond law is widely used today.

PRIMARY AND SECONDARY CRUSHERS

Gyratory Cone Crushers

These types of crushers employ a conical crushing element that gyrates in an eccentric manner in a shell

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120018807

resembling an inverted cone (Fig. 1). The material to be crushed enters the top of the crusher where the crushing surfaces are most widely spaced. The product becomes wedged and squeezed between the mantle and the hopper and is progressively broken down until it discharges through the narrow opening at the bottom of the crusher. Gyratory crushers are available in large sizes used for primary crushing and smaller sizes for secondary crushing of soft to medium-hard materials. These continuous-discharge crushers are available in capacities up to 3500 ton/hr and are more cost effective to operate than jaw crushers.

Jaw Crushers

These devices suited for coarse and intermediate crushing of large volumes of hard and semihard materials employ swinging jaws that work against a hardened stationary surface. The jaws, which are essentially flat, form a V-shaped crushing chamber with a wide inlet at the top and a narrow discharge at the bottom. The large feed material gets progressively reduced in size and falls down toward the narrow throat section.

There are three types of jaw crushers (Fig. 2)—the Blake, the Dodge, and the single-toggle type. In the popular double-toggle Blake jaw crusher (Fig. 3), the moving jaw is hinged at the top with the maximum movement at the bottom of this jaw. In the Dodge crusher, the swinging jaw is hinged at the bottom. This gives a fixed discharge opening, giving a more uniform product than the Blake crusher. The greatest movement acts on the large feed at the top of the jaw. The Dodge crusher is not used much in commercial large volume crushing operations. The single-toggle jaw crusher is pivoted at the top like the Blake crusher but has an additional movement in the downward direction. The Blake crusher costs more than the single-toggle machines but the operating costs are less for hard feed materials.

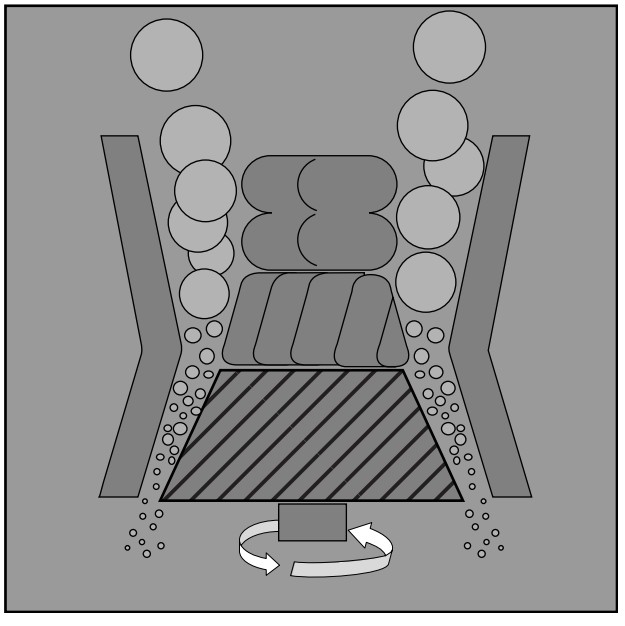


Fig. 1 Schematic representation of the operating principle of a gyratory cone crusher.

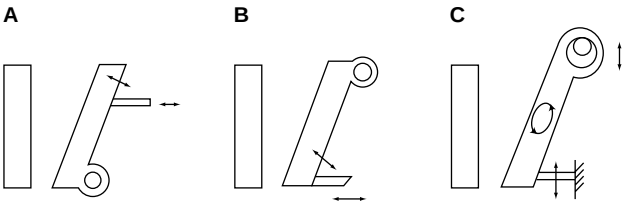


Fig. 2 The essential features of (A) Dodge, (B) Blake, and (C) single-toggle jaw crushers.

Taggart[4] proposed a rule-of-thumb that a jaw crusher would be the economically preferred size reduction equipment if the required throughput in tons per hour was less than the square of the crusher gap in inches. Depending on size, jaw crushers can handle up to a 48-in. size feed and create product smaller than half an inch. Jaw crushers are available in a variety of sizes up to about 1000 ton/hr and can be stationary, portable, or skid mounted.

Roll Crushers

These tertiary crushers employ smooth or toothed heavy-duty impact and abrasion-resistant steel-rimmed rolls. The rolls are mounted inline in a horizontal manner and turn toward each other at equal speeds to create a nip into which a friable feed material is introduced (Fig. 4). Heavy-duty compression springs with automatic reset are used to dampen crushing shock and to protect the crusher from tramp iron and oversize material. An adjustable screw that adjusts spring tension changes the crusher opening. A flywheel is used to even out pulses and economize on power consumption. These crushers have a theoretical maximum reduction ratio of 4 : 1 and will only crush materials to about 10 mesh. Roll crushers produce a controlled product size distribution without a lot of fines. The narrow particle size distribution is achieved by controlling a combination of variables including roll speed, gap measure, differential speed, feed rate, and roll surface.

Toothed single roll crushers that crush materials by working on a breaker plate are also available. The crushing teeth are set in segments to facilitate replacement.

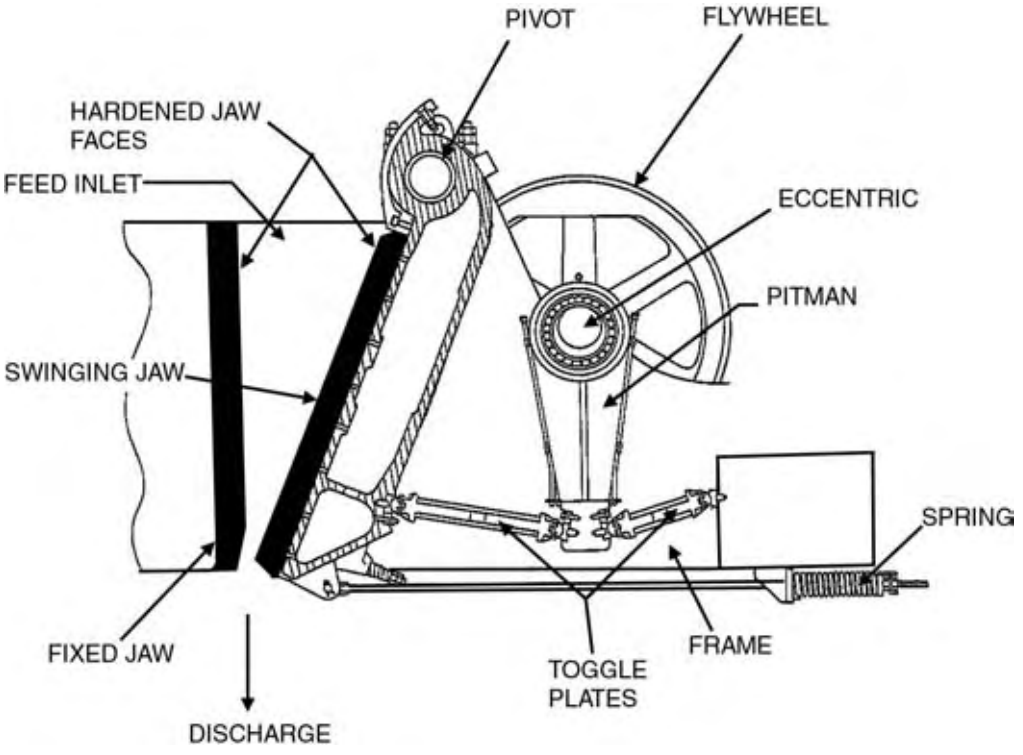


Fig. 3 Cross section of a Blake-type jaw crusher showing the nonchoking design of the swing jaw.

INTERMEDIATE AND FINE GRINDERS

Impact Mills

Hammer mills

Impact mills like the hammer mill use swinging hammers (typically running at 750–1800 rpm) to pulverize the solid feed by impact and attrition. As shown in Fig. 5, these hammers are mounted by pins on the periphery of a number of disks mounted on a horizontal rotating shaft and are free to swing. The hammers force the product against a rugged breaker plate. The product gets broken down until it is small enough to fall through a discharge grate at the bottom. For some applications like the grinding of dried animal byproducts, a vertical hammer mill that uses a vertical drive shaft with horizontal hammers and screens is more efficient. Hammer mills are used in primary, secondary, and tertiary crushing operations. These mills are relatively inexpensive but tend to produce a lot of fine material. They are best suited for soft or semihard materials, as harder materials tend to rapidly wear out the hammers.

Hammer mills can handle feed sizes of up to 10–20 in. and throughputs up to 500 ton/hr. Air-swept hammer

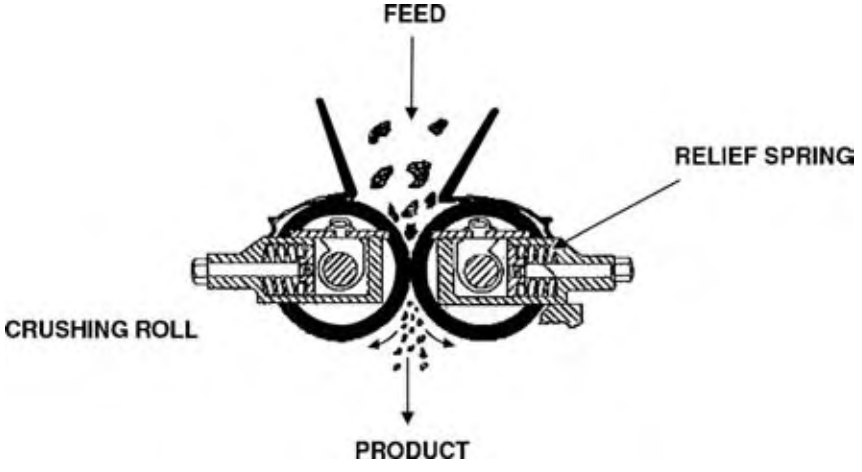


Fig. 4 Schematic of a smooth-roll crusher.

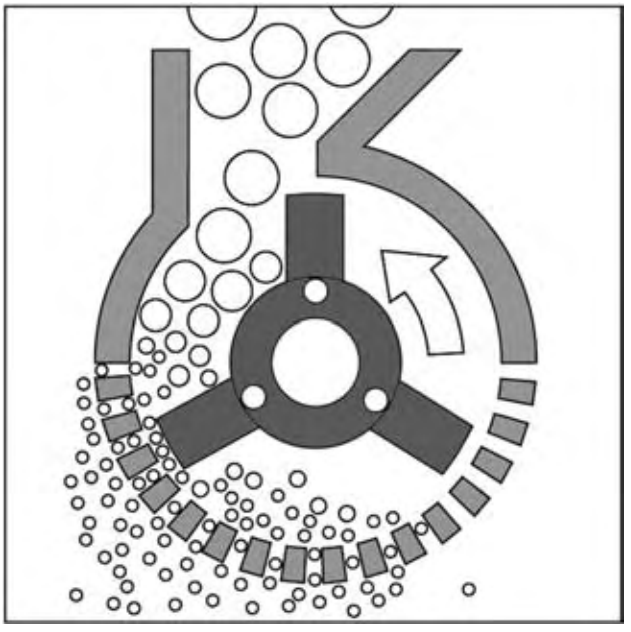


Fig. 5 Schematic representation of the operating principle of a hammer mill. (Courtesy of Sturtevant, Inc., Boston, MA.)

mills use air to help convey the particle product out of the mill.

Centrifugal pin mills

An example of a pin mill is shown in Fig. 6. It applies centrifugal forces to grind feed particles by impact. Feed entering the mill is divided into two streams that drop down on to a rotating plate. Centrifugal force directs the feed outward on to intermeshing pins or blocks mounted on the plate. Pin mills are available in capacities up to 200 ton/hr. These mills are economical, easy to operate, produce a uniform grind, are suitable for wet or dry grinding, and provide high throughput with low energy consumption.

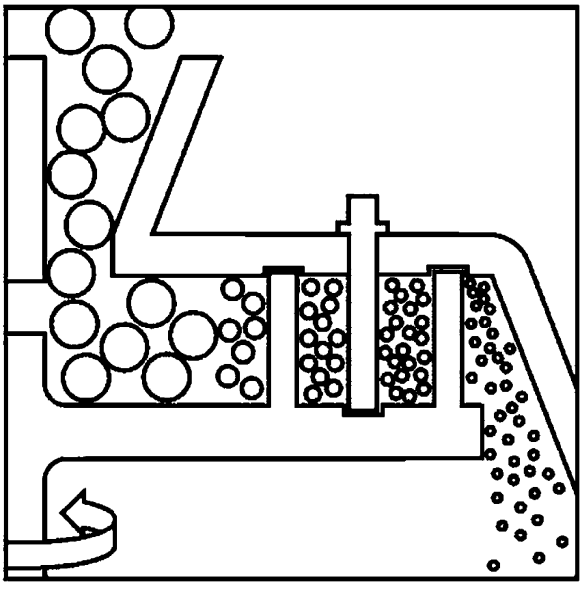


Fig. 6 Simpactor centrifugal pin-type impact mill. (Courtesy of Sturtevant, Inc., Boston, MA.)

Attrition/Disk Mills

These mills use a combination of impact, shear, and cutting action to grind materials between the replaceable wearing surfaces of two grinding disks. One or both of the disks can rotate; if both, they are run counter to each other. The distance between the disks can be adjusted to vary the product particle size. Fig. 7 is a schematic representation of a typical attrition mill. Liquid cooling of the disks is used with heat-sensitive materials to prevent degradation. Air is sometimes drawn through the mill to help remove product and prevent choking. Attrition mills can grind up to 8 ton/hr of product to a particle size passing 200 mesh.

Tumbling Mills

Ball mills

These mills employ attrition and impact to grind the product by tumbling in cylindrical mills partly filled with grinding media. The media can be round metal balls or nonmetallic pebbles that are 1/2 in. or larger in size. Ball mills for grinding hard materials are usually lined with heavy-duty steel alloy liners. Some ball mills employ internal baffles to prevent slippage of the grinding media over the internal shell surface.

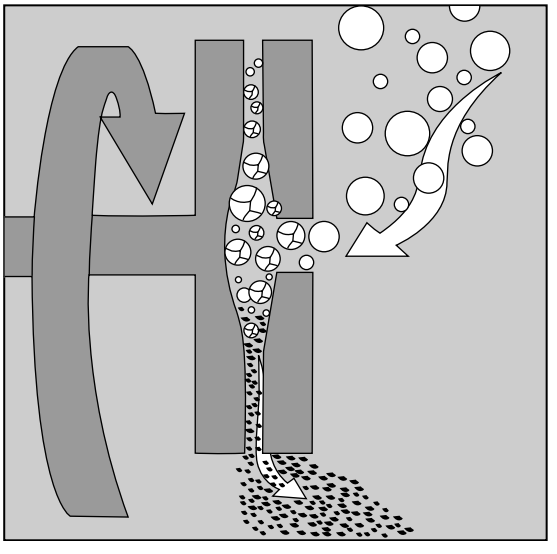


Fig. 7 Schematic of a disk-type attrition mill. (Courtesy of Sturtevant, Inc., Boston, MA.)

Mills for dry grinding are available with either full or semi air-swept capability.

Ball mills can operate either in a batch or a continuous mode. Batch mills typically use a charge of 50–55% of the mill volume, while continuously operating mills use a media charge of 40–45%. In the continuous mode, the mill can operate in such a mode that the product can leave the mill through a discharge grate. When operating in a continuous mode, the effluent of the grinding mill is sent through a classifier to separate out the oversize product to be reprocessed through the mill.

Commonly used types of grinding media are carbon steel, stainless steel, chrome steel, tungsten carbide, ceramic, or zirconia. Ball mills produce up to 50 ton/hr of powder substantially passing a 200-mesh screen. Temperature control can be achieved by the use of jacketed ball mills through which a heat transfer fluid is circulated.

Critical and Operating Speed. In operation, the balls in the mill are carried up in contact with each other and with the walls until the centrifugal force is overcome by the centripetal force. The critical speed N_{cr}, of a ball mill is the speed in rpm above which the grinding media will centrifuge and all milling effectively stops.

The critical speed of a mill is given by:

$$N_{cr} = 42.3/D^{0.5}$$

where D is the internal diameter of the mill in meters.

The operating speed of ball mills is usually 55–75% of the critical speed. Operating close to N_{cr} drastically reduces the effectiveness of the grinding action.

Conical ball mills like the Hardinge mill shown in Fig. 8 have their larger diameter closer to the feed inlet.

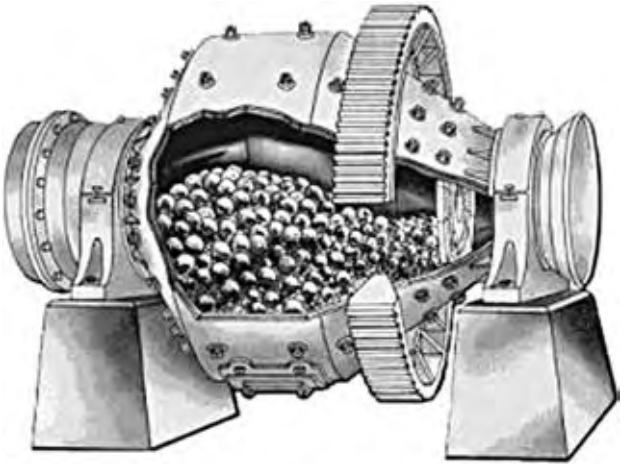


Fig. 8 Hardinge conical ball mill. (Courtesy of Metso Minerals, York, PA.)

The grinding media in the mill fall through different heights depending on their relative location. This provides a classifying action within the mill resulting in increased crushing energy efficiency.

Ball mills with centrifugal and planetary action are also available. Retsch manufactures a planetary ball mill in which multiple grinding chambers are turned around their own axes, and, in the opposite direction, around the common axis of a sun wheel that is driven by a rugged motor. This movement results in the superimposition of centrifugal forces that change constantly (Coriolis motion). The grinding media describe a semicircular motion, separate from the inside wall and collide with the opposite surface at high impact energy. Fig. 9 illustrates the forces encountered in the operation of this grinding mill that can reduce particles to submicron sizes through the imposition of impact and friction forces.

The Retsch centrifugal ball mill uses grinding chambers that move in a horizontal plane at speeds of 100–580 rpm. The centrifugal forces that are generated propel the grinding media against the inside walls of the mill where they roll over the product (Fig. 10). Size reduction is achieved by a combination of impact and friction. The mill is furnished with an automatic reversal system to counter any agglomeration effects and to enhance homogenization.

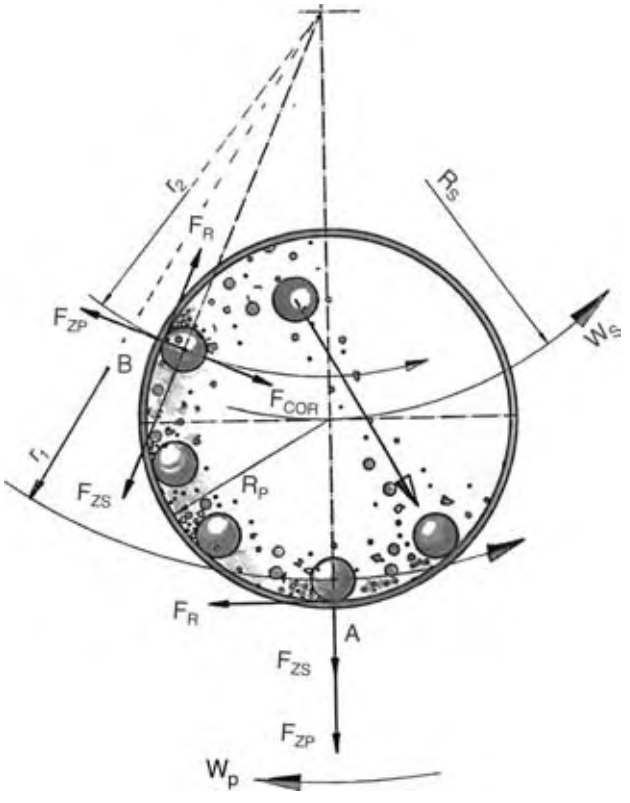


Fig. 9 Schematic of forces operating in a planetary ball mill. (Courtesy of Retsch, Inc., Newtown, PA.)

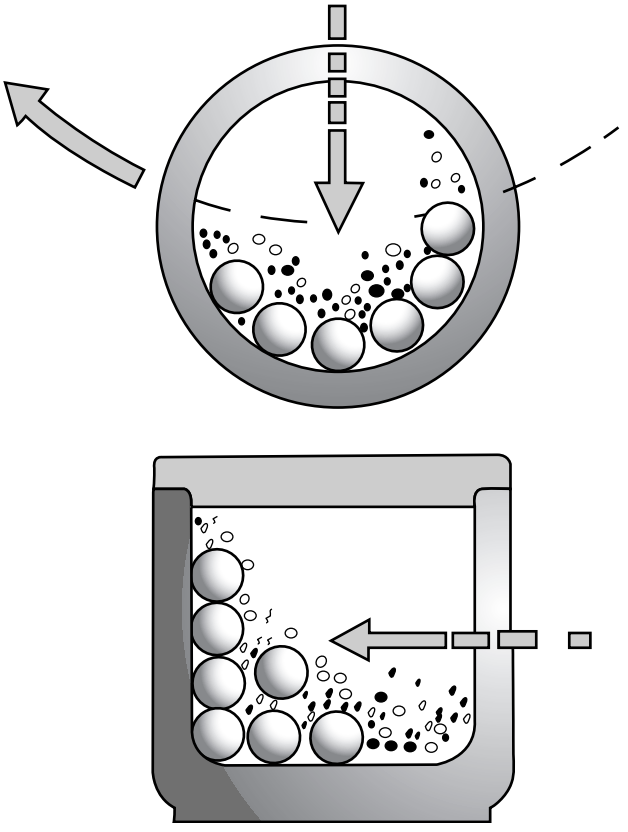


Fig. 10 Schematic of centrifugal ball mill action. (Courtesy of Retsch, Inc., Newtown, PA.)

Tube mills

Tube mills are basically finish-grinding ball mills with mill length several times the diameter. Tube mills can be made with a uniform diameter throughout the mill length or may be made with several compartments with different section diameters and lengths. Compartment mills are tube mills with slotted transverse partitions that separate grinding media of different sizes with the larger media working on coarser product.

Rod mills

Rod mills are regarded as intermediate grinding mills and are basically tube mills that employ grinding rods that are about as long as the mill. Rod mills employ rolling compression and attrition and produce very small amounts of oversize or fines. Feed sizes up to 50 mm (2 in.) can be reduced to product in the 5–10 mesh range in open circuit grinding and to a minus 35 mesh range in closed circuit grinding with a classifying device. These mills require operator attention to minimize rod misalignment or entangling during grinding.

Autogenous and semiautogenous mills

Autogenous mills are mechanically similar to ball mills but employ the product to be crushed as the grinding medium. These mills can be used for hard, soft, as well as sticky products. The product is discharged through a discharge grate designed to retain oversize product. Autogenous grinding reduces metal wear, eliminates secondary and tertiary crushing stages, and offers savings in capital and operating costs. A variation called air autogenous milling uses air to transport and classify the product and uses no discharge screens or grills that can become blocked. Semiautogenous mills use a combination of the feed, product, and grinding media to achieve the desired size reduction.

Rolling Compression Mills

These mills resemble a mortar and pestle in principle and use a rolling ball or roller member moving against the face of a ring or a casing to grind soft to medium-hard product at rates of up to 50 ton/hr. Pressure is applied by either using heavy springs or by the centrifugal force of the roller. Either the roller or ring may be stationary. Built-in air classification is used to improve the grinding efficiency of these mills. The common types of rolling compression mills are rolling-ring pulverizers, bowl mills, and roller mills. These mills are widely used for grinding coal, cement clinker, and limestone.

ULTRAFINE GRINDERS

Fluid-Energy Mills

These energy-intensive mills have no moving parts and use pressurized air, steam, or inert gases to grind particles to ultrafine sizes. The particles to be ground are suspended in a high-velocity fluid stream that carries the particles around in a circular or an elliptical path. The majority of the particle size reduction is accomplished by interparticle attrition with additional grinding caused by particles rubbing against the walls. The particles get classified as they go around the closed path provided by the mill. The larger particles are thrown toward the outer walls and stay in the mill, while the finer particles stay on the inner walls and are removed from the mill through an exit port. The effluent from the mill is sent to a particle separator to collect the ground product. No one jet mill is suitable for all process applications. Several configurations are available to serve specific needs: fluid bed, opposed jet, and multiple-port types.

The Jet-O-Mizer (Fig. 11) from Fluid Energy Processing can produce product with a 1–50 μ average size and a narrow particle size distribution. This

Fig. 11 Cutaway of Jet-O-Mizer showing fluid jets. (Courtesy of Fluid Energy Processing, Telford, PA.) *(View this art in color at www.dekker.com.)*

vertical jet mill can also be used for combined operations like grinding/blending and grinding/coating.

For finer finished product down to the submicron size range, the Sturtevant Micronizer jet mill or the

Fig. 13 Cutaway of the Micro-Jet mill showing fluid jets. (Courtesy of Fluid Energy Processing, Telford, PA.) *(View this art in color at www.dekker.com.)*

Fluid Energy Processing Micro-Jet grinders are suitable. The mills use a circular grinding chamber. The feed metered into the mill enters an air or a gas vortex created by precisely aligned jet nozzles along the mill periphery (Figs. 12 and 13). The tangential angle of the fluid flow causes interparticle impact resulting in size reduction. The desired product and oversize material are separated by air classification in the mill. The mill provides narrow particle size distribution with uniform shape and no heat build-up. Micronizers are available in capacities up to 5 ton/hr.

The Roto-Jet fluid-bed jet mill (Fig. 14) supplied by Fluid Energy Processing is designed to grind products to the 0.5–40 μ average size range with specific top

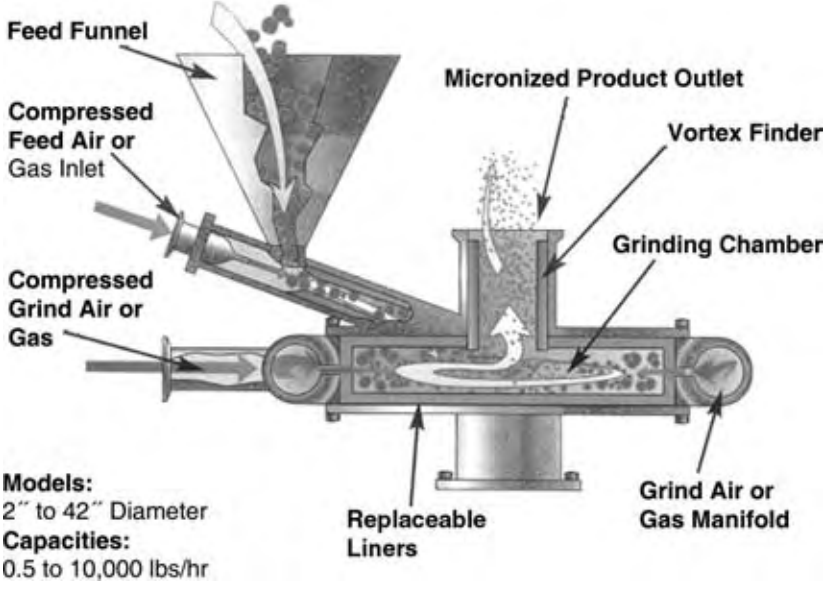


Fig. 12 Operating schematic of the Micronizer jet mill. (Courtesy of Sturtevant, Inc., Boston, MA.) *(View this art in color at www.dekker.com.)*

Fig. 14 Cutaway of Roto-Jet showing fluid jets. (Courtesy of Fluid Energy Processing, Telford, PA.) *(View this art in color at www.dekker.com.)*

and/or bottom size requirements. The machine uses a variable-speed rotor for tight control of product size.

Agitated Mills

Agitated mills are stirred vertical ball mills in which the grinding media are agitated by vibratory energy or by a rotating impeller that can run at speeds up to 1700 rpm. These mills are suited for both wet and dry grinding in batch or continuous mode. The fast grinding machines are energy efficient, compact, easy to operate, and are best suited for reduction of particles to submicron sizes. Particle size reduction is achieved by both shear and impact.

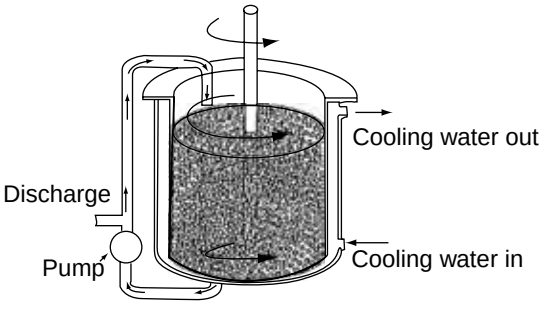


Fig. 15 Batch attritor. (Courtesy of Union Process, Akron, Ohio.)

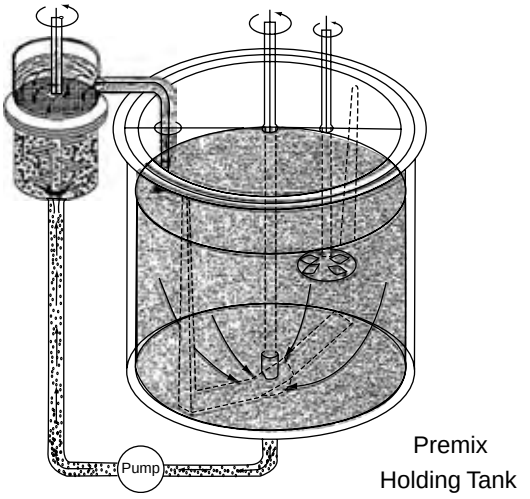


Fig. 16 Circulation attritor. (Courtesy of Union Process, Inc., Akron, Ohio.)

Attritors

The Szegvari attritor uses 1/8–3/8 in. grinding media agitated at speeds up to 350 rpm. Attritors are available that operate in three different modes:

1. Batch attritor shown in Fig. 15.
2. Circulation attritors that use an attritor in association with a holding tank that can hold about

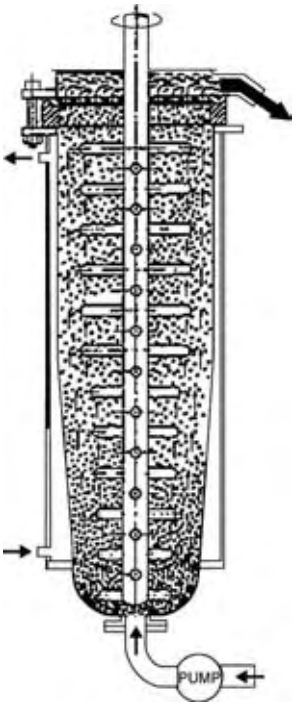


Fig. 17 Continuous attritor. (Courtesy of Union Process, Inc., Akron, Ohio.)

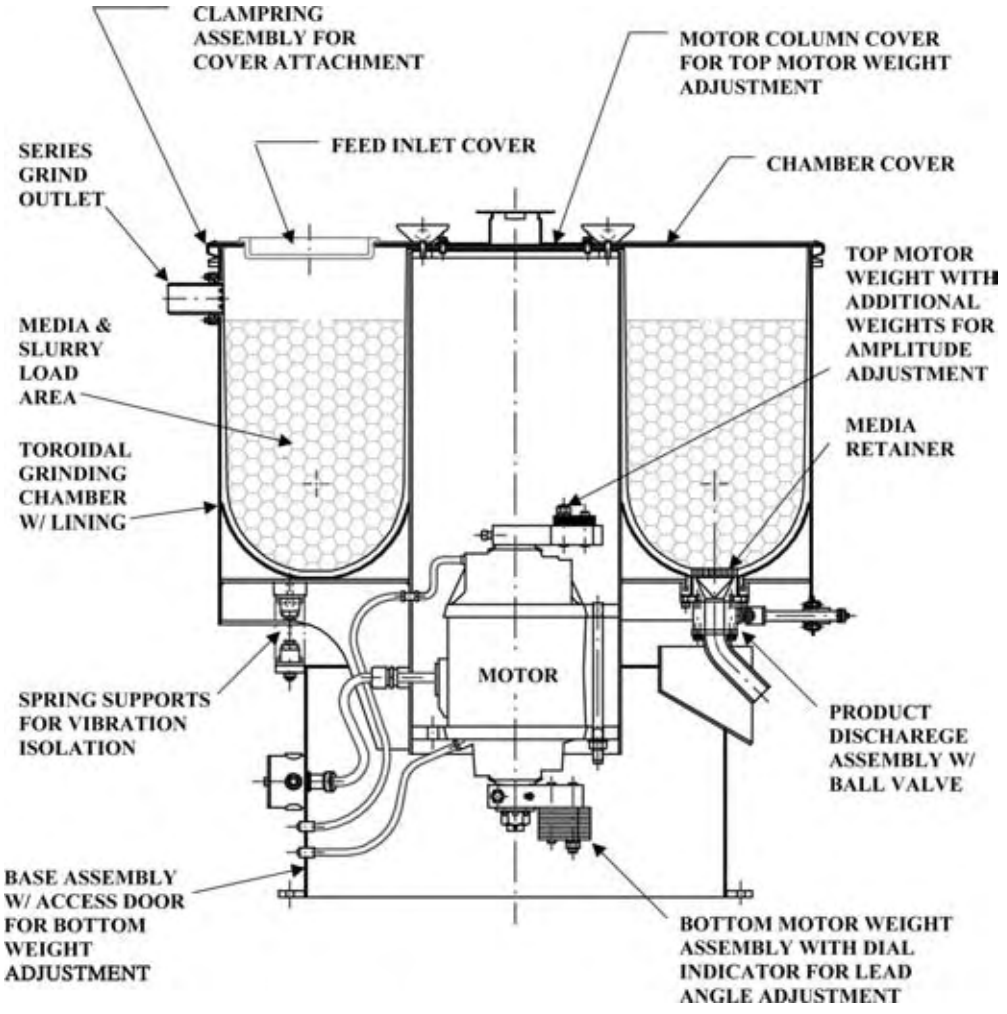


Fig. 18 Cross section of the Sweco vibro-energy mill. (Courtesy of Sweco, Florence, KY.) *(View this art in color at www.dekker.com.)*

10 times the volume of the attritor (Fig. 16). The contents of the holding tank are passed through the attritor multiple times until the desired reduction in size is achieved. This combination allows the use of small attritors for large grinding jobs through the use of high circulation rates.

4. Continuous attritors with the grinding time controlled by the feed pumping rate (Fig. 17).

A high-speed attritor that uses 0.5–3 mm grinding media and impeller speeds of up to 1700 rpm has been developed recently for dry grinding.

Vibratory mills

The Sweco vibro-energy mill (Fig. 18) applies high frequency, three-dimensional vibrations to the grinding chamber that contains small cylindrical grinding media. This helps produce an ultrafine product with a narrow particle size distribution.

The Vibra-Drum vibratory grinding mill supplied by General Kinematics uses a static grinding drum containing grinding media akin to a ball mill. Vibration energy is imparted to the drum from an external mechanism. A subresonant two-mass drive and spring system alternately stores and releases grinding power. Once in motion, the natural frequency design ensures that energy is only needed to move the grinding media as a fluid mass. This mill has lower capital investment, installation, maintenance, and energy costs as compared to conventional rotational mills.

Impact Mills with Internal Classification

An example of a classifying impact mill is the Powderizer marketed by Sturtevant, Inc. (Fig. 19). The feed to the mill enters the grinding chamber where it is broken by the action of impactors/pins mounted on a rotating disk. A column of air sweeps up the pulverized product

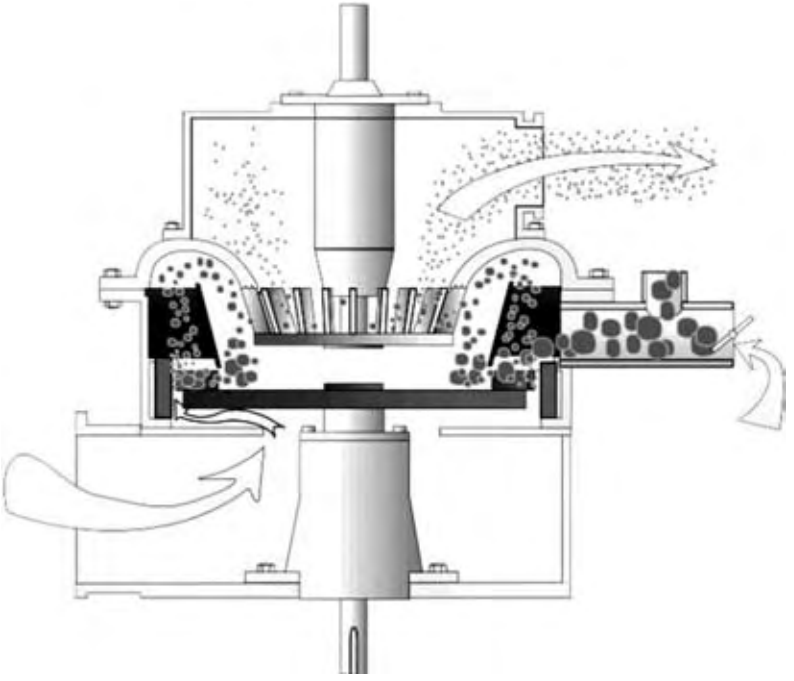


Fig. 19 Schematic of Powderizer air-swept impact mill. (Courtesy of Sturtevant, Inc., Boston, MA.)

through a rotating classifier that rejects oversize products that are returned back for further size reduction. Particle size can be adjusted by changing the classifier speed without shutting down the mill. The product can be pulverized to sizes below 10 μ with narrow size distribution. Pulverizers are available in sizes that can handle 20–7000 kg/hr. The Powderizer is also suitable for heat-sensitive materials.

ROTARY CUTTERS

Cutting mills are used for reducing the size of soft, medium-hard, tough, elastic, fibrous, and temperature-sensitive materials. Examples of products handled in such mills are electronic scrap, film, rubber, foil, foodstuffs, paper and paper products, textiles, and domestic waste. The mills cut, chop, or tear feed using a rotor-mounted set of blades or cutters. Fig. 20 shows the cutting action of the SM2000 cutting mill supplied by Retsch. The feed material is taken up by the rotor and is crushed by the stainless steel cutting strips inserted in the housing. Helically arranged reversible cutting plates of hard metal operate by successive cutting.

EQUIPMENT OPERATION

Open and Closed Circuit Grinding

Comminution is one of the most inefficient unit operations from the viewpoint of energy consumption. Continuous crushing and grinding equipment can operate in an open circuit mode, where the product being crushed passes through the mill only once. This mode of operation is inefficient as a considerable amount of energy is wasted in regrinding product that has reached the desired product size. A more efficient mode of operation is called closed circuit grinding (Fig. 21). In closed circuit grinding, the mill discharge is sent to size classification equipment that separates out the product that has reached the desired size.

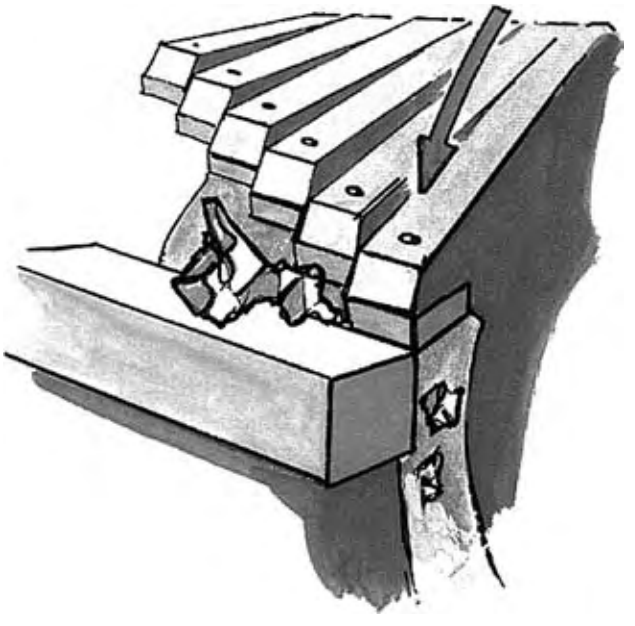


Fig. 20 Schematic showing the cutting action of the SM2000 cutting mill. (Courtesy of Retsch, Inc., Newtown, PA.)

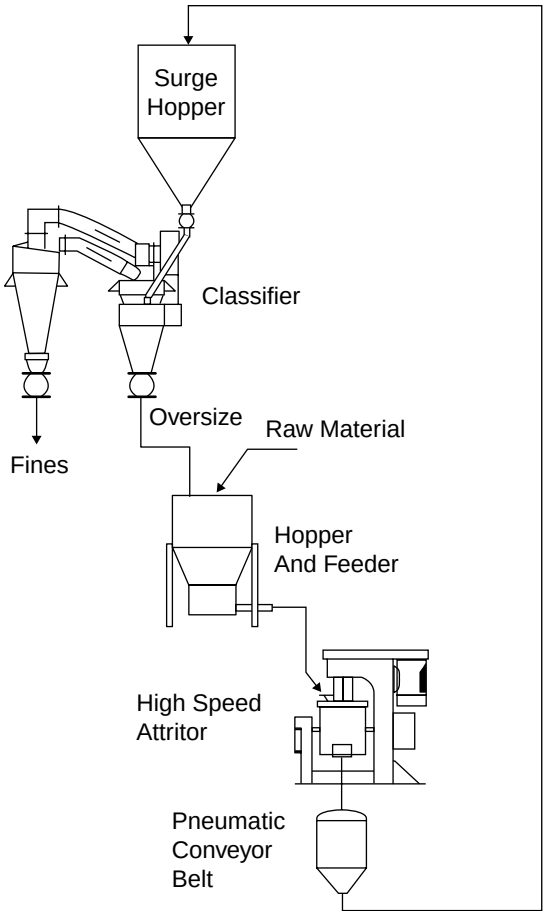


Fig. 21 Representation of a closed-circuit grinding operation employing an attritor. (Courtesy of Union Process, Inc., Akron, Ohio.)

Oversize material is returned to the mill for further size reduction. Closed circuit grinding is suitable for either wet or dry grinding of materials.

CONCLUSIONS

The greatest advancements in grinding technology in the recent past have been made in ultrafine grinding equipment like fluid-energy mills that can generate products in the micron and submicron ranges. This trend is expected to continue as comminution processes evolve to address technical issues in the production of ultrafine particles for high technology applications. As comminution is a very important unit operation in the conversion of run-of-mine raw materials, a considerable amount of effort was spent in the early and middle part of the last century to develop efficient size reduction equipment. The basic principles employed in the operation of crushing and grinding machines have not changed substantially over the past few decades. Saving energy, reducing costs, and cutting pollution have been the main areas targeted for continuous improvement in size reduction operations. The technical areas for future improvement of comminution processes fall into the following general categories: 1) Advanced sensors to provide integrated online physical and chemical characterization of feed and product to help in automated process control; 2) Utilization of the knowledge of real-time feed characteristics to optimize comminution processes; 3) Improved modeling of grinding mill operations like three-dimensional simulation of charge motion in tumbling mills; 4) Advanced abrasion-resistant milling media and surfaces; and 5) Novel or improved physical separation processes to be used in conjunction with size reduction equipment. As in the past, it is expected that future developments in conventional comminution will be evolutionary rather than revolutionary.

REFERENCES

1. Kick, F. *Das Gasetz der propertionalen Widerstande und siene Anwendung*; Arthur Felix: Leipzig, 1885.
2. Rittinger, P.R. *Lehrbuch der Aufbereitungskunde*; Ernst and Korn: Berlin, 1867.
3. Bond, F.C. The theory of comminution—Meeting of AIME, Mexico City, October, 1951. Trans. Am. Inst. Min. Metall. Pet. Eng. **1952**, *193*, 484–494.
4. Taggart, A.F. *Handbook of Mineral Dressing*; Wiley: New York, 1945.

Soave's Modified Redlich–Kwong Equation of State

J. Richard Elliott, Jr.
Department of Chemical Engineering, University of Akron, Akron, Ohio, U.S.A.

INTRODUCTION

Throughout the 1970s and 1980s, chemical engineering process design underwent a virtual revolution. By increasing computational speed and the accuracy of thermodynamic models, computer simulation transformed the design process from hand calculations with charts and tables to a modern engineering methodology. In 1996, the Chief Executive Officer of Dow Chemical, Frank Popoff, proclaimed, ''Process modeling is the single technology that has had the biggest impact on our business in the last decade.'' Soave's equation played a central role in that revolution. It was the first thermodynamic model to provide feasible accuracy for physical properties of a wide variety of compounds, over a wide range of conditions. Even now, applications and adaptations of Soave's equation abound. What follows is a short presentation of the nature of these applications and adaptations, how Soave's equation relates to other equations of state, and why Soave's approach has stood the test of time. An excellent reference on this subject is provided in Soave's own words with a greater emphasis on the historical context of Soave's development.[1]

THREE-PARAMETER CORRESPONDING STATES

The primary key to Soave's early success was the recognition of the need to extend from two-parameter to three-parameter corresponding states. With two parameters, it is possible to fit the critical temperature and critical pressure of any compound, but a general description of the vapor pressure curve is unattainable. The slope of the reduced vapor pressure ($P_r^{sat} \equiv P^{sat}/P_c$) curve varies substantially from one compound to another, but a two-parameter equation of state assumes that it is invariant. Because the vapor pressure varies as an exponential of temperature, small errors near the critical temperature are greatly amplified at reduced temperatures ($T_r \equiv T/T_c$) near 0.45.

Vapor pressure is a key property in modeling phase behavior because it represents the vapor–liquid equilibria (VLE) of the pure fluid. Accurate vapor pressure characterization is essential to VLE correlation. As the most common operation in chemical processing is distillation, accurate modeling of VLE had a tremendous impact on process modeling.

Characterization of the vapor pressure curve was readily accessible in the form of Pitzer's acentric factor. The definition of the acentric factor is

$$\omega \equiv -1 - \log_{10}(P^{sat}/P_c)_{|Tr=0.7}$$

Hence, the third parameter, ω, implicitly contains information about the vapor pressure, making vapor pressure prediction something like a circular loop. But Soave went beyond this simple observation. Wilson had previously recognized these issues,[2,3] but his equation met with limited success, especially at low reduced temperatures. Soave was careful to analyze the temperature dependence of his equation of state in great detail at the outset. He achieved this by introducing an adjustable parameter into the attractive contribution of the Redlich–Kwong[4] equation.

$$Z \equiv \frac{PV}{RT} = Z^{ig} + Z^{rep} + Z^{att} = 1 + \frac{b\rho}{1 - b\rho} - \frac{a}{bRT}\frac{b\rho}{1 + b\rho} \quad (1)$$

where ρ is the molar density, Z the compressibility factor, Z^{ig} the ideal gas contribution, Z^{rep} the repulsive contribution, Z^{att} the attractive contribution, P absolute pressure, T the absolute temperature, R the gas constant, and a and b are parameters characterizing each compound. The repulsive contribution approaches infinity as the density approaches a close-packed value. The attractive part depends on temperature and approaches zero at infinite temperature.

For pure fluids, a and b can be characterized by matching the critical criteria: $(\partial P/\partial\rho)_T = 0$; $(\partial^2 P/\partial\rho^2)_T = 0$. The resulting equations are

$$a = \frac{\alpha R^2 T_c^2}{P_c}\left(\frac{1}{9(2^{1/3} - 1)}\right); \quad b = \frac{RT_c}{P_c}\left(\frac{(2^{1/3} - 1)}{3}\right) \quad (2)$$

where α is a special temperature dependent adjustable parameter. Soave's principal modification of the Redlich–Kwong equation was to redefine α. Whereas,

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007869

the original Redlich–Kwong definition was effectively:

$$\alpha \equiv \sqrt{T_c/T} \tag{3}$$

Soave modified the definition of α to be:

$$\alpha \equiv \left[1 + \kappa\left(1 - \sqrt{T/T_c}\right)\right]^2 \tag{4}$$

where

$$\kappa = 0.480 + 1.574\omega - 0.176\omega^2 \tag{5}$$

Note that the discussion so far relates only to the model for pure fluids, while the primary application is for VLE of mixtures. Extending the model to mixtures requires characterization of the a and b parameters for the mixture. These characteristic equations are called mixing rules. They are discussed in the section titled ''Mixing Rules.''

Soave's method of developing this definition is instructive. First, he tabulated values of α that would exactly match the experimental vapor pressures of methane through n-hexane for $T_r \in [1.0, 0.3]$. Then, he plotted these with respect to temperature. As that plot did not generate linear correlations, he plotted several other candidate relations. By shortening the temperature range to $T_r \in [1.0, 0.45]$ and plotting $\sqrt{\alpha}$ vs. $\sqrt{T_r}$, Soave obtained a simple linear trend. He noted that α was constrained to unity at the critical temperature by its definition, and that the acentric factor establishes a second point for this linear trend, resulting in the form of Eq. (4). This form is sufficient for ~3% accuracy in the vapor pressure of hydrocarbons, if the critical properties and acentric factor are accurately known.

Note that the term ''critical properties'' in the present context refers to the critical temperature and pressure, not including the critical density. Matching the critical density might be considered an alternative version of three-parameter corresponding states that deserves a brief review in the present context. Hougen and coworkers explored this alternative for a number of years, but they focused on density in the critical region rather than vapor pressure. Density in the critical region is difficult to measure and fraught with theoretical challenges related to the nonclassical behavior of the coexisting densities in the critical region. Although relatively unimportant for most process simulations, density is important in the critical region for supercritical extractions. Despite these challenges, Lee and Kesler were able to develop an equation of state that fits the critical density and the vapor pressure curve. The Lee–Kesler equation interpolated between the equations of state for methane and octane in the form of:

$$Z = Z^{CH_4} + \omega(Z^{nC_8} - Z^{CH_4}) \tag{6}$$

By adapting the multiparameter Benedict–Webb–Rubin (BWR) equation for methane and n-octane as their basis, they approached an effective implementation of what might be called as four-parameter corresponding states. Unfortunately, their equation never achieved the widespread application of the Soave equation, despite having greater accuracy for liquid density and comparable accuracy for VLE. This observation may shed some light on an additional strength of the Soave equation. Soave's equation was much simpler than the multiparameter BWR two-compound interpolation form of the Lee–Kesler equation. Soave's equation can be rearranged into the form of a cubic polynomial, which can be solved analytically, eliminating the convergence difficulties that plagued early programmers. Thus, the popularity Soave's equation relative to the Lee–Kesler equation is a testament to the merits of a simpler form.

APPLICATIONS AND ADAPTATIONS OF SOAVE'S EQUATION

Applications of the Soave equation are too numerous to list comprehensively. Soave's original paper has been cited in roughly 1600 publications since 1980, and 88 times in 2003 alone. These citations include applications to petroleum recovery and refining, natural gas production, coal liquefaction, cryogenic gas separation, refrigeration, wax precipitation, hydrate formation, polymerization, interfacial tension, supercritical fluid processing, and petrochemical production. Interested readers should perform a ''cited reference'' search of the citation index to obtain a detailed list. Restricting the search to 2003 suffices to indicate the scope of coverage listed above.

Beyond applications reported in the literature, process simulation is ubiquitous throughout the chemical industry and academia. Every time users select the Soave, or SRK, thermodynamics model, they apply the Soave equation. One motivation for this selection is the long experience with the model and the compilation of correction factors, when the basic model is deficient. For example, binary interaction parameters (k_{ij}'s) have been compiled for a large number of binary mixtures to improve VLE correlation.[5] It is also possible to compensate for inaccuracies in density through volume translations.[6]

In light of the broad scope of applications and adaptations, the discussion here is restricted to general observations about the nature of adaptations and the accuracy to be expected. The Peng–Robinson equation

is included as an early adaptation as given below:

$$Z \equiv \frac{PV}{RT} = 1 + \frac{b\rho}{1 - b\rho} - \frac{a}{bRT}\frac{b\rho}{1 + 2b\rho + (b\rho)^2} \tag{7}$$

Some might consider the Peng–Robinson equation to be entirely independent of the Soave equation, but they have used exactly the same form for α that Soave developed in Eq. (4). Therefore, the only difference is the form of the denominator. The Peng–Robinson denominator results in slightly improved density correlations, especially for the low-boiling compounds of interest to the Gas Processors Association, sponsors of the Peng–Robinson research. In general however, density predictions are inaccurate, such that volume translations like those of Ref.[6] are still worthwhile. Moreover, comparisons generally indicate only marginal distinctions between the accuracy of the two equations for VLE correlations.[7,8] Hence adaptations like extended α correlations and advanced mixing rules are equally applicable to the original Soave form or the Peng–Robinson form with little distinction. We draw the line at the Peng–Robinson equation, however. Other equations of state might also be considered as incremental adaptations. For example, replacing the repulsive term with a more realistic hard sphere term might be considered to be a small adaptation, but that would be considered a major revision by this author.

One further adaptation of Soave's equation for pure fluids played a major role in maintaining the relevance of his model over the years. Soave himself was aware that his α correlation was deficient for mixtures involving hydrogen. This deficiency is not surprising when it is realized that, at room temperature, hydrogen's reduced temperature approaches $T_r \sim 10$. Graboski proposed an improved α correlation for hydrogen mixtures. Graboski also proposed a modified form of Eq. (5) based on optimizing the coefficients with a larger database than Soave's. Graboski's modification is occasionally referred to as the API equation. Soave's original α correlation is also deficient for polar compounds. Whereas the slope of the vapor pressure curve for hydrocarbons tends to drop at low temperatures, it remains relatively constant for polar compounds. Several authors proposed α adaptations for the Soave and Peng–Robinson equations. Notable references should include Stryjek and Vera,[9] Patel and Teja,[10] Soave,[1] and references cited therein. The disadvantage of these adapted α correlations is that they are no longer predictive. Whereas the acentric factor and Eq. (5) suffice for the original model, specific regressions must be performed for each compound of interest, when applying adapted α correlations. Stryjek and Vera[9] have tabulated the parameters for roughly 50 common compounds in their adaptation for the Peng–Robinson equation.

MIXING RULES

Another major adaptation of Soave's equation involved application to asymmetric nonideal mixtures. The original mixing rules proposed by Soave were in the quadratic form originally suggested by van der Waals as given below:

$$b = \sum_{i=1}^{NC} x_i b_i; \quad a = \sum_{i=1}^{NC}\sum_{j=1}^{NC} x_i x_j a_{ij};$$
$$a_{ij} = \sqrt{a_{ii}a_{jj}}(1 - k_{ij}) \tag{8}$$

Quadratic mixing rules are sufficient for symmetric deviations from ideality, but not for highly asymmetric mixtures like alcohols mixed with hydrocarbons. In symmetric mixtures, the k_{ij} parameter can be adjusted to characterize large nonidealities, but it primarily adjusts the magnitude of the Gibbs excess energy, and not the skewness. Negative values of k_{ij} indicate strong solvation interactions, as in inorganic acid + water mixtures. Positive values of k_{ij} indicate weak solvation interactions, as in acetone + hydrocarbon mixtures. In mixtures like alcohols + hydrocarbons, both the magnitude and skewness of the Gibbs excess curve are affected, because the alcohols associate except at infinite dilution.

Huron and Vidal[11] showed that equating the infinite pressure Gibbs energy of mixing to that of an activity model like the NRTL[12] or UNIQUAC[13] models provided a mixing rule that was sufficiently flexible to describe very complex phase behavior. With this modification, simple cubic equations like Soave's could be applied to nearly any kind of mixture at any conditions, including supercritical conditions. The Huron–Vidal mixing rule combined with NRTL activity model is illustrated below.

$$\frac{a}{b} = \sum_{i=1}^{NC} x_i \left[\frac{a_i}{b_i} - \frac{1}{\ln 2}\frac{\sum_{j=1}^{NC} x_j G_{ji} C_{ji}}{\sum_{k=1}^{NC} x_k G_{ki}}\right];$$
$$b = \sum_{i=1}^{NC} x_i b_i \tag{9}$$

$$G_{ji} = b_j \exp\left(-\alpha_{ji}\frac{C_{ji}}{RT}\right) \tag{10}$$

where C_{ji}, C_{ij}, and α_{ji} are adjustable binary parameters. (There is no relation between α_{ji} of the mixing rule and Soave's pure component α.)

A notable adaptation that combines complex mixing rules, the Stryjek–Vera α correlation, and the Peng–Robinson equation is the PRWS model of Wong

and Sandler.[14] The principal difference from the Huron–Vidal rule is that the Wong–Sandler rule degenerates to the quadratic mixing rule at low density. Technically, quadratic mixing at low density is required for thermodynamic consistency, but the small magnitude of nonidealities in the vapor phase means that the inconsistencies there have small impact on the accuracy of the correlation. The Wong–Sandler mixing rule is

$$b = \frac{Q}{1 - D} \tag{11}$$

$$\frac{a}{RT} = \frac{QD}{1 - D} \tag{12}$$

$$Q = \sum_i \sum_j x_i x_j \left(b - \frac{a}{RT}\right)_{ij} \tag{13}$$

$$D = \sum_{i=1}^{\mathrm{NC}} x_i \left[\frac{a_i}{b_i RT} - \frac{\sqrt{2}}{\ln(\sqrt{2} - 1)} \frac{\sum_{j=1}^{\mathrm{NC}} x_j G_{ji} C_{ji}}{\sum_{k=1}^{\mathrm{NC}} x_k G_{ki}}\right] \tag{14}$$

$$\left(b - \frac{a}{RT}\right)_{ij} = \left[\left(b_i - \frac{a_i}{RT}\right) + \left(b_j - \frac{a_j}{RT}\right)\right] \frac{(1 - k_{ij})}{2} \tag{15}$$

Note that k_{ij} values vary over large ranges $(-1, 0.5)$ in the PRWS model. An implementation of the PRWS model is available through the author's website.

A number of related efforts have studied the impact of applying the infinite pressure limit, or the zero pressure limit.[15] The zero pressure limit would appear to be closer to the basis applied in developing activity models. Once again, the more complex model is technically more correct, but the improvement in accuracy is small. A reasonable compromise in accuracy and simplicity is offered by the PRWS mixing rule.

In general, the Huron–Vidal mixing rule provides flexibility but the optimal parameters for the activity model alone are not optimal when adapted to the advanced mixing rule, necessitating regression of each specific binary mixture. Michelsen and Dahl,[16] on the other hand, showed how UNIFAC predictions could be implemented directly into the mixing rules, circumventing the regression step. This is the basis of the predictive SRK model of Holderbaum and Gmehling.[17] Gmehling has published a recent review that favors adaptation of a volume translated Peng–Robinson equation with temperature-dependent parameters in the complex mixing rules to improve excess enthalpy correlation.[18] His model is also capable of treating electrolytes and polymer solutions[19] within this general methodology. The inclusion of such a large number of adjustable parameters is made feasible by maintaining an extremely large database with perpetual global optimization of all parameters for all data.

An alternative to incorporating empirical activity models into the mixing rules is to seek an alternative physical explanation of solution nonidealities. Hydrogen bonding models are noteworthy in this regard. Most of the nonideality in the interaction between water and oil, for example, can be attributed to water hydrogen bonding while oil cannot. A hydrogen-bonding model would account for this observation explicitly, reducing the magnitude of the parameters in the mixing rules. Most inclusions of hydrogen bonding have also included corrections of the Van der Waals repulsive contribution,[20,21] and elimination of any parameter resembling Soave's α correlation. Such extensive generalizations are judged to be outside the scope of the this entry. On the other hand, Tassios and coworkers[22] have managed to include a hydrogen bonding contribution while maintaining the Soave model for the remainder of the equation of state. Unfortunately, including the hydrogen bonding contribution does not guarantee elimination of the need for complex mixing rules. Hence, the benefit of including an additional term in the model is not readily apparent. Ultimately, this benefit may be recognized as molecular modeling becomes integrated with chemical process and product design. Its future prospects are discussed below.

FUTURE

It is reasonable to ask, "What next? Is there anything left to be done?" Of course, the need for research is never ending. The primary criticism against the current incarnation of Soave's initiative is the lack of a strong relation between the form of Soave's equation and the rigorous physics that can be determined from statistical mechanics. For example, the van der Waals repulsive contribution is known to be inconsistent with molecular simulations.[23] Furthermore, there is no basis in statistical mechanics for the $\sqrt{T_r}$ dependency in Soave's α correlation. The presumption is that an improved physical basis would improve the capability for extrapolating the model to conditions at which experimental measurements are unavailable. So far, this presumption has been impossible to prove. By perpetually introducing new parameters and new data and reoptimizing, Gmehling and coworkers have demonstrated that adaptation of Soave's equation is

sufficiently flexible to correlate the vast majority of what is known about phase equilibria. One caveat in this regard would be variations related to "proximity effects." These are changes that depend on details of the molecular structure that first order group contributions cannot address. One example would be differences between *p*-xylene and *m*-xylene. One obvious answer to this problem is higher order corrections for proximity effects, and these are on the way. By proceeding in this manner, it will be many years before such adaptations of Soave's equation are supplanted.

Alternatives to Soave's approach and group contribution adaptations would be better focused on capabilities not offered by such approaches. For example, pure component properties like vapor pressure are assumed to be available when applying Soave's methodology. In the coming world of chemical product design, this assumption may not be satisfactory. Molecular simulation offers the prospect of being able to make these predictions for transport properties as well as equilibrium properties.[24–27] Proximity effects would also be naturally included within molecular modeling. While the National Research Council has estimated that such predictive capability may not be available for a "decade or two,"[28] viable preliminary versions may come much sooner than that.

The rise of molecular product design will ultimately lead to questions about molecular simulations of mixtures, which make it difficult to rationalize distinctions between molecular scale models and models that derive their justification primarily from macroscopic measurements. When this happens, systematic analysis that can be rationalized on both the molecular and macroscopic scales will take precedence. It is in this context that hydrogen-bonding models will begin to play a more significant role. Such a development would present little difficulty to researchers like Gmehling because they could easily add new terms to their model equations. The effort will be motivated by complementary goals that can be achieved through a consistent molecular perspective applied to phase equilibria, transport phenomena, adsorption, membrane permeation, ion exchange, protein folding, and a host of other diverse applications. Thus, phase equilibrium can teach us things about molecular interactions that can be applied in fields that are far removed from traditional phase equilibrium applications. This is where the future lies and Soave's equation has paved the way to such an ambitious goal.

CONCLUSIONS

The SRK equation has played a major role in the development of chemical process design and will continue to play a similar role. Adaptations may obscure the role of the original model, but it can be discerned through the recollection of a brief history. These adaptations include the Peng–Robinson equation, extended vapor pressure correlations, the Huron–Vidal and Wong–Sandler mixing rules, and group contribution extensions by Gmehling and coworkers. The strength of the SRK approach has been a heavy emphasis on correlating key engineering data. As future adaptations continue to recognize which data are key and accurately correlate larger databases, the SRK approach will be perpetuated for many years to come.

REFERENCES

1. Soave, G. 20 years of Redlich–Kwong equations of state. Fluid Phase Equilibr. **1993**, *82*, 345–359.
2. Wilson, G.M. Vapor–liquid equilibria correlated by means of a modified Redlich-Kwong equation of state. Adv. Cryog. Eng. **1964**, *9*, 168–176.
3. Wilson, G.M. Calculation of enthalpy data from a modified Redlich-Kwong equation of state. Adv. Cryog. Eng. **1966**, *11*, 392–400.
4. Redlich, O.; Kwong, N.S. On the thermodynamics of solutions. V. An equation of state. Fugacities of gaseous solutions. Chem. Rev. **1949**, *44*, 233.
5. Elliott, J.R.; Daubert, T.E. Revised procedures for phase equilibrium calculations with the Soave equation of state. Ind. Eng. Chem. Proc. Des. Dev. **1985**, *24*, 743.
6. Peneloux, A.; Rauzy, E.; Freze, R. A consistent correction for Redlich–Kwong–Soave volumes. Fluid Phase Equilibr. **1982**, *8*, 7.
7. Graboski, M.S.; Daubert, T.E. A modified Soave equation of state for phase equilibrium calculations. I. Hydrocarbon systems. Ind. Eng. Chem. Proc. Des. Dev. **1978**, *17*, 443.
8. Yang, J.; Griffiths, P.R.; Goodwin, A.R.H. Comparison of methods for calculating thermodynamic properties of binary mixtures in the sub and super critical state: Lee–Kesler and cubic equations of state for binary mixtures containing either CO_2 or H_2S. J. Chem. Thermo. **2003**, *35*, 1521–1539.
9. Stryjek, R.; Vera, J.H. PRSV—An improved Peng–Robinson equation of state for pure compounds and mixtures. Can. J. Chem. Eng. **1986**, *64*, 323–333.
10. Patel, N.C.; Teja, A.S. New cubic equation of state for fluids and fluid mixtures. Chem. Eng. Sci. **1982**, *37*, 463–473.
11. Huron, M.J.; Vidal, J. New mixing rules in simple equations of state for representing vapour–liquid equilibria of strongly non-ideal mixtures. Fluid Phase Equilibr. **1979**, *3*, 255.

12. Renon, H.; Prausnitz, J.M. Estimation of parameters for the NRTL equation for excess Gibbs energies of strongly nonideal liquid mixtures. Ind. Eng. Chem. Proc. Des. Dev. **1969**, *8*, 413.
13. Abrams, D.S.; Prausnitz, J.M. Statistical thermodynamics of liquid mixtures: a new expression for the excess Gibbs energy of partly or completely miscible systems. AIChE J. **1975**, *21*, 116.
14. Wong, D.S.H.; Sandler, S.I. A Theoretically correct mixing rule for cubic equations of state. AIChE J. **1992**, *38*, 671.
15. Twu, C.H.; Coon, J.E.; Bluck, D.; Tilton, B. CEOS/AE mixing rules from infinite pressure to zero pressure and then to no reference pressure. Fluid Phase Equilibr. **1999**, *158*, 271.
16. Michelsen, M.; Dahl, S. High-pressure vapor–liquid equilibrium with a UNIFAC-based equation of state. AIChE J. **1990**, *36*, 1829–1836.
17. Holderbaum, T.; Gmehling, J. PSRK: a group contribution equation of state based on UNIFAC. AIChE J. **1991**, *70*, 251–265.
18. Gmehling, J. Potential of group contribution methods for the prediction of phase equilibria and excess properties of complex mixtures. Pure Appl. Chem. **2003**, *75*, 875–888.
19. Wang, L.S.; Ahlers, J.; Gmehling, J. Development of a universal group contribution equation of state. 4. Prediction of vapor–liquid equilibria of polymer solutions with the volume translated group contribution equation of state. Ind. Eng. Chem. Res. **2003**, *42*, 6205–6211.
20. Chapman, W.G.; Gubbins, K.E.; Jackson, G.; Radosz, M. New reference equation of state for associating liquids. Ind. Chem. Eng. Res. **1990**, *29*, 1709.
21. Elliott, J.R.; Suresh, S.J.; Donohue, M.D. A simple equation of state for nonspherical and associating molecules. Ind. Eng. Chem. Res. **1990**, *29*, 1476.
22. Voutsas, E.C.; Kontogeorgis, G.M.; Yakoumis, I.V.; Tassios, D.P. Correlation of liquid–liquid equilibria for alcohol/hydrocarbon mixtures using the CPA equation of state. Fluid Phase Eq. **1997**, *132*, 61–75.
23. Honnnell, K.G.; Hall, C.K. A new equation of state for athermal chains. J. Chem. Phys. **1989**, *90*, 1841.
24. Chen, B.; Potoff, J.J.; Siepmann, J.I. Monte Carlo calculations for alcohols and their mixtures with alkanes. Transferable potentials for phase equilibria. 5. United-atom description of primary, secondary, and tertiary alcohols. J. Phys. Chem. B. **2001**, *105*, 3093–3104.
25. Martin, M.G.; Siepmann, J.I. Transferable potentials for phase equilibria. 1. United-atom description of *n*-alkanes. J. Phys. Chem. B. **1998**, *102*, 2569–2577.
26. Bourasseau, E.; Haboudou, M.; Boutin, A.; Fuchs, A.H.; Ungerer, P. New optimization method for intermolecular potentials: optimization of a new anisotropic united atoms potential for olefins: prediction of equilibrium properties. J. Chem. Phys. **2003**, *118*, 3020–3034.
27. Unlu, O.; Gray, N.; Gerek, Z.N.; Elliott, J.R. Transferable step potentials for the straight chain alkanes, alkenes, alkynes, ethers, and alcohols. Ind. Eng. Chem. Res. **2004**, *43*, 1788–1793.
28. Breslow, R.; Tirrell, M.V., Eds. *Beyond the Molecular Frontier—Challenges for Chemistry and Chemical Engineering*; The National Academies Press: Washington, DC, 2003.

Solid–Liquid Mixing: Numerical Simulation and Physical Experiments

Philippe A. Tanguy
Francis Thibault
Gabriel Ascanio
URPEI, Department of Chemical Engineering, Ecole Polytechnique, Montreal, Quebec, Canada

Edmundo Brito-de La Fuente
Departamento de Alimentos y Biotecnología, UNAM, México, México

INTRODUCTION

Solid–liquid mixing processes find numerous applications in industry. The modeling of these mixing operations using first principles is still limited, although some progress has been made in recent years. This study assesses the real capability of the network-of-zone approach in the case of a complex mixing problem involving a coaxial mixer. Coaxial mixers are very popular for the preparation of pastes and slurries in the chemical, food, and coating industries. The mixer considered is composed of two rotating shafts: a fast-driven shaft supporting an open impeller and a slow shaft driving a scraping anchor arm.

LITERATURE SURVEY

The dispersion of solids in liquids, the preparation of solid suspension in liquid media, and the make-down of pigment slurries in agitated vessels are typical solid–liquid mixing problems that find numerous applications in the process industries. For instance, they are involved in the preparation of paints and coatings, the manufacturing of food products, as well as suspension polymerization.

In solid–liquid mixing design problems, the main features to be determined are the flow patterns in the vessel, the impeller power draw, and the solid concentration profile versus the solid concentration. In principle, they could be readily obtained by resorting to the CFD (computational fluid dynamics) resolution of the appropriate multiphase fluid mechanics equations. Historically, simplified methods have first been proposed in the literature, which do not use numerical intensive computation. The most common approach is the dispersion–sedimentation phenomenological model. It postulates equilibrium between the particle flux due to sedimentation and the particle flux resuspended by the turbulent diffusion created by the rotating impeller. Based on this concept and assuming a one-dimensional distribution of particles along the vessel height, an analytical expression for the concentration profile can be derived:

$$\frac{\phi}{\bar{\phi}} = \frac{Pe_s}{1 - \exp(-Pe_s)} \exp(-Pe_s z/H) \quad (1)$$

where $\bar{\phi}$ represents the average volume fraction of the suspension, z the vertical coordinate, H the fluid height in the vessel, and Pe_s the Peclet number of the solid particles. The Peclet number is defined as

$$Pe_s = v_t \times H/D_{e,p} \quad (2)$$

where v_t is the settling velocity at equilibrium and $D_{e,p}$ the turbulent diffusion coefficient. This dimensionless number is a model fitting parameter, and it is generally correlated with the operating conditions and the physical properties of the suspension.[6] The model has been verified experimentally in a tank provided with four Rushton turbines (radial discharge impeller) at a very low solids concentration.[1,2] Several improvements have been proposed in the literature to make the model more general. Ferreira, Rasteiro, and Figueiredo[3] introduced new terms in Eq. (1) to take into account the radial variation of the particle concentration. Rasteiro, Figueiredo, and Friere[4] suggested the use of the Richardson and Zaki's expression[5] for the settling velocity, as this velocity is strongly dependent on the solid volume fraction in the suspension.

The application of CFD in the modeling of solid–liquid mixing is fairly recent. In 1994, Bakker et al.[6] developed a two-dimensional computational approach to predict the particle concentration distribution in stirred vessels. In their model, the velocity field of the liquid phase is first simulated taking into account the flow turbulence. Then, using a finite volume approach, the diffusion–sedimentation equation along with the convective terms is solved, which includes D_s, a

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120018056

turbulent diffusion coefficient of the particles, defined as:

$$D_s = \frac{\sqrt{k_t}}{3\pi n_p D_p^2} \tag{3}$$

where n_p is the overall volume of the particles, D_p their diameter, and k_t the turbulent kinetic energy density. As the authors did not use iterative coupling between the computation of the flow field and that of the solids concentration, the effect of the solids on the fluid mechanics in the vessel was not taken into account. As an example, they investigated the suspension of 20 μm particles with several radial discharge impellers. They showed that their approach was capable of predicting the solid–liquid interface in the vessel and the effect of the position and number of turbines on the solid distribution. Unfortunately, no experimental validation of this approach was carried out.

All the above models do not consider the particle–particle interactions, although these interactions influence the settling velocity and ignore the effect of the solid phase on the hydrodynamics in the vessel. As a consequence, the practical range of application is restricted to low solids concentrations.

A particle migration model was proposed by Gadala-Maria and Acrivos[7] to describe experimental shear-induced migration observations. This model allows for a better understanding of the shear effects on particle diffusion for concentrated suspensions.[8–11] Based on these studies, a conservation equation for the solid phase was established by Phillips, Amstrong, and Brown,[12] which takes into account convective transport, diffusion due to particle–particle interactions, and the variation of viscosity within the suspension, namely:

$$\frac{\partial \phi}{\partial t} + v \cdot \nabla \phi = \nabla \cdot \left(K_c a^2 \phi \nabla(|\dot{\gamma}|\phi)\right) + \nabla \cdot \left(K_\eta \frac{a^2}{\eta} |\dot{\gamma}| \phi^2 \frac{\partial \eta}{\partial \phi} \nabla \phi\right) \tag{4}$$

where ϕ is the volume fraction, v the suspension velocity, a the particle radius, $|\dot{\gamma}|$ the magnitude of the rate-of-strain tensor. K_c and K_η are empirical constants equal to 0.41 and 0.62, respectively. In order to describe the variation of the viscosity η with the particle volume fraction, the authors suggested the use of the Krieger–Dougherty phenomenological model.[13] This approach was applied to investigate the diffusion of suspensions consisting of poly(methylmethacrylate) monodisperse particles ($D_p = 675\,\mu m$) at high concentration ($\phi > 45\%$) in Newtonian silicon oil. Two geometries were tested: a Couette flow (flow between parallel plates with one plate in motion) and a Poiseuille flow (flow in a cylindrical channel). For these two cases, as the concentration varies only in the radial direction, several analytical expressions could be established for the solid volume fraction and the suspension velocity profile. The computational results were compared to concentration measurements based on nuclear magnetic resonance and a qualitative agreement was obtained. It should be noted here that in principle Eq. (4) can be applied at any solids concentration, but it is however restricted to noncolloidal systems.

A different numerical strategy to simulate multiphase mixing was introduced by Mann [14] and Mann and Hackett.[15] The idea of the method, called the network-of-zone, is to subdivide the flow domain in a set of small cells assumed to be mixed perfectly. The cells are allowed to exchange momentum and mass with their neighboring cells by convective and diffusive fluxes. Brucato and Rizzuti[16] and Brucato et al.[17] applied this idea to the modeling of solid–liquid mixing. An unsteady mass balance for the particles was derived to estimate the solid distribution in the vessel, namely:

$$V_c \frac{dC_c}{dt} = Q(C_p - C_c) + \sum_{i=1}^{2} \alpha Q(C_i - C_c) + \sum_{i=3}^{5} (v_t S_i C_i - v_t S_c C_c) \tag{5}$$

where the subscript c is the cell on which the balance is applied, the subscript i refers to the adjacent cells, and the subscript p denotes the feeding cell, i.e., the one yielding convective momentum, V is the cell volume, C the concentration in the cell, S the cell surface area where the particles settle from, Q the volume flow rate of fluid entering the cell, α an adjustable parameter that describes the turbulent diffusion, and v_t the particle settling velocity. In this model, the sedimentation and the diffusive flow were in the vertical direction, and the convective flow was radially oriented. These assumptions were justified on the basis of the radial discharge flow generated by the impellers. Brucato et al.[17] verified the model prediction with the experimental data of Fajner et al.[2] and Magelli et al.[1] It was shown that the axial concentration profile was very well predicted, however the validation was limited again to extremely low concentration suspensions.

The attractiveness of the network-of-zone method to compute solid–liquid mixing flows resides in its relative simplicity while being capable of capturing the main flow phenomena for a wide range of concentrations. The objective of the present work is to assess the real capability of this approach in the case of a complex mixing problem involving a coaxial mixer. Coaxial mixers are very popular for the preparation of pastes and slurries in the chemical, food, and coating industries. Another mixer setup is also tested,

which consists of a single marine propeller rotating in a vessel (see description in the appropriate section below). This setup is used to validate the numerical model.

DESCRIPTION OF THE COAXIAL MIXER SETUP

The experimental setup is shown in Fig. 1. It consists of the following items: (1) an AC drive with a nominal speed of 1760 rpm; (2) a gearbox with a reduction ratio 3.53:1 yielding a rotating speed of 500 rpm; (3) a torquemeter with a measurement range between 0 and 22.6 Nm (accuracy of $\pm$ 0.1% at full scale); and (4) a transparent vessel with a hemispherical bottom.

Table 1 Operating range of the agitators

Speed ratio N_c/N_a	N_a (rpm)	N_c (rpm)
0 (anchor only)	(0–125)	0
4	(0–125)	(0–500)
8	(0–62.5)	(0–500)
24	(0–20.8)	(0–500)
∞	0	(0–500)

The vessel diameter D_c and the height of the cylindrical section H_c are both equal to 40.64 cm, yielding a maximum volume of fluid of about 60 L. The impeller configuration is the following:

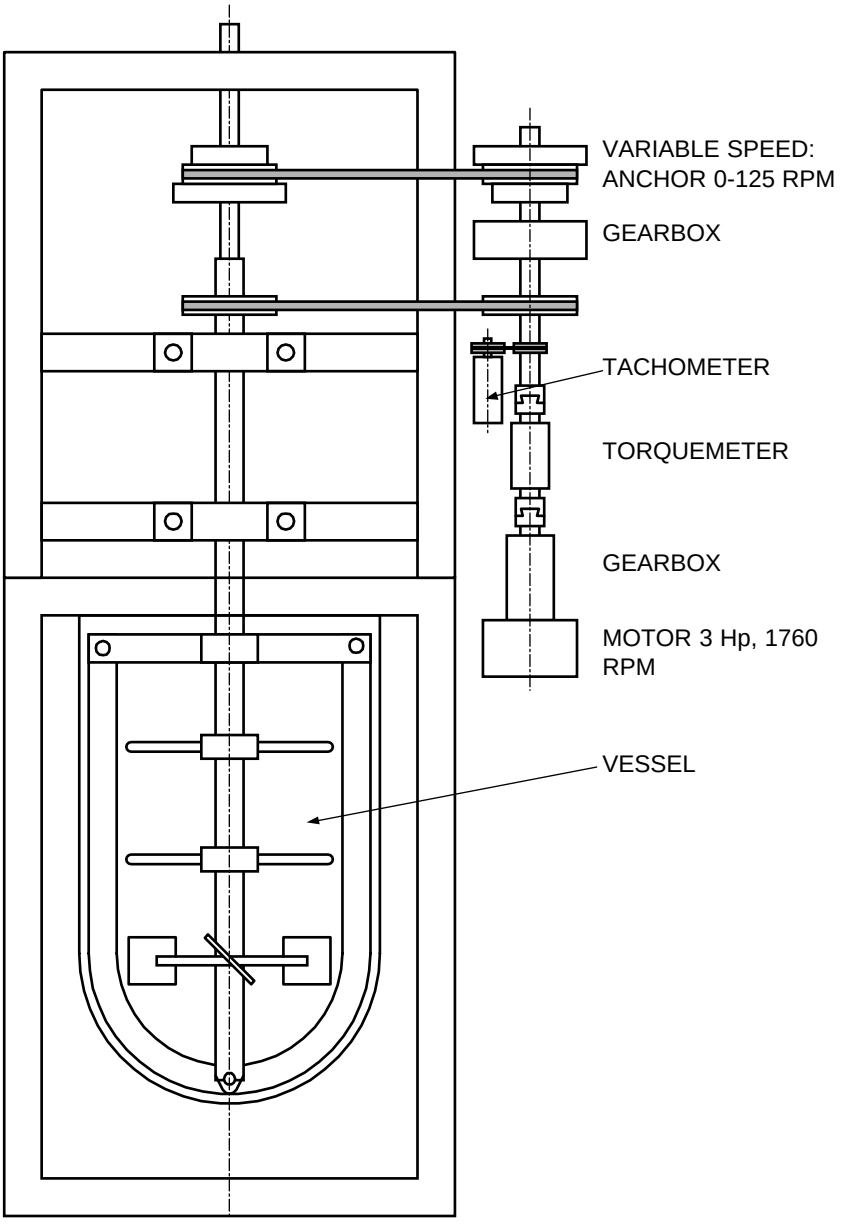


Fig. 1 Coaxial mixer setup.

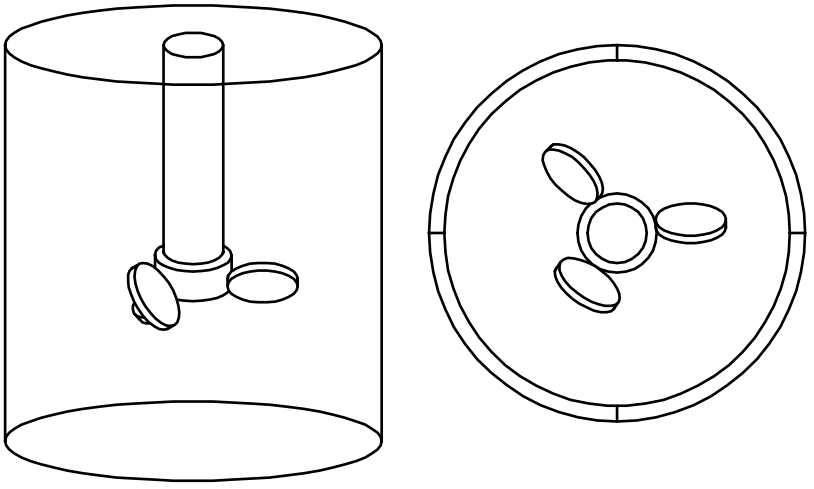


Fig. 2 Marine propeller mixer.

An anchor arm:

- Diameter $D_a = 36.83\,\text{cm}$
- Width $W_a = 3.81\,\text{cm}$
- Wall clearance $C_w = 1.9\,\text{cm}$

Four rigid rods at 90° used for pigment wetting:[18]

- Length $D_t = 23.77\,\text{cm}$
- Cross-section diameter $D_{st} = 0.95\,\text{cm}$

A pitched blade turbine with two blades at 45°:

- Diameter $D_p = 20.2\,\text{cm}$
- Width $W_p = 5\,\text{cm}$
- Length $L_p = 7\,\text{cm}$
- Bottom clearance $C_b = 20.32\,\text{cm}$.

This configuration yields the following dimensionless ratios $C_w/D_c = 0.047$, $D_a/D_c = 0.906$, and $W_a/D_a = 0.103$ for the anchor, $D_t/D_c = 0.585$ for the rods; and $D_p/D_c = 0.5$ and $C_b/D_c = 0.5$ for the pitched blade turbine.

In this coaxial mixer, the primary role of the anchor is to clean up the wall from any accumulated solid lumps and reincorporate them back in the bulk. It also acts as a moving baffle, hampering the creation of a vortex at the liquid free surface. The purpose of the pitched blade turbine is to provide axial pumping so as to promote the resuspension of the solids, and radial dispersion to avoid solids reagglomeration. Finally, the aim of the wetting rods is to ease hydrophobic pigment incorporation by avoiding the creation of surface lumps.[19]

Two driving shafts are installed in this mixer: a fast rotating shaft drives the four rods and the turbine in a counterclockwise direction at speed N_c, whereas a slow rotating shaft entrains the scraping anchor in the clockwise direction at speed N_a. The operating range used in this work was as described in Table 1. In order

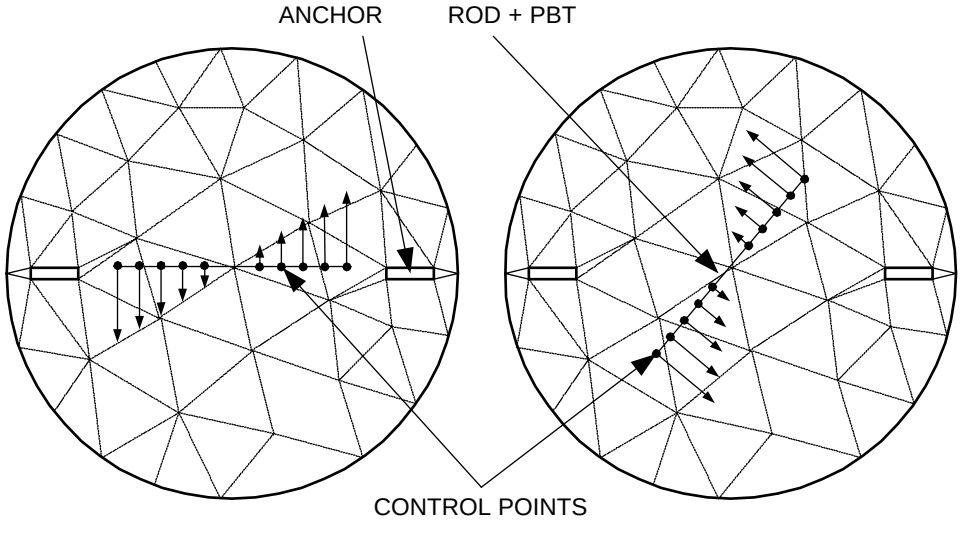


Fig. 3 Virtual finite elements method concept in 2-D.

to investigate the particle motion inside the coaxial mixer, a Newtonian solution of corn syrup with a viscosity of 1.05 Pa.s and a density of 1360 kg/m^3 was used in conjunction with red Ballotini glass beads. The beads had an average diameter of 1 mm and a density of 2500 kg/m^3. The maximum packing factor (ϕ_m) was 0.6.

MARINE PROPELLER MIXER

This simple mixing system (Fig. 2) involves a marine propeller in an unbaffled vessel (actually a laboratory beaker). The geometrical characteristics are the following:

- Distance shaft-impeller edge = 2.6 cm
- Blade diameter = 1.8 cm
- Bottom clearance = 2.6 cm

The propeller rotates clockwise in a down-pumping mode. The vessel (glass beaker) has a diameter of 7.2 cm and the fluid height is 6.5 cm corresponding to a stirred volume of 264 cm^3. The same fluid as the one considered in the coaxial mixer experiments was used. The rotating speed and the volume concentration of the particles were varied in order to investigate the resuspending mechanism in such a mixer. The operating conditions were as follows: $N = 173$ rpm and $\bar{\phi} = 2.8\%$; $N = 230$ rpm and $\bar{\phi} = 7.1\%$; and $N = 350$ rpm and $\bar{\phi} = 11.9\%$. For all the experiments, the particles were initially at rest on the tank bottom and the stirrer was suddenly set in motion.

Computational Model

Let us consider the incompressible flow of a suspension in a given domain Ω. The governing equations are:

$$\rho_m\left(\frac{\partial \mathrm{v}}{\partial t} + \mathrm{v}\cdot\mathrm{grad}\,\mathrm{v}\right) + \mathrm{grad}\,p + \mathrm{div}\,\tau = \rho_m g \quad \text{in } \Omega \tag{6}$$

$$\mathrm{div}\,\mathrm{v} = 0 \quad \text{in } \Omega \tag{7}$$

where v, p, g, and ρ are the velocity, pressure, gravity, and specific gravity, respectively. For a solid–liquid medium, the density ρ_m can be expressed as:

$$\rho_m = \rho_l(1 - \phi) + \rho_s\phi \tag{8}$$

where ρ_l and ρ_s represent the density of the liquid phase and the solid phase, respectively.

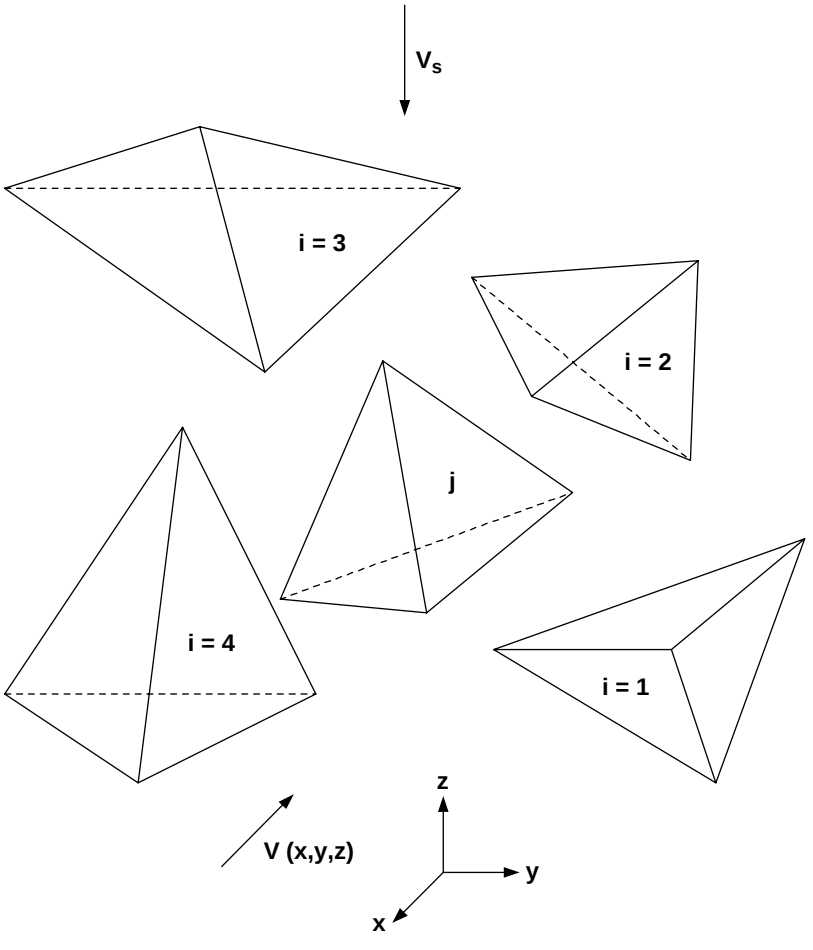


Fig. 4 Tetrahedral finite volumes.

The stress tensor τ in Eq. (6) is related to the rate-of-strain tensor by a rheological equation of state such as:

$$\tau = -2\eta_s\dot{\gamma} \tag{9}$$

where η_s is a function of $|\dot{\gamma}|$ and ϕ,

$$\dot{\gamma} = \tfrac{1}{2}[\text{grad v} + (\text{grad v})^{\text{T}}] \tag{10}$$

and

$$|\dot{\gamma}| = (\dot{\gamma} : \dot{\gamma})^{\frac{1}{2}} \tag{11}$$

The suspension viscosity η_s may or may not be a function of $|\dot{\gamma}|$ depending on the rheological behavior of the suspending medium. In a nondilute suspension, it is, however, always a function of the particle volume fraction ϕ. In this work, the Krieger–Dougherty model for a Newtonian suspension was used:

$$\eta_s(\phi) = \eta_l(1 - \phi/\phi_m)^{-1,82} \tag{12}$$

where η_l is the viscosity of the suspending liquid and ϕ_m the maximum packing factor.

For mathematical convenience, boundary conditions and initial conditions must be prescribed. For the simple marine propeller problem, a Lagrangian viewpoint was adopted. The frame of reference was attached to the propeller so that the propeller was fixed but the vessel was rotating. The boundary condition was then a zero velocity on the impeller, while the vessel wall rotated at $-\Omega_{\text{impeller}}$. The free surface was considered to be flat, therefore the normal velocity was zero and a shear-free condition was assumed. It should be noted that in the Lagrangian viewpoint, the frame of reference is in rotation. The fluid is therefore subjected to a constant acceleration and the momentum conservation equation [Eq. (6)] must be modified to account for centrifugal forces and Coriolis forces.[20] An advantage is, however, that the flow can be solved numerically at steady state provided the flow is fully periodic, which limits the computational efforts significantly.

In the case of the coaxial mixer, the rotation kinematics is much more complex since the two sets of agitators counter-rotate at different speeds. For the sake of simplicity, we decided to simulate the flow using the frame of reference of the anchor. In this Lagrangian viewpoint, the anchor is fixed but the vessel wall rotates at $-\Omega_{\text{anchor}}$ and the turbine rotates at $\Omega_{\text{anchor}} + \Omega_{\text{turbine}}$. In such a situation, contrary to the simple propeller problem, the resolution of the flow equations is time-dependent as the position of the central agitator changes with time.

The finite element method was used for the discretization of the flow equations. Considering the complex kinematics in the coaxial mixer and the associated change of topology at each time step, a new mesh should a priori be built for every topology considered in the time discretization. As a large number of time steps would be required to depict the agitator kinematics accurately, this approach would be a tremendous chore. To alleviate this difficulty, several alternatives have been proposed in the literature:

1. The description of the agitator by momentum sources or sinks inside the domain.[21] The major drawback of this method is the evaluation of a force equivalent to the representation of a body in rotation.
2. The arbitrary Euler–Lagrange method. It consists of moving the finite element mesh nodes with time.[22] This method works well as long as the mesh is not too distorted. In practice, remeshing is usually required after a few time steps.
3. Domain decomposition with sliding meshes.[23] This strategy is very popular in the finite volume literature and implemented in finite volume-based software (Fluent™, CFX™, and Star-CD™). The idea is to decompose the flow domain into several concentric cylindrical meshes and allow slipping of the meshes between the partitions. The continuity of the solution at the mesh interface is imposed by

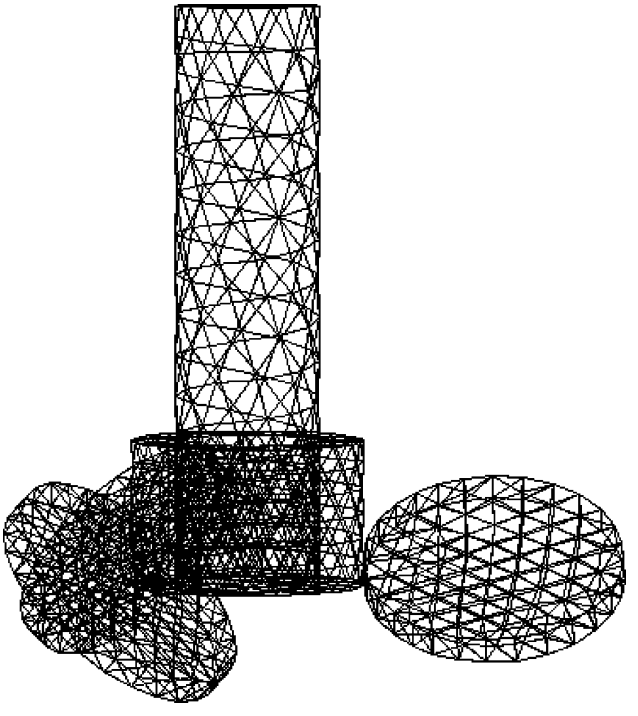


Fig. 5 Mesh of the propeller.

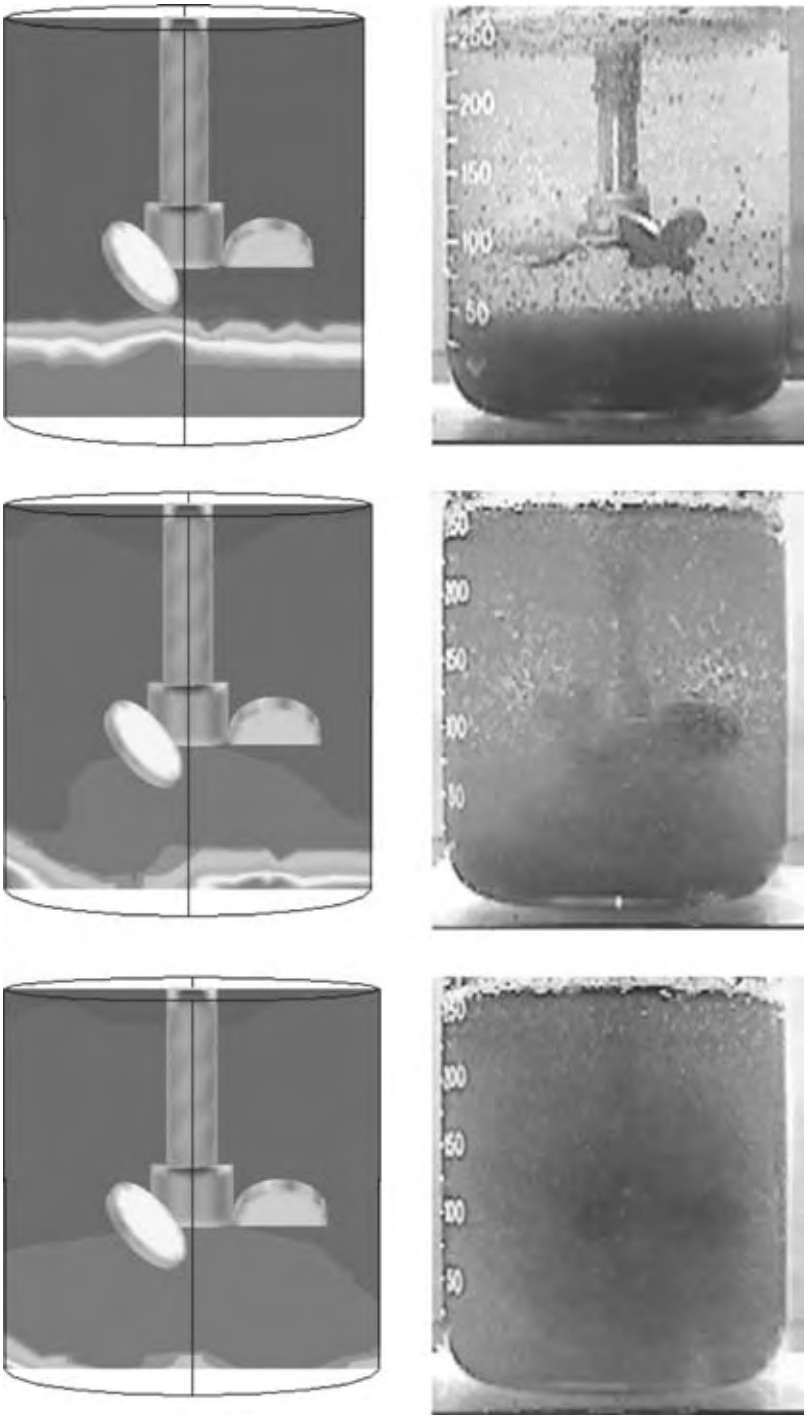


Fig. 6 Suspending mechanism as a function of time.

conservative interpolation. This method is powerful but limited to simple agitator configurations. It seems to work when very fine meshes are used although no error analysis has been published so far.

4. Mesh superimposition. This technique used in commercial finite element CFD software, like FIDAP^{TM} and PolyFlow^{TM}, consists of generating a volume mesh without the moving impellers and a surface mesh of each impeller. At each time step, the surface mesh is projected in the volume mesh and a procedure has been developed to determine if the nodes of the volume mesh are located inside the surface mesh. When it is the case, the velocity of the impeller is imposed on these nodes. This technique is fairly simple to implement, however, it does not allow a precise representation of the impeller shape.

Table 2 Variation of the average concentration with the rotation speed

N (rpm)	Average volume fraction	Predicted volume fraction
173	2.8%	0.5%
230	7.1%	1.7%
350	11.9%	7.3%

5. The virtual finite element method VFEM.[24] The method that was chosen for this work has been developed specifically to simulate flows around moving bodies in fixed enclosure. The principle of VFEM is the following: a volume mesh of the vessel with the anchor arm is first generated. Then, the surface of the moving bodies (in our case, the impeller and the rods) is meshed; the discretization nodes generated being stored as control nodes. Knowing the kinematics of the impellers, the velocity is imposed on these control nodes as constraints in the momentum equation and their position is updated with time according to the impeller rotation. A constrained optimization technique based on Lagrange multipliers and a penalty strategy is used to impose the velocity constraints. The VFEM approach was applied by Bertrand, Tanguy, and Thibault[24] to investigate the complex flow patterns in a planetary kneader and by Tanguy and Thibault[25] for the characterization of the hydrodynamics in a coaxial mixer provided by a helical ribbon and a Rushton turbine. Contrary to the mesh superimposition technique, the VFEM does not require that the control nodes coincide with the volume mesh nodes. They can be located inside finite elements as they are treated like external solicitations or optimization constraints. Fig. 3 illustrates the VFEM concept in 2-D. The above method was implemented in our POLY3-D^{TM} CFD finite element code using unstructured meshes made of tetrahedral elements. The reader is referred to Bertrand, Tanguy, and Thibault[24] for detailed information.

The modeling of solid–liquid mixing requires an additional equation to predict the dispersion of the solid phase in the vessel. As mentioned before, the network-of-zones approach was used in this work, which is based on unsteady mass balances on the solid phase carried out on a set of cells. In the literature, such mass balances are predominantly made on regular cells (finite volumes) in structured grids. The cells typically consist of quadrangles in 2-D and hexahedra in 3-D. In the present work, due to the use of unstructured grids, the mass balances were performed on the same elements as those used for the resolution of the flow equations, i.e., tetrahedral finite volumes (Fig. 4).

Let us consider the following mass balance on a tetrahedral element subjected to the velocity field $v(x,y,z)$ of the suspension (as computed from the solution of the flow equations) and a sedimentation velocity, namely:

$$V_j \frac{d\phi_j}{dt} = \sum_i Q_{cj,i}\phi^*_{cj,i} + Q_{sj,i}\phi^*_{sj,i} \tag{13}$$

where the subscript i represents the neighboring finite elements adjacent to the four sides of finite element j, V_j the volume of finite element j, $Q_{cj,i}$ the convective flux of solid particles going through a common face to elements i and j, $Q_{sj,i}$ the sedimentation flux going

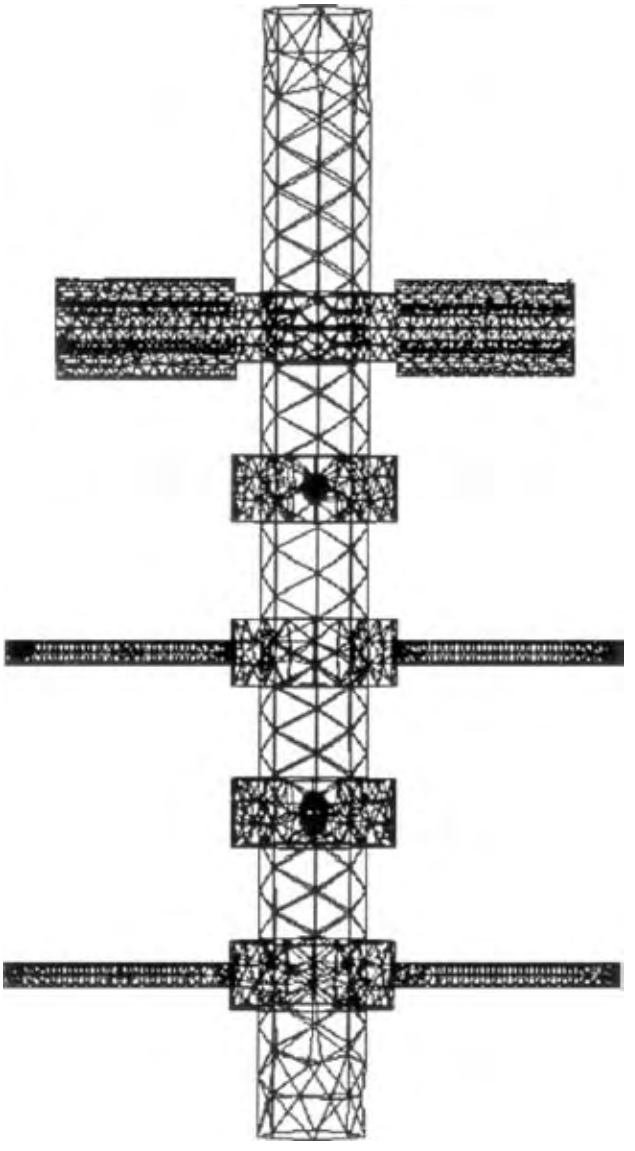


Fig. 7 Surface mesh of the moving impellers.

through a common face to elements i and j, and $\phi_c j^*{}_{,i}$ and $\phi_s j^*{}_{,i}$ the volume fraction of solid particles defined as follows:$\phi_c j^*{}_{,i} = \phi_I$, if $Q_{cj,i} > 0$ is the convective flux of particles leaving element i and entering element $j\phi_c\ j^*{}_{,i} = \phi_j$, if $Q_{cj,i} < 0$ is the convective flux of particles leaving element j and entering element $i\ \phi_s\ j^*{}_{,i} = \phi_i$, if $Q_{sj,i} > 0$ is the sedimentation flux of particles leaving element i and entering element $j\ \phi_s\ j^*{}_{,i} = \phi_j$, if $Q_{sj,i} < 0$ is the sedimentation flux of particles leaving element j and entering element i.

This model assumes perfect mixing inside the element, as the particles enter the cell with different concentrations but leave the cell with a homogeneous concentration. Moreover, inertia is neglected, as well as slipping, between the solid and liquid phases. In Eq. (13), the fluxes can be expressed using the following equations:

$$Q_{cj,i} = \int_{S_{j,i}} \left(v_{cj,i}(x,y,z) \cdot \vec{n}_{j,i}\right) \mathrm{d}S \tag{14}$$

$$Q_{sj,i} = \int_{S_{j,i}} \left(v_{sj,i} \cdot \vec{n}_{j,i}\right) \mathrm{d}S \tag{15}$$

where $\vec{n}$ represents the unit vector of each face pointing to inward element j, S the surface area of the face, v_c the suspension velocity, and v_s the settling velocity determined using the classical sedimentation relations,[26] namely:

$$\mathrm{Ar} = 24\,\mathrm{Re_p} \quad (\mathrm{Ar} < 4,8) \tag{16}$$

$$\mathrm{Ar} = 24\,\mathrm{Re_p} + 3,6\,\mathrm{Re_p}1.687 \quad (4,8 < \mathrm{Ar} < 10^5) \tag{17}$$

$$\mathrm{Ar} = 4/9\,\mathrm{Re_p^2} \quad (\mathrm{Ar} > 10^5) \tag{18}$$

where Ar is the Archimedes number defined as:

$$\mathrm{Ar} = \frac{4}{3} D_p^3 (\rho_s - \rho_l) \frac{\rho_l g}{\mu_l^2} \tag{19}$$

and $\mathrm{Re_p}$ is the particle Reynolds number defined as:

$$\mathrm{Re_p} = \frac{D_p v_l \rho_l}{\mu_l} \tag{20}$$

In order to compute the unsteady term $\mathrm{d}\phi/\mathrm{d}t$, the following expression can be used:

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{a_t \phi^{t+1} + b_t \phi^t + c_t \phi^{t-1}}{\Delta t} \tag{21}$$

where $a_t = 1$, $b_t = -1$ and $c_t = 0$ at the first time step (Euler implicit scheme), and $a_t = 3/2$, $b_t = -2$ and $c_t = 1/2$ for the subsequent steps (second order Gear scheme). After substitution in Eq. (13), the following implicit numerical problem is obtained:

$$V_j \left[\frac{a_t \phi_j^{t+1} + b_t \phi^t + c_t \phi^{t-1}}{\Delta t}\right] = \sum_i Q_{cj,i} \phi_{cj,i}^{*t+1} + Q_{sj,i} \phi_{sj,i}^{*t+1} \tag{22}$$

This problem can then be resolved with the appropriate initial conditions of the mixing problem considered.

Several mixing cases were considered in this work:

1. Simple solid–liquid mixing experiments for the validation of the numerical model
2. Hydrodynamic studies in the coaxial mixer without solid particles
3. Coaxial solid–liquid mixing experiments.

VALIDATION OF THE NUMERICAL MODEL

The propeller setup was used for this purpose. From a computational standpoint, a mesh of the vessel–propeller set was created containing 8746 elements yielding 54,333 velocity equations and 8746 concentration equations. The surface mesh of the propeller (Fig. 5) comprised 964 control points. A maximum of three control points per element was used to avoid locking. Unsteady state flow simulations were performed with a 1-s time step and three coupling iterations between the Navier–Stokes equations and the solid transport equation were required per time step. Steady state was deemed obtained when the solids concentration coefficient of variation did not change.

Fig. 6 illustrates the suspending mechanism versus time. At 350 rpm, all the particles are suspended

Table 3 Characteristics of the surface meshes

Surface mesh	Number of elements	Number of control points
Shaft	400	208
Mixing rods	5022	2279
PBT	1968	884

Table 4 Numerical and experimental values of K_p for an anchor

	K_p (numerical)	K_p (experimental)
This work	256	253
Tanguy, Thibault, and Brito de la Fuente[31]	206	199
Ho and Kwong[32]	—	215
Rieger and Novak[33]	—	206

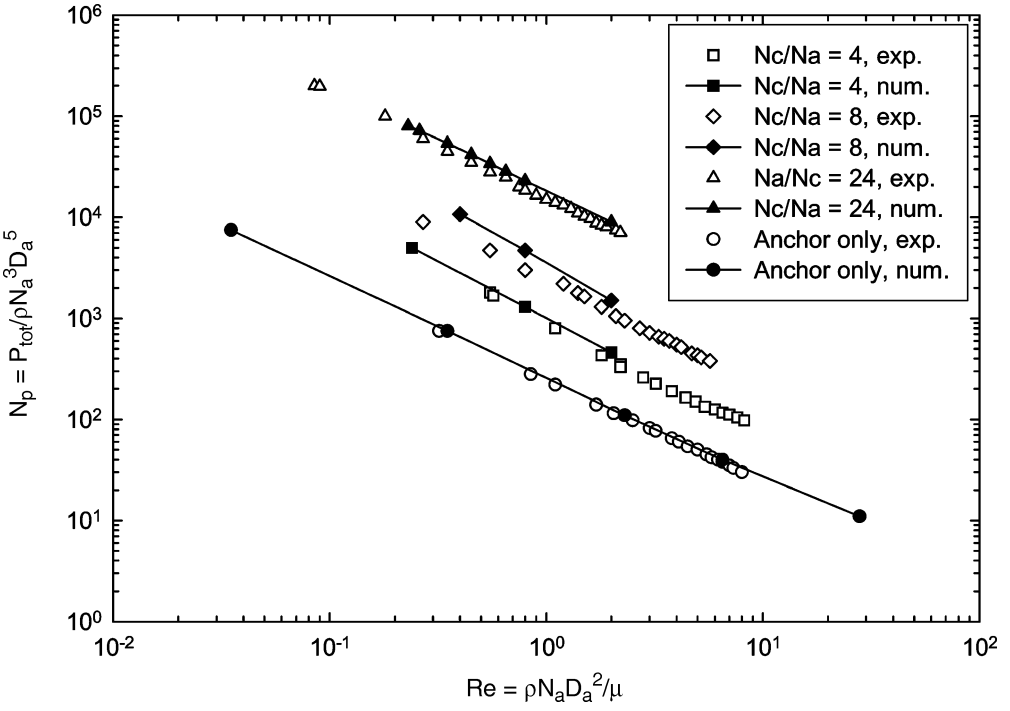


Fig. 8 Numerical and experimental power curve for three different speed ratios.

and uniformly distributed. An attempt has been made to evaluate the minimum suspending speed N_{js} using the work of Armenante and Nagamine.[27] We found that $N_{js} = 855$ rpm for an average particle concentration of 11.9% v/v, which seems to contradict our visual observations. However, as mentioned by Ibrahim and Nienow,[28] published N_{js} correlations largely overestimate the suspending velocity when the suspending medium viscosity is larger than 0.1 Pa.s. Fig. 6 also shows the time evolution of the computed volume fraction until stability is reached. As we do not know the experimental solids concentration distribution in the vessel, the numerical results can only be compared with the experimental results on a qualitative basis. We noted that the computation allows the prediction of the solid accumulation below the agitator. In agreement with the experiments, the network-of-zone model also predicts an increase in the resuspension mechanism when the rotation speed and/or the average concentration in the vessel increase (Table 2).

The accumulation of solids in the vessel bottom at equilibrium was also well captured. For instance at $N = 230$ rpm and a volume fraction of 7.1%, a solid layer with a volume fraction greater than 60% accumulates at the bottom and at the wall, which has been observed experimentally. At $N = 350$ rpm and a volume fraction of 11.9%, almost all the solid particles are suspended, except in a small region close to the bottom edge, again in agreement with the experiments. The performance of the computational model appears therefore satisfactory as fine hydrodynamic details can be predicted for intermediate concentration values. At very low concentration, as the solid layer thickness is reduced, a finer mesh would be required to enhance the model precision.

COAXIAL MIXER RESULTS AND DISCUSSION

Considering the prediction of the power consumption, classically, the power drawn by the impeller is expressed with power curves, i.e., plots of the power number N_p versus the Reynolds number, Re, where:

$$N_p = \frac{P}{\rho N^3 D^5} \tag{23}$$

$$\mathrm{Re} = \frac{\rho N D^2}{\mu} \tag{24}$$

Table 5 Numerical and experimental values of K_p for the coaxial mixer

Speed ratio, R_N	K_p (numerical)	K_p (experimental)
0	256	253
4	1003	817
8	2651	2284
24	17,411	16,486

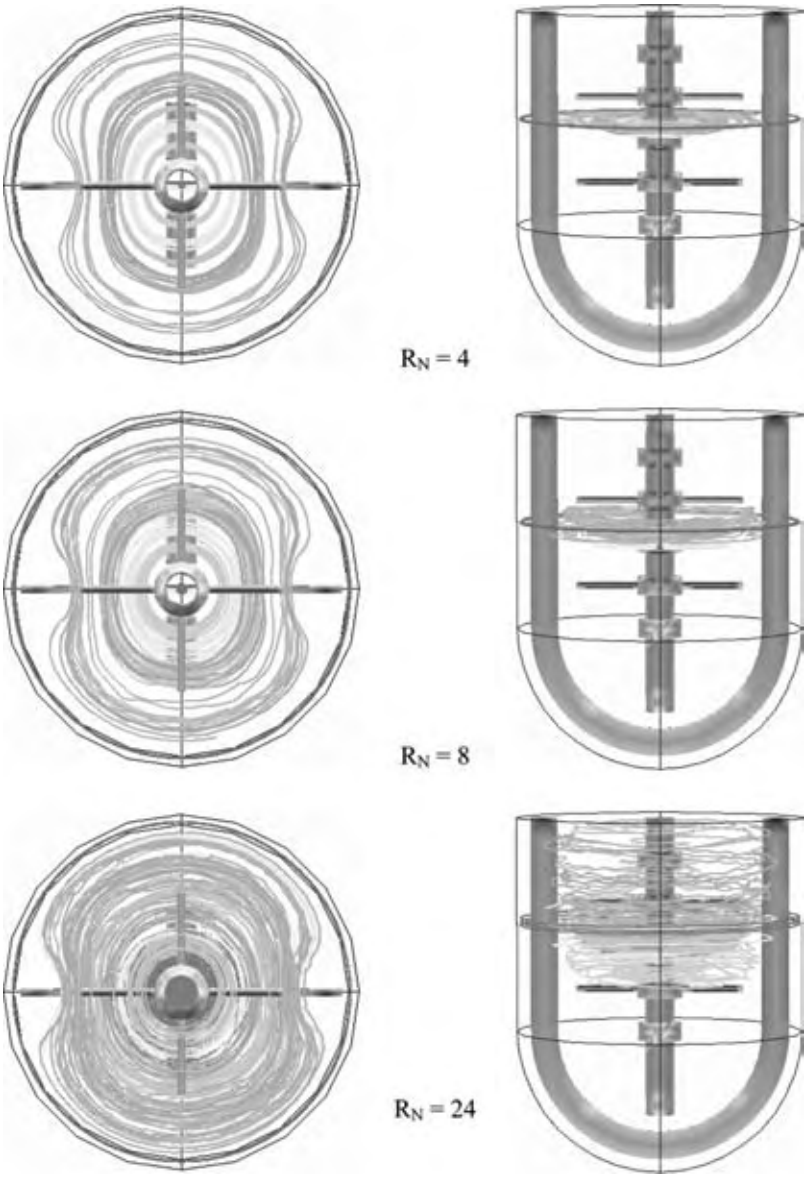


Fig. 9 Effect of speed ratio on the flow field (upper part of the vessel).

P is the mixing power, N and D the rotation speed (in rps) and the impeller diameter respectively, and μ and ρ the fluid viscosity and density. The following relations apply for the laminar and turbulent regimes, respectively:

$$K_p = N_p \quad \mathrm{Re} = \frac{P}{\mu N^2 D^3}, \quad \mathrm{Re} < 10 \tag{25}$$

$$N_p = C \quad \mathrm{Re} > 300 \tag{26}$$

where K_p and C depend only on the mixer geometry for a given fluid.

In a coaxial mixer, the selection of the characteristic speed and dimension that appear in the expression of these dimensionless numbers is ambiguous, as we have two different speeds (N_a and N_c) and three impeller diameters that can be considered. In this work, we used N_a and D_a as the characteristic parameters. The reader is referred to the discussion of this particular choice in Refs.[29,30]. The coaxial mixer power curve and the value of K_p have been obtained by numerical simulation, varying the impeller speed and the speed ratio. For each simulation, the velocity field was used to compute the power by a macroscopic energy balance, namely:

$$P = -\int\int\int_{\Omega} \tau : \nabla v d\Omega \tag{27}$$

Numerical simulations employed the virtual finite element method described above combined with a

Lagrangian flow description (anchor frame of reference). The finite element mesh (generated by IDEAS^{TM} software from SDRC) included 17,083 tetrahedral elements, yielding 106,295 velocities degrees of freedom. Twenty time steps per revolution (angular displacement of 18 degrees per time step) were used, the value of the time step depending on the revolution speed. Typically, the time step was in the range 0.03–0.135 s. For each time step, the control points located on the surface mesh of the moving impellers (Fig. 7 had to be updated. The number of control points required is given in Table 3. The flow simulations carried out for several revolutions showed that the flow was periodic. Moreover, it was found that only one single revolution was enough to obtain a stable, converged solution at low Reynolds number. Finally each simulation required between 12 and 24 CPU hours on an IBM RISC6000/550 server.

Table 4 shows a comparison of the computed and experimental values of K_p for the anchor only. An excellent agreement is obtained. The comparison with literature data shows that the computed values are larger than the data published. The difference is believed to originate from the shape of the vessel bottom (hemispherical in the present work, flat in the literature results).

We show in Fig. 8, a comparison of the numerical and experimental power curves for three different speed ratios ($R_N = N_c/N_a = 4$, 8 and 24). These results have been obtained with $\mu = 15$ Pa.s and $\rho = 1500$ kg/m^3. It can be seen that there is good agreement between the predicted and experimental

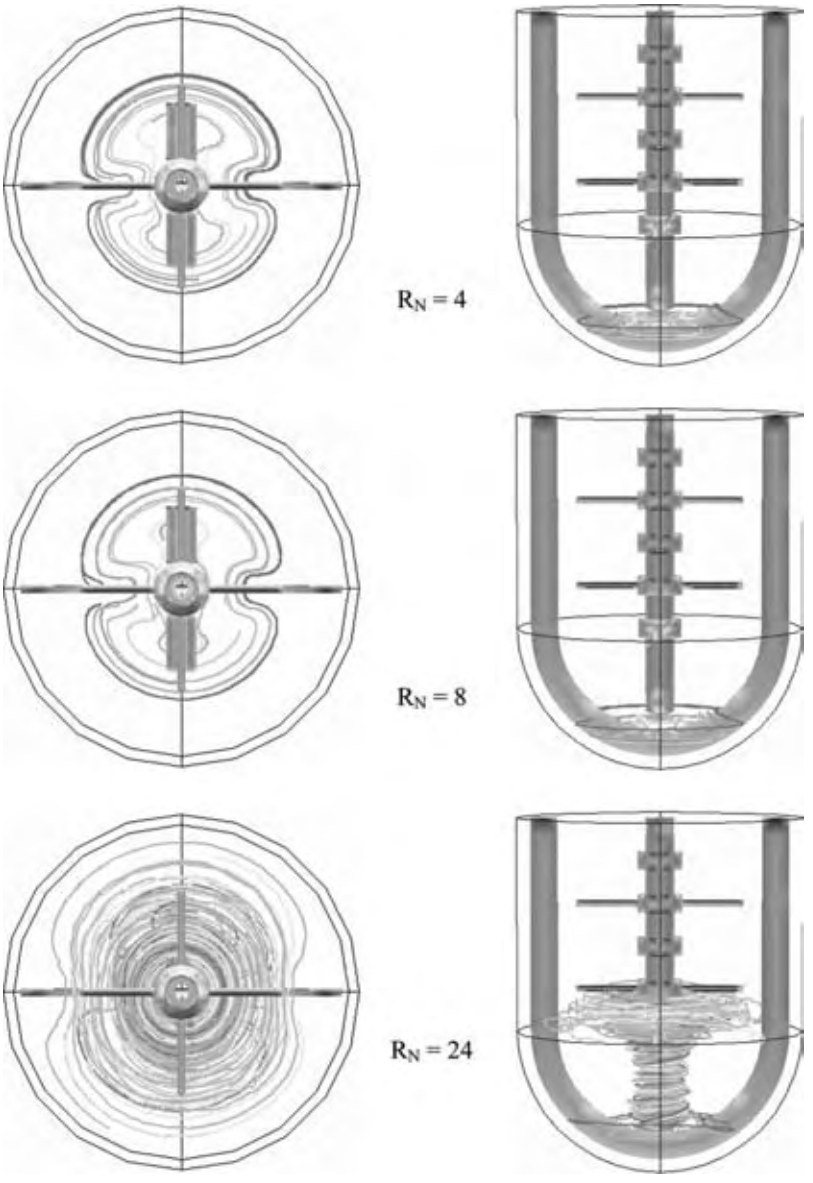


Fig. 10 Effect of speed ratio on the flow field (lower part of the vessel).

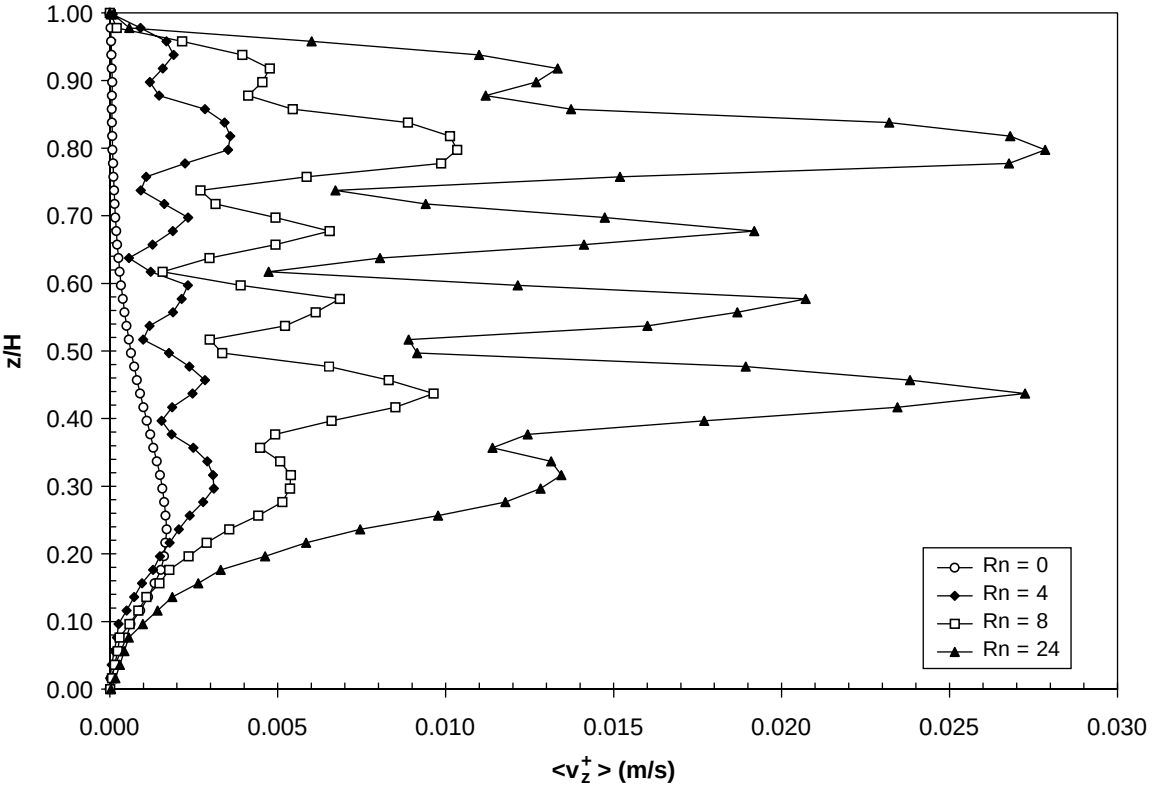


Fig. 11 Axial velocity in the upward direction as a function of the speed ratio.

values. The slope of –1 in the laminar region is captured by the simulation. The onset of the transition regime (decrease of the slope of the *Np* versus Re curve) occurs for a Re value below 10, and the threshold value is sensitive to the value of R_N. For a given Re value, the power increases with R_N. This result seems logical as the Reynolds number has been defined with the anchor parameters and when R_N increases, the central shaft speed rotates faster. This speed increase enhances the average shear rate in the tank, which entrains an increase in the power draw.

The values of the constant K_p of the coaxial mixer versus R_N were also established. Results are shown in Table 5. Here again, the agreement between the predictions and the experimental data is very good.

It is usual in laminar mixing simulations to represent the flow using tracer trajectories. The computation of such flow trajectories in a coaxial mixer is more complex than in traditional stirred tank modelling due to the intrinsic unsteady nature of the problem (evolving topology, flow field known at a discrete number of time steps in a Lagrangian frame of reference). Since the flow solution is periodic, a node-by-node interpolation using a fast Fourier transform of the velocity field has been used, which allowed a time continuous representation of the flow to be obtained. In other words, the velocity at node i was approximated with a Fourier series taking the following form:

$$v^i = a_0^i + a_n^i \cos nt + \sum_{k=1}^{n-1} (a_k^i \cos kt + b_k^i \sin kt) \tag{28}$$

where n is the number of harmonics and coefficients $a_k{}^i$, $k = 1, 2, \ldots, n-1$, and $b_k{}^i$, $k = 1,2, \ldots, n-1$ are obtained from the whole set of velocity values at node i during an impeller revolution. In practice, 10 harmonics were employed according to the Shannon sampling theorem.[34]

We show in Figs. 9 and 10 the effect of the speed ratio on the flow field for an anchor speed of 4.43 rpm. It can be seen that for the two injection points considered (upper part and lower part of the vessel), the radial dispersion increases significantly with the speed ratio. The axial dispersion is also enhanced but less dramatically. A speed ratio of four does not lead to good dispersion and therefore should yield longer mixing times. These results show the synergy between the anchor and the turbine, the axial pumping increasing with the speed ratio. To quantify this axial pumping, the axial velocity in the upward direction is plotted in Fig. 11 versus the speed ratio. An additional advantage of the coaxial mixer can be seen as the axial

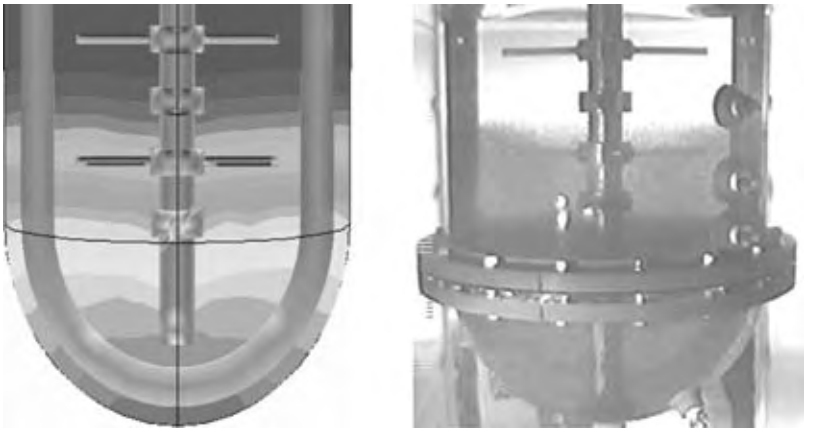


Fig. 12 Predicted and experimental solid volume fraction at equilibrium.

pumping in the upper part of the vessel is enhanced when the speed ratio is augmented.

We now turn to the prediction of the suspension mechanism of the ballotini versus the speed ratio in the coaxial mixer. The average volume concentration is 1%, and the solids are initially at rest in the tank bottom. The first case investigated corresponds to the motion of the sole anchor arm at a speed of 40 rpm. Simulations are carried out in the Lagrangian frame of reference (fixed anchor, rotating vessel). Fig. 12 shows the predicted and experimental solid volume fraction at equilibrium. The computation of the solid–liquid interface at the bottom is fairly well captured by the numerical model as in the previous case dealing with the propeller. The next case considered is the resuspension with the only central shaft in rotation at 160 rpm. This simulation has been carried out with the virtual finite element method described before. Steady state was achieved after 20 revolutions. In Fig. 13, we compare the numerical and experimental distribution of particles in the tank. The agreement is again noteworthy, and the computation predicts that no particle has been resuspended. In fact, although the rotation is counterclockwise, particles have moved in the clockwise direction and accumulated behind the anchor arm. The Mann model is therefore capable of

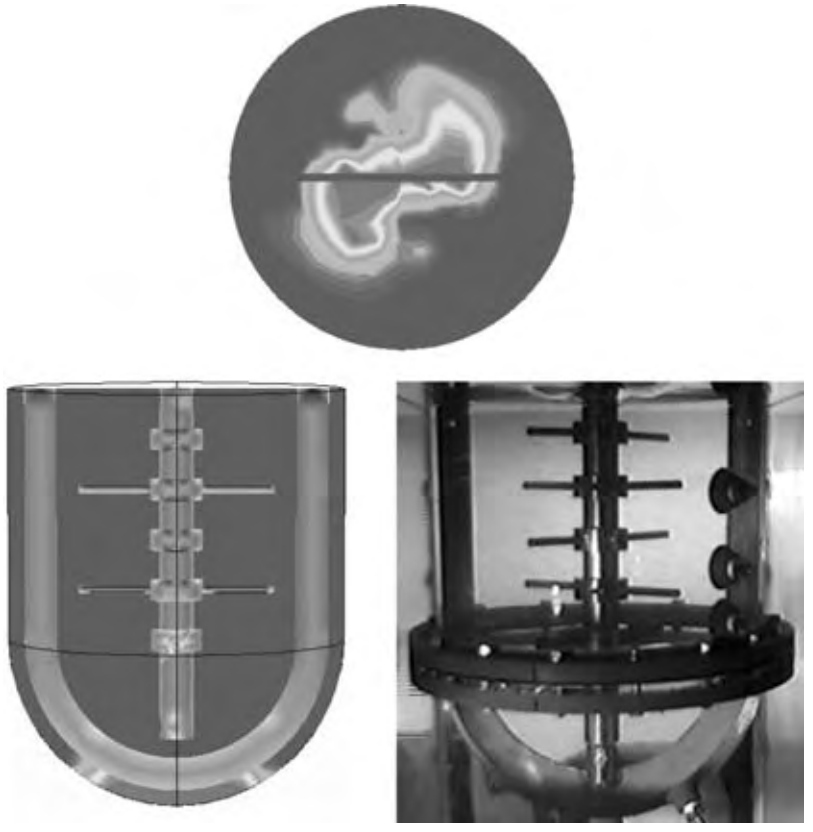


Fig. 13 Numerical and experimental distribution of particles.

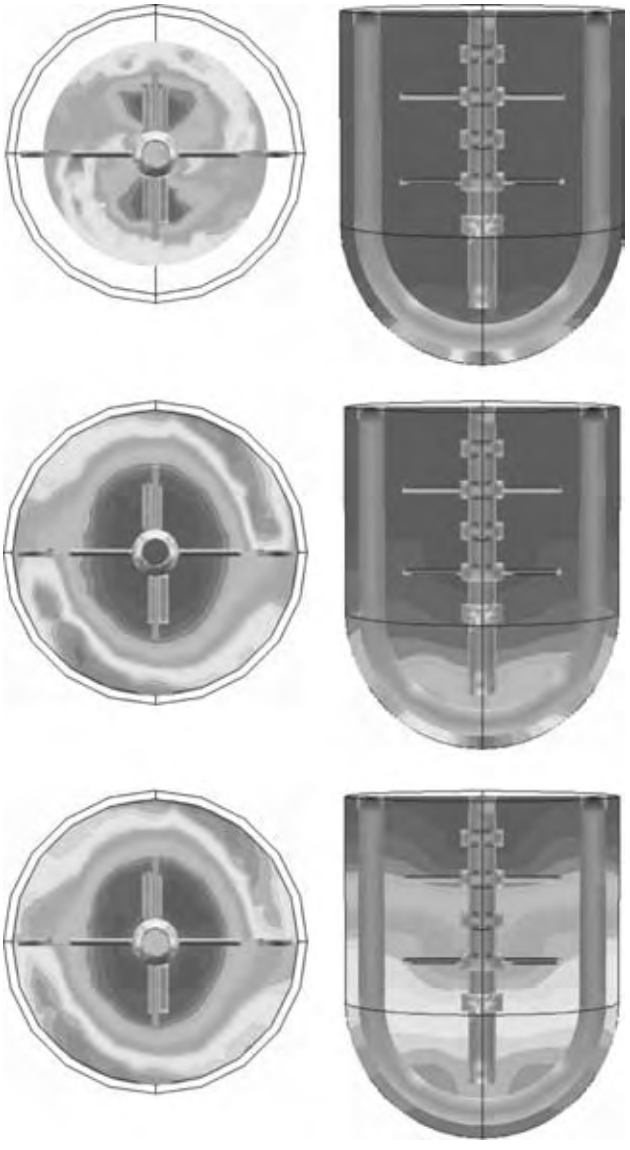


Fig. 14 Resuspending of ballotini a central shaft speed of 160 r.p.m. and a speed ratio of 4.

predicting this odd motion phenomenon at the bottom. Finally, we show in Fig. 14 the resuspending of ballotini for a rotation speed of the central shaft of 160 rpm and a speed ratio R_N of 4 ($N_c/N_a = 4$) after 60 revolutions. This number of revolutions is the maximum that we were capable of computing in a reasonable timeframe (a few days of CPU time). Interesting information can already be obtained from this snapshot result. First, an overconcentration is noticeable in the anchor arm wake in agreement with the experiment. This overconcentration decreases with time as the tank bottom becomes leaner in particle, thereby decreasing the number of particles that can be entrained in the wake. As far as the solid–fluid interface is concerned, the interface rises in the bulk with time but it has not yet reached its eventual position after 60 revolutions.

CONCLUSIONS

Solid–liquid mixing processes can be simulated with good precision when sound CFD methods are used. The application of a combination of the virtual finite element method and the network-of-zone approach was used in this work to analyze the complex flow and suspension mechanisms in a coaxial mixer. Experiments carried on the laboratory scale confirmed the validity of the predictions.

Coaxial mixing shows strong performance capabilities in the case of tough mixing problems involving complex rheology, which should prove more and more useful in industry. Tools are now available to design these systems without resorting to empirical rules.

REFERENCES

1. Magelli, F.; Fajner, D.; Nonentini, M.; Pasquali, G. Solid distribution in vessels stirred with multiple impellers. Chem. Eng. Sci. **1990**, *45*, 615–625.
2. Fajner, D.; Magelli, F.; Nocentini, M.; Pasquali, G. Solids concentration profiles in a mechanically stirred and staged column slurry reactor. Chem. Eng. Res. Des. **1985**, *63*, 235–240.
3. Ferreira, P.J.; Rasteiro, M.G.; Figueiredo, M.M. A new approach to measuring solids concentration in mixing tanks. Adv. Powder Tech. **1994**, *5* (1), 15–24.
4. Rasteiro, M.G.; Figueiredo, M.M.; Freire, C. Modelling slurry mixing tanks. Adv. Powder Tech. **1994**, *5* (1), 1–14.
5. Richardson, J.F.; Zaki, W.N. Sedimentation and fluidisation. Part I. Trans. Instn. Chem. Engrs. **1954**, *32*, 35–53.
6. Bakker, A.; Fasano, J.B.; Myers, K.J. Effects of flow pattern on the solids distribution in a stirred tank. I. Chem. E. Symp. Ser. **1994**, *136*, 1–8.
7. Gadala-Maria, F.; Acrivos, A. Shear-induced structure in a concentrated suspension of solid spheres. J. Rheol. **1980**, *24*, 799–814.
8. Leighton, D.; Acrivos, A. Viscous resuspension. Chem. Eng. Sci. **1986**, *41* (6), 1377–1384.
9. Leighton, D.; Acrivos, A. Measurement of shear-induced self-diffusion in concentrated suspensions of spheres. J. Fluid Mech. **1987**, *177*, 109–131.
10. Leighton, D.; Acrivos, A. The shear-induced migration of particles in concentrated suspensions. J. Fluid Mech. **1987**, *181*, 415–439.
11. Altobelli, S.A.; Givler, R.C.; Fukushima, E. Velocity and concentration measurements of

suspensions by nuclear magnetic resonance imaging. J. Rheol. **1991**, *35* (5), 721–735.
12. Phillips, R.J.; Armstrong, R.C.; Brown, R.A. A constitutive equation for concentrated suspensions that accounts for shear-induced particle migration. Phys. Fluids. **1992**, *A4* (1), 30–40.
13. Krieger, I.M.; Dougherty, T.J. A mechanism for non-Newtonian flow in suspension of rigid spheres. Trans. Soc. Rheol. **1959**, *3*, 137–152.
14. Mann, R. Gas-liquid stirred vessel mixers: toward a unified theory based on network-of-zones. Chem. Eng. Res. Des. **1986**, *64*, 23–34.
15. Mann, R.; Hackett, L.A. *Fundamentals of Gas–Liquid Mixing in a Stirred Vessel: An Analysis using Networks of Backmixed Zones*, Proceedings 6th European Conference on Mixing, Pavia, Italy, 1988; 321–328.
16. Brucato, A.; Rizzuti, L. In *The Application of the Network-of-Zones Model to Solid-Liquid Suspensions*, Proceedings 6th European Conference on Mixing, Pavia, Italy, 1988; 273–280.
17. Brucato, A.; Magelli, F.; Nocentini, M.; Rizzuti, L. An application of the network-of-zones model to solids suspension in multiple impeller mixers. Trans. Instn. Chem. Engrs. **1990**, *69*, Part A, 43–52.
18. Duquesnoy, J.A.; Thibault, F.; Tanguy, P.A. Dispersion of clay suspensions at high solids content. Private communication, 1995, MacMillan Bloedel, British Columbia.
19. Duquesnoy, J.A.; Tanguy, P.A.; Thibault, F.; Leuliet, J.C. A new pigment disperser for high solids paper coating colors. Chem. Eng. Technol. **1997**, *20*, 424–428.
20. Tritton, D.J. *Physical Fluid Dynamics*; Clarendon Press: Oxford, 1988; 325 pp.
21. Pelletier, D.H.; Schetz, J.A. Finite element Navier–Stokes calculation of three-dimensional turbulent flow near a propeller. AIAA **1986**, *24*, 1409–1416.
22. Dermidzic, I.; Peric, M. Finite volume method for the prediction of fluid flow in arbitrarily shaped domains with moving boundaries. Int. J. Numer. Methods. Fluids **1990**, *10*, 771–790.
23. Perng, C.Y.; Murthy, J. A sliding-mesh technique for simulation of flow in mixing tranks. ASME 93-WA-HT-33 **1993**, Old SRN049.
24. Bertrand, F.; Tanguy, P.A.; Thibault, F. A three-dimensional fictitious domain method for incompressible flow problems. Int. J. Num. Meth. Fluids **1997**, *25*, 719–736.
25. Tanguy, P.A.; Bertrand, F.; Labrie, R.; Brito-de la Fuente, E. Numerical modelling of the mixing of viscoplastic slurries in a twin-blade planetary mixer. Chem. Eng. Res. Des. **1997**, *74*, 499–504.
26. Coulson, J.M.; Richardson, J.F. *Chemical Engineering*, 3rd Ed.; Pergamon Press: New York, 1978.
27. Armenante, P.M.; Nagamine, E.U. Effect of low off-bottom impeller clearance on the minimum agitation speed for complete suspension of solids in stirred tanks. Chem. Eng. Sci. **1998**, *53* (9), 1757–1775.
28. Ibrahim, S.B.; Nienow, A.W. The effect of viscosity on mixing pattern and solid suspension in stirred vessels. I. Chem. E. Symp. Ser. **1994**, *136*, 25–32.
29. Tanguy, P.A.; Thibault, F. Power consumption in the turbulent regime for a coaxial mixer. Can. J. Chem. Eng. **2002**, *80*, 601–603.
30. Thibault, F.; Tanguy, P.A. Power draw characterization of coaxial mixer with Newtonian and non-Newtonian fluids. Chem. Eng. Sci. **2002**, *57*, 3861–3872.
31. Tanguy, P.A.; Thibault, F.; Brito de la Fuente, E. A new investigation of the Metzner–Otto concept for anchor mixing impellers. Can. J. Chem. Eng. **1996**, *74*, 222–228.
32. Ho, F.; Kwong, A. A guide to designing special agitators. Chem. Eng. **1973**, July issue, 94–104.
33. Rieger, F.; Novak, V. Power consumption of agitators in highly viscous non-Newtonian liquids. Trans. Instn. Chem. Engrs. **1973**, *51*, 105–111.
34. Ljung, L. *System Identification Theo for the User*; Prentice-Hall: Englewood, Cliffs, NJ, 1987.

Solid–Liquid Separation

Frank M. Tiller
Wenping Li
Department of Chemical Engineering, University of Houston, Houston, Texas, U.S.A.

INTRODUCTION

Solid–liquid separation (SLS) involves operation of solid and liquid systems with the objectives of:[1]

1. Recovering solids (the liquid being discarded).
2. Recovering liquid (the solids being discarded).
3. Recovering both solids and liquid.
4. Removing pollutants, solutes, micro-organisms, etc.

Solid–liquid separation is encountered in all stages of manufacturing processes ranging from raw material purification through product separation to waste management. Solid–liquid separation operations include screening, cake and deep-bed filtration, gravitational sedimentation, sedimenting and filtering centrifugation, expression (cake squeezing), hydro-cycloning, washing, membrane separation, flotation, etc. Selection of proper equipment and optimum operating conditions are among the most important problems faced by engineers involved in SLS. Because of the complex nature of fluid/particle systems and a general lack of fundamental training, SLS operations are often a problem area, or bottle neck in a plant. Various SLS operations are described as follows.[2]

SEDIMENTATION

Separation involving sedimentation is dependent upon settling velocity, which requires a difference in density between solid particles and the suspending liquid. Gravitational sedimentation operations are divided into clarification and thickening. Clarification involves dilute suspensions and frequently has the objective of liquid recovery. Thickening refers to solid recovery by forming more concentrated slurries. Particle size, liquid and particle densities, and liquid viscosity are important factors in sedimentation processes.

CAKE FILTRATION

In cake filtration (Fig. 1A), solids and liquid are separated by filter medium, which retains the solids as a cake and permits the liquid to pass through under pressure, vacuum or centrifugal forces. Once the cake is formed, it becomes the primary filter medium and particles finer than the openings of the medium can be separated.

CROSS-FLOW FILTRATION

In cross-flow filtration (Fig. 1B), shear forces are introduced at the cake surface to reduce cake thickness and total cake resistance. It is exclusively used in membrane separation applications to prevent fouling on membranes.

SCREENING

Screening (Fig. 1C) is a type of surface filtration to separate solids and liquid by screens with openings smaller than particle size. It is usually employed among the first few steps of a SLS process. Vibrations and other mechanism are often applied to avoid blinding of the screens.

DEEP-BED FILTRATION

In deep-bed filtration (Fig. 1D), particles are caught inside the filter medium. Examples of deep-bed filters are granular beds and some cartridge filters. Deep-bed filtration is used for dilute suspensions (<100 ppm) containing fine particles that are not easy to be removed by sedimentation or cake filtration.

CENTRIFUGATION

Centrifugation increases the separating driving force by developing centrifugal forces on particles. A difference in density between the liquid and the particles is also a prerequisite. Centrifuges are employed in filtering and sedimenting modes. The terms ''perforated bowl'' and ''solid bowl'' differentiate the two types of operations.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007766

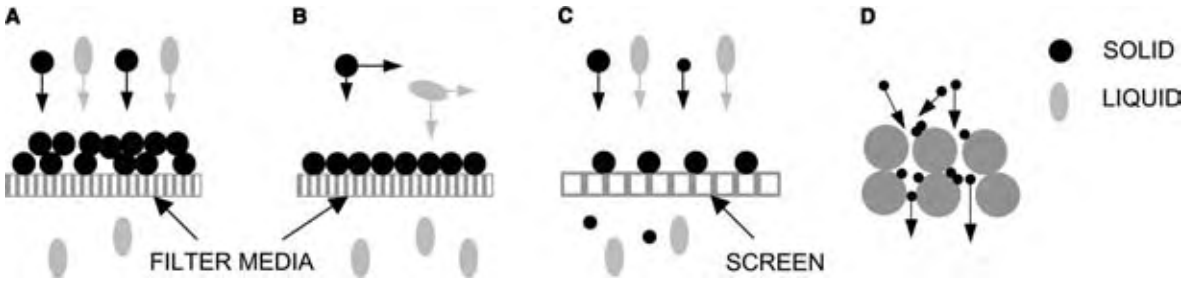


Fig. 1 Schematic mechanisms of four types of filtration: (A) cake filtration; (B) cross-flow filtration; (C) screening; (D) deep-bed filtration. *(View this art in color at www.dekker.com.)*

CAKE WASHING

Cake washing is a common operation for removal of soluble materials in filter cakes. There are displacement washing and dilution washing. A clean liquid is used in this step. Direct introduction of wash liquid behind the slurry is called displacement washing. As a rough guide, about half of the liquor in the cake is displaced at the full initial concentration. Breakthrough occurs at that point, and the concentration of the wash rapidly drops off as solutes gradually diffuse out of inaccessible pores. Dilution washing is accomplished by reslurrying the cake and refiltering the diluted slurry. Cocurrent or countercurrent washing can be used in dilution washing. Ideally, displacement washing would be used until the wash liquid reached a predetermined concentration, then the cake would be reslurried, and displacement washing would follow.

EXPRESSION

Expression operations include mechanic squeezing of a wet filter cake with stationary diaphragms, moving belts, rollers, or screws. For further deliquoring, blowing with gas is frequently used after expression.

MEMBRANES

Membranes are used for a wide variety of separations. A membrane serves as a barrier to some particles while allowing others to selectively pass through. The membrane pore size, shape, and electrostatic surface charge are fundamental to particle removal. Reverse osmosis (RO), nanofiltration (NF), ultrafiltration (UF), and microfiltration (MF) relate to separation of ions, macromolecules and particles in the 0.001–10 μm range.[3]

HYDROCYCLONES

Hydrocyclones are closely related to centrifuges in that centrifugal forces are employed for liquid and particles separation by introducing the slurry radially into the upper periphery of the cyclone at high velocity. solids are thrown out to the wall, and flow down the inclined walls, and exit at the bottom. In general, hydrocyclones operate as classifiers with large particles in the underflow and small particles in the overflow.

FLOTATION

Flotation is commonly employed for beneficiation of minerals. Frothers are added to the vigorously agitated mixture. Air bubbles attach to the collector-mineral particles, which then rise and are removed along with the froth or foam. The valuable components are separated from the waste or gangue by preferential floating of one of the components to the top while the other sinks to the bottom.

PARTICLE SIZE AND CLASSIFICATION OF SLS OPERATIONS

Sizes of particle are important in selecting the method of separation. Different authors do not necessarily use the same nomenclature in classifying separations in terms of ''size.'' A crude relationship of particle size to particle size measurements and methods of SLS and particle size analysis is provided in Fig. 2.[2]

As illustrated in Fig. 2, there is a weak relationship between particle size and the methods for separation. Particle size measurements are usually made on dilute dispersed systems, where each particle is separated from all other particles. However, in practice, particles are usually combined into aggregates with sizes much larger than the individual units. Solid–liquid separation techniques depend more on the size of the aggregates than the size of the dispersed particles. In addition to size and the degree of aggregation, other properties, such as shape, density, internal porosity and surface charge, and suspension concentration are also significant factors in choosing fluid/particle

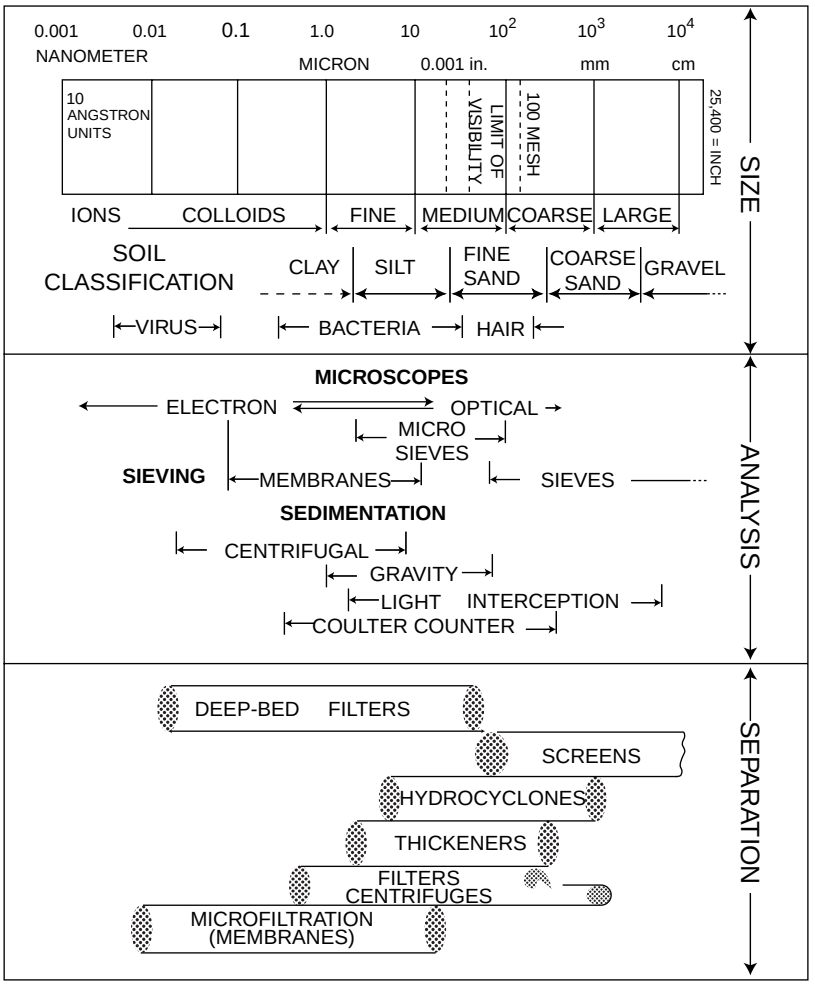


Fig. 2 Relationship of particle size and separation techniques.

separation systems. Characteristics of suspensions or slurries are fundamental to both natural and man-controlled processes.

FUNDAMENTAL OF SLS—FLOW THROUGH POROUS MEDIA

Mechanism of flow through porous media is fundamental in theoretical study of SLS process, such as filtration, thickening, centrifugation, expression, washing, etc. In the early study, with the development of fluid mechanics, interest was focused on flow in capillaries through incompressible sand beds. The beginning of present day theory can be traced to Hagen (1839), Poiseuille (1840), and Darcy (1856).[4]

Darcy's Law

Based on fluid mechanics, extensive theoretical and experimental studies by Poiseuille and Hagen led to a formula for viscous flow through capillaries, which can be written as

$$\frac{dp_L}{dx} = 32\mu\frac{u_L}{D_p^2} \tag{1}$$

where dp_L/dx is the pressure gradient, μ the viscosity, and u_L the average velocity in a tube of diameter D_p. For general viscous or turbulent flow in a circular tube

$$\frac{dp_L}{dx} = f\frac{\rho_L u_L^2}{2D_p} \tag{2}$$

where f is the friction factor of flow through the tube.

Darcy[4] carried on a series of experiments involving the flow of water through sand placed in a vertical iron pipe and found the flow rate to be proportional to the head of pressure, which is called Darcy's law as expressed by the following equation

$$\frac{dp_L}{dx} = \frac{\mu q_L}{K} = \frac{\mu u_L \varepsilon_L}{K} \tag{3}$$

where K is the permeability, m^2, and q_L is the superficial flow rate (volume of fluid per unit area of sand bed per unit time), $m^3/(m^2\,s)$. ε_L is the average porosity of

Table 1 Permeability of various materials

	m^2	$ft.^2$	darcies	m-darcies
Sand bed	E(10)	E(9)	100	100,000
Filter aids	E(13)	E(12)	E(1)	1.0
Clay	E(15)	E(14)	E(3)	0.01

sand bed, and the average pore velocity u_L is related to q and ε_L by

$$q_L = \varepsilon_L u_L \tag{4}$$

Petroleum and geotechnical engineers frequently use a unit of permeability called the *darcy*. One *darcy* approximately equals $10^{-12}\,m^2$. The order of magnitude of the permeability is provided in Table 1.

Basic Filtration Theory

Ruth modified Darcy's law for filtration in the form

$$\frac{dp_L}{dw} = \mu\alpha_w q_L \tag{5}$$

where w is the mass of dry solid deposited per unit area, kg/m^2, and α_w is the specific flow or filtration resistance in the mass basis, m/kg. The mass of dry solids is related to thickness by

$$dw = \rho_s(1 - \varepsilon_L)dx \tag{6}$$

Ruth's equation in mass basis required knowledge of density of solids. To get a better view of cake solids, and cake porosity, volume of cake solids instead of mass is used in Tiller's theory[2,5–7]

$$\frac{dp_L}{d\omega} = \mu\alpha q_L \tag{7}$$

in which ω is the volume of cake solids m^3/m^2, and α is the specific resistance of cake in volume basis, $1/m^2$. Eq. (6) is simplified as

$$d\omega = (1 - \varepsilon_L)dx \tag{8}$$

The volume fraction of solids (solidosity) ε_s is given by

$$\varepsilon_s = 1 - \varepsilon_L = d\omega/dx \tag{9}$$

In Eqs. (3), (5), or (7), the permeability K or specific resistance α_w or α are the key parameters to solve problem of flow through porous media. Many investigators have sought theoretical expressions from flow through incompressible beds by transforming parameters in pipe flow [Eqs. (1) and (2)] to flow through compactible porous media.

Flow Through Compactible Porous Media

Fig. 3 shows a simplified compactible filter cake. Darcy's law and a stress balance involving the accumulated friction drag F_s on the particulate structure (Fig. 3) are used to develop basic theory of flow through compactible porous media.

Assuming the particles are in point contact and the liquid pressure is effective over the entire cross-sectional area A, the force of the liquid at x will be $F_{Lx} = p_L A$. In the absence of measurable momentum change, the total force at the cake surface pA, because of the pump pressure p must be balanced by the force of the liquid, F_{Lx}, and the accumulated frictional stress F_s on the network of solid particles in the cake between x and L, then

$$pA = F_s + p_L A \tag{10}$$

Dividing by A and defining the effective stress as $p_s = F_s/A$ yields

$$p = p_L + p_s \tag{11}$$

where p is constant or a function of time and p_L and p_s are functions of x and t. At constant t,

$$dp_L + dp_s = 0 \tag{12}$$

Eqs. (11) and (12) must be modified if gravitational and centrifugal body forces are involved.

In Darcy's equations [Eqs. (3) and (7)], the permeability K and the specific resistance α are functions of the effective stress p_s in the particulate structure. Replacing dp_L by $-dp_s$ in accord with Eq. (12), and involving both spatial coordinate x and material

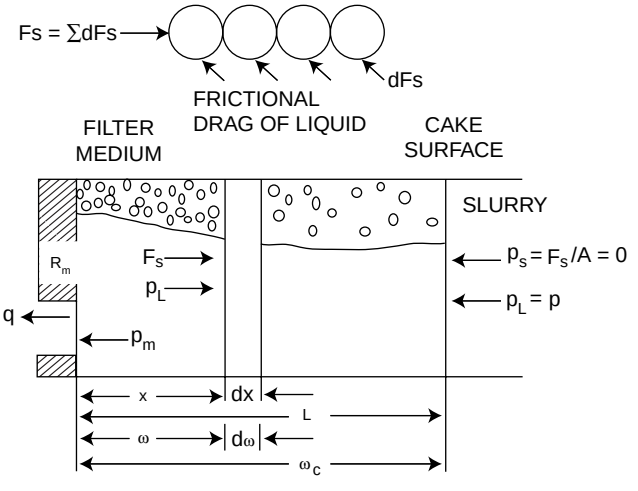


Fig. 3 Diagram of a cake with spatial, x, and material, ω, coordinates.

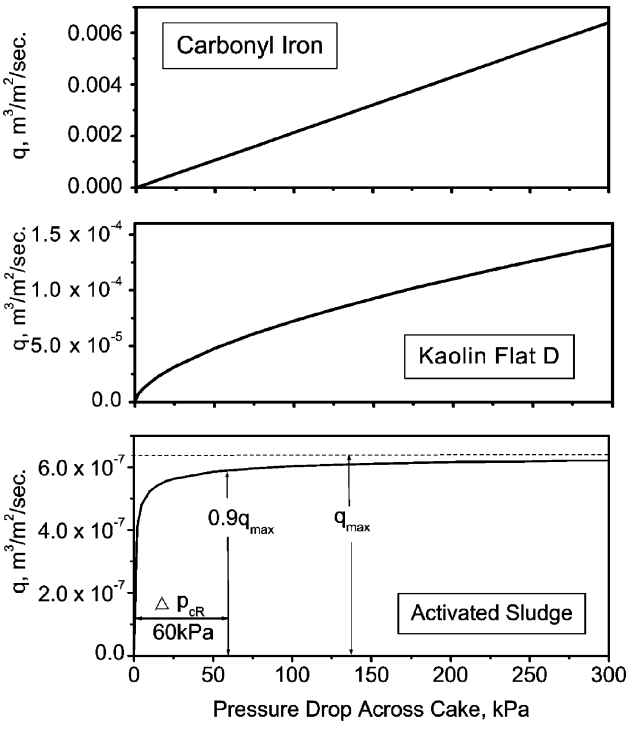


Fig. 4 Variation of q against Δp_c.

coordinate ω (Fig. 4) yields the Darcy's equation as:

$$q = \frac{K}{\mu}\frac{\mathrm{d}p_L}{\mathrm{d}x} = \frac{1}{\mu\alpha}\frac{\mathrm{d}p_L}{\mathrm{d}\omega} = -\frac{K}{\mu}\frac{\mathrm{d}p_s}{\mathrm{d}x} = -\frac{1}{\mu\alpha}\frac{\mathrm{d}p_s}{\mathrm{d}\omega} \tag{13}$$

Empirical constitutive equations relating the local permeability K, specific resistance, and solidosity to the effective stress p_s are

$$(\varepsilon_s/\varepsilon_{so})^{1/\beta} = (\alpha/\alpha_0)^{1/n} = (K/K_0)^{-1/\delta} = 1 + p_s/p_a \tag{14}$$

where α_0, K_0, ε_{so} are null stress values, and p_a is an empirical parameter. We have the relations

$$K\alpha\varepsilon = 1 \quad \text{and} \quad K_0\alpha_0\varepsilon_{so} = 1 \tag{15}$$

Combining the spatial Darcy and constitutive equations and integrating across the cake with limits of $(0, L)$ on x and $(0, \Delta p_c)$ on p_s yields[8]

$$\int_0^L \mu q \mathrm{d}x = \mu q L = \int_0^{\Delta p_c} K_0(1 + p_s/p_a)^{-\delta}\mathrm{d}p_s = \frac{K_0 P_a}{1-\delta}\left[\left(1 + \frac{\Delta p_c}{p_a}\right)^{1-\delta} - 1\right] \tag{16}$$

Repeating the process with the material coordinate form of the Darcy equation leads to

$$\mu q \omega_c = \int_0^{\Delta p_c} \frac{\mathrm{d}p_s}{\alpha_0(1 + p_s/p_a)^n} = \frac{P_a}{\alpha_0(1-n)} \times \left[\left(1 + \frac{\Delta p_c}{p_a}\right)^{1-n} - 1\right] \tag{17}$$

Dividing Eq. (17) by Eq. (16) produces an equation for the average cake solidosity:

$$\frac{\omega_c}{L} = \varepsilon_{sav} = \varepsilon_{so}\left(\frac{1-\delta}{1-n}\right) \times \frac{(1 + \Delta p_c/p_a)^{1-n} - 1}{(1 + \Delta p_c/p_a)^{1-\delta} - 1} = \varepsilon_{so}\left(\frac{\delta - 1}{n - 1}\right)\frac{1 - 1/(1 + \Delta p_c/p_a)^{n-1}}{1 - 1/(1 + \Delta p_c/p_a)^{\delta-1}} \tag{18}$$

where ω_c is the volume of solids/unit area in the entire cake, ε_{sav} the average solidosity in cake, and $1/\alpha_0 K_0$ has been replaced by ε_{so} in accord with Eq. (15). When medium resistance is negligible, the pressure drop across the medium can be neglected; and the pressure drop across the cake Δp_c can be replaced by the pump pressure p.

Table 2 Empirical constitutive parameters for five different materials

Materials	N	δ	ε_{s0}	K_0 (m^2)	p_a (Pa)
Carbonyl iron, grade E[15]	0	0	0.6	1.03E-13	—
Kaolin flat D[12]	0.4	0.52	0.14	2.4E-13	11
Mierlo biosolids[13]	1.83	2.3	0.03	8.3E-12	1000
Water treatment residue[14]	1.03	1.25	0.035	3.6E-14	170
Activated sludge[14]	1.4	1.66	0.05	5.53E-14	190

Cake Compactibility[8–11]

Classification of cake compactibility is based on values of n and δ as shown in the following table:

Incompressible	$n = 0$	$\delta = 0$
Moderately compactible	$n \approx 0.5$–0.6	$\delta \approx 0.6$–0.8
Highly compactible	$n \approx 0.7$–0.8	$\delta \approx 0.9$–1.0
Super-compactible	$n > 1$	$\delta > 1.0$

The empirical constitutive parameters for five materials with different compactibilities are shown in Table 2. The 9-μm spheres with $n = \delta = 0$, and $K = K_0 =$ constant correspond to an incompressible material. The Kaolin flat D[12] is moderately compactible and the Mierlo biosolids,[13] water treatment residue,[14] activated sludge (AS)[14] with n and δ exceeding unity are classified as super-compactible materials.

For super-compactible cakes, when $\delta > 1.0$ and $n > 1.0$, Eqs. (16)–(18) can be written better, respectively, as

$$q = \left[\frac{K_0 P_a}{\mu L(\delta - 1)}\right]\left[1 - 1/(1 + \Delta p_c/p_a)^{\delta-1}\right] \quad (19)$$

$$q\frac{P_a}{\alpha_0(1 - n)}\left[1 - 1/(1 + \Delta p_c/p_a)^{n-1}\right] \quad (20)$$

$$\varepsilon_{sav} = \varepsilon_{so}\left(\frac{\delta - 1}{n - 1}\right)\frac{1 - 1/(1 + \Delta p_c/p_a)^{n-1}}{1 - 1/(1 + \Delta p_c/p_a)^{\delta-1}} \quad (21)$$

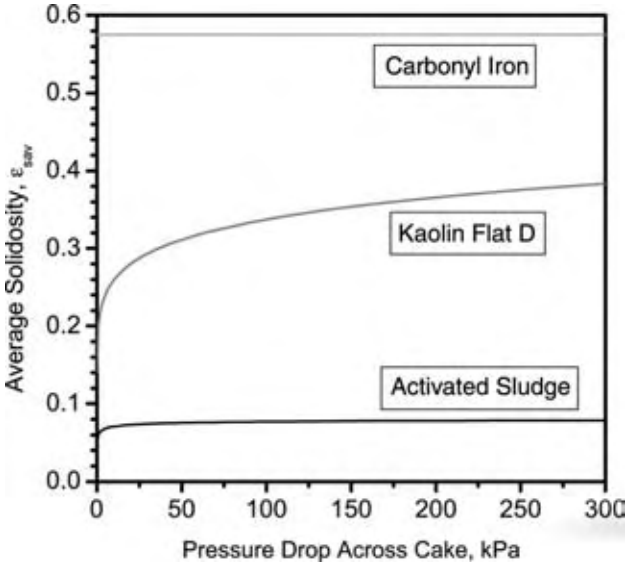


Fig. 5 Variation of ε_{sav} against Δp_c. *(View this art in color at www.dekker.com.)*

As Δp_c increases indefinitely in Eqs. (19)–(21), the term with Δp_c approaches zero, and the flow rate and the average solidosity reach constant. Plots of calculated values of flow rate and average solidosity as functions of pressure drop across cakes with $\omega_c = 0.02\,m^3/m^2$ for carbonyl iron, Kaolin flat D, and AS (in Table 2) are shown in Figs. 4 and 5. Whereas the flow rate q increases linearly with Δp_c for incompressible carbonyl iron, the flow rate increases with a power of Δp_c for compactible kaolin. For super-compactible materials with $n > 1$, and $\delta > 1$, increasing the pressure drop beyond some low value has negligible effect on the flow rate and average solidosity. The behavior of super-compactible materials that has little or no effect of pressure on either the flow rate or the average solidosity, when pressure is beyond some critical value, is unique and different compared to incompressible and moderately compactible materials. Special attention should be paid for filtration design and operation of this type of materials.

SLS SYSTEM[2,5,16]

Four Stages in SLS System

Solid–liquid separation systems generally consist of four stages, which are: 1) pretreatment to increase particle size; 2) solid concentration in thickeners; 3) solid separation in filters and centrifuges; and 4) post-treatment to remove solubles and reduce liquid content. Fig. 6 shows the relationship among these stages.

Solid–liquid suspension pretreatments are employed to improve performance of the following SLS processes. Coagulation and flocculation pretreatment for colloidal systems aim at increasing the effective particle size and particle settling velocity in sedimentation operations. Addition of $FeCl_3$, $Al_2(SO_4)_3$, acidifying, cationic polyelectrolytes, and other methods are used to produce flocculation of suspensions. Aging to allow slow reactions to occur, freezing, and addition of filter aids are representative of other pretreatment methods. It is relatively difficult to determine the best pretreatment practice. Large flocs generally mean a higher chemical cost and a lower capital investment.

In the second stage, the suspension is concentrated in a thickener. Although gravity thickeners dominate the field, cross-flow filter thickeners, hydrocyclones, and electrophoretic devices can be also used. A large fraction of the liquid can be removed economically in thickeners, thereby leading to smaller units in the third stage.

In the third stage, cakes are produced in diverse types of filters and centrifuges. Wet cakes and filtrate or centrate are the products of filters and centrifuges.

The fourth (post-treatment) stage involves further processing of both the filtrate and the wet solids from

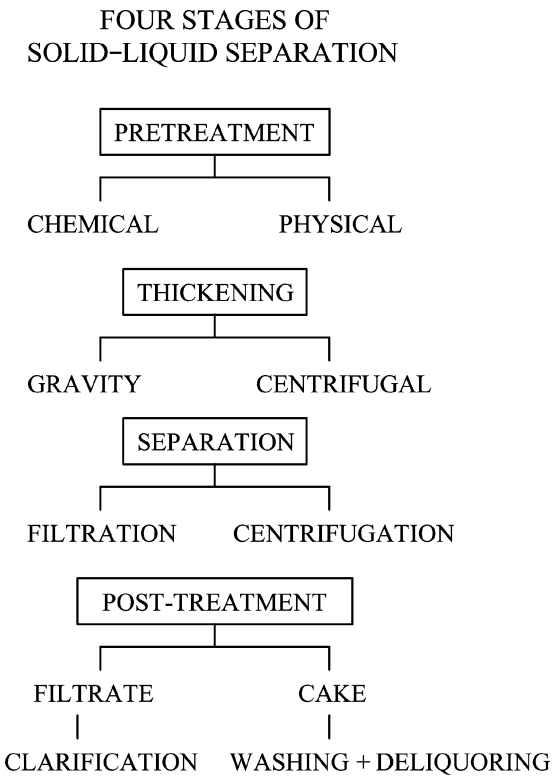


Fig. 6 Stages of solid–liquid separation.

stage 3. A small fraction of colloidal particles which migrate through the pores of both the cake and the medium appear as turbidity in the filtrate or centrate, and require some sort of deep-bed operation for further clarification in some applications. Specifications on the product liquid determine the need for additional processing.

For the cake, if the mother liquor in the slurry contains soluble substances, it may be necessary to wash the cake. Deliquoring is a process after cake washing. Decreasing the liquid content of cakes has become a major objective of many SLS processes. Although the term *dewatering* is widely used, *deliquoring* is more general and covers both aqueous and nonaqueous liquids. Deliquoring operations consist of expression (squeezing), blowing, sucking with a vacuum, and gravitational and centrifugal drainage.

Fractional Liquid (Solid) Recovery of a SLS System

In any separation process, three streams are involved as shown in Fig. 7 for a filter with cake formation, a cross flow filter without cake formation, and a thickener or clarifier based upon gravitational sedimentation. The filter is assumed to have a fixed volume into which slurry is fed and cake is formed as filtrate flows through the filter medium. The fraction of filtrate or overflow, compared to the liquid in the slurry, is called the fractional recovery of liquid, which can be obtained based upon material balance. The fractional recovery for cake filter, cross-flow filter, and thicker or clarifier are as follows:

$$\text{Fractional recovery} = \frac{V}{V_F(1 - \varphi_s)} = \left\{ \begin{array}{l} (1 - \varphi_s/\varepsilon_{sav})/(1 - \varphi_s)\ \text{(Filter)} \\ (1 - \varphi_s/\varphi_{sc})/(1 - \varphi_s)\ \text{(Cross-flow)} \\ (1 - \varphi_s/\varepsilon_{su})/(1 - \varphi_s)\ \text{(Thickener, clarifier)} \end{array} \right\} \tag{22}$$

The fractional removal, which is defined as the fraction of liquid removed in the cake (sediment) or concentrate then becomes

$$\text{Fractional removal} = \frac{\left\{ \begin{array}{l} V_c(1 - \varepsilon_{sav})\text{(Filter)} \\ V_c(1 - \varphi_c)\text{(Cross-flow)} \\ V_c(1 - \varepsilon_{su})\text{(Thickener, clarifier)} \end{array} \right\}}{V_F(1 - \varphi_s)} \tag{23}$$

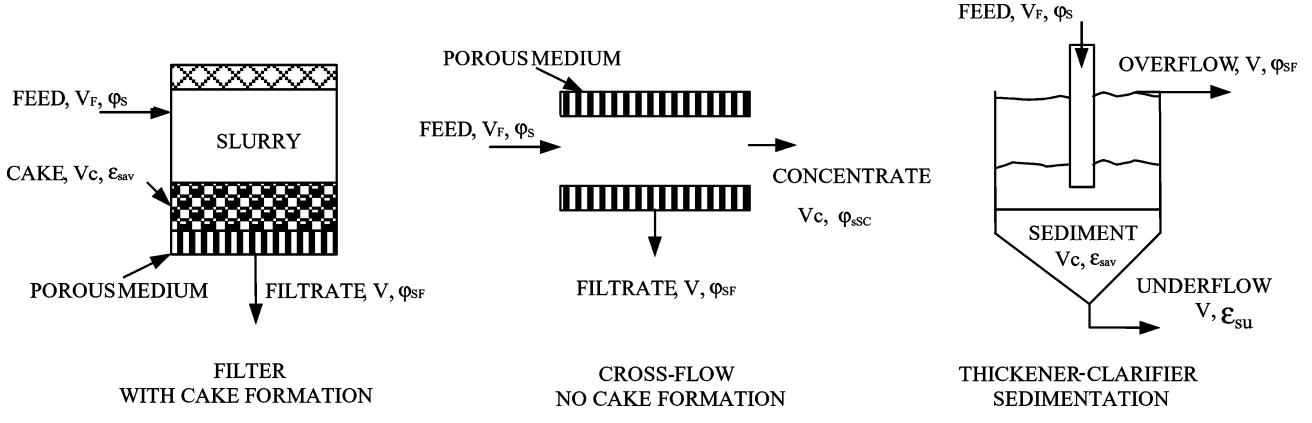


Fig. 7 Three streams involved in SLS systems.

Table 3 Fractional removal of liquid

Example	Stage	Vol.% solids	Void ratio (volume of liquid/ volume of solid)	Liquid removed (volume of liquid/ volume of solid)	Liquid removed (%)	Total liquid removed (%)
Example 1, feed vol.% solids = 1%	Feed	1	99	—	—	—
	Thickener	10	9	90	90.9	90.9
	Filter, centrifuge	25	3	6	6.1	97
	Deliquoring	50	1	2	2.0	99
Example 2, feed vol.% solids = 5%	Feed	5	19	—	—	—
	Thickener	10	9	10	52.6	52.6
	Filter, centrifuge	25	3	6	31.6	84.2
	Deliquoring	50	1	2	10.5	94.7

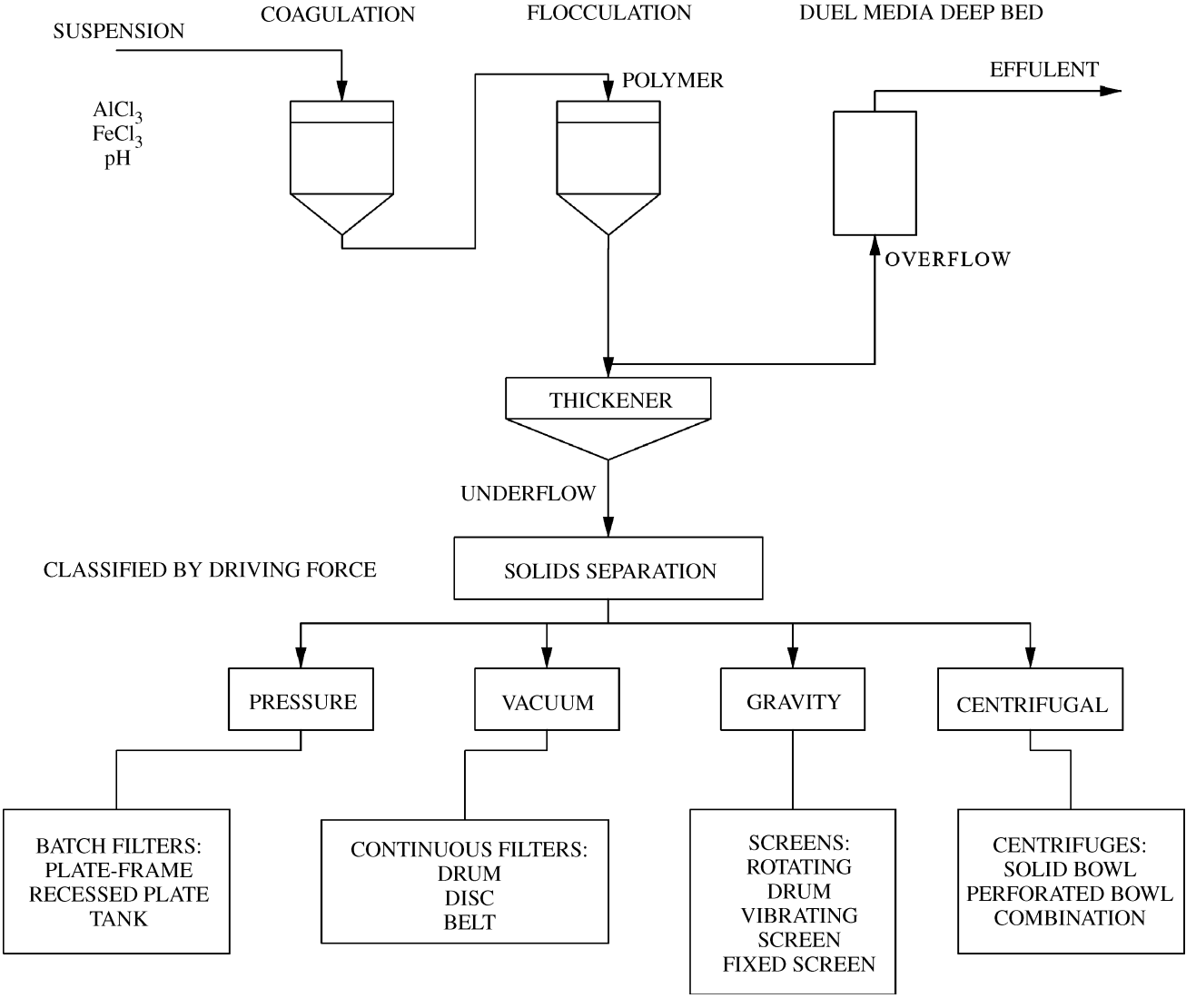


Fig. 8 Equipment used for the four stages of solid–liquid separation.

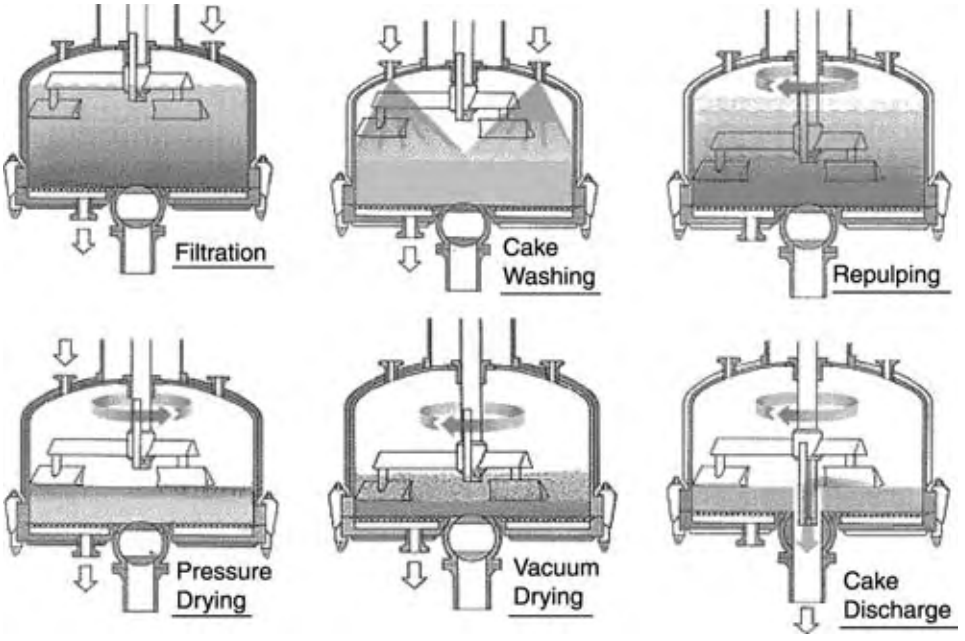


Fig. 9 Operations of nutsche filters. (Courtesy of the Rosenmund Guedu Group.) *(View this art in color at www.dekker.com.)*

Eqs. (22) and (23) are equivalent as "fractional recovery = 1 − fractional removal."

Liquid Recovery of Different Stages of SLS

It is instructive to consider the fraction of liquid removed in the second through the fourth stages shown in Fig. 6. Two examples with feeds containing 1.0 and 5.0% solids (volume), and going through sedimentation, filtration, and deliquoring will be considered. It is assumed that 1) the thickener produces a 10% by volume slurry; 2) a cake containing 25% by volume; and 3) the deliquored cake contains 50% by volume after expression, blowing with gas, or spinning in a centrifuge. To facilitate comparisons, the concentrations have been converted to void ratios. The results are shown in Table 3.

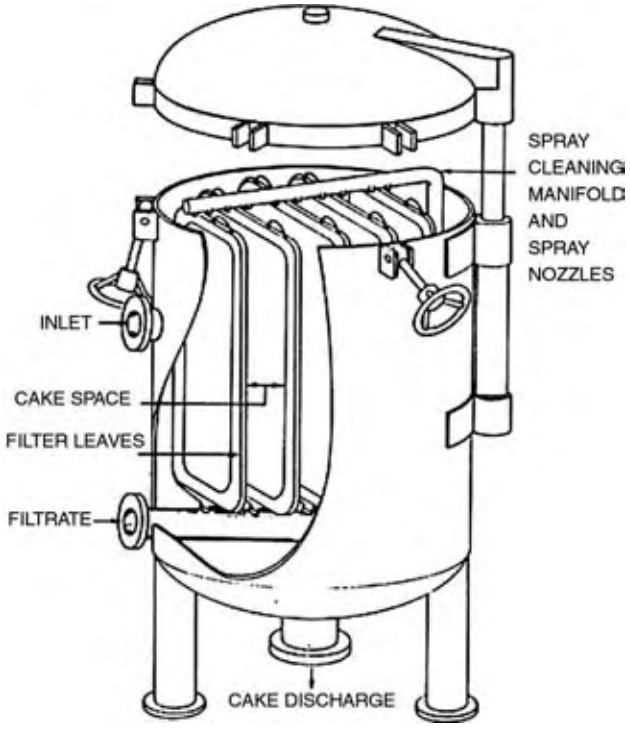


Fig. 10 Vertical tank-vertical leaf filter.

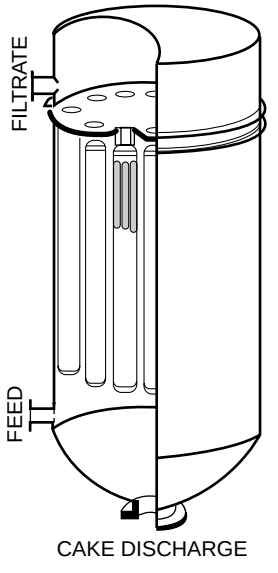


Fig. 11 Candle filter. *(View this art in color at www.dekker.com.)*

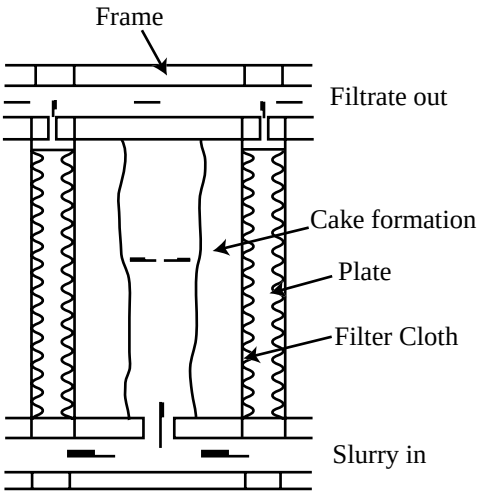


Fig. 12 Schematic of a plate-and-frame filter chamber.

From Table 3, the preponderance of liquid removed in the thickener is apparent. As slurries become dilute, larger fractions of the liquid are removed in the thickener operation. Small changes in the concentration of the stream from the thickener can have a large effect on the size of the filter or centrifuge. If the thickener effluent could be increased to 11% solids, the corresponding void ratio would be $89/11 = 8.09$. The amount of liquid to be removed in the filter or centrifuge would change from 6 (9–3) to 5.09 (8.09–3) volumes of liquid per unit volume of solid. The size of the filter or centrifuge could then be reduced about 15%.

Recognizing the importance of removing as much liquid as possible in the thickener, John Chandler pioneered a process in which the height of the red mud in the thickener was increased from the usual value of 1.0 to 10–15 m. The additional stress produced by the super-thick bed led to a much higher underflow concentrations. In some cases, the filter was actually eliminated.

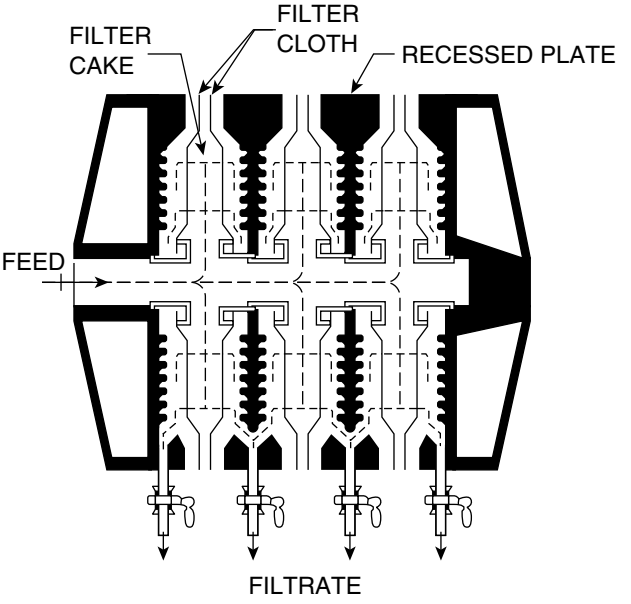


Fig. 14 The recessed plate provides space for cakes and eliminates the need for frames.

SLS EQUIPMENT[2,5,16]

An overview of some of the major types of equipment in relation to the four stages of SLS is provided in Fig. 8. After inorganic salts and polymers are used in the pretreatment, the resulting slurry is fed to a gravity thickener. The overflow from the thickener passes through a deep-bed for removal of fine particles. The underflow goes to the solids separation operations that are classified according to the driving force, i.e., gravity, vacuum, pressure, or centrifugal.

Gravity separation involving large particles is usually accomplished with screens that may be stationary, vibrating, or rotating cylinders. Large screens have slot openings greater than 0.5 in and small screens have openings of less than 0.5 in. Slot openings in rotary

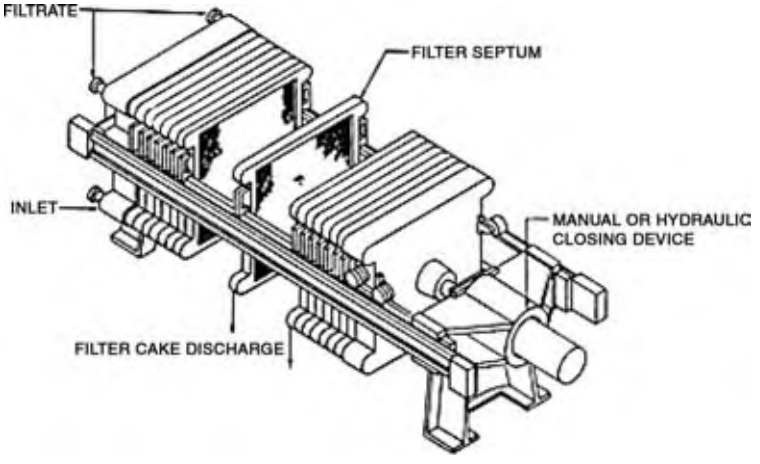


Fig. 13 Plate-and-frame filter.

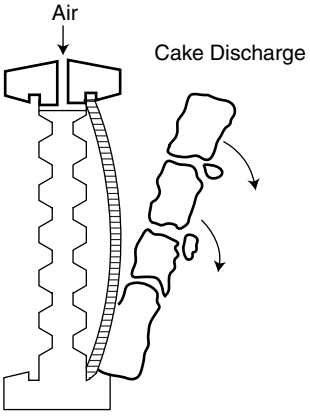


Fig. 15 Cake discharge in a membrane filter press.

screens run from 0.01 to 0.1 in. (254–2540 μm). Microscreens frequently fall in the 15–60 μm range.

In vacuum filtration, the driving force (~20 in Hg = 10 psi, 1 psi = 6895 Pa) is slightly higher than the gravity. The vacuum operation is frequently tied with continuous equipment such as drum, disc or belt filters, in which cake is removed continuously. As the permeability of cakes diminishes, pressure becomes an important element in producing a satisfactory flow rate.

Pressure filters are operated in the range of 30–60 psi and some times up to 100 psi. In contrast to vacuum filters, pressure filters normally operate in batch mode. The rate of cake buildup in batch operation is slow in comparison with continuous drum filters. The time required to dump the cake and clean and reassemble a pressure filter is called "dead time." It is a significant element in determining the capacity of a pressure filter.

In addition to vacuum and pressure, centrifugal forces are also used to increase driving force in separation of particles from liquids.

Solid–liquid separation equipment can also be classified according to principle of each unit operation. A list of equipment based on operating principle is given as follows and will be discussed.

- Batch pressure filters.
  a. Pressure vessel filters.
  b. Filter press.
- Continuous filters.
  a. Rotary drum and disk filters.
  b. Horizontal belt filters.
  c. Indexing belt filters.
- Deep-bed filters.
- Croos-flow filters.
- Membrane filters.
- Thickeners and clarifiers.
- Centrifuges.
- Hydrocyclones.
- Expression equipment.

Batch Pressure Filters[2,5]

Pressure filters are usually operated batch-wise. The batch pressure filters can be classified as tank (pressure vessel) filters or presses. Tank filters have different types of filter elements in pressure vessels. Presses consist of a series of filter surfaces (plates). The elements are mounted on a frame and are pressed together mechanically.

Tank filters

Tank (pressure vessel) filters are useful where noxious vapors are involved, and a completely closed system is desirable. Pressure vessel filters are divided into the following types:

1. Pressure nutsche filters (Fig. 9).
2. Leaf filters (Fig. 10).
3. Candle or tubular filters (Fig. 11).

Pressure Nutsche. Nutsche filters contain a single horizontal filtering surface in a pressure vessel. Gas is used to provide pressure for filtration. Because of the limited filtration area, they are often operated with thick cakes and are suitable for small batches. Automatic nutsches are available to perform reaction, crystallization, filtration, reslurry washing, drying, and cake discharging in the same vessel. Operations of Nutsche filter are shown in Fig. 9.

Leaf Filter. In comparison to nutsche filters, leaf filters provide more filtration area in the same volume of pressure vessel. They are more suitable for handling

Table 4 Brine sludge dewatering—Comparison of recessed press and membrane filter

Press type	Cake % solids	Cake structure	Cake thickness (mm)	Cake compr. strength (tpf)	Average rate (pph/ft2)
100 psig std. recess	59	Soft core	30	3.5	1.70
25 psig membrane	67	Very firm	25	>5.0	2.04

(From Ref. [20].)

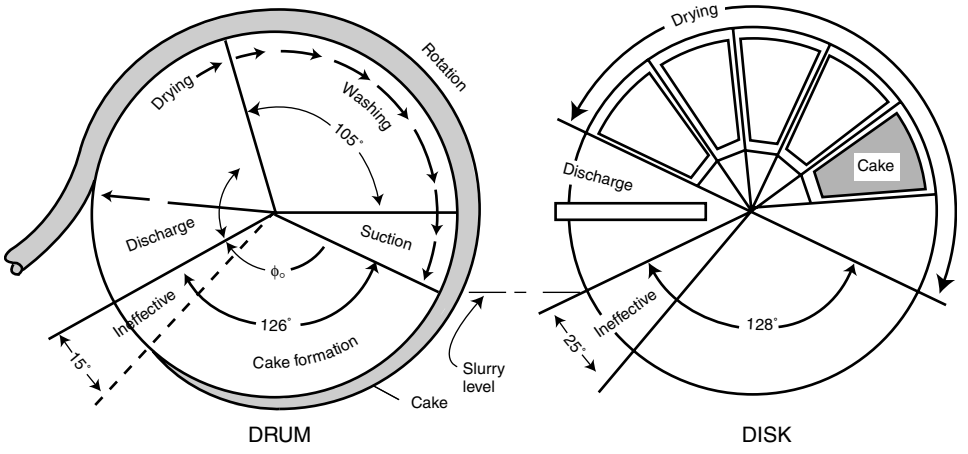


Fig. 16 Cycles for disc and drum filters show cake formation, suction, washing, drainage, and discharge areas. (From Ref.[21].)

larger quantities of slurry. Fig. 10 shows a vertical tank–vertical leaves filter. In general, cake can be discharged more easily from vertical leaf-type filters. During cake formation, to prevent dropping cake, the thickness is normally restricted to 3.5–4.0 cm. A space of about 2.0 cm must be maintained between the cakes to prevent arching and facilitate discharge.

Candle Filter.[17] Tubular filter elements contained in a matching vessel are known as candle filters. The actual filter vessel may contain one or more filter candles, and may be used as pressure or suction filters for the filtration of liquids and gases. A particular advantage offered is that candles may readily be changed to different types, to suit particular requirements or applications. A typical candle filter[18] is shown in Fig. 11. Materials of candles are selected to fit a particular process.

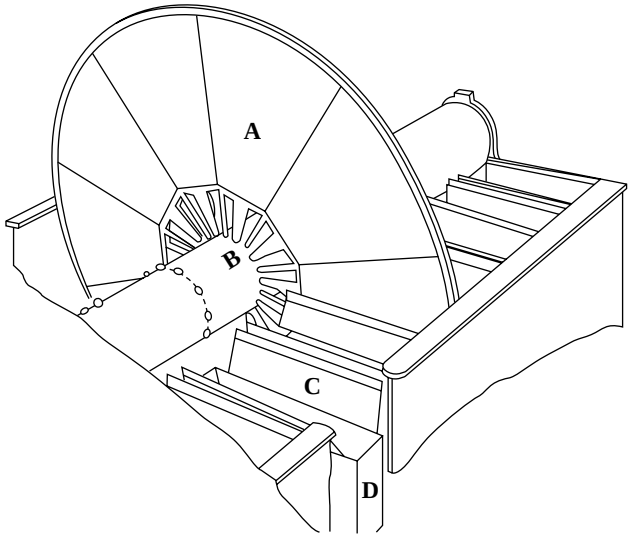


Fig. 17 Isometric view of disc filter.

Filter presses

The filter presses play a significant role in the SLS industry, where it is unnecessary to operate in a closed atmosphere. The most typical filter presses is the plate-and-frame filter press. The recessed filter press and the membrane filter press are revised, based on the plate-and-frame filter press.

Plate-and-Frame Filter Press. The major elements of the plate-and-frame version of the press are illustrated in Fig. 12. Feed to a plate-and-frame press is generally through openings at the bottom to reduce sedimenting tendencies. A filter medium (usually cloth or paper) placed over a grooved plate serves as the support for the cake, which is deposited in an adjoining frame. Plates and frames are alternated as shown in Fig. 13. The medium serves as a gasket when mechanical closure is affected. Where washing is desired, every other plate in a plate-and-frame press is constructed, so that liquid can be passed from one plate through the cake contained in the frame to the opposite plate. Thus, the wash passes through two cakes.

Recessed Plate Press.[19] Recessed plate press shown in Fig. 14 does not require a frame. The edges of the plate are extended outward leaving a space for a cake.

Membrane Filter Press. Membrane filter presses use impermeable, flexible membranes, or diaphragms to squeeze the cake for further cake deliquoring, as shown in Fig. 15. This type of filter provides less dead time in a filtration cycle, better washing, and drier cake compared to traditional plate-and-frame and recessed plate filter presses. A comparison of a recessed press operated at 100 psi and a membrane filter operated at 25 psi for sludge dewatering is shown in Table 4.[20]

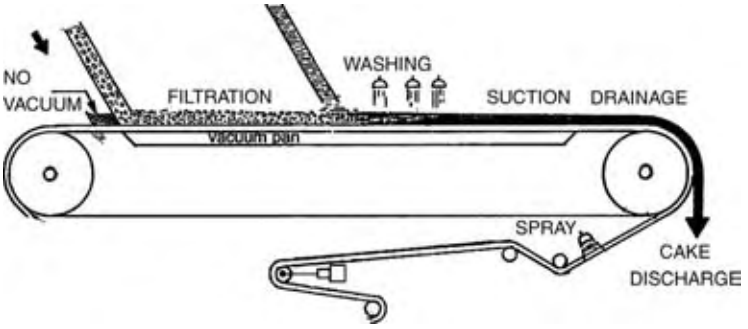


Fig. 18 Horizontal belt filter illustrating filtration, washing, and drainage stages.

Firmer cake with higher % cake solids, and faster filtrate rate were obtained for the membrane filter.

Continuous Filters[21]

Continuous filters work best on medium sized particles in the range of 5–50 μm. The larger particles generally encountered exert minor capillary forces, and cake drying or ''drainage'' can be accomplished by sucking air through the cakes under vacuum. Continuous filters are normally used for materials that are relatively concentrated and easy to filter with a cake buildup rate at cm/min. Types of continuous filter are characterized by different type of traveling filter surfaces. Rotary drum (Fig. 16), disc (Figs. 16 and 17), and horizontal belt filters (Fig. 18) will be discussed.

With drum filters (Fig. 16), the feed is below the filter surface; the slurry flows upward; and the cake faces downward. Disc filters (Figs. 16 and 17) have a vertical surface. Operations of drum and disk filters with different cycles including cake formation, washing, drying, and cake discharge are shown in Fig. 16.

A horizontal belt filter is shown in Fig. 18. It occupies more space than drum and disc filters for the same filter area. When settling is acute, it must be used in preference to the drum the disc filter. Unlike the drum and disk filters with strict requirements on cake formation, washing, drainage, and discharge time, the various stages on the horizontal filter are completely adjustable, and the entire filter surface can be utilized in contrast to the limitations of the disc and drum types.

Deep-Bed Filter

Deep-bed filters are employed for slurries with very dilute concentration less than 1000 ppm (parts per million by weight). The deep-bed has pores in which the fine particles are caught. Capture of particles in deep-beds depends upon transport mechanisms that carry the particles to the surface of the medium. As the deposit builds up, the permeability ultimately drops to a point where the bed must be regenerated or discarded. The deep beds are in the form of granular media (sand, crushed anthracite coal, garnet, usually backwashable)

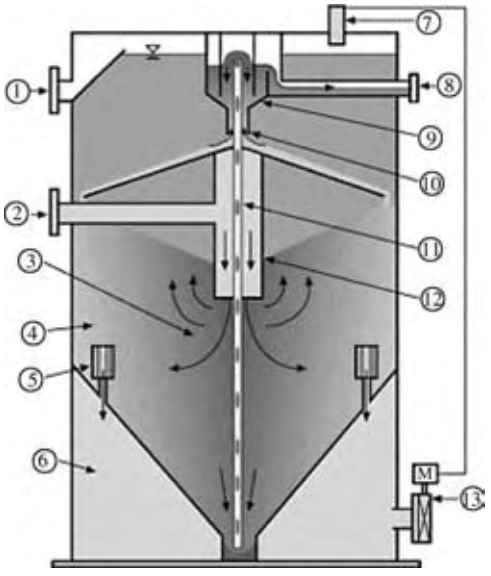


Fig. 19 Automatic backwash sand filter (Centra-flo™): 1, Overflow; 2, filter influent; 3, coarse media; 4, fine media; 5, filtrate nozzles; 6, filtrate chamber; 7, level controller; 8, filter reject; 9, washbox; 10, counter-current washer; 11, airlift; 12, central feed chamber. *(View this art in color at www.dekker.com.)*

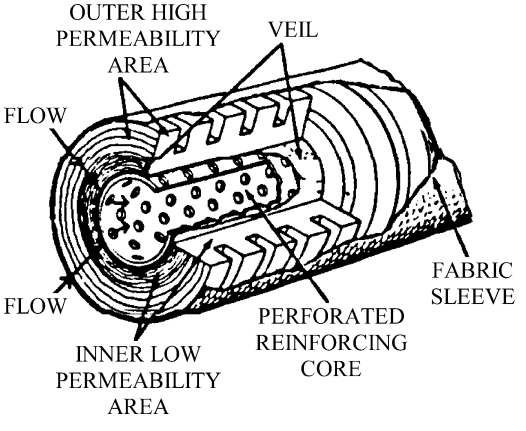


Fig. 20 Johns–Manville cartridge.

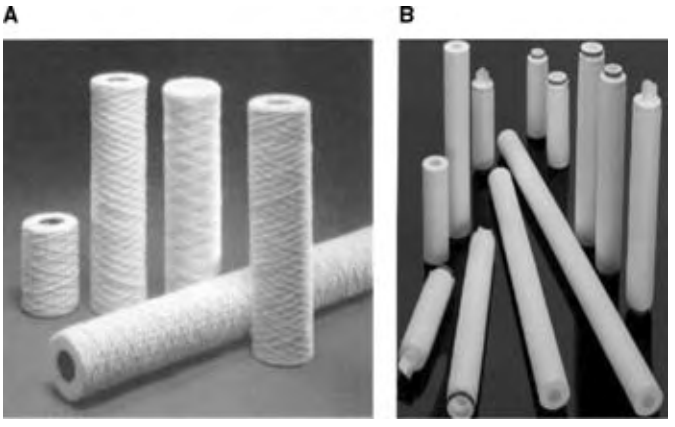


Fig. 21 Cartridge filters: (A) Wound cartridge and (B) melt blown cartridges. (From Ref.[22].) *(View this art in color at www.dekker.com.)*

(Figs. 1D and 19) or cartridges (usually disposable) (Figs. 20 and 21), which are cylinders containing a variety of materials for trapping the particles. An example of the wide application of cartridge filters includes removing particles from lubricating oils in automobiles and trucks.

Cross-Flow Filters

In cross-flow filtration (Fig. 1B) or delayed cake filtration, the slurry flows parallel to the cake surface with sufficient velocity to prevent partially or entirely the deposition of cake. It is used successfully to increase flow rate in membrane filtration. It is also employed for concentrating and recovering very fine particles in dilute suspensions when deep-bed or cake filtration would not applicable.

There are basically two types of cross-flow filter. One is without rotating element in which slurry is usually pumped into the filter in the direction parallel to the filter media to produce a cross-flow. Another type of cross-flow filter is equipped with rotating elements, such as rotary filter press with agitators or turbines attached to a rotary shaft as shown in Fig. 22.[23]

Membrane Filters[22]

Membrane separation has advantages of low energy consumption, effective multiple fine particle removal, and small waste stream. Developing of high performance and low cost membranes is a major factor in hindering the advances of membrane processes.

Based on the size of particles separated, membrane filtrations are categorized as MF (0.1–2.0 μm), UF (0.005–0.1 μm), NF (0.0005–0.005 μm) and RO (<0.5 nm). In UF and NF, a term called molecular weight cut off (MWCO—molecules with molecular weight greater than the MWCO will be rejected by this membrane) is also used for classification.

Membrane filters include flat sheet, spiral wound, and hollow fiber (Fig. 23). In addition, there are also rotating discs, annual gap, and vibrating disc systems for membrane cross-flow filtration.

Thickeners and Clarifiers[2,5,12]

Clarifiers are employed to remove small quantities of solids where the liquids are the main product.

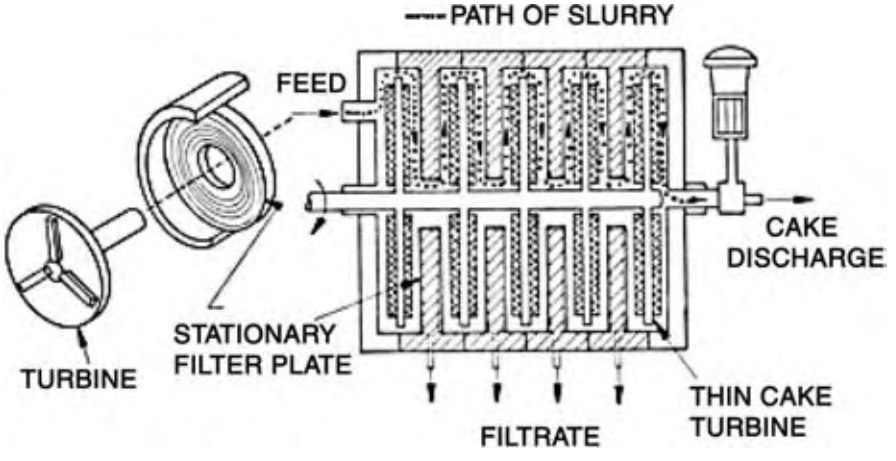


Fig. 22 Rotary filter press.

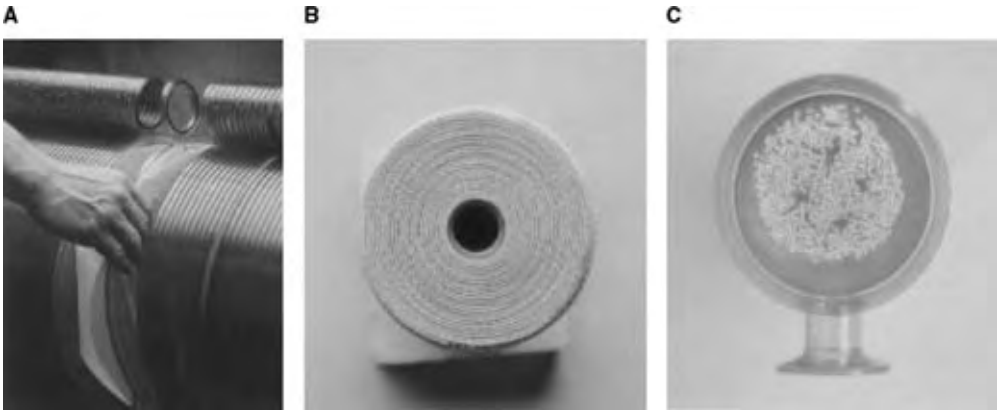


Fig. 23 Membrane filters: (A) Flat sheet; (B) spiral; (C) hollow fiber. (From Ref.[22].) *(View this art in color at www.dekker.com.)*

Thickeners are used to concentrate dilute suspensions in preparation for further separation in filters and centrifuges. As previously shown in Table 3, thickeners frequently lead to removal of a large fraction of the liquid in a slurry. A schematic continuous thickener is shown in Fig. 24. In the figure, a clear liquid overflows the top while a thickened sediment flows out of the bottom as underflow. Thickeners are widely used in fields such as wastewater, aluminum, coal, pulp and paper, cement, and sugar.

Centrifuges[2,5]

Centrifuges are operated as classifiers, sedimentors, and filters. Centrifugal classification and sedimentation are carried out in solid bowl units (frequently called decanters) as illustrated in Fig. 25. Most decanters are operated on a continuous basis. The slurry enters as shown in Fig. 25. The solids settle out and are conveyed toward discharge by a scroll, which is turning at a lower speed than the bowl. The clear liquid moves to the left and flows over a weir and is removed as "centrate."

In filtering centrifuges (Fig. 26), the solids settle out as in the decanter. However, the liquid flows in the same direction and out through the perforated bowl. As the liquid flows in the same direction as the solids, it increases the rate of deposition. In addition, as the liquid flows through the cake, the frictional effect results in sharply increased stresses on the cake. At low speeds, the stress because of the frictional flow may be larger than the centrifugal body forces.[24]

Hydrocyclones

The hydrocyclone is a type of fixed wall sedimenting centrifuge shown in Fig. 27. Compared with a sedimenting centrifuge (Fig. 25) in which the liquid is driven by the rotation of the inner wall of the centrifuge, the driving force in hydrocyclone is obtained by pumping the liquid tangentially into the upper

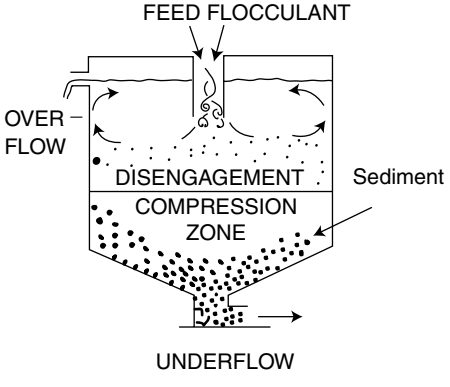


Fig. 24 A typical thickener.

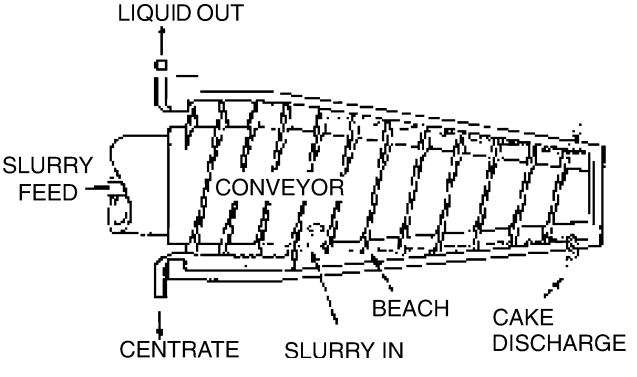


Fig. 25 Sedimenting centrifuge.

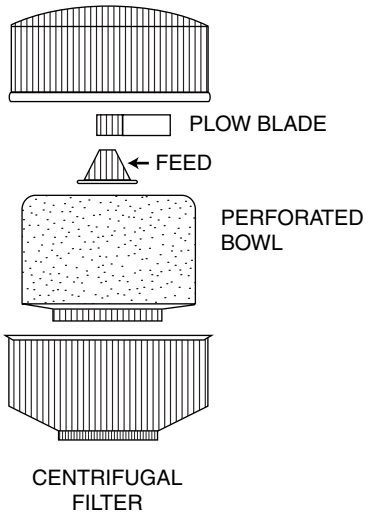


Fig. 26 Filtering centrifuge.

cylindrical section of the equipment. The total separation efficiency E_T is[22]

$$E_T = \frac{\text{Total solids in underflow}}{\text{Total solids in feed}} = \frac{U\varphi_{su}}{Q\varphi_{sF}} = \frac{U\varphi_{su}}{U\varphi_{su} + O\varphi_{so}} = \frac{Q\varphi_{sF} - O\varphi_{so}}{Q\varphi_{sF}} \quad (24)$$

where Q is the feed volumetric flow rate (m^3/s), U the underflow volumetric flow rate (m^3/s), O the overflow volumetric flow rate (m^3/s), φ_{sF} the solid concentration in the feed (volume fraction), φ_{su} the solid concentration in the underflow (volume fraction), and φ_{so} the solid concentration in the overflow (volume fraction).

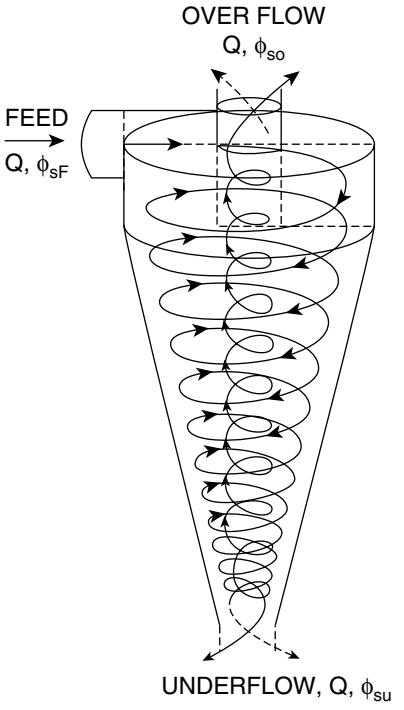


Fig. 27 Hydrocyclone.

Expression Equipment

Decreasing the liquid content of cakes has become a major objective of many SLS processes. Although the term ''dewatering'' is widely used, ''deliquoring'' is more general and covers both aqueous and

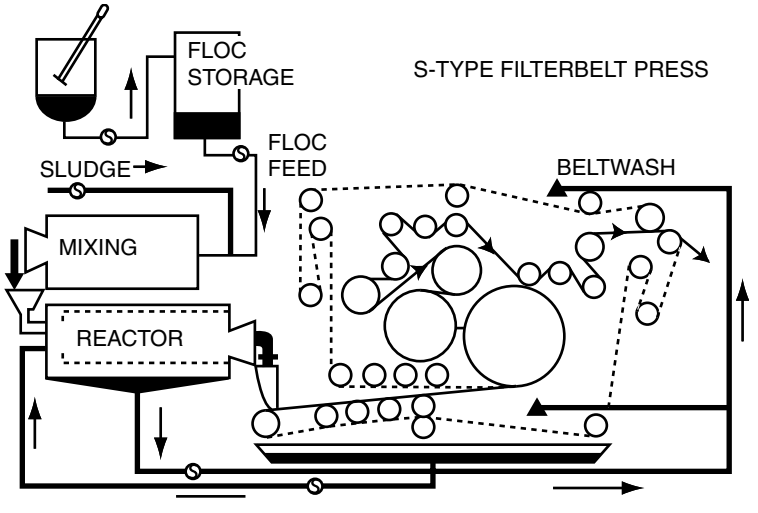


Fig. 28 S-type filter belt press.

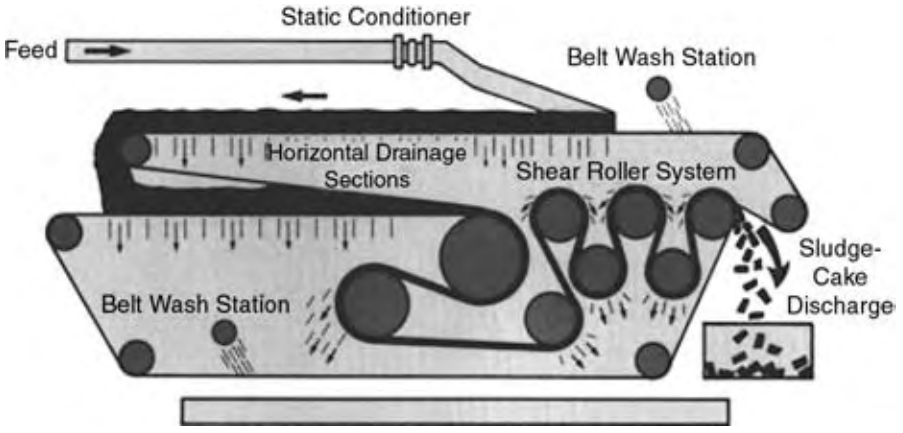


Fig. 29 Klampress belt press. (Courtesy of Ashbrook.)

nonaqueous liquids. Deliquoring operations consist of expression (squeezing), blowing, sucking with a vacuum, and gravitational and centrifugal drainage.

Fig. 28 shows the mechanically deliquoring of flocculated, fragile, and high porosity cakes by belt press filter. In the process, large flocs settle out rapidly on the porous belt, and the liquid flows through quickly. Sedimentation and filtration are accomplished quickly, but at a cost of producing a very soupy cake. Expression (the fourth stage of Fig. 5) of the liquid from the cake is essential. It is accomplished by the squeezing action of the two moving belts.

Fig. 29 shows a Klampress belt press manufactured by Ashbrook. The Klampress utilizes a combination of gravity and mechanical pressure to process porous compactible cakes. It is the type of equipment used for deliquoring of AS at many wastewater treatment plants.

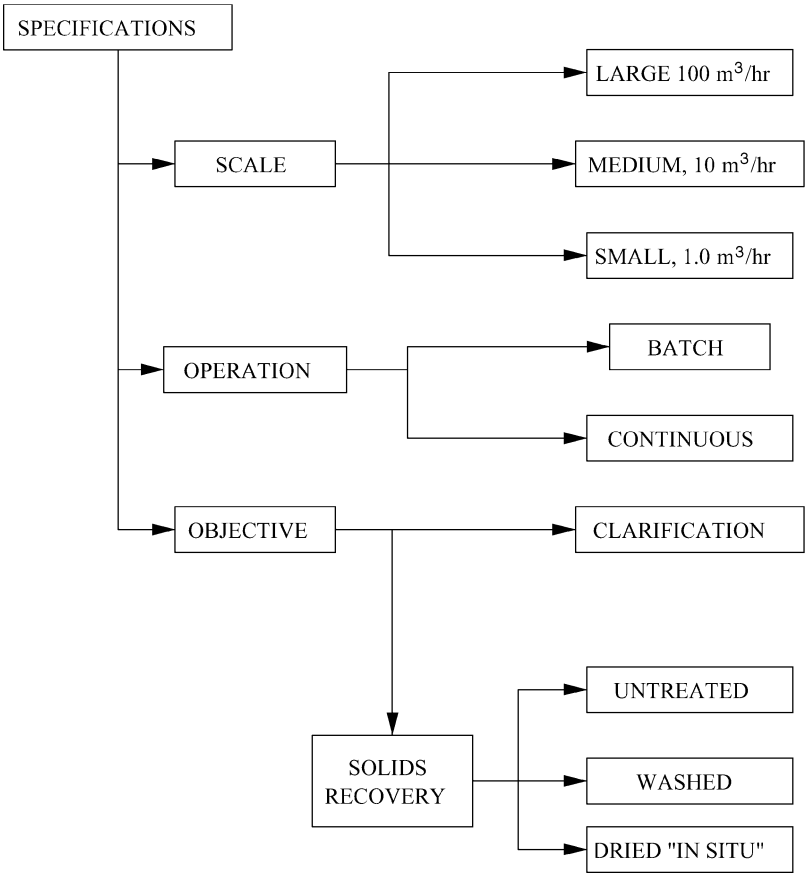


Fig. 30 Specification of scale, type of operation, and objective.

Table 5 Equipment profile

	Products			Feed				Performance		Costs		
	Solid in liquid	Liquid in solid	Wash	Solids (%)	Density	Particle size	Power	Space	Hold up	Initial cost	Operation cost	Maintenance cost
Filtration												
Drum	F	G	G–E	M–H	—	M	H	M	M	H	H	M
Disc	F	G	P–F	M	—	M	H	M	M	M–H	H	M
Horizontal	F	G	G–E	M–H	—	C	H	M	M	M	H	M
Precoat	E	NG	P–F	ppm	—	F	H–M	M	M	H	VH	M
Dresses	E	G–E	F–E	LOW	—	F	L	M	L	M	M–H	I
Leaf	E	F–E	F–E	LOW	—	F	L	L–M	M	M	M	I–M
Sedimentation												
Thickener	G–E	P	P	M	H	M	L	VH	VH	M–L	L	VL
Clarifier	G	P	VP	LOW	M	F	VL	VH	VH	M–L	L	VL
Centrifugation												
Disc	F–G	P	P	M	F	F	H	L	L	H	H	H
Solid bowl	P	F	P–F	M	M–F	M–F	H	L	L	M–H	H	H
Basket	P–F	E	E	—	M–C	M–C	H	L	L	M	H	H
Hydrocyclone												
Large	P	P–F	P	L–M	H	M–F	M–L	L	L	VL	M	H
Small multiple	P–F	P	VP	L	M–H	F	M–L	L	L	L	M	M
Screens	P	P–F	P	M–H	—	M–C	L	VL	VL	VL	M	M–H

(From Ref.[27].)
E: Excellent; H: high; F: fine; G: good; M: medium; F: fair; L: low; C: coarse; P: poor.
This table should be used as a crude approximation to be improved with experience.

SLS SYSTEM DESIGN[25–29]

Philosophy of SLS Design, Operation and Modification[25]

The facts governing the philosophy of design, operation, and modification of SLS system are as follows:

1. Many combinations of SLS equipment are capable of doing a specified job, and there are number of satisfactory and alternative choices of flocculants, filter aids, media, and filters or centrifuges.
2. The complexity of SLS phenomena frequently renders mathematical analysis inadequate. Testing and experience play predominant roles in SLS design.
3. Almost all solid–liquid systems change with time and treatment, and parameters in SLS design such as settling rate, permeability and porosity depend upon history of slurry. Therefore, continuous monitoring of operation and characteristics of suspension system is important.
4. For scale-up, certain tests, such as test for settling rate to evaluate flocculation, design of thickeners, and pressure-bomb or vacuum-leaf experiments to determine the rate of cake buildup, are easy to perform on a bench scale. On the other hand, for design of centrifuges, small-scale test results are not sufficient, and pilot or full-scale units are necessary for final design.

Selection of SLS equipment

Specifications

In approaching the initial selection of equipment, flows to be processed, filtrate clarity, and cake liquid, and solute contents must first be fixed. The product may be the solids, liquid, or both. Purchas[26] provided a conceptual model for initial specifications as illustrated in Fig. 30.

Comparative performance profiles

Fitch[27] discussed the general problem of matching process specifications to SLS equipment and the necessity of considering trade-offs involving such items as filtrate clarity, cake dryness, reliability, maintenance, versatility, and cost. A comparative equipment profile of the principle types of SLS is given in Table 5. A wide range of information about power, space, and costs as well as equipment performance parameters are provided in the table.

Rate of cake buildup

The rate of cake thickness growth is perhaps the single most important guide in selecting equipment. Growth rate determines whether pressure, vacuum or gravity filters, or solid or perforated bowl centrifuges are indicated. Initial choice of equipment is suggested in Table 6 as an approximate guide only.

The basic ideas of the table can be summarized as follows:

1. Rapid settling slurries lend themselves to gravity separation and screen-bowl centrifugation.
2. Medium speed filtering materials are suited to vacuum equipment and peeler centrifuges.
3. Slow filtering materials require pressure equipment or sedimenting centrifuges.
4. Dilute materials that result in high resistance cakes are generally best handled in deep-bed filters or mixed with filter aids.

Table 6 Slurry characteristics

Type	Buildup	Characteristics
Rapid filtering	CM/SEC	High concentration Fast setting Hard to suspend
Medium filtering	0.05–5.0 CM/HR	Forms cake easily Can be suspended Medium concentration
Slow filtering	0.1–5.0 CM/HR	Low concentration High resistance May blind media
Clarification	No cake	Very high resistance Less than 1000 ppm

(From Ref.[25].)

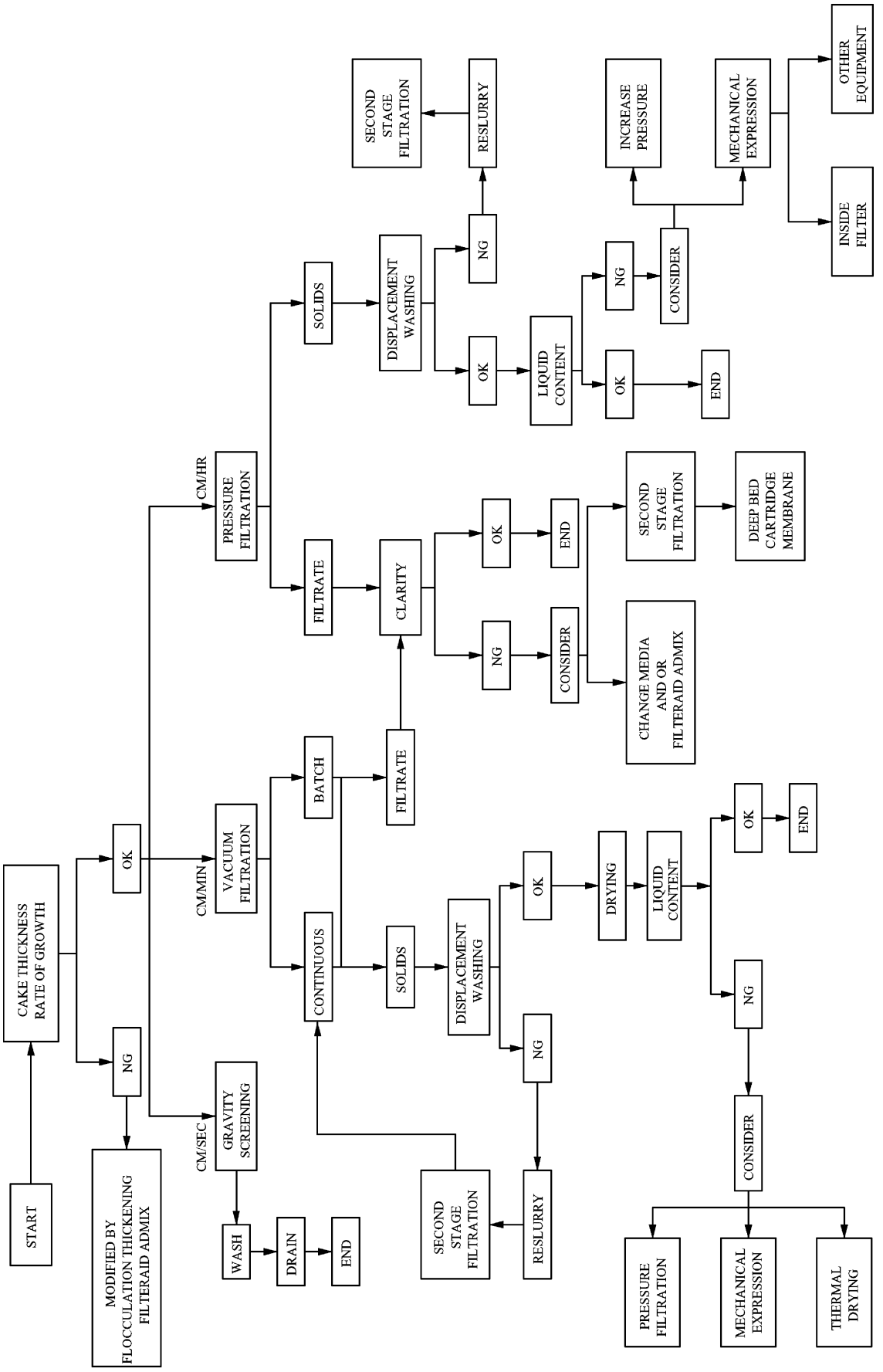


Fig. 31 Decision network for SLS system design. (From Ref.[25].)

Decision Network for Design of SLS Systems

A series of decisions must be made concerning the following factors in designing new or revising existing SLS systems:

1. Rate of cake buildup.
2. Filtrate clarity.
3. Solubles in cake.
4. Liquid content of cake.

In Fig. 31, a framework for making decisions at different junctures is provided. At the first step, a decision must be made as to whether or not pretreatment may increase the rate of filtration and enable the process to progress from pressure toward gravity filtration. Once a particular filter has been chosen, attention then focuses on the filtrate and wet cake. If the clarity is not satisfactory, a second stage deep-bed filter (cartridge or granular bed) may be utilized. Alternatively a tighter medium or a change in filter aid quality or quantity might provide a solution.

CONCLUSIONS

Solid–liquid separations are important operations in chemical engineering ranging from upstream raw material purification, to downstream product separation and waste managements. They are usually low cost, and low energy consumption. A successful operation of a SLS system upstream a major chemical engineering unit such as a catalyst bed will help to improve the efficiency the catalyst, and decrease overall cost of the whole process.

Particle size, shape, inter-particle forces, zeta potential, liquid surfactant phenomena, and liquid viscosity are important characteristics of a solid–liquid suspending system. Mechanism of flow through porous medium is fundamental to theories of sedimentation, filtration, centrifugation, and expression operations. Most solid–liquid materials are compactible. Unique and strange behavior of pressure filtration of compactible materials has been identified. More attention should be paid for separation of those materials.

Solid–liquid separation systems generally consist of four stages including pretreatment, solid concentration in thickeners, solid separation in filters or centrifuges, and post-treatment by expression and washing operations. There are different types of SLS equipment served for different functions in relation to the four stages. Product specification, characteristics of solid–liquid suspension, solid settling velocity, rate of cake buildup should be considered in equipment selection and SLS system design. Many combination of SLS equipment are capable of doing a specific job. Equipment cost, operating cost, separation efficiency, and separation specification should be considered in design and optimization of a SLS system and selection of flocculants, filter aids, filter medium, and separation equipments.

REFERENCES

1. Svarovsky, L. *Solid–Liquid Separation*, 3rd Ed.; Butterworths & Co Ltd., 1990.
2. Tiller, F.M.; Li, W. *Theory and Practice of Solid/Liquid Separation*, 4th Ed.; University of Houston, 2002.
3. Rushton, A.; Ward, A.S.; Holdich, R.G. *Solid–Liquid Filtration and Separation Technology*, 1st Ed.; VCH Verlagsgesellschaft mbH: Weinheim, Federal Republic of Germany, VCH Publishers, Inc.: New York, U.S.A., 1996.
4. Darcy, H. Determination of the laws of flow of water through sand. Appendix to *Histoire des Fontaines Publiques de Dijon*, 1856; 590–596 (republished in Fluid/Particle Sep. J. **1989**, *2*, 33–35).
5. Tiller, F.M.; Lloyd, P.J., et al. *Theory and Practice of Solid/Liquid Separation*, University of Houston, 1978.
6. Tiller, F.M.; Yeh, C.S. The role of porosity in filtration. Part 10. Deposition of compressible cakes on external radial surfaces. AIChE J. **1985**, *31*, 1241–1248.
7. Tiller, F.M. Tutorial: interpretation of filtration data I. Fluid/Particle Sep. J. **1990**, *3* (2).
8. Tiller, F.M.; Kwon, J.H. The role of porosity in filtration. Part 13. Behavior of highly compactible cakes. AIChE J. **1999**, *44*, 2159.
9. Tiller, F.M.; Li, W. Dangers of lab-plant scaleup for solid/liquid separation systems. Chem. Eng. Commun. **2002**, *189*, 1655–1677.
10. Tiller, F.M.; Li, W. Determination of the critical pressure drop for filtration of super-compactible cakes. Wat. Sci. Technol. **2001**, *44* (10).
11. Tiller, F.M.; Li, W. Explaining strange behavior of highly compactible materials. Chem. Process. **2000** September.
12. Tiller, F.M.; Chen, W. Limiting operating conditions for continuous thickeners. Chem. Eng. Sci. **1986**, *43*.
13. LaHeij, E.J. An Analysis of Sludge Filtration and Expression D. Eng. Diss., Technische Universiteit Eindhoven, Eindhoven, The Netherlands, 1994.

14. Kwon, J.H. Effects of Compressibility and Cake Clogging on Sludge Dewatering Characteristics. Ph.D. Diss., Seoul National University, Korea, 1995.
15. Grace, H.P. Resistance and compressibility of filter cakes (Parts I and II). Chem. Eng. Prog. **1953**, *49*, 303–318, 367–377.
16. Tiller, F.M.; Leu, W. Solid–Liquid Separation for Liquefied Coal Industries, Final Report for Project 1411-1, July, 1984.
17. Dickenson, C. *Filters and Filtration Handbook*; The Trade and Technical Press Limited: England, 1987.
18. Purchas, D.B. *Solid/Liquid Separation Technology* Upland Press Ltd.: Croydon CR9 1LB, England, 1981.
19. Alliot, E.A. Recessed plates, plate and frame, filter press: their construction and use. J. Soc. Chem. Ind. **1920**, *39*, 261–285T.
20. Mayer, E. Wastewater sludge dewatering with membrane filter presses. The 15th Annual AFS Technical Conference and Expo, 9–12 April 2002, Galveston, TX.
21. Tiller, F.M. Batch, continuous processes for cake filtration. Chem. Eng. **1974**, *29*.
22. Chen, W. Solid/liquid separation fundamentals and practice. AIChE Today Ser. **1996**.
23. Tiller, F.M. Characteristics of staged, delayed-cake filters. Filtration Sep. **1978**, May/June.
24. Tiller, F.M.; Horng, L.L. Hydraulic deliquoring of compressible filter cakes. Part I. Reverse flow in filter presses. AIChE J. **1983**, *29*, 297–305.
25. Tiller, F.M. Bench-scale design of SLS system. Chem. Eng. **1974** April 29, 117–119.
26. Purchas, D.B. *Solid/Liquid Separation Equipment Scale-up*; Uplands Press Ltd.: England, 1977.
27. Fitch, B. Filtration Sep. **1975**, *12*, 355–359, 482–483, 488, 553, 636–638.
28. Flood, J.E.; Porter, H.E.; Rennie, F.W. Filtration practice today. Chem. Eng. **1966** June 20, 163–181.
29. Purchas, D.B. *Industrial Filtration of Liquids*; CRC Press: Cleveland, 1967.

Solvent Refining Processes

Roland H. Heck
Princeton University, Princeton, New Jersey, U.S.A.

INTRODUCTION

Solvent refining can be a very effective method for separating chemicals or groups of chemicals based on their solubility or chemical affinity. As such, solvent refining processes are among the group broadly known as separation processes. For separation processes in general, the desired component or components of a feed are selectively concentrated in one phase, while the less desirable components concentrate in a second phase. In solvent refining, the solvent must be separated from the products so that the products can be recovered and the solvent can be recycled. A major issue with any solvent refining process is finding a relatively inexpensive solvent that can affect the desired separation and then can itself be recycled cost effectively. Although solvent refining is used in a number of industrial applications, the largest volume and perhaps the most economically important applications are in petroleum refining where they are used in the production of high quality lubricants and for removal of light aromatics from refinery steams. Accordingly, these processes will be the main focus of discussion that follows.

BACKGROUND

The most commonly employed and oldest of the separation processes is fractional distillation, a process in which compounds are separated based on their dissimilar volatility or boiling points. In fractional distillation, components are fractionated between the liquid and the gas phase, with the lighter, lower boiling components preferentially concentrating in the gas phase and the heavier, higher boiling components concentrating in the liquid phase. Fractional distillation equipment normally employs a cylindrical column in which the phases are separated and remixed a number of times as they move vertically through the column. It is important that the phases remix and separate a number of times in the column so that the phases are enriched in their desired component as they move toward the end of the column.

Solvent extraction is analogous to fractionation, except that two immiscible liquid phase are employed instead of gas and liquid phases. Extraction equipment usually involved a vertical cylindrical vessel in which the lower density liquid phase rises up the column, while the high density liquid phase moves down the column. These phases are remixed and resettled several times as they pass through the column, with the desired component concentrating toward one end of the column and the undesired component concentrating in the other end. Understanding the physical principles and design methodology of separation processes has long been the parlance of chemical engineers, and their textbooks on separation processes teach the mathematical and graphical analysis of both fractional distillation and extraction in an analogous fashion.[1]

Perhaps the earliest recorded attempt at solvent processing occurred in Vega, Rumania in 1909 where a process attributed to Erdeleanu was developed to reduce the smoke in flames produced when burning kerosene derived from Rumanian crude oils.[2] The solvent of choice in this operation was liquid sulfur dioxide. The first commercial application of this process was in Rouen, France in 1911.

Development efforts aimed at using solvent refining to improve the quality of lubricating oils in a manner similar to that of Erdeleanu soon followed.[3,4] Solvent dewaxing, solvent deasphalting, and solvent extraction processes to improve the cold flow properties, viscosity index, and oxidative stability of lube oils were commercialized in the 1930s and 1940s.

Efforts to identify better solvents and process configurations for solvent refining continue to this day; however, in recent years even more effort has been devoted to finding catalytic or membrane-based processes that could potentially replace the relatively expensive solvent-refining processes. Catalytic processes invariably involve hydroprocessing either to isomerize or crack linear alkanes and improve the cold flow properties, or to hydrocrack to selectively remove aromatics.[5] The literature on membrane processes to displace solvent refining is sparse;[6] however, this area holds the most promise if suitable membranes can be developed. Although a search for more cost-effective processes to replace solvent refining will continue, it is doubtful that they will displace solvent refining for several decades, even if successful.

The important commercial solvent-refining processes are covered every other year in a special Refining Handbook which appears in a monthly issue of

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120038527

Hydrocarbon Processing. In this handbook many licensors of currently available processes present a description of their process along with operating conditions and yields. Numerous solvent-refining processes, including those described below, are covered in this handbook.[7]

LUBE MANUFACTURING

A primary use of solvent refining is in the production of lubricating oils. There are a number of crude oil components that impart undesirable properties to lube oils, and their removal from lube base stocks is required if high quality lubes are to be produced. These undesirable components include waxes, heavy aromatics, asphaltenes, and organic compounds containing oxygen, sulfur, and nitrogen. For most of the last century, solvent refining has been found useful for the effective removal of these components from lube base stocks. The most difficult part of solvent extraction is identifying a suitable solvent that preferentially dissolves or rejects the desired component. This solvent must also be stable at the conditions to be employed and must be easily recovered for recycling. There are only a dozen or so solvents that have been used commercially in lube manufacturing over the last several decades.

Lube Treating

Lube treating generally refers to a solvent extraction process whereby aromatics and heteroatom-containing molecules (e.g., sulfur, nitrogen, oxygen) are removed to improve lube base stock properties. The most important lube properties impacted in solvent treating are viscosity index, oxidative stability, color, additive response, and sulfur content. The earliest lube-treating processes employed liquid sulfur dioxide as a solvent, but this has been largely phased out in favor of the organic compounds phenol, furfural, and n-methyl-2-pyrrolidone (NMP). Of these, NMP is considered the least toxic and several refiners have converted their phenol or furfural units to NMP in recent years.

Several companies have commercialized lube-treating processes that they now offer for licensing by others. For example, Bechtel offers a process based on the solvent n-methyl-2-pyrrolodine. In their literature, they offer the process flow diagram shown in Fig. 1.[8] Most of the process equipment in Fig. 1 is downstream of the extraction tower, illustrating the effort involved in recovering product and clean solvent for recycle. A major part of the expense associated with this and most solvent-refining processes is the energy cost. Bechtel gives the following estimates for typical utilities required per barrel of feed:

Fuel, absorbed, 10^3Btu	130
Electricity, k Wh	0.8
Steam, lb	8.1
Water, cooling (25°F), gal	550

The Bechtel process shown in Fig 1 has been employed in 13 licensed units, with two additional units being planned in 2001. Of the 13 units on stream in the year 2000, eight were conversions of existing phenol or furfural units. In a previous article on lube treating, Texaco claimed that over 100 units were at that time utilizing their furfural-based lube-treating process.[9] Although some furfural and phenol units have been converted to NMP, the preponderance of lube-treating units use furfural as solvent.

In the manufacture of lube base stocks, the product from solvent extraction using furfural, NMP, or other solvent is sent to lube dewaxing. The objective in dewaxing is to remove enough of the waxy components (e.g., long straight chain alkanes) to reduce the pour point of the base stock to acceptable levels (as low as −35°C). In dewaxing, the feed is mixed with ketone solvent and then chilled to a sufficiently low temperature to cause the wax to crystallize and drop out of the solution. The crystallized wax is recovered by filtration, and the solvent is recovered from the filtered product and recycled. The wax produced can also be a valuable product as it is the major source of waxes for consumer products, such as candles, crayons, and wax coatings. ExxonMobil offers their DILCHILL process for license and provided the flow diagram shown in Fig. 2.[10] A key unit in this scheme is the crystallizer tower in which ExxonMobil claims to have proprietary technology that achieves a number of mixing stages. The DILCHILL process can produce lube base stocks having low pour points from a wide range of stocks. As with any solids-handling process in refining, a great deal of know-how is involved in the detailed design of the equipment in this process.

Deasphalting

A principal source of the heaviest lube base stocks is vacuum-reduced crude (vacuum residuum). The primary application of the deasphalting process is to reduce the Conradson carbon residue (CCR), metals (Ni, V, etc.) content and viscosity of a vacuum residuum in order to produce an acceptable quality heavy lube base stock. The CCR content of a refinery stream is determined in an ASTM analytical test that involves

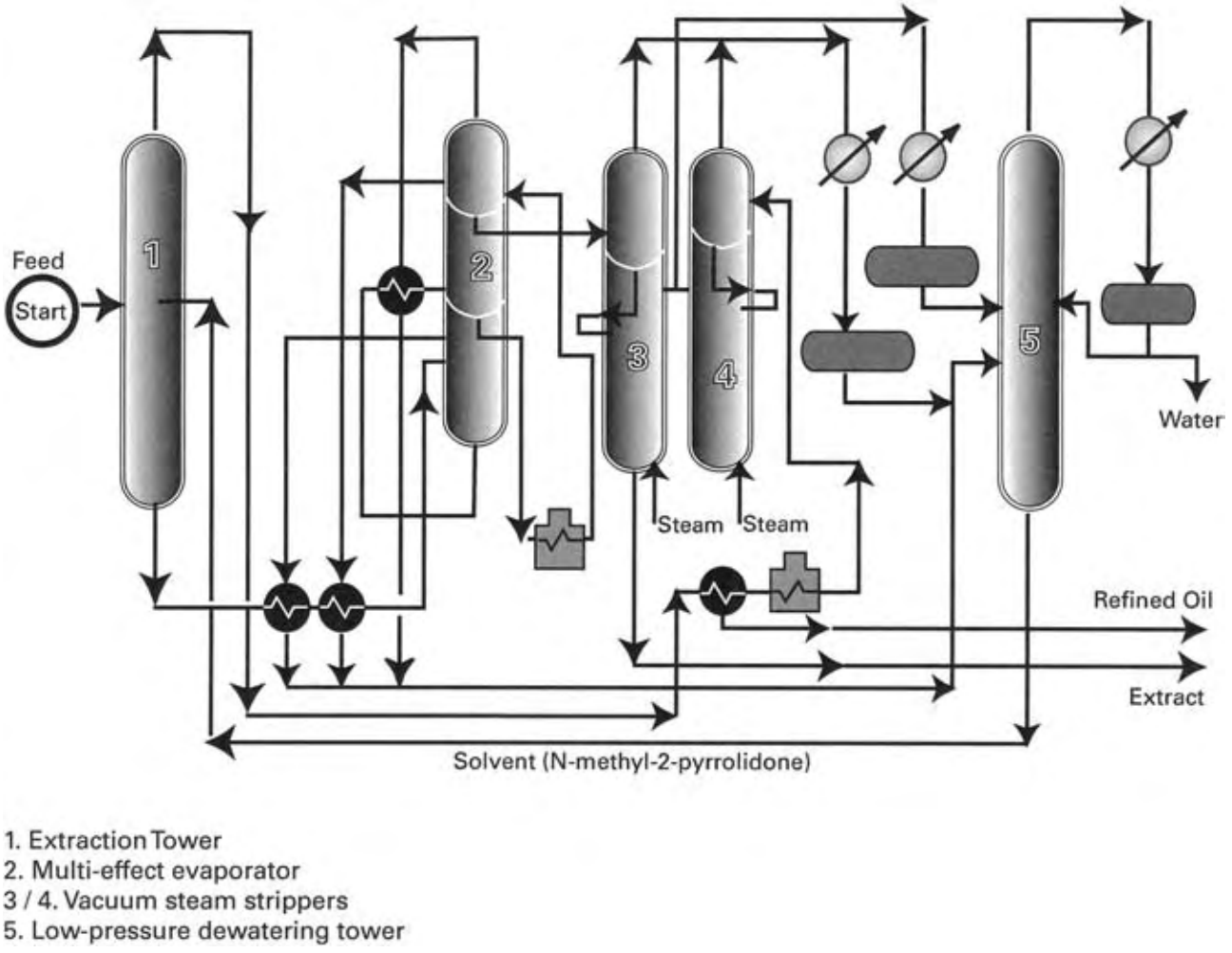


Fig. 1 Lube treating. (From Ref.[9].)

destructive distillation of the sample to dryness. The amount CCR (wt% solid residue) that is measured in this test correlates well with an oil's tendency to break down and form carbon deposits in many petroleum processes. Metals are usually associated with asphaltenes and are themselves undesirable lube components.

Deasphalting in lube manufacture is a precipitation process that usually employs propane as the solvent. Asphaltenes are large complex organic structures composed largely of aromatic rings. They contain primarily hydrogen and carbon atoms, but contain sulfur, nitrogen, nickel, vanadium, and iron, as well as trace amounts of many other atoms. Asphaltenes have a limited solubility in crude oil, but are believed to remain in solution as a result of their ability to absorb resins on their exterior surface. When lower molecular weight alkanes are mixed with a heavy oil, the resins desorb from the asphaltenes thus causing the asphaltenes to flocculate and drop out of solution. The lower the molecular weight of the alkane added, the greater the amount of the oil that precipitates as asphaltenes. Often, a little butane is included along with the propane in order to slightly modify the selectivity of the deasphalting process. The solvent to oil ratio for most deasphalting processes ranges from about 2 to 8 depending on the solvent employed, the feed stock, and the desired product quality.

In recent years, deasphalting has also been used to reduce the CCR, nickel, and vanadium content of petroleum residua as a pretreatment for catalytic cracking or hydrocracking. In this application, a heavier alkane solvent such as pentane or hexane is usually preferred, since it yields a higher volume of deasphalted oil of acceptable quality for catalytic cracking feed. All catalytic units can tolerate some level of CCR and metals in the feed; however, operation of any cracking unit is improved if they are removed. Each refiner determined the level of these contaminants acceptable for their cracking unit, but the feed to conventional catalytic crackers rarely contains more than a few parts per million of metals and a few weight percent CCR.

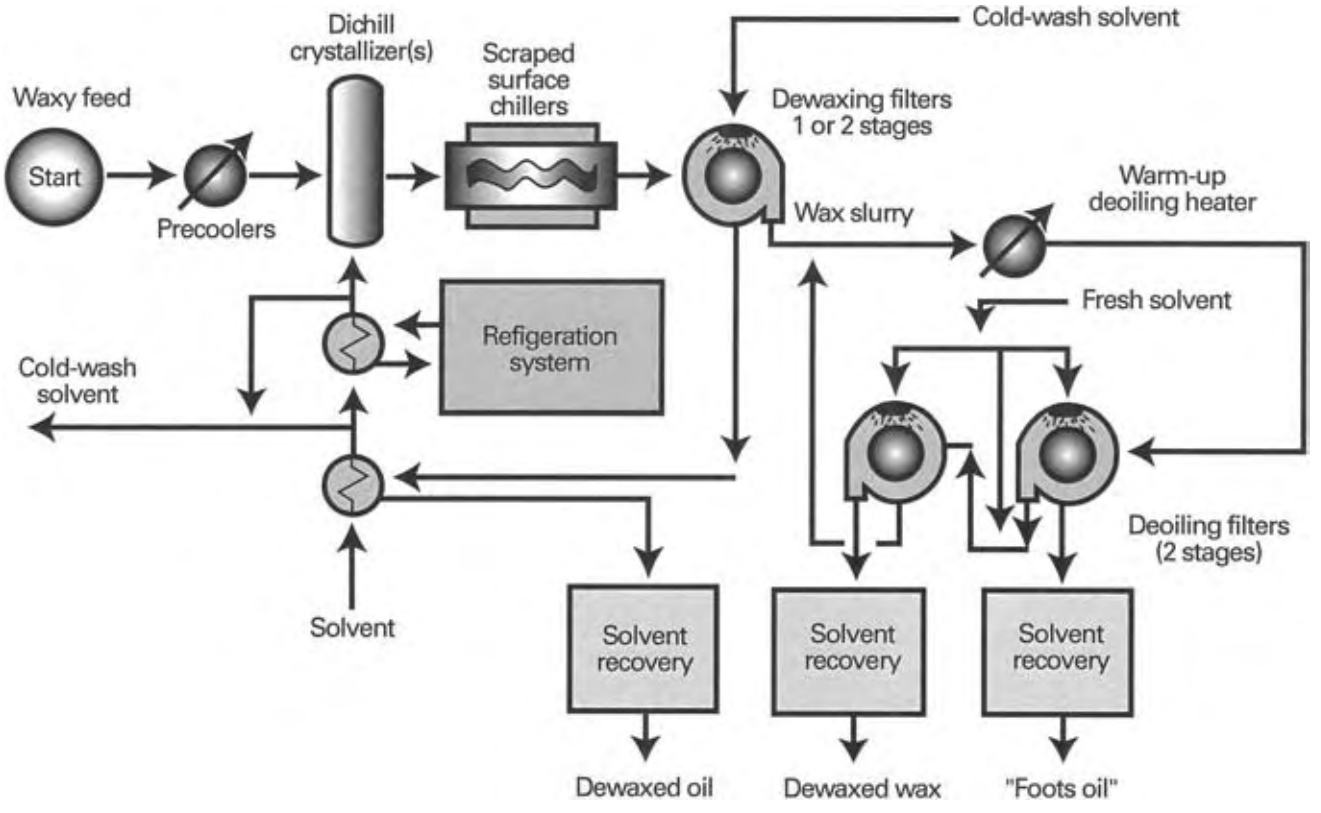


Fig. 2 Lube dewaxing. (From Ref.[9].)

Three licensors, Foster Wheeler USA Corp, Kellogg Brown & Root, Inc., and UOP LLC, offer flow diagrams and a discussion of their process in *Hydrocarbon Processing*.[11] These process vendors cumulatively report over 110 commercial applications of this technology. Each licensor offers their process for application in lube processing or in the preparation of catalytic cracker or hydrocracker feed. All vendors also discuss the use of supercritical conditions for solvent recovery. This is discussed later under ''Solvent Recovery.''

Alternative Routes to Lube Manufacturing

In recent years, efforts have been made to improve the yield and quality of lube manufacturing, while at the same time greatly improving operating costs by eliminating many of the solvent processes. Two different approaches are worthy of note. The first is to use a membrane separation process in lube dewaxing, and the second is to use catalytic hyrdroprocessing[5] for both dewaxing and lube treating.[6] Although the membrane concept has been around for several years, its use has been limited to that described in the above publication.

Although the hydrocatalytic lube-processing methods have been commercialized and are available from a number of licensors, still their application has been limited to date.

AROMATIC EXTRACTION

Benzene, toluene, and xylene (BTX) are valuable as chemical feedstocks and have been solvent extracted for this purpose from refinery streams for decades. Governmental regulations limiting the amount of benzene in gasoline have provided incentives for the extraction of aromatics from gasoline components in recent years and are found in most refineries today.

UOP, a worldwide leader in aromatics processing, has three aromatics extraction processes[12] (Udex, Sulfolane, and Carom^{TM}), which employ liquid–liquid extraction for the production of BTX from petroleum naphtha and other gasoline-blending components. The Udex process is the oldest and although there are many Udex units operating today Sulfolane and Carom have largely displaced Udex in new installations. The Carom process is a two-solvent process that can be employed to debottleneck or revamp an older glycol-based UDEX unit. There are 89 Udex/Carom units operating today. The Sulfolane process uses sulfolane as solvent and 136 of these units have been licensed since 1962. Other companies offer similar aromatics

extraction processes, with many of the processes employing a two-column system, the first being a liquid–liquid extraction column and the second an extractive distillation column. Sulfolane is effective both in extraction and extractive distillation.

Extraction and extractive distillation can also be accomplished using N-Forylmorpholin (NFM) as the solvent. Krupp Uhde[13] offers an extractive distillation process for license that employs NFM as the solvent. N-Forylmorpholin is employed in the extractor and in the distillation column to lower the vapor pressure of the aromatics relative to nonaromatics of a similar boiling point. In this way aromatics can be separated from nonaromatics in a very small and simple distillation column. N-Forylmorpholin has properties that make it an excellent solvent for this application. It is highly selective, thermally stable and has a boiling point in the range of the products to be separated. A flow diagram for this extractive distillation process is given in Fig. 3.

Aromatics concentrate in the solvent-rich stream that comes off the bottom of the extractive distillation column, while the nonaromatics concentrate in the column overhead. The main column bottoms are sent to a stripper column for solvent and aromatics recovery. Extensive heat exchange to recover energy from the hot solvent is required to achieve acceptably low energy consumption in this process. Krupp Uhde, the process licensor, claims that 45 commercial plants are operating with this technology. They also claim their process can produce excellent purity benzene (99.95%) and toluene (99.98%).

SOLVENT RECOVERY

As can be seen in the flow diagrams included in Fig. 3, solvent recovery is a very important part of any solvent process. Not only is there a large capital cost associated with solvent recovery, recovery by vaporization/distillation also requires a great deal of energy leading to high operating cost. A number of alternatives to straight distillation have been developed over the years in an effort to reduce the operating costs of solvent recovery in these processes. Three such developments are discussed below.

1. *Multieffect evaporization*: This is a classical engineering technique in which a solvent is removed by successive evaporation steps.[14] The solvent removed from the first step is used to heat the liquid which is again fed to the

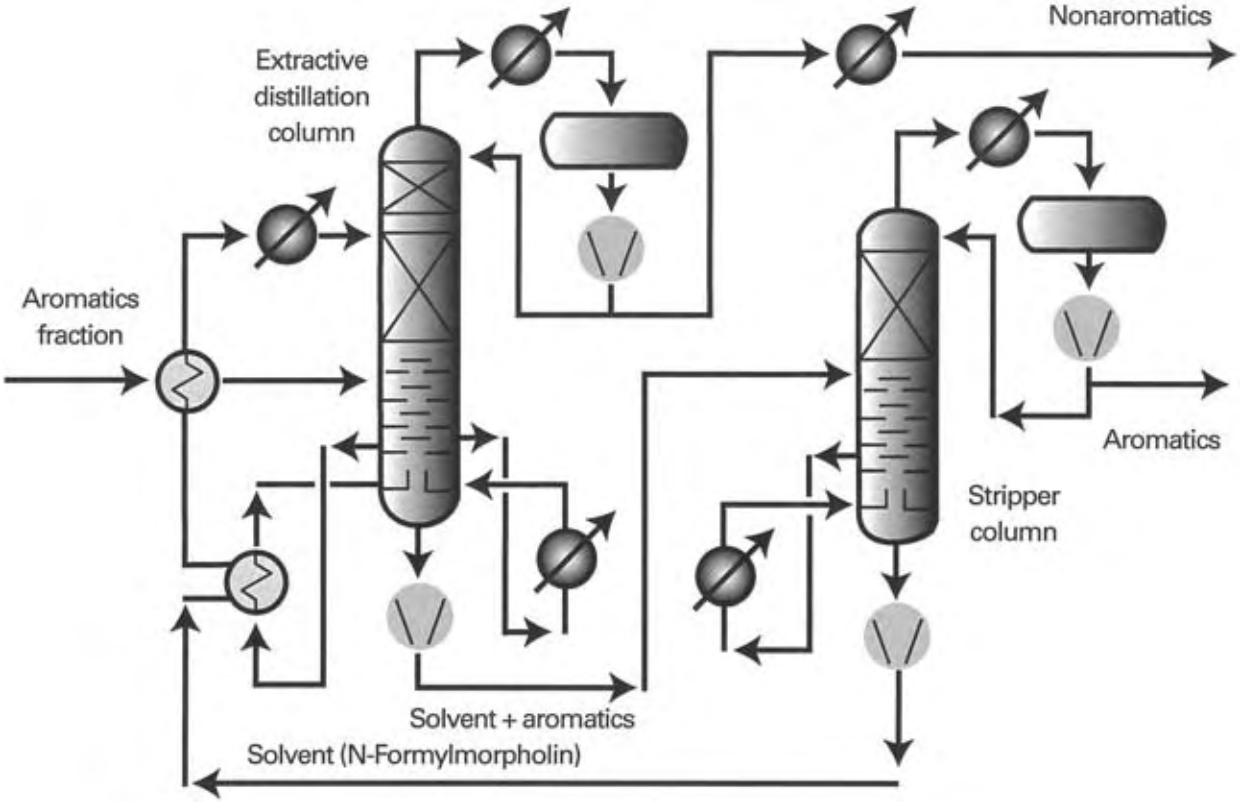


Fig. 3 Aromatics extractive distillation. (From Ref.[13].)

second evaporator where the pressure and boiling point are lower than in the previous evaporator. In this manner, the external heat applied to the first evaporator is employed a number of times. Although multiple-effect evaporators save energy, they add complexity and capital cost to a solvent recovery system. Both forward feed and backward feed have been employed in multieffect evaporators.

2. *Supercritical solvent separation*: This method has been employed to recover solvent in deasphalting.[15] In recovering solvent from deasphalted oil by this method, the deasphalted oil laden with solvent is pumped to a pressure slightly higher than that of the critical pressure of the solvent (e.g., propane critical pressure = 42.1 Atm). At this supercritical condition, the solvent is no longer completely miscible with the deasphalted oil and two phases result, the supercritical solvent phase and the deasphalted oil phase that is devoid of much of its solvent. These two phases can be separated in a simple liquid/liquid separator. The result is solvent recovery without solvent vaporization.
3. *Membrane separation*: The newest scheme for solvent separation/recovery involves the use of a proprietary polyimide membrane for recovering methyl-ethyl ketone and toluene from a solvent dewaxed oil filtrate.[6] A major advantage of this selective membrane process is that it removes about 25% of the cold solvent as a liquid at or near dewaxing temperatures (–18 to 0°C), thus debottlenecking the refrigeration unit by about 10%. Mobil and W.R. Grace jointly developed this new technology and utilized it commercially to debottleneck the refrigeration and solvent recovery system in Mobil's Beaumont, Texas refinery. An advantage of the process is its ability to remove solvent at or near dewaxing conditions, thus saving significantly on refrigeration/energy costs compared to an evaporative recovery method.

CONCLUSIONS

Solvent-refining processes belong to the general set of processes known as separation processes. Fractional distillation is usually the separation process of choice; however, when a separation is based on chemical type rather than molecular size or boiling point, solvent refining may be the preferred separation process. Solvent-refining processes have been extensively used in the refining industry for several decades for the production of high quality lubricants and for the removal of aromatics from gasoline. In the production of lubricant base stocks, hundreds of solvent-refining units are employed in a variety of licensable processes aimed at:

1. *Solvent dewaxing*: Solvent-enhanced precipitation to remove waxes and improve low temperature fluidity of lube base stocks.
2. *Solvent deasphalting*: Solvent-enhanced precipitation process to remove asphaltenes from heavy lube base stocks and for improving the feed quality of catalytic cracker feed.
3. *Lube treating*: Solvent extraction process for removal of aromatics and heteroatom containing molecules (e.g., sulfur, nitrogen, oxygen) to improve oxidative stability of lube base stocks.

Another important extraction process in refining is the removal of aromatic compounds from gasoline-blending stocks. This is done because the aromatics compounds have considerable value as chemical feedstocks and because environmental regulations greatly restrict the benzene content of gasoline. A number of commercial liquid–liquid extraction processes are employed in hundreds of commercial units that can accomplish this separation.

Often, the extraction is followed by extractive distillation. In these cases the solvent must not only preferentially extract aromatics, it must also lower the volatility of the aromatics so that they can be more easily separated from the solvent by distillation.

A major concern in solvent refining is the high energy consumption of these processes. Much of this is associated with recovery and recycling of solvent. The bulk of the process equipment in solvent refining is devoted to energy-efficient recovery and recycling of the solvent. Three energy-efficient solvent recovery processes are multi-effect evaporation, supercritical separation, and membrane separation.

In addition, efforts have been directed in recent years to catalytic processing as an alternative route particularly for removal of contaminants from lube base stocks. Although these processes are commercial and offer advantages over the solvent-refining processes for grass root applications, the advantages are usually not large enough to justify shutting down and replacing an operating solvent-refining unit. It is anticipated that solvent-refining processes will be important for separating undesirable components from hydrocarbon streams for many decades.

REFERENCES

1. King, C.J. *Separation Processes*; McGraw Hill: New York, 1980.
2. Hobson, G.D.; Pohl, W. *Modern Petroleum Technology*; Applied Science Publishers Ltd.: Essex, England, 1975.

3. Mair, B.J. *Chemical Technology of Petroleum*; Gruse, W.A., Stevens, D.R., Eds.; McGraw Hill: New York, 1960.
4. Treybal, R.E. *Liquid Extraction*; McGraw Hill: New York, 1963.
5. Banta, F.; Staffeld, P.O.; MacGuiness, M.P.; Wuest, R.G. Improved processes produce high quality lubes. 4th Annual Fuels and Lubes Asia Conference, Singapore, Jan 14–16, 1998.
6. Bhore, N.A. et al. (Mobil Oil Corp, Grace Davidson membranes). Oil Gas J. Nov 14, **1999**, *57*, 58.
7. *Hydrocarbon Processing*, Nov 2000, 87–142.
8. www.Bechtel.com/PDF/LubeBaseOil.pdf.
9. *Hydrocarbon Processing*, Nov 2000, 135.
10. *Hydrocarbon Processing*, Sep 1988, 89.
11. *Hydrocarbon Processing*, Nov 2000, 101–102.
12. www.UOP.com.
13. *Hydrocarbon Processing*, Nov 2000, 92.
14. *Perry's Chemical Engineering Handbook*, 6th Ed.; McGraw-Hill: New York, 1984; 11–31 and 11–40.
15. Winthrop, A.H. et al. (M.W. Kellogg), Advances in solvent deasphalting technology. National Petroleum Refining Association Annual Meeting, San Antonio, Texas, March 17–19, 1996, paper AM-96-55.

Solvents

Satish C. Mohapatra
Advanced Fluid Technologies, Inc., Dynalene Heat Transfer Fluids, Whitehall, Pennsylvania, U.S.A.

INTRODUCTION

Solvents are chemical compounds or mixtures that can dissolve other substances. In many applications, the substances need to be separated from the solvent after a process is over. Most of the solvents are in liquid form at room temperature and pressure. However, some fluids under certain temperature and pressure conditions can form ''supercritical solvents,'' which are neither liquid nor gas. These types of solvents are finding more applications in the recent times, because some of them are considered to be environmentally benign. Solvents are normally classified as follows:[1]

1. Nonpolar solvents (such as hexane and tetrachloromethane).
2. Solvents of low polarity (such as toluene and chloroform).
3. Aprotic dipolar solvents (such as acetone and *N,N*-dimethylformamide).
4. Protic and protogenic solvents (such as ethanol and nitromethane).
5. Basic solvents (such as pyridine and 1,2-diaminoethane).
6. Acidic solvents (such as 3-methylphenol and butanoic acid).

There are few other ways to classify fluids based on their dielectric constant, dipole moment, the hydrogen bond formation ability, acidity or basicity (relative to water) of the solvents, and extent of self-dissociation. Commercially available solvents can be obtained in several categories of purity, and the desired or required purity depends on their application. For example, a ''spectrophotometric'' grade solvent meets the requirement of not absorbing light at specified wavelength ranges. ''HPLC'' grade solvents are used in different liquid chromatography applications, where in addition to UV-transparency a low residue on evaporation is desired. Solvents used for ''electrochemical'' purposes should not contain any ionizable and electroactive, oxidizable or reducible impurities.

Another important aspect of solvents is their toxicity or their effect on humans and environment. Nowadays, the more hazardous solvents are being replaced with less hazardous ones. One example is the replacement of benzene (a known carcinogen) with toluene or xylene. Permissible exposure limit (PEL) has been determined for most of the widely used solvents. Immediate danger to life or health (IDLH) values have also been published for many toxic solvents (Table 1).[1,2] These values provide an indication of how toxic a solvent is when it evaporates into the atmosphere. Flammability of a solvent is determined from its flash point, autoignition temperature, and explosive limits (Table 1). The higher the flash point and autoignition temperature, the lesser is the flammability of the solvent. The toxic effect on ingestion is commonly reported in terms of LD_{50} in mg/kg body weight (Table 2).[2] This value indicates how much of the substance needs to be consumed into the body to cause mortality of 50% population.

TYPES OF SOLVENTS

In this entry, solvents are classified into four major categories, viz. organic solvents, aqueous solvents, supercritical solvents, and ionic liquids. This classification is based on the evolutionary approach to solvent design.

Organic Solvents

Organic solvents incorporate carbon and hydrogen in their molecule. Additionally, organic solvents can have oxygen, nitrogen, sulfur, chlorine, fluorine, phosphorous, and silicon atoms in the molecular structure. Incorporation of each of these atoms imparts some unique property to the solvent. For example, fluorination can make a resulting fluorinated hydrocarbon nonflammable and dense. Oxygenation can make a solvent more hydrophilic. Different categories of organic solvents do exist. They are hydrocarbons (aliphatic and aromatic), alcohols, ethers, ketones, carboxylic acids, esters, fluorocarbons, silanes and siloxanes, amines, amides, sulfides and sulfoxides, and phosphate esters.

Organic solvents are the most common solvents in a variety of applications ranging from cleaning to coating, and reaction and separation solvent.[1–5] These solvents

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008122

Table 1 Toxicity and flammability properties of solvents

Name	PEL	IDLH	Flash	Autoig.	Expl. lim.
n-Hexane	50	5000	−22	225	1.2–7.5
Benzene	1	2000	−11	560	1.3–7.1
Toluene	100	2000	4	480	1.3–7.3
o-Xylene	100	10,000	32	463	1.0–6.0
m-Xylene	100	10,000	29	465	1.1–7.0
p-Xylene	100	10,000	27	528	1.1–7.0
Ethylbenzene	100	2000	15	430	1.0–6.7
Cumene	50	44	424	none	0.9–6.5
Methanol	200	25,000	11	464	6.7–36
Ethanol	1000	13	423	none	3.3–19.0
n-Propanol	200	4000	25	371	2.1–13.5
i-Propanol	400	20,000	12	399	2.0–12.0
n-Octanol	50	35,000	68	300	1.2–9.3
1,2-Ethanediol	50	111	396	none	3.2–21.6
1,2-Propanediol	99	none	none	421	2.6–12.5
1,3-Propanediol	122			378	2.6–16.6
1,2-Butanediol	90			392	1.9–13.0
2,3-Butanediol (meso)	85			402	1.9–13.7
Diethyl ether	400		−45	160	1.8–36
Tetrahydrofuran	200		−17	224	1.8–11.8
Dioxane	25	200	12	180	2.0–22.2
Acetone	750	20,000	−18	465	2.6–12.8
Methyl-i-butyl ketone	50		17	448	1.4–7.5
Formic acid	5		69	480	18–57
Acetic acid	10		43	427	5.4–16.0
Propanoic acid			55	475	2.9–14.8
Methyl acetate	200		−10	502	3.1–16.0
Ethyl acetate	400	10,000	−4	427	2.2–11.0
Propyl acetate	200	8000	14	450	2.0–8.0
Chlorobenzene	75	2400	30	638	1.3–7.1
Chloroform	2	1000	none	none	none
1,1,1-Trichloroethane	350	1000	none	537	8.0–10.5
1,1,2-Trichloroethane	10	500	none	460	8.4–13.3
Trichloroethylene	50	1000	none	410	8.0–10.5
Morpholine	20	8000	38	310	1.8–10.8
Triethylamine	10	2000	8	232	1.2–8.0
Aniline	5		70	617	1.3–11.0
Diethanolamine	3		152	662	1.8–13.4
Acetonitrile	40	4000	6	524	4.4–16
Nitrobenzene	1		88	482	1.8
Formamide	20	none	175	none	7.0–29.3
Dimethylformamide	10	3500	58	445	2.2–15.2
Dimethylacetamide	10	none	63	354	1.8–13.8
Dimethyl sulfoxide	none	none	88	215	2.6–28.5
Ammonia	none	none	none	651	16–25

Units: (PEL) and IDLH in ppm; flashpoint and autoignition temperature in °C; explosive limits in volume %.
(From Refs.[1,2].)

Table 2 LD_{50} values for different solvents in mg/kg body weight

Solvent	LD_{50}	Solvent	LD_{50}
Benzene	4080	Ethylene carbonate	11,200
1-Propanol	1870	Triethyl phosphate	1370
1-Butanol	2680	1-Chlorobutane	5600
2-Chloroethanol	70	Chlorobenzene	3400
Phenol	340	Tetrachloromethane	5600
3-Methylphenol	828	1-Bromobutane	6700
Allyl alcohol	45	Butylamine	360
1,2-Propanediol	13,000	1,2-Diaminoethane	1850
1,5-Pentanediol	5900	Dibutylamine	770
Glycerol	19,300	Piperidine	11,000
Dibutyl ether	570	2-Methylpyridine	670
1,2-Dimethoxyethane	7000	Nitromethane	950
1,2-Dimethoxybenzene	1360	Propionitrile	40
Benzaldehyde	3260	Acrylonitrile	70
3-Pentanone	2140	Benzyl cyanide	350
Cyclohexanone	930	Benzonitrile	1400
Diisobutyl keyone	5750	Formamide	3100
Acetophenone	3000	*N,N*-Dimethylmacetamide	2580
Formic acid	1210	Tetramethylurea	1100
Acetic acid	3530	Sulfolane	2100
Propanoic acid	4290	Hexamethyl phosphoramide	2650

(From Ref.[2].)

are mostly inexpensive, their physical and chemical properties are well established,[1–4] and the process design is relatively simple. However, some of the popular organic solvents are very toxic, environmentally unfriendly, and could be carcinogenic. One example is the chlorinated compounds such as trichloroethylene, methylene chloride, and carbon tetrachloride. Because of this reason, chlorinated compounds and many other toxic solvents are slowly being phased out. These solvents are being replaced by more environmentally benign water-based solvents and supercritical solvents.

Aqueous Solvents

Aqueous solvents incorporate water as a major component in the solvent mixture. Water being a nontoxic, nonflammable, and inexpensive solvent, with excellent heat transfer properties is highly suitable for many of the applications. Water-based solvents are already used in certain types of paint and coating formulations. Water is also a major component in the household and industrial cleaning solvents.

The drawback of water as a solvent is that it cannot dissolve a variety of nonpolar compounds. In many aqueous solvent formulations, surfactants are used to emulsify the nonpolar components. One widely used example is emulsion polymerization for the manufacturing of latex. In this process, a nonpolar monomer such as styrene is emulsified with a surfactant first. Then, a free-radical initiator is added to the process, which starts the polymerization within the emulsion. This method creates colloidal polymer particles of very uniform shape and size that are used in a variety of applications.

Supercritical Solvents

A supercritical fluid is defined as a substance above its critical temperature (T_C) and critical pressure (P_C). The critical point represents the highest temperature and pressure at which the substance can exist as a vapor and liquid in equilibrium (Tables 3 and 4). Supercritical fluids are highly compressed fluids that combine properties of gases and liquids in a synergistic manner. Fluids such as supercritical xenon, ethane, and carbon dioxide (CO_2) offer a range of unusual chemical possibilities in both synthetic and analytical chemistry.

Supercritical CO_2 has been used for years to extract chemicals from foodstuffs. Commercial plants have

Table 3 Critical properties of supercritical fluids

Supercritical solvents	T_C(K)	P_C(MPa)	*dc* (g cm^{-3})
Nitrogen	126.20	3.40	0.157
Xenon	289.8	5.88	1.105
Methane	190.65	4.64	0.162
Ethylene	282.65	5.1	0.218
Ethane	308.15	4.88	0.203
Propane	369.95	4.26	0.219
n-Butane	425.15	3.80	0.228
Dichlorodifluoromethane	385.0	4.13	0.558
Carbon dioxide	304.20	7.39	0.468
Water	647.30	22.12	0.315
Dinitrogen oxide	309.60	4.26	0.450
Ammonia	405.55	11.4	0.236
Sulfur dioxide	430.35	7.87	0.525
Sulfur hexafluoride	318.70	3.76	0.736

(From Ref.[1].)

been scaled up for the extraction of caffeine from coffee and tea, as well as the extraction of biological products from herbs, spices, and vegetables, including soy beans, corn, and hops.[6]

In a typical process, liquid CO_2 is compressed and heated before being transferred to a vessel for extraction. Solute laden CO_2 is fed into a separator, where the pressure and temperature are adjusted to reduce the dissolving power of the CO_2, causing the extracts to be precipitated. Pressure adjustments can be performed in several stages, allowing different extracts to be isolated. Low density CO_2 is then fed to a refrigerated condenser and collected as liquid.

CO_2 is completely nonflammable. This property provides a tremendous advantage for some traditionally hazardous chemical processes and reactions. For example, fluoroethylene monomers used for the production of tetrafluoroethylene (TFE) are rendered nonexplosive when mixed with CO_2. In addition, highly reactive free-radical polymerization of these monomers can be carried out directly in a supercritical CO_2 continuous phase.

Widespread application of near-critical and supercritical CO_2 to process industries has been hindered up to now by a very weak solvent strength. In many applications where the physical and environmental properties of CO_2 would otherwise be favorable, the fact that CO_2 cannot dissolve or carry the required solute has blocked its application. One reason being the low dielectric constant (1.4–1.5) compared to the typical organic solvents (2.0–4.0). The very hydrophobic nature of CO_2 makes it a bad solvent for many solutes. Improving the solvent strength has been the focus of research in recent decades.

One approach to increasing CO_2 solvent strength relies on the fact that most volatile materials (such as alcohols, ketones, and hydrocarbons) are soluble in supercritical and near-critical CO_2. This allows the employment of a wide range of cosolvents to enhance the CO_2's solvation properties. The use of such cosolvent-modified CO_2 as a solvent medium is most recognized in the Unicarb spray-coating process commercialized by Union Carbide (now with Dow Chemical, Midland, MI) in the early 1990s. In this process, the majority of traditional solvents used in the spray coatings are replaced by supercritical CO_2. This process has been implemented in automotive and furniture industries.

Although the use of a cosolvent improves the solvation properties of CO_2, there are some drawbacks that were subsequently addressed. With the addition of these cosolvents, CO_2 loses its designation as a zero volatile organic carbon (VOC) solvent. Solvents are also not effective where heterogeneous processes are involved (such as dispersion polymerization). In these cases, a surfactant is required. Silicone and fluoroacrylate-based surfactants have been developed for stabilizing insoluble, dispersed polymer particles. Isolation of a dry powder is accomplished simply by depressurization. Although still in the early stages of implementation for dispersion processes, surfactant-modified CO_2 has a potential impact for almost all process technologies that rely on water or other solvent systems. In addition to the stabilization of polymer dispersions and latexes, surfactant-modified CO_2 is expanding into traditional-hydrocarbon-solvent, halogenated-solvent and aqueous-solvent cleaning processes.

CO_2-based solvent technology cannot be implemented in normal process equipment. More commonly, implementation requires installation of pressurized equipment in place of existing lines that operate at ambient conditions. Guidelines for design, operation, and maintenance of these systems are established by the International Standards Organization (ISO, Geneva), American Society of Mechanical Engineers (ASME, New York), American National Standards Institute (New York), the U.S. Occupational Safety and Health Administration (OSHA, Washington, D.C.), and CPI operating codes.

Another challenge that the high-pressure process puts on utilization of near-critical and supercritical CO_2 is a push toward batchwise processing methodologies. Batch pressurization and depressurization cycles are typical because ideally both the end product and the feed streams will be at ambient conditions. For example, decaffeination of raw coffee beans is usually done in a semibatch process. Beans are loaded in and out of the pressurized CO_2 solvent with lock-hopper type feed and discharge valves. In this design, engineers must consider the sealing methods for locks and chamber doors, as well as the potential cyclic load on

Table 4 Melting point, boiling point, critical temperature and density of solvents

Name	T_m	T_b	T_c	d
n-Hexane	177.8	341.9	507.7	0.6549
Benzene	278.7	353.2	562.2	0.8690
Toluene	178.2	383.8	591.8	0.8619
o-Xylene	248.0	417.6	630.3	0.8760
m-Xylene	225.3	412.3	617.1	0.8604
p-Xylene	286.4	411.5	616.2	0.8569
Ethylbenzene	178.2	409.3	617.2	0.8625
Cumene	177.1	425.6	631.1	0.8573
Methanol	175.5	337.7	512.6	0.7872
Ethanol	158.7	351.4	513.9	0.7848
n-Propanol	147.0	370.3	537.3	0.8003
i-Propanol	185.2	355.4	508.3	0.7815
n-Octanol	258.2	468.3	652.5	0.8222
1,2-Ethanediol	260.6	470.7	645.0	1.1104
1,2-Propanediol	213.0	460.8	626.0	1.0326
1,3-Propanediol	246.5	487.6	658.0	1.0497
1,2-Butanediol	220.0	467.0	680.0	0.9991
2,3-Butanediol (*meso*)	280.8	449.9	611.0	0.9871
Diethyl ether	156.9	307.6	466.7	0.7079
Tetrahydrofuran	164.8	339.1	540.2	0.8837
1,4-Dioxane	285.0	347.5	587.0	1.0281
Acetone	178.5	329.2	508.1	0.7849
Methyl-i-butyl ketone	189.0	390.6	571.5	0.7968
Formic acid	281.4	373.7	580.0	1.2145
Acetic acid	289.8	391.0	592.7	1.0443
Propanoic acid	252.5	414.3	612.7	0.9877
Methyl acetate	175.1	330.0	506.6	0.9283
Ethyl acetate	189.6	350.3	523.3	0.8945
Propyl acetate	178.2	374.7	549.7	0.8827
Chlorobenzene	227.6	404.8	632.4	1.1014
Chloroform	209.6	334.3	536.6	1.4793
1,1,1-Trichloroethane	242.8	347.2	545.0	1.3301
1,1,2-Trichloroethane	236.6	387.0	612.0	1.4314
Trichloroethylene	186.8	360.3	571.0	1.4599
Morpholine	268.4	402.1	618.0	0.9957
Triethylamine	158.5	362.0	535.4	0.7228
Aniline	267.2	457.6	699.0	1.0178
Diethanolamine	301.1	541.5	715.3	1.0947
Acetonitrile	229.3	354.8	545.5	0.7760
Nitrobenzene	278.9	484.0	732.0	1.1987
Formamide	275.7	483.7	771.0	1.1288
Dimethylformamide	212.7	426.2	596.6	0.9433
Dimethylacetamide	253.0	439.3	637.0	0.9368
Dimethyl sulfoxide	291.7	462.2	729.0	1.0958
Ammonia	195.4	239.7	405.7	0.6812

(From Refs.[1,2].)

chambers or systems that shuttle materials in or out of the process.

Ionic Liquids

This class of materials is ionic compounds that can remain in liquid state at room temperature. The ions in these compounds are, in general, large and bulky with the cations usually being organic and asymmetrical.[7,8] The forces holding such unwieldy components together are weak; hence, the crystal lattice tends to fall apart at relatively low temperatures. As a result, the salts form colorless fluids at temperatures ranging from 96 to 200°C. Additionally, their melting point, viscosity, density, and miscibility can be varied by changing the structure of the ions or, in binary systems, their relative ratios.

Theoretically, a trillion or more ionic liquids are possible. To make an ionic liquid, researchers can select it from dozens of small, negatively charged ions, such as hexafluorophosphate ($[PF_6]$-) and tetrafluoroborate ($[BF_4]$-), and hundreds of thousands of larger, positively charged ions, such as 1-hexyl-3-methylimidazolium or 1-butyl-3methylimidazolium.

Unlike typical organic solvents, ionic liquids tend not to give off vapors; hence they are less hazardous and more convenient in the laboratory, and are less likely to pose air pollution problems. Products and chemical catalysts can be recovered or extracted easily from ionic liquids and then the liquid can be recycled and used over and over again. Reactions that occur in organic solvents have been the standard way to make countless products. Now, many of these old reactions are carried out in these new solvents.

SOLVENT MIXTURES

Water is one of the most widely used solvents because of its availability, low cost, nontoxicity, and safety as well as its ability to dissolve a wide variety of substances. Sometimes when the solubility or any other property of water does not allow it to be used, a polar or a nonpolar organic solvent can be used. In certain applications, neat organic solvents fall short of the mark as far as their dissolving power or other properties are concerned. It is then necessary to use solvent mixtures. These mixtures can be binary (two components), tertiary (three components), or a multicomponent mixture. Many times, one of the components may be a solid. One very common example of this can be found in liquid chromatography where an electrolyte solution (buffer + salt) is used in many applications as a mobile phase.

In aqueous mixtures, the solvents that are completely miscible with water may be called cosolvents. The complete miscibility of solvents with water depends on their hydrophilic character, expressed as the log P. This is the logarithm of the partition coefficient of the substance considered as a solute between 1-octanol and water. Normally, the solvents with negative log P values are completely miscible with water, and solvents with $\log P > 1.3$ show very limited miscibility with water. Solvents with intermediate values ($0 < \log P < 1.3$) are somewhat miscible with water.

Mixtures of nonpolar solvents are normally characterized by the term ''solubility parameter'' (δ). The difference in solubility parameters of mixture components provides a measure of solution nonideality.[3] Mixtures of aliphatic hydrocarbons are nearly ideal, whereas mixtures of aliphatic hydrocarbon with aromatics show appreciable nonideality. Sometimes, it is difficult to predict the behavior of highly nonideal mixtures. Thermodynamic properties of binary and multicomponent mixtures have been dealt with extensively in the literature.[3,4]

PROPERTIES OF SOLVENTS

The proper choice of a solvent for a particular application depends on several factors, among which its physical properties are of prime importance. In most applications, the solvent is preferred to be in its liquid state under the temperature and pressure conditions at which it is been employed. Other properties, such as density, vapor pressure, heat capacity, surface tension, and transport properties, are also very important. Electrical, magnetic, and optical properties are, in some cases, relevant to the application. All these are physical properties of the solvents.

Chemical properties of solvents, like the physical properties, are very important. Solubility parameters or the solvency factors provides an idea of how good a solvent is to solubilize certain types of materials. Polarity, acid–base properties, hydrogen bonding ability, and water miscibility are important chemical properties of solvents and are somewhat related.

Physical Properties

Under ambient conditions, solvents are normally liquid, unless a supercritical solvent is considered. The freezing/melting points and boiling points of some widely used solvents are listed in Table 4.[1,2] A solvent could form solids if it is stored outside during winter, and the temperature drops below its freezing point. Several freeze thaw cycles can damage the performance of paints and coatings. Solvents with low boiling points

have high vapor pressure, which leads to a quick drying or separation time. But sometimes, slow drying is necessary for coating applications where a solvent with high boiling point may be utilized.

Specific heat capacity is another important property of solvents. This is defined by the energy required to raise one unit mass of the solvent by one degree. This term becomes important if a solvent is used in a reaction process, which is either exothermic or endothermic. Solvents with high specific heat capacity can absorb more heat per unit temperature rise, compared to those with low specific heat capacity.

The response of a solvent to an electric field depends on the intrinsic dipole moment of its molecules, and also on cooperative effects of adjacent dipoles, when these are correlated in the liquid. The dipole moment μ is the measure of the separation of the positive and negative centers of charge in the molecule, and is measured best for the solvent vapor, where such cooperative effects are absent.[1] When this is impractical because of low volatility of the solvent, then the dipole moment may be measured for a dilute solution with the solvent being the solute in an inert solvent. The inert solvent is normally chosen from a group of solvents with very low polarity, such as benzene, hexane, and tetrachloromethane. Solvents with highly symmetrical molecules have zero dipole moments, but electronegative atoms connected to aliphatic or aromatic skeletons cause the molecules to have finite dipole moments. Table 5 lists dipole moments for some common solvents.[1,2] It also lists dielectric constants (ε) for these solvents. The higher is the dielectric constant, the more polar is the solvent and vice versa. Low dielectric constant fluids, also known as ''dielectric fluid,'' are used in electronics industry as cleaning solvents for circuit boards. Examples are fluorocarbons (because of their nonflammability), high flash point hydrocarbons, and silicones.

Surface and transport properties of solvents are very important for solvents. Surface tension of a solvent shows how easy or difficult it would be to wet the surface on which the solvent is being applied. Low surface tension implies better wetting ability and vice versa. Water and other polar organic solvents have very high surface tension, whereas silicones, fluorocarbons, and aliphatic hydrocarbons have low surface tension. Solvents with low surface tension are easier to leak through threaded joints compared to those with high surface tension.

Viscosity is another important property of solvents. High viscosity implies high power requirement for the flow of the solvent. Low viscosity means the solvent will flow easily; however, it may not be desirable in many applications. For example, a paint may need a solvent with optimum viscosity, which will neither drip out of the brush easily nor will stick to the brush so

Table 5 Dipole moment (μ) and dielectric constants (ε) of common solvents

Name	μ	ε
n-Hexane	0.09	1.88
Benzene	0.00	2.27
Toluene	0.31	2.38
o-Xylene	0.45	2.57
m-Xylene	0.30	2.37
p-Xylene	0.00	2.27
Ethylbenzene	0.37	2.40
Cumene	0.39	2.38
Methanol	2.87	32.66
Ethanol	1.66	24.55
n-Propanol	3.09	20.45
i-Propanol	1.66	19.92
n-Octanol	1.76	10.34
1,2-Ethanediol	2.31	37.70
1,2-Propanediol	2.25	32.00
1,3-Propanediol	2.55	35.00
1,2-Butanediol	2.18	—
2,3-Butanediol (meso)	2.1	21.53
Diethyl ether	1.15	4.20
Tetrahydrofuran	1.75	7.58
Dioxane	0.45	2.21
Acetone	2.69	20.56
Methyl-i-butyl ketone	2.77	15.87
Formic acid	1.82	58.50
Acetic acid	1.68	6.15
Propanoic acid	1.68	3.37
Methyl acetate	1.68	6.68
Ethyl acetate	1.78	6.02
Propyl acetate	1.78	6.00
Chlorobenzene	1.69	5.62
Chloroform	1.15	4.89
1,1,1-Trichloroethane	1.70	7.25
1,1,2-Trichloroethane	1.55	7.29
Trichloroethylene	0.80	3.42
Morpholine	1.56	7.42
Triethylamine	0.66	2.42
Aniline	1.51	6.98
Diethanolamine	2.81	25.19
Acetonitrile	3.92	35.94
Nitrobenzene	4.22	37.78
Formamide	3.37	109.50
Dimethylformamide	3.82	36.71
Dimethylacetamide	3.72	37.78
Dimethyl sulfoxide	4.06	46.45
Ammonia	1.47	22.38

μ is Debey unit and ε is dimensionless.
(From Refs.[1,2].)

hard that it will be difficult to apply on a surface. Both surface tension and viscosity are temperature dependent. In formulations involving solvents, the viscosity behavior of the final formulation may be quite different from the solvent itself. Many times, solvent-based formulations such as paints exhibit non-Newtonian behavior with the viscosity being a strong function of shear stress. However, most of the pure solvents exhibit Newtonian behavior.

Chemical Properties

Chemical properties of solvents affect their usefulness in various applications. The solvent should selectively dissolve the desired solutes, should be inactive/inert in the chemical reactions, and solvate the transition states and products really well. This can be achieved by the proper blend of chemical properties such as solvency, polarity, hydrogen bond donation or acceptance ability, acidity or basicity, hydrophilicity, and redox properties.

A particularly common test for ranking hydrocarbon solvent strength is the kauri-butanol test. The kauri-butanol value (KB) of a solvent represents the maximum amount of the solvent that can be added to a stock solution of kauri resin (a fossil copal) in butyl alcohol without causing cloudiness. Because kauri resin is readily soluble in butyl alcohol but not in hydrocarbon solvents, the resin solution will tolerate only a certain amount of dilution. ''Stronger'' solvents such as toluene can be added in a greater amount (and thus have a higher KB value) than ''weaker'' solvents like hexane.

Another common test for solvency is called aniline point test. The aniline point is called the ''aniline point temperature,'' which is the lowest temperature (°F or°C) at which equal volumes of aniline ($C_6H_5NH_2$) and the oil form a single phase. The aniline point (AP) correlates roughly with the amount and the type of aromatic hydrocarbons in an oil sample. A low AP is indicative of higher aromatics, while a high AP is indicative of lower aromatics content. Diesel oil with AP below 120°F (49°C) is probably risky to use in oil-base mud. In general, the lower the aniline point, the more the number of unsaturants that are present and the higher the potential for swelling certain rubber compounds. The American Petroleum Institute has developed test procedures that are the standard for the industry.

Solvent polarity is related to the dipole moment as discussed in the earlier section. A solvent without a permanent dipole must be classified as nonpolar. However, a solvent may exhibit local polarity, if it possesses two mutually canceling dipoles. One such example is 1,4-dioxane, where the two oxygen atoms can participate in electron-pair donation to nearby acceptor atoms in solutes, although the molecule as a whole does not have a permanent dipole moment. Furthermore, highly polarizable molecules may interact via induced dipoles, so that polarizability may contribute to the chemical aspect of polarity.

The acidity and basicity properties of solvents play an important role in various processes, especially in chemical reactions. A solvent may accept or donate a proton, thereby increasing or decreasing the rate of a chemical reaction. This behavior is similar to that of a catalyst. Sometimes, a solvent mixture may be a better alternative than the single components.

Aqueous solubility of solvents is related to polarity and dipole moment. Many solvents are hygroscopic and need drying agents, such as molecular sieves to remove moisture from them. Solubility of solvents in water and vice versa are listed in Table 6. It is observed that the solvents with –OH, –C(O)Me, –COOH, –NH2, and –CN groups have high degree of miscibility with water, whereas hydrocarbons, esters, and chlorinated compounds have a low solubility in water.

As discussed earlier, solubility parameter is important when nonpolar solvents are mixed. Table 7 provides a list of molar liquid volume and solubility parameter for some common solvents.[3] Aromatics have a higher value of this parameter compared to the aliphatics.

APPLICATIONS OF SOLVENTS

Major applications of solvents are found in paints/coating and cleaning applications. However, they are also widely used in chemical processes (as solvent as well as raw material), inks, solvent extraction, heat transfer systems, and electrochemistry. Some solvents are also used as catalysts in a chemical reaction.

Paints/Coatings

Paints and coatings consume the largest amount of solvent among all the applications.[5] Because of regulatory concerns, the consumption of organic solvents in coating applications is being reduced and replaced with water-based as well as nonsolvent coatings (powder coating). Use of green solvents is being encouraged in recent years.

Solvents in paint or a coating serve multiple purposes. These include solubilization of resins and other ingredients, wetting, viscosity reduction, adhesion promotion, and gloss enhancement. Initially, the resin or polymer is dissolved in the solvent to form a continuous phase.

Table 6 Water miscibility of different solvents

Name	In water	Water in	log *P* O/W
n-Hexane	2.57e-6	5.31e-4	3.90
Benzene	4.13e-4	2.75e-3	2.13
Toluene	1.01e-4	171e-3	2.69
o-Xylene	2.97e-5	2.53e-3	3.12
m-Xylene	2.48e5	2.36e-3	3.20
p-Xylene	2.65e-5	2.68e-3	3.15
Ethylbenzene	2.58e-5	2.53e-3	3.15
Cumene	9.79e-6	2.01e-3	3.66
Methanol	miscible	miscible	−0.70
Ethanol	miscible	miscible	−0.25
n-Propanol	miscible	miscible	0.28
i-Propanol	miscible	miscible	0.13
n-Octanol	4.4e-5	2.75e-1	3.15
1,2-Ethanediol	miscible	miscible	−2.27
1,2-Propanediol	miscible	miscible	−1.41
1,3-Propanediol	miscible	miscible	—
1,2-Butanediol	—	—	—
2,3-Butanediol (meso)	miscible	miscible	−0.92
Diethyl ether	1.54e-2	5.76e-2	0.89
Tetrahydrofuran	miscible	miscible	0.46
1,4-Dioxane	miscible	miscible	−0.42
Acetone	miscible	miscible	−0.24
Methyl-i-butyl ketone	3.10e-3	9.72e-2	1.31
Formic acid	miscible	miscible	−0.54
Acetic acid	miscible	miscible	−0.24
Propanoic acid	miscible	miscible	0.32
Methyl acetate	7.31e-2	2.69e-1	0.18
Ethyl acetate	1.77e-2	1.29e-1	0.73
Propyl acetate	4.1e-3	1.45e-1	1.24
Chlorobenzene	7.83e-5	2.03e-03	2.84
Chloroform	1.24e-3	6.1e-3	1.94
1,1,1-Trichloroethane	1.78e-4	2.51e-3	2.36
1,1,2-Trichloroethane	5.96e-4	8.67e-3	1.89
Trichloroethylene	1.88e-4	2.29e-2	2.35
Morpholine	miscible	miscible	−1.08
Triethylamine	1.03e-2	2.13e-1	1.36
Aniline	6.72e-3	2.05e-1	0.90
Diethanolamine	7.8e-1	miscible	−1.43
Acetonitrile	miscible	miscible	−0.34
Nitrobenzene	2.78e-4	1.62e-2	1.85
Formamide	miscible	miscible	−0.97
Dimethylformamide	miscible	miscible	−1.01
Dimethylacetamide	miscible	miscible	−0.77
Dimethyl sulfoxide	miscible	miscible	−1.35
Ammonia	miscible	miscible	−1.49

All the quantities are dimensionless.
(From Refs.[1,2].)

Table 7 Molar volume and Hildebrand solubility parameter for solvents

Solvents	ν ($cm^3\ mol^{-1}$)	δ ($J\ cm^{-3})^{1/2}$
Perfluoro-*n*-heptane	226	12.3
Neopentane	122	12.7
Isopentane	117	13.9
n-Pentane	116	14.5
n-Hexane	132	14.9
1-Hexene	126	14.9
n-Octane	164	15.3
n-Hexadecane	294	16.3
Cyclohexane	109	16.8
Carbon tetrachloride	97	17.6
Ethyl benzene	123	18.0
Toluene	107	18.2
Benzene	89	18.8
Styrene	116	19.0
Tetrachloroethylene	103	19.0
Carbon disulfide	61	20.5
Bromine	51	23.5

(From Ref.[3].)

Dyes and pigments are then added to this solution to provide color. When a coating is applied, the solvent starts to evaporate leaving a thin continuous film on the surface. The solvent can also increase adhesion by softening the primer coat. In most of the coating applications, a solvent blend is used instead of a single solvent. Thus, each solvent component serves a particular purpose.

In water-based paints, the dominant solvent is water. There may be a small amount of other solvents to carry out certain functions. There are two types of water-based paints—one with latexes (composed of fine polymer particles dispersed in the solvent) and the other with water-soluble polymers. Latex paints are very common in the architectural market with flat, semigloss, and gloss coatings. More and more industrial coating applications are switching to water-based paints for environmental reasons.

Cleaning

The dry cleaning industry is one of the largest consumers of solvents. Perchloroethylene (PCE) is used in as high as 80% of the dry cleaning applications. Although this solvent is nonflammable, easy to recover, and exhibits good solvency, its environmental impact is severe. There is work underway to reduce the emission of this solvent from the dry cleaning plants. At the same time, many plants have switched to hydrocarbon-based solvents.

Another cleaning application involving solvents is the degreasing of metal parts and other objects in manufacturing plants and automotive repair facilities. Mineral oils and other high flash point hydrocarbon solvents are used in these applications.

Reaction Solvents

Chemical reactions normally require a solvent that can dissolve all the raw materials. Reaction kinetics and selectivity may be affected by the choice of solvent. Thermophysical properties of solvents are also very important for endothermic and exothermic reactions. Water, hydrocarbons, alcohols, ketones, chlorinated solvents, and amines are the most widely used solvents in the chemical process industry. In many applications, the solvent is consumed as a raw material. One example is the manufacture of esters, where alcohol takes part in the reaction.

Solvent Extraction

Solvent extraction is widely used in pharmaceutical and food processing industries. Oil seed extraction, manufacturing of neutraceuticals, decaffeinated coffee, intermediates, and some reactive-separation processes utilize solvent extraction. Hydrocarbons are common solvents for oil seed extraction. Supercritical solvents are gaining popularity in producing neutraceuticals and other active ingredients.

CONCLUSIONS

Innumerable production processes and applications rely on the proper selection of solvents. Organic solvents, although most widely used, are being slowly replaced with aqueous and supercritical solvents. Ionic liquids are still in the research stage and commercial applications are just beginning to appear. Solvent mixtures are used in many applications where a single solvent cannot do the job.

Both physical and chemical properties of the solvents are important. In addition, toxicity, flammability, and environmental friendliness should not be overlooked while selecting a solvent. All the properties are assessed to design the most optimum solvent for a particular need. The most widely used applications for solvents are paints/coatings, cleaning, reaction medium, and solvent extraction.

REFERENCES

1. Marcus, Y. *The Properties of Solvents*; John Wiley & Sons Ltd: New York, 1999.
2. Riddick, J.A.; Bunger, W.B.; Sakano, T.K. *Organic Solvents*; Wiley-Interscience: New York, 1986.
3. Prausnitz, J.M.; Lichtenthaler, R.N; de Azevedo, E.G *Molecular Thermodynamics of Fluid-phase Equilibria*; Prentice Hall: Upper Saddle River, NJ, 1999.
4. Marcus, Y. *Solvent Mixtures*; Marcel Dekker, Inc: New York, 2002.
5. Sullivan, D.A. Solvents, industrial. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 3rd Ed.; Wiley: New York, 1980; Vol. 22, 529–571.
6. McClain, J. Processing with supercritical solvents. Chemical Engineering **2000**, Feb.
7. Crabb, C. Exploring ionic liquids. Chemical Engineering **2001**, Mar.
8. Gorman, J. Faster, better, cleaner? Science News **2001**, *160* (10), Sept.

Sonochemical Reaction Engineering

David A. Bruce
Amarnath Nareddy
Department of Chemical and Biomolecular Engineering, Clemson University, Clemson, South Carolina, U.S.A.

INTRODUCTION

Sonochemical reaction processes, which are relatively new, use the transmission of ultrasonic waves (20 kHz to 1 MHz) through liquid media to initiate chemical transformations and enhance mass transfer. These processes are similar to other chemical treatment methods, such as those employing plasma, flame, and thermo chemistries, in that a large amount of energy is introduced to the material in a very short period of time. However, the highly reactive zones in a sonochemical process are micrometer sized and the bulk fluid, which is generally at ambient conditions, is not exposed to these harsh conditions for extended periods. The chemical effect of these sound waves is derived from the formation, growth, and sudden collapse of micrometer size bubbles via a process known as acoustic cavitation. The microbubbles formed during cavitation are slowly filled with vaporized liquids until they reach a critical size, whereupon they undergo violent collapse in less than a microsecond. This rapid implosion causes the gases inside the bubble to be adiabatically compressed, which leads to significant increases in temperature and pressure inside the cavity. In fact, the temperature inside a bubble during implosion can be as high as 5000 K and the pressure greater than 500 atm, which equates to heating rates of more than a billion degrees per second during bubble collapse. Such extreme conditions lead to the formation of highly reactive species that can readily react with chemicals in the surrounding liquid phase. Cavitation phenomena can also produce reactive radicals and ions via localized electrical discharges that result from positively and negatively charged ions becoming separated during microbubble oscillation and collapse. Interestingly, this entire reaction process occurs repeatedly in a fluid medium whose temperature and pressure are only slightly above ambient conditions. Thus, ultrasound irradiation can significantly improve reactions rates, while simultaneously allowing the overall process to operate at milder reaction conditions.

The unique reaction environment generated in sonochemical reactors has been shown to greatly enhance reaction rates for a variety of chemical transformations. However, the commercial scale application of sonochemical technologies is still limited. It is only recently that advances in electromechanical transducers have enabled high powered ultrasonic waves to be efficiently generated in industrial scale reactors. To date, sonication has been used to fabricate nanomaterials, epoxidize and oxidize unsaturated hydrocarbons, couple halogenated aromatics, and cause a variety of degradation reactions as well as a number of other organic and inorganic chemical reactions. Many of the reactive species formed by acoustic cavitation are free radicals; hence, reaction pathways that include the formation of free radicals are often accelerated by sonochemical processing. However, there are many factors, some of which are still not fully understood, which determine whether a given reaction process will be accelerated by ultrasonic irradiation. Therefore, much of the research and process development in this field is still Edisonian in nature.

HISTORY

Though sonochemistry is a relatively new field of study, cavitation phenomena have been of interest for over 200 years. Leonhard Euler first mentioned the effects of cavitation in 1754 and the first article on cavitation by Thornycroft and Barnaby was printed in 1895.[1] However, it was not until the work of Lord Rayleigh in 1917 and Loomis and others in 1927 that the physical, chemical, and biological effects of cavitation were fully described.[2–4] Later studies by Harvey et al. defined the underlying mass transfer phenomena controlling bubble growth (i.e., rectified diffusion), and shortly thereafter the first computer simulations of a cavitating bubble were reported by Noltingk and Neppiras.[5,6] A decade later, Naude and Ellis suggested that bubbles collapsing near solid surfaces undergo asymmetric implosions that lead to the formation of microjets (see Fig. 1).[7] These microjets are believed to cause solid surface pitting and can readily explain the microscale surface erosion that is observed with dispersed particulates (e.g., oxide catalysts), which leads to their decreasing in size during sonication. Since the 1980s, numerous studies have provided insights on the physical environment (temperature

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039788

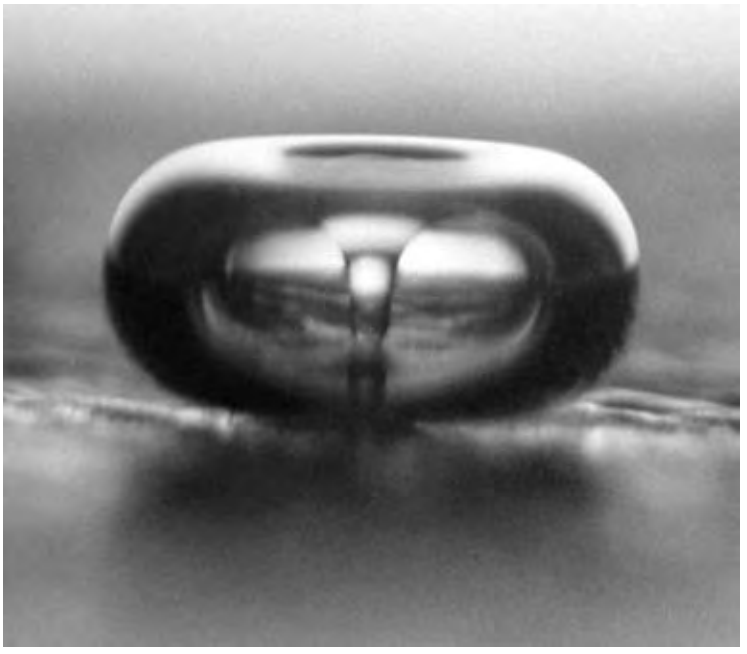


Fig. 1 Ultrasound induced collapse of a microbubble on a surface. (Courtesy of Crum, L.A. J. Physique Colloque **1979**, *40*, 285–288.)

and pressure) inside an imploding microbubble, as well as information about the types of reactive species formed in this environment (e.g., free radicals such as $H^{\bullet}$ and $OH^{\bullet}$).[8–10] During the same period, the production of low cost sonication equipment has led chemists to discover a wide array of chemical reactions whose rate is greatly enhanced or whose product distribution is significantly altered by the cavitation process.[11–16]

GENERATION OF ULTRASOUND WAVES

Ultrasound is sound energy with a frequency beyond human hearing. It is a form of acoustic radiation with a frequency range between 20 kHz and 500 MHz. Depending on the frequency, it is divided into three regions: power ultrasound (20–100 kHz), high frequency ultrasound (100 kHz–1 MHz), and diagnostic ultrasound (1–500 MHz). Diagnostic ultrasound has a low intensity and has long been used in the medical field for therapeutic, operative, and diagnostic procedures. Power and high frequency ultrasound, commonly called destructive ultrasound, are used in a variety of industries. Common industrial applications include: welding plastics, cleaning and decontamination, beneficiation of coal, soldering, and processing that includes emulsification, extraction, crystallization, and filtration.

In sonochemistry, physical and chemical changes are brought about by the usage of power ultrasound. Three types of transducers are used to generate power ultrasound: gas driven transducers, liquid driven transducers, and electromechanical transducers. Of these, electromechanical transducers are the most widely used and include piezoelectric transducers and magnetostrictive transducers.[12,17] Piezoelectric transducers convert electric energy into kinetic energy or sound waves. When piezoelectric materials, such as quartz or $BaTiO_3$, are subjected to an alternating electric field, corresponding dimensional changes in the crystal generate high energy mechanical vibrations (sound energy). These transducers are highly efficient and can be used over the entire range of ultrasonic frequencies, but they are generally restricted to small volume processes for short terms and often operate at a fixed frequency; hence, they are more commonly used for laboratory scale experiments or processes. The magnetostrictive transducers, on the other hand, are used when large volumes are to be processed and when the reaction times are long. The magnetostrictive transducers utilize the Joule effect in which a ferromagnetic material (e.g., nickel or iron) in an external magnetic field alternately expands and contracts producing sound waves. These transducers are very robust, have high driving force, and can be operated at elevated temperatures (250°C) for extended periods of time. These characteristics make the magnetostrictive transducers a good choice for largescale industrial applications.

ACOUSTIC CAVITATION

Acoustic cavitation is a nonlinear process that effectively concentrates the diffuse energy of sound in

liquids. The microbubbles formed during cavitation result from sound wave generated compression and rarefaction cycles, which cause the local fluid density to momentarily increase or decrease, respectively. When the local density decreases during a rarefaction cycle to a sufficiently low value, such that the spacing between molecules exceeds the cohesive forces in the liquid, voids or cavities are formed. These voids, during the subsequent compression and rarefaction cycles, grow because of inertial effects and the rectified diffusion of mass from the bulk liquid. The growing cavity eventually reaches a critical size, where the combination of internal pressure and surface tension effects no longer balance the compression forces generated by the next sound wave. At this instant, the bubble implodes. This bubble collapse leads to a highly reactive environment, localized in a micrometer sized region in the bulk fluid. To gain a better understanding, the three discrete stages of cavitation, viz., nucleation, bubble growth, and implosive collapse, are described below with respect to homogenous (liquid) sonochemical reaction systems.

Nucleation

Nucleation refers to the formation of bubbles or cavities and is the origin of all cavitation processes. These cavities are formed when the acoustic energy applied to the liquid is sufficient to overcome the attractive forces binding neighboring liquid molecules. This energy, which equates to the tensile strength of the liquid is of the order of 1000 atm for pure water. However, the presence of dissolved gases and suspended particulates can reduce this tensile strength to the order of 1 atm. These gas pockets serve as nucleation sites for ultrasound cavitation bubbles and are often called ''weak spots.'' Ultrasound transmitted through a liquid consists of alternate compression and rarefaction cycles. When the acoustic amplitude during the rarefaction cycle is sufficiently large and exceeds the average intermolecular bonding forces in the liquid, neighboring liquid molecules are pulled apart forming voids or cavities.[18] Depending on the vapor pressure of dissolved organics and inorganics, these voids can be rapidly filled with vapor content and form bubbles. Hence, bubble formation and the chemical effects of ultrasound can only be seen when there are impurities, such as dissolved gases, in the liquid and when the sound intensity exceeds the cavitation threshold (i.e., the surface tension of the liquid near weak spots).[12] Fig. 2 provides an illustration of the relationship between the acoustic pressure (which varies) and bubble creation, growth, and collapse processes.

Bubble Growth and Dynamics

Bubble growth is very complex, in that it is highly nonlinear with time and also different for each of the millions of bubbles formed in the liquid. Fundamentally, the gas bubbles generated almost instantaneously during nucleation are inherently unstable.[19] In the absence of further acoustic irradiation, large bubbles will float to the surface and burst, while small bubbles will slowly dissolve owing to the excess internal gas pressure. In the presence of ultrasound, a bubble can grow many times larger than its initial size over a period of time much longer than bubble nucleation or collapse. Depending on the acoustic intensity and frequency, these bubbles will form a stable cavity and oscillate around a mean radius for several acoustic cycles or they will continue to grow until they reach a critical size and collapse. In general, this bubble growth process occurs at sufficiently slow rates so that it can be considered isothermal; hence, extremely little or no reaction is thought to occur during the growth phase of the cavitation process.

Bubble growth with high intensity ultrasound arises primarily from inertial effects. At this high intensity,

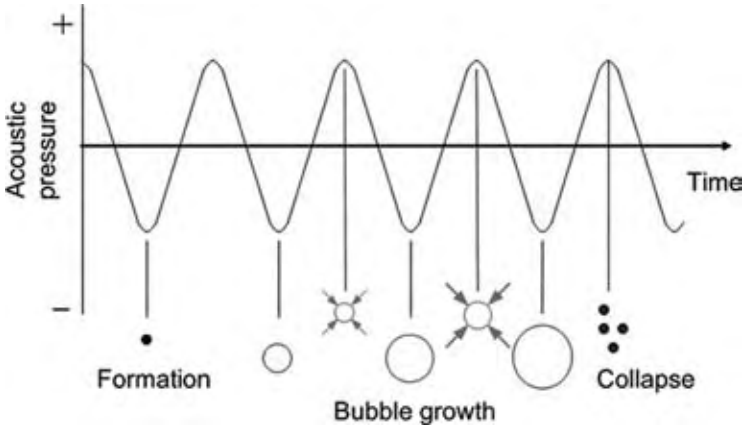


Fig. 2 The cavitation process: growth and collapse of a microbubble as a function of time and ultrasound frequency. *(View this art in color at www.dekker.com.)*

bubbles gain so much momentum during the negative pressure cycle that they will have no time to recompress during the positive pressure cycle and grow rapidly in the course of a single cycle of sound.[19] These bubbles, no longer in phase with the ultrasonic field, become unstable and implode in the subsequent compression cycle.

At low ultrasound intensities, bubble growth primarily occurs via rectified diffusion, which is the unequal mass transfer of species into the bubble during rarefaction and compression cycles. This phenomenon was first recognized by Harvey et al. during their experiments on animals, and Leighton has recently expounded on a well established theory that describes rectified diffusion in terms of an ''area effect'' and a ''shell effect.''[5,20] These two effects derive directly from basic mass transfer principles, which demonstrate that the rate of mass transfer is directly related to the surface area across which transfer can occur and the concentration (or more exactly, chemical potential) driving force.

The ''area effect'' can be described by the following. During rarefaction, when the bubble expands, concentrations and pressure inside the bubble decrease, and so dissolved gases and other volatile species diffuse from the liquid into the bubble. However, bubble compression yields high pressure and concentrations inside the bubble, which causes gas to diffuse from the bubble's interior into the bulk liquid. Because the surface area of the bubble is greater during the expansion phase than in the compression phase, the amount of gas diffused is greater during the expansion cycle. Therefore, over a complete cycle, there will be a net inflow of gas into the bubble causing it to grow over time, as depicted in Fig. 2.

Similarly, the ''shell effect'' arises because the diffusion rates of dissolved volatile species are proportional to their liquid phase concentration gradients. Consider a thin spherical shell of liquid surrounding an acoustically generated bubble. This liquid shell will change volume as the bubble pulsates, which in turn causes the concentration gradients of dissolved species to change, as shown in Fig. 3. Note that in Fig. 3, x_i and y_i equal the bulk concentration of the diffusing species i in the liquid shell and gas phases, $x_{i,\text{int}}$ and $y_{i,\text{int}}$ equal the concentration of species i at the gas–liquid interface, x_i^* equals the concentration of species i that would exist in the liquid phase if it were in equilibrium with a gas having concentration y_i, and y_i^* equals the concentration of species i that would exist in the gas phase if it were in equilibrium with a liquid having concentration x_i. At equilibrium conditions, there is no concentration gradient in the liquid surrounding the bubble; hence, no net mass transfer occurs. During bubble expansion (rarefaction), the shell contracts and the concentration of volatile species in the

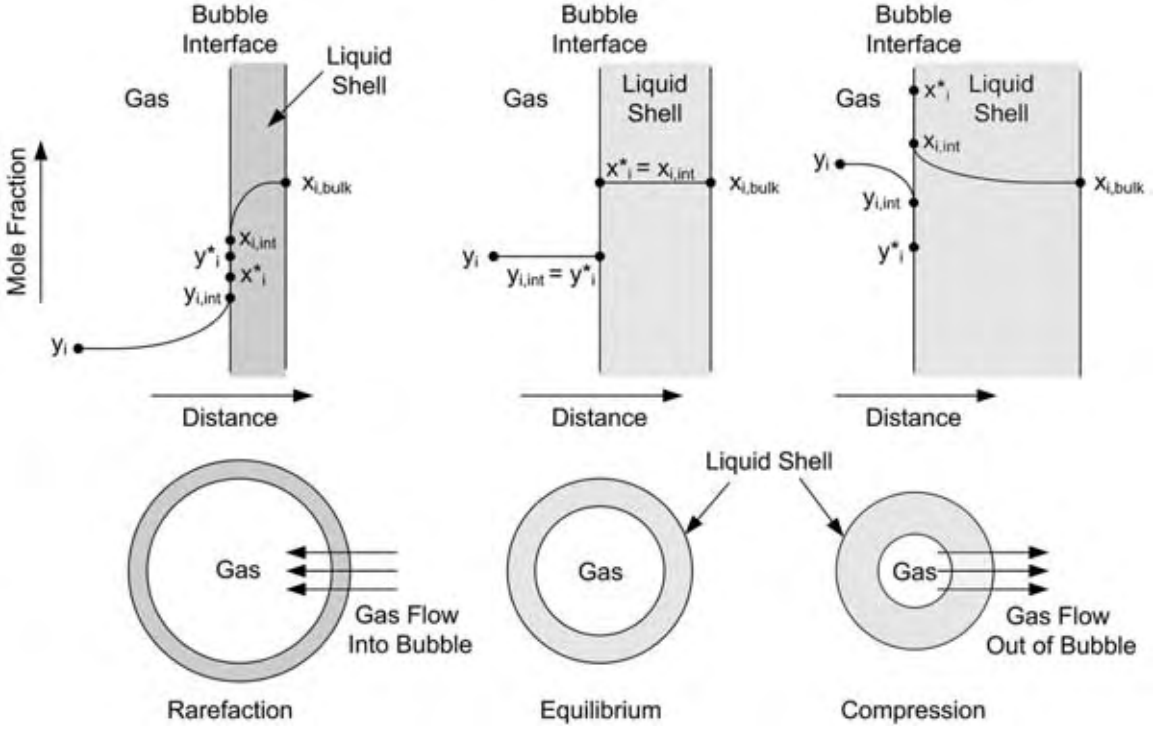


Fig. 3 Concentration profiles for species i in the gas and liquid phases of a microbubble undergoing rectified diffusion: the shell effect.

gas and surrounding liquid decreases. However, the concentration gradient is high because of the thinner shell and so the rate of diffusion of gas into the bubble from the bulk liquid is high. The opposite trend occurs when the bubble contracts. During bubble compression, the liquid shell expands and the concentration of volatile species in the gas and liquid shell increases. The increased shell thickness causes the concentration gradient to be small, thereby lowering the rate of gas diffusion. Hence, the amount of gas diffused is greater during the expansion phase because the high concentration gradient drives the gas a shorter distance, whereas in the compression phase the lower concentration gradient is driving gas a longer distance. Thus, both the area and shell effects lead to a net mass transfer of material into the bubble during each acoustic cycle.

Various mathematical models have been put forth to describe the rate of bubble growth and the threshold pressure for rectified diffusion.[14,20–25] The most widely used model quantifies the extent of rectified diffusion (i.e., the convection effect and bubble wall motion) by separately solving the equation of motion, the equation of state for the gas, and the diffusion equation. To further simplify the derivation, Crum and others made two assumptions: 1) the amplitude of the pressure oscillation is small, i.e., the solution is restricted to small sinusoidal oscillations, and 2) the gas in the bubble remains isothermal throughout the oscillations.[23,24] Given these assumptions, the wall motion of a bubble in an ultrasonic field with an angular frequency of $\omega = 2\pi f$ can be described by the Rayleigh–Plesset equation:

$$r\ddot{r} + \frac{3}{2}(\dot{r})^2 + \left(\frac{P_\infty + 2\sigma/r_0}{\rho}\right)\left(1 - \left(\frac{r_0}{r}\right)^{3\gamma}\right) - \frac{P_A}{\rho}\sin\omega t + r_0\omega_r b\dot{r} = 0 \quad (1)$$

In this expression, r and r_0 are, respectively, the instantaneous and equilibrium (i.e., when no sound field is acting on the liquid) values of the bubble radius; and $\dot{r}$ and $\ddot{r}$ represent, respectively, the first and second order time derivatives of the instantaneous bubble radius; ρ is the liquid density; γ is the polytropic exponent of the gas inside the bubble (i.e., the ratio of heat capacities, C_p/C_v); P_A is the acoustic pressure amplitude; P_∞ is the hydrostatic (ambient) pressure; b is the bubble pulsation damping term that accounts for thermal, viscous, and radiation effects; σ is the liquid surface tension; t is time; and ω_r is the resonance frequency of the bubble, which is defined by the equation below:

$$\omega_r^2 = \left(\frac{1}{\rho r_0^2}\right)\left(3\gamma P_\infty + \frac{2\sigma(3\gamma - 1)}{r_0}\right) \quad (2)$$

This equation of motion for bubble growth includes inertial terms that lead to a nonlinear solution; hence, the pressure threshold of bubble growth is dependent on the frequency of the sound field.

The diffusion of gas in and out of the bubble results from a linear response to the concentration driving force and can be adequately described by Fick's law of mass transfer:[26]

$$\frac{dC}{dt} = \frac{\partial C}{\partial t} + \nu\cdot\nabla C = D\nabla^2 C \quad (3)$$

where C is the molar concentration of gas in the liquid, ν is the liquid velocity, and D is the gas diffusivity in the liquid. By simultaneously solving the equation of state for the gas, the equation of motion for the bubble wall [Eq. (1)], and the equation for the diffusion of dissolved gas [Eq. (3)], it is possible to derive an expression for the mean rate of gas flow into the bubble:

$$\frac{dn}{dt} = 4\pi D C_\infty r_0 H\left[\langle r/r_0\rangle + r_0\left(\frac{\langle (r/r_0)^4\rangle}{\pi D t}\right)^{0.5}\right] \quad (4)$$

where the angled brackets represent time averages of the enclosed variables, n is the number of moles of gas in the bubble, C_∞ is the concentration of dissolved gas in the bulk liquid for the mean ambient pressure P_∞, and H is defined by the equation:

$$H = (C_i\ C_0) - \frac{\langle (R/R_0)^{4-3\gamma}(1 + 2\sigma/r_0P_\infty)\rangle}{\langle (R/R_0)^4\rangle} \quad (5)$$

Further refinements to theories involving bubble growth account for the fact that in a high energy cavitating system the cavitation effects are a result of the entire bubble population rather than a single bubble.[20] The acoustic effects of power ultrasound largely depend on the size distribution of bubbles in the liquid. Depending on the characteristics of the local field, there is a critical size range in which a free floating bubble must lie to undergo transient cavitation. The bubble nuclei formed can grow to provide intense cavitational effects or grow to a resonant size where it will oscillate stably for many cycles. Church predicted that the number of bubble nuclei growing by rectified diffusion to transient collapse decreased with increasing frequency.[25] Leighton also observed that the measured growth rates were often much greater than the predicted values and he attributed this to a phenomenon called "microstreaming."[20] Microstreaming continuously refreshes the liquid at the bubble wall by bringing the liquid from further out close to the bubble wall. This helps to counteract the depletion of gas occurring during rectified diffusion. Microstreaming

also plays a role in dissolution by removing the excess dissolved gas concentration from the region adjacent to the bubble wall.

Bubble Collapse and Splitting

The splitting or implosion of a transient cavity occurs when the bubble has reached a sufficiently large size so that the compression forces of an incoming sound wave exceed the stabilizing forces generated by the internal pressure and the surface tension of the bubble. The time required for bubble collapse is inversely related to the frequency of the ultrasonic wave and often occurs in only a few microseconds. Because the time required for thermal transport is significantly longer than that required for bubble collapse, the gases within the bubble undergo adiabatic compression, generating a localized ''hot spot,'' where temperatures can be as high as 5000 K. Further, the heating rates during bubble collapse are more than a billion degrees per second, and the sudden inrush of the surrounding fluid compresses the bubble contents to pressures of hundreds of atmospheres.[8,19] The high temperature is responsible for some of the production of highly reactive radicals, while the high pressure accounts for additional increases in chemical reactivity.[18] The maximum temperature and pressure reached during adiabatic collapse are given by the following equations:[27]

$$T_{\max} = T_0\left[\frac{(P_A + P_\infty)(\gamma - 1)}{P}\right] \quad (6)$$

$$P_{\max} = P\left[\frac{(P_A + P_\infty)(\gamma - 1)}{P}\right]^{\gamma/(\gamma - 1)} \quad (7)$$

where T_0 is the ambient temperature and P is the pressure in the bubble at its maximum size (sum of the vapor pressure of the liquid and the pressure of the gas in the bubble).

There also exists a second mechanism for producing reactive ions and radicals via cavitation, and it involves the concepts of localized charging and electrical discharge.[14] There is a tendency for a surface potential to be developed at the vapor–liquid interface of a microbubble because of orientational effects and the presence of charged species in the liquid (e.g., low concentration impurities). Following the Helmholtz model, this surface potential leads to the formation of an electrochemical double layer, which has an inner layer of weakly solvated ions and an outer layer of oppositely charged, fully solvated species that can more freely diffuse through the liquid. During bubble collapse or splitting, the ions at the gas–liquid interface (inner Helmholtz layer) become separated from the oppositely charged ions in the outer Helmholtz layer yielding local unbalanced electrical charges having a field strength above the critical electric field strength, which induces electrical microdischarges in the fluid. The discharge of these high energy electrons in the vicinity of other chemical species leads to the formation of highly reactive ions and free radicals.

During cavitation, both hot spot and electrical discharge phenomena lead to the formation of active intermediates (ions and free radicals) that can survive for extended periods and react with species in the liquid layer surrounding the recently collapsed microbubble.

REACTION ZONES OF SONOCHEMICAL REACTIONS

Electron paramagnetic resonance (EPR) and spin trapping studies have shown that there are three regions of sonochemical activity in liquid systems undergoing cavitation, as shown in Fig. 4.[28] Zone 1 contains a mixture of gas and vapor, where the amount of vapor in the bubble is directly related to the vapor pressure of species in the liquid phase. The high temperature and pressure generated during bubble collapse cause organic compounds to undergo pyrolysis reactions. Also, highly reactive radicals (e.g., $OH^\bullet$, $H^\bullet$, and $O^\bullet$) are generated in this zone because of the thermal dissociation of solvent (water) molecules in the vapor. The interfacial region, Zone 2, is also called the supercritical fluid region because the temperature and pressure in this zone during bubble collapse are above the critical temperature and pressure of the liquid (e.g., for water, 647 K and 221 bar). In this region, less

Fig. 4 Three reaction zones generated by acoustic cavitation. Zone 1 is the vapor phase, Zone 2 consists of the supercritical fluid region, and Zone 3 represents the bulk liquid. *(View this art in color at www.dekker.com.)*

volatile and less polar organic molecules are oxidized by the reactive radical species (e.g., $OH^{\bullet}$ and $H^{\bullet}$) generated in Zone 1. Zone 3 consists of the bulk liquid phase outside the bubble. This zone is near ambient temperatures and no primary sonochemical activity takes place in this zone.

The type of sonochemistry a given species will undergo is determined by its vapor pressure, polarity, and chemical composition (i.e., the nature and strength of its bonds). For example, hydrophobic compounds with high vapor pressure readily diffuse into the gas phase of growing microbubbles (assuming a polar solvent is used for the sonication reaction). Thus, these species very often undergo pyrolysis reactions in Zone 1 during bubble collapse, whereas less volatile hydrophobic species tend to accumulate at the vapor liquid interface during bubble growth and undergo reactions with reactive radical species generated in Zone 1 during bubble collapse.[29] Finally, nonvolatile, hydrophilic compounds do not readily diffuse into the vapor phase nor do they accumulate at the bubble interface; thus, they tend to be less affected by cavitation processes because they can react only with longlived radical species that diffuse out of Zone 1.

FACTORS AFFECTING SONICATION CHEMISTRY

This section briefly discusses the factors that significantly affect cavity formation and cavitational intensity (i.e., the temperature and pressure generated during collapse).

Acoustic Frequency

Put simply, acoustic energy manifests itself as an oscillating pressure wave. The frequency of this pressure fluctuation (defined as the number of oscillations per second) is known to directly affect the critical size and number of cavitation bubbles. At very high frequencies (1 MHz), cavitation reduces or ceases to occur because of the very short rarefaction and compression cycles that either reduce cavity formation or prevent bubble collapse. At high frequencies (100 kHz–1 MHz) more cavities are formed, but the intensity of cavitation is lower because the average bubble volume before collapse is small (i.e., less material compressed during collapse). At low frequencies (20–100 kHz) fewer cavities are formed, but the resonance bubble size before collapse is larger; hence, cavitation is more violent, leading to higher localized temperatures and pressures at the cavitation sites.[12] For example, sonic irradiation of water at 20 kHz creates resonant bubbles with a radius of 161 mm, whereas the bubble radius is 6.77 mm at 520 kHz.[30] Also, bubble lifetimes in water are shorter at higher frequencies (e.g., 3×10^{-7} sec at 514 kHz as compared to 3×10^{-5} sec at 20 kHz).[12] Thus, as acoustic frequency increases, production of cavities in the liquid increases (up to 1 MHz) and the intensity of cavitation decreases.[18]

Acoustic Intensity

Acoustic intensity (I) is a measure of the amount of energy transmitted to the liquid and is defined as the rate at which the acoustic energy passes across a unit area perpendicular to the direction of the propagating sound wave. It can be shown that the acoustic intensity is proportional to the square of the amplitude (P_A) of the acoustic wave divided by the density of the liquid (ρ) and the speed of sound in the liquid (c) (e.g., $c = 1500$ m/sec for water):[20]

$$I = \frac{P_A^2}{2\rho c} \tag{8}$$

To achieve any significant sonochemical effect, the sound intensity must be sufficiently high to overcome the cohesive forces in the liquid (i.e., break the intermolecular bonds) so that significant quantities of cavities can be formed. For degassed water at 25°C, the energy necessary to overcome the cohesive forces in the liquid is 6000 W/cm^2 at an acoustic frequency of 500 kHz. With increases in the intensity value, the sonochemical rate increases because of the large number of cavities formed. Also, the range of bubble sizes undergoing transient cavitation increases, increasing reaction rates. However, several experiments have shown that there is an upper limit to the power input. When the intensity is increased beyond this upper limit, bubbles grow so large in a single rarefaction cycle that they have insufficient time to collapse before the next rarefaction cycle.[16] Additionally, these large bubbles shroud the surface of the acoustic generator, reducing the coupling of sound energy to the liquid. Thus, sonochemical reaction rates increase with increasing acoustic intensity up to some limit where mechanical damage to the sound generator can occur or the presence of large bubbles reduces the overall efficiency of the cavitation process.[16,12]

External Pressure

An increase in the ambient (or hydraulic) pressure has two effects. First, there is an increase in the minimum acoustic intensity required to initiate cavitation, which occurs when the acoustic wave has sufficient amplitude to overcome both the tensile strength of the liquid and the external pressure. The second effect is the increase

in the intensity of cavitation collapse, leading to higher collapse temperatures and pressures. This is because the total pressure inside the cavitation bubble, just before collapse, is higher owing to the increased external pressure and applied ultrasound power. Even though the temperature and pressure in the final stages of collapse will increase, the overall sonochemical rate may or may not increase depending on the types of active intermediates formed and which reaction zone they are formed in. In contrast, as the external pressure is reduced, the severity of bubble collapse decreases and the dissolved gas content in the liquid decreases, eventually leading to a smaller number of microbubble nucleation sites (i.e., fewer cavitation bubbles). These effects combine to reduce the overall sonochemical reaction rate. Thus, the liquid surface tension and the nature of the specific chemical reaction determine the optimal external pressure for a sonochemical reaction.[12,16,18]

Bulk Temperature

Most homogenous reactions are accelerated by moderate increases in temperature; however, sonochemical reactions are often slowed down by increases in the bulk fluid temperature. This counterintuitive observation is easily explained once one considers the cavitation processes that lead to the formation of reactive radicals and ions. As the fluid temperature increases, the vapor pressure of the solvent and dissolved species increases, which leads to greater quantities of vapor entering into the gas phase of cavitation microbubbles. This increased quantity of vapor in the bubble cushions the collapse, lowering the temperature and pressure generated during bubble implosion, thereby lowering the number of reactive intermediates formed. Additionally, increased temperatures decrease the cavitational threshold for the liquid, which leads to the formation of large numbers of microbubbles that serve to dampen the passage of ultrasound energy through the liquid medium.[17] Hence, sonochemical reactions are usually carried out within a temperature range of 10–35°C.

Presence and Nature of Dissolved Gases

The presence of dissolved gas is essential for cavitation to occur in a liquid. The dissolved gas molecules disrupt intermolecular bonding between solvent molecules and hence, serve as nucleation sites for cavitation. There are three properties of dissolved gases that have significant influence on the degree of nucleation and cavitational intensity: solubility of gas in the liquid, ratio of specific heats (γ or C_P/C_V), and thermal conductivity (λ). More soluble gases reduce the cavitational effects because the bubbles formed redissolve before collapse can occur.[12] Also, the greater the solubility of the gas, the greater the amount that penetrates the bubble, thereby inducing the cushioning effect and lowering the intensity of shock waves released after collapse.[17,18] As the gas content in the liquid increases, both the cavitation threshold and the cavitational intensity are reduced.[18]

High degradation of organic compounds is achieved when a gas with high average specific heat ratio is employed. From Eqs. (6) and (7), the maximum temperature and pressure generated during adiabatic collapse increase with an increase in the value of γ, leading to higher degradation. This has been confirmed by sonochemical degradation studies of aqueous solutions of carbon tetrachloride, which showed that the initial rate of formation of free chlorine was less when the solution was saturated with nitrogen ($\gamma = 1.40$) instead of argon ($\gamma = 1.66$).[31] Solutions saturated with monatomic (inert) gases (Ar, Ne, He) give the highest rate with diatomic gases (O_2 and N_2) proceeding at an intermediate rate, and the lowest rate occurs for polyatomic gases (CO, etc.).

The thermal conductivity of the gas also has an important effect on cavitation intensity. Dissolved gases with high thermal conductivity can more rapidly dissipate heat generated during collapse to the surroundings, effectively reducing the maximum temperature (T_{max}) attained. Hence, even though helium ($\gamma = 1.66$), being an inert gas, has a high γ value, the maximum temperature attained during collapse is lower (with all the other conditions maintained the same) than that of nitrogen ($\gamma = 1.4$) because of the higher λ value of the former ($\lambda_{He} = 14.30 \times 10^{-2}\,W/m/K$ and $\lambda_{N2} = 2.52 \times 10^{-2}\,W/m/K$). Thus, greater cavitation intensity is achieved when the liquid is saturated with a high γ and low λ inert gas.

Solvent

The nature and strength of solvent intermolecular interactions (especially the presence of hydrogen bonding) can greatly influence the physical and chemical outcomes of ultrasound irradiation. These interactions determine the intensity of bubble collapse and the ease with which bubble nucleation can occur (recall that nucleation occurs when the pressure amplitude of the ultrasound wave exceeds the natural cohesive forces in the liquid). Increases in solvent viscosity and surface tension reduce the rate of bubble nucleation (i.e., fewer microbubbles are formed) but increase the intensity of bubble collapse (i.e., higher temperatures and pressures).[17,18] Some of the adverse effects of high surface tension can be overcome with the addition of small amounts of surfactants, which reduce the solvent surface tension and facilitate bubble nucleation.[12,17]

Solvent vapor pressure also has a significant effect on the cavitation phenomenon because the intensity of cavitation decreases as the vapor pressure of the solvent increases. This is because more vapor is enclosed in the microbubble, which cushions the collapse, leading to lower collapse temperatures and pressures. On the other hand, solvents with low vapor pressure tend not to diffuse into the growing microbubble; thereby reducing the size of the bubble, which lessens the intensity of bubble collapse.[17] Thus, a delicate balance of solvent properties must be achieved to attain the desired sonication conditions.

TYPES OF CAVITATION REACTIONS

Ultrasound irradiation can affect chemical reactions in two basic ways. It will either activate a new reaction or enhance the rate of an existing reaction pathway. Most primary activation processes involve reactions that occur in Zone 1 during bubble collapse, which are generally thought to be thermally initiated processes. During collapse, vaporized molecules are dissociated into smaller fragments (e.g., $H^{\bullet}$, Cl^{-}, $CH_3^{\bullet}$), which diffuse into the surrounding liquid media and react with other compounds, degrade into even smaller reactive species, or recombine to form low molecular weight compounds, such as CO_2, H_2O, and HCl. Also, solvent molecules are commonly dissociated into highly reactive radical species (e.g., $CH_3^{\bullet}$, $OH^{\bullet}$, $H^{\bullet}$). This reaction mechanism is the main degradation path for volatile organics with high vapor pressure and is predominant at lower ultrasound frequencies. The concentration of these high energy dissociation products is relatively small and their lifetimes short, but it is possible for them to diffuse to Zones 2 and 3 where they can react with other nonactivated species.

Secondary reactions involving free radicals (e.g., $OH^{\bullet}$ and $H^{\bullet}$), which were created by pyrolysis reactions in Zone 1, take place in the supercritical fluid region (Zone 2) and in the bulk liquid (Zone 3). These reactions involve species that strongly associate with the solvent or low volatility organics that cannot easily diffuse into the bubble. Higher ultrasound frequencies generally enhance these types of reactions. This is because at low ultrasound frequencies, the decomposition of the vapor gas mixture inside the microbubble is very efficient, yielding numerous free radicals. However, the proportion of radicals undergoing recombination to form unreactive species is also very high, whereas, at higher ultrasound frequencies fewer radicals are formed, but a greater percentage escape from the bubble; hence, the concentration of radical species exiting Zone 1 is greater at higher ultrasound frequencies.[32]

The number of organic reactions that have been initiated or enhanced by sonochemical methods is vast. Table 1 provides a brief listing of some of these. More detailed and thorough reviews of this subject can be found in the literature.[11,12,16,19,30–36] Sonochemical methods have proven to be efficient means to achieve a wide array of addition, isomerization, and oxidation reactions, while slightly fewer studies have examined the use of ultrasound for reduction and substitution reactions. In general, ultrasound irradiation has been shown to play varying roles in homogeneous and heterogeneous synthetic chemistry. For some reaction systems, the extreme conditions encountered during bubble collapse cause the initiation of a new reaction (e.g., the reduction of alkoxysilanes using $LiAlH_4$), whereas other reaction systems simply experience an increase in reaction rates that results from enhanced mixing and, in some cases, an increase in the number of highly reactive free radical intermediates (e.g., the oxidation of organics using $KMnO_4$). Most industrial applications of sonochemistry have involved the degradation of organic compounds, especially the destruction (pyrolysis or oxidation) of organics, such as aromatics and chlorinated organics, dissolved in water. These compounds are major contaminants in industrial and agricultural wastewater, and near-complete destruction (conversion to CO_2 and H_2O) of low molecular weight organics having high volatilities has been achieved. Further, dehalogenation reactions have been shown to occur with organic species having much lower volatilities (chlorobenzene, polychlorobiphenyls, etc.).

SONOLYSIS OF WATER

Though there are many important reactions initiated by ultrasound, water dissociation is by far the most important, and hence, is discussed here in detail. Weiss proposed and many have confirmed the mechanism shown below for the formation of $H^{\bullet}$ and $OH^{\bullet}$ radicals in Zone 1 during bubble collapse:[37]

$$H_2O \rightarrow H^{\bullet} + OH^{\bullet}$$

The formation of these free radical intermediates and subsequent sonochemical reactions carried out by them depend largely on the nature of the dissolved gases in the aqueous medium. Studies have shown that the sonolysis of distilled water saturated with different pure gases and their mixtures has different hydrogen peroxide yields and cavitation threshold values.[31] In general, these studies show that a gas with a high ratio of heat capacities and low thermal conductivity yields higher rates of production of hydrogen peroxide, indicating increased levels of free radical production

Table 1 A representative list of organic reactions that are initiated or enhanced via irradiation with ultrasound

Homogenous organic reactions	
Addition reactions	
Aldol condensations	Hydroboration reactions
Alkene hydrosilations	Knoevenagel condensations
Alkylation reactions	Michael addition reactions
Claisen–Schmidt condensations	Reformatsky reactions
Diels–Alder cyclizations	Ring opening polymerizations
Enantioselective Barbier reactions	Ullam coupling reactions
Esterification reactions	Wittig reactions to form olefin
Nucleophilic and electrophilic substitution reactions	
Ester hydrolysis	Halogenation of aromatics
Friedel–Crafts acylations	Nitration of aromatics
Friedel–Crafts alkylations	Solvolysis of halogenated alkanes
Oxidation reactions	
Alcohol oxidation to ketones and carboxylic acids	Oxidation of halogenated alkanes/aromatics
Epoxidation of olefins	Oxidation of arylalkyls
Reduction reactions	
Clemmensen reduction of carbonyls	Reduction of alkoxy silanes
Hydrogenation of carbonyls	Selective reduction of olefins
Degradation reactions	
Complete pyrolysis of organics	Polymer degradation
Dehalogenation reactions	Ozonolysis

via the thermal decomposition of water. The following free radical reactions were proposed for water upon ultrasonic irradiation under argon:[16]

$$H_2O \leftrightarrow H^\bullet + OH^\bullet$$
$$H^\bullet + H^\bullet \rightarrow H_2$$
$$OH^\bullet + OH^\bullet \rightarrow H_2O_2$$

Hydrogen peroxide is more readily formed when water contains oxygen, but the presence of oxygen is not necessary for hydrogen peroxide formation.[31] In the presence of oxygen, more H atoms are scavenged via the following reaction:

$$H^\bullet + O_2 \rightarrow HOO^\bullet$$

This reaction prevents the recombination of $H^\bullet$ and $OH^\bullet$ radicals to form water molecules and hence, increases the rate of other oxidation processes. In general, free radical reactions similar to those shown for water could occur with any organic or inorganic species capable of being present as a vapor during bubble collapse.

TYPES OF SONOCHEMICAL REACTORS

There are a variety of sonochemical reactor designs that have been developed for laboratory use as well as large scale production units. The method of ultrasound generation depends largely on the size of the reactor with piezoelectric transducers being used for small systems and magnetostrictive transducers more commonly found in large scale reactors. There are three basic types of sonochemical reactors: 1) reactors having the ultrasonic source in direct contact with the reacting fluid; 2) reactors with the ultrasound source mounted on the exterior walls of the reactor (i.e., a vibrating wall reactor); and 3) reactors that receive sound energy from an exterior coupling fluid that is in direct contact with the ultrasound source. There are several recent reviews of sonochemical reactors in the literature, and these works should be consulted before deciding on a particular sonochemical reactor configuration.[11,12,16] There are three major factors that must be taken into consideration when selecting the appropriate reactor type: cost, level of contaminants (from the erosion of the ultrasound source), and ultrasound energy density required.

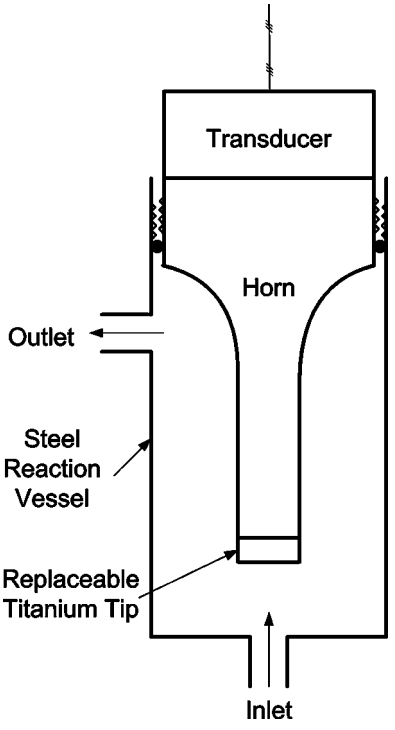


Fig. 5 Flow through reactor with direct sonication via an amplifying horn.

The first reactor type, which employs direct sonication, is represented by the lab scale system shown in Fig. 5. This flow reactor assembly consists of a titanium horn assembly, a piezoelectric transducer capable of variable energy output, and a flow through reaction chamber. This design provides for maximum utilization of the sound energy output from the transducer, but at a price. The replaceable tip of the titanium horn is prone to erosion and pitting, which can lead to a reduction in energy utilization as well as contamination of the process stream.

The second type of sonochemical reactor normally consists of a flat walled vessel that has multiple piezoelectric transducers mounted on the exterior surface of the reactor (in much the same way ultrasonic baths are arranged). These reactors have noncylindrical geometries because of difficulties associated with mounting the ceramic transducers on a curved surface (the transducer is normally glued to the exterior surface of the reactor). More optimized reactor designs of this type have multiple transducers located on opposing sides of the reactor, which creates a more uniform sound field for cavitation. An advantage of this reactor type is that corrosion byproducts from the ultrasound source cannot contaminate the reacting media; however, the coupling of the transducers to the reactor wall can be inefficient, and it is difficult to transmit high energy fluxes through the reactor walls.

The final reactor configuration type usually has the reacting fluid pass through a metal tube that is surrounded by a coupling fluid that is in direct contact with an ultrasound source. An efficient variant of this reactor type is the Branson sonochemical reactor shown in Fig. 6. This reactor configuration consists of a straight pipe for the reacting media and a series of external pipe sections that are arranged perpendicular to the reactant flow. This design is nonintrusive, hence there is little chance of contamination of the reagents; however, any design that uses probes

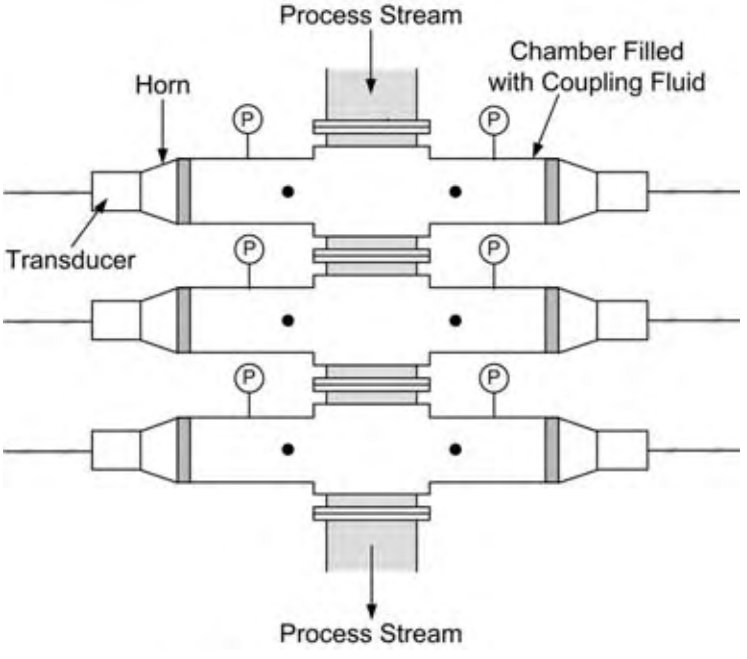


Fig. 6 Branson sonochemical reactor employing indirect sonication in a tubular configuration.

or horns to focus the acoustic energy is at a disadvantage because of tip erosion, which ultimately leads to greater maintenance costs.

CONCLUSIONS

Acoustic cavitation provides a relatively simple means for accessing high energy species in a reactor assembly that operates at near-ambient conditions. All sonochemical reaction processes are initiated by the introduction of high intensity sound waves into a liquid with moderately high surface tension. These high frequency pressure waves generate micrometer size bubbles that grow until they suddenly collapse (acoustic cavitation), generating extremely high pressures and temperatures (up to 1000 atm and 5000°C) in localized regions of the fluid. Such extreme thermal conditions lead to the formation of a variety of highly reactive free radicals and charged species. The interaction of these reactive intermediates with neighboring (cooler, low energy) species results in unique homogeneous, and in some cases heterogeneous, chemical conversions. To date, sonochemical transformations have been effectively employed in a wide variety of wastewater treatment applications; however, a number of efficiency and contamination issues have limited their use in the large scale production of organic chemicals and inorganic materials. As the technology for introducing high energy sound waves into liquids continues to advance, it is likely that sonochemical processing will become an ever more widely used tool to prepare high value added chemicals and materials.

REFERENCES

1. Thornycroft, J.I.; Barnaby, S.W. Torpedo-boat destroyers. Min. Proc. Inst. Chem. Eng. **1895**, *122* (4), 51–69.
2. Rayleigh, Lord. On the pressure develop in a liquid during collapse of a spherical cavity. Philos. Mag. Ser. 6, **1917**, *34* (200), 94–98.
3. Wood, R.W.; Loomis, A.L. The physical and biological effects of high frequency sound waves of great intensity. Philos. Mag. Ser. 7, **1927**, *4* (22), 417–436.
4. Richards, W.T.; Loomis, A.L. The chemical effects of high frequency sound waves. I. A preliminary study. J. Am. Chem. Soc. **1927**, *49*, 3086–3100.
5. Harvey, E.N.; Barnes, D.K.; McElroy, W.D.; Whitely, A.H.; Pease, D.C.; Cooper, K.W. Bubble formation in animals. J. Cell. Comp. Physiol. **1944**, *24*, 1–22.
6. Noltingk, B.E.; Neppiras, E.A. Cavitation produced by ultrasonics. Proc. Phys. Soc. **1950**, *63B*, 674–685.
7. Naude, C.F.; Ellis, A.T. On the mechanism of cavitation damage by non-hemispherical cavities in contact with a solid boundary. J. Basic Eng. **1961**, *83*, 648–656.
8. Suslick, K.S.; Hammerton, D.A.; Cline, R.E. The sonochemical hot spot. J. Am. Chem. Soc. **1986**, *108*, 5641–5642.
9. Suslick, K.S. Sonochemistry. Science **1990**, *247* (4949), 1439–1441.
10. Makino, K.; Mossoba, M.M.; Reisz, P. Chemical effects of ultrasound on aqueous solutions. Formation of hydroxyl radicals and hydrogen atoms. J. Phys. Chem. **1983**, *87*, 1369–1377.
11. Adewuyi, Y.G. Sonochemistry: environmental science and engineering applications. Ind. Eng. Chem. Res. **2001**, *40*, 4681–4715.
12. Thompson, L.H.; Doraiswamy, L.K. Sonochemistry: science and engineering. Ind. Eng. Chem. Res. **1999**, *38*, 1215–1249.
13. Gedanken, A. Using sonochemistry for the fabrication of nanomaterials. Ultrason. Sonochem. **2004**, *11*, 47–55.
14. Margulis, M.A.; Margulis, I.M. Mechanism of sonochemical reactions and sonoluminescence. High Energ. Chem. **2004**, *38* (5), 285–294.
15. Pandit, A.B.; Moholkar, V.S. Harness cavitation to improve processing. Chem. Eng. Prog. **1996**, *7*, 57–68.
16. Shah, Y.T.; Pandit, A.B.; Moholkar, V.S. *Cavitation Reaction Engineering*; Kluwer Academic/Plenum Publishers: New York, 1999.
17. Moholkar, V.S.; Shirgaonkar, I.Z.; Pandit, A.B. Cavitation and sonochemistry in the eyes of a chemical engineer. Indian Chem. Eng. B. **1996**, *38* (2), 81–93.
18. Lorimer, J.P.; Mason, T.J. Sonochemistry: part 1—the physical aspects. Chem. Soc. Rev. **1987**, *16*, 239–274.
19. Suslick, K.S. The chemical effects of ultrasound. Sci. Am. **1989**, *260*, 80–86.
20. Leighton, T.G. Bubble population phenomena in acoustic cavitation. Ultrason. Sonochem. **1995**, *2* (2), S123–S135.
21. Hsieh, D.Y.; Plesset, M.S. Theory of rectified diffusion of mass into gas bubbles. J. Acoust. Soc. Am. **1961**, *33* (2), 206–215.
22. Eller, A.; Flynn, H.G. Rectified diffusion during nonlinear pulsations of cavitation bubbles. J. Acoust. Soc. Am. **1965**, *37* (3), 493–503.
23. Crum, L.A.; Hansen, G.M. Generalized equations for rectified diffusion. J. Acoust. Soc. Am. **1982**, *72* (5), 1586–1592.
24. Crum, L.A. Rectified diffusion. Ultrasonics **1984**, *22L*, 215–223.

25. Church, C.C. A method to account for acoustic microstreaming when predicting bubble-growth rates produced by rectified diffusion. J. Acoust. Soc. Am. **1988**, *84* (5), 1758–1764.
26. Bird, R.B.; Stewart, W.E.; Lightfoot, E.N. *Transport Phenomena*; John Wiley & Sons: New York, 1960.
27. Neppiras, E.A. Acoustic cavitation. Phys. Rep. **1980**, *61* (3), 159–251.
28. Riesz, P.; Kondo, T.; Krishna, C.M. Sonochemistry of volatile and non-volatile solutes in aqueous solutions: E.P.R and spin trapping studies. Ultrasonics **1990**, *28*, 295–303.
29. Drijvers, D.; Van Langenhove, H.; Vervaet, K. Sonolysis of chlorobenzene in aqueous solution: organic intermediates. Ultrason. Sonochem. **1998**, *5* (1), 13–19.
30. De Visscher, A.; Van Langenhove, H. Sonochemistry of organic compounds in homogeneous aqueous oxidising systems. Ultrason. Sonochem. **1998**, *5* (3), 87–92.
31. Fitzgerald, M.E.; Griffing, V.; Sullivan, J. Chemical effects of ultrasonics—"Hot Spot'' chemistry. J. Chem. Phys. **1956**, *25* (5), 926–933.
32. Mason, T.J. Sonochemistry: current uses and future prospects in the chemical and processing industries. Philos. Trans. R. Soc. Lond. A. **1999**, *357*, 355–369.
33. Bremner, D.H. Recent advances in organic synthesis utilizing ultrasound. Ultrason. Sonochem. **1994**, *1* (2), S119–S124.
34. Low, C.M.R. Ultrasound in synthesis: natural products and supersonic reactions? Ultrason. Sonochem. **1995**, *2* (2), S153–S163.
35. Einhorn, C.; Einhorn, J.; Luche, J.L. Sonochemistry: the use of ultrasonic waves in synthetic organic chemistry. Synthesis **1989,** Nov, 787–813.
36. Ley, S.V.; Low, C.M.R. *Ultrasound in Synthesis*; Springer-Verlag: Berlin, 1989.
37. Weiss, J. Radiochemistry of aqueous solutions. Nature **1944**, *153*, 748–750.

Sorbent Technology

Shuguang Deng
Chemical Engineering Department, New Mexico State University, Las Cruces, New Mexico, U.S.A.

INTRODUCTION

This article covers the fundamentals, status, and future developments of sorbent materials and their applications in adsorptive separation and purification processes. A sorbent is usually a solid substance that adsorbs or absorbs another type of substance. It is the sorbent that makes a sorption process a unique and different separation and purification process from others. With the rapid development in novel sorbent materials and innovative cyclic adsorption processes, sorption has become a key separation process in many process industries including chemical, petrochemical, environmental, pharmaceutical, and electronic gases. A brief review of the fundamentals of adsorption and the basic requirements for sorbent materials is presented, followed with a summary of the status of commercial sorbents and their applications. The focus of this article is placed on recent advances in novel sorbent materials including oxide molecular sieves, sol–gel derived xerogels and aerogels, metal organic framework, hydrogen storage media, π-complexation and composite sorbents, and high-temperature sorbents for oxygen or carbon dioxide sorption. A concluding section outlines the future research needs and opportunities in sorbent technology development for new energy and environmental applications.

ADSORPTION MECHANISMS AND SORBENT MATERIALS

According to King, a mass separating agent is needed to facilitate separation for many separation processes.[1] The mass separating agent for adsorption process is the adsorbent, or the sorbent. Therefore, the characteristic of the sorbent directly decides the performance of any adsorptive separation or purification process. The basic definitions of adsorption-related terminologies are given in the following to clarify and standardize these widely used terms in this field.

Adsorption: The adhesion of molecules (as of gases, solutes, or liquids) to the surfaces of solid bodies or liquids with which they are in contact.

Absorption: The absorbing of molecules (as of gases, solutes, or liquids) into the solid bodies or liquids with which they are in contact.

Sorption: Formation from adsorption and absorption.

Adsorbent: A usually solid substance that adsorbs another substance on its surface.

Sorbent: A usually solid substance that adsorbs and absorbs another substance.

Adsorbate: Molecules (as of gases, solutes, or liquids) that are adsorbed on adsorbent surfaces.

Microporous: Pore size smaller than 20 Å.

Mesoporous: Pore size between 20 and 500 Å.

Macroporous: Pore size larger than 500 Å.

Adsorptive separation can be achieved through one of the following mechanisms. Understanding the fundamentals of adsorptive separation mechanisms will allow us to better design or modify sorbent materials to achieve their best possible separation performance.[2–4]

Adsorption equilibrium effect is because of the difference in the thermodynamic equilibria for each adsorbate–adsorbent interaction. The majority of adsorptive separation and purification processes are based on equilibrium effect. One example is to generate oxygen-enriched air or relatively pure oxygen (95%) from air using a zeolite molecular sieve 5A or 13X in either a pressure swing adsorption (PSA) or a vacuum swing adsorption (VSA) process. In this case, nitrogen is selectively adsorbed by the zeolite adsorbent, and oxygen is collected from the adsorption effluent stream.

Adsorption kinetics effect arises because of the difference of rates at which different adsorbate molecules travel into the internal structure of the adsorbent. There are only a few commercial successes using adsorption kinetic difference to achieve adsorptive separation of gases. The typical example is separation of nitrogen from air using a carbon molecular sieve (CMS). The CMS adsorbent has a similar adsorption equilibrium capacity for both nitrogen and oxygen, but the diffusivity of oxygen in CMS is at least 30 times larger than that of nitrogen in CMS.[5] High-purity

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007963

nitrogen can be recovered from the adsorption effluent stream in a PSA process because oxygen moves much faster than nitrogen into the micropores of CMS adsorbent. However, the cycle time of this CMS-based PSA process is much shorter than that of a typical PSA process based on adsorption equilibrium effect. This is because there will be no separation if both nitrogen and oxygen are allowed to reach adsorption equilibrium with the CMS adsorbent.

Molecular sieving effect, also called steric effect, is derived from the molecular sieving properties of some adsorbents with a microporous structure. In this case, the pore openings of the adsorbent structure are small enough to exclude large adsorbate molecules from penetrating the micropores of the adsorbent. This is the extreme case of the kinetic effect. There are several commercial applications based on this mechanism in adsorptive separation processes. One typical example is separating normal paraffin from *iso*-paraffin and aromatics in an adsorption process using zeolite 5A as an adsorbent. *n*-Paraffin, with a long straight chain, has a smaller effective diameter than the well-defined aperture of zeolite 5A. Therefore it adsorbs in the micropores of the adsorbent during the adsorption step, and is recovered from the adsorbed phase in the desorption step. A representative process for *n*-paraffin separation from naphtha and kerosene is UOP's Molex process that employs a simulated moving bed with binderless zeolite 5A as an adsorbent and light paraffin as a desorbent.[6]

We can define separation factor and selectivity as the ability of an adsorbent to separate molecule A from molecule B as:[7]

$$\text{Separation factor}: \alpha_{AB} = \frac{X_A/Y_A}{X_B/Y_B} \quad (1)$$

Here X_A, Y_A are strictly equilibrium mole fractions for component A in the adsorbed phase and adsorbate (fluid) phase, respectively; as are X_B, Y_B for component B. For equilibrium-based adsorptive separation process, the adsorbent selectivity is the same as the separation factor as defined in Eq. (1). Apparently, this definition is not applicable to other processes based on kinetic and steric effects. In a kinetically controlled adsorption process, the adsorbent selectivity depends on both equilibrium and kinetic effects. A simplified definition for adsorbent separation factor is given by Ruthven et al.:[8]

$$S_{AB} = \frac{K_A}{K_B}\sqrt{\frac{D_A}{D_B}} \quad (2)$$

where S_{AB} is the adsorbent selectivity, K is the adsorption equilibrium constant or isotherm slope, and D is the effective diffusivity. Although the above equation is strictly valid under the assumptions that components A and B have independent linear adsorption isotherms and independent diffusion process, it provides a good estimate of adsorbent selectivity for kinetically controlled processes.

Theoretically speaking, selectivity for adsorbents with a molecular sieving effect should be infinitely large because the larger molecules are excluded from getting into the adsorbent micropores. In reality, the adsorbent selectivity for steric effect is somewhat reduced by combining with the equilibrium effect from adsorption on the surface of large pores. So adsorption processes based on molecular sieving are usually considered as adsorption equilibrium effect.

Another very important adsorbent property affecting the adsorption process is the adsorption capacity because it determines the size of an adsorbent vessel, the amount of adsorbents required, and the related capital and operating costs. The requirements for commercial sorbents are discussed briefly as follows.

Characteristics of Sorbent Materials

Commercial sorbents used in cyclic adsorption processes should ideally meet the following requirements:

- Large selectivity derived from equilibrium, kinetic, or steric effect;
- Large adsorption capacity;
- Fast adsorption kinetics;
- Easily regenerable;
- Good mechanical strength;
- Low cost.

The above adsorbent performance requirements can simply transfer to adsorbent characteristic requirements as follows:

- Large internal pore volume;
- Large internal surface area;
- Controlled surface properties through selected functional groups;
- Controlled pore size distribution, preferably in micropore range;
- Weak interactions between adsorbate and adsorbent (mostly on physical sorbents);
- Inorganic or ceramic materials to enhance chemical and mechanical stability;
- Low-cost raw materials.

These basic requirements are usually proposed for adsorbents used in cyclic adsorption processes that are based on physical adsorption. There is an increasing demand for strong chemical adsorbents used in

purification processes to remove trace contaminants from main stream fluids such as the removal of very toxic contaminants from electronic process gas streams, and the removal of toxic, or radioactive species from contaminated water. In these cases, the sorbents are used as getter materials; no regeneration is needed, and instead, the spent sorbent materials are disposed of in designated areas regulated by government environmental policies.

COMMERCIAL SORBENTS AND APPLICATIONS

An excellent review and detailed coverage on commercial adsorbents and new adsorbent materials has been presented by Yang in his newly published monograph on adsorbents.[2] A very brief overview of existing commercial adsorbents is given here. Commercial sorbents that have been used in large-scale adsorptive separation and purification processes include activated carbon, zeolites, activated alumina, silica gel, and polymeric adsorbents. Although the worldwide sales of sorbent materials are relatively small as compared with other chemical commodities, sorbents and adsorption processes play a very important role in many process industries. The estimated worldwide sales of these sorbents are as follows:[2]

- Activated carbon: $1 billion
- Zeolite: $1.07 billion
- Activated alumina: $63 million
- Silica gel: $71 million
- Polymeric adsorbents: $50 million

Activated Carbon

Activated carbons are unique and versatile adsorbents because of their large surface area, microporous and mesoporous structure, universal adsorption effect, high adsorption capacity for many nonpolar molecules including organic molecules, and high degree of surface reactivity. They are used widely in industrial applications that include decolorizing sugar solutions, personnel protection, solvent recovery, volatile organic compound removal from air and water, water treatment, hydrogen and synthesis gas separation, and natural gas storage.[4,9,10] Activated carbons are produced in two main steps: carbonization of the carbonaceous raw materials at temperatures below 800 °C in the absence of oxygen, and activation of the carbonized products.[10] The properties of activated carbon depend largely on the nature of the raw materials, the activating agents and activation conditions. For gas-phase applications, activated carbons are usually made in pellets with mostly micropores; while for liquid-phase applications, activated carbon is produced in powder form with relatively large mesopores to enhance mass transfer rate in the carbons.

Fig. 1 compares the pore size distributions of major commercial adsorbents discussed in this section. Activated carbons have a broad pore size distribution like activated alumina and silica gel. Although activated carbon is thought to be "hydrophobic," it does adsorb

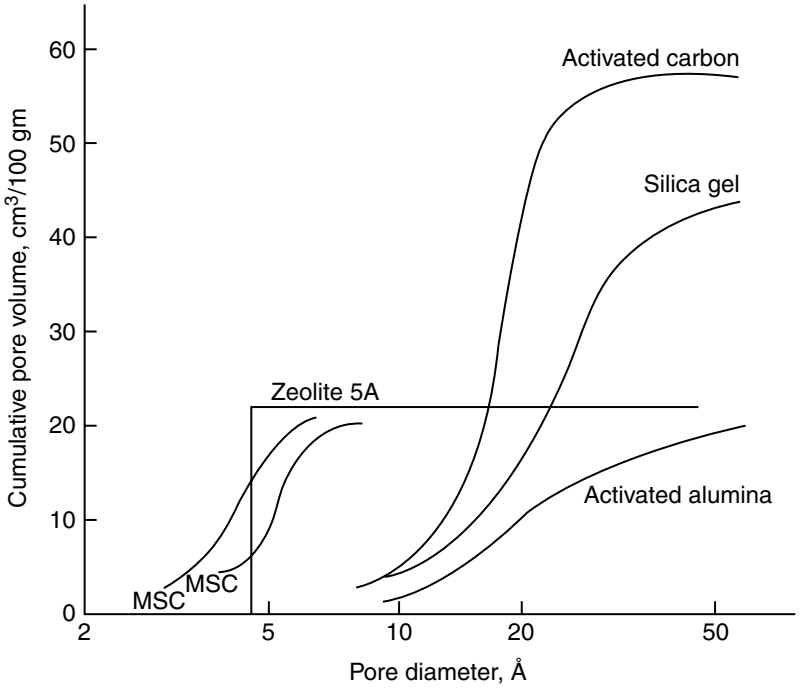


Fig. 1 Pore size distributions for activated carbon, silica gel, activated alumina, two molecular sieve carbons (MSCs), and zeolite 5A. (From Ref.[3].)

quite significant amount of water (>30 wt%) when relative humidity is higher than 50%. An example isotherm of water on activated carbon, along with water isotherms on other commercial adsorbents, is plotted in Fig. 2. The change from ''hydrophobic'' to ''hydrophilic'' on the activated carbon surface is attributed to the initial adsorbed water film on the carbon surface. This occurs because when the carbon surface is fully covered with a layer of water molecules, the adsorbed water molecules exhibit strong affinity to other polar molecules including water. Carbon molecular sieve (CMS) is a specially made carbonaceous material with very narrow pore size distribution (4–9 Å). The major application of CMS is in the generation of high-purity nitrogen from air in a PSA process. The representative physical properties of commercial adsorbents and their major applications are summarized in Tables 1 and 2, respectively.

Zeolites

Zeolites are porous crystalline aluminosilicates that are made of assemblies of SiO_4 and AlO_4 tetrahedra joined together through shared oxygen atoms. The general chemical formula for zeolites is:

$$M_{x/n}[(AlO_2)_x(SiO_2)_y]zH_2O \quad (3)$$

where x and y are integers with y/x (Si/Al ratio) equal or larger than 1; n is the valance of cation M, and z is the number of water molecules in each unit cell. The tetrahedra can be arranged in many different ways to form different crystalline structures. Some zeolites exist as minerals in nature, but all commercially important zeolites are synthetic. Zeolites are unique adsorbents owing to their special surface chemistries and crystalline pore structures. It should be pointed out that probably only 10% of $1 billion worldwide sales of zeolite is used as adsorbents; the majority of commercial zeolites are used as detergent additives (zeolite 4A), animal food additives (zeolite 4A), ion exchange, and catalyst supports. Among all commercial sorbents zeolites are probably the most extensively investigated and documented. Many excellent monographs and review articles are available.[2,11–13] Please refer to Tables 1 and 2 for properties and major applications of zeolites.

Activated Alumina

Activated alumina is a porous high-surface area form of aluminum oxide with the formula of $Al_2O_3 \cdot nH_2O$. Commercially, it is prepared either from thermal dehydration of aluminum trihydrate, $Al(OH)_3$, or directly from bauxite ($Al_2O_3 \cdot 3H_2O$), as a by-product of the Bayer process for alumina extraction from bauxite. Its surface is more polar than that of silica gel and, reflecting the amphoteric nature of aluminum, has both acidic and basic characteristics. Surface areas are in the range 250–350 m^2/g depending on the activation temperature and the source of raw materials. Because activated alumina has a higher capacity for water than silica gel at elevated temperatures it is used mainly as a desiccant for warm gases including air, but in many commercial applications it has now been replaced by zeolitic materials in a thermal swing

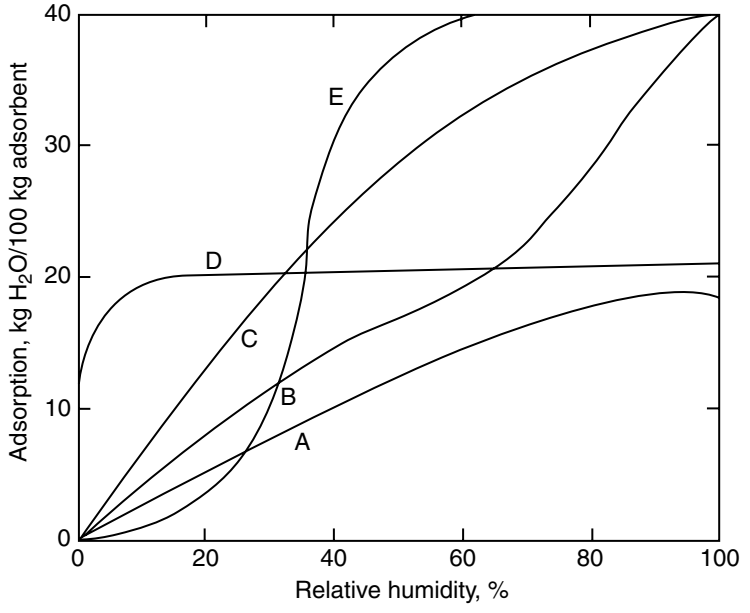


Fig. 2 Equilibrium sorption of water vapor from atmospheric air at 25 °C on: (A) alumina (granular), (B) alumina (spherical), (C) silica gel, (D) 5A zeolite, and (E) activated carbon. The vapor pressure at 100% relative humidity is 23.6 torr. (From Ref.[3].)

Table 1 Representative physical properties of commercial adsorbents

Adsorbent	Nature	Specific surface area (m^2/g)	Pore diameter (Å)	Porosity	Particle density (g/cm^3)
Activated carbon	Hydrophobic amorphous				
Small pore		400–1200	10–25	0.4–0.6	0.5–0.9
Large pore		200–600	>30	$\sim$0.5	0.6–0.8
Zeolite	Hydrophilic/hydrophobic crystalline	600–700	3–10	0.6	1.0
Activated alumina	Hydrophilic crystalline/x-ray amorphous	200–350	10–75	0.5	1.25
Silica gel	Hydrophilic/hydrophobic amorphous				
Small pore		750–850	22–26	0.47	1.09
Large pore		300–350	100–150	0.71	1.62
Polymeric adsorbent	Hydrophilic/hydrophobic	450–1100	25–90	0.5	1.25
Carbon molecular sieve	Hydrophilic	$\sim$400	3–9	0.5	1.0

adsorption (TSA) process. However, activated alumina has a low adsorption heat for water and other polar molecules as compared with zeolite; it is possible to regenerate activated alumina under PSA conditions. Activated alumina also demonstrates moderate adsorption affinity for carbon dioxide, which makes it a suitable sorbent for removing water and carbon dioxide from air in a PSA process. These adsorption properties of activated alumina have been explored extensively for air purification applications by industrial gas companies.[14–17] This is a perfect example to demonstrate the importance of sorbent regenerability over sorption capacity and selectivity in pressure swing adsorption processes. Activated alumina is also an excellent catalyst support. More applications and representative properties of activated alumina are listed in Tables 1 and 2.

Silica Gels

Silica gel is the most widely used desiccant because of its large adsorption capacity for water (40 wt%), as shown in Fig. 2, and easy for regeneration ($\sim$150°C, compared with 350°C for zeolites). Silica is a partially dehydrated polymeric form of colloidal silicic acid with the formula of $SiO_2 \cdot nH_2O$. Its water content, which is typically about 5 wt%, is presented in the chemically bonded hydroxyl groups. Silica is an amorphous material comprising spherical particles of 20–200 Å in size, which aggregate to form the sorbent with pore sizes in the range of 60–250 Å and surface areas of 100–850 m^2/g, depending on gel density. Its surface has mainly Si–OH and Si–O–Si polar groups; this is why it can be used to adsorb water, alcohols, phenols, amines, etc. by hydrogen bonding mechanisms. Other commercial applications include the separation of aromatics from paraffins, the chromatographic separation of organic molecules, and modified silica in chromatography columns.[2,18–20]

Polymeric Adsorbents

A wide range of synthetic, nonionic polymers are available for use as sorbents, ion-exchange resins, and particularly for analytical chromatography applications. Commercially available resins in bead form (typically 0.5 mm in diameter) are based usually on copolymers of styrene/divinyl benzene (DVB) and acrylic acid esters/divinyl benzene, and have a wide range of surface polarities, porosities, and macropore sizes. The porosities can be built through emulsion polymerization of relevant monomers in the presence of a solvent that dissolves the monomers and serves as a poor swelling agent for the polymer. This creates a polymer matrix with surface areas ranging up to 1100 m^2/g.[2,4] The major application of polymeric adsorbents is in water treatment. The macroporous polymeric resins can be modified by attaching different functional groups to mimic activated carbon, and to replace activated carbons for certain specific applications in food and pharmaceutical industries where color contamination by the black carbons of the final products is a major concern.

NEW DEVELOPMENTS IN SORBENT MATERIALS AND APPLICATIONS

The past two decades have witnessed major advances in new nanostructured sorbent materials including

Table 2 Selected applications of commercial sorbents

Adsorbent	Applications (the first molecule is the product)
Activated carbon	Hydrogen separation from syngas and hydrogenation processes Ethylene from methane and hydrogen Vinyl chloride monomer (VCM) from air Removal of odors from gases Recovery of solvent vapors Removal of SO_x, and NO_x Purification of helium Clean-up of nuclear off-gases Decolorizing of syrups, sugars, and molasses Water purification, including removal of phenol, halogenated compounds, pesticides, caprolactam, chlorine
Carbon molecular sieve	Nitrogen separation from air
Zeolite	Oxygen from air Drying of gases Removing water from azeotropes Sweetening sour gases and liquids Purification of hydrogen Separation of ammonia and hydrogen Recovery of carbon dioxide Separation of oxygen and argon Removal of acetylene, propane, and butane from air Separation of xylenes and ethyl benzene Separation of normal from branched paraffins Separation of olefins and aromatics from paraffins Recovery of carbon monoxide from methane and hydrogen Purification of nuclear off-gases Separation of cresols Drying of refrigerants and organic liquids Separation of solvent systems Pollution control, including removal of Hg, NO_x, and SO_x from gases Recovery of fructose from corn syrup
Activated alumina	Drying of gases, organic solvents, transformer oils Removal of HCl from hydrogen Removal of fluorine and boron–fluorine compounds in alkylation processes Removing of water and carbon dioxide from air in a PSA process
Silica gel	Drying of gases, refrigerants, organic solvents, transformer oils Desiccant in packings and double glazing Dew point control of natural gas
Polymeric adsorbents	Water purification, including removal of phenol, chlorophenols, ketones, alcohols, aromatics, aniline, indene, polynuclear aromatics, nitro- and chlor-aromatics, polychlorinated biphenyls (PCBs), pesticides, antibiotics, detergents, emulsifiers, wetting agents, kraftmill effluents, dyestuffs, and radionuclides Recovery and purification of steroids, amino acids and polypeptides Separation of fatty acids from water and toluene Separation of aromatics from aliphatics Separation of hydroquinone from monomers Recovery of proteins and enzymes Removal of colors from syrups Removal of organics from hydrogen peroxide
Clays (acid treated and pillared)	Removal of organic pigments Refining of mineral oils Removal of PCBs

(From Ref.[4].)

mesoporous molecular sieves, sol–gel-derived metal oxide xerogels and aerogels, metal organic framework, π-complexation and composite adsorbents, new carbonaceous materials (carbon nanotubes, carbon fibers, superactivated carbons), high-temperature ceramic sorbents, and strong chemical sorbent materials. Although these new sorbent materials have demonstrated promising sorption properties for many existing and new applications, systematic studies on synthesis methods and characterization of these new materials are necessary to fully explore and realize their potential as commercial sorbents. The review that follows aims at attracting more research efforts to develop novel sorbent materials to meet the increasing needs of new energy, environmental, and other emerging technologies.

Oxide Molecular Sieves

Microporous and mesoporous oxide molecular sieves that have the characteristics of large internal surface area and pore volume are ideal candidates for use as sorbent materials and catalyst supports of many heterogeneous catalysts. Oxide molecular sieves are generally synthesized by hydrothermal methods that involve both chemical and physical transformations within an amorphous oxide gel, often in the presence of a template species. The gel eventually converts to a crystalline material in which the template species and/or solvent molecules are guests within the channels and cages of an oxide host framework. A porous material is obtained upon removal of the guest molecules from the oxide framework. By manipulating the synthesis parameters, including starting precursors, synthesis temperature, pH, template species, drying, and calcination conditions, it is possible to tailor the pore size and shape of these porous materials for different applications. However, tailoring of porosity in oxide molecular sieves in terms of a priori structural design is extremely difficult because of the inherent complexity of the synthetic procedures employed.[21]

Recent advances and applications of oxide molecular sieves have been summarized in several review articles.[2,21–23] Microporous zeolite materials synthesized with molecular templates and their applications in host–guest chemistry have been covered elsewhere.[13] A new class of silicate/aluminosilicate mesoporous molecular sieves designated as M41S was discovered in the former Mobil research laboratory by extending the concept of zeolite templating with small organic molecules to large long-chain surfactant molecules.[24] A representative member of this family is MCM-41, which has a honeycomb-shaped hexagonal arrangement of uniform mesopores in the range of 15–100 Å, specific surface area of $1040\,m^2/g$, pore volume above $0.7\,cm^3/g$, and significantly high sorption capacity for hydrocarbons (49 wt% for *n*-hexane at 40 torr and 21°C, and 67 wt% for benzene at 50 torr and 25°C).[24] Other significant members of the M41S family include MCM-48 (cubic phase), MCM-50 (stabilized lamellar phase), SBA-1 (cubic phase), and SBA-2 (cubic phase).[21]

Although M41S type mesoporous oxide molecular sieves have exhibited unique properties of large surface area and exceptionally large pore volume ($> 0.7\,cm^3/g$), their large pore volume may not be attractive for gas sorption because the adsorbate–adsorbent interactions are not enhanced inside the internal pores of these materials.[2] Therefore, M14S type mesoporous oxide molecular sieves without surface modification are rarely used as sorbents. Significant research efforts were devoted to surface modification of M41S materials for different applications.[2] An amine-grafted MCM-48 sorbent, synthesized from tetraethoxysilane (TEOS), has been shown to have a surface area of $1389\,m^2/g$, a silanol number of 8, higher thermal stability than MCM-41, high adsorption selectivity, and high capacity for both carbon dioxide and hydrogen sulfide.[25]

Sol–Gel-Derived Xerogels and Aerogels

Sol–gel processing refers to the fabrication process of ceramic materials by preparation of a sol, gelation of the sol, and removal of the solvent.[26] Sols are dispersions of colloidal particles in a liquid solvent, and a gel is a solid matrix encapsulating a solvent. In a sol–gel process, the sol can be formed from a solution of colloidal powders or hydrolysis and condensation of alkoxides or salt precursors. In the latter approach, which is much more popular, primary particles of uniform size are formed and grow in a sol and connect to each other to form aggregates during gelation. These aggregates forming the network of the gel are broken apart into the primary particles in the drying step. Upon calcination and sintering, these primary particles are bound together strongly to form a very rigid solid network, and large interparticle space with uniform nanoscale pores is formed. *Xerogels* are obtained by drying the gels through evaporation at normal conditions under which capillary pressure causes shrinkage of the gel network, while *areogels* are produced by drying the wet gels at supercritical conditions where the liquid–vapor interface is eliminated, and relatively little shrinkage of the gel network occurs. Xerogels and aerogels typically have relatively large surface area, high porosity, and internal pore volume, and are ideal candidates as sorbent and catalyst support materials for many applications. The sol–gel process offers a very high flexibility to tailor xerogels and areogels for specific applications by manipulating the synthesis conditions.

Silica xerogel is probably the most studied and documented porous material in the sol–gel system.[27,28] Although silica has several crystalline forms, only amorphous silica gel is used as a desiccant (sorbent). A microporous silica that was synthesized with TEOS as precursor has an average pore size of 6.4 Å, pore volume of 0.24 cm^3/g, and Brunauer–Emmett–Teller (BET) surface area of 588 m^2/g.[29] However, this material lost about 90% of its microporosity when it was heated at 600°C for 30 hr. By doping with 1.5% of alumina, the thermal stability of this microporous silica was significantly improved.[29]

Crystalline sorbent materials including γ-alumina, zirconia, and titania were also synthesized using the sol–gel process in Lin's group;[29] the representative pore size distribution and pore texture data of xerogels of γ-alumina, zirconia, and titania are summarized in Fig. 3 and Table 3, respectively. As shown in Fig. 3, the pore size distributions of these materials are rather narrow, with an average pore diameter of about 3 nm. Such narrow size distribution and nanoscale average pore size are determined by the primary crystallite particles. The particles of the sol–gel-derived alumina, titania, and zirconia, owing to the Ostwald ripening mechanism,[26] are usually of nanoscale size; the uniform particle size distributions of γ-alumina crystallites are plate-shaped with size ranging from about 5 to 20 nm. The sol–gel-derived γ-alumina consists of such plate-shaped crystallite particles, which give rise to a relatively large surface area. Crystallites of tetragonal zirconia and rutile are of more spherical shape, with a crystallite size of about 15 and 11 nm, respectively.[29]

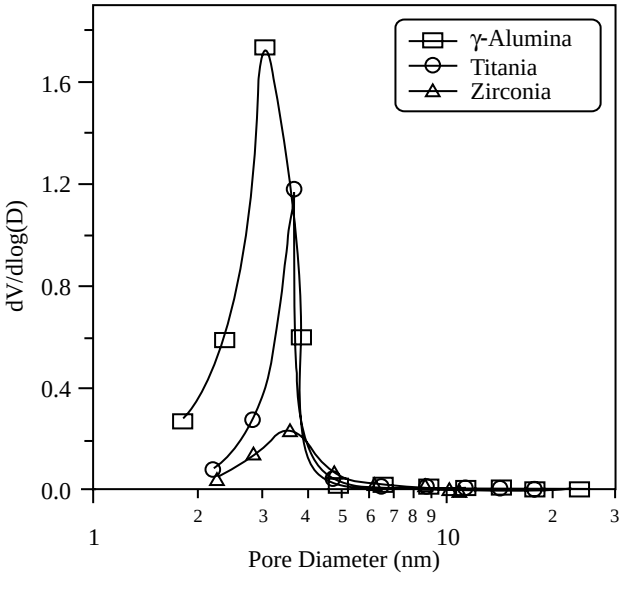


Fig. 3 Pore size distribution of sol–gel-derived alumina, zirconia, and titania. (From Ref.[29].)

Table 3 Pore texture sol–gel-derived alumina, zirconia, and titania (calcined at 450°C for 3 hr)

Xerogel	Average pore size (Å)	Pore volume (cm^3/g)	BET surface area (m^2/g)
γ-Al_2O_3	28	0.33	373
ZrO_2	38	0.11	57
TiO_2	34	0.21	147

One of the outstanding characteristics of sol–gel-derived γ-alumina xerogel is its excellent mechanical properties. Preparation of porous γ-alumina granules with good mechanical properties and desirable pore structure is of great importance in the development of novel catalysts and sorbents for various applications. The superior mechanical properties can be derived from the unique microstructure of the granule, which is defined by compacting small γ-alumina crystallite particles bound together by the bridges of the same material formed through coarsening or sintering. Such nanostructured γ-alumina can be prepared by combining the Yoldas process and the ''oil-drop'' method.[30–34] Table 4 compares the crush strength and attrition rate of sol–gel-derived γ-alumina xerogel granules with those of several commercial sorbents. It is clearly shown in Table 4 that the sol–gel-derived γ-alumina xerogel granules have excellent mechanical properties as compared with commercial sorbents. The excellent mechanical properties makes sol–gel-derived alumina granules very suitable for fluidized bed and other applications including separation and purification process for food and healthcare products that have very strict regulations on sorbent power contamination.

Sol–gel-derived xeorgel sorbents have been investigated for gas separation, purification, and environmental applications. γ-Alumina sorbents and membranes doped with cuprous and silver ions have been studied for selective adsorption or transfer of CO and ethylene through π-complexation.[35–37] Significant efforts have been devoted to explore the possibility of using CuO-doped γ-alumina sorbents for removing SO_x and NO_x from flue gas.[38–44] The sol–gel-derived CuO/γ-alumina sorbents have demonstrated high sorption capacity, high reactivity for SO_2, and high thermal and chemical stability. The excellent mechanical and desulfurization properties of sol–gel-derived sorbents make them ideal sorbent candidates for fluidized bed desulfurization process. However, the relatively high cost of sol–gel-derived alumina xerogels may prevent them from being used in many large-scale adsorption processes. Research efforts are needed to look for less expensive precursors to replace alkaoxides used in the Yoldas process.

Table 4 Comparison of crush strength and attrition rate of sol–gel-derived γ-alumina xerogel granules with commercial sorbents

Sorbents	Granular shape	Granular size (mm)	Average crush strength (N/granule)	Attrition rate (wt%/hr)
Sol–gel alumina	Spherical	2.0–2.5	160	0.033
Sol–gel alumina	Spherical	2.6–2.8	190	
Alcoa alumina (LD-350)	Spherical	4.0–4.6	42	0.177
UOP silicalite	Cylindrical	1.4–1.6	16	0.575
Degussa DAY zeolite	Cylindrical	3.5–3.5	40	0.073

Sol–gel-derived metal oxide xerogels were also investigated for water adsorption because most of these metal oxides are good sorbent candidates for desiccant applications.[45–47] Significant research works have been carried out to study the adsorption/complexation properties of heavy metal ions including mercury, Cu(II), $CdCl_2$, etc. in waste water on different sol–gel-derived xerogels.[48–54] The sol–gel-derived xerogels seem to be promising sorbent candidates for waste water treatment. Modified xerogel sorbents also showed promising adsorption properties for removing acid gas CO_2 and H_2S from natural gas, or as CO_2 storage sorbents.[25,55] There are several advantages of using xerogels for enzyme immobilization, including the opportunity to produce them in defined shapes or thin films and the ability to manipulate their physical characteristics including porosity, hydrophobicity, and optical properties.[56,57] Metal oxide composite xerogels can also adsorb methyl orange.[58] There are also reports on microporous and mesoporous carbon xerogels for gas separation and purification.[59,60]

As compared with xerogels, aerogels have larger surface area, larger pore volume, and higher porosity.[61–64] Alumina aerogels with a specific surface area as high as $1000\,m^2/g$, and a pore volume as high as $17.3\,cm^3/g$ have been synthesized by supercritical carbon dioxide drying, but a very limited information on their adsorption properties was found.[61,62] A super water adsorbent consisting of 17–30% of $CaCl_2$ doped on SiO_2 aerogel showed an effective reversible adsorption capacity of 100 wt%; the adsorption capacity of hydrophilic silica aerogels can be fully recovered after regeneration.[64–66] CaO- and MgO-modified SiO_2 aerogel sorbents can be used to capture pollution gases including CO_2, SO_2, CO, and NO_x emitted from power plants based on fossil fuels.[67] Several studies reported the use of aerogels as destructive sorbents for toxic gases and radionuclide removal from contaminated environments.[64,68–70] Carbon aerogels can also be made from carbon materials under supercritical carbon dioxide drying conditions; these carbon aerogels were studied for removing uranium and other inorganic ions from contaminated water.[71–73] Aerogels are special sorbent candidates with excellent pore texture, which may play a major role in environmental protection. However, more studies on their synthesis and adsorption properties are needed.

Metal Organic Framework (MOF)

Recently, Yaghi's group reported a novel crystalline nanoporous material that consists of metal atoms occupying the vertices of a lattice, with the lattice size, porosity, and chemical environment defined by the organic linker molecules that bind the metal atoms into a robust periodic structure.[74–76] These so-called metal organic framework (MOF) materials have been demonstrated to have an exceptionally high specific surface area of $4526\,m^2/g$, and find use as adsorbents for methane and as hydrogen storage materials.[74,77–80] A reticular synthesis method was developed to realize the bottom-up synthesis through top-down design logic by using inorganic, metal organic, and organic molecules to build frameworks and large molecules.[81] Well-defined molecular building blocks that will maintain their structural integrity throughout the construction process were used to build the MOF molecules. It allows remarkable control over composition and structure of the material formed and employs the full range of the molecular synthetic methods and compounds in the preparation of this new type of porous sorbent materials. The ability to molecularly engineer the lattice size, chemical environment, and possibly structure by careful choice of the metal centers and organic linkers offers the opportunity for the development of new types of sorbents that could potentially meet the Department of Energy (DOE) target for hydrogen storage and that can be used for other applications in separation and purification.

It is reported that metal organic framework-5 (MOF-5) of composition $Zn_4O(BDC)_3$ (BDC: 1,4-benzenedicarboxylate) with a cubic three-dimensional extended porous structure and octahedral Zn–O–C clusters with benzene links can adsorb hydrogen up to 4.5 wt% at 78 K, and 1.0 wt% at room temperature and pressure of 20 bar.[74,79] It is identified by inelastic neutron scattering spectroscopy of the rotational

transitions of the adsorbed hydrogen molecules that zinc and the BDC linker in MOF-5 are the two hydrogen binding sites responsible for hydrogen adsorption on this material. Higher hydrogen adsorption capacity at ambient temperature and 10 bar were observed on similar isoreticular metal organic framework-6 and -8 (IRMOF-6 and -8) having cyclobutylbenzene and naphthalene linkers.[79] A different microporous MOF sorbent [microporous metal coordination materials (MMOM)] was reported to have hydrogen sorption capacities (~1.0 wt% at room temperature and 48 bar) similar to those of the best single-wall carbon nanotubes.[80] The adsorbed hydrogen can be released when the gas pressure is reduced.

MOF sorbents have also been investigated for methane adsorption.[77] The reported methane storage capacity of MOF-6 is 155 cm^3 (STP)/cm^3 at 298 K and 36 atm, which is significantly higher than that of zeolite 5A (87 cm^3 (STP)/cm^3) and other coordination framework (213 cm^3 (STP)/cm^3).[77] Adsorption and desorption of carbon dioxide, nitrogen, and argon on a microporous manganese-based MOF sorbent has been reported.[78] Another interesting porous MOF sorbent, Cu-BTC (polymeric copper(II) benzene-1,3,5-tricarboxylate) with molecular sieve character, was studied for its sorption properties of various adsorbates including nitrogen, oxygen, carbon monoxide, carbon dioxide, nitrous oxide, methane, ethylene, ethane, and *n*-dodecane.[82,83] A detailed investigation of sorption thermodynamics was performed for carbon dioxide by a sorption-isosteric method. It was demonstrated that Cu-BTC sorbent can be used for the separation of carbon dioxide–carbon monoxide, carbon dioxide–methane, and ethylene–ethane mixtures. In addition, this sorbent can also be used to remove carbon dioxide, nitrous oxide, high molecular weight hydrocarbons, and moisture from ambient air before cryogenic separation to produce oxygen and nitrogen.[82]

Hydrogen Storage Media

The development of hydrogen-fueled transportation system and portable electronics will demand new materials that can store large amounts of hydrogen at ambient temperature and relatively low pressures with small volume, light weight, fast charging and discharging time, cyclic stability, and low cost. Table 5 summarizes the targets for hydrogen storage system for automotive applications set by USDOE. The hydrogen storage capacities are calculated as both weight and volume percentage of the storage system.[84] To achieve these goals, the hydrogen storage media (sorbent) should have a high reversible hydrogen sorption capacity, low weight and high packing density as well as fast sorption/desorption kinetics, and low cost.

Table 5 USDOE FreedomCAR hydrogen storage system targets

	Year		
Target factor	**2005**	**2010**	**2015**
Specific energy (MJ/kg)	5.4	7.2	10.8
Hydrogen (wt%)	4.5	6.0	9.0
Energy density (MJ/L)	4.3	5.4	9.72
System cost ($/kg/system)	9	6	3
Operating temperature (°C)	−20/50	−20/50	−20/50
Cycle life-time (adsorption/desorption cycles)	500	1000	1500
Flow rate (g/sec)	3	4	5
Delivery pressure (bar)	2.5	2.5	2.5
Transient response (sec)	0.5	0.5	0.5
Refueling rate (kg H_2/min)	0.5	1.5	2.0

(From Ref.[84].)

Hydrogen can be stored both physically and chemically in a confined vessel with or without the assistance of a storage media. The most commonly used methods for hydrogen storage are: gaseous and liquid hydrogen storage, solid state storage in complex metal hydrides, chemical storage materials, and in nanostructured materials.[2,85] The representative hydrogen storage capacities, hydrogen storage, and release conditions in various materials are summarized in Table 6.

Carbon nanotubes are probably the most investigated and documented hydrogen storage sorbent materials. Several excellent reviews on carbon nanotubes for hydrogen storage are available.[2,86] As shown in Table 6, the hydrogen storage capacities on representative carbon nanotubes are below 6 wt%, the most referred DOE target for 2010.[84,87,88] The following concerns about carbon nanotubes as hydrogen storage materials have driven research in this area to other directions:[85]

1. Difficult to meet the DOE's long-term target (9 wt%);
2. Mechanisms for hydrogen sorption in carbon nanotubes are not well understood;
3. Part of the adsorbed hydrogen can only be recovered at high temperatures;
4. Preparation and purification of carbon nanotubes involve complicated and expensive processes, which leads to high cost of carbon nanotubes;
5. Hydrogen storage capacity is quite sensitive to sorbent preparation conditions;
6. Mixed results on hydrogen adsorption capacity have been reported.

Table 6 Summary of hydrogen storage capacity of various nanostructured materials

Materials	H_2 storage capacity (wt%)	H_2 storage conditions	H_2 release conditions	Reference
Carbon nanotubes				
Single-walled	4.2	10 MPa, 300 K	1 bar, 300 K	[87]
Multi-walled	3.6	7 MPa, 298 K	1 bar, 298 K	[88]
Non-carbonaceous nanotubes				
BN	4.2	108 bar, 298 K	1 bar, 298 K	[89]
TiS_2	2.5	40 bar, 298 K	1 bar, 298 K	[90]
Microporous MOF				
MOF-5	4.5	0.75 bar, 77 K	1 bar, >77 K	[74]
MMOM	$\sim$1.0	48 bar, 298 K	1 bar, 298 K	[80]
Metal hydrides				
Mg_2NiH_4	3.6	1 bar, $\sim$500 K	1 bar, >528 K	[91]
$NaAlH_4$	8.0	90 bar, 403 K		[92]
$Mg(AlH_4)_2$	6.6		1 bar, >436 K	[93]
$LiBH_4$	13.3		1 bar, >473 K	[94]
Nitrides				
Li_3N	9.3	1 bar, 443–473 K	1 bar, >700 K	[95]
Clathrate/molecular compounds				
$H_2(H_2O)_2$	5.3	$\sim$1 bar, 77 K	1 bar, >77 K	[97]
$H_2(H_2O)$	10.0	$\sim$6000 bar, 190 K	<6000 bar, >190 K	[97]
$(H_2)_4(CH_4)$	33.3	$\sim$2000 bar, 77 K	<2000 bar, >77 K	[97]

Noncarbonaceous nanotubes including boron nitride (BN) and titanium sulfide (TiS_2) have been prepared and studied for hydrogen sorption.[89,90] Hydrogen storage capacity (2.5–4.5 wt%) similar to those for carbon nanotubes have been obtained on these noncarbonaceous materials. MOF-based sorbents for hydrogen sorption was discussed in the previous section. As suggested in Table 6, the hydrogen sorption capacities on MOF-5 and MMOM are lower than those on carbon nanotubes. However, MOF sorbents look more promising than carbon nanotubes as hydrogen storage media for the following reasons:

1. MOF is easy to make and is less expensive;
2. Sorption sites for hydrogen on MOF are better defined;
3. MOF sorbents may have extremely high specific surface area ($>4000\,m^2/g$);
4. It is possible to tailor the interaction between hydrogen and MOF by manipulating synthesis parameters including different building blocks.

Metal hydrides were widely investigated for hydrogen storage, and are believed to be ideal hydrogen storage system because they have the following characteristics:[2,84,91–94]

1. Relatively high hydrogen storage capacity at modest pressures as indicated in Table 6;
2. Fast hydrogen charging and discharging rates; and
3. Moderate temperature for hydrogen desorption.

However, metal hydrides also suffer from the following disadvantages as hydrogen storage materials:

1. High sensitivity to impurities in hydrogen (CO, H_2O, O_2, CO_2, and H_2S);
2. Storage capacity and rates decay with hydrogen charge–discharge cycles; and
3. Relatively high cost as compared with gaseous and liquid hydrogen storage methods.

Another interesting hydrogen storage material is lithium nitride (Li_3N), which shows 9.3 wt% useful hydrogen storage capacity between thermal swing cycles (473–700 K).[95] The requirement for high-temperature desorption will greatly limit its applications. Most recently, hydrogen clathrate hydrate and other

molecular compounds were found to have hydrogen storage capacities as high as 33.3 wt%.[96–98] This is a very innovative way to store hydrogen with exceptionally high capacity to meet the DOE long-term target. However, these clathrate and hydrogen storage compounds were synthesized at extremely high pressures and at liquid nitrogen temperature. It is unlikely these clathrate hydrates will be used for hydrogen storage until we find new clathrate hydrate compounds that can be synthesized and are stable at much lower pressures.

π-Complexation Sorbents and Composite Sorbents

A very good review article based on a panel study of status, future research needs, and opportunities for porous sorbent materials was published several years ago.[99] It was pointed out that very significant advances have been made in tailoring the porosity of porous sorbent materials in terms of size and shape selectivity. Relatively little progress has been achieved in terms of chemoselectivity of sorbents based on specific interactions between adsorbate molecules and functional groups in the sorbents. Incorporation of active sites into sorbents is of high priority in the development of sorbents.

The π-complexation bond is a weak chemical bond that is slightly stronger than van der Waals interaction, which governs physical sorption processes. Sorbents with π-complexation capability tend to have higher selectivity than other physical sorbents for certain adsorbate molecules. Several different types of π-complexation sorbents with Cu^+ or Ag^+ ions supported on different supports (SiO_2, γ-Al_2O_3, TiO_2, variety of zeolites, polymer resin, and activated carbon) were synthesized using different methods including thermal dispersion, wet-impregnation, sol–gel, microwave heating, ion-exchange zeolite, and ion-exchange resin.[34–36,99–105] It was found that the CO adsorption capacity increases with Cu^+ loading in an activated alumina supported sorbent.[100,101] To achieve the highest sorption capacity, the active species should be dispersed as a monolayer form.[99] The potential applications of these π-complexation sorbents include:[2]

1. Desulfurization of gasoline and diesel fuels;
2. Separation of olefins and paraffins;
3. CO separation from synthesis gases;
4. CO removal from hydrogen;
5. Removal of aromatics; and
6. Removal of volatile organic compounds (VOCs).

A π-complexation sorbent can also be viewed as a composite sorbent especially when the sorbent support contributes significantly to the adsorption. Composite sorbents are typically made by physically mixing the powders of constituent sorbents with different sorption properties; they tend to have multiple sorption sites for different adsorbate molecules. One example of a composite sorbent is a mixture of activated alumina and zeolites for removing moisture, carbon dioxide, and other trace components from air in an air-purification process prior to cryogenic air separation.[106–108] Conventionally, moisture is removed by activated alumina, carbon dioxide by zeolite 13X, and hydrocarbons by zeolite 5A.[107,108] Traditional air-purification processes employ multiple layers consisting of activated alumina, zeolite 13X, and optional zeolite 5A sorbents in a single vessel to achieve significant removal of moisture, carbon dioxide, and hydrocarbons from air. The major disadvantages of layered bed are nonuniform sorbent packing for a short sorbent layer, very significant temperature variation ($> 30°C$, sometimes called cold spots) between the zeolite and the activated alumina sorbent layers. The large temperature difference could upset the sorption process operation if it is designed to be operated isothermally. It is beneficial to have a single sorbent with multiple sorption features for different impurities and eliminate sorbent layering and temperature variations.

High-Temperature Ceramic O_2 Sorbents

Lin et al. disclosed in a U.S. patent a new group of sorbents for air separation and oxygen removal using oxygen-deficient perovskite-type ceramics as sorbents.[109] Perovskite-type ceramics are a group of metal oxides having the general formula of ABO_3. The ideal perovskite structure for ABO_3 is shown in Fig. 4. It consists of cubic array of corner-sharing BO_6 octahedra, where B is a transition metal ion. The A-site ion, interstitial between the BO_6 octahedra, may be occupied by an alkali, an alkaline earth, or a rare earth ion. Alternatively, the perovskite structure may be regarded as a cubic close packing of layers of AO_3 with B cations placed in the interlayer octahedral interstices.[110] This group of the sorbents can be viewed as chemisorbents that can selectively adsorb a considerable amount of oxygen at high temperatures ($> 300°C$), and theoretically has an infinitely high selectivity for oxygen over nitrogen or other nonoxygen species. The presence of other gases has negligible effect on the separation properties of these new sorbents. High-temperature membrane separation of oxygen has also received increasing interest from other industrial gas companies.[111–113] Development of high-temperature oxygen separation technology opens up several high-temperature applications of oxygen including syngas production, hydrogen production, and partial oxidation fuel reforming processes.

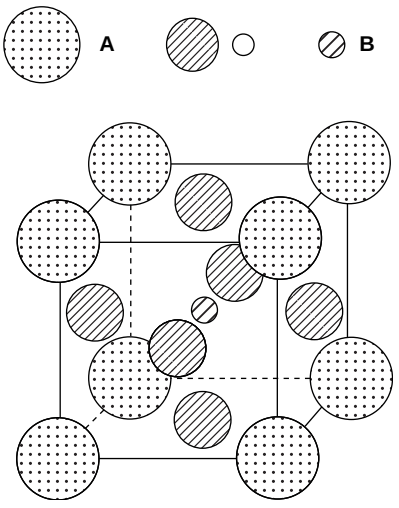


Fig. 4 Ideal perovskite structure for ABO_3 type oxides.

The oxygen equilibrium and kinetic properties of perovskite-type ceramics have been extensively studied primarily for applications as fuel cell electrodes and oxygen permeable membranes,[110] and only a few for oxygen sorption.[114–117] Oxygen nonstoichiometry (δ) occurs in some perovskite-type ceramics with B-site cations of variable oxidation states and A-site cations partially substituted by another cation with a lower oxidation state. Oxygen nonstoichiometry, or oxygen content, for a perovskite-type ceramic of a given composition is a function of temperature and oxygen partial pressure. Therefore, by changing temperature or oxygen partial pressure, the value of oxygen nonstoichiometry or the degree of oxygen vacancy in the material changes. Within a certain range of temperature and oxygen partial pressure the change of the oxygen nonstoichiometry does not affect the perovskite structure, and the change of the oxygen content in the material is a reversible process. The oxygen nonstoichiometry of the perovskite sorbents can be measured gravimetrically at different temperatures and oxygen partial pressures. Oxygen sorption capacity on the sorbent can then be calculated from the oxygen nonstoichiometry data once the initial state (zero sorption capacity) of the sorbent material is defined.[114] Figs. 5 and 6 are examples of oxygen nonstoichiometry of $La_{1-x}Sr_xCo_{1-y}Fe_yO_{3-\delta}$ perovskite oxide sorbents as a function of oxygen partial pressure or temperature, respectively.[114] The corresponding oxygen sorption isotherm of $La_{1-x}Sr_xCo_{1-y}Fe_yO_{3-\delta}$ perovskite oxide sorbents that were calculated from the oxygen nonstoichiometry data are shown in Fig. 7.[114] From these oxygen isotherms we can conceive a high-temperature vacuum swing sorption or temperature swing sorption process for oxygen separation or oxygen removing applications by using the $La_{1-x}Sr_xCo_{1-y}Fe_yO_{3-\delta}$ perovskite oxide sorbents. Future studies on perovskite oxide sorbents are needed to address the issues of slow desorption rate, potential sorbent structure stability in cyclic processes, and effective regeneration methods.

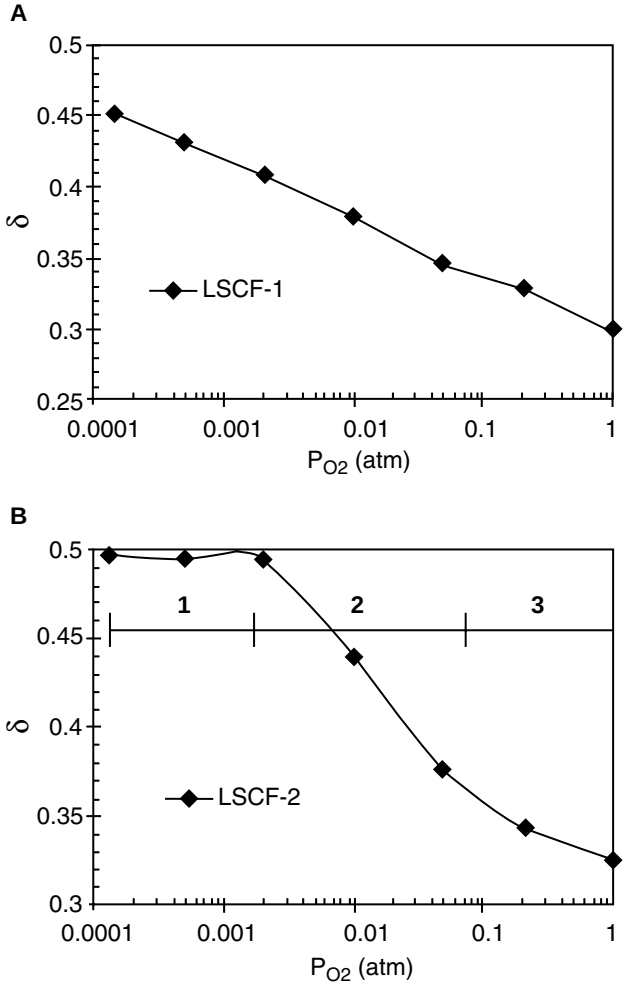


Fig. 5 Change of oxygen nonstoichiometry δ with oxygen partial pressure (LSCF-1. $La_{0.1}Sr_{0.9}Co_{0.5}Fe_{0.5}O_{3-\delta}$; LSCF-2, $La_{0.1}Sr_{0.9}Co_{0.9}Fe_{0.1}O_{3-\delta}$). (From Ref.[114].)

High-Temperature CO_2 Sorbents

Increased awareness of the global warming trend has led to worldwide concerns regarding ''greenhouse gas'' emissions. Greenhouse gases include CO_2, CH_4, and N_2O and are mostly associated with the production and utilization of fossil fuels, with CO_2 being the single greatest contributor to global warming. Significant research efforts are being devoted worldwide on looking for economical ways of mitigating CO_2 emission problem.[118–122] Carbon capture and sequestration costs can be considered in terms of four components: capture, compression, transport, and injection. Typically about 75% of this cost is attributable to capture

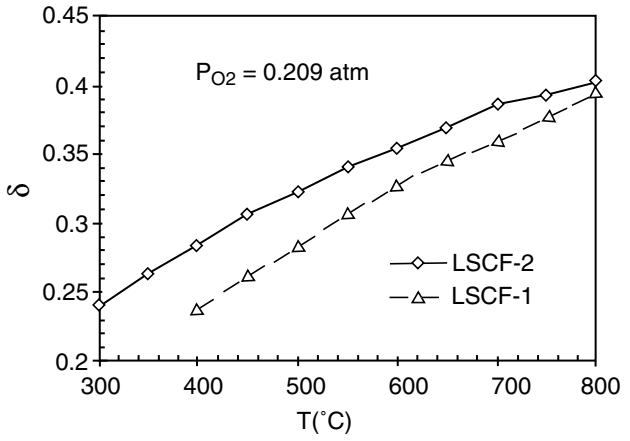


Fig. 6 Change of oxygen nonstoichiometry δ with temperature (LSCF-1, $La_{0.1}Sr_{0.9}Co_{0.5}Fe_{0.5}O_{3-\delta}$; LSCF-2, $La_{0.1}Sr_{0.9}Co_{0.9}Fe_{0.1}O_{3-\delta}$). (From Ref.[114].)

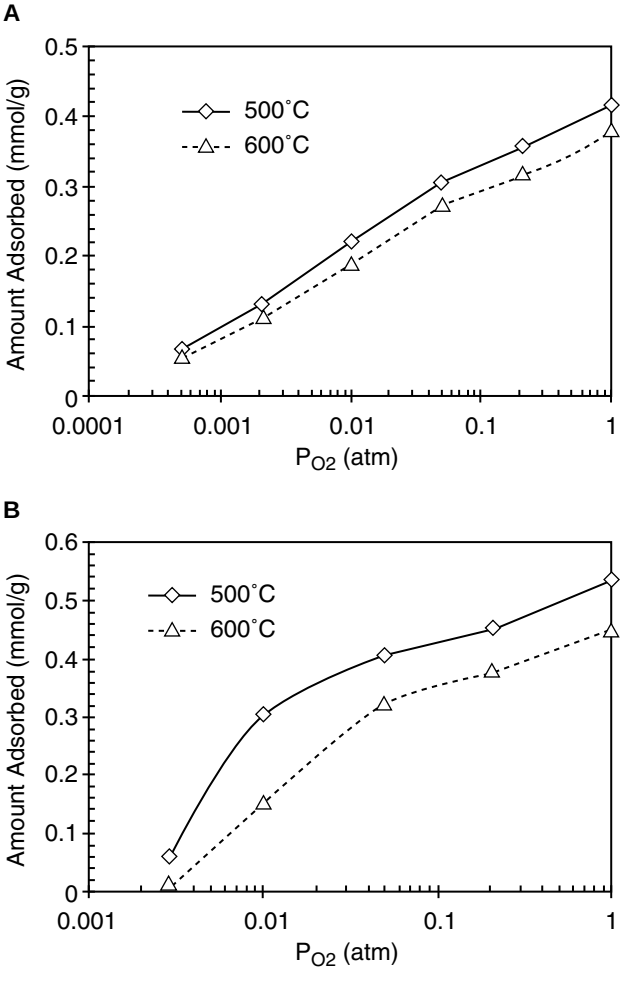


Fig. 7 Sorption isotherms of: (A) $La_{0.1}Sr_{0.9}Co_{0.5}Fe_{0.5}O_{3-\delta}$ and (B) $La_{0.1}Sr_{0.9}Co_{0.9}Fe_{0.1}O_{3-\delta}$ at 500 and 600°C. (From Ref.[114].)

and compression processes. Sorption of carbon dioxide on solid sorbents is receiving increased attention in view of the importance of both the removal and the recovery of carbon dioxide from flue gases.[123,124]

Physical sorbents for carbon dioxide separation and removal were extensively studied by industrial gas companies.[125–127] Zeolite 13X, activated alumina, and their improved versions are typically used for removing carbon dioxide and moisture from air in either a TSA or a PSA process.[125–128] The sorption temperatures for these applications are usually close to ambient temperature. There are a few studies on adsorption of carbon dioxide at high temperatures. The carbon dioxide adsorption isotherms on two commercial sorbents hydrotalcite-like compounds, EXM911 and activated alumina made by LaRoche Industries, are displayed in Fig. 8.[123,124] As shown in Fig. 8, LaRoche activated alumina has a higher carbon dioxide capacity than the EXM911 at 300°C. However, the adsorption capacities on both sorbents are too low for any practical applications in carbon dioxide sorption at high temperature. Conventional physical sorbents are basically not effective for carbon dioxide capture at flue gas temperature (>400°C). There is a need to develop effective sorbents that can adsorb carbon dioxide at flue gas temperature to significantly reduce the gas volume to be treated for carbon sequestration.

Only a handful of studies on high-temperature carbon dioxide sorbents have been published in the past few years.[123,124,129–133] It is believed that lithium zirconate (Li_2ZrO_3) is one of the most promising sorbent materials for carbon dioxide separation from flue gas at high temperature because it can absorb a large amount of carbon dioxide at around 400–700°C.[130,131] The carbon dioxide adsorption and desorption uptake curves on lithium zirconate are shown in Fig. 9.[131] As shown in this figure, about 20% carbon dioxide was captured by the lithium zirconate sorbent during sorption step at 500°C based on the following reaction:

$$Li_2ZrO_3 + CO_2 \rightarrow Li_2CO_3 + ZrO_2 \quad (4)$$

About 80% of adsorbed carbon dioxide can be desorbed with hot air 780°C. Addition of potassium carbonate (K_2CO_3) and Li_2CO_3 into Li_2ZrO_3 remarkably improves the CO_2 sorption rate of the Li_2ZrO_3-based sorbent materials. X-ray diffraction (XRD) analysis for phase and structural changes during the sorption/desorption process shows that the reaction between Li_2ZrO_3 and CO_2 is reversible.[131] Based on this work, a TSA process can be developed for carbon dioxide removal from flue gas using Li_2ZrO_3-type sorbent materials.

High-temperature carbon dioxide sorbents can also find applications in fuel reforming process to enhance fuel to hydrogen conversion efficiency. It was reported

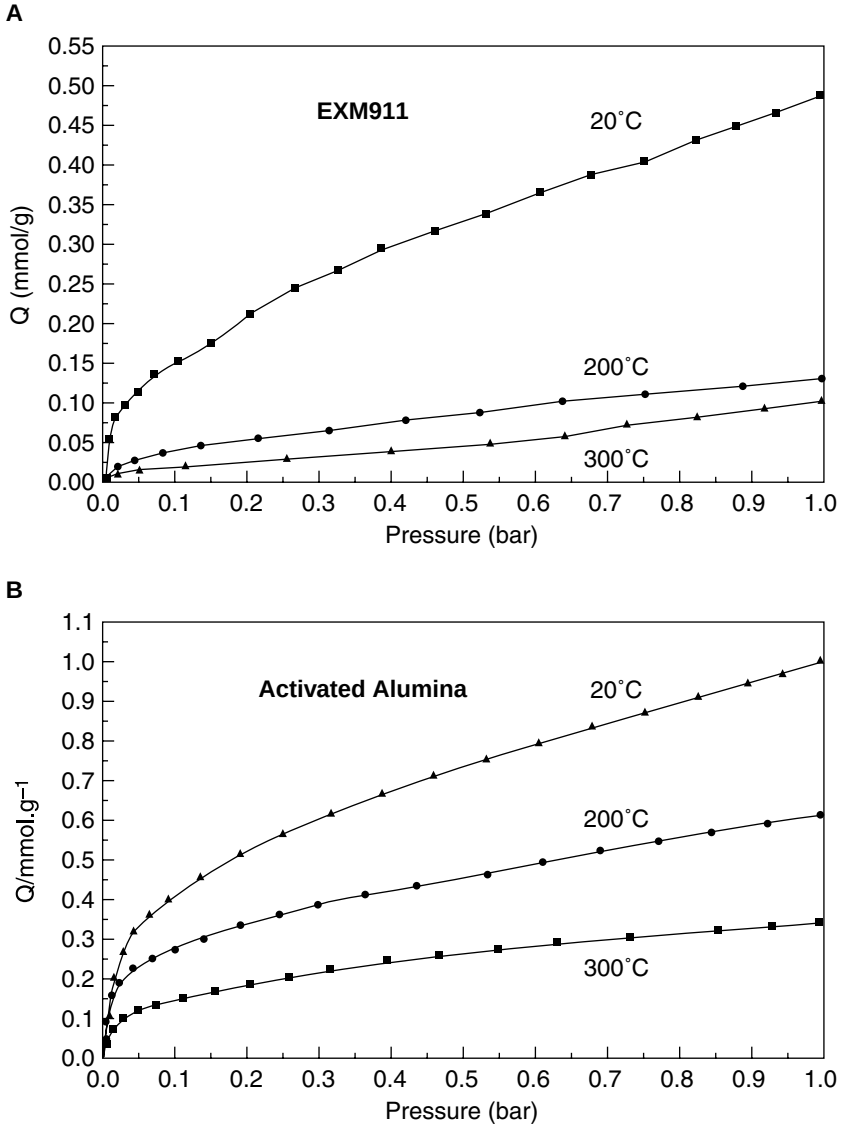


Fig. 8 Adsorption isotherms of carbon dioxide on commercial sorbents. (A) Hydrotalcite-like compound, EXM911; (B) LaRoche Industries activated alumina at 20, 200, and 300°C. (From Refs.[123,124].)

that sorption of carbon dioxide can enhance the production of hydrogen for a steam–methane reforming process using a mixture of Ni-based reforming catalyst and a Ca-based sorbent. The rates of the reforming, water-gas shift, and carbon dioxide removal reactions are sufficiently fast that combined reaction equilibrium was closely approached, allowing for >95 mol% hydrogen to be produced in a single step.[134]

CONCLUSIONS AND FUTURE DIRECTIONS

Existing commercial sorbents including activated carbon, zeolites, activated alumina, and silica gels will continue to play important roles in adsorptive separation and purification for current process industries in the near future. However, they cannot meet the needs of future technological developments in the new energy economy and the stringent environmental regulations. The newly developed nanostructured sorbent materials have shown some very promising features, but they are basically unexplored and systematic investigations are needed on both synthesis methods and adsorption characteristic studies. The following are the author's views on future research needs in both sorbent synthesis and applications:

1. Explore entirely new sorbent synthesis routes to better control of both sorbent pore texture and surface property.
2. Design new sorbent materials from basic building blocks and introduce active sorption sites according to sorbent–adsorbate interaction requirements. MOF material syntheses using

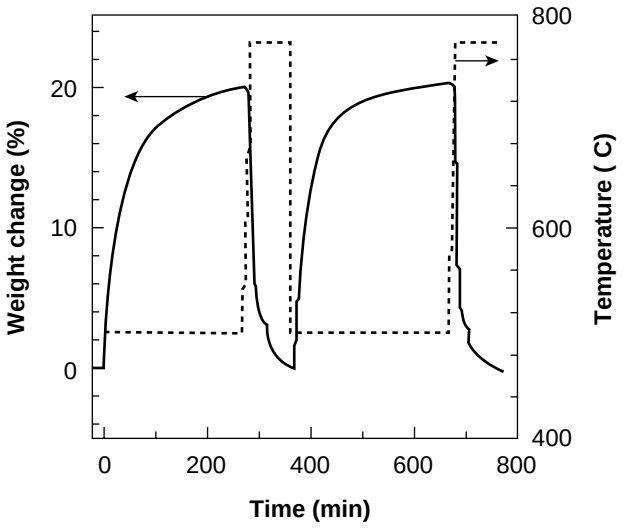


Fig. 9 CO_2 sorption and regeneration on the modified Li_2ZrO_3. Sorption process: 50% CO_2 balanced by dry air at 500°C. Desorption process: 50% CO_2 balanced by dry air at 780°C → dry air at 780°C. Gas flow rate: 150 ml/min. (From Ref.[131].)

the isoreticular method and sol–gel technique are two examples of this approach.

8. A better understanding of the relationship between sorbent–adsorbate interaction, sorption equilibrium, and kinetics through molecular simulation, and provide guidance for sorbent synthesis.

In terms of applications, new sorbents should be developed to meet the following pressing needs:

1. Deep desulfurization of fossil fuels for fuel cell application.
2. Hydrogen purification (H_2S, CO, and CO_2 removal).
3. Hydrogen and methane storage sorbents and processes.
4. Water treatment (arsenic, radionuclides and heavy metal ions and anions removal).
5. Air pollution control (SO_x, NO_x, and other toxic gases removal).
6. Chemisorbents as effective getter materials for toxic process gas and liquid streams.
7. Effective high-temperature carbon dioxide sorbents for carbon dioxide sequestration.

ACKNOWLEDGMENT

Professor Y.S. Lin is acknowledged for providing his publications and comments on high-temperature sorbents discussed in this entry.

REFERENCES

1. King, C.J. *Separation Processes*, 2nd Ed.; McGraw-Hill: New York, 1980.
2. Yang, R.T. *Adsorbents: Fundamentals and Applications*; John Wiley & Sons: Hoboken, NJ, 2003.
3. Yang, R.T. *Gas Separation by Adsorption Processes*; Butterworths: Boston, MA, 1987; 1–48.
4. Thomas, W.J.; Crittenden, B. *Adsorption Technology & Design*; Butterworth Heinemann: Oxford, 1998; 8–30.
5. Karger, J.; Ruthven, D.M. *Diffusion in Zeolites and Other Microporous Solids*; John Wiley & Sons: New York, 1992; 416–418.
6. Ruthven, D.M. *Principles of Adsorption and Adsorption Processes*; John Wiley & Sons: New York, 1984; 396–405.
7. Myers, A.L.; Prausnitz, J.M. Thermodynamics of mixed-gas adsorption. AIChE J. **1965**, *11* (1), 121–127.
8. Ruthven, D.M.; Farooq, S.; Knaebel, K.S. *Pressure Swing Adsorption*; VCH Publishers, Inc.: New York, 1994.
9. Patrick, J.W., Ed. *Porosity in Carbons: Characterization and Applications*; Halsted Press, an imprint of John Wiley and Sons: London, 1995.
10. Bansal, R.C.; Donnet, J.-B.; Stockli, F. *Activated Carbon*; Marcel Dekker, Inc.: New York, 1988.
11. Breck, D.W. *Zeolite Molecular Sieve*; Krieger Publishing Company: Marlbar, FL, 1974.
12. Barrer, R.M. *Zeolites and Clay Minerals as Sorbents and Molecular Sieves*; Academic Press: New York, 1978.
13. Auerbach, S.M.; Carrado, K.A.; Dutta, P.K., Eds. *Handbook of Zeolite Science and Technology*; Marcel Dekker, Inc.: New York, 2003.
14. Kumar, R Removal of Water and Carbon Dioxide from Atmospheric Air. U.S. Patent 4,711,645, Dec 8, 1987.
15. Jain, R Pre-purification of Air for Separation. U.S. Patent 5,232,474, Aug 3, 1993.
16. Jain, R.; Tseng, J.K Method and Apparatus for Producing Clean Dry Air having Application to Air Separation. U.S. Patent 6,077,488, Jun 20, 2000.
17. Deng, S.; Kumar, R.; Jain, R Air Purification Process with Thermal Regeneration. U.S. Patent 5,931,022, Aug 3, 1999.
18. Iler, R.K. *The Chemistry of Silica*; John Wiley & Sons: New York, 1979.
19. Unger, K.K. *Porous Silica, Its Properties and Use as a Support in Column Liquid Chromatography*; Elsevier: Amsterdam, The Netherlands, 1979.

20. Vansant, E.F.; Van Der Voort, P.; Vrancken, K.V. *Characterization and Chemical Modification of the Silica Surface*; Elsevier: Amsterdam, The Netherlands, 1995.
21. Barton, T.J.; Bull, L.M.; Klemperer, W.G.; Loy, D.A.; McEnaney, B.; Misono, M.; Monson, P.A.; Pez, G.; Scherer, G.W.; Vartuli, J.C.; Yaghi, O.M. Tailored porous materials. Chem. Mater. **1999**, *11* (10), 2633–2656.
22. Zhao, X.S.; Lu, G.Q. (Max) and Millar, G.J. Advances in mesoporous molecular sieve MCM-41. Ind. Eng. Chem. Res. **1996**, *35*, 2075–2090.
23. Guliants, V.V.; Carreon, M.A.; Lin, Y.S. Ordered mesoporous and macroporous inorganic films and membranes. J. Membr. Sci. **2004**, *235* (1/2), 53–72.
24. Beck, J.S.; Vartuli, J.C.; Roth, W.J.; Leonowicz, M.E.; Kresge, C.T.; Schmitt, K.D.; Chu, C.T.-W.; Olson, D.H.; Sheppard, E.W.; McCullen, S.B.; Higgins, J.B.; Schlenkert, J.L. A new family of mesoporous molecular sieves prepared with liquid crystal templates. J. Am. Chem. Soc. **1992**, *114*, 10834–10843.
25. Huang, H.Y.; Yang, R.T.; Chinn, D.; Munson, C.L. Amine-grafted MCM-48 and silica xerogel as superior sorbents for acidic gas removal from natural gas. Ind. Eng. Chem. Res. **2003**, *42* (12), 2427–2433.
26. Brinker, C.J.; Scherer, G.W. *Sol–Gel Science*; Academic Press, Inc.: San Diego, CA, 1990.
27. Iler, R.K. *The Chemistry of Silica*; John Wiley & Sons: New York, 1979.
28. Hench, L.L. *Sol–Gel Silica*; Noyes Publications: Westwood, NJ, 1998.
29. Lin, Y.S.; Deng, S. Sol–gel preparation of nanostructured adsorbents. In *Adsorption and Its Application in Industry and Environmental Protection*; Dabrowski, A., Ed.; Elsevier: Amsterdam, 1998; vol.120A, 653–686.
30. Yoldas, B.E. Transparent porous alumina. Am. Ceram. Soc. Bull. **1975**, *54*, 286–288.
31. Deng, S.G. Synthesis and Properties of Nanostructured Adsorbents for Gas Separation and Environmental Applications, Ph.D. dissertation, University of Cincinnati, Cincinnati, OH, 1996.
32. Deng, S.G.; Lin, Y.S. Granulation of sol–gel derived nano-scale alumina. AIChE J. **1997**, *43* (2), 505–514.
33. Buelna, G.; Lin, Y.S. Sol–gel derived nanoporous γ-alumina granules. Microporous Mesoporous Mater. **1999**, *30*, 359–369.
34. Buelna, G.; Lin, Y.S. Preparation of spherical alumina and copper oxide coated alumina sorbents by improved sol–gel granulation process. Microporous Mesoporous Mater. **2001**, *42*, 67–76.
35. Deng, S.; Lin, Y.S. Sol–gel preparation and properties of alumina adsorbents for gas separation. AIChE J. **1995**, *41* (3), 559–570.
36. Wang, Y.; Lin, Y.S. Sol–gel synthesis and gas adsorption properties of CuCl modified mesoporous alumina. J. Sol–Gel Sci. Technol. **1998**, *1*, 185–195.
37. Lin, Y.S.; Ji, W.; Wang, Y.; Higgins, R.J. Cuprous chloride modified nanoporous alumina membranes for ethylene–ethane separation. Ind. Eng. Chem. Res. **1999**, *38*, 2292–2298.
38. Deng, S.; Lin, Y.S. Sulfation and regeneration of sol-gel derived regenerative sorbents for flue gas desulfurization. In *AIChE Symp.*; Cohen, Y., Peters, R.W., Eds.; 1995; *91* (309), 32–39.
39. Deng, S.G.; Lin, Y.S. Synthesis, stability and sulfation properties of sol–gel derived regenerative sorbents for flue gas desulfurization. Ind. Eng. Chem. Res. **1996**, *35* (4), 1429–1437.
40. Lin, Y.S.; Deng, S.G. Removal of trace sulfur dioxide from a gas stream by regenerative sorption processes. Sep. Purif. Technol. **1998**, *13* (4), 65–77.
41. Wang, Z.-M.; Lin, Y.S. Sol–gel derived alumina alumina supported copper oxide sorbent for flue gas desulfurization. Ind. Eng. Chem. Res. **1998**, *37*, 4675–4681.
42. Yang, Z.; Lin, Y.S. Sol–gel synthesis of silicalite/γ-alumina granules. Ind. Eng. Chem. Res. **2000**, *39*, 4944–4948.
43. Buelna, G.; Lin, Y.S.; Liu, L.; Litster, J.D. Structural and mechanical properties of nanostructured granular alumina catalysts. Ind. Eng. Chem. Res. **2003**, *42*, 442–447.
44. Buelna, G.; Lin, Y.S. Combined removal of SO_2 and NO using sol–gel derived copper oxide coated alumina sorbents/catalysts. Environ. Technol. **2003**, *24* (9), 1087–1095.
45. Gordeeva, L.G.; Mrowiec-Bialon, J.; Jarzebski, A.B.; Lachowski, A.I.; Malinowski, J.; Aristov, Y.I. Selective water sorbents for multiple applications, 8. Sorption properties of $CaCl_2$–SiO_2 sol–gel composites. React. Kinet. Catal. Lett. **1999**, *66* (1), 113–120.
46. Rogojevic, S.; Jain, A.; Gill, W.N.; Plawsky, J. Moisture adsorption in nanoporous silica xerogels. Electrochem. Solid State Lett. **2002**, *5* (7), F22–F23.
47. Kittaka, S.; Uchida, N.; Kihara, T.; Suetsugi, T.; Sasaki, T. Interlayer water-molecules in vanadium pentoxide hydrate. 2. Effect of intercalated metal-ions on the adsorbability of water molecules.. Langmuir **1992**, *8* (1), 45–48.
48. Vaghetti, J.C.P.; Zat, M.; Bentes, K.R.S.; Ferreira, S.; Benvenutti, E.V.; Lima, E.C. 4-Phenylenediaminepropyl silica xerogel as a sorbent

for copper determination in waters by slurry-sampling ETAAS. Anal. At. Spectrom. **2003**, *8* (4), 376–380.
49. Burleigh, M.C.; Dai, S.; Barnes, C.E.; Xue, Z.L. Enhanced ionic recognition by a functionalized mesoporous sol–gel: synthesis and metal ion selectivity of diaminoethane derivative. Sep. Sci. Technol. **2001**, *36* (15), 3395–3409.
50. Pavan, F.A.; Lucho, A.M.S.; Goncalves, R.S.; Costa, T.M.H.; Benvenutti, E.V. Anilinepropyl-silica xerogel used as a selective Cu(II) adsorbent in aqueous solution. J. Colloid Interface Sci. **2003**, *263* (2), 688–691.
51. Khan, A.; Ahmad, S.; Zaidi, S.A.R.; Mahmood, F.; Khokhar, M.Y. Removal of mercury by 1-naphthylthiocarbamide doped xerogel using radiotracer technique. Sep. Sci. Technol. **2002**, *37* (13), 3099–3107.
52. Arenas, L.T.; Vaghetti, J.C.P.; Moro, C.C.; Lima, E.C.; Benvenutti, E.V.; Costa, T.M.H. Dabco/silica sol–gel hybrid material. The influence of the morphology on the $CdCl_2$ adsorption capacity. Mater. Lett. **2004**, *5* (6), 895–898.
53. Pavan, F.A.; Costa, T.M.H.; Benvenutti, E.V. Adsorption of $CoCl_2$, $ZnCl_2$ and $CdCl_2$ on aniline/silica hybrid material obtained by sol–gel method. Colloids Surf. A—Physicochem. Eng. Aspects **2003**, *226* (1–3), 95–100.
54. Pavan, F.A.; Lima, I.S.; Benvenutti, E.V.; Gushikem, Y.; Airoldi, C. Hybrid aniline/silica xerogel cation adsorption and thermodynamics of interaction. J. Colloid Interface Sci. **2004**, *275* (2), 386–391.
55. Deshpande, R.S.; Sharp-Goldman, S.L.; Bocarsly, A.B. Thermodynamics and kinetics of CO_2 adsorption on dehydrated palladium/cobalt-based cyanogels: a highly selective, fully reversible system for CO_2 storage. Langmuir **2002**, *18* (20), 7694–7698.
56. Aucoin, M.G.; Erhardt, F.A.; Legge, R.L. Hyperactivation of *Rhizomucor miehei* lipase by hydrophobic xerogels. Biotechnol. Bioeng. **2004**, *85* (6), 647–655.
57. Hayashi, J.; Watada, Y.; Muroyama, K. Preparation of mesoporous material having a hydrophobic surface by combining silica xerogel with resin using sol–gel method. Mater. Lett. **2001**, *50* (2/3), 87–91.
58. Wu, Z.J.; Ahn, I.S.; Lin, Y.X.; Huang, L.Y.; Lan, X.R.; Lee, K. Methyl orange adsorption by microporous and mesoporous TiO_2–SiO_2 and TiO_2–SiO_2–Al_2O_3 composite xerogels. Compos. Interfaces **2004**, *11* (2), 205–212.
59. Fuertes, A.B. Low-cost synthetic route to mesoporous carbons with narrow pore size distributions and tunable porosity through silica xerogel templates. Chem. Mater. **2004**, *1* (3), 449–455.
60. Yamamoto, T.; Endo, A.; Ohmori, T.; Nakaiwa, M. Porous properties of carbon gel microspheres as adsorbents for gas separation. Carbon **2004**, *42* (8/9), 1671–1676.
61. Teichner, S.L.; Nicolaon, G.A.; Vicarini, M.A.; Grades, G.E.E. Inorganic oxide areogels. Adv. Colloid Interface Sci. **1976**, *5*, 245–273.
62. Ayen, R.J.; Iacobucci, P.A. Metal oxide aerogel preparation by supercritical extraction. Rev. Chem. Eng. **1988**, *5* (1–4), 157–198.
63. Fanelli, A.J.; Verma, S.; Engelmann, T.; Burlew, J.V. Scale-up studies of alumina aerogel catalyst support. Ind. Eng. Chem. Res. **1991**, *30* (1), 126–129.
64. Stengl, V.; Bakardjieva, S.; Marikova, M.; Subrt, J.; Oplustil, F.; Olsanska, M. Aerogel nanoscale aluminum oxides as a destructive sorbent for mustard gas. Ceramics-Silikaty; **2003**, *47* (4), 175–180.
65. MrowiecBialon, J.; Jarzebski, A.B.; Lachowski, A.J.; Malinowski, J.J.; Aristov, Y.I. Effective inorganic hybrid adsorbents of water vapor by the sol–gel method. Chem. Mater. **1997**, *9* (11), 2486–2490.
66. Bostain, D.A.; Brenizer, J.S.; Norris, P.M. Neutron radioscopic measurement of water adsorption coefficients in aerogels. Res. Nondestruct. Eval. **2002**, *14* (1), 47–57.
67. Ahmed, M.S.; Attia, Y.A. Multi-metal oxide aerogel for capture of pollution gases from air. Appl. Therm. Eng. **1998**, *18* (9/10), 787–797.
68. Stengl, V.; Bakardjieva, S.; Marikova, M.; Subrt, J.; Oplustil, F.; Olsanska, M. Aerogel nanoscale magnesium oxides as a destructive sorbent for toxic chemical agents. Cent. Eur. J. Chem. **2004**, *2* (1), 16–33.
69. Khaleel, A.; Dellinger, B. FTIR investigation of adsorption and chemical decomposition of CCl_4 by high surface-area aluminum oxide. Environ. Sci. Technol. **2002**, *36* (7), 1620–1624.
70. Shaban, I.S.; Mikulaj, V. Sorption–desorption of radiocesium on various sorbents in presence of humic acid. J. Radioanal. Nucl. Chem. **1996**, *208* (2), 593–603.
71. Yamashita, J.; Ojima, T.; Shioya, M.; Hatori, H.; Yamada, Y. Organic and carbon aerogels derived from poly(vinyl chloride). Carbon **2003**, *41* (2), 285–294.
72. Coleman, S.J.; Coronado, P.R.; Maxwell, R.S.; Reynolds, J.G. Granulated activated carbon modified with hydrophobic silica aerogel—potential composite materials for the removal

of uranium from aqueous solutions. Environ. Sci. Technol. **2003**, *37* (10), 2286–2290.
73. Gabelich, C.J.; Tran, T.D.; Suffet, I.H. Electrosorption of inorganic salts from aqueous solution using carbon aerogels. Environ. Sci. Technol. **2002**, *36* (13), 3010–3019.
74. Rowsell, J.L.C.; Millward, A.R.; Park, K.S.; Yaghi, O.M. Hydrogen sorption in functionalized metal-organic frameworks. J. Am. Chem. Soc. **2004**, *126* (18), 5666–5667.
75. Eddaoudi, M.; Li, H.; Yaghi, O.M. Highly porous and stable metal-organic frameworks: structure design and sorption properties. J. Am. Chem. Soc. **2000**, *122*, 1391–1397.
76. Eddaoudi, M.; Moler, D.; Li, H.; Chen, B.; Reineke, T.; O'Keeffe, M.; Yaghi, O.M. Modular chemistry: secondary building units as a basis for the design of highly porous and robust metal-organic carboxylate frameworks. Acc. Chem. Res. **2001**, *34*, 319–330.
77. Eddaoudi, M.; Kim, J.; Rosi, N.; Vodak, D.; Watcher, J.; O'Keeffe, M.; Yaghi, O.M. Systematic design of pore size and functionality in isoreticular MOFs and their application in methane storage. Science **2002**, *295*, 469–472.
78. Dybtsev, D.N.; Chun, H.; Yoon, S.H.; Kim, D.; Kim, K. Microporous manganese formate: a simple metal-organic porous material with high framework stability and highly selective gas sorption properties. J. Am. Chem. Soc. **2004**, *126* (1), 32–33.
79. Rosi, N.; Eckert, J.; Eddaoudi, M.; Vodak, D.; Kim, J.; O'Keeffe, M.; Yaghi, O.M. Hydrogen storage in microporous metal-organic frameworks. Science **2003**, *300*, 1127–1129.
80. Pan, L.; Sander, M.B.; Huang, X.Y.; Li, J.; Smith, M.; Bittner, E.; Bockrath, B.; Johnson, J.K. Microporous metal organic materials: promising candidates as sorbents for hydrogen storage. J. Am. Chem. Soc. **2004**, *126* (5), 1308–1309.
81. Yaghi, O.M.; O'Keeffe, M.; Ockwig, N.W.; Chae, H.K.; Eddaoudi, M.; Kim, J. Reticular synthesis and the design of new materials. Nature **2003**, *423* (6941), 705–714.
82. Wang, Q.-M.; Shen, D.; Bülow, M.; Deng, S.; Fitch, F.R.; Lemcoff, N.O.; Semanscin, J. Metallorganic molecular sieve for gas separation and purification. Microporous Mesoporous Mater. **2002**, *55*, 217–230.
83. Wang, Q.-M.; Shen, D.; Bulow, M.; Lau, M.-L.; Fitch, F.R.; Deng, S. Metallo-organic Polymers for Gas Separation and Purification. U.S. Patent 6,491,740, , Dec 10, 2002.
84. Office of Science, U.S. Department of Energy, Basic Research Needs for the Hydrogen Economy, Feb. **2004**.
85. Seayad, A.M.; Antonelli, D.M. Recent advances in hydrogen storage in metal-containing inorganic nanostructures and related materials. Adv. Mater. **2004**, *16* (9/10), 765–777.
86. Ding, R.G.; Lu, G.Q.; Yan, Z.F.; Wilson, M.A. Recent advances in the preparation and utilization of carbon nanotubes for hydrogen storage. J. Nanosci. Nanotechnol. **2001**, *1* (1), 7–29.
87. Liu, C.; Fan, Y.Y.; Liu, M.; Cong, H.T.; Cheng, H.M.; Dresselhaus, M.S. Hydrogen storage in single-walled carbon nanotubes at room temperature. Science **1999**, *286* (5442), 1127–1129.
88. Lueking, A.; Yang, R.T. Hydrogen storage in carbon nanotubes: residual metal content and pretreatment temperature. AIChE J. **2003**, *49* (6), 1556–1568.
89. Tang, C.C.; Bando, Y.; Ding, X.X.; Qi, S.R.; Golberg, D. Catalyzed collapse and enhanced hydrogen storage of BN nanotubes. J. Am. Chem. Soc. **2002**, *124* (49), 14550–14551.
90. Chen, J.; Li, S.L.; Tao, L.; Shen, Y.T.; Cui, C.X. Titanium disulfide nanotubes as hydrogen-storage materials. J. Am. Chem. Soc. **2003**, *125* (18), 5284–5285.
91. Sandrock, G. A panoramic overview of hydrogen storage alloys from a gas reaction point of view. J. Alloys Compd. **1999**, *887*, 293–295.
92. Zaluska, A.; Zaluski, L.; Strom-Olsen, J.O. Structure, catalysis and atomic reactions on the nano-scale: a systematic approach to metal hydrides for hydrogen storage. Appl. Phys. A—Mater. Sci. Process. **2002**, *72* (2), 157–165.
93. Fichtner, M.; Fuhr, O.; Kircher, O. Magnesium alanate—a material for reversible hydrogen storage? J. Alloys Compd. **2003**, *356*, 418–422.
94. Zuttel, A.; Wenger, P.; Rentsch, S.; Sudan, P.; Mauron, P.; Emmenegger, C. $LiBH_4$: a new hydrogen storage material. J. Power Sources **2003**, *118* (1/2), 1–7.
95. Chen, P.; Xiong, Z.T.; Luo, J.Z.; Lin, J.Y.; Tan, K.L. Interaction of hydrogen with metal nitrides and imides. Nature **2002**, *420* (6913), 302–304.
96. Mao, W.L.; Mao, H.-K.; Goncharov, A.F.; Struzhkin, V.V.; Guo, Q.; Hu, J.; Shu, J.; Hemley, R.J.; Somayazulu, M.; Zhao, Y. Hydrogen cluster in clathrate. Science **2002**, *29* (5590), 2247–2249.
97. Mao, W.L.; Mao, H.-K. Hydrogen storage in molecular compounds. Proc. Natl. Acad. Sci. USA **2004**, *101* (3), 708–710.

98. Mao, W.L.; Mao, H.-K. Composition and Method for Hydrogen Storage. U.S. Patent 6,735,960 B2, 2004.
99. Xie, Y.C.; Tang, Y.Q. Spontaneous monolayer dispersion of oxides and salts onto surface of supports: applications to heterogeneous catalysis. Adv. Catal. **1990**, *37*, 1–43.
100. Golden, T.C.; Kratz, W.C.; Wilhelm, F.C Highly Dispersed Cuprous Compositions. U.S. Patent 5,126,310Jun 30, 1992.
101. Golden, T.C.; Kratz, W.C.; Wilhelm, F.C.; Pierantozzi, R.; Rokicki, A Highly Dispersed Cuprous Compositions. U.S. Patent 5,175,137, Dec 29, 1992.
102. Deng, S.G.; Lin, Y.S. Microwave heating synthesis of supported sorbents. Chem. Eng. Sci. **1997**, *52* (10), 1563–1575.
103. Takahashi, A.; Yang, R.T.; Munso, C.L.; Chinn, D. Cu(I)-Y-zeolite as a superior adsorbent for diene/olefin separation. Langmuir **2001**, *17*, 8405–8413.
104. Takahashi, A.; Yang, R.T.; Munso, C.L.; Chinn, D. Influence of Ag content and H_2S exposure on 1,3-butadiene/1-butene adsorption by Ag ion-exchanged Y-zeolites (Ag-Y). Ind. Eng. Chem. Res. **2001**, *40*, 3979–3988.
105. Yang, R.T.; Kikkinides, E.S. New sorbents for olefin paraffin separations by adsorption via π-complexation. AIChE J. **1995**, *41* (3), 509–517.
106. Golden, T.C.; Kalbassi, M.A.; Taylor, F.W.; Allam, R.J Use of Zeolites and Alumina in Adsorption Processes. U.S. Patent 5,779,767, Jul 14, 1998.
107. Deng, S.; Kumar, R.; Wolf, R.; Andrecovich, M.J Purification of Gases Using Multi-Composite Adsorbent. U.S. Patent 6,358,302, Mar 19, 2002.
108. Kumar, R.; Huggahalli, M.; Deng, S.G.; Andrecovich, M. Trace impurity removal from air. Adsorption **2003**, *9* (3), 243–250.
109. Lin, Y.S.; MacLean, D.L.; Zeng, Y High Temperature Adsorption Process. U.S. Patent 6,059,858, May 9, 2000.
110. Bouwmeester, H.J.M.; Burgraaf, A.J. Dense ceramic membranes for oxygen separation. In *Solid State Electrochemistry*; Gellings, P.J., Bouwmeester, H.J.M., Eds.; CRC Press: Boca Raton, FL, 1997; 481–553.
111. Dyer, P.N.; Richards, R.E.; Russek, S.L.; Taylor, D.M. Ion transport membrane technology for oxygen separation and syngas production. Solid State Ionics **2000**, *134*, 21–33.
112. Kang, D.; Srinivasan, R.S.; Thorogood, R.M.; Foster, E.P Integrated High Temperature Method for Oxygen ProductionU.S. Patent 5,516,359, May 16, 1996.
113. Prasad, R Advanced Membrane System for Separating Gaseous Mixtures. U.S. Patent 5,709,732, Jan 20, 1998.
114. Yang, Z.; Lin, Y.S.; Zeng, Y. High-temperature sorption process for air separation and oxygen removal. Ind. Eng. Chem. Res. **2002**, *41*, 2775–2784.
115. Yang, Z.; Lin, Y.S. A semi-empirical equation for oxygen nonstoichiometry of perovskite-type ceramics. Solid State Ionics **2002**, *150*, 245–254.
116. Yang, Z.; Lin, Y.S. Equilibrium of oxygen sorption on perovskite type ceramic sorbents. AIChE J. **2003**, *49*, 793–798.
117. Yang, Z.H.; Lin, Y.S. High temperature oxygen sorption in fixed-bed packed with perovskite-type ceramic sorbents. Ind. Eng. Chem. Res. **2003**, *42*, 4376–4381.
118. Draper, E.L.; Becker, R.A. *Research and Development Needs for the Sequestration of Carbon Dioxide as Part of a Carbon Management Strategy*; The National Coal Council: Washington, DC, 2000.
119. Azar, C.; Rodhe, H. Targets for stabilizing of atmospheric CO_2. Science **1997**, *276*, 1818–1819.
120. Bruant, R.G. Jr.; Guswa, A.J.; Celia, M.A.; Peters, C.A. Safe storage of CO_2 in deep saline aquifers. Environ. Sci. Technol. **2002**, *36*, 240–245.
121. Herzog, H.J. What future for carbon capture and sequestration? Environ. Sci. Technol. **2001**, *35*, 148–153.
122. Rao, A.B.; Rubin, E.A. Technical, economic, and environmental assessment of amine-based CO_2 capture. Environ. Sci. Technol. **2002**, *36*, 4467–4475.
123. Zou, Y.; Mata, V.; Rodrigues, A.E. Adsorption of carbon dioxide on basic alumina at high temperatures. J. Chem. Eng. Data **2000**, *45*, 1093–1095.
124. Zou, Y.; Mata, V.; Rodrigues, A.E. Adsorption of carbon dioxide onto hydrotalcite-like compounds (HTlcs) at high temperatures. Ind. Eng. Chem. Res. **2001**, *40*, 204–209.
125. Golden, T.C.; Taylor, F.W.; Johnson, L.M.; Malik, N.H.; Raiswell, C.J Purification of Air. U.S. Patent 610,659, Aug 22, 2000.
126. Monereau, C Use of an Activated Alumina for Removing the CO_2 from a Gas. U.S. Patent 6,379,430, Apr 30, 2002.
127. Kumar, R.; Huggahalli, M.; Bulow, M Thermal Swing Adsorption Process. U.S. Patent 6,432,171, Aug 13, 2002.

128. Ko, D.; Siriwardane, R.; Biegler, L.T. Optimization of a pressure-swing adsorption process using Zeolite 13X for CO_2 sequestration. Ind. Eng. Chem. Res. **2003**, *42*, 339–348.
129. Yong, Z.; Mata, V.; Rodrigues, A.E. Adsorption of carbon dioxide at high temperature–a review. Sep. Purif. Technol. **2002**, *26* (2/3), 195–205.
130. Nakagawa, K.; Ohashi, T. A novel method of CO_2 capture from high temperature gases. J. Electrochem. Soc. **1998**, *145* (4), 1344–1346.
131. Ida, J.-I.; Lin, Y.S. Mechanism of high-temperature CO_2 sorption on lithium zirconate. Environ. Sci. Technol. **2003**, *37* (9), 1999–2004.
132. Xiong, R.; Ida, J.-I.; Lin, Y.S. Kinetics of carbon dioxide sorption on potassium doped lithium zirconate. Chem. Eng. Sci. **2003**, *58*, 4377–4385.
133. Ida, J.-I.; Xiong, R.; Lin, Y.S. Synthesis and CO_2 sorption properties of pure and modified lithium zirconate. Sep. Purif. Technol. **2004**, *36*, 41–51.
134. Ortiz, A.L.; Harrison, D.P. Hydrogen production using sorption-enhanced reaction. Ind. Eng. Chem. Res. **2001**, *40*, 5102–5109.

Spinning Disk Reactor

R. J. J. Jachuck
Process Intensification and Clean Technology (PICT) Group, Department of Chemical Engineering, Clarkson University, Potsdam, New York, U.S.A.

J. R. Burns
Protensive Ltd., Bioscience Centre, Centre for Life, Newcastle Upon Tyne, U.K.

INTRODUCTION

The concept of process intensification aims to achieve enhancement in transport rates by orders of magnitude to develop multifunctional modules with a view to provide manufacturing flexibility in process plants. In recent years, advancement in the field of reactor technology has seen the development of catalytic plate reactors, oscillatory baffled reactors, microreactors, membrane reactors, and trickle-bed reactors. One such reactor that is truly multifunctional in characteristics is the spinning disk reactor (SDR). This reactor has the potential to provide reactions, separations, and good heat transfer characteristics.

SPINNING DISK REACTOR

The SDR technology utilizes the effects of centrifugal force, which is capable of producing highly sheared thin films (Fig. 1) on the surfaces of rotating disks or cones. Extensive heat and mass transfer studies using SDRs have shown that convective film heat transfer coefficients as high as $14\,kW/m^2$, K, and mass transfer coefficients, K_L, with values as high as 30×10^{-5} m/sec, and K_G values as high as 12×10^{-8} m/sec, can be achieved while providing micromixing and appropriate fluid dynamic environment for achieving faster reaction kinetics.[1] The size of the disk may range from 60 to 500 mm in diameter and the surface characteristic may be smooth, grooved, or meshed depending on the application and the throughput requirement. The rotational speeds may range from 100 to about 6000 rpm (typically around 1500 rpm). The SDR, which has been successfully used to perform free radical as well as condensation polymerizations, fast precipitation reactions for the production of mono-dispersed particles and catalyzed organic reactions, has the following characteristics:[2–4]

- Intense mixing in the thin liquid film.
- Short liquid residence time (may allow the use of higher processing temperatures).
- Plug flow characteristics.
- High solid/liquid heat/mass transfer.
- High liquid/vapor heat/mass transfer.

A schematic diagram of an SDR is shown in Fig. 2

For the purpose of clarity and continuity this entry has been subdivided into the following sections: hydrodynamics of liquid flow on a rotating surface, variation in SDR configuration, performance estimators, and process application of SDRs.

HYDRODYNAMICS OF LIQUID FLOW ON A ROTATING SURFACE

Synchronized Flow Model

The simplest model for flow over a rotating disk surface assumes that the liquid is rotating at the same speed as the disk itself and is thus fully synchronized with the disk rotation. Under these conditions, the centrifugal acceleration driving the liquid film across the disk surface at radius r can be simply estimated as $r\omega^2$. With this assumption, the flow over the disk can be made analogous to flow over an inclined surface.

Nusselt provided a simple model for laminar liquid flow down an inclined plane.[5] This assumed that the liquid had reached fully developed conditions in which drag due to viscous shear exactly balanced the weight of the film. Under these conditions, Nusselt showed that for a Newtonian fluid of kinematic viscosity, ν, film thickness, f, could be written in terms of the liquid flow rate, Q, moving over a vertically inclined surface of width, w, under a gravitational acceleration, g, using the following relationship:

$$f = \left(\frac{3\nu Q}{wg}\right)^{1/3} \tag{1}$$

By substitution of gravitational acceleration g with centrifugal acceleration $r\omega^2$ and the width of the inclined surface w with the perimeter of the disk at radius r, the following equation for film thickness on

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120022537

Fig. 1 View of sheared liquid films on an SDR. *(View this art in color at www.dekker.com.)*

a rotating disk surface can be written as [6]

$$f = \left(\frac{3\nu Q}{2\pi r^2\omega^2}\right)^{1/3} \tag{2}$$

This equation assumes that the liquid velocity can instantaneously adjust to the balance drag against centrifugal force. This assumption has been shown to be a reasonable one. Burns et al. and Eq. (2) can be used as the starting point for film thickness modeling.[7]

The fully synchronized flow model can be used to provide reasonable estimates for a wide range of measures that characterize the flow over the spinning disk. The first of these is the radial velocity of the film at radius r that can be calculated from

$$u = \frac{Q}{2\pi rf} = \left(\frac{1}{12\pi^2}\right)^{1/3}\left(\frac{Q^2\omega^2}{r\nu}\right)^{1/3} \tag{3}$$

For a fully developed flow profile the radial velocity can be used to provide an estimate of surface shear, S, using the following equation:

$$S = \frac{3u}{f} = \left(\frac{3}{2\pi}\right)^{1/3}\left(\frac{Qr\omega^4}{\nu^2}\right)^{1/3} \tag{4}$$

Integration of the equations for film thickness and velocity can be used to provide calculations for other global measures of conditions on the disk surface. The first of these is the measurement of the volume of liquid on the disk surface. This can be calculated by integration of Eq. (2) to give

$$V = \left(\frac{81\nu Q\pi^2}{16\omega^2}\right)^{1/3} R^{4/3} \tag{5}$$

Residence time of the liquid on the disk can be estimated from the volume calculation above and the liquid flow rate Q using the following equation:

$$t = \frac{V}{Q} = \left(\frac{81\pi^2R^4\nu}{16Q^2\omega^2}\right)^{1/3} \approx 3.68\left(\frac{R^4\nu}{Q^2\omega^2}\right)^{1/3} \tag{6}$$

A measure of average film thickness and velocity can also be generated from Eq. (5). In the case of average film thickness, f_{AV}, it is defined as the volume of liquid on the disk per unit disk area and characterizes film thickness for the system. In the case of average radial velocity it is defined as the radius divided by the residence time and again characterizes conditions for the process. These two measures are given by

$$f_{AV} = \left(\frac{81}{16\pi}\right)^{1/3}\left(\frac{Q\nu}{R^2\omega^2}\right)^{1/3} \approx 1.17\left(\frac{Q\nu}{R^2\omega^2}\right)^{1/3} \tag{7}$$

$$u_{AV} = \frac{R}{t} \approx 0.27\left(\frac{Q^2\omega^2}{R\nu}\right)^{1/3} \tag{8}$$

Eqs. (7) and (8) allow for a characteristic radial surface shear rate to be calculated as

$$S_{AV} = \frac{3u_{AV}}{h_{AV}} \approx 0.69\left(\frac{QR\omega^4}{\nu^2}\right)^{1/3} \tag{9}$$

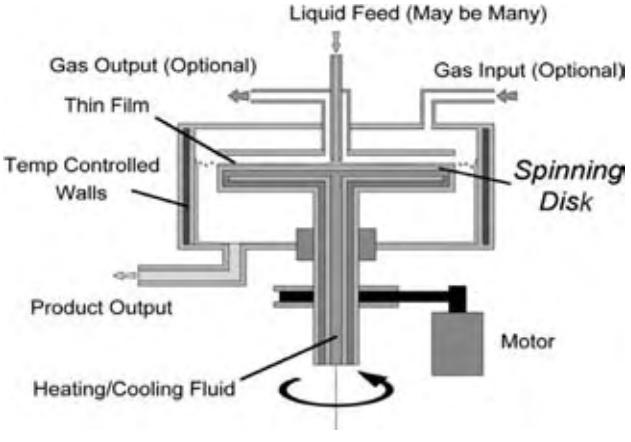


Fig. 2 A schematic view of an SDR. *(View this art in color at www.dekker.com.)*

Wetting of the disk surface

Wetting is an important aspect of the SDR. If the disk is not wetted then dry spots are created and rivulets are formed, which significantly reduce the transport rates achieved on the disk. Hartley and Murgatroyd provided a list of theoretical models for calculating the wetting film for liquid flows under gravity.[8] The equations derived for these models were based on physical principles rather than empirical data but have compared favorably with experimental results. Because

of this these models can be used for initial calculations of wetting of smooth spinning disk surfaces.

Two models for wetting of falling films were provided by Hartley and Murgatroyd.[8] Both these are based on the stability of dry patches rather than wetted films with the assumption that if a dry patch is not stable a wetting film should form. The first model is derived from minimizing surface energy and the second is from force balance at a contact line surrounding a dry spot. These are given by

Surface Energy Criterion

$$f_{\min} = 1.34\left(\frac{\sigma}{\rho}\right)^{1/5}\left(\frac{\nu}{g}\right)^{2/5} \tag{10}$$

Force Criterion

$$f_{\min} = 1.72\left(\frac{\sigma(1 - \cos\theta)}{\rho}\right)^{1/5}\left(\frac{\nu}{g}\right)^{2/5} \tag{11}$$

These equations produced very similar results with the exception that the force criterion approach included details of contact angle θ between the liquid and the surface and was therefore a more complete description. This theory can be related to the conditions on a spinning disk by the substitution of gravitation acceleration g with centrifugal acceleration $r\omega^2$. A stability parameter β for the film can then be given by comparing the film thickness at the edge of the disk with the minimum wetting film thickness above. This can be written as

$$\beta = \frac{f}{f_{\min}} \quad \text{at } r = R \text{ with } g = r\omega^2 \tag{12}$$

Using Eqs. (10) and (11) with Eq. (2) gives

$$\beta = 0.454\left(\frac{\rho^3 Q^5 \omega^2}{\sigma^3(1 - \cos\theta)^3 \nu R^4}\right)^{1/15} \tag{13}$$

Experimental data on falling films have suggested that this theory provides a very conservative measurement of minimum wetting film thickness as it is derived from the principle of the stability of dry patches rather than the breakdown of films. In general, it was observed that films could be maintained down to close to an order of magnitude lower than the above equations suggest, although under very contrived conditions.[8] Based on this, it can be assumed that

$\beta > 1 \Rightarrow$ Stable wetting film
$1 > \beta > 0.1 \Rightarrow$ Metastable film
$\beta < 0.1 \Rightarrow$ Rivulet flow

Examination of conditions used during a range of experimental work on SDR (Woods, Auone and Ramshaw, and Burns et al.) has shown that wetting films for β values within the ''metastable film'' region can be obtained.[7,9,10] In particular, the use of preflooding of the disk or the use of surface roughening can push the limit of the wetting film thickness down to the lower end of the metastable film region, further details of which will be discussed in the later sections. However, the above wetting parameter calculations can be used to examine how likely it is for a process to achieve a stable wetted disk surface.

Nonsynchronized Flow Model—Spin-up Zone

The synchronized flow model discussed in the previous sections can be used to provide estimates for general conditions on the spinning disk surface. However, to examine conditions close to the central feed requires a more complex two-dimensional model for fluid flow. In particular, this is required to provide information on the radial distance needed to achieve synchronization with the disk rotation, commonly referred to as the spin-up zone.[7]

A two-dimensional model for flow over a rotating disk, incorporating inertial and viscous influences, was described by Wood and Watts.[6] This was termed the Pigford model, owing to its origin. Burns et al. gave a comparison of the model with measurements of flow over a spinning disk.[7] A review of those findings and the model derivation are given below. The two-dimensional Pigford model can be written in terms of the radial and tangential flow velocity, u and v, respectively, relative to the disk using the following equations.

$$u\left(\frac{\partial u}{\partial r}\right) - \frac{v^2}{r} = -\left(\frac{12\pi^2 r^2 K_1 \nu}{Q^2}\right)u^3 \tag{14}$$

$$u\left(\frac{\partial v}{\partial r}\right) + \frac{uv}{r} = \left(\frac{12\pi^2 r^2 K_2 \nu}{Q^2}\right)u^2(r\omega - v) \tag{15}$$

where K_1 and K_2 were used as empirical correction factors to adjust the viscous terms in the equation. No specific data were given on the values of the constants K_1 and K_2, although it was assumed that they were close to unity. The solution of these equations was shown by Burns et al. to be strongly linked to the Eckman length scale λ, which was defined as[7]

$$\lambda = \left(\frac{Q^2}{\omega\nu}\right)^{1/4} \tag{16}$$

This scale was derived from the Eckman number that describes the ratio of viscous momentum transfer

through the film to that of the angular momentum of the film and can be used to gauge the ability for synchronized rotating flow to be achieved. Burns et al. found that K_1 had an unclear link with other experimental parameters and generally was in the region of 0.5–0.7 with an average of 0.61. In contrast, the value of parameter K_2 was found to be strongly linked to the Eckman length scale λ. The best fit of experimental results to the Pigford model was reported by Burns et al. as

$$K_1 = 0.61 \tag{17}$$

$$K_2 = \left(\frac{\lambda}{\lambda_0}\right) \tag{18}$$

where $\lambda_0 = 10.8\,\text{cm}$ and is an empirical constant whose units served only to balance the dimensions in Eq. (26). It is likely, however, that λ_0 may depend on other parameters not examined within the scope of the reported experiments, such as nozzle diameter, and may be a more complex function of several other dimensionless groups.

The experimentally determined values of parameters K_1 and K_2 allow the Pigford model to be used to examine the extent of the spin-up zone on the disk surface. Before this can be done, however, an unambiguous mathematical definition for the spin-up zone is required. This was provided by Burns et al. as the location at which the radial velocity stops accelerating and starts following the decelerating profile described by the synchronized flow model as shown in Eq. (3).[7]

Mathematically, this was defined as

$$\frac{\mathrm{d}u}{\mathrm{d}r} = 0 \quad \text{and} \quad \frac{\mathrm{d}^2u}{\mathrm{d}r^2} < 0 \text{ at } r = r_S \tag{19}$$

Numerical analysis of the Pigford model, combined with Eqs. (17) and (18), implied that the spin-up zone r_S based on this definition should be given by

$$r_S = 0.88\lambda\left(\frac{d}{2\lambda}\right)^{-0.025}\left(\frac{\lambda}{\lambda_0}\right)^{-0.37} \tag{20}$$

where d is the nozzle diameter used to inject the liquid. This result implies that the spin-up zone is approximately proportional to $\lambda^{2/3}$. A criterion for the accuracy of the synchronized flow model can also be derived from this analysis by comparison of the extent of the spin-up zone with the radius of the disk. This can be written simply as

$$\frac{r_S}{R} \ll 1 \Rightarrow \text{Synchronized flow} \tag{21}$$

Variation in Configuration

Surface structuring

Surface structuring can be used on SDR to achieve a variety of effects on the processing conditions. These can be categorized by the scale of the features being made to the disk. At the film-thickness scale or lower, surface structuring or roughening can be used to increase the wetting of the disk surface by the interference with the liquid–solid–gas interface contact. It can also be used to generate greater surface area at the base of the film and sites to promote small-scale convective mixing for situations with low-viscosity fluids and high flow rates.

Structuring at the larger than film-thickness scale can include surface grooves and channels and possible incorporation of surface meshes. These act to influence the velocity distribution across the disk allowing the liquid in some cases to detach and reimpact on the surface many times as it crosses the disk. This can be used to introduce a greater variation in the transport rates for the liquid flowing over the disk compared to the more steady conditions on the smooth disk surface.

At the large end of the scale changes in the configuration can be a change in the whole inclination of the surface to the centrifugal field the influence of which is discussed in the next few sections.

Spinning cone reactor

The SDR can be described as just one of the many reactor designs based on rotating surfaces. In particular, the SDR can be treated as a special case of the more general spinning cone reactor. These devices generally have a flat spinning disk close to the center followed by a conical section inclined to the centrifugal field. The equations describing the flow over these devices can, however, be shown to be more general versions of the previous equations given for the spinning disk reactor.

For the case of a spinning cone inclined at an angle α to the plane of rotation, film thickness, velocity, residence time, and surface shear can be related to that for the spinning disk using the following scaling relationships:

$$f_{\text{CONE}} = (\cos\alpha)^{-2/3} f_{\text{DISK}} \tag{22}$$

$$u_{\text{CONE}} = (\cos\alpha)^{2/3} u_{\text{DISK}} \tag{23}$$

$$t_{\text{CONE}} = (\cos\alpha)^{-5/3} t_{\text{DISK}} \tag{24}$$

$$S_{\text{CONE}} = (\cos\alpha)^{4/3} S_{\text{DISK}} \tag{25}$$

It should be noted that the relationships described above assume that the cone extends to the same radius

as the disk rather than the same length in the direction of flow and, hence, results in a strong influence on residence time due to slower flow combined with a longer distance to travel. It should also be noted that the flow velocity is given in Eq. (23) rather than the strict radial velocity and is therefore in the direction of the inclined surface. Eq. (24) gives an equivalent residence time scaling assuming that the cone extends to the axis of rotation. In reality, this is not often the case and the device will have a flat disk surface close to the center that leads to a cone further out. A more complete calculation for residence time can be given using the following equation:

$$t = \left(\frac{81\pi^2\nu}{16Q^2\omega^2}\right)^{1/3}\left(\frac{r_2^{4/3} - r_1^{4/3}}{\cos^{5/3}\alpha} + r_1^{4/3}\right) \tag{26}$$

where flow is over a flat disk up to a radius r_1 and then a cone inclined at an angle α up to a radius r_2. It should be noted that Eqs. (22)–(26) neglect the influence of gravity on the flow over the cone. Under most operating conditions, this will probably be a reasonable assumption; however, for systems with a vertical axis of rotation and low rotational speeds the validity of the previous equations may break down. A limit for the applicability of Eqs. (22)–(26) can be expressed as

$$\cos\alpha \gg \frac{R\omega^2}{g} \Rightarrow \text{Gravity may be neglected} \tag{27}$$

Nested cone reactor

A further modification of the rotating surface concept is that of the nested cone. This allows the use of inclined surfaces to the centrifugal acceleration in a similar manner to the cone but with two major differences. The first is that the design is more compact as the axial component of flow alternates as the liquid moves from conical surface to conical surface, which reduces the axial space required for the device. The second difference is that the liquid detaches from the surface as it moves from surface to surface, leading to periodic impacts and the potential for increased mixing at the point of impact. This can further be enhanced by the use of nested cones where the upper and the lower sections are moving at different rotational speeds, or even in opposite directions of rotation at the expense of increased power demand on the shaft.

Modeling of the flow on the conical surfaces can be approximated by the same scaling rules as shown in Eqs. (22), (23), and (25). However, residence time cannot be modeled in such a simple manner as it includes contributions both from flow over the conical sections and from detached flow between surfaces. In modeling the latter, the following equations can be used.

If it is assumed that the liquid leaving one conical section at a radius r_1 has a radial velocity component that is small compared to that of the rotational speed $r_1\omega$ of the surface, then the liquid can be assumed to leave the surface at a velocity $r_1\omega$ traveling at a tangent to the cone in the same direction as its rotation. Under these circumstances, the distance x traveled by the liquid before impacting on a second cone surface at radius r_2, assuming deflection due to the surrounding gas flow or gravity, can be written as

$$x = \sqrt{r_2^2 - r_1^2} \tag{28}$$

Assuming that the liquid travels between the surfaces without any change in velocity, as viewed from a nonrotating Cartesian frame of reference, the time required to travel between the cone surfaces can be written as

$$t = \frac{x}{r_1\omega} = \frac{1}{\omega}\sqrt{\frac{r_2^2}{r_1^2} - 1} \tag{29}$$

In comparison, the estimated time spent on the conical surface prior to this can be estimated from the following equation, which is a modified version of Eq. (26):

$$t = \left(\frac{81\pi^2\nu}{16Q^2\omega^2}\right)^{1/3}\left(\frac{r_1^{4/3} - r_0^{4/3}}{\cos^{5/3}\alpha}\right) \tag{30}$$

This assumes that the conical surface extends from radius r_0 to a radius r_1. This equation neglects the influence of gravity and is therefore subject to the same limitations as those expressed in Eq. (27). A total residence time for the nested cone system can therefore be calculated by a sum of Eqs. (28) and (29) for each conical surface. The velocity of the liquid hitting the second surface can also be calculated using the previous assumptions. The perpendicular component of velocity relative to the rotating surface as radius r_2 can be written as

$$u_{\text{imp}} = r_1\omega\sqrt{1 - \frac{r_1^2}{r_2^2}} \tag{31}$$

The tangential component of impact velocity on the second surface will also depend on whether it is moving at the same or a different rotational speed. Assuming the first surface is rotating at ω_1 and the second surface at ω_2 the tangential impact velocity can be written as

$$v_{\text{imp}} = \frac{r_1^2\omega_1 - r_2^2\omega_2}{r_2} \tag{32}$$

Thus, it can be seen that substantial impact velocities, and hence energy available for mixing, can be

Table 1 SDR—scaling influences

	Film thickness	Residence time	Flow velocity	Surface shear
Double liquid flow	26% increase	37% decrease	59% increase	26% increase
Double disk rotation	37% decrease	37% decrease	59% increase	152% increase
Ten times viscosity	115% increase	115% increase	54% decrease	78% decrease
Double disk radius	37% decrease	152% increase	21% decrease	26% increase
Disk to 30° cone	10% increase	27% increase	9% decrease	17% decrease
Disk to 75° cone	146% increase	851% increase	59% decrease	84% decrease

obtained by the use of surfaces moving at different rotational speeds even when the surfaces are in close proximity.

Scaling Rules for Hydrodynamic Properties

In the previous sections equations have been given to estimate a range of properties of the fluid flow over the rotating systems, spinning disks, and more generally spinning cones. These equations can be used to estimate general rules for examining the scaling of conditions from one rotating device to another and also from one set of operating conditions to another. Table 1 summarizes these implied scaling influences.

It should be noted that these rules assume that the liquid flowing over the system has Newtonian properties and is fully synchronized with the disk rotation. The influence of gravity is also neglected and it is assumed that the surface is completely wetted.

Performance Estimators

Estimating transport by diffusion

Owing to the nature of thin films generated by both spinning disks and spinning cones, transport by diffusion can become a significant method of transport of both heat and mass within the film. In the case of heat transfer this is especially true and will be the dominant mode of transport within the film. The equations for transport by diffusion have been studied and solved for many simple systems. One of the most comprehensive studies of this process was given by Crank in which the equations for diffusion are examined for several generalized systems.[11] In the solution of the diffusion equation one parameter is of paramount importance: that is, the Fourier number, *Fo*. This can be expressed as

$$Fo = \frac{Dt}{\delta^2} = \frac{DL}{u\delta^2} \tag{33}$$

where D is the diffusion coefficient for the process, t is the exposure time, and δ is the characteristic path length for diffusion. This can alternatively be written in terms of flow velocity u perpendicular to the process and the length traversed, L, in the direction of flow. In the case of thin films, the diffusion path length can be replaced by the film thickness f. However, for flow over a rotating surface film thickness and velocity are constantly changing and so an integrated form of Eq. (33) is required to examine the potential for diffusive transport through the rotating film. Fourier number for an SDR can be developed based on the following integration:

$$Fo = \int_0^R \frac{D}{uf^2}\,\mathrm{d}r \tag{34}$$

Combining Eqs. (2), (3), (22), (23), and (34) gives

$$Fo = \left(\frac{9\pi^4}{32}\right)^{1/3}\left(\frac{\omega^2 R^8 D^3}{\nu Q^4}\right)^{1/3}\cos^{2/3}\alpha \tag{35}$$

It should be noted that the diffusivity D in Eq. (35) is the molecular diffusivity in the case of mass transfer, which is typically of the order of $10^{-9}\,\mathrm{m}^2/\mathrm{sec}$ or thermal diffusivity defined as

$$D_{\mathrm{HEAT}} = \frac{k}{c_{\mathrm{P}}\rho} \tag{36}$$

in the case of heat transfer. For heat transfer, Eq. (34) will provide a measure of the capability of the film to reach the temperature of the solid surface assuming that it is maintained at a constant temperature. Burns and Jachuck showed that this parameter correlated well for a mass transfer limited process using the conversion of calcium hydroxide into calcium carbonate by diffusion of carbon dioxide through the liquid film on an SDR.[7] Results showed complete conversion for *Fo* in the region of 0.1–0.2. For nonreactive processes and heat transfer, however, efficient transport by diffusion should be expected for *Fo* in excess of 0.3 and preferably in excess of 1.

It can be seen from Eq. (35) that the performance for increased flows can be maintained by use of either higher rotational speed or increased surface area.

Eq. (34) also shows that inclined surfaces should yield slowly, decreasing diffusive transport capabilities compared to the spinning disk, owing to the thicker films. However, for reactions with some kinetic limitations the increased residence time may be more beneficial. It should also be noted that the results of mixing by collisions, motion over structured surfaces and through surface waves, are convective processes not considered in this formulation.

Estimating mass transfer coefficients

Correlations for mass transfer can be divided into diffusive and nondiffusive processes. The simpler of the two to predict is that of diffusive transfer and, in particular, the penetration of a species at either the solid–liquid or liquid–gas interface.

Gas–liquid mass transfer by diffusion

Estimation of liquid-side mass transfer coefficients for gas–liquid transport through the upper surface of the film can be most readily made by assumption of a purely diffusive transfer. Under these conditions, a mass transfer coefficient can be approximated using the following equation for mass transfer:[12]

$$k_{\mathrm{LG}} = \sqrt{\frac{4D}{\pi t}} \tag{37}$$

This assumes that a chemical species is penetrating into a static film over a time period of t with a diffusivity in the liquid of D. If it is assumed that the exposure time constant t in the equation is equal to the residence time of the liquid on a spinning disk surface, given by Eq. (6), then the liquid-side mass transfer coefficient k_{LG} for diffusion into the film can be estimated as

$$\begin{aligned} k_{\mathrm{LG}} &= \left(\frac{1024Q^2\omega^2D^3}{81\pi^5\nu R^4}\right)^{1/6} \\ &\approx 0.59\left(\frac{Q^2\omega^2D^3}{\nu R^4}\right)^{1/6} \end{aligned} \tag{38}$$

A measure of the transport performance can be given by considering the flux through the surface compared to the flux passing over the surface, which is the liquid flow rate. This produces the following dimensionless group:

$$\begin{aligned} \text{Transport performance} &= \frac{k_{\mathrm{LG}}\pi R^2}{Q} \\ &= 1.85\left(\frac{\omega^2D^3R^8}{\nu Q^4}\right)^{1/6} \end{aligned} \tag{39}$$

It may be suggested that the transport performance is proportional to *Fo* defined for the spinning surface given by Eq. (35) as the assumptions used for the diffusive processing are the same.

Liquid–solid mass transfer by diffusion

Mass transfer coefficients for the lower surface of a laminar film are strongly influenced by the shear at the liquid–solid interface. A solution for liquid–solid mass transfer coefficients for a diffusive process in a laminar film was provided by Bird et al. as the following equation:[12]

$$k_{\mathrm{LS}} = \frac{2D}{\Gamma(7/3)}\left(\frac{S}{9DL}\right)^{1/3} \tag{40}$$

where the function $\Gamma(7/3)$ is defined as

$$\Gamma(7/3) = \int_0^\infty x^{4/3}\,\mathrm{e}^{-x}\,\mathrm{d}x = 1.1906 \tag{41}$$

For a spinning disk, the standard model for falling film flow is complicated by the changing thickness and shear as the liquid flows over the disk. An approximation of this to conditions on a spinning disk surface can, however, be made by substitution of Eq. (9) for average liquid–solid surface shear into the above equation for mass transfer. If it is also assumed that the characteristic distance L traveled by the liquid is equal to that of the disk radius then an equation for the liquid–solid mass transfer coefficient k_{LS} can be written for an SDR as

$$k_{\mathrm{LS}} = 0.71\left(\frac{\omega^4D^6Q}{R^2\nu^2}\right)^{1/9} \tag{42}$$

A similar transport performance estimate can be provided for the liquid–solid process through the following equation:

$$\begin{aligned} \text{Transport performance} &= \frac{k_{\mathrm{LS}}\pi R^2}{Q} \\ &= 2.23\left(\frac{\omega^4D^6R^{16}}{\nu^2Q^8}\right)^{1/9} \end{aligned} \tag{43}$$

It should be noted that the grouping shown here is not identical to the *Fo* shown in Eq. (35) because of the strong influence of surface shear as well as molecular diffusion on the overall transport rates.

Mass transfer with surface wave convection—liquid/gas

The influence of surface waves on transport within a film has been shown to lead to enhanced mass transfer above that expected from pure diffusion.[10] This enhancement can be represented in terms of an improvement in the Sherwood number (Sh) as defined by

$$Sh = \frac{k_{LG} f_C}{D} \tag{44}$$

where f_C is a characteristic film thickness for the process. A series of publications have been produced modeling the structure of surface waves produced by liquid flow over a rotating disk surface. Sisoev et al. produced a series of publications examining the fluid dynamics of surface waves on an SDR.[13] Data describing surface waves produced by the disk rotation combined with the mass transfer results produced by Auone and Ramshaw were used to formulate a mathematical model for the mass transfer process.[9,10] Their results indicated that (G. M. Sisoev, O. K. Matar, and C. J. Lawrence, personal communication)

$$Sh = 0.154 Pe_K^{1/2} \tag{45}$$

where

$$Pe_K = \frac{Pe}{\kappa} = \left(\frac{\omega^2 f_C^4}{\nu D}\right)\left(\frac{\rho\omega^2 R^4}{\sigma f_C}\right)^{1/3} \tag{46}$$

and the characteristic film thickness is defined as

$$f_C = \frac{2}{3} f_{AV} = 0.54\left(\frac{Q\nu}{R^2\omega^2}\right)^{1/3} \tag{47}$$

Combining these equations mass transfer coefficient k_{LG} can be estimated as

$$k_{LG} = 0.154\left(\frac{f_C^5 \omega^8 R^4 \rho D^3}{\nu^3 \sigma}\right)^{1/6} \tag{48}$$

Transport performance can therefore be estimated as

$$\text{Transport performance} = \frac{k_{LG}\pi R^2}{Q} = 0.29\left(\frac{\omega^{14} R^{38} \rho^3 D^9}{\nu^4 \sigma^3 Q^{13}}\right)^{1/18} \tag{49}$$

It can be seen that performance under this model has a stronger positive dependence on rotational speed and a weaker negative influence from liquid flow rate owing to the increased surface wave generation under conditions of high speed and high flow. This can aid in the scale-up of spinning disk processes as transport rates are likely to improve in reality if scaling is based on the more simplistic diffusion-only models shown by Eqs. (38) and (39).

SCALING CALCULATIONS

In scaling up of an SDR process it is assumed that liquid properties, such as density and viscosity, are constant and that radius R is altered in the scaling to allow for greater throughput Q. The three parameters in the scaling are therefore liquid flow rate Q, rotational speed ω, and disk radius R. To examine the implications of scaling, the geometry scaling factor F for the system is used and is defined by

$$F = \frac{R_2}{R_1} \tag{50}$$

where a process on a disk radius R_1 is to be scaled to a disk of radius R_2. The following sections provide scaling methods that link the two remaining parameters to the scaling factor F.

Scaling with Residence Time and Film Thickness Constant

The first method of scaling is to preserve residence time and average film thickness. Using Eqs. (6) and (7) implies that

$$\left(\frac{R_1^4}{Q_1^2\omega_1^2}\right) = \left(\frac{R_2^4}{Q_2^2\omega_2^2}\right) \Rightarrow \text{Constant residence time} \tag{51}$$

$$\left(\frac{Q_1}{R_1^2\omega_1^2}\right) = \left(\frac{Q_2}{R_2^2\omega_2^2}\right) \Rightarrow \text{Constant film thickness} \tag{52}$$

Therefore, scaling for this requirement would be

$$Q_2 = F^2 Q_1 \tag{53}$$

$$\omega_2 = \omega_1 \tag{54}$$

This implies that flow rate is proportional to disk area while rotational speed should remain constant. Centrifugal acceleration would be increased in proportion to F for this system, as would radial velocity. Mass transfer performance based on the previous correlation would indicate increased performance from this scaling.

Scaling with Residence Time and Mass Transfer Performance Constant

If it is desired to scale while maintaining the same theoretical mass transfer performance and residence time then the following conditions must be met:

$$\left(\frac{R_1^4}{Q_1^2\omega_1^2}\right) = \left(\frac{R_2^4}{Q_2^2\omega_2^2}\right) \Rightarrow \text{Constant residence time} \tag{55}$$

This means that flow is scaled according to

$$Q_2 = Q_1\left(\frac{\omega_1}{\omega_2}\right)F^2 \tag{56}$$

Fixed mass transfer performance implies

$$\left(\frac{R_1^4\omega_1}{Q_1^2}\right) = \left(\frac{R_2^4\omega_2}{Q_2^2}\right) \tag{57}$$

(Model 1: liquid–gas diffusive)

$$\left(\frac{R_1^4\omega_1}{Q_1^2}\right) = \left(\frac{R_2^4\omega_2}{Q_2^2}\right) \tag{58}$$

(Model 2: liquid–solid shear/diffusion)

$$\left(\frac{R_1^{38}\omega_1^{14}}{Q_1^{13}}\right) = \left(\frac{R_2^{38}\omega_2^{14}}{Q_2^{13}}\right) \tag{59}$$

(Model 3: liquid–gas convective)

Therefore, the scaling rules for these three mass transfer models are

- Liquid–gas diffusive

$$Q_2 = F^2Q_1 \tag{60}$$

$$\omega_2 = \omega_1 \tag{61}$$

- Liquid–solid shear/diffusion

$$Q_2 = F^2Q_1 \tag{62}$$

$$\omega_2 = \omega_1 \tag{63}$$

- Liquid-gas with surface wave convection

$$Q_2 = F^{22/9}Q_1 \tag{64}$$

$$\omega_2 = F^{-4/9}\omega_1 \tag{65}$$

This implies that mass transfer by shear and by diffusion scales flow in proportion with disk area, while keeping rotational speed constant. If, however, convective transport due to surface waves is included the rotational speed is reduced as disk size is increased and flow is increased slightly more than proportional with disk area. However, wetting characteristics and absolute applied acceleration are not kept constant under these laws and will change as the disk is scaled up for a given mass transfer and residence time performance. Ultimately, their effects on the process will need to be studied before the scaling laws can be used; however, these can be used to provide a rough engineering guide to process capabilities.

Estimating Heat Transfer Performance for an SDR

In the previous sections we have discussed mass transfer into the thin centrifuged film occurring on a disk surface. For many processes diffusive transport can be seen to be a significant, if not the most significant, mechanism for transport. In the case of heat, transport by diffusion, that is, conduction, can be considered even more effective as diffusion coefficients for that process are substantially higher. Therefore, transport by conduction can be used as a good estimate for heat transfer performance through the film.

Substituting thermal diffusion for molecular diffusion allows the computation provided for liquid–solid mass transfer to be transferred to thermal diffusion using the following equation:

$$h_{\text{FILM}} = k_{\text{LS,THERM}}c_{\text{P}}\rho \tag{66}$$

where $k_{\text{LS,THERM}}$ is the equivalent transport coefficient for thermal diffusion as defined by Eq. (36). Combining this with Eq. (42) for transfer assuming a constant surface temperature gives an effective heat transfer coefficient for the film as

$$h_{\text{FILM}} = 0.71\left(\frac{\omega^4k^6Qc_{\text{P}}^3\rho^3}{R^2\nu^2}\right)^{1/9} \tag{67}$$

If the disk surface temperature is assumed to be held constant then the thermal power transfer capability into the film can be estimated using the following equation:

$$\text{Power} = 0.71\left(\frac{\omega^4k^6Qc_P^3\rho^3}{R^2\nu^2}\right)^{1/9}\pi R^2\Delta T_{\text{LMTD}} \tag{68}$$

where ΔT_{LMTD} is the log-mean temperature difference between the film and the disk surface. Heat transfer in practice, however, is limited by the design of the spinning disk system below the film. One common design route used for the reactors is the supply of heating or cooling using a heat transfer fluid. If it is assumed that this fluid can transfer heat to the underside disk of thickness d with a performance of h_{HTF}, then the overall heat transfer

coefficient U for the system can be estimated using

$$U = \left(\frac{1}{h_{\mathrm{FILM}}} + \frac{k_{\mathrm{DISK}}}{d} + \frac{1}{h_{\mathrm{HTF}}}\right)^{-1} \quad (69)$$

where k_{DISK} is the thermal conductivity of the spinning disk material.

Quantifying Power Input to the Fluid

Power consumed by a spinning disk and cone processes can be divided into various contributions. These include the power losses to bearing friction as well as drive inefficiencies, neither of which contributes directly to the power input of the fluid. A large component of power may also be required to initially spin-up the disk surface, this being a function of the disk mass and start-up time allowable. Power input to the fluid itself, however, may only be a fraction of the total power demand but is that which can be related to the processing conditions.

For a spinning disk and spinning cone reactor, power input to the fluid can be represented as that given to accelerate the liquid up to the tip speed of the disk. This can be written as

$$\text{Power} = \frac{\rho Q R^2 \omega^2}{2} \quad \text{(spinning disk/cone)} \quad (70)$$

Power is input to the fluid from a nested cone system that has nonsynchronized upper and lower rotating surfaces. An equation to estimate the power input from that design can be written as

$$\text{Power} = \frac{\rho Q R^2 \omega^2}{2} + \sum_{i=1}^{n}\left(\frac{u_{\mathrm{imp},i}^2 + v_{\mathrm{imp},i}^2}{2}\right)\rho Q \quad \text{(nested cone)} \quad (71)$$

where the nested cone has n impacts with relative velocities of u_{imp} and v_{imp} as given by Eqs. (31) and (32) before the liquid leaves at the edge of the system at a radius R. Energy input to the liquid per unit volume can also be calculated from the above equations using

$$\frac{\text{Energy}}{\text{Volume}} = \frac{\text{Power}}{Q} \quad (72)$$

This can be used to compare these reactors with that of a batch stirred reactor by using the following equation:

$$\frac{\text{Energy}}{\text{Volume}} = \frac{\text{Power} \times \text{Batch time}}{\text{Batch volume}} \quad (73)$$

Process Application of SDRs

The ability to handle viscous and inviscid fluids with considerable ease makes the SDR ideally suited for performing a range of process operations. It has considerable use in food, fine chemicals, polymers, energy, household products, water treatment, and the pharmaceutical industry. It is a unique treatment of fluids under the centrifugal field and has been shown to enhance rates of reaction by orders of magnitude. The SDR polymerization of styrene, butyl acrylate, and esters has been successfully demonstrated by Boodhoo and Jachuck.[2,3] Thin films on the disk also offer the opportunity to use ultraviolet radiation to trigger fast chemical reactions. Intense mixing within the films has been used to generate micro to submicrometer sized particles of barium sulfate and calcium carbonate.[14,15] Nano composites and continuous production of fine particles for high-value application are attractive possibilities with SDRs. The high mass transfer rates coupled with the ability to provide short sharp bursts of high temperature have been exploited by Protensive Limited to perform devolatilization of polymers or removal of lighter organic molecules from viscous polymer melts. In conclusion, it may be stated that SDR technology offers several opportunities for the processing sector, such as process flexibility, improved product quality, speed to market, just-in-time manufacturing, reduced footprint, improved inherent safety and energy efficiency, distributed manufacturing capability, and ability to use reactants at higher concentrations.

CONCLUSIONS

This chapter has provided a summary of the characteristics, potential applications, hydrodynamic equations, and transport expressions that govern the operation of SDR. The reader is encouraged to refer to the References particularly for more information on the application of SDR. Currently, these reactors are being studied extensively by industry under pilot-plant programs and it is expected that details of the case studies will be published soon.

NOTATION

- D Diffusivity of transferring quantity (molecule/heat/momentum) (m^2/sec)
- E Ekman number
- F Scaling factor
- g Acceleration due to applied field (gravitational/centrifugal) (m/sec^2)
- H Film thickness (m)
- K_1 Empirical constant for Pigford model

K_2 Empirical constant for Pigform model
Q Volumetric flow rate over the disk surface (m^3/sec)
r Radial position (m)
R Radius of spinning disk (m)
R_F Radius of central feed pipe (m)
R_S Radius of spin-up zone (m)
S Surface shear (sec^{-1})
t Residence time on the disk (sec)
u Velocity in the direction of flow or radial velocity (m/sec)
v Tangential flow velocity (m/sec)
x Mesh pore size (m)
α Angle of inclination of the cone (rad)
β Wetting factor
λ Ekman length scale (m)
σ Interfacial tension (kg/sec^2)
ρ Liquid density (kg/m^3)
θ Contact angle between liquid and disk surface (rad)
ω Rotational speed of the disk (rad/sec)
ν Kinematic viscosity (m^2/sec)

Common Subscripts

AV	Average over the disk
A	Average through the film
CONE	Conditions for a spinning cone
DISK	Conditions for a spinning disk
1	Initial condition
2	Scaled condition

REFERENCES

1. Jachuck, R.J.J.; Ramshaw, C. Process intensification: heat transfer characteristics of tailored rotating surfaces. In *Heat Recovery Systems & CHP*; Elsevier, 1994; Vol. 14, No. 5, 475–491.
2. Boodhoo, K.V.K.; Jachuck, R.J.J. Process intensification: spinning disk reactor for styrene polymerisation. In *Applied Thermal Engineering*; Elsevier, 2000; Vol. 20, 1127–1146.
3. Boodhoo, K.V.K.; Jachuck, R.J.J. Process intensification: spinning disk reactor for condensation polymerisation. In *Green Chemistry*; RSC Publishing, 2000; Vol. 4, 235–244.
4. Vicevic, M; Jachuck, R.J.J.; Scott, K. Process intensification for green chemistry: rearrangement of alpha-pinene oxide on the spinning disk. 4th International Conference on Process Intensification for the Chemical Industry Brugge, Belgium, Sep 10–12, 2001.
5. Nusselt, W. Die oberflächenkondensation des wasserdampfes. Z. Ver. Deut. Ing. **1916**, *60*, 541.
6. Wood, R.M.; Watts, B.E. The flow, heat, and mass transfer characteristics of liquid films on rotating disks. Trans. Inst. Chem. Eng. **1973**, *51*, 315–322.
7. Burns, J.R.; Ramshaw, C.; Jachuck, R.J.J. Measurement of liquid film thickness and the determination of spin-up radius on a rotating disk using an electrical resistance technique. Chem. Eng. Sci. **2003**, *58* (11), 2245–2253.
8. Hartley, D.E.; Murgatroyd, W. Criteria for the break-up of thin liquid layers flowing isothermally over solid surfaces. Int. J. Heat Mass Transfer **1964**, *7*, 1003–1015.
9. Woods, W.P. The Hydrodynamics of Thin Films Flowing over a Rotating Disk. Ph.D. Thesis, University of Newcastle-upon-Tyne, U.K. 1995.
10. Aoune, A.; Ramshaw, C. Process intensification: heat and mass transfer characteristics of liquid films on rotating disks. Int. J. Heat Mass Transfer **1999**, *42*, 2543–2556.
11. Crank, J. *The Mathematics of Diffusion*, 2nd Ed.; Clarendon Press: Oxford, 1975; 44–68.
12. Bird, R.B.; Stewart, W.E.; Lightfoot, E.N. *Transport Phenomena*; Wiley: New York, London, 1960; 551–552.
13. Sisoev, G.M.; Matar, O.K.; Lawrence, C.J. Axisymmetric wave regimes in viscous liquid film flow over a spinning disk. J. Fluid Mech. **2003**, *495*, 385–411.
14. Cafiero, L.M.; Chianese, A.; Baffi, G.; Jachuck, R.J.J. Process intensification: precipitation of barium sulfate using a spinning disk reactor. Ind. Eng. Chem. Res. **2002**, *41* (21), 5240–5246.
15. Trippa, G.; Jachuck, R.J.J. Process intensification: precipitation of calcium carbonate from carbonation reaction of lime water using a spinning disk reactor. ICheaP-6, Pisa, Italy, Jun 8–11, 2003.

Styrene

Guy B. Woodle
UOP LLC, Des Plaines, Illinois, U.S.A.

INTRODUCTION

Styrene is one of the most important aromatic monomers used for the manufacture of plastics. Small-scale commercial production of styrene began in the 1930s. Demand for styrene-based plastics has grown significantly, and in 2003 the worldwide annual production capacity was approximately 24.5 million metric tons.[1]

About 65% of styrene is used to produce polystyrene. Polystyrene is used in the manufacture of many commonly used products such as toys, household and kitchen appliances, plastic drinking cups, housings for computers and electronics, foam packaging, and insulation. Polystyrene finds such widespread use because it is relatively inexpensive to produce and is easy to polymerize and copolymerize, resulting in plastics with a broad range of characteristics. In addition to polystyrene, styrene is used to produce acrylonitrile–butadiene–styrene polymer, styrene–acrylonitrile polymer, and styrene–butadiene synthetic rubber (SBR).

The development of styrene technologies was mainly driven by demand for cheap synthetic rubber during and immediately after World War II. Between 5% and 10% of total styrene produced becomes a component of synthetic rubbers, which are copolymers of styrene and butadiene (SBR). Styrene copolymers containing acrylonitrile are specialty materials that are used for specific applications. Demand for styrene for the period 2004–2009 is estimated to grow at a rate of approximately 4% per year.[1]

PHYSICAL AND CHEMICAL PROPERTIES

Styrene is a colorless aromatic liquid. It is only very slightly soluble in water, but infinitely soluble in alcohol and ether. Additional properties are listed in Table 1.

Styrene is chemically reactive with the most important reaction being its polymerization to form polystyrene. Styrene can also copolymerize with other monomers, such as butadiene and acrylonitrile, to produce a variety of industrially important copolymers.

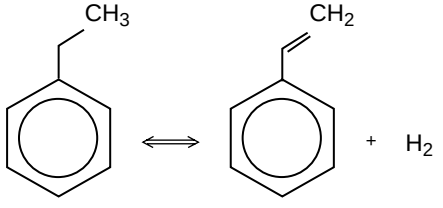
In addition to polymerization, styrene can undergo other types of reactions due to the chemical nature of its unsaturated side chain and aromatic ring. For example, styrene can be oxidized to form benzoic acid, benzaldehyde, styrene oxide, and other oxygenated compounds. Styrene oxide is used in the production of various cosmetics, perfumes, agricultural and biological chemicals.

REACTION KINETICS AND THERMODYNAMICS

Essentially all commercially produced styrene uses ethylbenzene (EB) as a feedstock. Between 85% and 90% of worldwide styrene production is based on EB dehydrogenation. The remaining 10–15% of styrene is obtained as a coproduct in a process to produce propylene oxide.

Ethylbenzene Dehydrogenation

Ethylbenzene is catalytically dehydrogenated in the presence of steam according to the equation:



The reaction is highly endothermic and conversion is limited in extent by equilibrium. The reaction equilibrium constant is defined as:

$$K\mathrm{eq} = (P\mathrm{sty} * P\mathrm{h2})/P\mathrm{eb}$$

where *P*sty is the partial pressure of styrene, *P*h2 is the partial pressure of hydrogen; and *P*eb is the partial pressure of ethylbenzene.

High temperature, steam dilution, and low system pressure produce an equilibrium more favorable to styrene. For endothermic vapor-phase reactions, the equilibrium constant increases with temperature and

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007970

Table 1 Physical properties of styrene

Molecular weight	104.152									
Specific gravity[a]	0.903									
Melting point, °C	−30.628									
Boiling point, °C	145.2									
Critical temperature, °C	373									
Critical pressure, atm	46.1									
Vapor pressure, mm°Hg at T °C	1	5	20	10	40	60	100	200	400	760
	−7.0	18.0	30.8	44.6	59.8	69.5	82.0	101.3	122.5	145.2

[a]Density is at 20°C referred to water at 4°C.
(From Perry, R.H., Green, D.W., Eds.; *Perry's Chemical Engineers Handbook*, 6th Ed.; McGraw-Hill: New York, 1984; 3-60 and Miller, S.A., Ed.; *Ethylene and Its Industrial Derivatives*; 901 pp.)

can be determined according to the following equation:[2]

$$\ln Keq = 16.12 - (15,350/T)$$

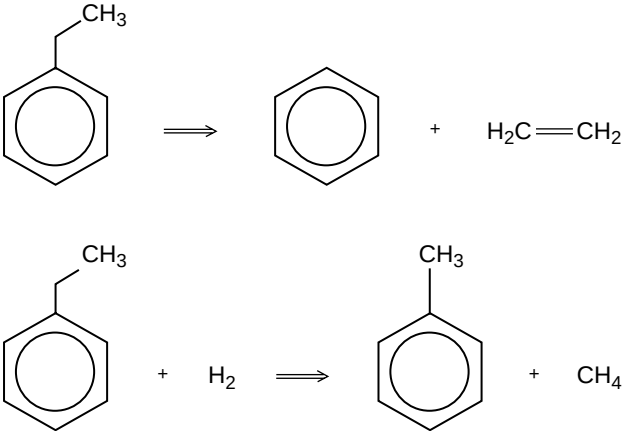
where Keq is the equilibrium constant in atmospheres and T is the temperature in K.

The equilibrium constant has the dimension of pressure since two moles of products are formed for each mole of EB converted. Therefore, a higher total pressure will shift the reaction equilibrium to the left and reduce EB conversion. Lower pressure results in greater EB conversion without an accompanying significant decrease in styrene selectivity.

Another method to create a positive shift in equilibrium is the use of steam dilution to reduce the partial pressures of EB, styrene, and hydrogen. Steam dilution provides the same effect as a reduction in total pressure.

Steam dilution has several other important benefits. First, steam supplies heat to the reacting mixture. Consequently, the drop in temperature for a given EB conversion is lower, allowing greater EB conversions to be obtained with the same inlet temperature. Second, a minimum amount of steam appears to keep the catalyst in the required oxidation state for high activity. The actual quantity of steam varies with the type of catalyst used. Third, steam is believed to suppress the deposition of carbonaceous material on the catalyst. If the carbonaceous material is allowed to accumulate, the catalyst will become fouled and its activity will decline to unacceptable levels.

The reaction feed mixture undergoes certain other reactions that are not equilibrium limited under typical operating conditions. Most important among these are the dealkylation reactions that result in the formation of benzene and ethylene or toluene and methane. Other reactions produce small amountsof α-methylstyrene and other high boiling components. The key dealkylation reactions can be described by the following equations:



Both methane and ethylene undergo steam reforming reactions according to the following equations:

$$CH_4 + H_2O \rightarrow CO + 3H_2$$

$$C_2H_4 + 2H_2O \rightarrow 2CO + 4H_2$$

The water–gas shift reaction also occurs and is generally near equilibrium at the reaction temperature:

$$CO + H_2O \leftrightarrow CO_2 + H_2$$

The combination of dealkylation, steam reforming, and water–gas shift side reactions should be avoided, if possible. In addition to losing valuable EB feed by dealkylation, the resultant net formation of carbon dioxide and hydrogen by this combination of reactions inhibits the primary dehydrogenation reaction. The net hydrogen formation gives an unfavorable shift in equilibrium, while the presence of carbon dioxide has a negative effect on dehydrogenation catalyst activity.[3]

Typically, there is less methane and ethylene present in the effluent of a reactor than would be expected from the benzene and toluene formation. Carbon monoxide is generally about 10 mol% of the total carbon oxides.

The critical operating and design parameters for EB dehydrogenation are discussed in the following paragraphs.

Reaction temperature

Because the dehydrogenation reaction is endothermic, the reaction mixture temperature decreases as the reaction proceeds. The reaction rate slows because of the closer approach to equilibrium and the decrease in kinetic reaction rate with the decreasing temperature. Furthermore, the equilibrium constant is less favorable at lower temperature. Therefore, in a normal design, about 80% of the temperature drop occurs in approximately the first third of the catalyst bed.

As a result, a high inlet catalyst temperature is required. However, high temperature also increases the rates of nonselective thermal reactions and dealkylation reactions, which form benzene and toluene by-products. In particular, as temperature is increased, the rate of benzene formation increases significantly relative to the rate of styrene formation. This means there is an effective upper limit to the inlet temperature if high styrene selectivity is a required criterion. Reaction temperature is generally adjusted by changing either the steam temperature or the steam-to-oil ratio.

Catalyst quantity

The amount of catalyst relative to EB feed is an important parameter for optimum reactor performance. Too little catalyst will prevent a close approach to equilibrium. If EB conversion is low, then distillation costs associated with recovery and recycle of the unconverted EB can become significant. With too much catalyst, the EB conversion reaches equilibrium before the outlet of the catalyst bed, while the side reactions continue leading to loss of selectivity.

The optimum catalyst quantity is achieved by balancing the EB conversion level and the styrene yield. Catalysts typically lose activity with time on-stream, which has the effect of decreasing the effective active catalyst quantity for reaction. Compensation for aging catalyst is achieved by adjusting other operating parameters, in particular, the reaction temperature.

Reaction pressure

Ethylbenzene dehydrogenation results in a significant increase in the volume of reactants due to the reaction stoichiometry. Lower pressure favors higher equilibrium conversion to styrene. Reaction pressure is established during the plant design at the lowest practical level. Modern commercial reactors operate below atmospheric pressure. Pressures as low as 300 mm Hg or lower are common. The key side reactions are largely independent of reaction pressure; hence, operating at lower pressures also provides higher styrene yield.

Steam dilution or steam-to-oil ratio

The main functions of steam dilution are to act as a diluent to reduce the hydrocarbon partial pressures, providing heat for the endothermic dehydrogenation reaction, and maintaining the catalyst's active surface in a desirable state. Increasing the steam-to-oil ratio has the net effect of improving the EB conversion and styrene yield. However, costs associated with generating and superheating the dilution steam also increase and eventually offset the reaction advantages.

Catalyst type and properties

Ethylbenzene dehydrogenation is generally catalyzed by a potassium-promoted iron oxide catalyst. The most widely used catalysts are composed of iron oxide, potassium carbonate, and various metal oxide promoters. Examples of metal oxide promoters include chromium oxide, cerium oxide, molybdenum oxide, and vanadium oxide.[4] The potassium component substantially increases catalyst activity relative to an unpromoted iron oxide catalyst. Potassium has been shown to provide other benefits. In particular, it reduces the formation of carbonaceous deposits on the catalyst surface, which prolongs catalyst life.

Properties such as catalyst size and shape also impact performance. In theory, smaller sized catalyst will increase reaction rates by providing more available catalyst surface area than larger sized catalyst. Small catalyst particles, however, have a disadvantage in that they result in greater pressure drop through a reactor and higher overall reaction pressures. To address this, catalyst developers have used specialized shapes, such as ribbed extrudates, to gain the advantage of increased surface area without incurring the penalty of increased pressure drop and reaction pressure.

The Sud-Chemie Group and Criterion Catalysts are the major catalyst developers and manufacturers for the styrene industry. Both companies offer a wide range of catalysts to suit individual processing needs. Ethylbenzene conversion, styrene selectivity, catalyst activity, and catalyst stability can be optimized by selecting the best catalyst or a combination of catalysts for a particular application. Dow and BASF manufacture proprietary catalysts, which have been mainly for use in their own respective technologies.

Propylene Oxide with Styrene Co-production

In the late 1960s, a method was discovered to produce propylene oxide by the epoxidation of propylene using organic hydroperoxides as the epoxidizing agent.[5] During the epoxidation reaction, the hydroperoxide is essentially converted to the corresponding alcohol, which in turn can be dehydrated to a more desirable coproduct. Styrene is coproduced in the form of this process that uses EB hydroperoxide as the epoxidizing agent. The chemistry of this process can be broken down into three main reactions as shown in Fig. 1.

The first step is oxidation of EB to form EB hydroperoxide. The oxidation is carried out in the liquid phase with a target EB conversion of approximately 13%.[6] Although higher conversions are attractive from an EB recovery and recycle standpoint, there is a significant disadvantage because the EB hydroperoxide selectivity declines sharply. The second step is epoxidation of propylene to form propylene oxide product and 1-phenylethanol. In the last step, the 1-phenylethanol is dehydrated to styrene and water. The dehydrated reaction mixture is typically stripped of light components and rerun in a styrene column to remove heavy by-products, resulting in a purified styrene product.

The design and operation of a propylene oxide/styrene process plant is complicated and includes numerous pieces of equipment. As a result, the total investment cost for a commercial-scale plant is about four times that of an EB dehydrogenation plant to produce the same quantity of styrene product.

COMMERCIAL PRODUCTION

Reactor Design

One important aspect of modern day EB dehydrogenation reactor design is managing the operating conditions to minimize thermal reactions. The major by-product from the thermal reaction of EB to styrene is benzene with significant subsequent conversions to a complex mixture of higher aromatics, such as anthracene and/or pyrene, as well as coke. Thermal reactions do not occur at a significant level below about 600°C, but become a considerable factor affecting overall yield when temperatures rise above 655°C.

One technique to reduce thermal reactions is to delay heating the EB to the reaction inlet temperature until the last possible moment before being exposed to the catalyst. The method involves superheating EB vapor, along with a portion of the dilution steam, to a temperature below approximately 580°C. The EB is vaporized with a certain amount of steam—commonly called primary steam—to suppress coking. The EB primary steam is combined with the major part of the dilution steam immediately prior to entering the dehydrogenation catalyst bed. The major portion of the dilution steam is generally referred to as main steam. The main steam is superheated to a temperature such that, when it is mixed with the EB and the primary steam, the total combined feed mixture reaches the desired catalyst inlet temperature.

Reactor design and catalyst bed configuration are key factors for controlling thermal reactions.

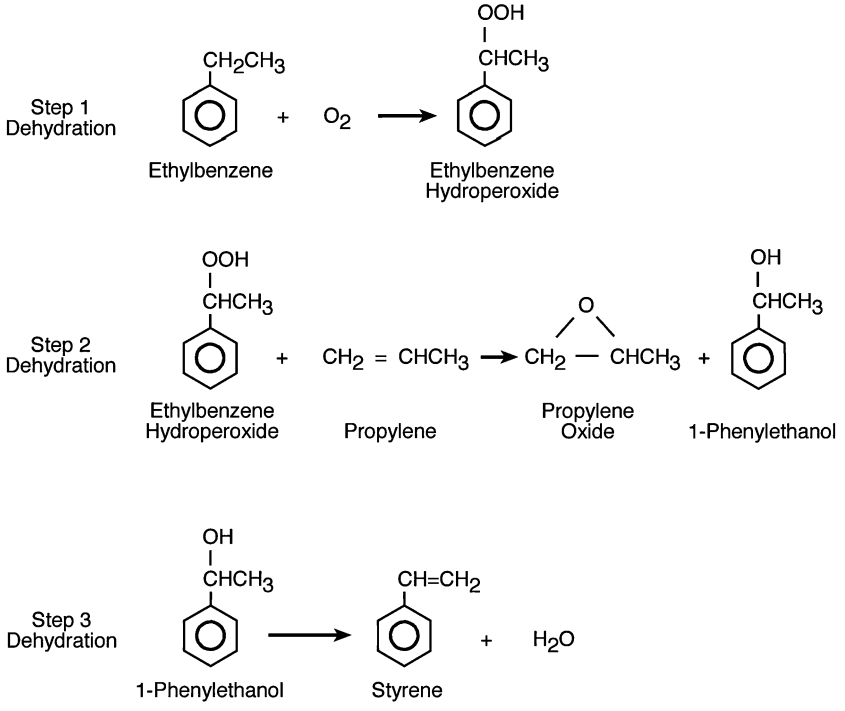


Fig. 1 Propylene oxide–styrene process chemistry.

Commercial adiabatic reactors are typically of radial flow construction with the flow path moving from in to out. This radial outflow geometry requires a much lower inlet volume to obtain proper distribution of the feed vapor through the catalyst bed than either an axial flow or a radial inflow reactor configuration. The radial flow reactor design also provides the advantage of low pressure drop since the flow path through the catalyst is much shorter relative to an axial flow reactor. To minimize thermal reactions, the reactor centerpipe diameter should be as small as possible to minimize residence time at the highest temperature throughout the reactor. However, too small a diameter will produce a high pressure drop through the centerpipe, potentially causing flow maldistribution and causing the feed vapor to enter the catalyst bed with a velocity that can result in erosion and attrition of catalyst particles.

A single-stage reactor with practical limits of temperature, pressure, and steam dilution is limited to 40–50% per pass conversion of EB. If the single-stage reactor effluent is reheated, the reaction mixture moves away from equilibrium allowing for higher EB conversion. When the reheated reaction mixture is fed to a second stage of catalyst, then total EB conversions of 60–75% per pass can be achieved. This process of reheating and adding catalyst stages can be repeated as frequently as economically feasible. With each additional reaction stage, however, a progressively smaller incremental EB conversion is achieved, generally with a corresponding decrease in styrene selectivity.

To obtain high EB conversions, typically two or three reactors are used in series with some type of reheating between the reactors to raise the temperature of the reaction mixture. Modern day commercial reactors are highly engineered. Designers use specialized computational fluid dynamics programs to study flow characteristics throughout a reactor.

Commercial Adiabatic Dehydrogenation Processes

Most commercial styrene plants are based on either the Lummus/UOP technology or the Fina/Badger technology. Dow Chemical is a major styrene producer and uses its own technology. These technologies are generally similar, but there are key differences in the details.

Lummus/UOP Classic SM™ Process

The first commercial plant based on the Lummus/Monsanto technology, which later became the Lummus/UOP technology, was commissioned in 1972. Since that time, more than 50 projects have been licensed with more than 40 plants in commercial operation as of 2004.

A typical Lummus/UOP Classic SM process flow diagram is shown in Fig. 2. Fresh and recycled EB are combined with steam and fed to the dehydrogenation reaction section of the plant. The reactor effluent is condensed and separated into off-gas, process condensate, and a dehydrogenated mixture. The hydrogen rich off-gas stream is recovered through an off-gas compressor for use as a fuel gas. The process condensate is stripped of organics and either recycled for use within the styrene plant or exported. The dehydrogenated mixture, consisting mainly of unconverted EB, styrene product, benzene, and toluene, is fed to the distillation section of the plant.

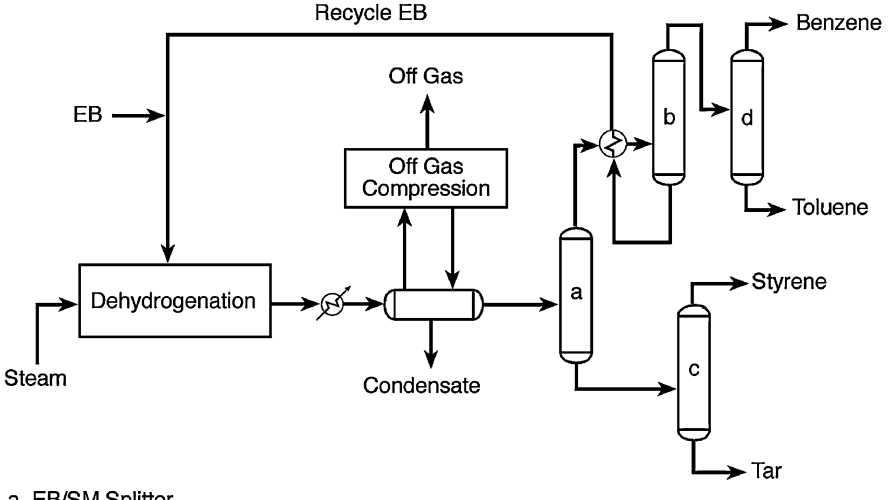


Fig. 2 Lummus/UOP classic SM process.

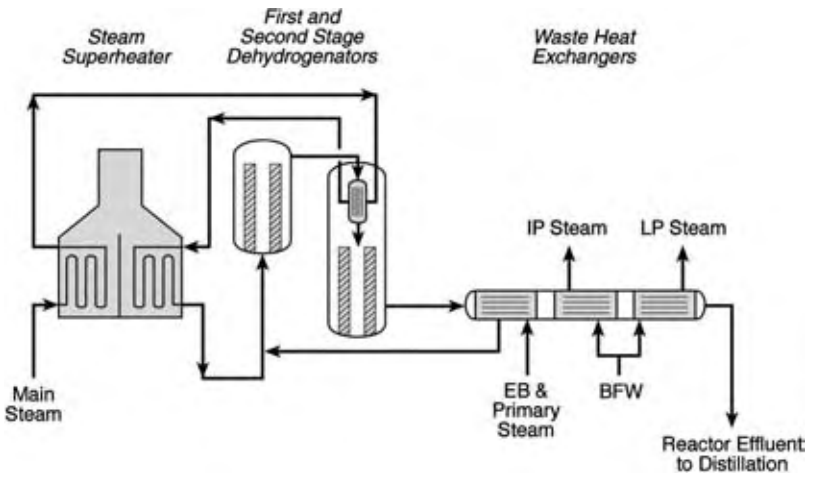


Fig. 3 Lummus/UOP classic SM process dehydrogenation section.

The main equipment in the dehydrogenation reaction section of a Lummus/UOP Classic SM plant includes a steam superheater, two dehydrogenation reactors, a series of waste heat exchangers, and an off-gas compressor (Fig. 3). The equipment is designed to minimize pressure drop from the dehydrogenation reactors inlet to the off-gas compressor.

The main steam is superheated and used to reheat the reaction mixture for the second stage dehydrogenator. The reaction mixture is reheated in a specially designed interchanger located inside the second stage dehydrogenator vessel shell. The cooled steam exiting the interchanger is reheated in the steam superheater prior to being fed to the first stage dehydrogenator. The superheated steam can range from 700°C to as high as approximately 850°C to achieve the desired inlet temperature for the first stage dehydrogenator.

Superheated main steam is mixed with the EB and the primary steam immediately before entering the first stage dehydrogenator. The reactor is designed to provide a uniform reaction mixture while minimizing residence time in the centerpipe to avoid thermal reactions. The reactor effluent is cooled in a series of three waste heat exchangers before final cooling and condensing.

The first stage of waste heat recovery is used to superheat the EB and the primary steam. Subsequent stages are used to generate steam at different pressures. Typically intermediate pressure steam and low pressure steam are generated, which are directed for use elsewhere in the styrene plant or larger EB–styrene complex.

Hydrogen and light hydrocarbons removed from the condensed reactor effluent are compressed and used as fuel gas in the steam superheater. The process steam from the reactor effluent stream is condensed and separated by gravity from the liquid hydrocarbon components. The condensate is stripped of hydrocarbons and revaporized for use as process steam.

The distillation section of a Lummus/UOP Classic SM plant consists of four distillation columns. The first column in the sequence splits the EB and the lighter components from styrene. The EB/styrene monomer (EB/SM) splitter is operated under vacuum and uses structured packing, such as Sulzer Mellapak Plus packing, to minimize temperature and polymer formation.[7] Polymerization inhibitors are injected into the splitter to restrict polymer formation, in particular into the bottom section of the column.

The overhead product from the EB/SM splitter is fed to an EB recovery column. The EB recovery column net bottoms' stream is recycled to the dehydrogenation section. Benzene and toluene by-products in the recovery column overhead stream are separated in a benzene/toluene splitter. Oftentimes, the benzene recovered in this scheme is recycled as feed to the upstream EB plant.

The EB/SM splitter bottoms' stream is fed to the SM column where the styrene is purified by removal of any heavy residual tars. Tertiary-butyl catechol (TBC) is injected into the overhead of the SM column, and the column is operated under vacuum to minimize polymer formation.

A unique feature of the Lummus/UOP Classic SM process is the noncompressive azeotropic heat recovery option.[8] In this option, the EB/SM splitter overhead vapor is used to boil an EB–water azeotrope mixture, which is then fed to the dehydrogenation reactors. The condensation of the splitter overhead vapor produces approximately 500 kcal/kg styrene. This energy savings potential makes the azeotropic heat recovery option economically attractive, in particular, in regions with moderate to high steam costs.

Lummus/UOP Smart SM™ Process

The Lummus/UOP Smart SM process is based on an oxidative reheat technology invented by UOP.[9] Although this technology can be used in the design of

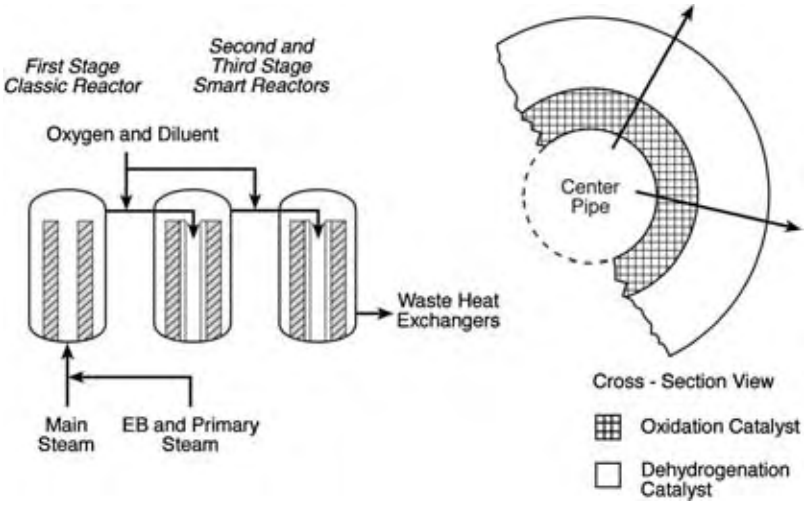


Fig. 4 Lummus/UOP smart SM process dehydrogenation section.

a grassroots plant, it is most commonly used in a revamp of an existing plant to increase styrene production by as much as 60% with minimal capital investment cost.

The Lummus/UOP Smart SM technology uses a specially designed reactor that contains two concentric catalyst zones. A cross-sectional view of the concentric oxidation and dehydrogenation catalyst beds is also shown in Fig. 4. In the first zone, hydrogen is selectively oxidized across a noble metal-containing catalyst. The direct combustion of hydrogen reheats the reaction mixture, which is directly fed into the second zone where the standard EB dehydrogenation reaction occurs. In addition to providing the full reheating requirement, another benefit of this technology is it shifts the reaction equilibrium in a favorable direction by removing the hydrogen by-product. This shift in equilibrium allows for higher EB conversion without a corresponding decrease in styrene yield.

The Lummus/UOP Smart SM technology was first commercialized in 1995 at Mitsubishi Chemical in Kashima, Japan. The Mitsubishi Chemical plant was designed with a dehydrogenation section containing two combination oxidation–dehydrogenation reactors as shown in Fig. 4.

The temperature rise in the oxidation zone is proportional to the amount of oxygen reacted across the catalyst bed. The oxygen is diluted in steam and the oxygen/steam mixture is well mixed to ensure the reaction mixture remains outside the flammability envelope at all times.

Fina/Badger Styrene Process

The Fina/Badger styrene process has evolved through many generations. The most recent design uses a flow diagram as shown in Fig. 5. Recycled and fresh EB

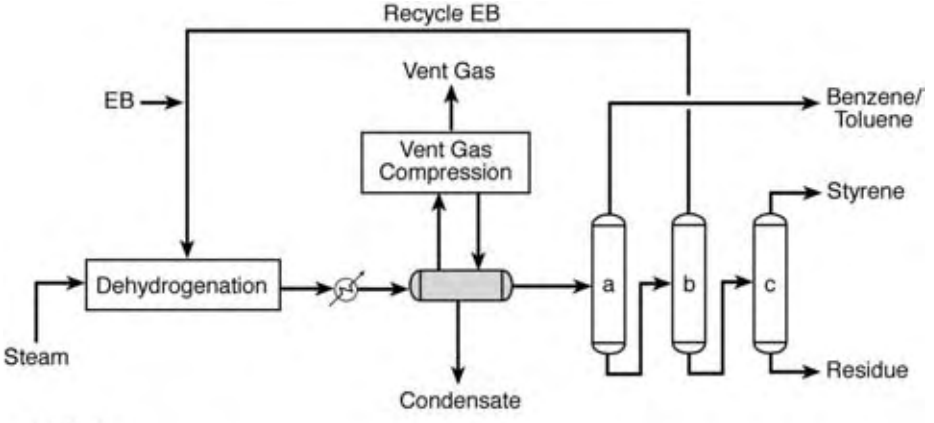


Fig. 5 Fina/Badger styrene process.

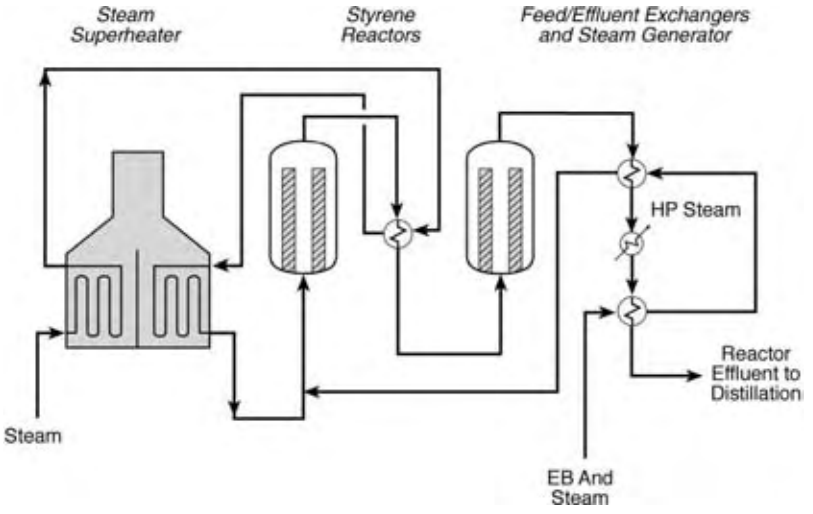


Fig. 6 Fina/Badger styrene process dehydrogenation section.

are mixed with steam and fed to the primary and the secondary dehydrogenation reactors. The reactor effluent is condensed and separated into vent gas, condensate, and hydrocarbon. The vent gas, the majority of which is hydrogen, is used as fuel gas. The condensate is stripped and used as feed water for steam generation. The hydrocarbon portion of the reactor effluent is fed to the distillation section of the plant, which consists of three distillation columns.

The main types of equipment in the dehydrogenation section of the plant are the steam superheater, the primary and secondary dehydrogenation reactors, and a series of feed/effluent exchangers (Fig. 6). High pressure steam is also generated by the recovery of heat from the reactor effluent stream.

The major portion of steam is superheated and used to reheat the reaction mixture for the secondary dehydrogenation reactor. As the cooled steam exits the reheater it is superheated again in the steam superheater, prior to being fed to the primary dehydrogenation reactor. The dehydrogenation reactors are designed to provide low pressure drop and uniform flow distribution. The reactor effluent is cooled in a series of three heat exchangers that heat the EB and steam feed to the reactors and generate steam.

The Fina/Badger distillation section consists of three distillation columns. All the columns are designed to operate under vacuum to minimize temperature and polymer formation. The first column in the sequence splits the benzene and toluene by-products from the unconverted EB and styrene product. The benzene and toluene mixture is typically sent to an integrated EB plant where it is further fractionated. In this case, the benzene by-product is ultimately consumed in the EB unit and the toluene becomes a by-product stream from the EB plant.

The EB recycle column separates the unconverted EB for recycle to the dehydrogenation reactors. Recent EB recovery columns use high efficiency packing to obtain minimum pressure drop through the column. This allows the column bottoms' temperature to be maintained below 100°C. This is an important aspect of the design as styrene polymerization becomes significant at temperatures higher than approximately 100°C.

The EB recovery column bottoms' stream is fed to a finishing column where the styrene is purified by the removal of any heavy residue. Tertiary-butyl catechol is injected into the overhead of the finishing column to prevent polymerization. Tertiary-butyl catechol is widely used to prevent styrene polymerization during storage.

In 1997, Fina/Badger joined with Shell Technology Ventures, a subsidiary of Shell Oil Company, to develop a reheating technology called Flameless Distributed Combustion (FDC) for application in EB dehydrogenation.[10] Flameless Distributed Combustion technology is patented by Shell Oil Company and was originally used as a heat injector for enhanced recovery of hydrocarbons from subterranean formations.

Flameless Distributed Combustion technology enables specific constraints in the conventional dehydrogenation system to be overcome, in particular designing for low steam-to-oil ratios. A low steam-to-oil ratio is desirable because of the substantial energy savings associated with superheating less steam. However, a practical lower steam-to-oil ratio limit exists due to the metallurgy of the steam superheater, steam transfer lines, and interstage reheater. Flameless Distributed Combustion allows for operation at molar steam-to-oil ratios less than 7:1 without a costly metallurgy upgrade. This is accomplished by heating the reaction mixture more directly through a combustion and convective heat transfer process.

Flameless Distributed Combustion technology, unlike the Lummus/UOP Smart SM technology, does not directly combust hydrogen from the reaction mixture; hence it does not obtain the benefit of a favorable shift in equilibrium.

Other Processes

Propylene oxide/styrene process

Aside from EB dehydrogenation, the only other commercial-scale production of styrene is through a propylene oxide/styrene process that produces roughly 15% of worldwide styrene. This technology was developed as an alternative to the chlorohydrin method for producing propylene oxide.

Styrene from Butadiene

Because the conventional EB dehydrogenation technologies are relatively mature, there is little room for significant additional reduction in production costs. This situation has motivated a lot of research toward using alternative, lower cost feedstocks for styrene production. One area that has been examined involves a two-step process to convert butadiene to styrene.

The first step of the process involves the cyclodimerization of butadiene to 4-vinylcyclohexene. The reaction is exothermic and can be catalyzed by either a copper-containing zeolite catalyst or an iron dinitrosyl chloride catalyst complex. Although both vapor-phase and liquid-phase processes have been studied, it appears that liquid-phase reactions are preferred because they achieve higher butadiene conversion levels. The second step is oxidative dehydrogenation of the 4-vinylcyclohexene to produce styrene. Dow has led the research effort in this area and has identified catalyst formulations that provide more than 90% conversion of 4-vinylcyclohexene with approximately 92% selectivity to styrene.[11]

Storage

Preventing polymerization is the key to successful styrene storage. Special handling and storage procedures are required to maintain the styrene product quality and to avoid a potentially dangerous situation involving uncontrolled polymerization.

During storage, styrene polymerization is prevented by maintaining low temperature and using an appropriate polymerization inhibitor. The industry standard styrene storage inhibitor is TBC and is typically used at concentrations between 10 ppm and 15 ppm. To be effective, TBC requires dissolved oxygen to be present in concentrations roughly equal to the TBC concentration.

In addition to adding TBC inhibitor, maintaining the styrene at the lowest practical temperature is critical to preserving product quality. Styrene storage facilities are generally maintained at temperatures below about 20°C, which allows for storage times of around 10 weeks. Even a 5°C increase in the storage temperature to 25°C can reduce the storage time to less than 4 weeks.[12] Tertiary-butyl catechol is added occasionally during storage to maintain the concentration in the desired range.

ECONOMICS

The cost of styrene production can be broken down into three main components: raw materials, utilities, and the fixed cost associated with the plant. The utilities cost includes fuel, electricity, steam, cooling water, catalyst, and chemical costs required to

Table 2 Styrene economics for conventional EB dehydrogenation process

	UNIT	Quantity UNIT/MT	Price $/UNIT	Cost $/MT
Produce				
Styrene	MT	1.0000	751	751.0
Raw materials				
Ethylene	MT	0.2912	629	183.2
Benzene	MT	0.7898	453	357.8
By-product credits				
Toluene	MT	0.0401	378	(15.1)
Light ends	MT	0.0401	289	(11.6)
Net feedstock costs				514.2
Utilities				95.0
Fixed cost				35
Total cost of production				644.2

Basis: North America, 2003

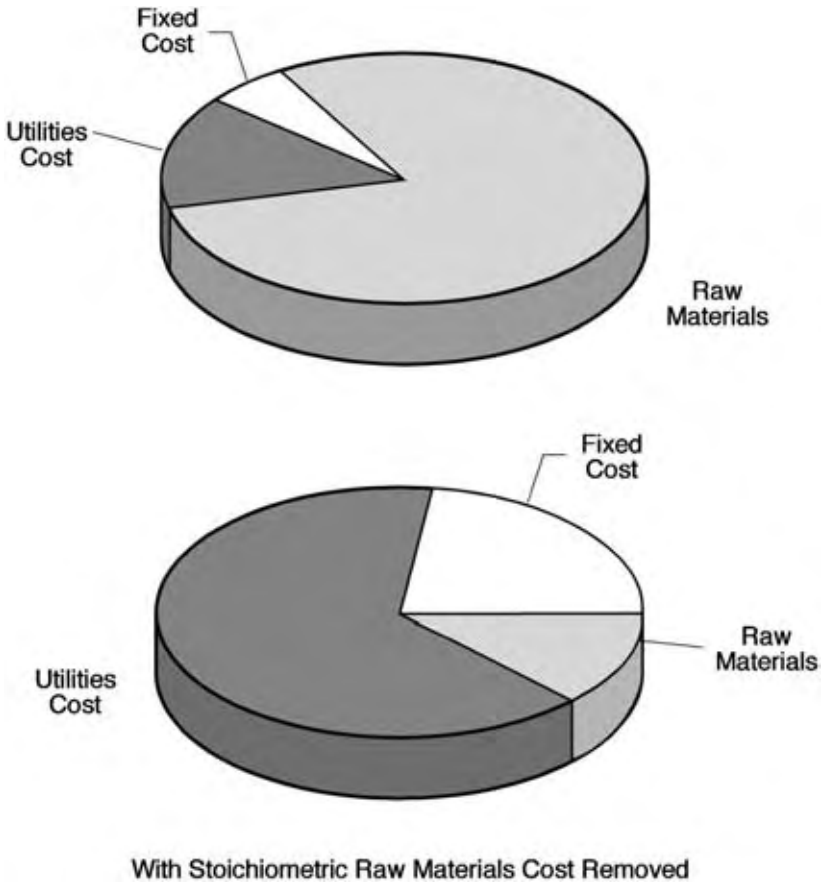


Fig. 7 Distribution of styrene production cost components.

operate the plant. The major cost components for styrene production using conventional adiabatic dehydrogenation process are listed in Table 2. The major cost of production is for the ethylene and benzene raw materials, which account for approximately 80% of the total cost of production. The benzene cost is the largest cost component; hence, the economics of styrene production are highly dependent on benzene price.

The raw materials cost has two components—one dictated by the stoichiometry and the other caused by

Table 3 Styrene economics for propylene oxide-styrene process

	UNIT	Quantity UNIT/MT	Price $/UNIT	Cost $/MT
Product				
Styrene	MT	1.0000	751	751.0
Raw materials				
Ethylene	MT	0.3135	629	197.2
Benzene	MT	0.8194	453	371.2
Propylene	MT	0.3541	465	164.6
Oxygen	MT	0.2529	43	10.9
By-product credits				
Light ends	MT	0.1000	289	(28.9)
Propylene oxide	MT	0.4500	1227	(552.1)
Tars	MT	0.0400	257	(10.3)
Net feedstock costs				152.6
Utilities				65.0
Fixed cost				95
Total cost of production				312.6

Basis: North America, 2003

yield losses occurring as a result of the process technology. If the unalterable stoichiometric raw material consumption is removed from the cost of production, the resultant distribution of cost components appears very different, as shown in Fig. 7. From this perspective, the raw materials' cost is only about 15% of the incremental cost of production and the utilities and fixed costs become dominant. Recent catalyst and process design improvements have reduced the variable costs of styrene production, while ever-increasing complexity and more stringent regulations have greatly increased the fixed costs. Other recent trends, such as larger plant capacities and globalization of the styrene market, have also resulted in higher fixed costs.[13]

The result of the shift of focus from variable to fixed costs is that plants are being designed for larger capacities. For example in 2003, typical new styrene plants in the Asia Pacific Region produced an average of 350 KMTA styrene per year, nearly double the capacity of typical plants started up just 5 years earlier. The drive to reduce fixed costs has led to numerous revamps of existing plants to substantially increase capacity. In many cases, capacity expansions on the order of 50% are being implemented.

The propylene oxide/styrene process, the only other commercial process for production of styrene, is a growing influence on the overall styrene market economics. When viewed from the perspective that styrene is the primary product and propylene oxide is a by-product, the economics of this process appear encouraging (Table 3). Depending on the credit value assigned to the propylene oxide coproduct, the total cost of styrene production can be approximately 50% of conventional EB dehydrogenation technology. Approximately 33% of the styrene capacity added between 1998 and 2003 was produced using propylene oxide/styrene technology. More recently, the trend appears to be reversing and propylene oxide/styrene processes are accounting for less of newly installed capacity. Although propylene oxide/styrene plants are built to produce propylene oxide, there is a profound impact on the styrene market supply/demand balance.

CONCLUSIONS

Since the first commercial-scale production in the 1930s, styrene, mainly through its derivatives, has become an integral part of life. Most people come in contact with numerous styrene-based products throughout the course of a normal day. Demand for styrene is expected to continue growing at a rate comparable to the gross domestic product growth rate.

The chemical processing technologies that have been developed are sophisticated, producing styrene to meet the demand at low cost. Research and development efforts are aimed at further improvements in existing technologies and identification of new technologies for styrene production opportunities.

REFERENCES

1. Lidback, A. Styrene—This is Not a Drill, 2004 World Petrochemical Conference, Chemical Marketing Associates, Inc.: Houston: TX, March 23–25, 2004.
2. Carra, S.; Forni, L. Kinetics of catalytic dehydrogenation of ethylbenzene to styrene. Ind Eng Chem Process Des. Dev. **1965**, *4* (3), 281–285.
3. Matsui, J.; Sodesawa, T.; Nozaki, F. Influence of carbon dioxide addition upon decay of activity of a potassium-promoted iron oxide catalyst for dehydrogenation of ethylbenzene. Appl. Catal. **1991**, *67*, 179–188.
4. Hirano, T. Roles of potassium in potassium-promoted iron oxide catalyst for dehydrogenation of ethylbenzene. Appl. Catal. **1986**, *26*, 65–79.
5. Kollar, J. Epoxidation Process. US Patent 3,351,635, November 7, 1967.
6. Chem Systems, Propylene Oxide 97/98–7, Tarrytown, New York, 1998.
7. Mullen, P. Enhancements in EB/SM technology. In *AIChE 2003 Spring Meeting Proceedings*; AIChE Spring National Meeting. Houston: TX, April 22–26, 2001.
8. Sardina, H. Dehydrogenation Process for Production of Styrene from Ethylbenzene Comprising Low Temperature Heat Recovery and Modification of the Ethylbenzene–Styrene Feed therewith. US Patent 4,628,136, December 9, 1986.
9. Imai, T. Dehydrogenation of Dehydrogenatable Hydrocarbons. US Patent 4,435,607, March 6, 1984.
10. Welch, V. Advanced styrene dehydrogenation with flameless distributed combustion. In *AIChE 2003 Spring Meeting Proceedings*; AIChE Spring National Meeting. Houston: TX, April 22–26, 2001.
11. Chem Systems PERP Report, Styrene from Butadiene, 93S3, Tarrytown, New York, 1995.
12. Technical Bulletin on Safe Handling & Storage of Styrene Monomer http://www.sterlingchemicals.com/SCI/WEBSITE/scihome.nsf/(WebContentByDocID)/0AF6F53F1881D0AA8626CD90082E36A?OpenDocument.
13. Ram, S. EB-SM splitter energy recovery options. In *Styrene Conference General Session*, Styrene Conference, Prague, Czech Republic, June 22–25, 2003; ABB Lummus Global, Sud-Chemie AG, and UOP LLC, 2003.

Styrene–Butadiene Rubber

Jing Peng
Department of Applied Chemistry, College of Chemistry, Peking University, Beijing, People's Republic of China

INTRODUCTION

Styrene–butadiene rubber (SBR) is a random polymer made from butadiene and styrene monomers. It possesses good mechanical property, processing behavior, and can be used like natural rubber. Moreover, some properties such as wear and heat resistance, aging, and curing property are even better than in natural rubber. Styrene–butadiene rubber was the first major synthetic rubber to be produced commercially. Now it has become the most common rubber with the largest production and consumption in the synthetic rubber industry. It can be widely used in tire, adhesive tape, cables, medical instruments, and all kinds of rubberware.

Styrene–butadiene rubber could be produced by using emulsion and solution process, thus it can be divided into emulsion-polymerized styrene–butadiene rubber (E-SBR) and solution-polymerized styrene–butadiene rubber (S-SBR). In this entry, we will describe their development and introduce their synthesis process, relationship between structure and property, processing property, blends, and applications.

EMULSION-POLYMERIZED STYRENE–BUTADIENE RUBBER

History

In the 1930s, I. G. Farbenindustrie in Germany prepared the first E-SBR known as Buna S. The American government in 1940 established the Rubber Reserve Company to start storage of natural rubber and a synthetic rubber program. These programs were expanded when the United States entered World War II. These E-SBR grades were called GR-S (government rubber styrene). Initially, the synthesis of E-SBR was focused on a hot polymerized (41°C) E-SBR. Production of a 23.5% styrene and 76.5% butadiene copolymer began in 1942. Cold polymerized E-SBR (5°C), which has significantly better physical properties than hot polymerized SBR, was developed in 1947. Thereafter, the oil extended E-SBR was produced in 1951.

Synthesis

The emulsion polymerization process has several advantages. It is normally carried out under mild reaction conditions that are tolerant to water in the absence of oxygen. The process is relatively resistant to impurities and amenable to using a range of functionalized and nonfunctionalized monomers. Additional benefits include the fact that emulsion polymerization gives high solid contents with low reaction viscosity and is a cost-effective process. The physical state of the emulsion (colloidal) system makes it easy to control the process. Thermal and viscosity problems are much less significant than in bulk polymerization.

Table 1 shows the raw materials required in the polymerization of E-SBR, which include monomers (styrene and butadiene), water, emulsifier, initiator system, modifier, terminal agent, and a stabilizer system. The original polymerization reactions were carried out in batch reactors in which all the ingredients were loaded to the reactor and the reaction was terminated after it had reached the desired conversion. Current commercial productions are run continuously by feeding reactants and polymerizing through a chain of reactors before terminating at the desired monomer conversion. The monomers are continuously metered into the reactor chains and emulsified with the emulsifiers and catalyst agents.

In cold polymerization, the most widely used initiator system is the redox reaction between chelated iron and organic peroxide using sodium formaldehyde sulfoxide (SFS) as a reducing agent [see Eqs. (1) and 2]. In hot polymerization, potassium persulfate is used as an initiator.

$$\mathrm{Fe(II)EDTA} + \mathrm{ROOH} \rightarrow \mathrm{Fe(III)EDTA} + \mathrm{RO^{\bullet}} + \mathrm{OH^{\bullet}} \quad (1)$$

$$\mathrm{Fe(III)EDTA} + \mathrm{SFS} \rightarrow \mathrm{Fe(II)EDTA} \quad (2)$$

Mercaptan is added to provide free radicals and to control the molecular weight distribution by terminating existing growing chains while initiating a new chain. The thiol group acts as a chain transfer agent to prevent the molecular weight from attaining the

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120019881

Table 1 Typical recipe for SBR emulsion polymerization

	Parts by weight	
Component	**Cold**	**Hot**
Styrene	25	25
Butadiene	75	75
Water	180	180
Emulsifier (FA, RA, MA)	5	5
Dodecyl mercaptan	0.2	0.8
Cumene hydroperoxide	0.17	—
$FeSO_4$	0.017	—
EDTA	0.06	—
$Na_4P_2O_7{\cdot}10H_2O$		1.5
Potassium persulfate		0.3
SFS	0.1	—
Stabilizer	Varying amount	Varying amount

FA, fatty acid; RA, rosin acid; MA, rosin acid/fatty acid.
(From Ref.[1].)

excessively high values possible in emulsion systems. The sulfur–hydrogen bond in the thiol group is extremely susceptible to attack by the growing polymer radical and thus loses a hydrogen atom by reacting with polymer radicals [Eq. (3)]. The $RS^{\bullet}$ formed will continue to initiate the growing of a new chain as shown in Eq. (4) below. The thiol prevents gel formation and improves the processability of the rubber.

$$P^{\bullet} + RSH \rightarrow P-H + RS^{\bullet} \quad (3)$$

$$RS^{\bullet} + M \rightarrow RS-M^{\bullet} \quad (4)$$

During polymerization, parameters such as temperature, flow rate, and agitation speed must be controlled carefully to get the right conversion. Polymerization is normally allowed to proceed to about 60% conversion in cold polymerization and 70% in hot polymerization before it is stopped with a terminal agent that reacts rapidly with the free radicals. Common terminal agents include sodium dimethyldithiocarbamate and diethyl hydroxylamine.

Once the latex is properly terminated, the unreacted monomers are removed from the latex. Butadiene is stripped by degassing the latex by means of flash distillation and reduction of system pressure. Styrene is removed by steam stripping the latex in a column. The latex is then stabilized with the appropriate antioxidant and transferred to blend tanks. In the case of oil-extended polymers or carbon black master batches, these materials are added as dispersions to the stripped latex. The latex is then transferred to finishing lines to be coagulated with sulfuric acid, sulfuric acid/sodium chloride, glue/sulfuric acid, aluminum sulfate, or amine coagulation aid. The type of coagulation system is selected depending on the end use of the product. Sulfuric acid/sodium chloride is used for general purposes. Glue/sulfuric acid is used for electrical grade and low water sensitivity SBR. Sulfuric acid is used for coagulations where low-ash polymer is required. Amine coagulating aids are used to improve coagulation efficiency and reduce production plant pollution. The coagulated crumb is then washed, dewatered, dried, baled, and packaged.

On the other hand, due to the difference in producing E-SBR, E-SBR grades have different properties. Hot emulsion polymerization is the original SBR process. The major characteristic of this process is that these grades have excellent processing behavior in terms of low mill shrinkage, good dimensional stability, and good extrusion characteristics. However, high content of microgels is also produced in hot polymerization; therefore, there is a trend toward using cold emulsion grades in many applications. Nevertheless, hot rubbers are still used in applications such as adhesives and flow modifiers for other elastomers where good flow properties are required.

Properties of E-SBR

E-SBR is commercially available in Mooney viscosities ranging from 30 to about 120 [ML (1 + 4) 125°C]. Lower Mooney viscosity E-SBR grades adhere more easily to the mill, incorporate fillers and oil more readily, show less heat generation during mixing, and are calendered more easily, shrink less, give higher extrusion rates, and have superior extrudate appearance than the higher Mooney viscosity grades. On the other hand, the high Mooney viscosity SBR grades have better green strength, less porosity in the vulcanizate, and accept higher filler and oil loadings.

As the molecular weight of the SBR increases, the vulcanizate resilience and the mechanical properties, particularly tensile strength and compression set, improve. The processability of SBR can be improved as its molecular weight distribution broadens. However, the formation of high molecular weight fractions with the increase in the average molecular weight can prevent improvements in the processability. This is due to the fact that the tendency for gel formation also increases at higher molecular weights.

In addition to the polymer viscosity, polymerization temperature also plays an important role in shaping the processability. Emulsion-polymerized SBR grades produced at low polymerization temperatures have less chain branching than those produced at higher temperatures. At an equivalent viscosity, cold polymerized E-SBR is normally easier to process than hot polymerized E-SBR, and this applies particularly

to a better banding on mills, less shrinkage after calendering, and a superior surface of green tire compounds. Hot rubbers give better green strength because they have more chain branching.

The styrene content of most emulsion SBR varies from 0% to 50%. The percentage of styrene in most commercially available E-SBR grades is 23.5%. In vulcanizates of SBR, as styrene content increases, dynamic properties and abrasion resistance decrease while traction and hardness increase. Polymerization temperature also affects the microstructure of E-SBR. In the cold polymerized E-SBR grades, the butadiene component has, on average, about 9% *cis*-1.4, 54.5% *trans*-1.4, and 13% of vinyl-1.2 structure. At the 23.5% styrene level, the glass transition temperature, T_g, of SBR is about –50°C. As the styrene content in the SBR increases, the glass transition temperature also increases. Rubbers with very low T_g values posses a high resilience and good abrasion resistance, but have poor wet traction. By contrast, those rubbers with high T_g (for instance, SBR 1721) exhibit a low resilience and poor abrasion resistance with an excellent wet traction.

The emulsifier remains in the rubber after coagulation and can also have an influence on the processability. Rosin acid emulsifiers impart better knitting, tack, and adhesion to the SBR polymer. Generally, polymers emulsified with rosin acid have better extrusion rates, slower cure rates, poorer heat resistance, and can cause mold fouling and polymer discoloration. Fatty acid emulsified SBR polymers generally have less tack, faster curing, and high tensile properties. A compromise of the above properties is obtained by using a mixed rosin acid/fatty acid emulsifier system.

Chemical Activity of E-SBR

It has been proved that incorporation of carboxylic acid groups in the polymeric chain has a significant effect on colloidal properties of latex, processability, and end-use property. Carboxylated styrene–butadiene latexes (XSBR) are prepared via batch emulsion copolymerization with different amounts of acrylic acid in the absence of emulsifier. They are among the most important polymeric colloids, and can be used as binder in paper coatings, carpet backing, paints, and nonwoven. There are several studies on the preparation and properties of XSBR latexes.

To improve the aging property of E-SBR, Parker and Roberts used the diimide reduction method to prepare hydrogenated E-SBR. In the system containing hydrazine hydrate, oxidant, and a metal-ion catalyst, hydrogenated E-SBR with 97% of hydrogenation could be obtained.[2] The hydrogenated E-SBR not only exhibited excellent ozone, oxidation, and UV resistance as expected, but also showed better mechanical properties in some circumstances than unhydrogenated E-SBR.

Moreover, E-SBR can be grafted with some polar monomers such as acrylic acid or organic chlorium to give modified E-SBR with good heat resistance and tensile strength.

Cure Properties and Processing

Styrene–butadiene rubber can be cured with a variety of cure systems including sulfur (accelerators and sulfur), peroxides, and phenolic resins. In addition, some papers have reported that the SBR could be vulcanized using gamma irradiation in the presence of polyfunctional monomers. The efficiencies of seven functional monomers toward radiation vulcanization of SBR are in the following order: tertamethylol methane tetraacrylate (ATMMT) = toluene diisocyanate (UA306T) > trimethylol propane trimethacrylate (TMPT) > diethylene glycol dimethacrylate (2G) > dipentaerthritol hexaacrylate (DPE6A) > hexamethylene diisocyanate (UA101H) > triallyl cyanurate (TAC).[3]

Processing of SBR compounds can be performed in a mill, internal mixers, or mixing extruders. Styrene–butadiene rubber compounds are cured in a variety of ways by compression, injection molding, hot air or steam autoclaves, hot air ovens, microwave ovens, and combinations of these techniques.

Types of E-SBR

There are a large variety of E-SBR types based on the styrene content, polymerization temperature, antioxidants, oil and carbon black content. Each of these basic classifications includes a variety of SBR polymer variations with respect to Mooney viscosities, coagulation types, emulsifier type, oil levels, and carbon black types and levels. Table 2 shows the basic series of E-SBR.

Table 2 The types of E-SBR

Series	Comments
1000	Hot polymerized polymers
1500	Nonextended cold polymerized polymers
1600	Non-oil-extended cold carbon black masterbatches
1700	Cold oil-extended polymers
1800	Cold oil-extended carbon black masterbatches
1900	Miscellaneous high styrene resin masterbatches

(From Ref.[1].)

Applications of E-SBR

Emulsion-polymerized SBR is predominantly used for the production of car and light truck tires and truck tire retread compounds. A complete list of the uses of SBR includes houseware mats, drain board trays, shoe soles and heels, chewing gum, food container sealants, tires, conveyor belts, sponge articles, adhesives and caulks, automobile mats, brake and clutch pads, hose, V-belts, flooring, military tank pads, hard rubber battery box cases, extruded gaskets, rubber toys, molded rubber goods, shoe soling, cable insulation and jacketing, pharmaceutical, surgical, and sanitary products, food packaging, etc. The typical applications of E-SBR polymers are presented in Table 3.

SOLUTION-POLYMERIZED STYRENE–BUTADIENE RUBBER

History

During the 1960s, Phillips produced the first solution-polymerized random SBR grades, which are named as Solprene X-40, commercially. In 1969, Firestone produced the commercial S-SBR grades named as Duradone. The original aim to produce SBR with lower styrene content than products made by emulsion polymerization is to counteract the increase in styrene price. Standard S-SBR grades now have comparable styrene content to emulsion types. These grades have superior mechanical properties to E-SBR.

Table 3 Typical applications of E-SBR polymers

	Hot E-SBR grades	Cold E-SBR		High styrene master batch	Black master batch
		Unextended	Oil extended		
Adhesives					
Type and label	•	•	•		
Caulking	•				
Laminating	•	•			
Mastic	•				
Panel	•				
Pressure sensitive	•	•			
Sealant	•				
Sprayable (cross-linked)	•				
Wall tile	•				
Automotive					
Tire treads		•			•
Apex/rim/flange				•	
Bead	•	•			
Carcass		•	•		
Retread	•	•	•		
Racing tires			•		
Mats		•	•	•	•
Miscellaneous					
Mechanical goods	•	•	•	•	•
Rolls		•	•		•
Gaskets		•	•	•	•
Belts/hoses		•	•		•
Hard rubber goods		•		•	
Cove base	•			•	
Floor tiles	•	•	•	•	
Footwear		•	•	•	•
Sponge		•	•	•	
Wire and cable (low ash)		•			•

(From Ref.[1].)

In 1983, the high *trans*-SBR grades were synthesized in America.

Solution-polymerized SBR grades have superior mechanical properties, particularly tensile strength, low rolling resistance, and handling, when used in tire applications. The ratio of butadiene configurations varies. Generally speaking, S-SBR grades have a lower *trans* and vinyl and a higher *cis* butadiene content than E-SBR grades. In initially making S-SBR grades, producers attempted to replicate the stereochemistry of E-SBR grades. However, solution polymerization differs from emulsion polymerization because of its flexibility and enables SBR grades with varying styrene–butadiene ratios and *cis*, *trans*, and vinyl contents to be produced by changing the catalyst and monomer ratios and reaction conditions. This enables S-SBR producers to synthesize grades specifically tailored for individual applications.

Synthesis

Solution-polymerized SBR is made by termination-free, anionic/live polymerization initiated by alkyl lithium compounds. Other lithium compounds are suitable (such as aryl, alkaryl, aralkyl, tolyl, xylyl lithium, and α/β-naphtyl lithium as well as their blends), but alkyl lithium compounds are the most commonly used in industry. The absence of a spontaneous termination step enables the synthesis of polymers possessing a very narrow molecular weight distribution and less branching. Carbon dioxide, water, oxygen, ethanol, mercaptans, and primary/secondary amines interfere with the activity of alkyl lithium catalysts, so the polymerization must be carried out in clean, near-anhydrous conditions. Stirred bed or agitated stainless steel reactors are widely used commercially.

Polymerization is carried out in a solution of inert aliphatic or aromatic solvent. The polymerization rate of butadiene in the presence of lithium-based catalysts is lower than styrene. However, when butadiene and styrene monomers are mixed, the rate of polymerization is reversed, resulting in the production of block copolymer with a high proportion of butadiene blocks. Block formation must be suppressed because the property requirements of traditional SBR markets cannot be met by block copolymers. Random copolymerization is encouraged by incorporating into the solution ''randomizing'' agents such as dialkyl and heterocyclic ethers, which act as a Lewis base on the catalyst, or by controlled monomer charging (i.e., some of the styrene is added later in the polymerization cycle). The resulting copolymer is precipitated, separated, dried, and baled.

By adjusting the reactivity of the initiator/modifier system toward metallation of the polymer backbone, Kerns and Henning have developed synthetic strategies that effectively control the relative level of branching and polydispersity in S-SBR. This development has allowed for the optimization of both the microstructure and the macrostructure of solution elastomers to meet the demands of a given application. They have taken the approach to control the polymer macrostructure by adjusting the initiator/modifier system so as to mediate the propensity for backbone metallation to occur. Several synthetic strategies were employed to impose control over the levels of branching in solution SBR grades. They found that the macrostructure is the independent variable and subtle differences can have profound effects on the rheology and, thus, mixing, extruding, and even physical properties of the resultant compound.[4]

Compared with the E-SBR, S-SBR has different structures and properties. Some typical characteristics of S-SBR grades are shown in Table 4.

Properties of S-SBR

The optimization of property has been achieved by conventional solution SBR technology. By modifying the way in which monomers are added, the polymerization conditions, the use of cocatalysts and randomizing agents, the proportion of *cis* and vinyl isomers, the chain structure of the resulting ''tailored'' polymer can be varied.

Generally, the Mooney viscosity of S-SBR is higher than E-SBR. Thus, it accepts higher filler and oil loadings. The effect of composition on the property of S-SBR is similar to that of E-SBR. As the styrene content increases, the rolling resistance, traction, and hardness increase, while the wear resistance decreases. With the increase of vinyl content, the wet traction increases.

Mechanical Property

Styrene–butadiene rubber is generally compounded with a vulcanization system, reinforcing filler (usually carbon black), processing/extending oil, and an antioxidant/stabilizer package prior to molding/fabrication. Standard S-SBR grades have comparable styrene content to emulsion types. These grades have superior mechanical properties to E-SBR. The typical mechanical properties of some commercial S-SBR grades are listed in Table 5.

Cerveny investigated the development of the dynamic glass transition in styrene–butadiene copolymers by dielectric spectroscopy in the frequency range from 10^{-2} to 10^{6} Hz. Two processes were detected and attributed to the alpha- and beta-relaxations. The alpha relaxation time has a non-Arrhenius temperature behavior that is highly dependent on styrene content

Table 4 The typical characteristics of some S-SBR grades

	Types							
	S-SBR							
	Low vinyl-1,2 structure		Medium vinyl-1,2 structure			High vinyl-1,2 structure		
Entry	Tufdene 2000R	NS 118	Solprene 1204	SL 552	NS 114	SL574	NS110	E-SBR 1500
Initiator	Alkyl lithium	Alkyl lithium	Alkyl lithium	Alkyl lithium	Alkyl lithium	Alkyl lithium	Alkyl lithium	Redox reaction
Styrene content (%)	25	17.5	25	24	23	15	12	23.5
Structure of butadiene								
cis-1,4 (%)	35		24	20		16		12
trans-1,4 (%)	52		40.5	40		27		68.5
vinyl-1,2 (%)	13	11.5	35.5	30	38	57	72.5	19.5
Molecular weight distribution	Narrow	Two peaks	Two peaks	Two peaks	Two peaks	Two peaks	Two peaks	Wide
Glass transition temperature (°C)	~−70	−75	~−50	−64	−53	−55	−28	~−60

while the beta relaxation time shows an Arrhenius behavior with an activation energy that is independent of styrene-content. Furthermore, the shape of the alpha-relaxation is strongly influenced by the styrene content while the shape of beta-relaxation is independent of styrene content. Someone interprets these results as follows. The observed beta-relaxation is primarily due to local motions of butadiene monomers and is therefore not affected by the presence of styrene. The alpha-relaxation, on the other hand, is highly sensitive to the styrene content owing to its cooperative character.[6]

Chemical Activity of S-SBR

Using chemical modification such as cyclization, chlorination, hydrogenation, and epoxidation, the different S-SBR derivatives could be obtained.

Controlled cyclization of SBR was achieved with the aid of cationic catalyst system based on diethylaluminum chloride ($AlEt_2Cl$) and benzyl chloride ($C_6H_5CH_2Cl$) and by working in xylene solution at high temperature ($T > 100°C$). Elastomers with low intrinsic viscosity, ready solubility, and free gel were produced by Wang et al. The cyclized products have been expected to be photoresists with high photosensitivity.[7]

Dichlorocarbene-modified SBR prepared by the alkaline hydrolysis of chloroform using cetyltrimethylammonium bromide as a phase-transfer agent resulted in a product that showed good mechanical properties, excellent flame and solvent resistance, and good thermal stability. The molecular weight of the polymers, determined by gel permeation chromatography, showed that chemical modification was accompanied by an increase in molecular weight. Proton nuclear magnetic

Table 5 Typical mechanical properties of S-SBR grades

Product	Tensile strength (MPa)	Stress at 300% elongation (MPa)	Elongation at break (%)
SE SLR-4601	19	15	350
SE SLR-4400	22	13	430
SE SLR-4610	19	9	550
SE SLR-4610H	21	9.4	570
SE SLR-4630	20	10.5	540
SE SLR-6410	23	9	630

(From Ref.[5].)

resonance and Fourier transform infrared (FTIR) studies revealed the attachment of chlorine through cyclopropyl rings to the double bond of butadiene.[8]

Because of the existence of unsaturation bonding in the SBR backbone, the material is susceptible to degradation under oxygen and ozone atmosphere. This drawback can be overcome by hydrogenating the copolymer to give hydrogenated styrene–butadiene copolymer. The hydrogenation of a copolymer of styrene and butadiene (SBR) catalyzed by $Ru(CH{=}CHPh)Cl(CO)(PCy_3)_2$ was experimentally investigated within the temperature range of 120–160°C at P_{H_2} of 300–1200 psi, and a catalyst concentration of 1.0–7.8 $\times$ 10^{-5} M. Special attention was paid to minimizing the catalyst metal residue and cross-linking in the product. The results indicated that high-quality hydrogenated SBR could be achieved without cross-linking and the metal residue was less than 7 ppm without post-treatment.[9] In addition, the hydrogenated SBR under certain conditions is more stable than its unsaturated precursors. They can be used as model compounds for predicting the unperturbed dimensions of polymer chains.

Cure Properties and Processing

Compared with E-SBR, the curing rate of S-SBR grades is faster by about 10–20%. Generally, the processing of S-SBR compounds is the same as for E-SBR. The difference is that the molecular chain of S-SBR cannot be broken in the internal mixer with high power; therefore, all the components can be added into the mixer simultaneously without precalendering.

Usually, the mechanical strength of raw S-SBR is lower than that of E-SBR; it is necessary to reduce the distance between two rolls by 10–20% to avoid the peeling off of the roll. On the other hand, S-SBR bonds easily on the cold roll. It will be better to ensure that the temperature difference between two rolls is about 5–10°C during the processing of S-SBR.

Applications of S-SBR

S-SBR grades have excellent balance between wet traction and rolling resistance; therefore, they can be used for low fuel consumption tire treads in all-season tires and high-quality rubber goods. In applications of S-SBR, carbon black and silica will often be added to enhance the property.

Blending

Rubber goods usually require a combination of properties that cannot be provided by one elastomer only and two or more polymer components have to be mixed to meet specific requirements such as lowering the compound cost, for ease of fabrication and to improve the performance of the industrial rubber. Natural rubber (NR) and SBR have been blended for a long time for these reasons. The tensile, hardness, and wear properties of the NR/SBR blends with the increment of NR percentages were greatly enhanced.[10]

For the system of natural rubber (NR) and styrene butadiene rubber (SBR) in a 1:1 ratio, the best properties were found when the NR/SBR mixture was prepared in such a way as to favor the vulcanization of the SBR phase while preserving the NR phase from excessive vulcanization.[11]

The morphology and mechanical and viscoelastic properties of a series of blends of NR and SBR latex blends were studied in the uncross-linked and cross-linked states. The morphology of the blends indicated a two-phase structure in which SBR is dispersed as domains in the continuous NR matrix when its content is less than 50%. A cocontinuous morphology was obtained at a 50/50 NR/SBR ratio and phase inversion was seen beyond 50% SBR when NR formed the dispersed phase. As the NR content and time of prevulcanization increased, the mechanical properties such as the tensile strength, modulus, elongation at break, and hardness increased. This was due to the increased degree of cross-linking that leads to the strengthening of the three-dimensional network. In most cases the tear strength values increased with the prevulcanization time. The effects of the blend ratio and prevulcanization on the dynamic mechanical properties of the blends were investigated at different temperatures and frequencies. All the blends showed two distinct glass transition temperatures, indicating that the system is immiscible. It was also found that the glass transition temperatures of vulcanized blends are higher than those of unvulcanized blends.[12]

Styrene butadiene rubber/epoxidized natural rubber (ENR) blends were prepared with an internal mixer, Haake Rheomix. The scorch time, t_2, and curing time, t_{90}, were found to decrease with increasing ENR composition in the blends. The mechanical properties such as tensile strength, tear strength, and tensile modulus, M_{300} (modulus at 300% elongation), increase with increasing ENR composition in the blends. However, elongation at break shows an opposite trend. At similar immersion times, SBR/ENR blends containing a higher ENR content exhibit better oil resistance.[13]

The mixing of incompatible polymers such as polyethylene terephthalate (PET) and SBR produces a blend with poor mechanical and impact properties, because polymeric phases interact weakly with each other and segregate. The use of SBR grafted with maleic anhydride (MAH) could increase the compatibility of the SBR–PET system by generating higher interactions and chemical bonds between the ingredients of the

blend. The induced compatibility is reflected in the 2.5-fold increase in the impact resistance of the blend as compared to that of pure PET. The grafting reaction to produce SBR-*g*-MAH is carried out by reactive extrusion using a reaction initiator benzoyl peroxide (BPO), and the extent of the reaction depends on the concentration of MAH and BPO. Results indicated the close relationship between processing conditions and microstructural parameters, such as particle diameter and interparticle distances of the dispersed rubber phase, which is necessary to achieve the best impact resistance.[14]

The morphology of ternary polystyrene/SBR/polyethylene (PS/SBR/PE) blends has been investigated in the limits of a constant content of the major component (PS: 75 wt%) while changing the weight ratio of the two minor constitutive polymers. A core-shell structure for the dispersed phase has been predicted from the spreading coefficients and observed by transmission electron microscopy. Actually, with the increase in the relative content of PE with respect to SBR, the structure of the dispersed phase changes from a multicore structure to a PE/SBR core-shell morphology. The size of the PE subphase in the mixed dispersed phase increases sharply at a PE content that corresponds to phase inversion in the parent SBR/PE binary blends. The ultimate mechanical properties of these blends are sensitive to the strength of the SBR interphase between PS and PE. Some synergism has been observed in the PE/SBR composition dependence of the tensile strengths at yield and break.[15]

Because SBR lacks the self-reinforcing qualities of natural rubber due to stress-induced crystallization, gum vulcanizates of SBR have lower tensile properties. The tensile property of E-SBR vulcanizates depends significantly on the type and amount of filler in the compound. Cured gum stocks have only 2.8–4.2 MPa in tensile strength, while fine particle carbon black loadings can produce tensile strength of 27.6 MPa. Though the compression set of some of the common E-SBR compounds is high, by proper compounding and blending, it is possible to obtain E-SBR vulcanizates with a low compression set. Some studies also indicated that filler affects the physical and mechanical properties of the blends. In such cases, the additives normally employed in rubber formulations are unevenly distributed, depending on the affinity of each compound to each polymeric phase. Thus, the dispersion of each one of these ingredients in the different rubbers will influence the rate and degree of vulcanization and, in consequence, the performance of the final composite.

Recently, nanocomposites were prepared with different grades of nitrile rubber with acrylonitrile contents of 19%, 34%, and 50%, with SBR (23% styrene content), and with polybutadiene rubber with Na-montmorillonite clay. The clay was modified with stearyl amine and was characterized by x-ray diffraction (XRD), FTIR spectroscopy, and transmission electron microscopy (TEM). At the degree of filler loading up to a certain level, mechanical properties of nanocomposites could be improved.[16]

A new kind of rubber powder with ''salami'' structure (RPS) was prepared by spray drying the mixture of SBR latex and nano-$CaCO_3$ slurry. It was found that RPS is an effective toughener with synergistic toughening effect on polypropylene (PP). The Izod impact strength of PP/RPS blend is not only higher than that of PP/rubber powder or PP/nano-$CaCO_3$ blends, but also higher than that of a PP/rubber powder/$CaCO_3$ blend. Transmission electron microscopy images showed that the microstructure of the PP/RPS blend is an ''island-sea'' structure with ''salami'' structure in RPS, in which nano-$CaCO_3$ particles are embedded in SBR particles. Perhaps this morphology resulted in the improvement of mechanical properties of blends.[17]

Technical Development of SBR

In recent years, SBR developments have been predominantly initiated by the tire industry. This is not surprising when several of the SBR producers have put their assets up for sale in recent times such as Polimeri in Europe, DSM in the United States, and others. This is partly due to the commoditization of SBR and also to butadiene price rises, which are difficult to pass for the tire producers. This has created severe pressure on margins for SBR producers. Therefore, the majority of recent developments have been undertaken by the tire companies themselves rather than the SBR producers who were more active in the past. Property optimization of SBR has been achieved, to some extent, by conventional S-SBR technology.[18]

The automotive industry is under continued pressure to improve the environmental performance and useful life of automotive components. S-SBR producers are responding to develop SBR grades with optimum combination of rolling resistance, wear resistance, blow-out resistance, chipping/chunking resistance, road traction under a variety of weather conditions, handling, noise transmission, and other performance properties for different tire applications.

There are many ways in which some improvements are being achieved:[18]

- Reformulating compounds and using high-performance additives in conjunction with ''tailored'' S-SBR grades.
- Developing novel additives/modifiers that can be added to the SBR at the compound stage.
- Further modification of polymerization conditions to enable both block and random copolymerization.

- Introduction of postpolymerization steps to obtain the copolymer with end group structure, thus significantly altering the properties of the resulting SBR grades.
- Introduction of postpolymerization steps to facilitate better interaction with the reinforcement system. This is one of the most radical developments affecting the rubber industry because it enables silica to significantly displace carbon black as the favored reinforcement for the applications.

CONCLUSIONS

With global demand at around 3.7 million tons per year, SBR is one of most important synthetic elastomers. Nearly 70% of SBR is consumed by the automobile industry for tires and tire products.

There is growing competition between emulsion SBR and solution SBR grades. S-SBR has superior physical properties (particularly wear resistance) and blendability with other rubbers. However, they have been failing to make significant inroads, owing to processing difficulties and higher price. Therefore, E-SBR grades are used widely in the markets due to their low cost. While S-SBR has been paid more and more attention, people are striving to reduce the cost of preparation of S-SBR. New synthesis procedure and more effective catalysts are required in the future.

REFERENCES

1. http://www.Azom.com/.
2. He, Y.; Daniels, E.S.; Klein, A.; El-Aasser, M.S. Hydrogenation of styrene-butadiene rubber (SBR) latexes. J. Appl. Polym. Sci. **1997**, *64*, 2047–2056.
3. Abdel-Aziz, M.M.; Youssef, H.A.; El-Miligy, A.A.; Yoshii, F.; Makuuchi, K. Effect of polyfunctional monomers on radiation vulcanization of styrene-butadiene rubber. J. Elastomers Plast. **1996**, *28*, 288–305.
4. Kerns, M.L.; Henning, S.K. Synthesis and rheological characterization of branched versus linear solution styrene-butadiene rubber. Rubber Chem. Technol. **2002**, *75*, 299–308.
5. http://www.dow.com/.
6. Cerveny, S.; Bergman, R.; Schwartz, G.A.; Jacobsson, P. Dielectric alpha- and beta-relaxations in uncured styrene butadiene rubber. Macromolecular **2002**, *35*, 337–342.
7. Wang, C.Y.; Huang, X.P.; Yang, J.H. Cationic cyclization of styrene-butadiene rubber. Eur. Polym. J. **2001**, *37*, 1895–1899.
8. Ramesan, M.T.; Alex, R. Dichlorocarbene modification of styrene-butadiene rubber. J. Appl. Polym. Sci. **1998**, *68*, 153–160.
9. Pan, Q.M.; Rempel, G.L. Hydroenation of styrene butadiene rubber catalyzed by RU(CH=CHPh)Cl $(CO)(PCy_3)_2$. Macromol. Rapid Commun. **2004**, *25* (8), 843–847.
10. Findik, F.; Yilmaz, R.; Koksal, T. Investigation of mechanical and physical properties of several industrial rubbers. Mater. Des. **2004**, *25* (4), 269–276.
11. Visconte, L.L.Y.; Martins, A.F.; Suarez, J.C.M.; Nunes, R.C.R. Different preparative modes for the incorporation of additives in NR/SBR blends. J. Appl. Polym. Sci. **2004**, *93* (2), 483–489.
12. Varkey, J.T.; Augustine, S.; Groeninckx, G.; Bhagawan, S.S.; Rao, S.S.; Thomas, S. Morphology and mechanical and viscoelastic properties of natural rubber and styrene butadiene rubber latex blends. J. Polym. Sci. B Polym. Phys. **2000**, *38* (16), 2189–2211.
13. Ismail, H.; Suzaimah, S.; Hairunezam, H.M. Curing characteristics, mechanical properties and oil resistance of styrene butadiene rubber/epoxidized natural rubber blends. J. Elastomers Plast. **2002**, *34* (2), 119–130.
14. Sanchez-Solis, A.; Estrada, M.R.; Cruz, J.; Manero, O. On the properties and processing of polyethylene terephthalate/styrene-butadiene rubber blend. Polym. Eng. Sci. **2000**, *40* (5), 1216–1225.
15. Luzinov, I.; Xi, K.; Pagnoulle, C.; Huynh-Ba, G.; Jerome, R. Composition effect on the core-shell morphology and mechanical properties of ternary polystyrene/styrene butadiene rubber polyethylene blends. Polymer **1999**, *40* (10), 2511–2520.
16. Sadhu, S.; Bhowmick, A.K. Preparation and properties of nanocomposites based on acrylonitrile-butadiene rubber, styrene-butadiene rubber, and polybutadiene rubber. J. Polym. Sci. B Polym. Phys. **2004**, *42* (9), 1573–1585.
17. Su, X.Q.; Hua, Y.Q.; Qiao, J.L.; Liu, Y.Q.; Zhang, X.H.; Gao, J.M.; Song, Z.H.; Huang, F.; Zhang, M.L. The relationship between microstructure and properties in PP/rubber powder/nano-$CaCO_3$ ternary blends. Macromol. Mater. Eng. **2004**, *289* (3), 275–280.
18. Alert PERP Program—New Report; http://www.Nexant.comAlert (accessed Jan 2004).

Superabsorbents

Takamasa Nonaka
Faculty of Engineering, Department of Applied Chemistry and Biochemistry, Kumamoto University, Kurokami, Kumamoto-shi, Japan

INTRODUCTION

Woven cloth, cotton wadding, cellulose fiber batt, papers, and foamed polyurethane have been used as traditional absorbent materials for water. These materials can absorb 1–20 g of water per gram material and the water absorbed is easily removed from the materials by applying low pressure. In recent years, superabsorbent polymers, which can absorb up to 1000 g of water per gram of polymer and up to about 100 g of dilute salt solution per gram of polymer and the water absorbed can hardly be removed from the polymers even by applying high pressure, have been prepared and commercially used in many applications.

Superabsorbents are superabsorbent polymers, which are loosely crosslinked hydrophilic polymers and have ionic charges in the polymers. They absorb water, swell, and retain aqueous solution above hundred times their own weight. Because of the excellent characteristics, the superabsorbents have been widely used in many applications such as disposable diapers, feminine napkins, soil additives in agricultural or horticultural applications, gel actuators, water-blocking tapes, materials for drug delivery system, absorbent pads, etc., where water absorbency or water retention is important.

HYDROPHILIC POLYMERS FOR WATER ABSORPTION

It is known that papers, cotton, and cellulose are natural hydrophilic polymers, which can absorb water, but are insoluble in water. Therefore, they have been used as disposable diapers, feminine napkins, etc. They absorb water by capillary action. Therefore, they absorb only a small amount of water and the water absorbed is easily removed by applying low pressure. Synthetic crosslinked poly(vinyl alcohol) and crosslinked poly(oxyethylene) are also hydrophilic polymers and they have almost the same water-absorption capacity of about 1–20 g water per gram of polymer as natural polymers. Those natural polymers and synthetic polymers do not essentially have ionic groups. On the other hand, superabsorbent polymers having anionic groups that were made from low crosslinked starch or synthetic polymers were found to have high water-absorption capacity of more than 100 g water per gram of polymer. The water absorbed can hardly be removed even by applying high pressure. Thus, superabsorbent polymers should be hydrophilic polymers and low crosslinked and have ionic groups such as anionic groups, cationic groups, or betaine in the polymers.

CLASSIFICATION OF SUPERABSORBENTS[1]

The first polymer hydrogels called as superabsorbents were prepared by hydrolysis of starch-graft-poly(acrylonitrile) at Northern Research Institute of Department of Agriculture of the U.S.A. in 1974. Since then, many superabsorbents consisting of modified natural hydrophilic polymers such as starch, cellulose, and alginic acids have been prepared. Most of them have anionic moieties such as poly(sodium acrylate) or poly(vinyl sulfonate). Now, many types of superabsorbents having not only anionic groups but also cationic groups or betaine have been prepared from both natural hydrophilic polymers and synthetic hydrophilic polymers. The classification of superabsorbents is listed in Table 1.

As given in Table 1, superabsorbents can be classified in four kinds of parts:[1]

1. Superabsorbent polymers should be essentially hydrophilic polymers. Therefore, natural polymers such as starch and cellulose or synthetic polymers such as poly(acrylate), poly(vinyl alcohol), poly(acrylamide), or poly(oxyethylene) are used as hydrophilic base polymers.
2. Superabsorbents are low crosslinked to prevent to dissolve in water. Several methods for crosslinking are used:
   a. Crosslinked polymers can be prepared by copolymerization of hydrophilic vinyl monomers with crosslinking monomers such as methylenebisacrylamide, polyethyleneglycol dimethacrylate, divinylbenzene, etc.
   b. Water-soluble polymers such as poly(vinyl alcohol) or poly(oxyethylene) can be crosslinked by annealing or the reaction with crosslinking reagent such as glutaraldehyde.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007973

Table 1 Classification of superabsorbents

Base polymer	Starch—graft copolymerization, carboxymethylation Cellulose—graft copolymerization, carboxymethylation Synthetic polymers—poly(acrylate), poly(vinyl alcohol), poly(acrylamide), poly(oxyethylene), cationic polymer
Method of crosslinking	Copolymerization with crosslinking monomers Crosslinking of water-soluble polymer Irradiation of radioactive ray Self-crosslinking Introduction of crystal structure
Method of introduction of ionic groups	Polymerization of ionic monomer Caboxymethylation of hydrophilic nonionic polymers Graft copolymerization of hydrophilic monomer to hydrophobic polymer Hydrolysis of nitrile or ester group
Shape of products	Powder—spherical Film—amorphous Fiber—long fiber, short fiber, nonwoven cloth

   c. Many polymers can be crosslinked by irradiation of radioactive rays such as γ-ray.
   d. Poly(sodium acrylates) are partially crosslinked by self-crosslinking during polymerization at high monomer concentration.
   e. Water-soluble polymers such as poly(vinyl alcohol) can be crosslinked by crystallization by annealing or freeze-drying.

3. Superabsorbents should have ionic groups in the polymers.
   a. Ionic groups can be introduced by polymerization of monomers having ionic groups.
   b. Ionic groups can be introduced by carboxymethylation of nonionic polymers such as starch and cellulose.
   c. Ionic groups can be introduced by graft copolymerization of monomers having ionic groups on nonionic polymers such as starch, cellulose, or poly(vinyl alcohol).
   d. Ionic groups can be introduced by graft copolymerization of monomers having nitrile groups or ester groups on starch or cellulose, followed by hydrolysis.
4. Superabsorbents are commercialized in many shapes of products. Most superabsorbents are amorphous powders. Film and fiber types of superabsorbents are also used.

PREPARATION OF SUPERABSORBENTS[2]

Up to the present, many types of superabsorbent polymers have been prepared and commercialized. Some commercial superabsorbents and their manufacturers are listed in Table 2. Several examples are described here:

1. Starch/acrylonitrile copolymers. These are the first superabsorbents, which were prepared at Northern Research Institute of Department of Agriculture of the U.S.A. in 1974. These were prepared by graft copolymerization of acrylonitrile on starch using ammonium cerium(IV) nitrate as initiator, followed by hydrolysis with sodium hydroxide aqueous solution. Therefore, these have anionic charges of carboxylic group. The copolymers are easily subject to disintegrate with microorganism, because basic polymers are natural polymer. Therefore, these polymers do not retain water for a long period.
2. Starch/acrylic acid copolymers. These were prepared by graft copolymerization of acrylic acid on starch. The crosslinking of the polymers was carried out by simultaneous graft copolymerization of crosslinking monomers or post-crosslinking with crosslinking reagent such as epichlorohydrin or glutaraldehyde. The polymers cannot retain water for a long time, because the copolymers are subject to disintegration with microorganism. The types of polymers are mostly amorphous powder.
3. Carboxymethylated celluloses. These were prepared by the reaction of celluloses with monochloroacetic acid, followed by crosslinking. These are mostly used in fiber form.
4. Poly(sodium acrylates). Poly(acrylic acids) or poly(sodium acrylates) were obtained by i) solution polymerization of acrylic acid or sodium

Table 2 Commercial superabsorbents and their manufacturers

	Manufacturer	Commercial name	Compositions
Europe	BASF	Luquasorb	Poly(sodium acrylate)
	Enka	Akucell	Carboxymethylcellulose
	Stockhausen	Favor	Poly(sodium acrylate)
	Unilever	Lyogel	Hydrolyzed starch/acrylonitrile copolymer
U.S.A.	Buckeye	CLD	Carboxymethylcellulose
	Dow Chemical	DWAL	Poly(sodium acrylate)
	Grain Processing	GPC	Hydrolyzed starch/acrylonitrile copolymer
	Henkel	SGP	Hydrolyzed starch/acrylonitrile copolymer
	Hercules	Aqualon	Carboxymethylcellulose
	National Starch	Permasorb	Poly(sodium acrylate)
	Super Absorbent	Magic Water Gel	Hydrolyzed starch/acrylonitrile copolymer
Japan	Nippon Shokubai	Aqualic	Poly(sodium acrylate)
	Sanyo Kasei	Sanfresh	Starch/acrylic acid copolymer
	Daicel	CMC Daicel	Carboxymethylcellulose
	Nippon Exlan	Espec	Acryl fiber/sodium acrylate composite fiber
	Sumitomo Seika	Aquakeep	Poly(sodium acrylate)

acrylate in water or by ii) reverse suspension polymerization in organic solvent. Crosslinked poly(acrylic acids) or poly(sodium acrylates) were prepared by copolymerization with crosslinking reagent such as methylenebisacrylamide or ethylene glycol dimethacrylate or post-crosslinking with diglycidyl compounds or higher alcohols. Crosslinked poly(acrylic acids) or poly(sodium acrylates) can also be obtained, when the polymerization is carried out at high concentration of the monomer in an aqueous solution. This crosslinking is called self-crosslinking. Crosslinked poly(sodium acrylates) are mostly prepared by solution polymerization and used as superabsorbents in hygienic application.

5. Vinyl alcohol/acrylic acid copolymers. Vinyl alcohol/sodium acrylate copolymers were obtained by hydrolysis of vinyl acetate/methylacrylate copolymers with alkaline solution. The vinyl alcohol/sodium acrylate copolymers obtained are insoluble in water, although they are not crosslinked with crosslinking monomer. This is because of crystal structure of vinyl alcohol moieties in the copolymers. Therefore, the vinyl alcohol/sodium acrylate copolymers have higher mechanical strength than crosslinked poly(sodium acrylates) in water.
6. 3-Methacrylamidepropyl trimethylammonium chloride/*N,N′*-methylenebisacrylamide (MBAAm) copolymers.[3] This type of crosslinked 3-methacrylamidepropyl trimethylammonium chloride/*N,N′*-methylenebisacrylamide copolymers were prepared by copolymerization of 3-methacrylamidepropyl trimethylammonium chloride and *N,N′*-methylenebisacrylamide (crosslinking monomer). The copolymers have ammonium groups, which exist as cationic polymers in water.
7. Trialkyl-4-vinylbenzyl phosphonium chloride/acrylamide copolymers.[4] This type of crosslinked trialkyl-4-vinylbenzyl phosphonium chloride (TRVB)/acrylamide (AAm) copolymers were prepared by copolymerization of TRVB, AAm, and MBAAm (crosslinking monomer) in dimethyl sulfoxide. Three TRVBs with different alkyl chain lengths (butyl, hexyl, and octyl) in phosphonium groups were used. They are abbreviated as TBVB, THVB, and TOVB, respectively. The copolymers obtained have phosphonium groups. Therefore, they are crosslinked cationic polymers.
8. Trialkyl-4-vinylbenzyl phosphonium chloride/*N*-isopropylacrylamide (NIPAAm) copolymers.[5] These types of crosslinked TRVB/NIPAAm copolymers were prepared by copolymerization of TRVB, NIPAAm, and MBAAm in dimethyl sulfoxide. Poly(NIPAAm) is a thermosensitive polymer, which has a lower critical solution temperature (LCST) at around 33°C. Therefore, the copolymers are crosslinked cationic polymers and thermosensitive polymers, which swell and deswell below and above the LCST (about 33°C) of poly(NIPAAm).

In Fig. 1, the structures of superabsorbent polymers (a)–(d), (g), and (h) are shown.

PROPERTIES OF SUPERABSORBENTS

Water-Absorption Capacity of Superabsorbent Polymers

The water-absorption capacity of superabsorbent polymers is usually measured by the following methods. The water-absorption capacity is greatly affected by not only the chemical structures of the polymers but also the external conditions such as kinds of solutions, temperature, pressure, etc. under which it is measured.

Tea bag method[6]

The water content of superabsorbent polymers is usually measured by a tea bag method as follows: The dried copolymers are put into a tea bag made of nonwoven fabric. Then they are soaked into excess deionized water or salt solutions for 24 hr at desired temperature. After the tea bag containing polymers are soaked in aqueous solution, the water unabsorbed into gels is removed by placing the tea bag in air for a short time. The water on the surface of the gels and the tea bag is immediately wiped with filter paper, and the weight (W_w) of the tea bag containing superabsorbent polymers is measured. The water content (Q) (H_2Og/g-polymer) of the polymers is calculated using Eq. (1):

$$Q = \frac{W_w - W_t - W_d}{W_d} \quad (1)$$

where W_t and W_d are the weights of the wet tea bag and the dried polymers, respectively.

In some cases, the water unabsorbed into superabsorbents can be removed by filtration or centrifugation.

UV absorption method[7]

A prescribed amount of superabsorbent polymer is soaked in a prescribed aqueous solution containing blue dextrin, which does not penetrate into superabsorbent polymers. The concentration of blue dextrin in outer solution after equilibrium swelling of the polymers is measured by UV spectrophotometer. And the

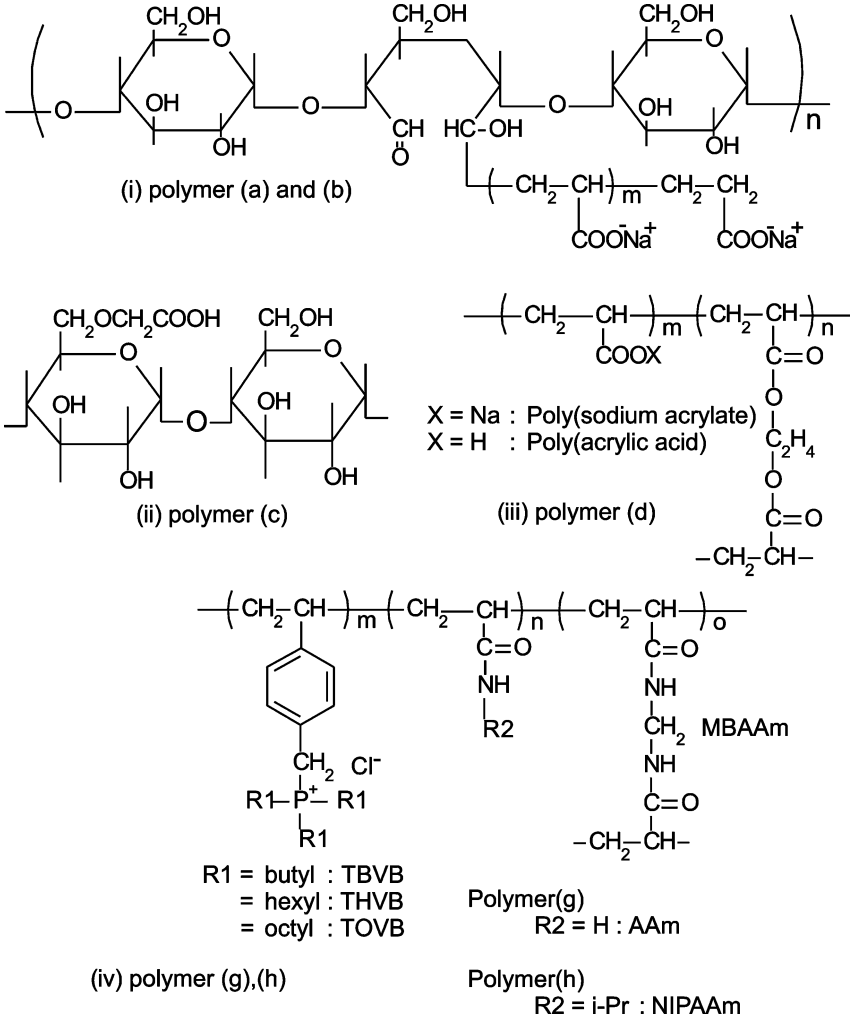


Fig. 1 The structure of super absorbent polymer (a)–(d), (g), and (h).

water-absorption capacity of the polymers is calculated from the difference of the concentration of dextrin in the presence of polymers and that in the absence of polymers.

Vortex time method[8]

A prescribed amount of superabsorbent polymer is mixed with a prescribed amount of the desired aqueous solution, which is stirred by means of a magnetic stirring bar in a small beaker. As the water absorption proceeds, the viscosity of the suspension increases until the stirring vortex disappears at the "sorption time," t_v. The swelling capacity at equilibrium (Q_{max}) is calculated by using Eq. (2):

$$t_v = \frac{-1}{k} \ln\left(1 - \frac{Q}{Q_{max}}\right) \tag{2}$$

where k is the first-order rate constant and Q is the swelling capacity (H_2Og/g-polymer) at any time t.

Mechanism of Water Absorption with Superabsorbents

As mentioned before, superabsobents should be essentially low crosslinked hydrophilic polymers and have ionic groups such as anionic groups, cationic groups, or betaine in the polymers. They usually have a high water-absorption capacity of more than 100 g-water/g-polymer even in dilute salt solution. In Fig. 2, swollen crosslinked poly (sodium acrylates) are illustrated.

In general, the swelling ratio of hydrogels, which corresponds to water content, can be expressed by Eq. (3):[9]

$$Q^{5/3} = \left\{\left[\frac{1}{2} \times \frac{i}{V_u} \times \frac{1}{S^{1/2}}\right]^2 + \frac{(1/2 - X_1)}{V_1}\right\} \Big/ \frac{\nu}{V_0} \tag{3}$$

where Q is the swelling ratio, i/V_u the charge density attached to polymer matrix, $((1/2) - X_1)$ the affinity between polymer matrix and water, $S^{1/2}$ the ionic strength of outer solution, and ν/V_0 the crosslinking density.

The first term $[(1/2) \times (i/V_u) \times (1/S^{1/2})]$ in this equation represents the osmotic pressure because of ions, the second term $\{[(1/2) - X_1]/V_1\}$ represents the affinity of polyelectrolyte for water, and the third term (V_0/ν) represents the degree of crosslinking. Therefore, this equation indicates that the water content of the superabsorbent polymers depends on the hydrophilicity, crosslinking density, charges of the crosslinked polymers, and the concentration of neutral salts in an aqueous solution.

The various effects on the water-absorption ability of the polymers were investigated with superabsorbent polymers having anionic or cationic groups. Several effects on the water-absorption capacity of the superabsorbent polymers having carboxyl groups(1) or phosphonium groups(4) are described in the following section.

Effect of ionic groups on the water-absorption ability of superabsorbent polymers

In Figs. 3A and B the effect of the content of ionic groups on the water-absorption ability of starch/acrylic acid copolymers [polymer (d)] and the

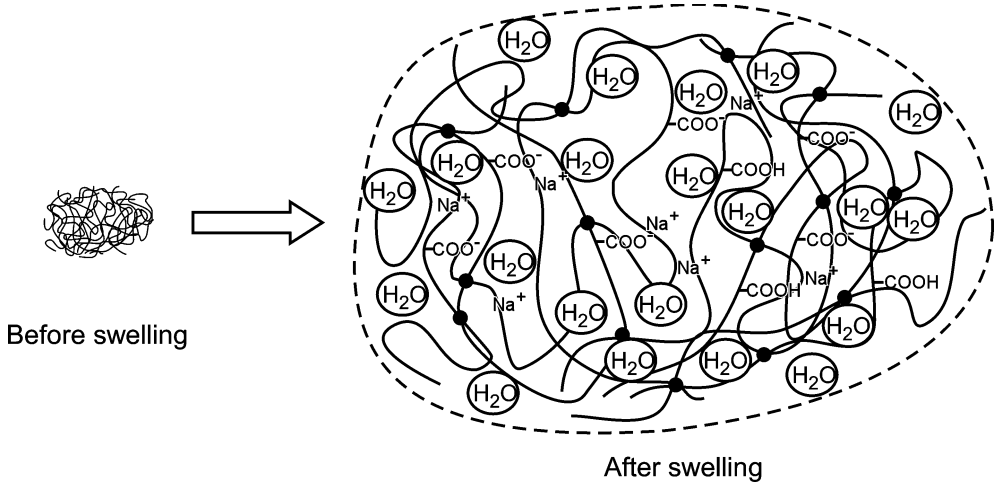


Fig. 2 Illustration of swollen crosslinked poly(sodium acrylate).

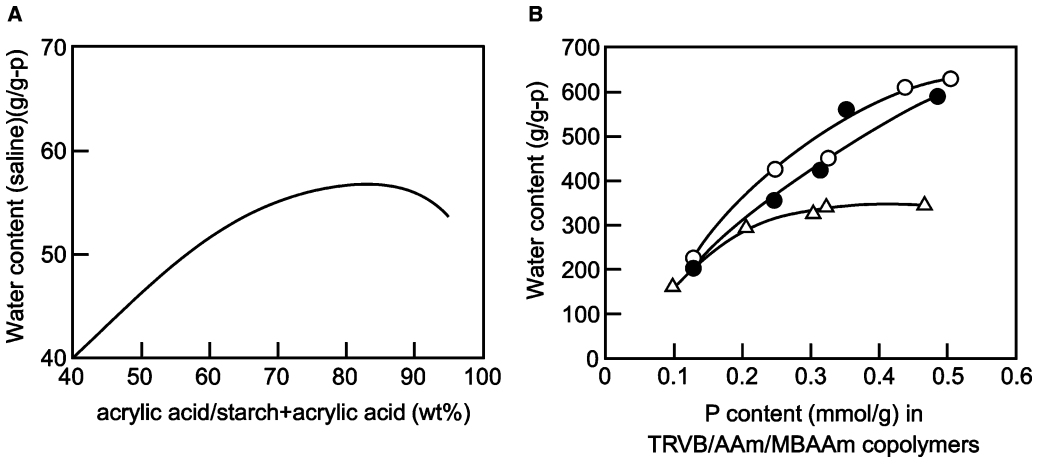


Fig. 3 Effect of ionic groups on the water absorption of (A) starch/acrylic acid and (B) TRVB/AAm/MBAAm [X : (97 - X) : 3] copolymer. Measured at room temperature in 0.9% NaCl aqueous solution. Copolymer: (○) TBVB–AAm–MBAAm, (●) THVB–AAm–MBAAm, and (△) TOVB–AAm–MBAAm. Measured at room temperature in deionized water.

TRVB/AAm/MBAAm copolymers [polymer (g)] containing phosphonium groups is shown, respectively. It is shown in Fig. 3A that the water content of the starch/acrylic acid copolymers increased with increasing content of acrylic acid, and then decreased above 90 wt.% of acrylic acid in the copolymers. In Fig. 3B, the water content of the TRVB/AAm/MBAAm copolymers having various contents of phosphonium groups in deionized water is shown. The water content increased with increasing phosphonium content in the copolymers, although the water contents of copolymers were different depending on the chain length of alkyl groups in phosphonium groups. This increase in water content with increasing content of ionic groups in the copolymers is because of the increasing osmotic pressure between the inner and outer sides of the polymer gels. It is also shown in Fig. 3B that the water content increased in the order of TBVB/AAm/MBAAm > THVB/AAm/MBAAm > TOVB/AAm/MBAAm copolymer.

This order is inversely proportional to the chain length of alkyl chain in the phosphonium groups in the copolymers. This indicates that hydrophilic polymers with shorter alkyl chain in phosphonium groups absorb more water, although they have almost the same ionic groups in the copolymers.

Effect of the degree of crosslinking on the water-absorption ability of superabsorbent polymers

In Figs. 4A and B the effect of the degree of crosslinking on the water-absorption ability of starch/acrylic acid copolymers [polymer (d)] and the TRVB/AAm/MBAAm copolymers [polymer (g)] containing phosphonium groups is shown, respectively.

It is shown in Fig. 4A that the water content of the starch/acrylic acid copolymers increased first, and then decreased with increasing content of crosslinking reagent, although the water content of starch/acrylic acid copolymers is different depending on the kind of crosslinking reagents. This indicates that hydrophilic copolymers should be crosslinked to some extent to retain high water content, and then the water content of the copolymers decreased with increasing degree of crosslinking of the copolymers.

In Fig. 4B, the water content of the TBVB/AAm/MBAAm copolymers having different degree of crosslinking and almost the same phosphonium groups in deionized water is shown. The water contents of AAm–MBAAm copolymer carrying no phosphonium groups are also shown for comparison.

The water content of the copolymers having phosphonium groups was fairly high and it decreased remarkably with increasing degree of crosslinking in the copolymers. However, the water content of the AAm–MBAAm copolymer carrying no phosphonium groups was fairly low compared with that of the copolymers having phosphonium groups and it decreased slightly with increasing degree of crosslinking of the copolymers. This result indicates that the introduction of ionic groups such as phosphonium groups into AAm–MBAAm zcopolymer is necessary to retain high water content.

In recent years, it is reported that the water-absorption capacity of crosslinked poly(sodium acrylates) can be increased by increasing the crosslinking density near the surface and by decreasing the crosslinking density at the inner side of superabsorbent copolymers.[10] It is also reported that this type of superabsorbent polymer has high mechanical strength.

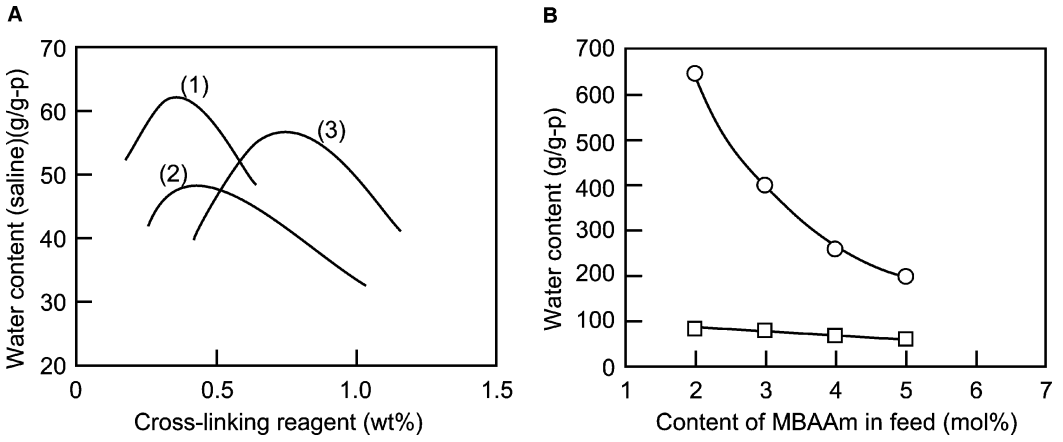


Fig. 4. Effect of degree of crosslinking on the water absorption of (A) starch/acrylic acid and (B) TBVB/AAm/MBAAm copolymer. (A) Crosslinking reagent: 1) MBAAm; 2) ethyleneglycol diglycidyl ether; and 3) poly(ethyleneglycol) diglycidyl ether. Measured in 0.9%NaCl solution. (B) (○) TBVB/AAm/MBAAm [2 : (98 - Y) : Y], (□) AAm/MBAAm. Measured in deionized water.

Effect of inorganic compounds in water on the water-absorption ability of superabsorbent polymers

In Figs. 5A and B the effect of inorganic compounds in water on the water-absorption ability of starch/acrylic acid copolymers [polymer (b)] and the TRVB/AAm/MBAAm copolymers [polymer (g)] containing phosphonium groups is shown, respectively.

It is shown in Fig. 5A that the water content of the starch/acrylic acid copolymers decreased sharply with increasing concentration of neutral salts (NaCl, $MgCl_2$) or NaOH in aqueous solutions up to about 1 wt.% and then decreased gradually.[11] The order of effect is as follows:

$$MgCl_2 > NaCl > NaOH$$

This result indicates that $MgCl_2$ containing divalent Mg^{2+} has a more pronounced effect than does NaCl containing monovalent Na^+.

The decrease in the water absorption by an addition of neutral salts is because of the decrease in the osmotic pressure between inner and outer sides of the polymer hydrogels in their aqueous solution.

In addition, it is known that the expansion of polymer chains because of the electrolytic repulsion between ionic groups of the polymers reduces, because an addition of these inorganic electrolytes shields the ionic charges of the polymers. This also results in the reduction of the water absorption of the superabsorbent polymers.

In NaOH solution, parts of acrylic acid groups in starch/acrylic acid copolymers are neutralized with NaOH, and sodium acrylates formed dissociate easily

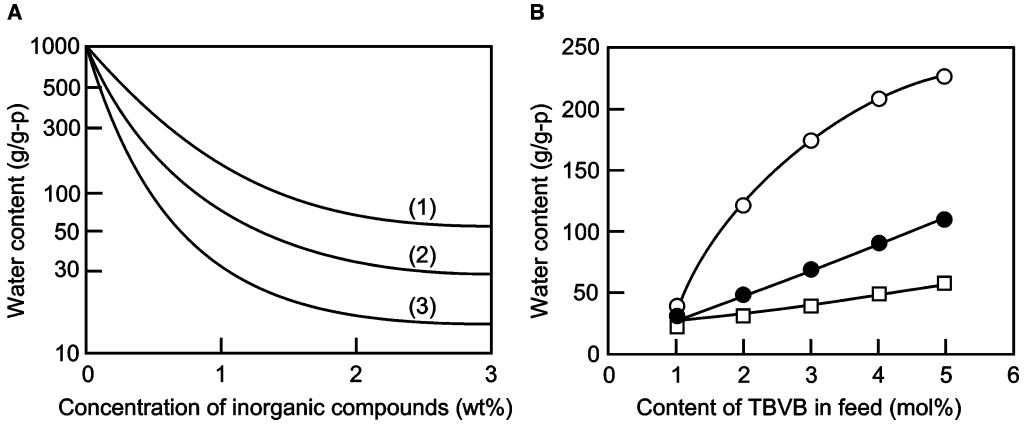


Fig. 5. Effect of addition of inorganic compounds on the water absorption of (A) starch/acrylic acid and (B) TBVB/AAm/MBAAm [X : (97 - X) : 3] copolymer. (A) Inorganic compounds: 1) NaOH; 2) NaCl; and 3) $MgCl_2$. (B) NaCl concentration (wt.%): (○) 0.009, (●) 0.09, and (□) 0.9.

in water. Therefore, starch/sodium acrylate copolymers become more polar hydrophilic polymers. This is the reason for the higher water content of the copolymers in NaOH solution than in other neutral salt solutions. However, the water content decreased also with increasing concentration of NaOH even in this aqueous solution.

In Fig. 5B, the water content of the TRVB/AAm/MBAAm [X : (97 - X) : 3] copolymers having different phosphonium groups and constant degree (3 mol%) of crosslinking in 0.009, 0.09, and 0.9 wt.% NaCl solutions is shown. The water content of TBVB/AAm/MBAAm copolymer increased with increasing TBVB content in the copolymers and decreased with increasing concentration of NaCl in solutions and they were considerably depressed in NaCl solutions above 0.09 wt.%. The degree of the decrease in water content by the addition of NaCl increased with increasing content in phosphonium groups in the copolymers. Therefore, the decrease in water content of the copolymers by an addition of inorganic compounds is because of both the decrease in the difference of osmotic pressure between the inner side and the outer side of the copolymer hydrogels and shielding effect of ionic charges of the copolymers by an addition of inorganic compounds.

Effect of physical structure of superabsorbent polymers on the water-absorption rate or water-absorption capacity

The water-absorption rate of superabsorbent polymers is affected by: 1) the specific surface area of the polymers; 2) the capillary action; and 3) the formation of fish-eyes in the polymer hydrogels.

1. Specific surface area of polymer hydrogels can usually be increased by decreasing the particle size of the polymers. In Fig. 6A, the relationship between water adsorption rate or absorption capacity and flaky particle size of crosslinked poly(sodium acrylates) is shown. The water adsorption rate increased with decreasing particle size in the range of particle size from 800 to 100 μm and then decreased. The decrease in the water-absorption rate with decreasing particle size (less than 100 μm) is because of the formation of fish-eyes in polymer hydrogels. The water-absorption rate can also be increased with decreasing apparent density (g/ml) of flaky polymer particles [Fig. 6B]. The decrease in apparent density of the polymers results in the increase in specific surface area of the polymers. This indicates that the water-absorption rate of the polymers increases with the specific surface areas of the polymers.
2. The water-absorption rate of the polymers can be increased by use of capillary action. It is reported that fibrous carboxymethyl celluloses, which were crosslinked with epichlorohydrin and had the carboxymethylation degree of 0.2–0.4, had high water-absorption rate of 20–40 ml/g-polymer/min.
3. The water-absorption rate of the polymers can be increased by mixing inorganic particles such as kaolin, talc, etc. with superabsorbent polymers. This mixing procedure results in the decrease in the formation of fish-eyes in polymer gels. In addition, the increase in the water-absorption rate is because of the fact that inorganic particles adsorbed on superabsorbent copolymers give space between polymer particles.

Effect of temperature or pH on water-absorption rate or -absorption capacity

The water-absorption rate of usual superabsorbent polymers can be increased by increasing the temperature, but the water-absorption capacity after equilibrium at

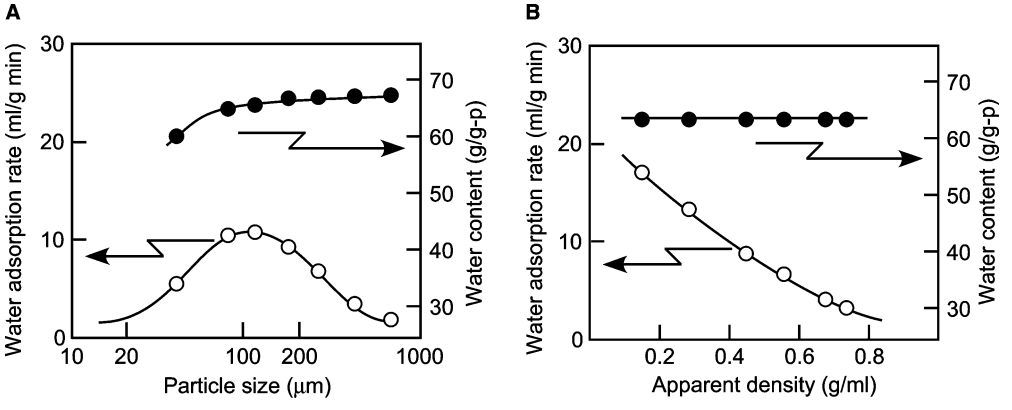


Fig. 6 Effect of particle size (A) and apparent density (B) on the water-absorption rate and water content of crosslinked poly(sodium acrylate).

each temperature is almost constant. The effect of pH on the water-absorption rate or -absorption capacity depends on the kind of ionic groups in the superabsorbent polymers. Usually the water-absorption capacity of superabsorbent polymers having weak acids such as acrylic acid increases with increasing pH. This is because of the formation of poly(acrylic salt) such as poly(sodium acrylate). On the other hand, the water-absorption capacity of superabsorbent polymers having weak bases such as amino groups increases with decreasing pH. This is because of the formation of ammonium groups by protonation of amino groups. In both cases, the water-absorption capacity decreases at too high pH or at too low pH. The water-absorption capacity of superabsorbent polymers having strong acids such as sulfonic acid groups or strong bases such as ammonium groups or phosphonium groups essentially does not depend on pH of outer solution, but even in this case, the water-absorption capacity decreases at too low and too high pH. The decrease in the water-absorption capacity at too high and low pH is because of both the decrease in the difference of osmotic pressure between the inner side and the outer side of the polymer hydrogels and shielding effect of ionic charges of the polymers.

Effect of organic solvent on the water absorption of superabsorbent polymers

In Fig. 7, the water absorption of crosslinked poly(sodium acrylate) in various alcohols/water mixtures is shown. The water absorption decreases with increasing alcohol content and decreases abruptly at above certain limit of alcohol content. The content at which the abrupt decrease occurs depends on the kinds of alcohols. The similar phenomena are also observed in other organic solvents such as acetone/water mixtures.

OTHER PROPERTIES OF SUPERABSORBENT POLYMERS

Evaporation of Water Absorbed in the Polymers[2]

Water imbibed with superabsorbent polymers can hardly be removed even by applying high pressure, but the water absorbed is gradually removed by drying in air. In Fig. 8, the time dependence of decrease of water absorbed with vinyl alcohol/acrylic acid copolymers by drying at the temperatures of 20°C, 50°C, and 100°C is shown.

It takes long time for water to evaporate from the superabsorbent polymers, although the rate of evaporation of water increased with increasing temperature. This result suggests that superabsorbent polymers can be used to retain water for a long time in soils for the agriculture in dry land.

Stability of Superabsorbent Polymers Against UV Light[1]

The stability of superabsorbent polymers against UV irradiation was evaluated by measuring the decrease of viscosity of swollen gels of poly(acrylate). In Fig. 9, the changes of the retention of the viscosity of the polymer gels UV-irradiated as a function of the degree of neutralization of acrylic acid moiety in crosslinked poly(acrylic acids) with alkali are shown.

The degree of the retention of the viscosity of swollen gels increased with increasing degree of

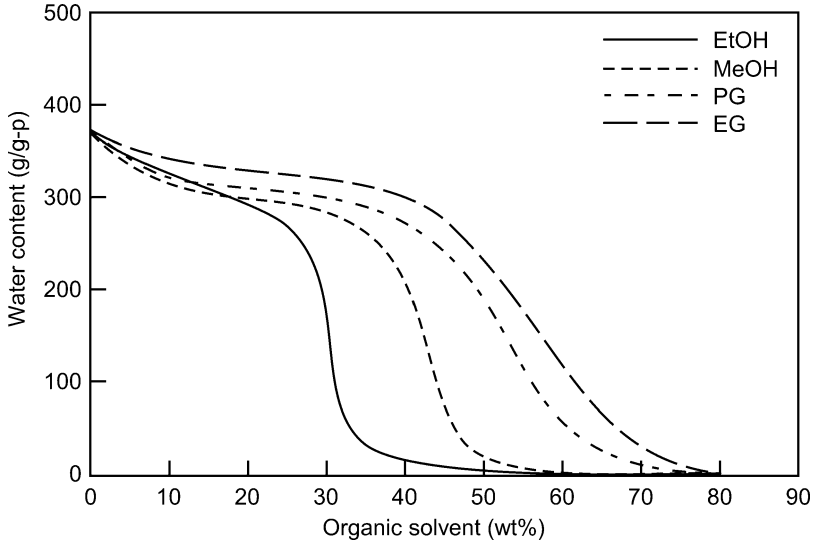


Fig. 7 Absorption of water in alcohol/water mixtures with crosslinked poly(sodium acrylate): EtOH, ethanol; MeOH, methanol; PG, propylene glycol; EG, ethylene glycol.

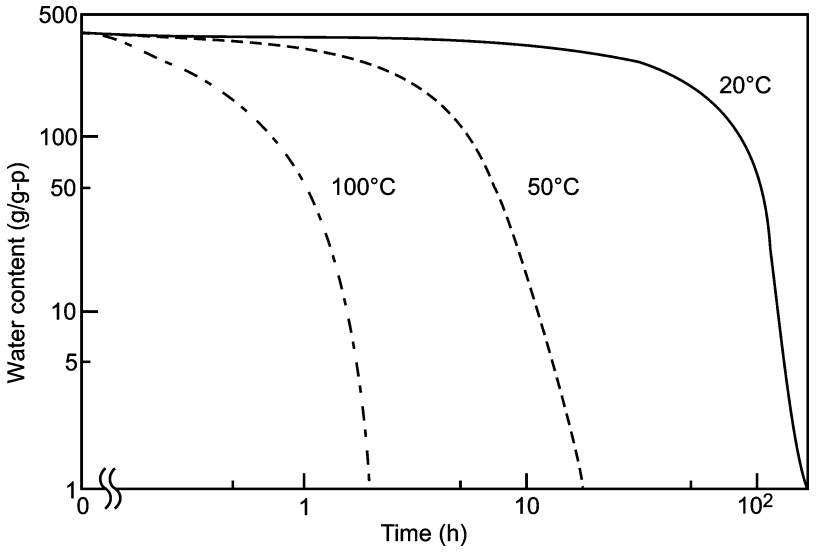


Fig. 8 Time dependence of the decrease of water absorbed with vinyl alcohol/acrylic acid copolymer at different temperatures.

neutralization of acid moiety in crosslinked poly (acrylic acids). It was found that poly(sodium acrylates), which was completely neutralized with alkali, are quite stable against UV light. It is reported that the significant decrease in the viscosity of crosslinked poly(acrylic acid) having a degree of neutralization less than 90% is because of the fact that radicals formed by UV irradiation to hydrogen of carboxyl group decomposed the main chain of poly (acrylic acids).

Adsorption of Ammonia with Superabsorbent Polymers Containing Carboxyl Groups[1]

Superabsorbent polymers consisting of poly(sodium acrylates) have partially carboxyl groups, which are weak acid groups. Therefore, they can adsorb ammonia, which is a weak base. It is shown in Fig. 10 that crosslinked poly(acrylic acid) adsorbs ammonia, which is formed by the decomposition of urine with urease.

Antibacterial Ability of Superabsorbent Polymers Containing Cationic Groups[5]

It is reported that cationic polymers having ammonium groups, pyridinium groups, or phosphonium groups exhibit antibacterial activity against bacteria such as *Escherichia coli* or *Staphylococcus aureus*.[12–14] It is shown in Fig. 11 that TBVB–NIPAAm–MBAAm copolymer [polymer (h)], which has phosphonium groups and high water-absorption ability, had high antibacterial activity against *S. aureus*.

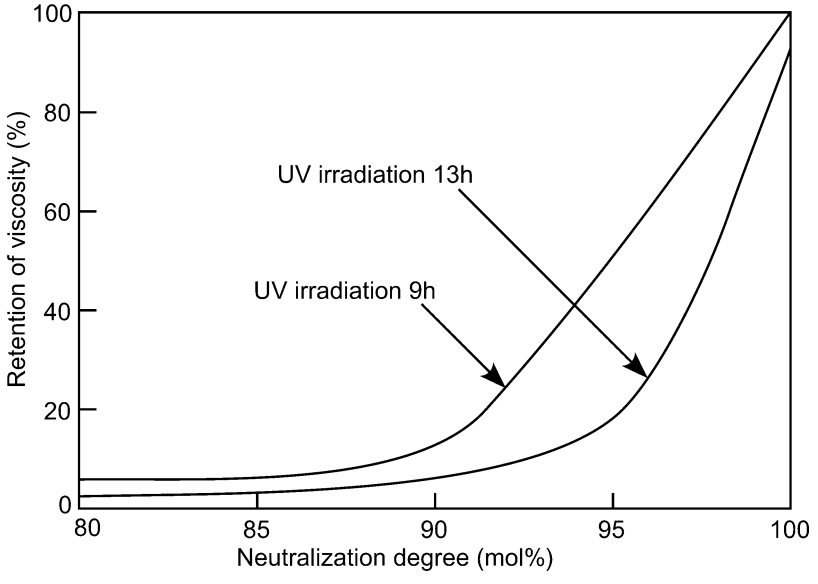


Fig. 9 Stability of swollen crosslinked poly (acrylate) neutralized with alkali against UV irradiation gel swollen about 50 times its own weight in water was used for this experiment.

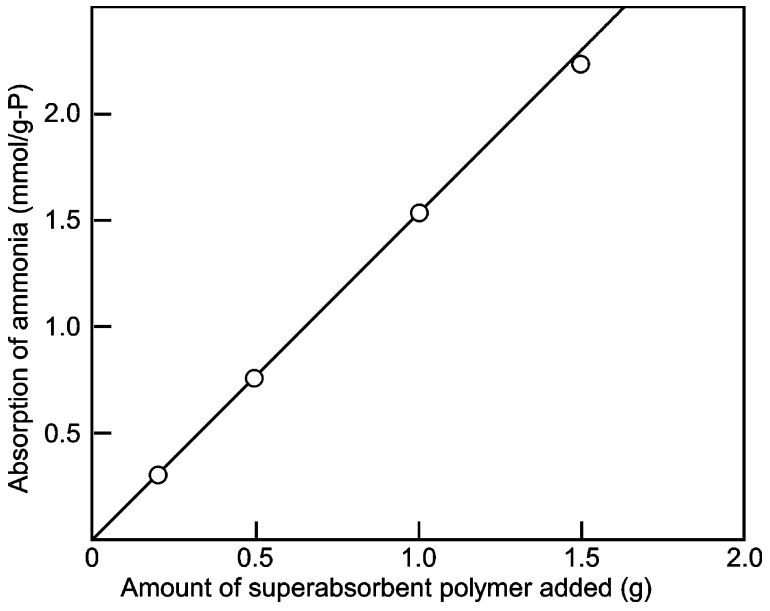


Fig. 10 Adsorption of ammonia with crosslinked poly(acrylic) acid. Conditions: urine, 50 ml; urease, 10 mg; 37°C; 2 hr.

It is also shown that the polymer having higher content of TBVB had higher antibacterial activity than did the copolymer with lower content of TBVB. They had the maximum antibacterial activity at 30°C and the water-absorption capacity of the copolymers decreased with increasing temperature and, in particular, decreased rapidly above 33°C, because the copolymers have LCST at around 33°C. The copolymers remaining in water can be separated by raising the temperature of aqueous solution above the LCST of the copolymers after using as an antibacterial reagent, because they are insoluble in water above the LCST of the copolymers, and they can be reused.

APPLICATION[15]

Superabsorbent polymers have been used in a variety of fields. In Table 3, the applications of superabsorbent polymers are listed.

Personal Hygiene Products[16–18]

The most widely spread use of superabsorbent polymers is in personal hygiene products such as disposable infant diapers, feminine sanitary napkins, and adult incontinence articles. In particular, over 90% of the total superabsorbent polymers are sold as infant

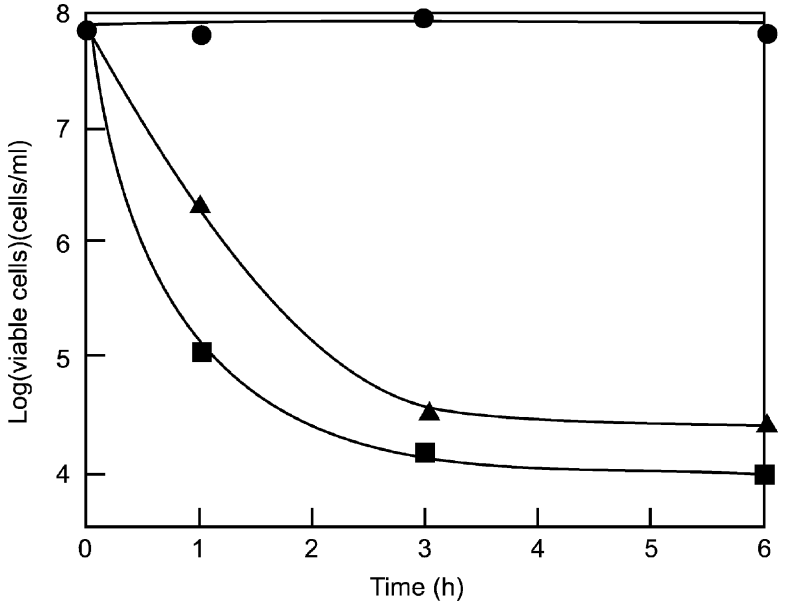


Fig. 11 Antibacterial activity of TBVB–NIPAAm–MBAAm copolymer against *S. aureus* in deionized water at 30°C: (●) blank, (▲) TBVB–NIPAAm–MBAAm (1 : 100 : 3), and (■) TBVB–NIPAAm–MBAAm (3 : 100 : 3) copolymer.

Table 3 Application of superabsorbent polymers

Hygiene products	Disposable infant diapers Feminine sanitary napkins Adult incontinence articles
Agriculture and horticulture	Water retention of water in soils Sheet for cultivating paddy Seed coating Artificial sphagnum
Food packing	Fresh maintenance of foods Drip absorption Water removal from foods Cold insulator
Civil engineering and construction industry	Dew prevention Sealing water of leakage Hardening sluge Drilling
Toiletry	Fragrant gel Sweat band Portable toilet Disposable pocket heater Milky liquid pack
Controlled release and medical field	Sheet for surgical operation Drug delivery Stupe Dressing for protection of wound
Electronics and cabling	Papers for ink jet recorder Prevention from water penetration into communication and light fiber cable
Others	Water sensor Artificial snow Fire extinguishing water Water swellable toys Paint for prevention of water leakage Removal of water in oil

diapers. In Fig. 12, the schematic diagram of diapers is given. A diaper consists of an absorbent core that is sandwiched between a porous top-sheet and impermeable back-sheet. The top-sheet is made of a porous, hydrophobic substance, e.g., polyester or propylene nonwoven fabric, and the back-sheet is a nonporous, hydrophobic substance, e.g., polyethylene film. The absorbent core has a composite structure, with the cellulose pulp fluff and superabsorbent polymers randomly mixed. The superabsorbent polymers are also used for feminine hygiene products or adult incontinence products in the same manner as in infant diapers.

Agricultural and Horticultural Application[19]

Superabsorbents can also help conserve water in a variety of agricultural and horticultural applications. Mainly, the polymers are used in the same way as much, to help the soil retain moisture as soil additives. The polymers are mixed into soil at a concentration of about 1 wt.%. The resulting soil mixture retains moisture longer, and plants live longer after germination.

Food Packaging[20]

Superabsorbent polymers are used as a liquid-absorber in food packaging systems. In these systems, the superabsorbent polymers absorb juice or water from fresh foods such as raw chicken, shellfish, and other meats or from frozen foods as they thaw. Chilled superabsorbent polymer gels may also be used as a dry-cooling medium. The water-swollen gel, contained in a durable plastic bag, is frozen and used to keep perishable foods cold. In addition to its liquid-water-absorption characteristic, superabsorbent polymers absorb water from the vapor state and therefore may be used to control humidity.

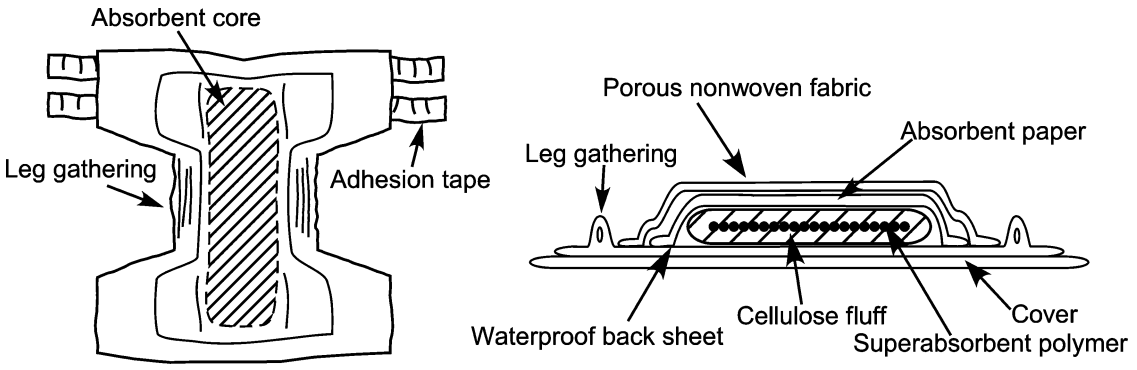


Fig. 12 Schematic diagrams of infant diapers.

Civil Engineering and Construction Industry[21–24]

Superabsorbent polymers have a number of uses in the civil engineering and the construction industry. One advantage of using superabsorbents in construction is to use the volume increase of the gel to form a barrier to further the water flow. Sealing composites made by blending a superabsorbent into a rubber such as chloroprene or thermoplastic elastomer such as poly (ethylene-co-vinyl acetate) have been developed for sealing around the joints of various building materials. The sealing composite may be used like mortar between the concrete blocks that make up the walls of the structure. If any gaps were left during construction or created by shifting after construction, the superabsorbent swells in any leaking water to fill any gaps and prevents water from leaking through the joint. A water-blocking construction filler that is composed of cement, water-absorbing polymer, and an asphalt emulsion has also been developed.

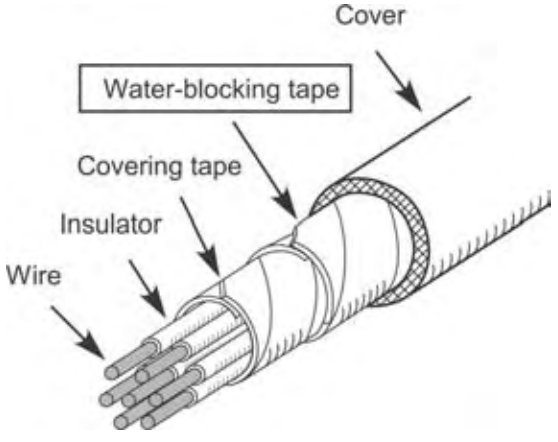


Fig. 13 Wire protected with a water-blocking tape containing superabsorbent polymer.

Electronics and Cabling[25,26]

The swelling property of superabsorbents is also applied to protect communication cables from water damages. Leaking water degrades the performance of fiberoptic communication cables and power transmission cables. Water-blocking tapes prevent intrusion of water into the cables. A flexible, water-blocking tape may be made by mixing a superabsorbent polymer and a polymeric binder, and then spreading the mixture on to a nonwoven fabric. Alternatively, the superabsorbent is mixed with a rubber such as butyl rubber and solvent and then coated onto a polyester tape. The tape is wrapped around the cable, beneath the plastic covering, as shown in Fig. 13.

Controlled Release and Medical Field[27,28]

Superabsorbent polymers may also be used to control the release of substances that are initially dissolved or trapped within the polymer phase, such as pesticides, fertilizers, and pharmaceuticals. Just as the absorption of water by superabsorbent polymer is caused by unequal activities of water inside the gel and in the external liquid phase, substances that are initially at a higher activity in the polymer will diffuse out of the particle and into the surroundings. The swelling of crosslinked poly(acrylic acid) and other acidic or basic polymers depends on the pH and ionic strength of the swelling medium, and the changes in pH or ionic strength may serve as switches for controlled release. Superabsorbent polymers are also used as materials in the sheet for surgical operation, stupe, or dressing for protection of wound.

Toiletry[29,30]

Superabsorbent polymers have been used in several cosmetic formulations. A skin-cream emulsion that

was prepared with a crosslinked superabsorbent in addition to branched poly(acrylic acid) thickener for cosmetic exhibited a long-term moisturizing effect than did the emulsion without the superabsorbent polymer. Superabsorbents have also been used as a component in a gel-form cosmetic face mask. Superabsorbent polymers are also used as materials for fragment gel, sweat band, portable toilet, etc.

Others[31–33]

1. Sensor: The swellability of superabsorbent polymer gels, their mechanical modulus and rubbery character, and their sensitivity to changes in water content, pH, and ionic strength make them suitable for use in various sensing systems.
2. Artificial snow: The fake snow can be made by mixing a superabsorbent polymer with up to 100 times its mass of water and then aerating the gel whole freezing it. The artificial snow is maintained with refrigeration system in indoor ski arenas. The frozen gel layer groomed to yield snow with a realistic feel, similar to ''powder'' snow. By using superabsorbent polymer in this way, the air temperature in the building can be at least 10°C higher than when using snow made from only water and it is more comfortable to the skiers.
3. In addition to this, superabsorbent polymers are used as additives in fire extinguishing water, materials for water-swellable toys, paint for prevention of water leakage, for removal of water in oil, etc.

CONCLUSIONS

Superabsorbents are superabsorbent polymers, which are loosely crosslinked hydrophilic polymers and have ionic charges. They absorb water, swell, and retain aqueous solutions up to 100 times their own weight. The water absorbed can hardly be removed even by applying high pressure.

Many types of superabsorbents have been prepared from various materials, by various methods, and in different shapes. They are modified natural hydrophilic polymers such as starch, cellulose, alginic acids, etc. and synthetic hydrophilic polymers such as poly (acrylic acid) and poly(sodium acrylate). Most of them have anionic moieties such as poly(sodium acrylate) and poly(vinyl sulfonate). Now, many types of superabsorbents having not only anionic groups but also cationic groups such as ammonium groups, phosphonium groups, or betaine have been prepared from both natural hydrophilic materials and synthetic hydrophilic polymers. Superabsorbents can be made in a variety of shapes such as powder, granules, fiber, sheet, etc.

It has been clarified that, to have high water-absorption ability, superabsorbent polymers should be essentially low crosslinked hydrophilic polymers having ionic charges. The water-absorption ability of the superabsorbent polymers depends on the hydrophilicity, crosslinking density, charges of the polymers, and the concentration of neutral salts in an aqueous solution. This suggests that much water invade into polymers because of high osmotic pressure between inside and outside of polymer gels in aqueous solutions and the electrostatic repulsive expansion of polymers. The synthesis of new type of superabsorbents, the modification of the superabsorbents synthesized to enhance their absorbency, gel strength, and absorption rate, and the development of new application of the superabsorbent have been investigated by many researchers in recent years.

REFERENCES

1. Masuda, F. What are superabsorbents? In *Superabsorbent Polymers*; Kyoritsu Shuppan: Tokyo, 1988; 1–11 (written in Japanese).
2. The Society of High Polymer Japan Ed., Superabsorbent polymers. In *Kobunshi Sinsozai Binran*; Maruzen: Tokyo, 1989; 228–235, (written in Japanese).
3. Farahani, E.V.; Vera, J.H.; Cooper, D.G.; Weber, M.E. Swelling of ionic gels in electrolyte solution. Ind. Eng. Chem. Res. **1990**, *29*, 554.
4. Nonaka, T.; Yamada, K.; Watanabe, T.; Kurihara, S. Preparation of superabsorbent polymer hydrogels from trialkyl-4-vinylbenzyl phosphonoium chloride-acrylamide-methylenebisacrylamide copolymer and their properties. J. Appl. Polym. Sci. **2000**, *78*, 1883–1884.
5. Nonaka, T.; Watanabe, T.; Kawabata, T.; Kurihara, S. Preparation of thermosensitive and superabsorbent polymer hydrogels from trialkyl-4-vinylbenzyl phosphonium chloride-*N*-isopropylacrylamide-*N*,*N*′-methylenebisacrylamide copolymers and their properties. J. Appl. Polym. Sci. **2001**, *79*, 115–124.
6. The Japan Industry Standards (JIS). Testing Method for Water Absorption Capacity of Super Absorbent Polymers, K7223, 1996.
7. Masuda, F. Properties of superabsorbent polymers. In *Superabsorbent Polymers*; Kyoritsu Shuppan: Tokyo, 1988; 51–80 (written in Japanese).
8. Makita, M.; Tanioku, S. Water-Absorbing Resins. U.S. Patent 4,587,308, May 6, 1986.
9. Flory, P.J. *Principle of Polymer Chemistry*; Cornell University Press: Ithaca, NY, 1953.

10. Tsubakimoto, T.; Shimomura, T.; Irie, Y. Absorbent Articles from Powdered Resins. U.S. Patent 4,666,983, May 19, 1987.
11. Masuda, Y. *Absorbent Polymer Gels, Gel Technology*; Abe, M., Murase, N., Suzuki, T., Eds.; Science Forum: Tokyo, 1997; 18–25.
12. Nakagawa, Y.; Hayashi, H.; Tawaratani, T.; Kourai, H.; Horie, T.; Shibasaki, I. Disinfection of water with quaternary ammonium salts insolubilized on a porous glass surface. Appl. Environ. Microbiol. **1984**, *47*, 513.
13. Nonaka, T.; Uemurra, Y.; Kurihara, S. Preparation of the resins containing quaternary ammonium groups from glycidyl methacrylate-1,4-divinylbenzene copolymer beads and antibacterial activity of the resins. Nippon Kagaku Kaishi **1994**, *12*, 1097–1106 (written in Japanese).
14. Nonaka, T.; Ohtsuka, T.; Kurihara, S. Preparation of the resins containing phosphonium groups from glycidyl methacrylate-1,4-divinylbenzene copolymer beads and antibacterial activity of the resins. Nippon Kagaku Kaishi **1995**, *1995* (7), 529–539 (written in Japanese).
15. Buchholz, F.L. Application of superabsorbent polymers. In *Modern Soperabsorbent Polymer Technology*; Buchholz, F.L., Graham, A.T., Eds.; Wiley-VCH: New York, 1997; 251–272.
16. Nishizawa, K.; Shirase, T.; Mizutani, H. Moisture-Permeable Disposable Diapers. U.S. Patent 4,306,559, December 22, 1981.
17. Harper, H.; Bashaw, R.; Atkins, B. Absorbent Product Containing a Hydrocelloidal Composition. U.S. Patent 3,669,103, June 13, 1972.
18. Kellenberger, S.R. Absorbent Products Containing Hydrogels with Ability to Swell Against Pressure. U.S. Patent 5,147,343, September 15, 1992.
19. Kazanskii, K.S.; Dubrovskii, S.A. Chemistry and physics of agricultural hydrogels. Adv. Polym. Sci. **1992**, *104*, 97–133.
20. DeGouw, A.M.; Prins, J.; Dingermas, L. Package for Food Stuffs, such as Shell Fish, Which While in the Packaged State Will Exude Liquid, and a Packaging Method. Eur. Patent 68,530, January 5, 1983.
21. Tsubakimoto, T.; Shimomura, T.; Kobayashi, H. Water absorbents. Jpn. Kokai Tokkyo Koho **1987**, July 3, 62–149, 335.
22. Suetsugu, M.; Sezaki, E.; Nakazato, T.; Isono, M. Water-absorbing compositions containing polyolefin thermoplastic elastomers. Jpn. Kokai Tokkyo Koho **1994**, June 07, 06–157, 839.
23. Shimomura, T.; Namba, T. Superabsorbent Polymers. In *Superasorbent Polymers, Science and Technology, Symposium Series 573*; Buchholz, F.L., Peppas, N.A., Eds.; American Chemical Society: Washington, DC, 1994; 112–115.
24. Moriyoshi, A.; Fukai, I.; Takeuchi, M. A composite material that solidifies in water. Nature **1990**, *344*, 230–232.
25. Bow, K.E. Electric Cable with Improved Water-Block. Eur. Patent. 24,631, March 11, 1981.
26. Hogari, K.; Ashiya, F. *Superabsorbent Polymres, Science and Technology, Symposium Series 573*; Buchholz, F.L., Peppas, N.A., Eds.; American Chemical Society: Washington, DC, 1994; 128–140.
27. Peppas, N.A. *Hydrogels in Medicine and Pharmacy*; CRC Press, Inc.: Boca Raton, FL, 1987; Vol. 1, 1–180.
28. Bronsted, H.; Kopecek, J. *Polyelectrolyte Gels, Properties, Preparation and Application*; Harland, R.S., Prud'homme, R.K., Eds.; American Chemical Society: Washington, DC, 1992; 285–304.
29. Kendall, J.M.; Maes, D.H.; Figueroa, R. Jr. Improved Skin-Moisturizing Emulsions Containing Water-Absorbent Resin Polymeritate. Eur. Patent 281,395, September 07, 1988.
30. Gueret, J.L.; Contamin, J.C.; Ayache, L. Sheet-Like Materials for the Treatment of Skin or Hair. Eur. Patent 309,309, March 29, 1989.
31. Sawahata, K.; Gong, J.P.; Osada, Y. Soft and wet touch-sensing system made of hydrogel. Macromol. Rapid Commun. **1995**, *16*, 713–716.
32. Bucceri, A. Method for Making Artificial Snow. U.S. Patent 4,742,958, May 10, 1988.
33. Morioka, K.; Nakahigashi, S. Construction plan of Tsudanuma skiing arenas with artificial snow and technology of making artificial snow. Refrigeration **1992**, *67*, 28 (written in Japanese).

Supercritical CO_2-Assisted Surface Coating Injection Molding

Masahiro Ohshima
Department of Chemical Engineering, Kyoto University, Nishikyo-ku, Kyoto, Japan

INTRODUCTION

Carbon dioxide (CO_2) becomes supercritical at pressures and temperatures above 7.38 MPa and 31.1°C. At the supercritical condition, the CO_2 has a higher diffusivity and a lower viscosity than liquids and the surface tension becomes absent, which allows a rapid penetration of molecule into the pores of heterogeneous matrices.[1] Taking advantage of these characteristics together with its nontoxicity, nonflammability, and inexpensiveness, CO_2 has been used in a variety of industries, i.e., food, pharmaceutical, fiber, and chemical and plastic industries. Moreover, the fact that CO_2 is a gas under ambient condition makes its removal from the plastic product possible without using the costly drying and solvent removal processes.[2] Because of this feature, it also has the potential of replacing many environmentally harmful solvents in industry.

A successful application reported in the polymer processing was the microcellular plastic foams,[3,4] where CO_2 was used as a physical foaming agent (PFA) to create microscale size cellular structure in plastics. After creating the cell structure in plastics, the CO_2 diffuses out and does not stay long in the plastics, unlike chemical browning agents. This feature gives some benefits of plastic recycling. In addition to the microcellular foaming, many potential applications of CO_2, such as polymer blending, additive impregnation, and surface modifications, have been investigated for polymer processing purposes.

In the polymer processing field, the usage of supercritical CO_2 ($scCO_2$) offers several exciting possibilities.[5] One major advantage of applying $scCO_2$ to polymer processing is that the processing conditions and the morphology of polymers can be controlled by $scCO_2$. This is because the presence of CO_2 in polymer changes the rheological and thermal properties in both the molten and solid states. For example, when CO_2 dissolves in polymer, the surface tension, shear viscosity, and the glass transition temperature are reduced and the crystallization rate is changed. This is called plasticization effect of CO_2. This effect provides many opportunities of improving the processing condition and quality of plastic products. Recently, by utilizing the plasticization effects of dissolved CO_2, an injection molding process called Asahi Mold Technology with CO_2 (AMOTEC) was invented to reduce surface roughness and prevent the surface defection in solidified plastic products,[6,7] where pressurized CO_2 is introduced in a mold cavity, the CO_2 is dissolved into polymer at the melt front of polymer running in the mold cavity, and the dissolved CO_2 reduces the polymer viscosity at the surface of the plastics owing to the plasticization effect. The viscosity reduction at the plastic surface increases the transferability of the mold shape on the plastics and improves the surface roughness. Extending and twisting the idea of the AMOTEC, a CO_2-assisted surface modification injection-molding technique was developed.[8] It utilizes the plasticization effect of CO_2 on polymer, solubility of low molecular compounds in $scCO_2$ and infusion mechanism of CO_2, and the low molecular compounds mixture into the injected polymer from its surface. By using a dye pigment as a low molecule to be dissolved in $scCO_2$, the surface of the solidified plastic products in the mold cavity can be dyed.[9] In this entry, fundamental properties of CO_2/polymer systems are briefly reviewed with some explanation of recent advancements in property measurements. The principle of the developed CO_2-assisted surface modification injection molding is then explained with a validation experiment.

FUNDAMENTAL PROPERTIES

Solubility of CO_2 in Polymer

Solubility and diffusivity of CO_2 in molten polymer are two fundamental properties for the proposed CO_2-assisted polymer processing. Several researchers have been studying the solubility of gases in polymers in conjunction with polymeric membrane and microcellular foaming. Durill and Griskey studied the solubility of CO_2 and nitrogen (N_2) gases into several polymers such as polypropylene (PP), low density polyethylene (LDPE), high density polyethylene (HDPE), and polystyrene (PS).[10,11] Liu and Prausnitz also investigated the solubility of several gases including CO_2 in PE and in ethylene-vinyl acetate (EVA). Their experiments were carried out in the temperature range from 125°C to 200°C under a fairly low-pressure

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120040332

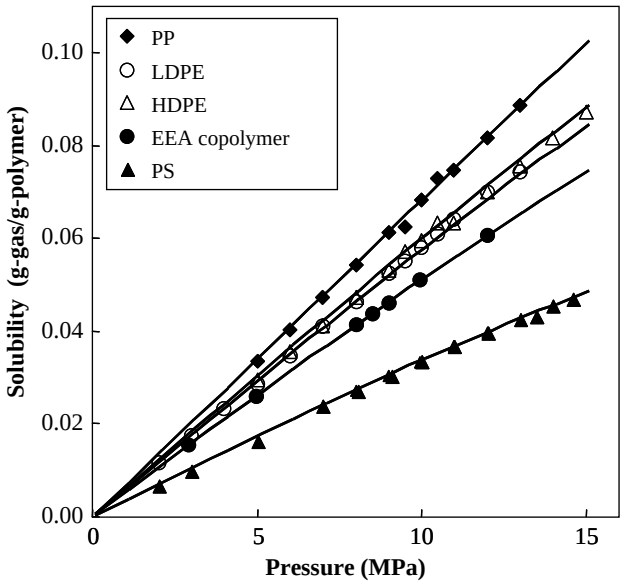


Fig. 1 Solubility of CO_2 in several polymers (at 200°C, a gravimetric method). (From Ref.[27].)

region (approximately 2.5 MPa).[12] Lee and Flumerfelt studied the solubility of N_2 in LDPE at the range from 120°C to 177°C and pressure up to 12.5 MPa.[13] Sato et al. studied the solubility of CO_2 and N_2 in PS, PP, and HDPE at 100°C to 200°C and pressure up to 20 MPa and in PVAc at 40°C to 100°C, pressure up to 17.5 MPa.[14,15,16,17]

Fig. 1 shows the solubility of CO_2 in LDPE, HDPE, PP, ethylene-ethylacrylate copolymer (EEA), and PS, which were measured at temperatures from 150°C to 200°C and pressure up to 12 MPa using a magnetic suspension balance (MSB).[18] As shown in Fig. 1, the solubility of gases in polymers increases as the saturation pressure increases and it follows the Henry's law. Thus, the Henry's constant is used to evaluate the solubility of low molecular gases in molten polymer systems as listed in Table 1. The solubility of CO_2 increases as temperature decreases. The van't Hoff equation is often used to describe temperature dependence of the solubility:

$$H_P(T) = H_O \exp\left(-\frac{E_a}{RT}\right) \quad (1)$$

where E_a is the heat of solution, R is the gas constant, and T denotes temperature. Note that the temperature dependency of the N_2 solubility in polymer shows a different behavior, i.e., the solubility of N_2 in molten polymer increases as temperature increases in a certain temperature range.[15] Furthermore, the Henry's law could not hold when the temperature is below the glass transition temperature due to the fact that the solution mechanism is changed to a combination of adsorption and dissolution.

Table 1 Henry's constant

	150 [Ref.]	180 [Ref.]	200 [Ref.]
PS		4.102[17]	3.200[17]
PS	4.710		3.320
PP		7.645[16]	6.604[16]
PP		7.180[a]	6.847
HDPE	7.313[16]	6.353[16]	5.710[12]
HDPE	6.807		5.885
LDPE	6.620		5.804
LDPE	6.715		5.890
PLA		5.786	5.342
PBS	6.861	5.639	
PC			3.467
EPDM			9.511

Henry constant (g-CO_2/g-polymer/MPa) $\times 10^3$.
[a]185°C.

Diffusivity of CO_2 in Polymer

In addition to the solubility, many researchers have also investigated the diffusivity of gas into polymer and several models have been proposed for estimating the diffusivity of gas in polymer. The most widely accepted concept for developing the models is based on the free volume. The free volume concept for diffusion has an origin at the theoretical work of Cohen–Turnbull, where the thermodynamic diffusion coefficient of the solute was given by an exponential function of inversed free volume fraction. Recently, Thran et al. analyzed the correlation between the free volume fraction and diffusion for six gases, O_2, N_2, CO_2, CH_4, He, and H_2, and 71 polymers.[19] They derived the exponential relationship between the specific free volume and the diffusion coefficient and reported empirical constants of the exponential correlation model. Extending the free volume model, Areerat et al. proposed a model for estimating the diffusivity of CO_2 in polymer,[18,27] where pressure, temperature, and concentration dependences of diffusivity were described by

$$D_{\text{mutual}} = \frac{x_2 D_{\text{self}}}{RT}\left(\frac{\partial \mu_1}{\partial \ln x_1}\right)_{T,P} = A\left(\frac{\partial \mu_1}{\partial \ln x_1}\right)_{T,P} x_2 \exp\left[\frac{-B}{(\hat{V}_{\text{mix}} - \hat{V}^0_{\text{mix}})}\right] \quad (2)$$

$$\hat{V}^0_{\text{mix}} = (1 - m_{CO_2})\hat{V}^0_{\text{poly}} + m_{CO_2}\hat{V}^0_{CO_2} \quad (3)$$

where A and B are parameters; x_2 is the mole fraction of polymer; $\hat{V}_{\text{mix}}$ is the specific volume of polymer/CO_2

mixture at a given temperature and pressure and $\hat{V}^0_{mix}$ is the occupied specific volume of the mixture at the absolute zero temperature, which is estimated by Eq. (3). $\hat{V}^0_{poly}$ and $\hat{V}^0_{CO_2}$ are the occupied specific volume of polymer alone and pure gas, respectively. m_{CO_2} is the weight fraction of CO_2 in the solution. $\hat{V}_{mix} - \hat{V}^0_{mix}$ is an estimate of the free volume at the given temperature, pressure, and CO_2 concentration. They proposed a scheme of calculating the $\hat{V}_{mix}$ from the swelling of polymer/CO_2 mixture, which could be estimated using an equation of state from the solubility data, pressure–volume–temperature (PVT) data of polymer alone.

Fig. 2 shows a temporal change in the weight of four different polymers, PP, LDPE, PS, and ethylene-propylene-diene rubber (EPDM) during sorption measurement of CO_2 at 200°C. Those data were acquired using a gravimetric measurement scheme, i.e., the magnetic suspension balance (MSB), which was the same apparatus used for solubility measurement. When the diffusivity of CO_2 is measured, CO_2 pressure in the MSB cell is increased in a stepwise manner 1.0 MPa from a saturated state, the temporal change in weight, as shown in Fig. 2, was obtained and the diffusion coefficients were then determined assuming the diffusion process follows the Fick's second law of diffusion. Fig. 3 shows the diffusivity of CO_2 in PS at different temperatures and its estimates.

The diffusivity is not only a function of temperature and pressure, but also concentration of dissolved CO_2. The diffusivity increases as the temperature increases and decreases as the pressure increases. When CO_2 dissolves, the polymer swells and its free volume increases. The diffusion coefficient then becomes larger as the dissolved CO_2 concentration becomes higher. Although the diffusivity does not change drastically with the concentration at temperature high enough to keep the polymer molten, the effect of dissolved CO_2 concentration on diffusivity becomes prominent around the glass transition temperature. Koga et al. investigated the CO_2 sorption into poly(methyl methacrylate) (PMMA) as well as PS polymers using in situ neutron reflectivity.[20] They observed that both PMMA and PS swell by approximately 30%, and at the same time, they observed an anomalous diffusion of CO_2 in these polymers at temperatures below their original glass transition temperature.

Diffusivity of $scCO_2$ and Low Molecular Compounds Mixture in Polymer

There are several papers discussing the diffusivity of one single penetrant in polymer. However, few papers could be found for measuring diffusivity of $scCO_2$ and low molecular compounds mixture into polymer, i.e., diffusivity of multicomponents in $scCO_2$, which is the key factor for determining the operating condition of the proposed surface coating injection molding techniques. In order to analyze the diffusion behavior of $scCO_2$ and low molecule mixture into polymer, an online measurement system using near infrared (NIR) spectroscopy was developed.[21]

Fig. 4 shows a schematic diagram of the NIR measurement system. The fiber optic probes of on-line Fourier transform NIR unit (FIR1000L, Yokogawa Electric CO.) were equipped with a high-pressure

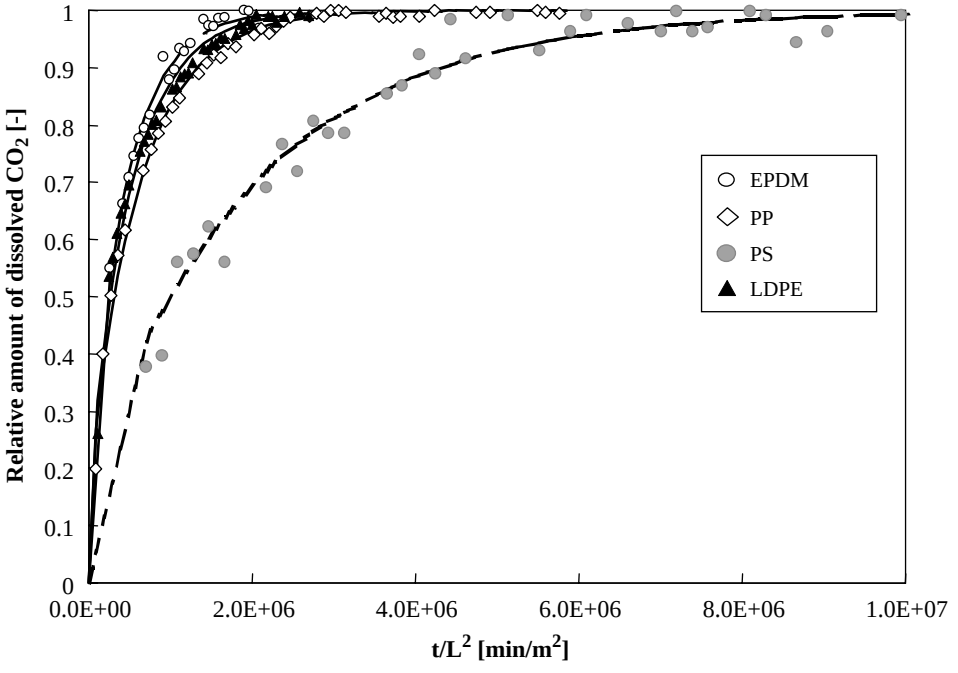


Fig. 2 Diffusivity of CO_2 in polymers (temperature at 200°C and pressure step 10–11 MPa for EPDM, PP, PS, 10.5–11 MPa for LDPE. L is the thickness of the sample).

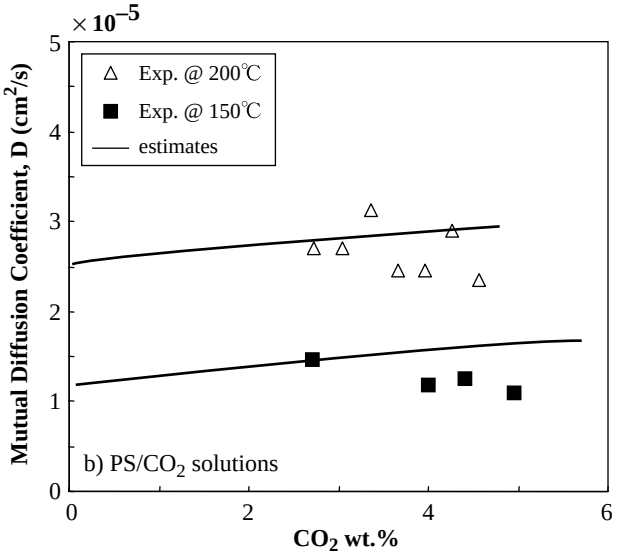


Fig. 3 Mutual diffusion coefficients of CO_2 in PS (experiments and estimates). (From Ref.[27].)

autoclave. The sample cell of the autoclave takes a cylindrical shape, in which a cylindrical sample polymer was charged and melted in the cell by increasing the temperature, so as to leave no space between the sample and the wall, while making an open space over the sample polymer. CO_2 alone, or a mixture of CO_2 and low molecule, was introduced to the open space and diffused into the molten polymer. The NIR light was transmitted through the molten polymer at a certain distance from the interface between the polymer and CO_2 or mixture. When some NIR light passed through a layer of solution having a concentration of chemical species, the power of the light is absorbed and attenuated by

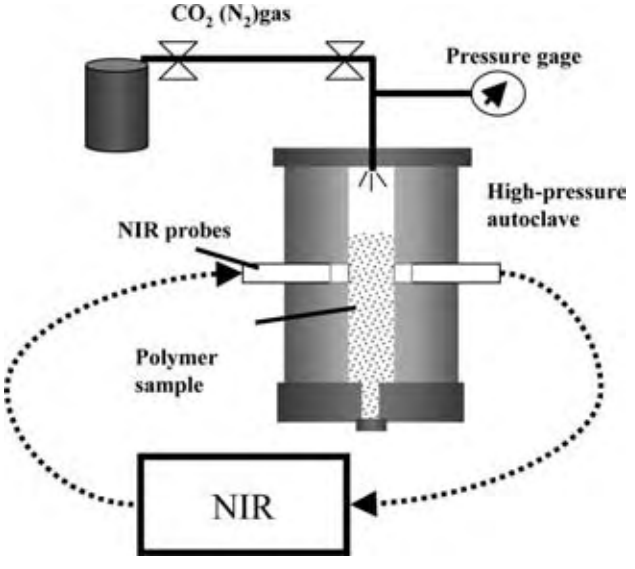


Fig. 4 NIR measurement system.

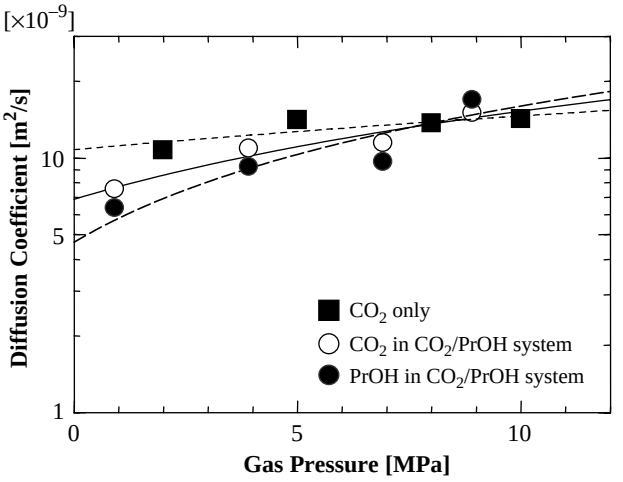


Fig. 5 Diffusion coefficients of CO_2 in LDPE measure by NIR at 175°C.

the chemical species due to the interaction between the photons and absorbing chemical bonds. The wavelength at which adsorption occurs depends on the species of chemical bonds. For example, CO_2 can be detected at 2019 nm. The absorbance, which is defined by logarithmic fraction of the light transmitted by the solution, is directly proportional to the path length through the solution and the concentration of the absorbing chemical bond in the solution. This is called the Lambert–Beer law. Nagata et al. clarified that the Lambert–Beer law could be established at the polymer CO_2 systems in the transmitted NIR measurements.[21]

Applying the law, the change in concentration of CO_2 and low molecular compounds in molten polymer could be monitored by measuring the NIR absorbance on-line at the autoclave. Fig. 5 shows the NIR measured diffusion coefficients of CO_2 alone, and those of CO_2 as well as propanol, of their mixture to LDPE (MI = 8.0 g/10 min; Mw = 1.05×10^5 g/mol; Mw/Mn = 6.94) at 175°C by changing the pressure from 0.1 MPa to a specified level. As can be seen in Fig. 5, both propanol (PrOH) and CO_2 of mixture could diffuse into the polymer even under the critical pressure of CO_2, i.e., 7.38 MPa. Since the propanol is volatile, it mixes with CO_2 in gas phase at 175°C. Therefore, the propanol could diffuse into polymer even at the low pressure. It could be worthwhile to note that the diffusivity of propanol was lower than that of CO_2 at low-pressure level, but it increased close to that of CO_2 over the critical pressure.

Nonvolatile molecules, such as dye pigments, cannot be infused into polymer without using $scCO_2$. Fig. 6 shows the change in absorbance at 1493 nm when CO_2 and dye pigment (disperse blue) mixture was introduced to the autoclave and diffused into the LDPE at 175°C by changing the CO_2 pressure from

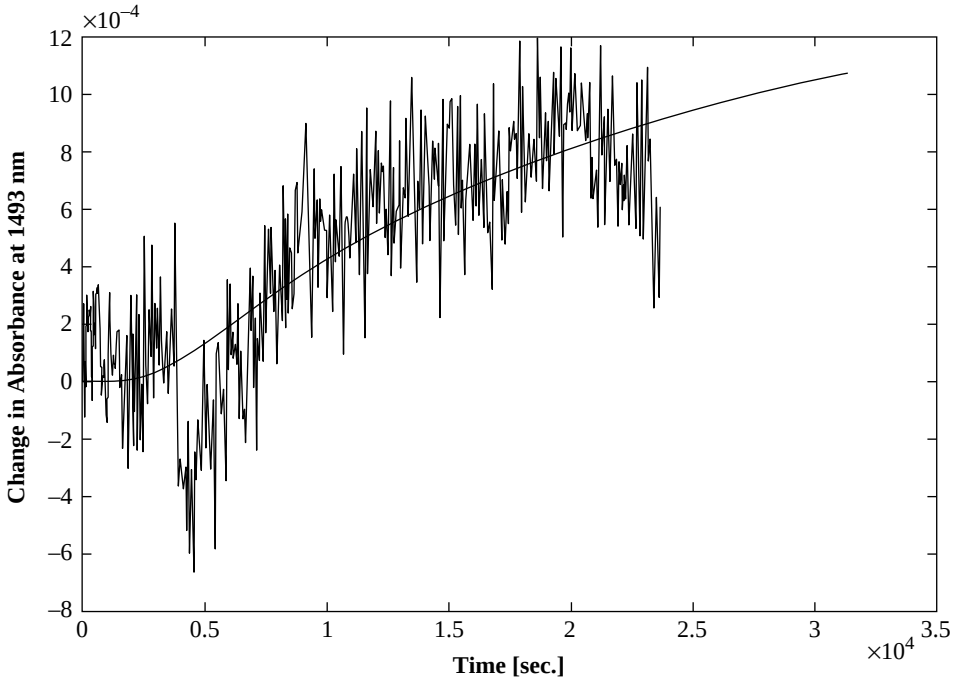


Fig. 6 Diffusion behavior disperse blue with CO_2 into LDPE measured by NIR at 175°C. *(View this art in color at www.dekker.com.)*

0.1 to 10 MPa. Although the signal over noise ratio was low, it was clearly observed that the blue pigment was diffusing into the polymer. As shown by the experimental data, nonvolatile molecules could be infused into polymer using $scCO_2$ as far as the molecule could be dissolved into $scCO_2$.

CO_2 Induced Viscosity Reduction

The other factor of controlling the proposed injection molding is the plasticization effect of CO_2, i.e., polymer viscosity reduction by CO_2 dissolution. During the last decade, the study of CO_2 induced viscosity reduction has been advanced drastically. Gerhardt et al. measured viscosity reduction of polydimethylsiloxane (PDMS)/CO_2 solutions by using a high-pressure plunger-type capillary rheometer.[22] Kwag et al. also measured the viscosity of PS/CO_2 solutions by employing the same technique as Gerhardt.[23] Lee et al. measured the viscosity of PS/CO_2 solutions.[24] Gendron et al. measured the viscosity of PS/hydrofluorocarbon (HFC) solutions, as well as PS/CO_2 solutions, and developed a viscosity prediction model for these mixtures by employing the William–Landle–Ferry (WLF) equation with a variable glass transition temperature, T_g.[25] They calculated T_g as a function of dissolved gas concentration by extending the Chow model. Royer et al. modified Gendron's model so that the pressure effect could be taken into account when the viscosity of PS/CO_2 mixtures was predicted.[26] Areerat et al. proposed the following practical model for predicting the viscosity of the polymer/CO_2 solutions.[27,28] The model was developed by integrating the Cross–Carreau model with Doolittle's equation by means of a variable free volume. The Cross–Carreau model was given by

$$\eta(\dot{\gamma},) = \frac{\eta_O}{\left(1+\left(\frac{\eta_O\dot{\gamma}}{\tau}\right)^a\right)^{\frac{(1-n)}{a}}} \tag{4}$$

where η and η_O are the shear and zero-shear viscosities, respectively; $\dot{\gamma}$ is the shear rate; and n, τ, and a are parameters which determine the shape of the viscosity-shear rate curve.

At high shear rates where $\left(\frac{\eta_O\dot{\gamma}}{\tau}\right) \gg 1$ can hold, Eq. (4) can be simplified into

$$\eta(\dot{\gamma}) \cong \eta_o^n\left(\frac{\dot{\gamma}}{\tau}\right)^{(n-1)} \tag{5}$$

It is assumed that the parameters n, τ, and a are constant and the viscosity reduction due to the CO_2 dissolution could be described only by the changes in zero-shear viscosity. The zero-shear viscosity, η_O, is given as a function of free volume fraction as described by

$$\eta_o = A\exp\left(\frac{B}{f(T,P,w_g)}\right) \tag{6}$$

where A and B are unique constant parameters for the polymer and they are determined when the polymer is given. $f(T,P,w_g)$ denotes the free volume fraction. Then, the free volume fraction is assumed to be a function of temperature, T, pressure, P, and weight percentage of dissolved CO_2, w_g.

Substituting Eq. (6) into Eq. (5) results in

$$\eta = \exp\left(C_3 + \frac{C_2}{f(T,P,w_g)}\right)\dot{\gamma}^{C_1} \quad (7)$$

where

$$C_1 = (n - 1)$$
$$C_2 = nB$$
$$C_3 = [n \ln A - (n - 1)\ln \tau]$$

Assuming that C_1, C_2, and C_3 are not affected by temperature, pressure, and dissolution of CO_2, they can be determined from a viscosity-shear rate curve of the neat polymer. Namely, the coefficient, C_1, which is equivalent to $n - 1$, can be determined by the slope of the viscosity and shear rate curve. The values of C_2 and C_3 can be determined from data of viscosity vs. free volume fraction of the neat polymer. The data of free volume fraction required for determining C_2 and C_3 can be obtained from PVT data of the neat polymer at temperatures and pressures where the viscosity measurements of the neat polymer are performed.

To predict the viscosity reduction, the change in the free volume fraction caused by CO_2 dissolution has to be calculated. Extending the definition of free volume fraction, the free volume fraction is given as a function of temperature, pressure, and dissolution of gas:

$$\begin{aligned} f(T,P,w_g) &= f_r + \left(\frac{\partial f}{\partial T}\right)_{P_r,T_r,w_{gr}} (T - T_r) \\ &\quad + \left(\frac{\partial f}{\partial P}\right)_{P_r,T_r,w_{gr}} (P - P_r) \\ &\quad + \left(\frac{\partial f}{\partial w_g}\right)_{P_r,T_r,w_{gr}} (w_g - w_{gr}) \quad (8) \\ &= f_r + (1 - f_r)\alpha(T - T_r) \\ &\quad - (1 - f_r)\beta(P - P_r) + \phi(w_g - w_{gr}) \end{aligned}$$

where f_r is a reference free volume fraction given at a reference temperature, T_r, pressure, P_r and weight percentage of dissolved CO_2, $w_g r$. $\alpha(:= \frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_{P_r,w_{gr}})$ is the thermal expansion coefficient, $\beta(:= -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_{P_r,w_{gr}})$ is the isothermal compressibility coefficient and ϕ is the gas expansion coefficient.

Since zero-CO_2 concentration is taken as a reference condition, the parameter f_r is not a function of CO_2. Then, f_r, α, and β can be determined from the PVT measurement of the neat polymer. There remains only gas concentration coefficient, φ, as a variable affected by CO_2 dissolution. The gas expansion coefficient, φ, is determined by solubility measurements, the models with these parameter values could predict the viscosity of polymer/CO_2 single-phase mixtures.[27]

Fig. 7 shows the CO_2 induced viscosity reductions of PP and PS polymers. The viscosities of the polymer/CO_2 mixtures were measured by a capillary rheometer equipped with a foaming extruder. As can be seen in Fig. 7, the viscosity of PS was reduced by 40% by dissolving CO_2 3.5 wt% and that of PP was reduced by 25%. The solid lines in Fig. 7 represent the estimates of the aforementioned models.

CO_2-ASSISTED SURFACE MODIFICATION INJECTION MOLDING

As shown in the previous section, CO_2 enhances the mobility of polymer and plasticizes the polymers. Furthermore, it can dissolve a variety of low molecular compounds, such as dye pigments, additive agents, monomer, and crystal nucleation agents, when it becomes a supercritical state. Integrating these characteristics of CO_2 can create an scCO_2-assisted surface coating injection molding technique. Two different injection schemes of coating the plastic surface are described in the following sections.

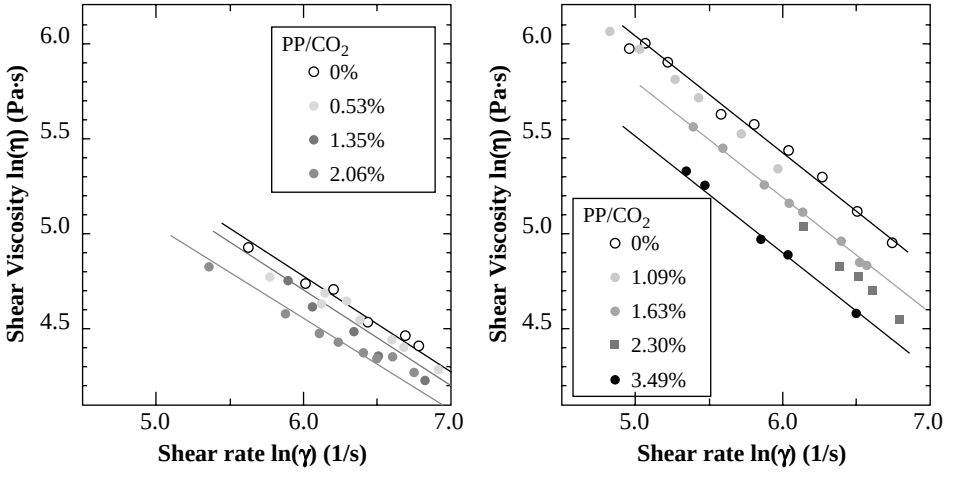


Fig. 7 CO_2-induced viscosity reduction at PS/CO_2 and PS/CO_2 systems at 200°C.

Direct Injection Scheme

One of the $scCO_2$-assisted surface coating schemes is illustrated in Fig. 8. $scCO_2$ is introduced into a mold cavity of an injection molder. On the line of introducing the $scCO_2$ from a high pressure CO_2 generator to the cavity, low molecular solutes, such as dye pigments, alcohols, and monomers, low molecular weight polymers are dissolved into the $scCO_2$. The cavity is pressurized and filled with the $scCO_2$ dissolving the substance (Fig. 8A). Then, the molten polymer is injected into the pressurized cavity (Fig. 8B). When the polymer is flowing in the cavity, the CO_2 and the low molecular solutes diffuse from the melt front of the flowing polymer. In the mold cavity, the injected polymer shows a so-called fountain flow. As shown in Fig. 9, the polymer at the melt front moves from the centerline of the stream to the cavity wall when it runs in the cavity. The polymer surface, where CO_2 as well as low molecular solute are dissolved, is carried toward the cavity walls on the fountain flow and is solidified at the wall, which normally remains below the transition temperature (Fig. 8C). The viscosity of the polymer surface containing the CO_2 is reduced owing to the plasticization effect. The reduced viscosity increases the transferability of polymer against the mold shape and reduces surface roughness. Furthermore, the dissolved CO_2 in polymer eventually diffuses out to atmosphere while low molecular substance remains in the products. This produces solidified plastic products whose surface is modified by the low molecular substances.

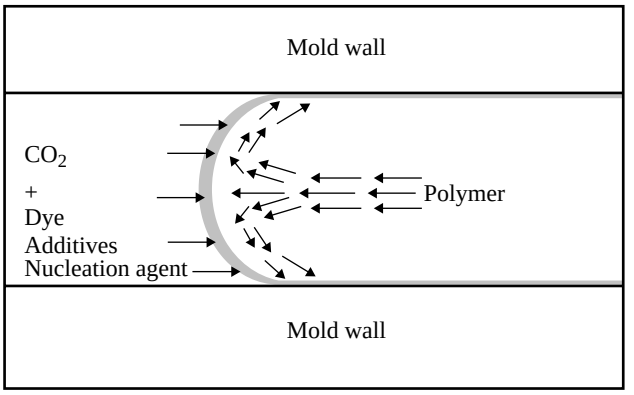


Fig. 9 Fountain flow behaviors of polymer and diffusion of CO_2 and additives into polymer melt front in a mold cavity. (From Ref.[9].)

Core-Back Molding Scheme

In the aforementioned direct injection scheme, the contact time, which is a period that the polymer front contacts $scCO_2$, becomes a critical factor of determining the thickness of modified polymer layer. The contact time, which is almost equivalent to the injection period at the method, cannot be made longer due to

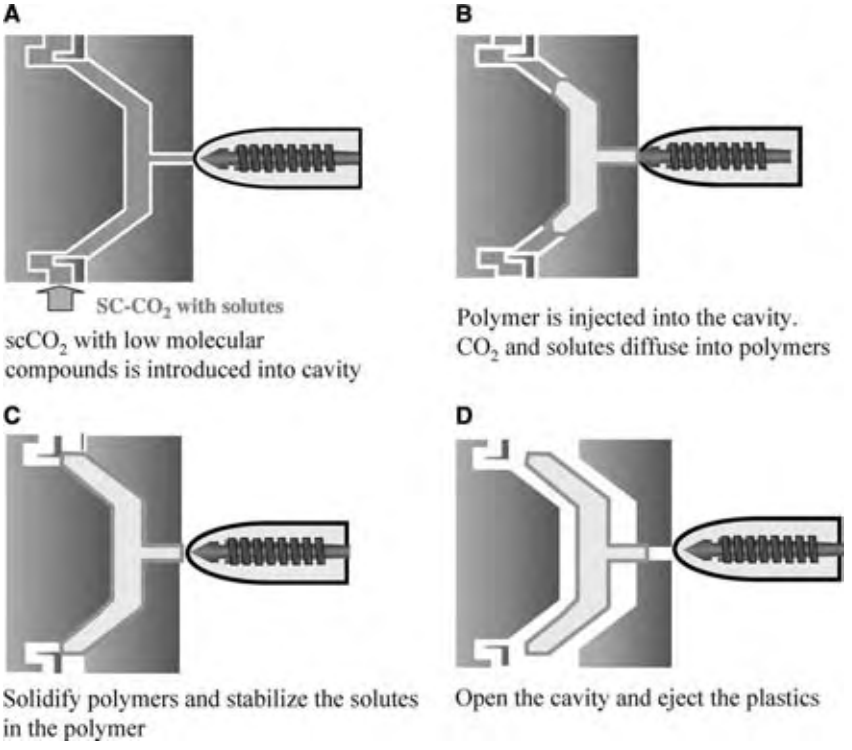


Fig. 8 Direct injection scheme of surface coating with $scCO_2$. (From Ref.[9].)

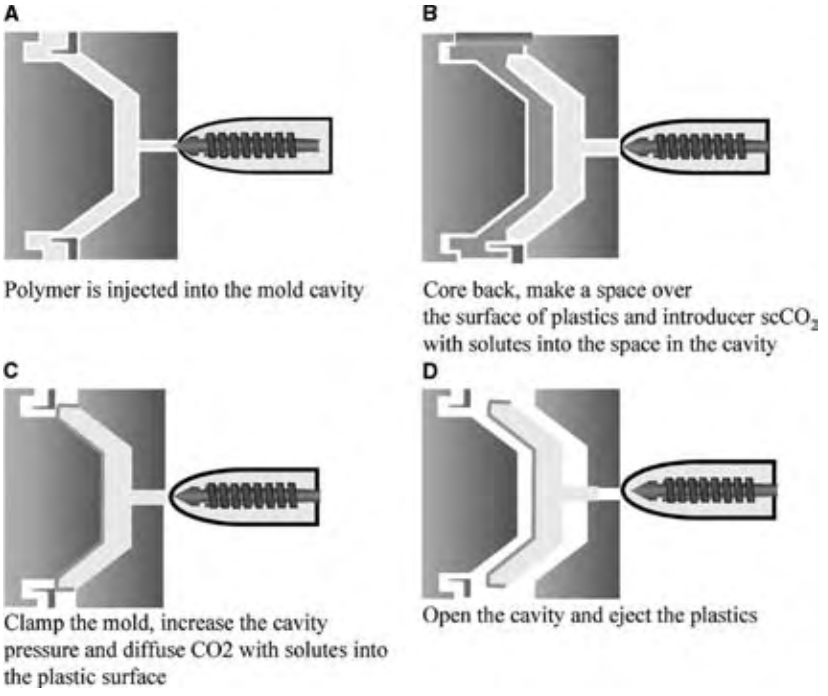


Fig. 10 Core-back injection molding of surface coating with $scCO_2$. (From Ref.[9].)

the fact that the lower injection speed reduces processability of the polymer. In order to manipulate the contact time of CO_2, solutes to polymer, and control the thickness of the polymer layer modified by a low molecular solute, the other scheme, which is named the core-back molding scheme, is developed as shown in Fig. 10. The polymer is injected into the cavity (Fig. 10A). At a certain moment in time, part of the mold is shifted back so as to make a narrow space while keeping the pressure sealing, and the $scCO_2$ containing the low molecular substance is introduced into the space (Fig. 10B). Then, part of the mold is clamped back and the pressure of CO_2 in the space is increased so as to dissolve both the CO_2 and the low molecular substance in the plastic in the mold (Fig. 10C). By varying the volume of the space and the pressurizing time, the thickness of the modifying layer can be changed.

EXPERIMENTAL RESULTS

Fig. 11 shows the experimental setup of the CO_2-assisted surface coating injection molding. A 35-ton injection molder was used. The molder cavity was sealed so that the cavity pressure can be kept at a higher value than critical pressure 7.38 MPa. The high pressure CO_2 was generated, by pumping the liquid CO_2, and introduced into a buffer tank. A low molecular compound to be dissolved in $scCO_2$, which

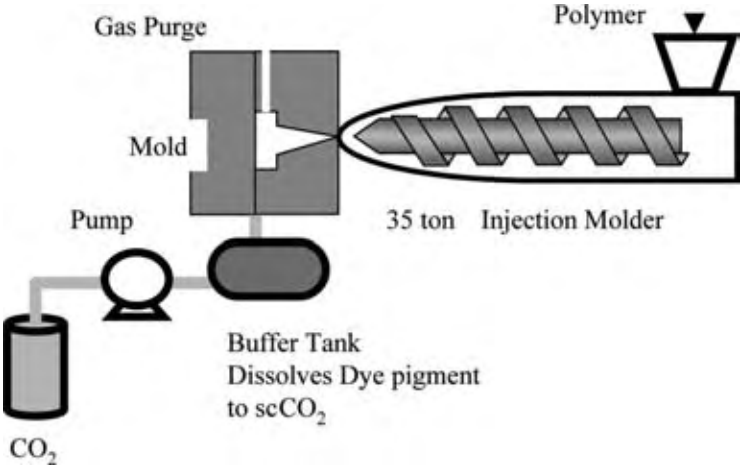


Fig. 11 Experimental setup for a surface coating injection molding. (From Ref.[9].)

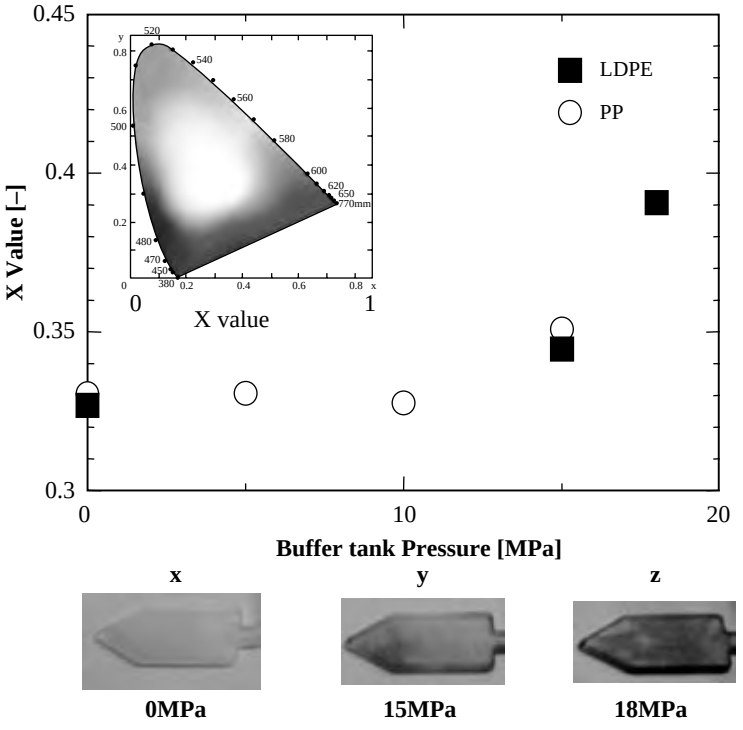


Fig. 12 The darkness of the red-colored surface of plastics. *(View this art in color at www.dekker.com.)*

in this experiment was Oil Red-O dye pigment, was placed inside the buffer tank. By keeping the buffer tank temperature higher than critical temperature and pressure of CO_2, the low molecular compound is dissolved in $scCO_2$ in the tank. A gas line for supplying $scCO_2$ and low molecule mixture is connected from the buffer tank to the mold.

The direct injection scheme of surface coating was performed by this experimental setup. Low density polyethylene (LDPE) and polypropylene (PP) were used as sample polymers and a disperse dye pigment, Oil Red-O, whose melting point is 120°C, was used as a low molecule to be dissolved in $scCO_2$ and coat the surface of LDPE and PP of injection products. The buffer tank temperature was kept at 60°C and the molder barrel temperature was 220°C, which is a normal temperature setup. By varying $scCO_2$ pressure at the buffer tank, the direct injection scheme was performed. The color of injected plastic product was evaluated using the x-y-z color system as shown in Fig. 12, which also shows a result of experiments. The X-value indicates the degree of red color. As can be seen, the red color of the plastic products' surface becomes darker as the pressure of $scCO_2$ increases over critical temperature.

CONCLUSIONS

In this entry, the fundamental aspects of CO_2/polymer systems, such as solubility, diffusivity of CO_2 in polymers, and CO_2 induced viscosity reduction, as well as the basic principles of supercritical CO_2 technique in polymer processing, were discussed. It was shown that utilization of these fundamental aspects of $scCO_2$ and polymer systems is one of the promising schemes and has great potential for inventing a new polymer processing technology. As an example, a novel CO_2-assisted surface coating injection-molding technique was introduced, which utilizes the plasticization and the solvation effects of $scCO_2$ for processing several plastic products. Two supercritical CO_2-assisted surface coating injection-molding schemes were proposed. By the developed technique, the surface of plastic products could be modified with low molecular substances and the new functionality can be added to plastic products.

REFERENCES

1. Kajimoto, O. Solution structure in supercritical fluids. In *Supercritical Fluids*; Arai, Y., Sako, T., Takebayashi, Y., Eds.; Materials Processing, Springer: Berlin, 2001; 1–65.
2. Adshiri, T.; Arai, K.; Kitamura, M.; Masuoka, H.; Sako, Y.; Takishima, S. Material processing using supercritical fluids. In *Supercritical Fluids*; Arai, Y., Sako, T., Takebayashi, Y., Eds.; Materials Processing, Springer, 2001; 280–345.
3. Okamoto, K.T. *Microcellular Processing*; Hanser: Munich, 2003.

4. Baldwin, D.F.; Park, C.B.; Suh, N.P. A microcellular processing study of poly(ethylene terephthalate) in the amorphous and semi-crystalline states. Polym. Eng. Sci. **1996**, *35* (4), 1446–1453.
5. Kazarian, S.G. Polymer processing with supercritical fluids. Polym. Sci. Ser. **2000**, *42* (1), 78–101.
6. Yamaki, H.; Matsuura, Y. Method for Injection-Molding of Amorphous Resin JP Patent. 1-245252 A, 1999.
7. Yasuda, K.; Yamaki, H. Injection molding technology with CO_2 as a plasticize of resin. Proc. of international workshop of foam processing and supercritical fluid aided polymer processing. Tokyo, 2003; 113–116.
8. Ohshima, M.; Yasuda, K. Surface Modifying Injection Molding Method, JP Patent 2003-320556, 2003.
9. Ohshima, M. Supercritical CO_2-assisted surface coating injection molding. Proceeding of ANTEC, Chicago, 2004; CD-ROM.
10. Durill, P.L.; Griskey, R.G. Diffusion and solution of gases in thermally softened or moten polymers; Part I. Development of technique and determination of data. AIChE J. **1966**, *12*, 1147–1151.
11. Durill, P.L.; Griskey, R.G. Diffusion and solution of gases in thermally softened or molten polymers; Part II. Relation of diffusivities and solubilities with temperature pressure and structural characteristics. AIChE J. **1969**, *15*, 106.
12. Liu, D.; Prausnitz, J.M. Solubility of gases and volatile liquids in polyethylene and in ethylene-vinyl acetate copolymers in the region 124–225°C. Ind. Eng. Chem. Fundam. **1976**, *15* (4), 330–335.
13. Lee, J.G.; Flumerfelt, R.W. Nitrogen solubilities in low-density polyethylene at high temperatures and high pressures. J. Appl. Polym. Sci. **1995**, *58* (12), 2213–2219.
14. Sato, Y.; Wang, M.; Takishima, S.; Masuoka, M.; Watanabe, T.; Fukasawa, Y. Solubility of butane and isobutane in molten polypropylene and polystyrene. Proceedings of 15th International Congress of Chemical and Process Engineering, Praha, Czech, Aug. 2002.
15. Sato, Y.; Yurugi, M.; Fujiwara, K.; Takishima, S.; Masuoka, H. Solubilities of carbon dioxide and nitrogen in polystyrene under high temperature and pressure. Fluid Phase Equilib. **1996**, *125*, 129–138.
16. Sato, Y.; Fujiwara, K.; Takikawa, T.; Sumarno; Takishima, S.; Masuoka, H. Solubilities and diffusion coefficients of carbon dioxide and nitrogen in polypropylene, high-density polyethylene, and polystyrene under high pressures and temperatures. Fluid Phase Equilib. **1999**, *162*, 261–276.
17. Sato, Y.; Takikawa, T.; Takishima, S.; Masuoka, H. Solubility and diffusion coefficients of carbon dioxide in poly(vinyl acetate) and polystyrene. Journal of Supercritical Fluids **2001**, *19*, 187–198.
18. Areerat, S.; Hayata, Y.; Katsumoto, R.; Kegasawa, T.; Egami, H.; Ohshima, M. Solubility of carbon dioxide in polyethylene/titanium dioxide composite under high pressure and temperature. J. Appl. Polym. Sci. **2002**, *86* (2), 282–288.
19. Thran, A.; Kroll, G.; Faupel, F. Correlation between fractional free volume and diffusivity of gas molecules in glassy polymers. J. Polym. Sci. Part B: Polym. Phys. **1999**, *37* (23), 3344–3358.
20. Koga, T.; Seo, Y.E.; Hu, X.; Shin, K.; Zhang, Y.; Rafailovich, M.H.; Sokolov, J.C.; Chu, B.; Satija, S.K. Dynamics of polymer thin films in supercritical carbon dioxide. Euro Phys Lett. **2002**, *60* (4), 559–565.
21. Nagata, T.; Tanigaki, M.; Ohshima, M. On-line NIR sensing of CO_2 Concentration for polymer extrusion foaming processes. Polym. Eng. Sci. **2000**, *40* (8), 1843–1849.
22. Gerhardt, L.J.; Manke, C.W.; Gulari, E. Rheology of polydimethylsiloxane swollen with supercritical carbon dioxide. J. Polym. Sci. Part B: Polym. Phys. **1997**, *35* (3), 523–534.
23. Kwag, C.; Manke, C.W.; Gulari, E. Rheology of molten polystyrene with dissolved supercritical and near-critical gases. J. Polym. Sci. Part B: Polym. Phys. **1999**, *37* (19), 2771–2781.
24. Lee, M.; Park, C.B.; Tzoganakis, T. Measurements and modeling of PS/supercritical CO_2 solution viscosity. Polym. Eng. Sci. **1999**, *39* (1), 99–109.
25. Gendron, R.; Daigneault, L. *Rheology of Thermoplastic Foam Extrusion Process In Foam Extrusion*; Lee, S.-T., Ed.; Technomic: Pennsylvania, 2000; 35–80.
26. Royer, J.R.; Gay, Y.J.; Desimone, J.M.; Khan, S.A. High-pressure rheology of polystyrene melts plasticized with CO_2: experimental measurement and predictive scaling relationships. J. Polym. Sci. Part B: Polym. Phys. **2000**, *38* (23), 3168–3180.
27. Areerat, S.; Funami, E.; Hayata, Y.; Nakagawa, D.; Ohshima, M. Measurement and prediction of diffusion coefficients of supercritical CO_2 in molten polymers. Polym. Eng. Sci. **2004**, *44* (10), 1915–1924.
28. Nagata, T.; Areerat, S.; Ohshima, M.; Tanigaki, M. Measurement and prediction of CO_2-induced viscosity reduction of polypropylene. Kagaku Kogaku Ronbunshu **2002**, *28* (6), 739–745.

Supercritical Fluid Extraction (SFE)

Ram B. Gupta
Department of Chemical Engineering, Auburn University, Auburn, Alabama, U.S.A.

INTRODUCTION

A fluid is supercritical when it is compressed beyond its critical pressure (P_c) and heated beyond its critical temperature (T_c). Supercritical fluid technology has emerged as an important technique for supercritical fluid extraction (SFE). In many of the industrial applications, it has replaced conventional solvent-based or steam extraction processes, mainly due to the quality and the purity of the final product and environmental benefits.

SUPERCRITICAL FLUIDS AND THEIR PROPERTIES

There are several fluids of interest as listed in Table 1.

In this list, carbon dioxide is the supercritical fluid of choice because it is nonflammable, nontoxic, inexpensive, and has mild critical temperature. Hence, much of the attention has been given to supercritical carbon dioxide for practical extraction applications.

No amount of compression can liquefy the supercritical fluid. In fact pressure can be used to continuously change the density from gas-like conditions to liquid-like conditions. Near the critical region, small changes in the pressure can give rise to large changes in the density. Fig. 1 shows how density of carbon dioxide is varied by pressure at different isotherms.

In addition to density, diffusivity of the supercritical fluids is higher than that of liquid solvents, and can be easily varied. For typical conditions, diffusivity in supercritical fluids is of the order of $10^{-3}\,cm^2/sec$ as compared to 10^{-1} for gases and 10^{-5} for liquids. Typical viscosity of supercritical fluids is of the order of $10^{-4}\,g/cm/sec$, similar to that of gases, and about 100-fold lower than that of liquids. High diffusivity and low viscosity provide rapid equilibration of the fluid to the mixture to be extracted, hence extraction can be achieved close to the thermodynamic limits. However, the main extraction benefit of supercritical fluids is their adjustable density that provides adjustable solvent strength. The compounds of choice can be dissolved/extracted in the supercritical fluid at high pressure and then this fluid mixture is carried to another vessel where simple lowering of the pressure precipitates the compound. A simple extraction scheme is shown in Fig. 2.

SUPERCRITICAL FLUID EXTRACTION SCHEMES

Most of the industrial SFE is for the extraction from a solid matrix. Because of the challenge in the pumping of the solids to high-pressure extractor, the process is usually carried out in a semibatch mode. A number of high-pressure extractors are used in parallel; the solid mixture is filled, extracted, and emptied in a sequential fashion, allowing almost a continuous stream of the extract product.

The lowering of pressure to precipitate the extract is not always necessary. For example, in the case of decaffeination of coffee beans, water can be used to extract caffeine from the CO_2/caffeine mixture, as caffeine is readily soluble in water (Fig. 3).

Additional benefits of the supercritical carbon dioxide extraction are: 1) oxygen-free system prevents oxidation of the extract; 2) low temperature minimizes the thermal degradation; 3) microbes or their spores are not soluble, hence aseptic extracts are obtained; and 4) solvent-free extract is obtained because CO_2 is gas at ambient and is not retained by the extract.

When designing SFE, the most important factor to consider is the solubility of the desired compound in supercritical carbon dioxide. Because of the benign (noninteracting) molecular nature of carbon dioxide, solubilities are usually small. Nonetheless, with the recycle of CO_2, multiple passes can be used to achieve the desired extraction. For illustration, solubilites of two compounds in orange skin, linalool, and limonene are shown in Figs. 4 and 5.[1] The solubility is highly dependent on the pressure and temperature of the supercritical carbon dioxide.

Solubility increases with the pressure owing to increase in the solvation power of the supercritical fluid. An increase in the temperature causes a drop in the density of the fluid (hence the solvation power); however, the substance itself is more volatile at the increased temperature. For the SFE, density, rather than pressure or temperature, appears to be the more natural variable for describing the solubilities. The solubility data in Fig. 4 are replotted vs. density in Fig. 5.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007975

Table 1 Critical constants and safety data for various supercritical solvents

Supercritical fluid	T_c (°C)	P_c (atm)	Safety hazard
Ethylene	9.3	49.7	Flammable gas
Trifluoromethane (fluoroform)	25.9	46.9	
Chlorotrifluoromethane	28.9	38.7	—
Ethane	32.3	48.2	Flammable gas
Carbon dioxide	31.1	72.8	—
Dinitrogen monoxide (laughing gas)	36.5	71.7	Not combustible but enhances combustion of other substances
Sulfur hexafluoride	45.5	37.1	—
Chlorodifluoromethane (HCFC 22; R 22)	96.4	48.5	Combustible under specific conditions
Propane	96.8	42.4	Extremely flammable
Ammonia	132.4	111.3	Flammable and toxic
Dimethyl ether (wood ether)	126.8	51.7	Extremely flammable
Trichlorofluoromethane (CFC 11; R 11)	198.0	43.5	—
Isopropanol	235.2	47.0	Highly flammable
Cyclohexane	280.3	40.2	Highly flammable
Toluene	318.6	40.6	Highly flammable
Water	374.0	217.7	—

SOLUBILITY ENHANCEMENT USING COSOLVENTS

The problem of the poor solubility of the extractant in supercritical carbon dioxide, can be overcome by the addition of a small amount of the cosolvent, such as methanol, ethanol, or acetone.[2] The carbon dioxide by itself is almost nonpolar (it has some polarity owing to its quadrupole moment); addition of a polar cosolvent that is fully miscible with supercritical carbon dioxide can provide a more polar solubilizing environment needed for the extraction of polar substances. For example, solubility of fish liver oil in supercritical carbon dioxide increases by addition of small amounts of ethanol (Fig. 6).[3] Depending on the amount of the cosolvent added, critical temperature and critical pressure of the binary mixture also change.

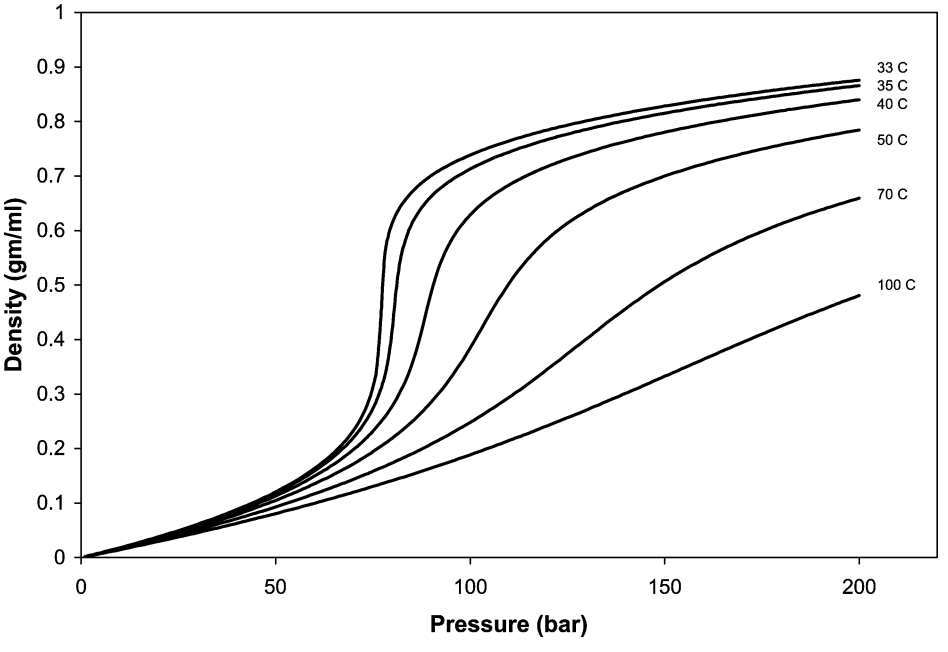


Fig. 1 Density dependence of carbon dioxide.

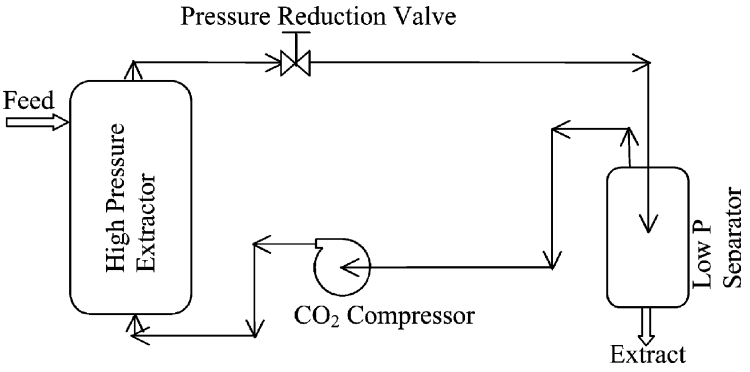


Fig. 2 Schematic of a typical supercritical fluid extraction process.

RAPID EXPANSION OF SUPERCRITICAL SOLUTION TO OBTAIN MICROPARTICLES

Once the solute is dissolved in the supercritical fluid, if the recovery of the solute is done by rapidly expanding the fluid, the solute can be recovered as fine particles or droplets. The pressure can be reduced as fast as the speed of sound, hence very high supersaturation can be achieved in a fraction of seconds. The solute precipitates out as a microparticles (if solid) or microdroplets (if liquid). Hence, it is possible to perform both extraction and micronization in a single step.

APPLICATIONS OF SUPERCRITICAL FLUID EXTRACTION

The SFE has been applied in a variety of industrial applications including: 1) foods: flavor extraction and concentration of flavors and fragrances, processing essential oils and edible oils; infusion of flavors and fragrances into the solid matrices; 2) neutraceuticals: vitamin extraction, antioxidant extraction, and the concentration of active ingredients; 3) petroleum: propane deasphalting, residuum oil extraction; 4) polymers: removal of monomers and oligomers, infusion of compounds into the polymeric matrix, and removal

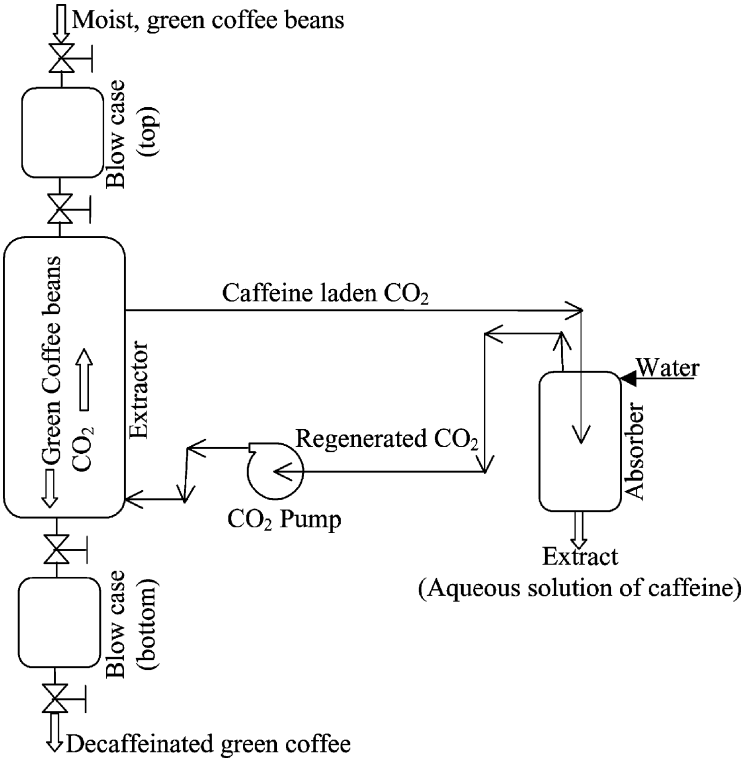


Fig. 3 Decaffeination of coffee beans using supercritical fluid extraction.

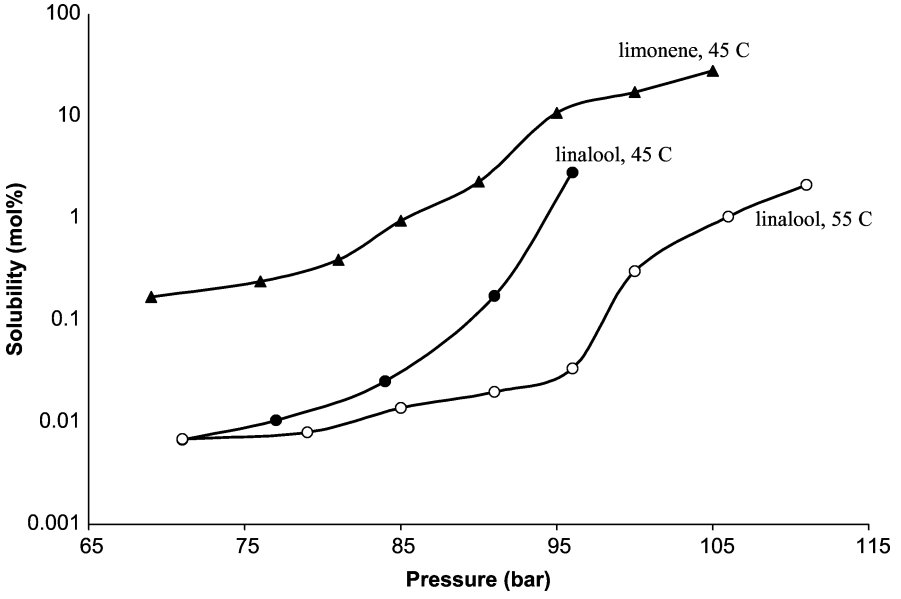


Fig. 4 Solubility of limonene and linalool in supercritical carbon dioxide. (From Ref.[1].)

of binder from powdered metals; 5) cleaning: precision machined components, silicon wafers, medical implants, and electronic components; 6) analytical: for extraction of the analytes from samples of food, cosmetics, polymers, and pharmaceuticals. Only a few of these applications are described here in detail.

DECAFFEINATION

Out of the above six major application areas, the applications in food, flavor, and fragrance have been heavily adopted by the industry. A good example is the coffee industry. Caffeine contents of coffee and tea are 3.5–5.0% and 1–3.5%, respectively, and depend on the soil, altitude, and climate of the plantation. There is a significant consumer demand for decaffeinated coffee that contains less than 0.4% caffeine. The SFE to decaffeinate coffee has been in industrial practice since more than two decades. Nearly 90% of the coffee consumed in the United States is now decaffeinated by supercritical CO_2 extraction.[4] Caffeine is extracted from the green coffee beans and then the aroma is simply developed later by roasting. In the case of decaffeination of tea, the extraction is conventionally performed on black tea. Because the enzymes in the green tea are to be protected, as they are needed in the development of the flavor and color at a later

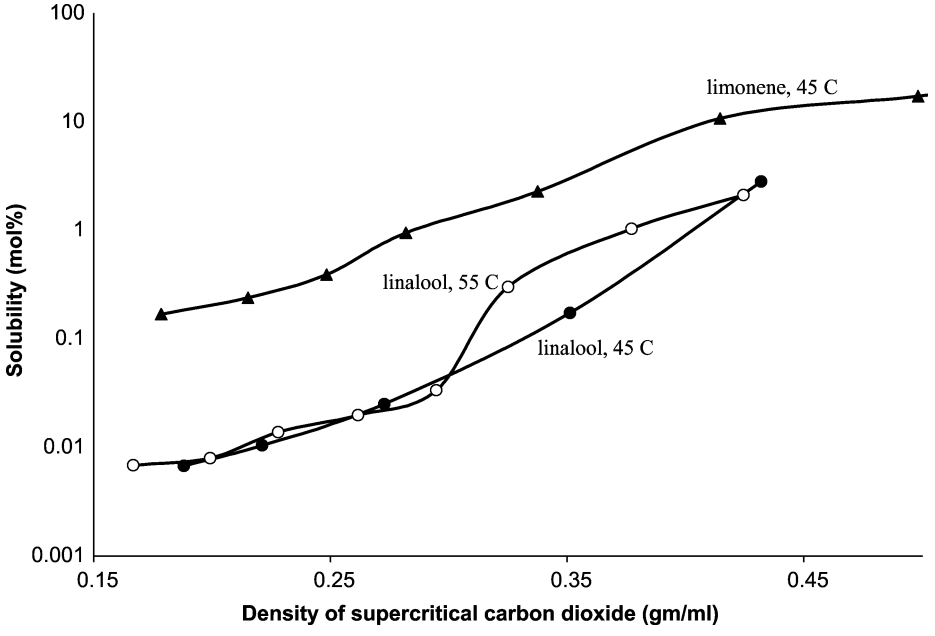


Fig. 5 Log of solubility vs. density shows a more linear behavior.

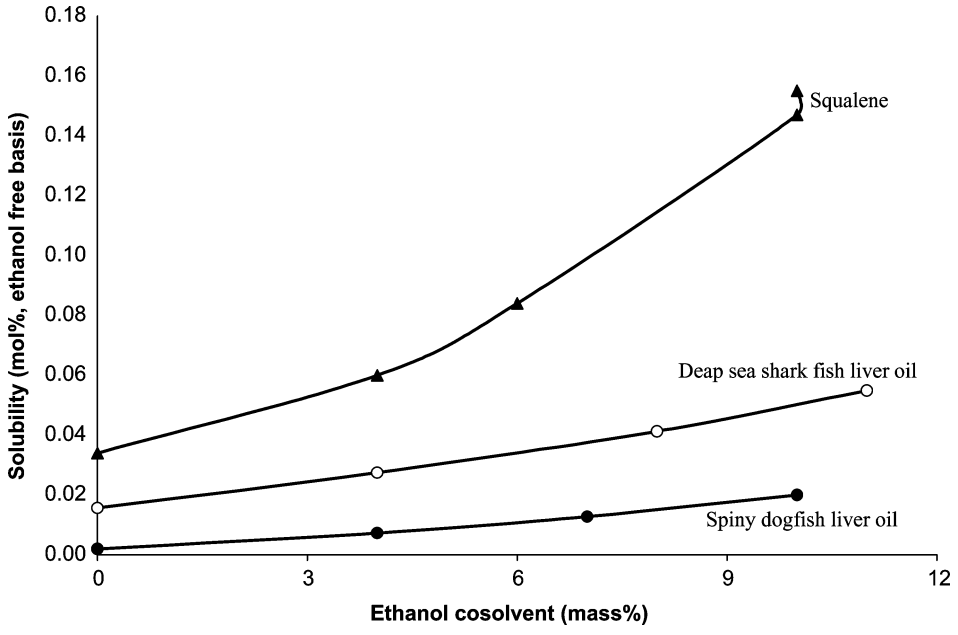


Fig. 6 Enhancement of fish oil solubility in supercritical carbon dioxide by addition of ethanol cosolvent, at 60°C and 250 bar. (From Ref.[3].)

stage, there were some concerns that supercritical CO_2 may inactivate the enzymes.

EXTRACTION OF FLAVORS AND FRAGRANCES

The flavors and fragrances extracted using supercritical carbon dioxide are significantly different from those extracted using steam distillation or solvent extraction. The SFE extract can almost be viewed as a new product, as usually the amount extracted is higher (Table 2) and the composition of the extract is somewhat different (Table 3), as more aromatic molecules are extracted. In many instances the extract using supercritical carbon dioxide is closer to the natural fragrance or flavor.[5–7] The higher SFE equipment cost can be easily offset by the higher yield of the product and the lower operating cost, as compared to steam distillation or solvent extraction.

Table 2 Comparison of percent yields of flavor and fragrance extracts from various natural products

Natural substance	Steam distillation (% yield)	Supercritical CO_2 (% yield)
Aniseed	2.1–2.8	7
Carrot	0.2–0.5	3.3
Cardamom	4–6	7
Clovebud	15–17	22
Cumin	2.3–3.6	14
Ginger	1.5–3.0	4.6
Garlic	0.06–0.4	4.6
Oregano	3–4	5
Pepper	1.0–2.6	8–18
Rosemary	0.5–1.1	7.5
Sage	0.5–1.1	4.3
Vetiver	0.5–1.0	1.0

(From Refs.[5,6].)

EXTRACTION OF EDIBLE OIL

There are several published studies on the use of supercritical carbon dioxide to extract oils from potato chips and other snack foods, motivated because of the increasing consumer awareness toward low-fat foods. About one-half of the oils in a potato chip containing about 40% oil can be extracted while maintaining the original flavor and texture. The extracted oil can be recovered by depressurization and reused in a subsequent frying operation. Supercritical carbon dioxide evenly dispersed the excess oil on the surface, thus removing the greasy taste. The processed fried snacks were found to have improved flavor and taste.[8]

POLYMER FRACTIONATION

Supercritical fluid can be used to extract low-molecular-weight polymers/oligomers from the given mixture, leaving the high-molecular-weight polymer behind. The pressure of the extract fluid is lowered sequentially in different vessels causing the precipitation of the polymer depending on its molecular weight. The lowest-molecular-weight polymer is precipitated the last. Thus, the broad-molecular-weight polymer mixture

Table 3 Composition of lavender extract by supercritical CO_2 and steam distillation

Chemical component	Supercritical CO_2 extract (%)	Steam distillation product (%)
1,8-Cineole	5.83	6.75
Linalool	25.29	35.31
Camphor	7.90	7.81
Borneol	2.30	2.98
4-Terpineol	3.79	3.34
Linalyl acetate	34.69	12.09
3,7-Dimethyl acetate	3.08	4.38
β-Farnescene	2.23	1.00
α-Bisabolol	2.09	3.76

(From Ref.[7].)

can be separated into the cuts of narrower-molecular-weight fractions. Carbon dioxide usually has extremely poor solubility for the polymers; however, it can be easily used to remove residual solvent or monomers. Also, the fractionation of softer polymers such as poly(ethylene glycol) is feasible using carbon dioxide.[9] Propane, ethylene, ethane fluids are needed for the fractionation of commercial polymers including poly-(methyl methacrylate), polyethylene using high pressures.[10] Isothermal pressure profiling in the liquid polyethylene/supercritical propane regime yielded 14 fractions with narrow polydispersity. This process can be used as an alternative to temperature rising elution fractionation, which was developed to fractionate linear low-density polyethylene on the basis of short-chain branching.

Supercritical fluid fractionation of hydroxyl-terminated polybutadiene was investigated by Watkins and Krukonis, as this polymer is used as a binder in solid propellants and the molecular weight plays an important role in the performance of the explosives.[11] The original polymer had a high polydispersity (ratio of weight average molecular weight to the number average molecular weight). On fractionation, 10 different cuts of narrower distribution were obtained in the pressure range 34–552 bar (Table 4).

WAX FRACTIONATION

Supercritical carbon dioxide, ethane, and propane have been examined for the fractionation of paraffin wax.[12] The original feed contained wax molecules with 10–35 carbon atoms. A narrower carbon distribution is needed in the printing ink, cosmetics, and pharmaceutical applications. Based on the cost analyses, vacuum distillation was proposed to be a cheaper option for light paraffin wax, whereas supercritical fractionation is a more attractive option for the fractionation of medium to heavy waxes.

Table 4 Fractionation of hydroxyl-terminated polybutadiene using supercritical propane at 130°C

Fraction	Pressure range (bar)	M_w	M_w/M_n
Original mixture	—	6,250	2.12
1	34–35	970	1.24
2	124–165	1,690	1.20
3	152–207	2,540	1.23
4	193–207	3,300	1.17
5	207–234	4,110	1.19
6	234–262	5,010	1.16
7	248–276	6,010	1.28
8	276–317	7,420	1.27
9	303–338	9,050	1.30
10	517–552	21,540	1.83

(From Ref.[11].)

EXTRACTION FROM ALCOHOLIC BEVERAGES

In addition to extraction from solids, supercritical fluids can be used to extract aromatic molecules from liquids. Señoráns et al. have utilized carbon dioxide to extract high-quality brandy aroma using a countercurrent supercritical fluid extractor.[13] The aroma quality is influenced by the extraction conditions. Medina and Martinez studied alcohol removal from beverages using supercritical carbon dioxide, to produce beverages with low-alcohol content but sufficient flavor, because of three key benefits: 1) water and salts are not appreciably removed by the carbon dioxide; 2) proteins and carbohydrates are not extracted or denatured; and 3) there is a good control in the aroma recovery. The alcohol removal efficiency increases with the extraction pressure; raffinate alcohol concentration can be reduced up to 3 wt.% at 250 bar and 40°C, from 6.2 wt.% in the feed.[14]

SAFETY AND HEALTH ISSUES

When dealing with supercritical carbon dioxide, there are two safety and health issues that are to be kept in mind when designing and operating the extractor: 1) the high pressure involved requires that the personnel are protected from the plant by proper isolating walls and 2) if carbon dioxide is released in the closed atmosphere, it can lead to asphyxiation, as it can replace the oxygen in the surroundings. If one is using a more flammable supercritical fluid (e.g., propane),

then the flammability becomes an additional concern. The concentrations should be not be allowed to fall between the explosive limits. In the case of ammonia, toxicity hazard also exists.

CONCLUSIONS

Supercritical fluid extraction offers several advantages over conventional extraction processes. The extraction is carried out at high pressures and then the extract is usually recovered by lowering the pressure, as the solubility is a strong function of fluid pressure. The compositions of the extracts are different from those from the liquid extraction. Supercritical fluid extraction has been well accepted for coffee decaffeination and is being applied in other food, cosmetics, and pharmaceutical applications. Supercritical carbon dioxide is an environmentally benign nonflammable fluid.

REFERENCES

1. Berna, A.; Chafer, A.; Monton, J.B. Solubilities of essential oil components of orange in supercritical carbon dioxide. J. Chem. Eng. Data **2000**, *45* (5), 724–727.
2. Dobbs, J.M.; Wong, J.M.; Lahiere, R.J.; Johnston, K.P. Modification of supercritical fluid phase behavior using polar cosolvents. Ind. Eng. Chem. Res. **1987**, *26* (1), 56–65.
3. Catchpole, O.J.; Grey, J.B.; Noermark, K.A. Fractionation of fish oils using supercritical CO_2 and CO_2 + ethanol mixtures. J. Supercrit. Fluids **2000**, *19* (1), 25–37.
4. McHugh, M.A.; Krukonis, V.J. *Supercritical Fluid Extraction: Principles and Practice*; Butterworth–Heinemann: Boston, 1994.
5. Mukhopadhyay, M. *Natural Extracts Using Supercritical Carbon dioxide*; CRC Press: Boca Raton, FL, 2000.
6. Moyler, D.A. Extraction of flavours and fragrances with compressed CO_2. In *Extraction of Natural Products Using Near-Critical Solvents*; King, M.B., Bott, T.R., Eds.; Blackie Academic & Professional (an imprint of Chapman Hall): Glasgow, U.K., 1994.
7. Reverchon, E.; Porta, G.D.; Senatore, F. Supercritical CO_2 extraction and fractionation of lavender essential oil and waxes. J. Agric. Food Chem. **1995**, *43* (6), 1654–1658.
8. Daneshvar, M.; Gulari, E. Supercritical-fluid fractionation of poly(ethylene glycols). J. Supercrit. Fluids **1992**, *5* (2), 143–150.
9. Inada, S.; Joji, O.; Giichi, T.; Schoichi, T.; Toshio, I.; Katsuyoshi, K.; Hirofumi, O.; Norio, Y. Improvement of taste of fried snack with subcritical or supercritical carbon dioxide. Jpn. Kokai Tokkyo Koho **1989**, 4. CODEN: JKXXAF JP 01243944 A2 19890928 Heisei.
10. Watkins, J.J.; Krukonis, V.J.; Condo P.D., Jr.; Pradhan, D.; Ehrlich, P. Fractionation of high density polyethylene in propane by isothermal pressure profiling and isobaric temperature profiling. J. Supercrit. Fluids **1991**, *4* (1), 24–31.
11. Crause, J.C.; Nieuwoudt, I. Paraffin wax fractionation: state of the art vs. supercritical fluid fractionation. J. Supercrit. Fluids **2003**, *27* (1), 39–54.
12. Watkins, J.J.; Krukonis, V.J. *Supercritical Fluid Processing of Propellants*, Technical Report PL-TR-91-3003; OLAC, Phillips Laboratory (AFSC): Edward Air Force Base, CA, December 1990.
13. Señoráns, F.J.; Ruiz-Rodríguez, A.; Ibáñez, E.; Tabera, J.; Reglero, G. Isolation of brandy aroma by countercurrent supercritical fluid extraction. J. Supercrit. Fluids **2003**, *26* (2), 129–135.
14. Medina, I.; Martinez, J.L. Dealcoholisation of cider by supercritical extraction with carbon dioxide. J. Chem. Technol. Biotechnol. **1997**, *68*, 14–18.

Supercritical Fluid Technology: Reactions

Aydin K. Sunol
Sermin G. Sunol
Naveed Aslam
University of South Florida, Tampa, Florida, U.S.A.

INTRODUCTION

Significant and steady inroads toward wider and more effective utilization of supercritical fluids have been made over the past three decades. Although the widely stated suitability for synthesis of high-value-added differentiated products is on the mark, continued interest in the displacement of key basic chemical processes with greener and superior supercritical ones is equally worthy owing to the longevity of such processes and the resulting socially responsible manufacturing practice in an era when the new major technology shifts and construction of new plants are far in between. Furthermore, a new wave of second-generation supercritical technologies started to emerge, particularly in materials, bringing forth new roles for dense gases. The motivation of this entry is to assess the current status of the reactions in supercritical fluids in an effort to extricate the challenges both in the current practice and in the potential areas into which we have not as yet ventured.

BACKGROUND

One of the significant paradigm shifts in chemical processing for the new millennium is the increased use of environmentally benign technology. The effectiveness of green solvents such as supercritical water and supercritical carbon dioxide for carrying out reactions, difficult separations, and materials processing is naturally very promising. These solvents are preferred due to their low cost, nontoxicity, nonflammability, and thermal stability. The effectiveness of the supercritical solvents is related to their state, critical temperature, and pressure. Supercritical fluids have the mobility of gases and the dissolving power of liquid solvents resulting in efficient penetration into porous matrices, high mass transfer and reaction rates, and high solvency. Furthermore, these properties are extremely sensitive to perturbations in temperature, pressure, and composition resulting in innovative processing concepts with tunable performance parameters suitable for devising creative processing strategies. Obviously, synergy between the physical characteristics of solvent and the conditions favorable for the desired chemistry is of paramount importance for the success of the application. Thus, over the last three decades, a spate of supercritical processes have been developed, particularly for manufacture of high-value-added products that are superior in performance and exhibit conscious regard for a more socially responsible manufacturing practice.

Despite the higher capital charges associated with relatively high pressures, the necessity to often add a new component into the processing environment, and operational challenges at conditions foreign to most process engineers, the interest in supercritical fluids had been growing steadily since the early 1980s beyond the select number of areas. We see applications in the food and beverage industries, pharmaceuticals, biomedical, microelectronic industries, textiles, forest products, petrochemicals, chemicals, environmental cleanup, syn-fuel production, polymeric materials, ceramics, auto industry, coatings and paint industry, energetic materials, and fuels.

More polymerization reactions carried out at supercritical conditions, select biomass conversion supercritical fluid technologies for hydrogen production, wider use of supercritical water oxidation processes, portfolio of self-assembly applications, a spate of opportunities in process intensification, many supercritical fluid aided materials synthesis applications, and numerous reactions for synthesis of specialty chemicals are expected for years to come.

The entry will start with fundamentals, focusing on the unique properties and the resulting opportunities covered in a generic fashion. Applications in different domains and reactive environments will follow. Some concluding remarks will summarize reactions in supercritical fluids and provide some perspectives for the years to come. Supercritical as well as other reactions can be facilitated through supercritical fluid aided synthesis and functionalization of catalysts. These are discussed elsewhere in the Encyclopedia.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120039329

FUNDAMENTALS

Properties and Opportunities

The unique properties of supercritical fluids make them attractive as a reaction medium as well. Although reactive supercritical fluids processes such as extraction of coal and supercritical water oxidation started to emerge in the 1970s and early 1980s, the first review on reactions in supercritical fluids was presented by Subramanian and McHugh, followed by many more recent and thorough reviews.[1–3] Earlier work on reactions at high pressure provided both fundamental understanding and technical readiness.[4–10]

The motivations for using supercritical fluids solvents in chemical reactions are many and most of the opportunities are due to the unique properties including density tuning as shown in Fig. 1.[5,11] In this section, these unique properties will be briefly discussed in the context of reactions, followed by the resulting opportunities and other related issues.

There are several reasons for carrying out the reactions at supercritical conditions. Naturally, some of the reasons are coupled. Nevertheless, they, in general, relate to favorable transport properties, unique solvency characteristics, favorable kinetic considerations, and their sensitivity to operating conditions (manipulated variables). These unique properties lead to opportunities in process synthesis, process intensification, and controllability. These advantages, coupled with the environmentally friendly nature of these processes, make reactions in supercritical fluids attractive. The effect of these properties on opportunities for these favorable reaction environments is summarized in Table 1.

Enhanced Reaction Rate

Reactions in supercritical media utilize high pressures. Therefore, the effect of pressure on reaction equilibrium as well as reaction rate plays an important role in supercritical phase reactions. Supercritical fluids that exhibit very high negative activation volumes for certain reactions will improve the rate and equilibrium conversion of the reaction.

The kinetics of the reaction can be explained in terms of the transition state theory. According to the theory, the reaction occurs via a transition state species M^* and the generic elementary reaction can be written as:

$$aA + bB + \cdots \leftrightarrow M^* \rightarrow \text{Products} \tag{1.1}$$

The effect of pressure on the rate constant is given as:

$$\left(\frac{\partial \ln k_x}{\partial P}\right)_T = \left(\frac{\partial \ln K_x^*}{\partial P}\right)_T + \left(\frac{\partial \ln \kappa}{\partial P}\right)_T = \frac{\Delta \overline{V}^*}{RT} + \left(\frac{\partial \ln \kappa}{\partial P}\right)_T$$

where k_x, rate constant; P, pressure; T, temperature; K_x^* mole fraction based equilibrium constant for reaction involving reactants and transition state κ_T, isothermal compressibility; $\Delta\overline{V}^*$, activation volume (difference between partial molar volume of activated complex and reactants), $\Delta \overline{V}^* = \overline{V}^* - a\overline{V}_A - b\overline{V}_B - \cdots$; R, universal gas constant.

The rate constant in the above equation is expressed in terms of pressure independent units (mole fraction). If the rate constant is expressed in terms of pressure

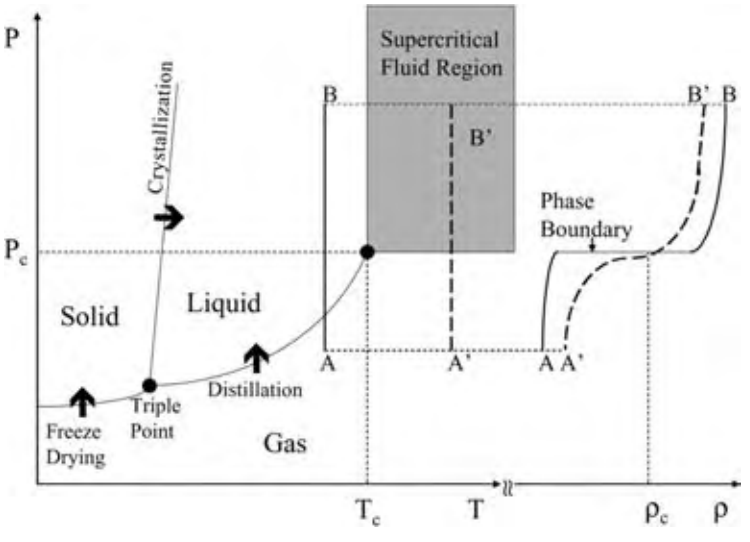


Fig. 1 Supercritical fluid region and density tuning.

Table 1 Effect of properties on opportunities for favorable reaction environments in supercritical media

	Homogenization	Tunability and control	Increased catalyst activity	Enhanced mass transfer	Increased selectivity	Ease of downstream separation
Enhanced reaction rate	Significant	Significant	Most significant	Not significant	Significant	Not significant
Enhanced solubility and selective solvation	Most significant	Significant	Significant	Significant	Significant	Significant
Transport properties (mobility)	Significant	Significant	Significant	Most significant	Significant	Significant
Sensitivity to operating conditions	Significant	Most significant	Significant	Significant	Significant	Significant

dependent units (such as concentration), the relevant equation is:

$$\left(\frac{\partial \ln k}{\partial P}\right)_T = -\frac{\Delta \overline{V}^*}{RT} + \left(\frac{\partial \ln \kappa}{\partial P}\right)_T + \kappa_T(1 - a - b - \cdots)$$

If the volume of activation is positive, the reaction is hindered by pressure. However, for high negative values of the volume of activation, the pressure enhances the rate of the reaction. Therefore, supercritical fluids that exhibit very high negative activation volumes for certain reactions will improve the rate of the reaction.

The volume of reaction, rather than activation, is crucial in determining the effect of pressure on the equilibrium constant:

$$\left(\frac{\partial \ln K_x}{\partial P}\right)_{T,x} = -\frac{\Delta \overline{V}_f}{RT}$$

where $\Delta \overline{V}_f$ is the reaction volume (difference between partial molar volume of products and reactants).

If the equilibrium constant is expressed in terms of pressure dependent units (such as concentration), the relevant equation is:

$$\left(\frac{\partial \ln K}{\partial P}\right)_{T,x} = -\frac{\Delta \overline{V}_f}{RT} + \kappa_T \sum \upsilon_i$$

where υ_i is the stoichiometric coefficient.

As the above equation implies, supercritical fluids that exhibit very high negative activation volumes for certain reactions will improve the equilibrium conversion of the reaction.

Enhanced Solubility

The enhancement factors for solubility of compounds, over ideal solubility, in supercritical fluids is typically around 10^4–10^6 and is discussed elsewhere. This enhancement implies that very heavy macromolecules can be solubilized in supercritical fluids to react or be selectively removed from the reaction environment. In another class of reactions, light noncondensable components such as hydrogen or oxygen will dissolve in supercritical fluids for effective hydrogenation or oxidation. Supercritical fluids can also dissolve in and expand liquids for a more favorable reaction and processing environment. This solubility behavior is synchronized with favorable reaction conditions through judicious selection of solvents and/or cosolvents.

Transport Properties (Mobility)

The supercritical fluids exhibit gas-like viscosities, diffusivities, and liquid-like densities. These favorable transport properties lead to enhanced mass transfer, permeation, and wetting characteristics. The mass transfer limited multiphase reactions will benefit from reduction of a number of phases, as in the case of most oxidation, hydrogenation, or replacement of the more viscous liquid phase with a supercritical or a less viscous expanded liquid phase. The mobility combined with tunability results in effective maintenance of catalyst activity in heterogeneous catalysis.

Sensitivity to operating conditions

The solubility including retrograde behavior, number and nature of coexisting phases, rate of phase creation, reaction rates, and transport properties are very sensitive to the operating conditions such as temperature, pressure, composition including cosolvents, and other external fields such as electric and magnetic as well as the rate of change of these operating conditions. The sensitivities to manipulated variables are usually coupled leading to novel processes and enable tuning of the process to impart the desired product properties.

Dramatic reduction of surface tension and reduction of dielectric constant affect the interfacial transport and solubility as well as reaction rate tuning, respectively.

Homogenization

Reactions that otherwise would be carried out in more than one phase (heterogeneous reactions) can be transformed to homogeneous ones, with the aid of supercritical fluids, where interphase transport limitations are eliminated. This is realized due to enhanced solubilities of the supercritical fluids. Typical examples are reactions in water (supercritical water can solubilize organic compounds), homogeneous catalytic reactions, reactions of organometallic compounds. Homogenizing one compound more than the other may also affect relative rates in complex reactions and enhance the selectivity.

In supercritical solutions, in a microscopic sense, molecules are nonuniformly distributed. There is aggregation of the solvent molecules around the solute and clusters are formed. The local clustering of solutes or solvents under supercritical condition increases the local concentration of the substrate or the catalyst in the solution and may result in enhanced reaction rates.

Supercritical fluids aid rapid diffusion of solutes or weakening of the solvation around reacting species. This may result in changes of the reaction pathways.

The homogenization could be effectively coupled with nucleation/crystallization to allow in-unit separation and phase transfer catalysis.

Increased Catalyst Activity

Some heterogeneous catalytic reactions are carried out in supercritical phase, to increase catalyst activity and life through in situ regeneration of surfaces with tuning of operation conditions. For example, supercritical fluids are capable of dissolving carbon, which may be irreversibly deposited on the catalyst otherwise.

Tunability and Control

Some properties of supercritical fluids can be monitored (manipulated) continuously by adjusting the temperature and pressure or density of the fluid. Dielectric constant is such a property and the solvent's dielectric constant can influence the rate of the reaction.

Supercritical fluids can be combined with polar cosolvent, which enhances the solubility selectivity and sensitivity of the environment to manipulated variables such as temperature and pressure. This combination also tunes the reaction to adjust the product distribution.[12]

Enhanced Mass Transfer

In many instances, reaction rates are limited by diffusion in the liquid phase. The rate of these reactions can be increased if the reaction is carried out in the supercritical phase. Typical examples are enzyme-catalyzed reactions as well as very fast reactions such as some free radical reactions. Selectivity considerations usually dominate in complex reactions. If some steps of the complex reaction are controlled by diffusion, changing the diffusivity changes relative rates of the reaction steps and affects the selectivity.

The mass transfer rates are also enhanced in porous media allowing effective in situ regeneration and removal of products/reactants as they become solubilized. The reduced viscosities play an important role here.

Ease of Downstream Separation

Another utility of supercritical fluids as the reaction medium is fractionation and purification of the products or removal of unreacted reactants from the product stream. Supercritical fluids can be used as a solvent or as an antisolvent in this instance.

In the case of homogeneous catalysis, employing supercritical fluids enables complete recovery of the expensive transition metal species. Also, these species may have environmentally unfriendly effects if not recovered completely. The combination of ionic liquids with the supercritical fluid can lead to product isolation as well as catalyst immobilization. Thus, catalysts can be recycled batchwise.[7]

Induced Reaction Selectivity

Supercritical carbon dioxide introduces changes in the selectivity of the reactions. These changes include chemical selectivity, as well as stereoselectivity. The selectivity changes originate from the solvent as well as the modification of the catalysts with the carbon supercritical carbon dioxide.

Energy Demand Reduction

The mechanical and thermal energy requirements depend on operating conditions, excursions in these conditions throughout the process, and heat integration. Naturally, we do have pressurization requirements

for all supercritical reactions, particularly high for supercritical water. Thermal energy loads for reactions at elevated conditions such as reactions in supercritical water are rather high as well. The energy demand reduction opportunities are due to favorable transport properties and enhanced rates as well as sensitivity to operating conditions bringing efficient recovery.

In instances where supercritical reactions displace mass transfer limited reactions, less stringent agitation or no agitation results in reduction of mechanical energy requirements. When enhanced reaction rates are coupled with higher mass and heat transfer rates, energy required and volume per unit product decrease substantially. Furthermore, efficient recycling and ease in downstream separation imply reduction in energy consumption. More systematic studies in heat and mass integration of these processes will eventually be done as industrial deployment of supercritical fluids reaches a critical threshold.

Safety

Safety issues often relate to process safety as well as end-product safety. The inherent process safety concern with supercritical fluids/reactions is due to pressurized inventory. For operation of these processes, effective pressurization, depressurization, and relief systems are a prerequisite. When these safety concerns are adequately handled, one sees the safety advantages that supercritical reactions and processes bring about. In most supercritical environments, which are water or carbon dioxide based, we replace flammable and toxic alternatives. The enhanced rates and increased throughputs imply smaller equipment and inventories, further reducing the safety hazards. Furthermore, gases like hydrogen, carbon monoxide, and oxygen can be handled more safely in the presence of supercritical carbon dioxide. This is crucial in hydrogenation as well as oxidation reactions. In the case of product safety, the end products are inherently contaminant free, posing less of a safety concern.

Naturally, a systematic study on safety of high-pressure reactions, particularly for runaway reactions is well overdue.

APPLICATIONS

Supercritical fluids have been utilized as a reaction medium from basic industries such as syn-fuels, biomass conversion, environmental remediation to high-value-added specialty chemicals and materials. An exposition of the underlying mechanism along with application domain will follow in this section. Homogeneous and heterogeneous reactions will be followed by biochemical and polymerization reactions, all with supercritical fluids. The typical examples for which original references are too numerous to cite individually are summarized in Tables 2 to 4.

Homogeneous Reactions in Supercritical Solvents

Homogeneous reactions carried out in supercritical fluids can be either catalytic or noncatalytic. The objective of carrying catalytic and noncatalytic reactions in supercritical fluids is to increase the overall rate of the reaction by eliminating the interfacial transport effects.

Homogeneous Reactions Catalyzed by Organometallic Compounds

Homogeneous catalysts have advantages over heterogeneous catalysts such as the possibility of carrying out the reaction under milder conditions, higher activity, and selectivity, ease of spectroscopic monitoring, and controlling the tunable reaction sites. A recent review by Noyori et al. discusses homogeneous catalytic reactions under supercritical conditions.[13] Examples of homogeneous catalytic reactions carried out under supercritical conditions are summarized in Table 2.

Because most organic reactants and products are not soluble in water, it cannot be utilized as the solvent in most catalytic reactions, although most catalytic materials are soluble in water. Therefore, liquid solvents for homogenization of catalytic reactions are usually organic solvents. An alternative to environmentally unacceptable organic solvents is a supercritical solvent that has several advantages over organic solvents. The aforementioned advantages are increased reaction rate, higher selectivity, and ease in separation of reactants, products, and the catalyst after the reaction. Because the properties of supercritical solvents are sensitive to operating conditions, reaction rate and selectivity are more readily tunable when the reaction is carried out in supercritical solvents.

Carbon dioxide is the supercritical solvent that is used most in homogeneous catalytic reactions. In addition to being environmentally acceptable (nontoxic, nonflammable), carbon dioxide is inert in most reactions, is inexpensive, and is available in large quantities. Its critical temperature is near ambient. Supercritical carbon dioxide dissolves nonpolar, nonionic, and low molecular mass compounds. However, addition of cosolvents enhances the solubility of many compounds in supercritical carbon dioxide.

When homogeneous reactions are carried out under supercritical conditions, liquid gas interfacial transport

Table 2 Homogeneous reactions in supecritical fluids

Reaction	Supercritical medium	Catalyst
Hydrogenation		
CO_2 to formic acid	CO_2	Ruthenium(II) phosphine complex
Asymmetric hydrogenation of tiglic acid	CO_2	Ruthenium catalyst
Asymmetric hydrogenation of enamides	CO_2	Cationic rhodium complex
Asymmetric hydrogenation of imines	CO_2	Iridium complex
Cyclopropene	CO_2	Manganese catalyst
3,3-Dimethyl-1,2-diphenyl cyclopropene	CO_2	Manganese catalyst
Isoprene	CO_2	Rhodium catalyst
C–C bond formation		
Ring opening metathesis polymerization	CO_2	Ruthenium catalyst
Ring closing metathesis of dienes to cyclic olefins	CO_2	Ruthenium catalyst
Cyclotrimerization of alkynes	CO_2	Cobalt catalyst
Olefination of arylhalides	CO_2	Palladium catalyst
Hydrovinylation of styrenes	CO_2	Nickel catalyst
Synthesis of 2-pyrones	CO_2	Nickel catalyst
Synthesis of cyclopentenes	CO_2	Cobalt catalyst
Cyclotrimerization of alkynes to substituted benzene derivatives	CO_2	Cobalt catalyst
Phenol and *p*-cresol alkylation	Water	None
Diels–Alder cycloaddition	Water	None/NaOH
Ring opening of 2,5-dimethylfuran	Water	Acid
Oxidation		
Cyclohexane	CO_2	Iron catalyst
Methylacrylate	CO_2	Palladium catalyst
Alkene epoxidation	CO_2	Molybdenum catalyst
2,3-Dimethylbutene epoxidation	CO_2	Molybdenum catalyst
Cyclooctene epoxidation	CO_2	Molybdenum catalyst
Cyclohexene epoxidation	CO_2	Molybdenum catalyst
2,3-Dimethylbutene epoxidation	CO_2	Molybdenum catalyst
Phenols	Water	None
Ethanol	Water	None
2-Propanol	Water	None
2-Butanol	Water	None
Chlorinated hydrocarbons	Water	None
Hydroformylation		
1-Octene	CO_2	Rhodium catalyst
Propylene	CO_2	Cobalt carbonyl
Styrene	CO_2	Rhodium catalyst
Carbonylation of arylhalides	CO_2	Palladium catalyst
Carbonylation of alkynes and alkenes	CO_2	Cobalt catalyst
Isomerization		
1-Hexene to 2-hexene	CO_2	Iron catalyst
Hydration/dehydration		
Conversion of *tert*-butyl alcohol to isobutylene	Water	None/H_2SO_4/NaOH
Dehydration of cyclohexanol, 2-methyl cyclohexanol, 2-phenylethanol	Water	Acid
Hydrolysis		
Esters to carboxylic acids and alcohols	Water	Autocatalytic
Nitriles to amides and then to acid	Water	Autocatalytic
Butyronitrile	Water	Autocatalytic
Polyethyleneterephthalate and polyurethane	Water	None
Diaryl ether to hydroxyarene	Water	None
Triglycerides into fatty acid	Water	None
Decomposition		
Cellulose and glucose decomposition	Water	None
Nitrobenzene	Water	None
4-Nitroaniline	Water	None
4-Nitrotoluene	Water	None

Table 3 Heterogeneous catalytic reactions in supercritical solvents

Reaction	Supercritical medium	Catalyst
Hydrogenation		
Fats and oils	Propane, CO_2	Supported platinum, palladium catalysts
Fats and oils	CO_2	Nickel catalyst
Acetophenone, cyclohexene	CO_2	Palladium on polysiloxane
Ethyl-pyruvate	CO_2	Pt/Al_2O_3
Fischer–Tropsch synthesis	*n*-Hexane, *n*-pentane, propane	Fe, Ru, Co/Al_2O_3, SiO_2
Dibenzothiophene hydrodesulfurization	Water	$Ni–Mo/Al_2O_3$
Dehydrogenation		
Cyclohexanol	Water	Pt catalyst
Oxidation		
Toluene	CO_2	Co/Al_2O_3
Propylene	CO_2	$CuI/Cu_2O/MnO_2$ on Al_2O_3
Benzyl alcohol	CO_2	Pd/Al_2O_3
Ethanol	CO_2	Pt/TiO_2
Methanol	CO_2	Iron oxide on Mo aerogel
Propene	SC reactant	CaI_2, CuI, Cu_2/MgO on Al_2O_3
Isobutane	SC reactant	SiO_2, TiO_2, Pd/carbon
1,4-Dichlorobenzene	Water	V_2O_5
NH_3	Water	MnO_2/CeO_2
Acetic acid	Water	Cu/Zn/Co oxide supported catalyst
Pyridine	Water	MnO_2/CeO_2, Pt/Al_2O_3, MnO_2/Al_2O_3
Phenol	Water	Cu/Zn/Co oxide, MnO_2/CuO, MnO_2
Chlorophenol	Water	CuO supported on zeolites
Cracking		
Heptane	CO_2	Zeolite
Rearrangement and isomerization		
1-Hexene	CO_2 and cosolvents	Pt/Al_2O_3
Xylene	SC reactant	Solid acid catalyst
Cyclohexene to methyl-cyclopentene	Water	Solid acid catalyst
Alkylation		
Benzene, ethylene, isopentane, isobutene isobutane	CO_2 or SC reactant	Zeolite
Metislene propene, propan2-ol	CO_2	Acid catalyst
Disproportionation		
Toluene to *p*-xylene, benzene	SC reactant	Zeolite
Ethylbenzene to benzene and diethylbenzene	Butane, pentane	Zeolite
Coupling		
Alkene-arene coupling	Water	Palladium catalyst
Esterification		
Oleic acid	CO_2	Biocatalyst
Glycerol and CO_2	CO_2	Zeolite
Hydroformylation		
Oct-1-ene	CO_2	Supported rhodium catalyst
Propylene	CO_2	Supported rhodium and iron catalyst
1-Hexene	CO_2	Rhodium catalyst

is eliminated, which is an advantage for reactions such as hydrogenation, where diffusion of gas into the liquid may be limiting the reaction rate. In asymmetric hydrogenation reactions, hydrogen and the supercritical solvent are miscible and this results in better enantioselectivity.[6] In Diels–Alder reactions, the advantage of the supercritical solvent is higher selectivity obtained rather than increased rate of the reaction due to the solvent. Although most of the oxidation reactions are carried out in supercritical water, recently, homogeneously catalyzed reactions in supercritical carbon dioxide are increasingly reported.

Table 4 Polymerization reactions in supercritical solvents

Polymerization mechanism	Substrate
Transition metal catalyzed, ring opening polymerization	Norbornene polymer, polycarbonate
Dispersion, cationic polymerization	Isobutylene polymer
Homogeneous/precipitation, cationic polymerization	Vinyl ether polymer
Homogeneous, free radical/cationic polymerization	Amorphous fluoropolymers
Precipitation, free radical polymerization	Vinyl polymer, semicrystalline fluoropolymers
Dispersion, free radical polymerization	Polyvinyl acetate and ethylene vinyl acetate copolymer

Homogeneous Reactions of Supercritical Water

Homogeneous reactions carried out at supercritical conditions within water are organo-metallic reactions and Diels–Alder reactions. Reactions in supercritical water are well studied and will be described in the following section.[14]

Supercritical water has a low dielectric constant compared to liquid water. The dielectric constant changes significantly with the density of the supercritical fluid. Also, the effect of hydrogen bonding is less pronounced at supercritical conditions, one consequence of which is high solubility of organics in supercritical water. When the homogeneous reaction is carried out in supercritical water, we get a high reactant concentration and the reaction can proceed in the absence of interphase mass transfer resistances. Also, ion dissociation constant of water is higher in the critical region and is lower as supercritical conditions are accessed. These properties also vary continuously in the supercritical region, so that they can be tuned during the reaction by changing temperature and/or pressure.

Examples of the homogeneous reactions in supercritical water are included in Table 2. Use of acid or base catalysts enhances the rates of some of the reactions.[14]

Homogeneous Noncatalytic Reactions in Supercritical Solvents

The use of supercritical fluids as reaction media for organometallic species is also investigated.[15] Reactions include photochemical replacement of carbon monoxide with N_2 and H_2 in metal carbonyls, where the reaction medium is supercritical xenon. Also, photochemical activation of C–H bonds by organometallic complexes in supercritical carbon dioxide is investigated. More recent studies on photochemical reactions also include laser flash photolysis of metal carbonyls in supercritical carbon dioxide and ethane and laser flash photolysis of hydrogen abstraction reaction of triplet benzophenone in supercritical ethane and CHF_3.[16,17]

Heterogeneous Reactions in Supercritical Solvents

Heterogeneous reactions in supercritical fluids can be catalytic or noncatalytic. Catalytic heterogeneous reactions are carried out on solid catalysts and are of great importance in the chemical process industries. The advantages of carrying out these reactions in a supercritical medium include enhanced interphase and intraparticle mass and heat transfer and in situ regeneration of catalyst and are described in the next section. Catalytic supercritical water oxidation also utilizes favorable properties of supercritical water, which are also discussed. Other heterogeneous reactions that will be described are fuels processing reactions and conversion and treatment of biomass.

Heterogeneous Catalytic Reactions in Supercritical Solvents

Obviously, a solid-catalyzed reaction takes place only on the active sites of the porous catalyst with the implication of some mass and heat transport steps prior to and after the reaction. The first step is the diffusion of the reactants through the film surrounding the catalyst particle to the external surface of the catalyst and diffusion of the reactants in the catalyst pore to the active site in the pores. These steps are limited by the diffusivity and viscosity of the reactants. In the case of a supercritical fluid phase reaction, the diffusivity is higher than the liquid diffusivity, the viscosity is less than the liquid viscosity, and, therefore, the rate of transfer to the active site will be higher. After adsorption, reaction, and desorption steps, the products have to diffuse out of the pore, and through the film surrounding the particle into the bulk fluid. The rates of these steps can be accelerated utilizing a supercritical medium for the reaction. Heat transfer effects are also important in a solid-catalyzed reaction. Higher thermal conductivity of supercritical fluids is an advantage as well.[18]

For two-phase reactions (typically hydrogenation and oxidation reactions), the first step is the diffusion

of the gas reactant to and through the gas–liquid interface and then into the bulk liquid. This mass transfer limitation is also eliminated if the reaction is carried out in a supercritical medium since the reactants are going to be in a single phase.

Supercritical fluids bring other benefits to solid-catalyzed reaction rate through eliminating or minimizing mass and heat transfer resistances. Supercritical solvents have the ability to regenerate the catalyst during the course of the reaction, which increases the catalyst life and activity, because undesirable deposits on the catalyst, such as carbon deposits, are soluble in the supercritical fluids. The rate of the intrinsic reaction is increased in supercritical fluids and by tuning the properties of the supercritical medium, one can control the selectivity.[18–20]

Supercritical fluids may also bring opportunities in downstream separation of the reactants and products. Examples of solid-catalyzed reactions in supercritical fluids are given in Table 3.

Catalytic supercritical water oxidation is an important class of solid-catalyzed reaction that utilizes advantageous solution properties of supercritical water (dielectric constant, electrolytic conductance, dissociation constant, hydrogen bonding) as well as the superior transport properties of the supercritical medium (viscosity, heat capacity, diffusion coefficient, and density). The most commonly encountered oxidation reaction carried out in supercritical water is the oxidation of alcohols, acetic acid, ammonia, benzene, benzoic acid, butanol, chlorophenol, dichlorobenzene, phenol, 2-propanol (catalyzed by metal oxide catalysts such as CuO/ZnO, TiO_2, MnO_2, $KMnO_4$, V_2O_5, and Cr_2O_3), 2,4-dichlorophenol, methyl ethyl ketone, and pyridine (catalyzed by supported noble metal catalysts such as supported platinum).[21,22]

Heterogeneous Noncatalytic Reactions in Supercritical Solvents

Use of the supercritical fluids as the reaction medium in syn-fuel processing is one of the earlier applications in the field. The advantage of the supercritical fluid as the reaction medium is again threefold. During thermal degradation of fuels (oil-shale, coal), primary pyrolysis products usually undergo secondary reactions yielding to repolymerization (coking) or cracking into gas phase. Both reactions decrease the yield of the desired product (oil). To overcome this problem, dense (supercritical) hydrogen donor (tetralin) or nonhydrogen donor (toluene) or inorganic (water) medium is used. Also, supercritical medium provides ease of transport in and out of the porous coal matrix. Finally, downstream processing (separation of the products) becomes an easier task, when supercritical medium is used. A review of the use of supercritical fluids in coal processing is given by Kershaw, while the mechanism is discussed by Sunol.[23,24]

The forest product applications in this category include biomass conversion and delignification for pulping purposes. Both provide unique opportunities, the first for hydrogen generation and later to obtain pulp within reaction times that are reduced almost two orders of magnitude.[25,26]

Polymerization Reactions in Supercritical Solvents

Supercritical carbon dioxide is a promising green alternative to traditional solvents in polymer synthesis because of gas-like transport properties and liquid-like solubility. It can be removed easily from the polymer solution by depressurization during drying of the polymer. It provides easy separation of the polymer from the unreacted monomers and catalysts. Finally, it also exhibits Lewis acid–base interactions with electron donating functional groups of polymer chains.[27] Examples of homogeneous and heterogeneous polymerization reactions carried out in supercritical carbon dioxide are given in Table 4.

Butane, pentane, and propane are also used as the reaction medium in polymer synthesis because carbon dioxide is not a strong solvent for most polymers.[28] Furthermore, some polymerization reactions (such as polyethylene synthesis) are carried out under supercritical conditions of the monomer.

Biochemical Reactions in Supercritical Solvents

Because of their tunable properties, supercritical solvents provide a useful medium for enzyme-catalyzed reactions.[29] The mechanism of enzyme-catalyzed reactions is similar to the mechanism described for solid-catalyzed reactions. External as well as internal transport effects may limit the reaction rate. Utilizing supercritical fluids enhances external transport rate due to increase in the diffusivity and therefore mass transfer coefficient. Internal transport rate depends on the fluid medium as well as the morphology of the enzyme. Supercritical fluids can alter both.

Water is known to be essential for the enzyme activity.[30] Small amounts of water enhance enzyme activity; however, excess water hinders the rate of some enzyme-catalyzed reactions. The active site concentration on enzymes, hence the enzyme activity, is found to be higher in the presence of hydrophobic supercritical fluids (ethane, ethylene) compared to hydrophilic supercritical carbon dioxide.

The effect of pressure on enzyme-catalyzed reactions can be explained in terms of the transition theory. Supercritical fluids that exhibit very high negative activation volumes for certain reactions are expected to improve the rate of the reaction.

Although, supercritical carbon dioxide has the advantage of being nontoxic and abundant, it is practically immiscible with water. Therefore, supercritical fluids used as the reaction medium in enzyme-catalyzed reactions include fluoroform, sulfur hexafluoride, and ethane, while lipases are the enzymes utilized in such reactions.[31]

In addition to advantages like higher initial reaction rate and higher conversion, supercritical fluids provide an easy separation of products and unreacted substrates. This ecologically safe recovery of the products is a unique advantage provided by supercritical carbon dioxide.[29]

CONCLUSIONS

A variety of chemical and biological reactions involving supercritical fluid technology are being explored and developed. They include polymerization reactions, biomass conversion, hydrogen production, applications of supercritical water oxidation, self-assembly applications, synthesis of specialty chemicals, manufacture of materials with tailored properties, and much more. These developments and new ones are expected to mature and be commercially deployed in years to come.

REFERENCES

1. Bartle, K.D.; Martin, T.G.; Williams, D.F. Chemical nature of a supercritical-gas extract of coal at 350 degrees C. Fuel **1975**, *54* (4), 226–235.
2. Modell, M. Processing Methods for the Oxidation of Organics in Supercritical Water. U.S. Patent 4,338,199, **1982**.
3. Subramanian, B; McHugh, M.A. Reactions in supercritical fluids—a review. Ind. Eng. Chem. Res. **1986**, *25*, 1–12.
4. Savage, P.E.; Gopalan, S; Mizan, T.I.; Martino, C.J.; Brock, E.E. Reactions at supercritical conditions: applications and fundamentals. AIChE J. **1995**, *41* (7), 1723–1778.
5. Clifford, A.A. Reactions in supercritical fluids. In *Supercritical Fluids, Fundamentals and Applications*; Kiran, E., Levent Sengers, J.M.H., Eds.; Kluvert Academic Publishers: Dordrecht, 1994; 449–479.
6. Ikariya, T.; Kayaki, Y. Supercritical fluids as reaction media or molecular catalysis. Catal. Surv. Jpn. **2000**, *4*, 39–50.
7. Leitner, W. Supercritical carbon dioxide as a green reaction medium for catalysis. Acc. Chem. Res. **2002**, *35*, 746–756.
8. Leitner, W. Carbon dioxide as an environmentally benign reaction medium for chemical synthesis. Appl. Organomet. Chem. **2000**, *14*, 809–814.
9. Oakes, R.S.; Clifford, A.A.; Rayner, C.M. The use of supercritical fluids in synthetic organic chemistry. J. Chem. Soc. Perkin Trans. **2001**, *1*, 917–941.
10. Sunol, A.K.; Sunol, S.G. Substitution of solvents by safer products and processes. In *Handbook of Solvents*; Wypych, G., Ed.; Chem Tec Publishing: Toronto, 2001; 1419–1459.
11. Wu, B.C.; Paspek, S.C.; Klein, M.T.; LaMarka, C. Reactions in and with supercritical fluids—a review. In *Supercritical Fluid Technology*; Bruno, T.J., Ely, J.F., Eds.; CRS Press: London, **1991**; 511–524.
12. Eckert, C.A.; Bush, D.; Brown, J.S.; Liotta, C.L. Tuning solvents for sustainable technology. Ind. Eng. Chem. Res. **2000**, *39*, 4615–4621.
13. Jessop, G.P.; Ikariya, T.; Noyori, R. Homogeneous catalysis in supercritical fluids. Chem. Rev. **1999**, *99*, 475–493.
14. Savage, P.E. Organic chemical reactions in supercritical water. Chem. Rev. **1999**, *99*, 603–621.
15. Poliakoff, M.; George, M.W.; Howdle, S.M. Inorganic and related chemical reactions in supercritical fluids. In *Chemistry under Extreme or Non-Classical Conditions*; van Eldik, R., Hubbard, D.C., Eds.; Wiley: New York, 1997; 189–218.
16. Ji, Q.; Eyring, E.M.; van Eldik, R.; Johnston, K.P. Laser flash photolysis studies of metal carbonyls in supercritical CO_2 and ethane. J. Phys. Chem. **1995**, *99*, 13461–13466.
17. Roberts, C.B.; Brennecke, J.F.; Chateauneuf, J.E. Solvation effects of reactions of triplet benzophenone in supercritical fluids. AIChE J. **1995**, *41*, 1306–1318.
18. Baiker, A. Supercritical fluids in heterogeneous catalysis. Chem. Rev. **1999**, *99*, 453–473.
19. Subramanian, B. Enhancing the stability of porous catalysts with supercritical reaction media. Appl. Catal. A Gen. **2001**, *212*, 199–213.
20. Subramanian, B.; Lyon, C.J.; Arunajatesan, V. Environmentally benign multiphase catalysis with dense phase carbon dioxide. Appl. Catal. B Environ. **2002**, *37*, 279–292.

21. Ding, Z.Y.; Frish, M.A.; Li, L.; Gloyna, E.F. Catalytic oxidation in supercritical water. Ind. Eng. Chem. Res. **1996**, *35*, 3257–3279.
22. Savage, P.E. Heterogeneous catalysis in supercritical water. Catal. Today **2000**, *62*, 167–173.
23. Kershaw, J.R. Supercritical fluids in coal processing. J. Supercrit. Fluids. **1989**, *2*, 35–45.
24. Sunol, A.K.; Beyer, G.H. Mechanism of supercritical extraction of coal. Ind. Eng. Chem. Res. **1990**, *29*, 842–849.
25. Vick Roy, J.R.; Converse, A.O. Biomass hydrolysis with sulfur dioxide and water in the region of the critical point. In *Supercritical Fluid Technology*; Penniger, J.M.L., Radosz, M., McHugh, M.A., Krukonis, V.J., Eds.; Elsevier: Amsterdam, 1985; 397–414.
26. Sunol, A.K. Supercritical Delignification of Wood. U.S. Patent 5,041,192, Aug 20, **1991**.
27. Kendall, J.L.; Canelas, D.A.; Young, J.L.; DeSimone, J.M. Polymerization in supercritical carbon dioxide. Chem. Rev. **1999**, *99*, 543–563.
28. Srinivasan, G.; Elliot, J.R. Microcellular materials via polymerization in supercritical fluids. Ind. Eng. Chem. Res. **1992**, *31*, 1414–1417.
29. Nakamura, K. Biochemical reactions in supercritical fluids. In *Supercritical Fluid Processing of Food and Biomaterials*; Rizvi, S.S.H., Ed.; Academic & Professional: London, 1994; 54–61.
30. Mesiano, A.J.; Beckman, E.J.; Russel, A.J. Supercritical biocatalysis. Chem. Rev. **1999**, *99*, 623–633.
31. Knez, Z.; Habulin, M. Compressed gases as alternative enzymatic-reaction solvents: a short review. J. Supercrit. Fluids. **2002**, *23*, 29–42.

Supercritical Water Oxidation

Ram B. Gupta
Department of Chemical Engineering, Auburn University, Auburn, Alabama, U.S.A.

INTRODUCTION

A fluid is supercritical when it is compressed beyond its critical pressure and heated beyond its critical temperature. Hence, water is supercritical at >374°C temperature and >22.1 MPa pressure. Supercritical water has liquid-like density and gas-like transport properties, and behaves very differently than it does at room temperature. For example, it is highly nonpolar, permitting complete solubilization of most organic compounds and oxygen. The resulting single-phase mixture does not have many of the conventional transport limitations that are encountered in multi-phase reactors. However, the polar species present, such as inorganic salts, are no longer soluble and start precipitating. The physiochemical properties of water, such as viscosity, ion product, density, and heat capacity, also change dramatically in the supercritical region with only a small change in the temperature or pressure, resulting in a substantial increase in the rates of chemical reactions. For example, Fig. 1 shows, how density, dielectric constant, and ionic product of water vary with temperature at 24 MPa.

From the above figure, it is interesting to see that the dielectric behavior of 200°C water is similar to that of ambient methanol, for 300°C water it is similar to that of ambient acetone, for 370°C water it is similar to that of methylene chloride, and for 500°C water it is similar to that of ambient hexane. In addition to the unusual dielectric behavior, transport properties of water are significantly different from the ambient water as shown in Table 1.

WASTE OXIDATION

Owing to the lack of mass transfer limitations and high thermal energy, oxidation in supercritical water is very fast. Usually, in less than 1 min of reaction time, complete oxidation of organics to CO_2 and mineral acids is achieved. Supercritical water has been successfully used to completely oxidize chemicals including polychlorinated biphenyls, organic solvents, and other industrial wastes. In fact, it is emerging a promising alternative to the incineration of aqueous organic waste streams and has received enormous interest during the last decade. Supercritical water oxidation (SWO) can be seen as a further development of the well-established wet air oxidation (WAO) process, running at temperatures up to 320°C and pressures up to 20 MPa. Treatment times in the WAO process are normally as high as several hours, and complete destruction of the organic material is seldom achieved, so a further waste treatment is necessary. In SWO, treatment times are in the range of seconds to a few minutes and more than 99% destruction is achieved in most cases. Incineration, WAO, and SWO are further compared in Table 2.

In a typical SWO operation, organic waste and high-pressure oxygen (or air, or H_2O_2) are fed to the reactor. The reaction products are then cooled and depressurized to collect the benign products (Fig. 2).

Owing to the use of excess oxygen, various elemental species from waste convert into their highest oxidation form. For example, carbon converts to CO_2, hydrogen converts to H_2O, nitrogen converts to HNO_3, sulfur converts to H_2SO_4, etc. Early reactors were simple tube reactors as shown in Fig. 3.

The above tube reactor is also designed to have intermittent ports for the entry of oxidant (e.g., oxygen) and/or quench water. This reactor design is cheaper than the other designs mainly owing to the ease of fabrication and installation. However, if the corrosive species are present then severe wall corrosion can occur. Hence, tube reactors are recommended for use in oxidation of waste that does not contain heteroatomic molecules (e.g., chlorine, sulfur, nitrogen, phosphorous). A typical flow diagram for SWO reactor is shown in Fig. 3.

Problems of SWO

Supercritical water oxidation has been successfully tested for the oxidation of a variety of waste including radioactive waste, rocket propellants, chemical warfare agents, pulp and paper waste sludge, polymer plant waste, organic waste, and municipal waste sludge. Oxidants used in the process include pure oxygen, compressed air, hydrogen peroxide, permanganates, and other industrial oxidants. The SWO process has had some success in

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007978

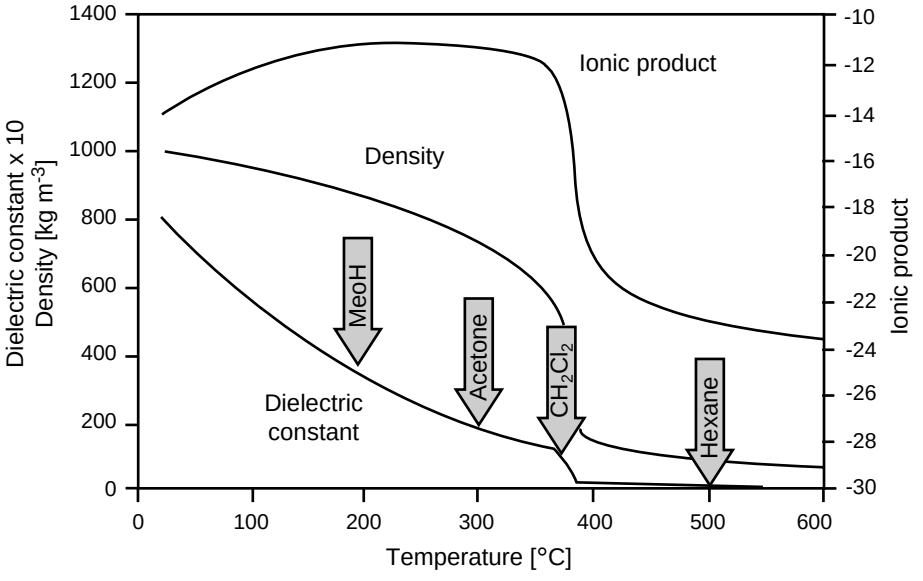


Fig. 1 Physical properties of water vs. temperature at 24 MPa. Dielectric constants of typical organic solvents at room temperature are also indicated. (From Ref.[2].)

military applications and some commercial installations. The lack of widespread industrial adaptation can be attributed to three major factors:[1,2]

1. Severe corrosion problems due to the formation of acids when elements such as S, P, Cl are present in the waste streams.
2. Serious plugging of the reactor, valves, and pipes caused by the precipitation of salts. Most waste streams contain salts that are not soluble in supercritical water.
3. Difficulty of scale-up and cost evaluation due to the lack of data on industrial scale SWO plants.

The corrosion of the heat exchanger is sometime more than that of the reactor, owing to intermediate temperatures in the heat exchanger. To avoid acid formation, a base can be added with the feed stream. The acids and base combine to yield salts, which then precipitate out causing the plugging problem.

Supercritical water oxidation poses a unique corrosion problem that has not been experienced in other chemical processes. Development of new materials is needed to address the problem. The material needs to withstand a chemically harsh environment along with the high temperature and pressure conditions. Even if some current materials are available for

Table 1 Comparison of ambient and supercritical waters

	Ambient water	Supercritical water
Dielectric constant	78	<5
Solubility of organic compounds	Very low	Fully miscible
Solubility of oxygen	6 ppm	Fully miscible
Solubility of inorganic compounds	Very high	~0
Diffusivity (cm^2 sec)	10^{-5}	10^{-3}
Viscosity (g/cm/sec)	10^{-2}	10^{-4}
Density (g/cm^3)	1	0.2–0.9

Table 2 Comparison of WAO, SWO, and incineration for waste oxidation

	Temperature (°C)	Pressure (bar)	Reaction time (min)	Efficiency (%)
WAO	150–325	20–200	15–120	20–97
SWO	380–650	220–500	1–2	60–99.9+
Incineration	490–1200	1	<1	75–99.9+

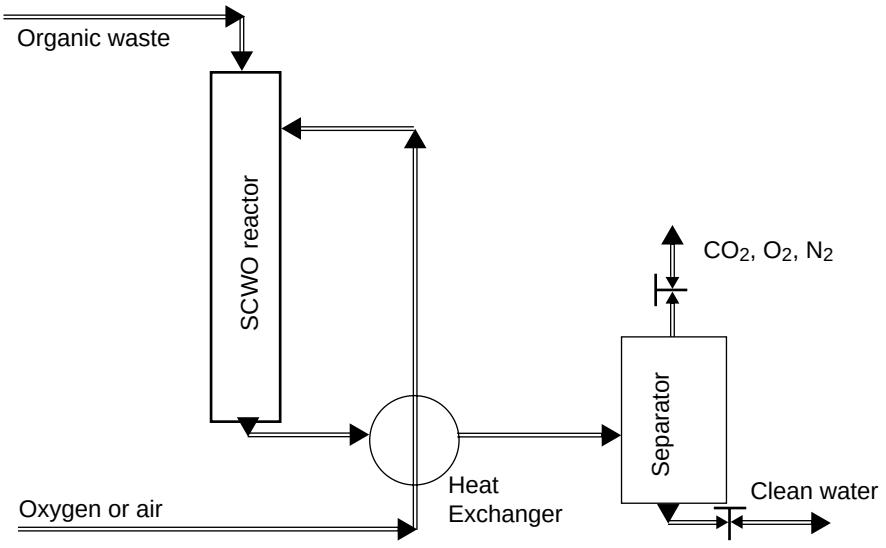


Fig. 2 Flow diagram for SWO process.

resisting one corrosive species, they fail in the presence of other corrosive species or mixtures of species. One such example is titanium, which is resistant to HCl but is prone to corrosion in the presence of sulfuric and phosphoric acids. Another aspect of the corrosion is study of the nature of the corrosive species under different processing conditions. At subcritical conditions, corrosive species exhibit extreme pH values resulting in a highly corrosive environment. Supercritical water with a lower density and dielectric constant reduces the extent of ionic species present in the system, thus reducing the extreme pHs. This guides the process condition to go far away from the critical point to avoid severe corrosion.

Different reactor concepts have evolved to meet the challenges posed by the highly reactive nature of supercritical water, as described below.

Reverse-flow tank reactor

The reverse-flow tank reactor is based on the use of a high-pressure tank with hot and cold zones (Fig. 4).[3] The high-pressure feed is injected downward into the center region where oxidation of the organics and precipitation of salts occur. The produced gases move up and leave through the exit at the top. The lower section is cooled by injecting cold water, and hence remains in the liquid phase in which salts dissolve. The salts are easily removed off the reactor as brine from the bottom outlet. With the reverse-flow tank reactor, oxidation of waste rich in salts is easier.

Transpiring wall reactor

The transpiring wall reactor addresses problems of both corrosion and salt plugging.[4] In this reactor, the platelet liner uniformly meters the flow of water to protect the liner from salt deposition and corrosion while providing a thermal and corrosion barrier for the pressure vessel (Fig. 5). This allows for higher reaction zone operating temperatures and shorter residence times. The reactor pressure boundary is exposed to controlled-temperature deionized water, resulting in a safer design. The platelet technology was borrowed from the cooling of high-pressure rocket engines from military and aerospace applications.

Microparticulate catalytic reactor

Keeping in mind the low cost of the tube reactor, Muthukumaran and Gupta have designed a reactor concept that uses tube reactor while addressing both corrosion and salt plugging concerns.[5] In this SWO reactor, in situ microparticles of a benign substance (e.g., sodium carbonate or other noncorrosive, nonsticky substances) are created so that the surface area of the microparticles is much greater than that of the reactor wall. The corrosive species generated in SWO deposit on the

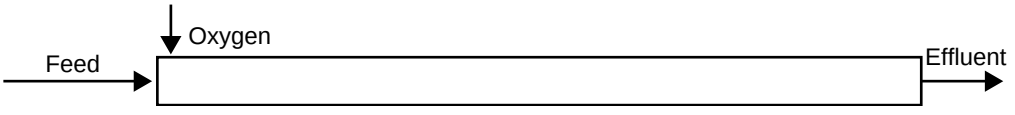


Fig. 3 Tube SWO reactor.

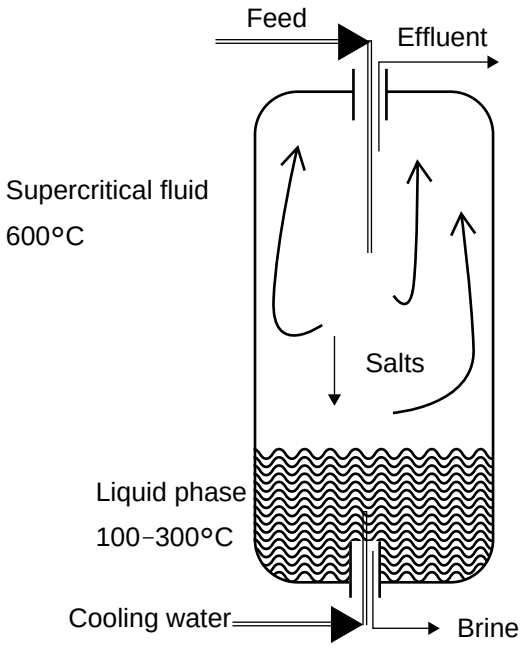


Fig. 4 Reverse-flow tank reactor with provision for the removal of salts.

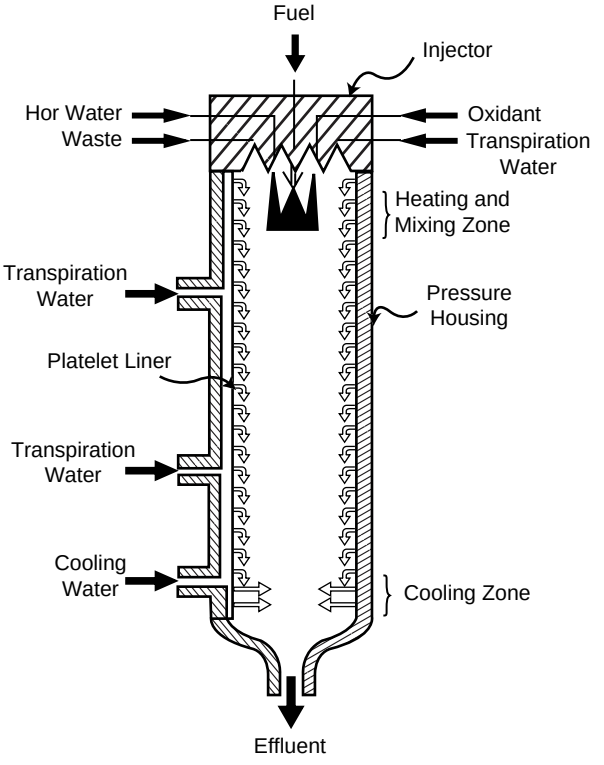


Fig. 5 Transpiring wall reactor. (From Ref.[4].)

microparticle surface and not the reactor wall. Thus, the reactor wall is protected. A schematic of the microparticulate reactor concept is shown in Fig. 6

Sodium carbonate is soluble in ambient water but is insoluble in supercritical water; the microparticles are generated as a result of the antisolvent effect of supercritical water. Because the surface area per unit mass for sodium carbonate microparticles is high, only a small amount of sodium carbonate needs to be injected in the reactor. For example, only 0.5 wt% sodium carbonate is able to provide a microparticle surface area about 100-fold that of the reactor wall.

In addition to protection against corrosion, the microparticles can also be highly catalytic for the oxidation of organics. The reaction time needed for complete oxidation can be reduced several fold. The significant catalytic activity is due to the newly generated irregular surface on the sodium carbonate microparticles. The organic molecules attach to the surface and become more susceptible to oxidation.

Additional reactor designs and flow configurations are described in the recent review paper by Marrone et al.[6]

Partial Oxidation in Supercritical Water

Complete oxidation in supercritical water occurs at a high rate; usually, complete conversion to carbon dioxide is achieved in the reaction time of about 1 min for most organics. New applications of supercritical water are emerging where only partial oxidation is carried out and oxygenated hydrocarbons are produced. Here, the reaction time has to be limited to the order of seconds.

Richter and Vogel oxidized cyclohexane in supercritical water for a varying reaction time from 5 to 80 sec range.[7] The partial oxidation resulted in the desired cyclohexanone product, but the yield was very low (Fig. 7). The by-products included carboxylic acids, carbon monoxide, and carbon dioxide.

The higher oxygenate by-products (carboxylic acids, CO, CO_2) are not desired; these were formed as a result of too much oxidation. These compounds were formed for even the lowest reaction time of 5 sec used in this

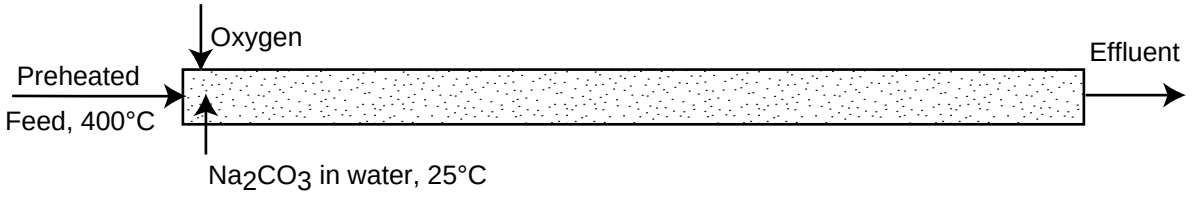


Fig. 6 Microparticulate catalytic reactor. (From Ref. [5].)

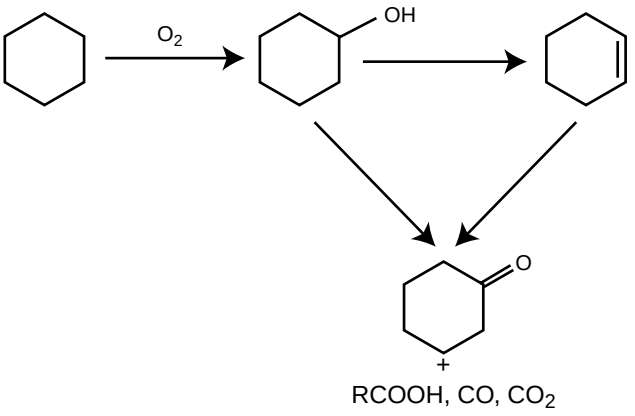


Fig. 7 Partial oxidation of cyclohexane in supercritical water.

study. Hence, even lower reaction times are required to limit the oxidation.

For the reaction time of less than 1 sec, a special quick-heating quick-quenching reactor (SWQH-QQ) was developed by Ikushima et al., as shown in Fig. 8.[8] The organic stream is mixed with the preheated water stream using a T-joint so that the process stream is heated in merely 0.05 sec. The reactor is composed of a small tubing of 50 μL volume. After the reaction the mixture is again quickly quenched to stop the reaction.

This reactor was tested for the production of ε-caprolactam from cyclohexanone-oxime (Fig. 9).

The various reaction conditions explored are listed in Table 3, where these are compared with the conventional techniques. Almost complete selectivity and yield were obtained when the oxidation was carried out for 0.7 sec in supercritical water containing 0.04–0.9 mM acid catalyst.

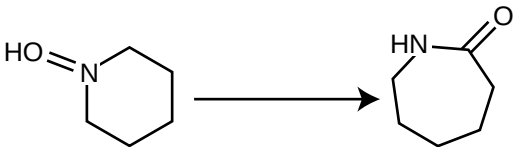


Fig. 9 Conversion of cyclohexanone-oxime to ε-caprolactam.

For commercial scale production, one would need to employ an array of microreactors. For example, a 50 mL microreactor can produce 0.25 kg of product per day, whereas an array of 1000 microreactors can produce 0.25 tons of product per day.

Safety and Health Issues

When dealing with supercritical water, there are two safety and health issues that are to be kept in mind when designing and operating the reactor: 1) the high pressure involved requires that personnel is protected from the plant by proper isolating walls and 2) the high temperature involved requires proper heat protection. Because of the concern of corrosion, reactor rupture is a possibility. Effluent water from the reactor should be checked for the presence of metal ions for on-line diagnostics of corrosion. In addition, periodic shut down and maintenance is very important. Additional safety concerns are due to the presence of high-pressure air or oxygen.

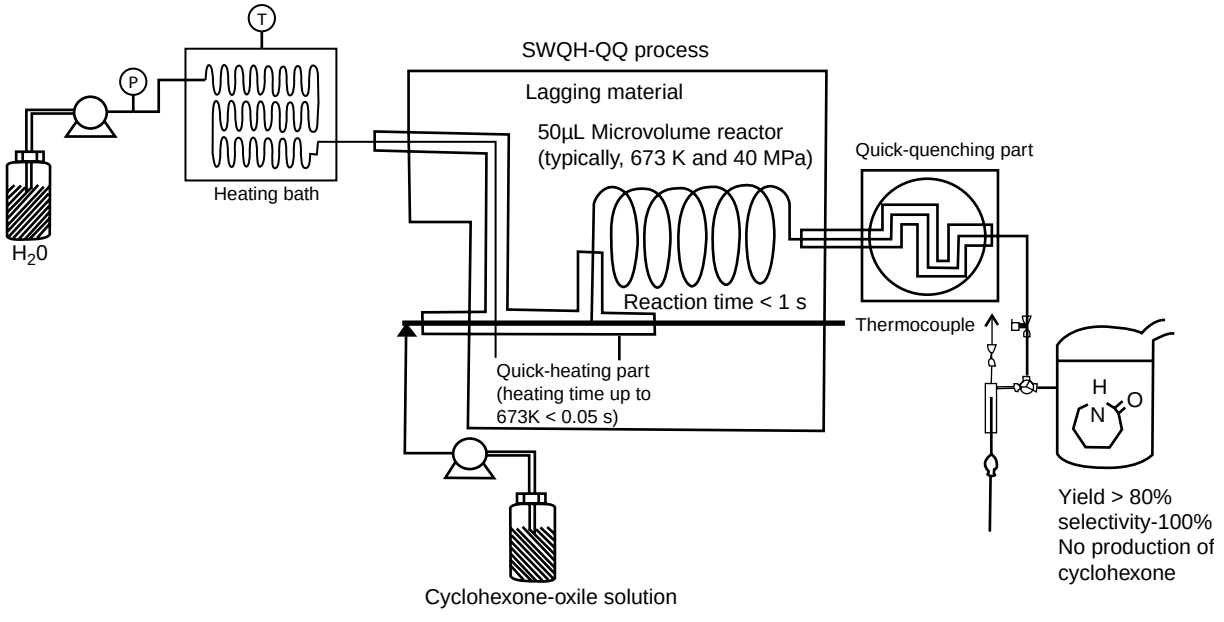


Fig. 8 Quick-heating quick-quenching supercritical oxidation reactor. (From Ref.[8].)

Table 3 Production of ε-caprolactam from cyclohexanone-oxime via various techniques

Reactor	T (°C)	P (bar)	Reaction time (sec)	Selectivity (%)	Yield (%)
Conventional					
7.7 M concentration H_2SO_4	110	1	5400		72.0
6.2 M concentration H_2SO_4	134	1	8160		96.0
$B_2O_3(20\%)/Al_2O_3$	200	0.003	6	75.0	72.0
High-silica MFI zeolite	350	0.1	3600	95.7	95.3
Subcritical					
Without QH-QQ	250	400	180	0	0
With QH-QQ	300	400	0.9	100	9.5
With QH-QQ	350	400	0.8	99.6	38.9
Supercritical					
Without QH-QQ	400	400	180	99.3	9.5
With QH-QQ	375	400	0.7	98.6	80.0
With QH-QQ	400	400	0.6	98.6	83.0
With QH-QQ	420	400	0.5	99.2	42.1
Supercritical + catalyst					
HCl with QH-QQ	375	400	0.7	99.3	99.3
H_2SO_4 with QH-QQ	375	400	0.7	99.5	99.5

CONCLUSIONS

Supercritical water offers a unique reaction environment. Most organic compounds and oxygen are fully miscible in supercritical water. Use of excess oxidant and about a minute of reaction time can usually cause complete oxidation of the organics to carbon dioxide. Hence, SWO is emerging as an alternative technology to incineration for waste treatment from a variety of sources. However, current challenges are due to reactor corrosion and plugging when acid forming compounds are present in the feed. If the oxidation is carried out with limited oxygen or for a limited time (say milliseconds), intermediate oxygenated products are obtained. The partial oxidation can be used to produce the oxygenated products.

REFERENCES

1. Hodes, M.; Marrone, P.A.; Hong, G.T.; Smith, K.A.; Tester, J.W. Salt precipitation and scale control in supercritical water oxidation—part A: fundamentals and research. J. Supercrit. Fluids **2004**, *29* (3), 265–288.
2. Kritzer, P.; Dinjus, E. An assessment of supercritical water oxidation (SCWO). Existing problems, possible solutions and new reactor concepts. Chem. Eng. J. (Amsterdam, Netherlands) **2001**, *83* (3), 207–214.
3. Swallow, K.C.; Killilea, W.R.; Malinowski, K.C; Staszak, C.N. The Modar process for the destruction of hazardous organic wastes—field test of a pilot-scale unit. Waste Manag. (Amsterdam, Netherlands) **1989**, *9* (1), 19–26.
4. Crooker, P.J.; Ahluwalia, K.S.; Fan, Z.; Prince, J. Operating results from supercritical water oxidation plants. Ind. Eng. Chem. Res. **2000**, *39* (12), 4865–4870.
5. Muthukumaran, P.; Gupta, R.E. Sodium-carbonate-assisted supercritical water oxidation of chlorinated waste. Ind. Eng. Chem. Res. **2000**, *39* (12), 4555–4563.
6. Marrone, P.A.; Hodes, M.; Smith, K.A.; Tester, J.W. Salt precipitation and scale control in supercritical water oxidation—part B: commercial/full-scale applications. J. Supercrit. Fluids **2004**, *29* (3), 289–312.
7. Richter, T.; Vogel, H. The partial oxidation of cyclohexane in supercritical water. Chem. Eng. Technol. **2002**, *25* (3), 265–268.
8. Ikushima, Y.; Sato, O.; Sato, M.; Hatakeda, R.; Arai, M. Innovations in chemical reaction processes using supercritical water: an environmental application to the production of ε-caprolactam. Chem. Eng. Sci. **2003**, *58* (3–6), 935–941.

Synthesis Gas

Kim Aasberg-Petersen
Haldor Topsøe A/S, Lyngby, Denmark

INTRODUCTION

Synthesis gas (syngas) is a gaseous mixture containing mainly hydrogen, carbon monoxide, and carbon dioxide in various amounts. In most cases, these three compounds constitute more than 90% of the syngas, but other components including methane and inert gases such as nitrogen and argon are often present in the mixture. Syngas is only seldom a product by itself, and its major role is as a key intermediate in the synthesis of a range of chemicals. Syngas may in principle be produced from any carbon-containing feedstock, but today only coal and hydrocarbons are used as raw materials to any significant extent. New plants are normally designed based on natural gas or other hydrocarbon feedstock, because the investment is only about one-third of that of a plant based on coal. The production of synthesis gas depends upon the type of feedstock and the desired end product.

Today, synthesis gas is mainly used for the production of ammonia (120×10^6 t/yr[a]) and methanol (30×10^6 t/yr[a]) followed by pure hydrogen for hydrotreating in refineries. Other current applications are in the production of higher alcohols by hydroformulation and a number of products including acetic acid, formaldehyde, dimethyl ether (DME), and methyl-*tert*-butyl ether (MTBE); in all cases methanol is used as a coreactant.

In recent years, new areas have emerged that may open the way for an increased production and use of synthesis gas. One example is the large-scale production of syngas using synthetic hydrocarbon fuels, produced by the Fischer–Tropsch synthesis. Similarly, a considerable effort is currently being undertaken in the development and commercialization of various types of fuel cells for both small- and large-scale power production. A successful market introduction will require efficient and cost effective technologies for converting the hydrocarbon feedstock into a syngas that is acceptable for the fuel cell. This entry describes the applications of syngas and focuses on the technologies for its production using hydrocarbon feedstock. The role of the catalysts and key parts of the technologies is outlined and the reasons for deactivation are emphasized. The main criteria for selection of the optimum route for syngas production for various applications are also described.

REQUIREMENTS TO SYNGAS PROPERTIES

The capital cost of the synthesis gas often accounts for more than 50% of the cost of the complete plant. The optimal choice of technology depends upon the size of the plant and the desired syngas composition for the subsequent synthesis. In Table 1, the optimum syngas composition for a range of final products is given.[1]

For most of the chemical products given in Table 1, the synthesis takes place at elevated pressures from typically 20 bar and above. Most syngas plants operate at 20–40 bar, although the thermodynamics shows that higher conversions can be obtained at lower pressures. However, the cost associated with larger equipment and syngas compression renders often syngas manufacture at low pressures prohibitive for these applications.

For fuel cell applications, the situation is often the reverse. With power as the end product, the compression of air should be minimized, especially in (small) plants in which coupling with turbine systems is not economical. The requirement for the syngas production in this case is a simple and low-cost design with a low pressure drop.

SYNGAS MANUFACTURE

The production of syngas from natural gas and other hydrocarbons takes place by steam reforming, partial oxidation, or combinations thereof.

Steam Reforming

Steam reforming is highly endothermic, involving the reactions listed in Table 2.[2] High temperatures and the addition of a significant amount of steam are needed to obtain maximum conversion as illustrated in Fig. 1.[1] Steam reforming is always accompanied by the shift Reaction (2) in Table 2, which is generally fast and can be considered in equilibrium at most conditions. Steam can be partially substituted for CO_2 to reduce the (H_2/CO) ratio in the product gas. This effect is illustrated in Fig. 2, which also shows that the methane conversion is almost unaffected as long as the $(H_2O + CO_2)/CH_4$ ratio is constant.

A considerable amount of heat input is needed to achieve sufficient conversion of methane at high

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007980

Table 1 Optimum syngas compositions[a]

Process	Optimum composition	Coreactants
Ammonia	$\frac{H_2}{N_2} = 3$	
Methanol	$M = \frac{H_2 - CO_2}{CO + CO_2} \cong 2$	
DME	$M = \frac{H_2 - CO_2}{CO + CO_2} \cong 2$	
High-temperature Fischer–Tropsch	$M = \frac{H_2 - CO_2}{CO + CO_2} \cong 2$	
Low-temperature Fischer–Tropsch	$\frac{H_2}{CO} \approx 2$	
Acetic acid	CO	Methanol
Higher alcohols	$\frac{H_2}{CO} = 1$	Olefins
Industrial hydrogen	99.99 H_2	
Hydrogen for PEFC[b]	<50 ppm CO	
Reducing gas (iron ore)	$\frac{H_2O + CO_2}{H_2 + CO + CO_2 + H_2O} \leq 0.05$	
Solid oxide fuel cell	H_2, CH_4, H_2O; 0% C_nH_m ($n \geq 2$)	

[a]Syngas composition for SOFC not included in Ref.[1].
[b]Polymer electrolyte fuel cell.

temperatures. As an example, above 10 MJ/Nm^3, CH_4 is needed for converting a feed of methane at 600°C and steam ($S/CH_4 = 2.5$) to its equilibrium composition at 900°C at 25 bar. Normally, the reaction takes place in high alloy tubes filled with nickel catalysts placed in a furnace. The heat is supplied by a series of burners placed on the walls of the furnace. The furnaces are designed with a variety of burner arrangements as illustrated in Fig. 3.[3] In industry today, the top-fired and side-fired reformers are, in most cases, used for new plants. The industrial reformer consists of a box-type radiant section including the burners and a convection zone to recover the waste heat of the flue gas that is leaving the radiant section.

The reformer tubes typically have a length of 10–13 m and outer diameters that do not exceed 15 cm. The tubes are mostly designed for maximum heat transfer taking into account the mechanical limitations when considering the heat flux and temperature. Tubular reformers are today designed for an average heat input exceeding 100,000 $kcal/m^2/hr$.[1]

The typical inlet temperatures to the reformer are 450–650°C and the exit temperatures typically range from 700°C to 950°C, depending on the type of feedstock and the application. A tubular reformer can be designed with capacities up to the equivalence of 300,000 Nm^3/hr of syngas.[1]

In general, the gas leaving a steam-reforming reactor is close to chemical equilibrium for Reaction (1) in Table 2. In industry, the approach to equilibrium at the outlet of the reformer tubes is expressed by a temperature difference defined by:

$$\Delta T_R = T(\text{exit catalyst}) - T(Q_R) \quad (5)$$

Table 2 Steam reforming reactions

Reaction no.	Steam reforming	$-\Delta H^0_{298}$ (kJ/mol)	$\ln K_p = A + B/T$[a]	
			A	*B*
1	$CH_4 + H_2O \rightarrow CO + 3H_2$	−206	30.420	−27106
2	$CO + H_2O \rightarrow CO_2 + H_2$	41	−3.798	4160
3	$CH_4 + CO_2 \rightarrow 2CO + 2H_2$	−247	34.218	−31266
4	$C_nH_m + nH_2O \rightarrow nCO + (n + \frac{m}{2})H_2$	−1175[b]		

[a]Standard state: 298 K, 1 bar.
[b]For *n*-heptane.
(From Ref.[2].)

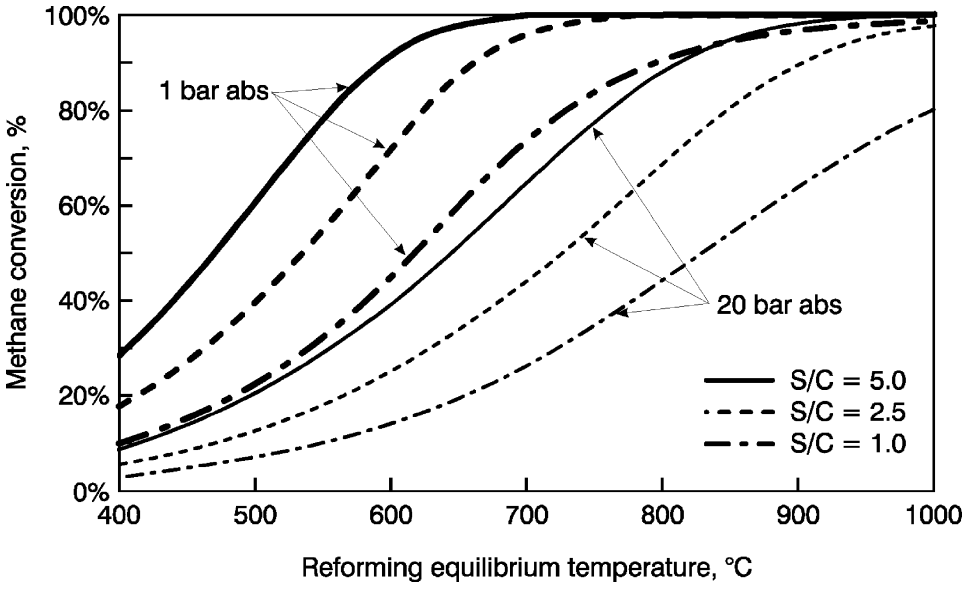


Fig. 1 Equilibrium methane conversion at various H_2O/CH_4 (S/C) ratios at 20 bar abs. (From Ref.[1].)

$T(Q_R)$ is the equilibrium temperature for the reaction corresponding to the reaction quotient Q_R, calculated by:

$$Q_R = \frac{P_{CO}\, P_{H_2O}^3}{P_{CH_4}\, P_{H_2O}}$$

P is the partial pressure of the gases at the reformer exit. T_R is usually less than 10–15°C throughout the catalyst lifetime. The thermal efficiency of a tubular steam reformer with waste heat recovery system approaches 95% of which 50–60% is transferred to the process.[4] The remaining heat is recovered by production of steam, preheating of air, feedstock, etc.

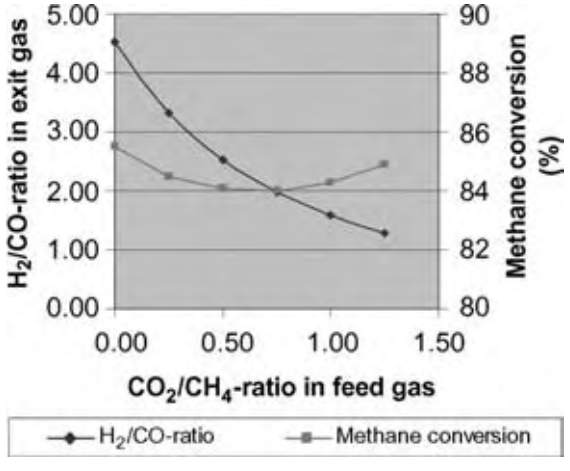


Fig. 2 Methane conversion and H_2/CO-ratios in combined CO_2- and steam reforming. $\frac{H_2O + CO_2}{CH_4} = 2.5$. *(View this art in color at www.dekker.com.)*

Other types of more compact steam reformers are also used. In most cases, the heat transfer is then accomplished by convection with the reformed gas itself, a flue gas, or by a combination as illustrated in Fig. 4.[5] In this case, about 80% of the heat is transferred to the process and export of steam from the plant can be reduced or avoided.

Another type of steam-reforming reactor that is attracting increasing attention is known as gas heated reformers or heat exchange reformers. In such reformers, heat is transferred by convection and the heat source is a hot process gas from another reformer or a partial oxidation reactor. A number of different installations of heat exchange reformers can be envisaged. In Fig. 5, the installation of a heat exchange reformer either in series or in parallel to an autothermal reformer (ATR) is illustrated.

In the series arrangement, all the feed pass through both the reformers. Both the amount of gas and the driving force for heat transfer are highest with a series scheme. The feed is split into two parts in the parallel arrangement, allowing the possibility of using different inlet gas compositions. The heating gas and the gas leaving the catalyst section are mixed in the scheme illustrated in Fig. 5. In principle, a parallel scheme without mixing the two gasses is also possible.

The use of heat exchange reformers decreases the steam export and the consumption of feed and fuel, in plants using only steam reforming. The efficiency is increased and the oxygen consumption is reduced in plants with a combination of partial oxidation and heat exchange reforming. As an example, process simulations show that the efficiency of a syngas generation unit in a plant for production of synthesis fuel increases approximately 5% by introducing a heat

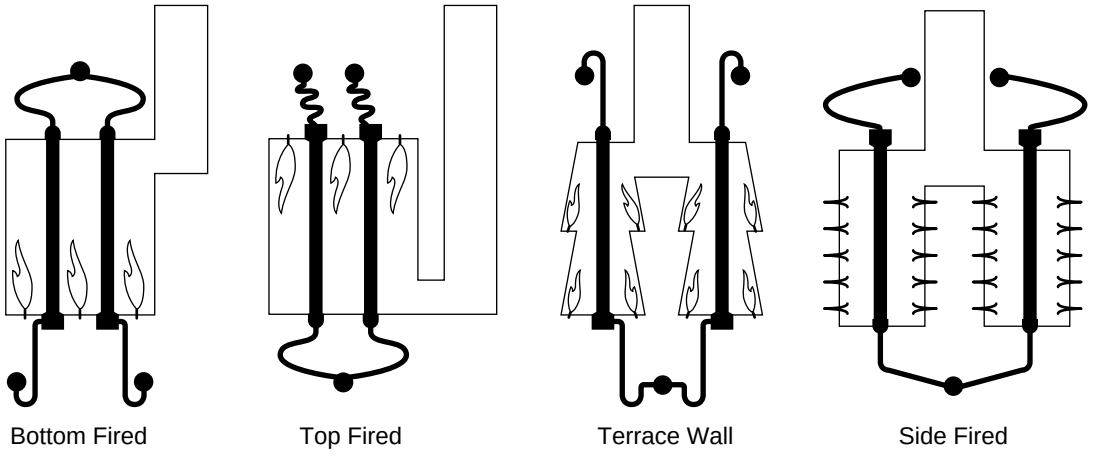


Fig. 3 Typical configurations of reformer furnaces. (From Ref.[3].)

exchange reformer in combination with autothermal reforming.[27] This combination also reduces the required amount of oxygen by 5–10% as compared to autothermal reforming alone.

A typical layout of the steam-reforming section of a syngas plant with hydrocarbon feedstock is illustrated in Fig. 6. The first step is purification of the feedstock to remove sulfur so as to avoid poisoning of the downstream reformer catalysts. This is typically accomplished in a two-step process. In the first step, organic sulfur compounds are converted into hydrogen sulfide by a hydrogenation catalyst. In the second step, H_2S is absorbed by zinc oxide by the following reaction:

$$ZnO + H_2S \Leftrightarrow ZnS + H_2O \quad (6)$$

The desulfurization typically takes place at temperatures between 200°C and 400°C, depending on the sulfur compounds of the feed. Carbon dioxide in the feed may react with hydrogen to generate steam, which may push the equilibrium of Reaction (6) to the left and increases the sulfur leakage. This must be considered in the design of the desulfurization section, and special methods exist to eliminate sulfur leakages in this case also.

In many (but not all) situations, an adiabatic prereformer is located upstream the main reformer to convert the higher hydrocarbons (hydrocarbon compounds with two or more carbon atoms) into an equilibrated mixture of carbon oxides, methane, hydrogen, and steam. Prereforming takes place at temperatures between 350°C and 550°C, depending on the feedstock. The prereforming step allows the use of higher preheat temperatures for the feed in the tubular reformer without the risk of cracking of the higher hydrocarbons, which may result in carbon formation on the catalyst and/or cause fouling of heat exchangers. A higher preheat temperature increases the energy efficiency and reduces the size of the steam reformer. Prereforming of LPG, naphtha, and refinery off-gas is practiced in the industry, allowing significant feedstock flexibility. The feasibility of prereforming of logistic fuels of interest to the fuel cell market has been demonstrated in pilot plant tests.[8]

The heat in the exit gas is recovered downstream the tubular reformer, usually by the production of steam, preheating of boiler feed water, etc. The final separation into the desired product compositions depends on the application. Pressure swing adsorption (PSA) is in most cases used if pure hydrogen is desired. Pure carbon monoxide can be obtained by cryogenic separation in a cold box. Adjustment of the H_2/CO ratio can be accomplished by polymer membrane modules with different selectivities for permeation of the two compounds.

Steam-Reforming Catalysts

Most steam-reforming catalysts are based on nickel as the active material. Also, cobalt and noble metals catalyze the steam-reforming reaction, but they are generally too expensive to find widespread use. A number of different carriers including alumina, magnesium–aluminum spinel, zirconia, and calcium aluminate are employed.

The activity of the catalyst is roughly proportional to the surface area of nickel used. In an adiabatic prereformer, a high activity is desired to maximize the space velocity. In a tubular reformer the activity may be of less significance because the reactor volume is settled by mechanical criteria. Most industrial tubular reformers operate at space velocities of 2000–4000 hr^{-1}. However, equilibrium conversion can be achieved at much higher space velocities, exceeding 10^4–10^5 $Nm^3/m^3/hr$, as determined by extrapolation

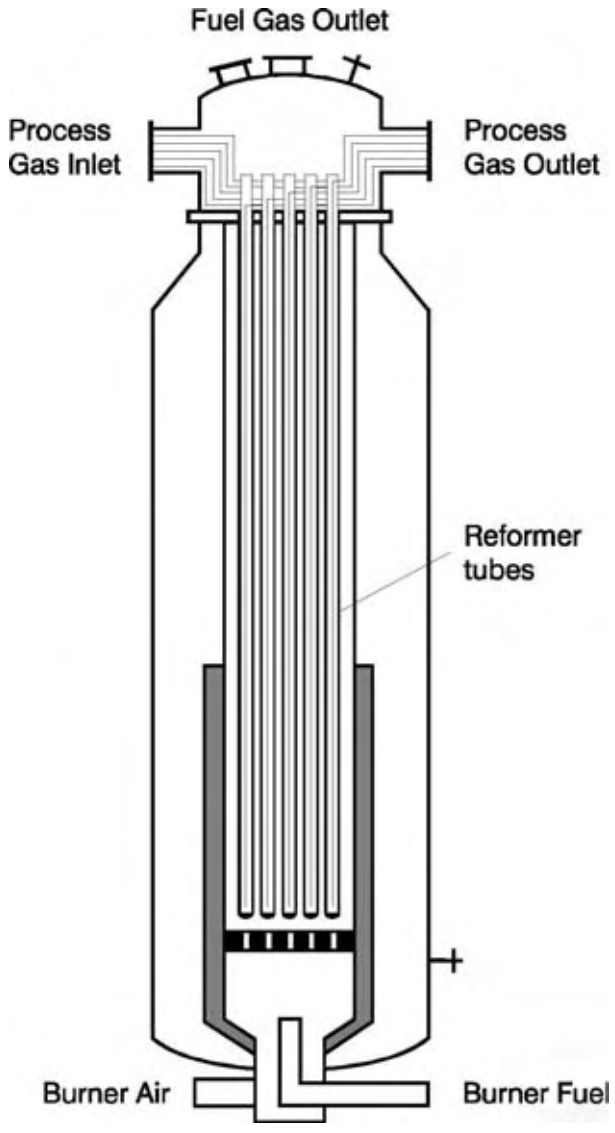


Fig. 4 Schematic illustration of a compact convection reformer. (From Ref.[5].)

of intrinsic rates.[1] Higher intrinsic rates can be obtained with noble metals where especially rhodium and ruthenium exhibit very high activities.[9]

Higher hydrocarbons are more reactive than methane, with aromatics showing the lowest reactivity as illustrated in Fig. 7.

Catalyst Deactivation

Steam-reforming catalysts may deactivate because of sintering, poisoning, or by carbon formation.

The mechanism for sintering is migration and coalescence of nickel particles on the carrier surface, leading to a smaller surface area.[11] Sintering is a complex process influenced by several parameters including chemical environment, catalyst structure and composition, and support morphology. Factors that enhance sintering include high temperature and high steam partial pressure.[11]

Steam-reforming catalysts are susceptible to sulfur poisoning. At reforming conditions, all sulfur compounds are converted to hydrogen sulfide, which is chemisorbed on the metallic surface

$$H_2S + Ni \leftrightarrows Ni - S + H_2 \quad (7)$$

This surface layer has a structure like a two-dimensional sulfide. The adsorption takes place well below the H_2S/H_2 ratio that is needed to form bulk nickel sulfides.

With an H_2S/H_2 ratio in the gas of 1 ppb, the equilibrium surface coverage of nickel at 500°C is around 70%. This means that all sulfur in the feed is quantitatively adsorbed on the nickel catalyst of a prereformer. The result is not only the deactivation of the prereformer catalyst even at very low sulfur levels, but also the protection of downstream catalysts from poisoning.[13] Sulfur uptake on the catalyst will initially take place as shell poisoning and because of pore diffusion restrictions, it may take years before sulfur reaches the center of the particle.[12]

Other components that poison the reformer catalysts include silica and alkali metals. Silica may act by blocking the pore mouth of the catalyst. The alkali metals reduce the turnover frequency of the catalyst.[12]

Steam reforming involves the risk of carbon formation especially when operating at low steam-to-carbon (S/C) ratios to decrease the H_2/CO ratio. Carbon may be formed from higher hydrocarbons, methane, or carbon monoxide as illustrated in Table 3.

A typical mechanism for carbon formation is the formation of whisker structures on the catalyst. Adsorbed hydrocarbons may react and form adsorbed carbon, which is dissolved in the nickel particle. After saturation, carbon may nucleate and grow with the nickel crystal at the tip.[13]

The whisker carbon has a higher energy than graphite.[13] Hence, operation in conditions at which thermodynamics would predict the formation of graphite is feasible without carbon formation on the catalyst. In Table 3, the equilibrium constant for whisker carbon formation from methane on a nickel catalyst is given. The equilibrium constant depends on the size of the nickel crystals, with smaller sizes resulting in lower values. The risk of whisker carbon formation from methane at equilibrium may be assessed by the so-called ''principle of equilibrated gas.'' According to this principle, carbon will form if there is a potential for carbon formation [according to Reaction (8) in Table 3] after equilibration of the gas mixture at the given temperature and pressure.[13] It can be shown that this thermodynamic carbon

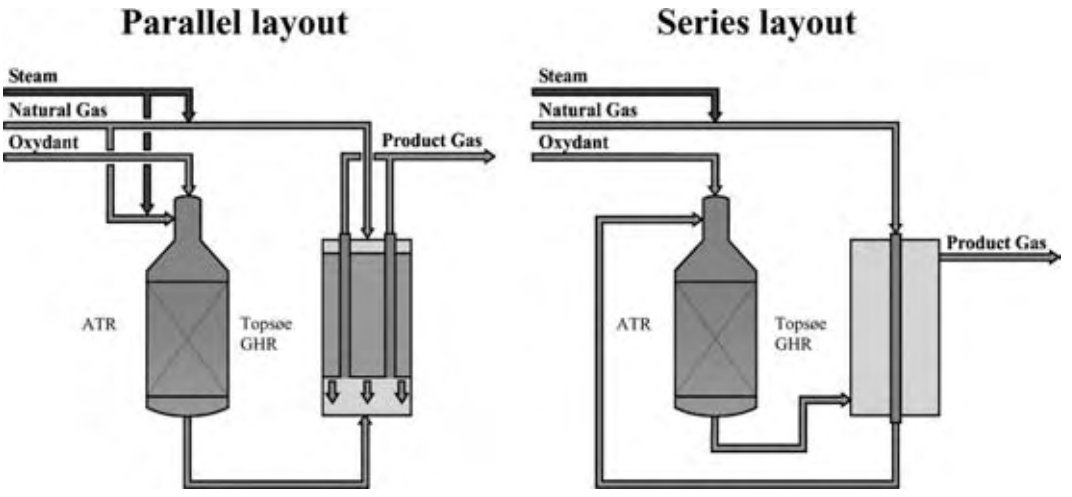


Fig. 5 Series and parallel arrangement of heat exchange reformers with ATR as the main reformer. (From Ref.[27].) *(View this art in color at www.dekker.com.)*

limit is a function of pressure, temperature, and the feed gas composition expressed as atomic ratios (O/C and H/C).[10]

Lower equilibrium constants for carbon formation from methane can be obtained by using noble metals.[9]

The irreversible formation of carbon from higher hydrocarbons on the catalyst may be in the form of either a type of gum layer or a whisker carbon. Gum formation is polymerization of higher hydrocarbons adsorbed on the catalyst surface into an encapsulating film on the nickel surface.[12] Low temperature, highly aromatic feedstock, and low amount of steam and hydrogen compared to the content of higher hydrocarbons are among the factors known to enhance the risk of gum formation. Deactivation by gum formation is mainly relevant for prereforming, but may be very fast especially at low temperatures.[7,12]

Higher hydrocarbons do not exist at equilibrium and any risk of whisker formation from these compounds can be disregarded at these conditions. Nevertheless whiskers may still form from higher hydrocarbons because at nonequilibrium conditions a potential for the irreversible carbon formation [e.g., Reaction (11) in Table 3] may exist. The formation of whisker carbon at these conditions depends on a kinetic balance between the rate of the carbon forming and steam-reforming reactions. A simplified reaction sequence outlining the kinetic balance is shown in Fig. 8. The key step is whether the adsorbed hydrocarbon species will react to form adsorbed carbon and whiskers or react with oxygen species to produce gas.[13]

The risk of carbon formation from higher hydrocarbons depends also on the type of hydrocarbon. Olefins will result in rapid carbon formation on the

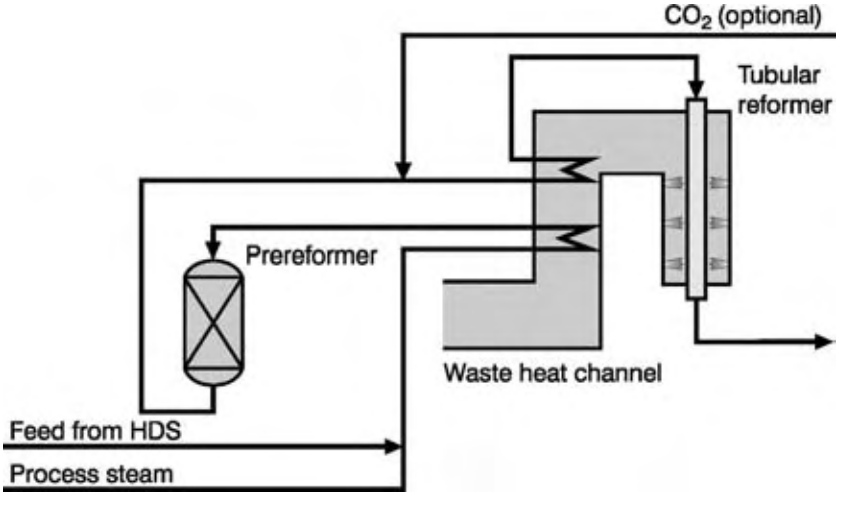


Fig. 6 Schematic layout of the steam-reforming section of a syngas plant with hydrocarbon feedstock. HDS, hydrodesulfurization. (From Ref.[7].) *(View this art in color at www.dekker.com.)*

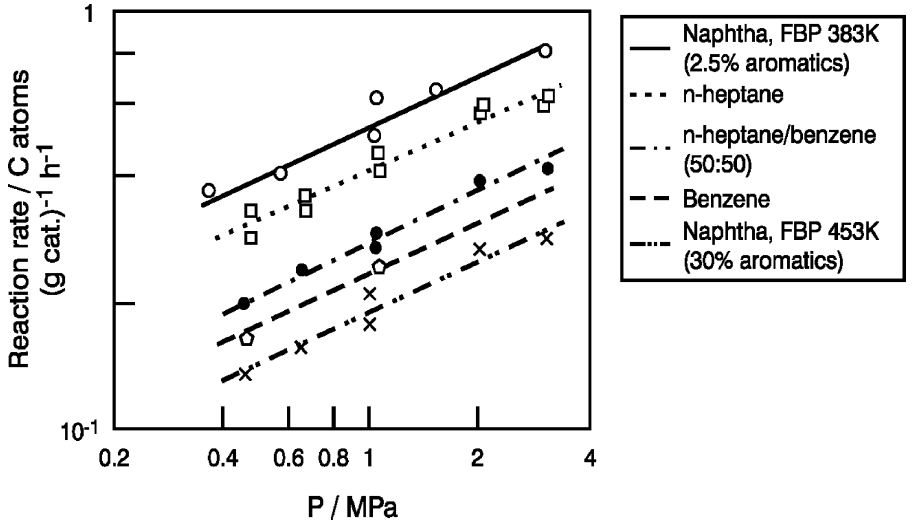


Fig. 7 Reactivity of liquid hydrocarbons. $H_2O/C = 4$ mol/mol C, 500°C, 20 bar. (From Ref.[10].)

catalyst, and aromatic compounds are in general more critical than paraffins.[10,13] Operation at low S/C ratios and at high temperatures increases the risk of whisker carbon formation from higher hydrocarbons. As indicated above, the risk of carbon formation must be assessed by a complex analysis involving the kinetics of both the carbon forming and the steam-reforming reactions.

It is imperative that carbon formation is avoided in tubular reformers. Carbon formation may lead to the breakdown of catalysts resulting in an uneven flow distribution between different tubes in the reformer, causing local overheating and shorter tube life. Formation of pyrolytic carbon may result in carbon deposits near the tube wall. This may lead to a reduced heat transfer coefficient and development of ''hot bands'' as evidenced by the reddish zone on the tube wall during operation.

One method that is used to prevent carbon formation is to promote the steam-reforming catalyst with alkali, usually potassium.[14] Alkali serves to increase the resistance to carbon formation on the catalyst and promotes the carbon (re)gasification.[14]

Carbon formation from carbon monoxide on the catalyst may take place according to Reaction (9) by a route similar to methane. However, decomposition of carbon monoxide may also occur on the surface of equipments such as heat exchangers downstream the tubular reformer (or the partial oxidation reactor). A given gas will have potential for carbon formation from carbon monoxide below a certain temperature by either Reaction (9) or Reaction (10) in Table 3. In this case, carburization of the metallic components may occur leading to a severe type of corrosion known as metal dusting. Metal dusting is, at least in its initial stages, often characterized by a series of spot attacks known as pits. Metal dusting may be very rapid and may lead to quick failure of equipment. Irrespective of the thermodynamic potential, metal dusting will not occur if the metal surface temperature is low enough.

In heat exchange reformers, special precautions must be taken to avoid corrosion. In such reformers the metal temperature is often high enough for metal dusting to occur and low enough for thermodynamic potential to exist. Special alloys or other techniques are needed to prevent metal dusting in these cases. This area is the subject of considerable research to widen the limits of operation to more severe operating conditions.

Table 3 Carbon forming reactions

Reaction no.	Steam reforming	$-\Delta H^0_{298}$ (kJ/mol)	$\ln K_p = A + B/T$[a]	
			A	B
8	$CH_4 \rightarrow C + 2H_2$	−75	12.68[b]	−10779[b]
9	$2CO \rightarrow C + CO_2$	172	−21.08	20486
10	$CO + H_2 \rightarrow C + H_2O$	131	−17.28	16326
11	$C_nH_m \rightarrow nC + {}^1/_2 mH_2$			

[a]Standard state: 298 K, 1 bar.
[b]Whisker carbon.
(From Ref.[2].)

$C_nH_m + 2*$	$\xrightarrow{k_A}$	$C_nH_y*_2$
$C_nH_y*_2 + *$	$\xrightarrow{k_H}$	$C_{n-1}H_z*_2 + CH_X*$
CH_X*	$\xrightarrow{k_c}$	$C* \longrightarrow$ whisker carbon
$C_x + OH*$	$\xrightarrow{k_d}$	gas
$CH_x* + OH*$	$\xrightarrow{k_g}$	gas
$H_2O + *$	$\xrightarrow{Kw}$	$OH* + \frac{1}{2}H_2$

Fig. 8 Simplified reaction sequence for steam reforming and carbon whisker formation. (From Ref.[13].)

Partial Oxidation

Several processes, including noncatalytic partial oxidation (POX), catalytic partial oxidation (CPO), and ATR may be used to perform partial oxidation of hydrocarbons. In all cases, some or all of the reactions listed in Table 4 are involved. Normally, these reactions are accompanied by the steam-reforming and shift reactions. The oxidation reactions are irreversible at all conditions of practical interest.

The feed to the partial oxidation reactor is a mixture of hydrocarbons, steam, and air or oxygen (or mixtures thereof). The reactor is in general adiabatic or autothermal and the exit gas is in many cases close to equilibrium at the exit temperature and pressure at chemical equilibrium. The exit composition can be determined based on the inlet temperature and composition, and on the assumption that all oxygen has reacted. In Fig. 9, product gas compositions are given at various conditions with oxygen as oxidant, assuming that chemical equilibrium is obtained.

The overall reaction in a partial oxidation reactor is highly exothermic. The desired reactions may be accompanied by thermal cracking of hydrocarbons or oxidative dehydrogenation into nonsaturated compounds including olefins, polyaromatics, and soot. The control of the heat balance and the formation of by-products are important considerations in the design of partial oxidation reactors.

Noncatalytic Partial Oxidation

In noncatalytic partial oxidation reactors for the production of syngas, the feed is substoichiometrically combusted with oxygen in a burner designed for intensive mixing of the reactants.[15] A small amount of steam may be added to the oxygen to moderate the flame temperature.[16] Plants based on POX are and have been used for the production of hydrogen for refineries and syngas for methanol (and ammonia) synthesis. No catalyst is included in the reactor.

A key part of the POX reactor is the burner. The burner must be able to withstand the highly severe conditions in the combustion chamber. A multiorifice coannular burner with alternative passages for feed and oxidant is described in Ref.[28].

Partial oxidation units may operate with feedstock, ranging from natural gas to heavy oil fractions such as asphalt.[15] Heavy feedstock may contain large amounts of both sulfur and heavy metals. In many cases these compounds are removed from the raw syngas downstream the POX reactors. Some residual carbon or soot may be formed in the combustion chamber of the reactor.[15] To avoid the carbon lay-down in the heat exchanger downstream the POX reactor, special coils and high gas velocity are used. The carbon may be removed downstream the heat exchanger in a suitable water wash or scrubbing system.[15,16]

Table 4 Reactions occurring in partial oxidation of methane

Reaction no.		
12	$CH_4 + \frac{1}{2}O_2 \rightarrow 2H_2 + CO$	$-\Delta H^0_{298} = +38\,kJ/mol$
13	$CH_4 + O_2 \rightarrow CO + H_2O + H_2$	$-\Delta H^0_{298} = +278\,kJ/mol$
14	$CH_4 + 3/2O_2 \rightarrow CO + 2H_2O$	$-\Delta H^0_{298} = +519\,kJ/mol$
15	$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O$	$-\Delta H^0_{298} = +802\,kJ/mol$

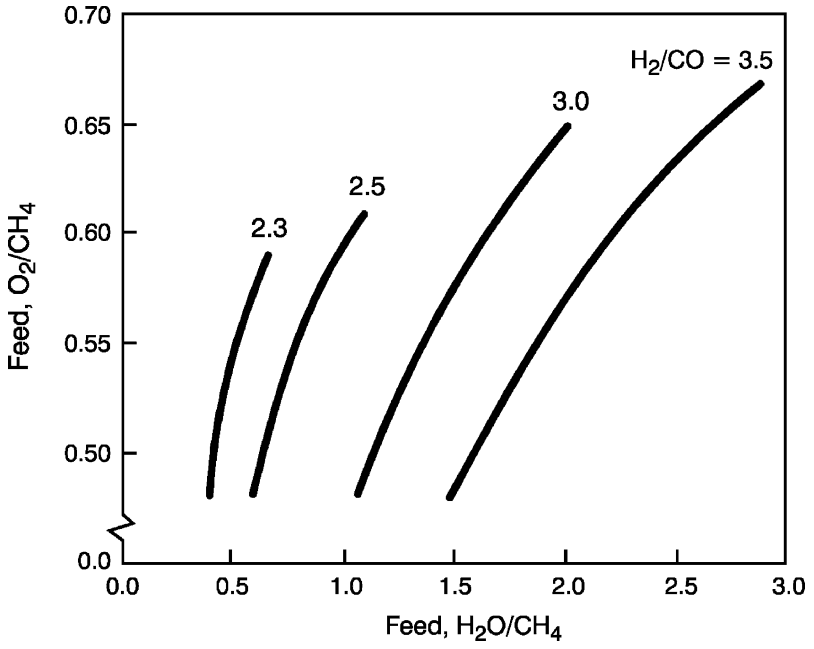


Fig. 9 H_2/CO ratios in exit gas from partial oxidation reactor. (From Ref.[22].)

The H_2/CO ratio produced by POX units is often 1.7–1.8 with natural gas as feed. Addition of steam to increase this ratio may not be feasible because the resulting lower exit temperature could substantially increase the amount of carbon formed in the reactor. In some cases the POX reactors may be combined with steam reforming in parallel if a higher H_2/CO ratio is required. This combination has been used in a plant for producing hydrocarbons in Malaysia.[29] A similar concept will be used for a large plant that is to produce synthetic hydrocarbons to be located in Qatar.

Catalytic Partial Oxidation

In catalytic partial oxidation, no burner is used. The hydrocarbon feedstock and the oxidant are mixed in an inlet zone and pass to the catalyst where the reaction takes place. Catalytic partial oxidation is today considered by many as a (potential) technology for producing syngas at very high space velocity and short contact time. As such, CPO is under development. It is considered by many as a promising technology for converting hydrocarbons into a syngas that is useful for fuel cells. Larger-scale applications may also be feasible.

A CPO reactor may operate adiabatically, and operation without preheat of the feedstock during normal operation is feasible. The reactor may be very compact, and complete conversion of oxygen has been demonstrated at atmospheric pressure with space velocities exceeding 500,000 hr^{-1}.[17] Catalytic partial oxidation can, at least with natural gas feed, be operated without steam in the feed. It has been catalytically proven that CPO at elevated inert temperature and pressure is feasible.[18,19]

The reactor is often divided into a mixing zone and a reaction zone with catalyst; both are normally refractory lined. Safety is a prime concern in the design of CPO reactors. The mixing of the hydrocarbon feed with air or oxygen involves the risk of thermal ignition and even explosion. Ignition may for example occur if the gas composition is between the lower and upper flammability limits of the mixture or if the temperature is above the autoignition temperature. In Table 5, the flammability limits for selected compounds in air are given.

The flammability limits depend on the temperature and pressure. At elevated temperatures and pressures, a given mixture may be flammable, even if this is not the case at ambient conditions. It is a key objective of the design of the inlet zone of a CPO reactor to prevent ignition and thermal combustion.

The surface temperatures of the catalyst may be very high. Some CPO reactors are designed with shields between the catalyst zone and the mixing zone

Table 5 Lower (LFL) and upper (UFL) flammability limits for hydrocarbons in air at ambient temperature and pressure

Fuel	LFL (vol.%)	UFL (vol.%)
Methane	5	15
Ethane	2.9	13
Propane	2.3	9.5
n-Octane	0.8	6.5

(From Ref.[20].)

to avoid ignition that could be caused by back radiation. The catalytic zone can be very compact because of the high rate of reaction. Catalysts primarily in the form of pellets and monoliths (or foams) are used. It may be desirable to use monoliths in cases (e.g., for fuel cells) where the pressure drop should be minimized.

Most CPO catalysts are based on noble metals on a ceramic carrier. The use of a number of different carriers including zirconia and alumina has been reported. Most of the noble metals and nickel are active for CPO, and rhodium is probably the superior metal and is in general superior to platinum at comparable conditions.[17]

The performance of a CPO reactor is, in the literature, often characterized by the hydrocarbon conversion and selectivities to carbon monoxide and hydrogen. Methane conversion and selectivities are often reported to be more than 80–90%. This corresponds in general to conditions at which the exit gas is close to equilibrium for the shift reaction and the methane steam-reforming reaction with a low value of ΔT_R in Eq. (5). The most likely reaction sequence is total oxidation in the initial part of the catalyst zone followed by other reactions including steam-reforming, shift, and possibly partial oxidation.

Autothermal Reforming

Autothermal reforming combines partial oxidation and adiabatic steam reforming for conversion of the hydrocarbon feedstock into synthesis gas free of soot and higher hydrocarbons. The ATR reactor design consists of burner, combustion chamber, and catalyst bed placed in a refractory lined vessel, as illustrated in Fig. 10. The hydrocarbon feedstock with steam is reacted with oxygen in a substoichiometric flame, often represented by Reaction (14) in Table 4. Typically, the molar ratio of oxygen (as O_2) to carbon in the hydrocarbon feed stream is 0.5–0.6, depending on the application. The steam-reforming and water gas shift reactions do also take place thermally to some extent in the combustion chamber. The temperature in the core of the flame may be higher than 2000°C and the design of the combustion chamber is made to minimize the transfer of heat to the burner. All oxygen is consumed by the reactions in the thermal zone.

Residual methane is present at the exit of the combustion zone. In the catalytic bed, the methane steam-reforming and the water shift reactions take place. The gas leaving the ATR reactor is in chemical equilibrium. Normally, the exit temperature is above 900–1100°C. The catalyst must withstand very severe conditions when exposed to very high temperatures and steam partial pressures. One example of an ATR catalyst is nickel supported by magnesium aluminum spinel. For compact design, the catalyst size and shape is optimized for a low pressure drop and high activity.

Autothermal reformer has been used for the production of syngas since the late 1950s. In the early years its main application was for production of syngas for methanol and ammonia. The ATR was operated at high S/C ratio to maximize the yield of hydrogen, especially for the latter application. Later developments in the 1990s have proven its operability at low S/C ratios. Industrial demonstration at S/C = 0.6 was carried out in 1999[26] and at these conditions, the first industrial plant was started in 2002.[22] This technology at low S/C ratio will also be used for syngas generation in a plant in Qatar for production of synthesis fuels with a capacity of 34,000 barrels/day, to be started in 2006.

A typical process diagram for producing syngas with ATR is shown in Fig. 11.[23] Adiabatic prereforming

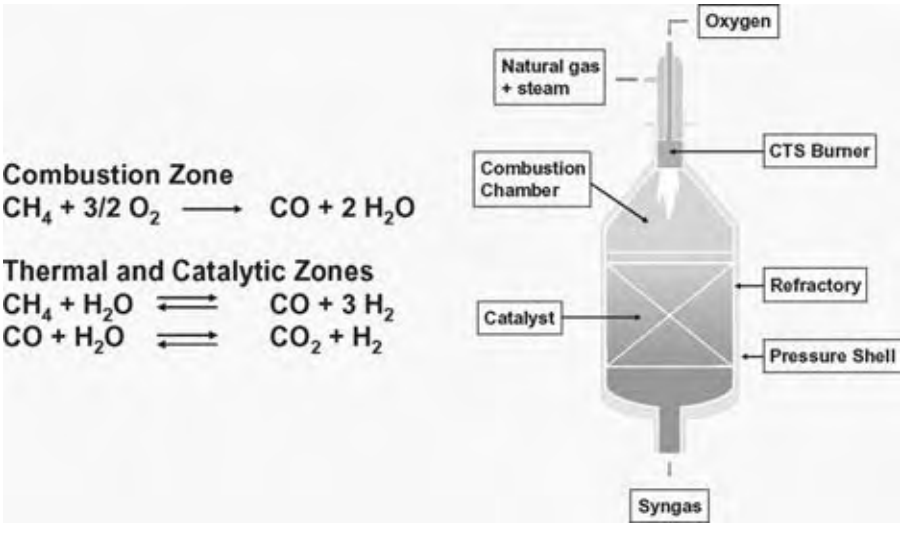


Fig. 10 Autothermal reforming reactor. *(View this art in color at www.dekker.com.)*

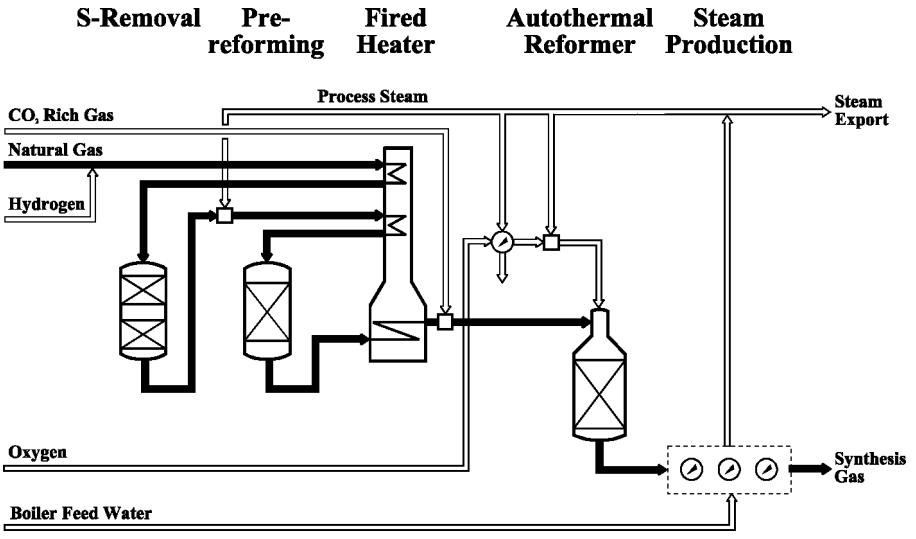


Fig. 11 Typical process diagram for an ATR-based plant for production of syngas. (From Ref.[23].)

of the hydrocarbon feedstock allows a higher preheat temperature without the risk of thermal cracking and reduces the oxygen consumption and the total plant investment. In the scheme given in Fig. 11, a fired heater is used for feed preheat. A carbon dioxide-containing stream can be added to the reactor feed stream to adjust the H_2/CO ratio in the syngas to the desired value. In a plant for the production of synthetic fuels, a tail gas from the Fischer–Tropsch synthesis can be used.

Autothermal reformer involves the risk of soot formation, especially when operating at low S/C ratios. The combustion in the flame proceeds through a large number of homogeneous radical reactions. The local stoichiometry in the flame zone will vary from being very fuel lean to very fuel rich. Fuel-rich combustion may lead to incomplete conversion of hydrocarbons in the feedstock into to nonsaturated compounds including ethylene and acetylene.[21,22] These compounds may be considered as precursors for the formation of polyaromatic hydrocarbons and soot, as illustrated by the simplified mechanism given in Fig. 12.[21] It is essential that the design and operating conditions are selected in such a way that the soot precursors are converted into syngas by the catalytic bed. One element is to ensure intensive mixing of the feed and oxidant by proper design of the burner.

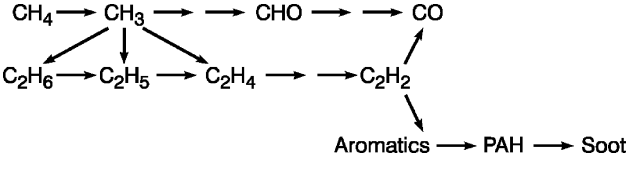


Fig. 12 Simplified model for the formation of soot in an ATR. (From Ref.[22].)

For a given design, the operating conditions determine whether or not soot is present in the exit gas from the ATR. High temperatures in the combustion chamber and sufficiently high amounts of steam in the feed suppress the formation of soot. Pilot plant experiments show that soot free operation at S/C ratios well below 0.6 is feasible.[6,21]

Ceramic Membrane Reforming

A large part of the cost of a syngas production plant based on partial oxidation is for the air separation unit. Continuous research efforts are going on for reducing the cost of this step. One possibility is to use oxygen-selective membranes integrated in a partial oxidation reactor, as illustrated in Fig. 13. Air is introduced

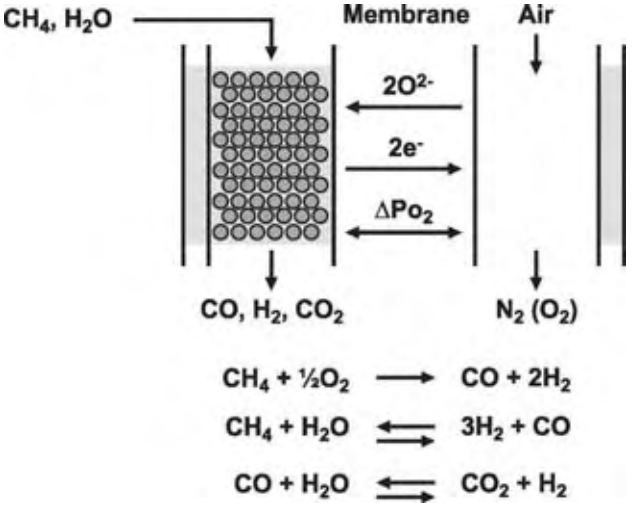


Fig. 13 Principle of ceramic membrane reforming.

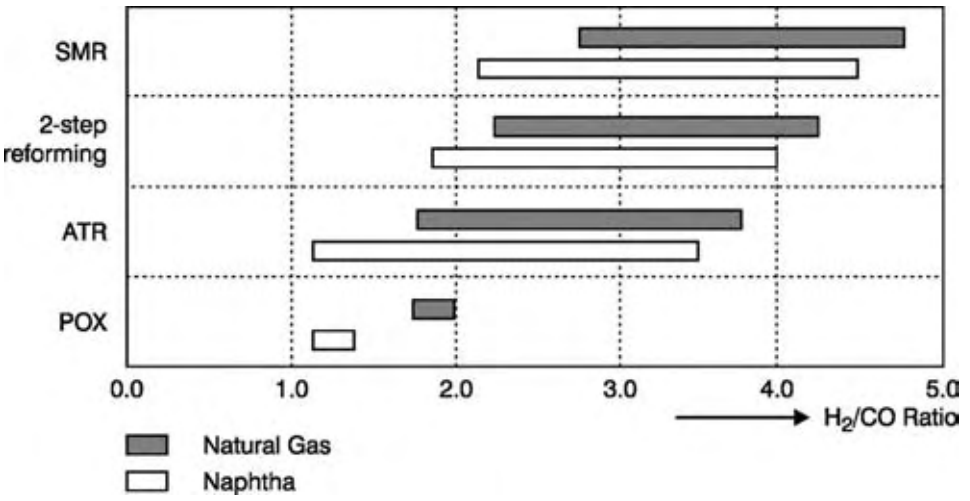


Fig. 14 H_2/CO ratios from various syngas production technologies. (From Ref.[1].)

at one side of the membrane through which oxygen with 100% selectivity is transported to the other side where it reacts with the hydrocarbons to produce syngas. This takes place at elevated temperatures that exceed 750°C or even higher than that. Membrane materials with oxygen conductivities in excess of $10\,Nm^3/m^2/hr$ have been identified, but several challenges remain for the development of this technology.

CHOICE OF TECHNOLOGY

The optimum technology for producing a syngas with a desired composition depends on the feedstock and the capacity. A very important parameter in selecting the best technology is the desired H_2/CO ratio. In Fig. 14, the syngas composition, which may be produced by various technologies with natural gas and naphtha feedstock, is illustrated.[1]

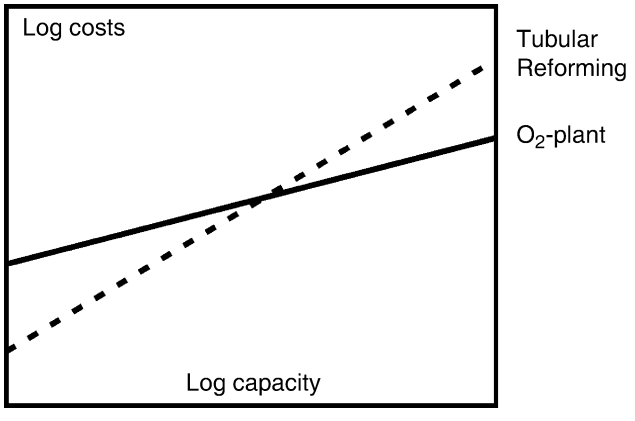


Fig. 15 Economy-of-scale for tubular reforming and air separation. (From Ref.[25].)

For small-scale applications, the cost may be more important than the efficiency of a given plant. In most cases, the partial oxidation reactors give a highly expensive plant layout because of the inherent, high cost of air separation (unless a low-cost source of oxygen is available). The use of air as oxidant makes the final separation very difficult, especially if carbon monoxide is needed either in a pure form or with hydrogen.

The situation may be different if a syngas for a high-temperature fuel cell (molten carbonate or solid oxide fuel cells) is needed. In this case, there is no need for separating the hydrogen and carbon monoxide from the nitrogen diluent when air is used as oxidant and methane can be processed directly by internal reforming in the fuel cell. The dilution with nitrogen may result in lower fuel cell efficiency and the use of partial oxidation will result in lower electrical efficiencies than with steam reforming.[a] However, this disadvantage may be more than offset by a simpler layout with fewer heat exchangers and no steam system.

For applications of larger scale, the selection of either steam-reforming or partial oxidation for the syngas section of a plant is often governed by the differences in the economy-of-scale of the tubular reformer and the oxygen plant as illustrated in Fig. 15. For the fairly low capacities, the cost of the air separation unit is prohibitive and steam reforming is the preferred choice. If possible and necessary, the H_2/CO ratio can be adjusted to the desired value by adding CO_2 either from import or by recirculation. Increasing

[a]Partial oxidation and steam reforming generally generate 3 and 4 moles of hydrogen per mole of methane, respectively (Tables 2 and 4). The difference is released in the form of heat in the case of partial oxidation.

the capacity of a tubular reformer means adding identical tubes and the scaling is therefore not as favorable as for an air-separation plant. At very large capacities, partial oxidation with air separation is the most economical technology. At intermediate capacities, tubular reforming followed by partial oxidation, which is known as secondary reforming, is often preferred.

CONCLUSIONS

Syngas is a key intermediate in the production of a number of important chemicals including ammonia, methanol, isocyanates, and higher alcohols. The desired end product determines the requirement of the syngas composition, ranging from pure hydrogen to pure carbon monoxide. In the coming years, the use and production of syngas may substantially increase if a market for synthetic fuels develops and/or fuel cells are commercialized.

The technologies for production of syngas from hydrocarbons are based on either steam-reforming or partial oxidation. In the former case, the hydrocarbons react with steam with considerable addition of heat to produce a syngas with a H_2/CO ratio of 3 or more. Partial oxidation may be carried out either thermally or catalytically (or by a combination) to produce a syngas with an H_2/CO ratio less than 2. Regardless of technology, CO_2 may be added to the feed to adjust the gas composition to a low H_2/CO ratio. In all cases, limits for the formation of carbon on catalysts or soot in the condensate must be considered to avoid rapid deactivation and low on-stream factors.

The choice of technology depends on the scale of operation and the desired end product. For small-scale syngas generators for fuel cells, partial oxidation may result in a simple layout, but steam reforming results in higher electrical efficiencies. For larger-scale syngas generators, the optimal technology is often dictated by differences in economy-of-scale of steam reforming and the air separation unit needed for partial oxidation plants. If the capacity is too low, the air separation unit may be too expensive; but in a larger scale, the superior economy-of-scale of the oxygen plant favors the partial oxidation technology.

The use of heat exchange reforming in which the heat is supplied by cooling a process gas is attracting increased attention. These reformers, working in combination with partial oxidation reactors, increase the plant energy efficiency and reduce the oxygen consumption. A main challenge in the further development of these types of heat exchange reformers is to prevent metal dusting. The use of heat exchange reforming may become more widespread in the coming years.

REFERENCES

1. Rostrup-Nielsen, J.R.; Sehested, J.; Nørskov, J.K. Hydrogen and synthesis gas by steam- and CO_2-reforming. Adv. Catal. (Elsevier Science, U.S.A.) **2002**, 65–139.
2. Rostrup-Nielsen, J.R.; Christiansen, L.J. *Chemical Reaction and Reactor Design*; Tamaki, M., Tominaga, H., Eds.; Maruzen Co., Ltd.: Tokyo, 1996, Chapter 5.2, 221.
3. Rostrup-Nielsen, J.R.; Dybkjaer, I.; Christiansen, L.J. Steam reforming opportunities and limits of the technology. In *Chemical Reactor Technology for Environmentally Safe Reactors and Products*; Kluwer Academic Publishers: The Netherlands, 1993; 249–281.
4. Rostrup-Nielsen, J.R.; Rostrup-Nielsen, T. Technology for large scale hydrogen production. Proceedings of the Sixth World Congress on Chemical Engineering, Melbourne, Australia, September 23–27, 2001.
5. Rostrup-Nielsen, T. Hydrocarbon Eng. **2002**, *7*, 51.
6. Aasberg-Petersen, K.; Bak Hansen, J.-H.; Christensen, T.S.; Dybkjaer, I.; Seier Christensen, P.; Stub Nielsen, C.; Winter Madsen, S.E.L.; Rostrup-Nielsen, J.R. Technologies for large-scale gas conversion. Appl. Catal. A: Gen. (Elsevier Science) **2001**, *221*, 379–387.
7. Rostrup-Nielsen, J.R.; Dybkjaer, I.; Christensen, T.S. Stud. Surf. Sci. Catal. **1998**, *113*, 81.
8. Piweth, M.; Larsen, J.S.; Christensen, T.S. Proceedings of the 1996 Fuel Cell Seminar; 780–783.
9. Rostrup-Nielsen, J.R.; Bak Hansen, J.-H. CO_2 reforming of methane over transition metals. J. Catal. **1993**, *144*, 38.
10. Rostrup-Nielsen, J.R. Catalytic steam reforming. In *Catalysis, Science and Technology*; Anderson, J.R., Boudart, M., Eds.; Springer: Berlin, 1983; Vol. 5, 1.
11. Sehested, J. Sintering of nickel steam reforming catalysts. Submitted to J. Catal.
12. Christensen, T.S. Adiabatic prereforming of hydrocarbons—an important step in syngas production. Appl. Catal. A: Gen. **1996**, *138*, 285.
13. Rostrup-Nielsen, J.R.; Sehested, J. Whisker carbon revisited. Stud. Surf. Sci. Catal. **2001**, *139*, 1.
14. Twigg, M., Eds. *Catalyst Handbook*; 1996.
15. Higman, C. Perspectives and Experience with Partial Oxidation of Heavy Residues. L'Hydrogène, Maillon Essentiel du Raffinage de Demain, L'Association Française des Techniques du Pétrol, Paris, 28 June 1994. http://www.higman.de/gasification/paris.pdf.
16. Madsack, H.-J. Make synthesis gas by residuum partial oxidation. Hydrocarbon Process. **1982**, 169.

17. Hickman, D.A.; Schmidt, L.D. Science **1993**, *259*.
18. Basini, L.; Aasberg-Petersen, K.; Guarinoni, A.; Ostberg, M.R. Catalytical partial oxidation of natural gas at elevated pressure and low residence time. Catal. Today **2001**, *64*, 9.
19. Hockmuth, J.K. Appl. Catal. B: Environ. **1992**, *1*, 89.
20. Bartok, W.; Sarofim, A.F. *Fossil Fuel Combustion*; John Wiley & Sons, 1991.
21. Christensen, T.S.; Primdahl, I.I. Improve syngas production using autothermal reforming. Hydrocarbon Process **1994**, March.
22. Aasberg-Petersen, K.; Christensen, T.S.; Stub Nielsen, C.; Dybkjaer, I. Recent developments in autothermal reforming and prereforming for synthesis gas production in GTL Applications. Submitted to Fuel Process. Technol. **2002**.
23. Christensen, T.S.; Ostberg, M.R.; Bak Hansen, J.-H. Process demonstration of autothermal reforming at low steam-to-carbon ratios for production of synthesis gas. Proceedings of the AIChE Annual Meeting, Reno, U.S.A., November 4–9, 2001.
24. Rostrup-Nielsen, J.R. New aspects of syngas production and use. Catal. Today **2000**, *63*, 159.
25. Dybkjaer, I.; Christensen, T.S. Syngas for large scale conversion of natural gas to liquid fuels. Stud. Surf. Sci. Catal. **2001**.
26. Ernst, W.S.; Venables, S.C.; Christensen, P.S.; Bertelsen, A.C. Hycrocarbon Process **2000**, *79* (3), 100.
27. Bakkerud, P.K.; Neergaard Gol, J.; Aasberg-Petersen, K.; Dybkjaer, I. Preferred synthesis gas production routes for GTL. Stud. Surf. Sci. Catal. 147. Natural Gas Conversion VII.
28. Patent Application WO 96/03345.
29. Dry, M.E.; Steynberg, A.P. Commercial FT process applications. Stud. Surf. Sci. Catal. *152*, Fischer–Tropsch Technology.

Tar Sand

James G. Speight
CD & W Inc., Laramie, Wyoming, U.S.A.

INTRODUCTION

Tar sand is sand saturated with a highly viscous hydrocarbonaceous material, not recoverable in its natural state through a well by ordinary production methods, including enhanced oil recovery methods. Thus, it is not surprising that the properties of bitumen from tar sand deposits are significantly different to the properties of conventional crude oil (Tables 1 and 2). Chemically, the material should perhaps be called bituminous sand rather than tar sand, since the organic matrix is bitumen, a hydrocarbonaceous material that consists of carbon and hydrogen with smaller amounts of nitrogen, oxygen, sulfur, and metals (especially nickel and vanadium).

Current recovery operations of bitumen in tar sand formations involve use of a mining technique. This is followed by bitumen upgrading and refining to produce a synthetic crude oil.

DISTRIBUTION

Tar sand deposits are distributed throughout the world in a variety of countries and the various tar sand deposits have been described as belonging to two types: 1) materials that are found in stratigraphic traps and 2) deposits that are located in structural traps. There are, inevitably, gradations and combinations of these two types of deposits and a broad pattern of deposit entrapment are believed to exist. There are no very large tar sand accumulations having more than 4 billion barrels (4×10^9 bbl) in place either in purely structural or in purely stratigraphic traps.

In Canada, the town of McMurray, about 240 miles north–north–east of Edmonton, Alberta lies at the eastern margin of the largest accumulation in the world There are, in effect, three major accumulations within the Lower Cretaceous deposits. The McMurray-Wabasca reservoirs are found toward the base of the formation and the deposit dips at between 5 ft and 25 ft per mile (1.5 m and 8 m per mile) to the south–west. The Bluesky-Gething sands overlie several unconformities between the Mississippian and Jurassic deposits.

In terms of specific geological and geochemical aspects of the formation, the majority of the work has, again, been carried out on the Athabasca deposit. Attention has repeatedly been focused on the variation in physical properties of crude oil produced in multiple-zone fields or in some instances within a single reservoir.

In the more localized context of the Athabasca deposit, inconsistencies arise presumably because of the lack of mobility of the bitumen at formation temperature (approximately 4°C, 39°F). For example, the proportion of bitumen in the tar sand increases with depth within the formation. Furthermore, the proportion of the nonvolatile asphaltenes or the nonvolatile asphaltic fraction (asphaltenes plus resins) in the bitumen also increases with depth within the formation that leads to reduced yields of distillate from the bitumen obtained from deeper parts of the formation. In keeping with the concept of higher proportions of asphaltic fraction (asphaltenes plus resins), variations (horizontal and vertical) in bitumen properties have been noted previously, as have variations in sulfur content, nitrogen content, and metals content. Obviously, the richer tar sand deposits occur toward the base of the formation, but the bitumen is generally of poorer quality.

The Canadian Melville Island deposits lie on the north shore of Marie Bay, Melville Island, some 1450 miles north of Edmonton. Deposits of Triassic sandstone of the Bjorne Formation are impregnated with a bituminous material.

The major tar sand deposits of the U.S.A. occur within and around the periphery of the Uinta Basin, Utah. These include the Asphalt Ridge, Sunnyside, Tar Sand Triangle, and Peor (PR) Springs deposits. Asphalt Ridge lies on the north–eastern margin of the central part of the Uinta Basin at the contact of the tertiary beds with the underlying Cretaceous Mesaverde Group. The Mesaverde Group is divided into three formations; two of which, the Asphalt Ridge sandstone and the Rim Rock sandstone are beach deposits containing the viscous bitumen.

The Californian deposits are concentrated in the coastal region west of the San Andreas Fault. The largest deposit is the Edna deposit, which is located midway between Los Angeles and San Francisco. The Sisquoc deposit (Upper Pliocene) is the second largest in California and occurs in a sandstone in which there are as many as eight individual tar sand units.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007983

Table 1 The properties of bitumen and conventional crude oil

Property	Bitumen	Conventional crude oil
Gravity, °API	8	35
Viscosity		
Centipoise at 100°F (38°C)	500,000	10
Centipoise at 210°F (99°C)	1700	
SUS at 100°F (38°C)	35,000	30
SUS at 210°F (99°C)	500	
Pour point, °F	50	0
Elemental analysis, % by weight		
Carbon	83	86
Hydrogen	10.6	13.5
Sulfur	4.8	0.1
Nitrogen	0.4	0.2
Oxygen	1	0.2
Fractional composition, % by weight		
Asphaltenes	19	5
Resins	32	10
Aromatics	30	25
Saturates	19	60
Metal parts per million		
Vanadium	250	10
Nickel	100	5
Carbon residue % by weight	14	5
Heating value btu/lb	17,500	19,500

Table 2 Distillation data (cumulative % by weight distilled) for bitumen and crude oil

Cut point		Cumulative % by weight distilled		
°C	°F	Athabasca	PR Spring	Leduc (Canada)
200	390	3	1	35
225	435	5	2	40
250	480	7	3	45
275	525	9	4	51
300	570	14	5	
325	615	26		
350	660	18	8	
375	705	22	10	
400	750	26	13	
425	795	29	16	
450	840	33	20	
475	885	37	23	
500	930	40	25	
525	975	43	29	
538	1000	45	35	
538+	1000+	55	65	

The Sisquoc deposit (Upper Pliocene) is the second largest in California and occurs in a sandstone in which there are as many as eight individual tar sand units. The third California deposit at Santa Cruz is located approximately 56 miles (90 km) from San Francisco.

South Texas holds the largest reserves in the state of tar sand deposits. The tar sand deposits occur in the San Miguel tar belt (Upper Cretaceous) mostly in Maverick and Zavala counties, as well as in the Anacadro limestone (Upper Cretaceous) of the Uvalde district. The Kentucky tar sand deposits are located at Asphalt, Davis-Dismal Creek, and Kyrock; they all occur in nonmarine Pennsylvanian or Mississippian sediments. The three deposits appear as stratigraphic traps and are thought to have received their bitumen or bitumen precursor from the Devonian Chattanooga shale. Tar sand deposits in New Mexico occur in the Triassic Santa Rosa sandstone, which is an irregularly bedded, fine- to medium-grained micaceous sandstone. Finally, the tar sand deposits in Missouri occur over an area estimated at 2000 square miles in Barton, Vernon, and Cass Counties.

Tar sand deposits in Venezuela occur in the Officina/Tremblador tar belt that are believed to contain bitumen-impregnated sands of a similar extent to those of Alberta, Canada. The Guanaco Asphalt Lake occurs in deposits that rest on a formation of mid-Pliocene age and is closely associated with the Guanaco crude oil field that produces heavy crude oil from shale and fractured argillite of the Upper Cretaceous group. The tar sands of the Orinoco deposit are located along the southern flanks of the eastern Venezuelan basin.

The Bemolanga (Madagascar) deposit is the third largest tar sand deposit presently known and extends over some 150 square miles in western Madagascar with a recorded overburden from zero to 100 ft (zero to 30 m). The average pay zone thickness is 100 ft (30 m) with a total bitumen in-place quoted at approximately two billion barrels (approximately 2×10^9 bbl).

The largest tar sand deposit in Europe is that at Selenizza, Albania, that also contains the Patos oil field throughout which there occurs extensive bitumen impregnation.

The Trinidad Asphalt Lake [situated on the Gulf of Paria, 12 miles west-south–west of San Fernando and 138 ft (43 m) above sea level] occupies a depression in the Miocene sheet sandstone.

The Rumanian deposits are located at Derna and occur (along with Tataros and other deposits) in a triangular section east and north–east of Oradia

between the Sebos Koros and Berrettyo rivers. The reservoir rock is nonmarine, representing freshwater deposition during a period of regression.

Tar sands occur at Cheildag, Kobystan and outcrop in the south flank of the Cheildag anticline; there are approximately 24 million barrels (24×10^6 bbl) of bitumen in place. Other deposits in the former USSR occur in the Olenek anticline (north–east of Siberia) and it has been claimed that the extent of asphalt impregnation in the Permian sandstone is of the same order of magnitude (in area and volume) as that of the Athabasca deposits. Tar sands have also been reported from sands at Subovka and the Notanebi deposit (Miocene sandstone) is reputed to contain 20% bitumen by weight. On the other hand, the Kazakhstan occurrence, near the Shubar-Kuduk oil field, is a bituminous lake with a bitumen content that has been estimated to be of the order of 95% by weight of the deposit.

Tar sands also occur in the Southern Llanos of Colombia below 1500 ft (457 m). The tar sands at Burgan in Kuwait and at the Inciarte and Bolivar coastal fields of the Maracaibo Basin are of unknown dimensions. Those at Inciarte have been exploited and are directly or closely associated with large oil fields. The tar sands of the Bolivar coastal fields are above the oil zones in Miocene beds and are in a lithological environment similar to that of the Officina-Tremblador tar belt. The tar sand deposits in the Leyte Islands (Philippines) are extreme samples of stratigraphic entrapment. Those of the Mefang Basin in Thailand are in Pliocene beds that overlie Triassic deposits and their distribution is stratigraphically controlled. There is a small accumulation at Chumpi, near Lima (Peru) from which a petroleum-type was distilled as a result of volcanic activity. Finally, tar sand deposits have also been recorded in Spain, Portugal, Cuba, Argentina, Thailand, and Senegal, but most are poorly defined and are considered to contain (in-place) less than 1 million barrels (1×10^6 bbl) of bitumen.

STRUCTURE

Tar sand is a mixture of sand, water, and bitumen with the sand component occurring predominantly as quartz. The arrangement of the sand, water, and bitumen has been assumed to be an arrangement whereby each particle of the sand is water-wet and a film of bitumen surrounds the water-wetted grains. The balance of the void volume is filled with bitumen, connate water, or gas; fine material, such as clay, occurs within the water envelope.

One additional aspect of the character of Athabasca tar sands that plays a role in research and practice is that the sand grains are not uniform in character. Grain-to-grain contact is variable and such a phenomenon influences attempts to repack mined sand, as may be the case in studies involving bitumen removal from the sand in laboratory-type in situ studies. This phenomenon also plays a major role in the expansion of the sand during processing where the sand to be returned to the mine site might occupy 120–150% of the volume of the original as-mined material.

PROPERTIES

Tar sand properties that are of general interest are bulk density, porosity, and permeability. Porosity is, by definition, the ratio of the aggregate volume of the interstices between the particles to the total volume and is expressed as a percentage. High-grade tar sand usually has a porosity in the range from 30% to 35%, that is somewhat higher than the porosity (5–25%) of most reservoir sandstone. The higher porosity of the tar sand has been attributed to the relative lack of mineral cement (chemically precipitated material that binds adjacent particles together and gives strength to the sand, which in most sandstone occupies a considerable amount of what was void space in the original sediment).

The bitumen content of the tar sand of the U.S.A and Canada varies from zero to as much as 22% by weight. There are, however, noted relationships between the bitumen, water, fines, and mineral contents for the Canadian tar sands. Similar relationships may also exist for the U.S.A. tar sands, but an overall lack of study has prevented the uncovering of such data.

For the Canadian tar sands, bitumen contents from 8% to 14% by weight may be considered as normal (or average). Bitumen contents above or below this range have been ascribed to factors that influence impregnation of the sand with the bitumen (or the bitumen precursor). There are also instances where bitumen contents in excess of 12% by weight have been ascribed to gravity settling.

COMPOSITION

Elemental Composition

The elemental composition of tar sand bitumen is generally constant and falls into a narrow range:

Carbon	3.4–0.5%
Hydrogen	0.4–0.2%
Nitrogen	0.4–0.2%
Oxygen	0.0–0.2%
Sulfur	0.0–0.5%
Metals (Ni and V)	1000 ppm

Chemical Composition

The precise chemical composition of bitumen is, despite the large volume of work performed in this area, largely speculative. In very general terms (and as observed from elemental analyses), tar sand bitumen is an extremely complex mixture of: 1) hydrocarbons; 2) nitrogen compounds; 3) oxygen compounds; 4) sulfur compounds; and 5) metallic constituents. However, this general definition is not adequate to describe the composition of petroleum as it relates to the behavior of the feedstocks.

Finally, during the fractionation of petroleum, the metallic constituents (metalloporphyrins and nonporphyrin metal chelates) are concentrated in the asphaltene fraction. The deasphaltened oils (petrolenes and maltenes) contain smaller concentrations of porphyrins than the parent materials and usually very small concentrations of nonporphyrin metals.

Fractional Composition

Bitumen can be separated into a variety of fractions using a variety of techniques that have been used since the beginning of petroleum science. In general, the fractions produced by these different techniques are called saturates, aromatics, resins, and asphaltenes. Much of the focus has been on the asphaltene fraction because of its high sulfur content and high coke-forming propensity.

The available evidence is specific to the Athabasca deposit. For example, bitumen obtained from the northern locales of the Athabasca deposit (Bitumount, Mildred-Ruth Lakes) has a lower amount (by weight values approximately 16–20%) of the nonvolatile asphaltene fraction than the bitumen obtained from southern deposits (Abasand, Hangingstone River; approximately 22–23% by weight asphaltenes). In addition, other data indicate that there is also a marked variation of asphaltene content in the tar sand bitumen with depth in the particular deposit.

Thus, bitumen from a specific deposit is not a uniform material. The chemical and physical (fractional) composition can vary not only with the location and age of the deposit but also with the depth.

Properties

The specific gravity of bitumen shows a fairly wide range of variation. The largest degree of variation is usually due to local conditions that affect material lying close to the faces, or exposures, occurring in surface tar sand deposits. There are also variations in the specific gravity of the bitumen found in beds that have not been exposed to weathering or other external factors.

Bitumen gravity primarily affects the upgrading requirements needed because of the low hydrogen content of the produced bitumen. The API gravity of known U.S.A tar sand bitumen ranges downward from about 14° API (0.973 specific gravity) to approximately 2° API (1.093 specific gravity). Although only a vague relationship exists between density (gravity) and viscosity, very-low-gravity bitumen generally has very high viscosity.

The pour point is the lowest temperature at which the bitumen will flow. The pour point for tar sand bitumen can exceed the natural temperature of tar sand deposits. It is important to consider because for efficient production a thermal extraction process to increase the reservoir temperature to beyond the pour point temperature must supply supplementary heat energy. Elements related to pour point are depth, bitumen viscosity, original reservoir temperature, and atomic hydrogen/carbon ratio.

Other properties such as calorific value, carbon residue, specific heat, softening point, flash point, molecular weight, and thermal conductivity are also used to determine the suitability of the bitumen for conversion options.

MINING TECHNOLOGY

The Athabasca Tar Sands deposit in Canada is the site of the only commercial tar sands mining operations. The Suncor mining and processing plant, located 20 miles north of Fort McMurray, Alberta, started production in the year 1967. The Syncrude Canada mining and processing plant, located five miles (8 km) away from the Suncor plant, started production in the year 1978. In both projects, about half of the terrain is covered with muskeg, an organic soil resembling peat moss, which ranges from a few inches to 23 ft (7 m) in depth. The major part of the overburden, however, consists of Pleistocene glacial drift and Clearwater Formation sand and shale. The total overburden varies from 23 ft to 130 ft (7–40 m) in thickness. The underlying tar sand strata averages about 150 ft (45 m), although typically 16–33 ft (5–10 m) must be discarded because of a bitumen content below the economic cut-off grade designated by either plant and generally on the order of 5% by weight.

In the Suncor pit design, the ore body is divided into two layers, or benches, each nominally 23 m high. The pit floor and the dividing plane between the upper and lower bench are roughly horizontal and 8000 ton/hr/hr bucket-wheel excavators are employed as the primary mining equipment. Tar sands loosened from the face of each bench by the bucket wheels are

discharged on to a series of conveyors. The overburden is stripped by an electric shovel that discharges to trucks for removal of the overburden material. Syncrude utilizes a single-bench design with 80 yd^3 (60 m^3) capacity draglines as the primary mining equipment. The draglines that pile the tar sands in windrows along the edge of the pit; four 64,000-ton/hr bucket wheels transfer the tar sands to a system of trunk conveyor belts that move the material to the extraction plant. The mining operations at the two plants differ by virtue of the choice of the primary mining equipment; the bucket-wheel excavators sit on benches whilst the draglines sit on the surface.

The bucket wheel excavators are units having a 33-ft (10-m) diameter digging wheel on the end of a long boom. Each wheel has a theoretical capacity of 9500 tons/hr, but the average output from digging is about 5000 tons/hr. Tar sand is transferred from the mine to the separation plant at a rate of 140,000 ton/day (140×10^6 per day) by a system of 60 inch wide (152 cm) conveyor belts and 72 inch (183 cm) trunk conveyors that operate at 1090 ft/min (333 m/min). The bucket wheel excavators are supplemented by front-end loaders that are used to dig overburden and load it through twin chutes on to 150-ton-capacity trucks. Additional equipment is used for maintaining the haul roads and for spreading and compacting the spoiled material. On the other hand, overburden may be stripped with 18 yd^3 hydraulically operated shovels and a fleet of 150-ton trucks.

The draglines are equipped with a 92 yd^3 (71 m^3) bucket at the end of a 364 ft (111 m) boom. They can be employed to dig both a portion of the overburden, which is free-cast into the mining pit, and the tar sand, which is piled in windrows behind the machine. Bucket wheel reclaimers, similar to bucket wheel excavators, load the tar sand from the windrows on to conveyor belts that transfer it to the plant.

Underground mining options have also been proposed, but for the moment have been largely discarded because of the fear of collapse of the formation on to any operators or equipment. This particular option should not, however, be rejected out of hand because a novel aspect or the requirements of the developer (which remove the accompanying dangers) may make such an option acceptable.

EXTRACTION TECHNOLOGIES

The Hot-Water Process

In terms of bitumen separation and recovery, the hot-water process is, to date, the only successful commercial process to be applied to bitumen recovery from mined tar sands in North America. Many process options have been tested with varying degrees of success and one of these options may even supersede the hot-water process.

The process utilizes the linear and nonlinear variation of bitumen density and water density, respectively, with temperature so that the bitumen that is heavier than water at room temperature becomes lighter than water at 80°C (180°F). Surface-active materials in the tar sand also contribute to the process. The essentials of the hot-water process involve conditioning, separation, and scavenging.

In the hot-water extraction process, the tar sand feed is introduced into a conditioning drum. In this step, the tar sand is heated, mixed with water, and agglomeration of the oil particles begins. The conditioning is carried out in a slowly rotating drum that contains a steam-sparging system for temperature control, as well as mixing devices to assist in lump size reduction and a size ejector at the outlet end. The tar sand lumps are reduced in size by ablation and mixing action. The conditioned pulp has the following characteristics: (1) solids 60% by weight to 85% by weight and (2) pH 7.5–8.5.

In the conditioning step, also referred to as mixing or pulping, tar-sand feed is heated and mixed with water to form a pulp of 60% by weight to 85% by weight solids at 80–90°C (175–196°F). The conditioned pulp is screened through a double-layer vibrating screen. Water is then added to the screened material (to achieve more beneficial pumping conditions) and the pulp enters the separation cell through a central feed well and distributor. The bulk of the sand settles in the cell and is removed from the bottom as tailing, but the majority of the bitumen floats to the surface and is removed as froth. A middlings stream (mostly of water with suspended fines and some bitumen) is withdrawn from approximately midway up the side of the cell wall. Clays accumulate in the middlings layer.

The separation cell acts like two settlers, one on top of the other. In the lower settler the sand settles down, whereas in the upper settler the bitumen floats. The bulk of the sand in the feed is removed from the bottom of the separation cell as tailings. A large portion of the feed bitumen floats to the surface of the separation cell and is removed as froth. A middlings stream consists mostly of water with some suspended fine minerals and bitumen particles.

The combined froth from the separation cell and scavenging operation contains an average of about 10% by weight mineral material and up to 40% by weight water. The dewatering and demineralizing is accomplished in two stages of centrifuging. In the first stage, the coarser mineral material is removed but much of the water remains. The feed then passes through a filter to remove any additional large-size mineral matter that would plug up the nozzles of the second stage centrifuges.

The third step in the hot-water process is scavenging. Froth flotation with air is usually employed. The scavenger froth is combined with the separation-cell froth to be further treated and upgraded to synthetic crude oil. Tailings from the scavenger cell join the separation-cell tailings stream and go to waste.

Froth from the hot-water process may be mixed with a hydrocarbon diluent, e.g., coker naphtha, and centrifuged. The Suncor process employs a two-stage centrifuging operation and each stage consists of multiple centrifuges of conventional design installed in parallel. The bitumen product contains 1% by weight to 2% by weight mineral (dry bitumen basis) and 5% by weight to 15% by weight water (wet diluted basis). Syncrude also utilizes a centrifuge system with naphtha diluent.

Other Processes

The proposed cold-water process for bitumen separation has been developed to the point of small-scale continuous pilot plants. The process uses a combination of cold water and solvent and the first step usually involves disintegration of the tar sand charge that is mixed with water, diluent, and reagents. The diluent may be a petroleum distillate fraction such as kerosene and is added in approximately a 1:1 weight ratio to the bitumen in the feed. The pH is maintained at 9–9.5 by the addition of wetting agents and approximately 0.77 kg of soda ash per ton of tar sand. The effluent is mixed with more water, and in a raked classifier the sand is settled from the bulk of the remaining mixture. The water and oil overflow the classifier and are passed to thickeners where the oil is concentrated. Clay in the tar-sand feed has a distinct effect on the process; it forms emulsions that are hard to break and are wasted with the underflow from the thickeners.

The sand-reduction process is a cold-water process without solvent. In the first step, the tar-sand feed is mixed with water at approximately 20°C (68°F) in a screw conveyor in a ratio of 0.75 ton to 3 ton per ton of tar sand (the lower range is preferred). The mixed pulp from the screw conveyor is discharged into a rotary-drum screen, which is submerged in a water-filled settling vessel. The bitumen forms agglomerates that are retained by an 840-micron (20-mesh) screen and withdrawn as oil product. The sand readily passes through the 840-micron (20 mesh) screen and is withdrawn as waste stream.

The spherical agglomeration process resembles the sand-reduction process. Water is added to tar sands and the mixture is ball-milled. The bitumen forms dense agglomerates of 75% by weight to 87% by weight bitumen, 12% by weight to 25% by weight sand, and 1% by weight to 5% by weight water.

An anhydrous solvent extraction process for bitumen recovery has been attempted and usually involves the use of a low-boiling hydrocarbon. The process generally involves up to four steps. In the mixer step, fresh tar sand is mixed with recycle solvent that contains some bitumen and small amounts of water and mineral. Solvent-to-bitumen weight ratio is adjusted to approximately 0.5. The drain step consists of a three-stage countercurrent wash. Settling and draining time is approximately 30 min for each stage.

The other aboveground method of separating bitumen from tar sands after the mining operation involves direct heating of the tar sand without previous separation of the bitumen. Thus, the bitumen is not recovered as such, but is an upgraded overhead product.

UPGRADING

The quality of tar sand bitumen is low compared to that of conventional crude oil and heavy oil. The high carbon residue of bitumen dictates that considerable amounts of coke will be produced during thermal refining (Table 3).

The current oil sands bitumen upgrading processes for the production of synthetic crude oil (Table 4) begin with diluted bitumen being processed through the diluent recovery units. The diluent recovery units are atmospheric distillation units that serve three purposes: 1) distill off diluent naphtha and return it to the froth treatment process; 2) distill off light gas oil and send it directly to a light gas oil hydrotreater; and 3) produce hot atmospheric topped bitumen as feedstock for vacuum distillation unit and downstream bitumen conversion processes.

The vacuum distillation unit cuts deeper into the atmospheric topped bitumen. It distills off the remaining light gas oil and heavy gas oil that are sent directly to hydrotreaters. The remaining vacuum topped bitumen can be blended with atmospheric topped bitumen and then sent to bitumen conversion units.

Table 3 Predicted coke yields from various feedstocks

API gravity of feedstock	Carbon residue % by weight	Coke yield	
		Delayed coking	Fluid coking
2	30	45	35
6	20	36	23
10	15	28	17
16	10	18	12
26	5	9	3

Table 4 Properties of synthetic crude oil from Athabasca bitumen

Property		Bitumen	Synthetic crude oil
Gravity, API		8	32
Sulfur, % by weight		4.8	0.2
Nitrogen, % by weight		0.4	0.1
Viscosity centipoise at 100°F		500,000	10
Distillation profile, % by weight			
°C	°F		
0	30	0	5
30	85	0	30
220	430	1	60
345	650	17	90
550	1020	45	100

Coking Processes

The overall upgrading process by which bitumen is converted to liquid products is accomplished in two steps. The first step is the primary upgrading process that improves the hydrogen-to-carbon ratio by either carbon removal or hydrogen addition, cracking the bitumen to lighter products which are more easily processed downstream. The secondary upgrading process involves hydrogenation of the primary products and is the means by which sulfur and nitrogen are removed from the primary products.

There are two processes that have been applied to the production of liquids from Athabasca bitumen. In this respect, these processes are proven but are not necessarily the best or ultimate processes. Delayed coking is practised at the Suncor (formerly Great Canadian Oil Sands) plant, whereas Syncrude employs a fluid coking process which produces less coke than the delayed coking in exchange for more liquids and gases (Table 5).

Product Upgrading

The primary liquid product is then hydrotreated (secondary conversion or refining) to remove sulfur and nitrogen (as hydrogen sulfide and ammonia, respectively) and to hydrogenate the unsaturated sites exposed by the conversion process. It may be necessary to employ separate hydrotreaters for light distillates and medium-to-heavy fractions; for example, the heavier fractions require higher hydrogen partial pressures and higher operating temperatures to achieve the desired degree of sulfur and nitrogen removal. Commercial applications have therefore been based on the separate treatment of two or three distillate fractions at the appropriate severity to achieve the required product quality and process efficiency.

Table 5 Potential emissions from a tar sand plant

Source of emission	Potential contaminants
Process waste water	Suspended solids
	Dissolved solids
	Organic compounds
	Ammonium compounds
	Inorganic (metal) compounds
Upgrader area runoff	Suspended solids
Upgrader heaters	Particulate matter
	Carbon oxides
	Sulfur oxides
	Nitrogen oxides
Coke storage pile	Suspended solids
	Sulfur
Solid waste landfill	Suspended solids
	Organic compounds
	Inorganic (metal) compounds
	Sulfur
Power plant	Particulate matter
	Carbon oxides
	Sulfur oxides
	Nitrogen oxides
Sulfur plant	Particulate matter
	Hydrogen sulfide
	Sulfur oxides

Other Processes

There are several other processes that have received some attention for bitumen upgrading. These processes include partial upgrading (a form of thermal deasphalting), flexicoking, the Eureka process, and various hydrocracking processes. In addition, a partial coking or thermal deasphalting process provides a minimal upgrading of bitumen.

Most hydrocracking processes start with an upflow reactor system in which the 524°C (975°F) material is cracked or converted. To prevent coking, the processes operate at high pressure with direct contact between the bitumen feed and circulating hydrogen. Hydrocracking processes include the H-Oil process, the LC-Fining process, the Veba-Combi Cracking process, and the Chiyoda process.

ENVIRONMENTAL ASPECTS

A major aspect of tar sand resource development is the influence of the various integral parts of the development on the surrounding environment and recognition of the various emissions from the plant (Table 5).

Thus, regulations could clearly state conditions for resource development in terms of:

1. Foliage removal-to what extent and how or to where.
2. Flora-endangered species should be transplanted.
3. Fauna removal to another (similar) location.
4. Till removal.

The mining operation poses a potential threat to the environment because:

1. Large ''holes'' have to be dug to retrieve the ore.
2. The overburden must first be removed from the immediate mine site

The first constraint is an inevitable part of the mining operation, but regulations may require that the site be returned to at least the original (unspoiled) condition. This is being achieved in Alberta by the techniques in which the mineral matter from the hot water separation process is returned to the mine site as fill material, thereby bringing the ground level back to the original premine level. In a similar manner, overburden handling can also be achieved with minimal stockpiling of the earth. Revegetation programs will complete the return of the site to the ''predevelopment'' state.

One of the greatest problems that has emerged from tar sand resource development is the disposal (and control) of the tailings streams that arise from the hot-water separation. The extent of this problem had not been appreciated until the commercialization of the tar sands commenced.

There are three major aspects of tailings disposal that need to be addressed:

1. Each ton of oil band in-place has a volume of approximately 16 ft^3; the tailings has a volume of approximately 22 ft^3, a volume gain of approximately 40%.
2. The clay discharged as part of the tailings stream does not settle and, therefore, limits the amount of the stream that can he used as recycle water.
3. The tailings stream contains bitumen.

Thus, tailings disposal requires a larger hole than that from which the tar sand originated or the level of the land will/must be raised at site abandonment. Until a satisfactory means of clay settling can be achieved that will allow reuse of the water, the tailings ponds will grow in size with the life of the plant. Currently, a recycle stream from the tailings ponds produces some 40–60% of the daily water requirements of the tar sand plants; the remainder is makeup water from the nearby Athabasca River.

CONCLUSIONS

Coking processes have served well in the production of synthetic crude oil from tar sand bitumen. However, new processes for the conversion of bitumen will probably be used in place of the coking processes with some degree of hydrocracking as a primary conversion step. Other processes may replace or augment the deasphalting processes in many refineries.

When catalytic processes are employed, complex molecules (such as those that may be found in the original asphaltene fraction) or those formed during the process, are not sufficiently mobile (or are too strongly adsorbed by the catalyst) to be saturated by hydrogenation.

BIBLIOGRAPHY

Gray, M.R. *Upgrading Petroleum Residues and Heavy Oils*; Marcel Dekker Inc.: New York, 1994.

Meyer, R.F., Ed. *Heavy Crude and Tar Sands – Hydrocarbons for the 21st Century*; Caracas, Venezuela: Petróleos de Venezuela S.A., 1991.

Speight, J.G. *Fuel Science and Technology Handbook*, Speight, J.G., Ed.; Marcel Dekker Inc.: New York, 1990; Part II, Chapters 12–16.

Speight, J.G. *The Chemistry and Technology of Petroleum*; 3rd Ed.; Marcel Dekker Inc.: New York, 1999.

Speight, J.G. *The Desulfurization of Heavy Oils and Residua*, 2nd Ed.; Marcel Dekker Inc.: New York, 1999.

Speight, J.G.; Ozum, B. *Petroleum Refining Processes*; Marcel Dekker Inc.: New York, 2002.

Theoretical Aspects of Liquid Crystals and Liquid Crystalline Polymers

T

James J. Feng
Department of Chemical and Biological Engineering and Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada

INTRODUCTION

Liquid crystallinity refers to an intermediate state of matter where the molecules exhibit a degree of order that is between that of ordinary liquids and solids. Typically, these molecules have an elongated shape and an intrinsic tendency toward alignment. Yet, the inter-molecular forces are not so strong as to bind them into a regular lattice as in crystalline solids. Thus, the orientational order among the molecules makes the material anisotropic and ''crystalline,'' while the lack of strong positional order allows the material to flow like ordinary fluids. This liquid–solid duality gives rise to much of the ''anomalous'' behavior of liquid crystalline materials. When understood, however, their unique dynamics can be harnessed to produce high-performance materials with unique properties.

In this entry, we review the state of the art in theoretical modeling and computation of the flow and rheology of low-molecular-weight liquid crystals (LCs) and liquid crystalline polymers (LCPs). The latter can be viewed as macromolecular liquid crystals, and the significance of the molecular weight will be made clear shortly. We restrict our scope to nematics, one of the several types of LCs with no positional order altogether;[1] they seem to have the most important applications and have been the focus of theoretical efforts. Owing to the long-standing academic and industrial interest in these materials, their dynamics have been summarized in several reviews, e.g., Refs.[1–4]. While overlapping those in certain aspects, we emphasize constitutive modeling and numerical simulation of defects and texture. The latter nicely demonstrates the interplay between macroscopic flow and mesoscopic orientation, a hallmark of liquid crystalline dynamics. Due to space limitation, we will attempt to construct a coherent framework instead of an exhaustive literature review. Based on their conceptual origins, we discuss continuum theories and molecular theories separately. In addition, a theory for liquid crystalline materials has been proposed based on nonequilibrium thermodynamics.[5] As the theoretical framework and mathematical representation are quite independent of the other theories, we will not discuss this theory but refer the reader to Beris and Edwards' monograph.[5]

Because of similar molecular attributes, LCs and LCPs share enough common features to be discussed in this single entry. A significant difference is that the large LCP molecules have a much longer relaxation time. Thus, the molecular conformation of LCPs is easily disturbed by flow and deformation. Rheologically, therefore, the material exhibits ''molecular viscoelasticity'' as macromolecular fluids normally do. Small-molecule LCs relax so fast that their molecular configuration remains almost always at equilibrium; there is no viscoelasticity. Both LCs and LCPs resist spatial distortion to their orientational pattern. This tendency is known as ''distortional elasticity.'' The continuum theories were originally developed for small-molecule LCs, capturing distortional elasticity but not molecular viscoelasticity. Molecular theories, on the other hand, have evolved to contain both ingredients. Not surprisingly, a properly constructed molecular theory should reduce to a continuum theory in the limit of vanishing molecular relaxation time.

CONTINUUM THEORIES

The tendency of LCs to resist and recover from distortion to their orientation field bears clear analogy to the tendency of elastic solids to resist and recover from distortion of their shape (strain). Based on this idea, Oseen, Zocher, and Frank established a linear theory for the distortional elasticity of LCs.[1] Ericksen[6,7] incorporated this into hydrostatic and hydrodynamic theories for nematics, which were further augmented by Leslie[8] with constitutive equations. The Leslie–Ericksen theory has been the most widely used LC flow theory to date.

Frank Elasticity

Viewing the LC as an anisotropic continuum, a pseudo-vector $\boldsymbol{n}$ of unit length, called the director, is used to indicate the orientation field. Thus, orientational

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030566

distortion is described by the spatial gradient $\nabla \boldsymbol{n}(\boldsymbol{r})$. Retaining quadratic terms, one may write the free energy density for orientation distortion as:[1]

$$F_d = \frac{K_1}{2}(\nabla \cdot \boldsymbol{n})^2 + \frac{K_2}{2}(\boldsymbol{n} \cdot \nabla \times \boldsymbol{n})^2 + \frac{K_3}{2}|\boldsymbol{n} \times \nabla \times \boldsymbol{n}|^2 \quad (1)$$

where the coefficients K_1, K_2, and K_3 are elastic constants corresponding to three canonical forms of orientational distortion: splay, twist, and bend.[1] The K's have the dimension of force, and F_d has the dimension of energy per unit volume. When the three K's are equal, the free energy takes on a particularly simple form:

$$F_d = \frac{K}{2}\nabla \boldsymbol{n} : (\nabla \boldsymbol{n})^{\mathrm{T}} \quad (2)$$

Thus the ''one-constant approximation'' is often used in theoretical analysis.

Leslie–Ericksen Theory

When $\boldsymbol{n}(\boldsymbol{r})$ is disturbed, say by an external magnetic field, an elastic torque arises:

$$h = -\frac{\delta F_d}{\delta \boldsymbol{n}} = -\frac{\partial F_d}{\partial \boldsymbol{n}} + \nabla \cdot \left(\frac{\partial F_d}{\partial \nabla \boldsymbol{n}}\right) \quad (3)$$

Known as the molecular field, $\boldsymbol{h}$ tends to restore $\boldsymbol{n}$ to its former orientation. Minimizing the total free energy shows that an equilibrium is achieved when $\boldsymbol{h}$ is parallel to $\boldsymbol{n}$. Furthermore, when the LC is deformed, the variation of $F_d(\boldsymbol{n},\nabla \boldsymbol{n})$, with respect to the strain, gives rise to an elastic stress known as the Ericksen stress:[1]

$$\sigma^{\mathrm{E}} = -\frac{\partial F_d}{\partial(\nabla \boldsymbol{n})} \cdot (\nabla \boldsymbol{n})^{\mathrm{T}} \quad (4)$$

The above two equations form the core of Ericksen's theory of LC hydrostatics.

Leslie recognized from early experiments that the anisotropy of the materials calls for multiple viscosity coefficients corresponding to different orientation of the LC relative to the flow. Combining this idea with the Ericksen theory leads to the Leslie–Ericksen (LE) theory, which comprises two elements: one describing the evolution of $\boldsymbol{n}(\boldsymbol{r})$ in a flow field, and the other prescribing an extra stress tensor due to the evolving $\boldsymbol{n}(\boldsymbol{r})$ field.

The evolution of $\boldsymbol{n}(\boldsymbol{r})$ is governed by a balance between viscous and elastic torques:

$$\boldsymbol{h} = \gamma_1 \boldsymbol{N} + \gamma_2 \boldsymbol{D} \cdot \boldsymbol{n} \quad (5)$$

where γ_1 and γ_2 are viscosity coefficients,

$$\boldsymbol{N} = \frac{d\boldsymbol{n}}{dt} - \boldsymbol{\Omega} \cdot \boldsymbol{n} \quad (6)$$

is the rotation of $\boldsymbol{n}$ relative to the background fluid, $\boldsymbol{\Omega} = [(\nabla \boldsymbol{v})^{\mathrm{T}} - \nabla \boldsymbol{v}]/2$ and $\boldsymbol{D} = [(\nabla \boldsymbol{v})^{\mathrm{T}} + \nabla \boldsymbol{v}]/2$ being the rotation and strain rate tensors. The molecular field $\boldsymbol{h}$ is given by Eq. (3) in terms of $\boldsymbol{n}$. The Leslie–Ericksen stress tensor is written as

$$\sigma^{\mathrm{LE}} = \sigma^{\mathrm{E}} + \alpha_1 \boldsymbol{D} : \boldsymbol{nnnn} + \alpha_2 \boldsymbol{nN} + \alpha_3 \boldsymbol{Nn} + \alpha_4 \boldsymbol{D} + \alpha_5 \boldsymbol{nn} \cdot \boldsymbol{D} + \alpha_6 \boldsymbol{D} \cdot \boldsymbol{nn} \quad (7)$$

where the α's are viscosity coefficients, related to the γ's of Eq. (5) by $\gamma_1 = \alpha_3 - \alpha_2$ and $\gamma_2 = \alpha_2 + \alpha_3 = \alpha_6 - \alpha_5$ (see Ref.[1] for explanation).

We must point out two related limitations of the LE theory. First, it applies to small-molecule LCs and to LCPs in the limit of vanishing strain rate. This is because the LE theory uses a vector $\boldsymbol{n}$ to represent the orientation state of the fluid, tacitly assuming that the molecular orientation distribution stays at its equilibrium state. This is reasonable when the molecular relaxation time is much shorter than the characteristic time of the flow. Second, the theory does not allow orientational defects, which would be singularities in the $\boldsymbol{n}$ field. In reality, LCs and LCPs tend to have a high density of defects.[1,9,10] Near the defect core, large spatial gradients distort the molecular orientation distribution, thus invalidating the LE theory.

Predictions of the Leslie–Ericksen Theory for Shear Flows

The LE theory has been applied to simple shear, Poiseuille, and nonviscometric flows.[4,11] If $\alpha_3/\alpha_2 > 0$, the LE theory predicts a steady ''flow-aligning'' solution in shear flows. If $\alpha_3/\alpha_2 < 0$, however, no steady solution exists, and $\boldsymbol{n}$ rotates continuously. The latter type, known as ''tumbling'' nematics, exhibits much richer dynamics. We will focus on instabilities in sheared tumbling LCs, for these reveal the most interesting physics, especially regarding the nucleation of defects. The key parameter in this problem is the Ericksen number: $Er = \gamma_1 VH/K_1$, V and H being the characteristic velocity and length, respectively. It represents the interplay between the viscous torque exerted by shear and the elastic torque emanating from the wall anchoring.

Consider the shear flow geometry in Fig. 1, with three different anchoring conditions on the top and bottom planes. The initial condition is a uniform $\boldsymbol{n}$ field consistent with the wall anchoring. Cases (a) and (b) are similar in which $\boldsymbol{n}$ lies initially in the y–z plane. In both cases, the LE theory predicts an in-plane tumbling instability and an out-of-plane twist instability.[12]

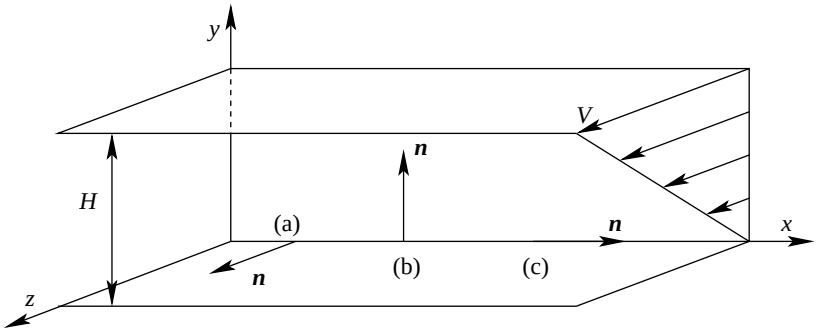


Fig. 1 Geometry for shear flow, with three possible anchoring conditions on the top and bottom planes: (a) ***n*** fixed along the flow direction (planar anchoring); (b) ***n*** fixed along the velocity gradient (homeotropic anchoring); and (c) ***n*** fixed along the vorticity direction (log rolling).

If ***n*** is restricted to the y–z plane during the shear, a steady ''windup'' solution obtains for low shear rates, with ***n*** rotating the most at the center and less toward the walls. This windup picture becomes unstable at a critical shear rate, where the director tumbles discontinuously to a new solution with reduced elastic energy. The critical Er depends on material parameters, and falls roughly between 10 and 100. Mathematically, this instability is represented by the existence of multiple in-plane solutions at certain ranges of the shear rate.[13] An example is shown in Fig. 2.

If one relaxes the constraint of in-plane orientation, a second instability appears in roughly the same Er range, with the director being driven out of the flow-gradient (y–z) plane toward the vorticity (x) direction.[12] This twist instability is accompanied by secondary flows orthogonal to the primary flow. Critical Er values for the tumbling and twist instabilities have been determined,[12] and detailed bifurcation diagrams constructed.[14] With increasing Er, ***n*** approaches the vorticity direction except for a thin layer next to the walls; the whole field approaches the initial condition in Fig. 1 with the anchoring condition (c).

This ''log-rolling'' configuration, be it the initial state with condition (c) or the result of a twist instability with conditions (a) or (b), is itself unstable to the formation of counter-rotating pairs of rolls aligned with the flow (Fig. 3A). This roll-cell instability was first recognized by linear analyses.[15,16] Along the vorticity direction, the rolls disturb the ***n*** field periodically and produce alternating dark and light stripes parallel to the primary flow, which have been experimentally confirmed.[17–19]

Experiments demonstrate that at even higher Er, the rolls become unstable and irregular. Ultimately, defect lines called disclinations form in the flow direction. As the linear analysis concerns the behavior of infinitesimal disturbances, the growth of the instability and further bifurcations are inaccessible to such analyses. This motivated Feng, Tao, and Leal[20] to carry out a direct numerical simulation of a sheared nematic. Using the LE theory, with the one-constant approximation, they predicted a cascade of instabilities illustrated in Fig. 3. Steady state rolls first appear at $Er = 2368$. The director twists toward the flow (z) direction at the center of the cells. With increasing Er, the secondary flow and the director twisting intensify,

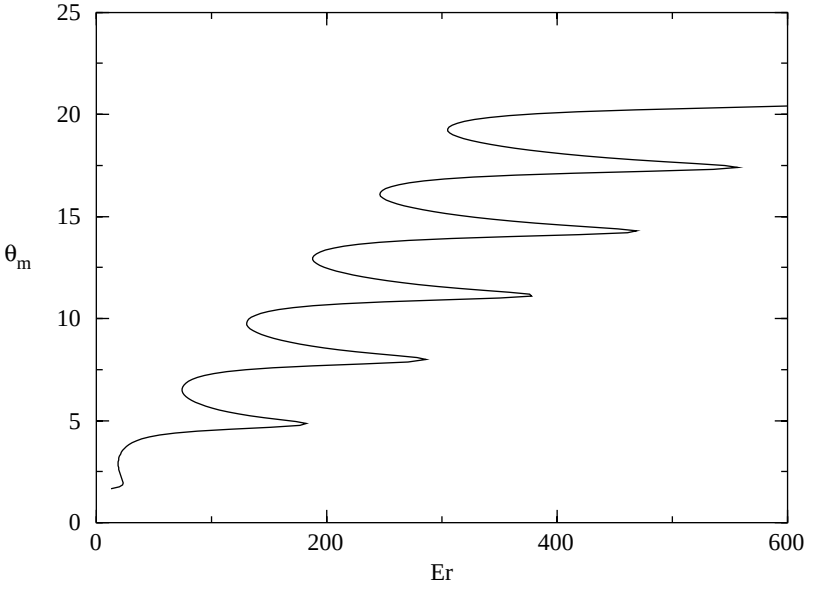


Fig. 2 The maximum orientation angle θ_m (in radians) as a function of Er for in-plane windup solutions of the LE theory using parameters for 8CB.[12] Multiple values of θ_m indicate multiple solutions at one shear rate, the jumps among which are the tumbling instability.

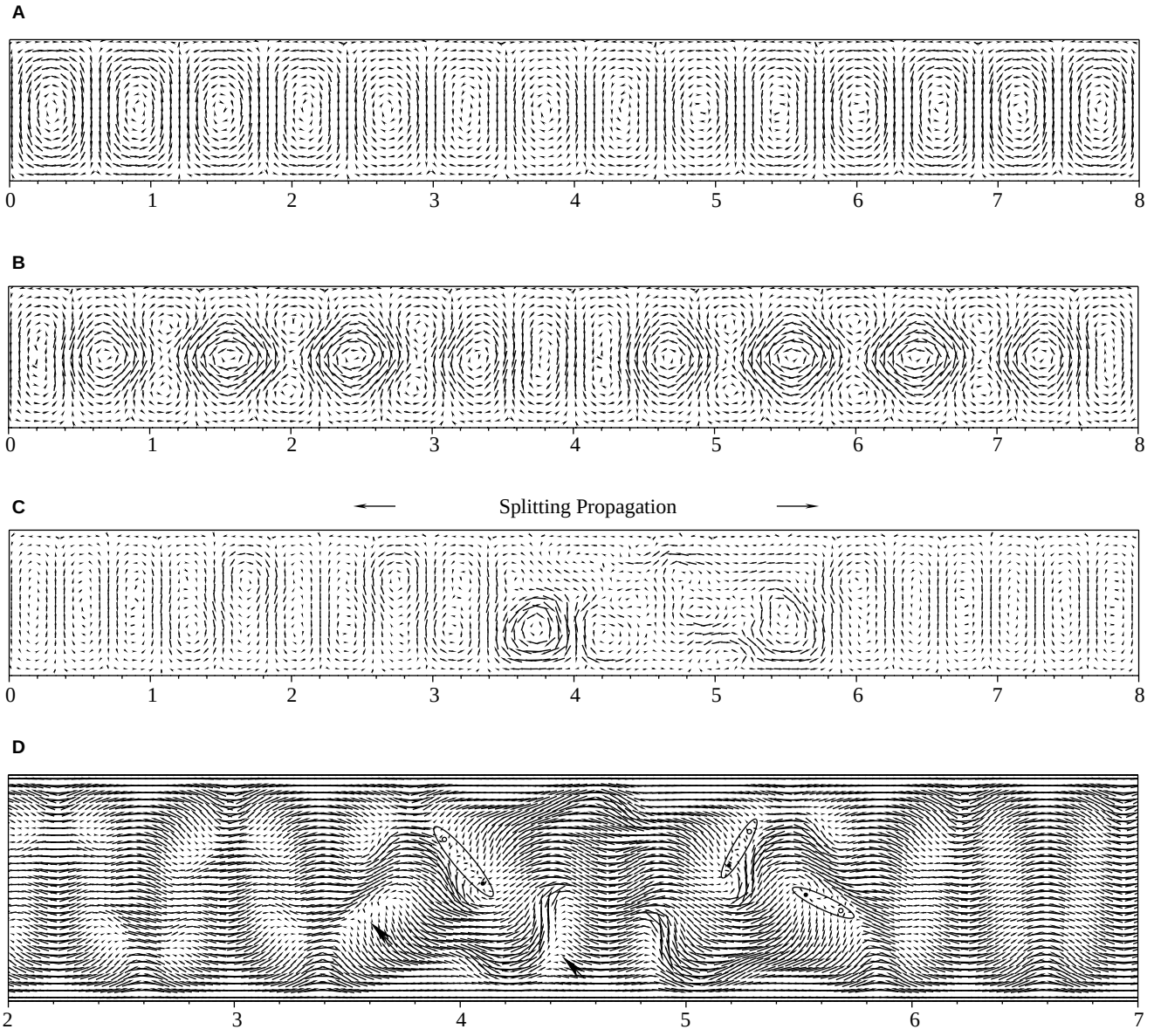


Fig. 3 Secondary flows in the x–y plane showing instabilities in a sheared nematic: (A) velocity vectors showing steady roll cells at $Er = 2602$; (B) a snapshot of oscillating roll cells at $Er = 4336$; (C) cell splitting at $Er = 8672$; and (D) ridges, indicated by arrows, break up to produce pairs of ± 1 defects marked by the ellipses. (From Ref.[20].)

while the wavelength of the dominant mode shrinks. When Er reaches 4026, the steady solution gives way to a time-periodic one: the roll cells begin to oscillate (Fig. 3B). When Er exceeds 6867, the cells split and evolve into an irregular pattern (Fig. 3C). Meanwhile, the director field forms ''ridges,'' namely elongated regions where $\boldsymbol{n}$ is mostly aligned with the primary flow (Fig. 3D). If the local secondary flow is favorable, the ridges break up to produce pairs of ± 1 disclinations of the escaped type with a diffuse core. This is the first detailed understanding of the process of flow-induced defect formation. More recently, Tao and Feng[21] extended this work to allow three unequal elastic constants, and explored the interaction among the three modes of distortion during the nucleation of defects. Note that the LE theory is generally incapable of describing defects. Fortunately, the disclinations produced by broken roll cells in both experiments[18] and numerical simulations[20,21] are nonsingular lines with a diffuse core.

Generalizations of the Leslie–Ericksen Theory

The inability of the LE theory to describe orientational defects has motivated efforts to generalize it.

The origin of the difficulty is the assumption of an equilibrium orientation distribution. In reality, the large spatial gradients at the defect would distort the molecular orientation distribution and severely reduce the local order parameter.

Ericksen[22] introduced a variable scalar order parameter; this amounts to assuming a spheroidal orientation distribution, with the axis of symmetry along the unit-length director. Liu and Walkington[23] proposed a conceptually related model by allowing the director length to vary. Its deviation from unity incurs a penalty in the Ginzburg–Landau form, and the Frank energy with the one-constant approximation can be rewritten as

$$F_{CW} = K\left[\frac{1}{2}\nabla \boldsymbol{n} : (\nabla \boldsymbol{n})^{T} + \frac{(\boldsymbol{n}\cdot\boldsymbol{n} - 1)^2}{4\delta^2}\right] \quad (8)$$

The small parameter δ indicates the extent of the defect region where $|\boldsymbol{n}|$ deviates from unity; singularity is avoided as $|\boldsymbol{n}|$ shrinks. Thus, $|\boldsymbol{n}|$ serves the role of Ericksen's scalar parameter. Tsuji and Rey[24,25] represented the molecular orientation distribution by a second-rank tensor. Thus, the distribution is allowed an ellipsoidal shape with three distinct eigenvalues. The free energy assumes the Landau–DeGennes form.[26] Now defects correspond to points where the two largest eigenvalues are equal and a unique director cannot be defined. This theory has been used to study the appearance and evolution of spatial textures.[24,25] Finally, the Larson–Doi theory[27] is an ingenious generalization of the LE idea to a larger length scale to account for polydomain textures.

MOLECULAR THEORIES

The starting point of a molecular constitutive theory is a simple mechanical model for the molecule that captures its most salient traits. Thus, flexible polymer molecules have been represented by elastic dumbbells and bead-spring chains,[28] and rigid polymers by rigid dumbbells[28] and rigid rods.[29] For its simplicity, the evolution of the model molecule is easily described by a convection-diffusion equation. Then a Fokker–Planck equation is written for the probability distribution function of an ensemble of these molecules. Finally, the macroscopic stress tensor is derived in terms of the distribution function. This kinetic theory framework was pioneered by Kirkwood (see, for example, Ref.[29]).

Nematic Potentials

The orientational order in a nematic derives from inter-molecular forces of various origins. We represent such molecular interactions by a mean-field potential energy known as the nematic potential. The Onsager and the Maier–Saupe potentials have played prominent roles in the development of the subject.

The Onsager potential is based on the excluded-volume effect among rod-like molecules.[1] Consider an ensemble of rigid rods of uniform length L and diameter b, with a number density ν and an orientation distribution function $\Psi(\boldsymbol{u})$ for the molecular orientation $\boldsymbol{u}$. For a test molecule oriented in $\boldsymbol{u}$, the effect of all the other molecules can be represented by a mean-field Onsager potential:

$$V_{ON}(\boldsymbol{u}) = \nu kT \int \Psi(\boldsymbol{u}')\beta(\boldsymbol{u},\boldsymbol{u}')d\boldsymbol{u}' \quad (9)$$

where $\beta(\boldsymbol{u},\boldsymbol{u}') = 2bL^2|\boldsymbol{u}\times\boldsymbol{u}'|$ represents the excluded-volume interaction between two molecules along $\boldsymbol{u}$ and $\boldsymbol{u}'$, k is the Boltzmann constant, T is temperature, and the integration is over a unit sphere representing all possible orientations.

The Maier–Saupe potential is a phenomenological model originally proposed for thermotropic small-molecule LCs.[1] It is obtained by replacing the excluded-volume interaction in Eq. (9) by

$$\beta(\boldsymbol{u},\boldsymbol{u}') = \text{const} - \beta_1 bL^2(\boldsymbol{u}\cdot\boldsymbol{u}')^2 \quad (10)$$

where β_1 is a constant. Then, the Maier–Saupe nematic potential can be written as:

$$\begin{aligned} V_{MS}(\boldsymbol{u}) &= \text{const} - \frac{3}{2}UkT\boldsymbol{u}\boldsymbol{u} : \int \Psi(\boldsymbol{u}')\boldsymbol{u}\boldsymbol{u}'d\boldsymbol{u}' \\ &= \text{const} - \frac{3}{2}UkT\boldsymbol{u}\boldsymbol{u} : \boldsymbol{S}, \end{aligned} \quad (11)$$

where

$$U = \frac{2}{3}\beta_1 \nu bL^2 \quad (12)$$

is called the nematic strength, and

$$S = \int \Psi(\boldsymbol{u}')\boldsymbol{u}'\boldsymbol{u}'d\boldsymbol{u}' \quad (13)$$

is the second-order moment of the orientation distribution.

With either nematic potential, the equilibrium orientation distribution $\Psi_{eq}(\boldsymbol{u})$ can be computed via a self-consistency condition, and the isotropic-to-nematic transition has been analyzed in terms of nematic strength parameters.[1,29]

Doi Theory for Monodomain LCPs: Molecular Viscoelasticity

Doi developed a dynamic theory for rigid-rod polymers using the Onsager potential.[29,30] Subsequent applications mostly used the Maier–Saupe potential for mathematical simplicity; we will illustrate Doi's theory using the latter. Consider the evolution of the orientation distribution $\Psi(\boldsymbol{u})$ for an ensemble of rods in a linear flow with a constant velocity gradient $\nabla \boldsymbol{v}$. $\Psi(\boldsymbol{u})$ is assumed to be spatially uniform; the nematic is a single crystal, or "monodomain." The conservation of probability leads to the following Fokker–Planck equation:

$$\frac{\partial \Psi}{\partial t} = -\Re \cdot (\boldsymbol{u} \times \boldsymbol{\kappa} \cdot \boldsymbol{u} \Psi) + D_r \Re \cdot \left[\Re \Psi + \frac{1}{kT} \Psi \Re V_{\mathrm{MS}} \right] \tag{14}$$

where the rotational operator $\Re = \boldsymbol{u} \times (\partial / \partial \boldsymbol{u})$, and $\boldsymbol{\kappa} = (\nabla \boldsymbol{v})^{\mathrm{T}}$. D_r is a "pre-averaged" rotational diffusivity for the rods.[29] The nematic potential V_{MS} enters as it modifies the rotation of rods by exerting a torque $-\Re V_{\mathrm{MS}}$. An evolution equation for the second-moment tensor $\boldsymbol{S}$ follows from Eq. (14):

$$\frac{\partial \boldsymbol{S}}{\partial t} = -6D_r \left(\boldsymbol{S} - \frac{\delta}{3} \right) + 6D_r U (\boldsymbol{S} \cdot \boldsymbol{S} - \boldsymbol{S} : \boldsymbol{Q}) + \kappa \cdot \boldsymbol{S} + \boldsymbol{S} \cdot \kappa^{\mathrm{T}} - 2\kappa : \boldsymbol{Q} \tag{15}$$

where $Q = \int \boldsymbol{uuuu} \Psi(\boldsymbol{u}) d\boldsymbol{u}$, and $\boldsymbol{\delta}$ is the second-rank unit tensor. An elastic stress tensor can be derived by the virtual work principle:

$$\sigma_E = 3\nu kT [\boldsymbol{S} - U(\boldsymbol{S} \cdot \boldsymbol{S} - \boldsymbol{S} : \boldsymbol{Q})] \tag{16}$$

In addition, a viscous stress tensor arises from viscous friction on the rods:

$$\sigma_V = \frac{\nu}{2} \zeta_r \boldsymbol{\kappa} : Q \tag{17}$$

where ζ_r is a rotational friction constant defined by Doi and Edwards.[29]

For Eqs. (15)–(17) to be a self-contained rheological theory, the fourth-order moment $\boldsymbol{Q}$ has to be related to the second-moment $\boldsymbol{S}$ by a closure approximation. Such a closed theory describes the LCP orientation by the second-rank tensor $\boldsymbol{S}$, and the director can be identified as the eigenvector for the largest eigenvalue. Doi[30] introduced a decoupling approximation: $\boldsymbol{S}:\mathbf{Q} = \boldsymbol{S}:\boldsymbol{SS}$, which turns out to be unsatisfactory as it artificially suppresses director tumbling.[31] More sophisticated closure models have since appeared, and their impact on the theory's prediction has been carefully examined (for example, Refs.[32,33].) The best closure models to date preserve most of the qualitative features of the theory.

The Doi theory captures the molecular viscoelasticity of LCP, i.e., the relaxation of the orientation distribution under flow. But it completely ignores distortional elasticity and is limited to monodomains. The assumption of spatial uniformity underlies all its key elements: the nematic potential, the kinetic equation, and the elastic stress tensor. Therefore, its successes are restricted to situations where distortional elasticity is insignificant.

One such success is the prediction of anomalous normal stress differences in shear flow. Measurements indicate that the normal stress differences N_1 and N_2 undergo two sign changes as the shear rate increases. For small and large shear rates, $N_1 > 0$, $N_2 < 0$ as expected of flexible polymers. In an intermediate range, however, $N_1 < 0$, $N_2 > 0$.[34] When Eq. (14) is solved without closure approximations, the Doi theory predicts three regimes of director dynamics with increasing shear rate: tumbling, wagging, and steady alignment.[31,35,36] The first sign change occurs within the tumbling regime, while the second occurs in the steady alignment regime. These transitions are linked to the spreading or narrowing of the orientation distribution Ψ. For N_1 and N_2, quantitative agreement has been achieved between prediction and measurements.[34] Using the closure-approximated Doi theory, Feng and Leal[37] simulated inhomogeneous LCP flows, and showed that director tumbling can lead to defect-like patterns. They also predicted a flow-orientation instability resembling experimental observations in channel flows.[38]

Complete Theories with Molecular Viscoelasticity and Distortional Elasticity

To add distortional elasticity to the Doi theory, one has to start with a more general nematic potential that accounts for spatial gradients. Marrucci and Greco[39] derived such a potential for nonlocal interaction among rigid-rod molecules in a distorted nematic material. For a test molecule oriented along $\boldsymbol{u}$ at position $\boldsymbol{r}$, Marrucci and Greco delineated an interaction region V enveloping the test molecule. Spatial averaging of the molecular interaction inside a spherical V produces a nonlocal nematic potential:

$$V_{\mathrm{MG}}(\boldsymbol{u}) = \text{const} - \frac{3}{2} UkT \left(\boldsymbol{S} + \frac{\Lambda^2}{24} \nabla^2 \boldsymbol{S} \right) : \boldsymbol{uu} \tag{18}$$

where the interaction length $\Lambda \sim V^{1/3}$ is the only free parameter in the theory. Eq. (18) is a molecularly-based generalization of the Maier–Saupe potential.

Following the procedure in deriving the Doi theory, one arrives at a kinetic equation for $\boldsymbol{S}$:[40,41]

$$\frac{\partial \boldsymbol{S}}{\partial t} + \boldsymbol{v}\cdot\nabla S = -6D_r\left(\boldsymbol{S} - \frac{\delta}{3}\right) + 6D_r U(\boldsymbol{S}\cdot\boldsymbol{S} - \boldsymbol{S}:\boldsymbol{Q}) + \boldsymbol{\kappa}\cdot\boldsymbol{S} + \boldsymbol{S}\cdot\boldsymbol{\kappa}^{\mathrm{T}} - 2\boldsymbol{\kappa}:\boldsymbol{Q} + \frac{D_r U\Lambda^2}{8} \times (\nabla^2\boldsymbol{S}\cdot\boldsymbol{S} + \boldsymbol{S}\cdot\nabla^2\boldsymbol{S} - 2\nabla^2\boldsymbol{S}:\boldsymbol{Q}) \quad (19)$$

where the last line represents the consequences of the spatial gradients. Derivation of the elastic stress tensor involves subtleties in treating the nonlocal interaction among molecules. Feng, Sgalari, and Leal[41] generalized the virtual work principle to cover a finite volume of the LCP, and imposed a condition of zero work on the boundary. This leads to an elastic stress tensor:

$$\sigma_E = 3\nu kT[\boldsymbol{S} - U(\boldsymbol{S}\cdot\boldsymbol{S} - \boldsymbol{S}:\boldsymbol{Q})] - \frac{\nu kTU\Lambda^2}{8} \times \left(\boldsymbol{S}\cdot\nabla^2\boldsymbol{S} - \nabla^2\boldsymbol{S}:\boldsymbol{Q} + \frac{\boldsymbol{P} - \nabla\nabla\boldsymbol{S}:\boldsymbol{S}}{4}\right) \quad (20)$$

where $P_{ij} = (\partial S_{kl}/\partial x_i)(\partial S_{lk}/\partial x_j)$. With a proper closure approximation, Eqs. (19) and (20) constitute a ''complete'' constitutive theory for LCPs that incorporates both molecular viscoelasticity and distortional elasticity.

This theory has two notable features. The nonlocality of molecular interaction is reflected by the ellipticity of Eq. (19) [cf. Eq. (15)]. Thus, the LCP configuration is globally coupled by distortional elasticity. In addition, the elastic stress tensor is asymmetric. The mean-field torque on LCP molecules amounts to a ''volume torque'' on the material, which modifies the usual conservation of angular momentum. The antisymmetric part of the stress tensor precisely balances the volume torque computed by averaging the molecular torque.[41]

Eq. (18) is actually a special form of the Marrucci–Greco potential obtained by taking a spherical interaction region around a test molecule. This corresponds, in the limit of weak distortion, to the one-constant approximation in Frank elasticity. Accounting for the molecular length in an oblong interaction region, Marrucci and Greco[39] derived a more general potential that produces three unequal elastic constants. Wang[42] has built this general nematic potential into a dynamic theory. Finally, we note several alternatives to the kinetic approach of LCP rheology.[5,24,43] Despite the different starting points, the complete theories obtained are essentially the same.

Predictions of the Complete Theories

As the complete theories include both molecular viscoelasticity and distortional elasticity, an outstanding question is: do they predict the correct dynamics for texture evolution that have eluded previous simpler models? So far, efforts to answer this have been limited to numerical simulations on shear flows.

Kupferman, Kawaguchi, and Denn[40] assumed no streamwise (along z in Fig. 1) or spanwise (along x) variations and a two-dimensional orientation with all molecules lying on the y–z plane. A one-dimensional ''windup'' picture emerges, similar to the Leslie–Ericksen predictions.[13] At low shear rates, a steady state is reached. At higher shear rates, the director continues to tumble in the interior of the domain, with the near-wall distortion being released periodically. A disappointing feature of the result is the lack of an inherent texture length scale that would reflect the balance between elastic and viscous torques.[3]

Rey and Tsuji[24,25] permitted two-dimensional variations of orientation on the y–z plane, but decoupled the flow field from the kinetic equation so that the linear velocity profile is fixed and unperturbed by LCP stress. Results reveal the conflict between the tumbling tendency far from the anchoring walls and the fixed orientation at the walls, which is resolved by periodic nucleation of a pair of defects. Again, no inherent texture length scale emerges. The common features of the above solutions—a windup structure and periodic appearance and annihilation of $\pm 1/2$ defects—are similar to the prediction of the original Doi theory.[32,37,38] Thus, a key expectation of the complete theories, namely the prediction of a texture length independent of macroscopic geometry, is not fulfilled, perhaps owing to the simplifications in these simulations.

Sgalari, Leal, and Feng[44] sought to relax the geometric and physical restrictions in simulating shear flows. The constitutive and momentum equations are fully coupled, with the flow affecting the LCP orientation, and the resulting stress tensor modifying the flow in return. Two-dimensional variations on the y–z plane are allowed, though the configuration tensor $\boldsymbol{S}$ is assumed to be symmetric about the shear plane as in Tsuji and Rey.[24] Finally, the dependence of the rotational diffusivity on the order parameter, the so-called tube dilation effect,[29] is included. Upon startup of shear, the interior of the domain tumbles as seen before. In time, however, this tumbling is arrested by distortional elasticity, and narrow strips containing much-reduced order emerge along the flow direction. These strips are roughly parallel to each other and are taken to be disclination lines. A characteristic texture length h_{text} is identified from the Fourier spectrum of the order parameter profiles. h_{text} refines in time and approaches an equilibrium level in about 100 strain

units (Fig. 4A). Marrucci's argument for the inherent texture length,[3] carried over to the complete molecular theory, yields the following scaling:

$$h_{\text{text}} \propto \Lambda\left(\frac{U}{De}\right)^{1/2} \tag{21}$$

where De is the Deborah number of the flow. The numerical data appear to be consistent with the square-root scaling (Fig. 4B). Sgalari, Leal, and Feng[44] further removed the restriction on $\boldsymbol{S}$ being symmetric about the shear plane. Then, the complete theory predicts an out-of-plane twist instability resembling the predictions of the LE theory.[12]

More recently, Sgalari, Leal, and Meiberg[45] and Klein and Leal[46] examined the instability of simple shear to secondary flows in the x–y plane. The initial instability takes on a similar form to that predicted by the LE theory, with the appearance and subsequent breakup of roll cells.[20] But the complete theory allows these authors to probe director turbulence and the Deborah-number cascade at much higher shear rates.[18] The authors argued that $\pm 1/2$ strength "thin" defects would eventually appear in addition to the escaped ± 1 disclinations noted before.

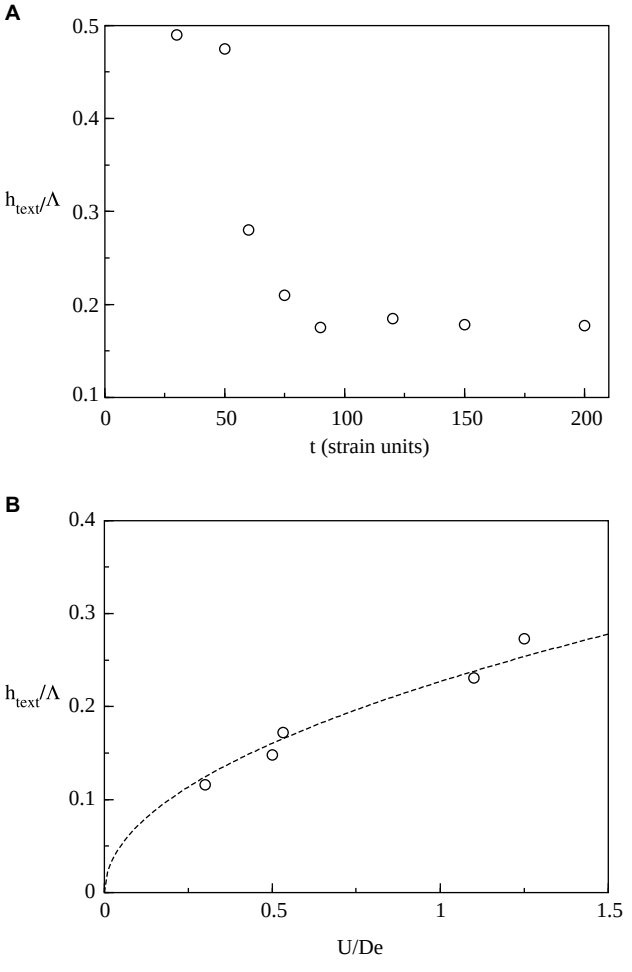


Fig. 4 (A) Evolution of h_{text} in time for $U = 8$, $De = 15$, $Er = 8 \times 10^6$, $\Lambda/H = 10^{-3}$ and (B) scaling for h_{text} when U/De is varied. The data points are numerical results and the dotted line indicates the square-root law of Eq. (21). (From Ref.[44].)

Connection Between Molecular and Continuum Theories

As alluded to in the introduction to this entry, the LE theory was conceived for small molecule LCs while molecular theories are intended for LCPs. LC molecules retain their equilibrium orientation distribution. LCPs are susceptible to disturbances to their distribution function $\Psi(\boldsymbol{u})$; its temporal relaxation gives rise to molecular viscoelasticity, while its spatial gradient produces distortional elasticity. A natural question is whether the molecular theories reduce properly to the continuum LE theory in the limiting case of an undisturbed orientation distribution. This situation arises in the "weak flow limit" where the flow is weak ($De \ll 1$) and spatial distortions are small ($|\nabla \boldsymbol{S}| \ll 1/\Lambda$). In this limit, Feng, Sgalari, and Leal[41] has proved that their version of the complete theory reduces properly to the LE theory.

Their proof relies on the fact that in the limit of small spatial distortions, The Marrucci–Greco potential of Eq. (18) reduces to the Frank elastic energy of Eq. (2) with

$$K = \frac{1}{8}\nu kTUS_{\text{eq}}^2\Lambda^2 \tag{22}$$

S_{eq} being the equilibrium order parameter. Feng, Sgalari, and Leal[41] followed the unusual perturbation procedure outlined by Kuzuu and Doi,[47] in which the base state is undetermined because a uniform single crystal does not have a preferred orientation. The kinetic equation and stress tensor reduce precisely to the LE forms, with the phenomenological elastic and viscous constants determined by equilibrium molecular parameters.

As indicated above, such an agreement is perhaps expected. On the other hand, it is remarkable that a rather complex phenomenological theory postulated for an LC continuum can be reconciled with an even more complex molecular theory built on the concept of intermolecular potential. Perhaps the only other such happy instance is the agreement between the continuum Oldroyd-B model for viscoelastic liquids and the molecular model based on a dilute suspension of linear Hookean dumbbells in a Newtonian solvent.[28]

The idea that the LE theory applies to weak flows, while the molecular theories to strong flows provides a convenient framework for interpreting experimental

observations. Larson and Mead[18] identified the so-called Ericksen number (*Er*) and the Deborah number (*De*) cascades for sheared nematic polymers. At lower shear rates, the molecular distribution remains largely undisturbed although the preferred orientation–the director–rotates. The dynamics are thus dictated by *Er*, the ratio between viscous and distortional elastic effects. The LE simulations of Feng, Tao, and Leal[20] and Tao and Feng[21] fall in this cascade, and their results indeed correlate well to observations in the *Er* cascade. Higher flow rates start to distort the molecular distribution, and molecular viscoelasticity becomes significant once *De* rises to O(1).[18] This is the *De* cascade, where distortional elasticity no longer has a major effect on certain macroscopic properties such as rheology. That is why the original Doi theory, with distortional elasticity omitted entirely, is able to reproduce the anomalous normal stress differences to quantitative accuracy.[34] For other features such as defect generation, distortional elasticity remains locally important. This is the case for the simulations of Sgalari et al.[44,45] and Klein and Leal.[46]

CONCLUSIONS

In this brief review, we strive to construct a coherent picture of our current theoretical understanding of the flow and rheology of small-molecule and polymeric nematic liquid crystals. Owing to space limitations, we have presented results selectively, based more on the need to tell a somewhat coherent story than the significance of the work.

To sum up, there have been two types of constitutive theories for LCs and LCPs: one based on phenomenological modeling that treats the material as a continuum, and the other based on a molecular picture and a statistico-mechanical approach. The molecular theories contain the continuum LE theory as a limiting case for weak flows and small spatial distortions. Shear-flow simulations using the LE theory capture qualitative features of the Ericksen number cascade, including shear-induced defect formation. More recent simulations using a complete molecular theory have reproduced some experimental observations at higher shear rates extending into the Deborah number cascades. In particular, an inherent texture length emerges and follows a scaling expected from balancing the viscous and elastic effects.

ACKNOWLEDGMENTS

Acknowledgment is made to the Donors of The Petroleum Research Fund, administered by the American Chemical Society, for partial support. This work was also supported by grants from the US-NSF (CTS-0229298; CTS-9984402), the NSERC and the Canada Research Chair program and the NNSF of China (No. 20174024 and No. 20490220).

REFERENCES

1. de Gennes, P.G.; Prost, J. *The Physics of Liquid Crystals*, 2nd Ed.; Oxford University Press: New York, 1993.
2. Larson, R.G. *The Structure and Rheology of Complex Fluids*; Oxford University Press: New York, 1999.
3. Marrucci, G.; Greco, F. Flow behavior of liquid crystalline polymers. Adv. Chem. Phys. **1993**, *86*, 331–404.
4. Rey, A.D.; Denn, M.M. Dynamical phenomena in liquid–crystalline materials. Ann. Rev. Fluid Mech. **2002**, *34*, 233–266.
5. Beris, A.N.; Edwards, B.J. *Thermodynamics of Flowing Systems with Internal Microstructure*; Oxford University Press: New York, 1994.
6. Ericksen, J.L. Hydrostatic theory of liquid crystals. Arch. Rational Mech. Anal. **1962**, *9*, 371–378.
7. Ericksen, J.L. On equations of motion for liquid crystals. Quart. J. Mech. Appl. Math. **1976**, *29* (2), 203–208.
8. Leslie, F.M. Some constitutive equations for liquid crystals. Arch. Rational Mech. Anal. **1968**, *28*, 265–283.
9. Kleman, M. Defects in liquid–crystalline polymers. MRS Bull. **1995**, *20* (9), 23–28.
10. O'Rourke, M.J.E.; Thomas, E.L. Morphology and dynamic interaction of defects in polymer liquid crystals. MRS Bull. **1995**, *20* (9), 29–36.
11. Chandrasekhar, S. *Liquid Crystals*, 2nd Ed.; Cambridge University Press: New York, 1992.
12. Zuniga, I.; Leslie, F.M. Shear-flow instabilities in non-flow-aligning nematic liquid crystals. Liq. Cryst. **1989**, *5*, 725–734.
13. Carlsson, T. Theoretical investigation of the shear flow of nematic liquid crystals with the Leslie viscosity $\alpha_3 > 0$: hydrodynamic analogue of first order phase transition. Mol. Cryst. Liq. Cryst. **1984**, *104*, 307–334.
14. Han, W.H.; Rey, A.D. Dynamic simulations of shear-flow-induced chirality and twisted-texture transitions of a liquid–crystalline polymer. Phys. Rev. E. **1994**, *49*, 597–613.
15. Manneville, P.; Dubois-Violette, E. Shear flow instability in sheared nematic liquids: theory steady simple shear flows. J. Phys. Paris **1976**, *37*, 285–296.
16. Larson, R.G. Roll-cell instabilities in shearing flows of nematic polymers. J. Rheol. **1993**, *37*, 175–197.

17. Larson, R.G.; Mead, D.W. Development of orientation and texture during shearing of liquid–crystalline polymers. Liq. Cryst. **1992**, *12*, 751–768.
18. Larson, R.G.; Mead, D.W. The Ericksen number and Deborah number cascades in sheared polymeric nematics. Liq. Cryst. **1993**, *15*, 151–169; Corrigendum **1996**, *20*, 265.
19. Mather, P.T.; Pearson, D.S.; Larson, R.G. Flow patterns and disclination-density measurements in sheared nematic liquid crystals. II. Tumbling 8CB. Liq. Cryst. **1996**, *20*, 527–538.
20. Feng, J.J.; Tao, J.; Leal, L.G. Roll cells and disclinations in sheared nematic polymer. J. Fluid Mech. **2001**, *449*, 179–200.
21. Tao, J.; Feng, J.J. Effects of elastic anisotropy on the flow and orientation of sheared nematic liquid crystals. J. Rheol. **2003**, *47* (4), 1051–1070.
22. Ericksen, J.L. Liquid crystals with variable degree of orientation. Arch. Rational Mech. Anal. **1991**, *113*, 97–120.
23. Liu, C.; Walkington, N.J. Approximation of liquid crystal flows. SIAM J. Numer. Anal. **2000**, *37*, 725–741.
24. Tsuji, T.; Rey, A.D. Effect of long range order on sheared liquid crystalline materials. Part 1: compatibility between tumbling behavior and fixed anchoring. J. Non-Newton. Fluid Mech. **1997**, *73*, 127–152.
25. Rey, A.D.; Tsuji, T. Recent advances in theoretical liquid crystal rheology. Macromol. Theory Simul. **1998**, *7*, 623–639.
26. De Gennes, P.G. Phenomenology of short-range-order effects in the isotropic phase of nematic materials. Phys. Lett. **1969**, *35A*, 454–455.
27. Larson, R.G.; Doi, M. Mesoscopic domain theory for textured liquid–crystalline polymers. J. Rheol. **1991**, *35*, 539–563.
28. Bird, R.B.; Curtis, C.F.; Armstrong, R.C.; Hassager, O. *Dynamics of Polymeric Liquids. Vol. 2. Kinetic Theory*; Wiley: New York, 1987.
29. Doi, M.; Edwards, S.F. *The Theory of Polymer Dynamics*; Oxford University Press: New York, 1986.
30. Doi, M. Molecular dynamics and rheological properties of concentrated solutions of rodlike polymers in isotropic and liquid crystalline phases. J. Polym. Sci. Polym. Phys. Edn. **1981**, *19*, 229–243.
31. Larson, R.G. Arrested tumbling in shearing flows of liquid crystal polymers. Macromolecules **1990**, *23*, 3983–3992.
32. Feng, J.; Chaubal, C.V.; Leal, L.G. Closure approximations for the Doi theory: which to use in simulating complex flows of liquid–crystalline polymers? J. Rheol. **1998**, *42*, 1095–1119.
33. Grosso, M.; Maffettone, P.L.; Halin, P.; Keunings, R.; Legat, V. Flow of nematic polymers in eccentric cylinder geometry: influence of closure approximations. J. Non-Newton. Fluid Mech. **2000**, *94* (2–3), 119–134.
34. Magda, J.J.; Baek, S.G.; DeVries, K.L.; Larson, R.G. Shear flows of liquid crystal polymers: measurements of the second normal stress difference and the Doi molecular theory. Macromolecules **1991**, *24*, 4460–4468.
35. Marrucci, G.; Maffettone, P.L. Description of the liquid–crystalline phase of rodlike polymers at high shear rates. Macromolecules **1989**, *22*, 4076–4082.
36. Suen, J.K.; Nayak, R.; Armstrong, R.C.; Brown, R.A. A wavelet-Galerkin method for simulating the Doi model with orientation-dependent rotational diffusivity. J. Non-Newton. Fluid Mech. **2003**, *114* (2–3), 197–228.
37. Feng, J.; Leal, L.G. Simulating complex flows of liquid–crystalline polymers using the Doi theory. J. Rheol. **1997**, *41*, 1317–1335.
38. Feng, J.; Leal, L.G. Pressure-driven channel flows of a model liquid–crystalline polymer. Phys. Fluids **1999**, *11*, 2821–2835.
39. Marrucci, G.; Greco, F. The elastic constants of Maier–Saupe rodlike molecule nematics. Mol. Cryst. Liq. Cryst. **1991**, *206*, 17–30.
40. Kupferman, R.; Kawaguchi, M.N.; Denn, M.M. Emergence of structure in a model of liquid crystalline polymers with elastic coupling. J. Non-Newton. Fluid Mech. **2000**, *91*, 255–271.
41. Feng, J.J.; Sgalari, G.; Leal, L.G. A theory for flowing nematic polymers with orientational distortion. J. Rheol. **2000**, *44* (5), 1085–1101.
42. Wang, Q. A hydrodynamic theory for solutions of nonhomogeneous nematic liquid crystalline polymers of different configurations. J. Chem. Phys. **2002**, *116* (20), 9120–9136.
43. Lhuillier, D. Thermo-mechanical modelling of nematic polymers. In *Continuum Thermodynamics*; Maugin, G.A., Drouot, R., Sidoroff, F.S., Eds.; Kluwer: Dordrecht, 2000; 237–246.
44. Sgalari, G.; Leal, L.G.; Feng, J.J. The shear flow behavior of LCPs based on a generalized Doi model with distortional elasticity. J. Non-Newton. Fluid Mech. **2002**, *102*, 361–382.
45. Sgalari, G.; Leal, L.G.; Meiberg, E. Texture evolution of sheared liquid crystalline polymers: numerical predictions of roll-cells instability, director turbulence, and striped texture with a molecular model. J. Rheol. **2003**, *47* (6), 1417–1444.
46. Klein, D.H.; Leal, L.G. Computational studies of the shear flow behavior of models for liquid crystalline polymers. AIChE Annual Meeting, San Francisco, November 16–21, 2003.
47. Kuzuu, N.; Doi, M. Constitutive equation for nematic liquid crystals under weak velocity gradient derived from a molecular kinetic equation. J. Phys. Soc. Japan **1983**, *52*, 3486–3494.

Thermal Analysis Techniques: Overview

John O. Hill
La Trobe University, Melbourne, Victoria, Australia

INTRODUCTION

The techniques that constitute the field of ''thermal analysis'' are reviewed in terms of the changes in a specified physical property of a material being monitored as a function of temperature, the instrumentation involved, and selected applications in the fields of materials science and chemical processing. The basic thermal analysis techniques—thermogravimetric analysis (TGA), differential thermal analysis (DTA), differential scanning calorimetry (DSC), thermomechanical analysis (TMA), and dynamic mechanical thermal analysis (DMTA) —together with some ''less common'' thermal analysis techniques such as thermal emanation analysis (TEA), thermosonimetry (TS), and dielectric thermal analysis (DETA) are discussed in detail. The application of these techniques in various fields indicates that ''thermal analysis'' is recognized as a major analytical science in its own right and together with calorimetry comprises the core of the experimental thermal sciences.

BACKGROUND

The generic title ''thermal analysis'' is inclusive of a range of analytical techniques, which all measure the change in a physical property of a material as a function of temperature. Thermal analysis can now be considered as a ''mature science,'' having been continuously developed for more than 11 decades. A variety of developments were seen particularly in the last decade. Many of these developments are directly related to the associated techniques, which include new analytical methodologies such as temperature modulation, sample controlled regimes, on-line data analysis, and ''robotic'' operational regimes. Major instrumental developments have also taken place, which enabled the analysis of ''small'' sample sizes over an ''extended'' temperature range. Essentially, the development in the field of thermal analysis over the last decade has revolutionized materials science and the chemical processing industry, because thermal analysis techniques have been intensively used to study and characterize the thermal behavior of a wide range of materials, of both natural and synthetic origins. They have played a pivotal role in both the processing of chemical materials and in quantifying the quality control procedures inherently associated with such operations on an industrial scale. The aim of this overview is to discuss the full range of thermal analysis techniques in terms of the type of measurement involved in each technique, the application range of that technique, and the interpretation of data obtained on using the technique. The applications are chosen with specific reference to the broad concept of ''chemical processing,'' so as to align with the overall objectives of this encyclopedia.

THE BASIC THERMAL ANALYSIS TECHNIQUES

It has become increasingly difficult to differentiate between the fields of thermal analysis and calorimetry, and some thermal analysis techniques can be classified as calorimetric techniques as well. One such example is the ''differential scanning calorimetry.'' Thermal methods of analysis have been developed in accordance with the need to study the changes in the properties of a material as a consequence of heating, whereas calorimetric methods measure such heat changes directly. Thermal analysis techniques and the associated fields of application are uniquely complementary. Thermal analysis investigates changes in a wide range of physical properties of a material with change in temperature. The change in property may be physical, such as a mass, linear, or three-dimensional change, melting, crystalline transition, or vaporization, or it may be a chemical change involving a reaction that alters the chemical structure of a material. The change may also be biological when related to the metabolism, interaction, or decomposition. Thus, the application field of thermal analysis is wide, diverse, and of paramount analytical significance in a wide range of areas, particularly in chemical processing.

Hemmingner[1] has recently defined ''thermal analysis'' as ''the analysis of a change in a property of a sample which is related to an imposed temperature alteration.'' The ''temperature alteration'' can involve a stepwise change from one constant temperature to another, a linear rate of change of temperature, modulation of a

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120041306

constant or linearly changing temperature with constant frequency and amplitude, and uncontrolled heating or cooling. The direction of temperature change may involve either heating or cooling and the modes of operation of temperature change may be combined in any sequence—adding to the diversity of the operational modes of thermal analysis techniques. The temperature may also be programmed to maintain a constant rate of reaction—leading to ''sample controlled thermal analysis.''

As there are many properties of a material that can be measured, the number and range of thermal analysis techniques is large. However, it is generally recognized that there are five basic thermal analysis techniques—TGA, DTA, DSC, TMA, and DMTA.

Thermal analysis instrumentation is complex but all thermal analysis instruments have common features, as shown schematically in Fig. 1.

Such instruments usually comprise four major parts: the sample in a container or holder, sensors (transducers) to detect and measure a particular property of the sample as a function of temperature, a sample enclosure within which the experimental parameters, such as temperature, pressure, and gas atmosphere, can be controlled, and a computer to control the experimental parameters and to collect and process the data from the sensors to produce meaningful results and cumulative records. Measurements are usually continuous and the heating rate is often, but not necessarily, linear with time. The results of such measurements are generated as thermal analysis curves and the features or characteristics of these curves, such as peaks, discontinuities, and slope changes, are related to thermal events in the sample, thereby characterizing its thermal behavior. Many analytical techniques require samples in a specified ''form,'' thereby imposing a pretreatment process. Pretreatment essentially destroys the structure of the matrix containing the sample. Thermal analysis methods generally require no significant pretreatment, and thereby the integrity of the sample is retained and its

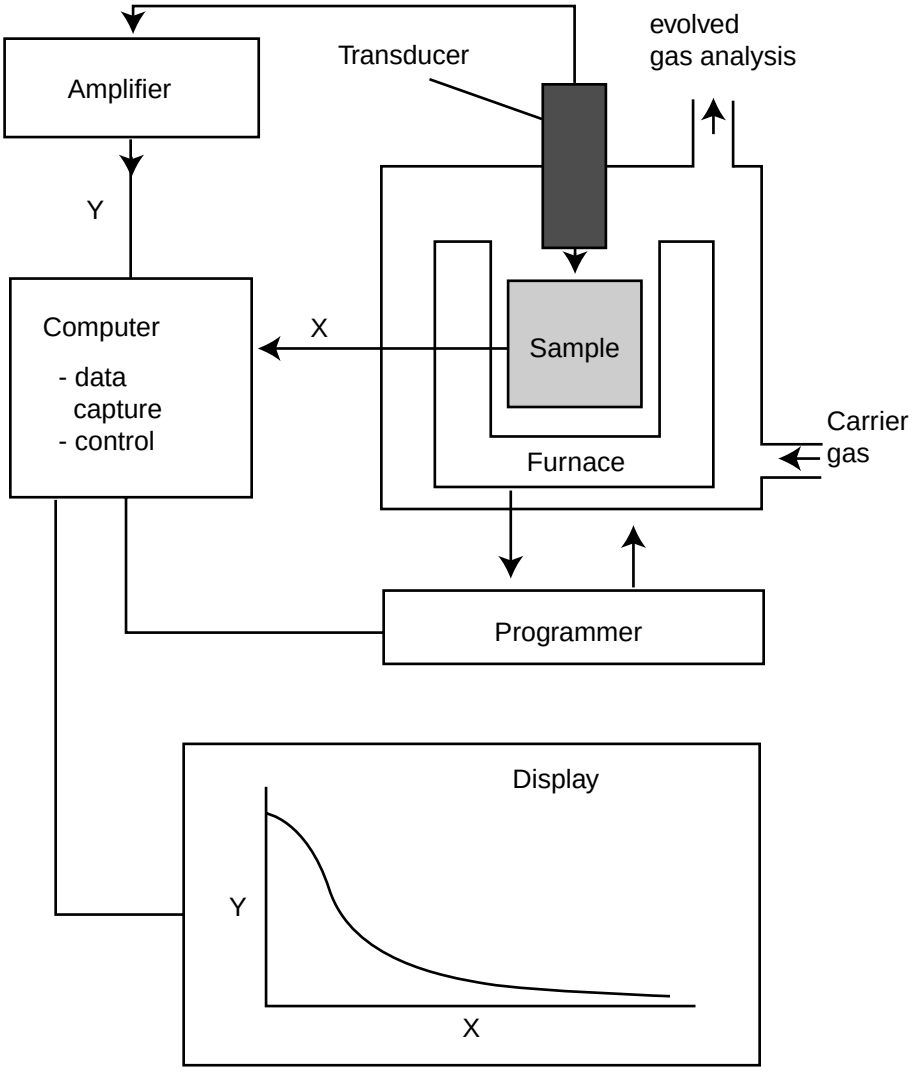


Fig. 1 Schematic diagram of a thermal analysis system.

thermal and molecular history is not changed prior to analysis. The information received from the subsequent analysis can therefore be related directly to the sample, ''as received,'' and to the situation or process where it is actually used. This feature has major significance in the chemical processing technology.

It is important in thermal analysis, as for any other analytical techniques, to obtain reproducible results. However, there are many experimental variables associated with thermal analysis and it is necessary to define at least five of these for each thermal analysis investigation. These five variables are the nature of the sample and its container (crucible), the heating (or cooling) rate, the sample-chamber atmosphere, and the sample mass. Some thermal analysis techniques require the recording of other variables, such as the load on the sample in thermomechanical analysis. In thermal analysis, the ''samples'' are usually in the solid state, but liquids can also be studied using special sample preparation techniques. Gases are not normally studied by thermal analysis techniques.

It is generally accepted that the ''basic'' thermal analysis techniques are TGA, DTA, DSC, and the thermomechanical techniques such as TMA and DMTA. Thermogravimetric analysis involves the measurement of the change in mass of the sample with change of temperature, DTA involves the measurement of the temperature difference between the sample and a thermally inert reference material with change of temperature, DSC involves the measurement of the differential power input to the sample with change of temperature, TMA involves measurement of the linear dimensional change of the sample (under static load) with change of temperature, and DTMA involves measurement of the dimensional change of the sample (under dynamic load) with change of temperature. Differential scanning calorimetry is essentially ''quantitative DTA.'' Thermogravimetric analysis and DTA may be combined into a single thermal analysis system, giving rise to the simultaneous thermal analysis (STA). It is also possible to couple TGA with DSC to produce a simultaneous TGA–DSC system. There are advantages and disadvantages of simultaneous thermal analysis systems. The principal advantage of such systems is that two types of measurement are made on the same sample over the same time interval. The disadvantage is that the sensitivity of the coupled thermal analysis systems is generally less than the individual systems operated independently and, particularly for TGA–DSC, the effective operational temperature range is restricted.

Thermal analysis systems require calibration both in terms of temperature and the physical property being measured, so as to detect and eliminate the inherent systematic errors associated with the analysis. Thus, for example, TGA systems require both temperature and mass calibration in accordance with the defined procedures for calibration of thermal analysis systems.[2]

Thermogravimetric analysis involves the measurement of the change in sample mass with change in temperature, using a ''thermobalance.'' A thermobalance is a combination of a suitable electronic microbalance with a furnace, a temperature programmer, and an on-line computer for control of the overall experiment and data analysis. The system allows for the sample to be simultaneously weighed and programmed and heated or cooled, and the mass, time and temperature data to be recorded and processed. The balance should be in a suitably enclosed system so that the nature and pressure of the atmosphere surrounding the sample can be controlled. The microbalance mechanism is maintained at or close to the ambient temperature throughout the experiment.

In TGA, mass loss is observed if a thermal event involves the loss of a volatile component. Processes such as melting do not involve any mass loss. These may be studied using DTA and DSC. Thermogravimetric analysis data are presented as a plot of mass, m vs. temperature, T, or time, t. The TGA curve shape is sigmoid, indicating that most of the mass loss occurs at a particular temperature or over a narrow temperature range where the curve has greatest slope. However, the thermal event commences before the region of greatest mass loss and continues thereafter. A more convenient representation of the character of the overall thermal event is to give the derivative of the original TGA curve as dm/dt, or rate of mass loss, and to plot this vs. temperature, T, or time, t. Thereby, the DTG curve is obtained and thus, the spread of the thermal event over a broad temperature range appears as a peak. The DTG curve is of particular assistance if two or more thermal events overlap within similar temperature ranges. Double peaks or a shoulder on a main peak appear in such cases. Slow processes, superimposed with other fast processes, appear as gradient changes in the DTG curve. Thus the DTG curve can kinetically characterize overlapping thermal events. Further, the DTG peak area is directly proportional to the corresponding mass loss; therefore, relative mass losses may be compared. The peaks in a DTG curve can act as ''fingerprints'' of components in a mixture by comparison with DTG curves of the suspected individual components.

It is common practice to check the performance of a TGA system by running a sample of calcium oxalate monohydrate. This salt is known to thermally decompose in three stages over well-defined temperature ranges. The first step involves the loss of the single water of hydration molecule followed sequentially by the conversion of anhydrous calcium oxalate to calcium carbonate with the loss of carbon monoxide, and thence the decomposition of calcium carbonate to calcium oxide with the evolution of carbon dioxide.

Thermogravimetric analysis applications are extensive and diverse and include oxysalt decompositions, natural and synthetic polymer characterization, metal oxidation and corrosion analysis, compositional analysis of coals, polymers, and rubbers, study of glass-making processes, and a wide range of chemical processing phenomena. Thermogravimetric analysis has been used very successfully to study the kinetics of chemical processes.[3,4] A quintessential TGA application has been the determination of the proximate analysis of coal. The procedure involves the in situ determination of the volatile components under an inert gas atmosphere and the combustible materials under an oxidizing atmosphere. The TGA system is fitted with a ''gas-switching'' devise to effect the overall analysis as a single run. The principal advantage of the TGA procedure is that the proximate analysis of a coal sample is achieved in less than 1 hr, whereas the conventional procedure, using classical analytical procedures, takes several hours. Using a 10 mg sample, the moisture loss is determined by heating to 110°C and the volatiles are determined by heating to 900°C, both determined in an atmosphere of pure, oxygen-free nitrogen. The carbon content is then measured by switching to an oxygen atmosphere and the residue mass gives the ash content. The entire analysis takes less than 1 hr.

Differential thermal analysis is the simplest and most widely used thermal analysis technique. The difference in temperature between the sample and a reference material is measured while both are subjected to the same heating program. In ''classical'' DTA, as represented schematically in Fig. 2A, a single block with symmetrical cavities or inserts for the sample and reference is heated in the surrounding furnace. The block is effectively a heat sink and ensures a measurable differential temperature signal during a thermal event. In conventional DTA, as represented schematically in Fig. 2B, the sample and reference material are contained in separate crucibles and both are subjected to the same temperature program.

If an endothermic event occurs in the sample, the temperature of the sample lags that of the reference, which follows the heating program. A typical DTA curve for such a thermal event is shown in Fig. 2C. If an exothermic process occurs in the sample, the response will be in the opposite direction. The negative peak shown in Fig. 2C is termed an ''endotherm'' and is characterized by its onset temperature. The area under a DTA peak is proportional to the enthalpy associated with the thermal event. Because such proportionality is not linear with temperature, only approximate enthalpy data can be obtained from DTA. The nature of the reference material is important. It must undergo no thermal events over the DTA operational temperature range and it must not chemically react with its container. Alumina has been

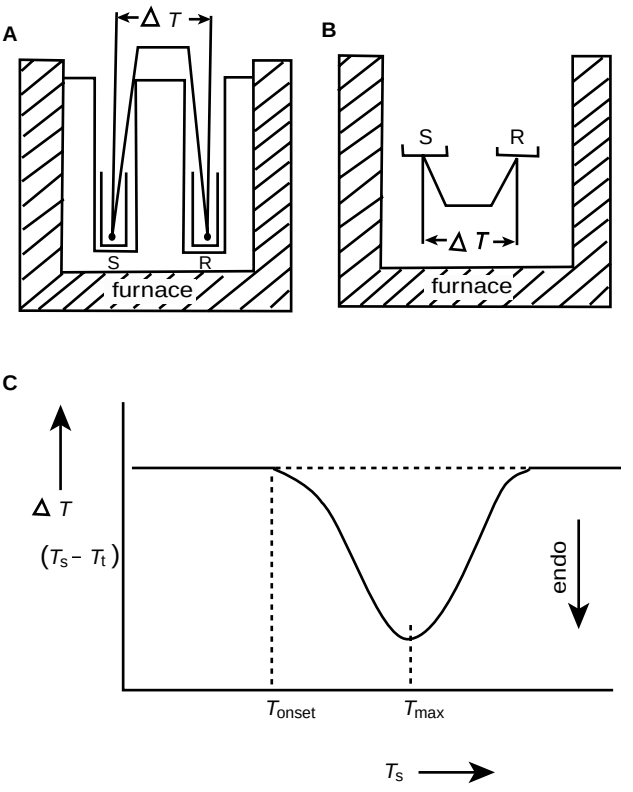


Fig. 2 Schematic diagram of a classical DTA system (A), a heat-flux DTA system (B), and a typical DTA curve (C).

extensively used as the reference material in DTA. Differential thermal analysis systems are calibrated using the ICTAC certified reference materials[2] consisting of a suite of metals with certified melting points.

Differential thermal analysis applications are extensive and wide ranging, which include the determination of melting points and the characterization of polymers, glasses, ceramics, oils, fats, waxes, clays, minerals, coals, lignite, wood, liquid crystals, explosives, propellants, pyrotechnics, pharmaceuticals, biological materials, metals, alloys, natural products, and synthetic catalysts. This list is not exhaustive. The most significant application of DTA has been the characterization of polymers and the most extensive thermal analysis study of synthetic polymers relates to polyethyleneterephthalate (PET).[5] The DTA curve of PET can be used to determine the glass transition temperature, the crystallization temperature range, and the melting range of the crystalline phase. Thus, DTA is a simple method for the complete thermal analysis of polymers and, additionally, assesses polymer thermal stability.

There are two types of DSC—heat-flux DSC and power-compensated DSC. The former is essentially a quantitative DTA. In power-compensated DSC, the aim is to maintain the sample and reference material at the same temperature throughout the controlled temperature program. The difference in the independent

supplies of power to the sample and reference material (alumina) is then recorded vs. the programmed reference temperature. A schematic of a power-compensated DSC is shown in Fig. 3, together with a representative DSC curve.

Thermal events in the sample thus appear as deviations from the baseline, in either an endothermic or exothermic direction, depending upon whether more or less energy is supplied to the sample compared with the reference material. The operational temperature range of DSC is less than that of DTA—being typically subambient to 750°C. Temperature and energy calibration of DSC instruments is achieved using the ICTAC certified reference materials.[2]

Differential scanning calorimetry has an application range similar to that of DTA but also offers additional application possibilities, the most important of which are the determinations of heat capacity and thermal conductivity of a wide range of materials. This facility makes DSC a very significant analytical technique in chemical processing. Differential thermal analysis is also used to characterize polymers and to study the curing characteristics of resins. It therefore has a widespread application in materials science. As an example, DSC has been used to identify the constituents of a plastic waste containing low density polyethylene (LPDE), high density polyethylene (HPDE), polypropylene (PP) and poly(tetrafluoroethylene) (PTFE)—''Teflon.''[6] The DSC peak areas are directly proportional to amounts of each constituent present in the waste mixture and thus, from a single DSC analysis, a complete analysis of such a sample is possible.

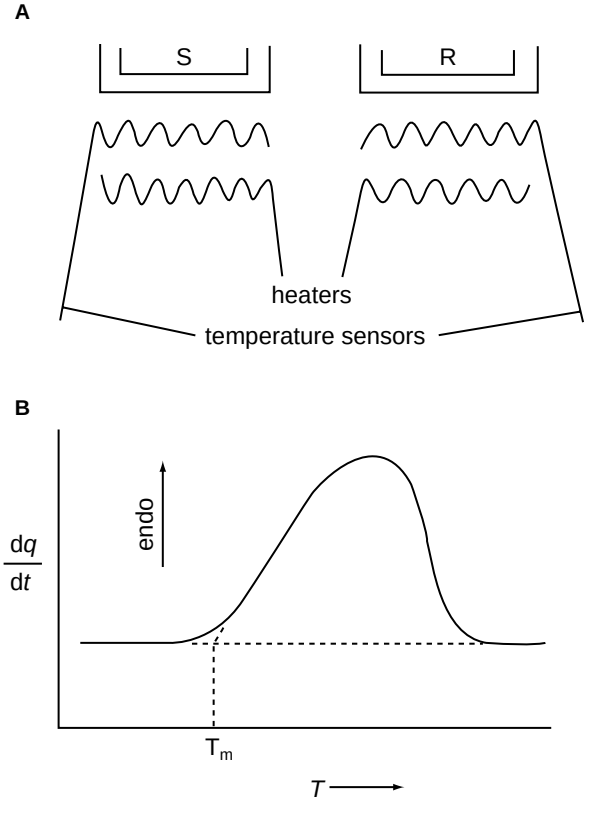


Fig. 3 Schematic diagram of a power-compensated DSC system (A) and a typical DSC curve (B).

Thermomechanical analysis embraces four different modes of measuring the dimensional change of a material as a function of temperature—penetration, extension, flexure, and torsion. The closely related technique of thermodilatometry is effected by measuring the expansion and contraction of a material under negligible load. Additional mechanical data may be obtained by measuring the penetration (the expansion or contraction of a sample while under compression) as a function of temperature. These techniques, together with flexure and torsion measurements, are collectively classified as TMA and are of paramount importance in materials characterization and testing. Thermomechanical analysis systems are calibrated using materials of known coefficient of expansion, such as aluminum,[7] and it is necessary to compensate for the many dimensional changes in the instrument as heating takes place. Temperature calibration in TMA is effected by using a suite of metals, indium, tin and bismuth.[7]

Applications of TMA are numerous and are largely confined to materials science. Thermomechanical analysis is widely used to measure the coefficient of thermal expansion of materials, to determine the glass transition and melting and crystallization characteristics of polymers, and for the determination of the structural integrity of films and fibers. Thus, TMA has been applied to study the change in length of a sample of silicone gum rubber under different applied loads.[8] At zero force, a change in length of the sample is apparent at $-60°C$ and is because of the sample undergoing a change from glassy to rubbery behavior. At this temperature, the polymer chains acquire additional degrees of mobility that is manifested as an increase in thermal expansion coefficient. The glass transition temperature can be defined by the intercept of tangents to the linear portions of the length vs. temperature plot, above and below this region. When a force is applied to the specimen, the probe deforms the material in inverse proportion to its stiffness. Below the glass transition, the polymer is rigid and is able to resist the applied force, and thus its deformation is negligible. Above the glass transition, the polymer becomes soft and the probe penetrates into the bulk of the specimen. The temperature at which this occurs is the ''softening temperature'' and is dependent on the applied force.

In dynamic mechanical analysis (DMA), the sample is subjected to a periodically varying stress (of usually sinusoidal or angular frequency). The response of the sample to this treatment provides information on the

''stiffness'' of the material, as quantified by its elastic moduli, and its ability to dissipate energy, as measured by its ''damping.'' For a viscoelastic material, the strain resulting from the periodic stress will also be periodic, but will be out of phase with the applied stress as a result of energy dispersion as heat or damping in the sample. The phase angle between the stress and the strain is an important parameter in DMA.

Dynamic mechanical analysis is routinely used to investigate the morphology of polymers, composites, and a wide range of other materials. The technique is particularly sensitive to low energy transitions, which are not readily observed by DSC. Many of these processes are time-dependent and by using a range of mechanical deformation frequencies, the kinetic characteristics of these processes can be investigated. A typical DMA application is the study of the thermal properties of polyesters.[5] The aliphatic polyester, poly(caprolactone), is typically highly crystalline and melts between 50°C and 60°C. However, the sample is not completely crystalline and contains domains of amorphous material which undergoes a glass–rubber transition at −40°C. The peaks in Tan delta at −85°C and −130°C correspond to the beta and gamma transitions of the polymer, respectively, and are because of the motion of short lengths of the polymer backbone rather than to a large scale increase in mobility that accompanies the glass–rubber transition. It is difficult to measure this type of behavior by DSC but DMA is able to detect such minor transitions, as it is able to monitor those that involve dissipation of mechanical energy as heat.

LESS COMMON THERMAL ANALYSIS TECHNIQUES

''Less common'' thermal analysis techniques encompass those that monitor changes in the ''less obvious'' properties of a material as a function of temperature and hence these are specialized techniques, which require specialized and sophisticated equipment. ''Less common techniques'' are not synonymous with ''less used'' techniques, but relate to a field of thermal analysis that is less well developed as a result of being highly specialized.

Emanation thermal analysis (ETA) involves the measurement of the release of inert (usually radioactive) gas from a solid sample, as a function of temperature. The rate of such gas release is essentially an indication of the changes taking place in the sample, and a comparison of ETA data with those of other thermal analysis techniques, particularly TGA and evolved gas analysis (EGA), provides information on the microstructure of the sample material. However, most of the solid samples studied by ETA are ''spiked'' with inert gas, which is trapped at lattice imperfections and the latter also provide diffusion pathways for adsorbed gases. It is this spiking procedure that characterizes ETA as a less common TA technique. Several methodologies are available to effect the required ''spiking.'' These are essentially divided into two groups: techniques for introducing the parent nuclide of an inert gas and techniques for introducing the inert gas itself. The former group of spiking techniques is preferable, because labeling the sample with the parent nuclide gives an enhanced time frame over which the inert gas can be desorbed from the sample, as it is formed continuously in situ. With the second group of techniques, the absorbed inert gas may be lost from the sample over the time frame of a single run. The development of ETA has been largely because of Balek and Tolgyseey.[9]

Emanation thermal analysis is usually carried out in conjunction with other TA techniques, most notably DTA and EGA. In this context, ETA can be considered as a ''coupled TA technique.'' Carrier gas at a constant flow rate is used to carry the released gas from the sample to appropriate detectors—usually radioactive counting devices. In the case of desorbed radon, a scintillation counter is used, whereas Geiger counters are used for krypton, xenon, and argon.

An ETA curve is a plot of ''emanating power'' E as a function of time and the E/time relationship has been developed and defined by Balek and Tolgyseey.[9] A major difficulty of ETA is that preparation and handling of samples require sophisticated radiochemical facilities coupled with all the associated precautions. However, the net amounts of radioactive gas incorporated into samples are so small that the evolved gas, after dilution with the carrier, does not pose a significant hazard.

The applications of ETA are numerous and have been discussed by Balek,[10] and one of the major applications is the characterization of powders. Emanating power is related directly to surface area and hence changes in grain size and the occurrence of sintering during heating can readily be detected. Kinetic parameters related to sintering can also be derived. Solid-state phase changes can also be manifested as changes in the emanating power. In this context, ETA data can be correlated with DTA, DSC, and TMA data that are related to phase changes of solid-state materials. Emanation thermal analysis has also been applied to study solid–gas reactions, such as the oxidation of metals and the reduction of metal oxides. In this respect, ETA data are useful for studying surface phenomena and hence ETA has potential for future applications in the study of corrosion and catalytic activity and catalytic processes. Solid–solid reactions, such as spinel formation between zinc oxide and iron(1 1 1) oxide can be studied using ETA, by

labeling ZnO with a radioactive thorium isotope. It is apparent that both the reactant oxides interact in a series of stages involving an initial surface diffusion process and a final bulk volume diffusion process. These processes are not readily detected by other TA techniques such as DTA or TMA, but are clearly revealed by increased emanation with clear definitions of the corresponding temperature ranges. The reactivity of oxides can be readily determined by ETA in terms of the mutual relativities toward spinel formation. Emanation thermal analysis therefore has major significance as an analytical technique in materials science.

Thermosonometry involves the measurement of sound intensity emitted by a sample as a function of temperature. Sound emission originates from the release of thermal stresses in the sample, such as movement of dislocations, generation and propagation of cracks, nucleation of new phases, relaxation processes, and discontinuous changes in physical properties. For example, at a glass transition temperature, a discontinuous change in free volume generates elastic waves that cause an acoustic effect. Such stress relief processes are not usually detectable by conventional TA techniques because of the low associated energy, and hence TS has many unique applications, such as the detection of radiation damage, crystal defect content, and degree of annealing of polymeric samples. It is also complementary to TMA for the detection of mechanical effects associated with melting, dehydration, and decomposition and for studying a range of solid-state phase changes.

Thermosonometry is usually carried out in conjunction with DTA and sound energy is emitted as mechanical vibrations prior to and during the thermal events associated with the sample over defined temperature ranges. Such sonic activity is detected by a specially adapted stethoscope. The mechanical waves are converted to electrical signals by conventional piezoelectric transducers. The stethoscope is constructed of fused silica for operation up to 1000°C and for higher operating temperatures, ceramics or noble metals are used. The sample is contained in the DTA sample head, which in turn acts as an acoustic transformer and is connected via a transmitting rod to a piezoelectric sensor, fixed on a recoil foundation and a seismic mount to prevent interference from external noise. A typical TS–DTA system has been described in detail by Clark.[11] Several special aspects of the instrumentation enhance sensitivity. The waveguide system transmits the acoustic emissions from the heated sample to the transducer at an ambient temperature so as to nullify the effect of temperature on the transducer, and the sample temperature thermocouple is located close to, but not in direct contact with the sample. This is because, the insertion of the sensor into the sample can cause mechanical damping effects, which directly affect acoustic activity.

The output of a TS experiment consists of a rapid cascade of decaying signals that may be recorded in several ways: as the number of signals of peak amplitude greater than a set threshold value in a given time, the time for which the signal amplitude exceeds the threshold value, the time frequency that signals pass through a preselected voltage level in a positive direction, or the root-mean-square amplitude of energy or a set of frequencies. In general, TS data are complex and difficult to interpret. Attempts have been made to relate frequency distributions to processes (thermal events) occurring in the sample. In the simplest interpretation, frequency distributions can be used as ''fingerprints'' of sample origin. Superficial interpretation of TS data is also possible, in that ''bubble release'' is associated with low frequencies and crystal fracture or distortion is associated with high frequencies. In general, TS is a powerful technique for rationalizing a wide variety of solid-state phenomena and is hence a valuable analytical technique in materials science.

Significant TS applications have been described by Butteridge et al.,[12] Bellchamber et al.,[13] and Shimada,[14] and one example is included here. Thermosonimetry has been applied to study the dehydroxylation behavior of kaolin, which show two regions of acoustic activity. The low temperature region corresponds to dehydroxylation and the high temperature region corresponds to recrystalization as metakaolinite. The relevant temperature regions of acoustic activity are defined by the coupled DTA analysis.

Thermoelectrometry or thermoelectrical analysis (TEA) is the generic title for a group of techniques involving the measurement of electrical properties of a sample as a function of temperature. The electrical properties commonly measured in TEA are conductance, capacitance, and dielectric properties. The most prominent TEA technique is DETA, which involves measurements of both the capacitance and the conductance of a sample as a function of time, temperature, and frequency. The former is a measure of the ability of the sample to store charge, whereas the latter is a measure of its ability to transfer charge. Four parameters are associated with DETA: the permittivity, the loss factor, the dissipation factor, and the ionic conductivity. These parameters provide information related to molecular motion within the sample and, as for most TA techniques, it is the variation of these parameters during thermal events that is of primary interest, rather than their absolute values. Hence, in DETA, the sample is subjected to an oscillating sinusoidal electric field and the applied voltage produces a polarization within the sample and causes a current to flow, which in turn leads the electric field

by a phase difference delta. Dielectric thermal analysis is somewhat related to thermomechanical techniques, in that the net current flow is similar in nature to the deformation induced by applied mechanical stress and represents a measure of the freedom of charge carriers in the sample to respond to the applied field. Also, DETA is able to differentiate between different types of charge carriers in the sample. Dipoles attempt to orientate with the applied electric field, whereas ions, which are often present as impurities, migrate toward the electrode of opposite polarity. Dielectric thermal analysis is therefore significant as an analytical technique for a wide range of ''quality control'' processing operations.

The instrumentation for DETA has been described in detail by Price.[15] The sample is presented as a thin film, typically no more than 1–2 mm thick, sandwiched between two parallel plates so as to form a simple electrical capacitor. A grounded electrode surrounding one plate, known as a ''guard ring,'' is incorporated to nullify stray electric fields. The sample temperature sensor is a thermocouple or a platinum resistance thermometer, which is placed in contact with one of the plates. The sample is connected to the inverting input of an operational amplifier, configured as a current-to-voltage converter with capacitive feedback. The measured phase shift of this network is directly correlated with the dielectric properties of the sample. There are two operational modes in DETA: using a continuous range of frequencies (50 Hz to 1 MHz) at preselected temperatures or over a continuous range of temperatures at selected frequencies. The latter is the more common operating mode for the study of phase transitions and chemical reactions in a sample. Temperature calibration of DETA is effected by measuring the melting transition of a low molar mass organic material, such as benzoic acid.

Dielectric thermal analysis is a highly specialized thermal analysis technique, but is particularly useful for detecting polar impurities in nonpolar materials such as commercial polymers. Thus, DETA and the thermoelectrometric techniques, in general, occupy a specialized niche in the TA application field in the context of addressing a specific problem in the microstructure of solids. For example, the thermal analysis of thin films can be problematic because of the very small amount of material involved. The glass transition behavior of adhesives affects the tack and bonding behavior, and the analysis of adhesive films by DSC is difficult unless the film is removed from any backing support material. Dielectric thermal analysis of such films is usually undertaken by swelling the adhesive in a suitable solvent and scraping it off the supporting substrate. However, the intrinsic properties of the adhesive may change in this process, even if the sample is subsequently dried to remove residual solvent. Dielectric thermal analysis, however, is ideally suited to direct analysis of thin films and so long as the support does not undergo any thermal transitions in the temperature region of interest, DETA can measure the thermal properties of the adhesive in situ, without any sample pretreatment. The DETA curves of a thin layer of pressure-sensitive adhesive on poly(ethyleneterephthalate) film at different electrical field frequencies have been interpreted by Price.[16]

Dielectric thermal analysis essentially involves monitoring the viscosity of a system via its ability to store and transport electrical charge. Changes in the degree of alignment of dipoles, together with the ion mobility, provide information relating to physical transitions in the sample and to changes in properties such as viscosity, rigidity, reaction rate, and cure state. Dielectric thermal analysis is particularly useful for characterizing the cure process of polymers. Polymers contain ionic impurities and the application of a voltage between two electrodes creates an electric field, which forces such ions to migrate to the oppositely charged electrode. However, the conductivity of the material depends on the magnitude of the viscous drag experienced by these ionic impurities. Ions moving through polymerized materials have low mobility and conductivity corresponding to the high viscosity of the medium. As the degree of cure in synthetic polymers is proportional to the degree of polymer cross-linking, the cure process can be monitored by monitoring the conductivity of the material throughout the entire curing process and DETA is able, more so than any other TA technique, to determine the ''cure endpoint.'' Dielectric thermal analysis applications to the curing process of synthetic polymers have been described in detail elsewhere.[16]

Dielectric thermal analysis can also be used very effectively for the in situ monitoring of the UV degradation of adhesives such as tinted selfadhesive plastic films, used to screen out direct sunlight from penetrating buildings and vehicles. Attached to the window glass, the adhesive directly receives solar radiation and must be stabilized against photodegradation by a suitable blend of polymer and stabilizer. A modified form of DETA, in which the sample temperature sensor is an interdigitized single surface dielectric sensor, can be used to study such systems.[16] The window film samples are mounted on glass plates for accelerated ageing. A small hole is cut into the plastic film and the interdigitized sensor is applied to the exposed surface of the adhesive after freezing the plate to aid the removal of the backing. Measurements of the dielectric and permittivity factors are made at an ambient temperature, as a function of frequency. Such measurements correlate UV aging of the film with changes in dielectric behavior of the adhesive and hence, optimized polymer/stabilizer formulations can

be identified. Dielectric thermal analysis is thus a powerful analytical technique in the processing of a wide range of polymer preparations.

CONCLUSIONS

This brief overview about the field of thermal analysis has concentrated on discussing some of the experimental techniques that constitute this analytical domain. Techniques such as differential thermal analysis were developed more than a century ago, whereas the dielectric thermal analysis technique was developed less than a decade ago. However, both the basic and the less common thermal analysis techniques have the same objective—to determine the change in a physical property of a material as a function of temperature. The power of these techniques lies in the diversity of physical properties of a material which can be investigated, the wide range of materials which can be studied, the wide temperature range over which variation in such physical properties can be determined, and the wide range of ''conditions'' in which the thermal properties and thermal behavior of a material can be evaluated. This overview provides an insight into the diversity, complexity, authenticity, and superiority of this category of analytical techniques, which collectively have revolutionized the studies of materials and chemical processing over several decades and promise unique analytical opportunities in support of the nanotechnological age.

REFERENCES

1. Hemmingner, W. Recommendations of the ICTAC Nomenclature Committee. ICTAC News 1998, December, 106–122.
2. Richardson, M.J.; Charsley, E.L. *Handbook of Thermal Analysis and Calorimetry*; Elsevier: Amsterdam, 1998; Chapter 13.
3. Galwey, D. Is the science of thermal analysis based on solid foundations? A literature appraisal. Thermochim. Acta **2004**, *413*, 139–183.
4. Reading, M. The kinetics of heterogeneous solid state decomposition reactions: a new way forward. Thermochim. Acta **1988**, *135*, 37–57.
5. Haines, P.J., Ed. *Principles of Thermal Analysis and Calorimetry*; Royal Society of Chemistry: Cambridge, U.K., 2002; Chapter 3, 61 pp.
6. Brown, M.E. *Introduction to Thermal Analysis—Techniques and Applications*; Kluwer: Amsterdam, 2001; Chapter 4, 84 pp.
7. Brown, M.E., Ed. *Handbook of Thermal Analysis and Calorimetry*; Elsevier: Amsterdam, 1988; Vol. 1, Chapter 6.
8. Earnest, C.M. Assignment of glass transition temperatures using thermomechanical analysis. In *Assignment of the Glass Transitions*; Seyler, R.J., Ed.; American Society for Testing and Materials: Philadelphia, U.S.A., 1994; ASTM STP 1249, 75–87.
9. Balek, V.; Tolgyseey, J. *Emanation Thermal Analysis in Comprehensive Analytical Chemistry*; Elsevier: Amsterdam, 1984; Vol. XXIC.
10. Balek, V. Emanation thermal analysis and its application potential. Thermochim. Acta **1991**, *192*, 1–17.
11. Clark, G.M. Instrumentation for thermosonimetry. Thermochim. Acta **1978**, *27*, 19–25.
12. Butteridge, D.; Joslin, M.T.; Lilley, T. Acoustic emissions from chemical reactions. Anal. Chem. **1981**, *53*, 1064–1073.
13. Bellchamber, R.M.; Betteridge, D.; Collins, M.P.; Lilley, T.; Wade, A.P. Qualitative study of acoustic emission from a model chemical process. Anal. Chem. **1986**, *58*, 1873–1877.
14. Shimada, S. A study of the thermal decomposition of sodium and potassium perchlorates by acoustic emission thermal analysis. Thermochim. Acta **1990**, *163*, 313–316.
15. Price, D.M. Thermomechanical, dynamic mechanical and dielectric methods. In *Principles of Thermal Analysis and Calorimetry*; Haines, P.J., Ed.; Royal Society of Chemistry: Cambridge, U.K., 2002; Chapter 4, 105 pp.
16. Price, D.M. Thermomechanical, dynamic mechanical and dielectric methods. In *Principles of Thermal Analysis and Calorimetry*; Haines, P.J., Ed.; Royal Society of Chemistry: Cambridge, U.K., 2002; Chapter 4, 115–116, 123–126.

Thermal Cracking of Hydrocarbons

Tom Chunghu Tsai
The Dow Chemical Company, Houston, Texas, U.S.A.

Lyle F. Albright
School of Chemical Engineering, Purdue University, West Lafayette, Indiana, U.S.A.

INTRODUCTION

Thermal cracking of ethane, propane, butane, naphthas, gas oils, and/or vacuum gas oils is the main process employed for the production of ethylene and propylene; butadiene and benzene, toluene, and xylenes (BTX) are also produced. Thermal cracking of these hydrocarbons is also called pyrolysis of hydrocarbons. Ethylene is the organic chemical produced worldwide in the largest amounts and has been called keystone to the petrochemical industry.[1] This technology is well documented in the literature.[2–8] Somewhat similar thermal cracking processes are used to produce vinyl chloride monomer (VCM) from ethylene dichloride (EDC), styrene from ethylbenzene, and allyl chloride from propylene dichloride (PDC). Production of charcoal and coke from wood and coal is actually a pyrolysis process, but it is not discussed here.

The worldwide ethylene capacity in year 2000 was approximately 100 million metric tons per year and even greater production is being considered. Propylene capacity is approximately 60 million metric tons per year for which about half is produced by thermal cracking and the other half by refinery operations. The capacities of BTX are approximately 40, 20, and 30 million tons per year, respectively. Considerable BTX production is obtained by thermal cracking of naphthas and gas oils. The largest single train ethylene plant currently produces approximately 1.2 million tons per year (or 2.6 billion pounds/year). That was technically impossible or uneconomical 10 years ago. Recent improvements have been made as follows: a) thermal cracking or pyrolysis technology in the furnace plus coil metallurgy; b) separation and recovery techniques including the compression and refrigeration systems, and distillation towers often filled with packing material; and c) process control and optimization requiring major computer capabilities. The choices of feedstocks for the production of ethylene and other olefins are of major importance. Modern ethylene units are often designed so that different hydrocarbon feedstocks can be used. A modern plant often costs between US $500 million and 2 billion depending on the production capacity, the feedstocks, and the products obtained. Countries with considerable natural gas ethane and propane prefer using them as feedstocks; such areas include the United States, Canada, and the Middle Eastern countries. Europe and Japan, however, have and prefer liquid feedstocks (naphthas and gas oils). Ethylene operators typically employ sophisticated computer models and optimization techniques in choosing feedstocks.

Coil reactors employed in current thermal cracking furnaces are surrounded by hot combustion gases. The coils are often several vertical sections of high-alloy steel tubes connected by U-bends. In other cases, the coil is merely a single vertical tube. The gaseous feed to each coil is a mixture of the hydrocarbon plus steam. A chemical additive to minimize coking is also often incorporated in the feed stream. The gas mixture entering the coil is often at temperatures of 650–700°C. As the reactant gases pass through the coil, they are heated to approximately 850–950°C. As a result, thermal cracking occurs rapidly, often in about 0.06–0.6 sec depending on the design and the operation of the furnace. Since thermal cracking is highly endothermic and the temperature of the gas mixture is increased to a considerable extent, much heat is transferred–first, mainly by radiation from hot combustion gases to the outer surfaces of the coils; second, by conduction through the walls of the coil and any coke deposited on the inner coil surface; and third, by both convection and radiation to the gas mixture in the coil. Temperatures of the coil walls are frequently in the 900–1100°C range. Pressures of the gases in the coil frequently vary from 68 to 400 kPa, all above atmospheric pressure. The weight ratio of steam to hydrocarbon is often 0.25–0.30 with ethane and propane feeds, and 0.5–1.0 with naphthas and gas oils. The material of construction for the coil, usually a high-alloy steel, must be carefully selected in order to obtain three to five years' life expectancy. Flow velocities in the coil are always high, often at about 0.5–0.7 Mach number. Fuel requirements for the furnaces are relatively high, and methods to minimize them are considered later. Due mainly to the high temperatures, coke (essentially solid

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007934

carbon) is always formed; it is a highly undesirable by-product. The coke formation and the collection on the coil surfaces increase the heat-transfer resistances and the furnace has to be shutdown at intervals for decoking. Both complete shutdown of the furnace, or the so-called on-line decoking, reduce the ethylene production, and increase production cost. Coke formation increases carburization of the coil material, shortening the life of the coil. In addition to the production of the desired products reported above, hydrogen and methane are also produced. When naphthas and gas oils are used as feeds, fuel oil is also produced in significant amounts. These by-products can be used to supplement the main fuel for the furnaces (generally natural gas). Large amounts of steam and electricity are also required in the overall thermal cracking process, which includes in addition to furnaces, large compressors for the refrigeration system and to pressurize the product streams and distillation towers. Fig. 1 indicates the total energy requirements for ethylene production with different feedstocks and different ethylene yields. More propylene and more BTX are produced with liquid feedstocks. As indicated, energy demands per unit weight of ethylene decrease as the ethylene production increases. In this entry, we discuss the fundamentals of thermal cracking, the mechanistic modeling, and modern ethylene plant design and operation. New trends of development in thermal cracking and ethylene technology are discussed.

FUNDAMENTALS AND THERMAL CRACKING CHEMISTRY

Basic Chemistry and Mechanistic Modeling

Thermal cracking reactions can be divided into initiation steps that generate free radicals, chain propagation steps that form the desired products, and termination of free radicals. This classification of reaction steps permits the mechanistic modeling of the complicated reactions. The main chemical reactions are illustrated in Table 1.[9] Almost all reactions during thermal cracking occur in the gas phase and involve free radicals. Reactions 1 and 2 are key initiation reactions forming an increase of the number of gaseous free radicals. Reactions 3–16 are rapid propagation steps with no increase in the number of gaseous free radicals. They result in the production of ethylene, propylene, hydrogen, methane, plus other by-products (not shown in Table 1). Reactions 17 and 18 are termination steps with a decrease of gaseous radicals.

The main reactions for liquid feedstocks (naphthas, gas oils, etc.) are sometimes summarized according to the types and amounts of paraffins (n-paraffins and iso-paraffins), olefins, naphthenes (monocyclic naphthenes and multicyclic naphthenes), and aromatics (naphthenic aromatics and pure aromatics).[6] The basic chemistry and modeling efforts have been reported by numerous investigators including Dente et al.,[10–14]

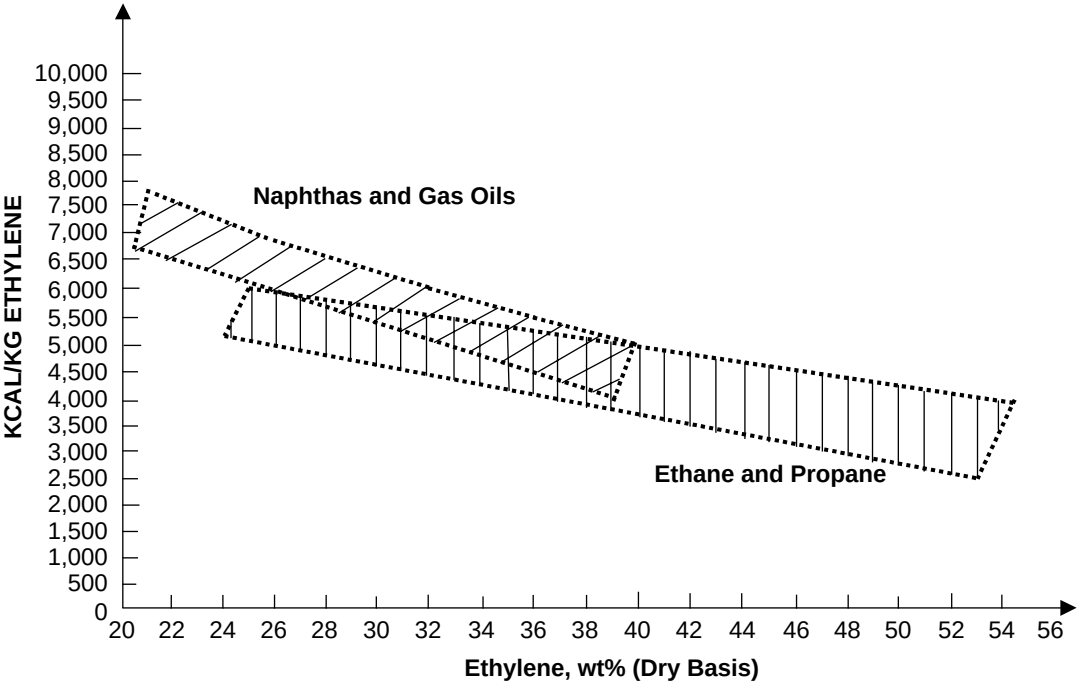


Fig. 1 Energy requirement of modern ethylene plant (total energy for pyrolysis furnaces, compressor units, and separation units).

Table 1 Mechanistic reaction model for ethane–propane cocracking

Reaction no.	Reactions	Reaction rate, r_i[a]	Kinetic parameters	
			A_i	E_i
1	$C_2H_6 \rightarrow 2\ CH_3^{\bullet}$	$r_1 = k_1\ [C_2H_6]$	5.185×10^{16}	90.75
2	$C_3H_8 \rightarrow CH_3^{\bullet} + C_2H_5^{\bullet}$	$r_2 = k_2\ [C_3H_8]$	2.074×10^{16}	87.63
3	$C_2H_6 + CH_3^{\bullet} \rightarrow CH_4 + C_2H_5^{\bullet}$	$r_3 = k_3\ [C_2H_6]\ [CH_3^{\bullet}]$	3.941×10^{11}	8.371
4	$C_2H_6 + H^{\bullet} \rightarrow H_2 + C_2H_5^{\bullet}$	$r_4 = k_4\ [C_2H_6]\ [H^{\bullet}]$	7.537×10^{10}	19.64
5	$C_2H_5^{\bullet} \rightarrow C_2H_4 + H^{\bullet}$	$r_5 = k_5\ [C_2H_5^{\bullet}]$	1.013×10^{14}	51.85
6	$C_3H_8 + H^{\bullet} \rightarrow H_2 + 1\text{-}C_3H_7^{\bullet}$	$r_6 = k_6\ [C_3H_8]\ [H^{\bullet}]$	5.096×10^{10}	7.545
7	$C_3H_8 + H^{\bullet} \rightarrow H_2 + 2\text{-}C_3H_7^{\bullet}$	$r_7 = k_7\ [C_3H_8]\ [H^{\bullet}]$	2.401×10^{10}	10.76
8	$C_3H_8 + CH_3^{\bullet} \rightarrow CH_4 + 1\text{-}C_3H_7^{\bullet}$	$r_8 = k_8\ [C_3H_8]\ [CH_3^{\bullet}]$	2.813×10^{10}	6.705
9	$C_3H_8 + CH_3^{\bullet} \rightarrow CH_4 + 2\text{-}C_3H_7^{\bullet}$	$r_9 = k_9\ [C_3H_8]\ [CH_3^{\bullet}]$	2.119×10^{9}	10.47
10	$C_3H_8 + C_2H_5^{\bullet} \rightarrow C_2H_6 + 1\text{-}C_3H_7^{\bullet}$	$r_{10} = k_{10}\ [C_3H_8]\ [C_2H_5^{\bullet}]$	3.440×10^{8}	13.07
11	$C_3H_8 + C_2H_5^{\bullet} \rightarrow C_2H_6 + 2\text{-}C_3H_7^{\bullet}$	$r_{11} = k_{11}\ [C_3H_8]\ [C_2H_5^{\bullet}]$	3.973×10^{8}	13.49
12	$1\text{-}C_3H_7^{\bullet} \rightarrow C_2H_4 + CH_3^{\bullet}$	$r_{12} = k_{12}\ [1\text{-}C_3H_7^{\bullet}]$	2.119×10^{13}	33.81
13	$2\text{-}C_3H_7^{\bullet} \rightarrow C_3H_6 + H^{\bullet}$	$r_{13} = k_{13}\ [2\text{-}C_3H_7^{\bullet}]$	3.195×10^{13}	50.17
14	$C_3H_6 + H^{\bullet} \rightarrow 1\text{-}C_3H_7^{\bullet}$	$r_{14} = k_{14}\ [C_3H_6]\ [H^{\bullet}]$	9.705×10^{9}	4.698
15	$C_3H_6 + H^{\bullet} \rightarrow 2\text{-}C_3H_7^{\bullet}$	$r_{15} = k_{15}\ [C_3H_6]\ [H^{\bullet}]$	1.151×10^{11}f	1.250
16	$C_2H_4 + H^{\bullet} \rightarrow C_2H_5^{\bullet}$	$r_{16} = k_{16}\ [C_2H_4]\ [H^{\bullet}]$	4.559×10^{9}	1.945
17	$CH_3^{\bullet} + CH_3^{\bullet} \rightarrow C_2H_6$	$r_{17} = k_{17}\ [CH_3^{\bullet}]\ [CH_3^{\bullet}]$	1.349×10^{10}	0
18	$C_2H_5^{\bullet} + H^{\bullet} \rightarrow C_2H_6$	$r_{18} = k_{18}\ [C_2H_5^{\bullet}]\ [H^{\bullet}]$	5.185×10^{10}	0

[a] k_i is the reaction velocity constant of reaction i expressed by the Arrhenius equation: $k_i = (A_i)^*(\exp[-E_i/(RT)])$ where A_i is the frequency factor of the *i*th reaction in the reaction model (s^{-1} or $1 \times s^{-1} \times mol^{-1}$); E_i is the activation energy of the *i*th reaction in the reaction model (kcal $\times$ mol^{-1}); *R* is the universal gas constant; and *T* is the reaction temperature, °K.

Allara and Edelson,[15,16] and Albright and McConnell.[17,18]

Tsai and Zhou[9] have presented a somewhat simplified mechanistic model for ethane and propane cocracking, as shown in Table 1. Their investigation indicated the positive effects relative to ethylene and propylene production for this cocracking, savings in the separation of ethane and propane from LPG mixtures, and feedstock reduction for constant olefins production due to higher olefin selectivity. Their model, however, does not include any surface reactions that may be important plus reactions causing coke formation.

Shorter residence times in the coils, such as 0.06–0.12 sec, employed in millisecond units, tend to give higher yields of ethylene. To obtain the shorter residence times and also high conversions, higher temperatures are needed. Table 2 indicates typical yields for a millisecond unit and a unit operated at 0.4–0.6 sec residence times.

Steam is used as the diluent to reduce hydrocarbon partial pressure and hence to improve selectivity toward olefins. Steam also gasifies (via oxidation) part of the coke deposited on the tubular reactor wall producing CO and CO_2 and hence promotes coke removal from the inner surfaces of the coils.

Temperature, Pressure, Residence Time, Steam Diluent, and Conversion Energy

During the evolution of the thermal cracking process, various parameters, such as the average residence time and average hydrocarbon partial pressure, were found to be important. Computing systems employing complicated mathematical techniques are used to solve stiff differential equations. As a result, operations can be optimized relative to product distribution, selectivity, yield, and economic profits. Design and operation of the thermal cracking units are of critical importance. Many thermal cracking units employ two furnaces (or reaction zones) with one convection zone (see Fig. 2). Thermal cracking occurs in the coils in the radiant zone. In the convection zones, heat is transferred as follows: first, to preheat and vaporize the hydrocarbons to be cracked; second, to preheat the steam to be mixed with the hydrocarbon feeds; third, to preheat the entering fuel (often natural gas) and air (used to burn the fuel); fourth, to preheat the boiler feed water; fifth, to produce steam from boiler feed water; and sixth, to superheat the steam. Preheating of the fuel and air results in higher flame temperatures, which increases both the rate and the amount of heat transfer to the reacting gases in the coils.

Table 2 Typical yields of millisecond versus conventional short residence time heaters

Furnace type	Conventional[1]		Millisecond[6]	Conventional[1]	Millisecond[6]
Feedstocks	Ethane		Ethane	Naphthas	Naphthas
Residence time	0.4–0.6		0.06–0.12	0.4–0.6	0.06–0.12 (sec)
Conversion (%)	50	60	59.12		
Pyrolysis yields (wt %)					
H_2	3.06	3.55	3.80	0.9	1.1
CH_4	2.60	4.20	2.43	15.8	14.9
CO[a]	—	—	0.07	—	—
CO_2[a]	—	—	—	—	—
C_2H_2	0.12	0.25	0.52	0.4	0.9
C_2H_4	41.65	48.20	48.62	28.6	32.2
C_2H_6	50.00	40.00	40.88	3.9	3.0
C_3H_4	0.10	0.02	0.02	0.7	1.3
C_3H_6	0.89	1.11	0.92	15.0	14.3
C_3H_8	0.14	0.17	0.19	0.4	0.3
C_4H_6	0.50	1.07	1.18	4.4	5.6
C_4H_8	0.25	0.21	0.14	4.2	3.8
C_4H_{10}	0.35	0.27	0.22	0.5	0.4
C_5s	0.20	0.27	0.38		
C_6H_6	0.20	0.48	0.35		
C_7H_8	0.03	0.06	0.09		
C_8H_8	—	—	0.03		
C_8H_{10}					
C_6-C_8 PONA	—	0.14	0.04		
C_5-200°C				21.7	18.9
C_9-200°C	—	—	0.06		
Fuel oil	—	—	0.06	3.5	3.3
Total	100.0	100.0	100.0	100.0	100.0

[a]CO and CO_2 are reported based on carbon balance.

Furnace Coil Material of Construction and Selection

Tensile strength, creep strength, creep ductility, and hydrogen embrittlement are all important considerations in the selection of the high-alloy steel for the thermal cracking coils. Three requirements of the coil material in the radiant section are as follows: 1) excellent carburization resistance at high temperature; 2) high tensile strength combined with good ductility; and 3) low rates of creep. The following three alloys are widely used: 1) high silicon HK-40; 2) modified HP-40; and 3) modified 28/48 Cr-Ni alloy. These steels are employed at metal temperatures up to approximately 1065°C, 1120°C, and 1200°C, respectively.

As high-alloy steels are used, the surface compositions change with use; such changes occur in a matter of hours or even minutes. First, major increases occur in the chromium and manganese content. At or near the surface, the chromium content often increases from about 20% to 50–70%. The manganese content increases from <1% to 15–20%. Meanwhile the iron and nickel concentrations decrease from 40–50% and 25–35% to less than 5%. Significant increases of the surface compositions of silicon, aluminum, and titanium also occur sometimes. Second, the surface layer, often in 0–3 μm range, becomes largely metal oxides (and sometimes metal sulfides). Third, part of the surface layer frequently spalls off exposing sublayers with high iron and nickel concentrations. Conventional decoking significantly contributes to the spalling process that corrodes the inner surfaces of the coils, reduces coil life, and promotes coke formation (to be discussed later). Surfaces containing high concentrations of chromium and low concentrations of iron and nickel are desired since they minimize coke collection and/or formation on the inner surfaces of the coil.

Surface Pretreatment or Treatment

Understanding of the migration rates of metal elements in the coil walls has been utilized to create surface compositions that minimize coke formation.[19–21] Stainless steel coils for thermal cracking have been exposed to gases with low oxidative potentials at temperatures up to 1200°C for about 50 hrs.[22,23]

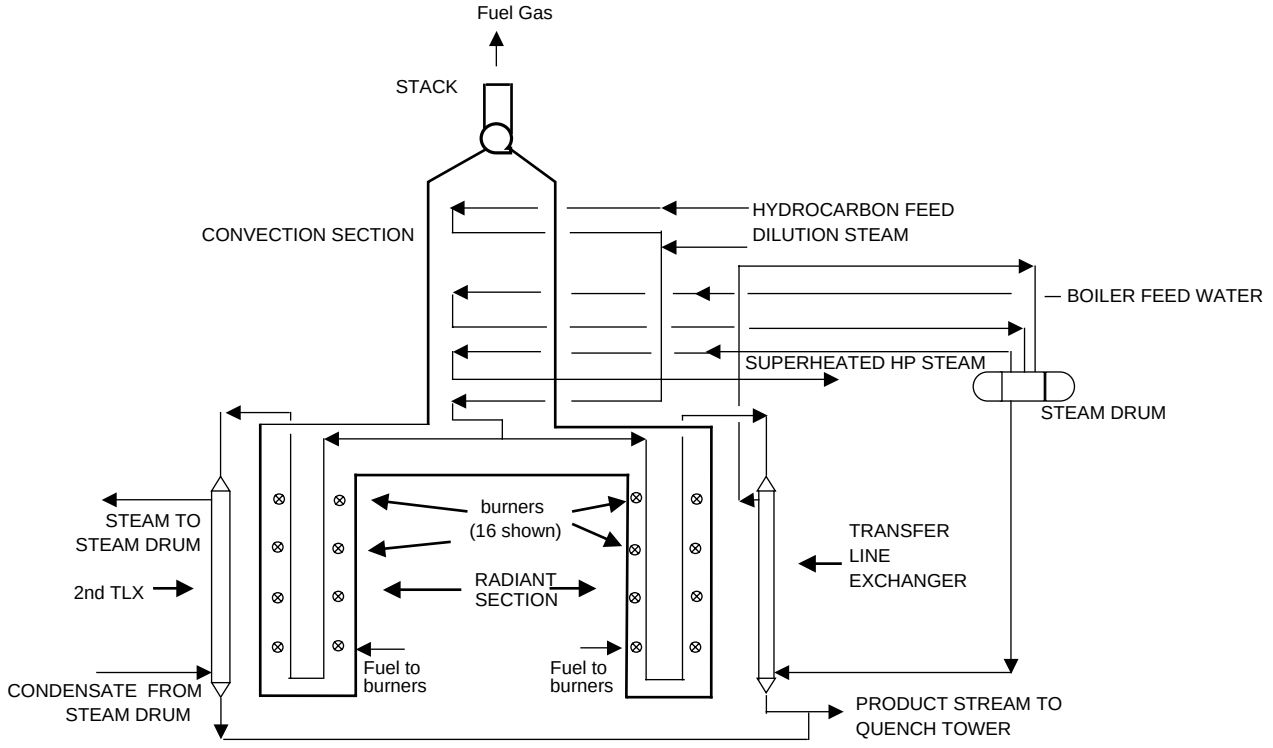


Fig. 2 Typical cracking furnaces.

The resulting coils in one plant test increased the run lengths before decoking from 31 days to 516 days.[24] Sulfiding the reactor surface by adding to the feed-stream sulfur-containing compounds, such as CS_2, dimethyldisulfide, or mercaptans, has been beneficial in reducing coke formation.[25] Some new coking inhibitors also contain rare earth elements such as Ce or La.[26]

Coking and Decoking

In the steam cracking of hydrocarbons, a small portion of the hydrocarbon feed gases decomposes to produce coke that accumulates on the interior walls of the coils in the radiant zone and on the inner surfaces of the transferline exchanger (TLX). Albright et al.[27–31] identified three mechanisms for coke formation. Mechanism 1 involves metal-catalyzed reactions in which metal carbides are intermediate compounds and for which iron and nickel are catalysts. The resulting filamentous coke often contains iron or nickel positioned primarily at the tips of the filaments. This filamenteous coke acts as excellent collection sites for coke formed by mechanisms 2 and 3. Mechanism 2 results in the formation of tar droplets in the gas phase, often from aromatics. These aromatics are often produced by trimerization and other reactions involving acetylene. Some, but not all, of these droplets collect on the inner walls where they decompose forming spherical or semispherical coke deposits. Mechanism 3 first involves the reactions of gaseous microspecies with the free radicals on the coke surfaces. These microspecies include acetylene, ethylene or other olefins, butadienes, and free radicals such as methyl, ethyl, or benzyl radicals; second, C—H bonds located on the surface break to produce coke. Both amorphous and graphite cokes form and collect on inner surfaces of the coils and TLXs. More graphitic coke appears to be formed near the interface next to the coil wall. Amorphous coke, however, is often predominant at the interface layer near the gaseous process hydrocarbon stream.

Zou[25] have published information pertaining to the mechanism and the kinetic models of coke deposition during both thermal cracking of propane and atmospheric gas oils (AGOs). Their findings, plus those of several other investigators, neglect the role of surface reactions. There is the need to better explain how and why the surface compositions and roughness of the coils have, on occasion, such a major effect on rates of coke that forms and/or collects on the inner walls of coils.

Analysis of coke samples from an industrial furnace cracking ethane and propane revealed some interesting phenomena. Coke samples from both the coils and the TLXs generally, if not always, contain appreciable amounts of metal and various elements including

sulfur, calcium, and silicon. Scanning electron microscopy plus EDAX, polarized light microscopy, and transmission electron microscopy have been employed to investigate coke samples, surfaces of coils; determine the metal contents; view the metal deposits; and determine the character of coke deposits, variations of coke deposits, and the porosity of the coke.

Plant data indicate that first the inner diameters of the coils increase and the wall thicknesses decrease as the coils are used. Spalling of the inner surfaces is apparently a major factor. The wall thickness sometimes decreased by as much as 30%. Decreases of the wall thickness may however vary from 5% to 30% around the circumference of the coil. Second, the roughness of surface varies greatly with location. Part of these variations likely depends on local overheating (and burner locations) relative to the coil.

TYPICAL COMMERCIAL UNITS

Contractors that specialize in the design and construction of cracking furnaces for ethylene plants include Technip, KBR, ABB Lummus, Stone and Webster, and Linde A.G. The height of the furnaces may vary from approximately 30 to 50 m. Currently, almost all tubes (or coils) are arranged vertically inside a combustion chamber. The capacity of a single furnace has reached well over 120,000 metric tons/furnace/year. Fig. 2 illustrates a typical industrial cracking furnace. A typical modern cracking furnace employs two radiant sections, one convection section, and two transfer line exchangers (TLXs).

Radiant Section and Combustion Chamber

Thermal cracking reactions occur in the tubular reactors positioned inside the combustion chambers. This is generally referred to as the radiant section. Coils of modern thermal cracking furnace are all arranged in a vertical fashion with double side radiant heating. Some furnaces also have floor burners. Thermal cracking coils are generally classified into the following four types (see Fig. 3):[32–34]

1. Coil A illustrates the conventional single diameter coil. This type contains 6–8 vertical tubes joined by 5–7 U-bends. The tube diameters range from approximately 100 to 125 mm and the total lengths may vary from approximately 65 to 75 m.
2. Coil B contains vertical tubes of different diameters. This type of coil can vary from 2 to 4 passes. The tube diameters range from approximately 38 to 83 mm and the total coil lengths range from approximately 20 to 45 m. This type of coil starts with smaller diameter tubes connected to larger diameter tubes.
3. Coil C represents the varying diameter split coils. This type of coil can vary from 4 to 6 passes. The number of tubes per pass can be 4-2-1-1, 8-4-1-1 for 4-pass coils and 4-2-1-1-1-1 for 6-pass coils. Diameters of the first pass are normally the smallest, being approximately 50–65 mm. The last pass tube is normally larger in diameter, being approximately 145–165 mm. The total coil lengths may range from approximately 44–62 m. These split coils begin with smaller diameter tubes that have higher surface to volume ratio to enhance the heat-transfer rates. This way, it can bring the process fluids to the cracking temperature more rapidly.
4. Coil D is employed in ''millisecond'' furnaces. They are straight tubes with diameters varying from approximately 25–38 mm and with lengths ranging from approximately 10–12 m.

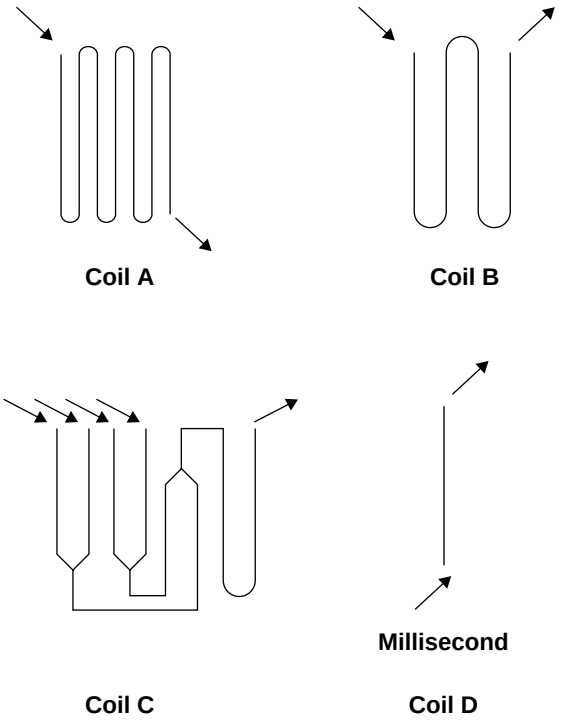


Fig. 3 Typical thermal cracking coils.

More details of thermal cracking coils and furnace design are given in Ref.[6].

Hydrocarbon and steam mixture flows from the convection section into the radiant section at approximately 525–700°C, depending on the feedstocks and the heater design. This is called the crossover temperature. For ethane and propane thermal cracking, the crossover temperature is approximately 700°C. For vacuum gas oil (VGO) and AGO cracking,

the crossover temperature is perhaps 525–550°C. The coil outlet temperature is generally in the range of 750–950°C, depending on the feedstocks cracked and the severity desired. Typical thermal cracking furnace characteristics are summarized in Table 3.

Shorter residence time and lower hydrocarbon partial pressure improve the selectivity toward ethylene at a constant conversion. Conversion is defined as follows for single components in either wt% or mol%:

$$\text{Conversion} = (C_{in} - C_{out})/C_{in}$$

for mixtures, the conversion can be defined as the weighted average of the individual component conversion if the componential analysis is known, i.e., $X = \sum W_i X_i$, where X_i is the conversion for the ith feed component, and W_i is the weighing factor (usually weight fraction or mole fraction). The term severity is used to describe the conversion or the extent or depth of hydrocarbon cracking. For liquid feeds, various methods such as n-pentane conversion, kinetic severity factor,[2] or molecular collision parameter[35–37] are employed to measure cracking severity. Tsai[37] has reported a quantitative yield correlation as a function of MCP (molecular collision parameter) for AGO cracking. For other complex multicomponent feedstocks of unknown composition, conversion or cracking severity has been defined as a function of molecular weights of the thermal cracking feedstock and the product effluent as follows:[7]

$$\alpha = [(m_f - m_e) - 1]/(\gamma - 1)$$

where m_f is the molecular weight of feed, m_e is the molecular weight of dry (steam-free) effluent, and γ is the molal expansion factor, i.e., moles of effluent produced per mole of feed converted. The empirical expression, $\gamma = (m_f/24.5)$, has been used. Another expression of conversion for feedstock of unknown composition uses the hydrogen content instead:

$$\alpha = (Y - 1)/(CY - 1)$$

where $Y = (H - 6)/(HF - 6)$; H is the hydrogen content in C_5 plus product, wt%; HF is the hydrogen content of the feed, wt%; and C is a constant for any given feedstock.

Other parameters, such as coil outlet temperature, propylene to methane ratio, propylene to ethylene ratio, or cracking severity index, have been used to indicate cracking severity. These parameters are, to some degree, dependent on feed properties and operating conditions.

For some liquid feedstocks such as naphthas, the componential composition is often obtained by gas chromatography (GC) and/or mass spectrometry (MS). For gas oils or heavier feedstocks, it is impossible to obtain the desired analysis. Paraffins, olefins, naphthenes, aromatics (PONA) grouping is sometimes used as a means of feed characterization. For gas oils, Bureau of Mines Correlation Index (BMCI) has been used as a parameter for feed characterization. Since the 1980s, nuclear magnetic resonance (NMR) spectroscopy has been used to characterize heavy feedstocks.

The radiant coils are always cleaned by burning the coke with mixtures of steam and air in different proportions and at different temperatures. Some radiant coil decoking is for a fixed period of time such as 12 or 24 hrs. Some decoking operations are monitored to determine CO or CO_2 in the effluent. Since hydrogen in the coke burns off very rapidly, the initial concentration of oxygen is often kept low to avoid overheating the coils. Usually, all coils in a given heater are decoked, and the effluent of decoking is sent to a decoking pot or to a firebox for burning. One patent[38] claims single-coil decoking, while other coils in the same furnace are cracking hydrocarbons; that is a true on-line decoking. However, the decoking effluent streams contain CO, CO_2, and carbon particles that affect the downstream operations. On-line decoking is probably limited in practice.

Table 3 Characteristics of modern thermal cracking furnaces

Ethylene production capacity, tons/furnace/year	**20,000–120,000**
Numbers of coils	2–180
Tube diameter, mm	25–200
Coil length, m	10–80
Coil (Gas) outlet temperature, °C	750–950
Tube wall temperature, °C (clean)	900–1100
Maximum tube wall temperature, °C	1040–1150
Average heat flux, kw/m_2 (of outside surface area)	50–120
Coil outlet pressure, kPa	168–300
Residence time, sec	0.06–0.60

CONVECTION SECTION

A conventional typical convection section often has tube bundles for fuel gas or air preheat, feed preheat, boiler feed water preheat, steam generation, and steam superheating. In the convection section, the heat transfer is mainly gas-to-gas heat transfer and the overall heat-transfer coefficients are relatively low. Finned tubes are generally used to improve heat-transfer rates. Material for the convection section tubes varies from carbon steel to a high temperature alloy. Sometimes, high-alloy tubes are positioned in the lower section

close to the radiant chamber. Condensation of acidic flue gases must be considered in the design of the convection section and the selection of tube material. Corrosion sometimes occurs at the tube inlet causing tube leaks. Fouling in both the inside and the outside of the tubes may also occur. The inside fouling is normally cleaned by burning with steam and air mixture. The outside of the tubes is normally cleaned by steam. On top of the convection section, stacks and induced draft (ID) fans help create a draft for flue gas that enhances heat transfer for the tube bundles.

Transfer Line Exchangers

The thermal cracking products from the radiant coils pass through TLXs (or TLEs) for steam production and heat recovery. From there, the pyrolysis gas leaves the TLXs at approximately 300–400°C for ethane, propane, or light naphthas cracking, and at approximately 550–650°C for heavy naphthas, gas oils, or VGO cracking. The transfer line exchangers outlet temperature is the lowest at the beginning of the operating cycle when the exchangers are clean. To minimize further thermal cracking, gas temperatures must be reduced rapidly.

In all modern ethylene plants, TLXs recover much of the energy in the furnace effluent product streams. Several vendors offer high pressure shell and tube TLXs. These include Struthers-Wells of the United States; Schmidtsche Heissdamp GmbH and Borsig of Germany; and Mitsubishi, Mitsui, and Babcock-Hitachi of Japan. More discussion on those individual TLXs is given in Ref.[1]. Both Mitsubishi and Mitsui TLXs differ drastically from other designs. Mitsubishi offers a TLX with an integral steam drum and cyclone for vapor–liquid separation. The pyrolysis gas flows in the shell side. Decoking of both the coil and the TLX is claimed to occur in one separation. The Mitsui TLX (or quench cooler) uses three concentric tubes as the tube element and requires steam-air decoking to clean the TLXs.

During decoking of the radiant coils, the TLXs can also be partially decoked. For complete decoking, the furnace is usually cooled down and the TLXs are separated from the coils and hydrojetted with high pressure water. In some cases, the coke in the TLXs can also be burnt off, and hence no mechanical cleaning is required.

The temperatures at the end of a run of the TLX are generally in the range of 600–650°C, depending on the specific design, material of construction, and the type of TLX used. For cracking of ethane and/or propane, the inner film fouling resistance and the outer film fouling resistance are often approximately 0.0000245 m^2-Hr-°K/kJ. The thermal conductivity of the TLX coke is approximately 20% of that of the coke in the coils. Different types of coke are apparently formed; perhaps the cokes are dehydrogenated to various degrees. Therefore, it is very important to select the proper TLX tube material of construction allowing higher TLX outlet temperature, which would increase the capacity and the furnace and TLX run lengths as well.

Product Separation, Recovery, and Purification

Fig. 4 represents a simplified schematic flow diagram for an ethane and/or propane cracking ethylene plant. The separation sequences are shown as follows:

1. *Quench tower.* This is accomplished by quenching the product effluent stream with water. Condensed hydrocarbons and water from the quench tower are separated in a coalescer. Hydrocarbons from this coalescer and the drips from the first stage compressor discharge drums are sent to the drip stripper to separate lighter hydrocarbons from the pyrolysis gasoline and fuel oil. Water from the coalescer is transferred to the process water stripper to separate the hydrocarbons before sending the water to the biopond for wastewater treatment. Quench tower design is a challenge. A considerable amount of water is recirculated, and coke particles in the water make this stream quite erosive requiring special pumps.
2. *Compression and condensation.* The pyrolysis gas leaving the quench tower is compressed to approximately 35 atmospheres in a four or five stage centrifugal compressor (normally referred to as cracked gas compressor). The number of stages is determined by the maximum temperature for the material of construction of the cracked gas compressor and the fouling tendency of the pyrolysis gas. The compressor consists of two or three compressor casings driven by a single or double extraction/condensing turbines depending on plant size. Between each stage of the compressor are condensers and drums that separate the condensed hydrocarbons and water.
3. *Acid gas removal.* Acid gases (CO_2 and H_2S) are removed by absorption after the third or the fourth compression stage. This is the optimum location since the gas volume is significantly reduced. Scrubbing with caustic solution, monoethanolamine (MEA), or diethanolamine (DEA) is generally used. More details can be found in the literature.[1,7,8]
4. *Water removal.* Complete removal of water vapor from the pyrolysis gas is generally

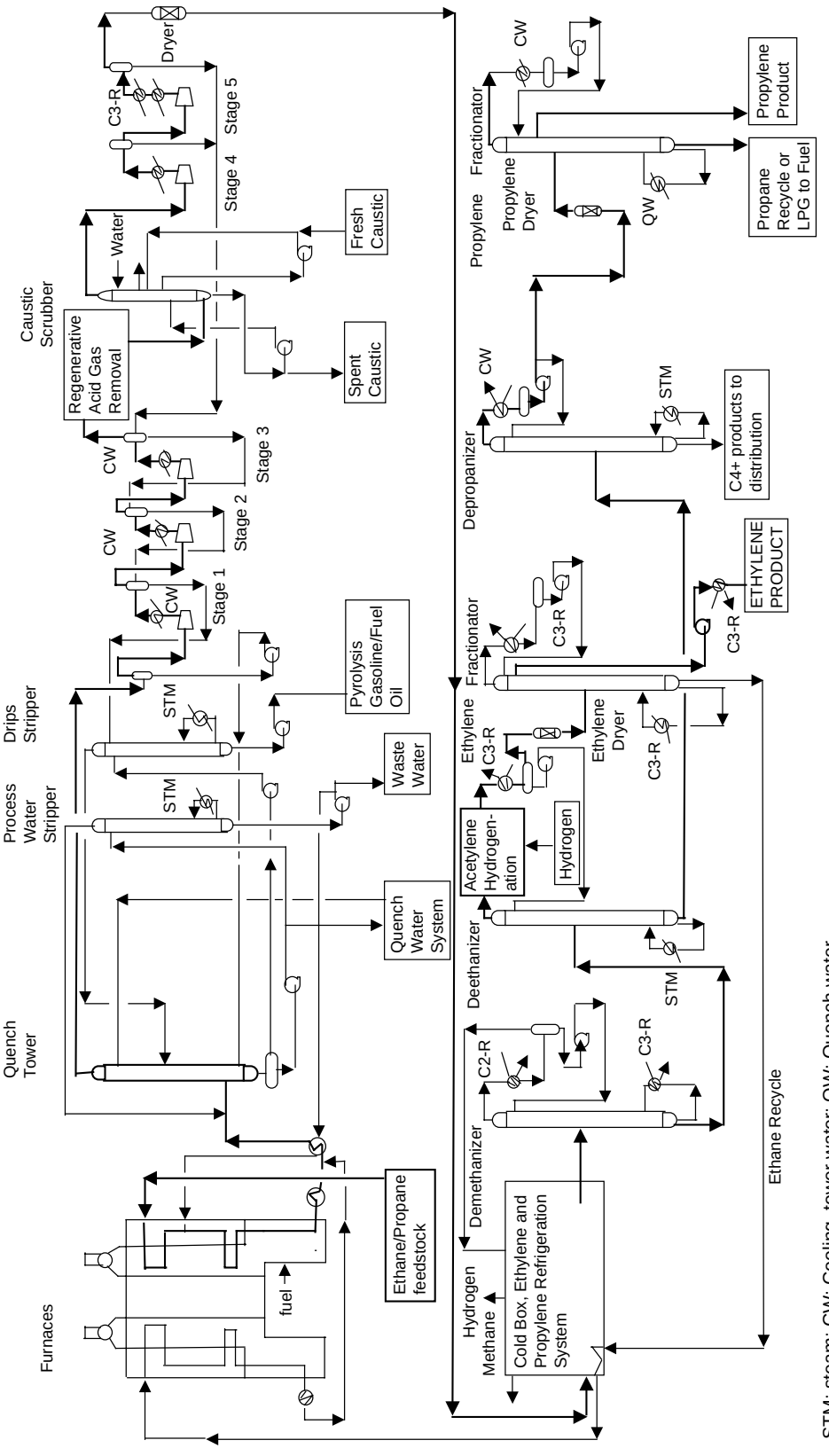


Fig. 4 Simplified schematic flow diagram for a typical ethane/propane cracking ethylene plant.

STM: steam; CW: Cooling tower water; QW: Quench water
C2-R: Ethylene refrigerant; C3-R: Propylene refrigerant

achieved using dryers packed with molecular sieves. One dryer is operated while the other is in the regeneration mode.

5. *Cold box and refrigeration system.* After the acid gas and water removal, the pyrolysis gas is cooled and condensed to approximately –165°C; only hydrogen and some methane remain in the vapor phase. The feed locations are determined via process simulation. Hydrogen and methane are drawn from the lowest temperature stage separator and sent to thermal cracking furnaces as fuel.
6. The demethanizer is designed for complete separation of methane from ethylene and heavier components. The demethanizer is normally operated at approximately 7 atm. The demethanizer overhead consists of methane, plus relatively small amounts of hydrogen, carbon monoxide, and traces of ethylene. Brazed aluminum plate-fin exchangers are used for the multipass cryogenic heat-transfer services. Some ethylene plants employ high pressure demethanizers operated at approximately 35 atm. They are generally combined with either a front-end deethanizer or depropanizer.
7. *Deethanizer and ethylene fractionator (ethylene/ethane splitter).* The C_2 and heavier hydrocarbons from the bottom of the demethanizer are sent to the deethanizer operated at approximately 25 atm. It is either a trayed tower or a packed column. Deethanizer overhead consists of C_2 hydrocarbons and the bottom products are C_3 and heaviers.
8. *Acetylene hydrogenation.* In modern ethylene plants, acetylene is generally hydrogenated to ethylene and ethane in a palladium catalyst bed. The reaction is quite exothermic, and intermediate cooling is required. The effluent normally contains less than 1 ppmv of acetylene. This is generally referred to as back-end acetylene hydrogenation, which exerts higher selectivity and more precise temperature control. Front-end acetylene hydrogenation is also practiced in the stream of an intermediate compressor. Front-end acetylene hydrogenation uses a different catalyst, and a deethanizer or a depropanizer column is ahead of the demethanizer; this scheme has become of increased importance in the recent past. In most acetylene hydrogenation reactor operation, some carbon monoxide is injected to monitor the catalyst activity and the reaction rate, especially after changeover of the fresh catalyst. The catalyst is deactivated over time because of deactivation; the catalyst is periodically regenerated. After acetylene hydrogenation, the dried gas enters the ethylene fractionator to separate ethylene from ethane. A section is provided near the top of the fractionator for removal of residual hydrogen, carbon monoxide, and methane. A closed loop heat pump is sometimes used, which uses propylene refrigerant as the coolant in the reflux condenser. Ethane is normally recycled back to the thermal cracking furnaces to be cracked.
9. *Depropanizer and propylene fractionator (propylene/propane splitter).* The deethanizer bottom is sent to the depropanizer where the C_3 hydrocarbons are the overhead product and the C_4 and heavier hydrocarbons are removed from the bottom. The overhead of the depropanizer is sent to the propylene fractionator where propylene is separated from propane. A two-tower propylene fractionator produces polymer grade propylene (99.5% plus). For the naphthas or gas oil cracking plants, the depropanizer bottom stream is normally further processed in a debutanizer, which separates the C_4 and lighter in the overhead, and the pyrolysis gasoline and heavier in the bottom. Also, the depropanizer overhead stream generally is sent to a methylacetylene and propadiene (MAPD) hydrogenation reactor, where MA is hydrogenated to propylene and propane.

Process Control and Optimization

Recent improvements of process controls have resulted in improved operations and higher profits. Real-time optimization of thermal cracking furnaces based on the feedstock and product flow rates, pricing, and fuel cost has become a reality. Implementation of pinch technology, process separation network analysis, and optimization has made significant reductions in energy demands.

ALTERNATE ETHYLENE PRODUCTION PROCESSES

The search continues for better and more economical processes for the production of ethylene. Those processes include catalytic thermal cracking, methanol to ethylene, oxidative coupling of methane, advanced cracking technology, adiabatic cracking reactor, fluidized bed cracking, membrane reactor, oxydehydrogenation, ethanol to ethylene, propylene disproportionation, and coal to ethylene. Much work is still needed before any such process can compete with current processes.

One process that potentially could be used employs a catalyst and is operated at approximately 1 ms.

It operates at 950°C and oxidizes ethane at almost 70% conversion with approximately 85% selectivity.[39] Major concerns include safe operation of potentially explosive ethane and air mixtures at elevated temperatures. If the process uses air instead of pure oxygen, then separating nitrogen from ethylene can be costly. Scale-up of such a catalytic reactor and the life of the catalyst are concerns.

CONCLUSIONS

Demand for ethylene in the world continues to grow. Capacities of modern ethylene plants and furnaces continue to increase in a highly competitive market. Current research concentrates on the following areas: a) selective cracking that produces higher yields; b) better metallurgy that allows higher temperature operation, shorter residence times, and higher conversions; c) coking or anticoking technology that minimize or eliminate coke formation and increase run lengths; d) advanced process control and optimization including real-time optimization; e) more efficient separation techniques such as new packing materials in distillation columns; f) new process sequences and networks that reduce energy demands and result in higher quality products; and g) improved compressor and refrigeration systems.

ACKNOWLEDGMENT

The authors wish to thank The Dow Chemical Company for their permission to publish this entry.

REFERENCES

1. Kniel, L.; Winter, O.; Stork, K. *Ethylene–Keystone to the Petrochemical Industry, Chemical Industries–2*; Marcel Dekker, Inc.: New York and Basel, 1980.
2. Zdonik, S.B.; Green, E.J.; Hallee, L.P. Manufacturing ethylene. *The Oil & Gas Journal* **1970**.
3. Albright, L.F.; Crynes, B.L. *Industrial and Laboratory Pyrolysis*; ACS Symposium Series 32; American Chemical Society: Washington, D.C., 1976.
4. Albright, L.F.; Crynes, W.L.; Corcoran, W.H. *Pyrolysis: Theory and Industrial Practice*; Academic Press Inc.: New York, 1983.
5. Albright, L.F.; Crynes, B.L.; Nowak, S. *Novel Production Methods for Ethylene, Light Hydrocarbons, and Aromatics*; Marcel Dekker Inc.: New York, 1992.
6. Zhou, R. *Fundamentals of Pyrolysis in Petrochemistry and Technology*; CITIC Publishing House: Beijing, 1993; CRC Press Inc.: Boca Raton, Florida, 1993.
7. Kniel, L.; Winter, O.; Tsai, T.C. Ethylene. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 3rd Ed.; John Wiley & Sons Inc.: New York, 1980; Vol. 9, 393–431.
8. Sundaram, K.M.; Shreehan, M.M.; Oleszewski, E.F. Ethylene. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th Ed.; John Wiley & Sons Inc.: New York, 1994; Vol. 9, 877–915.
9. Tsai, T.C.; Zou, R.; Zou, J. Model predicts yields from ethane–propane cocracking; Oil & Gas J. **1987**, 38.
10. Dente, M.E.; Ranzi, E.M.; Antolini, G.; Losco, F. Event Eur. Fed. Chem. Eng.; 97th, Florence, Italy Paper II-4, 1970.
11. Dente, M.E.; Ranzi, E.M.; Goossens, A.G. CACE '79; Symposium Comput. Appl. Chem. Eng.; 12th, Montreux, Switzerland, 1979; Paper 2A-1.
12. Dente, M.E.; Ranzi, E.M.; Barendregt, S.; Goossens, A.G.; AICHE 72nd Annual Meeting; San Francisco, California, 1979; Paper 38-b.
13. Dente, M.E.; Ranzi, E.M.; Pierucci, S.J.; Biardi, G.E.; Goossens, A.G.; Losco, F. ICP, Ind. Chim. Petrolifera **1979**, *7* (2), 67.
14. Dente, M.E.; Ranzi, E.M. Mathematical modeling of hydrocarbon pyrolysis reactions. In *Pyrolysis: Theory and Industrial Practice*; Albright, L.F., Crynes, B.L., Corcoran, W. H., Eds.; Academic Press: New York, 1983.
15. Allara, D.L.; Edelson, D. Int. J. Chem Kinet. **1975,** *7*, 479.
16. Edelson, D.; Allara, D.L. AIChE J. **1973**, *19*, 638.
17. Albright, L.F.; McConnell, C.F. *Thermal Hydrocarbon Chemistry*. Oblad, A.G.; Davis, H.G.; Eddinger, R.T., Eds.; Advances in Chemistry Series, No. 183; American Chemical Society: Washington, D.C., 1979; Chapter 12.
18. Albright, L.F.; McConnell, C.F.; Welther, K. *Thermal Hydrocarbon Chemistry*. Oblad, A.G.; Davis, H.G.; Eddinger, R.T., Eds.; Advances in Chemistry Series, No. 183; American Chemical Society: Washington, D.C., 1979; Chapter 10.
19. Tsai, T.C.; Albright, L.F. Surface reactions occurring during pyrolysis of light paraffins. In *Industrial and Laboratory Pyrolysis*; Albright, L.F. Crynes, B.L., Eds.; ACS Symposium Series, 32; American Chemical Society: Washington, D.C., 1976; Chapter 16.
20. Luan, Ta-Chi. Reduction of Coke Deposition in Ethylene Furnaces; Ph. D. Thesis; Purdue University: West Lafayette, Indiana, August 1993.

21. Albright, L.F. Improved Operation of Pyrolysis Unit: Reduced Formation and Collection of Coke in Ethylene Units; Paper presented at AIChE Meeting, Atlanta, Georgia, March 6, 2000.
22. Benum, L.W.; Wong, W.; Oballa, M.C. Treatment of Furnace Tubes. U.S. Patent No. 5,630,887, May 20, 1997.
23. Benum, L.W.; Wong, W.; Oballa, M.C. Surface Treatment of Stainless Steel. Canadian Patent Application CA 2164021, Filed on November 29, 1995.
24. Benum, L. Achieving longer furnace runs at NOVA chemicals. Paper presented at the Spring 2002 AIChE National Meeting, New Orleans, Louisiana, March 2002.
25. Zou, R. Fundamentals of pyrolysis in petrochemistry and technology. In *Coke Deposition During Hydrocarbon Pyrolysis and Its Inhibition*; CITIC Publishing House: Beijing; CRC Press Inc.: Boca Raton, Florida, 1993; Chapter 6.
26. Forester, D.R. U.S. Patent No. 4,962,264, October 9, 1990.
27. Albright, L.F.; Marek, J.C. Mechanistic model for formation of coke in pyrolysis units producing ethylene. Ind. Eng. Chem. Res. **1988**, *27*, 751–754. American Chemical Society; first in a series of three in this issue.
28. Albright, L.F.; Marek, J.C. Ind. Eng. Chem. Res. **1988**, *27*, 755–759. American Chemical Society; second in a series of three in this issue.
29. Albright, L.F.; McConnell C.F.; Welther, K. *Thermal Hydrocarbon Chemistry*; Oblad, A.G., Davis, H.G., Eddinger, R.T., Eds. Advances in Chemistry Series 183; American Chemical Society: Washington, D.C., 1979; 175–191.
30. Baker, R.T.K. personal Communications. **1985**.
31. Albright, L.F., Tsai, T.C-H. Importance of surface reactions in pyrolysis units. In *Pyrolysis: Theory and Industrial Practice*; Albright, L.F; Crynes, B.L.; Corcoran, W.H., Eds.; Academic Press: New York, 1983; Chapter 10.
32. Fernandez-Baujin, J.M.; Solomon, S.M. An industrial application of pyrolysis technology: Lummus SRT III module. In *Industrial and Laboratory Pyrolysis;* Albright, L.F., Crynes, B.L., Eds.; ACS Symposium Series 32; American Chemical Society: Washington, D.C., 1976; Chapter 20.
33. Leftin, H.P.; Newsome, D.S.; Wolff, T.J.; Yarze, J.C. Pyrolysis of naphtha and of kerosene in the Kellogg millisecond furnace. In *Industrial and Laboratory Pyrolysis;* Albright, L.F; Crynes, B.L., Eds.; ACS Symposium Series 32; American Chemical Society: Washington, D.C., 1976; Chapter 21.
34. Mathis, J.F. Intensified Use of Petroleum and Natural Gas for Petrochemicals; Tenth World Petroleum Congress, September 1979.
35. Witt, R.H.; Wall, F.M. Theoretical analysis of gas oil pyrolysis and product yield correlation. Paper Presented at AIChE 70th Annual Meeting, New York, 1977.
36. Lohr, B.; Schwab, W. Linde Rep. Sci. Technol. **1979**, *30*, 18.
37. Tsai, T.C. Yield correlation for AGO cracking are reliable. Oil & Gas J. **1985**, March 25, 96–101.
38. European Patent No. 81,101,665.8 and Germany Patent Application 010,000 (March 151980), (to BASF).
39. Bodke, A.S.; Olschki, D.A.; Schmidt, L.D.; Ranzi, E. High selectivities to ethylene by partial oxidation of ethane. Science **1999**, *285*, 712–715.

Thermal Desorption

Timothy P. Sullivan
Randolph Air Force Base, Texas, U.S.A.

INTRODUCTION

There are currently thousands of sites in the U.S.A. containing soil contaminated with volatile organic compounds (VOCs), semivolatile organic compounds (SVOCs), polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), pesticides, mercury, and mixed waste (radioactive and hazardous waste). One method used frequently to remediate these sites is thermal desorption.

Thermal desorption heats the contaminated soil to desorb and volatilize the contaminants, which are then either recovered or destroyed. This entry discusses the specifics of thermal desorption, including desorption kinetics, soil pretreatment requirements, types of thermal desorption units (TDUs), off-gas treatment, solids treatment, as well as system cost and performance.

EXTENT OF SOIL CONTAMINATION

The development of remediation technologies is driven by the large number of contaminated sites and the tremendous costs associated with cleanup. In the U.S.A., site cleanup is driven by two primary laws—the Recourse Conservation and Recovery Act (RCRA), which regulates operational sites, and the Comprehensive Environmental Response Compensation and Liability Act (CERCLA), which regulates historical sites (Superfund sites). Currently, there are 11,303 sites either on CERCLA's National Priority List (NPL) or under review for possible inclusion.[1] The numbers are even more striking in the underground storage tank (UST) portion of the RCRA program. The US Environmental Protection Agency (USEPA) estimates that as of March 2002, there have been 422,573 releases from USTs and that approximately 145,000 cleanups have yet to be completed.[2]

One technology suitable for remediating these sites is thermal desorption, which heats soils and sludge to remove volatile contaminants such as VOCs, SVOCs, PAHs, PCBs, pesticides, mercury, and mixed waste for recovery or further treatment. The success of thermal desorption is evidenced by how frequently it has been used since its initial demonstration. From fiscal years (October 1 through September 30) 1982 through 1999, thermal desorption was used in 14% of Superfund sites requiring ex situ source control, 2% of Superfund sites requiring in situ source control, and 9% of total Superfund sites.[3] (Please note that the USEPA classifies in situ thermal desorption as thermally enhanced recovery.) The utilization of thermal desorption in the RCRA cleanup program is similar, with thermal desorption employed at 9% of cleanup sites through August 1998.[4] As thermal desorption is used more frequently, it is important to understand the operational advantages, as well as limitations, of these systems.

APPLICABILITY OF THERMAL DESORPTION

Results of cleanup actions at various sites illustrate that thermal desorption is effective in treating soil and sludge contaminated with VOCs, SVOCs, PAHs, and PCBs under the right conditions.[5] Thermal desorption is also effective in treating soils contaminated with organic pesticides, although it is not considered effective in treating wastes contaminated with inorganic or metal-containing pesticides.[6] While not its primary use, thermal desorption may also be used to treat mercury and mixed (radioactive and hazardous) wastes. Studies conducted by the US Department of Energy indicate that thermal desorption is effective in treating mercury waste and that thermal desorption would be effective in removing hazardous constituents from mixed waste without volatilizing radioactive material.[7,8] This would separate the mixed waste into radioactive waste and hazardous waste allowing them to be managed separately.

DESORPTION KINETICS

An understanding of thermal desorption begins with an explanation of desorption kinetics. In thermal desorption systems, the portion of the contaminant in the liquid phase is removed as the soil temperature is increased above the boiling point of water and the contaminant. However, any portion of the contaminant bound to the soil must still be desorbed prior to removal.

The desorption of contaminants from soils is influenced by three primary parameters: temperature, soil moisture content, and soil type. A recent study evaluating

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007731

the temperature dependence of desorption looked at the sorption of PAHs to sediment and determined that PAHs have a high activation energy when bound to coal particles. The high activation energy indicates that the desorption of PAHs increases significantly with temperature.[9] Another study analyzed the desorption of motor oil and PCBs from porous materials and determined that PCBs have an onset temperature of 350°C below which desorption does not occur.[10] These results imply that it is essential to operate thermal desorption systems, such that the temperature of the soil exceeds the desorption onset temperature and ensures full desorption and volatilization of the contaminants of concern. A sufficient operating temperature is needed to ensure contaminant volatilization and is typically determined during trial runs of the thermal desorption system using contaminated soil from the actual cleanup site.

Soil moisture content also plays a critical role in the effectiveness of thermal desorption. At levels below 20%, soil moisture plays a beneficial role by removing contaminants through steam distillation. When the soil is heated, the water in the soil undergoes a phase change to steam. During the phase change, contaminants with lower boiling points boil with the water and partition into the gas phase where they are removed from the soil with the steam.[11] While steam distillation is beneficial at lower moisture contents, higher soil moisture contents inhibit thermal desorption systems as additional energy is required to volatilize the water. This also increases the cost of treatment, the complexity of the air pollution control (APC) system, and the residence time of the soil in the thermal desorption unit.

The type of soil also plays an important role in determining how readily contaminants will desorb when treated with thermal desorption. Studies conducted on harbor sediment indicate organic contaminants exhibit preferential sorption to coal and coal-derived particles over sand, silt, and clays.[12] Researchers found that while coal/wood derived particles only accounted for 5% of particles in the test soil, 62% of total PAHs were sorbed to these particles. Successive extractions were conducted and revealed that only 8% of the PAHs sorbed to the coal/wood-derived particles desorbed within three months, while 80% of PAHs sorbed to clay and silt particles desorbed within one month.[12] Another study demonstrated that it is more difficult to desorb contaminants from soils with a high humic content. In this study, aeration of a slurry containing PAHs revealed that 20% of the naphthalene bound to inorganic material was removed after 30 min of aeration, while only 10% of the naphthalene bound to humic material was removed.[13]

THERMAL DESORPTION SYSTEMS

Thermal desorption systems can be broken into three key components: pretreatment, the TDU, and post-treatment (Fig. 1) as discussed below. Additionally, thermal desorption systems are often used in conjunction with other treatment technologies as part of a treatment train. At previous RCRA and Superfund sites, thermal desorption has been used in conjunction with bioremediation, dechlorination, incineration, soil vapor extraction, soil washing, and solidification/stabilization.[3,4] In these cases, the combination of technologies was required to meet the site's cleanup goals.

Pretreatment

Pretreatment involves analyzing and preparing the waste stream for treatment in the TDU. The key steps involved are analyzing the waste, soil mixing, segregating large particles, and dewatering.

Under an RCRA or a Superfund cleanup, the waste is characterized prior to selection of the cleanup technology. With this information, the following pretreatment samples are taken as appropriate: contaminants of concern, total petroleum hydrocarbons (TPH), chlorinated pesticides, PCBs, chlorinated solvents, mercury, and/or radionuclides.[14] These samples are then used to determine the required removal efficiency of the cleanup technology to ensure the cleanup goals can be met. Prior to using a thermal desorption system (if selected), the waste must be sampled to spatially characterize the contamination so that the soil that is

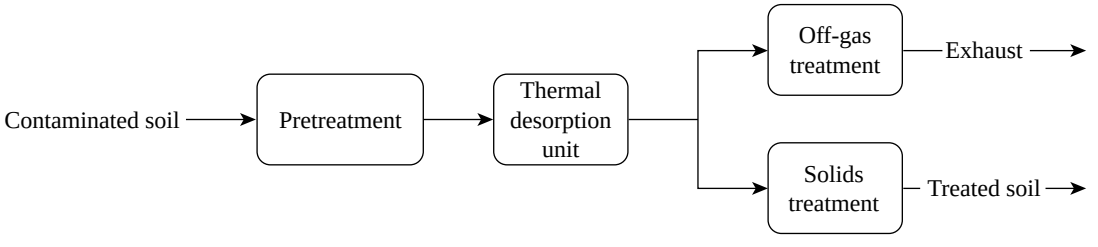


Fig. 1 Steps in a thermal desorption system.

most difficult to treat will be used during test runs of the TDU to establish the unit's operating parameters.[14]

Pretreatment samples are also useful at sites where wide variations in the contaminant concentration occur and soil mixing is utilized. Soil mixing is used to ensure a relatively homogeneous feedstock for the TDU so that it can operate effectively and efficiently. Soil mixing has been used at several cleanup sites to prevent clumping of soil, which impairs treatment, and to prevent temperature spikes within the TDU and consequent variations in off-gas properties.[5]

Segregation of larger particles is necessary to maintain the consistency of the feedstock and to ensure treatment goals are met. Particles greater than 5 cm in diameter can interfere with the TDU in the following ways: first, they can clog the waste feed system leading to lengthy system downtimes and operational upsets;[15] second, larger particles take longer to heat to the desired treatment temperature and may require a longer residence time for complete treatment. Since the TDU operates at a fairly constant feed rate, these larger particles are not likely to meet the treatment goals. Problems with larger particles are typically addressed by either crushing or passing the feedstock through a screen and removing the larger pieces for separate cleaning or disposal as hazardous waste.

Soils with a moisture content above 20% must also be dewatered or treated to prevent excessive energy costs and residence times in the TDU. Dewatering may be carried out with a filter press, air drying, or gravity separation.[5] Alternatively, contaminated soil is sometimes treated with lime to lower the moisture content and improve the material handling characteristics.[5] Whichever method is chosen, the goal is to achieve a soil moisture content of approximately 8–12%.[16] This allows for the benefits of steam distillation without excessive fines in the off-gas as would happen if the soil were too dry.

Thermal Desorption Unit

Once the contaminated material is pretreated, it is treated in the TDU. There are two broad categories of TDUs: in situ (also referred to as thermally enhanced recovery) and the more common ex situ. In situ involves heating the soil in place to remove contaminants, while ex situ units require the excavation of contaminated soil for treatment in the TDU.

In situ units

The two types of in situ TDUs are thermal blankets and thermal wells. Thermal blankets operate by placing heating elements directly on the contaminated soil, covering the area with an impermeable barrier to minimize fugitive emissions, and collecting the off-gas for further treatment. In one field test, a 3.05×3.05 m heater was installed at a site contaminated with hexadecane. In this system, approximately $5382\ W/m^2$ of electrical power was supplied to the soil for $2\frac{1}{2}$ days causing the soil to reach 576°C at the surface, 345°C at a depth of 15.2 cm, 195°C at 30.5 cm, and 102°C at 45.7 cm.[17] The results of this test indicate that 99.94% of the hexadecane was removed from the first 15.2 cm of soil, 98% was removed from the soil at depths of between 15.2 and 30.5 cm, and 75% was removed from depths of between 30.5 and 45.7 cm. Interestingly, in this demonstration, only 50 g of hexadecane was found in the off-gas, with the remainder being oxidized or pyrolized in the soil.[17] These results indicate that thermal blankets have the potential to successfully treat surface and near-surface volatile contamination. They also have the advantage of not requiring excavation, which involves increased cost and fugitive emissions.

While thermal blankets may treat contaminants near the surface, thermal wells offer the ability to treat contamination at much greater depths. In thermal well systems, heater wells are inserted into the contaminated soil to volatilize contaminants in situ. The generated off-gas is then collected in combination heater–vacuum wells and sent to an APC system. The number of heater–vacuum wells, and the well spacing, depend on site-specific parameters, but typically, one-third of the wells are heater–vacuum wells and the well spacing is 1.8–2.1 m.[11] With heaters operating at up to 800°C, thermal well systems are able to achieve soil temperatures of 300–500°C between wells.[18] In order to fully remediate contaminated sites, the wells are inserted up to 0.9 m below the contaminated layer and operate for 30–60 days depending on the soil characteristics and contaminants of concern.[19]

Much like thermal blankets, thermal well systems do not require costly excavation and they also offer additional benefits. They have been used to treat contaminants at depths up to 5.5 m below the surface and much of the contaminants are destroyed in situ through oxidation or pyrolysis reactions.[19] Furthermore, thermal well systems offer uniform heating and consequent treatment of contaminants is effective across a wide range of soil types. The long residence time favors desorption mechanisms that may be time dependent.[18,19]

Ex situ units

Ex situ units are used more commonly than in situ units for two primary reasons. First, in many locations, groundwater intrusion can limit the effectiveness and increase the cost of in situ units. This can be overcome

in ex situ units by dewatering the soil and treating the water in a wastewater treatment plant. Second, if the soil also contains metals, ex situ units can be used as the first step in treatment with solidification/stabilization used to immobilize the metals. The most common types of ex situ units are batch units, thermal screws, and rotary kilns.

Batch units

Batch TDUs are typically only used for smaller applications because their throughputs are often much lower than other thermal desorption systems. These systems do have the benefit of requiring less pretreatment of the contaminated media because feedstock consistency and particle size are not as crucial to control as they are in continuous feed units. Unlike continuous feed units, it is easy to vary the residence time and/or temperature between batches to handle different contaminant concentrations in the feedstock. Also, there is no material feeding equipment to clog with larger particles, although it is still not economical to treat extremely large particles in a batch process. However, particle sizes up to 15.2 cm have been treated effectively by batch units.[5]

After limited pretreatment, the batch unit is loaded with contaminated soil and sealed to allow generation of a vacuum. The amount of vacuum applied depends on the contaminants being treated. A system used to treat VOCs typically operates at a vacuum of up to 8 kPa, whereas one used to treat mixed waste containing mercury operates at a vacuum of 84.7 kPa.[5,7] In the presence of this vacuum, contaminants are volatilized at lower temperatures. Due to this and the lack of oxygen, combustion is prevented.[5] The vacuum is also beneficial when treating mercury waste as no sweep gas is used in the system, thereby allowing volatilized mercury to be condensed on cold-water impingers and recovered.[7] This helps to prevent fugitive emissions from the system, which is especially critical when dealing with mercury-contaminated wastes due to its high toxicity. If not properly controlled, mercury emissions from the thermal desorption system can reach concentrations high enough to cause unhealthy conditions for onsite workers and neighboring populations.[14]

With the vacuum applied, the feedstock is indirectly heated up to the operating temperature. Operating temperatures vary widely depending on the waste being treated. In one system, the operating temperature was limited to 82.2°C with a 60-min residence time for clay soils and a 120–150 min residence time for highly contaminated soils to prevent the volatilization of radioactive metals.[5] A system treating mercury-containing wastes, however, was operated at temperatures up to 750°C for 10–20 min.[7]

Thermal screws

In contrast to batch units, thermal screws operate with continuous feed as hollow flight rotary screws mix, heat, and convey the contaminated feedstock through the system. Either hot oil or steam is circulated through the flights to indirectly heat the feedstock. Typically, systems using hot oil can achieve operating temperatures up to 260°C while systems utilizing steam are limited to temperatures of 177°C.[20] These limitations make thermal screw systems more appropriate for treating contaminants with lower boiling points.

Operationally, the feedstock is loaded into the TDU with earthmoving equipment and then mixed, heated, and conveyed with one to four screws arranged either in series or parallel. Multiple screws are operated in series to increase the soil's residence time, and are operated in parallel to increase the system's throughput up to 25 tons per hour (tph).[21] The residence time is highly variable and depends on the soil moisture content, soil type, and contaminants being treated, but residence times up to 90 min have been reported.[5] Once the contaminants are removed from the soil, they are carried to the APC system by a sweep gas such as steam.

Notably, when thermal screw systems employ multiple interlocking screws, they are susceptible to damage from larger particles lodging between the screws. This can damage the flights and has even damaged the drive gearbox, although reversing the screws can clear minor jamming.[15] In one commercial system, this is avoided by screening the feed material down to 2.5 cm instead of the more common 5 cm.[21]

Despite the temperature limitations of thermal screw systems, there are some advantages. First, since thermal screws operate at a lower temperature, they require less fuel and do not combust the contaminants, thereby allowing for contaminant recovery. Second, since these systems are indirectly heated, the volume of off-gas generated is lower. Finally, the interlocking screws increase mixing of the feedstock, which improves treatment.[15]

Rotary kilns

The most common type of TDU is the rotary kiln where the contaminated soil is fed into a rotating drum at a slight incline. As the drum rotates, internal flights agitate the soil and convey it through the drum to the outlet. The drum is typically constructed of either carbon steel or stainless steel and is sized to meet the amount of soil to be remediated with drum sizes ranging from 6.1 m long by 1.5 m in diameter to 14.3 m long by 2.6 m in diameter.[5] In order to prevent fugitive emissions, induced draft fans are utilized to maintain a

negative pressure of 0.01–0.05 kPa throughout the system.[5]

Heating of the soil is achieved by burners located either inside the shell or near the outside surface operating on natural gas, propane, or fuel oil. Burner sizes vary between 6.15 and 117.12 MW.[5] When the combustion gases from the burner directly contact the soil, the unit is called direct-fired, and when the combustion gases heat the outer shell of the rotary kiln, the unit is called indirect-fired. A notable difference in the two firing methods is how they are perceived by regulatory agencies. The USEPA considers direct-fired rotary kilns incinerators and requires them to be permitted as such, while indirect-fired systems are considered a ''miscellaneous unit.''[22] Indirect units also allow the use of an inert sweep gas, such as nitrogen, to prevent the combustion of contaminants if the formation of dioxins and furans is a concern or if there is a desire to recover the volatilized contaminants. An additional distinction between direct and indirect-fired units is the residence time, which averages 15–20 min for direct-fired units and up to 2 hr for indirect-fired units.[5] During operation, the residence time is adjusted by altering the drum's rotational speed and incline.

The flow of the combustion gases in direct-fired systems may be concurrent (i.e., the gases flow in the same direction as the feedstock) or countercurrent (i.e., the gases flow opposite to the movement of the feedstock). The benefits of a concurrent system are higher treatment temperatures, longer residence times, and treatment of fines. Countercurrent systems offer more efficient heat transfer, and lower volumes and temperatures of off-gas.[20] In large part due to the beneficial aspects of the off-gas and consequent simplified APC system, countercurrent systems are used more commonly in rotary kilns.

There is a wide variation in the system throughput for rotary kiln TDUs with values ranging from 5 to 150 tph. However, most systems operate in the 15–30 tph range.[5] The actual throughput achieved depends not only on the design capacity of the rotary kiln, but also on the treatment criteria. Treatment criteria determine the maximum soil depth in the kiln and the required residence time.

All rotary kilns operate to achieve a desired soil output temperature to ensure complete treatment of the contaminated soil. The specific soil exit temperature is determined during startup testing to ensure the soil is properly treated in the TDU. This temperature generally falls into two groups based on residence time: (1) less than or equal to 260°C and a residence time of approximately 6–7 min and (2) greater than 260°C and a residence time of approximately 15–20 min.[5]

The benefits of using a rotary kiln over other ex situ TDUs include the ability to achieve high system throughputs and high soil temperatures during a fairly short residence time. This allows the treatment of multiple soil types and higher boiling point contaminants. Another important factor is that rotary kiln technology is mature and available from multiple vendors, thereby increasing availability and lowering cost.

System monitoring and safety interlocks

During operation, the TDU's operating parameters are measured and recorded to ensure correct operation and to prevent accidental releases of contaminants. While there are several possible parameters to monitor, the following are recommended where applicable: treated soil exit temperature, vacuum in the TDU, pressure drop in the APC, waste feed rate, afterburner temperature, off-gas exit temperature from the TDU, stack gas velocity and temperature, and the flow rate and pH of acid–gas scrubber liquor.[14]

The selected parameters can be tied in to a safety interlock system that provides a warning to the operator, adjusts system parameters, or shuts the system down during an upset to prevent the accidental release of contaminants. Instantaneous system shutdown is recommended if the following upsets are detected: primary burner or heater system failure; blower failure or loss of negative pressure; or a pressure drop in the APC outside allowable range. Shutdown after a 10-min delay is recommended for the following situations: soil exit temperature below limit; high carbon monoxide level in oxidizer exhaust; waste feed rate exceeds limit; or a significant change in the gas velocity through the APC. A final interlock will shut the system down in 30 sec to 2 min if the oxidizer temperature is below the minimum allowable level.[14]

Off-Gas Treatment

After exiting the TDU, the off-gas contains volatilized contaminants that are then either destroyed in an oxidizer (afterburner) or collected for further treatment. Treatment in a thermal oxidizer is a fairly common option in thermal desorption systems and is discussed first.

Thermal destruction

While there are several possible APC configurations to destroy volatilized contaminants, a typical system is shown in Fig. 2. The first step in the process involves removal of fines entrained in the off-gas using either a cyclone separator or a baghouse.

Cyclone separators operate by directing the air stream into a conical shell where the off-gas circulates downward along the outer edge of the shell. Due to their inertia, particles strike the shell and fall to the

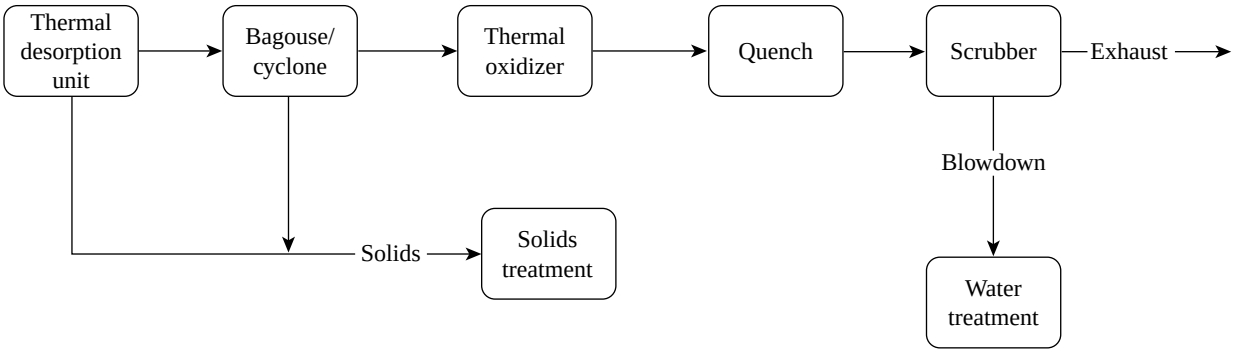


Fig. 2 Off-gas treatment with thermal destruction of contaminants.

bottom of the cone where they are collected. The air stream makes several rotations through the cone before changing direction at the bottom of the cone and exiting through the top of the cyclone. Cyclone separators may be operated individually or in series and can achieve particulate removal efficiencies of 90–95%, although they are not effective in removing particulate smaller than 10 μm.[23]

Baghouses are used to control particulate matter emissions to submicron size at control efficiencies greater than 99%.[23] In baghouses, gases are passed through several fabric bags and the particulate matter is deposited on the bag. During normal operation, the bags are periodically cleaned and the removed fines are combined with soils exiting the TDU for further treatment.

The bags may be made of several different materials depending on the off-gas temperature and the presence of acid gases. Common fabrics used in baghouses include fiberglass, Nomex®, Ryton®, and Teflon®[a] with their operating temperatures and acid gas resistance shown in Table 1.[24] After choosing the appropriate bag material, the bags are sized using the primary design criterion for baghouses, the air-to-cloth (A/C) ratio. The A/C ratio measures the ratio of actual cubic feet of air flow to the square feet of cloth area. In thermal desorption systems, typical values of A/C are 0.76 actual cubic meters per minute per square meter ($acmm/m^2$) for woven fabric bags and 2.13–3.05 $acmm/m^2$ for felt bags, which can lead to bags as large as 0.3×9.1 m.[24] The final important parameter in baghouse systems is the pressure drop across the bags, which is typically 0.5–3.0 kPa.[25]

After exiting the baghouse, the off-gas is directed to the thermal oxidizer consisting of a refractory lined cylinder with a burner. The burner rapidly heats the off-gas to oxidize the organic contaminants. The required organics destruction efficiency in the oxidizer is generally determined on a case by case basis during permitting, however, at a minimum, regulators use the standard Destruction and Removal Efficiency (DRE) specified under RCRA regulations of 99.99% for principle organic hazardous constituents, and 99.9999% for wastes containing dioxins and furans [40 Code of Federal Regulations (CFR) Parts 264 and 265]. Regulations promulgated under the authority of the Toxic Substances Control Act also require a DRE of 99.9999% for wastes containing PCBs (40 CFR Part 761).

The DRE in a thermal oxidizer is affected by the three T's of time, temperature, and turbulence. In order to achieve the required destruction efficiencies, the burner heats the off-gas to temperatures between 760°C and 982°C in a turbulent mixing environment where the contaminants are held for a residence time of up to 2 sec.[5] The turbulent environment is critical to ensuring thorough mixing of the off-gas in the combustion chamber thereby preventing any short circuiting where off-gas exits the combustion chamber prior to being treated.

RCRA also limits the amount of HCl contained in the thermal oxidizer exhaust to 4 lb/hr or less than 1%, whichever is less (40 CFR 264 and 265). Achieving this requires a quench process followed by an acid gas scrubber. The quench process lowers the gas temperature to the operating temperature of the acid gas scrubber, which must be below the boiling point of the acid gases to prevent revolatilization. Quench systems operate by spraying water into the gas stream to cool the gas as the water evaporates.

In the acid gas scrubber, water is again sprayed into the gas stream. However, the gas is forced through a packed bed of acid-resistant material, such as polyvinylchloride, to increase its contact with the water. By forcing the gas through this tortuous path, larger amounts of HCl are absorbed by the water. Removal efficiencies of 70% are achievable when using plain water and 93–96% when using water with an alkaline

[a]Teflon® and Nomex® are registered trademarks of DuPont. Ryton® is a registered trademark of Phillips Petroleum Company. The mention of companies and trade names does not constitute endorsement or recommendation for use.

Table 1 Operating temperature and acid gas resistance for various bag materials[c]

Bag material	Operating temperature (°C)	Acid gas resistance
Fiberglass	260	Fair to good
Nomex®[a] (Aramid)	191	Fair
Ryton®[b] (Polyphenylene sulfide)	191	Excellent
Teflon®[a] (Polytetrafluoroethylene)	232	Inert except for fluorine

[a]DuPont registered trademark.
[b]Phillips Petroleum Company registered trademark.
[c]From Ref.[24].

material in it.[23] The wastewater from the acid gas process can then either be neutralized or treated in a wastewater treatment plant depending on its characteristics. As this represents the last step in the APC system, the off-gas leaves the scrubber through a stack and is vented to atmosphere.

Contaminant Recovery

An alternative to destroying contaminants in a thermal oxidizer is to recover them for further treatment. As shown in Fig. 3, the APC system for contaminant recovery also begins with a baghouse/cyclone to remove particulate matter. However, the off-gas is then directed to a condenser instead of a thermal oxidizer.

Condensers operate by cooling the off-gas stream, leading to the condensation of the volatilized contaminants. A typical system consists of a tube and a shell heat exchanger where the off-gas is passed through the shell containing refrigerant tubes. As refrigerant circulates through the tubes, the off-gas is cooled and the contaminants condense on the outside of the tubes. These contaminants then gravity drain to the bottom of the shell where they are recovered for further treatment. The removal efficiency of these systems depends on the contaminants in the off-gas, but efficiencies of 90% are attainable when using chilled water, ammonia, or chlorofluorocarbons as the refrigerant.[26] (Note: The Montreal Protocol phases out production of chlorofluorocarbons and they are no longer used in new equipment.)

After exiting the condenser, the off-gas still contains enough contaminants to be of concern; therefore the gas is then directed to a carbon adsorption unit as a final polishing step. In these systems, the off-gas is passed through a bed of activated carbon (either powdered or granular) and any entrained organic contaminants are adsorbed on to the surface of the carbon. The unit is sized based on the inlet contaminant concentration, gas flow rate, and desired outlet concentration. It should be noted that special designs are generally not required for thermal desorption systems because fixed bed adsorption units to treat VOC streams of 5000 ppm at 2832 standard cubic meters per minute (scmm) to an outlet concentration of 25 ppm are commercially available.[27] Periodic monitoring of VOCs in the adsorber exhaust will show when the carbon is saturated with VOCs. At this point, the carbon may either be regenerated through heating or disposed of

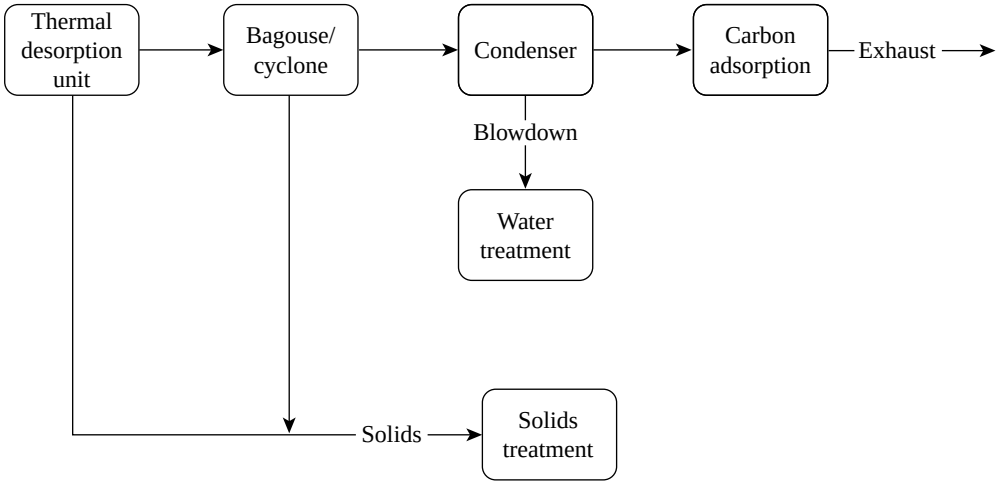


Fig. 3 Off-gas treatment with system with contaminant recovery.

as hazardous waste and replaced. After passing through the activated carbon bed, the off-gas is vented to the atmosphere through a stack.

Emission limits and monitoring

In addition to the minimum DRE and the limits on HCl for thermal oxidizers already mentioned, regulatory agencies will place emission limits on the exhaust gas depending on local regulations and the contaminants of concern. The most commonly regulated emissions for thermal oxidizer exhaust are particulate matter (PM), CO, HCl, sulfur oxides (SO_x), VOCs, total hydrocarbons (THC), and nitrogenoxides (NO_x). Alternatively, regulators typically place limits on PM, THC, and/or VOC emissions from contaminant recovery systems.[5]

Regulators verify that the emissions are within limits in different ways for thermal oxidizers and contaminant recovery systems. Since thermal oxidizers involve combustion, the combustion process parameters of CO, CO_2, and O_2 are often monitored continuously to verify that complete combustion of contaminants occurs in the thermal oxidizer.[5] The acceptable range for these parameters is determined during a performance test where the optimal operating conditions for treating the contaminants of concern are determined. Additionally, regulatory agencies typically require periodic exhaust gas sampling to further demonstrate that emissions limits are being met.

In contaminant recovery systems, regulators typically do not require continuous monitoring of the exhaust gas. As with thermal oxidizers, emissions are monitored during an initial performance test to determine the appropriate operating parameters. In these systems, the assumption is that exhaust gas emissions will meet the regulatory limits provided the system is operated in the same manner and within the limits established during the performance test. Regulators may, however, require periodic exhaust gas sampling to verify this assertion.

Solids Treatment

The second waste stream generated from the TDU is the soil. After exiting the TDU, the first step in secondary treatment of the soils involves adding water to both cool the soil and minimize the generation of fugitive dust. After this step, the soil is typically placed in piles for analysis to determine if the treatment standards were met and whether or not the soil is hazardous.

Depending on the results of the analysis, one of three things happens to the soil. In the case of nonhazardous soil, typically no further treatment is necessary and the soil is used as backfill on-site. Soil that fails to meet the cleanup goals is processed through the TDU again and reanalyzed. If the soil is hazardous but within the cleanup goals, which occurs when sufficient metals are present, it is either disposed of off-site as a hazardous waste or treated further with a method such as solidification/stabilization.[3]

COST AND PERFORMANCE

The cost of thermal desorption treatment can depend heavily on the amount of soil treated because the cost per ton of soil treated decreases as economics of scale are realized with increasing soil volume. An analysis conducted by the USEPA using 2000-yr dollars shows that for systems treating less than 15,000 tons of soil, the cost is between $125 and $275 per ton; for systems treating between 15,000 and 30,000 tons of soil, the cost is between $100 and $125 per ton; and for systems treating over 30,000 tons, the cost is between $60 and $100 per ton.[28] These costs compare favorably with 2002-yr dollars amounts for on-site incineration, the most comparable technology. For incineration systems treating less than 15,000 tons of soil, the cost ranges from $770 to $3200 per ton; when treating between 15,000 and 30,000 tons of soil, the cost ranges from $650 to $3300 per ton; and when treating over 30,000 tons of soil, the cost ranges from $230 to $800 per ton.[5]

The amount of soil treated is not the only factor affecting cost, however. The following are the most commonly identified factors affecting the cost of treatment: moisture content of the soil, soil characteristics, amount of debris with the waste, initial contaminant concentration, target contaminant concentration, and utility rates.[29]

While cost is an important consideration when selecting a remediation technology, the primary consideration is the technology's effectiveness in treating the contaminants of concern. Specific treatment goals are determined for each site based on acceptable residual risk, but it is useful to look at the range of removal efficiencies. Table 2 shows that thermal desorption has proven effective in treating a wide range of contaminants, which explains why thermal desorption has been used at several sites.

Table 2 Thermal desorption removal efficiencies[a]

Contaminant	High (%)	Low (%)
Polycyclic aromatic hydrocarbons	97.8723	92.5
Polychlorinated biphenyls	99.9999	98
Volatile organic compounds	99.9970	98.8889
Pesticides	99.9839	97.8723

[a]Developed from Ref.[5].

CONCLUSIONS

After its initial development as an innovative technology for treating soil contaminated with VOCs, SVOCs, PAHs, PCBs, mercury, and pesticides, thermal desorption has become a preferred method for treating these wastes. Thermal desorption has been successfully used in several sites, consistently achieving removal efficiencies greater than 97%, at an attractive cost.

The future of thermal desorption as a remediation technology is promising, and it has already been used in over 146 sites in the U.S.A. As of 2003, the USEPA's REmediation And CHaracterization Innovative Technologies (REACH IT) database lists 77 vendors offering thermal desorption services.[29] The expectation is that more and more contaminated sites will be identified, as insurers routinely require environmental investigations prior to the transfer of industrial properties and regulatory agencies continue to promulgate increasingly stringent cleanup standards. Since the USEPA identified thermal desorption as a presumptive remedy for treating many contaminated soils, its use has risen and will continue to grow and be refined and improved.[30]

REFERENCES

1. Superfund Accomplishment Figures, Summary: Fiscal Year (FY) 2002.; http://www.epa.gov/superfund/action/process/numbers.htm (accessed January 2003).
2. UST Program Facts. U.S. Environmental Protection Agency: Washington, DC, 2002.
3. *Treatment Technologies for Site Cleanup: Annual Status Report*, 10th Ed.; EPA-542-R-01-004; U.S. Environmental Protection Agency: Washington, DC, 2001.
4. *Treatment Technologies for Site Cleanup: Annual Status Report*, 9th Ed.; EPA-542-R-99-001; U.S. Environmental Protection Agency: Washington, DC, 1999.
5. *FRTR Cost and Performance Remediation Case Studies and Related Information*, 3rd Ed.; EPA-542-C-02-004; U.S. Environmental Protection Agency: Washington, DC, 2002.
6. Koustas, R.N.; Fischer, D. Review of separation technologies for treating pesticide-contaminated soil. J. Air Waste Manag. Assoc. **1998**, *48* (5), 434–440.
7. Morris, M.I.; Osborne-Lee, I.W.; Hulet, G.A. *Demonstration of New Technologies Required for the Treatment of Mixed Waste Contaminated with ≥260 ppm Mercury*; ORNL/TM-2000/147; U.S. Department of Energy: Oak Ridge, TN, 2002.
8. Cavanaugh, R.; Molina, M.; Anderson, C.; Athy, A.; Bardacke, P.; Budnitz, R.J.; Glickman, G.L.; Resnikoff, M.; Till, C. *Report of The Secretary of Energy Advisory Board's Panel on Emerging Technological Alternatives to Incineration*; U.S. Department of Energy: Washington, DC, 2000.
9. Ghosh, U.; Talley, J.W.; Luthy, R.G. Particle-scale investigation of PAH desorption kinetics and thermodynamics from sediment. Environ. Sci. Technol. **2001**, *35* (17), 3468–3475.
10. El-Shoubary, Y.; Woodmansee, D.E.; Shilling, N.Z. Sorption and desorption of contaminants from different host matrices. Environ. Progress. **1993**, *13* (1), 37–44.
11. Baker, R.S.; LaChance, J.C. In-situ thermal destruction (ISTD) performance relative to dioxins. In *Current Practices in Oxidation and Reduction Technologies for Soil and Groundwater (in press)*, Presented at the 2nd International Conference on Oxidation and Reduction Technologies for Soil and Groundwater, Al-Ekabi, H., Ed.; ORTs-2, Toronto, Ontario, Canada, Nov. 17–21, 2002.
12. Upal, G.; Gillette, J.S.; Luthy, R.G.; Zare, R.N. Microscale location, characterization, and association of polycyclic aromatic hydrocarbons on harbor sediment particles. Environ. Sci. Technol. **2000**, *34* (9), 1729–1736.
13. Zeng, Y.; Hong, P.K.A. Slurry-phase ozonation for remediation of sediments contaminated by polycyclic aromatic hydrocarbons. J. Air Waste Manag. Assoc. **2002**, *52* (1), 58–68.
14. *Technical Guidelines for On-site Thermal Desorption of Solid Media and Low Level Mixed Waste Contaminated With Mercury and/or Hazardous Chlorinated Organics*; Interstate Technology and Regulatory Cooperation Work Group, Low Temperature Thermal Desorption Work Team: Washington, DC, 1998.
15. *Low Temperature Thermal Treatment (LT$^{3®}$) Technology,* Roy F. Weston, Inc., Applications Analysis Report; EPA-540-AR-92-019; U.S. Environmental Protection Agency's Risk Reduction Engineering Laboratory: Washington, DC, 1992.
16. Rosta, P.R.; Rubin, D.K. Thermal desorption heats up as waste cleanup technology. ENR. **1993**, *233* (7), 36–37.
17. Stegemeier, G.L.; Vinegar, H.J. Soil remediation by surface heating and vacuum extraction. Proceedings of 1995 SPE/EPA Exploration and Production Environmental Conference, Houston, TX, March 27–29, 1995; Society of Petroleum Engineers: Richardson, TX, Paper Number SPE 29771.
18. Baker, R.S.; Kuhlman, M. A description of the mechanisms of in-situ thermal destruction (ISTD) reactions. In *Current Practices in Oxidation and Reduction Technologies for Soil and*

Groundwater (in press), Presented at the 2nd International Conference on Oxidation and Reduction Technologies for Soil and Groundwater, Al-Ekabi, H., Ed.; ORTs-2, Toronto, Ontario, Canada, Nov. 17–21, 2002.

19. Vinegar, H.J.; Stegemeier, G.L.; Carl, F.G.; Stevenson, J.D.; Dudley, R.J. In situ thermal desorption of soils impacted with chlorinated solvents. Proceedings of the Annual Meeting of the Air and Waste Management Association, St. Louis, MO, Jun 20–24, 1999; A&WMA: Pittsburgh, PA, 1999; Paper Number 99–450.
20. *How to Evaluate Alternative Cleanup Technologies for Underground Storage Tank Sites, A Guide for Corrective Action Plan Reviewers*; EPA-510-B-95-007; U.S. Environmental Protection Agency: Washington, DC, 1995.
21. Young, C.A.; Hatch, H.; Leibbert, J. On-site thermal desorption of PCP- and dioxin-contaminated soil at the Coleman-Evans wood preserving site, Whitehorse, Florida. Proceedings of the 2000 International Conference on Incineration and Thermal Treatment Technologies(IT3), Portland, OR, May 8–12, 2000; University of Maryland: College Park, MD, 2000.
22. Shapiro, M. *Clarification on the Distinction Between Thermal Desorbers and Incinerators*; 9489.1994(01); U.S. Environmental Protection Agency: Washington, DC, 1994.
23. *Air CHIEF*; EPA-454-C-01-003; U.S. Environmental Protection Agency: Washington, DC, 2001.
24. Turner, J.H.; McKenna, J.D.; Vatavuk, W.M. Baghouses and filters. In *EPA Air Pollution Cost Control Manual,*; 6th Ed.; EPA-452-B-02-001; U.S. Environmental Protection Agency: Research Triangle Park, NC, 2002; 564–621.
25. Merritt, R.L.; Bush, P.V. Status and future of baghouses in the utility industry. J. Air Waste Manag. Assoc. **1997**, *47* (6), 704–709.
26. Shareef, G.S.; Vatavuk, W.M. Refrigerated condensers. In *EPA Air Pollution Cost Control Manual*, 6th Ed.; EPA-452-B-02-001; U.S. Environmental Protection Agency: Research Triangle Park, NC, 2002; 236–280.
27. Vatavuk, W.M.; Klotz, W.L.; Stallings, R.L. Carbon adsorbers. In *EPA Air Pollution Cost Control Manual*, 6th Ed.; EPA-452-B-02-001; U.S. Environmental Protection Agency: Research Triangle Park, NC, 2002; 195–235.
28. *Remediation Technology Cost Compendium—Year 2000*; EPA-542-R-01-009; U.S. Environmental Protection Agency: Washington, DC, 2001.
29. EPA REACH-IT; http://www.epareachit.org/index.html (accessed Jan 2003).
30. *Presumptive Remedies: Policy and Procedures*; EPA-540-F-93-047; U.S. Environmental Protection Agency: Washington, DC, 1993.

Thermal Stability of Chemical Reactors

Haishan Zheng
Jason M. Keith
Department of Chemical Engineering, Michigan Technological University, Houghton, Michigan, U.S.A.

INTRODUCTION

Even the simplest of chemical reactors can exhibit complex steady-state behavior, including multiplicity and ignition–extinction hysteresis. In addition, there can be dynamic phenomena which can occur during reactor startup or in response to perturbations in operating variables in the approach to or from a steady state, or within a limit cycle or chaotic attractor. In the last couple of decades, a great deal of attention has been focused on purposely operating chemical reactors in unsteady-state conditions. To properly design such systems requires a knowledge of the thermal stability of the chemical reactor. After a brief historical review of the classical stability problem of a batch reactor and a continuous stirred tank reactor, an overview of thermal stability in two classes of transiently operated reactors is given: the periodic reverse-flow reactor (where thermal runaway is to be avoided) and the monolithic catalytic converter/diesel particulate filter (in which thermal ignition is desired to reduce vehicle emissions while maximizing fuel economy).

CLASSICAL MODELS

Reactor designs are characterized as either homogeneous or heterogeneous. Typically, homogeneous reactors are well mixed stirred tanks (either batch or continuous), but can also be tubular reactors. They are widely used in the chemical industry from pilot plant to full-scale production. Examples include decomposition of azomethane, production of ethylene glycol, and the copolymerization of styrene and butadiene.

Thermal runaway or instability can occur when exothermic reactions are carried out in homogeneous reactors, when the rate of heat generation is faster than the rate of heat removal. In this case, a small change in an operating variable can induce a large variation in the system behavior. In practice, operation in a parametrically sensitive region is usually to be avoided. However, the reaction engineer must still be aware of these regions to ensure ''safe'' operation. Therefore, there have been many published studies on the derivation of runaway criteria and analysis of the stability of steady states, which is the focus of this section.

The Batch Reactor

Stability and dynamics of chemical reactions within batch systems date back to the pioneering work of Russian scientist Nikolay Semenov[1] on chain reactions and combustion, for which he shared the 1956 Nobel Prize with Sir Cyril Hinshelwood. There have since been considerable literature contributions, the most widely studied system being the irreversible first order reaction $A \rightarrow B$ with Arrhenius kinetics $k(T) = k_o \exp(-E/RT)$. The mass balance (in dimensionless variables) within the reactor yields:

$$\frac{dy}{d\tau_b} = (1-y)\exp\left\{\frac{\theta}{1 + \theta/\gamma}\right\} \tag{1}$$

The energy balance is:

$$Le\frac{d\theta}{d\tau_b} = B(1-y)\exp\left\{\frac{\theta}{1 + \theta/\gamma}\right\} - \alpha(\theta - \theta_c) \tag{2}$$

where $y = (C_{A,f} - C_A)/C_{A,f}$ is the reactor conversion and $\theta = (T - T_f)\gamma/T_f$ is the dimensionless reactor temperature, with $\gamma = E/RT_f$ as the dimensionless activation energy. Time is rendered dimensionless with the reaction time at the initial state $t_r = 1/k(T_f)$, such that $\tau_b = tk_o \exp(-\gamma)$. Also, $B = (-\Delta H)C_{A,f}\gamma/(\rho C_p T_f)$ is the dimensionless adiabatic temperature rise and $\alpha = hAt_r/(V\rho C_p)$ is a ratio of the reaction time to the characteristic time for heat removal from the reaction mixture t_{co}. It must be noted that the ratio B/α is often called the Semenov number. The parameter Le is a ratio of the heat capacitance of the reaction vessel and its contents to that of the vessel contents, such that $Le = 1$ implies a negligible contribution of the reactor vessel.

The most customary assumption is that of negligible reactant consumption (i.e., $y = 1$) such that Eq. (1) disappears and one is only concerned with the relationship between the heat generation curve

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007987

$B\exp(\theta/(1 + \theta/\gamma))$ and the heat removal line $\alpha(\theta - \theta_c)$. It has been shown that bifurcations or critical conditions exist when the heat generation and heat removal curves are tangent to each other, such that the following pair of equations is satisfied:

$$\exp\left(\frac{\theta}{1 + \theta/\gamma}\right) = \frac{\alpha}{B}(\theta - \theta_c) \tag{3}$$

and its derivative

$$\exp\left(\frac{\theta}{1 + \theta/\gamma}\right)\frac{1}{(1 + \theta/\gamma)^2} = \frac{\alpha}{B} \tag{4}$$

This is illustrated in Fig. 1 below (with parameters $B = 1$, $\gamma = 3$ and $\theta_c = -1.75$) for two particular heat removal lines (the upper dashed line with $\alpha = 0.968$ and the lower dashed line with $\alpha = 0.199$). Three basic patterns arise: one "safe," stable steady state (to the left of the upper dashed line), multiple steady states (in between the two dashed lines, where the lower steady state is stable and "safe," the intermediate steady state is unstable, and there may be a stable but "unsafe" steady state at high temperature) and an "unsafe" region (to the right of the lower dashed line). The upper, "unsafe" steady state will only exist when $\theta/\gamma \gg 1$ (which is usually not the case for industrially relevant reactions). In this limit, the heat generation curve levels off asymptotically to a value of $B\exp(\gamma)$. It is also noted that the classical result obtained by Semenov[1] for "safe" operation in the limit of high activation energy (i.e., $\gamma \to \infty$) when $\theta_c = 0$ is $\alpha/B > e$.

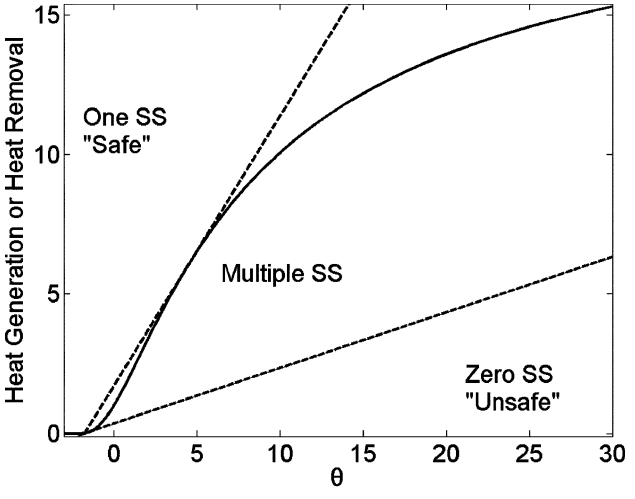


Fig. 1 Multiple steady states and their classifications in a batch reactor with negligible reactant consumption. The solid line is the heat generation curve, with $B = 1$, $\gamma = 3$, and $\theta_c = -1.75$. The upper dashed line is the heat removal line with $\alpha = 0.968$ and the lower dashed line is the heat removal line with $\alpha = 0.199$. [Data from Eq. (3).]

Since this initial work, analysis of these batch systems has been further expanded to include reactant consumption, beginning with the work of Rice, Allen, and Campbell.[2] Furthermore, an excellent study of stability with a generalized nth order reaction rate and the effect of the heat capacity of the reactor walls (when $Le \neq 1$) was presented by Balakotaiah, Kodra, and Nguyen.[3] They verified previous work which showed the boundary to runaway behavior occurs when two inflection points appear in the reaction trajectory between the initial and final states. In the limit of $\gamma \to \infty$ and $\theta_c = 0$, the "safe" criteria under adiabatic conditions ($\alpha = 0$) is given as $B < Le(1 + \sqrt{n})^2$ and for highly exothermic reactions ($B \gg 1$) with large cooling ($\alpha \gg 1$) the "safe" criteria approaches Semenov's classical result $\alpha/B > e$.

The Continuous Stirred Tank Reactor

The classical problem of steady-state multiplicity in a continuous stirred tank reactor (CSTR) was brought to popular attention in 1953 in the theoretical article by Van Heerden.[4] Large amounts of experimental work which measured these steady states were performed by the group of Schmitz beginning in 1970.[5] Schmitz also wrote two excellent reviews on multiplicity, stability, and sensitivity of steady states in chemical reactors[6] and the application of bifurcation theory to determine the presence of steady-state multiplicity in chemical reactors.[7] Even these reviews are not inclusive and it is our intention in this subsection to only provide a background to the novice in reactor design.

Consider a well-mixed stirred tank reactor with a first order, exothermic reaction A → B. A material balance about the reactor gives the following:

$$\frac{dy}{d\tau_c} = -y + Da(1 - y)\exp\left\{\frac{\theta}{1 + \theta/\gamma}\right\} \tag{5}$$

and the energy balance is given as:

$$Le\frac{d\theta}{d\tau_c} = -\theta + BDa(1 - y)\exp\left\{\frac{\theta}{1 + \theta/\gamma}\right\} - \beta(\theta - \theta_c) \tag{6}$$

where y, θ, γ, and B are defined in the previous subsection. The symbol $Da = t_c/t_r$ is a ratio of the residence time $t_c = V/F$ to the reaction time $t_r = 1/k_o\exp(-\gamma)$ at the feed conditions, and $\beta = hAt_c/(V\rho C_p)$ is a ratio of the residence time to the characteristic time t_{co} for heat removal from the reaction mixture. For the CSTR, time is rendered dimensionless as $\tau_c = t/t_c$.

The most widely occuring multiple solution in the published theoretical and experimental literature is an ignition–extinction plot with three steady states. As seen in Fig. 2, the upper and lower steady states

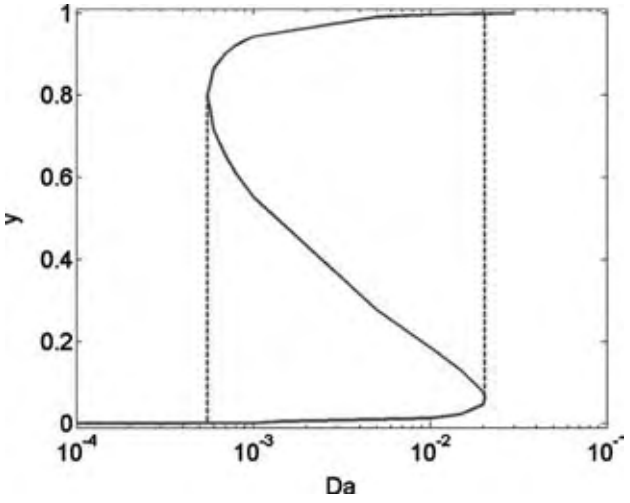


Fig. 2 Steady-state hysteresis in a CSTR with $\beta = 0$, $B = 20$, $\gamma = 20$, and $\theta_c = 0$. [Data from Eqs. (5) and (6) at steady state.]

are stable to small perturbations, but the middle steady state is unstable. Uppal, Ray, and Poore[8] analyzed the system of Eqs. (5) and (6) and found six different regions of multiplicity with nine different types of phase plots. The dynamical features include limit cycles and separatrices. They continued their analysis by redefining Da and β in terms of the residence time t_c.[9] Thus, they used $Da = Da_o/\tau_c$ and $\beta = \beta_o/\tau_c$ where $Da_o = k_o \exp(-\gamma)t$ and $\beta_o = hAt/(V\rho C_p)$ and added an additional adjustable parameter τ_c. Surprisingly, this seemingly minor change in the system led to increased bifurcation behavior, such as isolated steady states (''isolas'') and ''mushrooms.'' Beyond these studies, the complexity is seen to increase dramatically with multiple reactions and reaction orders, as well as within polymerization reactions.

Balakotaiah, Kodra, and Nguyen[3] also studied the CSTR. They found that the boundary between the insensitive and runaway region is where there are two limit points (a point of infinite slope – there are two such limit points in Fig. 2) in the reaction path connecting the initial and final states. They found identical criteria to the batch reactor in the special limit of $\gamma \to \infty$ and $\theta_c = 0$. That is, the ''safe'' criteria under adiabatic conditions ($\alpha = 0$) is given as $B < (1 + \sqrt{n})^2$. It is noted that there is no dependence on the Lewis number, Le, as there is no contribution due to the heat capacity of the reactor walls at steady state. Also, for the case of $B \gg 1$ and $\alpha \gg 1$, the ''safe'' criteria is $\alpha/B > e$.

Other Systems

Other systems also reviewed by Schmitz[5,6] include distributed reactor models such as catalyst pellets and fixed bed reactors, which also display interesting dynamical and steady-state behavior. Balakotaiah, Kodra, and Nguyen[3] also present stability criteria for homogenous tubular reactors. The equations for such a reactor are given by:

$$\frac{\partial y}{\partial \tau_b} + \frac{t_r}{t_c}\frac{\partial y}{\partial \xi} = \frac{t_r}{t_m}\frac{\partial^2 y}{\partial \xi^2} + (1 - y)\exp\left\{\frac{\theta}{1 + \theta/\gamma}\right\} \tag{7}$$

for mass conservation and

$$Le\frac{\partial \theta}{\partial \tau_b} + \frac{t_r}{t_c}\frac{\partial \theta}{\partial \xi} = \frac{t_r}{t_h}\frac{\partial^2 \theta}{\partial \xi^2} + B(1 - y)\exp\left\{\frac{\theta}{1 + \theta/\gamma}\right\} - \frac{t_r}{t_{co}}(\theta - \theta_c) \tag{8}$$

for energy conservation, where ξ is the dimensionless reactor length. It is noted that the tubular reactor has the three timescales from the CSTR: the characteristic reaction time t_r, the cooling time t_{co}, and the residence time t_c. In addition, there are two additional timescales: the mass dispersion time $t_m = L^2/D$, and the axial heat dispersion time $t_h = L^2/\kappa$. Depending upon the relative importance of these timescales, the stability behavior of the tubular reactor can be broken up into several classifications:

- A cooled plug flow reactor when axial dispersion is negligible, a usual assumption for tubular reactors with no mixing ($t_r, t_{co}, t_c \ll t_m, t_h$)
- A batch reactor when axial dispersion and convection are both negligible, as there is no outflow from the reactor ($t_r, t_{co} \ll t_c, t_m, t_h$)
- A CSTR when dispersion is very large and the conversion and temperature are independent of position due to perfect mixing ($t_m, t_h \ll t_c, t_r, t_{co}$).

In addition, it was found that the plug flow reactor had the same asymptotic behavior as the batch reactor, with reactor heat capacity unimportant ($Le \neq 1$) and the only difference being the interchange of time (batch) with reactor length (plug flow). The authors also stated that the uncertainty involved in determining kinetic and heat transfer parameters is greater than the differences between the respected criteria for batch, CSTR, and plug flow, and that one single thermal stability criterion based upon the CSTR would be more than adequate for all homogeneous reactor models. This paper[3] also described runaway limits for catalytic reactors and for multiple reactions.

Doraiswamy[10] presented a very detailed review on the role of heterogeneous catalytic systems on inducing multiplicity and instability. In particular, this review focuses on the role that the Langmuir isotherm and the Langmuir–Hinshelwood rate equations (resulting

from adsorption, surface reaction, and desorption) have on inducing instabilities in chemical reactors. Additional complexities are introduced through accounting for multilayer adsorption and lateral interactions between adsorbed molecules, and nonideal surfaces.

CURRENT APPLICATIONS OF THERMAL STABILITY

This section focuses on some current applications of thermal stability theory.

Reverse-Flow Reactors

The reverse-flow chemical reactor (RFR) has been shown to be a potentially effective technique for many industrial chemical processes, including oxidation of volatile organic compounds such as propane, propylene, and carbon monoxide; removal of nitrogen oxides; sulfur dioxide oxidation or reduction; production of synthesis gas; methanol formation; and ethylbenzene dehydration into styrene. An excellent introductory article in the topic is given by Eigenberger and Nieken[11] on the effect of the kinetic reaction parameters, reactor size, and operating parameters on RFR performance. A detailed review that summarizes the applications and theory of RFR operation is given by Matros and Bunimovich.[12]

A reverse-flow reactor operates by periodically opening and closing valves so that the gas flow through the RFR bed alternates back and forth. After a long startup time, the RFR can reach a cyclic steady state, where the temperature profile is identical to that during the previous full cycle. This behavior shows that heat storage during lean conditions is a strong advantage of RFR operation versus a conventional packed bed (single-pass) reactor, as seen in the following example.

Consider the comparison between an adiabatic RFR and a single-pass reactor to treat a feed with carbon monoxide (with data available in Table 1 and performance depicted in Fig. 3). Invoking the commonly cited psuedohomogeneous model, each reactor is described by Eqs. (7) and (8). The reactor is preheated to an ignited steady state by introducing CO at a large concentration $C_{in,COseed} = 100C_{in,CO}$ for $t < t_f$ (3.33 h). For times $t > t_f$, the flow is reversed with a switching time t_{rev} and a feed concentration $C_{in,CO}$ (such a stream with a low concentration may be an effluent from a chemical process elsewhere in the plant, and may need to be treated before exiting the plant). The reactors appear to behave similarly for times $t_f < t < t_f + t_{rev}$ even though the gases flow in different directions.

It is noted that the higher maximum temperature within the RFR is due to wrong-way behavior, a purely transient effect which was shown for a single-pass reactor by Pinjala, Chen, and Luss[13] in reverse-flow reactors by many investigators including Purwono[14] and Keith, Leighton, and Chang.[15] Upon flow reversal, the temperature in the feed stream is lower than that near the entrance of the reactor. Thus, a temperature step propagates into the reactor. Within this step there is some reactant that has not been depleted. Because the reaction rate is higher at the elevated temperature, there is a localized heat release at the step in the form of a thermal spike and the maximum reactor temperature increases.

It is also noted, for this example, that only after a time $t_f + t_{rev}$ do the reactors diverge in performance. In the RFR, the flow has again switched, and about one-third of the reactor is ignited for the remainder of the simulation and is able to treat the feed gas

Table 1 Data for carbon monoxide oxidation in a reverse-flow reactor

T_{in}	400 K
t_f	3.3 hr
$C_{in,CO}$	5.0×10^{-9} mol/cm^3
$C_{in,other}$	Trace
$\Delta T_{ad,COmax} = \Delta H C_{in,COmax}/(\rho_g C_{p,g})$	2.2 K
t_{rev}	2 hr

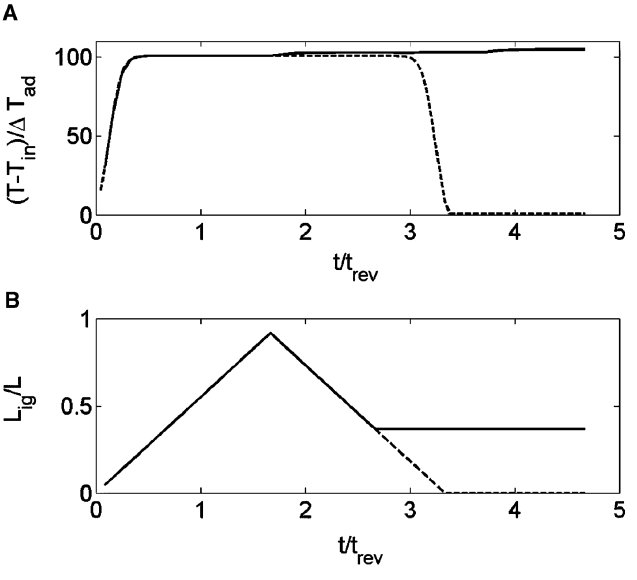


Fig. 3 Comparison of RFR (solid lines) with single-pass reactor (dashed lines). (A) the maximum temperature within the reactor and (B) the ignited length of the reactor. (Data taken from the numerical model in Ref.[15].)

adequately. The single pass reactor finally extinguishes after a time $2\,t_f$ when the maximum temperature within the reactor is equal to the feed temperature. Under these conditions, the feed gas may not be treatable and can exit the reactor unabated.

For systems where the adiabatic temperature rise is low (as is the case considered here) the thermal spikes introduced by the flow reversals do not dramatically affect the reactor performance. However, the concentration of feed streams to such treatment reactors can fluctuate to a high level which can result in a high temperature thermal spike developing within the reactor. Pinjala, Chen, and Luss[13] characterized this dynamic response and showed that reactor runaway could occur within the single-pass reactor. Their work is directly applicable to the RFR as the forced oscillations in the gas flow direction can result in a thermal spike formation at the beginning of each half cycle. Thus, there is a need to understand thermal stability within these systems. Further complicating the matter is the fact that the temperature spikes are very narrow and are thus difficult to detect using thermocouples or other sensors imbedded within the reactor.

Eigenberger and Nieken[11] demonstrated ignition–extinction hysteresis within the RFR. They illustrated that the maximum bed temperature extinguishes to a low, unignited cyclic steady state for high values of t_{rev}. One must reduce the value of t_{rev} below this extinction point to reignite the reactor to the upper cyclic steady state.

Often the ignited steady-state temperature can be high enough to damage the catalyst or lead to ''unsafe'' reactor operation. Purwono et al.[14] performed experiments and developed a numerical simulation of a two recuperator/catalyst bed reverse-flow reactor system. They measured runaway for propene oxidation and for carbon monoxide oxidation. They identified the reactor peak temperature as the primary indicator of reactor runaway. Keith, Leighton, and Cheng[15] stabilized the reactor by increasing the thermal dispersion by a factor of 200 by placing metal rods into the bed, parallel with the gas flow. The thermal dispersion that arises is a transient phenomenon, akin to the classical Taylor–Aris dispersion mechanism.[16,17] It smoothes out the temperature spikes after they form and allows for a lower temperature cyclic steady state within the reactor. This work also developed analytical expressions for the reactor maximum temperature as a function of cycle number which can be used to predict the time for runaway (if extremely high pollutant feed conditions exist) or extinction (if there is a very low concentration of pollutant in the feed stream).

Khinast and Luss presented a rigorous method for constructing parameter maps of the RFR which contain different bifurcation diagrams.[18] Using the Damköhler number as the bifurcation variable, five different regions were found for a single exothermic first-order reaction (these contained hysteresis, isola, and mushroom varieties). In addition, using the adiabatic temperature rise as the bifurcation variable, seven regions are found when there are two independent exothermic first-order reactions (these contained hysteresis and double limit varieties and five of these regions had three stable periodic states). Khinast, Gurumoorthy, and Luss also identified regimes where cooled RFR exhibit complex quasi-periodic states.[19] The maximum temperature within the reactor in many of these states was high which could possibly lead to catalyst deactivation or ''unsafe'' operation.

Catalytic Converters

Monolithic catalytic converters continue to receive attention in the literature because of their applications in air pollution control and clean energy production. They differ from packed-bed reactors in their configuration as there are many parallel channels coated with a layer of catalyst. The flow in the channels is typically laminar. Because of its large void fraction, it is expected that the temperature transients will exhibit a significant impact on the performance of the monolith, particularly with respect to thermal stability.

The light-off behavior of catalytic monolith is characterized by the presence of multiple steady states with an S-shaped hysteresis locus, which is similar to that of the CSTR. The catalytic monolith initially operates on the low conversion steady-state branch at the monolith entrance. As the fluid flows through the channel, the energy produced by the exothermic reaction is carried downstream in the channel. Beyond a point where the rate of heat generation is equal to the rate of heat removal, the catalytic monolith jumps to the high conversion branch. However, this ignition point may be beyond the finite length of the monolith for low inlet temperatures, which means no ignition occurs. With an increase in the inlet temperature, the ignition point should move into the reactor and then towards the monolith entrance.

Much effort has been made to study this light-off behavior of catalytic monolith. Oh and Cavendish[20] studied the response of the monolith to a step increase in the feed stream temperature by using a one-dimensional two-phase (gas and solid) model. They tracked the cross-sectional average temperature and concentration in each phase and used heat and mass transfer coefficients to describe interphase transport. The results indicated that the light-off occurs at the monolith entrance for a sufficiently high inlet exhaust temperature. For a lower inlet exhaust temperature, the light-off occurs in the downstream section, and the

thermal front moves upstream due to solid phase heat conduction. They also pointed out wrong-way behavior in catalytic monoliths, as described in the previous subsection and reported by Pinjala, Chen, and Luss.[13] Under certain operating conditions, a decrease in the inlet exhaust temperature can result in a downstream ignition within the monolith.

It should be emphasized that Oh and Cavendish[20] assumed that the reactions only occur on the surface of the channel wall. This assumption is less realistic for a layer of washcoat (typically γ-alumina) dispersed with catalyst applied on to the wall surface. Ramanathan, Balakotaiah, and West[21] showed that the diffusion in the washcoat has a profound influence on the light-off behavior of a monolith converter. They derived an analytical light-off criterion based on a one-dimensional two-phase model with position-dependent heat and mass transfer coefficients. The derivation of this criterion is based on the two key assumptions: a positive exponential approximation (i.e., the Frank–Kameneskii approximation) and negligible reactant consumption in the fluid phase. The light-off is defined as the occurrence of multiple steady states with the attainment of the ignited steady state. Here, we discuss only the results of their analysis, without going into the details of their derivation.

For a first-order reaction, their light-off criterion (deemed ''unsafe'' by the classical definition, but desired in this application) accounting for washcoat diffusion is given by

$$m\left(\phi_s\sqrt{\Lambda}\right)\left[\frac{n(Pe_h)B\phi_s^2}{eP} + \frac{4B\phi_s^2}{Le_f Nu_{H,\infty}}\right]^{-1} < e \qquad (9)$$

where m is a function of the product of the Thiele modulus ϕ_s and the square root of the dimensionless washcoat thickness Λ that accounts for the washcoat diffusion; n is a function of the axial heat Peclet number, Pe_h, that accounts for solid conduction; B is the dimensionless adiabatic temperature rise, P is the transverse Peclet number (a ratio of transverse diffusion time to the convection time), Le_f is the fluid Lewis number, a ratio of heat to mass diffusivities; and $Nu_{H,\infty}$ is the asymptotic value of the Nusselt number with a constant flux boundary condition.

The first term in the sum within Eq. (9) represents the ignition locus of the homogenous plug flow model. It can be seen that, in the case of infinite conductivity in the solid phase and negligible washcoat thickness (i.e., $n \to e$ as $Pe_h \to 0$ and $m \to 1$ as $\Lambda \to 0$), the above lightoff (''unsafe'') criterion reduces to the classic Semenov criterion $P/B\phi_s^2 < e$. It can also be seen from the above criterion that for a small value of P(i.e., long monolith or a low flow rate), the criterion for the homogenous plug flow model can be used. This principle has been used to derive the light-off criterion for monolithic diesel particulate filters in the case of low exhaust flow, which we will discuss in the next subsection.

The second term in the sum within Eq. (9) accounts for the heterogeneous contribution (i.e. washcoat diffusion). Ramanathan, Balakotaiah, and West[21] pointed out that ignition will occur at the back end when the first term is greater than one and the second term is much less than one. Furthermore, if the second term is greater than one, the ignition occurs at the leading edge independent of the value of the first term. If both terms are less than one but the sum is greater than one, the ignition will occur in the middle of the converter. It is worth noting that, for a given reaction kinetics and a fixed washcoat thickness, the dynamic behavior of the monolith can be classified into three regions: no ignition, ignition without washcoat diffusion limitations, and ignition with washcoat diffusion limitations. Their studies showed that the influence of the washcoat thickness on ignition becomes constant in the washcoat diffusion limitation region. In the region without washcoat diffusion limitations, ignition becomes easier as the washcoat thickness increases.

Catalytic monoliths have been extensively used in the control of automobile emissions. A large quantity of emissions are produced in the cold-start period. Thus, it is of practical interest to design a quick light-off catalytic monolith converter. Leighton and Chang[22] have shown that the Taylor–Aris dispersion mechanism[16,17] has a profound influence on the light-off time of the catalytic converter. They found that the ignition time depends on two key parameters: a Damköhler parameter χ, which is a ratio of axial dispersion to axial convection during the characteristic time for ignition, and the degree of monolith subcooling η. For a small value of χ (i.e., slow reaction), the Taylor–Aris dispersion smoothes the thermal front and delays the ignition process, and the ignition occurs downstream near the thermal front. For a large value of χ (i.e., fast reaction) the ignition occurs before the Taylor–Aris dispersion takes effect. Thus, the thermal dynamics of the leading edge of the monolith can be assumed to be homogenous, and ignition occurs at the leading edge of the monolith. The analytical expressions for the light-off time for these two distinct mechanisms have been derived based on Oh and Cavendish's[20] one-dimensional two-phase model. By use of these analytical expressions, it is found that preheating the exhaust flow is an effective way to reduce the light-off time for leading edge ignition. The leading edge ignition is preferred because the upstream propagation of the thermal front is slow when ignition occurs downstream. The ignition dynamics for downstream and leading edge ignition are shown in Fig. 4.[23]

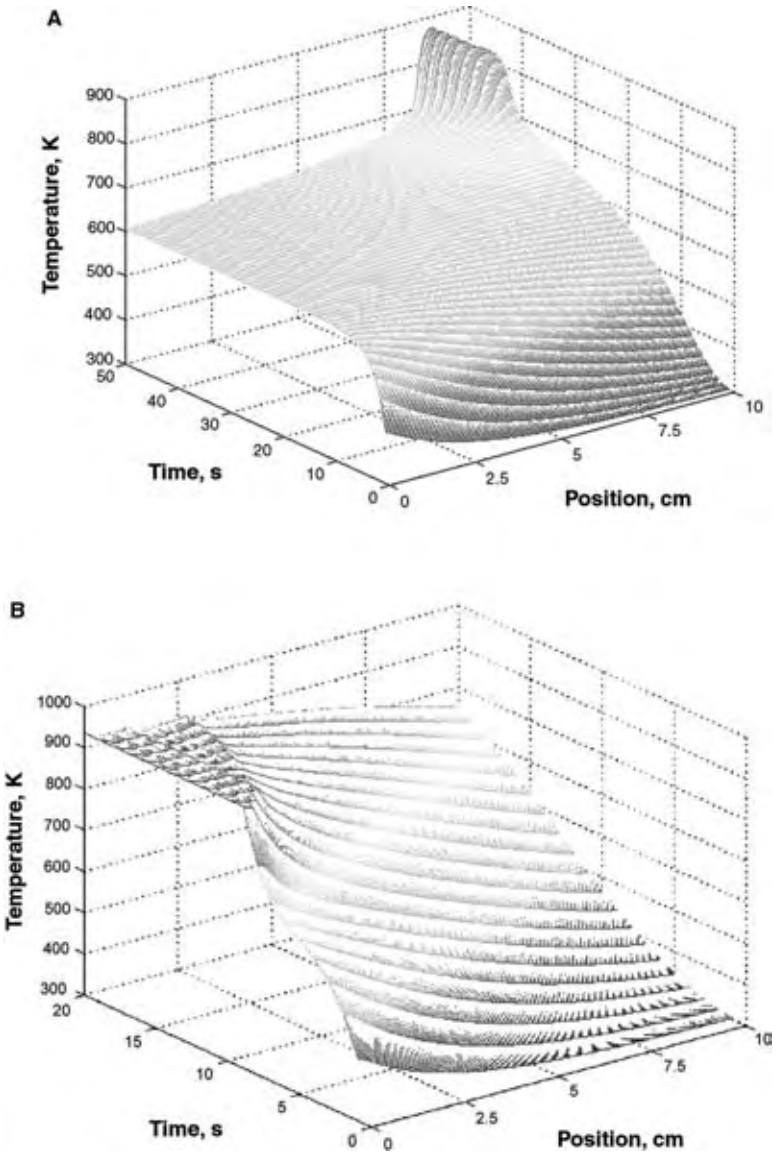


Fig. 4 Schematic illustrating downstream (A) and leading-edge (B) ignition within a catalytic converter. (Data taken from the numerical model in Ref.[23].)

Diesel Particulate Filters

Monolithic diesel particulate filters (DPF) are widely used in diesel particulate emission control. They consist of many parallel channels which are alternately plugged at either end in order to force the exhaust gases through the porous ceramic walls. The particulates are deposited on the inside wall of the inlet channel to form a thin, porous soot bed. Once a sufficient mass of particulates is collected, it is burned off to regenerate the filter. Thus, in order to achieve a successful regeneration, the DPF should operate in the thermal runaway region.

Many theoretical and experimental efforts have been devoted to study thermal runaway in the DPF during regeneration. The classic transient pseudohomogeneous model by Bissett and Shadman[24] is based on the assumption of a uniform flow distribution through the whole monolith channel. Although this assumption is only applicable in the case of high exhaust gas flow rates, this model successfully predicts the regeneration trends observed in experiments. McCabe and Sinkevitch[25] studied the ignition of soot beds by using the RVVR (reactor to visualize and video regeneration) system. They pointed out that there exists a critical value of the soot temperature beyond which thermal runaway occurs. Furthermore, Hayashi et al.[26] has shown that complete regeneration cannot be achieved without ignition, and that the efficiency of regeneration can be strongly affected by the inlet gas temperature, exhaust gas flow rate, and oxygen concentration.

Recently, Zheng and Keith[27] derived an analytical expression for the ignition time in a DPF and also showed that there exists a critical inlet exhaust gas temperature for thermal ignition to occur. Interest in

practical applications is usually in determining the critical value of the inlet gas temperature. Zheng and Keith[28] were the first to publish an explicit ignition criterion to predict the critical inlet gas temperature for the DPF. They simplified the model of Bissett and Shadman[24] by assuming a constant oxygen concentration through the soot layer. Although this assumption is violated after ignition, it is usually satisfied before ignition over most operating conditions and therefore introduces little error in the derivation of ignition criteria. With this assumption in mind, and using the Frank–Kamenetskii approximation for the combustion kinetics $\exp(-E/RT) \approx \exp(-E/RT_f)\exp(E(T - T_f)/RT_f^2)$(it is noted that this approximation is equivalent to the large γ limit) such that $\exp(-E/RT) \approx \exp(-\gamma)\exp(\theta)$, the reactor dynamics in the DPF system are described by the energy conservation equation

$$\frac{d\theta}{d\tau_d} = \frac{B_d Z e^{\theta} - \delta\theta}{1 + Le_d Z} \tag{10}$$

and the mass conservation equation

$$\frac{dZ}{d\tau_d} = -Ze^{\theta} \tag{11}$$

Since in the DPF the gas flows through a solid bed, with a non-Arrhenius reaction rate given by $k = k_o T \exp(-E/RT)$, we redefine, for the DPF, the following parameters: the characteristic reaction time $t_{r,d}$ (such that $\tau_d = t/t_{r,d}$), the characteristic time for thermal convection $t_{c,d}$ (such that the cooling parameter $\delta = t_{r,d}/t_{c,d}$), the dimensionless adiabatic temperature rise B_d, and the Lewis number, Le_d, as the ratio of the total heat capacity of the soot bed to that of the substrate wall. $Z = w/w_b$ is the dimensionless soot layer thickness, and θ is the dimensionless temperature, defined earlier.

The simplified model of the DPF presented in Eqs. (10) and (11) is similar to the classic batch reactor model. Accordingly, the ignition (''unsafe'') criterion for thermal runaway in a batch reactor can be directly used for the DPF system, such that

$$\frac{\delta}{B_d} < e \tag{12}$$

It is well known that Semenov theory is a good approximation for a highly exothermic reaction systems, i.e., $B_d \to \infty$. Zheng and Keith[28] have shown that Eq. (12) is too conservative when $B_d \ll 10$, which may occur in DPF systems. Thus, the effect of reactant consumption on ignition criterion needs to be considered.

Rice, Allen, and Campbell[2] first studied the effect of reactant consumption on thermal runaway. They defined the critical point as that which possesses a single inflection point before the maximum temperature in the temperature–time plane. Thus, a positive second-order derivative of temperature with respect to time before the maximum temperature is reached implies that the temperature increase with time is accelerated. This is necessary for thermal runaway to occur. Many other mathematical definitions of the critical point exist. The most widely used criteria was formulated in the $\theta - z$ plane by Adler and Enig:[29]

$$\frac{d^2\theta}{dZ^2} = 0, \quad \frac{d^3\theta}{dZ^3} = 0 \tag{13}$$

Zheng and Keith[28] applied this definition to derive an explicit ignition (''unsafe'') criterion accounting for the reactant consumption

$$\frac{\delta}{B_d} + f(Le_d)B_d^{-g(Le_d)} < e \tag{14}$$

where for most cases in DPF applications $f(Le_d) = 6.0$ and $g(Le_d) = 0.64$. This ignition criterion has a very good agreement with simulations of Bissett and Shadman's model in the prediction of the critical inlet gas temperature, as shown in Fig. 5. It is noted that Eq. (14) collapses into Eq. (12) in the limit as $B_d \to \infty$. Thus, the second term in Eq. (14) is due to reactant consumption.

Zheng and Keith[28] also studied the effect of the parameter Le_d on the critical value of the system temperature. They pointed that the critical temperature

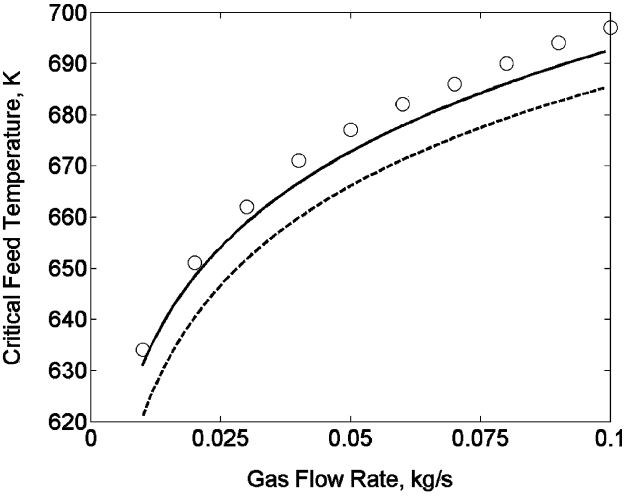


Fig. 5 The critical inlet gas temperature as a function of the gas flow rate. Simulation data, plotted as circles, the solid line is Eq. (14), and the dashed line is Eq. (12). (Data taken from the numerical model in Ref.[24].)

decreases as Le_d increases, which is reasonable since a large value of Le_d indicates a large amount of diesel particulate soot. This work also studied the influence of the various design variables and operating conditions on ignition, and concluded that a thin channel wall, a high total filtration area, a high initial loading, a low gas flow rate, and a high oxygen concentration are desirable for regeneration in the DPF. Combined with the explicit prediction of ignition time reported by Zheng and Keith,[27] the ignition criteria can be used to optimize regeneration strategies in the DPF system.

The Future of Thermal Stability Analysis

The discussion on stability analysis within DPF systems has been limited to the case of a high gas flow rate where axial thermal gradients within the DPF are negligible. This may not always be the case in practical applications. Thus, the location of the ignition within the DPF can strongly affect its overall performance. A leading-edge ignition is desired because it will lead to complete regeneration of the entire filter by convective gas flow. A downstream ignition may lead to flow channeling within the DPF, where the gas bypasses the area at the front of the filter where the particulate loading is high (as seen in Fig. 6). However, particulate may continue to accumulate at the leading edge of the DPF. Upon the next regeneration, a large heat release may crack or melt the DPF substrate and render it ineffective. The development of stability criteria that account for spatial and temporal variations in temperature within the DPF is expected to follow the work performed on automobile catalytic converters, particularly the influence of thermal dispersion. Through numerical and analytical modeling efforts, specific criteria will be formulated to ensure ''safe,'' long-term operation of the DPF.

In addition, with the growth in computational power it is easier to perform complex simulations and bifurcation analyses of multiphase reaction systems. Building upon prior experience with such techniques,[18,19] Wang et al.[30] recently measured multiplicity in multiphase bubble columns, finding four limit points in a plot of maximum temperature as a function of the Damköhler number. These results shows that three stable states exist, which they classify as cold, warm, and hot. The first two steady states are akin to the ignition–extinction behavior illustrated in the CSTR. The final steady state is due to multiplicity in the enhancement factor. Additional studies on detailed behavior such as this may help reduce byproduct formation and lead to more efficient industrial chemical processing.

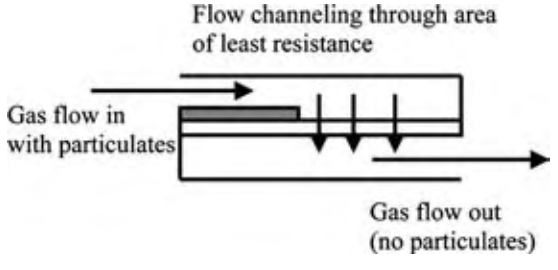


Fig. 6 Cartoon illustrating flow channeling in a DPF after ignition via the downstream mechanism. For proper performance, DPF filters should always be ignited at the leading edge.

CONCLUSIONS

Thermal stability of chemical reactors is a classic yet active area within chemical engineering science. Considerable research has focused on determining ''safe'' operating criteria for batch, CSTR, and tubular reactors. Current work has been directed towards understanding thermal stability in the presence of multiple phases (fluid/solid and gas/liquid) and multiple reactions with realistic, complex reaction rates expressions. The advent of computational methods has allowed for this field to continue to thrive. A sound understanding of these principles may help improve industrial reactor performance by reducing waste and costly separation operations and help maintain a clean environment.

ACKNOWLEDGMENT

Discussions with Professor Dan Crowl at Michigan Technological University were greatly appreciated in the preparation of this article.

NOTATION

A	Surface area for cooling by heat exchange
B	Dimensionless adiabatic temperature rise
C_A	Concentration
C_p	Heat capacity
D	Mass diffusivity
Da	Damköhler number, Da $= t_c/t_r$
DPF	Diesel particulate filter
E	Activation energy
F	Flow rate
h	Heat transfer coefficient
ΔH	Heat of reaction
k	Reaction rate constant
k_o	Pre-exponential factor
L	Reactor length

Le	Lewis number, a thermal capacitance ratio
Le_f	Fluid Lewis number, ratio of heat to mass diffusivity
$Nu_{H,\infty}$	Asymptotic value for the Nusselt number for heat transfer with a constant flux boundary condition
P	Transverse Peclet number
Pe_h	Axial heat Peclet number
R	Gas constant
RFR	Reverse-flow reactor
T	Temperature
ΔT_{ad}	Adiabatic temperature rise
t	time
t_c	Residence time
t_{co}	Cooling time
t_f	Thermal front time
t_h	Thermal dispersion time
t_m	Mass dispersion time
t_r	Reaction time
t_{rev}	Reversal time
V	Reactor volume
w	Soot layer thickness
w_b	Soot layer thickness prior to regeneration
y	Conversion
Z	Dimensionless soot layer thickness

Subscripts

c	Coolant
d	Values for DPF
f	Initial (batch) or feed (CSTR) value
ig	Ignited portion
in	Feed value
seed	Seed concentration used to ignite a RFR

Greek

α	Ratio of reaction time to cooling time
β	Ratio of residence time to cooling time
γ	Dimensionless activation energy
δ	Cooling parameter for the DPF
ϕ_s	Thiele modulus
η	Degree of monolith subcooling
Λ	Dimensionless washcoat thickness
κ	Axial thermal diffusivity
ρ	Density
θ	Dimensionless temperature
τ_b	Dimensionless time, $\tau_b = t/t_r$
τ_c	Dimensionless time, $\tau_c = t/t_c$
χ	Damköhler parameter for the catalytic converter
ξ	Dimensionless reactor length

REFERENCES

1. Semenov, N.N. Zur theorie des verbrennungsprozesses. Zhurnal Fizich. Khim. **1928**, *48*, 571–581.
2. Rice, O.K.; Allen, A.O.; Campbell, H.C. The induction period in gaseous thermal explosions. J. Am. Chem. Soc. **1935**, *57*, 2212–2222.
3. Balakotaiah, V.; Kodra, D.; Nguyen, D. Runaway limits for homogeneous and catalytic reactors. Chem. Eng. Sci. **1995**, *50*, 1149–1171.
4. Van Heerden, C. Autothermic processes. Ind. Eng. Chem. **1953**, *45*, 1242–1247.
5. Vejtasa, S.A.; Schmitz, R.A. An experimental study of steady state multiplicity and stability in an adiabatic stirred reactor. AIChE J. **1970**, *16*, 410–419.
6. Schmitz, R.A. Multiplicity, stability, and sensitivity of states in chemically reacting systems – a review. Adv. Chem. Ser. **1975**, *148*, 156–211.
7. Razon, L.F.; Schmitz, R.-A. Multiplicities and instabilities in chemically reacting systems – a review. Chem. Eng. Sci. **1987**, *42*, 1005–1047.
8. Uppal, A.; Ray, W.H.; Poore, A.B. On the dynamic behavior of continuous stirred tank reactors. Chem. Eng. Sci. **1974**, *29*, 967–985.
9. Uppal, A.; Ray, W.H.; Poore, A.B. The classification of the dynamic behavior of continuous stirred tank reactors – influence of reactor residence time. Chem. Eng. Sci. **1976**, *31*, 205–214.
10. Doraiswamy, L.K. Catalytic reactions and reactors – a surface science approach. Prog. Surf. Sci. **1991**, *37*, 1–277.
11. Eigenberger, G.; Nieken, U. Catalytic combustion with periodic flow reversal. Chem. Eng. Sci. **1988**, *43*, 2109–2115.
12. Matros, Y.S.; Bunimovich, G.A. Reverse-flow operation in fixed bed catalytic reactors. Catalys. Rev. Sci. Eng. **1996**, *38*, 1–68.
13. Pinjala, V.; Chen, Y.C.; Luss, D. Wrong-way behavior of packed-bed reactors: II. Impact of thermal dispersion. AIChE J. **1988**, *34*, 1663–1672.
14. Purwono, S.; Budman, H.; Hudgins, R.R.; Silveston, P.L.; Matros, Y.S. Runway in packed bed reactors operating with periodic flow reversal. Chem. Eng. Sci. **1994**, *49*, 5473–5487.
15. Keith, J.M.; Leighton, D.T.; Chang, H.-C. A new design of reverse-flow reactors with enhanced thermal dispersion. Ind. Eng. Chem. Res. **1999**, *38*, 667–682.
16. Taylor, G.I. Dispersion of soluble matter in solvent flowing slowly through a tube. Proc. Roy. Soc. Lond. **1953**, *A219*, 186–203.
17. Aris, R. On the dispersion of a solute in a fluid flowing through a tube. Proc. Roy. Soc. Lond. **1956**, *A235*, 67–77.
18. Khinast, J.; Luss, D. Mapping regions with different bifurcation diagrams of a reverse-flow reactor. AIChE J. **1997**, *43*, 2034–2047.

19. Khinast, J.; Gurumoorthy, A.; Luss, D. Complex dynamic features of a cooled reverse-flow reactor. AIChE J. **1998**, *44*, 1128–1140.
20. Oh, S.H.; Cavendish, J.C. Transients of monolithic catalytic converters: responses to step changes in feed stream temperature as related to controlling automobile emissions. Ind. Eng. Chem. Prod. Res. Dev. **1982**, *21*, 29–37.
21. Ramanathan, K.; Balakotaiah, V.; West, D. Light-off criterion and transient analysis of catalytic monoliths. Chem. Eng. Sci. **2003**, *58*, 1381–1405.
22. Leighton, D.T.; Chang, H.-C. A theory for fast-igniting catalytic converters. AIChE J. **1995**, *41*, 1898–1914.
23. Keith, J.M.; Chang, H.-C.; Leighton, D.T. Designing a fast-igniting catalytic converter system. AIChE J. **2001**, *47*, 650–663.
24. Bissett, E.J.; Shadman, F. Thermal regeneration of diesel particulate monolithic filters. AIChE J. **1985**, *31*, 753–758.
25. McCabe, R.W.; Sinkevitch, R.M. A laboratory combustion study of diesel particulates containing metal additives. SAE paper 860011, **1986**.
26. Hayashi, K.; Ogura, Y.; Kobashi, K.; Sami, H.; Fukami, A. Regeneration capability of wall flow monolith diesel particulate filter with electric heater. SAE paper 900603, **1990**.
27. Zheng, H.; Keith, J.M. A new design for efficient diesel particulate trap regeneration. AIChE J. **2004**, *50*, 184–191.
28. Zheng, H.; Keith, J.M. Ignition analysis of wall-flow monolith diesel particulate filters. Catalysis Today **2004**, *98*, 403–412.
29. Adler, J.; Enig, J.W. The critical conditions in thermal explosion theory with reactant consumption. Combust. Flame, **1964**, *8*, 97–103.
30. Wang, J.; Bindal, A.; Leib, T.M.; Khinast, J.G. Analysis of the complex nonlinear behavior of reacting bubble flows: steady-state multiplicity. Chem. Eng. Sci. **2004**, *59*, 5575–5585.

Thermogravimetric Analysis

John O. Hill
La Trobe University, Melbourne, Victoria, Australia

INTRODUCTION

Thermogravimetric analysis (TGA or TG) is considered to be one of the five basic thermal analysis techniques. It involves the measurement of change of sample mass with change of temperature. In TGA, mass loss is observed if a thermal event involves loss of a volatile component. Chemical reactions, such as combustion, involve mass losses, whereas physical changes, such as melting, do not. The latter may be studied by differential thermal analysis (DTA) or differential scanning calorimetry (DSC), both of which measure the variation of heat flux in a sample with variation of temperature. Both physical and chemical changes in a sample can be measured simultaneously by using simultaneous thermal analysis (STA or TGA–DTA), which is a combination of TGA and DTA or, less commonly, DSC (TGA–DSC). Simultaneous thermal analysis involves measurement of both mass loss and heat flux of a sample simultaneously with variation of temperature. However, thermal analysis techniques cannot be used in general for the identification or characterization of a sample. A more detailed analysis of the thermal behavior of a sample is possible if the volatile components released from a thermal decomposition can be identified. While TGA provides mass loss data related to such volatiles, DTA gives the temperature range over which such volatiles are released; but neither data set unambiguously identifies such volatiles. This can be achieved by evolved gas analysis (EGA), which is a combination of TGA–DTA system and a gas analysis system such as an infrared spectrophotometer (TGA–DTA–FTIR), a gas chromatograph, or a mass spectrometer (TGA–DTA–MS). Recently, TGA has become refined and considerably augmented as the principal component of sample controlled thermal analysis (SCTA). In SCTA, the heating rate is varied so as to produce a constant rate of mass loss. Thus, the rate of mass loss, rather than its magnitude, controls the overall analysis process. It essentially overcomes the problems associated with linear heating rates in conventional thermal analysis techniques such as TGA, which can cause significant temperature and pressure gradients in the sample. Judicious use of heating regimes, which are modified in some way by the reaction rate, can greatly reduce these problems, giving enhanced resolution in analytical and characterization investigations.

The aim of this entry is to discuss the primary thermal analysis technique, TGA, in terms of the associated instrumentation, calibration procedures, and application range. Related techniques, such as TGA–DTA, EGA, SCTA, and modulated temperature TGA, are also discussed, as these reflect the significant developments of the TGA family of techniques that have occurred over the last decade.

THERMOGRAVIMETRIC ANALYSIS

Thermogravimetric analysis involves the measurement of change of sample mass with change of temperature using a ''thermobalance.'' A thermobalance is a combination of a suitable electronic microbalance with a furnace, a temperature programmer, and an online computer for overall experimental control and data collection and processing. The system allows for the sample to be simultaneously weighed and programmed to be heated or cooled, and the mass, time, and temperature data to be recorded and processed. The balance should be in a suitably enclosed system so that the nature and pressure of the atmosphere surrounding the sample can be controlled. The microbalance is shielded from the furnace environment and is maintained at ambient temperature throughout the analysis.

In TGA, mass loss is observed if a thermal event involves loss of a volatile component. Chemical changes in the sample, such as decomposition, combustion, and dehydration, involve mass loss (or mass gain) and may be thus studied by TGA. The data are presented as a plot of mass, m, vs. temperature, T, or time, t. The curve shape is sigmoid, indicating that most of the mass loss occurs at a particular temperature or over a narrow temperature range where the curve slope is greatest. However, the thermal event commences before the region of greatest mass loss and continues thereafter. A more convenient representation of the character of the overall thermal event is to compute the derivative of the original TGA curve as dm/dt, or rate of mass loss and to plot this vs. temperature, T, or time, t. Thereby, the DTG curve is obtained and the spread of the thermal event over a broad temperature range appears as a peak. The DTG curve is of particular assistance if two or more thermal events overlap within similar temperature ranges. Double peaks or

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007988

a shoulder on a main peak appears in such cases. Slow processes, with other fast processes superimposed, appear as gradient changes in the DTG curve. Thus, the DTG curve can kinetically characterize overlapping thermal events. Further, the DTG peak area is directly proportional to the corresponding mass loss, so relative mass losses may be compared. The peaks in a DTG curve can act as ''fingerprints'' of components in a mixture by comparison with DTG curves of the suspected individual components.

Thermal analysis systems require calibration prior to routine use. In TGA, calibration for mass is carried out by calibrating the microbalance using a set of standard weights, as for any balance system. Temperature calibration is effected by measuring the Curie point temperatures of a suite of International Confederation for Thermal Analysis and Calorimetry (ICTAC) Certified Reference Materials, which have well-defined Curie points.[1]

It is a common practice to check the performance of a TGA system by running a sample of calcium oxalate monohydrate. This salt is known to thermally decompose in three stages over well-defined temperature ranges. The first step involves the loss of a single water of hydration molecule followed sequentially by the conversion of anhydrous calcium oxalate to calcium carbonate with the loss of carbon monoxide and thence the decomposition of calcium carbonate to calcium oxide with the evolution of carbon dioxide.

The applications of TGA are extensive and diverse and include oxy-salt decompositions, natural and synthetic polymer characterization, metal oxidation and corrosion analysis, compositional analysis of coals, polymers, and rubbers, study of glass materials, foodstuffs, catalytic materials, biological materials, and a wide range of chemical processing phenomena. It has been used very successfully to study the kinetics of chemical processes;[2,3] however, there is much controversy surrounding this application, particularly in terms of relating TGA data to reaction kinetics models.

Thermogravimetric analysis has been used extensively to determine the thermal stability of materials. It provides useful information on the storage of explosives and on the shelf-life of drugs, foodstuffs, and biological materials such as tobacco and grain. By using an atmosphere of air or oxygen, the conditions under which processes such as metal oxidation, metal fatigue, spontaneous combustion of pyrotechnics, and degradation of synthetic polymers can be determined.

Thermogravimetric analysis has been widely used to study the thermal decomposition of oxy-salts, such as metal oxalates and metal sulfates. Dollimore, Griffiths and Nicholson[4] have reported TGA data for a wide range of metal oxalates. In an atmosphere of air, these all decompose in three stages, similar to the thermal decomposition of calcium oxalate monohydrate. However, the decomposition temperature ranges differ widely in accordance with the nature of the metal involved and the wide variation in thermal stability of the intermediate metal carbonates. Aluminium oxalate is anomalous in that its nonstoichiometric water of hydration is retained in the final residue, which corresponds to ''hydrated alumina.''

Other oxy-salts loose water of hydration in a stepwise sequence when heated. Copper sulfate pentahydrate loses water of hydration in three steps with the trihydrate as the intermediate.[5] Such studies are of importance for differentiating between coordinated water and water of hydration.

Because the effective operating temperature range for TGA is subambient to 1500°C and higher with special instrumental adaptations, the technique is of particular significance in materials science and in chemical processing. It has been widely used to study the oxidation characteristics of metals. Metal oxidation involves a mass gain instead of a mass loss and the process is controlled by transport across a layer of oxide formation. Metals other than gold spontaneously convert into the corresponding oxide when heated in dry oxygen, because of the metals being less thermodynamically stable than the corresponding oxide. Essentially, the oxidation process ceases long before the metal is completely converted to the oxide because of the formation of a surface oxide layer, which impedes oxygen transport through it. Thus, passivation prevails and the oxidation process is retarded. The temperature range and rate of oxide formation vary widely with the metal involved, as is the case for magnesium and niobium.[6] Such studies are of critical importance in metallurgy and more particularly in metal fatigue studies and, more generally, in studies of metal catalysts and metal surface phenomena.

A variation of TGA is thermomagnetometry (TM). A thermomagnetometry system is obtained by placing the sample in a magnetic field—achieved by using a permanent magnet as the sample chamber. The apparent mass of the sample corresponds to the sum of the actual sample mass and the applied magnetic force. The external magnetic field may be applied periodically so that TGA and TM curves can be compared. Warne and Gallagher[7] have published recent reviews of TM and its applications to minerals[8] and inorganic materials.[9] It is of particular significance in the investigation of the thermal characteristics of ferromagnetic materials. Thus, for example, TM has been applied by Gallagher[10] to examine the spinel phases formed during the decomposition of siderite. Siderite samples heated in nitrogen show the onset of decomposition at 400°C. As the wustite (FeO) formed originally is oxidized by the evolved carbon dioxide, magnetite is formed. The apparent mass gain in the TM curve is because of the progressive growth and crystallization

of the magnetite nuclei. This feature is not detectable by TGA alone and illustrates the additional advantages of TM over TGA. When the temperature exceeds the Curie point of magnetite, the TM curve coincides with the TGA curve in the absence of the applied magnetic field. In oxygen, the wustite intermediate is oxidized so rapidly to hematite that the strongly magnetic spinel phase is not formed. Such studies are of major significance in metallurgy and are of particular relevance to the quantitative characterization of ferromagnetic chemical processes.

SIMULTANEOUS THERMAL ANALYSIS

Simultaneous thermal analysis essentially involves the combination of TGA and DTA, although combination of TGA with differential scanning calorimetry (TGA–DSC) is also possible but less frequently undertaken. The pioneering simultaneous system was the ''derivatograph,'' developed by the Paulik brothers in 1955.[11] This type of instrumentation involves the mounting of a DTA head onto the thermobalance suspension or rise rod, which is connected directly to the microbalance. Modern simultaneous systems are capable of a resolution of the order of 1 μg with sample masses ranging from 5 to 100 mg. As for all thermal analysis techniques, temperature calibration is essential and is usually achieved by using the suite of ICTAC Certified Reference Materials[12] and/or the ICTAC Curie point temperature standards.[1]

There are advantages of simultaneous systems over independent TGA and DTA (or DSC). Fundamentally, it is difficult to correlate results obtained on different portions of a sample with different techniques because of ''sample'' and ''instrument'' factors, whereas, in general, results obtained on the same sample by two thermal analysis techniques simultaneously are reproducible. In general, simultaneous thermal analysis may be viewed as yielding data, which are more significant than the sum of the two sets of data obtained from the two noncoupled techniques because a synergistic effect operates. This effect is readily apparent—TGA can only detect thermal events associated with a mass change and hence a phase change such as ''melting'' is not detected by this technique but melting is readily detected as an endothermic change in DTA. Thus, simultaneous thermal analysis shows no sample mass change associated with the melting endotherm. Further, DTA (or DSC) alone is unable to differentiate between types of sample enthalpy changes such as melting and decomposition, whereas simultaneous thermal analysis is able to produce such differentiation because decomposition is associated with a mass change. Thus, simultaneous thermal analysis is able to characterize a thermal event both qualitatively and quantitatively. The synergistic effect operating in simultaneous thermal analysis is readily apparent from an examination of a material, which melts with decomposition. The overall event is readily characterized by simultaneous thermal analysis, whereas TGA and DTA alone can only partially define such a transformation. Other obvious advantages of simultaneous thermal analysis are ''time saving'' (only need to do one experiment instead of two), ''sample saving'' (only require one portion of a sample instead of two) and ''capital cost savings'' (only need one instrument instead of two). However, generally, the sensitivity of individual thermal analysis techniques is decreased on combination because of essential compromises in instrumental design. This is particularly apparent with the combination of TGA and DSC. Essentially, the major advantage of simultaneous thermal analysis is associated directly with the unambiguous correlation of two sets of thermal analysis data obtained on the same sample subjected to the same temperature regime. Peak measurements from DTA can be correlated with maximum rate of mass loss, which is determined from the corresponding TG curve, and DTA peak areas can be correlated with single process thermal events in cases where the corresponding TGA curves show that no decomposition or sublimation is taking place. These features are particularly advantageous in the high-temperature studies of complex materials.

Simultaneous thermal analysis has been widely applied to study the thermal behavior of minerals, coals, ceramics, polymers, pharmaceuticals, and superconductor materials. For example, the combustion of coal is highly sensitive to experimental conditions, and thermal analysis experiments on coal are usually associated with low reproducibility, which makes them of little value in the processing of coal and its derived products. However, STA can overcome these problems and is particularly useful for the determination of the thermal characteristics of coal samples. In general terms,drying of the sample is evident from the TGA curve and is complete at around 200°C, and the large exotherm at 500°C because of oxidation can be attributed to the ''dry weight'' of the coal sample. Comparison of the simultaneous TGA and DTA curves reveals that a substantial proportion of the oxidation occurs during a period of mass gain, a feature that would be difficult to ascertain by other means. Spontaneous ignition of coal samples can also be studied very effectively by simultaneous thermal analysis.

EVOLVED GAS ANALYSIS

Evolved gas analysis is a ''coupled'' thermal analysis technique and essentially characterizes the gases evolved during a thermal analysis investigation and hence augments the information obtained. The most common EGA techniques emerge from the coupling

of TGA and/or DTA with a gas chromatograph and/or a mass spectrometer. The coupling relies on a flowing purge gas stream through the system such that the evolved gases are swept into the coupled gas analysis system. The most comprehensive EGA system is a TGA–DTA–GC–MS system, which combines the advantages of simultaneous thermal analysis with the gas separation ability of GC and the identification ability of MS. Moreover, TGA–MS and TGA–FTIR systems are available commercially. For all EGA systems, the direct coupling of an analytical instrument to a thermal analysis system poses significant challenges and the interface has to recognize the different conditions associated with the individual instruments involved and the different time intervals required for sampling. For example, in TGA–MS, a TGA system usually operates at ambient pressure, whereas an MS operates at high vacuum. The most significant problem to overcome in this context is the possible condensation of vapors in the interface. Heating of the latter may remedy this problem but may also cause decomposition of the evolved gaseous components.

Fundamentally, EGA systems require powerful software control and data analysis systems, which, in addition to overall hardware control and displaying and analyzing output data, may also store reference libraries of MS or FTIR spectra to assist identification of the gaseous products of a thermal decomposition. However, EGA represents the most comprehensive group of thermoanalytical systems available for the characterization of the thermal behavior of materials.

Evolved gas applications are widespread and diverse. Coordination chemistry has been considerably enhanced by thermal analysis studies of a wide range of metal complexes. The emphasis of these studies has been to derive thermal decomposition mechanisms for such metal complexes. Prior to the development of EGA, such reported mechanisms were suspect because the volatile decomposition products and the solid intermediates were, in general, not identified directly but characterized indirectly on the basis of mass loss data. This procedure is flawed when multiple gaseous decomposition products and nonstoichiometric intermediates are involved. The thermal analysis of one such group of metal complexes, the metal dithiocarbamates, has been extensively studied and reviewed by Hill and Magee[13] and these are definitive studies in this field.

Evolved gas analysis, particularly in the form of TGA–DTA–MS, has obvious synthetic polymer applications. It has been applied to study the thermal behavior of homopolymers, copolymers, polymeric blends, composites, residual polymers, solvents, additives, and toxic degradation polymers. In the latter context, hydrogen chloride evolution from heated polyvinylchloride materials is readily quantified by TGA–DTA–MS and such data are of major significance in the design of fire-resistant polymeric materials. Pyrotechnic materials have also been studied by STA–MS. A complex sequence of thermal events relates to the decomposition of these materials involving interactions between the nitrocellulose, perchlorate, and metal components with periodic release of carbon dioxide and oxygen. Only by EGA is it possible to rationalize the thermal behavior of such materials. Various EGA systems have also been used to study materials of environmental significance such as contaminated soils, waste (polymeric) products, and packaging materials. Such studies are important in the context of chemical processing and recycling strategies.

An unusual application of EGA has been chosen for selected discussion. The EGA of a graphite–calcite schist has been studied by Warrington.[14] The DTA curve for this material under oxidizing conditions, together with the EGA curves for carbon dioxide evolution obtained in oxidizing and inert conditions have been interpreted. The large exothermic DTA peak reveals a superimposed endothermic peak because of calcite dissociation. Subtraction of the EGA peak recorded in nitrogen and arising from calcite dissociation from that in an oxidizing atmosphere gives the amount of carbon dioxide emerging for the graphite oxidation, which subsequently allows a quantitative estimation of the graphite content of the material. This is a very elegant and subtle application of EGA to a naturally occurring material.

SAMPLE CONTROLLED THERMAL ANALYSIS

Sample controlled thermal analysis is the generic title for a group of thermal analysis techniques in which the heating rate is varied so as to produce a constant rate of mass loss. It is a major development and refinement of TGA and is thus discussed in this review. The various forms and applications of SCTA have been comprehensively reviewed by Sorenson and Rouquerol.[15,16] It essentially overcomes the problems associated with linear heating rates in conventional thermal analysis techniques, which can cause significant temperature and pressure gradients in the sample and thus give rise to ambiguous data. Variation of the sample temperature to maintain a constant reaction rate essentially ensures that pseudo-equilibrium conditions are maintained throughout the sample over the entire temperature range scanned. A constant rate of reaction is maintained by using the measured mass loss as feedback to the furnace temperature controller. The resulting conditions for the analysis correspond to near isothermal and hence the so-called ''quasi-isothermal controlled rate thermal analysis'' results, which is more appropriately termed ''sample controlled thermal analysis.''

Several variations of SCTA have been developed by Parkes et al.,[17] in which the reaction rate for gas–solid reactions is controlled by programming the concentration of reactive gas whilst keeping the temperature constant. Also, further variants involve resolution in the temperature and time domains. Time-resolved SCTA techniques are useful in situations where it is necessary to isolate and collect gaseous products and/or the residual solids of a thermal analysis experiment for subsequent identification by other analytical techniques. Also, SCTA has been developed for the investigation of ''large'' samples of the order of 1 g—thereby characterizing SCTA as a ''preparative technique.'' A further variant within the SCTA family is ''Hi-Res'' TG^{TM}, in which the heating rate is reduced when the system crosses the upper threshold set for the reaction rate. The changes in heating and cooling rates are controlled by adjustment of the values of two numerical parameters: ''resolution'' and sample ''selfheating'' and ''sensitivity.'' As SCTA essentially eliminates sample ''selfheating'' and inhomogeneous reaction within a sample, the resolution and the sensitivity of SCTA are generally greater than those associated with the conventional TGA technique.

The essential difference between conventional TGA and SCTA is in the input to the control system. In conventional thermal analysis, a linear sample heating rate is imposed, whereas in SCTA, the heating rate is related directly to the reaction rate.

In addition, SCTA offers a unique possibility for ''preparative scale'' thermal analysis whilst maintaining high levels of sensitivity and resolution. Such an SCTA system has recently been developed.[18] The focus of the system is a microbalance and a high-temperature water-cooled furnace, located above the balance and allowing operation to 1000°C. The sample is contained in a silica or platinum crucible and sits on top of a rise rod connected directly to the balance pan. Samples in excess of 1 g can be investigated, but routinely, samples of 500 mg are usually involved. Grinding of powdered samples to mesh size 150–250 μm is advantageous. The sample temperature is monitored by a thermocouple located centrally within the furnace tube. Radiation shields attached to the thermocouple

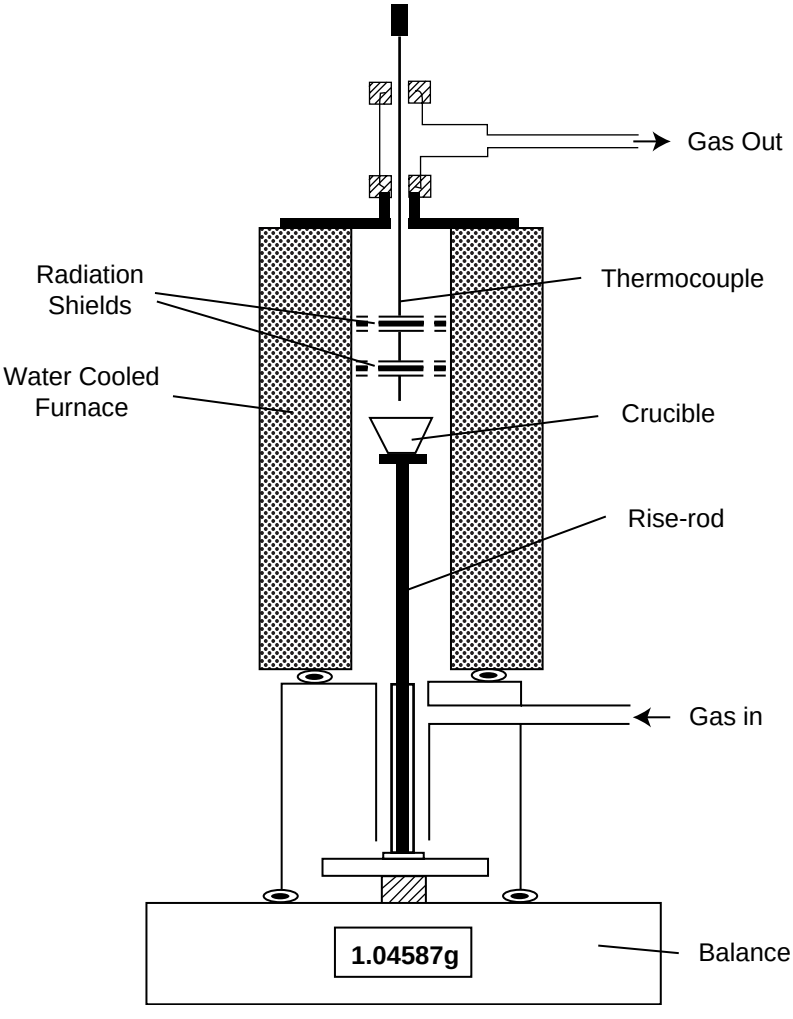


Fig. 1 Schematic diagram of an SC-TGA system.

reduce convection currents within the sample region. A gas diffusion chamber, situated between the balance and the furnace, permits experiments to be carried out in an upward flow of gas at a constant flow rate via continuously pumping the system. Although the overall system is not sealed, this net gas flow prevents air entering the system and results from the differential flow rates at the inlet and at the outlet, as measured by individual flow meters. Volatile decomposition products can be trapped using a combination of filters and activated charcoal and liquid nitrogen traps. A schematic diagram of this SCTA system is shown in Fig. 1 and a block diagram of the associated data acquisition and control systems is shown in Fig. 2.

The temperature of the furnace is continuously varied by a temperature controller, and the outputs from the balance and from the sample temperature thermocouple are logged by a computer fitted with a 16-bit dynamic range data acquisition board. The signals are processed via a program derived using a combination of Virtual Basic and C++ and the derivative of the TGA curve is used to control the heating program via the temperature controller.

Several SCTA modes are available with this system, including linear heating TGA and proportional heating TGA. In addition, the control software has a ''stop mass'' facility, by which ''products'' can be produced at any preselected degree of reaction, thereby conforming to the ''preparative'' aspect of the system. Modifications to the control system give rise to ''time-resolved SCTA,'' incorporating peak slope heating and dynamic rate modes. This ''in-house'' construction is thus a very versatile SCTA system, which incorporates the principal advantages of the ''sample controlled'' methodology.

Applications of SCTA take advantage of the enhanced uniform reaction environment provided by this technique, as compared with conventional thermal analysis techniques. A major problem with conventional thermal analysis techniques is that the imposed linear sample heating rates can cause temperature and pressure gradients within the sample. Thus, sample ''selfheating'' results, and this poses an experimental difficulty for rapid, highly exothermic decompositions. Essentially, in conventional TGA, it is impossible to unambiguously define the ''sample temperature'' throughout the analysis. In SCTA, ''selfheating'' of the sample is eliminated because it is the sample decomposition rate that determines the rate of heating and not the external heating program. Thus, with SCTA, thermal events associated with a material can be studied under controlled conditions and this feature characterizes SCTA as an enhanced thermal analysis technique, particularly suited to the study of microthermal events, reversible thermal events, and the kinetics of a wide variety of chemical and biochemical phenomena.

The Centre for Thermal Studies and the Centre for Applied Catalysis at the University of Huddersfield,

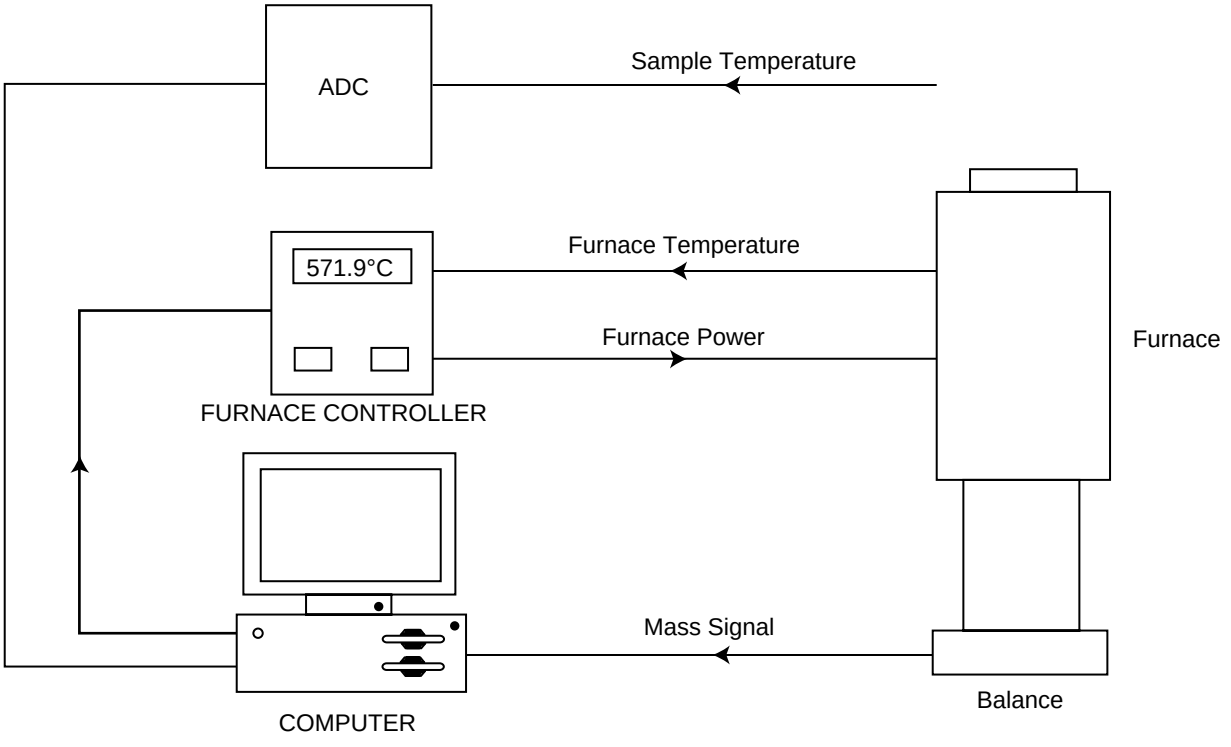


Fig. 2 SC-TGA control system.

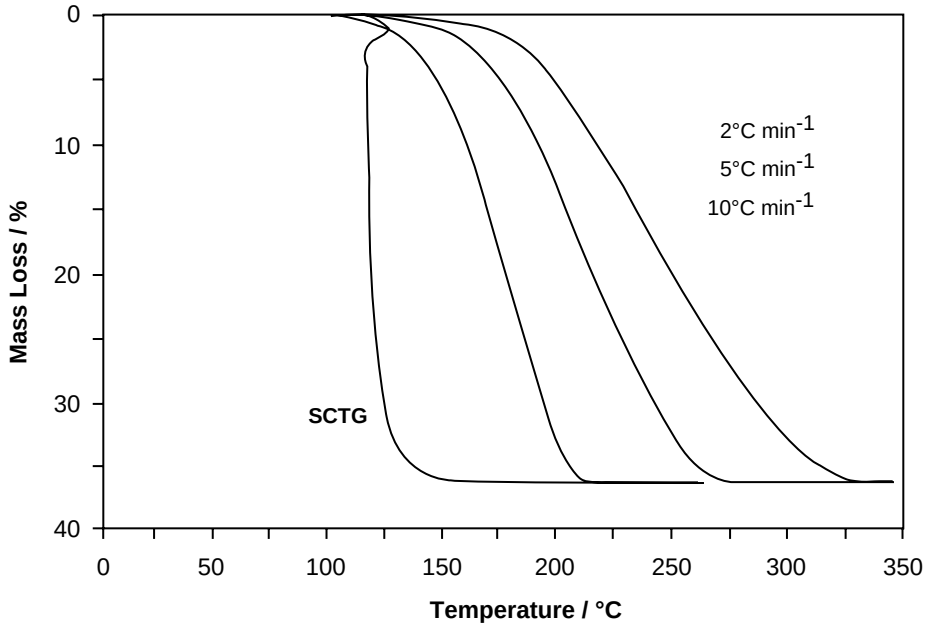


Fig. 3 SC-TGA and TGA curves of sodium bicarbonate.

U.K., have been pioneers in developing the ''preparative scale'' SCTA system described above and investigating its potential to study the thermal behavior of a wide range of materials. The principal advantage of SCTA over conventional TA techniques is that with the former, thermal events can be studied under controlled conditions. A simple example of ''controlled decomposition'' is the thermal decomposition of sodium bicarbonate, and SCTGA and TGA curves for this process are shown in Fig. 3.

The most striking feature of the (proportional heating) SCTGA curve is controlled decomposition over a narrow temperature range. The corresponding TGA curves show a reduction of the decomposition temperature range with decreasing heating rate, but the SCTGA decomposition temperature range corresponds to zero heating rate of the sample and hence, via SCTGA, the most realistic thermal decomposition temperature ranges are obtained. This feature is of particular significance for samples that violently decompose or ignite, such as pyrotechnics and coals in an oxidizing atmosphere. The primary advantage of the sample controlled approach in thermal analysis in reducing both the temperature and the gas concentration gradients in a sample is illustrated by the oxidation of a Drayton (Australia) coal sample (Fig. 4).

Even at a 500 mg sample mass level, a controlled oxidation is apparent and spontaneous ignition is

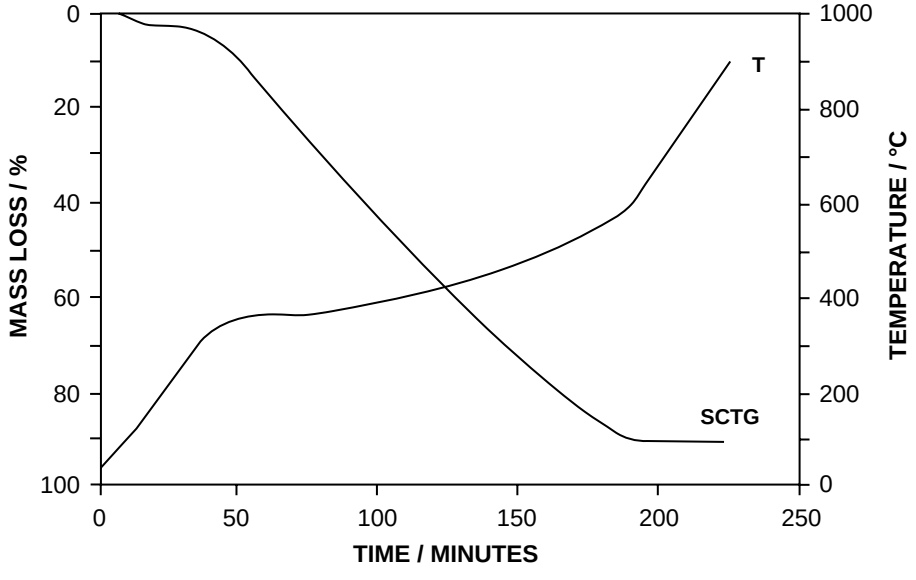


Fig. 4 SC-TGA curve for the oxidation of Drayton coal.

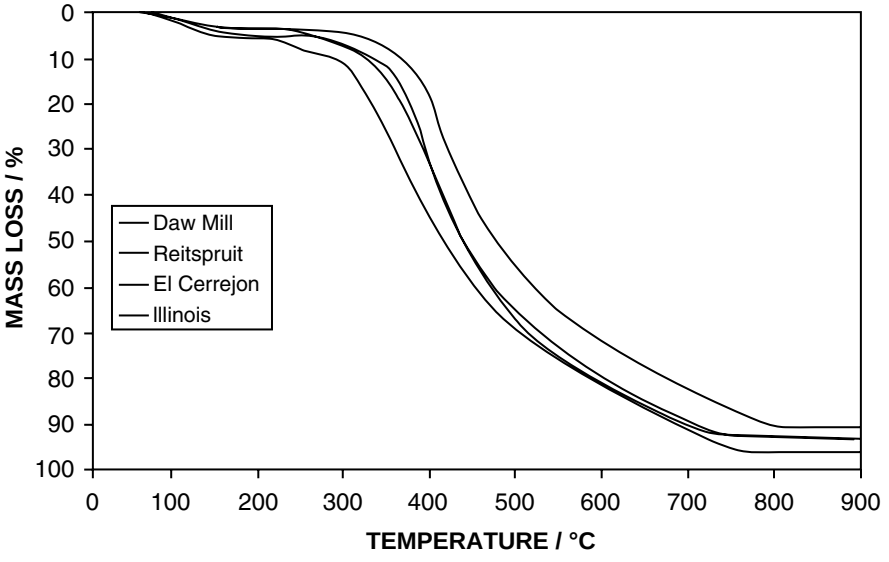


Fig. 5 SC-TGA curves for the oxidation of different coals.

avoided during the analysis. In conventional TGA using the same conditions, selfheating of the sample leads to ignition and hence a loss of meaningful data after ignition, coupled with ''apparent'' decomposition temperature ranges. Similarly, the controlled oxidation of four different coals is shown in Fig. 5.

Although the SCTGA profiles are similar in character and reveal controlled oxidation, the subtle differences in oxidation behavior of these coals are clearly revealed. Such differences are not readily deciphered from TGA curves because the dominating feature is ignition, which effectively masks the intermediate thermal events.

Coal pyrolysis can also be studied by SCTGA conducted in a nitrogen atmosphere. The proportional SCTGA curve for the pyrolysis of a sample of Illinois (U.S.A.) coal is shown in Fig. 6.

Initial pyrolysis occurs over a relatively narrow temperature range and the slow nature of the later stages of pyrolysis is indicated by the temperature program returning to maximum heating rate. The ''stop-mass'' facility of the SCTGA system provides ''preparative opportunities'' and in the context of coal pyrolysis, chars can be produced at any stage of the pyrolysis process, which can subsequently be characterized by other analytical techniques. Further, the oxidation of such chars can be monitored by SCTGA as shown in Fig. 7 for a Drayton coal char. Prior to the development of SCTGA, coal char thermal analysis was beset with controversy because of the indecisive nature of the derived data.

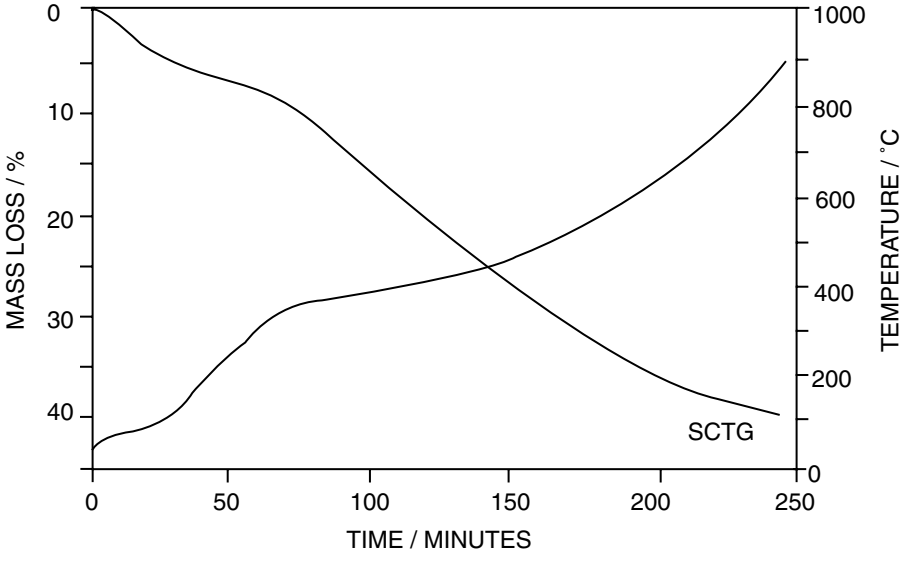


Fig. 6 SC-TGA curve for the pyrolysis of Illinois coal.

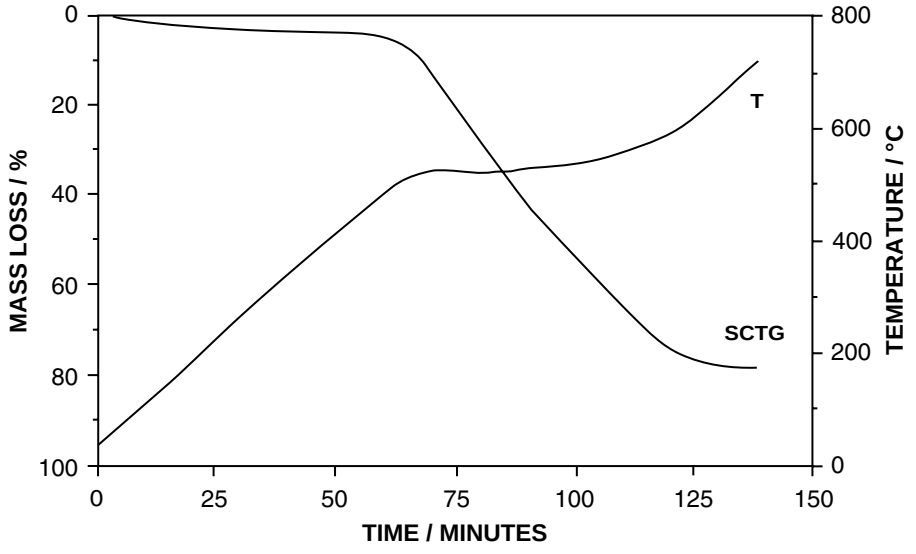


Fig. 7 SC-TGA curve for the oxidation of a Drayton coal char.

The high reactivity of finely divided zirconium powder toward aerial oxidation makes conventional TA studies difficult to perform under controlled conditions. Fig. 8 shows an SCTGA curve for a 10 mg sample of zirconium powder, which, in a TGA–DTA experiment under similar conditions but using a linear heating rate, shows ignition at about 410°C. However, it is apparent that under proportional heating SCTGA conditions, a smooth oxidation reaction process results, which is 50% complete below 400°C. These types of studies are most suitable for pyrotechnic mixtures, because reliable ignition data are generated.

The possibility of performing sample controlled thermomagnetometry is illustrated by an SCTM experiment on ICTAC nickel, with a permanent magnet placed above the thermobalance furnace (Fig. 9).

During the heating program, the sample oscillates between the magnetized and demagnetized states. If a zero cooling rate is applied, this transition can be made to take place over a narrow temperature range, thereby characterizing the magnetic property of the metal.

The scope of temperature- and time-resolved SCTGA has been revealed by Parkes, Barnes and Charsley,[17] via studies of the decomposition of inorganic salts. Temperature-resolved SCTA is particularly useful for resolving the dehydration characteristics of copper sulfate pentahydrate. A typical linear heating (LH) TGA profile for this salt is shown in Fig. 10, and is compared with a proportional heating (PH) SCTGA profile over the same temperature range.

The typical 2 : 2 : 1 water loss is shown in both profiles but the resolution of these peaks is superior

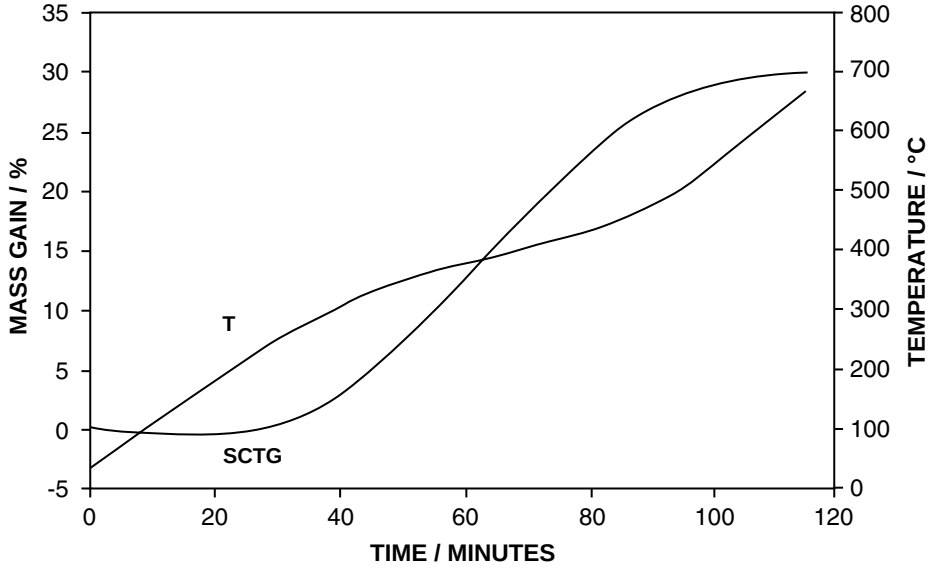


Fig. 8 SC-TGA curve for the oxidation of zirconium metal powder.

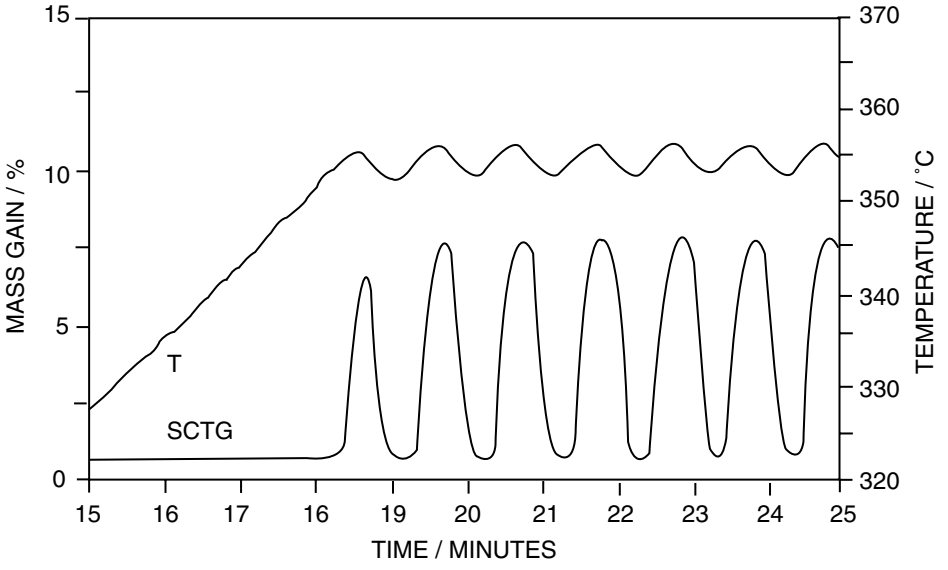


Fig. 9 SC-TM curve for "ICTAC Nickel."

in the SCTGA profile. Further, the latter shows how the heating rate is reduced over each dehydration peak and increased between the peaks, which is the defining feature of temperature-resolved SCTA. Further, temperature-resolved SCTA can be used in mechanistic studies of complex reactions. The decomposition of ammonium hydrogen carbonate is associated with the coevolution of ammonia, water, and carbon dioxide.

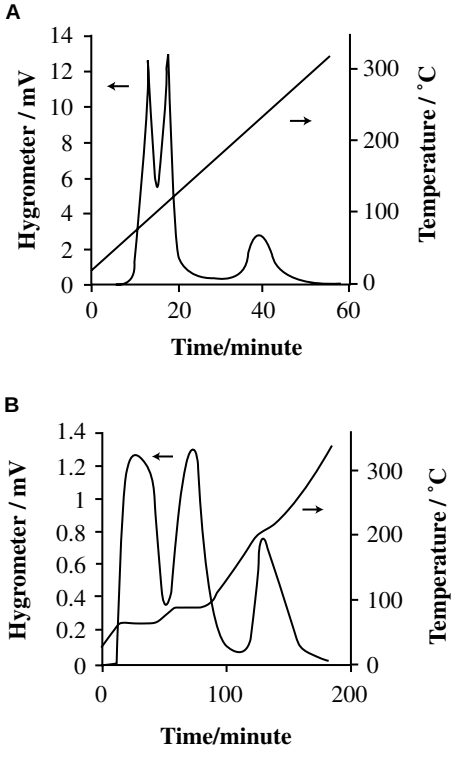


Fig. 10 LH (A) and PH (B) curves for the dehydration of copper sulfate pentahydrate.

Proportional heating SCTA in conjunction with mass spectrometry was applied to study this decomposition process, and linear heating and proportional heating TA curves are shown in Fig. 11.

A single decomposition peak, centered at 125°C, is apparent from the linear heating profile, whereas the proportional heating profile shows the presence of a nucleation stage prior to decomposition, thereby providing mechanistic indications of the decomposition mode.

An application of the time-resolved SCTA technique is shown in Fig. 12, relating to linear heating (LH) and peak slope heating (PSH) of a 1 : 2 mixture of copper hydroxycarbonate and zinc hydroxycarbonate.

Decomposition of this mixture yields both carbon dioxide and water, and the latter was monitored using a hygrometer. The principal feature of time-resolved SCTA is shown by the PSH profile. It is apparent that the switch to zero heating rate on the falling edge of each peak is rapid, whereas the increase in heating rate as the first decomposition approaches completion is slower. Also, the switching point between heating and nonheating is at the same relative point on each peak despite their difference in size. The enhanced resolution of the decomposition peaks without distortion of shape is apparent from the PSH profile and this feature is indicative of the ability of PSH to time-resolved gaseous decomposition products for "online" evolved gas analysis.

It should however be noted that although both proportional heating SCTA and peak-resolved SCTA provide enhanced resolution in the temperature and time domains, respectively, both of these modes are typically associated with increased experimental times compared with those of conventional TA experiments. This feature at present eliminates SCTA as an analytical technique for the rapid characterization of

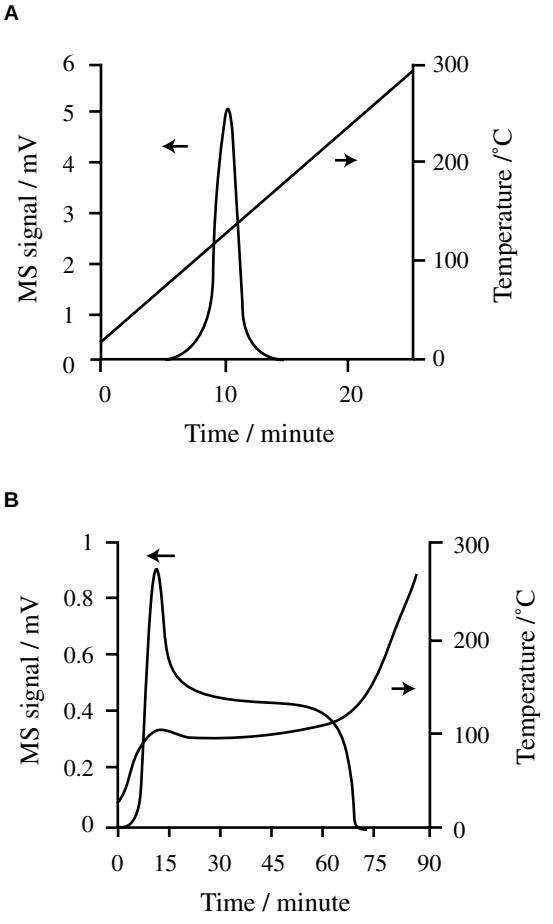


Fig. 11 LH (A) and PH (B) curves of ammonium hydrogen carbonate.

materials and in quality control applications. However, these techniques offer many other advantages, most notably in providing mechanistic and kinetic information on the thermal decomposition behavior of materials.

TA Instruments (U.S.A.) have recently produced a commercial SCTA system—known as ''hi-res'' TG or ''dynamic rate'' TG.[19] The basic principle of this system is that there is a marked decrease in heating rate when the system crosses a preset upper reaction rate threshold. The changes in heating and cooling rates are controlled by adjusting the values of two numerical parameters, referred to as ''resolution'' and ''sensitivity.'' Dynamic rate TG varies the heating rate smoothly and continuously in response to the rate of sample decomposition so that the resolution of the associated mass change is maximized. This mode allows rapid heating in regions where no thermal transitions occur and proportionally slow heating during thermal events. An illustrative example of the advantages of ''dynamic rate'' TG application is a study of a sodium/potassium bicarbonate mixture. It is apparent from the dynamic rate TG curve that enhanced resolution of the decomposition peaks is achieved in a time-frame of less than 10% of the conventional TA experiment.

MODULATED TEMPERATURE THERMOGRAVIMETRIC ANALYSIS

Modulated temperature thermogravimetry (MT-TGA or MT-TG) is a new TA methodology, which has particular application to the study of reaction kinetics.[20] Thermogravimetric analysis has been applied routinely to derive kinetic parameters for a wide range of reactions and this large and complex field has been comprehensively reviewed by Galwey[2] and Reading.[3] In MT-TGA, a sinusoidal temperature modulation is superimposed on the conventional linear heating rate. This modulated temperature stimulus results in an oscillatory response in the rate of mass loss, deconvolution of which via a real time discrete Fourier transformation produces the desired kinetic parameters. A typical MT-TGA profile of polytetrafluoroethylene (PTFE) is shown in Fig. 13.[20]

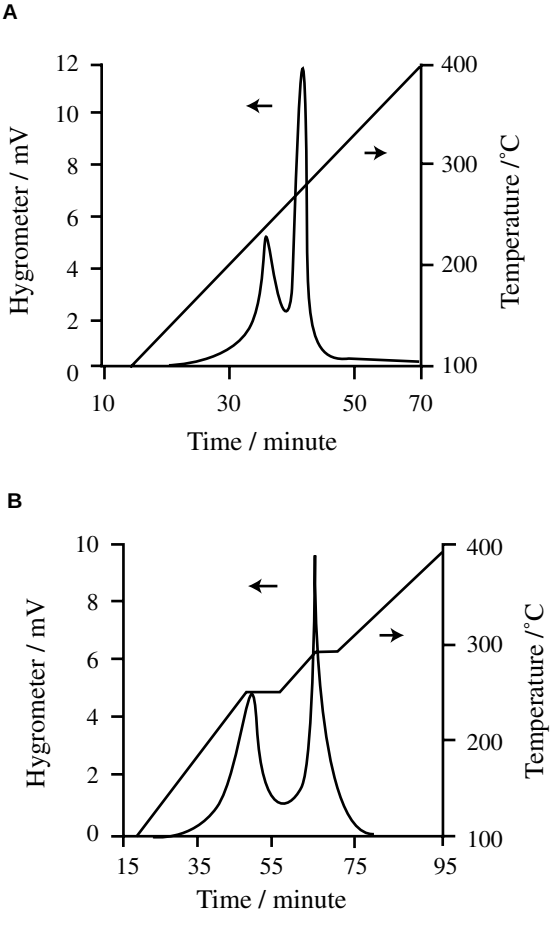


Fig. 12 LH (A) and PSH (B) curves of a 1:2 mixture of copper and zinc hydroxycarbonates.

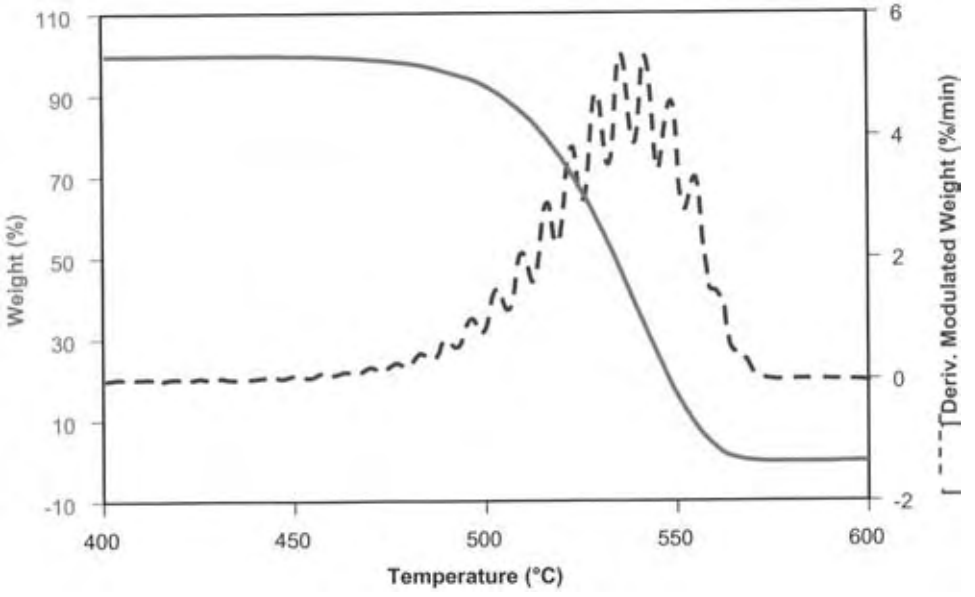


Fig. 13 TGA and MT-TGA curves of PTFE. *(View this art in color at www.dekker.com.)*

Derivation of kinetic parameters by conventional TGA is exceedingly difficult when multistep decomposition reactions are involved. Because the resolution of MT-TGA is much greater than that of conventional TGA, greater resolution of individual thermal events is achieved with the former, thereby simplifying the subsequent kinetic analysis of the overall decomposition process.

CONCLUSIONS

Conventional thermogravimetric analysis has considerably diversified over the last decade. Thermogravimetric analysis has successfully been combined with differential thermal analysis (DTA) to give simultaneous thermal analysis. The principal advantage of TGA–DTA is that both TGA and DTA are obtained simultaneously under identical experimental conditions for a single sample of material. This allows a direct comparison of the TGA and DTA curves and for greater characterization of thermal events at defined temperatures or over defined temperature ranges. The coupling of TGA with another analytical technique, such as mass spectrometry, has produced EGA. The principal advantage of EGA is that volatiles evolved from a thermal analysis of a sample are analyzed ''online'' by a directly coupled mass spectrometer. Thus, EGA further enhances the analytical potential of conventional TGA. The principal disadvantage of conventional TGA, ''selfheating'' of the sample, is overcome with sample controlled thermal analysis. In SCTA, the sample is heated slowly through thermal events and rapidly between them. Such a differential heating program leads to increased sensitivity and resolution in TGA but also to a much increased time-frame for the analysis. Dynamic rate TG appears to have addressed both of these features and hence has much potential as a ''high resolution/rapid'' thermal analysis system, which, unlike SCTA, can be applied for rapid and reproducible thermal analysis of a wide range of complex materials. Finally, modulated temperature thermogravimetric analysis has enhanced potential for the kinetic analysis of thermal decomposition reactions over conventional TGA because of its greater resolution of thermal events.

REFERENCES

1. Haines, P.J. *Principles of Thermal Analysis and Calorimetry*; Royal Society of Chemistry: Cambridge, U.K., 2002; Chapter 2, 20.
2. Galwey, D. Is the science of thermal analysis based on solid foundations? A literature appraisal. Thermochim. Acta **2004**, *413*, 139–183.
3. Reading, M. The kinetics of homogeneous solid state decomposition reactions: a new way forward. Thermochim. Acta **1998**, *135*, 37–57.
4. Dollimore, D.; Griffiths, D.C.; Nicholson, D. The thermal decomposition of oxalates (11). Thermogravimetric analysis of various oxalates in air and nitrogen. J. Chem. Soc. **1963**, 2617–2623.
5. Brown, M.E. *Introduction to Thermal Analysis—Techniques and Applications*; Kluwer: Amsterdam, 2001; Chapter 3, 46.
6. Heal, G.R. Thermogravimetry and derivative thermogravimetry. In *Principles of Thermal*

Analysis and Calorimetry; Haines, P.J., Ed.; Royal Society of Chemistry: Cambridge, U.K., 2002; Chapter 2, 38.
7. Warne, S.J.; Gallagher, P. Thermomagnetometry. Thermochim. Acta **1987**, *110*, 269–279.
8. Gallagher, P.; Warne, S.J. Thermomagnetometry and the thermal decomposition of siderite. Thermochim. Acta **1981**, *43*, 253–267.
9. Warne, S.J. Thermal analysis—a resurgence in the earth sciences with applied, industrial and environmental aspects. Thermochim. Acta **1991**, *192*, 19–28.
10. Gallagher, P. Applications of thermal analysis to the study of inorganic materials. Thermochim. Acta **1993**, *214*, 1–7.
11. Paulik, F.; Paulik, J. Simultaneous thermoanalytical examinations by means of the derivatograph. In *Comprehensive Analytical Chemistry*; Svehla, G., Ed.; Elsevier: Amsterdam, 1981; Vol. X11(A).
12. Richardson, M.J.; Charsley, E.L. *Handbook of Thermal Analysis and Calorimetry*; Elsevier: Amsterdam, 1998; Chapter 13.
13. Hill, J.O.; Magee, R.J. The thermochemistry of metal dithiocarbamate complexes and related compounds. Rev. Inorg. Chem. **1981**, *3*, 141–197.
14. Warrington, S.B. Simultaneous thermal analysis techniques. In *Principles of Thermal Analysis and Calorimetry*; Haines, P.J., Ed.; Royal Society of Chemistry: Cambridge, U.K., 2002; Chapter 2, 20.
15. Sorenson, O.T.; Rouquerol, J. *Sample Controlled Thermal Analysis*; Kluwer: Amsterdam, 2003.
16. Sorenson, O.T. Quasi-isothermal methods in thermal analysis. Thermochim. Acta **1981**, *50*, 163–175.
17. Parkes, G.M.B.; Barnes, P.A.; Charsley, E.L. New concepts in sample controlled thermal analysis: resolution in the time and temperature domains. Anal. Chem. **1999**, *71*, 2482–2487.
18. Charsley, E.L.; Rooney, J.J.; Hill, J.O.; Parkes, G.M.P.; Barnes, P.A.; Dawson, E.A. Development and applications of a preparative scale sample controlled thermogravimetric system. J. Therm. Anal. Calorim. **2003**, *72*, 1091.
19. Heal G.R. Thermogravimetry and derivative thermogravimetry. In *Principles of Thermal Analysis and Calorimetry*; Haines, P.J., Ed.; Royal Society of Chemistry: Cambridge, U.K., 2002; Chapter 2, 51–53.
20. Buckman, N.; Blaine, R.; Dallas, G. Modulated thermogravimetry—a new approach for obtaining kinetic parameters. Chem. Aust. **1998**, *68* (2), 22–23.

Thermomechanical Analysis

Kevin P. Menard
PerkinElmer Thermal Laboratory, University of North Texas Materials Science Department, Denton, Texas, U.S.A.

INTRODUCTION

Thermomechanical analysis (TMA) is the technique of measuring the dimensional changes in a specimen as a function of time or temperature. It can be argued that rheology and traditional mechanical tests should be included in this classification. In its purest form, the changes in a material's dimensions under minimal load are recorded and used as an indicator of the changes in the material's free volume. These data allow the calculation of a material's expansivity or coefficient of thermal expansion (CTE) as well as detection of transitions in the material. Thermomechanical analysis on inorganic glass was the first measurement of glass transition and still remains the best technique for that measurement in many applications. It is often said to be more sensitive to glass transition than differential scanning calorimetry (DSC) by an order of magnitude.

Pressure volume temperature (PVT) instruments are designed to probe more deeply into a material's free volume. They can be considered a subset of TMA and are treated here as such. Capable of applying literally tons of force and of reaching high temperatures, they are used to collect information of how pressure and temperature affect the volume of a material and how the transitions in a material shift as a function of pressure.

THEORY AND OPERATION OF TMA

The basis of TMA is the change in the dimensions of a sample as a function of temperature. A simple way of looking at TMA is as a very sensitive micrometer. It is believed to have developed from hardness or penetration tests and was used on polymers in 1948.[1] Subsequently, it has developed into a powerful tool in the analytical laboratory. Thermomechanical analysis measurements record changes caused by alterations in the free volume of a polymer.[2] While the latter tends to be preferred by engineers and rheologists, in contrast to chemists and polymer physicists who lean towards the former, both descriptions are equivalent in explaining behavior. Changes in free volume, v^f, can be monitored as a volumetric change in the polymer: by the absorption or release of heat associated with that change; loss of stiffness; increased flow; or by the change in relaxation time. The free volume of a polymer, v^f, is known to be related to viscoelasticity,[3] aging,[4] penetration by solvents,[5] and impact properties.[6] Defined as the space a molecule has for internal movement, it is schematically shown in Fig. 1.

The T_g in a polymer corresponds to the expansion of the free volume allowing greater chain mobility above this transition (Fig. 2). Seen as an inflection or bend in the thermal expansion curve, this change in TMA can be seen to cover a range of temperatures, of which the T_g temperature is an indicator calculated by an agreed upon method (Fig. 3). This fact seems to be forgotten by inexperienced users, who often worry why perfect agreement is not seen in the value of the T_g when comparing different methods. The width of the T_g can be as important as an indicator of changes in the material as the actual temperature.

Experimentally, TMA consists of an analytical train that allows precise measurement of position and can be calibrated against known standards. A temperature control system of a furnace, heat sink, and temperature-measuring device (most commonly a thermocouple) surrounds the samples. Fixtures to hold the sample during the run are normally made out of quartz because of its low CTE, although ceramics and invar steels may also be used. Fixtures are commercially available for expansion, three-point bending or flexure, parallel plate, and penetration tests (Fig. 4).

APPLICATIONS OF TMA

Applications of TMA are in many ways the simplest of the thermal techniques. Only the change in the size or position of a sample is being measured. However, they are also incredibly important in supplying information needed to design and process everything from chips to food products to engines. A sample of ASTM methods for TMA is shown in Table 1. Because of the sensitivity of modern TMA, it is often used to measure T_g's that are difficult to obtain by DSC, for example, those of highly cross-linked thermosets.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120028330

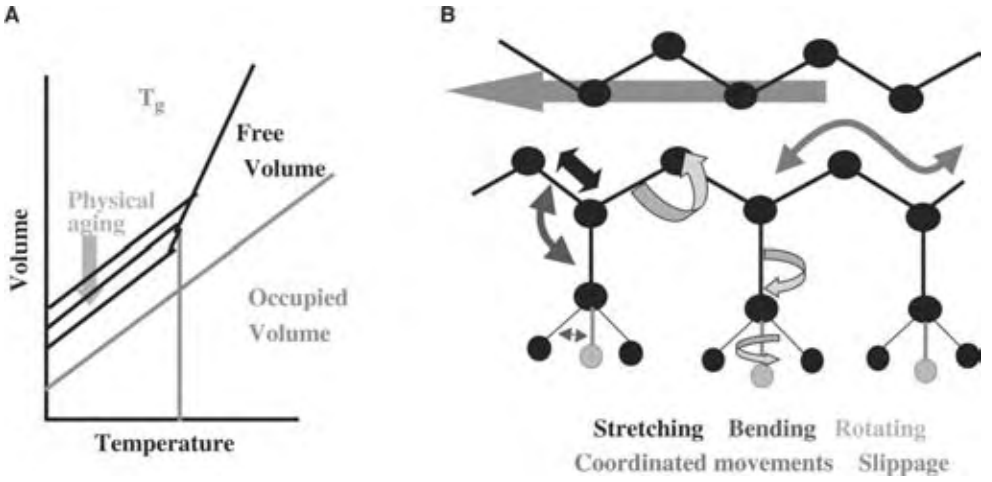


Fig. 1 Free volume, v^f, in polymers: (A) the relationship of free volume to transitions, and (B) a schematic example of free volume and the crankshaft model. Below the T_g in (A) various paths with different free volumes exist depending on heat history and processing of the polymer, where the path with the least free volume is the most relaxed. (B) shows the various motions of a polymer chain. Unless enough free volume exists, the motions cannot occur. (From Menard K.; *Dynamic Mechanical Analysis: A Practical Introduction*, CRC Press: Boca Raton, 1999).

Expansion and CTE

Thermomechanical analysis allows the calculation of thermal expansivity[a] from the same data set as used to calculate the T_g. Since many materials are used in contact with a dissimilar material in the final product, knowing the rate and amount of thermal expansion helps design around mismatches that can cause failure in the final product. These data are only available when the T_g is collected by thermal expansion, not by the flexure or penetration method. This is in many ways the simplest or most essential form of TMA measurement. A sample is prepared with parallel top and bottom surfaces and is allowed to expand under minimal load (normally 5 mN or less, ideally 0 mN) as it is slowly heated and/or cooled. The CTE is calculated by:

$$\alpha_1 = 1/l_0(\delta l/\delta T)F$$

where α_l is the linear coefficient of thermal expansion, l_0 is the original length, δl is the change in length and δT the change in temperature. F indicates that this is done under constant force. Once this value is obtained, it can be used to compare other materials used in the same produce. Large differences in the CTE can lead to motors binding, solder joints failing, composites splitting on bond lines, or internal stress build up. The T_g is obtained from the same data by measuring the inflection point in changes of the baseline slope. As a material's CTE changes dramatically at T_g, one would expect this to be an easily detected transition. It can be; however, for highly cross-linked materials, the T_g can be so broad and the change in CTE so slight as to be undetectable. Other approaches, such as flexure testing, are then used. Different T_g values are shown for each mode of testing[7] (Fig. 5) and the methods used to measure the Tg by TMA.

If the material is heterogeneous or anisotropic, it will have different thermal expansions depending on the direction in which they have been measured. For example, a composite of graphite fibers and epoxy will show three distinct thermal expansions corresponding to the x, y, and z directions. Blends of liquid crystals and polyesters show a significant enough difference between directions for the orientation of the crystals to be determined by TMA.[8] Similarly, oriented fibers and films have a different thermal expansivity in the direction of orientation to that in the unoriented direction. This is normally addressed by recording the CTE in the x, y and z directions (Fig. 6A). Bulk measurements or volumetric expansion can be made by dilatometry as shown in Fig. 6B and this is discussed in detail below.

Expansion studies can also be run on samples immersed in solvents to measure the swelling of a polymer. This test is commonly used with rubbers to measure the cross-link density of the rubber.[9] As cross-linking increases, the amount of swelling will decrease. Special fixtures are commercially available

[a]Thermal expansivity is often referred to as the coefficient of thermal expansion or CTE by polymer scientists and in the older literature. The terms are used interchangeably in this paper.

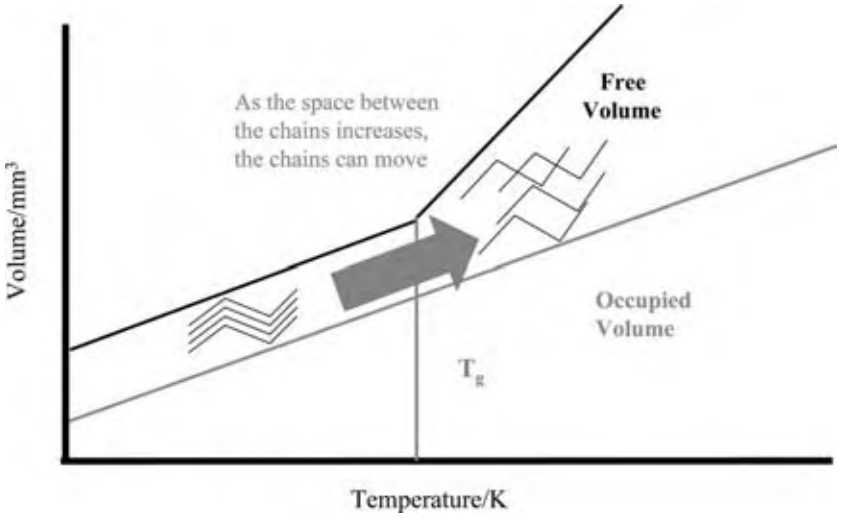


Fig. 2 The increase in free volume is caused by increased energy absorbed in the chains and this increased free volume permits the various types of chain movement to occur.

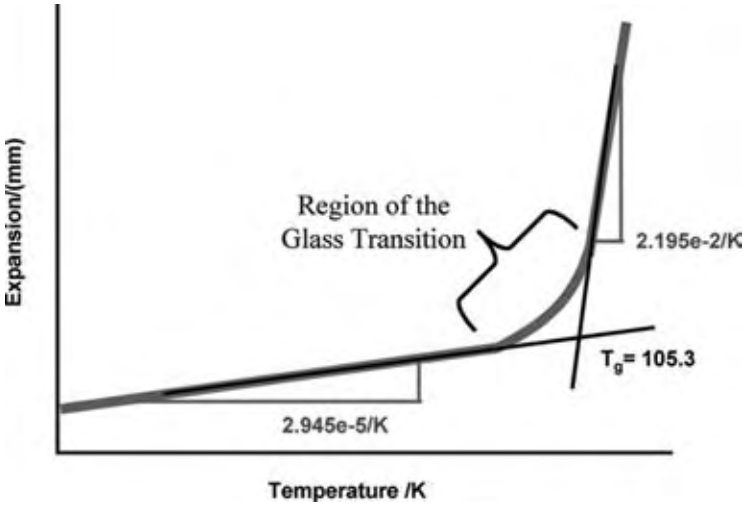


Fig. 3 The Tg is a region, shown here, between the points where the tangents depart from the curve. The Tg, by convention, is taken as the intersection of those two tangents.

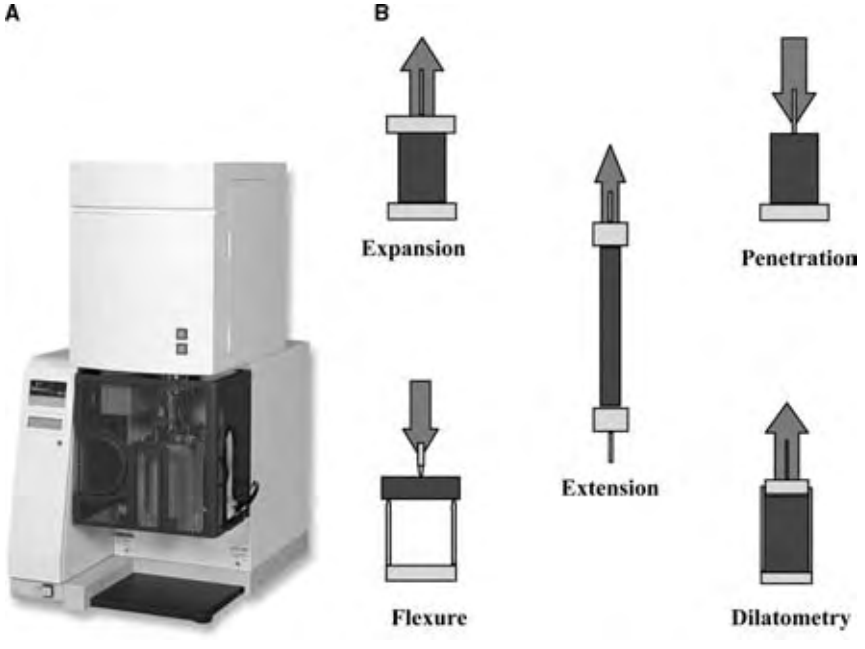


Fig. 4 A PerkinElmer TMA 7 is shown in (A) as a schematic. (Courtesy of Perkin-Elmer Instruments, Norwalk, CT.) Various test geometries are shown in (B). Normally these are made of quartz glass to take advantage of its low CTE values.

Table 1 ASTM tests for the TMA

D3386	CTE of electrical insulating materials by TMA
E 228-95	CTE by TMA with silica dilatometer
E 473-94	Terminology for thermal analysis
E 831-93	CTE of solids by TMA
E 1363-97	Temperature calibration for TMA
E 1545-95(a)	T_g by TMA
E 1824-96	T_g by TMA in tension

for this, or standard fixtures may be used in a lined furnace. A sample is immersed in oil and the degree of swelling measured. Fig. 7A shows the results from a rubber swelling study.

Flexure and Penetration

Thermomechanical analysis methods are used in geometries more commonly associated with traditional mechanical testing to increase sensitivity or to mimic other tests. The most common of these are the flexural and penetration modes. Flexure studies involve loading a thin beam, often a splinter of material, with a constant load of 100 mN or more and heating until the sample deflects. This is similar to a heat distortion test and is often used for QC as it gives clear, easily detected transitions.

Thin coatings are often better handled by penetration tests, when a small glass probe is used to penetrate the sample. This technique involves moderate force and allows the measurement of the transitions of material as thin as 10 microns. Baseline subtraction is often used to remove any effects from the substrate's thermal expansion during the test. As mentioned above, these methods often give different values for T_g, and O'Neill et al.[7] carried out an extensive study on how the method of testing affected the Tg. They reported that the expansion method gives the lowest value for T_g by TMA, and that the flexure method tends to give a T_g that is 5°C higher. By the penetration method, the T^g tends to be 10°C higher than the expansion method.

Dilatometry and Bulk Measurements

Another approach to anisotropic materials is to measure the bulk expansion of material using dilatometry (Fig. 6). The technique was used extensively to study initial rates of reaction for bulk styrene polymerization in the 1940s,[10] an experiment which the author has used in his thermal analysis class on TMA. By immersing the sample in a fluid (normally silicon oil) or

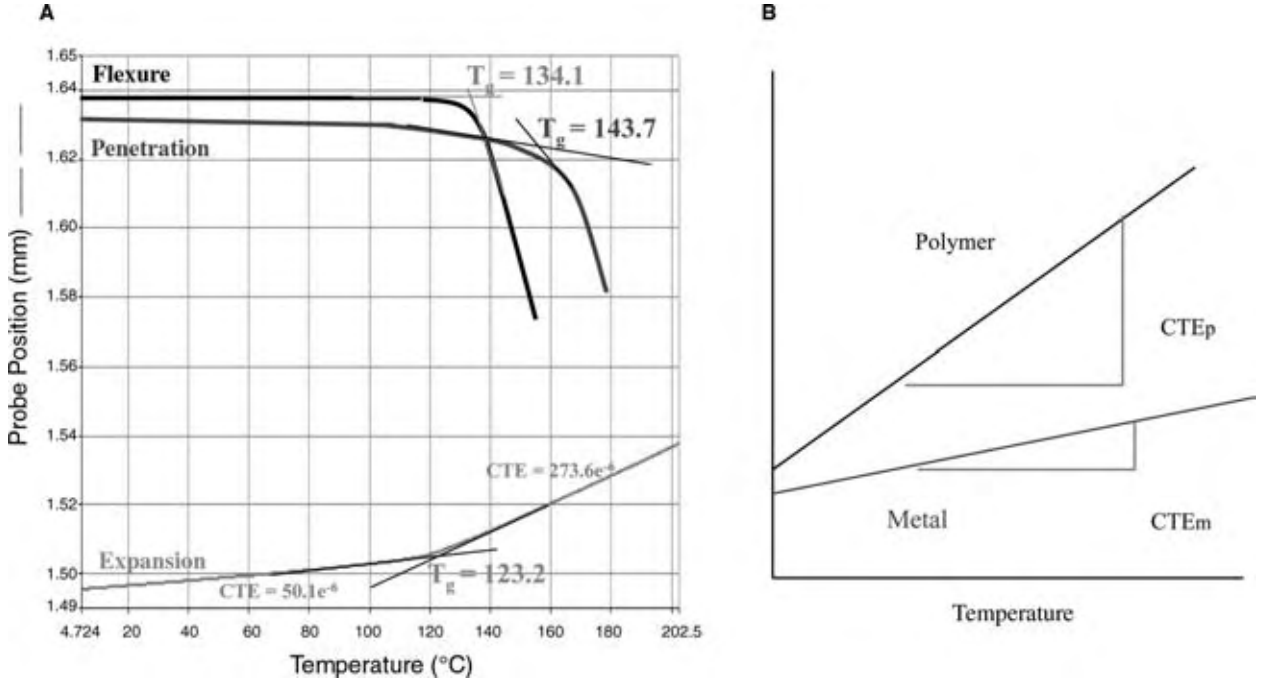


Fig. 5 Different methods of measuring the Tg in the TMA give different values as shown in (A) the overlay of the penetration, flexure and expansion runs above. In (B) we see the comparison of a polymer to a metal CTE run.

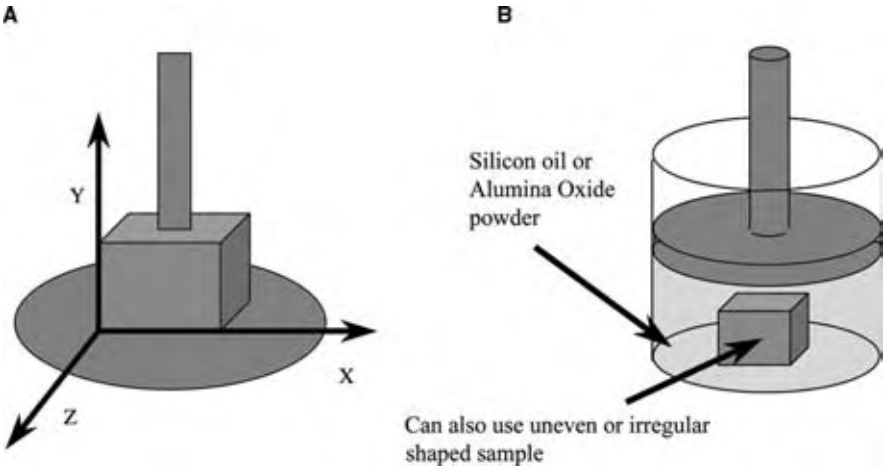


Fig. 6 Heterogeneous samples require the CTE to be determined in the x, y and z planes (A) or in bulk to obtain a volumetric expansion in the dilatometer (B). *(View this art in color at www.dekker.com.)*

powder (normally Al_2O_3) in the dilatometer, the expansions in all directions are converted to a vertical movement, which is measured by TMA. This technique has enjoyed a renaissance in the last few years because modern TMAs make it easier to perform than previously. It has been particularly useful for studying the contraction of a thermoset during its cure.[11,12] The technique itself is rather simple: a sample is immersed in either a fluid, such as silicone oil, or buried in alumina oxide in the dilatometry and run through the temperature cycle. If a pure liquid or a monomer is used, the dilatometer is filled with that liquid rather than the silicon oil or alumina oxide. A CTE can be determined and Speyer[13] reports that, as well as discussing the technique in detail.

PVT Relationship Studies

The temperature of a polymer's melting, glass transition, crystallization, and its solid–state annealing are all known to have pressure dependencies.[14] High-pressure instruments, such as the Gnomix Inc. (Colarado, U.S.A.), have been developed to study PVT relationships in polymers. In these experiments, the sample is placed in an incompressible fluid and then the desired pressure is applied. Full details of this technique, as well as a collection of P–V–T relationships for a wide range of polymers up to 200 MPa (~30,000 psi) and 400°C, have recently been published.[15] These data have mainly been collected isothermally to report the effects of pressure and temperature on the volume of

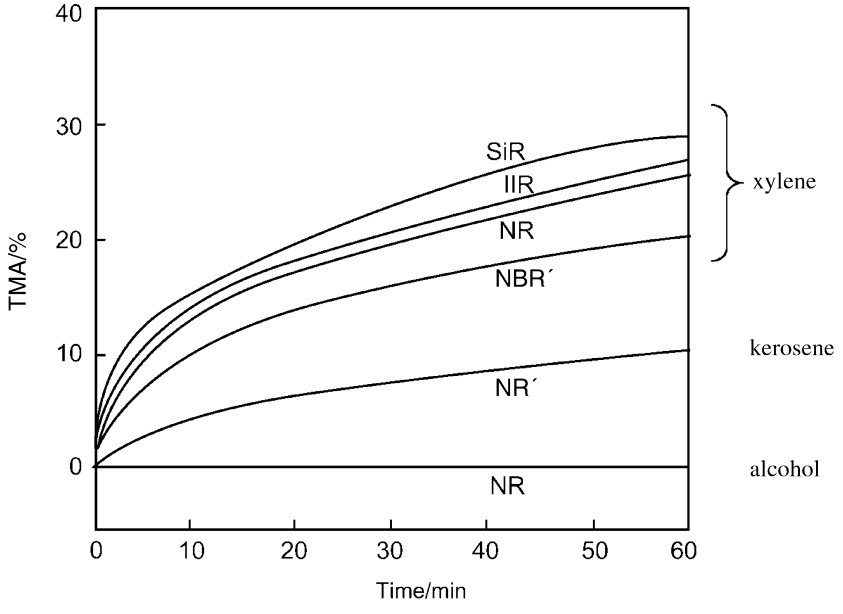


Fig. 7 Swelling of rubber samples in organic solvents measured in the Diamond TMA. NR, natural rubber; NBR, nitrile butyl rubber; SiR, silicon rubber; IIR, isobutyl rubber. (Courtesy of the PerkinElmer Instruments, Norwalk, CT.)

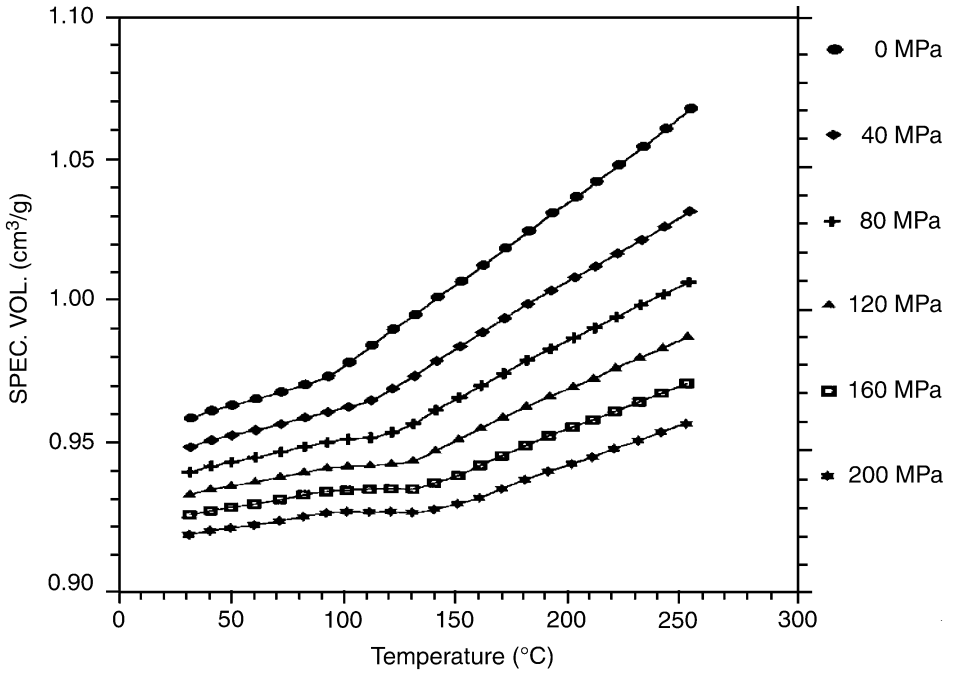


Fig. 8 The effect of pressure of the T_g of a polystyrene (M ~ 1.1ev, D ~ 1.06). (From *Zoller's Standard Pressure–Volume–Temperature Data for Polymers*, used with the permission of the Technomics Publishing Company, Inc.)

the polymer and to monitor the respective changes in melt and glass transitions. For example, data on polymer liquid crystals have been obtained as a function of the concentration of the liquid crystal (rigid) constituent in a series of copolymers.[16]

The measurement of the volume of the polymer above and below the glass transition under high pressure is an attempt to determine the occupied volume of the polymer. As the pressure is increased, the glass transition, which occurs in the free volume, becomes generally greater and broader (Fig. 8) as the material changes between two different types of glass before the T_g is reached. This pressure sensitivity can be useful in determining the processing condition. While the Gnomix can apply pressures that greatly exceed any high pressure differential scanning

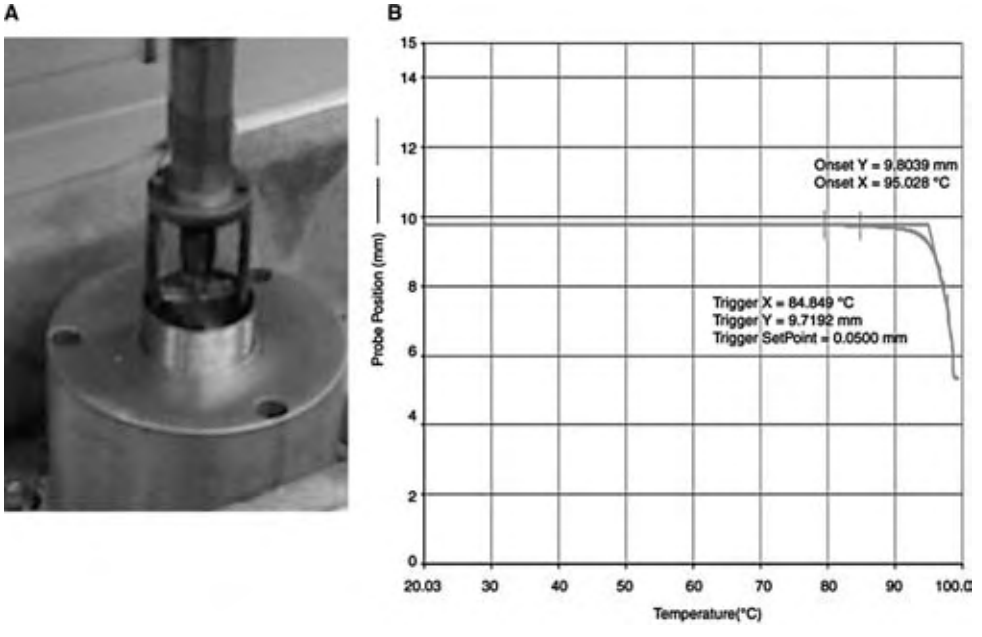


Fig. 9 Heat Distortation Tests, which look for sample flexure while the sample is immersed in oil, can be run in the TMA. (A) shows a TMA 7 with a furnace liner. (B) shows the results of the HDT runs.

calorimeter (HP DSC), the experimental times are long and DSC may be more useful for studies up to ~6.8 MPa (~1000 psi). The agreement between these techniques is quite good[16] and hence HP DSC, within its limits, is a suitable alternative in a number of cases to a P–V–T instrument.

Mechanical Tests

A wide variety of tests is performed in TMA, which are adapted from physical tests that were used before the instrument became commonly available.[17] These tests may also be modeled or mimicked in TMA, such as heat distortion (Fig. 9) and softening points.[17] Methods to obtain the modulus,[17] compressive viscosity,[18] and penetrative viscosity[19] have been developed. Many of these methods, such as ASTM D648 for example, will specify the stress the sample needs to be exposed to during the run. In D684, a sample is tested at 66 and 264 psi. Most TMAs on the market today have software available that allows them to generate stress—strain curves and to run creep—recovery experiments.[b] Some are also capable of limited types of stress relaxation studies (for example a constant gauge length test[20]).

CONCLUSIONS

While TMA is one of the older and simpler forms of thermal analysis, its importance is in no way diminished by its age. Advances in DSC technology and the appearance of dynamic mechanical analysis (DMA) as a common analytical tool have decreased the use of it for measuring glass transitions, but nothing else allows the measurement of CTE as readily as TMA. In addition, the ability to run standardized material test methods at elevated temperatures easily makes TMA a reasonable alternative to larger mechanical testers. As the electronic, biomedical, and aerospace industries continue to push the operating limits of polymers and their composites, this information will become even more important. During the last 5 years a major renewed interest in dilatometry and volumetric expansion has been seen. Other thermomechanical techniques will also likely be developed or modernized as new problems arise.

[b]See, for example, the product literature for Perkin Elmer's Diamond TMA and PYRIS™ software. The TMA and the DMA can both be used for running simple mechanical tests like stress-strain curves, creep-recovery, heat set and stress relaxation. Other vendors have similar packages.

ACKNOWLEDGMENTS

The authors wish to acknowledge the PerkinElmer Thermal Business Unit and the Material Science Department of the University of North Texas for their support and assistance. In addition, the help and advice of Professor Witold Brostow of the Materials Science Department of UNT, and especially my graduate students, John White and Dr Bryan Bilyeu, is greatly appreciated.

REFERENCES

1. Kargin, V.A.; Mikhailov, N.V. Mecanism of lengthening and shortening oriented fibers. Dokl. Akad. Nauk SSSR **1948**, *62*, 239.
2. Bird, R.; Curtis, C.; Armstrong, R.; Hassenger, O. *Dynamics of Polymer Fluids*, 2nd Ed.; Wiley: New York, 1987.
3. (a) Ferry, J.D. *Viscoelastic Properties of Polymers*, 3rd Ed.; Wiley: New York, 1980 and (b) Aklonis, J.J.; McKnight, W.J. *Introduction to Polymer Viscoelasticity*, 2nd Ed.; Wiley: New York, 1983.
4. (a) Struik, L.C.E. Physical Aging in Amorphous Polymers and Other Material; Elsevier: New York, 1978; (b) Struik, L.C.E. *Failure of Plastics;* Brostow, W., Corneliussen, R.D., Eds.; Hanser: New York, 1986; (c) Matsuoka, S. *Failure of Plastics;* Brostow, W., Corneliussen, R.D., Eds.; Hanser: New York, 1986; and (d) Matsuoka, S. *Relaxation Phenomena in Polymers;* Hanser: New York, 1992.
5. Vrentas, J.D.; Duda, J.L.; Huang, J.W. Regions of Fickian diffusion in polymer-solvent systems. Macromolecules **1986**, *19*, 1718.
6. Brostow, W.; Macip, M.A. Impact transition temperatures and polymer densities. Macromolecules **1989**, *22*, 2761.
7. Curran, G.; Rogers, J.; O'Neal, H.; Welch, S.; Menard, K.J. Comparison of the glass transition value for a graphite-epoxy composite by differential scanning calorimetry (DSC), thermomechanical analysis (TMA), and dynamic mechanical analysis (DMA). Adv. Mater. **1995**, *26*, 49.
8. Brostow, W.; Arkinay, A.; Ertepinar, H.; Lopez, B. Phase structure and phase diagrams on polymer liquid crystal systems: copolymers of PET and p-hydrobenzoic acid. POLYCHAR-3 Proceedings **1993**, *3*, 46.
9. (a) Mark, J., Ed.; *Physical Properties of Polymers Handbook*, American Institute of Physics, Woodbury, NY, 1996; (b) Beerblower, A.; Dickey, J.

Measurement of the effect of lubricating oils on engine gasket materials. Am. Soc. Lubr. Eng. Trans. **1969**, *12*.

10. Boundy, R.; Boyer, R.; Stoesser, S. *Styrene, Its Polymers, Copolymers, and Derivatives*; Reinhold Publishing: New York, 1952.
11. Snow, A.; Armistead, J.J. A simple dilatometer for thermoset cure shrinkage and thermal expansion measurements. Appl. Polymer Sci. **1994**, *52*, 401.
12. Bilyeu, B.; Brostow, W.; Menard, K. Determination of volume changes during cure via void elimination and shrinkage of an epoxy prepeg using a quartz dilatometer cell. Polimery **2001**, *46*, 794.
13. Speyer, R. *Thermal Analysis of Materials*; Marcel Dekker: New York, 1993, 165 pp.
14. Zoller, P.; Fakhreddine, Y. PVT Studies: a versatile tool for polymers. Thermochim Acta **1994**, *238*, 397.
15. Zoller, P.; Walsh, D. *Standard Pressure–Volume–Temperature Data for Polymers*; Technomic Publishing Co.; Lancaster, 1995.
16. Berry, J.; Brostow, W.; Hess, M.; Jacobs, E. PVT relationships in a series of longitudinal polymer liquid crystal composites with varying mesogen concentration. Polymer **1998**, *39*, 4081.
17. (a) Cassel, R.; Fyans, R. Applications of thermomechanical analysis in product development. Industrial Research **1977**, Aug44; and (b) Cassel, R. Beakley, D. Use of thermal analysis methods in foam research and development. Thermal Application Study **1978**, *29*.
18. (a) Webber, P.; Savage, J.J. Measurement of the viscosity of a chacogenide glass by the parallel plate technique. Mater. Sci. **1983**, *23*, 783; (b) Tong, H.; Appleby-Hougham, G.J. Predication of thermoset viscosity using TMA. Appl. Polymer Sci. **1986**, *31*, 2509.
19. McLin, M.; Angell, A. Viscosity of salt-in-polymer solutions near the glass transition temperature by penetrometry. Polymer **1996**, *37*, 4703.
20. Daley, C.; Menard, K. Measurement of shrinkage forces of synthetic fibers held at constant length during solvent exposure. North American Thermal Analysis Society Notes **1994**, *26*, 56.

Thermosets: Materials, Processes, and Waste Minimization

Kwok-Wai Lem
Sean A. Curran
Steve Sund
Mina Gabriel
Honeywell International Inc., Morristown, New Jersey, U.S.A.

INTRODUCTION

One of the most significant developments in materials technology was the discovery of polymers. Over the last century, they have played an increasing role in a wide range of industries. ''Cheap plastic'' polymers have evolved into ''high technology composites'' for a myriad of structural applications. Polymers can be broadly classified into two major groups: thermoplastics and thermosets. Thermoplastics soften reversibly, can be easily deformed (or high damage tolerance), and made to flow when heated. They then harden to retain their shape on cooling from the liquid to the solid state, and may be reprocessed by reheating. Being able to reprocess the polymer is a significant economic advantage for thermoplastics. However, in use, thermoplastic polymers are susceptible to creep and deformation.

In contrast, thermosets harden irreversibly during processing. They form permanent crosslinks, i.e., covalent bonds between polymer chains that can only be broken by degrading the material. Heat, light, electricity, or catalysts to name a few can initiate such crosslinking reactions. Regardless of how crosslinking is initiated, the formation of permanent crosslinks makes it impossible to reprocess the material. This means that any mistake during processing generates waste. However, the advantages of thermosets in applications requiring complex shapes, high stability, structural integrity, or very high strength justify the care needed to avoid excessive waste.

The combination of processability and performance has driven significant growth in thermoset applications where dimensional stability and accuracy are critical. For example, thermosets (usually UV or e-beam cured) are the materials of choice as the ''medium'' for rapid prototyping. Similarly, the combination of processability, dimensional stability, and strength have driven the use of thermosets for automotive components, structural members in aircraft, and braking systems in both automotive and aircraft applications.

OVERVIEW OF THERMOSET TECHNOLOGY

Since the introduction of the first commercial thermoset, Bakelite, based on phenol formaldehyde condensation, a wide range of thermoset materials have been introduced. These are typically designed for specific properties related to their chemistry and processability. Some commercially important thermosets include phenolics, ureas, melamines, epoxy resins, unsaturated polyesters, silicones, rubbers, polyurethanes, acrylics, cyanates, polyimides, and benzocyclobutenes.[1]

Thermosets are particularly effective in the manufacture of large components. They and their composites are formed from precursors such as reactive monomers or low molecular weight oligomers. Because of their low molecular weight, these precursors flow easily, and can fill large molds without requiring large injection pressures. Once the mold is filled, the precursor crosslinks to form a solid product. While the part may shrink during cure, proper design allows for this shrinkage.

Typically, thermoplastics, because of their high melt viscosity, require higher temperatures and injection molding pressures, or expensive mold fabrication costs to design multiple gate systems and therefore are not cost effective materials for large parts. In addition to cost, the use of multiple gates may also lead to a less structurally sound part as a result of knit lines.

Thermosets have also found wide applicability where there is a need for high strength, low density, good heat resistance, and electrical performance. These properties can be enhanced by the addition of fillers, such as fibers, minerals, and glass, to form a composite structure. Alternatively, interpenetrating networks (IPN) can be formed to take advantage of the beneficial properties of two network polymers simultaneously. Typical fiber-reinforced thermosetting composites uses are automotive body panels, under-the-hood automotive applications, household goods, breaker switch boxes in the electrical industry, etc.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007991

Manufacturing, Processing, and Design

The first step in manufacturing is usually a process to wet the fillers and disperse them in the resin. This is followed by preparation of a preform which can be then cured or prepregs which can be laminated together. For example, thermoset tapes can be laminated with the direction of filler orientation varying between layers to distribute the strength of the reinforcement uniformly. The resins in these prepregs can be semi-solid (i.e., wet lay up) or solid (dry lay-up). In either case, the application of heat enhances resin flow, binding the layers together. Ultimately the entire layer structure becomes a single fused mass on complete cure.

The typical manufacturing processes for thermoset composites are summarized in Table 1 They may be divided into three broad categories: (1) manual; (2) semiautomatic; and (3) automatic.[2,3] Manual processes include hand lay-up, spray-up, pressure bagging, and autoclave molding. Semiautomatic processes include cold pressing, hot pressing, and compression molding of SMC (sheet molding compound) and BMC (bulk molding compound), and resin injection. Automatic processes are those such as pultrusion, filament winding, centrifugal casting, resin transfer molding, injection molding, reaction injection molding, and short cycle compression molding.

Fig. 1 is a schematic diagram of various composite-manufacturing processes following the design methodology process map given in Fig. 2. The description of some of these processes is summarized in Table 1. Karger-Kocsis[4] provided a very good overview of the manufacturing methods for thermoplastic composites, illustrating the inter-relationship between matrix, reinforcement, and preform/prepreg (Fig. 3). Their work can also apply to thermoset composites. Furthermore, Karger-Kocsis' scheme can be further extended to design the process and the final products with a simple combinatory approach as outlined in Table 2. For example, the combination route in line 1 of Table 2 corresponds to the use of chopped or continuous fiber (A1 in Fig. 3) combined with polymer (D1 in Fig. 3), processed to formulate the desired intermediate, and then processed to make the final part. Thus, a variety of combinations and approaches achieve the same nominal result. It is incumbent on the scientist and engineer to optimize for properties and cost.

The most significant aspect of a well-designed fiber-reinforced composite is that its end use performance may be predicted from the properties of each constituent.[5] In recent years, Quality by Design and Design for Six Sigma have advocated a set of predictive methodologies and protocols that ensure the delivered goods meet the market demands.[6–11] Elastic constants can be predicted for continuous reinforced composites using the Rule of Mixtures; discontinuous or chopped fiber-reinforced composites can also be modeled very well. Predictive relationships between constituents making up the composite and the resultant mechanical properties are well known. These methods are, therefore, invaluable additions to the Quality by Design and Design for Six Sigma efforts.

With advances in computer technology in recent years, simulation has often provided a very powerful tool to predict the end-use performance from the selected component properties and to establish structure-performance relations with a certain degree of confidence. This approach has been termed ''in silico'' design as opposed to in vitro or in vivo.[12–14] In silico experimentation – the ability to explore simulations rapidly reduces the need for in vivo experimentation: slow, multiple iterations of trial and error for design of new composite structures. The combination of design tools and computer modeling will drive development of new thermosets, with improved performance. This includes new structures based on known materials and the development of new chemistries for selected applications.

Optimization of the cure cycle is inextricably linked to the design of processes for specific applications. Typically, this involves a clear understanding of the influence of temperature and time used to cure the thermosets. This is key to predicting resin flow and the degree of cure.[15–19] Consequently, much attention has been placed on the curing process/morphology/structure relationship of thermosets.[20–22]

Resins and Fillers

Fig. 2 illustrates the selection process that may be used to design a composite material with specific, targeted performance characteristics. Thermosets have been the backbone of the advanced composites industry since it has gained commercial acceptance in the early 1960s. We can segment composites (Fig. 4), into two major classes: (1) high volume composites, i.e., commodity products and (2) high performance composites.[4] The type of matrix, composition, and fiber/filler interphase determines the performance of a composite. The concentration, type, and orientation of the fiber reinforcement affect the strength and stiffness. Random mat or chopped fibers are frequently used in high volume composites at maximum fiber loading (>30% by volume). Continuous oriented fibers are required for high performance composites. In both cases, performance depends strongly on the structure and performance of the component parts.

Table 3 summarizes the key properties of selected major types of thermosetting resins.[1,23–25] Their cross-linking reactions can occur by radical chain addition

Table 1 Typical molding processes in thermoset composites

Molding process	Description
Prepreg molding	Prepreg molding is in many respects the next step up from hand lay up. With prepreg molding the resin content of the finished component can be accurately controlled, which cannot always be said for hand lay up. Also, woven or unidirectional fiber reinforcements are used, rather than chopped strand mat. The reason is that they be aligned in the required orientation.
Compression molding	Compression molding is the most common method of forming thermosetting materials. It is simply the squeezing of a thermosetting preform into a desired shape by application of heat and pressure to the material in a mold at a desired duration time. Molding preform mixed with fillers and chopped fibers to strengthen the finished product, is put directly into the open mold cavity. The mold is then closed, pressing down on the preform and causing it to flow throughout the mold. It is while the heated mold is closed that the thermosetting material undergoes a chemical change which permanently hardens it into the shape of the mold.
Transfer molding	Transfer molding is most generally used for thermosets. This method is like compression molding in that the preform is cured into an infusible state in a mold under heat and pressure. It differs from compression molding in that the thermosetting preform is heated to a point of plasticity before it reaches the mold and is forced into a closed mold by means of a hydraulically operated plunger. The molten molding material in transfer molding flows around these metal parts without causing them to shift position. Transfer molding was developed to facilitate the molding of intricate products with small deep holes or numerous metal inserts.
Vacuum-assisted molding	In conventional SMC processes, charge loading and press closure speeds are selected with the primary aim of forcing out entrapped air. With vacuum assistance, the charge is selected to rapidly cover the mold surface and provide faster closure speeds without entrapping air. The increase in mold surface coverage by the charge has several additional benefits such as elimination of wave patterns and flow lines, better localized strength of components due to better retention of fiber orientation.
Injection molding	Injection molding is a major process in making thermoplastic materials. With proper modifications, the injection process is sometimes used for thermosetting plastics. In injection molding, material is put into a hopper which feeds into a heating chamber. A plunger pushes the material through this long heating chamber, where the material is softened to a fluid state. At the end of this chamber there is a nozzle which abuts firmly against an opening into a cool, closed mold. The fluid material is forced at high pressure through this nozzle into the cold mold. As soon as the material cools (in thermoplastics) or heated (in thermosets) to a solid state, the mold opens and the finished part is ejected from the press.
Resin transfer molding	In the resin transfer molding (RTM) process, a low viscosity resin is transferred into a closed mold containing all the appropriate reinforcements and inserts as a preform. The air is normally evacuated from the mold, allowing the use of low resin injection pressures and epoxy molds. Manhole covers, compressor casings, car doors, and propeller blades have all been manufactured by RTM.
Reaction injection molding	Reaction injection molding (RIM) is a relatively new processing technique that has rapidly taken its place alongside more traditional methods. Unlike liquid casting, the two liquid components, polyols and isocyanates, are mixed in a chamber at relatively low temperatures (75°–140°F) before being injected into a closed mold. An exothermic reaction occurs, and consequently RIM requires far less energy usage than any other injection molding system. The three major types of polyurethane RIM systems are rigid structural foam, low-modulus elastomers and high-modulus elastomers. Reinforced RIM (R-RIM) consists of the addition of such materials as chopped or milled glass fiber to the polyurethane to enhance stiffness and to increase modulus, thus expanding the range of applications.

(Continued)

Table 1 Typical molding processes in thermoset composites *(Continued)*

Molding process	Description
Pultrusion	In the pultrusion process, dry reinforcements are impregnated with a specially prepared low viscosity liquid resin system and drawn through a die heated to about 120–150°C where curing occurs. The solid laminate, which has assumed the shape of the die, is withdrawn by a series of haul off grippers, and is cut to length or coiled. Pultrusion is unique among the processes under consideration in that it is capable of producing complex components on a continuous basis. The process can basically produce any shape that can be extruded. It is also not allied to any one industry and applications range from civil engineering to electrical. These factors combine to give pultrusion one of the highest predicted growth rates of all composite processes.
Filament winding	Filament winding is a process of high speed precise laying down of resin impregnated continuous fibers on to a mandrel. The mandrel can be any shape. Pressure vessels, pipes, and drive shafts have all been manufactured using filament winding. Multi-axis winding machines can also be used. The process is usually computer controlled and the reinforcement can be oriented to match the design loads. The fibers may be impregnated with resin before winding (wet winding), pre-impregnated (dry winding), or post impregnated. Wet winding has the advantages of using the lowest cost materials with long storage life and low viscosity. The prepreg systems produce parts with more consistent resin content and can often be wound faster.

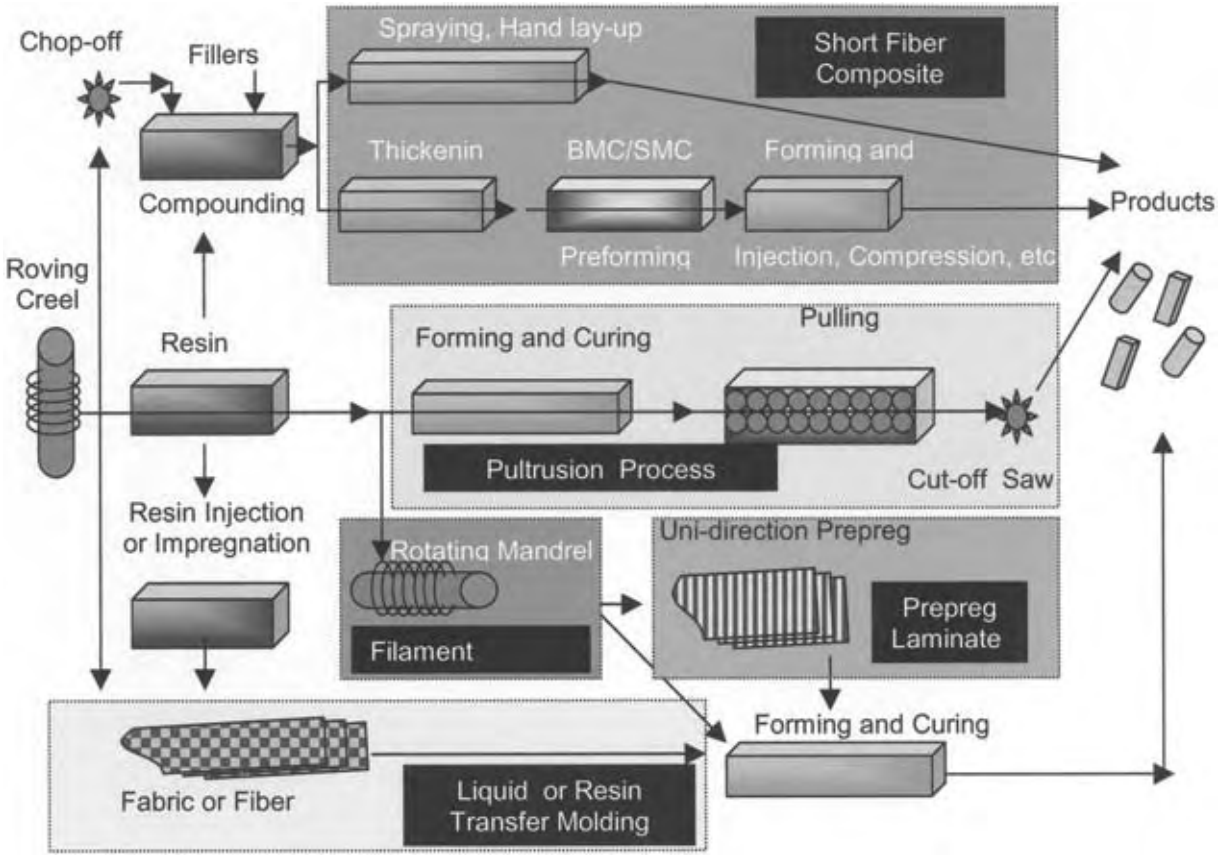


Fig. 1 Schematic diagram of the various typical composite manufacturing processes.

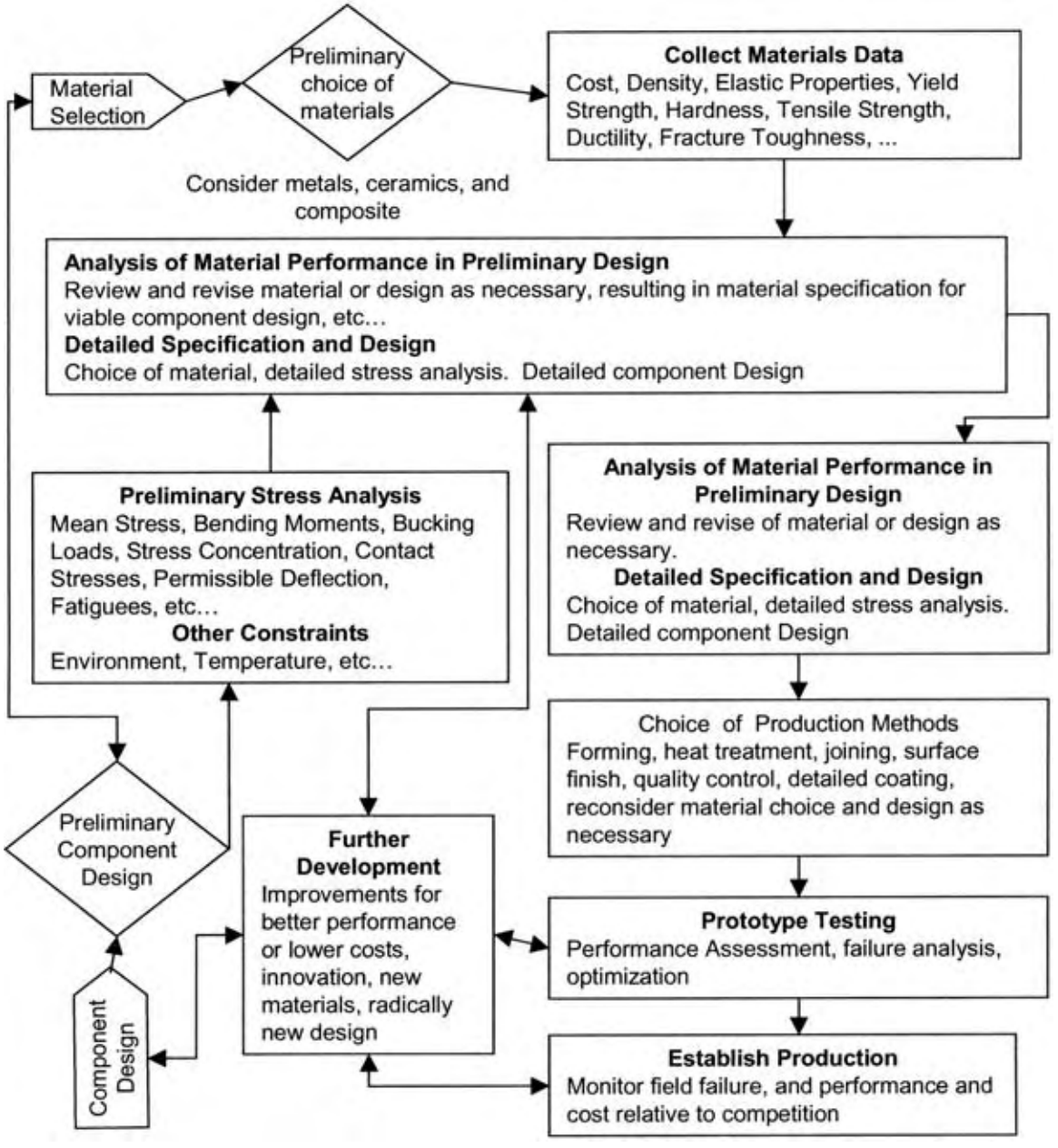


Fig. 2 Design methodology. (From Refs.[4,5].)

(unsaturated polyester resins, vinyl ester resins, benzocyclobutene, etc.), ionic ring opening (epoxies), condensation (urethanes, phenolics, polyimides, such as bismaleimides), and cyclotrimerization (cyanates, isocyanurates, etc.).

Table 4 compares the mechanical properties of various organic fibers, graphite/carbon fibers, ceramic fibers, and glass fibers, and lists commercially important applications. One major application of organic fibers, such as Spectra® and aramid fibers, is in ballistic armor.[26,27] In contrast, graphite, ceramic fibers, and glass fibers are primarily used in structural applications. One of the reasons that organic fibers find little structure applications is because of their poor compressive strength. (see Table 4 in Ref.[28].)

CHARACTERIZATION

The final morphology of a network system depends greatly on the flow properties and phase separation of the reactive components as seen in Fig. 5.[16–18,29] Naturally, design of appropriate processes requires the ability to directly investigate the progress of reaction. This includes monitoring the chemistry of cure with spectroscopic techniques such as FTIR, NMR,

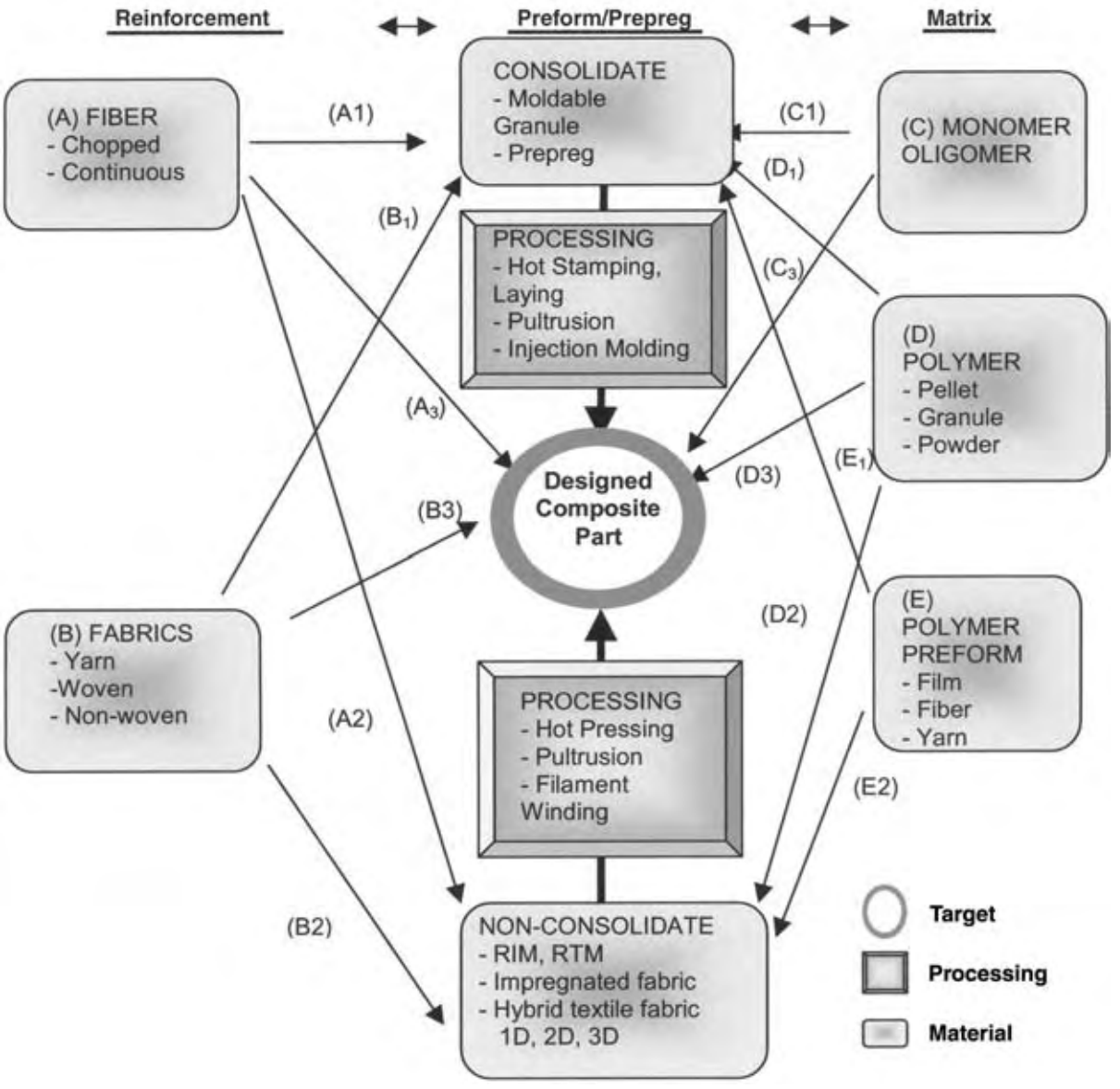


Fig. 3 Interrelationship of reinforcement-preform/prepreg-matrix. (From Refs.[4,29].)

UV-Vis, thermal analysis such as DMA, TMA, or rheological measurements. Fig. 5 illustrates the application of rheological techniques to monitor cure. Fig. 5A shows the relation of viscosity versus extent of reaction for a thermoset system, whereas Fig. 5B gives the relation of temperature versus extent of reaction in a liquid crystal thermoset. As an illustration, the effect of curing temperature on the structure of a cured liquid crystalline thermoset is given in Table 5.[30] Gillham[31] has postulated that the curing of a thermoset resin can be expressed in terms of a time–temperature transformation diagram (TTT), in which the entire curing cycle is represented by the resin rheology as a function of time and isothermal temperature. Similar to Fig. 5, the TTT diagram is divided into four sections: liquid, gelled rubber, ungelled glass, and gelled glass. This concept has found application to describe the structure–property relationship of thermoset epoxy resins using torsional braid analysis.

RECENT ADVANCES

With recent advances in the mixing of different types of thermosets and/or thermoplastics, the technology of IPN has provided very unique properties and performance.[32–34] IPN is a method to produce very special network polymer blends. Two materials which do not react with each other are blended and subsequently cured in place. The blended network polymers

Table 2 A combinatory approach to establish the matrix-fiber-process relationship

Combination route	Process	(Intermediate) products	Further processing to final part	Potential process to make helmets (given in the example)
(A_1 + D_1)	Melt compounding Melt impregnation Powder impregnation	Chopped fiber-reinforced granule UD prepreg, tape UD fiber tow	Injection, extrusion Molding Tape laying and winding Filament winding	Yes
(A_2 + E_2)	Commingling Interweaving, braiding	Mixed yarn Textile fabric	Filament winding, pultrusion Hot pressing, pultrusion	Yes
(A_3 + C_3)	Chopped fiber Reinforced reaction Injection molding, RRIM	—	No	Yes
(B_1 + E_1)	Film stacking	Glass mat-reinforced sheet GMT, SMC, TMC	Thermoforming Hot pressing	Yes
(A_2 + D_2)	Powder or slurry impregnation	Impregnated fleece	Thermoforming	Yes
(B_2 + D_2)		Impregnated textile fabric	Hot pressing	
(B_2 + E_2)	Tufting, needle-punching	Hybrid textile fabric	Thermoforming Hot pressing	Yes
(B_3 + C_3)	Structural RIM	—	No	Yes
(C_3 + D_3)	Molding of self-reinforced polymers	—	No (unsolved process for molecular composites)	??

UD, unidirectional; GMT, glass mat thermoplastic; TMC, thick molding compound; SMC, sheet molding compound.
(From Refs.[4,29].)

share the same region of space (macroscopic volume) of the sample. Thus one has an intimate mixture of two normally immiscible systems. Consequently, the IPN exhibits a combination of properties from the two networks. This approach can generate unique advantages such as crack resistance,[34] improved modulus, strength, and hydrolytic resistance.[35] One can even generate organic/inorganic hybrid IPNs to impart specific properties such as permeability reduction.[36]

Advances in liquid crystalline thermosets in recent years,[30,37–40] have resulted in some fascinating thermosetting systems. Control of the liquid crystal type and morphology allows one to generate a wide range of properties. For example, the properties can be tuned from electrical insulators to conductors with addition of conductive fillers, while maintaining desirable high temperature performance.[39] In addition, can provide control of the degree of shrinkage on cure.[40]

Another area of current research is development of nanocomposites. Incorporation of nanometer-sized fillers in thermosets allows even broader application of thermoset technology. The processing techniques are quite similar to conventional thermosets, but the effect of filler can be quite dramatic. For example, incorporation of only 2.5% by weight of nanometer-sized silicates in a cyanate resin led to a marked improvement in physical and thermal properties while imparting a 30% increase in both the modulus and toughness.[29,41]

As discussed earlier, while the scale of the fillers is substantially different, nanocomposite materials concepts and technology are very similar to those of conventional composite materials. This is clearly demonstrated in the case of new thermosets for nonlinear optical (NLO) applications,[37,38,42,43] where nanocomposite of liquid crystalline thermosets, IPNs, and simple filled thermosets are evaluated. Tripathy et al.[42] discussed four different ways to prepare nonlinear optical polymers. (1) The polymer matrix is doped with NLO moieties in a guest/host system; (2) In side-chain polymer systems, NLO polymers with active moieties are covalently bonded as pendant groups; (3) In the main chain polymer, the chromophores are incorporated as parts of the main polymer backbone to enhance the temporal stability of the NLO properties; and (4) Stability of the optical noninearity in sol–gel-based thermosets is related to

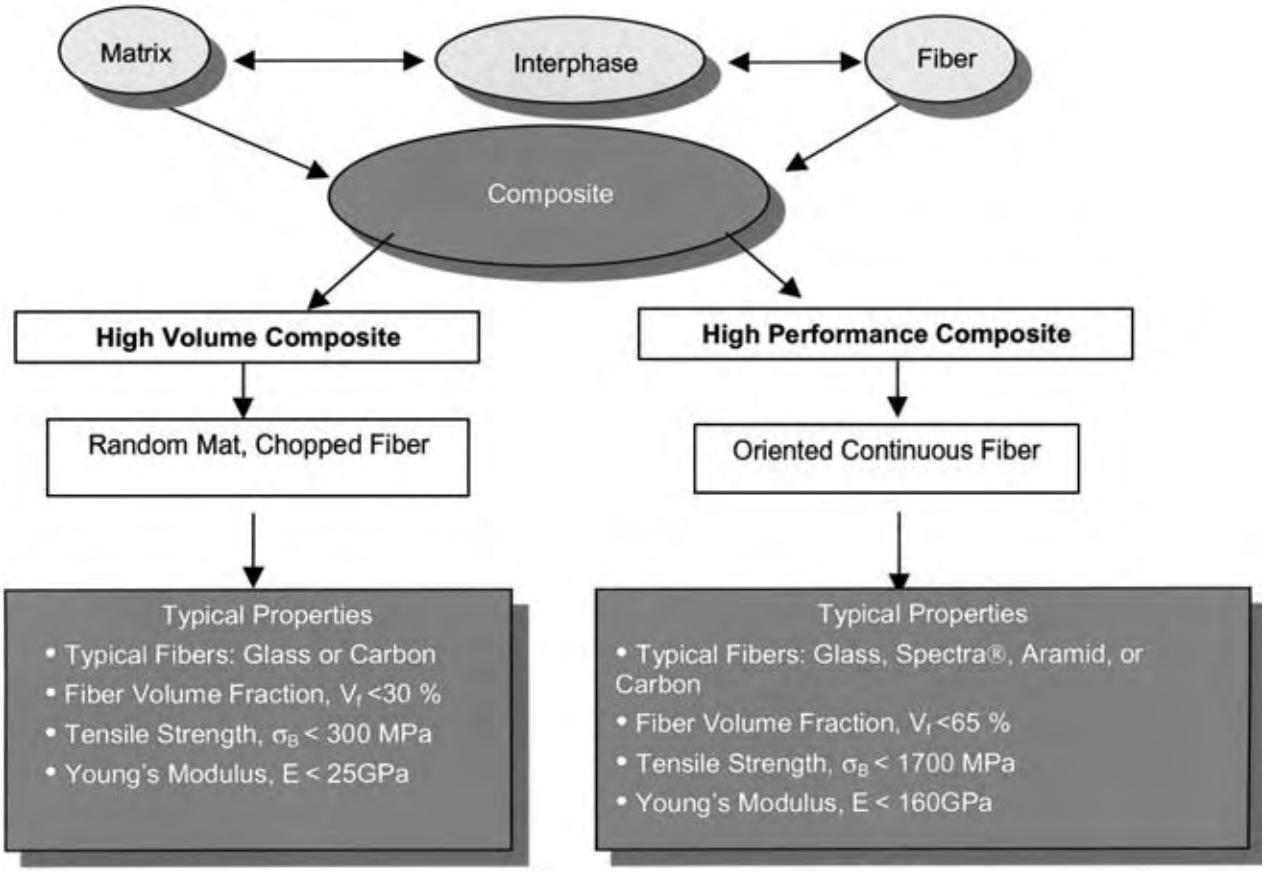


Fig. 4 Classification of composites. (From Ref.[29].)

the increasing crosslinking density. It was found that sol-gel technology can be used to prepare three-dimensional network glasses of optically clear and low loss properties at high temperatures.

The final structure can be further manipulated by employing IPN technology and/or using the AC electric field technique.[38] It has been reported that the combination of high T_g (glass transition temperature) polymers such as polyimides, high crosslink density, and permanent chain entanglement results in high temporal stability at elevated temperatures.[42,43]

By varying the structure of various component groups of the molecule, one can understand how to control the liquid crystallinity and the resultant properties of the materials (see the Fig. 5B).[44] A successful development of a practical device for NLO application will require optimization of properties in all aspects in materials and processes, such as optical loss, optical power handling, processability, and reproducibility.

WASTE MINIMIZATION

Finally, waste generation is a critical limitation for any industrial process.[45] Waste generation is unavoidable so waste minimization becomes a fundamental requirement for economic feasibility. Waste generation inherently decreases the overall value of any material. A clear understanding of the materials and processes is central to waste reduction. For example, development of (1) Laboratory definition of a robust model for cure kinetics lead to; (2) robust on-line cure monitors, which coupled to; (3) robust process controllers for manufacturing can significantly reduce the waste in production of glass filled epoxies.

Senge and Carstedt[45] recently offered a view in Fig. 6 of why industry produces waste. They suggested that a synthetic process can emulate nature to reduce the waste using a cyclic industrial system. This is accomplished by focusing on three key aspects of the manufacturing process: (1) resource productivity; (2) cleaning products; and (3) remanufacturing, recycle, and compost. Clearly in Fig. 6, preventing the waste generation from production, use, and disposal in the first place can eliminate the waste. One example is the industrial recycling of nylon 6 carpet patented by Honeywell (formally AlliedSignal), Inc. and DSM.[46,47] This process has addressed and overcome the economic, technical, and logistical barriers to commercialize a closed-loop recycling process to recover caprolactam

Table 3 Typical thermosetting resins of commercial applications

			Thermal properties					Mechanical properties		
Resin	**Features**	**Curing reaction**	**UUT, °C**	**MCUT, °C**	***Tg*, °C**	**TC, W/m.k**	**CTE, 10^{-5} °K**	**E, MPa**	**σ, MPa**	**K_{IC} J/m²**
Unsaturated polyester	Low viscosity, easy processing, lowest cost, limited mechanical and thermal properties	Radical	120		80–140					
Vinyl ester, acrylics	Similar to unsaturated polyester, but with improved properties and chemical resistance, and have higher cost	Radical	120		100–160					
Phenolics	Low cost, good high temperature resistant, fair mechanical properties 1. Novolacs–require additional formaldehyde for further curing 2. Resoles–Selfcondensation, thermally activated or acid catalyzed	Condensation	180	110	—	0.35	3–5	5600–12,000	69–207	
Epoxies	Very versatile. Wide range of formulations available for whole range of processes and applications. Excellent balance of mechanical, electrical, thermal, and environmental resistance properties. Poor toughness without impact modification	Ionic ring-opening chain polymerization catalyzed by tertiary amine (anionic) or BF_3 complexes (cationic) Stepwise reactions with curing agents such as amines, acids or anhydrides, and phenols	180	130	170–206	0.88	1.1–3.5			54

(Continued)

Table 3 Typical thermosetting resins of commercial applications *(Continued)*

Resin	Features	Curing reaction	Thermal properties					Mechanical properties		
			UUT, °C	MCUT, °C	*Tg*, °C	TC, W/m.k	CTE, 10^{-5} °K	E, MPa	σ, MPa	K_{IC} J/m^2
Cyanates	Epoxy-like processing at higher temperatures. Higher cost, low dielectric, good mechanical properties and toughness	Self-cyclotrimerization catalyzed by transition metal carboxylates and nonylphenol	180–220		254–280					60–210
Polyimides	BMI: Convenient processing, but at high temperatures. Very brittle without impact modification. Very good hot-wet performance	Chain polymerization through the double bonds in maleimides. Stepwise additions with, or allyl or propenyl modifiers	200–230	400	240–300			3700		400
	PMR: Excellent high temperature resistance, more brittle to BMI, uses of toxic component	Condensation of amine with acid ester to form amide-acid, then imide, followed by crosslinking through norbornene end-group	290–315							200

Upper use temperature, UUT (°C); maximum continous use temperature, MCUT (°C); thermal conductivity, TC (W/m.k); coefficient of thermal expansion, CTE (10^{-5} °K); tensile modulus, E (Mpa).

Table 4 Typical properties of reinforcement fibers and their commercial applications

	Spectra® (PE)[a]		Aramid		Carbon		Glass-S	Boron	Al_2O_3
			LM	HM	HS	HM			
Yarn specifications, d/fil	900	1000	1500/1000	1500/1000	1730/5100	1630/3000			
Filament diameter, μ	38	27	12	12	7	7	7		
Density, g/cm^3	0.97	0.97	1.44	1.44	1.81	1.81	2.50	2.5–2.6	3.7
Tensile strength, Gpa	2.59	3.00	2.80	2.80	3.10	2.40	4.6	3.5	415
Tensile modulus, GPa	117	172	62	124	228	379	90	1.7	350–380
Tensile elongation,%	3.5	2.7	3.6	2.8	1.20	0.60	5.40		
Specific tensile strength, 10^6 in.	10.7	12.4	7.8	7.8	6.8	5.4	7.4	4.3–4.5	1.8
Specific tensile modulus, 10^8 in.	486	714	173	346	507	846	140	309–321	380–412
Compressive strength, GPa			0.32–0.46		0.48–2.8		1.1	5.9	6.9
Commercial applications									
Structural	??		Significant		Very significant		Very significant	Significant	Very significant
Armor	Very significant		Very significant		??		Significant	??	??
Tire/elastomer	??		Very significant		??		Significant	??	??
Others	Very significant		Very significant		Very significant		Very significant	OK	Very significant

[a]Spectra® is a registered trademark of Honeywell International.
(From Refs.[26,33].)

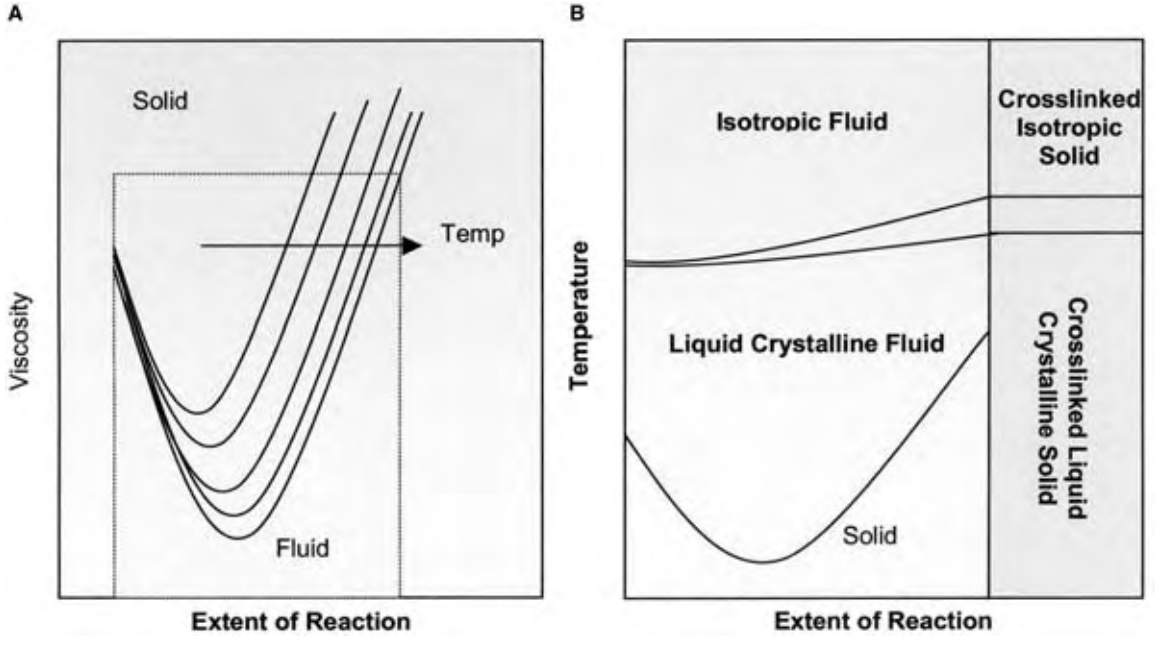


Fig. 5 Schematic diagrams of effect of extent of reaction on viscosity and morphology in curing of thermosets. (A), Effect of extent of reaction on viscosity. (From Refs.[11–14].) (B) Effect of extent of reaction on morphology. (From Refs.[30,37,38].)

Table 5 Effect of curing temperature on the structure

Liquid crystalline thermoset system:
1. Liquid crystalline thermoset prepolymer: dihydroxy methylstilbene epoxy (Mn, 3600)
2. Curing agent: 4,4′-methylene dianiline

Sample	Curing temperature,°C	Molecular order	Clearing transition temperature,°C
Liquid crystalline thermoset epoxy prepolymer	Not applicable	Nematic to isotropic	176
Cured isotropic thermoset	190	Thermoset	157
Cured liquid crystalline thermoset	140	Nematic	192

(From Ref.[30].)

from waste nylon 6 articles. Nylon carpet is a composite structure containing 45% nylon fiber on a calcium carbonatefilled polypropylene backing. The fiber and backing are held together by a cure thermoset styrene butadiene (SBR) latex adhesive. The waste product derived from the medium-pressure depolymerization of nylon 6 carpet from this process is a composite of 65% calcium carbonate and 18% polypropylene dispersed in

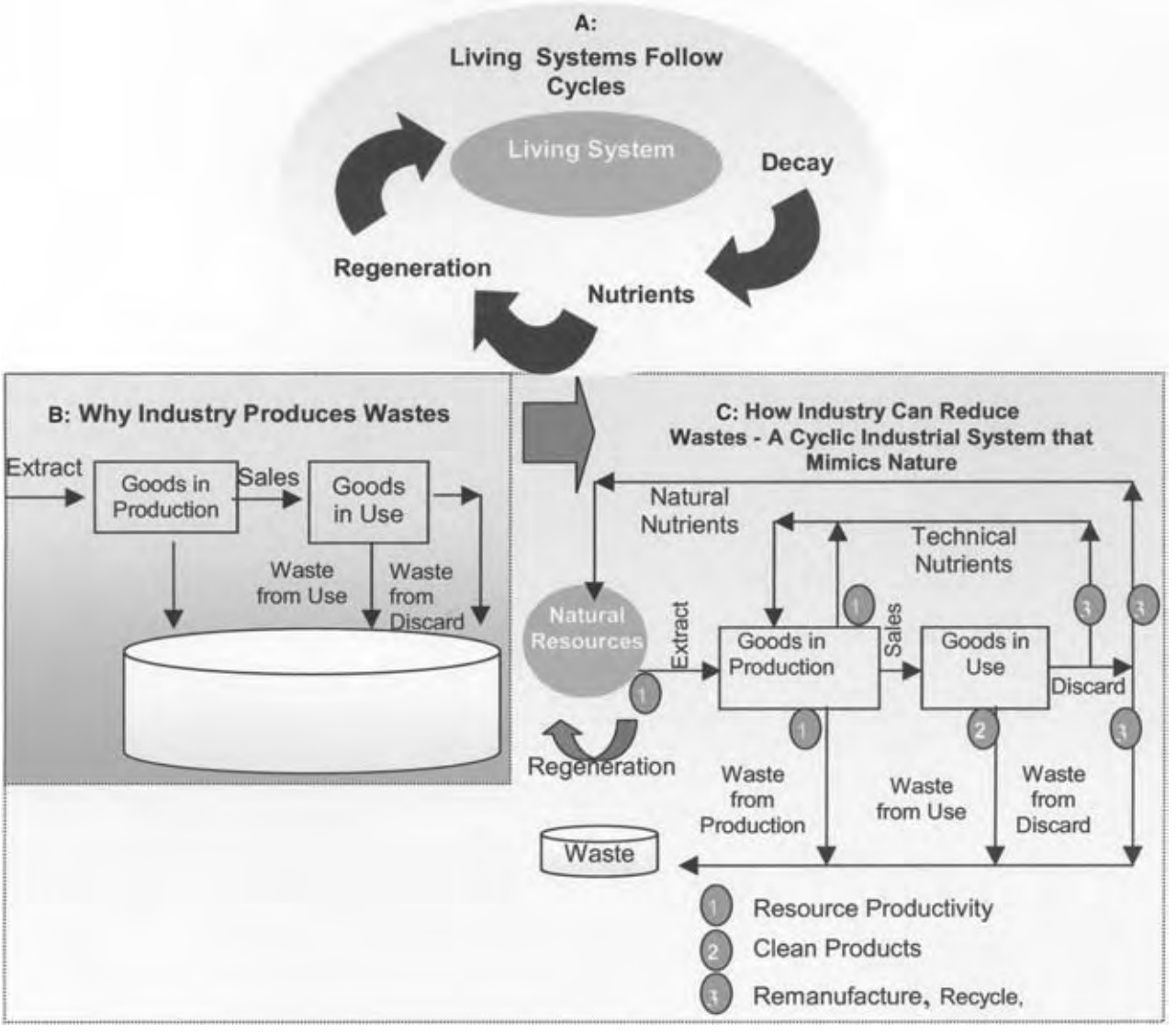


Fig. 6 Waste generation and reduction from current industry. (From Ref.[45].)

a styrene butadiene thermoset. This residue material can be used as a filler in road asphalt, asphalt roof membranes, molding compounds, and plastic lumber.[48,49] Thus, we have a real example of the closed loop recycle process as postulated by Senge and Carstedt.

Gutowski[50] has examined the product induced material flows through the product manufacturing system, and has suggested several research strategies to reduce material-related environmental loads. Fig. 7A gives a typical schematic of a materials processing flow map. Fig. 7B gives a typical fabrication process of flat laminate sheet found commonly in composite manufacturing.[50] In Fig. 7B, we can see that waste can be generated from the raw materials supply, from each step in the process, and from scrap. Sometimes, the total waste generated from these processes could be as high as 25% based on the raw materials.

The net flow balance in Fig. 7A, can be described by Eq. (1):

$$\begin{aligned}\text{Performance} = {} & af(\text{critical component selection} \\ & - \text{its wastes}) \\ & + bf(\text{process} - \text{its wastes}) \\ & + cf(\text{structure} - \text{its wastes}) \\ & + df(\text{products} - \text{its wastes}) \\ & - \text{performance wastes}\end{aligned} \tag{1}$$

$$\begin{aligned}\text{Performance} = {} & af(\text{critical component selection}) \\ & + bf(\text{process}) \\ & + cf(\text{structure}) \\ & + df(\text{products}) \\ & - \Sigma_{\text{wastes}} \text{ all sources}\end{aligned} \tag{2}$$

where a, b, c, d are the constants. In term of a continuous flow process, we can rewrite Eqs. (1) and (2) into Eq. (3)

$$P(x) = \int \lambda_j V(x_j)dx_j - \int \omega_j W(x_j)dx_j \tag{3}$$

where, $P(x)$ is a value performance function, $V(x_j)$ is the value generating function at component x_j stage or phase j, and $W(x_j)$ is the waste generating function at component x_j, and λ_j, and ω_j are constants. The variation of $V(x_j)$, $W(x_j)$, λ_j, and ω_j greatly affects the value of $P(x)$. From Eq. (3), we can see that wastes generated from the materials flow negatively to impact a wide range of factors. These range from the selection of materials and from innovation design to environmental concerns as seen in Figs. 6 and 7.

$V(x_j)$ is not restricted with any limitations in Eq. (3) and it is valid to include the feedback loops in Fig. 6. Thus, both processes in Figs. 6 and 7 can be described by Eq. (3) with different forms of $V(x_j)$ and $W(x_j)$, and different values of λ_j and ω_j.

Furthermore, this form of expression is not new and it has found uses in many applications in mathematics and science.[51–62]

The definition of wastes can be viewed as the gap(s) between the full (theoretical) potential of the resources supplied and the actual delivered. As can be seen in the examples in Table 6, the defects found in a fiber structure impair our ability to obtain much less than half of its theoretical tensile modulus.[28] We need to point out that the concepts in Figs. 6 and 7, and Eqs. (1)–(3) have been the background of the so called ''Six Sigma'' methodology used to minimize the wastes in manufacturing industries over the last decade.[6–10] Six Sigma is a strategy that was developed to accelerate improvements in processes, products, and services, radically reduce manufacturing and/or administrative costs, and improve quality by relentlessly focusing on eliminating waste and reducing defects and variation. Since the material processing is a very broad field, we have limited our focus in the discussion to thermosets and their fiber-reinforced composites. However, the same principle will apply for any materials development effort.

PROPERTY OPTIMIZATION CASE STUDY

We discussed ''in silico'' experimentation above. In particularly demanding applications, it is far more efficient to design the composite for the application before spending excessive time on trial and error. We will demonstrate this with a case study. On the basis of the design methodology illustrated in Fig. 2 and the processes given in Table 2, we will discuss the design of a helmet for ballistic protection. Figs. 8–10 show the design and simulated deformation of a helmet under ballistic impact.

Honeywell International Inc.'s Spectra Shield® composite has been used as the ballistic material for a lightweight helmet system since the mid-1990's. Spectra Shield composite is made from Honeywell's Spectra® fiber, one of the world's lightest and strongest fibers (see Table 4), which is, pound-for-pound 10 times stronger than steel. The use of this material in helmet construction results in substantially increased ballistic protection, while reducing weight. Before field use of Spectra Shield® material for this application, modeling of the composite helmet under ballistic impact was performed. The modeling was carried out to highlight the increased protection under the ballistic impact of Spectra fiber compared to metal. Finite element analysis (FEA) of the helmet configuration impacted by a standard projectile round to the front

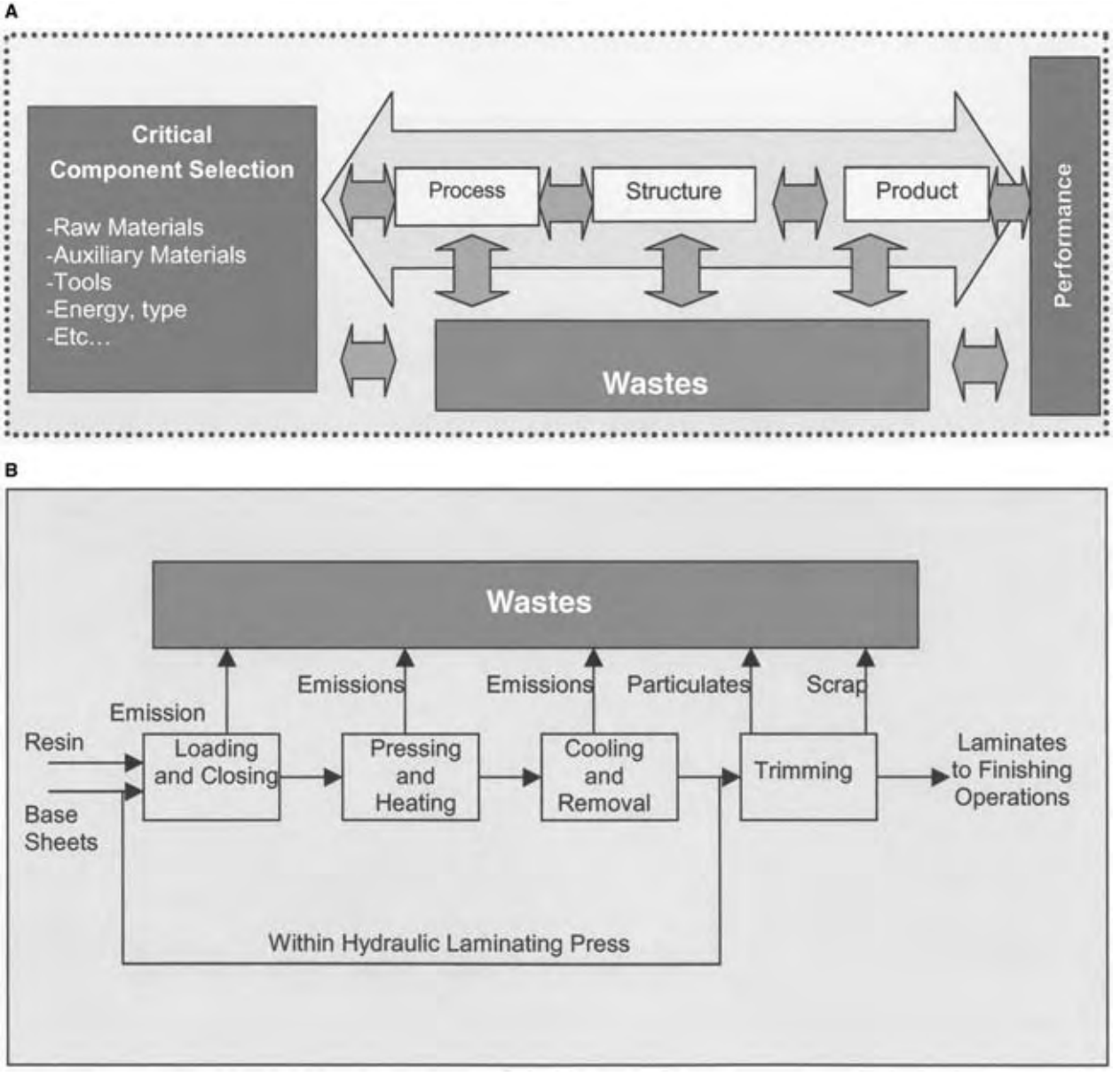


Fig. 7 Materials flow relationship. (A) Relationship of "component section–process–structure–product–performance." (B) schematic of a flat laminate sheet production process. (From Ref.[50].)

Table 6 Tensile modulus of several ordered polymers

No	Materials	Molecular structure	Theoretical (Gpa)	Actual (Gpa)	Gap	
					GPa	%
1	Poly(p-phenylene-2,6-benzo[1,2-d:45-d′] bisoxazole (PBO)	Cis	730–670	360	370–310	55–43
		Trans	707–620		347–260	60–37
2	Poly(p-phenylene-2,6-benzo[1,2-d:45-d′] bisthiazole (PBZT)	Cis	610–600	325	285–275	48–45
		Trans	605–525		280–200	53–33
3	Polyethylene		360–320	172–117	243–148	76–41
4	Graphite		1500	600–70	1430–900	95–60

(From Ref.[28].)

Fig. 8 Typical layout of a helmet.

Fig. 9 Helmet design for finite element analysis.

of the helmet was performed using the Abaqus FEA program (Abaqus Inc. Pawtucket, Rhode Island, U.S.A.).

A FEA analysis can be linear or nonlinear, but for our application nonlinear effects are significant and have to be taken into account. In a finite element analysis there are at least four primary causes of nonlinearity.

1. *Material nonlinearity.* This nonlinearity is present because of the fiber orientation in the composite. The effect of the Spectra fiber reinforcement on the composite structure was taken into account by a user subroutine which added the oriented fibers effect to the base material of the composite.
2. *Geometric nonlinearity.* This nonlinearity is applicable due to the relatively large deformations anticipated. As the helmet is a curved structure, the effect of this curvature on the stress distribution has to be taken into account.
3. *Force application nonlinearity.* This nonlinearity is applicable due to the expected nonlinear displacement loading as a function of time. A FEA contact analysis was required because we idealized the system as a nondeformable projectile impacting a deformable helmet.
4. *Boundary condition nonlinearity.* This nonlinearity is also applicable in FEA contact problems. This nonlinearity is expected during the projectile impact on the helmet due to the boundary conditions changing as a function of time and loading.

The model used consisted of a half helmet with three element layers of composite material through the thickness. A half model was used to take advantage of the symmetry present in the problem. The elements used were eight-noded brick elements because these elements can represent the deformation sufficiently. Anisotropic material properties were defined for each element using a user subroutine and impact was designated at the front of the helmet.

After the solution was obtained the results were analyzed. Each of the three principal stresses (σ_1, σ_2,

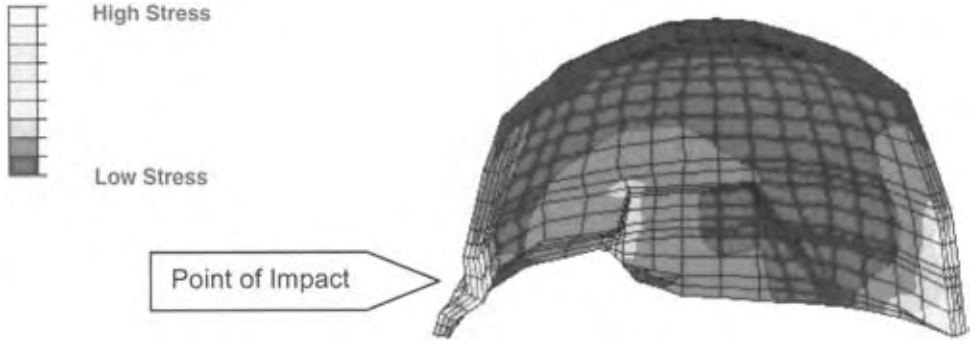


Fig. 10 Finite element analysis results for stress during ballistic impact on the helmet.

σ_3) are available in most FEA software packages and stresses are usually averaged by the FEA software packages to provide more accurate stress values when mapped (contoured) on to the mesh. A good first cut to the understanding of analysis results is the use of the von Mises stress (effective or equivalent stress). Fig. 10 shows the von Mises stress contour mapped to the FEA mesh in pounds/square inch (PSI).

The simulation shows that the helmet would deform at the location of impact, but the stresses would be distributed out and around the impact location thereby reducing penetration probability as compared with a metal helmet. Thus, we have used a combination of an understanding of material properties with computer simulation to deliver the required performance.

CONCLUSIONS

Environmental and economic drivers will accelerate the development of products generating minimal wastes. As Senge and Carstedt have pointed out, an understanding of the utilization of resource productivity is essential to the development of an industry that can reduce waste. An understanding of thermoset processing is of vital importance in many industries such as integrated chip manufacturers, aerospace technologies, automotive manufacturers, etc. Traditionally, design and control of these processes has relied on trial and error methods due to the complexity of the reacting systems. With the recent advances of modeling, kinetic tests, and chemorheological measurement, characterization of these complex reacting systems have found a foundation in the prediction of the processing-structure and performance of thermosetting systems. However, there is still a long way to go. Most of the test methods described in the paper by Halley and Mackay[15] have one thing in common: there are very few data to determine the waste generation function in material control. Some of the key drivers for waste are well known. In thermosets for example, compounded resin consistency is a key input variable for waste minimization. Since the final structure of the thermoset controls the performance, there is a need to have a measurement system to evaluate resin reactivity reliably. One such measurement system has been reported by Kranbuehl et al.[63] They developed an in-line frequency-dependent electromagnetic sensor to monitor cure in polyimide and epoxy thermosets. The sensor was then incorporated into a closed loop controller to monitor and control cure during manufacturing. Similarly, we have shown that phenolic resin composition directly affects carbon composite performance.[64] Control of the base resin system allowed production of carbon-carbon composites with minimal failures due to cracking. These types of capable measurement systems must be extended to the whole field of thermoset manufacturing so that the design engineer can reliably predict the effect of waste. If we want to minimize the waste generation function, and maximize the value generating function in materials processing, we need to develop a clear, consistent way to predict and minimize the waste generation function.

ACKNOWLEDGMENTS

The authors thank Tony Signorelli, Alan Levy, Richard Wilson, Janice Sund, Virginia Szigeti, Lori Wagner, Tom Izod, and Sunil Kasavan for their helpful comments and suggestions.

REFERENCES

1. Brydson, J.A. *Plastics Materials*, 6th Ed.; Butterworth—Heinemann: New York, 1999.
2. Newman, S. Introduction to composite materials technology: mass production techniques. In *Composite Materials Technology–Processes and Properties*; Mallick, P.K., Newman, S., Eds.; Hanser Publishers: New York, 1990; 10–24.
3. Crawford, R.J. *Plastic Engineering*, 3rd Ed.; Butterworth-Heinemann: New York, 1998.
4. Karger-Kocsis, J. Composites (structure, properties, and manufacturing). In *Polymeric Materials Encyclopedia*; Salamone, J.C., Ed.; CRC Press: New York, 1996; Vol. 2, 1378–1383.
5. Ashby, M.F.; Jones, D.R.H. *Engineering Materials 2–An Introduction to Microstructures, Processing, and Design*; Pergamon Press: New York, 1988.
6. Harry, M.J.; Lawson, J.R. *Six Sigma Producibility Analysis and Process Characterization*; Addison-Wesley: New York, 1992.
7. Curran, S.; Lem, K.W.; Sund, S.; Gabriel, M. Six sigma design–an overview of design for six sigma. In *Encyclopedia of Chemical Processing*; Lee, S., Ed.; Marcel Dekker, Inc.: New York, 2006 (in press)
8. Berryman, M.L. Transform your organization into one that's world class. Six Sigma Forum Magazine **2002**, *2* (1).
9. Creveling, C.M.; Slutsky, J.L.; Antis, D., Jr. *Design for Six Sigma – In Technology and Product Development*; Pearson Education, Inc.: New Jersey, 2003.
10. Bicerano, J. *Prediction of Polymer Properties*; Marcel Dekker, Inc.: New York, 1993.
11. Mark, J.E. *Physical Properties of Polymer Handbook*; American Institute of Physics: New York, 1996.

12. Rajan, K. An informatics infrastructure for in silico materials science. NIST Combinatorial Methods Center NCMC3, May 23, 2003.
13. Iden, R.; Schrof, W.; Hadeler, J.; Lehmann, S. Combinatorial materials research in the polymer industry: speed versus flexibility. Macromolec. Rapid Commun. **2003**, *24* (1), 63–72.
14. Yarovsky, I.; Evans, E. Computer simulation of structure and properties of crosslinked polymers: application to epoxy resins Polymer **2001**, *43* (3), 963–969.
15. Halley, P.J.; Mackay, M.E. Chemorheology of thermosets–an overview. Polym. Eng. Sci. **1996**, *36* (5), 593–609.
16. Lem, K.W. Rheology and Curing Behavior of Thermosetting Polyester Resins. Ph.D. Dissertation, Polytech. Inst. New York, Brooklyn: New York, U.S.A., 1983.
17. Lem, K.W.; Han, C.D. Thermokinetics of unsaturated polyester and vinyl ester resins. Polym. Eng. Sci. **1984**, *24* (3), 175–184.
18. Mallick, P.K. Compression molding. In *Composite Materials Technology–Processes and Properties*; Mallick, P.K., Newman, S., Eds.; Hanser Publishers: New York, 1990; 67–102.
19. Li, M.; Tucker, C.L. Optimal curing of thermoset matrix composites: thermochemical and consolidation consideration. Polym. Compos. **2002**, *23*, 739–757.
20. Racich, J.L.; Koutsky, J.A. Boundary layers in thermosets. In *Chem. Prop. Crosslinked Polym,* Proc. ACS Symp; Labana, S.S., Ed.; 1977; 303–323.
21. Mijovic, J.; Tsay, L.L. Correlations between dynamic mechanical properties and nodular morphology of cured epoxy resins. Polymer **1981**, *22* (7), 902–906.
22. Dusek, K. Are cured thermoset resins inhomogeneous? Angew. Makromol. Chemie **1996**, *240*, 1–15.
23. Feldman, D. Composites–thermosetting polymers. In *Polymeric Materials Encyclopedia*; Salamone, J.C., Ed.; CRC Press: New York, 1996; 2, 1383–1389.
24. Bigg, D.M. Thermoplastic matrix composites. In *International Encyclopedia of Composites*; Lee, S.M., Ed.; VCH Publishers: New York, 1991; Vol. 6, 30–33.
25. Scola, D.A. Thermoset composites, advanced. In *International Encyclopedia of Composites*; Lee, S.M., Ed.; VCH Publishers: New York, 1991; Vol. 6, 34–48.
26. Prevorsek, D.C. Spectra®: The latest entry in the field of high-performance fibers. In *Handbook of Fiber Science and Technology*; Lewin, M., Ed.; Marcel Dekker, Inc.: New York, 1996; Vol. 3, 1–170.
27. Gray, W.; Zabel, P.H. Armor, composite. In *International Encyclopedia of Composites*; Lee, S.M., Ed.; VCH Publishers: New York, 1991; Vol. 6, 125–140.
28. Kumar, S. Ordered polymer fibers. In *International Encyclopedia of Composites*; Lee, S.M., Ed.; VCH Publishers: New York, 1991; Vol. 4, 51–74.
29. Ganguli, S.; Dean, D.; Derrick, J.; Kelvin, P.; Price, G.; Vaia, R. Mechanical properties of intercalated cyanate ester-layered silicate nanocomposites. Polymer **2003**, *44* (4), 1315–1319.
30. Su, W.-F.A. Thermosets (main chain liquid crystalline polymers). In *Polymeric Materials Encyclopedia*; Salamone, J.C., Ed.; CRC Press: New York, 1996; Vol. 11, 8375–8380.
31. Gillham, J.K. *The Role of Polymer Matrix in Processing and Structural Properties of Composites*; Seferis, J. C., Nicolais, L., Eds.; Plenum Press: New York, 1983; 127–145.
32. Sperling, L.H.; Mishra, V. Interpenetrating polymer networks (overview). In *Polymeric Materials Encyclopedia*; Salamone, J. C., Ed.; CRC Press: New York, 1996; Vol. 5, 3292–3301.
33. Frisch, H.L.; Xue, Y.P. Interpenetrating polymer networks (rubber-based). In *Polymeric Materials Encyclopedia*; Salamone, J. C., Ed.; CRC Press: New York, 1996; Vol. 5, 3302–3308.
34. Bascom, W.; Gweon, S.Y.; Grande, G. Multiphase matrix for carbon fiber composites. Adv. Chem. Ser. **1993**, *233*, 519–537.
35. Cook, W.D.; Dean, K.; Forsythe, J. Cure, rheology and properties of IPN thermosets for composite applications. Materials Forum **2001**, *25*, 30–59.
36. Yoshihara, T. Substrates Having Gas-barrier Organic-Inorganic Hybrid Polymer Layers and their Manufacture. Patent JP 2003094572, 2003.
37. Qin, H.; Mather, P.T. Optical rheology of new liquid crystalline thermosets (lcts): influence of shear on disclination texture. Mat. Res. Soc. Symp. Proc. **2002**, *709*, CC8.4.1–7.
38. Korner, H.; Shiota, A.; Ober, C.K. The processing of LC thermosets in orienting external fields. Mat. Res. Soc. Symp. Proc. **1996**, *425*, 149–160.
39. Wadahara, E.; Ishibashi, S.; Nagashima, Y. Polymer Conductor Composition and Molded Materials thereof. Patent JP 2001067933, 2001.
40. Dershem, S.; Yang, K. Low Shrinkage Thermosetting Resin Compositions and use in Low Shrinkage Die Attach Pastes. Patent WO 2002028813, 2002.
41. Ritzenthaler, S.; Court, F.; David, L.; Girard-Reydet, E.; Leibler, L.; Pascault, J.P. ABC Triblock copolymers/epoxy-diamine blends. 1. Keys to achieve nanostructured thermosets. Macromolecules **2002**, *35* (16), 6245–6254.

42. Tripathy, S.; Chen, J-I.; Marturunkakul, S.; Kumar, J. Nonlinear optical materials. In *Polymeric Materials Encyclopedia*; Salamone, J.C., Ed.; CRC Press: New York, 1996; 6, 4587–4596.
43. Shannon, P.J.; Gibbons, W.M.; Sun, S.T. Nonlinear optical polymers, thermosets. In *Polymeric Materials Encyclopedia*; Salamone, J.C., Ed.; CRC Press: New York, 1996; 6, 4605–4611.
44. Sek, D. Structural variations of liquid crystalline polymer macromolecules: review. Acta Polymerica **1988**, *39* (11), 599–607.
45. Senge, P.M.; Carstedt, G. Innovating our way to the next industrial revolution. MIT Sloan Management Rev. 2001(Winter), 24–38.
46. AlliedSignal, Inc.; DSM. Innovative Green Chemistry for Sustainable Manufacture of Caprolactam. Program Proposal to U.S. Environmental Protection Agency's Presidential Green Chemistry Challenge Awards Program. 1999 (December).
47. Sifniades, S; Levy, A.B.; Hendrix, J.A.J. Processes for Depolymerization Nylon-Containing Whole Carpet to Form Caprolactam. US Patent 5,929,234, July 27, 1999, US Patent 5,932,724, August 3, 1999.
48. Pollution Prevention Assistance Division, Dept. of Natural Resources; AlliedSignal, Inc.; Georgia Department of Transportation; Georgia Environmental Facilities Authority. Demonstration of Polymer By-Product Utilization as an Asphalt Modifier. A proposal Submitted to U.S. Department of Energy for National Industrial Competitiveness Through Energy, Environment, and Economics. Solicitation No. DE-PS36-98GO10294. October 20, 1998.
49. Lem, K.W.; Letton, A.; Izod, T.P.J.; Lupton, F.S.; Bedwell, W.B. Composition Containing Caprolactam-Free Residue from Depolymerization of Nylon 6 Carpet and Use Thereof in Paving Asphalt, Plastic Lumber and Crack Sealants. US Patent 6,214,908, April, 10, 2001, USP 6,414,066, July 2, 2002.
50. Gutowski, T.G. Environmentally benign manufacturing and ecomaterials; product induced material flows. Mater Trans. **2002**, *43* (3), 359–363.
51. Vaklieva-Bancheva, N.G.; Shopova, E.G.; Ivanov, B.B. Application of Fourier transformation for waste minimization in batch plants. 1. Analysis of production recipes. Hung J. Ind. Chem. **2002**, *30* (3), 199–206.
52. O'Reilly, A.J. Batch reactor optimization, profitability vs. waste minimization. Chem. Eng. Res. Des. **2002**, *80* (A6), 587–596.
53. Zhang, Q.-Y. Multiple objectives application approach to waste minimization. J. Zhejiang Univ. Sci. **2002**, *3* (4), 405–411.
54. Ciantar, C.; Hadfield, M.; Howarth, G. Case studies to assist integrating waste prevention in product design. MechE Conference Transactions; Engineering for Profit from Waste **2001**, *9*, 201–210.
55. Cochrane, T.; Smith, J.A. Designing processes and products to minimize wastes produced. MechE Conference Transactions; Engineering for Profit from Waste **2001**, *9*, 137–148.
56. Henningsson, S.; Smith, A.; Hyde, K. Minimizing material flows and utility use to increase profitability in the food and drink industry. Trends Food Sci. Technol. **2001**, *12* (2), 75–82.
57. Page, P.G. Efficient cost management through chemical conservation and waste minimization for the electroplating industry. Proc. AESF Ann. Tech. Conf. **1997**, *84*, 321–327.
58. Han, C.; Stephanopoulos, G.; Liu, Y.A. Knowledge-based approach in process synthesis. Recents Progres en Genie des Procedes-Simulation, Optimisation et Commande, SIMO'96. **1996**, *10* (49), 1–13.
59. Basta, N. Design hazards out with process simulation. Environ. Eng. World **1996**, *2* (3), 28–29.
60. Hilaly, A.K.; Sikdar, S.K. Process simulation tools for pollution prevention. Chem. Eng. **1996**, *103* (2), 98–105.
61. Edgar, T.F.; Huang, Y.L. Artificial intelligence approach to synthesis of a process for waste minimization. In *Emerging Technologies in Hazardous Waste Management IV*; ACS Symp Ser, 1994; Vol. 554, 96–113.
62. Berglund, R.L.; Snyder, G.E. Minimize waste during design. Hydrocarb. Proc., Int. Ed. **1990**, *69* (4), 39–42.
63. Kranbuehl, D.E.; Hood, D.K.; Rogozinski, J.; Barksdale, R.; Loos, A.C.; McRae, D. FDEMS sensing for automated intelligent processing of polyimides. Proc. ASME Mat. Div. **1995**, *2*, 1017–1046.
64. Curran, S.; Walker, T.B.; Brambilla, R. Characterization of phenolic resins for carbon composites. In *POLY-471*, 213th ACS National Meeting, San Francisco, April 13–17, 1997.

Thin Film Processes in MEMS and NEMS Technologies

W. R. Ashurst
C. Carraro
R. Maboudian
University of California–Berkeley, Berkeley, California, U.S.A.

INTRODUCTION

The integration of miniaturized mechanical components with microelectronic components has resulted in a new technology, known as microelectromechanical systems (MEMS), which extends the benefits of microelectronic fabrication to sensing and actuating functions. Several MEMS devices are in commercial use, such as accelerometers, pressure sensors, digital mirror displays, and gyroscopes. The first fully integrated single-chip MEMS accelerometer was manufactured in 1991 by Analog Devices. By September 2002, Analog Devices had sold more than 100 million MEMS accelerometers and gyroscopes. With size shrinking below the micrometer scale, nanoelectromechanical systems (NEMS) are also being realized.

Generally, MEMS and NEMS technologies are able to exploit properties that scale favorably with decreasing size. While this ability provides a unique capability, it also poses challenges to the fabrication and reliability of these devices. Conventional machining and assembly techniques cannot be easily applied to microsystems. Moreover, because of the dominance of surface effects on the micro- and nanoscale, strong adhesion, friction, electrostatic charging, and wear have been shown to be issues crucial to this technology. In this article, we present an overview of MEMS and NEMS fabrication processes, reliability issues, and ways to improve reliability with special emphasis on the chemical processes involved.

FABRICATION PROCESSES FOR MEMS AND NEMS

Historically, silicon and silicon based materials have formed the basis for the well-established integrated circuit (IC) technology. As a consequence, polycrystalline silicon (polysilicon) is one of the most commonly used structural materials for MEMS. Comparatively, silicon is a good structural material for the microscale.[1] However, the inherent mechanical nature of MEMS devices brings about a new level of complexity to their production and reliability compared to standard integrated circuits. Micromechanical structures can be fabricated from a variety of methods, including bulk micromachining, LIGA (a German acronym for lithography, electroplating, and molding), and surface micromachining.[2] Surface micromachining is one of the most common methods for MEMS fabrication, and it involves the deposition, patterning, and etching of thin films–processes that are commonplace in the IC industry.[1,3] As such, surface micromachining leverages heavily from the technology of the IC industry, and allows for the production of sophisticated microstructures with parallel fabrication and high fabrication yield. It should be noted that silicon is not always chosen as the structural material. Some commercialized MEMS products use aluminum as the structural material, e.g., the Texas Instruments DMD.™

A simplified surface micromachining process diagram is illustrated in Fig. 1. The process begins with a silicon wafer, which has an electrical isolation layer (typically silicon nitride) deposited on the surface. Next, the first polysilicon (poly) layer is deposited. This will form the ground plane and actuation pad for the device in this illustration, a cantilever beam. This poly layer is then lithographically patterned and anisotropically etched (using, e.g., plasma etching). The sacrificial (spacer) layer, typically silicon oxide, is then deposited over the patterned poly layer. The sacrificial layer is also lithographically patterned, and anisotropically etched to reveal access windows to the underlying poly layer. The structural poly layer is deposited, lithographically patterned, and anisotropically etched to complete the structure. Now, although the fabrication of the device is complete, it is not usable until the sacrificial layer is removed. This is accomplished by using an isotropic etch, to dissolve the sacrificial layer and leave a ''released'' device.

The general process for MEMS fabrication can also be applied to NEMS. Owing to the dimensions involved in NEMS, the patterning processes used differ from that of MEMS. For NEMS patterning, e-beam lithography or atomic force microscopy (AFM) writing can be used. E-beam lithography uses a highly focused electron beam to ''write'' on a thin film. The action of the electron beam alters the film material to the extent that it can be removed while leaving unaltered material intact. In this respect e-beam lithography is

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120037168

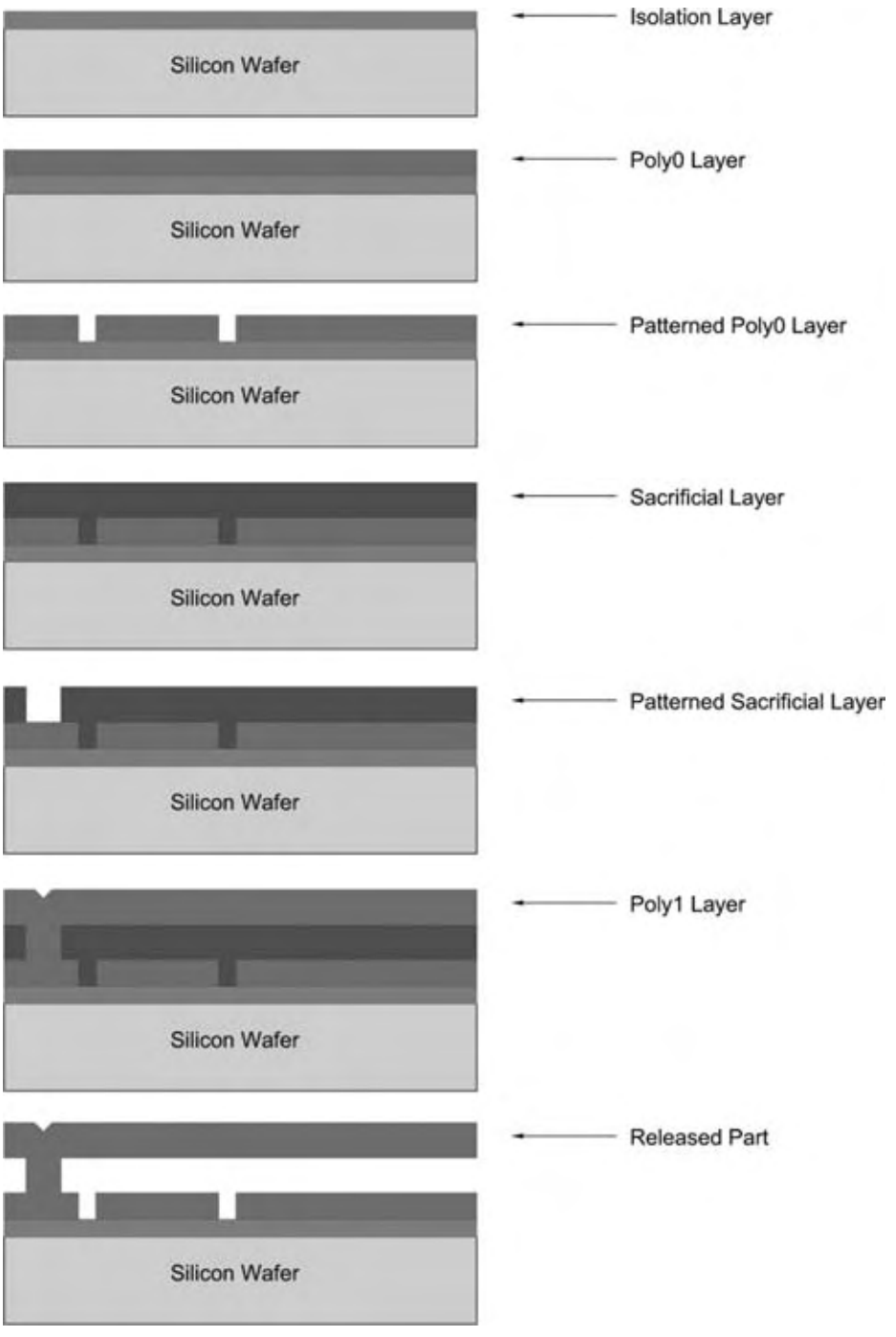


Fig. 1 Basic diagram of the surface micromachining process. *(View this art in color at www.dekker.com.)*

similar to conventional lithography. However, the e-beam must be traced over the mask (a serial process) and is not flash exposed as in conventional lithography. Atomic Force Microscopy writing can be done in several ways. One way is to apply a bias to the sharp tip and alter the material under the tip by exposing it to the high electric field. This may lead to localized oxidation, which can be used to transfer the pattern to the substrate. Alternatively, the tip can be scraped across the surface (especially on soft coatings) and mechanically remove material from the desired areas. Similar to e-beam lithography, AFM writing is a serial process. In addition to nanoscale writing, the so-called ''spacer lithography'' can be used to precisely define nanometer-wide gaps. This process consists of the conformal deposition of a nanometer thin film over an exposed step surface. Afterward, the plan surfaces are anisotropically etched, leaving only the nanometer thin ''spacer'' material on the sidewalls of the vertical step. Unlike e-beam lithography

and AFM writing, spacer lithography is a parallel process and is carried out over the whole surface of the sample at once.

CHEMICAL PROCESSES USED IN FABRICATION

There are a variety of chemical processes used in the manufacturing of MEMS and ICs. These include deposition and etching processes. Some of the more common examples of these processes are described briefly in the sections that follow.

Chemical Vapor Deposition

One of the most common deposition processes encountered in the fabrication of microsystems is chemical vapor deposition (CVD) (see, e.g., Ref.[2]). This broad category encompasses processes including low-pressure (LP), atmospheric-pressure (AP), plasma-assisted (PA), ion beam-assisted (IBA), and laser-assisted (LA) CVD. Regardless of the particular type of CVD, the process consists of exposing a target substrate to a supply of a gas (or gas mixture) and activating that gas to produce a chemical reaction at the surface of the substrate. This activation method may include heating, striking a plasma, laser irradiation, ion beams, or combinations thereof. The process parameters are carefully controlled so that the deposited material films are of high quality, and have the desired stress and composition properties.

In silicon based MEMS processing, common CVD films include polysilicon, silicon oxide, and silicon nitride. For polysilicon films (usually the structural layer), an LPCVD pyrolysis method is generally used with silane (SiH_4) as the source gas [see Eq. (1)]. To obtain a uniform film across the wafer, the process is carried out at low pressure to ensure that the deposition is surface reaction controlled and not diffusion limited. Typical process temperatures are in the range of 580–650°C, and pressures between 0.1 and 0.4 Torr.

$$SiH_4 \xrightarrow{\Delta} Si + 2H_2 \quad (1)$$

Another important material is silicon nitride, which is usually employed as an electrical isolation layer. It is typically deposited by the LPCVD or PECVD method. The process gases are dichlorosilane $SiCl_2H_2$ and ammonia NH_3 [see Eq. (2)] and the process parameters are in the range of 700–900°C and 0.2–0.5 Torr.

$$SiCl_2H_2 + NH_3 \xrightarrow{\Delta} Si_3N_4 + 2HCl + \frac{3}{2}H_2 \quad (2)$$

Silicon oxide, typically a sacrificial layer, can be deposited using PECVD and LPCVD methods. There are many source gases that can be used to produce silicon oxide. These include mixtures of silane and oxygen, tetraethoxysilane and oxygen, dichlorosilane and N_2O, and silane and N_2O. An example reaction is shown in Eq. (3). Because there are many options for source gases, the process parameters may range from 400 to 900°C and from 0.1 to 0.5 Torr.

$$SiH_4 + O_2 \xrightarrow{\Delta} SiO_2 + 2H_2 \quad (3)$$

Etching

To create micromachines, films that have been deposited must be patterned and etched to reveal the desired structures. Often, it is important to etch these structures with vertical sidewalls (anisotropic etching). In this case, most pattern transfer operations (lithography and etch) are carried out using plasma etching. Conceptually, this process is the reverse of deposition. The etching process consists of exposure of the patterned and masked substrate to a low-pressure plasma. The reactive species and ions preferentially etch those areas that are not masked, resulting in the definition of features on the surface. The key to plasma etching is that the products of the reaction of the activated gas and the material to be etched must be volatile (see e.g., Ref.[2]).

Reactive ion etching (RIE) and deep reactive ion etching (DRIE) are common examples of plasma etching. In this incarnation, the substrate to be etched is placed on a powered electrode in a plasma chamber. Process gases are admitted into the chamber and a plasma is struck. Because the substrate is directly in the ion flux of the plasma, the ions impinge on the surface and may participate in chemistry. For example, RIE of polysilicon may use SF_6 as the reactive gas, and etches the Si by a reaction with fluoride ions to form the volatile product SiF_4, as illustrated by Eqs. (4) and (5).[a]

$$SF_6 \xrightarrow{*} SF_5^+ + F^- \quad (4)$$

$$Si + 4F^- \xrightarrow{*} SF_4 \quad (5)$$

The anisotropic nature of the etching by RIE and DRIE is a result of the directionality of the impinging ions. Sometimes it is useful to etch very deep channels or holes. Using conventional DRIE, some tapering of the sidewall profile is expected. However, an etching process has been developed where sequential etch/passivate steps are performed. This etching process, called the Bosch process, uses conventional DRIE methods as process gases for a short period of time, interrupts

[a]Plasma chemistries are complex and these equations are meant only to indicate an example process.

this process with a fluorocarbon polymer deposit phase, and resumes DRIE. Because of the directionality of the etching process, the polymer on the plan surfaces is quickly removed and the etching process resumes. However, the material removal on the sidewalls is slower, and therefore sidewall surfaces are not readily etched. This process allows very deep features to be etched with little tapering. Because of the way this process is carried out, however, the sidewall surfaces are generally scalloped, and this is a characteristic feature of the Bosch etch process.

RELEASE PROCESSES

While MEMS technologies make much use of the preexisting integrated circuit fabrication processes, there is at least one critical step that is unique to MEMS processing. This step, called the release step, refers to the processing stage that frees the microstructure from the sacrificial layers that were used during the fabrication of the device. This is usually accomplished by exposing the micromachine to an etchant fluid that will selectively and completely remove the sacrificial material. In the case of polysilicon based microstructures, the sacrificial layers are typically silicon oxide, and the etchant fluid is an aqueous solution of HF. In some cases, an etching solution of concentrated HF and HCl is used. The presence of HCl helps to protect the silicon nitride from etching by HF. Next, the etchant is completely rinsed away with deionized water. If, at this point, the released structures are simply removed from the rinse liquid and air-dried, they are almost inevitably found adhering to each other. This phenomenon is called ''release stiction'' and is caused by the action of the curved receding air–liquid interface as it passes over the structures. Alternate processes are typically carried out to avoid the problem of release stiction.

One alternate approach is to change the water meniscus shape from concave (wetting) to convex (nonwetting), so that capillary forces are reduced.[4] This is accomplished by altering the composition of the surface by chemically grafting hydrophobic molecules to the surface. This approach is discussed in greater detail in the sections that follow. Other approaches involve the avoidance of liquid–vapor interfaces altogether through supercritical fluid, freeze sublimation drying, and dry-release methods such as the use of vapor HF etching. Fig. 2 shows a P–T diagram that depicts the concepts of supercritical drying and freeze sublimation. Supercritical drying is a process that begins with released microstructures, which have been kept submerged in liquid since their sacrificial layer etch, and then placed in a high-pressure chamber filled with a short alcohol (methanol, ethanol, and isopropanol) solvent. This solvent is then completely

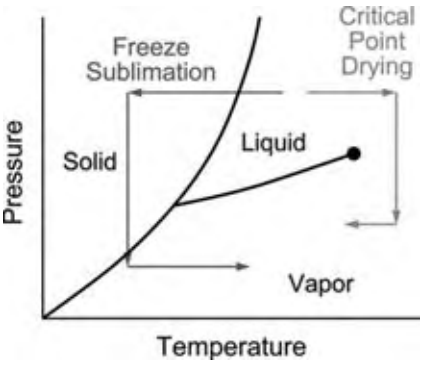


Fig. 2 Diagram depicting the phase paths of critical point drying and freeze- sublimation drying. Note that neither of these processes involve crossing the liquid–vapor line. *(View this art in color at www.dekker.com.)*

displaced by liquid carbon dioxide (CO_2). The chamber is pressurized and heated to cause the liquid CO_2 to become supercritical. The chamber is then carefully vented and the structures are removed.[5] Freeze sublimation drying is similar to supercritical drying. This process also begins with released microstructures, which are maintained in a solvent, such as isopropanol. This solvent is then displaced by another solvent (water/isopropanol mixture), which is then frozen. The vapor above this solid is pumped away by a vacuum pump until the solid is completely sublimated and the process is complete.[6] Vapor HF etching is a process that completely avoids the issue of solvent displacements, as the sacrificial layer etch is carried out by a gaseous isotropic etchant. Usually, the samples are placed above a solution of HF and water and the vapors above that solution are used to perform the etch.[7]

Polymer ashing is another process related method. This process is fairly complex and involves patterning of a polymer layer during the release. First, structures are partially released by a timed etch. Next, a polymer film is deposited onto the partially released structures. This film is patterned into support posts that hold the structure in position as the remainder of the sacrificial layer is etched away. Because the polymer support structures hold the devices in place, there is no concern for special drying techniques. Finally, the polymer supports are burned away, typically by ashing in an oxygen plasma.[8] This leaves behind fully released and free-standing microstructures.

SURFACE PROCESSING FOR INCREASED RELIABILITY

Surface microstructures typically have lateral dimensions of 50–500 μm with thicknesses of 0.1–2.5 μm, and are offset 0.1–2 μm from the substrate. The large

surface-area-to-volume ratios of surface and bulk micromachined micromechanisms lead to the dominance of surface and interfacial forces over body forces.[8–12] For commercial viability and industrial growth to continue, micromechanical systems must be built with high yields and reproducible device properties and must exhibit reliability over the expected device lifetime. The traditional high yields experienced in the IC industry extend to MEMS production only up to the point of microstructure release.

The release techniques discussed here do not prevent adhesion from occurring during micromachine operation. Microstructure surfaces may come into contact unintentionally through acceleration or electrostatic forces, or intentionally in applications where surfaces impact or shear against each other. When adhesive attractions exceed restoring forces, surfaces permanently adhere to each other causing device failure—a phenomenon known as ''in-use stiction.''

Stiction is not the only reliability issue that plagues microsystems. Owing to the dimensions of the devices and roughness of the surfaces involved, the mechanical properties of the structural material are very important. Although silicon is a good structural material for the microscale, owing to its high elastic modulus and low density, small area contacts (as in contact points between two rough surfaces) can generate enormous contact pressures, which can create plastic deformations in silicon. For example, the hubs on microgears can become visibly cluttered with wear debris after a few hundred thousand cycles. These microgears seldom last more than a few million cycles before the buildup of wear debris and increased friction irreversibly bind the gear.[13] Current research on the issue of wear seeks to understand the fundamental mechanisms responsible, as well as to address the practical issues by integrating hard materials into the micromachining process. Adhesion, friction, and wear are collectively referred to as tribology.

Owing to the extensive infrastructure related to the chosen materials base (silicon), and the relative ease in which silicon based microsystems can be produced, a convenient approach to overcoming some of these reliability issues is to apply coatings for stiction, friction, and wear control to the devices after they are produced. This is analogous to painting buildings or automobiles after they are constructed to make them more weather resistant. Therefore, facile and effective coating processes that satisfy the needs of the microsystems for reliable performance have been developed.

Methods to Quantify MEMS Tribology

To quantitatively study stiction, friction, and wear, micromechanical test devices have been designed and fabricated. One of the most common test devices for stiction quantification is the cantilever beam array.[8] With this device, stiction is tested by electrostatically bringing the beam into contact with the substrate and releasing the electrostatic force. By examination of the behavior of the interaction of the beam with the substrate, one can determine an apparent work of adhesion, which is related to stiction.

Other devices have been designed for friction and wear testing. Although there are many types of these devices, they generally consist of a movable structure and a fixed structure. To test friction, the movable structure is contacted against the fixed structure under a prescribed contact (normal) load and a tangential force is ramped up. The device is carefully monitored until a slip occurs and the tangential force at slippage is known. The coefficient of static friction can be determined by knowledge of the applied normal load and the tangential force required to produce a slip. Generally, devices that are used for friction testing can also be used for wear testing. Instead of a single slip event, the movable part is cycled against the fixed part under a given normal load. At the end of the test, scanning electron microscopy or AFM can be used to inspect the contacting parts to gage the amount of wear.

Self-Assembled Monolayer Coatings to Control MEMS Tribology

Coatings that are to be applied to micro- and nanostructures must satisfy a series of constraints related to their properties and the process used to generate them. They must be conformal and uniform because all surfaces of the device must be coated with the same amount of coating to avoid the introduction of extraneous stresses. Coatings must be very thin for design fidelity reasons, and so that it does not adversely influence the operation of the device. (For example, a coating must not adversely affect the resonance frequency or quality factor for a MEMS resonator.) They must be thermally and chemically stable because the device is expected to last a long time and endure packaging steps. The coating process must be compatible with the device and should be easily implemented.

In light of these constraints, an effective chemical modification for anti-stiction treatments involves the application of a molecular film to the micromachine surface. This is most often accomplished through a process known as self-assembled monolayer (SAM) deposition.[14] Self-assembled monolayers are molecular films that spontaneously form on a (usually pretreated) surface upon exposure to a reactive precursor molecule. SAM precursors generally consist of three main parts: a terminal group, a backbone, and a head group. Fig. 3 shows these parts on a model SAM

precursor molecule, octadecyltrichlorosilane (OTS). The head group is a chemically reactive group that is chosen to bind the molecule to the MEMS surface. The backbone is often an *n*-alkyl chain (i.e., $-(CH_2)_n-$) whose function is to assist in the assembly of a well-packed monolayer film. The terminal group is the part of the molecule that comprises the new surface after monolayer self assembly and imparts the desired surface functionality to the coating. For adhesion control, a highly hydrophobic surface is desired, and the terminal groups of choice are hydrophobic (water disliking) $-CH_3$ or $-CF_3$ groups (the $-CF_3$ terminal groups result in monolayer coatings that have nonstick behavior similar to Teflon®).

It has been demonstrated that, when properly integrated into the microstructure release process, SAM coatings deposited from solution can eliminate release stiction, reduce in-use stiction, and decrease both friction and wear in microengines.[11,12] Several classes of organic films have been explored as antistiction agents. For aluminum surfaces, perfluoroalkanoic acids have been successfully employed. On silicon surfaces, chloroilane based films, which include alkyl- and perfluoroalkyl-trichlorosilane SAMs, dialkyldichlorosilane, tris-dimethylaminosilanes, and primary alkene, alcohol and aminebased molecular films have been tested.

Perfluoroalkanoic acids

Perhaps the best example of antistiction technology for MEMS to date is the coating process employed by Texas Instruments on the DMD™ device. The DMD is a MEMS device, which consists of an array of a million or more rotatable aluminum mirrors. Photoresist is used as the sacrificial layer and is removed at the release step using a remote plasma containing oxygen and fluorine species. The device has contacting surfaces in relative motion, which are susceptible to adhesion, friction, and wear, and require lubrication. About 50 different lubrication schemes were investigated for DMD, ranging from SAMs to fluids to solid lubricants.[15] The most successful ones reportedly are perfluorinated *n*-alkanoic acids ($C_nF_{2n-1}O_2H$), which form self-assembled monolayers on aluminum oxide surfaces (see Fig. 4). Within this class of SAMs, perfluorodecanoic acid (PFDA, $n = 10$) was found to be the lubricant of choice to minimize the friction coefficient and the possibility of thermal decomposition. To keep moisture out and create a background pressure of PFDA, hermetic chip package is used. When properly lubricated, devices have operated for more than 350 billion cycles.

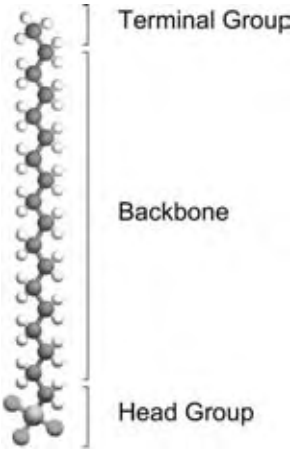


Fig. 3 An example SAM precursor molecule (OTS) with the major parts of the molecule labeled. Alternatively, the terminal group and backbone may be collectively referred to as the "pendant" group. *(View this art in color at www.dekker.com.)*

Chlorosilane based monolayers

Among the SAMs suitable for coating silicon, the OTS based variety is the most widely used. Some of the properties for the OTS SAM, of relevance to MEMS, are listed in Table 1. Although there is much debate concerning the true structure of the OTS monolayer on silicon oxide, Fig. 5 illustrates a simplified conceptual model of the film.

In addition to the OTS precursor molecule, there are many other molecules of the form $RSiCl_3$ and $R,R'SiCl_2$ that are used to produce oriented hydrophobic monolayers on silicon surfaces, with R and R′ denoting an aliphatic carbon chain. It has been demonstrated that the most effective chlorosilane reagents to produce hydrophobic coatings on oxidized silicon surfaces are perfluorinated alkyltrichlorosilanes.[19] Indeed, lower values for the apparent work of adhesion were reported for 1*H*,1*H*,2*H*,2*H*-perfluorodecyltrichlorosilane [$CF_3(CF_2)_7(CH_2)_2SiCl_3$; FDTS] in comparison to OTS.[17] Table 1 shows an interfacial property comparison among different surface preparations.

Although these SAM coatings have been shown to effectively alleviate both release and in-use stiction, they possess a number of limitations intrinsically related to their chemistry. A serious limitation arises from the ability of the precursor molecule to polymerize.[20] As long as the precursor molecule has a functionality greater than one, bulk polymerization can occur, and the higher the functionality, the greater the likelihood for polymerization. This is potentially dangerous for micromachines in that large particulates, such as polymerized clusters of SAM precursor molecules (which can be several micrometers

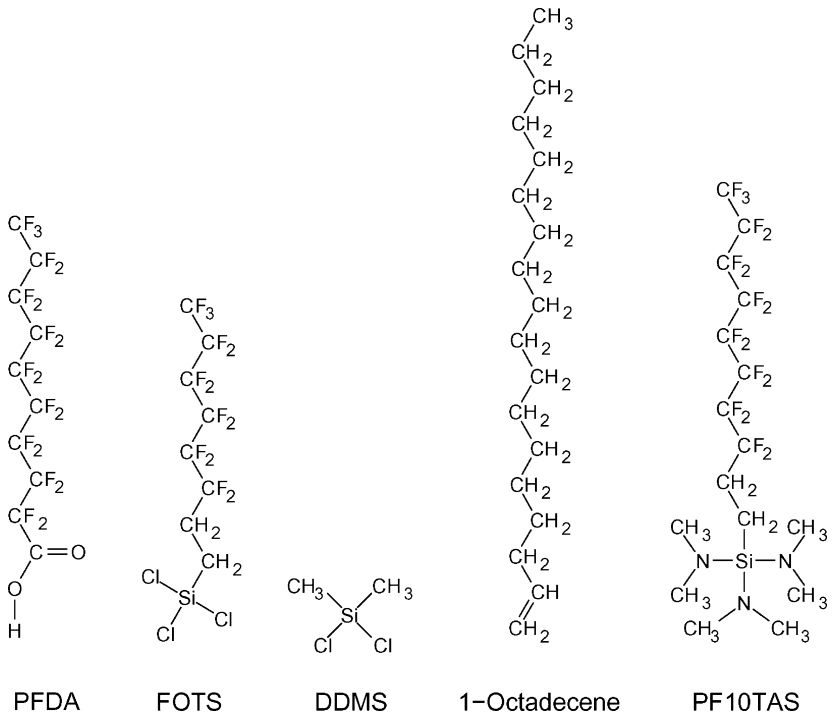


Fig. 4 Structural formulas for perfluorodecanoic acid (PFDA) perfluorooctyltrichlorosilane (FOTS), dimethyldichlorosilane (DDMS), 1-octa decene, and perfluorodecyl-tris-(dimethylamino)silane (PF10TAS).

in diameter), can mechanically interfere with the device operation. Unfortunately, there is no satisfactory method for removal of the polymerized clusters once they have agglomerated on the surfaces of the substrate or micromachines. Another limitation of long-chain chlorosilane SAM coatings (such as OTS) arises from the coating procedure, which must be performed from the liquid phase because of the extremely low vapor pressure of the precursor.

Recent developments in the chlorosilane based monolayer technology address some of these issues by performing the coating process in the vapor phase. Vapor phase processing eliminates the use of organic solvents and greatly simplifies handling of the samples. Moreover, the stoichiometry of the precursor molecules can be more precisely controlled. It has been demonstrated that monolayer films that are produced from the precursor tridecafluoro-1,1,2,2-tetrahydrooctyltrichlorosilane [$CF_3(CF_2)_5(CH_2)_2SiCl_3$; FOTS] in a low-pressure CVD style reactor exhibit low adhesion energies.[19] Fig. 4 shows the structure of FOTS. Additionally, in situ plasma cleaning of the sample as well as in situ measurement of the film growth provide excellent process uniformity, reproducibility, and monitoring capability.[21] This approach demands that the microstructures undergo some form of dry release process, such as critical point drying or vapor HF etching, before receiving the monolayer coating.

The dimethyldichlorosilane [$(CH_3)_2SiCl_2$; DDMS)] monolayer has also been proposed as a promising surface coating for MEMS. Fig. 4 shows the structure of DDMS. The monolayer has been compared to the OTS SAM with respect to the film properties and their effectiveness as antistiction coatings for micromechanical structures.[16,22] While water and hexadecane contact angles are comparable, the apparent work of adhesion for the DDMS monolayer is somewhat higher than that for OTS (Table 1). Furthermore,

Table 1 Physical property data for various surface treatments

Surface treatment	Contact angle (°)		Work of adhesion (mJ/m^2)	Coefficient of static friction	Thermal stability in air (°C)	Particulate formation	Selective to Si	Reference
	Water	Hexadecane						
OTS	110	38	0.012	0.07	225	High	No	[16]
FDTS	115	68	0.005	0.10	400	Very high	No	[17]
DDMS	103	38	0.045	0.28	400	Low	No	[16]
Octadecene	104	35	0.009	0.05	200	Negligible	Yes	[18]
Oxide	0–30	0–20	20	1.1	—	—	—	[16]

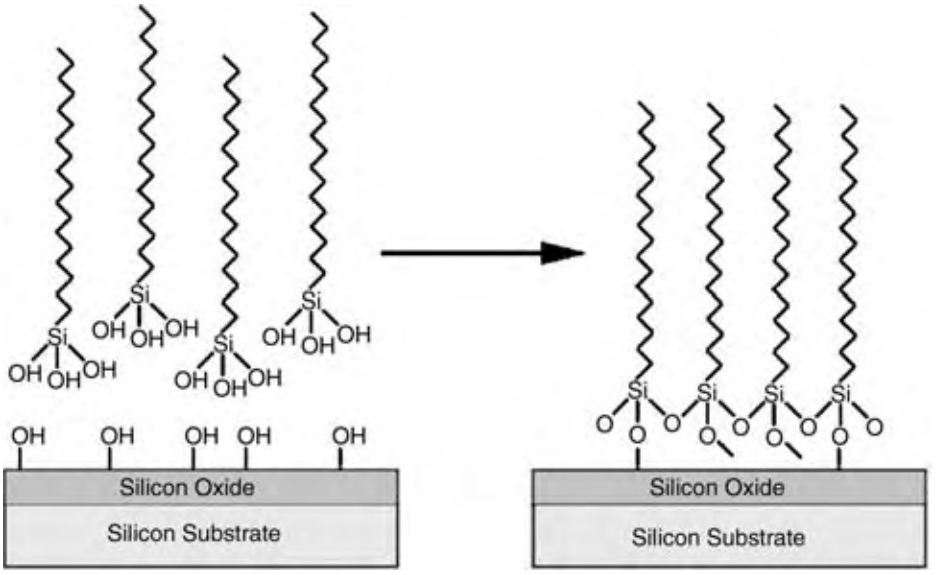


Fig. 5 A simplified diagram of the formation of an OTS monolayer. The first reaction (not shown) is the complete hydrolysis of the OTS molecule, resulting in the formation of 3HCl and the trisilanol form of OTS shown in the figure. Subsequently, water elimination reactions result in Si–O–Si linkages between adjacent OTS molecules and/or the oxidized substrate.

coefficient of static friction data indicate that the DDMS films are not as effective at lubrication as the OTS SAMs are, although both exhibit much improvement over chemical oxide. However, AFM data show that the samples that receive DDMS films accumulate fewer particles during processing than those that get the OTS SAM treatment. The thermal stability of the DDMS film in air far exceeds that of the OTS SAM, as the DDMS remains very hydrophobic to temperatures well upward of 400°C (Fig. 6).

Alkene based monolayers

Another surface modification technique involves the free radical reaction of a primary alkene (e.g., $C_{16}H_{33}CH{=}CH_2$, 1-octadecene) with hydrogen terminated silicon.[23] Fig. 4 shows the structure of 1-octadecene while Fig. 7 displays a simplified diagram of the resulting monolayer structure. This monolayer coating has several key advantages over OTS and FDTS based SAMs:

1. The coating does not produce HCl at any stage in the monolayer formation whereas chlorosilanebased chemistry does.
2. The coating does not require the formation of an intervening oxide layer.
3. The film formation procedure for alkene based monolayers is simpler than for chlorosilane based SAMs for two main reasons. First, the surface oxidation step is eliminated. Second, the coating solution does not need to be conditioned before use, as water is not a reagent in this process.
4. The coating process is much more robust because it is essentially insensitive to relative humidity.

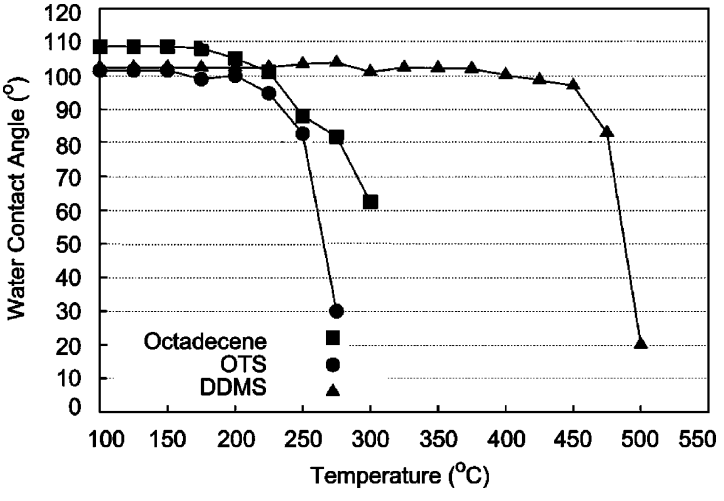


Fig. 6 Water contact angle as a function of substrate temperature. The substrate was exposed to high temperature for 5 min in laboratory air. Note the extreme stability of the DDMS monolayer in comparison to the OTS and octadecene monolayers.

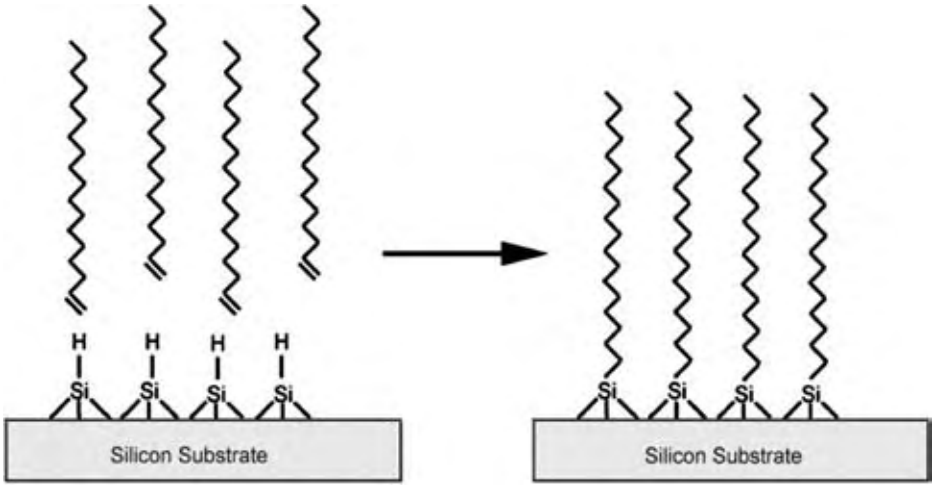


Fig. 7 A simplified diagram of the formation of an octadecene monolayer. The precursor molecule (1-octadecene) is bound directly to the silicon, with no oxide layer.

5. The coated surfaces have much fewer particulates in comparison to those coated with OTS.
6. The coating process can be made selective to coat only exposed silicon by generating radicals using a radical initiator.

These improvements have been achieved without sacrificing the antistiction characteristics of the film, in that water and hexadecane contact angles, apparent work of adhesion, and coefficient of static friction data are found to be similar to those of OTS, (see Table 1).[18]

In some devices, the lubrication properties of a film may be more critical than in others. Fig. 8 shows a complex torque multiplying transmission. In this example, the large number of gears brings the issue of friction and lubrication to the foreground. When the device is in operation, every gear experiences sliding contact at its hub, as well as at every tooth intermeshing location. Initial releases of the device that employed oxidized surfaces were not viable. Attempts at surface modification with OTS were marginally successful, but devices ultimately failed because of the high degree of particle contamination and binding friction on the gear hubs. To date, the only devices of this type to successfully and reliably operate are those that receive octadecene films as part of their release process.

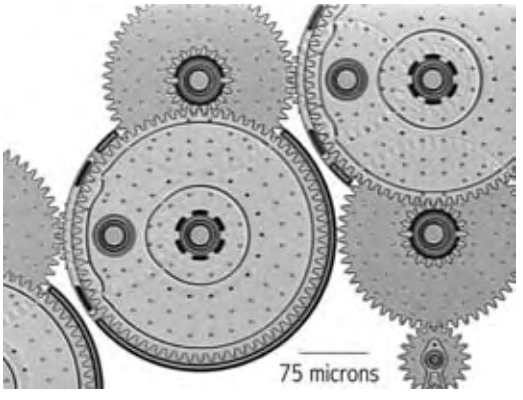


Fig. 8 Micrograph of a complex micromechanical system. Here, there are a large number of sliding contacts (gear hubs and intermeshing gear teeth) that must be properly lubricated for the system to function.

tris-Dimethylaminosilane based monolayers

Another binding chemistry that has been used to successfully attach perfluoroalkyl groups to micromachines surfaces is based on the precursors (tridecafluoro-1,1,2,2,-tetrahydrooctyl)tris-dimethylamino silane (PF8TAS) and (heptadecafluoro-1,1,2,2,-tetrahydrodecyl)tris-dimethylamino silane (PF10TAS).[24] Fig. 4 shows a structural diagram of PF10TAS. These precursors are not commercially available, but can be synthesized from their corresponding trichlorosilanes (FOTS and FDTS, respectively) and dimethylamine. It should be noted that the aminosilane precursors are extremely sensitive to water and must be kept rigorously anhydrous.

The process for applying the aminosilane to micromachines is essentially the same as that used for the chlorosilanes. However, an important distinction is that there is no water vapor added to the chamber during the deposition. Although detailed process parameters are not given, the apparatus used to deposit the aminosilane based monolayers is also similar to that which is used for chlorosilane deposition.[24]

The aminosilane coating PF8TAS has been characterized on Si(100) and on microengines. The AFM analysis on coated Si(100) confirms that the process does not generate particles. Ellipsometry measurements conducted on coated Si(100) in controlled humidity

environment revealed that the PF8TAS films were effective at preventing water sorption.[24]

Thin hard coatings for wear resistance

Although much progress has been made in the area of monolayer films for stiction and friction control, these techniques are not effective at controlling wear. Currently, the mechanism for wear on the micro- and nanoscale is not fully understood, and research has shown that wear in micromachines depends on a number of factors, including several related to the environmental conditions of the device. Although the full details about the mechanism of wear may not be well known, it is generally accepted that one approach to combat the problem of wear is to integrate harder materials at the contacting surfaces. This means that, for silicon structures, materials like diamond, SiC, W, Al_2O_3, or others must be coated on the micromachine surfaces. Of these materials, diamond is the hardest. However, for some applications diamond may not be the best choice as exposure to moderate temperatures in oxidizing ambients can etch diamond. A more chemically inert material with very high hardness is silicon carbide.

Silicon carbide is well known as an attractive material for demanding mechanical and high-temperature applications. Thin SiC coatings share many of the desirable properties of bulk SiC, such as exceptional tribological properties and corrosion resistance. Unfortunately, the conventional method for heteroepitaxial growth of 3C-SiC on Si (based on a dual-precursor CVD) requires high growth temperatures (typically $> 1000°C$).[25] Because released polycrystalline silicon (polysilicon) microstructures cannot withstand such high temperatures without some deformation or warpage, the conventional SiC CVD method cannot be used to coat existing polysilicon devices.

Another CVD process based on the single source precursor 1,3-disilabutane (DSB) has recently been developed.[26] The primary advantage of this method is that the deposition temperature can be much lower than in the conventional CVD method. In fact, the pyrolysis reaction of DSB [shown in Eq. (6) below] can be carried out at temperatures as low as 650°C. However, a deposition temperature of 800°C is chosen for micromachine coatings because it leads to reasonable growth rate, good uniformity, and low film stress.[27] Here, it is demonstrated that this CVD method is suitable for postprocessing of released polysilicon micromachine devices and leads to enhanced reliability and lifetime of microstructures.

$$SiH_3CH_2SiH_2CH_3 \xrightarrow{\Delta} 2SiC + 5H_2 \quad (6)$$

The integration of hard material coatings into the micromachining process scheme is also complicated by the need to coat all surfaces of the device. This requirement leads to deposition processes that occur after the micromachines are fabricated and released. It therefore becomes very important that the coating process be completely uniform and conformal as application of uneven film thickness can alter the stresses on a device and cause it to deform. Recently, much research has focused on atomic layer deposition (ALD) methods. The ALD method is, in principle, capable of depositing perfectly uniform and conformal thin coatings with atomic level control of thickness. Briefly, the method utilizes self-limiting, surface absorption reactions of alternating species to form the desired film one atomic layer at a time (see, e.g., Ref.[28]).

The viability of applying ALD films of Al_2O_3 to released micromachines has been shown.[28,29] It has been indicated that ALD Al_2O_3 films show promise for reduced friction and wear and reduced electrical shorting.[28,29] Another hard material of interest is titania (TiO_2). Owing to its optical, photochemical, and catalytic properties, as well as its biocompatibility and activity in certain sensor applications, TiO_2 has received attention in materials research. TiO_2 is also reported to have low friction and wear, and is chemically stable.[30]

In addition to these methods, the selective deposition of tungsten metal has been explored as an antiwear coating on polysilicon microstructures.[31] This coating is accomplished by heating polysilicon microstructures in a tungsten hexafluoride (WF_6) gas at about 450°C [see Eqs. (7) and (8)]. In this manner, only exposed silicon is coated with W, and the reaction is self-limiting because the deposition of W obscures the underlying Si. W coated micromachines have been shown to exhibit lower wear and improved lifetime vs. uncoated microstructures.[31]

$$2WF_6 + 3Si \xrightarrow{\Delta} 2W + 3SiF_4 \quad (7)$$

$$WF_6 + 3Si \xrightarrow{\Delta} W + 3SiF_2 \quad (8)$$

CONCLUSIONS

Thin film processes and the chemistry of surfaces play important roles in the technological implementation of MEMS and NEMS devices. Thin film deposition of a variety of materials is a mature technology, whereas MEMS surface coating technologies have only recently come of age.

Unlike the commercially employed Si and Al, the development of structural layers that can perform

reliably without further coatings is still in its infancy. Silicon carbide and diamond are emerging as the most promising technologies in this regard.

REFERENCES

1. Petersen, K. Silicon as a mechanical material. Proc. IEEE Electron. Devices **1982**, *70*, 420–457.
2. Madou, M.J. *Fundamentals of Microfabrication*; CRC Press: Boca Raton, 1997.
3. Howe, R.T. Surface micromachining for microsensors and microactuators. J. Vacuum Sci. Technol. B **1988**, *6*, 1809–1813.
4. Abe, T.; Messner, W.C.; Reed, M.L. Effects of elevated temperature treatments in microstructure release procedures. J. Microelectromech. Syst. **1995**, *4*, 66–75.
5. Mulhern, G.T.; Soane, D.S.; Howe, R.T. Supercritical carbon dioxide drying of microstructures. In *Technical Digest*, 7th International Conference on Solid-State Sensors and Actuators, Jun 1993; 296–299.
6. Guckel, H.; Sniegowski, J.J.; Christenson, T.R.; Raissi, F. The application of fine-grained polysilicon to mechanically resonant transducers. Sens. Actuators A **1990**, 346–351.
7. Anguita, J.; Briones, F. HF/H_2O vapor etching of SiO_2 sacrificial layer for large-area surface-micromachined membranes. Sens. Actuators A **1998**, *64* (3), 247–251.
8. Mastrangelo, C.H. Adhesion-related failure mechanisms in micromechanical devices. Tribology Lett. **1997**, *3* (3), 223–238.
9. Komvopoulos, K. Surface engineering and microtribology for microelectromechanical systems. Wear **1996**, *200*, 305–327.
10. Tas, N.; Sonnenberg, T.; Jansen, H.; Legtenberg, R.; Elwenspoek, M. Stiction in surface micromachining. J. Micromech. Microeng. **1996**, *6*, 385–397.
11. Maboudian, R.; Ashurst, W.R.; Carraro, C. Tribological challenges in micromechanical systems. Tribology Lett. **2002**, *12* (2), 95–100.
12. Maboudian, R.; Carraro, C. Surface chemistry and tribology of MEMS. Annu. Rev. Phys. Chem. **2004**, *55*, 35–54.
13. Tanner, D.M. Reliability of surface micromachined microelectromechanical actuators. 22nd International Conference in Microelectronics, IEEE Electron Devices Society: Nis, Yugoslavia, May 2000; 97–104.
14. Ulman, A. *An Introduction to Ultrathin Organic Films*; Academic Press, Inc: San Diego, 1991.
15. Henck, S.A. Lubrication of Digital Micromirror Devices™. Tribology Lett. **1997**, *3* (3), 239–247.
16. Ashurst, W.R.; Yau, C.; Carraro, C.; Maboudian, R.; Dugger, M.T. Dichlorodimethylsilane as an anti-stiction monolayer for MEMS: a comparison to the octadecyltrichlorosilane self assembled monolayer. J. Microelectromech. Syst. **2001**, *9* (4), 41–49.
17. Srinivasan, U.; Houston, M.R.; Howe, R.T.; Maboudian, R. Alkyltrichlorosilane-based self assembled monolayer films for stiction reduction in silicon micromachines. J. Microelectromechan. Syst. **1998**, *7* (2), 252–260.
18. Ashurst, W.R.; Yau, C.; Carraro, C.; Lee, C.; Kluth, G.J.; Howe, R.T.; Maboudian, R. Alkene based monolayer films as anti-stiction coatings for polysilicon MEMS. Sens. Actuators A **2001**, *91*, 239–248.
19. Banga, R; Yarwood, J; Morgan, A.M.; Evans, B.; Kells, J. FTIR and AFM studies of the kinetics and self-assembly of alkyltrichlorosilanes and (perfluoroalkyl)trichlorosilanes onto glass and silicon. Langmuir **1995**, *11* (11), 4393–4399.
20. Maboudian, R.; Ashurst, W.R.; Carraro, C. Self-assembled monolayers as anti-stiction coating for mems: characteristics and recent progress. Sens. Actuators A. **2000**, *82*, 219–223.
21. Mayer, T.M.; de Boer, M.P.; Shinn, N.D.; Clews, P.J.; Michalske, T.A. Chemical vapor deposition of fluoroalkylsilane monolayer films for adhesion control in microelectromechanical systems. J. Vac. Sci. Technol. B. **2000**, *18* (5), 2433–2440.
22. Kim, B.H.; Chung, T.D.; Oh, C.H.; Chun, K. A new organic modifier for anti-stiction. J. Microelectromech. Syst. **2001**, *10* (1), 33–40.
23. Sung, M.M.; Kluth, G.J.; Yauw, O.W.; Maboudian, R. Thermal behavior of alkyl monolayers on silicon surfaces. Langmuir **1997**, *13* (23), 6164–6168.
24. Hankins, M.G.; Resnick, P.J.; Clews, P.J.; Mayer, T.M.; Wheeler, D.R.; Tanner, D.M.; Plass, R.A. Vapor deposition of amino-functionalized self-assembled monolayers on MEMS. In *Proceedings of SPIE: Reliability, Testing, and Characterization of MEMS/MOEMS II*; Ramesham, R., Tanner, D.M., Eds.; SPIE; 2003; Vol. 4980, 238–247.
25. Mehregany, M.; Zorman, C.A.; Roy, S.; Fleischman, A.J.; Wu, C.H.; Rajan, N. Silicon carbide for microelectromechanical systems. Int. Mater. Rev. **2000**, *45* (3), 85–108.
26. Stoldt, C.R.; Carraro, C.; Ashurst, W.R.; Gao, D.; Howe, R.T.; Maboudian, R. A low-temperature CVD process for silicon carbide MEMS. Sens. Actuators A **2002, 97–98**, 410–415.
27. Gao, D.; Wijesundara, M.B.J.; Howe, R.T.; Maboudian, R. Characterization of residual

strain in SiC films deposited using 1,3-disilabutane for MEMS application. J. Microlithography Microfabrication Microsyst. **2003**, *2* (4), 259–264.
28. Mayer, T.M.; Elam, J.W.; George, S.M.; Kotula, P.G.; Goeke, R.S. Atomic-layer deposition of wear-resistant coatings for microelectromechanical devices. Appl. Phys. Lett. **2003**, *82* (17), 2883–2885.
29. Hoivik, N.D.; Elam, J.W.; Linderman, R.J.; Bright, V.M.; George, S.M.; Lee, Y.C. Atomic layer deposited protective coatings for microelectromechanical systems. Sens. Actuators A **2003**, *103* (1–2), 100–108.
30. Ashurst, W.R.; Jang, Y.J.; Magagnin, L.; Carraro, C.; Sung, M.M.; Maboudian, R. Nanometer-thin titania films with SAM-level stiction and superior wear resistance for reliable MEMS performance. Proceedings of the 17th IEEE International Conference on MEMS, 2004, 153–156.
31. Mani, S.S.; Fleming, J.G.; Sniegowski, J.J.; de Boer, M.P.; Irwin, L.W.; Walraven, J.A.; Tanner, D.M.; Dugger, M.T. Chemical vapor deposition coating for micromachines. In *New Methods, Mechanisms and Models of Vapor Deposition*; Materials Research Society Symposium Proceedings; 2000; Vol. 616, 21–26.

Thin Film Science and Technology

T. L. Alford
Department of Chemical and Materials Engineering, Arizona State University, Tempe, Arizona, U.S.A.

J. Kouvetakis
Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona, U.S.A.

J. W. Mayer
Department of Chemical and Materials Engineering, Arizona State University, Tempe, Arizona, U.S.A.

INTRODUCTION

Thin film science and technology is the deposition and characterization of layered structures, typically less than a micron in thickness, which are tailored from the atomic scale upwards to achieve desired functional properties. Deposition is the synthesis and processing of thin films under controlled conditions of chemical processing. Chemical vapor deposition (CVD) and gas-phase molecular beam epitaxy (MBE) are two processes that allow control of the composition and structure of the films. Characterization is the instrumentation that use electrons, X-ray, and ion beams to probe the properties of the film. Epitaxial films of semiconductors are used for their electronic properties to emit light in the infrared (IR) and the ultraviolet rays.

The remaining work discusses two techniques in thin film analysis, Rutherford backscattering spectrometry (RBS) and X-ray diffractrometry with emphasis on strain measurements. Rutherford backscattering spectrometry is illustrated with analysis of silicide formation as an example of thin film reactions. Silicon–germanium–carbon films serve as an example of strain calculations.

The critical aspect of modern thin film technology is the growth of epitaxial layers.[1] We illustrate the importance of epitaxial growth and strain-engineering in producing light emitting films on silicon substrates.

RUTHERFORD BACKSCATTERING SPECTROMETRY

The characterization of thin film structures is now a well-described laboratory technique.[2,3] Fig. 1 is a schematic representation of sources and detectors used in film analysis. The techniques provide film thickness and composition, lattice structure, and epitaxy and strain.

Rutherford backscattering spectrometry analysis determines how the composition varies as a function of depth and is used to characterize thin films and thin film reactions. During ion-beam analysis, the incident particle (typically a proton or helium ion) penetrates into the thin film and undergoes inelastic collisions, with target electrons, and loses energy as it transverses the sample.

Scattering Kinematics

During the penetration of the helium ions a small fraction undergo elastic collisions with the target atom, which defines the backscattering signal. Fig. 2 shows a schematic representation of the geometry of an elastic collision between a projectile of mass M_1 and energy E_0 with a target atom of mass M_2 initially at rest. After collision, the incident ion is scattered back through an angle ϑ and emerges from the sample with an energy E_1. The target atom after collision has a recoil energy E_2. There is no change in target mass, because nuclear reactions are not involved and energies are nonrelativistic. For $M_1 < M_2$, the ratio of the projectile energies, the kinematic factor (K) is given by the following:

$$\mathrm{K} = \frac{E_1}{E_0} = \left[\frac{(M_2^2 - M_1^2 \sin^2 \vartheta)^{\frac{1}{2}} + M_1 \cos \vartheta}{M_2 + M_1}\right]^2 \quad (1)$$

Eq. (1) shows that the energy of the backscattered particle is a function of the incident particle and target atom masses, the scattering angle, and incident energy.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120018083

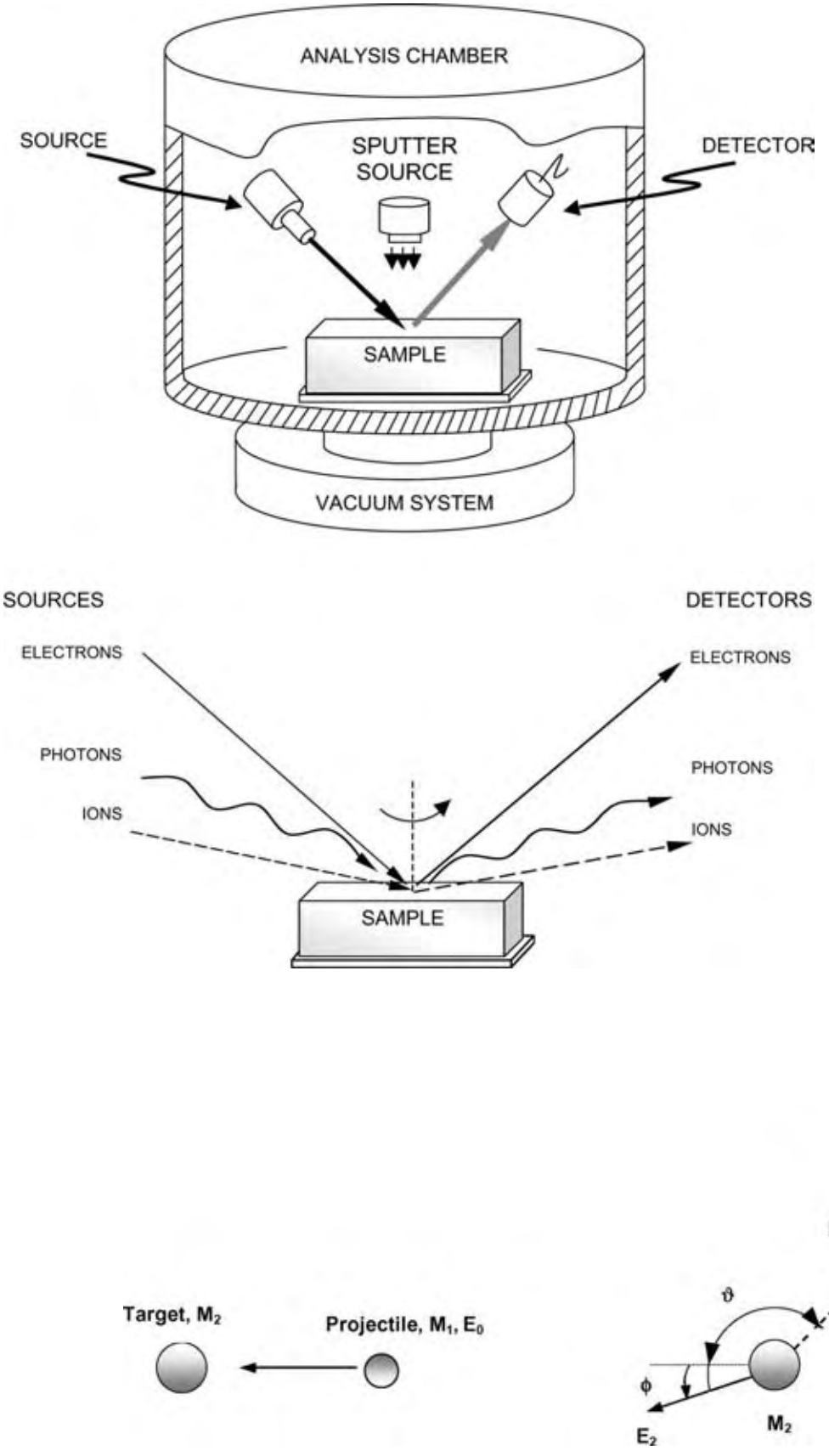


Fig. 1 Schematic of radiation sources and detectors in thin film analysis techniques. Analytical probes are represented by almost any combination of source and detected radiation, i.e., photons in and photons out or ions in and photon out.

Fig. 2 Schematic representation of an elastic collision between a particle of mass M_1 and initial energy E_0 and a target atom of mass M_2. After the collision, the projectile and target atoms have energies of E_1 and E_2, respectively.

Scattering Cross-Section

The identity of target elements is established by the energy of the scattered particles after an elastic collision. The number of target atoms is given by measuring the yield Y, the number of backscattered particles for a given value of incident particles Q. The detector's solid angle is given as Ω. The areal density, the number of atoms per unit area, N_S, is determined from the scattering cross-section $\sigma(\vartheta)$ by:

$$N_S = \frac{Y}{\sigma(\vartheta)\Omega Q} \tag{2}$$

This is shown schematically in Fig. 3. A narrow beam of fast particles impinges on a thin uniform target that is wider than the beam. At a scattering angle ϑ from the direction of incidence, an ideal detector is located that counts each particle scattered in the differential solid angle $d\Omega$. In the simplest approximation, the scattering cross section is given by:

$$\sigma(\vartheta) = \left(\frac{Z_1 Z_2 e^2}{4E}\right)^2 \cdot \frac{1}{\sin^4 \frac{\vartheta}{2}} \tag{3}$$

The cross-sections are relatively large, such that one can detect submonolayers of most heavy mass elements on silicon. For example, the yield of 2.0 MeV helium ions from $10^{14}\,cm^{-2}$ silver atoms (approximately 1/10th of a monolayer) is 800 counts for a current of 100 nano-amperes for 15 min and detector area of 5 msr. This represents a large signal for a small amount of atoms on the surface.

Depth Scale

Light ions, such as helium, lose energy through inelastic collision with atomic electrons. In backscattering spectrometry, where the elastic collision takes place at depth t below the surface, there is energy loss along the inward path and on the outward path as shown in Fig. 4. The total is given by the relationship:

$$\Delta E = \Delta t\left(K\frac{dE}{dX}\bigg|_{in} + \frac{1}{\cos\vartheta} \cdot \frac{dE}{dX}\bigg|_{out}\right) = \Delta t[S] \tag{4}$$

where dE/dX is the rate of energy loss with distance and $[S]$ is the energy loss factor. The particle loses

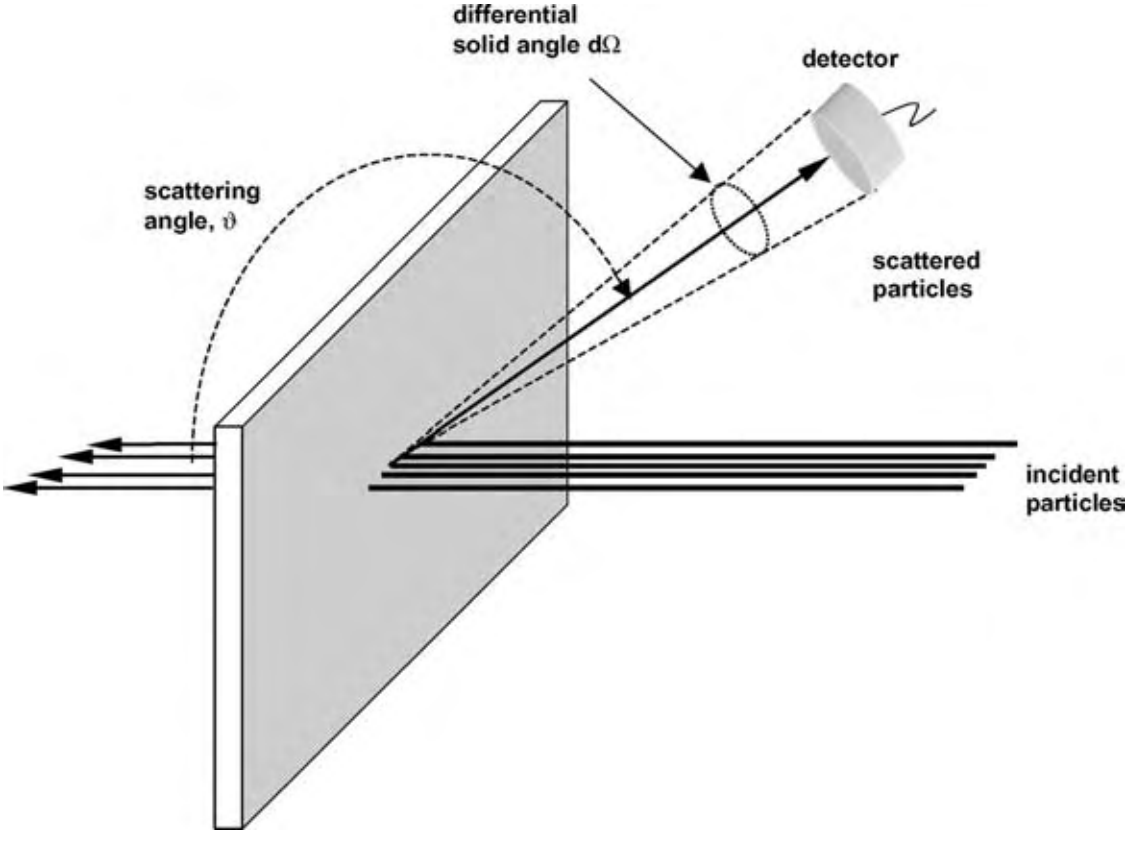


Fig. 3 Schematic of a typical scattering geometry. Only particles that are scattered within the solid angle Ω spanned by the solid-state detector are detected.

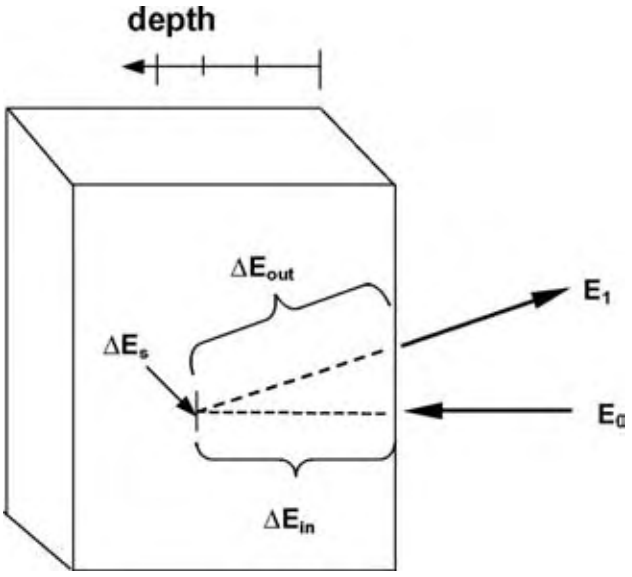


Fig. 4 Energy loss components for a projectile that scatters from depth t. The particle loses energy ΔE_{in} via inelastic collisions with electrons along the inward path. There is energy loss ΔE_s in the elastic scattering process at depth t. There is energy lost to inelastic collisions ΔE_{out} along the outward path. For an incident energy E_0, the energy of the exiting particle is $E_1 = E_0 - \Delta E_{in} - \Delta E_s - \Delta E_{out}$.

energy ΔE_{in} via inelastic collisions with electrons along the inward path. There is energy loss ΔE_s in the elastic scattering process at depth t. There is energy loss due to inelastic collisions ΔE_{out} along the outward path. For an incident energy E_0 the energy of the exiting particle is $E_1 = E_0 - \Delta E_{in} - \Delta E_s - \Delta E_{out}$.

Thin Film Reactions

Heat treatment of deposited thin metal films on Si leads to the formation of silicides. The analysis of these silicide layers is a strong demonstration of the power of Rutherford backscattering thin film technology. The thin metal films react with the Si at temperatures substantially below those indicated in equilibrium phase diagraph. The phase Ni_2Si forms at temperatures around 300°C, where the equilibrium phase diagram shows high temperatures above 800°C for this phase. This is clearly an interface reaction. Rutherford backscattering is a pioneering tool in the silicide formation studies.

We illustrate the analysis of the formation of Ni-silicide in Fig. 5, which plots backscattering yield versus the kinetic energy of the backscattered particles whose incident energy was 2.0 MeV. The solid line is the spectrum from a 200 nm layer of Ni deposited on Si. The thickness is indicated by the width of the energy signal. After heat treatment at 250°C, there is a step in the rear edge of the Ni signal and the forward edge in the Si signal. Analysis of the ratios of the heights of the Ni to Si signals indicates a composition of Ni_2Si. The phase Ni_2Si was confirmed by X-ray diffraction measurements (not shown). Upon further heat treatment, the width of the energy steps increased indicating that the compound has grown in thickness. By measurement of the thickness of the film as function of time and temperature, one can determine the growth mechanism (diffusion limited with $t^{1/2}$ growth rate). Other silicides, such as $CrSi_2$, exhibit limited interfacial interaction (growth proportional to t). One can also implant inert markers of Ar markers into the films and measure the displacement of the markers during heat treatment. If the marker displaces toward the surface, the metal is the moving species. If the marker displaces deeper into the sample, then the silicon is the moving species. Further examples of these studies are given in Tu, Mayer, and Feldman.[1]

ION CHANNELING

When a single-crystal sample is mounted on a goniometer such that a major crystallographic axis of the sample is aligned within approximately 0.1° or 0.5° of the incident beam, the crystal lattice can steer the trajectories of the incident ions penetrating into the crystal. It is this steering of the incident energetic beam that is known as ion channeling; the atomic rows and planes are guides that steer the energetic ions along the channels between rows and planes. The channeled ions do not approach close to the lattice atoms to be backscattered. Hence, the range (the total distance that a ion penetrates a solid) is increased by several fold. The reduced probability of scattering results in a two-orders of magnitude reduction in the yield by an aligned spectrum compared to that when the incident ions are misaligned from the lattice atoms, a random spectrum.

Fig. 6 shows schematically a random and aligned spectrum for MeV helium ions incident on silicon. The characteristic feature of the aligned spectrum is the peak at the high energy end of the spectrum. This peak is a result of ions scattered from the outermost layer of atoms directly exposed to the incident beam. This peak is referred to as the surface peak. Behind the surface peak, at lower energies, the aligned spectra drops to a value of 1/40th of the silicon random spectrum indicating that nearly 97% of the incident ions are channeled and do not make close impact collisions with the lattice atoms. The rise in the aligned spectrum at lower energies represent the ions that are dechanneled, deflected from the steering by the lattice

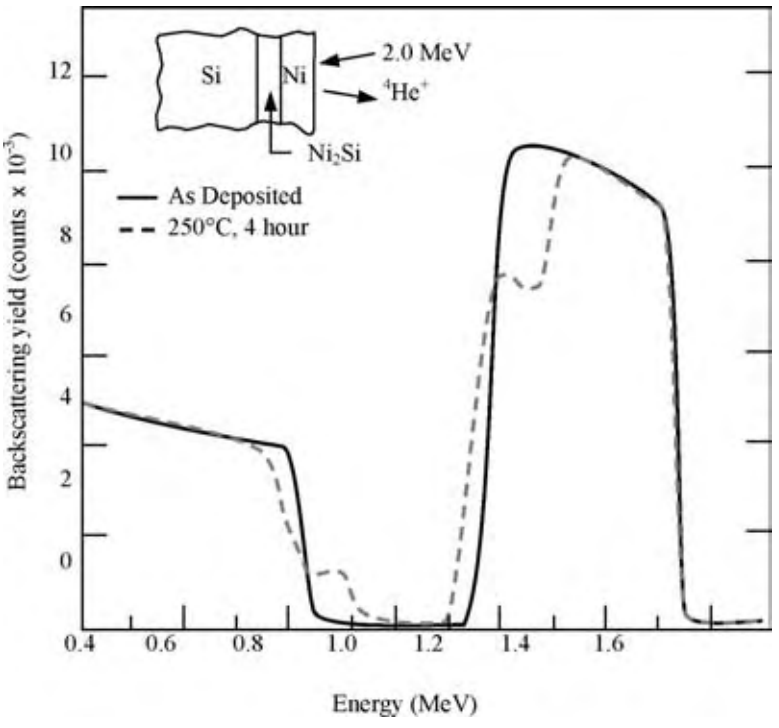


Fig. 5 Rutherford backscattering spectra of a 200-nm Ni thin film on a silicon substrate before and after annealing. *(View this art in color at www.dekker.com.)*

atoms, which can then collide in close impact collisions with the lattice atoms and hence directly contribute to the backscattering spectra. The ratio between the aligned minimum yield and random yield for the same channel number gives the χ_{min} value. The crystalline quality is given by the values of χ_{min} at a specific energy, e.g., it is equal to 0.028 for Si when using 2.0 MeV helium ions.

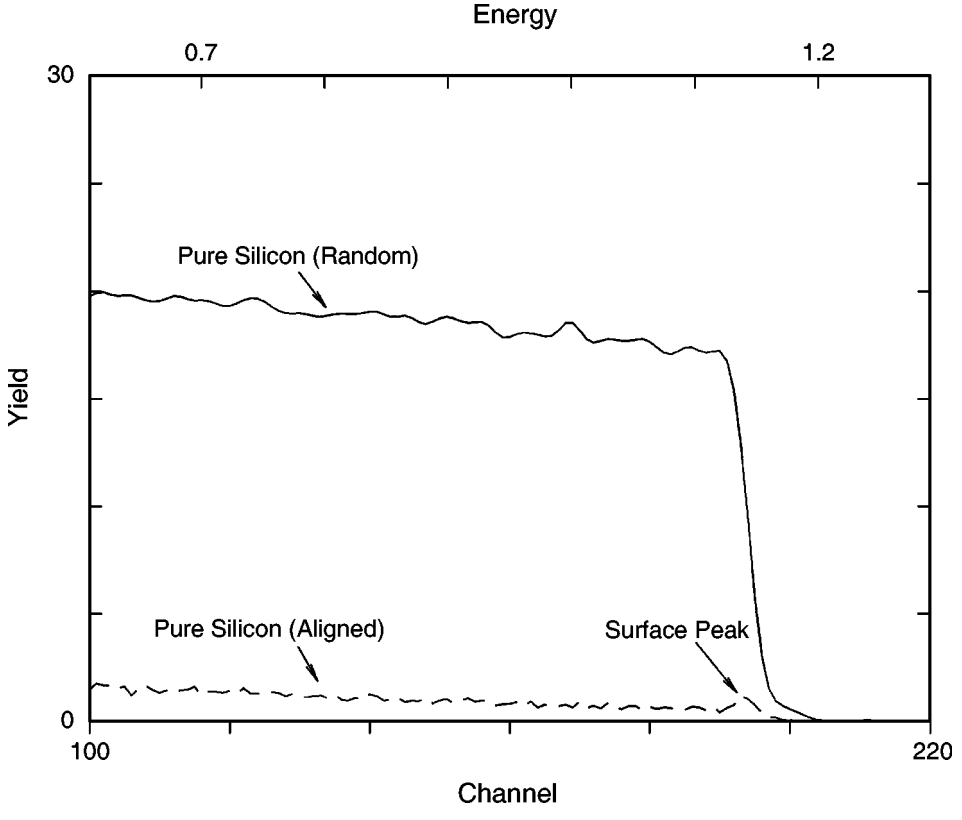


Fig. 6 Random and aligned (channeled) backscattering spectrum from a single crystal. The surface peak corresponds to the small peak at the high energy end of the spectrum signal. The yield behind the peak is reduced because the atoms are shielded from close encounter elastic collisions from the ion beam that is channeled along the axial rows of the crystal.

The application of channeling to RBS is used to determine the amount of damage in ion-implanted single-crystal silicon and the lattice location of ion-implanted dopant atoms. One important example of the contribution of channeling to integrated circuit technology is the analysis of damage evolution during thin film deposition. Modern day ion channeling analysis is done in unison with transmission electron microscopy and/or high resolution X-ray diffraction analysis.

X-RAY DIFFRACTOMETRY

Bragg derived a description for coherent scattering from an array of periodic scattering sites, i.e., atoms in a crystalline solid. The scalar description of diffraction considers the case of monochromatic radiation impinging on two sheets of atoms in the crystal spaced d_{hkl} between reflecting planes. The wavelength λ of the radiation is smaller than the interatomic spacing d_{hkl} of the specific (hkl) planes.

Bragg invoked the Law of Reflectivity (or Reflections) that states that the scattering incident angle and exiting angle must be equal, $\vartheta_{in} = \vartheta_{out}$ under the condition of coherent scattering. The wavelets scattered by the atoms combine to produce constructive inference if the total path difference $2^*\Delta P$ for the reflected waves equals integer (n) multiples of λ:

$$n\lambda = 2^*\Delta P = 2^* d_{hkl} \sin\vartheta \quad (5)$$

Hence, Bragg's Law, $n\lambda = 2d_{hkl}\sin\vartheta$ defines the condition for diffraction.

The typical X-ray spectrum is a plot of intensity verses angle, e.g., 2ϑ. The phase can be indentified by comparing the spectrum to the Powder Diffraction File compiled by the international Center for Diffraction Data, formerly known as Joint Committee on Powder Diffraction Standards.

Strain Measurements

The strain in a layer is determined by comparing the perpendicular and parallel lattice spacings (a) of the film to that of the underlying substrate a_{sub}, determining if they are larger or smaller and by how much. Fig. 7 shows a Si film containing Ge (large circles) on a substrate of Si to illustrate the perpendicular and parallel lattice constant. The film is psedomorphic because the lattice lines up with that of the substrate atoms. These values are expressed as $[\Delta a/a]_\perp$ and $[\Delta a/a]_{||}$, for the difference in the perpendicular and parallel lattice constants, respectively. Since these differences are measured relative to the substrate, the

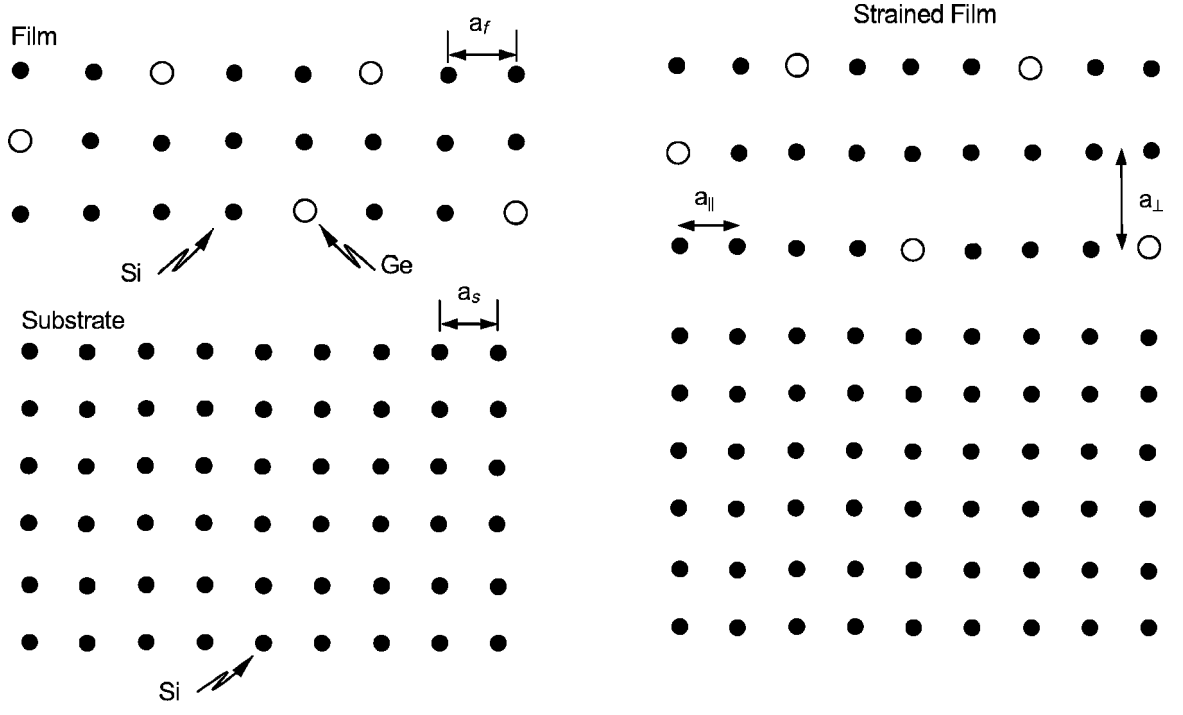


Fig. 7 Schematic of a Si film containing Ge (large circles) on a substrate of Si to illustrate the perpendicular and parallel lattice constant.

substrate lattice spacings must be known. From these parameters, a strain value is calculated by determining the difference in the lattice spacings of the film from its bulk relaxed state.

Silicon–Germanium–Carbon Films on Silicon

The application of X-ray diffraction in thin film samples is illustrated by diffraction measurement of epitaxial layers of Si–Ge–C on silicon. Initially, a scan is taken about a plane parallel to the surface; this is called a symmetrical scan, since ϑ_{in} and ϑ_{out} are identical. In Fig. 8, the Si substrate has an (0 0 1) orientation and the first allowed reflection that is parallel to the surface is the (0 0 4) reflection. The symmetrical scan in this case provides information about the perpendicular spacing ($a_\perp$) between the places. To obtain information about both $a_\perp$ and parallel spacing ($a_{||}$), an asymmetrical scan is taken. This involves scanning a plane that is not parallel to the surface. In Si this is typically the $(2\,2\,\bar{4})$ and (2 2 4) reflection as shown in Fig. 9. In one case the spectrum is a glancing incidence (2 2 4) reflection and the other is a $(2\,2\,\bar{4})$ glancing exiting reflection. In each of these scans, note the existence of a peak from the Si substrate and from the Si–Ge–C film. Satellite peaks above and below the primary reflection are due to thickness interference fringes. In strain calculations, the substrate is assumed to be unstrained and hence all measurements are made relative to the substrate's lattice constant.

Calculation of Strain

The perpendicular and parallel lattice constants are calculated by first measuring the angular peak separation from glancing incident {2 2 4} reflections, ω_1, and the glancing exit $(\bar{2}\,\bar{2}\,4)$ reflections, ω_2. The deviation of the Bragg angle between the substrate and layer, $\Delta\vartheta$, is calculated from

$$\Delta\vartheta = 0.5(\omega_1 - \omega_2) \tag{6}$$

Due to tetragonal distortion in the epitaxial layer, the angle between the (2 2 4) planes and the surface in the substrate will not be the same as this angle in the film. This difference, $\Delta\psi$ was calculated from:

$$\Delta\psi = 0.5(\omega_1 - \omega_2) \tag{7}$$

The perpendicular lattice constant of the film is calculated by

$$[\Delta a/a]_\perp = \Delta\vartheta \cot(\vartheta_B) + \Delta\psi \tan(\psi) \tag{8}$$

where ϑ_B is the Bragg angle for the reflection measured and ψ is the angle between the (2 2 4) planes and the surface. Both angles refer to the substrate. The parallel lattice constant of the film is calculated from:

$$[\Delta a/a]_{||} = -\Delta\vartheta \cot(\vartheta_B) - \Delta\psi \cot(\psi) \tag{9}$$

Negative values for $[\Delta a/a]_\perp$ and $[\Delta a/a]_{||}$ indicate that the dimension in the film is larger than that of the substrate. When the perpendicular lattice constant

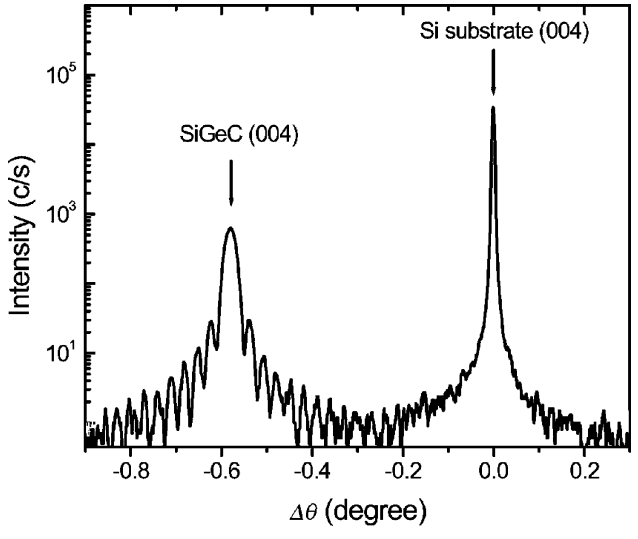


Fig. 8 Rocking curve XRD of an epitaxial layers of Si–Ge–C layers on silicon. The symmetrical scan is of the first allowed reflection that is parallel to the surface, i.e., the (0 0 4) reflection. Note the existence of a peak from the Si substrate and from the Si–Ge–C film.

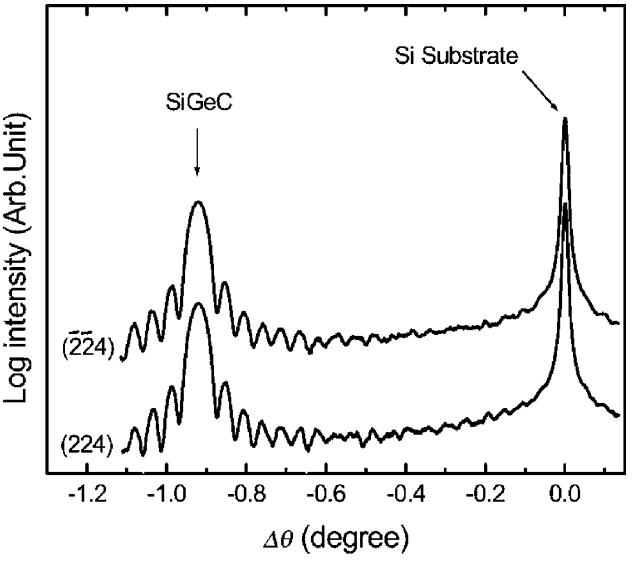


Fig. 9 A glancing incidence (2 2 4) reflection and a glancing existing reflection (2 2 4) of an epitaxial layers of Si–Ge–C layers on silicon. Note the existence of a peak from the Si substrate and from the Si–Ge–C film.

for a pseudomorphic film is larger than the substrate, the film has a larger unit cell and is in compression for $[\Delta a/a]_{\perp} = 0.016638$ and $[\Delta a/a]_{||} = 0.016618$.

GROWTH OF THIN FILMS BASED ON Si–Ge–Sn SYSTEM BY CVD AND GAS PHASE MBE

Silicon is the most desirable substrate for the growth of semiconductor materials. Virtually defect-free Si wafers are available at low cost, and the range of applications of any semiconductor grown on Si can be enhanced by integration with silicon-based microelectronics. The growth of $Si_{1-x}Ge_x$ films on Si(1 0 0) has been the subject of intensive studies over the past two decades owing to their many important applications in high speed microelectronic devices.[4–7] A number of different methods have been used for the heteroepitaxial growth of $Si_{1-x}Ge_x/Si$, but the two most commonly employed techniques are molecular beam epitaxy utilizing solid Si and Ge, and ultrahigh vacuum chemical vapor deposition (UHV-CVD) or gas-source molecular beam epitaxy (GSMBE) utilizing SiH_4 and GeH_4 or Si_2H_6 and Ge_2H_6. There are two ultimate, but also diverse, objectives in the growth of $Si_{1-x}Ge_x/Si$. The first is the achievement of defect-free $Si_{1-x}Ge_x$ layers, which may take the form of strained layer superlattices, while the second is the growth of self-assembled coherent $Si_{1-x}Ge_x$ islands or quantum dots.

It has been known for many years—on theoretical grounds—that the Si–Ge–Sn alloy system should have very interesting properties, especially as high efficiency IR devices. This has stimulated intense experimental efforts to grow such compounds, but for many years the resulting material quality has been incompatible with device applications. Recently, we have achieved growth of device quality Sn–Ge and Si–Ge–Sn films using novel CVD methods. This is an important development for several reasons. First, Sn_xGe_{1-x} alloys have been predicted to undergo a transition from indirect to direct gap semiconductors so that this material may lead to the first direct-gap semiconductor fully integrated with Si technology.[7–9] Second, device-quality Sn_xGe_{1-x} layers of arbitrary thickness can be deposited directly on Si and these can be used as ''virtual substrates'' for the growth of $Ge_{1-x-y}Si_xSn_y$ ternary analogs. The fabrication of $Ge_{1-x-y}Si_xSn_y$ makes it possible to decouple strain and band gap engineering to achieve unique systems cover a wide range of operating wavelengths in the IR and new device structures that lead to novel photonic devices based entirely on group IV materials.[10]

In the following sections, we present the fabrication and discuss properties of device quality, strain-free Sn_xGe_{1-x} films, as well as strained engineered $Ge_{1-x-y}Si_xSn_y$ layers grown over time directly on Si wafers. We also describe synthesis of films and nanometer-scale islands of $Si_{1-x}Ge_x$ grown on Si(1 0 0) substrates via a unique single-source molecular precursor method. This new approach allows precise control of concentration and structure at the atomic level and it is particularly useful for development of compositionally homogeneous and uniform assemblies of nanoscale structures.

Growth of Si–Ge–Sn on Si for Strain-Balanced Heterostructures Via CVD

The binary Sn_xGe_{1-x} alloys are grown by a specially developed CVD method.[11–17] The combination of SnD_4 with high-purity H_2 (15–20% by volume) remains remarkably stable at 22°C for extended time periods. This formulation provides the simplest possible CVD source of Sn atoms for the growth of novel Sn–Ge systems. Depositions were conducted in a custom-built ultra-high-vacuum CVD reactor on Si(0 0 1) wafers. Growth temperatures between 250° and 350° produced thick films (50–500 nm) with Sn concentrations up to 20%, as measured by Rutherford backscattering. Fig. 10 shows a comparison between random and aligned RBS spectra for a $Sn_{0.02}Ge_{0.98}$ sample and a $Sn_{0.12}Ge_{0.88}$ sample. The ratio χ_{min} between the aligned and random peak heights is 4% in the $x = 0.02$ sample and about 30% in the $x = 0.12$ sample for both Ge and Sn. This provides proof that Sn occupies substitutional sites in the average diamond structure. The $\chi_{min} = 4\%$ value closely approaches the practical limit of about 3% for structurally perfect Si, which is unprecedented for a binary crystal grown directly on Si. The microstructural properties of the films were investigated by XTEM. Electron micrographs demonstrating nearly defect-free growth of $Sn_{0.06}Ge_{0.94}$ are shown in Fig. 11. The images show that the predominant defects accommodating the large misfit between the alloys and the Si substrate are Lomer edge dislocations at the interface with no dislocation cores propagating to the film surface. These are parallel to the interface plane and do not degrade the film quality. The surfaces of the films are very smooth, continuous, and atomically flat. Electron diffraction and high-resolution X-ray studies (including rocking curves and reciprocal space maps) show a monotonically increasing average lattice constant as a function of the Sn-concentration, with no evidence for a significant tetragonal distortion or strain. The full width at half maximum of the X-ray peaks range from 0.25 to 0.50° indicating tightly aligned crystal mosaics. Reciprocal space maps of the (0 0 4) reflection show that there is no epilayer tilt between the Si and the GeSn (0 0 4) Bragg planes. Comparisons of the (2 2 4) grazing incidence and

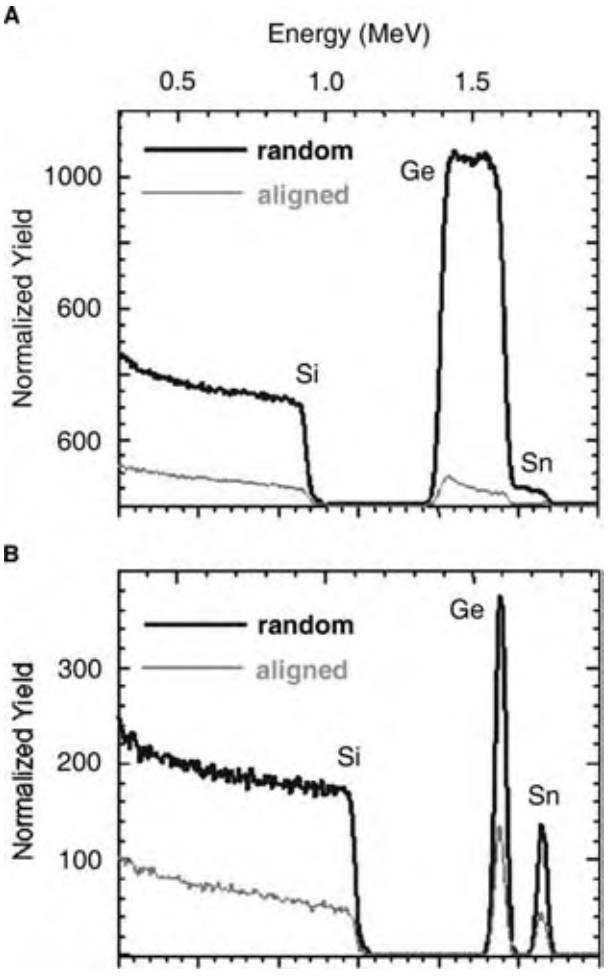


Fig. 10 (A) RBS aligned and random spectra of $Ge_{0.98}Sn_{0.02}$ with near perfect crystallinity. Channeling of Sn and Ge approaches the theoretical limit of pure Si. (B) Aligned and random spectra for $Ge_{0.88}Sn_{0.12}$ show the same χ_{min} for both Ge and Sn indicating that the entire Sn content is substitutional. (From Ref.[11].)

grazing exit $\omega/2\theta$ scans, show that the GeSn layers are 100 relaxed and this is confirmed with reciprocal space maps of the (0 0 4) peak and the off axis and (2 2 4) reflection. Specular X-ray reflectivity scans (reflectivity vs. sample angle ω) were obtained to characterize the density and layer thickness and determine the average roughness of the substrate film interface. The data for samples with 5% Sn content are shown by the curve in Fig. 12. Note that the reflectivity drops off steeply at 1100 sec, yielding the density of the Ge–Sn layer to be 5.32 g/cm^3,which is close to Ge. The fringes indicate that the thickness of the Ge–Sn layer is 57.8 nm, which is virtually identical to the RBS value of 58 nm. The fringes are reduced in amplitude at larger angles and disappear at about 6000 sec due to minor surface roughness. The best fit shows a roughness thickness for the substrate Si and the Ge–Sn layer to be 1.2 nm and 0.8 nm, respectively. The presence of a thin oxide layer on the film surface reduces the amplitude of the reflectivity at larger sample angles.

The ternary $Ge_{1-x-y}Sn_xSi_y$ are grown on Si (1 0 0) via $Ge_{1-x}Sn_x$ alloy buffer layers.[18,19] The depositions are carried out via ultra-high-vacuum chemical vapor deposition of unimolecular hydrides with direct Si–Ge bonds [SiH_3GeH_3, $Si(GeH_3)_4$, $(GeH_3)_2SiH_2$] with SnD_4. The crystal structure, elemental distribution, and morphological and bonding properties of the $Ge_{1-x-y}Sn_xSi_y$/$Ge_{1-x}Sn_x$ heterostructures are characterized by high-resolution TEM, including electron energy loss nanospectroscopy, high resolution X-ray diffraction, AFM and Raman. These techniques demonstrate growth of perfectly epitaxial, uniform, and highly aligned layers with atomically smooth surfaces and monocrystalline structures. The Raman spectral shifts are consistent with lattice expansion produced by the Sn incorporation into Si–Ge tetrahedral sites. The films possess a variable and controllable range of compositions (10–20% Si, 60–85% Ge, and 1–15% Sn), and exhibit lattice constants above and below that of bulk. A material with a tunable lattice constant above and below that of Ge is synthesized directly onto Si substrates. Representative RBS spectra and high-resolution TEM micrographs are shown in Fig. 13. The data are obtained from a sample with concentrations $Si_{0.20}Ge_{0.72}Sn_{0.08}$ grown on Si(1 0 0) via a $Ge_{0.98}Sn_{0.02}$

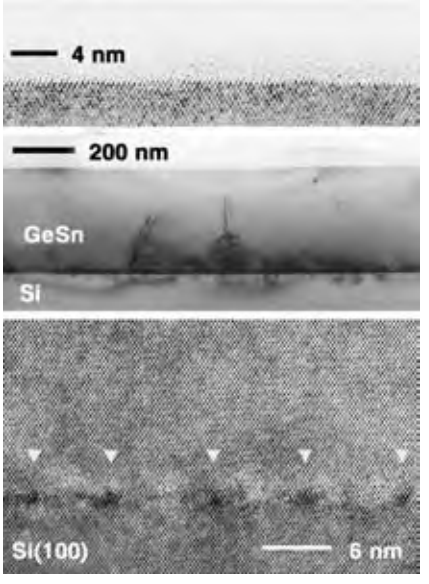


Fig. 11 Cross-sectional electron micrographs of $Ge_{0.94}Sn_{0.06}$. Top panel shows atomically flat film surface morphology, middle panel shows the exceptional uniformity of the film thickness. Bottom panel is a high-resolution electron micrograph of the interface region showing virtually perfect epitaxial growth. Arrows indicate the location of misfit dislocations. (From Ref.[11].) *(View this art in color at www.dekker.com.)*

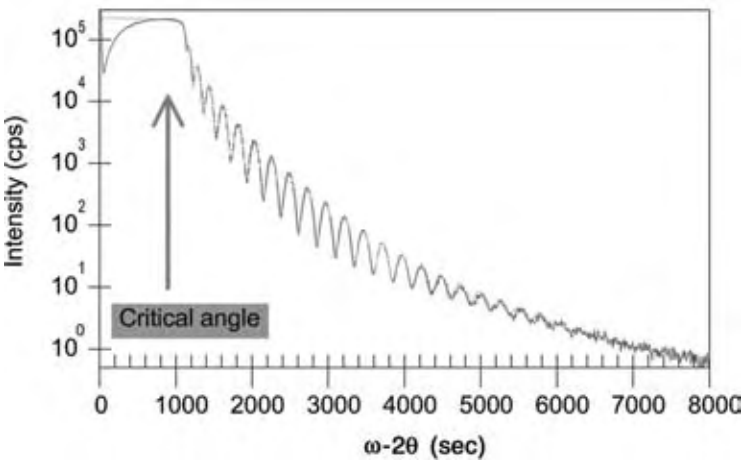


Fig. 12 High-resolution X-ray reflectivity of a $Ge_{0.95}Sn_{0.05}$ film grown on Si(1 0 0). The reflectivity data give a layer thickness of 58–60 nm and density close to that of Ge (2% larger). (Courtesy of Dr. Stefan Zollner.) *(View this art in color at www.dekker.com.)*

buffer layer at 320°C. The high degree of RBS He ion channeling indicates aligned, single-phase material in which the constituent elements in the heterostructure occupy random substitutional sites in the same average diamond cubic lattice. Note that the extent of channeling is identical for Si, Ge, and Sn indicating substitutionality of the elements in both the Ge–Sn buffer layer and the Si–Ge–Sn film of the sample. Fig. 13 shows a dark field image of the entire $Si_{0.20}Ge_{0.72}Sn_{0.08}/Ge_{0.98}Sn_{0.02}/Si$ heterostructure indicating highly uniform layers free from threading defects. High-resolution micrographs of the $Si_{0.20}Ge_{0.72}Sn_{0.08}/Ge_{0.98}Sn_{0.02}/Si(1\,0\,0)$ interfaces show highly commensurate microstructures. The $Si_{0.20}Ge_{0.72}Sn_{0.08}$ and $Ge_{0.98}Sn_{0.02}$ layers are nearly lattice matched and their interface is virtually defect-free as shown Fig. 13D. The large lattice mismatch between $Ge_{0.98}Sn_{0.02}$ and Si(1 0 0) is accommodated by periodic edge dislocations located at the interface as shown in Fig. 13C.

The strain properties of these materials have been investigated by high resolution XRD. Studies involving samples that span a wide range of concentrations show fully strained (tensile and compressively strained), as well as relaxed Si–Ge–Sn films are obtained on strain-free Ge–Sn buffer layers. These results show that strain engineering can be achieved in Si–Ge–Sn heterostructures and multilayers by tuning the lattice parameter of the Ge–Sn buffer layer.

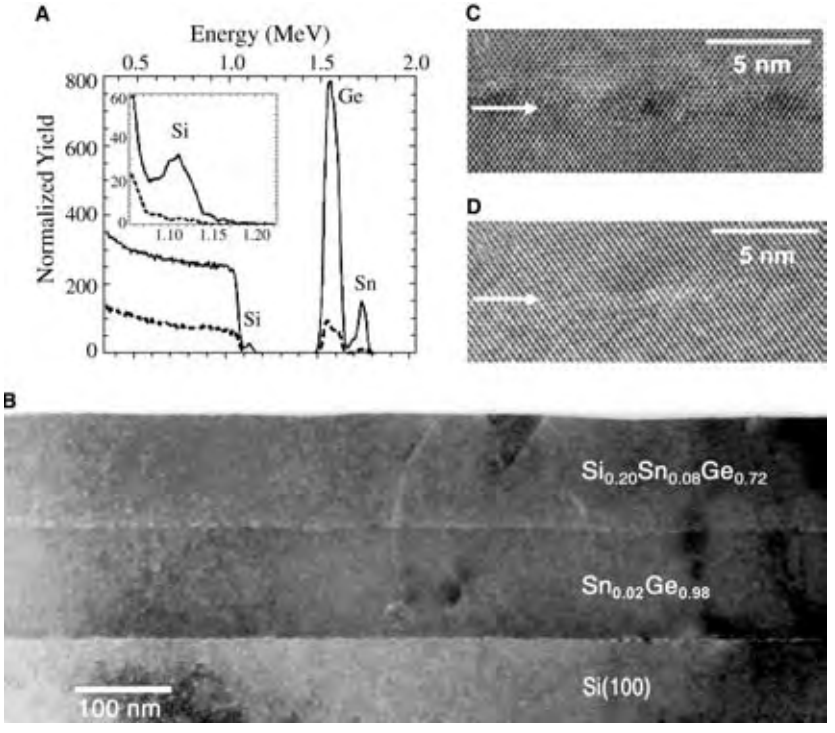


Fig. 13 RBS aligned and random spectra of $Si_{0.20}Sn_{0.08}Ge_{0.72}$ epilayer and $Sn_{0.02}Ge_{0.98}$ buffer layer showing a highly aligned heterostructure. Inset: (A) magnified view of the Si peak indicating complete substitutionality of Si in the Sn–Ge lattice; (B) XTEM of the entire heterostructure; (C) $Si/Sn_{0.02}Ge_{0.98}$ interface (indicated by arrow); and (D) $Si_{0.20}Sn_{0.08}Ge_{0.72}/Sn_{0.02}Ge_{0.98}$ interface. (From Ref.[19].)

A host of novel strained engineered optical and electronic devices have been designed based on this concept and are currently being fabricated and tested. It is interesting to note that the strain is reliably robust up to at least 400–500°C (400°C is the growth temperature of the films).

Growth of SiGe and Si_4Ge Epitaxial Materials by Gas Phase MBE of H_3SiGeH_3 and $Ge(SiH_3)_4$

Growth of $Si_{1-x}Ge_x$ films and nanometer scale islands has been demonstrated by using single source molecular hydrides containing direct Si–Ge bonds.[20] The growth process occurs via single source GSMBE directly on Si(1 0 0) and the concentration x of the film is predetermined by tailoring the composition of the molecular precursor. The technique was demonstrated by growing $Si_{0.5}Ge_{0.5}$ and $Si_{0.8}Ge_{0.2}$ epitaxial films and coherent islands on Si(1 0 0) via thermal dehydrogenation of H_3SiGeH_3 and $Ge(SiH_3)_4$, respectively, between 475 and 700°C. Note that the entire content of Si and Ge of the precursors is incorporated in the film. The major advantages of using single molecular sources with preformed Si–Ge bonds for the MBE growth of $Si_{1-x}Ge_x$ lie not only in the a priori control over the composition, but also the film morphology can be controlled by the adjustment of a single kinetic parameter, i.e., the flux rate of the precursor, at a given temperature.

The growth of the $Si_{0.5}Ge_{0.5}$ and $Si_{0.8}Ge_{0.2}$ films on Si(1 0 0) substrates is conducted in the sample chamber of a low-energy electron microscope (LEEM) where in situ observation of the GSMBE growth process takes place in real time. Fig. 14 shows a sequence of frame-captured LEEM video images over a field of view of 8 μm of $Si_{0.5}Ge_{0.5}$ growth on Si(100)-(2x1) at 550°C using H_3Si-GeH_3. The images were taken with the ($^1/_2$, 0) diffraction beam such that the (2 × 1) and (1 × 2) terraces separated by single-height atomic steps alternate in contrast from dark to bright due to the rotation of the dimer reconstruction across a step. The last LEEM frame taken at 46 min showing island formation was a bright-field image acquired using the (0, 0) beam. In Fig. 14, contrast reversal in the (2 × 1) and (1 × 2) terraces was observed during the time lapse from 0 to 190 sec as growth of $Si_{0.5}Ge_{0.5}$ proceeded. Such contrast reversal is indicative of layer-by-layer growth. The growth of a full $Si_{0.5}Ge_{0.5}$ monolayer (ML) is completed every 30 sec. The layer-by-layer growth continued after the completion of 6 ML, after which the LEEM contrast of the surface became very diffuse. The loss of contrast was due to new layers growing on top of incomplete layers, and this kind of growth results in a rough surface even though the domains are two-dimensional.

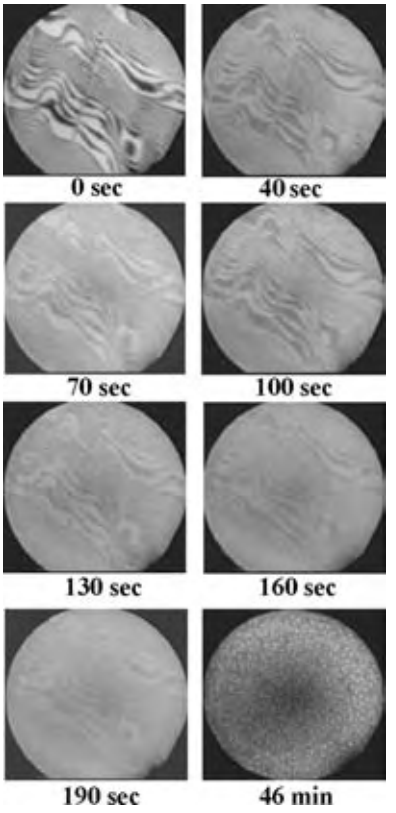


Fig. 14 Frame-captured LEEM video images showing Stranski–Krastanov growth of $Si_{0.5}Ge_{0.5}$. The elapsed time during growth is indicated under each frame. Approximate Si–Ge coverage for each frame: 0 ML at 0 sec; 1 ML at 40 sec; 2 ML at 70 sec; 3 ML at 100 sec; 4 ML at 130 sec; 5 ML at 160 sec; 6 ML at 190 sec; and 3D islands at 46 min. Field of view is 8 μm. (From Ref.[20].)

Three-dimensional (3D) islands began to appear after about 18 min of growth. The last frame shown in Fig. 14 is a bright-field LEEM image taken after 46 min of growth with the 3D islands appearing as bright spots. The LEEM results indicate that the growth mode of the $Si_{0.5}Ge_{0.5}$ is Stranski–Krastanov. AFM images show that the $Si_{0.5}Ge_{0.5}$ islands are primarily large dome-shaped structures. The dome shape of the islands with (1 1 3) and (15 3 23) facets is clearly demonstrated in the XTEM images, which also show that the islands are completely coherent. Similar islands, grown at 600 and at 700°C are also found to be completely coherent demonstrating the fact that the single source approach is capable of producing large coherently strained dome-shaped islands at a wide range of conditions.

Growth of $Si_{0.8}Ge_{0.2}$ films on Si(1 0 0) was conducted using the $Ge(SiH_3)_4$ gaseous precursor. In situ real-time LEEM observations indicated that the Stranski–Krastanov growth mode was also in operation in this case. AFM and XTEM images of films

grown at 500°C indicate growth of completely coherent islands with uniform size and spatial distribution. The images of films gown at 600 and 700°C show that large (200 and 400 nm) faceted dome-shaped islands are produced at both temperatures. These are both strain-free and dislocated. Selected area electron diffraction patterns show a lattice constant of 0.548 ± 0.001 nm, very close to the ideal 0.5476 nm for $Si_{0.8}Ge_{0.2}$. High spatial resolution electron energy loss spectroscopy (EELS) used to verify that the composition of the $Si_{1-x}Ge_x$ films indeed reflects the stoichiometry of the unimolecular precursor used for growth. Typical EELS composition line-scans across the interface from the substrate to the island using a electron beam with 1 nm diameter are given in Fig. 15. The EELS composition profile of a $Si_{0.5}Ge_{0.5}$ island grown at 700°C in Fig. 15 shows an almost constant 50% Ge within the island. Similarly, the EELS composition profile of a $Si_{0.8}Ge_{0.2}$ island shows a nearly constant 20% Ge. No segregation of Ge is observed in any of our EELS composition profiles of the islands.

The single-source MBE method growth of $Si_{1-x}Ge_x$ films permit control of composition x at the atomic level via the design of the single-source gaseous precursor containing precise atomic arrangements with direct Si–Ge bonds. Uniform composition reflecting the stoichiometry of the precursor is observed in all cases without any segregation of either Ge or Si. The growth of $Si_{1-x}Ge_x$ films proceeds via the Stranski–Krastanov growth mode. Morphological control of the size and shape of the islands is achieved by simple adjustments of the flux rate of the precursor and the growth temperature.

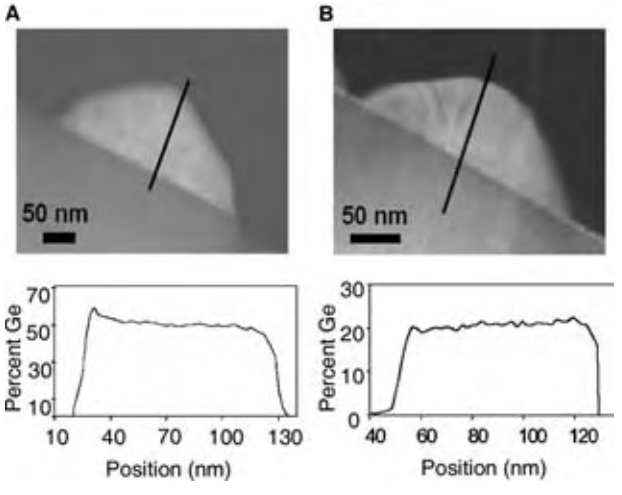


Fig. 15 Ge composition line scans determined by high-resolution EELS on (A) a $Si_{0.5}Ge_{0.5}$ island grown at 700°C, and (B) a $Si_{0.8}Ge_{0.2}$ island grown at 500°C. The profiles shown below each XTEM image correspond to the line across the substrate/island interface. Scan direction is from the substrate to the vacuum. (From Ref.[20].)

CONCLUSIONS

Thin film science and technology is a dynamic field. Illustrated by growth of thin films for silicon technology for optical and high speed microelectronics devices. Accepted technology is readily available. The key is control of the composition and structure of thin films and islands fully integrated with Si technology. The Ge–Si–Sn system makes it possible to decouple strain and ban-gap engineering to each achieve unique photonic devices.

In this work, the focus has been based entirely on group IV materials. This excitement can also be realized in photonic and high speed microelectronics. Growth methods and analysis are also readily available.

ACKNOWLEDGMENT

The authors acknowledge support from NSF in their investigations.

REFERENCES

1. Tu, K.-N.; Mayer, J.W.; Feldman, L.-C. *Electronic Thin Film Science*; Macmillan Publishing Company: New York, 1992.
2. Feldman, L.C.; Mayer, J.W. *Fundamental of Surface and Thin Film Analysis*; PTR Prentice Hall: New Jersey, 1986.
3. Nastasi, M.; Mayer, J.W.; Hirvonen, J. *Ion–Solid Interactions: Fundamentals and Applications*; Cambridge University Press: UK, 1996.
4. Patton, G.L.; Harame, D.L.; Strock, J.-M.; Meyerson, B.S.; Scilla, G.-S. Graded-SiGe-base, poly emitter heterojunction bipolar transistors. IEEE Electron Device Lett. **1989**, *10*, 534–536.
5. Mooney, P.M.; Chu, J.O. SiGe Technology: Heteroepitaxy and high-speed microelectronics. Ann. Rev. Mater. Sci. **2000**, *30*, 335–362.
6. Tromp, R.M.; Ross, F.M. Advances in situ electron microscopy: Growth of SiGe on Si. Ann. Rev. Mater. Sci. **2000**, *30*, 431–449.
7. Jenkins, D.W.; Dow, J.D. Electronic properties of metastable Ge_xSn_{1-x} alloys. Phys. Rev. B. **1987**, *36*, 7994.
8. Mader, K.-A.; Baldereschi, A.; von Kanel, H. Band structure and instability of Ge_xSn_{1-x} alloys. Solid State Commun. **1989**, *69*, 1123.
9. Soref, R.-A.; Friedman, L. Direct-gap Ge/GeSn/Si and GeSn/Ge/Si heterostructures, Superlattice Microstruct **1993**, *14*, 189.
10. Soref, R.A.; Menendez, J.; Kouvetakis, J. Strained Engineered Direct-gap Ge/Sn_xGe_{1-x} Heterodiode and Mult-quantum Well Photodetectors, Lasers, Emitters, and Modulators Grown

on $Sn_ySi_xGe_{1-y-z}$ Buffered Silicon. US Patent 6,897,471, 2005.
11. Bauer, M.R.; Taraci, J.; Tolle, J.; Chizmeshya, A.V.G.; Zollner, S.; Smith, D.-J.; Menendez, J.; Hu, C.W.; Kouvetakis, J. Ge-Sn semiconductors for band-gap and lattice engineering. Appl. Phys. Lett. **2002**, *81*, 2992–2994.
12. Taraci, J.; Zollner, S.; McCartney, M.-R.; Menéndez, J.; Santana, M.-A.; Smith, D.-J.; Haaland, A.; Tutukin, A.V.; Gundersen, G.; Wolf, G.; Kouvetakis, J. Synthesis of silicon-based infrared semiconductors in the Ge-Sn system using molecular chemistry methods. J. Am. Chem. Soc. **2001**, *123* (44), 10980–10987.
13. Chizmeshya, A.V.G.; Bauer, M.; Kouvetakis, J. Experimental and theoretical study of deviations from Vegards Law in the $Ge_{1-x}Sn_x$ system. Chem. Mater. **2003**, *15*, 2511–2519.
14. Li, S.F.; Bauer, M.R.; Menéndez, J.; Kouvetakis, J. Scaling law for the compositional dependence of Raman frequencies in GeSi and SnGe alloys. Appl. Phys. Lett. **2004**, *84*, 867–869.
15. Bauer, M.R.; Chizmeshya, A.V.G.; Menendez, J.; Kouvetakis, J. Tunable band structure in diamond cubic tin germanium alloys grown on Si. Solid State Commun. **2003**, *127*, 355–359.
16. Cook, C.S.; Zollner, S.; Kouvetakis, J.; Tolle, J.; Menendez, J. Optical constants and interband transitions of $Ge_{1-x}Sn_x$ alloys ($x < 0.2$) grown on Si, Thin Solid Films **2003**, 455–456, 217–221.
17. Bauer, M.R.; Cook, C.S.; Zollner, S.; Crozier, P.; Chizmeshya, A.V.G.; Kouvetakis, J. GeSn superstructured materials for Si-based optoelectronics. Appl. Phys. Lett. **2003**, *83*, 3489–3491.
18. Bauer, M.R.; Ritter, C.; Crozier, P.; Menendez, J.; Ren, J.; Kouvetakis, J. Synthesis of ternary Si-Ge-Sn semiconductors on Si(1 0 0) via Sn_xGe_{1-x} buffer layers. Appl. Phys. Lett. **2003**, *83* (9), 2163–2165.
19. Aella, P.; Cook, C.; Tolle, J.; Zollner, S.; Chizmeshya, A.; Kouvetakis, J. Structural and optical properties of $Sn_xSi_yGe_{1-x-y}$ alloys. Appl. Phys. Lett. **2004**, *84*, 888–890.
20. Hu, C.W.; Taraci, J.L.; Tolle, J.; Bauer, M.R.; Crozier, P.A.; Tsong, I.S.T.; Kouvetakis, J. Synthesis of highly coherent SiGe and Si_4Ge nanostructures by single-source molecular beam epitaxy of H_3SiGeH_3 and $Ge(SiH_3)_4$. Chem. Mater. **2003**, *15* (19), 3569–3572.

Thin Liquid Film Deposition

Myung S. Jhon
Department of Chemical Engineering and Data Storage Systems Center, Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A.

Thomas E. Karis
Hitachi Global Storage Technologies, San Jose Research Center, San Jose, California, U.S.A.

INTRODUCTION

Nano-films play a vital role in all aspects of human existence. This encompasses the spectrum extending from biological systems to small and large-scale electromechanical devices. Cell membranes encapsulate cells, while a film of tears lubricates eye movement. Grease lubricates the automobile wheel bearings by gradually releasing a thin film of oil into the ball-race interface. Due to the pervasiveness of nano-films in our life, many aspects of their physics and chemistry are being revealed through increasingly sophisticated methods of scientific investigation.

This entry focuses on the state-of-the-art in two related aspects of liquid nano-film physics, spreading and dewetting. The solid/thin liquid film system must be designed to either dewet and evaporate, as with water on automotive car finish, or to wet and spread as in well-designed lubrication systems. For example, when molecularly thin liquid films are employed to lubricate gears in a micro-electromechanical system, or in the head (slider) disk interface of a hard disk drive, certain properties are essential for them to function without an oil change for many years of continuous operation. These devices cannot be flooded with lubricant, or the small parts stick together. Lubricant must flow to where it is needed in asperity contacts by spreading, while at the same time the viscosity must be high enough to provide boundary lubrication and to avoid evaporation and spattering by dewetting or spin-off from the rapidly moving parts. This is especially true for open systems such as a hard disk drive. The lubricant spreads and the vapor pressure is suppressed by energetically favorable interaction with the surfaces, so there is little or even no evaporation at elevated temperature.

Here, we employ examples primarily from the magnetic recording industry to show how the physics of nano-films is derived from the properties of bulk liquids through the thermodynamics of interaction with the solid surface. The nano-film spreading rate is derived from the film thickness dependence of the surface energy. The nano-film viscosity is related to the nano-film vapor pressure, both of which increase due to dispersion force in the limit of molecularly thin film. Vapor pressure and evaporation are related to dewetting and capillary wave roughness through the surface energy for a typical perfluoropolyether (PFPE) lubricant Zdol which has a polar hydroxyl group on each end of the chain. Finally, the thermodynamic model is tied together by a description of the contemporary molecular dynamics (MD) and Monte Carlo (MC) simulation which graphically elucidates images of the chain conformation.

BACKGROUND AND MATERIALS

A fundamental understanding of the spreading behavior of the liquid film is required for the proper design of a number of engineering systems. Micro/nanoscopic spreading behavior is quite different from macroscopic behavior due to the difference in the driving forces between these two cases. Forces, such as surface tension gradient and gravity, drive spreading in macroscopic films.[1] However, a disjoining pressure gradient is the driving force for spreading in thin liquid films that have a thickness on the order of nanometer.[2] Although the spreading of liquid films on solid surfaces in the macroscopic regime has been studied extensively, the spreading phenomenon in thin films remains to be clearly understood and is the subject of this entry.

Nanoscale confined liquids are an important subject due to their ubiquitous nature and future industrial applications. The functionalities of polymer chain and solid surfaces, for example, are key control factors in determining the material design of lubricant used in nanotribology. Materials having constituents with dimensions on the nanometer scale behave remarkably different than when in bulk state, which has led to a new paradigm we now refer to as nanotechnology. Due to broad technological interests, numerous studies on nanoscale confined liquids have been investigated, both theoretically and experimentally, by scientists and engineers from a variety of backgrounds, including data storage, semiconductor devices, catalysis, polymer engineering and science, tribology, robotics, and medicine.[3]

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008083

The spreading behavior of small drops of polydimethylsiloxane (PDMS) on silica surfaces has been studied by many investigators including Cazabat et al.[4] Novotny[5] investigated the spreading of polyperfluoropropylene oxides on silica surfaces using scanning microellipsometry (SME) and scanning photoemission spectroscopy. The results revealed that the surface diffusion coefficient increases as the film thickness decreases down to 1 nm and remained constant below that thickness. The spreading characteristics of PFPEs have been investigated using SME equipped with temperature and humidity controls. They studied the effects of initial film thickness, end-group functionality, and molecular weight on silica surfaces[6,7] and on amorphous carbon surfaces.[8,9]

Molecularly thin PFPE films are mainly referred to in this entry which examines and illustrates the essence of the thin liquid film, especially the role of molecule-surface coupling. In particular, two examples of these materials are the commercially available Fomblin® PFPE Z and Zdol random copolymers with the linear backbone chain structure

$$X\text{—}[(OCF_2CF_2)_p\text{—}(OCF_2)_q]\text{—}O\text{—}X\ (p/q \cong 2/3)$$

where X (the ends-group) is CF_3 in PFPE Z, CF_2CH_2OH in Zdol (Fig. 1A). Notice that Zdol has a hydroxyl group at the end of the chain, which exhibits moderate interactions with solid surfaces.

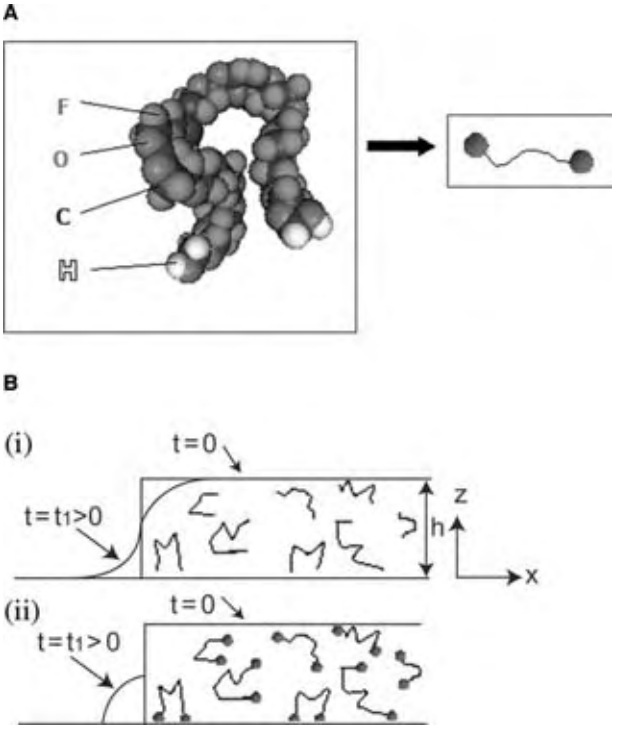


Fig. 1 (A) The molecular structure of PFPEs and its simplified model and (B) The spreading profiles as time progresses ($t = 0$ to t_1) from ''thought experiment'': (i) Z and (ii) Zdol. The coordinate system is represented in B(i). *(View this art in color at www.dekker.com.)*

EXPERIMENTATION AND MESOSCOPIC ANALYSIS

In this section, experiments on spreading properties using SME for various PFPE/solid surface pairs are discussed. The rheological characterization of PFPEs is also given. Further interrelationships among SME spreading profiles, rheology, surface energy or disjoining pressure, and dewetting, as well as a qualitative interpretation based on mesoscopic thermodynamics and transport phenomena, are provided.

Scanning Microellipsometry

To demonstrate the essence of the spreading phenomena, we invoke the ''thought experiments'' for different lubricant/surface and lubricant/lubricant interactions, which are shown in Fig. 1B.

In the experiments of O'Connor et al.,[7] monodisperse PFPE Z and Zdol, which were fractionated via supercritical fluid extraction in CO_2, were dip-coated onto the surface of silicon wafers as shown in Fig. 2A. Film thickness was controlled by altering the PFPE concentration and draw rate. An SME apparatus (Fig. 2B) was used to measure the thickness of the film as it spread with time. The coated wafer was placed on a pedestal-like plate housed in an environmental chamber with slits for passage of the incident and reflected beams. The chamber was mounted on a stage, which translated the sample across the beam area, and the thickness profiles were measured in the controlled temperature and humidity environment. The spreading profile obtained from SME strongly depended on the PFPE molecule–surface interactions, and provided a fingerprint for each pairing.

Typical SME thickness profiles for monodisperse Z and Zdol, as examined by O'Connor et al.,[6,7] are shown in Figs. 3A and B. As the film spreads with time, the spreading front travels along the surface of the silicon wafers. The sharp ''spike'' for the PFPE Z profile in Figs. 3A was observed to decay over time. The ''spike'' at the step did not affect the spreading rate, and could be avoided by eliminating free surface vibrations and decreasing the solvent evaporation rate by dip-coating with the solution in the bottom of a tall and thin container. The spreading of PFPE Zdol exhibits a characteristic layering structure or shoulder with a height on the order of the radius of gyration for the PFPE molecules and the separation of PFPE molecules from the initial film layer at a sharp boundary. Thicker films of Zdol appear

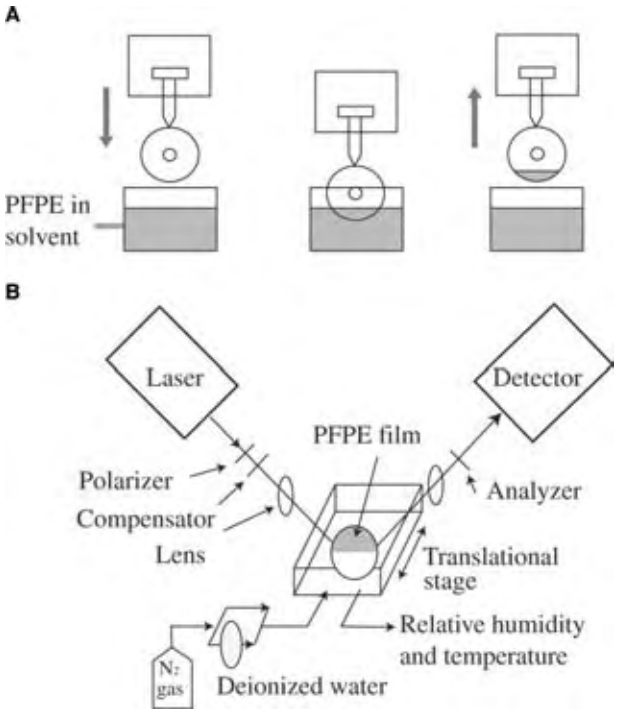


Fig. 2 (A) Partially dip-coating a disk in a PFPE solution and (B) schematic of the SME apparatus. The PFPE film is shaded. *(View this art in color at www.dekker.com.)*

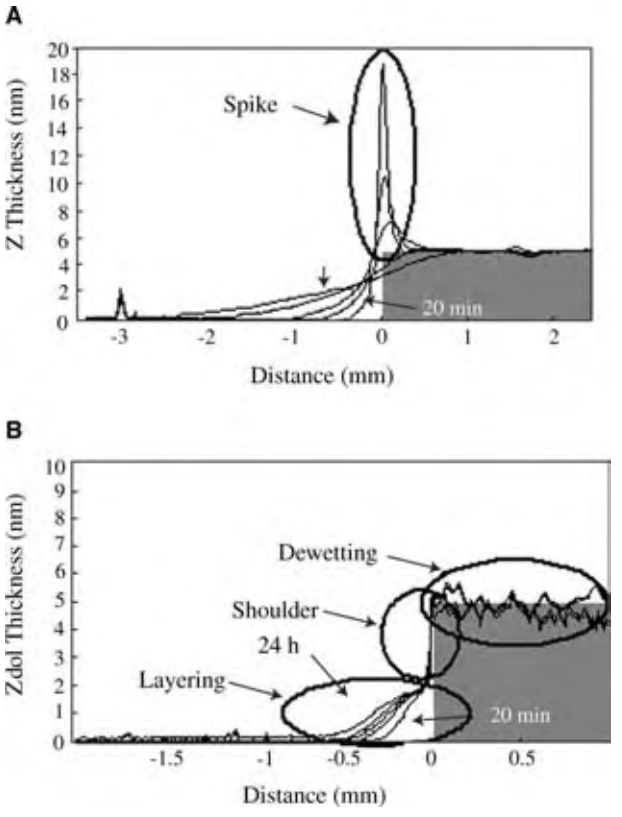


Fig. 3 SME spreading profiles for PFPE on silicon wafers: (A) Z (M_w = 13,800 g/mol) and (B) Zdol (M_w = 3100 g/mol). Both have the initial thickness of 5 nm for times of 20 min, 1.5, 3, 12, and 24 h at 26°C.

to experience partial "dewetting" as indicated by the rough appearance in the SME scan (Fig. 3B). Dewetting phenomena (investigated for mogul dynamics[10] in magnetic recording disk lubrication) are important for the nanoscale, multiphase system design and will be discussed in the section entitled Capillary Waves and Dewetting. Similar PFPE spreading experiments on carbon coated magnetic recording disks were performed by Ma et al.,[8,9] a few years after the observations made by O'Connor et al. In their measurement, however, mass build-up for PFPE Z was not observed, and dewetting characteristics for Zdol were somewhat suppressed. These phenomena clearly demonstrated the relevance of surface energy driven flow effects in nanoscale spreading processes, and will play a critical role in design criteria for nanodevices and/or sensors. Ma et al.[11] examined spreading characteristics for various carbon surfaces (α-C/hydrogenated/nitrogenated), illustrating the fingerprint for the various surface/PFPE molecular coupling.

Mesoscopic Thermodynamic and Transport Models

Disjoining pressure

The remarkable similarity of the PDMS and PFPE spreading profiles suggests that interactions between the polar entities of the liquid and the surface are at the root of the "anomalous" spreading in these liquids. In the case of PFPE terminated with functional end-groups, evidence for strong interactions between the end-groups and the polar active sites on the carbon surface has been reported.

The free energy function provides a mesoscopic framework for understanding the terraced spreading phenomena. Since the surface energy is defined as the free energy per unit area, the total disjoining pressure (Π) for these fluids can be derived from the experimental surface energy data by:

$$\Pi = -\frac{\partial}{\partial h}\left(\gamma^d + \gamma^p\right) \quad (1)$$

where, γ^d and γ^p are the dispersive and polar components of the surface energy, respectively, and h is the film thickness. The regression fit to the surface energy data, shown as the smooth curves in Fig. 4A and B, were numerically differentiated to obtain the disjoining pressure. The total disjoining pressure, as well as the individual contributions from the dispersive and polar components, is shown in Fig. 4C. Notice that

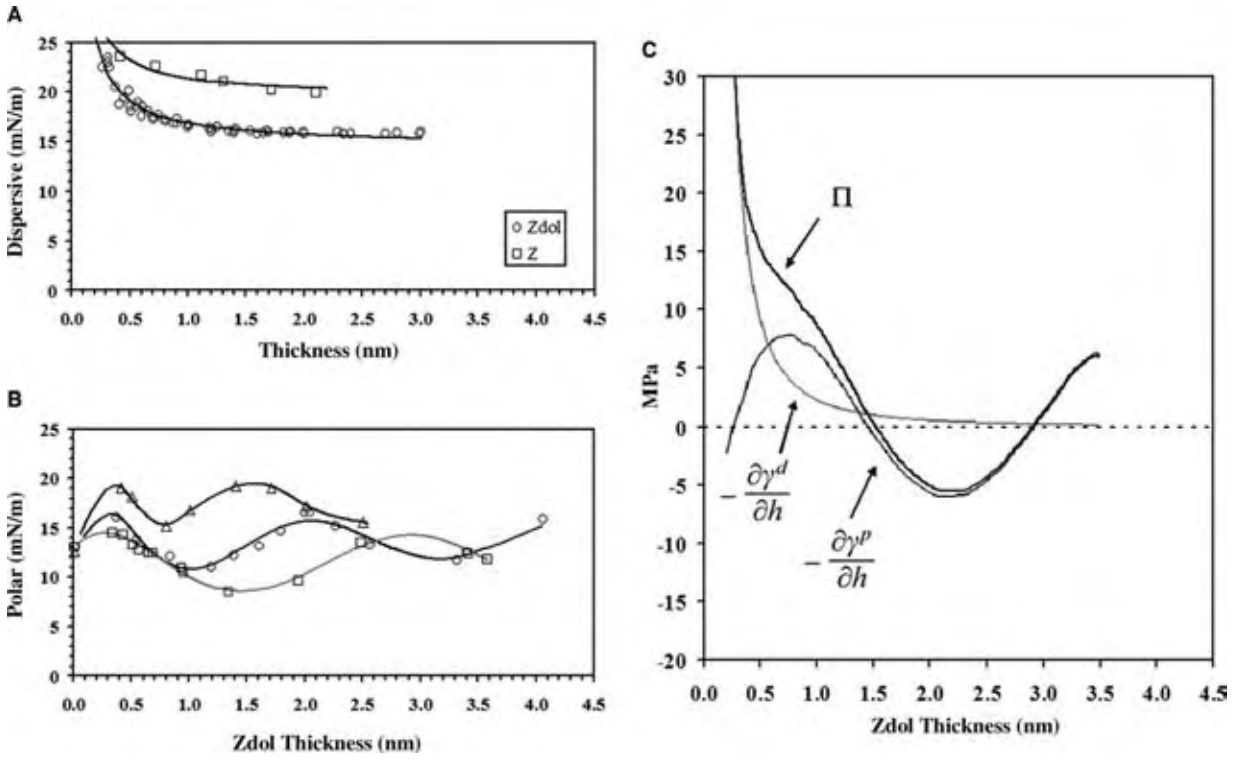


Fig. 4 The components of the surface energy measured on hydrogenated carbon overcoated thin film magnetic recording media: (A) The dispersive component of the surface energy for PFPE Z and Zdol; (B) the polar component of the surface energy for PFPE Zdol with molecular weight (M_W) of 1100 (△) 1600 (○), and 3100 (□) g/mol; and (C) the disjoining pressure as a function of film thickness for PFPE Zdol (M_W is 3100 g/mol).

γ^d decreases monotonically with h. Below film thickness of approximately 0.5 nm, Π at each molecular weight is dominated by γ^d, which increases rapidly with decreasing film thickness and is largely independent of molecular weight. The γ^p, however, oscillates with film thickness and becomes larger in magnitude than γ^d as h increases. Oscillations in γ^p provide an additional contribution to Π for PFPE Zdol that produces alternating regions of stable and unstable film thickness. The sum of the two contributions gives rise to oscillations in the total disjoining pressure. The shoulder height in the Zdol spreading profile corresponds to the film thickness at which $\partial\Pi/\partial h < 0$ and Π changes sign from positive to negative. The shoulder heights for two of the Zdol fractions were close to those expected from analysis of the disjoining pressure isotherm.[12]

Spreading characteristics from the mesoscopic transport model

To quantify the spreading characteristics, we examined the thickness-dependent diffusion coefficient $D(h)$, extracted from the SME via two methods. The first method utilizes the spreading data at constant height (isoheight), while the second method utilizes the entire spreading profile. The first method does not assume that the spreading is a diffusive process, whereas the second method does. The length L, which the leading edge of the advancing lubricant front traveled after time t, is defined as the difference between the initial position of the leading edge and its position in each of the subsequent profiles. To quantify the spreading rate, L and the corresponding times t (obtained from the profiles in Fig. 3B) are plotted. As shown in Fig. 5A, the L–t plot (in log) can be fit with the two piecewise straight lines for the entire range of t, i.e., $L \propto t^{\alpha}$. We found that $\alpha \cong 1$ at short times (Regime I), and after a gradual transition, $\alpha \cong 1/2$ (Regime II; diffusion region). Therefore, the surface diffusion coefficient D is estimated from the data in Regime II ($L \propto t^{1/2}$) alone via the relationship $D = L^2/t$.

We adopted a second method[9] called the Boltzmann–Matano interface method [see Eq. (5)], which is more accurate for the purely diffusive process. Notice that Ma et al.[8,9] successfully employed this method, since the spreading profiles they measured exhibit little mass buildup at the front and negligible dewetting.

The shape of spreading profiles can be derived from the diffusion coefficient D as follows: Since the flux vector $\mathbf{j}$ is proportional to the concentration (or related

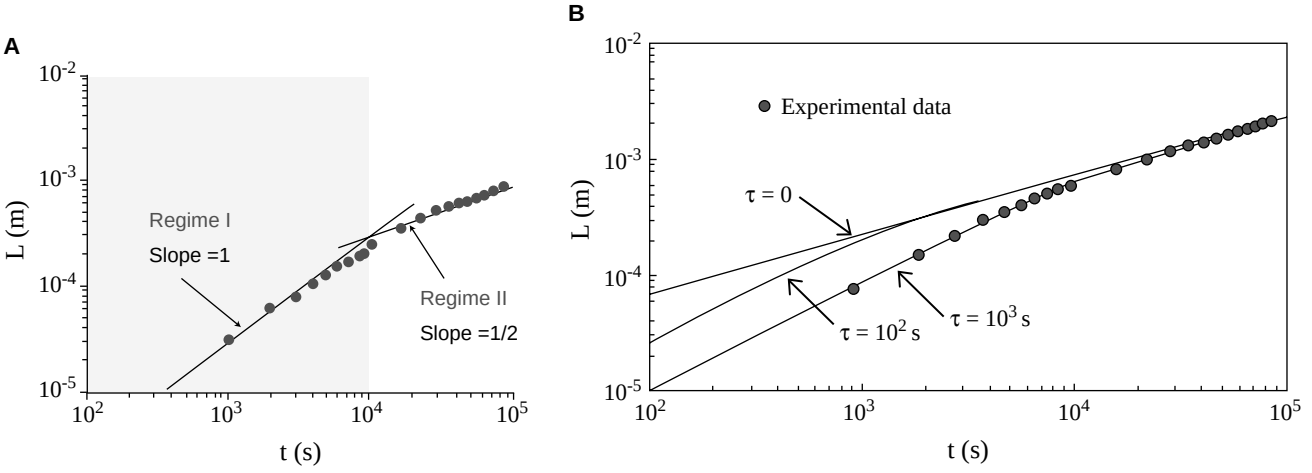


Fig. 5 (A) Travel length (L) as a function of time (t) obtained from Fig. 3B for monodisperse Zdol; $T = 26°C$, under dry nitrogen and (B) L–t plot results using the modified diffusion equation (Eq. 7) for several values of τ. *(View this art in color at www.dekker.com.)*

to film thickness, h) gradient ∇h we obtain,

$$\mathbf{j}(\mathbf{r},t) = -D(h)\nabla h(\mathbf{r},t) \tag{2}$$

Further, the mass conservation (or continuity) equation gives:

$$\frac{\partial h(\mathbf{r},t)}{\partial t} + \nabla \bullet \mathbf{j}(\mathbf{r},t) = 0 \tag{3}$$

By combining Eqs. (2) and (3), we obtain

$$\frac{\partial h(x,t)}{\partial t} = \nabla \bullet [D(h)\nabla h(x,t)] \tag{4}$$

where ∇ is the ''del'' operator and $h(x,t)$ is the film thickness.

On the other hand, the diffusion coefficient can be evaluated from Eq. (4). This 1-D analysis is called the Boltzmann–Matano technique:[9]

$$D(h) = -\frac{1}{2t}\left(\frac{\mathrm{d}x}{\mathrm{d}z}\right)_{z=h}\int_0^h x\,\mathrm{d}z,$$
$$\text{with the constraint } \int_0^{h_0} x\,\mathrm{d}z = 0 \tag{5}$$

where h_0 is the initial film thickness far from the step. Note that h and the coordinates x and z are shown in Fig. 1B(i). However, as shown from the L–t plot, the spreading cannot be described by the diffusive concept alone.

Since the theories developed by Novotny[5] [Eq. (4)] and Mate[13] (hydrodynamic model) fail to describe the $L \propto t$ behavior exhibited in Regime I, a simple mesoscopic model for the overall L–t behavior is described below. (The more rigorous molecular simulation is described later.) By introducing a time lag denoted as τ (which may be related to the relaxation time of the PFPEs), the Fick's constitutive equation can be modified as:[14]

$$\mathbf{j}(\mathbf{r},t+\tau) = e^{\tau\frac{\partial}{\partial t}}\mathbf{j}(\mathbf{r},t) = -D(h)\nabla h(\mathbf{r},t) \tag{6}$$

To derive Eq. (6), we imposed causality into the relationship between $\mathbf{j}$ and ∇h inspired by the microscale heat transfer theory or Cattaneo equation.[15]

By combining Eqs. (3) and (6), and assuming $\tau\partial/\partial t$ to be small, we obtained the following modified diffusion equation:

$$e^{\tau\frac{\partial}{\partial t}}\frac{\partial h}{\partial t} \cong \tau\frac{\partial^2 h(r,t)}{\partial t^2} + \frac{\partial h(r,t)}{\partial t} = \nabla \bullet [D(h)\nabla h(r,t)], \tag{7}$$

where τ characterizes the crossover behavior between Regimes I and II. Notice that Eq. (7) reduces to Eq. (4) when $\tau = 0$. We found that the transition between these two regimes depends on the value of τ. Increasing τ shifts the transition to a later time. By setting $\tau = 10^3$ s in Eq. (7), we attained an excellent fit for the experimental L–t data (Fig. 5B).

Relationship between diffusion coefficient and disjoining pressure

As will be shown later, the measured disjoining pressure Π can be used to qualitatively describe the spreading profile. Before that, the relationship between Π and D is established. Note that Π and D can also be calculated by MD simulation

The relationship between position dependent viscosity $\eta(z)$, Π, and $D(h)$ is developed by generalizing the hydrodynamic model (Reynolds equation):

$$D(h) = \int_0^h \frac{(h-z)^2}{\eta(z)}\left(\frac{\partial \Pi}{\partial h}\right) dz \tag{8}$$

or by substituting constant viscosity assumption, i.e., $\eta(z) = \eta_\infty$ in Eq. (8), we obtain Eq. (9) developed by Mate:[13]

$$D(h) = -\frac{h^3}{3\eta_\infty}\frac{\partial \Pi}{\partial h} \tag{9}$$

In deriving Eq. (9), we used the no-slip boundary condition at the lubricant/solid boundary. If we were to generalize the above analysis with the partial-slip boundary condition, e.g., $\partial v/\partial z = \beta v$ ($\beta \equiv$ slip parameter) instead of no-slip condition with $\eta(z) = \eta_\infty$, Eq. (9) could be modified as:

$$D(h) = -\frac{1}{\eta_\infty}\left(\frac{h^2}{\beta} + \frac{h^3}{3}\right)\frac{\partial \Pi}{\partial h} \tag{10}$$

We could use a simplified form of $\eta(z) = \eta_\infty f(z)$. Here η_∞ is the bulk viscosity, and $f(z)$ can be experimentally determined. Here, z is the distance normal to the solid surface. A partial justification for the above functional form can be drawn from the temperature dependence of the surface diffusion coefficient and the bulk viscosity,[7] or the fly stiction correlation with the bulk viscosity. To develop a rigorous hydrodynamic model, we need better rheological data and more details are given in the following section on Rheology Measurement.

An alternative approach to describe thin film spreading is the jump diffusion model developed by Karis and Tyndall.[16] In their analysis, the flow rate is calculated by integrating the velocity $\mathbf{v}$ throughout the depth of the film

$$\mathbf{j} = \int_0^h \mathbf{v}\, dz \tag{11}$$

The velocity ($\mathbf{v}$) is taken to be the drift velocity of the fluid particles in a disjoining pressure gradient, $\nabla\Pi$. The velocity is therefore proportional to the disjoining pressure gradient with the proportionality constant m representing the mobility [the units of m are (m/sec)/(Pa/m)], and

$$\mathbf{v} = m\nabla\Pi \tag{12}$$

The flow rate is calculated from the velocity by integrating through the film thickness according to Eq. (11)

$$\mathbf{j} = mh\nabla\Pi = mh\frac{\partial \Pi}{\partial h}\nabla h \tag{13}$$

$$\text{or} \quad D(h) = mh\frac{\partial \Pi}{\partial h} \text{ (in 1-D form)} \tag{13}'$$

In the integration of Eq. (11) to derive Eq. (13), the mobility was assumed to be independent of the distance from the surface between $z = 0$ and $z = h$, and therefore should be considered an ''effective'' mobility. In general, m is a function of end group chemisorption, and m could be rigorously calculated with the MD simulation. The differential equation that defines the 1–D spreading profile is obtained via substitution of Eq. (13)′ into the continuity Eq. (4).

$$\frac{\partial h}{\partial t} = \frac{\partial}{\partial x}\left(D(h)\frac{\partial h}{\partial x}\right) = \frac{\partial}{\partial x}\left[\left(mh\frac{\partial \Pi}{\partial h}\right)\frac{\partial h}{\partial x}\right] \tag{14}$$

Eq. (14) is solved numerically to calculate spreading profiles of nonpolar and polar PFPEs from the disjoining pressure measured from contact angles.[16] The calculated profiles, shown in Fig. 6A for Zdol, are in quantitative agreement with those measured by SME, Fig. 6B. The only adjustable parameter was $m\gamma_\infty/d_0$, where γ_∞ is the bulk surface tension of the lubricant, and d_0 is a characteristic dimension of the PFPE chain. Here $d_0 \approx 0.4$ nm, which is approximately the van der Waals diameter of the PFPE chain. For Zdol, $\gamma_\infty \approx 24$ mN/m.

A previously undescribed region for thin film spreading, which is particularly relevant to dewetting, is the region of film thickness h where $\Pi > 0$ and $\partial\Pi/\partial h > 0$, is predicted from this analysis. This region does not always show up in the spatially averaged spreading profiles measured by SME. It is referred to as the reverse flow region, where the free energy is lowered by decreasing the local film thickness. Thus, the flow in this region tends to decrease rather than increase the film thickness. A similar effect is observed in spinodal decomposition and is referred to as ''up-hill'' diffusion.[17]

Rheology Measurement

The bulk rheological properties of the PFPEs, including the melt viscosity, storage modulus, and loss modulus at several different temperatures, have been widely reported via steady shear and dynamic oscillation tests.[18] In this entry, the focus is on the confined geometry effects on viscosity.

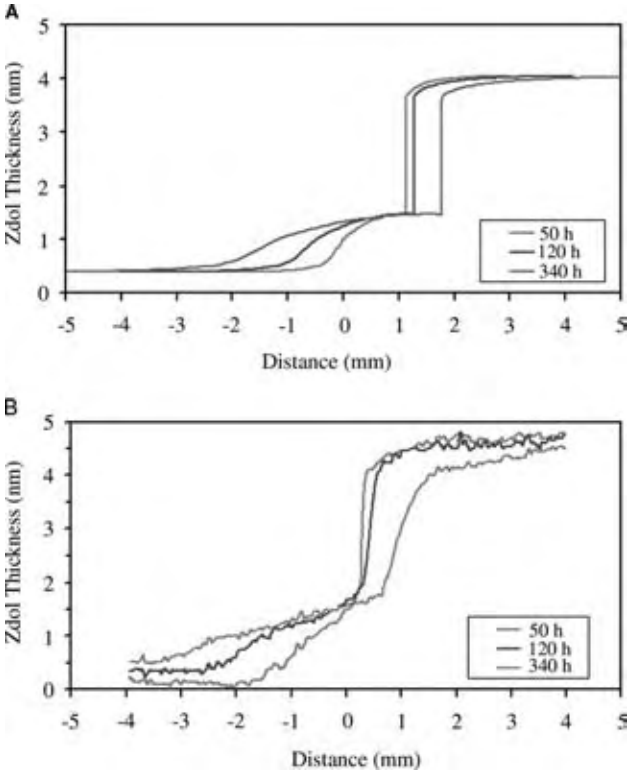


Fig. 6 (A) Calculated and (B) measured spreading profiles for Zdol. M_W is 3100 g/mol. *(View this art in color at www.dekker.com.)*

Dispersive interaction has a dramatic effect on the molecular layers closest to the surface, and can be explained in terms of the rate theory for viscous flow.[19] Within the rate theory, a flow event comprises the transition of a flow unit from its normal or quiescent state, through a flow-activated state, to a region of lower free energy in an external stress field. For small molecules, the flow unit is the whole molecule, while for longer chains, the flow unit is a segment of the entire molecule. By analogy with chemical reaction rate theory, there is a flow-activation enthalpy, ΔH_{vis}, and entropy, ΔS_{vis}, for transition into the flow-activated state.

A flow unit is approximated by a particle in a box, with the energy being partitioned among rotational and translational degrees of freedom, which govern the transition probability. On this basis, the viscosity $\eta = (Nh_p/V_l)\exp(\Delta G_{vis}/RT)$, where N is Avogadro's number, h_p is the Planck constant, V_l is the molar volume of the liquid, and $\Delta G_{vis} = \Delta H_{vis} - T\Delta S_{vis}$ is the flow-activation free energy. Here, $\Delta H_{vis} = \Delta E_{vis} + \Delta(pV)_{vis}$, where ΔE_{vis} is the flow-activation energy and $\Delta(pV)_{vis}$ is the pressure–volume work. At constant pressure, $\Delta(pV) = p\Delta V_{vis}$. For PFPE Z, the flow-activation volume $\Delta V_{vis} \approx 0.1\,\text{nm}^3$,[20] which is equivalent to a spherical region ≈ 0.6 nm in diameter. At ambient pressure (100 kPa), $\Delta(pV)_{vis} \approx 6.2$ J/mol, near ambient conditions, $\Delta H_{vis} \approx \Delta E_{vis}$. Therefore, the viscosity is given by:

$$\eta = \left(\frac{Nh_p}{V_l}\right)\exp\left[\frac{(\Delta E_{vis} - T\Delta S_{vis})}{RT}\right] \tag{15}$$

A regression fit to the bulk viscosity as a function of temperature,[19] provided $\Delta E_{vis} = 34.7$ kJ/mol) and $\Delta S_{vis} = 9.87$ J/mol-K. The flow-activation energy is close to that reported for bulk Zdol with a molecular weight of 3100 in Refs.[6,7]. A positive value for the flow activation entropy of bulk Zdol means that the entropy of the flow unit increases on going into the flow-activated state.

Changes in the lubricant flow-activation energy and entropy near the solid surface cause changes in the viscosity with decreasing film thickness. The flow-activation energy near a solid surface is estimated from the thin film vaporization energy as follows. In an ideal gas, the chemical potential μ (or partial molar Gibbs free energy) is given by:

$$d\mu = RTd\ln P \tag{16}$$

where P is the partial pressure of the lubricant in the vapor phase and R is the gas constant. The chemical potential per unit volume in the lubricant film $\mu/V_l = \Pi$. The ratio of the film surface vapor pressure, $P^0(h)$ to the vapor pressure of the bulk lubricant, $P^0(h)/P^0(\infty)$, is derived by integrating Eq. (16):

$$\mu(h) - \mu(\infty) = RTln\left[P^0(h)/P^0(\infty)\right] \tag{17}$$

The reference state is taken to be the chemical potential and vapor pressure of the bulk lubricant: $u(\infty) = 0$ and $P^0(\infty)$ is the vapor pressure of the bulk liquid. The surface chemical potential is approximated by the unretarded atom-slab dispersive interaction energy:

$$\mu = -\frac{V_l A}{6\pi h^3} \tag{18}$$

The dispersive interaction coefficient A is also referred to as the Hamaker constant, and $A = 10^{-19}$ J for Zdol.

The dispersive component of the vaporization energy near a solid surface is approximately given by Eq. (18). The vaporization energy is the energy required to remove a molecule from the liquid without leaving a hole behind. The free volume needed for a flow unit to make a transition into the flow-activated state is less than the size of the entire molecule. It is found that the ratio $n \equiv \Delta E_{vap,\infty}/\Delta E_{vis,\infty} > 3$, where $\Delta E_{vap,\infty}$ and $\Delta E_{vis,\infty}$ are the vaporization- and

flow-activation energy of the bulk liquid, respectively. Thus, the flow-activation energy near the surface is approximated by:

$$\Delta E_{vis} = \Delta E_{vis,\infty} - \mu/n \quad (19)$$

For linear chains longer than 5 or 10 carbon atoms, n increases due to the onset of segmental flow. In practice, n is experimentally determined from the measured values of ΔE_{vap} and ΔE_{vis}. For PFPE Zdol 4000, $\Delta E_{vap,\infty} = 166\,\text{kJ/mol}$, giving $n \approx 4.8$, which is consistent with segmental flow.

In order to calculate the thin film viscosity with Eq. (15), the flow-activation entropy near the surface is also needed. An experimental flow-activation entropy is calculated from the spin-off data[19] with Eqs. (15) and (19) as follows. The experimental η vs. h is determined from the dh/dt during air shear induced flow on a rotating disk. Eq. (19) is then solved for ΔS_{vis} versus h using Eq. (19) for ΔE_{vis}. Below 2.3 nm, $T\Delta S_{vis} \approx -1.9\,\text{kJ/mol}$, which corresponds to the critical configurational entropy change for flow ($-R\ln 2 \approx -5.76\,\text{J/mol-K}$).

Capillary Waves and Dewetting

Liquids are permeated by thermal fluctuations at finite temperatures. In quiescent bulk liquids, thermal fluctuations give rise to density fluctuations and diffusion, and these enable fluids to spread and flow. Molecular drift and shear flow, as discussed above, are present when an energy gradient is superimposed on the thermal fluctuations. According to the rate theory of flow, the likelihood that a thermal fluctuation will lead to a displacement event is determined by the ratio of the free energy change of the event taking place to the thermal energy. The same analogy holds for vaporization. Thermal fluctuations are present at the liquid vapor interface, as well as in the bulk, and governed by the same thermodynamic principles as diffusion and flow.

However, the present discussion focuses on molecularly structured liquids with at least 20 monomer units. Thermal fluctuations leading to cooperative motion in polymers pertain to segments consisting of several monomer units. A measure of the degree of segmental flow is provided by the ratio of the vaporization energy to the flow-activation energy, which is approximately four to five for Zdol 4000. If the entire molecule is the flow unit, then this ratio is closer to three. Therefore, with typical PFPE lubricant molecules, the size of the cooperative groups participating in thermal fluctuations is less than the entire chain.

Thermal fluctuations at the free liquid surface of molecularly thin hexane films on solid surfaces have been found to produce a fluctuating surface roughness, which is detected by X-ray reflectivity and diffuse scattering. For hexane, the whole molecule is the flow unit. Thus, for hexane, density fluctuations near the surface involve molecular rearrangement. In contrast, for PFPE, density fluctuations near the surface involve segmental rearrangements. Similar X-ray reflectivity and diffuse scattering measurements of PFPEs on carbon overcoats and silica have been reported by Toney, Mate, and Leach.[21] The surface smoothens as the lubricant thickness on carbon is increased, which probably reflects the lubricant filling in porosity of the overcoat.[22] Surface density fluctuations observed with fully bonded end-groups show that these fluctuations are segmental rather than cooperative motions of the entire chain.

Surface roughness induced by surface density fluctuations as a function of film thickness is estimated for the PFPE Zdol 3100 on amorphous carbon for which the spreading profile is shown in Fig. 3. The rms-roughness is given by the capillary wave continuum theory[23] as:

$$\sigma_{rms} = \sqrt{\frac{k_B T}{4\pi\gamma_\infty} \ln\left[1 + \frac{2\pi^3\gamma_\infty h^4}{a^2 A}\right]} \quad (20)$$

where a is a molecular length scale and k_B is the Boltzmann constant. The length scale $a \approx 0.75\,\text{nm}$ is roughly estimated from 60% of the maximum stable film thickness in the thin film region of Fig. 6. The peak to valley roughness, denoted $\sigma_{pp} = (2\sqrt{2})\sigma_{rms}$, is more relevant to magnetic recording slider disk spacing considerations. Fig. 7 shows the roughness due to thermal fluctuations at the free surface. The stable and unstable regions for PFPE Zdol with molecular weight of 3100 are superimposed in the roughness curve. The dashed region of the roughness curve denotes the unstable region, which is the step height in Fig. 6, and the region between the zero crossing and the first maximum of Π in Fig. 4C.

Separation into two stable regions of film thickness is not always observed when the mean film thickness is within the unstable region, even though the spreading film will not spontaneously flow to form a film with thickness in the unstable region. Films are routinely dip-coated and remain at the metastable thickness within the unstable region during measurement of the contact angle to determine the surface energy shown in Fig. 4A and B. No transition was observed in the viscosity during spin-off as the film thickness made a transition through the unstable region. As noted by Toney, Mate, and Leach,[21] thermal fluctuations of the free surface, at least at room temperature, do not appear to initiate the phase transition into two stable

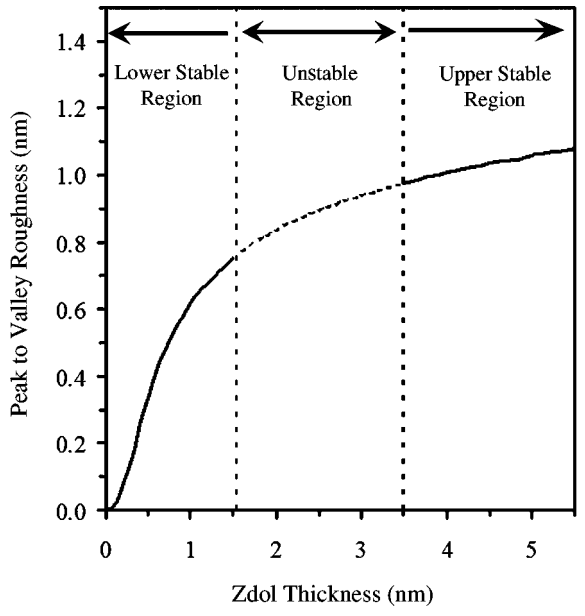


Fig. 7 Calculated peak to valley surface roughness due to thermal fluctuations at the liquid vapor interface of Zdol ($M_W = 3100\,g/mol$) as a function of the average film thickness.

regions of film thickness (i.e., dewetting) from the metastable state.

This is because the fluctuations represent only a small perturbation to the overall molecular configuration of the PFPE chain. For PFPE Zdol with a molecular weight of 3100, the number average degree of polymerization is approximately 31, and the contour length is 12 nm. By analogy with the number of atoms in a flow unit of linear alkanes, the fluctuation comprises two to four chain monomers, or 6–12% of the PFPE chain contour length. Much larger cooperative fluctuations, or external perturbations, are needed to nucleate the spinodal decomposition. However, thermal fluctuations in the lower stable region of film thickness may play a role in the critical flying height for the onset of bridge formation.[24]

SIMULATION

The immediate goal of the simulation is to construct reliable tools that accurately describe the static and dynamic behavior of thin liquid films. As should be apparent from the preceding sections, it is extremely difficult to measure the flow-activation energy, disjoining pressure, and dewetting due to thermal fluctuations accurately or directly. The molecular simulation provides a complementary tool and is described in detail in this section.

A multitude of references deal with the general methodology behind molecular simulations.[25] In this entry, the simple lattice-based, simple reactive sphere (SRS) MC techniques for examining the fundamentals of PFPE film structure are introduced first. An off lattice-based, bead-spring MC model is introduced later to capture the deailed internal structure of the PFPE molecules, and the molecular dynamics method is implemented for a full-scale nanostructural dynamic analysis of thin PFPE liquid films.

SRS Model

We adopted a spin analogy/lattice gas model, or SRS model, as shown in Fig. 8A, which illustrates the oversimplified molecular structure, while still capturing the essence of the molecule/surface interactions describing SME profiles. Notice that similar techniques using the Ising model have been previously investigated to study other physical systems.

MC simulations based on the SRS model were pioneered by Ma et al.[26] to explain the peculiar spreading profiles of PFPEs on amorphous carbon surfaces shown in Fig. 3 via the adoption of four different interactions: molecule/molecule, molecule/surface, end-group/end-group, and end-group/surface (a molecule is denoted as a backbone in the absence of polar end-groups). Molecules are approximated as reactive spheres, where a spin $S = 1$ is assigned to an occupied lattice site (for a vacant site, $S = 0$).

A comparison between MC simulation results for molecules with non-polar (e.g., PFPE Z) and polar (e.g., PFPE Zdol) end-groups is illustrated in Fig. 1. A drastic difference is apparent between Fig. 1B(i) and B(ii): the PFPE Z profile is relatively smooth and spreads more rapidly, while the Zdol profile exhibits a complex layering structure. These results are qualitatively consistent with the experimental observations given in Fig. 3, which reveal that molecules with polar end-groups demonstrate a first layer that is one molecule thick (the thickness of the first layer is on the order of the PFPE diameter of gyration in the bulk state), with subsequent layers approximately twice the thickness of the first layer. This simulation qualitatively explained the experimental spreading data on hydrogenated and nitrogenated carbon surfaces[11] by adjusting the screening length (descriptive for end-group/end-group coupling), which is related to the hydrogen and nitrogen content. Using the SRS model, the L–t plot was constructed by plotting the distance traveled from the initial sharp boundary versus the number of MC steps (the iso-height in the spreading profile). The simulated L–t response shows a distinct transition between short and long times, which agrees with the experimentally observed L–t behavior shown in Fig. 5. The long-term behavior exhibits $L \propto t^{1/2}$, thus meeting the criteria for the surface diffusion assumption.[27]

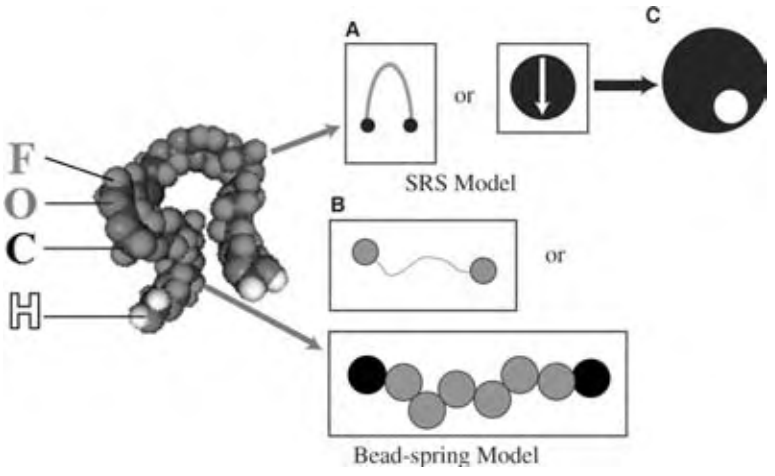


Fig. 8 Molecular model of PFPE Zdol: (A) SRS model with a discrete spin state; (B) bead-spring model, where gray spheres represent the backbone and black spheres represent the polar end-groups; and (C) SRS model with a continuous spin state (small end group sphere with the large backbone sphere). *(View this art in color at www.dekker.com.)*

The 3D results for PFPE Zdol, Fig. 9A provides more insight into the layer coupling. Due to the simplicity in this SRS model with lattice sites, an occupancy spin, and an end-group orientation spin (toward or away from the surface), we can qualitatively investigate the 3D steady state surface structure of PFPE films, describe its dewetting phenomenon and complex surface morphology, and examine its properties as a function of interaction parameters, surface loading, and initial surface structure. To illustrate the dewetting phenomenon as a function of end-group/end-group and end-group/surface interaction energies, simulations were conducted on a 3D lattice.[28] For an initial condition, we randomly distributed a layer with 90, 100, and 110% loading onto the surface to represent a very rough, highly damaged layer as seen for the 100% initial loading condition in Figs. 9B and C. If the end-group/end-group interactions are too strong, the layer does not relax, and the amount that a layer will relax is determined by the end-group interaction strength. The fractal dimension along with the simulated surface morphology described above can have a potential use for fingerprint analyses of PFPE molecule/surface pairs.

To better examine the disjoining pressure driving force behind the nanoscale spreading process, it is more advantageous to adopt a continuous spin state SRS model for an off-lattice scenario. The end-groups in this case are

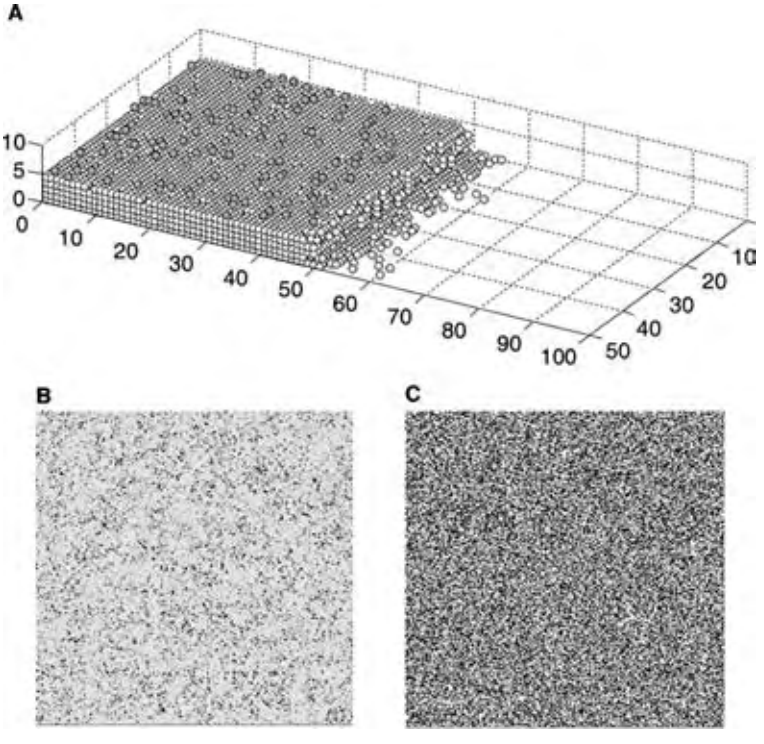


Fig. 9 (A) A typical simulation results for spreading 3-D dewetting simulation with (B) initial condition for 100% layer loading; and (C) relaxed film with a rough, dewetted surface. Black indicates a bare surface and white indicates peaks.

represented as smaller, internal spheres on the molecular backbone sphere surface, as shown in Fig. 8C.

Monte Carlo Simulation with Bead-Spring Model

To represent the molecular structure with reasonable accuracy, as well as to reduce computational time, the coarse-grained, bead-spring model (Fig. 8B) was employed to model a PFPE molecule. This simplifies the detailed atomistic information while preserving the essence of the internal molecular structure. The off-lattice MC technique was used to examine nanoscale PFPE lubricant film structures and stability with internal degrees of freedom.[29]

In this model, a PFPE molecule is composed of a finite number of beads with different physical or chemical properties. For simplicity, we assume that all the beads, including the end-beads, have the same radius. Lennard-Jones and van der Waals potentials were used for nonpolar bead–bead and bead–wall interactions, respectively. For polar interactions, exponential potential functions were added to both end-bead end-bead and end-bead wall interactions. For the bonding potential between adjacent beads in the chain, a finitely extensible nonlinear elastic model was used. For example, PFPE Zdol can be characterized differently from PFPE Z by assigning the end-bead a polarity originating from the hydroxyl group in the chain end.

The steady-state nanoscopic properties were examined and the result agrees qualitatively with the simulation results obtained via the SRS model. Our results provide a direct interpretation of the experimental surface energy data for PFPE films with functional end-groups.[12] The most remarkable achievement of MC simulation is its ability to predict equilibrium properties, including the surface morphology visualization. For example, in our simulation, stable films did not experience dewetting or film rupture. However, a rougher surface morphology was observed for smaller molecular weights (Fig. 10B) and strong end-bead functionalities (Fig. 10C).

To establish a qualitative relationship between the orientation and the layer structure, we examined the number of layers and the end-bead density. We found that the adsorption of functional end-beads results in an alternate ordering in the subsequent layers, i.e., upward orientation in the second layer, downward orientation in the third, etc. Our result provides a direct interpretation of experimental surface energy data for PFPE films with functional end-groups. The nondispersive component of surface energy exhibited an oscillatory pattern with increasing film thickness and was shown to be approximately proportional to end-group density, as demonstrated in our study.[29] Our simulation results also suggest that the density variation of the end-groups is related to the characteristic behavior of the surface energy of PFPE films.

Not only we can interpret experimental data, but we can also predict the nanostructure which is unobtainable experimentally. For example, we found that the observed expansion of layer thickness is attributed not to the bond stretching, but to the temperature dependence of the intermolecular interaction and excluded volume effect.

Molecular Dynamics Simulation

So far, we have demonstrated that the MC simulation (lattice-based SRS model and off-lattice bead-spring model) results are in qualitative agreement with the experimental results. A complementary approach is MD simulation using the bead-spring model.

Our preliminary MD simulations provided similar results as the MC method for the calculation of the static properties of confined PFPE nanofilms, especially the radius of gyration and end-bead density profiles. The anisotropic molecular conformation and experimental layering structures in the film were also verified. MD simulations[30] provide a powerful tool for examining the dynamics of nanofilms through correlation functions by tracking the trajectories of molecules, including the space and velocity coordinates. MD simulations, therefore, are suitable for

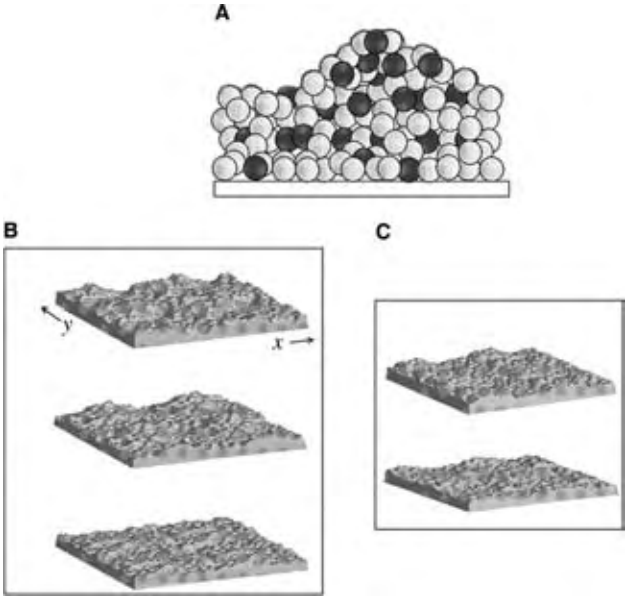


Fig. 10 Morphology of PFPE films from MC simulations: (A) schematic of simulated surface roughness for; (B) molecular weight dependence [$N_p = 10$ (upper), 15 (middle), 20 (bottom)]; and (C) end-bead functionality dependence [$\varepsilon_w^p = \varepsilon^p = 1$ (upper) and $\varepsilon_w^p = \varepsilon^p = 4$ (bottom)]. Here, N_p is the number of monomers and ε_w^p and ε^p are polar energy parameters.

calculating the transport coefficients from the correlation functions. In our preliminary study, self-diffusion coefficients of PFPE nanofilms were investigated with the Einstein relationship or Green–Kubo formula, to examine the effects of molecular weight, end-group functionality, external surface interaction strength, polydispersity, and temperature. Several autocorrelation functions were examined to study molecular motions. The Kohlrausch–Williams–Watts function and wavelet analysis were employed to interpret the molecular structure of PFPE nanofilms. Insight into physics of the disjoining pressure was revealed by calculating the internal pressure tensor during the simulation of nanofilms.

CONCLUSIONS

We reviewed fundamental scientific tools as well as potential applications relevant to the thin liquid film technology. We focused on understanding the behavior of molecularly-thin lubricant films relevant to the emerging field of nanotechnology, especially for achieving durability and reliability in nanoscale devices. The topics on the experimentation and theory for the physicochemical properties of ultra-thin PFPE films were reviewed by examining liquid film thickness from submonolayer to multilayer. By systematically tuning the end-group strength for PFPE, we can examine and control physicochemical properties for thin liquid films in various PFPE/solid surface combinations. Methods for extracting spreading properties from the SME and rheological properties (we only discussed surface-enhanced viscosity) of PFPEs are shown. The interrelationships among SME spreading profiles, surface energy or disjoining pressure, rheology, and tribology (dewetting and density/thermal fluctuations) are discussed from the viewpoint of thermodynamics. Mesoscopic theories, including microscale mass transfer, are introduced to interpret qualitatively thin PFPE film dynamics. Simulation tools, including a lattice-based SRS model, the off-lattice bead-spring MC, and MD methods were reviewed. These tools accurately describe the static and dynamic behaviors of liquid nano-films. The simulation results are consistent with experimental findings and are thus suitable for describing nanoscale molecular mechanisms in thin film fluid dynamics.

REFERENCES

1. Teletzke, G.F.; Davis, H.T.; Scriven, L.E. How liquids spread on solids. Chem. Eng. Commun. **1987**, *55*, 41–82.
2. Mate, C.M.; Marchon, B. Shear response of molecularly thin liquid films to an applied air stress. Phys. Rev. Lett. **2000**, *85*, 3902–3905.
3. Roco, M. C.; Bainbridge, W. S., Eds. *Converging Technologies for Improving Human Performance*; Kluwer Academic Publisher: Norwell, MA, 2003.
4. Cazabat, A.M.; Fraysse, N.; Heslot, F.; Carles, P. Spreading at the microscopic scale. J. Phys. Chem. **1990**, *94*, 7581–7585.
5. Novotny, V.J. Migration of liquid polymers on solid surfaces. J. Chem. Phys. **1990**, *92*, 3189–3196.
6. Min, B.G.; Choi, J.W.; Brown, H.R.; Yoon, D.Y.; O'Connor, T.M.; Jhon, M.S. Spreading characteristics of thin liquid films of perfluoropolyalkylethers on solid surfaces: effects of chain-end functionality and humidity. Trib. Lett. **1995**, *1*, 225–232.
7. O'Connor, T.M.; Jhon, M.S.; Bauer, C.L.; Min, B.G.; Yoon, D.Y.; Karis, T.E. Surface diffusion and flow activation energies of perfluoropolyalkylethers. Trib. Lett., **1995**, *1*, 219–223.
8. Ma, X.; Gui, J.; Smoliar, L.; Grannen, K.; Marchon, B.; Jhon, M.S.; Bauer, C.L. Spreading of perfluoropolyalkylether films on amorphous carbon surfaces. J. Chem. Phys. **1999**, *110* (6), 3129–3137.
9. Ma, X.; Gui, J.; Smoliar, L.; Grannen, K.; Marchon, B.; Bauer, C.L.; Jhon, M.S. Complex terraced spreading of perfluoropolyalkylether films on carbon surfaces. Phys. Rev. E **1999**, *59* (1), 722–727.
10. Pit, R.; Marchon, B.; Meeks, S.; Velidandla, V. Formation of lubricant ''moguls'' at the head/disk interface. Trib. Lett. **2001**, *10*, 133–142.
11. Ma, X.; Gui, J.; Grannen, K.J.; Smoliar, L.A.; Marchon, B.; Jhon, M.S.; Bauer, C.L. Spreading of pfpe lubricants on carbon surfaces: effect of hydrogen and nitrogen content. Trib. Lett. **1999**, *6* (1), 9–14.
12. Tyndall, G.W.; Karis, T.E.; Jhon, M.S. Spreading profiles of molecularly thin perfluoropolyether films. Trib. Trans. **1999**, *42* (3), 463–470.
13. Mate, C.M. Application of disjoining and capillary pressure to liquid lubricant films in magnetic recording. J. Appl. Phys. **1992**, *72*, 3084–3090.
14. O'Connor, T.M.; Back, Y.-R.; Jhon, M.S.; Min, B.G.; Yoon, D.Y.; Karis, T.E. Surface diffusion of thin perfluoropolyalkylether films. J. Appl. Phys. **1996**, *79* (8), 5788–5790.
15. Dolak, Y.; Hillen, T. Cattaneo models for chemosensitive movement—numerical solution and pattern formation. J. Mathematical Biology **2003**, *46* (5), 460–478.

16. Karis, T.E.; Tyndall, G.W. Calculation of spreading profiles for molecularly-thin films from surface energy gradients. J. Non-Newtonian Fluid Mech. **1999**, *82*, 287–302.
17. Porter, D.A.; Easterling K.E., Eds. *Phase Transformations in Metals and Alloys*; Van Nostrand Reinhold Co. Ltd.: Workingham, Berkshire, England, 1981; Chapter 2.
18. Ferry, J. D. *Viscoelastic Properties of Polymers*, 3rd Ed.; John Wiley & Sons: New York, NY, 1980; 241–254.
19. Karis, T.E.; Marchon, B.; Flores, V; Scarpulla, M. Lubricant spin-off from magnetic recording disks. Trib. Lett. **2001**, *11* (3–4), 151–159.
20. Cantow, M.J.R.; Barrall, E.M., Jr.; Wolf, B.A.; Geerissen, H. Temperature and pressure dependence of the viscosities of perfluoropolyether fluids. J. Polym. Sci. Polym. Phys. **1987**, *25*, 603–609.
21. Toney, M.F.; Mate, C.M.; Leach, K.A. Roughness of molecularly thin perfluoropolyether polymer films. Appl. Phys. Lett. **2000**, *77* (20), 3296–3298.
22. Karis, T.E. Water adsorption on thin film media. J. Colloid Interface Sci. **2000**, *225*, 196–203.
23. Braslau, A.; Pershan, P.S.; Swislow, G.; Ocko, B.M.; Als-Nielsen, J. X-ray reflectivity studies of a microemulsion surface. Phys. Rev. A **1988**, *38*, 2457–2470.
24. Karis, T.E.; Nyak, U.V. Liquid nanodroplets on thin film magnetic recording disks. Trib. Trans. **2004**, *47* (1), 103–110.
25. Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithm to Applications*, 2nd Ed.; Academic Press: New York, NY, 2002.
26. Ma, X.; Bauer, C.L.; Jhon, M.S.; Gui, J.; Marchon, B. Monte Carlo simulations of liquid spreading on a solid surface. Phys. Rev. E **1999**, *60* (5), 5795–5801.
27. Vinay, S.J.; Phillips, D.M.; Lee, Y.S.; Schroeder, C.M.; Ma, X.; Kim, M.C.; Jhon, M.S. Simulation of ultra-thin lubricant films spreading over various carbon overcoats. J. Appl. Phys. **2000**, *87* (9), 6164–6166.
28. Phillips, D.M.; Jhon, M.S. Dynamic simulation of nanoscale lubricant films. J. Appl. Phys. **2002**, *91* (10), 7577–7579.
29. Izumisawa, S.; Jhon, M.S. Molecular simulation of thin polymer films with functional endgroups. J. Chem. Phys. **2002**, *117* (8), 3972–3977.
30. Jhon, M.S.; Izumisawa, S.; Guo, Q.; Phillips, D.M.; Hsia, Y.T. Simulation of nanostructured lubricant films. IEEE Trans. Mag. **2003**, *39* (2), 754–758.

Thiochemicals: Mercaptans, Sulfides, and Polysulfides

Jeffrey H. Yen
Vijay R. Srinivas
Gary S. Smith
Arkema Inc., King of Prussia, Pennsylvania, U.S.A.

INTRODUCTION

The word ''thiochemicals'' is derived from the Greek word ''theion.'' Generally speaking, thiochemicals are sulfur analogs of oxygen-containing compounds; for example, thiochemicals containing sulfur in the +2 oxidation state, mercaptans and alkyl sulfides, are sulfur analogs of alcohols and ethers, respectively.

As the sulfur atom is a homolog of the oxygen atom, present in the VI Group of the Periodic Table, there are some similarities in character and occurrence in nature. The differences, however, are dramatic—oxygen being a colorless gas essential to sustain human life, while sulfur is a yellow solid. The most fundamental molecules that contain oxygen and sulfur, respectively, are water and hydrogen sulfide. Again, water is essential for all living beings and hydrogen sulfide is a deadly and poisonous gas. Many complex molecules of biological interest contain sulfur, and the capability of the sulfur atom to both oxidize and reduce is used extensively in the normal everyday biological processes. Generally, thiochemicals can be classified as follows:

Sulfur (II) compounds

- Basic thiochemicals: examples are sulfur dioxide, hydrogen sulfide, sodium sulfide, sodium thiosulfate, etc.
- Mercaptans (thiols): RSH; examples are methyl mercaptan, ethyl mercaptan, etc.
- Sulfides: RSR′; examples are dimethyl sulfide, diethyl sulfide, etc.
- Polysulfides: $RS_x R'$; dimethyl disulfide, diethyldisulfide, *t*-nonyl polysulfide, etc.
- Sulfenic acids RSOH and derivatives: benzenesulfenic acids/salts, benzenesulfenyl chloride.

Sulfur (IV) compounds

- Sulfoxides RS(O)R and derivatives; dimethylsulfoxide.
- Sulfinic acids and derivatives: RS(O)OH, RS(O)X; benzenesulfinic acid, *p*-toluenesufinyl chloride, other arenesulfinate esters and salts.
- Sulfite esters: $(RO)_2SO$; ethylene sulfite.

Sulfur (VI) compounds

- Sulfones: $R\text{-}SO_2\text{-}R$; sulfolane, dimethyl sulfone, etc.
- Sulfonic acids and derivatives; RSO_2OH, RSO_2X; methanesulfonic acid, *p*-toluenesulfonic acid, methanesulfonyl chloride, methanesulfonamide, etc.

To describe the full range of thiochemicals is beyond the scope of this entry, which will, therefore, place emphasis on the most common thiochemicals. This entry will cover mercaptans, sulfides, and polysulfides, while the next entry will cover sulfonic acids and its derivatives, some sulfoxides, mercaptoacids and their derivatives. The safety, health, and environmental issues for both entries are discussed at the end of this chapter.

MERCAPTANS

Mercaptans and alkyl sulfides are the sulfur analogs of alcohols and ethers, respectively. They can be characterized by their extremely unpleasant odor. These compounds play an important role in biological systems as well as in the application of chemistry to everyday life. Some of the alkyl sulfides are found in many plant and animal oils, and are minor components of petroleum distillates, shale oil, and coal tar.

Chemical and Physical Properties

The properties of mercaptans and alcohols are quite different, although they appear to be similar in nature. The bond dissociation energy of the S–H bond is over 10/kcal/mol lower than the corresponding O–H bond. The ease of free-radical hydrogen abstraction from a mercaptan supports this fact and permits these compounds to be included in preparative free-radical chemistry. The above fact, along with the different boiling points observed for the mercaptans compared to the corresponding alcohols and the differences in their acidities, helps explain why the chemistry of these two classes of compounds differs in so many ways.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120033979

Table 1 lists some properties for various straight-chain alphatic mercaptans (C_2–C_{12}). The boiling points range from 34°C to 277°C, while their solubility in water is the greatest for ethyl mercaptan (6.76/g/L) and decreases as the aliphatic chain length increases.

Manufacturing Technology

There are several methods for the preparation of mercaptans and sulfides.[1] It is our intention here to emphasize only the most facile methods that are safe to operate commercially, use inexpensive raw materials, produce minimum by-products, and allow for simple and easy by-product handling. While the sources of sulfur for thiolation can be many, the most common ones are hydrogen sulfide (H_2S), elemental sulfur, and carbon disulfide (CS_2). H_2S is a component of sour natural gas and can be separated for use in the manufacture of mercaptans and other sulfur chemicals. Additionally, H_2S is generated in the many hydrodesulfurizing units in a refinery. With minimal purification this could be used for the manufacture of thiochemicals. In many instances H_2S is manufactured by either the Girdler or the Folkins processes:

$$\text{Girdler process} \quad H_2 + S \xrightarrow{450^\circ C} H_2S$$

$$\text{Folkins process} \quad CH_4 + S \longrightarrow 2H_2S + CS_2$$

Mercaptans and sulfides are manufactured mainly by the thiolation of alcohols and by the addition of H_2S to olefins.

From alcohols via heterogenous catalysis

Alcohols react with H_2S in the presence of acid or base catalysts to give mercaptans. The nucleophilic substitution of alcohols by H_2S occurs at around 300°C on an alumina-type catalyst impregnated with alkali metal oxides such as Na_2O, K_2O or transition metal oxides such as WO_3. Phosphotungstate alkaline salts on alumina have also been used.

$$ROH + H_2S \xrightarrow[\text{catalyst}]{\text{ca. } 300^\circ C} RSH + H_2O$$

Mercaptans can further react with alcohols, under the same conditions, to produce sulfides. Therefore, the H_2S/ROH ratio must be high to optimize mercaptan selectivity over sulfide selectivity. The temperature also plays a key role in the selectivity and conversion. As the molecular weight of the alcohol increases, the temperature of the reaction needs to be minimized and controlled in the range of 270–400°C, to avoid dehydration of the alcohol to an olefin. Mercaptans ranging from methylmercaptan to dodecylmercaptan have been industrially synthesized by thiolation of primary alcohols. This route is generally used to make primary mercaptans from primary alcohols.

Thiolation of secondary alcohols is not the preferred route to synthesize secondary mercaptans, because of the facile dehydration that occurs under reaction conditions of the secondary alcohols. H_2S can also react with aliphatic or cyclic oxides, viz., ethers; to produce the corresponding thioethers or sulfides. This is one of the preferred ways to make tetrahydrothiophene or thiophene.

From olefins

The catalytic addition of H_2S to olefins follows the Markovnikov rule, with the –SH moeity attaching itself to the carbon atom connected to the least number of hydrogen atoms. With linear olefins, having terminal or internal double bonds, acid catalysis using zeolites, cation-exchange resins, and silica–alumina yields secondary mercaptans. The preferred route for secondary mercaptans is the H_2S addition to olefinic compounds.

$$RCH{=}CHR' + H_2S \longrightarrow RH\overset{\displaystyle SH}{\overset{|}{C}}{-}CH_2R'$$

With branched olefins, one usually obtains tertiary mercaptans as the major product. For example, tertiary mercaptans such as *t*-dodecyl mercaptan and *t*-nonyl mercaptan are manufactured starting from propylene oligomers and using the acid-catalyzed addition of H_2S. In the presence of strong acids, electrophilic addition of H_2S and mercaptans to olefins takes place, yielding Markovnikov adducts. Because divalent sulfur compounds are stronger nucleophiles than alcohol or water, the addition of mercaptans to an intermediate carbonium ion takes place quite readily.[2] This type of reaction yields secondary mercaptans and secondary thioethers from alpha olefins. Since 1965, the ready availability of alpha olefins has opened the possibility of utilizing several of their basic reactions known for a long time. Friedel Crafts catalysts such as anhydrous aluminium chloride, fluoboric acid, mixtures of hydrogen fluoride and boron trifiuoride, and their hydrocarbon complexes have been used for the addition of H_2S to olefins—specifically decene.[3,4] Silica–alumina has been used to make high-molecular-weight mercaptans, which have found use as rubber modifiers.[5] Elemental sulfur along with the bases ammonia or alkylamine or rubber vulcanizing agents such as mercaptobenzothiazole, thiurams, and dithiocarbamates have been used as catalysts to make mercaptans and sulfides from olefins and hydrogen sulfides.[6–12] Acid cation exchangers (e.g., wet sulfonated styrene-divinylbenzene copolymers in the acid form or more advantageously in combination

Table 1 Properties of mercaptans

Product	CAS No.	Formula	Purity[a] (% min.)	Color[b] (APHA max.)	Distillation[c] range(°C)		Molecular wL (calc.)	Density at 20° (Kg/L)	Flash point °C (cc)[d]	Vapor pressure at 20°C (mbar)
					IBP min.	95% max.				
Methyl mercaptan	74-93-1	CH_3SH	99.5	15	4.5	7.5	48.1	0.870	<−18	1870
Ethyl mercaptan	75-08-1	C_2H_5SH	99.3	40	34.4	36.1	62.1	0.837	<−18	575
n-Propyl mercaptan	107-03-09	*n*-C_3H_7SH	98.0	20	66.2	69.4	76.2	0.840	<−18	160
Isopropyl mercaptan	75-33-2	$(CH_3)_2CHSH$	97.0	−15	51.0	56.1	76.2	0.814	<−18	305
n-Butyl mercaptan	109-79-5	*n*-C_4H_9SH	98.0	15	96.3	100.0	90.2	0.843	−7	50
t-Butyl mercaptan	75-66-1	*t*-C_4H_9SH	99.0	15	62.1	67.8	90.2	0.801	<−18	205
n-Hexyl mercaptan	111-31-9	*n*-$C_6H_{13}SH$	96.0	15	149.0	158.0	118.2	0.841	32	5
Cyclohexyl mercaptan	1569-69-3	$C_6H_{11}SH$	99.0	15	155.0	161.0	116.2	0.946	29	5
n-Octyl mercaptan	111-88-6	*n*-$C_8H_{17}SH$	98.0	15	194.0	203.0	146.3	0.845	52	1
t-Nonyl mercaptan	25360-10-5	*t*-$C_9H_{19}SH$	97.0	15	188.7	201.0	160.3	0.848	51	1
n-Dodecyl mercaptan	112-55-0	*n*-$C_{12}H_{25}SH$	98.0	20	269.0	285.0	202.4	0.844	93	1
t-Dodecyl mercaptan	25103-58-6	*t*-$C_{12}H_{25}SH$	98.5	15	227.0	248.0	202.4	0.857	87	1
Pennfloat[d]	Mixture	Mixture	—	N/A	N/A	N/A	Mixture	0.873	>95	0.1
sec-Octyl mercaptan	3001-66-9	$(CH_3)_2CHC_5H_{10}SH$	—	—	—	—	—	—	—	—
n-Decyl mercaptan	143-10-2	*n*-$C_{10}H_{21}SH$	96	15	235	245	174.3	0.849	65	10
n-Tetradecyl mercaptan	2079-95-0	*n*-$C_{14}H_{29}SH$	96.04	40	298	320	230	0.850	>95	1.0
D-Limonene dimercaptan	4802-20-4	$C_{10}H_{20}S_2$	95.0	100	—	—	204	1.034	>120	1

[a]Determined by gas chromatography or titrimetric analysis.
[b]APHA by ASTM D-1209.
[c]ASTM D-1078.
[d]Tag closed cup.

with sulfonated phenol–formaldehyde resins) have been used for making mercaptans and sulfides from olefins.[13–15] Acidified clays such as montmorillonite, cation-exchanged zeolites, and aluminum mercaptides have resulted in catalyzing the Markovnikov addition of H_2S and mercaptans to olefins to give mercaptans and sulfides, respectively.[16–19] When used, trifluoromethanesulfonic acid not only catalyzes the addition of H_2S to olefins but also polymerizes them to give higher-molecular-weight mercaptans than those corresponding to the initial olefin used.[20] The solid supported catalysts are now preferred. The Lewis acids that were formerly used had many disadvantages, such as the need to wash them away from the product mixture, leading to corrosion in specific areas of the process, and high catalyst consumption. The new catalysts give products of high purity and consistency, which can be easily separated by distillation. Consistency is required in many of their applications, especially when the mercaptans are used as chain transfer agents in polymerization. The above process manufactures tertiary mercaptans such as, *tert*-butyl, *t*-octyl, *t*-nonyl, and *t*-dodecyl mercaptan.

The majority of the normal mercaptans, e.g., *n*-dodecyl mercaptan and *n*-octyl mercaptan are currently manufactured by the direct addition of hydrogen sulfide and mercaptans to alpha olefins. It has been well established that in the presence of free-radical initiators or ultraviolet (UV) light of a frequency that would activate H_2S and mercaptans, H_2S and mercaptans add to olefins in an anti-Markovnikov fashion by a mechanism similar to the well-known inverse addition of hydrogen bromide to olefins.[21–24] Several free radicals and UV light catalyze this reaction.[25,26] Olefins exposed to air even for short periods contain enough peroxides to initiate this reaction, while radical inhibitors (e.g., hydroquinone) retard it. A general mechanism may be written, involving a secondary free-radical intermediate that undergoes chain transfer to give a primary mercaptan or sulfide from the addition of H_2S or mercaptan, respectively, to an alpha olefin.

$$RCH{=}CH_2 + H_2S \xrightarrow[\text{radical initiator}]{\text{hv or}} RCH_2CH_2SH \quad \text{(mercaptan)}$$

$$RCH{=}CH_2 + RCH_2CH_2SH \longrightarrow RCH_2CH_2SCH_2CH_2R \quad \text{(sulfide)}$$

If one desires to optimize the yield of mercaptan, there needs to be a large excess of H_2S used in the process. Acetophenone and its derivatives such as 2,2′-diethoxyacetophenone and benzoin ethyl ether, and organic phosphites are a few photoinitiators that can be used to make the above process more efficient.[27,28] A photochemical process starting from α-olefins has industrially manufactured primary mercaptans ranging from C_3 to C_{12} carbons.[29,30]

Other methods for preparation of mercaptans and sulfides

Nucleophilic substitution by salts of H_2S will react with halogenated compounds to make mercaptides that can be acidified to mercaptans. The most convenient salts are sodium and ammonium hydrosulfides. Mercaptoacetic acid is made by the reaction of ammonium hydrosulfide with monochloroacetic acid followed by acidification.

Applications

Mercaptans in general have been used in a wide array of applications, ranging from the production of agricultural chemicals and polymers to miscellaneous use in specific end-user consumer products.[31–33] The existing U.S. market for C_3–C_{12} mercaptans derived from C_3–C_{12} alpha olefins is approximately \$15–20 million/yr. Overall market for mercaptans is over \$300 million. The single largest volume mercaptan produced is methyl mercaptan.

Agricultural chemicals

Several compounds with high biological activity have one or more sulfur atoms. Mercaptans are used to introduce this sulfur atom into these compounds. They have probably been used because they couple high biological activity with good biodegradability. Methyl mercaptan, e.g., is a primary component in the manufacture of methionine, $CH_3SCH_2CH_2CH(NH_2)COOH$, an essential amino acid used to fortify poultry feed. In terms of herbicides and insecticides, the market is relatively mature in the western hemisphere; its growth, however, will come from the development of new and sophisticated agricultural chemicals and their increased use abroad especially in Asia. Propyl mercaptan is used in the manufacture of several herbicides.[34] Another major agricultural area where mercaptans and their derivatives find use is as pesticides or insecticides.[35] Mercaptans, sulfides, sulfoxides, and sulfones are used as plant growth protectors and stimulators.[36,37]

Detergents

Polyoxyethylene mercaptans, obtained from the reaction between an alkyl mercaptan and ethylene oxide, have been demonstrated to be good nonionic

surface-active agents.[38] In more recent years, this class of detergents has become more complex, probably because of the increased availability of inexpensive and sophisticated starting materials such as alpha olefins.[39] Finally, the salts and esters of long-chain aliphatic thioethercarboxylic acids act as wetting agents, emulsifiers, and thickeners for aqueous products.[40]

Lubricants

There are many organosulfur compounds, which have been used in formulating lubricant additives. The sulfide, disulfide, and polysulfide linkages, incorporated into novel compositions-of-matter, serve various functions, such as corrosion inhibition, oxidation resistance, high-temperature stability, and static prevention.[41]

Pharmaceuticals

Organosulfur compounds have been incorporated into various different classes of molecules, which have pharmaceutical utility as anti-inflammatory agents, analgesics, antiarrhythmic agents, antibiotics, antiobesity agents, and so on.[42,43]

Polymers

The uses of mercaptans in polymers fall into three major categories: chain transfer agents, additives such as stabilizers against heat or UV light, and monomers that incorporate an alkylmercapto group into their structure. Mercaptans *t*-dodecyl, *n*-dododecyl, etc. are excellent chain transfer agents used to control molecular weight of several different kinds of polymers, styrene butadiene rubber, acrylonitrile–butadiene–styrene, polyacrylates, to name a few.[33,44,45]

Gas odorants

Mercaptans and sulfides are used extensively to odorize natural gas.[46] The low-odor threshold limit values of most mercaptans and sulfides make them ideally suited for this purpose. Very small quantities can thus be used so as to not contribute to sulfur dioxide emissions when burned. Pure products such as tetrahydrothiophene, *tert*-butyl mercaptan, methylethyl sulfide, or mixtures/blends of these products and others such as isopropyl mercaptan, diethyl sulfide, dimethyl sulfide, ethyl mercaptan, and *n*-propyl mercaptan are used.

SULFIDES AND DISULFIDES

Chemical and Physical Properties

The common chemical and physical properties of commercial sulfides are shown in Table 2.[26,27]

Manufacturing Technology

A commonly used method to prepare sulfides is to react a chloride and sodium mercaptide. Mercaptans and alcohols will react in the presence of acidic catalysts, at elevated temperatures (300–400°C), to produce sulfides:

$$RCl + R'SNa \longrightarrow RSR' + NaCl$$
$$ROH + R'SH \longrightarrow RSR' + H_2O$$

Mercaptans can be added to olefins to produce sulfides. Markovnikov addition occurs readily in the presence of acidic catalysts, at relatively mild temperatures, especially with branched, unsymmetrical (i.e., polarized) olefins.

$$R_2C{=}CH_2 + R'SH \longrightarrow RH\,\overset{\displaystyle SR'}{\overset{|}{C}}\,{-}CH_3$$

Anti-Markovnikov additions occur in the presence of free-radical catalysts such as peroxides, azo

Table 2 Properties of commercial sulfides

Product	Physical state	Physical and chemical properties	Uses
Dimethyl sulfide CH_3SCH_3 CAS: 75-18-3 EINECS: 200-846-2	Water white and strongly odorous liquid	BP = 37.3°C $d(20/4)$ = 0.840	LPG gas odorant, methylation agent
Diethyl sulfide $C_2H_5SC_2H_5$ CAS: 352-93-2 EINECS: 206-526-9	Water white and strongly odorous liquid	BP = 92°C $d(20/4)$ = 0.840	Gas odorant, sulfiding additive for hydrotreating catalysts and anticooling additive for steamcrackers

(From Refs.[47,48].)

compounds, or UV light. This reaction proceeds readily, even at room temperature or lower.

$$RCH{=}CH_2 + R'SH \longrightarrow RCH_2CH_2SR'$$

So readily does this reaction occur, that care must be taken to avoid using olefins having appreciable peroxide levels when the Markovnikov addition product is desired, to prevent the formation of isomeric sulfides.

Disulfides are manufactured by the oxidation of unhindered mercaptans. This reaction is one of the most facile reactions in organic chemistry. On prolonged exposure to air, mercaptans become contaminated with several percent of the corresponding disulfides. Primary mercaptans oxidize more readily than secondary; while tertiary mercaptans oxidize the least readily.

$$2RSH + \frac{1}{2}O_2 \longrightarrow RSSR + H_2O$$

The oxidation is greatly accelerated by an alkaline medium. The mildest of oxidizing agents affect this reaction. Iodine will convert mercaptans quantitatively to their disulfides. Because overoxidation does not occur in this case, iodine can be used for the quantitative determination of mercaptans.

$$2RSH + I_2 \longrightarrow RSSR + 2HI$$

Under the proper mild conditions, chlorine can also be used, but overoxidation and sidereactions often result in some yield loss. Elemental sulfur can be employed, but it must be used in a molar deficiency to prevent the formation of tri- and higher polysulfides as by-products. Conversely, higher-rank polysulfides can be retrograded back to disulfides by reaction with mercaptans.

$$2RSH + S \longrightarrow RSSR + H_2S$$
$$RSSR + S \longrightarrow RSSSR$$
$$2RSS_XSR + 2RSH \longrightarrow RSSR + RSS_{X-1}SR + H_2S$$

Disulfides can also be produced by passing mixtures of the mercaptan and air (or dilute oxygen in an inert gas) over copper or cobalt oxide/molybdenum oxide catalysts at elevated temperatures.

Applications

Sulfides find use (Table 3) in the refinery and petrochemical industries.[49–55] Some refiners use dimethyl sulfide as a sulfiding agent to convert the metal oxide hydrotreating catalysts to their sulfide form.[56] In ethylene manufacture, ethane or other hydrocarbon liquid like naphtha is reacted at high temperatures with steam. The process would make an unacceptable amount of coke and produce too much carbon

Table 3 Properties of commercial disulfides

Product	Physical state	Physical and chemical properties	Uses
Dimethyldisulfide CH_3SSCH_3 CAS: 624-92-0 EINECS: 210-671-0	Colorless to pale yellow and strongly odorous liquid	BP = 109.6°C $d(20/4)$ = 1.063	Thiomethylation agent, sulfiding agent for catalysts, anticoking agent for steam crackers, deplugging agent for sour gas wells
Diethyldisulfide $C_2H_5SSC_2H_5$ CAS: 110-81-6 ElNECS: 203-805-7	Colorless to pale yellow and strongly odorous liquid	BP = 154°C $d(20/4)$ = 0.993	Thioethylation agent
Dibenzyldisulfide $C_6H_5CH_2SSCH_2C_5H_4$ CAS: 150-60-7 ElNECS: 205-764-0	Pink colored crystals	MP = 60–72°C BP = 210–216°C (24 mbar) $d(20/4)$ = 1.300	Lub and grease additive
Disec-butyl disulfide sec-C_4H_9SS-sec-C_4H_9 CAS: 5943-30-6 EINECS: 227-702-1	Colorless liquid	BP = 212°C $d(20/4)$ = 0.950	Synthesis of intermediates
Di *tert*-dodecyl disulfide *tert*-$C_{12}H_{25}$SS-*tert*-$C_{12}H_{25}$ CAS: 27458-90-8 EINECS: 248-468-7	Pale yellow and strongly odorous liquid	BP = 321°C $d(20/4)$ = 0.911	Antioxydant for polymers

(From Refs.[47,48].)

monoxide if no sulfur compounds are added. Therefore, almost all of the steam cracking furnaces in the world use a sulfur compound that can generate H_2S under conditions of operation to minimize coke and CO. Some of the lower-molecular-weight sulfides have been used as additives to aid in the metal mining industry.[57] Methyl ethyl sulfide is used as a gas odorant, in combination with other mercaptans and sulfides.

Among the disulfides, dimethyl disulfide (DMDS) is a very versatile product experiencing broad growth. Most recently, it has been shown to have superior nematocidal activity and will soon find application as a broad-spectrum fumigant. As a thiomethylation agent, it is used to produce methylthiophenols, which have found use as pesticide, pharmaceutical, and antioxidant intermediates. The synthetic method used is the electrophilic substitution of phenol with dialkyl disulfides, resulting in the formation of both *ortho*- and *para*-thiosubstituted products. Work continues to improve the regiospecificity of this reaction. With its lower odor and vapor pressure, it is safer to use instead of methyl mercaptan in some specific applications.

Dimethyl disulfide has been found to be the most efficient catalyst sulfiding agent in the refining industry. It is more efficient and easier to handle than competing products such as CS_2 and H_2S. In the refining industry, it is largely used for the sulfidation of metal oxide catalysts to metal sulfides, known to have sustained activity for various treatments such as hydrodesulfurization, hydrodenitrefication, and hydrocracking of crude petroleum. Dimethyl disulfide is also used to modify the selectivity of platinum-based catalysts in reforming and to prevent coke formation in hydrodealkylation processes. The clean breakdown of DMDS at relatively low temperatures into simple products like methyl mercaptan, H_2S, and methane, on typical hydrotreating catalysts, makes it the agent of choice for catalyst sulfiding. It is also used as a sulfur carrier in the steamcracking of hydrocarbons to produce ethylene and propylene. Dimethyl disulfide generates H_2S, which is known to minimize catalytic coke and also CO, a gaseous by-product that poisons the catalysts used downstream to partially hydrogenate acetylene, coproduced during ethylene manufacture.

Dimethyl disulfide is also an excellent ''solvent'' for sulfur. In fact, DMDS chemically reacts with sulfur in the presence of a base catalyst to form dimethylpolysulfide.[31]

$$CH_3SSCH_3 + S \xrightarrow[\text{catalyst}]{\text{base}} CH_3S_XCH_3$$

$$CH_3S_XCH_3 + H_2S \longrightarrow CH_3SH + (X-2)S$$

This property makes DMDS the agent of choice for the deplugging of sour gas wells ($H_2S < 60\%$). In the bottom of wells where pressure and temperature are high, sulfur is soluble in the gas, but at the top, where pressure and temperature decrease, the sulfur precipitates in the piping, thus plugging the flow of gas. Catalyzed DMDS, known as SulfaHitec®, a product made by Arkema, can be injected batchwise to unplug sour gas wells and to keep pipes and formations free of precipitated sulfur.

POLYSULFIDES

Chemical and Physical Properties

Most polysulfides with a sulfur rank higher than three, are mixtures where the different sulfur ranks coexist in equilibrium (Table 4). In many cases elemental sulfur is also involved in the above equilibrium. Changes in temperature and pressure sometimes alter this equilibrium and precipitate sulfur. Viscosity of the polysulfides also is a function of temperature, increasing dramatically with decreasing temperature. In most cases heating polysulfides results in their decomposition to the alkyl mercaptans. Some of the aromatic polysulfides are tacky solids. Many trisulfides can be isolated as pure compounds, and exhibit unique chemical properties. They are the only polysulfides that are not corrosive to copper.

Manufacturing Technology

Mercaptans can be reacted with elemental sulfur at moderate temperatures (50–120°C) in the presence of alkaline catalysts, such as trialkylamines, to produce polysulfides. The first step in the reaction sequence is always disulfide formation (with liberation of H_2S). The subsequent insertion of additional sulfur atoms occurs readily. The number of sulfur atoms inserted will depend on the molar ratio of sulfur to mercaptan employed, according to the stoichiometry of the following equations:

$$2RSH + 2S \longrightarrow RSSSR + H_2S$$

$$2RSH + 3S \longrightarrow RS_4R + H_2S$$

$$2RSH + 4S \longrightarrow RS_5R + H_2S$$

Solid catalysts such as zeolite (aluminosilicate), heteropolyacids and their alkali or alkaline earth metal salts on a metal oxide support and basic resins have been used as catalysts to manufacture polysulfides.

Polysulfides above RS_5R are generally unstable. The disulfides and trisulfides can be produced in fairly pure form by using appropriate reaction conditions and slightly less than the stoichiometric amounts of sulfur,

Table 4 Chemical and physical properties of poly-*ter*-alkylphenol disulfides

Product	Physical state	Physical and chemical properties	Uses
Poly-*tert*-amylphenol disulfide CAS: 68555-98-6 EINECS: conforms	Brown, slightly tacky solid	Liquid density at 25°C = 1150–1200 kg/m^3 Flash Point (closed cup) = 235°C Sulfur content ≥ 21.8% ≤ 23.8% Softening point ≥ 50°C ≤ 60°C	Nitrosamine-free vulcanizing agents; stabilizer of rosin resins
Poly-*tert*-amylphenol disulfide CAS: 68555-98-6 EINECS: conforms	Reddish brown to dark brown pastilles	Liquid density = 1200 kg/m^3 Flash point (closed cup) = 171°C Sulfur content ≥ 27% ≤ 29% Softening point ≥ 95°C ≤ 115°C	Nitrosamine-free vulcanizing agents; stabilizer of rosin resins
Poly-*tert*-amylphenol disulfide Mixed with 25% Sipernat CAS: 68555-98-6 and 112926-00-8 EINECS: conforms	Beige free flowing powder	Pour/lap density at 25°C = 360 kg/m^3 Sulfur content ≥ 20% ≤ 22% Water content ≤ 2% 99 wt% passes 840 μm sieve 98 wt% passes 300 μm sieve	Nitrosamine-free vulcanizing agents
Poly-*tert*-butylphenol disulfide CAS: 60303-68-6 EINECS: conforms	Yellow to light brown pastilles	Solid density at 20°C = 805 kg/m^3 Liquid density at 80°C = 1175 kg/m^3 Flash point (closed cup) > 250°C Decomposition temperature = 217°C Sulfur content ≥ 29.5% ≤ 31.5% Softening point ≥ 95°C ≤ 115°C	Nitrosamine-free vulcanizing agents
Poly-*tert*-butylphenol disulfide mixed with stearic acid CAS: 60303-68-6 and 57-11-4 EINECS: conforms	Yellow to light brown pastilles	Decomposition temperature > 200°C Flash point (closed cup) > 200°C Sulfur content ≥ 26.4% ≤ 28.4% Softening point ≥ 75°C ≤ 95°C Stearic acid content ≥ 9% ≤ 1.1%	Vulcanizing agents

(From Refs.[47,48].)

but the higher-rank $-S_4-$ and $-S_5-$ polysulfides are generally produced as mixtures containing the tri–, tetra–, and pentasulfides. The higher dialkyl pentasulfides, such as *tert*-nonyl-S_5-*tert*-nonyl (TNPS), on standing, tend to precipitate some sulfur and equilibrate at an average sulfur ''rank'' of about 4.7.

Applications

The higher dialkyl polysulfides, such TNPS, find use as extreme-pressure lube additives, especially for cutting oils; the lower dialkyl polysulfides, such as $(CH_3)_2S_x$, as sulfur-donor agents, for example, to convert phosphites to thiophosphates:

$$(RO)_3P + R'S_XR' \longrightarrow (RO)_3P{=}S + R'S_{X-1}R'$$

Disulfides and polysulfides are useful for sulfiding hydrotreating catalysts, used in petroleum refining to convert metal oxides to the preferred metal sulfides. Hydrotreating is an essential process in the refining of petroleum. It removes heteroatoms, nitrogen, and sulfur from crude oil and its fractions, formulated into gasoline and diesel (Table 5).

SAFETY, HEALTH, AND ENVIRONMENTAL ISSUES

6.1 Material Handling[47]

Mercaptans are flammable. Normal precautions used in handling any flammable chemical are applicable. Flares used to burn vapors to reduce pressure on a storage tank should be situated as far away from the general storage area as practical. The flare should have a suitable flame arrestor. All potential sources of fire, flame, and sparks in the immediate area should be sought out and eliminated.

Plastic gloves and chemical goggles should be worn to protect the skin and eyes when handling mercaptans. If mercaptan contacts the skin or eyes, flush the affected area immediately and thoroughly with water. In handling mercaptans, avoid spills and leaks of liquid or vapor, avoid prolonged breathing of mercaptan vapors. Although the odor of mercaptan will normally become extremely disagreeable before the vapors reach dangerous concentrations, the nose may become temporarily desensitized after exposure. A self-contained breather-type mask is recommended for protection when working in areas of high vapor concentration or for prolonged exposure to lower concentrations of vapor.

Table 5 Chemical and physical properties of polysulfides

Commercial product	Physical state	Physical and chemical properties	Uses
TPS 20 Di-*tert*-dodecyl trisulfide CAS: 68425-15-0 EINECS: 270-335-7	Pale yellow and slightly odorous liquid	BP > 200°C (decomposition) $d(20/4) = 0.95$	Extreme pressure additive for lubricating and cutting oils
TPS 32 Di-*tert*-decyl pentasulfide CAS: 68425-15-0 EINECS: 270-335-7	Pale yellow and slightly odorous liquid	BP > 200°C (decomposition) $d(20/4) = 1.01$	Extreme pressure additive for lubricating and cutting oils; cross-linking agent for elastomers.
TPS 37 Di-*tert*-nonyl pentasulfide CAS: 68425-16-1 EINECS: 270-336-2	Yellow and slightly odorous liquid	BP > 200°C (decomposition) $d(20/4) = 1.024$	Extreme pressure additive for lubricating and cutting oils; cross-linking agent for elastomers
TPS 44 Di-*tert*-butyl polysulfide CAS: 68937-96-2 EINECS: 273-103-3	Yellow and slightly odorous liquid	BP > 178°C (decomposition) $d(20/4) = 1.007$	Extreme pressure additive for gear oils and cutting oils
TPS 54 Di-*tert*-butyl polysulfide CAS : 68937-96-2 EINECS: 273-103-3	Yellow and slightly odorous liquid	BP > 170°C (decomposition) $d(20/4) = 1.082$	Extreme pressure additive for cutting oils

(From Refs.[47,48].)

To prevent accidental ignition of mercaptan vapors, employees working in handling and storage areas should not wear metal heel- or toe-plates on shoes. Nonsparking tools should be used when working on mercaptan equipment or containers.

The general precautions discussed in handling mercaptans are applicable to Thiophene and tetrahydrothiophene. Di-*t*-nonyl polysulfide is a relatively safe nontoxic material. It can be handled in the same manner as any fuel oil.

Materials of Construction[47]

Steel, stainless steel, and copper-free steel alloys are the preferred materials of construction for mercaptan service. In particular, stainless steel should be used for any vessel or line that is to be open and exposed to air frequently. Aluminum is also suitable for mercaptan use provided the pressure rating of aluminum equipment or piping is sufficient to meet the pressure requirements of the application. Iron or carbon steel is less acceptable than stainless steels or aluminum although it can be used if appropriate measures are taken to condition the iron or carbon steel equipment before putting it into service. Allowing a small amount of mercaptan to stand in it for a period of time and subsequently keeping the equipment under a dry, inert atmosphere can prevent corrosion of the equipment. The hazard in using iron or carbon steel is the formation of iron–sulfur complexes, which are pyrophoric and constitute a potential fire hazard because of the flammability of mercaptans. If iron or carbon steel is used in mercaptan service, thorough washing with water during disassembly and cleaning is absolutely necessary.

Sulfonic acids are corrosive. Do not use copper or copper-bearing alloys for mercaptan service. Mercaptans readily attack these metals and are contaminated by them. Mercaptans are odorous; therefore, storage tanks should not be vented to the atmosphere. Venting should be to a flare or a scrubber to remove the mercaptan.

Toxicity[47]

All alkyl mercaptans have distinctive mercaptan odors, and their presence in air is quite evident. While this may present a problem in community relations, it is a built-in warning system. It is almost impossible to have an undetected mercaptan leak. However, there is a danger in ignoring the mercaptan odor. The lower-molecular-weight mercaptans, in high concentrations, tend to deaden the sense of smell; therefore, dangerous concentrations of vapors could go undetected.

The lower-molecular-weight mercaptans have the strongest odors. In very low concentrations (ppm range) they are recognized as gas leaks, because they are commonly used to odorize fuel gas. In intermediate concentrations their pungent odor has been likened to that of garlic, rotting cabbage, and other decomposed

organic matter. In higher concentrations, the desensitization effect takes place, and an alcoholic or somewhat pleasant odor may be experienced.

Although some oral toxicity data are available, inhalation is the most likely route of entry into the body. Most mercaptans have an LC_{50} greater than 2000 ppm. This is above the normally accepted ''Toxic Substance'' classification. The exception, *t*-octyl mercaptan, has been tested recently and has been found to be a ''highly toxic substance'' with LC_{50} of less than 200 ppm.

No cumulative or permanent sensitization effects have been reported from exposure to mercaptans. Long-term exposure to high concentrations of the lower-molecular-weight mercaptans can result in headache, nausea, loss of consciousness, and even death. This is possible because of olfactory desensitization. However, death from mercaptan inhalation is extremely rare, and if the victim is removed to fresh air, even if unconscious, recovery is complete.

CONCLUSIONS

Thiochemicals are among one of the more important intermediates and end-user chemicals, starting from a relatively simple molecule such as H_2S to complex molecules such as polysulfides. Many thiochemicals that are commercially manufactured provide building blocks for more specialized ones. Thiochemicals enjoy applications as intermediates in the manufacture of agricultural chemicals, pharmaceuticals, detergents, lubricants, etc. Some thiochemicals have played important roles, as solvents, chain transfer agents in polymer synthesis, electroplating solutions, presulfiding agents for hydrotreating catalysts, coke and CO inhibitors in the steam cracking process, etc.

ACKNOWLEDGMENT

The authors acknowledge the support and information provided by Arkema Inc.

REFERENCES

1. Reid, F.F. *Organic Chemistry of Bivalent Sulfur*; Chemical Publishing Company: New York, 1958; Vol. 1.
2. Ohno, A.; Oae, S. Thiols. In *Organic Chemistry of Sulfur*; Oae, S., Ed.; Plenum Press: New York, 1977.
3. Bell, R.T.; Thacker, C.M U.S. Patent 2,498,872, 1950.
4. Bell, R.T U.S. Patent 2,531,602, 1950.
5. Schulze, W.A U.S. Patent 2,502,596, 1950.
6. Jones, S.O.; Reid, E. J. Am. Chem. Soc. **1938**, *60*, 2452.
7. Louthan, R.P U.S. Patent 3,221,056, 1965.
8. Doss, R.C U.S. Patent 3,419,614, 1968.
9. Warner, P.F U.S. Patent 3,114,776, 1963.
10. Lang, A.; Vannel, P U.S. Patent 3,333,008, 1967.
11. Pennsalt Chem Corp. U.S. Patent 2,951,875, 1960.
12. Elf Aquitaine European Patents 101,356, 1982.
13. Ipatieff, V.N.; Pines, H.; Friedman, B.S. J. Am. Chem. Soc. **1938**, *60*, 273.
14. Macho, V Czech Patent 185,469, 1980.
15. Arretz, E.; Mirassou, A.; Landoussy, C.; Auge, P. Fr. Demande FR. 2,531,426, 1984.
16. Kubicek, D.H Belgium Patent 886,261, 1981.
17. Onyestyak, O.; KaHo, D.; Papp, J.; Detrekoy, E. Hung. Teijes Hungarian Patent 29,972, 1984.
18. Fried, H.E Fur. Patent 122,654, 1984.
19. Hahn, W German Patent 1,110,631, 1961.
20. Fields, E.K U.S. Patent 4,347,384, 1982.
21. Kharasch, M.S.; Read, A.T.; Mayo, F.R. Chem. Md. **1938**, *752*.
22. Posner, H. Ber. Dtsch. Chem. **1905**, *38*, 646.
23. Ashworth, F.; Burkhardt, G.N. J. Chem. Soc. **1928**, 1791–1802.
24. Grattan, D.W.; Locke, J.M.; Wallis, S.R. J. Chem. Soc. Perkin Trans. **1973**, *1*, 2264.
25. Mayo, F.R.; Walling, C. Chem. Rev. **1940**, *27*, 387.
26. Evans, E.A.; Vaughan, W.E.; Rust, F.F U.S. Patent 2,376,675, 1945; U.S. Patent 2,411,961, 1946; British Patent 567,524, 1946.
27. Dimmig, D.A U.S. Patent 4,140,604, 1979; European Patent 5400, 1979.
28. Olivier, J.; Souloumiac, J.; Suberlucq, J U.S. Patents 4,233,128, 1980, 4,443,310, 1984.
29. Buchholz, B U.S. Patent 3,652,680, 1972.
30. Bierker, G.L.; Kiovnick, A U.S. Patent 4,043,886, 1977.
31. Pennwalt Corporation Publication S-222B; Brandrup, J., Immergut, E.H., *Polymer Handbook,* 2nd Ed.; Wiley–Interscience: New York, 1975.
32. Mayo, F.R. J. Am. Chem. Soc. **1943**, *65*, 2324.
33. Pryor, W.A. *Mechanisms of Sulfur Reactions*; McGraw-Hill: New York, 1962; 82–88.
34. Allesandrin, C.G U.S. Patent 4,119,659, 1978.
35. Punja, N U.S. Patents 4,429,153, 1984, 4,370,346, 1983.
36. Mathew, C.T.; Ulmer, H.E U.S. Patent 3,393,331, 1976.
37. Fankhauser, E.; Sturm, E U.S. Patent 4,459,152, 1984.
38. Lemaire, H. *Nonionic Surfactants*; Schick, M.J., Ed.; Marcel Dekker: New York, 1966.
39. Onopchenko, A.; Schultz, J U.S. Patents 4,102,932, 1978, 4,009,211, 1977.

40. Suzuki, S U.S. Patent 4,172,211, 1979.
41. Mawn, P.E. Outlook for lube oil additives. Impact (L760704), 1976, July 29.
42. Carson, J U.S. Patent 4,213,905, 1980.
43. Fujimoto, T.; Kondo, K.; Suda, M.; Tunemoto, D U.S. Patent 4,268,442, 1981.
44. Pennwalt Corporation Publication S-222B; Brandrup, J.; Immergut, E.H. *Polymer Handbook*, 2nd Ed.; Wiley–Interscience: New York, 1975.
45. Mayo, F.R. J. Am. Chem. Soc. **1943**, *65*, 2324.
46. Nevers, A.D U.S. Patent 3,826,631, 1974.
47. Arkema Corp. (formerly Pennwalt Corporation). *Organosulfur Intermediates*, 2000; Arkema corp.: Philadelphia, PA, 1979.
48. Arkema Corp. (formerly Pennwalt Corporation) *Organic Chemicals Division Product Catalog* 2000; Arkema Corp.: Philadelphia, PA, 1989.
49. Pennwalt European Patents 171,092, 1986, 202, 420, 1986.
50. Elf Aquitaine European Patents 269,517, 1988, 337,837, 1989; Lubrizol Corp. U.S. Patent 4,344,854, 1982.
51. EI. Du Pont de Nemours U.S. Patents 4,792,633, 1988, 4,599,451, 1986.
52. Vottero, C.; Labat, Y.; Poirier, J.-M. European Patent 318,394, 1989.
53. Lindstrom, M Atochem North America. U.S. Patent 5,028,343, 1991.
54. Clark, P.D.; Lesage, K.L. Alberta Sulfur Res. Ltd. Q. Bull. **1987**, *24* (1–2), 23–43, *27* (3), 41–43 **1990**.
55. Labat, Y. U.S. Patents 4,728,447, 1998; 4,827,040, 1989.
56. Halle, H. Oil Gas J. **1980**, *70*.
57. Wiechers, A. U.S. Patent 4,211,644, 1980.

Thiochemicals: Mercapto Acids and Organosulfur (IV/VI) Compounds

Jeffrey H. Yen
Gary S. Smith
Vijay R. Srinivas
Arkema Inc., King of Prussia, Pennsylvania, U.S.A.

INTRODUCTION

As discussed in the previous entry, ''Thiochemicals: Mercaptans, Sulfides, and Polysulfides,'' thiochemical is derived from the Greek word ''theion.'' Generally speaking, thiochemicals containing sulfur formally in the +2 oxidation state are analogs of oxygen-containing compounds; for example, mercaptans and alkyl sulfides are sulfur analogs of alcohols and ethers, respectively. Unlike oxygen, sulfur can also exist in the formal +4 and +6 oxidation states, thus affording a vast array of functionalities. Generally, thiochemicals can be classified into the following categories:

Sulfur (II) compounds

- Basic thiochemicals: Examples are sulfur dioxide, hydrogen sulfide, sodium sulfide, sodium thiosulfate, etc.
- Mercaptans (thiols): RSH; examples are methyl mercaptan, ethyl mercaptan, etc.
- Sulfides: RSR′; examples are dimethyl sulfide, diethyl sulfide, etc.
- Polysulfides: RS_xR'; dimethyl disulfide, diethyldisulfide, *t*-nonyl polysulfide, etc.
- Sulfenic acids RSOH and derivatives: benzenesulfenic acids/salts, benzenesulfenyl chloride.

Sulfur (IV) compounds

- Sulfoxides: RS(O)R and derivatives; dimethylsulfoxide.
- Sulfinic acids and derivatives: RS(O)OH, RS(O)X; benzenesulfinic acid, *p*-toluenesufinyl chloride, other arenesulfinate esters, and salts.
- Sulfite esters: $(RO)_2SO$; ethylene sulfite.

Sulfur (VI) compounds

- Sulfones: $R\text{-}SO_2\text{-}R$; sulfolane, dimethyl sulfone, etc.
- Sulfonic acids and derivatives: RSO_2OH, RSO_2X; methanesulfonic acid (MSA), *p*-toluenesulfonic acid (PTSA), methanesulfonyl chloride (MSC), methanesulfonamide, etc.

To describe the full range of thiochemicals is beyond the scope of this entry, which will therefore emphasize information on the most common thiochemicals of commerce. This entry will cover mercaptoacids, sulfonic acids and their derivatives, some sulfoxides, etc., while the previous entry covers mercaptans, sulfides, and polysulfides. The safety, health, and environmental issues for both entries are discussed at the end of the previous entry.

MERCAPTO ACIDS, SALTS, ESTERS

Chemical and Physical Properties

Mercaptocarboxylic acids are difunctional molecules with both carboxyl and thiol functionality. The general formula is $HS\text{-}R\text{-}CO_2H$, with R being alkyl or aryl. Both the carboxyl and thiol groups are acidic. For mercaptoacetic acid and thioglycolic acid the pK_a values are 3.6 and 10.2, respectively. The physical properties of commercial mercaptoacids, esters, and salts are given in Table 1. The carboxylic acid group can be readily converted to afford salts or esters. Mild oxidants such as I_2 convert the thiol to the disulfide. The generated disulfides can undergo thiol/disulfide exchange to afford unsymmetrical disulfides. The thiol group also reacts readily with aldehydes or ketones to afford dithioketals, with acyl chlorides or anhydride to afford thioesters, and with α,β-unsaturated carbonyl and nitrile compounds. As discussed below, the thiol readily reacts with alkyl halides to form dialkyl sulfides.

Manufacturing Technology

Mercaptoacetic acid, also known as thioglycolic acid, is prepared from the continuous reaction of chloroacetic acid with two molar equivalents of aqueous ammonium or metal hydrosulfide. The reaction is run under autogenous H_2S overpressure, derived from the neutralization of the chloroacetic acid with one equivalent

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016066

Table 1 Physical properties of selected mercapto carboxylic acids, esters, and derivatives

Name	Structure	MW	Melting point (°C)	Boiling point (°C/torr)	CAS Reg. No.
Mercaptoacetic acid (thioglycolic acid)	$HSCH_2CO_2H$	92.12	−16	123/29	68-11-1
2-Mercaptopropionic acid (thiolactic acid)	$HSCH(CH_3)CO_2H$	106.1		85/5	79-42-5
3-Mercaptopropionic acid	$HS(CH_2)_2CO_2H$	106.1	15.5	111/15	107-96-0
Thiodiglycolic acid	$S(CH_2CO_2H)_2$	150.2	128		123-93-3
Dithiodiglycolic acid	$(SCH_2CO_2H)_2$	182.2	100–102		505-73-7
Thiodipropionic acid	$S(CH_2CH_2CO_2H)_2$	178.2	130		111-17-1
Dithiodipropionic acid	$(SCH_2CH_2CO_2H)_2$	210.3	152–156		1119-62-6
Methyl mercaptoacetate		106.1		148/760	2365-48-2
Ethyl mercaptoacetate		120.2		155/760	623-51-8
Butyl mercaptoacetate		148.2		110/40	10047-28-6
2-Ethylhexyl mercaptoacetate		204.3		106/3	7659-86-1
Isooctyl mercaptoacetate (mixture of isomers)		204.3		107/4	25103-09-7
Methyl 3-mercaptopropionate		120.2		166/760	2935-90-2
Butyl 3-mercaptopropionate		162.3		101/12	16215-21-7
2-Ethylhexyl 3-mercaptopropionate		218.4		85–87/2	50448-95-8
Isooctyl 3-mercaptopropionate (mixture of isomers)		218.4		95/2.2	30774-01-7
Ethylene glycol dimercaptoacetate		210.3		137/2	123-81-9
Ethylene glycol dimercaptopropionate		238.3		101/0.3	22504-50-3
2-Mercaptoethyl tallate	$HSCH_2CH_2O_2C$–R (R = unsaturated C_{18}–C_{20})				68440-24-4

of hydrosulfide. Because H_2S ($pK_{a1} = 7.0$) has approximately 1000 times the acidity of the thiol group on the formed mercaptoacetic acid, the presence of the H_2S inhibits formation of the mercaptoacetate dianion and subsequent reaction with chloroacetate to form the sulfide. The crude mercaptoacetate salt from the reaction is acidified, extracted into organic solvents, and ultimately recovered by vacuum distillation. Principal by-products are the halide salts.

$$ClCH_2CO_2H + 2MSH \longrightarrow HSCH_2CO_2M + MCl + H_2S \quad (1)$$

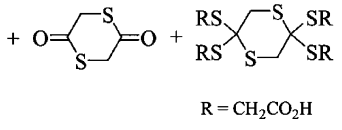
$$HSCH_2CO_2M + HCl \longrightarrow HSCH_2CO_2H + MCl \quad M = NH_4, Na, K, Ca \quad (2)$$

$$HSCH_2CO_2^- + HS^- \rightleftharpoons {}^-SCH_2CO_2^- + H_2S \quad (3)$$

$$ClCH_2CO_2^- + {}^-SCH_2CO_2^- \longrightarrow S(CH_2CO_2^-)_2 + Cl^- \quad (4)$$

The product mercaptoacetic acid will self-esterify on storage or heating. The self-esterification products are primarily the linear dimer ($x = 1$), along with some cyclic dimer, oligoesters ($x = 2+$), and solid orthothioester. These self-esterification reactions, with the exception of the formation of the orthothioesters, can be readily reversed in water with dilute acid or base catalyst.

$$HSCH_2CO_2H \overset{\Delta,\,H^+}{\rightleftharpoons} HS[CH_2C(O)S]_xCH_2CO_2H + H_2O$$



$$R = CH_2CO_2H \quad (5)$$

Most mercaptoacetic acid is sold in the form of esters, primarily for use in organotin stabilizers for poly(vinyl chloride) (PVC). The esterification is performed in batch or continuous mode, primarily using 2-ethylhexanol or isooctanol.

$$HSCH_2CO_2H + ROH \xrightarrow{H+} HSCH_2CO_2R + H_2O \quad R = \text{2-ethylhexyl, isooctyl, butyl, etc.} \quad (6)$$

Purified mercaptoacetic acid is sold in anhydrous or 85% aqueous forms, most often for subsequent conversion to the esters or to the ammonium, sodium, potassium, or calcium salts. 3-Mercaptopropionic acid is produced from metal hydrosulfides and either acrylic acid or acrylonitrile. Mercaptoethyl tallate is another mercapto-ester used in commercial organotin stabilizers. It is manufactured by a standard esterification of mercaptoethanol and tall oil, a mixture of fatty acids.

Applications

Mercaptoacetate salts have extensive applications in personal care, particularly in cold-waving of human hair, hair straightening, and as depilatories. Essentially, the dianionic mercaptide reacts with the cystine (disulfide) cross-links between the keratin polypeptide chains via mercaptan–disulfide exchange. This allows deformation of the keratin with resultant curling or straightening of the hair. The cystine disulfide linkages are then restored by use of a ''neutralizer,'' actually an oxidant such as H_2O_2, sodium perborate, or sodium or potassium perbromate. Similarly, mercaptoacetic acid has found applications in leather treatments, including dehairing and softening of hides, and in the treatment of wool fibers.

Mercaptoacetate and mercaptopropionate esters are key raw materials in the manufacture of organotin stabilizers for use in PVC. The mercapto-esters are reacted with organotin chlorides under two-phase aqueous conditions, with the liquid products being isolated by phase separation and stripping of residual water.[1,2] The generated organotin mercaptides are the workhorse thermal stabilizers for processing of PVC. Methyl and octyltin mercaptide stabilizers are used in PVC for food-contact applications because these materials offer very low migration from the PVC and low overall toxicity. Butyltin mercaptides are employed in the production of PVC films, sheet, injection moldings, floor tiles, and wall coverings, as well as for pipe extrusions and siding containing high levels of TiO_2.

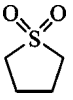
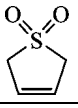
Other applications for mercaptoacids and esters include the following. Mercaptoacetic acid serves as a cocatalyst with strong mineral acid or sulfonic acid resins for the condensation of phenol and acetone in the manufacture of bisphenol A. The acid and its ammonium and ethanolamine salts are components in descaling compositions for removal of iron oxides. Mercaptoacetic acid is used in oil drilling where it acts as an iron sequesterant and an acidizing agent. In concrete and cement applications, polymeric melamines and mercaptoacetic acid have been combined in a new class of ''superplasticizers'' that impart high fluidity, reduced water content, and improved mechanical properties. In this application, it is replacing sulfananilic acid. Alkanolamine salts of the dithiodipropionic acid are available as extreme pressure lubricants. Similar salts of dithiodiglycolic acid are also available.

SULFOXIDES AND SULFONES

Chemical and Physical Properties

The physical properties of the predominant alkyl sulfoxides and sulfones of commerce are listed in Table 2.

Table 2 Physical properties of selected sulfoxides and sulfones

Name	Formula	MW	Melting point (°C)	Boiling point (°C/torr)	CAS Reg. No.
Dimethyl sulfoxide	CH_3SOCH_3	78.1	18.6	189.0/760	67-68-5
Dimethyl sulfone	$CH_3SO_2CH_3$	94.1	108–110	238/760, sublimes 90–100/13	67-71-0
Sulfolane		120.2	28.5	287.3/760	126-33-0
3-Sulfolene		118.2	65	>110/760 (dec.)	77-79-2

Of these, dimethyl sulfoxide (DMSO) and sulfolane (tetramethylene sulfone) represent the largest volumes.

The physical, chemical, and solvent properties of DMSO and sulfolane are well documented.[3–5,6] Dimethyl sulfoxide decomposes only slowly at its normal boiling point (189°C) in air, exhibiting only 2% decomposition to $(CH_3)_2S$, $(CH_3)_2SO_2$, CH_3SH, CH_3SSCH_3, and $(CH_3S)_2CH_2$. Its decomposition (vide infra) is accelerated by acids, anhydrides, acid chlorides, amides, and glycols. Dimethyl sulfoxide has many applications as a solvent and reactant (vide infra).

Sulfones possess extraordinary thermal and chemical stability. Sulfolane is thermally stable to 220°C decomposing to SO_2 and polymeric materials. Dimethyl sulfone decomposes by C–S fission at 510–640°C at 0.7 torr. Sulfones are generally nonreactive materials except under extreme conditions. Nonetheless, sulfones with multiple chlorine or fluorine substituents can be hydrolyzed to sulfonic acids or cleaved by halogen to afford the sulfonyl halides.

Manufacturing Technology

While a number of synthetic routes are available to various sulfoxides, the primary methods for commercial production of DMSO involve oxidation of dimethyl sulfide by oxides of nitrogen or by air in the presence of NO_x catalyst.[3–5] Dimethyl sulfoxide is both the product and the reaction solvent. To alleviate the potential for exothermic, and potentially explosive, runaway reactions in these oxidations, the feed rate for dimethyl sulfide is adjusted to ensure complete conversion and, thus, low instantaneous concentrations at any time. Alternate oxidants for the conversion of sulfides to sulfoxides include nitric acid, H_2O_2/acetic acid, peracids, and halogen/water.[7]

$$(CH_3)_2S + NO_2 \longrightarrow (CH_3)_2SO + NO \quad (7)$$

$$(CH_3)_2S + \tfrac{1}{2}O_2 \xrightarrow{\text{cat.NO}_X} (CH_3)_2SO \quad (8)$$

In all oxidations of sulfides to sulfoxides, yield loss because of overoxidation to form the sulfones is common unless temperatures are low. Another source of yield loss in aliphatic sulfoxides with α-hydrogens arises from the Pummerer reactions, which occur at temperatures greater than 80°C under acidic conditions or in the presence of acid anhydrides/halides. The Pummerer reactions are quite general, leading to products of the general formula $RSCR'_2Z$, with Z being acetate, chloride, hydroxide, etc.

Because sulfoxide sulfur is in an intermediate oxidation state, formally +4, it undergoes a small degree of thermal disproportionation to the corresponding sulfone and sulfide, at elevated temperatures near or above the normal boiling point. The formation of dimethyl sulfide is in part responsible for the garlic-like odor in aged DMSO, and can be minimized by storage under air.

Ozone, H_2O_2, HNO_3, and O_2 with NO_x catalysts are all effective oxidants for converting sulfides to sulfones, albeit much more slowly and at higher temperatures than were used to make sulfoxides. Alkyl aryl sulfones are typically produced by the Friedel–Crafts reaction of an aromatic hydrocarbon with an alkanesulfonyl halide or anhydride. Symmetrical aromatic sulfones are generally prepared by the sulfonation of aromatic hydrocarbon with SO_3, H_2SO_4/SO_3, Cl_2SO_2, $ClSO_3H$, etc.[3–5] The corresponding aromatic sulfonic acids or sulfonyl chlorides are coproducts of the latter reaction.

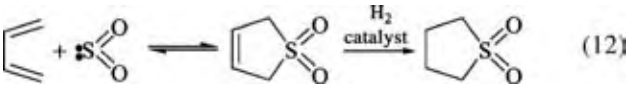
$$R_2SO \xrightarrow{[O]} R_2SO_2 \quad (9)$$

$$RSO_2X + ArH \xrightarrow[X = Cl,Br,OSO_2R]{\text{Lewis Acid Catalyst}} RSO_2Ar + HX \quad (10)$$

$$R_2SO \xrightarrow{\Delta} R_2SO_2 + R_2S \quad (11)$$

3-Sulfolene is produced from butadiene and SO_2, via classical cycloaddition chemistry, with hydrogenation

of the sulfolene affording the sulfolane. Sulfolane is available as a crystalline anhydrous material, or with 3% added water as a freezing point depressant.



(12)

Applications

Both DMSO and sulfolane are extensively used in the chemical, pharmaceutical, polymer, and electronics industries as polar aprotic solvents, with unique properties such as a high dielectric constant, high polarity, and high miscibility with organic and aqueous materials. For example, sulfolane finds use in the refining industry for the separation of benzene, toluene, and xylene (BTX) fractions from paraffins.

The use of DMSO as a reactant and as a solvent has been discussed in detail. Because of its highly polarized S^{+}–O^{-} bond, DMSO is a strong solvator of water, inorganic salts, and most moderately to highly polar organic molecules. Because of its ready permeation through the skin and strong solvating properties, it is used as a carrier solvent for transdermal delivery of pharmaceuticals. As a reaction solvent, DMSO serves to activate NH, OH, and CH functional groups in a variety of reaction chemistries. Dimethyl sulfoxide oxidizes aromatic thiols to disulfides at room temperature.

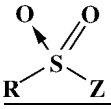
SULFONIC ACID AND DERIVATIVES

Chemical and Physical Properties

Sulfonic acids and their most common derivatives can be represented by the general formula $RSO_2 - Z$, with Z defined as shown below (R and R′ = substituted or unsubstituted aliphatic, aromatic or heterocyclic groups, and M = metal, ammonium, phosphonium).

Table 3 Estimated acidities for sulfonic acids and inorganic mineral acids

	pK_a in H_2O		pK_a in H_2O
CH_3SO_3H	−1.9	HNO_3	−1.5
$CH_3CH_2SO_3H$	−1.7	HCl	<−2.0
$CH_3CH_2CH_2SO_3H$	−1.5	$HOSO_3H$	−3.0
p-$CH_3C_6H_5SO_3H$	−2.7	CH_3OSO_3H	−3.5
$C_6H_5SO_3H$	−2.8	FSO_3H	−5.6
$HO_3SCH_2SO_3H$	−3.0		
p-$O_2NC_6H_5SO_3H$	−3.8		
CF_3SO_3H	−5.5		
p-$H_3N^+C_6H_4SO_3^-$	3.2	HEPES	7.5
$H_3N^+CH_2CH_2SO_3^-$	4.9	MES	6.1
		MOPS	7.2



Z	
Sulfonic acid	OH
Sulfonate salts	OM
Sulfonate esters and thiolsulfonates	OR′, SR′
Sulfonic anhydride	OSO_2R', OC(O)R
Sulfonyl halide	F, Cl, Br, I
Sulfonamides	NH_2, NHR′, NR'_2, NHM, NR′M
Sulfonimides	NH-SO_2R′, NR′-SO_2R′, NH-C(O)R′, NR-C(O)R′, $NMSO_2R'$, NMC(O)R′

As a class, sulfonic acids (RSO_3H) are the strongest uncharged organic acids. Their acidities are similar to those for the strong mineral acids. Selected physical properties for these materials are given in Tables 3, 4A and 4B. Selected physical properties for the alkanesulfonyl halides and anhydrides in common commercial and laboratory use are provided in Table 5.

Methanesulfonyl chloride is the largest-volume alkanesulfonyl chloride. Benzenesulfonyl chloride is the only aromatic sulfonyl halide with significant commercial production, primarily as a feedstock for the manufacture of *N*-butyl benzenesulfonamide. All sulfonyl chlorides are poorly water soluble, which limits their hydrolysis except at elevated temperatures or in the presence of a homogenizing agent such as a cosolvent, surfactant or phase-transfer agent.

The common commercial or laboratory alkanesulfonamides are listed in Table 6. Of these, methanesulfonamide, *N*-methyl methanesulfonamide, *N*-butyl benzenesulfonamide, and trifluoromethanesulfonimide are in commercial production, the latter in the form of its lithium salt. Measured pK_a values are available for a variety of aromatic and akanesulfonamides and sulfonamides along with a discussion on substituent effects.[3–5]

Manufacturing Technology

Alkanesulfonic acids and alkanesulfonyl chlorides from thiols and disulfides

Methanesulfonic acid and methanesulfonyl chloride (MSC) are the alkanesulfonic acids and alkanesulfonyl chlorides produced in the largest volumes.

Table 4A Physical properties of selected aliphatic and aromatic sulfonic acids

Name	Formula	MW	Melting point (°C)	Boiling point (°C/torr)	CAS Reg. No.
MSA	CH_3SO_3H	96.10	19 11 ($\cdot 1H_2O$) −52 ($\cdot 3H_2O$)	167/10	75-75-2
Methanedisulfonic acid	$HO_3SCH_2SO_3H$	176.2	96–100		503-40-2
Ethanesulfonic acid	$CH_3CH_2SO_3H$	110.1	−17 4.3 ($\cdot 1H_2O$)	123/1	594-45-6
Propane-1-sulfonic acid	$CH_3(CH_2)_2SO_3H$	124.2	7.5	130/3	5284-66-2
Butane-1-sulfonic acid	$CH_3(CH_2)_3SO_3H$	138.2	−15.2	149/1	2386-47-2
2-Methylpropane-2-sulfonic acid	$(CH_3)_3CSO_3H$	144.2	113		16794-13-1
Pentane-1-sulfonic acid	$CH_3(CH_2)_4SO_3H$	152.2	15.9	163/1.3	35452-30-3
Hexane-1-sulfonic acid	$CH_3(CH_2)_5SO_3H$	166.2	16.1	174/1.3	13595-73-8
Octane-1-sulfonic acid	$CH_3(CH_2)_7SO_3H$	194.3			394-72-7
Sulfoacetic acid	$HO_2CCH_2SO_3H$	140.1	84–88	245 (dec.)	123-43-3
L-(−)-Camphorsulfonic acid		232.3	198 (dec.)		35963-20-3
D-(+)-Camphorsulfonic acid		232.3	196–200 (dec.)		3144-16-9
Phenylmethanesulfonic acid	$C_6H_5CH_2SO_3H$	172.2			100-87-8
Benzenesulfonic acid	$C_6H_5SO_3H$	158.2	65–66 45–46 ($\cdot 1H_2O$) 43–44 ($\cdot 1.5H_2O$)	172/0.1	98-11-3
o-Toluenesulfonic acid	*o*-$CH_3C_6H_4SO_3H$	172.2	62.1 ($\cdot 2H_2O$)		88-20-0
m-Toluenesulfonic acid	*m*-$CH_3C_6H_4SO_3H$	172.2			617-97-0
PTSA	*p*-$CH_3C_6H_4SO_3H$	172.2	106–107 38 (metastable) 106 ($\cdot 1H_2O$) 93 ($\cdot 3H_2O$)	140/20	104-15-4
m-Xylene-*m*-sulfonic acid	$(CH_3)_2C_6H_3SO_3H$	188.2			18023-22-8
n-Dodecylbenzenesulfonic acid (mixed isomers)	$C_{12}H_{23}C_6H_4SO_3H$	326.5		>204/1	27176-87-0
p-Nitrobenzenesulfonic acid	*p*-$O_2NC_6H_4SO_3H$	203.2			138-42-1
Naphthalene-1-sulfonic acid	α-$C_{10}H_7SO_3H$	208.2	90 (dehy.)		85-47-2
Naphthalene-2-sulfonic acid	β-$C_{10}H_7SO_3H$	208.2	124–125 ($\cdot 1H_2O$) 83 ($\cdot 3H_2O$)		120-18-3
Phenol-2-sulfonic acid	*o*-$HOC_6H_4SO_3H$	174.2	50 ($\cdot \frac{1}{2} H_2O$)		609-46-1
Phenol-3-sulfonic acid	*m*-$HOC_6H_4SO_3H$	174.2			585-38-6
Phenol-4-sulfonic acid	*p*-$HOC_6H_4SO_3H$	174.2			98-67-9

Table 4B Physical properties of selected sulfonic acids

Name	Formula	MW	Melting point (°C)	Boiling point (°C/torr)	CAS Reg No.
Fluorinated organic sulfonic acids					
Trifluoromethanesulfonic acid	CF_3SO_3H	150.1	−40 34 (·1H_2O)	162/760 217/760 (·1H_2O)	1493-13-6
Pentafluoroethanesulfonic acid	$CF_3CF_2SO_3H$	200.1		178/760	354-88-1
Heptafluoropropane-1-sulfonic acid	$CF_3(CF_2)_2SO_3H$	150.1		196/760	423-41-6
Nonafluorobutane-1-sulfonic acid	$CF_3(CF_2)_3SO_3H$	300.1		210–212/760	375-73-5
Undecafluoropentane-1-sulfonic acid	$CF_3(CF_2)_4SO_3H$	350.1		224–226/760	2706-91-4
Tridecafluorohexane-1-sulfonic acid	$CF_3(CF_2)_5SO_3H$	400.1		238–239/760	355-46-4
Heptadecafluorooctane-1-sulfonic acid	$CF_3(CF_2)_7SO_3H$	500.1		120/3	1763-23-1
Aminosulfonic acids					
Sulfanilic acid	p-$H_3N^+C_6H_4SO_3^-$	173.2	288		121-57-3
Taurine	$H_3N^+CH_2CH_2SO_3^-$	125.1	317 (dec)		107-35-7

Methanesulfonic acid is primarily manufactured by the batch or continuous oxidation of the methyl mercaptan or dimethyl disulfide with chlorine in saturated aqueous hydrochloric acid.[3–5,8,9] This chemistry is also the basis for much of the worldwide production of MSC, with photochlorination of methane (vide infra) being the most significant commercial alternative. Other alkanesulfonyl chlorides and sulfonic acids have also been produced in lesser quantities by the Cl_2/H_2O oxidation. Reaction yields are typically 92–100%. The HCl by-product separates as vapor from the saturated reaction mixture. It is absorbed into water to afford a concentrated hydrochloric acid coproduct.

$$RSH + 3Cl_2 + 2H_2O \xrightarrow[\text{aqueous HCl}]{10-35^\circ C} RSO_2Cl + 5HCl \quad (13)$$

$$RSO_2Cl + H_2O \xrightarrow[\text{aqueous HCl}]{>85^\circ C} RSO_3H + HCl \quad (14)$$

When the desired product is the alkanesulfonyl chloride RSO_2Cl, reaction, temperatures are maintained at 10–35°C to inhibit hydrolysis of the poorly soluble sulfonyl chlorides. Extraction with clean

Table 5 Physical properties of selected sulfonyl halides and anhydrides

Name	Formula	MW	Melting point (°C)	Boiling point (°C/torr)	CAS Reg. No.
Methanesulfonyl anhydride	$(CH_3SO_2)_2O$	174.2	69–70	125/4	7143-01-3
Methanesulfonyl fluoride	CH_3SO_2F	98.10		124/760	558-25-8
Methanesulfonyl bromide	CH_3SO_2Br	159.0			41138-92-5
MSC	CH_3SO_2Cl	114.6	−33	161/730 46–47/9	124-63-0
Ethanesulfonyl chloride	$CH_3CH_2SO_2Cl$	128.6		51/4	594-44-5
1-Propanesulfonyl chloride	$CH_3(CH_2)_2SO_2Cl$	142.6	−44	50–51/4 75/13	10147-36-1
Isopropanesulfonyl chloride	$(CH_3)_2CHSO_2Cl$	142.6	−47	59–61/11	10147-37-2
1-Butanesulfonyl chloride	$CH_3(CH_2)_3SO_2Cl$	156.6		81/9	2386-60-9
tert-Butanesulfonyl chloride	$(CH_3)_3CSO_2Cl$	156.6	95–95.5	80/15	10490-22-9
Octanesulfonyl chloride	$CH_3(CH_2)_7SO_2Cl$	212.7	14.6	122/4	7795-95-1
Octanesulfonyl fluoride	$CH_3(CH_2)_7SO_2F$	196.3			40630-63-5
Fluorinated organic sulfonyl halides and anhydrides					
Trifluoromethanesulfonyl anhydride	$(CF_3SO_2)_2O$	282.1		81–83/745	358-23-6
Trifluoromethanesulfonyl chloride	CF_3SO_2Cl	168.5		29–32.9/760	421-83-0
Trifluoromethanesulfonyl fluoride	CF_3SO_2F	152.1		−23/760	335-05-7
Nonafluorobutanesulfonyl fluoride	$CF_3(CF_2)_3SO_2F$	302.1		65/760	375-72-4
Perfluorooctanesulfonyl fluoride	$CF_3(CF_2)_7SO_2F$	502.1		154–155/760	307-35-7

Table 6 Physical properties of selected sulfonamides

Name	Formula	MW	Melting point (°C)	Boiling point (°C/torr)	CAS Reg No.
Methanesulfonamide	$CH_3SO_2NH_2$	95.1	92		3144-09-0
Methanesulfonimide	$(CH_3SO_2)NH$	173.2	155		5347-82-0
N-Methyl methanesulfonamide	$CH_3SO_2NHCH_3$	109.2	3	153/15	1184-85-6
N-Phenyl methanesulfonamide	$CH_3SO_2NHC_6H_5$	171.2	93–97		1197-22-4
Ethanesulfonamide	$CH_3CH_2SO_2NH_2$	109.2	62		1520-70-3
Propane-1-sulfonamide	$CH_3(CH_2)_2SO_2NH_2$	123.2	56–58	105–106/0.1	24243-71-8
Propane-2-sulfonamide	$(CH_3)_2CH_2SO_2NH_2$	123.2	67.5		
Benzenesulfonamide	$C_6H_5SO_2NH_2$	157.2	150–152		98-10-2
N-Butyl benzenesulfonamide	$C_6H_5SO_2NHC_4H_9$	213.3		314/760	3622-84-2
Trifluoromethanesulfonamide	$CF_3SO_2NH_2$	149.1	117–119		421-85-2
N-Trifluoromethanesulfonyl trifluoromethanesulfonimide	$(CF_3SO_2)_2NH$	281.1	46–57		82113-65-3
N-Phenyl trifluoromethanesulfonamide	$CF_3SO_2NHC_6H_5$	225.2	67		456-64-4
Nonafluorobutanesulfonamide	$CF_3(CF_2)_3SO_2NH_2$	299.1			30334-69-1
Perfluorooctanesulfonamide	$CF_3(CF_2)_7SO_2NH_2$	499.2			754-91-6

concentrated HCl removes any H_2SO_4 or RSO_3H impurities, while residual HCl and water are removed by evaporation under vacuum. Alternately, the sulfonyl chloride can be vacuum distilled to separate the impurities.

The alkanesulfonic acids are produced at elevated temperatures to ensure complete hydrolysis of the intermediate RSO_2Cl, with residual HCl being stripped from the reaction product. Methanesulfonic acid and other alkanesulfonic acids are typically obtained as 60–70% aqueous solutions. Similarly, isolated alkanesulfonyl chlorides can be hydrolyzed to afford the aqueous or anhydrous sulfonic acids.[10] Anhydrous nonfluorinated alkanesulfonic acids can also be prepared by evaporative removal of water from the aqueous acids.[11]

Nitric acid has also been shown to be an effective oxidant for converting thiols or disulfides to the corresponding sulfonic acids.[12,13] When this reaction is performed in the presence of HCl at low temperature, the sulfonyl chlorides can also be obtained. The use of nitric acid oxidant has traditionally resulted in higher levels of overoxidation, affording H_2SO_4 as an impurity. In electroplating applications, the major market for MSA, the specification for H_2SO_4 is typically less than 150 ppm and occasionally much lower. In one recently launched commercial MSA process involving nitric acid oxidant, the H_2SO_4 levels in the product were minimized by performing the reaction in two sequential reactors, the first operating at 80–120°C where the bulk of the conversion occurs, the second at 130–150°C. Residual HNO_3 and NO_x were removed by evaporative stripping. Distillation of the MSA then afforded the final product with low impurities. In this process the HNO_3 was regenerated from the nitric oxide by-product by reaction with air. Similarly, the use of atmospheric O_2, catalytic HNO_3/NO_x, and Br^-/Br_2cocatalyst to oxidize aliphatic thiols and disulfides to alkanesulfonic acids and alkanesulfonyl chlorides has also been described.[14]

$$\frac{1}{2}\mathrm{RSSR} + \frac{5}{3}\mathrm{HNO_3} \xrightarrow[\mathrm{H_2O}]{80-150^\circ\mathrm{C}} \mathrm{RSO_3H} + \frac{5}{3}\mathrm{NO} + \frac{1}{3}\mathrm{H_2O} \qquad (15)$$

Alkanesulfonic acids can be prepared from the thiols or disulfides and air using cocatalytic DMSO and HBr.[15] The use of aqueous DMSO in the absence of air, but in the presence of a halogen or hydrogen halide catalyst, readily converts most aliphatic, aromatic, and heterocyclic thiols or disulfides to the corresponding sulfonic acid.[16,17] In effect, DMSO oxidizes the hydrohalide to the molecular halogen, which then reacts with the organosulfur substrates. Water serves as a proton and oxygen source, and inhibits the Pummerer-type decomposition of the DMSO.

Other oxidants for the conversion of alkanethiols and disulfide alkanesulfonic acids have been described. H_2O_2, in combination with catalysts, e.g., peroxycarbonates, alkanesulfonic acids, molybdates, tungstates, or HCl, was effective in producing alkanesulfonic acids.[18,19] At lower temperatures (25–35°C) the combination of H_2O_2 catalytic HCl afforded the alkanesulfonyl chlorides. Kinetic studies of the photooxidation with O_2 in acetonitrile and other solvents have been reported and the relative stabilities of the various

oxysulfur intermediates have been evaluated by ab initio methods.[20,21] Electrolytic oxidation of alkanethiols and disulfides has also been described.[22]

$$RSH + 3(CH_3)_2SO \xrightarrow[\text{catalyst HX or } X_2]{100-130^\circ\ H_2O} RSO_3H + 3(CH_3)_2S \quad (16)$$

$$2HX + (CH_3)_2SO \longrightarrow X_2 + H_2O + (CH_3)_2S \quad (17)$$

Alkanesulfonic acids and alkanesulfonyl chlorides from aliphatic hydrocarbons

Sulfoxidation of saturated hydrocarbons with SO_3, or SO_2 with O_2, is an effective but nonselective method for producing alkanesulfonic acids, invariably producing a mixture of alkanesulfonic acid products.[23] Despite the poor selectivity, this procedure has been used extensively in the manufacture of long-chain alkanesulfonate salts as surfactants since the 1940s. Methane is a particularly nonreactive hydrocarbon and early efforts to sulfoxidate CH_4 using mercury salt catalysts afforded mixtures of MSA, methanedisulfonic acid, and methyl esters.[24] More recent efforts have demonstrated highly selective sulfoxidations of methane with SO_3, or air and SO_2, to afford MSA, as well as a related sulfochlorination of methane with SO_2Cl_2 to make MSC.[25–28]

Photochemical sulfochlorination of saturated hydrocarbons has been used in the production of long-chain alkanesulfonyl chlorides, again as precursors to the salts as surfactants. These processes involve free radical intermediates. Analogous to the sulfoxidation of long-chain alkanes, these processes also provide mixtures of alkanesulfonyl chloride and chloroalkane products.[29] Nonetheless, pure products can be obtained for short C_1–C_3 alkyl chain lengths. A high-pressure gas phase process for the reaction of CH_4 with Cl_2 and excess SO_2 is used in one commercial manufacturing route to MSC.[30]

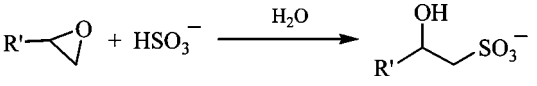
$$CH_4 + Cl_2 + SO_2 \xrightarrow[\text{gas phase}]{h\nu\ 32^\circ C} CH_3SO_2Cl + HCl \quad (18)$$

Alkanesulfonate salts from sulfite and alkyl halides, alkyl sulfate, olefins, or epoxides

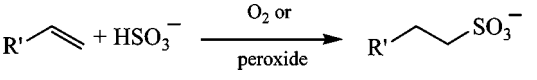
Alkyl halides, alkyl sulfonates, and alkyl sulfates undergo nucleophilic displacement by aqueous sulfite to afford sulfonic acid salts under very mild conditions.[3–5] This chemistry is traditionally referred to as the Strecker reaction. It is currently used in the commercial production of disodium methanedisulfonic acid from CH_2Cl_2. Similarly, metal hydrogen sulfite also reacts with epoxides to afford 2-hydroxyethanesulfonates, the basis for the production of sodium isethionate (sodium 2-hydroxyethanesulfonate) used in the manufacture of surfactants. Free-radical addition of metal hydrogen sulfites to alkenes or alkynes is an alternate route to 2-alkyl- or 2-aryl-substituted ethanesulfonates or to geminal disulfonates.

$$RX + SO_3^{2-} \xrightarrow{H_2O} RSO_3^- + X^- \quad (19)$$

R = alkyl, X = Cl, Br, OSO_3R, OSO_2R



(20)

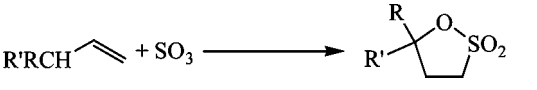


(21)

R' = H, alkyl, aryl

Sultones from olefins

Olefins react with SO_3 under free-radical conditions to afford cyclic sulfonate esters, i.e., sultones.[3–5] The initially formed products are believed to be the highly strained 1,2-adducts, which then rearrange to the 1,3-adducts. This is the commercial route to propanesultone.

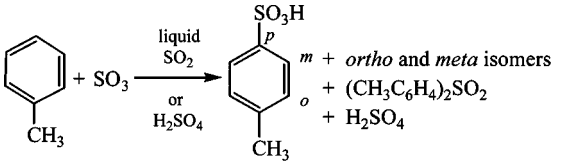


(22)

Aromatic sulfonic acids from aromatic hydrocarbons

The traditional manufacturing processes for making aromatic sulfonic acids involve sulfonation of an aromatic hydrocarbon (or polymers) by falling film technology using vaporized SO_3 in air. Alternate sulfonating agents include oleum ($H_2SO_4 + SO_3$), $ClSO_3H$, SO_3 in solvents, and SO_2 with air. A complete description is beyond the scope of this entry, and for a more thorough overview interested readers should see Ref.[31]. The industrial preparation of PTSA illustrates one approach. Toluene is reacted with SO_3 in liquid SO_2 in a falling film reactor.[32] The final reaction mixture is treated with water to separate unreacted toluene and convert residual SO_3 to H_2SO_4. The PTSA

product is then isolated by crystallization and drying.

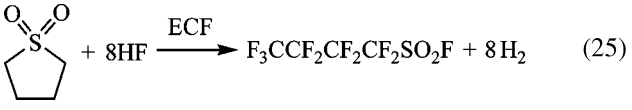


PTSA (23)

Fluorinated alkanesulfonic acids

Perfluoroalkanesulfonyl fluorides and of related materials are manufactured by the electrochemical fluorination of the corresponding alkanesulfonyl fluorides or cyclic sulfones.[3–5] This electrochemical synthesis results in replacement of all C–H bonds in the feedstock. The perfluorinated alkanesulfonyl fluoride products are neutralized to make the anhydrous salt, then acidified and distilled to afford the anhydrous sulfonic acid. Alkyl chain degradation in the electrochemical cell becomes more pronounced at longer chain lengths. Addition of short-chain divalent sulfur compounds (e.g., thiols, sulfides) to the cell inhibits buildup of tarry materials and loss of efficiency.

$$C_8H_{17}SO_2F + 17HF \xrightarrow{ECF} C_8F_{17}SO_2F + 17H_2 \quad (24)$$



(25)

Sulfonamides, sulfonimides, and sulfonate esters from sulfonyl halides

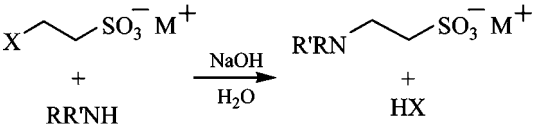
Aliphatic and aromatic sulfonamides are prepared from the corresponding sulfonyl halide (RSO_2X) and ammonia, 1°- or 2°-amines.[3–5] The reaction requires alkaline conditions and thus a stoichiometric amount of additional base to neutralize the by-product hydrohalide (see below Eq. (27)). This base may be a second molecule of the reactive amine, a 3°-amine, or an inorganic base. The reaction can be performed in organic or aqueous solvents, although excessive temperatures must be avoided in water to prevent hydrolysis of the sulfonyl chloride. The sulfonamide products are very polar materials and separation from the by-product salts is invariably a key processing issue. The general approach is to choose a reaction solvent with good solubility for the sulfonamide and poor solubility for the by-product salt, thus permitting removal of the salt by filtration. Such solvents have included nitroalkanes, toluene, chlorobenzene, tetrahydrofuran, other ethers, and nitriles.[33] Running the reaction in water and recovery of the alkanesulfonamide by extraction have also been demonstrated.[34] Alternately, the separation of aqueous mixtures of alkanesulfonamides and ammonium chloride by electrodialysis is also feasible.[35] Alkanesulfonate esters are similarly prepared from the alkanesulfonyl chloride, the appropriate alcohol, and base.[36]

Symmetrical *N*-sulfonyl sulfonamides, i.e., sulfonimides, can be prepared from the amine and two equivalents of sulfonyl chloride and base. The unsymmetrical *N*-sulfonyl sulfonamides and the related *N*-acyl sulfonamides can be obtained by the reaction of a sulfonamide or carboxamide with an acid halide.

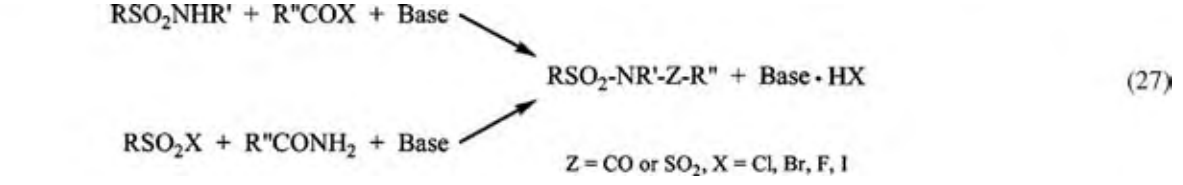
$$2RSO_2X + R'NH_2 + 2\ \text{Base} \longrightarrow (RSO_2)NR' + 2\ \text{Base} \cdot HX \quad (26)$$

Aminoalkanesulfonates

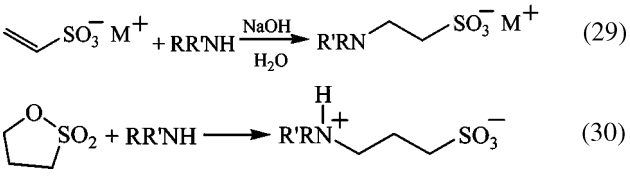
The 2-aminoethanesulfonates (i.e., taurines) are the prepared reaction of ammonia or 1°- or 2°-amines with metal 2-haloalkanesulfonate, vinylsulfonate, or hydroxyethanesulfonate salts under alkaline conditions.[37,38] The aminoethanesulfonate salts are then converted to the zwitterionic acids by cation exchange or acidification. 3-Aminopropanesulfonic acids are directly prepared by the reaction of propanesultone with the amine.



X = OH or halide (28)



(27)



(29)

(30)

Interconversion of Sulfonic Acid Derivatives

A variety of procedures have been identified for the interconversion of sulfonyl derivatives. Because of their high acidity, sulfonic acids can readily be neutralized with inorganic bases or organic amines to form salts. Conversion of metal salts back to the acid form is often desirable in synthetic or electroplating applications. Direct acidification of alkali metal sulfonates with HCl affords the sulfonic acid with precipitation of the formed metal halide salt. An alternative approach is to convert the salt back to the acid by electrodialysis.[39] Recovery and recycle of MSA from spent metal plating baths have been described.[36]

Sulfonic acids and alkali metal salts can be readily converted to the sulfonyl chlorides using a variety of reagents, including thionyl chloride, phosgene, phosphorous pentahalides, or phosphorous oxychloride. Sulfonyl bromides can be obtained from the corresponding chlorides by sequential reaction with alkaline Na_2SO_3 and Br_2. Sulfonyl fluorides are obtained from sulfonyl chlorides by reaction with aqueous KF. Aromatic sulfonyl chlorides can be obtained from the sulfonamides by reaction with PCl_5, but this is not applicable to the aliphatic analogs.

Applications

Sulfonic acids and their derivatives are used in innumerable industrial applications in chemical synthesis, electroplating of metals, surfactants, ion-exchange resins, and preparation of dyes, animal feeds, pesticides, and pharmaceuticals.

Methane sulfonic acid is used as an electrolyte for electroplating of tin onto sheet steel, for plating tin and tin/lead alloy onto nickel or other base metal substrates in the manufacture of lead frames and bump-contacts for microelectronic devices.[36] It can also be used for copper deposition during the manufacture of microprocessors. Other alkanesulfonic acids have also found use in electroplating applications. Disodium methanedisulfonate and other alkanedisulfonate salts are used in chrome plating.[40] As discussed previously, several processes for the recovery and recycle of alkanesulfonic acids from spent metal plating baths have been described.

Because of their high acidity, sulfonic acids are extensively used as Bronsted acid catalysts for esterification, aromatic alkylation, and etherification. The primary sulfonic acids in these applications include MSA, PTSA, trifluoromethanesulfonic acid (triflic acid), and nonafluorobutanesulfonic acid (nonaflic acid), as well as the aromatic and fluorinated alkanesulfonic acid resins. In these applications, they compete against the traditional strong inorganic acids such as H_2SO_4, HCl, or HF. Triflic acid and MSA are both liquids and are easily handled in a manufacturing environment. This lends these materials to use as both acid catalyst and reaction solvent.[41] Methane sulfonic acid bound to poly(vinylpyridine) has been shown to be an effective esterification catalyst that can be readily recovered and recycled.[42] Because of its much higher cost, the commercial uses of triflic acid are generally limited to reactions where its higher acidity is a requirement, e.g., the potential replacement for HF or H_2SO_4 in paraffin alkylation in petroleum refining. *p*-Toluenesulfonic acid is a solid at ambient temperatures and poorly soluble in some organic matrices at low temperature. Thus, it is recyclable by filtration.

In addition to its synthetic utility, MSA is employed as an amine-salt former in pharmaceutical final dosage forms. Other sulfonic acids are also used as pharmaceutical salt formers, but much less frequently.

The major commercial application of aromatic sulfonic acid salts is anionic surfactants.[43] These are predominately linear long-chain alkyl benzene sulfonates and the naphthalene analogs, or α-olefinsulfonates. Short-chain alkylarenesulfonates are used in liquid detergent formulations as coupling agents, solubilizers, and hydrotropes where high concentrations of organic surfactants and inorganic compounds must be kept in aqueous solution. Fatty acid esters of sodium isethionate are mild surfactants unaffected by hard water, but with limited hydrolytic stability. Cocoyl isethionate is the principle ingredient in detergent bars for personal use.

Aromatic sulfonic acid and perfluoroalkanesulfonic acids resins are widely used as ion-exchange resins in water treatment and multiple other industrial applications. In the form of membranes, they are routinely used in electrochemical cells, particularly in electroplating of metals and in battery applications. The lithium salts of trifluoromethanesulfonic acid and *N*-trifluoromethanesulfonyl trifluoromethanesulfonamide are both employed as electrolytes in secondary battery applications.

Taurine is an essential dietary nutrient in felines and is routinely added to packaged food for domestic cats.[3–5] Other aminoalkanesulfonic acids, e.g., *N*-2-hydroxyethylpiperazine-*N*′-2-ethanesulfonic acid (HEPES), 2-[*N*-morpholino]ethanesulfonic acid (MES), or 3-(*n*-morpholino)propanesulfonic acid (MOPS), are commonly referred to as biological buffers or as ''Good's'' buffers.

These are intensively used in fermentation and protein separation.[37,38]

Sulfonyl chlorides have extensive uses in organic synthesis in the preparation of sulfonamides and sulfonate esters. Methanesulfonyl chloride is a key raw material in the synthesis of critical components for photographic color developing formulations, as well as for herbicides and pharmaceuticals.

Methanesulfonamide and *N*-methyl methanesulfonamide are used as chemical intermediates in the manufacture of agrochemical herbicides and fungicides.[44,45] Methyl methanesulfonamide is also used in the synthesis of the anticholesterol agent rosuvastatin.[46] *N*-Butyl benzenesulfonamide serves as a plasticizer for polyamide resins.

CONCLUSIONS

Thiochemicals are among the important intermediates and end-user chemicals, varying from relatively simple molecules, e.g., H_2S, to complex molecules, e.g., polysulfides. Their manufacturing routes vary and depend highly on their properties and quantity of interest. Their intermediate applications include the manufacturing of agricultural products, pharmaceuticals, detergents, lubricants, etc. Some thiochemicals have played important roles in solvents, polymer synthesis, electroplating, presulfiding of hydrotreating catalysts, coke prevention in steam cracking process, etc.

ACKNOWLEDGMENT

The authors acknowledge the support and information provided by Arkema Inc.

REFERENCES

1. Fisch, M.H.; Bacalogulu, R.; Biesiada, K.; Brecker, L.R. Mechanism of organotin stabilization of poly(vinyl chloride). 1. The structure and equilibria of alkyltin alkyl mercaptopropionates and their compatibility with PVC. In *Plastics: Plastics on My Mind*, Proceedings Antec 1998 Brookfield, Society of Plastic Engineering; Vol. 3, 33291.
2. Fisch, M.H.; Bacalogulu, R.; Biesiada, K.; Brecker, L.R. Mechanism of organotin stabilization of poly(vinyl chloride). 2. Significance for PVC stabilization of structure and equilibria of alkyltin alkyl thioglycolates. In *Plastics: Bridging the Millenia*, Proceedings Antec 1999, Society of Plastic Engineering; Vol. 3, 33296.
3. Patai, S., et al. *The Chemistry of Sulfonic Acids, Esters, and Their Derivatives*; Patai, S, Rappoport, Z., Eds.; John Wiley and Sons: New York, 1991.
4. Patai, S., et al. *Supplement S–The Chemistry of Sulfur Containing Functional Groups*; Patai, S., Rappoport, Z., Eds.; John Wiley and Sons: New York, 1993.
5. Patai, S., et al. *The Chemistry of Sulphones and Sulphoxides*; Patai, S., Rappoport, Z., Eds.; John Wiley and Sons: New York, 1988.
6. Wypych G., Ed.; *Knovel Solvents: A Properties Database*; ChemTec Publishing: Toronto, Ontario, 2000.
7. Flick, E.W. Ed.; Organic sulfur compounds. In *Industrial Solvents Handbook*, 5th Ed.; Noyes Data Corp.: Westwood, NJ, 1998; 222–237.
8. Guertin, R.M Method of Preparing Alkyl Sulfonyl Chloride. U.S. Patent 3,626,004, Dec 7, 1971.
9. Guertin, R.M Method of Preparing Alkane Sulfonic Acids. British Patent 1,350,328, Apr 18, 23, 1974.
10. Ollivier, J.; Lagaude, C.H.; Baptiste, H.; Larrouy, M Process for the Production of Alkanesulphonic AcidsU.S. Patent 4,859,373, Aug 22, 1989.
11. Comstock, P.; Keys, K.M Preparation of Anhydrous Alkanesulfonic Acid. U.S. Patent 4,938,846, Jul 3, 1990.
12. Eiermann, M.; Tragut, C.; Ebel, K Method of Producing Alkanesulfonic Acid. U.S. Patent 6,531,629, Mar 11, 2003.
13. Bechtel, M.; Thiel, J Improved Liquid Distributor for Distillation Columns Consisting of Glass. German Patent DE 10139587, Feb 20, 2003; Chem. Abstr. *138*, 155467.
14. Chen, J Preparation of Alkane Sulfonic Acids and Sulfonyl Chlorides by Oxidation of Alkanethiols and Dialkyl Disulfides. U.S. Patent 6,124,497, , Sep 26, 2000.
15. Husain, A.; Wheaton, G.A Production of Alkanesulfonic Acids by Oxidation of Alkanethiols or Dialkyl Disulfides in the Presence of DMSO. European Patent 424,616, Feb 2, 1991; Chem. Abstr. *115*, 8085.
16. Toland, W.G Bromide Ion Promoted Oxidation of Sulfide-Sulfur by Lower Dialkyl Sulfoxides. U.S. Patent 3,428,671, Feb 18, 1969.
17. Lowe, O.G Oxidation of Thiols and Disulfides to Sulfonic Acids. U.S. Patent 3,948,922, Apr 6, 1976.
18. Schreyer, G.; Geiger, F.; Hensel, J A Process for the Production of Alkyl Sulfonic Acids. U.S. Patent 4,239,696, Dec 16, 1980 (and cited references therein).
19. Husain, A.; Wheaton, G.A. Oxidation of Thiols, Disulfides, and Thiolsulfonates. U.S. Patent 4,956,494, Sep 11, 1990.

20. Robert-Banchereau, E.; Lacombe, S.; Ollivier, J.; Micheau, J.C.; Lavabre, D. Kinetic modeling of the photooxidation of dimethyldisulfide in the liquid phase. Int. J. Chem. Kinet. **1997**, *29* (11), 825–834.
21. Lacombe, S.; Loudet, M.; Dargelos, A.; Robert-Banchereau, E. Oxysulfur compounds derived from dimethyl disulfide: an ab-initio study. J. Org. Chem. **1999**, *63*, 2281–2291.
22. Gardner, D.M.; Wheaton, G.A Preparation of Alkanesulfonyl Halides and Alkanesulfonic Acids. U.S. Patent 5,035,777, Jul 30, 1991 (and references therein).
23. Bost, H.W Sulfoxidation Process. U.S. Patent 3,413,337, Nov 26, 1968.
24. Snyder, J.C.; Grosse, A.V Reaction of Methane with Sulfur Trioxide. U.S. Patent 2,493,038, Jan 3, 1950.
25. Basickes, N.; Hogan, T.E.; Sen, A. Radical-initiated functionalization of methane and ethane in fuming sulfuric acid. J. Am. Chem. Soc. **1996**, *118*, 13111.
26. Mukhopadhyay, S.; Bell, A.T. A High-yield approach to the sulfonation of methane to methanesulfonic acid initiated by H_2O_2 and a metal chloride. Angew. Chem. Int. Ed. **2003**, *42*, 2990–2993 (and references cited therein).
27. Mukhopadhyay, S.; Zarella, M.; Bell, A.T.; Srinivas, R.V.; Smith, G.S. Synthesis of methanesulfonyl chloride (msc) from methane and sulfuryl chloride. Chem. Commun. **2004**, 472–473.
28. Richards, A.K Anhydrous Conversion of Methane and Other Alkanes into Methanol and Other Derivatives Using Radical Pathways and Chain Reactions with Minimal Waste Products. World Patent Application 2004/041399A2, May 21, 2004.
29. Hertel, O.; Schlecht, H.; Schneider, R. Production of Substituted Alkanes. U.S. Patent 3,911,004, , Oct 7, 1975.
30. Ollivier, J.; Baptiste, H.; Laqaude, C.H.; Larrouy, M Procédé et Appareil pour la Sulfochloration Photochimique d'Alcanes Gazeux. European Patent Application 194,931, Mar 14, 1985; Chem. Abstr. *106*, 20289.
31. Lindner, O.; Rodefeld, L. Benzenesulfonic acids and their derivatives. In *Ullmanns Encyclopedia of Industrial Chemistry*; John Wiley and Sons: New York, 2002.
32. Wu, J.C.; Wang, B.H.; Zhang D.L. Song, G.F.; Yuan, J.T.; Liu, B.F. Production of *p*-toluenesulfonic acid by sulfonating toluene with gaseous sulfur trioxide. J. Chem. Technol. Biotechnol. **2001**, *76*, 619–623.
33. Smith, G.S.; Cordova, R.; Overgaard, T.H.; Budrick, M.T.; Brown, S.M Preparation of alkanesulfonamides with low residual ammonium impurities, U.S. Patent 5,599,983, Feb 4, 1997 (and references cited therein).
34. Brown, S.M.; Muxworthy, J.P.; Gott, B.D Process for the Production of Sulfonamides. World Patent Application WO 98/25890, Nov 27, 1997 (and citations referenced therein).
35. Gancet, C.; Lauranson, D.; Perie, F Desalination of aqueous sulphonamide solutions. U.S. Patent 6,036,830, Mar 14, 2000.
36. Gernon, M.D.; Wu, M.; Buszta, T.; Janney, P. Environmental benefits of methanesulfonic acid: comparative properties and benefits. Green Chem. **1999**, 127–140 (and references therein).
37. Good, N.E.; Winget, G.D.; Winter, W. Hydrogen ion buffers for biological research. Biochemistry **1974**, *5*, 467–477.
38. Good, N.E.; Izawa, S. Hydrogen ion buffers. In *Photosynthesis*; Methods Enzymol; San Pietro, A., Ed.; 1971; Vol. 24B, 53–68.
39. Gavach, C.; Gancet, C.; Mirassou, A.; Perie, F Regeneration of Acids, Particularly Strong Organic Acids, Using Bipolar Membranes. U.S. Patent 5,993,629, Nov 30, 1999.
40. Newby, K.R Protection of Lead-Containing Anodes During Chromium Electroplating. U.S. Patent 5,176,813, Jan 5, 1993.
41. Kieczykowski, G.R Process for Preparing 4-Amino-1-Hydroxybutylidene-1,1-Bisphosphonic Acid (ABP) or Salts Thereof. U.S. Patent 5,019,651, May 28, 1991.
42. Gancet, C Preparation of Esters of Carboxylic Acids Directly From Carboxylic Acids and Alcohols Using a Catalyst System Comprising a Sulfonic Acid and a Polymer-Bound Tertiary Amine. European Patent 1,167,337, Jan 2, 2002.
43. Lynn, J.L.; Bory, B.H. Surfactants. In *Kirk-Othmer Encyclopedia of Chemical Technology*; 1998 Electronic Release; Wiley–VCH: Weinheim, Germany, 1997.
44. Schlegel, G Preparation of Pyrimidyl-Substituted Sulfonylureas. European Patent Application EP 560178, Sep 15, 1993; Chem. Abstr. *120*, 8613.
45. Madsen, H.B.; Klemmensen, H.K.-A N-Phenyl or N-Pyridyl Sulfamides and Methanesulfonamides. . U.S. Patent 4,148,901, Apr 10, 1979.
46. Hirai, K.; Ishiba, T.; Koike, H.; Watanbe, M Pyrimidine Derivatives. U.S. Patent 5,260,440, Nov 9, 1993.

Tissue Engineering

Shang-Tian Yang
Clayt Robinson
Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio, U.S.A.

INTRODUCTION

Tissue engineering is a combination of biology and engineering for producing biological substitute structures that can reconstitute cellular or tissue function, which is lost, declining, or insufficient. Theorized for a long time, the field has evolved greatly within recent decades to meet the challenges of disease and a limited organ donor supply. For tissue substitutes to be successful in replacing lost function, they must accurately and reliably perform the desired function that is dependent on a multitude of variables. Parameters that must be carefully investigated include the cell source, the polymeric scaffold support, the tissue culture protocol, and the implantation procedure. Tissue engineering research covers a wide range of applications including many tissue substitutes, cell therapy, and diagnostic modeling. Although faced with as yet unmet challenges, tissue engineering stands to provide momentous advancements to health in the 21st century.

First coined in 1987, tissue engineering, as a field, is defined by two main objectives. The first objective is to apply methods and principles of engineering and life sciences to understand the biological tissue construct.[1] In addition to deciphering the vast catalog of cellular functions, the interconnections of a cellular population, regulated by signaling pathways, and the interactions of a cellular composite with the noncellular surrounding environment must be understood. Since the three-dimensional structure and organization of tissue components are integral in defining proper function of the tissue, this environmental arrangement must be mimicked to produce a tissue substitute that functions appropriately. A fundamental understanding of a tissue environment structure and how it enables or affects the tissue function is essential. The second objective is to apply established knowledge of the tissue construct toward developing biological substitutes to restore, maintain, or improve natural function to tissues or organs that are structurally or physiologically altered.[1]

Included in this entry is a review of the development of tissue engineering, from theorized concepts and early experiments through advancements made in developing tissue substitutes in the recent past. A review of biological systems and components precedes a thorough discussion of the components utilized in tissue engineering constructs. The entry concludes with tissue engineering applications and the challenges that remain for full realization of the field's potential.

MOTIVATION

Motivation for further development in the field of tissue engineering is to save the lives of patients suffering from organ failure or loss, who are waiting for donor organs to become available, and those suffering from debilitating diseases, such as Parkinson's disease, in which essential cell or tissue function deteriorates or is lost over time. Further, people are victims to catastrophic events in which tissue repair or replacement is a grave concern with, commonly, no immediate remedy.

As a result, the annual health care costs in the United States alone exceed $400 billion for patients with tissue loss or organ failure. The cost includes 8 million surgical procedures and 40–90 million hospital days annually.[2] In 2002, the waiting list for organ recipients exceeded 81,700 candidates, with over 38,600 people added to the list. However, only 12,800 organs were transplanted in 2002 owing to a severe shortage of donors.[3] With a limited organ supply, accumulating costs, and limited medical procedures available, there is a void to fill.

FROM FICTION TO REALITY

There are several examples of tissue engineering well before the field was named in 1987. Stories that include the replacement of a person's body parts with those from another person are found in literature centuries old. Possibly, the first tale of a tissue engineering procedure is found in the Bible, when Eve is given life from the rib of Adam. In 1818, Mary Shelley wrote *Frankenstein* in which the title character is given life through the compilation of body parts garnered from donor corpses. In the legend of St. Cosmas and St. Damien, about 200 A.D., the two physicians performed a procedure in which the gangrenous leg of a

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120006824

man was successfully replaced with the leg of a recently deceased man.[4] More recently, Hollywood has produced films that portray ideas from tissue engineering. In 1991, the aptly titled *Body Parts* was released, in which a man loses his arm in an accident and receives a transplant from a recently executed inmate. Although successfully transplanted, incompatibility becomes an issue as the transplanted arm retains the murderous personalities of its donor.

The collection of stories that involve tissue engineering concepts shows the promise and spectacular possibilities that the future could bring. However, rarely do these stories fully conceptualize or even mention the challenges involved in performing these acts in reality. Until the advancement of tissue engineering in the 1970s, replacements for bodily tissues were prostheses made of wood, ceramics, and plastics. Replaced body parts included arms, legs, eyes, ears, teeth, and noses.

Experimentation with animals became more common for studying the growth of tissue. In the 1930s, the work done by Bisceglie became one of the earliest documented tissue engineering procedures.[2] Two important concepts were displayed including the transplantation of mouse tumor cells into the abdomen of a pig, while the cells were encased within a polymer membrane structure. First, the results showed that cells could survive within a foreign environment without rejection by the host immune system. Second, the procedure introduced the encapsulation approach. The semi-permeable polymer membrane system allowed for nutrient and waste fluxes into and out of the membrane system, respectively. Concurrently, the membrane selectively rejected the passage of immune system molecules and proteins, thus protecting the enclosed cells. In the work performed by Chick and coresearchers, islet cells were encapsulated within a semi-permeable membrane and transplanted into animal models to provide glucose level control as a cure for diabetes.[2]

In the laboratory setting, the first tissue to be reconstituted was skin because of its relatively simple two-dimensional structure. During the late 1970s and 1980s, artificial skin was created using skin cells distributed within natural collagen or collagen–glycosaminoglycan composite support structures. The growth of tissue engineering in the 1990s and early 2000s is due to interdisciplinary advancements in the fields of engineering, genomics, proteomics, cell biology, and material science.[5] As understanding of the important relationship between tissue structure and function became fully realized, three-dimensional synthetic polymers were utilized to mimic the bodily in vivo environment as a support for cells to attach and grow on. The liver was the first tissue to be cultivated using these three-dimensional constructs owing to its relatively simple composition. Utilizing these newly developed tools, a multitude of tissues have been or are being studied and mimicked with tissue-engineered products for possible future usage in a wide array of applications that is discussed later in this chapter.

CELLULAR SYSTEMS BIOLOGY

For a tissue substitute to function properly, many biological aspects of the tissue and component cells must be understood. Some aspects include the extracellular matrix, cell-specific gene expression and surface markers, cell growth parameters, population arrangement and behavior, and the immune system.

Tissues and Organs

An organ is a component of the human body system made of one or more tissue types and has specific jobs within the complex network. Tissue is a collection of similar cells and the surrounding supportive environment that together perform specific tasks. There are four basic categories of tissues: epithelium, which constitutes surfaces such as skin, connective tissue, muscle tissue, and nerve tissue.[4]

Extracellular Matrix

Tissue in vivo consists of cells that are engaged in an environment, called the extracellular matrix (ECM), that provides support to allow for proper cell function, cell–scaffold interactions, and tissue morphology. The ECM directly affects or controls cell shape, function, viability, and population structure. The ECM that supports the tissue structure and cells in vivo is a complex network of collagens, glycoproteins, such as fibronectin and laminin, hyaluronic acid, proteoglycans, glycosaminoglycans, and elastins to which cells adhere and interact. The three-dimensional ECM has a two-way interaction with the cells. The ECM surface properties and molecules provide cell-surface receptor-mediated signals to influence cellular spatial organization, migration, growth, differentiation, and death.[6] Important cell receptors that interact with the ECM include integrin and cadherin adhesion receptors.[7] The cells influence the ECM by remodeling the structure and secreting new ECM components.[8]

Genes and Proteins

The genes on chromosomes within each cell are the blueprints of the human body. The pattern in which the genes are expressed determines the cell type and behavior. Gene expression, which leads to protein

production, is a highly controlled process dependent on signals originating from other cellular components within the same cell, externally from other cells within the immediate environment, or from distant locations, such as part of the endocrine system. External signals are transduced, or passed, with the aid of surface molecules that are receptors for the signal molecules. Other molecules and proteins are expressed to perform cell-specific functions within the cell, or externally following secretion. The presence of these surface receptors or expressed functional proteins is characteristic of the cell type and can be used as cell markers for identifying proper cell function.

Cell Growth

As cells grow, or proliferate, they proceed through a highly controlled process called the cell cycle. The cell cycle consists of four phases during which the DNA set is accurately replicated and subsequently divided into two complete sets partitioned into two new cells derived from dividing the original cell via mitosis. Regulatory proteins guard progression between phases to ensure cellular readiness and DNA integrity preservation. Errors in this regulatory process can lead to DNA mutations and uncontrolled growth, both characteristics of cancer. With specific signals present, cells exit the cell cycle for a maturation process called differentiation before re-entering the cell cycle. As cells differentiate, their function becomes more defined and limited.

Morphology and Arrangement

Cellular morphology and population arrangement within the tissue culture environment are dependent on the support surface properties. Cells that are highly proliferating appear round and smooth. As cells adhere to a surface, spreading occurs across the surface, and the cells flatten. As population of cells increases, interaction and rearrangement occur among the cells. In a two-dimensional environment, interactions are limited, but in a three-dimensional environment, cellular aggregates can form within the structure of the supporting material. Aggregation and interaction among cells in all directions mimic the body environment and allows for similar cellular function.

Immune System

The immune system defends the body from infection and illness using cellular and molecular components to detect and clear objects from the body that are not identified as normal. When objects such as invading viruses, micro-organisms, implanted materials, and mutated cancerous cells are targeted by the immune system, the resulting immune rejection mechanism kills, disrupts, or encloses the foreign object to prevent further harm to the body. Cells are screened by immune system antibody molecules or T-cells to verify whether the surface molecules are native or foreign. A key determinant of whether a cell is native or foreign is the major histocompatibility complex genes contained within the cell and expressed on the cell surface. Therefore, for cells to be compatible and to avoid immune rejection in a new host, the similarity of the major histocompatibility complex genes must be high. This is important in determining the success of a tissue-engineered product within the host.

TISSUE ENGINEERING CONSTRUCT COMPONENTS

As summarized in Fig. 1, the components that must be customized based on the application include cell source, scaffold parameters, and cell culture procedures.

Cell Sources

The cells utilized for producing a tissue substitute come from multiple sources. When a specific cell type is needed to culture a certain tissue type, the availability of the cell type, the means of obtaining the cells, procedures for maintaining and multiplying the cell population in culture, and immune rejection upon transplantation must all be considered. Avoiding rejection is a huge challenge in the development of a tissue substitute, thus autologous cells obtained from the same person to whom the transplant will be given are ideal as the host body does not reject autologous cells. However, supply of autologous cells is frequently the problem. For example, in many situations, a large enough population of healthy cells is unavailable due to the extent of disease.[9] Further, the harvesting of cells from one location may cause long-term harm to the donor site. Once obtained, autologous cells are cultured, or grown, in a laboratory environment, or in vitro in the presence of a proper support structure and nutrient supply until a larger population is present. The autologous cells are then administered to the donor patient at the necessary site.

When autologous cells are not appropriate or available, the cells can come from either another donor of the same species or a different species. The transplantation of cells between similar or dissimilar species is called allotransplantation or xenotransplantation, respectively. Although the supply of these types of grafts is plentiful, immune rejection is very common, so some additional strategy must be utilized to avoid rejection. Cells provided by family members, or

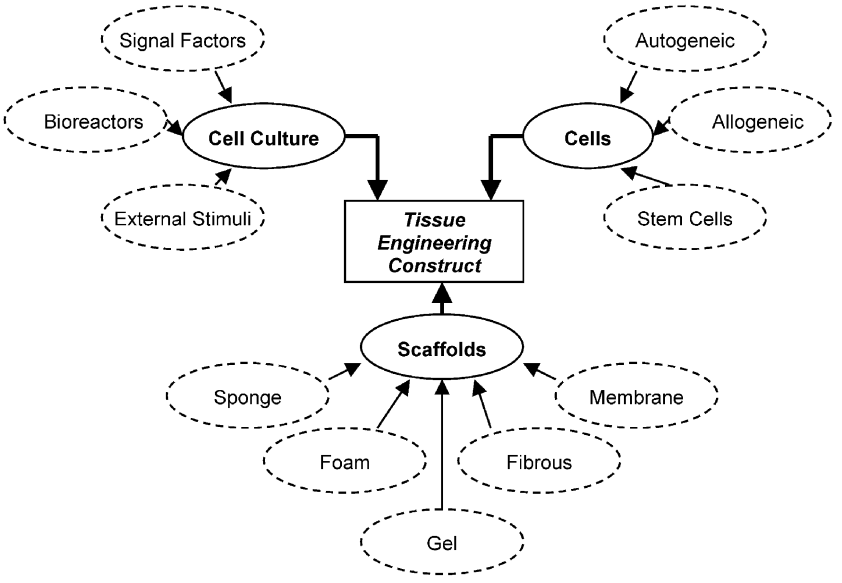


Fig. 1 Overview of the components utilized in tissue engineering constructs.

allogeneic cells, tend to be the most compatible source in order to avoid rejection due to similar gene sets, including, specifically, the histocompatibility complex genes. Xenografts provide an additional challenge since xenogeneic cells may contain components that are infectious when introduced into a human. This fact, in addition to ethical and moral issues regarding utilizing nonhuman parts in a human, has led to general unpopularity of this procedure.[9]

A fourth cell source is stem cells. Stem cells are characterized by the ability to proliferate indefinitely and develop into different cell types, or pluripotent, depending on the stem cell origin and given the appropriate signals. Embryonic stem cells are present as an embryo first begins to develop and differentiate to form all components of the human body. These stem cells allow for generation of any tissue cell type, however, ethics and regulations limit their usage. An adult retains a limited supply of adult stem cells in the bone marrow and in tissues throughout the body. Most of these progenitor cells are partially differentiated into a lineage of cell types, but remain multipotent to develop into a more limited range of cell types. Examples are neural stem cells, hematopoietic stem cells, and mesenchymal stem cells.

Owing to the limited supply of stem cells in the body, specialized techniques are necessary to find and separate them from the mixture. Stem cells may be isolated from an extracted tissue mass by digesting the ECM structure surrounding the cells, and then detecting the cells based on signature biomolecular expression profiles as a ''fingerprint'' for stem cells. Once identified, the stem cells can be separated from the total tissue population by passing the cells through a system that selectively removes the stem cells while allowing the remainder of cells to exit the system separately. Selective retention technology systems extract stem cells from the mixed population by customizing the system to have a high affinity for expressed stem cell surface molecules or to secondary molecules previously bound selectively to the cell surface. Example stem cell surface markers are CD34 and CD45 for hematopoietic stem cells and Oct-4 and SSEA-3 for embryonic stem cells.

Once harvested, stem cells can be cultured in specific biological, chemical, and physical stimuli to differentiate into the cell type of interest. Upon expansion in culture to a large enough population, the cells may be transplanted as therapy. However, strategies to avoid rejection are necessary unless the cells are autologous or cultured in such a way to disguise the fact that the cells are from a different source.[9] The difficulty lies in providing the correct composition, amounts, and timing of the stimuli to direct the differentiation to the desired cell type. A current approach is to direct the transformation of the stem cells in vitro until the cells are within one or two steps of the complete differentiation destination. The final transformation steps are accomplished in vivo, after transplantation to the site of the desired cell type where signals are provided to complete the differentiation.[10]

One concern with stem cell usage is that transplanted cells derived from stem cells can be tumorigenic owing to undifferentiated stem cells present in the population that proliferate uncontrollably.[9] The use of stem cells, however, is promising because of the limited cell supply for many tissues. Cells that are not autogeneous must be able to avoid immune rejection. Somatic cell and nuclear transfer procedures could provide cells that function appropriately and

yet will not be rejected.[9] The production of a universal donor cell source that can be used in any patient while avoiding immune rejection is a future goal. This probably involves altering the expression of or hiding the histocompatibility complex surface markers. Different types of stem cells are also being investigated as a source for multiple cell types if the differentiation can be precisely directed and controlled.

Scaffolds

Studies performed in vitro with cells growing on a two-dimensional surface have observed isolated cell function performance, such as proliferation, glycolysis, respiration, and gene expression, by optimizing the media nutrient, hormone, and growth factor compositions. However, the proper regulation and control of these functions are dependent on cellular interactions present within a three-dimensional structure.[1] Therefore, scaffolds are essential for creating tissue substitutes that mimic in vivo function.

Scaffolds can be foams, sponges, gels, membranes, or fibrous materials (Fig. 2). They are categorized as natural, synthetic, or a combination of both. Table 1 provides a list of scaffold materials and applications utilized in tissue engineering.[11–14] Natural biomaterials, such as collagen, are inherently equipped for cell interaction, but have the disadvantages of limited adaptability and customizable processing as well as relatively scarce availability compared to synthetic biomaterials.[15] The most commonly utilized synthetic biodegradable materials are poly(glycolic acid) (PGA), poly(lactic acid) (PLA), and poly(lactic co-glycolic acid) (PLGA), a blend of the former two polymers. Synthetic polymers can be used for the culture of many cell types, but it remains difficult to culture some cell types, such as nerve cells, on synthetic polymers.

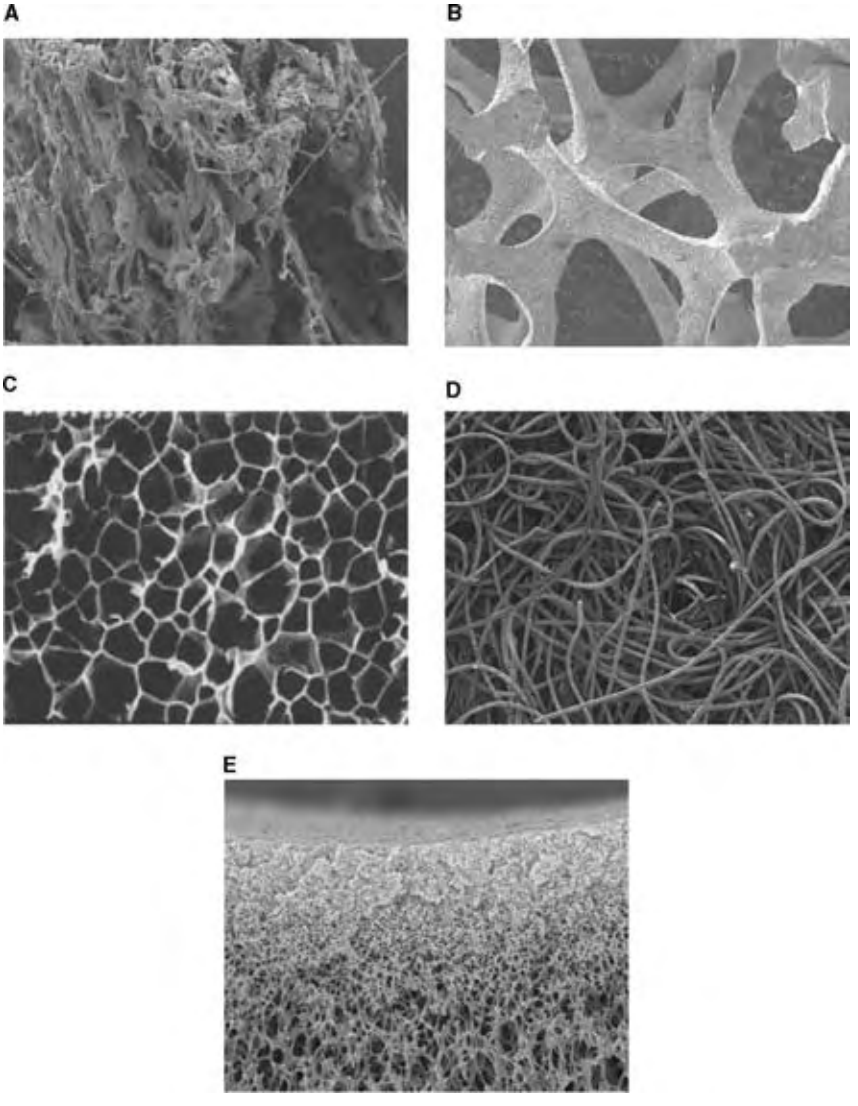


Fig. 2 Five scaffold types: (A) sponge; (B) foam; (C) gel; (D) fibrous; and (E) membrane.

Table 1 Commonly utilized scaffold materials

Polymer	Application	References
Natural		
Type I collagen	Skin, bone, cartilage, tendon, nerve, kidney, cornea, vessels	[7,11,12]
Alginate	Cartilage, muscle, soft tissue	[12]
Chitosan	Encapsulation, membranes	[11]
Fibrin	Cartilage	[12]
Laminin	Epithelial tissues, islets	[7]
Hyaluronic acid	Medical devices	[11]
Polyhydroxyalkanoates (PHA)	Skin, drug delivery, sutures	[11]
Isolated ECM from bone and small intestine	Bone, blood vessels, ureters	[7]
Synthetic		
Poly(esters) Poly(glycolic acid) (PGA) Poly(lactic acid) (PLA) Poly(caprolactone) (PCL) Poly(lactic-co-glycolic) (PLGA)	Cartilage, bone, muscle, nerve, blood vessel, valves, bladder, liver, cardiac tissue, drug delivery, sutures	[11,12]
Poly(anhydride)	Bone, drug delivery	[12]
Poly(hydroxybutyrate)	Valves	[7]
Poly(vinyl alcohol) (PVA)	Cartilage, nerve	[12]
Poly(ethylene glycol) (PEG)	Cartilage	[12]
Poly(ethylene terephthalate) (PET)	Cornea, blood vessels	[13,14]
Expanded poly(tetrafluoroethylene) (e-PTFE)	Cornea, blood vessels	[13,14]
Poly(propylene fumarate)	Bone, cardiovascular tissue	[12]

Advancement in three-dimensional polymer processing makes customization possible for polymer composition, mechanical strength, cell–surface attachment interactions, degradation rates, and high cell density.[15]

Scaffold properties influence a plurality of cell culture aspects including proliferation, differentiation, adhesion, migration, gene expression, and function. Characteristics on three size scales influence these aspects. On the macroscopic scale, the scaffold is conformed to a specific shape and size to direct the formation of a three-dimensional structure. For example, the scaffold utilized for blood vessel regeneration would be tubular in shape in order to direct cell growth and tissue morphology accordingly. A three-dimensional matrix has a high surface area to volume ratio that allows for high-density cell populations and sufficient space for nutrient transfer. The mechanical strength of the scaffold may be an important consideration depending on whether the implanted tissue is subject to a large amount of stress, as for example, with cartilage.

On the microscale, the porosity and pore structure regulate cell penetration, migration, interaction, and growth. Optimal porosities allow for penetration of cell seeding suspensions throughout the scaffold resulting in uniform distributions. The successful mixing and distribution of cells within a scaffold result in chondrocytes functioning properly by producing ECM molecules at high cellularity for enhanced cartilage strength characteristics of the tissue.[16] The morphogenesis of the developing tissue is influenced by the allowable migration of cells. The pore size distribution relates to the migration ability because it determines the amount of space available. The porosity of the scaffold and the size of the pores affect the supply of nutrients and mediation of the waste concentrations via fluid and mass transfer mechanisms. Transfer considerations are increasingly important as high cell density cultures are obtained which limit the available space for fluid and nutrient transport. In fibrous scaffolds, the fiber diameter and affiliated surface curvature affect the spreading ability of attached cells. Spreading allows cells to increase proliferation and this is regulated by fiber dimensions. Additionally, the diameter affects the degree of cell–cell interactions allowable around the fiber which are necessary for proper tissue function. Patterning of the scaffold surface, such as grooves, directs cell adhesion as well as cell growth and function for certain cell types.[17]

On the nanoscale, the surface chemistry of the scaffold must recreate the important cell–ECM properties of adhesion and control. Biocompatibility of the scaffold surface with cells is key for allowing adhesion and migration of cells. The amino acid sequence of arginine–glycine–aspartic acid (RGD) has been identified on fibronectin and other ECM glycoproteins as a key adhesion domain, and the design of synthetic scaffolds incorporating the peptide has been successful

in improving adhesion and cytocompatibility.[18] The organization of RGD peptides on the scaffold surface affects adhesion as well, with a clustered arrangement optimal rather than randomly positioned RGD peptides. Scaffolds can be supplemented with binding or signal molecules bound to the polymer surface. For enhanced proliferation and differentiation, scaffolds can be designed to release growth factors efficiently. Neural cells were cultivated in rats on a scaffold equipped with degrading beads that released nerve growth factor in a controlled manner.

If a scaffold is transplanted, the rate of biodegradability is important to ensure that the scaffold remains to support a transplant until a natural ECM replaces it. The biodegradation or resorption rate is a function of the scaffold composition, structure, and the mechanical load present at the site of transplantation.[7] The necessary rate at which the scaffold is degraded varies according to the tissue type. For example, slow degradation is allowable in bone tissue, whereas in other tissues chronic inflammation may occur if the rate is too low.[7] It is important that the degradation by-products are nontoxic to the body.

Cell Culture

To develop a tissue in culture for use as a tissue substitute, the tissue cell density must be high (commonly 10^9 cells/ml) and uniform within the scaffold. In neomorphogenesis, cells are brought into contact with a porous scaffold and they form a structure together. The cell seeding process must be optimized to achieve uniformity of cell distribution within the scaffold, to maximize utilization of the cells, and to minimize seeding process time in order to avoid damage to the cells. To seed a scaffold, cells and scaffolds are incubated together to allow for adhesion to take place. Dynamic seeding protocols incorporate mixing or flow to distribute cells throughout the scaffold efficiently. When introduced, cells attach to the scaffold surfaces if the surface chemistry of the scaffold is compatible. Scaffolds can be pre-treated to alter the surface chemistry thus allowing improved compatibility with the cells. For example, increasing the concentration of hydrophilic compounds of the surface will improve cell adhesion.

The medium utilized to provide complete nutrition for the growth of different cell types is based on a standard minimum essential nutrient composition. Most media consist of a sugar source, minerals, vitamins, and amino acids. Serum, such as fetal bovine serum, commonly supplements the medium to enhance cell growth. Growth factors and cytokines are utilized to accelerate cell growth through interaction with specific cell receptors. Differentiation inducers are added to direct the differentiation pathway of stem cells. Important parameters to control include the pH, pCO_2, and pO_2 of the media.

There are several types of bioreactor designs utilized for cultivating new tissue growth, and four are shown schematically in Fig. 3. The environment within may be static or mixed using internally designed or externally applied mechanisms. Static environments rely on diffusion as the mass transfer mechanism for nutrient supply. Mixing within a culture vessel provides convective flow of oxygen and nutrients to the cells while removing waste from the surroundings. Mixing-induced shear stress levels are an important consideration since they may cause cell death. The simplest design is the plate or Petri dish. The spinner flask is larger in scale and has an internal agitation mechanism to provide a uniform nutrient concentration within the medium and the enclosed cell–scaffold construct. Perfused bioreactors are culture environments in which media are circulated within a closed system past the immobilized cell–scaffold components. The continuous flow allows for uniform nutrient supply with enhanced mass transfer. Long-term stability of the culture is attained using continuous perfusion reactors.[19] Hollow fiber bioreactors are specialized perfusion bioreactor designs in which a semipermeable membrane in a tubular configuration creates an interior and exterior region. Media can pass through the interior region of the hollow fiber axially, while the membrane allows nutrient and product transfer into and out of the extracapillary region, respectively, where cells are located growing on the exterior surface of the membrane.[20]

As a novel bioreactor design, rotating-wall bioreactors spin on an axis, and the enclosed cell–scaffold constructs tumble within the rotating microgravity environment. Tissue culture in microgravity has been shown to improve cellular aggregation and produce highly differentiated tissue products.[21] Rotating-wall bioreactors avoid the high shear stress found within bioreactors that have agitation devices. As a result, altered gene expression favors improved aggregation resulting in aggregates up to ten times larger in diameter than those attained in conventional bioreactors. Additionally, necrosis of cells within the center of the aggregates due to mass transfer limitations is not seen. The stability of tissue constructs after removal from the microgravity environment remains to be proven in order to realize in vivo utilization.[21]

Other stimuli may be incorporated into the culture environment to cultivate proper tissue function. Mechanical or electrical stimulation, provided at frequencies simulating in vivo conditions, have been shown to improve the resulting properties of the cultivated tissue.[19] For example, pulsatile conditions that simulate a beating heart are utilized in the culture of blood vessels resulting in improved strength and function relative to cultures lacking this stimulus.[2]

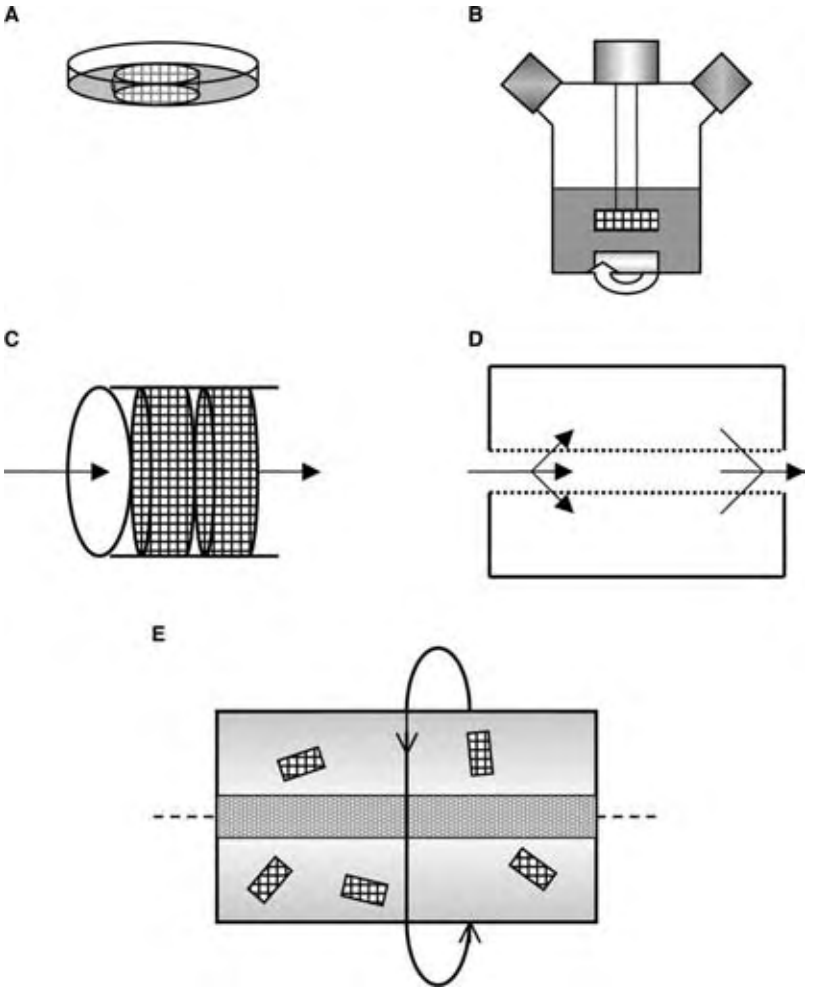


Fig. 3 Tissue engineering culture environment designs: (A) plate; (B) spinner flask; (C) perfusion; (D) hollow fiber; and (E) rotating wall bioreactor.

In cell culture, the essential nutrients must be present in the medium and be able to flow or diffuse to the cell membranes to allow for a viable culture. The bioreactor design needs to enable the cell population to expand to the cell density of the in vivo tissue; however, nutrient limitations deter this when cell growth decreases the space available within the scaffold for nutrient diffusion. Additionally, stagnant build-up of waste by-products increases acidity and harmful conditions. Therefore, the viability of a culture decreases within heavily populated regions of the scaffold. Supplying nutrients to these interior regions is a challenge, and currently, a tissue thickness of over 1 mm is not maintainable without cell death in the core of the tissue mass. In vivo, nutrient transfer within a tissue is achieved by a process called angiogenesis. As tissue mass increases, vasculature is established by promoting blood vessel growth within a starvation zone. A similar process of angiogenesis established within an in vitro culture would permit large-scale tissue growth. A tissue transplant can promote angiogenesis by providing angiogenic growth factors, such as vascular endothelial growth factor, along with the transplant that are released over time to allow for high cell density tissue regeneration.

Many tissues consist of more than one cell type. The fact that most major organs in the human body consist of more than one tissue and cell type adds to the complexity in recreating a functional organ replacement. The proper physiological function of these tissues depend on the interactions between these multiple cell types, so developing a tissue substitute with the ability to restore function should be a heterogeneous culture consisting of multiple cell types organized appropriately. Coculture of multiple cell types within one environment in order to accomplish this is very difficult. Currently, two different cell types are cultured together by growing each cell type in layers with membranes. The membranes allow for signal passage without direct cell–cell contact. The concept of ''organ printing'' may permit the generation of heterogeneous, vascularized, and three-dimensional organ constructs using a computer-controlled ''printing'' device that deposits multiple cell types, biomaterials, and other tissue components layer by layer to form an organized structure.[22]

TRANSPLANTATION AND CRYOPRESERVATION

When implanting a tissue substitute or therapy device, immune rejection by the host against the foreign implant is the primary concern unless a biocompatible scaffold with autologous cells is used. Several strategies exist to circumvent this challenge, depending on cell type, source, and desired function. Three specific strategies are immune system therapy before and after the transplantation procedure, gene modification in the cells prior to tissue development to allow immune system acceptance, and immunoisolation.[23] Traditionally, similar to organ transplants, the complete immune system is suppressed for a period of time following the transplant to increase the chances of the transplant eventually being accepted. There is a key 5–10 week period in which the successful integration of the transplant or graft rejection is determined.[23] However, during this period of immunosuppression, the patient is susceptible to other illnesses. A novel strategy involves building a tolerance in the host for a cell type prior to transplantation. By first performing a bone marrow transplant using donor hematopoietic cells, the host will circulate immune system components that match the future donor. Then, when the transplant occurs, the donor tissue is tolerated.[10] Immunoisolation strategies incorporate a semipermeable membrane enclosing the cell-based device, protecting the cells from immune recognition and yet allowing the transport of therapeutic cell-derived compounds to emanate from the device.

When a tissue-engineered product is not to be used for transplantation upon its creation, cryopreservation allows for long-term storage until future application. Cryopreservation is commonly used for preparing cells for storage, but its utility for storage of tissues remains in development. Cryopreservation is a major focus for researchers, though, as it is a key component for improving the marketability of tissue-engineered products, allowing the development of an on-demand tissue supply, preservation of tissue genetic stability, and establishment of production quality control archives.[24] During the cryopreservation process, both the freezing and thawing procedures are equally important in regulating water displacement and replacement, respectively, while maintaining cellular and tissue integrity. Cryoprotectants and thermal processing protocols are utilized in this process. Commonly utilized cryoprotectants are dimethyl sulfoxide and glycerol, which remove intracellular water to avoid damaging ice crystal formation. Owing to the larger scale of tissues compared to cells, the challenges are greater, including the induction of chemical and thermal gradients within tissue, which must be resolved by utilizing optimized mass and heat transfer operations, respectively. Technologies to improve these processes and to monitor tissue parameters for performance control and modeling will further help to develop applicable cryopreservation protocols.

APPLICATIONS

Tissue engineering applications can incorporate the aforementioned components in various combinations to achieve specific goals. There are five general applications of tissue engineering:

1. The development of human tissues in vitro for future implantation into the body to replace lost tissue function.
2. In vivo tissue regeneration by transplantation of a seeded or unseeded scaffold to aid in regeneration at a deficient site in the body.
3. Development of an in vivo or extracorporeal device that supplements reduced tissue function. The in vivo device is encapsulated within a semipermeable membrane to allow for provision of a therapeutic molecule to the site while protecting the cells from the host immune system. An externally positioned device would provide deficient tissue function compounds through a tube directed to the body site while avoiding cellular contact with the immune system.
4. Establishment of an environment for expanding a cell population that is later extracted from the scaffold for implementation within the body in a cell-based or gene therapy application.
5. Development of a model to promote in vivo-like function of a population of cells in vitro for studying tissue development, pathology, pharmacology, and toxicology projects instead of using animal models.

The future financial outlook for tissue engineering products have varied greatly with some predicting an $80 billion market in 2000.[2] However, owing to increasingly evident challenges involved in developing tissue replacements and a lack of realization of previous tissue engineering product success predictions, more recent market estimates have been around $15 billion annually.[5] Currently, over 20 different tissues have been researched for a tissue engineering application. A selection of these applications is shown in Table 2 and Fig. 4.[26–33] To successfully design a tissue substitute that mimics the normal in vivo counterpart, the cells must perform similar functions with the support of a biomaterial with appropriate biocompatibility, degradation, and strength characteristics and, subsequently, restore functionality of the tissue to the system.

Table 2 Selected tissue engineering applications

Tissue	Why	Cells	Scaffold	Application	Goal	References
Blood vessels	Vessel occlusion	Endothelial	PET, e-PTFE	Replacement	Thromboresistance	[13]
Bone	Disease, radiation	Periosteal, osteoblast, mesenchymal stem cells	PGA, PLA, calcium alginate	Repair, replacement, regeneration	Restored mechanical properties	[25,26]
Brain	Parkinson's and Huntington's diseases	Neurons	Membrane	Replacement, encapsulation	Dopamine production, neural network re-established	[27]
Breast	Lumpectomy, mastectomy	Smooth muscle cells (SMC), fibroblast, chondrocytes	PGA, PLA, PCL	Replacement	Flexible transplant with induced angiogenesis	[28]
Heart valves	Valvular disease	Endothelial, fibroblast	PGA	Repair, replacement	Durability over mechanical devices	[29]
Cornea	Disease, trauma	Epithelial	e-PTFE, PMMA, PVA	Repair, replacement	Elasticity, refractive properties, transparency, curvature	[14]
Pancreas	Diabetes	Islets of Langerhans	Alginate/poly(L-lysine)	Encapsulated, restore	Active glucose concentration control	[30]
Kidney	Renal failure	Endothelial	Hollow fiber	Support device, replacement	Blood filtration, homeostasis maintenance	[31]
Tendons and ligaments	Injury	Fibroblasts	Collagen	Repair, replacement	Mechanical durability	[32]
Red blood cells	Disease, shortage	Artificial with hemoglobin	Membrane	Micro-encapsulation	Avoid removal of artificial cells from circulation	[33]
Liver	Acute liver failure	Porcine hepatocytes	Polysulfone hollow fiber	Extracorporeal support device	Detoxification activity, bridge-to-transplantation	[20]

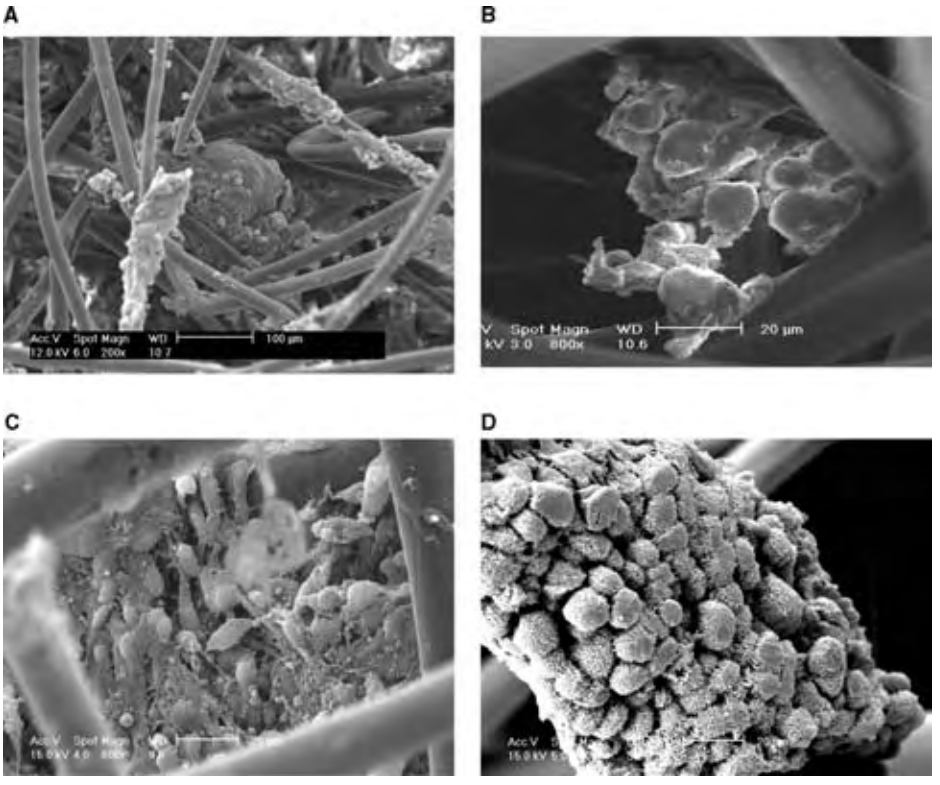


Fig. 4 (A) In vitro modeling of colon cancer; (B) production of cord blood cells; (C) astrocyte culture for cell-based therapy of Parkinson's disease; and (D) placenta model using trophoblast cells for transport.

Skin

Skin was the first tissue to be produced as a tissue substitute. The relatively simple two-dimensional, bi-layered structure of skin primarily consists of keratinocytes, fibroblasts, and ECM components. Skin functions include roles as a protective barrier, fluid and heat regulator, and immune system reconnaissance for early warning of dangers. Common features are its flexibility, elasticity, and strength. The major application is to develop skin substitutes for use on burn victims, especially when skin from the same individual is not available for autogenous transplantation. The main role for the regeneration of skin is to re-establish the barrier function with the dermis and epidermis. Other functions can be developed in vivo via migration of other component cells from surrounding areas into the regenerated skin. Wound healing can be achieved by transplanting an unseeded scaffold at the wound site promoting migration of surrounding cells into the scaffold. Artificial skin substitutes were the first tissue engineering products to reach the market. One available dermal substitute on the market is Dermagraft®, produced by Smith & Nephew (Florida, U.S.A.), which consists of fibroblasts, extracellular matrix, and a bioabsorbable scaffold. When applied, usually for healing diabetic foot ulcers, healthy cells surrounding the wound, including keratinocytes, migrate into the scaffold to fully reconstitute healthy skin with natural barrier properties.[5]

Liver

The liver was the first tissue engineered in three-dimensional scaffolding. Proper differentiated function of a tissue-engineered construct containing hepatocytes includes the production of albumin and the completion of urea and bilirubin metabolism. The cells used to seed the scaffold should be highly proliferative in order to develop a high cell density tissue. Owing to the size of the liver, nutrient diffusion limitations are a concern, and, therefore, vascularization of liver construct is necessary for blood to supply nutrients within the liver mass. When the engineered liver is transplanted to the host site, vascularization promoters can be included to initiate the blood vessel migration into the new tissue. Hepatocytes have also been cultured in hollow-fiber membranes to supply liver function from a device outside the body. The HepatAssist System (Circe Biomedical, Inc., Massachusetts, U.S.A.) is an extracorporeal device supporting liver function in patients waiting for transplantation.[20] The device utilizes porcine hepatocytes immobilized in the outer space of a hollow-fiber design to perform liver functions while plasma is circulated inside the hollow fiber.

The polysulfone membrane allows for protein and toxin transport across the membrane while preventing the passage of foreign cells into the patient. Thus, hepatocytes in the device do not have to be compatible with the host immune system since there is no direct contact.

Bone

Autogeneous bone grafts are difficult because there is a limited supply of cells, and donor sites are subject to post-graft morbidity and bone deformations. Tissue-engineered bone structures consist of osteoblasts and osteocytes, which re-form in three-dimensional scaffolds to produce signature bone structures and features such as load-bearing ability and osteocalcin secretion for osteoblasts. The progenitor cells utilized in the osteogenesis can provide developmental signals such as bone morphogenetic proteins, which induce and promote bone growth. Periosteal cells, from the outer bone membrane periosteum, were seeded on calcium alginate hydrogel to allow for calcium extraction by the cells and to make customized bone-shaped molds for transplantation into bone defect sites.[25]

Gene Therapy

Gene therapy is introducing DNA, which codes for an essential protein, into cells that are lacking productivity of the protein. Gene therapy is an alternative to supplying the deficient region with the protein itself or replacing the defunct tissue with a functional tissue substitute. If stem cells are utilized, the regenerating source of corrected, functional cells will eventually replace expired, deficient cells. DNA is carried and introduced into cells using a delivery system which is commonly viral or plasmid based.[34] Once the gene is expressed within the cell, persistence of the essential protein supply is key. Accompanying the huge potential of this process are important safety concerns with regards to the implementation of the DNA correctly within the chromosomes. Incorrect incorporation of the gene could alter important control genes or other essential genes. Once the successfully transduced cells are selected, the cell population is expanded in vitro.

In Vitro Studies

Tissue engineering can also be used to develop laboratory tools using human cells to perform pathological, developmental, pharmacological, and toxicological studies. Tissue cultured as an in vitro model can be used to study tissue or organ development processes including signaling and control mechanisms. Additionally, toxicology and carcinogen studies on tissue models provide insight without utilizing animal models.[35] Cancer models can be used to study tumor biology, control, and progression as well as use in cancer treatment studies.[36]

Cell-Based Technologies

Additional cell-based technologies are being developed by utilizing the knowledge of cells and their interaction with materials attained from tissue engineering research. Biosensors take advantage of the specificity of cell surface receptors for select target molecules and the high signal amplification for detecting low levels of chemical or biological agents. The brain consists of a neural network that processes a large number of signals. Neurons organized on a support in specific patterns could produce in vitro neural networks that pass signals similarly to the electronic structure on microchips.[37]

CURRENT CHALLENGES

Although there has been some limited success in producing tissue substitutes, especially artificial skin, the overall tissue engineering aim of creating tissue that can replace, maintain, or improve deficient in vivo tissue is not an easy task. Multiple challenges remain in providing tissue products to meet demand. These include creating a greater supply of cells to avoid immune rejection, such as a universal donor cell supply. Advancements in biomaterial production and bioreactor design are necessary to provide customized support and environments to replicate the complex in vivo environment. When developing complex, three-dimensional tissues, challenging issues include promoting angiogenesis to overcome tissue size limitations, the heterogeneous coculture of multiple cell types within one tissue, and the cryopreservation of the resulting tissue before transplantation.

CONCLUSIONS

Through the multidisciplinary application of biological and medical knowledge and engineering skills, tissue engineering allows the development of tissues in the human body to restore, repair, or improve deficient biological components. Generally, utilizing cells, supportive scaffolds, and designed culture systems, tissue development processes face many challenges in reconstituting in vivo-like performance. Although many challenges remain, the field is a major research focus, and medical demand is strong, both ensuring that

tissue engineering stands to provide momentous advancements to health in the 21st century.

REFERENCES

1. Ma, T. Fiber-Based Bioreactor Systems In Mammalian Cell Culture and Tissue Engineering Human Trophoblast Cells; Dissertation; The Ohio State University: Columbus, OH, 1999.
2. Langer, R. Tissue engineering. Mol. Ther. **2000**, *1* (1), 12–15.
3. http: //www.optn.org/latestData/step2.asp (accessed July 2003).
4. Schulthesis, D.; Bloom, D.A.; Wefer, J.; Jonas, U. Tissue engineering from Adam to the zygote: historical reflections. World J. Urol. **2000**, *18*, 84–90.
5. Flanagan, N. Engineering methods for tissue and cell repair. Genet. Eng. News **2003**, *23* (8), 10.
6. Martins-Green, M. Dynamics of cell–ECM interactions. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 33–55.
7. Griffith, L.G. Biomaterials. In *WTEC Panel on Tissue Engineering Research*: *Final Report*; McIntire, L.V., Ed.; Academic Press: New York, 2002; 9–22.
8. Hubbell, J.A. Matrix effects. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 237–250.
9. Heath, C.A. Cells for tissue engineering. Trends Biotechnol. **2000**, *18*, 17–19.
10. Brower, V. Stem cell research and development advances in face of challenges. Genet. Eng. News **2003**, *23* (9), 1.
11. Pachence, J.M.; Kohn, J. Biodegradable polymers. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 263–277.
12. Elisseeff, J.H.; Langer, R.; Yamada, Y. Biomaterials for tissue engineering. In *Tissue Engineering and Biodegradable Equivalents*: *Scientific and Clinical Applications*; Lewandrowski, K.-U., Wise, D., Trantolo, D., Gresser, J.D., Yaszemski, M.J., Altobelli, D.E., Eds.; Marcel Dekker: New York, 2002; 1–24.
13. Xue, L.; Greisler, H.-P. Blood vessels. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.-P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 427–446.
14. Trinkaus-Randall, V. Cornea. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.-P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 471–491.
15. Langer, R. Selected advances in drug delivery and tissue engineering. J. Control. Release **1999**, *62*, 7–11.
16. Vunjak-Novakovic, G.; Freed, L.; Biron, R.J.; Langer, R. Effects of mixing on the composition and morphology of tissue-engineered cartilage. Am. Inst. Chem. Eng. J. **1996**, *42*, 850–860.
17. Saltzman, W.M. Cell interactions with polymers. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 221–235.
18. Griffith, L.G.; Naughton, G. Tissue engineering—current challenges and expanding opportunities. Science **2002**, *295*, 1009–1014.
19. Freed, L.E.; Vunjak-Novakovic, G. Tissue engineering bioreactors. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 143–156.
20. Mullon, C.; Soloman, B.A. HepatAssist liver support system. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 553–558.
21. Unsworth, B.R.; Lelkes, P.I. Tissue assembly in microgravity. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 157–164.
22. Mironov, V.; Boland, T.; Trusk, T.; Forgacs, G.; Markwald, R.R. Organ printing: computer-aided jet-based 3D tissue engineering. Trends Biotechnol. **2003**, *21* (4), 157–161.
23. Hardin-Young, J.; Teumer, J.; Ross, R.N.; Parenteau, N.L. Approaches to transplanting engineered cells and tissues. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 281–291.
24. Karlsson, J.O.M.; Toner, M. Cryopreservation. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 293–307.
25. Vacanti, C.A.; Bonassar, L.J.; Vacanti, J.P. Structural tissue engineering. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 671–682.
26. Bruder, S.P.; Caplan, A.I. Bone regeneration through cellular engineering. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 683–696.
27. Wahlberg, L.U. Brain implants. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 773–783.

28. Lee, K.Y.; Halberstadt, C.R.; Holder, W.D.; Mooney, D.J. Breast reconstruction. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 409–423.
29. Love, J.W. Cardiac prostheses. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 455–467.
30. Wang, T.G.; Lanza, R.P. Bioartificial pancreas. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 495–507.
31. Humes, H.D. Renal replacement devices. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 645–653.
32. Goulet, F.; Rancourt, D.; Cloutier, R.; Germain, L.; Poole, A.R.; Auger, F.A. Tendons and ligaments. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 711–722.
33. Chang, T.M.S. Red blood cell substitutes. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 601–610.
34. Fradkin, L.G.; Ropp, J.D.; Warner, J.F. Gene-based therapeutics. In *Principles of Tissue Engineering*, 2nd Ed.; Lanza, R.P., Langer, R., Vacanti, J., Eds.; Academic Press: New York, 2000; 385–405.
35. Li, A.P. Screening for human ADME/Tox drug properties in drug discovery. Drug Discov. Today **2001**, *6* (7), 357–366.
36. Chung, L.W.K.; Zhau, H.E.; Wu, T.T. Development of human prostate cancer models for chemoprevention and experimental therapeutics studies. J. Cell. Biochem. Suppl. **1997**, *28/29*, 174–181; Supplement.
37. Mrksich, M. Cell-based technologies: non-medical applications. In *WTEC Panel on Tissue Engineering Research*: *Final Report*; McIntire, L.V., Ed.; Academic Press: New York, 2002; 61–69.

Trace Elements

Ian D. Brindle
Brock University, St. Catharines, Ontario, Canada

INTRODUCTION

It is a truism that nothing in the universe is free from trace elements. As analytical techniques have improved with the development of ever more sophisticated measuring devices, the impact of trace and ultratrace concentrations of elements has been revealed much more clearly over the last two decades or so. We know from nutritional studies the importance of ultratrace concentrations of several elements including, for example, cobalt in cyanocobalamine (vitamin B12), where the daily requirement for this vitamin is only 1 μg/day for an adult.

Some elements commonly occur in ores of other elements and so can become problems, even where they are present at ultratrace levels. These elements can have deleterious or beneficial effects upon the properties of the major element that is won from the ore. The impact of trace elements on properties of materials, even when they are present at ultratrace concentrations, can be striking, and understanding their influence on materials is important. Regulations, either environmental or prescriptive, will continue to have a major impact on our interest in trace elements.

ENVIRONMENTAL IMPACTS

Treatment of many materials results in the liberation of the trace elements into the environment, which can have an impact on health. Coal is a particularly useful example of a major source of trace elements poured into the environment from coal combustion. Coal contains an alphabet soup of trace elements, including arsenic, mercury, uranium, selenium, and chromium. Pyrite is a ubiquitous mineral found in coal, but coal can also contain a variety of other mineral phases. West Virginia coal, for example, includes clay minerals such as kaolinite (35%) and illite (35%), quartz (18%), pyrite (7%), and calcite (3%).[1] A number of projects that utilize coal for power generation while minimizing the impact on the environment have been described. An excellent example is the SNOX^{TM} (trademark owner Haldor Topsoe) demonstration project, which utilizes high-sulfur coal (2.8%).[2] The demonstration project of this technology, equally funded by the U.S. Department of Energy and participants at a total cost of U.S.$ 31.5 million, was able to reduce NO_x to nitrogen (>90% reduction) and to oxidize SO_2 to SO_3 (>95% transformation) and ultimately to make sulfuric acid for sale, and hence minimizing the amount of acid rain produced by the system. Mercury was not retained at all by the system and boron was only partially retained. Selenium and cadmium, normally problematic elements, were recovered in the processing of the spent gases. Particulate emissions were reduced by 99%. Mineral content of coal is variable and each new source must be evaluated closely for its trace element composition. For some applications, it may be a case where the raw material should not be processed. Thus, in Guizhou province of the People's Republic of China, coal is mined from a deposit that is very high in arsenic.[3] Germanium is also relatively high in coal. The highest reported value of germanium found in coal ash was 1.1% in a sample from a particular seam of coal in Durham, U.K. Chalcophilic elements are usually associated with sulfur minerals in coal, and some success in removing these elements prior to burning the coal has been achieved by washing the coal and removing the heavier mineral components.

Other fuels are also susceptible to contamination with trace elements at low, but significant, concentrations. A report prepared for the United States Environmental Protection Agency (USEPA) describes several sources of contamination from traces of mercury in oils and natural gas.[4] Elemental mercury is present in liquid petroleum oil or in natural gas, condensates at parts per million level, and must be removed from the product stream to prevent it from reacting with metallic components, particularly during the cryogenic treatments, where liquid mercury can be condensed from the gas phase. Mercury concentrations can vary considerably, and the EPA report describes one gas reservoir in Texas, where the concentration of mercury is sufficiently high to lead to the conclusion that the gas is in equilibrium with elemental mercury in the subsurface reservoir. In terms of processing of petroleum and natural gas, mercury is found in a variety of compartments and in a variety of forms or species including alkylmercury and inorganic mercury salts such as mercury halides and mercury sulfides (insoluble and largely found in suspension or in drilling wastes). Table 1, which presents predominant species in a variety of matrices, is adapted from the report.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007994

Table 1 Approximate distribution and abundance of mercury compounds in hydrocarbons

	Coal	Natural gas	Gas condensate	Crude oil
Hg^0	T	D	D	D
$(CH_3)_2Hg$	[a]	T	T, (S[a])	T, (S[a])
$HgCl_2$	S[a]	N	S	S
HgS	D	N	Suspended	Suspended
HgO	T[a]	N	N	N
CH_3HgCl	[a]	N	T[a]	T[a]

Abundances, expressed as percentages of total Hg concentrations: D (dominant), greater than 50 percent of total; S (some), 10-50%; T (trace), less than 1%; N (none), rarely detected.
[a]Inconclusive data.
(From EPA-600/R-01-066.)

In view of the widely varying toxicities of these species and, in particular, the unexpectedly huge toxicity of dimethylmercury, for which we have only one reported human death (Professor Karen Wetterhahn of Dartmouth College, who died as a result of a single and very limited exposure to dimethylmercury in 1997), the significance of these trace concentrations of ''species'' must be considered when the health of workers who will be exposed to these compounds is evaluated.

An element that is also found in natural gas represents the world's best source for the element and that is helium which, in some U.S. and polishgas supplies, can reach 7% by weight. This valuable gas is collected by liquefaction of the natural gas, which leaves the remaining gas significantly enriched in helium. Helium originates from radioactive decay and, because it is a fugitive gas and can escape the earth's atmosphere, it is not practical to recover helium from the air.

TRACE ELEMENTS IN METALS

Trace concentrations of a variety of elements have huge effects upon metals and can, in many cases, determine the metals' fitness for various purposes in manufacture and use or the traces may disqualify a particular form of the metal from use either because of the changes in physical or chemical properties or because it may affect the end use or disposal of the metal.

Iron

Iron is one of the most recycled metals and the recycling process can introduce unwanted elements into the alloys and can play havoc when they are ignored. Iron, cast into the engine blocks of internal combustion engines, is often made from gray iron or nodular iron. Nodular cast iron has a crystallinity and hardness that enables the casting to support the continuous motion of pistons and piston rings, suitably lubricated, for more than 100,000 miles without significant wear. Reproducible production of nodular cast iron is clearly of great economic importance. One trace element that has been identified in guaranteeing the quality of nodular cast iron is lanthanum. Although not required as a pure additive, several rare earth elements confer advantages on the manufacturing process. Often, lanthanum is added as an ingredient in a rare earth mix of elements, which typically includes cerium at higher concentrations.

Much of the control of microscopic structure of iron depends on the management of the carbon content of the metal. Carbon at levels less than 5% precipitates in a variety of forms and can also react with iron and other metals in the alloy to make carbides. Trace elements can significantly alter the form in which carbon occurs, and manipulation and control of trace elements are major determinants of quality standards for categorizing irons by the American Society for Testing Materials (ASTM). The structure of precipitated carbon determines the properties of the alloy, in particular its tensile strength. Lead at levels in iron that exceed 50 parts per million (ppm) results in the formation of Widmanstätten graphite, which confers a mossy or fuzzy appearance to graphite flakes in the iron. Without an appropriate modifier, Widmanstätten graphite reduces the tensile strength of iron to less than 15,000 pounds per square inch (psi). Addition of cerium, in a process usually called ''inoculation,'' reduces the effect of lead to more manageable levels. Nitrogen also affects the form of carbon found in iron. Normally, nitrogen equilibrium concentrations are less than 70 ppm. When the values exceed 150 ppm, the graphite form in gray iron is ''fatter'' than it would be in the presence of low concentrations of nitrogen. Titanium, present at trace concentrations, reverses this trend and can restore graphite flakes to normal, filamentous morphology.

Boron-containing low-carbon steels have a number of advantages that are specifically because of the

presence of boron. Subtle effects relate to boride concentration appearing at grain boundaries. In combination with niobium, boron confers several advantageous properties to the so-called bainitic steels. Thus boron low-carbon steels have improved hardness, strength, weldability, and corrosion resistance, which has made them the steel of choice in machinery manufacture, oil pipelines, and maritime applications (bridges, drilling rigs, and ships). Furthermore, as they can be rapidly cooled, these steels are relatively energy efficient materials.

Trace concentrations of silicon and aluminum tend to decrease the strength and hardness of iron by increasing the ratio of ferrite to pearlite. Nickel, copper, and tin work in the contrary direction and, by increasing the amount of pearlite, increase the strength and hardness of iron. In some cases, not surprisingly, trace elements can confer benefits for some purposes and liabilities with others. The impact of trace elements on creep deformation and fracture was investigated in depth.[5] The conclusions were uncertain as to the effects of trace concentrations of several elements (P, As, Sn, etc.), but they determined that traces of titanium decrease embrittlement caused by trace elements. Other work suggested that arsenic and antimony reduce the ductility of steel and increase brittleness; antimony also improves the resistance of iron to corrosion.[6] Welz and Melcher[7] noted that traces of bismuth improve the machining properties of steel up to a point but, even at low levels, steels may break during cold working. Tin has a deleterious effect upon the hot working of steel.

As we are better informed about the fate of materials in the environment by our greater knowledge of the cycling of elements in the environment, the present-day regulators are increasingly concerned about ''cradle to grave'' management of materials. One critical aspect of such regulations is the anticipation of issue around the disposal of wastes at the end of the life of materials. Thus, some iron and steel manufacturers are anticipating the fate of iron and steels when they re-enter the environment as they rust and disintegrate. One striking example will serve to illustrate this case. A study was commissioned by the U.S. Department of Energy to review the fate of ferritic steels with regard to shallow burial or recycling.[8] The anticipated problems relate to the activation, by neutron absorption, of various elements adventitiously or deliberately present in the steel. The radioactive isotopes, generated by neutron activation, limit the disposal or recycling options, and elements including silver, molybdenum, and niobium were identified as potentially problematic. Niobium is deliberately added to iron, as noted above, and because it is readily activated, it can render radioactivity to the iron. Other elements, including silver and gold, appear at trace concentrations and are likely carried into the irons and steels from the processing of the iron ores, which can be problematic when activated by the neutron flux from a nuclear reactor.

Copper

High purity copper is an essential component in the electronics industry, and a number of trace elements can decrease the conductivity of copper. As electrical resistance translates into energy losses as heat (I^2R), manufacturers are anxious to ensure the lowest resistance of copper wiring in circuitry so that they can reduce both the energy and heat load in electrical and electronic equipment. Most trace impurities increase the resistivity of copper and so manufacturers look to develop the highest purity copper for applications in the electronics industry. Cadmium appears to have the least effect upon copper resistivity.

Zinc

The dry cell battery industry is a major consumer of zinc. The amount of current that can be withdrawn from a dry cell battery is limited because of polarization at the surface of the zinc, caused by the overpotential, which is in turn caused by hydrogen at the surface. This overpotential was traditionally reduced by incorporating an oxidizing agent (manganese dioxide) to react with the hydrogen. Incorporating small amounts of mercury in the battery extends battery life by limiting a phenomenon known as local action. Local action is caused by trace impurities in zinc, which set up independent galvanic cells that react and reduce the useful life of the battery. Incorporation of mercury in the zinc is proposed to separate the impurities from the zinc and thereby reduces the independent galvanic activity. As mercury has become increasingly regulated, indium has taken its place in dry cell batteries. Another concern for battery manufacturers is the concentration of the hydride-forming elements that appear to be responsible for significant losses in performance of dry cell batteries even at low concentrations. One of the products manufactured by the primary producers of zinc is a 30% m/v $ZnSO_4$ solution. The detection limits desired by the manufacturers for hydride-forming elements in this concentrated solution are 10 μg/L for arsenic, antimony, bismuth, tin, selenium, and tellurium, and 2 μg/L for germanium.[9]

CATALYST POISONING BY TRACE ELEMENT

Coal combustion remains the single largest source of energy used in the generation of electricity in the

U.S.A. Energy production in 2000 was 1968 TW hr and coal represented 51% of the total. By 2010, energy production is anticipated to rise to around 4600 TW, with coal burning being responsible for about 2200 TW hr or 48% of the total previously noted.[10] Coal is a source of many trace elements; while mercury may be a problem for human health, other elements create problems in other areas. As the large amounts of coal are burned (around 10 million tons per year for some power stations), efforts are underway to improve the quality of emissions to fulfill a mandate to use ''clean coal'' technology. Selective catalytic reduction (SCR) is employed to reduce NO_x emissions (a component of acid rain) into the environment from burning coal. Arsenic in the coal is usually oxidized to arsenic trioxide (As_2O_3), which has a significant vapor pressure in the flue gases and poisons SCR catalysts. Arsenic removal is possible when there are sufficient calcium oxide equivalents in the coal, which can react with the arsenic trioxide to give nonvolatile calcium arsenate [$Ca_3(AsO_4)_2$]. The desirable presence of calcium-containing minerals in coal in this application suggests that minerals should be retained in coal to deal with arsenic. This retention, however, appears to run contrary to the desirability of removing minerals as a source of both arsenic and other toxic elements noted above. As coal washing is used to discriminate coal from noncoal, it may be difficult to separate calcium minerals that confer benefits from iron sulfide minerals that are problematic.

Polyolefin production is enhanced by a new generation of metallocene cocatalysts that are used together with Ziegler–Natta catalysts. These metallocene catalysts are poisoned by the presence of oxygen in the feedstock. The presence of oxygen can also be problematic by causing some crosslinking, chain branching, and delay of induction of polymerization. Typical olefin feedstock has oxygen maintained at levels below 50 parts per billion (ppb).

Reforming catalysts are poisoned by sulfur and halogens. The complexity of reactions involved in reforming means that catalysts can work to the advantage of some reactions over others. This advantage can work against the process for, for example, the production of hydrogen. The Boudouard reaction, for example, removes carbon from the reforming process. Some catalysts need to be poisoned to some degree. Sulfur, present in most natural gas, is an appropriate poison for some catalysts, and it is not totally removed from the process stream, as it poisons the catalyst for the Boudouard reaction without materially affecting the reforming reactions. Several important reforming reactions are presented in Scheme 1.

Sulfur, chlorine, and phosphorus are classed as temporary poisons in reforming reactions, because the activity of the catalyst is recovered when the poison is removed from the gas stream. Permanent poisoning results from the presence of certain metal vapors including zinc, lead, and arsenic. Alkali and alkaline earths are also permanent poisons. The occurrence of permanent poisons is a rare event in reforming reactions. Sulfur often occurs at levels below 20 ppm, therefore sulfur removal may not be necessary, but the levels of sulfur should be monitored.

Steam reforming:	$C_nH_{2n+2} + nH_2O \rightarrow (2n+1)H_2 + nCO$
Watergas shift reaction:	$CO + H_2O \rightarrow CO_2 + H_2$
Boudouard reaction:	$2CO \rightarrow CO_2 + C$
Beggs reaction:	$CO + H_2 \rightarrow H_2O + C$
Carbon dioxide reforming:	$C_nH_{2n+2} + nCO_2 \rightarrow 2nCO + (n+1)H_2$

Scheme 1 Reforming reactions.

Carbon monoxide is a common impurity in hydrogen that has been generated by a reforming process. Portable power sources, used by the military for energy generation in remote locations, use a methanol reforming reaction as shown in the following equation:

$$CH_3OH + H_2O \rightarrow CO_2 + 3H_2 \quad (1)$$

Another significant reaction can also occur, which is shown in the following equation:

$$CH_3OH \rightarrow CO + 2H_2 \quad (2)$$

This reaction, or the hydrocarbon reforming reactions noted above, can generate sufficient carbon monoxide to poison the catalysts used for hydrogen-based fuel cells that are used to generate electricity. Thus, carbon monoxide is a trace component of concern, and as the hydrogen economy is further developed and reformation reactions may become more important in the development of hydrogen fuels, technologies for the efficient removal of carbon monoxide must be developed.

One of the most widely known examples of catalyst poisoning is taken from the automobile industry. Though tetra-ethyl lead has been removed from essentially all gasoline in North America, the ban on leaded gasoline is not worldwide, and leaded and unleaded gasoline is available in many countries. Catalytic converters, which contain precious metals like platinum, palladium, and rhodium, are used to both reduce NO_x and oxidize CO and unburned hydrocarbons. Lead irreversibly destroys the catalytic ability of the converter. Concentrations of lead in leaded gasoline are nominally 150 mg/L.

TRACE ELEMENTS IN SOLUTION

Treatment of solutions, when they are used to crystallize compounds, is done in various ways to change the size and the nature of crystals formed. Sudden cooling usually results in the formation of small crystals, whereas crystals formed by slow cooling tend to be large. In addition, rapid precipitation often results in the incorporation of traces of contaminants in the crystals. Finely divided crystals have a large surface area and can be used advantageously as a way to remove impurities from solutions. This technique has been used for a variety of purposes, from the removal of trace elements from drinking water to the removal of radionuclides from solutions derived from spent nuclear materials. Magnesium hydroxide has also been used analytically as a method to preconcentrate elements. Using this technique, Brindle et al.[11] were able to preconcentrate germanium and determine its concentration in the range of 1 pg/ml.

The impact that trace concentrations of elements in solutions can have on precipitation reactions is quite striking and can be problematic, or can confer economic benefits. Impurities at the levels of 10^{-3} to 10^{-4} M, which consist of relatively large organic molecules, have an inhibitory effect upon the growth of crystals but do not appear to fully inhibit their growth. Concentrations of inorganic complexes, even if they are low, can inhibit or ''poison'' the growth of crystals by adsorption at a surface step rather than their being adsorbed as a surface monolayer. Sears proposed such a mechanism for the inhibition of the growth of potassium bromide and lithium fluoride in the presence of FeF_6^{3-} at concentrations in the range of 10^{-5} to 10^{-6} M.[12]

In addition to affecting the crystal growth, these foreign ions can affect also the shape of crystals.[14] At higher concentrations, impurities can cause a variety of changes. If the impurity is isostructural with the major component, the morphology of the resulting crystal will be intermediate between the two forms. Where adsorption of ions onto a crystal surface takes place, small changes in the lattice position of ions on a facet of the crystal farthest from the center of symmetry have the largest effects on the surface energy, which in turn can modify the habit of the crystal. It appears to be a general principle that the adsorbed ions have a minor impact upon habit, and the ions that insert themselves into the crystal lattice have the greatest effect in altering crystal habit.

The normal crystal habit of many commodity chemicals is incompatible with their use. Not all crystalline materials can flow readily. In applications where a smooth-flowing powder is required, the natural shapes of crystals may not lend themselves to being free-flowing, and ''clumping'' can occur, which can clog feed tubes and hoppers. Thus if one can alter the crystal habit to a form which allows a more free-flowing powder, time and product can be saved. The modification of habit has occupied process chemists for many years, and a number of high-volume chemicals are routinely treated with habit modifiers to improve the flow characteristics, especially of these commercial products.

A crystal has several faces, and the nature of interactions with the surrounding solutions, including the impact of supersaturation, solvent, temperature, and the level of impurities in solution, varies widely according to the energies of the various faces. Because crystals expose a different chemical environment at each face, it should not be surprising that impurities will have a major impact on the crystal habit—all other conditions remaining the same. Face-specific interactions of impurities, even when they are present at low concentrations, can be considerable. As discussed in this entry, impurities tend to reduce the rate of growth at the site of adsorption, thus allowing other facets to grow at their expense. Manufacturers usually prefer granular or prismatic crystal habit, but other forms, including needle and flaky crystals, are sometimes desirable. There are thousands of papers in the literature that describe the modification of crystal habit, and this entry can only touch upon a few.

Highly charged and/or complex ions appear to have the greatest effect upon crystal habit. Thus, Al^{3+}, Cr^{3+}, Fe^{3+}, and $Fe(CN)_6^{4-}$ are commonly used in a variety of applications to modify habit. A number of excellent monographs have been published.[13] This reference also includes a number of sources that provide further reviews that discuss the impact of trace concentrations upon crystallization. In Table 2, the range of initial morphologies, habit modifiers, and the crystal forms that can result from trace concentrations of several different trace modifiers are shown.

TRACE ELEMENTS IN WATER

Drinking water is obtained from surface sources, including lakes and rivers, and from the subsurface aquifers that may be shallow or deep, ancient or relatively recent. Shallow aquifers are often recharged by rainwater at an aquifer recharge zone. Both surface and subsurface waters can be contaminated by toxic elements. Sometimes, surface waters can become contaminated ultimately by subsurface water and can create major environmental and potential health problems.

An example of the contamination of surface water by toxic elements is the case of northern California's problem with selenium in surface waters. This problem is acute in the area south of San Francisco and in the Kesterton National Wildlife Refuge. Groundwater flows through seleniferous formations and is used

Table 2 Habit modification by addition of trace ions or chemicals

Commodity	Normal habit	Habit modifier	New habit
Ammonium alum	Octahedral	Borax	Cubes
NH_4Cl	Dendrites	Cd^{2+}, Ni^{2+}	Cubes
$NH_4H_2PO_4$	Needles	Al^{3+}, Fe^{3+}, Cr^{3+}	Tapered prisms
$(NH_4)_2SO_4$	Prisms	Fe^{3+}	Irregular crystals
$MgSO_4{\cdot}7H_2O$	Needles	Borax	Prisms
$AgNO_3$	Plates	Sodium oleate	Dendrites
Potassium alum	Octahedral	Borax	Cubes
KCN	Cubes	Fe^{3+}	Dendrites
KCl	Cubes	$Fe(CN)_6^{4-}$	Dendrites
	Cubes	$PbCl_2$	Octahedral
K_2SO_4	Rhombic prisms	Fe^{3+}	Irregular needles
NaBr	Cubes	$Fe(CN)_6^{4-}$	Dendrites
NaCN	Cubes	Fe^{3+}	Dendrites
NaCl	Cubes	$Fe(CN)_6^{4-}$	Dendrites
	Cubes	$Na_6P_4O_{13}$	Octahedral
$NaClO_3$	Cubes	$S_2O_6^{2-}$	Octahedral

(From Ref.[4].)

mostly for irrigation. The surplus flow eventually finds its way into the wetlands where it is highly toxic to migratory birds and other wildlife that depend on the wetland for food and protection.

In a similar fashion, arsenic is found in groundwater in many countries including Argentina, Bangladesh, Malaysia, and Thailand, and also in some areas of the U.S.A. The levels of arsenic in the untreated groundwater in some parts of Bangladesh are sufficiently high to cause a number of health problems including keratosis and cancer. Regulations on the levels of arsenic in drinking water have been steadily modified and the acceptable levels have been lowered over the last three decades. Many municipal laboratories are currently incapable of measuring the levels of arsenic below 10 ppb, which is creating problems for regulatory agencies charged with protecting public health.

Aluminum, in the form of alum, is used in the treatment of water to remove trace elements and particulate matter. The reaction involves the hydrolysis of the aluminum ion to form aluminum hydroxide, which is a flocculent precipitate with a very large specific surface area that is useful for the adsorption of ions and also for the collection of microscopic particles that are carried down by the precipitate. In the management of water treatment plants, the kinetics of the reactions are taken into consideration in designing plants so that there is sufficient time to allow all of the aluminum to hydrolyze and precipitate as the hydroxide. When there are upset conditions, aluminum can be carried into the water supply. As the public is concerned about Alzheimer's disease and as there has been some considerable discussion about the involvement of aluminum in the disease, operators of water treatment plants need to be mindful of this transfer reaction. Some water treatment plants use iron(III) chloride (ferric chloride) instead of aluminum. The hydrolysis reaction in this case tends to be less problematic and the precipitation forms more rapidly.

Although water treatment tends to remove any traces of lead from water, lead still appears at the consumer's tap and can be a cause for concern. There are several sources for this lead. In the early years of municipal water distribution, leader pipes, i.e., the pipes that deliver water from the street main to the house, were made from lead. With the development of flexible copper piping, leaders in new houses were constructed of this material. As the technologies used in water treatment have changed, the protective layers of hydroxides and hydroxycarbonates have been eroded. In recent years, water in some areas, such as Washington, DC, has become increasingly contaminated with unacceptable levels of lead. These municipalities are now faced with replacing lead leaders with copper—in part because of increasingly stringent standards for lead in drinking water.

Other trace elements in drinking water become unacceptable because they undergo species transformation during the water treatment process. Both chromium and bromine as chromium(III) and bromide, respectively, are benign species and their presence in drinking water does not pose a hazard. As water is treated in the water treatment plant by strongly oxidizing reagents such as ozone, used to sterilize the water

and render it germfree, chromium and bromine can be transformed into chromium(VI) (as chromate, CrO_4^-) and bromate (BrO_3^-), respectively. Both of these compounds are carcinogenic. Thus, drinking water standards for chromium and bromide are put into effect to control their likely transformation into the more toxic species by chemical treatment.

TRACE ELEMENTS IN FOOD PROCESSING

Trace elements affect foods in a number of ways. Their effects also vary and can diminish or enhance the toxicity of trace concentrations of elements. As a result of the increased level of food processing that is done to increase the stability, shelf life, etc., many nutritionists feel that modern, highly processed foods are in fact missing many essential trace elements and that consumers in developed countries are showing deficiencies in a number of trace elements such as chromium and zinc that they would normally get through the husks, germs, and other parts of plants that are disposed off.

In wines, traces of iron, which are picked up, perhaps, from processing and/or storage, or copper, which are picked up from mildew sprays, such as Bordeaux mixture, affect the oxidative stability of wines by acting as the redox shuttles as they transfer between oxidation states. Winemakers discovered that adding ferricyanide to wine, in a process known as ''blue fining,'' precipitates copper and iron and thereby reduces their concentrations below 1 ppm, which is considered to be acceptable. Critical control of ferricyanide addition is necessary, as cyanide is also a contaminant that must be measured. Where vineyards have replaced cherry and apple orchards, low concentrations of arsenic have started to appear; but they are present at very low concentrations in high quality wines. The arsenic appears from arsenical compounds such as lead and calcium arsenates that were used for many decades as pesticides on apples and cherry orchards.

A bizarre manifestation of trace element interference with food processing was described in the early 1980s by Procter and Gamble food scientists. Their development of a lemon chocolate cake created very negative responses from the tasting panels that were set up to determine the acceptability of new formulae for foods.[15] In this case, the problem was iodine, which was present as iodide in the salt used in the formulation. The iodide was oxidized to iodine during the cooking process, which in turn reacted with one of the flavoring ingredients, cresol. This resulted in the formation of iodocresol, which has a very strong medicinal taste. This taste was responsible for stimulating the gag reflex in some members of the tasting panels. On the effect of iodide in various foods, UNICEF commissioned a report.[16]

It is rapidly becoming clear that trace elements in foodstuffs acquire greater significance when they can undergo important metabolic transformations in which simple inorganic salts are converted by enzyme systems within the organism or the ingestion of previously transformed compounds that may appear in a biomagnification scenario. A number of elements are particularly noteworthy in this regard, such as mercury, arsenic, and tin.

Mercury tends to occur at highest concentrations in fish, particularly the predatory fish at the top of the food chain. The toxicity of mercury compounds, as noted above, varies radically according to the chemical species. It is well known at this point that most of the mercury in fish is present as methylmercury. Bloom[17] reported that more than 90% of the mercury in fish muscle is present as methylmercury. The action of methylmercury is particularly troubling for women in the first trimester of pregnancy. Methylmercury, it is believed, coordinates with cysteine in the blood and this coordinated form can cross the blood–brain barrier, where it can result in a number of serious neurological problems and is implicated in mental retardation in babies. The mother typically exhibits no symptoms. The most infamous case of methylmercury poisoning amongst an adult population was in the Japanese community of Minamata. In the 1950s, methylmercury, a byproduct from the manufacture of acetaldehyde, was discharged into the bay at Minamata where it was rapidly incorporated into fish and shellfish. The effect on the population gave rise to the ''Minamata disease,'' which was characterized by tremors, hallucination, and death amongst the subsistence fishing families living around the bay. Similar but reduced effects were observed in northern Ontario amongst the aboriginal Cree tribes along the English–Wabigoon river systems, which were contaminated by phenylmercury from a local paper mill that had used phenylmercury salts as slimicides. Fish consumption guidelines are issued in many areas where mercury, and hence methylmercury, is a likely health problem. Recognizing the health benefits that derive from eating fish, consumption guidelines are often specific about the frequency of consumption recommended and particular warnings relating to women in the first trimester of pregnancy or women who intend to become pregnant. Dimethylmercury has not been reported as being present in fish, although there are reports of its presence in ultratrace concentrations in mangrove swamps and human breath.

Arsenic is another element that occurs in a variety of seafoods; both vegetable and animal sources contain varying amounts of several organic arsenic compounds as well as traces of inorganic arsenic. Measuring the total arsenic concentration in crab or lobster gives an alarmingly high number, but arsenic is present in the

Table 3 Arsenic species commonly found in seafood

Formula	Name
AsO_4^{3-}	Arsenate [As(V)]
AsO_3^{3-}	Arsenite [As(III)]
$(CH_3)AsO_3^{2-}$	Methylarsenate [As(V)]
$(CH_3)AsO_2^{2-}$	Methylarsenite [As(III)]
$(CH_3)_2AsO_2^{2-}$	Dimethylarsenate [As(V)]
$(CH_3)_2AsO^-$	Dimethylarsenite [As(III)]
$(CH_3)_3As^+CH_2CH_2OH$	Arsenocholine
$(CH_3)_3As^+CH_2CO_2^-$	Arsenobetaine

form of arsenobetaine, a compound whose acute toxicity has been estimated to be similar to table salt. Arsenobetaine is a water-soluble compound and is usually excreted through the urine within 48 hr. Commonly found arsenic species are listed in Table 3.

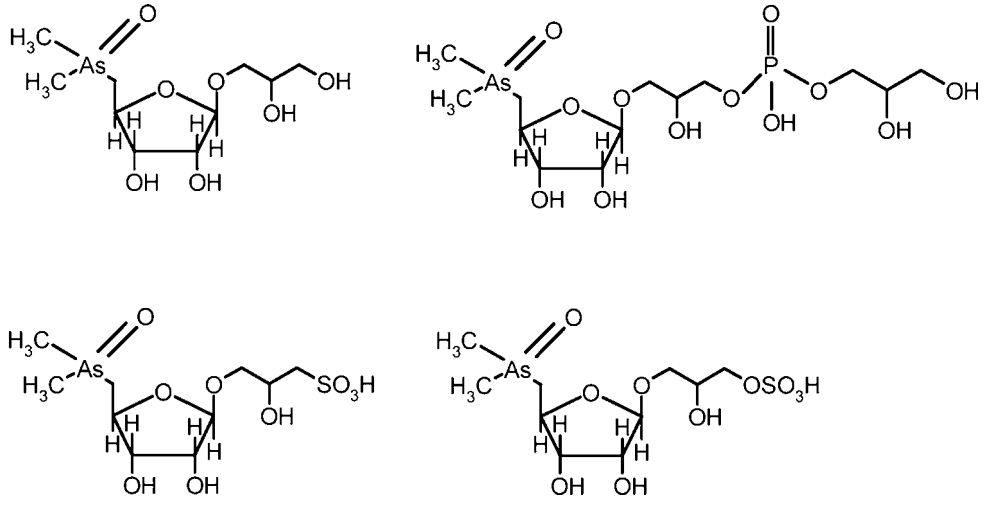
Arsenic also finds its way into the seaweed, which is eaten in a number of foods by the population in Japan, Wales, and Canada. Arsenosugars contain a dimethylarsenic moiety and these sugars also appear in creatures that graze on seaweed. Not all arsenosugars have been identified, but some of the species are illustrated in Scheme 2.

Tin appears in seafood as butylated tin compounds and as methyltin compounds in leachate from plastics. Tributyltin was used for many years to control barnacles and other marine growths on the hulls of ships. These organisms attached to the ships increase the drag on the ships and make them less energy efficient. Tributyltin leaches from the paint into the surrounding waters and can reach levels of concern in confined areas, such as harbors. Filter feeding organisms, such as oysters and mussels, as well as detritus feeders accumulate butyltin compounds and these compounds appear to have estrogenic effects. International food and health organizations have proposed standards between 0.25 and 1.6 μg/K of the body weight for the human population. Although there is not much clinical evidence, researchers suspect that organotins have an adverse effect upon the immune system. Methyltins appear as stabilizers in a wide variety of plastics. But there appears to be no strong evidence to suggest that they migrate readily into food and water that are in contact with plastics, which has been heat-stabilized with organotin compounds. There is evidence that inorganic tin is methylated in the environment.

CONCLUSIONS

Trace elements have huge impacts in many aspects of processing. Trace elements' effects upon crystallization, whether in alloys or in commodity chemicals, are of enormous significance in forming the final product. Our understanding of the effects of trace elements will develop further as the techniques for their determination improve.

Concern about the form in which trace elements appear in food or water will lead to increased demands for techniques that will determine the concentrations of the traces of species of elements. In this discussion, it is clear that the toxicities of different chemical species in which elements can be found vary widely. Although we may feel that the water we drink and the food we eat are well characterized, we should also know that our knowledge of trace elements and their impact at these minuscule concentrations is relatively recent. To protect human and animal health adequately, this area of development will continue to gain importance.



Scheme 2 Arsenosugars found in seaweed and in animals that consume seaweed.

The impact of trace elements in a variety of scenarios shows that we need to remain vigilant about their impact on materials that are processed and used in commerce. The anticipation of ''cradle to grave'' management of materials may provoke increasingly stringent regulations for the disposal of wastes. The effects of trace elements in foods will grow in importance not only from a nutritional perspective but also from a security perspective and the consequences of new regulations upon monitoring and regulation. Although trace elemental concentrations are important, speciation of elements will continue to grow in significance.

REFERENCES

1. http://www.wvgs.wvnet.edu/www/datastat/te/Glossary.htm (accessed May 25, 2005).
2. http://www.netl.doe.gov/cctc/summaries/snox/snoxtmdemo.html (accessed May 25, 2005).
3. Liu, J.; Zheng, B.; Aposhian, H.V.; Zhou, Y.; Chen, M.L.; Zhang, A.; Waalkes, M.P. Chronic arsenic poisoning from burning high arsenic coal in Guizhou, China. Environ. Health Perspect. **2002**, *110* (2), 119–122.
4. National Risk Management Research Laboratory. Mercury in Petroleum and Natural Gas: Estimation of Emissions from Production, Processing, and Combustion, EPA/600/R-01/066; United States Environment Protection Agency Office of Air Quality Planning and Standards: Washington, DC, 2001; 28.
5. Larouk, Z.; Pilkington, R. Creep formation and fracture of a Cr/Mo/V bolting steel containing selected trace element additions. Metal. Mater. Trans. A **1999**, *30A*, 2049–2058.
6. Fleming, H.D.; Ide, R.G. Determination of volatile hydride-forming metals in steel by atomic absorption spectrometry. Anal. Chim. Acta **1976**, *83*, 67–82.
7. Welz, B.; Melcher, M. Determination of antimony, arsenic, bismuth, selenium, tellurium and tin in metallurgical samples using the hydride atomic absorption technique. 1. Analysis of low alloy steels. Spectrochim. Acta B **1981**, *36*, 439–462.
8. Klueh, R.L.; Cheng, E.T.; Grossbeck, M.L.; Bloom, E.E. Impurity Content of Reduced-Activation Ferritic Steels and the Effect on the Reduced-Activation Characteristics. U.S. Department of Energy Office of Fusion Energy Sciences Contract DE-AC05-96OR22464, 1998.
9. Rigby, C; Brindle, I.D. Determination of arsenic, antimony, bismuth, germanium, tin, selenium, and tellurium in 30% zinc sulphate solution by hydride generation inductively coupled plasma atomic emission spectrometry. J. Anal. Atomic Spectrom. **1999**, *14*, 253–258.
10. http://www.nrcan.gc.ca/es/es/energypicture/chap6_e.cfm (accessed May 25, 2005).
11. Brindle, I.D; Brindle, M.E.; Le, X.-C.; Chen, H. Preconcentration by coprecipitation.Part I. Rapid method for the determination of ultratrace amounts of germanium in natural waters by hydride generation—atomic emission spectrometry. J. Anal. Atomic Spectrom. **1991**, *6*, 129–132.
12. Cahn, J.W.; Hilling, W.B.; Sears, G.W. The molecular mechanisms of crystallization. Acta Mettalurgica **1964**, *12*, 1421–1439.
13. http://www.lut.fi/~hhatakka/docit/impure.html (accessed May 26, 2005).
14. Mullin, J.W. *Crystallization*, 3rd Ed.; Butterworth-Heinmann: London, 1993; 255.
15. Sevenants, M.R.; Sanders, R.A. Anatomy of an of-flavor investigation: the ''Medicinal'' cake mix. Anal. Chem. **1985**, *56*, 293A–298A.
16. Westnb, C.E.; Merx, R.J.H.M.; de Koning, F.L.H.A. *Effect of Iodized Salt on the Color and Taste of Food*; UNICEF: New York, PD/95/009, 1995.
17. Bloom, N.S. On the chemical forms of mercury in edible fish and marine invertebrate tissues. Can. J. Fish Aquat. Sci. **1992**, *49*, 1010–1017.

BIBLIOGRAPHY

Hans Löffler, Ed. *Structure and Structure Development of Al–Zn Alloys*; Academie Verlag: Berlin (VCH Publishers, Inc.: New York), 1995.

István Pais, J.; Benton Jones, Jr. *The Handbook of Trace Elements*; St. Lucie Press: Boca Raton, FL, 1997.

Jancic, S.J.; Grootscholten, P.A.M. *Industrial Crystallization*; Delft University Press: The Netherlands (D. Riedel Publishing Company: Dordrecht, The Netherlands), 1984.

Transmission Electron Microscopy for Materials Science

Rolf Erni
Nigel D. Browning
Department of Chemical Engineering and Materials Science, University of California Davis, Lawrence Berkeley National Laboratory, Berkeley, California, U.S.A.

INTRODUCTION

Transmission electron microscopy (TEM) comprises a complete repertoire of imaging, diffraction, and analysis techniques. The combination of direct imaging as well as local structural and chemical analyses makes TEM a powerful tool to study materials down to the atomic level. TEM is used to characterize the microstructure of materials, the constitution of phases and nanoparticles, the local arrangement of atoms, and particularly to study crystal defects such as grain boundaries, dislocations, precipitates, and their impact on the physical properties of the solid. The increasing need to study materials on the atomic scale, which is primarily given by the scaling down of electronic devices, the optimization of catalysts, the development of nanostructures, and also by the goal to obtain a basic understanding of physical and mechanical properties of solids, can in many cases be met by selectively applying one or a combination of experimental TEM techniques.

In a transmission electron microscope, a highly coherent electron beam passes through a thin sample. The electron beam interacts with the sample and is transferred to the specimen's exit plane. The electron wave at the exit plane is magnified in order to form an image or alternatively a diffraction pattern of the sample.

A brief depiction of a transmission electron microscope is provided first. A short section about electron scattering qualitatively describes what types of electron–atom interactions are relevant for TEM. The most common experimental techniques are then explained consecutively. Diffraction mode including nanodiffraction and convergent beam electron diffraction (CBED) are explained first. Direct imaging techniques, such as bright-field (BF) and dark-field (DF) imaging, as well as high-resolution transmission electron microscopy (HRTEM) are dealt with in the following section. The third part of the experimental techniques is about scanning transmission electron microscopy (STEM) and Z-contrast imaging. Finally, analytical methods such as energy-dispersive X-ray spectroscopy (EDS), electron energy-loss spectroscopy (EELS), and energy-filtered imaging are discussed.

MICROSCOPE SETUP

Starting at the top of a microscope column (Fig. 1), two types of electron sources are common, thermionic sources, i.e., W and LaB_6 cathodes, and field-emission guns (FEG). Compared to thermionic sources, field-emission sources show higher brightness as well as coherence and a smaller energy spread,[1] which significantly increases the information limit of the microscope. Below the electron source, the emitted electrons are accelerated. The acceleration voltage U of a microscope defines the primary energy of the electrons and hence the wavelength of the electron radiation. Most microscopes used in materials science are operated between 100 and 400 kV corresponding to an electron wavelength λ between 3.7 and 1.6 pm, which is given by

$$\lambda = 4\pi^2 h \sqrt{2m_e eU\left(1 + \frac{eU}{2m_e c^2}\right)} \quad (1)$$

where h is Planck's constant, e is the elementary charge, m_e is the rest mass of the electron and c is the speed of light in vacuum.

The optical part of the microscope starts below the accelerator. The magnetic electron lenses can be divided into three main lens systems: condenser, objective, and projector lens system (Fig. 1). The condenser lens system, including C1 and C2 aperture, is used to form the illuminating electron beam in front of the objective lens. The objective lens, which focuses the electron beam to the specimen plane, consists of two parts, termed pole pieces. The sample is located between these pole pieces. The objective aperture, mainly used for BF and DF imaging, is in the back focal plane of the objective lens. The selected area aperture, used for selected area diffraction (SAD), is in the first image plane below the objective lens. A stack of projector lenses magnifies the electron wave below the sample. Depending on its setting, either an image or a diffraction pattern is formed on the fluorescence screen or the recording medium. State-of-the-art microscopes are equipped with charge coupled device cameras;[2] however, imaging plates and more frequently films are also used.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030635

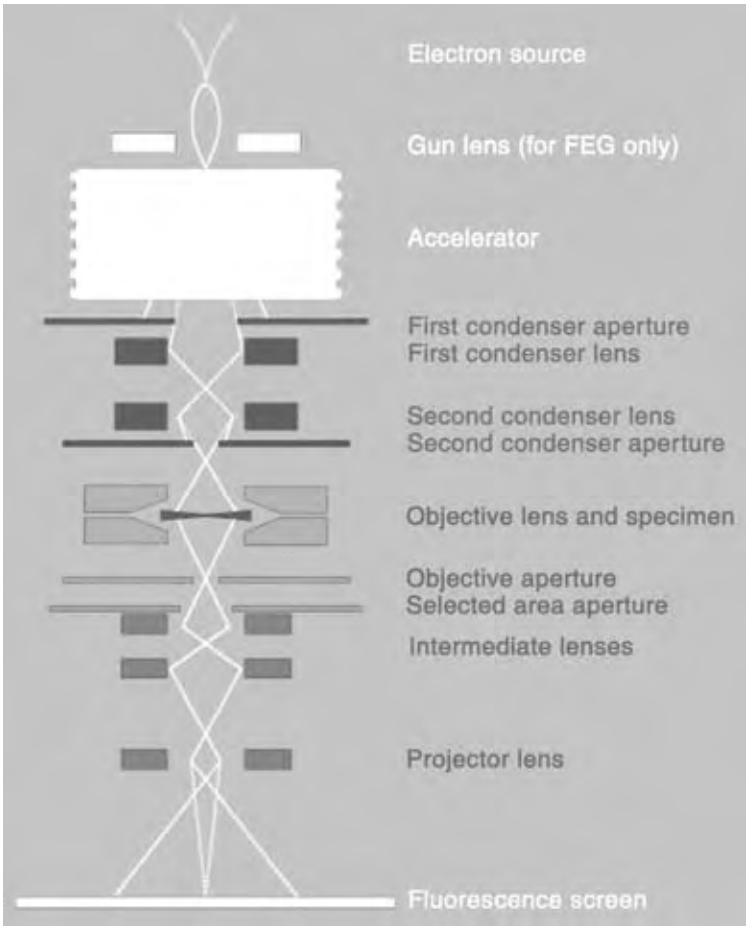


Fig. 1 Microscope setup. The electron beam is indicated as a green line. The projector lens forms either an image (solid line) or a diffraction pattern (dashed line) on the screen. *(View this art in color at www.dekker.com.)*

The resolution in TEM is limited by lens aberrations. In contrast to optical microscopy, where by serially ordering concave and convex lenses, aberrations can be compensated and hence the wavelength of the radiation is resolution limiting; in TEM lens aberrations cannot be compensated since concave electron lenses are not feasible.[3] The objective lens is the crucial part for image defining the microscope's resolution. The quality of the objective lens is described by the constants of spherical C_S(~0.5–3 mm) and chromatic aberration C_C (~1–2 mm). Recently, microscopes equipped with complex correctors for the spherical aberration have become available.[4]

ELECTRON–ATOM INTERACTIONS

Two types of electron–atom interactions have to be considered: elastic and inelastic interactions (Fig. 2). Generally, the interaction between electrons and matter is strong. Thin samples have, thus, to be prepared in order to make them electron transparent.[5] An electron-transparent foil has a thickness between 10 and ~200 nm and for perforated samples, a wedge-shaped thickness profile is typical.

Elastic Interaction

Elastic scattering is a result of the electrostatic interaction between the incident electrons and the atoms in the sample. Electrons elastically scattered on passing through the sample are used to form an image or a diffraction pattern. Elastic scattering of an electron by an atom is described by the elastic scattering factor f_e. The scattering factor is a function of the scattering angle θ, which in the Mott formula is written as:[6]

$$f_e(\theta) = \frac{\gamma m_e e^2}{128\pi^5 h^2 \varepsilon_o} \frac{\lambda^2}{\sin^2\theta}[Z - f_x(\theta)] \quad (2)$$

where γ is the relativistic factor, ε_o is the permittivity of free space, and Z is the atomic number of the element. The function $f_x(\theta)$, whose Fourier transform describes the distribution of the electrons surrounding the nucleus, corresponds to the X-ray scattering factor.[6] Eq. (2) can be interpreted as follows: the momentum of an incident electron is affected by the electrostatic potential of the nucleus, described by the term Z, and by the electrostatic potential of the electrons surrounding the nucleus, described by $f_x(\theta)$. Including the prefactor, each of these terms shows a different θ-dependence. For small and

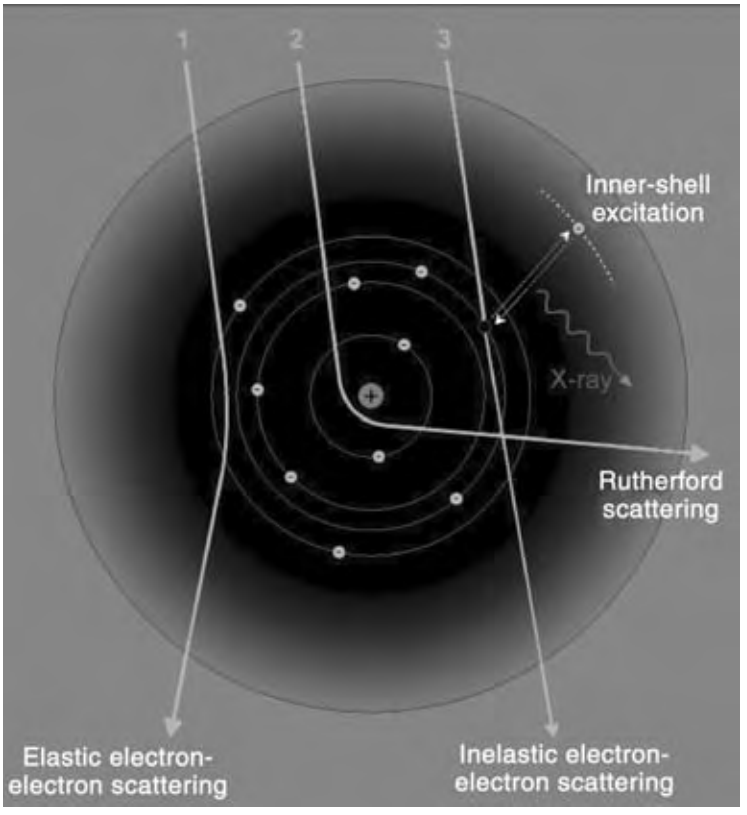


Fig. 2 Electron–atom interactions. (A) Elastic electron–electron interaction dominates the scattering intensity at low and medium scattering angles; (B) Rutherford scattering at the nucleus causes high-angle scattering; and (C) electrons can excite atom-bonded electrons from the ground state to higher unoccupied states or to the vacuum level, element specific X-rays are produced when the excited electron returns to the ground state. *(View this art in color at www.dekker.com.)*

medium scattering angles, scattering is dominated by electron–electron interactions (Fig. 2). However, since $f_x(\theta)$ rapidly decreases with increasing scattering angle, elastic scattering at the nucleus, i.e., Rutherford scattering, becomes important at large scattering angles (Fig. 2). With Eq. (2) and the first Born approximation, the elastic differential scattering cross-section of an atom is given by:

$$\frac{d\sigma_e}{d\Omega}(\theta) = |f_e(\theta)|^2 \quad (3)$$

where Ω is the solid angle and σ_e is the total elastic scattering cross-section. Compared to elastic X-ray $\left(f_x^2 \approx 10^{-26}\right)$ and neutron $(b^2 \approx 10^{-28})$ scattering, the total elastic scattering cross-section for electron scattering is large ($\sigma_e \approx 10^{-19}$). The probability that an electron is scattered more than once has to be taken into account by considering dynamical diffraction.[6]

Inelastic Interaction

Electrons also interact inelastically with the sample. Incident electrons can excite plasmons and phonons, and atom-bonded electrons can be excited from the ground state to higher unoccupied electron states.[7] In the de-excitation process, i.e., when the excited electron returns to the unoccupied core state (Fig. 2), element-specific X-rays are produced which are analyzed by EDS.

Inelastic interactions with atom-bonded, inner-shell electrons cause element-specific electron energy losses. The collective excitation of outer-shell electrons gives rise to energy losses corresponding to the plasmon energy of the solid.[7] The electron energy distribution below the sample can be measured by EELS. Phonon scattering, referred to as quasi-elastic scattering, causes a very small relative energy change (~25 meV). The electron energy distribution is hardly affected by phonon scattering. However, the change of the electron momentum caused by phonon scattering is observable, particularly as thermal diffuse scattering (TDS) in diffraction pattern.[8]

Radiation Damage

Inelastic and elastic interactions between the incident electrons and the sample can cause radiation damage. By elastic interactions, electrons can transfer a certain amount of kinetic energy to a nucleus.[9] The maximum energy that can be transferred, called maximum recoil energy, depends on the primary electron energy and the atomic weight of the element. If the maximum recoil energy exceeds the minimum energy required for displacing an atom, radiation damage, i.e., the creation of point defects, likely occurs.[10] By inelastic

electron–electron and elastic electron–nucleus interactions, electrons deposit energy to the sample. As an additional damage mechanism, the sample can be heated and even be destroyed by the amount of energy transferred to the sample.

ELECTRON DIFFRACTION

In diffraction mode, the projector lens system is adjusted in order to image the electron wave located at the back focal plane of the objective lens. What is seen on the screen is the intensity of this electron wave, which for coherent elastic scattering is called an electron diffraction pattern.

Diffraction Spots, Bragg's Law, and the Reciprocal Lattice

For crystalline samples under plane-wave illumination, a diffraction pattern is observed as a spot pattern. The individual spots depend on the crystal orientation, its structure factor, and obey Bragg's law. Bragg's law states that the difference between the scattered $\mathbf{k}$ and the incident $\mathbf{k}_0$ wave vector is equal to a vector $\mathbf{g}$ of the reciprocal lattice:[6]

$$\mathbf{k} - \mathbf{k}_0 = \mathbf{g} \tag{4}$$

For elastic scattering, the absolute value of the scattered and the incident electron wave vector are equal, given by the reciprocal value of the wavelength, $|\mathbf{k}| = |\mathbf{k}_0| = 1/\lambda$. The angle between scattered and incident wave vector is the scattering angle θ and the vector $\mathbf{q}$, given by

$$\mathbf{q} = \mathbf{k} - \mathbf{k}_0 \tag{5}$$

is the scattering vector. For elastic scattering in a crystal and according to Bragg's law (Eq. 4), only scattering vectors corresponding to reciprocal lattice vectors are allowed, $\mathbf{q} = \mathbf{g}$. The reciprocal lattice is related to the crystal lattice; a general reciprocal lattice vector $\mathbf{g}$ can be written as the sum of primitive vectors, $\mathbf{g} = h\mathbf{g}_1 + k\mathbf{g}_2 + l\mathbf{g}_3$, where h, k and l are independent integers termed Miller indices.[11] The set of primitive reciprocal lattice vectors defines the reciprocal lattice obeying the condition $\mathbf{g}_i \cdot \mathbf{a}_j = \delta_{ij}$, where $\mathbf{a}_j$ is a primitive crystal lattice vector and δ_{ij} is the Kronecker delta symbol ($\delta_{ij} = 0$ for $i \neq j$ and $\delta_{ij} = 1$ for $i = j$).

Ewald Construction

The general relation for elastic scattering in Eq. (5) can be visualized by using the concept of the Ewald sphere.[11] The incident wave vector points in beam direction to the origin of the reciprocal lattice, i.e., $hkl = \{0\,0\,0\}$. For elastic scattering, the set of scattered wave vectors $\mathbf{k}$ form a sphere surrounding the source point of $\mathbf{k}_0$ (Fig. 3). The points of intersection of Ewald sphere and reciprocal lattice form the set of allowed scattering vectors obeying Bragg's law, $\mathbf{q} = \mathbf{g}$. These points are observed as diffraction spots. The orientation of the crystal, its reciprocal lattice and the size of the Ewald sphere define which spots appear in the diffraction pattern. Tilting a sample under an invariant illumination means to rotate the reciprocal lattice around the origin of the reciprocal lattice, whereas the Ewald sphere remains unaltered.

Two points have to be considered when using the Ewald construction for electron diffraction in TEM. First, for the electron wavelength (~2 pm) is much smaller than a typical lattice spacing (~0.4 nm), the radius of the Ewald sphere is much larger than the spacing between nearby reciprocal lattice points, $|\mathbf{k}| \square |\mathbf{g_i}|$. For small and medium scattering angles, the curvature of the Ewald sphere is almost negligible. Second, the reciprocal lattice "points" of a thin foil are elongated perpendicularly to the foil plane and form rods.

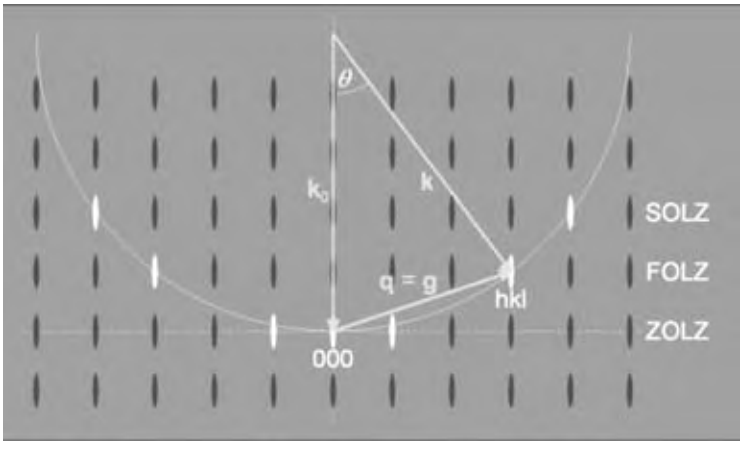


Fig. 3 Ewald construction. The white half-circle indicates the Ewald sphere in two dimensions. The points of intersection between the reciprocal lattice rods and the Ewald sphere form the set of reciprocal lattice points (bright) which obey Bragg's law and appear as diffraction spots in the diffraction pattern. Zero-, first- and second-order Laue zone are indicated. For electron diffraction in TEM, the ratio between the radius of the Ewald sphere and the reciprocal lattice unit is larger than visualized in the figure. *(View this art in color at www.dekker.com.)*

This circumstance is called shape effect.[1] The rods themselves are modulated according to the interference function.[6] What is seen in an electron diffraction pattern are the points of intersection between Ewald sphere, which approaches a plane, and the reciprocal lattice rods (Fig. 3). For small and medium scattering angles, an electron diffraction pattern basically corresponds to one plane of reciprocal space normal to the incident wave vector. This plane containing the origin of the reciprocal lattice is called zero-order Laue zone (ZOLZ). At large scattering angles, the curvature of the Ewald sphere causes the appearance of contributions of higher-order Laue zones (HOLZ).

Kikuchi Lines

Besides diffraction spots, which are caused by coherent elastic scattering, a diffraction pattern also contains contributions of incoherently scattered electrons. Quasi-elastic phonon scattering for instance, which is incoherent scattering, causes TDS, which with increasing sample thickness becomes apparent as a diffuse background. Combined incoherent and coherent elastic scattering gives rise to faint lines. These Kikuchi lines appear in pairs consisting of a deficient and an excess line.[12] Particularly for thicker samples, they can even dominate the contrast features of an electron diffraction pattern. Similar to diffraction spots, Kikuchi lines reflect the symmetry of the crystal.

Selected Area Diffraction

Diffraction spots reveal the symmetry and the spacing of the reciprocal lattice. Since both of them are directly related to the crystal lattice, diffraction patterns recorded for different crystal orientations can be used to determine crystal symmetry and lattice parameters. In order to see sharp diffraction spots, the sample has to be illuminated by a plane wave. A plane-wave illumination warrants that only one incident wave vector $\mathbf{k}_0$ goes into the elastic scattering relation, $\mathbf{k} - \mathbf{k}_0 = \mathbf{q}$. A plane-wave illumination, however, means that the entire sample is uniformly illuminated. The information contained in such a diffraction pattern is not localizable.

For a local analysis of the crystal structure, i.e., for the study of individual grains or selected phases, a selected area aperture can be inserted. Only the sample area selected by the aperture contributes to the diffraction pattern (Fig. 1). A plane-wave illumination, i.e., the appearance of sharp diffraction spots, can thus be maintained (Fig. 4). This technique is called SAD.[13] The smallest area that can be selected is about 0.5 μm given by the smallest selected area aperture.

Convergent Beam Illumination

In case a diffraction pattern of a smaller area is required, which is frequently the case when nanoscale

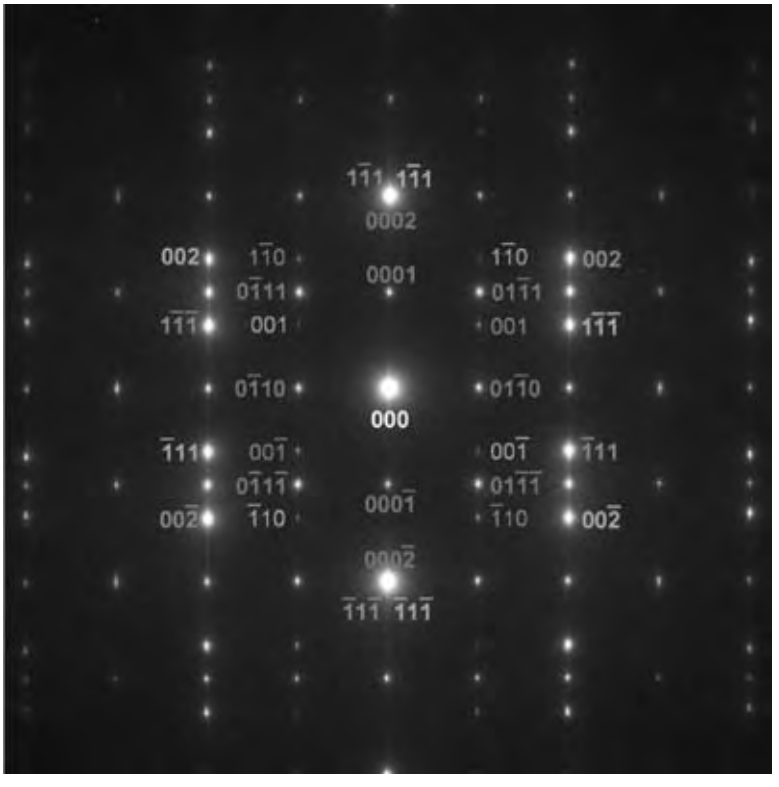


Fig. 4 Selected area diffraction. SAD pattern of uniaxially aligned lamellar γ/α_2 titanium aluminide consisting of hexagonal α_2-Ti_3Al lamellae with $D0_{19}$ structure and tetragonal (slightly distorted cubic) γ-TiAl lamellae with $L1_0$ structure. γ-TiAl is present in two twin variants causing two sets of reflections (green, blue). The spots labeled in red are caused by α_2-Ti_3Al. Main reflections are labeled in bold font, superstructure reflections of the tetragonal phases are in normal font. The interfaces between the individual lamellae are fully coherent which causes the overlap of certain diffraction spots. Incident beam direction for the tetragonal phases is {1 1 0}, for the hexagonal phase {1 1 0}. The streaks in y-direction are caused by the lamellar structure of the material. (*View this art in color at www.dekker.com.*)

materials are studied, a convergent beam is formed, which is focused to a small area. The area illuminated by a convergent beam is in the nanometer range. For a convergent beam, a continuous set of incident wave vectors $\mathbf{k}_0$, each forming an independent Ewald sphere, has to be considered. A convergent beam illumination thus results in the appearance of diffraction disks. Due to the appearance of HOLZ lines (Kikuchi lines of higher-order Laue zones), a convergent beam diffraction pattern contains information, which cannot be accessed by SAD. From CBED, lattice parameters, point and space groups, local lattice strains, and the sample thickness can precisely be measured.[14]

For FEG transmission electron microscopes, it is possible to demagnify the electron beam to a subnanometer electron probe. The diameter of the electron probe determines the sample area contributing to the (nano-) diffraction pattern. Electron nanodiffraction (END) makes possible the recording of a diffraction pattern from areas smaller than 1 nm.[15] Therefore, END is the appropriate technique to study the structure of nanoparticles, as for instance nanotubes and quantum dots. Although a nanoprobe is a convergent beam, the convergence angle is usually smaller than in CBED. As will be shown, END is fundamental for STEM.

IMAGING

Transmission electron microscopy images show a projection of the sample in beam direction. The three-dimensional information is projected to a two-dimensional intensity map. Direct imaging techniques are commonly divided into three groups: 1) conventional imaging covering BF and DF imaging techniques; 2) high-resolution imaging; and 3) STEM, which will be dealt with in the next section.

Electrons passing through a crystalline sample are elastically scattered according to Bragg's law. At the focal plane of the objective lens and of any other lens in the projector lens system, the angular distribution of the diffracted beams is observed. This angular distribution is studied when doing electron diffraction. At the corresponding planes where these beams interfere, an image is formed. An image is thus formed by the interference of Bragg-diffracted beams, however, each diffracted beam contains part of the full image information and can be used to form an individual image. This is the fundamental idea behind TEM imaging.

Diffraction Contrast and Mass-Thickness Contrast

Though each diffracted beam contains part of the full image information, the information contained in the individual beams differs. Defects as for instance dislocations, precipitates, and grain boundaries may locally change the diffraction conditions. Owing to distortions, the Bragg condition may locally be breached and a certain diffracted beam will not be excited from a particular area. Forming an image using this beam would result in low intensity for the area the beam is not excited. This type of contrast is termed diffraction contrast. Additionally, due to the presence of phases consisting of different elements and/or thickness changes, the attenuation of the forward scattered beam (i.e., the $\{0\ 0\ 0\}$ reflection) may locally vary. This position-dependent attenuation of the forward scattered beam gives rise to mass-thickness contrast. If an image is formed with the forward scattered beam only, areas containing heavy elements and thicker areas will show lower intensity.

Diffraction and mass-thickness contrast are both caused by an intensity change of a diffracted beam over the field of view. Since the intensity, specifically the amplitude, of a beam causes these types of image contrast, diffraction and mass-thickness contrast are referred to as amplitude contrast.[1]

Although this concrete explanation of the amplitude contrast is appropriate for a basic understanding, it does not account for all image features. It is based on the kinematical approach to explain electron diffraction. In the kinematical approach, an electron can be scattered once and once it is scattered it will not change its momentum. For very thin samples of light elements, this approximation is sometimes justifiable.

Dynamical Diffraction

Due to the strong Coulomb interaction between electrons and atoms, electron diffraction is generally treated by dynamical diffraction. In the dynamical approach,[6] an electron can be scattered more than once. As a consequence of multiple scattering, the forward scattered beam exchanges its intensity with the diffracted beams and each diffracted beam exchanges its intensity with any other diffracted beam. The intensity of a diffracted beam is thus not a smooth function of the specimen thickness; it is strongly modulated and the modulation period, which is termed extinction distance, depends on the elements in the sample as well as on the particular reflection. For heavy elements, the extinction distance is generally shorter, whereas for light elements, it is larger. Typical extinction distances are in the range of 10–$\sim$200 nm.[16]

Conventional Imaging

Conventional imaging is usually performed at low or medium magnification. Though the interference of all

Bragg diffracted beams forms an image, it is often of low contrast. In order to make use of the amplitude contrast, a single diffracted beam can be selected to form an image. By inserting an objective aperture at the back focal plane of the objective lens (Fig. 1) and selecting one particular beam, an image dominated by amplitude contrast can be formed. This is done by switching to diffraction mode, centering the aperture on the reflection, and switching back to imaging mode. Using the forward scattered beam, which for thin samples has the strongest intensity, is called BF imaging. Forming an image with any other beam, the imaging technique is called DFimaging. Bright-Field and Dark-Field images are in a qualitative way complementary; what is bright in the BF image appears with low intensity in the DF image and vice versa. The complementary application of BF and DF imaging is frequently used to locate (ordered) phases, which show certain (super)structure reflections the bulk material does not show. By choosing a reflection characteristic for a certain phase, the phase that causes the reflection will appear with high intensity and can clearly be located (Fig. 5).

Because the intensity of a diffracted beam is modulated as a function of the crystal thickness, using one diffracted beam to form an image results for a wedge-shaped sample in alternating bright and dark lines (Fig. 6A). The spacing of these thickness contours, depends on the angle of the wedge and the extinction distance of the selected beam.

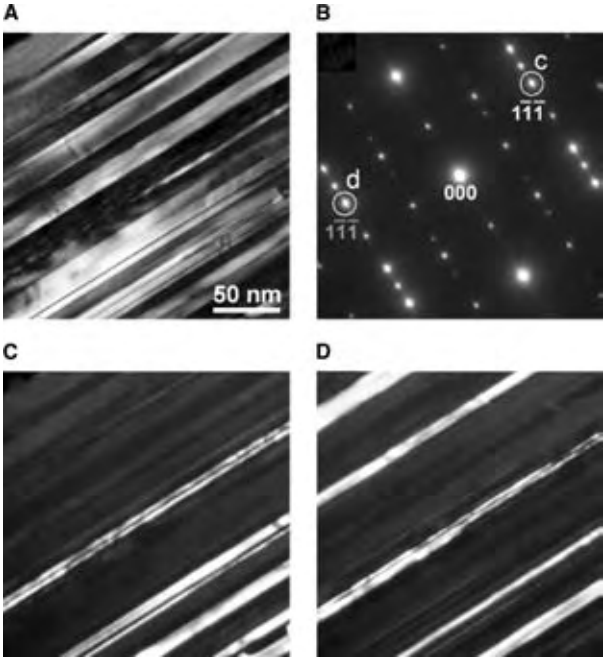


Fig. 5 Bright-field and dark-field imaging: (A) BF image of lamellar γ/α_2 titanium aluminide; (B) corresponding SAD pattern (see Fig. 4); (C) and (D) DF images of the reflections marked in (B). Each of these reflections is characteristic for one twin variant of tetragonal γ-TiAl appearing with high intensity in the corresponding DF image. *(View this art in color at www.dekker.com.)*

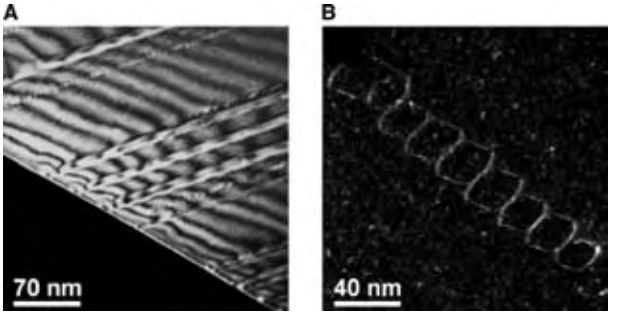


Fig. 6 Weak-beam DF imaging. Weak-beam DF images of: (A) lamellar titanium aluminide showing thickness contours and (B) a spiral dislocation in Al-3 at% Ag, the white speckling is caused by silver-rich Guinier–Preston zones. *(View this art in color at www.dekker.com.)*

Weak-beam DF imaging is an imaging technique which is based on a special diffraction condition, i.e., a two-beam case; besides the forward scattered beam, only one Bragg-diffracted beam is excited which is used to form an image.[1] Because weak-beam DF imaging is highly sensitive to the local diffraction conditions, defects as for instance dislocations can be imaged with high contrast (Fig. 6B).

High-Resolution Imaging

High-resolution transmission electron microscopy can be understood as a general information-transfer process. The incident electron wave, which for HRTEM is ideally a plane wave with its wave vector parallel to a zone axis of the crystal, is diffracted by the crystal and transferred to the exit plane of the specimen. The electron wave at the exit plane contains the structure information of the illuminated specimen area in both the phase and the amplitude.. This exit-plane wave is transferred, however affected by the objective lens, to the recording device. To describe this information transfer in the microscope, it is advantageous to work in Fourier space with the spatial frequency of the electron wave as the relevant variable. For a crystal, the frequency spectrum of the exit-plane wave is dominated by a few discrete values, which are given by the most strongly excited Bloch states,[17] respectively, by the Bragg-diffracted beams.

An ideal information transfer is described by a constant, frequency-independent transfer function with a value of one. A transmission electron microscope is a nonideal information channel, the individual spatial frequencies of the exit-plane wave are differently affected by the transfer and interfere. The complex

transfer process, which is a function of the defocus, lens aberrations, and the degree of coherency, is described by the transmission cross coefficient.[18,19] Both amplitude and phase of the exit-plane wave are transferred, however they are being mixed up. What is finally seen in a high-resolution micrograph (Fig. 7) is dominated by the phase of the exit-plane wave. Therefore, HRTEM is known as phase-contrast imaging. It is the imaginary part of the transmission cross coefficient, called transfer function, which is crucial for the imaging process (Fig. 8).

Due to the highly coherent imaging process in HRTEM, images can be recorded over a certain defocus range; there is not a well-defined value in which the sample is in focus. However, because the ideal case of a plane wave illumination is usually not met, the convergence angle of the incident beam and additionally the energy spread of the electron source constrict the coherency. Owing to this partial coherence, the transfer function is damped at high spatial frequencies (Fig. 8).

Optimizing the transfer function of a microscope means approaching the ideal case of a frequency-independent transfer, i.e., a constant phase shift over a large spatial frequency range. Working at Scherzer condition,[20] which gives a criterion for an optimum defocus $\Delta f_{\text{Scherzer}}$ balancing the effect of spherical aberration C_S against defocus Δf:

$$\Delta f_{\text{Scherzer}} = -\sqrt{\frac{4}{3} C_S \lambda} \tag{6}$$

results in a broad low-pass with almost constant (negative) value. The phase of the diffracted beams is similarly shifted up to a maximum spatial frequency where the transfer function becomes zero (Fig. 8). This first zero-crossing at Scherzer condition defines the microscope's point-resolution:

$$\rho_S = 0.65 \sqrt[4]{C_S \lambda^3} \tag{7}$$

As a consequence of the higher coherence in FEG electron microscopes, information beyond the point resolution significantly contributes to the image. In addition to the point resolution, it is adequate to define an information limit. The information limit corresponds to the highest spatial frequency, which contributes to the coherent imaging process. For FEG electron microscopes, there is a pronounced gap between point resolution and information limit (Fig. 8). However, owing to the strongly oscillating behavior of the transfer function, the information beyond the point resolution is not directly interpretable and suffers from image delocalization. Image delocalization means that image details are displaced from their true locations in the specimen and blurred over a certain area. Image delocalization depends on the defocus and the spatial frequency. In addition to the Scherzer condition, there is another criterion, called Lichte defocus Δf_{Lichte}, which minimizes the overall image delocalization up to a maximum spatial frequency $k_{\max}$.[21]

$$\Delta f_{\text{Lichte}} = -\frac{3}{4} C_S (k_{\max} \lambda)^2 \tag{8}$$

Although image delocalization is minimized when working at Lichte defocus, it is still present and has to be considered when analyzing high-resolution micrographs of FEG microscopes. Numerical methods, such as reconstruction of the exit-plane wave by analyzing focal series of HRTEM micrographs, are used to access the full information up to the information limit.[22]

There are different criteria to optimize the transfer function in HRTEM; the two cited are: Scherzer defocus for a nearly frequency-independent phase shift and Lichte defocus for minimizing image delocalization. The highly coherent imaging process, particularly in the case of FEG microscopes, causing image delocalization and focus-dependent contrast reversal[1] complicates a direct image interpretation. What is seen in a HRTEM micrograph strongly depends on the transfer function, i.e., on the experimental conditions, and therefore, has to be considered thoroughly.

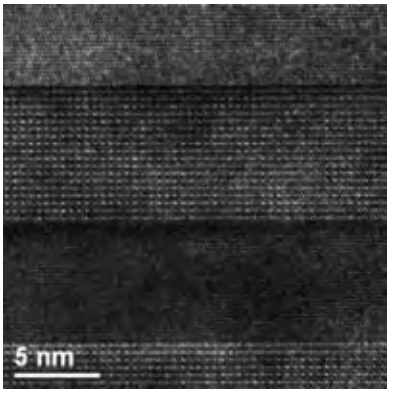


Fig. 7 High-resolution transmission electron microscopy. HRTEM micrograph of lamellar γ/α_2 titanium aluminide. From top to bottom, first twin variant of tetragonal γ-TiAl, hexagonal α_2-Ti_3Al, second twin variant of γ-TiAl and again α_2-Ti_3Al. Incident beam direction for the tetragonal phases is {1 1 0}, for the hexagonal phase {1 1 0}. *(View this art in color at www.dekker.com.)*

SCANNING TRANSMISSION ELECTRON MICROSCOPY

Imaging Process

The incremental way an image is acquired in STEM is fundamentally different from the single-shot imaging

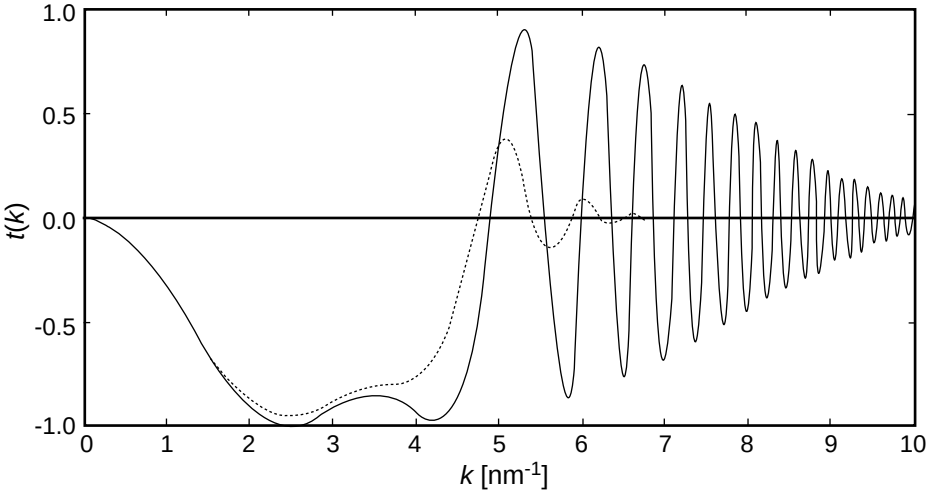


Fig. 8 Transfer function. The microscope transfer function $t(k)$ as a function of the spatial frequency k calculated for Scherzer conditions at 300 kV. Dashed line, microscope equipped with a thermionic source (LaB_6) and C_S = 1.3 mm; full line, FEG microscope, C_S = 1.2 mm.

process in normal transmission mode. In scanning mode, the electron beam is demagnified in the condenser lens system to a small, convergent electron probe (~1–2 Å), which is scanned across the sample. At each scan position, the electron probe is locally scattered on passing through the sample. The microscope is in diffraction mode; i.e., each scan position produces a nanodiffraction pattern of overlapping diffraction disks. On scanning the electron probe, the position of the diffraction pattern remains invariant. Scan detectors, which measure the electron current of a part of its diffraction pattern for each scan position, are either located below the retractable screen or can be inserted above it. Transmitted electrons are detected as a function of the electron-probe position and the scattering angle. The electron current measured for a single scan position reflects the integral intensity of a part of its diffraction pattern. This value is finally represented as the corresponding pixel intensity in the STEM image.

Scan detectors are distinguished according to the scattering-angle in whose range electrons are being detected. Bright-field detectors measure the forward scattered beam, annular dark-field detectors measure the integral intensity of low-order diffracted beams, and high-angle annular dark-field (HAADF) detectors collect electrons scattered to high angles over a large scattering angle range (Fig. 9).

Z-Contrast Imaging

According to the principle of reciprocity,[23] TEM and STEM mode are equivalent in a qualitative way. For instance, what is observed in a HRTEM image is similar to what is seen in a high-resolution BF STEM image. From the principle of reciprocity also follows that a large electron source in TEM mode is equivalent to the use of a large STEM detector that integrates the intensity over a large scattering angle range. A large, although fictitious electron source implies loss of (spatial) coherence and therefore, by using a large electron detector in STEM incoherently formed images can be recorded. This is realized when doing HAADF STEM (Fig. 9). The large HAADF detector area significantly reduces the coherence of the imaging process. Both TDS and the convergence of the electron beam reduce the coherence even further.[24] Contrast reversal and delocalization, caused by the coherent imaging process in (HR)TEM, are therefore absent when doing HAADF STEM. The incoherent nature of HAADF STEM allows for a direct, unambiguous image interpretation. Additionally, since the HAADF detector measures electrons scattered to high angles, Rutherford scattering (i.e., the Z-contribution in Eq. (2) dominates. The signal recorded approaches a Z^2 behavior, where Z is the atomic number of an element. HAADF STEM, also referred to as Z-contrast imaging,[25] is thus a chemical-sensitive imaging technique (Fig. 10).

Spatial Resolution

The spatial resolution in STEM mode is given by the size, i.e., the full width at half maximum (FWHM), of the electron probe, which depends on the electron wavelength, the convergence angle, the defocus, and the constant of spherical aberration C_S of the objective lens. The optimum probe size can be set according to Scherzer incoherent conditions,[20] where the optimum values for defocus $\Delta f_{S,inc}$ and semi-convergence angle $\alpha_{S,inc}$ are given by:

$$\Delta f_{S,inc} = -\sqrt{C_S \lambda} \quad \text{and} \quad \alpha_{S,inc} = \sqrt[4]{\frac{4\lambda}{C_S}}. \tag{9}$$

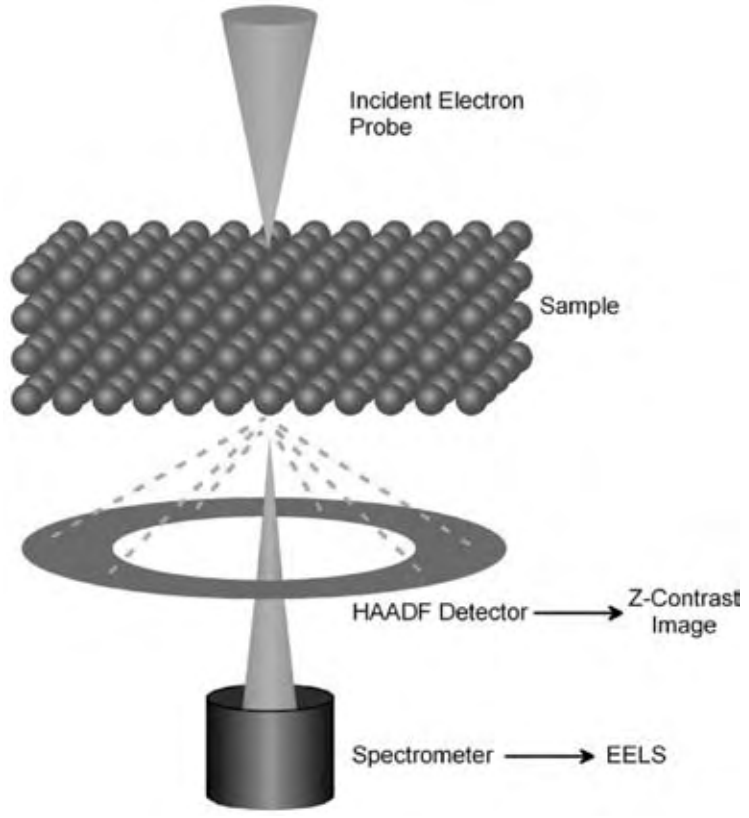


Fig. 9 Z-contrast imaging and EELS. The electron probe is scanned across the sample. For each scan position, the HAADF detector collects the high-angle scattering intensity. The intensity of one scan-position is represented as the corresponding pixel intensity in the STEM image. The forward scattered beam is not affected by the detector and can be used for EELS. *(View this art in color at www.dekker.com.)*

The electron probe size (FWHM), specifically the spatial resolution, the becomes:

$$\rho_{\mathrm{STEM}} = 0.43\sqrt[4]{C_{\mathrm{S}}\lambda^3} \tag{10}$$

For a well-aligned electron probe, which can be done by observing the electron Ronchigram of an amorphous specimen area,[26] atomic resolution STEM images are feasible for FEG microscopes (Fig. 10).

ANALYTICAL TECHNIQUES

The most common analytical techniques used in (S)TEM are EDS and EELS. The combination of STEM and EDS and/or EELS can be used to analyze the specimen locally; when stopping the scan process, the electron probe can be positioned at the point of interest.

EDS analyzes element characteristic X-rays caused by inelastic electron–atom interactions. It is mainly used to measure the composition of the sample. The spatial

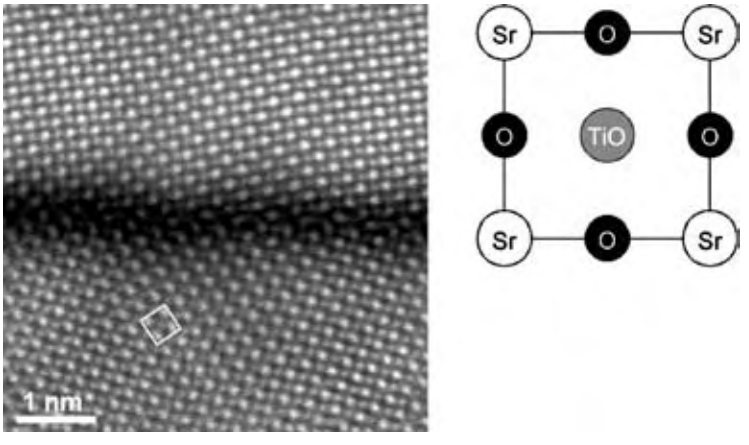


Fig. 10 High-resolution Z-contrast imaging. Z-contrast image of a grain boundary in $SrTiO_3$ (perovskite structure) recorded in {0 0 1} direction. One of the unit cells framed in the micrograph is illustrated on the left, the Sr columns (bright) are at the corner of the unit cell, in the center there is a TiO column. The pure oxygen columns, black in the model, are not observable in the Z-contrast image. The atomic number (Z) contrast is apparent; with increasing atomic number (Z) of the elements, the intensity increases. *(View this art in color at www.dekker.com.)*

resolution of EDS carried out in scanning transmission mode is given by the excited material volume, which in projection is in the nanometer range. EDS lacks the possibility to detect light elements,[1] i.e., elements of atomic numbers smaller than about six.

EELS measures the electron energy distribution below the sample. Two types of spectrometer are common, Ω-spectrometers and post-column energy filters.[7] Ω-spectrometers, located between objective lens and screen, are advantageous when analyzing angle-dependent inelastic scattering in diffraction pattern. As an independent add-on, post-column filters are mounted below the microscope column. Post-column filters use a magnetic prism, which deflects the electron trajectories as a function of their energy. The use of a post-column spectrometer is particularly advantageous when doing Z-contrast imaging in combination with EELS (Fig. 9). The forward scattered beam, which is not affected by the HAADF detector, can be analyzed in the spectrometer with a spatial resolution down to the atomic level.[27]

A typical EEL spectrum consists of the zero-loss peak, caused by elastically and quasi-elastically scattered electrons, plasmon-loss peaks, a downward-sloping background and element-specific ionization edges superimposed on this background.[7] The characteristic ionization edges can be used for a chemical analysis with the advantage compared to EDS that light elements down to He can be detected.[7] Due to the better energy-resolution in EELS (<1 eV) compared to EDS (~100 eV), EELS provides additional information. By studying the fine structure of ionization edges, information about chemical bonds and oxidation states of atoms, as well as the electronic structure of the material can be accessed.

Energy-filtered transmission electron microscopy (EFTEM) combines the analytical capability with imaging. Instead of forming an image with the entire electron distribution, an energy-slit can be inserted in an EEL spectrometer. The energy slit selects electrons with a defined energy loss, which are used to form an image. When positioning the energy slit on the zero-loss peak, images are filtered from inelastic contributions. The energy-slit can also be positioned on element-specific ionization edges. Performing an appropriate background correction, the distribution of an element can be mapped quickly.[28]

CONCLUSIONS

Transmission electron microscopy is an essential tool to study materials and their defects down to the atomic scale. Atomic resolution is feasible in HRTEM as well as in STEM mode, however the real strength of TEM is the combination of different imaging, diffraction, and analysis techniques all able to be performed in a single microscope.

The performance of (scanning) transmission electron microscopes is being steadily improved. Combined STEM/TEM microscopes equipped with correctors for the spherical aberration (already available), and in the near future also for the chromatic aberration (will), improve the spatial resolution in HRTEM. Aberration correctors for the electron probe forming optical part, allow for sub-Å electron probes when working in scanning transmission mode.[29] The improved spatial resolution in HRTEM and STEM is advantageous in studying materials of small lattice spacing containing light elements. However, increasing the spatial resolution is particularly of importance to make (S)TEM a more quantitative tool in materials science. Reconstructed electron exit-plane waves[22] and HRTEM micrographs[4] can be quantitatively analyzed providing information about the composition of individual atomic columns. The combination of sub-Å resolution Z-contrast imaging and EELS[27] is being used for instance to measure the chemical composition of dislocation cores and the intrinsic bonding states related to point defects. Electron tomography[30] on an atomic scale will fill the general lack of (S) TEM, which basically provides projected, two-dimensional information. Microscopes already available equipped with a monochromator allow for sub-100 meV energy-resolution in EELS. The experimentally accessed fine structure of ionization edges can directly be compared with electronic structure calculations. Valence EELS with a monochromated electron beam is used to study band gaps and inter- and intraband transitions locally. Not only are the microscopes being improved, but the repertoire of experimental techniques, not fully covered in this article (but see for example Refs.[1,13]) and the implementation of different techniques in one microscope, are being developed further as well.

ACKNOWLEDGMENTS

We would like to thank Prof. Dr. G. Kostorz, Institute of Applied Physics, ETH Zürich (Switzerland), for providing the titanium aluminide sample and for permitting the use of the corresponding figures in this article. This work was supported by the U.S. Department of Energy under grant No. DE-FG02-03ER46057.

REFERENCES

1. Williams, D.B.; Carter, C.B. *Transmission Electron Microscopy*; Plenum Press: New York, 1996.

2. De Ruijter, W.J. Imaging properties and applications of slow-scan charge-coupled device cameras suitable for electron microscopy. Micron **1995**, *26* (3), 247–275.
3. Spence, J.C.H. *Experimental High-Resolution Electron Microscopy*, 2nd Ed.; Oxford University Press: New York, 1988.
4. Jia, C.L.; Lentzen, M.; Urban, K. Atomic-resolution imaging of oxygen in perovskite ceramics. Science **2003**, *299*, 870–873.
5. Goodhew, P.J. *Specimen Preparation for Transmission Electron Microscopy of Materials*; Oxford University Press: New York, 1984.
6. Schwartz, L.H.; Cohen, J.B. *Diffraction from Materials*, 2nd Ed.; Springer Verlag: Berlin, 1987.
7. Egerton, R.F. *Electron Energy-Loss Spectroscopy in the Electron Microscope*, 2nd Ed.; Plenum Press: New York, 1996.
8. Gjønnes, J. Disorder and defect scattering, thermal diffuse scattering, amorphous materials. In *Electron Diffraction Techniques*; Cowley, J.M., Ed.; Oxford University Press: New York, 1993; Vol. 2, 223–259.
9. Jouffrey, B. Sur quelques aspects des collisions électron-atome: cas élastique et inélastique. In *Cours de l'Ecole de Microscopie Electronique en Science des Matériaux*; Jouffrey, B., Bourret, A., Colliex, C., Eds.; CNRS: Paris, 1983; 85–184.
10. Corbett, J.W. *Electron Radiation Damage in Semiconductors and Metals, Solid State Phys. Suppl. 7*; Seitz, F., Turnbull, D., Eds.; Academic Press: New York, 1966.
11. McKie, D.; McKie, C. *Essentials of Crystallography*; Blackwell Scientific Publications: Oxford, 1986.
12. Cowely, J.M. Electron diffraction: an introduction. In *Electron Diffraction Techniques*; Cowley, J.M., Ed.; Oxford University Press: New York, 1992; Vol. 1, 1–74.
13. Reimer, L. *Transmission Electron Microscopy*; Springer Verlag: Berlin, 1989.
14. Eades, J.A. Convergent-beam diffraction. In *Electron Diffraction Techniques*; Cowley, J.M., Ed.; Oxford University Press: New York, 1992; Vol. 1, 313–359.
15. Cowley, J.M. Application of nanodiffraction. Micron **2004**, *35* (5), 345–360.
16. Humphreys, C.J.; Bithell, E.G. Electron diffraction theory. In *Electron Diffraction Techniques*; Cowley, J.M., Ed.; Oxford University Press: New York, 1992; Vol. 1, 75–151.
17. Metherell, A.J.F. Diffraction of electrons by perfect crystals. In *Electron Microscopy in Materials Science*; Valdrè, U., Ruedl, E., Eds.; Commission of the European Communities: Brussels, 1975; 401–552.
18. Ishizuka, K. Contrast transfer of crystal images in TEM. Ultramicroscopy **1980**, *5*, 55–65.
19. Wade, R.H.; Frank, J. Electron microscope transfer functions for partially coherent axial illumination and chromatic defocus spread. Optik **1977**, *49* (1), 81–92.
20. Scherzer, O. The theoretical resolution limit of the electron microscope. J. Appl. Phys. **1948**, *20*, 20–29.
21. Lichte, H. Optimum focus for taking electron holograms, Ultramicroscopy **1991**, *38*, 13–22.
22. Kisielowski, C.; Hetherington, C.J.D.; Wang, Y.C.; Kilaas, R.; O'Keefe, M.A.; Thust, A. Imaging columns of the light elements carbon, nitrogen and oxygen with sub ångstrom resolution. Ultramicroscopy **2001**, *89*, 243–263.
23. Pogany, A.P.; Turner, P.S. Reciprocity in electron diffraction and microscopy. Acta Cryst. A, **1968**, *24*, 103–109.
24. Nellist, P.D.; Pennycook, S.J. The principles and interpretation of annular dark-field Z-contrast imaging. Adv. Imag. Electr. Phys. **2000**, *113*, 147–203.
25. Pennycook, S.J. Scanning transmission electron microscopy: Z contrast. In *Handbook of Microscopy*; Amelinckx, S., van Dyck, D., van Landuyt, J., van Tendeloo, G., Eds.; VCH: Weinheim, 1997; Vol. 2, 595–620.
26. James, E.M.; Browning, N.D. Practical aspects of atomic resolution imaging and analysis in STEM. Ultramicroscopy **1999**, *78*, 125–139.
27. Browning, N.D.; Chisholm, M.F.; Pennycook, S.J. Atomic-resolution chemical analysis using a scanning transmission electron microscope. Nature **1993**, *366*, 143–146.
28. Hofer, F.; Grogger, W.; Warbichler, P.; Papst, I. Quantitative energy-filtering transmission electron microscopy. Microchim. Acta **2000**, *132*, 273–288.
29. Batson, P.E.; Dellby, N.; Krivanek, O.L. Sub-ångstrom resolution using aberration corrected electron optics. Nature **2002**, *418*, 617–620.
30. Midgley, P.A.; Weyland, M. 3D electron microscopy in the physical sciences: the development of Z-contrast and EFTEM tomography. Ultramicroscopy **2002**, *96*, 413–431.

Tubular Reactors: Reactor Types and Selected Process Applications

T

Patrick L. Mills
Chemical Science and Engineering Laboratory, DuPont Company, Wilmington, Delaware, U.S.A.

Joseph M. Lambert Jr.
Parr Instrument Company, Moline, Illinois, U.S.A.

INTRODUCTION

Tubular reactors are commonly used in laboratory, pilot plant, and commercial-scale operations. Because of their versatility, they are used for heterogeneous reactions as well as homogeneous reactions. They can be run with cocurrent or counter-current flow patterns. They can be run in isothermal or adiabatic modes and can be used alone, in series, or in parallel. Tubular reactors can be empty, packed with inert materials for mixing, or packed with catalyst for improved reactions. It is often the process that will dictate the design of the reactor, as discussed in this entry.

BACKGROUND

Definition

A tubular reactor is a device for conducting chemical transformations of reactants to one or more desired products. Typical products that may be generated with this device include organic or inorganic chemicals, petroleum-derived products, biological materials, and polymers. In the simplest embodiment, a homogenous fluid containing a reactant is introduced into one end of a cylindrical tube at a constant flow rate, where it undergoes a chemical transformation as it is transported through the tube by convection until the product is removed from the other end (Fig. 1). The local rate of reaction of a species, which corresponds to a given position measured relative to the tube inlet, will generally vary with both concentration and temperature. Under ideal conditions, the following three phenomena characterize the reactor hydrodynamics: 1) no fluid mixing occurs in the axial coordinate, i.e., in the direction of fluid flow; 2) complete mixing occurs in the radial direction; and 3) the fluid velocity is uniform across the reactor cross-section.

Application Areas

Commercial-scale tubular reactors are primarily used in the manufacture of petroleum-based products, bulk chemicals, fine and specialty chemicals, pharmaceuticals, and polymers and in environmental remediation. The types of reactions that may be conducted can occur either in the presence or in the absence of a catalyst and may involve more than one phase, such as gaseous phase, liquid phase, and solid phase. When a catalyst is used, it may generally exist as either a soluble organometallic complex or a solid heterogeneous catalyst. The catalyst either can follow the motion of a continuous phase or may be fixed in a place on either an active or an inert support. Examples of reactions that can be realized in tubular reactors include partial and total oxidation, hydrogenation, oxidative or nonoxidative dehydrogenation, hydroformylation, carbonylation, oxidative carbonylation, and polymerization. Tubular reactor types that are used in the above applications include single and multistage adiabatic fixed-bed reactors with either unidirectional or reverse flow, single or multitubular fixed-bed reactors, pipeline reactors, and other special-purpose tubular-type contactors. The scope of applications for these and various other reaction classes has been summarized.[1–3] Several other reviews and the references cited therein provide details of more recent work from both chemistry and process perspectives.[4–8]

Modes of Operation

Tubular reactors can be operated using various modes. These modes also serve to distinguish between various tubular reactor types. The number and type of phases that are present and whether or not a catalyst is used are the criteria for distinguishing between these various modes. The complexity of operation and prediction of reactor performance generally increases when more than one phase is present and when either soluble

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007998

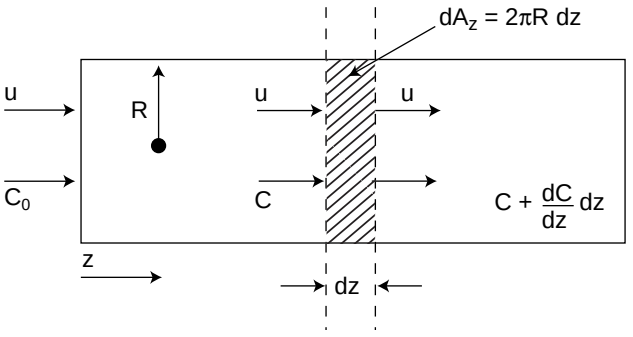


Fig. 1 The ideal tubular reactor.

homogeneous catalysts or solid-phase heterogeneous catalysts promote the reactions.

Homogeneous-phase mode

The simplest mode of operation occurs when the reaction mixture consists of a single phase (i.e., gas, liquid, or solid) during its sojourn through the tubular reactor environment and a catalytic material is not utilized to effect the reaction. The reactor is then referred to as a ''homogenous-phase noncatalytic'' tubular reactor. Some examples of reactions that are realized in this mode are the thermal cracking of hydrocarbons, the noncatalytic combustion of fuels, liquid-phase acid–base reactions, and ethylene polymerization.[9,10] These reactions are typically driven by the formation or the induced generation of one or more radicals that result in a sequence of either elementary or nonelementary steps. The empty tube may also be filled with inert packing to promote radial fluid mixing and heat transfer. Typical inert packings that might be used are illustrated in Fig. 2A–C. In Fig. 2A, nearly all of the commercially available randomly dumped packings are shown in a variety of materials of construction. Those constructed of plastic might find application in liquid-phase reactions where the expected upper temperature limit would not exceed the safe operating temperature of the material. In Fig. 2B, a subset that would be suitable for higher temperature operation, as they are constructed of either metal or ceramic, is shown. In Fig. 2C, several motionless mixers that could be inserted into a pipe or tube are shown. The mixers are designed to promote fluid mixing and heat transfer without the use of moving parts, such as motor-driven agitators. For multiphase mixtures, the mixers are intended to generate local turbulence and interfacial area between the various phases.

When the reaction mixture remains as a single phase, but a catalyst is used to accelerate the reaction, then the reactor is referred to as a ''homogenous-phase catalytic tubular reactor.'' One common application occurs when the reactant is a liquid mixture in which a soluble catalytic material is added to accelerate the conversion of the reactants into the desired product. Esterification reactions in which an alcohol is reacted with a carboxylic acid in the presence of an inorganic acid, such as sulfuric acid, are well-known examples. Pure or multicomponent mixtures of gases can also be used as reactants, and these may react at the active sites of a catalytic material that is deposited on the walls of the reactor tube. The tube can also be filled with a stationary heterogeneous catalyst in the form of a sphere, extrudate, tablet, hollow cylinder, or any other shape (Fig. 3) to form a randomly filled packed bed. Structured packing that has a well-defined geometrical representation, but with surfaces that are washed-coated with a heterogeneous catalyst can also be utilized (Fig. 4). The catalyst may also assume the form of a finely divided powder that is fluidized by either a moving gas phase or a liquid phase,[11] such as those used in fluidized catalytic cracking of crude oil[12] or olefin polymerization.[13] Examples of reactions that use heterogeneous catalysts are quite numerous and include gas-phase hydrogenation, partial oxidation, ammoxidation, catalytic cracking, and emission control chemistries.

Heterogeneous-phase mode

The other distinct mode of operation for a tubular reactor occurs in applications where more than one phase is present in the reaction mixture, e.g., gas and liquid reactants. The products from the reaction can be gases, liquids, or solids where the latter can exist as crystalline or amorphous materials. Either aqueous or organic-based solvents are sometimes included in the reaction medium to control the concentrations of reaction species, to provide increased thermal capacity for highly exothermic systems, or to alter solubility properties for subsequent catalyst recovery or product separations and recovery. This type of reaction is often termed a ''multiphase'' reaction, owing to the presence of multiple interacting phases in the reaction environment. In most practical applications of this mode, either a soluble organometallic complex or a solid heterogeneous catalyst is utilized to transform the reactants into the desired product or products.

When a soluble catalyst is used, the multiphase mixture can be contacted in an open empty tube, and mixing of the various phases occurs through an interaction that is induced by local fluid dynamical phenomena. Alternatively, the tube may be filled with inert packing in an attempt to increase radial mixing for improved multiphase contacting and to promote heat transfer between the fluid and the reactor walls. The inert packing may be either randomly dumped into the tube or inserted in the form of the structured variety as illustrated in Figs. 2 and 4, respectively.

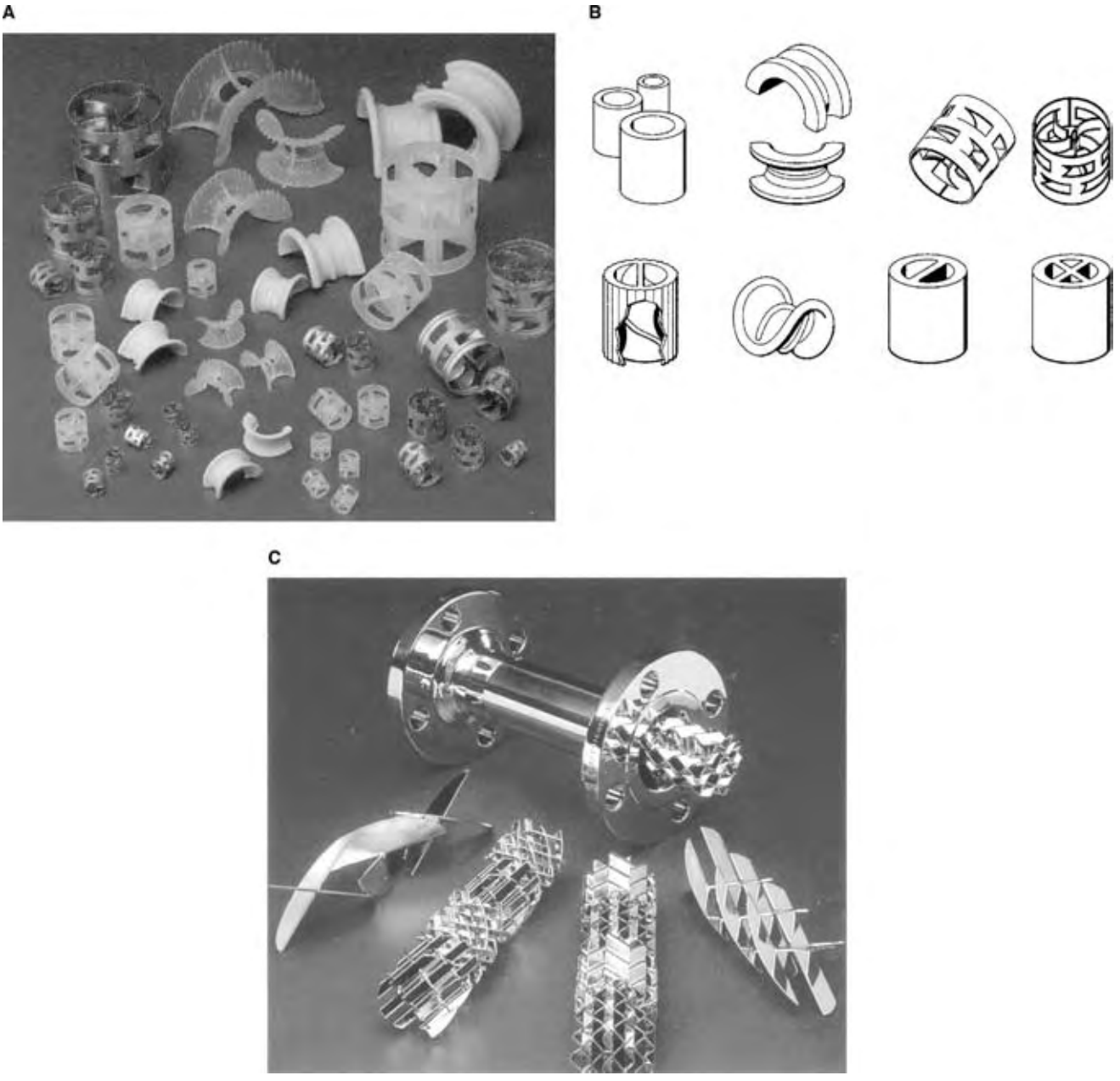


Fig. 2 Examples of randomly dumped inert packings for tubular reactors. (A) Packings for tubular reactors constructed from plastics, metals, and ceramics; (B) packings for tubular reactors constructed from metals and ceramics for higher temperature applications; and (C) inline motionless mixers constructed of metal.

For example, in one version of the Wacker process used for the oxidation of ethylene to acetaldehyde with soluble palladium–copper complexes, a tubular reactor containing ceramic rings is used to promote gas–liquid contacting and to impart a gas–liquid flow pattern that approaches ideal tubular reactor behavior.

Multiphase reactions that occur in the presence of a heterogeneous catalyst can be realized using a variety of tubular reactor types that can be classified according to the state of motion of the catalyst. When the catalyst remains stationary inside the tube and has the shape of a larger particle, such as those illustrated in Fig. 3, then the reactor is generally called a ''heterogeneous-phase fixed catalytic tubular reactor'' or ''multiphase fixed-bed reactor.''

Gas and liquid reactants can be contacted in fixed-bed tubular reactors according to one of several possible so-called ''gas–liquid contacting patterns.'' The three most commonly used contacting patterns are illustrated in Fig. 5. These include trickle-bed reactors (TBRs) with either downward cocurrent (Fig. 5A) or counter-current (Fig. 5B) gas–liquid flow,

Fig. 3 Heterogeneous catalyst shapes.

and packed bubble column reactors with cocurrent upward gas–liquid flow (Fig. 5C). Within this classification, TBRs are the most commonly used reactor in commercial operations as this contacting mode allows various flow regimes to be generated. This reactor is also the most flexible with respect to accommodating variations in liquid and gas flow rates that may be encountered in a manufacturing environment subject to pressure drop constraints. Counter-current flow operation of a TBR with a downflow of gas and an upflow of liquid provides the opportunity for selective removal or stripping of volatile byproducts from the liquid phase that may act as catalyst inhibitors. An example for this is the removal of H_2S and other light sulfur-containing compounds in hydrodesulfurization processes that can otherwise inhibit the local heterogeneous reaction rate. Counter-current gas–liquid contacting can also lead to a larger net driving force for local gas–liquid mass transfer for chemistries where the reaction rate is limited by the gaseous reactant and this species exists as a dilute component in the gaseous feed. Most importantly, with proper catalyst selection, counter-current operation provides the opportunity for in situ separation via catalytic or reactive distillation.[14–22] Upflow packed bubble column reactors (Fig. 5C), or the so-called ''packed bubble-flow reactors,'' are used when it is desirable to ensure a complete external wetting of the catalyst and to achieve the highest possible liquid holdups. The latter condition is desirable for reactions where control of thermal effects may be an issue, owing to the higher heat capacity of the liquid and the cooling effect associated with simultaneous vaporization and condensation.

A variation of the heterogeneous-phase mode occurs when the catalyst is used in the form of a powder so that it more or less follows the motion of the liquid phase. The catalyst powder usually has an irregular shape with a mean diameter ranging from 10 to 400 μm depending upon the particular application, with 50 μm being typical of most operations. Multiphase tubular reactors that use these powdered catalysts are called ''heterogeneous-phase slurry catalytic tubular reactors'' or ''multiphase slurry reactors.'' The more common mode of operation is one where the multiphase contacting between the reactants and powdered catalyst occurs in either a mechanically agitated vessel or a bubble column reactor where the latter has a length-to-diameter aspect ratio from

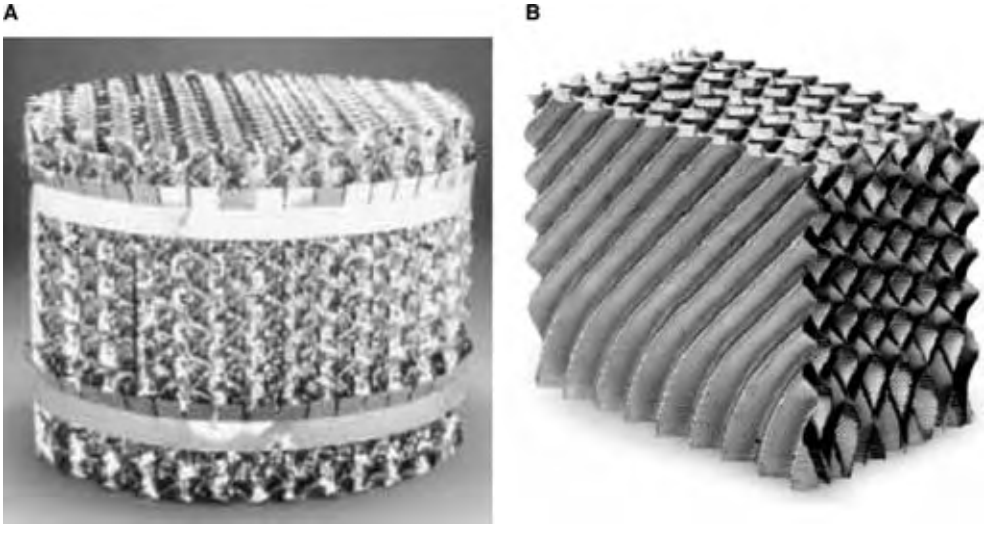


Fig. 4 Structured catalyst packings; A) Metal-Max structured packing and B) Montz-Pak type M structured packing.

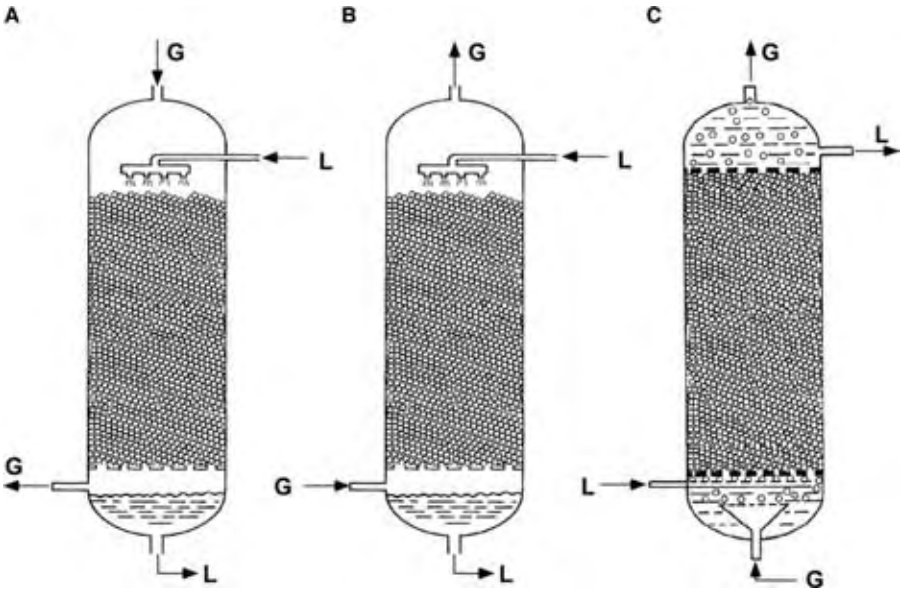


Fig. 5 Fixed-bed reactors with gas–liquid flow. (A) Trickle-bed reactor with cocurrent downflow; (B) trickle-bed reactor with counter-current flow; and (C) packed bubble-flow reactor with cocurrent upflow.

5 to 20. However, other applications exist where the multiphase reactants and powdered catalyst are contacted in tubes with narrower diameter, as part of an attempt to approach the flow pattern of an ideal tubular reactor. In some cases, various types of internals, such as sieve trays, may be added to alter the mixing characteristics of the various reacting phases.

REACTOR DESCRIPTIONS AND PROCESS APPLICATIONS

Introduction

The objective of this section is to provide a brief overview of selected chemistries and processes that are based upon various tubular reactor designs to illustrate more practical aspects. As the partial oxidation process is a key manufacturing technology that utilizes various tubular reactor designs, most of the emphasis will be placed here. The extension of the same concepts to other chemistries, such as hydrogenation reactions, is based upon similar principles.

The least complex class of reactors used for gas-phase catalytic oxidation reactions is adiabatic fixed-bed reactor. The number of processes that employ adiabatic reactors for oxidation processes is somewhat limited. The single-stage or single-bed adiabatic fixed-bed reactor has the simplest design. Also included in the adiabatic reactor class is the multistage design, even though heat is exchanged between adjacent stages. Finally, the most recently developed reactor type in the adiabatic class is the reverse-flow reactor. This reactor has been mainly used in the former Soviet Union to oxidize SO_2 to SO_3.

Unlike catalytic reactors that have heat exchange within the reactor, the distinguishing feature of the adiabatic reactor is that the heat generated by an exothermic reaction goes unchecked. The resulting steady-state temperature profile is a monotonically increasing function of distance from the reactor inlet. The temperature rise asymptotically approaches the adiabatic temperature rise. Temperatures can exceed the adiabatic temperature rise in the unsteady-state adiabatic reverse-flow reactor, which is described here. Because of the potential severity of the heat effects, oxidation reactions carried out in adiabatic reactors form a special subclass. Not only must the catalyst tolerate a temperature rise of several hundreds of degrees, but also that the high temperature must not negatively impact the desired product selectivity. If these two requirements are met, then the designer can take advantage of the high rate of reaction at the high temperature without significant penalty. Moreover, the relative simplicity of the adiabatic reactors, when compared with their cooled counterparts, can be exploited in design, scale-up, and operation.

Commercial manufacturing operations for various inorganic or organic chemicals that are used as chemical intermediates, monomers, or final products are often based upon partial oxidation processes using air, enriched air, or pure oxygen. The reactions upon which these processes are based are conducted in either the liquid phase or the vapor phase, using homogenous or heterogeneous catalysts. The key factors that determine the phase in which the reaction is conducted include the reactant volatility, thermal stability of the reactants and products, specific reaction rate, space–time yield, process safety, and process economics. Examples of inorganic chemicals that are produced using vapor-phase

partial oxidation include nitric acid, sulfuric acid, and hydrogen cyanide. The oxychlorination of HCl is also an example for vapor-phase partial oxidation process. Typical organic chemicals that are commercially manufactured using vapor-phase oxidation processes include ethylene oxide, acrolein, acrylic acid, methacrolein, methacrylic acid, maleic anhydride (MAN), and phthalic anhydride. Ammoxidation reactions, which occur when a mixture of air and ammonia reacts with a hydrocarbon to form a nitrile, are related to these as many of the catalysts used for oxidation are also utilized for this application. A summary of process aspects for selected chemistries is included in the next few sections.

Single-Stage Adiabatic Fixed-Bed Reactors

Description

A typical single-stage adiabatic reactor is shown in Fig. 6. Significant pressure drop can be created as packed bed length increases, the particle size is reduced, or the flow rate is increased. To minimize the pressure drop, a variety of other reactor designs for single-stage adiabatic reactors have been implemented as shown in Fig. 7. These include the simple catalyst bed,[23] the disk or shallow-bed reactor,[23,24] and the radial-flow reactor.[24] The shallow-bed reactor is used in fast reactions where complete conversion can be achieved in a very short contact time. The shallow-bed affords minimal pressure drop, although flow distribution can be an imposing problem.[24] The radial-flow reactor is used to reduce pressure drop as well.

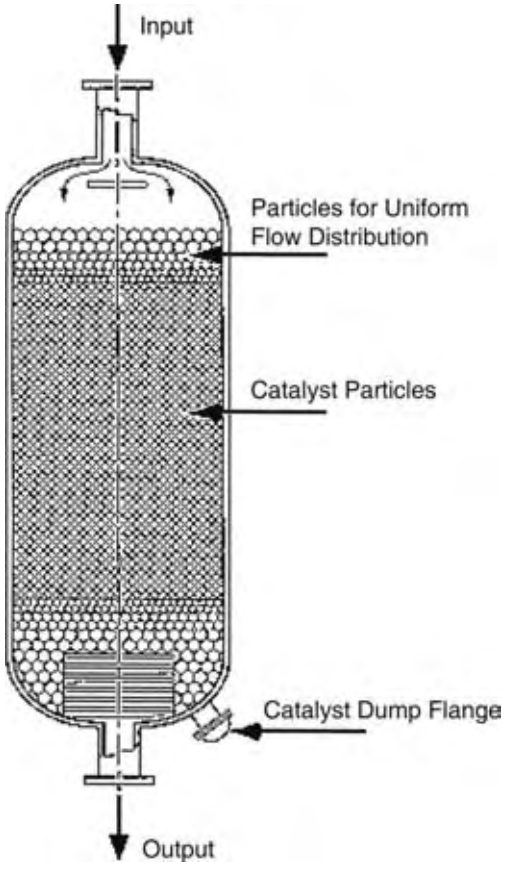


Fig. 6 Typical adiabatic catalytic fixed-bed reactor.

Single-stage adiabatic reactor commercial processes

The oxidation of ammonia to nitric oxide over Pt/Rh is the key step in the manufacture of nitric acid and is realized using a single-stage adiabatic reactor. The partial oxidation of methane to synthesis gas using a millisecond contact-time reactor has not been demonstrated on a commercial scale, but is an interesting example of a novel reactor operating design concept.[25] It is an attractive alternative to the conventional methane steam-reforming route to synthesis gas.

Nitric Acid Process. Nitric acid (HNO_3) is produced by the absorption of nitrogen dioxide (NO_2) in water. The primary route for the manufacture of NO_2 is based on the series oxidation of both ammonia and NO using a feed gas containing excess oxygen. A schematic of a typical high-pressure nitric acid plant is shown in Fig. 8.[26] The oxidation of ammonia is carried out in a reactor containing a Pt/Rh gauze at temperatures between 850°C and 950°C. The following two overall reactions occur:

$$NH_3 + 1.25O_2 \rightarrow NO + 1.5H_2O + 226\ kJ \quad (1)$$

$$NH_3 + 0.75O_2 \rightarrow 0.5N_2 + 1.5H_2O + 317\ kJ \quad (2)$$

Ammonia oxidation can also form nitrous oxide (N_2O) in small amounts. The desired NO producing reaction [Eq. (1)] is extremely fast as the contact time is about 1 msec and is limited by gas–solid mass transfer. For this reason, both the gas linear velocity and flow uniformity are critical issues. The yield to NO is an increasing function of temperature and a slightly decreasing function of total pressure. The NO yield is 98% at atmospheric pressure and 50°C, while it is 96% at 8 atm and 900°C. The loss of precious metal catalyst is an important consideration at higher temperatures.

The high-pressure process reaction operates at 110 psig and offers reduced equipment size at the expense of slightly lower NO selectivity. In the split-pressure process, the ammonia conversion is carried out at an intermediate pressure of about 30 psig, while the water absorption is carried out at higher pressure. The NO that is formed in the ammonia converter

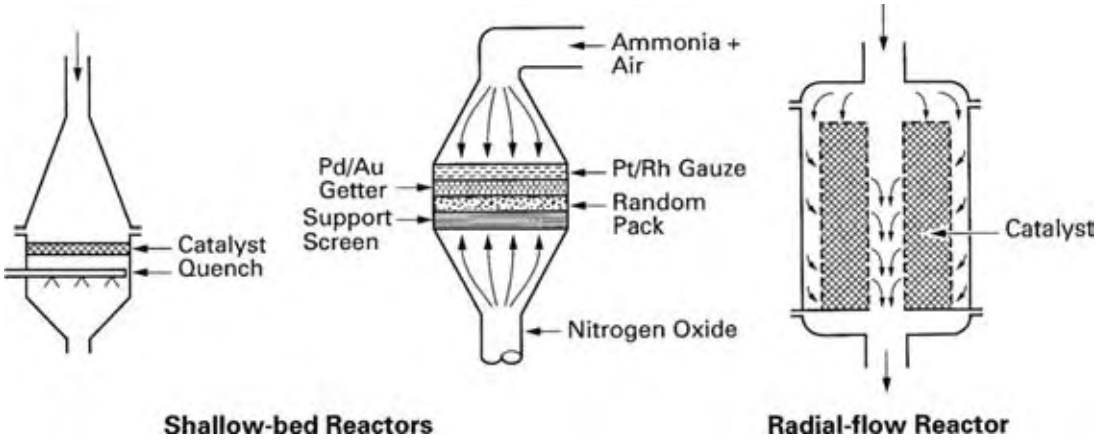


Fig. 7 Three single-stage adiabatic fixed-bed reactor types.

oxidizes further into NO_2:

$$NO + 0.5O_2 \rightarrow NO_2 + 57\ kJ \quad (3)$$

This reversible exothermic reaction is thermodynamically favored at low temperature. Finally, the resulting mixture of NO and NO_2 passes through a condenser and then a cooled absorption column to produce HNO_3. During the reaction between NO_2 and water, NO is formed, which is reoxidized into NO_2.

Multistage Adiabatic Fixed-Bed Reactors

Description

There are two classes of reaction systems for which the single-stage adiabatic reactor is incapable of satisfying the demands that can be placed upon reactant conversion and selectivity. The first class is reversible exothermic reactions. Multiple stages with interstage cooling are required for these reactions to achieve an acceptable conversion level with a reasonable reactor volume. The classical oxidation example is that of SO_2 to SO_3, which is described in more detail below. The second class is the partial oxidation of hydrocarbons in which the desired product selectivity is significantly sensitive to both temperature and the oxygen concentration. More specifically, yield losses to carbon oxides, which result from sequential or parallel side reactions, undermine the goal of achieving high product selectivity at a reasonable hydrocarbon conversion per pass. Most hydrocarbon partial oxidation reactions have this feature. Two illustrative commercial processes include the silver-catalyzed oxidation of methanol to formaldehyde and the catalytic oxidation of monomethylformamide to methyl isocyanate.

Multistage adiabatic reactor commercial processes

Important commercial reactions that are carried out in multistage adiabatic reactors include: 1) the oxidation of SO_2 to SO_3 over a vanadium pentoxide catalyst, which is the key step in sulfuric acid manufacture, and 2) the silver-catalyzed oxidation of methanol to

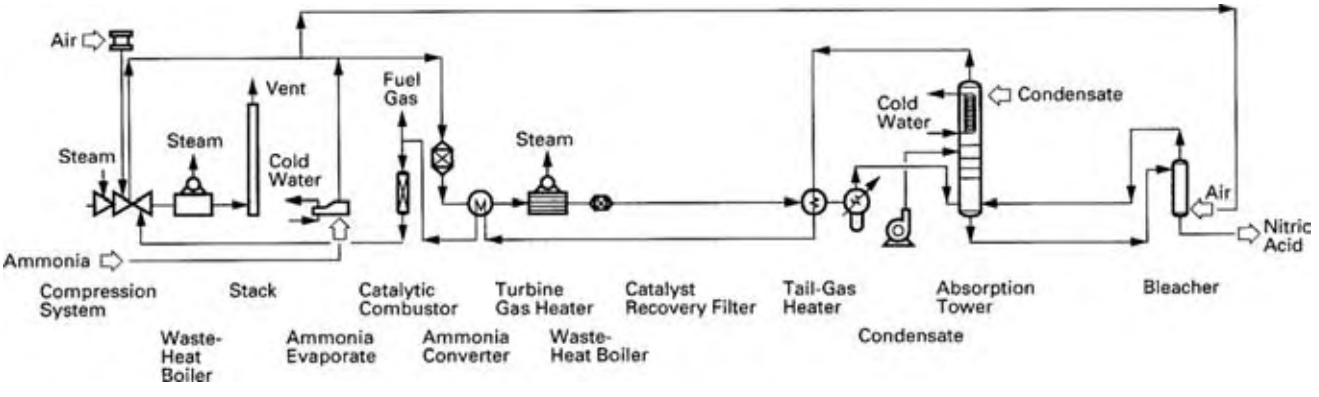


Fig. 8 High-pressure ammonia oxidation process. (From Ref.[26].)

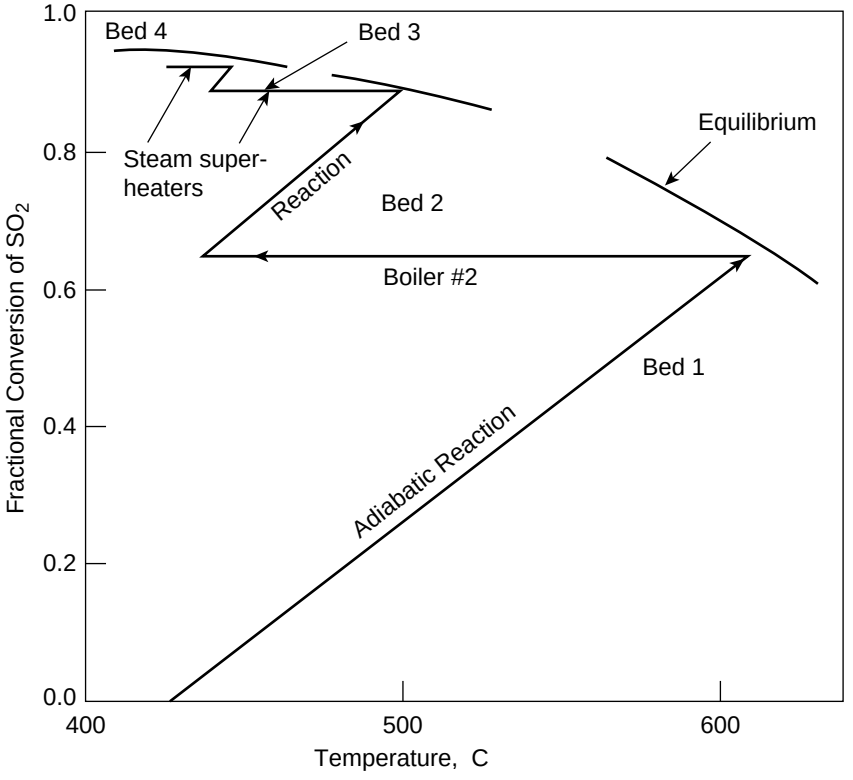


Fig. 9 Typical conversion vs. temperature profile for SO_2 oxidation in a multistage reactor system.

formaldehyde. A summary of the key process aspects for oxidation of SO_2 to SO_3 is given here.

Sulfur Dioxide to Sulfur Trioxide Process. The manufacture of sulfuric acid involves the oxidation of elemental sulfur to SO_2, followed by the catalytic oxidation of SO_2 to SO_3 over vanadium pentoxide. The next step involves the absorption of SO_3 with water to form H_2SO_4. The SO_2 oxidation reaction to SO_3 and the subsequent absorption of SO_3 to form H_2SO_4 are described by the following two reactions:

$$SO_2 + 0.5O_2 \rightarrow SO_3 \quad (4)$$

$$SO_3 + H_2O \rightarrow H_2SO_4 \quad (5)$$

A key attribute of the oxidation of SO_2 to SO_3 is that it is both exothermic and reversible. As shown in the conversion vs. temperature plot in Fig. 9, low temperature

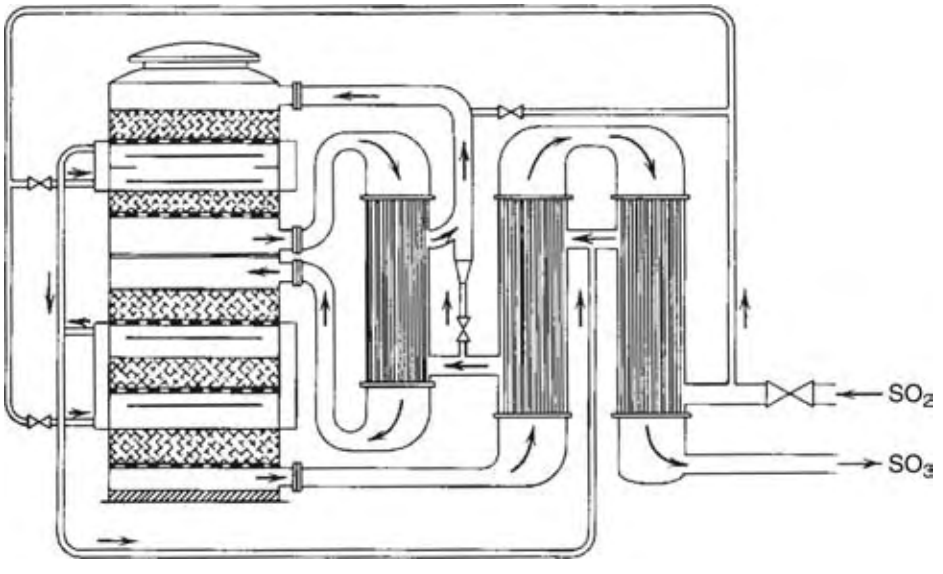


Fig. 10 Multibed adiabatic reactor for SO_3 synthesis. (From Ref.[27].)

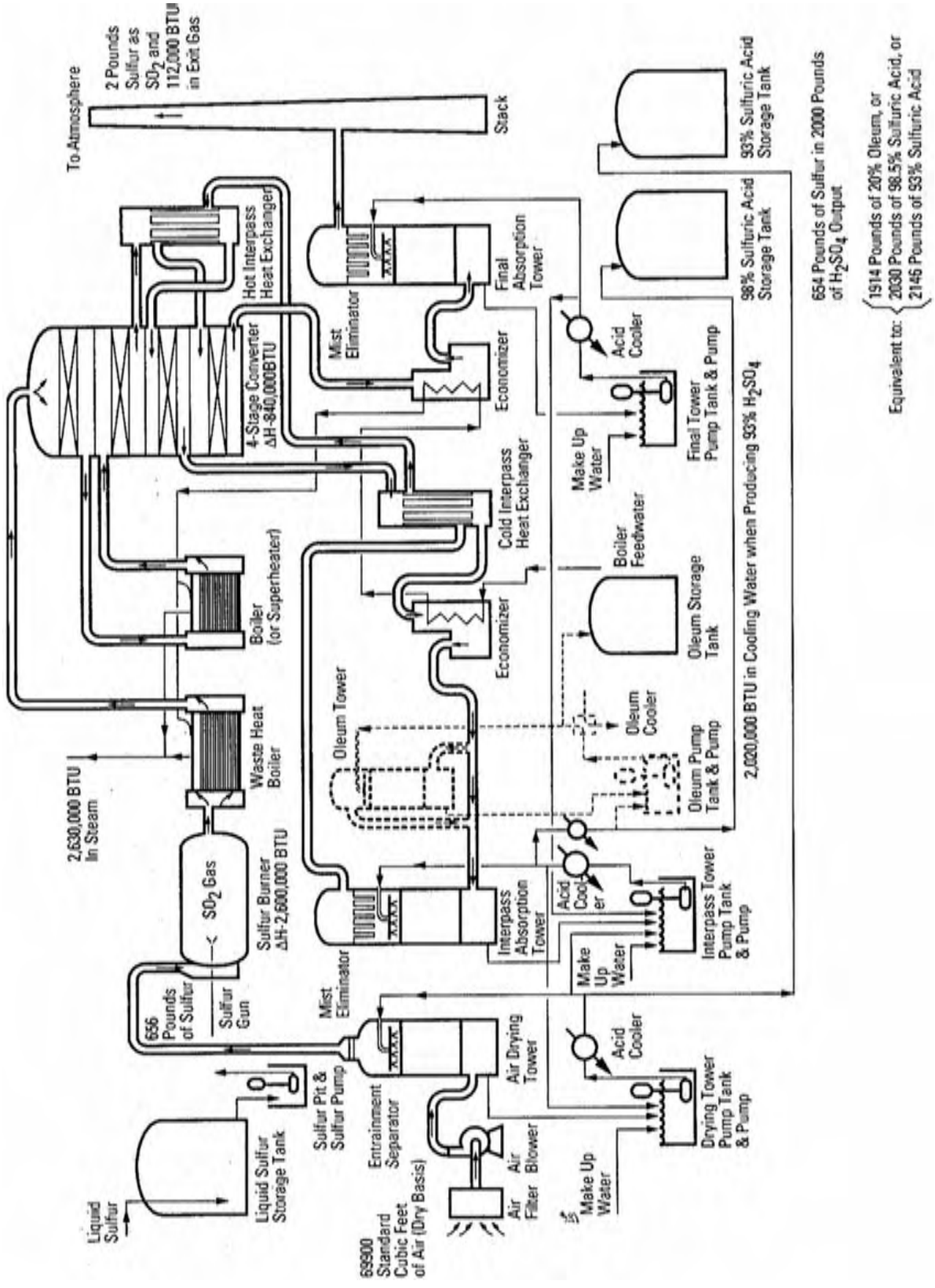


Fig. 11 Process flow sheet for a sulfuric acid process. (From Ref.[28].)

favors equilibrium. This is problematic because the rate of reaction is an increasing function of temperature. For this reason, the reaction is carried out in a multistage adiabatic reactor. In the first stage, a large fraction of the SO_2 conversion is carried out. A high temperature is desirable in the first reactor stage because the feed is free of the reaction product SO_3, which allows the exploitation of the beneficial kinetic effect of a higher temperature. The reactor temperature, which increases linearly with conversion, becomes sufficiently high in the first bed so that the reaction mixture must be cooled to confront the equilibrium constraint. In practice, several stages are employed. Interstage cooling can be accomplished with heat exchangers or by using the so-called "cold shots" of air.

Fig. 10 shows a typical SO_2 multistage reactor system manufactured by Zieren-Chemiebau.[27] This particular design has feed–effluent heat exchangers and cold-shot interstage cooling.

More elaborate multistage reactors have been developed as a result of environmental pressures to reduce sulfur emissions in sulfuric acid plants. One such process is shown in Fig. 11.[28] The final stage of the reactor in this process is fed with a stream in which a fraction of the SO_3 product has been recovered. This enables the final stage to push the SO_2 conversion closer to completion.

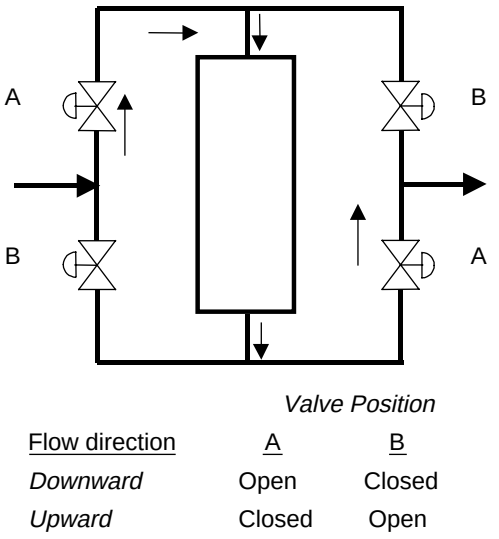


Fig. 12 Schematic of the reverse-flow reactor.

Reverse-Flow Adiabatic Fixed-Bed Reactors

Description

The reverse-flowreactor is a novel reactor design that was originally conceived by Boreskov and Matros[29,30]

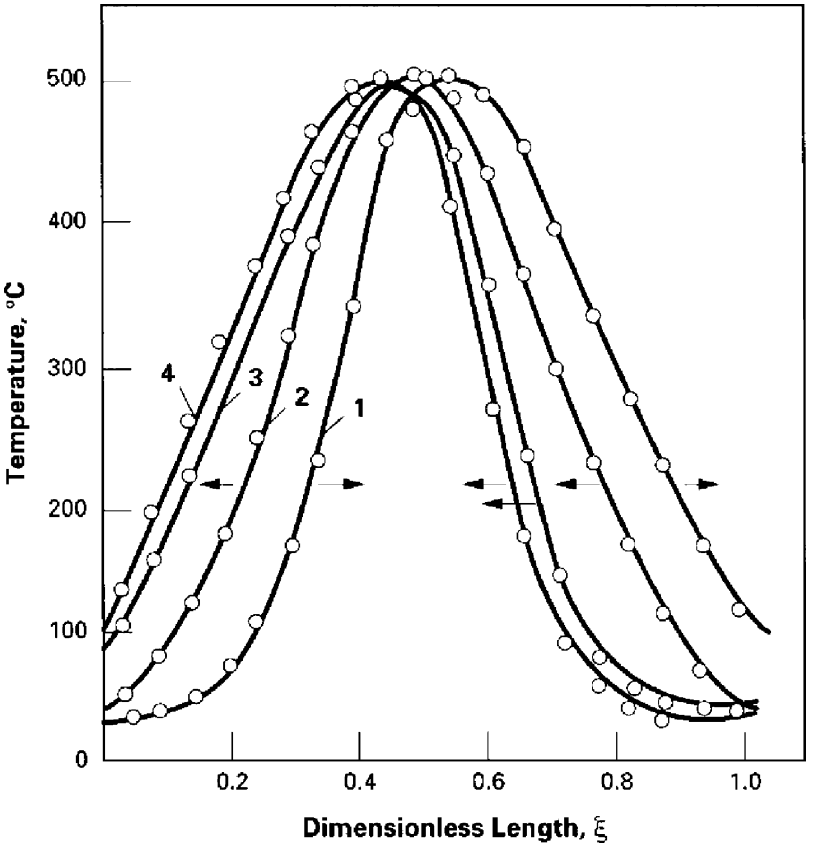


Fig. 13 Measured temperature profiles during SO_2 oxidation in a reverse-flow reactors. (From Ref.[29].)

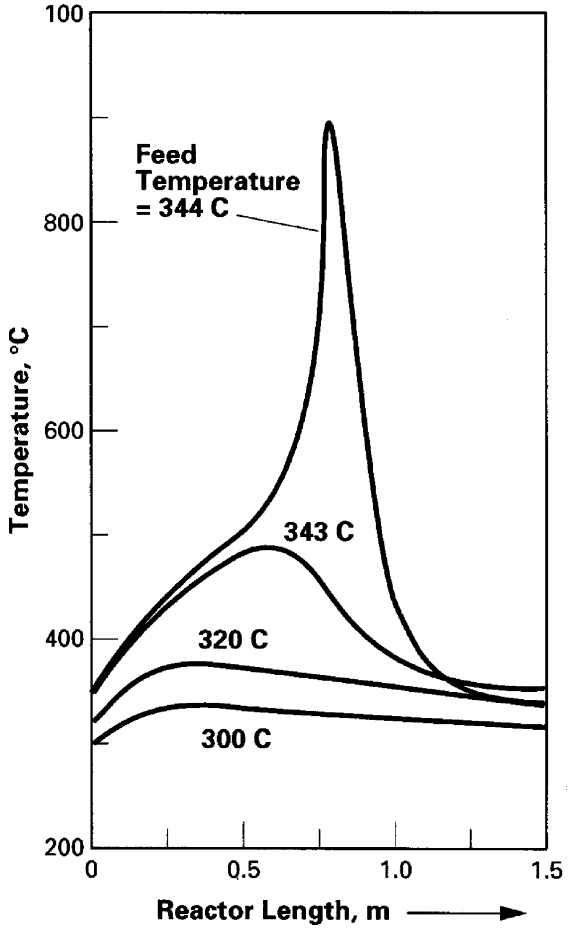


Fig. 14 Effect of reactor feed temperature on predicted temperature profiles for a gas-phase tubular reactor. (From Ref.[24].)

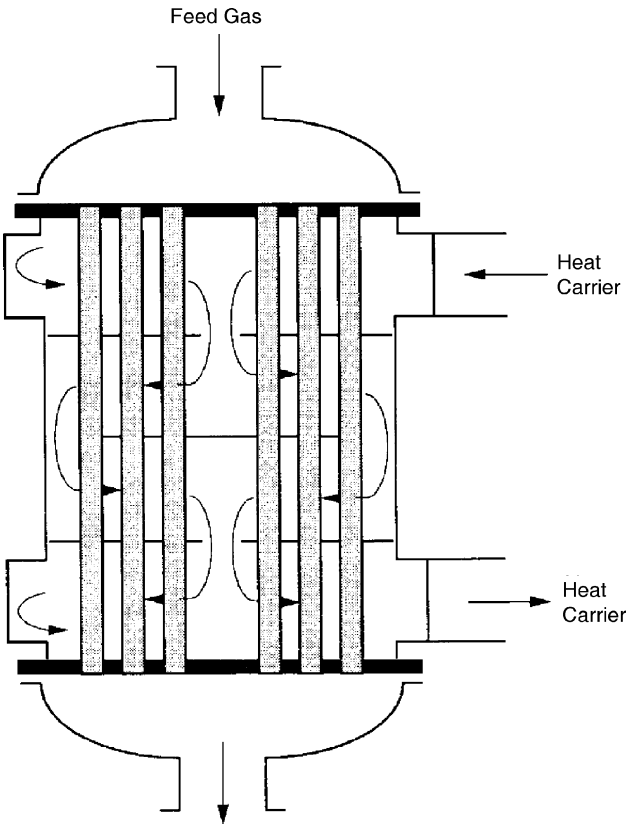


Fig. 15 Simplified schematic of a fixed-bed multitubular reactor.

and Matros et al.[31] As shown in Fig. 12, it is a standard adiabatic fixed-bed reactor, in which the feed gas enters through the opposing ends of the reactor in a deliberate periodic fashion. The main idea is to utilize the exothermic heat of reaction as efficiently as possible within the catalyst bed itself.

Operation of the reactor for pollutant abatement has been described as follows:[32,33] "After a sufficient portion of the packing has been heated to temperatures higher than the ignition temperature, the burner can be turned off. Thereafter, cold polluted air enters into the packing, where the hot bed heats the air so that catalytic oxidation takes place. The introduction of cold air leads to a progressive cooling of the inlet portion of the bed; and, as a result, to a continuous displacement of the temperature front, a so-termed migrating combustion zone. Without further action, the temperature front would move out of the reactor after a certain time, thereafter, the reaction would be extinguished. To prevent extinction, the direction of flow through the fixed bed is periodically reversed with the help of valves. As a result, the portion of packing that has cooled down is heated up again by the combustion zone moving in the opposite direction. Hence, the two end regions of the packing act as regenerative heat exchangers. After a considerable number of flow reversals, a periodically steady state is established in the fixed bed in which, in one-half period, the zone of reaction moves a distance upward, whereas, in the next half period, the zone moves the same distance downward."

The ability to capture the hot spot within the bed by flow reversal relies on the large difference in the characteristic time for convective mass (flow) and conductive energy transport. Flow switching can be easily accomplished, as this occurs on a time-scale that is much shorter than the characteristic time of transit of a creeping hot spot to traverse the length of the reactor.

Reverse-flow adiabatic reactor commercial processes

The reverse-flow reactor can be used to carry out the complete combustion of pollutants contained in air streams and to carry out reversible exothermic reactions. For the former application, the efficient exchange of

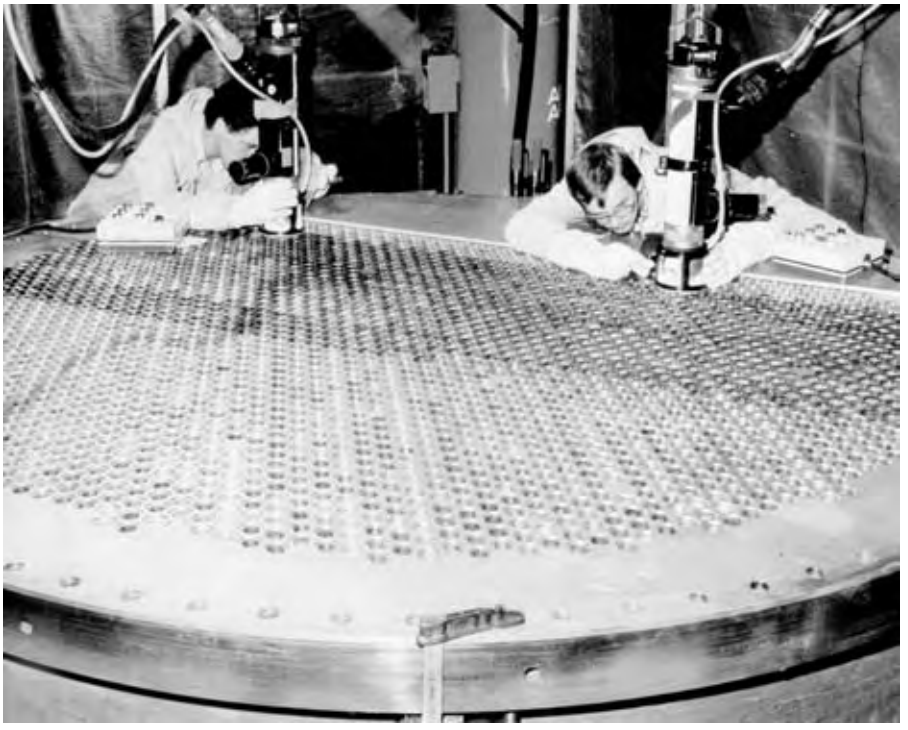


Fig. 16 Multitubular reactor tube sheet.

energy enables the complete combustion of pollutants in small concentrations in air streams with low feed temperature. For the latter application, the pseudosteady-state temperature profile that is established within the bed is close to the optimal profile for achieving high conversion in an equilibrium-limited situation. It is this application that has relevance for oxidation to produce useful chemical intermediates or final products.

Sulfur Dioxide to Sulfur Trioxide Process. In Fig. 13, representative temperature profiles within a bed of vanadium oxide particles during flow reversal for SO_2 oxidation to SO_3 are shown.[29] Profile 1 was measured during flow from left to right through the bed just before the flow reversal. Profiles 2–4 were measured after the flow reversal during flow from right to left. The latter three were measured at various time intervals to show how

Fig. 17 Multitubular reactor used for the air oxidation of *o*-xylene to phthalic anhydride.

they migrate down the bed after the flow reversal. The profiles reveal the very large temperature rise that can be sustained (approximately 400°C) compared with the adiabatic temperature rise (approximately 70–80°C) for the 1.7% SO_2 mixture in air.

Multitubular Fixed-Bed Reactors

Description

Owing to their thermodynamic behavior, almost all partial oxidation reactions require more precise temperature control than that provided by the staged adiabatic reactor. The selectivity of the desired product usually decreases with an increase in temperature beyond a certain practical limit. This behavior suggests that the activation energy of the desired reaction is less than the activation energies for the remaining nonselective reactions. In such cases, ineffective temperature control can lead to hot spots and a subsequent loss in the desired product selectivity. An illustration of the importance of thermal energy management in this type of reactor is provided in Fig. 14. Here, temperature profiles in a gas-phase catalytic fixed-bed reactor, as predicted from a detailed mathematical model for a single reaction, are compared using various feed temperatures.[24] Increasing the feed temperature by 1°C, i.e., from 343°C to 344°C, causes the maximum temperature to increase from 480°C to approximately 900°C. The implications of such sensitive behavior in terms of process operations, process control, and mechanical engineering are obvious.

A simplified schematic of a typical multitubular reactor design is shown in Fig. 15. The mechanical design generally parallels that used for multipass heat exchangers except that the scale is often significantly larger. It consists of a bank of parallel tubes whose nominal inside diameter typically varies from 25 to 40 mm. The tubes are usually laid out on a triangular pitch on a tube sheet where the pitch distance typically varies from 1.5 to 3 tube diameters. A typical tube sheet for a smaller multitubular reactor is shown in Fig. 16. The tube bundle, which is welded on both ends into the tube sheet, is contained inside in a shell through which a suitable heat transfer fluid is circulated. Typical heat transfer fluids include water, pressurized steam, molten salts, and commercially available oils. The multitubular arrangement provides sufficient heat transfer area and reduces the effective radial heat transfer distance by the use of inside tube diameter to particle aspect ratios that are in the order of 2–5. A commercial-scale reactor design used for the manufacture of phthalic anhydride is shown in Fig. 17. Multiple reactors of this type are often operated in parallel when a single large-scale reactor cannot meet the required plant process design rate.

Some of the key attributes of the multitubular fixed-bed reactor are provided in Table 1. The corresponding attributes of the fluidized-bed reactor, which is an alternative reactor type for carrying out oxidation reactions, are also included for comparison.

Multitubular fixed-bed reactor commercial processes

Multitubular reactors are mainly used in gas-phase partial oxidation processes, such as the air oxidation of light olefins, paraffins, and aromatics. Examples of chemistries where these reactors are used include the partial oxidation of methanol to formaldehyde, ethylene to ethylene oxide, ethylene and acetic acid to vinyl acetate, propylene to acrolein and acrylic acid, butane to maleic anhydride, isobutylene to methacrolein and methacrylic acid, and *o*-xylene to phthalic anhydride. An overview of the multitubular reactor process for the partial oxidation of *n*-butane to maleic anhydride is given here.

Table 1 Fixed-bed. vs. fluidized-bed reactors for selective hydrocarbon oxidation reactions

Parameter	Fixed-bed reactor	Fluidized-bed reactor
Hydrocarbon concentration	Below flammability limit	Flammable region possible
Oxygen concentration	Large excess	Near stoichiometric
Temperature control	Hot spot	Nearly isothermal
Catalyst effectiveness	Poor to average	Good
Catalyst attrition	Minimal	Possible problem
Catalyst charging	Complex	Straightforward
Catalyst cost	Least expensive	More expensive
Gas flow pattern	Approaches plug-flow	Flow-regime dependent
Solid flow pattern	Fixed	Flow-regime dependent
Design and scale-up	Well-established	System-dependent
Capital investment	Expensive	Less expensive

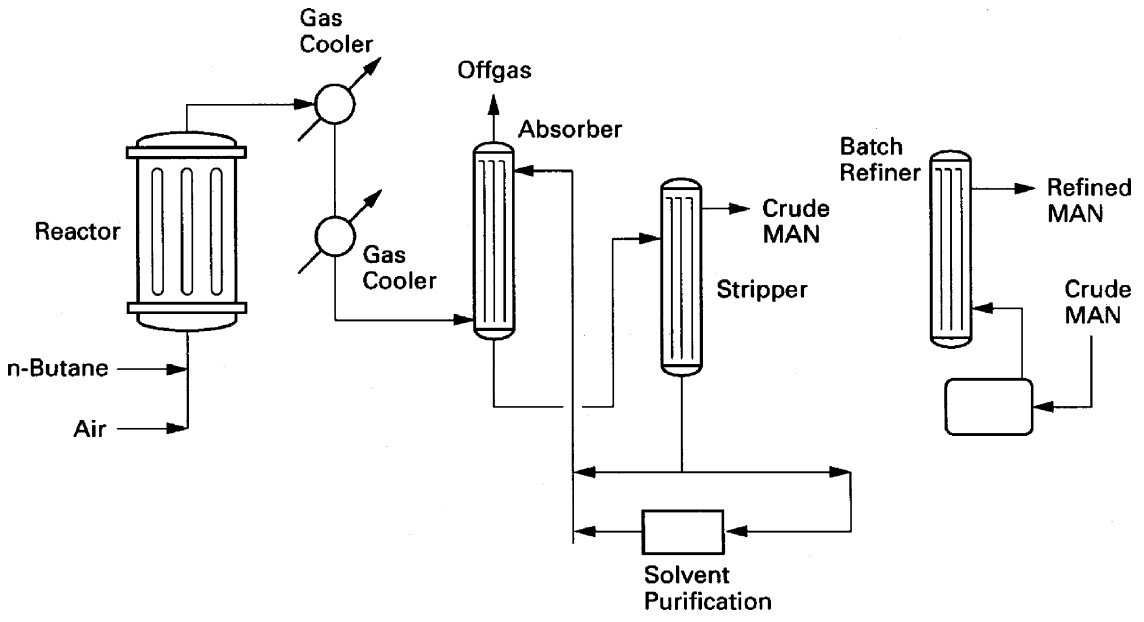


Fig. 18 Process flowsheet for *n*-butane oxidation to maleic anhydride using multitubular reactors. (From Refs.[35,36].)

Butane to Maleic Anhydride Process. Maleic anhydride is an unsaturated diacid that is used in the manufacture of several key products.[34] The chemistry, applications, and process technology of MAN have been described in several comprehensive reviews.[35,36] The greatest use of MAN is in the formation of unsaturated polyester resins. The polymeric resin is formed from the reaction between MAN, ethylene glycol, and a vinyl monomer. The production of MAN is carried out in most of the existing processes by a selective oxidation of *n*-butane over a vanadium–phosphorus oxide (VPO) catalyst. The main selective and nonselective reactions that occur are

$$C_4H_{10} + 3.5O_2 \rightarrow C_4H_2O_3 + 4H_2O + 1236\ \text{kJ} \quad (6)$$

$$C_4H_{10} + 6.5O_2 \rightarrow 4CO_2 + 5H_2O + 2656\ \text{kJ} \quad (7)$$

$$C_4H_{10} + 4.5O_2 \rightarrow 4CO + 5H_2O + 1521\ \text{kJ} \quad (8)$$

The ability to achieve a 60% to 75% yield of MAN at 80% to 90% butane conversion is remarkable given the complexity of the selective reaction, because it involves the abstraction of fourteen hydrogen atoms, the incorporation of three oxygen atoms, and ring closure. Moreover, the selective and nonselective reactions are quite exothermic.

A schematic of the Huntsman fixed-bed MAN process is shown in Fig. 18.[35,36] The process consists of a fixed-bed multitubular reactor, energy recovery units, a MAN absorber/stripper section, and a refinery section. As in other processes discussed in this section, thousands of tubes are used in the fixed-bed reactor to decrease undesirable radial temperature gradients.

Recent advances in fixed-bed-based MAN processes include the development of reactors that can operate in the flammable regime and of the total butane recycle processes that can give overall process yields between 65% and 75%. The ability to safely use butane feed compositions that are above the lower flammability limit of approximately 1.8% in air can translate into increased production rate of MAN, which has obvious advantages from a process economics perspective.

CONCLUSIONS

Tubular reactors have been reviewed from the perspective of various reactor types and selected process applications. The tubular reactor configuration is generally used to contact one or more fluid phases so that high reactant conversion and product selectivity can be achieved in a single pass or by several reactors operating in a series or parallel arrangement. The reactor is generally characterized by a geometry whose length-to-diameter aspect ratio is chosen so that the flow pattern of the fluid, in the case of single-phase operation, approaches ideal plug-flow. The counterpart to the tubular reactor is the continuous-flow stirred-tank reactor whose ideal flow pattern approaches perfect back mixing. Tubular reactors are also widely used for reactions involving multiple flowing phases in the presence of a stationary or moving catalyst, in which case the flow patterns of the individual phases

and the interaction of these phases with the catalyst must be considered as part of reactor engineering issues. Tubular reactors that are used to conduct gas–solid-catalyzed reactions or reactions involving gases and liquids over fixed beds of catalyst always involve complex engineering issues from the perspective of mechanical design, heat transfer, mass transfer, and fluid mechanics. For this reason, each application must be treated on a case-by-case basis, where the interactions between reaction kinetics, transport effects, hydrodynamics, mixing, and heat transfer are accounted for through the use of appropriate reaction engineering models. The engineering science associated with tubular reactor design is well developed for homogeneous single-phase systems, but the heterogeneous systems still have various challenges that are the subject of active research. Use of advanced modeling techniques, such as computational fluid dynamics, along with new flow visualization experimental tools, is generating new insights into both the local and macroscopic phenomena. Tubular reactor design is expected to be more reliable, as these new tools gain more validation.

REFERENCES

1. Doraiswamy, L.K.; Sharma, M.M. *Heterogeneous Reactions*: *analysis, Examples, and Reactor Design. Gas–Solid and Solid–Solid Reactions*; John Wiley: New York, 1984; Vol. 1.
2. Doraiswamy, L.K.; Sharma, M.M. *Heterogeneous Reactions*: *Analysis, Examples, and Reactor Design. Fluid–Fluid–Solid Reactions*; John Wiley: New York, 1984; Vol. 2.
3. Sharma, M.M. Multiphase reactors in the manufacture of fine chemicals. Chem. Eng. Sci. **1988**, *43* (8), 1749–1758.
4. Chaudhari, R.V.; Mills, P.L. Multiphase reactors for fine chemicals and pharmaceuticals. La Chimca e l 'Industria **2000**, *82*, 539–548.
5. Duduković, M.P.; Larachi, F.; Mills, P.L. Multiphase catalytic reactors: a perspective on current knowledge and future trends. Catal. Rev. Sci. Eng. **2002**, *44* (1), 123–246.
6. Duduković, M.P.; Larachi, F.; Mills, P.L. Multiphase reactors—revisited. Chem. Eng. Sci. **1999**, *54*, 1975–1995.
7. Lerou, J.J.; Harold, M.P.; Mills, P.L. Industrial heterogeneous gas-phase oxidation processes. Proceedings of the First Annual NIOK Conference, van Santen, R. World Scientific Publishing: Amsterdam, 1994.
8. Mills, P.L.; Chaudhari, R.V. Multiphase catalytic reactor engineering and design for pharmaceuticals and chemicals. Catal. Today **1997**, *37*, 367–404.
9. Froment, G.F. Kinetic modeling of complex processes. Thermal cracking and catalytic hydrocracking. In *Chemical Reactor Technology for Environmentally Safe Reactors and Products*; de Lasa, H.I, Dogu, G., Ravella, A., Eds.; Kluwer Academic Publishers: Dordrecht, 1992; 409–424.
10. Kirkbir, F.; Kisakurek, B. Dynamic analysis of an ethane cracking reactor. Chemical Reactor Design and Technology, Proceedings of the NATO Advanced Study Institute on Chemical Reactor Design and Technology; de Lasa, H.I., Ed.; Martinus Nijhoff: Dordrecht, 1986; 779–794.
11. Yerushalmi, J. Applications of fluidized beds. In *Handbook of Multiphase Systems*; Hetsroni, G., Ed.; Hemisphere Publishing Corporation: Washington, D.C., 1982; 8-152–8-240.
12. King, D. Engineering of fluidized catalytic crackers. In *Chemical Reactor Technology for Environmentally Safe Reactors and Products*; de Lasa, H.I., Dogu, G., Ravella, A., Eds.; Kluwer Academic Publishers: Dordrecht, 1992; 17–50.
13. Zacca, J.J.; Debling, J.A.; Ray, W.H. Reactor residence time distribution effects on the multistage polymerization of olefins. I. Basic principles and illustrative examples. polypropylene. Chem. Eng. Sci. **1996**, *51* (21), 4859–4886.
14. Beckmann, A.; Nierlich, F.; Popken, T.; Reusch, D.; von Scala, C.; Tuchlenski, A. Industrial experience in the scale-up of reactive distillation with examples from C_4-chemistry. Chem. Eng. Sci. **2002**, *57* (9), 1525–1530.
15. Tuchlenski, A.; Beckmann, A.; Reusch, D.; Dussel, R.; Weidlich, U.; Janowsky, R. Reactive distillation—industrial applications, process design and scale-up. Chem. Eng. Sci. **2001**, *56* (2), 387–394.
16. Krishna, R. Reactive separations: more ways to skin a cat. Chem. Eng. Sci. **2002**, *57* (9), 1491–1504.
17. Taylor, R.; Krishna, R. Modelling reactive distillation. Chem. Eng. Sci. **2000**, *55* (22), 5183–5229.
18. Solokhin, A.V.; Blagov, S.A. Reactive-distillation is an advanced technique of reaction process operation. Chem. Eng. Sci. **1996**, *51* (11), 2559–2564.
19. Stankiewicz, A. Reactive separations for process intensification: an industrial perspective. Chem. Eng. Process. **2003**, *42* (3), 137–144.
20. Schoenmakers, H.G.; Bessling, B. Reactive and catalytic distillation from an industrial perspective. Chem. Eng. Process. **2003**, *42* (3), 145–155.
21. Noeres, C.; Kenig, E.Y.; Gorak, A. Modelling of reactive separation processes: reactive absorption and reactive distillation. Chem. Eng. Process. **2003**, *42* (3), 157–178.
22. Kreul, L.U.; Gorak, A.; Dittrich, C.; Barton, P.I. Dynamic catalytic distillation: advanced

simulation and experimental validation. Comp. Chem. Eng. **1998**, *22* (1), S371–S378.
23. Walas, S.M. Reactors and chemical kinetics. In *Encyclopedia of Chemical Processing and Design*; McKetta, J.J., Ed.; Marcel-Dekker: New York, 1994.
24. Eigenberger, G. Fixed-bed reactors. In *Ullmans's Encyclopedia of Industrial Chemistry, Principles of Chemical Reaction Engineering and Plant Design*; Elvers, B., Hawkins, S., Schulz, G., Eds.; VCH Publishers: Weinheim, 1992; 200–237.
25. Huff, M.P.; Tornianen, P.M.; Schmidt, L.D. Partial oxidation of alkanes over noble metal coated monoliths. Catal. Today **1994**, *21* (1), 113–128.
26. Newman, D.J. Nitric acid. In *Kirk-Othmer Encyclopedia of Chemical Technology*; Mark, H.F., Ed.; Wiley: New York, 1981; 853–871.
27. Froment, G.F.; Bischoff, K.B. *Chemical Reactor Analysis and Design*, 2nd Ed.; Wiley Series in Chemical Engineering; John Wiley & Sons: New York, 1990; 664.
28. Strickland, R.W. Sulfur and sulfuric acid. In *Riegel's Handbook of Industrial Chemistry,* 9th Ed.; Kent, J.A., Ed.; Van Nostrand Reinhold: New York, 1992; Chapter 14, 458–479.
29. Boreskov, G.K.; Matros, Y.S. Unsteady State Performance of Heterogeneous Catalytic Reactions. Catal. Rev. Sci. Eng. **1983**, *25* (4), 551–590.
30. Boreskov, G.K.; Matros, Y.S. Fixed catalyst bed reactors operated in steady- and unsteady-state conditions. In *Recent Advances in the Engineering Analysis of Chemically Reacting Systems*; Doraiswamy, L.K., Ed.; Halsted Press: New York, 1984; 142–155.
31. Matros, Y.S.; Noskov, A.S.; Chumachenko, V.A.; Goldman, O.V. Theory and application of unsteady catalytic detoxication of effluent gases from sulfur dioxide, nitrogen oxides and organic compounds. Chem. Eng. Sci. **1988**, *43* (8), 2061–2066.
32. Eigenberger, G.; Nieken, U. Catalytic cleaning of polluted air: Reaction engineering problems and new solutions. Int. Chem. Eng. **1994**, *34*, 4–16.
33. Nieken, U.; Kolios, G.; Eigenberger, G. Control of the ignited steady state in autothermal fixed-bed reactors for catalytic combustion. Chem. Eng. Sci. **1994**, *49* (24), 5507–5518.
34. Felthouse, T.R. Growing maleic anhydride through its uses and processes. The Chemist **1998**, *75* (5), 22–31.
35. Felthouse, T.R.; Burnett, J.C., Jr.; Mitchell, S.F.; Mummey, M.J. Maleic anhydride, maleic acid, and fumaric acid. In *Kirk-Othmer Encyclopedia of Chemical Technology*; Kroschwitz, J.I., Howe-Grant, M., Eds.; John Wiley & Sons: New York, 1995; 893–928.
36. Felthouse, T.R.; Burnett, J.C., Jr.; Horrell, B.; Mummey, M.J.; Kuo, Y.-J. Maleic anhydride, maleic acid, and fumaric acid. In *Kirk-Othmer Online*; Seidel, A., Ed.; John Wiley & Sons: New York; 2001.

Twin-Screw Extrusion

Paul Andersen
Coperion Corporation, Ramsey, New Jersey, U.S.A.

INTRODUCTION

The term ''Twin-screw Extrusion'' encompasses a wide range of material process applications. Likewise, these applications are not all processed on a single type or configuration of twin-screw extruder, but utilize a variety of twin-screw equipment design concepts. For example, a vinyl profile such as might be used for a window casement, or a wood fiber composite decking board, can be directly extruded from a premix formula using a conical counter-rotating twin-screw extruder.[1] Wood fiber composites are also compounded and extruded as deck and other profiles from parallel counter-rotating fully intermeshing twin-screws, as well as co-rotating intermeshing twin-screws with a single-screw discharge.[2,3] Many breakfast cereals are cooked and extruded from co-rotating units.[4] Additionally, thermoplastic polyurethane is polymerized and compounded, solvent is devolatilized from polymerization products, and water is removed from the washed recycled polymers on the co-rotating twin-screw.[5–7] Also, highly filled products such as detergents, ceramics, fertilizer, nanocomposites (both masterbatch and full recipe formulations), polymer magnets, and propellants (explosives) are compounded on the co-rotating machine.[8–11] The above examples are in addition to the more traditional mixing/compounding operations for producing filled and reinforced plastics, color and additive masterbatch, blends and alloys, thermoplastic vulcanizates (TPV), powder coating, toner, adhesives, sealants, and curable rubber.[12–19] The long list of processes already noted only scratches the surface of the possible applications for a twin-screw extruder. Unfortunately, there is a tendency to treat the various twin-screw geometries as one general category. This is very misleading. Each type of twin-screw extruder has a very distinct operating principle (mechanism) and set of process applications where it is successfully implemented. There are also many cross-over applications. This entry will review the various types of commercial co-rotating and counter-rotating twin-screw extruder systems and define the operating principles and processing characteristics of each type. Wherever possible, process examples will be included for illustration.

EXTRUSION BASICS

Before going into details of the different machine geometries and operating mechanisms, it is useful to look at the common features. Each contains a motor, coupling, gearbox, process section, and shaping device, such as a strand or profile die (Fig. 1). This die is either directly coupled to the end of the process section or an intermediate pressure generation unit such as a single-screw extruder. The process section is built up from modular components (both barrels and screw elements). The only notable exception is the counter rotating profile extruder (both parallel and conical) that normally has a single piece barrel and a solid screw. The process section for a typical compounding line can normally be broken down into unit operations that would include:

- Introduction of the feed material
- Solids conveying
- Melting (softening, phase transformation)
- Mixing (dispersive, distributive)
- Additive incorporation
- Degassing/devolatilization
- Discharge pressurization
- Discharge shaping.

TYPES OF TWIN-SCREW OPERATING MECHANISMS

There are several basic design characteristics that provide differentiation among twin-screw extruders. The two most obvious are co-rotating vs. counter-rotating, and intermeshing vs. partial and non-intermeshing. Parallel vs. conical shafts is also a variable, but there is really only one type of conical machine, that is, a counter-rotating, fully intermeshing design.

Fig. 2 shows the cross-section and profile of both the counter and co-rotating intermeshing geometry. The counter rotating design has a closed chamber in the intermesh. Therefore this system is crosswise and lengthwise closed (i.e., there is no path for material transfer from chamber to chamber either along the screw axis or from screw to screw). Consequently, material is force conveyed down the barrel in ''C'' shaped sections. This segregated forced transport translates into good conveying and pressure build up characteristics, but reduced longitudinal mixing (residence time distribution). The co-rotating system has an open intermesh area. Therefore, material is transferred from screw to screw, but each time with a down channel offset with respect to the starting point.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120007999

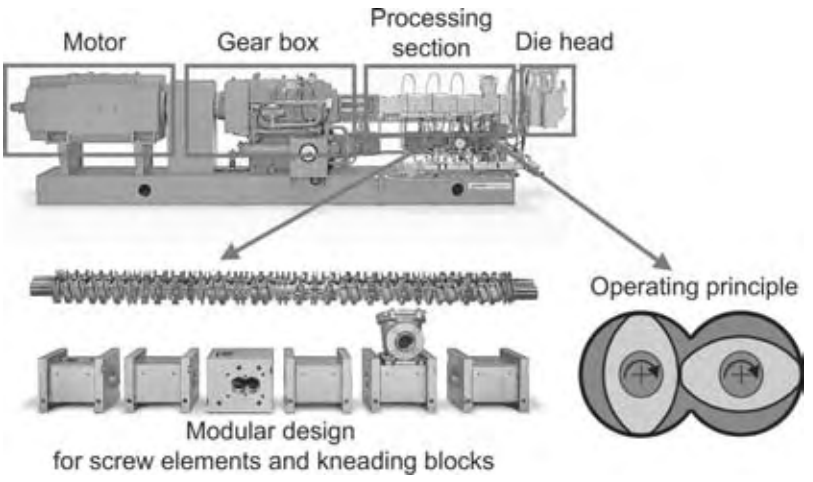


Fig. 1 Basic layout and main components of the twin-screw extruder with drive power available from 10 kw to 12 Mega-watts and rates from 5 kg/hr to 75 tons/hr. (Courtesy of Coperion Corporation.) *(View this art in color at www.dekker.com.)*

This means that there is a continuous path along the screw axis from inlet to discharge. However, there is no path for material to flow from channel to channel along a single- screw shaft for screw bushings, but there is one for kneading blocks and other special elements. Therefore, the co-rotating fully intermeshing geometry is lengthwise open, but depending on screw profile (screw elements vs. kneading blocks) it is either crosswise closed or open. Compared to the counter rotating geometry this flow mechanism has reduced conveying characteristics, reduced pressure build up capabilities, but good longitudinal mixing. Non-intermeshing counter-rotating systems are always open lengthwise and crosswise and therefore have good distributive mixing or homogenization characteristics, but an extremely poor pumping efficiency.

CO-ROTATING FULLY INTERMESHING SYSTEMS

Geometry

The main degrees of freedom in twin-screw extruder design are geometry (cross-section profile of screw pair), power (torque capacity) and speed (rpm). The essential geometry of any co-rotating twin-screw extruder is defined by two key parameters: 1) centerline distance [a] between the shafts, and 2) outer diameter to inner (root) diameter ratio [OD/ID], see Fig. 3. For a fixed centerline, the OD/ID ratio defines the free volume of the extruder.

The co-rotating twin-screw extruder is commercially available with both 2- and 3-lobe cross-section geometry

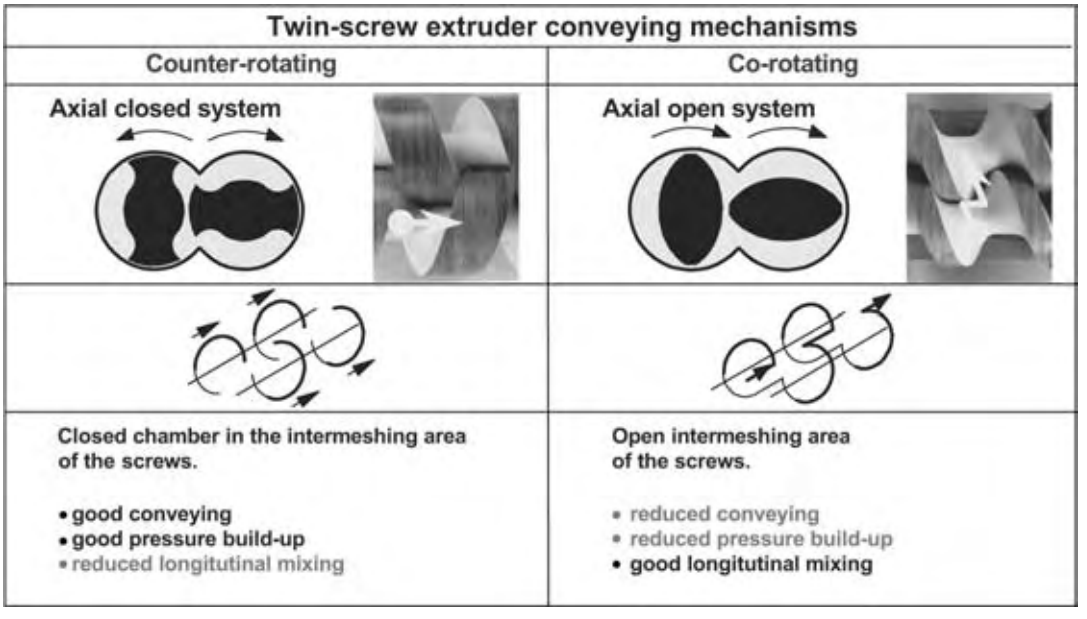


Fig. 2 Comparison of twin-screw conveying mechanisms. (Courtesy of Coperion Corporation.) *(View this art in color at www.dekker.com.)*

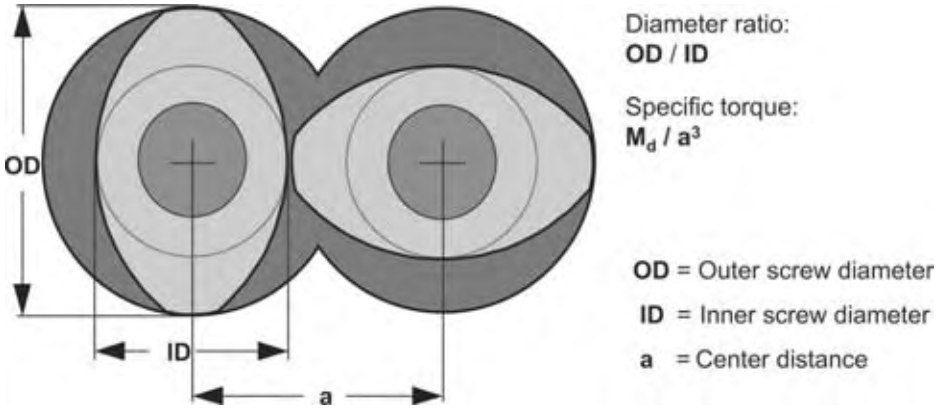


Fig. 3 Characteristic dimensions of an intermeshing twin-screw extruder. (Courtesy of Coperion Corporation.) *(View this art in color at www.dekker.com.)*

(Fig. 4). Geometry constraints limit the 3-lobe design to a more shallow channel depth (lower OD/ID ratio) than 2-lobe units with identical centerline distances. This means that the 3-lobe machine can have a greater shaft diameter for power transmission, impose higher average shear rates on material, but have substantially less free volume. In practice, because of the low free volume relative to the available power, the 3-lobe extruder is often volume, i.e., rate, limited for many mixing tasks.

Within the geometric constraints for self-wiping, the OD/ID ratio can be specified to impart a specific average shear rate, define a free volume, or determine allowable shaft diameter (power transmission). While a 2-lobe machine could be designed to have a low OD/ID ratio, as in the 3-lobe unit, this would not be very sensible. The purpose of creating a 2-lobe unit is to have a machine that would be less likely to be volume limited, but still have the power transmission

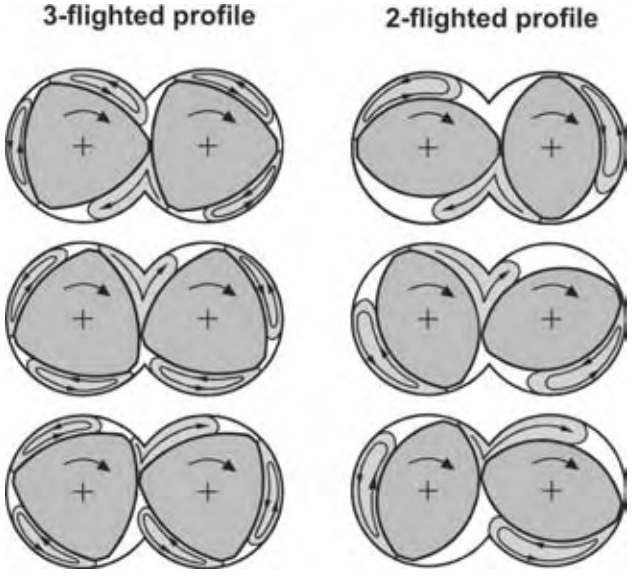


Fig. 4 Self-wiping in the two and three lobed co-rotating twin-screw extruders. (Courtesy of Coperion Corporation.) *(View this art in color at www.dekker.com.)*

capacity necessary to accomplish the desired mixing tasks. The greater the OD/ID ratio, the lower the average shear rate, the higher the unit free volume, and the thinner the shaft available for power transmission. Therefore, the appropriate balance required between power transmission and available free volume must be determined. Fig. 5 shows the progression of available power and volume for five co-rotating twin-screw extruders with the same centerline distance. The OD/ID ratio increases from 1.22 to 1.80 (greater volume) and the specific torque [power/volume ratio expressed as torque capacity/centerline distance cubed (M/a^3)] more than triples. (The allowable power transmission increases for the 1.55 OD/ID unit with thinner shafts as a result of a shaft/element geometry interface improvement. The energy is transferred through 24 splines as opposed to one or two keys). All five machines can be used for similar processing tasks, however, the screw configuration and operating conditions would be different, and the resulting rate typically would be lower for the machines with the smaller OD/ID ratio.

As the power transmission capacity of the extruder gearbox has increased, the ability to process at higher rpm has kept pace. While at the beginning of the 1990s typical processing speeds were between 300 and 600 rpm, today, these same materials are processed at up to 1200 rpm and greater.[20,21]

Elements

Material transport in intermeshing, co-rotating twin-screw extruders is generally dependent on drag flow. The screws pick up the material as they rotate and, where the two screws meet, a complete transfer of the material from one screw to the other takes place, see Fig, 4. The tip of one screw wipes the flanks and roots of the other screw, resulting in a self-wiping action. As the material is transferred from one screw to the other, the direction of material flow is changed and new material surfaces are created with each screw revolution.

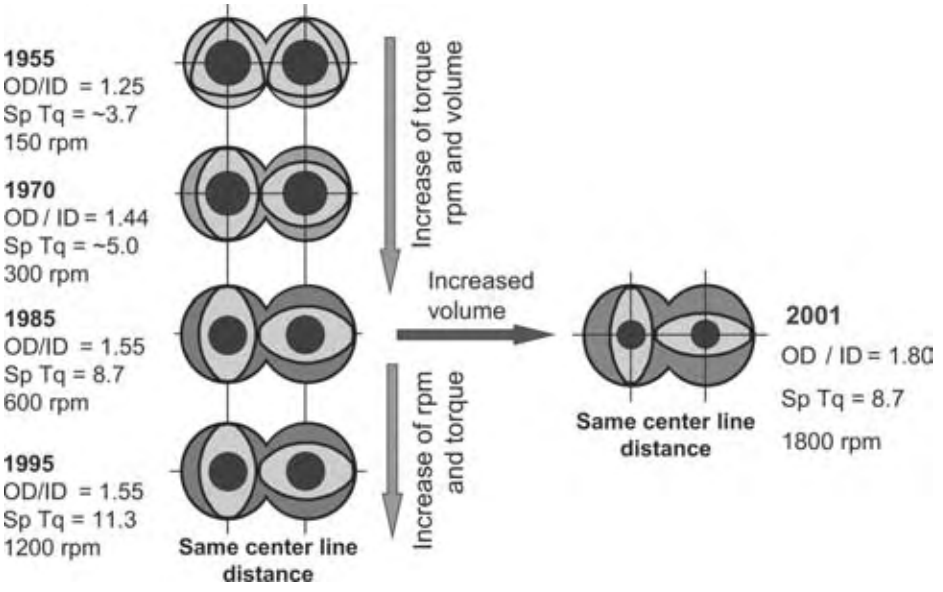


Fig. 5 Development of cross-section, torque and rpm. (Courtesy of Coperion Corporation.) *(View this art in color at www.dekker.com.)*

Screw Bushings

Standard screw bushings are constructed with pitches ranging from approximately 1/2D to 2.0D, where D is the machine diameter, see Fig. 6. High pitch elements might typically be used in feed or devolatilization areas of the extruder. Narrow pitch is used in areas where compaction of material and 100% fill is desired, such as before kneading blocks or seals, or between unit operations (i.e., feeding and vacuum devolatilization). Up to approximately 2.5D, increasing element pitch results in a decrease in residence time and degree of fill, a more narrow residence time distribution, increased drag flow capacity, but also increased sensitivity to pressure flow. That is, as the pitch of an element is increased, drag flow conveys material in the down-channel direction at a faster rate. However, if there is a restrictive force placed in the flow path, the higher pitch element is less effective in building up the pressure necessary to push material past the restriction.

Reverse pitch elements are used to generate back-pressure and therefore create sections of 100% fill that, for example, can be used to separate unit operations, or totally fill a mixing section.

Kneading Blocks and Special Elements

The basic building blocks for mixing in the co-rotating, intermeshing type twin-screw extruder are kneading blocks and special mixing bushings. Special bushings include slotted elements, toothed mixing elements, and blister rings, or the self-wiping equivalent element. Standard conveying type screw bushings are also used in certain circumstances.

Just as screw bushings are characterized by pitch (i.e., flight angle), kneading blocks can be characterized

Fig. 6 Examples of screw bushings (top) and kneading blocks (bottom). (Courtesy of Coperion Corporation.) *(View this art in color at www.dekker.com.)*

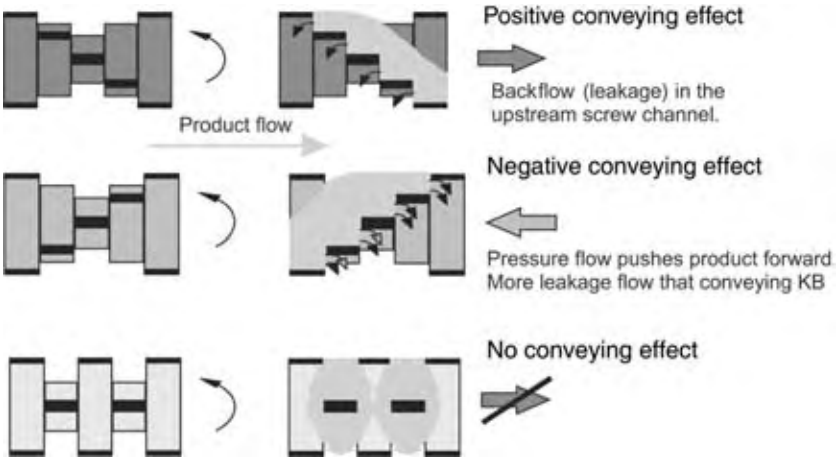


Fig. 7 Working principle of kneading blocks. (Courtesy of Coperion Corporation.) *(View this art in color at www.dekker.com.)*

by individual disc length (width) and stagger angle between successive discs, see Fig. 7. Kneading blocks introduce both a distributive and dispersive mixing component into the system. The relative intensity of each component depends upon disc length as well as stagger angle. For constant stagger angle, increasing the disc width leads to an increase in the dispersive mixing component per unit mixing length, but at the expense of distributing mixing (stream splitting). In other words, as a disc gets wider, it can trap particles more easily in the high shear region, but there are fewer discs per unit length to distribute them. For a constant disc width, the greater the stagger angle the more back mixing and the higher the residence time distribution. Finally, just as there are reverse pitch elements, there are reverse stagger kneading blocks. As with reverse pitch elements, reverse kneading blocks are used to create backpressure. They are however, less restrictive than the reverse pitch elements.

Toothed/gear type elements are the most commonly used of the special mixing elements. The number of teeth around the circumference, as well as the tooth angle, defines them. The former contributes to stream splitting for generation of interfacial surface and the latter, conveying capacity. The main function of these elements is to provide the maximum amount of distributive mixing (little if any dispersive mixing) with minimal energy input.

COUNTER-ROTATING FULLY INTERMESHING

There are three functional variants of the counter-rotating intermeshing extruder, low speed, slow speed, and medium speed. The ''low'' rpm profile extruder is closest in design concept to the screw pump from which all the counter-rotating intermeshing extruders originate. The function of this extruder is to take a premixed powder formulation, melt it with minimal energy introduction, and build pressure to force the melt through some sort of shaping device, such as a profile die. These machines are generally about 25 l/d and have a single feed location. There are a number of different screw concepts. In one variant, the screw consists of thin flight elements at the feed throat for feed intake and initial transport down-channel.[22] The flight width of these elements takes up only a small percentage of the channel width on the opposite screw. The screw then transitions to wide flight elements that take up the entire channel width of the opposite screw. This effectively cuts off any transfer of material from one screw channel to the other, both in the cross-channel and down-channel direction. It also reduces the free volume in the screw channel. This causes the material to be compressed. Thus, as the screw rotates, the material particles push and rub against each other. This interaction causes heat generation, deformation, and eventual melt formation. However, due to the constraints on melt flow, mixing, especially dispersive mixing, is limited. A second screw variant[22] would be to have only fully conjugated screws in the extruder. In this configuration, the pitch along the length of the machine is decreased to implement compression and therefore induce material to melt. Unless the channels are full, the material is not compressed as it is conveyed downstream. Therefore, these machines are typically flood fed. Starve feeding causes the point at which there is sufficient compression to melt the material to move further downstream. Because of the fully conjugated screws, another limitation is screw speed and, therefore, rate for a given machine diameter. Typical speeds for one of these units is 25–50 rpm. This is limited because of the mechanical separation forces between the screw resulting from calendaring effect when material passes between the screw. As rpm is increased, the separation forces also rise causing the screw shafts to rub against the barrel wall.

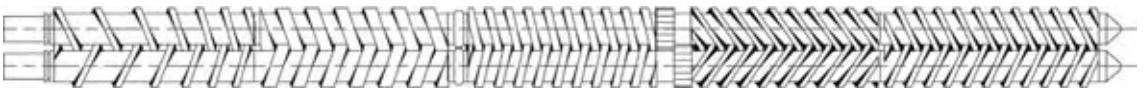


Fig. 8 Traditional compounding counter-rotating screw geometry. (Courtesy of Leistritz Corporation.)

There are exceptions to the above design concept. Most notable is the ''Woodtruder'' that has an l/d up to 40 for some designs. The wood fiber is starve fed into barrel one and molten polymer is introduced downstream by a side arm extruder.[2]

The second variant is a compounding extruder.[23] This unit uses the basic design described above but opens up the intermesh clearances a little and adds mixing elements at various locations along the length of the machine, see Fig. 8. This allows a significant fraction of the material to flow through the high shear ''calendar'' gap that exists between the screws, but still maintains closed pumping chambers. There are several types of conveying elements (all based on mono lobe geometry) and two basic mixing element designs. The conveying elements consist of nonconjugated narrow flight units, fully conjugated units of various pitch and flight width, and transition units. The mixing elements are: 1) intermeshing shearing or blister rings for dispersive mixing; and 2) blister rings with slots cut in the axial direction for more distributive mixing. As a compounding unit, this machine is composed of modular components and can be constructed with downstream feed locations and vent zones. It can also have an l/d up to 60, although a more typical value is between 30 and 40. A significant number of these machines have been sold into the color masterbatch market where high shear is typically required for good pigment dispersion. Since they also have good pressure generation capabilities, they can also be used in profile applications where additional melt mixing is required. The typical maximum speed for these machines is 150 or so rpm.

Recently, a more open geometry that maintains good dispersive mixing characteristics, but permits material to pass between the elements at reduced separation force, has been developed.[23] This allows the machine to be run at much higher speeds of up to 600 rpm. The new mixing element geometry consists of long ''lobal'' elements that mesh together like gear teeth and squeeze material between them, see Fig. 9. This action provides dispersive mixing. The elements can have almost any number of lobes. The more lobes, the greater the number of squeezing actions per revolution. A typical element would have six lobes. Lobal elements can also have slots for distributive mixing cut across the lobes. Additionally, there are also distributive mixing elements composed of circumferential rows of axially offset teeth that wipe past each other in the intermesh zone.

CONICAL COUNTER-ROTATING FULLY INTERMESHING

Conical twins are mostly used in profile extrusion applications. The operating mechanism of these extruders is based on the same positive displacement screw pump conveying principle as is used in low rpm parallel counter-rotating fully intermeshing extruders. As a result, they display a low energy melting mechanism and good pressure generation capability while maintaining

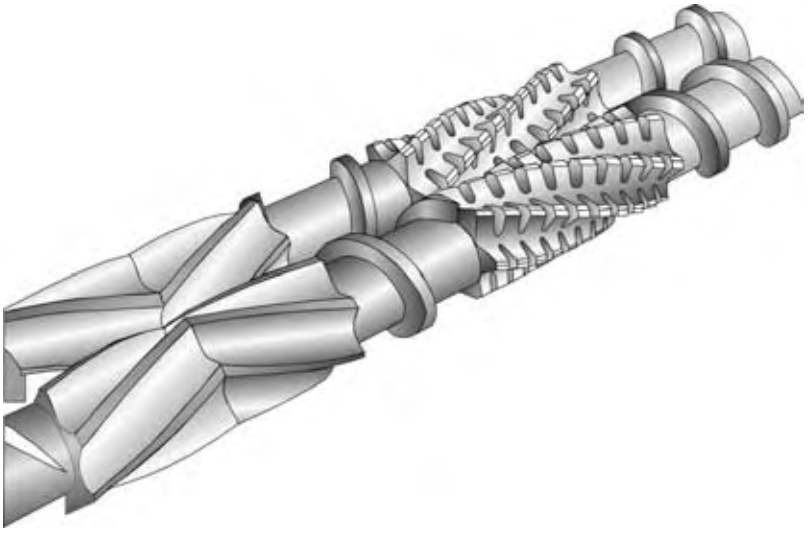


Fig. 9 ''Lobal'' mixing elements. (Courtesy of Leistritz Corporation.)

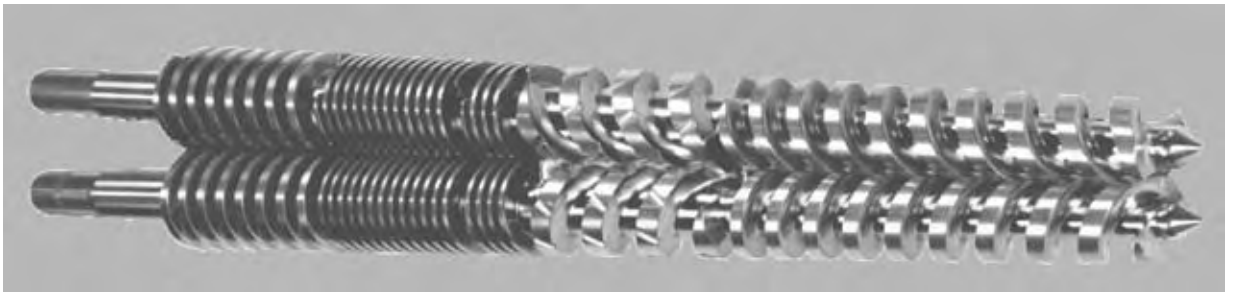


Fig. 10 Screw profile for Conical Twin-screw. (Courtesy of Cincinnati Milacron.) *(View this art in color at www.dekker.com.)*

uniform material flow. This is ideal for heat and shear sensitive materials such as PVC. The unit melts, blends, degasses, and builds up pressure to force the PVC through the die with a minimal temperature rise.

The machine typically has an l/d in the low 20s. The screws are tapered from the inlet to the discharge. A typical ratio of inlet to outlet diameter is two.[24] The screw is a one-piece construction. It is generally divided into five zones: feed, initial compression, compression melting, venting, and metering discharge, see Fig. 10. The feedstock is most often a powder preblend of all ingredients. It is generally flood fed into the large diameter feed opening, then conveyed downstream. To improve the rate for more difficult feedstocks, a crammer can be utilized.[1] As the material is transported down the machine length, the reduction in screw pitch and barrel diameter compress the powder and put it under increasing pressure. Compression, material circulation in the ''C'' shaped chamber that generates frictional heat, and heat transfer from both the barrel and the cored screw introduces enough energy to melt the material. Even as the material melts, the circulation within the chamber is not sufficient to impart any significant mixing of the material. This is why the feedstock is generally premixed powder that only needs to be fused together. However, in some configurations a compression mixing section is used to promote melting and dispersive mixing.[1] Once the material is melted, it enters a vent zone for removal of trapped air, moisture, and other volatiles. As with all other intermeshing extruders, in the vent zone the screw pitch is increased to achieve a partially filled section. Finally, the material is metered and discharged through a profile die.

COUNTER-ROTATING NON-INTERMESHING TWIN-SCREW

In the most simplistic view, the counter-rotating non-intermeshing twin-screw extruder could be viewed as two tangential single-screws sitting side by side in a barrel with an open apex area between them, see Fig. 10. Unlike intermeshing twin-screws, the non-intermeshing extruder can use variation in channel depth as an operating variable, as in a single-screw extruder. The feed and vent sections have deeper screw channels and for pressure generation the channel depth is reduced. Also, in some designs, one shaft is extended to serve as a discharge screw.[25] Unlike the single-screw, the non-intermeshing counter-rotating twin-screw has very good distributive mixing characteristics. This is due to material reorientation that occurs in the apex as the material flow from each screw converges and is transferred from one screw to the other.[26]

Just as in other twin-screw extruders, mixing is accomplished or enhanced by incorporation of special screw elements, see Fig. 11. Dispersive mixing is introduced by placing blister rings in the screw configuration. These elements are typically 1.5D long. Blister rings have two effects that must be balanced. First, the greater the diameter of the ring, the more shear energy is introduced to help disperse particles or agglomerates. However, the larger diameter causes a greater backpressure that must be overcome by the conveying action of the upstream screw configuration. For increased distributive mixing, a reverse pitch element is used. This restricts material flow, increases residence time, and residence time distribution, which in turn increases distributive mixing.

Screws can be placed in the barrel so that the flight tips match. That is, the two flight tips are across from each other in the apex. The other configuration is to have them staggered, typically 180°. When the tips are matched, then the conveying forces of the two screws are working together to drag the material forward, thus minimizing the amount of material that slips backward. If the screws are staggered, then the material has an easier path to avoid being conveyed or dragged forward.

Since the screws do not intermesh, the non-intermeshing extruder is not subjected to some of the same mechanical tolerance constraints as the other twin-screw extruders. This enables the non-intermeshing machine to be built to an almost unlimited l/d ratio. A 72 to 1 l/d unit was used for a number of devolitalization experiments.[27] The enhanced length capability provides a mechanism for the increased internal volume.

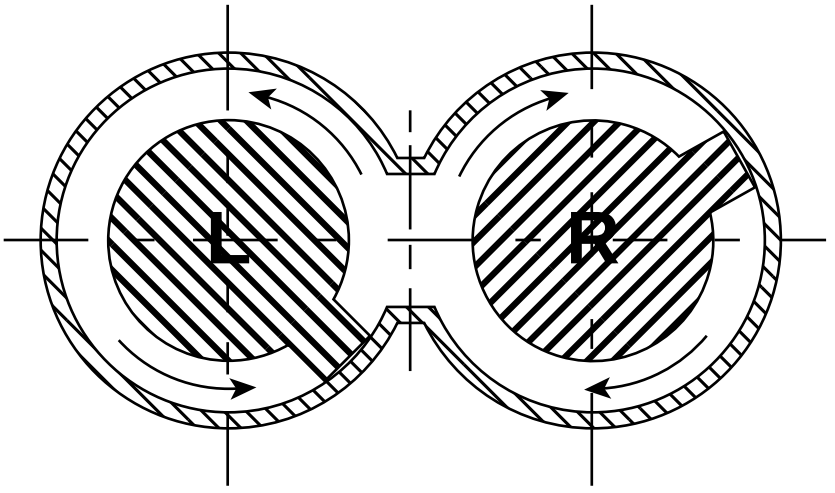


Fig. 11 Crosssection of counter-rotating non-intermeshing extruder. (Courtesy of the Polymer Processing Institute.)

Combine this with the fact that, for the same diameter, a non-intermeshing geometry has approximately 25% greater internal free volume per unit length than an intermeshing unit. This means that material in the non-intermeshing machine can have a significant residence time. This is beneficial for reactive extrusion, dewatering, and devolitalization processes. However, the lack of self-wiping of one screw against the other creates the potential for stagnation zones that can allow material to degrade and potentially slough off to give black specks. Running the extruder screws at a high fill ratio and with proper screw configuration will minimize this potential problem.

COMPARISON BETWEEN CO-ROTATING AND COUNTER-ROTATING OPERATION

The obvious difference between the co- and counter-rotating extruder is that in the co-rotating geometry both screws rotate in the same direction, either clockwise or counter clockwise. In the counter-rotating geometry, they rotate in opposite directions. In addition to geometry, this simple difference impacts the way the two screws interact.

Apex and Intermesh Region

For the counter-rotating systems, the screws in the intermesh area roll-off one another. They both have the same angular velocity at the horizontal axis where the two shafts meet, but a different circumferential speed. The screw flight tip has a larger diameter than the screw root into which it nests. Therefore, the flight tip rolls passed the screw root in a calendering type action. Co-rotating screws, however, do not roll-off one another. In contrast to the rolling motion of the counter-rotating screws, a scraping movement takes place, where one crest edge wipes a screw flank tangentially with equally high relative velocity.

The open area in the apex region between the two screws is significantly different. In the fully intermeshing counter-rotating system, the two screws effectively interlock in the intermesh. This means that no material transfer from one screw to the other takes place and, therefore, the distributive mixing efficiency is greatly reduced. Co-rotating screws form V-shaped apex areas, with 4–5 times more volume. Material can therefore transfer from one screw to the other. This results in a renewal of material layers and surfaces and a better degree of distributive mixing.

Self-Wiping Fully Intermeshing Screws

Self-wiping screws are desirable, and frequently necessary, to prevent material from adhering to the screw root and subsequently degrading. Both counter-rotating and co-rotating fully intermeshing screws are self-wiping. However, self-wiping is achieved in different ways in each system.

In counter-rotating screws, the roll-off process between the screw crest and screw root and between the screw flanks creates a calendar effect. The shear velocity available to wipe the boundary layers is proportionately lower due to the small relative velocity between the surfaces.

In co-rotating screws, one crest edge wipes the flanks of the other screw with a tangentially oriented, constant relative velocity. Because of the higher relative velocity in co-rotating screws, there is a sufficiently high shear velocity to wipe the boundary layers. Consequently, a more efficient and uniform self-wiping action is achieved with the co-rotating screws.

Allowable Operational Speeds and Their Effect on Wear

In counter-rotating screws, the rolling motion of the screws and the resultant calendar effect produces

pressure between the surfaces. This pressure tries to push the shafts apart. The parting forces on the two shafts can cause considerable wear. Since the wear of the screws increased proportionally to the relative velocity of the shafts, it becomes obvious that the intermeshing, self-wiping, counter-rotating machines should operate at relatively low rpm.[23] However, the newer ''lobal'' geometry with greater gaps between the screws, gives up some of the pumping characteristics of the traditional design but can turn at up to 600 rpm.

In co-rotating screws, there is no calendaring effect between the crest and the flank of the screw. The screws are suspended in the plastic melt and, since there is no tendency to push them apart, there is considerably less wear. Therefore, co-rotating machines can be operated at much higher screw speeds with greater throughputs. For example, a 58 mm co-rotating, self-wiping, twin-screw extruder running at 1200 rpm can process 2000 plus lbs/hr of glass filled nylon.[20]

UNIT OPERATIONS—COMPOUNDING

The primary application of twin-screw extruders is profile extrusion and compounding. As indicated, profile extrusion is normally limited to melting of a preblended feedstock. Compounding is significantly more complicated and involves additional unit operations. Additionally, co-rotating machines are the primary geometry used in compounding applications.[23] Therefore, most of the comments in the following sections are related to this geometry.

Feeding

Twin-screw extruders (with the exception of most counter-rotating profile extruders) are designed to be starve fed. Therefore, throughput is independent of screw speed. This permits the processor to control residence time, degree of fill, and specific energy input (kw/kg).

Feeding locations are not limited to the first barrel section. Depending on the feedstock or recipe, feed streams, either solid or liquid, may be introduced at a number of locations along the process section. The only general exception is the conical and parallel counter-rotating twin-screw for profile extrusion. As mentioned previously, this extruder geometry does not lend itself to good mixing. Therefore, all ingredients are generally preblended and introduced at a single feed location.

Upstream Feeding

In compounding operations, the feedstock introduced at the first barrel section is usually a polymeric solid. The form is typically pellets, flake, or powder. Most pellets and flake are easy to feed since they have a high bulk density, their individual dimensions are sufficiently small with respect to extruder channel depth, and they do not lubricate the screw and thus impede the drag flow mechanism. Standard profile screw elements with a pitch of 1–2D are appropriate.

In situations where the particle size is large with respect to the channel depth ($>1/2\,h$), the material has a lower bulk density, or absolute maximum feed capacity is needed, special elements with increased volume are useful. The self-wiping profile of the pushing flight (and sometimes the trailing flight) has been transformed into a square channel profile. This modification accomplishes two functions. First, it directs more force acting on the material to be in the down-channel direction. Second, it creates up to 40% additional free volume in the element.

Powders that tend to fluidize, especially those such as silica which have a very low initial bulk density, are significantly more difficult to feed than pellets or flakes. The first step to successful feeding depends upon eliminating (or at least minimizing) fluidization, and controlling the separation of already entrained air, prior to material entering the extruder. In general, the vertical drop should be as short as possible. Also, direct the feed to the down-turning section of the screw. This is the apex region for the counter-rotating machine and the barrel wall for the co-rotating extruder. Ideally the screw configuration should be designed to allow air to travel down channel for removal through a vent section rather than be trapped and forced to flow countercurrent and back out the feed throat. However, in order to be effective, a plastification zone typically must be backed up by a restrictive element. This element then blocks the air from traveling downstream. In this case, special elements such as increased volume (undercut) or single flight (SF) elements play an important role in feed introduction. The choice of either increased volume or SF elements depends upon the amount and degree of air separation required. For a relatively low amount of air, the undercut element again provides a greater free volume and a more open path for air to flow backwards. However, as more air is forced back to the feed throat, the material fluidizes and the undercut elements lose their effectiveness. Under these circumstances the SF should be considered. Unlike the undercut element, which has greater free volume, the SF has approximately 15% lower free volume per unit length than a standard element. The SF elements do not allow air to flow easily in a countercurrent direction and, therefore, force it to flow past the restriction in the plastification section. These elements function in this manner as a result of the severe flow restrictions through the apex caused by the wide crest. These crests create a positive

displacement flow greater than in any other co-rotating twin-screw element.

Downstream Feeding

In many compounding operations it is necessary to split the feed streams. This may be required in order to: 1) achieve disperse phase size of an impact modifier; 2) retain aspect ratio of reinforcing fiber filler; or 3) obtain high level of loading for either low bulk density filler or incompatible low viscosity additives.

The most efficient way to add solids in a downstream location is to use a twin-screw side feeder. The twin-screw side feeder is normally a co-rotating device with high OD/ID ratio ($\sim$1.8). It has several advantages over single-screw side or top feeders. First, the twin-screw has better solids conveying characteristics, since a twin-screw device does not totally rely on drag flow. The twin-screw also has a wider longitudinal as well as a larger total cross-sectional discharge opening than a single-screw side feeder, and therefore provides lower pressure and more uniform feed introduction. Typically there is a vent above the side feed opening in the barrel. This permits entrained air to escape. For very low bulk density material containing significant air, the vent opening is moved one barrel upstream to avoid material particles getting entrained in the air as it exits the extruder. By having the air travel upstream before discharge, requires it to travel a path around the screw channel. This disrupts the air flow and causes the solid to disengage.

Liquid additives are mostly added downstream of the plastification section because they tend to lubricate the pellets or cause powders to agglomerate in the feed throat. If a significant amount of liquid is to be incorporated, it can be added at several locations. The most effective method for low viscosity liquid incorporation is to inject into a fully filled distributive mixing section. This requires a pressure injection valve and positive displacement pump. For small amounts of compatible liquid, non-pressurized injection into a low degree of fill area of the screw configuration may also be acceptable.

Plastification Mechanisms

Plastification of polymeric material requires energy to be transferred from an outside source into the material. In the twin-screw extruder, this energy transfer occurs through both mechanical and thermal mechanisms. However, as the extruder gets larger, the surface to volume ratio decreases significantly. Therefore, mechanical energy transfer is the dominant mechanism for plastification.

Thermal Heat Transfer

Normally, thermal energy is introduced through electric heaters surrounding the barrel or heat transfer medium that is pumped through barrel bores. As mentioned previously, smaller extruders can introduce a greater percentage of energy through heat transfer. Conversely, they can also loose a higher percentage of heat through heat transfer. If too much heat is lost in this manner, then on scale up to a production unit, the same percentage of heat cannot be removed. This will result in a higher discharge temperature. It is therefore very important that laboratory or development extruders run as close to adiabatic conditions as possible.

Mechanical Heat Transfer

In most cases, the majority of energy required for plastification comes from the mechanical input of the screw configuration. In counter-rotating intermeshing systems, material is compressed and deformed by reduction of screw pitch. In co-rotating systems, kneading blocks are the primary tool used to accomplish this task. The amount of mechanical energy introduced depends upon not only the number of kneading blocks, but also the configuration within the plastification zone, the screw rpm, and throughput rate. Specific mechanical energy (Sme), the energy introduced per pound or kilo of product, increases with rpm, especially when the plastification section is backed-up with a restrictive screw element. A second restrictive element further increases Sme.

Mixing Mechanisms

Mixing requirements for polymer compounding can be divided into two basic disciplines—dispersive and distributive. Dispersive mixing breaks down a particle into smaller units, while distributive mixing homogenizes the spatial relationship of the particles (whether dispersed or not).

The basic building blocks for mixing in the co-rotating twin-screw are kneading blocks and special mixing bushings. These special elements include toothed mixing units and blister rings, both standard geometry and self-wiping.

It is not surprising that the mixing element selection for distributive and dispersive blending processes is different. The distributive mixing profile uses narrow kneading blocks to maximize the number of flow divisions per machine length.

In dispersive mixing, wide disc kneading blocks are used to increase shear stress applied to the material. These elements are typically backed up with a restrictive

kneading block or screw bushing to increase the degree of fill as well as residence time. The number of restrictive elements that can be used is limited due to the resulting temperature buildup and potential material degradation. Therefore, in order to transmit increased stress to the material, a number of shorter mixing sections (1–2D long) are often more efficient than one long mixing section as this permits elastic materials to relax in the conveying sections between the mixing areas.

For very low viscosity products, introduce sufficient energy for dispersive or distributive mixing by adjusting the feed sequence, such that only a small portion of the solvent or dilution oil is added in the upstream part of the twin-screw extruder and therefore high mechanical stress transfer can be achieved. The remainder of the diluent is introduced in the latter sections of the machine.

Devolitalization/Degassing

In compounding lines, devolatilization typically involves removal of entrained air, moisture contained in the incoming feedstock, or byproducts from any reaction between recipe components. The amounts are generally less than 1% or 2% and typically less than 1%. Therefore, the staged vacuum set-ups and stripping techniques used in devolatilization processes where 10% to 60% volatile must be removed, are not necessary. However, even in compounding, devolatilization is a critical step, because while it is necessary to remove these volatile, it serves as a location for the potential introduction of contamination. The screw configuration in the vent area is typically comprised of screw bushing with a pitch of between 1 and 2D. This design permits sufficient residence time at a lower degree of fill to remove volatiles while maintaining the polymer within the bounds of the screw flight.

Discharge

Material is discharged from most twin-screw compounding operations as pellets or specialized forms, such as sheets, tubes, ropes, or other more complicated profiles. In order to push material through these dies, the machine must generate the appropriate pressure. This can range from 200 psi or less for strand dies, and up to 5000 or so psi for sheet and other profile dies. When the discharge pressure exceeds 2000 psi, or precise product gauge control is critical, a single-screw extruder or melt gear pump is typically used to generate the pressure in the co-rotating twin-screw. In large-scale polyolefin pelletizing operations, rates can be dramatically increased by integrating a gear pump into the system.[28] Also, fiber spinning and sheet extrusion operations typically use a gear pump at the end of the machine. For fully intermeshing counter-rotating machines, no device for discharge assistance is required.

CONCLUSIONS

There is a wide range of process applications for twin-screw extrusion. The two most prolific are profile extrusion and compounding. The former utilizes traditional counter-rotating geometry with either parallel or conical screws. The majority of compounding applications are run on co-rotating fully intermeshing extruders. The other twin-screw machine types, while not as widely used, each have specific areas where they are the preferred geometry. For a more detailed analysis of the operating mechanisms of the various geometries please see Refs.[29–33].

REFERENCES

1. Weinrich, C. Direct extrusion of wood flour/plastics composites on conical counterrotating twin screw extruders. In *Wood-Plastic Conference,* Conference Proceedings, Baltimore, MD, December 5–6, 2000; 207–215.
2. Dardenne, D.S. Woodtruder™ (Patents Pending) system for extrusion of wood fiber polymeric composites. In *Wood-Plastic Conference,* Conference Proceedings, Baltimore, MD, December 5–6, 2000; 187–188.
3. Jackson, S.M. Advanced process technology for manufacturing wood-polymer composites. In *Wood-Plastic Conference,* Conference Proceedings, Baltimore, MD, December 5–6, 2000; 21–28.
4. Wiedmann, W. Improving product quality through twin-screw extrusion and closed-loop quality control. In *Food Extrusion Science and Technology*; Kokini, J.L., Ho, C., Karwe, M.V., Eds.; Marcel Dekker: New York, 1992; 539–570.
5. Brauer, F. Continuous production of polyurethanes on a twin-screw extruder reactor. In *PPS Summer Meeting* Conference Abstracts, Amherst, MA, August 16–17, 1989; 7D.
6. Uberall, M.; Kapfer, K.; Brauer, F.; Trippler, R. Devolitalization operations for high volume polyolefin solutions. In *Polyolefins IX,* Conference Proceedings, Houston, TX, February 26–March 1, 1995; 323–338.
7. Andersen, P.G.; Kite-Powell, K.K. Method of Removing Liquids from Solids. US Patent 5,151,026, September 29, 1992.
8. Friedrich, R. Powder compression with co-rotating twin-screw extruders. J. Powder Bulk Solids Technol. **1980**, *4* (4), 27–32.

9. Beecher, E.D.; Carr, M.E.; Grillo, J.G. Starch compounding on co-rotating twin screw extruder: starch as an encapsulation medium for a controlled release application. SPE Antec Proceedings, Boston, MA, May 7–11,1995; 2037–2041.
10. Andersen, P.G. Twin screw extrusion guidelines for compounding nanocomposites. SPE Antec Proceedings, San Francisco, CA, May 5–8, 2002; 219–223.
11. Friedrich, R. Comparison between processing polymers and propelling charge powders on continuous co-rotating twin-screw mixers, 18th International Annual Confernece of the ICT, July 1–3, 1987.
12. Kapfer, K.; Schneider, H. Production of compounds with high filler or fiber loading on screw kneaders. SPE Antec Proceedings; Nashville, TN, May 4–8, 2003; CD-ROM.
13. Andersen, P.G.; Dickens, D. Selection criteria for concentrate and masterbatch compounding. AMI Proceedings Thermoplastic Concentrates 2002, New Orleans, LA, January 16–18, 2002.
14. Kapfer, K. Current developments in twin-screw design and its application in the preparation of polymer blends. SPE Antec Proceedings, Atlanta, GA, April 18–21, 1988; 96–101.
15. Laughner, M.; Parikh, D.; Walton, K. New developments in metallocene ethylene elastomers for automotive applications. In *Polyolefins 2001*, Proceedings, Houston, TX, February 25–28, 2001; 339–367.
16. Liberto, L.P. Ed. *Powder Coating: The Complete Finishers Handbook*; Powder Coating Institute: Alexandria, VA, 1999; p 12.
17. Charpentier, J.C. Compounding color toners. Proceedings Toners & Photoreceptors 2001, Santa Barbara, CA, June 10–13, 2001; Session 13.
18. Beecher, E. Techniques for short run continuous compounding of adhesives and sealants. Presented at ASC 2001 Fall Convention. New Orleans, LA, October 21–24, 2001.
19. Burbank, F.R.; Jackson, S.M. New twin-screw element design for elastomer compounding. SPE Antec Proceedings, New York, NY, May 2–6, 1999, 225–229.
20. Andersen, P.; Haering, E.; Kapfer, K. Understanding high rate and high speed compounding on co-rotating twin-screw extruders. SPE Antec Proceedings, Toronto, Canada, April 27–May 2, 1997; CD-ROM.
21. Kapfer, K.; Haering, E. Deeper screw flights offer new opportunities for co-rotating twin-screw extruders. SPE Antec Proceedings, San Francisco, CA, May 5–8, 2001; CD-ROM.
22. White, J.L. Modular counter-rotating twin screw extruders: intermeshing and tangential. In *Polymer Mixing: Technology and Engineering*; White, J.L., Coran, A.Y., Moet, A., Eds.; Hanser: Munich, 2001; 167.
23. Thiele, W.C. Counterrotating intermeshing twin-screw extruders. In *Plastics Compounding: Equipment and Processing*; Todd, D.B., Ed.; Hanser: Munich, 1998; 46–70.
24. Brown, T. Extrusion of rigid PVC products. In *Vinyl RETEC and Tutorial '96 Technical Papers*; Cincinnati, OH, October 15–16, 1996; 57–67.
25. Skidmore, R.H. Method of Separating an Insoluble Liquid from Polymer Composition. US Patent 3,742,093, June 26, 1973.
26. Howland, C.; Erwin, L. Mixing in counter rotating tangential twin screw extruders. SPE Antec Proceedings, Chicago, IL, May 2–5, 1983; 113–116.
27. Foster, R.W.; Lindt, J.T. Bubble free devolitalization in counter-rotating non-intermeshing twin screw extruders. Polym. Engng. Sci. **1990**, *30* (7), 424–435.
28. Schuler, E.W. Energy efficiency in high volume plasctics compounding operations. SPE Antec Proceedings, Chicago, IL, May 2–5, 1983, 924–927.
29. Herrmann, H.; Jakopin, S. A comprehensive analysis of multi-screw extruder mechanisms. SPE Antec Proceedings, Montreal, Canada, April 25–28, 1977; 481–486.
30. White, J.L. *Twin Screw Extrusion*; Hanser: Munich, 1990.
31. White, J.L.; Coran, A.Y.; Moet, A., Eds.; *Polymer Mixing: Technology and Engineering*; Hanser: Munich, 2001.
32. Manas-Zloczower, I.; Tadmor, Z., Eds.; *Mixing and Compounding of Polymers: Theory and Practice*; Hanser: Munich, 1993.
33. Todd, D.B., Ed. *Plastics Compounding: Equipment and Processing*; Hanser: Munich, 1998.

Use of Lipases to Isolate Polyunsaturated and Oxygenated Fatty Acids and Form Value-Added Ester Products

Douglas G. Hayes
Department Biosystems Engineering and Environmental Science, University of Tennessee, Knoxville, Tennessee, U.S.A.

INTRODUCTION

Polyunsaturated and oxygenated fatty acids, obtained from triacylglycerols (TAG) of several different plant and animal species, are valuable materials feedstock for value-added products in a variety of industries: food, pharmaceutical, cosmetics, and paints and coatings. The acyl species, their chemical structure, and their most abundant sources are summarized in Table 1. In contrast to inexpensive C_{16} and C_{18} saturated and Δ9-unsaturated acyl groups, such as palmitic (16:0), stearic (18:0), oleic (18:1-9c), linoleic (18:2-9c, 12c), and α-linolenic acid (ALA; 18:3-9c, 12c,15c), recovered from the oil of soybean and other common sources, and C_4–C_{16} saturates from palm oil and milk fat, polyunsaturated and oxygenated acids are derived from less common sources, and particularly for polyunsaturated fatty acid (PUFA), are typically present at only 20–40% purity.

Although common C_4–C_{18} acyl groups are readily isolated and modified by chemical means, PUFA and oxygenated fatty acids are susceptible to degradation at their double bonds and oxygenated groups, respectively. Biocatalytic means of isolation and modification provide mild operating conditions (low temperature and pressure, near-neutral pH, and the absence of toxic materials such as catalysts and harmful solvents), which will prevent degradation and promote an environmentally-friendly workplace.[1,2] The positional, substrate, and regioselectivity of lipases will lead to a narrower product distribution compared to chemical methods. The higher selling costs of PUFA and oxygenated fatty acid-derived products will help offset the high materials costs of enzymes, in contrast to common C_{16} and C_{18} acyl groups and their esters. This entry will briefly review the applications of PUFA and oxygenated fatty acids and the biocatalytic properties of lipases in a broad sense, discuss the lipase-mediated isolation of PUFA and oxygenated free fatty acids (FFA), and review lipase-catalyzed synthesis of PUFA and hydroxy FFA-enriched ester products and their applications.

APPLICATIONS OF POLYUNSATURATED AND OXYGENATED FATTY ACIDS

Δ4–Δ6 PUFA such as docosahexaenoic (DHA), eicosapentaenoic (EPA), arichidonic (AA), and γ-linolenic (GLA) and their derivatives are gaining popularity as food additives, or ''nutraceuticals,'' due to the many medical and nutritional benefits they present.[3] (DHA, EPA, and ALA are frequently grouped together in application-oriented discussions as ''n-3'' fatty acids, since each contains a double bond located three positions away from the terminal carbon atom on the acyl chain and provide similar medical benefits.) For instance, a literature search on patents involving DHA and AA from the year 1999 to the present resulted in 1014 and 878 hits, respectively. Of great benefit to world population suffering from an epidemic of obesity, a diet rich in PUFA is known to decrease blood TAG levels, blood pressure, and low density lipoprotein (or ''bad'') cholesterol. They also have many applications relating to brain and nervous system development in infants, treatment of arthritis and osteoporosis, and possess anti-inflammatory properties. In addition to the applications listed above, AA dietary supplements have been reported to increase muscle mass and reduce intravenous hemorrhaging in premature infants. AA, DHA, EPA, GLA, and ALA possess *cis* double bonds in positions whereby they can serve as oxidation substrates for the enzyme lipoxygenase to produce hydroperoxides, which are valuable precursors of prostaglandins, leukotrienes, and flavor ingredients: ketones, alcohols, and aldehydes, and are chiral synthons for combinatorial libraries employed in drug discovery. Although not technically a PUFA, petroselinic acid (18:1-6c) is listed in this category because it possesses a double bond positioned at C_6, and shares many of the same nutritional and medical applications as DHA and EPA. Petroselinic acid is employed as a cosmetic ingredient for hair growth and skin rehydration products. Conjugated linoleic acids, CLAs, lipids that occur naturally in milk fat and beef tallow are becoming increasingly popular materials with applications similar to those

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120030531

Table 1 Polyunsaturated and oxygenated free fatty acyl species

FFA product	Chemical structure	Sources
Arachidonic (AA)	20:4-5c, 8c, 11c, 14c	Single-cell oils (25%)
Crepenynic	18:1-9c, 12-yne	*Crepis alpine* oil
Docosahexaenoic (DHA)	22:6-4c, 7c, 10c, 13c, 16c, 19c	Fish oils; single-cell oils (5–20%)
Eicosapenaenoic (EPA)	20:5-5c, 8c, 11c, 14c, 17c	Fish oils; single-cell oils (5–20%)
γ-Linolenic (GLA)	18:3-6c, 9c, 12c	Blackcurrant (*Ribes nigrum*, 15–20%) borage (*Borago officinalis*, 20–25%), or evening primrose (*Oenothera biennis*, 8–14%) oil
Petroselinic	18:1-6c	Coriander (*Coriandrum sativum*) oil (15.3%)
Conjugated linoleic acids (CLAs)	18:2-9c, 11t; 18:2-10t, 12c	Milk fat; beef tallow; modification of ricinoleic acid; free radical isomerization of α-linolenic acid (equimolar mixture of two CLA isomers)
	20:1-5c, 22:2-5c,13c	Meadowfoam (*limnanthes alba*) oil (83.5%)
	20:3-5c, 11c, 14c, 20:4-5c, 11c, 14c, 17c	*Biota orientalis* oil (15.6%)
Dimorphecolic	S-18:2-10t, 12t,OH-9	*Dimorphotheca pluvialis* oil (61.8%)
Lesquerolic and auricolic	R-22:1-11c, OH-14, R-22:2-11c, 17c OH-14	*Lesquerella fendleri* oil (57%)
Ricinoleic	R-18:1-9c, OH-14	Castor (*Ricinus comminus*) oil (90%)
Vernolic	18:1-9c,-epoxy-12,13c	*Vernonia galamensis* (77–81%) or *Euphorbia lagascae* (60-65%) oil

listed above for DHA, with additional employment in the treatment of diabetes.[4] An equimolar mixture of the CLAs 18:2-9c, 11t and 18:2-10t, 12c are synthesized from free radical isomerization of ALA.[4] Long-chain PUFA from meadowfoam and *Biota orientalis* oils (Table 1) have applications as oleochemicals, surfactants, and cosmetics; *B. orientalis* PUFA may have activity in lipid metabolism. Crepenynic acid has received attention as a chemical feedstock, including for use in paints and coatings, due to its unique carbon–carbon triple bond.

The second category, hydroxy- and epoxy-containing acyl groups (and their esters), have numerous applications as chemical feedstocks in lubricants, paints and coatings, food and cosmetics emulsifiers, nylon synthesis, laxatives, disinfectants, etc.[5,6] The utility of ricinoleic acid and its derivatives, for instance, the most commonly recognized member of this category, has existed for at least a century, demonstrated by US patents issued nearly 100 years ago, and 577 patents issued between the years 1990 and 2004.

LIPASES

Lipases play the specific role of forming and hydrolyzing ester bonds involving long-chain carboxylic, or fatty, acids (Fig. 1).[7] Lipase-catalyzed hydrolysis, or ''lipolysis,'' typically occurs at liquid–liquid or liquid–solid interfaces, while esterification and acyl exchange (reactions 2 and 3–5 of Fig. 1, respectively) are catalyzed by solid-phase lipase dispersed in low-water, nonaqueous, media.[8] Lipases possess regio-, stereo-, and substrate selectivity to control the product distribution.[9] For instance, many lipases are strongly regioselective toward primary hydroxyl groups, allowing them to utilize the 1- and 3- position of glycerol as substrate but not the 2-position. Lipases which fit this description are categorized as ''1,3-regioselective'' while those that do not are ''random.'' The strong stereoselectivity of lipases has made them common tools in the separation of racemic acid or alcohol mixtures in the pharmaceutical industry. The fatty acyl substrate selectivity of lipases is based on chain length and the position, type (*cis* or *trans*), and number of double bonds. For instance, lipases discriminate against substrates with double bonds near the carbonyl terminus, such as the Δ4–Δ6 PUFA. *Geotrichum candidum* lipase has unique substrate selectivity, with a strong preference toward Δ9 acyl groups, such as oleic, linoleic, and α-linolenic acids. There are several lipases commercially available which possess high activity and thermo-, storage, and operational stability resulting from enhanced knowledge of the fundamental

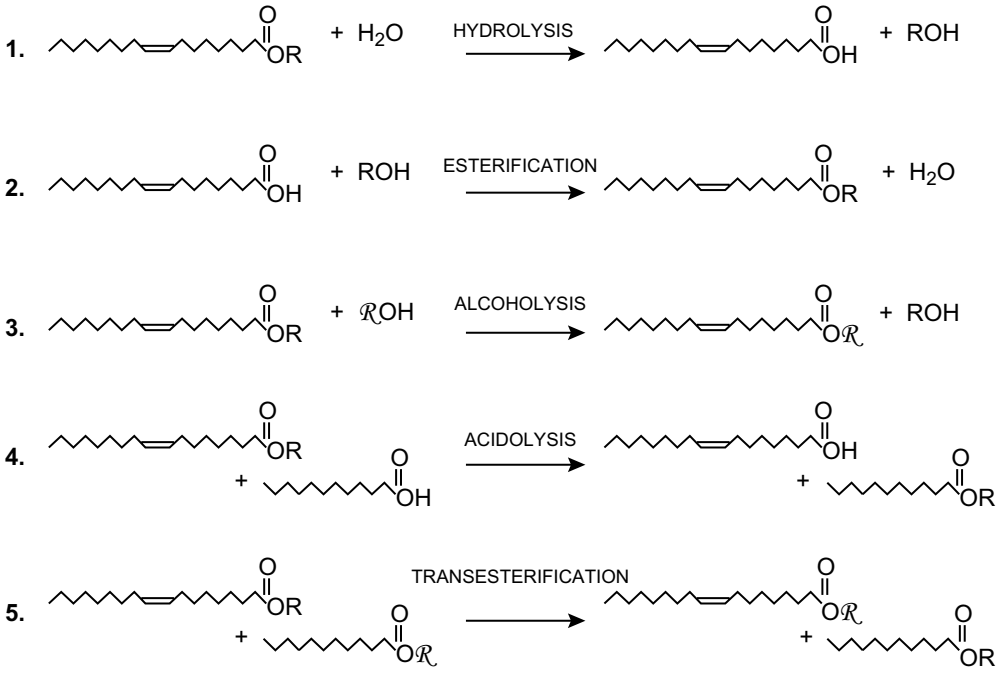


Fig. 1 Reactions catalyzed by lipases. (From Ref.[6].)

biochemistry of lipases, isolation of lipases from extremophiles, expression of lipase genes in recombinant hosts, directed evolution, and enhanced immobilization technology (e.g., Novozyme™, Lipozyme RM IM™, and Lipozyme TL IM, immobilized *Candida antarctica*, *Rhizomucor miehei*, and *Thermomyces lauginosus* lipases, respectively, from Novozymes, Inc., Franklinton, North Carolina, U.S.A.).[7]

ACYL MIGRATION PLAGUES 1,3-POSITIONAL SELECTIVITY OF LIPASES

1,3-Selective lipases are particularly useful for lipolysis of TAG that possess acyl groups of interest only in the 1- and 3-glycerol positions and for forming ''structured'' TAG, both of which are discussed later. However, acyl migration, a non-enzymatic intramolecular acyl exchange within monoacyl- and diacyl-glycerol, or MAG and DAG, respectively, greatly reduces the anticipated yield and purity. Specifically, the following two reactions occur:

$$\text{I.}\quad 2-MAG \rightleftharpoons 1-(3-)MAG$$

$$\text{II.}\quad 1,2-DAG \rightleftharpoons 1,2-(2,3-)DAG$$

Reviewed elsewhere,[10] acyl migration is frequently catalyzed by electrostatically charged materials in the reaction medium or enzyme preparation. Its occurrence increases with the water content of the reaction medium and temperature. Although acyl migration is difficult to prevent, its extent is significantly lessened when the residence time of the reaction medium in the presence of the biocatalyst is minimized.

THE IMPORTANCE OF LIPID SEPARATIONS IN ENZYMATIC PROCESSES

High yield of PUFA incorporation can only occur when the acyl donor is highly enriched in PUFA due to the relatively poor selectivity of all lipases toward Δ4–Δ6 FFA, with *Pseudomonas* sp. and *C. antarctica* lipase being the least discriminatory. Thus, when synthesizing structured or PUFA-enriched TAG from raw sources such as seed or fish oil or lard, one must employ a PUFA source that is highly pure in order to achieve a high yield. Thus, lipid separation steps are a very important component of the overall process. The most commonly employed lipid separation steps integrated with biocatalysis are listed in Table 2. Molecular distillation is perhaps the most common and rapid of the separation methods, but can degrade double bonds and involves a high energy cost.

Saponification is the preferred choice for isolation of FFA from esters, but must be employed in a buffered solution when applied to hydroxy fatty acyl mixtures. Crystallization and precipitation can be slow

Table 2 Separation methods that accompany lipase-catalyzed reactions

Method	Selectivity	Application with enzyme reactions
Short-path molecular distillation	Separates molecules by their molecular weight	Fractionation of MAG, DAG, TAG, and fatty acid esters from lipolysate or reaction medium
Saponification	Removes FFA from esters	Downstream removal of FFA
Crystallization and precipitation	Fractionates based on melting point	In situ removal of long-chain and hydroxy FFA and polyol monoesters
Urea inclusion compounds	Fractionation of polyunsaturates from saturates	Downstream processing of lipase-catalyzed PUFA-enriched FFA
High performance and low pressure liquid chromatography (adsorption)	Separation of lipid classes, separation of lipids by molecular weight and degree of unsaturation	Impractical for most preparative or large-scale processes
Solvent extraction (''countercurrent distribution'')	Separation of neutral lipid classes and PUFA from saturated FFA	Readily scaled up, but poorly selective

processes that require high-energy costs due to the low temperatures and process time required; thus, they are impractical for large-scale processes. Exceptions to this rule may be the isolation of MAG, polyol monoesters, and long-chain saturated and hydroxy FFA and their derivatives, which have melting point temperatures slightly below ambient (10–20°C). A cold trap can be implemented into the bioreactor design for continuous isolation of MAG, thermodynamically driving the desired reaction in the forward direction.

Chromatographic methods provide the greatest selectivity among the methods listed in Table 2; however, chromatography is very difficult to adapt to large-scale processing and is quite expensive. Solvent extraction, a technique that isolates FFA by their degree of saturation and/or the presence or absence of hydroxy groups, is poorly selective. It was employed frequently in the 1940s–1960s in the form of multistage, countercurrent, contacting equipment (known as ''countercurrent distribution'') before its replacement by chromatographic techniques for analytical and preparative scale applications. Urea inclusion compounds (UICs) have also been employed for over 50 years to separate lipids primarily by the degree of saturation.[11] UICs represent a process that has a high selectivity (greater than solvent extraction but less than molecular distillation or chromatography), can be readily scaled up, and involves much lower energy costs.[11] UIC-based fractionation is complementary to lipase-catalyzed selective esterification to purify PUFA.[11] Moreover, lipase-catalyzed esterification of an FFA mixture will remove common Δ9-unsaturated C_{18} FFA and UIC-based fractionation will remove saturated and monounsaturated FFA, resulting in a PUFA-enriched FFA product.[11]

ISOLATION OF FFA BY LIPOLYSIS

Traditionally, FFA are isolated from a degummed and bleached seed oil by the Colgate–Emery process, which employs steam at high pressure and temperature (typically 250°C and 5 MPa) to cleave the TAG ester bonds. The high temperature of this process is known to degrade PUFA and oxygenated FFA. In addition, the Colgate–Emery process consumes a large amount of energy, about 800 kJ of energy per kg of oil. Although lipolysis may serve as a low-cost and environmentally safe alternative to the Colgate–Emery process for FFA isolation, its widespread use in the oleochemical industry is blocked by the high material costs of enzymes. Lipolysis will have a greater impact upon PUFA and oxygenated FFA isolation due to the susceptibility of double bonds, epoxy, and hydroxyl groups to degradative side-reactions and the higher selling costs of PUFA-related products.

Partial isolation of PUFA with Δ4–Δ6 unsaturation occurs during lipolysis due to the inability of lipases to rapidly utilize TAG that contain PUFA groups.[12] Moreover, the PUFA are recovered from the MAG, DAG, and TAG molecules of the reaction mixture, while common acyl groups are preferentially released, then removed from the glycerides by saponification. (Alternatively, TAG are subjected to alcoholysis, reaction 3 of Fig. 1. The ester products are removed from the MAG/DAG/TAG by short-path molecular distillation. Since lipolysis and alcoholysis yield similar degrees of purification, selection between the two is based on economic considerations.) *C. rugosa* lipase has been employed in most instances for selective lipolysis. Due to the inability of many lipases to hydrolyze TAG with 2 or 3 Δ4–Δ6 PUFA groups, glyceride

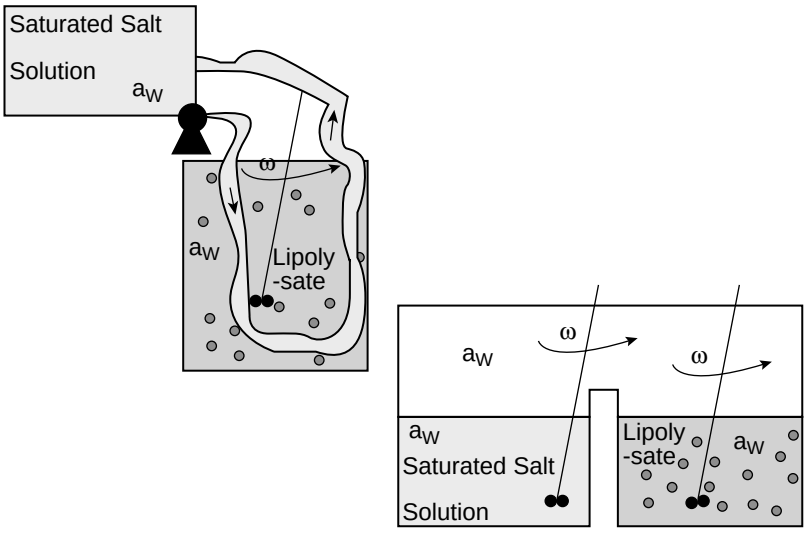


Fig. 2 Bioreactor configurations for lipolysis that employ saturated salt solutions of high water activity to maintain water saturation. *(View this art in color at www.dekker.com.)*

products contain PUFA at only moderately high concentrations (~40–75%) and yield (60–90%). The purity and yield are not strongly affected by reaction medium type nor by addition of fresh lipase during the time course of reaction. *Pseudomonas* sp. lipases generally produced the highest degree of lipolysis (79% or greater), but exhibited a much lower degree of discrimination against PUFA acyl groups than most other lipases.

1,3-Selective lipolysis is employed to isolate the hydroxy FFA listed in Table 1 because, unlike "random" lipolysis, poly(hydroxy acid) byproduct formation cannot be catalyzed. The employment of 1,3-selective lipases is particularly beneficial for recovery of lesquerolic and dimorphecolic acids from lesquerella and dimorphotheca oils, respectively, because the hydroxyl acyl groups are located solely in the 1- and 3-glycerol positions within TAG. Lipolysis of lesquerella oil by Lipozyme RM IM occurred at 66.7% yield and produced lesquerolic and auricolic FFA at a combined purity of 80% among the FFA.[13] A greater extent of lipolysis resulted in a greater yield at the cost of lower purity due to the occurrence of acyl migration.[13] FFA enriched in hydroxyl, epoxy, and long-chain ($>C_{20}$) cannot be isolated by a simple saponification procedure; the alkali extractant must be buffered at pH 10. Alternatively, hydroxy FFA-rich products are isolated by crystallization through use of cold trap placed on-line in a continuous-mode bioreactor.[14]

When designing a bioreactor for lipolysis for preparative scale or larger, the use of immobilized lipases is essential to facilitate reusability and recovery of the biocatalyst from the reaction medium. The major issue to decide is the relative amounts of water and TAG, and the means of contacting the two poorly miscible liquids. The most common approach, to employ rapid stirring, resulting in either water-in-oil or oil-in-water emulsions, provides a high degree of interfacial area, but promotes poor dispersion of the immobilized biocatalyst and hinders rapid separation of liquid phases and the immobilized lipase. Microemulsions, nanometer-sized dispersions formed by the addition of a surfactant and alkane, provide excellent interfacial area and interfacial mass transport but are not recommended because of the occurrence of product inhibition and the very difficult separation of the water and oil phases due to the surfactant. Two approaches are recommended as alternatives to the employment of emulsions. The first is to operate with a water-saturated oil phase, with saturation maintained by the use of a compartmentally separated saturated salt solution (Fig. 2). The principle which drives water transport from the saturated salt solution to the lipophilic reaction solution is the thermodynamic requirement of equal water activity (i.e., chemical potential) for all phases. Hence, water consumed by the reaction will be replenished by the saturated salt solution. The water content of the lipolysate will increase during the course of reaction for a given water activity due to the formation of FFA, a surface-active agent. A salt which yields a high water activity, such as NaCl, should be employed. Two configurations are presented in Fig. 2, one where a saturated salt solution is circulated through the lipolysate through semi-permeable tubing of hydrophobic material. The second configuration employs a common vapor head space between the saturated salt solution and the reservoir. The second alternative to emulsion-based systems is to employ membrane bioreactors, in which the membrane compartmentalizes the water and oil phases. Both flat membrane and hollow microfiber configurations have been employed.[15] Membrane bioreactors allow for continuous-mode operation and the continuous removal of glycerol, the latter of which will

thermodynamically drive hydrolysis over its reverse reaction, esterification. However, the barrier to mass transport presented by the membrane and the low activity of lipases immobilized to the membrane surface lessen enthusiasm for this approach. A nitrogen-rich, oxygen-poor reactor head space is recommended for all lipolytic reactions that produce PUFA or hydroxy FFA to reduce oxidative degradation.

ISOLATION OF PUFA VIA LIPASE-CATALYZED SELECTIVE ESTERIFICATION

In contrast to lipolysis, the use of selective lipase-catalyzed esterification (reaction 2 of Fig. 1) in conjunction with a purification method is very successful for isolating PUFA.[16] Due to the discrimination most lipases possess against Δ4–Δ6 FFA (particularly from *Rhizopus* sp., *G. candidum*, *Candida rugosa*, and *Chromobacterium viscosum*), acyl groups such as the C18 Δ9s are rapidly esterified, while the Δ4–Δ6 PUFA are esterified much more slowly. As a first step, FFA are derived from a TAG source, for instance, by employing lipolysis. Next, selective esterification is operated in a stirred batch reactor containing the FFA source, an acyl acceptor (in stoichiometric excess), lipase, and perhaps a solvent. The acyl acceptor chain length (n-C_{12} reported optimal[17]), substrate ratio, temperature, water activity level, biocatalyst type and concentration, and extent of reaction are variables that should be optimized by the user in preliminary small-scale batch reactions.[18] Optimal conditions represent a position on the hyperbolic plot of purity versus yield that is dictated by economic factors. The time course of reaction is monitored so that the reaction can be stopped upon reaching the optimal extent of reaction by removal of the lipase via filtration or sedimentation. Saponification is then employed to isolate the PUFA-rich FFA from the reaction mixture. As an alternative to saponification, FFA, esters, and unreacted alcohol can be fractionated by short-path molecular distillation if long-chain acyl acceptors are used. An additional cycle of lipase-catalyzed selective esterification applied to the recovered PUFA-enriched FFA fraction will improve the purity of PUFA at the expense of reduced yield. For PUFA isolation from FFA mixtures derived from their natural sources (Table 1), typical purities and yields are 73–95% and 60–90%, respectively. To improve the PUFA purity, urea inclusion compounds (UICs) are frequently employed, as discussed above. Lipase-catalyzed alcoholysis (reaction 3 of Fig. 1) or hydrolysis of fatty acid methyl or ethyl ester mixtures is equally effective in discriminating against Δ4–Δ6 acyl groups. Assuming that the acyl acceptor substrate and acceptor group in the ester substrate in a selective alcoholysis reaction are significantly different in chain length, Δ4–Δ6 polyunsaturated fatty acyl-rich esters can be isolated by short-path molecular distillation. Lipase-catalyzed esterification to isolate PUFA has been scaled up to the 1–10 kg level.[18]

Lipase-selective esterification has recently helped fractionate PUFA species present in the same mixture. Fish oil-derived FFA was esterified with glycerol using Lipozyme IM in solvent-free media to fractionate DHA and EPA.[19] The unesterified FFA contained 78% DHA and only 3% EPA, with a 79% recovery of DHA; the esters contained the majority of the EPA and the other acyl groups, with a 91% recovery of EPA.[19] Lipases have also successfully fractionated the two most abundant species of CLAs: 18:2-9c,11t and 18:2-10t,12c from an equimolar mixture derived from acid-catalyzed isomerization of ALA.[20]

STRUCTURED AND PUFA-ENRICHED TAG

Structured TAG are defined as TAG molecules that contain mixtures of short- (C_1–C_4), medium- (C_6–C_{12}), and long- (C_{14} or higher) chain acyl groups, with a given acyl group types frequently being confined to either the 1-(3-) or 2-position on the glycerol backbone.[21] (TAG rich in medium chain acyl groups are clinical pharmaceuticals for patients with lipid absorption or digestion disorders and high- and rapid-energy sources for athletes since medium chain acyl groups are readily metabolized via the portal vein and generally are not stored in adipose cells for long-term use.) Structured TAG have many applications as nutraceuticals. Commercial examples include Salatrim™ from Danisco Cultor (dietary agents consisting of TAG with at least one long-chain saturate, e.g., 18:0, and one short-chain acyl group), Caprenin™ and Caprucin™ from Procter and Gamble, dietary agents containing C_8, C_{10}, and behenic (22:0), or erucic (22:1-13) acyl groups, respectively, Captex™ 810-D from Abitec, Columbus, Ohio, U.S.A. (cosmetic agents containing TAG enriched in C_8, C_{10}, and 18:2 acyl groups), and Structolipid™ from Fresenius-Kabi AB, Sweden (parenteral nutrients formed by random interesterification of coconut and soybean oils). All of these products contain TAG with random distribution of the named acyl groups among the three acylglycerol positions.

The employment of 1,3-selective lipases leads to structured TAG where specific acyl groups are confined to either the 1-(3-) or 2-acylglycerol position. One of the earliest examples is cocoa butter substitute formed by 1,3-selective lipase-catalyzed acidolysis (reaction 1 of Fig. 3) of palm oil midfraction by palmitic acid (resulting in the replacement of 1-, 3-dipalmityl, 2-oleyl TAG by 1-(3-) palmityl, 2-oleyl, 3- (1-) steryl TAG and 1-, 3-disteryl, 2-oleyl TAG, both of which are abundant in cocoa butter). The product

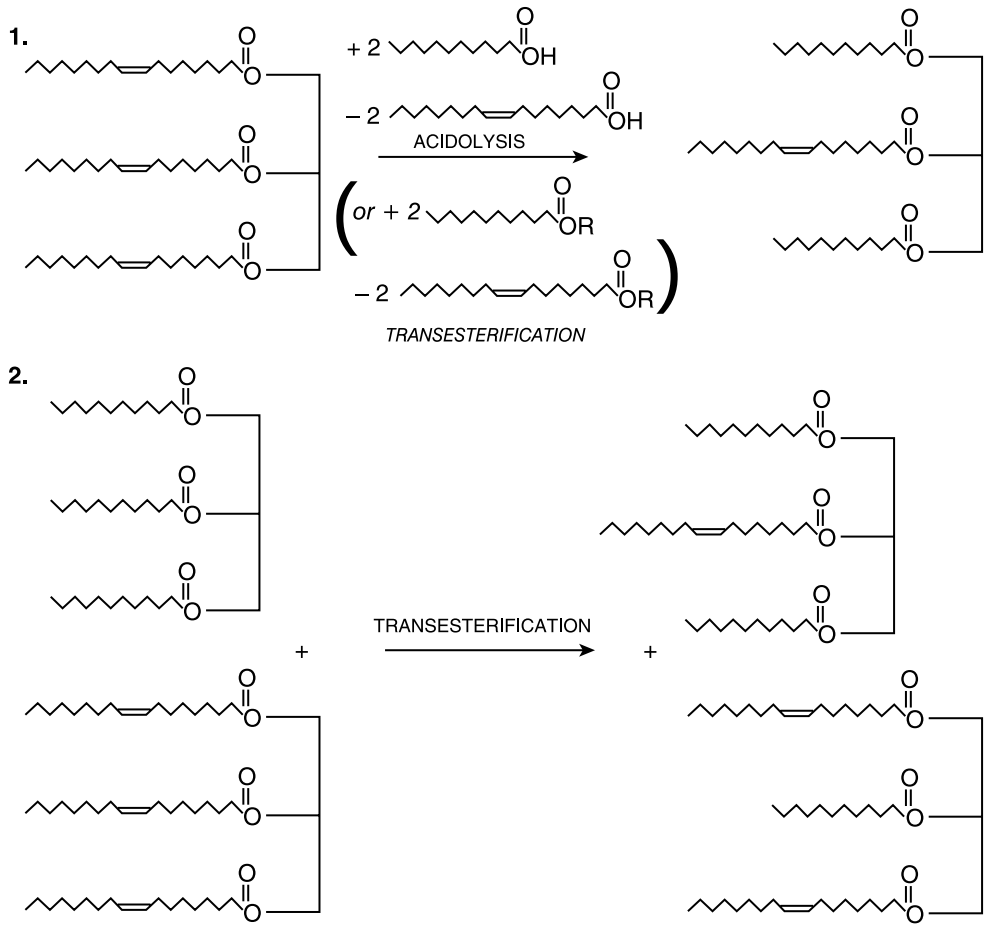


Fig. 3 Reactions catalyzed by 1,3-selective lipases to synthesize structured TAG. (From Ref.[6].)

Bohenin™, consisting mostly of 1,3-behenyl, 2-oleyl TAG, is synthesized using lipases by Fuji Oil, Osaka, Japan, as a cocoa butter improver. A second product area is human milk fat substitute for use in formula for both prematurely born and full-term infants. Mammalian milk frequently contains TAG possessing a palmitic acyl group at the 2-position. The commercial product Betapol™ from Loders Croklaan, Wormerveer, Netherlands, a TAG product enriched in 1,3-dioleyl, 2-palmityl TAG, recently received the rating ''generally regarded as safe (GRAS)'' by the US Food and Drug Administration (FDA). Betapol™, the first commercial product employing lipases to create structured TAG, improves the absorption of dietary fat and calcium by infants, and reduces constipation. Infant formula that contains PUFA-acyl groups at the 1- and 3-acylglycerol position is suggested to enhance brain and nervous system development.[22] Positioning of PUFA groups in the outer acylglycerol positions allows for proper digestion by infants, since only the 2-position acyl group is strongly absorbed in vivo.[22] This fact suggests that structured TAG containing PUFA at the 2-position and medium-chain acyl groups at the 1- and 3-positions may be a useful nutraceutical, providing the benefits described above for Salatrim™, Caprenin™, and Caprucin™, with the added benefit of an essential FFA being adsorbed by the body. Such a structured TAG has been suggested as a nutrition source for patients with pancreatic deficiencies. However, a problem recently discovered for structured TAG prepared by 1,3-selective lipases is the product's relatively low oxidative stability.[23]

Synthesis of a structured TAG with PUFA confined to either the 1-(3-) or 2-position requires 1,3-selective lipases to interesterify a TAG using FFA or fatty acid esters (reaction 1), or to transesterify two different populations of TAG molecules (Fig. 3). Alternatively, a two-step approach can be employed, where a TAG is first subjected to 1,3-selective lipolysis, resulting in 2-MAG that is carefully isolated, then crystallized, to prevent acyl migration.[24] The second step is 1,3-selective esterification of the 2-MAG.[24] Such an approach results in a 72–85% yield and >95%

purity.[24] In addition, the acidolysis of a medium acyl chain-rich TAG by PUFA-rich FFA was enhanced by use of a membrane bioreactor in which the released medium-chain FFA were selectively removed by their permeation through the membrane.[25] To enrich a common TAG mixture in PUFA, either 1,3- selective or ''random'' lipases are employed to catalyze interesterification using PUFA-enriched FFA (acidolysis, reaction 4 of Fig. 1) or fatty acid methyl or ethyl ester (transesterification, reaction 5 of Fig. 1). PUFA-enriched MAG/DAG/TAG, produced via selective lipolysis (discussed above), can replace the TAG substrate, resulting in a TAG highly enriched in PUFA. An alternative approach to produce PUFA-enriched TAG is direct esterification between glycerol and PUFA-rich acyl groups.

When conducting interesterification for structured and PUFA-enriched TAG synthesis, the three main goals are high activity, low occurrence of hydrolysis (resulting in DAG and perhaps MAG), and, for structured lipid synthesis, low occurrence of acyl migration, because this reaction will lead to formation of TAG with undesired structure. One must optimize reaction parameters: residence time, reaction temperature, and fatty acyl donor-to-TAG substrate ratio, to arrive at a desired position between the two extremes of high percent interesterification and low degrees of acyl migration and hydrolysis.[26] An optimal water activity also exists: suboptimal water content reduces biocatalytic activity, while high water content leads to increased hydrolysis and acyl migration.[26] A continuous-mode packed bed bioreactor is recommended because it lowers residence time of reaction mixture in the presence of immobilized lipase, which leads to reduced acyl migration.[26]

PUFA AND HYDROXY ACYL-RICH *MAG*, *DAG*, AND POLYOL ESTERS

MAG are well-known biodegradable and biocompatible emulsifiers in the foods, dairy, cosmetics (e.g., in toothpastes), and pharmaceutical industries.[27–29] 1-(3-) MAG are easily modified chemically to produce MAG sulfates, cosmetic surfactants which yield low irritability. As discussed above, 2-MAG are feedstock for the lipase-catalyzed synthesis of structured lipids. Approval of MAG as GRAS was recently granted by the FDA. DAG are cocoa butter additives that reduce the extent of crystallization or ''blooming.'' Isomerically pure DAG are possible feedstocks for the synthesis of glycerylphospholipids, glycolipids, and pharmaceuticals.[29] Current industrial methods to produce MAG and DAG (e.g., glycerolysis directed by a heterogeneous catalyst) involve high temperatures (~180–220°C or 30–160°C if a high vacuum pressure of 200–400 Pa is applied) that may produce byproducts that promote off-flavors, -odors, or -colors, and a broad product distribution containing various MAG, DAG, and TAG species.[27,29] Molecular distillation must be applied to remove the impurities, which increases operating costs and can result in further chemical degradation. Therefore, enzymatic preparation of MAG and DAG may gain further interest if energy costs continue to increase and will be of particular interest for products that contain degradation-susceptible PUFA and oxygenated acyl groups. PUFA-enriched MAG (and polyol monoesters, discussed below) are food emulsifiers that provide essential FFA nutrients. Hydroxy acyl-rich MAG are waxy materials employed in lipsticks.

Lipase-catalyzed synthesis has been employed to synthesize FFA partial and poly-esters of glycols such as ethylene and propylene glycol, neopentanol, trimethylolpropane, and polyglycerol as a low-cost and low-temperature alternative to chemical processing for applications as biodegradable and biocompatible lubricants and emulsifiers in the food and cosmetics industry. Saccharide acyl acceptors (e.g., fructose, glycols, sucrose, xylose, maltose, and trehalose) have received the most attention of the polyols due to the narrow product distribution provided by regioselective lipases.[30] For instance, 1,3-selective lipase-catalyzed esterification of 1,2-propanediol results in only the 1-monoester. The substrate ratio, water activity, and reaction medium polarity can also be selected to control the relative proportion of mono- and di-esters. As a second example, lipases esterify only the primary OH groups of sucrose, at the 6, 6′ and 1′ positions, with the three hydroxyls listed in the order of preference by lipases. Chemical synthesis of sucrose esters requires temperatures above 100°C and results in a broad product distribution.

Lipases catalyze MAG (or polyol ester) synthesis via esterification or glycerolysis (alcoholysis) in nonaqueous media (Fig. 4). The challenge to overcome when conducting lipase-catalyzed reactions is the poor miscibility of glycerol (or polyol) and the acyl donor-rich lipophilic phase. The employment of polar solvents (e.g., *tert*-butanol, acetone, and, recently, room temperature ionic liquids) increases miscibility and the product distribution in favor of mono- rather than di- or poly-esters, but decreases the catalytic rates because of the partitioning away of essential water molecules from lipase to the solvent. Other approaches include the use of polyhydric alcohol complexation agents such as phenylboronic acid, the use of protective groups such as isopropylidene (which for glycerol covalently blocks access to two of its three hydroxyls, resulting in regioisomerically pure 1-(3-) MAG), water-in-oil microemulsions, and the suspension of silica gel saturated with glycerol in the reaction medium,

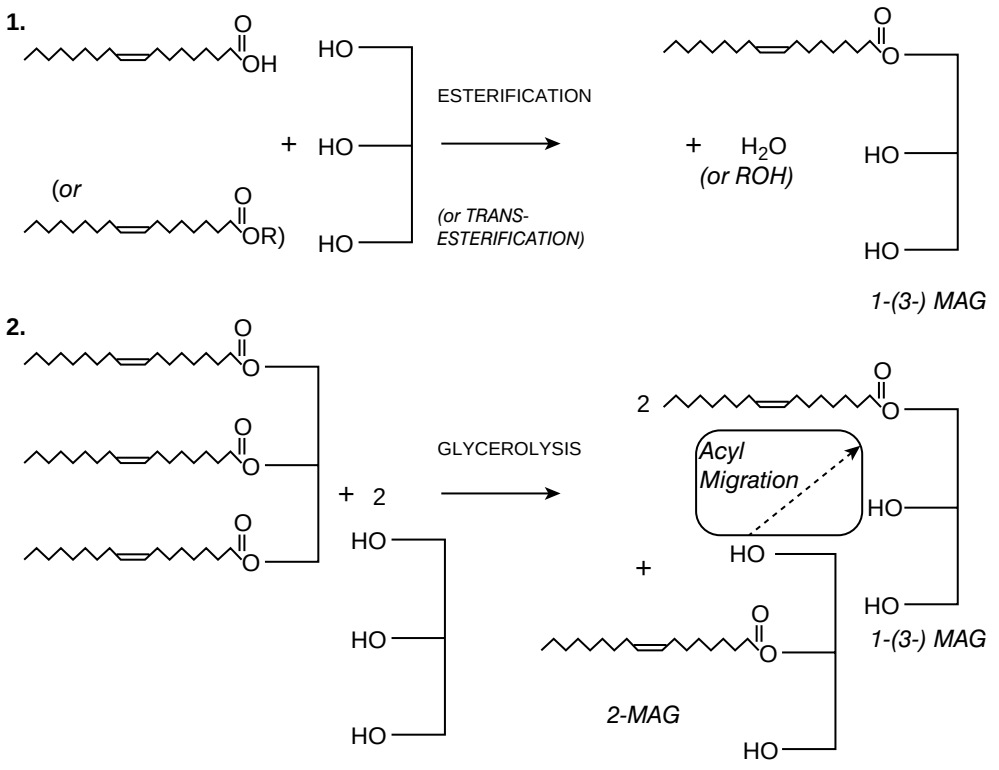


Fig. 4 Reactions catalyzed by 1,3-selective lipases to synthesize monoacylglycerol (MAG). (From Ref.[6].)

all of which are impractical for large-scale manufacturing. Of note, glycosides (saccharide alkyl ethers) can be readily formed by chemical and enzymatic means (glucosidases) to solve the problem of poor substrate miscibility for saccharides. *Penicillium camembertii* lipase-catalyzed esterification will produce MAG and DAG, but not TAG due to its unique substrate selectivity.

POLY (HYDROXY FATTY ACID) ESTERS

Random lipases are known to catalyze the formation of poly(hydroxy) fatty acids, primarily ricinoleic acid and its product via hydrogenation, 12-hydroxystearic acid; however, polymerization also occurs readily by chemical processing.[31] The value of lipase-catalyzed processing is the ability of 1,3-selective lipases to permit esterification of the free carboxyl moiety without cleavage of the ester bonds that join together the hydroxyacyl monomeric units. Esterification improved the physical properties of poly(hydroxy acids) as lubricant materials, evidenced by their reduced viscosity change with temperature, i.e., increased viscosity index.[31] Monoesters of polyglycerol and poly(ricinoleic acid) are well known emulsifiers for the food industry. Star polymers have been synthesized recently from polyol and poly(ricinoleic acid), e.g., pentaerythritol-poly(ricinoleic acid) tetraester, using lipases in nonaqueous media. Such materials have very good lubricant properties and may have utility as drug delivery vehicles.[32]

STERYL ESTERS

Lipases have also been employed to isolate tocopherols and sterols, important antioxidants from ''deodorizer distillate,'' a by-product formed during the deodorization step of seed oil purification, by hydrolyzing MAG and DAG in the ''distillate,'' and by esterifying sterols into sterol esters.[33] The latter serves as antioxidant in non-polar food products such as margarine (e.g., Benacol™ from the Raisio Group, Finland and ''Take Control''™ from Unilever, Englewood Cliffs, New Jersey, U.S.A.). PUFA-sterol esters would provide the benefits of essential fatty acid intake and antioxidant protection.

PEROXIDATED FFA

Peroxy fatty acids, R-COOOH, formed from the esterification of FFA or alcoholysis of fatty acid methyl or

ethyl esters with hydrogen peroxide, H_2O_2, are oxidants employed in the chemical industry for epoxidation of double bonds, conducting Baeyer-Villiger oxidation reactions, and hydroxylation of aromatic rings and amines. Traditionally, this reaction occurs via catalysis using a strong acid; however, for long-chain and unsaturated FFA, harsh operating conditions (strong acids and high temperature) are required and by-products frequently occur. Thus, biocatalysis may be a safer alternative.[34] Lipase-catalyzed peroxidation of PUFA yields epoxy FFA, of potential use for paints, coatings, and disinfection agents (sporicides).[34] Epoxy FFA occur due to the isomerization, or ''self-epoxidation'' of peroxy FFA. For instance, peroxidation of ALA results in peroxy-ALA, which isomerizes, or ''self-epoxidizes,'' into a mixture of 9,10-epoxy, 12,13-epoxy, and 15,16-epoxy FFA. Lipase-catalyzed peroxidation typically results in a 70–90% yield of epoxy-FFA.[34] Hydroxy and peroxy acids are also formed from the chemical rearrangement of hydroperoxides, a class of lipid products formed via lipoxygenase-catalyzed oxidation of PUFA,[35] a topic that is beyond the scope of this entry.

Although enzymes are very susceptible to denaturation by H_2O_2, Novozyme^{TM} appears to be reasonably stable. To reduce denaturation, fed-batch addition of 30–60% aqueous H_2O_2 to the pure FFA or alkyl ester is employed. Typical solvents used for this reaction are toluene and dichloromethane.

CONCLUSIONS

The employment of lipases to isolate polyunsaturated and oxygenated FFA and transform them into ester products will expand in the near future, especially for purifying PUFA and synthesizing structured TAG. Biocatalytic modification will become more competitive with chemical processing if energy costs continue to rise. Advances in lipase bioreactor design and process scale-up technology will be an area of focus in future research, including the implementation of reactive separations. Along the same lines, multistep processing schemes composed of biocatalytic and separation steps to convert raw lipid feedstock into products of interest such as structured TAG, have improved in their sophistication during the past five years, with the trend expected to continue, particularly with regard to the integration of reaction and separation steps. Bioreactor design will also be developed for multiple enzyme systems, for instance, the combination of random lipases and lipoxygenases to transform TAG into hydroperoxides. Another example may be a multistage bioreactor where random lipases hydrolyze PUFA-enriched TAG in an upstream stage, and 1,3-selective lipases catalyze the acidolysis of a tripalmitin feed stream with the PUFA-enriched FFA transported from the upstream portion of the process in a downstream stage, resulting in an infant formula nutraceutical.

Although acyl migration and hydrolysis side reactions cannot be prevented in manufacturing TAG, future bioreactor designs will facilitate the programming of operational parameters such as water activity, solvent system polarity, and temperature to improve yield and recovery. The ability to program the water activity and other parameters effectively will require improved capabilities in kinetic modeling and means of in situ control and rapid measurement of substrate and product concentration. The in situ measurement of water activity is being developed.[36] It is highly desired that a lipase be produced that will be less discriminatory against Δ4–Δ6 and long-chain saturated FFA to improve the rate and extent of PUFA acyl incorporation into ester products. Advancements in directed evolution, extremophile discovery and screening, and recombinant DNA technology will lead to more active and stable lipases that possess narrower substrate selectivity.

In conclusion, although a mature field, lipase-catalyzed lipid modifications will require further improvement along the lines given in the preceding paragraphs for the further use of PUFA and hydroxy acyl material as chemical and biological intermediates. Although alternate technologies (lipid purification, chemical modification, transgenic plants,[37] and microbial bioconversions[38]) will improve, the selectivity of lipase-based processing will continue to make it a viable choice as unit operation in a lipid ''biorefinery'' as a substitute for petrochemical-based processes.

REFERENCES

1. Bornscheuer, U.T. Ed. *Enzymes in Lipid Modification*; Wiley-VCH: Weinheim, Germany, 2000.
2. Kuo, T.M.; Gardner, H.W. *Lipid Biotechnology*; Marcel-Dekker: New York, 2002.
3. Li, D.; Bode, O.; Drummond, H.; Sinclair, A.J. Omega-3 (n-3) fatty acids. In *Lipids for Functional Foods and Nutraceuticals*; Gunstone, F.D., Ed.; Oily Press Lipid Library: Bridgwater, UK, 2003; 225–262.
4. Pariza, M.W.; Park, Y.; Cook, M.E. The biologically active isomers of conjugated linoleic acid. Prog. Lipid Res. **2001**, *40* (4), 283–298.
5. Schwitzer, M.K. Castor oil. Proceedings World Conference on Oleochemicals: Into the 21st Century; Applewhite, T.H., Ed.; AOCS Press: Champaign, IL, 1991; 111–118.
6. Hayes, D.G. Enzyme-catalyzed modification of oilseed materials to produce eco-friendly

products. J. Am. Oil Chem. Soc. **2004**, *81* (12), 1077–1103.
7. Jaeger, K.E.; Dijkstra, B.W.; Reetz, M.T. Bacterial biocatalysts: molecular biology, three-dimensional structures, and biotechnological applications of lipases. Ann. Rev. Microbiol. **1999**, *53* (1), 315–351.
8. Halling, P. Enzymic conversions in organic and other low-water media. In *Enzyme Catalysis in Organic Synthesis*; 2nd Ed.; Drauz, K., Waldmann, H., Eds.; Wiley-VCH: Weinheim, Germany, 2002; 259–285.
9. Villeneuve, P.; Foglia, T.A. Lipase specificities: potential application in lipid bioconversions. Inform **1997**, *8* (6), 640–651.
10. Fureby, A.M.; Virto, C.; Adlercreutz, P.; Mattiasson, B. Acyl group migrations in 2-monoolein. Biocatal. Biotransform. **1996**, *14* (2), 89–111.
11. Hayes, D.G. Purification of free fatty acids via urea inclusion compounds. In *Handbook of Functional Lipids*; Akoh, C.C., Ed.; CRC Press: Boca Raton, FL, 2005; 77–88.
12. Shimada, Y.; Sugihara, A.; Tominaga, Y. Enzymatic enrichment of polyunsaturated fatty acids. In *Lipid Biotechnology*; Kuo, T.M., Gardner, H.W., Eds.; Marcel Dekker: New York, 2002; 493–515.
13. Hayes, D.G.; Kleiman, R. 1,3-specific lipolysis of *Lesquerella fendleri* oil by immobilized and reverse micellar encapsulated lipases. J. Am. Oil Chem. Soc. **1993**, *70* (11), 1121–1127.
14. Derksen, J.T.P.; Krosse, A.M.; Tassignon, P.; Cuperus, F.P. Lipase-catalyzed production of functionalized fatty acids from *Dimorphotheca pluvialis* seed oil. Mededelingen van de Faculteit Landbouwwetenschappen, Universiteit Gent **1992**, *57* (4a), 1741–1747.
15. Malcata, F.X.; Reyes, H.R.; Garcia, H.S.; Hill, C.G., Jr.; Amundson, C.H. Immobilized lipase reactors for modification of fats and oils – a review. J. Am. Oil Chem. Soc. **1990**, *67* (12), 890–910.
16. Shimada, Y.; Sugihara, A.; Tominaga, Y. Enzymatic purification of polyunsaturated fatty acids. J. Biosci. Bioeng. **2001**, *91* (6), 529–538.
17. Shimada, Y.; Sugihara, A.; Nakano, H.; Kuramoto, T.; Nagao, T.; Gemba, M.; Tominaga, Y. Purification of docosahexaenoic acid by selective esterification of fatty acids from tuna oil with Rhizopus delemar lipase. J. Am. Oil Chem. Soc. **1997**, *74* (2), 97–101.
18. Shimada, Y. Application of lipase reactions to separation and purification of useful materials. Inform **2001**, *12* (12), 1168–1174.
19. Halldorsson, A.; Kristinsson, B.; Glynn, C.; Haraldsson, G.G. Separation of EPA and DHA in fish oil by lipase-catalyzed esterification with glycerol. J. Am. Oil Chem. Soc. **2003**, *80* (9), 915–921.
20. Nagao, T.; Yamauchi-Sato, Y.; Sugihara, A.; Iwata, T.; Nagao, K.; Yanagita, T.; Adachi, S.; Shimada, Y. Purification of conjugated linoleic acid isomers through a process including lipase-catalyzed selective esterification. Biosci. Biotechnol. Biochem. **2003**, *67* (6), 1429–1433.
21. Osborn, H.T.; Akoh, C.C. Structured lipids–novel fats with medical, nutraceutical, and food applications. Compreh. Rev. Food Sci. Food Safety **2002**, *1* (3), 93–103.
22. Yamane, T. Lipase-catalyzed synthesis of structured triacylglycerols containing polyunsaturated fatty acids: monitoring the reaction and increasing the yield. In *Enzymes in Lipid Modification*; Bornscheuer, U.T., Ed.; Wiley-VCH: Weinheim, Germany, 2000; 148–169.
23. Timm-Heinrich, M.; Skall Nielsen, N.; Xu, X.; Jacobsen, C. Oxidative stability of structured lipids containing C18:0, C18:1, C18:2, C18:3 or CLA in sn2-position – as bulk lipids and in milk drinks. Innovat. Food Sci Emerg. Technol. **2004**, *5* (2), 249–261.
24. Schmid, U.; Bornscheuer, U.T.; Soumanou, M.M.; McNeill, G.P.; Schmid, R.D. Highly selective synthesis of 1,3-oleoyl-2-palmitoylglycerol by lipase catalysis. Biotechnol. Bioeng. **1999**, *64* (6), 678–684.
25. Xu, X.; Skands, A.; Jonsson, G.; Adler-Nissen, J. Production of structured lipids by lipase-catalysed interesterification in an ultrafiltration membrane reactor. Biotechnol. Lett. **2000**, *22* (21), 1667–1671.
26. Xu, X. Enzymatic production of structured lipids: process reactions and acyl migration. Inform **2000**, *11* (10), 1121–1131.
27. Bornscheuer, U.T. Lipase-catalyzed synthesis of monoacylglycerols. Enzyme Microb. Technol. **1995**, *17* (7), 578–586.
28. Boyle, E. Monoglycerides in food systems: current and future uses. Food Technol. **1997**, *51* (8), 52–5456, 58–59.
29. Diks, R.M.M.; Bosley, J.A. The exploitation of lipase selectivities for the production of acylglycerols. In *Enzymes in Lipid Modification*; Bornscheuer, UT, Ed.; Wiley-VCH: Weinheim, Germany, 2000; 3–22.
30. Sarney, D.B.; Vulfson, E.N. Enzymatic synthesis of sugar fatty acid esters in solvent-free media. In *Enzymes in Nonaqueous Solvents: Methods and Protocols (Methods in Biotechnology Vol. 15)*; Vulfson, E.N., Halling, P.J., Holland, H.L., Eds.; Humana Press: Totawa, 2001; 531–543.

U

31. Hayes, D.G. The catalytic activity of lipases toward hydroxy acids (a review). J. Am. Oil Chem. Soc. **1996**, *73* (5), 543–549.
32. Hayes, D.G. Lipase-catalyzed synthesis of polyhydric alcohol-poly(ricinoleic acid) ester star polymers. Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.) **2005**, *46* (1), 276–277.
33. Weber, N.; Weitkamp, P.; Mukherjee, K.D. Cholesterol-lowering food additives: lipase-catalysed preparation of phytosterol and phytostanol esters. Food Res. Intl. **2002**, *35* (2/3), 177–181.
34. Rüsch gen. Klaas, M.; Warwel, S. Lipase-catalyzed peroxy fatty acid generation and lipid oxidation. In *Enzymes in Lipid Modification*; Bornscheuer, U.T., Ed.; Wiley-VCH: Weinheim, Germany, 2000; 116–127.
35. Iacazio, G.; Martini-Iacazio, D. Properties and applications of lipoxygenases. In *Enzymes in Lipid Modification*; Bornscheuer, U.T., Ed.; Wiley-VCH: Weinheim, Germany, 2000; 337–359.
36. Kang, I.J.; Rezac, M.E.; Pfromm, P.H. Membrane permeation based sensing for dissolved water in organic micro-aqueous media. J. Membr. Sci. **2004**, *239* (2), 213–217.
37. Sayanova, O.V.; Napier, J.A. Eicosapentaenoic acid: biosynthetic routes and the potential for synthesis in transgenic plants. Phytochemistry **2004**, *65* (2), 147–158.
38. Kuo, T.M.; Kaneshiro, T.; Hou, C.T. Microbiological conversions of fatty acids to value-added products. In *Lipid Biotechnology*; Kuo, T.M., Gardner, H.W., Eds.; Marcel Dekker: New York, 2002; 605–628.

Vapor–Liquid–Solid Synthesis of Nanowires

Brian A. Korgel
Tobias Hanrath
Forrest M. Davidson, III
The University of Texas, Austin, Texas, U.S.A.

INTRODUCTION

Fundamental aspects of vapor–liquid–solid (VLS) semiconductor nanowire growth are presented here. The synthesis of VLS semiconductor has been extended to different reaction media and pathways from the early chemical vapor deposition (CVD) approach, including solution–liquid–solid (SLS) and supercritical fluid–liquid–solid (SFLS), laser-catalyzed growth, and vapor–liquid–solid-epitaxy. The properties of nanowires grown by these VLS embodiments are compared. In this entry, VLS growth of nanowire heterostructures and oriented and hyperbranched arrays is examined. In addition, surface passivation and functionalization are assessed, and the importance of these techniques in the progress toward VLS produced nanowire devices is detailed.

BACKGROUND

Semiconductor nanowires are ultrahigh aspect ratio (>1000) crystals that are micrometers in length and <50 nm in diameter. Because of their nanometer diameter and high aspect ratio, they exhibit unique size-dependent optical, electronic, and mechanical properties. When the nanowire diameter is smaller than the Bohr exciton diameter, quantum confinement occurs, changing the optical and electronic properties of the semiconductor. These properties might be exploited in new optical sensor, light-emitting diode, or field effect transistor (FET) applications. Semiconductor nanowires are also extremely flexible, unlike bulk ceramics, bending and flexing like polymers, giving them the ability to combine and interface with plastic substrates and organic electronic materials for flexible display and computing applications. Their very high surface area-to-volume ratios make their properties very sensitive to surface chemistry and to subtle changes in the environment, which may be exploited in high sensitivity sensor applications. Recently, new synthetic approaches for varying the composition—either along the length of the nanowire or radially in the form of core/shell structures—have been developed, which adds another degree of freedom for materials property design. This capability is particularly important for electronic applications, such as computing architectures that require controlled doping, such as p–n–p type heterostructures, and metal–insulator–semiconductor (MIS) devices. Aside from having an immense technological potential, semiconductor nanowires provide a model materials system for experimental studies of fundamental quantum mechanical concepts.

The VLS Growth Mechanism

In the 1960s, Wagner proposed the VLS growth mechanism for crystalline silicon whiskers.[1] In the literature, crystalline wires with diameters exceeding 100 nm have generally been referred to as "whiskers," whereas wires with diameters <100 nm have been called "nanowires." Noncrystalline wires are generally called "fibers." And a "wire" typically has an aspect ratio of at least 100. Since its initial proposal, VLS growth has been observed in a broad range of crystallization environments, including liquids and supercritical solvents—VLS growth is even believed to be responsible for the formation of Fe whiskers on the surface of lunar rock samples.[2]

Wagner, in his early work, formed single crystal Si whiskers by CVD of SiI_2 or $SiCl_4$ on Si(1 1 1) substrates in the presence of metal impurities, such as gold (Au). In VLS growth, the metal impurity promotes whisker crystallization through the formation of a liquid metal : Si alloy droplet. Instead of depositing epitaxially on the underlying substrate, Si adsorbs to the surface of the metal impurity and dissolves into it. An example of the binary equilibrium phase diagram for Au–Si is shown in Fig. 1A. It exhibits a low-temperature liquid eutectic (~360°C), which is lower than the typical CVD temperatures in the range of 500–800°C. Therefore, as Si deposits on the substrate surface, it dissolves into the Au metal "seed" until reaching supersaturation when it nucleates and recrystallizes from the particle surface in the form of a whisker. Under the appropriate deposition conditions, sidewall growth on the whisker surface will occur at a slower rate when compared with the rate of crystallization from the metal seed particle.

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120041403

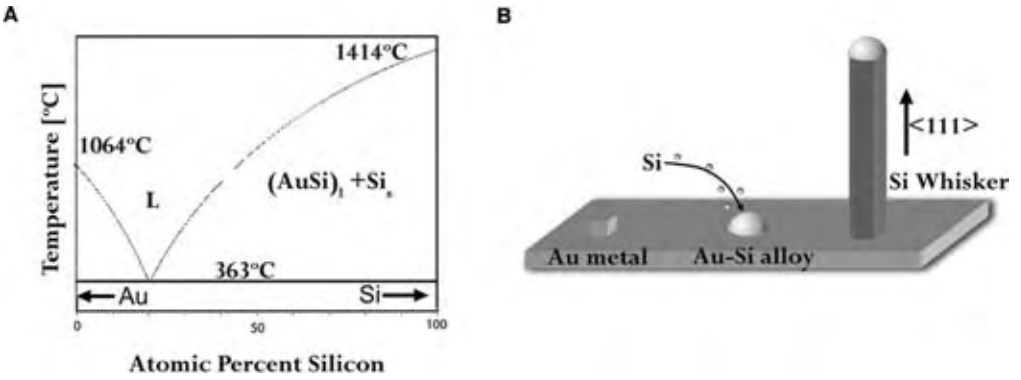


Fig. 1 (A) Binary equilibrium phase diagram for Au–Si. (From Massalski, Binary Alloy Phase Diagrams, Secondary Binary Alloy Phase Diagrams, 1986) and (B) schematic illustration of the VLS growth mechanism. (From Ref.[1].) *(View this art in color at www.dekker.com.)*

Au forms a relatively low-temperature eutectic with many different semiconductors, including Ge and GaAs and, therefore, has been used extensively to seed the growth of many different semiconductor whiskers. However, there is nothing unique about the properties of Au for VLS growth, and high-quality crystalline semiconductor whiskers have been grown using many different metals as seeds, including Ag, Cu, Ni, Pd, and Pt.[1] The only requirement for VLS growth is that the deposition temperature should be greater than the metal : semiconductor eutectic.

In Fig. 1B, the VLS process is illustrated. At reaction temperatures exceeding the eutectic, Si vapor dissolves in the Au seed to form a liquid alloy droplet. As more gaseous Si dissolves in the alloy, the droplet becomes supersaturated, and Si crystallizes on the droplet surface. The shape of the droplet interface and the surface tension difference between the liquid alloy and the solid semiconductor promote crystallization in one direction. In many cases, the liquid alloy droplet displaces off the surface and ''rides'' on the top of the vertically growing whisker—crystallization continues to occur at the liquid–solid interface as the metal seed is continuously fed with Si from the gas phase.

For high-quality whiskers with a consistent diameter along their lengths, the CVD deposition parameters must be optimized to minimize sidewall deposition and to promote relatively rapid whisker growth. The deposition conditions must be at sufficiently high temperature and precursor concentration to feed sufficient semiconductor to the liquid metal : semiconductor seeds to maintain consistent whisker growth. Fluctuations in semiconductor concentration, or temperature, lead to defects in the whisker, such as changes in the growth direction, bends in the wire, or even the formation of an amorphous phase. The semiconductor feed rate, however, must be slow enough to limit unwanted sidewall deposition on the surface of the whisker, which leads to polycrystallinity and large variations in diameter along the whisker length. The ''nutrient'' concentration must sustain steady whisker growth, while not inducing sidewall deposition or homogeneous nucleation on the substrate.

Crystallographic characterization has revealed that nearly all Si whiskers produced by Au-seeded VLS by CVD are oriented with their long axes in the $\langle 1\,1\,1\rangle$ direction. Some whiskers are observed to be oriented with their long axes extended in the $\langle 2\,1\,1\rangle$ or $\langle 1\,1\,0\rangle$ axis. This observation is somewhat unexpected, as the rate of epitaxial Si deposition on monolithic substrates by CVD occurs much faster on $\{2\,1\,1\}$ and $\{1\,1\,0\}$ surfaces than on $\{1\,1\,1\}$ surfaces. The preferential $\langle 1\,1\,1\rangle$ growth direction of Au-seeded Si whiskers is believed to relate to the formation of an energetically favorable flat $\{1\,1\,1\}$ interface with the liquid alloy seed droplet.

Measurements of whisker crystallization rates by VLS have revealed that the linear growth rate is proportional to the whisker diameter, with larger fibers growing faster than smaller ones.[3] The observed diameter-dependence has been attributed to differences in seed saturation concentration resulting from the curved interface of the seeds. The Gibbs–Thomson equation provides the relationship between the sphere diameter, d, and the chemical potential difference between the nutrient phase (e.g., molecular Si in the vapor) and the liquid alloy droplet $\Delta\mu$:

$$\Delta\mu = \Delta\mu_0 - (4\Omega\gamma_{\mathrm{SF-L}})\left(\frac{1}{d}\right) \quad (1)$$

where $\Delta\mu_0$ is the chemical potential difference at a flat interface, $\gamma_{\mathrm{SF\text{-}L}}$ the surface energy of the vapor–liquid alloy interface, and Ω the atomic volume of the liquid alloy.

As the smaller seed particles have a higher solubility, the seed supersaturation, or the Si chemical potential between that dissolved in the liquid droplet and

adsorbed to the solid crystal, is lower. In fact, Eq. (1) predicts a lower size limit for the seed when VLS growth becomes thermodynamically unfavorable. Recent experiments measuring Si/Ge[4] and Si[5] nanowire growth rates as a function of seed diameter have verified slower growth from small seeds (vide infra).

In early VLS whisker growth studies, there was no attempt to control the liquid alloy droplet size. Consequently, whisker diameters were large, ranging from 100 nm to 0.2 mm. The smallest whisker diameters that could be produced were limited by the minimum stable liquid droplet diameter (~100 nm). In 2000, researchers began to apply size-selected nanometer-size colloidal metal particles as seeds to promote semiconductor nanowire growth.[6,7]

Solution-Phase Nanowire Growth

Nearly 30 years after the discovery and development of VLS whisker growth, the drive toward nanoscale miniaturization of electronic devices and the development of ''bottom-up'' nanocircuit assembly, in which integrated devices are ''self-assembled'' from prefabricated nanocrystal building blocks, pushed research on alternative synthetic strategies for nanometer-scale semiconductor crystals. In 1993, two years after the discovery of single-wall carbon nanotubes, Heath and LeGoues reported the synthesis of highly anisotropic crystalline Ge nanowires with diameters ranging from 7 to 30 nm and lengths up to 10 μm.[8] Their synthesis involved the solvothermal reduction of $GeCl_4$ and phenyl-$GeCl_3$ by Na metal in hexane at 275°C and 100 atm. Although the authors did not attribute nanowire formation to a VLS growth mechanism, most likely liquid Na clusters formed in the reaction ''soup'' promoted anisotropic nanowire crystallization. Interestingly, the Na-seeded Ge nanowires exhibited a ⟨1 1 0⟩ growth direction with a high density of ⟨1 1 1⟩ twins extending parallel to the wire axis, qualitatively different than the ⟨1 1 1⟩ oriented Si and Ge whiskers produced by VLS from Au seeds. The yield of Ge nanowires from this synthesis was relatively low, only ~10–20%, which may have been because of a low reaction yield of Ge in these solution conditions, or the very low Ge : Na ratio used (~1 : 10) in the synthesis—typical semiconductor : metal ratios in most current syntheses range from 10 : 1 to 1000 : 1.

A couple of years later, Buhro and coworkers observed polycrystalline InP, InAs, and GaAs nanowire formations in hydrocarbon solvents from organometallic precursors at relatively low temperatures of ~200°C—well below the Au : Si eutectic temperature.[9] The nanowires had diameters ranging from 10 to 150 nm and their formation was attributed to a VLS-like growth mechanism. The organometallic precursors they used were similar to those used for CVD growth of III–V semiconductor thin films, such as tri-*tert*-butylindane (or gallane) and tri-*tert*-butylphosphine (or arsine). Solution-phase syntheses are limited by the solvent boiling point (~350°C for trioctylphosphine oxide), compared with the typical CVD synthesis temperatures of 500–800°C. At the low temperatures in the solution, reactions of tri-*tert*-butyl precursors are plagued by the incomplete elimination of the alkyl ligand from the Group III precursor, resulting in reaction products composed primarily of organometallic oligomers instead of the crystalline semiconductor material. To promote the reaction at these low temperatures, Buhro and coworkers added small amounts (~10%) of protic reagents, such as methanol, thiophenol, or diethylamine, to catalytically assist the elimination of the alkyl ligands from the Group III and V precursors and to successfully generate the III–V semiconductors at low temperatures (≤200°C).[9] It is well known that the In and Ga precursors decompose to metallic In and Ga at relatively low temperatures—in the range at which these reactions are carried out. Therefore, during the reaction, the III–V semiconductor is produced, but Group III precursor decomposition can also produce Ga or In metal. Both Ga and In have low melting points and form very low-temperature eutectics with GaAs, InAs, GaP, and InP. Therefore, in the reaction, metallic In (or Ga) forms and drives the crystallization of the III–V semiconductors at the relatively low temperature of ~200°C. Because precursor decomposition occurred in solution as opposed to the vapor phase, this metal-seeded semiconductor nanowire growth mechanism was termed ''solution–liquid–solid'' growth.

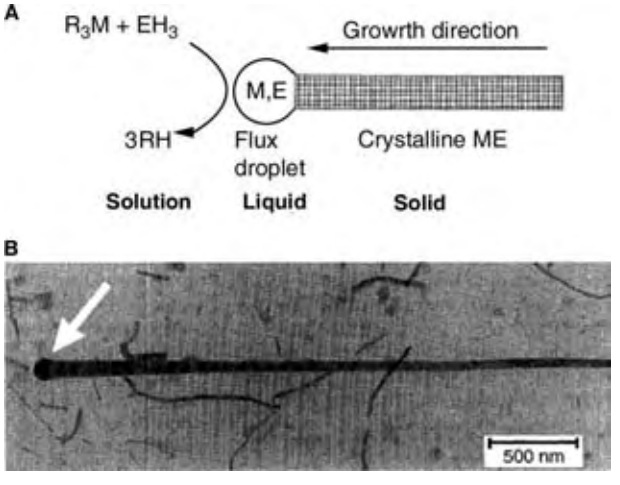


Fig. 2 (A) Schematic of SLS growth mechanism. M and E are elements of the composite semiconductor material. (B) TEM image of an InP whisker. The white arrow points at the In flux particle at the whisker's tip. (From Ref.[9].)

The SLS process was illustrated in Fig. 2A—Ga metal evolves as a reaction byproduct and serves as the metal seed to crystallize the semiconductor nanowire. In Fig. 2B, a transmission electron microscope (TEM) image of an InP nanowire crystallized from liquid In (mp = 157°C) droplets is shown. Note the metal droplet at the end of the nanowire (white arrow in Fig. 2B). The TEM image reveals a tapered nanowire shape with the nanowire diameter decreasing with increasing distance from the seed particle. This tapered shape indicates that the In particle at the tip of the growing InP nanowire grows larger as the nanowire grows. Increasing In accumulates in the seed metal tip over the course of the reaction. The SLS process was later applied to other compound semiconductor nanowires. However, it has been found nearly impossible to limit the size distribution of the metal seeds that are generated in situ during wire growth and that the polydisperse seeds that grow during nanowire elongation give rise to polycrystalline nanowires with diameters that vary by at least one order of magnitude in the sample.

CVD Nanowire Growth

In the gas phase, Westwater et al. narrowed the Si whisker diameter into the nanometer-size range by VLS growth by decomposing SiH_4 over a Si substrate coated with a very thin (0.6 nm) Au film.[10] The nanometer thickness of the thin Au film limited the nanowire diameter to the nanoscale. Their comprehensive survey of critical synthesis parameters, including Si partial pressure and temperature, showed that Si nanowires could be grown with diameters as narrow as 10 nm and successfully demonstrated that CVD–VLS could produce crystalline semiconductor whiskers with nanoscale diameter. However, the small diameter nanowires produced in their studies were of relatively poor quality, because it was found that narrow diameter (~20 nm) wires could only be produced at low Si partial pressure (1 Torr) and low temperature (440°C). The low-temperature slow growth conditions gave nanowires a high density of kinks and crystallographic defects. Si nanowires grown under low Si partial pressure (0.01 Torr) and high temperature (600°C) were much higher quality single crystal nanowires with limited crystallographic defects; however, the nanowires produced under these conditions were very large with diameters of approximately 120 nm. Despite some of the apparent remaining challenges in producing high-quality nanometer diameter Si nanowires by CVD–VLS, Westwater and coworkers revealed the potential to grow nanowires on selected areas of a substrate, as they lithographically patterned the gold films to selectively grow nanowires in desired regions. These studies provided a glimpse into the possibility of integrating VLS nanowire growth into standard Si complementary metal-oxide semiconductor (CMOS) electronics processing.

Using CVD–VLS, Kamins et al. grew Si nanowires from solid $TiSi_2$ seed islands on a Si substrate.[11] Ti-containing seed islands were formed on the substrate via CVD of $TiCl_4$ in a hydrogen atmosphere at 640°C, and in select cases a 920°C thermal anneal was used to form $TiSi_2$ islands. These islands were employed as nucleation sites for Si nanowires, which were subsequently grown by introducing either SiH_4 or $SiCl_2H_2$ at growth temperatures ranging between 640°C and 920°C. At low-growth temperatures (640°C), a small number of long crystalline nanowires formed with the majority of the reaction product consisting of short wires with multiple crystallographic defects. At high temperatures (920°C), the quality of the nanowires, in terms of crystallinity, was better, but significant homogeneous Si particle nucleation and film deposition on the substrate occurred, along with substantial sidewall Si growth on the radial nanowire surface. Nanowires produced in the higher temperature were very large with diameters in the ~100 nm range.

By comparing the nanowire growth when using $SiCl_2H_2$ and more reactive SiH_4, Kamins et al. identified the importance of the reaction chemistry on the quality of the nanowires. At higher temperature, Si decomposition on the substrate is very fast (the diffusion-limited deposition regime), and it is nearly impossible to control Si deposition on the substrate and on the sidewalls of the nanowires. At lower temperature, the decomposition of the Si precursor on the substrate is relatively slow (the surface reaction limited regime), and Si deposition is nearly eliminated on the substrate. Ti catalyzes the decomposition of the Si precursor and helps to promote selective nanowire growth. The rate of nanowire growth at lower temperature, however, was relatively slow and it was difficult to achieve high-quality single crystal Si nanowires. Wang and Dai later also showed that the balance between substrate nucleation and deposition and metal-seeded nanowire formation relates to the precursor reactivity. They showed that H_2 could be added to suppress the GeH_4 decomposition rate in the Au nanocrystal-seeded Ge nanowire synthesis by CVD–VLS.[12] This ''additive'' prevented sidewall growth and deposition on the substrate to give much higher quality narrow diameter nanowires. They were also able to grow Ge nanowires at very low temperatures, ~275°C, which helped to limit unwanted substrate and sidewall deposition. This is also the strategy used in the metal nanocrystal-seeded growth of single-wall carbon nanotubes by CVD—a precursor is identified that is reactive enough to produce tubes, but unreactive on the bare substrate.

The work of both Kamins et al. and Dai et al. revealed the importance of the nanocrystal seed in

promoting crystallization and its potential role in the reaction kinetics. $TiSi_2$-nucleated Si nanowire growth was actually observed at temperatures more than 600°C below the lowest Ti–Si eutectic, indicating that metal-assisted nanowire growth can occur from solid nanometer-size seed particles by a vapor–solid–solid (VSS) mechanism.[11] Nanowire growth from solid seed particles is made possible by sufficiently rapid diffusion of the nutrient material (Si) either through the solid seed particle or along its surface. Mathur et al. have reported VSS growth of Ge nanowires from polished Fe substrates at 385°C—approximately 500°C below the lowest temperature eutectic in the Fe : Ge phase diagram.[13] For binary semiconductors such as GaAs, CVD nanowire growth can proceed at temperatures below the eutectic by a VSS pathway if there is preferential solubility of the Group III metal in the seed, as is the case for Ga and gold. For example, Samuelson and coworkers combined nanoscale spatially resolved elemental characterization with in situ heating experiments on Au-nucleated GaAs nanowires to show that the seed particle contained only Ga and Au and remained solid during nanowire growth, implying that nutrient delivery proceeds by both solid (for Ga) and surface (for As) diffusion.[14] It is possible, however, to have VLS nanowire growth at temperatures below the metal : semiconductor eutectic on the phase diagram as a result of melting point depression in nanometer-size particles. For example, Wang and Dai attributed their Au particle-assisted VLS-CVD growth of Ge nanowires at 275–80°C below the Au : Ge eutectic—to such a melting point reduction in nanoscale seed particles.[12]

A significant advance in metal-nucleated semiconductor nanowire synthesis was made by Morales and Lieber by combining VLS growth with nanometer-size aerosol metal seeds generated by laser ablation.[15] In the initial conception of this laser-catalyzed growth (LCG) process, a pulsed-laser ablated a $Si_{0.9}Fe_{0.1}$ target at 1200°C to generate nanometer-size $Si_{1-x}Fe_x$ liquid clusters (see Fig. 3A). These supersaturated clusters underwent phase separation to precipitate Si nanowires with diameters ranging from 6 to 20 nm and lengths up to 30 μm (Fig. 3B). Similar to previous reports of VLS grown Si nanowires, the LCG approach yielded Si nanowires grown in the ⟨1 1 1⟩ crystallographic direction. Expanding upon this approach, Duan and Lieber demonstrated LCG synthesis of essentially all main group elemental and binary semiconductor materials.[16] For example, Figs. 3C and

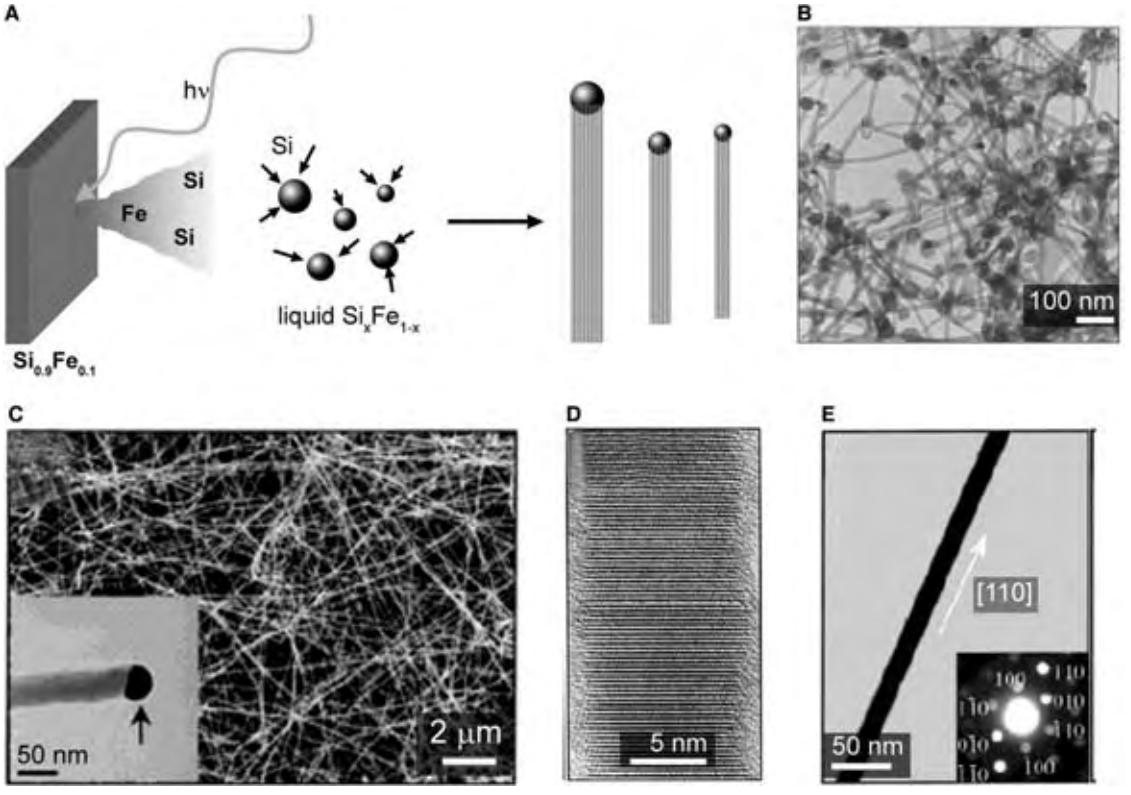


Fig. 3 (A) Schematic of the LCG model. (B) TEM image of Si nanowires produced via LCG. (From Ref. [15].) (C) SEM image of GaP nanowires, single seed particle composed mostly of Au is shown in the inset. (D) HRTEM image of $GaAs_{0.6}P_{0.4}$ nanowire with ⟨1 1 1⟩ growth axis. (E) CdSe nanowire with 18 nm diameter (wurtzite crystal structure is indicated in the inset). (From Ref.[16].) *(View this art in color at www.dekker.com.)*

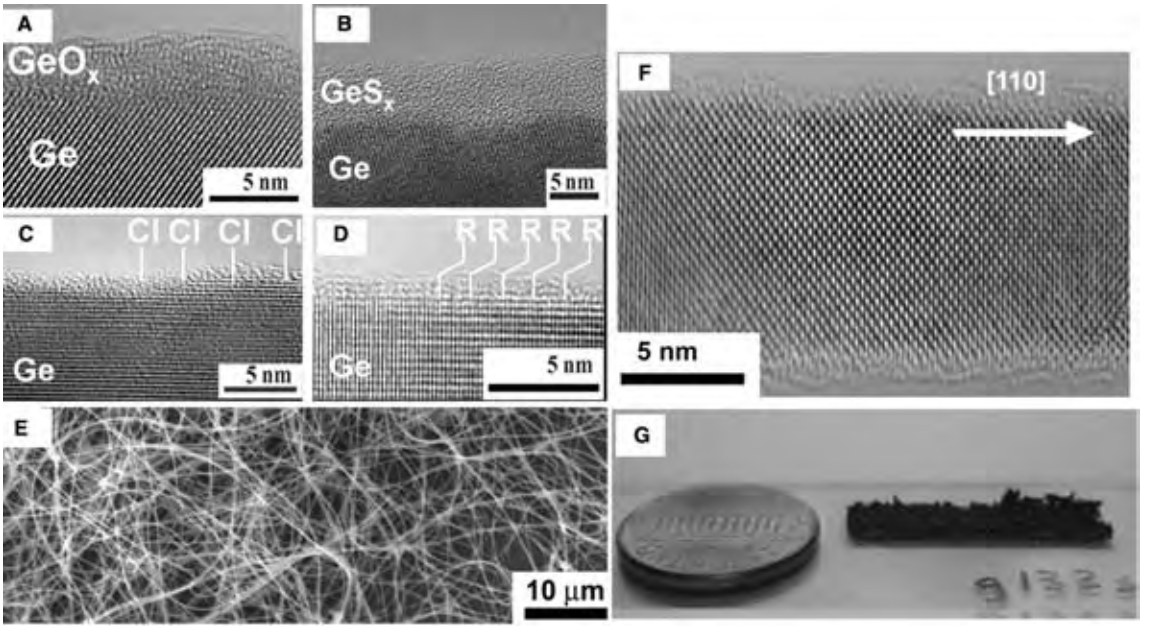


Fig. 4 (A–D) Several HRTEM images of Ge nanowire surfaces: (A) The native oxide surface layer. (B) GeS surface layer derived by in situ treatment with $(NH_4)_2S$. (C) Chloride terminated surface resulting from HCl etching. (D) An organic monolayer passivated surface by thermally initiated hydrogermylation. (E) An SEM image showing a large ensemble of long, straight Ge nanowires. (From Ref.[35].) (F) An HRTEM showing the $\{1\,1\,0\}$ growth direction. (G) A photograph of large quantities of Ge nanowires appearing as a black powder. *(View this art in color at www.dekker.com.)*

3D show SEM and HRTEM images of an ensemble of GaP nanowires and a $GaAs_{0.6}P_{0.4}$nanowire oriented along the $\langle 1\,1\,1 \rangle$ axis. In Fig. 3E, a CdSe nanowire with $\langle 1\,1\,0 \rangle$ growth direction as an example of a Group II–VI compound semiconductor is shown. These studies highlight the generality of metal-seeded semiconductor nanowire growth to virtually any semiconductor, provided the appropriate precursors, seed metal, and reaction temperatures can be identified.

Supercritical Fluid Solution Growth

Although the seed particle sizes in LCG and SLS growth are in the nanometer-size range, the particle size distribution cannot be controlled, as the in situ generation of seed particles leads to broad log–normal size distributions and affords little control of the nanowire diameter in the preparation. These limitations have been addressed in two general ways by research groups: 1) by depositing size-selected colloidal metal nanocrystals onto substrates to seed CVD–VLS nanowire and 2) by performing SLS-type nanowire growth in solution by feeding in prefabricated size-selected metal seed particles. One limitation of the SLS approach is the relatively narrow temperature range available for nanowire growth. Squalene boils at about 400°C, which is about the maximum allowable temperature for SLS growth. One way to increase the temperature range available for SLS growth is to use pressurized solvents, in particular organic solvents heated and pressurized above their critical point—supercritical fluids. In 2000, we demonstrated the use of size-selected organic monolayer stabilized Au nanocrystals to seed Si nanowire synthesis in hexane.[6] To achieve the necessary high-temperature growth conditions in solution, we pressurized the solvent well above its critical pressure, which enabled temperatures between 350°C and 650°C to be reached. In this temperature range, organosilanes, such as diphenylsilane can be used as a Si precursor, decomposing to Si. For example, diphenylsilane and Au nanocrystals have been used to grow high-quality crystalline Si nanowires in hexane at ~450°C with diameters as small as 4 nm in diameter. This SFLS approach was later extended to other semiconductors, including Ge,[17] GaAs,[18] and GaP,[19] and has been scaled up to a continuous flow reactor configuration capable of producing a gram of crystalline nanowires in a single reaction.[17]

In addition to providing a potentially technologically important synthetic route to semiconductor nanowires capable of meeting the high throughput and high-quality needs to meet the potential future technological demands for semiconductor nanowire applications, SFLS synthesis also provides a good experimental testbed for studying the fundamental interplay between kinetic and thermodynamic factors that underlie VLS growth. A wide range of semiconductor

precursor molecules can be studied in solution, as limitations of volatility are not an issue. For example, Korgel and coworkers conducted a systematic study of various organosilane precursors and found that phenylsilanes produce high-quality nanowires with high product yield, while alkylsilanes decompose much too slowly at 450°C to yield nanowires. These studies provide important information that might be applied in CVD processes, which have been particularly prone to unwanted sidewall deposition because of very fast precursor decomposition. Highly reactive trisilane, on the other hand, degrades too rapidly and overwhelms the Au-mediated nanowire crystallization with homogeneous nucleation and colloidal growth of amorphous Si particles.[20] The SEM image in Fig. 4E shows the high quality and large quantities (in Fig. 4G) of Ge nanowires produced by SFLS. The TEM image shows a Ge nanowire.

Also, SFLS has been applied to the synthesis of binary semiconductors, such as GaAs and GaP; however, the development of the synthesis has proven to be relatively challenging. Unlike in the gas-phase, there is very little we know about precursor decomposition rates in high-temperature supercritical fluids, which creates challenges because high-quality nanowire synthesis requires precursor decomposition kinetics that are balanced to ensure that materials with the appropriate stoichiometry are produced. The growth of Au nanocrystal-nucleated GaP nanowires via the SFLS approach using a stoichiometric ratio of tri-*tert*-butylgallium and *tris*-(trimethylsilyl) phosphine can result in excess Ga metal that accumulates at the tips of growing nanowires, which can under some conditions supplant Au as the metal catalyst.[19] In another conception of the SFLS process, Au nanocrystals were molecularly tethered to the Si substrate inside the reactor. The combination of this technique with the continuous flow SFLS reactor enabled the elucidation of the influence of critical synthesis parameters, including reactor pressure, temperature, residence time, precursor concentration, and precursor choice on the nanowire growth kinetics and morphology.[21]

Nanowire Growth Seeds

The use of size-selected Au nanocrystals has been employed extensively by many researchers to control the size of VLS-grown nanowires (Fig. 5A).[6,7,12,14,17,18,21,27] For CVD–VLS nanowire growth, colloidal particles are first size-selected in solution and then drop cast or spun cast onto the substrate prior to nanowire deposition. Nanowire diameter control is very important, as the optical, electronic, and mechanical properties of nanowire depend on the nanowire diameter. The histograms shown in Fig. 5B illustrate that the nanowire diameter

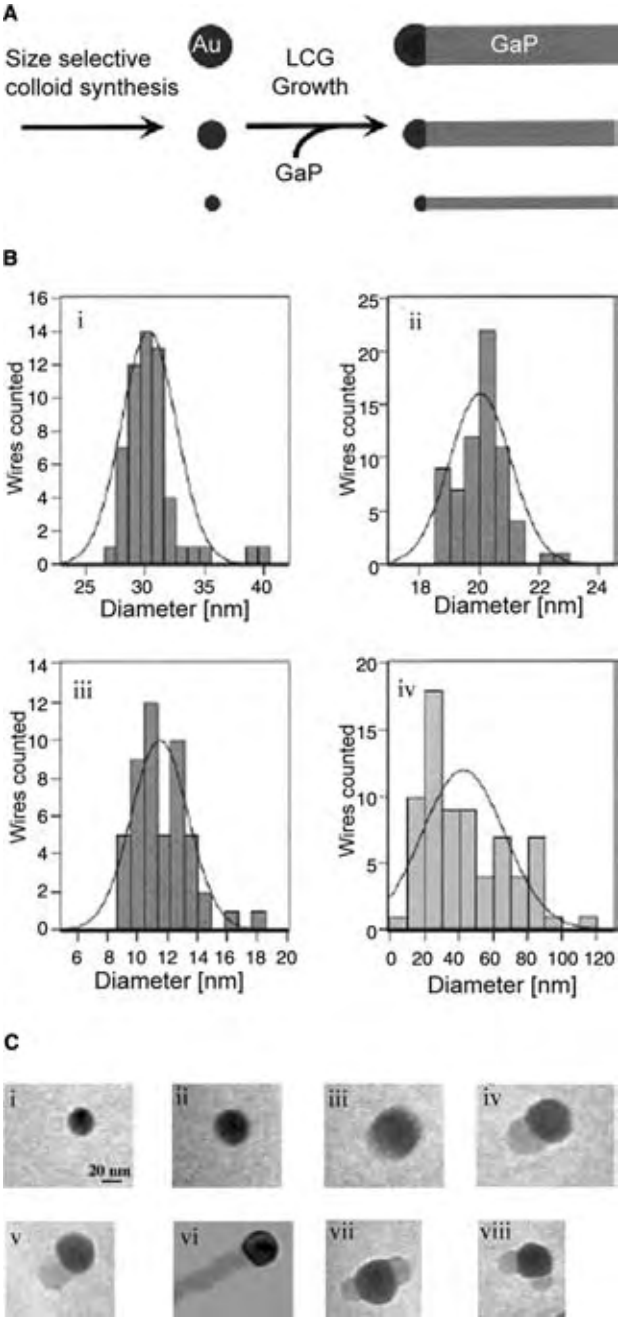


Fig. 5 (A) Schematic illustrating control of nanowire diameter distribution through use of Au nanocrystal seeds with different diameter distributions. (B) Diameter distribution histograms of GaP nanowires nucleated by: i) 28.2 nm, ii) 18.5 nm, iii) 8.4 nm diameter average sized Au seed nanocrystal, and iv) the broad diameter distribution obtained with LCG growth. (From Ref.[7].) (C) In situ TEM images of Au-nucleated Ge nanowire growth: i) solid Au nanocrystal at 500°C, ii) initial alloying at 800°C, iii) liquid Au/Ge alloy, iv) nucleation of Ge nanocrystal on the alloy surface, v) Ge nanocrystal elongates with further condensation, vi) eventually forms a wire, vii and viii) two nucleation events from a single alloy droplet. (From Ref.[23].)

corresponds reasonably well to the diameter distribution of the size-selected Au seed nanocrystals.

Although still under investigation, the nanowire growth direction also appears to be influenced by the seed particle diameter. Lieber and coworkers found that the growth direction of Au-seeded Si nanowires depended on the nanowire diameter.[22] Small Si nanowires with diameters ranging from 3 to 10 nm appeared to favor the ⟨1 1 0⟩ growth direction; whereas, larger nanowires in the diameter range of 10–20 nm favored the ⟨2 1 1⟩ growth direction, and even larger nanowires (20–30 nm) favored the characteristic ⟨1 1 1⟩ growth direction observed in Si whiskers. Lieber correlated the Si nanowire growth direction with the appearance of a flat {1 1 1} Si/Au interface, which gave rise to a V-shaped Au/Si interface in smaller diameter ⟨1 1 0⟩-oriented nanowires. Ge nanowires developed through SFLS exhibit predominantly ⟨1 1 0⟩ oriented growth, which appears to relate to crystallographic faceting of the nanowire nucleus and hexagonal sidewall faceting along the length of the nanowires as opposed to the Au/Ge interface.[28] From Au nanocrystals on a TEM grid, Wu and Yang made real-time TEM observations of VLS Ge nanowire from resistively vaporizing Ge (800–900°C).[23] As shown in Fig. 5C, these in situ imaging experiments captured the VLS growth process, including the liquefaction of the solid Au seed particle upon alloy formation and the emergence of the Ge crystal.

Au has been the most widely used metal to seed semiconductor nanowire growth because of the general availability of relatively size monodisperse nanocrystals, their chemical inertness, and the relatively low-temperature eutectic formed with a broad range of semiconductor materials. Unfortunately, Au is generally unacceptable in Si microelectronics, as it forms deep traps in Si that seriously degrade device performance. Other metals, such as Fe, Ni, Co, and Ti, are more compatible with Si electronics but exhibit significantly higher eutectic temperatures. Several researchers have explored the use of these alternate metals as replacements for Au seed nanocrystals.[11,13,24] For example, Korgel and coworkers recently exploited the catalytic properties of colloidal Ni nanocrystals to direct the growth of Si nanowires by SFLS.[24]

Complex Structure Nanowires

In a further extension of LCG growth, Lieber and coworkers[25] and Yang and coworkers[4] independently demonstrated the preparation of nanowires with structurally complex radial or axial heterostructures. Radial or core-shell heterostructure nanowires were formed by depositing layers on a core nanowire (Fig. 6A).[25] Using this approach, homoepitaxial growth of B-doped Si shells on intrinsic Si and heteroepitaxial

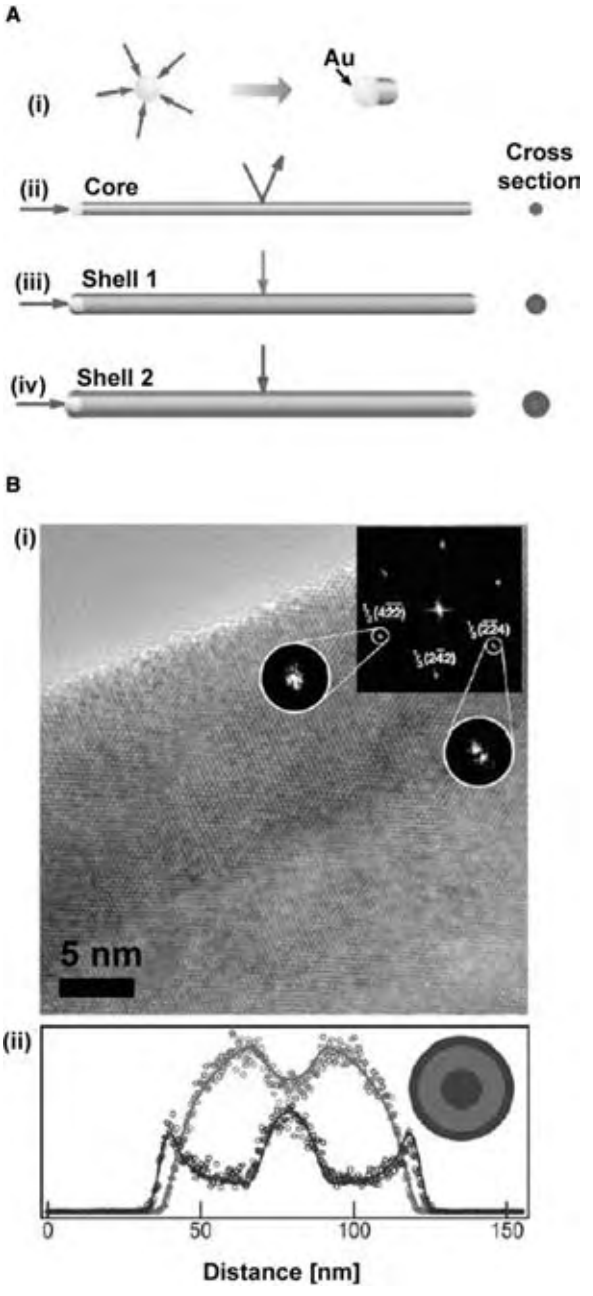


Fig. 6 (A) Schematic illustration of core-shell nanowire synthesis: i) gaseous reactants (red) catalytically decompose on the surface of a gold nanocrystal leading to nucleation and nanowire growth, ii) one-dimensional growth is maintained as reactant decomposition on the gold catalyst is strongly preferred, iii) synthetic conditions are altered to induce homogeneous reactant decomposition on the nanowire surface, and iv) multiple shells are grown by repeated modulation of reactants. (B) i) HRTEM image of Si–Ge–Si core shell nanowire showing the epitaxial interface, and ii) cross-sectional elemental mapping of a double shell structure with an intrinsic silicon core (diameter, 20 nm) intrinsic germanium inner shell (thickness 30 nm) and p-type silicon outer shell (4 nm); silicon is blue circles and germanium is red circles. (From Ref.[25].) *(View this art in color at www.dekker.com.)*

growth of Ge–Si, Si—Ge, and Ge–Si–Ge core shell structures was demonstrated. The TEM image and compositional profile shown in Fig. 6B illustrates the well-defined heteroepitaxial Ge–Si interface between the core and the shell structures. Heterostructure nanowires hold great potential for nanowire based electronic devices, as the multistep synthesis allows controlled modulation of the electrical properties of the core or the shell materials.[29]

Nanowire heterostructures with axially modulated compositions were achieved by switching the laser to ablate a different material target or through alternating the vapor phase molecular precursor being fed into a CVD reactor.[4] Both Group III–V and Si/SiGe SiGe heterostructure ''superlattice'' nanowires have been produced, as shown in Fig. 7A. By using well-defined switching intervals, Yang and coworkers measured the nanowire growth rate and verified its relation to the nanowire diameter as governed by the Gibbs–Thompson equation. The nanowires are especially intriguing for optoelectronic applications. For example, it has been shown that by modulating dopant fluxes it is possible to create an InP p–n heterojunction for use as the active component in polarized nano-LEDs.[30]

In another conception of nanowire heterostructures with axially modulated composition, Lieber and coworkers generated nanowires with lithographically defined segments of alternating metallic NiSi and semiconducting Si.[31] Ni was thermally evaporated on lithographically defined segments of a Si nanowire on a substrate. A thermal anneal diffused Ni into the Si nanowire to form the silicide NiSi, resulting in single nanowire NiSi/Si heterostructures with well-defined segments as shown in Figs. 7A and 7D. Heterostructures such as these modulate electrical properties along the nanowire length, enabling the fabrication of electrical devices with multiple source and drain electrodes defined on a single nanowire. The structural complexity of individual nanowires can also be increased, as branched and hyperbranched nanowires have been generated, as shown in Fig. 8A.[26] Secondary and tertiary branches were epitaxially grown from a central nanowire by decorating the nanowire surface with additional Au seed nanocrystals and repeating the VLS growth process. Multistep modification of VLS

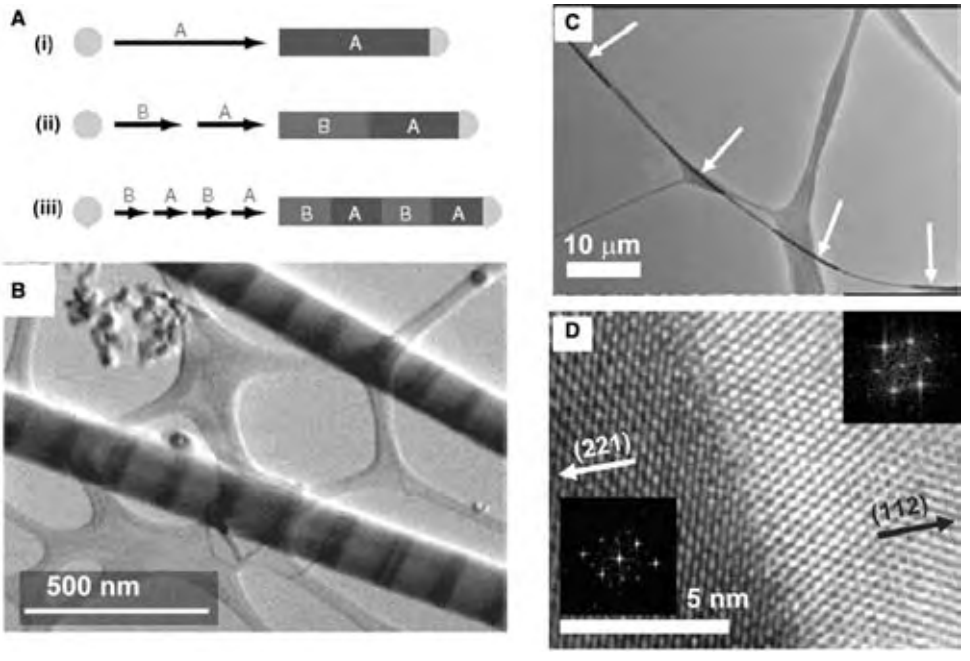


Fig. 7 (A) Synthesis of axial nanowire heterostructures: i) Au nanocrystal catalyst nucleates and directs one-dimensional semiconductor nanowire (blue) growth with the catalyst remaining at the terminus of the nanowire, ii) upon completion of the first growth step, a different material (red) can be grown from the end of the nanowire, and iii) repetition of steps i) and ii) leads to a compositional superlattice within a single nanowire. (From Ref.[30].) (B) STEM image of two Si/SiGe nanowires in bright field mode. (From Ref. [4].) (C) TEM image of a NiSi/Si heterostructured nanowire (the bright segments of the nanowire correspond to silicon and the dark segments, which are highlighted with arrows, correspond to NiSi). (D) High-resolution TEM image of the junction between NiSi and Si showing an atomically abrupt interface. (From Ref.[31].) In the insets, two-dimensional Fourier transforms of the image depicting the [1 1 0] and [1 1 1] zone axes of NiSi and Si, are shown, where the arrows highlight the growth fronts of the NiSi(2 2 1) and Si(1 1 2). *(View this art in color at www.dekker.com.)*

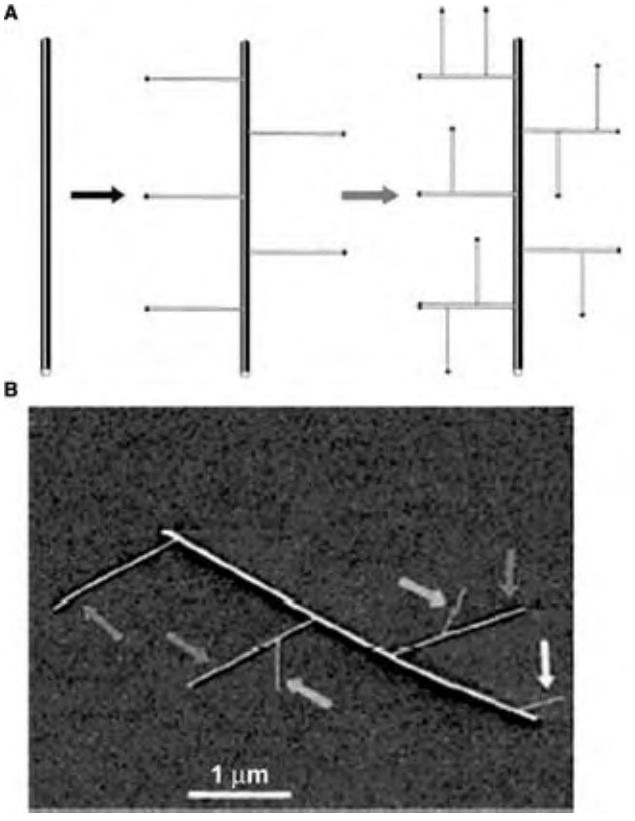


Fig. 8 (A) Schematic illustrating the multistep syntheses of branched and hyperbranched NW structures. Red, green, and blue arrows/colors signify the growth of the backbone, first-generation, and second-generation NWs, respectively. (B) SEM image of a hyperbranched SiNW structure. The first-generation and second-generation branches are indicated by orange and blue arrows, respectively. Yellow arrow indicates a 10 nm SiNW (from second generation) grown from the backbone. (From Ref.[26].) *(View this art in color at www.dekker.com.)*

processes can provide significant control over the structural and chemical complexity of the nanowires (Fig. 8B).

Integrating Nanowire Synthesis with Existing Microelectronics

The vast majority of proposed technological applications of semiconductor nanowires require multiscale integration with existing microelectronic platforms, and research efforts have been intensively directed at making nanowire syntheses compatible with conventional microelectronic processes. Some efforts have been directed at reducing nanowire growth temperatures.[12,27] Other efforts have focused on lithographic patterning for the directed deposit of nanowires.[12,32] Recently, epitaxial nanowire growth has been demonstrated, with

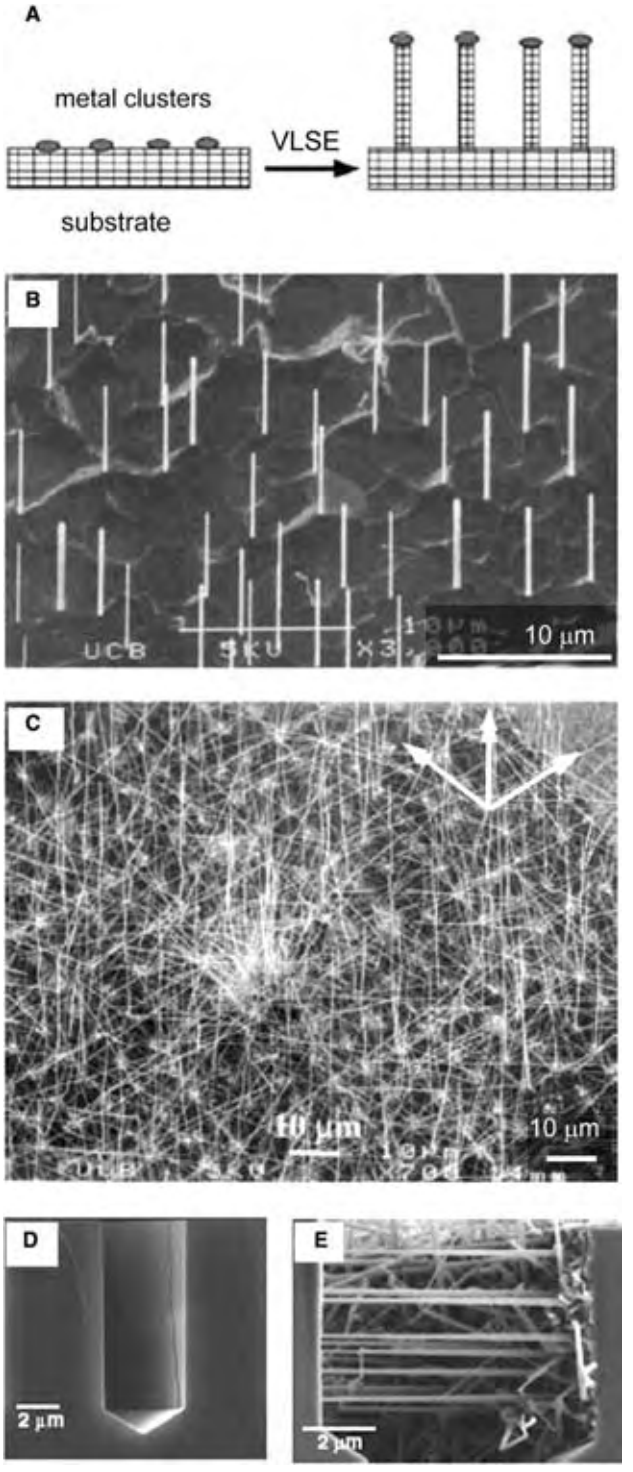


Fig. 9 (A) Schematic illustration of the VLSE process. (B) Tilted SEM image of vertical Si nanowire array grown on a (1 1 1) Si wafer. (C) Tilted SEM image of Si nanowire array grown on Si(1 0 0). (From Ref.[33].) Three of the four equivalent ⟨1 1 1⟩ directions are indicated by the white arrows. (D) Cross-sectional SEM images of a 4 μm-wide, anisotropically etched trench in a Si(1 1 0) wafer. (E) Au-catalyzed, lateral epitaxial nanowire growth across an 8 μm-wide trench, connecting to opposing sidewall. (From Ref.[34].)

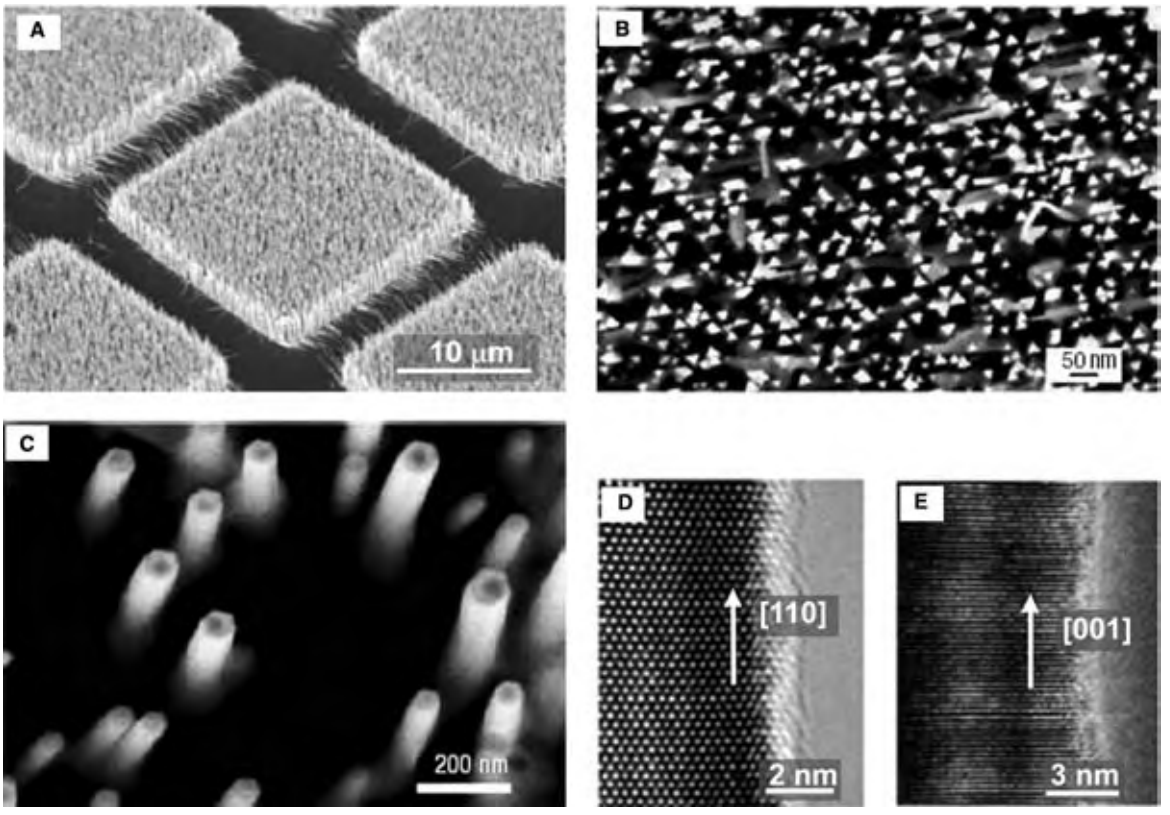


Fig. 10 (A) SEM image of GaN nanowire arrays grown on patterned (1 0 0) γ-$LiAlO_2$ substrate. (B) SEM image of [1 1 0] oriented GaN nanowires with triangular cross-sections grown from a (1 0 0) γ-$LiAlO_2$ substrate. (C) SEM image of [0 0 1] oriented GaN nanowires with hexagonal cross-section grown from a (1 1 1) MgO substrate. (D) HRTEM image of the lattice structure of a [1 1 0] oriented wire as seen in (B). (E) HRTEM image of the lattice structure of a [0 0 1] oriented wire as seen in (C). (From Ref.[32].) *(View this art in color at www.dekker.com.)*

nanowire crystallographic orientation matching the substrate orientation.[32,33] Si nanowires grown via vapor–liquid–solid-epitaxy (VLSE) preferred $\langle 1\,1\,1 \rangle$ growth axes, growing vertically on Si(1 1 1) substrates and along four equivalent $\langle 1\,1\,1 \rangle$ directions oriented 54.7° relative to the surface normal when grown on a Si(1 0 0) substrate[33] (Figs. 9A–C). In a related work, Williams and coworkers later demonstrated the epitaxial VLS growth of high density Si nanowire arrays between two vertical sidewall surfaces, as seen in Figs. 9D and 9E[34].

Nanowire Surface Chemistry

The enormous surface to volume ratio of the nanowires profoundly impacts the nanowire properties, making it imperative to control the surface chemistry. Several chemical routes have been explored for passivating or functionalizing nanowire surfaces.[35,36] Hanrath and Korgel reported a comprehensive investigation of Ge nanowire surface chemistry modification including oxidation, sulfidation, hydride and chloride termination, and organic monolayer passivation (Figs. 4A–D).[35] By combining SFLS synthesis with thermally initiated hydrogermylation reactions, organic monolayers were covalently attached to the nanowire surfaces (Fig. 4D). Monolayer-passivated nanowires exhibit improved dispersibility in solvents—an important requirement for many nanowire device fabrication processes—and enhanced chemical stability.[37] Electron transport was also found to be particularly sensitive to the surface chemistry,[38,39] a characteristic that several researchers have exploited to create high sensitivity environmental sensors. For example, nanowires with antibody-decorated surfaces exhibited conductivity changes that were sensitive to the presence of influenza or adenovirus, enabling the detection of single binding/unbinding events.[40]

Controlling the crystallographic growth direction is a highly desirable synthetic goal for a broad range of technological applications because of the anisotropy of fundamental properties, particularly the electron mobility and bandgap. In addition to size, shape, and composition, control over the nanowire growth direction offers another parameter that can be tuned to adjust the nanomaterial properties. In 2004, Yang

and coworkers reached a new milestone in semiconductor nanowire synthesis when they combined the VLSE process with careful selection of a lattice matching substrate to achieve control over the crystallographic orientation of GaN nanowire arrays.[32] Their experiments showed that wurtzite GaN nanowires epitaxially grown on (1 0 0) γ-$LiAlO_2$ substrates exhibited triangular cross-sections and a [1 1 0] crystallographic growth axis (Figs. 10B and 10D). Similar epitaxial growth on (1 1 1) MgO substrates, which exhibits good epitaxial match to GaN(0 0 1), resulted in [0 0 1] oriented GaN nanowires with hexagonal cross-sections (Figs. 10C and 10E).

CONCLUSIONS

The progress in the field of semiconductor nanowires over the past decade has been rapid. The field has evolved to a point where relatively complex, both structurally and chemically, functional nanomaterials with control over the size, shape, composition, surface chemistry, and crystallographic orientation can be produced. These synthetic advances have enabled nanowires to be integrated into a broad range of prototype devices including logic gates, data storage devices, optoelectronic devices, nanowire lasers, and chemical and biological sensors.

REFERENCES

1. Wagner, R.S. *Whisker Technology*; Wiley: New York, 1970.
2. Carter, J.L. VLS (vapor–liquid–solid): newly discovered growth mechanism on the lunar surface?. Science **1973**, *181*, 841–842.
3. Givargizov, E.I. Fundamental aspects of VLS growth. J. Crystal Growth **1975**, *31*, 20–30.
4. Wu, Y.; Fan, R.; Yang, P. Block-by-block growth of single-crystalline Si/Si–Ge superlattice nanowires. Nano Lett. **2002**, *2* (2), 83–86.
5. Westwater, J.; Gosain, D.P.; Usui, S. Si nanowires grown via the vapor–liquid–solid reaction. Phys. Status Solidi A: Appl. Res. **1998**, *165* (1), 37–42.
6. Holmes, J.D.; Johnston, K.P.; Doty, R.C.; Korgel, B.A. Control of thickness and orientation of solution-grown silicon nanowires. Science **2000**, *287* (5457), 1471–1473.
7. Gudiksen, M.S.; Wang, J.; Lieber, C.M. Synthetic control of the diameter and length of single crystal semiconductor nanowires. J. Phys. Chem. B **2001**, *105* (19), 4062–4064.
8. Heath, J.R.; LeGoues, F.K. A liquid solution synthesis of single crystal germanium quantum wires. Chem. Phys. Lett. **1993**, *208*, 263–266.
9. Trentler, T.J.; Hickman, K.M.; Goel, S.C.; Viano, A.M.; Gibbons, P.C.; Buhro, W.E. Solution–liquid–solid growth of crystalline III–V semiconductors: an analogy to vapor–liquid–solid growth. Science **1995**, *270*, 1791–1796.
10. Westwater, J.; Gosain, D.P.; Tomiya, S.; Usui, S.; Ruda, H. Growth of silicon nanowires via gold/silane vapor–liquid–solid reaction. J. Vac. Sci. Technol. B **1997**, *15* (3), 554.
11. Kamins, T.I.; Williams, R.S.; Basile, D.P.; Harris, J.S. Ti-catalyzed Si nanowires by chemical vapor deposition: microscopy and growth mechanisms. J. Appl. Phys. **2001**, *89*, 1008–1016.
12. Wang, D.; Dai, H. Low-temperature synthesis of single-crystal germanium nanowires by chemical vapor deposition. Angew. Chem., Int. Ed. **2002**, *41* (24), 4783–4786.
13. Mathur, S.; Shen, H.; Sivakov, V.; Werner, U. Germanium nanowires and core-shell nanostructures by chemical vapor deposition of $[Ge(C_5H_5)_2]$. Chem. Mater. **2004**, *16* (12), 2449–2456.
14. Persson, A.I.; Larsson, M.W.; Stenstroem, S.; Ohlsson, B.J.; Samuelson, L.; Wallenberg, L.R. Solid-phase diffusion mechanism for GaAs nanowire growth. Nature Mater. **2004**, *3*, 677–681.
15. Morales, A.M.; Lieber, C.M. A laser ablation method for the synthesis of crystalline semiconductor nanowires. Science **1998**, *279*, 208–211.
16. Duan, X.; Lieber, C.M. General synthesis of compound semiconductor nanowires. Adv. Mater. **2000**, *12* (4), 298–302.
17. Hanrath, T.; Lu, X.; Johnston, K.P.; Korgel, B.A. Growth of single crystal nanowires. U.S.A. Provisional Appl. 60/485,244, 2004.
18. Davidson F.M., III; Schricker, A.D.; Wiacek, R.J.; Korgel, B.A. Supercritical fluid–liquid–solid synthesis of gallium arsenide nanowires seeded by alkanethiol-stabilized gold nanocrystals. Adv. Mater. **2004**, *16* (7), 646–649.
19. Davidson F.M., III; Wiacek, R.; Korgel, B.A. Supercritical fluid–liquid–solid synthesis of gallium phosphide nanowires. Chem. Mater. **2005**, *17*, 230–233.
20. Lee, D.C.; Hanrath, T.; Korgel, B.A. Role of precursor decomposition kinetics in silicon nanowire synthesis in organic solvents. Angew. Chem. Int. Ed. **2005**, *44*, 3573–3577.
21. Lu, X.; Hanrath, T.; Johnston, K.P.; Korgel, B.A. Growth of single-crystal silicon nanowires in supercritical solution from tethered gold particles on a silicon substrate. Nano Lett. **2003**, *3* (1), 93–99.
22. Wu, Y.; Cui, Y.; Huynh, L.; Barrelet, C.J.; Bell, D.C.; Lieber, C.M. Controlled growth and structures of molecular-scale silicon nanowires. Nano Lett. **2004**, *4* (3), 433–436.

23. Wu, Y.; Yang, P. Direct observation of vapor–liquid–solid nanowire growth. J. Am. Chem. Soc. **2001**, *123* (13), 3165–3166.
24. Tuan, H.Y.; Lee, D.C.; Hanrath, T.; Korgel, B.A. Catalytic solid-phase seeding of silicon nanowires by nickel nanocrystals in organic solvents. Nano Lett. **2005**, *5*, 681–684.
25. Lauhon, L.J.; Gudiksen, M.S.; Wang, D.; Lieber, C.M. Epitaxial core-shell and core-multishell nanowire heterostructures. Nature **2002**, *420* (6911), 57–61.
26. Wang, D.; Qian, F.; Yang, C.; Zhong, Z.; Lieber, C.M. Rational growth of branched and hyperbranched nanowire structures. Nano Lett. **2004**, *4* (5), 871–874.
27. Greytak, A.B.; Lauhon, L.J.; Gudiksen, M.S.; Lieber, C.M. Growth and transport properties of complementary germanium nanowire field-effect transistors. Appl. Phys. Lett. **2004**, *84* (21), 4176–4179.
28. Hanrath, T.; Korgel, B.A. Crystallography and surface faceting of germanium nanowires. Small **2005**, *1*, 717–721.
29. Qian, F.; Li, Y.; Gradecak, S.; Wang, D.; Barrelet, C.J.; Lieber, C.M. Gallium nitride-based nanowire radial heterostructures for nanophotonics. Nano Lett. **2004**, *4*, 1975–1979.
30. Gudiksen, M.S.; Lauhon, U.J.; Wang, J.; Smith, D.C.; Lieber, C.M. Growth of nanowire superlattice structures for nanoscale photonics and electronics. Nature **2002**, *415* (6872), 617–620.
31. Wu, Y.; Xiang, J.; Yang, C.; Lu, W.; Lieber, C.M. Single-crystal metallic nanowires and metal/semiconductor nanowire heterostructures. Nature **2004**, *430* (6995), 61–65.
32. Kuykendall, T.; Pauzauskie, P.J.; Zhang, Y.; Goldberger, J.; Sirbuly, D.; Denlinger, J.; Yang, P. Crystallographic alignment of high-density gallium nitride nanowire arrays. Nature Mater. **2004**, *3* (8), 524–528.
33. Wu, Y.; Yan, H.; Huang, H.; Messer, B.; Song, J.H.; Yang, P. Inorganic semiconductor nanowires: rational growth, assembly, and novel properties. Chem. Eur. J. **2002**, *8* (6), 1261.
34. Islam, M.S.; Sharma, S.; Kamins, T.I.; Williams, R.S. Ultrahigh-density silicon nanobridges formed between two vertical silicon surfaces. Nanotechnology **2004**, *15* (5), L5–L8.
35. Hanrath, T.; Korgel, B.A. Chemical surface passivation of Ge nanowires. J. Am. Chem. Soc. **2004**, *126*, 15466–15472.
36. Cui, Y.; Wei, Q.; Park, H.; Lieber, C.M. Nanowires nanosensors for highly sensitive and selective detection of biological and chemical species. Science **2001**, *293*, 1289.
37. Hanrath, T.; Korgel, B.A. Germanium nanowire transistors: a comparison of electrical contacts patterned by electron beam lithography and beam-assisted chemical vapor deposition. J. Nanoeng. Nanosyst. **2005**, *218*, 25–34.
38. Hanrath, T.; Korgel, B.A. Influence of surface states on electron transport through intrinsic Ge nanowires. J. Phys. Chem. B **2005**, *109*, 5518–5524.
39. Wang, D.C.; Wang, Q.; Cao, J.; Farmer, D.B.; Gordon, R.G.; Dai, H. Surface chemistry and electrical properties of germanium nanowires. J. Am. Chem. Soc. **2004**, *126*, 11602–11611.
40. Patolsky, F.; Zheng, G.; Hayden, O.; Lakadamyali, M.; Zhuang, Z.; Lieber, C.M. Electrical detection of single viruses. Proc. Natl. Acad. Sci. **2004**, *101* (39), 14017–14022.

Water Gas Shift Reaction

Wolfgang Ruettinger
Oleg Ilinich
Engelhard Corporation, Iselin, New Jersey, U.S.A.

INTRODUCTION

Water gas shift (WGS) refers to the reaction of carbon monoxide with steam to yield carbon dioxide and hydrogen:

$$CO + H_2O \leftrightarrow CO_2 + H_2$$
$$\Delta H = -41.1 \text{ kJ/mol}$$

It is a mildly exothermic reaction, with an adiabatic temperature rise of 8–10°C per percent (wet gas) CO converted to CO_2 and H_2. It was first used industrially at the beginning of the 20th century in hydrogen production via coal gasification, as a part of ammonia synthesis by the Haber-Bosch process.[1,2] Water gas ($CO + H_2$) was produced by blowing steam over hot coal via an endothermic reaction $C + H_2O \rightarrow CO + H_2$. To maintain high temperature of the coal (above approximately 1000°C), steam was periodically cut off and air was blown through the bed of coal, producing carbon monoxide via the exothermic reaction $2C + O_2 \rightarrow 2CO$. Therefore, the gas mixture exiting the water gas converter contained CO and hydrogen plus small amounts of nitrogen and CO_2. Carbon monoxide was catalytically shifted to produce more hydrogen. Residual CO was removed by absorption in copper liquor, CO_2 was removed by caustic scrubbing, and the resulting hydrogen/nitrogen mixture was used in ammonia synthesis. Most hydrogen plants today consist of hydrodesulfurization (HDS) units and zinc oxide beds for feed purification, followed by steam reformers (SRs) and WGS reactors.

A catalyst consisting of iron oxide stabilized by chromium oxide was used in the shift reaction and was patented by Bosch and Wild in 1914.[3] This catalyst is largely unchanged and is still in use in industry as the high temperature shift (HTS) catalyst. Since the 1960s, copper-based low temperature shift (LTS) catalysts have been introduced in order to reduce the CO content of reformate to <0.5%. They are primarily important in the ammonia industry, where a high hydrogen yield is economically important.[4] Additionally it enables CO clean-up by the methanation reaction to protect the Fe-based ammonia synthesis catalyst from CO poisoning. If properly used, CO concentrations as low as 0.1% are possible in the resulting reformate. The WGS reaction is primarily used today in hydrogen and synthesis gas production for production of ammonia and methanol. Besides the large industrial production of hydrogen and synthesis gas, small-scale hydrogen production for fuel cells, hydrogen filling stations, as well as on-site hydrogen generation, is becoming a field of growing interest.[5] New catalysts have been developed for these applications, and we will describe some of the recent development later in this chapter.

These catalysts have to fulfil new safety requirements as they may be operated in the consumer's home, and also will experience a very different duty cycle than industrial catalysts. Another field that has attracted interest is sulfur tolerant shift catalysts. Catalysts have been developed for high sulfur concentrations in the feed gas since the 1960s,[6,7] but have not found large-scale application up to the present. The fact is that most ammonia, methanol, and hydrogen plants are based on natural gas or naphtha feedstocks, which have relatively low content of sulfur-containing compounds, and therefore do not require these more expensive catalysts. HDS is typically used to reduce sulfur levels in the feed gas to ppm to ppb levels to protect the steam reforming and downstream catalysts.

THERMODYNAMICS

The water gas shift reaction is a mildly exothermic reversible reaction, and CO conversion is limited by the thermodynamic equilibrium (see Fig. 1).

$$K_p = \exp\{(4577.8K/T) - 4.22\}$$
$$K_p = \{p(CO_2) * p(H_2)\}/\{p(CO) * p(H_2O)\}$$

The equilibrium constant K_p (see equations above) decreases with increasing temperature. Thus, low temperatures favor product formation in the WGS reaction. The equilibrium can also be shifted towards the right by increasing the steam concentration or removing a product from the reaction mixture (i.e., hydrogen through a hydrogen permeable membrane).

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008007

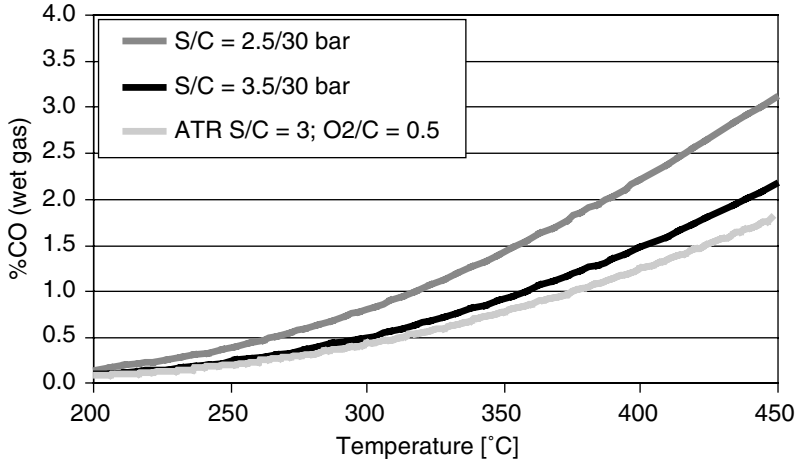


Fig. 1 WGS equilibrium curves for different reformates.

Pressure does not influence the equilibrium constant, since the same number of moles and therefore the same volume of gases exist on both side of the reaction equation.

Industrially, the shift reaction is run in adiabatic reactors.[7] Thus, there exists the dilemma of slow catalyst kinetics where the equilibrium is favorable (low temperature) and fast catalyst kinetics where the equilibrium is unfavorable (high temperature). This problem is solved by operating the reaction in one to three sequential reactors with inter-stage cooling and subsequently lower inlet temperatures.

Let us examine how the gas composition influences the WGS equilibrium CO concentration. If the feed gas comes from a steam reformer (SR) where methane and water react producing CO and hydrogen,

$$CH_4 + H_2O \leftrightarrow CO + 3H_2$$

the SR equilibrium CO concentrations are dependent on the steam to carbon ratio and pressure at which the reformer is run as well as the hydrocarbon feedstock. Some typical SR gas compositions are shown in Table 1 and the equilibrium gas composition for the WGS reaction is shown in Fig. 1.

The other methods of gas generation involve catalytic partial oxidation (CPO)

$$CH_4 + \tfrac{1}{2}O_2 \rightarrow CO + 2H_2$$

and autothermal reforming (a combination of CPO and SR). The CO concentration achievable using autothermal reforming (ATR) reformate as the feed gas are lower than with SR reformate, but the gas is diluted with ~25% nitrogen. The CO concentration in CPO feed gas can be as high as 50% and care has to be taken not to exceed the upper temperature limit of the catalyst due to the exotherm of the WGS reaction.

CATALYSTS, KINETICS, MECHANISMS, AND OPERATING WINDOWS

This section gives a brief overview of HTS and low temperature shift (LTS) catalyst formulations, their use in the water gas shift processes, as well as the reaction mechanisms and kinetics (Fig. 2).

Table 1 Typical reformate gas compositions of steam reformate and ATR reformate

Component	S/C = 2.5, 1 bar, 700°C exit temperature	S/C = 2.5, 30 bar, 850°C exit temperature	S/C = 3.5, 30 bar, 850°C exit temperature	S/C = 2.5, 30 bar, 850°C exit temperature
Feed gas	CH_4	CH_4	CH_4	Naphtha
CO	12.1	9	7.35	10.9
CO_2	5.9	5.3	5.7	7.5
H_2	60	48	44.85	43.3
CH_4	0.25	6.2	3.35	5.2
N_2	—	—	—	—
H_2O	21.75	31.5	38.75	33.1

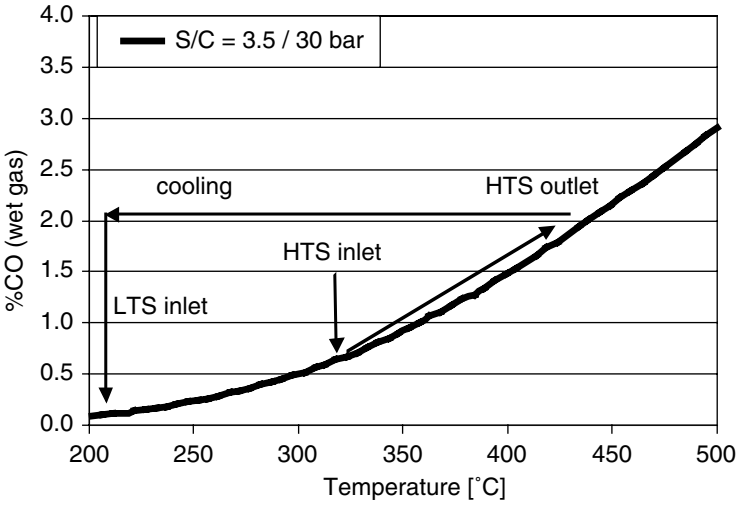


Fig. 2 Typical operating regime of HTS and LTS catalysts.

High Temperature Shift Catalysts

The first stage of the WGS process is HTS which operates at relatively high inlet temperatures, typically in the range of approximately 320–360°C.[6,7] These temperatures are dictated by the activities and selectivities of iron-chrome (or *ferrochrome*) catalysts that are being used in the HTS stage. In 90 years that have passed since its first use, the catalyst compositions and manufacturing procedures have gone through numerous changes to adopt the catalysts to new feedstocks (natural gas has now largely replaced coal) and process requirements (e.g., faster startup; high pressure), and to improve their activity.

HTS Catalyst Formulation

The HTS catalysts are typically produced by precipitation of aqueous solution of iron sulfate and chromium sulfate with sodium carbonate or sodium hydroxide, followed by careful washing to remove essentially all of the residual sulfate. Washing is an important step in the catalyst manufacture, since the residual sulfate converts into hydrogen sulfide when the catalyst is reduced during the process startup, and hydrogen sulfide is a poison to the LTS catalyst located downstream of the HTS bed.[8]

Addition of chromium is essential for stabilizing the catalyst activity over long periods of time on stream, which typically averages several years of continuous operation.[7,9] Chromium was reported to retard sintering of iron oxides in the catalyst under HTS reaction conditions, thus minimizing the loss of catalytically active surface.

Subsequent steps of the catalyst manufacture involve drying, calcining, and tabletting. Typically, tablets are 5–10 mm in diameter. Diffusion limitations to the reaction rate occur in the larger pellets increasingly above 350°C. However, larger pellets have superior strength. Therefore, the choice of the pellet size can be very important. After calcinations, the catalyst consists of approximately 90–95% Fe_2O_3 and 5–10% chromium oxides, which is predominantly Cr_2O_3 with some CrO_3 also present. The dominant phase in the freshly calcined catalyst is a solid solution of Cr^{3+} in α-Fe_2O_3 (hematite).[10] Cr^{+6} in the catalyst is undesirable because it represents a serious health hazard as a carcinogen. Besides, CrO_3 is water soluble and can leach out of the catalyst, e.g., if the WGS process steam condenses over the oxidized catalyst. In addition, reduction of CrO_3 to Cr_2O_3 is highly exothermic and can lead to overheating of the catalyst and activity loss due to sintering of the active phase. Therefore, the calcination step has to be carefully controlled to minimize the amount of Cr^{+6} in the catalyst.

The iron-chrome catalyst was improved by addition of Cu to increase activity and selectivity (by suppressing CH_4 generation).[8,11,12] The possible elimination of Cr from Fe-based WGS catalysts has also been addressed.[13] However, Fe–Cr catalysts have been used for a long time and a change of catalyst by industry will inevitably be slow.

Reduction of Iron-Chrome Catalyst in HTS Process

To become active, this catalyst needs to be reduced. The reduction converts α-Fe_2O_3 (hematite) into Fe_3O_4 (magnetite), which is the catalytically active phase, and CrO_3 reduces to Cr_2O_3. The reduction is typically performed during the HTS reactor startup, and it should be carefully controlled because of a significant heat release resulting from the exothermic

nature of the reactions involved, which may damage the catalyst:

$$3Fe_2O_3 + H_2 \rightarrow 2Fe_3O_4 + H_2O \quad \Delta H = -9.6\,kJ/mol$$

$$3Fe_2O_3 + CO \rightarrow 2Fe_3O_4 + CO_2 \quad \Delta H = -50.7\,kJ/mol$$

$$2CrO_3 + 3H_2 \rightarrow Cr_2O_3 + 3H_2O \quad \Delta H = -697.9\,kJ/mol$$

$$2CrO_3 + 3CO \rightarrow Cr_2O_3 + 3CO_2 \quad \Delta H = -821.3\,kJ/mol$$

Besides, over-reduction with formation of FeO and metallic iron should be avoided since these phases are not catalytically active in WGS and they catalyze undesired side reactions (i.e., methanation and CO disproportionation). Therefore, the reduction is always conducted in the presence of steam and with relatively low concentrations of the reductants in the process gas. Steam acts as a mild oxidant, thereby stabilizing the magnetite phase.

Fe–Cr HTS Catalyst Kinetics

The mechanism and kinetics of the WGS reaction over Fe–Cr catalysts have been the subject of numerous publications.[7,10,14–16] Despite intense investigations, still there is no full agreement as to the reaction mechanism. The two competing approaches are a redox (regenerative) mechanism first proposed by Kulkova and Temkin[17] as early as 1949 which presumes reduction of an oxide center (O) by a CO molecule yielding CO_2 and a vacant surface center (*), followed by reoxidation of the vacant center by water that produces hydrogen and regenerates the oxide center for the catalytic cycle.

$$CO + (O) \rightarrow CO_2 + (^*)$$
$$H_2O + (^*) \rightarrow H_2 + (O)$$

The alternative approach implies that adsorbed CO reacts with a surface hydroxyl group to yield a surface bound formate species.[18] The formate then decomposes in the presence of steam to CO_2 and H_2.

$$H_2O \rightarrow OH(ads) + H(ads)$$
$$CO(ads) + OH(ads) \rightarrow COOH(ads)$$
$$COOH(ads) \rightarrow CO_2 + H(ads)$$
$$2H(ads) \rightarrow H_2$$

The published evidence is overwhelmingly supportive of the redox mechanism,[6,7,19–22] although some authors still favor the adsorptive (associative) mechanism.[10,18] The existing discrepancy in conclusions probably results from differences in the experimental conditions employed in different studies such as temperatures, concentrations, and nature of catalysts.

The earliest kinetic equation published in the open literature was the one suggested by Kulkova and Temkin.[17] It was based on the hypothesis of the redox mechanism, with reduction of the oxide center (O) by a CO molecule being the rate-controlling step:

$$r = k_+P_{CO}(P_{H_2O}/P_{H_2})^{1/2} - k\_P_{CO_2}(P_{H_2}/P_{H_2O})^{1/2}$$

where r is the reaction rate, k_+ and k_- are rate constants for forward and reverse reactions, respectively, and P_i is the partial pressure of a given reactant.

A more accurate kinetic equation covering a broader range of process variables was later proposed by Shchibrya et al.[23] This equation was based on the same hypothesis of a redox mechanism but without rate-controlling steps.

In later years, a number of equations were suggested by different authors. A comprehensive analysis of kinetic equations published before 1980 can be found in Ref.[10] An alternative to the fundamental equations based on mechanistic concepts, e.g., proposed in Refs.[17,24], is an empirical equation of the following power-law type:

$$r = k(P_{CO})^a(P_{H_2O})^b(P_{CO_2})^c(P_{H_2})^d(1 - \beta),$$

where k is the rate constant, P_i is the partial pressure of a given reactant, a, b, c, d are the exponents, $\beta = P_{CO2}P_{H2}/K P_{CO} P_{H2O}$, and K is the WGS equilibrium constant.

The equations of the latter type, with the exponents determined under a sufficiently broad range of the process parameters, are good practical tools for describing the WGS process.

Low Temperature Shift Catalysts

The second stage in WGS processes is LTS, which was introduced in the industry in the 1960s and is now widely used in hydrogen plants. Rapid success of the LTS process was due to the use of copper-based catalysts that are more active at lower temperatures than iron-chrome HTS catalysts and therefore enable low equilibrium CO concentrations in the gas exiting the reactor, which increases the yield of the hydrogen production process.

Although copper-based catalysts have long been known to have good WGS activities, sensitivity of those catalysts to poisons that were present in the coal-derived gas precluded them from being employed industrially. It is only due to a massive change from coal gasification to hydrocarbon steam reforming which produces much purer synthesis gas, that copper-based catalysts entered the scene of WGS processes. Since

then, these catalysts have gone through significant improvements that have made them much more active and robust. However, the thermal stability of the LTS catalysts is still inferior to the HTS catalyst. This is why an HTS catalyst is needed. If the steam reformate containing approximately 10% CO enters an LTS catalyst, the WGS reaction exotherm would bring the exit temperature to unacceptable levels.

LTS Catalyst Formulations

The first industrially used LTS catalyst consisted of copper and zinc oxides,[7] and these components are still present in the catalysts that operate in today's plants. The latest formulations of LTS catalyst are significantly more active than their predecessor and have a much longer operating life. This progress was achieved by introducing additional components in the catalyst to stabilize small crystallites of copper (the active component) from sintering under the process conditions. In earlier versions of LTS catalysts an additional component was chromium,[23,25] later replaced by aluminum since Cu–Zn–Al oxide system has shown better stability. Today copper-zinc-alumina-based catalysts are widely used in the LTS stage of hydrogen plants. They are typically made from aqueous solutions of corresponding nitrates via continuous precipitation by sodium hydroxide, sodium carbonate, or ammonium bicarbonate under controlled pH.[6] The other operations in the catalyst manufacture include washing of the precipitate to remove sodium, followed by drying, calcination, and tabletting. Optimum content of copper oxide in the final Cu-Zn–Al catalyst is about one third of the total weight, while the amounts of zinc and aluminum oxides in different commercial catalysts vary. Thus, the catalysts produced by ICI were reported to have 53–34% ZnO and 15-33% Al_2O_3.[7] The pellet size varies by 3–8 mm in different commercial catalysts. Since the reaction is pore diffusion limited under most reaction conditions, the size and shape of the particles is optimized to give maximum surface area while still having the required strength and pressure drop.

Reduction of Copper-Based Catalysts in LTS process

Copper-based LTS catalysts are supplied to customers in their oxide form. Analogous to the iron-chrome HTS catalysts, LTS catalysts have to be activated by reduction prior to their use:

$$CuO + H_2 \rightarrow Cu + H_2O \qquad \Delta H = -80.8\ \text{kJ/mol}$$

Reduction of the catalysts can be a lengthy procedure for which experienced plant operators are needed.[7,8] For conventional LTS catalysts, adiabatic temperature rise resulting from the above reaction can be as high as 500–600°C. Such extreme temperatures are unacceptable for the LTS section primarily because of the catalyst sintering which occurs above approximately 260°C.[7] Therefore the reduction is always carried out using low concentrations of hydrogen and/or CO in inert gas such as nitrogen or natural gas, while carefully monitoring the catalyst bed temperatures and consumption of the reducing agent(s). The use of steam in the reducing gas is known to cause sintering of the copper crystallites decreasing the activity and shortening the catalyst lifetime. To achieve optimal performance of the catalyst, the reduction should be conducted in dry gas, and care should be taken not to exceed 260°C in the bed.

Cu-based LTS Catalysts Kinetics

Publications dealing with LTS over Cu-based catalysts offer even more conflicting opinions on the reaction mechanism than those considering Fe–Cr catalyzed HTS. Over the years, both redox and associative mechanisms have gained essentially equal support. When copper-based catalysts appeared on the industrial scene, the Cu-based LTS catalyst kinetics was found to obey equations similar to those proposed for HTS,[26] and initially that was the ground for conclusions about LTS occurring via redox mechanism. Further studies produced evidence both conflicting with this hypothesis[27] and supporting it.[22,28–31] It seems likely that depending on LTS process conditions and catalyst properties, the reaction can proceed either via associative or via redox mechanism.

SULFUR-TOLERANT WGS CATALYSTS

The use of Fe–Cr and especially Cu–Zn–Al WGS catalysts in present-day hydrogen plants is directly connected with moderate levels of sulfur-containing compounds in natural gas and naphtha that almost completely displaced coal as the feedstock. It is likely, however, that the incentives for the use of fossil fuels rich in sulfur can be revitalized in the future. If this scenario comes into play, sulfur-tolerant catalysts will be a must for WGS process. Such catalysts are already developed, but so far they have only found a limited use in some particular cases where feed gases with high concentrations of CO and sulfur compounds had to be converted. The best known of this type are cobalt-molybdenum compositions, which are usually supported over alumina and may be promoted with alkali

metals to increase their activity. These catalysts are completely sulfur tolerant; moreover, they have to be sulfided to show maximum activity. Sulfided cobalt-molybdenum catalysts are reported to operate between 230°C and 450°C. Good reviews on compositions and catalytic behavior of these catalysts can be found in the literature.[6,7,10]

NEW CATALYST DEVELOPMENTS FOR EMERGING APPLICATIONS

In recent years, the area of hydrogen production has seen the emergence of new trends stimulated by growing societal concern about environmental protection and efficient use of natural resources. In this aspect, hydrogen has solidly acquired the reputation of being the fuel of the future, with potential uses ranging from internal combustion engines to fuel cells. To meet these expectations, hydrogen fuel will have to become readily available to millions of customers, just like gasoline is at present. This goal requires some important changes to be implemented in the hydrogen industry. The major trend in those upcoming changes is that hydrogen has to be produced by converting fuels in numerous small-scale environmentally-benign and low-maintenance processes. This totally new approach requires new types of catalysts capable of withstanding frequent start–stops under harsh conditions when no special precautions common to the existing hydrogen plants will be available. Therefore new catalysts in particular must be much less prone to self-heating than the conventional WGS catalysts. In addition, the ''green'' image of the emerging applications implies that the new catalysts should be devoid of components harmful to humans and the environment. Development of such catalysts is briefly reviewed below.

Base Metal Catalysts

In developing water-gas shift catalysts for fuel cell applications, efforts have been primarily focused on the LTS rather than the HTS process, since the latter yields too high outlet CO concentration for subsequent CO clean-up steps. New base metal LTS catalysts have been developed specifically for the purpose of fuel processing for fuel cells.[32] These catalysts do not require special precautions during reduction of a fresh reactor charge, as well as during discharge of a spent catalyst. They are more resistant against frequent start–stops typical of fuel cell applications, where traditional Cu–Zn LTS catalysts deactivate because of water condensation on the catalyst or air exposure. However, their activity is somewhat inferior to that of the conventional Cu-Zn-Al catalysts.

Another type of LTS catalyst includes nitrides, carbides, and borides of Group VI and Group VIII metals, with molybdenum carbide apparently being the most active of this group.[33] Those are highly experimental catalysts and are not in production.

Precious Metal Catalysts

Precious metal-based (PM) catalysts have long been known to be active in WGS.[34] investigated catalytic properties of Pt, Ru, Pd, Os, and Ir supported over alumina, silica, and active carbon by running the WGS reaction at atmospheric pressure in the temperature range 270–380°C. They found that alumina-supported catalysts were the most active, with turnover frequencies of the metals at 300°C changing in the order $Ru > Pt > Os > Pd \approx Rh > Ir$.

In recent years, the interest toward PM catalysts has grown from essentially academic to much more practical, as a result of efforts to develop on-board fuel processors for cars. Operational conditions envisaged for such fuel processors preclude the existing Cu–Zn–Al catalysts from being applicable, while PM catalysts are believed to be a viable candidate. Platinum-based catalysts with a great variety of promoters, modifiers, and supports, as well as their preparation conditions seem to be most widely explored in research papers and patent applications. Pt supported on ceria containing support materials have been widely explored for the WGS reaction.[35–38]

In recent years, significant attention has also been given to gold-containing catalysts.[39,40] However, PM catalysts have not yet become an industrial reality, and different opinions regarding their viability have been expressed.[41,42] At present, these catalysts are still too expensive and their activity needs to be improved to assure good performance at low temperatures (~200°C). However, PM catalysts offer important advantages such as insensitivity to frequent start–stops, thermal stability, and operational safety. Besides, significant cost savings and environmental benefits can result from PM recycling.

Reaction Kinetics for Precious Metal Catalysts

Reaction orders and activation energies have been determined for Pt/ceria catalysts by several authors.[37,38,43] There is an agreement that the reaction order with respect to CO is approximately 0 at 200°C. Therefore high CO concentrations do not speed up the reaction for Pt-based catalysts at low temperature, as opposed to Cu-based catalysts which have approximately first-order kinetics. Activation energy estimates range from approximately 46 kJ/mol[37] to 80 kJ/mol.[34]

PRACTICAL ASPECTS

In industrial hydrogen and synthesis gas production, the WGS reaction is used to:

1. Adjust the CO/H_2 ratio for the specific purpose of the synthesis gas (methanol production, Fischer-Tropsch synthesis)
2. Increase the hydrogen yield of the process (i.e., ammonia synthesis; refinery hydrogen production)
3. Reduce the CO concentration to a level amenable to clean-up by preferential oxidation catalysts (fuel cell applications).

Water gas shift converters in industry are largely unchanged from their original design: adiabatic fixed bed reactors with particulate catalysts. There are only two kinds of these reactors in use in industry today: HTS and LTS. They differ in the operating temperatures and catalysts which they use. Two reactors are needed to convert the majority of CO, because of the equilibrium limitations of the process (see previous section). The inlet temperature of the HTS reactor is typically around 320°C. The outlet temperature rises to about 400–450°C because of the reaction exotherm. The gases are cooled to about 200°C before entering the LTS reactor, where the final 2–3% CO is partially converted to CO_2 and H_2.

HTS Reactors

The HTS reactor converts the majority of CO (from about 12% for SR plants to 45% for partial oxidation and coal gasification plants) to CO_2 and H_2 and experiences the majority of the 100–400°C exotherm associated with CO conversion. This conversion can be done in one step (for SR gas) or two to three steps with addition of quench water or inter-cooling (for high CO content gas, i.e., partial oxidation or coal gasification) because of the associated exotherm of CO conversion.[8] Inlet HTS temperatures are typically kept in the range of approximately 320–360°C, which is the compromise determined by such criteria as the catalyst activity, equilibrium outlet CO concentration, catalyst lifetime, reactor materials cost, etc. In particular, higher operating temperatures lead to a more rapid decrease in the catalyst activity due to sintering of magnetite (Fe_3O_4) crystallites, which accelerates with temperature. Typically, the lifetime of the iron-chrome HTS catalyst amounts to 3–5 years and depends on how the catalyst has been handled in a given plant. Aging of the catalyst requires the corresponding gradual increase of the reactor inlet temperature to compensate for the activity decrease. This increase has to be carefully controlled, since unnecessarily high temperature spikes can shorten the catalyst life. Despite all precautions, however, the exit CO concentration eventually gets to a too high level, which necessitates a replacement of the aged catalyst by a fresh charge.

The exit concentration depends on the inlet gas composition and temperature and usually lies between 1% and 4% (dry gas). The space velocity is kept between 500 hr^{-1} and 5000 hr^{-1}.

Poisoning of HTS Catalyst

The HTS catalysts in modern plants are fairly resistant against poisoning. The most common poison almost inevitably present in the feedstocks is sulfur. The iron-chrome catalysts are less sensitive to sulfur compared to the other catalysts used in hydrogen generation processes, such as steam reforming and LTS catalysts, therefore the levels of sulfur-containing compounds in the feed gas that are tolerant for those catalysts do not affect the activity of the HTS catalyst. Moreover, the HTS catalysts, even if sulfided with Fe_3O_4 being converted into FeS (this may occur in the coal-based processes where high concentrations of sulfur-containing compounds are common), still retain the WGS activity at about one half that of Fe_3O_4.[7] For this reason the iron-chrome HTS catalyst is a good choice in processes where high concentrations of sulfur-containing compounds in the feed gas can be anticipated. It also serves as a sulfur guard bed for the more expensive Cu-Zn catalyst.

Other known chemical poisons for the HTS catalyst are halides, although under normal operating conditions they are not present in the feed at an appreciable concentration. Decay in the catalytic activity was also observed with the feed gas which contained minor amounts of unsaturated hydrocarbons, oxygen, and nitric oxides.[7] Under the HTS conditions these compounds were converted into a heavy carbonaceous residue deposited on to the surface of the catalyst, blocking access of the reactants to the catalytic surface.

Commercial HTS catalysts are mechanically quite strong. However, in an industrial process environment, the HTS catalyst can suffer from mechanical factors that deteriorate its performance, such as steam condensation leading to a gradual disintegration of the catalyst pellets, and deposition of foreign components (e.g. particulate matter from corrosion of the process equipment). Increase in the pressure drop across the catalyst bed resulting from these factors is yet another factor determining the catalyst lifetime.

Discharge of HTS Catalyst

The HTS catalyst deactivated beyond an acceptable level has to be discharged from the reactor and

replaced by a fresh charge. Due to a highly exothermic reoxidation of the magnetite phase

$$2Fe_3O_4 + \tfrac{1}{2}O_2 \rightarrow 3Fe_2O_3$$
$$\Delta H = -232.2 \text{ kJ/mol}$$

which results in an adiabatic temperature rise of approximately 450°C, the reduced catalyst should not be exposed to air unless special measures are taken to avoid dangerous overheating. Thus, on the reactor cool down, the catalyst is usually kept under a reducing or inert atmosphere, then discharged under nitrogen and slowly oxidized by gradually admitting air into the catalyst storage vessel. Another procedure for safe catalyst discharge includes filling up the cooled reactor with water and removing the wet catalyst.

Sulfided catalyst must be handled with exceptional precautions, since oxidation of iron sulfide is extremely exothermic:

$$6\,FeS + 13\tfrac{1}{2}O_2 \rightarrow 2\,Fe_2(SO_4)_3 + Fe_2O_3$$
$$\Delta H = -5379 \text{ kJ/mol}$$

Therefore the sulfided catalyst initially needs to be steamed to convert the sulfide into magnetite.

Operation of LTS Catalyst

Temperature dependence of the equilibrium CO concentrations and sintering of copper crystallites (the active phase) at elevated temperatures dictate that the LTS stage be operated at the lowest possible temperature. It is common practice to keep the inlet temperature at least at about 20°C above the dew point of the feed gas, which can be as high as 180–200°C. If water condenses on the catalyst, it can lead to breakage of particles and deactivation of the catalyst.

The catalyst bed temperature increases in the direction of gas flow due to the WGS reaction exotherm. Typical temperature gradients in the bed are about 20–30°C. The lifetime and state of activity of the catalyst is conveniently monitored by the temperature profile through the adiabatic bed. As the reaction front moves through the bed when the catalyst ages, so does the temperature rise from the reaction (Fig. 3).

Poisoning of LTS Catalysts

The lifetime of the LTS catalysts in industrial processes is about 3–4 years. During this time, the catalyst slowly loses its activity due to inevitable sintering and poisoning of copper. To be able to compensate for deactivation and to operate the LTS stage for at least 3 years with required low outlet CO concentrations, the actual charge of LTS catalyst is always overestimated by a factor of about three or more, relative to the amount that would be needed if the activity did not decrease. Sulfur- and halide-containing compounds, the known poisons for copper-based LTS catalysts, gradually decrease the activity even if their concentrations in the feed are very low. The sensitivity to these substances for many years precluded copper-based catalysts from being used in industrial WGS processes. It is only with the massive transfer of hydrogen industry from coal gasification to steam reforming of natural gas and other hydrocarbons, which produces much cleaner gas, that copper-based LTS catalysts became a viable option. The effects of poisons on Cu-based LTS catalysts is described in detail in Refs.[7,8].

Another problem is the formation of methanol over the LTS catalyst. The methanol accumulates in the process condensate and in the gas entering the CO_2 removal system.[44] Therefore, emissions from the plant can become an environmental concern. Catalysts doped with Cs have been developed to address this problem.[44,45] These catalysts reduce the production of methanol by nearly 90%.

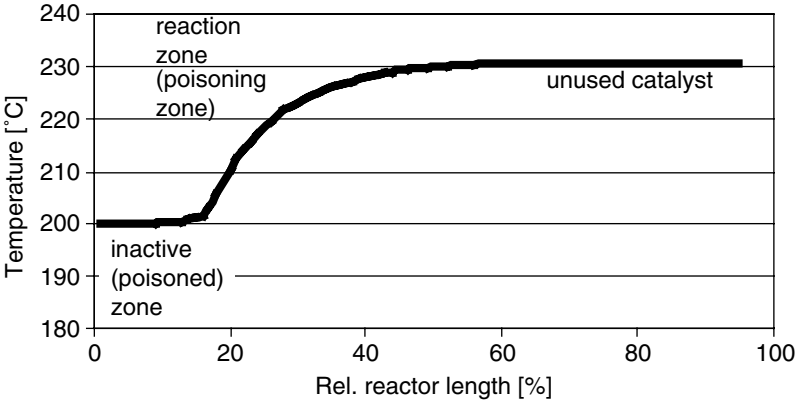


Fig. 3 Typical temperature profile through an LTS bed.

Table 2 Catalyst requirement for different H_2 generation applications

Catalyst properties	Industrial H_2 generation	Small scale H_2 generation
Cost	Very important	Important
Size (volume)	Important	Very important
Sensitivity to water condensation	Accommodated by process	Insensitivity required
Sulfur tolerance	Desired	Desired
Attrition resistance	Important	Important
Reduction requirements	Accommodated by process	Use of process gas
Air sensitivity	Accommodated by process	Insensitivity required

Discharge of LTS Catalyst

After several years of operation at the LTS stage, the CO level at the reactor exit begins to increase due to catalyst deactivation. This signals the need for the catalyst to be replaced and the plant shut down. Discharge of a spent catalyst from the reactor requires special precautions due to the strongly exothermic character of its reoxidation by air, which may generate an exotherm as high as 800–900°C, unless special measures are taken to moderate the temperature rise. Therefore, the common LTS catalyst discharge procedure includes purging the reactor with nitrogen, while cooling it down to below 50°C, followed by the discharge under nitrogen flow with immediate spraying of the catalyst with water to prevent rapid reoxidation. An alternate procedure for a safe catalyst discharge is analogous to the one employed for HTS catalysts and includes filling up the cooled reactor with water and removing the wet catalyst.

SIZE VERSUS COST—NEW CATALYSTS FOR ON-SITE H_2 GENERATION AND FUEL CELL REFORMERS

The industrial means of hydrogen production for methanol synthesis, ammonia synthesis, Fischer-Tropsch synthesis, and refinery hydrogen production have one thing in common: they are tightly controlled processes, run by trained plant operators in an industrial setting.[46] For the emerging applications in hydrogen production, such as fuel cells and small-scale hydrogen production, the operating conditions are different. Often, fuel processors and small hydrogen generators will be operated by inexperienced personnel.[5] Moreover, the typical "duty cycle" of operation for the catalyst is very different. For most industrial processes, this duty cycle consists of continuous steady-state operation with infrequent plant shutdowns. These planned shutdowns are tightly controlled by trained plant personnel and the catalyst is cautiously treated so that it retains its

Table 3 Limitation of WGS catalysts compared to new PM monoliths

	HTS catalyst (FeCr)	LTS catalyst (CuZn)	PM catalyst
Form	3–10 mm pellets	3–6 mm pellets	Monolith supported
Activity	Not active below ~350°C	High activity starting below 200°C	Highly active above 250–300°C
Thermal stability	Thermally stable to ~500°C	Loss of activity above ~260–280°C	Thermally stable to >400°C
Space velocity	<8,000 h^{-1}	<6,000 h^{-1}	>20,000 h^{-1}
Poisons	Low sensitivity to sulfur	Deactivates upon exposure to sulfur and halogens	Tolerant to some poisons
Other considerations	Self-heating material in the reduced state Cr(VI) can be present Inexpensive material Insensitive to sulfur; used to polish sulfur to very low levels	Needs to be carefully reduced Self-heating material in the active (i.e., reduced) state Deactivates by water condensation and air exposure Needs careful discharge from the reactor (self-heating catalyst) Reaction is diffusion limited	Insensitive to start-stop and air exposure No need for pre-reduction Kinetics limit usefulness of the catalyst below 250°C

activity after restart. Industrial startup/shutdown procedures may involve purging with nitrogen and preheating the catalyst bed before restart.

On the other hand, in fuel cell applications, the H_2 generator (reformer) is often tied to the demand for electricity or hot water and therefore operation at night is often not required. These systems may shut down daily (for home systems) or weekly (for small businesses). At night, there is no plant operator available to monitor the proper shutdown and startup conditions. Therefore the catalyst has to be able to withstand repeated water condensation, to ensure fast startup and to respond well to transient conditions due to changing hydrogen demand.

The size of catalytic reactors is a factor of minor importance compared to catalyst and process cost in the industrial environment. This changes as we look at potential future applications of hydrogen production: fuel cell reformers, hydrogen filling stations and on-site hydrogen generation. In some of these applications, size, simplicity, and durability are equally or even more important than minimal cost.

Table 2 compares requirements for the catalysts in industrial hydrogen generation with small-scale production of hydrogen. Table 3 compares the properties of Fe–Cr and Cu–Zn with new precious metal catalyst formulations on monolith supports. A clear match can be seen between the desired properties of catalysts for small-scale H_2 production and actual properties of precious metal catalysts.

CONCLUSIONS

Water gas shift catalysis has been practiced for 90 years in hydrogen and syngas production. High- and low-temperature WGS is run in industry in adiabatic reactors using Fe–Cr– and Cu-Zn-based formulations. The established catalysts have not changed dramatically in the last 40 years, but continue to improve marginally.

New applications for WGS catalysis are slowly emerging. These are small-scale hydrogen production and fuel cell reformers. For these new applications, a new generation of catalysts, including precious metal-based monolith catalysts, is being developed, since the traditional catalysts do not fit the application profile (duty cycle, size and safety requirements).

Sulfur-tolerant shift catalysts have been developed, but are not being used because of the development of other sulfur-sensitive technologies and their higher cost as compared to traditional catalysts.

ACKNOWLEDGMENT

We would like to thank Robert J. Farrauto for helpful discussions.

REFERENCES

1. Holdermann, K. Im Banne der Chemie: Carl Bosch Leben und Werk. Econ-Verlag: Düsseldorf, 1953.
2. Topham, S.A. The history of the catalytic synthesis of ammonia. In *Catalysis: Science and Technology*; Anderson, J.R., Boudart, M., Eds.; Springer-Verlag: New York, 1985; 1–50.
3. Bosch, C.; Wild, W. Canada Patent 153,379, 1914.
4. Khanna, V.K. Hydrogen generation – a brief review. Hydrocarb. Proc. **2001**, *86* (July) 86-A – 86-D.
5. Ladebeck, J.R.; Wagner, JP. Catalyst development for water-gas shift. In *Handbook of Fuel Cells – Fundamentals, Technology and Applications*; Vielstich, W., Gasteiger, H.A., Lamm, A., Eds.; John Wiley & Sons: West Sussex, England, 2003; Vol. 3, 190–201.
6. Kochloefl, K. Water gas shift and COS removal. In *Handbook of Heterogeneous Catalysis*; Ertl, G., Knözinger, H., Weitkamp, J., Eds.; VCH Verlagsgesellschaft: Weinheim, 1997; Vol. 4, 1831–1843.
7. Lloyd, L.; Ridler, D.E.; Twigg, M.V. The water gas shift reaction. In *Catalyst Handbook*, 2nd Ed.; Twigg, M.V., Ed.; Wolfe Publishing: Frome, 1989; 283–338.
8. Hawker, P.N. Shift CO plus steam to H_2. Hydrocarb. Proc. **1982**, *61* (4), 183–187.
9. Farrauto, R.J.; Bartholomew, C.H. *Fundamentals of Industrial Catalytic Processes*: Blackie Academic & Professional, 1997.
10. Newsome, D.S. The water-gas shift reaction. Catal. Rev.- Sci. Eng. **1980**, *21* (2), 275–318.
11. Huang, D.C.; Braden, J.L. High temperature shift catalyst. EU Patent 0353453, 1990.
12. Andreev, A.; Idakiev, V.; Mihajlova, D.; Shopov, D. Iron-based catalysts for the water-gas shift reaction promoted by first-row transition metal oxides. Appl. Catal. **1986**, *22* (2), 385–387.
13. De Araujo, G.-C.; Rangel, M.C. An environmentally friendly dopant for the high-temperature shift catalyst. Catal. Today **2000**, *62*, 201–207.
14. Bahlbro, H. An investigation on the kinetics of the conversion of CO with H_2O over iron oxide based catalysts. In *The Haldor Topsoe Laboratory*; Gjellerup: Copenhagen, 1969.
15. Ruthven, D.M. Activity of commercial water gas shift catalysts. Can. J. Chem. Eng. **1969**, *47* (3), 327–331.
16. Chinchen, G.C.; Logan, R.H.; Spencer, M.S. Water-gas shift reaction over an iron oxide chromium oxide catalyst I: Mass transport effects. Appl. Catal. **1984**, *12* (1), 69–88.
17. Kulkova, N.V.; Temkin, M.I. Kinetics of the reaction of carbon monoxide conversion with steam. Zh. Fiz. Chim. (USSR) **1949**, *23* (6), 695–713.

18. Oki, S.; Happel, J.; Hnatow, M.; Kaneko, Y. The mechanism of the water-gas shift reaction over iron oxide catalyst. Proc. 5th Int. Congress Catal., Hightower, J., Ed.; North-Holland, 1973; 173–183.
19. Boreskov, G.K.; Yur'eva, T.M.; Sergeeva, A.S. Mechanism of carbon monoxide conversion on an iron-chromium catalyst. Kinetika i Kataliz. **1970**, *11* (6), 1476–9.
20. Kubsh, J.E.; Dumesic, J.A. In situ gravimetric studies of the regenerative mechanism for water-gas shift over magnetite: Equilibrium and kinetic measurements in CO and H gas mixtures. AIChE Journal **1982**, *28* (5), 793–800.
21. Retwisch, D.G.; Dumesic, J.A. The effect of metal-oxygen bond strength on properties of oxides: II. Water-gas shift over bulk oxides. App. Catal. **1986**, *21*, 97–109.
22. Rhodes, C.; Hutchings, G.J.; Ward, A.M. Water-gas shift reaction, finding the mechanistic boundary. Catal. Today **1995**, *23*, 43–58.
23. Shchibrya, G.G.; Alekseev, A.M.; Chesnokova, R.V.; Lyudkovskaya, B.G. The study on preparation and calcination of zinc-chrome-copper catalyst for carbon monoxide conversion with steam. Kinet. Katal. **1971**, *12*, 1186.
24. Shchibrya, G.G.; Morozov, N.M.; Temkin, M.I. Kinet. Katal. **1965**, *6*, 1057.
25. Nielsen, P.E.H.; Hansen, J.B. Haldor Topsoe A/S, assignee. Denmark Patent 4980-80, November 18, 1980.
26. Cherednik, E.M.; Morozov, N.M.; Temkin, M.I. Kinet. Katal. **1965**, *6*, 1115.
27. van Herweijnen, T.; de Jong, W.A. Kinetics and mechanism of the CO shift on Cu/ZnO. I. Kinetics of the forward and reverse CO shift reactions. J. Catal. **1980**, *63*, 83–93.
28. Fiolitakis, E.; Hoffmann, H. Dependence of the kinetics of the low-temperature water-gas shift reaction on the catalyst oxygen activity as investigated by wavefront analysis. J. Catal. **1983**, *80*, 328–339.
29. Chinchen, G.C.; Spencer, M.S.; Waugh, K.C.; Wahn, D.A. J. Chem. Soc. Faraday Trans. **1987**, *83* (1), 2193.
30. Ovesen, C.V.; Stolze, P.; Nørskov, J.K.; Campbell, C.T. A kinetic model of the water gas shift reaction. J. Catal. **1992**, *134*, 445–468.
31. Koryabkina, N.A.; Phatak, A.A.; Ruettinger, W.F.; Farrauto, R.J.; Ribeiro, F.H. Determination of kinetic parameters for the water gas shift reaction on copper catalysts under realistic conditions for fuel cell applications. J. Catal. **2003**, *217* (1), 233–239.
32. Ruettinger, W.; Ilinich, O.; Farrauto, R.J. A new generation of water gas shift catalysts for fuel cell applications. J. Power Sources **2003**, *118*, 61–65.
33. Patt, J.; Moon, D.J.; Phillips, C.; Thompson, L. Molybdenum carbide catalysts for water-gas shift. Catal. Lett. **2000**, *65*, 193–195.
34. Grenoble, D.C.; Estadt, M.M.; Ollis, D.F. The chemistry and catalysis of the water gas shift reaction 1. The kinetics over supported metal catalysts. J. Catal. **1981**, *67*, 90–102.
35. Jacobs, G.; Patterson, P.M.; Graham, U.M.; Sparks, D.; Davis, B.H. Low temperature water-gas shift, kinetic isotope effect observed for decomposition of surface formates for Pt/ceria catalysts. Appl. Catal. A, General **2004**, *269*, 63–73.
36. Silver, R.G. UTC Fuel Cells, LLC, assignee. Shift converter having an improved catalyst composition and method for its use. US Patent 6,455,182, 2002.
37. Bunluesin, T.; Gorte, R.J.; Graham, G.W. Studies of the water-gas-shift reaction on ceria-supported Pt, Pd and Rh, implications for oxygen-storage properties. Appl. Catal. B **1998**, *15*, 107–114.
38. Hilaire, S.; Wang, X.; Luo, T.; Gorte, R.J.; Wagner, J. A comparative study of water-gas shift reaction over ceria supported metallic catalysts. Appl. Catal. A **2004**, *258*, 271–276.
39. Fu, Q.; Weber, A.; Flytzani-Stephanapoulos, M. Nanostructured Au-CeO_2 catalysts for low-temperature water-gas shift. Catal. Lett. **2001**, *77*, 87–95.
40. Andreeva, D.; Idakiev, V.; Tabakova, T.; Ilieva, L.; Falaras, P.; Bourlinos, A.; Travlos, A. Low-temperature water-gas shift reaction over Au/CeO_2 catalysts. Catal. Today **2002**, *72*, 51–57.
41. Zalc, J.M.; Sokolovskii, V.; Löffler, D.G. Are noble metal based water-gas shift catalysts practical for automotive fuel processing? J. Catal. **2002**, *206*, 169–171.
42. Xue, E.; O'Keeffe, M.; Ross, J.R.H. Water-gas shift conversion using a feed with a low steam to carbon monoxide ratio and containing sulfur. Catal. Today **1996**, *30*, 107–118.
43. Koryabkina, N.; Ribeiro, F.; Ruettinger, W. Determination of kinetic parameters for water-gas shift catalysts under realistic conditions for fuel cell applications. In *Fuel Cell Technology: Opportunities and Challenges*, AICHE Spring National Meeting, March 10–14, New Orleans, 2002; 92–97.
44. Carstensen, J. Reduce methanol formation in your hydrogen plant. Hydrocarb. Proc. **1998**, (3), 100C-D.
45. Storgaard, L.; Nielsen, H.C. Catalysts for cost effective hydrogen production. PTQ **1999** (July), 127–133.
46. Twigg, M.V. *Catalyst Handbook*, 2nd Ed.; Frome: Wolfe Publishing, 1989.

Water Reclamation

Mark A. Kuehne
Norman N. Li
Richard Q. Song
Maxwell Tsai
Jane C. Li
NL Chemical Technology, Inc., Mount Prospect, Illinois, U.S.A.

INTRODUCTION

With the increasing world population and limited water resources, water reclamation is becoming increasingly important in providing and maintaining sustainable water supplies. Water reclamation may be defined as the treatment of wastewater to make it suitable for one or more beneficial uses. One of the most important water treatment technologies used in water reclamation is membrane technology. Therefore, this entry is focused on membrane technology and its use for water reclamation.

Membrane technology used in water reclamation includes five major membrane types: reverse osmosis, nanofiltration, ultrafiltration, microfiltration, and liquid membranes. These five types of membranes are discussed briefly, and examples of their applications in municipal and industrial wastewater reclamation is also described.

REVERSE OSMOSIS MEMBRANES

When two saline solutions are separated by a semipermeable membrane, water will pass through the membrane from the side with lower salt concentration to the side with higher salt concentration. This spontaneous passage of water is called osmosis. One of the most important examples of osmosis is transport of water through cell membranes. The driving force of the osmosis process is the gradient of the chemical potential across the membrane. The water passage across the membrane continues until the chemical potential of the water is equal on both sides. At equilibrium, the pressure difference between the two sides of the membrane is equal to the osmotic pressure difference. As a rule of thumb, every 100 mg/L of dissolved salts produces roughly 1 psi of osmotic pressure.

If a pressure higher than the osmotic pressure difference is applied to the side with high salt concentration, the water flow can be reversed. This process is termed reverse osmosis (RO), also known as hyperfiltration. This phenomenon makes the separation of water from saline solution possible. Fig. 1 provides a schematic drawing of the osmosis and reverse osmosis processes.

Reverse osmosis membrane is a semipermeable membrane, which is permeable to water but not to other species, such as dissolved salts. The observed salt rejection R of an RO membrane is defined as

$$R = 1 - c_p/c_f$$

where c_f is the solute concentration in the feed and c_p is the solute concentration in the permeate.

Reverse osmosis membranes are usually synthetic membranes and are made of polymers. Depending on their structures, RO membranes can be classified as either asymmetric or composite.

An asymmetric membrane has a very thin dense top layer (or skin) with a thickness of 0.1–0.5 μm. A porous sublayer with a thickness of approximately 50–150 μm supports the dense top layer. The thin dense skin facing the feed solution acts as the selective layer, allowing water passage but rejecting dissolved solids. The resistance to mass transfer across the membrane is also mainly determined by the thin top layer. In asymmetric membranes, the selective top layer and the porous support layer are made of the same polymer material. Asymmetric membranes can be obtained by phase inversion, a technique in which a polymer in solution is transformed in a controlled manner from a liquid into a solid form. The top skin layer and the porous support layer are formed in a single-step process.

In composite RO membranes, the selective top layer and the porous support layer are usually made of different polymeric materials. The selective top layer is formed on the porous support in a second step, typically by an interfacial polymerization reaction.[1] For example, a commercially available thin film composite RO membrane is made by coating a porous polysulfone support with a polyamide thin film formed by the interfacial reaction of *m*-phenylenediamine and 1,3,5-benzenetricarbonyl trichloride. Details regarding membrane structures can be found elsewhere in the

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120016186

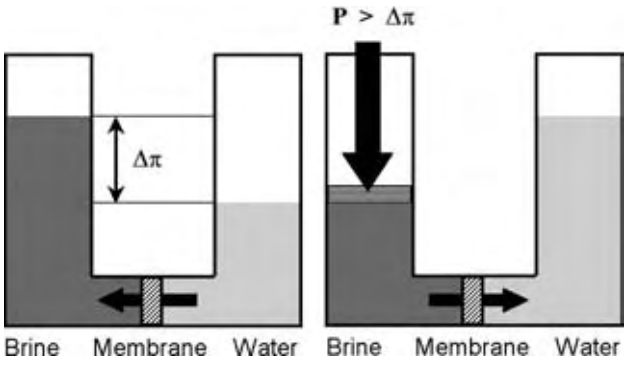


Fig. 1 Schematic of osmosis and reverse osmosis processes.

literature.[2–4] A schematic diagram of a thin film composite membrane is provided in Fig. 2.

Reverse osmosis membranes can be further classified according to their applications or processes as either seawater (SW) membranes or brackish water (BW) membranes. Seawater membranes are high-pressure membranes because seawater desalination is carried out at high feed pressure (5.5–10 MPa). Seawater membranes have very high salt rejection (>99.5% NaCl rejection) and moderate water flux. Brackish water membranes, on the other hand, are low-pressure membranes. The operating pressure of BW membranes is usually between 1.4 and 4.0 MPa. The inorganic salt rejection of BW membrane is approximately 95–99.5%, and the water flux is higher than that of SW membrane. When RO membranes have fixed charge groups (such as carboxylate anions), improved salt rejection can be obtained by means of Donnan exclusion. In principle, RO membrane is suitable for treatment of organic as well as aqueous solutions. However, problems with membrane stability under actual operation conditions make organic solvent applications very limited.

Among others, spiral-wound and hollow-fiber modules are two commonly available RO membrane products on the market. Currently, spiral-wound modules have approximately 80% of the total RO membrane market. Thin film composite membrane

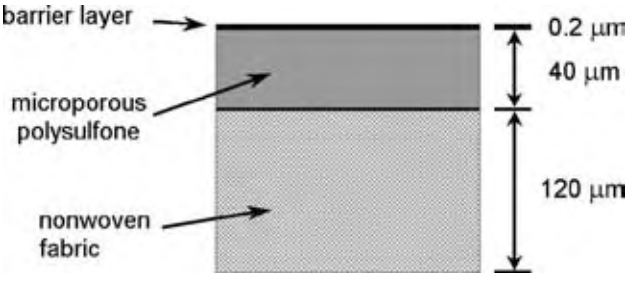


Fig. 2 Schematic diagram of a thin film composite membrane.

is usually sold in spiral-wound modules. Two flat membrane sheets are glued (sandwiched) together with a product spacer (permeate spacer) in between to form an envelope. The envelope is sealed on three sides with the fourth open side attached to a product tube. Many of these envelopes are then rolled around a product tube, with feed spacers between the facing membrane surfaces. Under pressure, the feed stream enters the module from one end and exits from the other end. The product water is collected from the center tube. Fig. 3 provides a cutaway view of a spiral-wound module.

Reverse osmosis membrane is widely used in seawater and brackish water desalination processes. Compared to traditional distillation, there is no energy-intensive phase change involved in membrane processes. Therefore, desalination with RO membrane is more energy efficient. In addition to the traditional desalination processes, RO membranes have also found wide application in industrial and municipal wastewater treatment, in pure water production for the electronic and pharmaceutical industries, and in the food industries.

Current RO membrane research is focused mainly on improving membrane performance and membrane resistance to fouling. Fouling is the deposition of components from the feed stream onto the membrane surface, which reduces membrane flux. Membrane fouling is a major obstacle for the efficient use of membrane technology. Membrane fouling can dramatically reduce process efficiency and, ultimately, also shorten membrane life. Methods to improve fouling resistance of membranes include chemical modifications of membranes and improvements in module design. The membrane surface can be modified to suit specific applications. For instance, making the membrane surface more hydrophilic will improve membrane biofouling characteristics; a smooth surface will reduce the tendency of particle deposition onto the membrane surface. Introducing surface charge is found to be helpful in certain types of applications. Because most colloidal particles, proteins, and cells are negatively charged in aqueous solution, one can expect that membrane with negative charges will be more fouling resistant.[5]

Another area of continuing interest and need is to provide RO membranes having high flux at low pressure, while maintaining high solute rejection characteristics.[6,7] The major advantages of using such membranes are low energy consumption and low equipment cost. Compared to the first generation of asymmetric cellulose acetate RO membranes operated at 3 to 4 MPa, current low-pressure RO membranes can be run at approximately 1 MPa. Research efforts to further enhance the permeation rate of RO membranes and reduce energy requirements are ongoing.[8–10]

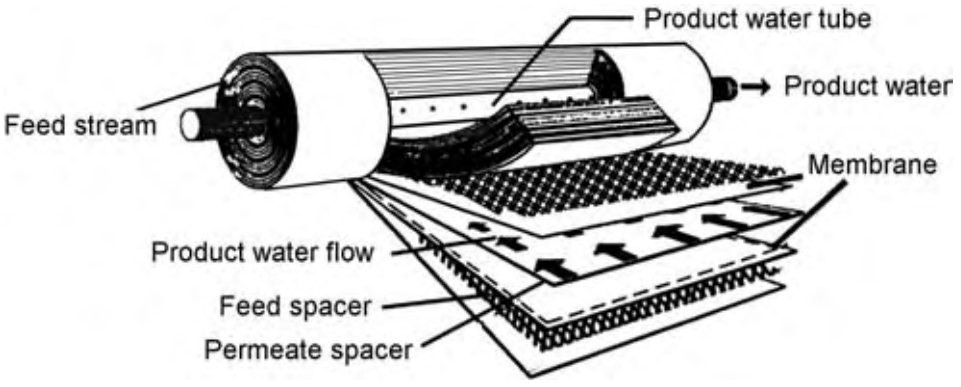


Fig. 3 Cutaway view of a spiral-wound membrane module.

NANOFILTRATION MEMBRANES

Nanofiltration (NF) membrane is a category of membrane between reverse osmosis and ultrafiltration (UF). An NF membrane can reject organic molecules with molecular weights (mol wt) in the range of 300 to 1000, a lower range that can be separated by UF. However, NF membranes typically have much higher flux than RO membranes at low pressures (0.5 MPa or lower), and they have lower rejection than RO membranes for neutral molecules below 150 mol wt and for monovalent inorganic salts.[11] For example, the NaCl rejection of most NF membranes is in the range of only 30–90%.

Most NF membranes have a large concentration of negatively charged groups, such as carboxylate ions, covalently bonded to their surfaces. These surface anions give NF membranes a high rejection for salts with multivalent anions, such as sulfates, carbonates, and phosphates. Nanofiltration membranes are commonly used as water softening membranes, because they can very effectively remove most hard water components, i.e., carbonates and sulfates of calcium and magnesium. NF membranes designed for this purpose are usually rated in terms of their $MgSO_4$ rejection, which is typically in the range of 95–99%.

Other applications of NF membrane include organics removal from surface water, radium removal from ground water, sulfate removal from seawater, and food and pharmaceutical applications such as concentration of dilute solutions and desalting of cheese whey.

Like RO membranes, many NF membranes are polyamide thin film composite membranes. These membranes can be prepared by interfacial reaction of piperazine with 1,3,5-benzenetricarbonyl trichloride and/or isophthaloyl dichloride, or by treating polyamide thin film composite RO membranes with compounds such as mineral acids, to increase their flux and lower their salt rejection.[2] A few ceramic NF membranes have also been developed. New methods to produce NF membrane are a subject of ongoing research, as exemplified by two recently reported methods: filling the pores of microfiltration (MF) membrane with cross-linked poly(4-vinylpyridinium) salts,[12] and grafting charged groups onto the surface of UF membrane.[13]

ULTRAFILTRATION AND MICROFILTRATION MEMBRANES

Ultrafiltration and microfiltration membranes are both porous membranes but with two different ranges of pore size. Microfiltration membranes are used to separate particles in the range of 0.02–10 μm. Ultrafiltration membranes are used to remove much smaller species, such as dissolved macromolecules and colloids with 1000–500,000 mol wt.

Ultrafiltration membranes have pore diameters in the range of 1–100 nm and separate primarily by size exclusion. In this process, a liquid containing dissolved or suspended matter is driven through the membrane by an applied pressure difference across the membrane, and any species larger than the pore size of the membrane is rejected (prevented from passing through the membrane). The observed rejection R of a species is given by

$$R = 1 - c_p/c_r$$

where c_p is the concentration of the species in the permeate liquid that passes through the membrane, and c_r is the concentration of the same species in the retentate stream, which remains on the feed side of the membrane.

The pore size of a UF membrane is described by its nominal molecular weight cutoff (MWCO), defined as the molecular weight of the smallest species for which the rejection is 90% or greater. Although the nominal

MWCO is a useful measure of a membrane's average pore size, it is only an approximate indication of how a UF membrane may perform in a given application. In reality, the rejection level for a solute molecule depends not only on its molecular weight, but also on its shape, its molecular orientation, and operating conditions such as feed pressure and solute concentration. To eliminate the effect of molecular orientation in MWCO measurement, spherical macromolecules such as globular proteins are recommended as test solutes. Although there is at present no generally accepted standardization of the procedure for MWCO measurement, the factors that affect the measurement have been reviewed in detail, and recommended operating conditions have been proposed.[14]

Two phenomena that complicate UF processes are concentration polarization and fouling. During filtration, the solution upstream of the membrane becomes enriched in the retained molecules or particles. At high permeation flux rates, mixing and diffusion in the upstream liquid might not be rapid enough to maintain uniform concentration, resulting in a layer of highly concentrated solution next to the membrane. This effect is called concentration polarization, and it leads to lower flux and lower apparent rejection, because the effective solute concentration at the membrane surface is higher than that of the bulk solution. Fouling is the deposition of components from the feed stream onto the surface and in the pores of the membrane, which can cause total pore blockage or an effective reduction of pore diameter, and leads to lower flux and in some cases, increased rejection. Fouling can be partially reversed by membrane cleaning, but some fouling is also irreversible, eventually making membrane replacement necessary.

Although some inorganic membranes are available, most UF membranes are made of polymers. The earliest UF membranes were made of cellulose acetate, but today, the most widely used polymers are polysulfone and polyethersulfone, which are preferred because of their higher resistance to extremes of pH and temperature. Other polymers used include polyvinylidenefluoride, polyacrylonitrile, and polyamides.[15]

Most UF membranes are asymmetric, having a thin separating layer or ''skin layer'' with small pores on one side of the membrane, and a much thicker layer with larger pores below the membrane which provides structural support with minimum flow resistance. Asymmetric membranes are manufactured by wet phase inversion casting. In this process, a casting solution of a polymer in a water-miscible solvent is spread in a thin layer onto a flat surface and then immersed in water. The water causes extraction of solvent and precipitation of the polymer as a porous flat sheet. The skin layer is formed on the upper surface that was in direct contact with water, and the underlying porous structure has progressively larger pores with increasing distance from the skin layer. The skin layer pore size and MWCO are determined by the casting solution formulation (polymer concentration, type of solvent, and additives) and by the phase inversion casting process conditions.

Flat sheet UF membrane is usually manufactured with a polyester or polyolefin nonwoven fabric backing for added durability, and membrane sheets are rolled into spiral-wound membrane modules (see Fig. 3). Ultrafiltration membranes are also commonly manufactured as hollow membrane fibers and sold as modules containing bundles of fibers. Both spiral-wound and hollow fiber membrane modules are usually operated in crossflow filtration mode, wherein the feed liquid flows tangential to the surface of the membrane under pressure. The high tangential feed velocity helps to sweep particles away from the membrane surface and reduce concentration polarization and fouling. In crossflow filtration, only a portion of the feed passes through the membrane and the remaining liquid exits as a separate retentate stream, carrying rejected components away and out of the membrane module.

Microfiltration membranes are similar to UF membranes but have larger pores. Microfiltration membranes are used to separate particles in the range of 0.02–10 μm from liquid or gas streams. Commercial MF membranes are made from a wide variety of materials including polymers, metals, and ceramics. A wide variety of membrane module designs are available including tubular, spiral wound, pleated sheet, hollow fiber, and flat sheet designs. Some modules are best suited for crossflow filtration,[16] and others are designed for dead-end filtration.[17] In dead-end filtration, the feed liquid flows normal to the surface of the membrane, and retained particles build up with time as a cake layer on the membrane surface or within the pores of the membrane.

LIQUID MEMBRANES AND FACILITATED TRANSPORT MEMBRANES

Liquid membranes are ultra-thin films of water or an organic liquid that are stabilized by the presence of surfactant molecules and other additives. This film is placed between two miscible phases: one is a feed phase, and the other is a phase receiving the materials that are transferred from the feed phase through the membrane phase. This means that if the feed and the receiving phases are both aqueous, an organic liquid membrane will be needed, whereas if the feed and the receiving phases are organic, the liquid membrane will need to be aqueous. In commercial applications, liquid membranes are prepared via emulsion techniques as agglomerations of micrometer-sized droplets.

This emulsification technique aims to maximize membrane surface area. The internal phase of the droplets can be placed inside the pores of polymeric films, or inside hollow fibers.[18–21] The membrane phase can contain surfactants or no surfactants for easy coalescence. This aspect of the liquid membrane system will be discussed in more detail in the application section.

The transfer of materials through liquid membranes is based on two facilitated transport mechanisms. In the type 1 facilitation, the solute reacts with a reagent dissolved in the receiving phase, or the internal phase in an emulsion liquid membrane system. This reaction should be selected so that it produces a nonpermeating compound. The purpose is to effectively reduce the concentration of the solute to zero in the receiving phase, thereby facilitating the continuous transfer of the materials. The type 2 facilitation uses a complexing agent for the transferred compound in the membrane phase. The complexing agent reacts with the transferred compound inside the membrane phase to form a complex, which permeates through the membrane phase to the receiving phase, or the internal phase in an emulsion liquid membrane system. The transfer through the membrane phase is based on the concentration gradient of the complex molecules, and the decomplexation at the interface between the membrane and the receiving phase is based on a counter-diffusion of ions. The counter-diffusing ions from the receiving phase are usually either hydrogen ions or hydroxide ions from the acid or base materials placed in the receiving phase (Fig. 4). The facilitated transport will stop when the amount of the acid or base material is consumed.[21–23]

There have been several modified systems since the invention of liquid membranes, including a facilitated transport mechanism. One of them is to disperse the receiving solution in an organic membrane phase on one side of a porous hollow fiber.[24,25] Two plants to treat contaminated groundwater were built and operated based on this revised liquid membrane system. More discussion about this application is given below in the application section.

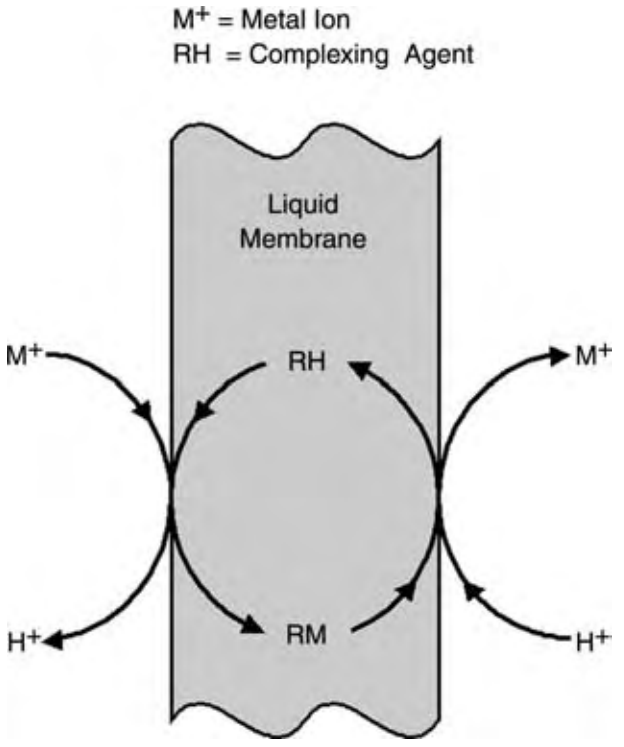


Fig. 4 Facilitated transport in a liquid membrane system with complexing agent in the membrane phase and acid in the receiving phase.

MEMBRANE PROCESSES FOR WATER RECLAMATION

With world population increase and scarcity of water sources, the key to providing a sustainable development of water resources is to carry out water reclamation. The concept of water reclamation is to treat wastewater from different sources to meet different water quality requirements of various water applications economically. USEPA provided guidelines for nonpotable water reuse,[26] which stated various conventional technologies commonly used in treating municipal wastewater such as primary and secondary treatment, and specified the water quality level required for various applications such as irrigation, industrial cooling processes, and groundwater recharge. As for industrial wastewater reuse, the American Institute of Chemical Engineers Center for Waste Reduction Technologies published a book[27] to provide guidelines for a systematic approach, which was summarized by Zinkus.[28]

With the advancement of membrane technologies and engineering in the integration of MF, UF, NF, and RO membrane systems, and possibly a liquid membrane system, wastewater reclamation can provide high quality water for underground water replenishment and direct household nonpotable use, and source water for ultra-pure water applications. Membrane technology provides several advantages over conventional treatment processes:

1. It is simpler to apply in treating a wide range of contaminants in wastewater effectively.
2. A membrane system occupies a smaller footprint. Efficient use of floor space is critical in meeting the future expansion needs of municipal wastewater treatment plants, which are typically located within densely populated areas.
3. Modular design of membrane treatment systems allows easier future expansion.
4. Less coagulation or flocculation chemicals are required in the treatment processes.

Membrane technology will play a major role in water treatment and wastewater reclamation, as indicated in *Desalination and Water Purification Technology Roadmap*,[29] and will require continued research and development in:

1. reducing the operational cost of membrane systems;
2. enhancing membrane fouling resistance to reduce membrane cleaning frequency; and
3. providing better on-line monitoring of membrane module and system integrity to ensure a 100% safety barrier in rejecting bacteria and viruses.

It should be noted that, in developing a membrane process for treating a specific type of wastewater efficiently and effectively, the Six Sigma quality management program was suggested to be incorporated in the development process.[30]

Municipal Wastewater Reclamation

In the Orange County Sanitary District of California, the wastewater treatment plant (Fig. 5) treats approximately 200 million gallons per day (MGD) of municipal wastewater through preliminary treatment (bar screens and grit processing to separate solid waste from wastewater), primary treatment (a physical/chemical treatment process to settle small solids), and secondary treatment (an aerobic bio-treatment process consisting of trickling filter and activated sludge plants to treat dissolved matter with activated sludge). Effluent from secondary treatment is discharged into the ocean. Because of excessive exploitation of underground water in this area, however, a barrier to prevent seawater intrusion was necessary. In 1976, Water Factory 21 was built to provide highly treated (purified) water as this water barrier. This treatment plant reclaimed 15 MGD of the secondary treatment effluent from Orange Country Sanitary District's wastewater plant by pretreatment (flocculation, recarbonation, and multi-media filtration), followed by division into two treatment processes: 6 MGD was purified by an RO membrane system, and 9 MGD was treated with activated carbon and disinfected with chlorine. This highly treated wastewater (with quality of potable water) was then blended with underground water before being injected into the ground as a water barrier for seawater intrusion prevention.

To provide a sustainable water resource in Orange County, a 100,000 acre-foot per year Ground Water Replenishment System (GWRS) was proposed[31] on the basis of the success of the Water Factory 21 project. Because of limitations in land space available to this treatment plant, a new pretreatment process using a membrane system (MF and UF system) was tried and approved by virtue of its superior performance, cost effectiveness, and smaller footprint than the conventional pretreatment process.[32] The first phase of GWRS, a 71,600 acre-foot wastewater reclamation project using a new integrated membrane process consisting of an MF system followed by an RO system, and last by a UV disinfection system, has been under construction since 2001.

Membrane technology also shows its potential in replacing the secondary activated sludge treatment step. Although using a membrane system as a replacement for sedimentation in the activated sludge process was tried in the late 1960s, current membrane bioreactor (MBR) systems provide more robust engineering design to overcome membrane fouling.[33] Compared to the conventional activated sludge process, the MBR provides complete solids removal, a significant disinfection capacity, high treatment rate, and high efficiency of organic and nutrient removal. Therefore, the MBR strengthens the future role of membrane systems in wastewater recycling and reuse applications. Currently, more than 500 commercial MBR systems are in operation worldwide.

Industrial Wastewater Reclamation

In addition to wastewater effluent treatment regulations as a driving force for industrial wastewater treatment, other factors such as water and wastewater management costs, operation costs, and recovery and recycle of processing chemicals play critical roles in determining industrial wastewater reclamation needs and treatment processes.

With 16 million tons/day of water usage, the paper and pulp industry in the USA generates enormous wastewater pollution problems.[34] Characteristics of wastewater from pulp and paper are high BOD and COD, and high suspended solids content. The industrial "end of pipe" abatement approach is to treat all wastewater together, using primary treatment to settle the solids and to biological processes such as activated sludge to reduce BOD and COD content. Industry has tried membrane technology to reclaim wastewater, thereby reducing water consumption and wastewater pollution. Without significant pretreatment of wastewater, however, traditional hollow fiber, spiral-wound, or tubular membrane modules would be fouled within a very short time. Nuortila-Jokinen[35] used shear enhanced membrane modules to overcome the membrane fouling problem, thereby allowing membrane processes to treat and reuse pulp mill effluent effectively. Shear enhanced membrane modules used a vibration or spinning force to produce 150,000 s^{-1}

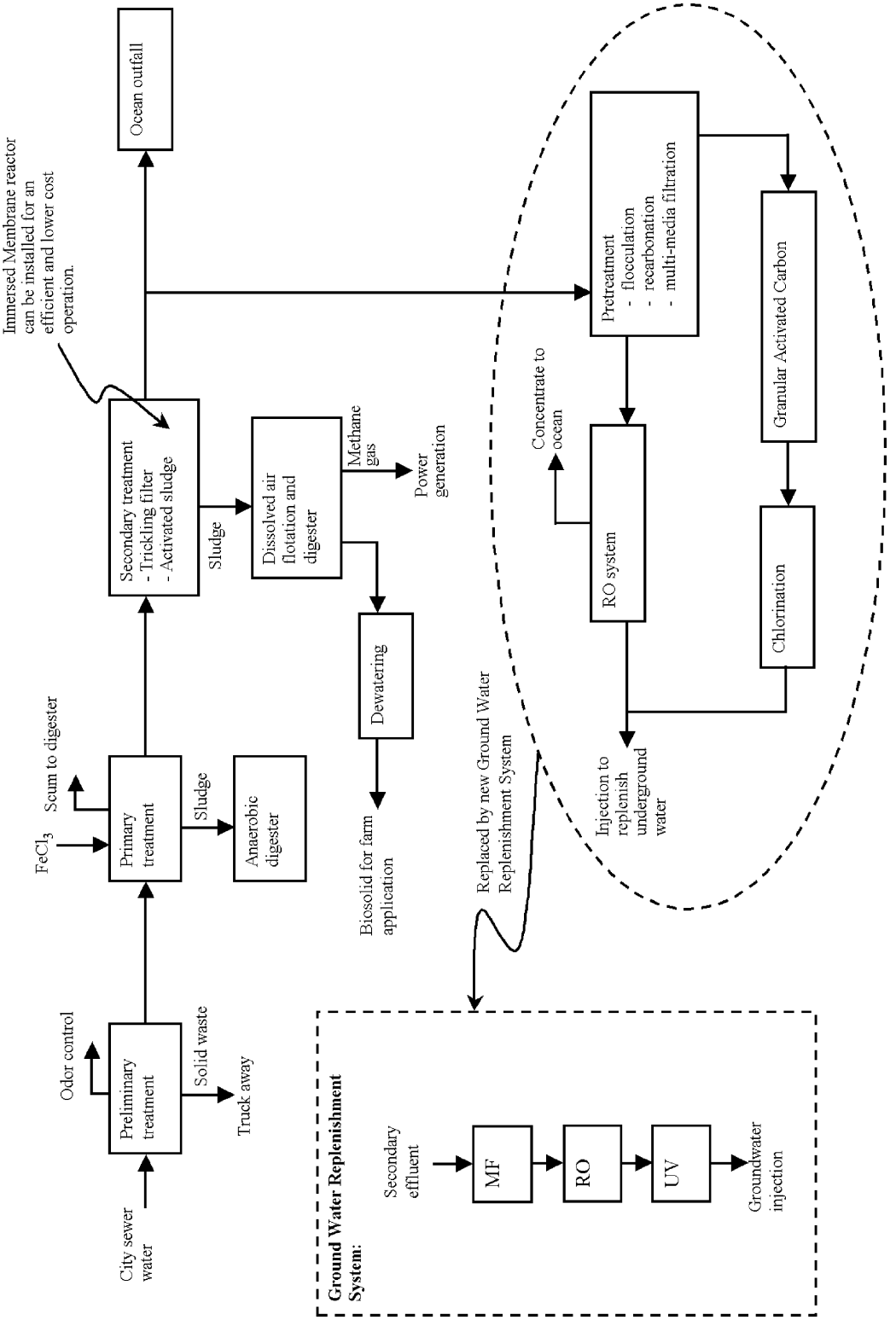


Fig. 5 An overall block diagram of the Orange County municipal wastewater treatment process and Water Factory 21 process.

shear rate at the membrane surface, which was three to five times the rate attainable in crossflow systems.[36] This dispersal force prevented foulants from plugging up the membrane surface and resulted in a significantly higher permeation rate and longer operation time between cleanings. Wagner[37] described the application of membranes for wastewater treatment in a paper manufacturing plant. Wastewater generated during the process of paper formation was rather white because of the high content of suspended solids. This white water from the paper mill was reclaimed by the use of a plate and frame UF filter. Wagner also noted that as more water was recycled in a pulp or paper mill, more salt built up. Salt concentration had to be reduced in order to decrease the possibility of corrosion in machinery and pipelines. Nanofiltration or RO membranes showed capability in significantly reducing salt content. Using UF membrane, fresh water consumption of the pulp and paper industry could generally be reduced by 50%, while NF or RO membrane could provide a further 50% reduction.

Two commercial size plants for groundwater treatment based on liquid membrane technology in general, and the supported liquid membrane using hollow fibers in particular, were built and operated in Baltimore, U.S.A.[20] Specifically, the purpose of the two plants is for hexavalent chromium cleanup. One plant went into commercial operation in March 1999 and the other approximately about a year later. The liquid membrane system in these two plants is able to reduce metal-ion concentration from 100–1000 ppm range to approximately 0.05 ppm and, meanwhile, produce a concentrated chromium solution, which is the spent strip solution, at approximately 20% Cr (VI). This concentration is suitable for sale for reuse.

Other commercialized major applications for liquid membranes are the recovery of zinc from rayon plant effluents,[38] the removal of phenol from industrial wastewater streams,[39] and the treatment of cyanide-containing wastewater from gold leaching solutions.[40,41] The process of extraction of uranium from phosphoric acid was successfully developed in a pilot plant and is waiting for commercial opportunity.[42]

Membrane technology also helped an aluminum can manufacturer (Fig. 6) treat processes wastewater to reduce ever-increasing fresh water and wastewater discharge fees.[43] Oily wastewater from the can forming process and other facilities processes was filtered by a UF unit. The UF permeate combined with can washing wastewater was further treated by an NF unit. Three RO units were used to purify the incoming municipal water and the NF permeates. Using such an integrated membrane system, this aluminum can manufacturer recovered approximately 88% of the process water.Cassano[44] reviewed the potential of using UF and NF in the treatment of aqueous solutions from leather processing industries. Water is used in almost every step during leather manufacturing processes. Through the application of various types of membranes, process water can be either recycled directly or used in other steps of leather manufacturing. Polysulfone spiral-wound UF modules with an MWCO of 20,000 could maintain the sulfide concentration during the unhairing steps, and NF element could concentrate the tanning solution for reuse instead of discharge into the wastewater treatment plant.

Semiconductor wafer production also consumes a large quantity of fresh water. The Philips San Antonio facility was producing 150 mm wafers and planned to move into 200 mm wafer production in early 2000. One task it encountered was the need to increase its supply of high purity water. Its project team evaluated several options, including expanding the fresh water supply and recycling waste wafer rinse water. The team

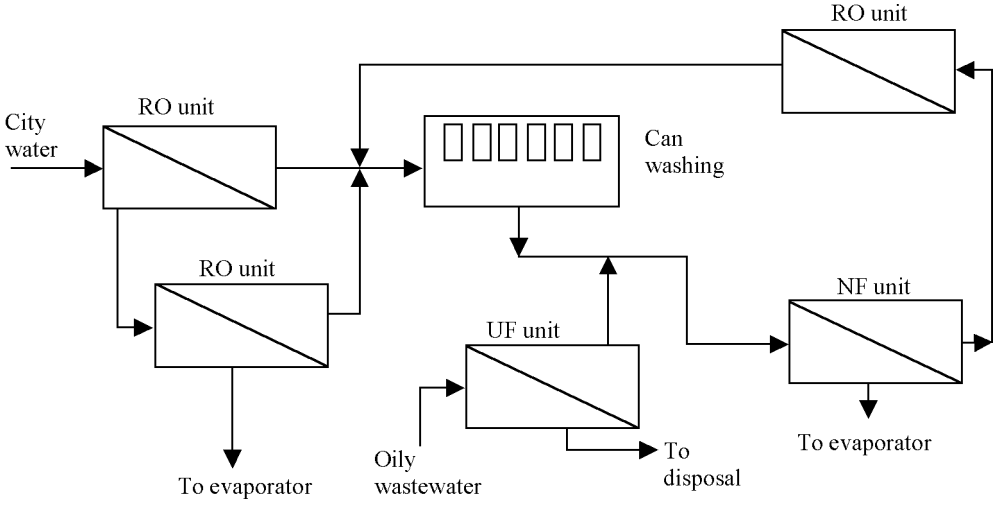


Fig. 6 Integrated membrane system used in a can manufacturing plant.

concluded that reclaiming and recycling the process wastewater would meet both the project's critical schedule milestones as well as its budget.[45] A 180 gpm double pass RO system was installed with weak acid cation resin as RO pretreatment and post-treatment. This system provided a maximum 87% recovery rate with 93% or better rejection of waste rinse water TOC.

CONCLUSIONS

Water reclamation, the treatment of wastewater to meet the water quality standards of various applications economically, is becoming increasingly important in view of the increasing world population and scarcity of fresh water sources. The major technology used for water reclamation is membrane technology. This entry gives an overview of the major membrane types used for water reclamation: reverse osmosis, nanofiltration, ultrafiltration, microfiltration, and liquid membranes. Applications of these membranes in municipal and industrial wastewater reclamation have been described.

REFERENCES

1. Morgan, P.W. Interfacial polycondensation in unstirred systems. In *Condensation Polymers*: *by Interfacial and Solution Methods*; Interscience Publishers: New York, 1965; 19–64.
2. Petersen, R. Composite reverse osmosis and nanofiltration membranes. J. Membr. Sci. **1993**, *83* (1), 81–150.
3. Mulder, M. *Basic Principles of Membrane Technology*; Kluwer Academic Publishers: Boston, 1991; 10–12.
4. Strathmann, H. Synthetic membranes and their preparation. In *Handbook of Industrial Membrane Technology*; Porter, M., Ed.; Noyes Publications: Park Ridge, NJ, 1990; 12–1445–49.
5. Chen, V.; Fane, A.G.; Fell, C.J.D. The use of anionic surfactants for reducing fouling of ultrafiltration membranes: their effects and optimization. J. Membr. Sci. **1992**, *67* (2–3), 249–261.
6. Li, N.N.; Kuehne, M.A.; Petersen, R.J. High Flux Reverse Osmosis Membrane. US Patent 6,162,358, December 19, 2000, NL Chemical Technology, Inc.
7. Li, N.N.; Kuehne, M.A.; Petersen, R.J. A novel high flux membrane. ICOM '99: the 1999 International Congress on Membranes & Membrane Processes, Toronto, Canada, June 12–18, 1999; North American Membrane Society, 1999; Abstract Number 236.
8. Kuehne, M.A.; Song, R.Q.; Li, N.N.; Petersen, R.J. Flux enhancement in TFC RO membranes. Environ. Prog. **2001**, *20* (1), 23–26.
9. Kuehne, M.A.; Li, N.N.; Petersen, R.J. Flux enhancement in TFC membranes. ICOM '99: the 1999 International Congress on Membranes & Membrane Processes, Toronto, Canada, June 12–18, 1999; North American Membrane Society, 1999; Abstract Number 82.
10. Song, R.Q.; Ho, W.S.; Li, N.N.; Petersen, R.J. Polymeric TFC membrane formation studies. ICOM '99: the 1999 International Congress on Membranes & Membrane Processes, Toronto, Canada, June 12–18, 1999; North American Membrane Society, 1999; Abstract Number 469.
11. Eriksson, P. Nanofiltration extends the range of membrane filtration. Environ. Prog. **1988**, *7* (1), 58–62.
12. Mika, A.M.; Childs, R.F.; Dickson, J.M. Ultra-low pressure water softening: a new approach to membrane construction. Desalination **1999**, *121* (2), 149–158.
13. Béquet, S.; Remigy, J.-C.; Rouch, J.-C.; Espenan, J.-M.; Clifton, M.; Aptel, P. From ultrafiltration to nanofiltration hollow fiber membranes: a continuous UV-photografting process. Desalination **2002**, *144* (1–3), 9–14.
14. Cheryan, M. Membrane properties. In *Ultrafiltration and Microfiltration Handbook*; Technomic Publishing Company: Lancaster, Pennsylvania, 1998; 89–110.
15. Kulkarni, S.S.; Funk, E.W.; Li, N.N. Ultrafiltration membranes. In *Membrane Handbook*; Ho, W.S., Sirkar, K.K., Eds.; Van Nostrand Reinhold: New York, 1992; 408–431.
16. Mir, S.; Michaels, S.L.; Goel, V; Kaiser, R. Crossflow microfiltration: applications, design, and cost. In *Membrane Handbook*; Ho, W.S., Sirkar, K.K., Eds.; Van Nostrand Reinhold: New York, 1992; 571–594.
17. Goel, V; Accomazzo, M.A.; DiLeo, A.J.; Meier, P.; Pitt, A.; Pluskal, M. Deadend microfiltration: applications, design, and cost. In *Membrane Handbook*; Ho, W.S., Sirkar, K.K., Eds.; Van Nostrand Reinhold: New York, 1992; 506–570.
18. Li, N.N. Separating Hydrocarbons with Liquid Membranes. US Patent 3,410,794, November 12, 1968.
19. Li, N.N.; Cahn, R.P.; Shrier, A.L. Removal of Organic Compounds by Liquid Membrane. US Patent 3,617,546, November 2, 1971.
20. Jacoby, M. Norman Li Wins Perkin Medal. Chem. Eng. News (March 6, 2000), *78* (10), 60–61.
21. Ho, W.S.; Li, N.N. Emulsion liquid membranes: definitions and theory. In *Membrane*

Handbook; Ho, W.S., Sirkar, K.K., Eds.; Van Nostrand Reinhold: New York, 1992; 597–655.
22. Cahn, R.P.; Li, N.N. Commercial applications of emulsion liquid membranes. In *Separation and Purification Technology*; Li, N.N, Calo, J.M., Eds.; Marcel Dekker: New York, 1992; 195–212.
23. Bartsch, R.A.; Charewicz, W.A.; Kang, S.I.; Walkowiak, W. Proton-coupled transport of alkali metal cations across liquid membranes by ionizable crown ethers. In *Liquid Membranes: Theory and Applications*; Noble, R.D., Way, J.D., Eds.; ACS Symp. Ser. No. 347; American Chemical Society: Washington, D.C., 1987; 86–97.
24. Ho, W.S.; Poddar, T.K. New membrane technology for removal and recovery of chromium from waste waters. Environ. Prog. **2001**, *20* (1), 44–52.
25. Ho, W.S.; Wang, B.; Neumuller, T.E.; Roller, J. Supported liquid membranes for removal and recovery of metals from waste waters and process streams. Environ. Prog. **2001**, *20* (2), 117–122.
26. EPA. *Guidelines for Water Reuse*, EPA/625/R-92/004; Office of Compliance, US Environmental Protection Agency: Washington, Dc, 1992.
27. Byers, W.; Doerr, W.; Krishnan, R.; Peters, D. *How to Implement Industrial Water Reuse*; American Institute of Chemical Engineers: New York, 1995.
28. Zinkus, G.A.; Byers, W.D.; Doerr, W.W. Identify appropriate water reclamation technologies. Chem. Eng. Prog. **1998**, *94* (5), 19–31.
29. U.S. Bureau of Reclamation; Sandia National Laboratories. *Desalination and Water Purification Technology Roadmap—A Report of the Executive Committee*, Desalination & Water Purification Research & Development Program Report #95; 2003.
30. Song, R.Q.; Li, J.C.; Kuehne, M.; Tsai, M.; Li, N. Development of an advanced membrane for water treatment. In *Water Purification and Reuse*. Potsdam, Germany, June 8–13, 2003; Engineering Conferences International: Brooklyn, New york, 2003.
31. http://www.gwrsystem.com (accessed April 2001).
32. Lesile, G.L.; Dunivin, W.R.; Gabillet, P.; Conklin, S.R.; Mills, W.R.; Sudak, R.G. Pilot testing of microfiltration and ultrafiltration upstream of reverse osmosis during reclamation of municipal wastewater. Proceedings of American Desalting Association Biennial Conference, Monterey, California, August 1996.
33. Stephenson, T.; Judd, S.; Jefferson, B.; Brindle, K. *Membrane Bioreactors for Wastewater Treatment*; IWA Publishing: London, England, 2000.
34. EPA. *Profile of the Pulp and Paper Industry*, EPA/310-R-95-015; Office of Compliance, US Environmental Protection Agency: Washington, D.C., 1995.
35. Nuortila-Jokinen, J.; Kuparinen, A.; Nyström, M. Tailoring an economical membrane process for internal purification in the paper industry. Desalination **1998**, *119* (1–3), 11–19.
36. Culkin, B.; Plotkin, A.; Monroe, M. Solve membrane fouling problems with high-shear filtration. Chem. Eng. Prog. **1998**, *94* (1), 29–33.
37. Wagner, J. Membrane technology in wood, pulp and paper industries. In *Membrane Technology in Water and Wastewater Treatment*; Hillis, P., Ed.; The Royal Society of Chemistry: Cambridge, UK, 2000; 233–240.
38. Draxler, J.; Marr, R.; Prötsch, M. Commercial-scale extraction of zinc by emulsion liquid membranes. In *Separation Technology,* Engineering Foundation Conference, Schloss Elmau, Germany, April 27–May 1, 1987; Li, N.N., Strathmann, H., Eds.; American Institute of Chemical Engineers: New York, 1988; 204–214.
39. Zhan, X.; Liu, J.; Fan, Q.; Lian, Q.; Zhang, X.; Lu, T. Industrial application of liquid Mmembrane separation for phenolic wastewater treatment. In *Separation Technology,* Engineering Foundation Conference, Schloss Elmau, Germany, April 27–May 1, 1987; Li, N.N., Strathmann, H., Eds.; American Institute of Chemical Engineers: New York, 1988; 190–203.
40. Treatment of Cyanide-Containing Waste Water from Gold Mine Operation by Liquid Membrane Technology. News release in *Kexue Bao (Newspaper of Science)*, China, October 16, 1987.
41. Jin, M.; Zhang, Y. Study on extraction of gold and cyanide from alkaline cyanide solution by liquid membrane. Proceedings, The 1990 International Congress on Membranes and Membrane Processes, ICOM '90, Chicago, August 20–24, 1990; American Institute of Chemical Engineers: New York, 1990; Vol. 1, 676–678.
42. Bock, J.; Klein, R.R.; Valint, P.L.; Ho, W.S. Liquid membrane extraction of uranium from wet process phosphoric acid. In *Sulfuric/Phosphoric Acid Plant Operations*; American Institute of Chemical Engineers: New York, 1982; 175–183.
43. Abolmaali, B.; Yassine, I.; Capone, P. Water recovery from an aluminum can manufacturing process using spiral wound membrane elements. In *Membrane Technologies for Industrial and Municipal Wastewater Treatment and Reuse*; Water Environment Federation: Alexandria, VA, 2000; 51–56.
44. Cassano, A.; Molinari, R.; Romano, M.; Drioli, E. Treatment of aqueous effluents of the leather industry by membrane processes. J. Membr. Sci. **2001**, *181* (1), 111–126.
45. Weems, J.; Sohns, R.; Bell, C.; Pandya, K. Water reuse. Ultrapure Water **2003**, *20* (4), 25–36.

Wide Band-Gap Electronics Materials

Mark A. Prelas
Krishnendu Saha
Nuclear Science and Engineering Institute, University of Missouri–Columbia, Columbia, Missouri, U.S.A.

INTRODUCTION

A wide band-gap material is a semiconductor that has a band-gap approximately greater than 2 eV. Some of the most prominent wide band-gap materials are GaP (2.26 eV), 3C–SiC (2.36 eV), 6H–SiC (3.0 eV), 4H–SiC (3.23 eV), GaN (3.2 eV), diamond (5.46 eV), AlN (6.2 eV), and BN (6.1–6.4 eV). These materials have applications in high-efficiency optoelectronic devices such as blue and UV light emitting diodes (LEDs) and lasers, as well as high-power, high-temperature, and high-frequency electronic devices. Electronic devices formed in wide band-gap materials operate at high temperatures without suffering from intrinsic conduction effects because of the wide energy band-gap. Some wide band-gap materials emit and detect short-wavelength light, which has applications in blue LEDs and nearly solar blind UV photodetectors. Wide band-gap materials, in general, have a high breakdown voltage, which allows them to withstand a voltage gradient much greater than that of Si or GaAs without undergoing avalanche breakdown. This property makes wide band-gap materials excellent for applications in very high-voltage, high-power devices such as diodes, power transistors, power thyristors, and surge suppressors, as well as high-power microwave devices. Because of their excellent thermal properties, devices made of wide band-gap materials can be placed close together, allowing a high device packing density for integrated circuits. Packing density is dependent on molecular size of the element in the chip and an example of packing density is a few million transistors in an area of 4 mm^2 with spacing between elements of approximately 1.5 μm. Diamond, for example, has the highest known thermal conductivity (five times larger than copper). Wide band-gap materials can operate at very high power levels and still dissipate even large amounts of excess heat. Many wide band-gap materials have a high electron drift velocity, which allows them to operate at high frequencies (RF and microwave). The electron drift velocity is the velocity of electrons in semiconductors under the influence of electric field. Wide band-gap materials have a high resistance to chemical attack, which allows them to be used in corrosive environments. Additionally, wide band-gap materials are resistant to radiation, which makes them ideal for devices that require radiation hardening, such as components for satellites and spacecraft.

BACKGROUND

Diamond

Diamond is discussed in Diamond and Diamond-Like Film Applications.

III–V Materials (Nitrides)

Wide band-gap materials have always been a subject of interest for electronic applications.[1] Light emitting diodes with blue and green light emitting capability have been in the market and ultraviolet and blue laser diodes , high-speed transistors, and ultraviolet photodetector technology have been demonstrated. The III–V materials (nitrides) especially stand out. The properties like high electron mobility, high current carrying capability, high thermal capability, high temperature operation, and high breakdown field are important attributes of III–V materials. Electron mobility and hole mobility are the measure of scattering of electron or holes in semiconductors. The first report of synthesizing a small GaN crystal was made by Johnston and Parsons in 1932.[2] Grimmeiss and Koelmans in 1959 performed luminescence studies.[3] In 1969, Maruska and Tietjen succeeded in growing the first single-crystal GaN on a sapphire substrate by hydride vapor phase epitaxy (HVPE).[4] This work was a major breakthrough because of the difficulty of growing large bulk GaN single crystals and because the available methods relied on heteroepitaxial growth. Heteroepitaxial growth is a method in which a thin layer of single-crystal material is deposited on a single-crystal substrate, the chemical composition of the depositing material being different from that of the substrate. Additionally, it was found that GaN has a direct-transition band structure with band-gap energy of about 3.39 eV, at room temperature. This caused an acceleration in research on GaN as seen in Fig. 1, period A, which lists the number of worldwide

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120025854

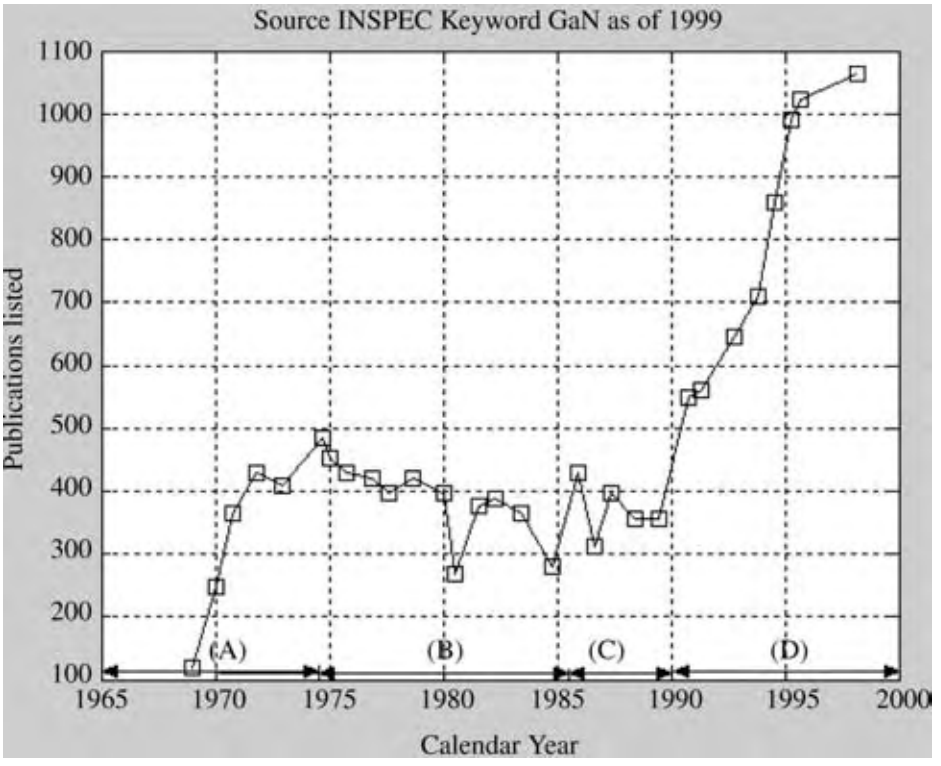


Fig. 1 Number of worldwide publications on nitrides over the years. (From Ref.[1].)

publications related to GaN in any given year. By 1971, Dingle et al. demonstrated optically pumped UV stimulated emission from a GaN crystal at 2 K.[5] The demonstration used Zn as a deep acceptor.

By changing doping of the Zn level, various colors can be produced and blue, green, yellow, or red can be emitted. The first blue LED using a metal–insulator–semiconductor structure was developed by Pankove et al. in 1971.[6] The basic studies of the physical properties of GaN crystals occurred in this period. Ejder reported energy dispersion of the refractive index of GaN in 1971.[7] Then, studies of luminescence including exciton recombination were reported by several researchers.[8,9] Through these studies some basic properties like positions of exciton ground state transition energies of GaN and the lowest band-gap of GaN were clarified. However, the crystalline quality of GaN in those days was not sufficient to characterize the intrinsic properties. Moreover, because of the high n-type background, p-type conduction was very difficult to obtain. N-Type is the lattice where there is an extra electron while p-type is where there is an one electron less in the lattice. Because of this limitation many researchers discontinued Group III nitride research and as a result, the number of publications gradually decreased during this period as seen in Fig. 1, period B. With the introduction of metalorganic vapor phase epitaxy, which is preferred compared to molecular beam epitaxy (MBE), dramatic improvement of crystalline quality of nitrides, was obtained. This caused a revolutionary breakthrough in period C, which resulted in extensive research from period C to period D.

AlN, GaN, InN, and their alloys are all wide band-gap semiconductors. They crystallize into both wurtzite and zincblende polytypes.[10] Wurtzite GaN, AlN, and InN have direct room temperature band-gaps of 3.4, 6.2, and 1.9 eV, respectively. In cubic form, GaN and InN have direct band-gaps while AlN has an indirect energy band-gap. Alloys of GaN with AlN and InN can form materials with a wide range of energy band-gaps. Group III nitrides span a continuous range of direct band-gap energies throughout much of the visible spectrum into the ultraviolet wavelength. Short-wavelength optoelectronic device application is one of the reasons for the increase in research papers after 1990. Group III nitride optoelectronic devices, like LEDs and lasers, can be fabricated for green, blue, and UV wavelengths. The LEDs have applications as elements for full color displays and also as elements for signal and illumination applications. Used as coherent sources they have applications in high-density optical read and write device. In the latter application, because of the fact that the diffraction limited optical storage density increases roughly quadratically as the probe laser wavelength is reduced, nitride based coherent sources at wavelengths down to UV are attracting attention. For optical storage applications these devices have shown storage and retrieval property of a large

number of images and large video files. Currently, optical storage devices use lasers made from InGaAlP heterostructures whose wavelengths are limited to about 550 nm. ZnSe alloys have recently been explored that operate in fringes of green and blue wavelength. These devices have a short lifetime probably owing to stacking faults, which is in the range of about $10^5/cm^2$. The potential short life of ZnSe because of mechanical instability is a significant concern. GaN and/or its alloys can overcome these problems because of longer lifetime and fill the green, blue, and UV regions of the spectrum. During the past several decades, lasers and LEDs have expanded in terms of both range of wavelengths and brightness. Introduction of a bright blue emitter paved the way for full color displays. The low power consumption of LEDs will allow full color display to extend battery life and reduce battery weight.

Both n- and p-type doping is possible in SiC and it has an excellent power amplification device performance. As a result it is a direct competitor of the nitrides in this application. Nitrides form direct band-gap heterostructures allowing the placing of carriers at the interface and carrier confinement. In addition, nitrides can form good ohmic contacts, which are imperative for power devices.

Many wide band-gap semiconductors are produced by epitaxial growth. A brief history of applications of these techniques is given below.[11] The first event of importance was the success of Al-ferov in 1969 in fabricating a continuous-wave laser diode that operated at room temperature by liquid phase epitaxy. For this, he received the Nobel Prize in physics in 2000 with two other scientists. Liquid phase epitaxy, which was proposed by Nelson, makes high-quality thin film semiconductors.[12] In III–V LPE, the compounds are grown from Group III metal solutions on a substrate. This produces highly stochiometric, thin, pure, and perfect films. The next important event in the history of III–V epitaxy is the invention of the high electron mobility transistor (HEMT) by Mimura et al., which was made by MBE.[13] They used the idea of modulation doping that was originally proposed by Dingle et al.[14] The growth technology of high-quality and ultrathin III–V films of about 10 nm thickness was required for the fabrication of HEMT. At that time, such films could be grown only by MBE. Almost at the same time, the technology of metal organic chemical vapor deposition (MOCVD) was developed and used for growing high-quality thin films. Without MBE and MOCVD, neither HEMT nor high-performance laser diodes such as quantum well (QW) and multi-quantum well (MQW) lasers could be made.

Recent successes in growing high-quality GaN and related compounds were achieved by Amano et al. by using a low-temperature buffer layer on sapphire in MOCVD.[15] The improvement in crystal quality realized made it possible to get p-type GaN. The success of p-GaN was a real breakthrough for making blue to ultraviolet lasers and LEDs.

PROPERTIES OF WIDE BAND-GAP MATERIALS

General Properties

The properties like high electron mobility, high current carrying capability, high thermal capability, high temperature operation, and high breakdown field) of wide band-gap materials are superior to silicon. The Keyes figure of merit (KFM) takes into account the power density dissipation for closely packed integrated circuits. High thermal conductivity is an important element for the KFM. The KFM is based on V_{sat} (electron saturation velocity), σ_t (thermal conductivity), and ε_r (dielectric constant). The relative value of the KFM is related to the speed of the transistor in the material as given below:

$$\text{KFM} = \sigma_t (V_{sat}/\varepsilon_r)^{0.5}$$

The higher the value of KFM, the higher will be the speed of the transistor (Table 1).

Table 1 Properties of some wide band-gap semiconductors

Material	Band-gap (eV)	σ_t (300 K) (W/cm)	ε_r	V_{sat} (cm/sec)	KFM ($W\,cm^{-1/2}sec^{-1/2}$)	Ratio to silicon
Si	1.1	1.5	11.8	1.0×10^7	13.8×10^2	1.0
GaN	3.5	1.5	9.5	2.5×10^7	24.3×10^2	1.76
αSiC(6H)	3.0	5.0	10.0	2.0×10^7	70.7×10^2	5.12
βSiC(4H)	3.3	5.0	9.7	2.5×10^7	80.3×10^2	5.8
Diamond	5.47	20.0	5.5	2.7×10^7	444.0×10^2	32.2
BN	6.1	5.7	3.3	3.1×10^7	174.7×10^2	12.7
AlN	6.2–6.4	3.0	9.0	3.0×10^7	54.8×10^2	4.0

(From NSM Archive, Http://www.ioffe.rssi.ru/SVA/NSM/Semicond/.)

The Johnson figure of merit (JFM) is used to compare materials for power microwave applications. Larger values indicate superior performance in microwave power applications. The higher the value, the higher will be the power capability. This figure of merit is the product of the breakdown voltage (V_b) and the electron saturation velocity (V_{sat}).

$$\text{JFM} = (V_b V_{sat})/2\pi$$

Wide band-gap materials have an excellent JFM, about two orders of magnitude higher than silicon, making them excellent candidates for high-power microwave electronics (Table 2).

Overall, the properties of wide band-gap materials have many advantages over Si, GaAs, and related semiconductor materials (Table 3). Wide band-gap materials have a much higher breakdown voltage and a higher thermal conductivity than silicon and related semiconductors. In addition when you compare the electron mobility (Si: 1450 $cm^2/V/sec$, GaAs: 8500 $cm^2/V/sec$) and hole mobility (Si: 450 $cm^2/V/sec$, GaAs: 400 $cm^2/V/sec$) of standard semiconductor materials to that of diamond (electron mobility: 2200 $cm^2/V/sec$, hole mobility: 2000 $cm^2/V/sec$), there are some very distinct advantages that come to light. The large band-gap difference makes these materials suitable for high-temperature electronics, power electronics, and radiation-hardened electronics.

The material properties of some individual wide band-gap materials are discussed below.

SPECIFIC MATERIAL PROPERTIES

GaP

Gallium phosphide is primarily used for the manufacture of red and green diodes. Its band structure is direct. Some properties of wide band-gap materials are given in Table 2.

GaP has a number of potential shallow-level donors (n-type dopant). These include sulfur (substituted for P-activation energy of 0.107 eV), selenium (substituted for P-activation energy of 0.105 eV), tellurium (substituted for P-activation energy of 0.093 eV), lithium (substituted for P-activation energy of 0.091 eV), germanium (substituted for Ga-activation energy of 0.204 eV), silicon (substituted for Ga-activation energy of 0.085 eV), tin (substituted for Ga-activation energy of 0.072 eV), and lithium (substituted for Ga-activation energy of 0.061 eV). Of these potential donors, typical commercial GaP wafers use sulfur (carrier concentration of $1.15 \times 10^{17}/cm^3$ and an electron mobility between 80 and 130 $cm^2/V/sec$) or tellurium (carrier concentration of $1.7 \times 10^{17}/cm^3$ and an electron mobility between 100 and 140 $cm^2/V/sec$). Wafer sizes are 50–80 mm.

GaP also has a number of potential shallow-level acceptors (p-type dopant). These include germanium (substituted for P-activation energy of 0.265 eV), carbon (substituted for P-activation energy of 0.0543 eV), silicon (substituted for P-activation energy of 0.210 eV), beryllium (substituted for Ga-activation energy of 0.0566 eV), cadmium (substituted for Ga-activation energy of 0.1022 eV), magnesium (substituted for Ga-activation energy of 0.0599 eV), and zinc (substituted for Ga-activation energy of 0.0697 eV). Of these potential acceptors, typical commercial GaP wafers use zinc (carrier concentration of $1.2 \times 10^{17}/cm^3$ and a hole mobility between 50 and 80 $cm^2/V/sec$). Wafer sizes are 50–80 mm.

3C–SiC

3C–SiC is the sole cubic polytype among the many SiC polytypes. It has the highest electron mobility of the SiC polytypes. Its band structure is indirect. Most of the commercial applications of SiC are with 4H–SiC and 6HSiC. 3C–SiC has a potential shallow-level donor (n-type dopant) with nitrogen (activation energy of 0.06–0.1 eV).

3C–SiC has several potential shallow-level acceptors (p-type dopant) with aluminum (activation energy of 0.26 eV), gallium (activation energy of 0.344 eV), and boron (activation energy 0.735 eV).

Table 2 Johnson figure of merit (JFM) of some wide band-gap semiconductors

Material	Band-gap (eV)	V_b(V/cm)	V_{sat}(cm/sec)	JFM (V/sec)	Ratio to silicon
Si	1.1	3×10^5	1.0×10^7	3×10^{12}	1.0
GaN	3.5	3×10^7	2.5×10^7	8.0×10^{14}	267
αSiC(6H)	3.0	3×10^7	2.0×10^7	6.0×10^{14}	200
βSiC(4H)	3.3	3×10^7	2.5×10^7	8.0×10^{14}	267
Diamond	5.47	1×10^7–1×10^8	2.7×10^7	2.7×10^{14}–27×10^{14}	90–900
BN	6.1	1.2×10^7	3.1×10^7	3.72×10^{14}	124
AlN	6.2–6.4	2×10^7	3.0×10^7	6.0×10^{14}	200

Table 3 Properties of some wide band-gap materials

Property	GaP	3C–SiC	6H–SiC	4H–SiC	GaN	ZnO	Diamond	AlN	BN
Crystal structure	Zinc blende (cubic)	Zinc blende (cubic)	Wurtzite	Wurtzite	Wurtzite	Wurtzite	Diamond	Wurtzite	Zinc blende
Group of symmetry	T_d^2-$F43m$	T_d^2-$F43m$	C_{6v}^4-$P6_3mc$	C_{6v}^4 -$P6_3mc$	C_{6v}^4 -$P6_3mc$	C_{6v}^4 -$P6_3mc$	O_h^7 -$Fd3m$	C_{6v}^4 -$P6_3mc$	T_d^2 -$F43m$
Number of atoms per cubic centimeter	4.9×10^{22}				8.9×10^{22}		1.7×10^{23}	9.6×10^{22}	
Debye temperature (K)	445	1200	1200	1300	600		1860	1150	1700
Density (g/cm^3)	4.14	3.166	3.21		6.15	5.642	3.515	3.255	3.48
Dielectric constant (static)	11.1	9.72	9.66	9.66	8.9	8.75	5.7	9.14	7.1
Dielectric const (high frequency)	9.11	6.52	6.52	6.52	5.35	3.75		4.84	4.5
Effective longitudinal electron mass (m_l)	$1.12m_0$	$0.68m_0$	$0.20m_0$	$0.29m_0$	$0.20m_0$		$1.40m_0$	$0.4m_0$	$0.35m_0$
Effective transverse electron mass (m_t)	$0.22m_0$	$0.25m_0$	$0.42m_0$	$0.42m_0$	$0.20m_0$		$0.36m_0$		$0.24m_0$
Effective heavy hole mass (m_h)	$0.79m_0$				$1.4m_0$		$2.12m_0$	$3.53m_0$	$0.37m_0$
Effective light hole mass (m_{lp})	$0.14m_0$				$0.3m_0$		$0.70m_0$	$3.53m_0$	$0.150m_0$
Electron affinity (eV)	3.8				4.1		−0.070	0.6	4.5
Lattice constant (Å)	5.4505	4.3596	$A = 3.0730$, $b = 10.053$	$a = 3.0730$, $b = 10.053$	$a = 3.189$, $c = 5.186$	$a = 4.75$, $c = 2.92$	3.567	$A = 3.11$, $c = 4.98$	3.6157
Optical phonon energy (meV)	51	102.8	104.2	104.2	91.2		160	99.2	130
Band-gap (eV)	2.26	2.26	3.0	3.3	3.5	3.37	5.47	6.2	6.2–6.4
Breakdown voltage (MV/cm)	~1.1	~2	~3	~3	~3		1–10	1.2	2
Electron mobility ($cm^2/V/sec$)	250	1000	380	800	300	80	2200	300	200
Hole mobility ($cm^2/V/sec$)	150	50	40	140	350		2000	14	500
Melting point (°C)	1457	2830	2830	2830	2500	1977	4373	3273	2973
Thermal conductivity (W/cm/°C)	1.1	4.9	4.9	4.9	1.3	0.54	20	2.85	7.4
Hardness Mohs scale	5	9.2	9.2	9.2		4	10		9.5

6H–SiC

6H–SiC is an indirect band-gap material that is produced commercially. 6H–SiC along with 4H–SiC is used for short-wavelength optoelectronic, high-temperature, radiation-resistant, and high-power/high-frequency electronic devices. High-quality SiC is grown commercially by companies, such as CREE Research Inc., of Raleigh, NC, by bulk growth methods (4H and 6H structures) as well as by chemical vapor deposition (CVD). A p–n structure in SiC can be achieved by various methods. Some of these methods are sublimation epitaxy, container-free liquid phase epitaxy, CVD, ion implantation of aluminum, and boron diffusion. Many of the problems, which plagued SiC in the past, have been overcome. The micropipe defect density has been reduced to less than one per square centimeter. The crystal size is now being produced in 3 in. production wafers and large by CREE.

6H–SiC has potential shallow-level donors (n-type dopant) with nitrogen (activation energy of 0.085–0.125 eV) and phosphorous (activation energy of 0.085 eV). 6H–SiC has several potential shallow-level acceptors (p-type dopant) with aluminum (activation energy of 0.239 eV), gallium (activation energy of 0.317 eV), beryllium (activation energy of 0.320 eV), and boron (activation energy of 0.31–0.38 eV).

4H–SiC

4H–SiC is an indirect band-gap material that is produced commercially. As stated previously, 4H–SiC along with 6H–SiC are used for short-wavelength optoelectronic, high-temperature, radiation-resistant, and high-power/high-frequency electronic devices.

4H–SiC has potential shallow-level donors (n-type dopant) with nitrogen (activation energy of 0.059–0.102 eV), titanium (activation energy of 0.13 eV), and chromium (activation energy of 0.15–0.18 eV). 4H–SiC has several potential shallow-level acceptors (p-type dopant) with aluminum (activation energy of 0.19 eV), gallium (activation energy of 0.267 eV), and boron (activation energy of 0.647 eV).

GaN

Gallium nitride is a direct band-gap material that is primarily used in green, blue, and UV LEDs and lasers, white-light sources, high-power RF and microwave sources, and high-temperature devices. Its band structure is direct. Some of the properties of gallium nitride are given in Table 3. GaN has been the extensively studied material for optoelectronic applications among all III–V nitrides. Blue LEDs and lasers made of GAN are obtained by overcoming constraints like doping and heteroepitaxial growth.[16,17] Gallium nitride (GaN) substrates are being grown by MOCVD, MBE, and HVPE. GaN is typically grown on sapphire (Al_2O_3), 6H–SiC, and ZnO. Most as-grown GaN (and InN) films exhibited high n-type conductivity because of native defects but p-type conductivity cannot be obtained.[16] p-Type GaN was achieved by doping with Mg, and GaN p–n homojunction and blue light emitting devices were demonstrated.[18,19] GaN LEDs are now being made commercially. The alloy of $Al_xGa_{1-x}N$ has also been developed for blue to UV emitters.[18,19] However, only films with a small amount of Al ($x \sim 0.1$ for p-type and $x < 0.4$) for n-type can be doped successfully.[18–20]

GaN has a number of potential shallow-level donors (n-type dopant). An N vacancy is believed to be the dominant donor. The activation energy of the N vacancy is still under debate. Of these potential donors, typical commercial GaN wafers are naturally a donor because of the N vacancy. Wafer sizes are about 50 mm.

Both magnesium and beryllium have been used to make p-type GaN; however, there is an unintentional incorporation of oxygen and silicon in the GaN growth process.[16,18,19]. p-Type materials are possible with beryllium or vanadium (in the Ga position with approximate activation energy of 0.236 eV), magnesium or vanadium (in the Ga position with approximate activation energy of 0.236 eV), and zinc or vanadium (in the Ga position with approximate activation energy of 0.232 eV). There are a number of deep-level acceptors.

ZnO

ZnO is a direct band-gap material that has gained interest for optoelectronic applications. It is an II–VI semiconductor, which gives it advantages in performance because II–VI semiconductors are not easily degraded through defects. ZnO is naturally an n-type material.[21] Its primary uses are for the polycrystalline form for piezoelectric transducers, varistors, phosphors, and transparent conducting films. Bulk crystals of ZnO have been fabricated by various methods including vapor phase growth, hydrothermal growth, and melt growth. Additionally, epitaxial growth of ZnO single-crystal films has been demonstrated by various techniques such as pulsed laser deposition, MBE, MOCVD, and HVPE.

Acceptor levels in ZnO can be created with hydrogen and potentially sulfur and nitrogen.[22,23] A p–n junction in ZnO has been formed.[24] Various light emitting structures have been formed using ZnO including LEDs, and a UV nanowire laser.[25,26]

Diamond

Diamond is an indirect band-gap material that is used in grinding, polishing, cutting, wear resistance, tribology, acoustics, optics, thermal management, and electronic applications (see section on Diamond and Diamond-Like Materials). It has been used in electronic applications as a heat sink for large-scale integrated circuits. Recently, an 81 GHz high-frequency diamond device was made by Nippon Telegraph and Telephone Corp. (NTT).[27] This promises to make possible amplification *e* in the millimeter-wave band from 30 to 300 GHz. In addition, efforts by the Japanese government to initiate a joint research project with industry in fiscal year 2003 to develop diamond-based semiconductors indicates the progress that has been made in diamond electronics over the last decade.[28] Diamond, because of its superior properties is envisioned as an advanced chip technology that could one day replace silicon as the base for superfast, high-voltage semiconductors. There are two keys to diamond technology, the most challenging is making electronic grade high-quality synthetic diamond single crystals. NTT will focus its technology on a proven means of fabricating millimeter electronic single crystals. Other breakthroughs may lead to a source of large single crystalline diamond. A company called Genesis is producing large high-quality HPHT diamond crystals that may lead to a source of diamond wafers for the electronic industry. A company called Apollo is producing gem quality single-crystal diamond that may be economical for the diamond electronics business.[29] The second is in the formation of complex device structures. A fundamental device, a p–n junction, was fabricated in 1997 using a method called field enhanced diffusion with optical activation the details of which are given in Ref.[30]. Field enhanced diffusion with optical activation lends itself well to the fabrication of complex integrated circuits.[31]

Donors for diamond are challenging. Lithium has been identified as a donor with an activation energy of 0.2 eV (searching did not reveal) using field enhanced diffusion with optical activation.[32] However, the lithium concentration was limited to about 5×10^{17} atoms/cm^3. Other methods of introducing lithium into the diamond, such as in situ doping or ion implantation, did not produce n-type material owing to the defects introduced by the method. When dopant atoms are introduced into the semiconductor layers during its growth, most commonly during epitaxial growth of semiconductor layer, it is called in situ doping. Field enhanced diffusion with optical activation is a low-energy method that maintains the quality of the crystal and thus is able to limit the formation of defects. Another method of creating n-type conductivity has been reported. In this process, deuterating boron-doped diamond in a plasma at about 550°C converts the material from p-type to n-type with conductivities as high as 2 S/cm at room temperature.[33] Nitrogen is also a donor but it is deep level with activation energies of 1.7 and 4 eV, which may be due to substitutional impurity, simple aggregates, or platelets. Phosphorus is a donor with reported activation energies between 0.84 and 1.16 eV. Sulfur is a donor that has received a great deal of attention recently. Dr. Matt West demonstrated that diamond doped by the field enhanced diffusion with optical activation method has activation energy of 0.9 eV.[34]

Boron is an acceptor in diamond with an activation energy of 0.37 eV. Natural type II b diamond is an example of a boron doped p-type diamond. There are many commercial sources of synthetic boron doped-diamond. Polycrystalline boron doped diamond films, made by CVD, have been made for electrode applications. High-pressure high-temperature diamonds with boron doping are also available.[35]

AlN

Aluminum nitride is direct band-gap material that has a very wide band-gap (6.2 eV). Also, it has a high thermal conductivity, high electrical resistivity, high acoustic velocity, high thermal stability, and high chemical resistance, and radiation stability. These properties make AlN suitable for UV optical devices, surface acoustic wave (SAW) devices, electrical insulators, or passive layers in microelectronics. Such a device can operate in a harsh environment with high temperatures and/or radiation. However, it is very difficult to dope AlN with impurities to make it to n- or p-type semiconductors. Also as grown AlN films do not show any n- or p-type characteristics. Although both n- and p-type doping of aluminum nitride have been reported as far back as 1967, almost no recent paper can be found that verifies these results.[36] Field enhanced diffusion with optical activation has been used for doping AlN and some promising results with magnesium as the p-type dopant and silicon as the n-type dopant.[37] Additional potential donors include C (Al), Ge (Al), and Se (N) while additional potential acceptors include Be (Al), C (N), Ca (Al), and Hg (Al).

BN

Despite the excellent properties of cubic boron nitride, it is a very difficult crystal to grow. Thus far the crystal size has been limited to micrometer-type size. Little is known about the ultimate potential of cubic boron nitride in electronics. Potential donors include Si with activation energy of 0.24 eV, C with activation energy of 0.28 eV, and sulfur with activation energy of

0.41 eV.[38] A potential acceptor is Be with activation energy of 0.19 eV.[39]

APPLICATIONS

There are many industries that utilize wide band-gap materials. The primary applications of wide band-gap materials are identified below.

Grinding and Polishing

Grinding and polishing is one of the oldest applications for wide band-gap materials primarily owing to the property of hardness that some of these materials possess (e.g., diamond). SiC and cubic boron nitride, in addition to diamond, have found a commercial market in grinding and polishing, primarily for ferromagnetic materials with high carbon solubility.[39,40]

Tool and Die

Because of the hardness of band-gap materials such as diamond, SiC, and boron nitride, wide band-gap materials have found applications in the tool and die industry.[40,41].Wide band-gap materials have been successfully used for coating field emission tips in SEM with results showing reduction in turn-on voltage by 100 V and more uniform emission during low-voltage operation.[42]

Electronics

One of the major markets for wide band-gap materials is in electronics. Specifically, they are suitable for and have been used for heat sinks (diamond), short wavelength optoelectronic devices (GaP, GaN, SiC), high-temperature electronics (SiC, GaN), radiation resistant devices, and high-power/high-frequency electronic devices (diamond, GaN, SiC).[40,43–45] Recent research showed that Mn-doped GaN can be used for spintronic applications.[46] Atomically flat technology developed by NASA for SiC and GaN WBG material can introduce a new dimension of application for WBG materials.

Heat Sinks

Because of diamond's exceptional heat conductivity (about five times that of copper), it has been used as a heat sink for large-scale integrated circuits.[40]

Insulators

Because of the high breakdown voltage, wide band-gap material can and have been used as electrical insulators. There has been a great deal of interest in pressed nanophase SiC and diamond powders as an electrical insulator for capacitors.

Semiconductors

All of the wide band-gap materials have potential applications as semiconductors, ranging from high-temperature, high-power, and high-radiation resistance applications to high-speed integrated circuits to high-frequency devices. Recent research has shown that wide band-gap materials like ilmenite-hematite (IH) can be used for low-voltage varistor application for space research.[47] Semiconductor technology based on wide band-gap material composites has been widely used for MEMS devices.[48]

Optoelectronic Devices

Wide band-gap semiconductors have been used for green, blue, and UV LEDs and lasers, and white-light sources.

CONCLUSIONS

As the wide band-gap electronic field matures, a wide variety of optoelectronic devices, high-temperature devices (for automotives, combustion chambers, smoke stacks, furnaces, etc.: 150°C to >600°C), radiation hardened devices (for satellites, nuclear power plants, etc.), high-power devices (high-voltage switching, power transistors, etc.), high-speed integrated circuits (LSI, CPU, etc.), and high-frequency devices (microwave amplification, communication, etc.) will be developed. Wide band-gap materials are positioned well to play an important role in the vibrant electronics industry over the next several decades.

REFERENCES

1. Isamu, A. Renaissance and progress in nitride semiconductors—my personal hisory of nitride research. Materials Research Society Symposium Proceedings, 2001; Vol. 639.
2. Johnson, W.C.; Parsons, J.B. Nitrogen compounds of gallium. I, II. J. Phys. Chem. **1932**, *36*, 2588.

3. Grimmeiss, H.G.; Koelmans, H. Emission near the absorption edge and other emission effects of GaN. Zeitschrift fuer Naturforschung **1959**, *14a*, 264–271.
4. Maruska, H.P.; Tietjen, J.J. The preparation and properties of vapor-deposited single-crystal-line GaN. Appl. Phys. Lett. **1969**, *15*, 327.
5. Dingle, R.; Shaklee, K.L.; Leheny, R.F.; Zetterstrom, R.B. Stimulated emission and laser action in gallium nitride. Appl. Phys. Lett. **1971**, *19*, 5.
6. Pankove, J.I.; Miller, E.A.; Richman, D.; Berkeyheiser, J.E. Electroluminescence in GaN. J. Lumin. **1971**, *4*, 63.
7. Ejder, E. Refractive index of gallium nitride. Physica Status Solidi A: Applied Research **1971**, *6* (2), 445–448.
8. Monemar, B. Fundamental energy gap of GaN from photoluminescence excitation spectra. Phys. Rev. **1974**, *B10*, 676.
9. Jacob, G.; Boulou, M.; Bois, D. GaN electroluminescent devices: Preparation and studies. J. Lumin. **1978**, *17*, 263.
10. Mohammad, S.N.; Morkoc, H. Progress and prospects of Group III nitride semicondeuctors. Prog. Quant. Electr. **1996**, *20* (5/6), 361–325.
11. Nishinaga, T.; Naritsuka, S. Epitaxial Growth of III–V Compounds. Department of Material Science and Engineering, Meijo University, Japan, 2003; 1–501.
12. Kressel, H.; Dunse, J.U.; Nelson, H.; Hawrylo, F.Z. Luminescence in silicon-doped gallium arsenide grown by liquid-phase epitaxy. J. Appl. Phys. **1968**, *39* (4), 2006–2011.
13. Mimura, T.; Hiyamizu, S.; Fujii, T.; Nanbu, K. A new doped GaAs/n-Al*x*Ga1-*x* as heterojunctions. Jpn. J. Appl. Phys. **1980**, *19L*, 225.
14. Dingle, R.; Stormer, H.L.; Gossard, A.C.; Wiegmann, W. Electron mobilities in modulation-doped semiconductor heterojunction superlattices. Appl. Phys. Lett. **1978**, *33*, 665.
15. Amano, H.; Sawaki, N.; Akasaki, I.; Toyoda, Y. Metalorganic vapor phase epitaxial growth of a high quality GaN film using an AlN buffer layer. Appl. Phys. Lett. **1986**, *48*, 353.
16. Strite, S.; Morkoc, H. GaN, AlN, and InN: a review. J. Vac. Sci. Technol. B **1992**, *10* (4), 1237–1266.
17. Davis, R.F. III-V nitrides for electronic and optoelectronic applications. Proc. IEEE **1991**, *79* (5), 702–712.
18. Akasaki, I.; Amano, H. Perspective of the UV/blue light emitting devices based on GaN and related compounds. Optoelectron. Devices Technol. **1992**, *7* (1), 49–56.
19. Tanaka, T.; Watanabe, A.; Amano, H.; Kobayashi, Y.; Akasaki, I.; Yamazaki, S.; Koike, M. *p*-Type conduction in Mg-doped GaN and $Al_{0.08}Ga_{0.92}N$ grown by metalorganic vapor phase epitaxy. Appl. Phys. Lett. **1994**, *65* (5), 593–594.
20. Yoshida, S.; Misawa, S.; Gonda, S. Properties of Al/*x*/Ga/1-*x*/N films prepared by reactive molecular beam epitaxy. J. Appl. Phys. **1982**, *53* (10), 6844–6848.
21. Jayaraj, M.K; Aldrin, A.; Manoj, R. Transparent conducting zinc oxide thin film prepared by off-axis rf magnetron sputtering. Bull. Mater. Sci. **2002**, *25* (3), 227–230.
22. Chris, G.; Van de Walle. Hydrogen as a cause of doping in zinc oxide. Phys. Rev. Lett. **2000**, *85* (5), 1012–1015.
23. Cruz-Vázquez, C.; Rocha-Alonzo, F.; Burruel-Ibarra, S.E.; Inoue, M.; Bernal, R. Fabrication and characterization of sulfur doped zinc oxide thin films. Superficies y Vacío 13, Dec 2001; Sociedad Mexicana de Ciencia de Superficies y de Vacío, 2001; 88–91.
24. Ohta, H.; Kawamura, K.; Orita, M.; Hirano, M.; Sarukura, N.; Hosono, H. Current injection emission from a transparent *p*-*n* junction composed of *p*-$SrCu_2O_2$/*n*-ZnO. Appl. Phys. Lett. **2000**, *77* (4), 475–477.
25. Look, D.C. Recent advances in ZnO materials and devices. Mater. Sci. Eng. **2001**, *B80*, 383–387.
26. Huang, M.H.; Mao, S.; Feick, H.; Yan, H.; Wu, Y.; Kind, H.; Weber, E.; Russo, R.; Yang, P. Room-temperature ultraviolet nanowire nanolasers. Science **2001**, *292*, 1897–1899.
27. Hara, Y. NTT verifies diamond semiconductor operation at 81 GHz. EE Times **2003**, Aug 26.
28. Miyake, K. Japanese government to launch diamond chip project. Info World **2002**, Dec 27.
29. http://www.wired.com/wired/archive/11.09/diamond_pr.html.
30. Popovici, G.; Melnikov, A.; Varichenko, V.V.; Sung, T.; Prelas, M.A.; Wilson, R.G.; Loyalka, S.K. Diamond ultraviolet photovoltaic cell obtained by lithium and boron doping. J. Appl. Phys. **1997**, *81* (5), 2429–2431.
31. Popovici, G.; Sung, T.; Prelas, M. Forced diffusion in diamond: a review. J. Chem. Vap. Deposition **1994**, *3*, 115–132.
32. Sung, T. Doping Diamond by Forced Diffusion. Ph.D. thesis, University of Missouri, Columbia, May, 1996.
33. High-conductivity n-type diamond demonstrated, photonics spectra. Presstime Bull. **2003**, Aug.
34. Matthew, W. Diffusion of Sulfur into Natural Diamond: Characterization and Applications in Radiation Detection. University of Missouri–Columbia, Dec 1999.
35. Ekimov, E.A.; Sidorov, V.A.; Bauer, E.D.; Mel′nik, N.N.; Curro, N.J.; Thompson, J.D.; Stishov, S.M.

Superconductivity in Diamond; Vereshchagin Institute for High Pressure Physics, Russian Academy of Sciences: Troitsk, Russia. Los Alamos National Laboratory, Condensed Matter, 2004; 1–13.
36. Chu, T.L.; Ing, D.W.; Noreika, A.J. Epitaxial growth of aluminum nitride. Solid-State Electron. **1967**, *10*, 1023.
37. Prelas, M.; Ghosh, T.; Tompson, R. *Direct Conversion of Radioisotope Energy to Electricity*. Final Report for DOE contract, DE FG07-001D13927, August 2003.
38. http://www.ioffe.rssi.ru/SVA/NSM/Semicond/Semicond/ BN/bandstr.html#Donors.
39. Subramanian, K.; Shanbhag, V.R. Abrasive applications of diamond. In *Handbook of Industrial Diamonds and Diamond Films*; Marcel Dekker: 1998; 1023–1042.
40. Prelas, M.; Popovici, G.; Bigelow, K. *Handbook of Industrial Diamond and Diamond Films*; Marcel Dekker: 1998; 1023–1042.
41. Hay, R.A.; Galimberri, J.M. Cutting and wear applications. In *Handbook of Industrial Diamonds and Diamond Films*; Marcel Dekker: 1998; 1135–1147.
42. Graf, T.; Goennenwein, S.T.B.; Brandt, M.S. Prospects for carrier-mediated ferromagnetism in GaN. Walter Schottky Institut, Technische Universitaet Muenchen: Garching, Germany. Phys. Status Solidi B: Basic Res. **2003**, *239* (2), 277–290.
43. Dreifus, D.L.; Fox, B.A. Active devices. In *Handbook of Industrial Diamonds and Diamond Films*; Marcel Dekker: 1998; 1043–1072.
44. Eden, R.C. Applications of diamond in computers. In *Handbook of Industrial Diamonds and Diamond Films*; Marcel Dekker: 1998; 1073–1102.
45. Brandes, G.R. Diamond vacuum electronics. In *Handbook of Industrial Diamonds and Diamond Films*; Marcel Dekker: 1998; 1103–1127.
46. Hajra, M.; Chubun, N.N.; Chakhovskoi, A.G.; Hunt, C.E.; Liu, K.; Murali, A.; Risbud, S.H.; Tyler, T.; Zhirnov, V. Field emission characterization of silicon tip arrays coated with GaN and diamond nanoparticle clusters. Electrical and Computer Engineering Department, University of California, Davis, CA. J. Vacuum Sci. Technol. B: Microelectron. Nanometer Struct.—Process. Meas. Phenom. **2003**, *21* (1), 458–463.
47. Padmini, P.; Pulikkathara, M.; Wilkins, R.; Pandey, R.K. Neutron radiation effects on the nonlinear current–voltage characteristics of ilmenite-hematite ceramics. Department of Electrical and Computer Engineering, The University of Alabama, Tuscaloosa, AL. Appl. Phys. Lett. **2003**, *82* (4), 586–588.
48. Luchinin, V.V.; Korlyakov, A.V.; Vasilev, A.A.; Jandjgava, G.I.; Prosorov, S.V.; Solomatin, A.K.; Sorokin, A.V.; Kucherkov, S.G.; Severov, L.A.; Ponomarev, V.K. SiC-AlN-composition-based MEMS. Proceedings of SPIE—The International Society for Optical Engineering, 1999; Microtechnology Center, St. Petersburg State Electrotechnical University: St. Petersburg, Russia. Indo-Russian Workshop on Micromechanical Systems); 141–145.

Zeolite Membranes

Yushan Yan
Zijian Li
Shuang Li
Christopher Lew
Department of Chemical and Environmental Engineering, University of California at Riverside, Riverside, California, U.S.A.

INTRODUCTION

Zeolites and zeolite-type materials (hereafter referred to as zeolites for simplicity) are a class of crystalline oxides with uniform channels and cages ranging from 0.3 to 2 nm. Structurally, a zeolite consists of 3-dimensionally linked tetrahedra and each tetrahedron has one tetrahedrally coordinated atom (hereafter referred to as T-atom) (e.g., Si, Al, P, B, Ga, Ge, Zn, Be, etc.) at the center and four oxygen atoms at the corners. Each corner oxygen atom is shared by two neighboring tetrahedra. Depending on the T-atom, the zeolite framework can be charged or neutral. When the framework is charged, balancing ions are needed, and this is the origin of the ion exchange capability of zeolites. According to their structural symmetry and topology, zeolites are classified into different framework types. Each framework type is assigned a unique three-letter code (e.g., MFI) by the International Zeolite Association (IZA). These structure codes are generally from the zeolite names (e.g., MFI from Mobil Five) and do not follow any sort of established naming procedure. More than 150 framework types with numerous compositional variations have been verified and approved by the IZA.

Zeolite materials are used commercially as shape/size selective catalysts in the petrochemical and petroleum refining industry, and as molecular sieving separation media for gases and hydrocarbons. For both applications, zeolites are used in powder composite form such as pellets and granules. In this entry, we focus on zeolite membranes. We define zeolite membranes as a continuous phase of zeolite-based materials (pure zeolite or composite) that separate two spaces. Zeolite membranes are generally uniform thin films attached to a porous or a nonporous substrate. They can also be self-standing without a substrate. Note that we have included zeolite films and layers on nonporous substrate in this entry because we believe many of the synthesis strategies and applications reported for those nonporous substrates are easily transferred to a porous substrate to prepare a zeolite membrane.

In this entry, we briefly discuss the types of the zeolite membranes and then focus on new applications that have been demonstrated recently. New developments are also included and analyzed, which are followed by some concluding remarks and future directions.

TYPES OF ZEOLITE MEMBRANES

Polycrystalline Zeolite Membranes

Polycrystalline zeolite membranes consist of intergrown zeolite crystals with no apparent cracks or pinholes[1] (Fig. 1A). These films are composed of only zeolite (i.e., there are no non-zeolite components such as amorphous silica or polymer). They are normally supported on a substrate although free-standing films have also been synthesized.[2] Membranes can be prepared on different substrates such as silicon wafer, quartz, porous alumina, carbon, glass, stainless steel (SS), gold, etc. Polycrystalline films are primarily prepared by hydrothermal synthesis methods including in situ crystallization,[1] seeded growth,[3] and vapor transport,[4] and have potential use in all of the applications discussed in this entry.

Zeolite Matrix Composite Membrane

A zeolite matrix composite membrane is defined as a membrane in which zeolite crystals are imbedded in a solid matrix that is either inorganic (e.g., silica and carbon)[5–8] or organic (e.g., polymer)[9,10] (Fig. 1B). Inorganic matrix membranes are often supported, while organic polymer matrices are normally free-standing. The selection between an inorganic and an organic phase as the imbedding matrix strongly depends on the intended application. For example, an inorganic matrix is ideal for applications where thermal stability is required,[5] whereas an organic polymer matrix is the better choice when flexibility is desired.[9,10] These composite films are usually

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120025023

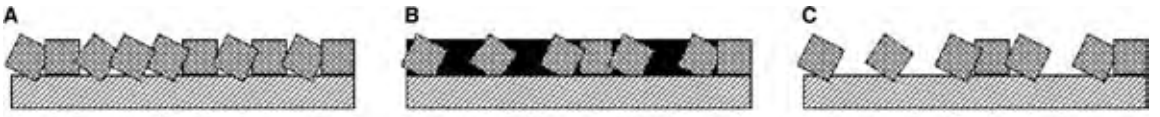


Fig. 1 Schematic of the three types of zeolite membranes: (A) a polycrystalline zeolite membrane; (B) a zeolite matrix composite membrane; and, (C) a zeolite crystal layer.

prepared by wash-coating, sol-gel processing, and polymer film casting.

Zeolite Crystal Layer

Another form of zeolite membranes is a zeolite crystal layer that consists of isolated crystals deposited on a solid substrate[11] (Fig. 1C). The substrate can be a variety of materials such as metal, ceramic, or silicon wafer. Crystal layers have to be supported. There has been exciting fundamental research carried out in this area, however, demonstrated applications have been limited to sensors. The organic linker approach appears very promising for the preparation of these types of membranes.[12]

APPLICATIONS OF ZEOLITE MEMBRANES

Zeolite membranes have been demonstrated for many applications. Applications such as separation membranes, membrane reactors, adsorption, and catalysis have been covered in several reviews.[13–16] In this entry, we focus on new applications including sensors, low-dielectric constant (low-k) films, corrosion resistant coatings, hydrophilic coatings, heat pumps, and thermoelectrics.

Sensors

Chemical sensors are important for industrial process control and environmental monitoring. They also find critical use in medical and defense applications. Two of the critical performance parameters of a sensor are selectivity and sensitivity. When integrated with an appropriate sensor platform (e.g., quartz crystal microbalance or electrochemical quartz crystal microbalance), a zeolite film can improve the selectivity and sensitivity of the sensor due to its ability to selectively adsorb a component out of a mixture. A large number of publications have appeared in this area, and an in-depth review on zeolite films for chemical sensor applications up to 1996 was provided as a section in a recent review by Bein.[11] Here, we highlight a recent development of selective chemical sensor based on continuous oriented zeolite films.

To achieve selective adsorption and quick response, the zeolite film is required to be continuous (i.e., free of non-zeolitic pores) and thin, and have proper orientation. For facile mass transport, a monocrystal film is also preferred because there are no grain boundaries in the mass transport direction, which have proved to decrease diffusion in zeolite by orders of magnitude. Strong adhesion between the zeolite film and the sensor surface is also critical for the durability of the sensor. Recently, a *b*-oriented continuous pure-silica-zeolite (PSZ) MFI monolayer film was successfully synthesized on SS and silicon wafers using a very facile direct in situ crystallization method.[17–19] As-synthesized *b*-oriented films have good continuity and thin monocrystal layer thickness. The thin film thickness and preferred crystal orientation can reach a fast adsorption and desorption equilibrium, which can thus increase the stability of the sensor. In addition, the superior adhesion of this film to the substrate ensures high durability. All these features make this type of film a better candidate as sensor coating to achieve higher selectivity, sensitivity, stability, and durability.[11] The molecular sieving property of this continuous *b*-oriented PSZ MFI monocrystal film was demonstrated in a zeolite modified electrode (ZME) configuration using redox molecules of different sizes (Fig. 2).[20] $Ru(NH_3)_6^{3+}$ (diameter $\sim$5.5 Å) was able to traverse the film, while $Co(phen)_3^{2+}$ (diameter $\sim$13.0 Å) was completely excluded. This film has excellent adhesion to the electrode and is stable in strong acidic conditions.

Low-k Dielectrics

Low-k (k for dielectric constant) dielectric materials have been identified as one of the most difficult challenges for interconnects in future generation integrated circuits (ICs). Many materials have been proposed, studied, or are under commercial development as potential candidates for low-k dielectrics. Among these materials, two major classes are dense organic polymers and porous inorganic based materials. It has been shown that some dense, organic polymers could easily have a k value between 2 and 3, but there are concerns about their low thermal stability and low heat conductivity. Also, due to their low mechanical strength, polymeric materials may have potential problems with the chemical and mechanical polishing (CMP) process.

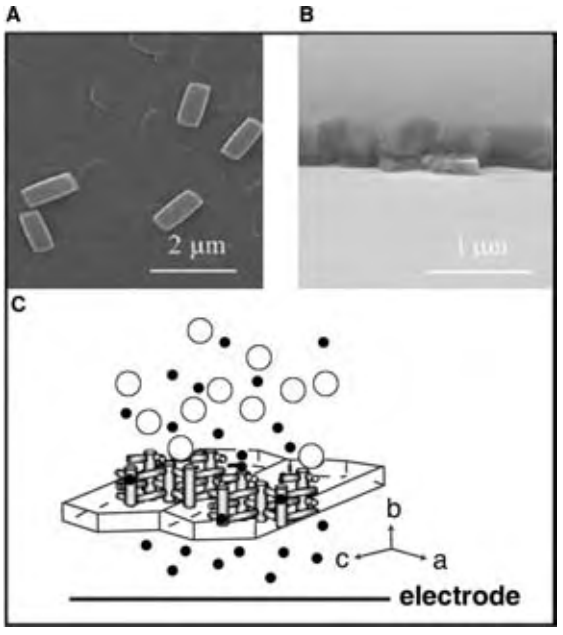


Fig. 2 (A) SEM top view of *b*-oriented pure-silica-zeolite (PSZ) MFI monocrystal-thick film; (B) cross-sectional view of the sample in (A) after slight polishing; (C) schematic illustration of molecular sieving in *b*-oriented PSZ MFI monocrystal-thick film. A distance was intentionally kept between the MFI film and the electrode so that the sieving behavior can be illustrated more clearly. (From Ref.[20].)

Current dielectric material is dense silica that has a k value of 4. Porous silica has a lower k because of the incorporation of air. Sol-gel silica and mesoporous silica have been studied as possible low-k materials. However, these materials have drawbacks such as low mechanical strength and low heat conductivity. Sol-gel silica also has randomly occurring large pores that can cause electrical breakdown. High mechanical strength is needed for the new low-k materials to be compatible with the CMP process.

Very recently, zeolites have been demonstrated to be a promising low-k material (Fig. 3). Zeolites have very uniform pores and their pore size falls into the micropore range (<2 nm), which is significantly smaller than the IC feature size. Therefore, the problem of electrical breakdown should be mitigated. Zeolites have a high heat conductivity and mechanical strength due to their dense crystalline structure.[5,21–25] Pure-silica-zeolite MFI membranes were first developed by in situ crystallization. These films had a k value of 2.7 and were highly hydrophobic. The elastic modulus was 30–40 GPa by nanoindentation using a membrane of 0.43 μm thick. This value is much higher than sol-gel silica or mesoporous silica films. Note that a modulus of 6 GPa is usually considered to be a threshold value for industrially viable low-k dielectrics.

Although the membranes formed by in situ crystallization are mechanically strong, the k value is still high. Also, the solution phase deposition is of concern to the semiconductor industry. Another method was recently developed, which is a simple spin-on process of zeolite nanoparticle suspension. The spin-on process is simple and preferred by the semiconductor manufacturers, and therefore represents a significant step toward the integration of zeolite materials into microelectronic devices. More recently, k values of PSZ MFI films can be lowered to 1.8 by incorporating extra mesopores from organic porogens.[22] The k value of the spin-on PSZ MFI films could be further reduced to 1.6 by using PSZ nanoparticle suspensions with high crystallinity.[23]

PSZs with low framework density (FD) and small pore opening windows are preferred for low-k applications.[25] By having lower FDs, the same k can be achieved with higher microporosity, which translates into better mechanical strength. We recently extended PSZ low-k materials from MFI (FD = 18.4) to MEL (FD = 17.4). We prepared MEL nanoparticle suspensions with high crystallinity (relative crystallinity ~70%). The k values obtained from these suspensions can be as low as 1.5. Organic-functionalized PSZ MFI low-k films are prepared.[25] With organic groups substituting hydroxyl groups, the films are less hydrophilic.

Corrosion-Resistant Coatings

Aluminum alloys are widely used in the aerospace industry because they are light and mechanically

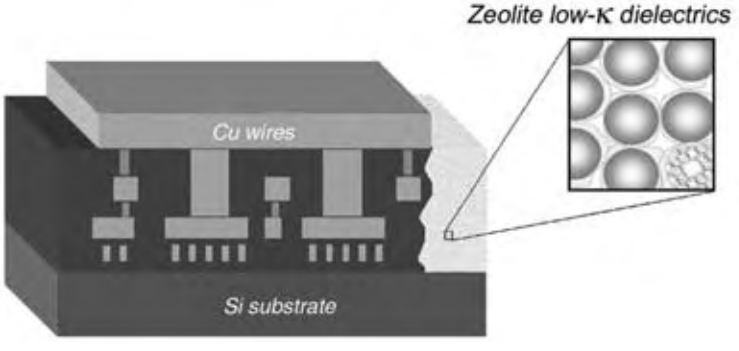


Fig. 3 Schematic illustration of spin-coated zeolite low-k membranes from a nanoparticle suspension as insulator in the interconnect of a microprocessor. *(View this art in color at www.dekker.com)*

strong. However, some of the aerospace aluminum alloys (e.g., 2024-T3) are very prone to corrosion. In order to combat corrosion of aluminum alloys, chromic acid anodization and chromate conversion have thus far proved most effective. Both processes, however, use and release hexavalent chromium, a proven human carcinogen, causing serious environmental and worker health and safety concerns. Thus, a chromium-free alternative with equivalent or superior corrosion performance has long been desired.

The idea of corrosion-resistant zeolite coatings may be counter intuitive, but is well supported by several well-known facts about zeolites. First, high-silica and pure-silica zeolites are thermally and chemically stable. For example, many high-silica zeolites are thermally stable up to 1000°C, and pure-silica zeolites are stable in all mineral acids except hydrofluoric acid. Thus, the framework of zeolites contains the needed corrosion resistance. Second, most of the high-silica zeolite syntheses use template molecules, and these molecules are eventually trapped inside the zeolite pores. Normally, for separation and catalysis applications, these molecules have to be removed by high temperature calcination. It was recognized very early on that as-synthesized zeolite membranes could be gas-tight if they were defect-free. This suggested that as-synthesized zeolite films could be corrosion resistant.

It has been recently shown that as-synthesized template-containing high-silica MFI coatings on aluminum alloys have superior corrosion resistance to chromate conversion coatings in strong acids, bases, as well as pitting aggressive media,[26] and the in situ crystallization coating deposition process can coat surfaces of complex shape and in confined spaces. In addition, the thermal and mechanical properties of the high-silica MFI coatings were shown to be satisfactory.[27] It was also shown that corrosion resistance is possible for all high-silica zeolite coatings such as MFI, BEA, and MTW,[28,29] and high-silica MFI coatings can be extended to other aluminum alloys such as 6061-T6.

Hydrophilic Coating

An environmental control system for a manned spacecraft typically includes a condenser for controlling cabin temperature and humidity. As the moisture-laden air is cooled in the small channels of the condenser, water normally condenses out as water droplets. In a microgravity environment, these water droplets tend to become entrained in the air stream and carried back into the cabin, failing to achieve water separation. In such an instance, the cabin air could become foggy. To effectively remove water from the air, heat exchanger surfaces could be coated with a hydrophilic zeolite coating[30] so that the condensation mode changes from drop-wise to film-wise (Fig. 4). The water film can then be removed by vacuum sipping.

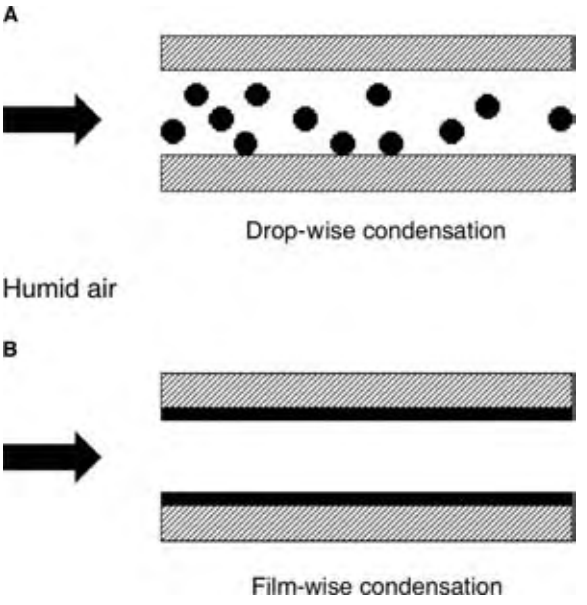


Fig. 4 Schematic illustration of the use of hydrophilic zeolite coating to change condensation mode from drop-wise without zeolite film (A) to film-wise with zeolite film (B).

In situ crystallization was used, and clearly preferred, as it can coat surfaces of complex shape and in confined spaces. In addition, it is a low temperature process (e.g., 165°C) and is compatible with most metal alloys. Another significant advantage for a zeolite coating is that an anti-microbial characteristic can be added readily through ion exchange of metal ions such as silver. Anti-microbial function is desirable for hydrophilic coatings as damp surfaces tend to grow bacteria and fungi. Furthermore, once depleted, the anti-microbial capability of zeolite coatings can easily be regenerated by another ion-exchange without the need of a potentially costly recoating. Hydrophilic membranes made from zeolite A (LTA) with a low silicon to aluminum ratio have been silver exchanged and are effective at killing a large number of cells over a short amount of time (Fig. 5).[31]

Heat Pump

It is well known that water adsorption in zeolites releases heat while water desorption absorbs heat. This phenomenon has been exploited for the design of a heat pump for cooling applications using waste heat or solar energy. Currently, zeolite pellets are used, and the heat and mass transfer are inefficient, leading to bulky pumps. Recently, there have been renewed interests in zeolite coating heat pumps.[32] A new in situ

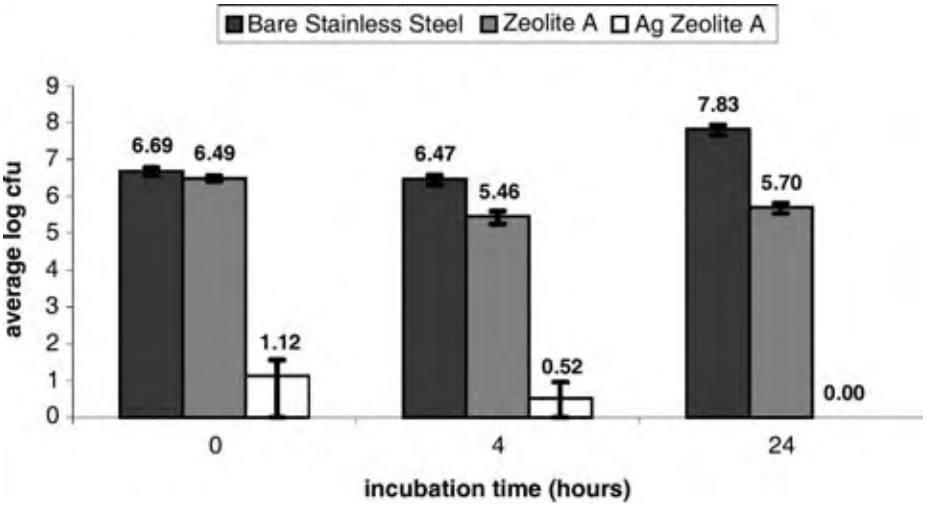


Fig. 5 Surviving colony forming units on bare stainless steel zeolite A coated stainless steel, and silver ion-exchanged zeolite A over a 24 hr incubation period. (From Ref.[31].)

crystallization method for zeolite coatings was developed which suppresses the reaction in the bulk and promotes it on the substrate by applying a temperature difference between the reaction mixture and the substrate. The substrate was heated directly, while the reaction mixture was kept at a lower temperature by means of a water bath. The method was tested for zeolite A coatings on SS plates from a clear aluminosilicate synthesis solution. Zeolite layer thickness and wall thickness of the heat exchanger tubes are essential to optimize the cycle durations. The optimum zeolite layer thickness was found to be 75–150 μm, depending on the wall thickness of the heat exchanger tubes, which led to almost a two-fold power increase from the adsorption heat pump.

Thermoelectrics

Thermoelectrics are all solid-state devices that can directly convert heat to electricity (power generation), or use electricity to pump heat from cold to hot (refrigeration), both without using any moving parts or hazardous compressed fluids. Although the phenomena was discovered in the early 19th century, it was only after the mid-20th century that technology brought practical applications for thermoelectric devices. Thermoelectric devices are lightweight and compact, and have extremely fast responses, and therefore are particularly useful for applications such as microelectronics and semiconductor lasers where conventional compression based cooling is very difficult or impossible. Applications of thermoelectric devices are presently limited because of their low efficiency. The efficiency of a thermoelectric device is usually characterized by a dimensionless figure-of-merit, ZT. Current thermoelectric materials have a ZT of 1. A thermoelectric cooler with a ZT of 1 operates at only 10% of Carnot efficiency, whereas a thermoelectric cooler with a ZT of 4 would achieve 30% of Carnot efficiency, which is comparable to the efficiency of a home refrigerator.

Very recently, theoretical calculations for isolated single nanowires predict over a ten-fold increase in the figure-of-merit. In order to realize these size-effects, wires must be synthesized with diameters less than the thermal de Broglie wavelength (typically less than 10 nm), which is smaller than what is currently accessible, by lithography techniques. Also, in order to transfer meaningful amounts of thermal energy, a practical device must consist of an array of a large number of nanowires in parallel. One way to potentially synthesize these nanowire arrays is to use the pores of microporous and mesoporous materials as a mold to template the diameter and orientation of wires of suitable thermoelectric materials. A recent review has been published discussing the relevant physics, from which a model was derived to assess the feasibility of using microporous and mesoporous frameworks in the fabrication of thermoelectric devices.[33] The model accounted for the possible deleterious effects of thermal conduction through the microporous or mesoporous framework and the presence of bulk thermoelectric material (that may result from defects in the microporous or mesoporous framework) in parallel with the wires on the thermoelectric figure-of-merit. Simulation results were reported for SBA-15, MCM-41, and zeolites VFI, LTL, and LTA embedded with Bi_2Te_3 nanowires. The results showed that the microporous zeolite frameworks may yield figures-of-merit much larger than one, while thermal conduction through most mesoporous frameworks reduces the figure-of-merit below the current level using traditional thermoelectric materials.

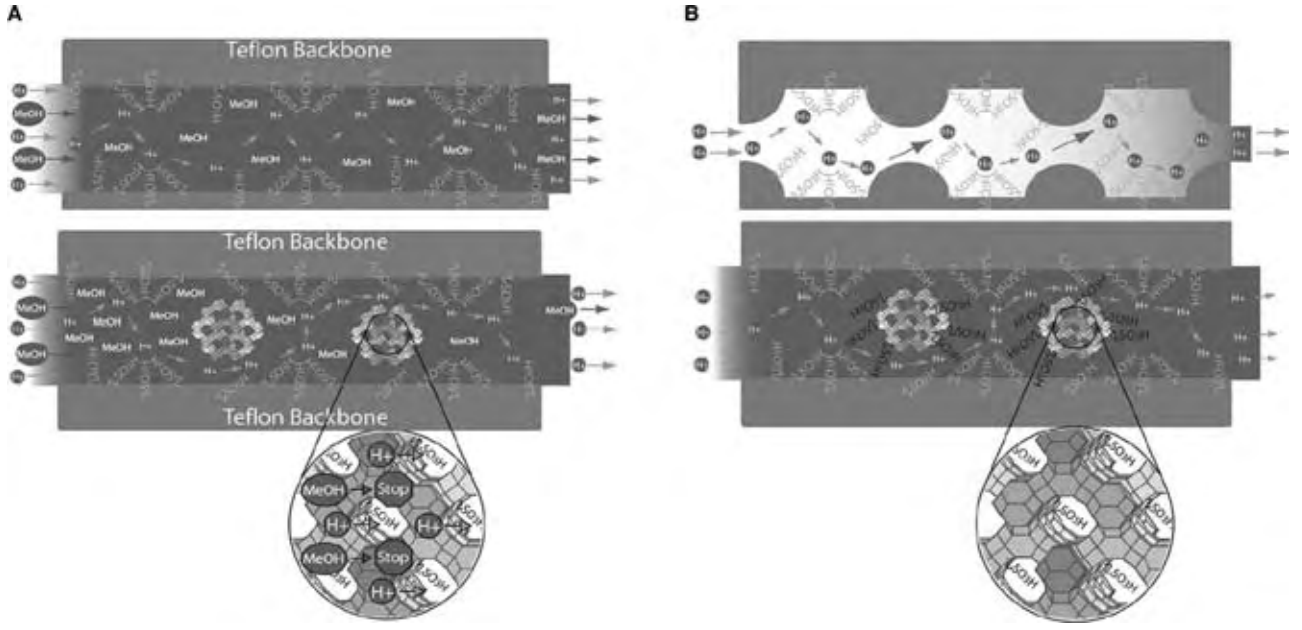


Fig. 6 Illustration of Nafion-acid functionalized zeolite Beta nanocomposite membranes helping to increase the proton conductivity and decrease the methanol crossover: (A) H_2O and CH_3OH diffusion reduced by zeolite flow resistance; (B) sulphonic acid functionalized zeolite nanoparticles increases proton conductivity of composite membranes. *(View this art in color at www.dekker.com.)*

NEW DEVELOPMENTS

Proton Exchange Membranes for Fuel Cells

Direct methanol fuel cells (DMFCs) are potentially compact, high efficiency power sources for portable applications. One of the major problems of DMFCs is methanol crossover, meaning that methanol molecules can permeate through the Nafion® membrane from anode to cathode, where it is chemically oxidized. This reduces the fuel efficiency and also causes mixed potential on the cathode. It has been recently demonstrated that acid functionalized zeolite beta Nafion nanocomposite membranes can take advantage of the molecular sieving property and high proton conductivity of acid functionalized zeolite beta nanocrystals, and are shown to have higher proton conductivity and lower methanol crossover (e.g., <50%) than Nafion[10]

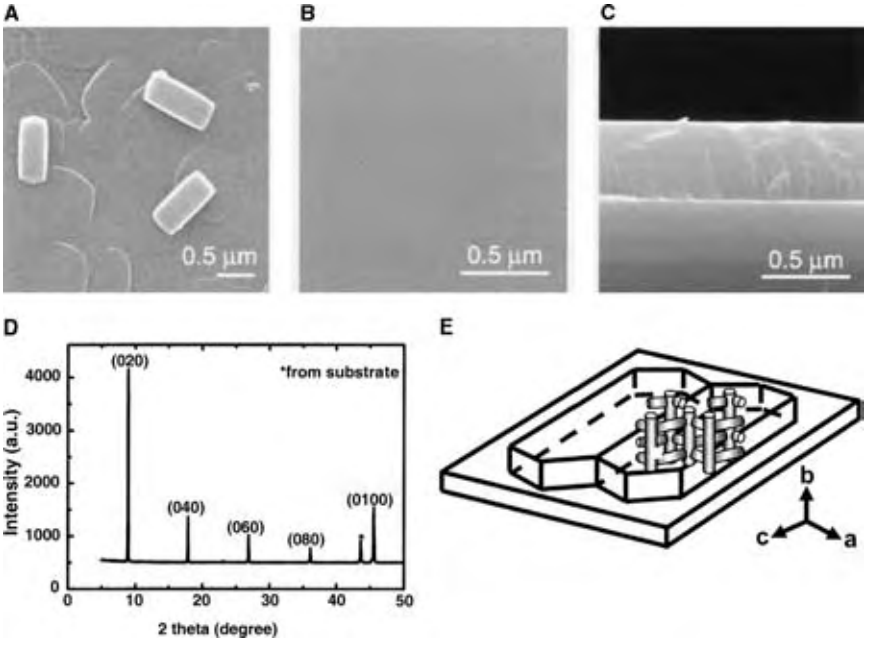


Fig. 7 (A, B, C) SEM micrographs of silicalite-1 films on silicon wafer by in-situ crystallization: (A) before polishing, top view; (B) after polishing, top view; (C) after polishing, cross sectional view; (D) XRD pattern of the silicalite-1 film showing that all the straight channels are vertical to the silicon substrate; (E) a schematic showing the orientation of the crystals vs. the substrate. (From Refs.[34–36].)

(Fig. 6). Polymer–zeolite nanocomposite membrane exploits the unique size of zeolite nanocrystals (e.g., adjustable diameter from 1580 nm) and has the potential to synergistically combine the advantages of polymers and zeolites while overcoming the shortcomings of both. Highly crystallized zeolite beta nanocrystals were hydrothermally synthesized with phenethyl organic functional groups. Sulfonic acid groups were then added to the phenethyl side chains with a concentrated sulfuric acid treatment. Acid functionalized zeolite beta nanocrystals were mixed with Nafion in suspension and cast into a membrane in a glass vessel. This nanocomposite membrane also has the potential to be able to operate at high temperatures (e.g., 150°C) for hydrogen proton exchange membrane (PEM) fuel cells.

Oriented Zeolite Films

Almost all applications for zeolite films could benefit from a preferred orientation. This is especially critical for zeolites with anisotropic pore systems. One particularly interesting *b*-oriented MFI film is shown in Fig. 7. It is a continuous monolayer film (0.4–0.5 μm thick) by in situ crystallization on metal substrates (SS and Al alloy)[34–36] or on silicon wafers.[19] Since MFI has straight channels in the *b*-direction, this film is extremely well suited for membrane applications.[38] It has been found that there is positive thermal expansion in the *a*- and *c*-directions and negative expansion in the *b*-direction; thus, a *b*-oriented MFI film helps to relieve thermal stress. The thinness is also important for relieving thermal stress. Also, because this is a monolayer film, it does not have any grain boundaries in the direction perpendicular to the substrate, leading to minimized resistance for mass transport. This film could also be useful for sensor, magnetic, optical, and thermoelectric applications as the template for nanowire arrays.

Patterned Zeolite Films

The ability to pattern zeolite films is critical for many of their potential applications, especially optical, magnetic, and electronic applications that have not been adequately explored so far. The first patterned zeolite film was achieved by using zeolite nanocrystals as building blocks combined with soft lithography[38] (Fig. 8). Convection-assisted assembly was also demonstrated for the generation of surface patterned zeolite films although the patterns are not very regular[6] (Fig. 8). Recently, using organic linkers, Yoon et al. have demonstrated that microcrystals can also be patterned and oriented. Although these films have very sophisticated patterns, they may find limited use as they are either not continuous or adhesion is not very strong. There is a critical need to fabricate patterned oriented continuous films on a variety of substrates. This was recently accomplished for *b*-oriented MFI monocrystal layer film on gold patterned silicon wafer by taking advantage of the fact that selective deposition can be achieved on Si over Au.[39]

Zeolite Nanocrystals as Building Blocks

Zeolite nanocrystals have been demonstrated to be versatile building blocks for constructing hierarchical porous structures.[40–42] The use of nanoparticles as building blocks allows mild processing conditions

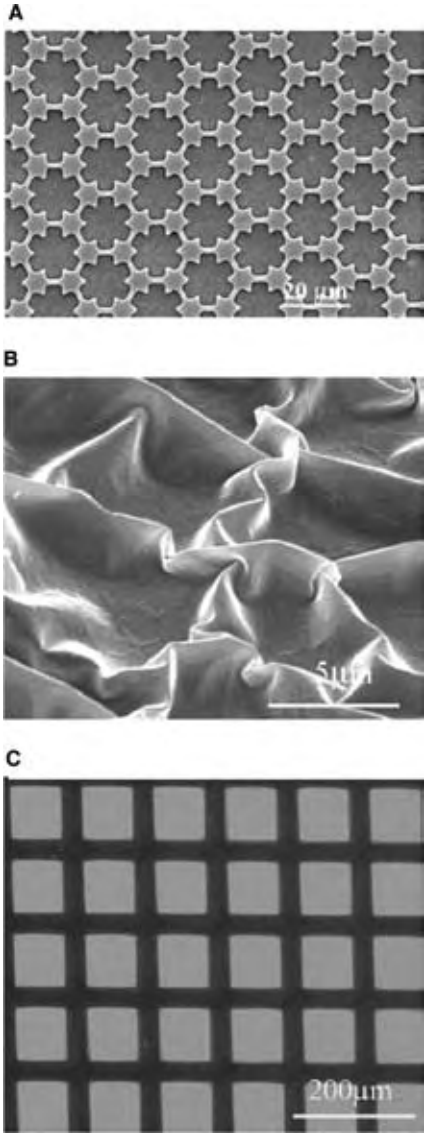


Fig. 8 Patterned zeolite films by using: (A) soft-lithography; (B) convection-assisted assembly of zeolite nanoparticles; and (C) selective deposition on Si on a wafer pre-patterned with gold. (From Refs.[6,38,39].)

and innovative strategies to be used that would have been difficult or impossible otherwise with the conventional hydrothermal synthesis methods such as in situ crystallization, seeded growth, and vapor phase transport. The fabrication of 3-D structures with designed shapes via gel-casting[42] and transparent zeolite films[40] (Fig. 9) are examples that clearly illustrate the utilities and versatility of this approach.

Specifically for the preparation of zeolite films, zeolite nanoparticles could be used directly to form zeolite films by self-assembly.[38] They could also be combined nicely with the seeded growth and vapor phase transport method to produce high quality polycrystalline zeolite films. There is clear evidence that small zeolite nanoparticles are preferred for producing compact continuous films.

Another important application for zeolite nanoparticles is the preparation of zeolite matrix composite films. Micrometer-sized crystals have long been used in these types of films, using techniques such as wash-coating and sol-gel processing. However, the use of nanoparticles in these films appears to offer many unparalleled advantages in terms of achievable film thickness, uniformity, and processability. The low-k dielectric films[21,22,43] and polymer-zeolite nanocomposite membranes[9,44] are two examples.

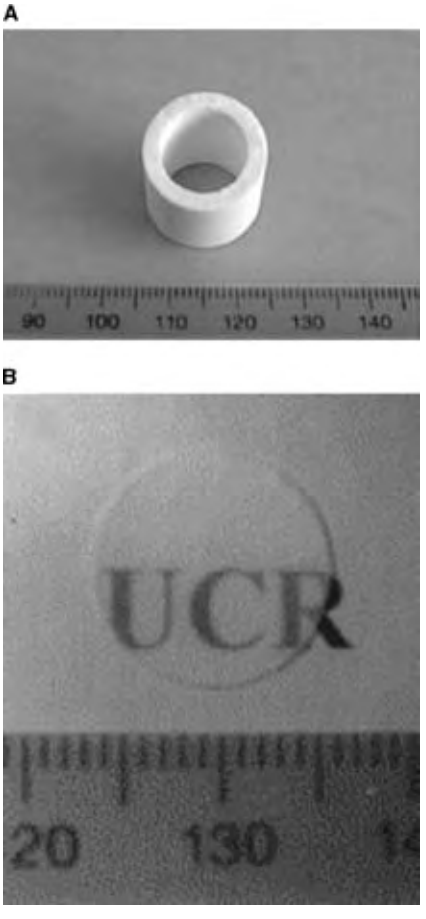


Fig. 9 Zeolite structures constructed by using zeolite nanocrystals as building blocks.. (From Refs.[40,42].)

Another interesting application is microreactors. Zeolite adsorbents and catalysts have been an important part for the conventional reaction-separation system. As the efforts to miniaturize the reactors evolve, using zeolite nanocrystals as building blocks could be an ideal way to introduce zeolites into these microreactor systems.

CONCLUSIONS

Since the first paper on ZME using large crystals,[45] the field of zeolite membranes has come a long way during the last decade from almost nonexistence in the early 1980s to a vibrant field. Many applications show promises and one commercial application has already been realized today (i.e., zeolite A membrane for water separation). The key to this success has been the persistent work of a large number of researchers around the world during the last 10 yrs. The single biggest driver behind this collective effort so far appears to be the potential application of zeolite membranes for separation and membrane reactors. However, non-conventional applications (i.e., non-separation and non-catalysis) are quickly being added and promise to offer even more exciting opportunities (e.g., zeolite low-k films and polymer–zeolite nanocomposite membranes for fuel cells). As shown in this entry, zeolite-based nanostructured membranes could take on several different configurations and be prepared by a number of synthesis methods. Current strategies for oriented membranes are highly specific for a particular zeolite or a particular orientation. More general methods for preparing oriented membranes are still critically needed. The use of nanoparticles as building blocks has broadened our capability significantly and will continue to bring exciting opportunities to the field. Although not covered in this entry, additional new applications such as optical, magnetic, and electric devices are being, and will continue to be explored. As the research of zeolite membranes draws more and more attention, novel applications will emerge.

ACKNOWLEDGMENTS

We wish to thank our sponsors for financial support and our colleagues in the Yan Group at University of California at Riverside for contributing to our work that is included in this entry.

Sponsors: AlliedSignal Inc., Honeywell International, Advanced Micro Devices, Asahi Kasei Corporation, Pacific Fuel Cells Corp., Engelhard

Corporation, The Technology for Sustainable Environment (TSE) program of the US Environmental Protection Agency, The Strategic Environmental Research and Development Program (SERDP) of the US Department of Defense, University of California—Discovery Grant (UC-DISCOVERY), University of California—Toxic Substance Research & Teaching Program (UC-TSR&TP), University of California—Energy Institute (UC-EI), California Energy Commission (CEC), Riverside Public Utilities (RPU), and NIRT/NSF (CTS-0404376).

Colleagues: Dr. Z. Wang, Dr. X. Cheng, Dr. A. Mitra, Dr. L. Huang, Dr. H. Wang, Dr. X. Wang, T. Cao, D. Beving, B. Holmberg, C. Demmelmaier, and A. McDonnell.

REFERENCES

1. Yan, Y.; Davis, M.E.; Gavalas, G.R. Preparation of Zeolite ZSM-5 membranes by in-situ crystallization on porous.alpha.-Al2O3. Ind. Eng. Chem. Res. **1995**, *34* (5), 1652–1661.
2. Haag, W.O.; Tsikoyiannis, J.G. Membrane composed of a pure molecular sieve. U.S. Patent 5019263, 1991.
3. Lovallo, M.C.; Tsapatsis, M. Preferentially oriented submicron silicalite membranes. AIChE J. **1996**, *42* (11), 3020–3029.
4. Matsukata, M.; Nishiyama, N.; Ueyama, K. Zeolitic membrane synthesized on a porous alumina support. J. Chem. Soc. Chem. Comm. **1994**, 339–340.
5. Wang, Z.; Mitra, A.; Wang, H.; Huang, L.; Yan, Y. Pure silica zeolite films as low-k dielectrics by spin-on of nanoparticle suspensions. Adv. Mater. **2001**, *13* (19), 1463–1466.
6. Wang, H.; Wang, Z.; Huang, L.; Mitra, A.; Yan, Y. Surface patterned porous films by convection-assisted dynamic self-assembly of zeolite nanoparticles. Langmuir **2001**, *17* (9), 2572–2574.
7. Wang, H.; Huang, L.; Holmberg, B.A.; Yan, Y. Nanostructured zeolite 4A molecular sieving air separation membranes. Chem. Comm. **2002**, 1708–1709.
8. Li, S.; Li, Z.; Yan, Y. Ultra-low-k pure-silica zeolite MFI films using cyclodextrin as porogen. Adv. Mater. **2003**, *15* (18), 1528–1531.
9. Wang, H.; Holmberg Brett, A.; Yan Y. Homogeneous polymer–zeolite nanocomposite membranes by incorporating dispersible template-removed zeolite nanocrystals. J. Mater. Chem. **2002**, *12* (12), 3640–3643.
10. Holmberg, B.A.; Wang, H.; Norbeck, J.M.; Yan, Y. Nafion/acid functionalized zeolite nanocomposite fuel cell membranes. Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.) **2004**, *45* (1), 24–25.
11. Bein, T. Synthesis and applications of molecular sieve layers and membranes. Chem. Mater. **1996**, *8* (8), 1636–1653.
12. Ha, K.; Lee, Y.-J.; Lee, H.J.; Yoon, K.B. Facile assembly of zeolite monolayers on glass, silica, alumina, and other zeolites using 3-Halopropylsilyl reagents as covalent linkers. Adv. Mater. **2000**, *12* (15), 1114–1117.
13. Caro, J.; Noack, M.; Kolsch, P.; Schafer, R. Zeolite membranes—state of their development and perspective. Micro. Meso. Mater. **2000**, *38* (1), 3–24.
14. Tavolaro, A.; Drioli, E. Zeolite membranes. Adv. Mater. **1999**, *11* (12), 975–996.
15. Yan, Y.; Wang, H. Nanostructured zeolite films. Encyclopedia of Nanoscience and Nanotechnology **2004**, *7*, 763–781.
16. Tsapatsis, M.; Xomeritakis, G.; Hillhouse, H.; Nair, S.; Nikolakis, V.; Bonilla, G.; Lai, Z. Zeolite membranes. Cattech **2000**, *3* (2), 148–163.
17. Wang, Y.J.; Tang, Y.; Wang, X.D.; Dong, A.G.; Shan, W.; Gao, Z. Fabrication of hierarchically structured zeolites through layer-by-layer assembly of zeolite nanocrystals on diatom templates. Chem. Lett. **2001**, *11*, 1118–1119.
18. Wang, Z.B.; Yan, Y.S. Controlling crystal orientation in zeolite MFI thin films by direct in situ crystallization. Chem. Mater. **2001**, *13* (3), 1101–1107.
19. Wang, Z.; Wang, H.; Mitra, A.; Huang, L.; Yan, Y. Pure-silica zeolite low-k dielectric thin films. Adv. Mater. **2001**, *13* (10), 746–749.
20. Li, S.; Wang, X.; Beving, D.; Chen, Z.W.; Yan, Y.S. Molecular sieving in a nanoporous b-oriented pure-silica-zeolite MFI monocrystal film. J. Am. Chem. Soc. **2004**, *126* (13), 4122–4123.
21. Wang, Z.; Wang, H.; Mitra, A.; Huang, L.; Yan, Y. Pure-silica zeolite low-k dielectric thin films by spin-on process. Stud. Surf. Sci. Cata. **2001**, *135*, 3217–3224.
22. Li, S.; Li, Z.J.; Yan, Y.S. Ultra-low-k pure-silica zeolite MFI films using cyclodextrin as porogen. Adv. Mater. **2003**, *15* (18), 1528–1531.
23. Li, Z.J.; Li, S.; Yan, Y.S. Effects of crystallinity in spin-on pure-silica-zeolite MFI low dielectric constant films. Adv. Func. Mater. **2004**, *14* (10), 1019–1024.
24. Li, Z.; Li, S.; Medina, D.I.; Yan, Y. Low-k Films from Pure-Silica-Zeolite MEL Nanoparticle Suspension. J. Phys. Chem. B. **2005**, *109* (18), 8652–8658.
25. Li, S.; Yan, Y. Organic-functionalized pure-silica-zeolite MFI ultra low-k films. Chem. Mater. **2005**, *17* (7), 1851–1854.

26. Cheng, X.; Wang, Z.; Yan, Y. Corrosion-resistant zeolite coatings by in-situ crystallization. Electrochem. Solid-State Letts. **2001**, *4* (5), B23–B26.
27. Wang, H.; Wang, Z.; Cheng, X.; Mitra, A.; Huang, L.; Yan, Y. Characterization of corrosion-resistant zeolite coatings on Al alloys. Stud. Surf. Sci. Catal. **2001**, *135*, 3448–3455.
28. Mitra, A.; Wang, Z.; Cao, T.; Wang, H.; Huang, L.; Yan, Y. Synthesis and corrosion resistance of high-silica zeolite MTW, BEA, and MFI coatings on steel and aluminum. J. Electrochem. Soc. **2002**, *149* (10), B472–B478.
29. Mitra, A.; Kirby, C.W.; Wang, Z.; Huang, L.; Wang, H.; Huang, Y.; Yan, Y. Synthesis of pure-silica MTW powder and supported films. Micro. Meso. Mate. **2002**, *54* (1–2), 175–186.
30. Yan, Y. Hydrophilic zeolite coatings. U.S. Patent 6500490, 2002.
31. McDonnell, A.M.P.; Beving, D.; Wang, A.; Chen, W.; Yan, Y. Hydrophilic and antimicrobial zeolite coatings for gravity independent water separation. Adv. Fun. Mater. **2004**, *15* (2), 336–340.
32. Erdem-Senatalar, A.; Tatlier, M.; Urgen, M. Preparation of zeolite coatings by direct heating of the substrates. Micro. Meso. Mater. **1999**, *32* (3), 331–343.
33. Hillhouse, H.W.; Tuominen, M.T. Modeling the thermoelectric transport properties of nanowires embedded in oriented microporous and mesoporous films. Micro. Meso. Mater. **2001**, *47* (1), 39–50.
34. Wang, Z.; Yan, Y. Controlling crystal orientation in zeolite MFI thin films by direct in situ crystallization. Chem. Mater. **2001**, *13* (3), 1101–1107.
35. Wang, Z.; Yan, Y. Oriented zeolite MFI monolayer films on metal substrates by in situ crystallization. Micro. Meso. Mater. **2001**, *48* (1–3), 229–238.
36. Li, S.; Li, Z.; Bozhilov, K.N.; Chen, Z.; Yan, Y. TEM investigation of formation mechanism of monocrystal-thick b-oriented pure silica zeolite MFI film. J. Am. Chem. Soc. **2004**, *126* (34), 10732–10737.
37. Lai, Z.; Bonilla, G.; Diaz, I.; Nery, J.G.; Sujaoti, K.; Amat, M.A.; Kokkoli, E.; Terasaki, O.; Thompson, R.W.; Tsapatsis, M.; Vlachos, D.G. Microstructural optimization of a zeolite membrane for organic vapor separation. Science **2003**, *300* (5618), 456–460.
38. Huang, L.; Wang, Z.; Sun, J.; Miao, L.; Li, Q.; Yan, Y.; Zhao, D. Fabrication of ordered porous structures by self-assembly of zeolite nanocrystals. J. Am. Chem. Soc. **2000**, *122* (14), 3530–3531.
39. Li, S.; Demmelmaier, C.; Itkis, M.; Liu, Z.; Haddon, R.C.; Yan, Y. Micropatterned oriented zeolite monolayer films by direct in situ crystallization. Chem. Mater. **2003**, *15* (14), 2687–2689.
40. Huang, L.; Wang, Z.; Wang, H.; Sun, J.; Li, Q.; Zhao, D.; Yan, Y. Hierarchical porous structures by using zeolite nanocrystals as building blocks. Micro. Meso. Mater. **2001**, *48* (1–3), 73–78.
41. Jung, K.T.; Shul, Y.G. Preparation of ZSM-5 zeolite film and its formation mechanism. J. Membr. Sci. **2001**, *191* (1–2), 189–197.
42. Wang, H.; Huang, L.; Wang, Z.; Mitra, A.; Yan, Y. Hierarchical zeolite structures with designed shapes by gelcasting of colloidal nanocrystal suspensions. Chem. Comm. **2001**, 1364–1365.
43. Wang, Z.; Mitra, A.; Wang, H.; Huang, L.; Yan, Y. Pure silica zeolite films as low-k dielectrics by spin-on of nanoparticle suspensions. Adv. Mater. **2001**, *13* (19), 1463–1466.
44. Wang, H.; Huang, L.; Holmberg Brett, A.; Yan, Y. Nanostructured zeolite 4A molecular sieving air separation membranes. Chem. Comm. **2002**, 1708–1709.
45. Marshall, C.E. Use of zeolitic membrane electrodes. J. Phys. Chem. **1939**, *43*, 1155–1164.

Ziegler–Natta Catalysis

John C. Chadwick
Dutch Polymer Institute (DPI), Laboratory of Polymer Chemistry, Eindhoven University of Technology, Eindhoven, The Netherlands

INTRODUCTION

Ziegler–Natta catalysts play a dominant role in polyolefins manufacture. More than 50 million tonnes per annum of polyethylene and polypropylene are now produced by means of Ziegler–Natta catalysis. Since the first discoveries, more than 50 years ago, many breakthroughs and innovations have been made in catalyst and process chemistry and technology, leading to ever more efficient manufacturing processes, and also to increasing control over polymer structure and properties.

This entry describes the development of successive generations of Ziegler–Natta catalysts and highlights the main factors determining catalyst activity, stereospecificity, and the relationship between catalyst type and polymer molecular weight, molecular weight distribution, and chain microstructure. Important mechanistic aspects of propylene polymerization using $MgCl_2$-supported Ziegler–Natta catalysis are outlined, and the current position of Ziegler–Natta catalysts in polyolefin production is considered in relation to developments in the area of homogeneous catalysis.

BACKGROUND

Ziegler–Natta catalysis stems from breakthrough discoveries made by Karl Ziegler and Giulio Natta in the early 1950s. In 1953, Ziegler and coworkers at the Max Planck Institute in Mülheim were investigating the ''Aufbau'' reaction in which triethylaluminum reacts with ethylene to give higher aluminum trialkyls.[1] Unexpectedly, one experiment led to the formation of 1-butene and it turned out that this dimerization reaction had been catalyzed by traces of nickel present as a contaminant in the reactor. Soon afterwards, a revolutionary breakthrough was achieved when combinations of transition metal compounds and aluminum alkyls were found which could polymerize ethylene under mild conditions, yielding high-density polyethylene. In 1954, Giulio Natta and coworkers at Milan Polytechnic were successful in polymerizing propylene with the Ziegler catalyst combination $TiCl_4$/$AlEt_3$ and in fractionating the resulting polymer to obtain and characterize isotactic polypropylene.[2] This demonstration of stereoregular polymerization led to an explosive growth of new olefin- and diene-based polymers, and Ziegler and Natta were jointly awarded the Nobel Prize for chemistry in 1963.

The first-generation Ziegler–Natta catalysts used in early manufacturing processes for polypropylene (PP) comprised $TiCl_3$ and cocrystallized $AlCl_3$, resulting from reduction of $TiCl_4$ with Al or an aluminum alkyl. At low Al/Ti ratios, this reaction yields titanium trichloride as a solid precipitate. The β-$TiCl_3$ formed can be converted to the γ form by heating.[3] The latter catalyst has much higher stereoregulating ability in propylene polymerization, while β-$TiCl_3$ is an effective catalyst for the production of cis-1,4-polyisoprene. The cocatalyst used in propylene polymerization was $AlEt_2Cl$ (DEAC). Catalyst activity was relatively low, giving polymer yields of around 1 kg PP/g cat., necessitating removal (de-ashing) of catalyst residues from the polymer. In many cases, extractive removal of atactic polymer was also required.

In the 1970s, an improved $TiCl_3$ catalyst for PP was developed by Solvay.[4] Catalyst preparation involved reduction of $TiCl_4$ using DEAC, followed by treatment with an ether and $TiCl_4$. The ether treatment results in the removal of $AlCl_3$ from $TiCl_3.nAlCl_3$, while treatment with $TiCl_4$ effects a phase transformation from β to δ-$TiCl_3$ at a relatively mild temperature ($<100°C$).[5] Using catalysts of this type, it was possible to obtain PP yields in the range 5–20 kg/g cat. in 1–4 hrs polymerization in liquid monomer.[6] Commercial implementation of second-generation catalysts was, however, overshadowed by the advent of third- and later-generation magnesium chloride-supported catalysts, described later in the sections entitled ''Ziegler–Natta Catalysts for Ethylene (Co) Polymerization'' and ''Ziegler–Natta Catalysts for Polypropylene.''

POLYMER CHAIN GROWTH

The essential characteristic of Ziegler–Natta catalysis is the polymerization of an olefin or diene, using a combination of a transition metal compound and a base metal alkyl cocatalyst, normally an aluminum alkyl. The function of the cocatalyst is to alkylate the transition metal, generating a transition metal–carbon

Encyclopedia of Chemical Processing DOI: 10.1081/E-ECHP-120008015

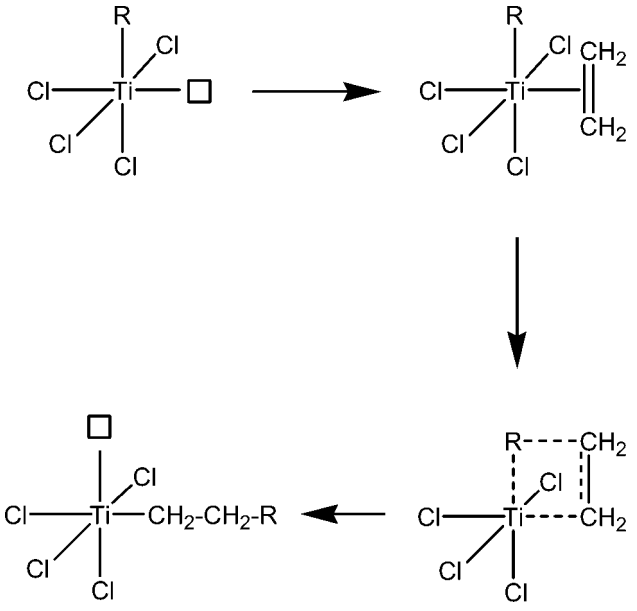


Fig. 1 Cossee–Arlman mechanism for polymerization.

bond. It is also essential that the active center contains a coordination vacancy. Chain propagation takes place via the Cossee–Arlman mechanism,[7] in which coordination of the olefin at the vacant coordination site is followed by chain migratory insertion into the metal carbon bond, as illustrated in Fig. 1.

Regulation of polyolefin molecular weight is effected by the use of hydrogen as the chain transfer agent. Chain transfer can also occur via β-hydrogen transfer from the growing chain to the transition metal or to monomer, and, to a lesser extent, via alkyl exchange with the cocatalyst.

In propylene polymerization using titanium chloride catalysts, chain propagation takes place via primary (1,2-) insertion of the monomer. For isospecific propagation, there must be only one coordination vacancy and the active site must be chiral. Corradini and coworkers have demonstrated that the asymmetric environment of the active site forces the growing chain to adopt a particular orientation so as to minimize steric interactions with (chlorine) ligands present on the catalyst surface.[8] This in turn leads to one particular prochiral face of the incoming monomer being preferred, as illustrated in Fig. 2, leading to isotactic polymer.

An elegant demonstration of the above mechanism has been provided by Zambelli et al.,[9] who showed that the first insertion of propylene into a Ti-CH_3 bond generated by chain transfer with Al alkyl using the system $TiCl_3/Al(CH_3)_3$ is *not* stereospecific, whereas the second insertion (i.e., into Ti-isobutyl) *is* stereospecific.

In contrast to the isospecific titanium-based catalysts, vanadium-based catalysts give predominantly syndiotactic polypropylene. At very low polymerization temperature (–78°C), living polymerization can be obtained using homogeneous catalysts obtained by the reaction of a vanadium compound (e.g. VCl_4 or a V(III) β-diketonate) with R_2AlCl.[10,11] With these catalysts, syndiospecific propagation occurs via secondary (2,1-) insertion of the monomer. The overall stereo- and regioregularity of the polymer is poor, comprising not only syndiotactic blocks resulting from secondary insertions, but also short, atactic blocks arising from sequences of primary insertions. This polymer has not been developed commercially, but vanadium catalysts are used in ethylene (co) polymerization, outlined in the section entitled "Ziegler–Natta Catalysts for Ethylene (Co) Polymerization" below. C_s-symmetric metallocene catalysts[12] have been developed for the production of syndiotactic polypropylene having significantly higher chain regularity.

POLYMER PARTICLE GROWTH

A very important feature of any heterogeneous catalyst used in slurry and gas-phase processes for polyolefin production is particle morphology. Typically, heterogeneous Ziegler–Natta catalysts have particle sizes in the range 10–100 μm. Each particle comprises of millions of primary crystallites with sizes of up to approximately 15 nm. On contacting the catalyst components, at the start of polymerization, cocatalyst and monomer diffuse through the catalyst particle and polymerization takes place on the surface of each primary crystallite within the particle. As solid crystalline polymer is formed, the primary crystallites are pushed outwards and apart as the particle grows, analogous to the expanding universe. The particle shape is retained, and this phenomenon is therefore referred to as replication (Fig. 3). Ideally, the catalyst particle should have spherical morphology and controllable porosity.

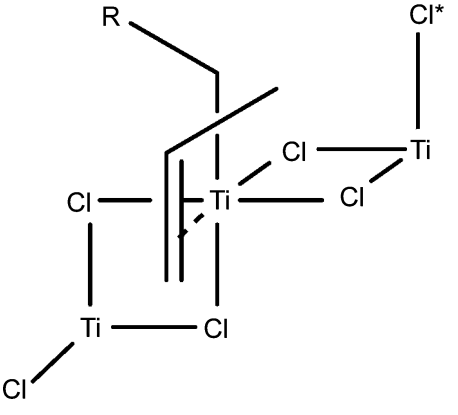


Fig. 2 Model for stereospecific polymerization of propylene. The orientation of the growing chain is influenced by the chlorine atom marked with an asterisk.

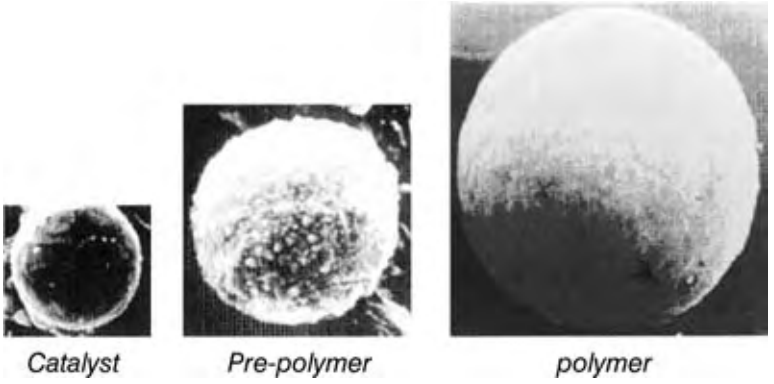


Fig. 3 ''Replication'' phenomenon during polymerization. *(View this art in color at www.dekker.com.)*

It is important that the mechanical strength of the catalyst is high enough to prevent disintegration, but low enough to allow progressive expansion as polymerization proceeds.[13]

Under appropriate polymerization conditions, polymer particles can be obtained with an internal morphology that can range from a compact to a porous structure.[14] In what is termed Reactor Granule Technology (RGT), porous polymer particles can be produced which can then function as a microreactor for the polymerization of other monomers within the solid matrix. A polypropylene skin encloses the second polymer phase, thereby preventing coalescence of particles in which the second phase is an amorphous, low-melting material. Reactor Granule Technology is able to provide products ranging from superstiff, high-fluidity PP homopolymers to stiff/impact or clear/impact heterophasic copolymers and supersoft alloys, produced using the *Catalloy* process.[13,15] The feasibility of producing heterophasic alloys containing up to 70% of elastomeric copolymers arises from the use of a controlled porosity catalyst and the ability to control the porosity of the growing polymer particle during the early stages of polymerization. Prepolymerization is applied to give the particles sufficient heat capacity to withstand injection into a gas-phase polymerization step.

Several models have been put forward to explain the mechanism of particle growth during polymerization.[16] One of the most popular models is the ''multigrain model,'' put forward by Ray et al.,[17] in which the monomer diffuses into the catalyst macroparticle and polymerises on the surface of the microparticles within, causing progressive expansion of the macroparticle as polymerization proceeds. An investigation by Kakugo et al.[18] of nascent polymer morphology obtained using a $TiCl_3$ catalyst, showed that the polymer particle comprised numerous globules, each of which contained some tens of much smaller primary particles. Recently, a model for particle growth with $MgCl_2$-supported catalysts has been proposed by Cecchin,[15,19] who has also provided evidence for polymer ''subglobule'' formation within the growing particle. Again, these subglobules originate from clusters of primary crystallites, but the crystallites themselves are pushed to the surface of each subglobule as the polymer forms. This model explains the fact that, in the preparation of heterophasic copolymers via propylene homopolymerization followed by ethylene/propylene copolymerization, the rubbery E/P copolymer is formed at the surface of the homopolymer subglobules, gradually filling up the pores within the particle. Evidence for drifting of catalyst microparticles to the surface of polymer (sub) globules has been provided by scanning electron microscopy studies of prepolymerized catalyst particles.[20]

ZIEGLER–NATTA CATALYSTS FOR ETHYLENE (CO) POLYMERIZATION

Ziegler–Natta catalysts are widely used in the production of high-density and linear low-density polyethylene (HDPE and LLDPE). More than half the world production of HDPE and over 90% of LLDPE is based on Ziegler–Natta catalysts, although increased use of metallocene and other single-site catalysts is expected throughout the next decade.

The most important titanium-based catalysts for HDPE and LLDPE are those comprising a titanium component on magnesium chloride or on a magnesium chloride-containing support. Towards the end of the 1960s, catalysts obtained by reaction of $TiCl_4$, or derivative thereof, with a magnesium compound such as Mg(OH)Cl, $Mg(OH)_2$, $Mg(OH)_2$, or $MgCl_2$, were found to give very high activity in ethylene polymerization, eliminating the need for deashing of the polymer.[13,21] The most effective support was found to be active magnesium chloride, prepared by comilling of $MgCl_2$ and titanium halides or by chlorination of organomagnesium compounds.[22] Numerous catalyst systems and methods of preparation have been

disclosed,[23] and the characteristics of magnesium chloride as a support for Ziegler–Natta catalysts are discussed in more depth later in the section entitled ''Ziegler–Natta Catalysts for Polypropylene''. Magnesium chloride can also be used in combination with a silica support, for example, by impregnation of the porous support with a solution of $MgCl_2$ and $TiCl_4$ in tetrahydrofuran.[24]

An important manufacturing process for HDPE which makes use of high-mileage catalysts is the cascade process, in which polymerization reactors in series are used to give reactor blends with improved properties for film and pipe applications.[25] Broad molecular weight distribution can be obtained by the use of different hydrogen concentrations in each reactor. In addition, the process can be designed to give low molecular weight homopolymer in the first reactor and a high molecular weight copolymer in the second. The high molecular weight copolymer chains function as tie molecules linking the crystalline, homopolymer domains, thereby leading to high stress crack resistance of the polymer. This process allows an ''inverse'' comonomer distribution to be obtained, in the sense that the comonomer is in the high molecular weight fraction, counteracting the general tendency of Ziegler–Natta catalysts to incorporate the comonomer mainly in the low molecular weight chains. The latter feature is an important consideration in Ziegler–Natta catalyst design for LLDPE, where the best catalysts are those that have relatively uniform active center distribution.

Vanadium catalysts have also been developed for polyethylene and ethylene-based copolymers, particularly ethylene-propylene-diene rubbers (EPDM). Homogeneous (soluble) vanadium catalysts produce relatively narrow molecular weight distribution polyethylene, whereas supported V catalysts give broad MWD.[26]

Ethylene polymerization, in contrast to the polymerization of propylene and other alpha-olefins, is often influenced by diffusion limitations, which occur if the monomer reactivity in polymerization is high, relative to diffusivity through the catalyst particle. This can result in the formation of an ''onion'' type internal particle morphology as polymerization first takes place at the external surface of the particle, particle growth occurring step by step as the monomer reaches the inner parts of the catalyst particle. This mechanism of particle growth is associated with a kinetic profile in which an initial induction period is followed by an acceleration period, after which, in the absence of chemical deactivation, a stationary rate is obtained. Ethylene polymerization activity can be increased by first carrying out a prepolymerization with propylene, which results in a significant lowering in the monomer diffusion barrier in the subsequent ethylene polymerization.[27]

ZIEGLER–NATTA CATALYSTS FOR POLYPROPYLENE

The manufacture of polypropylene (PP) is dominated by high-activity $MgCl_2$-supported Ziegler–Natta catalysts. The development and implementation of $MgCl_2$-supported catalysts in bulk (liquid monomer) and gas-phase processes has led to the advent of simple, low-cost (non-de-ashing, nonextracting) manufacturing processes for PP,[28] global production of which now exceeds 30 million tonnes per annum.

The basis for the development of the high-activity supported catalysts lay in the discovery in the late 1960s of ''activated'' $MgCl_2$ able to support $TiCl_4$ and give high catalyst activity, and the subsequent discovery in the mid-1970s of electron donors (Lewis bases) capable of increasing the stereospecificity of the catalyst so that (highly) isotactic polypropylene could be obtained.[21,22,29,30]

Catalyst Structure and Composition

In the early stages of $MgCl_2$-supported catalyst development, activated magnesium chloride was prepared by ball milling in the presence of ethyl benzoate, leading to the formation of very small ($\leq$3 nm thick) primary crystallites within each particle.[5] Nowadays, however, the activated support is prepared by chemical means, such as via complex formation of $MgCl_2$ and an alcohol, or by reaction of a magnesium alkyl, or alkoxide with a chlorinating agent or $TiCl_4$. Many of these approaches are also effective for the preparation of catalysts having controlled particle size and morphology. For example, the cooling of emulsions of molten $MgCl_2.n$EtOH in paraffin oil gives almost perfectly spherical supports, which are then converted into the catalysts.[28] A typical catalyst preparation involves reaction of the $MgCl_2.n$EtOH support with excess $TiCl_4$ in the presence of an ''internal'' electron donor.

High-activity Ziegler–Natta catalysts comprising $MgCl_2$, $TiCl_4$, and an internal donor, are typically used in combination with an aluminum alkyl cocatalyst such as $AlEt_3$ and an ''external'' electron donor added in polymerization. The first catalyst systems containing ethyl benzoate as internal donor were used in combination with a second aromatic ester such as methyl p-toluate as external donor.[30] These were followed by catalysts containing a diester (e.g., diisobutyl phthalate) as internal donor, used in combination with an alkoxysilane external donor of type $RR'Si(OMe)_2$ or $RSi(OMe)_3$.[31] The combination $MgCl_2/TiCl_4$/phthalate ester–AlR_3–alkoxysilane represents the most widely used catalyst system in PP manufacture. The most effective alkoxysilane donors for high isospecificity are methoxysilanes containing

relatively bulky alkyl groups branched at the position alpha to the silicon atom,[32,33] such as cyclohexyl(methyl) dimethoxysilane and dicyclopentyldimethoxysilane.[34] Of these, the latter gives particularly high stereospecificity[35] and broader molecular weight distribution.[36]

The function of the internal donor in $MgCl_2$-supported catalysts is to stabilize the small primary crystallites of magnesium chloride formed in the catalyst preparation and to control the amount and distribution of $TiCl_4$ in the final catalyst. Activated magnesium chloride has a disordered structure comprising very small lamellae. Giannini[22] has indicated that, on preferential lateral cleavage surfaces, the magnesium atoms are coordinated with four or five chlorine atoms, as opposed to six chlorine atoms in the bulk of the crystal. These lateral cuts correspond to (110) and (100) faces of $MgCl_2$ (Fig. 4).

It has been proposed that bridged, dinuclear Ti_2Cl_8 species can coordinate to the (100) face of $MgCl_2$ and give rise to the formation of chiral, isospecific active species (Fig. 5), it being pointed out that Ti_2Cl_6 species formed by reduction on contact with $AlEt_3$ would resemble analogous species in $TiCl_3$ catalysts.[37] Accordingly, it has been suggested[28] that a possible function of the internal donor is preferential coordination on the more acidic (110) face of $MgCl_2$, such that this face is prevailingly occupied by donor and the (100) face is prevailingly occupied by Ti_2Cl_8 dimers.

Analytical studies[38] have indicated that a monoester internal donor such as ethyl benzoate is coordinated to $MgCl_2$ and not to $TiCl_4$. In the search for donors giving catalysts having improved performance, it was considered[39] that bidentate donors should be able to form strong chelating complexes with

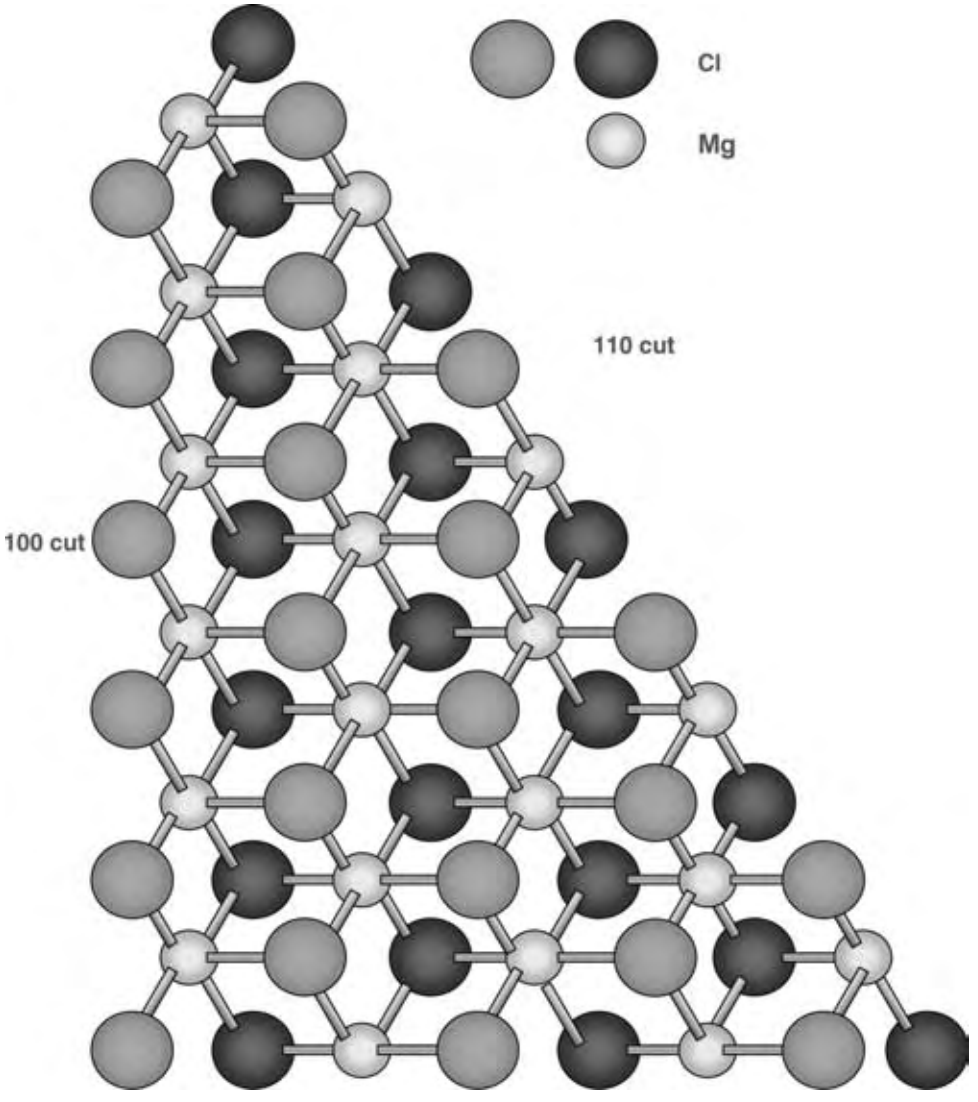


Fig. 4 Model of a $MgCl_2$ layer showing the (100) and (110) lateral cuts. (From Ref.[13].) *(View this art in color at www.dekker.com.)*

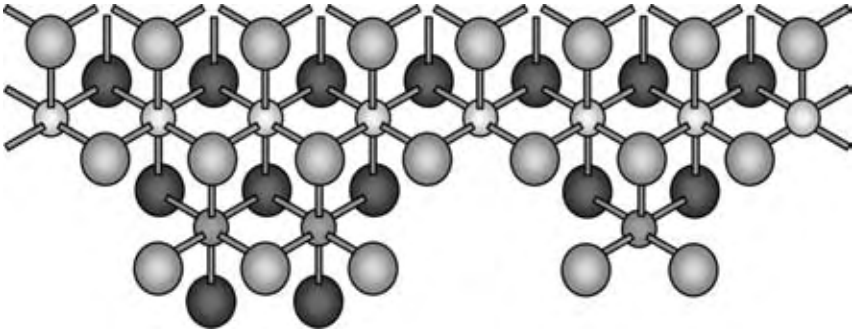


Fig. 5 Model showing dimeric and monomeric Ti species on a (100) lateral cut of $MgCl_2$. (From Ref.[13].) *(View this art in color at www.dekker.com.)*

tetracoordinate Mg atoms on the (110) face of $MgCl_2$, or binuclear complexes with two pentacoordinate Mg atoms on the (100) face. This led to the development of the $MgCl_2/TiCl_4$/phthalate ester catalysts, used, as indicated above, in combination with an alkoxysilane as external donor. The requirement for an external donor when using catalysts containing a benzoate or phthalate ester is due to the fact that, when the catalyst is brought into contact with the cocatalyst, a large proportion of the internal donor is lost as a result of alkylation and/or complexation reactions, which, in the absence of an external donor, would lead to poor stereospecificity. When the external donor is present, contact of the catalyst components leads to replacement of the internal donor by the external donor. The most active and stereospecific systems were found to be those which allowed the highest incorporation of external donor,[40] the effectiveness of a catalyst system depending more on the combination of donors rather than on the individual internal or external donor.

More recently, research on $MgCl_2$-supported catalysts has led to systems not requiring the use of an external donor. This required the identification of bidentate internal donors, which not only had the right oxygen–oxygen distance for effective coordination with $MgCl_2$ but which, unlike phthalate esters, were not removed from the support on contact with $AlEt_3$. It was found[39,41–43] that certain 2,2-disubstituted-1,3-dimethoxypropanes met all these criteria. The best performance was obtained when bulky substituents in the 2-position resulted in the diether having a most probable conformation[44] with an oxygen–oxygen distance in the range 2.8–3.2 Å. Successive ''generations'' of high-activity $MgCl_2$-supported catalyst systems for polypropylene are summarised below:

$MgCl_2/TiCl_4$/ethyl benzoate – AlR_3 – aromatic ester

$MgCl_2/TiCl_4$/phthalate ester – AlR_3 – alkoxysilane

$MgCl_2/TiCl_4$/diether – AlR_3

Catalyst performance has improved considerably with each generation. The polypropylene yield obtained under typical polymerization conditions (liquid monomer, in the presence of hydrogen, 70°C, 1–2 hrs) has increased from 15–30 kg/g cat. for the third generation ethyl benzoate-containing catalysts to 30–80 kg/g cat. for the fourth generation phthalate-based catalysts. With the fifth generation catalysts containing a diether as internal donor, yields of 80–160 kg/g cat. can be achieved. These different catalysts also display different kinetic profiles in propylene polymerization. The catalysts containing a diether as internal donor exhibit very stable activities during polymerization. A low rate of catalyst decay during polymerization is also obtained with the catalyst system $MgCl_2/TiCl_4$/phthalate ester–AlR_3–alkoxysilane, whereas the system $MgCl_2/TiCl_4$/ethyl benzoate–AlR_3–aromatic ester has a very high initial polymerization activity, but also a high decay rate, limiting the final polymer yield. The rapid decay in activity can, at least, partially be ascribed to the use of an ester as external as well as internal donor, the ester being able to react with titanium–hydrogen bonds formed in chain transfer with hydrogen, generating Ti–O bonds inactive for chain propagation.[45]

Most recently, a further family of $MgCl_2$-supported catalysts has been developed[15,46] in which the internal donor is a succinate rather than a phthalate ester. As is the case with the phthalate-based catalysts, an alkoxysilane is used as external donor. The essential difference between these catalysts is that the succinate-based systems produce polypropylene having much broader molecular weight distribution, discussed below in the section entitled ''Catalyst/Polymer Relationship.''

Mechanistic Aspects

It is well established that effective external donors not only increase the isotactic index of the polymer (the proportion of polymer insoluble in boiling heptane or in xylene at 25°C), but they can also increase in absolute terms the amount of isotactic polymer formed. This has been demonstrated by Kashiwa[47] for the catalyst system $MgCl_2/TiCl_4$–$AlEt_3$. An increase in the molecular

weight and stereoregularity of the isotactic fraction was also noted. Similar trends are apparent with catalyst systems of type $MgCl_2/TiCl_4$/phthalate ester–AlR_3–alkoxysilane.[48] Kakugo[49] has used elution fractionation to demonstrate that the external donor not only decreases ''atactics'' formation, but also increases the degree of steric control at isospecific sites. This is also apparent from determination of the stereoregularity of the first insertion step in propylene polymerization.[50]

Recently, significant advances have been made in understanding the fundamental factors determining the performance of state-of-the-art $MgCl_2$-supported catalysts. Studies by Busico et al.[51] have shown that the chain irregularities in isotactic polypropylene prepared using Ziegler–Natta catalysts are not randomly distributed along the chain but are clustered. Therefore, the chain can contain, in addition to highly isotactic blocks, sequences which can be attributed to weakly isotactic (isotactoid) and to syndiotactic blocks. This implies that the active site can isomerise very rapidly (during the growth time of a single polymer chain, i.e., in less than a second) between three different propagating species. The same sequences are present, but in different amounts, in both the soluble and the insoluble fractions. The polymer can therefore be considered to have a stereoblock structure in which highly isotactic sequences alternate with defective isotactic (isotactoid) and syndiotactoid sequences. Fig. 6 illustrates a mechanistic model in which the relative contributions of these sequences can be related to site transformations involving the presence or absence of steric hindrance in the vicinity of the active species.

If it is considered that the steric hindrance in the vicinity of the active species can result from the presence of a donor molecule, and that the coordination of such a donor is reversible, the above model provides us with an explanation for the fact that strongly-coordinating, stereorigid donors typically give stereoregular polymers in which the highly isotactic sequences predominate. Several types of active species in which the presence of a donor in the vicinity of the active Ti atom is necessary for high isospecificity have been proposed,[52] although the exact structure of the active species is still by no means resolved. The model illustrated in Fig. 6 has also been used as a basis for quantum mechanical calculation of activation energies for primary (1,2-) and secondary (2,1-) insertion of the propene monomer into the growing chain.[53] The lowest energy insertion path, in accordance with the Corradini mechanism,[8] is the 1,2-insertion shown in Fig. 7(A). Of the paths for 2,1-insertion, illustrated in Figs. 7(B) and (C), the latter is prohibited by steric repulsion with the ligand (L1 of Fig. 6). However, a low stereoselectivity of 2,1-insertion with a donor-free catalyst[53] indicates the presence of active species in which the ligand is missing.

In PP production, hydrogen is used as chain transfer agent for polymer molecular weight control. Catalysts containing a diether donor show particularly high sensitivity to hydrogen, such that relatively little hydrogen is required for molecular weight control. This effect can be ascribed to chain transfer after the occasional 2,1- rather than the usual 1,2-insertion, a 2,1-insertion slowing down a subsequent monomer insertion and therefore increasing the probability of chain transfer.[54] Reactivation of ''dormant'' (2,1-inserted) species via chain transfer with hydrogen also explains the frequently observed activating effect of hydrogen in propylene polymerization, giving polymer yields which may be around three times those observed in the complete absence of hydrogen. These conclusions have been based on the ^{13}C NMR determination of the relative proportions of *i*-Bu and *n*-Bu terminated chains, resulting from chain transfer with hydrogen after primary and secondary insertion respectively:

$$TiCH_2CH(CH_3)[CH_2CH(CH_3)]_nPr + H_2 \rightarrow TiH + iBuCH(CH_3)[CH_2CH(CH_3)]_{n-1}Pr$$

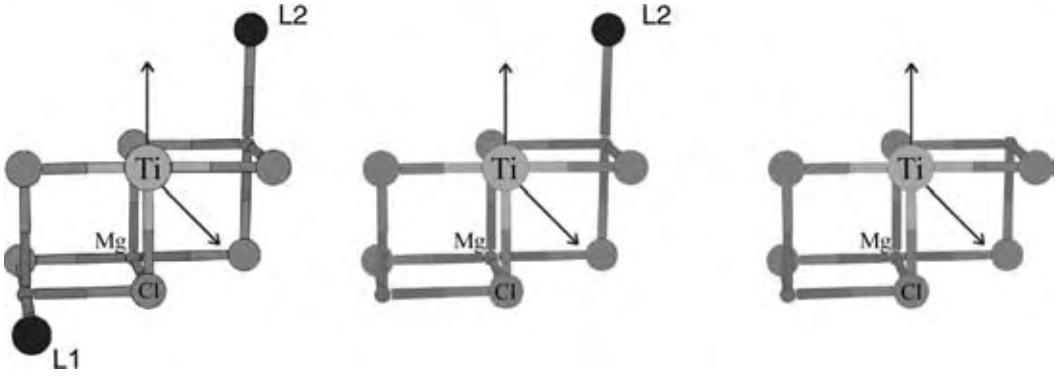


Fig. 6 Model of possible active species for highly isotactic, isotactoid and syndiotactic propagation. (From Ref.[53].) *(View this art in color at www.dekker.com.)*

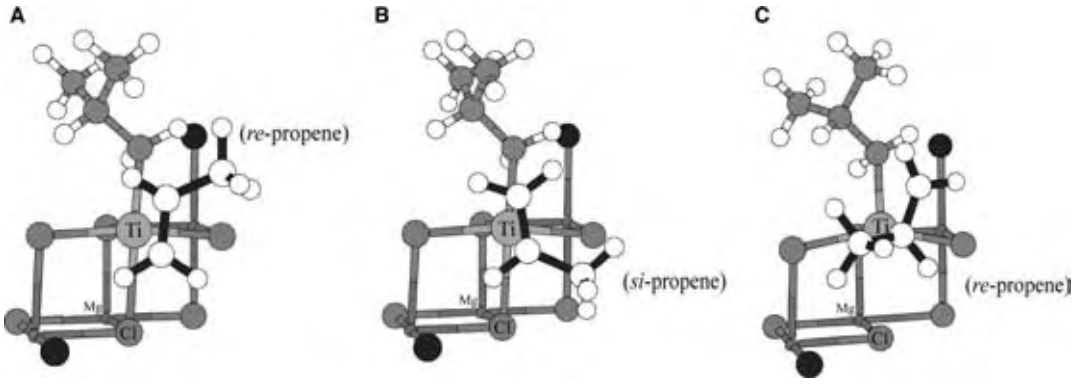


Fig. 7 Optimized geometries for (A): 1,2-insertion, and (B) and (C): 2,1-insertion of propylene. (From Ref.[53].) *(View this art in color at www.dekker.com.)*

$$TiCH(CH_3)CH_2[CH_2CH(CH_3)]_nPr + H_2 \rightarrow TiH + nBuCH(CH_3)[CH_2CH(CH_3)]_{n-1}Pr$$

Other studies have demonstrated that chain transfer is dependent not only on *regio-*, but also on *stereo-*selectivity.[35] This is in keeping with the tendency that, with catalyst systems of type $MgCl_2/TiCl_4$/phthalate ester–AlR_3–alkoxysilane, the silanes which give the most stereoregular isotactic polymer also give relatively low hydrogen response. Effective chain transfer with hydrogen also appears to be dependent on the presence of species able to promote the dissociation of H_2 to atomic hydrogen, studies by Terano and coworkers[55] having shown that, under stopped-flow conditions, hydrogen is only effective as a chain transfer agent when catalyst and cocatalyst have been precontacted.

CATALYST/POLYMER RELATIONSHIP

By varying the catalyst composition, and in particular the nature of the electron donors (esters, silanes, diethers) present in the catalyst system, it is possible to control the polypropylene tacticity, molecular weight, and molecular weight distribution to produce a range of polymers having the processing and end-use properties required for very different applications. Ziegler–Natta catalysts typically give broader molecular weight distributions than are obtained with homogeneous (metallocene) catalysts. This is because Ziegler–Natta catalysts contain a range of different active centers, each center giving different relative rates of chain propagation and chain transfer. The proportion of Ti which is catalytically active in $MgCl_2$-supported catalyst systems for PP is relatively low, values reported by different groups ranging from 1% or less[56] to more than 20%.[57] The propagation rate constant, k_p, for isospecific active sites, is around an order of magnitude greater than for weakly-specific sites[56,57] and increases significantly in the presence of hydrogen,[58] in accordance with the reactivation of ''dormant'' (2,1-inserted) centers by chain transfer.

The donors present in the catalyst system play an active role in the formation or modification of isospecific sites, and the polymer molecular weight distribution depends on the relative contribution and hydrogen response (i.e., sensitivity to chain transfer with hydrogen) of each type of active site. The characteristics of different catalyst systems with regard to PP molecular weight distribution are as follows:[15]

Internal donor	*External donor*	M_w/M_n
Diether	–	5–5.5
Phthalate	Alkoxysilane	6.5–8
Succinate	Alkoxysilane	10–15

It can be seen that the diether-based catalysts give relatively narrow molecular weight distribution. A narrow molecular weight distribution, and relatively low molecular weight, is advantageous in fibre spinning applications. In contrast, extrusion of pipes and thick sheets requires high melt strength, and, therefore, relatively high molecular weight and broad molecular weight distribution. A broad molecular weight distribution, along with high isotactic stereoregularity, is also beneficial for high crystallinity and, therefore, high rigidity. The new succinate-based catalysts enable very broad MWD PP homopolymers to be produced in a single reactor and are also of interest for the production of heterophasic copolymers having an improved balance of stiffness and impact strength, taking into account that the incorporation of a rubbery (ethylene/propylene) copolymer phase into a PP homopolymer matrix increases impact strength, but, at the same time, leads to decreased stiffness.

The relatively narrow PP molecular weight distributions obtained using diether-based catalysts can be attributed to the fact that, in these systems, even the most highly stereospecific active sites are not totally regiospecific. A proportion of approximately one secondary insertion for every 2000 primary insertions at highly isospecific sites has been noted for the system $MgCl_2/TiCl_4$/diether–AlR_3.[54] The probability of chain transfer with hydrogen after a secondary insertion is such that this is sufficient to prevent the formation of very high molecular weight chains, taking into account that the highest molecular weight fraction of the polymer is formed by the active species having the highest isospecificity. The broader molecular weight distributions obtained with catalysts containing ester internal donors are likely to be due to the presence of (some) isospecific active sites having very high regiospecificity and, therefore, lower hydrogen sensitivity. Such results illustrate the profound effect of catalyst regio- and stereospecificity distribution on both molecular weight control and polymer MWD and properties.

POLYMERIZATION OF OTHER MONOMERS

In addition to their widespread use in the production of polyethylene and polypropylene, Ziegler–Natta catalysts play an important role in the production of poly-1-butene, and are also widely used in the manufacture of synthetic rubbers such as *cis*-1,4-polybutadiene and *cis*-1,4-polyisoprene, the synthetic equivalent of natural rubber. Titanium-, cobalt-, nickel- and neodymium-based Ziegler–Natta catalyst systems are used in the manufacture of butadiene rubber. Isoprene rubber is produced using β-$TiCl_3$[59] typically prepared by the reaction of approximately equimolar quantities of $TiCl_4$ and Al*i*Bu_3 in the presence of a small quantity of an ether. Recently, it has been shown that almost perfectly stereoregular *cis*-1,4-polyisoprene can be synthesized using lanthanide metallocene catalysts.[60]

Cis-1,4-polymerization of conjugated dienes requires the presence of two coordination vacancies on the transition metal atom, allowing bidentate coordination of the diene. β-$TiCl_3$ has a fiber-like structure in which the titanium atoms in the lattice are octahedrally coordinated to six chlorine atoms. The terminal titanium atoms are, however, incompletely coordinated and are linked to four or five chlorine atoms. Alkylation of the tetracoordinated titanium atoms will generate the double-vacancy species active in isoprene polymerization. Stereospecificity in diene polymerization can change dramatically if one of the coordination vacancies is blocked by a Lewis base. An interesting illustration of this[61] is the addition of an external donor in isoprene polymerization with $TiCl_4$–$AlEt_3$ or $MgCl_2/TiCl_4$–$AlEt_3$, which changes the catalyst stereospecificity to give mainly *trans*-1,4- rather than *cis*-1,4-polyisoprene. At the same time, a notable increase in isospecificity in propylene polymerization is observed.

$TiCl_3$-based and $MgCl_2$-supported catalysts have been developed for the production of poly-1-butene. With $TiCl_3$, the use of $AlEt_2I$ as cocatalyst was found to give higher isotacticity than $AlEt_2Cl$.[62] Much higher polymerization activity, as well as high isotacticity and broad molecular weight distribution, is obtained using $MgCl_2$-supported catalysts, for example, the catalyst system $MgCl_2/TiCl_4$/di-isobutyl phthalate–$AlEt_3$–alkoxysilane.[63]

Ziegler–Natta catalysts have also been developed for the polymerization of 4-methyl-1-pentene[64] and higher alpha-olefins. Polymerization activity decreases with increasing steric bulk of the monomer. For example, with the catalyst system $MgCl_2/TiCl_4$/ethyl benzoate–$AlEt_3$–ethyl benzoate, the relative activities in propylene, 1-butene, and 4-methyl-1-pentene polymerization were 100:80:15.[65] For catalyst systems of type $MgCl_2/TiCl_4$/phthalate ester–AlR_3–alkoxysilane, the type of silane required is dependent on the steric bulk of the monomer. An active center having high stereospecificity in propylene polymerization may be too sterically hindered for effective polymerization of a bulkier monomer, propylene/1-butene copolymerization studies having shown[66] that the incorporation of 1-butene into the polymer chain decreases with increasing site stereospecificity. This phenomenon is also illustrated by the fact that nonbulky alkoxysilanes, such as Me_3SiOMe, are effective donors in 4-methyl-1-pentene polymerization,[67] whereas such donors are relatively ineffective in propylene polymerization.

FUTURE TRENDS IN ZIEGLER–NATTA CATALYSIS

Progress in (heterogeneous) Ziegler–Natta catalysis has continued unabated over the last 50 years, while the last 20 years have seen the advent of homogeneous (metallocene and other single-site) catalysts.[68] However, despite the enormous research effort and many advances made in the field of homogeneous catalysis, polyolefins manufacture is still dominated by Ziegler–Natta systems.

Homogeneous catalysts, in which each catalytic center is identical, give polymers having very narrow molecular weight and tacticity distribution and, in the case of copolymers, random comonomer incorporation. The latter feature is particularly important in polymers such as LLDPE, while a uniform stereo- or regiodefect distribution in PP can give improved optical and barrier properties. However, utilization of

these catalysts in slurry or gas-phase processes requires immobilization on a suitable support material.[69]

Inherently heterogeneous Ziegler–Natta catalysts have the advantage of excellent morphology control, while the nonuniform nature of the active species results in polymers having molecular weight distributions ranging from very broad to relatively narrow. For PP, this has opened the way to a range of polymers suitable for widely differing applicational areas. The most important variable determining the suitability of a catalyst for a particular application is the internal/external donor combination and continuing research in this area is highlighted by the recent development of succinate-based catalysts.[15,46] Very high activity, notably with diether-based catalysts, has already been achieved, but if the proportion of active centers in a typical Ziegler–Natta PP catalyst is taken to be less than 10% of the Ti present, it is clear that there is still considerable scope for further improvement.

There are many differences between Ziegler–Natta and metallocene catalysts, not the least of which is that in a $MgCl_2$-supported $TiCl_4$ catalyst the Ti centers are octahedral, whereas the metallocene complexes are tetrahedral. However, homogeneous single-site catalysts having octahedral geometry have now been developed. Complexes containing phenoxy-imine ligands have been developed that give exceptionally high activity in ethylene polymerization.[70] Other complexes, containing tetradentate ligands, have been found[71] to give isotactic poly(1-hexene) and the stereoselectivity of such C_2-symmetric complexes in propylene polymerization has been shown[72] to mimic the behavior of Ziegler–Natta catalysts, albeit with much lower activity. The accessibility of well-defined octahedral complexes for olefin polymerization therefore provides further mechanistic insight in Ziegler–Natta catalysis, and molecular modeling is playing an increasingly important role in defining the nature of the active species in both heterogeneous and homogeneous systems.[8]

CONCLUSIONS

Continual improvements in catalyst activity, selectivity, and particle morphology obtained with successive generations of Ziegler–Natta catalysts have paved the way to efficient polyolefin manufacturing processes and to ever-increasing control over polymer structure and properties. Different catalysts are required for different applications, and the characteristics of Ziegler–Natta catalysts for polypropylene are strongly dependent on the type of electron donor(s) used. Recent developments have led to new $MgCl_2/TiCl_4$/donor catalysts for polypropylene, with diether donors giving relatively narrow PP molecular weight distribution and succinate donors giving broad molecular weight distribution.

Ziegler–Natta catalysts are complex systems and the exact structure of the various active species is still by no means fully understood. Nevertheless, significant advances in basic understanding have recently been made with regard to the role of the electron donor and the effects of the stereo- and regiospecificity of monomer insertion on polymer chain growth and chain transfer. Recent developments in homogeneous (single-site) catalysis have provided further mechanistic insight, with molecular modeling playing an increasingly important role.

Polyolefins manufacture is still dominated by Ziegler–Natta catalysts, and this situation is likely to remain for some time, despite the many advances being made in the field of metallocene and related single-site catalysis. Ziegler–Natta and single-site catalysts can be regarded as complementary rather than competitive systems, which together provide the basis for a wide range of polymers with closely controlled molecular structure.

REFERENCES

1. Wilke, G. Fifty years of Ziegler catalysts: consequences and development of an invention. Angew. Chem. Int. Ed. **2003**, *42*, 5000–5008.
2. Corradini, P. The discovery of isotactic polypropylene and its impact on pure and applied science. J. Polym. Sci: Part A: Polym. Chem. **2004**, *42*, 391–395.
3. Boor, J., Jr. *Ziegler–Natta Catalysts and Polymerizations*; Academic Press: New York, 1979.
4. Hermans, J.P.; Henrioulle, P. Catalytic complexes US Patent 4,210,738, March 21, 1972.
5. Goodall, B.L. Polypropylene; catalysts and polymerization aspects. In *Polypropylene and other Polyolefins. Polymerization and Characterization*; Van der Ven, S., Ed.; Elsevier: Amsterdam, 1990; 1–133.
6. Bernard, A.; Fiasse, P. New Solvay SB 12 $TiCl_3$ polypropylene catalyst. In *Catalytic Olefin Polymerization*; Keii, T., Soga, K., Eds.; Elsevier: Amsterdam, 1990; 405–423.
7. Arlman, E.J.; Cossee, P. Ziegler–Natta catalysis III. Stereospecific polymerization of propene with the catalyst system $TiCl_3$-$AlEt_3$. J. Catal. **1964**, *3*, 99–104.
8. Corradini, P.; Guerra, G.; Cavallo, L. Do new century catalysts unravel the mechanism of stereocontrol on old Ziegler–Natta catalysts? Acc. Chem. Res. **2004**, *37*, 231–241.
9. Zambelli, A.; Sacchi, M.C.; Locatelli, P.; Zannoni, G. Isotactic polymerization of α-olefins:

stereoregulation for different reactive chain ends. Macromolecules **1982**, *15*, 211–212.
10. Doi, Y.; Suzuki, S.; Soga, K. A perfect initiator for ''living'' coordination polymerization of propene: tris(2-methyl-1,3-butanedionato)vanadium/diethylaluminum chloride system. Makromol. Chem., Rapid Commun. **1985**, *6*, 639–642.
11. Zambelli, A.; Sessa, I.; Grisi, F.; Fusco, R.; Accomazzi, P. Syndiotactic polymerization of propylene: single-site vanadium catalysts in comparison with zirconium and nickel. Macromol. Rapid Commun. **2001**, *22*, 297–310.
12. Ewen, J.A.; Jones, R.L.; Razavi, A.; Ferrara, J.D. Syndiospecific propylene polymerization with Group 4 metallocenes. J. Amer. Chem. Soc. **1988**, *110*, 6255–6256.
13. Simonazzi, T.; Giannini, U. Forty years of development in Ziegler–Natta catalysis: from innovations to industrial realities. Gazz. Chim. Ital. **1994**, *124*, 533–541.
14. Galli, P. The breakthrough in catalysis and processes for olefin polymerization: innovative structures and a strategy in the materials area for the twenty-first century. Progr. Polym. Sci. **1994**, *19*, 959–974.
15. Cecchin, G.; Morini, G.; Pelliconi, A. Polypropylene product innovation by reactor granule technology. Macromol. Symp. **2001**, *173*, 195–209.
16. McKenna, T.F.; Soares, J.B.P. Single particle modelling for olefin polymerization on supported catalysts: a review and proposals for future developments. Chem. Eng. Sci. **2001**, *56*, 3931–3949.
17. Hutchinson, R.A.; Chen, C.M.; Ray, W.H. Polymerization of olefins through heterogeneous catalysis X: modeling of particle growth and morphology. J. Appl. Polym. Sci. **1992**, *44*, 1389–1414.
18. Kakugo, M.; Sadatoshi, H.; Yokoyama, M.; Kojima, K. Transmission electron microscopic investigation of nascent polypropylene particles using a new staining method. Macromolecules **1989**, *22*, 547–551.
19. Cecchin, G.; Marchetti, E.; Baruzzi, G. On the mechanism of polypropene growth over $MgCl_2/TiCl_4$ catalyst systems. Macromol. Chem. Phys. **2001**, *202*, 1987–1994.
20. Pater, J.T.M.; Weickert, G.; Loos, J.; van Swaaij, W.P.M. High precision prepolymerization of propylene at extremely low reaction rates—kinetics and morphology. Chem. Eng. Sci. **2001**, *56*, 4107–4120.
21. Kashiwa, N. The discovery and progress of $MgCl_2$-supported $TiCl_4$ catalysts. J. Polym. Sci.: Part A: Polym. Chem. **2004**, *42*, 1–8.
22. Giannini, U. Polymerization of olefins with high activity catalysts. Makromol. Chem., Suppl. **1981**, *5*, 216–229.
23. Choi, K-Y.; Ray, W.H. Recent developments in transition metal catalyzed olefin polymerization—a survey. 1. Ethylene polymerization. J. Macromol. Sci.—Rev.Macromol. Chem. Phys. **1985**, *C25*, 1–55.
24. Goerke, G.L.; Wagner, B.E.; Karol, F.J. Process for Copolymerizing Ethylene. Eur. Patent 4,647, March 30, 1979.
25. Böhm, L.L.; Bilda, D.; Breuers, W.; Enderle, H.F.; Lecht, R. The microreactor model-guideline for PE-HD process and product development. In *Ziegler Catalysts. Recent Scientific Innovations and Technological Improvements*; Fink, G., Mülhaupt, R., Brintzinger, H.H., Eds.; Springer-Verlag: Berlin, 1995; 387–400.
26. Karol, F.J.; Cann, K.J.; Wagner, B.E. Developments with high-activity titamium, vanadium, and chromium catalysts in ethylene polymerization. In *Transition Metals and Organometallics as Catalysts for Olefin Polymerization*; Kaminsky, W., Sinn, H., , Eds.; Springer-Verlag: Berlin, 1988; 149–161.
27. Zakharov, V.A.; Bukatov, G.D.; Barabanov, A.A. Recent data on the number of active centers and propagation rate constants in olefin polymerization with supported ZN catalysts. Macromol. Symp. **2004**, *213*, 19–28.
28. Moore, E.P., Jr. *Polypropylene Handbook–Polymerization, Characterization, Properties, Processing, Applications*; Hanser Publishers: Munich, 1996.
29. Galli, P.; Luciani, L.; Cecchin, G. Advances in the polymerization of polyolefins with coordination catalysts. Angew. Makromol. Chem. **1981**, *94*, 63–89.
30. Barbè, P.C.; Cecchin, G.; Noristi, L. The catalytic system Ti-complex/$MgCl_2$ Adv. Polym. Sci. **1987**, *81*, 1–81.
31. Parodi, S.; Nocci, R.; Giannini, U.; Barbè, P.C.; Scatà, U. Components and Catalysts for the Polymerization of Olefins. Eur. Patent 45,977, August 13, 1981.
32. Seppälä, J.V.; Härkönen, M.; Luciani, L. Effect of the structure of external alkoxysilane donors on the polymerization of propene with high activity Ziegler–Natta catalysts. Makromol. Chem. **1989**, *190*, 2535–2550.
33. Proto, A.; Oliva, L.; Pellecchia, C.; Sivak, A.J.; Cullo, L.A. Isotactic-specific polymerization of propene with supported catalysts in the presence of different modifiers. Macromolecules **1990**, *23*, 2904–2907.

34. Ishimaru, N.; Kioka, M.; Toyota, A. Process for polymerising olefins and polymerisation catalyst therefore. Eur. Patent 350,170, June 13, 1989.
35. Chadwick, J.C.; van Kessel, G.M.M.; Sudmeijer, O. Regio- and stereospecificity in propene polymerization with $MgCl_2$-supported Ziegler–Natta catalysts: effects of hydrogen and the external donor. Macromol. Chem. Phys. **1995**, *196*, 1431–1437.
36. Chadwick, J.C. Advances in propene polymerization using $MgCl_2$-supported catalysts: fundamental aspects and the role of electron donors. Macromol. Symp. **2001**, *173*, 21–35.
37. Busico, V.; Corradini, P.; De Martino, L.; Proto, A.; Savino, V.; Albizzati, E. Polymerization of propene in the presence of $MgCl_2$-supported Ziegler–Natta catalysts, 1: the role of ethyl benzoate as ''internal'' and ''external'' base. Makromol. Chem. **1985**, *186*, 1279–1288.
38. Terano, M.; Kataoka, T.; Keii, T. A study on the states of ethyl benzoate and $TiCl_4$ in $MgCl_2$-supported high-yield catalysts. Makromol. Chem. **1987**, *188*, 1477–1487.
39. Albizzati, E.; Giannini, U.; Morini, G.; Smith, C.A.; Zeigler, R.C. Advances in propylene polymerization with $MgCl_2$ supported catalysts. In *Ziegler Catalysts. Recent Scientific Innovations and Technological Improvements*; Fink, G., Mülhaupt, R., Brintzinger, H.H., Eds.; Springer-Verlag: Berlin, 1995; 413–425.
40. Sacchi, M.C.; Tritto, I.; Shan, C.; Mendichi, R.; Zannoni, G.; Noristi, L. Role of the pair of internal and external donors in magnesium chloride-supported Ziegler–Natta catalysts. Macromolecules **1991**, *24*, 6823–1626.
41. Albizzati, E.; Barbè, P.C.; Noristi, L.; Scordamaglia, R.; Barino, L; Giannini, U.; Morini, G. Components and Catalysts for the Polymerization of Olefins. Eur. Patent 361,494, September 29, 1989.
42. Morini, G.; Cristofori, A. Diethers Suitable for Use in the Preparation of Ziegler–Natta Catalysts. Eur. Patent 728,724, February 21, 1996.
43. Albizzati, E.; Giannini, U.; Morini, G.; Galimberti, M.; Barino, L.; Scordamaglia, R. Recent advances in propylene polymerization with $MgCl_2$-supported catalysts. Macromol. Symp. **1995**, *89*, 73–89.
44. Barino, L.; Scordamaglia, R. Steric equivalence between internal and external donors as polymerization stereoregulators: a molecular mechanics study. Macromol. Symp. **1995**, *89*, 101–111.
45. Albizzati, E.; Galimberti, M.; Giannini, U.; Morini, G. The chemistry of magnesium chloride supported catalysts for polypropylene. Makromol. Chem., Macromol. Symp. **1991**, *48/49*, 223–238.
46. Morini, G.; Balbontin, G.; Gulevich, Y.; Duijghuisen, H.; Kelder, R.; Klusener, P.A.; Korndorffer, F. Components and catalysts for the polymerization of olefins. Int. Patent WO 00/63261, April 12, 2000.
47. Kashiwa, N.; Yoshitake, J.; Toyota, A. Studies on propylene polymerization with a highly active $MgCl_2$-supported $TiCl_4$ catalyst system. Polym. Bull. (Berlin) **1988**, *19*, 333–338.
48. Chadwick, J.C. Effects of electron donors in super high activity catalysts for polypropylene. In *Ziegler Catalysts. Recent Scientific Innovations and Technological Improvements*; Fink, G., Mülhaupt, R., Brintzinger, H.H., Eds.; Springer-Verlag: Berlin, 1995; 427–440.
49. Kakugo, M.; Miyatake, T.; Naito, Y.; Mizunuma, K. Microtacticity distribution of polypropylene prepared with heterogeneous Ziegler–Natta catalysts. Macromolecules **1988**, *21*, 314–319.
50. Sacchi, M.C.; Shan, C.; Locatelli, P.; Tritto, I. Stereochemical investigation of the insertion step in $MgCl_2$-supported Ziegler–Natta catalysts. The Lewis base activation effect. Macromolecules **1990**, *23*, 383–386.
51. Busico, V.; Cipullo, R.; Monaco, G.; Talarico, G.; Vacatello, M.; Chadwick, J.C.; Segre, A.L.; Sudmeijer, O. High-resolution ^{13}C NMR configurational analysis of polypropylene made with $MgCl_2$-supported Ziegler–Natta catalysts. 1. The ''model'' system $MgCl_2/TiCl_4$-2,6-dimethylpyridine/$Al(C_2H_5)_3$. Macromolecules **1999**, *32*, 4173–4182.
52. Barino, L.; Scordamaglia, R. Modeling of isospecific Ti sites in $MgCl_2$-supported heterogeneous Ziegler–Natta catalysts. Macromol. Theory Simul. **1998**, *7*, 407–419.
53. Busico, V.; Cipullo, R.; Polzone, C.; Talarico, G.; Chadwick, J.C. Propene/ethene-[1-^{13}C] copolymerization as a tool for investigating catalyst regioselectivity. 2. The $MgCl_2/TiCl_4$-AlR_3 system. Macromolecules **2003**, *36*, 2616–2622.
54. Chadwick, J.C.; Morini, G.; Albizzati, E.; Balbontin, G.; Mingozzi, I.; Cristofori, A.; Sudmeijer, O.; van Kessel, G.M.M. Aspects of hydrogen activation in propene polymerization using $MgCl_2/TiCl_4$/diether catalysts. Macromol. Chem. Phys. **1996**, *197*, 2501–2510.
55. Mori, H.; Tashino, K.; Terano, M. Study of the chain transfer reaction by hydrogen in the initial stage of propene polymerization. Macromol. Rapid Commun. **1995**, *16*, 651–657.
56. Bukatov, G.D; Zakharov, V.A. Propylene Ziegler–Natta polymerization: numbers and propagation rate constants for stereospecific and non-stereospecific centers. Macromol. Chem. Phys. **2001**, *202*, 2003–2009.

57. Tait, P.J.T.; Zohuri, G.H.; Kells, A.M.; McKenzie, I.D. Kinetic studies in propene polymerization using magnesium dichloride supported Ziegler–Natta catalysts. In *Ziegler Catalysts. Recent Scientific Innovations and Technological Improvements*; Fink, G., Mülhaupt, R., Brintzinger, H.H., Eds.; Springer-Verlag: Berlin, 1995; 343–362.
58. Bukatov, G.D.; Goncharov, V.S.; Zakharov, V.A. Number of active centers and propagation rate constants in the propene polymerization on supported Ti-Mg catalysts in the presence of hydrogen. Macromol. Chem. Phys. **1995**, *196*, 1751–1759.
59. Schoenberg, E.; Marsh, H.A.; Walters, S.J.; Saltman, W.M. Polyisoprene. Rubber Chem. Technol. **1979**, *52*, 526–604.
60. Kaita, S.; Doi, Y.; Kaneko, K.; Horiuchi, A.C.; Wakatsuki, Y. An efficient gadolinium metallocene-based catalyst for the synthesis of isoprene rubber with perfect 1,4-cis microstructure and marked reactivity difference between lanthanide metallocenes toward dienes as probed by butadiene-isoprene copolymerization catalysis. Macromolecules **2004**, *37*, 5860–5862.
61. Soga, K.; Sano, T.; Yamamoto, K.; Shiono, T. The role of additives on the improvement of the isotacticity of polypropylene—a possible interpretation. Chem. Lett. (Japan) **1982**, 425–428.
62. Van der Ven, S. *Polypropylene and other Polyolefins. Polymerization and Characterization*; Elsevier: Amsterdam, 1990.
63. Cecchin, G.; Collina, G.; Covezzi, M. Polybutene-1 (co)polymers and process for their preparation. US Patent 6,306,996, March 4, 1999
64. Lopez, L.C.; Wilkes, G.L.; Stricklen, P.M.; White, S.A. Synthesis, structure and properties of poly(4-methyl-1-pentene). J. Macromol. Sci., Rev. Macromol. Chem. Phys. **1992**, *C32*, 301–406.
65. Kashiwa, N.; Yoshitake, J. Polymerizations of α-olefins and styrene with $MgCl_2$-supported titanium catalyst system: $MgCl_2$/ $TiCl_4$/$PhCO_2Et$ with $AlEt_3$/$PhCO_2Et$. Polym. Bull. (Berlin) **1984**, *11*, 485–489.
66. Sacchi, M.C.; Fan, Z.-Q.; Forlini, F.; Tritto, I.; Locatelli, P. Use of different alkoxysilanes as external donors in $MgCl_2$-supported Ziegler–Natta catalysts to obtain propene/1-butene copolymers with different microstructure. Macromol. Chem. Phys. **1994**, *195*, 2805–2816.
67. Kashiwa, N.; Fukui, K. Process for production of 4-methyl-1-pentene polymer or copolymer. US Patent 4,659,792, December 23, 1985
68. Kaminsky, W. The discovery of metallocene catalysts and their present state of the art. J. Polym. Sci.: Part A: Polym. Chem. **2004**, *42*, 3911–3921.
69. Hlatky, G.G. Heterogeneous single-site catalysts for olefin polymerization. Chem. Rev. **2000**, *100*, 1347–1376.
70. Makio, H.; Kashiwa, N.; Fujita, T. FI Catalysts: a new family of high performance catalysts for olefin polymerization. Adv. Synth. Catal. **2002**, *344*, 477–493.
71. Tshuva, E.Y.; Goldberg, I.; Kol, M. Isospecific living polymerization of 1-hexene by a readily available nonmetallocene C_2-symmetrical zirconium catalyst. J. Am. Chem. Soc. **2000**, *122*, 10706–10707.
72. Busico, V.; Cipullo, R.; Ronca, S.; Budzelaar, P.H.M. Mimicking Ziegler–Natta catalysts in homogeneous phase, 1: C_2-symmetric octahedral Zr(IV) complexes with tetradentate [ONNO]-type ligands. Macromol. Rapid Commun. **2001**, *22*, 1405.

Index

Volume 1: Pages 1–684; *Volume 2:* Pages 685–1366; *Volume 3:* Pages 1367–2002; *Volume 4:* Pages 2003–2662; *Volume 5:* Pages 2663–3260.

Volume 1

Encyclopedia of Chemical Processing

EDITED BY
Sunggyu Lee

about the book...

Collecting information of vital interest to chemical, polymer, electrical, mechanical, and civil engineers, as well as chemists and chemical researchers, this Encyclopedia supplies nearly 350 articles on current design, engineering, and manufacturing practices—offering expertly written articles on technologies at the forefront of the field to maximize and enhance the research and production phases of current and emerging chemical manufacturing practices and techniques.

A complete reconceptualization of the classic reference series—*the Encyclopedia of Chemical Processing and Design,* whose first volume published in 1976—this resource offers extensive A-Z treatment of the subject in five simultaneously published volumes...completely new material on the design of key unit operations involved with chemical processes; the design, unit operation, and integration of reactors and separation systems; process system peripherals such as pumps, valves, and controllers; analytical techniques and equipment; and pilot plant design and scale-up criteria...comprehensive indexing of all five volumes to appear in the backmatter of each tome...well-researched sections on automation, equipment, design and simulation, reliability and maintenance, separations technologies, and energy and environmental issues...authoritative contributions on chemical processing equipment, engineered systems, and laboratory apparatus currently utilized in the field... expert overviews on key engineering science topics in property predictions, measurements and analysis, novel materials and devices, and emerging chemical fields...and gold standard articles on the methods, practices, products, and standards recently influencing the chemical industries.

about the editor...

SUNGGYU LEE is the C.W. LaPierre Professor and Chairman of the Department of Chemical Engineering at the University of Missouri–Columbia. He is the author or coauthor of several books and over 400 archival publications. He received 20 U.S. patents in fields of chemical process technologies. A specialist in chemical reaction engineering and process analysis, and an active member of the American Institute of Chemical Engineers, Dr. Lee has designed more than 25 pilot, commercial, and demonstration plants, and advised companies such as B. F. Goodrich, Water Technologies Limited, and Northern Technology International. He received the B.S. (1974) and M.S. (1976) degrees in chemical engineering from Seoul National University, Korea, and the Ph.D. degree (1980) in chemical engineering from Case Western Reserve University, Cleveland, Ohio.

DK500X

ISBN 0-8247-5500-6

90000

9 780824 755003

Taylor & Francis
Taylor & Francis Group
www.taylorandfrancisgroup.com

270 Madison Avenue
New York, NY 10016

2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK